Notes on Elementary Probability

Liviu I. Nicolaescu
University of Notre Dame

Last modified on December 5, 2019.

Notes for the undergraduate probability class at the University of Notre Dame.
Started November 25, 2015. Completed January 3, 2016.
Last modified on **December 5, 2019**.

Introduction

These are notes for the undergraduate probability course at the University of Notre Dame.[1] It covers the topics required for the actuaries Exam P.

Teaching this class I discovered that when first encountering probability it is more productive to learn how to use the main theoretical results than knowing their proofs. For this reason there are very few "theoretical" proofs in this text. Instead, we illustrate each important concept with many we hope illuminating examples. In particular, we have included more that 160 exercises, of varied difficulty, and their complete solutions.

Probability comes alive during simulations. The book contains a very basic introduction to R and the codes of several R simulations I have presented in class.

I have taught this course for several years and I have incorporated in this book many of the remarks and questions I received from my students. In particular, I used the explanations and arguments that seem to resonate the most with my audiences. The various parts written in fine print represent topics that I covered in class only if I had enough time.

[1] Started November 25, 2015. Completed January 3, 2016. Last modified on December 5, 2019.

Notation

- We denote by $\mathbb{R}$ the set of real numbers.
- We denote by $\mathbb{Z}$ the set of integers
$$\mathbb{Z} = \{0, \pm 1, \pm 2, \dots\}.$$
- We denote by $\mathbb{N}$ the set of natural numbers,
$$\mathbb{N} = \{1, 2, \dots\}.$$
- We denote by $\mathbb{N}_0$ the set of nonnegative integers
$$\mathbb{N}_0 = \{0\} \cup \mathbb{N} = \{0, 1, 2, \dots\}.$$
- $x \gg y$ signifies that x is a lot bigger than y
- A notation such as $x :=$ *bla-bla-bla* or *bla-bla-bla* $=: x$ indicates that the symbol x denotes the quantity defined to be whatever *bla-bla-bla* means. E.g. $\sqrt{2} :=$ *the positive number whose square is* 2
- $\forall$ signifies *for any*, *for all* etc.
- $\exists$ signifies *there exists*, *there exist.*
- $\Rightarrow$ stands for the term *implies.*
- i.i.d. = *independent identically distributed*

The Greek Alphabet

A	α	Alpha	N	ν	Nu
B	β	Beta	Ξ	ξ	Xi
Γ	γ	Gamma	O	o	Omicron
Δ	δ	Delta	Π	π	Pi
E	ε	Epsilon	P	ρ	Rho
Z	ζ	Zeta	Σ	σ	Sigma
H	η	Eta	T	τ	Tau
Θ	θ	Theta	Υ	υ	Upsilon
I	ι	Iota	Φ	φ	Phi
K	κ	Kappa	X	χ	Chi
Λ	λ	Lambda	Ψ	ψ	Psi
M	μ	Mu	Ω	ω	Omega

Contents

Chapter 1

Sample spaces, events and probability

1.1. Probability spaces

Here are some examples of chance experiments/phenomena to have in mind. We will discuss more sophisticated ones as we progress in our investigation of probability.

(i) Flip a coin once. The outcome is either **H**eads or **T**ails, and it is not predictable.

(ii) Flip a coin twice. The possible outcomes are HH, HT, TT, TH, and again, they are not predictable.

(iii) Roll a die. The possible outcomes are $1, 2, 3, 4, 5, 6$, but they are not predictable.

(iv) Roll a pair of *distinguishable* dice, say a red die and a green die. The possible outcomes are the pairs

$$(n_1, n_2), \;\; n_1, n_2 \in \{1, 2, 3, 4, 5, 6\}.$$

(v) The number of light bulbs that need to be replaced during a fixed time period (say 5 years) is an unpredictable quantity which could take any value $0, 1, 2, \ldots$.

(vi) The life span of a light bulb or a machinery is an unpredictable quantity which could be any nonnegative number.

(vii) The amount of damages an insurance company must pay over a calendar year is an unpredictable quantity that can take any value in $[0, \infty)$

(viii) Throw a dart at a circular dartboard of given radius $r > 0$. The unpredictable outcome could be any point in the disc

$$D_r := \{\, (x,y) \in \mathbb{R}^2;\ \ x^2 + y^2 \leq r^2 \,\}.$$

Roughly speaking, the *sample space* of a random experiment/phenomenon is the set S of all possible outcomes of that experiment. For example, the sample space in the above experiments are

- (i) → $\{H, T\}$,
- (ii) → $\{HH, HT, TT, TH\}$,
- (iii) → $\{1, 2, 3, 4, 5, 6\}$
- (iv) → $\{\, (n_1, n_2);\ \ n_1, n_2 = 1, 2, 3, 4, 5, 6 \,\}$,
- (v) → $\{0, 1, 2, \dots\}$,
- (vi) → $[0, \infty)$,
- (vii) → $[0, \infty)$,
- (viii) → D_r.

Often, when dealing with chance phenomena, we are interested only in certain events. E.g., we may want to know if the sum of observed numbers is 7. This can happen if and only if the outcome of the roll belongs to the set

$$S_7 = \{\, (1,6), (2,5), (3,4), (4,3), (5,2), (6,1) \,\}.$$

In general, we define an *event* of a random experiment to be *a subset of the sample space*. The sample space S itself is an event called the *sure event*. The empty set $\emptyset$ is called the *impossible* event.

In concrete situations the events are described by properties:

- *the event of flipping a coin three times and obtaining at least two heads.* This corresponds to the subset $\{HHH, HHT, HTH, THH\}$ of the sample space.
- *the event that the damages paid by an insurance company are bigger than a given threshold* etc.

Formally, events are sets and, as such, we can operate with them. These set operations have linguistic counterparts.

- The *union* $A \cup B$ of sets corresponds to the linguistic OR, "*A or B*". A word of warning. *This is not an exclusive "OR"*, meaning that A could happen, B could happen or both A and B could happen.
- The *intersection* $A \cap B$ of sets corresponds to the linguistic AND "*A and B*".
- The *complement* A^c of a set A corresponds to the linguistic negation NOT, "*not A*".

- The *set difference* $A \setminus B$ corresponds to the linguistic "*A but not B*".
- The empty set $\emptyset$ corresponds to the linguistic *impossible event*

Example 1.1. Best of 7 final, Boston Ruins vs. Montreal Canadiens. For $k = 1, \dots, 7$ we define B_k to be the event: *Boston wins* game k. The event "*Boston loses game* 1*, but wins game* 2 *and* 3" is described mathematically by the set

$$B_1^c \cap B_2 \cap B_3.$$

The event "*Boston wins the series with at most one loss*" is encoded mathematically by the set

$$\begin{gathered}\big(B_1 \cap B_2 \cap B_3 \cap B_4\big) \cup \big(B_1^c \cap B_2 \cap B_3 \cap B_4 \cap B_5\big) \cup \big(B_1 \cap B_2^c \cap B_3 \cap B_4 \cap B_5\big)\\ \cup\big(B_1 \cap B_2 \cap B_3^c \cap B_4 \cap B_5\big) \cup \big(B_1 \cap B_2 \cap B_3 \cap B_4^c \cap B_5\big).\end{gathered}$$ □

Definition 1.2. Consider a random experiment with sample space S. A *probability function* or *probability distribution* associated to this experiment is a function that assigns to each event $E \subset S$ a real number $\mathbb{P}(E)$, called *the probability of* E, satisfying the following properties.

(i) (*Positivity*)

$$0 \leq \mathbb{P}(E) \leq 1, \;\; \forall E \subset S.$$

(ii) (*Normalization*) $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(S) = 1$.

(iii) (*Countable additivity*) If $E_1, E_2, \dots$ is a sequence of pairwise disjoint events,

$$E_i \cap E_j = \emptyset, \;\; \forall i \neq j,$$

then

$$\mathbb{P}\Big(\bigcup_{n\geq 1} E_n\Big) = \sum_{n\geq 1} \mathbb{P}(E_n).$$

A *probability space* is a pair $(S, \mathbb{P})$ consisting of a sample space S and a probability distribution $\mathbb{P}$ on S. □

Remark 1.3. (a) The above definition of probability function is too restrictive, but it captures the main features of the modern concept of probability. The modern period of probability theory begins with the 1933 groundbreaking monograph [11] of the Russian mathematician A.N. Kolmogorov. However, as W. Feller superbly demonstrates in his gem [5], one can still ask and answer many interesting questions without a full adoption of Kolmogorov's point of view.

(b) The way one associates a probability distribution to a random phenomenon is based on empirical data and/or "reasonable" assumptions. The philosophical meaning of the concept of probability is still being debated.[1] Statements such as

[1] There are essentially to views on probability, the *frequentist* view
https://en.wikipedia.org/wiki/Frequentist_probability, and *Bayesian* view
https://en.wikipedia.org/wiki/Bayesian_probability

"*the probability of getting tails when flipping a fair coin is* 50%" can be understood as saying that if we flip a coin many, many times, then roughly half the time we will get tails. However, statements such as "*there is a 30% chance of rain tomorrow*" may have different meanings to different people.

The probability of an event can be viewed as measuring the "amount of information" we have about that event. □

Definition 1.4. An event E in a probability space $(S, \mathbb{P})$ is called *almost sure* (a.s.) if $\mathbb{P}(E) = 1$. An event E is called *improbable* if $\mathbb{P}(E) = 0$. □

Example 1.5. (a) Let us associate a probability function to the experiment of rolling one "fair" die. The attribute "fair" is meant to indicate that all the possible 6 outcomes are "equally likely" so each should have a probability of 1 in 6 of occurring. In this case

$$S = \{1, \dots, 6\},$$

and for every event $E \subset S$ we have $\mathbb{P}(A) = \frac{|E|}{6} = \frac{\#E}{6}$, where $|E|$ or $\#E$ denote the cardinality of E, the number of elements of the set E.

One can simulate rolling a die on a computer. For example, to simulate 30 consecutive rolls of a die one can use the following R command[2]

```
sample(1:6, 30, replace=TRUE)
```

Intuitively, if we roll a die a very large number of times (say 6 million times), then would should expect that the number 1 will show up roughly one sixth of the time, i.e., "close" to one million times; see Figure 1.1.

(b) If the sample space S is finite and consists of N elements, then the *uniform probability distribution* is the probability distribution $\mathbb{P}_{\text{unif}}$ such that all the elementary outcomes are equally likely, i.e.,

$$\mathbb{P}_{\text{unif}}\big(\{s\}\big) = \frac{1}{N}, \;\; \forall s \in S.$$

In this case, for any event $E \subset S$ we have

$$\mathbb{P}_{\text{unif}}\big(E\big) = \frac{\#E}{N}.$$

Intuitively, $\mathbb{P}_{\text{unif}}\big(E\big)$ represents the fraction of the sample space occupied by the event (subset) E.

(c) In general, if the sample space S is discrete, i.e., finite or countable, then we can produce probability functions on S as follows. Choose a function (weight) $w : S \to (0, \infty)$ such that

$$Z_w := \sum_{s \in S} w(s) < \infty.$$

[2] Try this R command on your computer and see how many 6-s you get.

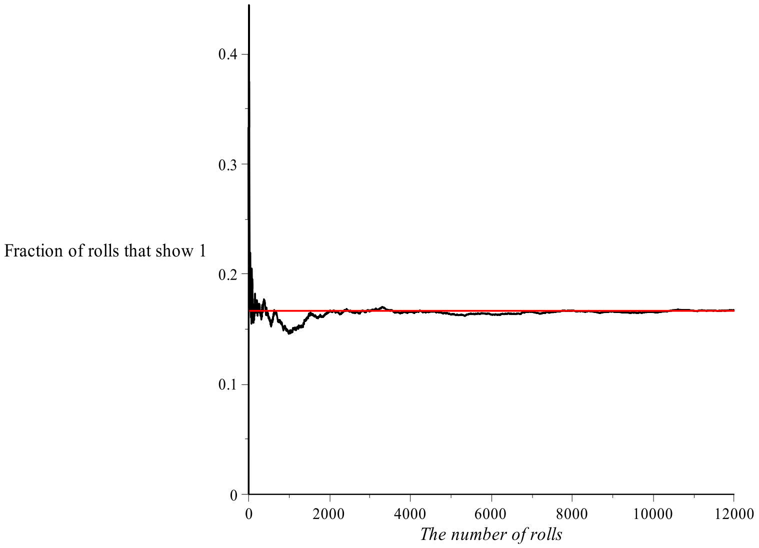

Figure 1.1. *Simulating* 12000 *rolls of a fair die and recording the frequency of* 1*'s. The horizontal line at altitude* 1/6 *is the theoretically prescribed probability of getting* a 1.

Define

$$p_w(s) := \frac{1}{Z_w} w(s), \quad \forall s \in S,$$

and think of $p_w(s)$ as the probability of the outcome s occurring. The probability of an event $X \subset S$ occurring is then

$$\mathbb{P}_w(X) = \sum_{x \in X} p_w(x) = \frac{1}{Z_w} \sum_{x \in X} w(x).$$

Note that,

$$\mathbb{P}_w(\{s\}) = p_w(s) = \frac{w(s)}{Z_w}$$

so, the larger $w(s)$, the more likely is the event $\{s\}$ will occur.

When S is a finite set consisting of N elements and $w(s) = 1$, $\forall s \in S$, the resulting probability function is the uniform probability distribution. Later we will discuss various other weights w that appear in concrete problems.

(d) If the sample space is a compact interval $S = [a, b]$, then the *uniform probability distribution* $\mathbb{P}_{unif}$ on S associates to an event $A \subset [a, b]$ the "fraction of the length of $[a, b]$ occupied by A",

$$\mathbb{P}_{\text{unif}}(A) = \frac{\text{total length}\,(A)}{\text{length}\,(S)} = \frac{\text{total length}\,(A)}{(b-a)}.$$

For example, if $S = [-1, 2]$, then the probability that a uniform random number in $[-1, 2]$ is negative, is $1/3$. Note that the probability of the event "a uniform random number in this interval is equal to 0.5” is 0. Thus, *the event that a random number in this interval has a precise given value is possible, but improbable.*

Similarly, if S is a region in the plane such as a disk or a square, then the *uniform probability distribution* $\mathbb{P}_{\text{unif}}$ on S associates to an event $A \subset [a,b]$ the “fraction of the area of S occupied by A”,

$$\mathbb{P}_{\text{unif}}(A) = \frac{\text{total area}\,(A)}{\text{area}\,(S)}.$$

Suppose that we throw at random a dart at a circular board, and all points are equally likely to be hit. This means that the probability of hitting a given region inside the board is proportional to its area. In particular, the probability of hitting the center is 0, so almost surely, we will never hit the center. Hitting the center is an *improbable* event, yet it is *not* an impossible event. □

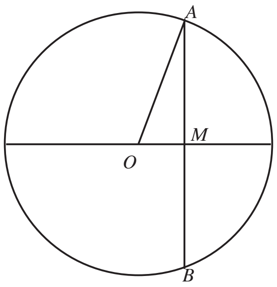

Figure 1.2. *The length of a chord AB on a circle is determined by the distance to the center O of its midpoint M.*

Example 1.6 (Bertrand’s “Paradox”)**.** Let us find the probability that a *random chord* of a circle of unit radius has a length greater than $\sqrt{3}$, the side of an inscribed equilateral triangle. Let us describe two possible solutions to this problem.

Solution 1. The length of the chord depends only on its distance from the center of the circle and not on its direction. For the chord to have length $> \sqrt{3}$, the distance from the center of the circle to the chord must be $< \frac{1}{2}$. If this distance is chosen uniformly in the interval $[0,1]$, we deduce that the sought probability is $\frac{1}{2}$.

Solution 2. Any cord is uniquely determined by its center. Assume that its midpoint is uniformly distributed in the unit circle. For the chord to have length $> \sqrt{3}$, its midpoint must be located within of disk of radius $1/2$ centered at the origin. The area of this disk is $\frac{\pi}{4}$ and occupies $\frac{1}{4}$ of the area of the disk of radius 1. Thus the sought probability is $\frac{1}{4}$.

One question jumps at us. Which of the two solutions above is the correct one? The answer is: *both of them are correct*! The reason is that in the initial formulation of our question the concept of *random* chord was not specified. There are different natural choices of randomness when sampling chords, leading to different answers. This example shows the need to describe precisely the concept of randomness used in a concrete situation. □

Here are a few useful consequences of the properties of a probability distribution.

Proposition 1.7. *If $\mathbb{P}$ is a probability function on a sample space S, the the following hold.*

(i) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$, $\forall A \subset S$.

(ii) $\mathbb{P}(A \setminus B) = \mathbb{P}(A) - \mathbb{P}(A \cap B)$, $\forall A, B \subset S$.

(iii) *(Inclusion-Exclusion Principle)*

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B), \;\; \forall A, B \subset S. \tag{1.1}$$

(iv) *(DeMorgan)*

$$\mathbb{P}(A^c \cap B^c) = 1 - \mathbb{P}(A \cup B). \tag{1.2}$$

(v) $\forall A, B \subset S$, $A \subset B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$.

□

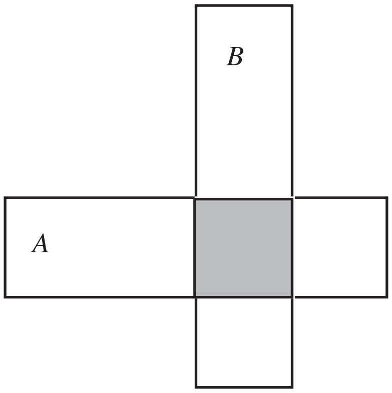

Figure 1.3. *The area of the union $A \cup B$ is the area of A + the area of B − the area of the overlap $A \cap B$.*

Proof. The equality (i) follows from the disjoint union $S = A \cup A^c$ so

$$1 = \mathbb{P}(S) = \mathbb{P}(A) + \mathbb{P}(A^c).$$

We deduce

$$0 \leq \mathbb{P}(A^c) = 1 - \mathbb{P}(A) \Rightarrow \mathbb{P}(A) \leq 1.$$

To prove (ii) note from Figure 1.3 that we have a disjoint union $A = (A \backslash B) \cup (A \cap B)$. Hence

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

The equality (iii) is proved in a similar fashion. We have a disjoint union

$$A \cup B = (A \setminus B) \cup (A \cap B) \cup (B \setminus A),$$

so that

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

The De Morgan law follows from the equality

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

The last equality follows from the fact that A and $B \setminus A$ are disjoint and $B = A \cup (B \setminus A)$. □

Example 1.8. If the chance of raining on Saturday is 50% and the chance of raining on Sunday is 50% , can one conclude that the chance of raining during the weekend is 100%?

Define the events A = "it will rain on Saturday", B = "it will rain on Sunday". Then the event "it will rain during the weekend" is $A \cup B$, and the inclusion-exclusion principle implies

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$= 0.5 + 0.5 - P(A \cap B) = 1 - P(A \cap B).$$

This shows that one cannot conclude that $P(A \cup B) = 1$. It shows that if $\mathbb{P}(A \cap B) > 0$, i.e., if the probability that it will rain on both Saturday and Sunday is positive, then the probability that it will rain on weekend is < 1. □

The inclusion-exclusion formula applies to more general situations. Given three events A_1, A_2, A_3, then

$$\begin{aligned} \mathbb{P}(A_1 \cup A_2 \cup A_3) = {} & \mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3) \\ & -\mathbb{P}(A_1 \cap A_2) - \mathbb{P}(A_1 \cap A_3) - \mathbb{P}(A_2 \cap A_3) \\ & +\mathbb{P}(A_1 \cap A_2 \cap A_3). \end{aligned} \tag{1.3}$$

Indeed, using (1.1) we deduce

$$\begin{aligned} \mathbb{P}\big(\, A_1 \cup A_2 \;\cup\; A_3 \,\big) &= \mathbb{P}(A_1 \cup A_2) + \mathbb{P}(A_3) - \mathbb{P}\big(\, (A_1 \cap A_2) \cap A_3\big) \\ &= \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2) + \mathbb{P}(A_3) - \mathbb{P}\big(\, (A_1 \cap A_3) \cup (A_2 \cup A_3)\,\big) \\ &= \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2) + \mathbb{P}(A_3) - \mathbb{P}(\, A_1 \cap A_3 \,) - \mathbb{P}(\, A_2 \cap A_3 \,) + \mathbb{P}(\, A_1 \cap A_2 \cap A_3 \,). \end{aligned}$$

More generally, if $A_1,\dots,A_n$ are n events, then we have the most general *inclusion-exclusion formula*

$$\boxed{\begin{aligned}\mathbb{P}\big(\,A_1\cup A_2\cup\cdots\cup A_n\,\big) = &\sum_{i=1}^{n}\mathbb{P}(A_i)\\ &-\sum_{1\leq i<j\leq n}\mathbb{P}(A_i\cap A_j)\\ &+\sum_{1\leq i<j<k\leq n}\mathbb{P}(A_i\cap A_j\cap A_k)\\ &\cdots\end{aligned}} \tag{1.4}$$

A sequence of events $(A_n)_{n\in\mathbb{N}}$ is called *increasing* if

$$A_1\subset A_2\subset\cdots.$$

In this case we set

$$\lim_{n\to\infty}A_n:=\bigcup_{n\in\mathbb{N}}A_n.$$

A sequence of events $(A_n)_{n\in\mathbb{N}}$ is called *decreasing* if

$$A_1\supset A_2\supset\cdots.$$

In this case we set

$$\lim_{n\to\infty}A_n:=\bigcup_{n\in\mathbb{N}}A_n.$$

From the countable additivity of a probability function we deduce immediately the following useful fact.

Proposition 1.9. *If $(A_n)_{n\in\mathbb{N}}$ is either an increasing sequence of events, or a decreasing sequence of events then*

$$\mathbb{P}\Big(\lim_{n\to\infty}A_n\Big)=\lim_{n\to\infty}\mathbb{P}(A_n). \tag{1.5}$$

□

1.2. Finite sample spaces and counting

Suppose that the sample space S consists of n elements,

$$S:=\{s_1,\dots,s_n\}.$$

Denote by $\mathbb{P}$ the uniform probability distribution on S. In this case, all outcomes are equally likely and the probability of an event A is given by the classical formula

$$\boxed{\mathbb{P}(A)=\frac{\#A}{n}=\frac{\text{the number of favorable outcomes}}{\text{the number of possible outcomes}}}. \tag{F/P}$$

Above, an outcome is called *favorable to the event* A if it belongs to the set A. Computing the probability of an event with respect to the discrete uniform distribution reduces to a *counting problem*.

Example 1.10. Consider a randomly chosen family with three children. What is the probability that they have exactly two girls? Here we tacitly assume that all distributions of genders among the three children are equally likely.

To decide this, let us first introduce the symbols b for boy, and g for girl. We first compute all the possible outcomes or gender distributions. Such an outcome is encoded by a string of three b's or g's arranged in decreasing order of their ages or, equivalently, in the order they were born. There are 8 possible outcomes

$$bbb, bbg, bgb, \boxed{bgg}, gbb, \boxed{gbg}, \boxed{ggb}, ggg.$$

Above, we have boxed the favorable outcomes, so the the probability that there are exactly 2 girls is $\frac{3}{8}$.

We want to point out that this computation is based on a non-mathematical assumption, namely that the probability of having a male offspring is equal to the probability of having a female offspring. This is more or less true for the human species, but it is not necessarily true for other species.

This problem involved rather small sample spaces. If the sample space is larger, the problem gets more complicated. Think of the related problem, that of a probability that a family with six children has exactly two girls? The techniques we will develop in this section will describe a few simple principles that will allow us to answer such question in an organized fashion. □

Theorem 1.11 (Multiplication principle)**.** *If we perform, in order, r experiments so that the number of possible outcomes of the k-th experiment is n_k, then the number of possible outcomes of this ordered sequence of experiments is $n_1 n_2 \cdots n_r$.* □

Example 1.12. (a) Suppose that we roll a die 3 times. For each roll there are 6 possible outcomes so the total number of possible outcomes is $6^3 = 216$. Each outcome is a triplet (i, j, k), $i, j, k \in \{1, \dots, 6\}$.

(b) Up there in the Sky there is an inexhaustible box containing baby boys and baby girls. Every time The Stork (see Figure 1.4) gets an order for a baby, she picks a baby at random from the Box-up-in-the-Sky, and both genders are equally likely to be picked up. Every order has thus two equally likely outcomes. The multiplication principle shows that if a family orders successively 6 babies, there are $2^6 = 64$ possible outcomes (gender distributions). □

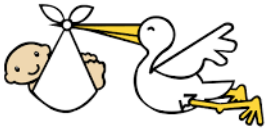

Figure 1.4. *How many baby girls in a family with 6 kids?*

Example 1.13 (Sampling with replacement)**.** We have an urn containing n balls labeled 1 through n. A *sampling with replacement* is the experiment consisting of

- extracting one ball at random from the urn,
- recording the label of the extracted ball,
- and then placing the extracted ball back in the urn.

Suppose that we perform, *in order*, k samplings with replacement. The outcome of such an ordered sequence of experiments is an ordered list of integers

$$(\ell_1, \ell_2, \dots, \ell_k),\ \ 1 \leq \ell_i \leq n.$$

The number of possible outcomes is thus n^k. In Example 7.4 we explain how to simulate in R the samplings with replacement. □

Example 1.14 (Sampling without replacement)**.** We have a box containing n balls labeled 1 through n. A *sampling without replacement* consists of

- extracting one ball at random from the urn,
- recording the label of the extracted ball,
- and then throwing away the extracted ball away.

Suppose that we perform, *in order*, k samplings without replacement. The first sampling without replacement has n possible outcomes. The second sampling without replacement has $n-1$ possible outcomes, because when we sample the urn for the second time, there are only $n-1$ balls left. The third sampling without replacement has $n-2$ possible outcomes. The k-th sampling without replacement has $n-(k-1) = n-k+1$ possible outcomes. The outcome of k successive samplings without replacement is called an *arrangement* of k objects out of n (possible objects). Thus the number of arrangements of k objects out of n is the total number of possible outcomes of an ordered sequence of k samplings without replacement is

$$\boxed{A_{k,n} = n(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!}}, \tag{1.1}$$

where, for any nonnegative integer m, we set

$$m! := \begin{cases} 1, & m = 0, \\ 1\cdot 2\cdots m, & m > 0. \end{cases}$$

The number $m!$ is called m *factorial.* For later usage, we introduce the notation

$$\boxed{(x)_k := x(x-1)\cdots(x-k+1),\ \ \forall x \in \mathbb{R}}.$$

This function is usually referred to as the *falling factorial* or the *Pochhammer symbol.*[3]

Thus the number of k samplings without replacements of n labeled objects is $A_{k,n} = (n)_k$. Note that

$$(10)_3 = 10\cdot 9\cdot 8,\ \ (22)_5 = \underbrace{22\cdot 21\cdot 20\cdot 19\cdot 18}_{\text{Decreasing 5 consecutive numbers starting at 22}}.$$

In Example 7.6 it is explained how to use R to compute $(n)_k$ and to simulate samplings without replacement. □

Example 1.15 (Lottery)**.** The country of Utopia organizes a lottery. The organizers use an urn containing balls labeled 0 to 99. Five balls are successively drawn. The winner is the person that guesses all the numbers in the order they were drawn. The odds of winning this lottery are 1 in $(100)_5 = 9,034,502,400$, roughly 1 in 9 billion.

By comparison, the odds of dying due to an asteroid impact are[4] 1 in 79 million, about 100 times higher. According to the National Safety Council[5], in 2016, the odds of an American being hit by lightning were 1 in $175,000$ (more than 50 thousand times higher than winning the lottery). The odds death due to an air incident were 1 in 9700, while the odds of an American dying due to firearm discharge were about 1 in 8000. The odds of death by firearm assault were 1 in 358, while the odd of death from heart disease or cancer were 1 in 7.□

Example 1.16 (The birthday problem)**.** We want to find the probability that, in a group of k people labeled 1 through k, selected at random, there are two people born on the same day of the year. We plan to use the formula (F/P). We

[3] Some authors denote the Pochhammer symbol $(x)_k$ by $(x)^{\underline{k}}$.

[4] *A crash course in probability*, The Economist, Jan.29, 2015 http://www.economist.com/blogs/gulliver/2015/01/air-safety.

[5] http://www.nsc.org/learn/safety-knowledge/Pages/injury-facts-chart.aspx

assume that a year consists of 365 days (so we neglect leap years) and, moreover, a person is equally likely to be born on any day of the year.[6]

A possible outcome consists of an ordered list of k numbers in the set

$$\{1, 2, \dots, 365\}.$$

This is precisely the outcome of k samplings with replacement from a box with 365 labeled balls: the first extracted ball gives the birthday of the first person in the group, the second extracted ball gives the birthday of the second person in the group etc. Thus, the possible number of outcomes is 365^k.

Denote by A_k the event "*at least two people in the group have the same birthday*". Its complement is the event A_k^c, "*no two persons in the group have the same birthday*". Then

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

so it suffices to count the number of outcomes favorable to the event A^c. This is equal with the number of outcomes of an ordered string of k samplings *without* replacement from a box with 365 birthdays. This number is $365 \cdot 364 \cdots (365 - k + 1)$. If we set $q_k := \mathbb{P}(A_k^c)$, then

$$q_k = \frac{365 \cdot 364 \cdots (365 - k + 1)}{365^k} = \frac{365}{365} \cdot \frac{364}{365} \cdots \frac{365 - (k-1)}{365}$$

$$= \left(1 - \frac{0}{365}\right)\left(1 - \frac{1}{365}\right) \cdots \left(1 - \frac{k-1}{365}\right) = \prod_{j=0}^{k-1}\left(1 - \frac{j}{365}\right).$$

Hence

$$p_k = \mathbb{P}(A_k) = 1 - q_k = 1 - \prod_{j=0}^{k-1}\left(1 - \frac{j}{365}\right).$$

For example,

$$p_{22} \approx 0.475, \quad p_{23} \approx 0.507, \quad p_{30} \approx 0.706, \quad p_{51} \approx 0.9744.$$

Figure 1.5 depicts the dependence of p_k on k. □

Example 1.17 (Permutations)**.** A *permutation* of r objects labeled $1, \dots, r$ is a way of arranging these objects successively, one after the other. For example, there are 2 permutations of 2 labeled objects, $(1, 2)$ and $(2, 1)$, and there are 6 permutations of 3 objects

$$(1, 2, 3), \ (1, 3, 2), \ (2, 1, 3), \ (2, 3, 1), \ (3, 1, 2), \ (3, 2, 1).$$

[6] That is not really the case. The odds of being born in August are higher than the odds of being born in any other month of the year. Apparently, the least common birthday is May 22. https://www.yahoo.com/parenting/why-the-most-babies-are-born-in-the-summer-128339451207.html

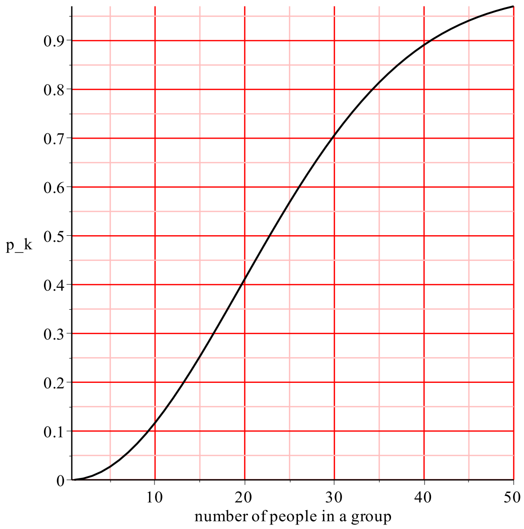

Figure 1.5. *The Birthday Problem: p_k is the probability that, in a random group of k people, at least two have the same birthday.*

Formally, a permutation of objects labeled $1, \dots, r$ is a bijection

$$\ell : \{1, \dots, r\} \to \{1, \dots, r\},$$

where $\ell(k)$ is the label of the object placed in the k-th position of the permutation. From this point of view, the object called $\ell(1)$ is placed first, followed by the object labeled $\ell(2)$ etc. Thus, the permutation $(3, 1, 2)$ corresponds to the bijection

$$1 \mapsto 3 = \ell(1), \;\; 2 \mapsto 1 = \ell(2), \;\; 3 \mapsto 2 = \ell(3).$$

If we put r labeled objects in a box, then we can obtain a permutation of these objects as follows. Extract one object from the box, record its label ℓ_1 and then put the object on the table. Extract the second object from the box that now contains $(r-1)$ objects, record its label ℓ_2, and then put this object on the table, next to ℓ_1. Continue in this fashion until the box is empty. On the table we will then have a permutation $\ell_1, \ell_2, \dots, \ell_r$ of these objects.

This shows that a permutation of r labeled objects can be viewed as a string of r samplings without replacement of these objects. The number of such strings of samplings is

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

This shows that

$$\boxed{\text{The number of permutations of } r \text{ labeled objects is } r!.}$$

For example, 5 people can be arranged in a line in $5! = 120$ ways. The factorial $r!$ grows very fast with r.

Remark 1.18. A very convenient way of estimating the size of $r!$ for large r is *Stirling's formula* [**5**, §II.9]

$$\boxed{r! \sim \sqrt{2\pi r}\left(\frac{r}{e}\right)^r \quad \text{as } r \to \infty}, \tag{1.2}$$

where we recall that the asymptotic notation $x_r \sim y_r$ as $r \to \infty$ signifies that

$$\lim_{r\to\infty} \frac{x_r}{y_r} = 1.$$

□

In Example 7.7 we explain how to use R to generate random permutations and compute $r!$. □

Example 1.19 (Combinations)**.** To understand the concept of *combination* consider the following question: how many 5-card hands can we get from a regular 52-card deck?

The problem can be recast in a more general context. Suppose that we have a box containing n labeled balls and we extract k balls *simultaneously*, $k \leq n$. The outcomes of such an extraction is called a *combination* of k objects out of n (possible objects). Two combinations are considered identical if they the sets of extracted balls are identical. We denote by $\binom{n}{k}$ or C_n^k (read n *choose* k) the number of *combinations of k objects out of n.*

We can extract these k balls successively, one by one, and at the end *forget about the order* in which they were extracted. The number of such k successive extractions is the number of arrangements of k balls out of n,

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

Two arrangements of k objects out of n can lead to the same combination because different successions of extractions could end up extracting the same balls, but in a different order. Thus, each permutation of k extracted balls is a possible outcome of a succession of k extractions. Since the number of permutations of k objects is $k!$ we deduce that for each combination of k balls there are exactly $k!$ arrangements of k balls out of n yielding that combination and therefore

$$\boxed{C_n^k = \binom{n}{k} = \frac{1}{k!}(n)_k = \frac{n!}{k!(n-k)!}.} \tag{1.3}$$

For example, the number of possible 5-card hands out of a deck of 52 is

$$\binom{52}{5} = \frac{52\cdot 51\cdot 50\cdot 49\cdot 48}{1\cdots 2\cdot 3\cdot 4\cdot 5} = \frac{52\cdot 51\cdot 50\cdot 49\cdot 48}{10\cdot 12}$$
$$= 52\cdot 51\cdot 5\cdot 49\cdot 4 = 20\cdot 52\cdot 51\cdot 49 = 2,598,960.$$

This number can be computed R using the command

```
choose(52,5)
```

The binomial coefficients can be conveniently arranged in the so called *Pascal triangle*

$$\begin{array}{cccccccccc}
\binom{0}{k}: & & & & & 1 & & & & \\
\binom{1}{k}: & & & & 1 & & 1 & & & \\
\binom{2}{k}: & & & 1 & & 2 & & 1 & & \\
\binom{3}{k}: & & 1 & & 3 & & 3 & & 1 & \\
\binom{4}{k}: & 1 & & 4 & & 6 & & 4 & & 1 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots
\end{array}$$

Observe that each entry in the Pascal triangle is the sum of the neighbors immediately above it. This translates into the equality[7]

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}, \ \ \forall 0 < k < n. \tag{1.4}$$

Thus

$$\binom{5}{3} = \frac{5!}{2!3!} = \frac{5 \cdot 4}{2!} = 10 = 6 + 4 = \binom{4}{2} + \binom{4}{3}.$$

Furthermore, the Pascal triangle is symmetric with respect to the middle vertical axis. This translates into the equality

$$\boxed{\binom{n}{k} = \binom{n}{n-k}, \ \ \forall 0 \leq k \leq n}.$$

Thus

$$\binom{52}{5} = \binom{52}{52-5} = \binom{52}{47}. \qquad \square$$

Example 1.20 (Combinations and colorings)**.** Another convenient way of interpreting combinations is through the concept of colorings.

Suppose that, after extracting the k balls, we paint them red, and then we paint black the $(n-k)$ balls left in the box. Thus, a combination of k objects out of n is a way of coloring the balls with two colors, red and black, so that k are red, and $(n-k)$ are black. The number of such colorings is therefore $\binom{n}{k}$.

Let us look at the total number of possible colorings of n balls $1, \dots, n$ with two colors, red and black, so some balls are colored red, and the other are colored black. To find this number note that a coloring is the outcome of the following

[7] Can you give a proof of (1.4) that does not rely on the equality (1.3) but instead uses the combinatorial interpretation of $\binom{n}{k}$?

succession of n experiments. Take ball 1. It must be colored with one of 2 possible colors. Make a choice. Repeat this with balls 2 through n. Since each experiment has 2 possible outcomes and we perform n experiments, we deduce from the multiplication principle that the number of possible outcomes of such successions of experiments is 2^n.

On the other hand,

$$\begin{aligned} 2^n = &\text{number of colorings with 0 red balls}\\ &+\text{number of colorings with 1 red ball}\\ &+\text{number of colorings with 2 red balls}+\cdots \end{aligned}$$

We deduce

$$2^n = \binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n} = \sum_{k=0}^{n}\binom{n}{k}. \tag{1.5}$$

The last equality can be expressed in terms of the Pascal triangle as saying that the sum of the numbers on a row of the Pascal triangle is an appropriate power of 2.

More generally, *the number of colorings of n objects with m colors*

$$c_1, \dots, c_m$$

such that n_1 objects are colored with c_1, n_2 objects are colored with c_2 etc, and $n_1 + n_2 + \cdots + n_m = n$, is

$$\boxed{\binom{n}{n_1,\dots,n_m} = \binom{n_1+\cdots+n_m}{n_1,\dots,n_m} = \frac{n!}{n_1!\cdots n_m!}}.$$

In particular,

$$\boxed{\binom{n_1+n_2}{n_1,n_2} = \binom{n_1+n_2}{n_1} = \binom{n_1+n_2}{n_2}}.$$

As an application of the above formula, suppose that we have a class consisting of 45 students. In how many we can assign grades A, B, C, D, F so that 10 students get A's, 20 students get B's, 10 students get C's, 3 students get D and 2 students get F? The answer is

$$\binom{45}{10,20,10,3,2} = \frac{45!}{10!20!10!3!2!}.$$

In Example 7.8 we explain how to use R to simulate random combinations. □

Example 1.21. Consider again the situation in Example 1.12(b), that of families that have 6 children, and we ask what is the probability that one such family has exactly two girls.

The children in such a family are labelled 1 through 6, according to the order in which they were born. We have seen that there are $2^6 = 64$ possible gender types of families with 6 children. Assigning genders to children can be viewed as "coloring" the kids 1 through 6 with one of two colors: **b**(oy) or **g**(irl). Thus the number families with exactly two girls correspond to coloring of 6 objects with two colors **b** and **g** so that exactly two objects have the color **g**. Thus number of such colorings is

$$\binom{6}{2} = \frac{6 \cdot 5}{1 \cdot 2} = 15.$$

Thus, the probability that a family with 6 children has exactly two girls is

$$\frac{15}{64} \approx 0.2343. \qquad \square$$

Example 1.22 (Newton's binomial formula)**.** Let n be a positive integer. The *Newton binomial formula* is the very useful identity

$$\boxed{(x+y)^n = \binom{n}{0}x^n + \binom{n}{1}x^{n-1}y + \binom{n}{2}x^{n-2}y^2 + \cdots + \binom{n}{n}y^n.} \tag{1.6}$$

Because of this identity, the numbers $\binom{n}{k}$ are often referred to as *binomial coefficients*.

For example, when $n = 2$ this formula reads

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

when $n = 3$ it reads

$$(x+y)^3 = x^3 + 3x^2y + 3xy^+y^3,$$

while for $n = 4$ it reads

$$(x+y)^4 = x^4 + 4x^3y + 6x^2y^2 + 4xy^3 + y^4.$$

To prove formula (1.6) note that

$$(x+y)^n = \underbrace{(x+y)\cdot(x+y)\cdots(x+y)}_{n}$$

$$= \boxed{?}x^n + \boxed{?}x^{n-1}y + \boxed{?}x^{n-2}y^2 + \cdots + \boxed{?}x^{n-k}y^k + \cdots + \boxed{?}y^n,$$

where $\boxed{?}$ stands for a coefficient to be determined.

Let us explain how we expand the power $(x+y)^n$, i.e., determine the mysterious coefficients $\boxed{?}$. We have n boxes, each containing the variables x, y

$$\underbrace{\boxed{x,y}, \boxed{x,y}, \ldots, \boxed{x,y}}_{n}$$

The terms in the expansion of $(x+y)^n$ are obtained as follows.

- From each of the above boxes extract one of the variables it contains and the multiply the extracted variables to obtain a monomial of the type $x^{n-k}y^k$.
- Do this in all the possible ways and add the resulting monomials.

We can color-code the extraction process. We color black the boxes from which we extract the variable x and red the boxes from which we extract the variable y. An extraction yields the monomial $x^{n-k}y^k$ if and only if we paint k of the boxes red, and $(n-k)$ of the boxes black. The number of such colorings is $\binom{n}{k}$. Hence, in the expansion of $(x+y)^n$, we have

$$\boxed{?}\,x^{n-k}y^k = \binom{n}{k}x^{n-k}y^k.$$

This proves Newton's formula.

Note that if in (1.6) we let $x = y = 1$, then we deduce

$$2^n = (1+1)^n = \binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \binom{n}{3}\cdots. \tag{1.7}$$

This is precisely the identity (1.5).

If in (1.6) we let $x = 1$ and $y = -1$ we deduce

$$0 = (1-1)^n = \binom{n}{0} - \binom{n}{1} + \binom{n}{2} - \binom{n}{3} + \cdots$$

If we add this to (1.7) we deduce

$$2^n = 2\binom{n}{0} + 2\binom{n}{2} + 2\binom{n}{4} + \cdots$$

so that

$$2^{n-1} = \binom{n}{0} + \binom{n}{2} + \binom{n}{4} + \cdots. \tag{1.8}$$

□

For every real number x and any nonnegative integer $k = 1, 2, \dots,$ we set

$$\boxed{\binom{x}{k} := \frac{(x)_k}{k!} = \frac{x(x-1)\cdots(x-k+1)}{k!}}.$$

Example 1.23 (Poker hands)**.** You are dealt a poker hand, 5 cards, without replacement out of 52. We refer to the site

https://en.wikipedia.org/wiki/List_of_poker_hands

for precise definitions of the various poker hands.

(i) What is the probability that you get exactly k-hearts?

(ii) What is the most likely number of hearts you will get?

(iii) What is the probability that you will get two pairs, but not a fullhouse?

Denote by p_k the probability that you get exactly k hearts. The number of possible outcomes is $\binom{52}{5}$. There are 13 hearts and $39 = 52 - 13$ non-hearts in a deck. To get k hearts means that you also get $5-k$ non-hearts. There are $\binom{13}{k}$ ways of choosing k hearts out of 13 and $\binom{39}{5-k}$ non-hearts so

$$p_k = \frac{\binom{13}{k}\binom{39}{5-k}}{\binom{52}{5}}.$$

We deduce

$$p_0 \approx 0.22,\ p_1 \approx 0.41,\ p_2 \approx 0.27,\ p_3 \approx 0.08,\ p_4 \approx 0.0107,\ p_5 \approx 0.0004.$$

Thus, you are more likely to get 1 heart.

A deck of card consists of 13 (face) values and 4 suits. If your hand consists of two pairs, but not a full house, then you have 3 values in your hand.

There are $\binom{13}{3}$ ways of choosing these values. Once these values are choses there are $\binom{3}{2}$ ways of choosing the two values that come in pairs. Once these values are chosen there are $\binom{4}{2}$ ways of chosing the suits for each pair and $\binom{4}{1}$ of choosing the suit for the single values. Thus the number of possible 2-pair-hands is

$$\binom{13}{3}\binom{3}{2}\binom{4}{2}^2\binom{4}{1} = \frac{13\cdot 12\cdot 11}{6}\cdot 3\cdot 6^2\cdot 4 = 13\cdot 2\cdot 11\cdot 12\cdot 36 = 123,552.$$

The total possible number of 5-card hands is $\binom{52}{5} = 2,598,960$ so the probability of getting two pairs is

$$\frac{123552}{2598960} \approx 0.047. \qquad \square$$

Example 1.24. An urn contains 10 Red balls, 10 White balls and 10 Blue balls. You draw 5 balls random, without replacement. What is the probability that you do not get all the colors?

Denote by R the event *no red balls*, by W the event *no white balls*, and by B the event *no blue balls*. We are interested in the probability of the event $R\cup W\cup B$. We will compute the probability of this event by using the inclusion-exclusiion principle (1.4). Hence

$$\mathbb{P}(R\cup W\cup B) = \mathbb{P}(R) + \mathbb{P}(W) + \mathbb{P}(B)$$

$$-\mathbb{P}(R\cap W) - \mathbb{P}(W\cap B) - \mathbb{P}(B\cap R) + \mathbb{P}(R\cap W\cap B).$$

Now notice a few things.

- $R\cap B =$ "all the balls are White", $W\cap B =$ "all the balls are Red", $R\cap W =$ "all the balls are Blue".
- $\mathbb{P}(R\cap W\cap B) = 0$.
- Because there are equal numbers of balls of different colors, we have

$$\mathbb{P}(R) = \mathbb{P}(W) = \mathbb{P}(B),\ \ \mathbb{P}(R\cap W) = \mathbb{P}(W\cap B) = \mathbb{P}(B\cap R).$$

Hence

$$\mathbb{P}(R \cup W \cup B) = 3\mathbb{P}(R) - 3\mathbb{P}(R \cap W)$$

To compute $\mathbb{P}(R)$ and $\mathbb{P}(R \cap W)$ we use formula (F/P). The number of possible outcomes of a five ball extraction out of 30 is $\binom{30}{5}$. The number of 5-ball extractions with no red balls is $\binom{20}{5}$, and the number of 5-ball extractions with no red or white balls is $\binom{10}{5}$. Hence

$$\mathbb{P}(R \cup W \cup B) = 3\frac{\binom{20}{5} - \binom{10}{5}}{\binom{30}{5}} \approx 0.321. \qquad \square$$

Example 1.25 (Moivre-Maxwell-Boltzmann)**.** Suppose that we randomly distribute 30 gifts to 20 people labelled 1 through 20. In doing so, some people will get more than one gift, and some people may not get any gift. What is the probability that at least one person will receive no gift.

Denote by E the event $E =$ "*at least one person does not receive any gift*". For $k = 1, 2, \dots, 20$, we denote by E_k the event "the person k does not receive any gift". Then

$$E = E_1 \cup E_2 \cup \cdots \cup E_{20}.$$

The inclusion-exclusion formula implies

$$\mathbb{P}(E) = \underbrace{\sum_{1\le i\le 20} \mathbb{P}(E_1)}_{S_1} - \underbrace{\sum_{1\le i<j\le 20} \mathbb{P}(E_i \cap E_j)}_{S_2}$$

$$+ \underbrace{\sum_{1\le i<j<j<k\le 20} \mathbb{P}(E_i \cap E_j \cap E_k)}_{S_3} - \cdots$$

The sum S_1 consists of 20 terms, one for each person. Each of these terms is equal to the probability that a *given* person receives no gift,

$$\mathbb{P}(E_1) = \cdots = \mathbb{P}(E_{20}) = \left(\frac{19}{20}\right)^{30}.$$

To see this note that there are 19^{30} ways of distributing 30 gifts among the 19 people other than the first person. Similarly, there are 20^{30} ways of distributing 30 gifts to 20 people.

In general, for $m = 1, 2, \dots, 19$, the sum S_m consists of $\binom{20}{m}$ terms, one term for each subcollection of m persons. The corresponding term is equal to the probability that no person in that subcollection receives a gift. Equivalently, this is the probability that all the 30 gifts go to the $20 - m$ people outside this subcollection. This probability is

$$\left(\frac{20-m}{20}\right)^{30}.$$

Thus

$$S_m = \binom{20}{m}\left(\frac{20-m}{20}\right)^{30}.$$

and

$$\mathbb{P}(E) = S_1 - S_2 + S_3 - \cdots = \sum_{m=1}^{19}(-1)^{m+1}\binom{20}{m}\left(\frac{20-m}{20}\right)^{30} \approx 0.9986.$$

This is a rather surprising conclusion. Although there are more gifts than persons, the probability that at least one person will receive no gift is very close to 1. □

Remark 1.26. Suppose that we distribute g gifts to N people. Then the probability that one of them will not receive a gift is

$$f(N,g) = \sum_{m=1}^{N-1}(-1)^{m+1}\binom{N}{m}\left(\frac{N-m}{N}\right)^{g}.$$

This is typically a long sum if N is large and you can use R or MAPLE to compute it. Here is a possible R implementation.

```
options(digits=12)
f<-function(N,g){
  x<-0
  for(m in 1:(N-1)){
    x<-x+ (-1)^(m+1)*choose(N,m)*(1-m/N)^g
    }
 cat("The probability that one person
 does not get a  gift  is ", x, sep="")
 }
```

The result Example 1.25 can be found using the R command

```
 f(20,30)
```

Example 1.27 (Derangements and matches)**.** This is an old and famous problem in probability that was first considered by Pierre-Remond Montmort. It is sometimes referred to as Montmort's *matching problem* in his honor. It has an amusing formulation, [**17**, II.4].

A group of n increasingly inebriated sailors on shore leave is making its unsteady way from pub to pub. Each time the sailors enter a pub they take off their hats and leave them at the door. On departing for the next pub, each intoxicated sailor picks up one of the hats at random. What is the probability p_n that no sailor retrieves his own hat?

A *derangement* is said to occur if no sailor picks up his own hat. Denote by D thus event A *match* occurs if at least one sailor picks up his own hat. Denote by M this event. Thus, we are asked what is the probability $\mathbb{P}(D)$. Note that $D = M^c$ so that

$$\mathbb{P}(D) = 1 - \mathbb{P}(M)$$

so it suffices to find the probability of a match. For $k = 1, \dots, n$ denote by M_k the event "*the k-th sailor picks up his own hat*". Clearly

$$M = M_1 \cup M_2 \cup \cdots \cup M_n.$$

To compute the probability of M we will use the inclusion-exclusion principle (1.4) to deduce that

$$\mathbb{P}(M) = \underbrace{\sum_{i=1}^n \mathbb{P}(M_i)}_{S_1} - \underbrace{\sum_{i<j} \mathbb{P}(M_i \cap M_j)}_{S_2} + \underbrace{\sum_{i<j<k} \mathbb{P}(M_i \cap M_j \cap M_k)}_{S_3} - \cdots .$$

The sum S_1 consists of n terms $\mathbb{P}(M_1), \dots, \mathbb{P}(M_n)$ and they all are equal to each other because any two sailors have the same odds of getting their own hats. (Here we tacitly assume that all sailors display similar behaviors.) Hence

$$S_1 = n\mathbb{P}(M_1).$$

The sum S_2 consists of $\binom{n}{2}$ terms, one term for each group of two sailors. These terms are equal to each other because the probability that two of the sailors pick up their own hats is equal to the probability of any other two sailors pick up their own hats. Hence

$$S_2 = \binom{n}{2}\mathbb{P}(M_1 \cap M_2).$$

Similarly, the term S_k consists of $\binom{n}{k}$ terms, one term for each group of k sailors, and these terms are equal to each other. Hence

$$S_k = \binom{n}{k}\mathbb{P}(M_1 \cap \cdots \cap M_k).$$

We deduce

$$\mathbb{P}(M) = n\mathbb{P}(M_1) - \binom{n}{2}\mathbb{P}(M_1 \cap M_2) + \binom{n}{3}\mathbb{P}(M_1 \cap M_2 \cap M_3) - \cdots .$$

To compute the probability $\mathbb{P}(M_1 \cap \cdots \cap M_k)$, i.e., the probability that the sailors $1, \dots, k$ pick up their own hats we use the formula (F/P).

The number of possible outcomes is equal to the number of permutations of n objects (hats), i.e., $n!$. The number of favorable outcomes is the number of permutations of $(n-k)$ objects (the first k hats have returned to their rightful owners). Hence

$$\mathbb{P}(M_1 \cap \cdots \cap M_k) = \frac{(n-k)!}{n!},$$

$$S_k = \binom{n}{k}\frac{(n-k)!}{n!} = \frac{n!}{k!(n-k)!} \cdot \frac{(n-k)!}{n!} = \frac{1}{k!}.$$

We deduce

$$\mathbb{P}(M) = \frac{1}{1!} - \frac{1}{2!} + \frac{1}{3!} - \cdots,$$

$$\mathbb{P}(D) = 1 - \mathbb{P}(M) = 1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \cdots = \sum_{k=0}^{n} \frac{(-1)^k}{k!}.$$

Note that as $n \to \infty$ we have

$$\mathbb{P}(D) \to \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} = \frac{1}{e} \approx 0.367.$$

Intuitively, this says that if the number of sailors is large, then the probability of a derangement is > 0.36. □

Example 1.28 (Combinations with repetitions)**.** Suppose we have k *identical* balls that we want to place in n *distinguishable* boxes labeled $B_1, \dots, B_n$. We are allowed to place more than one ball in any given box, and some boxes may not contain any ball. In how many ways can we do this?

Such a placement of balls can be encoded by a string

$$\underbrace{1, \dots, 1}_{k_1}, \underbrace{2, \dots, 2}_{k_2}, \dots, \underbrace{n, \dots, n}_{k_n}$$

where k_1 denotes the number of balls in box B_1 etc. Note that

$$k_1 + \cdots + k_n = k.$$

Such a distribution of balls is called a *combination with repetition* of k objects out of n. We denote by $\left(\!\binom{n}{k}\!\right)$ (read n *multi-choose* k) the number of such combinations.

To find the number of such combinations with repetition, imagine that we have $(n-1)$ separating vertical walls arranged successively along a horizontal line and producing in this fashion $(n-1)$ chambers $C_1, \dots, C_n$; see top of Figure 1.6. Now place k_i balls in the chamber C_i arranged successively along the line; see bottom of Figure 1.6.

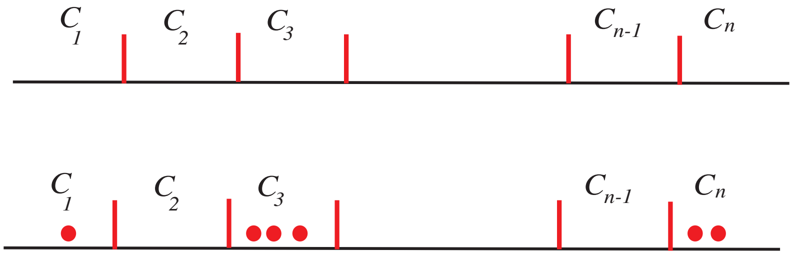

Figure 1.6. *Walls and balls.*

Along the line we have now produced $n+k-1$ points: k of these points mark the location of the balls, and $n-1$ of these points mark the locations of the walls.

Suppose now that we have $n+k-1$ points on the line arranged in increasing order

$$x_1 < x_2 < \cdots < x_{n+k-1}.$$

We can use such an arrangement of points to obtain a distribution of k balls in n chambers as follows.

- Choose $n-1$ of these points and mark them w. This is where we will place the *walls.*
- Mark the remaining points b. This is where we will place the *balls.*

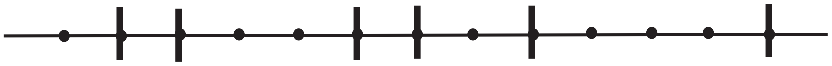

Figure 1.7. *We've placed* 7 *balls, in* 7 *chambers delimited by* 6 *walls.*

For example, in Figure 1.7 we have placed 1 ball in the first chamber, no balls in chamber 2, 2 balls in chamber 3, no balls in chamber 4, 1 ball in chamber 5, 3 balls in chamber 6, and no balls in chamber 7. Thus, a placement of k identical balls in n boxes corresponds to a choice of $n-1$ locations out of $n+k-1$ where we are to place the walls. Thus

$$\boxed{\left(\!\!\binom{n}{k}\!\!\right) = \binom{n+k-1}{n-1} = \binom{n+k-1}{k}}. \qquad \square$$

1.3. Conditional probability, independence and Bayes' formula

Example 1.29. Somebody rolls a pair of dice and you are on the lookout for the event

$$A = \text{“the outcome is a double”}.$$

Suppose you are also told that the event $B :=$“*the sum of the numbers is* 10” has occurred. What could be the probability of A given that B has occurred?

Clearly, the extra information that we have, cuts down the number of possible outcomes to

$$\{\,(4,6),(5,5),(6,4)\,\}.$$

Of these outcomes, the only one is favorable, $(5,5)$, so the probability of A given B ought to be $\frac{1}{3}$. □

1.3.1. Conditional probability. The above simple example is the motivation for the following *fundamental* concept.

Definition 1.30 (Conditional probability)**.** Suppose that $(S, \mathbb{P})$ is a probability space. If $A, B \subset S$ are two events such that $\mathbb{P}(B) \neq 0$, then *the conditional probability of A given B* is the real number $\mathbb{P}(A|B)$ defined by

$$\boxed{\mathbb{P}(A|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}}. \tag{1.9}$$

□

If A and B are as in Example 1.29, then

$$\mathbb{P}(A) = \frac{6}{36} = \frac{1}{6},\ \ \mathbb{P}(B) = \frac{3}{36} = \frac{1}{12},\ \ \mathbb{P}(A \cap B) = \frac{1}{36},$$

so that

$$\mathbb{P}(A|B) = \frac{\frac{1}{36}}{\frac{1}{12}} = \frac{12}{36} = \frac{1}{3}.$$

From the definition of conditional probability we obtain immediately the following very useful *multiplication formula*

$$\boxed{\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A)}. \tag{1.10}$$

If we think of probability as a measure of degree of belief, we can think of conditional probability as an update of that degree, in the light of new information.

Example 1.31. Alice and Bob are playing a gambling game. Each rolls one die and the person with higher numbers wins. If they tie, they roll again. If Alice just won, what is the probability that she rolled a 5?

Let A be the event "*Alice wins*" and R_i the event "*she rolls an i*". We are looking for the probability $\mathbb{P}(R_5|A)$. If we write the outcomes with Allice' s roll first, then the event A is

$$\begin{array}{ccccc}
(2,1) & (3,1) & (4,1) & (5,1) & (6,1) \\
 & (3,2) & (4,2) & (5,2) & (6,2) \\
 & & (4,3) & (5,3) & (6,3) \\
 & & & (5,4) & (6,4) \\
 & & & & (6,5)
\end{array}$$

Thus A has $1 + 2 + 3 + 4 + 5 = 15$ favorable outcomes, while $R_5 \cap A$ has only 4 favorable outcomes so that

$$\mathbb{P}(R_5|A) = \frac{\mathbb{P}(R_5 \cap A)}{\mathbb{P}(A)} = \frac{4}{15} \approx 0.266.$$ □

Example 1.32. From a deck of cards draw four cards at random, without replacement. If you get j aces, draw j cards from another deck. What is the probability of getting exactly 2 aces from each deck?

Define the events

$$A = \text{“two aces from the first deck”},$$

$$B = \text{“two aces from the second deck”}.$$

We are interested in the probability of the event $A \cap B$. Using (1.10) we deduce

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

Note that

$$\mathbb{P}(A) = \frac{\binom{4}{2}\binom{48}{2}}{\binom{52}{4}} \approx 0.0249, \;\; \mathbb{P}(B|A) = \frac{\binom{4}{2}}{\binom{52}{2}} \approx 0.0045$$

so that

$$\mathbb{P}(A \cap B) \approx 0.0001131. \qquad \square$$

Example 1.33 (The two-children paradox)**.** A family has two children. Consider the following situations

(i) One of the children is a boy.

(ii) One of the children is a boy born on a Thursday.

Let us compute, in each case the probability that both children are boys. Denote by B the event “a child is a boy”, by B_T he event “a child is a boy born on a Thursday”, by B^* the event “a child is a boy not born on a Thursday” and by G “a child is a girl”. We will we use compound events such that BG signifying the first child is a boy and the second is a girl.

(i) We are interested in the probability

$$p_i = \mathbb{P}(BB|BB \cup BG \cup GB) = \frac{\mathbb{P}(BB)}{\mathbb{P}(BB) + \mathbb{P}(BG) + \mathbb{P}(GB)}$$

$$= \frac{\frac{1}{4}}{\frac{1}{4} + \frac{1}{4} + \frac{1}{4}} = \frac{1}{3} \approx \boxed{0.333}.$$

(ii) We are interested in the probability

$$p_{ii} := \mathbb{P}\big(\, BB|B_TG \cup GB_T \cup B_TB \cup B^*B_T\,\big)$$

$$= \frac{\mathbb{P}(B_TB) + \mathbb{P}(B^*B_T)}{\mathbb{P}(B_TG) + \mathbb{P}(GB_T) + \mathbb{P}(B_TB) + \mathbb{P}(B^*B_T)}.$$

Observe that

$$\mathbb{P}(B_T) = \frac{1}{14}, \;\; \mathbb{P}(B^*) = \frac{6}{14}.$$

We deduce

$$p_{ii} = \frac{\frac{1}{14}\frac{1}{2} + \frac{6}{14}\frac{1}{14}}{2\frac{1}{14} \cdot \frac{1}{2} + \frac{1}{14}\frac{1}{2} + \frac{6}{14}\frac{1}{14}} = \frac{13}{27} \approx \boxed{0.481}.$$

The inequality $p_{ii} > p_i$ may seem surprising at a first look. Knowing that one of the children is a boy born on a Thursday seems to increase the odds that the other child is also a boy, although the above argument does not seem to distinguish between Thursday or Wednesday!

However, if we think of probability as quantifying the amount of information about an event, this inequality seems more more palatable: the information that the family has a boy born on a Thursday is much more precise than the information that the family has at least a boy and thus one can expect more accurate inferences in the second case. □

Definition 1.34. Suppose that S is a sample space and $A \subset S$ is an event. A *partition* of the event A is a collection of events $(A_n)_{n\geq 1}$ with the following properties.

(i) The events $(A_n)_{n\geq 1}$ are mutually disjoint (exclusive), i.e.,

$$A_n \cap A_m = \emptyset, \;\; \forall m \neq n.$$

(ii) The event A is the union of the events A_n,

$$A = \bigcup_{n\geq 1} A_n.$$

□

Proposition 1.35. *Suppose that S is a sample space, $\mathbb{P}$ is a probability function on S and $B \subset S$ is such that $\mathbb{P}(B) \neq 0$. Then the correspondence $A \mapsto \mathbb{P}(A|B)$ defines a probability function on S, i.e., the following hold.*

(i) $0 \leq \mathbb{P}(A) \leq \mathbb{P}(A|B) \leq 1$, *for any event $A \subset S$.*

(ii) *If $B \subset A \subset S$, then $\mathbb{P}(A|B) = 1$. In particular, $\mathbb{P}(S|B) = 1$.*

(iii) *If $(A_n)_{n\geq 1}$ is a partition of the event A, then*

$$\boxed{\mathbb{P}(A|B) = \sum_{n\geq 1} \mathbb{P}(A_n|B)}.$$

□

The fact that $A \mapsto \mathbb{P}(A|B)$ is a probability distribution implies the following equalities satisfies by all probability functions.

Corollary 1.36.

$$\boxed{\mathbb{P}(A^c|B) = 1 - \mathbb{P}(A|B), \;\; \forall A \subset S},$$

$$\boxed{\mathbb{P}(A_1 \cup A_2|B) = \mathbb{P}(A_1|B) + \mathbb{P}(A_2|B) - \mathbb{P}(A_1 \cap A_2|B)}.$$ □

1.3.2. Independence.

The notion of conditional probability is intimately related to the subtle and very important concept of independence.

Definition 1.37. Suppose that $(S, \mathbb{P})$ is a probability space. Two events $A, B \subset S$ are called *independent* and we write this $A \perp\!\!\!\perp B$, if

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

Note that if $\mathbb{P}(A) = 0$ or $\mathbb{P}(B) = 0$, then $0 \leq \mathbb{P}(A \cap B) \leq \min(\mathbb{P}(A), \mathbb{P}(B)\} = 0$ so the events A and B are independent. If $\mathbb{P}(A)\mathbb{P}(B) \neq 0$, then

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

so A, B *are independent if and only if*

$$\boxed{\mathbb{P}(A|B) = \mathbb{P}(A)}.$$

Proposition 1.38. *If the events A, B are independent, then the events A, B^c are also independent.*

Proof. We know that $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. We have

$$\mathbb{P}(A \cap B^c) = \mathbb{P}(A \setminus A \cap B) = \mathbb{P}(A) - \mathbb{P}(A \cap B) = \mathbb{P}(A) - \mathbb{P}(A)\mathbb{P}(B)$$

$$= \mathbb{P}(A)\big(1 - \mathbb{P}(B)\big) = \mathbb{P}(A)\mathbb{P}(B^c).$$

□

Example 1.39. Flip a fair coin twice and consider the events

$$A = \{\text{head in the first flip}\} = \{HT, HH\},$$

$$B = \{\text{head in the second flip}\} = \{TH, HH\},$$

$$C = \{\text{the two flips yield different results}\} = \{TH, HT\}.$$

Note that $A \cap B = \{HH\}$ so

$$\mathbb{P}(A \cap B) = \frac{1}{4} = \mathbb{P}(A)\mathbb{P}(B)$$

so the events A, B are independent. Similarly, it is easy to show that any two of the above events are independent. Hence, these events are *pairwise independent.* However, it does not seem quite right to say that the three events A, B, C are independent since C is not independent of $A \cap B$. Indeed

$$A \cap B \cap C = \emptyset$$

so $0 = \mathbb{P}(C \cap A \cap B) \neq \mathbb{P}(C)\mathbb{P}(A \cap B) = \frac{1}{8}$. □

Definition 1.40 (Independence)**.** (a) The sequence of events

$$A_1, A_2, \dots, A_n, \dots$$

is called *independent* if, for any $k \geq 1$, and for any $i_1 < \cdots < i_k$ we have

$$\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \cdots \mathbb{P}(A_{i_k}).$$

(b) The sequence of *collections of events* $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_n, \dots$ is called *independent* if any sequence of events $\{A_1, A_2, \dots\}$ such that $A_n \in \mathcal{C}_n$ for any n, is independent. □

Let us observe that the sequence of events $\{A_1, A_2, \dots, A_n, \dots,\}$ is independent if and only if the sequence of collections of events

$$\{A_1, A_1^c\}, \{A_2, A_2^c\}, \dots, \{A_n, A_n^c\}, \dots$$

is independent.

Example 1.41. (a) Suppose we roll a die until the first 6 appears. What is the probability that this occurs on the n-th roll, $n = 1, 2, \dots$? The event of interest is

$$B_n = \{\text{the first 6 appears in the } n\text{-th roll}\}.$$

Consider the event

$$A_k = \{\text{the } k\text{-th roll yields a 6}\}. \tag{1.11}$$

It is reasonable to assume that the rolls are independent of each other, i.e., the result of a roll is not influenced by and does not influence other rolls. This implies that the events $A_1, \dots, A_n$ are independent. Observing that B_n occurs if during the first $(n-1)$ rolls we did not get a 6 and we got a six on the nth roll, i.e.,

$$B_n = A_1^c \cap \cdots \cap A_{n-1}^c \cap A_n$$

we deduce from the independence assumption that

$$\mathbb{P}(B_n) = \mathbb{P}(A_1^c) \cdots \mathbb{P}(A_{n-1}^c)\mathbb{P}(A_n).$$

Now observe that $\mathbb{P}(A_k) = \frac{1}{6}$. We deduce

$$\mathbb{P}(B_n) = \frac{1}{6}\left(\frac{5}{6}\right)^{n-1}.$$

(b) Suppose we roll a die 10 times. Denote by N the number of times we get a 6. What is the probability $\mathbb{P}(N = 2)$?

For $1 \leq i < j \leq j$ denote by B_{ij} the event that we get 6 at the i-th and j-th roll, and no 6 otherwise. We have

$$\mathbb{P}(B_{ij}) = \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^8,$$

because B_{ij} is the intersection of 10 independent events A_i, A_j and A_k^c, $k \neq i, j$, where A_i is described by (1.11). Two of these events have probability $\frac{1}{6}$ and the remaining 8 have probability $\frac{5}{6}$.

The event $\{N = 2\}$ is the union of the disjoint events B_{ij}, $1 \leq i < j \leq 10$. There are $\binom{10}{2}$ such events and all have the same probability. Hence

$$\mathbb{P}(N = 2) = \binom{10}{2} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^8 . \qquad \square$$

Definition 1.42. Suppose A, B, C are three events in a sample space $(S, \mathbb{P})$ such that $\mathbb{P}(C) > 0$. We say that A, B are *conditionally independent* given C if

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

□

Proposition 1.43 (Markov property)**.** *Suppose A_+, A_-, A_0 are three events in a sample space $(S, \mathbb{P})$ such that $\mathbb{P}(A_- \cap A_0) > 0$. Then A_+, A_- are conditionally independent given A_0 if and only if*

$$\mathbb{P}(A_+ | A_- \cap A_0) = \mathbb{P}(A_+ | A_0). \tag{1.12}$$

Proof. We have

$$\mathbb{P}(A_+ | A_- \cap A_0) = \frac{\mathbb{P}(A_+ \cap A_- \cap A_0)}{\mathbb{P}(A_- \cap A_0)} = \frac{\mathbb{P}(A_+ \cap A_- | A_0)\mathbb{P}(A_0)}{\mathbb{P}(A_- | A_0)\mathbb{P}(A_0)}$$

$$= \frac{\mathbb{P}(A_+ \cap A_- | A_0)}{\mathbb{P}(A_- | A_0)}.$$

We see that

$$\mathbb{P}(A_+ | A_- \cap A_0) = \mathbb{P}(A_+ | A_0) \Longleftrightarrow \frac{\mathbb{P}(A_+ \cap A_- | A_0)}{\mathbb{P}(A_- | A_0)} = \mathbb{P}(A_+ | A_0)$$

$$\Longleftrightarrow \mathbb{P}(A_+ \cap A_- | A_0) = \mathbb{P}(A_+ | A_0)\mathbb{P}(A_- | A_0).$$

□

Remark 1.44. In applications A_+ is a future event, A_0 is a present event and A_- is a past event. The Markov property is often phrased as follows

The future is independent of the past given the present if and only if the probability of the future given past and present is equal to the probability of the future given the present. □

1.3.3. The law of total probability.

Example 1.45. Alex flips a *fair* coin $n+1$ times coins and Betty flips that coin n times. Alex wins if the number of heads he gets is strictly greater than the number Betty gets. What is the probability that Alex will win?

The situation seems biased in favor of Alex since he's allowed one coin flip more than Betty so, from this perspective, it seems that the Alex' winning probability ought to be better than $1/2$.

Denote by X and respectively Y the number of Heads of Alex and respectively Betti $\underline{\textit{after } n \textit{ steps}}$. At that moment Alex has one more coin flip to go.

There are three possibilities: $X > Y$, $X = Y$ and $X < Y$. Set

$$p := \mathbb{P}(X > Y).$$

On account of symmetry,

$$\mathbb{P}(Y > X) = \mathbb{P}(X > Y) = p.$$

Hence

$$\mathbb{P}(X = Y) = 1 - 2p.$$

In the first case $X > Y$ Alex has already won. In the third case, $X < Y$, he cannot win. If we denote by A the event "*Alex wins*", then we conclude that

$$\mathbb{P}(A) = \mathbb{P}(A, X > Y) + \mathbb{P}(A, X = Y) + \underbrace{\mathbb{P}(A, X < Y)}_{=0}$$

(use multiplication formula)

$$= \mathbb{P}(A|X > Y)\mathbb{P}(X > Y) + \mathbb{P}(A|X = Y)\mathbb{P}(X = Y)$$

$$= \mathbb{P}(X > Y) + \frac{1}{2}\mathbb{P}(X = Y) = p + \frac{1}{2}(1 - 2p) = \frac{1}{2}.$$

This shows that Alex and Betty have equal chances of winning, even though Al is allowed one extra flip! □

The computations in Example 1.45 used a very simple but potent principle.

Theorem 1.46 (Law of total probability)**.** *Suppose that $(S, \mathbb{P})$ is a probability space and the events $B_1, \dots, B_n, \dots$ form a partition of S such that $\mathbb{P}(B_k) \neq 0$, $\forall k = 1, \dots, n$. Then for any event $A \subset S$ we have*

$$\boxed{\mathbb{P}(A) = \mathbb{P}(A|B_1)\mathbb{P}(B_1) + \cdots + \mathbb{P}(A|B_n)\mathbb{P}(B_n) + \cdots}. \tag{1.13}$$

Proof. We have

$$A = \big(A \cap B_1\big) \cup \big(\, A \cap B_2\,\big) \cup \cdots$$

and the sets $\big(A \cap B_1\big), \big(\, A \cap B_2\,\big), \dots$ are pairwise disjoint. Hence

$$\mathbb{P}(A) = \mathbb{P}\big(A \cap B_1\big) + \mathbb{P}\big(\, A \cap B_2\,\big) + \cdots$$

$$= \mathbb{P}\big(A|B_1\big)\mathbb{P}(B_1) + \mathbb{P}\big(\,A|B_2\,\big)\mathbb{P}(B_2) + \cdots$$

□

Remark 1.47. In concrete situations the partition $B_1, B_2, \dots, B_n, \dots$ comes in the guise of a classification by type: any outcome of the random event can only by of one and only one of the types $B_1, \dots, B_n, \dots$.

Example 1.48. Suppose that we have an urn containing b black balls and r red balls. A ball is drawn from the urn and discarded. Without knowing its color, what is the probability that a second ball drawn is black?

For $k = 1, 2$ denote by B_k the event "*the k-th drawn ball is black*". We are asked to find $\mathbb{P}(B_2)$. The first drawn ball is either black (B_1) or not black (B_1^c). From the law of total probability we deduce

$$\mathbb{P}(B_2) = \mathbb{P}(B_2|B_1)\mathbb{P}(B_1) + \mathbb{P}(B_2|B_1^c)\mathbb{P}(B_1^c).$$

Observing that

$$\mathbb{P}(B_1) = \frac{b}{b+r} \ \text{ and } \ \mathbb{P}(B_1^c) = \frac{r}{b+r},$$

we conclude

$$\mathbb{P}(B_2) = \frac{b-1}{b+r-1}\cdot\frac{b}{b+r} + \frac{b}{b+r-1}\cdot\frac{r}{b+r} = \frac{b(b-1)+br}{(b+r)(b+r-1)}$$
$$= \frac{b(b+r-1)}{(b+r)(b+r-1)} = \frac{b}{b+r} = \mathbb{P}(B_1).$$

Thus, the probability that the second extracted ball is black is equal to the probability that the first extracted ball is black. This seems to contradict our intuition because when we extract the second ball the composition of available balls at that time is different from the initial composition.

This is a special case of a more general result, due to S. Poisson, [**1**, Sec. 5.3].

> *Suppose in an urn containing b black and r red balls, n balls have been drawn first and discarded without their colors being noted. If another ball is drawn drawn next, the probability that it is black is the same as if we had drawn this ball at the outset, without having discarded the n balls previously drawn.*

To quote John Maynard Keynes[8], [**10**, p.394],

> This is an exceedingly good example of the failure to perceive that a probability cannot be influenced by the *occurrence* of a material event but only by such *knowledge* as we may have, respecting the occurrence of the event.

[8] John Maynard Keynes (1883-1946) was an English economist widely considered to be one of the most influential economists of the 20th century and the founder of modern macroeconomics. `https://en.wikipedia.org/wiki/John_Maynard_Keynes`

□

Example 1.49. Consider again the situation in Example 1.27 with n inebriated sailors. Label the sailors $S_1, \dots, S_n$ and assume that, as they exit a pub, they wait in line to pick a hat at random from the ones available. Thus S_1 picks a hat uniformly random from the n available hats, S_2 picks a hat uniformly random from the $(n-1)$ available hats etc. We assume that no sailors pays attention to what hats where picked before him. Denote by p_k the probability that the sailor S_k picks his own hat. Clearly $p_1 = \frac{1}{n}$. What about $p_2, p_3, \dots$?

Denote by H_k the event "*the sailor S_k picked his own hat*" and by A_k, $k > 1$, the event "*none of the sailors $S_1, \dots, S_{k-1}$ picked S_k's hat*".

For $k > 1$ we have

$$p_k = \mathbb{P}(H_k) = \mathbb{P}(H_k|A_k)\mathbb{P}(A_k) + \mathbb{P}(H_k|A_k^c)\mathbb{P}(A_k^c).$$

Note that $\mathbb{P}(H_k|A_k^c) = 0$ because if any of the first $(k-1)$ sailors picked S_k's hat, the chances that S_k picks his own hat are nil. Hence

$$p_k = \mathbb{P}(H_k|A_k)\mathbb{P}(A_k).$$

Now observe that

$$\mathbb{P}(H_k|A_k) = \frac{1}{n-k+1}$$

because the sailor S_k has at its disposal $n-(k-1)$ hats, and we know that one of them is his.

To compute $\mathbb{P}(A_k)$ we use (F/P). The number of outcomes favorable to A_k is equal to the number of *ordered* samplings without replacement of $(k-1)$ hats from a box containing $(n-1)$ hats, i.e., all the hats, but S_k's hat. This number is equal to the number of arrangements of $(k-1)$ objects out of $(n-1)$ possible, i.e.,

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

Similarly, the number of possible outcomes is equal to the number of arrangements of k objects out of n.

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

so that

$$\mathbb{P}(A_k) = \frac{\frac{(n-1)!}{(n-k)!}}{\frac{n!}{(n-k+1)!}} = \frac{(n-1)!}{(n-k)!} \cdot \frac{(n-k+1)!}{n!} = \frac{n-k+1}{n}.$$

Thus

$$p_k = \frac{1}{n-k+1} \cdot \frac{n-k+1}{n} = \frac{1}{n} = p_1, \quad \forall k = 1, 2, \dots, n. \tag{1.14}$$

We have reached the same surprising conclusion as in the previous example, reinforcing Keynes' remark.

To better appreciate the surprising nature of the above result consider the following equivalent formulation.

Suppose the n inebriated sailors board a plane. They enter successively and each picks a seat at random from the available ones. The above result shows that the probability that the 5th sailor pick his assigned seat equals the probability that the 20th sailor picks his assigned seat, which in turn is equal to the probability that the 1st sailor picks his assigned seat, i.e., $\frac{1}{n}$ □

Example 1.50 (The Monty Hall Problem)**.** Contestants in the show *Let's make a deal* were often placed in situations such as the following: you are shown three doors. Behind one door is a car; behind the other two doors are donkeys.

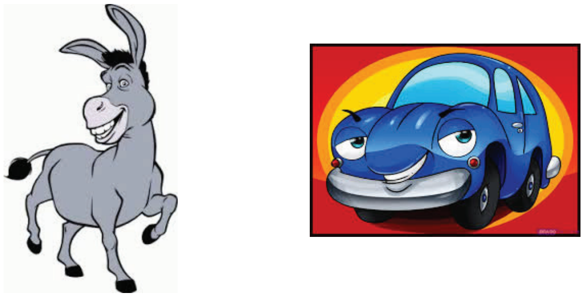

Figure 1.8. *Donkey and car*

You pick a door, but you don't open. **Label that door #** 1. To build some suspense the host opens up one of the two remaining doors to reveal a donkey. What is the probability that there is a car behind the door # 1 that you chose? Should you switch curtains and pick the third, unopened, door if you are given the chance?

Many people argue that the two unopened doors are the same, so they each will contain the car with probability 1/2, and hence there is no point in switching. As we will now show, this naive reasoning is incorrect.

To compute the answer, we will make the following assumptions.[9]

A_1: The host knows behind what door is the car hidden,

A_2: The host always chooses to show you a donkey.

A_3: If there are two unchosen doors with donkeys, then the host chooses one at random, by tossing a fair coin.

Denote by C_1 the event "*there is a car behind door* #1" and by C_3 the event "*there is a car behind the third, unopened, door*". The probability of winning by

[9] Under different assumptions one gets different answers!

switching is $\mathbb{P}(C_3)$. Note that

$$\mathbb{P}(C_1) + \mathbb{P}(C_3) = 1.$$

We will compute $\mathbb{P}(C_3)$ by relying on the law of total probability. We have

$$\begin{aligned}\mathbb{P}\big(C_3\big) &= \mathbb{P}\big(C_3|C_1\big)\mathbb{P}(C_1) + \mathbb{P}\big(C_3|C_1^c\big)\mathbb{P}\big(C_1^c\big)\\ &= 0\cdot\frac{1}{3} + 1\cdot\frac{2}{3} = \frac{2}{3}.\end{aligned}$$

Thus the probability of winning by switching your choice after a door with a donkey is revealed is $\frac{2}{3}$. In particular $\mathbb{P}(C_1) = \frac{1}{3}$. Thus, the switching strategy increases the chances of winning by a factor of 2. □

Example 1.51 (Craps)**.** In this game, a player rolls a pair of dice.

- If the sum is 2, 3, or 12 on his first roll, the player loses.
- If the sum is 7 or 11, he wins.
- If the sum belongs to $J = \{4, 5, 6, 8, 9, 10\}$, this number becomes his "point", and he wins if he "makes his point" that is, his number comes up again before he throws a 7.

What is the probability that the player wins?

Let W denote the event "*the player*" wins. For $s = \{2, \dots, 12\}$ denote by B_s the event "*the first roll of the dice yields the sum s*". We set

$$p_s := \mathbb{P}(B_s).$$

Note that

s	2	3	4	5	6	7	8	9	10	11	12
p	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

(1.15)

From the law of total probability we deduce

$$\mathbb{P}(W) = \sum_{s=2}^{12} \mathbb{P}(W|B_s)\mathbb{P}(B_s).$$

Observe the following.

- If $k = 7, 11$, then $\mathbb{P}(W|B_k) = 1$.
- If $k = 2, 3, 12$, then $\mathbb{P}(W|B_k) = 0$.

We denote by J the set of points, $J = \{4, 5, 6, 9, 9, 10\}$. We deduce

$$\mathbb{P}(W) = \mathbb{P}(B_7) + \mathbb{P}(B_{11}) + \sum_{j\in J}\mathbb{P}(W|B_j)\mathbb{P}(B_j) = \frac{6}{36} + \frac{2}{36} + \sum_{j\in J}\mathbb{P}(W|B_j)p_j. \tag{1.16}$$

For $s \in \{2, \dots, 12\}$ we denote by T_s the number of rolls until we get the first s. Note that for any $j \in J$ we have

$$\mathbb{P}(W|B_j) = \mathbb{P}(T_j < T_7).$$

We have

$$\mathbb{P}(T_j < T_7) = \sum_{k=1} \mathbb{P}(T_j = k, T_7 > k).$$

The event $\{T_j = k, T_7 > k\}$ occurs if during the first $k-1$ rolls the player did not get a 7 or j and he got a j at the k-th roll. The probability q_j of not getting a 7 or a j is $q_j = 1 - p_7 - p_j$. Since the successive rolls are independent we deduce

$$\mathbb{P}(T_j = k, T_7 > k) = q_j^{k-1}p_j$$

so that

$$\mathbb{P}(T_j < T_7) = \sum_{k=1}^{\infty} q_j^{k-1} p_j = p_j \Big(1 + q_j + q_j^2 + \cdots \Big) = p_j \cdot \frac{1}{1-q_j}$$
$$= p_j \cdot \frac{1}{1-(1-p_7-p_j)} = \frac{p_j}{p_j+p_7}.$$

From (1.16) we deduce

$$\mathbb{P}(W) = \frac{8}{36} + \sum_{j\in J} \frac{p_j^2}{p_j+p_7}.$$

Using the table (1.15) we deduce

$$\mathbb{P}(W) = \frac{8}{36} + 2\left(\frac{(3/36)^2}{3/36+6/36} + \frac{(4/36)^2}{4/36+6/36} + \frac{(5/36)^2}{5/36+6/36} \right)$$
$$= \frac{8}{36} + \frac{2}{36}\left(\frac{9}{9} + \frac{16}{10} + \frac{25}{11} \right) \approx 0.4929. \qquad \square$$

Example 1.52 (A before B)**.** The computation of the probability $\mathbb{P}(T_j < T_7)$ in the previous example is a special case of the following more general problem.

A random experiment is performed repeatedly and the outcome of an experiment is independent of the outcomes of the previous experiments. While performing these experiments we keep track of the occurrence of the mutually exclusive events A and B. We assume that A and B have positive probabilities. *What is the probability that A occurs before B?* For example if we roll a pair of dice, A could be the event "*the sum is* 4" and B could be the event "*the sum is* 7". In this case

$$\mathbb{P}(A) = \frac{3}{36} = \frac{1}{12},\ \ \mathbb{P}(B) = \frac{6}{36} = \frac{1}{6}.$$

To answer this question we distinguish two cases.

1. $B = A^c$. Thus, $\mathbb{P}(A \cup B) = 1$, so during an experiment with probability 1 either A or B occurs. Thus A occurs before B iff and only if A occurs at the first trial so, in this case, the probability that A occurs before B is $\mathbb{P}(A)$.

2. $B \neq A^c$. Denote by E the event "A *occurs before* B". Set $C = (A \cup B)^c$ so C signifies that neither A, nor B occurs. Note that

$$\mathbb{P}(C) = 1 - \mathbb{P}(A \cup B) = 1 - \mathbb{P}(A) - \mathbb{P}(B).$$

The collection A, B, C is a partition of the sample space. We condition on the first trial. Thus, either, A, or B, or C occurs. If A occurs then E occurs as well. If B occurs then E^c occurs. If C occurs, then neither A, nor B occurred, we wipe the slate clean, and we're back to where we started, as if we have not performed the first trial. From the law of total probability we deduce

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

The above discussion shows that

$$\mathbb{P}(E|A) = 1,\ \ \mathbb{P}(E|B) = 0,\ \ \mathbb{P}(E|C) = \mathbb{P}(E).$$

Hence

$$\mathbb{P}(E) = \mathbb{P}(A) + \mathbb{P}(E)\mathbb{P}(C) \Rightarrow \mathbb{P}(E)\big(1 - \mathbb{P}(C)\big) = \mathbb{P}(A)$$

$$\Rightarrow \mathbb{P}(E) = \frac{\mathbb{P}(A)}{1 - \mathbb{P}(C)} = \frac{\mathbb{P}(A)}{\mathbb{P}(A) + \mathbb{P}(B)}. \qquad \square$$

Example 1.53 (Gambler's Ruin)**.** Ann decided to gamble. She chose a two-player game of chance with the winning probability p and losing probability $q = 1 - p$. She gets one dollar for every win and pays Bob one dollar for every loss. We denote by β the *bias* of the game defined by

$$\beta = \frac{q}{p}.$$

Thus the game is fair if $\beta = 1$, i.e., $p = q = \frac{1}{2}$, it is biased in favor of Ann if $\beta < 1$, i.e., $q < p$, and it is biased in favor of Bob if $\beta > 1$, i.e., $q > p$.

Ann starts with an amount a of dollars and she decided to play the game until whichever of the following two outcomes occurs first.

- Her fortune reaches a level N prescribed in advance.
- She is ruined, i.e., her fortune goes down to zero.

You can think that N is equal to the combined fortunes of Ann and Bob so when Ann's fortune reaches N Bob is ruined. Obviously when her fortune reaches 0, she is ruined. The number p is the winning probability of Ann, while $q = 1 - p$ is Bob's winning probability.

Denote by W_a "*Ann's starts with an initial amount* $= a$ *(in dollars), and her fortune reaches the level* N *before she is ruined*". Set

$$w_a := \mathbb{P}(W_a).$$

We will refer to w_a as the winning probability.

Denote by R_a the event "*Ann's starts with* a *dollars and is ruined before her fortune reaches the level* N*.*" Set

$$r_a := \mathbb{P}(R_a).$$

We will refer to r_a the *ruin probability.* The events W_a and R_a are disjoint so that

$$w_a + r_a \leq 1.$$

A priori we could not exclude the possibility $w_a + r_a < 1$. This could happen if the probability that Ann plays forever, without getting ruined or reaching the level N were positive. The computations below will show that the probability of this happening is zero. (This is an example of possible but improbable event.)

We assume that both a and N are nonnegative integers, $N \geq a$. We distinguish several cases.

(a) *The game is fair* i.e., $p = \frac{1}{2}$ and $\beta = 1$. (You can think that Ann tosses a coin: tails she wins, heads, the house wins. We condition on the first game. If she loses it (with probability 1/2), her fortune goes down to $a - 1$, and if she wins is (with the same probability), her fortune goes up to $a + 1$ and the process starts anew.

Let us denote by W the event "*Ann wins the first game*" and by L the event "*Ann loses her first game*". The law of total probability then implies

$$\begin{aligned} w_a &= \mathbb{P}(W_a) = \mathbb{P}(W_a|W)\mathbb{P}(W) + \mathbb{P}(W_a|L)\mathbb{P}(L),\\ r_a &= \mathbb{P}(R_a) = \mathbb{P}(R_a|W)\mathbb{P}(W) + \mathbb{P}(R_a|L)\mathbb{P}(L). \end{aligned} \tag{1.17}$$

Now observe that

$$\begin{aligned} \mathbb{P}(W_a|W) = \mathbb{P}(W_{a+1}) = w_{a+1},\;\; \mathbb{P}(W_a|L) = \mathbb{P}(W_{a-1}) = w_{a-1},\\ \mathbb{P}(R_a|W) = \mathbb{P}(R_{a+1}) = r_{a+1},\;\; \mathbb{P}(R_a|L) = \mathbb{P}(R_{a-1}) = r_{a-1}. \end{aligned} \tag{1.18}$$

We deduce

$$w_a = \frac{1}{2}(w_{a-1} + w_{a+1}),\;\; r_a = \frac{1}{2}(r_{a-1} + r_{a+1}).$$

Equivalently, this means

$$w_{a+1} - w_a = w_a - w_{a-1},\;\; r_{a+1} - r_a = r_a - r_{a-1}. \tag{1.19}$$

This shows that both sequences $(w_a)_{a=1,\dots,N}$ and $(r_a)_{a=1,\dots,N}$ are arithmetic progressions. Note that

$$w_0 = 0,\;\; w_N = 1,\;\; r_0 = 1,\;\; r_N = 0.$$

The ratio of (w_a) is $w_1 - w_0 = w_1$. Thus

$$w_a = w_0 + aw_1 = aw_1.$$

From the equality $1 = w_N = Nw_1$ we deduce

$$w_1 = \frac{1}{N},\;\; w_a = \frac{a}{N},\;\; a = 0, 1, \dots, N. \tag{1.20}$$

The ratio of the arithmetic progression (r_a) is $r_1 - r_0 = r_1 - 1$. Hence

$$r_a = r_0 + a(r_1 - 1) = 1 + a(r_1 - 1).$$

From the equality $r_N = 0$ we deduce

$$0 = r_N = 1 + N(r_1 - 1) \Rightarrow r_1 - 1 = -\frac{1}{N},$$

$$r_a = 1 - \frac{a}{N} = \frac{N - a}{N} = 1 - w_a. \tag{1.21}$$

Observing that $N - a$ is Bob's fortune, we see Ann's ruin probability is equal to Bob's winning probability. Note that

$$\lim_{N\to\infty} r_a(N) = 1.$$

Figure 1.9 depicts a computer simulation of a sequence of fair games. Ann

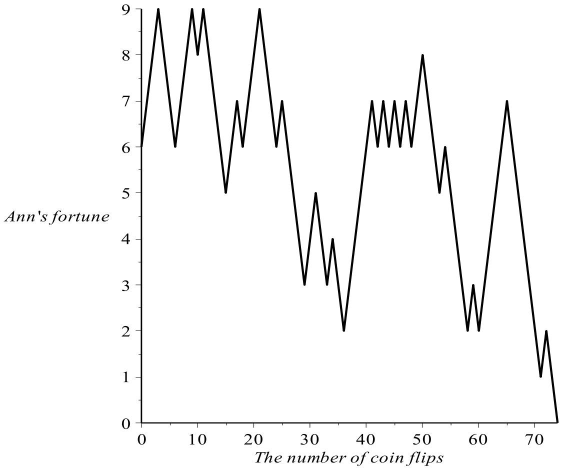

Figure 1.9. *Gambler's ruin simulation with $N = 10$, $a = 6$, $p = 0.5$.*

started with \$6 and has decided to end the game if she reaches \$10 or is ruined, whichever comes first. In this example, and it took over 70 games for Ann to lose her fortune without reaching the \$10 goal. In this case, her chances of reaching the \$10 goal are 60%. In Example 3.34 we will analyze how many games it takes on average for Ann to win.

(b) *The game is biased in favor of Ann*, i.e., $p > q$ (and thus $\beta \neq 1$). The equality (1.17) continues to hold, but in this case we have $\mathbb{P}(W) = p$, $\mathbb{P}(L) = q$. The equality (1.18) takes a different form in this case

$$w_a = pw_{a+1} + qw_{a-1}, \;\; r_a = pr_{a+1} + qr_{a-1}. \tag{1.22}$$

Taking into account that $p + q = 1$ we deduce

$$0 = pw_a a + 1 + qw_{a-1} - w_a = pw_a a + 1 + qw_{a-1} - (p+q)w_a$$
$$= p(w_{a+1} - w_a) - q(w_a - w_{a-1}).$$

A similar argument shows that

$$p(r_{a+1} - r_a) = q(r_a - r_{a-1}).$$

Now set

$$d_a = w_a - w_{a-1}, \;\; \forall a = 1, \dots N.$$

We can rewrite the above equality as

$$pd_{a+1} = qd_a \Longleftrightarrow d_{a+1} = \frac{q}{p} d_a = \beta d_a.$$

Thus the sequence $d_1, \dots, d_N$ is a geometric progression with ratio β so that

$$d_a = \beta^{a-1} d_1, \;\; \forall a = 1. \dots N.$$

On the other hand, we have

$$w_0 + d_1 + d_2 + d_3 + \cdots + d_n$$
$$= w_0 + (w_1 - w_0) + (w_2 - w_1) + (w_3 - w_2) + \cdots + (w_a - w_{a-1}) = w_a.$$

Hence

$$w_a = d_1 \frac{1 - \beta^a}{1 - \beta} = w_1 \frac{1 - \beta^a}{1 - \beta}. \tag{1.23}$$

To find w_1 we need to use the boundary condition $w_N = 1$ and we deduce

$$w_1 \frac{1-\beta^N}{1-\beta} = w_N = 1,$$

so that

$$w_1 = \frac{1-\beta}{1-\beta^N}, \;\; w_a = \frac{1-\beta^a}{1-\beta^N} = \frac{1-\left(\frac{q}{p}\right)^a}{1-\left(\frac{q}{p}\right)^N}.$$

A similar argument shows that the ruin probability is

$$r_a = r_a(N) = \frac{1-\left(\frac{p}{q}\right)^{N-a}}{1-\left(\frac{p}{q}\right)^N} = 1 - w_a = 1 - \frac{1-\beta^a}{1-\beta^N}.$$

Note that

$$\lim_{N\to\infty} r_a(N) = \begin{cases} 1, & \beta > 1, \\ \beta^a, & \beta < 1. \end{cases}$$ □

1.3.4. Bayes' formula.

Example 1.54. Two candidates A and B ran in a mayoral election. The candidate A received 55% of the votes, while B received the remaining 45% of the votes. When analyzing the youth vote, the pollster discovered that 40% of A's voters where under 35 years of age, while only 20% of B's voters where under 35. What percentage of people under 35 voted for A?

Denote by Y the event "*the person is under* 35", by A the event "*the person voted for A*" and by B the event "*the person voted for B*". The question then asks to compute the conditional probability $\mathbb{P}(A|Y)$, i.e., the probability that the person voted for A given that it is under 35. The information extracted by the pollster reads

$$\mathbb{P}(A) = 0.55, \;\; \mathbb{P}(B) = 0.45, \;\; \mathbb{P}(Y|A) = 0.4, \;\; \mathbb{P}(Y|B) = 0.2.$$

We have

$$\mathbb{P}(A|Y) = \frac{\mathbb{P}(A\cap Y)}{\mathbb{P}(Y)} \overset{(1.10)}{=} \frac{\mathbb{P}(Y|A)\mathbb{P}(A)}{\mathbb{P}(Y)} \overset{(1.13)}{=} \frac{\mathbb{P}(Y|A)\mathbb{P}(A)}{\mathbb{P}(Y|A)\mathbb{P}(A)+\mathbb{P}(Y|B)\mathbb{P}(B)}$$

$$= \frac{0.55\cdot 0.4}{0.55\cdot 0.4 + 0.2\cdot 0.45} \approx 0.70.$$

Thus, approximatively 70% of people under 35 voted for A. □

The argument used in the above example is a special case of the following versatile result.

Theorem 1.55 (Bayes' Formula). *Suppose that* $(S, \mathbb{P})$ *is a probability space and* $(B_n)_{n\geq 1}$ *is a partition of* S *such that*

$$\mathbb{P}(B_n) > 0, \;\; \forall n \geq 1.$$

Then, for any event $A \subset S$, and any $k \geq 1$, we have

$$\boxed{\mathbb{P}(B_k|A) = \frac{\mathbb{P}(A|B_k)\mathbb{P}(B_k)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B_k)\mathbb{P}(B_k)}{\sum_{n\geq 1}\mathbb{P}(A|B_k)\mathbb{P}(B_k)}}. \tag{1.24}$$

Proof. We have

$$\mathbb{P}(B_k|A) \overset{(1.10)}{=} \frac{\mathbb{P}(A\cap B_k)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B_k)\mathbb{P}(B_k)}{\mathbb{P}(A)} \overset{(1.13)}{=} \frac{\mathbb{P}(A|B_k)\mathbb{P}(B_k)}{\sum_{n\geq 1}\mathbb{P}(A|B_k)\mathbb{P}(B_k)}.$$

□

Corollary 1.56. *Suppose that $(S, \mathbb{P})$ is a probability space and B is an event such that $0 < \mathbb{P}(B) < 1$. Then for any event A such that $\mathbb{P}(A) \neq 0$ we have*

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

Proof. Observe that $B_1 = B, B_2 = B^c$ is a partition of S. Clearly

$$\mathbb{P}(B), \mathbb{P}(B^c) > 0.$$

Now apply (1.24). □

Example 1.57 (Bayesian inference: **M**aximum **A** **P**osteriori)**.** Suppose doctors are asked to report the number of cases of smallpox and chickenpox, and the symptoms they observed. This survey shows that 90% of the patients with smallpox have spots and 80% of the patients with chickenpox have spots. We can write this as conditional probabilities

$$\mathbb{P}(\text{spots}|\text{smallpox}) = 0.9,\;\; \mathbb{P}(\text{spots}|\text{chickenpox}) = 0.8.$$

These are called the *likelihoods* of smallpox and respectively chickenpox.

In diagnosing a disease a doctor needs to take into account the conditional probabilities

$$\mathbb{P}(\text{smallpox}|\text{spots}),\;\; \mathbb{P}(\text{chickenpox}|\text{spots})$$

called *a posteriori estimates*.

If initially, or *a priori*, we assume that

$$\mathbb{P}(\text{smallpox}) = \mathbb{P}(\text{chickenpox}) = p,$$

then we deduce from the Bayes' formula that

$$\mathbb{P}(\text{smallpox}|\text{spots}) = \frac{\mathbb{P}(\text{spots}|\text{smallpox})\mathbb{P}(\text{smallpox})}{\mathbb{P}(\text{spots})} = 0.9\frac{p}{\mathbb{P}(\text{spots})},$$

$$\mathbb{P}(\text{chickenpox}|\text{spots}) = \frac{\mathbb{P}(\text{spots}|\text{chickenpox})\mathbb{P}(\text{chickenpox})}{\mathbb{P}(\text{spots})} = 0.8\frac{p}{\mathbb{P}(\text{spots})}.$$

This analysis shows that when observing spots, the patient is more likely to have smallpox. In this case, the decision was based on the likelihoods, and the largest one decided which disease is more likely. This strategy is called MLE or *Maximum Likelihood Estimates.* The probabilities $\mathbb{P}(\text{smallpox}), \mathbb{P}(\text{chickenpox})$ are called *priors* and the doctor made the a priori assumption that they are equal.

Suppose now that the doctor is aware that statistical data also show that the priors are

$$\mathbb{P}(\text{smallpox}) = 0.01,\;\; \mathbb{P}(\text{chickenpox}) = 0.1.$$

Then

$$\mathbb{P}(\text{smallpox}|\text{spots}) = \frac{0.9 \cdot 0.01}{\mathbb{P}(\text{spots})} = \frac{0.009}{\mathbb{P}(\text{spots})},$$
$$\mathbb{P}(\text{chickenpox}|\text{spots}) = \frac{0.8 \cdot 0.1}{\mathbb{P}(\text{spots})} = \frac{0.08}{\mathbb{P}(\text{spots})}.$$

Thus, with a different assumptions on priors, we reach a different conclusion. This strategy is called MAP or *Maximum A Posteriori* estimate.

The above decision making process is an example of what is commonly referred to as *Bayesian inference*. This principle is playing an increasingly bigger part in Machine Learning and Artificial Intelligence. □

Example 1.58. The polygraph is an instrument used to detect physiological signs of deceptive behavior. Although it is often pointed out that the polygraph is not a lie detector, this is probably the way most of us think of it. For the purpose of this example, let us retain this notion. Let us assume that the polygraph test is indeed very accurate and that it decides "lie" or "truth" correctly with probability 0.95. Now consider a randomly chosen individual who takes the test and is determined to be lying. What is the probability that this person did indeed lie?

Consider the event L "*the person lies*" and the event L_P "*the polygraph says the person lied*". We are are interested in the conditional probability $\mathbb{P}(L|L_P)$. Note that L^c signifies that the person is truthful, while L_P^c signifies that the polygraph says the person is truthful. We know that

$$P(L_P|L) = 0.95 = P(L_P^c|L^c) = 0.95.$$

In particular

$$\mathbb{P}(L_P|L^c) = 1 - \mathbb{P}(L_P^c|L^c) = 0.05.$$

The quantity $\mathbb{P}(L_P|L)$, the probability that the detector says you're lying, given that you lied is called the *likelihood* of lying. We set

$$\lambda := \mathbb{P}(L).$$

The quantity λ determines the *prior* information. Bayes' formula implies

$$\mathbb{P}(L|L_P) = \frac{\mathbb{P}(L_P|L)\mathbb{P}(L)}{\mathbb{P}(L_P|L)\mathbb{P}(L) + \mathbb{P}(L_P|L^c)\mathbb{P}(L^c)} = \frac{0.95\lambda}{0.95\lambda + 0.05(1-\lambda)}.$$

The graph of $\mathbb{P}(L|L_P)$ as a function of the prior $\lambda = \mathbb{P}(L)$ is depicted in Figure 1.10.

The probability λ is typically very small. If we assume that only 1 in 1000 people lies, i.e., $\lambda = 0.001$, then

$$p(\lambda) \approx 0.018.$$

This tells that *the probability that a person lied, given that the polygraph says so is very small, $<2\%$!*

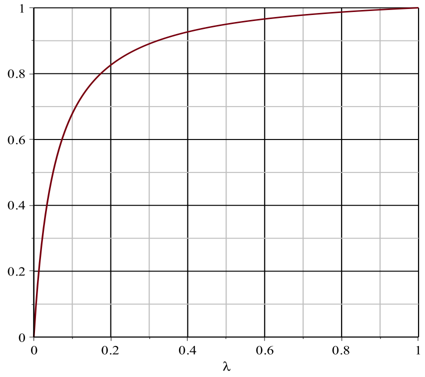

Figure 1.10. *The graph of $p(\lambda)$.*

We can turn this on its head and conclude that the probability that a person told the truth, given that the polygraph said otherwise is very large, $> 98\%$. The graph of $p(\lambda)$ shows that $p(\lambda) > 0.90$ if $\lambda > 0.3$. Thus the polygraph works well on a sample with large number of liars, but is not that good on truthful people.□

Example 1.59. Approximately 1% of women aged $40 - 50$ years have breast cancer. A woman with breast cancer has a 90% chance of a positive test from a mammogram, while a woman without cancer has a 10% chance of a (false-)positive result. What is the probability that a woman has breast cancer given that she just had a positive test?

We denote by B the event "*the woman has breast cancer*" and by T the event "*the woman tested positive for breast cancer*". We know that

$$\mathbb{P}(B) = 0.01,\ \ \mathbb{P}(B^c) = 0.99,\ \ \mathbb{P}(T|B) = 0.9,\ \ \mathbb{P}(T|B^c) = 0.1.$$

We are asked to find $\mathbb{P}(B|T)$. Bayes' formula implies

$$\mathbb{P}(B|T) = \frac{\mathbb{P}(T|B)\mathbb{P}(B)}{\mathbb{P}(T|B)\mathbb{P}(B) + \mathbb{P}(T|B^c)\mathbb{P}(B^c)} = \frac{0.9 \cdot 0.01}{0.9 \cdot 0.01 + 0.1 \cdot 0.99}$$

$$= \frac{0.009}{0.009 + 0.099} = \frac{9}{9 + 99} \approx 0.083.$$

This answer is somewhat surprising. Indeed, when 95 physicians were asked the question "What is the probability a woman has breast cancer given that she just had a positive test", their average answer was 75%. The two statisticians who carried out this survey indicated that physicians were better able to see the answer when the data were presented in frequency format. Ten out of $1,000$ women have breast cancer. Of these 9 will have a positive mammogram. However, of the remaining 990 women without breast cancer, 99 will have a positive test, and again we arrive at the answer $9/(9 + 99)$. □

1.4. Exercises

Exercise 1.1. A certain thick and asymmetric coin is tossed and the probability that it lands on the edge is 0.1. If it does not land on the edge, it is twice as likely to show heads as tails. What is the probability that it shows heads?

Exercise 1.2. Let A and B be two events such that

$$\mathbb{P}(A) = 0.3, \ \ \mathbb{P}(A \cup B) = 0.5, \text{ and } \mathbb{P}(A \cap B) = 0.2.$$

Find

(i) $\mathbb{P}(B)$,

(ii) the probability that A but not B occurs,

(iii) $\mathbb{P}(A \cap B^c)$,

(iv) $\mathbb{P}(A^c)$,

(v) the probability that B does not occur, and

(vi) the probability that neither A nor B occurs.

Exercise 1.3. Let A be the event that "it rains on Saturday" and B the event that "it rains on Sunday". Suppose that $\mathbb{P}(A) = \mathbb{P}(B) = 0.5$. Furthermore, let p denote the probability that it rains on both days. Express the probabilities of the following events as functions of p:

(i) it rains on Saturday but not Sunday.

(ii) It rains on one day but not the other.

(iii) It does not rain at all during the weekend.

Exercise 1.4. The probability in Exercise 1.3(b) is a decreasing function of p. Explain this intuitively.

Exercise 1.5. People are asked to assign probabilities to the events "rain on Saturday" "rain on Sunday", "rain both days", and "rain on at least one of the days". Which of the following suggestions are consistent with the probability axioms:

(i) 70%, 60%, 40%, and 80%,

(ii) 70%, 60%, 40%, and 90%,

(iii) 70%, 60%, 80%, and 50%, and

(iv) 70%, 60%, 50%, and 90%?

Exercise 1.6. You are asked to select a password for a Web site. It must consist of five lowercase letters and two digits in any order. How many possible such passwords are there if (a) repetitions are allowed, and (b) repetitions are not allowed?

Exercise 1.7. An Indiana license plate consists of three letters followed by three digits. Find the probability that a randomly selected plate has (a) no duplicate letters, (b) no duplicate digits, (c) all letters the same, (d) only odd digits, and (e) no duplicate letters and all digits equal.

Exercise 1.8. "A thousand monkeys, typing on a thousand typewriters will eventually type the entire works of William Shakespeare" is a statement often heard in one form or another. Suppose that one monkey presses 10 keys at random.What is the probability that he types the word HAMLET if he is (a) allowed to repeat letters, and (b) not allowed to repeat letters? (Assume that the typewriter has precisely 26 symbols.)

Exercise 1.9. Four envelopes contain four different amounts of money. You are allowed to open them one by one, each time deciding whether to keep the amount or discard it and open another envelope. Once an amount is discarded, you are not allowed to go back and get it later. Compute the probability that you get the largest amount under the following different strategies: (a) You take the first envelope. (b) You open the first envelope, note that it contains the amount x, discard it and take the next amount which is larger than x (if no such amount shows up, you must take the last envelope). (c) You open the first two envelopes, call the amounts x and y, and discard both and take the next amount that is larger than both x and y.

Exercise 1.10. On a chessboard (8×8 squares, alternating black and white), you place three chess pieces at random. What is the probability that they are all (a) in the first row, (b) on black squares, (c) in the same row, and (d) in the same row and on the same color?

Exercise 1.11. In a regular coordinate system, you start at $(0, 0)$ and flip a fair coin to decide whether to go sideways to $(1, 0)$ (East) or up to $(0, 1)$ (North). You continue in this way, and after n flips you have reached the point (j, k), where $j + k = n$. What is the probability that (a) all the j steps sideways came before the k steps up, (b) all the j steps sideways came either before or after the k steps up, and (c) all the j steps sideways came in a row?

Exercise 1.12. An urn contains n red balls, n white balls, and n black balls. You draw k balls at random without replacement (where $k \leq n$). Find an expression for the probability that you do not get all colors.

Exercise 1.13. You are dealt a poker hand.[10] What is the probability of getting (a) royal flush, (b) straight flush, (c) four of a kind, (d) full house, (e) flush?

Exercise 1.14. From the integers $1, \ldots, 10$, three numbers are chosen at random without replacement. (a) What is the probability that the smallest number is 4?

[10] For the definition of poker hands see https://en.wikipedia.org/wiki/List_of_poker_hands

(b) What is the probability that the smallest number is 4 and the largest is 8?
(c) If you choose three numbers from $1, \dots, n$, what is the probability that the smallest number is j and the largest is k for possible values of j and k?

Exercise 1.15. An urn contains n white and m black balls. You draw repeatedly at random and without replacement. What is the probability that the first black ball comes in the k-th draw, $k = 1, 2, \dots, n+1$?

Exercise 1.16. A city with 6 districts has 6 robberies in one week. Assume that robberies are located randomly, and all districts are equally likely to be robbed.

(i) What is the probability that that some district had more than one robbery?

(ii) Answer the same question in the case when the city has 10 districts and was robbed 8 times.

(iii) In which of the above two cases the probability that one district was robbed more than once is larger?

Exercise 1.17. Eggs are delivered to a restaurant by the gross (1 gross = 12 dozen). From each gross, a dozen of eggs are chosen at random. If none are cracked, the gross is accepted, and if more than one egg is cracked, the gross is rejected. If exactly one egg is cracked, an additional dozen eggs from the same gross are inspected. If this lot has no cracked eggs, the entire gross is accepted, otherwise it is rejected. Suppose that a gross has eight cracked eggs. What is the probability that it is accepted?

Exercise 1.18. Let A and B be disjoint events. Show that

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

Exercise 1.19. Let A, B, and C be three events such that $\mathbb{P}(B \cap C) > 0$. Show that

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

and that

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

Exercise 1.20. Let A, B, and C be independent events. Show that A is independent of both $B \cap C$ and $B \cup C$.

Exercise 1.21. A coin has probability p of showing heads. Flip it three times and consider the events A, "at most one tails" and B, "all flips are the same". For which values of p are A and B independent?

Exercise 1.22. A politician considers running for election and has decided to give it two tries. He figures that the current conditions are favorable and that he has about a 60% chance of winning this election and, win or lose the first election, he has a 50% chance in the next election. However, if he does win this election, he estimates that there ought to be a 75% chance of being reelected.

(i) Find the probability that he wins both elections.

(ii) Find the probability that he wins the first election and loses the second.

(iii) If you learn that he won the second election, what is the probability that he won the first election?

(iv) If he loses the first election, what is the probability that he wins the second?

Exercise 1.23. In December 1992, a small airplane crashed in a residential area near Stockholm, Sweden. In an attempt to calm the residents, the airport manager claimed that they should now feel safer than before since the probability of two crashes is much smaller than the probability of one crash and hence it has now become less likely that another crash will occur in the future. What do you think of his argument?

Exercise 1.24. Bob and Joe are working on a project. They each have to finish their individual tasks to complete the project and work independent of each other. When Bob is asked about the chances of him getting his part done, Joe getting his part done, and then both getting the entire project done, he estimates these to be 99%, 90%, and 95%, respectively. Is this reasonable?

Exercise 1.25. You roll a die and consider the events A, “the number you get is even”, and B, “you get at least 2”. Find $\mathbb{P}(B|A)$ and $\mathbb{P}(A|B)$.

Exercise 1.26 (The prosecutor’s fallacy)**.** [11] Let G be the event that an accused is guilty and T the event that some testimony is true. Some lawyers have argued that $\mathbb{P}(G|T) = \mathbb{P}(T|G)$. Prove that this is the case if and only if $\mathbb{P}(G) = \mathbb{P}(T)$.

Exercise 1.27. We are given 20 urns $U_1, U_2, \dots, U_{20}$, each containing 19 balls, such that U_1 contains 19 green balls, U_2 contains 1 red ball and 18 green, U_3 contains two red balls and 17 green etc. We select an urn at random and the we sample without replacement two balls. What is the probability that the second ball we sample is green?

Exercise 1.28 (The prisoner’s dilemma)**.** Three prisoners, Al, Bob, and Charlie, are in a cell. At dawn two will be set free and one will be hanged, but they do not know who will be chosen. The guard offers to tell Al the name of one of the other two prisoners who will go free but Al stops him, screaming, “No, don’t! That would increase my chances of being hanged to 1/2.” Is Al correct in his assessment? Justify your answer.

Exercise 1.29. You roll a die twice and record the largest number (if the two rolls give the same outcome, this is the largest number).

[11] The prosectors made this error during the famous Dreyfus affair.

(i) Given that the first roll gives 1, what is the conditional probability that the largest number is 3?

(ii) Given that the first roll gives 3, what is the conditional probability that the largest number is 3?

Exercise 1.30. Roll two fair dice. Let A_k be the event that the first die gives k, and let B_n be the event that the sum is n. For which values of n and k are A_k and B_n independent?

Exercise 1.31. You are offered to play the following game: A roulette wheel is spun eight times. If any of the 38 numbers $(0, 00, 1-36)$ is repeated, you lose $10, otherwise you win $10. Should you accept to play this game? Argue by computing the relevant probability.

Exercise 1.32. Consider the following simplified version of the birthday problem in Example 1.16. Divide the year into "winter half" and "summer half." Suppose that the probability that an individual is born in the winter half is p. What is the probability that two people are born in the same half of the year? For which value of p is this minimized?

Exercise 1.33. Consider three dice, A, B, and C, numbered on their six sides as follows:

$$\begin{aligned} A &: 1, 1, 5, 5, 5, 5, \\ B &: 3, 3, 3, 4, 4, 4, \\ C &: 2, 2, 2, 2, 6, 6. \end{aligned}$$

If all three dice are rolled at once, which is the most likely to win?

Exercise 1.34. Three fair dice are rolled. Given that there are no 6s, what is the probability that there are no 5s?

Exercise 1.35. Suppose that parents are equally likely to have (in total) one, two or three children. A girl is selected at random. What is the probability that the family has no older girl? (Assume that the genders of the children are independent and are equally likely to be male or female.)

Exercise 1.36. The distribution of blood types in the United States according to the ?ABO classification? is O: 45%, A:40%, B: 11%, and AB: 4%. Blood is also classified according to Rh type, which can be negative or positive and is independent of the ABO type (the corresponding genes are located on different chromosomes). In the U.S. population, about 84% are Rh positive. Sample two individuals at random and find the probability that

(i) both are A negative,

(ii) one of them is O and Rh positive, while the other is not,

(iii) at least one of them is O positive,

(iv) one is Rh positive and the *other* is not AB,

(v) they have the same ABO type, and

(vi) they have the same ABO type and different Rh types.

Exercise 1.37. In a blood transfusion, you can always give blood to somebody of your own ABO type (see Exercise 1.36). Also, type O can be given to anybody and those with type AB can receive from anybody (people with these types are called universal donors and universal recipients, respectively). Suppose that two individuals are chosen at random. Find the probability that

(i) neither can give blood to the other,

(ii) one can give to the other but not vice versa,

(iii) at least one can give to the other, and

(iv) both can give to each other.

Exercise 1.38. In the United States, the overall chance that a baby survives delivery is 99.3%. For the 15% that are delivered by cesarean section, the chance of survival is 98.7%. If a baby is not delivered by cesarean section, what is its survival probability?

Exercise 1.39. Bob is flying from O'Hare to Sydney with a stopover at LAX. He knows that at O'Hare luggages are mishandled with probability $p = \frac{1}{100}$ and if they are correctly handled at O'Hare, the probability that they are mishandled at LAX is also $p = \frac{1}{100}$.

(i) What the probability that his luggage was mishandled at O'Hare given that its luggage was missing in Sydney?

(ii) What the probability that his luggage was mishandled at LAX given that its luggage was missing in Sydney?

Exercise 1.40. You roll a die and flip a fair coin a number of times determined by the number on the die. What is the probability that you get no heads?

Exercise 1.41. You have three pieces of string and tie together the ends two by two at random.

(i) What is the probability that you get one big loop?

(ii) Generalize to n pieces of string.

Exercise 1.42. We sample with replacement a regular deck of cards until we get an ace, or we get a spade but not the ace of spades. What is the probability that the ace comes first?

Exercise 1.43. From a deck of cards, draw four cards at random without replacement. If you get k aces, draw k cards from another deck. What is the probability to get exactly k aces from the first deck and exactly n aces from the second deck?

Exercise 1.44. Graduating students from a particular high school are classified as *weak* or *strong*. Among those who apply to college, it turns out that 56% of the weak students but only 39% of the strong students are accepted at their first choice. Does this indicate a bias against strong students?

Exercise 1.45. A box contains two regular quarters and one fake two-headed quarter.

(i) You pick a coin at random. What is the probability that it is the two-headed quarter?

(ii) You pick a coin at random, flip it, and get heads. What is the probability that it is the two-headed quarter?

Exercise 1.46. Two cards are chosen at random without replacement from a deck and inserted into another deck. This deck is shuffled, and one card is drawn. If this card is an ace, what is the probability that no ace was moved from the first deck?

Exercise 1.47. A transmitter randomly sends the bits 0 and 1 to a receiver. Each bit is received correctly (0 as 0, 1 as 1) with probability 0.9. Bits are received correctly independent of each other and, on the average, twice as many 0s as 1s are being sent.

(i) If the sequence 10 is sent, what is the probability that 10 is received?

(ii) If the sequence 10 is received, what is the probability that 10 was sent?

Exercise 1.48. Consider two urns, one with 10 balls numbered 1 through 10 and one with 100 balls numbered 1 through 100. You first pick an urn at random, then pick a ball at random, which has number 5.

(i) What is the probability that it came from the first urn?

(ii) What is the probability in (i) if the ball was instead chosen randomly from all the 110 balls?

Exercise 1.49. The serious disease D occurs with a frequency of 0.1% in a certain population. The disease is diagnosed by a method that gives the correct result (i.e., positive result for those with the disease and negative for those without it) with probability 0.99. Mr Smith goes to test for the disease and the result turns out to be positive. Since the method seems very reliable, Mr Smith starts to worry, being "99% sure of actually having the disease." Show that this is not the relevant probability and that Mr Smith may actually be quite optimistic.

Exercise* 1.50. In Example 1.53 suppose that Bob's fortune is infinite, Ann starts with one dollar, and her winning probability is $p > \frac{1}{2}$. What is the probability that she eventually goes broke?

Exercise 1.51. Four red balls and two blue balls are placed at random into two urns so that each urn contains three balls. What is the probability of getting a blue ball in the following instances.

(i) You select a ball at random from the first urn?

(ii) You select an urn at random and then select a ball from it at random?

(iii) You discard two balls from the second urn and select the last ball?

Exercise 1.52. In a factory, if the most recent accident occurred exactly k days before today, then the probability that an accident occurs today is p_k ; there is no accident with probability $1 - p_k$. During the n successive days immediately after an accident, what is the probability that

(i) There are no accidents?

(ii) There is exactly one accident?

Chapter 2

Random variables

2.1. Some general facts

Loosely speaking a *random variable* (or *rv*) is a numerical quantity associated to the outcome of a random experiment. As such, it is a random quantity. Here is a more formal description.

Definition 2.1. A *random* variable is a function $X : S \to \mathbb{R}$, where $(S, \mathbb{P})$ is probability space. □

The following examples illustrate this idea.

Example 2.2. (a) Roll a pair of dice. The *sum of the numbers* on the two dice is a random variable. Its values can be any number $\{2, 3, \dots, 12\}$.

(b) Roll a pair of dice $1,000$ times. The *number of times we get* 7 is a random variable. Its value can be any of the numbers $\{0, 1, \dots, 1000\}$.

(c) Roll a pair of dice until you get a sum of 7. The *number of rolls needed to get a* 7 is a random variable. It can take any value $\{1, 2, \dots, \infty\}$.

(d) The *lifetime of a lightbulb* is a random variable. It can take any value in the interval $[0, \infty)$. □

The random variables in Example 2.2(a),(b),(c) are *discrete* random variables, while the random variable in (d) is *continuous.*

Convention. Random variables are to be denoted by *Capital letters*

$A, B, C, \dots, Y, Z.$

Definition 2.3. Let X be a random variable. The *cumulative distribution function*(cdf) of X is the function

$$F : \mathbb{R} \to [0,1], \quad F(x) := \mathbb{P}(X \leq x). \qquad \square$$

Definition 2.4 (Quantiles)**.** Suppose that X is a random variable with cumulative distribution function F_X. For any number $p \in (0,1]$ the *p-quantile* of X, denoted by $Q_X(p)$ is the smallest number x_0 such that

$$F_X(x_0) = \mathbb{P}(X \leq x_0) \geq p.$$

Thus, x_0 is the p-quantile of X, $x_0 = Q_X(p)$, if

- $\mathbb{P}(X \leq x_0) \geq p$ and
- $\mathbb{P}(X \leq x) < p$, fo any $x < x_0$.

The *median* of X is the 0.5-quantile. The p-quantile defines a function $Q_X : (0,1] \to \mathbb{R}$ that is a sort of inverse of the cdf F_X. □

Remark 2.5. In statistics, the term *percentile* is often used when referring to quantiles. For example, the median is the 50-th percentile. A number x_0 is the 28-th percentile of the random variable X if x_0 is the 0.28-quantile of X, i.e.,

$$\mathbb{P}(X \leq x_0) \geq 0.28$$

and, if $x_1 < x_0$, then $\mathbb{P}(X \leq x_1) < 0.28$. □

Example 2.6. Suppose that H denotes the height (in inches) of a random individual in the country of Lilliput. The statement "the height 1 inch is the 60-th percentile of H" signifies two things:

- at least 60% of Lilliputians have height ≤ 1, and
- less than 60% have height ≤ 0.9999.

□

Definition 2.7 (Independence)**.** Fix a sample space $(S, \mathbb{P})$.

(i) The random variables $X_1, \dots, X_n : S \to \mathbb{R}$ are called *independent* if, for any numbers $x_1, \dots, x_n \in (-\infty, \infty]$, the events

$$\{\, X_1 \leq x_1 \,\}, \dots, \{\, X_n \leq x_n \,\}$$

are independent. Equivalently, for any numbers $x_1, \dots, x_n \in (-\infty, \infty]$,

$$\mathbb{P}(X_1 \leq x_1, \dots X_n \leq x_n) = \mathbb{P}(X_1 \leq x_1) \cdots \mathbb{P}(X_n \leq x_n). \tag{2.1}$$

(ii) An infinite sequence of random variables $X_n : S \to \mathbb{R}$, $n = 1, 2, \dots$, is called *independent* if $X_1, \dots, X_n$ are independent for any n. The sequence is called *iid* (independent, identically distributed) if it is independent and the random variables have the same cdf, i.e.,

$$\mathbb{P}(X_i \leq x) = \mathbb{P}(X_j \leq x), \;\; \forall x \in \mathbb{R}, \;\; i, j.$$

□

We will use the notation $X \perp\!\!\!\perp Y$ to indicate that the random variables X and Y are independent.

Remark 2.8. One can prove that the random variables $X_1, \dots, X_n$ are independent *if and only if* for "any"[1] sets $B_1, \dots, B_n \subset \mathbb{R}$ the events

$$\big\{\, X_1 \in B_1 \,\big\}, \dots \{X_n \in B_n\,\big\}$$

are independent, i.e.,

$$\boxed{\mathbb{P}\big(\, X_1 \in B_1, \dots X_n \in B_n \,\big) = \mathbb{P}\big(\, X_1 \in B_1 \,\big) \cdots \mathbb{P}\big(\, X_n \in B_n \,\big)}. \tag{2.2}$$

$$B_1 = (-\infty, x_1], \dots, B_n = (-\infty, x_n],$$

then the equality (2.2) becomes (2.1). □

2.2. Discrete random variables

A random variable X is called *discrete* if its range is a finite or countable set of real numbers $x_1, x_2, \dots, x_n, \dots$.

Definition 2.9. (a) Let X be a discrete random variable with range

$$\mathscr{X} = \{x_1, x_2, \dots\}.$$

The *probability mass function* (pmf) (or the *law*) of X is the function $p_X : \mathscr{X} \to [0,1]$ given by

$$p_X(x_k) := \mathbb{P}(X = x_k).$$

A *mode* of X is a value x^* in the range $\mathscr{X}$ where $p(x)$ has a local maximum.

(b) Two discrete random variables X, Y with ranges $\mathscr{X}$ and respectively $\mathscr{Y}$ are called *equivalent*, and we denote this $X \sim Y$, if $\mathscr{X} = \mathscr{Y}$ and $p_X = p_Y$. □

Example 2.10. Let G be the number of girls in a random family with three children. The range of G is $\mathscr{G} = \{0,1,2,3\}$ and its pmf is

$$p : \{0,1,2,3\} \to [0,1], \;\; p(0) = p(3) = \frac{1}{8}, \; p(1) = p(2) = \frac{3}{8}.$$

The median of G is 1: 50% of families with 3 children have at most one girl. The modes of G are 1 and 2 as they are the most likely values. We can encode this in a pie chart as in Figure 2.1.

[1] The term "*any*" needs to be taken with a grain of salt due to some rather subtle foundational issues.

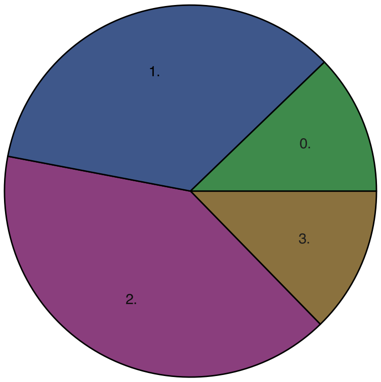

Figure 2.1. *The distribution of the number of girls in a random sample of* 1000 *families with* 3 *children.*

The cdf of this random variable is

$$F(x) = \begin{cases} 0, & x < 0, \\ p(0) = \frac{1}{8}, & x \in [0,1), \\ p(0) + p(1) = \frac{1}{2}, & x \in [1,2), \\ p(0) + p(1) + p(2) = \frac{7}{8}, & x \in [2,3), \\ 1, & x \geq 3. \end{cases} \tag{2.3}$$

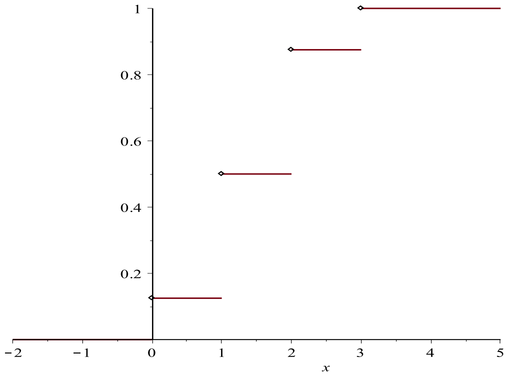

Figure 2.2. *The graph of the cdf* $F(x)$.

The 0.6 quantile of the random variable G is the smallest number $x_0 \in \{0,1,2,3\}$ such that $F(x_0) \geq 0.6$. From (2.3) we see that 2 is the 0.6-quantile. More generally, the quantile function of G, $Q_G : (0,1] \to \{0,1,2,3\}$, is given by

$$Q_G(p) = \begin{cases} 0, & p \in (0,1/8] \\ 1, & p \in (1/8,1/2], \\ 2, & p \in (1/2,7/8], \\ 3, & p \in (7/8,1]. \end{cases}$$

Let us observe that the cdf of G is *right*-continuous, while the quantile function is *left*-continuous. (This is true for all discrete random variables.)

In Example 7.9 we explain how to simulate in R, custom discrete random variables such as this. □

Example 2.11. Suppose we roll a die and we declare the result a success, if we get a 6 and failure otherwise. We let X be the random variable with range $\{0,1\}$, where $X=1$ indicates success, and $X=0$ indicates failure. The pmf of X is

$$p_0 = \mathbb{P}(X=0) = \frac{5}{6}, \;\; p_1 = \mathbb{P}(X=1) = \frac{1}{6}.$$ □

Example 2.12. Suppose we roll a die n times and X denotes the number of times we get a 6. As in the previous example, we will call success the event of getting a 6. Then X is a random variable with range

$$\{\, 0,1,\dots,n \,\}.$$

To compute its pmf we consider the *independent* events $A_1,\dots,A_n$, where A_k is the event "*we have success at the k-th roll*". The event $\{X=k\}$ can be described as the event of having exactly k successes in n independent trials. The independence of the event $A_1,\dots,A_n$ probability of having successes at trials $t_1,\dots,t_k$, but at no other trials is

$$\left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{n-k}.$$

We deduce

$$\mathbb{P}(X=k) = \binom{n}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{n-k}$$

because there are exactly $\binom{n}{k}$ ways of choosing k trials out of n and declare them the successful trials. Each such choice has the same probability $\left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{n-k}$. □

Example 2.13. If we roll a die and we denote by X the number of rolls until we have our first success, i.e., we get the first 6, then X is a discrete random variable with range $\{1,2,\dots\}$. The computation in Example 1.41 shows that its pmf is

$$\mathbb{P}(X=n)=\left(\frac{5}{6}\right)^{n-1}\frac{1}{6},\;\;\forall n=1,2,\dots\,. \tag{2.4}$$

If F denotes the cdf of X, then

$$F(x)=\mathbb{P}(X=1)+\mathbb{P}(X=2)+\cdots+\mathbb{P}(X=n),\;\;\forall n\in\mathbb{N},\;\;x\in[n,n+1).$$

For n a positive integer, $F(n)$ is the probability that the first 6 appears after at most n trials. Thus $1-F(n)$ is the probability that there is no 6 during the first n trials. The probability of this event is

$$\mathbb{P}(X>n)=\left(\frac{5}{6}\right)^{n},$$

so that

$$F(n)=1-\left(\frac{5}{6}\right)^{n}. \qquad \square$$

Proposition 2.14. *Suppose that $\mathscr{X}\subset\mathbb{R}$ is a finite or countable set,*

$$\mathscr{X}=\{x_1,x_2,\dots\}.$$

A function $p:\mathscr{X}\to[0,1]$ is the probability distribution of a discrete random variable with range $\mathscr{X}$ if and only if it satisfies the normalization condition

$$\sum_{x\in\mathscr{X}}p(x)=1. \tag{2.5}$$

□

2.2.1. Fundamental examples of discrete random variables.

Example 2.15 (Discrete Uniform Distribution)**.** Suppose that $\mathscr{X}\subset\mathbb{R}$ is a finite set consisting of precisely n elements

$$\mathscr{X}=\{\,x_1,\dots,x_n\,\}.$$

The *discrete uniform distribution* assigns to each value x_k the same probability of occurring

$$p(x_k)=\frac{1}{n}. \qquad \square$$

Example 2.16 (Bernoulli trials)**.** Fix a probability space $(S,\mathbb{P})$ and an event $E\subset S$ with probability $\mathbb{P}(E)=p\in[0,1]$. We refer to E as "success" and to p as the probability of success, i.e., the probability that the event E occurs. (For example, we can declare success if we get a 6 in a roll of a fair die, in which case $p=1/6$.) We set $q:=1-p$ and we think of q as the probability of failure. An experiment in which we observe if the given event E has occurred is called a *Bernoulli trial*.

We can encode a Bernoulli trial as the *indicator function* of the event E, i.e., the function

$$I_E : S \to \{0,1\},\ \ I_E(s) = \begin{cases} 1, & s \in E, \\ 0, & s \in S \setminus E. \end{cases}$$

The indicator function I_E is the random variable that takes value 1, if E occurs and 0 otherwise. The probability mass function of I_E is then

$$\mathbb{P}(I_E = 0) = q,\ \ \mathbb{P}(I_E = 1) = p.$$

Any discrete random variable X with the above probability mass function is called a *Bernoulli* random variable with success probability p and we will indicate this by $X \sim \mathrm{Ber}(p)$. □

Example 2.17 (Binomial distributions)**.** Suppose we perform a sequence of n independent Bernoulli trials each with the same probability of success p. Let X denote the number of successes observed during these n trials. Then X is a discrete random variable with range $\{0,1,2,\dots,n\}$. Arguing as in Example 2.12 we deduce that

$$\mathbb{P}(X = k) = \binom{n}{k} p^k q^{n-k},\ \ \forall k = 0,1,\dots,n. \tag{2.6}$$

This probability distribution is called the *binomial distribution* corresponding to n trials and success probability p. We will denote by $\mathrm{Bin}(n,p)$ this pmf and we will use the notation by $X \sim \mathrm{Bin}(n,p)$ to indicate that X is a discrete random variable with this pmf. Note that $\mathrm{Ber}(p) = \mathrm{Bin}(1,p)$.

The probability mass of $\mathrm{Bin}(n = 10, p = 0.4)$ is depicted in Figure 2.3. It has only one mode, located at 4. This is the most probable value of $\mathrm{Bin}(10, 0.4)$. The r.v. in Example 2.12 is $\mathrm{Bin}(n, 1/6)$.

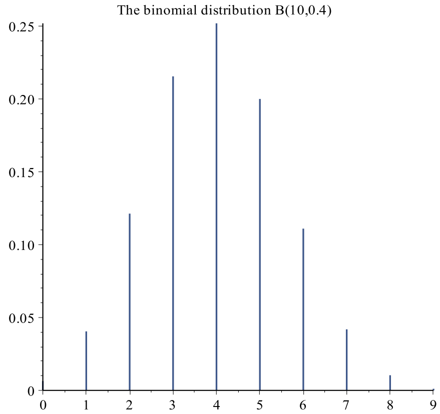

Figure 2.3. *The binomial distribution* $\mathrm{Bin}(10, 0.4)$.

The graph of the CDF of $\text{Bin}(n = 10, p = 1/3)$ is depicted in Figure 2.4.

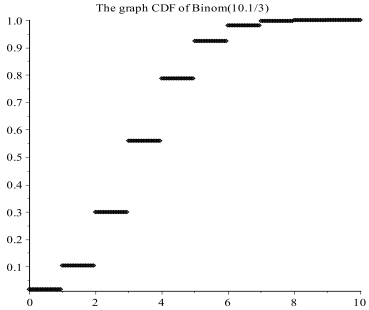

Figure 2.4. *The cdf of binomial distribution* $\text{Bin}(10, 1/3)$.

The binomial random variable can be mechanically simulated with the famous *Galton quincunx* or *Galton board*[2]; see Figure 2.5.

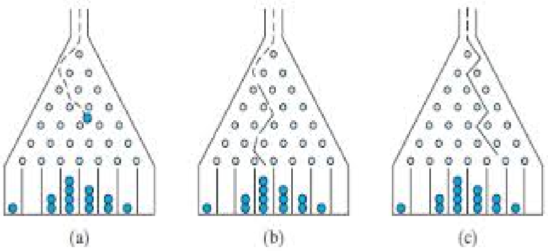

Figure 2.5. *Galton's board*

There is another convenient way of viewing a binomial random variable $X \sim \text{Bin}(n, p)$. Fix an event E with probability $\mathbb{P}(E) = p$. Suppose that we perform a sequence of n independent trials. For $k = 1, \dots, n$ we denote by X_k the random variable that is equal to 1 if the event E has occurred at the k-th trial and equal to zero if it did not.

The variables $X_1, \dots, X_n$ are independent and have the same pmf, namely $\text{Ber}(p)$. Clearly $X_1 + \cdots + X_n$ is the number of successes in n independent trials, i.e.,

$$X = X_1 + \cdots + X_n.$$

Thus, *any binomial random variable* $X \sim \text{Bin}(n, p)$ *is the sum of* n *independent Bernoulli random variables with the same probability of success* p. □

[2] For an interactive computer simulation of a Galton board we refer to the Math is Fun site http://www.mathsisfun.com/data/quincunx-explained.html

Example 2.18 (Propagation of lies)**.** Suppose we have a sequence of 101 people,

$$P_1, \dots, P_{100}, P_{101}.$$

The first 100 people are liars and its is known that they lie independently of each other with equal probability $\frac{1}{2}$.

The person P_1 is told an information, communicates it (truthfully or falsely with equal probabilities) to P_2 who in turn communicates it the same fashion to P_3 and so on until P_{101} receives some information. We want to find out what is the probability that the information that reaches P_{101} is the correct information. We denote by E this event.

1st Method. Note that the transmission consisted of 100 person-to-person communications. We denote by L the number of lies among these 100 communications. Note that $L \sim \mathrm{Bin}(100, 1/2)$.

Note that the information received by P_{101} is the true information if the number of lies during the transmission is even. Thus

$$\begin{aligned}\mathbb{P}(E) &= \mathbb{P}(L=0) + \mathbb{P}(L=2) + \mathbb{P}(L=4) + \cdots \\ &= \binom{n}{0}\frac{1}{2^n} + \binom{n}{2}\frac{1}{2^n} + \binom{n}{4}\frac{1}{2^n} + \cdots \\ &= \frac{1}{2^n}\left(\binom{n}{0} + \binom{n}{2} + \binom{n}{4} + \cdots\right) \overset{(1.8)}{=} \frac{1}{2^n}\cdot 2^{n-1} = \frac{1}{2}.\end{aligned}$$

2nd Method. Here is an alternative way of proving the above equality. The propagation of information can be viewed as a length 100 sequence of T and L's,

$$1 \xrightarrow{L} 2 \xrightarrow{T} 3 \xrightarrow{L} \cdots .$$

where a T/F on the k-th arrow signifies that the k-th transmission was unaltered/changed. All sequences of 100 T's and F's have the same likelihood of occurring namely $\frac{1}{2^n}$.

There is a one-to-one correspondence between the strings of transmissions that with and even number of lies and those with an odd number of lies described by flipping the symbol on the first arrow in the string. This shows that the probability that L is odd is equal to the probability that L is even and these two probabilities add up to 1. □

Example 2.19 (Geometric distributions)**.** Suppose that we perform a sequence of independent Bernoulli trials with success probability p until we get the first success. We let T denote the epoch when we record the first success. Then T is a discrete random variable with range $\{1, 2, \dots\}$. Arguing as in Example 2.13 we deduce that

$$\mathbb{P}(T = n) = pq^{n-1}, \;\; \forall n = 1, 2, \dots \tag{2.7}$$

The above probability distribution is called the *geometric distribution* with probability of success p. We will use the notation $T \sim \text{Geom}(p)$ to indicate that X is a r.v. with such a distribution. The random variable in Example 2.13 is geometric with success probability $1/6$.

The geometric distribution enjoys the *memoryless property*:

> *for $n_0, n > 0$, the conditional probability that we will observe the first success after more than n_0+n trials, given that we performed n trials with recording a success, is independent of n.*

More precisely

$$\boxed{\mathbb{P}(T > n_0 + n | T > n) = \mathbb{P}(T > n_0), \ \ \text{for any } n, n_0 \in \mathbb{N}}. \tag{2.8}$$

□

Example 2.20 (Negative binomial distributions)**.** Fix a natural number k. Suppose that we independently repeat a Bernoulli trial with probability of success p until we register k successes. Denote by T_k the epoch when we register the k-th success. Clearly $\mathbb{P}(T_k < k) = 0$. Note that $T_k = N \geq k$ if and only if at the N-th trial we registered a success and during the previous $N-1$ trials we registered exactly $k-1$ successes. There are $\binom{N-1}{k-1}$ possibilities for the epochs when these $(k-1)$ successes took place. Using the independence and (2.6) we deduce that

$$\mathbb{P}(T_k = N) = p\binom{N-1}{k-1}p^{k-1}q^{N-k} = \binom{N-1}{k-1}p^k q^{N-k}, \ \ N \geq k. \tag{2.9}$$

The above pmf is called the *negative binomial distribution* with probability of success p, and number of successes k. We will use the notation $\text{NegBin}(k,p)$ to denote this pmf and we will use the notation $X \sim \text{NegBin}(k,p)$ to indicate that X is a random variable with this distribution. Note that in the case $k = 1$ we obtain the geometric distribution, i.e., $\text{Geom}(p) \sim \text{NegBin}(1,p)$.

Let us observe that *if $X_1, \dots, X_k$ are independent geometric random variables, with the same probability of success p, then*

$$X_1 + \cdots + X_k \sim \text{NegBin}(k,p).$$

In Example 7.10 we explain how to operate in R with these random variables.

□

Example 2.21 (Banach's Problem)**.** In an address honoring Stefan Banach,[3] Hugo Steinhaus[4] made a humorous reference to the smoking habits of the famous mathematician.

An eminent mathematician fuels a smoking habit by keeping matches in both trouser pockets. When impelled by need, he reaches a hand into a randomly selected pocket and grubs about for a match. Suppose he starts with n matches in each pocket. What is the probability that when he first discovers a pocket to be empty of matches the other pocket contains exactly m matches?

Denote by L_m the event "the first empty pocket is the left one and there are m matches remaining in the right pocket". Denote by R_m the event "the first empty pocket is the right one and there are m matches remaining in the left pocket". We are looking for the probability of $L_m \cup R_m$. Clearly these events are disjoint and are equally likely so

$$\mathbb{P}(L_m \cup R_m) = \mathbb{P}(L_m) + \mathbb{P}(R_m) = 2\mathbb{P}(L_m).$$

Model this as a sequence of Bernoulli trials with success probability $p = \frac{1}{2}$ of choosing the left pocket and failure probability $q = \frac{1}{2}$ of choosing the right pocket. The event L_m occurs after $N = n+1+n-m = 2n+1-m$ trials during which we registered exactly $n+1$ successes, with the $(n+1)$-th occurring at the last trial.

We are looking at a negative binomial random variable T_{n+1} , the epoch of the $(n+1)$-th success, and we are interested in the probability
$\mathbb{P}(T_{n+1} = 2n+1-m)$. Using (2.9) with $N = 2n+1-m$ and $k = n+1$ we deduce

$$\begin{aligned}\mathbb{P}(L_m) = \mathbb{P}(T_{n+1} = 2n+1-m) &= \binom{2n+1-m-1}{n+1-1}\frac{1}{2^{2n+1-m}}\\ &= \binom{2n-m}{n}\frac{1}{2^{2n+1-m}}.\end{aligned}$$

Thus the probability we are seeking is

$$2\mathbb{P}(L_m) = 2\binom{2n-m}{n}\frac{1}{2^{2n+1-m}} = \binom{2n-m}{n}\frac{1}{2^{2n-m}}. \qquad \square$$

Example 2.22 (The hypergeometric distribution)**.** Suppose that we have a bin containing w white balls and b black balls. We select n balls at random from the bin and we denote by X the number of white balls among the selected ones. This is a random variable with range $0, 1, \dots, n$ called the *hypergeometric random*

[3] Stefan Banach (1892-1945) was a Polish mathematician who is generally considered one of the world's most important and influential 20th-century mathematicians.
https://en.wikipedia.org/wiki/Stefan_Banach

[4] Hugo Steinhaus (1887-1972) was a Polish mathematician and educator. He is credited with "discovering" the mathematician Stefan Banach, with whom he gave a notable contribution to functional analysis.
https://en.wikipedia.org/wiki/Hugo_Steinhaus

variable with parameters w, b, n. We will use the notation $X \sim \mathrm{HGeom}(w, b, n)$ to indicate this and we will refer to its pmf as the *hypergeometric distribution*. For example, if A is the number of aces in a random poker hand, then $A \sim \mathrm{HGeom}(4, 48, 5)$.

To compute $\mathbb{P}(X = k)$ when $X \sim \mathrm{HGeom}(w, b, n)$ we use the formula (F/P). The number of possible outcomes is

$$\binom{w+b}{n}.$$

A favorable outcome for the event $X = k$ is determined by a choice of k white balls (out of w) and another independent choice of $n - k$ black balls (out of b). Thus, the number of favorable outcomes is

$$\binom{w}{k}\binom{b}{n-k},$$

so that

$$\boxed{\mathbb{P}(X = k) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}}}. \qquad \square$$

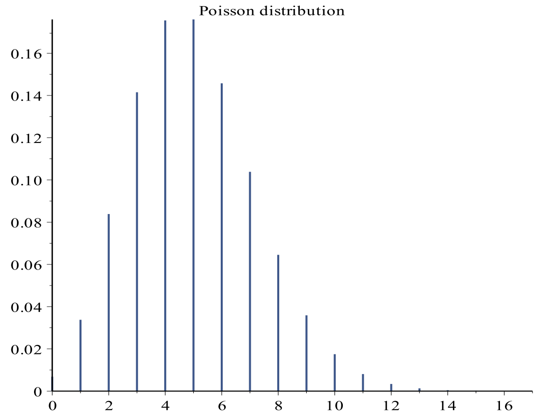

Figure 2.6. *The Poisson distribution with parameter* $\lambda = 5$.

Example 2.23 (The Poisson distribution)**.** A *Poisson random variable* with parameter $\lambda > 0$ is a discrete random variable with range $\{0, 1, 2, \dots\}$ and probability mass function

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

We will denote by $\mathrm{Poi}(\lambda)$ this pmf and we will use the notation $X \sim \mathrm{Poi}(\lambda)$ to indicate that the pmf of X is $\mathrm{Poi}(\lambda)$.

Note that the numbers $(p_k(\lambda))_{k\geq 0}$ do indeed satisfy the normalization condition (2.5)

$$p_0(\lambda)+p_1(\lambda)+p_2(\lambda)+\cdots = e^{-\lambda}\underbrace{\left(1+\frac{\lambda}{1!}+\frac{\lambda^2}{2!}+\cdots\right)}_{=e^{\lambda}} = 1.$$

In Example 7.10 we explain how to operate with Poisson variables in R.

The Poisson distribution typically models the occurrence of rare events in a given unit of time. Then $p_k(\lambda)$ would be the probability that k of these events took place during that unit of time. One can show that for *fixed* λ, if n is very large so $p_n=\lambda/n$ is very small, then

$$\operatorname{Bin}(n,p_n)\approx \operatorname{Poi}(\lambda),\quad np_n=\lambda. \tag{2.10}$$

We describe below a simple argument justifying (2.10). Fix a natural number k and a positive number λ. Suppose that we run the same experiment a large number N of times. The probability of success in each experiment is assumed very small

$$p\approx\frac{\lambda}{N}.$$

As usual, set $q=1-p$. The probability of having exactly k successes in this long sequence of N trials is then

$$\begin{aligned}\binom{N}{k}p^kq^{N-k} &\approx \frac{N(N-1)\cdots(N-k+1)}{k!}\frac{\lambda^k}{N^k}\left(1-\frac{\lambda}{N}\right)^{N-k}\\ &= \frac{N(N-1)\cdots(N-k+1)}{N^k}\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{N}\right)^{N-k}\\ &= 1\cdot\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\cdots\left(1-\frac{k-1}{N}\right)\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{N}\right)^{N-k}\end{aligned}$$

If we keep k fixed, but let $N\to\infty$, we have

$$1\cdot\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\cdots\left(1-\frac{k-1}{N}\right)\to 1.$$

Next observe that

$$\left(1-\frac{\lambda}{N}\right)^{N-k} = \Bigg[\underbrace{\left(1-\frac{\lambda}{N}\right)^{-\frac{N}{\lambda}}}_{x_N}\Bigg]^{-\frac{\lambda}{N}(N-k)},$$

$$\lim_{N\to\infty}x_N=e,\quad \lim_{N\to\infty}\frac{\lambda}{N}(N-k)=\lambda,$$

so that

$$\lim_{N\to\infty}\left(1-\frac{\lambda}{N}\right)^{N-k}=e^{-\lambda},$$

$$\lim_{N\to\infty}\binom{N}{k}p^kq^{N-k}=\frac{\lambda^k}{k!}e^{-\lambda}=p_k(\lambda). \tag{2.11}$$

Thus $p_k(\lambda)$ is an approximation for the probability of occurrence of k rare events. □

Remark 2.24. The result (2.10) was proved by Siméon D. Poisson in 1837. Perhaps the first person to put Poisson's result to use was von Bortkewitsch in his analysis, published in 1898 of the number of deaths of Prussian officers who were kicked by their steeds. We refer to [**17**, VIII.6] for many interesting applications of this law. □

Example 2.25 (Overbooking)**.** Empirical data suggest that about 12% of all booked passengers do not show up at the gate due to cancellations and no-shows. If an airline sells 110 tickets for a flight that seats 100 passengers, what is the probability that the airline overbooked, i.e., it sold more tickets than seats?

We consider the event "*passenger shows-up at the gate*". The empirical data show that the probability of this event is $1-0.12 = 0.88$. We assume that the 110 passengers booked for that flight make independent decisions, and each decides with probability 0.88 to show up at the gate.

We are thus dealing with a binomial experiment, with 110 trials and probability of success 0.88. Denote by X the number of passengers that show-up at the gate. The probability $F(x) := \mathbb{P}(X \leq x)$ is the cumulative distribution function of $\mathrm{Bin}(110, 0.88)$. The probability that the flight was overbooked is then

$$\mathbb{P}(X > 100) = 1 - \mathbb{P}(X \leq 100) = 1 - F_X(100).$$

The computation of the cumulative distribution of the binomial random variables has been implemented in R. More precisely, $F_X(100)$ is computed using the command

```
pbinom(100,110, 0.88)
```

and yields the result

$$F_X(100) \approx 0.8633, \;\; \mathbb{P}(X > 100) \approx 0.1366.$$

Suppose that for a 100-seat flight the airline would like to sell the maximum number of tickets such that the chance of overbooking is less than 5%. Thus we need to find $n > 100$ such that if $X \sim \mathrm{Bin}(n, 0.88)$, then $\mathbb{P}(X > 100) < 0.05$. Using a trial and error approach that can be implemented in R using the command

```
for(i in 101:115){
  print(i)
  print(1-pbinom(100,i,0.88))
}
```

we deduce that for $n = 109$ sold tickets we have

$$\mathbb{P}(X > 100) = 0.0823$$

while for $n = 108$ sold tickets we have

$$\mathbb{P}(X > 100) = 0.0449.$$

Thus, if the company books 108 passengers, the odds of overbooking are less than 5%, i.e., fewer than 1 in 20 flights will be overbooked. □

2.2.2. Probability generating functions. The statistics of a random variable with range contained in the set $0, 1, 2, \dots$ can be efficiently encoded in a single function called the probability generating function associated to the random variable. Any statistical quantity associated to such a random variable can be expressed in terms of its probability generating function.

Definition 2.26 (The probability generating function)**.** Suppose that X is a discrete random variables whose range is contained in the set $\mathbb{N}_0 = \{0, 1, 2, \dots\}$.

For $n \in \mathbb{N}_0$ we set $p_n := \mathbb{P}(X = n)$. The *probability generating function* or *pgf* of X is the function

$$G_X : [0,1] \to \mathbb{R}, \;\; G_X(s) = p_0 + p_1 s + \cdots + p_n s^n + \cdots = \sum_{n=0}^{\infty} p_n s^n. \tag{2.12}$$

□

Remark 2.27. The series in (2.12) has nonnegative terms and it is convergent for any $s \in [0,1]$. Note that

$$G_X(1) = p_0 + p_2 + \cdots + p_n + \cdots = 1.$$

□

Example 2.28. (a) The probability generating function (pgf) of the random variable G in Example 2.10 is

$$\underbrace{\frac{1}{8}}_{\mathbb{P}(G=0)} s^0 + \underbrace{\frac{3}{8}}_{\mathbb{P}(G=1)} s^1 + \underbrace{\frac{3}{8}}_{\mathbb{P}(G=2)} s^2 + \underbrace{\frac{1}{8}}_{\mathbb{P}(G=3)} s^3 = \frac{1}{8}(1+s)^3 = \left(\frac{1+s}{2}\right)^3.$$

(b) The pgf of the Bernoulli variable Ber(p)with probability of success p is

$$\boxed{G_{\mathrm{Ber}(p)}(s) = q + ps, \;\; q = 1 - p}.$$

(c) The pgf of a *binomial random variable* $\mathrm{Bin}(n,p)$ corresponding to n independent Bernoulli trials with success probability p (see Example 2.17) is

$$G_{\mathrm{Bin}(n,p)}(s) = \mathbb{P}(X=0) + \mathbb{P}(X=1)s + \mathbb{P}(X=2)s^2 + \cdots + \mathbb{P}(X=n)s^n$$

$$= \binom{n}{0} q^n + \binom{n}{1} q^{n-1} ps + \binom{n}{2} q^{n-2} p^2 s^2 + \cdots + \binom{n}{n} p^n s^n.$$

Using Newton's binomial formula (1.6) we deduce

$$\boxed{G_{\mathrm{Bin}(n,p)}(s) = (q+ps)^n = G_{\mathrm{Ber}(p)}(s)^n}. \tag{2.13}$$

(d) The pgf of a *geometric random variable* Geom(p) with success probability p (see Example 2.19) is

$$\begin{aligned} G_{\mathrm{Geom}(p)}(s) &= ps + pqs^2 + pq^2s^3 + \cdots + pq^{n-1}s^n + \cdots \\ &= sp(1 + qs + q^2s^2 + \cdots (qs)^{n-1} + \cdots \boxed{= \frac{ps}{1-qs}}. \end{aligned} \tag{2.14}$$

(e) Let NegBin(k,p) be the negative binomial distribution corresponding to the waiting time of k successes in a sequence of Bernoulli trials with success probability p. Then its pgf is

$$\boxed{G_{\mathrm{NegBin}(k,p)}(s) = (ps)^k(1-qs)^{-k} = \left(\frac{ps}{1-qs}\right)^k = G_{\mathrm{Geom}(p)}(s)^k}. \tag{2.15}$$

The equality

$$G_{\mathrm{NegBin}(k,p)}(s) = G_{\mathrm{Geom}(p)}(s)^k.$$

is no accident. It is a manifestation of a more general principle that we will discuss later in Section **??** when we give a very simple proof of (2.15).

Here is an elementary, but more involved proof of (2.15). Using (2.9) we deduce

$$G_{\mathrm{NegBin}(k,p)}(s) = \sum_{n\geq k}\binom{n-1}{k-1}p^kq^{n-k}s^n = p^kq^{-k}\sum_{n\geq k}\binom{n-1}{k-1}(qs)^n$$

$(n = k+m)$

$$= p^kq^{-k}\sum_{m\geq 0}\binom{k+m-1}{k-1}(qs)^{k+m} = (ps)^k\sum_{m\geq 0}\binom{k+m-1}{m}(qs)^m.$$

The last sum can be dramatically simplified. To see how, start with the well known equality

$$(1-x)^{-1} = 1 + x + x^2 + \cdots, \quad |x| < 1.$$

Derivating this equality we obtain

$$(1-x)^{-2} = 1 + 2x + 3x^2 + \cdots, \quad |x| < 1.$$

Derivating again we deduce

$$2(1-x)^{-3} = 2 + 3\cdot 2\,x + 4\cdot 3\,x^2 + \cdots, \quad |x| < 1.$$

Derivating $(k-1)$ times we obtain

$$\begin{aligned} (k-1)!(1-x)^{-k} &= 1\cdot 2\cdots(k-1)\,(1-x)^{-k} \\ &= (k-1)! + k(k-1)\cdots 2\,x + (k+1)(k)\cdots 3\,x^2 + \cdots, \quad |x| < 1. \end{aligned}$$

Hence

$$\begin{aligned} (1-x)^{-k} &= \frac{(k-1)!}{(k-1)!} + \frac{k(k-1)\cdots 2}{(k-1)!}x + \frac{(k+1)(k)\cdots 3}{(k-1)!}x^2 + \cdots \\ &= \binom{k-1}{k-1} + \binom{k}{k-1}x + \binom{k+1}{k-1}x^2 + \cdots \\ &= \sum_{m\geq 0}\binom{m+k-1}{k-1}x^m = \sum_{m\geq 0}\binom{m+k-1}{m}x^m, \quad |x| < 1. \end{aligned}$$

This proves (2.15).

(f) The probability generating function of a *hypergeometric* random variable $\mathrm{HGeom}(w,b,n)$ (see Example 2.22) is

$$\boxed{G_{\mathrm{HGeom}(w,b,n)}(s) = \frac{1}{\binom{w+b}{n}}\sum_{k=0}^{w}\binom{w}{k}\binom{b}{n-k}s^k}. \tag{2.16}$$

(g) The probability generating function of a *Poisson random variable* $\mathrm{Poi}(\lambda)$ with parameter $\lambda > 0$ (see Example 2.23) is

$$G_{\mathrm{Poi}(\lambda)}(s) = p_0(\lambda) + p_1(\lambda)s + p_2(\lambda)s^2 + \cdots = e^{-\lambda} + \frac{\lambda}{1!}e^{-\lambda}s + \frac{\lambda^2}{2!}e^{-\lambda}s^2 + \cdots$$
$$= e^{-\lambda}\left(1 + \frac{\lambda s}{1!} + \frac{(\lambda s)^2}{2!} + \cdots + \frac{(\lambda s)^k}{k!} + \cdots\right) = e^{-\lambda}\cdot e^{\lambda s}.$$

Hence

$$\boxed{G_{\mathrm{Poi}(\lambda)}(s) = e^{\lambda(s-1)}}. \tag{2.17}$$

□

2.2.3. Statistical invariants of discrete random variables. The pmf of a discrete random variable contains most of the useful information concerning that random variable. We want to describe below certain numerical invariants of a discrete random variable that are defined in terms of its pmf but are easier to manipulate and often can be computed even when we do not have precise information about the pmf.

Definition 2.29. Suppose that X is a discrete random variable with range $\mathscr{X} = \{x_1, x_2, \dots,\}$ and probability mass function $p : \mathscr{X} \to [0,1]$.

(a) Let $s \in [1,\infty)$. We say that X is *s-integrable* and we write this $X \in L^s$, if

$$\sum_{x\in\mathscr{X}} |x|^s p(x) < \infty. \tag{2.18}$$

We say that X is *integrable* if it it is 1-integrable, i.e.,

$$\boxed{\sum_{x\in\mathscr{X}} |x|p(x) < \infty}. \tag{2.19}$$

(b) If X is integrable, then we define its *expectation* or *mean* to be the real number

$$\boxed{\mathbb{E}[X] := \sum_{x\in\mathscr{X}} xp(x)}.$$

Often, the mean of a discrete random variable X is denoted by the Greek letter μ.

(c) If $n \in \{1, 2, 3, \dots\}$ and X is n-integrable, then we define its *n-th moment* to be the quantity

$$\boxed{\mu_n[X] := \sum_{x\in\mathscr{X}} x^n p(x)}.$$

Note that $\mu_1[X] = \mathbb{E}[X]$. □

Remark 2.30. If a discrete random variable is s-integrable for some $s \geq 1$, then it is r-integrable for any $r \in [1, s]$. □

Example 2.31 (Casino-craps)**.** A casino owner is willing to run craps is his casino he could be reasonably sure that it would yield a profit of \$ 100 per day. The gambler's winning probability is 0.4929 while the house's winning probablity is $1 - 0.4929 = 0.5071$.

Assuming that the house earns a dollar if the gambler loses a game of craps and pays a dollar otherwise, we see that the winning per game of craps is a random variable X with values ± 1 and pmf

$$p = \mathbb{P}(X = 1) = 0.5071, \;\; q = 1 - p = \mathbb{P}(X = -1) = 0.4929.$$

The expectation of this random variable is

$$\mu = \mathbb{E}[X] = 1 \cdot p + (-1) \cdot q = 0.0142$$

This is the expected house winning per game. If the house runs $10,000$ craps games per day, then it can expect to win \$ 142 per day. The following R program simulates the average winning in a sequence of $n = 1000$ games of craps.

```
#Fix the gambler's winning probability
p=0.4929
# The casino's winning probability
q=1-p
prob=c(p,q)
#define the random variable: Casino gets 1 dollar every time it wins
and pays one dollar when it loses
X=c(-1,1)
#the mean of X
mu=sum(X*prob)
mu
# The number of games per day is n
n=10000
"expected winning  per day is"
n*mu
#simulate n games; store the winnings in the vector w
w=sample(X, n, replace=TRUE, prob)
#cummulative winnings
cumwin=cumsum(w)
#average winning per game
avwin=cumwin/(1:n)
plot(1:n , avwin, type="l", xlab="Number of games", ylab="Running average",
main="Average house profite for the game of craps")
abline(h=mu,col="red")
```

Definition 2.32. Let X be a discrete random variable with range $\mathscr{X}$ and probability mass function $p:\mathscr{X}\to[0,1]$. If X is 2-integrable and its mean is $\mu=\mu_1[X]$, then we define its variance to be the quantity

$$\boxed{\boldsymbol{var}[X]=\sum_{x\in\mathscr{X}}(x-\mu)^2p(x)}. \tag{2.20}$$

The *standard deviation* of X is defined to be the quantity

$$\sigma[X]=\sqrt{\boldsymbol{var}[X]}.$$ □

Remark 2.33. Above, the square integrability of X guarantees that the variance of X is finite, even when the range of X is unbounded. □

Proposition 2.34. *Let X be a discrete random variable with range $\mathscr{X}$ and probability mass function $p:\mathscr{X}\to[0,1]$. If X is 2-integrable, then*

$$\boxed{\boldsymbol{var}[X]:=\mu_2[X]-\mu_1[X]^2}. \tag{2.21a}$$

$$\boxed{\boldsymbol{var}[cX]=c^2\,\boldsymbol{var}[X]},\quad \forall c\in\mathbb{R}. \tag{2.21b}$$

Proof. Set $\mu:=\mathbb{E}[X]$. We have

$$\begin{aligned}\boldsymbol{var}[X]&=\sum_{x\in\mathscr{X}}(x-\mu)^2p(x)=\sum_{x\in\mathscr{X}}(x^2-2\mu x+\mu^2)p(x)\\&=\underbrace{\sum_{x\in\mathscr{X}}x^2p(x)}_{=\mu_2(X)}-2\mu\underbrace{\sum_{x\in\mathscr{X}}xp(x)}_{=\mu}+\mu^2\underbrace{\sum_{x\in\mathscr{X}}p(x)}_{=1}=\mu_2[X]-\mu^2.\end{aligned}$$

Next observe that

$$\begin{aligned}\boldsymbol{var}[cX]&:=\mu_2[cX]-\mu_1[cX]^2=c^2\mu_2[X]-c^2\mu_1[X]^2\\&=c^2(\mu_2[X]-\mu_1[X]^2)=c^2\,\boldsymbol{var}[X].\end{aligned}$$

□

Example 2.35. A discrete random variable with *finite* range is integrable because in this case the sum (2.18) consists of finitely many terms.

For example, if G is the random variable in Example 2.10 describing the number of girls in a random family with three children, then its range is $\{0,1,2,3\}$ and its expectation is

$$\mathbb{E}[G]=0p(0)+1p(1)+2p(2)+3p(3)=\frac{3}{8}+\frac{6}{8}+\frac{3}{8}=\frac{12}{8}=\frac{3}{2}=1.5.$$

We should interpret this by saying that the average number of girls in a family with 3 children is 1.5 girls. The variance of G is

$$0^2p(0)+1^2p(1)+2^2p(2)+3^2p(3)-\frac{9}{4}=\frac{3}{8}+\frac{4\cdot 3}{8}+\frac{9\cdot 1}{8}-\frac{9}{4}=\frac{24}{8}-\frac{9}{4}=\frac{3}{4}.$$ □

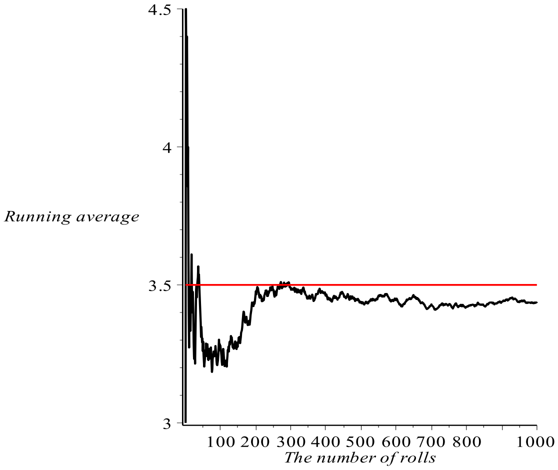

Figure 2.7. *Simulating* 1000 *rolls of a fair die..*

Example 2.36. Let D be the random variable describing the number we get after rolling one fair die. Its range is $\{1,2,3,4,5,6\}$, and each number is equally likely. We deduce that

$$\mathbb{E}[D] = \frac{1+2+3+4+5+6}{6} = \frac{21}{6} = \frac{7}{2} = 3.5.$$

One can interpret this as follows. Roll the die a large number N of times so we get the numbers $d_1, d_2, \dots, d_N$. Then, with high confidence, the average number

$$\frac{d_1 + \cdots + d_N}{N}$$

is close to 3.5. This phenomenon is depicted in Figure 2.7 which shows what is the average number we obtain after $n = 1, \dots, 1000$ rolls. The variance of D is

$$\boldsymbol{var}[D] = \frac{1^2+2^2+3^2+4^2+5^2+6^2}{6} - \frac{49}{4} = \frac{35}{12} \approx 2.916. \qquad \square$$

Example 2.37. Suppose we roll a pair of dice and S is the sum of the two numbers we observe. The range of S is $\{2,3,\dots,12\}$. We have

$$\begin{aligned}\mathbb{E}[S] &= 2\mathbb{P}(S=2) + 3\mathbb{P}(S=3) + \cdots + 7\mathbb{P}(S=7) + \ldots + 12\mathbb{P}(S=12)\\ &= \frac{2\cdot 1 + 3\cdot 2 + \cdots + 7\cdot 6 + 8\cdot 5 + \cdots + 12\cdot 1}{36}\\ &= \frac{2+6+12+20+30+42+40+36+30+22+12}{36} = \frac{252}{36} = 7.\end{aligned}$$

Proposition 2.38. *Suppose that X is an integrable, discrete, random variable with range contained in $0,1,2,\dots$, then we have*

$$\boxed{\mathbb{E}[X] = \sum_{n\geq 0} \mathbb{P}(X > n)}. \tag{2.22}$$

Proof. We have

$$\sum_{n\geq 0}\mathbb{P}(X>n)=\mathbb{P}(X>0)+\mathbb{P}(X>1)+\mathbb{P}(X>2)+\cdots$$

$$=\underbrace{\mathbb{P}(X=1)+\mathbb{P}(X=2)+\mathbb{P}(X=3)+\cdots}_{\mathbb{P}(X>0)}$$

$$+\underbrace{\mathbb{P}(X=2)+\mathbb{P}(X=3)+\mathbb{P}(X=4)+\cdots}_{\mathbb{P}(X>1)}$$

$$+\underbrace{\mathbb{P}(X=3)+\mathbb{P}(X=4)+\mathbb{P}(X=5)+\cdots}_{\mathbb{P}(X>2)}$$

$$+\cdots$$

$$=\mathbb{P}(X=1)+2\mathbb{P}(X=2)+3\mathbb{P}(X=3)+\cdots=$$

□

Example 2.39. Consider again situation in Example 2.13. Suppose that X is the number of rolls of a die until the first 6 appears. This is a geometric random variable with success probability $1/6$. Its range is $\{1,2,\dots\}$ and, as explained in Example 2.13 we have

$$\mathbb{P}(X>n)=\left(\frac{5}{6}\right)^n.$$

Hence

$$\mathbb{E}[X]=1+\frac{5}{6}+\left(\frac{5}{6}\right)^2+\cdots+\left(\frac{5}{6}\right)^n+\cdots=\frac{1}{1-\frac{5}{6}}=6.$$

Thus the mean (or expectation) of X is 6.

This is a very intuitive conclusion: the chance of getting a 6, when rolling a die, is 1 in 6 so, on average, would should expect 6 rolls of the die until we get the first 6. □

Proposition 2.40. *Suppose that X is a discrete random variable with range contained in $\mathbb{N}_0=\{0,1,2,\dots\}$. If $G_X(s)$ is the pgf of X, then*

$$G_X(1)=1 \tag{2.23a}$$

$$\mathbb{E}[X]=G_X'(1). \tag{2.23b}$$

$$\mu_2[X]=G_X''(1)+G_X'(1), \tag{2.23c}$$

$$\boldsymbol{var}[X]=G_X''(1)+G_X'(1)-G_X'(1)^2. \tag{2.23d}$$

Proof. We have

$$G_X(s) = p_0 + p_1 s + p_2 s^2 + \cdots + p_n s^n + \cdots,$$

so

$$G_X(1) = p_0 + p_1 + p_2 + \cdots = 1.$$

Next,

$$G_X'(s) = 0p_0 + 1p_1 + 2p_2 s + \cdots + np_n s^{n-1} + \cdots,$$

$$G_X'(1) = 0p_0 + 1p_1 + 2p_2 + \cdots + np_n + \cdots = \mathbb{E}[X].$$

Finally,

$$G_X''(s) = \sum_{n\geq 0} n(n-1)p_n s^{n-2},$$

so that

$$\begin{aligned} G_X''(1) &= \sum_{n\geq 0} n(n-1)p_n = \sum_{n\geq 0} n^2 p_n - \sum_{n\geq 0} np_n \\ &= \mathbb{E}[X^2] - \mathbb{E}[X] = \mu_2[X] - G_X'(1). \end{aligned}$$

This proves (2.23c). The equality (2.23d) now follows from (2.23b) and (2.20). □

Example 2.41 (Binomial distributions)**.** Suppose that $X \sim \mathrm{Bin}(n,p)$ is a *binomial random variable* corresponding to n independent Bernoulli trials with success probability p. Then its probability generating function is (see Example 2.28(c))

$$G_X(s) = (q+ps)^n, \;\; q = 1-p.$$

We have

$$G_X'(s) = np(q+ps)^{n-1}, \;\; \mathbb{E}[X] = G_X'(1) = np(p+q)^{n-1} = np.$$

Next,

$$G_X''(s) = n(n-1)p^2(q+ps)^{n-2}, \;\; G_X''(1) = n(n-1)p^2.$$

$$\boldsymbol{var}[X] = n(n-1)p^2 + np - (np)^2 = np - np^2 = np(1-p) = npq.$$

We can rewrite this as

$$\boxed{\mathbb{E}[\mathrm{Bin}(n,p)] = np, \;\; \boldsymbol{var}[\mathrm{Bin}(n,p)] = npq}, \tag{2.24}$$

Let us observe a simple consequence of the above equality namely

$$np = \mathbb{E}\big[\, \mathrm{Bin}(n,p)\,\big] = \binom{n}{1}p^1q^{n-1} + \cdots + k\binom{n}{k}p^kq^{n-k} + \cdots + n\binom{n}{n}q^n. \tag{2.25}$$

□

Example 2.42 (The geometric distribution)**.** Suppose that $T_1 \sim \text{Geom}(p)$ is the waiting time for the first success in a sequence of independent Bernoulli trials with success probability p. The r.v. T_1 is a *geometric random variable* with success probability p. Its probability generating function is given by (2.14),

$$G_{T_1}(s) = \frac{ps}{1-qs}, \quad q = 1-p.$$

We have

$$G'_{T_1}(s) = \frac{p(1-qs) - ps(-q)}{(1-qs)^2} = \frac{p}{(1-qs)^2},$$

so that

$$\mathbb{E}[X] = G'_{T_1}(1) = \frac{p}{(1-q)^2} = \frac{1}{p}.$$

This is in perfect agreement with the special case $p = 1/6$ analyzed in Example 2.39 using a different method. Next

$$G''_{T_1}(s) = \frac{2pq}{(1-qs)^3}, \quad G''_{T_1}(1) = \frac{2q}{p^2}$$

so that

$$\boldsymbol{var}[T_1] = \frac{2q}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{2q}{p^2} + \frac{p-1}{p^2} = \frac{q}{p^2}.$$

Hence

$$\boxed{\mathbb{E}[\text{Geom}(p)] = \frac{1}{p}, \quad \boldsymbol{var}[\text{Geom}(p)] = \frac{q}{p^2}}. \tag{2.26}$$

□

Example 2.43 (Negative binomial distributions)**.** The probability distribution of the waiting time T_k to observe k successes in a sequence of Bernoulli trials with success probability p, is the so called *negative binomial distribution* corresponding to k successes with probaility p. We write this $T_k \sim \text{NegBin}(k,p)$. Using (2.15) we deduce as above that

$$\mathbb{E}[T_k] = G'_{T_k}(1) = \frac{k}{p} = k\mathbb{E}[T_1].$$

A similar computation shows that

$$\boldsymbol{var}[T_k] = k\,\boldsymbol{var}[T_1] = \frac{kq}{p^2}.$$

Hence

$$\boxed{\mathbb{E}[\text{NegBin}(k,p)] = \frac{k}{p}, \quad \boldsymbol{var}[\text{NegBin}(k,p)] = \frac{kq}{p^2}}. \tag{2.27}$$

□

Example 2.44 (Hypergeometric distributions)**.** Suppose that $X\sim \mathrm{HGeom}(w,b,n)$ is a *hypergeometric random variable*; see Example 2.22. Its probability generating function is

$$G_X(s)=\frac{1}{\binom{N}{n}}\sum_{k=0}^{w}\binom{w}{k}\binom{b}{n-k}s^k,\;\; N:=w+b.$$

We can identify $G_X(s)$ as the coefficient of x^n in the polynomial

$$Q(s,x)=\frac{1}{\binom{N}{n}}(1+sx)^w(1+x)^b.$$

We write this

$$G_{\mathrm{HGeom}(w,b,n)}(s)=\mathrm{Coeff}\left(x^n:\;\frac{1}{\binom{N}{n}}(1+sx)^w(1+x)^b\right). \tag{2.28}$$

We have

$$\frac{\partial Q}{\partial s}(s,x)=\frac{wx(1+x)^b}{\binom{N}{n}}(1+sx)^{w-1},$$

The mean of X is $G_X'(1)$ and it is equal to the coefficient of x^n in

$$\frac{\partial Q}{\partial s}(1,x)=\frac{wx}{\binom{w+b}{n}}(1+x)^{N-1}=\frac{w\binom{N-1}{n-1}}{\binom{N}{n}}=\frac{wn}{N}=\frac{wn}{w+b}.$$

Hence

$$\boxed{\mathbb{E}\big[\;\mathrm{HGeom}(w,b,n)\;\big]=\frac{w}{w+b}\cdot n}. \tag{2.29}$$

We have

$$\frac{\partial^2 Q}{\partial s^2}(s,x)=\frac{w(w-1)x^2(1+x)^b}{\binom{N}{n}}(1+sx)^{w-2}.$$

Next, $G_X''(1)$ is the coefficient of x^n in

$$\frac{\partial^2 Q}{\partial s^2}(1,x)=\frac{w(w-1)x^2(1+x)^{N-2}}{\binom{N}{n}}$$

so

$$G_X''(1)=\frac{w(w-1)\binom{N-2}{n-2}}{\binom{N}{n}}=w(w-1)\frac{\frac{(N-2)!}{(n-2)!(N-n)!}}{\frac{N!}{n!(N-n)!}}=w(w-1)\frac{n(n-1)}{N(N-1)}=\mu\frac{(w-1)(n-1)}{N-1}.$$

We deduce

$$\begin{aligned}\boldsymbol{var}[X]=G_X''(1)+G_X'(1)-G_X'(1)^2&=\mu\frac{(w-1)(n-1)}{N-1}+\mu-\mu^2\\&=\mu\cdot\frac{(w-1)(n-1)+N-1}{N-1}-\mu^2.\end{aligned}$$

□

Example 2.45 (Poisson distributions)**.** Suppose that $X \sim \mathrm{Poi}(\lambda)$ is a *Poisson random variable* with parameter λ; see Example 2.23. Its probability generating function is (see Example 2.28(g)

$$G_X(s) = e^{\lambda(s-1)}.$$

We have

$$G_X'(s) = \lambda e^{\lambda(s-1)},\;\; G_X''(s) = \lambda^2 e^{\lambda(s-1)},\;\; G_X''(1) = \lambda^2$$

so that

$$\mathbb{E}[X] = G_X'(1) = \lambda,\;\; G_X''(1) = \lambda^2,\;\; \boldsymbol{var}[X] = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

Hence,

$$\boxed{\mathbb{E}\big[\,\mathrm{Poi}(\lambda)\,\big] = \lambda,\;\; \boldsymbol{var}\,\big[\,\mathrm{Poi}(\lambda)\,\big] = \lambda}. \qquad \square$$

2.2.4. Functions of a discrete random variable. Suppose that X is a discrete random variable with range $\mathscr{X}$ and probability mass function $p : \mathscr{X} \to [0,1]$. Then, for any function $f : \mathbb{R} \to \mathbb{R}$, we get a new random variable $f(X)$ with the property that if the value of X for a random experiment is x, then the value of $f(X)$ for the same experiment is $f(x)$. Note that

$$\mathbb{P}\big(\,f(X) = y\,\big) = \sum_{\substack{x\in\mathscr{X}\\ f(x)=y}} p(x).$$

Example 2.46. Suppose that X is a discrete random variable with range

$$\mathbb{Z} = \{0, \pm 1, \pm 2, \dots\},$$

and pmf

$$p_n := \mathbb{P}(X = n),\quad \forall n \in \mathbb{Z}.$$

Then X^2 is the random variable with range

$$\{0^2, 1^2, 2^2, 3^2, \dots\} = \{0, 1, 4, 9, \dots\},$$

and pmf

$$\mathbb{P}(X^2 = 0) = p_0,\;\; \mathbb{P}(X^2 = 1) = p_1 + p_{-1},\;\; \mathbb{P}(X^2 = 2^2) = p_{-2} + p_2, \dots . \qquad \square$$

Theorem 2.47 (The law of the subconscious statistician)**.** [5] *Suppose that X is a discrete random variable with range*

$$\mathscr{X} = \{x_1, x_2, \dots\}$$

and pmf

$$p(x_k) = \mathbb{P}(X = x_k),\quad k = 1, 2, \dots .$$

Then, for any function f, if the expectation of $f(X)$, it is given by

$$\boxed{\mathbb{E}[\,f(X)\,] = \sum_{k\geq 1} f(x_k)p(x_k)}. \qquad \square$$

[5] Sometime this is also known as the law of the *unconscious* statistician or LotUS.

Corollary 2.48 (Linearity of the expectation). *Suppose that X is a discrete random variable and f, g are functions. Then*

$$\boxed{\mathbb{E}\big[\,f(X)+g(X)\,\big]=\mathbb{E}\big[\,f(X)\,\big]+\mathbb{E}\big[\,g(X)\,\big]}.$$

In particular, for any real constants a, b we have

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

Proof. Suppose that the range of X is

$$\mathscr{X}=\{x_1,x_2,\dots\}$$

and its pmf is

$$p_k=\mathbb{P}(X=x_k),\;\;k=1,2,\dots.$$

Then

$$\sum_{k\geq 1}\big(\,f(x_k)+g(x_k)\,\big)p_k=\sum_{k\geq 1}f(x_k)p_k+\sum_{k\geq 1}g(x_k)p_k=\mathbb{E}\big[\,f(X)\,\big]+\mathbb{E}\big[\,g(X)\,\big].$$

□

Using (2.20) we obtain the following result.

Corollary 2.49. *Suppose that X is a 2-integrable discrete random variable with mean μ. Then*

$$\boxed{\mu_n[X]=\mathbb{E}[X^n],\;\;\forall n\in\{1,2,\dots\}},$$

$$\boxed{\boldsymbol{var}[X]=\mathbb{E}[X^2]-\mathbb{E}[X]^2=\mathbb{E}\big[\,(X-\mu)^2\,\big]}.$$ □

Moreover, for any $a, b\in\mathbb{R}$ we have

$$\boxed{\boldsymbol{var}[aX+b]=a^2\,\boldsymbol{var}[X]}.$$

Proof. Only the last equality requires a proof. Denote by μ_X the mean of X. We set $Y:=aX+b$. From Corollary 2.48 we deduce that the mean μ_Y of Y is $\mu_Y=a\mu_X+b$. Hence

$$Y-\mu_Y=aX+b-(a\mu_X+b)=a(X-\mu_X),$$

so

$$\boldsymbol{var}[Y]=\mathbb{E}\big[\,(Y-\mu_Y)^2\,\big]=\mathbb{E}[a^2(X-\mu_X)^2\,\big]=a^2\mathbb{E}\big[\,(X-\mu_X)^2\,\big]=a^2\,\boldsymbol{var}[X].$$

□

Corollary 2.50. *Suppose that X is a discrete random variable with range contained in $\{0,1,2,\dots\}$. We denote by $G_X(s)$ its probability generating function. Then, for any $s\in[0,1]$ we have*

$$G_X(s)=\mathbb{E}[s^X]. \tag{2.30}$$

Proof. For $n = 0, 1, 2, \dots$ we set $p_n := \mathbb{P}(X = n)$. The law of subconscious statistician shows that

$$\mathbb{E}[s^X] = p_0 s^0 + p_1 s^1 + p_2 s^2 + \cdots = G_X(s).$$

□

Corollary 2.51 (Monotonicity of the expectation)**.** *Suppose that X is a discrete r.v. with range $\mathscr{X}$ and $f, g : \mathscr{X} \to \mathbb{R}$ are functions such that*

$$f(x) \leq g(x), \ \ \forall x \in \mathscr{X}.$$

Then

$$\mathbb{E}\big[\, f(X)\,\big] \leq \mathbb{E}\big[\, g(X)\,\big].$$

Proof. Let $p : \mathscr{X} \to [0,1]$ denote the pmf of X. From the law of the subconscious statistician we deduce

$$\mathbb{E}\big[\, f(X)\,\big] = \sum_{x\in\mathscr{X}} f(x)p(x) \leq \sum_{x\in\mathscr{X}} g(x)p(x) = \mathbb{E}\big[\, g(X)\,\big].$$

□

Example 2.52 (Markov inequality)**.** Suppose that Y is a discrete random variable with range $\mathscr{Y}$ consisting only of nonnegative numbers and pmf $p(y)$. Then we have the *classical Markov inequality*

$$\boxed{Y \geq 0 \Rightarrow \mathbb{P}\big(\, Y > c\,\big) \leq \frac{1}{c}\mathbb{E}\big[\, Y\,\big], \ \ \forall c > 0}. \tag{2.31}$$

Let $c > 0$

$$\mathbb{P}(Y > c) = \sum_{y\in\mathscr{Y},\, y>c} p(y)$$

so

$$c\mathbb{P}(Y > c) = \sum_{y\in\mathscr{Y},\, y>c} cp(y) < \sum_{y\in\mathscr{Y},\, y>c} yp(y) \leq \sum_{y\in\mathscr{Y}} yp(y) = \mathbb{E}[Y].$$

□

Theorem 2.53 (Chebyshev's inequality)**.** *Suppose that X is a 2-integrable discrete random variable. Denote by μ its expectation, $\mu = \mathbb{E}[X]$, by v its variance $v = \mathbb{E}[(X-\mu)^2]$ and by σ its standard deviation, $\sigma = \sqrt{v}$. Then for any positive numbers c, r we have*

$$\boxed{\mathbb{P}\big(\, |X-\mu| > c\sigma\,\big) \leq \frac{1}{c^2}}. \tag{2.32a}$$

$$\boxed{\mathbb{P}\big(\, |X-\mu| > r\,\big) \leq \frac{\sigma^2}{r^2}}. \tag{2.32b}$$

Proof. Let $\mathscr{X}$ be the range of X and $p : \mathscr{X} \to [0,1]$ be the associated probability mass function. We have

$$c^2\sigma^2\mathbb{P}\big(\,|X-\mu|>c\sigma\,\big) = c^2\sigma^2 \sum_{x\in\mathscr{X},\,|x-\mu|>c\sigma} p(x) = \sum_{x\in\mathscr{X},\,|x-\mu|>c\sigma} c^2\sigma^2 p(x)$$

$$\leq \sum_{x\in\mathscr{X},\,|x-\mu|>c\sigma} (x-\mu)^2 p(x) \leq \sum_{x\in\mathscr{X}} (x-\mu)^2 p(x) = v = \sigma^2.$$

This proves (2.32a) The inequality (2.32b) is obtained by choosing c such that $c\sigma = r$, i.e., $c = \frac{r}{\sigma}$.

□

Remark 2.54. In more pedestrian terms, Chebyshev's inequality states that the probability that the actual value of a random variable deviates from the mean by a multiple c of the standard deviation is $1/c^2$. For example, if $c = 100$, then the odds that X is more than 100 deviations away from its mean are less than 1 in $10,000$. Note that if the standard deviation is very, very small, then 100 standard deviations is small quantity. Hence the probability of deviating from the mean by a small quantity is small if the standard deviation is very, very small. □

2.3. Continuous random variables

2.3.1. Definition and basic invariants. The lifespan of a bulb can be any nonnegative *real* number. This random quantity does not have a discrete range and it is in some sense "continuous". Here is a more precise definition.

Definition 2.55. A random variable X is called *continuous* if there exists a function

$$p : \mathbb{R} \to [0,\infty)$$

such that

$$\mathbb{P}(X \leq x) = \int_{-\infty}^{x} p(s)ds, \quad \forall x \in \mathbb{R}.$$

The function p is called the *probability density function* (pdf) of X. The random variable is said to be *concentrated on an interval* I if $p(x) = 0$ for all $x \notin I$. □

Remark 2.56. (a) If X is a continuous r.v. with probability density function p, then $p(x)$ has the following statistical interpretation:

the probability that the value of X is located in the infinitesimal interval $[x, x+dx]$ is $p(x)dx$.

The *cumulative distribution function* (cdf) of X is

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} p(s)ds.$$

In particular, we have

$$\boxed{1 = \mathbb{P}(X < \infty) = \int_{-\infty}^{\infty} p(x)dx}. \tag{2.33}$$

Moreover $F'(x) = p(x)$ for all but countably many x.

(b) Given a fixed real number c, we have

$$\boxed{\mathbb{P}(X < c) = \mathbb{P}(X \leq c) = F_X(c)}.$$

Indeed,

$$\mathbb{P}(X < c) = \lim_{\varepsilon \searrow 0} \mathbb{P}(X \leq c - \varepsilon) = \lim_{\varepsilon \searrow 0} \int_{-\infty}^{c-\varepsilon} p(s)ds = \int_{-\infty}^{c} p(s)ds = \mathbb{P}(X \leq c).$$

(c) For any real numbers $a \leq b$ we have

$$\boxed{\begin{aligned} \mathbb{P}(a \leq X \leq b) &= \mathbb{P}(X \leq b) - \mathbb{P}(X < a) \\ &= F_X(b) - F_X(a) = \int_a^b p(s)ds. \end{aligned}} \tag{2.34}$$

(d) Given a fixed real number c, the probability that $X = c$ is zero. Indeed

$$\mathbb{P}(X = c) = \mathbb{P}(c \leq X \leq c) = \mathbb{P}(X \leq x) - \mathbb{P}(X < c) = 0. \qquad \square$$

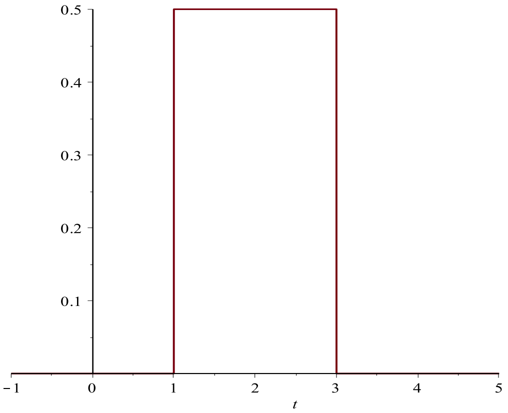

Figure 2.8. *The probability density function of a random variable uniformly distributed on* $[1, 3]$.

The equality (2.33) has a converse.

Proposition 2.57. *Any function* $p : \mathbb{R} \to [0, \infty)$ *satisfying*

$$\int_{-\infty}^{\infty} p(x)dx = 1$$

is the pdf of some continuous random variable. $\square$

Definition 2.58. Suppose that X is a continuous random variable with probability density

$$p:\mathbb{R}\to[0,\infty)$$

and s is a real number ≥ 1.

(a) We say that X is *s-integrable* and we write this $X\in L^s$ if

$$\int_{-\infty}^{\infty}|x|^s p(x)dx<\infty.$$

We say that X is *integrable* if it it is 1-integrable, i.e.,

$$\int_{\mathbb{R}}|x|p(x)<\infty. \tag{2.35}$$

We say that X is *square integrable* if it is 2-integrable, i.e.,

$$\int_{\mathbb{R}}|x|^2p(x)<\infty.$$

(b) If X is integrable, then we define its *expectation* or *mean* to be the real number

$$\boxed{\mathbb{E}[X]:=\int_{\mathbb{R}}xp(x)dx=\int_{-\infty}^{\infty}xp(x)dx}.$$

Often, the mean of a discrete random variable X is denoted by the Greek letter μ.

(c) If $n\in\{1,2,3,\dots\}$ and X is n-integrable, then we define its *n-th moment* to be the quantity

$$\boxed{\mu_n[X]:=\int_{\mathbb{R}}x^np(x)dx}.$$

Note that $\mu_1[X]=\mathbb{E}[X]$.

(d) If X is square integrable, and $\mu=\mathbb{E}[X]$, then we define the *variance* of X to be the quantity

$$\boxed{\boldsymbol{var}[X]:=\int_{\mathbb{R}}(x-\mu)^2p(x)dx}. \tag{2.36}$$

The *standard deviation* of X is defined to be the quantity

$$\boxed{\sigma[X]=\sqrt{\boldsymbol{var}[X]}}.$$ □

Remark 2.59. If a continuous random variable is s-integrable for some $s\geq 1$, then it is r-integrable for any $r\in[1,s]$. □

Proposition 2.60. *If X is square integrable, then*

$$\boxed{\boldsymbol{var}[X]=\mu_2[X]-\mu_1[X]^2}.$$

Proof. We have

$$\begin{aligned}\boldsymbol{var}[X] &= \int_{\mathbb{R}}(x-\mu)^2 p(x)dx = \int_{\mathbb{R}}(x^2-2\mu x+\mu^2)p(x)dx\\ &= \int_{\mathbb{R}} x^2 p(x)dx - 2\mu \underbrace{\int_{\mathbb{R}} xp(x)dx}_{=\mu} + \mu^2 \underbrace{\int_{\mathbb{R}} p(x)dx}_{=1}\\ &= \mu_2[X] - 2\mu^2 + \mu^2 = \mu_2[X] - \mu_1[X]^2.\end{aligned}$$

□

Proposition 2.61. *Suppose that X is a* nonnegative *continuous random variable. Then*

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > x)dx.$$

Idea of proof. Let $p(x)$ be the pdf of X. Denote by $F(x)$ the cdf of X. Then

$$\mathbb{P}(X > x) = 1 - F(x) \Rightarrow \frac{d}{dx}\mathbb{P}(X > x) = -F'(x) = -p(x)dx.$$

Integrating by parts we have

$$\int_0^\infty \mathbb{P}(X > x)dx = \underbrace{x\mathbb{P}(X > x)\Big|_{x=0}^{x=\infty}}_{=0} + \int_0^\infty xp(x) = \mathbb{E}[X].$$

□

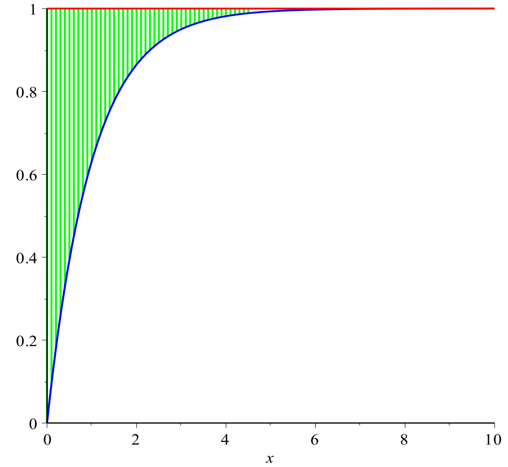

Figure 2.9. *The expectation of a nonnegative random variable (with cdf depicted in blue) is the area of the (green) shaded region.*

Proposition 2.61 has a simple geometric interpretation: the expectation of a nonnegative random variable is the area of the region to the right of the y-axis, between the graph of the cdf $F(x)$ and the horizontal line $y = 1$; see Figure 2.9.

Theorem 2.62 (The law of the subconscious statistician). *Suppose that X is a continuous random variable with probability density $p(x)$ and $gf(x)$ is a function such that $f(X)$ is either a continuous random variable, or a discrete random variable. If $g(X)$ is integrable, then*

$$\mathbb{E}[f(X)] = \int_{\mathbb{R}} g(x)p(x)dx. \tag{2.37}$$

□

The next results are immediate consequences of the law of subconscious statistician.

Corollary 2.63. *Suppose that X is a continuous, integrable random variable with mean μ. Then for any $a, b \in \mathbb{R}$ we have (linearity of expectation)*

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b. \tag{2.38}$$

If X is k-integrable, $c \in \mathbb{R}$, then

$$\boxed{\mu_k[X] = \mathbb{E}[X^k]}, \quad \mu_k[cX] = c^k \mu_k[X]. \tag{2.39}$$

If X is square integrable, then

$$\boxed{\boldsymbol{var}[X] = \mathbb{E}\big[\ (X-\mu)^2\ \big] = \mathbb{E}[X^2] - \mathbb{E}[X]^2}. \tag{2.40a}$$

$$\boxed{\boldsymbol{var}[cX] = c^2\, \boldsymbol{var}[X], \ \ \forall c \in \mathbb{R}}. \tag{2.40b}$$

□

Corollary 2.64 (Monotonicity of the expectation). *Suppose that X is a continuous r.v. with $f, g : \mathbb{R} \to \mathbb{R}$ are functions such that*

$$f(x) \leq g(x), \ \ \forall x \in \mathbb{R}.$$

and each of the random variables $f(X)$ or $g(X)$ is either discrete or continuous. Then

$$\mathbb{E}\big[\,f(X)\,\big] \leq \mathbb{E}\big[\,g(X)\,\big].$$

□

Proposition 2.65 (Markov Inequality). *Suppose that Y is an integrable continuous,* nonnegative *random variable. Then for any $c > 0$ we have*

$$\boxed{\mathbb{P}(Y > c) \leq \frac{1}{c}\mathbb{E}\big[\,Y\,\big]}. \tag{2.41}$$

Proof. Denote by $p_Y(y)$ the density of Y. Since Y is nonnegative, $p_Y(y) = 0$ for $y < 0$. Hence

$$\mathbb{E}[Y] = \int_0^\infty y p_Y(y) dy.$$

Then

$$c\mathbb{P}(Y > c) = \int_c^\infty c p_Y(y) dy \leq \int_c^\infty y p_Y(y) dy \leq \int_0^\infty y p_Y(y) dy = \mathbb{E}[Y].$$

□

Theorem 2.66 (Chebyshev's inequality)**.** *Suppose that X is a 2-integrable continuous random variable. Denote by μ its expectation, $\mu = \mathbb{E}[X]$, by v its variance $v = \mathbb{E}[(X-\mu)^2]$ and by σ its standard deviation, $\sigma = \sqrt{v}$. Then for any positive numbers c, r we have*

$$\boxed{\mathbb{P}\big(\, |X-\mu| \geq c\sigma \,\big) \leq \frac{1}{c^2}}. \tag{2.42a}$$

$$\boxed{\mathbb{P}\big(\, |X-\mu| > r \,\big) \leq \frac{\sigma^2}{r^2}}. \tag{2.42b}$$

□

Proof. Consider the nonnegative random variable $Y = (X-u)^2$. Observe that $|X-\mu| > r$ if and only if $Y \geq r^2$. Markov's inequality implies that, for all $c > 0$,

$$\mathbb{P}(|X-\mu| > r) = \mathbb{P}(Y > r^2) < \frac{1}{r^2}\mathbb{E}\big[\, Y \,\big]$$

$$= \frac{1}{r^2}\mathbb{E}\big[\, (X-\mu)^2 \,\big] = \frac{1}{r^2}\boldsymbol{var}[X] = \frac{\sigma^2}{r^2}.$$

This proves (2.42b). The inequality (2.42a) is obtained by letting $r = c\sigma$ in (2.42b).

□

2.3.2. Important examples of continuous random variables. subh

We describe below a few frequently encountered continuous random variables.

Example 2.67 (Uniform distribution)**.** A continuous r.v. X is said to be *uniformly distributed* on the finite interval $[a,b]$ and we write this

$$X \sim \mathrm{Unif}(a,b)$$

if its pdf is

$$p(x) = \frac{1}{b-a} \times \begin{cases} 1, & x \in [a,b], \\ 0, & x \notin [a,b]. \end{cases}$$

From the definition it follows that

$$\int_{-\infty}^{\infty} p(x)dx = \int_a^b p(x)dx = \frac{1}{b-a}\int_a^b dx = 1.$$

Using (2.34) we deduce that for any interval $I = [c, d] \subset [a, b]$, the probability that X belongs to I is proportional to the length $(d - c)$ of the interval. Indeed

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

This shows that the probability that X takes a value in a given interval does not depend on the location of the interval, but only of its size. Figure 2.8 depicts the graph of this probability density in the special case $[a, b] = [1, 3]$.

We see that X is s-integrable for any $s \geq 1$. Moreover, for any natural number k, the k-th momentum of X is

$$\mu_k[X] = \frac{1}{b-a}\int_a^b x^k dx = \frac{b^{k+1} - a^{k+1}}{(k+1)(b-a)}.$$

In particular, its mean is

$$\mu = \mathbb{E}[X] = \mu_1[X] = \frac{b^2 - a^2}{2(b-a)} = \frac{b+a}{2}.$$

Let us note that $\frac{b+a}{2}$ is the midpoint of the interval $[a, b]$. Its second momentum is

$$\mu_2[X] = \frac{b^3 - a^3}{3(b-a)} = \frac{a^2 + ab + b^2}{3}$$

and its variance is

$$\boldsymbol{var}[X] = \mu_2[X] - \mu_1[X]^2 = \frac{a^2 + ab + b^2}{3} - \frac{(b+a)^2}{4}$$

$$= \frac{4a^2 + 4ab + 4b^2 - 3a^2 + 6ab - 3b^2}{12} = \frac{(b-a)^2}{12}.$$

□

Example 2.68 (Exponential distribution)**.** A continuous random variable T is said to be *exponentially distributed with parameter* $\lambda > 0$, and we write this

$$T \sim \mathrm{Exp}(\lambda)$$

if its pdf is

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

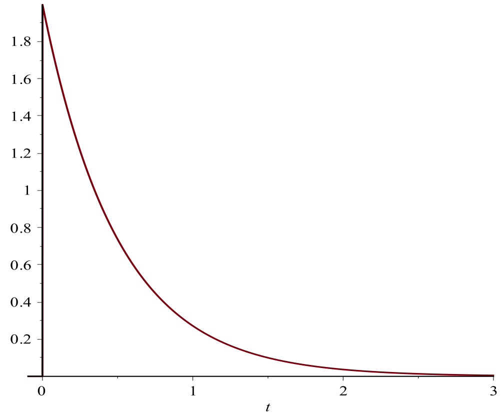

Figure 2.10. *Exponential distribution with parameter $\lambda = 2$.*

The parameter λ is called the *rate* of the exponential random variable. Let us observe that the above function $p(x)$ is a pdf because it is nonnegative and

$$\int_{-\infty}^{\infty} p(t)dt = \lambda \int_0^{\infty} e^{-\lambda t}dt = \lambda \left(-\frac{1}{\lambda} e^{-\lambda t} \right) \Big|_{t=0}^{t=\infty} = 1.$$

Figure 2.10 depicts a portion of the graph of the pdf of an exponentail random variable with parameter $\lambda = 2$.

Denote by $F(t)$ the cumulative distribution function of the random variable T. Clearly $F(t) = 0$, for any $t \leq 0$. For $t > 0$ we have

$$F(t) = \int_0^t \lambda e^{-\lambda s} = \lambda \left(-\frac{1}{\lambda} e^{-\lambda s} \right) \Big|_{s=0}^{s=t} = 1 - e^{-\lambda t}.$$

Typically T models the lifetime of a product (think light bulb, computer, car), or waiting times between two consecutive events, e.g., waiting time for a bus, waiting time between two consecutive orders at an online store etc. The function

$$G(t) = 1 - F(t) = \mathbb{P}(T > t)$$

is called the *survival function* and it gives the probability that the product will last for more that t units of time. Note that

$$\boxed{\mathbb{P}(T > t) = \lambda \int_t^{\infty} e^{-\lambda s}ds = e^{-\lambda t}}.$$

The usefulness of the exponential distribution is due to the *memoryless property*:

> *for $t_0, t > 0$, the conditional probability that the object will last more than $t_0 + t$ units of time, given that it lasted at least t units of time is independent of t.*

More precisely

$$\boxed{\mathbb{P}(T > t_0 + t | T > t) = \mathbb{P}(T > t_0), \ \text{ for any } t, t_0 > 0}. \tag{2.43}$$

Indeed

$$\mathbb{P}(T > t_0 + t | T > t) = \frac{\mathbb{P}(T > t_0 + t, T > t)}{\mathbb{P}(T > t)}$$

$$= \frac{\mathbb{P}(T > t_0 + t)}{\mathbb{P}(T > t)} = \frac{-\lambda(t_0 + t)}{e^{-\lambda t}} = e^{-\lambda t_0} = \mathbb{P}(T > t_0).$$

The exponential random variables are the only continuous random variables with the memoryless property. We refer to [9] for more surprising properties of the exponential distribution.

We see that T is s-integrable for any $s \geq 1$. Moreover, for any natural number k, the k-th momentum of X is

$$\mu_k[T] = \lambda \int_0^\infty t^k e^{-\lambda t} dt$$

($x = \lambda t$, $t = \lambda^{-1} x$, $t^k = \lambda^{-k} x^k$, $dt = \lambda^{-1} dx$)

$$= \lambda^{-k} \int_0^\infty x^k e^{-x} dx$$

(see Proposition 2.71(iv))

$$= \lambda^{-k} k!.$$

In particular

$$\boxed{\mathbb{E}[T] = \mu_1[T] = \frac{1}{\lambda}}, \ \ \mu_2[T] = \frac{2}{\lambda^2}, \ \ \boxed{\boldsymbol{var}[T] = \mu_2[T] - \mu_1[T]^2 = \frac{1}{\lambda^2}}.$$

If T happens to model the lifetime of a light bulb, then the expected lifetime is $\frac{1}{\lambda}$ so that λ is measured in [unit-of-time]$^{-1}$. We can then interpret λ as describing how many light bulbs we expect to replace in a given unit of time. This explains the terminology "rate" used when referring to λ.

The *half-life* of an exponential random variable $T \sim \mathrm{Exp}(\lambda)$ is its *median*, i.e., its 0.5-quantile. The half-life is then the smallest solution $h = h(\lambda)$ of the equation

$$\mathbb{P}(T \leq h) = \frac{1}{2}.$$

Equivalently, this means

$$e^{-\lambda h} = \mathbb{P}(T > h) = \frac{1}{2}$$

which shows that

$$\boxed{h(\lambda) = \frac{\ln 2}{\lambda}}. \tag{2.44}$$

When we say that a radioactive material has a life time h_0 we mean that the lifetime of this material until it has decayed completely is an exponential random variable $X \sim \mathrm{Exp}(\lambda)$ such that $h(\lambda) = h_0$. In Example 4.33 we give another statistical interpretation to the concept of half-life that will perhaps give a more convincing explanation for the name *half-life*. □

Remark 2.69. The geometric random variables are intimately related to the exponentially distributed ones. Fix a small positive number and assume that every δ seconds we perform a Bernoulli experiment with probability of success p_δ. Denote by T the time we have to wait until the first success. This is a geometrically distributed discrete r.v. and the probability that the waiting time T is longer than $t = n\delta$ is

$$G_\delta(t) = \mathbb{P}(T > t) = (1-p_\delta)^n = (1-p_\delta)^{\frac{t}{\delta}} = \Big(\big(1-p_\delta\big)^{-\frac{1}{p_\delta}} \Big)^{-\frac{p_\delta}{\delta} t}.$$

We assume that p_δ is proportional to the duration δ, $p_\delta = \lambda\delta$. We deduce

$$G_\delta(t) = \Big(\big(1-\lambda\delta\big)^{-\frac{1}{\lambda\delta}} \Big)^{-\lambda t}.$$

Since

$$\lim_{\delta \searrow 0} (1-\lambda\delta)^{-\frac{1}{\lambda\delta}} = e,$$

we deduce that

$$\lim_{\delta \searrow 0} G_\delta(t) = e^{-\lambda t}.$$

The quantity $G(t) = e^{-\lambda t}$ the survival function of an exponentially distributed random variable with parameter λ. □

To describe the next examples of continuous random variables we need to survey a few facts of classical analysis.

Definition 2.70 (Gamma and Beta functions)**.** The *Gamma function* is the function

$$\Gamma : (0,\infty) \to \mathbb{R}, \quad \boxed{\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt}. \tag{2.45}$$

The *Beta function* is the function of two positive variables

$$\boxed{B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}}, \quad x, y > 0.$$ □

In the sequel we will need to know a few basic facts about the Gamma and Beta functions. For proofs we refer to [12, Chap. 1].

Proposition 2.71. *The following hold.*

(i) $\Gamma(1) = 1$.

(ii) $\Gamma(x+1) = x\Gamma(x)$, $\forall x > 0$.

(iii) *For any* $n = 1, 2, \ldots$ *we have*

$$\Gamma(n) = (n-1)!. \tag{2.46}$$

(iv) $\boxed{\Gamma(1/2) = \sqrt{\pi}}$.

(v) *For any* $x, y > 0$ *we have* Euler's formula

$$\boxed{\int_0^1 s^{x-1}(1-s)^{y-1}ds = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} = B(x,y)}. \tag{2.47}$$

□

The equality (iv) above reads

$$\sqrt{\pi} = \Gamma(1/2) = \int_0^\infty e^{-t}t^{-1/2}dt$$

$(t = x^2,\ t^{-1/2} = x^{-1}\ dt = 2xdx)$

$$= 2\int_0^\infty e^{-x^2}dx = \int_{-\infty}^0 e^{-x^2}dx + \int_0^\infty e^{-x^2}dx = \int_{-\infty}^\infty e^{-x^2}dx.$$

If we make the change in variables $x = \frac{s}{\sqrt{2}}$ so that $x^2 = \frac{s^2}{2}$ and $dx = \frac{1}{\sqrt{2}}ds$, then we deduce

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

From this we obtain the fundamental equality

$$\boxed{\frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty e^{-\frac{x^2}{2}}dx = 1}. \tag{2.48}$$

Example 2.72 (The normal distributions)**.** A continuous random variable X is said to be *normally distributed* or *Gaussian with parameters* (μ, σ^2) if its pdf is

$$\boxed{\gamma_{\mu,\sigma}(y) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(y-\mu)^2}{2\sigma^2}}}.$$

The equality (2.48) shows that

$$\int_{-\infty}^\infty \gamma_{\mu,\sigma}(x)dx = 1,$$

so that $\gamma_{\mu,\sigma}$ is indeed a probability density. **This distribution plays a fundamental role in probability** and we will have more to say about it.

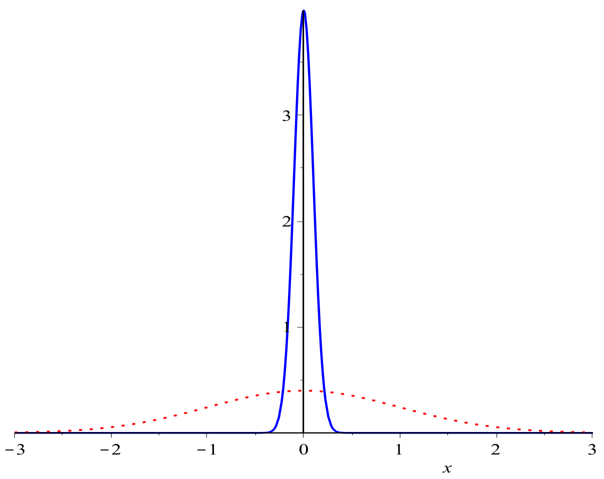

Figure 2.11. *The graph of* $\gamma_{0,\sigma}$ *for* $\sigma = 1$ *(dotted red curve) and* $\sigma = 0.1$ *(continuous blue curve).*

In Figure 2.11 we have depicted the graph of $\gamma_{0,\sigma}$ for $\sigma = 1$ and $\sigma = 0.1$. Note that the smaller σ corresponds to the sharper peak. The graphs of $\gamma_{0,\sigma}$ are called *Gauss bells.*

We will use the notation

$$X \sim N(\mu, \sigma^2)$$

to indicate that X is a continuous random variable with the above probability density. We will see later in Example 2.78 that μ is the mean of X, σ is its standard deviation and σ^2 its variance, i.e.,

$$\boxed{X \sim N(\mu, \sigma^2) \Rightarrow \mathbb{E}[X] = \mu, \;\; \boldsymbol{var}[X] = \sigma^2}.$$

Here we verify this assertion in the special case $\mu = 0$ and $\sigma = 1$. In this case we say that that X *standard normal* or *standard Gaussian*, and we indicate this $X \sim N(0, 1)$.

If $X \sim N(0, 1)$, then the pdf of X is

$$\gamma(x) = \gamma_{0,1} = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

This shows that X is s-integrable for any $s \geq 1$. Moreover, for any $k \in \{1, 2, \dots\}$, the k-th momentum of X is

$$\mu_k[X] = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} x^k e^{-\frac{x^2}{2}} dx.$$

Observe that when k is odd, the integral above is equal to zero because the function $f(x) = x^k e^{-\frac{x^2}{2}}$ is odd, $f(-x) = -f(x)$. In particular,

$$\mathbb{E}[X] = \mu_1[X] = 0.$$

For k even, $k = 2n$ we have

$$\mu_{2n}[X] = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} x^{2n} e^{-\frac{x^2}{2}} = \frac{2}{\sqrt{2\pi}} \int_0^\infty x^{2n} e^{-\frac{x^2}{2}} dx$$

$(r = x^2/2,\ x = \sqrt{2r},\ dx = (2r)^{-1/2},\ x^{2n} = (2r)^n)$

$$= \frac{2^{n+1/2}}{\sqrt{2\pi}} \int_0^\infty r^{n-1/2} e^{-r} dr = \frac{2^n}{\sqrt{\pi}} \Gamma(n+1/2).$$

In particular

$$\mu_2[X] = \frac{2}{\sqrt{\pi}} \Gamma(3/2).$$

Since

$$\Gamma(3/2) = \frac{1}{2}\Gamma(1/2) = \frac{\sqrt{\pi}}{2}$$

we deduce

$$\boxed{X \sim N(0,1) \Rightarrow \mathbb{E}[X] = 0, \ \ \boldsymbol{var}[X] = 1}. \tag{2.49}$$

The cumulative distribution function of a standard normal variable is

$$\boxed{\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt}, \tag{2.50}$$

and it is typically expressed in terms of the *error function*

$$\boxed{\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt = \frac{1}{\sqrt{\pi}} \int_{-x}^x e^{-t^2} dt}. \tag{2.51}$$

More precisely

$$\boxed{\Phi(x) = \frac{1}{2} + \frac{1}{2} \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)}.$$

While there is no closed formula for computing $\Phi(x)$, the value of $\Phi(x)$ can be very well approximated for any x. These computations are included in long tables frequently used by statisticians.[6] We can use R to compute $\Phi(x)$. For example, $\Phi(1.58)$ is computed using the R command

```
pnorm(1.58)
```

Example 2.73 (Gamma distributions)**.** The *Gamma distributions* with parameters ν, λ are defined by the probability densities $g_\nu(x;\lambda)$, $\lambda, \nu > 0$ given by

$$g_\nu(x;\lambda) = \begin{cases} \frac{\lambda^\nu}{\Gamma(\nu)} x^{\nu-1} e^{-\lambda x}, & x > 0, \\ 0, & x \leq 0. \end{cases} \tag{2.52}$$

[6] We recommend the Math is Fun site on the standard normal distribution. There you can interactively find the values of $\Phi(x)$ for many x-s.
https://www.mathsisfun.com/data/standard-normal-distribution-table.html

From the definition of the Gamma function we deduce that $g_\nu(x);\lambda$ is indeed a probability density

$$\int_0^\infty g_\nu(x;\lambda)dx = 1.$$

Note that $g_1(x;\lambda)$ is the exponential distribution with parameter λ. We will use the notation $X \sim \text{Gamma}(\nu,\lambda)$ to indicate that the probability density function of X is a Gamma distribution with parameters ν,λ. The parameter ν is sometimes referred to as the *shape* parameter. Figure 2.12 may explain the reason for this terminology.

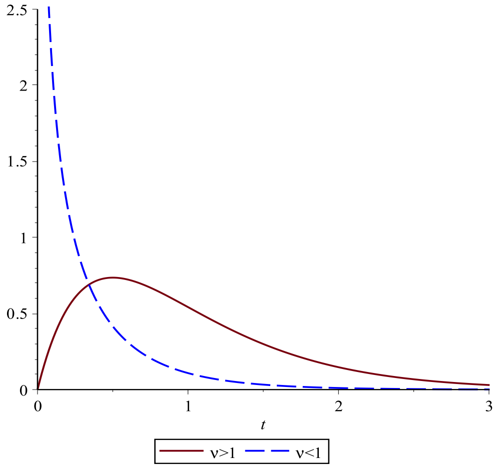

Figure 2.12. *The graphs of $g_\nu(x;\lambda)$ for $\nu > 1$ and $\nu < 1$.*

For $n = 1, 2, 3, \dots$ the distribution $\text{Gamma}(n,\lambda)$ has a simple probabilistic interpretation. If the waiting time T for a certain event is exponentially distributed with rate λ, e.g., the waiting time for a bus to arrive, then the waiting time for n of these events to occur independently and in succession is a $\text{Gamma}(n,\lambda)$ random variable. We will prove this later in Proposition 5.14.

The distribution $g_{n/2}(x;1/2)$, where $n = 1, 2, \dots,$ plays an important role in statistics it also known as the *chi-squared distribution with n degrees of freedom*. One can show that if $X_1, \dots, X_n$ are independent standard normal random variables, the the random variable

$$X_1^2 + \cdots + X_n^2$$

has a chi-squared distribution of degree n.

If $X \sim \mathrm{Gamma}(\nu, \lambda)$ is a Gamma distributed random variable, then X is s-integrable for any $s \geq 1$. Moreover, for any $k \in \{1, 2, \dots\}$ we have

$$\mu_k[X] = \frac{\lambda^\nu}{\Gamma(\nu)} \int_0^\infty x^{k+\nu-1} e^{-\lambda x} dx$$

$(x = \lambda^{-1}t,\ dx = \lambda^{-1}dt,\ \lambda x = t,\ x^{k+\nu-1} = \lambda^{-(k+\nu-1)}t^{k+\nu-1})$

$$= \frac{1}{\lambda^k \Gamma(\nu)} \int_0^\infty t^{k+\nu-1} e^{-t} dt = \frac{\Gamma(k+\nu)}{\lambda^k \Gamma(\nu)}.$$

Hence

$$\boxed{\mu_k[X] = \frac{\Gamma(k+\nu)}{\lambda^k \Gamma(\nu)},\ \ X \sim \mathrm{Gamma}(\nu, \lambda)}. \tag{2.53}$$

We deduce

$$\mathbb{E}[X] = \mu_1[X] = \frac{\Gamma(\nu+1)}{\lambda \Gamma(\nu)} = \frac{\nu}{\lambda},$$

$$\boldsymbol{var}[X] = \mu_2[X] - \mu_1[X]^2 = \frac{\Gamma(\nu+2)}{\lambda^2 \Gamma(\nu)} - \frac{\nu^2}{\lambda^2}$$

$$= \frac{k(k+1) - k^2}{\lambda^2} = \frac{\nu}{\lambda^2}.$$

Hence

$$\boxed{X \sim \mathrm{Gamma}(\nu, \lambda) \Rightarrow \mathbb{E}[X] = \frac{\nu}{\lambda},\ \ \boldsymbol{var}[X] = \frac{\nu}{\lambda^2}}.$$

Example 2.74 (Beta distributions)**.** The *Beta distribution* with parameters $a, b > 0$ is defined by the probability density function

$$\beta_{a,b}(x) = \frac{1}{B(a,b)} \times \begin{cases} x^{a-1}(1-x)^{b-1}, & x \in (0,1), \\ 0, & \text{otherwise.} \end{cases}$$

We will use the notation $X \sim \mathrm{Beta}(a,b)$ to indicate that the pdf of X is a Beta distribution with parameters a, b.

Suppose that $X \sim \mathrm{Beta}(a,b)$. Then

$$\mathbb{E}[X] = \frac{1}{B(a,b)} \int_0^1 x^a (1-x)^{b-1} dx = \frac{B(a+1,b)}{B(a,b)}$$

$$= \frac{\Gamma(a+1)\Gamma(a+b)}{\Gamma(a)\Gamma(a+b+1)} = \frac{a}{a+b},$$

$$\mathbb{E}[X^2] = \frac{1}{B(a,b)} \int_0^1 x^{a+1}(1-x)^{b-1} dx = \frac{\Gamma(a+2)\Gamma(a+b)}{\Gamma(a)\Gamma(a+b+2)}$$

$$= \frac{a(a+1)}{(a+b)(a+b+1)}.$$

Hence

$$\boldsymbol{var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{a}{a+b}\left(\frac{a+1}{a+b+1} - \frac{a}{a+b}\right)$$

$$= \frac{a}{a+b} \cdot \frac{(a+1)(a+b) - a(a+b+1)}{(a+b)(a+b+1)} = \frac{ab}{(a+b)^2(a+b+1)}.$$

We summarize the above results

$$\boxed{\mathbb{E}[X] = \frac{a}{a+b},\ \ \boldsymbol{var}[X] = \frac{ab}{(a+b)^2(a+b+1)},\ \ X \sim \mathrm{Beta}(a,b).} \tag{2.54}$$

The cdf of $X \sim \mathrm{Beta}(a,b)$ is

$$\mathbb{P}(X \leq x) = B_{a,b}(x) := \frac{1}{B(a,b)} \int_0^x t^{a-1}(1-t)^{b-1}dt, \;\; x \in [0,1]. \tag{2.55}$$

The function $B_{a,b}(x)$ is called the *incomplete Beta function*[7] with parameters $a,b > 0$. Note that

$$\begin{gathered} \frac{d}{dt}\Big(t^a(1-t)^b \Big) = at^{a-1}(1-t)^b - bt^a(1-t)^{b-1} \\ = at^{a-1}(1-t)^{b-1} - at^a(1-t)^{b-1} - bt^a(1-t)^{b-1} \\ = at^{a-1}(1-t)^{b-1} - (a+b)t^a(1-t)^{b-1}. \end{gathered}$$

Integrating from 0 to x we deduce

$$x^a(1-x)^b = a\int_0^x t^{a-1}(1-t)^{b-1}dt - (a+b)\int_0^x t^a(1-t)^{b-1}dt$$

so

$$\frac{x^a(1-x)^b}{aB(a,b)} = B_{a,b}(x) - B_{a+1,b}(x). \tag{2.56}$$

Arguing in a similar fashion we deduce

$$\frac{x^a(1-x)^b}{bB(a,b)} = B_{a,b+1}(x) - B_{a,b}(x). \tag{2.57}$$

If we add the above two equalities we deduce

$$B_{a,b+1}(x) - B_{a+1,b}(x) = \frac{a+b}{abB(a,b)}x^a(1-x)^b. \tag{2.58}$$

Suppose now that a,b are natural numbers. Then

$$\frac{a+b}{abB(a,b)} = \frac{(a+b)\Gamma(a+b)}{a\Gamma(a)b\Gamma(b)} = \frac{(a+b)!}{a!b!} = \binom{a+b}{a}.$$

Now fix a natural number a. For any $0 < a < n$ we have

$$B_{a,n-a+1}(x) - B_{a+1,n-a}(x) = \binom{n}{a}x^a(1-x)^{n-a}.$$

Hence

$$B_{k,n-k+1}(x) - B_{n,1}(x) = \sum_{a=k}^{n-1}\binom{n}{a}x^a(1-x)^{n-a}$$

Since

$$B_{n,1}(x) = x^n$$

we deduce

$$\boxed{B_{k,n+1-k}(x) = \sum_{a=k}^{n}\binom{n}{a}x^a(1-x)^{n-a}}. \tag{2.59}$$

Thus, if $X \sim \mathrm{Bin}(n,p)$, then

$$\mathbb{P}(X \geq k) = B_{k,n+1-k}(p). \tag{2.60}$$

□

Example 2.75 (The Cauchy distribution)**.** The *Cauchy distribution* is defined by the probability density

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

In Figure 2.13 we have depicted the Cauchy density against the normal density. The Cauchy distribution

[7] Other authors use the notation $I_{a,b}(x)$ for the incomplete Beta function. Our notation follows A. Rényi's [16].

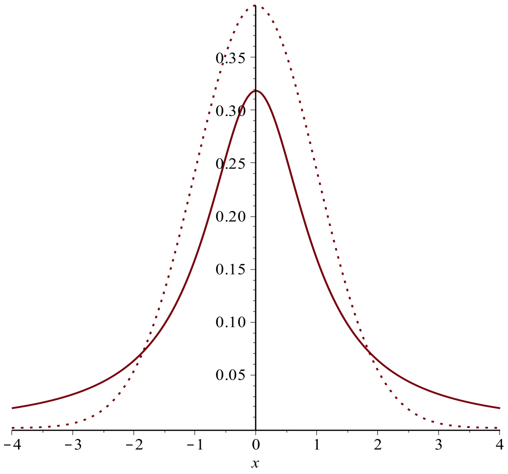

Figure 2.13. *The Cauchy distribution (continuous line) vs. normal distribution (dotted line).*

is not integrable since the improper integral

$$\int_{\mathbb{R}} \frac{|x|}{\pi(1+x^2)} dx$$

is divergent. □

2.3.3. Functions of continuous random variables. If X is a continuous random variable and $g : \mathbb{R} \to \mathbb{R}$ is a function, then $g(X)$ is another random variable. Under certain assumptions on g, the random variable $g(X)$ is also continuous.

Example 2.76. Suppose that $g(x) = -x$, and X is a continuous random variable with probability density $p(x)$. Then $Y = g(X) = -X$ is also a continuous random variable with density $q(y) = p(-y)$. To see this we compute the cumulative probability function of Y

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(-X \leq y) = \mathbb{P}(X \geq -y) = \int_{-y}^{\infty} p(x)dx.$$

Derivating with respect to y we get

$$F'_{-X}(y) = q(y) = p(-y).$$ □

Example 2.77. Suppose that $g(x) = ax+b$, $a > 0$ and X is a continuous random variable with probability density $p(x)$. Then $Y = g(X)$ is also a continuous random variable. To find its probability density q we compute its cumulative distribution function

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(aX + b \leq y) = \mathbb{P}(aX \leq y - b)$$

$$\overset{a\geq 0}{=} \mathbb{P}\left(X \leq \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right).$$

Derivating with respect to y we get

$$\boxed{q(y) = F_Y'(y) = \frac{d}{dy}F_X\left(\frac{y-b}{a}\right) = \frac{1}{a}p\left(\frac{y-b}{a}\right)}.$$

Observe that the function $h(y) = \frac{y-b}{a}$ is the inverse function $g^{-1}(y)$ and $a = g'(x)$.

□

Example 2.78 (Gaussian random variables)**.** Suppose that X is a standard normal random variable, $X \sim N(0,1)$. Then for any $\sigma > 0$ and $\mu \in \mathbb{R}$ the random variable $Y = \sigma X + \mu$ is Gaussian with parameter (μ, σ^2), $Y \sim N(\mu, \sigma^2)$.

Indeed, the probability density of X is

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

Using Example 2.77 with $g(x) = \sigma x + \mu$ we deduce that the pdf of Y is

$$\boxed{\boldsymbol{\gamma}_{\mu,\sigma}(y) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(y-\mu)^2}{2\sigma^2}}},$$

i.e., $Y \sim N(\mu, \sigma^2)$. Invoking (2.38) and (2.40b) we conclude that

$$\mathbb{E}[Y] = \sigma\mathbb{E}[X] + \mu = \mu, \quad \boldsymbol{var}[Y] = \sigma^2\,\boldsymbol{var}[X] = \sigma^2.$$

Conversely, it follows from Example 2.77 with $g(y) = (y-\mu)/\sigma$ that if $Y \sim N(\mu, \sigma^2)$, then

$$X = \frac{1}{\sigma}(Y - \mu) \sim N(0,1).$$

We deduce that, for any real number u we have

$$\mathbb{P}\big(\,Y \leq \mu + u\sigma\,\big) = \mathbb{P}\left(\frac{1}{\sigma}(Y-\mu) \leq \frac{\mu + u\sigma - \mu}{\sigma}\right) = \mathbb{P}(X \leq u) = \Phi(u), \quad (2.61)$$

where we recall that $\Phi(x)$ denotes cdf of the standard normal r.v. described explicitly in (2.50). This shows that in practice, the standard deviation σ is the most convenient measuring stick. For example, we deduce from (2.61) that, for any real numbers $u < v$ we have

$$\boxed{Y \sim N(\mu, \sigma^2) \Rightarrow \mathbb{P}\big(\,\mu + u\sigma \leq Y \leq \mu + v\sigma\,\big) = \Phi(v) - \Phi(u)}. \quad (2.62)$$

Before the computers became ubiquitous, the values of Φ were stored in large statistical tables. In R, the values of Φ are accessible using the command

```
pnorm()
```

Thus, $\Phi(0.112)$ is accessed using the command

```
pnorm(0,112)
```

that yields $\Phi(0.112) \approx 0.5445883$. □

Example 2.79. Suppose that $X \sim N(0,\sigma)$ is a normal random variable with mean 0 and standard deviation σ. The pdf of X is the function

$$\gamma_{0,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{s^2}{2\sigma^2}}$$

depicted in Figure 2.11. In this case we have

$$\mathbb{P}(|X| \geq c\sigma) \overset{(2.62)}{=} 2\big(1 - \Phi(c)\big), \;\; \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{s^2}{2}} ds.$$

$$\begin{aligned} \mathbb{P}(|X| \geq \sigma) = 2\big(\,1 - \Phi(1)\,\big) &\approx 0.3173, \\ \mathbb{P}(|X| \geq 2\sigma) &\approx 0.0455, \\ \mathbb{P}(|X| \geq 3\sigma) &\approx 0.0026. \end{aligned}$$

Thus, the probability of X being at least 3 standard deviations away from its mean 0 is about 2 in 1000. Compare this with Chebyshev's inequality prediction that these odds are at most 1 in 9. Let us observe that

$$\Phi_\sigma(\sigma) \approx 0.84, \;\; \Phi_\sigma(2\sigma) \approx 0.97, \;\; \Phi_\sigma(3\sigma) \approx 0.99.$$

We can interpret this by saying that 1 standard deviation is the 84-th percentile of X, 2 standard deviations is the 97-th percentile and 3 standard deviations is the 99-th percentile.

In Figure 2.11 we have depicted the graphs of the Gaussian distributions $\gamma_{0,\sigma}$ for two values of σ, $\sigma = 1$ and $\sigma = 0.1$. The sharper-peak-graph correspond to the smaller standard deviation $\sigma = 0.1$.

The sharp peak in Figure 2.11 is a manifestation of the phenomenon described in Remark 2.54: a small standard deviation suggests that the large deviations from the mean are highly unlikely.

□

Example 2.80. Suppose that X is a continuous random variable with probability density $p(x)$ and $n \in \{1,2,\dots\}$. If $g(x) = x^n$, then $Y = g(X)$ is also a continuous random variable. For simplicity we discuss only the case $n = 2$, i.e., $g(x) = x^2$.

Note that $\mathbb{P}(Y < 0) = 0$. To find its probability density q we compute its cumulative distribution function for $y > 0$

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

Derivating with respect to $y > 0$ we get

$$q(y) = F_Y'(y) = \frac{d}{dy}\Big(F_X(\sqrt{y}) - F_X(-\sqrt{y}) \Big) = \frac{1}{2\sqrt{y}}\big(p(\sqrt{y}) + p(-\sqrt{y}) \big), \;\; y > 0.$$ □

The next result is a generalization of both Example 2.76 and 2.77.

Theorem 2.81 (Method of transformation)**.** *Suppose that X is a continuous random variable with probability density $p(x)$, and I is an interval containing the range of X. If $g : I \to \mathbb{R}$ is a differentiable function such that $g'(x) \neq 0$, $\forall x \in I$, then $Y = g(X)$ is also a continuous random variable with probability density*

$$q(y) = |h'(y)|p\big(\,h(y)\,\big),$$

where h is the inverse function of g.

Proof. Since $g'(x) \neq 0$, $\forall x \in I$ we deduce from the intermediate value property for derivatives that g' does not change sign on I. Thus either g' is everywhere positive, or everywhere negative. We assume that g' is everywhere positive so g is increasing. (The case of decreasing g can be dealt with in a similar fashion.) We compute the cumulative distribution function of Y,

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y).$$

Since g is increasing we deduce that $g(X) \leq y$ if and only if $X \leq g^{-1}(y) = h(y)$. Thus

$$F_Y(y) = \mathbb{P}\big(\,X \leq h(y)\,\big) = F_X\big(\,h(y)\,\big).$$

The result is obtained derivating with respact to y the above equality. □

Example 2.82 (Simulating exponential random variables)**.** Modern computers have ways of simulating uniformly distributed random variables by using so called *random number generators*. Using this random number generator one can then simulate many other random variables. We explain the general principle in the special case of the exponential variable with parameter $\lambda > 0$.

Fix a uniformly distributed random variable $X \sim \mathrm{Unif}(0,1)$. Thus the probability density of X is

$$p(x) = \begin{cases} 1, & x \in [0,1], \\ 0, & x \notin [0,1]. \end{cases}$$

We seek a function g such that $Y = g(X)$ is exponentially distributed $Y \sim \mathrm{Exp}(\lambda)$. The function g has to be an increasing function $g : (0,1) \to (0,\infty)$ such that

(i) $\lim_{x\to 0} g(x) = 0$, $\lim_{x\to 1} g(x) = \infty$.

(ii) If $h(y)$ is the inverse of g, then

$$h'(y) = h'(y)p(h(y)) = \lambda e^{-\lambda y} = F_Y'(y), \;\; \forall y \in (0,\infty).$$

We deduce from (ii) that $h(y) = F_Y(y) + c$ for some constant c. Condition (i) implies $h(0) = 0$ so $c = -F_Y(0)$. On the other hand, since $Y \sim \mathrm{Exp}(\lambda)$ we have

$$F_Y(y) = \int_0^y \lambda e^{-\lambda t} dt = 1 - e^{-\lambda y}$$

so $F_Y(0) = 0$

$$h(y) = F_Y(y) = 1 - e^{-\lambda y}.$$

The function $g(x)$ is the inverse of $h(y)$ and can be found by solving for y the equation

$$x = h(y) = 1 - e^{-\lambda y}.$$

We deduce

$$e^{-\lambda y} = 1 - x \Rightarrow g(x) = y = -\frac{1}{\lambda}\ln(1-x).$$

Thus, if the random quantity X is uniformly distributed on $[0,1]$, then the random quantity $Y = -\frac{1}{\lambda}\ln(1-X)$ is exponentially distributed with parameter λ. □

Remark 2.83 (Quantiles/Percentiles)**.** Suppose that X is a random variable (discrete or continuous) with cumulative distribution function F_X. Recall that for any number $p \in [0,1]$ we defined the *p-quantile* of X denoted by $Q_X(p)$, to be smallest number x_0 such that

$$\mathbb{P}(X \leq x_0) \geq p.$$

Thus x_0 is the p-quantile of X if

$$\mathbb{P}(X \leq x_0) \geq p,\;\; \mathbb{P}(X \leq x) < p,\;\; \forall x < x_0.$$

If X is a continuous random variable with probability density $p(x)$, then the p-quantile is the smallest number x_0 such that

$$p = \int_{-\infty}^{x_0} p(s)ds.$$

If moreover, F_X is an invertible function, and then $Q_X(p) = F_X^{-1}(p)$.

One should think of the p-quantile as a function $Q_X : [0,1] \to \mathbb{R}$. One can show that if $U \sim \text{Unif}(0,1)$, then the random variable $Q_X(U)$ has the same cumulative distribution function as X. We have seen this principle at work in Example 2.82.

□

2.4. Exercises

Exercise 2.1. The discrete random variable X has cdf F that is such that

$$F(x) = \begin{cases} 0, & x < 1,\\ F(x) = \frac{1}{3}, & 1 \leq x < 3,\\ F(x) = 1, & x \geq 3.\end{cases}$$

Find (a) $F(2)$, (b) $\mathbb{P}(X > 1)$, (c) $\mathbb{P}(X = 2)$, and (d) $\mathbb{P}(X = 3)$.

Exercise 2.2. Roll two dice and find the pmf of X if X is (a) the smallest number and (b) the difference between the largest and the smallest numbers.

Exercise 2.3. The random variable X has pmf $p(k) = c/2^k$, $k = 0, 1, \dots$ Find(a) the constant c, (b) $\mathbb{P}(X > 0)$, and (c) the probability that X is even.

Exercise 2.4. Five cards are drawn at random from a deck of cards. Let X be the number of aces. Find the pmf of X if the cards are drawn (a) with replacement and (b) without replacement.

Exercise 2.5. A fair coin is flipped twice. Let X be the number of heads minus the number of tails. Find the pmf and sketch the cdf of X.

Exercise 2.6. The game of chuck-a-luck is played with three dice, rolled independently. You bet one dollar on one of the numbers 1 through 6 and, if exactly k of the dice show your number, you win k dollars $k = 1, 2, 3$ (and keep your wagered dollar). If no die shows your number, you lose your wagered dollar. What is your expected loss?

Exercise 2.7. The demand for a certain weekly magazine at a newsstand is a random variable with pmf $p(i) = (10 - i)/18$, $i = 4, 5, 6, 7$. If the magazine sells for $\$a$ and costs $\$2a/3$ to the owner, and the unsold magazines cannot be returned, how many magazines should be ordered every week to maximize profit in the long run?

Exercise 2.8. Find the expected number of different birthdays amongst four people selected at random.

Exercise 2.9. In a game, Ann gives Bob three fair quarters to flip. Bob will keep those which land heads and return those landing tails. However, if all three quarters land tails then Bob must pay Ann \$2. Find the expected value and variance of Bob's net gain.

Exercise 2.10. An urn contains 10 balls labelled 1 through 10. We draw without replacement 3 balls and we denote by X the smallest label among the three extracted balls. Find the pmf and the mean of X.

Exercise 2.11. A drunken man has 5 keys, one of which opens the door to his office. He tries the keys at random, one by one and independently. Compute the expectation and variance of the number of tries required to open the door if the wrong keys (a) are not eliminated; (b) are eliminated.

Exercise 2.12. An object is hidden randomly in one of ten covered boxes numbered from 1 to 10. You search for it by randomly lifting the lids. Find the expected number of lids you need to lift until you locate the object.

Exercise 2.13. A belt conveys tomatoes to be packed. Each tomato is defective with probability p, independently of the others. Each is inspected with probability r; inspections are also mutually independent. If a tomato is defective and inspected, it is rejected.

(i) Find the probability that the n-th tomato is the k-th defective tomato.

(ii) Find the probability that the n-th tomato is the k-th rejected tomato.

(iii) Given that the $(n+1)$-th tomato is the first to be rejected, let X be the number of its predecessors that were defective. Find $\mathbb{P}(X = k)$, the probability that X takes the value k, and $\mathbb{E}[X]$.

Exercise 2.14. The random variable X has a binomial distribution with $\mathbb{E}[X] = 1$ and $\boldsymbol{var}[X] = 0.9$. Compute $\mathbb{P}(X > 0)$.

Exercise 2.15. Roll a die 10 times. What is the probability of getting (a) no 6s, (b) at least two 6s, and (c) at most three 6s.

Exercise 2.16. Let X be the number of 6s when a die is rolled six times, and let Y be the number of 6s when a die is rolled 12 times. Find (a) $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ and (b) $\mathbb{P}(X > \mathbb{E}[X])$ and $\mathbb{P}(Y > \mathbb{E}[Y])$.

Exercise 2.17. A fair coin is flipped n times. What is the probability of getting a total of k heads if (a) the first flip shows heads, (b) the first flip shows tails, and (c) at least one flip shows heads?

Exercise 2.18. A fair coin is flipped 10 times. Each time it shows heads, Ann gets a point; otherwise Bob gets a point.

(i) What is the most likely final result?

(ii) Which is more likely: that it ends $5-5$ or that somebody wins $6-4$?

(iii) If Ann wins the first three rounds, what is the probability that she ends up the winner?

(iv) If Ann wins the first four rounds, what is the probability that Bob never takes the lead?

(v) What is the probability that the lead changes four times?

Exercise 2.19. A multiple-choice test consists of six questions, each with four alternatives. At least four correct answers are required for a passing grade. What is the probability that you pass if you (a) guess at random; (b) know the first three answers, and guess on the rest; (c) for each question know the correct answer with probability $\frac{1}{2}$, otherwise guess at random? (d) In (c) how high must should the probability that you know an answer be to ensure that, with at least 95% certainty you will pass? (e) For (a)-(c), find the mean and variance of the number of correct answers.

Exercise 2.20. A restaurant has 15 tables, and it is known that 70% of guests who make reservations actually show up. To compensate for this, the restaurant has a policy of taking more than 15 reservations, thus running a risk to become overbooked. How many reservations can they take to limit this risk to at most 5%?

Exercise 2.21. On average, how many games of bridge are necessary before a player is dealt three aces? (A bridge hand is 13 randomly selected cards from a standard deck.)

Exercise 2.22. A fair coin is flipped repeatedly. What is the probability that the fifth tail occurs before the tenth head?

Exercise 2.23. Ann rolls a fair die until she gets a 6. Bob then rolls the same die until he rolls an even number. Find the probability that Ann rolls the die more times than Bob.

Exercise 2.24. From a panel of prospective jurors, 12 are selected at random. If there are 200 men and 160 women on the panel, what is the probability that more than half of the jury selected are women?

Hint. Think hypergeometric distribution.

Exercise 2.25. Suppose $X \sim \mathrm{Geom}(p)$. Find the probability that X is even.

Exercise 2.26. The number of customers X who call a certain toll-free number in a minute has a Poisson distribution with mean $\lambda = 2$. A minute is classified as "idle" if there are no calls and"busy" otherwise. (a) What is the probability that a given minute is busy? (b) Let Y be the number of calls during a busy minute. Find the pmf of Y and $\mathbb{E}[Y]$. (c) If a minute is idle, what is the expected number of busy minutes before the next idle minute? What assumptions are you making?

Exercise 2.27. Insects of a certain type lay eggs on leaves such that the number of eggs on a given leaf has a Poisson distribution with mean $\lambda = 1$. For any given leaf, the probability is 0.1 that it will be visited by such an insect, and leaves are visited independent of each other. (a) What is the probability that a given leaf has no eggs? (b) If a leaf is inspected and has no eggs, what is the probability that it has been visited by an insect? (c) If 10 leaves are inspected and none have any eggs, what is the probability that at least one leaf has been visited by an insect?

Exercise 2.28. They say that many are called and few are chosen. Suppose that the number of people called is $\mathrm{Poi}(\lambda)$. Each person called is chosen independently by flipping a fair coin: Heads you're chosen, Tail you're not. Show that the number of chosen people is $\mathrm{Poi}(\lambda/2)$.

Exercise 2.29. Let X be the Poisson random variable with parameter λ. Show that the the maximum of $\mathbb{P}(X = i)$ occurs at $\lfloor \lambda \rfloor$, the greatest integer less than or equal to λ.

Hint. Show that $p(i) = \frac{\lambda}{i} p(i-1)$ and use this to work out when $p(i)$ is increasing or decreasing.

Exercise 2.30. A prize is randomly placed in one of ten boxes, numbered from 1 to 10. You search for the prize by asking yes-no questions. Find the expected number of questions until you are sure about the location of the prize, under each of the following strategies.

(a) An enumeration strategy: you ask questions of the form“is it in box k?”.

(b) A bisection strategy: you eliminate as close to half of the remaining boxes as possible by asking questions of the form “is it in a box numbered less than or equal to k”?

Exercise 2.31. The probability density function of a random variable X is given by

$$f(x) = \begin{cases} \frac{c}{\sqrt{1-x^2}}, & -1 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

(i) Find the value of c.

(ii) Find the cumulative distribution function of X.

Hint. (i) Use Proposition 2.57.

Exercise 2.32. Let X be a random variable with probability density function

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

Find the expectation and variance of X.

Exercise 2.33. You bid on an object at a silent auction. You know that you can sell it later for 100 and you estimate that the maximum bid from others is uniform on $[70, 130]$ (for convenience, you assume that it is continuous, thus disregarding the possibility of two equal bids). How much should you bid to maximize your expected profit, and what is the maximum expected profit?

Exercise 2.34. A stick measuring one yard in length is broken into two pieces at random. Compute the expected length of the longest piece.

Exercise 2.35. Jobs arrive at a computer such that the time T between two consecutive jobs has an exponential distribution with mean 10 seconds. Find

(i) $\boldsymbol{var}[T]$,

(ii) $\mathbb{P}(T \leq 5)$,

(iii) the probability that the next job arrives within 5 seconds given that the last job arrived 25 seconds ago,

(iv) $\mathbb{P}(T > \mathbb{E}[T])$.

Exercise 2.36. A large number of lightbulbs are turned on in a new office building. A year later, 80% of them still function, and 2 years later, 30% of the original light- bulbs still function. Does it seem likely that the lifetimes follow an exponential distribution?

Exercise 2.37. Let $X \sim \text{Exp}(\lambda)$ and let $Y = \lambda X$. Show that $Y \sim \text{Exp}(1)$.

Exercise 2.38. The element nobelium has a half-life of 58 min. Let X be the lifetime of an individual nobelium atom. Find

(i) $\mathbb{P}(X > 30)$,

(ii) $\mathbb{P}(X \leq 60 | X > 30)$,

(iii) $\mathbb{E}[X]$ and

(iv) $\boldsymbol{var}[X]$.

Hint. Use (2.44).

Exercise 2.39. Let $T \sim \text{Exp}(\lambda)$ and let $X = [T] + 1$ ("$[x]$" denoting the integer part of the real number x). Show that $X \sim \text{Geom}(1 - e^{-\lambda})$ (success probability $1 - e^{-\lambda}$). If T is the lifetime of a component, what could be the interpretation of X?

Exercise 2.40. Let X have a normal distribution with mean $\mu = 200$ and standard deviation $\sigma = 10$. Find

(i) $\mathbb{P}(X \leq 220)$,

(ii) $\mathbb{P}(X \leq 190)$,

(iii) $\mathbb{P}(X > 185)$,

(iv) $\mathbb{P}(X > 205)$,

(v) $\mathbb{P}(190 < X < 210)$,

(vi) $\mathbb{P}(180 \leq X \leq 210)$.

Hint. Use (2.62).

Exercise 2.41. Two species of fish have weights that follow normal distributions. Species A has mean 20 and standard deviation 2; species B has mean 40 and standard deviation 8. Which is more extreme: a 24 pound A-fish or a 48 pound B-fish?

Exercise 2.42. Let X be an exponentially distributed random variable,

$$X \sim \text{Exp}(3).$$

Find the probability density function of $Y = \ln X$.

Exercise 2.43. Suppose X is a normal random variable, $X \sim N(\mu, \sigma^2)$ Show that for any $c, b \in \mathbb{R}$, $c \neq 0$, the random variable $cX + b$ is normal,

$$cX + b \sim N(c\mu + b, c^2\sigma^2).$$

Exercise 2.44. Suppose that $X \sim N(0, 1)$. Show that $X^2 \sim \text{Gamma}(1/2, 1/2)$.

Exercise 2.45. Let $X \sim \text{Unif}(-1, 1)$. Find the probability density function of the random variable $Y = X^2$.

Exercise 2.46. Let $X \sim \operatorname{Exp}(1)$. Define

$$Y = \begin{cases} X, & X \leq 1 \\ 1/X & X > 1. \end{cases}$$

Find the probability density function of Y.

Exercise 2.47. Suppose that $X \sim N(0,1)$. Prove that for any $x > 0$ we have

$$\frac{x}{x^2+1}\phi(x) \leq \mathbb{P}(\, X > x \,) \leq \frac{1}{x}\phi(x) \tag{2.63}$$

where $\phi(x)$ is the pdf of X,

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

Exercise 2.48. For $n \geq 1$ let X_n be the continuous random variable with probability density function

$$f(x) = \begin{cases} \frac{c_n}{x^{n+1}}, & x \geq c_n \\ 0 & \text{otherwise.} \end{cases}$$

The X_n-s are the *Pareto random variables* and are used in the study of income distributions.

(i) Calculate c_n, $n \geq 1$.

(ii) Find $\mathbb{E}[X_n]$, $n \geq 1$.

(iii) Determine the density function of $Z_n = \ln X_n$, $n \geq 1$.

(iv) For what values of m does $\mathbb{E}\big[X_n^{m+1}\big]$ exist?

Hint. (i) Use Proposition 2.57.

Exercise 2.49. Prove the claims in Corollary 2.63.

Chapter 3

Multivariate discrete distributions

3.1. Discrete random vectors

Let X, Y be random variables defined on the same probability space $(S, \mathbb{P})$, $X, Y : S \to \mathbb{R}$. Often we are forced to treat the pair (X, Y) as an entity and thus we get a *random vector*

$$(X, Y) : S \to \mathbb{R}^2.$$

In this section we investigate the case when both X and Y are discrete random variables.

Suppose that the range of X is $\mathscr{X}$ and the range of Y is $\mathscr{Y}$. Then the range of (X, Y) is contained in the Cartesian product $\mathscr{X} \times \mathscr{Y}$ and the statistics of the random vector (X, Y) are determined by the *joint probability mass function* (joint pmf). This is the function

$$p : \mathscr{X} \times \mathscr{Y} \to [0, 1], \;\; p(x, y) = \mathbb{P}(X = x, Y = y).$$

Observe that

$$\sum_{(x,y) \in \mathscr{X} \times \mathscr{Y}} p(x, y) = 1. \tag{3.1}$$

The probability mass function of X (denoted p_X) and the probability mass function of Y (denoted by p_Y) are commonly referred to as the *marginal probability mass functions* or the *marginals* of the discrete random vector (X, Y).

Example 3.1. Suppose that X and Y are discrete random variables with ranges $\mathscr{X}$ and respectively $\mathscr{Y}$. Consider their pmf's

$$p_X : \mathscr{X} \to [0, 1], \;\; p_Y : \mathscr{Y} \to [0, 1].$$

The random variables X, Y are independent (Definition 2.7) if and only if

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

Thus, *the discrete random variables X, Y are independent if and only if, the joint pmf $p_{X,Y}$ of the random vector (X, Y) is the product of the pmf's of X and Y*

$$p_{X,Y}(x, y) = p_X(x)p_Y(y). \qquad \square$$

Example 3.2. Suppose that we draw 2 cards out of a regular deck of 52. We denote by X the number of Hearts drawn and by Y the number of Queens. In this case $\mathscr{X} = \mathscr{Y} = \{0, 1, 2\}$. Denote by $p(x, y) = \mathbb{P}(X = x, Y = y)$ the joint probability mass function of (X, Y).

There are 13 Hearts and 4 Queens, and exactly one of the Queens is the Queen of Hearts. This will count both as a Heart and as a Queen. All together, there are

- 16 cards that are Hearts or Queens,
- 36 cards that are neither Hearts, nor Queens,
- 12 Hearts that are not Queens, and
- 3 Queens that are not Hearts.

In particular, in the pair of drawn cards we cannot have two Queens and two Hearts so

$$p(2, 2) = 0.$$

We set

$$N = \binom{52}{2} = 26 \cdot 51 = 1326.$$

Then, using (F/P) we deduce

$$p(0, 0) = \frac{\binom{36}{2}}{N} = \frac{630}{N}.$$

$$p(0, 1) = \frac{36 \cdot 3}{N} = \frac{108}{N},\ \ p(0, 2) = \frac{\binom{3}{2}}{N} = \frac{3}{N},$$

$$p(1, 0) = \frac{12 \cdot 36}{N} = \frac{432}{N},\ \ p(2, 0) = \frac{\binom{12}{2}}{N} = \frac{66}{N},$$

$$p(2, 1) = \frac{12}{N},\ \ p(1, 2) = \frac{3}{N},$$

$$p(1, 1) = \frac{12 \cdot 3 + 1 \cdot 36}{N} = \frac{72}{N}.$$

A simple computation shows that

$$630 + 108 + 3 + 432 + 66 + 12 + 3 + 72 = 1326 = N$$

so the above probabilities add up to 1 as in (3.1) showing that we have exhausted all the possibilities.

y=Queens				
2	$\frac{3}{N}$	$\frac{3}{N}$	0	
1	$\frac{108}{N}$	$\frac{72}{N}$	$\frac{12}{N}$	
0	$\frac{630}{N}$	$\frac{432}{N}$	$\frac{66}{N}$	
$p(x,y)$	0	1	2	x=Hearts

Table 3.1. Describing a joint pmf by a rectangular array

It is convenient to organize the above results in the probability Table 3.2. Adding the elements on each column of this table we deduce

$$p(0,0)+p(0,1)+p(0,2)=\frac{630+108+3}{N}=\frac{741}{1326}.$$

On the other hand, $\mathbb{P}(X=0)$ is the probability that when we draw a pair of cards we get no Heart. Since there are 39 non-Heart cards we deduce

$$\mathbb{P}(X=0)=\frac{\binom{39}{2}}{N}=\frac{39\cdot 19}{1326}=\frac{741}{1326}.$$

Thus,

$$\mathbb{P}(X=0)=p(0,0)+p(0,1)+p(0,2).$$

We deduce in a similar fashion that

$$\mathbb{P}(X=1)=p(1,0)+p(1,1)+p(1,2),\;\;\mathbb{P}(X=2)=p(2,0)+p(2,1)+p(2,2),$$

$$\mathbb{P}(Y=0)=p(0,0)+p(1,0)+p(2,0),\;\;\mathbb{P}(Y=1)=p(0,1)+p(1,1)+p(2,1),$$

$$\mathbb{P}(Y=2)=p(0,2)+p(1,2)+p(2,2).$$

Hence, if we add the elements in the same column of the table we obtain the pmf of X, and if we add the elements in the same row of the table we obtain the pmf of Y. □

The last observations in Example 3.2 are manifestations of the following general fact.

Proposition 3.3 (Marginals)**.** *Suppose that X,Y are two discrete random variable, with ranges $\mathscr{X}$ and respectively $\mathscr{Y}$ and probability mass distributions p_X and respectively p_Y. If $p(x,y)$ is the joint pmf of the random vector (X,Y), then*

$$p_X(x)=\sum_{y\in\mathscr{Y}}p(x,y),\;\;p_Y(y)=\sum_{x\in\mathscr{X}}p(x,y). \tag{3.2}$$

□

The equality (3.2) is another way of writing the obvious equalities

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

The first sum corresponds to the summation over columns in Table 3.2, while the second sum corresponds to the summation over rows in Table 3.2.

Definition 3.4. (a) The range of a discrete random vector (X,Y) consists of the pairs of real numbers (x,y) such that

$$\mathbb{P}(X=x,Y=y)\neq 0.$$

In other words, the range $\mathcal{R}$ of (X,Y) is the set of all possible values of the pair of random variables (X,Y). The joint pmf of (X,Y) is thus uniquely determined by its restriction to the range.

(b) A discrete random vector (X,Y) with range $\mathcal{R}$ is said to be *uniformly distributed* if the restriction to $\mathcal{R}$ of the joint pmf is a *constant function.* □

Remark 3.5. The range $\mathcal{R}$ of a discrete random vector (X,Y) can be visualized as a collection of points in the plane. The joint pmf of (X,Y) assigns a positive number to each point in this collection. You can think of the number assigned to a point as the weight of that point. The sum of the weights of all the points in the range is equal to 1.

If the random vector (X,Y) is uniformly distributed, this means that each point in the range is equally likely to be sampled. In particular, we deduce that, in this case, the range $\mathcal{R}$ consists of a finite number n of points and the probability of each point in $\mathcal{R}$ is $\frac{1}{n}$.

The range of X is the projection of the collection $\mathcal{R}$ on the x-axis, and the range of Y is the projection of $\mathcal{R}$ on the y-axis. The pmf p_X of X is obtained as follows: $p_X(x_0)$ is the sum of the weights of the points in the collection $\mathcal{R}$ that project to the point x_0 on the x-axis.

The range of the random vector (X,Y) in Example 3.2 consists of 8 points depicted in red in Figure 3.1. Their coordinates are

$$(0,0),\ (0,1),\ (0,2),\ (1,0),\ (1,1),\ (1,2),\ (2,0),\ (2,1).$$

If X,Y are discrete random variables with ranges $\mathscr{X}$ and respectively $\mathscr{Y}$, then the range of of (X,Y) is contained in $\mathscr{X}\times\mathscr{Y}$. Example 3.2 (see Figure 3.1) shows that *the range of* (X,Y) *need not be equal to* $\mathscr{X}\times\mathscr{Y}$. □

Example 3.6. Suppose that a chicken lays a random number of eggs N, where $N\sim\operatorname{Poi}(\lambda)$. A given egg hatches with probability p. Denote by X the number of hatched eggs and by Y the number of eggs that did not hatch so that $X+Y=N$.

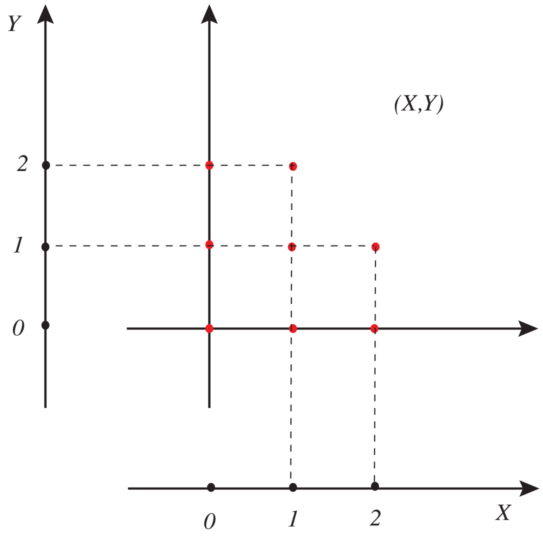

Figure 3.1. *The range of the discrete random vector in Example 3.2.*

We want to compute the joint pmf of (X, Y). The law of total probability implies

$$\begin{aligned}\mathbb{P}(X = i, Y = j) &= \sum_{n=0}^{\infty} \mathbb{P}(X = i, Y = j|N = n)\mathbb{P}(N = n)\\ &= \mathbb{P}(X = i, Y = j|N = i + j)\mathbb{P}(N = i + j).\end{aligned}$$

Observing that

$$\mathbb{P}(X = i, Y = j|N = i + j) = \mathbb{P}(X = i|N = i + j) = \binom{i+j}{i} p^i q^j, \quad q = 1 - p.$$

we deduce

$$\mathbb{P}(X = i, Y = j)\binom{i+j}{i} p^i q^j e^{-\lambda}\frac{\lambda^{j+j}}{(i + j)!} = e^{-\lambda}\frac{(\lambda p)^i(\lambda q)^j}{i!j!}$$

$(e^{-\lambda} = e^{-\lambda p}e^{-\lambda q})$

$$= e^{-\lambda p}\frac{(\lambda p)^i}{i!}e^{-\lambda q}\frac{(\lambda q)^j}{j!}.$$

We can now compute the pmf's of X and Y using the equality (3.2). We have

$$\begin{aligned}\mathbb{P}(X = i) = \sum_{j=0}^{\infty} e^{-\lambda p}\frac{(\lambda p)^i}{i!}e^{-\lambda q}\frac{(\lambda q)^j}{j!} &= e^{-\lambda p}\frac{(\lambda p)^i}{i!}\underbrace{\sum_{j=0}^{\infty} e^{-\lambda q}\frac{(\lambda q)^j}{j!}}_{=1}\\ &= e^{-\lambda p}\frac{(\lambda p)^i}{i!},\end{aligned}$$

$$\mathbb{P}(Y=j)=\sum_{i=0}^{\infty}e^{-\lambda p}\frac{(\lambda p)^i}{i!}e^{-\lambda q}\frac{(\lambda q)^j}{j!}=e^{-\lambda q}\frac{(\lambda q)^j}{j!}\underbrace{\sum_{i=0}^{\infty}e^{-\lambda p}\frac{(\lambda p)^i}{i!}}_{=1}$$

$$=e^{-\lambda q}\frac{(\lambda q)^j}{j!}.$$

In particular, we deduce $X\sim\text{Poi}(\lambda p)$, $Y\sim\text{Poi}(\lambda q)$ and $X\perp\!\!\!\perp Y$. □

Example 3.7. Phone calls arrive to a call center such that the number of phone calls in a minute has a Poisson distribution with mean 4, i.e., $\lambda=4$. Conditional on the total number of callers, a caller is female with probability 0.5. In a given minute, let X be the number of female callers, Y the total number of male callers callers and N the total number of callers

$$N=X+Y\sim\text{Poi}(4).$$

Set $\lambda=4$, $p=0.5$, $q=1-p=0.5$. Arguing exactly as in Example 3.6 above we deduce

$$\mathbb{P}(X=i,Y=j)=e^{-\lambda p}\frac{(\lambda p)^i}{i!}e^{-\lambda q}\frac{(\lambda q)^j}{j!}$$

and

$$X\sim\text{Poi}(p\lambda)=\text{Poi}(2),\;\;Y\sim\text{Poi}(2),\;\;X\perp\!\!\!\perp Y$$

□

Theorem 3.8 (The law of the subconscious statistician)**.** *Suppose that X,Y are two discrete random variables, with ranges $\mathscr{X}$ and respectively $\mathscr{Y}$. If $p(x,y)$ is the joint pmf of the random vector (X,Y) and $g(x,y)$ is a function of two variables defined on a region containing $\mathscr{X}\times\mathscr{Y}$, then the expectation of the random variable $g(X,Y)$ is*

$$\boxed{\mathbb{E}\big[\,g(X,Y)\,\big]=\sum_{x\in\mathscr{X},y\in\mathscr{Y}}g(x,y)p(x,y)}.$$

□

Corollary 3.9 (Linearity of expectation)**.** *(a) Suppose that X,Y are two discrete random variables. Then*

$$\mathbb{E}[X+Y]=\mathbb{E}[X]+\mathbb{E}[Y]. \tag{3.3}$$

(b) Suppose that $X_1,\dots,X_n$ are discrete random variables and $c_1,\dots,c_n$ are real constants. Then

$$\boxed{\mathbb{E}\big[\,c_1X_1+\cdots+c_nX_n\,\big]=c_1\mathbb{E}[X_1]+\cdots+c_n\mathbb{E}[X_n]}. \tag{3.4}$$

Proof. (a) Suppose that the ranges of X, Y are $\mathscr{X}$ and respectively $\mathscr{Y}$ and the joint pmf of (X, Y) is $p(x, y)$. Applying the law of subconscious statistician to the function $g(x, y) = x + y$ we deduce

$$\begin{aligned}\mathbb{E}[X+Y] &= \sum_{x\in\mathscr{X},\, y\in\mathscr{Y}} (x+y)p(x,y) = \sum_{x\in\mathscr{X},\, y\in\mathscr{Y}} xp(x,y) + \sum_{x\in\mathscr{X},\, y\in\mathscr{Y}} yp(x,y)\\ &= \sum_{x\in\mathscr{X}} x \underbrace{\left(\sum_{y\in\mathscr{Y}} p(x,y)\right)}_{p_X(x)} + \sum_{y\in\mathscr{Y}} y \underbrace{\left(\sum_{x\in\mathscr{X}} p(x,y)\right)}_{p_Y(y)}\\ &= \sum_{x\in\mathscr{X}} xp_X(x) + \sum_{y\in\mathscr{Y}} yp_Y(y) = \mathbb{E}[X] + \mathbb{E}[Y].\end{aligned}$$

(b) The equality (3.4) follows inductively from (3.3). We skip the proof. □

The remarkable feature of (3.4) is that *the random variables* $X_1, \dots, X_n$ *need not be independent*! Let us illustrate the versatility of (3.4) on some simple examples.

Example 3.10. Suppose that 20 of birds labelled $1, \dots, 20$ are sitting on a circle facing its center. At a given moment, each bird turns randomly and with equal probability to the left or right to see who is his neighbor. We denote by N the number of birds not seen by either of their neighbors. We want to compute the expectation of N. For $k = 1, \dots, 20$ we denote by A_k the event "*the k-bird is not seen by either of its neighbors*". We denote by I_{A_k} the indicator of A_k, i.e., the random variable that is equal to 1 if A_k has occurred and 0 otherwise; see Example 2.16. Note that

$$\mathbb{E}[I_{A_k}] = \mathbb{P}(A_k) = 0.5 \cdot 0.5 = 0.25,$$

and

$$N = I_{A_1} + \cdots + I_{A_{20}}.$$

Hence

$$\mathbb{E}[N] = \mathbb{E}[I_{A_1}] + \cdots + \mathbb{E}[I_{A_{20}}] = 20 \cdot \mathbb{P}(A_1) = 5. \qquad \square$$

Example 3.11 (The matching problem revisited)**.** Consider again the matching problem discussed first in Example 1.27. Given the n drunken sailors picking randomly their hats, we denote by N_n the number of matches, i.e., the number of sailors that end up picking their own hat. We want to compute the expectation of N_n.

As in Example 1.49, for $k = 1, \dots, n$, we denote by H_k the event "*sailor k pick his own hat*". In equation (1.14) of Example 1.49 we have shown that

$$\mathbb{P}(H_k) = \frac{1}{n}, \;\; \forall k = 1, \dots, n.$$

We denote by I_{H_k} the indicator of H_k, i.e., the random variable that has value 1 if H_k occurs, and value 0 otherwise. We see that I_{H_k} is a Bernoulli random variable with winning probability $\frac{1}{n}$ so

$$\mathbb{E}\big[\, I_{H_k}\,\big] = \mathbb{P}(H_k) = \frac{1}{n},\ \ \forall k = 1, \dots, n.$$

Note that

$$N_n = I_{H_1} + \cdots + I_{H_n},$$

and the linearity of expectation property (3.4) implies

$$\mathbb{E}\big[\, N_n\,\big] = \mathbb{E}\big[\, I_{H_1}\,\big] + \cdots + \mathbb{E}\big[\, I_{H_n}\,\big] = \underbrace{\frac{1}{n} + \cdots + \frac{1}{n}}_{n \text{ times}} = 1.$$

One can show[1] that, as n goes to infinity, the random variable N_n approaches a Poisson random variable with mean 1. More precisely, we have

$$\lim_{n\to\infty} \mathbb{P}(N_n = k) = \frac{e^{-1}}{k!}. \qquad \square$$

Definition 3.12 (Covariance and correlation)**.** Suppose that X, Y are discrete random variables with ranges $\mathscr{X}$ and respectively $\mathscr{Y}$. Let $p(x, y)$ be the joint pmf of the random vector (X, Y). We assume additionally that X and Y are 2-integrable, μ_X is the mean of X and μ_Y.

(i) The *covariance* of X, Y is

$$\begin{aligned}\boldsymbol{cov}[X,Y] &= \mathbb{E}\big[\,(X-\mu_X)(Y-\mu_Y)\,\big] = \mathbb{E}\big[\,XY\,\big] - \mu_Y\mathbb{E}[X] - \mu_X\mathbb{E}[Y] + \mu_X\mu_Y\\ &= \mathbb{E}[XY] - \mu_X\mu_Y = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].\end{aligned}$$

(ii) The *correlation coefficient* of X, Y is the number

$$\boxed{\rho[X,Y] = \frac{\boldsymbol{cov}[X,Y]}{\sqrt{\boldsymbol{var}[X]\,\boldsymbol{var}[Y]}}}.$$

(iii) The discrete random variables X, Y are called *uncorrelated* if

$$\rho[X,Y] = 0 = \boldsymbol{cov}[X,Y]. \qquad \square$$

Proposition 3.13. *Suppose that X, Y are discrete random variables and a, b are real numbers. Then*

$$\boldsymbol{var}[aX+bY] = a^2\,\boldsymbol{var}[X] + b^2\,\boldsymbol{var}[Y] + 2ab\,\boldsymbol{cov}[X,Y]. \tag{3.5}$$

Proof. Let μ_X, μ_Y denote the means of X and respectively Y. Using (3.4) we deduce that the mean μ of $aX + bY$ is

$$\mu = \mathbb{E}[aX+bY] = a\mathbb{E}[X] + b\mathbb{E}[Y] = a\mu_X + b\mu_Y.$$

[1] See http://www.randomservices.org/random/urn/Matching.html

Then

$$(aX+bY)-\mu = aX+bY-(a\mu_X+b\mu_Y) = a(X-\mu_X)+b(Y-\mu_Y)$$

$$\boldsymbol{var}[aX+bY] = \mathbb{E}\big[\,(aX+bY-\mu)^2\,\big].$$

We have

$$(aX+bY-\mu)^2 = \big(\,a(X-\mu_X)+b(Y-\mu_Y)\,\big)^2$$

$$= a^2(X-\mu_X)^2+b^2(Y-\mu_Y)^2+2ab(X-\mu_X)(Y-\mu_Y).$$

Hence

$$\boldsymbol{var}[aX+bY] = \mathbb{E}\big[\,a^2(X-\mu_X)^2+b^2(Y-\mu_Y)^2+2ab(X-\mu_X)(Y-\mu_Y)\big]$$

$$\overset{(3.4)}{=} a^2\mathbb{E}\big[\,(X-\mu_X)^2\,\big]+b^2\mathbb{E}\big[\,(Y-\mu_Y)^2\,\big]+2ab\mathbb{E}\big[\,(X-\mu_X)(Y-\mu_Y)\,\big]$$

$$= a^2\,\boldsymbol{var}[X]+b^2\,\boldsymbol{var}[Y]+2ab\,\boldsymbol{cov}[X,Y].$$

□

From (3.5) we obtain immediately the following useful result.

Corollary 3.14. *If X and Y are uncorrelated discrete random variables, then*

$$\boldsymbol{var}[X+Y] = \boldsymbol{var}[X]+\boldsymbol{var}[Y].$$ □

Example 3.15. Suppose that $(S,\mathbb{P})$ is a probability space and $A_1,A_2\subset S$ are two events with probabilities

$$\mathbb{P}(A_i)=p_i,\;\; i=1,2.$$

As usual we set $q_i=1-p_i$, $i=1,2$. The indicator functions

$$I_{A_1},I_{A_2}:S\to\mathbb{R},\;\; I_{A_i}(s)=\begin{cases}1, & s\in A_i,\\ 0, & s\in S\setminus A_i,\end{cases},\quad i=1,2,$$

are Bernoulli random variables, $I_{A_i}\sim \mathrm{Ber}(p_i)$. Observing that $I_{A_1}I_{A_2}=I_{A_1\cap A_2}$ and

$$\mathbb{E}[I_{A_1}]=p_1,\;\; \mathbb{E}[I_{A_2}]=p_2,$$

we deduce that

$$\boldsymbol{cov}[I_{A_1},I_{A_2}] = \mathbb{E}\big[\,I_{A_1}I_{A_2}\,\big]-\mathbb{E}[I_{A_1}]\mathbb{E}[I_{A_2}]$$

$$= \mathbb{E}[I_{A_1\cap A_2}]-p_1p_2 = \mathbb{P}(A_1\cap A_2)-\mathbb{P}(A_1)\mathbb{P}(A_2).$$

Thus, the random variables I_{A_1} and I_{A_2} are uncorrelated if and only if the events A_1 and A_2 are independent. □

Proposition 3.16. *Suppose that the discrete random variables X and Y are independent. Then, for any functions f and g the random variables $f(X)$ and $g(Y)$ are independent and*

$$\boxed{\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)] \cdot \mathbb{E}[g(Y)]}. \tag{3.6}$$

In particular, if the independent random variables X and Y are uncorrelated

$$\boldsymbol{cov}[X, Y] = 0,$$

and

$$\boxed{\boldsymbol{var}[X + Y] = \boldsymbol{var}[X] + \boldsymbol{var}[Y]}. \tag{3.7}$$

Proof. Denote by p_X and respectively p_Y the pmf's of X and Y. Since these random variables are indepedent, the joint pmf of the random vector (X, Y) is

$$p(x, y) = p_X)x)p_Y(y).$$

Using the law of the subconscious statistician we deduce

$$\mathbb{E}[f(X)g(Y)] = \sum_{x\in\mathscr{X},\, y\in\mathscr{Y}} f(x)g(y)p(x, y) = \sum_{x\in\mathscr{X},\, y\in\mathscr{Y}} f(x)g(y)p_X(x)p_Y(y)$$

$$= \left(\sum_{x\in\mathscr{X}} f(x)p_X(x)\right)\left(\sum_{y\in\mathscr{Y}} g(y)p_Y(y)\right) = \mathbb{E}[f(X)] \cdot \mathbb{E}[g(Y)].$$

In the special case

$$f(X) = X - \mu_X, \;\; g(Y) = Y - \mu_Y,$$

where μ_X and μ_Y are the expectations of X and respectively Y, then

$$\mathbb{E}[f(X)]\mathbb{E}[g(Y)] = 0$$

and

$$\boldsymbol{cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y] = 0.$$

Hence X and Y are uncorrelated. The equality (3.7) now follows from (3.5). □

Arguing inductively we deduce the following useful result.

Corollary 3.17. *If the discrete random variables $X_1, \dots, X_n$ are independent, then*

$$\boxed{\boldsymbol{var}[X_1 + \cdots + X_n] = \boldsymbol{var}[X_1] + \cdots + \boldsymbol{var}[X_n]}. \tag{3.8}$$

□

Example 3.18. We have seen that two discrete independent random variables X, Y are not correlated, i.e., $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. However *there exists uncorrelated random variables that are dependent.* Here is such an example.

Consider the discrete random variables X, Y with joint pmf $p(x, y)$ described by the table below

y				
1	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{1}{20}$	
0	$\frac{2}{20}$	$\frac{8}{20}$	$\frac{2}{20}$	
-1	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{1}{20}$	
$p(x,y)$	-1	0	1	x

Observe that

$$\mathbb{P}(X=-1)=\mathbb{P}(Y=-1)=\frac{1}{20}+\frac{2}{20}+\frac{1}{20}=\frac{4}{20}=\frac{1}{5}=\mathbb{P}(X=1)=\mathbb{P}(Y=1).$$

$$\mathbb{P}(X=0)=\mathbb{P}(Y=0)=\frac{2}{20}+\frac{8}{20}+\frac{2}{20}.$$

We have

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

and, similarly, $\mathbb{E}[Y]=0$. We deduce

$$\begin{gathered}\boldsymbol{cov}[X,Y]=\mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]=\mathbb{E}[XY]\\ =1\cdot\big(\,\mathbb{P}(X=-1,Y=-1)+\mathbb{P}(X=1,Y=1)\,\big)\\ +(-1)\cdot\big(\,\mathbb{P}(X=-1,Y=1)+\mathbb{P}(X=1,Y=-1)\,\big)=0.\end{gathered}$$

Thus, the random variables X,Y are uncorrelated. On the other hand

$$\mathbb{P}(X=1,Y=1)=\frac{1}{20}\neq\mathbb{P}(X=1)\cdot\mathbb{P}(Y=1)=\frac{1}{25}.$$

Hence, the random variables X, Y are *dependent.* □

Remark 3.19 (Predictors/Estimators)**.** If $\boldsymbol{cov}[X,Y]>0$ we say that the random variables are *positively correlated.* If $\boldsymbol{cov}[X,Y]<0$, then the random variables are said to be *negatively correlated.* Negative correlation suggests that when one of the random variables has large values, the other has small values. Positive correlation suggest that the two random variables tend to have large and small values at the same time. This can be argued as follows.

Suppose that we are interested in a random variable Y, but we can only have information about a random variable X. In statistics, a function $g(X)$ of X is referred to as a *predictor* or *estimator* of Y based on X. The *mean square error of a predictor* is the quantity

$$\mathbb{E}\big[\,(\,Y-g(X)\,)^2\,\big].$$

This measures how far is the predicted value $g(X)$ from the actual value Y.

A predictor $g(X)$ is called *linear* if it has the form $g(X)=mX+b$, where $m,b\in\mathbb{R}$. A *best linear predictor* is a linear predictor that produces the smallest mean square error among all the *linear* predictors. The *linear regression formula* states that the best linear predictor of Y based on X is

$$\boxed{L_Y(X) := \mu_Y + \frac{\rho\sigma_Y}{\sigma_X}\big(\,X - \mu_X\,\big)}, \tag{3.9}$$

where μ_X, μ_Y are the means of X respectively Y, σ_X, σ_Y are the standard deviations of X and Y and ρ is the correlation coefficient of X and Y. Loosely speaking, $L_Y(X)$ is the best linear approximation of Y given X.

Note that if $\rho > 0$, i.e., the random variables X and Y are *positively correlated* then $L_Y(X)$ is an *increasing function* of X. Analogously, if $\rho < 0$, i.e., the random variables X and Y are *negatively correlated* then $L_Y(X)$ is a *decreasing function* of X. □

Example 3.20. Suppose that $X \sim \mathrm{Bin}(n, p)$. As we have remarked earlier, X is the sum of n independent Bernoulli variables with probability of success p

$$X = X_1 + \cdots + X_n, \;\; X_k \sim \mathrm{Ber}(p), \;\; \forall k = 1, \dots, n.$$

Hence

$$\mathbb{E}[X] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] = \underbrace{p + \cdots + p}_{n\ times} = np.$$

This provides another confirmation of the first half the equality (2.24) obtained by more laborious means.

Also, due to the independence of the random variables $X_1, \dots, X_n$, we have

$$\boldsymbol{var}[X] = \boldsymbol{var}[X_1] + \cdots + \boldsymbol{var}[X_n] = n\,\boldsymbol{var}[X_1] = np(1-p).$$

This confirms the second half of (2.24). □

Example 3.21 (The coupon collector problem)**.** The coupon collector's problem arises from the following scenario. Suppose that each box of cereal contains one of n different coupons. Once you obtain one of every type of coupon, you can send in for a prize. Assuming that the coupon in each box is chosen independently and uniformly at random from the n possibilities and that you do not collaborate with others to collect coupons, how many boxes of cereal must you buy before you obtain at least one of every type of coupon?

Let X denote the number of boxes bought until at least one of every coupon is obtained. We want to determine $\mathbb{E}[X]$. For $i = 1, \dots, n-1$ denote by X_i the number of boxes you bought while you had exactly i coupons. The first box you bought contained one coupon. Then you bought X_1 boxes containing the coupon you already had. After $1 + X_1$ boxes you have two coupons. Next you bought X_2 boxes containing one of the two coupons you already had etc. Hence

$$X = 1 + X_1 + \cdots + X_{n-1}.$$

Let us observe first that for $i = 1, \cdots, n-1$ we have $X_i \sim \mathrm{Geom}(p_i)$, $p_i = \frac{n-i}{n}$, $q_i = 1 - p_i = \frac{i}{n}$. Observe next that the random variables X_i are *independent*.

Indeed, at the moment you have i coupons, a success occurs when you buy one of the remaining $n - i$ coupons. The probability of buying one such coupon is thus $\frac{n-i}{n}$. Think of buying a box at this time as a Bernoulli trial with success probability $\frac{n-i}{n}$. The number X_i is then equal to the number of trials until you register the first success. In particular,

$$\mathbb{E}[X_i] = \frac{1}{p_i} = \frac{n}{n-i}, \quad \boldsymbol{var}[X_i] = \frac{\frac{i}{n}}{\frac{(n-i)^2}{n^2}} = \frac{ni}{(n-i)^2}$$

From the linearity of expectation we deduce

$$\begin{aligned}\mathbb{E}[X] = 1 + \mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_{n-1}] &= \frac{n}{n} + \frac{n}{n-1} + \frac{n}{n-2} + \cdots + \frac{n}{1} \\ &= n \underbrace{\left(1 + \frac{1}{2} + \cdots + \frac{1}{n-1} + \frac{1}{n}\right)}_{=:H_n}.\end{aligned}$$

Also

$$\begin{aligned}\boldsymbol{var}[X] &= \boldsymbol{var}[X_1] + \cdots + \boldsymbol{var}[X_{n-1}] \\ &= \sum_{i=1}^{n-1} \frac{ni}{(n-i)^2} = n\left(\frac{n-1)}{1^2} + \frac{(n-2)}{2^2} + \cdots + \frac{n-(n-1)}{(n-1)^2}\right) \\ &= n^2\left(\frac{1}{1^2} + \frac{1}{2^2} + \cdots + \frac{1}{(n-1)^2}\right) - n\left(\frac{1}{1} + \cdots + \frac{1}{n-1}\right) \\ &= n^2\left(\frac{1}{1^2} + \frac{1}{2^2} + \cdots + \frac{1}{(n-1)^2} + \frac{1}{n^2}\right) - n\left(\frac{1}{1} + \cdots + \frac{1}{n-1} + \frac{1}{n}\right) \\ &= n^2\left(\frac{1}{1^2} + \frac{1}{2^2} + \cdots + \frac{1}{(n-1)^2} + \frac{1}{n^2}\right) - nH_n.\end{aligned}$$

One can show that there exists a mysterious constant[2] $\boldsymbol{\gamma} \approx 0.577$ such that

$$\lim_{n\to\infty} (H_n - \log n) = \boldsymbol{\gamma}.$$

Thus the expected number of boxes needed to collect all the n coupons is about $n \log n + n\boldsymbol{\gamma}$.

On the other hand, a famous result of Euler states that

$$\frac{1}{1^2} + \frac{1}{2^2} + \cdots + \frac{1}{n^2} + \cdots = \frac{\pi^2}{6}.$$

[2] The constant $\boldsymbol{\gamma}$ is called the Euler-Mascheroni constant and $\boldsymbol{\gamma} = 0.5772\ldots$. For more details see the Wikipedia page `https://en.wikipedia.org/wiki/Euler-Mascheroni_constant`

This shows that

$$\lim_{n\to\infty}\frac{\boldsymbol{var}[X_n]}{n^2}=\frac{\pi^2}{6}-\lim_{n\to\infty}\frac{H_n}{n}=\frac{\pi^2}{6}.$$

We can use the coupon collector problem to solve another birthday problem: how many people should you expect to meet until you know a person born in each day of the year. The answer is $365\cdot H_{365}$. Using the R command

```
sum(365/(1:365))
```

we deduce

$$365\cdot H_{365}\approx 2364.646.$$

□

Remark 3.22. The coupon collector problem has many important uses in computer sciences and much more is known. It turns out that the odds that the number of boxes is a lot bigger than the expectation $n\log n$ are very small for large n. More precisely, for any $c>0$ we have [14, Thm.3.8]

$$\lim_{n\to\infty}\mathbb{P}[X>n\log n+cn]=p(c)=1-e^{-e^{-c}}.$$

For example, when $c=20$ we have $p(c)\approx 2\cdot 10^{-9}$.

Example 3.23 (Balls-in-bins)**.** This problem has origins in theoretical computer science. We have m balls that we randomly distribute among N bins, each ball being equally likely to be placed in any of the N bins. We seek to understand the random number X_m of nonempty bins. More precisely we want to compute the mean and the variance of this random variable.

Here are a few equivalent descriptions of this problem. Suppose that m people located on the ground floor of a tall building enter an elevator. We know that there are N stories above the ground level and each passenger is equally likely to get out at any of N stories above. We tacitly assume that at each stop no person gets in but at least one person gets out of the elevator. In this case X_m is the number of stops until everyone gets out.

Equivalently, suppose that Santa comes to a kindergarden attended by N kids. Santa carries m gifts in his bags and distributes all of them randomly among the kids so that each kid is equally likely to receive any of the m gifts. During this process it is possible that some kids will receive more than one gift, while some will receive none. We will refer to the kids that have received a gift as "*lucky*" and to the remaining kids as "*unlucky*". Hence X_m can be identified with the number of lucky kids.

When $N=365$, we can think of the bins as the days of the year, and then X_m is the number of different birthdays in a random group of m people: each of these m persons was "parachuted" randomly by the delivery stork in a bin labeled by his/her birthday. In Exercise 2.8 you were asked to compute $\mathbb{E}[X_4]$.

For $k = 1, \dots, N$, we denote by B_k the event that the k-th bin is nonempty after all the m-balls were randomly distributed. Denote by I_{B_k} the indicator of B_k. Then

$$X_m = I_{B_1} + \cdots + I_{B_N}.$$

Hence

$$\mathbb{E}[X_m] = \sum_{k=1}^{N} \mathbb{E}[I_{B_k}] = \mathbb{P}(B_1) + \cdots + \mathbb{P}(B_N).$$

All the above probabilities are equal and we denote by p their common value. Hence

$$\mathbb{E}[X_m] = Np.$$

Observe that B_1^c is the event that the first bin is empty after all the m balls were distributed. The probability that one of the balls is placed in a bin *other than* B_1 is $\frac{N-1}{N}$. Hence

$$\mathbb{P}(B_1^c) = \left(\frac{N-1}{N}\right)^m, \quad \boxed{p = 1 - \mathbb{P}(B^c) = 1 - \left(\frac{N-1}{N}\right)^m}.$$

In particular, we deduce

$$\boxed{\mathbb{E}[X_m] = N\left(1 - \left(\frac{N-1}{N}\right)^m\right)}. \tag{3.10}$$

Next observe that

$$\boldsymbol{var}[X_m] = \mathbb{E}[X_m^2] - \mathbb{E}[X_m]^2.$$

We have

$$X_m^2 = \Big(I_{B_1} + \cdots + I_{B_N}\Big)^2$$
$$= I_{B_1}^2 + \cdots + I_{B_N}^2 + 2\sum_{j<k} I_{B_j} I_{B_k} = I_{B_1} + \cdots + I_{B_N} + 2\sum_{j<k} I_{B_j \cap B_k}.$$

Hence

$$\mathbb{E}[X_m^2] = \mathbb{E}\big[\, I_{B_1}\,\big] + \cdots + \mathbb{E}\big[\, I_{B_N}\,\big] + 2\sum_{j<k} \mathbb{E}\big[\, I_{B_j \cap B_k}\,\big]$$
$$= N\mathbb{E}\big[\, I_{B_1}\,\big] + 2\binom{N}{2}\mathbb{E}[I_{B_1 \cap B_2}\,\big] = N\mathbb{E}\big[\, I_{B_1}\,\big] + N(N-1)\mathbb{E}[I_{B_1 \cap B_2}\,\big],$$

since all the events $B_j \cap B_k$ have the same probability. We set

$$r := \mathbb{P}(B_1 \cap B_2).$$

The inclusion-exclusion principle shows that

$$r = \mathbb{P}(B_1 \cap B_2) = \mathbb{P}(B_1) + \mathbb{P}(B_2) - \mathbb{P}(B_1 \cup B_2) = 2p - \mathbb{P}(B_1 \cup B_2).$$

Now observe that

$$\mathbb{P}(B_1 \cup B_2) = 1 - \mathbb{P}\Big(\big(\, B_1 \cup B_2\,\big)^c\Big) = 1 - \mathbb{P}(B_1^c \cap B_2^c\,\big).$$

The probability $\mathbb{P}(B_1^c \cap B_2^c)$ that both bin 1 and 2 are empty after all the m balls were distributed is

$$\mathbb{P}(B_1^c \cap B_2^c) = \left(\frac{N-2}{N}\right)^m.$$

Hence

$$\boxed{r = 2p - 1 - \left(\frac{N-2}{N}\right)^m}.$$

so

$$\mathbb{E}[X_m^2] = Np + N(N_1)r$$

and

$$\boxed{\boldsymbol{var}[X_m] = Np + N(N_1)r - (Np)^2}. \qquad \square$$

3.2. Conditioning

Suppose that X and Y are discrete random variables with ranges $\mathscr{X}$ and respectively $\mathscr{Y}$. For $x \in \mathscr{X}$, we define the *conditional pmf of* Y *given that* $X = x$ to be the function

$$p_{Y|X=x}(-) : \mathscr{Y} \to [0,1], \ \ p_{Y|X=x}(y) := \mathbb{P}(Y = y | X = x).$$

If $p(x,y) = \mathbb{P}(X = x, Y = y)$ is the joint pmf of (X,Y), then

$$\boxed{p_{Y|X=x}(y) = \frac{p(x,y)}{p_X(x)}}, \tag{3.11}$$

where p_X is the pmf of X. Note that

$$\sum_{y\in\mathscr{Y}} p_{Y|X=x}(y) = \frac{1}{p_X(x)} \sum_{y\in Y} p(x,y) \overset{(3.2)}{=} 1.$$

Thus, the conditional pmf $p_{Y|X=x}(y)$ is the pmf of a discrete random variable with range contained in $\mathscr{Y}$. We denote this random variable by $Y|X = x$. The expectation of the conditioned random variable $(Y|X = x)$ is *the number*

$$\boxed{\mathbb{E}[Y|X = x] = \sum_y y p_{Y|X=x}(y) \overset{(3.11)}{=} \frac{1}{p_X(x)} \sum_{y\in\mathscr{Y}} y p(x,y)}.$$

We will refer to it as the *conditional expectation of* Y *given that* $X = x$. From the above we deduce the following consequence

Corollary 3.24. *The discrete random variables* X *and* Y *are independent if and only if*

$$p_{Y|X=x}(y) = p_Y(y), \ \ \forall x \in \mathscr{X}, \ \ y \in \mathscr{Y}.$$

Moreover, if X *and* Y *are independent, then* $\mathbb{E}[Y|X = x] = \mathbb{E}[Y]$.

Proof.

$$p_{Y|X=x}(y) = p_Y(y) \Longleftrightarrow \frac{p(x,y)}{p_X(x)} = p_Y(y) \Longleftrightarrow p(x,y) = p_X(x)p_Y(y).$$

If X and Y are independent, then

$$\mathbb{E}[Y|X=x] = \sum_{y\in\mathcal{Y}} y p_{Y|X=x}(y) = \sum_{y\in\mathcal{Y}} y p_Y(y) = \mathbb{E}[Y].$$

□

Example 3.25. Let us look again at the situation analyzed in Example 3.7, where N is the total number of callers, X the number of female callers and Y is the number of male callers at a call center. As observes in Example 3.7, Y is independent of X and $X, Y \sim \mathrm{Poi}(2)$. Thus

$$\mathbb{P}(X=i, Y=j) = \mathbb{P}(X=i)\mathbb{P}(Y=j).$$

The conditional pmf of N given $X = i$ is

$$p_{N|X=j}(i+j) = \mathbb{P}(N = i+j|X=j)$$

$$= \frac{\mathbb{P}(X=i, Y=j)}{\mathbb{P}(X=j)} = \mathbb{P}(Y=j) = e^{-2}\frac{2^j}{j!}.$$

Thus, given that the number of female callers is i, the total number of callers k is equal to i plus an independent random number (j) of male callers, distributed as Poi(2). □

Example 3.26. Suppose that $X \sim \mathrm{Bin}(m,p)$ and $Y \sim \mathrm{Bin}(n,p)$ are independent binomial random variables with the same probability of success p. We want to compute the conditional pmf of X given that $X+Y=r$, i.e., the function

$$\mathbb{P}(X=j|X+Y=r).$$

Note that we have

$$\mathbb{P}(X=j|X+Y=r) = \frac{\mathbb{P}(X=j, Y=r-j)}{\mathbb{P}(X+Y=r)}.$$

Observing that $X+Y \sim \mathrm{Bin}(m+n,p)$ we deduce

$$\mathbb{P}(X=j|X+Y=r) = \frac{\binom{m}{j}p^j q^{m-j}\binom{n}{r-j}p^{r-j}q^{n-r+j}}{\binom{n+m}{r}p^r q^{n+m-r}} = \frac{\binom{m}{j}\binom{n}{r-j}}{\binom{n+m}{r}}.$$

Thus the conditional distribution of X given that $X+Y=r$ is HGeom(m,n,r). In particular, the statistic of $(X|X+Y=r)$ *is independent of the success probability p.* □

The linearity property of expectation, i.e., the equality (3.4), has a conditional counterpart. More precisely, given discrete random variables $X, Y_1, \dots, Y_n$ and constants $c_1, \dots, c_n$, we have

$$\boxed{\mathbb{E}\big[\, c_1 Y_1 + \cdots + c_n Y_n \,|\, X = x\,\big] = c_1 \mathbb{E}[Y_1 \,|\, X = x] + \cdots + c_n \mathbb{E}[Y_n \,|\, X = x]}. \tag{3.12}$$

Definition 3.27. We denote by $\mathbb{E}[Y|X]$ the <u>*random variable*</u> that takes the value $\mathbb{E}[Y|X = x]$ when $X = x$. We will refer to its as the *conditional expectation of Y given X.* □

The random variable $\mathbb{E}[Y|X]$ *is a function of X* and the law of subconscious statistician shows that its expectation is

$$\mathbb{E}\big[\,\mathbb{E}[Y|X]\,\big] = \sum_{x \in \mathscr{X}} \mathbb{E}[Y|X = x] p_X(x)$$

$$= \sum_{x \in \mathscr{X}} \mathbb{E}\big[\,Y|X = x\,\big] \cdot \left(\sum_y y p_{Y|X}(y|x) \right) p_X(x)$$

$$= \sum_{x \in \mathscr{X},\, y \in \mathscr{Y}} y \underbrace{p_{Y|X}(y|x) p_X(x)}_{p(x,y)} = \sum_{x \in \mathscr{X},\, y \in \mathscr{Y}} y p(x,y)$$

$$= \sum_{y \in \mathscr{Y}} y \sum_{x \in \mathscr{X}} p(x,y) \overset{(3.2)}{=} \sum_{y \in \mathscr{Y}} y p_Y(y) = \mathbb{E}[Y].$$

We have thus shown that

$$\boxed{\mathbb{E}[Y] = \sum_{x \in \mathscr{X}} \mathbb{E}\big[\,Y \,|\, X = x\,\big] p_X(x)}. \tag{3.13}$$

Note that the above formula makes no reference to the joint pmf of (X, Y); all we need to know is the conditional pmf $p_{Y|X=x}(y)$. The above equalities generalize the law of total probability, Theorem 1.46.

Example 3.28. An urn contains 999 balls labelled $1, \dots, 999$. Draw at random a ball from the urn and let L denote its label. Next, roll a die L times and record the number of times N that you get a 6. We want to compute the expectation of N, i.e., the expected number of 6's we get.

Set $p := \frac{1}{6}$ and $q := \frac{5}{6}$. Note that if we condition on L we have

$$(N|L = \ell) \sim \mathrm{Bin}(\ell, p)$$

so

$$\mathbb{E}[N|L = \ell] = p\ell = \frac{\ell}{6}.$$

Using (3.13) we deduce

$$\mathbb{E}[N] = \mathbb{E}[N|L = 1]\mathbb{P}(L = 1) + \cdots + \mathbb{E}[N|L = 999]\mathbb{P}(L = 999)$$

$$= \frac{1}{6\cdot 999} + \frac{2}{6\cdot 999} + \cdots + \frac{999}{6\cdot 999} = \frac{\frac{1000\cdot 999}{2}}{6\cdot 999} = \frac{500}{6} \approx 83.33. \qquad \square$$

Remark 3.29. The conditional expectation of Y given X is a very subtle and fundamental concept with many uses. It is essentially the best information about Y given the knowledge of X. As indicated above, the conditional expectation $\mathbb{E}[Y|X]$ is a *function of* X

$$\mathbb{E}[Y|X] = g_0(X)$$

Using the language of *predictors* or *estimators* in Remark 3.19 we can say that $\mathbb{E}[Y|X]$ is a *predictor* of Y given X. It turns out that $g_0(X) = \mathbb{E}[Y|X]$ is the *best predictor* in the sense that it has the *smallest mean square error* among *all* predictors, i.e., for any other predictor $g(X)$ we have

$$\mathbb{E}\big[\,(\,Y - g_0(X)\,)^2\,\big] \leq \mathbb{E}\big[\,(Y - g(X)\,)^2\big]. \qquad \square$$

Example 3.30. Suppose that Y is a discrete random variable on the sample space S with pmf p_Y, and B is an event with probability $p = \mathbb{P}(B) > 0$.

The indicator function I_B is a random variable with range $\{0,1\}$. The conditional expectation $\mathbb{E}[Y|I_B = 1]$ is denoted $\mathbb{E}[Y|B]$ and it is called the *conditional expectation of* Y *given* B. We define the *conditional pmf* $p_{Y|B}$ by the equality

$$\boxed{p_{Y|B}(y) := \mathbb{P}\big(Y = y|B\big)}.$$

We have

$$\boxed{\mathbb{E}[Y|B] = \sum_y y\mathbb{P}[Y = y|B] = \sum_y y p_{Y|B}(y).} \tag{3.14}$$

$\square$

We have the following generalization of the law of total probability (1.13).

Proposition 3.31. *Suppose that the events $A_1, \dots, A_n \subset S$ partition of the sample space S., i.e., their union is S and they are mutually disjoint. Suppose next that $Y : S \to \mathbb{R}$ is a discrete random variable with range $\mathcal{Y}$.*

$$\boxed{\mathbb{E}\big[Y\big] = \mathbb{E}[Y|A_1]\mathbb{P}(A_1) + \cdots + \mathbb{E}[Y|A_n]\mathbb{P}(A_n).} \tag{3.15}$$

Proof. Consider the random variable

$$X = I_{A_1} + 2I_{A_2} + \cdots + nI_{A_n}.$$

Note that $A_k = \{X = k\}$. The equality (3.15) becomes

$$\mathbb{E}\big[Y\big] = \mathbb{E}[Y|X = 1]\mathbb{P}(X = 1) + \cdots + \mathbb{E}[Y|X = n]\mathbb{P}(X = n).$$

This is clearly a special case of (3.13). $\square$

Example 3.32 (Coin patterns)**.** A *coin pattern* is an ordered string of symbols

$$p_1, p_2, \dots, p_k \in \{H, T\},\ \ H = \text{ Heads},\ \ T = \text{ Tails}.$$

For example, HH and HT are patterns.

Fix a pattern $\boldsymbol{p} = (p_1, \dots, p_k)$. We flip a fair coin until we notice for the first time the pattern $\boldsymbol{p}$ appearing in successive flips. We denote by $W_{\boldsymbol{p}}$ the number of flips (or waiting time) until we first observe the pattern $\boldsymbol{p}$. We want to compute the expectation of $W_{\boldsymbol{p}}$ in two special cases $\boldsymbol{p} = HT$ and $\boldsymbol{p} = HH$.

To compute the expectations $E_{HT} := \mathbb{E}[W_{HT}]$ and $E_{HH} := \mathbb{E}[W_{HH}]$ we condition on the face F_k that shows up the k-th flip. We have

$$E_{HT} = \mathbb{E}\big[\, W_{HT} | F_1 = T\,\big]\mathbb{P}(F_1 = T) + \underbrace{\mathbb{E}\big[\, E_{HT} | F_1 = H\,\big]}_{=:E_{HT|H}}\mathbb{P}(F_1 = H)$$

$$= \frac{1}{2}\Big(\mathbb{E}[W_{HT}|F_1 = T] + E_{HT|H}\Big).$$

Note that

$$\mathbb{E}\big[\, W_{HT} | F_1 = T\,\big] = \mathbb{E}[W_{HT}] + 1 = E_{HT} + 1,$$

because, if the first flip is not H, then it is as if we have to start the game all over again, having wasted one flip. Thus

$$E_{HT} = \frac{1}{2}\Big(E_{HT} + 1 + E_{HT|H}\Big) \Rightarrow E_{HT} = 1 + E_{HT|H}. \tag{3.16}$$

On the other hand

$$E_{HT|H} = \mathbb{E}\big[\, W_{HT} | F_1 = H\,\big] = \underbrace{\mathbb{E}\big[\, W_{HT} | F_1 = H, F_2 = T\,\big]}_{=2}\mathbb{P}(F_2 = T)$$

$$+ \underbrace{\mathbb{E}\big[\, W_{HT} | F_1 = H, F_2 = H\,\big]}_{=E_{HT|H}+1}\mathbb{P}(F_2 = H) = 1 + \frac{1}{2}\big(E_{HT|H} + 1\big).$$

Hence

$$E_{HT|H} = 3,\ \ \boxed{E_{HT} = 4}.$$

Similarly

$$E_{HH} = \mathbb{E}\big[\, W_{HH} | F_1 = T\,\big]\mathbb{P}(F_1 = T) + \underbrace{\mathbb{E}\big[\, W_{HH} | F_1 = H\,\big]}_{=:E_{HH|H}}\mathbb{P}(F_1 = H)$$

$$= \frac{1}{2}\big(\, E_{HH} + 1 + E_{HH|H}\Big) \Rightarrow E_{HH} = E_{HH|H} + 1.$$

$$E_{HH|H} = \underbrace{\mathbb{E}\big[\, W_{HH} | F_1 = F_2 = H\,\big]}_{=2}\mathbb{P}(F_2 = H)$$

$$+ \underbrace{\mathbb{E}\big[\, W_{HH} | F_1 = H, F_2 = T\,\big]}_{=H_{HH}+2}\mathbb{P}(F_2 = T)$$

$$= 1 + \frac{1}{2}(E_{HH} + 2) = \frac{1}{2}E_{HH} + 2.$$

Hence

$$E_{HH} = \frac{1}{2}\big(\,E_{HH|H} + 2\,\big) + 1 \Rightarrow \boxed{E_{HH} = 6}.$$

The above computations lead to a rather surprising conclusion: although each face of a fair coin is equally likely to occur, the pattern HH is less frequent than the pattern HT since on average it takes more flips to observe HH than to observe HT. □

Remark 3.33. There exists a very beautiful formula that describes the expected waiting time of any pattern of heads and tails.

Suppose that $\boldsymbol{p} = (p_1, \dots, p_n)$ is and H-and-T pattern. We will refer to n as the length of the pattern. E.g. $\boldsymbol{p}_0 = HHTHT$, is a pattern of length 5.

The *right/left-truncation* of a pattern $\boldsymbol{p}$ is the pattern $R(\boldsymbol{p})$ (respectively $L(\boldsymbol{p})$) obtained from $\boldsymbol{p}$ by by removing its rightmost (respectively leftmost) entry

$$L(\boldsymbol{p}) = (p_2, \dots, p_n), \;\; R(\boldsymbol{p}) = (p_1, \dots, p_{n-1}).$$

E.g.,

$$L(HHTHT) = HTHT, \;\; R(HHTHT) = HHTH.$$

For any natural number k we denote by $L^k(\boldsymbol{p})$ the pattern obtained from $\boldsymbol{p}$ after k left-truncations. We define $R^k(\boldsymbol{p})$ is an analogous way. We set

$$\epsilon_k = \epsilon_k(\boldsymbol{p}) = \begin{cases} 1, & R^k(\boldsymbol{p}) = L^k(\boldsymbol{p}), \\ 0, & R^k(\boldsymbol{p}) \neq L^k(\boldsymbol{p}). \end{cases}$$

Then, the expected time $T(\boldsymbol{p})$ to first observe the pattern $\boldsymbol{p}$ of length n is

$$\tau(\boldsymbol{p}) = 2^n + \epsilon_1(\boldsymbol{p})2^{n-1} + \epsilon_2(\boldsymbol{p})2^{n-2} + \cdots + \epsilon_{n-1}(\boldsymbol{p})2. \tag{3.17}$$

For example, when $\boldsymbol{p} = HH$, then

$$L(HH) = R(HH), \; \epsilon_1 = 1,$$

and thus

$$\tau(HH) = 2^2 + 2 = 6,$$

which agrees with our computations. When $\boldsymbol{p} = HTH$, then

$$L(\boldsymbol{p}) = TH, \;\; R(\boldsymbol{p}) = HT, \;\; \epsilon_1 = 0,$$

$$L^2(\boldsymbol{p}) = H, \;\; R^2(\boldsymbol{p}) = H, \;\; \epsilon_2 = 1,$$

so the expected time to observe the pattern HTH is

$$\tau(HTH) = 2^3 + 2 = 10. \tag{3.18}$$

One of the most elegant proofs of (3.17) is due to S.-Y.R. Li [13] and uses a very clever betting strategy. For a rather elementary description of this strategy we refer to the nice exposition in [8, p.428].

The numbers $\tau(\boldsymbol{p})$ make their appearance in the Penney-ante game describe so eloquently by Martin Gardner in [6].

Here is a simple R code that can be used to compute the expected waiting time $\tau(\boldsymbol{p})$ to observe a pattern $\boldsymbol{p}$.

```
#patt is the pattern defined as a vector
tau<-function(patt){
n<-length(patt)
m<-n-1
t<-2^n
for (i in 1:m){
j<-n-i
k<-i+1
t<-t+ any(patt[1:j]==patt[k:n])*2^(n-i)
}
t
}
```

For example, to compute the waiting time for the pattern HTH use the command

```
x<-c(1,0,1)
tau(x)
```

In Example 7.19 we describe the R code that can be used to simulate this random experiment. □

Example 3.34 (Gambler's ruin revisited)**.** Consider again the situation in Example 1.53. Recall that Ann plays a sequence of two-player game of chance with Bob. She starts with a fortune a, Bob-s fortune is $N-a$. Ann's winning probability is p and Bob's is $q = 1-p$. Every time Ann wins, she gets a dollar from Bob and every time she losses, she gives Bob a dollar. The sequence of games ends when either player is out of money. Here we consider only the special case of a fair game, $p = q = \frac{1}{2}$.

Let T_a denote the number of of games Ann and Bob play until one of them is ruined, assuming that Ann's fortune is a. Clearly $T_0 = T_N = 0$. We set $t_a := \mathbb{E}[T_a]$, we denote by X the random variable which is equal to 1 if "Ann wins her first game", and it is equal to -1 if "Ann loses her first game". Observe that

$$\mathbb{P}(X=1) = \mathbb{P}(X=-1) = \frac{1}{2}.$$

For $0 < a < N$ we have

$$t_a = \mathbb{E}[T_a] = \mathbb{E}[T_a | X=1]\mathbb{P}(X=1) + \mathbb{E}[T_a | X=1]\mathbb{P}(X=-1).$$

Now observe that

$$\mathbb{E}[T_a|X=1]=1+\mathbb{E}[T_{a+1}],\;\;\mathbb{E}[T_a|X=-1]=1+\mathbb{E}[T_{a-1}],$$

so that

$$t_a=\frac{1}{2}\Big(t_{a+1}+1+t_{a-1}+1\Big)=1+\frac{1}{2}\Big(t_{a+1}+t_{a-1}\Big).$$

We deduce

$$\boxed{t_{a+1}-t_a=t_a-t_{a-1}-2,\;\;\forall 1<a<N.}$$

Thus

$$t_2-t_1=t_1-t_0-2=t_1-2,\;\;t_3-t_2=t_2-t_1-2=t_1-4,$$
$$t_a-t_{a-1}=t_1-2(a-1),\;\;\forall a=1,\dots,N.$$

Observe that

$$\begin{aligned}0=t_N-t_0&=(t_1-t_0)+(t_2-t_0)+\cdots+(t_N-T_{N-1})\\&=\underbrace{t_1+(t_1-2)+(t_1-4)+\cdots+\big(t_1-2(N-1)\big)}_{N\text{ terms}}\\&=Nt_1-2\big(1+2+\cdots+(N-1)\big)=Nt_1-N(N-1).\end{aligned}$$

This proves that

$$t_1=N-1$$

and we deduce

$$\begin{aligned}t_a=t_a-t_0&=(t_1-t_0)+(t_2-t_1)+\cdots+(t_a-t_{a-1})\\&=\underbrace{t_1+(t_1-2)+\cdots+\big(t_1-2(a-1)\big)}_{a\text{ terms}}\\&=at_1-a(a-1)=a(N-1)-a(a-1)=a(N-a).\end{aligned}$$

Thus

$$\boxed{\mathbb{E}[T_a]=a(N-a).}$$ □

Example 3.35 (Balls-in-bins)**.** We consider again the balls-in-bins problem investigated in Example 3.23. One formulation of that problem involved Santa randomly distributing m gifts to N kids. We denote by X_m the number of lucky kids, i.e., kids that have received at least one gift. We want to present a computation of $\mathbb{E}[X_n]$ using the conditioning technique. Set $x_m:=\mathbb{E}[X_m]$. We will argue inductively.

Clearly $x_0=0$ and $x_1=1$. We next attempt to compute X_{m+1} by conditioning on X_m, $m\geq 1$. We have

$$x_{m+1}=\mathbb{E}[X_{m+1}]=\sum_{k>0}\mathbb{E}\big[\,X_{m+1}|X_m=k\,\big]\mathbb{P}(X_m=k).$$

Now observe that

$$\begin{aligned}\mathbb{E}[X_{m+1}|X_m=k]&=\mathbb{E}\big[\,X_m|X_m=k\,\big]+\mathbb{E}\big[\,X_{m+1}-X_m|X_m=k\,\big]\\&=k+\mathbb{E}\big[\,X_{m+1}-X_m|X_m=k\,\big].\end{aligned}$$

To compute $\mathbb{E}\big[\,X_{m+1}-X_m|X_m=k\,\big]$ we argue as follows.

We know that Santa had already distributed m gifts to k kids. At this moment there are k lucky kids and $(N-k)$ unlucky ones. The probability that the $(m+1)$-th gift will go to one of the k lucky kids is $\frac{k}{N}$, and the probability that this gift will go to one of the $(N-k)$ unlucky kids is $\frac{N-k}{N}$. Thus

$$\mathbb{E}\big[\,X_{m+1}-X_m|X_m=k\,\big]=0\frac{k}{N}+1\frac{N-k}{N}=1-\frac{k}{N}.$$

Hence

$$\mathbb{E}\big[\,X_m|X_m=k\,\big]=k+1-\frac{k}{N}=1+k\left(1-\frac{1}{N}\right).$$

For simplicity we set

$$r:=1-\frac{1}{N}$$

so that

$$\mathbb{E}\big[\,X_m|X_m=k\,\big]=1+kr$$

and

$$x_{m+1}=\sum_{k>0}(1+kr)\mathbb{P}(X_m=k)=\underbrace{\sum_{k>0}\mathbb{P}(X_m=k)}_{=1}+r\underbrace{\sum_{k>0}k\mathbb{P}(X_m=k)}_{=\mathbb{E}[X_m]}.$$

Hence we deduce that for any $m>0$ we have

$$x_{m+1}=1+rx_m. \tag{3.19}$$

Recalling that $x_1=1$ we deduce that $x_2=1+r$. Using this information again in (3.19) we deduce

$$x_3=1+rx_2=1+r(1+r)=1+r+r^2.$$

Arguing inductively we deduce

$$x_m=1+r+\cdots+r^{m-1}=\frac{1-r^m}{1-r}.$$

Using the equality $r=1-\frac{1}{N}$ we deduce $1-r=\frac{1}{n}$

$$x_m=N(1-r^m)=N\left(1-\left(1-\frac{1}{N}\right)^m\right)=\frac{N^m-(N-1)^m}{N^{m-1}}. \qquad \square$$

3.3. Multi-dimensional discrete random vectors

The discussion in the previous section has an obvious higher dimensional counterpart.

Definition 3.36. Suppose that $X_1,\dots,X_n$ are discrete random variables with ranges

$$\mathscr{X}_1,\dots,\mathscr{X}_n$$

and pmf-s $p_1,\dots,p_n$. The *joint pmf* of the n-dimensional random vector $(X_1,\dots,X_n)$ is the function

$$p:\mathscr{X}_1\times\cdots\times\mathscr{X}_n\to[0,1],$$

defined by

$$p(x_1,\dots,x_n)=\mathbb{P}(X_1=x_1,\dots,X_n=x_n). \qquad \square$$

Proposition 3.37. *The discrete random variables $X_1, \dots, X_n$ with ranges*

$$\mathscr{X}_1, \dots, \mathscr{X}_n,$$

and pmf-s $p_1, \dots, p_n$ are independent if and only if the joint pmf $p(x_1, \dots, x_n)$ of the the discrete random vectors $(X_1, \dots, X_n)$ satisfies the equality

$$p(x_1, \dots, x_n) = p_1(x_1) \cdots p_n(x_n), \;\; \forall (x_1, \dots, x_n) \in \mathscr{X}_1, \times \cdots \times \mathscr{X}_n. \qquad \square$$

Corollary 3.38. *Suppose that random variables $X_1, \dots, X_n$ are independent. Then the following hold.*

(i) *For any $1 \leq k \leq n-1$, and any functions*

$$f(x_1, \dots, x_k), \;\; g(x_{k+1}, \dots, x_n),$$

the random variables $f(X_1, \dots, X_k)$ and $g(X_{k+1}, \dots, X_n)$ are independent.

(ii) *For any functions $f_1(x_1), \dots, f_n(x_n)$ the random variables*

$$f_1(X_1), \dots, f_n(X_n)$$

are independent.

□

Proposition 3.39 (The law of the subconscious statistician)**.** *Suppose that $X_1, \dots, X_n$ are discrete random variables with ranges $\mathscr{X}_1, \dots, \mathscr{X}_n$ and joint pmf $p(x_1, \dots, x_n)$.*

(i) *If $f : \mathscr{X}_1 \times \cdots \times \mathscr{X}_n \to \mathbb{R}$ is a function, then*

$$\mathbb{E}\big[\, f(X_1, \dots, X_n) \,\big] = \sum_{(x_1, \dots, x_n) \in \mathscr{X}_1 \times \cdots \times \mathscr{X}_n} f(x_1, \dots, x_n) p(x_1, \dots, x_n).$$

(ii) *If the random variables $X_1, \dots, X_n$ are independent, then for any functions $f_k : \mathscr{X}_k \to \mathbb{R}$, $k = 1, \dots, n$, the random variables*

$$f_1(X_1), \dots, f_n(X_n)$$

are independent and we have

$$\mathbb{E}\left[\prod_{k=1}^{n} f(X_k)\right] = \prod_{k=1}^{n} \mathbb{E}\, f(X_k)\,]. \tag{3.20}$$

□

3.4. Exercises

Exercise 3.1. Let (X, Y) be uniform on the four points $(0,0), (1,0), (1,1), (2,1)$.

(i) Find the marginal pmfs of X and Y.

(ii) For which joint pmf of (X, Y) are X and Y uniform on their respective ranges?

Exercise 3.2. Is it true in general that $p(x, y) \leq p_X(x)$ for all x, y?

Exercise 3.3. Consider a family with three children. Let X be the number of daughters and Y the number of sons. Find the joint pmf of (X, Y).

Exercise 3.4. Suppose you roll a die until you get a 6 and record the number N of rolls it took. Then flip a coin N times. Denote by X the number of heads and by Y the number of tails.

(i) Find the joint pmf of (X, Y).

(ii) Find the marginal distributions p_X of X and p_Y of Y.

(iii) Are the random variables X, Y independent?

Hint. For (ii) you need to use the fact that for any $x \in (-1, 1)$ we have

$$\begin{aligned}\left(\frac{1}{1-x}\right)^{k+1} &= \frac{1}{k!}\frac{d^k}{dx^k}\left(\frac{1}{1-x}\right) = \frac{1}{k!}\frac{d^k}{dx^k}\left(1 + x + x^2 + \cdots\right) \\ &= \binom{k+0}{k}x^0 + \binom{k+1}{k}x^1 + \binom{k+2}{k}x^2 + \cdots\end{aligned} \tag{3.21}$$

Exercise 3.5 (The two-envelope paradox)**.** One envelope contains b dollars the other $2b$ dollars. The amount $b > 0$ is *unknown*. A player selects an envelope at random and she opens it. Let X be the amount she observed in this envelope, and Y the amount in the, yet unopened, envelope. Adopt the strategy of switching to the unopened envelope with probability

$$p(x) = \frac{e^{-x}}{e^{-x} + e^{x}} = \frac{1}{1 + e^{2x}}.$$

(For example, if the player observes that in the envelope she chose there are \$10, then she'll switch envelope with probability $p(10)$.) Denote by Z the amount the player receives by adopting this random switch strategy.

(i) Show that

$$\mathbb{E}[X] = \mathbb{E}[Y] = \frac{3b}{2}.$$

(ii) Show that

$$\mathbb{E}\left[\frac{Y}{X}\right] = \frac{5}{4}(> 1).$$

(iii) Show that $\mathbb{E}[Z] > \mathbb{E}[X]$.

Remark. Note the surprising contradictory conclusions. Part (i) shows that, on average, if the player does not switch she will make the same amount of money as if she switched automatically upon opening the envelope she picked. Part (ii) shows that, on average, the ratio between the amount of money in the the unopened envelope to the amount of money in the open envelope is > 1. Part (iii) shows that if she adopts the random switching strategy then, on average, she will make more money than by adopting the no-switch strategy or the automatic switch strategy!!!

Exercise 3.6. A system consists of four components which function independently with probabilities $0.9, 0.8, 0.6$ and respectively 0.6. Let X denote the number of components that work.

(i) Find $\mathbb{E}[X]$.

(ii) Find $\boldsymbol{var}[X]$.

(iii) Find $\mathbb{P}(X > 0)$.

(iv) Find $\mathbb{P}[X = 1]$.

Exercise 3.7. Suppose that n man-woman couples go to a meeting. The individuals (men and women) are randomly placed at a round table.

(i) For $k = 1, 2, \dots, n$ find the probability the k-th woman sits near her partner.

(ii) Denote by N the number of couples that have neighboring seats. Find $\mathbb{E}[N]$.

Hint. For (ii) use the linearity of the expectation (3.4).

Exercise 3.8. Suppose that n birds are arranged in a circle, $n \geq 3$. At a given moment each bird turns at random to the left or right, with equal probabilities and pecks the corresponding neighboring bird. Find the expected number of unpecked birds.

Exercise 3.9. (a) Suppose you are performing the following random experiment: you flip a fair coin until you get a head. Record the number F of flips it took, and then roll a fair die F times and record the number N of 6-s you obtain. Find the expectation $\mathbb{E}[N]$ of N.

(b) Suppose you perform another experiment: you roll a fair die until you get a 6. Record the number R of rolls, and then flip a fair coin R times, and record the number N of heads you get. Find the expectation $\mathbb{E}[N]$ of N.

Hint. Use (3.13).

Exercise 3.10. Suppose that a player gambles according to the following strategy at a game of coin tossing: he always bets "tail"; if "head" occurs he doubles his stake in the next coin toss. He plays until "tail" occurs for the first time. What is his expected gain?

Exercise 3.11. Let X and Y be independent and have the same geometric distribution with success probability p. Find the conditional distribution of X given $X + Y = n$. Explain intuitively.

Exercise 3.12. Suppose that N_1 and N_2 are independent and Poisson distributed with parameters λ_1 and respectively λ_2. Let $N := N_1 + N_2$.

(i) Show that $N \sim \mathrm{Poi}(\lambda_1 + \lambda_2)$.

(ii) Show that $N_1 | N = n \sim \mathrm{Bin}\left(n, \frac{\lambda_1}{\lambda_1+\lambda_2}\right)$.

Hint (i) At some point you will need to invoke Newton's binomial formula (1.6).

Exercise 3.13. Flip a fair coin repeatedly and wait for the first occurrence of 3 consecutive heads, HHH. Find the expected number of flips until this occurs.

Hint. Imitate the strategy in Example 3.32.

Exercise 3.14. Draw three cards without replacement from a deck of cards. Let H be the number of hearts and S the number of spades drawn. Find $\rho(H, S)$.

Exercise 3.15. Let A and B be two events. The degree of dependence between A and B can then be measured by the correlation between their indicators I_A and I_B. Suppose $\mathbb{P}(A) = \mathbb{P}(B) = \frac{1}{2}$. Express the correlation coefficient $\rho(I_A, I_B)$ in terms of $\mathbb{P}(A|B)$.

Exercise 3.16. Two fair dice are rolled. Find the joint probability mass function of X and Y when

(i) X is the largest value obtained on any die and Y is the sum of the values;

(ii) X is the value on the first die and Y is the larger of the two values;

(iii) X is the smallest and Y is the largest value obtained on the dice.

Exercise 3.17. Suppose that 2 balls are successive chosen without replacement from an urn consisting of 5 white and 8 red balls. Let X_i equal 1 if the i-th ball selected is white, and let it equal 0 otherwise. Find the joint probability mass function of (X_1, X_2).

Exercise 3.18. Repeat Exercise 3.17 when the ball selected is replaced in the urn before the next selection.

Exercise 3.19. Suppose we draw (without replacement) two tickets from a hat that contains tickets numbered $1, 2, 3, 4$. Let X be the first number drawn and Y be the second. Find the joint distribution of X and Y.

Exercise 3.20. Consider a sequence of independent Bernoulli trials, each of which is a success with probability p. Let X_1 be the number of failures preceding the first success, and let X_2 be the number of failures between the first two successes. Find the joint probability mass function of X_1 and X_2.

Exercise 3.21. A company consists of $m = 24$ men and $w = 30$ women. Each of the employee is to be randomly promoted with probability 1/3 independently of the other employees.

(i) Find the expected number of women that will be promoted given that the total number of employees promoted was 15.

(ii) Find the expected number of women that will be promoted.

Exercise 3.22. A electronics store owner figures that 45% of the customers entering his store will purchase computers, 15% percent will purchase a smart TV set, and 40% will just be browsing. If 5 customers enter his store on a given day, what is the probability that he will sell exactly 2 computers and 1 smart TV set on that day?

Exercise 3.23. The joint probability mass function of (X, Y) is given by

$$p(1,1) = \frac{1}{8}, \;\; p(1,2) = \frac{1}{4}, \;\; p(2,1) = \frac{1}{8}, \;\; p(2,2) = \frac{1}{2}.$$

(i) Compute the conditional mass function of X given $Y = i$, $i = 1, 2$.

(ii) Are X and Y independent?

Exercise 3.24. Choose a number X at random from the set of numbers $\{1, 2, 3, 4, 5\}$. Now choose a number at random from the subset of numbers $\leq X$, that is, from $\{1, \ldots, X\}$. Call this second number Y.

(i) Find the joint mass function of (X, Y).

(ii) Find the marginal pdf of Y.

(iii) Are X and Y independent? Why?

Exercise 3.25. Two dice are rolled. Let X and Y denote, respectively, the largest and smallest values obtained. Compute the conditional mass function of Y given $X = i$, for $i = 1, 2, \ldots, 6$. Are X and Y independent? Why?

Exercise 3.26. A fair die is successively rolled. Let X and Y denote, respectively, the number of rolls necessary to obtain a 6 and a 5. Find

(i) $\mathbb{E}[X]$.

(ii) $\mathbb{E}[X|Y = 1]$.

(iii) $\mathbb{E}[X|Y = 5]$.

Exercise 3.27. Urn 1 contains 5 white and 6 black balls, while urn 2 contains 8 white and 10 black balls. Two balls are randomly selected from urn 1 and are put into urn 2. If 3 balls are then randomly selected from urn 2, compute the expected number of white balls in the trio.

Hint. Denote by N_1 the number of white balls drawn from the first urn and by N_2 the number of white balls drawn from the 2 urn, after we have added the balls drawn form the first. We have

$$\mathbb{E}[N_2] = \sum_{k=0}^{2} \mathbb{E}[N_2|N_1 = j]\mathbb{P}(N_1 = j).$$

Exercise 3.28. (a) A monkey has a bag with four apples, three bananas, and two pears. He eats fruit at random until he takes a fruit of a kind he has eaten already. He throws that away and the bag with the rest. What is the expected the number of fruit eaten? (b) Suppose that the bag contain n of each m types of fruits, $m \leq n$. Compute the expected number of fruits the monkey eats if he follows the same procedure as above.

Exercise 3.29. The number of people who enter an elevator on the ground floor is a Poisson random variable with mean λ. If there are N floors above the ground floor, and if each person is equally likely to get off at any one of the N floors, independently of where the others get off, compute the expected number of stops that the elevator will make before discharging all of its passengers.

Hint. Condition on the number M of people that enter the elevator, $M \sim \text{Poi}(\lambda)$ and then use the result in the balls-in-bins problem, Example 3.23.

Exercise 3.30. A coin having probability p of coming up Heads is continually flipped until both heads and tails have appeared.

(i) Find the expected number of flips.

(ii) Find the probability that the last flip lands on heads.

Exercise 3.31. A biased coin having probability p of coming up Heads is flipped $n \geq 2$ times. Flip this coin 101 times. Let C_n denote the number of changes in the string of flips, i.e., flips whose outcome is different of the outcome of the previous flip.

(i) Find $\mathbb{E}[C_2]$.

(ii) Find $\mathbb{E}[C_3]$.

(iii) Find $\mathbb{E}[C_n]$.

Exercise 3.32. A group of 20 people consisting of 10 men and 10 women is randomly arranged into 10 pairs of 2 each.

(i) Compute the expectation and variance of the number of pairs that consist of a man and a woman.

(ii) Now suppose the 20 people consist of 10 married couples. Compute the mean and variance of the number of married couples that are paired together.

Exercise 3.33. The number of accidents that a person has in a given year is a Poisson random variable with mean Λ. However, suppose that the value of Λ changes from person to person, being equal to 2 for 60 percent of the population and 3 for the other 40 percent. If a person is chosen at random, what is the probability that he will have

(i) 0 accidents;

(ii) exactly 3 accidents in a certain year?

Exercise 3.34. A company puts five different types of prizes in their cereal boxes, one in each box and in equal proportions. How many boxes should you expect to buy until you collect all five prizes?

Exercise 3.35 (D. Bernoulli)**.** An urn R contains n red balls and an urn B contains n black balls. At each stage a ball is selected at random from each urn and then they are swapped. Show that the mean number of red balls in urn R at stage k is $\frac{1}{2}\big(1+(1-2/n)^k\big)$.

Exercise 3.36 (G. Polya)**.** An urn U contains r_0 red balls and g_0 red balls. At each stage a ball is selected at random from the urn, we observe its color, we return it to the urn and then we add another ball of the same color. We denote by b_n the total number of balls in U at stage n, by R_n the number of red balls and by G_n the number of green balls at stage n. Finally, we denote by C_n the "concentration" of red balls at stage n,

$$C_n = \frac{R_n}{b_n} = \frac{R_n}{R_n+G_n}.$$

(i) Show that

$$\mathbb{E}\big[\,C_{n+1}|R_n = i\,\big] = \frac{i}{b_n}.$$

(ii) Show that

$$\mathbb{E}[C_n] = \frac{r_0}{r_0+g_0},\ \ \forall n \in \mathbb{N}.$$

Chapter 4

Multivariate continuous distributions

4.1. Two-dimensional continuous random vectors

Suppose that X, Y are two random variables defined on the same sample space S. We say that X, Y are *jointly continuous* if there exists a two-variable function $p : \mathbb{R}^2 \to \mathbb{R}$ such that, for any real numbers x, y, we have

$$\mathbb{P}(X \leq x,\ Y \leq y) = \int_{-\infty}^{y}\int_{-\infty}^{x} p(s,t)dsdt.$$

One can show that the above condition is equivalent with a stronger requirement: for any subset $B \subset \mathbb{R}^2$ we have

$$\boxed{\mathbb{P}\big(\,(X,Y) \in B\,\big) = \int_B p(x,y)dxdy}.$$

The function p is called the *joint probability density* (joint pdf) of the random variables X, Y. In this case we also say that the pair (X, Y) is a *continuous random vector* or *point*.

The *joint cumulative distribution function* (joint cdf) of the random vector (X, Y) is the function

$$F_{X,Y} : \mathbb{R}^2 \to [0,1],\ \ F(x,y) = \mathbb{P}(X \leq x,\ Y \leq y) = \int_{-\infty}^{y}\int_{-\infty}^{x} p(s,t)dsdt.$$

The pdf p has a simple interpretation: the quantity $p(x,y)dxdy$ is equal to the probability that the random point (X,Y) is located in the infinitesimal rectangle $[x,x+dx]\times[y,y+dy]$, i.e.,

$$p(x,y)dxdy = \mathbb{P}\big(\, X\in[x,x+dx],\;\; Y\in[y+dy]\,\big).$$

Proposition 4.1. *Suppose that X,Y are jointly continuous random variables with joint pdf $p(x,y)$ and joint cdf F. Then X and Y are continuous random variables. Their pdf-s p_X and respectively p_Y are called the* marginal pdf-s *or the* marginals *of p and are given by*

$$\boxed{p_X(x)=\int_{\mathbb{R}} p(x,y)dy,\;\; p_Y(y)=\int_{\mathbb{R}} p(x,y)dx}. \tag{4.1}$$

Moreover

$$\boxed{p(x,y)=\frac{\partial^2}{\partial x\partial y}F_{X,Y}(x,y)}. \tag{4.2}$$

□

Let us observe that a function $f:\mathbb{R}^2\to\mathbb{R}$ is the joint pdf of a continuous random vector (X,Y), if and only if

$$f(x,y)\geq 0,\;\; \forall x,y\in\mathbb{R}, \tag{4.3a}$$

$$\int_{\mathbb{R}^2} f(x,y)dxdy = 1. \tag{4.3b}$$

When we describe a function $f(x,y)$ with the properties (4.3a, 4.3b), we are implicitly describing a continuous random vector, though we may not mention this explicitly.

Example 4.2 (Independent random variables)**.** Recall (see Definition 2.7) that the random variables X,Y are independent if and only if for any $A,B\subset\mathbb{R}$ we have

$$\mathbb{P}(X\in A, Y\in B)=\mathbb{P}(X\in A)\mathbb{P}(Y\in B).$$

If the random variables X,Y are continuous with pdf-s p_X and respectively p_Y, then they are independent if and only if the random vector (X,Y) is continuous and its joint pdf $p(x,y)$ is the product of its marginals, i.e., it satisfies the equality

$$p(x,y)=p_X(x)p_Y(y),\;\; \forall x,y\in\mathbb{R}.$$

We want to compute the cdf and pdf of their sum $S=X+Y$.

$$\begin{aligned}\mathbb{P}(S\leq s)=\mathbb{P}(X+Y\leq s) &= \int_{x+y\leq s} p_X(x)p_Y(y)dxdy\\ &= \int_{\mathbb{R}}\left(\int_{-\infty}^{s-x} p_Y(y)dy\right)p_X(x)dx\end{aligned}$$

$$= \int_{\mathbb{R}} \mathbb{P}(Y \leq s - x) p_X(x) dx.$$

If we derivate with respect to s the last equality we deduce

$$\boxed{p_{X+Y}(s) = \int_{\mathbb{R}} p_Y(s-x) p_X(x) dx}. \tag{4.4}$$

□

Example 4.3 (Uniform distribution on a planar region)**.** Suppose that $D \subset \mathbb{R}^2$ is a planar region with finite area. The uniform distribution on D is described by the density

$$p(x,y) = \frac{1}{\text{area}(D)} I_D(x,y),$$

where I_D the indicator function of D,

$$I_D : \mathbb{R}^2 \to \mathbb{R}, \;\; I_D = \begin{cases} 1, & (x,y) \in D, \\ 0, & (x,y) \notin D. \end{cases}$$

Thus, the probability that a random planar point with this distribution belongs to a region $A \subset D$ is equal to the *fraction* of the area of D occupied by A. The specific location of A in D or the shape of A play no role in this case.

For illustration suppose that D is the unit disk in the plane

$$D := \bigl\{ (x,y) \in \mathbb{R}^2; \;\; x^2 + y^2 \leq 1 \bigr\}.$$

Then area$(D) = \pi$ so

$$\frac{1}{\pi} I_D(x,y)$$

is the joint pdf of a random vector (X,Y), the random vector *uniformly distributed in* D. The marginals of this random vector are computed using (4.1). For $|x| > 1$ we have $p_X(x) = 0$, while for $|x| \leq 1$ we have

$$p_X(x) = \frac{1}{\pi} \int_{\mathbb{R}} I_D(x,y) dy = \frac{1}{\pi} \int_{|y| \leq \sqrt{1-x^2}} dy = \frac{2}{\pi} \sqrt{1-x^2}.$$

Similarly,

$$p_Y(y) = \frac{2}{\pi} \times \begin{cases} 0, & |y| > 1, \\ \sqrt{1-y^2}, & |y| \leq 1. \end{cases}$$

□

Example 4.4 (Buffon needle problem, 1777)**.** *A needle of length $L < 1$ is placed at random on a plane ruled by parallel lines at unit distance apart. What is the probability $p = p(L)$ that the needle intersects one of these lines?*

Think of the ruling lines as horizontal, one of them being the x axis. To decide the whether the needle intersects one of the lines we need to know two parameters.

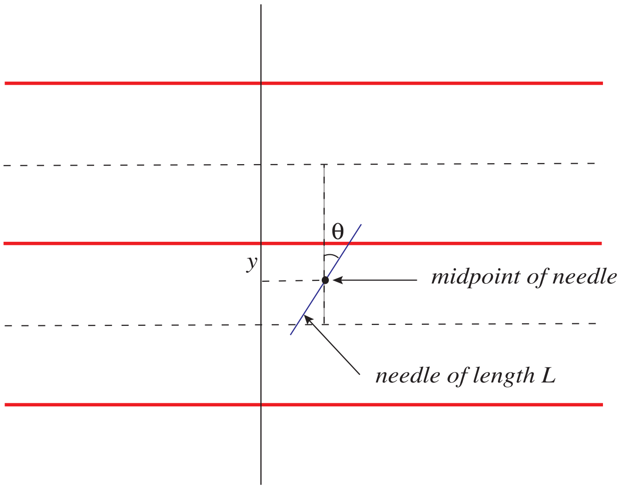

Figure 4.1. *The plane is ruled by (red) horizontal lines,* 1 *unit apart and we randomly drop a needle of length* $L<1$.

- The *signed* distance y between the midpoint of the needle and the closest line, measured vertically. Thus $y\in[-1/2,1/2]$. (The distance y is negative when the center of the needle is below the closest line.)
- The angle $\theta\in[-\pi/2,\pi/2]$ the needle makes with the vertical axis.

We interpret the randomness of the location of the needle as stating that the distance y is independent of the angle θ and they are uniformly distributed in their respective ranges. Thus (y,θ) is uniformly distributed in the rectangle

$$R=\left\{(y,\theta)\in\mathbb{R}^2;\ -\frac{1}{2}\leq y\leq\frac{1}{2},\ -\frac{\pi}{2}\leq\theta\leq\frac{\pi}{2}\right\}.$$

The joint pdf of the random vector (*signed distance, angle*) is

$$f(y,\theta)=\frac{1}{\text{Area}\,(R)}I_R(y,\theta)=\frac{1}{\pi}I_R(y,\theta),$$

where I_R is the indicator function of R.

The length of the projection of the needle on the vertical axis is $L\cos\theta$. We deduce that the needle intersects the closest line if and only of $|y|\leq\frac{1}{2}L\cos\theta$. Thus the intersection probability is

$$p(L)=\frac{1}{\pi}\iint_{|y|\leq\frac{L}{2}\cos\theta}I_R(y,\theta)dyd\theta=\frac{1}{\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\left(\int_{-\frac{L}{2}\cos\theta}^{\frac{L}{2}\cos\theta}dy\right)d\theta$$

$$=\frac{L}{\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\cos\theta d\theta=\frac{2L}{\pi}.$$

Using this formula we deduce

$$\boxed{\pi = \frac{2L}{p(L)}}.$$

If we can compute $p(L)$ by some means, then we can also compute π. This raises the real possibility of computing π by performing random experiments, for example, by tossing a needle very large number of times N. Denote by f_N the number of times the needle crosses a line. The Law of Large Numbers Theorem 6.1 shows that if N is large then

$$\frac{f_N}{N} \approx p(L) = \frac{2L}{\pi} \Rightarrow \pi \approx \frac{2LN}{f_N}.$$

Unfortunately we need to toss the needle more than one million times to get the first two decimals of π. We refer to Example 7.17 for an R-code that simulates the Buffon problem. □

Theorem 4.5 (Law of the subconscious statistician). *If (X,Y) is a continuous random vector with joint pdf $p(x,y)$ and $f(x,y)$ is a function of two variables, then*

$$\boxed{\mathbb{E}\big[\,f(X,Y)\,\big] = \iint_{\mathbb{R}^2} f(x,y)p(x,y)dxdy}. \qquad \square$$

Corollary 4.6 (Linearity of expectation). *If (X,Y) is a continuous random vector then*

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

Proof. Denote by $p(x,y)$ the joint pdf of (X,Y) and by p_X, p_Y the pdf-s of X and respectively Y. Then

$$\mathbb{E}[X+Y] = \iint_{\mathbb{R}^2}(x+y)p(x,y)dxdy = \iint_{\mathbb{R}^2} xp(x,y)dxdy + \iint_{\mathbb{R}^2} yp(x,y)dxdy$$

Observe that

$$\iint_{\mathbb{R}^2} xp(x,y)dxdy = \int_{\mathbb{R}} \underbrace{\left(\int_{\mathbb{R}} p(x,y)dy\right)}_{p_X(x)} xdx = \int_{\mathbb{R}} xp_X(x)dx = \mathbb{E}[X].$$

The equality

$$\iint_{\mathbb{R}^2} yp(x,y)dxdy = \mathbb{E}[Y]$$

is proved in a similar fashion. □

Corollary 4.7. *If the continuous random variables are independent, then for any functions $f(x)$ and $g(y)$ we have*

$$\boxed{\mathbb{E}\big[\,f(X)g(Y)\,\big] = \mathbb{E}\big[\,f(X)\,\big]\mathbb{E}\big[\,g(Y)\,\big]}.$$

Proof. Denote by p_X, p_Y the pdf-s of X and respectively Y. Then

$$\mathbb{E}\big[\, f(X)g(Y)\,\big] = \iint_{\mathbb{R}^2} f(x)g(y)p_X(x)p_Y(y)dxdy$$

(use Fubini theorem)

$$= \left(\int_{\mathbb{R}} f(x)p_X(x)dx\right)\left(\int_{\mathbb{R}} g(y)p_Y(y)dy\right) = \mathbb{E}\big[\, f(X)\,\big]\mathbb{E}\big[\, g(Y)\,\big].$$

□

Example 4.8. Suppose that T_0 and T_1 are two *independent* exponential random variables with parameters λ_0 and respectively λ_1 so that

$$\mathbb{P}(T_0 \geq t) = e^{-\lambda_0 t},\;\; \mathbb{P}(T_1 \geq t) = e^{\lambda_1 t}. \tag{4.5}$$

For concreteness, think that T_0 and T_1 are the lifetimes of two laptops, L_0 and L_1. Denote by T the first moment one of these laptops dies, i.e.,

$$T = \min(T_0, T_1).$$

Let N be the indicator random variable

$$N = \begin{cases} 0, & T = T_0, \\ 1, & T = T_1. \end{cases}$$

Thus, the first laptop to die is L_N. Let us observe that T is also an exponential variable with parameter $\lambda_0 + \lambda_1$. Indeed

$$\mathbb{P}(T \geq t) = \mathbb{P}(T_0 \geq t, T_1 \geq t)$$

(T_0 and T_1 are independent)

$$= \mathbb{P}(T_0 \geq t)\mathbb{P}(T_1 \geq t) \overset{(4.5)}{=} e^{-(\lambda_0+\lambda_1)t}.$$

Next we compute the probability that the laptop L_0 dies first, i.e., $\mathbb{P}(N = 0)$. We have

$$\mathbb{P}(N = 0, T \geq t) = \mathbb{P}(T_1 > T_0 \geq t) = \iint_{t \leq x_0 < x_1} \lambda_0 e^{-\lambda_0 x_0} \lambda_1 e^{-\lambda_1 x_1} dx_0 dx_1$$

$$= \lambda_0 \lambda_1 \int_t^\infty \left(\int_{x_0}^\infty e^{-\lambda_1 x_1} dx_1\right) e^{-\lambda_0 x_0} dx_0$$

$$= \lambda_0 \int_t^\infty e^{-(\lambda_0+\lambda_1)x_0} dx_0 = \frac{\lambda_0}{\lambda_0+\lambda_1} e^{-(\lambda_0+\lambda_1)t}$$

$$= \frac{\lambda_0}{\lambda_0+\lambda_1} \mathbb{P}(T \geq t).$$

Similarly

$$\mathbb{P}(N = 1, T \geq t) = \frac{\lambda_1}{\lambda_0+\lambda_1} \mathbb{P}(T \geq t).$$

We deduce

$$\mathbb{P}(N=0)=\mathbb{P}(N=0,T>0)=\frac{\lambda_0}{\lambda_0+\lambda_1},\ \ \mathbb{P}(N=1)=\frac{\lambda_1}{\lambda_0+\lambda_1}.$$

Note we have proved something more, namely

$$\begin{gathered}\mathbb{P}(N=0,T\geq t)=\mathbb{P}(N=0)\mathbb{P}(T\geq t),\\ \mathbb{P}(N=1,T\geq t)=\mathbb{P}(N=1)\mathbb{P}(T\geq t).\end{gathered}$$

In other words, the random variables N and T *are independent*!!! Let us explain why this is surprising.

Suppose that the two laptops have rather different expected life times, e.g.,

$$\mathbb{E}[T_0]=1,\ \ \mathbb{E}[T_1]=20.$$

Suppose that we observe that one of them dies after say 0.5 units of time, i.e. $T\in[0.5,0.5+dt]$. One might be tempted to conclude that $N=0$, i.e., the laptop that died first is the laptop L_0 since it has a shorter expected lifespan. This is however an illegitimate inference. Indeed, since T and N are independent, we cannot draw any conclusion about N from any information about T. □

Definition 4.9. Suppose that (X,Y) is a continuous random vector. Denote by μ_X, μ_Y the mean of X and respectively Y. The *covariance* of (X,Y) is the number

$$\boxed{\boldsymbol{cov}[X,Y]=\mathbb{E}\big[\,(X-\mu_X)(Y-\mu_Y)\,\big]}.$$

The *correlation coefficient* of (X,Y) is the number

$$\boxed{\rho[X,Y]=\frac{\boldsymbol{cov}[X,Y]}{\sqrt{\boldsymbol{var}[X]\cdot\boldsymbol{var}[Y]}}}. \qquad \square$$

If $p(x,y)$ is the joint pdf of the continuous random vector (X,Y) in the above definition, then

$$\begin{aligned}\boldsymbol{cov}[X,Y]&=\iint_{\mathbb{R}^2}(x-\mu_X)(y-\mu_Y)p(x,y)dxdy\\ &=\iint_{\mathbb{R}^2}xyp(x,y)dxdy-\mu_X\mu_Y=\mathbb{E}[XY]-\mu_X\mu_Y.\end{aligned}$$

Corollary 4.10. *If the continuous random variables X and Y are* <u>*independent*</u>*, then $\boldsymbol{cov}[X,Y]=0$ and*

$$\boxed{\boldsymbol{var}[X+Y]=\boldsymbol{var}[X]+\boldsymbol{var}[Y]}.$$

Proof. Denote by μ_X,μ_Y the mean of X and respectively Y. We have

$$\boldsymbol{cov}(X,Y)=\mathbb{E}[XY]-\mu_X\mu_Y=\mathbb{E}[X]\mathbb{E}[Y]-\mu_X\mu_Y=0.$$

Observe that the mean of $X+Y$ is $\mu_X+\mu_Y$. If we denote by p_X and respectively p_Y the pdf of X and respectively Y, then the jopint pdf of (X,Y) is $p_X(x)p_Y(y)$ and we have

$$\begin{gathered}\boldsymbol{var}[X+Y]=\mathbb{E}\big[\,(X+Y-\mu_X-\mu_Y)^2\,\big]\\ =\iint_{\mathbb{R}^2}\big(\,x+y-\mu_X-\mu_Y\,\big)^2p_X(x)p_Y(y)dxdy\\ =\iint_{\mathbb{R}^2}(x-\mu_X)^2p_X(x)p_Y(y)dxdy+\iint_{\mathbb{R}^2}(y-\mu_Y)^2p_X(x)p_Y(y)dxdy\\ +2\iint_{\mathbb{R}^2}(x-\mu_X)(y-\mu_Y)p_X(x)p_Y(y)dxdy\\ =\boldsymbol{var}[X]+\boldsymbol{var}[Y]+2\,\boldsymbol{cov}[X,Y]=\boldsymbol{var}[X]+\boldsymbol{var}[Y].\end{gathered}$$

□

Example 4.11 (Bivariate normal distribution)**.** Let $\sigma_x,\sigma_y,\mu_x,\mu_y,\rho$ be real constant such that $\sigma_x,\sigma_y>0$ and $\rho\in(-1,1)$. Define

$$Q(x,y)=\frac{1}{1-\rho^2}\left(\left(\frac{x-\mu_x}{\sigma_x}\right)^2-2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)+\left(\frac{y-\mu_y}{\sigma_y}\right)^2\right)$$

A continuous random vector (X,Y) with joint pdf

$$p(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}e^{-\frac{1}{2}Q(x,y)} \tag{4.6}$$

is called *normally distributed*. The pdf (4.6) is called *bivariate normal*.

A rather long and tedious computation[1] shows that

$$X\sim N(\mu_x,\sigma_x),\quad Y\sim N(\mu_y,\sigma_y),\quad \rho[X,Y]=\rho. \tag{4.7}$$

For this reason, two random variables X,Y with joint pdf (4.6) are said to be *jointly normal*. □

Example 4.12 (Transformation of random vectors)**.** Suppose that (X,Y) is a continuous random vector with joint pdf $p_{X,Y}(x,y)$, i.e.,

$$\mathbb{P}\big(\,X\in[x+dx],\quad Y\in[y,y+dy]\,\big)=p_{X,Y}(x,y)dxdy.$$

Suppose that we are given an *injective* transformation $\mathbb{R}^2\to\mathbb{R}^2$

$$(x,y)\mapsto(\,u,v\,)$$

where $u=f(x,y)$ and $v=g(x,y)$ are functions such that the Jacobian

$$J(x,y)=\det\begin{bmatrix} f'_x & f'_y\\ g'_x & g'_y\end{bmatrix}$$

is not zero. To compute the joint pdf of the random vector $(U,V)=(u(X,Y),v(X,Y))$ proceed as follows.

[1] There is a "cleaner" way of proving (4.7) but it relies on more sophisticated linear algebra.

(i) Solve for x and y the equations $u = f(x, y)$, $v = g(x, y)$. The solutions x, y can be viewed as functions of u, v, $x = x(u, v)$, $y = y(u, v)$.

(ii) Compute the Jacobian

$$J(u, v) = \det \begin{bmatrix} x'_u & x'_v \\ y'_u & y'_v. \end{bmatrix}.$$

(iii) Using the change of variables formula in double integrals we deduce that the joint pdf of the random vector $(U, V) = (u(X, Y), v(X, Y))$ satisfies

$$\boxed{p_{U,V}(u, v) = p_{X,Y}(x(u), y(u)) \cdot |J(u, v)| = p_{X,Y}(x, y) \cdot |J(x, y)|^{-1}}.$$

Suppose that X and Y positive random variables and $p(x, y)$ is the joint pdf of (X, Y). We want to compute the joint pdf of the vector $(U, V) = (X, Y/X)$.

We have $f(x, y) = x$, $g(x, y) = y/x$. Solving for x, y the equation $u = x$, $v = y/x$ we deduce $x = u$, $y = xv = uv$ and

$$J(u, v) = \det \begin{bmatrix} x'_u & x'_v \\ y'_u & y'_v. \end{bmatrix} = \det \begin{bmatrix} 1 & 0 \\ v & u \end{bmatrix} = u.$$

Thus

$$p_{U,V}(u, v) = f(x, y)|u| = f(u, uv)|u|.$$ □

4.2. Conditioning

Suppose that Y is a continuous random variable with cdf F_Y and pdf p_Y.

Definition 4.13. Let B an event with *positive* probability. The random variable $Y|B$, referred to as Y *conditioned on* B, is the random variable with cdf

$$F_{Y|B}(y) = \mathbb{P}\big(\, Y \leq y|B\,\big) = \frac{\mathbb{P}(B \cap (F_Y \leq y)\,)}{\mathbb{P}(B)}.$$

If $Y|B$ happens to be continuous, then we denote by $p_{Y|B}$ is pdf and by $\mathbb{E}[Y|B]$ its expectation. We will refer to $\mathbb{E}[Y|B]$ as the *conditional expectation of* Y *given* B. □

Example 4.14. Suppose that Y is independent of the event B i.e., the random variables Y and I_B are independent. Then

$$F_{Y|B}(y) = F_Y(y) \text{ and } \mathbb{E}[Y|B] = \mathbb{E}[Y].$$ □

Example 4.15. Suppose that B is the event $B = \{Y \geq 2\}$.

$$F_{Y|B}(y) = \mathbb{P}[Y \leq y|Y \geq 2] = \begin{cases} 0, & y < 2 \\ \frac{F_Y(y) - F_Y(2)}{1 - F_Y(2)}, & y \geq 2. \end{cases}$$

In this case we have

$$p_{Y|B}(y) = \frac{1}{1 - F_Y(2)} \times \begin{cases} 0, & y < 2 \\ p_Y(y), & y \geq 2, \end{cases}$$

$$\mathbb{E}\big[\, Y|Y \geq 2\,\big] = \frac{1}{1 - F_Y(2)} \int_2^\infty p_Y(y)dy.$$ □

Example 4.16. Suppose that the random vector (X, Y) is uniformly distributed in the region R between the parabola $6x(1-x)$ and the x-axis; see Figure 4.2,

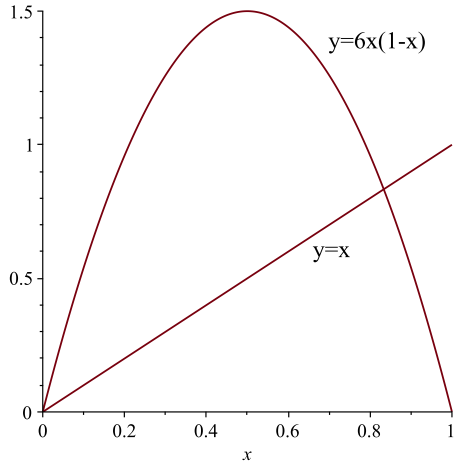

Figure 4.2. *The region R between the parabolla $y = 6x(1-x)$ and the x-axis.*

Observing that

$$\text{area}\,(R) = \int_0^1 6x(1-x)dx = 1$$

we deduce that the joint pdf of (X, Y)

$$p(x,y) = \begin{cases} 1, & 0 \leq x \leq 1,\ 0 \leq y \leq 6x(1-x), \\ 0, & \text{otherwise.} \end{cases}$$

Consider the event

$$B = \{Y > X\}.$$

We want to compute

$$\mathbb{E}[X|B] = \mathbb{E}[X|Y > X].$$

We have

$$F_{X|B}(x) = \frac{\mathbb{P}(\{X \leq x\} \cap B)}{\mathbb{P}(B)}$$

To perform this computation we observe that B correspond to the region between the line $y = x$ and the parabola $y = 6x(1-x)$; see Figure 4.2. To find where these two curves intersect we need to solve the equation

$$x = 6x(1-x) \Longleftrightarrow 6x^2 - 5x = 0 \Longleftrightarrow x = 0, \frac{5}{6}.$$

Thus if $(x,y)\in B$, then $0\leq x\leq \frac{5}{6}$. We have

$$\mathbb{P}(B)=\int_0^{\frac{5}{6}}\big(6x(1-x)-x\big)dx=\int_0^{\frac{5}{6}}(5x-6x^2)dx=\frac{125}{216}.$$

$$\mathbb{P}(\{X\leq x\}\cap B)=\int_0^x\left(\int_t^{6t(1-t)}dy\right)dt=\int_0^x(5t-6t^2)dt$$

Hence

$$p_{X|B}(x)=\frac{d}{dx}F_{X|B}(x)=\frac{1}{\mathbb{P}(B)}(5x-6x^2),\ ];0\leq x\leq\frac{5}{6}.$$

We conclude

$$\mathbb{E}[X|Y>X]=\frac{216}{125}\int_0^{\frac{5}{6}}x(5x-6x^2)dx=\frac{5}{12}.$$

□

We have the following continuous counterpart of Proposition 3.31 that generalizes the law of total probability.

Proposition 4.17. *Suppose that the events $A_1,\dots,A_n\subset S$ partition the sample space S, i.e., their union is S and they are mutually disjoint. Suppose next that $Y:S\to\mathbb{R}$ is a continuous random variable. Then*

$$\boxed{\mathbb{E}\big[Y\big]=\mathbb{E}[Y|A_1]\mathbb{P}(A_1)+\cdots+\mathbb{E}[Y|A_n]\mathbb{P}(A_n).}\tag{4.8}$$

□

Suppose that X is another random variable. We would like to condition Y on the event $X=x$. The above discussion applies only in the case when $\mathbb{P}(X=x)\neq 0$. However, if X is continuous, this positivity condition fails. Fortunately, there are things that we can do when (X,Y) is jointly continuous with joint pdf $p(x,y)$ and marginals $p_X(x)$ and respectively $p_Y(y)$.

Definition 4.18. The *conditional pdf* of Y given that $X=x$ is

$$p_{Y|X=x}(y):=\begin{cases}\frac{p(x,y)}{p_X(x)}, & p_X(x)\neq 0,\\ 0, & p_X(x)=0.\end{cases}\tag{4.9}$$

□

Intuitively, but less rigorously we have

$$p_{Y|X=x}(y)dy=\mathbb{P}\big(\,Y\in[y,y+dy]\,\big|X\in[x,x+dx]\,\big)dx$$
$$=\frac{\mathbb{P}(Y\in[y,y+dy],X\in[x,x+dx])}{\mathbb{P}(X\in[x,x+dx])}=\frac{p(x,y)dxdy}{p_X(x)dx}.$$

Observe that

$$\int_{\mathbb{R}} p_{Y|X=x}(y)dy = \int_{\mathbb{R}} \frac{p(x,y)}{p_X(x)}dy = \frac{1}{p_X(x)}\int_{\mathbb{R}} p(x,y)dy = 1,$$

so the function $y \mapsto p_{Y|X=x}(y)$ is the pdf of a continuous random variable. We denote this random variable $Y|X = x$ and we will refer to it as Y *conditioned on* $X = x$. Its expectation is

$$\boxed{\mathbb{E}\big[\,Y|X = x\,\big] = \int_{\mathbb{R}} y p_{Y|X=x}(y)dy}.$$

We will refer to the number $\mathbb{E}\big[\,Y|X = x\,\big]$ as the *conditional expectation of* Y *given that* $X = x$.

More generally, for any function $f : \mathbb{R} \to \mathbb{R}$, the conditional expectation of $f(Y)$ given that $X = x$ is the number

$$\boxed{\mathbb{E}[\,f(Y)|X = x] := \int_{\mathbb{R}} f(y)p_{Y|X=x}(y)\,dy}.$$

The next result is an immediate consequence of (4.9) and it is a continuous version of the law of total probability, Theorem 1.46.

Proposition 4.19. *For any $y \in \mathbb{R}$ we have*

$$p_Y(y) = \int_{\mathbb{R}} p_{Y|X=x}(y)p_X(x)dx.$$

Moreover, for any function f we have

$$\mathbb{E}[\,f(Y)\,] = \int_{\mathbb{R}} f(y)p_Y(y)dy = \int_{\mathbb{R}}\left(\int_{\mathbb{R}} f(y)p_{Y|X=x}(y)\,dy\right)p_X(x)dx. \tag{4.10}$$

In particular, for any $B \subset \mathbb{R}$ we have

$$\mathbb{P}(Y \in B) = \mathbb{E}[I_B(y)] = \int_{\mathbb{R}}\left(\int_B p_{Y|X=x}(y)dy\right)p_X(x)dx, \tag{4.11}$$

where we recall that I_B is the indicator function

$$I_B(y) = \begin{cases} 1, & y \in B,\\ 0, & y \notin B.\end{cases}$$

□

The equality (4.10) can be rewritten in the more compact form.

$$\boxed{\mathbb{E}\big[\,f(Y)\,\big] = \int_{\mathbb{R}} \mathbb{E}\big[\,f(Y)|X = x\,\big]p_X(x)dx}. \tag{4.12}$$

We denote by $\mathbb{E}[f(Y)|X]$ the <u>random variable</u> that takes the value $\mathbb{E}[\,f(Y)|X = x]$ when $X = x$. As in the discrete case, the random variable $\mathbb{E}[Y|X]$ is called the *conditional expectation of Y given X*. We can rewrite (4.12) in the even more compact form

$$\mathbb{E}\big[\,f(Y)\,\big] = \mathbb{E}\big[\,\mathbb{E}[f(Y)|X]\,\big].$$

Moreover, one can show that

$$\mathbb{E}[X+Y|Z] = \mathbb{E}[X|Z] + \mathbb{E}[Y|Z] \tag{4.13}$$

Example 4.20. Consider the random point (X,Y) uniformly distributed in the triangle (see Figure 4.3)

$$T = \big\{(x,y)\in\mathbb{R}^2;\;\; x,y\geq 0,\; x+y\leq 1\,\big\}.$$

The area of this triangle is $\frac{1}{2}$ and thus the joint pdf of (X,Y) is

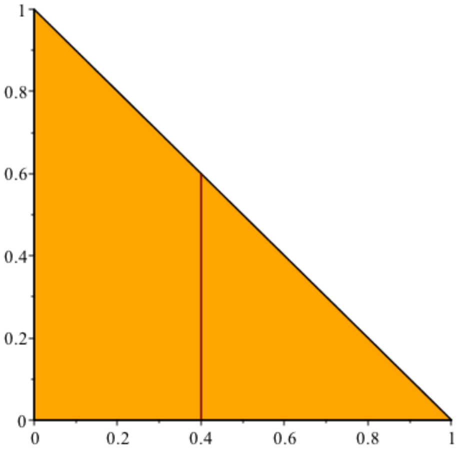

Figure 4.3. *The triangle T.*

$$p(x,y) = \begin{cases} 2, & (x,y)\in T,\\ 0, & \text{otherwise.}\end{cases}$$

The (marginal) density of X is

$$p_X(x) = \int_{\mathbb{R}} p(x,y)dy = \begin{cases} \int_0^{1-x} 2dy, & x\in[0,1],\\ 0, & \text{otherwise,}\end{cases} = 2(1-x)I_{[0,1]}(x).$$

We deduce that

$$p_{Y|X=x}(y) = \frac{p(x,y)}{p_X(x)} \sim \text{Unif}(0,1-x)$$

Then

$$\mathbb{E}[\,Y|X=x\,] = \mathbb{E}\big[\;\text{Unif}(0,1-x)\,\big] = \frac{1-x}{2}$$

and

$$\mathbb{E}[Y|X] = \frac{1}{2}(1-X). \qquad \square$$

Example 4.21. Suppose that (X, Y) is uniformly distributed in the unit disk. Then p is the uniform distribution on the unit disk as in Example 4.3

$$p(x,y) = \frac{1}{\pi} \times \begin{cases} 1, & x^2+y^2 \leq 1, \\ 0, & x^2+y^2 > 1. \end{cases}$$

Then the conditional pdf of Y given $X = x_0$ is the uniform distribution on the vertical line obtained by intersection the unit disk with the vertical line $x = x_0$, i.e.,

$$p_{Y|X=x_0} \sim \mathrm{Unif}\Big(-\sqrt{1-x_0^2},\ \sqrt{1-x_0^2}\Big),$$

for $|x_0| < 1$.

Denote by R the distance from the point (X, Y) to the origin and by Θ the angle it forms with the x-axis. More precisely

$$X = R\cos\Theta,\ \ Y = R\sin\Theta,\ \ 0 \leq \Theta \leq 2\pi,\ \ R \geq 0.$$

Let $F(r,\theta)$ denote the joint cdf of (R, Θ),

$$F(r,\theta) = \mathbb{P}(R \leq r, \Theta \leq \theta) = \frac{1}{\pi}\mathrm{area}\,\big(\,S_{r,\theta}\,\big)$$

where $S_{r,\theta}$ denotes the sector swept by a radius of length r origin rotating an angle θ about the origin. We have

$$\mathrm{area}\,(S_{r,\theta}) = \frac{r^2\theta}{2},\ \ F(r,\theta) = \frac{r^2\theta}{2\pi}.$$

The joint pdf of (R, Θ) is

$$p(r,\theta) = \frac{\partial^2}{\partial r\partial\theta}F(r,\theta) = \frac{r}{\pi}.$$

In particular,

$$p_\Theta(\theta) = \int_0^1 p(r,\theta)dr = \frac{1}{\pi}\int_0^1 rdr = \frac{1}{2\pi}$$

so that $\Theta \sim \mathrm{Unif}(0, 2\pi)$. Note that

$$p_{R|\Theta=\theta_0}(r) = \frac{p(r,\theta_0)}{p_\Theta(\theta_0)} = 2r,\ \ r \in [0,1],\ \ \mathbb{E}[R|\Theta=\theta_0] = \int_0^1 2r^2dr = \frac{2}{3}.$$

Thus, given that the random point (X, Y) is located on a given ray, the distance to the origin is not uniformly distributed on $(0, 1)$: the point is more likely to be closer to the boundary of the disk than to its center. □

Example 4.22. Suppose that X, Y are two independent identically distributed random variables with common pdf $p(x)$. We set $Z = X + Y$. We want to compute $\mathbb{E}[X|Z = z]$ and $\mathbb{E}[X|Z]$. We deduce from (4.13) that

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

Since X, Y are identically distributed, independent and $Z = X + Y$ is symmetric in X and Y we expect that $\mathbb{E}[X|Z] = \mathbb{E}[Y|Z]$ so we suspect that

$$\mathbb{E}[X|Z] = \mathbb{E}[Y|Z] = \frac{1}{2}Z.$$

Let us verify this "suspicion". Denote by $\rho(x, y)$ the joint pdf of (X, Y) so that

$$\rho(x, y) = p(x)p(y).$$

If $F(x, z)$ denotes the joint cdf of (X, Z), then

$$F(x, z) = \mathbb{P}(X \leq x, Z \leq z) = \mathbb{P}(X \leq x, X + Y \leq z)$$

(use Fubini)

$$= \int_{-\infty}^{x} \mathbb{P}(Y \leq z - s)p_X(s)ds = \int_{-\infty}^{x} F_Y(z - s)p(s)ds.$$

If $p(x, z)$ denotes the joint pdf of (X, Z), then

$$\rho(x, z) = \frac{\partial}{\partial z}\left(\frac{\partial}{\partial x}F(x, z)\right) = \frac{\partial}{\partial z}F_Y(z - x)p(x) = p(z - x)p(x).$$

Hence

$$p_Z(z) = \int_{\mathbb{R}} p(z - x)p(x)dx.$$

Thus

$$p_{X|Z=z}(x) = \frac{p(z - x)p(x)}{p_Z(z)},$$

$$\mathbb{E}[X|Z = z] = \frac{1}{p_Z(z)} \int_{\mathbb{R}} xp(z - x)p(x)dx.$$

We have

$$\int_{\mathbb{R}} xp(z - x)p(x)dx \overset{x=t+z/2}{=} \int_{\mathbb{R}} \left(t + \frac{z}{2}\right) p(z/2 - t)p(z/2 + t)dt$$

$$= \underbrace{\int_{\mathbb{R}} tp(z/2 - t)p(z/2 + t)dt}_{I_1} + \frac{z}{2} \underbrace{\int_{\mathbb{R}} p(z/2 - t)p(z/2 + t)dt}_{I_2}.$$

Note that $I_1 = 0$ because the integrand $f(t) = tp(z/2 - t)p(z/2 + t)$ is odd in the variable t

$$f(-t) = -f(t).$$

On the other hand, if we make the change in variables $t = x - z/2$ we deduce

$$I_2 = \int_{\mathbb{R}} p(z - x)p(x)dx = \int_{\mathbb{R}} p_Z(z)dz = 1.$$

Hence

$$\mathbb{E}[X|Z=z]=\frac{z}{2},\ \ \mathbb{E}[X|Z]=\frac{1}{2}Z.$$

Let us now consider the special case when X,Y are independent exponential random variables with the same parameter λ so $p(X)=\lambda e^{-\lambda x}$.Thus

$$\rho(x,z)=\lambda^2e^{-\lambda_z}\times\begin{cases}1, & 0<x\leq z,\\ 0, & x>z\geq 0,\end{cases}$$

$$p_Z(x)=\lambda^2\int_0^z e^{-\lambda z}dx=z\lambda^2e^{-\lambda z}.$$

$$p_{X|Z=z}=\frac{1}{z}\times\begin{cases}1, & 0<x\leq z,\\ 0, & 0<z<z.\end{cases}.$$

This shows that $p_{X|Z=z}(x)\sim\text{Unif}(0,z)$. □

Example 4.23 (Gaussian regression formula)**.** Consider again the normal random vector (X,Y) in Example 4.11. In this case, $\mathbb{E}[Y|X]$ is a *linear function* of X, more precisely

$$\mathbb{E}[Y|X]=\mu_Y+\frac{\rho\sigma_Y}{\sigma_X}(\,X-\mu_X\,).$$

This shows that if X,Y are jointly normal, the best predictor $\mathbb{E}[Y|X]$ of Y based on X coincides with the best *linear* predictor described in (3.9).

For example, if X,Y are independent Gaussian variables, $X,Y\sim N(0,\sigma^2)$ and $Z=X+Y$, then (X,Z) is jointly Gaussian. Then $\mu_Z=\mu_X=0$,

$$\sigma_Z=\sqrt{\boldsymbol{var}[X+Y]}=\sqrt{\boldsymbol{var}[X]+\boldsymbol{var}[Y]}=\sqrt{2}\sigma,$$

$$\boldsymbol{cov}[X,Z]=\mathbb{E}[X,X+Y]=\mathbb{E}[X^2]=\sigma^2,\ \ \rho=\rho(X,Y)=\frac{1}{\sqrt{2}},\ \ \mathbb{E}[X|Z]=\frac{1}{2}Z.$$

□

4.3. Multi-dimensional continuous random vectors

Definition 4.24. Suppose that $X_1,\dots,X_n$ are n random variables. We say that the random vector $(X_1,\dots,X_n)$ is *continuous* if there exists a function of n variables $p(x_1,\dots,x_n)$ such that, for any $B\subset\mathbb{R}^n$ we have

$$\mathbb{P}\big(\,(X_1,\dots,X_n)\in B\,\big)=\int_B p(x_1,\dots,x_n)dx_1\cdots dx_n.$$

The function p is called the *joint pdf* of the random vector. In this case, the components $X_1,\dots,X_n$ are themselves continuous random variables and their pdf-s are called the *marginals* of the joint pdf p. □

Theorem 4.25 (Law of the subconscious statistician)**.** *Suppose that*

$$(X_1, \dots, X_n)$$

is a continuous n-dimensional random vector with joint pdf $p(x_1, \dots, x_n)$. Then for any function of n variables $f(x_1, \dots, x_n)$ we have

$$\boxed{\mathbb{E}\big[\, f(X_1, \dots, X_n)\,\big] = \int_{\mathbb{R}^n} f(x_1, \dots, x_n)p(x_1, \dots, x_n)dx_1 \cdots dx_n}. \qquad \square$$

The above result has the following very useful consequence.

Corollary 4.26. *Suppose that*

$$(X_1, \dots, X_n)$$

is a continuous n-dimensional random vector. Then

$$\boxed{\mathbb{E}\big[\, c_1X_1 + \cdots + c_nX_n \,\big] = c_1\mathbb{E}[X_1] + \cdots + c_n\mathbb{E}[X_n]}, \tag{4.14}$$

for any real constants $c_1, \dots, c_n$

Recall (see Definition 2.7) that the random variables $X_1, \dots, X_n$ are called *independent* if, for any sets $A_1, \dots, A_n \subset \mathbb{R}$, we have

$$\mathbb{P}(X_1 \in A_1, \dots X_n \in A_n) = \mathbb{P}(X_1 \in A_1) \cdots \mathbb{P}(X_n \in A_n).$$

Proposition 4.27. *(a) The continuous random variables $X_1, \dots, X_n$ with pdf-s $p_{X_1}, \dots, p_{X_n}$ are independent if and only if the random vector $(X_1, \dots, X_n)$ is continuous and its joint pdf $p(x_1, \dots, x_n)$ satisfies the equality*

$$p(x_1, \dots, x_n) = p_{X_1}(x_1)p_{X_2}(x_2) \cdots p_{X_n}(x_n).$$

(b) If the continuous random variables $X_1, \dots, X_n$ are <u>independent</u>, then

$$\boxed{\boldsymbol{var}[X_1 + \cdots + X_n] = \boldsymbol{var}[X_1] + \cdots + \boldsymbol{var}[X_n]}, \tag{4.15}$$

and, for any functions $f_1(x_1), \dots, f_n(x_n)$, the random variables

$$f_1(X_1), \dots, f_n(X_n)$$

are independent and

$$\mathbb{E}\big[\, f_1(X_1) \cdots f_n(X_n) \,\big] = \mathbb{E}\big[\, f_1(X_1)\,\big] \cdots \mathbb{E}\big[\, f_n(X_n)\,\big]. \qquad \square$$

Example 4.28. You are selling your car and receive independent consecutive bids, one bid per unit of time. The bidders do not know each other's bids and for each bid you need to decide immediately whether or not to take it. If you decline, you cannot accept the offer later. You have to decide on two strategies.

Strategy 1. Reject the first bid and then accept the next bid that is greater than the first bid.

Strategy 2. Reject the first bid and then accept the first bid that is greater than the immediately preceding one.

How long can you expect to wait in each case ? Assume that the bids $X_1, X_2, \dots, X_n, \dots,$ are iid continuous random variables. Denote by $F(x)$ the common cdf of the variables X_i and by $f(x)$ their common pdf so that

$$F(x) = \int_0^x f(s)ds.$$

Denote by N the moment you accept the bid. In both cases we have $N \geq 2$ and thus

$$\mathbb{P}(N > 0) = \mathbb{P}(N > 1) = 1$$

According to (2.22) we have

$$\mathbb{E}[N] = \mathbb{P}(N > 0) + \mathbb{P}(N > 1) + \mathbb{P}(N > 2) + \cdots .$$

Strategy 1. In this case, for $n \geq 2$, we have

$$\begin{aligned} \mathbb{P}(N > n) = \mathbb{P}(X_2, \dots, X_n < X_1) = \int_0^\infty \mathbb{P}(X_2, \dots, X_n < x_1) f(x_1) dx_1 \\ = \int_0^\infty f(x_1)dx_1 \int_{0 \leq x_1, \dots, x_n \leq x_1} f(x_2) \cdots f(x_n) dx_2 \cdots dx_n \\ = \int_0^{x_1} F(x_1)^{n-1} f(x_1) dx_1 = \frac{1}{n} F(x_1)^n \Bigg|_{x=0}^{x_1=\infty} = \frac{1}{n}. \end{aligned}$$

Hence, using the first strategy we have

$$\mathbb{E}[N] = 1 + 1 + \frac{1}{2} + \frac{1}{3} + \cdots = \infty.$$

Let us compute the expected value of the bid X_N in the special case when the bids are uniformly distributed on $[0,1]$

$$f(x) = \begin{cases} 1, & x \in [0,1] \\ 0, & \text{otherwise.} \end{cases}$$

We set

$$E_n := \{N = n\} = \{0 \leq x_2, \dots, x_{n-1} \leq x_1 \leq x_n\}, p_n := \mathbb{P}(E_n).$$

We deduce

$$\mathbb{E}[X_N] = \sum_{n=2}^\infty \mathbb{E}[X_n | E_n] \mathbb{P}(E_n).$$

We have

$$\begin{aligned} \mathbb{E}[X_N | E_n] = \frac{1}{\mathbb{P}(E_n)} \int_{E_n} x_n dx_1 \cdots dx_n \\ = \frac{1}{p_n} \int_0^\infty \left(\int_0^{x_n} \left(\int_{\substack{x_j \in [0, x_1] \\ j=2,\dots,n}} dx_2 \cdots dx_{n-1} \right) dx_1 \right) x_n dx_n \\ = \frac{1}{p_n} \int_0^\infty dx_n \left(\int_0^{x_n} x_1^{n-2} dx_1 \right) x_n dx_n \end{aligned}$$

$$= \frac{1}{(n-1)p_n}\int_0^1 x_n^n dx_n = \frac{1}{(n+1)(n-1)p_n}.$$

Hence

$$\mathbb{P}(X_n|E_n)\mathbb{P}(E_n) = \frac{1}{(n+1)(n-1)}. = \frac{1}{2}\left(\frac{1}{n-1} - \frac{1}{n+1}\right),$$

$$\mathbb{E}[X_N] = \frac{1}{2}\sum_{n\geq 2}\left(\frac{1}{n-1} - \frac{1}{n+1}\right) = \frac{5}{6} \approx 0.833.$$

Strategy 2. If $n \geq 1$, then $N > n$ if and only if you have rejected the first n bids, i.e.,

$$X_1 > \cdots > X_{n-1} > X_n.$$

The probability of this event is

$$\int_0^\infty f(x_1)dx_1 \int_0^{x_1} f(x_2)dx_2 \cdots \int_0^{x_{n-2}} f(x_{n_1})dx_{n-1} \int_0^{x_{n-1}} f(x_n)dx_n.$$

For any $t > 0$ and $n \geq 1$ we set

$$P_n(t) :=$$

$$= \int_0^t f(x_1)dx_1 \int_0^{x_1} f(x_2)dx_2 \cdots \int_0^{x_{n-2}} f(x_{n-1})dx_{n-1} \int_0^{x_{n-1}} f(x_n)dx_n$$

so that

$$\mathbb{P}(N > n) = \lim_{t\to\infty} P_n(t).$$

Note that

$$P_1(t) = \int_0^t f(x_1)dx_1 = F(t),$$

$$P_2(t) = \int_0^t f(x_1)dx_1 \int_0^{x_1} f(x_2)dx_2 = \int_0^t f(x_1)F(x_1)dx_1 = \frac{1}{2}F(t)^2,$$

since $f(x) = F'(x)$. Next observe that

$$P_3(t) = \int_0^t f(x_1)P_2(x_{n-1})dx_1 = \frac{1}{2}\int_0^t f(x_1)F(x_1)^2 dx_1 = \frac{1}{3!}F(t)^3.$$

Arguing inductively we deduce

$$P_n(t) = \frac{1}{n!}F(t)^n \Rightarrow \mathbb{P}(N > n) = \frac{1}{n!}, \;\; \forall n \geq 1.$$

We deduce

$$\mathbb{E}[N] = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \cdots = e.$$

Let us compute the expected value of the bid X_N in the special case when the bids are uniformly distributed on $[0,1]$. We set

$$S_n := \{N = n\} = \{x_1 \geq x_2 \geq \cdots \geq x_{n-1} \leq x_n\}, a_n := \mathbb{P}(S_n).$$

We deduce

$$\mathbb{E}[X_N] = \sum_{n=2}^\infty \mathbb{E}[X_n|S_n]\mathbb{P}(S_n).$$

We have

$$\mathbb{E}[X_N|S_n] = \frac{1}{q_n}\int_{S_n} x_n dx_1\cdots dx_n$$

$$= \frac{1}{q_n}\int_{1\geq x_1\geq\cdots\geq x_{n-1}\geq 0}\left(\int_{x_{n-1}}^1 x_n dx_n\right) dx_1\cdots dx_{n-1}$$

(we set $x_0 := 1$)

$$= \frac{1}{q_n}\int_0^1 dx_1\int_0^{x_1} dx_2\cdots\int_0^{x_{n-2}}\frac{1}{2}(1-x_{n-1}^2)dx_{n-1}$$

$$= \frac{1}{2q_n}\int_0^1 dx_1\int_0^{x_1} dx_2\cdots\int_0^{x_{n-2}} dx_{n-1}$$

$$-\frac{1}{2q_n}\int_0^1 dx_1\int_0^{x_1} dx_2\cdots\int_0^{x_{n-2}} x_{n-1}^2 dx_{n-1}$$

$$= \frac{1}{2q_n}\left(\frac{1}{(n-1)!}-\frac{2!}{(n+1)!}\right).$$

Thus

$$\mathbb{E}[X_N] = \frac{1}{2}\sum_{n\geq 2}\frac{1}{(n-1)!}-\sum_{n\geq 2}\frac{1}{(n+1)!} = \frac{1}{2}(e-1)-(e-2-1/2) = 2-\frac{e}{2}\approx 0.640.$$

Thus using the first strategy we have to wait for a very long time to accept a bid but the expected bid in this case is bigger than the expected bid using the second strategy when we have to a considerably shorter period of time to accept a bid. To add to the difficulty of making a decision note that, using the first strategy, the probability that the wait is > 100 bids is quantifiably small $\mathbb{P}(N > 100) = \frac{1}{N}$. □

Example 4.29 (Gaussian random vectors)**.** A continuous random vector $(X_1,\dots,X_n)$ is Gaussian if there exist a *symmetric, positive definite* $n\times n$ matrix $C = (c_{ij})_{1\leq,i,j\leq n}$ and a vector

$$\boldsymbol{\mu} = (\mu_1,\dots,\mu_n)\in\mathbb{R}^n$$

such that the joint pdf has the form

$$p(\boldsymbol{x}) = \frac{1}{\sqrt{\det 2\pi C}}e^{-\frac{1}{2}Q_C(\boldsymbol{x}-\boldsymbol{\mu})},$$

where $\boldsymbol{x} = (x_1,\dots,x_n)$,

$$Q_C(\boldsymbol{y}) = \left\langle C^{-1}\boldsymbol{y},\boldsymbol{y}\right\rangle,\ \ \forall\boldsymbol{y}\in\mathbb{R}^n,$$

$\langle -,-\rangle$ inner product in $\mathbb{R}^n$. The matrix C has a simple statistical interpretation, more precisely

$$c_{ij} = \boldsymbol{cov}[X_i,X_j],\ \ \forall i,j.$$

The component X_i is a normal variable, $X_i\sim N(\mu_i,c_{ii})$. □

Example 4.30 (Linear transformations of random vectors)**.** Suppose that $\boldsymbol{X} = (X_1,\dots,X_n)$ is a continuous random vectors with joind pdf $p(x_1,\dots,x_n)$ and $A:\mathbb{R}^n\to\mathbb{R}^n$ is an *invertible* linear transformation $\boldsymbol{x}\mapsto\boldsymbol{y} = A\boldsymbol{x}$,

$$\boldsymbol{x} = (x_1,\dots,x_n),\ \ \boldsymbol{y} = (y_1,\dots,y_n),$$

$$y_1 = a_{11}x_1+\cdots+a_{1n}x_n,$$

$$y_2 = a_{21}x_1+\cdots+a_{2n}x_n,$$

$$\vdots$$

$$y_n = a_{n1}x_1+\cdots+a_{nn}y_n.$$

Then the random vector $\boldsymbol{Y} = A\boldsymbol{X}$ described by

$$\begin{aligned} Y_1 &= a_{11}X_1 + \cdots + a_{1n}Y_n, \\ Y_2 &= a_{21}X_1 + \cdots + a_{2n}X_n, \\ &\vdots \\ Y_n &= a_{n1}X_1 + \cdots + a_{nn}X_n, \end{aligned}$$

is a *continuous* random vector with joint pdf $q(\boldsymbol{y})$ described by

$$q(\boldsymbol{y}) = \frac{1}{|\det A|}p(\boldsymbol{x}) = \frac{1}{|\det A|}p(A^{-1}\boldsymbol{y}), \;\; \boldsymbol{y} = A\boldsymbol{x}. \qquad \square$$

4.4. Order statistics

Suppose that $X_1, \dots, X_n$ are independent continuous random variables. Because these variables are *continuous* the probability that $X_i = X_j$ for some $i \neq j$ is zero. So, almost surely, there is an unambiguous way to reorder these variables in increasing order

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}.$$

Thus, $X_{(1)}$ is the smallest and $X_{(n)}$ is the largest of the variables $X_1, \dots, X_n$. The random veector $(X_{(1)}, \dots, X_{(n)})$ is called the *order statistics* of the random vector $(X_1, \dots, X_n)$. Although the variables $X_1, \dots, X_n$ are independent, the variables $X_{(1)}, \dots, X_{(n)}$ are obviously not.

Suppose additionally that the random variables $X_1, \dots, X_n$ are identically distributed, $f(x)$ is their common pdf and $F(x)$ is their common cdf. The joint pdf of the random vector $\big(X_{(1)}, \dots, X_{(n)}\big)$ is

$$p(x_1, \dots, x_n) = n! \times \begin{cases} f(x_1) \cdots f(x_n), & \text{if } x_1 \leq x_2 \leq \cdots \leq x_n, \\ 0, & \text{otherwise.} \end{cases}$$

To understand this formula, note that given $x_1 < \cdots < x_n$, then

$$(X_{(1)}, \dots, X_{(n)}) = (x_1, \dots, x_n)$$

if and only of there exists a permutation ϕ of $1, \dots, n$ such that

$$X_1 = x_{\phi(1)}, \dots, X_n = x_{\phi(n)}.$$

There are $n!$ such permutations ϕ and each of the $n!$ possibilities

$$\big(x_{\phi(1)}, \dots, x_{\phi(n)}\big)$$

is equally likely to occur because the random variables $X_1, \dots, X_n$ are identically distributed.

Denote by $F_{(j)}(x)$ the cdf of $X_{(j)}$,

$$F_{(j)}(x) = \mathbb{P}\big(X_{(j)} \leq x\big).$$

To compute $F_{(j)}(x)$ we consider the *independent events*

$$A_1 = \{X_1 \leq x\},\ A_2 = \{X_2 \leq x\}, \dots, A_n = \{X_n \leq x\}.$$

The indicator functions I_{A_k} are independent Bernoulli random variables with success probability $F(x)$. Hence,

$$Y = I_{A_1} + \cdots + I_{A_n} \sim \operatorname{Bin}(n,p),\ \ p = F(x).$$

Note that $X_{(j)} \leq x$ if and only if at least j of the variables $X_1, \dots, X_n$ are $\leq x$ or, equivalently, $Y \geq j$. Hence

$$F_{(j)}(x) = \sum_{k=j}^{n} \binom{n}{k} F(x)^k \big(\, 1 - F(x)\,\big)^{n-k}. \tag{4.16}$$

From the equality (2.59) we deduce

$$F_{(j)}(x) = B_{j,n+1-j}\big(\, F(x)\,\big), \tag{4.17}$$

where $B_{a,b}(x)$ denotes the incomplete Beta function defined in (2.55). In particular,

$$\boxed{F_{(1)}(x) = 1 - \big(\, 1 - F(x)\,\big)^n,\ \ F_{(n)}(x) = F(x)^n}. \tag{4.18}$$

Example 4.31. Suppose that $X_1, \dots, X_n$ are independent random variables uniformly distributed on $[0,1]$. In this case $F(x) = x$ so

$$F_{(j)}(x) = B_{j,n+1-j}(x)$$

Thus, $X_{(j)} \sim \operatorname{Beta}(j, n+1-j)$. In particular

$$\mathbb{E}[X_{(j)}] = \frac{j}{n+1},\ \ \forall j = 1, \dots, n.$$

More generally if, $X_1, \dots, X_n \sim \operatorname{Unif}(0,L)$ then

$$\frac{1}{L} X_{(j)} \sim \operatorname{Beta}(j, n+1-j),\ \ \mathbb{E}[X_{(j)}] = \frac{j}{n+1} L.$$

□

Example 4.32. Suppose that $X_1, \dots, X_n$ are independent exponential random variables with the same rate λ. In this case

$$1 - F(x) = e^{-\lambda x}$$

and

$$F_{X_{(1)}}(x) = 1 - (e^{-\lambda x})^n = 1 - e^{-n\lambda x}$$

so that $X_{(1)} \sim \operatorname{Exp}(n\lambda)$. Suppose for example that an institution purchases n laptops, where n is assumed to be large (say 1000) and the liftemes of the computers are independent exponential random variables with the same parameter λ, say $\lambda = 1$. Then each of them is expected to last $\mathbb{E}[\,\operatorname{Exp}(1)\,] = 1$ unit of time. $X_{(1)}$ is the waiting time until the first the laptop breaks. This is an exponential random variable with rate $n = 1000$. The expected time when the first laptop will break is then

$$\mathbb{E}[X_{(1)}] = \frac{1}{1000}.$$

This is very small! On the other hand the probability that the first computer to die will do so after the expected time is rather large e^{-1}. This indicates that, although each laptop is expected to last one unit of time, we expect the first laptop to die in a rather short period of time. However, since the standard deviation is of the same size as the mean, large deviations from the mean are probable. □

Example 4.33 (Half-life revisited)**.** An institution has purchased a large number N of computers. For simplicity we assume that N is even, $N = 2n$. The lifetimes of the N computers are i.i.d. exponential random variables $T_1, \dots, T_N \sim \mathrm{Exp}(\lambda)$. Consider the order statistics of this collection,

$$T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(N)}.$$

Thus $T_{(1)}$ is the waiting time until the first computer dies, and $T_{(n)} = T_{(N/2)}$ is the waiting time until half the computers die. In (2.44) of Example 2.68 we defined the half-life of $\mathrm{Exp}(\lambda)$ to be

$$h(\lambda) = \frac{\ln 2}{\lambda}$$

Then

$$\lim_{N\to\infty} \mathbb{E}\big[\, T_{(N/2)} \,\big] = h(\lambda). \tag{4.19}$$

We describe below two proofs of (4.19). We denote by $F(t)$ the common cdf of $T_1, \dots, T_N$,

$$F(t) = 1 - e^{-\lambda t}.$$

Denote by $F_n(t)$ the cdf of $T_{(n)}$. We deduce from (4.16) that

$$F_n(t) = \sum_{k=n}^{2n} \binom{2n}{k} (1 - e^{-\lambda t})^k e^{-(2n-k)\lambda t}.$$

Note that

$$\mathbb{P}(T_{(n)} > t) = 1 - F_n(t).$$

From the equality

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

we deduce

$$1 - F_n(t) = \sum_{k=0}^{n-1} \binom{2n}{k} (1 - e^{-\lambda t})^k e^{-(2n-k)\lambda t}.$$

From Proposition 2.61 we deduce

$$\mathbb{E}[T_{(n]} = \int_0^\infty \mathbb{P}(T_{(n)} > t) dt = \sum_{k=0}^{n-1} \binom{2n}{k} \int_0^\infty (1 - e^{-\lambda t})^k e^{-(2n-k)\lambda t}.$$

We now make the change in variables $x = e^{-\lambda t}$ so $t = -\frac{1}{\lambda}\ln x$ and we deduce

$$\mathbb{E}[T_{(n)}] = \sum_{k=0}^{n-1}\binom{2n}{k}\frac{1}{\lambda}\int_0^1 (1-x)^k x^{2n-k}\frac{dx}{x}$$

$$= \frac{1}{\lambda}\sum_{k=0}^{n-1}\binom{2n}{k}\int_0^1 (1-x)^k x^{2n-k-1}dx \overset{(2.47)}{=} \frac{1}{\lambda}\sum_{k=0}^{n-1}\binom{2n}{k}\frac{\Gamma(k+1)\Gamma(2n-k)}{\Gamma(2n+1)}$$

$$\overset{(2.46)}{=} \frac{1}{\lambda}\sum_{k=0}^{n-1}\binom{2n}{k}\frac{(2n-k-1)!k!}{(2n)!} = \frac{1}{\lambda}\sum_{k=0}^{n-1}\frac{(2n)!}{k!(2n-k)!}\frac{(2n-k-1)!k!}{(2n)!}$$

$$= \frac{1}{\lambda}\sum_{k=0}^{n-1}\frac{1}{(2n-k)} = \frac{1}{\lambda}\left(\frac{1}{2n}+\frac{1}{2n-1}+\cdots+\frac{1}{n+1}\right).$$

Using Riemann sums one can show that

$$\lim_{n\to\infty}\left(\frac{1}{2n}+\frac{1}{2n-1}+\cdots+\frac{1}{n+1}\right) = \int_1^2 \frac{1}{x}dx = \ln 2.$$

Hence as $N = 2n \to \infty$ we have

$$\lim_{N\to\infty}\mathbb{E}[T_{(N/2)}] = \frac{\ln 2}{\lambda},$$

Recall that This proves (4.19),

We present below an alternate approach to the computation of $\mathbb{E}[T_{(n)}]$. The joint pdf of the random vector

$$\boldsymbol{T} = (T_{(1)},\dots,T_{2N)})$$

$$p(t_1,\dots,t_N) = N!\times\begin{cases}\lambda^N e^{\lambda(t_1+\cdots+t_N)}, & \text{if } t_1\leq t_2\leq\cdots\leq t_N,\\ 0, & \text{otherwise.}\end{cases}$$

Form the new random vector $\boldsymbol{Y} = (Y_1, Y_2,\dots,Y_N)$ defined by

$$Y_1 = NT_{(1)},\;\; Y_2 = (N-1)\big(T_{(2)}-T_{(1)}\big),\cdots$$

$$Y_{k+1} = (N-k)\big(T_{(k+1)}-T_{(k)}\big),\cdots, Y_N = \big(T_{(N)}-T_{(N-1)}\big)$$

The random vector $\boldsymbol{Y}$ is obtained from the random vector $\boldsymbol{T}$ via the linear transformation $\boldsymbol{Y} = A\boldsymbol{T}$ described by

$$y_1 = Nt_1$$

$$y_2 = -(N-1)t_1 + (N-1)t_2$$

$$y_3 = -(N-3)t_2 + (N-1)t_3$$

The matrix describing A is lower triangular and has determinant $\det A = N!$. Thus A is invertible. Its inverse is found by observing that

$$T_{(k+1)} - T_k = \frac{1}{N-k}Y_k$$

so that

$$t_1 = \frac{1}{N}y_1,\;\; t_1 = \frac{1}{N}y_1 + \frac{1}{N-1}y_2,\cdots$$

$$t_{k+1} = \frac{1}{N-k}y_{k+1}+\cdots+\frac{1}{N}y_1,\;\; k = 0,1,\dots,N-1.$$

We observe next that

$$t_1+\cdots+t_N = \frac{1}{N}y_1 + \left(\frac{1}{N}y_1+\frac{1}{N-1}y_2\right)+\left(\frac{1}{N}y_1+\frac{1}{N-1}y_2+\frac{1}{N-2}y_3\right)+\cdots$$

$$= y_1 + y_1 + \cdots + y_N.$$

Since $t_1\leq t_2\leq\cdots\leq t_N$ if and only if $y_1, y_2,\dots,y_N\geq 0$ we deduce from Example 4.30 that $\boldsymbol{Y}$ is a continuous random vector and its joint pdf is

$$q(y_1,\dots,y_N) = \lambda^n\begin{cases}e^{-\lambda(y_1+\cdots+y_N)}, & y_1,\dots,y_N\geq 0,\\ 0, & \text{otherwise}\end{cases}$$

$$= \begin{cases} (\lambda e^{-\lambda y_1}) \cdots (\lambda e^{-\lambda y_N}), & y_1, \dots, y_N \geq 0, \\ 0, & \text{otherwise} \end{cases}$$

This shows that the components of $Y_1, \dots, Y_N$ are independent random variables,

$$Y_k \sim \operatorname{Exp}(\lambda), \ \ \forall k = 1, \dots, N.$$

Hence

$$\mathbb{E}\big[\, T_{(k+1)} - T_{(k)} \,\big] = \frac{1}{N-k} \boldsymbol{E}[Y_k] = \frac{1}{\lambda(N-k)}.$$

Recalling that $N = 2n$, we deduce

$$\mathbb{E}\big[\, T_{(n)} \,\big] = \mathbb{E}\big[\, [T_{(1)}\big] + \mathbb{E}\big[\, T_{(2)} - T_{(1)} \,\big] + \mathbb{E}\big[\, T_{(n)} - T_{(n-1)} \,\big]$$
$$= \frac{1}{\lambda}\left(\frac{1}{2n} + \frac{1}{2n-1} + \cdots + \frac{1}{n+1} \right).$$

□

4.5. Exercises

Exercise 4.1. The joint probability density function of X and Y is given by

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

(i) Find c.

(ii) Find the marginal densities of X and Y.

(iii) Find $\mathbb{E}[X]$.

Exercise 4.2. Let X and Y be nonnegative, *independent* continuous random variables.

(i) Show that

$$\mathbb{P}(X < Y) = \int_0^\infty F_X(y) p_Y(y) dy.$$

where $p_X(x)$ and $p_Y(y)$ are the pdf-s of X and respectively Y?

(ii) What does this become if $X \sim \operatorname{Exp}(\lambda_1)$ and $Y \sim \operatorname{Exp}(\lambda_2)$?

Exercise 4.3. (a) Suppose that X_1, X_2 are two independent Gamma distributed random variables $X_1 \sim \operatorname{Gamma}(\nu_1, \lambda)$, $X_2 \sim \operatorname{Gamma}(\nu_2, \lambda)$. Show that

$$X_1 + X_2 \sim \operatorname{Gamma}(\nu_1 + \nu_2, \lambda).$$

(b) Suppose that $X_1, X_2, \dots, X_n \sim \operatorname{Exp}(\lambda)$ are independent identically distributed exponential variables. Show that $X_1 + \cdots + X_n \sim \operatorname{Gamma}(n, \lambda)$.

Hint. (a) Use (4.4) and Proposition 2.71. (b) Use (a) and argue by induction.

Exercise 4.4. Let X have a uniform distribution on $(0,1)$, and given that $X = x$, let the conditional distribution of Y be uniform on $(0, 1/x)$.

(i) Find the joint pdf $f(x,y)$ and sketch the region where it is positive.

(ii) Find $f_Y(y)$, the marginal pdf of Y, and sketch its graph.

(iii) Compute $\mathbb{P}(X > Y)$.

Exercise 4.5. Adam and Billy Bob have agreed to meet at 12:30. Assume that their arrival times are independent random variables, Adam's uniformly distributed between 12:30 and 1:00 and Billy Bob's uniformly distributed between 12:30 and 1:15.

(i) Compute the probability that Billy Bob arrives first.

(ii) Compute the probability that the one who arrives first must wait more than 10 min.

Exercise 4.6. Consider the quadratic equation $x^2+Bx+C=0$ where B and C are independent and have uniform distributions on $[-n,n]$. Find the probability that the equation has real roots. What happens as $n \to \infty$?

Exercise 4.7. Let X and Y be independent and $\sim$ Exp(1). Find

$$\mathbb{E}\big[\, e^{-(X+Y)/2} \,\big].$$

Exercise 4.8. Water flows in and out of a dam such that the daily inflow is uniform on $[0,2]$ (megaliters) and the daily outflow is uniform on $[0,1]$, independent of the inflow. Each day the surplus water (if there is any) is collected for an irrigation project. Compute the expected amount of surplus water in a given day.

Exercise 4.9. Let X and Y be independent and $\sim$ Unif$[0,1]$. Find (a) $\mathbb{E}[XY]$, (b) $\mathbb{E}[X/Y]$, (c) $\mathbb{E}[\ln(XY)]$, and (d) $\mathbb{E}[\ |Y-X|\ ]$.

Exercise 4.10. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with mean μ and variance σ^2, let $S_n = X_1 + \cdots + X_n$, and let $\tilde{X} = S_n/n$ (called the sample mean). Find $\mathbb{E}[\ \tilde{X}\ ]$ and $\boldsymbol{var}[\ \tilde{X}\ ]$.

Exercise 4.11. Let X and Y be nonnegative and have joint pdf f and let $Z = Y/X$.

(i) Express the joint pdf of (X, Z) in terms of f.

(ii) If X and Y are independent Exp(1), find the joint pdf of (X, Z) and the marginal pdf of Z.

Hint. Have a look at Example 4.12.

Exercise 4.12. Let X be a continuous random variable with pdf p and let b be a real number. Show that

$$\mathbb{E}\big[\, X | X > b \,\big] = \frac{\int_b^\infty x p(x) dx}{\mathbb{P}(X > b)}.$$

Exercise 4.13. A saleswoman working for a company sells goods worth $X \times \$1000$ per week, where X is Unif$[0,2]$. Of this, she must pay the company back up to \$800 and gets to keep the rest. Compute her expected profit

(i) in a given week, and

(ii) in a week when she makes a profit.

Hint. Use Exercise 4.12.

Exercise 4.14. Let U and V be independent and Unif$[0,1]$ and let

$$X = \min(U,V),\ \ Y := \max(U,V).$$

Find $\boldsymbol{cov}[X,Y]$ and comment on its sign.

Exercise 4.15. Let X and Y be independent and uniform on $[0,1]$. Let A be the area and C the circumference of a rectangle with sides X and Y. Find the correlation coefficient of A and C.

Exercise 4.16. Suppose that three contestants on a quiz show are each given the same question, and that each answers it correctly, independently of the others, with probability p. But the difficulty of the question is itself a random variable, so let us suppose, for the sake of illustration, that p is uniformly distributed over the interval $(0,1]$.What is the probability that exactly two of the contestants answer the question correctly?

Exercise 4.17. Let$X$ have a Poison distribution with parameter λ. Supose λ itself is random, following an exponential density with parameter θ.

(i) What is the marginal distribution of X?

(ii) Determine the conditional density for λ given $X = k$.

Exercise 4.18. Let

$$f(x,y) = \begin{cases} 24xy, & x,y \geq 0,\ \ x+y \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

(i) Show that $f(x,y)$ is the joint pdf of a continuous random vector (X,Y).

(ii) Find the marginal pdf of X.

(iii) Find $\mathbb{E}[X]$ and $\mathbb{E}[Y]$.

(iv) Find $\mathbb{E}[Y|X=x]$, $x \in (0,1)$.

Exercise 4.19. Suppose that the continuous random variables X,Y have a joint distribution $\rho(x,y)$ satisfying the symmetry condition $\rho(x,y) = \rho(y,x)$. Set $Z := X+Y$ and denote by $p_X, p_Y.p_Z$ the pdf-s of X,Y and respectively Z.

(i) Show that $p_X = p_Y$.

(ii) Show that

$$p_Z(z) = \int_{\mathbb{R}} \rho(z-x, z)dx.$$

(iii) Show that

$$\mathbb{E}[X|Z=z] = \frac{z}{2}.$$

Exercise 4.20. Pick a point uniformly random inside the unit disk in the plane centered at the origin; see Example 4.3. Denote by R the distance to the origin from this random point. Compute the pdf, the mean and the variance of the random variable R. □

Exercise 4.21. Three points X_1, X_2, X_3 are selected uniformly and independently at random in the interval $[0,1]$. What is the probability that X_2 lies between X_1 and X_3?

Exercise 4.22. Let U, X, S be three random variables such that U, X are independent and $U \sim \mathrm{Unif}(0,1)$. Denote by $p_X(x)$ the pdf of X and by $p_S(s)$ the pdf of S. Suppose that there exists a positive constant a such that

$$0 \leq p_S(x) \leq a p_X(x), \quad \forall x \in \mathbb{R}.$$

Prove that

$$\mathbb{P}\big(X \leq x \big| aUp_X(X) \leq p_S(X)\big) = \int_{-\infty}^{x} p_S(s)ds.$$

Exercise 4.23. The joint density function of (X,Y) is given by

$$f(x,y) = \frac{1}{y}e^{-(y+x/y)}, \quad x>0, \quad y>0.$$

Find $\mathbb{E}[X]$, $\mathbb{E}[Y]$, $\mathbb{E}[X^2|Y=y]$, and show that $\boldsymbol{cov}[X,Y]=1$.

Exercise 4.24. Let X and Y have joint density function

$$f(x,y) = e^{-x(y+1)}, \quad x>0, \quad 0<y<e-1.$$

(i) Find and describe the conditional distribution of X given $Y=y$.

(ii) Find $\mathbb{E}[X|Y=y]$ and $\mathbb{E}[X|Y]$.

(iii) Find $\mathbb{E}[X]$.

Exercise 4.25. Let $X \sim \mathrm{Unif}(0,1)$. Suppose that Y is a random variable such that $Y|X=x \sim \mathrm{Unif}(0,x)$.

(i) Find the joint pdf of (X,Y).

(ii) Find $\mathbb{P}(Y<1/4)$ by conditioning on X.

(iii) Find $\mathbb{P}(Y<1/4)$ by using the marginal density of Y.

Exercise 4.26. Let $X_1, \ldots, X_n$ be independent exponential random variables having a common parameter λ. Determine the distribution of the random variable $Y = \min(X_1, \ldots, X_n)$.

Hint. Note that $\mathbb{P}(Y > y) = \mathbb{P}(X_1 > y, \ldots, X_n > y)$.

Exercise 4.27 (A, Rényi, [16])**.** Suppose that there are 9 barbers working in a hair salon. One hair cut takes 10 min. At a given moment t a new customer enters the salon and observes that all the 9 barbers are busy and 3 more customers are waiting. The 9 busy barbers will complete their jobs at moments of time $T_1, \ldots, T_n \in [t, t+10]$. Assuming that $T_1, \ldots, T_9$ are independent and uniformly distributed in $[t, t+10]$ find the expected waiting time until his turn comes.

Exercise 4.28. Let $X_1, \ldots, X_n$ be independent and identically distributed random variables having cumulative distribution function F and density f . The quantity

$$M = \frac{1}{2}\big[\, X_{(1)} + X_{(n)} \,\big],$$

defined to be the average of the smallest and largest values in $X_1, \ldots, X_n$, is called the *midrange of the sequence*. Show that its cumulative distribution function is

$$F_M(m) = n \int_{-\infty}^{m} \big[\, F(2m - x) - F(x) \,\big]^{n-1} f(x) dx.$$

Exercise 4.29. Two points X and Y are chosen uniformly random and independently from the segment $[0, 1]$. Given $\ell \in (0, 1)$, what is the probability that $|X - Y| < \ell$?

Exercise* 4.30. The weights of n items are $X_1, \ldots, X_n$, assumed independent $\mathrm{Unif}(0, 1)$. Mary and John each have a bin (or suitcase) which can carry total weight 1. Mary likes to pack in her bin only the heaviest item. John likes to pack the items in order $1, 2, \ldots, n$, packing each item if it can fit in the space remaining. Denote by W_M the weight of Mary's suitcase and by W_J the weight of John's suitcase. Find the pdf-s of W_M and W_J and then compute the expectations of these random variables.

Hint. You need to know that, for any $c > 0$ and any positive integer k we have

$$\int_{\substack{x_1, \ldots, x_k \geq 0 \\ x_1 + \cdots + x_k \leq c}} = \frac{c^k}{k!}.$$

Chapter 5

Generating functions

5.1. The probability generating function

Recall (see Definition 2.26) that if X is a discrete random variable with range $\mathscr{X}$ contained in $\mathbb{N}_0 = \{0, 1, 2, \dots\}$ and

$$p_n = \mathbb{P}(X = n), \;\; n = 0, 1, 2, , \dots$$

then its probability generating function (pgf) is

$$G_X(s) = p_0 + p_1 s + p_2 s^2 + \cdots, \;\; s \in [0, 1].$$

According to (2.30), we have

$$G_X(s) = \mathbb{E}\big[\, s^X \,\big].$$

Note that if $X_1, \dots, X_n$ are discrete random variables with ranges contained in $\mathbb{N}_0$, then the range of the sum $X_1, \dots, X_n$ is also contained in $\mathbb{N}_0$. Moreover

$$s^{X_1 + \cdots + X_n} = s^{X_1} \cdots s^{X_n}.$$

If additionally the variables $X_1, \dots, X_n$ are independent, then (3.20) implies that

$$\mathbb{E}\big[\, s^{X_1 + \cdots + X_n} \,\big] = \mathbb{E}\big[\, s^{X_1} \cdots s^{X_n} \,\big] = \mathbb{E}\big[\, s^{X_1} \,\big] \cdots \mathbb{E}\big[\, s^{X_n} \,\big].$$

We have thus proved the following result.

Proposition 5.1. *If the discrete random variables $X_1, \dots, X_n$ are independent and their ranges are contained in $\mathbb{N}_0$, then*

$$G_{X_1 + \cdots + X_n}(s) = \prod_{k=1}^{n} G_{X_k}(s). \tag{5.1}$$

□

Example 5.2. (a) Suppose that $X_1, \dots, X_n$ are independent Bernoulli variables with the same success probability p. We set $q = 1 - p$ and

$$X := X_1 + \cdots + X_n.$$

Thus X is the number of successes in a string of n independent Bernoulli trials with success probability p, i.e., $X \sim \mathrm{Bin}(n, p)$. Since

$$G_{X_k}(s) = p + qs, \ \ \forall k = 1, \dots, n,$$

we deduce from (5.1) that

$$G_{\mathrm{Bin}(n,p)} = G_X(s) = (p + qs)^n.$$

This is in perfect agreement with our earlier computation (2.13).

(b) Suppose that $X_1, \dots, X_k$ are independent geometrically distributed random variables $X_i \sim \mathrm{Geom}(p)$, $\forall i = 1, \dots, p$. The sum $T_k = X_1 + \cdots + X_k$ is the the number independent of Bernoulli trials with success probability p until we get the k successes, i.e., $T_k \sim \mathrm{NegBin}(k, p)$. As usual, set $q = 1 - p$. From (2.14) we deduce

$$G_{X_i}(s) = \frac{qs}{1 - ps}, \ \ \forall i = 1, \dots, k.$$

We deduce from deduce from (5.1) that

$$G_{\mathrm{NegBin}(k,p)}(s) = \left(\frac{qs}{1 - ps}\right)^k.$$

This is in perfect agreement with (2.15).

(c) Suppose $X_1, \dots, X_n$ are independent Poisson random variables,

$$X_k \sim \mathrm{Poi}(\lambda_k), \ \ k = 1, \dots, n.$$

From (2.17) we deduce that

$$G_{X_k}(s) = G_{\mathrm{Poi}(\lambda_k)}(s) = e^{\lambda_k(s-1)}, \ \ k = 1, \dots, n.$$

We deduce from deduce from (5.1) that

$$G_{X_1 + \cdots + X_n}(s) = \prod_{k=1}^n e^{\lambda_k(s-1)} = e^{(\lambda_1 + \cdots + \lambda_n)(s-1)} = G_{\mathrm{Poi}(\lambda_1 + \cdots + \lambda_n)}(s).$$

Hence, the sum of independent Poisson variables is also a Poisson variable. □

Theorem 5.3 (Wald's formula). *Suppose that $X_1, X_2, \dots$ are independent identically distributed (iid) discrete random variables with ranges contained in $\mathbb{N}_0$. Denote by $G_X(s)$ their common pgf. Let N be another discrete random variable independent of the the X_i, with range contained in $\mathbb{N}_0$ and with pfg $G_N(s)$. Then the pgf of*

$$S_N = X_1 + \cdots + X_N$$

is

$$G_{S_N}(s) = G_N\big(G_X(s)\big). \tag{5.2}$$

Moreover, if

$$\mu = \mathbb{E}[X_i], \;\; \sigma^2 = \boldsymbol{var}[X_i],$$

then

$$\mathbb{E}[S_N] = \mathbb{E}[N]\mu, \;\; \boldsymbol{var}[S_N] = \mathbb{E}[N]\sigma^2 + \boldsymbol{var}[N]\mu^2. \tag{5.3}$$

Proof. For any $n \in \mathbb{N}_0$ we set $p_n = \mathbb{P}(N = n)$. Note that

$$G_N(s) = \sum_{n=0}^{\infty} p_n s^n$$

and for any $n \in \mathbb{N}_0$ we have

$$G_{S_n}(s) = G_X(s)^n.$$

Using (3.13) we deduce that

$$G_{S_N}(s) = \mathbb{E}\big[\, s^{S_N}\,\big] = \sum_{n=0}^{\infty} \mathbb{E}\big[\, s^{S_N} | N = n\,\big]\mathbb{P}(N = n) = \sum_{n=0}^{\infty} p_n \mathbb{E}\big[\, s^{S_n} | N = n\,\big].$$

Since N is independent of the S_n for any n we deduce from Corollary 3.24 that

$$\mathbb{E}\big[\, s^{S_n} | N = n\,\big] = \mathbb{E}[s^{S_n}] = G_{S_n}(s).$$

Hence

$$G_{S_N}(s) = \sum_{n=0}^{\infty} p_n G_{S_n}(s) = \sum_{n=0}^{\infty} p_n G_X(s)^n = G_N\big(G_X(s)\big).$$

To prove (5.3) we use Proposition 2.40. We have

$$\mathbb{E}[S_N] = G'_{S_N}(1) = G'_N(G_X(1))G'_X(1) = G'_N(1) \cdot G'_X(1) = \mathbb{E}[N]\mu.$$

Next,

$$G''_{S_N}(s) = G''_N(G_X(s))G'_X(s)^2 + G'_N(G_X(s))G''_X(1),$$

so

$$G''_{S_N}(1) = G''_N(1)G'_X(1)^2 + G'_N(1)G''_X(1) = G''_N(1)\mu^2 + \mathbb{E}[N]G''_X(1).$$

We deduce from (2.23d) that

$$\begin{aligned}
\boldsymbol{var}[S_N] &= G''_{S_N}(1) + G'_{S_N}(1) - G'_{S_N}(1)^2 \\
&= G''_N(1)\mu^2 + \mathbb{E}[N]G''_X(1) + \mathbb{E}[N]\mu - \mathbb{E}[N]^2\mu^2 \\
&= \big(\,\boldsymbol{var}[N] + \mathbb{E}[N]^2 - \mathbb{E}[N]\,\big)\mu^2 + \mathbb{E}[N](\sigma^2 + \mu^2 - \mu) + \mathbb{E}[N]\mu - \mathbb{E}[N]^2\mu^2 \\
&= \mathbb{E}[N]\sigma^2 + \boldsymbol{var}[N]\mu^2.
\end{aligned}$$

□

Here is a simple application of Wald's formula.

Example 5.4. Suppose that customers arrive at a rural convenience store such that the number of customers in a not so busy hour has a Poisson distribution with mean 5. Each customer buys a number of lottery tickets, independent of other customers, and this number has a Poisson distribution with mean 2.

Denote by N the number of customers arriving in one hour so $N \sim \text{Poi}(5)$. Denote by T_i the number of lottery tickets bought by the customer i, so that $T_i \sim \text{Poi}(2)$, $\forall i$. The total number of tickets sold in one hour is

$$T = T_1 + \cdots + T_N$$

so that

$$G_T(s) = G_N(\text{Poi}(2)) = e^{5(G_{\text{Poi}(2)(s)}-1)} e^{5(e^{2(s-1)}-1)}.$$

We have

$$\mathbb{P}[T=0] = G_T(0) = e^{-5(1+e^{-2})} \approx 0.0034.$$

Next

$$\mathbb{E}[T] = \mathbb{E}[N]\mathbb{E}[X_i] = 10,$$

$$\boldsymbol{var}[N] = 5, \;\; \boldsymbol{var}[X_i] = 2, \;\; \boldsymbol{var}[T] = 5\,\boldsymbol{var}[X] + 5 \cdot \mathbb{E}[X]^2 = 30. \qquad \square$$

Example 5.5 (Branching processes)**.** The *branching processes* or the *Galston-Watson processs* was first introduced in 1873 by Francis Galton and was concerned with the extinction of family names in the British peerage. It was first successfully attacked in 1874 in a joint work of Galton with the Reverend Henry Watson. It since found many applications in physical and biological sciences. We follow closely the presentation in [**7**, Sec.5.4].

Suppose that a population of bacteria evolves in generations. We denote by Z_n the number of members of the n-th generation. This is a random quantity. Each bacteria in the n-th generation gives birth to new family of members of the next generation, the $(n+1)$-th generation; see Figure 5.1. The number of members of this new family is the random variable Z_{n+1}.

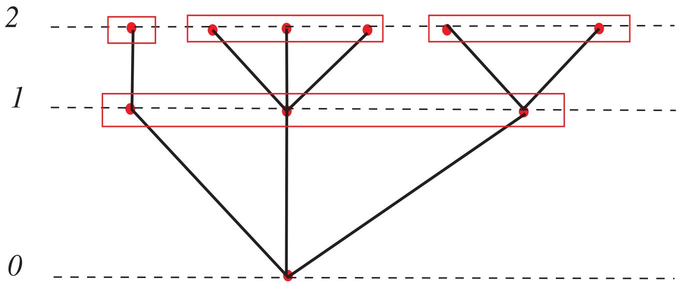

Figure 5.1. *The growth of of a population of bacteria from one generation to another*

We make the following assumptions.

(i) $Z_0 = 1$

(ii) The family sizes of each bacteria are independent random variables.

(iii) The family sizes of each bacteria have the same pmf as a fixed random variable B with values in $\{0, 1, 2, \dots\}$. We denote by p_B the pmf of B and by $G_B(s)$ its pgf

$$G_B(s) = p_0 + p_1 s + p_2 s^2 + \cdots, \;\; p_n := \mathbb{P}(B = n), \;\; n = 0, 1, 2, \dots .$$

We set

$$\mu := \mathbb{E}[B] = G_B'(1).$$

We want to understand the behavior of Z_n as $n \to \infty$ and in particular we want to study if the population can become extinct. The extinction event E happens if $Z_n = 0$ for some n. Thus

$$E = \bigcup_{n \geq 0} \{Z_n = 0\}.$$

Note that if $Z_n = 0$ then $0 = Z_{n+1} = Z_{n+2} = \cdots$ so that

$$\{Z_0 = 0\} \subset \{Z_1 = 0\} \subset \cdots \subset \{Z_n = 0\} \subset \cdots$$

so that

$$\mathbb{P}(E) = \lim_{n \to \infty} \mathbb{P}(Z_n = 0).$$

Denote by G_n the pgf of Z_n. If $\beta_{n,1}, \dots, \beta_{n,Z_n}$ are the bacteria of the n-th generation, then for any $i = 1, \dots, Z_n$, the bacterium $\beta_{n,i}$ gives birth to a number $B_{n,i}$ bacteria of the next generation. According to our assumptions, the random variables $B_{n,1}, \dots, B_{n,Z_n}$ are independent with the same pmf p_B. Thus

$$Z_{n+1} = B_{n,1} + \cdots + B_{n,Z_n}.$$

From Wald's formula (5.2) we deduce

$$G_{n+1}(s) = G_{Z_{n+1}}(s) = G_B\big(G_{Z_n}(s)\big) = G_B\big(G_n(s)\big).$$

We deduce

$$\mathbb{E}[Z_{n+1}] = G_{n+1}'(1) = G_B'\big(G_n(1\big)G_n'(1) = G_B'(1)G_n'(1) = \mu\mathbb{E}[Z_n].$$

Since $\mathbb{E}[Z_0] = 1$ we conclude

$$\mathbb{E}[Z_n] = \mu^n.$$

We deduce that if $\mu < 1$, then

$$\lim_{n \to \infty} \mathbb{E}[Z_n] = 0.$$

From Markov's inequality (2.31) we deduce

$$\mathbb{P}(Z_n > 0) = \mathbb{P}(Z_n \geq 1) \leq \mathbb{E}[Z_n] = \mu^n \to 0 \;\text{ as }\; n \to \infty.$$

Hence

$$1 \geq \mathbb{P}(Z_n = 0) = 1 - \mathbb{P}(Z_n > 0) \geq 1 - \mu^n \to 1 \;\text{ as }\; n \to \infty$$

so

$$\mathbb{P}(E) = 1.$$

This should not be too surprising: if the expected number of offsprings of a single bacterium is < 1, then in the long run this species of bacteria will become extinct. One can then expect that if $\mu > 1$, then the species will not become extinct. Indeed, the expected size of the n-th generation grows exponentially

$$\mathbb{E}[Z_n] = \mu^n \to \infty \text{ as } n \to \infty.$$

However, the probability that $Z_n = 0$ could be nonnegligible. We set

$$\rho_n := \mathbb{P}(Z_n = 0) = G_n(0). \tag{5.4}$$

Note that

$$\rho_n = \mathbb{P}(Z_n = 0) \leq \mathbb{P}(Z_{n+1} = 0) = \rho_{n+1},$$

so that the sequence (ρ_n) is increasing. It is also bounded above by 1 so it is convergent. We want to show that if $p_0 = \mathbb{P}(B = 0) > 0$ and $\mu = \mathbb{E}[B] > 1$, then the sequence converges to a number $\rho \in (0, 1)$. Clearly $\rho = \mathbb{P}(E)$. In other words, the extinction probability is nonzero but strictly smaller than 1 so the probability that the species will survive in perpetuity is also positive. To prove this we need the following technical result.

Lemma 5.6. *If $p_0 = G_B(0) > 0$ and $\mu = G_B'(1) > 1$, then the equation*

$$x = g_B(x)$$

has a unique solution ρ in the interval $(0, 1)$.

We will not give a formal proof of this fact. Instead we will provide the geometric intution behind it.

The solutions of the equation $x = g_B(x)$ correspond to the intersections of the graph pf G_B with the diagonal line $y = x$; see Figure 5.2. The intersection of the graph of G_B with the y axis is the point $(0, G_B(0)) = (0, p_0)$ which is above the diagonal. The slope of the line $y = x$ is 1. Since $\mu = G_B'(1) > 1$ and $G_B(x)$ is convex, i.e., $G_B''(x) \geq 0$, we deduce that near the corner $(1, 1)$ the graph of G_B is below the diagonal; see Figure 5.2. Thus the graph of G_B starts above the diagonal and eventually it reaches points below the diagonal. The intermediate value theorem then implies that the graph of G_B must cross the diagonal.

We can now prove the claimed facts concerning the probabilities

$$\rho_n := \mathbb{P}(Z_n = 0) = G_n(0).$$

Note that since $0 < \rho$ and G_B is increasing we have

$$p_0 = G_B(0) < G_B(\rho) = \rho,$$

$$\rho_2 = G_B(\rho_1) < G_B(\rho) = \rho.$$

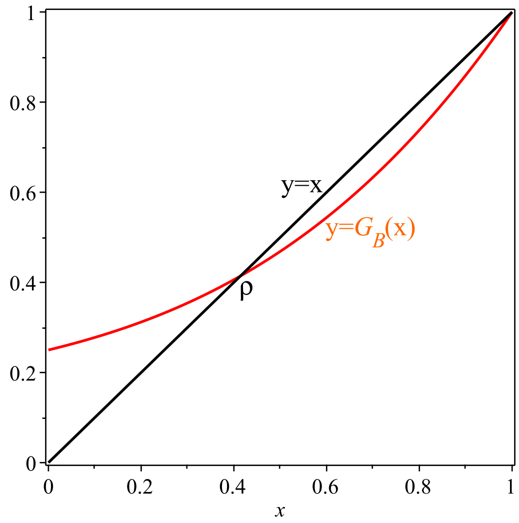

Figure 5.2. *The equation $x = G_B(x)$ has a unique solution ρ.*

Continuing in this fashion we deduce $\rho_n < \rho < 1$, for any n. Hence the sequence (ρ_n) is increasing and bounded above by ρ so it is convergent. We denote by ρ_∞ its limit. Note that $\rho_\infty \leq \rho < 1$. Letting $n \to \infty$ in the equality

$$\rho_{n+1} = G_B(\rho_n)$$

we deduce

$$\rho_\infty = G_B(\rho_\infty), \;\; \rho_\infty \in [0,1).$$

Lemma 5.6 then implies that $\rho_\infty = \rho$.

For example suppose that

$$\mathbb{P}(B=0) = \frac{1}{4}, \;\; \mathbb{P}(B=1) = \frac{3}{8}, \;\; \mathbb{P}(B=2) = \frac{3}{8}, \;\; \mathbb{P}(B>2) = 0.$$

Then

$$G_B(x) = \frac{2+3x+3x^2}{8}.$$

We have

$$p_0 = \frac{1}{4} \;\; \mu = \mathbb{E}[B] = \frac{9}{8} > 1.$$

The equation $G_B(x) = x$ is equivalent to the quadratic equation

$$3x^2 + 3x + 2 = 8x,$$

i.e.

$$3x^2 - 5x + 2 = 0.$$

Its roots are

$$x_\pm = \frac{5 \pm 1}{6}$$

and we deduce $\rho = x_- = \frac{2}{3}$. □

5.2. Moment generating function

Definition 5.7 (Moment generating function)**.** Let X be a random variable (could be either discrete, or continuous) which is s-integrable for any $s > 1$. The *moment generating function* (or mgf) of X is the function

$$M_X(t) = \sum_{k=0}^{\infty} \frac{\mu_k[X]}{k!} t^k, \tag{5.5}$$

where $\mu_k(X)$ is the k-th moment of X, $\mu_k[X] = \mathbb{E}\big[\, X^k \,\big]$. □

Remark 5.8. Even if all the the moments of X exist, it is not clear that the series in the right-hand side of (5.5) is convergent for some $t \neq 0$. In the sequel, when talking about the mgf of a random variable, we will implicitly assume that the series is convergent for some $t_0 \neq 0$. The theory of power series then implies that it is convergent for any t such that $|t| \leq |t_0|$. □

Proposition 5.9. *Let X be a random variable. Then*

$$M_X(t) = \mathbb{E}\big[\, e^{tX} \,\big]. \tag{5.6}$$

Proof. We have

$$M_X(t) = \sum_{k=0}^{\infty} \mathbb{E}\big[X^k\big] \frac{t^k}{k!} \overset{(3.4)}{=} \mathbb{E}\left[\sum_{k=0}^{\infty} \frac{(tX)^k}{k!}\right] = \mathbb{E}\big[e^{tX}\big].$$

□

The importance of the moment generating function comes from the fact that it completely determines the statistics of a random variable. More precisely we have the following nontrivial result.

Theorem 5.10. *Suppose that X and Y are two random variables. If*

$$M_X(t) = M_Y(t)$$

for all t in an open interval containing 0, then $X \sim Y$, i.e., X and Y are identically distributed, i.e.,

$$\mathbb{P}(X \leq c) = \mathbb{P}(Y \leq c) \;\; \textit{for any } c \in \mathbb{R}.$$ □

Proposition 5.11. *(a) Suppose that X is a random variable with mgf $M_X(t)$. Then*

$$\mu_k[X] = \mathbb{E}\big[X^k\big] = M_X^{(k)}(0).$$

(b) If $X_1, \dots, X_n$ are independent random variables, then

$$M_{X_1+\cdots+X_n}(t) = M_{X_1}(t) \cdots M_{X_n}(t). \tag{5.7a}$$

Proof. (a) Follows by differentiating the equality

$$M_X(t) = \sum_{k=0}^{\infty} \mathbb{E}\big[X^k\big]\frac{t^k}{k!}.$$

(b) Observe that the random variables $e^{tX_1}, \dots, e^{tX_n}$ are independent and

$$e^{t(X_1+\cdots+X_n)} = e^{tX_1}\cdots e^{tX_n}.$$

Now conclude by invoking (3.20). □

Example 5.12. (a) Suppose that $X \sim \mathrm{Ber}(p)$. The law of subconscious statistician implies

$$M_{\mathrm{Ber}(p)}(t) = \mathbb{E}\big[\,e^{tX}\,\big] = q + pe^t.$$

(b) Suppose that $X \sim \mathrm{Bin}(n,p)$. Then X is a sum of n independent identically distributed random variables $X_1, \dots, X_n \sim \mathrm{Ber}(p)$. We deduce

$$\boxed{M_{\mathrm{Bin}(n,p)}(t) = (q + pe^t)^n}.$$

(c) Suppose that $X \sim \mathrm{Geom}(p)$. As usual, set $q = 1-p$. The law of subconscious statistician implies

$$\mathbb{E}\big[\,e^{tX}\,\big] = \sum_{n=1}^{\infty} e^{tn}q^{n-1}p = pe^t\sum_{n=0}^{\infty} e^{(n-1)t}q^{n-1} = pe^t\sum_{k=0}^{\infty}\big(qe^t\big)^k = \frac{pe^t}{1-qe^t}.$$

Hence

$$\boxed{M_{\mathrm{Geom}(p)} = \frac{pe^t}{1-qe^t}}.$$

(d) Suppose that $X \sim \mathrm{NegBin}(k,p)$. Then X is a sum of k independent identically distributed random variables $X_1, \dots, X_n \sim \mathrm{Geom}(p)$. We deduce

$$\boxed{M_{\mathrm{NegBin}(k,p)} = \left(\frac{pe^t}{1-qe^t}\right)^k}.$$

(e) Suppose that $X \sim \mathrm{Poi}(\lambda)$. Then

$$\mathbb{E}\big[\,e^{tX}\,\big] = e^{-\lambda}\sum_{n=0}^{\infty} e^{tn}\frac{\lambda^n}{n!} = e^{-\lambda}e^{\lambda e^t} = e^{\lambda(e^t-1)}.$$

Hence

$$\boxed{M_{\mathrm{Poi}(\lambda)}(t) = e^{\lambda(e^t-1)}}.$$

□

Example 5.13. (a) Suppose that $X \sim \mathrm{Unif}(a,b)$. Then

$$M_X(t) = \mathbb{E}\big[\,e^{tX}\,\big] = \frac{1}{b-a}\int_a^b e^{tx}dx = \frac{e^{tb}-e^{ta}}{tb-ta}.$$

(b) Suppose that $X \sim \text{Exp}(\lambda)$. Then

$$M_X(t) = \mathbb{E}\big[\, e^{tX} \,\big] = \lambda \int_0^\infty e^{tx - \lambda x} dx = \frac{\lambda}{\lambda - t}$$

Thus

$$\boxed{X \sim \text{Exp}(\lambda) \Rightarrow M_X(t) = \frac{\lambda}{\lambda - t}}.$$

(c) Suppose that $X \sim N(0,1)$

$$M_X(t) = \mathbb{E}\big[\, e^{tX} \,\big] = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{tx - \frac{x^2}{2}} dx$$

$$= \frac{e^{t^2/2}}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-t^2/2 + tx - \frac{x^2}{2}} dx = \frac{e^{t^2/2}}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{(x-t)^2}{2}} dx = e^{\frac{t^2}{2}}.$$

Suppose that $Y \sim N(\mu, \sigma^2)$. We can write $Y = \sigma X + \mu$, $X \sim N(0,1)$. Then

$$M_Y(t) = \mathbb{E}\big[\, e^{tY} \,\big] = \mathbb{E}\big[\, e^{t(\sigma X + \mu)} \,\big] = \mathbb{E}\big[\, e^{t\mu} e^{t\sigma X} \,\big]$$

$$= e^{t\mu} \mathbb{E}\big[\, e^{t\sigma X} \,\big] = e^{t\mu} M_X(\sigma t) = e^{t\mu} \cdot e^{\sigma^2 t^2/2} = e^{\frac{\sigma^2}{2} t^2 + \mu t}.$$

Thus

$$\boxed{Y \sim N(\mu, \sigma^2) \Rightarrow M_Y(t) = e^{\frac{\sigma^2}{2} t^2 + \mu t}}.$$

(d) If $X \sim \text{Gamma}(\nu, \lambda)$, then

$$M_X(t) = \frac{\lambda^\nu}{\Gamma(\nu)} \int_0^\infty x^{\nu-1} e^{-\lambda x} e^{tx} dx = \frac{\lambda^\nu}{\Gamma(\nu)} \int_0^\infty x^{\nu-1} e^{-(\lambda - t)x} dx$$

$(y = (\lambda - t)x$, $dx = \frac{1}{(\lambda - t)} dy)$

$$= \frac{\lambda^\nu}{\Gamma(\nu)(\lambda - t)^\nu} \int_0^\infty y^{\nu-1} e^{-y} dy = \left(\frac{\lambda}{\lambda - t} \right)^\nu.$$

Thus

$$\boxed{X \sim \text{Gamma}(\nu, \lambda) \Rightarrow M_X(t) = \left(\frac{\lambda}{\lambda - t} \right)^\nu}.$$ □

From Example 5.13(b),(d) we deduce the following useful resuly.

Proposition 5.14. *Suppose that $X_1, \dots, X_n$ are i.i.d. exponential random variables*

$$X_1, \dots, X_n \sim \text{Exp}(\lambda).$$

Then

$$X_1 + \cdots + X_n \sim \text{Gamma}(n, \lambda).$$

Proof. We have

$$M_{X_1+\cdots+X_n}(t) = \left(\frac{\lambda}{\lambda - t}\right)^n .$$

Thus $X_1 + \cdots + X_n$ has the same mfg as a Gamma(n, λ)-variable. We can now conclude by invoking Theorem 5.10. □

Example 5.15 (Poisson processes)**.** Suppose that we have a have a stream of events occurring in succession at random times $S_1 \leq S_2 \leq S_3 \leq \cdots$ such that the waiting times between two successive occurrences

$$T_1 = S_1,\;\; T_2 = S_2 - S_1, \dots, T_n = S_n - S_{n-1}, \dots$$

are i.i.d. exponential random variables $T_n \sim \operatorname{Exp}(\lambda)$, $n = 1, 2, \dots$.

It may help to think of the sequence (T_n) as waiting times for a bus to arrive: once the n-th bus has left the station, the waiting time for the next bus to arrive is an exponential random variable T_{n+1} independent of the preceding waiting times. From this point of view, S_n is the arrival time of the n-th bus.

For $t > 0$ we denote by $N(t)$ the number of buses that have arrived at the station during the time interval $[0, t]$. This is a discrete random variable with range $\{0, 1, 2, 3, \dots\}$. To find its pmf we need to compute the probabilities

$$\mathbb{P}(N(t) = n),\;\; n = 0, 1, 2, \dots .$$

We have

$$\mathbb{P}(N(t) = 0) = \mathbb{P}(T_1 > t) = e^{-\lambda t} = \text{the survival function of } \operatorname{Exp}(\lambda).$$

If $n > 0$, then $N(t) = n$ if and only if the n-th buss arrived sometime during the interval $[0, t]$, i.e., $S_n \leq t$, but the $(n+1)$-th bus has not arrived in this time interval. We deduce

$$\mathbb{P}(N(t) = n) = \mathbb{P}\big(\, \{S_n \leq t\} \setminus \{S_{n+1} \leq t\}\,\big) = \mathbb{P}(S_n \leq t) - \mathbb{P}(S_{n+1} \leq t).$$

If we denote by $F_n(t)$ the cdf of S_n, then we can rewrite the above equality in the form

$$\mathbb{P}(N(t) = n) = F_n(t) - F_{n+1}(t).$$

Using Proposition 5.14 we deduce $S_{n+1} \sim$ Gamma$(n+1, \lambda)$ so that,

$$F_{n+1}(t) = \frac{\lambda^{n+1}}{\Gamma(n+1)} \int_0^t s^{\nu} e^{-\lambda s} ds = \frac{\lambda^{n+1}}{n!} \int_0^t s^n e^{-\lambda s} ds.$$

For $n > 0$, we integrate by parts to obtain

$$F_{n+1}(t) = -\left(\frac{\lambda^n}{n!} s^n e^{-\lambda s}\right)\Bigg|_{s=0}^{s=t} + \frac{\lambda^n}{(n-1)!} \int_0^t s^{n-1} e^{-\lambda s} ds$$

$$= -\frac{(t\lambda)^n}{n!} e^{-\lambda t} + F_n(t).$$

Hence

$$\mathbb{P}(N(t)=n)=F_n(t)-F_{n+1}(t)=\frac{(t\lambda)^n}{n!}e^{-\lambda t},\ \ n>0.$$

This shows that $N(t)$ is a Poisson random variable, $N(t)\sim \mathrm{Poi}(t\lambda)$.

The collection of random variables

$$\big\{\, N(t),\ \ t\geq 0\big\},$$

is called the *Poisson process* with intensity λ. □

5.3. Chernoff bounds

We have developed all the tools to needed to give you a taste of a topic of current interest. We will do this on a rather special case which exhibits the most important features of the general case.

Suppose that X is the discrete random variable with range $\{-1,1\}$ and pmf

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

Think that you own a casino where only one kind of game of chance is played. The casino wins \$1 with probability p and looses \$1 with probability q. The expected win at the end of one game is

$$w:=\mathbb{E}\big[\,X\,\big]=p-q.$$

Thus, if the casino is to stay in business, its winning probability p should be bigger than the loosing probability. We assume this to be the case so $w>0$.

Suppose that during a year a large number N of independent games are played at this casino. Denote by $X_1,\dots,X_N$ the winning at these game. The total winning for the year is

$$W=W_N=X_1+\cdots+X_N.$$

Its expectation is

$$w_N:=\mathbb{E}\big[\,W_N\,\big]=Nw.$$

The mfg of X is

$$M_X(t)=\mathbb{E}\big[\,e^{tX}\,\big]=pe^t+qe^{-t}.$$

Since the N games played during a year are independent we deduce

$$M_W(t)=\big(pe^t+qe^{-t}\big)^N.$$

Fix a number $r\in(0,1)$. (Think secretly that $r=0.99$). Let us try to estimate the probability that the total winning during a year is $<rw_N$. (Secretly, this means the winning is less than 99% of the theoretically expected profit.)

Set

$$Z:=W_N-w_N.$$

Note that $\mathbb{E}\big[\,Z\,\big]=0$ and $W_N<rw_N$ if and only if $Z<(r-1)w_N$ or, $-Z>(1-r)w_N$.

Markov's inequality (2.31) applied to the positive random variable $Y=e^{-tZ}$ implies that

$$\begin{gathered}\mathbb{P}\big(-Z>(1-r)w_N\big)=\mathbb{P}\big(e^{-tZ}>e^{t(1-r)w_N}\big)\\ =\mathbb{P}\big(Y>e^{t(1-r)Nw}\big)\leq\frac{1}{e^{t(1-r)Nw}}\mathbb{E}\big[\,Y\,\big]\\ =e^{-t(1-r)Nw}\mathbb{E}\big[\,e^{t(Nw-W_N)}\,\big]=e^{trNw}\mathbb{E}\big[\,e^{-tW_N}\,\big]\\ =e^{trNw}\big(pe^{-t}+qe^{t}\big)^N=\left(e^{trw}\big(pe^{-t}+qe^{t}\big)\right)^N,\ \ \forall t\geq 0.\end{gathered}$$

We obtain in the fashion the *Chernoff bounds*

$$\mathbb{P}\big(W_N<rNw\big)\leq\left(\underbrace{\inf_{t\geq 0}e^{trw}\big(pe^{-t}+qe^{t}\big)}_{=:c(r)}\right)^N,\ \ \forall r\in(0,1).\tag{5.8}$$

Note that we can rewrite this as

$$\mathbb{P}\big(W_N<rNw\big)\leq c(r)^N.\tag{5.9}$$

Thus, if $c(r)<1$, the quantity $c(r)^N$ goes to zero really fast showing that the probability that the yearly profit is less than a fraction r of the theoretical expectation is exponentially small, i.e., very, very small for large N.

Let us estimate $c(r)$.

Proposition 5.16.

$$\log c(r)=-\frac{1-rw}{2}\log\frac{\frac{1-rw}{2}}{q}-\frac{1+rw}{2}\log\frac{\frac{1+rw}{2}}{p}<0.\tag{5.10}$$

In particular, $c(r)<1$.

.

Proof. Define

$$h:(0,\infty)\to\mathbb{R},\ \ h(r)=\inf_{t\geq 0}f_r(t),$$

where

$$f_r(t):=\log\Big(e^{trw}\big(pe^{-t}+qe^{t}\big)\Big)=trw+\log\big(pe^{-t}+qe^{t}\big).$$

Then $c(r)=e^{h(r)}$. Note that $f_r(0)=0$ so $h(r)\leq 0$. Moreover

$$\lim_{t\to\infty}f_r(t)=\infty$$

since $p>q$. Hence f_r attains its minimum somewhere on $[0,\infty)$.

To find the minimum of f_r we use Fermat's principle. We have

$$f_r(t)=trw+\log e^{-t}\big(p+qe^{2t}\big)=t(rw-1)+\log\big(p+qe^{2t}\big),$$

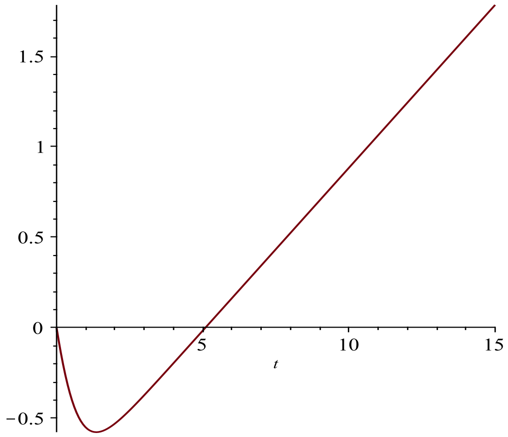

Figure 5.3. *The graph of $f_r(t)$, $p = 0.6$, $q = 0.4$, $r = 0.9$.*

$$f_r'(t) = rw - 1 + \frac{2qe^{2t}}{p + qe^{2t}} = rw - 1 + \frac{2qe^{2t}}{p + qe^{2t}}.$$

Note that since $p + q = 1$ we deduce $w = 1 - 2q$

$$f_r'(0) = rw - 1 + 2q = rw - w = w(r - 1) < 0$$

so $f_r(t) < f_r(0) = 0$, if t is sufficiently small; see Figure **??**. In particular, this shows that

$$h(r) = \inf_{t \geq 0} f_r(t) < f_r(0) = 0$$

so

$$c(r) = e^{h(r)} < 1.$$

For simplicity we set $x := e^{2t}$ so $t = \frac{1}{2}\log x$. Thus

$$f_r'(t) = rw - 1 + \frac{2qx}{p + qx}.$$

Thus $f_r'(t) = 0$ if and only if

$$2qx = (p + qx)(1 - rw) \Rightarrow qx\big(1 + rw\big) = p\big(1 - rw\big)$$

$$\Rightarrow qx = p\frac{1 - rw}{1 + rw}, \quad x = \underbrace{\frac{p}{q}\frac{1 - rw}{1 + rw}}_{=:x_*}.$$

We see that f_r has a unique critical point $t_* = \frac{1}{2}\log x_*$ in $(0, \infty)$ which therefore must be the global minimum. Then

$$h(r) = f_r(t_*) = t_*(rw - 1) + \log(p + qx_*)$$

$$= t_*(rw - 1) + \log p\left(1 + \frac{1 - rw}{1 + rw}\right)$$

$$= t_*(rw - 1) + \log\frac{2p}{1 + rw}$$

$$= \frac{rw - 1}{2}\left(\log\frac{p}{q} + \log(1 - rw) - \log(1 + rw)\right) + \log(2p) - \log(1 + rw)$$

$$= -\frac{1 - rw}{2}\log(1 - rw) - \frac{1 + rw}{2}\log(1 + rw) + \log 2$$

$$+ \frac{rw - 1}{2}\log\frac{p}{q} + \log p$$

$$\begin{aligned}
&= -\frac{1-rw}{2}\log\frac{1-rw}{2} - \frac{1+rw}{2}\log\frac{1+rw}{2} \\
&\qquad + \frac{1-rw}{2}\log q + \frac{1+rw}{2}\log p. \\
&= -\frac{1-rw}{2}\log\frac{\frac{1-rw}{2}}{q} - \frac{1+rw}{2}\log\frac{\frac{1+rw}{2}}{p}.
\end{aligned}$$

□

Example 5.17 (Casino business)**.** Let us see what Chernoff's bounds tell us about running a casino. To see the power of the inequality (**??**) we look at a special case when the probability of winning is $p = 0.55$ and the casino runs $n = 1,000,000$ games a year. With an expected profit of \$0.1 per game one can expect a profit \$100,000 per year.

We fix $r = 0.90$ and we ask what is the probability of making less than $r \cdot 100,000 = 90,000$ dollars per year. In this case we have

$$c(r) \approx 0.9999495294 \text{ and } c^{1,000,000} \approx 1.203216470 \times 10^{-22}!$$

To better appreciate how incredible this conclusion is let us mention that the chance that a killer asteroid will hit the Earth[1] in a given year is one in 100 million or 10^{-8}. The inequality (**??**) shows that it is 100 *trillion* (!!!) more likely to be hit by a killer asteroid during a given year than to make less than 90% of the theoretically predicted profit if one million games are played. This for practical purposes a sure thing with a caveat: you have to get people playing.

We include below the R script that lead to the above numerical conclusion.

```
#p is the winning probability
#q is  the losing probability
#N is the number of games   played in the casino
#r is the fraction of the theoretical mean we expect to win
chernoff<-function(p,r,N){
  q<-1-p
  w<-p-q
  u<-r*w
  h<-((1-u)/2)*log((1-u)/(2*q))+((1+u)/2)*log( (1+u)/(2*p))
  c<-exp(-h)
  Y<-r*w*N
  ch<-c^N
  cat("The probability of earning less than ", Y,
  "  dollars is smaller than ", ch, ".", sep="")
}
```

If you run the command

```
chernoff(0.55,0.9, 100000)
```

[1] See this BBC Science Focus article
https://www.sciencefocus.com/space/what-is-the-chance-of-an-asteroid-hitting-earth/

you will receive the output

```
The probability of earning less than 90000  dollars
is smaller than 1.203157e-22.
```

□

The Chernoff bounds express quantitatively a concentration-near-the-mean phenomenon which, surprisingly, is more ubiquitous than one suspects. This has many useful application computer science in the form of Monte Carlo methods, randomized algorithms, machine learning etc.

The next chapter discusses two classical examples of concentration phenomena closely related to the setup in this section. More precisely, we will discuss the laws of large numbers and central limit theorems.

5.4. Exercises

Exercise 5.1. Let X be a discrete random variable with range the positive integers and pmf $p(i) = \frac{2}{3^i}$, for $i \geq 0$. Find the moment generating function $M_X(t)$ and use this to calculate $\mathbb{E}[X]$ and $\boldsymbol{var}[X]$.

Exercise 5.2. Let X be a random variable with range $\{0, 1, 2\}$ such that

$$\mathbb{E}[X] = 1, \;\; \mathbb{E}[X^2] = \frac{3}{2}.$$

(i) Find the pgf of X.

(ii) Find $\mathbb{E}[X^5]$.

Exercise 5.3. Let X denote the number of coin tosses until the first occasion when successive tosses show HTH. Show that the probability generating function of X is

$$G_X(s) = \mathbb{E}\big[\, s^X \,\big] = \frac{s^3}{8 - 8s + 2s^2 - s^3}.$$

Exercise 5.4. Let X be a continuous random variable with range $[0, 1]$ and pdf $f(x) = 6x(1 - x)$ for $0 \leq x \leq 1$. Find the moment generating function $M_X(t)$ and use this to calculate $\mathbb{E}[X]$ and $\boldsymbol{var}[X]$.

Exercise 5.5. Let $X_1, X_2, \ldots, X_n$ be independent geometric random variables each with the same parameter p. Find the moment generating function $M_Y(t)$ of $Y = X_1 + \cdots + X_n$ and use this to find the distribution of Y.

Exercise 5.6. Let $X_1, X_2, \ldots, X_n \sim \text{Gamma}(\nu, \lambda)$ be independent Gamma distributed random variables, each with the same parameters $\nu, \lambda > 0$. Show that

$$X_1 + \cdots + X_n \sim \text{Gamma}(n\nu, \lambda).$$

Exercise 5.7. Let X, Y and Z be independent Poisson random variables with parameters λ_1, λ_2 and λ_3 respectively. For $y = 0, 1, \ldots, t$ calculate

$$\mathbb{P}(Y = y | X + Y + Z = t).$$

Exercise 5.8. Suppose that X_1, X_2 are independent normal random variables, $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$. Show that

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

Exercise 5.9. Let $X \sim N(1, 2)$ and $Y \sim N(4, 7)$ be independent random variables. Find the probability of the following events:

(a) $X + Y > 0$; (b) $X - Y < 2$; (c) $3X + 4Y > 20$.

Exercise 5.10. College student IQ is normally distributed with mean 110 and standard deviation 4. Find the probability that the average IQ of 10 randomly selected students is at least 112.

Exercise 5.11. The lifetime of car mufflers is normally distributed with mean 3 years and standard deviation 1 year. Suppose that a family buys two new cars at the same time.

(i) What is the probability that the muffler on one car needs changing at least 1 year before the muffler on the other car?

(ii) What is the probability that one car needs two new mufflers before the muffler on the other car has to be replaced?

Exercise 5.12. Suppose that $X_1, X_2, \ldots$ are independent geometric random variables with success probability p and N is a Poisson random variable with mean λ independent of the X_i. Find the probability generating function of

$$Y = X_1 + \cdots + X_N.$$

Chapter 6

Limit theorems

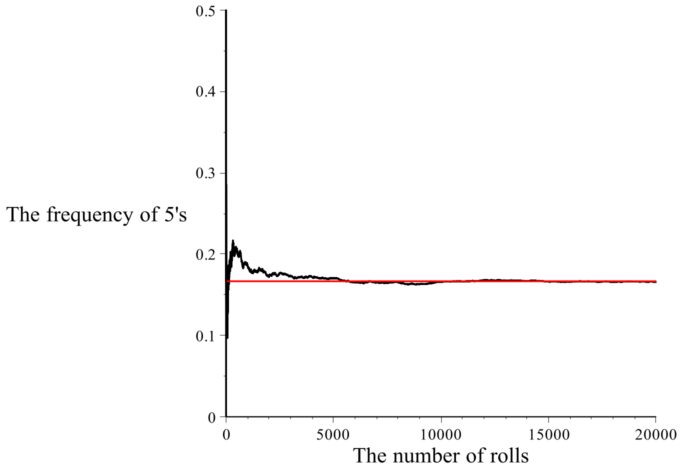

Figure 6.1. *Simulating* 20,000 *rolls of a fair die and recording the frequency of* 5*'s. The horizontal line at altitude* 1/6 *is the theoretically prescribed probability of getting* a 5.

6.1. The law of large numbers

We have indirectly alluded to the law of large numbers early on in Example 1.5(a) when we observe that if we roll a fair die a large number of times then we expect that the number 5 will show up roughly one sixth of the time; see Figure 6.1. For each $n \in \mathbb{N}$ denote by X_n the Bernoulli random variable that takes value 1 if the we get a five at the n-th roll and 0 otherwise.

Note that the the random variables $X_1, X_2, \dots$ are i.i.d. (independent and identically distributed). In particular, they all have the same mean

$$\mu = \mathbb{E}\big[\, X_k \,\big] = \frac{1}{6},\ \ \forall k = 1, 2, \dots .$$

Moreover, the sum $X_1 + \cdots + X_n$ describes the number of fives that we get after the first n rolls of the die. The average

$$\overline{X}_n = \frac{1}{n}\big(X_1 + \cdots + X_n\big)$$

represents the fraction of the first n rolls that yielded a 5 and it is usually referred to as the *sample mean* or the *empirical mean*, i.e., the average observed experimentally. It is itself a random quantity.

The numerical experiment depicted in Figure 6.1 suggests that for n large the empirical mean $\overline{X}_n$ is very close to the theoretical mean $\mu = \frac{1}{6}$. The Law of Large Numbers (or LLN) gives a precise meaning to this heuristic.

Theorem 6.1 (The Law of Large Numbers). *Suppose that $X_1, X_2, \dots$ is a sequence of i.i.d. random variables with mean μ and finite variance σ^2. We denote by $\overline{X}_n$ the empirical mean*

$$\overline{X}_n := \frac{1}{n}\big(\, X_1 + \cdots + X_n \,\big).$$

Then, for any $\varepsilon > 0$ we have

$$\boxed{\mathbb{P}\big(\, |\overline{X}_n - \mu| > \varepsilon \,\big) \leq \frac{\sigma^2}{n\varepsilon^2}}. \tag{6.1}$$

In particular

$$\boxed{\lim_{n\to\infty} \mathbb{P}\big(\, |\overline{X}_n - \mu| > \varepsilon \,\big) = 0,\ \ \forall \varepsilon > 0}. \tag{6.2}$$

In other words, for any fixed small number $\varepsilon > 0$ the probability that the sample mean $\overline{X}_n$ deviates from the theoretical mean by more than $\varepsilon > 0$ is extremely small if n is extremely large. Thus, for large n is is extremely unlikely that $\overline{X}_n$ differs from μ by more than ε.

Proof. Observe that

$$\mathbb{E}\big[\, X_1 + \cdots + X_n \,\big] = n\mu \rightarrow \mathbb{E}\big[\, \overline{X}_n \,\big] = \mu.$$

Moreover, since the variables X_i are *independent* we have

$$\boldsymbol{var}\big[\, X_1 + \cdots + X_n \,\big] = n\, \boldsymbol{var}[X_1] = n\sigma^2.$$

We deduce from (2.21b) and (2.40b) that

$$\boldsymbol{var}\big[\, \overline{X}_n \,\big] = \frac{1}{n^2}\, \boldsymbol{var}\big[\, X_1 + \cdots + X_n \,\big] = \frac{\sigma^2}{n}.$$

Hence, the standard deviation of $\overline{X}_n$ is

$$\bar{\sigma}_n = \sqrt{\boldsymbol{var}[\overline{X}_n]} = \frac{\sigma}{\sqrt{n}}.$$

Observe that, for n very large the standard deviation $\bar{\sigma}_n$ is very small so the random variable is highly concentrated near its mean.

Fix $\varepsilon > 0$. From Chebyshev's inequality (2.32b) or (2.42b) we deduce

$$\mathbb{P}\big(\,|\overline{X}_n - \mu| > \varepsilon\,\big) \leq \frac{\boldsymbol{var}[\bar{X}_n]}{\varepsilon} = \frac{\sigma^2}{n\varepsilon^2} \to 0 \;\text{ as } n \to \infty.$$

□

Definition 6.2. Let (Y_n) be a sequence of random variables and μ a constant. We say that Y_n *converges in probability* to μ, and we write this $Y_n \xrightarrow{P} \mu$ if for any $\varepsilon > 0$ we have

$$\lim_{n\to\infty} \mathbb{P}(|Y_n - \mu| > \varepsilon) = 0.$$ □

Thus, the law of large numbers states that the sample means of a sequence of i.i.d. random variables with finite variance converge in probability to their common mean. It validates the vague but intuitive idea that "superposing independent observations average out the noise".

Remark 6.3. (a) We want to mention that finiteness of the variance is not needed for the Law of Large Numbers to hold. The proof in this more general case is considerably more ingenious. For details we refer to [**5**, X.2].

(b) The Law of Large Number as formulated above is a special case of the so called *Strong Law of Large Numbers* or SLLN which states that *the empirical mean $\bar{X}_n$ converges almost surely to the theoretical mean μ.*[1]

The event "$\bar{X}_n \to \mu$ as $n \to \infty$" can be given the set theoretic description

$$\bigcap_{\varepsilon>0} \bigcup_{N\geq 1} \bigcap_{n\geq N} \big\{\,|\bar{X}_n - \mu| < \varepsilon\,\big\}$$

and SLLN states that the above event has probability 1. This result is called the *Strong* Law because almost sure convergence implies convergence in probability. For its difficult proof we refer to [**5**, X.4].

Example 6.4 (The Monte Carlo method)**.** Consider a function of two variables $f(x, y)$ defined over the square $\boldsymbol{S} = [0, 1] \times [0, 1]$. We want to describe a probabilistic method of computing the double integral

$$\iint_{\mathbb{S}} f(x, y)dxdy.$$

[1] The precise statement of the SLLN requires more sophisticated mathematical concepts.

Suppose that X, Y are independent random variable uniformly distributed on $[0,1]$,

$$X, Y \sim \text{Unif}(0,1).$$

Then the random vector (X,Y) is uniformly distributed over the square $\boldsymbol{S}$ since its pdf is $dxdy$, $x, y \in [0,1]$. Let Z denote the random variable $Z := f(X,Y)$. The law of the subconscious statistician shows that

$$\mathbb{E}\big[\, Z \,\big] = \iint_{\boldsymbol{S}} f(x,y)dxdy.$$

Suppose that $X_1, Y_1, X_2, Y_2, \dots$ are independent random variables uniformly distributed on $[0,1]$. Then the random variables

$$Z_1 = f(X_1, Y_1),\;\; Z_2 = f(X_2, Y_2),$$

are independent and have the same distributions as Z. The (strong) law of large numbers shows that the empirical mean

$$\bar{Z}_N = \frac{1}{N}\big(\, Z_1 + Z_2 + \cdots + Z_N \big)$$

converges almost surely to $\mathbb{E}\big[\, Z \,\big]$ as $N \to \infty$.

We can use this for practical computations as follows: choose a large number of independent, uniformly distributed random samples

$$(X_1, Y_1), \dots, (X_N, Y_N)$$

of $\boldsymbol{S}$. Then, for N very large, with probability 1,

$$\iint_{\boldsymbol{S}} f(x,y)dxdy \approx \frac{1}{N}\sum_{k=1}^{N} f(X_k, Y_k).$$

In Example 7.18 we explain how to implement the Monte Carlo method in R. □

6.2. The central limit theorem

Suppose that

$$X_1, X_2, \dots, X_n, \dots$$

are i.i.d. random variables with mean μ, standard deviation σ and variance σ^2. The computations in the previous section show that the sum

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

has mean $n\mu$, variance $n\sigma^2$ and standard deviation $\sigma\sqrt{n}$,

$$\mathbb{E}[S_n] = n\mu,\;\; \boldsymbol{var}[S_n] = n\sigma^2,\;\; \sigma[S_n] = \sigma\sqrt{n}.$$

The *Central Limit Theorem* (or CLT) states that, for large n, S_n is very close to a normal random variable with the same mean $n\mu$ and variance $n\sigma^2$.

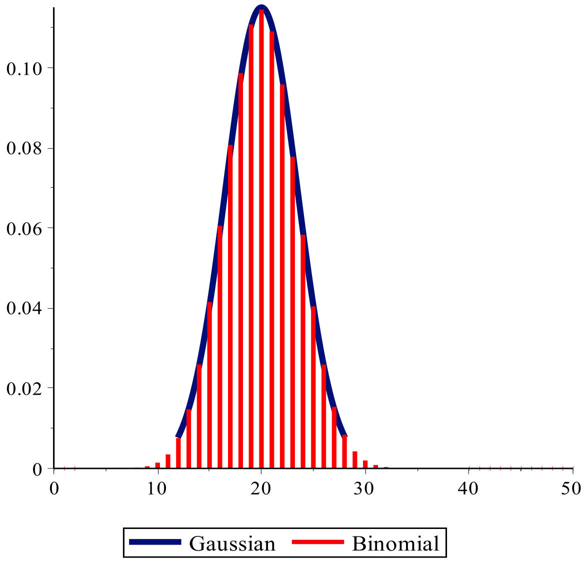

Figure 6.2. *The pmf of* $\mathrm{Bin}(50, 0.4)$ *is very close to the pdf of the Gaussian random variable* Y *with the same mean and variance.*

This can be easily seen in Figure 6.2. This depicts the special case when the random variables X_j are Bernoulli random variables, $X_k \sim \mathrm{Ber}(0.4)$, $\forall k \geq 1$. The sum S_n is then a binomial random variable $S_n \sim \mathrm{Bin}(n, 0.4)$. It has mean $0.4n$ and variance $0.24n$.

In Figure 6.2 we have depicted the pmf of S_{50} (red vertical segments) and the thick (blue) curve is the graph of the pdf of a normal random variable $Y \sim N(0.4 \cdot 50, 0.24 \cdot 50)$. The fact that the curve follows closely the profile of the pmf of $\mathrm{Bin}(50, 0.4)$ is no accident. It is a manifestation of the CLT.

To formulate the Central Limit Theorem (CLT) precisely we begin with a few simple observation. Note that the sample mean

$$\bar{X}_n = \frac{1}{n} S_n$$

has mean μ and variance

$$\bar{\sigma}_n^2 = \boldsymbol{var}\big[\,\bar{X}_n\,\big] = \frac{1}{n^2}\boldsymbol{var}\big[\,S_n\,\big] = \frac{\sigma^2}{n}.$$

It is thus highly concentrated near its mean, and as the Law of Large Numbers shows, it converges to μ in probability and almost surely.

The rescaled random variable

$$Z_n := \frac{1}{\bar{\sigma}_n}\big(\,\bar{X}_n - \mu\,\big) = \frac{1}{\sigma\sqrt{n}}\big(S_n - n\mu\,\big),$$

has mean 0 and standard deviation 1. The CLT states that the random variables Z_n approach a standard normal random variable $Y \sim N(0,1)$.

Theorem 6.5 (Central Limit Theorem)**.** *Suppose that*

$$X_1, X_2, \dots, X_n, \dots$$

are i.i.d. random variables with mean μ*, standard deviation* σ *and variance* σ^2*. Set*

$$S_n := X_1 + \cdots + X_n,$$
$$Z_n := \frac{1}{\sigma\sqrt{n}}\big(S_n - n\mu\big) = \frac{1}{\sigma[S_n]}\big(S_n - \mathbb{E}[S_n]\big).$$

If

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

denotes the cdf of $N(0,1)$*, then, for any* $x \in \mathbb{R}$*, we have*

$$\lim_{n\to\infty} \mathbb{P}\big(Z_n \leq x\big) = \Phi(x). \qquad \square$$

Remark 6.6. (a) The Central Limit Theorem shows that the normal distribution occupies a special place in probability since the empirical means of *any* sequence of i.i.d. square integrable random variables converge in probability to a normal random variable.

(b) The central limit theorem can be substantially improved if we have additional information on the random variables X_n. The *Berry-Essen theorem* shows that if

$$\rho := \mathbb{E}\big[\,|X_n - \mu|^3\,\big] < \infty,$$

then

$$\big|\,\mathbb{P}\big(Z_n \leq x\big) - \Phi(x)\,\big| \leq \frac{\rho}{\sigma^3\sqrt{n}},\ \ \forall x \in \mathbb{R},\ \ n = 1, 2, \dots .$$

For example, if n is 1 million, $n = 10^6$, the $\sqrt{n} = 10^3 = 1{,}000$ so

$$\mathbb{P}\big(Z_n \leq x\big) \approx \Phi(x) \pm 0.001.$$

Remark 6.7 (z-score)**.** The central limit essentially states that, for large n, the probability that S_n belongs to a given interval is very close to the probability that a normal random variable $N(n\mu, n\sigma^2)$ belongs to the same interval. If we choose this interval of the form

$$\big(n\mu - a\hat{\sigma}_n, n\mu + b\hat{\sigma}_n\big],\ \ \hat{\sigma}_n = \sigma[S_n] = \sigma\sqrt{n},$$

then the central limit theorem implies

$$\boxed{\mathbb{P}\big(a\hat{\sigma}_n < S_n - n\mu \leq b\hat{\sigma}_n\big) \approx \Phi(b) - \Phi(a)}. \tag{6.3}$$

Equivalently, this means

$$\mathbb{P}\big(\alpha < S_n - n\mu < \beta\big) \approx \Phi\left(\frac{\beta}{\sigma\sqrt{n}}\right) - \Phi\left(\frac{\alpha}{\sigma\sqrt{n}}\right). \tag{6.4}$$

In practice is convenient to rephrase (6.3) in terms of *z-score* defined as

$$\boxed{z = \frac{\text{observed value} - \text{expected value}}{\text{standard deviation}}}.$$

The numbers a, b in (6.3) are examples z-scores. Using the fact that $\mathbb{E}[S_n] = n\mu$ we deduce

$$\boxed{\mathbb{P}\Big(n\mu + z_*\sqrt{n}\sigma \leq S_n \leq n\mu + z^*\sqrt{n}\sigma\Big) \approx \Phi(z^*) - \Phi(z_*)}.$$

The error in the above approximation is roughly the size $\frac{1}{\sqrt{n}}$. For small n this error is rather large. E.g. for $n = 100$, the error could be as large as 0.1. □

Example 6.8 (Histogram correction)**.** Suppose that we roll a die $n = 120$ times. We denote by S_n the number fives rolled. We have

$$S_n \sim \operatorname{Bin}(n, 1/6),\ \ \mathbb{E}[S_n] = \frac{n}{6} = 20,\ \ \sigma[S_n] = \sqrt{\frac{5\cdot 120}{36}} = \sqrt{\frac{50}{3}} \approx 4.0824$$

Thus the theory tells us that we should expect $\mathbb{E}[S_n] = 20$ with a standard deviation of ≈ 4.0824.

Let us estimate what is the probability p_1 that after 120 rolls the number of fives is at most 21, i.e.,

$$\mathbb{P}(S_n \leq 21).$$

The z-score corresponding to 21 is

$$z^* := \frac{21-20}{4.0824} \approx 0.2449.$$

The central limit theorem then states the probability p_1 should be close to $\Phi(0.2449)$. We can compute this number using the R command

```
pnorm(0.2449)
```

which yields the approximation $p_1 \approx 0.5967$. We can find the actual value of p_1 using the R command

```
pbinom(21,120,1/6)
```

which yields the "real"[2] value

$$p_1 = 0.6520062.... \tag{6.5}$$

This is quite different.

To improve the approximation we use the *histogram correction* trick. Instead of finding a normal (Gaussian) approximation to $\mathbb{P}(S_n \leq 21)$ we seek a normal approximation to $\mathbb{P}(S_n \leq 21.5)$. In this case the z-score is

$$z^* := \frac{21.5-20}{4.0824} \approx 0.3674.$$

[2] The output of R is not truly the real value. It is however a very good approximation.

We can compute $\Phi(0.3674)$ using the R command

```
pnorm(0.3674)
```

which yields the new approximation

$$p_1 \approx 0.6433$$

which is much closer to the real probability (6.5).

The histogram correction trick should be used anytime the random variable S_n is *integer valued*. Thus, instead of approximating $\mathbb{P}(S_n \leq k)$, where k is an integer, we approximate the probability $\mathbb{P}(S_n \leq k + 0.5)$.

While $\mathbb{P}(S_n \leq k) = \mathbb{P}(S_n \leq k + 0.5)$, the normal approximations of the probabilities of the events $\{S_n \leq k\}$ and $\{S_n \leq k + 0.5)$ are different.

Similarly, instead of approximating the probability of the event $\{S_n \geq k)$, k integer, we approximate the probability of the event $\{S_n \geq k - 0.5\}$. For n very very large, the histogram correction trick does not significantly improve the normal approximation. □

Example 6.9 (Casino business, again)**.** Let us see what the central limit theorem says about the casino business problem introduced in Example **??**. Recall the setup.

Suppose you run a casino consisting of 50 identical and independent slot machines. A player earns \$1 with probability $p = 0.45$ and loses a dollar with probability $q = 1 - p = 0.55$. It is known that each machine is played 200 games per day, for 365 days each year. We wan to estimate the probability that the yearly profit[3] from these machines is smaller than \$350, 000.

Denote by X the profit per game from one machine. The house's profit is \$1 when the player loses, and its "profit" is negative $\$-1 = 1 - 2$, when the player wins.

$$\mathbb{E}[X] = -1p + (1 - p) = 1 - 2p = 0.1.$$

Note that $\mathbb{E}[X^2] = 1$. The standard deviation is

$$\sigma = \sqrt{1 - (1 - 2p)^2} = \sqrt{0.9} \approx 0.9949.$$

The number of games per year is

$$n = 50 \times 200 \times 365 = 3,650,000$$

so the expected profit is \$365, 000 with a standard deviation

$$\sqrt{n}\sigma \approx 1900.21.$$

[3] This is a simplified model. We exclude many costs such as utilities, employees etc. when computing the profit.

In this case the z score is

$$z = \frac{350,000 - 365,000}{1900.21} \approx -7.89.$$

The central limit then predicts that the probability that the yearly profit is $< 350,000$ to be close to $\Phi(-7.89) \approx 10^{-15}$!!!

The above computations are organized in the R procedure below.

```
#p is the player's winning probability
#g is the number of games per day per machine
#m is the number of machines
#r is ratio  of expected profit
pr<-function(p,g,m,r){
  x<-1-2*p
  s<-sqrt(1-x^2)
  n<-m*g*365
  X<-n*x
  S<-sqrt(n)*s
  Y<-r*X
  Z<-(Y-X)/S
  ratio<-r*100
  lik<-pnorm(Z)
  cat("The probability of earning
  less than ", Y, " is approximatively  ",lik, ".",sep="")
}
```

To find the probability of earning less than $350,000 using the above function, first compute the ratio

$$r = \frac{350000}{365000} \approx 0.9589$$

and then use the R command

```
pr(0.45, 200, 50, 0.9589).
```

The result will have the form

```
The expected profit is 365000. The probability of earning
less than 349998.5 is 1.490469e-15.
```

Suppose we tweak the machines, lowering the winning probability to $p = 0.4$ and seek the probability of earning less that 99% of the expected profit. We invoke the command

```
pr(0.4, 200, 50, 0.99)
```

and the answer we get is

```
The expected profit is 730000. The probability of earning
less than 722700 is 4.813881e-05.
```

It is interesting to compare this approximation with the Chernoff bounds. When we run the R-script discussed in Example **??**

```
chernoff(0.55,0.9589, 3650000)
```

we obtain the result

```
The probability of earning less than 349998.5
 is smaller than 3.01995e-14.
```

The central limit approximation predicts a smaller probability. It is however only an approximation amd the Chernoff bounds offer some information on the size of the error of this approximation. □

6.3. Exercises

Exercise 6.1. Fifty numbers are rounded off to the nearest integer and then summed. If the individual round-off errors are uniformly distributed over $(-.5, .5)$, approximate the probability that the resultant sum differs from the exact sum by more than 3.

Exercise 6.2. A certain component is critical to the operation of an electrical system and must be replaced immediately upon failure. If the mean lifetime of this type of component is 100 hours and its standard deviation is 30 hours, how many of these components must be in stock so that the probability that the system is in continual operation for the next 2000 hours is at least 0.95?

Exercise 6.3. Estimate the probability that the average of 150 random points from the interval $[0, 1]$ lies within 0.02 of the midpoint 0.5.

Exercise 6.4. An insurance company has $10,000$ automobile policyholders. The expected yearly claim per policy- holder is \$ 240, with a standard deviation of \$ 800. Approximate the probability that the total yearly claim exceeds \$ 2.7 million.

Exercise 6.5. Each time that Jim charges an item to his credit card, he rounds the amount to the nearest dollar in his records. If he has used the credit card 300 times in the last 12 months, what is the probability that his record differs from the total expenditure by at most 10 dollars?

Exercise 6.6. Suppose that, whenever invited to a party, the probability that a person attends with a guest is $1/3$, the probability a person attends alone is $1/3$, and the probability that a person does not attend at all is $1/3$. A company invites all 300 of its employees to a Christmas party. Use the central limit theorem to estimate the probability that at least 320 will attend.

Exercise 6.7. A fair coin is flipped repeatedly. Approximate the probability of flipping 25 heads before 50 tails.

Exercise 6.8. On each bet, a gambler loses \$1 with probability 0.7, loses \$2 with probability 0.2, or wins \$10 with probability 0.1. Approximate the probability that his cummulative winning of the gambler after his first 100 bets is negative.

Chapter 7

A very basic introduction to R

This is not an introduction to programming in R. It mainly lists a few basic tricks that you might find useful in dealing with simple probability problems. First, here is how you install R on your computers

For Mac users

https://cran.r-project.org/bin/macosx/

For Windows users

https://cran.r-project.org/bin/windows/base/

Next, install R Studio (the Desktop version). This is a very convenient interface for using R.

https://www.rstudio.com/products/RStudio/

(*Install first R and then R Studio.*) You can also access RStudio and R in the cloud

https://www.rollapp.com/app/rstudio

The site

http://www.people.carleton.edu/~rdobrow/Probability/

has a repository of many simple R programs (or R scripts) that you can use as models.

The reader familiar with the basics of programming will have no problems learning the basics of R. This introduction is addressed to such a reader. We list some of commands most frequently used probability and we have included several examples so the reader learns the R-syntax of the basic operations that enter a program. R-Studio comes with links to various freely availableweb sources for R-programming. A commercial source that I find very useful is "*The Book of R*", [2].

Example 7.1 (Operations with vectors)**.** The workhorse of R is the object called *vector*. An n-dimensional vector is essentially an element in $\mathbb{R}^n$. An n-dimensional vector in R can be more general in the sense that its entries need not be just numbers.

To generate in R the vector $(1, 2, 4.5)$ and then naming it x use the command

```
x<-c(1,2,4.5)
```

To see what the vector x is type

```
x
```

If you want to add an entry to x, say you want to generate the longer vector $(1, 2, 4.5, 7)$, use the command

```
c(x,7)
```

For long vectors this approach can be time consuming. This can be accelerated if the entries of the vector x are subject to patterns. For example, the vector of length 22 with all entries equal to the same number, say 1.5, can be generated using the command

```
rep(1.5, 22)
```

To generate the vector listing in increasing order all the integers between -2 and 10 (included) use the command

```
(-2):10
```

To generate the vector named x consisting of 25 equidistant numbers staring at 1 and ending at 7 use the command

```
 x<-seq(from=1, to=7, length.out=25)
```

To add all the entries of a vector $x = (x_1, \dots, x_n)$ use the command

```
 sum(x)
```

Suppose we want to add all the natural numbers from 50 to 200. We first store these numbers in a vector called x

```
 x<-50:200
```

The sum of all these numbers is then computed using the command

```
sum(x)
```

The result is $18, 875$.

You can sort the entries of a vector, if they are numerical. For example

```
> z<-c(1,4,3)
> sort(z)
[1] 1 3 4
```

A very convenient feature of working with vectors in R is that the basic algebraic operations involving numbers extend, component wise. For example, if z is the above vector, then

```
 > z^2
[1]  1 16  9
> 2^z
[1]  2 16  8
```

□

Example 7.2 (Logical operators)**.** These are operators whose output is a TRUE or FALSE or a vector whose entries are TRUE/FALSE.

For example, the command $2 < 5$ returns TRUE. On the other hand if x is the vector $(2, 3, 7, 8)$, then the command $x < 5$ return

```
TRUE,TRUE,  FALSE, FALSE.
```

In R the logicals TRUE/FALSE also have arithmetic meaning,

$$\text{TRUE} = 1, \ \ \text{FALSE} = 0.$$

The output of $x < 5$ is a vector whose entries are TRUE/FALSE. To see how many of the entries of x are < 5 use the command

```
sum(x<5)
```

Above $x < 5$ is interpreted as a vector with 0/1-entries. When we add them we count how many are equal to 1 or, equivalently, how many of the entries of x are < 5.

The R language also has two very convenient logical operators any and all. When we apply any to a vector with TRUE/FALSE entries it returns TRUE if at least one of the entries of v are TRUE and returns FALSE otherwise. When we apply all to a vector v with TRUE/FALSE entries it returns TRUE if *all* of the entries of v are TRUE and returns FALSE otherwise. □

Example 7.3 (Functions in R)**.** One can define and work with functions in R. For example, to define the function

$$f(q) = 1 + 6q + 10q^2(1 - q)^4$$

use the command

```
f<-function(q) (1+4*q+10*q^2)*(1-q)^4
```

To find de value of f at $q = 0.73$ use the command

```
f(0.73)
```

To display the values of f at all the points

$$0,\ 0.01,\ 0.02,\ 0.03, \dots,\ 0.15,\ 0.16$$

use the command

```
x<-seq(from=0, to=0.16, by=0.01)
f(x)
```

To plot the values of f over 100 equidistant points in the interval $[2, 7]$ use the command

```
x<-seq(from=2, to=7, length.out=100)
y<-f(x)
plot(x,y, type="l")
```

Here is how we define in R the indicator function of the unit disc in the plane

$$I_D(x, y) = \begin{cases} 1, & x^2 + y^2 \leq 1, \\ 0, & x^2 + y^2 > 1. \end{cases}$$

```
indicator<-function(x,y) if(1 >= x^2+y^2)  1 else 0
```

Another possible code that generates this indicator function is

```
indicator<-function(x,y) as.integer(x^2+y^2<= 1)
```

Above, the command as.integer converts TRUE/FALSE to 1/0. □

Example 7.4 (Samples with replacement)**.** For example, to sample *with replacement* 7 balls from a bin containing balls *labeled* 1 through 23 use the R command

```
 sample(1:23,7, replace=TRUE)
```

The result is a 7-dimensional vector whose entries consists of 7 numbers sampled with replacement from the set $\{1, \dots, 23\}$. Similarly, to simulate rolling a fair die 137 times use the command

```
 sample(1:6,137, replace=TRUE)
```

Example 7.5 (Rolling a die)**.** Let us show how to simulate rolling a die a number n of times and then count how many times we get 6. Suppose $n = 20$. We indicate this using the command

```
 n<-20
```

We now roll the die n times and store the results in a vector x

```
x<-sample(1:6, n, replace=TRUE)
```

Next we test which of the entries of x are equal to 6 and store the results of these 20 tests in a vector y

```
y<-x==6
```

The entries of y are True or False, depending on whether the corresponding entry of x was equal to 6 or not. To find how many entries of y are T use the command

```
sum(y)
```

The result is equal to the number of 6s we got during the string of 20 rolls of a fair die. □

Example 7.6 (Samples without replacement)**.** To sample without replacement 7 balls from an urn containing balls labeled 1 through 23 use the R command

```
sample(1:23, 7)
```

The number of possible samples above is $(27)_7$ and to compute it use the R command

```
prod(21:27)
```

□

Example 7.7 (Permutations)**.** To sample a random permutation of 7 objects use the R command

```
sample(1:7,7)
```

To sample 10 random permutations of 7 objects use the R command

```
for (i in 1:10 ) print(sample(1:7,7))
```

To compute 7! in R use the command

```
factorial(7)
```

□

Example 7.8 (Combinations)**.** Sampling random m-element subsets out of an n-element set possible is possible in R. For example, to sample 4 random subsets with 2 elements out of a 7-element set possible the following command

```
replicate(4, sort( sample(1:7, 2) ))
```

The sampled sets will appear as columns. To compute $\binom{52}{5}$ in R use the command

```
choose(52,5)
```

To compute $\left(\!\!\binom{n}{k}\!\!\right)$ is is convenient to define a function

```
mchoose<-function(n,k) {choose(n+k-1,k)}
```

To compute $\left(\!\!\binom{10}{8}\!\!\right)$ use the command

```
mchoose(10,8)
```

□

Example 7.9 (Custom discrete distribution)**.** We can produce custom discrete random variables in R.

Suppose that we want to simulate a discrete random variable X whose values are (in increasing order)

$$x_1 = 0.1,\ \ x_2 = 0.2,\ \ x_3 = 0.3,\ \ x_4 = 0.7$$

with probabilities

$$p_1 = 1/3,\ \ p_2 = 1/6,\ \ p_3 = 1/4,\ \ p_4 = 1/4.$$

The R-commands below describe how to compute the mean and the variance of X and how to sample X.

```
X<-c(0.1,0.2,0.3,0.7) # stores the values  of X in increasing order.
prob<-c(1/3,1/6,1/4,1/4) # stores the probabilities.
sum(prob) # If this is 1 prob is a pmf. Otherwise check prob.
m<-sum(X*prob) # computes the mean of X and stores in m.
v<-sum((X^2)*prob) -m^2# computes the variance of X
# and stores it in v.
m # produces the value of the mean.
v # produces the variance of X.
sample(X,15,replace=TRUE, prob) # produces 15 random
#samples of X.
cumsum(prob) # computes the values of the cdf of X at
# x_1,x_2,...
```

In R the symbol # indicates a comment. It is only for the programer/user benefit. Anything following a # is not treated by R as a command. □

Example 7.10 (Useful discrete distributions)**.** The standard discrete distributions are implemented in R.

The distribution	The R command
The binomial distribution Bin(n, p)	binom(n,p)
The geometric distribution Geom(p)	geom(p)
The negative binomial distribution NegBin(k, p)	nbinom(k,p)
The Poisson distribution Poi(λ)	pois(lambda)

The R library however uses rather different conventions

(i) The geometric distribution in R is slightly different from the one described in these notes. In R, the range of Geom(p) variable T is $\{0, 1, \dots\}$ and its pmf is $\mathbb{P}(T = n) = p(1-p)^n$. In the present course notes, a geometric random variable has range $\{1, 2 \dots\}$ and its pmf is $\mathbb{P}(T = n) = p(1-p)^{n-1}$; see Example 7.12.

(ii) In R the equality $\mathbf{nbinom}(k, p) = n$ represents the number of *failures* until we register the k-th success; see Example 7.13.

The above commands by themselves mean nothing if they are not accompanied by one of the prefixes

- d produces the density or $\boxed{\textit{pmf}}$.
- p produces the $\boxed{\textit{cdf}}$.
- r produces random $\boxed{\textit{samples}}$.
- q produces $\boxed{\textit{quantiles}}$.

□

You can learn more details using R's help function. The examples below describe some concrete situations.

Example 7.11 (Binomial)**.** For example, suppose that $X \sim \text{Bin}(10, 0.2)$, i.e., X is the number of successes in a sequence of 10 independent Bernoulli trials with success probability 0.2.

To find the probability $\mathbb{P}(X = 3)$ use the R command

```
dbinom(3,10,0.2)
```

If $F_X(x) = \mathbb{P}(X \leq x)$ is the cdf of X, then you can compute $F_X(4)$ using the R command

```
pbinom(4,10,0.2)
```

To generate 253 random samples of X use the command

```
rbinom(253,10,0.2)
```

To find the 0.8-quantile of X use the R command

```
qbinom(0.8,10,0.2)
```

□

Example 7.12 (Geometric)**.** Suppose now that $T \sim \text{Geom}(0.2)$ is the waiting time until the first success in a sequence of independent Bernoulli trials with success probability $p = 0.2$.

To find the probability $\mathbb{P}(T = 3)$ use the command

```
dgeom(3-1,0.2)
```

To find the probability $\mathbb{P}(T \leq 4)$ use the command

```
pgeom(4-1,0.2)
```

To generate 253 random samples of T use the command

```
1+rgeom(253,0.2)
```

To find the 0.8-quantile of T use the R command

```
qgeom(0.8,0.2)+1
```

Example 7.13 (Negative Binomial)**.** Suppose that $T \sim \text{NegBin}(8, 0.2)$ is the waiting time for the first 8 successes in a string of Bernoulli trials with success probability.
To find the probability $\mathbb{P}(T = 12)$ use the R command

```
dnbinom(12-8,8,0.2)
```

You can compute $\mathbb{P}(T \leq 14)$ using the R command

```
pnbinom(14-8,8,0.2)
```

To generate 253 random samples of T use the command

```
8+rnbinom(253,8,0.2)
```

To find the 0.8-quantile of T use the R command

```
8+qnbinom(0.8,8,0.2)
```

□

Example 7.14 (Poisson)**.** Suppose that $X \sim \text{Poi}(0.2)$ is a Poisson random variable with parameter $\lambda = 0.2$.
To find the probability $\mathbb{P}(X = 3)$ use the command

```
dpois(3,0.2)
```

To find the probability $\mathbb{P}(X \leq 4)$ use the command

```
ppois(4,0.2)
```

To generate 253 random samples of X use the command

```
rpois(253,0.2)
```

To find the 0.8-quantile of X use the R command

```
qpois(0.8,0.2)
```

□

Example 7.15 (Continuous distributions in R)**.** The continuous distributions $\mathrm{Unif}(a, b)$, $\exp_\lambda$ and $N(\mu, \sigma^2)$ can be simulated in R by invoking

```
unif(min=a, max=b)
```

```
exp(rate=lambda)
```

```
norm(mean=mu, sd=sigma)
```

where sd:=standard deviation.

To invoke the standard normal random variable one could use the shorter command

```
norm
```

□

As in the case of discrete distributions, we utilize these commands with the prefixes $d-$, $p-$, $q-$ and $r-$ that have the same meaning as in R-Session 7.10. Thus $d-$ will generate the pdf, $p-$ the cdf, $r-$ generates a random sample, and $q-$ produces quantiles.

Example 7.16. Here are some concrete examples. To find the probability density of $\exp_3$ at $x = 1.7$ use the command

```
dexp(1.7, 3)
```

To find the probability density of $N(\mu = 5, \sigma^2 = 7)$ at $x = 2.6$ use the command

```
dnorm(2.6,5, sqrt(7))
```

To produce 1000 samples from $\mathrm{Unif}(3, 13)$ use the command

```
runif(1000,3,13)
```

□

Example 7.17 (Buffon's needle problem)**.** The R program below uses the Buffon needle problem to find an approximation of π

```
L<-0.7 # L is the length of the needle. It is <1.
N<-1000000 # N is the number of times we throw the needle.
f<-0
#the next loop simulates the tossing of
#N random needles and computes
# the number f  of times they intersect a line

for (i in 1:N){
  y<-runif(1, min=-1/2,max=1/2) #this locates
  # the center of the needle
  t<-runif(1, min=-pi/2,max=pi/2)#this determines
  #the   inclination of the needle
  if ( abs(y)< 0.5*L*cos(t) ) f<-f+1 }
#f/N  is the empirical  frequency
"the aproximate value of pi is";  (N/f)*2*L
```

□

Example 7.18 (Monte Carlo)**.** The R-command lines below implement the Monte Carlo strategy for computing a double integral over the unit square

```
# Monte Carlo integration of the function f(x,y)
#over the rectangle [a,b] x[c,d]
# First we describe the function
f<- function(x,y) sin(x*y)
# Next, we describe the region of integration [a,b]x[c,d]
a=0
b=1
c=0
d=1
# Finally, we decide the number N  of sample points in
# the region of integration
N=100000
#S will store the integral
S=0
for (i in 1:N){
  x<- runif(1,a,b) #we sample a point uniformly  in [a,b]
  y<- runif(1,c,d) #we sample a point uniformly  in [c,d]
  S<-S+f(x[1],y[1])
}
’the integral is’; (b-a)*(d-c)*S/N
```

The next code describes a Monte-Carlo computation of the area of the unit circle.

```
nsim<-1000000#nsim is the number of simulations
x<-runif(nsim,-1,1)#we choose nsim uniform samples
#in the interval (-1,1) on the x axis
y<-runif(nsim,-1,1)#we choose nsim uniform samples
#in the interval (-1,1) on the y axis
area<-4*sum(x^2+y^2<1)/nsim
"the area of the unit circle is very likely"; area
```

Example 7.19. We describe below the R code that can be used to simulate the problem discussed in Example 3.32. Namely it determines empirically the waiting time to observe a coin pattern. It deals with a rather more general situation. We deal with a "die" with L equally likely faces. When $L = 2$ this die can be viewed as a fair coin, while for $L = 6$ deals with a traditional.

We play m games. A game consists of rolling this generalized die until we observe the pattern patt. After each game the code records the number of rolls of this generalized die required to observe the pattern.

The code will return the average of these m empirical waiting times and graphs the cumulative averages of these empirical waiting times.

```
#patt is  a vector that encodes the pattern
#m is the number of games we play
#L number of ‘faces of the die
#the output is a vector  of cumulative frequencies


Tpattern<-function(patt, m, L){
  k<-length(patt)
 T<-c()
 for (i in 1:m){
x<-sample(1:L,k,replace=TRUE)
n<-k
while ( all(x[(n-k+1):n]==patt)==0 ){
x<-c(x, sample(1:L,1,replace=TRUE) )
 n<-n+1
}
T<-c(T,n)
 }
f<-cumsum(T)/(1:m)
 f
}
```

Now let's use the above function you need to declare the value L, the number of sides of the die, choose a pattern patt and decide on the number m of games you want to play. For example if you flip a coin then $L = 2$. If the pattern is HH and you want to play 240 games to find empirically the waiting time then use the comands

```
m<-240
patt<-c(1,1)
y<-Tpattern(patt,m,2)
"the mean waiting time to pattern";patt; "is"; y[m]
```

```
plot(1:m,y,type="l", xlab="Number of games",
 ylab="Running average of waiting time to pattern")
abline(h=y[m],col="red")
```

□

Basic invariants of frequently used probability distributions

$$X \sim \mathrm{Bin}(n,p) \Longleftrightarrow \mathbb{P}(X=k) = \binom{n}{k} p^k q^{n-k}, \ \ k=0,1,\dots,n, \ \ q=1-p.$$

$$\mathrm{Ber}(p) \sim \mathrm{Bin}(1,p).$$

$$X \sim \mathrm{NegBin}(k,p) \Longleftrightarrow \mathbb{P}(X=n) = \binom{n-1}{k-1} p^k q^{n-k}, \ \ n=k,k+1,\dots$$

$$\mathrm{Geom}(p) \sim \mathrm{NegBin}(1,p).$$

$$X \sim \mathrm{HGeom}(w,b,n), \ \ \mathbb{P}(X=k) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}}, k=0,1,\dots,w.$$

$$X \sim \mathrm{Poi}(\lambda), \ \ \lambda>0 \Longleftrightarrow \mathbb{P}(X=n) = e^{-\lambda}\frac{\lambda^n}{n!}, \ \ n=0,1,\dots$$

$$X \sim \mathrm{Unif}(a,b) \Longleftrightarrow p_X(x) = \frac{1}{b-a} \times \begin{cases} 1, & x \in [a,b], \\ 0, & x \notin [a,b]. \end{cases}$$

$$X \sim \mathrm{Exp}(\lambda), \ \ \lambda>0 \Longleftrightarrow p_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \ge 0, \\ 0, & x < 0. \end{cases}$$

$$X \sim N(\mu,\sigma^2), \ \ \mu \in \mathbb{R}, \ \ \sigma>0 \Longleftrightarrow p_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \ \ x \in \mathbb{R}.$$

$$X \sim \mathrm{Gamma}(\nu,\lambda) \Longleftrightarrow p_X(x) = \begin{cases} \frac{\lambda^\nu}{\Gamma(\nu)} x^{\nu-1} e^{-\lambda x}, & x > 0,\\ 0, & x \leq 0, \end{cases} \quad \nu,\lambda > 0.$$

$$X \sim \mathrm{Beta}(a,b) \Longleftrightarrow p_X(x) = \frac{1}{B(a,b)} \times \begin{cases} x^{a-1}(1-x)^{b-1}, & x \in (0,1),\\ 0, & \text{otherwise.} \end{cases}$$

Name	Mean	Variance	pgf	mgf
$\mathrm{Ber}(p)$	p	pq	$(q+ps)$	pe^t
$\mathrm{Bin}(n,p)$	np	npq	$(q+ps)^n$	$p^n e^{nt}$
$\mathrm{Geom}(p)$	$\frac{1}{p}$	$\frac{q}{p^2}$	$\frac{ps}{1-qs}$	$\frac{pe^t}{1-qe^t}$
$\mathrm{NegBin}(k,p)$	$\frac{k}{p}$	$\frac{kq}{p^2}$	$\left(\frac{qs}{1-ps}\right)^k$	$\left(\frac{pe^t}{1-qe^t}\right)^k$
$\mathrm{Poi}(\lambda)$	λ	λ	$e^{\lambda(s-1)}$	$e^{\lambda(e^t-1)}$
$\mathrm{HGeom}(w,b,n)$	$\frac{w}{w+b}\cdot n$	$*$	(2.28)	$*$
$\mathrm{Unif}(a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	NA	$\frac{e^{tb}-e^{ta}}{tb-ta}$
$\mathrm{Exp}(\lambda)$	λ^{-1}	λ^{-2}	NA	$\frac{\lambda}{\lambda-t}$
$N(\mu,\sigma^2)$	μ	σ^2	NA	$\exp\left(\frac{\sigma^2}{2}t^2+\mu t\right)$
$\mathrm{Gamma}(\nu,\lambda)$	$\frac{\nu}{\lambda}$	$\frac{\nu}{\lambda^2}$	NA	$\left(\frac{\lambda}{\lambda-t}\right)^\nu$
$\mathrm{Beta}(a,b)$	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$	NA	$*$

Basic facts about the Gamma function.

$$\Gamma : (0,\infty) \to \mathbb{R}, \quad \boxed{\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt}.$$

(i) $\Gamma(1) = 1$.

(ii) $\Gamma(x+1) = x\Gamma(x)$, $\forall x > 0$.

(iii) $\Gamma(n) = (n-1)!$, $\forall n = 1,2,3,\dots$.

(iv) $\Gamma(1/2) = \sqrt{\pi}$.

(v) For any $x, y > 0$ we have

$$B(x,y) := \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} = \int_0^1 s^{x-1}(1-s)^{y-1}ds = \int_0^\infty \frac{u^{x-1}}{(1+u)^{x+y}}du. \tag{A.1}$$

The function $\Gamma(x)$ grows very fast as $x \to \infty$. Its asymptotics is governed by the *Stirling formula*

$$x\Gamma(x) \sim \sqrt{2\pi x}\left(\frac{x}{e}\right)^x \quad \text{as } x \to \infty. \tag{A.2}$$

Note that for $n \in \mathbb{N}$ the above estimate reads

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \quad \text{as } n \to \infty. \tag{A.3}$$

Appendix B

Solutions to homework problems

Bibliography

[1] K. L. Chung, F. AitSahlia: *Elementary probability theory: with stochastic processes and an introduction to mathematical finance*, Springer Verlag, 2003.

[2] T.M. Davies: *The Boof of R*, No Starch Press, 2015, `https://nostarch.com/bookofr`

[3] R.P. Dobrow: *Probability with Applications and R*, John Wiley & Sons, 2014.

[4] R. Durrett: *Elementary Probability for Applications*, Cambridge University Press, 2009.

[5] W. Feller: *An Introduction to Probability Theory and its Applications*, vol.1, 3rd Edition, John Wiley & Sons, 1968.

[6] M. Gardner: *Mathematical games*, Scientific American , Vol. 231, No. 4 (October 1974), pp. 120-125. `https://www.jstor.org/stable/10.2307/24950199`

[7] G. R. Grimett, D. R.Stirzaker: *Probability and Random Processes*, 3rd Edition, Oxford University Press, 2001.

[8] C. M. Grinstead, J.L. Snell: *Introduction to Probability: A Second Revised Edition*, Amer. Math. Soc. 1997. This book is also freely available in electronic form at `http://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/book.html`

[9] S. Karlin, H.M. Taylor: *An Introduction to Stochastic Modeling*, 3rd Edition, Academic Press, 1998.

[10] J. M. Keynes: *A Treatise on Probability*, MacMillan and Co. London, 1921. Available at Project Guttenberg `http://www.gutenberg.org/ebooks/32625`.

[11] A.N. Kolmogorov: *Grundsbegriffe der Wahrscheinlichkeitrechnung*, 1933. English Translation: *Foundations of the Theory of Probability*, Chelsea Publishing Co. 1956.

[12] N.N. Lebedev: *Special Functions and Their Applications*, Dover, 1972.

[13] S.-Y. R. Li: *A martingale approach to the study of occurrence of sequence patterns in repeated experiments*, Ann. Prob. **8**(1980), 1171-1176.

[14] R. Motwani, P. Raghavan: *Randomized Algorithms*, Cambridge University Press, 1995.

[15] P. Olofsson, M. Andersson: *Probability, Statistics and Stochastic Processes*, 2nd Edition, John Wiley & Sons, 2012.

[16] A. Rényi: *Probability Theory*, Dover Publications, 2007.

[17] S. S. Venkatesh: *The Theory of Probability. Explorations and Applications*, Cambridge University Press, 2013.

Index