THE UNIVERSITY OF QUEENSLAND

A Short Introduction to Probability

Prof. Dirk P. Kroese
Department of Mathematics

Contents

Preface

These notes form a comprehensive 1-unit (= half a semester) second-year introduction to probability modelling. The notes are not meant to replace the lectures, but function more as a *source of reference*. I have tried to include proofs of all results, whenever feasible. Further examples and exercises will be given at the tutorials and lectures. To completely master this course it is important that you

1. visit the lectures, where I will provide many extra examples;
2. do the tutorial exercises and the exercises in the appendix, which are there to help you with the "technical" side of things; you will learn here to apply the concepts learned at the lectures,
3. carry out random experiments on the computer, in the simulation project. This will give you a better intuition about how randomness works.

All of these will be essential if you wish to understand probability beyond "filling in the formulas".

Notation and Conventions

Throughout these notes I try to use a uniform notation in which, as a rule, the number of symbols is kept to a minimum. For example, I prefer q_{ij} to $q(i,j)$, X_t to $X(t)$, and $\mathbb{E}X$ to $\mathbb{E}[X]$.

The symbol ":=" denotes "is defined as". We will also use the abbreviations r.v. for *random variable* and i.i.d. (or iid) for *independent and identically and distributed*.

I will use the sans serif font to denote probability distributions. For example Bin denotes the binomial distribution, and Exp the exponential distribution.

Numbering

All references to Examples, Theorems, etc. are of the same form. For example, Theorem 1.2 refers to the second theorem of Chapter 1. References to formula's appear between brackets. For example, (3.4) refers to formula 4 of Chapter 3.

Literature

- Leon-Garcia, A. (1994). *Probability and Random Processes for Electrical Engineering*, 2nd Edition. Addison-Wesley, New York.
- Hsu, H. (1997). *Probability, Random Variables & Random Processes.* Shaum's Outline Series, McGraw-Hill, New York.
- Ross, S. M. (2005). *A First Course in Probability*, 7th ed., Prentice-Hall, Englewood Cliffs, NJ.
- Rubinstein, R. Y. and Kroese, D. P. (2007). *Simulation and the Monte Carlo Method*, second edition, Wiley & Sons, New York.
- Feller, W. (1970). *An Introduction to Probability Theory and Its Applications*, Volume I., 2nd ed., Wiley & Sons, New York.

Chapter 1

Random Experiments and Probability Models

1.1 Random Experiments

The basic notion in probability is that of a **random experiment**: an experiment whose outcome cannot be determined in advance, but is nevertheless still subject to analysis.

Examples of random experiments are:

1. tossing a die,
2. measuring the amount of rainfall in Brisbane in January,
3. counting the number of calls arriving at a telephone exchange during a fixed time period,
4. selecting a random sample of fifty people and observing the number of left-handers,
5. choosing at random ten people and measuring their height.

Example 1.1 (Coin Tossing) The most *fundamental* stochastic experiment is the experiment where a coin is tossed a number of times, say n times. Indeed, much of probability theory can be based on this simple experiment, as we shall see in subsequent chapters. To better understand how this experiment behaves, we can carry it out on a digital computer, for example in Matlab. The following simple Matlab program, simulates a sequence of 100 tosses with a fair coin(that is, heads and tails are equally likely), and plots the results in a bar chart.

```
x = (rand(1,100) < 1/2)
bar(x)
```

Here x is a vector with 1s and 0s, indicating Heads and Tails, say. Typical outcomes for three such experiments are given in Figure 1.1.

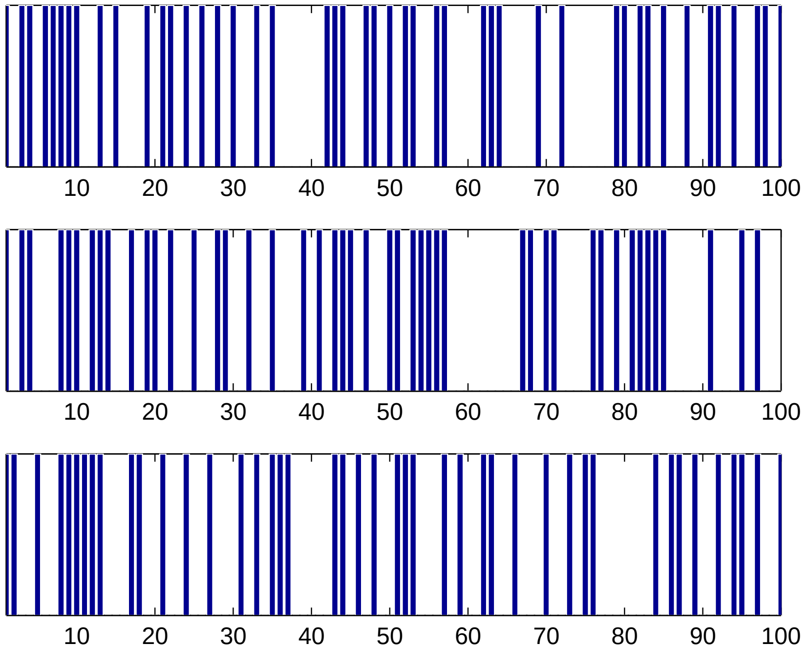


Figure 1.1: Three experiments where a fair coin is tossed 100 times. The dark bars indicate when "Heads" (=1) appears.

We can also plot the average number of "Heads" against the number of tosses. In the same Matlab program, this is done in two extra lines of code:

```
y = cumsum(x)./[1:100]
plot(y)
```

The result of three such experiments is depicted in Figure 1.2. Notice that the average number of Heads seems to converge to 1/2, but there is a lot of random fluctuation.

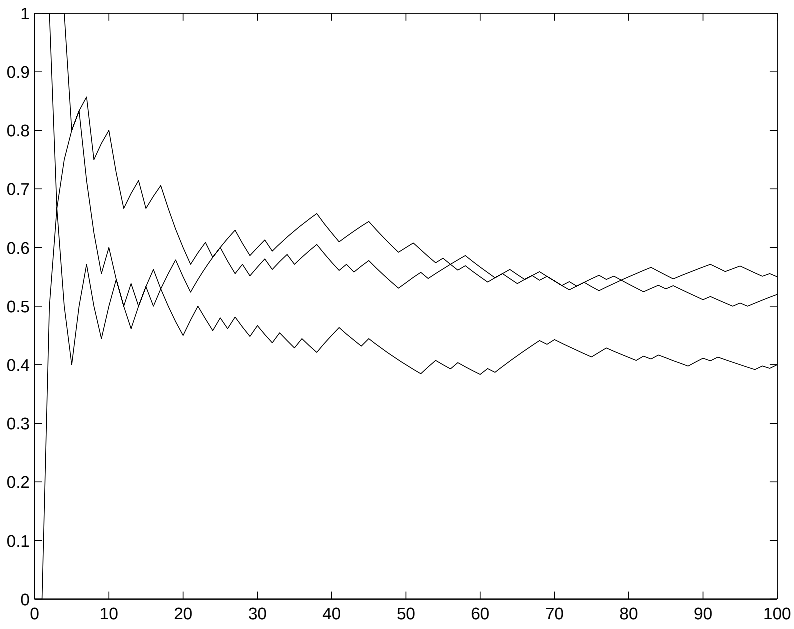


Figure 1.2: The average number of heads in n tosses, where $n = 1, \ldots, 100$.

Example 1.2 (Control Chart) Control charts, see Figure 1.3, are frequently used in manufacturing as a method for *quality control*. Each hour the average output of the process is measured — for example, the average weight of 10 bags of sugar — to assess if the process is still "in control", for example, if the machine still puts on average the correct amount of sugar in the bags. When the process $>$ *Upper Control Limit* or $<$ *Lower Control Limit* and an alarm is raised that the process is out of control, e.g., the machine needs to be adjusted, because it either puts too much or not enough sugar in the bags. The question is how to set the control limits, since the random process naturally fluctuates around its "centre" or "target" line.

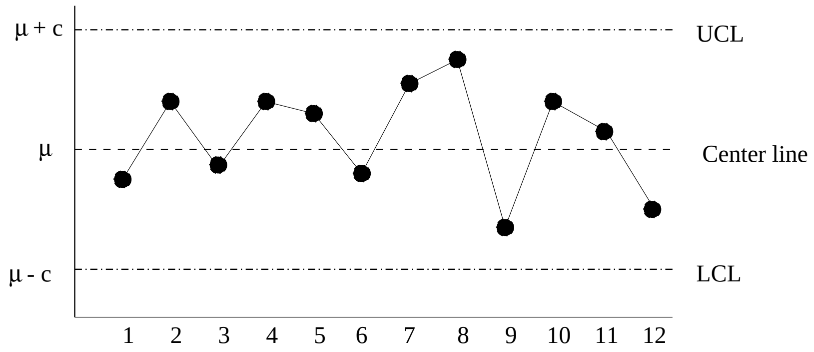


Figure 1.3: Control Chart

Example 1.3 (Machine Lifetime) Suppose 1000 identical components are monitored for failure, up to 50,000 hours. The outcome of such a random experiment is typically summarised via the cumulative lifetime table and plot, as given in Table 1.1 and Figure 1.3, respectively. Here $\hat{F}(t)$ denotes the proportion of components that have failed at time t. One question is how $\hat{F}(t)$ can be modelled via a continuous function F, representing the lifetime distribution of a typical component.

t (h)	failed	$\widehat{F(t)}$	t (h)	failed	$\widehat{F(t)}$
0	0	0.000	3000	140	0.140
750	22	0.020	5000	200	0.200
800	30	0.030	6000	290	0.290
900	36	0.036	8000	350	0.350
1400	42	0.042	11000	540	0.540
1500	58	0.058	15000	570	0.570
2000	74	0.074	19000	770	0.770
2300	105	0.105	37000	920	0.920

Table 1.1: The cumulative lifetime table

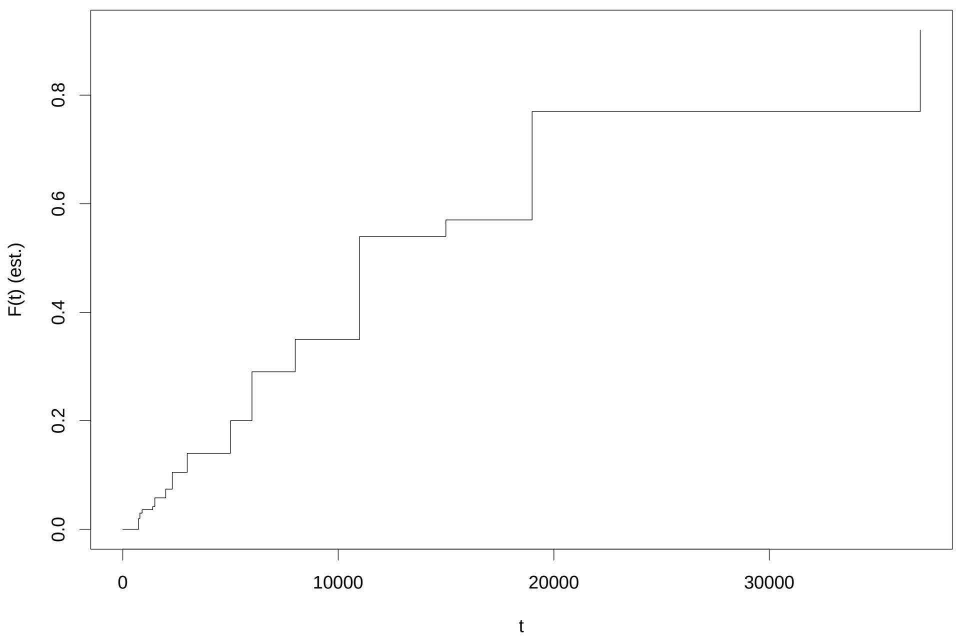


Figure 1.4: The cumulative lifetime table

Example 1.4 A 4-engine aeroplane is able to fly on just one engine on each wing. All engines are unreliable.

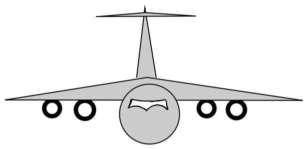


Figure 1.5: A aeroplane with 4 unreliable engines

Number the engines: 1,2 (left wing) and 3,4 (right wing). Observe which engine works properly during a specified period of time. There are $2^4 = 16$ possible outcomes of the experiment. Which outcomes lead to "system failure"? Moreover, if the probability of failure within some time period is known for each of the engines, what is the probability of failure for the entire system? Again this can be viewed as a random experiment.

Below are two more pictures of randomness. The first is a computer-generated "plant", which looks remarkably like a real plant. The second is real data depicting the number of bytes that are transmitted over some communications link. An interesting feature is that the data can be shown to exhibit "fractal" behaviour, that is, if the data is aggregated into smaller or larger time intervals, a similar picture will appear.

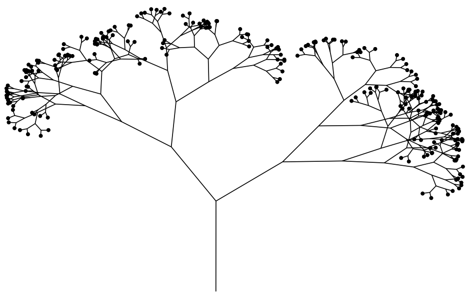


Figure 1.6: Plant growth

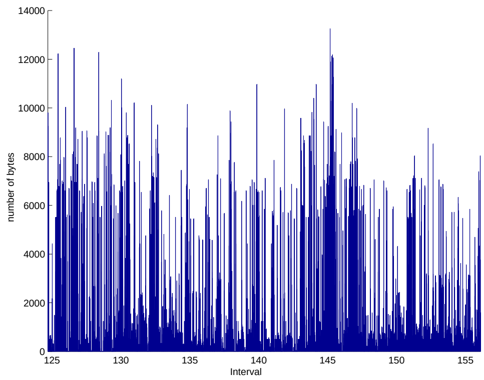


Figure 1.7: Telecommunications data

We wish to describe these experiments via appropriate mathematical models. These models consist of three building blocks: a *sample space*, a set of *events* and a *probability*. We will now describe each of these objects.

1.2 Sample Space

Although we cannot predict the outcome of a random experiment with certainty we usually can specify a set of possible outcomes. This gives the first ingredient in our model for a random experiment.

Definition 1.1 The **sample space** Ω of a random experiment is the set of all possible outcomes of the experiment.

Examples of random experiments with their sample spaces are:

1. Cast two dice consecutively,

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

2. The lifetime of a machine (in days),

$$\Omega = \mathbb{R}_+ = \{ \text{ positive real numbers } \}.$$

3. The number of arriving calls at an exchange during a specified time interval,

$$\Omega = \{0,1,\cdots\} = \mathbb{Z}_+ .$$

4. The heights of 10 selected people.

$$\Omega = \{(x_1,\ldots,x_{10}), x_i \geq 0, i = 1,\ldots,10\} = \mathbb{R}_+^{10} .$$

   Here $(x_1,\ldots,x_{10})$ represents the outcome that the length of the first selected person is x_1, the length of the second person is x_2, et cetera.

Notice that for modelling purposes it is often easier to take the sample space larger than necessary. For example the actual lifetime of a machine would certainly not span the entire positive real axis. And the heights of the 10 selected people would not exceed 3 metres.

1.3 Events

Often we are not interested in a single outcome but in whether or not one of a *group* of outcomes occurs. Such subsets of the sample space are called **events**. Events will be denoted by capital letters $A, B, C, \ldots$. We say that event A **occurs** if the outcome of the experiment is one of the elements in A.

Examples of events are:

1. The event that the sum of two dice is 10 or more,
$$A = \{(4,6),(5,5),(5,6),(6,4),(6,5),(6,6)\}.$$
2. The event that a machine lives less than 1000 days,
$$A = [0,1000)\ .$$
3. The event that out of fifty selected people, five are left-handed,
$$A = \{5\}\ .$$

Example 1.5 (Coin Tossing) Suppose that a coin is tossed 3 times, and that we "record" every head and tail (not only the number of heads or tails). The sample space can then be written as

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}\ ,$$

where, for example, HTH means that the first toss is heads, the second tails, and the third heads. An alternative sample space is the set $\{0,1\}^3$ of binary vectors of length 3, e.g., HTH corresponds to (1,0,1), and THH to (0,1,1).

The event A that the third toss is heads is

$$A = \{HHH, HTH, THH, TTH\}\ .$$

Since events are sets, we can apply the usual set operations to them:

1. the set $A \cup B$ (A **union** B) is the event that A *or* B *or* both occur,
2. the set $A \cap B$ (A **intersection** B) is the event that A *and* B both occur,
3. the event A^c (A **complement**) is the event that A does *not* occur,
4. if $A \subset B$ (A is a **subset** of B) then event A is said to *imply* event B.

Two events A and B which have no outcomes in common, that is, $A \cap B = \emptyset$, are called **disjoint** events.

Example 1.6 Suppose we cast two dice consecutively. The sample space is $\Omega = \{(1,1),(1,2),\ldots,(1,6),(2,1),\ldots,(6,6)\}$. Let $A = \{(6,1),\ldots,(6,6)\}$ be the event that the first die is 6, and let $B = \{(1,6),\ldots,(1,6)\}$ be the event that the second dice is 6. Then $A \cap B = \{(6,1),\ldots,(6,6)\} \cap \{(1,6),\ldots,(6,6)\} = \{(6,6)\}$ is the event that both die are 6.

It is often useful to depict events in a **Venn diagram**, such as in Figure 1.8

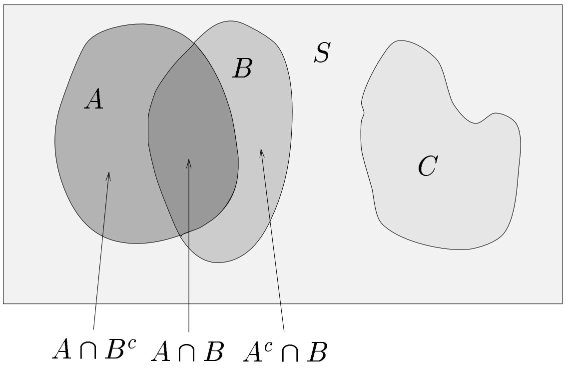


Figure 1.8: A Venn diagram

In this Venn diagram we see

(i) $A \cap C = \emptyset$ and therefore events A and C are disjoint.

(ii) $(A \cap B^c) \cap (A^c \cap B) = \emptyset$ and hence events $A \cap B^c$ and $A^c \cap B$ are disjoint.

Example 1.7 (System Reliability) In Figure 1.9 three systems are depicted, each consisting of 3 unreliable components. The *series* system works if and only if (abbreviated as iff) *all* components work; the *parallel* system works iff *at least one* of the components works; and the 2-out-of-3 system works iff at least 2 out of 3 components work.

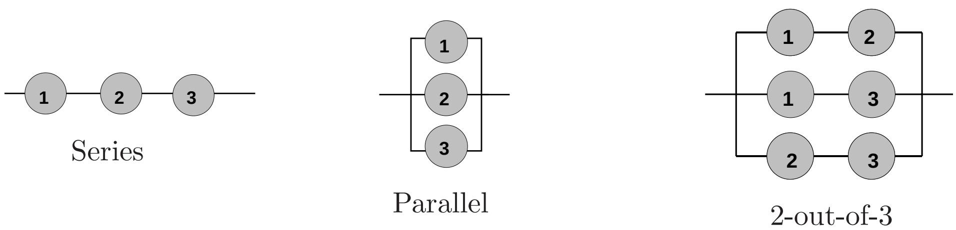


Figure 1.9: Three unreliable systems

Let A_i be the event that the ith component is functioning, $i = 1, 2, 3$; and let D_a, D_b, D_c be the events that respectively the series, parallel and 2-out-of-3 system is functioning. Then,

$$D_a = A_1 \cap A_2 \cap A_3 \ ,$$

and

$$D_b = A_1 \cup A_2 \cup A_3 \ .$$

Also,

$$\begin{aligned} D_c &= (A_1 \cap A_2 \cap A_3) \cup (A_1^c \cap A_2 \cap A_3) \cup (A_1 \cap A_2^c \cap A_3) \cup (A_1 \cap A_2 \cap A_3^c) \\ &= (A_1 \cap A_2) \cup (A_1 \cap A_3) \cup (A_2 \cap A_3) \; . \end{aligned}$$

Two useful results in the theory of sets are the following, due to **De Morgan**: If $\{A_i\}$ is a collection of events (sets) then

$$\left(\bigcup_i A_i\right)^c = \bigcap_i A_i^c \tag{1.1}$$

and

$$\left(\bigcap_i A_i\right)^c = \bigcup_i A_i^c \; . \tag{1.2}$$

This is easily proved via Venn diagrams. Note that if we interpret A_i as the event that a component works, then the left-hand side of (1.1) is the event that the corresponding parallel system is not working. The right hand is the event that at all components are not working. Clearly these two events are the same.

1.4 Probability

The third ingredient in the model for a random experiment is the specification of the probability of the events. It tells us how *likely* it is that a particular event will occur.

Definition 1.2 A probability $\mathbb{P}$ is a rule (function) which assigns a positive number to each event, and which satisfies the following **axioms**:

Axiom 1: $\mathbb{P}(A) \geq 0$.
Axiom 2: $\mathbb{P}(\Omega) = 1$.
Axiom 3: For any sequence $A_1, A_2, \ldots$ of *disjoint* events we have

$$\boxed{\mathbb{P}(\bigcup_i A_i) = \sum_i \mathbb{P}(A_i)} \; . \tag{1.3}$$

Axiom 2 just states that the probability of the "certain" event Ω is 1. Property (1.3) is the *crucial* property of a probability, and is sometimes referred to as the **sum rule**. It just states that if an event can happen in a number of different ways *that cannot happen at the same time*, then the probability of this event is simply the sum of the probabilities of the composing events.

Note that a probability rule $\mathbb{P}$ has exactly the same properties as the common "area measure". For example, the total area of the union of the triangles in Figure 1.10 is equal to the sum of the areas of the individual triangles. This

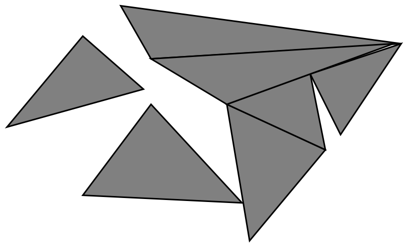


Figure 1.10: The probability measure has the same properties as the "area" measure: the total area of the triangles is the sum of the areas of the idividual triangles.

is how you should interpret property (1.3). But instead of measuring areas, $\mathbb{P}$ measures probabilities.

As a direct consequence of the axioms we have the following properties for $\mathbb{P}$.

Theorem 1.1 Let A and B be events. Then,

1. $\mathbb{P}(\emptyset) = 0$.
2. $A \subset B \Longrightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$.
3. $\mathbb{P}(A) \leq 1$.
4. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
5. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Proof.

1. $\Omega = \Omega \cap \emptyset \cap \emptyset \cap \cdots$, therefore, by the sum rule, $\mathbb{P}(\Omega) = \mathbb{P}(\Omega) + \mathbb{P}(\emptyset) + \mathbb{P}(\emptyset) + \cdots$, and therefore, by the second axiom, $1 = 1 + \mathbb{P}(\emptyset) + \mathbb{P}(\emptyset) + \cdots$, from which it follows that $\mathbb{P}(\emptyset) = 0$.
2. If $A \subset B$, then $B = A \cup (B \cap A^c)$, where A and $B \cap A^c$ are disjoint. Hence, by the sum rule, $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \cap A^c)$, which is (by the first axiom) greater than or equal to $\mathbb{P}(A)$.
3. This follows directly from property 2 and axiom 2, since $A \subset \Omega$.
4. $\Omega = A \cup A^c$, where A and A^c are disjoint. Hence, by the sum rule and axiom 2: $1 = \mathbb{P}(\Omega) = \mathbb{P}(A) + \mathbb{P}(A^c)$, and thus $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
5. Write $A \cup B$ as the disjoint union of A and $B \cap A^c$. Then, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \cap A^c)$. Also, $B = (A \cap B) \cup (B \cap A^c)$, so that $\mathbb{P}(B) = \mathbb{P}(A \cap B) + \mathbb{P}(B \cap A^c)$. Combining these two equations gives $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

■

We have now completed our model for a random experiment. It is up to the modeller to specify the sample space Ω and probability measure $\mathbb{P}$ which most closely describes the actual experiment. This is not always as straightforward as it looks, and sometimes it is useful to model only certain *observations* in the experiment. This is where *random variables* come into play, and we will discuss these in the next chapter.

Example 1.8 Consider the experiment where we throw a fair die. How should we define Ω and $\mathbb{P}$?

Obviously, $\Omega = \{1, 2, \ldots, 6\}$; and some common sense shows that we should define $\mathbb{P}$ by

$$\mathbb{P}(A) = \frac{|A|}{6}, \quad A \subset \Omega,$$

where $|A|$ denotes the number of elements in set A. For example, the probability of getting an even number is $\mathbb{P}(\{2,4,6\}) = 3/6 = 1/2$.

In many applications the sample space is *countable*, i.e. $\Omega = \{a_1, a_2, \ldots, a_n\}$ or $\Omega = \{a_1, a_2, \ldots\}$. Such a sample space is called **discrete**.

The easiest way to specify a probability $\mathbb{P}$ on a discrete sample space is to specify first the probability p_i of each **elementary event** $\{a_i\}$ and then to define

$$\mathbb{P}(A) = \sum_{i:a_i \in A} p_i \ , \quad \text{for all } \ A \subset \Omega.$$

This idea is graphically represented in Figure 1.11. Each element a_i in the sample is assigned a probability weight p_i represented by a black dot. To find the probability of the set A we have to sum up the weights of all the elements in A.

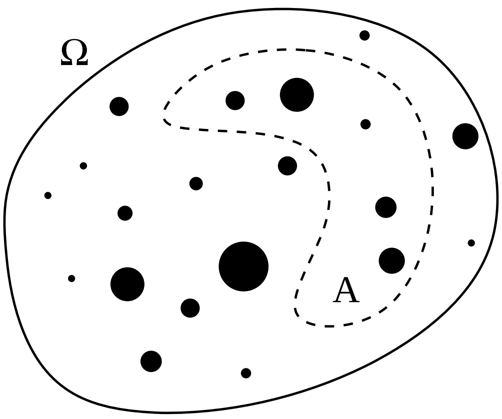


Figure 1.11: A discrete sample space

Again, it is up to the modeller to properly specify these probabilities. Fortunately, in many applications all elementary events are *equally likely*, and thus the probability of each elementary event is equal to 1 divided by the total number of elements in Ω. E.g., in Example 1.8 each elementary event has probability $1/6$.

Because the "equally likely" principle is so important, we formulate it as a theorem.

Theorem 1.2 (Equilikely Principle) If Ω has a finite number of outcomes, and all are equally likely, then the probability of each event A is defined as

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

Thus for such sample spaces the calculation of probabilities reduces to *counting* the number of outcomes (in A and Ω).

When the sample space is not countable, for example $\Omega = \mathbb{R}_+$, it is said to be **continuous**.

Example 1.9 We draw at random a point in the interval $[0, 1]$. Each point is equally likely to be drawn. How do we specify the model for this experiment?

The sample space is obviously $\Omega = [0, 1]$, which is a continuous sample space. We cannot define $\mathbb{P}$ via the elementary events $\{x\}$, $x \in [0, 1]$ because each of these events must have probability 0 (!). However we can define $\mathbb{P}$ as follows: For each $0 \leq a \leq b \leq 1$, let

$$\mathbb{P}([a, b]) = b - a \ .$$

This completely specifies $\mathbb{P}$. In particular, we can find the probability that the point falls into any (sufficiently nice) set A as the *length* of that set.

1.5 Counting

Counting is not always easy. Let us first look at some examples:

1. A multiple choice form has 20 questions; each question has 3 choices. In how many possible ways can the exam be completed?

2. Consider a horse race with 8 horses. How many ways are there to gamble on the placings (1st, 2nd, 3rd).

3. Jessica has a collection of 20 CDs, she wants to take 3 of them to work. How many possibilities does she have?

4. How many different throws are possible with 3 dice?

To be able to comfortably solve a multitude of counting problems requires a lot of experience and *practice*, and even then, some counting problems remain exceedingly hard. Fortunately, many counting problems can be cast into the simple framework of drawing balls from an urn, see Figure 1.12.

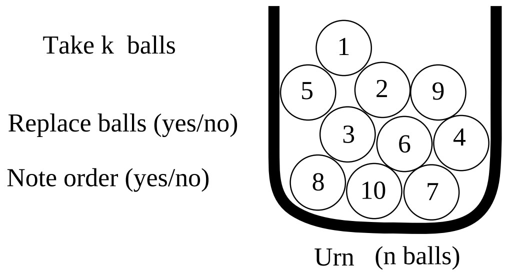


Figure 1.12: An urn with n balls

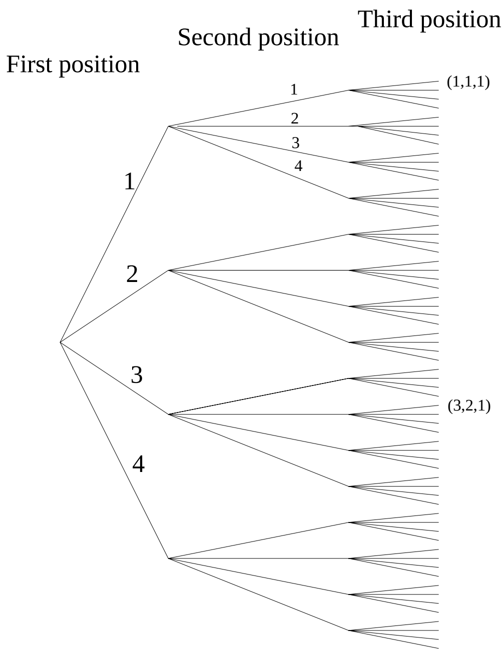
Consider an urn with n different balls, numbered $1, \ldots, n$ from which k balls are drawn. This can be done in a number of different ways. First, the balls can be drawn one-by-one, or one could draw all the k balls at the same time. In the first case the **order** in which the balls are drawn can be noted, in the second case that is not possible. In the latter case we can (and will) still assume the balls are drawn one-by-one, but that the order is not noted. Second, once a ball is drawn, it can either be put back into the urn (after the number is recorded), or left out. This is called, respectively, drawing with and without **replacement**. All in all there are 4 possible experiments: (ordered, with replacement), (ordered, without replacement), (unordered, without replacement) and (ordered, with replacement). The art is to recognise a seemingly unrelated counting problem as one of these four urn problems. For the 4 examples above we have the following

1. Example 1 above can be viewed as drawing 20 balls from an urn containing 3 balls, noting the order, and with replacement.
2. Example 2 is equivalent to drawing 3 balls from an urn containing 8 balls, noting the order, and without replacement.
3. In Example 3 we take 3 balls from an urn containing 20 balls, not noting the order, and without replacement
4. Finally, Example 4 is a case of drawing 3 balls from an urn containing 6 balls, not noting the order, and with replacement.

Before we proceed it is important to introduce a notation that reflects whether the outcomes/arrangements are ordered or not. In particular, we denote ordered arrangements by *vectors*, e.g., $(1, 2, 3) \neq (3, 2, 1)$, and unordered arrangements by *sets*, e.g., $\{1,2,3\} = \{3,2,1\}$. We now consider for each of the four cases how to count the number of arrangements. For simplicity we consider for each case how the counting works for $n = 4$ and $k = 3$, and then state the general situation.

Drawing with Replacement, Ordered

Here, after we draw each ball, note the number on the ball, and put the ball back. Let $n = 4, k = 3$. Some possible outcomes are $(1,1,1), (4,1,2), (2,3,2), (4,2,1), \ldots$ To count how many such arrangements there are, we can reason as follows: we have three positions $(\cdot, \cdot, \cdot)$ to fill in. Each position can have the numbers 1,2,3 or 4, so the total number of possibilities is $4 \times 4 \times 4 = 4^3 = 64$. This is illustrated via the following tree diagram:

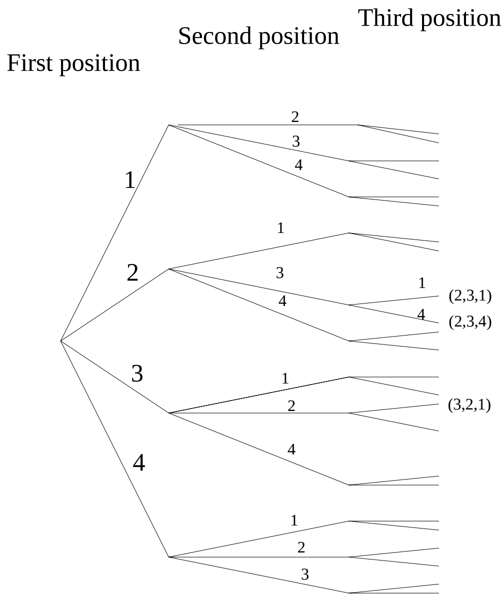


For general n and k we can reason analogously to find:

> The number of ordered arrangements of k numbers chosen from $\{1,\ldots,n\}$, with replacement (repetition) is n^k.

Drawing Without Replacement, Ordered

Here we draw again k numbers (balls) from the set $\{1,2,\ldots,n\}$, and note the order, but now do not replace them. Let $n = 4$ and $k = 3$. Again there are 3 positions to fill $(\cdot, \cdot, \cdot)$, but now the numbers cannot be the same, e.g., (1,4,2),(3,2,1), etc. Such an ordered arrangements called a **permutation** of size k from set $\{1,\ldots,n\}$. (A permutation of $\{1,\ldots,n\}$ of size n is simply called a permutation of $\{1,\ldots,n\}$ (leaving out "of size n"). For the 1st position we have 4 possibilities. Once the first position has been chosen, we have only 3 possibilities left for the second position. And after the first two positions have been chosen there are 2 positions left. So the number of arrangements is $4\times3\times2 = 24$ as illustrated in Figure 1.5, which is the same tree as in Figure 1.5, but with all "duplicate" branches removed.



For general n and k we have:

> The number of permutations of size k from $\{1,\ldots,n\}$ is ${}^nP_k = n(n-1)\cdots(n-k+1)$.

In particular, when $k = n$, we have that the number of ordered arrangements of n items is $n! = n(n-1)(n-2)\cdots 1$, where $n!$ is called n-**factorial**. Note that

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

Drawing Without Replacement, Unordered

This time we draw k numbers from $\{1,\ldots,n\}$ but do not replace them (no replication), and do not note the order (so we could draw them in one grab). Taking again $n = 4$ and $k = 3$, a possible outcome is $\{1,2,4\}$, $\{1,2,3\}$, etc. If we noted the order, there would be ${}^nP_k$ outcomes, amongst which would be (1,2,4),(1,4,2),(2,1,4),(2,4,1),(4,1,2) and (4,2,1). Notice that these 6 permutations correspond to the single unordered arrangement $\{1,2,4\}$. Such unordered arrangements without replications are called **combinations** of size k from the set $\{1, \ldots, n\}$.

To determine the number of combinations of size k simply need to divide ${}^nP_k$ be the number of permutations of k items, which is $k!$. Thus, in our example $(n = 4, k = 3)$ there are $24/6 = 4$ possible combinations of size 3. In general we have:

The number of combinations of size k from the set $\{1, \ldots n\}$ is

$$ {}^nC_k = \binom{n}{k} = \frac{{}^nP_k}{k!} = \frac{n!}{(n-k)!\,k!} \,. $$

Note the two different notations for this number. We will use the second one.

Drawing With Replacement, Unordered

Taking $n = 4, k = 3$, possible outcomes are $\{3,3,4\}$, $\{1,2,4\}, \{2,2,2\}$, etc. The trick to solve this counting problem is to represent the outcomes in a different way, via an ordered vector $(x_1, \ldots, x_n)$ representing how many times an element in $\{1, \ldots, 4\}$ occurs. For example, $\{3,3,4\}$ corresponds to $(0,0,2,1)$ and $\{1,2,4\}$ corresponds to $(1,1,0,1)$. Thus, we can count how many distinct vectors $(x_1, \ldots, x_n)$ there are such that the sum of the components is 3, and each x_i can take value 0,1,2 or 3. Another way of looking at this is to consider placing $k = 3$ balls into $n = 4$ urns, numbered 1,. . . ,4. Then $(0,0,2,1)$ means that the third urn has 2 balls and the fourth urn has 1 ball. One way to distribute the balls over the urns is to distribute $n - 1 = 3$ "separators" and $k = 3$ balls over $n - 1 + k = 6$ positions, as indicated in Figure 1.13.

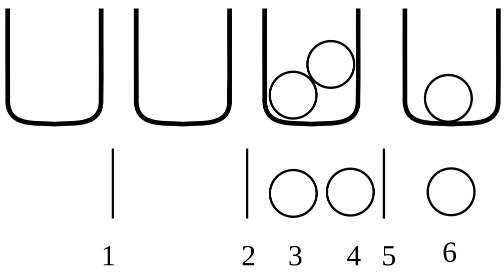


Figure 1.13: distributing k balls over n urns

The number of ways this can be done is the equal to the number of ways k positions for the balls can be chosen out of $n - 1 + k$ positions, that is, $\binom{n+k-1}{k}$. We thus have:

The number of different sets $\{x_1, \ldots, x_k\}$ with $x_i \in \{1, \ldots, n\}$, $i = 1, \ldots, k$ is

$$ \binom{n+k-1}{k} \,. $$

Returning to our original four problems, we can now solve them easily:

1. The total number of ways the exam can be completed is $3^{20} = 3,486,784,401$.
2. The number of placings is $^8P_3 = 336$.
3. The number of possible combinations of CDs is $\binom{20}{3} = 1140$.
4. The number of different throws with three dice is $\binom{8}{3} = 56$.

More examples

Here are some more examples. Not all problems can be directly related to the 4 problems above. Some require additional reasoning. However, the counting principles remain the same.

1. In how many ways can the numbers 1,...,5 be arranged, such as 13524, 25134, etc?

   Answer: $5! = 120$.

2. How many different arrangements are there of the numbers 1,2,...,7, such that the first 3 numbers are 1,2,3 (in any order) and the last 4 numbers are 4,5,6,7 (in any order)?

   Answer: $3! \times 4!$.

3. How many different arrangements are there of the word "arrange", such as "aarrnge", "arrngea", etc?

   Answer: Convert this into a ball drawing problem with 7 balls, numbered 1,...,7. Balls 1 and 2 correspond to 'a', balls 3 and 4 to 'r', ball 5 to 'n', ball 6 to 'g' and ball 7 to 'e'. The total number of permutations of the numbers is 7!. However, since, for example, (1,2,3,4,5,6,7) is identical to (2,1,3,4,5,6,7) (when substituting the letters back), we must divide 7! by $2! \times 2!$ to account for the 4 ways the two 'a's and 'r's can be arranged. So the answer is $7!/4 = 1260$.

4. An urn has 1000 balls, labelled 000, 001, ..., 999. How many balls are there that have all number in ascending order (for example 047 and 489, but not 033 or 321)?

   Answer: There are $10 \times 9 \times 8 = 720$ balls with different numbers. Each triple of numbers can be arranged in $3! = 6$ ways, and only one of these is in ascending order. So the total number of balls in ascending order is $720/6 = 120$.

5. In a group of 20 people each person has a different birthday. How many different arrangements of these birthdays are there (assuming each year has 365 days)?

   Answer: $^{365}P_{20}$.

Once we've learned how to count, we can apply the equilikely principle to calculate probabilities:

1. What is the probability that out of a group of 40 people all have different birthdays?

   Answer: Choosing the birthdays is like choosing 40 balls with replacement from an urn containing the balls 1,...,365. Thus, our sample space Ω consists of vectors of length 40, whose components are chosen from $\{1,\ldots,365\}$. There are $|\Omega| = 365^{40}$ such vectors possible, and all are *equally likely*. Let A be the event that all 40 people have different birthdays. Then, $|A| = {}^{365}P_{40} = 365!/325!$ It follows that $\mathbb{P}(A) = |A|/|\Omega| \approx 0.109$, so not very big!

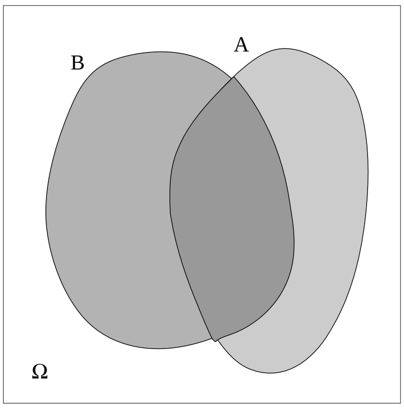
2. What is the probability that in 10 tosses with a fair coin we get exactly 5 Heads and 5 Tails?

   Answer: Here Ω consists of vectors of length 10 consisting of 1s (Heads) and 0s (Tails), so there are 2^{10} of them, and all are *equally likely*. Let A be the event of exactly 5 heads. We must count how many binary vectors there are with exactly 5 1s. This is equivalent to determining in how many ways the positions of the 5 1s can be chosen out of 10 positions, that is, $\binom{10}{5}$. Consequently, $\mathbb{P}(A) = \binom{10}{5}/2^{10} = 252/1024 \approx 0.25$.

3. We draw at random 13 cards from a full deck of cards. What is the probability that we draw 4 Hearts and 3 Diamonds?

   Answer: Give the cards a number from 1 to 52. Suppose 1–13 is Hearts, 14–26 is Diamonds, etc. Ω consists of unordered sets of size 13, without repetition, e.g., $\{1,2,\ldots,13\}$. There are $|\Omega| = \binom{52}{13}$ of these sets, and they are all equally likely. Let A be the event of 4 Hearts and 3 Diamonds. To form A we have to choose 4 Hearts out of 13, and 3 Diamonds out of 13, followed by 6 cards out of 26 Spade and Clubs. Thus, $|A| = \binom{13}{4} \times \binom{13}{3} \times \binom{26}{6}$. So that $\mathbb{P}(A) = |A|/|\Omega| \approx 0.074$.

1.6 Conditional probability and independence



How do probabilities change when we know some event $B \subset \Omega$ has occurred? Suppose B has occurred. Thus, we know that the outcome lies in B. Then A will occur if and only if $A \cap B$ occurs, and the relative chance of A occurring is therefore

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

This leads to the definition of the **conditional probability** of A given B:

$$\boxed{\mathbb{P}(A \,|\, B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}} \tag{1.4}$$

Example 1.10 We throw two dice. Given that the sum of the eyes is 10, what is the probability that one 6 is cast?

Let B be the event that the sum is 10,

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

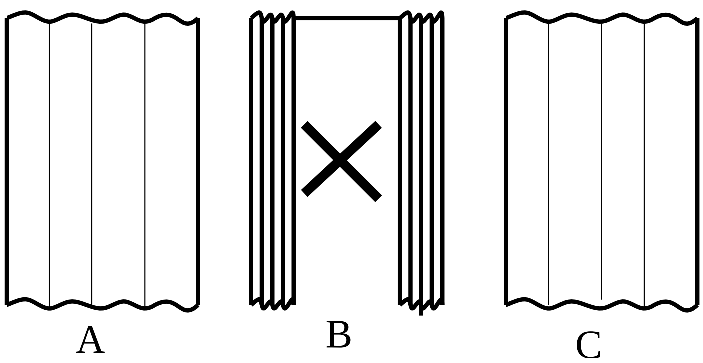
Let A be the event that one 6 is cast,

$$A = \{(1,6), \ldots, (5,6), (6,1), \ldots, (6,5)\}.$$

Then, $A \cap B = \{(4,6), (6,4)\}$. And, since all elementary events are equally likely, we have

$$\mathbb{P}(A \,|\, B) = \frac{2/36}{3/36} = \frac{2}{3}.$$

Example 1.11 (Monte Hall Problem) This is a nice application of conditional probability. Consider a quiz in which the final contestant is to choose a prize which is hidden behind one three curtains (A, B or C). Suppose without loss of generality that the contestant chooses curtain A. Now the quiz master (Monte Hall) always opens one of the other curtains: if the prize is behind B, Monte opens C, if the prize is behind C, Monte opens B, and if the prize is behind A, Monte opens B or C with equal probability, e.g., by tossing a coin (of course the contestant does not see Monte tossing the coin!).



Suppose, again without loss of generality that Monte opens curtain B. The contestant is now offered the opportunity to switch to curtain C. Should the contestant stay with his/her original choice (A) or switch to the other unopened curtain (C)?

Notice that the sample space consists here of 4 possible outcomes: Ac: The prize is behind A, and Monte opens C; Ab: The prize is behind A, and Monte opens B; Bc: The prize is behind B, and Monte opens C; and Cb: The prize is behind C, and Monte opens B. Let A, B, C be the events that the prize is behind A, B and C, respectively. Note that $A = \{Ac, Ab\}$, $B = \{Bc\}$ and $C = \{Cb\}$, see Figure 1.14.

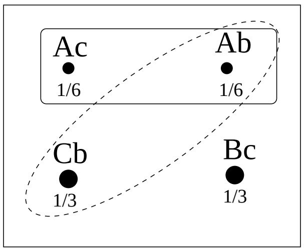


Figure 1.14: The sample space for the Monte Hall problem.

Now, obviously $\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C)$, and since Ac and Ab are equally likely, we have $\mathbb{P}(\{Ab\}) = \mathbb{P}(\{Ac\}) = 1/6$. Monte opening curtain B means that we have information that event $\{Ab, Cb\}$ has occurred. The probability that the prize is under A given this event, is therefore

$$\mathbb{P}(A \,|\, \text{B is opened}) = \frac{\mathbb{P}(\{Ac, Ab\} \cap \{Ab, Cb\})}{\mathbb{P}(\{Ab, Cb\})} = \frac{\mathbb{P}(\{Ab\})}{\mathbb{P}(\{Ab, Cb\})} = \frac{1/6}{1/6 + 1/3} = \frac{1}{3}.$$

This is what we expected: the fact that Monte opens a curtain does not give us any extra information that the prize is behind A. So one could think that it doesn't matter to switch or not. But wait a minute! What about $\mathbb{P}(B \,|\, \text{B is opened})$? Obviously this is 0 — opening curtain B means that we know that event B cannot occur. It follows then that $\mathbb{P}(C \,|\, \text{B is opened})$ must be 2/3, since a conditional probability behaves like any other probability and must satisfy axiom 2 (sum up to 1). Indeed,

$$\mathbb{P}(C \,|\, \text{B is opened}) = \frac{\mathbb{P}(\{Cb\} \cap \{Ab, Cb\})}{\mathbb{P}(\{Ab, Cb\})} = \frac{\mathbb{P}(\{Cb\})}{\mathbb{P}(\{Ab, Cb\})} = \frac{1/3}{1/6 + 1/3} = \frac{2}{3}.$$

Hence, given the information that B is opened, it is twice as likely that the prize is under C than under A. Thus, the contestant should switch!

1.6.1 Product Rule

By the definition of conditional probability we have

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\, \mathbb{P}(B \,|\, A). \qquad (1.5)$$

We can generalise this to n intersections $A_1 \cap A_2 \cap \cdots \cap A_n$, which we abbreviate as $A_1 A_2 \cdots A_n$. This gives the **product rule** of probability (also called *chain rule*).

Theorem 1.3 (Product rule) Let $A_1, \ldots, A_n$ be a sequence of events with $\mathbb{P}(A_1 \ldots A_{n-1}) > 0$. Then,

$$\boxed{\mathbb{P}(A_1 \cdots A_n) = \mathbb{P}(A_1)\, \mathbb{P}(A_2 \,|\, A_1)\, \mathbb{P}(A_3 \,|\, A_1 A_2) \cdots \mathbb{P}(A_n \,|\, A_1 \cdots A_{n-1}).} \qquad (1.6)$$

PROOF. We only show the proof for 3 events, since the $n > 3$ event case follows similarly. By applying (1.4) to $\mathbb{P}(B \mid A)$ and $\mathbb{P}(C \mid A \cap B)$, the left-hand side of (1.6) is we have,

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

which is equal to the left-hand size of (1.6). ■

Example 1.12 We draw consecutively 3 balls from a bowl with 5 white and 5 black balls, without putting them back. What is the probability that all balls will be black?

Solution: Let A_i be the event that the ith ball is black. We wish to find the probability of $A_1A_2A_3$, which by the product rule (1.6) is

$$\mathbb{P}(A_1)\,\mathbb{P}(A_2 \mid A_1)\,\mathbb{P}(A_3 \mid A_1A_2) = \frac{5}{10}\,\frac{4}{9}\,\frac{3}{8} = 0.083.$$

Note that this problem can also be easily solved by counting arguments, as in the previous section.

Example 1.13 (Birthday Problem) In Section 1.5 we derived by counting arguments that the probability that all people in a group of 40 have different birthdays is

$$\frac{365 \times 364 \times \cdots \times 326}{365 \times 365 \times \cdots \times 365} \approx 0.109. \tag{1.7}$$

We can derive this also via the product rule. Namely, let A_i be the event that the first i people have different birthdays, $i = 1, 2, \ldots$. Note that $A_1 \supset A_2 \supset A_3 \supset \cdots$. Therefore $A_n = A_1 \cap A_2 \cap \cdots \cap A_n$, and thus by the product rule

$$\mathbb{P}(A_{40}) = \mathbb{P}(A_1)\mathbb{P}(A_2 \mid A_1)\mathbb{P}(A_3 \mid A_2)\cdots\mathbb{P}(A_{40} \mid A_{39})\,.$$

Now $\mathbb{P}(A_k \mid A_{k-1} = (365 - k + 1)/365$ because given that the first $k-1$ people have different birthdays, there are no duplicate birthdays if and only if the birthday of the k-th is chosen from the $365 - (k-1)$ remaining birthdays. Thus, we obtain (1.7). More generally, the probability that n randomly selected people have different birthdays is

$$\mathbb{P}(A_n) = \frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times \cdots \times \frac{365 - n + 1}{365}, \quad n \geq 1\,.$$

A graph of $\mathbb{P}(A_n)$ against n is given in Figure 1.15. Note that the probability $\mathbb{P}(A_n)$ rapidly decreases to zero. Indeed, for $n = 23$ the probability of having no duplicate birthdays is already less than 1/2.

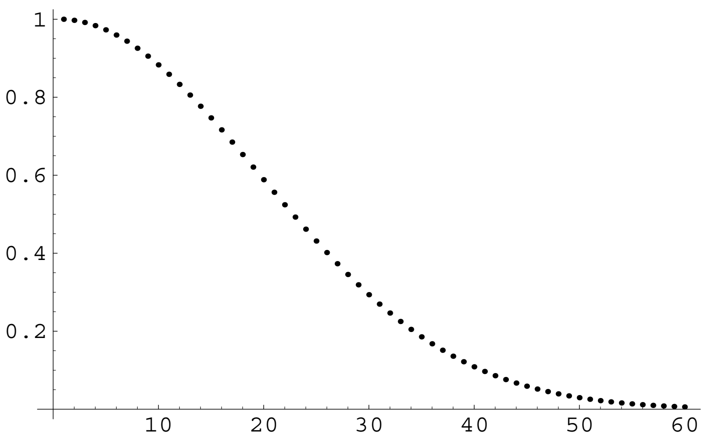


Figure 1.15: The probability of having no duplicate birthday in a group of n people, against n.

1.6.2 Law of Total Probability and Bayes' Rule

Suppose $B_1, B_2, \ldots, B_n$ is a **partition** of Ω. That is, $B_1, B_2, \ldots, B_n$ are disjoint and their union is Ω, see Figure 1.16

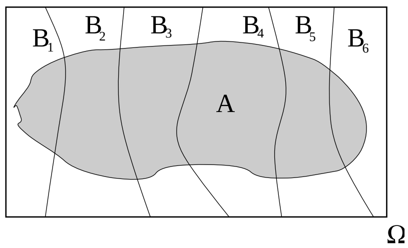


Figure 1.16: A partition of the sample space

Then, by the sum rule, $\mathbb{P}(A) = \sum_{i=1}^n \mathbb{P}(A \cap B_i)$ and hence, by the definition of conditional probability we have

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

This is called the **law of total probability**.

Combining the Law of Total Probability with the definition of conditional probability gives **Bayes' Rule**:

$$\boxed{\mathbb{P}(B_j|A) = \frac{\mathbb{P}(A|B_j)\,\mathbb{P}(B_j)}{\sum_{i=1}^n \mathbb{P}(A|B_i)\mathbb{P}(B_i)}}$$

Example 1.14 A company has three factories (1, 2 and 3) that produce the same chip, each producing 15%, 35% and 50% of the total production. The probability of a defective chip at 1, 2, 3 is 0.01, 0.05, 0.02, respectively. Suppose someone shows us a defective chip. What is the probability that this chip comes from factory 1?

Let B_i denote the event that the chip is produced by factory i. The $\{\mathcal{B}_\rangle\}$ form a partition of Ω. Let A denote the event that the chip is faulty. By Bayes' rule,

$$\mathbb{P}(B_1 \,|\, A) = \frac{0.15 \times 0.01}{0.15 \times 0.01 + 0.35 \times 0.05 + 0.5 \times 0.02} = 0.052 \ .$$

1.6.3 Independence

Independence is a very important concept in probability and statistics. Loosely speaking it models the *lack of information* between events. We say A and B are *independent* if the knowledge that A has occurred does not change the *probability* that B occurs. That is

$$A, B \text{ independent} \Leftrightarrow \mathbb{P}(A|B) = \mathbb{P}(A)$$

Since $\mathbb{P}(A|B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$ an alternative definition of independence is

$$\boxed{A, B \text{ independent} \Leftrightarrow \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)}$$

This definition covers the case $B = \emptyset$ (empty set). We can extend the definition to arbitrarily many events:

Definition 1.3 The events $A_1, A_2, \ldots,$ are said to be **(mutually) independent** if for any n and any choice of distinct indices $i_1, \ldots, i_k$,

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

Remark 1.1 In most cases independence of events is a **model assumption**. That is, we assume that there exists a $\mathbb{P}$ such that certain events are independent.

Example 1.15 (A Coin Toss Experiment and the Binomial Law) We flip a coin n times. We can write the sample space as the set of binary n-tuples:

$$\Omega = \{(0, \ldots, 0), \ldots, (1, \ldots, 1)\} \ .$$

Here 0 represent Tails and 1 represents Heads. For example, the outcome $(0, 1, 0, 1, \ldots)$ means that the first time Tails is thrown, the second time Heads, the third times Tails, the fourth time Heads, etc.

How should we define $\mathbb{P}$? Let A_i denote the event of Heads during the ith throw, $i = 1, \ldots, n$. Then, $\mathbb{P}$ should be such that the events $A_1, \ldots, A_n$ are *independent*. And, moreover, $\mathbb{P}(A_i)$ should be the same for all i. We don't know whether the coin is fair or not, but we can call this probability p $(0 \leq p \leq 1)$.

These two rules completely specify $\mathbb{P}$. For example, the probability that the first k throws are Heads and the last $n-k$ are Tails is

$$\begin{aligned}\mathbb{P}(\{(1,1,\ldots,1,0,0,\ldots,0)\}) &= \mathbb{P}(A_1)\cdots\mathbb{P}(A_k)\cdots\mathbb{P}(A_{k+1}^c)\cdots\mathbb{P}(A_n^c)\\ &= p^k(1-p)^{n-k}.\end{aligned}$$

Also, let B_k be the event that there are k Heads in total. The probability of this event is the sum the probabilities of elementary events $\{(x_1,\ldots,x_n)\}$ such that $x_1+\cdots+x_n=k$. Each of these events has probability $p^k(1-p)^{n-k}$, and there are $\binom{n}{k}$ of these. Thus,

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

We have thus discovered the **binomial distribution**.

Example 1.16 (Geometric Law) There is another important law associated with the coin flip experiment. Suppose we flip the coin until Heads appears for the first time. Let C_k be the event that Heads appears for the first time at the k-th toss, $k=1,2,\ldots$. Then, using the same events $\{A_i\}$ as in the previous example, we can write

$$C_k = A_1^c \cap A_2^c \cap \cdots \cap A_{k-1}^c \cap A_k,$$

so that with the product law and the mutual independence of $A_1^c,\ldots,A_k$ we have the **geometric law**:

$$\begin{aligned}\mathbb{P}(C_k) &= \mathbb{P}(A_1^c)\cdots\mathbb{P}(A_{k-1}^c)\,\mathbb{P}(A_k)\\ &= \underbrace{(1-p)\cdots(1-p)}_{k-1 \text{ times}}p = (1-p)^{k-1}\,p\,.\end{aligned}$$

Chapter 2

Random Variables and Probability Distributions

Specifying a model for a random experiment via a complete description of Ω and $\mathbb{P}$ may not always be convenient or necessary. In practice we are only interested in various *observations* (i.e., numerical measurements) of the experiment. We include these into our modelling process via the introduction of *random variables.*

2.1 Random Variables

Formally a **random variable** is a *function* from the sample space Ω to $\mathbb{R}$. Here is a concrete example.

Example 2.1 (Sum of two dice) Suppose we toss two fair dice and note their sum. If we throw the dice one-by-one and observe each throw, the sample space is $\Omega = \{(1,1), \ldots, (6,6)\}$. The function X, defined by $X(i,j) = i + j$, is a random variable, which maps the outcome (i,j) to the sum $i + j$, as depicted in Figure 2.1. Note that all the outcomes in the "encircled" set are mapped to 8. This is the set of all outcomes whose sum is 8. A natural notation for this set is to write $\{X = 8\}$. Since this set has 5 outcomes, and all outcomes in Ω are equally likely, we have

$$\mathbb{P}(\{X = 8\}) = \frac{5}{36}.$$

This notation is very suggestive and convenient. From a non-mathematical viewpoint we can interpret X as a "random" variable. That is a variable that can take on several values, with certain probabilities. In particular it is not difficult to check that

$$\mathbb{P}(\{X = x\}) = \frac{6 - |7 - x|}{36}, \quad x = 2, \ldots, 12.$$

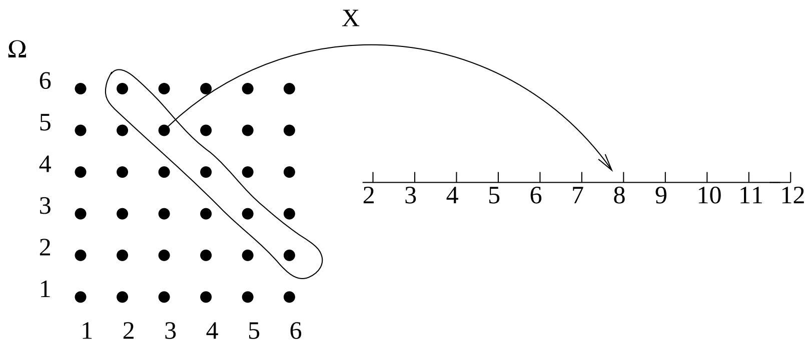


Figure 2.1: A random variable representing the sum of two dice

Although random variables are, mathematically speaking, *functions*, it is often convenient to view random variables as observations of a random experiment that has not yet been carried out. In other words, a random variable is considered as a measurement that becomes available once we carry out the random experiment, e.g., *tomorrow*. However, all the *thinking* about the experiment and measurements can be done *today*. For example, we can specify today exactly the probabilities pertaining to the random variables.

We usually denote random variables with *capital* letters from the last part of the alphabet, e.g. X, $X_1, X_2, \ldots, Y, Z$. Random variables allow us to use natural and intuitive notations for certain events, such as $\{X = 10\}$, $\{X > 1000\}$, $\{\max(X, Y) \leq Z\}$, etc.

Example 2.2 We flip a coin n times. In Example 1.15 we can find a probability model for this random experiment. But suppose we are not interested in the complete outcome, e.g., (0,1,0,1,1,0...), but only in the total number of heads (1s). Let X be the total number of heads. X is a “random variable” in the true sense of the word: X could lie anywhere between 0 and n. What we are interested in, however, is the *probability* that X takes certain values. That is, we are interested in the **probability distribution** of X. Example 1.15 now suggests that

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n\,. \tag{2.1}$$

This contains all the information about X that we could possibly wish to know. Example 2.1 suggests how we can justify this mathematically Define X as the function that assigns to each outcome $\omega = (x_1, \ldots, x_n)$ the number $x_1 + \cdots + x_n$. Then clearly X is a random variable in mathematical terms (that is, a function). Moreover, the event that there are exactly k Heads in n throws can be written as

$$\{\omega \in \Omega : X(\omega) = k\}.$$

If we abbreviate this to $\{X = k\}$, and further abbreviate $\mathbb{P}(\{X = k\})$ to $\mathbb{P}(X = k)$, then we obtain exactly (2.1).

We give some more examples of random variables without specifying the sample space.

1. The number of defective transistors out of 100 inspected ones,
2. the number of bugs in a computer program,
3. the amount of rain in Brisbane in June,
4. the amount of time needed for an operation.

The set of all possible values a random variable X can take is called the **range** of X. We further distinguish between discrete and continuous random variables:

Discrete random variables can only take *isolated* values.

For example: a count can only take non-negative integer values.

Continuous random variables can take values in an *interval*.

For example: rainfall measurements, lifetimes of components, lengths, ... are (at least in principle) continuous.

2.2 Probability Distribution

Let X be a random variable. We would like to specify the probabilities of events such as $\{X = x\}$ and $\{a \leq X \leq b\}$.

If we can specify all probabilities involving X, we say that we have specified the **probability distribution** of X.

One way to specify the probability distribution is to give the probabilities of all events of the form $\{X \leq x\}$, $x \in \mathbb{R}$. This leads to the following definition.

Definition 2.1 The **cumulative distribution function** (cdf) of a random variable X is the function $F : \mathbb{R} \to [0, 1]$ defined by

$$F(x) := \mathbb{P}(X \leq x), \ \ x \in \mathbb{R}.$$

Note that above we should have written $\mathbb{P}(\{X \leq x\})$ instead of $\mathbb{P}(X \leq x)$. From now on we will use this type of abbreviation throughout the course. In Figure 2.2 the graph of a cdf is depicted.

The following properties for F are a direct consequence of the three Axiom's for $\mathbb{P}$.

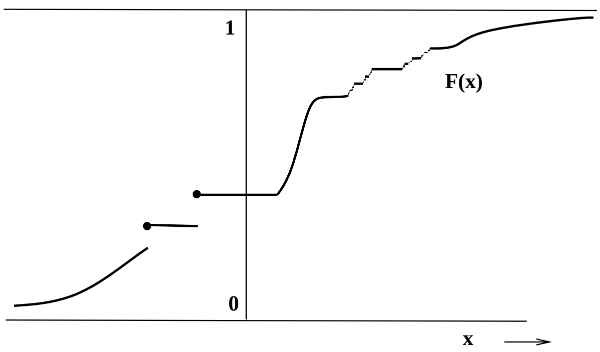


Figure 2.2: A cumulative distribution function

1. F is right-continuous: $\lim_{h\downarrow 0} F(x+h) = F(x)$,
2. $\lim_{x\to\infty} F(x) = 1$; $\lim_{x\to-\infty} F(x) = 0$.
3. F is increasing: $x \le y \Rightarrow F(x) \le F(y)$,
4. $0 \le F(x) \le 1$.

Proof. We will prove (1) and (2) in STAT3004. For (3), suppose $x \le y$ and define $A = \{X \le x\}$ and $B = \{X \le y\}$. Then, obviously, $A \subset B$ (for example if $\{X \le 3\}$ then this implies $\{X \le 4\}$). Thus, by (2) on page 14, $\mathbb{P}(A) \le \mathbb{P}(B)$, which proves (3). Property (4) follows directly from the fact that $0 \le \mathbb{P}(A) \le 1$ for any event A — and hence in particular for $A = \{X \le x\}$. ∎

Any function F with the above properties can be used to specify the distribution of a random variable X. Suppose that X has cdf F. Then the probability that X takes a value in the interval $(a, b]$ (excluding a, including b) is given by

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

Namely, $\mathbb{P}(X \le b) = \mathbb{P}(\{X \le a\} \cup \{a < X \le b\})$, where the events $\{X \le a\}$ and $\{a < X \le b\}$ are disjoint. Thus, by the sum rule: $F(b) = F(a) + \mathbb{P}(a < X \le b)$, which leads to the result above. Note however that

$$\begin{aligned}\mathbb{P}(a \le X \le b) &= F(b) - F(a) + \mathbb{P}(X = a) \\ &= F(b) - F(a) + F(a) - \lim_{h\downarrow 0} F(a-h) \\ &= F(b) - \lim_{h\downarrow 0} F(a-h).\end{aligned}$$

In practice we will specify the distribution of a random variable in a different way, whereby we make the distinction between *discrete* and *continuous* random variables.

2.2.1 Discrete Distributions

Definition 2.2 We say that X has a **discrete** distribution if X is a discrete random variable. In particular, for some finite or countable set of values $x_1, x_2, \ldots$ we have $\mathbb{P}(X = x_i) > 0$, $i = 1, 2, \ldots$ and $\sum_i \mathbb{P}(X = x_i) = 1$. We define the **probability mass function** (pmf) f of X by $f(x) = \mathbb{P}(X = x)$. We sometimes write f_X instead of f to stress that the pmf refers to the random variable X.

The easiest way to specify the distribution of a discrete random variable is to specify its pmf. Indeed, by the sum rule, if we know $f(x)$ for all x, then we can calculate all possible probabilities involving X. Namely,

$$\boxed{\mathbb{P}(X \in B) = \sum_{x \in B} f(x)} \tag{2.2}$$

for any subset B of the range of X.

Example 2.3 Toss a die and let X be its face value. X is discrete with range $\{1, 2, 3, 4, 5, 6\}$. If the die is fair the probability mass function is given by

x	1	2	3	4	5	6	$\sum$
$f(x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

Example 2.4 Toss two dice and let X be the largest face value showing. The pmf of X can be found to satisfy

x	1	2	3	4	5	6	$\sum$
$f(x)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$	1

The probability that the maximum is at least 3 is $\mathbb{P}(X \geq 3) = \sum_{x=3}^{6} f(x) = 32/36 = 8/9$.

2.2.2 Continuous Distributions

Definition 2.3 A random variable X is said to have a **continuous distribution** if X is a continuous random variable for which there exists a *positive* function f with *total integral 1*, such that for all a, b

$$\boxed{\mathbb{P}(a < X \leq b) = F(b) - F(a) = \int_a^b f(u)\, du.} \tag{2.3}$$

The function f is called the **probability density function** (pdf) of X.

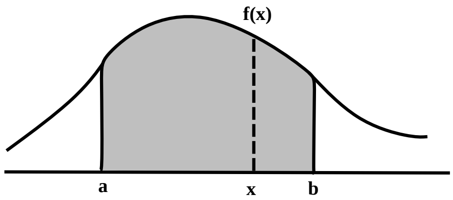


Figure 2.3: Probability density function (pdf)

Note that the corresponding cdf F is simply a *primitive* (also called anti-derivative) of the pdf f. In particular,

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(u)\, du.$$

Moreover, if a pdf f exists, then f is the *derivative* of the cdf F:

$$f(x) = \frac{d}{dx} F(x) = F'(x)\ .$$

We can *interpret* $f(x)$ as the "density" that $X = x$. More precisely,

$$\mathbb{P}(x \leq X \leq x + h) = \int_{x}^{x+h} f(u)\, du \approx h\, f(x)\ .$$

However, it is important to realise that $f(x)$ is *not a probability* — is a probability *density*. In particular, if X is a continuous random variable, then $\mathbb{P}(X = x) = 0$, *for all* x. Note that this also justifies using $\mathbb{P}(x \leq X \leq x + h)$ above instead of $\mathbb{P}(x < X \leq x + h)$. Although we will use the same notation f for probability mass function (in the discrete case) and probability density function (in the continuous case), it is crucial to understand the difference between the two cases.

Example 2.5 Draw a random number from the interval of real numbers $[0, 2]$. Each number is equally possible. Let X represent the number. What is the probability density function f and the cdf F of X?

Solution: Take an $x \in [0, 2]$. Drawing a number X "uniformly" in [0,2] means that $\mathbb{P}(X \leq x) = x/2$, for all such x. In particular, the cdf of X satisfies:

$$F(x) = \begin{cases} 0 & x < 0, \\ x/2 & 0 \leq x \leq 2, \\ 1 & x > 2. \end{cases}$$

By differentiating F we find

$$f(x) = \begin{cases} 1/2 & 0 \leq x \leq 2, \\ 0 & \text{otherwise} \end{cases}$$

Note that this density is *constant* on the interval $[0, 2]$ (and zero elsewhere), reflecting that each point in [0,2] is equally likely. Note also that we have modelled this random experiment using a continuous random variable and its pdf (and cdf). Compare this with the more "direct" model of Example 1.9.

Describing an experiment via a random variable and its pdf, pmf or cdf seems much easier than describing the experiment by giving the probability space. In fact, we have not used a probability space in the above examples.

2.3 Expectation

Although all the probability information of a random variable is contained in its cdf (or pmf for discrete random variables and pdf for continuous random variables), it is often useful to consider various numerical characteristics of that random variable. One such number is the *expectation* of a random variable; it is a sort of "weighted average" of the values that X can take. Here is a more precise definition.

Definition 2.4 Let X be a *discrete* random variable with pmf f. The **expectation** (or expected value) of X, denoted by $\mathbb{E}X$, is defined by

$$\mathbb{E}X = \sum_x x\,\mathbb{P}(X = x) = \sum_x x\,f(x)\,.$$

The expectation of X is sometimes written as μ_X.

Example 2.6 Find $\mathbb{E}X$ if X is the outcome of a toss of a fair die.

Since $\mathbb{P}(X = 1) = \ldots = \mathbb{P}(X = 6) = 1/6$, we have

$$\mathbb{E}X = 1\,(\frac{1}{6}) + 2\,(\frac{1}{6}) + \ldots + 6\,(\frac{1}{6}) = \frac{7}{2}.$$

Note: $\mathbb{E}X$ is not necessarily a possible outcome of the random experiment as in the previous example.

One way to interpret the expectation is as a type of "expected profit". Specifically, suppose we play a game where you throw two dice, and I pay you out, in dollars, the sum of the dice, X say. However, to enter the game you must pay me d dollars. You can play the game as many times as you like. What would be a "fair" amount for d? The answer is

$$\begin{aligned} d &= \mathbb{E}X = 2\,\mathbb{P}(X = 2) + 3\,\mathbb{P}(X = 3) + \cdots + 12\,\mathbb{P}(X = 12) \\ &= 2\frac{1}{36} + 3\,\frac{2}{35} + \cdots + 12\,\frac{1}{36} = 7\,. \end{aligned}$$

Namely, in the long run the fractions of times the sum is equal to 2, 3, 4, ... are $\frac{1}{36}, \frac{2}{35}, \frac{3}{36}, \ldots$, so the average pay-out per game is the weighted sum of 2,3,4,... with the weights being the probabilities/fractions. Thus the game is "fair" if the average profit (pay-out - d) is zero.

Another interpretation of expectation is as a *centre of mass*. Imagine that point masses with weights $p_1, p_2, \ldots, p_n$ are placed at positions $x_1, x_2, \ldots, x_n$ on the real line, see Figure 2.4.

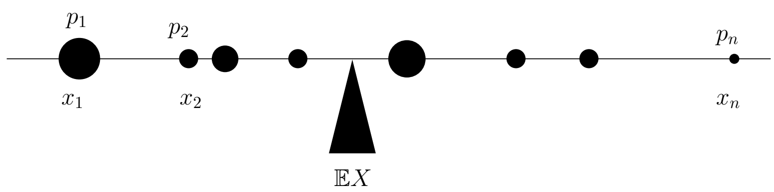


Figure 2.4: The expectation as a centre of mass

Then there centre of mass, the place where we can "balance" the weights, is

$$\text{centre of mass} = x_1\,p_1 + \cdots + x_n\,p_n,$$

which is exactly the expectation of the discrete variable X taking values $x_1, \ldots, x_n$ with probabilities $p_1, \ldots, p_n$. An obvious consequence of this interpretation is that for a *symmetric* probability mass function the expectation is equal to the symmetry point (provided the expectation exists). In particular, suppose $f(c+y) = f(c-y)$ for all y, then

$$\begin{aligned}
\mathbb{E}X &= c\,f(c) + \sum_{x>c} x f(x) + \sum_{x<c} x f(x) \\
&= c\,f(c) + \sum_{y>0} (c+y) f(c+y) + \sum_{y>0} (c-y) f(c-y) \\
&= c\,f(c) + \sum_{y>0} c\,f(c+y) + c\sum_{y>0} f(c-y) \\
&= c\sum_{x} f(x) = c
\end{aligned}$$

For continuous random variables we can define the expectation in a similar way:

Definition 2.5 Let X be a *continuous* random variable with pdf f. The **expectation** (or expected value) of X, denoted by $\mathbb{E}X$, is defined by

$$\mathbb{E}X = \int_x x\,f(x)\,dx\ .$$

If X is a random variable, then a function of X, such as X^2 or $\sin(X)$ is also a random variable. The following theorem is not so difficult to prove, and is entirely "obvious": the expected value of a function of X is the weighted average of the values that this function can take.

Theorem 2.1 If X is *discrete* with pmf f, then for any real-valued function g

$$\mathbb{E}\, g(X) = \sum_x g(x)\, f(x)\ .$$

Similarly, if X is *continuous* with pdf f, then

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

PROOF. We prove it for the discrete case only. Let $Y = g(X)$, where X is a discrete random variable with pmf f_X, and g is a function. Y is again a random variable. The pmf of Y, f_Y satisfies

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

Thus, the expectation of Y is

$$\mathbb{E}Y = \sum_y y f_Y(y) = \sum_y y \sum_{x:g(x)=y} f_X(x) = \sum_y \sum_{x:g(x)=y} y f_X(x) = \sum_x g(x) f_X(x)$$

■

Example 2.7 Find $\mathbb{E}X^2$ if X is the outcome of the toss of a fair die. We have

$$\mathbb{E}X^2 = 1^2\,\frac{1}{6} + 2^2\,\frac{1}{6} + 3^2\,\frac{1}{6} + \ldots + 6^2\,\frac{1}{6} = \frac{91}{6}.$$

An important consequence of Theorem 2.1 is that the expectation is "linear". More precisely, for any real numbers a and b, and functions g and h

1. $\mathbb{E}(a\,X + b) = a\,\mathbb{E}X + b\ .$
2. $\mathbb{E}(g(X) + h(X)) = \mathbb{E}g(X) + \mathbb{E}h(X)\ .$

PROOF. Suppose X has pmf f. Then 1. follows (in the discrete case) from

$$\mathbb{E}(aX + b) = \sum_x (ax + b) f(x) = a \sum_x x\, f(x) + b\, \sum_x f(x) = a\,\mathbb{E}X + b\ .$$

Similarly, 2. follows from

$$\mathbb{E}(g(X) + h(X)) = \sum_x (g(x) + h(x)) f(x) = \sum_x g(x) f(x) + \sum_x h(x) f(x)$$
$$= \mathbb{E}g(X) + \mathbb{E}h(X)\ .$$

The continuous case is proved analogously, by replacing the sum with an integral. ■

Another useful number about (the distribution of) X is the *variance* of X. This number, sometimes written as σ_X^2, measures the *spread* or dispersion of the distribution of X.

Definition 2.6 The **variance** of a random variable X, denoted by $\text{Var}(X)$ is defined by

$$\text{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2 \; .$$

The square root of the variance is called the **standard deviation**. The number $\mathbb{E}X^r$ is called the rth **moment** of X.

The following important properties for variance hold for discrete or continuous random variables and follow easily from the definitions of expectation and variance.

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

Proof. Write $\mathbb{E}X = \mu$, so that $\text{Var}(X) = \mathbb{E}(X-\mu)^2 = \mathbb{E}(X^2 - 2\mu X + \mu^2)$. By the linearity of the expectation, the last expectation is equal to the sum $\mathbb{E}(X^2) - 2\mu\mathbb{E}X + \mu^2 = \mathbb{E}X^2 - \mu^2$, which proves 1. To prove 2, note first that the expectation of $aX + b$ is equal to $a\mu + b$. Thus,

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

■

2.4 Transforms

Many calculations and manipulations involving probability distributions are facilitated by the use of *transforms*. We discuss here a number of such transforms.

Definition 2.7 Let X be a *non-negative* and *integer-valued* random variable. The **probability generating function** (PGF) of X is the function $G : [0,1] \to [0,1]$ defined by

$$G(z) := \mathbb{E}\, z^X = \sum_{x=0}^{\infty} z^x \, \mathbb{P}(X = x) \; .$$

Example 2.8 Let X have pmf f given by

$$f(x) = \mathrm{e}^{-\lambda}\,\frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots$$

We will shortly introduce this as the *Poisson* distribution, but for now this is not important. The PGF of X is given by

$$\begin{aligned} G(z) &= \sum_{x=0}^{\infty} z^x \, \mathrm{e}^{-\lambda} \, \frac{\lambda^x}{x!} \\ &= \mathrm{e}^{-\lambda} \sum_{x=0}^{\infty} \frac{(z\lambda)^x}{x!} \\ &= \mathrm{e}^{-\lambda} \mathrm{e}^{z\lambda} = \mathrm{e}^{-\lambda(1-z)} \, . \end{aligned}$$

Knowing only the PGF of X, we can easily obtain the pmf:

$$\mathbb{P}(X = x) = \frac{1}{x!} \frac{d^k}{dz^x} G(z) \bigg|_{z=0} .$$

Proof. By definition

$$G(z) = z^0 \mathbb{P}(X = 0) + z^1 \, \mathbb{P}(X = 1) + z^2 \, \mathbb{P}(X = 2) + \cdots$$

Substituting $z = 0$ gives, $G(0) = \mathbb{P}(X = 0)$; if we differentiate $G(z)$ once, then

$$G'(z) = \mathbb{P}(X = 1) + 2z \, \mathbb{P}(X = 2) + 3z^2 \, \mathbb{P}(X = 3) + \cdots$$

Thus, $G'(0) = \mathbb{P}(X = 1)$. Differentiating again, we see that $G''(0) = 2 \, \mathbb{P}(X = 2)$, and in general the n-th derivative of G at zero is $G^{(n)}(0) = x! \, \mathbb{P}(X = x)$, which completes the proof. ■

Thus we have the **uniqueness** property: two pmf's are the same if and only if their PGFs are the same.

Another useful property of the PGF is that we can obtain the moments of X by differentiating G and evaluating it at $z = 1$.

Differentiating $G(z)$ w.r.t. z gives

$$\begin{aligned} G'(z) &= \frac{d \, \mathbb{E} z^X}{dz} = \mathbb{E} X z^{X-1} \, . \\ G''(z) &= \frac{d \, \mathbb{E} X z^{X-1}}{dz} = \mathbb{E} X(X-1) z^{X-2} \, . \\ G'''(z) &= \mathbb{E} X(X-1)(X-2) z^{X-3} \, . \end{aligned}$$

Et cetera. If you're not convinced, write out the expectation as a sum, and use the fact that the derivative of the sum is equal to the sum of the derivatives (although we need a little care when dealing with infinite sums).

In particular,

$$\boxed{\mathbb{E} X = G'(1),}$$

and

$$\boxed{\mathrm{Var}(X) = G''(1) + G'(1) - (G'(1))^2 \, .}$$

Definition 2.8 The **moment generating function** (MGF) of a random variable X is the function, $M : I \to [0, \infty)$, given by

$$M(s) = \mathbb{E}\, \mathrm{e}^{sX} \; .$$

Here I is an open interval containing 0 for which the above integrals are well defined for all $s \in I$.

In particular, for a discrete random variable with pmf f,

$$M(s) = \sum_x \mathrm{e}^{sx} f(x),$$

and for a continuous random variable with pdf f,

$$M(s) = \int_x \mathrm{e}^{sx} f(x)\, dx \; .$$

We sometimes write M_X to stress the role of X.

As for the PGF, the moment generation function has the *uniqueness property*: Two MGFs are the same if and only if their corresponding distribution functions are the same.

Similar to the PGF, the moments of X follow from the derivatives of M:

If $\mathbb{E}X^n$ exists, then M is n times differentiable, and

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

Hence the name moment generating function: the moments of X are simply found by differentiating. As a consequence, the variance of X is found as

$$\mathrm{Var}(X) = M''(0) - (M'(0))^2.$$

Remark 2.1 The transforms discussed here are particularly useful when dealing with **sums** of **independent** random variables. We will return to them in Chapters 4 and 5.

2.5 Some Important Discrete Distributions

In this section we give a number of important discrete distributions and list some of their properties. Note that the pmf of each of these distributions depends on one or more parameters; so in fact we are dealing with *families* of distributions.

2.5.1 Bernoulli Distribution

We say that X has a **Bernoulli** distribution with success probability p if X can only assume the values 0 and 1, with probabilities

$$\mathbb{P}(X = 1) = p = 1 - \mathbb{P}(X = 0) \; .$$

We write $X \sim \mathsf{Ber}(p)$. Despite its simplicity, this is one of the most important distributions in probability! It models for example a single coin toss experiment. The cdf is given in Figure 2.5.

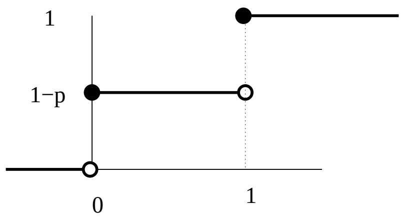


Figure 2.5: The cdf of the Bernoulli distribution

Here are some properties:

1. The expectation is $\mathbb{E}X = 0\mathbb{P}(X = 0) + 1\mathbb{P}(X = 1) = 0 \times (1-p) + 1 \times p = p$.
2. The variance is $\text{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \mathbb{E}X - (\mathbb{E}X)^2 = p - p^2 = p(1-p)$. (Note that $X^2 = X$).
3. The PGF is given by $G(z) = z^0(1-p) + z^1 p = 1 - p + zp$.

2.5.2 Binomial Distribution

Consider a sequence of n coin tosses. If X is the random variable which counts the total number of heads and the probability of "head" is p then we say X has a **binomial** distribution with parameters n and p and write $X \sim \mathsf{Bin}(n, p)$. The probability mass function X is given by

$$f(x) = \mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n. \tag{2.4}$$

This follows from Examples 1.15 and 2.2. An example of the graph of the pmf is given in Figure 2.6

Here are some important properties of the Bernoulli distribution. Some of these properties can be proved more easily after we have discussed multiple random variables.

1. The expectation is $\mathbb{E}X = np$. This is a quite intuitive result. The expected number of successes (heads) in n coin tosses is np, if p denotes the probability of success in any one toss. To prove this, one could simply evaluate the sum
$$\sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x},$$
but this is not elegant. We will see in chapter 4 that X can be viewed as a sum $X = X_1 + \cdots + X_n$ of n independent $\mathsf{Ber}(p)$ random variables,

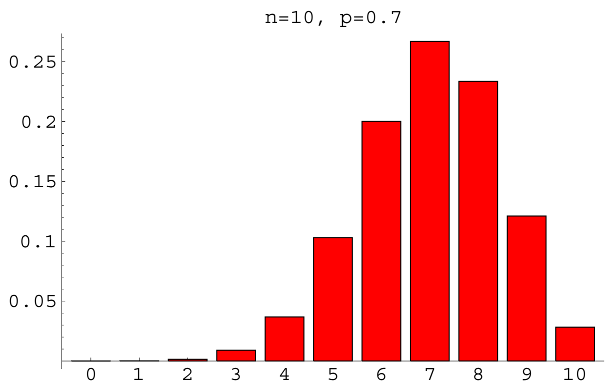


Figure 2.6: The pmf of the $\mathsf{Bin}(10, 0.7)$-distribution

where X_i indicates whether the i-th toss is a success or not, $i = 1, \ldots, n$. Also we will prove that the expectation of such a sum is the sum of the expectation, therefore,

$$\mathbb{E}X = \mathbb{E}(X_1 + \cdots + X_n) = \mathbb{E}X_1 + \cdots + \mathbb{E}X_n = \underbrace{p + \cdots p}_{n \text{ times}} = np \; .$$

2. The variance of X is $\mathrm{Var}(X) = np(1-p)$. This is proved in a similar way to the expectation:

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) \\ &= \underbrace{p(1-p) + \cdots p(1-p)}_{n \text{ times}} = np(1-p) \; .\end{aligned}$$

3. The probability generating function of X is $G(z) = (1 - p + zp)^n$. Again, we can easily prove this after we consider multiple random variables in Chapter 4. Namely,

$$\begin{aligned}G(z) &= \mathbb{E}z^X = \mathbb{E}z^{X_1 + \cdots + X_n} = \mathbb{E}z^{X_1} \cdots \mathbb{E}z^{X_n} \\ &= (1 - p + zp) \times \cdots \times (1 - p + zp) = (1 - p + zp)^n \; .\end{aligned}$$

However, we can also easily prove it using Newton's binomial formula:

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k \, b^{n-k} \; .$$

Specifically,

$$G(z) = \sum_{k=0}^{n} z^k \binom{n}{k} p^k \, (1-p)^{n-k} = \sum_{k=0}^{n} \binom{n}{k} (z\,p)^k (1-p)^{n-k} = (1-p+zp)^n \; .$$

Note that once we have obtained the PGF, we can obtain the expectation and variance as $G'(1) = np$ and $G''(1) + G'(1) - (G'(1))^2 = (n-1)np^2 + np - n^2p^2 = np(1-p)$.

2.5.3 Geometric distribution

Again we look at a sequence of coin tosses but count a different thing. Let X be the number of tosses needed before the first head occurs. Then

$$\mathbb{P}(X = x) = (1-p)^{x-1}p, \quad x = 1, 2, 3, \ldots \tag{2.5}$$

since the only string that has the required form is

$$\underbrace{ttt\ldots t}_{x-1}\,h$$

and this has probability $(1-p)^{x-1}p$. See also Example 1.16 on page 28. Such a random variable X is said to have a **geometric** distribution with parameter p. We write $X \sim \mathsf{G}(p)$. An example of the graph of the pdf is given in Figure 2.7

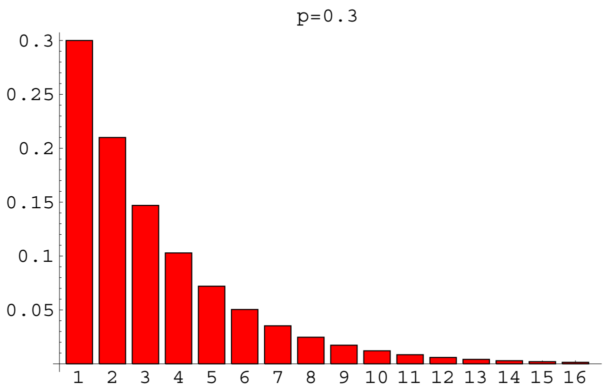


Figure 2.7: The pmf of the $\mathsf{G}(0.3)$-distribution

We give some more properties, including the expectation, variance and PGF of the geometric distribution. It is easiest to start with the PGF:

1. The PGF is given by

$$G(z) = \sum_{x=1}^{\infty} z^x p(1-p)^{x-1} = z\,p\sum_{k=0}^{\infty}(z(1-p))^k = \frac{z\,p}{1-z\,(1-p)},$$

   using the well-known result for *geometric sums*: $1 + a + a^2 + \cdots = \frac{1}{1-a}$, for $|a| < 1$.

2. The expectation is therefore

$$\mathbb{E}X = G'(1) = \frac{1}{p},$$

   which is an intuitive result. We expect to wait $1/p$ throws before a success appears, if successes are generated with probability p.

3. By differentiating the PGF twice we find the variance:

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

4. The probability of requiring more than k tosses before a success is

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

   This is obvious from the fact that $\{X > k\}$ corresponds to the event of k consecutive failures.

A final property of the geometric distribution which deserves extra attention is the **memoryless property**. Think again of the coin toss experiment. Suppose we have tossed the coin k times without a success (Heads). What is the probability that we need more than x additional tosses before getting a success. The answer is, obviously, the same as the probability that we require more than x tosses if we start from scratch, that is, $\mathbb{P}(X > x) = (1-p)^x$, irrespective of k. The fact that we have already had k failures does not make the event of getting a success in the next trial(s) any more likely. In other words, the coin does not have a memory of what happened, hence the word memoryless property. Mathematically, it means that for any $x, k = 1, 2, \ldots$,

$$\mathbb{P}(X > k + x \,|\, X > k) = \mathbb{P}(X > x)$$

Proof. By the definition of conditional probability

$$\mathbb{P}(X > k + x \,|\, X > k) = \frac{\mathbb{P}(\{X = k + x\} \cap \{X > k\})}{\mathbb{P}(X > k)}.$$

Now, the event $\{X > k + x\}$ is a subset of $\{X > k\}$, hence their intersection is $\{X > k + x\}$. Moreover, the probabilities of the events $\{X > k + x\}$ and $\{X > k\}$ are $(1-p)^{k+x}$ and $(1-p)^k$, respectively, so that

$$\mathbb{P}(X > k + x \,|\, X > k) = \frac{(1-p)^{k+x}}{(1-p)^k} = (1-p)^x = \mathbb{P}(X > x),$$

as required. ■

2.5.4 Poisson Distribution

A random variable X for which

$$\mathbb{P}(X = x) = \frac{\lambda^x}{x!}\, \mathrm{e}^{-\lambda}, \quad x = 0, 1, 2, \ldots, \tag{2.6}$$

(for fixed $\lambda > 0$) is said to have a **Poisson** distribution. We write $X \sim \mathsf{Poi}(\lambda)$. The Poisson distribution is used in many probability models and may be viewed as the "limit" of the $\mathsf{Bin}(n, \mu/n)$ for large n in the following sense: Consider a coin tossing experiment where we toss a coin n times with success probability λ/n. Let X be the number of successes. Then, as we have seen $X \sim \mathsf{Bin}(n, \lambda/n)$. In other words,

$$\begin{aligned}\mathbb{P}(X = k) &= \binom{n}{k}\left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!}\,\frac{n \times n-1 \times \cdots \times n-k+1}{n \times n \times \cdots \times n}\left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k}\end{aligned}$$

As $n \to \infty$, the second and fourth factors go to 1, and the third factor goes to $\mathrm{e}^{-\lambda}$ (this is one of the defining properties of the exponential function). Hence, we have

$$\lim_{n\to\infty} \mathbb{P}(X = k) = \frac{\lambda^k}{k!}\,\mathrm{e}^{-\lambda},$$

which shows that the Poisson distribution is a limiting case of the binomial one. An example of the graph of its pmf is given in Figure 2.8

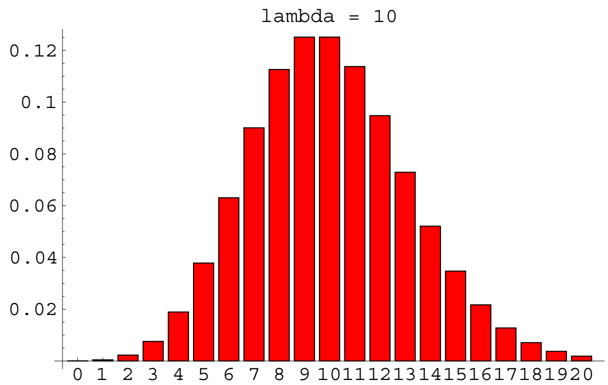


Figure 2.8: The pdf of the $\mathsf{Poi}(10)$-distribution

We finish with some properties.

1. The PGF was derived in Example 2.8:
$$G(z) = \mathrm{e}^{-\lambda(1-z)}\,.$$

2. It follows that the expectation is $\mathbb{E}X = G'(1) = \lambda$. The intuitive explanation is that the mean number of successes of the corresponding coin flip experiment is $np = n(\lambda/n) = \lambda$.

3. The above argument suggests that the variance should be $n(\lambda/n)(1 - \lambda/n) \to \lambda$. This is indeed the case, as
$$\mathrm{Var}(X) = G''(1) + G'(1) - (G'(1))^2 = \lambda^2 + \lambda - \lambda^2 = \lambda\,.$$
Thus for the Poisson distribution the variance and expectation are the same.

2.5.5 Hypergeometric Distribution

We say that a random variable X has a **Hypergeometric distribution** with parameters N, n and r if

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

for $\max\{0, r + n - N\} \leq k \leq \min\{n, r\}$.

We write $X \sim \mathsf{Hyp}(n, r, N)$. The hypergeometric distribution is used in the following situation.

Consider an urn with N balls, r of which are red. We draw at random n balls from the urn without replacement. The number of red balls amongst the n chosen balls has a $\mathsf{Hyp}(n, r, N)$ distribution. Namely, if we number the red balls $1, \ldots, r$ and the remaining balls $r+1, \ldots, N$, then the total number of outcomes of the random experiment is $\binom{N}{n}$, and each of these outcomes is equally likely. The number of outcomes in the event "k balls are red" is $\binom{r}{k} \times \binom{N-r}{n-k}$ because the k balls have to be drawn from the r red balls, and the remaining $n - k$ balls have to be drawn from the $N - k$ non-red balls. In table form we have:

	Red	Not Red	Total
Selected	k	$n - k$	n
Not Selected	$r - k$	$N - n - r + k$	$N - n$
Total	r	$N - r$	N

Example 2.9 Five cards are selected from a full deck of 52 cards. Let X be the number of Aces. Then $X \sim \mathsf{Hyp}(n = 5, r = 4, N = 52)$.

k	0	1	2	3	4	$\sum$
$\mathbb{P}(X = k)$	0.659	0.299	0.040	0.002	0.000	1

The expectation and variance of the hypergeometric distribution are

$$\mathbb{E}X = n\,\frac{r}{N}$$

and

$$\text{Var}(X) = n\,\frac{r}{N}\left(1 - \frac{r}{N}\right)\frac{N - n}{N - 1}\,.$$

Note that this closely resembles the expectation and variance of the binomial case, with $p = r/N$. The proofs will be given in Chapter 4 (see Examples 4.7 and 4.10, on pages 73 and 77).

2.6 Some Important Continuous Distributions

In this section we give a number of important continuous distributions and list some of their properties. Note that the pdf of each of these distributions depends on one or more parameters; so, as in the discrete case discussed before, we are dealing with *families* of distributions.

2.6.1 Uniform Distribution

We say that a random variable X has a **uniform** distribution on the interval $[a, b]$, if it has density function f, given by

$$f(x) = \frac{1}{b-a}, \quad a \le x \le b \,.$$

We write $X \sim \mathsf{U}[a, b]$. X can model a randomly chosen point from the interval $[a, b]$, where each choice is equally likely. A graph of the pdf is given in Figure 2.9.

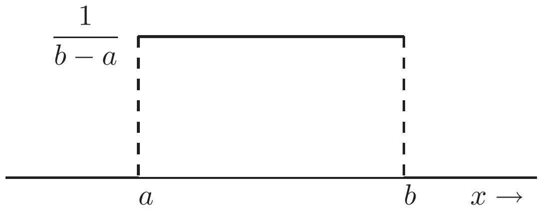


Figure 2.9: The pdf of the uniform distribution on $[a, b]$

We have

$$\mathbb{E}X = \int_a^b \frac{x}{b-a}\,dx = \frac{1}{b-a}\left[\frac{b^2-a^2}{2}\right] = \frac{a+b}{2} \,.$$

This can be seen more directly by observing that the pdf is symmetric around $c = (a+b)/2$, and that the expectation is therefore equal to the symmetry point c. For the variance we have

$$\begin{aligned} \mathrm{Var}(X) &= \mathbb{E}X^2 - (\mathbb{E}X)^2 = \int_a^b \frac{x^2}{b-a}\,dx - \left(\frac{a+b}{2}\right)^2 \\ &= \ldots = \frac{(a-b)^2}{12} \,. \end{aligned}$$

A more elegant way to derive this is to use the fact that X can be thought of as the sum $X = a + (b-a)U$, where $U \sim \mathsf{U}[0, 1]$. Namely, for $x \in [a, b]$

$$\mathbb{P}(X \le x) = \frac{x-a}{b-a} = \mathbb{P}\left(U \le \frac{x-a}{b-a}\right) = \mathbb{P}(a + (b-a)U \le x) \,.$$

Thus, we have $\mathrm{Var}(X) = \mathrm{Var}(a + (b-a)U) = (b-a)^2\mathrm{Var}(U)$. And

$$\mathrm{Var}(U) = \mathbb{E}U^2 - (\mathbb{E}U)^2 = \int_0^1 u^2 du - \left(\frac{1}{2}\right)^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12} \,.$$

2.6.2 Exponential Distribution

A random variable X with probability density function f, given by

$$f(x) = \lambda\, \mathrm{e}^{-\lambda\, x}, \quad x \geq 0 \tag{2.7}$$

is said to have an **exponential** distribution with parameter λ. We write $X \sim \mathsf{Exp}(\lambda)$. The exponential distribution can be viewed as a continuous version of the geometric distribution. Graphs of the pdf for various values of λ are given in Figure 2.10.

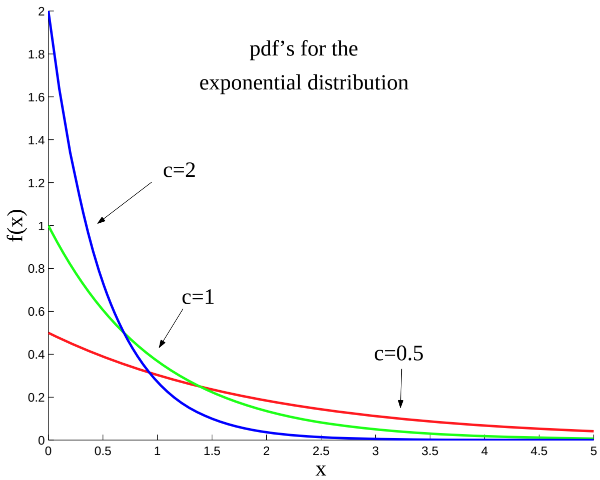


Figure 2.10: The pdf of the $\mathsf{Exp}(\lambda)$-distribution for various λ (c should be λ).

Here are some properties of the exponential function:

1. The moment generating function is

$$M(s) = \int_0^\infty \mathrm{e}^{sx} \lambda \mathrm{e}^{-\lambda x} dx = \lambda \int_0^\infty \mathrm{e}^{-(\lambda - s)x}\, dx = \lambda \left[\frac{-\mathrm{e}^{-(\lambda-s)x}}{\lambda - s} \right]_{x=0}^{\infty}$$
$$= \frac{\lambda}{\lambda - s},$$

for $s < \lambda$.

2. From the moment generating function we find by differentiation:

$$\mathbb{E}X = M'(0) = \left. \frac{\lambda}{(\lambda - s)^2} \right|_{s=0} = \frac{1}{\lambda}\,.$$

Alternatively, you can use partial integration to evaluate

$$\mathbb{E}X = \int_0^\infty \underbrace{x}_{\substack{\downarrow \\ 1}} \underbrace{\lambda \mathrm{e}^{-\lambda x}}_{\substack{\downarrow \\ -\mathrm{e}^{-\lambda x}}}\, dx = \left[-\mathrm{e}^{-\lambda x} \right]_0^\infty + \int_0^\infty \mathrm{e}^{-\lambda x}\, dx = 0 + \left[\frac{-\mathrm{e}^{-\lambda x}}{\lambda} \right]_0^\infty = \frac{1}{\lambda}\,.$$

3. Similarly, the second moment is $\mathbb{E}X^2 = M''(0) = \frac{2\lambda}{(\lambda-s)^3}\big|_{s=0} = 2/\lambda^2$, so that the variance becomes
$$\text{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}\,.$$
4. The cdf of X is given by
$$F(x) = \mathbb{P}(X \leq x) = \int_0^x \lambda \mathrm{e}^{-\lambda u} du = \Big[-\mathrm{e}^{-\lambda u}\Big]_0^x = 1 - \mathrm{e}^{-\lambda x}, \quad x \geq 0.$$
5. As consequence the tail probability $\mathbb{P}(X > x)$ is exponentially decaying:
$$\mathbb{P}(X > x) = \mathrm{e}^{-\lambda x}, \quad x \geq 0\,.$$

The most important property of the exponential distribution is the following

Theorem 2.2 (Memoryless Property) Let X have an exponential distribution with parameter λ. Then for any $s, t > 0$
$$\mathbb{P}(X > s + t \,|\, X > s) = \mathbb{P}(X > t), \quad \text{for all } s, t \geq 0. \tag{2.8}$$

Proof. By (1.4)
$$\begin{aligned}\mathbb{P}(X > s + t \,|\, X > s) &= \frac{\mathbb{P}(X > s + t,\, X > s)}{\mathbb{P}(X > s)} = \frac{\mathbb{P}(X > s + t)}{\mathbb{P}(X > s)} \\ &= \frac{\mathrm{e}^{-\lambda(t+s)}}{\mathrm{e}^{-\lambda s}} = \mathrm{e}^{-\lambda t} = \mathbb{P}(X > t),\end{aligned}$$
where in the second equation we have used the fact that the event $\{X > s + t\}$ is contained in the event $\{X > s\}$ hence the intersection of these two sets is $\{X > s + t\}$. ■

For example, when X denotes the lifetime of a machine, then given the fact that the machine is alive at time s, the remaining lifetime of the machine, i.e. $X - s$, has the same exponential distribution as a completely new machine. In other words, the machine has no memory of its age and does not "deteriorate" (although it will break down eventually).

It is not too difficult to prove that the exponential distribution is the *only* continuous (positive) distribution with the memoryless property.

2.6.3 Normal, or Gaussian, Distribution

The normal (or Gaussian) distribution is the most important distribution in the study of statistics. We say that a random variable has a **normal** distribution with parameters μ and σ^2 if its density function f is given by
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\,\mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},\; x \in \mathbb{R}. \tag{2.9}$$

We write $X \sim \mathsf{N}(\mu, \sigma^2)$. The parameters μ and σ^2 turn out to be the expectation and variance of the distribution, respectively. If $\mu = 0$ and $\sigma = 1$ then

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

and the distribution is known as a **standard normal** distribution. The cdf of this latter distribution is often denoted by Φ, and is tabulated in Appendix B. In Figure 2.11 the probability densities for three different normal distributions have been depicted.

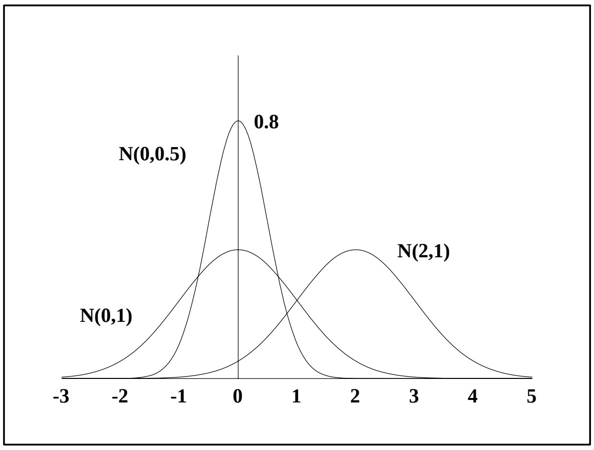


Figure 2.11: Probability density functions for various normal distributions

We next consider some important properties of the normal distribution.

1. If $X \sim \mathsf{N}(\mu, \sigma^2)$, then

$$\frac{X - \mu}{\sigma} \sim \mathsf{N}(0, 1)\ . \tag{2.10}$$

Thus by subtracting the mean and dividing by the standard deviation we obtain a standard normal distribution. This procedure is called **standardisation**.

PROOF. Let $X \sim \mathsf{N}(\mu, \sigma^2)$, and $Z = (X - \mu)/\sigma$. Then,

$$\begin{aligned}
\mathbb{P}(Z \le z) &= \mathbb{P}((X-\mu)/\sigma \le z) = \mathbb{P}(X \le \mu + \sigma z) \\
&= \int_{-\infty}^{\mu+\sigma z} \frac{1}{\sigma\sqrt{2\pi}}\, \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \\
&= \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-y^2/2} dy \qquad [\text{c.o.v. } y = \tfrac{x-\mu}{\sigma}] \\
&= \Phi(z)\ .
\end{aligned}$$

Thus Z has a standard normal distribution. ∎

Standardisation enables us to express the cdf of any normal distribution in terms of the cdf of the standard normal distribution. This is the reason why only the table for the standard normal distribution is included in the appendix.

2. A trivial rewriting of the standardisation formula gives the following important result: If $X \sim \mathsf{N}(\mu, \sigma^2)$, then

$$X = \mu + \sigma Z, \qquad \text{with } Z \sim \mathsf{N}(0,1) .$$

In other words, any Gaussian (normal) random variable can be viewed as a so-called *affine* (linear + constant) transformation of a standard normal random variable.

3. $\mathbb{E}X = \mu$. This is because the pdf is symmetric around μ.

4. $\mathrm{Var}(X) = \sigma^2$. This is a bit more involved. First, write $X = \mu + \sigma Z$, with Z standard normal. Then, $\mathrm{Var}(X) = \mathrm{Var}(\mu + \sigma Z) = \sigma^2 \mathrm{Var}(Z)$. Hence, it suffices to show that the variance of Z is 1. Consider $\mathrm{Var}(Z) = \mathbb{E}Z^2$ (note that the expectation is 0). We have

$$\mathbb{E}Z^2 = \int_{-\infty}^{\infty} z^2 \frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-z^2/2} dz = \int_{-\infty}^{\infty} z \times \frac{z}{\sqrt{2\pi}}\, \mathrm{e}^{-z^2/2} dz$$

By writing the last integrand in this way we can apply partial integration to the two factors to yield

$$\mathbb{E}Z^2 = \left[z\, \frac{-1}{\sqrt{2\pi}}\, \mathrm{e}^{-z^2/2} \right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-z^2/2}\, dz = 1 ,$$

since the last integrand is the pdf of the standard normal distribution.

5. The moment generating function of $X \sim \mathsf{N}(\mu, \sigma^2)$ is given by

$$\mathbb{E}\mathrm{e}^{sX} = \mathrm{e}^{s\mu + s^2\sigma^2/2}, \quad s \in \mathbb{R}. \tag{2.11}$$

PROOF. First consider the moment generation function of $Z \sim \mathsf{N}(0,1)$. We have

$$\begin{aligned}\mathbb{E}\mathrm{e}^{sZ} &= \int_{-\infty}^{\infty} \mathrm{e}^{sz} \frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-z^2/2} dz = \mathrm{e}^{s^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-(z-s)^2/2} dz \\ &= \mathrm{e}^{s^2/2},\end{aligned}$$

where the second integrand is the pdf of the $\mathsf{N}(s,1)$ distribution, which therefore integrates to 1. Now, for general $X \sim \mathsf{N}(\mu, \sigma^2)$ write $X = \mu + \sigma Z$. Then,

$$\mathbb{E}\mathrm{e}^{sX} = \mathbb{E}\mathrm{e}^{s(\mu + \sigma Z)} = \mathrm{e}^{s\mu} \mathbb{E}\mathrm{e}^{s\sigma Z} = \mathrm{e}^{s\mu} \mathrm{e}^{\sigma^2 s^2/2} = \mathrm{e}^{s\mu + \sigma^2 s^2/2} .$$

■

More on the Gaussian distribution later, especially the multidimensional cases!

2.6.4 Gamma- and χ^2-distribution

The **gamma** distribution arises frequently in statistics. Its density function is given by

$$f(x) = \frac{\lambda^\alpha x^{\alpha-1} \mathrm{e}^{-\lambda x}}{\Gamma(\alpha)}, \quad x \geq 0, \tag{2.12}$$

where Γ is the Gamma-function defined as

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1}\, \mathrm{e}^{-u}\, du, \; \alpha > 0.$$

Parameter α is called the **shape** parameter, and λ is called the **scale** parameter. We write $X \sim \mathsf{Gam}(\alpha, \lambda)$.

Of particular importance is following special case: A random variable X is said to have a **chi-square** distribution with n ($\in \{1, 2, \ldots\}$) **degrees of freedom** if $X \sim \mathsf{Gam}(n/2,\ 1/2)$. We write $X \sim \chi^2_n$. A graph of the pdf of the χ^2_n-distribution, for various n is given in Figure 2.12.

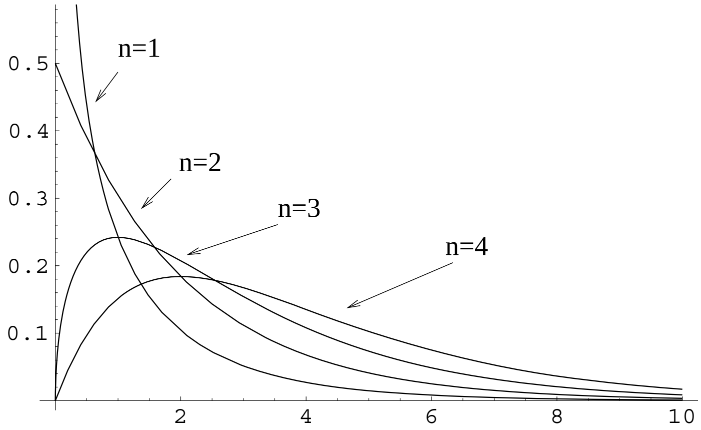


Figure 2.12: Pdfs for the χ^2_n-distribution, for various degrees of freedom n

We mention a few properties of the Γ-function.

1. $\Gamma(a + 1) = a\, \Gamma(a)$, for $a \in \mathbb{R}_+$.
2. $\Gamma(n) = (n - 1)!$ for $n = 1, 2, \ldots$.
3. $\Gamma(1/2) = \sqrt{\pi}$.

The moment generating function of $X \sim \mathsf{Gam}(\alpha, \lambda)$ is given by

$$\begin{aligned} M(s) = \mathbb{E}\, \mathrm{e}^{sX} &= \int_0^\infty \frac{\mathrm{e}^{-\lambda x}\, \lambda^\alpha\, x^{\alpha-1}}{\Gamma(\alpha)}\, \mathrm{e}^{sx}\, dx \\ &= \left(\frac{\lambda}{\lambda - s}\right)^\alpha \int_0^\infty \frac{\mathrm{e}^{-(\lambda-s)x}\, (\lambda - s)^\alpha\, x^{\alpha-1}}{\Gamma(\alpha)}\, dx \\ &= \left(\frac{\lambda}{\lambda - s}\right)^\alpha . \end{aligned} \tag{2.13}$$

As a consequence, we have

$$\mathbb{E}X = M'(0) = \frac{\alpha}{\lambda}\left(\frac{\lambda}{\lambda - s}\right)^{\alpha+1}\Bigg|_{s=0} = \frac{\alpha}{\lambda},$$

and, similarly,

$$\mathrm{Var}(X) = \frac{\alpha}{\lambda^2}\,.$$

Chapter 3

Generating Random Variables on a Computer

3.1 Introduction

This chapter deals with the execution of random experiments via the computer, also called **stochastic simulation**. In a typical stochastic simulation, randomness is introduced into simulation models via independent uniformly distributed random variables, also called **random numbers**. These random numbers are then used as building blocks to simulate more general stochastic systems.

3.2 Random Number Generation

In the early days of simulation, randomness was generated by *manual* techniques, such as coin flipping, dice rolling, card shuffling, and roulette spinning. Later on, *physical devices*, such as noise diodes and Geiger counters, were attached to computers for the same purpose. The prevailing belief held that only mechanical or electronic devices could produce "truly" random sequences. Although mechanical devices are still widely used in gambling and lotteries, these methods were abandoned by the computer-simulation community for several reasons: (a) Mechanical methods were too slow for general use, (b) the generated sequences cannot be reproduced and, (c) it has been found that the generated numbers exhibit both bias and dependence. Although certain modern physical generation methods are fast and would pass most statistical tests for randomness (for example, those based on the universal background radiation or the noise of a PC chip), their main drawback remains to be their lack of repeatability. Most of today's random number generators are not based on physical devices, but on simple algorithms that can be easily implemented on a computer, are fast, require little storage space, and can readily reproduce a given sequence of random numbers. Importantly, a good random number generator captures all the important statistical properties of true random sequences, even though the sequence is generated by a deterministic algorithm. For this reason, these generators are sometimes called *pseudorandom*.

The most common methods for generating pseudorandom sequences use the so-called *linear congruential generators*. These generate a deterministic sequence of numbers by means of the recursive formula

$$X_{i+1} = aX_i + c \pmod{m} , \tag{3.1}$$

where the initial value, X_0, is called the *seed*, and the a, c and m (all positive integers) are called the *multiplier*, the *increment* and the *modulus*, respectively. Note that applying the modulo-m operator in (3.1) means that $aX_i + c$ is divided by m, and the remainder is taken as the value of X_{i+1}. Thus, each X_i can only assume a value from the set $\{0, 1, \ldots, m-1\}$, and the quantities

$$U_i = \frac{X_i}{m} , \tag{3.2}$$

called *pseudorandom numbers*, constitute approximations to the true sequence of uniform random variables. Note that the sequence $\{X_i\}$ will repeat itself in at most m steps, and will therefore be periodic with period not exceeding m. For example, let $a = c = X_0 = 3$ and $m = 5$; then the sequence obtained from the recursive formula $X_{i+1} = 3X_i + 3 \pmod{5}$ is $X_i = 3, 2, 4, 0, 3$, which has period 4, while $m = 5$. In the special case where $c = 0$, (3.1) simply reduces to

$$X_{i+1} = a\,X_i \pmod{m} . \tag{3.3}$$

Such a generator is called a *multiplicative congruential generator*. It is readily seen that an arbitrary choice of X_0, a, c and m will not lead to a pseudorandom sequence with good statistical properties. In fact, number theory has been used to show that only a few combinations of these produce satisfactory results. In computer implementations, m is selected as a large prime number that can be accommodated by the computer word size. For example, in a binary 32-bit word computer, statistically acceptable generators can be obtained by choosing $m = 2^{31} - 1$ and $a = 7^5$, provided the first bit is a sign bit. A 64-bit or 128-bit word computer computer will naturally yield better statistical results.

Most computer languages already contain a built-in pseudorandom number generator. The user is typically requested only to input the initial seed, X_0, and upon invocation, the random number generator produces a sequence of independent uniform $(0, 1)$ random variables. We, therefore assume in this chapter the availability of a "black box", capable of producing a stream of pseudorandom numbers. In Matlab, for example, this is provided by the `rand` function.

Example 3.1 (Generating uniform random variables in Matlab) This example illustrates the use of the `rand` function in Matlab, to generate samples from the $\mathsf{U}(0, 1)$ distribution. For clarity we have omitted the "`ans = `" output in the Matlab session below.

```
>> rand                  % generate a uniform random number
    0.0196
>> rand                  % generate another uniform random number
    0.823
>> rand(1,4)             % generate a uniform random vector
    0.5252  0.2026  0.6721  0.8381
rand('state',1234)       % set the seed to 1234
>> rand                  % generate a uniform random number
    0.6104
rand('state',1234)       % reset the seed to 1234
>> rand
    0.6104               % the previous outcome is repeated
```

3.3 The Inverse-Transform Method

In this section we discuss a general method for generating one-dimensional random variables from a prescribed distribution, namely the *inverse-transform method*.

Let X be a random variable with cdf F. Since F is a nondecreasing function, the inverse function F^{-1} may be defined as

$$F^{-1}(y) = \inf\{x : F(x) \geq y\} \ , \quad 0 \leq y \leq 1 \ . \tag{3.4}$$

(Readers not acquainted with the notion inf should read min.) It is easy to show that if $U \sim \mathsf{U}(0,1)$, then

$$X = F^{-1}(U) \tag{3.5}$$

has cdf F. Namely, since F is invertible and $\mathbb{P}(U \leq u) = u$, we have

$$\mathbb{P}(X \leq x) = \mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x) \ . \tag{3.6}$$

Thus, to generate a random variable X with cdf F, draw $U \sim \mathsf{U}(0,1)$ and set $X = F^{-1}(U)$. Figure 3.1 illustrates the inverse-transform method given by the following algorithm.

Algorithm 3.1 (The Inverse-Transform Method)

1. *Generate U from* $\mathsf{U}(0,1)$.
2. *Return $X = F^{-1}(U)$.*

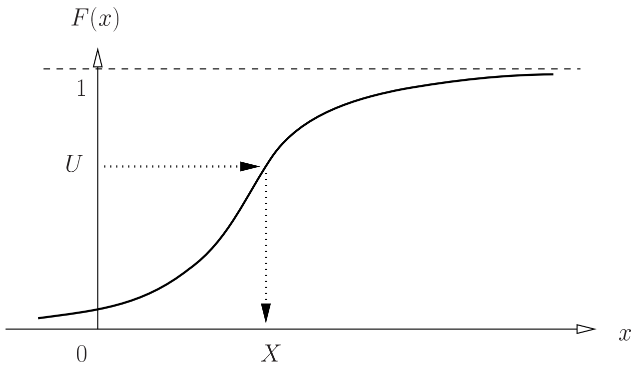


Figure 3.1: The inverse-transform method.

Example 3.2 Generate a random variable from the pdf

$$f(x) = \begin{cases} 2x, & 0 \leq x \leq 1 \\ 0, & \text{otherwise.} \end{cases} \tag{3.7}$$

The cdf is

$$F(x) = \begin{cases} 0, & x < 0 \\ \int_0^x 2x\,dx = x^2, & 0 \leq x \leq 1 \\ 1, & x > 1. \end{cases}$$

Applying (3.5), we have

$$X = F^{-1}(U) = \sqrt{U}, \quad 0 \leq u \leq 1 \; .$$

Therefore, to generate a random variable X from the pdf (3.7), first generate a random variable U from $\mathsf{U}(0,1)$, and then take its square root.

Example 3.3 (Drawing From a Discrete Distribution) Let X be a discrete random variable with $\mathbb{P}(X = x_i) = p_i, \; i = 1, 2, \ldots$, with $\sum_i p_i = 1$. The cdf F of X is given by $F(x) = \sum_{i:x_i \leq x} p_i, \; i = 1, 2, \ldots$; see Figure 3.2.

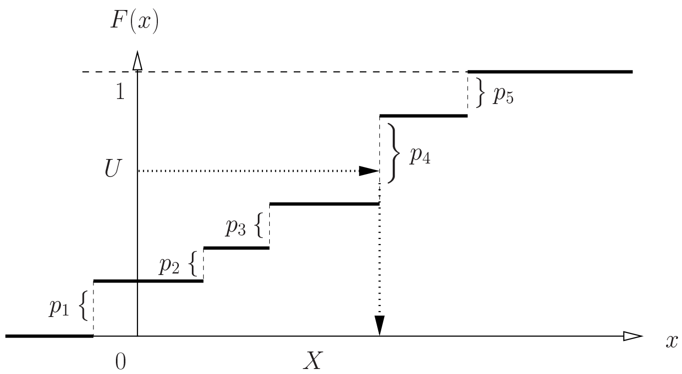


Figure 3.2: The inverse-transform method for a discrete random variable.

The algorithm for generating a random variable from F can thus be written as follows:

Algorithm 3.2 (The Inverse-Transform Method for a Discrete Distribution)

1. *Generate* $U \sim \mathsf{U}(0,1)$.
2. *Find the smallest positive integer,* k*, such that* $U \leq F(x_k)$ *and return* $X = x_k$.

In matlab drawing from a probability vector $(p_1, \ldots, p_n)$ can be done in one line:

```
min(find(cumsum(p)> rand));
```

Here `p` is the vector of probabilities, such as (1/3,1/3,1/3), `cumsum` gives the cumulative vector, e.g., (1/3, 2/3, 1), `find` finds the indices i such that $p_i > r$, where r is some random number, and `min` takes the smallest of these indices.

Much of the execution time in Algorithm 3.2 is spent in making the comparisons of Step 2. This time can be reduced by using efficient search techniques.

In general, the inverse-transform method requires that the underlying cdf, F, exist in a form for which the corresponding inverse function F^{-1} can be found analytically or algorithmically. Applicable distributions are, for example, the exponential, uniform, Weibull, logistic, and Cauchy distributions. Unfortunately, for many other probability distributions, it is either impossible or difficult to find the inverse transform, that is, to solve

$$F(x) = \int_{-\infty}^{x} f(t)\, dt = u$$

with respect to x. Even in the case where F^{-1} exists in an explicit form, the inverse-transform method may not necessarily be the most efficient random variable generation method.

3.4 Generating From Commonly Used Distributions

The next two subsections present algorithms for generating variables from commonly used continuous and discrete distributions. Of the numerous algorithms available we have tried to select those which are reasonably efficient and relatively simple to implement.

Exponential Distribution

We start by applying the inverse-transform method to the exponential distribution. If $X \sim \mathsf{Exp}(\lambda)$, then its cdf F is given by

$$F(x) = 1 - \mathrm{e}^{-\lambda x}, \qquad x > 0 \ . \tag{3.8}$$

Hence, solving $u = F(x)$ in terms of x gives

$$F^{-1}(u) = -\frac{1}{\lambda} \log(1-u) \ .$$

Noting that $U \sim \mathsf{U}(0,1)$ implies $1 - U \sim \mathsf{U}(0,1)$, we obtain the following algorithm.

Algorithm 3.3 (Generation of an Exponential Random Variable)

1. *Generate* $U \sim \mathsf{U}(0,1)$.
2. *Return* $X = -\frac{1}{\lambda} \ln U$ *as a random variable from* $\mathsf{Exp}(\lambda)$.

There are many alternative procedures for generating variables from the exponential distribution. The interested reader is referred to Luc Devroye's book *Non-Uniform Random Variate Generation*, Springer-Verlag, 1986. (The entire book can be downloaded for free.)

Normal (Gaussian) Distribution

If $X \sim \mathsf{N}(\mu, \sigma^2)$, its pdf is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \ \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \ , \qquad -\infty < x < \infty \ , \tag{3.9}$$

where μ is the mean (or expectation) and σ^2 the variance of the distribution.

Since inversion of the normal cdf is numerically inefficient, the inverse-transform method is not very suitable for generating normal random variables, and some other procedures must be devised instead. We consider only generation from $\mathsf{N}(0,1)$ (standard normal variables), since any random $Z \sim \mathsf{N}(\mu, \sigma^2)$ can be represented as $Z = \mu + \sigma X$, where X is from $\mathsf{N}(0,1)$. One of the earliest methods for generating variables from $\mathsf{N}(0,1)$ is the following, due to Box and Müller. A justification of this method will be given in Chapter 5, see Example 5.6.

Algorithm 3.4 (Generation of a Normal Random Variable, Box and Müller Approach)

1. *Generate two independent random variables,* U_1 *and* U_2*, from* $\mathsf{U}(0,1)$.

2. *Return two independent standard normal variables, X and Y, via*

$$X = (-2 \ln U_1)^{1/2} \cos(2\pi U_2) \; , \\ Y = (-2 \ln U_1)^{1/2} \sin(2\pi U_2) \; . \tag{3.10}$$

An alternative generation method for $\mathsf{N}(0,1)$ is based on the acceptance–rejection method. First, note that in order to generate a random variable Y from $\mathsf{N}(0,1)$, one can first generate a nonnegative random variable X from the pdf

$$f(x) = \sqrt{\frac{2}{\pi}}\, \mathrm{e}^{-x^2/2}, \qquad x \geq 0, \tag{3.11}$$

and then assign to X a random sign. The validity of this procedure follows from the symmetry of the standard normal distribution about zero.

To generate a random variable X from (3.11), we bound $f(x)$ by $C\, g(x)$, where $g(x) = \mathrm{e}^{-x}$ is the pdf of the $\mathsf{Exp}(1)$. The smallest constant C such that $f(x) \leq Cg(x)$ is $\sqrt{2\mathrm{e}/\pi}$, see Figure 3.3. The efficiency of this method is thus $\sqrt{\pi/2\mathrm{e}} \approx 0.76$.

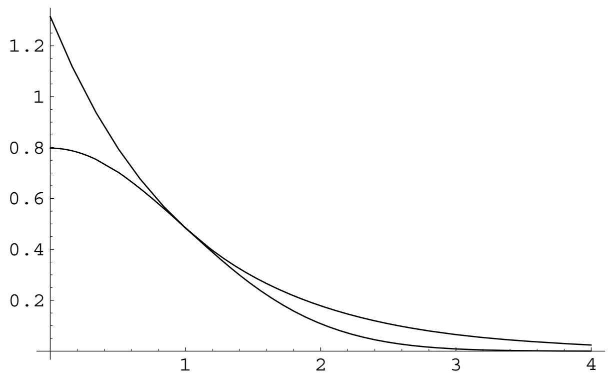


Figure 3.3: Bounding the positive normal density

Bernoulli distribution

If $X \sim \mathsf{Ber}(p)$, its pmf is of the form

$$f(x) = p^x (1-p)^{1-x}, \qquad x = 0, 1, \tag{3.12}$$

where p is the success probability. Applying the inverse-transform method, one readily obtains the following generation algorithm:

Algorithm 3.5 (Generation of a Bernoulli Random Variable)

1. *Generate $U \sim \mathsf{U}(0,1)$.*
2. *If $U \leq p$, return $X = 1$; otherwise return $X = 0$.*

In Figure 1.1 on page 6 typical outcomes are given of 100 independent Bernoulli random variables, each with success parameter $p = 0.5$.

Binomial distribution

If $X \sim \mathsf{Bin}(n,p)$ then its pmf is of the form

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n\ . \tag{3.13}$$

Recall that a binomial random variable X can be viewed as the total number of successes in n independent Bernoulli experiments, each with success probability p; see Example 1.15. Denoting the result of the i-th trial by $X_i = 1$ (success) or $X_i = 0$ (failure), we can write $X = X_1 + \cdots + X_n$ with the $\{X_i\}$ iid $\mathsf{Ber}(p)$ random variables. The simplest generation algorithm can thus be written as follows:

Algorithm 3.6 (Generation of a Binomial Random Variable)

1. *Generate iid random variables* $X_1, \ldots, X_n$ *from* $\mathsf{Ber}(p)$.
2. *Return* $X = \sum_{i=1}^n X_i$ *as a random variable from* $\mathsf{Bin}(n,p)$.

It is worthwhile to note that if $Y \sim \mathsf{Bin}(n,p)$, then $n-Y \sim \mathsf{Bin}(n, 1-p)$. Hence, to enhance efficiency, one may elect to generate X from $\mathsf{Bin}(n,p)$ according to

$$X = \begin{cases} Y_1 \sim \mathsf{Bin}(n,p), & \text{if } p \le \frac{1}{2} \\ Y_2 \sim \mathsf{Bin}(n, 1-p), & \text{if } p > \frac{1}{2}\ . \end{cases}$$

Geometric Distribution

If $X \sim \mathsf{G}(p)$, then its pmf is of the form

$$f(x) = p\,(1-p)^{x-1}, \qquad x = 1, 2 \ldots\ . \tag{3.14}$$

The random variable X can be interpreted as the the number of trials required until the first success occurs, in a series of independent Bernoulli trials with success parameter p. Note that $\mathbb{P}(X > m) = (1-p)^m$.

We now present an algorithm which is based on the relationship between the exponential and geometric distributions. Let $Y \sim \mathsf{Exp}(\lambda)$, with λ such that $q = 1-p = \mathrm{e}^{-\lambda}$. Then, $X = \lfloor Y \rfloor + 1$ has a $\mathsf{G}(p)$ distribution — here $\lfloor\ \rfloor$ denotes the integer part. Namely,

$$\mathbb{P}(X > x) = \mathbb{P}(\lfloor Y \rfloor > x - 1) = \mathbb{P}(Y > x) = \mathrm{e}^{-\lambda x} = (1-p)^x\ .$$

Hence, to generate a random variable from $\mathsf{G}(p)$, we first generate a random variable from the exponential distribution with $\lambda = -\ln(1-p)$, truncate the obtained value to the nearest integer and add 1.

Algorithm 3.7 (Generation of a Geometric Random Variable)

1. *Generate* $Y \sim \mathsf{Exp}(-\ln(1-p))$
2. *Return* $X = 1 + \lfloor Y \rfloor$ *as a random variable from* $\mathsf{G}(p)$.

Chapter 4

Joint Distributions

Often a random experiment is described via more than one random variable. Examples are:

1. We select a random sample of $n = 10$ people and observe their lengths. Let $X_1, \ldots, X_n$ be the individual lengths.

2. We flip a coin repeatedly. Let $X_i = 1$ if the ith flip is "heads" and 0 else. The experiment is described by the sequence $X_1, X_2, \ldots$ of coin flips.

3. We randomly select a person from a large population and measure his/her weight X and height Y.

How can we specify the behaviour of the random variables above? We should not just specify the pdf or pmf of the individual random variables, but also say something about the "interaction" (or lack thereof) between the random variables. For example, in the third experiment above if the height Y is large, we expect that X is large as well. On the other hand, for the first and second experiment it is reasonable to assume that information about one of the random variables does not give extra information about the others. What we need to specify is the **joint distribution** of the random variables.

The theory for multiple random variables is quite similar to that of a single random variable. The most important extra feature is perhaps the concept of *independence* of random variables. Independent random variables play a crucial role in stochastic modelling.

4.1 Joint Distribution and Independence

Let $X_1, \ldots, X_n$ be random variables describing some random experiment. We can accumulate the X_i's into a row vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ or column vector

$\boldsymbol{X} = (X_1, \ldots, X_n)^T$ (here T means transposition). $\boldsymbol{X}$ is called a **random vector**.

Recall that the distribution of a *single* random variable X is completely specified by its cumulative distribution function. Analogously, the joint distribution of $X_1, \ldots, X_n$ is specified by the **joint cumulative distribution function** F, defined by

$$F(x_1, \ldots, x_n) = \mathbb{P}(\{X_1 \le x_1\} \cap \cdots \cap \{X_n \le x_n\}) = \mathbb{P}(X_1 \le x_1, \ldots, X_n \le x_n),$$

If we know F then we can in principle derive any probability involving the X_i's. Note the abbreviation on the right-hand side. We will henceforth use this kind of abbreviation throughout the notes.

Similar to the 1-dimensional case we distinguish between the case where the X_i are discrete and continuous. The corresponding joint distributions are again called discrete and continuous, respectively.

4.1.1 Discrete Joint Distributions

To see how things work in the discrete case, let's start with an example.

Example 4.1 In a box are three dice. Die 1 is a normal die; die 2 has no 6 face, but instead two 5 faces; die 3 has no 5 face, but instead two 6 faces. The experiment consists of selecting a die at random, followed by a toss with that die. Let X be the die number that is selected, and let Y be the face value of that die. The probabilities $\mathbb{P}(X = x, Y = y)$ are specified below.

	y						
x	1	2	3	4	5	6	$\sum$
1	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{3}$
2	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{9}$	0	$\frac{1}{3}$
3	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	0	$\frac{1}{9}$	$\frac{1}{3}$
$\sum$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

The function $f : (x, y) \mapsto \mathbb{P}(X = x, Y = y)$ is called the *joint pmf* of X and Y. The following definition is just a generalisation of this.

Definition 4.1 Let $X_1, \ldots, X_n$ be *discrete* random variables. The function f defined by $f(x_1, \ldots, x_n) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$ is called the **joint probability mass function** (pmf) of $X_1, \ldots, X_n$.

We sometimes write $f_{X_1,\dots,X_n}$ instead of f to show that this is the pmf of the random variables $X_1, \dots, X_n$. Or, if $\boldsymbol{X}$ is the corresponding random vector, we could write $f_{\boldsymbol{X}}$ instead.

Note that, by the sum rule, if we are given the joint pmf of $X_1, \dots, X_n$ we can in principle calculate *all possible probabilities* involving these random variables. For example, in the 2-dimensional case

$$\mathbb{P}((X,Y) \in B) = \sum_{(x,y)\in B} \mathbb{P}(X = x, Y = y) ,$$

for any subset B of possible values for (X, Y). In particular, we can find the pmf of X by summing the joint pmf over all possible values of y:

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

The converse is *not* true: from the individual distributions (so-called **marginal** distribution) of X and Y we cannot in general reconstruct the joint distribution of X and Y. We are missing the "dependency" information. E.g., in Example 4.1 we cannot reconstruct the inside of the two-dimensional table if only given the column and row totals.

However, there is one *important exception* to this, namely when we are dealing with *independent* random variables. We have so far only defined what independence is for *events*. The following definition says that random variables $X_1, \dots, X_n$ are independent if the events $\{X_1 \in A_1\}, \dots, \{X_n \in A_n\}$ are independent for any subsets $A_1, \dots, A_n$ of $\mathbb{R}$. Intuitively, this means that any information about one of them does not affect our knowledge about the others.

Definition 4.2 The random variables $X_1, \dots, X_n$ are called **independent** if for all $A_1, \dots, A_n$, with $A_i \subset \mathbb{R}$, $i = 1, \dots, n$

$$\mathbb{P}(X_1 \in A_1, \dots, X_n \in A_n) = \mathbb{P}(X_1 \in A_1) \cdots \mathbb{P}(X_n \in A_n) .$$

The following theorem is a direct consequence of the definition above.

Theorem 4.1 Discrete random variables $X_1, \dots, X_n$, are independent if and only if

$$\boxed{\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = \mathbb{P}(X_1 = x_1) \cdots \mathbb{P}(X_n = x_n),} \qquad (4.1)$$

for all $x_1, x_2, \dots, x_n$.

Proof. The necessary condition is obvious: if $X_1, \dots, X_n$ are independent random variables, then $\{X_1 = x_1\}, \dots, \{X_n = x_n\}$ are (mutually) independent events. To prove the sufficient condition, write

$$\mathbb{P}(X_1 \in A_1, \dots, X_n \in A_n) = \sum_{x_1\in A_1} \cdots \sum_{x_n\in A_n} \mathbb{P}(X_1 = x_1, \dots, X_n \in x_n).$$

Then, if (4.1) holds, the multiple sum can be written as

$$\sum_{x_1 \in A_1} \mathbb{P}(X_1 = x_1) \cdots \sum_{x_n \in A_n} \mathbb{P}(X_n \in x_n) = \mathbb{P}(X_1 \in A_1) \cdots \mathbb{P}(X_n \in A_n),$$

which implies that $X_1, \ldots, X_n$ are independent random variables. ■

Example 4.2 We repeat the experiment in Example 4.1 with three ordinary fair dice. What are now the joint probabilities in the table? Since the events $\{X = x\}$ and $\{Y = y\}$ are now independent, each entry in the pmf table is $\frac{1}{3} \times \frac{1}{6}$. Clearly in the first experiment not *all* events $\{X = x\}$ and $\{Y = y\}$ are independent (why not?).

Example 4.3 (Coin Flip Experiment) Consider the experiment where we flip a coin n times. We can model this experiments in the following way. For $i = 1, \ldots, n$ let X_i be the result of the ith toss: $\{X_i = 1\}$ means Heads, $\{X_i = 0\}$ means Tails. Also, let

$$\mathbb{P}(X_i = 1) = p = 1 - \mathbb{P}(X_i = 0), \quad i = 1, 2, \ldots, n\ .$$

Thus, p can be interpreted as the probability of Heads, which may be known or unknown. Finally, assume that $X_1, \ldots, X_n$ are *independent*.

This completely specifies our model. In particular we can find any probability related to the X_i's. For example, let $X = X_1 + \cdots + X_n$ be the total number of Heads in n tosses. Obviously X is a random variable that takes values between 0 and n. Denote by A the set of all binary vectors $\boldsymbol{x} = (x_1, \ldots, x_n)$ such that $\sum_{i=1}^n x_i = k$. Note that A has $\binom{n}{k}$ elements. We now have

$$\begin{aligned}
\mathbb{P}(X = k) &= \sum_{\boldsymbol{x} \in A} \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) \\
&= \sum_{\boldsymbol{x} \in A} \mathbb{P}(X_1 = x_1) \cdots \mathbb{P}(X_n = x_n) = \sum_{\boldsymbol{x} \in A} p^k (1-p)^n \\
&= \binom{n}{k} p^k (1-p)^n\ .
\end{aligned}$$

In other words, $X \sim \mathsf{Bin}(n, p)$. Compare this to what we did in Example 1.15 on page 27.

Remark 4.1 If $f_{X_1,\ldots,X_n}$ denotes the joint pmf of $X_1, \ldots, X_n$ and f_{X_i} the marginal pmf of X_i, $i = 1, \ldots, n$, then the theorem above states that independence of the X_i's is equivalent to

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$$

for all possible $x_1, \ldots, x_n$.

Remark 4.2 An *infinite* sequence $X_1, X_2, \ldots$ of random variables is called independent if for any finite choice of parameters $i_1, i_2, \ldots, i_n$ (none of them the same) the random variables $X_{i_1}, \ldots, X_{i_n}$ are independent.

Multinomial Distribution

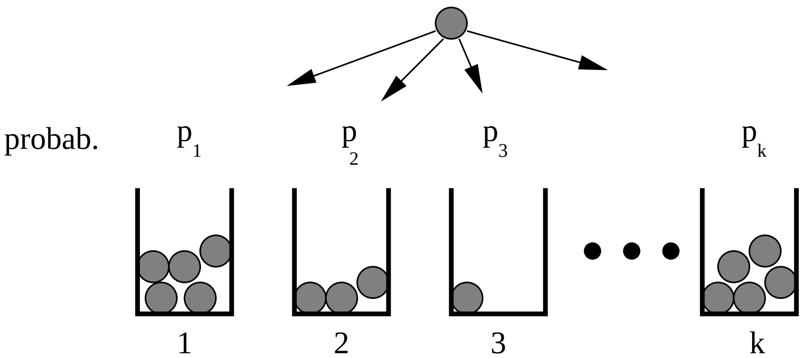
An important discrete joint distribution is the multinomial distribution. It can be viewed as a generalisation of the binomial distribution. First we give the definition, then an example how this distribution arises in applications.

Definition 4.3 We say that $(X_1, X_2, \ldots, X_k)$ has a **multinomial** distribution, with parameters n and $p_1, p_2, \ldots, p_k$, if

$$\mathbb{P}(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1!\, x_2! \cdots x_k!}\, p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} , \tag{4.2}$$

for all $x_1, \ldots, x_k \in \{0, 1, \ldots, n\}$ such that $x_1 + x_2 + \cdots + x_k = n$. We write $(X_1, \ldots, X_k) \sim \mathsf{Mnom}(n, p_1, \ldots, p_k)$.

Example 4.4 We independently throw n balls into k urns, such that each ball is thrown in urn i with probability p_i, $i = 1, \ldots, k$.



Let X_i be the total number of balls in urn i, $i = 1, \ldots, k$. We show that $(X_1, \ldots, X_k) \sim \mathsf{Mnom}(n, p_1, \ldots, p_k)$. Let $x_1, \ldots, x_k$ be integers between 0 and n that sum up to n. The probability that the *first* x_1 balls fall in the first urn, the *next* x_2 balls fall in the second urn, etcetera, is

$$p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} .$$

To find the probability that there are x_1 balls in the first urn, x_2 in the second, etcetera, we have to multiply the probability above with the number of ways in which we can fill the urns with $x_1, x_2, \ldots, x_k$ balls, i.e. $n!/(x_1!\, x_2! \cdots x_k!)$. This gives (4.2).

Remark 4.3 Note that for the *binomial* distribution there are only *two* possible urns. Also, note that for each $i = 1, \ldots, k$, $X_i \sim \mathsf{Bin}(n, p_i)$.

4.1.2 Continuous Joint Distributions

Joint distributions for continuous random variables are usually defined via the joint pdf. The results are very similar to the discrete case discussed in

Section 4.1.1. Compare this section also with the 1-dimensional case in Section 2.2.2.

Definition 4.4 We say that the continuous random variables $X_1, \ldots, X_n$ have a **joint probability density function** (pdf) f if

$$\mathbb{P}(a_1 \le X_1 \le b_1, \ldots, a_n \le X_n \le b_n) = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} f(x_1, \ldots, x_n)\, dx_1 \cdots dx_n$$

for all $a_1, \ldots, b_n$.

We sometimes write $f_{X_1,\ldots,X_n}$ instead of f to show that this is the pdf of the random variables $X_1, \ldots, X_n$. Or, if $\boldsymbol{X}$ is the corresponding random vector, we could write $f_{\boldsymbol{X}}$ instead.

We can interpret $f(x_1, \ldots, x_n)$ as a continuous analogue of a pmf, or as the "density" that $X_1 = x_1$, $X_2 = x_2$, $\ldots$, and $X_n = x_n$. For example in the 2-dimensional case:

$$\begin{aligned}&\mathbb{P}(x \le X \le x + h,\ y \le Y \le y + h)\\ &= \int_x^{x+h} \int_y^{y+h} f(u, v)\, du\, dv \approx h^2 f(x, y)\ .\end{aligned}$$

Note that if the joint pdf is given, then in principle we can calculate *all probabilities*. Specifically, in the 2-dimensional case we have

$$\mathbb{P}((X, Y) \in B) = \int\int_{(x,y)\in B} f(x, y)\, dx\, dy\ , \tag{4.3}$$

for any subset B of possible values for $\mathbb{R}^2$. Thus, the calculation of probabilities is reduced to *integration*.

Similarly to the discrete case, if $X_1, \ldots, X_n$ have joint pdf f, then the (individual, or marginal) pdf of each X_i can be found by integrating f over all other variables. For example, in the two-dimensional case

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

However, we usually cannot reconstruct the joint pdf from the marginal pdf's unless we assume that the the random variables are *independent*. The definition of independence is exactly the same as for discrete random variables, see Definition 4.2. But, more importantly, we have the following analogue of Theorem 4.1.

Theorem 4.2 Let $X_1, \ldots, X_n$ be continuous random variables with joint pdf f and marginal pdf's $f_{X_1}, \ldots, f_{X_n}$. The random variables $X_1, \ldots, X_n$ are independent if and only if, for all $x_1, \ldots, x_n$,

$$\boxed{f(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n).} \tag{4.4}$$

Example 4.5 Consider the experiment were we select randomly and independently n points from the interval [0,1]. We can carry this experiment out using a calculator or computer, using the *random generator*. On your calculator this means pushing the RAN# or Rand button. Here is a possible outcome, or **realisation**, of the experiment, for $n = 12$.

0.9451226800	0.2920864820	0.0019900900	0.8842189383	0.8096459523
0.3503489150	0.9660027079	0.1024852543	0.7511286891	0.9528386400
0.2923353821	0.0837952423			

A model for this experiment is: Let $X_1, \ldots, X_n$ be independent random variables, each with a uniform distribution on [0,1]. The joint pdf of $X_1, \ldots, X_n$ is very simple, namely

$$f(x_1, \ldots, x_n) = 1, \quad 0 \le x_1 \le 1, \ \ldots, \quad 0 \le x_n \le 1 \ ,$$

(and 0 else). In principle we can now calculate any probability involving the X_i's. For example for the case $n = 2$ what is the probability

$$\mathbb{P}\left(\frac{X_1 + X_2^2}{X_1 X_2} > \sin(X_1^2 - X_2)\right) ?$$

The answer, by (4.3), is

$$\iint_A 1 \, dx_1 \, dx_2 = \text{ Area } (A),$$

where

$$A = \left\{(x_1, x_2) \in [0,1]^2 \ : \ \frac{x_1 + x_2^2}{x_1 x_2} > \sin(x_1^2 - x_2)\right\} \ .$$

(Here $[0,1]^2$ is the unit square in $\mathbb{R}^2$).

Remark 4.4 The type of model used in the previous example, i.e., $X_1, \ldots, X_n$ are independent and all have the same distribution, is the most widely used model in statistics. We say that $X_1, \ldots, X_n$ is a **random sample** of **size** n, from some given distribution. In Example 4.5 $X_1, \ldots, X_n$ is a random sample from a $\mathsf{U}[0,1]$-distribution. In Example 4.3 we also had a random sample, this time from a $\mathsf{Ber}(p)$-distribution. The common distribution of a random sample is sometimes called the **sampling distribution**.

Using the computer we can generate the outcomes of random samples from many (sampling) distributions. In Figure 4.1 the outcomes of a two random samples, both of size 1000 are depicted in a **histogram**. Here the x-axis is divived into 20 intervals, and the number of points in each interval is counted. The first sample is from the $\mathsf{U}[0,1]$-distribution, and the second sample is from the $\mathsf{N}(1/2, 1/12)$-distribution. The matlab commands are:

```
figure(1)
hist(rand(1,1000),20)
figure(2)
hist(1/2 + randn(1,1000)*sqrt(1/12),20)
```

Note that the true expectation and variance of the distributions are the same. However, the "density" of points in the two samples is clearly different, and follows that shape of the corresponding pdf's.

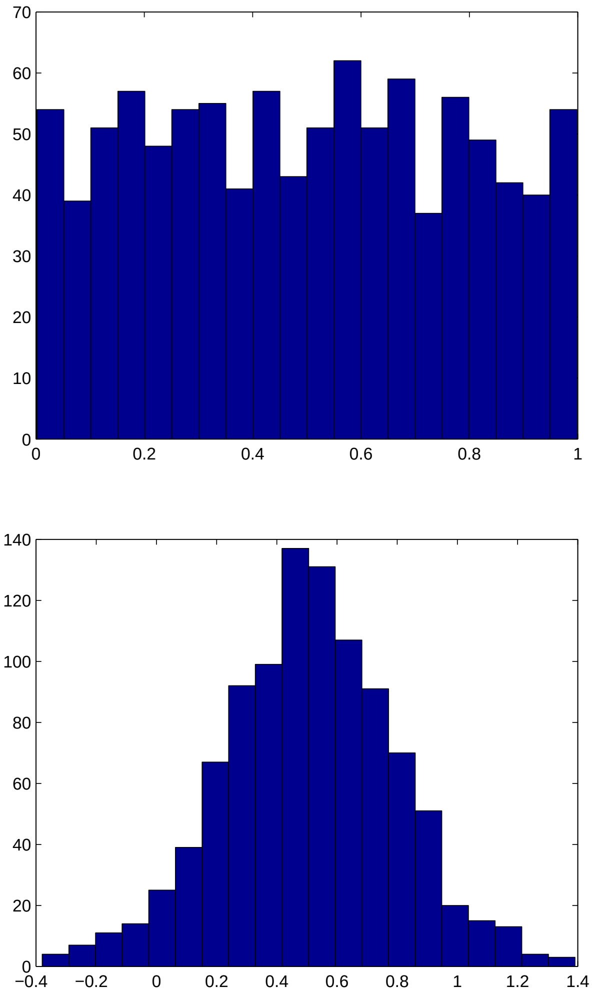


Figure 4.1: A histogram of a random sample of size 100 from the $\mathsf{U}[0,1]$-distribution (above) and the $\mathsf{N}(1/2, 1/12)$-distribution (below).

4.2 Expectation

Similar to the 1-dimensional case, the expected value of any real-valued function of $X_1, \ldots, X_n$ is the weighted average of all values that this function can take.

Specifically, if $Z = g(X_1, \ldots, X_n)$ then in the discrete case

$$\mathbb{E}Z = \sum_{x_1} \cdots \sum_{x_n} g(x_1, \ldots, x_n)\, f(x_1, \ldots, x_n)\ ,$$

where f is the joint pmf; and in the continuous case

$$\mathbb{E}Z = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \ldots, x_n) f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n\ ,$$

where f is the joint pdf.

Example 4.6 Let X and Y be continuous, possibly *dependent*, r.v.'s with joint pdf f. Then,

$$\begin{aligned}
\mathbb{E}(X+Y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x+y) f(x,y)\, dxdy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f(x,y)\, dxdy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f(x,y)\, dxdy \\
&= \int_{-\infty}^{\infty} x f_X(x)\, dx + \int_{-\infty}^{\infty} y f_Y(y)\, dy \\
&= \mathbb{E}X + \mathbb{E}Y\ .
\end{aligned}$$

The previous example is easily generalised to the following result:

Theorem 4.3 Suppose $X_1,\ X_2,\ \ldots,\ X_n$ are discrete or continuous random variables with means $\mu_1, \mu_2, \ldots, \mu_n$. Let

$$Y = a + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n$$

where $a,\ b_1,\ b_2,\ \ldots,\ b_n$ are constants. Then

$$\boxed{\begin{aligned}
\mathbb{E}Y &= a + b_1 \mathbb{E}X_1 + \cdots + b_n \mathbb{E}X_n \\
&= a + b_1 \mu_1 + \cdots + b_n \mu_n
\end{aligned}}$$

Example 4.7 (Expectation of $\mathsf{Bin}(n,p)$ and $\mathsf{Hyp}(n,r,N)$) We can now prove that the expectation of a $\mathsf{Bin}(n,p)$ random variable is np, without having to resort to difficult arguments. Namely, if $X \sim \mathsf{Bin}(n,p)$, then X can be written as the sum $X_1+\cdots+X_n$ of iid $\mathsf{Ber}(p)$ random variables, see Example 4.3. Thus,

$$\mathbb{E}X = \mathbb{E}(X_1 + \cdots + X_n) = \mathbb{E}X_1 + \cdots + \mathbb{E}X_n = \underbrace{p + \cdots p}_{n \text{ times}} = np\ , \qquad (4.5)$$

because the expectation of each X_i is p.

Notice that we do not use anywhere the independence of the $\{X_i\}$. Now, let $X \sim \mathsf{Hyp}(n,r,N)$, and let $p = r/N$. We can think of X as the total number of red balls when n balls are drawn from an urn with r red balls and $N - r$ other balls. Without loss of generality we may assume the balls are drawn one-by-one. Let $X_i = 1$ if the i-th balls is red, and 0 otherwise. Then, again $X_1 + \cdots + X_n$, and each $X_i \sim \mathsf{Ber}(p)$, but now the $\{X_i\}$ are *dependent*. However, this does not affect the result (4.5), so that the expectation of X is $np = nr/N$.

Another important result is the following.

Theorem 4.4 If $X_1, \ldots, X_n$ are *independent*, then

$$\boxed{\mathbb{E}X_1 X_2 \cdots X_n = \mathbb{E}X_1 \mathbb{E}X_2 \cdots \mathbb{E}X_n \ .}$$

Proof. We prove it only for the 2-dimensional continuous case. Let f denote the joint pdf of X and Y, and f_X and f_Y the marginals. Then, $f(x, y) = f_X(x) f_Y(y)$ for all x, y. Thus

$$\begin{aligned} \mathbb{E}XY &= \iint x\, y\, f(x,y)\, dx\, dy = \iint x\, y\, f_X(x)\, f_Y(y)\, dx\, dy \\ &= \int x f_X(x) dx \int f_Y(y) dy = \mathbb{E}X\mathbb{E}Y \ . \end{aligned}$$

The generalization to the n-dimensional continuous/discrete case is obvious. ■

Theorem 4.4 is particularly handy in combination with transform techniques. We give two examples.

Example 4.8 Let $X \sim \mathsf{Poi}(\lambda)$, then we saw in Example 2.8 on page 38 that its PGF is given by

$$G(z) = \mathrm{e}^{-\lambda(1-z)} \ . \tag{4.6}$$

Now let $Y \sim \mathsf{Poi}(\mu)$ be independent of X. Then, the PGF of $X + Y$ is given by

$$\mathbb{E}[z^{X+Y}] = \mathbb{E}[z^X]\, \mathbb{E}[z^Y] = \mathrm{e}^{-\lambda(1-z)} \mathrm{e}^{-\mu(1-z)} = \mathrm{e}^{-(\lambda+\mu)(1-z)} \ .$$

Thus, by the uniqueness property of the MGF, $X + Y \sim \mathsf{Poi}(\lambda + \mu)$.

Example 4.9 The MGF of $X \sim \mathsf{Gam}(\alpha, \lambda)$ is given, see (2.13) on page 52, by

$$\mathbb{E}[\mathrm{e}^{sX}] = \left(\frac{\lambda}{\lambda - s}\right)^{\alpha} \ .$$

As a special case, the moment generating function of the $\mathsf{Exp}(\lambda)$ distribution is given by $\lambda/(\lambda - s)$. Now let $X_1, \ldots, X_n$ be iid $\mathsf{Exp}(\lambda)$ random variables. The MGF of $S_n = X_1 + \cdots + X_n$ is

$$\mathbb{E}[\mathrm{e}^{sS_n}] = \mathbb{E}[\mathrm{e}^{sX_1} \cdots \mathrm{e}^{sX_n}] = \mathbb{E}[\mathrm{e}^{sX_1}] \cdots \mathbb{E}[\mathrm{e}^{sX_n}] = \left(\frac{\lambda}{\lambda - s}\right)^{n} \ ,$$

which shows that $S_n \sim \mathsf{Gam}(n, \lambda)$.

Definition 4.5 (Covariance) The **covariance** of two random variables X and Y is defined as the number

$$\mathrm{Cov}(X,Y) := \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y).$$

The covariance is a measure for the amount of linear dependency between the variables. If small values of X (smaller than the expected value of X) go together with small values of Y, and at the same time large values of X go together with large values of Y, then $\mathrm{Cov}(X,Y)$ will be *positive*. If on the other hand small values of X go together with large values of Y, and large values of X go together with small values of Y, then $\mathrm{Cov}(X,Y)$ will be *negative*.

For easy reference we list some important properties of the variance and covariance. The proofs follow directly from the definitions of covariance and variance and the properties of the expectation.

1.	$\mathrm{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2$.
2.	$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$.
3.	$\mathrm{Cov}(X,Y) = \mathbb{E}XY - \mathbb{E}X\mathbb{E}Y$.
4.	$\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$.
5.	$\mathrm{Cov}(aX + bY, Z) = a\,\mathrm{Cov}(X,Z) + b\,\mathrm{Cov}(Y,Z)$.
6.	$\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.
7.	$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$.
8	X and Y indep. $\Longrightarrow \mathrm{Cov}(X,Y) = 0$.

Table 4.1: Properties of variance and covariance

Proof. Properties 1. and 2. were already proved on page 38. Properties 4. and 6. follow directly from the definitions of covariance and variance. Denote for convenience $\mu_X = \mathbb{E}X$ and $\mu_Y = \mathbb{E}Y$.

3. $\mathrm{Cov}(X,Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY - Y\mu_X - X\mu_Y + \mu_X\mu_Y] = \mathbb{E}XY - \mu_Y\mu_X - \mu_X\mu_Y + \mu_X\mu_Y = \mathbb{E}XY - \mu_X\mu_Y$.

5. Using property 3. we get

$$\begin{aligned}\mathrm{Cov}(aX + bY, Z) &= \mathbb{E}[(aX + bY)Z] - \mathbb{E}[aX + bY]\,\mathbb{E}[Z] \\ &= a\,\mathbb{E}[XZ] + b\,\mathbb{E}[YZ] - a\,\mathbb{E}[X]\,\mathbb{E}[Z] - b\,\mathbb{E}[Y]\,\mathbb{E}[Z] \\ &= a\{\mathbb{E}[XZ] - \mathbb{E}[X]\mathbb{E}[Z]\} + b\{\mathbb{E}[YZ] - \mathbb{E}[Y]\mathbb{E}[Z]\} \\ &= a\,\mathrm{Cov}(X,Z) + b\,\mathrm{Cov}(Y,Z).\end{aligned}$$

7. By property 6. we have

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

By property 5. we can expand this to

$$\begin{aligned}\mathrm{Cov}(X+Y, X+Y) &= \mathrm{Cov}(X, X+Y) + \mathrm{Cov}(Y, X+Y)\\ &= \mathrm{Cov}(X,X) + \mathrm{Cov}(X,Y) + \mathrm{Cov}(Y,X) + \mathrm{Cov}(Y,Y),\end{aligned}$$

where we have also used the symmetry property 4. to expand the second argument of the covariance. Now, by properties 4. and 6. we can simplify the above sum to $\mathrm{Var}(X)+\mathrm{Var}(Y)+2\mathrm{Cov}(X,Y)$, which had to be shown.

8. If X and Y are independent, then $\mathbb{E}XY = \mathbb{E}X\mathbb{E}Y$, so that $\mathrm{Cov}(X,Y) = \mathbb{E}XY - \mathbb{E}X\mathbb{E}Y = 0$.

■

A scaled version of the covariance is given by the **correlation coefficient**.

$$\rho(X,Y) := \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}.$$

Theorem 4.5 The correlation coefficient always lies between -1 and 1.

PROOF. Let a be an arbitrary real number, and denote the standard deviations of X and Y by σ_X and σ_Y. Obviously the variance of $\pm aX + Y$ is always non-negative. Thus, using the properties of covariance and variance

$$\mathrm{Var}(-aX+Y) = a^2\sigma_X^2 + \sigma_Y^2 - 2a\mathrm{Cov}(X,Y) \geq 0.$$

So that after rearranging and dividing by $\sigma_x\sigma_Y$, we obtain

$$\rho(X,Y) \leq \frac{1}{2}\left(\frac{a\,\sigma_X}{\sigma_Y} + \frac{\sigma_Y}{a\sigma_X}\right).$$

Similarly,

$$\mathrm{Var}(aX+Y) = a^2\sigma_X^2 + \sigma_Y^2 + 2a\mathrm{Cov}(X,Y) \geq 0,$$

so that

$$\rho(X,Y) \geq -\frac{1}{2}\left(\frac{a\,\sigma_X}{\sigma_Y} + \frac{\sigma_Y}{a\sigma_X}\right).$$

By choosing $a = \sigma_Y/\sigma_X$, we see that $-1 \leq \rho(X,Y) \leq 1$. ■

In Figure 4.2 an illustration of the correlation coefficient is given. Each figure corresponds to samples of size 40 from a different 2-dimensional distribution. In each case $\mathbb{E}X = \mathbb{E}Y = 0$ and $\mathrm{Var}(X) = \mathrm{Var}(Y) = 1$.

As a consequence of properties 2. and 7., we have

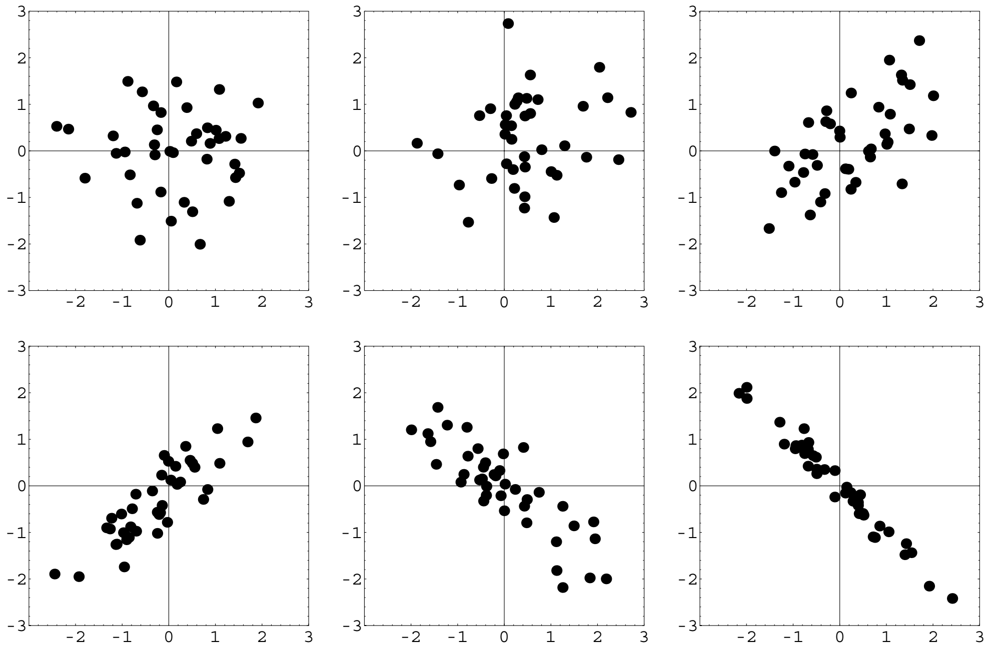


Figure 4.2: *Illustration of correlation coefficient. Above:* $\rho = 0$, $\rho = 0.4$, $\rho = 0.7$. *Below:* $\rho = 0.9$, $\rho = -0.8$, $\rho = -0.98$.

Theorem 4.6 Suppose $X_1, X_2, \ldots, X_n$ are discrete or continuous *independent* random variables with variances $\sigma_1^2, \sigma^2, \ldots, \sigma_n^2$. Let

$$Y = a + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n$$

where a, b_1, b_2, $\ldots$, b_n are constants. Then

$$\boxed{\begin{aligned} \mathrm{Var}(Y) &= b_1^2\,\mathrm{Var}(X_1) + \cdots + b_n^2\,\mathrm{Var}(X_n) \\ &= b_1^2\,\sigma_1^2 + \cdots + b_n^2\,\sigma_n^2 \end{aligned}}$$

Proof. By virtue of property 6., and repetitive application of property 5., we have (note that the constant a does not play a role in the variance):

$$\begin{aligned} \mathrm{Var}(Y) &= \mathrm{Cov}(b_1X_1 + b_2X_2 + \cdots + b_nX_n,\ b_1X_1 + b_2X_2 + \cdots + b_nX_n) \\ &= \sum_{i=1}^{n} \mathrm{Cov}(b_iX_i, b_iX_i) + 2\sum_{i<j} \mathrm{Cov}(b_iX_i, b_jX_j)\ . \end{aligned}$$

Since $\mathrm{Cov}(b_iX_i, b_iX_i) = b_i^2\mathrm{Var}(X_i)$ and all covariance term are zero because of the independence of X_i and $X_J j$ $(i \neq j)$, the result follows. ■

Example 4.10 (Variance of $\mathsf{Bin}(n,p)$ and $\mathsf{Hyp}(n,r,N)$) Consider again Example 4.7 where we derived the expectation of $X \sim \mathsf{Bin}(n,p)$ and $X \sim \mathsf{Hyp}(n,r,N)$

by writing X as

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

of independent (in the binomial case) or dependent (in the hypergeometric case) $\mathsf{Ber}(p)$ random variables, where $p = r/N$ in the hypergeometric case. Using Theorem 4.6, the variance of the binomial distribution follows directly from

$$\mathrm{Var}(X) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) = n\mathrm{Var}(X_1) = np(1-p).$$

For the hypergeometric case must include the covariance terms as well:

$$\mathrm{Var}(X) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) + 2\sum_{i<j} \mathrm{Cov}(X_i, X_j) \ .$$

By symmetry all the $\mathrm{Cov}(X_i, X_j)$ are the same ($i \neq j$). Hence,

$$\mathrm{Var}(X) = n\mathrm{Var}(X_1) + n(n-1)\mathrm{Cov}(X_1, X_2) \ .$$

Since $\mathrm{Var}(X_1) = p(1-p)$, and $\mathrm{Cov}(X_1, X_2) = \mathbb{E}X_1X_2 - p^2$, it remains to find $\mathbb{E}X_1X_2 = \mathbb{P}(X_1 = 1, X_2 = 1)$, which is

$$\mathbb{P}(X_1 = 1, X_2 = 1) = \mathbb{P}(X_1 = 1)\mathbb{P}(X_2 = 1 \,|\, X_1 = 1) = p\,\frac{r-1}{N-1} \ .$$

Simplifying gives,

$$\mathrm{Var}(X) = np(1-p)\frac{N-n}{N-1} \ .$$

Expectation Vector and Covariance Matrix

Let $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ be a random vector. Sometimes it is convenient to write the expectations and covariances in vector notation.

Definition 4.6 For any random vector $\boldsymbol{X}$ we define the **expectation vector** as the vector of expectations

$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T := (\mathbb{E}X_1, \ldots, \mathbb{E}X_n)^T \ .$$

The **covariance matrix** Σ is defined as the matrix whose (i,j)th element is

$$\mathrm{Cov}(X_i, X_j) = \mathbb{E}(X_i - \mu_i)(X_j - \mu_j) \ .$$

If we define the expectation of a vector (matrix) to be the vector (matrix) of expectations, then we can write:

$$\boldsymbol{\mu} = \mathbb{E}\boldsymbol{X}$$

and

$$\Sigma = \mathbb{E}(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T \ .$$

Note that $\boldsymbol{\mu}$ and Σ take the same role as μ and σ^2 in the 1-dimensional case. We sometimes write $\boldsymbol{\mu}_{\boldsymbol{X}}$ and $\Sigma_{\boldsymbol{X}}$ if we wish to emphasise that $\boldsymbol{\mu}$ and Σ belong to the vector $\boldsymbol{X}$.

Remark 4.5 Note that any covariance matrix Σ is a *symmetric* matrix. In fact, it is *positive semi-definite*, i.e., for any (column) vector $\boldsymbol{u}$, we have

$$\boldsymbol{u}^T \Sigma \boldsymbol{u} \geq 0 .$$

To see this, suppose Σ is the covariance matrix of some random vector $\boldsymbol{X}$ with expectation vector $\boldsymbol{\mu}$. Write $\boldsymbol{Y} = \boldsymbol{X} - \boldsymbol{\mu}$. Then

$$\begin{aligned} \boldsymbol{u}^T \Sigma \boldsymbol{u} &= \boldsymbol{u}^T \mathbb{E} \boldsymbol{Y}\boldsymbol{Y}^T \boldsymbol{u} = \mathbb{E}\, \boldsymbol{u}^T \boldsymbol{Y}\boldsymbol{Y}^T \boldsymbol{u} \\ &= \mathbb{E} \left(\boldsymbol{Y}^T\boldsymbol{u}\right)^T \boldsymbol{Y}^T\boldsymbol{u} = \mathbb{E} \left(\boldsymbol{Y}^T\boldsymbol{u}\right)^2 \geq 0. \end{aligned}$$

Note that $\boldsymbol{Y}^T\boldsymbol{u}$ is a random variable.

4.3 Conditional Distribution

Suppose X and Y are both discrete or both continuous, with joint pmf/pdf $f_{X,Y}$, and suppose $f_X(x) > 0$. The *conditional pdf/pmf* of Y given $X = x$ is defined as

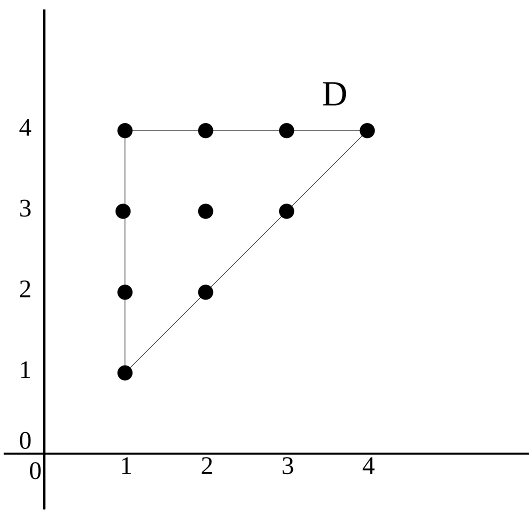
$$f_{Y|X}(y \,|\, x) = \frac{f_{X,Y}(x,y)}{f_X(x)}, \text{ for all } y \,. \tag{4.7}$$

We can interpret $f_{Y|X}(\cdot \,|\, x)$ as the pmf/pdf of Y given the information that X takes the value x. For discrete random variables the definition is simply a consequence or rephrasing of the conditioning formula (1.4) on page 23. Namely,

$$f_{Y|X}(y \,|\, x) = \mathbb{P}(Y = y \,|\, X = x) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(X = x)} = \frac{f_{X,Y}(x,y)}{f_X(x)} \,.$$

Example 4.11 We draw "uniformly" a point (X, Y) from the 10 points on the triangle D below. Thus, each point is equally likely to be dawn. That is, the joint pmf is The joint and marginal pmf's are easy to determine:

$$f_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y) = \frac{1}{10}, \; (x,y) \in D,$$



The marginal pmf's of X and Y are

$$f_X(x) = \mathbb{P}(X = x) = \frac{5-x}{10},\ x \in \{1,2,3,4\},$$

and

$$f_Y(y) = \mathbb{P}(Y = y) = \frac{y}{10},\ y \in \{1,2,3,4\}.$$

Clearly X and Y are not independent. In fact, if we know that $X = 2$, then Y can only take the values $j = 2, 3$ or 4. The corresponding probabilities are

$$\mathbb{P}(Y = y \mid X = 2) = \frac{\mathbb{P}(Y = y, X = 2)}{\mathbb{P}(X = 2)} = \frac{1/10}{3/10} = \frac{1}{3}.$$

In other words, the conditional pmf of Y given $X = 2$ is

$$f_{Y \mid X}(y \mid 2) = \frac{f_{X,Y}(2, y)}{f_X(2)} = \frac{1}{3}, \quad y = 2, 3, 4\ .$$

Thus, given $X = 2$, Y takes the values 2,3 and 4 with equal probability.

When X is *continuous*, we can not longer directly apply (1.4) to define the conditional density. Instead, we define first the **conditional cdf** of Y given $X = x$ as the limit

$$F_Y(y \mid x) := \lim_{h \to 0} F_Y(y \mid x < X \leq x + h)\ .$$

Now, (1.4) can be applied to $F_Y(y \mid x < X \leq x + h)$ to yield

$$F_Y(y \mid x < X \leq x + h) = \frac{\int_{-\infty}^{y} \int_{x}^{x+h} f_{X,Y}(u, v)\, du\, dv}{\int_{x}^{x+h} f_X(u)\, du}$$

Now, for small h the integral $\int_{x}^{x+h} f_{X,Y}(u, v)\, du$ is approximately equal to $h f_{X,Y}(x, v)$ plus some small term which goes to zero faster than h. Similarly, $\int_{x}^{x+h} f_X(u)\, du \approx h f_X(x)$ (plus smaller order terms). Hence, for $h \to 0$, the limit of $F_Y(y \mid x < X \leq x + h)$ is

$$F_Y(y \mid x) = \frac{\int_{-\infty}^{y} f_{X,Y}(x, v)\, dv}{f_X(x)}.$$

Note that $F_Y(y \mid x)$ as a function of y has all the properties of a cdf. By differentiating this cdf with respect to y we obtain the conditional pdf of Y given $X = x$ in the continuous case, which gives the same formula (4.7) as for the discrete case.

Since the conditional pmf (pdf) has all the properties of a probability mass (density) function, it makes sense to define the corresponding *conditional expectation* as (in the continuous case)

$$\mathbb{E}[Y \mid X = x] = \int y\, f_{Y|X}(y \mid x)\, dy\ .$$

Conditional pmf's and pdf's for more than two random variables are defined analogously. For example, the conditional pmf of X_n given $X_1, \ldots, X_{n-1}$ is given by

$$f_{X_n \,|\, X_1,\ldots,X_{n-1}}(x_n \,|\, x_1, \ldots, x_{n-1}) = \frac{\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)}{\mathbb{P}(X_1 = x_1, \ldots, X_{n-1} = x_{n-1})} \,.$$

Chapter 5

Functions of Random Variables and Limit Theorems

Suppose $X_1, \ldots, X_n$ are the measurements on a random experiment. Often we are interested in certain *functions* of the measurements only, rather than all measurements themselves. For example, if $X_1, \ldots, X_n$ are the repeated measurements of the strength of a certain type of fishing line, then what we are really interested in is not the individual values for $X_1, \ldots, X_n$ but rather quantities such as the average strength $(X_1 + \cdots + X_n)/n$, the minimum strength $\min(X_1, \ldots, X_n)$ and the maximum strength $\max(X_1, \ldots, X_n)$. Note that these quantities are again random variables. The distribution of these random variables can in principle be derived from the joint distribution of the X_i's. We give a number of examples.

Example 5.1 Let X be a continuous random variable with pdf f_X, and let $Y = aX + b$, where $a \neq 0$. We wish to determine the pdf f_Y of Y. We first express the cdf of Y in terms of the cdf of X. Suppose first that $a > 0$. We have for any y

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(X \leq (y - b)/a) = F_X((y - b)/a) \; .$$

Differentiating this with respect to y gives $f_Y(y) = f_X((y-b)/a)\,/a$. For $a < 0$ we get similarly $f_Y(y) = f_X((y-b)/a)\,/(-a)$. Thus in general

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right) \; . \tag{5.1}$$

Example 5.2 Let $X \sim \mathsf{N}(0,1)$. We wish to determine the distribution of $Y = X^2$. We can use the same technique as in the example above, but note first that Y can only take values in $[0, \infty)$. For $y > 0$ we have

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) = \mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}) = 2F_X(\sqrt{y}) - 1 \; . \end{aligned}$$

Differentiating this with respect to y gives

$$\begin{aligned} f_Y(y) &= 2\, f_X(\sqrt{y})\, \frac{1}{2\sqrt{y}} = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\sqrt{y})^2\right) \frac{1}{\sqrt{y}} \\ &= \frac{(1/2)^{1/2}\, y^{-1/2}\, \mathrm{e}^{-y/2}}{\Gamma(1/2)}. \end{aligned}$$

This is exactly the formula for the pdf of a χ^2_1-distribution. Thus $Y \sim \chi^2_1$.

Example 5.3 (Minimum and Maximum) Suppose $X_1, \ldots, X_n$ are independent and have cdf F. Let $Y = \min(X_1, \ldots, X_n)$ and $Z = \max(X_1, \ldots, X_n)$. The cdf of Y and Z are easily obtained. First, note that the maximum of the $\{X_i\}$ is less than some number Z if and only if *all* X_i are less than z. Thus,

$$\mathbb{P}(Z \le z) = \mathbb{P}(X_1 \le z, X_2 \le z, \ldots, X_n \le z) = \mathbb{P}(X_1 \le z)\mathbb{P}(X_2 \le z) \cdots \mathbb{P}(X_n \le z),$$

where the second equation follows from the independence assumption. It follows that

$$F_Z(z) = (F(z))^n \ .$$

Similarly,

$$\mathbb{P}(Y > y) = \mathbb{P}(X_1 > y, X_2 > y, \ldots, X_n > y) = \mathbb{P}(X_1 > y)\mathbb{P}(X_2 > y) \cdots \mathbb{P}(X_n > y),$$

so that

$$F_Y(y) = 1 - (1 - F(y))^n \ .$$

Example 5.4 In Chapter 3 we saw an important application of functions of random variables: the inverse-transform method for generating random variables. That is, $U \sim \mathsf{U}(0,1)$, and let F be continuous and strictly increasing cdf. Then $Y = F^{-1}(U)$ is a random variable that has cdf F.

We can use simulation to get an idea of the distribution of a function of one or more random variables, as explained in the following example.

Example 5.5 Let X and Y be independent and both $\mathsf{U}(0,1)$ distributed. What what does the pdf of $Z = X + Y$ look like? Note that Z takes values in (0,2). The following matlab line draws 10,000 times from the distribution of Z and plots a histogram of the data (Figure 5.1

```
hist(rand(1,10000)+rand(1,10000),50)
```

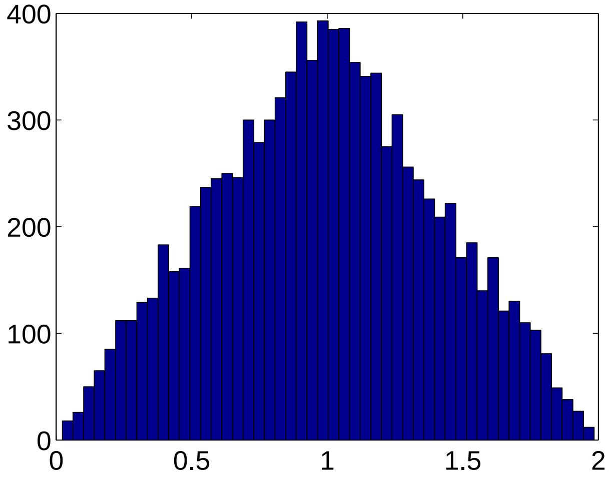


Figure 5.1: Histogram for adding two uniform random variables

This looks remarably like a triangle. Perhaps the true pdf of $Z = X + Y$ has a triangular shape? This is indeed easily proved. Namely, first observe that the pdf of Z must be symmetrical around 1. Thus to find the pdf, it suffices to find its form for $z \in [0, 1]$. Take such a z. Then, see Figure 5.2,

$$F_Z(z) = \mathbb{P}(Z \le z) = \mathbb{P}((X,Y) \in A) = \iint_A f(x,y)\, dx dy = \text{area}(A) = \frac{1}{2}\, z^2 \,.$$

where have used the fact that the joint density $f(x, y)$ is equal to 1 on the square $[0, 1] \times [0, 1]$. By differentiating the cdf F_Z we get the pdf f_Z

$$f_Z(z) = z, \quad z \in [0,1],$$

and by symmetry

$$f_Z(z) = 2 - z, \quad z \in [1,2],$$

which is indeed a triangular density. If we rescaled the histogram such that the total area under the bars would be 1, the fit with the true distribution would be very good.

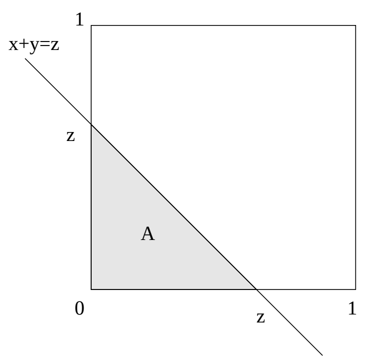


Figure 5.2: The random point (X, Y) must lie in set A

Linear Transformations

Let $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ be a (column) vector in $\mathbb{R}^n$ and A an $(n \times m)$-matrix. The mapping $\boldsymbol{x} \mapsto \boldsymbol{z}$, with

$$\boldsymbol{z} = A\boldsymbol{x}$$

is called a **linear transformation**. Now consider a *random* vector $\boldsymbol{X} = (X_1, \ldots, X_n)^T$, and let

$$\boldsymbol{Z} = A\boldsymbol{X} \; .$$

Then $\boldsymbol{Z}$ is a random vector in $\mathbb{R}^m$. Again, in principle, if we know the joint distribution of $\boldsymbol{X}$ then we can derive the joint distribution of $\boldsymbol{Z}$. Let us first see how the expectation vector and covariance matrix are transformed.

Theorem 5.1 If $\boldsymbol{X}$ has expectation vector $\boldsymbol{\mu_X}$ and covariance matrix $\Sigma_{\boldsymbol{X}}$, then the expectation vector and covariance matrix of $\boldsymbol{Z} = A\boldsymbol{X}$ are respectively given by

$$\boldsymbol{\mu_Z} = A\boldsymbol{\mu_X} \tag{5.2}$$

and

$$\Sigma_{\boldsymbol{Z}} = A\, \Sigma_{\boldsymbol{X}}\, A^T \; . \tag{5.3}$$

Proof. We have $\boldsymbol{\mu_Z} = \mathbb{E}\boldsymbol{Z} = \mathbb{E}A\boldsymbol{X} = A\mathbb{E}\boldsymbol{X} = A\boldsymbol{\mu_X}$ and

$$\begin{aligned}
\Sigma_{\boldsymbol{Z}} &= \mathbb{E}(\boldsymbol{Z} - \boldsymbol{\mu_Z})(\boldsymbol{Z} - \boldsymbol{\mu_Z})^T = \mathbb{E}A(\boldsymbol{X} - \boldsymbol{\mu_X})(A(\boldsymbol{X} - \boldsymbol{\mu_X}))^T \\
&= A\mathbb{E}(\boldsymbol{X} - \boldsymbol{\mu_X})(\boldsymbol{X} - \boldsymbol{\mu_X})^T A^T \\
&= A\, \Sigma_{\boldsymbol{X}}\, A^T
\end{aligned}$$

which completes the proof. ■

From now on assume A is an *invertible* $(n \times n) - matrix$. If $\boldsymbol{X}$ has joint density $f_{\boldsymbol{X}}$, what is the joint density $f_{\boldsymbol{Z}}$ of $\boldsymbol{Z}$?

Consider Figure 5.3. For any fixed $\boldsymbol{x}$, let $\boldsymbol{z} = A\boldsymbol{x}$. Hence, $\boldsymbol{x} = A^{-1}\boldsymbol{z}$. Consider the n-dimensional cube $C = [z_1, z_1 + h] \times \cdots \times [z_n, z_n + h]$. Let D be the *image* of C under A^{-1}, i.e., the parallelepiped of all points $\boldsymbol{x}$ such that $A\boldsymbol{x} \in C$. Then,

$$\mathbb{P}(\boldsymbol{Z} \in C) \approx h^n\, f_{\boldsymbol{Z}}(\boldsymbol{z}) \; .$$

Now recall from linear algebra that any n-dimensional rectangle with "volume" V is transformed into a n-dimensional parallelepiped with volume $V\,|A|$, where $|A| := |\det(A)|$. Thus,

$$\mathbb{P}(\boldsymbol{Z} \in C) = \mathbb{P}(\boldsymbol{X} \in D) \approx h^n |A^{-1}|\, f_{\boldsymbol{X}}(\boldsymbol{x}) = h^n |A|^{-1}\, f_{\boldsymbol{X}}(\boldsymbol{x})$$

Letting h go to 0 we conclude that

$$\boxed{f_{\boldsymbol{Z}}(\boldsymbol{z}) = \frac{f_{\boldsymbol{X}}(A^{-1}\boldsymbol{z})}{|A|}, \quad \boldsymbol{z} \in \mathbb{R}^n.} \tag{5.4}$$

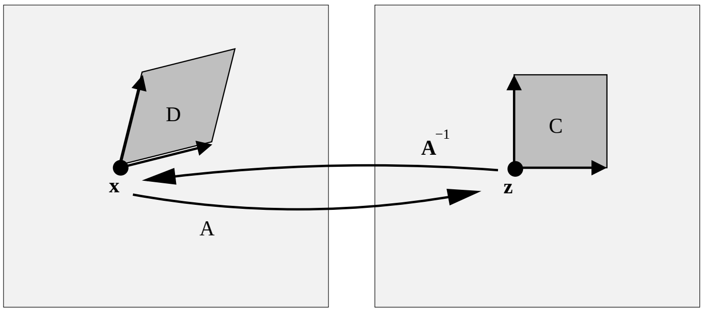


Figure 5.3: Linear transformation

General Transformations

We can apply the same technique as for the linear transformation to general transformations $\boldsymbol{x} \mapsto \boldsymbol{g}(\boldsymbol{x})$, written out:

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \mapsto \begin{pmatrix} g_1(\boldsymbol{x}) \\ g_2(\boldsymbol{x}) \\ \vdots \\ g_n(\boldsymbol{x}) \end{pmatrix} .$$

For a fixed $\boldsymbol{x}$, let $\boldsymbol{z} = \boldsymbol{g}(\boldsymbol{x})$. Suppose $\boldsymbol{g}$ is invertible, hence, $\boldsymbol{x} = \boldsymbol{g}^{-1}(\boldsymbol{z})$. Any infinitesimal n-dimensional rectangle at $\boldsymbol{x}$ with volume V is transformed into a n-dimensional parallelepiped at $\boldsymbol{z}$ with volume $V\,|J_{\boldsymbol{x}}(\boldsymbol{g})|$, where $J_{\boldsymbol{x}}(\boldsymbol{g})$ is the **matrix of Jacobi** at $\boldsymbol{x}$ of the transformation $\boldsymbol{g}$:

$$J_{\boldsymbol{x}}(\boldsymbol{g}) = \begin{bmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \vdots & \cdots & \vdots \\ \frac{\partial g_n}{\partial x_1} & \cdots & \frac{\partial g_n}{\partial x_n} \end{bmatrix} .$$

Now consider a random column vector $\boldsymbol{Z} = \boldsymbol{g}(\boldsymbol{X})$. Let C be a small cube around $\boldsymbol{z}$ with volume h^n. Let D be image of C under $\boldsymbol{g}^{-1}$. Then, as in the linear case,

$$\mathbb{P}(\boldsymbol{Z} \in C) \approx h^n\, f_{\boldsymbol{Z}}(\boldsymbol{z}) \approx h^n |J_{\boldsymbol{z}}(\boldsymbol{g}^{-1})|\, f_{\boldsymbol{X}}(\boldsymbol{x}) \; .$$

Hence, we have the transformation rule

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = f_{\boldsymbol{X}}(\boldsymbol{g}^{-1}(\boldsymbol{z}))\;|J_{\boldsymbol{z}}(\boldsymbol{g}^{-1})|, \;\; \boldsymbol{z} \in \mathbb{R}^n. \tag{5.5}$$

(Note: $|J_{\boldsymbol{z}}(\boldsymbol{g}^{-1})| = 1/|J_{\boldsymbol{x}}(\boldsymbol{g})|$)

Remark 5.1 In most coordinate transformations it is $\boldsymbol{g}^{-1}$ that is given — that is, an expression for $\boldsymbol{x}$ as a function of $\boldsymbol{z}$, — rather than $\boldsymbol{g}$.

Example 5.6 (Box-Müller) Let X and Y be two independent standard normal random variables. (X, Y) is a random point in the plane. Let (R, Θ) be the corresponding polar coordinates. The joint pdf $f_{R,\Theta}$ of R and Θ is given by

$$f_{R,\Theta}(r,\theta) = \frac{1}{2\pi}\,\mathrm{e}^{-r^2/2}\,r\ , \ \text{for } r \geq 0 \text{ and } \theta \in [0, 2\pi).$$

Namely, specifying x and y in terms of r and θ gives

$$x = r\cos\theta \quad \text{and} \quad y = r\sin\theta\ . \tag{5.6}$$

The Jacobian of this coordinate transformation is

$$\det\begin{bmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{bmatrix} = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = r\ .$$

The result now follows from the transformation rule (5.5), noting that the joint pdf of X and Y is $f_{X,Y}(x,y) = \frac{1}{2\pi}\,\mathrm{e}^{-(x^2+y^2)/2}$. It is not difficult to verify that R and Θ are independent, that $\Theta \sim \mathsf{U}[0,2\pi)$ and that $\mathbb{P}(R > r) = \mathrm{e}^{-r^2/2}$. This means that R has the same distribution as $\sqrt{V}$, with $V \sim \mathsf{Exp}(1/2)$. Namely, $\mathbb{P}(\sqrt{V} > v) = \mathbb{P}(V > v^2) = \mathrm{e}^{-v^2/2}$. Both Θ and R are easy to generate, and are transformed via (5.6) to independent standard normal random variables.

5.1 Jointly Normal Random Variables

In this section we have a closer look at normally distributed random variables and their properties. Also, we will introduce normally distributed random *vectors*.

It is helpful to view normally distributed random variables as simple transformations of standard normal random variables. For example, let $X \sim \mathsf{N}(0,1)$. Then, X has density f_X given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-\frac{x^2}{2}}.$$

Now consider the transformation

$$Z = \mu + \sigma X.$$

Then, by (5.1) Z has density

$$f_Z(z) = \frac{1}{\sqrt{2\pi\sigma^2}}\,\mathrm{e}^{-\frac{(z-\mu)^2}{2\sigma^2}}.$$

In other words, $Z \sim \mathsf{N}(\mu,\sigma^2)$. We could also write this as follows, if $Z \sim \mathsf{N}(\mu,\sigma^2)$, then $(Z-\mu)/\sigma \sim \mathsf{N}(0,1)$. This **standardisation** procedure was already mentioned in Section 2.6.3.

Let's generalise this to n dimensions. Let $X_1,\ldots,X_n$ be independent and standard normal random variables. The joint pdf of $\boldsymbol{X} = (X_1,\ldots,X_n)^T$ is given by

$$f_{\mathbf{X}}(\boldsymbol{x}) = (2\pi)^{-n/2}\mathrm{e}^{-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x}}, \quad \boldsymbol{x} \in \mathbb{R}^n. \tag{5.7}$$

Consider the transformation

$$\boldsymbol{Z} = \boldsymbol{\mu} + B\,\boldsymbol{X}, \tag{5.8}$$

for some $(m \times n)$ matrix B. Note that by Theorem 5.1 $\boldsymbol{Z}$ has expectation vector $\boldsymbol{\mu}$ and covariance matrix $\Sigma = BB^T$. Any random vector of the form (5.8) is said to have a **jointly normal** (or multi-variate normal) distribution. We write $\boldsymbol{Z} \sim \mathsf{N}(\boldsymbol{\mu}, \Sigma)$.

Suppose B is an *invertible* $(n \times n)$-matrix. Then, by (5.4) the density of $\boldsymbol{Y} = \boldsymbol{Z} - \boldsymbol{\mu}$ is given by

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{|B|\sqrt{(2\pi)^n}}\,\mathrm{e}^{-\frac{1}{2}(B^{-1}\boldsymbol{y})^T B^{-1}\boldsymbol{y}} = \frac{1}{|B|\sqrt{(2\pi)^n}}\,\mathrm{e}^{-\frac{1}{2}\boldsymbol{y}^T(B^{-1})^T B^{-1}}\boldsymbol{y}\,.$$

We have $|B| = \sqrt{|\Sigma|}$ and $(B^{-1})^T B^{-1} = (B^T)^{-1}B^{-1} = (BB^T)^{-1} = \Sigma^{-1}$, so that

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{\sqrt{(2\pi)^n\,|\Sigma|}}\,\mathrm{e}^{-\frac{1}{2}\boldsymbol{y}^T\Sigma^{-1}\boldsymbol{y}}\,.$$

Because $\boldsymbol{Z}$ is obtained from $\boldsymbol{Y}$ by simply adding a constant vector $\boldsymbol{\mu}$, we have $f_{\boldsymbol{Z}}(\boldsymbol{z}) = f_{\boldsymbol{Y}}(\boldsymbol{z} - \boldsymbol{\mu})$, and therefore

$$\boxed{f_{\boldsymbol{Z}}(\boldsymbol{z}) = \frac{1}{\sqrt{(2\pi)^n\,|\Sigma|}}\,\mathrm{e}^{-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{z}-\boldsymbol{\mu})}, \quad \boldsymbol{z} \in \mathbb{R}^n\,.} \tag{5.9}$$

Note that this formula is very similar to the 1-dimensional case.

Example 5.7 Consider the 2-dimensional case with $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$, and

$$B = \begin{pmatrix} \sigma_1 & 0 \\ \sigma_2\rho & \sigma_2\sqrt{1-\rho^2} \end{pmatrix}\,. \tag{5.10}$$

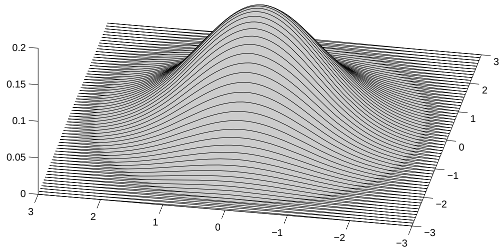
The covariance matrix is now

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\,. \tag{5.11}$$

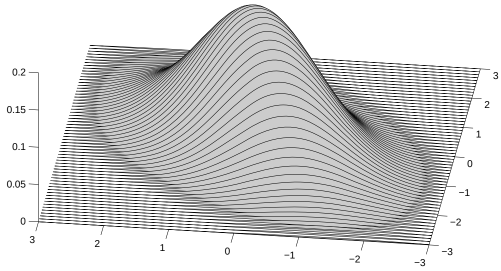
Therefore, the density is

$$f_{\mathbf{Z}}(\boldsymbol{z}) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\,\exp\Bigg\{-\frac{1}{2(1-\rho^2)} \left(\frac{(z_1-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(z_1-\mu_1)(z_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(z_2-\mu_2)^2}{\sigma_2^2}\right)\Bigg\}. \tag{5.12}$$

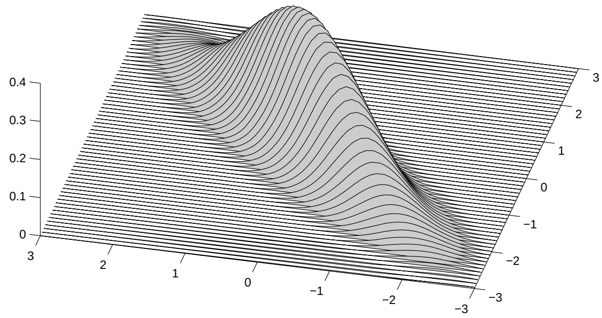
Here are some pictures of the density, for $\mu_1 = \mu_2 = 0$ and $\sigma_1 = \sigma_2 = 1$, and for various ρ.



$\rho = 0$



$\rho = 0.5$



$\rho = 0.9$

We say that $(Z_1, Z_2)^T$ has a **bivariate normal** distribution. Note that in this example $\mathbb{E}Z_i = \mu_i, i = 1, 2$. Moreover, since we have chosen B such that the covariance matrix has the form (5.11), we have $\text{Var}(Z_i) = \sigma_i^2, i = 1, 2$, and $\rho(Z_1, Z_2) = \rho$. We will see shortly that Z_1 and Z_2 both have normal distributions.

Compare the following with property 8 of Table 4.1.

Theorem 5.2 If Z_1 and Z_2 have a jointly normal distribution then

$$\text{Cov}(Z_1, Z_2) = 0 \quad \Longrightarrow \quad Z_1 \text{ and } Z_2 \text{ are independent.}$$

Proof. If $\text{Cov}(Z_1, Z_2) = 0$, then B in (5.10) is a diagonal matrix. Thus, trivially $Z_1 = \sigma_1 X_1$ and $Z_2 = \sigma_2 X_2$ are independent. ■

One of the most (if not the most) important properties of the normal distribution is that linear combinations of independent normal random variables are normally distributed. Here is a more precise formulation.

Theorem 5.3 If $X_i \sim \mathsf{N}(\mu_i, \sigma_i^2)$, independently, for $i = 1, 2, \ldots, n$, then

$$\boxed{Y = a + \sum_{i=1}^{n} b_i X_i \sim \mathsf{N}\left(a + \sum_{i=1}^{n} b_i \mu_i, \sum_{i=1}^{n} b_i^2 \sigma_i^2\right)\ .} \tag{5.13}$$

PROOF. The easiest way to prove this is by using moment generating functions. First, recall that the MGF of a $\mathsf{N}(\mu, \sigma^2)$-distributed random variable X is given by

$$M_X(s) = \mathrm{e}^{\mu s + \frac{1}{2}\sigma^2 s^2} .$$

Let M_Y be the moment generating function of Y. Since $X_1, \ldots, X_n$ are independent, we have

$$\begin{aligned} M_Y(s) &= \mathbb{E}\exp\{as + \sum_{i=1}^n b_i X_i s\} \\ &= \mathrm{e}^{as}\prod_{i=1}^n M_{X_i}(b_i s) \\ &= \mathrm{e}^{as}\prod_{i=1}^n \exp\{\mu_i(b_i s) + \frac{1}{2}\sigma_i^2 (b_i s)^2\} \\ &= \exp\{sa + s\sum_{i=1}^n b_i\mu_i + \frac{1}{2}\sum_{i=1}^n b_i^2\sigma_i^2 s^2\}, \end{aligned}$$

which is the MGF of a normal distribution of the form (5.13). ■

Remark 5.2 Note that from Theorems 4.3 and 4.6 we had already established the expectation and variance of Y in (5.13). But we have now found that the *distribution* is normal.

Example 5.8 A machine produces ball bearings with a $\mathsf{N}(1, 0.01)$ diameter (cm).The balls are placed on a sieve with a $\mathsf{N}(1.1, 0.04)$ diameter. The diameter of the balls and the sieve are assumed to be independent of each other.
Question: What is the probability that a ball will fall through?

Answer: Let $X \sim \mathsf{N}(1, 0.01)$ and $Y \sim \mathsf{N}(1.1, 0.04)$. We need to calculate $\mathbb{P}(Y > X) = \mathbb{P}(Y - X > 0)$. But, $Z := Y - X \sim \mathsf{N}(0.1, 0.05)$. Hence

$$\mathbb{P}(Z > 0) = \mathbb{P}\left(\frac{Z - 0.1}{\sqrt{0.05}} > \frac{-0.1}{\sqrt{0.05}}\right) = \Phi(0.447) \approx 0.67 ,$$

where Φ is the cdf of the $\mathsf{N}(0, 1)$-distribution.

5.2 Limit Theorems

In this section we briefly discuss two of the main results in probability: the Law of Large Numbers (LLN) and the Central Limit Theorem (CLT). Both are about sums of independent random variables.

Let $X_1, X_2, \ldots$ be independent and identically distributed random variables. For each n let

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

Suppose $\mathbb{E}X_i = \mu$ and $\text{Var}(X_i) = \sigma^2$. We assume that both μ and σ^2 are finite. By the rules for expectation and variance we know that

$$\mathbb{E}S_n = n\,\mathbb{E}X_1 = n\mu$$

and

$$\text{Var}(S_n) = n\,\text{Var}(X_1) = n\sigma^2.$$

Moreover, if the X_i have moment generating function M, then the MGF of S_n is simply given by

$$\mathbb{E}e^{s(X_1+\cdots+X_n)} = \mathbb{E}e^{sX_1}\cdots\mathbb{E}e^{sX_n} = [M(s)]^n\ .$$

The law of large numbers roughly states that S_n/n is close to μ, for large n. Here is a more precise statement.

Theorem 5.4 ((Weak) Law of Large Numbers) If $X_1, \ldots, X_n$ are independent and identically distributed with expectation μ, then for all $\epsilon > 0$

$$\boxed{\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) = 0\ .}$$

PROOF. First, for any $z > 0$, and any positive random variable Z we have

$$\begin{aligned}\mathbb{E}Z &= \int_0^z t f(t)\,dt + \int_z^\infty t f(t)\,dt \geq \int_z^\infty t f(t)\,dt \\ &\geq \int_z^\infty z f(t)\,dt = z\,\mathbb{P}(Z \geq z)\ ,\end{aligned}$$

from which follows immediately the following **Markov inequality:** if $Z \geq 0$, then for all $z > 0$,

$$\mathbb{P}(Z \geq z) \leq \frac{\mathbb{E}Z}{z}\ . \tag{5.14}$$

Now take $Z = (S_n/n - \mu)^2$ and $z = \epsilon^2$. Then,

$$\mathbb{P}(Z^2 \geq \epsilon^2) \leq \frac{\mathbb{E}(S_n/n - \mu)^2}{\epsilon^2}$$

The left-hand size of the above equation can also be written as $\mathbb{P}(|S_n/n - \mu| \geq \epsilon)$, and the right-hand side is equal to the variance of S_n/n, which is σ^2/n. Combining gives,

$$\mathbb{P}(|S_n/n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2},$$

for any $\epsilon > 0$. As $n \to \infty$ the quotient $\frac{\sigma^2}{n\epsilon^2}$ tends to zero, and therefore $\mathbb{P}(|S_n/n - \mu| \geq \epsilon)$ goes to zero as well, which had to be shown. ∎

There is also a **strong law of large numbers**, which implies the weak law, but is more difficult to prove. It states the following:

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{S_n}{n} = \mu\right) = 1,$$

as $n \to \infty$, meaning that the set of outcomes ω such that $\frac{S_n(\omega)}{n} \to \mu$, has probability one. In other words if we were to run a computer simulation, then all paths that we would simulate would converge to μ.

The Central Limit Theorem says something about the approximate *distribution* of S_n (or S_n/n). Roughly it says this:

The sum of a large number of iid random variables has approximately a **normal** *distribution*

Here is a more precise statement.

Theorem 5.5 (Central Limit Theorem) If $X_1, \ldots, X_n$ are independent and identically distributed with expectation μ and variance $\sigma^2 < \infty$, then for all $x \in \mathbb{R}$,

$$\boxed{\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le x\right) = \Phi(x)},$$

where Φ is the cdf of the standard normal distribution.

In other words, S_n has approximately a normal distribution with expectation $n\mu$ and variance $n\sigma^2$.

Proof.(Sketch) Without loss of generality assume $\mu = 0$ and $\sigma = 1$. This amounts to replacing X_n by $(X_n - \mu)/\sigma$. A Taylor-expansion of the MGF around $s = 0$ yields

$$M(s) = \mathbb{E}\,\mathrm{e}^{sX_1} = 1 + s\mathbb{E}X_1 + \frac{1}{2}s^2\,\mathbb{E}X_1^2 + o(s^2) = 1 + \frac{1}{2}\,s^2 + o(s^2),$$

where $o(\cdot)$ is a function for which $\lim_{x\downarrow 0} o(x)/x = 0$. Because the $X_1, X_2, \ldots$ are i.i.d., it follows that the MGF of $S_n/\sqrt{n}$ satisfies

$$\begin{aligned}
\mathbb{E}\exp\left(s\frac{S_n}{\sqrt{n}}\right) &= \mathbb{E}\exp\left(\frac{s}{\sqrt{n}}(X_1 + \cdots X_n)\right) = \prod_{i=1}^{n}\mathbb{E}\exp\left(\frac{s}{\sqrt{n}}X_i\right) \\
&= M^n\left(\frac{s}{\sqrt{n}}\right) = \left[1 + \frac{s^2}{2n} + o(\frac{s^2}{n})\right]^n.
\end{aligned}$$

For $n \to \infty$ this converges to $\mathrm{e}^{s^2/2}$, which is the MGF of the standard normal distribution. Thus, it is plausible that the cdf of $S_n/\sqrt{n}$ converges to Φ. To make this argument rigorous, one needs to show that convergence of the moment generating function implies convergence of the cdf. Moreover, since for some distributions the MGF does not exist in a neighbourhood of 0, one needs to replace the MGF in the argument above with a more flexible transform, namely the Fourier transform, also called characteristic function: $r \mapsto \mathbb{E}\mathrm{e}^{irX}$, $r \in \mathbb{R}$. ■

To see the CLT in action consider Figure 5.4. The first picture shows the pdf's of $S_1, \ldots, S_4$ for the case where the X_i have a $\mathsf{U}[0,1]$ distribution. The second

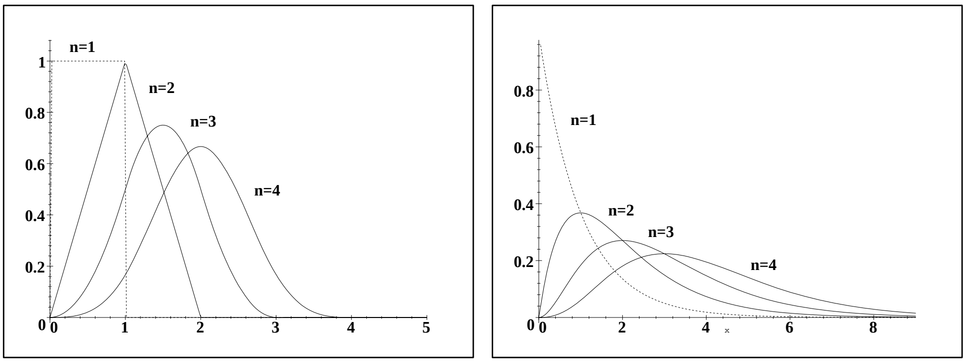


Figure 5.4: Illustration of the CLT for the uniform and exponential distribution

show the same, but now for an $\mathsf{Exp}(1)$ distribution. We clearly see convergence to a bell shaped curve.

The CLT is not restricted to continuous distributions. For example, Figure 5.5 shows the cdf of S_{30} in the case where the X_i have a Bernoulli distribution with success probability 1/2. Note that $S_{30} \sim \mathsf{Bin}(30, 1/2)$, see Example 4.3.

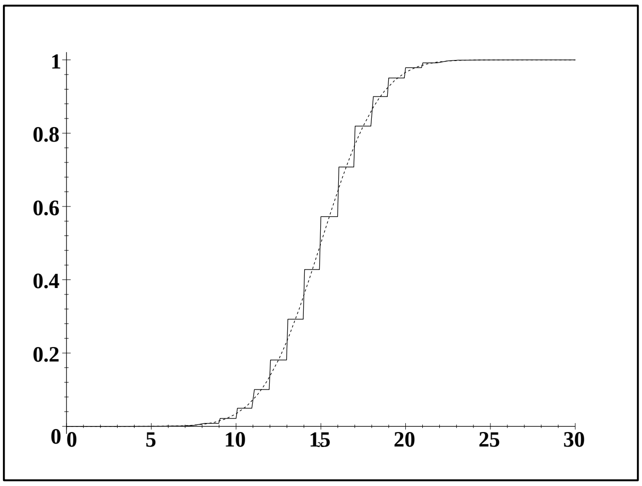


Figure 5.5: The cdf of a $\mathsf{Bin}(20, 1/2)$-distribution and its normal approximation.

In general we have:

Theorem 5.6 Let $X \sim \mathsf{Bin}(n, p)$. For large n we have

$$\mathbb{P}(X \leq k) \approx \mathbb{P}(Y \leq k),$$

where $Y \sim \mathsf{N}(np, np(1-p))$. As a rule of thumb, the approximation is accurate if both np and $n(1-p)$ are larger than 5.

Appendix A

Exercises and Solutions

A.1 Problem Set 1

1. We toss a fair coin three times.
   (a) Find the sample space, if we observe the exact sequences of Heads (= 1) and Tails (= 0).
   (b) Find the sample space, if we observe only the total number of Heads.
2. Assign a probability to all elementary events in the sample spaces 1.(a) and 1.(b).
3. We randomly select 3 balls from an urn with 365 balls, numbered 1, ...,365, noting the order.
   (a) How many possible outcomes of the experiment are there, if we put each ball back into the urn before we draw the next?
   (b) Answer the same question as above, but now if we *don't* put the balls back.
   (c) Calculate the probability that in case (a) we draw 3 times the same ball.
4. Let $\mathbb{P}(A) = 0.9$ and $\mathbb{P}(B) = 0.8$. Show that $\mathbb{P}(A \cap B) \geq 0.7$.
5. What is the probability that none of 54 people in a room share the same birthday?
6. Consider the experiment of throwing 2 fair dice.
   (a) Find the probability that both dice show the same face.
   (b) Find the same probability, using the extra information that the sum of the dice is not greater than 4.
7. We draw 3 cards from a full deck of cards, noting the order. Number the cards from 1 to 52.

(a) Give the sample space. Is each elementary event equally likely?

(b) What is the probability that we draw 3 Aces?

(c) What is the probability that we draw 1 Ace, 1 King and 1 Queen?

(d) What is the probability that we draw no pictures (no A,K,Q,J)?

8. We draw at random a number in the interval [0,1] such that each number is "equally likely". Think of the *random generator* on you calculator.

   (a) Determine the probability that we draw a number less than 1/2.

   (b) What is the probability that we draw a number between 1/3 and 3/4?

   (c) Suppose we do the experiment two times (independently), giving us two numbers in [0,1]. What is the probability that the sum of these numbers is greater than 1/2? Explain your reasoning.

9. Select at random 3 people from a large population. What is the probability that they all have the same birthday?

10. We draw 4 cards (at the same time) from a deck of 52, not noting the order. Calculate the probability of drawing one King, Queen, Jack and Ace.

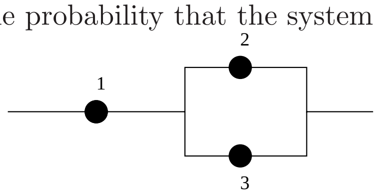
11. In a group of 20 people there are three brothers. The group is separated at random into two groups of 10. What is the probability that the brothers are in the same group?

12. How many binary vectors are there of length 20 with exactly 5 ones?

13. We draw at random 5 balls from an urn with 20 balls (numbered 1,. . . ,20), without replacement or order. How many different possible combinations are there?

14. In a binary transmission channel, a 1 is transmitted with probability 2/3 and a 0 with probability 1/3. The conditional probability of receiving a 1 when a 1 was sent is 0.95, the conditional probability of receiving a 0 when a 0 was sent is 0.90. Given that a 1 is received, what is the probability that a 1 was transmitted?

15. Consider the following system. Each component has a probability 0.1 of failing. What is the probability that the system works?



16. Two fair dice are thrown and the smallest of the face values, Z say, is noted.

(a) Give the pmf of Z in table form:

z	$*$	$*$	$*$	$\cdots$
$\mathbb{P}(Z=z)$	$*$	$*$	$*$	$\cdots$

(b) Calculate the expectation of Z.

17. In a large population 40% votes for A and 60% for B. Suppose we select at random 10 people. What is the probability that in this group exactly 4 people will vote for A?

18. We select "uniformly" a point in the unit square: $\{(x,y) : 0 \le x \le 1,\ 0 \le y \le 1\}$. Let Z be the largest of the coordinates. Give the cdf and pdf of Z and draw their graphs.

19. A continuous random variable X has cdf F given by,

$$F(x) = \begin{cases} 0, & x < 0 \\ x^3, & x \in [0,1] \\ 1 & x > 1\ . \end{cases}$$

(a) Determine the pdf of X.
(b) Calculate $\mathbb{P}(1/2 < X < 3/4)$.
(c) Calculate $\mathbb{E}[X]$.

A.2 Answer Set 1

1. (a) $\Omega = \{(0,0,0), \ldots, (1,1,1)\}$.
   (b) $\Omega = \{0,1,2,3\}$.

2. (a) $\mathbb{P}(\{(x,y,z)\}) = 1/8$ for all $x,y,z \in \{0,1\}$.
   (b) $\mathbb{P}(\{0\}) = 1/8, \mathbb{P}(\{1\}) = 3/8, \mathbb{P}(\{2\}) = 3/8, \mathbb{P}(\{3\}) = 1/8$.

3. (a) $|\Omega| = 365^3$.
   (b) $|\Omega| = 365 \times 364 \times 363$.
   (c) $365/365^3 = 1/365^2$.

4. $\mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cup B)$. Since $\mathbb{P}(A \cup B) \le 1$, we have $\mathbb{P}(A \cap B) \ge \mathbb{P}(A) + \mathbb{P}(B) - 1 = 1.7 - 1 = 0.7$.

5. $\frac{365 \times 364 \times \cdots \times (365-54+1)}{365^{54}} \approx 0.016$.

6. (a) 1/6.
   (b) Let A be the event that the dice show the same face, and B the event that the sum is not greater than 4. Then $B = \{(1,1),(1,2),(1,3)(2,1),(2,2),(3,1)\}$, and $A \cap B = \{(1,1),(2,2)\}$. Hence, $\mathbb{P}(A \mid B) = 2/6 = 1/3$.

7. (a) $\Omega = \{(1,2,3), \ldots, (52,51,50)\}$. Each elementary event is equally likely.

(b) $\frac{4\times3\times2}{52\times51\times50}$.

(c) $\frac{12\times8\times4}{52\times51\times50}$.

(d) $\frac{40\times39\times38}{52\times51\times50}$.

8. (a) 1/2.

   (b) 5/12.

   (c) 7/8.

9. $1/365^2$, see question 3 (c).

10. $|\Omega| = \binom{52}{4}$ (all equally likely outcomes. Note that the outcomes are repsesented as (unordered sets), e.g., $\{1,2,3,4\}$. Let A be the event of drawing one K, Q, J and Ace each. Then, $|A| = \binom{4}{1}\times\binom{4}{1}\times\binom{4}{1}\times\binom{4}{1} = 4^4$. Thus, $\mathbb{P}(A) = 4^4/\binom{52}{4}$.

11. Suppose we choose 10 people to go in group 1 (the rest go in group 2). The total number of ways this can be done is $\binom{20}{10}$. Let A be the event that the brothers belong to the same group. The number of ways in which they can be chosen into group 1 is: $\binom{17}{7}$. The number of ways they can be chosen into group 2 is the same, $\binom{17}{10} = \binom{17}{7}$. Thus, $\mathbb{P}(A) = 2\binom{17}{7}/\binom{20}{10}$.

12. $\binom{20}{5}$, because we have to choose the 5 positions for the 1s, out of 20 positions.

13. $\binom{20}{5}$

14. Let B be the event that a 1 was sent, and A the event that a 1 is received. Then, $\mathbb{P}(A\,|\,B) = 0.95$, and $\mathbb{P}(A^c\,|\,B^c) = 0.90$. Thus, $\mathbb{P}(A^c\,|\,B) = 0.05$ and $\mathbb{P}(A\,|\,B^c) = 0.10$. Moreover, $\mathbb{P}(B) = 2/3$ and $\mathbb{P}(B^c) = 1/3$. By Bayes:

$$\mathbb{P}(B\,|\,A) = \frac{\mathbb{P}(A\,|\,B)\mathbb{P}(B)}{\mathbb{P}(A\,|\,B)\mathbb{P}(B) + \mathbb{P}(A\,|\,B^c)\mathbb{P}(B^c)} = \frac{0.95\times\frac{2}{3}}{0.95\times\frac{2}{3} + 0.10\times\frac{1}{3}}$$

15. Let A_i be the event that component i works, $i = 1,2,3$, and let A be the event that the system works. We have $A = A_1 \cap (A_2 \cup A_3)$. Thus, by the independence of the $A_i's$:

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A_1)\times\mathbb{P}(A_2\cup A_3)\\ &= \mathbb{P}(A_1)\times[\mathbb{P}(A_2)+\mathbb{P}(A_3)-\mathbb{P}(A_2\cap A_3)]\\ &= \mathbb{P}(A_1)\times[\mathbb{P}(A_2)+\mathbb{P}(A_3)-\mathbb{P}(A_2)\mathbb{P}(A_3)]\\ &= 0.9\times[0.9+0.9-0.81]\end{aligned}$$

16. (a)

z	1	2	3	4	5	6
$\mathbb{P}(Z=z)$	11/36	9/36	7/36	5/36	3/36	1/36

   (b) $\mathbb{E}[Z] = 1\times11/36+2\times9/36+3\times7/36+4\times5/36+5\times3/36+6\times1/36$.

17. Let X be the number that vote for A. Then $X \sim \mathsf{Bin}(10, 0.4)$. Hence, $\mathbb{P}(X=4) = \binom{10}{4}(0.4)^4(0.6)^6$.

18. The region where the largest coordinate is less than z is a square with area z^2 (make a picture). Divide this area by the area of the unit square (1), to obtain $\mathbb{P}(Z \leq z) = z^2$, for all $z \in [0, 1]$. Thus,

$$F(z) = \mathbb{P}(Z \leq z) = \begin{cases} 0 & z < 0 \\ z^2 & 0 \leq z \leq 1 \\ 1 & z > 1\,. \end{cases}$$

Differentiate to get the pdf:

$$f(z) = \begin{cases} 2\,z & 0 \leq z \leq 1 \\ 0 & \text{otherwise}\,. \end{cases}$$

19. (a)

$$f(x) = \begin{cases} 3\,x^2 & 0 \leq x \leq 1 \\ 0 & \text{otherwise}\,. \end{cases}$$

(b) $\int_{1/2}^{3/4} f(x)\,dx = F(3/4) - F(1/2) = (3/4)^3 - (1/2)^3$.

(c) $\mathbb{E}[X] = \int_0^1 x\,3x^2\,dx = 3\int_0^1 x^3\,dx = 3/4$.

A.3 Problem Set 2

1. In a binary communication channel, 0s and 1s are transmitted with equal probability. The probability that a 0 is correctly received (as a 0) is 0.95. The probability that a 1 is correctly received (as a 1) is 0.99. Suppose we receive a 0, what is the probability that, in fact, a 1 was sent?

2. Throw two fair dice one after the other.
   (a) What is the probability that the second die is 3, given that the sum of the dice is 6?
   (b) What is the probability that the first die is 3 and the second not 3?

3. We flip a fair coin 20 times.
   (a) What is the probability of exactly 10 Heads?
   (b) What is the probability of 15 or more Heads?

4. We toss two fair dice until their sum is 12.
   (a) What is probabillity that we have to wait exactly 10 tosses?
   (b) What is the probability that we do not have to wait more than 100 tosses?

5. We independently throw 10 balls into one of 3 boxes, numbered 1,2 and 3, with probabilities 1/4, 1/2 and 1/4, respectively.

(a) What is the probability that box 1 has 2, box 5 has 10 and box 3 has 3 balls?

(b) What is the probability that box 1 remains empty?

6. Consider again the experiment where we throw two fair dice one after the other. Let the random variable Z denote the sum of the dice.

   (a) How is Z defined as a function?

   (b) What is the pmf of Z?

7. Consider the random experiment of question 4. Let X be the number of tosses required until the sum of the dice is 12. Give the pmf of X.

8. We draw at random and uniformly a point from the interior of a circle with radius 4. Let R be the distance of this point to the centre of the circle.

   (a) What is the probability that $R > 2$?

   (b) What is the pdf of R?

9. Let $X \sim \mathsf{Bin}(4, 1/2)$. What is the pmf of $Y = X^2$?

10. Let $X \sim \mathsf{U}[0, 1]$. What is the pdf of $Y = X^2$?

11. Let $X \sim \mathsf{N}(0, 1)$, and $Y = 1 + 2X$. What is the pdf of Y?

12. Let $X \sim \mathsf{N}(0, 1)$. Find $\mathbb{P}(X \leq 1.4)$ from the table of the $\mathsf{N}(0, 1)$ distribution. Also find $\mathbb{P}(X > -1.3)$.

13. Let $Y \sim \mathsf{N}(1, 4)$. Find $\mathbb{P}(Y \leq 3)$, and $\mathbb{P}(-1 \leq Y \leq 2)$.

14. If $X \sim \mathsf{Exp}(1/2)$ what is the pdf of $Y = 1 + 2X$? Sketch the graph.

15. We draw at random 5 numbers from 1,... 100, *with replacement* (for example, drawing number 9 twice is possible). What is the probability that exactly 3 numbers are even?

16. We draw at random 5 numbers from 1,... 100, *without replacement*. What is the probability that exactly 3 numbers are even?

17. A radioactive source of material emits a radioactive particle with probability 1/100 in each second. Let X be the number of particles emitted in one hour. X has approximately a Poisson distribution with what parameter? Draw (with the use of a computer) or sketch the pmf of X.

18. An electrical component has a lifetime X that is exponentially distributed with parameter $\lambda = 1/10$ per year. What is the probability the component is still alive after 5 years?

19. A random variable X takes the values $0, 2, 5$ with probabilities $1/2, 1/3, 1/6$, respectively. What is the expectation of X?

20. A random variable X has expectation 2 and variance 1. Calculate $\mathbb{E}[X^2]$.

21. We draw at random a 10 balls from an urn with 25 red and 75 white balls. What is the expected number of red balls amongst the 10 balls drawn? Does it matter if we draw the balls with or without replacement?

22. Let $X \sim \mathsf{U}[0,1]$. Calculate $\mathrm{Var}(X)$.

23. If $X \sim \mathsf{U}[0,1]$, what is the expectation of $Y = 10 + 2X$?

24. We repeatedly throw two fair dice until two sixes are thrown. What is the expected number of throws required?

25. Suppose we divide the population of Brisbane (say 1,000,000 people) randomly in groups of 3.

    (a) How many groups would you expect there to be in which all persons have the same birthday?

    (b) What is the probability that there is at least one group in which all persons have the same birthday?

26. An electrical component has an exponential lifetime with expectation 2 years.

    (a) What is the probability that the component is still functioning after 2 years?

    (b) What is the probability that the component is still functioning after 10 years, given it is still functioning after 7 years?

27. Let $X \sim \mathsf{N}(0,1)$. Prove that $\mathrm{Var}(X) = 1$. Use this to show that $\mathrm{Var}(Y) = \sigma^2$, for $Y \sim \mathsf{N}(\mu, \sigma^2)$.

28. let $X \sim \mathsf{Exp}(1)$. Use the Moment Generating Function to show that $\mathbb{E}[X^n] = n!$.

29. Explain how to generate random numbers from the $\mathsf{Exp}(10)$ distribution. Sketch a graph of the scatterplot of 10 such numbers.

30. Explain how to generate random numbers from the $\mathsf{U}[10,15]$ distribution. Sketch a graph of the scatterplot of 10 such numbers.

31. Suppose we can generate random numbers from the $\mathsf{N}(0,1)$ distribution, e.g., via Matlabs `randn` function. How can we generate from the $\mathsf{N}(3,9)$ distribution?

A.4 Answer Set 2

1. $\frac{0.01 \times \frac{1}{2}}{0.01 \times \frac{1}{2} + 0.95 \times \frac{1}{2}}$ (Bayes' rule).

2. (a) $\frac{1}{5}$ (conditional probability, the possible outcomes are $(1,5),(2,4),\ldots,(5,1)$. In only one of these the second die is 3).
   (b) $\frac{1}{6} \times \frac{5}{6}$ (independent events).

3. (a) $\binom{20}{10}/2^{20} \approx 0.176197$.
   (b) $\sum_{k=15}^{20} \binom{k}{20}/2^{20} \approx 0.0206947$.

4. (a) $\left(\frac{35}{36}\right)^9 \frac{1}{36} \approx 0.021557$. (geometric formula)
   (b) $1 - \left(\frac{35}{36}\right)^{100} \approx 0.94022$.

5. (a) $\frac{10!}{2!\,5!\,3!}\left(\frac{1}{4}\right)^2\left(\frac{1}{2}\right)^5\left(\frac{1}{4}\right)^3 = \frac{315}{4096} \approx 0.076904$. (multinomial)
   (b) $(3/4)^{10} \approx 0.0563135$.

6. (a) $Z((x,y)) = x + y$ for all $x, y \in \{1,2,\ldots,6\}$.
   (b) Range: $S_X = \{2,3,\ldots,12\}$. Pmf: $\mathbb{P}(X = x) = \frac{6-|x-7|}{36}, \quad x \in S_X$.

7. $\mathbb{P}(X = x) = \left(\frac{35}{36}\right)^{x-1} \frac{1}{36}, \quad x \in \{1,2,\ldots\}$. (geometric formula)

8. (a) $\frac{\pi 16 - \pi 4}{\pi 16} = \frac{3}{4}$.
   (b) First, the cdf of R is $F_R(r) = \pi\, r^2/(\pi 16) = r^2/16$, $r \in (0,4)$. By differentiating we obtain the pdf $f(r) = \frac{r}{8}, \quad 0 < r < 4$.

9. $S_Y = \{0,1,4,9,16\}$. $\mathbb{P}(Y = k^2) = \mathbb{P}(X = k) = \frac{\binom{4}{k}}{16}, \quad k = 0,1,\ldots,4$.

10. $S_Y = [0,1]$. For $0 \le y \le 1$ we have $\mathbb{P}(Y \le y) = \mathbb{P}(X^2 \le y) = \mathbb{P}(X \le \sqrt{y}) = \sqrt{y}$. Thus, $f_Y(y) = \frac{1}{2\sqrt{y}}, \quad 0 < y \le 1$.

11. $Y \sim \mathsf{N}(1,4)$. (affine transformation of a normal r.v. gives again a normal r.v.)

12. $\mathbb{P}(X \le 1.4) = \Phi(1.4) \approx 0.9192$. $\mathbb{P}(X > -1.3) = \mathbb{P}(X < 1.3)$ (by symmetry of the pdf — make a picture). $\mathbb{P}(X < 1.3) = \mathbb{P}(X \le 1.3) = \Phi(1.3) \approx 0.9032$.

13. $\mathbb{P}(Y \le 3) = \mathbb{P}(1 + 2X \le 3)$, with X standard normal. $\mathbb{P}(1 + 2X \le 3) = \mathbb{P}(X \le 1) \approx 0.8413$. $\mathbb{P}(-1 \le Y \le 2) = \mathbb{P}(-1 \le X \le 1/2) = \mathbb{P}(X \le 1/2) - \mathbb{P}(X \le -1) = \Phi(1/2) - (1 - \Phi(1)) \approx 0.6915 - (1 - 0.8413) = 0.5328$.

14. $f_Y(y) = f_X((y-1)/2) = \mathrm{e}^{-(y-1)/4}, \quad y \ge 1$.

15. Let X be the total number of "even" numbers. Then, $X \sim \mathsf{Bin}(5, 1/2)$. And $\mathbb{P}(X = 3) = \binom{5}{3}/32 = 10/32 = 0.3125$

16. Let X be the total number of "even" numbers. Then $X \sim \mathsf{Hyp}(5, 50, 100)$. Hence $\mathbb{P}(X = 3) = \frac{\binom{50}{3}\binom{50}{2}}{\binom{100}{5}} = 6125/19206 \approx 0.318911$.

17. $\lambda = 3600/100 = 36$.

18. $\mathrm{e}^{-5/10} \approx 0.3679$.

19. 3/2.

20. 5

21. It does matter for the distribution of the number of red balls, which is $\mathsf{Bin}(10, 1/4)$ if we replace and $\mathsf{Hyp}(10, 25, 100)$ if we don't replace. However the expectation is the same for both cases: 2.5.

22. $\mathrm{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2$. By symmetry $\mathbb{E}X = 1/2$. And $\mathbb{E}X^2 = \int_0^1 x^2\, dx = 1/3$. So $\mathrm{Var}(X) = 1/12$.

23. $10 + 2\mathbb{E}X = 11$.

24. 36 (expectation for the $\mathsf{G}(1/36)$ distribution).

25. Let N be the number of groups in which each person has the same birthday. Then $N \sim \mathsf{Bin}(333333, 1/365^2)$. Hence (a) $\mathbb{E}N \approx 2.5$, and (b) $\mathbb{P}(N > 0) = 1 - \mathbb{P}(N = 0) = 1 - (1 - 1/365^2)^{333333} \approx 0.92$. [Alternatively N has approximately a $\mathsf{Poi}(2.5)$ distribution, so $\mathbb{P}(N = 0) \approx \mathrm{e}^{-2.5}$, which gives the same answer 0.92.]

26. First note $\lambda = 1/2$. Let X be the lifetime.
    (a) $\mathbb{P}(X > 2) = \mathrm{e}^{2/2} = \mathrm{e}^{-1} \approx 0.368$.
    (b) $\mathbb{P}(X > 10 \,|\, X > 7) = \mathbb{P}(X > 3) = \mathrm{e}^{-1.5} \approx 0.223$ (memoryless property).

27. $\mathrm{Var}(X) = \mathbb{E}X^2 - \mathbb{E}X = \mathbb{E}X^2 = \int_{-\infty}^{\infty} \frac{x^2 \mathrm{e}^{-x^2/2}}{\sqrt{2\pi}}\, dx = \int_{-\infty}^{\infty} x\, \frac{x\mathrm{e}^{-x^2/2}}{\sqrt{2\pi}}\, dx = \int_{-\infty}^{\infty} \frac{\mathrm{e}^{-x^2/2}}{\sqrt{2\pi}}\, dx = 1$. Note that we have used partial integration in the fifth equality. For general $Y \sim \mathsf{N}(\mu, \sigma^2)$, write $Y = \mu + \sigma X$, so that $\mathrm{Var}(Y) = \mathrm{Var}(\mu + \sigma X) = \sigma^2 \mathrm{Var}(X) = \sigma^2$.

28. $M(s) = \lambda/(\lambda - s)$. Differentiate: $M'(s) = \lambda/(\lambda - s)^2$, $M''(s) = 2\lambda/(\lambda - s)^3$, $\ldots$, $M^{(n)}(s) = n!\, \lambda/(\lambda - s)^{n+1}$. Now apply the moment formula.

29. Draw $U \sim \mathsf{U}(0, 1)$. Return $X = -\frac{1}{10} \log U$.

30. Draw $U \sim \mathsf{U}(0, 1)$. Return $X = 10 + 5U$.

31. Draw $X \sim \mathsf{N}(0, 1)$. Return $Y = 3 + 3X$.

A.5 Problem Set 3

1. Consider the random experiment where we draw independently n numbers from the interval [0,1]; each number in [0,1] being equally likely to be drawn. Let the independent and $\mathsf{U}[0, 1]$-distributed random variables $X_1, \ldots, X_n$ represent the numbers that are to be drawn.

(a) Let M be the smallest of the n numbers, and $\bar{X}$ the average of the n numbers. Express M and $\bar{X}$ in terms of $X_1, \ldots, X_n$.

(b) Determine the pdf of M.

(c) Give the expectation and variance of $\bar{X}$.

2. The joint pmf of X and Y is given by the table

x \ y	1	3	6	8
2	0	0.1	0.1	0
5	0.2	0	0	0
6	0	0.2	0.1	0.3

(a) Determine the (marginal) pmf of X and of Y.

(b) Are X and Y independent?

(c) Calculate $\mathbb{E}[X^2 Y]$.

3. Explain how, in principle we can calculate

$$\mathbb{P}\left(\frac{X_1^2 + \sin(X_2)}{X_1^2 X_2} > 1\right) ,$$

if we know the joint pdf of X_1 and X_2.

4. Suppose $X_1, X_2, \ldots, X_n$ are independent random variables, with cdfs $F_1, F_2, \ldots,$ F_n, respectively. Express the cdf of $M = \max(X_1, \ldots, X_n)$ in terms of the $\{F_i\}$.

5. Let $X_1, \ldots, X_6$ be the weights of 6 people, selected from a large population. Suppose the weights have a normal distribution with a mean of 75 kg and a standard deviation of 10 kg. What do $Y_1 = 6X_1$ and $Y_2 = X_1 + \cdots + X_6$ represent, physically? Explain why Y_1 and Y_2 have different distributions. Which one has the smallest variance?

6. Let $X \sim \mathsf{Bin}(100, 1/4)$. Approximate, using the CLT, the probability $\mathbb{P}(20 \leq X \leq 30)$.

7. Let X and Y be independent and $\mathsf{Exp}(1)$ distributed. Consider the coordinate transformation

$$x = uv , \qquad y = u - uv \qquad (\text{thus } u = x + y \text{ and } v = x/(x + y)) .$$

Let $U = X + Y$ and $V = X/(X + Y)$.

(a) Determine the Jacobian of the above coordinate transformation.

(b) Determine the joint pdf of U and V.

(c) Show that U and V are independent.

8. A random vector (X, Y) has joint pdf f, given by

$$f(x,y) = 2\,\mathrm{e}^{-x-2y}, \quad x > 0, y > 0 \,.$$

   (a) Calculate $\mathbb{E}[XY]$.
   (b) Calculate the covariance of $X + Y$ and $X - Y$.

9. Consider the random experiment where we make repeated measurements of the voltage across a resistor in an electric circuit. Let $X_1, \ldots, X_n$ be the voltage readings. We assume that the $X_1, \ldots, X_n$ are independent and normally distributed with the same (unknown) mean μ and (unknown) variance σ^2. Suppose $x_1, \ldots, x_n$ are the outcomes of the random variables $X_1, \ldots, X_n$. Let $\bar{X} = (X_1 + \cdots + X_n)/n$.

   (a) How would you estimate the unknown parameter μ if you had the data $x_1, \ldots, x_n$?
   (b) Show that $\mathbb{E}\bar{X} = \mu$.
   (c) Show that $\mathrm{Var}\bar{X}$ goes to 0 as $n \to \infty$.
   (d) What is the distribution of $\bar{X}$?
   (e) Discuss the implications of (b) and (c) for your estimation of the unknown μ.

10. Let $X \sim \mathsf{Bin}(100, 1/2)$. Approximate, using the CLT, the probability $\mathbb{P}(X \geq 60)$.

11. Let X have a uniform distribution on the interval $[1, 3]$. Define $Y = X^2 - 4$. Derive the probability density function (pdf) of Y. Make sure you also specify where this pdf is zero.

12. Let $X \sim \mathsf{U}(1, 3)$. Define $Y = 4X + 5$. What is the distribution of Y?

13. Let $Y \sim \mathsf{N}(2, 5)$.

   (a) Sketch the pdf and cdf of Y.
   (b) Calculate $\mathbb{P}(Y \leq 5)$. [Use the table for the cdf of the $\mathsf{N}(0, 1)$-distribution.]
   (c) Let $Z = 3Y - 4$. Determine $\mathbb{P}(1 \leq Z \leq 5)$. [Use the table for the cdf of the $\mathsf{N}(0, 1)$-distribution.]

14. Some revision questions. Please make sure you can comfortably answer the questions below *by heart*.

   (a) Give the formula for the pmf of the following distributions:
      i. $\mathsf{Bin}(n, p)$,
      ii. $\mathsf{G}(p)$,
      iii. $\mathsf{Poi}(\lambda)$.
   (b) Give the formula for the pdf of the following distributions:

i. $\mathsf{U}(a,b)$,
ii. $\mathsf{Exp}(\lambda)$,
iii. $\mathsf{N}(0,1)$.

(c) Give examples of random experiments where the distributions in (a) and (b) above occur.

15. Random variables $X_1, X_2, \ldots$ are independent and have a standard normal distribution. Let $Y_1 = X_1$, $Y_2 = X_1 + X_2$, and, generally, $Y_n = X_1 + \cdots + X_n$, $n = 1, 2, \ldots$.

    (a) Sketch a typical outcome of $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ for large n.
    (b) Determine $\mathbb{E}Y_n$ and $\mathrm{Var}(Y_n)$.
    (c) Determine $\mathrm{Cov}(Y_m, Y_n)$, $m \leq n$.

16. A lift can carry a maximum of 650 kg. Suppose that the weight of a person is normally distributed with expectation 75 kg and standard deviation 10 kg. Let Z_n be the total weight of n randomly selected persons.

    (a) Determine the probability that $Z_8 \geq 650$.
    (b) Determine n such that $\mathbb{P}(Z_n \geq 650) \leq 0.01$.

17. The thickness of a printed circuit board is required to lie between the specification limits of 0.150 - 0.004 and 0.150 + 0.004 cm. A machine produces circuit boards with a thickness that is normally distributed with mean 0.151 cm and standard deviation 0.003 cm.

    (a) What is the probability that the thickness X of a circuit board which is produced by this machine falls within the specification limits?
    (b) Now consider the mean thickness $\bar{X} = (X_1 + \cdots + X_{25})/25$ for a batch of 25 circuit boards. What is the probability that this batch mean will fall within the specification limits? Assume that $X_1, \ldots, X_{25}$ are independent random variables with the same distribution as X above.

18. We draw n numbers independently and uniformly from the interval [0,1] and note their sum S_n.

    (a) Draw the graph of the pdf of S_2.
    (b) What is approximately the distribution of S_{20}?
    (c) Calculate the probability that the *average* of the 20 numbers is greater than 0.6.

19. Consider the following game: You flip 10 fair coins, all at once, and count how many Heads you have. I'll pay you out the squared number of Heads, in dollars. However, you will need to pay me some money in advance. How much would you prepare to give me if you could play this game as many times as you'd like?

A.6 Answer Set 3

1. (a) $M = \min(X_1, \ldots, X_n)$ and $\bar{X} = (X_1 + \cdots + X_n)/n$.
   (b) $\mathbb{P}(M > x) = \mathbb{P}(X_1 > x, \cdots, X_n > x) =$ (by indep. of $X_1, \ldots, X_n$) $\mathbb{P}(X_1 > x) \cdots \mathbb{P}(X_n > x) = (1-x)^n$, for all $0 \le x \le 1$. Hence the cdf of M is given by $F_M(x) = 1 - (1-x)^n$, $0 \le x \le 1$. Consequently, the pdf of M is given by $f_M(x) = n(1-x)^{n-1}$, $0 \le x \le 1$.
   (c) $\mathbb{E}\bar{X} = \mathbb{E}\left(\frac{X_1+\cdots+X_n}{n}\right) = \frac{1}{n}(\mathbb{E}X_1 + \cdots + \mathbb{E}X_n) = \frac{1}{n} n \mathbb{E}X_1 = \frac{1}{2}$.
   $\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{X_1+\cdots+X_n}{n}\right) \underbrace{=}_{\text{by indep.}} \frac{1}{n^2}(\mathrm{Var}X_1 + \cdots + \mathrm{Var}X_n) = \frac{1}{n^2} n \mathrm{Var}(X_1) = \frac{1}{12n}$.
2. (a) $\mathbb{P}(X = 2) = 0.2$, $\mathbb{P}(X = 5) = 0.2$, $\mathbb{P}(X = 6) = 0.6$.
   $\mathbb{P}(Y = 1) = 0.2$, $\mathbb{P}(Y = 3) = 0.3$, $\mathbb{P}(Y = 6) = 0.2$, $\mathbb{P}(Y = 8) = 0.3$.
   (b) No. For example $\mathbb{P}(X = 2, Y = 1) = 0 \neq \mathbb{P}(X = 2) \cdot \mathbb{P}(Y = 1)$
   (c) $\mathbb{E}[X^2 Y] = 0.1(2^2 3) + 0.1(2^2 6) + 0.2(5^2 1) + 0.2(6^2 3) + 0.1(6^2 6) + 0.3(6^2 8) = 138.2$.
3. By integrating the joint pdf over the region $A = \{(x_1, x_2) : \frac{x_1^2 + \sin(x_2)}{x_1^2 x_2} > 1\}$.
4. $\mathbb{P}(M \le m) = \mathbb{P}(X_1 \le x_1, \ldots, X_n \le x_n) = \mathbb{P}(X_1 \le x_1) \cdots \mathbb{P}(X_n \le x_n) = F_1(m) \cdots F_n(m)$.
5. Y_2 represents the sum of the weights of 6 different people, whereas Y_1 represents 6 times the weight of 1 person. The expectation of both Y_1 and Y_2 is 6×75. However, the variance of Y_1 is $6^2 \mathrm{Var}(X_1) = 3600$, whereas the variance of Y_2 is 6 times smaller: $6\mathrm{Var}(X_1) = 600$. Thus, $Y_1 \sim \mathsf{N}(75, 3600)$ and $Y_2 \sim \mathsf{N}(75, 600)$.
6. $\mathbb{P}(20 \le X \le 30) \approx \mathbb{P}(20 \le Y \le 30)$, with $Y \sim \mathsf{N}(100 \times \frac{1}{4}, 100 \times \frac{1}{4} \times \frac{3}{4}) = \mathsf{N}(25, 75/4)$. We have

   $$\mathbb{P}(20 \le Y \le 30) = \mathbb{P}\Big(\frac{20-25}{\sqrt{75/4}} \le Z \le \frac{30-25}{\sqrt{75/4}}\Big) = \mathbb{P}(-1.1547 \le Z \le 1.1547),$$

   where $Z \sim \mathsf{N}(0,1)$. The cdf of the standard normal distribution in 1.1547 is $\mathbb{P}(Z \le 1.1547) = \Phi(1.1547) = 0.875893$. Hence, $\mathbb{P}(-1.1547 \le Z \le 1.1547) = \Phi(1.1547) - (1 - \Phi(1.1547)) = 2\Phi(1.1547) - 1 = 0.752$. [The exact answer is 0.796682.]
7. (a) The Jacobian is

   $$\left|\det \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix}\right| = \left|\det \begin{pmatrix} v & u \\ 1-v & -u \end{pmatrix}\right| = |-u| = u\ .$$

   (b) The joint pdf of U and V follows from the transformation rule:

   $$f_{U,V}(u,v) = f_{X,Y}(x,y)\, u = \mathrm{e}^{-(x+y)}\, u = \mathrm{e}^{-u}\, u,$$

   for $u \ge 0$ and $0 \le v \le 1$.

(c) By integrating over u we see that V is uniformly distributed over $[0,1]$, and by integrating over v we find that $f_U(u) = u\mathrm{e}^{-u}$, $u \geq 0$. Thus, the joint pdf of U and V is equal to the marginal pdfs of U and V, and hence U and V are independent.

8. Note that $f(x,y)$ can be written as the product of $f_1(x) = \mathrm{e}^{-x}, x \geq 0$ and $f_2(y) = 2\mathrm{e}^{-2y}, y \geq 0$. It follows that X and Y are independent random variables, and that $X \sim \mathsf{Exp}(1)$ and $Y \sim \mathsf{Exp}(2)$.

   (a) Because X and Y are independent: $\mathbb{E}[XY] = \mathbb{E}[X] \times \mathbb{E}[Y] = 1 \times 1/2 = 1/2$.

   (b) $\mathrm{Cov}(X+Y, X-Y) = \mathrm{Cov}(X,X) - \mathrm{Cov}(X,Y) + \mathrm{Cov}(Y,X) - \mathrm{Cov}(Y,Y) = \mathrm{Var}(X) - \mathrm{Var}(Y) = 1 - 1/4 = 3/4$.

9. (a) Take the average $\bar{x} = (x_1 + \cdots + x_n)/n$.

   (b) $\mathbb{E}\bar{X} = \mathbb{E}[(X_1 + \cdots + X_n)/n] = \frac{1}{n}(\mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n]) = \frac{1}{n}\,n\mu = \mu$.

   (c) $\mathrm{Var}(\bar{X}) = \mathrm{Var}[(X_1 + \cdots + X_n)/n] = \frac{1}{n^2}(\mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_n]) = \frac{1}{n^2}\,n\sigma^2 = \sigma^2/n$. This goes to 0 as $n \to \infty$.

   (d) $\mathsf{N}(\mu, \sigma^2/\mathsf{n})$.

   (e) The larger n is, the more accurately μ can be approximated with $\bar{x}$.

10. Similar to question 6: $\mathbb{P}(X \geq 60) \approx \mathbb{P}(Y \geq 60)$, with $Y \sim \mathsf{N}(50, 25)$. Moreover, $\mathbb{P}(Y \geq 60) = \mathbb{P}(Z \geq (60-50)/5) = \mathbb{P}(Z \geq 2) = 1 - \Phi(2) = 0.02275$, where $Z \sim \mathsf{N}(0,1)$.

11. First draw the graph of the function $y = x^2 - 4$ on the interval [1,3]. Note that the function is increasing from -3 to 5. To find the pdf, first calculate the cdf:

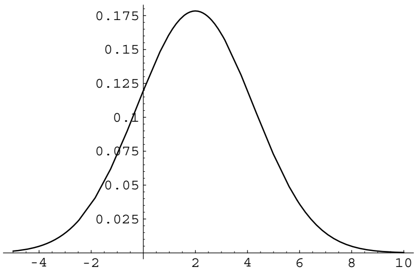
$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(X^2 - 4 \leq y) = \mathbb{P}(X \leq \sqrt{y+4}) = F_X(\sqrt{y+4}), \quad -3 \leq y \leq 5\,.$$

Now take the derivative with respect to y:

$$f_Y(y) = \frac{d}{dy}F_X(\sqrt{y+4}) = f_X(\sqrt{y+4}) \times \frac{1}{2\sqrt{4+y}} = \frac{1}{4\sqrt{4+y}}, \quad -3 \leq y \leq 5\,.$$

12. $\mathsf{U}(9, 17)$.

13. (a)

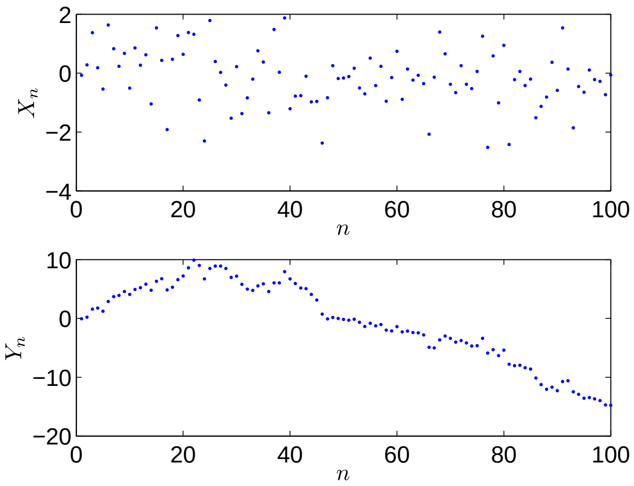


   (b) $\mathbb{P}(Y \leq 5) = \Phi((5-2)/\sqrt{5}) \approx 0.9101$.

(c) $Z \sim \mathsf{N}(3\times 2-4, 3^2\times 5) = \mathsf{N}(2,45)$. $\mathbb{P}(1 \le Z \le 5) = \mathbb{P}((1-2)/\sqrt{45} \le V \le (5-1)/\sqrt{45})$, with $V \sim \mathsf{N}(0,1)$. The latter probability is equal to $\Phi(4/\sqrt{45}) - (1-\Phi(1/\sqrt{45})) \approx 0.2838$.

14. See the notes.

15. (a)



(b) $\mathbb{E}[Y_n] = n\mathbb{E}[X_1] = 0$, and $\mathrm{Var}(Y_n) = n\mathrm{Var}(X_1) = n$.

(c) $\mathrm{Cov}(Y_m, Y_n) = \mathrm{Cov}(Y_m, Y_m + X_{m+1} + \cdots + X_n) = \mathrm{Var}(Y_m) + 0 = m$.

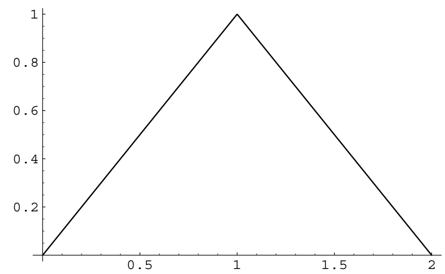
16. (a) $Z_8 \sim \mathsf{N}(8\times 75, 8\times 100) = \mathsf{N}(600, 800)$. $\mathbb{P}(Z_8 \ge 650) = 1 - \mathbb{P}(Z_8 \le 650) = 1 - \Phi((650-600)/\sqrt{800}) = 1 - \Phi(1.7678) \approx 0.0385$.

(b) $\mathbb{P}(Z_n \ge 650) = 1 - \Phi((650 - n75)/\sqrt{n100})$. For $n = 8$ the probability is 0.0385. For $n = 7$ it is much smaller than 0.01. So the largest such n is $n = 7$.

17. (a) $\mathbb{P}(0.150-0.004 \le X \le 0.150+0.004) = \mathbb{P}((0.150-0.004-0.151)/0.003 \le Z \le (0.150+0.004-0.151)/0.003) = \mathbb{P}(-1.66667 \le Z \le 1) = \Phi(1) - (1-\Phi(1.66667)) \approx 0.794$, where $Z \sim \mathsf{N}(0,1)$.

(b) Note first that $\bar{X} \sim \mathsf{N}(0.151, (0.003)^2/25)$. Thus, $\mathbb{P}(0.150 - 0.004 \le \bar{X} \le 0.150 + 0.004) = \mathbb{P}((0.150-0.004-0.151)/(0.003/5) \le Z \le (0.150+0.004-0.151)/(0.003/5)) = \mathbb{P}(-1.66667\times 5 \le Z \le 5) = \Phi(5) - (1 - \Phi(8.3333)) \approx 1$.

18. (a)



(b) $\mathsf{N}(10, 20/12)$, because the expectation of $\mathsf{U}(0,1)$ is $1/2$ and the variance is $1/12$.

(c) $\mathbb{P}(\bar{X} > 0.6) = \mathbb{P}(X_1 + \cdots + X_{20} > 12) \approx \mathbb{P}(Y > 20)$, with $Y \sim \mathsf{N}(10, 20/12)$. Now, $\mathbb{P}(Y > 12) = 1 - \mathbb{P}(Y \leq 12) = 1 - \Phi((12 - 10)/\sqrt{20/12}) = 1 - \Phi(1.5492) \approx 0.0607$.

19. The payout is X^2, with $X \sim \mathsf{Ber}(10, 1/2)$. The expected payout is $\mathbb{E}X^2 = \mathrm{Var}(X) + (\mathbb{E}X)^2 = 2.5 + 5^2 = 27.5$. So, if you pay less than 27.5 dollars in advance, your expected profit per game is positive.

Appendix B

Sample Exams

B.1 Exam 1

1. Two fair dice are thrown and the sum of the face values, Z say, is noted.

   (a) Give the pmf of Z in table form: [4]

z	$*$	$*$	$*$	$\cdots$
$\mathbb{P}(Z = z)$	$*$	$*$	$*$	$\cdots$

   (b) Calculate the variance of Z. [4]

   (c) Consider the game, where a player throws two fair dice, and is paid $Y = (Z - 7)^2$ dollars, with Z the sum of face values. To enter the game the player is required to pay 5 dollars.

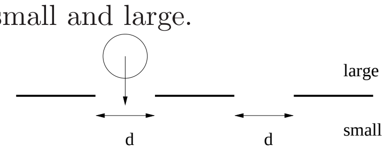
   What is the expected profit (or loss) of the player, if he/she plays the game 100 times (each time paying 5 dollars to play)? [4]

2. Consider two electrical components, both with an expected lifetime of 3 years. The lifetime of component 1, X say, is assumed to have an exponential distribution, and the lifetime of component 2, Y say, is modeled via a normal distribution with a standard deviation of 1/2 years.

   (a) What is the probability that component 1 is still functioning after 4.5 years, given that it still works after 4 years? [6]

   (b) What is the probability that component 2 is still functioning after 4.5 years, given that it still works after 4 years? [6]

3. A sieve with diameter d is used to separate a large number of blueberries into two classes: small and large.

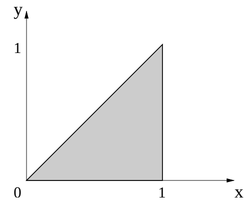


   Suppose the diameters of the blueberries are normally distributed with an expectation $\mu = 1$ (cm) and a standard deviation $\sigma = 0.1$ (cm).

(a) How large should the diameter of the sieve be, so that the proportion of large blueberries is 30%? [6]

(b) Suppose that the diameter is chosen such as in (a). What is the probability that out of 1000 blueberries, fewer than 280 end up in the "large" class? [6]

4. We draw a random vector (X, Y) non-uniformly from the triangle $(0,0)$–$(1,0)$–$(1,1)$



in the following way: First we draw X uniformly on $[0,1]$. Then, given $X = x$ we draw Y uniformly on $[0, x]$.

(a) Give the conditional pdf of Y given $X = x$. Specify where this conditional pdf is 0. [3]

(b) Find the joint pdf of X and Y. [4]

(c) Calculate the pdf of Y and sketch its graph. [5]

5. We draw n numbers independently and uniformly from the interval [0,1] and note their sum S_n.

(a) Draw the graph of the pdf of S_2. [3]

(b) What is approximately the distribution of S_{20}? [4]

(c) Calculate the probability that the *average* of the 20 numbers is greater than 0.6. [5]

B.2 Exam 2

1. Two fair dice are thrown and the smallest of the face values, Z say, is noted.

(a) Give the probability mass function (pmf) of Z in table form: [3]

z	*	*	*	...
$\mathbb{P}(Z = z)$	*	*	*	...

(b) Calculate the expectation of $1/Z$. [2]

(c) Consider the game, where a player throws two fair dice, and is paid Z dollars, with Z as above. To enter the game the player is required to pay 3 dollars.

What is the expected profit (or loss) of the player, if he/she plays the game 100 times (each time paying 3 dollars to play)? [3]

2. Let U and V be independent random variables, with $\mathbb{P}(U = 1) = \mathbb{P}(V = 1) = 1/4$ and $\mathbb{P}(U = -1) = \mathbb{P}(V = -1) = 3/4$. Define $X = U/V$ and $Y = U + V$.

   (a) Give the joint pmf of X and Y. [4]

   (b) Calculate $\text{Cov}(X, Y)$. [4]

3. In a binary transmission channel, a 1 is transmitted with probability 1/4 and a 0 with probability 3/4. The conditional probability of receiving a 1 when a 1 was sent is 0.90, the conditional probability of receiving a 0 when a 0 was sent is 0.95.

   (a) What is the probability that a 1 is received? [3]

   (b) Given that a 1 is received, what is the probability that a 1 was transmitted? [5]

4. Consider the probability density function (pdf) given by

$$f(x) = \begin{cases} 4\,\mathrm{e}^{-4(x-1)}, & x \geq 1\,, \\ 0 & x < 1\,. \end{cases}$$

   (a) If X is distributed according to this pdf f, what is the expectation of X? [3]

   (b) Specify how one can generate a random variable X from this pdf, using a random number generator that outputs $U \sim \mathsf{U}(0, 1)$. [5]

5. A certain type of electrical component has an exponential lifetime distribution with an expected lifetime of 1/2 year. When the component fails it is immediately replaced by a second (new) component; when the second component fails, it is replaced by a third, etc. Suppose there are 10 such identical components. Let T be the time that the last of the components fails.

   (a) What is the expectation and variance of T? [3]

   (b) Approximate, using the central limit theorem and the table of the standard normal cdf, the probability that T exceeds 6 years. [3]

   (c) What is the exact distribution of T? [2]

Appendix C

Summary of Formulas

1. **Sum rule:** $\boxed{\mathbb{P}(\bigcup_i A_i) = \sum_i \mathbb{P}(A_i),}$ when $A_1, A_2, \ldots$ are disjoint.

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

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

4. **Cdf** of X: $F(x) = \mathbb{P}(X \leq x), \ x \in \mathbb{R}$.

5. **Pmf** of X: (discrete r.v.) $f(x) = \mathbb{P}(X = x)$.

6. **Pdf** of X: (continuous r.v.) $f(x) = F'(x)$.

7. For a discrete r.v. X:

$$\boxed{\mathbb{P}(X \in B) = \sum_{x \in B} \mathbb{P}(X = x).}$$

8. For a continuous r.v. X with pdf f:

$$\boxed{\mathbb{P}(X \in B) = \int_B f(x)\, dx.}$$

9. In particular (continuous), $F(x) = \int_{-\infty}^{x} f(u)\, du$.

10. Similar results 7-8 hold for random vectors, e.g.

$$\mathbb{P}((X,Y) \in B) = \iint_B f_{X,Y}(x,y)\, dx\, dy\ .$$

11. Marginal from joint pdf: $f_X(x) = \int f_{X,Y}(x,y)\, dy$.

12. **Important discrete distributions**:

Distr.	pmf	$x \in$
$\mathsf{Ber}(p)$	$p^x(1-p)^{1-x}$	$\{0,1\}$
$\mathsf{Bin}(n,p)$	$\binom{n}{x} p^x(1-p)^{n-x}$	$\{0,1,\ldots,n\}$
$\mathsf{Poi}(\lambda)$	$\mathrm{e}^{-\lambda} \frac{\lambda^x}{x!}$	$\{0,1,\ldots\}$
$\mathsf{G}(p)$	$p(1-p)^{x-1}$	$\{1,2,\ldots\}$
$\mathsf{Hyp}(n,r,N)$	$\frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$	$\{0,\ldots,n\}$

13. **Important continuous distributions**:

Distr.	pdf	$x \in$
$\mathsf{U}[a,b]$	$\frac{1}{b-a}$	$[a,b]$
$\mathsf{Exp}(\lambda)$	$\lambda\,\mathrm{e}^{-\lambda x}$	$\mathbb{R}_+$
$\mathsf{Gam}(\alpha,\lambda)$	$\frac{\lambda^\alpha x^{\alpha-1}\mathrm{e}^{-\lambda x}}{\Gamma(\alpha)}$	$\mathbb{R}_+$
$\mathsf{N}(\mu,\sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}}\,\mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	$\mathbb{R}$

14. **Conditional probability:** $\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)}$.

15. **Law of total probability**:

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

with $B_1, B_2, \ldots, B_n$ a partition of Ω.

16. **Bayes' Rule:** $\mathbb{P}(B_j|A) = \frac{\mathbb{P}(B_j)\,\mathbb{P}(A|B_j)}{\sum_{i=1}^n \mathbb{P}(B_i)\,\mathbb{P}(A|B_i)}$.

17. **Product rule**:

$$\boxed{\mathbb{P}(A_1\cdots A_n) = \mathbb{P}(A_1)\,\mathbb{P}(A_2 \mid A_1)\cdots\mathbb{P}(A_n \mid A_1\cdots A_{n-1}).}$$

18. **Memoryless property** (Exp and G distribution):
$\mathbb{P}(X > s+t \mid X > s) = \mathbb{P}(X > t),\ \forall s,t.$

19. **Independent events**: $\mathbb{P}(A\cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$.

20. **Independent r.v.'s:** (discrete)

$$\boxed{\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{k=1}^n \mathbb{P}(X_k = x_k)\ .}$$

21. **Independent r.v.'s:** (continuous)

$$\boxed{f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \prod_{k=1}^n f_{X_k}(x_k)\ .}$$

22. **Expectation** (discr.): $\mathbb{E}X = \sum_x x\,\mathbb{P}(X = x)$.

23. (of function) $\mathbb{E}\,g(X) = \sum_x g(x)\,\mathbb{P}(X = x)$.

24. **Expectation** (cont.): $\mathbb{E}X = \int x\,f(x)\,dx$.

25. (of function) $\mathbb{E}\,g(X) = \int g(x)f(x)\,dx$,

26. Similar results 18–21 hold for random vectors.

27. **Expected sum :** $\mathbb{E}(aX + bY) = a\,\mathbb{E}X + b\,\mathbb{E}Y$.

28. **Expected product** (only if X, Y independent):
$$\mathbb{E}[X\,Y] = \mathbb{E}X\,\mathbb{E}Y.$$

29. **Markov inequality:** $\mathbb{P}(X \geq x) \leq \frac{\mathbb{E}X}{x}$.

30. $\mathbb{E}X$ **and Var**(X) **for various distributions**:

	$\mathbb{E}X$	$\mathrm{Var}(X)$
$\mathsf{Ber}(p)$	p	$p(1-p)$
$\mathsf{Bin}(n,p)$	np	$np(1-p)$
$\mathsf{G}(p)$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$\mathsf{Poi}(\lambda)$	λ	λ
$\mathsf{Hyp}(n, pN, N)$	np	$np(1-p)\frac{N-n}{N-1}$
$\mathsf{U}(a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$\mathsf{Exp}(\lambda)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
$\mathsf{Gam}(\alpha,\lambda)$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$
$\mathsf{N}(\mu,\sigma^2)$	μ	σ^2

31. n**-th moment:** $\mathbb{E}X^n$.

32. **Covariance:** $\mathrm{cov}(X,Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y)$.

33. **Properties of Var and Cov**:

$\mathrm{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2$.
$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$.
$\mathrm{cov}(X,Y) = \mathbb{E}XY - \mathbb{E}X\mathbb{E}Y$.
$\mathrm{cov}(X,Y) = \mathrm{cov}(Y,X)$.
$\mathrm{cov}(aX + bY, Z) = a\,\mathrm{cov}(X,Z) + b\,\mathrm{cov}(Y,Z)$.
$\mathrm{cov}(X,X) = \mathrm{Var}(X)$.
$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{cov}(X,Y)$.
X and Y independent $\Longrightarrow \mathrm{cov}(X,Y) = 0$.

34. **Probability Generating Function (PGF)**:

$$G(z) := \mathbb{E}\,z^N = \sum_{n=0}^{\infty} \mathbb{P}(N = n) z^n, \qquad |z| < 1\,.$$

35. **PGFs for various distributions:**

$\mathsf{Ber}(p)$	$1 - p + zp$
$\mathsf{Bin}(n,p)$	$(1 - p + zp)^n$
$\mathsf{G}(p)$	$\frac{z\,p}{1 - z\,(1-p)}$
$\mathsf{Poi}(\lambda)$	$\mathrm{e}^{-\lambda(1-z)}$

36. $\mathbb{P}(N = n) = \frac{1}{n!}\, G^{(n)}(0)$. (n-th derivative, at 0)

37. $\mathbb{E}N = G'(1)$

38. $\mathrm{Var}(N) = G''(1) + G'(1) - (G'(1))^2$.

39. **Moment Generating Function (MGF)**:

$$M(s) = \mathbb{E}\,\mathrm{e}^{sX} = \int_{-\infty}^{\infty} \mathrm{e}^{sx}\, f(x)\, dx \,,$$

$s \in I \subset \mathbb{R}$, for r.v.'s X for which all moments exist.

40. **MGFs for various distributions:**

$\mathsf{U}(a,b)$	$\frac{\mathrm{e}^{bs}-\mathrm{e}^{as}}{s(b-a)}$
$\mathsf{Gam}(\alpha,\lambda)$	$\left(\frac{\lambda}{\lambda-s}\right)^{\alpha}$
$\mathsf{N}(\mu,\sigma^2)$	$\mathrm{e}^{s\mu+\sigma^2 s^2/2}$

41. **Moment property**: $\mathbb{E}X^n = M^{(n)}(0)$.

42. $M_{X+Y}(t) = M_X(t)\, M_Y(t)$, $\forall t$, if X, Y independent.

43. If $X_i \sim \mathsf{N}(\mu_i, \sigma_i^2)$, $i = 1, 2, \ldots, n$ (independent),
$a + \sum_{i=1}^n b_i\, X_i \sim \mathsf{N}\left(a + \sum_{i=1}^n b_i\, \mu_i,\ \sum_{i=1}^n b_i^2\, \sigma_i^2\right)$.

44. **Conditional pmf/pdf**
$f_{Y\,|\,X}(y\,|\,x) := \frac{f_{X,Y}(x,y)}{f_X(x)}, \quad y \in \mathbb{R}.$

45. The corresponding **conditional expectation** (discrete case):
$\mathbb{E}[Y\,|\,X = x] = \sum_y y\, f_{Y\,|\,X}(y\,|\,x).$

46. **Linear transformation**: $f_{\boldsymbol{Z}}(\boldsymbol{z}) = \frac{f_{\boldsymbol{X}}(A^{-1}\boldsymbol{z})}{|A|}$.

47. **General transformation**: $f_{\boldsymbol{Z}}(\boldsymbol{z}) = \dfrac{f_{\boldsymbol{X}}(\boldsymbol{x})}{|J_{\boldsymbol{x}}(g)|}$, with $\boldsymbol{x} = g^{-1}(\boldsymbol{z})$, where $|J_{\boldsymbol{x}}(g)|$ is the Jacobian of g evaluated at $\boldsymbol{x}$.

48. Pdf of the **multivariate normal** distribution:

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = \frac{1}{\sqrt{(2\pi)^n\, |\Sigma|}}\, \mathrm{e}^{-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{z}-\boldsymbol{\mu})} \,.$$

Σ is the covariance matrix, and $\boldsymbol{\mu}$ the mean vector.

49. If $\boldsymbol{X}$ is a column vector with independent $\mathsf{N}(0,1)$ components, and B is a matrix with $\Sigma = BB^T$ (such a B can always be found), then $\boldsymbol{Z} = \boldsymbol{\mu} + B\boldsymbol{X}$ has a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ.

50. **Weak Law of Large Numbers:**

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \epsilon\right) = 0, \quad \forall \epsilon\ .$$

51. **Strong Law of Large Numbers**:

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{S_n}{n} = \mu\right) = 1\ .$$

52. **Central Limit Theorem**:

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right) = \Phi(x),$$

where Φ is the cdf of the standard normal distribution.

53. **Normal Approximation to Binomial:** If $X \sim \mathsf{Bin}(n, p)$, then, for large n, $\mathbb{P}(X \leq k) \approx \mathbb{P}(Y \leq k)$, where $Y \sim \mathsf{N}(np, np(1-p))$.

Other Mathematical Formulas

1. Factorial. $n! = n\,(n-1)\,(n-2)\cdots 1$. Gives the number of *permutations* (orderings) of $\{1, \ldots, n\}$.
2. Binomial coefficient. $\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$. Gives the number *combinations* (no order) of k different numbers from $\{1, \ldots, n\}$.
3. Newton's binomial theorem: $(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k\, b^{n-k}$.
4. Geometric sum: $1 + a + a^2 + \cdots + a^n = \frac{1-a^{n+1}}{1-a}$ $(a \neq 1)$.
   If $|a| < 1$ then $1 + a + a^2 + \cdots = \frac{1}{1-a}$.
5. Logarithms:
   (a) $\log(x\,y) = \log x + \log y$.
   (b) $\mathrm{e}^{\log x} = x$.
6. Exponential:
   (a) $\mathrm{e}^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$.
   (b) $\mathrm{e}^x = \lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n$.
   (c) $\mathrm{e}^{x+y} = \mathrm{e}^x\,\mathrm{e}^y$.
7. Differentiation:
   (a) $(f+g)' = f' + g'$,
   (b) $(fg)' = f'g + fg'$,
   (c) $\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}$.

   (d) $\frac{d}{dx}x^n = n\,x^{n-1}$.
   (e) $\frac{d}{dx}\mathrm{e}^x = \mathrm{e}^x$.
   (f) $\frac{d}{dx}\log(x) = \frac{1}{x}$.

8. Chain rule: $(f(g(x)))' = f'(g(x))\,g'(x)$.

9. Integration: $\int_a^b f(x)\,dx = [F(x)]_a^b = F(b) - F(a)$, where $F' = f$.

10. Integration by parts: $\int_a^b f(x)\,G(x)\,dx = [F(x)\,G(x)]_a^b - \int_a^b F(x)\,g(x)\,dx$. (Here $F' = f$ and $G' = f$.)

11. Jacobian: Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be an n-dimensional vector, and $g(\boldsymbol{x}) = (g_1(\boldsymbol{x}), \ldots, g_n(\boldsymbol{x}))$ be a function from $\mathbb{R}^n$ to $\mathbb{R}^n$. The *matrix of Jacobi* is the matrix of partial derivatives: $(\partial g_i/\partial x_j)$. The corresponding determinant is called the *Jacobian*. In the neighbourhood of any fixed point, g behaves like a *linear transformation* specified by the matrix of Jacobi at that point.

12. Γ function: $\Gamma(\alpha) = \int_0^\infty u^{\alpha-1}\,\mathrm{e}^{-u}\,du$, $\alpha > 0$. $\Gamma(\alpha+1) = \alpha\,\Gamma(\alpha)$, for $\alpha \in \mathbb{R}_+$. $\Gamma(n) = (n-1)!$ for $n = 1, 2, \ldots$. $\Gamma(1/2) = \sqrt{\pi}$.

Appendix D

Statistical Tables

- **Standard normal distribution**
- **Binomial distribution with** $p = 1/2$

Standard normal distribution

This table gives the cummulative distribution function (cdf) Φ of a $\mathsf{N}(0,1)$-distributed random variable Z.

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

The last column gives the probability density function (pdf) φ of the $\mathsf{N}(0,1)$-distribution

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

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09	$\phi(z)$
0.0	5000	5040	5080	5120	5160	5199	5239	5279	5319	5359	0.3989
0.1	5398	5438	5478	5517	5557	5596	5636	5675	5714	5753	0.3970
0.2	5793	5832	5871	5910	5948	5987	6026	6064	6103	6141	0.3910
0.3	6179	6217	6255	6293	6331	6368	6406	6443	6480	6517	0.3814
0.4	6554	6591	6628	6664	6700	6736	6772	6808	6844	6879	0.3683
0.5	6915	6950	6985	7019	7054	7088	7123	7157	7190	7224	0.3521
0.6	7257	7291	7324	7357	7389	7422	7454	7486	7517	7549	0.3332
0.7	7580	7611	7642	7673	7704	7734	7764	7794	7823	7852	0.3123
0.8	7881	7910	7939	7967	7995	8023	8051	8078	8106	8133	0.2897
0.9	8159	8186	8212	8238	8264	8289	8315	8340	8365	8389	0.2661
1.0	8413	8438	8461	8485	8508	8531	8554	8577	8599	8621	0.2420
1.1	8643	8665	8686	8708	8729	8749	8770	8790	8810	8830	0.2179
1.2	8849	8869	8888	8907	8925	8944	8962	8980	8997	9015	0.1942
1.3	9032	9049	9066	9082	9099	9115	9131	9147	9162	9177	0.1714
1.4	9192	9207	9222	9236	9251	9265	9279	9292	9306	9319	0.1497
1.5	9332	9345	9357	9370	9382	9394	9406	9418	9429	9441	0.1295
1.6	9452	9463	9474	9484	9495	9505	9515	9525	9535	9545	0.1109
1.7	9554	9564	9573	9582	9591	9599	9608	9616	9625	9633	0.0940
1.8	9641	9649	9656	9664	9671	9678	9686	9693	9699	9706	0.0790
1.9	9713	9719	9726	9732	9738	9744	9750	9756	9761	9767	0.0656
2.0	9772	9778	9783	9788	9793	9798	9803	9808	9812	9817	0.0540
2.1	9821	9826	9830	9834	9838	9842	9846	9850	9854	9857	0.0440
2.2	9861	9864	9868	9871	9875	9878	9881	9884	9887	9890	0.0355
2.3	9893	9896	9898	9901	9904	9906	9909	9911	9913	9916	0.0283
2.4	9918	9920	9922	9925	9927	9929	9931	9932	9934	9936	0.0224
2.5	9938	9940	9941	9943	9945	9946	9948	9949	9951	9952	0.0175
2.6	9953	9955	9956	9957	9959	9960	9961	9962	9963	9964	0.0136
2.7	9965	9966	9967	9968	9969	9970	9971	9972	9973	9974	0.0104
2.8	9974	9975	9976	9977	9977	9978	9979	9979	9980	9981	0.0079
2.9	9981	9982	9982	9983	9984	9984	9985	9985	9986	9986	0.0060
3.0	9987	9987	9987	9988	9988	9989	9989	9989	9990	9990	0.0044
3.1	9990	9991	9991	9991	9992	9992	9992	9992	9993	9993	0.0033
3.2	9993	9993	9994	9994	9994	9994	9994	9995	9995	9995	0.0024
3.3	9995	9995	9995	9996	9996	9996	9996	9996	9996	9997	0.0017
3.4	9997	9997	9997	9997	9997	9997	9997	9997	9997	9998	0.0012
3.5	9998	9998	9998	9998	9998	9998	9998	9998	9998	9998	0.0009
3.6	9998	9998	9999	9999	9999	9999	9999	9999	9999	9999	0.0006

Example: $\Phi(1.65) = \mathbb{P}(Z \le 1.65) = 0.9505$

Binomial distribution with $p = 1/2$

Below is a list of $F(c) = \mathbb{P}(X \leq c) = \sum_{k=0}^{c} \binom{n}{k} \left(\frac{1}{2}\right)^n$ if $X \sim \mathsf{Bin}\left(n, \frac{1}{2}\right)$.
(Only the first 4 digits are given, and probabilities ≥ 0.99995 have been omitted.)

n	c	$F(c)$	n	c	$F(c)$	n	c	$F(c)$	n	c	$F(c)$	n	c	$F(c)$
9	0	0020	14	0	0001	17	13	9936	21	0	0000	23	18	9987
	1	0195		1	0009		14	9988		1	0000		19	9998
	2	0898		2	0065		15	9999		2	0001			
	3	2539		3	0287					3	0007	24	0	0000
	4	5000		4	0898	18	0	0000		4	0036		1	0000
	5	7461		5	2120		1	0001		5	0133		2	0000
	6	9102		6	3953		2	0007		6	0392		3	0001
	7	9805		7	6047		3	0038		7	0946		4	0008
	8	9980		8	7880		4	0154		8	1917		5	0033
				9	9102		5	0481		9	3318		6	0113
10	0	0010		10	9713		6	1189		10	5000		7	0320
	1	0107		11	9935		7	2403		11	6682		8	0758
	2	0547		12	9991		8	4073		12	8083		9	1537
	3	1719		13	9999		9	5927		13	9054		10	2706
	4	3770					10	7597		14	9608		11	4194
	5	6230	15	0	0000		11	8811		15	9867		12	5806
	6	8281		1	0005		12	9519		16	9964		13	7294
	7	9453		2	0037		13	9846		17	9993		14	8463
	8	9893		3	0176		14	9962		18	9999		15	9242
	9	9990		4	0592		15	9993					16	9680
				5	1509		16	9999	22	0	0000		17	9887
11	0	0005		6	3036					1	0000		18	9967
	1	0059		7	5000	19	0	0000		2	0001		19	9992
	2	0327		8	6964		1	0000		3	0004		20	9999
	3	1133		9	8491		2	0004		4	0022			
	4	2744		10	9408		3	0022		5	0085	25	0	0000
	5	5000		11	9824		4	0096		6	0262		1	0000
	6	7256		12	9963		5	0318		7	0669		2	0000
	7	8867		13	9995		6	0835		8	1431		3	0001
	8	9673					7	1796		9	2617		4	0005
	9	9941	16	0	0000		8	3238		10	4159		5	0020
	10	9995		1	0003		9	5000		11	5841		6	0073
				2	0021		10	6762		12	7383		7	0216
12	0	0002		3	0106		11	8204		13	8569		8	0539
	1	0032		4	0384		12	9165		14	9331		9	1148
	2	0193		5	1051		13	9682		15	9738		10	2122
	3	0730		6	2272		14	9904		16	9915		11	3450
	4	1938		7	4018		15	9978		17	9978		12	5000
	5	3872		8	5982		16	9996		18	9996		13	6550
	6	6128		9	7728					19	9999		14	7878
	7	8062		10	8949	20	0	0000					15	8852
	8	9270		11	9616		1	0000	23	0	0000		16	9461
	9	9807		12	9894		2	0002		1	0000		17	9784
	10	9968		13	9979		3	0013		2	0000		18	9927
	11	9998		14	9997		4	0059		3	0002		19	9980
							5	0207		4	0013		20	9995
13	0	0001	17	0	0000		6	0577		5	0053		21	9999
	1	0017		1	0001		7	1316		6	0173			
	2	0112		2	0012		8	2517		7	0466			
	3	0461		3	0064		9	4119		8	1050			
	4	1334		4	0245		10	5881		9	2024			
	5	2905		5	0717		11	7483		10	3388			
	6	5000		6	1662		12	8684		11	5000			
	7	7095		7	3145		13	9423		12	6612			
	8	8666		8	5000		14	9793		13	7976			
	9	9539		9	6855		15	9941		14	8950			
	10	9888		10	8338		16	9987		15	9534			
	11	9983		11	9283		17	9998		16	9827			
	12	9999		12	9755					17	9947			

Index