STAT 2857 — Probability and Statistics I

by Hristo Sendov

Contents

1 Sample spaces, random variables, events **2**

2 Set operations **4**

3 Finite and infinite sets **6**

4 Definition of probabilities **8**

4.1 Properties of probabilities . . . 9
4.2 Probabilities on finite sample spaces . . . 12
4.3 Probabilities on countably infinite sample spaces . . . 13
4.4 Probabilities on uncountably infinite sample spaces . . . 13
4.5 The inclusion-exclusion formula . . . 14
4.6 Examples . . . 15

5 Combinatorics **17**

5.1 Counting ordered objects when we choose with replacement . . . 18
5.2 Counting ordered objects when we choose without replacement . . . 19
5.3 Counting unordered objects when we choose without replacement . . . 21
5.3.1 The multinomial theorem . . . 23
5.4 Counting unordered objects when we choose with replacement . . . 25
5.5 Binomial distribution . . . 28
5.6 More examples . . . 30
5.6.1 Sample spaces made of sequences . . . 36

6 Conditional probability and independence **38**

6.1 Conditional probability . . . 38
6.2 Independence . . . 40
6.3 Bayes' theorem . . . 45

7 Random variables **49**

7.1 Density function . . . 49
7.2 Cumulative distribution function . . . 50

8 Expected value & variance **55**

9 Important distributions 62
9.1 The Poisson Distribution 62
9.2 The exponential distribution 68
9.3 The geometric distribution 69
9.4 The negative binomial distribution 71
9.5 The hypergeometric distribution 72
9.6 The normal distribution 74
9.7 Gamma distribution 82
9.8 Beta function 84

10 Jointly distributed random variables 85

11 Independent random variables 92
11.1 Sums of independent random variables 95
11.1.1 Discrete random variables 95
11.1.2 Continuous random variables 95
11.1.3 Using moment generation functions 99

12 Covariance and correlation 100
12.1 Covariance 100
12.1.1 Examples and applications 102
12.2 Correlation 106

13 Limit theorems 108
13.1 The central limit theorem 112

14 Conditional distributions 114

15 Appendix A: The Prosecutor's Fallacy 117

1 Sample spaces, random variables, events

Definition 1. Suppose we have an experiment whose outcome, denoted by small Greek letter ω, depends on chance. The *sample space* of the experiment is the set of all possible outcomes.

We generally denote a sample space by the capital Greek letter Ω. The sample space is either finite or infinite. If the sample space is infinite, it can be countably infinite or not. A set is countably infinite if its elements can be ordered in a sequence so, looking at one element, one can tell which is the next one. First we shall consider chance experiments with a finite number of possible outcomes, that is

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}.$$

For example, we roll a die and the possible outcomes are $\Omega = \{1, 2, 3, 4, 5, 6\}$ corresponding to the side that turns up. That is, $\omega_1 = 1, \omega_2 = 2, \ldots, \omega_6 = 6$. We toss a coin with possible outcomes H (heads) and T (tails). That is $\Omega = \{H, T\}$ or $\omega_1 = H$ and $\omega_2 = T$. If we toss a coin twice one after another then the possible outcomes are $\Omega = \{HH, HT, TH, TT\}$. If we pick a card at random from a deck of cards then the possible outcomes are $\Omega = \{2 - heart, 2 - spade, 2 - club, 2 - diamond, \ldots, ace - heart, ace - spade, ace - club, ace - diamond\}$, a total of 52 outcomes. If we choose a person at random on the streets of London, then there are more than $400,000$ possible outcomes of our experiment. How you decide to describe them is up to you—you may use their names (but some names may repeat) or just their SIN numbers. If you pick a person at random from the whole planet, then we may safely assume that there are infinitely many possible outcomes. If we select at random a grain of sand from a beach then we may safely assume that there are infinitely many possible outcomes of the experiment.

After the experiment is performed, the outcome of the experiment ω is measured or just observed and the value is of this measurement is denoted by $X(\omega)$. The value $X(\omega)$ is usually a real number. Thus X is just a function from Ω to $\mathbb{R}$.

For example, if we roll a die then we may be interested in the number of dots on the upper side of the die. In that case the measurement is just counting the number of dots, that is $X(1) = 1$, $\cdots$, $X(6) = 6$. If we toss a coin twice then we may want to measure the number of times heads comes up. Then $X(HH) = 2, X(HT) = 1, X(TH) = 1, X(TT) = 0$. If we select at random a grain of sand from a beach and measure its width precisely then we may safely assume that the result will be any number from the interval $[0, 1]$, where 1 stands for 1 millimetre. In that case we have infinitely many possible values of X—any number in $[0, 1]$. If we choose a person at random on the streets of London we may be interested in their weight. Say we stop Mary, then $X(Mary) =$ *the weight of Mary.* It may be convenient to assume that the weight of a person can take any value in the interval $[1, 300]$ where the measuring unit is a kilogram.

Definition 2. The function X defined on the set Ω taking values in $\mathbb{R}$ is called *a random variable.* [1]

And here comes the first shock: there is nothing random about a random variable X. It is just a function from Ω that takes real numbers as values.

Definition 3. Any subset A of the sample space Ω is called an *event.*[2]

For example, we roll a die and the possible outcomes are $\Omega = \{1, 2, 3, 4, 5, 6\}$. The event $A = \{2, 4, 6\}$ corresponds to the statement that the result of the roll is an even number. If the random variable X is just counting the number of dots on the die, then the event A can also be described by saying that X is even. If we toss a coin twice one after another then $\Omega = \{HH, HT, TH, TT\}$. The event "at least one tail came up" is $A = \{TH, HT, TT\}$. The event "no tail came up" is $B = \{HH\}$. If we perform the experiment by tossing two coins and the result is HT, then we say that the event A occurred or that the event B did not occur.

Since events are just subsets of Ω, we need to review the rules and notation for working with sets.

[1] This is not entirely precise and more advanced courses in probability will state that the function X must satisfy additional requirements, but for us this definition will be sufficient.

[2] This is not entirely precise and more advanced courses in probability will state that only certain subsets are events and other subsets are not. For us this definition will be sufficient.

2 Set operations

Let Ω be a set (a sample space). The symbol $\emptyset$ denotes the empty set, the set without any element. Sometimes it is also denoted by $\{\}$ (curly brackets with nothing between them). Let A and B be subsets of Ω, denoted $A \subset \Omega$ and $B \subset \Omega$. We say that A is subset of B, denoted by $A \subset B$ if whenever $\omega \in A$ we have $\omega \in B$. Other notations are $A \subseteq B$, $B \supset A$, $B \supseteq A$. If $A \subset B$ and $B \subset A$, then $A = B$. Note that we always have $\emptyset \subset A$ and $\emptyset \subset B$, that is, the empty set is a subset of any set. If $A \subset B$ and $B \subset C$, then $A \subset C$. Common operations with sets are

union: $A \cup B := \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}$
the set of all outcomes that are in A or in B or both;

intersection: $A \cap B := \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$
the set of all outcomes that are in both A and B;

set-minus: $A \setminus B := \{\omega \in \Omega : \omega \in A \text{ and } \omega \notin B\}$
the set of all outcomes that are in A but not in B;

complement: $A^c := \{\omega \in \Omega : \omega \notin A\}$.
the set of all outcomes that are not in A.

symmetric difference: $A \Delta B := \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B \text{ but } \omega \notin A \cap B\}$
the set of all outcomes that are in A or in B but not in both;

Often $A \setminus B$ is also denoted by $A - B$. Illustrate by Venn diagrams.

The rules for operating with union are

$$\begin{aligned} A \cup B &= B \cup A \\ A \cup A &= A \\ A \cup \emptyset &= A \\ A \cup \Omega &= \Omega \\ \text{if } A \subset B \text{ then } A \cup B &= B. \end{aligned}$$

We can have union of more than two events. If $A_1, \dots, A_n$ are events, then

$$\bigcup_{i=1}^{n} A_i := \text{ the set of outcomes in any of } A_1, A_2, \dots, A_n.$$

We can have union of infinite number of events $A_1, A_2, \dots$, denoted by $\bigcup_{i=1}^{\infty} A_i$. It does not matter in what order we form the union of events:

$$A \cup B \cup C = (A \cup B) \cup C = A \cup (B \cup C).$$

The rules for operating with intersection are

$$\begin{aligned} A \cap B &= B \cap A \\ A \cap A &= A \end{aligned}$$

$$\begin{aligned} A \cap \emptyset &= \emptyset \\ A \cap \Omega &= A \\ \text{if } A \subset B \text{ then } A \cap B &= A. \end{aligned}$$

We can have intersection of more than two events. If $A_1, \dots, A_n$ are events, then

$$\bigcap_{i=1}^{n} A_i := \text{ the set of outcomes in every } A_1, A_2, \dots, A_n.$$

We can have intersection of infinite number of events $A_1, A_2, \dots$, denoted by $\bigcap_{i=1}^{\infty} A_i$. It does not matter in what order we form the intersection of events:

$$A \cap B \cap C = (A \cap B) \cap C = A \cap (B \cap C).$$

The rules for operating with complement are

$$\begin{aligned} (A^c)^c &= A \\ \emptyset^c &= \Omega \\ \Omega^c &= \emptyset \\ A \cup A^c &= \Omega \\ A \cap A^c &= \emptyset. \end{aligned}$$

Let us show that if $A \subset B$, then $B^c \subset A^c$. Indeed, suppose $A \subset B$. We need to show that $B^c \subset A^c$. Take any $\omega \in B^c$, that is ω is not in B. But B contains all the elements of A, hence ω is not in A also. That is $\omega \in A^c$. We are done.

Exercise 4. Show that for any events A, B, and C, we have

(i) $A \setminus B = A \cap B^c$;

(ii) $A = (A \cap B) \cup (A \cap B^c)$;

(iii) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$;

(iv) $(A \cup B)^c = A^c \cap B^c$;

(v) $(A \cap B)^c = A^c \cup B^c$;

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

Show each item above both using a Venn diagram and with a rigorous proof (imitating the argument in the paragraph above this exercise).

Definition 5. Two subsets A and B of Ω are called *disjoint* if $A \cap B = \emptyset$. A sequence of subsets $A_1, A_2, A_3, \dots$ of Ω is called disjoint, or *pairwise disjoint*, if the intersection of any two members of the sequence is empty, that is, $A_i \cap A_j = \emptyset$ for any $i \neq j$ where $i, j \in \{1, 2, 3, \dots\}$.

3 Finite and infinite sets

A set is *finite*, well, if it has finitely many elements, otherwise it is called infinite.

Lemma 6. Suppose $\Omega = \{\omega_1, \dots, \omega_n\}$ is a set with n elements. There are exactly 2^n different subsets (or events) of Ω.

Proof. A subset A of Ω can be specified by stating exactly which elements of Ω are in A and which are not. Consider a $0,1$-vector $(x_1, \dots, x_n)$ with n coordinates, that is $x_i \in \{0,1\}$ for every $i = 1, 2, \dots, n$. Every such vector describes a subset A of Ω. Indeed, we define ω_i to be in A if $x_i = 1$ and ω_i not to be in A if $x_i = 0$. Conversely for any subset A of Ω there is a $0,1$-vector that describes A in the above way. Since there are 2^n different $0,1$ vectors with n coordinates, there are 2^n subsets of Ω. □

The set of all subsets of Ω will be denoted by 2^Ω. For example, if $\Omega = \{a, b, c\}$ has three elements then

$$2^\Omega = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a,b\}, \{a,c\}, \{b,c\}, \{a,b,c\}\}$$

has $2^3 = 8$ elements. Each element of 2^Ω is a subset of Ω.

In the case when Ω has infinitely many elements, it has infinitely many subsets. Unfortunately, some infinities are larger than other infinities. A set with infinitely many elements is called *countably infinite* if its elements can be ordered in a sequence.

Example 7. The natural numbers form a countably infinite set since we can order them in a sequence $1, 2, 3, \dots$

Example 8. The integers (positive and negative) $\dots - 3, -2, -1, 0, 1, 2, 3, \dots$ are also countably infinite because we can order them as

$$0, 1, -1, 2, -2, 3, -3, \dots$$

Example 9. The rational numbers (those that can be written as a ratio of two integers, say $1/2$ or $345/45$, or $-3/4$) are also countably infinite. Note that the rational numbers are dense in the sense that every interval (a, b), no matter how small or large contains a rational number. A priori, by looking at the real number line one cannot tell what is the next rational number after, say $1/2$. Their placement on the real number line does not show immediately how to order them in a sequence. Here is how one can order them in a sequence. We will do that only for the positive rational numbers for added simplicity.

	1	2	3	4	5	$\cdots$
1	1/1	1/2	1/3	1/4	1/5	$\cdots$
2	2/1	2/2	2/3	2/4	2/5	$\cdots$
3	3/1	3/2	3/3	3/4	3/5	$\cdots$
4	4/1	4/2	4/3	4/4	4/5	$\cdots$
5	5/1	5/2	5/3	5/4	5/5	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

We look at the diagonals in this table that go from north-east to south-west and list the numbers in them one diagonal after another:

$$1/1, 1/2, 2/1, 1/3, 2/2, 3/1, 1/4, 2/3, 3/2, 4/1, 1/5, 2/4, 3/3, 4/2, 5/1, \ldots$$

Many numbers in this sequence are repeated, for example $1/1 = 2/2 = 3/3 = 1$ or $1/2 = 2/4$. We delete all repetitions leaving only the first instance of a repeated number to obtain

$$1/1, 1/2, 2/1, 1/3, 3/1, 1/4, 2/3, 3/2, 4/1, 1/5, 5/1, \ldots$$

We obtained a sequence of all positive rational numbers, which is what we wanted.

Lemma 10. Union of countably many sets each one of which has countably many elements is countably infinite.

Proof. We have countably many sets, that is we can order them in a sequence $A_1, A_2, A_3, \ldots$ Each set A_i has countably many elements, say $A_1 = \{a_1, a_2, a_3, \ldots\}$, $A_2 = \{b_1, b_2, b_3, \ldots\}$, $A_3 = \{c_1, c_2, c_3, \ldots\}$, and so on. We need to show that we can order the elements of $\bigcup_{i=1}^{\infty} A_i$ in a sequence as well. We use an idea analogous to the one presented above. Place the elements of the sets A_i in rows one after another

	1	2	3	4	5	$\cdots$
A_1	a_1	a_2	a_3	a_4	a_5	$\cdots$
A_2	b_1	b_2	b_3	b_4	b_5	$\cdots$
A_3	c_1	c_2	c_3	c_4	c_5	$\cdots$
A_5	d_1	d_2	d_3	d_4	d_5	$\cdots$
A_6	e_1	e_2	e_3	e_4	e_5	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

We look at the diagonals in this table that go from north-east to south-west and list the numbers in them one diagonal after another:

$$a_1, a_2, b_1, a_3, b_2, c_1, a_4, b_3, c_2, d_1, \ldots$$

This sequence contains the all elements in the union $\bigcup_{i=1}^{\infty} A_i$. □

Definition 11. A random variable X defined on a sample space Ω is called *discrete* if X takes a finitely many or countably many different values. If the random variable X takes more than countably many different values, for example, if it takes any value in an interval (a, b), then X is called *continuous* random variable.

Note that if Ω is finite or countably infinite set, then any random variable X on Ω is necessarily discrete.

An argument very similar to the one in Example 9 shows the next property.

Proposition 12. *If X and Y are two discrete random variables, then so is $X + Y$.*

Proof. Let $\{x_1, x_2, \ldots\}$ be all the values that X takes. We can order them in a sequences (possibly finite one) since X is discrete random variable. Let $\{y_1, y_2, \ldots\}$ be all the values that Y takes. Then, the values that $X + Y$ takes are given in the interior of the following table

	x_1	x_2	x_3	x_4	x_5	$\cdots$
y_1	x_1+y_1	x_2+y_1	x_3+y_1	x_4+y_1	x_5+y_1	$\cdots$
y_2	x_1+y_2	x_2+y_2	x_3+y_2	x_4+y_2	x_5+y_2	$\cdots$
y_3	x_1+y_3	x_2+y_3	x_3+y_3	x_4+y_3	x_5+y_3	$\cdots$
y_4	x_1+y_4	x_2+y_4	x_3+y_4	x_4+y_4	x_5+y_4	$\cdots$
y_5	x_1+y_5	x_2+y_5	x_3+y_5	x_4+y_5	x_5+y_5	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

We already know how to order all values $x_i + y_j$ in a sequence. □

Example 13. The real numbers are infinitely many and are more than countable. That is, the real numbers cannot be ordered in a sequence. Let us see that the real numbers in the interval $(0,1)$ cannot be ordered in a sequence. The decimal representation of every real number in $(0,1)$ is of the form $0.a_1a_2a_3\ldots$. where the digits $a_1, a_2, a_3, \ldots$ in the decimal representation are integers between 0 and 9. Suppose the real numbers in $(0,1)$ can be listed in a sequence

$$\begin{aligned}&0.a_1a_2a_3a_4\ldots\\&0.b_1b_2b_3b_4\ldots\\&0.c_1c_2c_3c_4\ldots\\&0.d_1d_2d_3d_4\ldots\\&\vdots\end{aligned}$$

Consider a number $0.xyzt\ldots$ constructed in such a way that $x \neq a_1$, $y \neq b_2$, $z \neq c_3$, $t \neq d_4$, and so on. This number will be in the interval $(0,1)$ and will not be in the list since it differs from the first number in the list by its first digit after the decimal point; it differs from the second number in the list by its second digit, and so on. This contradiction shows that our assumption that the real numbers in $(0,1)$ can be ordered in a sequence is wrong.

4 Definition of probabilities

We are observing (or performing) an event (or experiment) with random outcomes and we defined Ω to be the set of all possible outcomes of the event. How should we assign probabilities (that is, positive numbers) so that we "model" the random outcomes of the event? The most natural way is to use proportions, as follows. Take an event $A \subset \Omega$ and perform the experiment many, many times, say N times and count the number of times the outcome was in the set A. The ratio

$$\frac{\text{the number of times the outcome is in the set } A}{N}$$

is aways going to be a positive number. In particular, if $A = \Omega$ then we have

$$\frac{\text{the number of times the outcome is in the set } \Omega}{N} = 1.$$

Moreover, if A and B are two disjoint events, then

$$\frac{\text{the number of times the outcome is in the set } A \cup B}{N}$$

$$
\begin{aligned}
&= \frac{\text{the number of times the outcome is in the set } A}{N} \\
&\quad + \frac{\text{the number of times the outcome is in the set } B}{N}.
\end{aligned}
$$

So, if we want our notion of probabilities to "model" the proportion of times an event occurs, it must have the above three properties. This naturally leads to the following definition.

Definition 14. A *probability measure* or *probability distribution* P on Ω is a function defined on 2^Ω taking real values that satisfies the following three properties

(i) $P(\Omega) = 1$;

(ii) For every event $A \subset \Omega$, $P(A) \geq 0$;

(iii) For every sequence $A_1, A_2, A_3, \ldots$ of disjoint events

$$P\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} P(A_i).$$

Clearly, there are potentially many functions P that satisfy the above three properties so there is much freedom in choosing a probability measure on Ω. Here is an example of a rather strange measure, called point measure. Fix any point $\omega \in \Omega$ and define $P(A) = 1$ if $\omega \in A$ and $P(A) = 0$ if $\omega \notin A$. Check that this is also a measure.

We will discuss other ways for choosing a measure that "makes sense" in this course.

4.1 Properties of probabilities

There are some properties of the probability function that follow immediately from the definition.

Property 1. $P(\emptyset) = 0$.

Proof. Let $A_1 := \emptyset$, $A_2 := \emptyset$, $A_3 := \emptyset$, $\ldots$ This is a disjoint sequence of events. Their union is $\emptyset$. So by the third property in Definition 14 we have

$$P(\emptyset) = P\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} P(A_i) = \sum_{i=1}^{\infty} P(\emptyset).$$

The only possible value of $P(\emptyset)$ that satisfies that equality is 0. □

Property 2. If $A_1, A_2, \ldots, A_n$ are disjoint then

$$P\Big(\bigcup_{i=1}^{n} A_i\Big) = \sum_{i=1}^{n} P(A_i).$$

Proof. Let $A_{n+1} := \emptyset$, $A_{n+2} := \emptyset$, $A_{n+3} := \emptyset$, ... The whole sequence of events $A_1, A_2, \ldots, A_n, A_{n+1}, A_{n+2}, \ldots$ is disjoint and $\bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{\infty} A_i$. Then

$$P\Big(\bigcup_{i=1}^{n} A_i\Big) = P\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} P(A_i) = \sum_{i=1}^{n} P(A_i),$$

where in the last equality we used that $P(\emptyset) = 0$. □

Property 3. $P(A^c) = 1 - P(A)$.

Proof. The events A and A^c are disjoint and $A \cup A^c = \Omega$. So

$$P(A) + P(A^c) = P(A \cup A^c) = P(\Omega) = 1.$$

□

Property 4. If $A \subset B$ then $P(A) \leq P(B)$.

Proof. Note that $B = A \cup (A^c \cap B)$ and that the events A and $A^c \cap B$ are disjoint. Then

$$P(B) = P(A) + P(A^c \cap B) \geq P(A),$$

where in the last inequality we used that $P(A^c \cap B) \geq 0$. □

Property 5. For every event $A \subset \Omega$, we have $0 \leq P(A) \leq 1$.

Proof. Since $\emptyset \subseteq A$, by Properties 1 and 4, we get

$$0 = P(\emptyset) \leq P(A).$$

Since $A \subseteq \Omega$, by Propertiy 4 and the definition of P, we get

$$P(A) \leq P(\Omega) = 1.$$

Combining the two inequalities completes the proof. □

Property 6. For any events $A, B \subset \Omega$, we have

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{1}$$

Proof. Note first that

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

and the sets in the above union: $(A \cap B^c)$, $(A \cap B)$, and $(A^c \cap B)$ are disjoint. Hence by the properties of the probability measure

$$P(A \cup B) = P(A \cap B^c) + P(A \cap B) + P(A^c \cap B). \tag{2}$$

Next, note that

$$A = (A \cap B^c) \cup (A \cap B) \text{ and}$$

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

Again, since the events $(A \cap B^c)$, $(A \cap B)$, and $(A^c \cap B)$ are disjoint we have

$$\begin{aligned} P(A) &= P(A \cap B^c) + P(A \cap B) \text{ and} \\ P(B) &= P(A \cap B) + P(A^c \cap B). \end{aligned}$$

Adding these last equalities together we obtain

$$P(A) + P(B) = P(A \cap B^c) + P(A \cap B) + P(A \cap B) + P(A^c \cap B) = P(A \cup B) + P(A \cap B),$$

where in the last equality we used (2). We are done. □

Formula (1) is called *inclusion-exclusion* formula for two events. There is a useful corollary of the last property.

Corollary 15. For any events $A, B \subset \Omega$, we have

$$P(A \cup B) \leq P(A) + P(B).$$

Proof. Since $P(A \cap B) \geq 0$, removing this term from the right-hand side of (1) gives the desired inequality. □

It is important to keep in mind that there is nothing in the properties of a probability function to prevent the existence of events A that are non-empty and $P(A) = 0$. In other words, the fact that $P(A) = 0$ does not imply that A is the empty set. Similarly, the fact that $P(A) = 1$ does not imply that $A = \Omega$. That is, there may be events $A \subset \Omega$ for which $P(A) = 1$. In this regard, we have the following properties.

Corollary 16. *a) If A, B are events with $P(A) = P(B) = 0$, then $P(A \cup B) = 0$.*
b) If A, B are events with $P(A) = P(B) = 1$, then $P(A \cap B) = 1$.

Proof. a) This follows from Corollary 15:

$$0 \leq P(A \cup B) \leq P(A) + P(B) = 0.$$

b) By Property 3, we have $P(A^c) = P(B^c) = 0$. Hence, by part a) we have $P(A^c \cup B^c) = 0$. But $A^c \cup B^c = (A \cap B)^c$, and by Property 3 again we get

$$1 - P(A \cap B) = P((A \cap B)^c) = 0.$$

We are done □

We now apply the properties of a probability function in a particular situation.

Example 17. *Suppose that $P(A) \geq 0.9$, $P(B) \geq 0.8$, and $P(A \cap B \cap C) = 0$. Show that $P(C) \leq 0.3$.*

Solution. There are several ways to solve this problem. Here is one. Apply (but kind of in reverse) the inclusion exclusion formula to the events $(A \cap B)$ and C

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

The left-hand side is zero since $(A \cap B) \cap C = A \cap B \cap C$. Next, we apply the inclusion exclusion formula to $P(A \cap B)$ again:

$$\begin{aligned} 0 &= P(A \cap B) + P(C) - P((A \cap B) \cup C) \\ &= P(A) + P(B) - P(A \cup B) + P(C) - P((A \cap B) \cup C). \end{aligned}$$

Solving for $P(C)$ gives

$$\begin{aligned} P(C) &= P(A \cup B) + P((A \cap B) \cup C) - P(A) - P(B) \\ &\leq P(A \cup B) + P((A \cap B) \cup C) - 0.9 - 0.8 \leq 1 + 1 - 0.9 - 0.8 \\ &= 0.3. \end{aligned}$$

□

Suppose X is a random variable on Ω, that is X is a function $X : \Omega \to \mathbb{R}$. We often use the following shorthand notation. Let $x \in \mathbb{R}$ be a fixed number.

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

That is, the probability that the random variable X takes value x is the probability of the event $\{\omega \in \Omega : X(\omega) = x\}$.

4.2 Probabilities on finite sample spaces

Suppose that $\Omega = \{\omega_1, \ldots, \omega_n\}$ is a sample space with n outcomes. What do we need to define a probability measure P on Ω? Every event A in Ω has finitely many elements so it can be expressed as the (finite) union of disjoint events each one of which has one elements. For instance, if $A = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ then

$$A = \{\omega_1\} \cup \{\omega_2\} \cup \{\omega_3\} \cup \{\omega_4\}.$$

By the rules of probability measures

$$P(A) = P(\{\omega_1\}) + P(\{\omega_2\}) + P(\{\omega_3\}) + P(\{\omega_4\}).$$

Thus, it is enough to define

$$p_i := P(\{\omega_i\}) \text{ for all } i = 1, 2, \ldots, n$$

and we will be able to compute the probabilities of all events in Ω. For example, with A defined above $P(A) = p_1 + p_2 + p_3 + p_4$. What properties should these numbers p_i satisfy? By Definition 14 they must satisfy

$$\begin{aligned} &p_i \geq 0 \text{ for all } i = 1, 2, \ldots, n \\ &\sum_{i=1}^{n} p_i = 1. \end{aligned}$$

Note that $\sum_{i=1}^{n} p_i = P(\Omega)$.

4.3 Probabilities on countably infinite sample spaces

Suppose that $\Omega = \{\omega_1, \omega_2, \omega_3, \ldots\}$ is a sample space with countably infinite outcomes. What do we need to define a probability measure P on Ω. Every event A in Ω has finitely many countably infinitely many elements so it can be expressed as the finite union or countable union of disjoint events each one of which has one elements. For instance, if $A = \{\omega_{i_1}, \omega_{i_2}, \omega_{i_3}, \ldots\}$ is a subset of Ω then

$$A = \{\omega_{i_1}\} \cup \{\omega_{i_2}\} \cup \{\omega_{i_3}\} \cup \cdots .$$

By the rules of probability measures

$$P(A) = P(\{\omega_{i_1}\}) + P(\{\omega_{i_2}\}) + P(\{\omega_{i_3}\}) + \cdots .$$

Thus, it is enough to define

$$p_i := P(\{\omega_i\}) \text{ for all } i = 1, 2, 3, \ldots,$$

and we will be able to compute the probabilities of all events in Ω. For example, with A defined above $P(A) = p_{i_1} + p_{i_2} + p_{i_3} + \cdots$. What properties should these numbers p_i satisfy? By Definition 14 they must satisfy

$$p_i \geq 0 \text{ for all } i = 1, 2, 3, \ldots,$$
$$\sum_{i=1}^{\infty} p_i = 1.$$

Note that $\sum_{i=1}^{\infty} p_i = P(\Omega)$.

4.4 Probabilities on uncountably infinite sample spaces

We cannot assign probabilities to *individual* elements of Ω if Ω has uncountably many number of elements, as we did in the finite and countable cases, as expect to get something sensible. To see why, take for example $\Omega = (0, 1)$. Suppose for every $x \in (0, 1)$ we assign a non-negative number $P(\{x\}) := p_x$. Now check if $P(\Omega) = 1$ as we did in the other two cases:

$$P(\Omega) = P(\cup_{x \in (0,1)} \{x\}) = \sum_{x \in (0,1)} P(\{x\}) = \sum_{x \in (0,1)} p_x.$$

There is no way to be sure that the last sum is 1, because the sum has uncountably many terms. Ordinary Calculus works with finite sums or countable sums (called infinite series). When we try to add uncountable many terms together many paradoxes may arise and we will avoid it.

Thus, to assign probabilities on uncountably infinite sample spaces we need fresh ideas. Two such ideas will be discussed later in these notes: 1) a way to assign probabilities on sample spaces of infinite sequences, and 2) assigning probabilities using density functions.

4.5 The inclusion-exclusion formula

We now turn our attention to the generalization of the identity

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

holding for any two events A and B. The *inclusion-exclusion* formula says that, in order to find the probability that at least one of n events A_i occurs, first add the probability of each event, then subtract the probabilities of all possible two-way intersections, add the probability of all three-way intersections, and so forth.

Theorem 18 (Inclusion-exclusion formula)**.** Let P be a probability measure on a sample space Ω. Then, for any n events $A_1, A_2, \ldots, A_n$ we have

$$\begin{aligned} P(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_{i=1}^{n} P(A_i) - \sum_{1 \le i < j \le n} P(A_i \cap A_j) + \sum_{1 \le i < j < k \le n} P(A_i \cap A_j \cap A_k) - \cdots \\ + (-1)^{n-1} P(A_1 \cap A_2 \cap \cdots \cap A_n) \end{aligned}$$

Proof. We prove this formula by induction. If we want to prove that a statement holds for all values of $n = 1, 2, 3, \ldots$, then we first prove it for $n = 1$, then if assuming that the statement holds for $n = m$ we manage to prove it for $n = m + 1$, we are done. The inclusion exclusion formula does hold trivially for $n = 1$ and even more, we have proved it before for $n = 2$. So assume now that it holds for $n = m$, that is the probability $P(A_1 \cup \cdots \cup A_m)$ is given by the stated formula. We use this to show the formula for $m + 1$ events:

$$\begin{aligned} P\Big(\bigcup_{i=1}^{m+1} A_i\Big) &= P\Big(\Big(\bigcup_{i=1}^{m} A_i\Big) \cup A_{m+1}\Big) \\ &= P\Big(\bigcup_{i=1}^{m} A_i\Big) + P(A_{m+1}) - P\Big(\Big(\bigcup_{i=1}^{m} A_i\Big) \cap A_{m+1}\Big) \\ &= P\Big(\bigcup_{i=1}^{m} A_i\Big) + P(A_{m+1}) - P\Big(\bigcup_{i=1}^{m} (A_i \cap A_{m+1})\Big) \end{aligned}$$

We write out the formula for the two m-element unions, using the induction hypothesis:

$$\begin{aligned} &= \Big(\sum_{i=1}^{m} P(A_i) - \sum_{1 \le i < j \le m} P(A_i \cap A_j) + \sum_{1 \le i < j < k \le m} P(A_i \cap A_j \cap A_k) - \cdots\Big) + P(A_{m+1}) \\ &- \Big(\sum_{i=1}^{m} P(A_i \cap A_{m+1}) - \sum_{1 \le i < j \le m} P(A_i \cap A_j \cap A_{m+1}) + \sum_{1 \le i < j < k \le m} P(A_i \cap A_j \cap A_k \cap A_{m+1}) - \cdots\Big) \\ &= \sum_{i=1}^{m+1} P(A_i) - \sum_{1 \le i < j \le m+1} P(A_i \cap A_j) + \sum_{1 \le i < j < k \le m+1} P(A_i \cap A_j \cap A_k) - \cdots \end{aligned}$$

The induction is complete. □

4.6 Examples

Example 19. John and Mary are taking a mathematics course. The course has only three grades: A, B, and C. The probability that John gets a B is 0.3. The probability that Mary gets a B is 0.4. The probability that neither gets an A but at least one gets a B is 0.1. What is the probability that at least one gets a B but neither gets a C?

Solution. Find the sample space first. John and Mary are taking the exam independently of each other, so each one of them may receive any of the grades A, B, or C. There are 9 possible outcomes, given in the table

$$\begin{aligned}\big\{\{A,A\}, \quad \{A,B\}, \quad \{A,C\},\\ \{B,A\}, \quad \{B,B\}, \quad \{B,C\},\\ \{C,A\}, \quad \{C,B\}, \quad \{C,C\}\big\}\end{aligned}$$

where the first grade is Johns and the second is Mary's. You are given that

$$\begin{aligned}P(\{\{B,A\},\{B,B\},\{B,C\}\}) &= 0.3\\ P(\{\{A,B\},\{B,B\},\{C,B\}\}) &= 0.4\\ P(\{\{B,B\},\{B,C\},\{C,B\}\}) &= 0.1.\end{aligned}$$

We have to find $P(\{\{A,B\},\{B,B\},\{B,A\}\}) =$? In detail, we have

$$\begin{aligned}P(\{\{A,B\},\{B,B\},\{B,A\}\}) &= P(\{A,B\}) + P(\{B,B\}) + P(\{B,A\})\\ &= \big(P(\{A,B\}) + P(\{B,B\}) + P(\{C,B\})\big) + \big(P(\{B,A\}) + P(\{B,B\}) + P(\{B,C\})\big)\\ &\qquad - \big(P(\{C,B\}) + P(\{B,B\}) + P(\{B,C\})\big)\\ &= P(\{\{B,A\},\{B,B\},\{B,C\}\}) + P(\{\{A,B\},\{B,B\},\{C,B\}\}) - P(\{\{B,B\},\{B,C\},\{C,B\}\})\\ &= 0.3 + 0.4 - 0.1 = 0.6.\end{aligned}$$

Note that even though we do not know the probability of some of the intermediate events, by grouping them together, we obtained events whose probability we know. □

Example 20. Consider the experiment that consists of rolling a pair of dice.
a) What is the sample space?
b) What is the probability measure?
c) What is the probability that the result is a six and a one?
d) What is the probability that the sum of the dice is x, where $x = 2, 3, \ldots, 12$?

Solution. a) The sample space is **not**

$$\begin{aligned}\big\{\{1,1\},&\{1,2\},\{1,3\},\{1,4\},\{1,5\},\{1,6\},\{2,2\},\{2,3\},\{2,4\},\{2,5\},\{2,6\},\\ &\{3,3\},\{3,4\},\{3,5\},\{3,6\},\{4,4\},\{4,5\},\{4,6\},\{5,5\},\{5,6\},\{6,6\}\big\}\end{aligned}$$

which is the set of all pairs of numbers between 1 and 6. (Having 15 elements.) Rather, the sample space is the set of all *ordered* pairs of numbers between 1, and 6. (Having 36 elements.) Huh?

What is the difference? The difference is that the sets $\{1,2\}$ and $\{2,1\}$ are the same, they both contain the same elements. But the ordered pairs $(1,2)$ and $(2,1)$ are not the same. (Think about the numbers 12 and 21, they are different. Numbers are ordered pairs of digits.) Thus, the sample space is

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

In math notation, the sample space is

$$\Omega = \{(i,j) : 1 \leq i,j \leq 6\}.$$

b) Since each value of i is equally likely and there should be no connection between the dice, it makes sense to assume that each pair is equally likely, that is,

$$p_{ij} := P((i,j)) := 1/36. \tag{3}$$

c) The event "the result is a six and a one" is $\{(1,6),(6,1)\}$, so

$$P(\{(1,6),(6,1)\}) = p_{16} + p_{61} = 2/36 = 1/18.$$

d) Let A be the event that the sum of the dice is x, where $x = 2, 3, \ldots, 12$, that is $A = \{(i,j) : i + j = x\}$. The easiest way to find this is to list all cases.

If $x = 2$, then $A = \{(1,1)\}$ and $P(A) = 1/36$.
If $x = 3$, then $A = \{(1,2),(2,1)\}$ and $P(A) = 2/36$.
If $x = 4$, then $A = \{(1,3),(2,2),(3,1)\}$ and $P(A) = 3/36$.
If $x = 5$, then $A = \{(1,4),(2,3),(3,2),(4,1)\}$ and $P(A) = 4/36$.
If $x = 6$, then $A = \{(1,5),(2,4),(3,3),(4,2),(5,1)\}$ and $P(A) = 5/36$.
If $x = 7$, then $A = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$ and $P(A) = 6/36$.
If $x = 8$, then $A = \{(2,6),(3,5),(4,4),(5,3),(6,2)\}$ and $P(A) = 5/36$.
If $x = 9$, then $A = \{(3,6),(4,5),(5,4),(6,3)\}$ and $P(A) = 4/36$.
If $x = 10$, then $A = \{(4,6),(5,5),(6,4)\}$ and $P(A) = 3/36$.
If $x = 11$, then $A = \{(5,6),(6,5)\}$ and $P(A) = 2/36$.
If $x = 12$, then $A = \{(6,6)\}$ and $P(A) = 1/36$. □

If you play a betting game trying to guess the sum of two dice, then you should always bet that the sum will be 7 since then the probability of being right is highest.

Example 21. In Example 20 one may consider another probability measure on the sample space. Suppose that the two dice are entangled just like elementary particles may be entangled in Quantum Mechanics. This means that if the first die lands on x the second die will also land on the same number x. In that case the probability measure (3) will be

$$p_{ij} := P((i,j)) = \begin{cases} 1/6 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

In this case, each die is fair (the probability of each die landing on x is $1/6$) but the pair is not.

We now have some of the tools needed to accurately describe sample spaces and to assign probability functions to those sample spaces. Nevertheless, in some cases, the description and assignment process is somewhat arbitrary. Of course, it is to be hoped that the description of the sample space and the subsequent assignment of a probability function will yield a model which accurately predicts what would happen if the experiment were actually carried out. As the following examples show, there are situations in which "reasonable" descriptions of the sample space do not produce a model which fits the data.

In Feller's book,[3] a pair of models is given which describe arrangements of certain kinds of elementary particles, such as photons and protons. It turns out experiments have shown that certain types of elementary particles exhibit behavior which is accurately described by one model, called *"Bose-Einstein statistics,"* while other types of elementary particles can be modelled using *"Fermi-Dirac statistics."* Feller says:

> We have here an instructive example of the impossibility of selecting or justifying probability models by a priori arguments. In fact, no pure reasoning could tell that photons and protons would not obey the same probability laws.

We now give some examples of this description and assignment process.

Example 22. In the quantum mechanical model of the helium atom, various parameters can be used to classify the energy states of the atom. In the triplet spin state ($S = 1$) with orbital angular momentum 1 ($L = 1$), there are three possibilities, 0, 1, or 2, for the total angular momentum (J). (It is not assumed that the reader knows what any of this means; in fact, the example is more illustrative if the reader does not know anything about quantum mechanics.) We would like to assign probabilities to the three possibilities for J. The reader is undoubtedly resisting the idea of assigning the probability of 1/3 to each of these outcomes. She should now ask herself why she is resisting this assignment. The answer is probably because she does not have any "intuition" (i.e., experience) about the way in which helium atoms behave. In fact, in this example, the probabilities 1/9, 3/9, and 5/9 are assigned by the theory. The theory gives these assignments because these frequencies were observed *in experiments* and further parameters were developed in the theory to allow these frequencies to be predicted.

5 Combinatorics

There are four main counting techniques that we will cover. In increasing order of difficulty, they are

- Counting ordered objects when we choose with replacement;
- Counting ordered objects when we choose without replacement;
- Counting unordered objects when we choose without replacement;
- Counting unordered objects when we choose with replacement.

[3] W. Feller, *Introduction to Probability Theory and Its Applications* vol. 1, 3rd ed. (New York: John Wiley and Sons, 1968), p. 41

5.1 Counting ordered objects when we choose with replacement

Suppose there are n different objects and we need to make r selections. After we select an object we record it and return it in, so that it can be selected again. The order in which the objects appear is important. In how many ways can these r selections be made? Each time we select, there are n possibilities, thus there are

$$n^r$$

ways to make the selections in this case.

Example 23. How many six letter strings can be formed using the letters A, B, C, D, E, with repetitions permitted. For example ABBACC counts just as well as ACCEDE, even though the former is not an English word.

Solution. We must choose which letter to put in each of six different positions; thus $5^6 = 15,625$ answers the question. □

Example 24. Show that there are at least two people in Columbus, Ohio, who have the same three initials.

Solution. Assuming that each person has three initials, there are 26 possibilities for a person's 1-st initial, 26 for the second, and 26 for the third. Therefore, there are $26^3 = 17,576$ possible sets of initials. This number is smaller than the number of people living in Columbus, Ohio; hence, there must be at least two people with the same three initials. □

A slightly more general case is the following. Consider an experiment that takes place in r stages and is such that the number of outcomes n_i at the i-th stage is independent of the outcomes of the previous stages. We want to count the number of ways that the entire experiment can be carried out. There are n_1 ways to carry out the 1st stage; for each of these n_1 ways, there are n_2 ways to carry out the second stage; for each of these n_2 ways, there are n_3 ways to carry out the third stage, and so forth. Then the total number of ways in which the entire task can be accomplished is given by the product

$$n_1 n_2 \cdots n_r.$$

Example 25. You are eating at Emile's restaurant and the waiter informs you that you have (a) two choices for appetizers: soup or juice; (b) three for the main course: a meat, fish, or vegetable dish; and (c) two for dessert: ice cream or cake. How many possible choices do you have for your complete meal?

It will often be useful to use a tree diagram when studying probabilities of events relating to experiments that take place in stages and for which we are given the probabilities for the outcomes at each stage. Suppose that at every outcome of every stage, we know what the probabilities are for the different outcomes of the next stage to occur. How should we find the probabilities of the final outcomes at the end of the r-th stage?

Example 26. Assume that the owner of Emile's restaurant has observed that 80 percent of his customers choose the soup for an appetizer and 20 percent choose juice. Of those who choose soup, 50 percent choose meat, 30 percent choose fish, and 20 percent choose the vegetable dish. Of those who choose juice for an appetizer, 30 percent choose meat, 40 percent choose fish, and 30 percent choose the vegetable dish. Find the probabilities that a random customer will choose a particular two course meal.

Let $\Omega = \{\omega_1, \dots, \omega_n\}$ be a finite sample space. A probability measure P on Ω is said to be uniform if $P(\{\omega_k\}) = 1/n$ for all $k = 1, 2, \dots, n$. In that case, calculating the probability of an event A reduced to counting the number of elements in A.

$$P(A) = P\big(\cup_{\omega \in A} \{\omega\}\big) = \sum_{\omega \in A} P(\{\omega\}) = \sum_{\omega \in A} \frac{1}{n} = \frac{|A|}{n},$$

where $|A|$ denotes the number of elements of A.

In Example 26, choosing a customer at "random" means that every one has equal probability of being selected. Thus, we can try to work with the uniform measure on the set of all customers. Unfortunately, we do not know how many customers there are, that is, we do not know n. Instead, we are given 6 disjoint events: the sets of customers who ordered soup-meat, soup-fish, soup-veggies, juice-meat, juice-fish, and juice-veggies and the proportion of customers in each event. With this information, we can think that the sample space is made up of these 6 elements having respective probabilities: 0.8×0.5, 0.8×0.3, 0.8×0.2, 0.2×0.5, 0.2×0.3, and 0.2×0.2.

Definition 27 (Uniform distribution)**.** Let n be a positive integer. Let X be the random variable taking n possible values $\{1, 2, \dots, n\}$. We say that X has a *uniform distribution* if it satisfies $P(X = k) = 1/n$ for all $k = 1, 2, \dots, n$.

Note that calculating probabilities involving a uniform random variable reduces to counting. Indeed, if $A \subset \{1, 2, \dots, n\}$, then

$$P(X \in A) = P\big(\cup_{a \in A} \{X = a\}\big) = \sum_{a \in A} P(X = a) = \sum_{a \in A} \frac{1}{n} = \frac{|A|}{n},$$

where $|A|$ denotes the number of elements of A. That is, all we have to do to find $P(X \in A)$ is count the elements of A and then divide them by n.

Note that in order to work with the random variable X we do not really have to know the sample space Ω on which it is defined. In fact, it could be finite or infinite. All that we need and know are the probabilities of the events $\{X \in A\} \subset \Omega$ for a subset A of the range of X. (In Definition 27, the range of X is the set $\{1, 2, \dots, n\}$.)

5.2 Counting ordered objects when we choose without replacement

Definition 28. Let A be any finite set. A *permutation* of A is a one-to-one mapping of A onto itself.

To specify a particular permutation we list the elements of A and, under them, show where each element is sent by the one-to-one mapping. For example, if $A = \{a, b, c\}$ then a possible permutation would be

$$\sigma := \begin{pmatrix} a & b & c \\ b & c & a \end{pmatrix}.$$

By the permutation σ, a is sent to b, b is sent to c, and c is sent to a. The condition that the mapping be one-to-one means that no two elements of A are sent, by the mapping, into the same element of A. We can put the elements of our set in some order and rename them 1, 2,. . .,n. Then, a typical permutation of the set $A = \{a_1, a_2, a_3, a_4\}$ can be written in the form

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix},$$

indicating that a_1 went to a_2, a_2 to a_1, a_3 to a_4, and a_4 to a_3. If we always choose the top row to be 1, 2, 3, 4 then, to prescribe the permutation, we need only give the bottom row, with the understanding that this tells us where 1 goes, 2 goes, and so forth, under the mapping. When this is done, the permutation is often called a rearrangement of the n objects 1, 2, 3, . . ., n. For example, all possible permutations, or rearrangements, of the numbers $A = \{1, 2, 3\}$ are:

$$123, 132, 213, 231, 312, 321.$$

It is an easy matter to count the number of possible permutations of n objects. By our general counting principle, there are n ways to assign the 1-st element, for each of these we have $n-1$ ways to assign the second object, $n-2$ for the third, and so forth. This proves the following theorem.

Theorem 29. The total number of permutations of a set A of n elements is given by $n! := n(n-1)(n-2)\cdots 2\cdot 1$.

The expression 0! is defined to be 1 to make certain formulas come out simpler.

Definition 30. *Let A be an n-element set, and let r be an integer between 0 and n. Then a r-permutation of A is an ordered listing of a subset of A of size r.*

Using the same techniques as in the last theorem, the following result is easily proved. The total number of r-permutations of a set A of n elements is given by

$$n(n-1)(n-2)\cdots(n-r+1).$$

If we multiply and divide the last expression by $(n-r)!$ we see that

$$n(n-1)(n-2)\cdots(n-r+1) = \frac{n!}{(n-r)!}.$$

Note that when $r = n$, an n-permutation becomes just a permutation.

Example 31. Five people of a group of 100 are to receive one of five prizes, each prize having a distinctive name. In how many ways can the five recipients be selected from the 100 candidates?

Solution. We are not told the names of the prizes, but they might just be ”First Prize,” ”Second Prize,” etc. It is thus clear that the problem of choosing in order five people to receive five identical prizes is equivalent to the problem of choosing five people to receive five different prizes. Thus our problem amounts to choosing $r = 5$ out of $n = 100$ people. The order counts since the prizes are different and are ordered as first, second, and so on. The selection is without replacement because one person may receive only one prize. The answer is thus 100!/95! = 9,034,502,400. □

5.3 Counting unordered objects when we choose without replacement

Let A be a set with n elements. We want to count the number of distinct subsets of the set A that have exactly r elements. A set is an unordered object, that is, the order of its elements does not matter. (Here r can be $0, 1, \ldots$ up to n.) The number of distinct subsets with r elements that can be chosen from a set with n elements is denoted by $\binom{n}{r}$ and is pronounced "n choose r." The number $\binom{n}{r}$ is called a *binomial coefficient*. This terminology comes from an application to algebra which will be discussed later.

Example 32. Let $A = \{a, b, c\}$ then the different subsets of A are

$$\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}.$$

Hence $\binom{3}{0} = 1$, $\binom{3}{1} = 3$, $\binom{3}{2} = 3$, and $\binom{3}{3} = 1$.

Clearly, we always have

$$\binom{n}{0} = \binom{n}{n} = 1,$$

because there is exactly one subset of every set with 0 elements, namely the empty set, and exactly one subset with n elements, namely the set itself.

Theorem 33. The binomial coeffcients are given by the formula

$$\binom{n}{r} = \frac{n(n-1)\cdots(n-r+1)}{r!}. \tag{4}$$

Proof. Each subset of size r of a set of size n can be ordered in $r!$ ways. Each of these orderings is a r-permutation of the set of size n. The number of r-permutations is $n(n-1)\cdots(n-r+1)$, so the number of subsets of size r is

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

□

The above formula can be written as

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}, \tag{5}$$

which shows immediately that

$$\binom{n}{r} = \binom{n}{n-r}. \tag{6}$$

Another point that should be made concerning formula (4) is that if it is used to define the binomial coeffcients, then it is no longer necessary to require n to be a positive integer. The variable r must still be a non-negative integer under this definition. This idea is useful when extending the Binomial Theorem to general exponents (more on that later). The Binomial Theorem for non-negative integer exponents is given shortly.

Theorem 34. For integers n and r with $0 < r < n$, the binomial coefficients satisfy

$$\binom{n}{r} = \binom{n-1}{r} + \binom{n-1}{r-1}.$$

Proof. A mindless proof of this identity is to utilize formula (5):

$$\begin{aligned}\binom{n-1}{r} + \binom{n-1}{r-1} &= \frac{(n-1)!}{r!(n-r-1)!} + \frac{(n-1)!}{(r-1)!(n-r)!} = \frac{(n-r)(n-1)! + r(n-1)!}{r!(n-r)!} \\ &= \frac{n(n-1)!}{r!(n-r)!} = \binom{n}{r}.\end{aligned}$$

□

The last identity, together with the fact that $\binom{n}{0} = \binom{n}{n} = 1$, allows us to compute the binomial coefficients inductively by the so-called *Pascal triangle*.

$$\begin{array}{ccccccccccc}
 & & & & & \binom{0}{0} & & & & & \\
 & & & & \binom{1}{0} & & \binom{1}{1} & & & & \\
 & & & \binom{2}{0} & & \binom{2}{1} & & \binom{2}{2} & & & \\
 & & \binom{3}{0} & & \binom{3}{1} & & \binom{3}{2} & & \binom{3}{3} & & \\
 & \binom{4}{0} & & \binom{4}{1} & & \binom{4}{2} & & \binom{4}{3} & & \binom{4}{4} & \\
\binom{5}{0} & & \binom{5}{1} & & \binom{5}{2} & & \binom{5}{3} & & \binom{5}{4} & & \binom{5}{5} \\
\vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots
\end{array}$$

$$\begin{array}{ccccccccccc}
 & & & & & 1 & & & & & \\
 & & & & 1 & & 1 & & & & \\
 & & & 1 & & 2 & & 1 & & & \\
 & & 1 & & 3 & & 3 & & 1 & & \\
 & 1 & & 4 & & 6 & & 4 & & 1 & \\
1 & & 5 & & 10 & & 10 & & 5 & & 1 \\
\vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots
\end{array}$$

Theorem 35 (Binomial theorem)**.** The quantity $(a+b)^n$ can be expressed in the form

$$(a+b)^n = \sum_{r=0}^{n} \binom{n}{r} a^r b^{n-r}. \tag{7}$$

Proof. To see that this expansion is correct, write

$$(a+b)^n = (a+b)(a+b)\cdots(a+b).$$

When we multiply this out we will have a sum of terms each of which results from a choice of an a or b for each of n factors. When we choose r a's and $(n-r)$ b's, we obtain a term of the form $a^r b^{n-r}$. To determine such a term, we have to specify r of the n terms in the product from which we choose the a. This can be done in $\binom{n}{r}$ ways. Thus, collecting these terms in the sum contributes a term $\binom{n}{r} a^r b^{n-r}$. □

The following example gives a useful alternative view.

Example 36. In how many ways can you order n letters A and m letters B in a sequence?

Solution. At first it appears that this question asks us to count ordered objects, that is, sequences, but first looks may be deceiving. A sequence of n letters A and m letters B is $n+m$ letters long. All that we have to do is choose where to place the letters A. That uniquely defines the sequence since at the rest of the places we place the B's. That is, we need to select n positions out of the $n+m$ positions. So the answer is

$$\binom{n+m}{n}.$$

Incidentally, this is equal to $\binom{n+m}{m}$ — the number of sequences we would have gotten if we were choosing the places of the letters B first. (See Formula (6).) □

Poker players sometimes wonder why a *four of a kind* beats a *full house.* A poker hand is a random subset of 5 elements from a deck of 52 cards. A hand has four of a kind if it has four cards with the same value—for example, four sixes or four kings. It is a full house if it has three of one value and two of a second—for example, three twos and two queens. Let us see which hand is more likely.

Example 37. Compute the probability of obtaining four of a kind and full house in a random poker hand.

Solution. Assuming that each poker hand is equally likely to occur, we work with the uniform measure on the set of all poker hands. There are $\binom{52}{5}$ pocket hands. How many hands have four of a kind? There are 13 ways that we can specify the value for the four cards. For each of these, there are 48 possibilities for the 5-th card. Thus, the number of four-of-a-kind hands is $13 \cdot 48 = 624$. Since the total number of possible hands is $\binom{52}{5} = 2598960$, the probability of a hand with four of a kind is $624/2598960 = 0.00024$.

Now consider the case of a full house. How many such hands are there? There are 13 choices for the value which occurs three times; for each of these there are $\binom{4}{3} = 4$ choices for the particular three cards of this value that are in the hand. Having picked these three cards, there are 12 possibilities for the value which occurs twice; for each of these there are $\binom{4}{2} = 6$ possibilities for the particular pair of this value. Thus, the number of full houses is $13 \cdot 4 \cdot 12 \cdot 6 = 3744$, and the probability of obtaining a hand with a full house is $3744/2598960 = 0.0014$. Thus, while both types of hands are unlikely, you are six times more likely to obtain a full house than four of a kind. □

5.3.1 The multinomial theorem

Now, we generalize the arguments just presented. Choosing r objects out of n, without replacement, amounts to dividing the objects into two categories — those that are chosen and those that are not. In exactly the same way we can consider dividing the objects into more than two categories. We do retain the condition that the number of objects in each category be determined in advance.

Suppose that there are r categories, and that we plan to put n_1 objects into the first category, n_2 into the second, and so on. If we can do this at all, we must have

$$n_1 + n_2 + \cdots + n_r = n,$$

since each object is placed in just one category. For example, if $n = 3$ and $r = 2$ with $n_1 = 2$ and $n_1 = 1$, then there are three possible ways for dividing the three objects A, B, and C into two groups with two objects in the first and one object in the second group:

Group 1	Group 2
$\{A, B\}$,	$\{C\}$;
$\{A, C\}$,	$\{B\}$;
$\{B, C\}$,	$\{A\}$.

Example 38. Let $n_1, n_2, \ldots, n_r$ be r non-negative integers such that $n_1 + n_2 + \cdots + n_r = n$. A set of n distinct objects is to be divided into r distinct groups or sizes n_1, n_2, …, n_r respectively. How many different ways are there to do that?

Solution: There are $\binom{n}{n_1}$ ways to put n_1 objects in the first group. For each choice for the first group, there are $\binom{n-n_1}{n_2}$ possible ways to put n_2 (of the remaining objects) objects in the second group. For each choice of the first two groups, there are $\binom{n-n_1-n_2}{n_3}$ possible choices for the third group and so on. Hence there are

$$\begin{aligned}
&\binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdots\binom{n-n_1-n_2-\cdots-n_{r-1}}{n_r} \\
&= \frac{n!}{n_1!(n-n_1)!}\frac{(n-n_1)!}{n_2!(n-n_1-n_2)!}\frac{(n-n_1-n_2)!}{n_3!(n-n_1-n_2-n_3)!}\cdots\frac{(n-n_1-n_2-\cdots-n_{r-1})!}{n_r!(n-n_1-n_2-\cdots-n_{r-1}-n_r)!} \\
&= \frac{n!}{n_1!n_2!\cdots n_r!},
\end{aligned}$$

possible ways to do the selection. □

Define the notation

$$\binom{n}{n_1, n_2, \ldots, n_r} := \frac{n!}{n_1!n_2!\cdots n_r!}.$$

These numbers are called *multinomial coefficients* because of their use in the following general formula

Theorem 39 (Multinomial theorem)**.** For any numbers $a_1, a_2, \ldots, a_r$ we have

$$(a_1 + a_2 + \cdots + a_r)^n = \sum_{\substack{(n_1,n_2,\ldots,n_r) \text{ s.t.} \\ n_1+n_2+\cdots+n_r=n \\ n_1\geq 0,\ldots,n_r\geq 0}} \binom{n}{n_1, n_2, \ldots, n_r} a_1^{n_1}a_2^{n_2}\cdots a_r^{n_r}. \tag{8}$$

The sum is over all non-negative integer vectors $(n_1, n_2, \ldots, n_r)$ such that $n_1 + n_2 + \cdots + n_r = n$.

We are not going to prove the Multinomial theorem, but every educated person should know it. Note that when $r = 2$, Theorem 39 reduces to the binomial theorem because the multinomial coefficients reduce to the binomial coefficients. Indeed, if $n_1 + n_2 = n$ then

$$\binom{n}{n_1, n_2} = \frac{n!}{n_1!n_2!} = \frac{n!}{n_1!(n-n_1)!} = \binom{n}{n_1},$$

and $a_1^{n_1}a_2^{n_2} = a_1^{n_1}a_2^{n-n_1}$.

Another particular case that arises often is $n = 2$. In that case the multinomial formula reduces to (why?)

$$(9) \qquad (a_1 + a_2 + \cdots + a_r)^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j = \sum_{i=1}^{r} a_i^2 + \sum_{i=1}^{r}\sum_{\substack{j=1 \\ j\neq i}}^{r} a_i a_j = \sum_{i=1}^{r} a_i^2 + 2\sum_{i<j} a_i a_j,$$

where the last sum is over all pairs of integers (i, j), each between 1 and r, such that $i < j$.

A very useful alternative view is given by the next example.

Example 40. In how many ways can the letters of the word $MISSISSIPPI$ be arranged?

Solution. Clearly the order of the letters in a “word” is important, so it seems that the question asks you to count ordered objects, but that is deceiving. There are 4 different letters: M — repeated once; I — repeated 4 times; S — repeated 4 times; and P — repeated 2 times. There are a total of 11 letters. The different letters are going to be our groups, that is $r = 4$. The positions of the 11 letters are going to be our objects, that is $n = 11$. Next, we let $n_1 := 1$, $n_2 := 4$, $n_3 := 4$, and $n_4 := 2$. Thus, we have

$$n_1 + n_2 + n_3 + n_4 = 11.$$

An arrangement of the letters, can be viewed as assigning the positions (the objects) to the letters (the groups). For example, the arrangement $MISSISSIPPI$ assigned (position) 1 to the first group (M); assigned (positions) 2, 5, 8, and 11 to the second group (I); assigned (positions) 3, 4, 6, and 7 to the third group (S) and finally positions 9 and 10 to the fourth group (P).

With this interpretation, we see that there are

$$\frac{n!}{n_1!n_2!n_3!n_4!} = \frac{11!}{1!4!4!2!} = 34,650$$

ways to arrange the letters. □

Notice, finally, that Example 40 is a generalization of Example 36. Indeen, Example 40 asks us to count the number of ways in which we can place one M, four I’s, four S’s, and two P’s in a sequence.

5.4 Counting unordered objects when we choose with replacement

Suppose there are r objects, from which we want to select n times. Every time we select an object, we record it, and return it back, so that it could be selected again on the next selection. How many different selections are there if we only care about how many times each object was selected? (That is, we do not care about the order in which the objects are selected.) Let us look at an example.

Example 41. In how many ways can 14 chocolate bars be distributed among five children? To see how this fits into the current situation, think that each time we hand out one of the 14 bars, we choose one of the five children to receive it. Thus, it this case, $r = 5$ and $n = 14$.

On a sheet of paper draw r bins. This can be achieved by drawing $r - 1$ vertical lines. The first bin is to the left of the left-most vertical line, the second bin is between the first and the second vertical line, and so on. The r-th bin is to the right of the right-most vertical line. Next, we start selecting. Every time an object is selected, we put a star in the bin corresponding to that object. For convenience, we may order the stars on a horizontal row. When we finish with the selections, we have stars and bars arranged in a row.

For example, say $r = 5$ and $n = 12$, then the arrangement

$$\ast\ast \mid \ast\ast\ast\ast || \ast\ast\ast \mid \ast\ast\ast$$

indicates that the first object was selected 2 times, the second — 4 times, the third — 0 times, and the fourth — 3 times, and the fifth — 3 times.

Each way of making n selections from r objects, with replacement, disregarding the order of selection, corresponds to just one such arrangement of the stars and bars. Conversely, each such arrangement corresponds to a certain selection. Thus, the number we seek is simply the number of ways to put the stars and bars in order. As noted above, there are $r - 1$ bars. Since there is a star for each selection, there are n stars. The number of ways of arranging $r - 1$ bars and n stars in a row was discussed in Example 36. It is

$$\binom{n+r-1}{r-1} = \binom{n+r-1}{n}.$$

Here is a variation of the above argument.

Example 42. Find the number of different integer vectors $(n_1, n_2, \ldots, n_r)$ satisfying

$$n_1 + n_2 + \cdots + n_r = n \text{ and } n_i > 0 \text{ for all } i = 1, 2, \ldots, r.$$

Solution. Think of the integer number n as the sum of n ones: $n = \underbrace{1 + 1 + \cdots + 1}_{n \text{ times}}$. We want to find in how many ways we can divide the last sum into r non-empty groups. For example, if $n = 5$ and $r = 3$, there are the following possibilities

$$1|1|111 \quad 1|11|11 \quad 1|111|1 \quad 11|1|11 \quad 11|11|1 \quad 111|1|1$$

corresponding to the solutions

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

For $n = 5$ and $r = 3$ these are all positive integer solutions of $n_1 + n_2 + n_3 = 5$. In general, from the example above, we see that we have to select the positions of $r - 1$ dividers among the $n - 1$ spaces between adjacent 1's. There are

$$\binom{n-1}{r-1}$$

ways to do that. □

Yet another way to phrase the last example is: find the number of ways to distribute n indistinguishable from each other objects into r different non-empty boxes.

In the binomial expansion (7) the sum is for j from 0 to n. That is, there are $n+1$ terms in the sum. A natural question to ask is to find the number of terms in the sum (8). Here is the answer.

Example 43. Find the number of different integer vectors $(n_1, n_2, \dots, n_r)$ satisfying

$$n_1 + n_2 + \cdots + n_r = n \text{ and } n_i \geq 0 \text{ for all } i = 1, 2, \dots, r.$$

We give two different solutions since each one shows a different side of the problem.

Solution 1. Think of the integer number n as the sum of n ones: $n = \underbrace{1 + 1 + \cdots + 1}_{n \text{ times}}$. We want to find in how many ways we can divide the last sum into r possibly empty groups. For example, if $n = 3$ and $r = 3$, there are the following possibilities

$$||111 \quad |111| \quad 111|| \quad |1|11 \quad |11|1 \quad 1||11 \quad 1|11| \quad 11|1| \quad 11||1 \quad 1|1|1$$

corresponding to the solutions

$$(0,0,3) \quad (0,3,0) \quad (3,0,0) \quad (0,1,2) \quad (0,2,1) \quad (1,0,2) \quad (1,2,0) \quad (2,1,0) \quad (2,0,1) \quad (1,1,1).$$

As before, we have $r-1$ dividers, and n ones. But this time the dividers may appear next to each other. (That corresponds to a 0 in the solution.) So, the question asks, in how many ways we can order in a sequence $r-1$ bars and n ones. The answer was given in Example 36. It is

$$\binom{n+r-1}{r-1}.$$

Solution 2. From every non-negative solution of $n_1+n_2+\cdots+n_r = n$ we can obtain a positive solution of $m_1 + m_2 + \cdots + m_r = n + r$ by defining $m_i = n_i + 1$ for $i = 1, 2, \dots, r$. Conversely, from every positive solution of $m_1 + m_2 + \cdots + m_r = n + r$ we obtain a non-negative solution of $n_1+n_2+\cdots+n_r = n$ by defining $n_i := m_i - 1$ for $i = 1, 2, \dots, r$. This correspondence is one-to-one. So, the number of non-negative solutions of $n_1 + n_2 + \cdots + n_r = n$ is the same as the number of positive solutions of $m_1 + m_2 + \cdots + m_r = n + r$. The number of the latter is given by Example 42 after replacing n by $n + r$. It is

$$\binom{n+r-1}{r-1}.$$

This concludes the example. □

Another way to phrase the last example is: the number of ways to distribute n indistinguishable from each other objects into r different boxes (some of them possibly empty).

5.5 Binomial distribution

Definition 44 (Bernoulli trials)**.** A *Bernoulli trials* process is a sequence of n chance experiments such that

(i) Each experiment has two possible outcomes, which we may call success and failure.

(ii) The probability p of success on each experiment is the same for each experiment, and this probability is not affected by any knowledge of previous outcomes. The probability q of failure is given by $q = 1 - p$.

Example 45. The following are Bernoulli trials processes:

(i) A coin is tossed ten times. The two possible outcomes are heads and tails. The probability of heads on any one toss is 1/2.

(ii) An opinion poll is carried out by asking 1000 people, randomly chosen from the population, if they favour the Equal Rights Amendment—the two outcomes being yes and no. The probability p of a yes answer (i.e., a success) indicates the proportion of people in the entire population that favour this amendment.

(iii) A gambler makes a sequence of 1-dollar bets, betting each time on black at roulette at Las Vegas. Here a success is winning 1 dollar and a failure is losing 1 dollar. Since in American roulette the gambler wins if the ball stops on one of 18 out of 38 positions and loses otherwise, the probability of winning is $p = 18/38 = 0.474$.

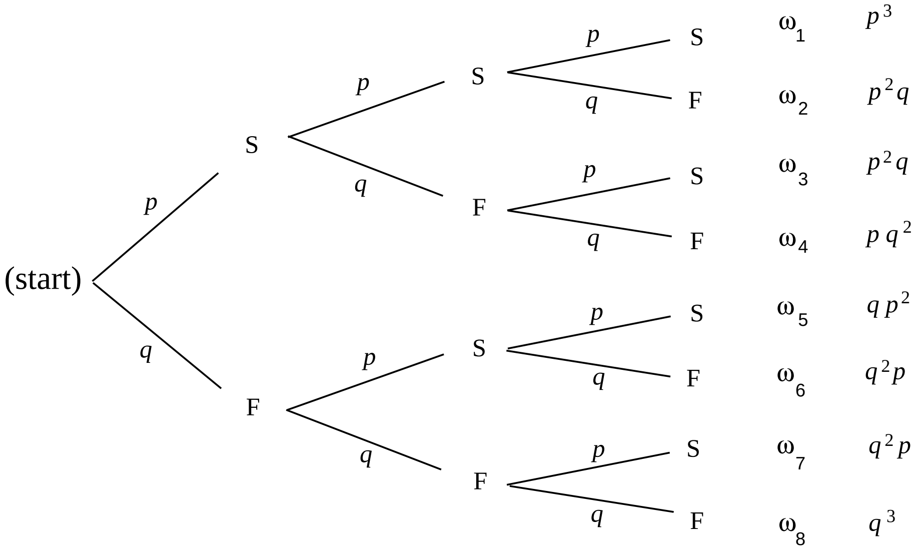


Figure 1: Tree diagram of three Bernoulli trials

To analyze a Bernoulli trials process, we choose as our sample space a binary tree and assign a probability distribution to the paths in this tree. Suppose, for example, that we have three Bernoulli trials. The possible outcomes are indicated in the tree diagram shown in Figure 1. The outcomes of the process are all possible ordered triple of S's and F's. For example, ω_3 represents the outcomes SFS. The probabilities assigned to the branches of the tree represent the probability for each individual trial. Since we have assumed that outcomes on any one trial do not affect those on another, we assign the same probabilities at each level of the tree. Our frequency interpretation of probability would lead us to expect a fraction p of successes on the first experiment; of these, a fraction q of failures on the second; and, of these, a fraction p of successes on the third experiment. This suggests assigning probability pqp to the outcome ω_3. And so on.

We shall be particularly interested in the probability that in n Bernoulli trials there are exactly j successes. We denote this probability by $b(n, p, j)$. Let us calculate the particular value $b(3, p, 2)$ from our tree measure. We see that there are three paths which have exactly two successes and one failure, namely ω_2, ω_3, and ω_5. Each of these paths has the same probability p^2q. Thus $b(3, p, 2) = 3p^2q$. Considering all possible numbers of successes we have

$$\begin{aligned}
b(3, p, 0) &= q^3, \\
b(3, p, 1) &= 3pq^2, \\
b(3, p, 2) &= 3p^2q, \\
b(3, p, 3) &= p^3.
\end{aligned}$$

We can, in the same manner, carry out a tree measure for n experiments and determine $b(n, p, r)$ for the general case of n Bernoulli trials.

Theorem 46. Given n Bernoulli trials with probability p of success on each experiment, the probability of exactly r successes is

$$b(n, p, r) = \binom{n}{r} p^r q^{n-r}, \tag{10}$$

where $q = 1 - p$.

Proof. We construct a probability measure on the all outcomes of a n-stage binary tree as described above. We want to find the sum of the probabilities for all sequences of n S's and F's which have exactly r successes and $n - r$ failures. Each such path is assigned a probability $p^r q^{n-r}$. How many such paths are there? To specify a path, we have to pick, from the n possible trials, a subset of r to be successes, with the remaining $n - r$ outcomes being failures. We can do this in $\binom{n}{r}$ ways. Thus the sum of the probabilities is given by the right-hand side of (10). □

Example 47. A fair coin is tossed six times. What is the probability that exactly three heads turn up?

Solution. The answer is

$$b(6, 0.5, 3) = \binom{6}{3} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^3 = \frac{20}{64} = 0.3125.$$

Example 48. A die is rolled four times. What is the probability that we obtain exactly one 6?

Solution. We treat this as Bernoulli trials with success = "rolling a 6" and failure = "rolling some number other than a 6." Then $p = 1/6$, and the probability of exactly one success in four trials is

$$b(4, 1/6, 1) = \binom{4}{1}\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^3 = 0.386.$$

Definition 49 (Binomial distribution)**.** Let n be a positive integer, and let p be a real number between 0 and 1. Let X be the random variable taking $n+1$ possible values $\{0, 1, 2, \ldots, n\}$. We say that X has a binomial distribution if it satisfies $P(X = r) = b(n, p, r)$.

The reason why this distribution is called binomial is due to its close relationship with the binomial theorem. A simple corollary of the binomial theorem we obtain that the probabilities of the binomial distribution sum up to one. Indeed, let n be a positive integer and p be a real number between 0 and 1, and let $q := 1 - p$, then

$$\sum_{r=0}^{n} b(n, p, i) = \sum_{r=0}^{n} \binom{n}{r} p^r q^{n-r} = (p+q)^n = 1^n = 1.$$

In addition, we also have the following interesting relationships:

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

and

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

5.6 More examples

We learned four basic counting techniques. Some examples below will use them. The difficulty usually lies in the fact that a problem often uses a mixture of these techniques. Sometimes different mixtures lead to the correct solution. Seeing in what order to apply the four basic counting techniques, and applying them properly is an art. To become better at it, you may want to try the problems after Chapter 2 in [1], all problems there have answers at the end of the text.

Example 50. In how many ways can you place the 3 identical kangaroos in 7 adjacent cells so that each kangaroo is alone in a cell, and no 2 kangaroos are neighbours? See the following figure.

Solution. There are 7 cells. So, after placing the 3 kangaroos, there will be 4 cells left and we want these 4 cells to divide the kangaroos. The three kangaroos, define four "bins" around them we have to place the cells. But we want the second and third bin to be non-empty. That is, we are looking for the number of integer solutions of

$$n_1 + n_2 + n_3 + n_4 = 4 \text{ where } n_1 \geq 0, n_2 > 0, n_3 > 0, n_4 \geq 0.$$

The solutions of this system are in one-to-one correspondence with the solutions of the system

$$m_1 + m_2 + m_3 + m_4 = 6 \text{ where } m_1 > 0, m_2 > 0, m_3 > 0, m_4 > 0.$$

(If you start with a solution of the first system, just add 1 to n_1 and n_4 to get a solution of the second system. Conversely, if you start with a solution of the second system, just subtract 1 from m_1 and m_4 to get a solution of the first system. See Solution 2 of Example 43). Now, Example 42 tells us that there are $\binom{5}{3} = 10$ solutions of the second system. □

Example 51. In how many ways can you place the kangaroos Annie, Betty, and Clara in 7 adjacent adjacent cells so that each kangaroo is alone in a cell, and no 2 kangaroos are neighbours?

Solution. Now the kangaroos are different, so placing them $\_A\_B\_C\_$ is different than placing them $\_B\_C\_A\_$. First, there are 3! ways to order (to permute) the three kangaroos. Once we have fixed the order, then we start counting in how many different ways we can separate them. The argument from the previous example applies to each of the 6 permutations of the kangaroos. So, the answer to this problem is 6 × 10 = 60.

Example 52. In a restaurant, six people ordered roast beef, three ordered turkey, two ordered pork chops, and one ordered vegetarian. The 12 servings are brought from the kitchen and distributed randomly among the people.

a) What is the probability that everyone gets the correct order?

b) What is the probability that no one gets the correct order?

Solution. The phrase "the 12 servings are brought from the kitchen and distributed randomly among the people" means that all 12! permutations of the dishes are equally likely. That is, each permutation of the dishes occurs with probability 1/12!.

a) We have to count, how many are the permutations in which every person gets their order. The six beef dishes, can be delivered to the six people who ordered them in 6! possible ways. For every such correct delivery, there are 3! ways in which the people who ordered turkey get turkey. For each such correct delivery of the turkey, there are 2! ways in which the pork is delivered correctly, and then only one way for the vegetarian. Thus, there are 6!3!2!1! permutations that deliver the dishes correctly. Hence the probability that everyone gets the correct order is

$$\frac{6!3!2!1!}{12!} = \frac{8,640}{479,001,600} = 0.0000180375.$$

Incidentally, the answer can be written as $1/\binom{12}{6,3,2,1}$.

b) We have to count, how many are the permutations in which nobody gets their order. Here are the dishes

$$B_1, B_2, B_3, B_4, B_5, B_6, T_1, T_2, T_3, P_1, P_2, V,$$

where B stands for a beef dish, T stands for a turkey dish, and V stands for the vegetarian. In order for the beef guys not to obtain their dish, they must receive the other six dishes, which may be permuted among them in 6! possible ways. This leaves us with six beef dishes to be given to the six turkey, pork, and vegetarian clients of the restaurant. The beef dishes can be delivered to the six non-beef guys in another 6! ways. Thus, there are 6!6! permutations in which nobody gets their order. The answer is

$$\frac{6!6!}{12!} = \frac{518,400}{479,001,600} = 0.00108225.$$

In this example, we were a bit lucky that the beef guys were exactly equal to the non-beef guys. The problem would be more difficult if we had, say 5 beef guys and 2 vegetarians, with everything else kept the same. □

Example 53 (The game of craps)**.** The game of craps is played by trowing two fair dice. The game has two phases: "come-out" and "point". In the "come-out" phase the player rolls the dice and a roll with sum 2, 3 or 12 is called "craps" or "crapping out", and the player loses. A "come-out" roll of 7 or 11 is a "natural", and the player wins. The other possible throws in the "come-out" phase are those with sums: 4, 5, 6, 8, 9, and 10. If the shooter rolls one of these numbers on the "come-out" roll, this establishes the "point" in the next phase. To win in the "point" phase, the "point" number must be rolled again before a 7. What is the probability of a win?

Solution. Let us call the two phases phase one and phase two. To win one needs to win either in phase one or in phase two. Denote these events by W_1 and W_2. So the event of winning is $W := W_1 \cup W_2$ with the union being disjoint, hence $P(W) = P(W_1) + P(W_2)$. First, we have

$$P(W_1) = P(\text{throw sum 7 or 11}) = \frac{6}{36} + \frac{2}{36} = \frac{2}{9}.$$

To win in the second phase, one needs to throw a number $N \in \{4, 5, 6, 8, 9, 10\}$ in the first phase and then in the second phase throw N again before 7. The event W_2 can occur only if we throw 4 in phase one and then throw 4 before 7 in phase two; or if we throw 5 in phase one and then throw 5 before 7 in phase two; and so on. Denote these events by $W_4, W_5, W_6, W_8, W_9, W_{10}$. Thus, we have the disjoint union

$$W_2 = W_4 \cup W_5 \cup W_6 \cup W_8 \cup W_9 \cup W_{10}.$$

We compute the probabilities of these six events.

$$\begin{aligned} P(W_4) &= P(N = 4 \text{ and then throw 4 before 7}) = P(N = 4)P(\text{throw 4 before 7}) \\ &= \frac{3}{36}\frac{3/36}{3/36 + 6/36} = \frac{1}{36}. \end{aligned}$$

For the last equality, we used Example 20 that the probabilities to throw $4, 5, 6, 7, 8, 9, 10$ are $3/36, 4/36, 5/36, 6/36, 5/36, 4/36, 3/36$, respectively, and we used the result in Example 59. Analogously, we compute

$$\begin{aligned} P(W_5) &= \frac{4}{36}\frac{4/36}{4/36 + 6/36} = \frac{2}{45}; \\ P(W_6) &= \frac{5}{36}\frac{5/36}{5/36 + 6/36} = \frac{25}{396}; \\ P(W_8) &= \frac{5}{36}\frac{5/36}{5/36 + 6/36} = \frac{25}{396}; \end{aligned}$$

$$P(W_9) = \frac{4}{36}\frac{4/36}{4/36+6/36} = \frac{2}{45};$$
$$P(W_{10}) = \frac{3}{36}\frac{3/36}{3/36+6/36} = \frac{1}{36}.$$

Thus

$$P(W) = \frac{2}{9} + \frac{1}{36} + \frac{2}{45} + \frac{25}{396} + \frac{25}{396} + \frac{2}{45} + \frac{1}{36} = \frac{244}{495} = 0.49292.$$

Example 54 (The Birthday Problem)**.** There are r random people in a room. What is the probability that two of them will have the same birthday?

Solution. Assume that there are 365 possible birthdays for each person (we ignore leap years). Order the people from 1 to r. The fact that we select people at random is interpreted to mean that a person's birthday could be on any of the 356 days of the year with equal probability. That is the probability that the first person is born on Jan 1, is 1/365. The sample space consists of all possible sequences of length r of birthdays each chosen as one of the 365 possible dates. There are 365 possibilities for the 1-st element of the sequence, and for each of these choices there are 365 for the second, and so forth, making 365^r possible sequences of birthdays. The fact that we select r people at random is interpreted to mean that the probability of any sequence of birthdays of length r occurring in the room is $1/365^r$. Note that we have

$$P(\text{two people will have the same birthday}) = 1 - P(\text{they all have different birthdays}).$$

It turns out that it is easier to calculate the second probability. We need to answer the following question. How many sequences of length r are there with distinct integers from 1 to 365? For such a sequence, we can choose any of the 365 days for the 1-st element, then any of the remaining 364 for the second, 363 for the third, and so forth, until we make r choices. For the r-th choice, there will be $365 - r + 1$ possibilities. Hence, the total number of sequences with no duplications is

$$365 \cdot 364 \cdot 363 \cdots (365 - r + 1).$$

Hence

$$P(\text{they all have different birthdays}) = \frac{365 \cdot 364 \cdot 363 \cdots (365 - r + 1)}{365^r} = \frac{365!}{365^r(365-r)!},$$

or

$$P(\text{two people will have the same birthday}) = 1 - \frac{365!}{365^r(365-r)!}.$$

You may compute this formula with Maple to get values for different r's:

r	P
21	0.4436883351
22	0.4756953073
23	0.5072972342
24	0.5383442578
30	0.7063162426
40	0.8912318098
50	0.9703735796
60	0.9941226609

At $r = 23$ the probability becomes more than 0.5 and at 60 people it is almost certain that two people will have the same birthday. □

In the birthday problem we assumed that all possible sequences of r birthdays have the same probability. That is, the sequences of length r have uniform probability measure on them.

In the birthday problem, we have assumed that birthdays are equally likely to fall on any particular day. Statistical evidence suggests that this is not true. However, it is intuitively clear (but not easy to prove) that this makes it even more likely to have a duplication with a group of r people.

The birthday problem becomes easier if we slightly modify it by requiring that exactly one pair of people have the same birthday and the rest have different birthdays.

Example 55. There are r random people in a room. What is the probability that exactly two of them will have the same birthday?

Solution. We need to count all sequences of length r of numbers from between 1 and 365 having exactly one pair of equal elements. We can choose the pair having the same birthday in $r(r-1)/2$ ways (why?), and once the pair is fixed their birthday may be on any of the 365 days of the year. The next person in the sequence has to be born on a different day and that can happen in 364 ways, for the next person, there are 363 possible remaining days, and so on for the last person there will be $(365 - r + 2)$ possible days. Hence the number of our sequences is

$$\frac{r(r-1)}{2} 365 \cdot 364 \cdots (365 - r + 2) = \frac{r(r-1)}{2} \frac{365!}{(365 - r + 1)!}.$$

Since a sequence of length r is chosen at random with equal probability, the probability of a sequence to be selected is $1/365^r$. Thus, the answer to the problem is

$$\frac{r(r-1)}{2} \frac{365!}{365^r (365 - r + 1)!}.$$

There are many interesting problems that relate to properties of a permutation chosen at random from the set of all permutations of a given finite set. For example, since a permutation is a one-to-one mapping of the set onto itself, it is interesting to ask how many points are mapped onto themselves. We call such points *fixed points* of the mapping.

Example 56 (The fixed-point problem)**.** Find the probability that a random permutation does not contain a fixed point.

More picturesque versions of the fixed-point problem are:

1) In a restaurant n hats are checked and they are hopelessly scrambled; what is the probability that no one gets his own hat back? (The hat check problem.)

2) You have arranged the books on your book shelf in alphabetical order by author and they get returned to your shelf at random; what is the probability that exactly k of the books end up in their correct position? (The library problem.)

Solution. Recall that a permutation is a one-to-one map of a set $A = \{a_1, a_2, \ldots a_n\}$ onto itself. Let A_i be the event that the i-th element a_i remains fixed under this map. We are going to calculate

$$P(A_1 \cup A_2 \cup \cdots \cup A_n).$$

using the inclusion-exclusion formula. The probability that a random permutation does not contain a fixed point is

$$1 - P(A_1 \cup A_2 \cup \cdots \cup A_n).$$

If we require that a_i is fixed, then the map of the remaining $(n-1)$ elements provides an arbitrary permutation of $(n-1)$ objects. Since there are $(n-1)!$ such permutations, $P(A_i) = (n-1)!/n! = 1/n$. Note that this probability does not depend on i. In the same way, to have a particular pair (a_i, a_j) fixed, we can choose any permutation of the remaining $(n-2)$ elements; there are $(n-2)!$ such choices and thus

$$P(A_i \cap A_j) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)}.$$

Note that this probability does not depend on i or j. Similarly, for any three events A_i, A_j, and A_k, we have

$$P(A_i \cap A_j \cap A_k) = \frac{(n-3)!}{n!} = \frac{1}{n(n-1)(n-2)}.$$

And so on. Now we substitute these probabilities in the inclusion-exclusion formula.

$$\begin{aligned}
P(A_1 \cup A_2 \cup \cdots \cup A_n) &= \sum_{i=1}^{n} \frac{1}{n} - \sum_{1 \le i < j \le n} \frac{1}{n(n-1)} + \sum_{1 \le i < j < k \le n} \frac{1}{n(n-1)(n-2)} - \cdots \\
&\quad + (-1)^{n-1} \frac{1}{n(n-1)(n-2)\cdots 2 \cdot 1} \\
&= \binom{n}{1}\frac{1}{n} - \binom{n}{2}\frac{1}{n(n-1)} + \binom{n}{3}\frac{1}{n(n-1)(n-2)} - \cdots \\
&\quad + (-1)^{n-1}\binom{n}{n}\frac{1}{n(n-1)(n-2)\cdots 2 \cdot 1} \\
&= 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + (-1)^{n-1}\frac{1}{n!}.
\end{aligned}$$

This is the probability that a randomly chosen permutation has at least one fixed point. Thus the probability that a randomly chosen permutation has **no** fixed points is

$$P(\text{no fixed points}) = 1 - P(\text{at least one fixed point}) = \frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^n \frac{1}{n!}.$$

Recall from Calculus that

$$e^x = \frac{x^0}{0!} + \frac{x^1}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} + \cdots .$$

Thus, when $x := -1$ we get

$$e^{-1} = \frac{1}{2!} - \frac{1}{3!} + \cdots + \frac{(-1)^n}{n!} + \cdots \approx 0.3678794.$$

Therefore, the probability that there is no fixed point, i.e., that none of the n people gets his own hat back, is equal to the sum of the first n terms in the expression for e^{-1}. This series converges very fast. After $n = 9$ the probability is essentially equal to e^{-1} to six significant figures. □

5.6.1 Sample spaces made of sequences

Example 57. Suppose you flip a fair coin once after another for ever. (The probability of obtaining heads, or tails, on any flip is 1/2.)
a) What is the probability that you flip the first head on the N-th try?
b) What is the probability of getting a head on or before the k-th flip?
c) What is the probability that you flip a head on an even trial?

Solution. Flipping a coin one is modelled by the sample space $\{H, T\}$ with probability measure $P(H) = P(T) = 1/2$. But now our *big* experiment consists of flipping a coin once after another for ever. Thus, the result of the experiment is a sequence of H's and T's. Hence, the set of all possible outcomes, call it Ω^*, of the experiment is the set of all possible sequences of H's and T's. This sample space is uncountably infinite, so we cannot assign probability measure to it by assigning a probability to each of its elements, see Subsection 4.4. We are going to describe now how to assign probabilities to certain events in Ω^* that is intuitive and, most importantly correct. (This correctness claim has to be taken for granted in this course.) The events of Ω^* are sets of sequences.

What is the natural probability to assign to the set of all sequences that start with an H? Imagine that 1000 people flip a coin in a sequence for ever. Right after the first flip about 500 of them will flip an H and the rest a T, what ever happens after that is not important for our event. Thus, we assign probability 1/2 to this event.

What is the natural probability to assign to the set of all sequences that have an H on the 47-th position? Imagine that 1000 people flip a coin in a sequence for ever. What they flip on the first try is not important for the event, nor is what they flip on the second flip and so on. But on the 47-th flip about a half of the people will flip an H and the rest a T, and after that it is not important again what will happen. Thus, we assign probability 1/2 to this event as well.

What is the natural probability to assign to the set of all sequences that have an H on the 2-nd position and a T on the 4-th position? Imagine that 1000 people flip a coin in a sequence for ever. What they flip on the first try is not important for the event and about 500 of the people will flip an H on the second attempt. The third flip is not important, while on the 4-th attempt about a half of those 500 will flip a T. After that it is not important again what will happen. Thus, we assign probability $(1/2) \cdot (1/2) = 1/4$ to this event.

And so on.

a) What is the probability that you flip the first head on the N-th try? This means that we are interested in the event, call it E_N made up of all sequences that have a T on positions $1, 2, 3, \dots, N-1$ and an H on position N. Following the logic from above, we have imposed N conditions for the outcome of the these N flips. So, the natural probability of this event is $1/2^N$.

b) What is the probability of getting an H on or before the k-th flip? This means that we are interested in the event made up of all sequences that have an H *for the first time* on position 1, or 2, or $\dots$ k. But this event is the disjoint union $\cup_{i=1}^k E_i$, so

$$P(\cup_{i=1}^k E_i) = \sum_{i=1}^k P(E_i) = 1/2 + 1/4 + \cdots + 1/2^k = (2^k - 1)/2^k.$$

c) What is the probability that you flip an H on an even trial? This means that we are interested in the event made up of all sequences that have an H *for the first time* on position 2, or

4, or … $2k$ and so on. But this event is the disjoint union $\cup_{i=1}^{\infty} E_{2i}$, so

$$P(\cup_{i=1}^{\infty} E_{2i}) = \sum_{i=1}^{\infty} P(E_{2i}) = 1/4 + 1/16 + 1/64 + \cdots$$
$$= 1/4 \sum_{i=0}^{\infty} (1/4)^i = (1/4)/(1-1/4) = 1/3.$$

The solution is complete. □

Exercise 58. Suppose you flip an unfair coin once after another for ever. (The probability of obtaining a head on any flip is 1/3.)
a) What is the probability that you flip the first head on the N-th try?
b) What is the probability of getting a head on or before the k-th flip?
c) What is the probability that you flip a head on an even trial?

Example 59. An experiment has outcomes $\Omega = \{\omega_1, \omega_2, \ldots\}$ occurring with probabilities $P(\omega_i) = p_i$. Perform the experiment repeatedly. What is the probability that ω_i occurs before ω_j, where $i \neq j$.

Solution. We have a little experiment with outcomes $\Omega = \{\omega_1, \omega_2, \ldots\}$ occurring with probabilities $P(\omega_i) = p_i$, but we repeat it endlessly, one experiment after another. So, we end up with a *big* experiment having an outcome that is a sequence of ω's. The sample space of this big experiment is the set of all possible sequences of ω's. (Make sure you read Example 57 before you proceed.)

We are interested in the event, call it E, of all sequences in which ω_i occurs before ω_j. But the first time ω_i occurs in a sequence could be on position 1, or 2, or any. Let E_n be the event consisting of all sequences in which ω_i occurs *for the first time* on the n-th position and ω_j does not occur on any of the previous $(n-1)$ positions. Note that the events $E_1, E_2, \ldots$ are disjoint (that is there is no sequence of ω's that is in say E_1 and E_2) and

$$E = \bigcup_{i=1}^{\infty} E_i.$$

The probability that neither ω_i nor ω_j occurs in one little experiment is $1 - p_i - p_j$. Thus, E_n is the event consisting of all sequences in which neither ω_i nor ω_j occur on positions $1, 2, \ldots, n-1$, and ω_i occurs on position n. Thus,

$$P(E_n) = (1 - p_i - p_j)^{n-1} p_i.$$

Thus, we finally have

$$P(E) = P\Big(\bigcup_{n=1}^{\infty} E_n\Big) = \sum_{n=1}^{\infty} P(E_n) = \sum_{n=1}^{\infty} (1 - p_i - p_j)^{n-1} p_i$$
$$= p_i \sum_{n=1}^{\infty} (1 - p_i - p_j)^{n-1} = p_i \frac{1}{1 - (1 - p_i - p_j)} = \frac{p_i}{p_i + p_j}.$$

Example 60 (Problem of the points)**.** Perform a sequence of binomial trials, resulting in a success with probability p and a failure with probability $1-p$. What is the probability that r successes occur before m failures?

Solution. There are only two outcomes at each trial: S and F but we perform a sequence of such trials resulting is a sequence of S's and F's. So, we are dealing with a *big* experiment having an outcome that is a sequence of S's and F's. The sample space of this experiment is the set of all possible sequences of S's and F's.

We are interested in the event, call it E, of all sequences in which r S's occur before m F's.

Now, we claim that r successes occur before m failures, if and only if there are at least r successes in the first $r+m-1$ trials.

Indeed, if there are at least r successes in the first $r+m-1$ trials, there could be at most $m-1$ failures in those $r+m-1$ trials. Thus, r successes occur before m failures.

Suppose now that r successes occur before m failures, and mark the position where the r-th success occurs. That is, there are already r successes so far and at most $m-1$ failures. Say, there are exactly $k \leq m-1$ failures. In the remaining trials $r+k+1, r+k+2, \ldots, r+m-1$ there may be more successes. Hence in the first $r+m-1$ trials there will be at least r successes.

Thus, our event E breaks down into a disjoint union of events, $E_k, k = r, r+1, \ldots, r+m-1$, where E_k is the event consisting of all sequences with exactly k successes in the first $r+m-1$ trials. The rest of the trials in the sequence do not really matter now. Thus, $P(E_k) = b(r+m-1, p, k)$, and we have

$$P(E) = P\Big(\bigcup_{k=r}^{r+m-1} E_k\Big) = \sum_{k=r}^{r+m-1} P(E_k) = \sum_{k=r}^{r+m-1} b(r+m-1, p, k)$$
$$= \sum_{k=r}^{r+m-1} \binom{r+m-1}{k} p^k (1-p)^{r+m-1-k}.$$

The last example occupies an important place in the history of probability theory. Imagine two players put up stakes and play some game, with the stakes to go to the winner of the game. An interruption requires them to stop before either has won, and when each has some sort of a "partial score." How should the stakes be divided? Suppose when the game was interrupted John needed r more points to win it, while Mary needed m more points to win it. John wins each point with probability p, and Mary wins each point with probability $1-p$. Then, the probability that John wins the game is $P_{r,m} := \sum_{k=r}^{r+m-1} b(r+m-1, p, k)$. Hence, when the game was interrupted, it would be fair for John to receive a proportion of the stakes equal to $P_{r,m}$ and the rest of the stakes to go to Mary.

6 Conditional probability and independence

6.1 Conditional probability

Suppose that we toss two fair dice. There are 36 possible outcomes each one with probability of occurring 1/36. Suppose that we observe that the first die is a 4. Given this information, what is the probability that the sum of the two dice is equal to 9? We reason as follows. Given that the initial die is a 4, it follows that there can be at most 6 possible outcomes of our experiment, namely, $(4,1)$, $(4,2)$, $(4,3)$, $(4,4)$, $(4,5)$, and $(4,6)$. Since each of these outcomes originally had the same probability of occurring, the outcomes should still have equal probabilities. That is, given that the first die is a 4, the (conditional) probability of each of the outcomes $(4,1)$, $(4,2)$, $(4,3)$, $(4,4)$, $(4,5)$, and $(4,6)$ is $1/6$, whereas the (conditional) probability of the other 30 points in the sample space is now 0. Hence the desired probability will be $1/6$. If we let A and B denote, respectively, the event that the sum of the dice is 9 and the event that the first die is a 4, then the probability just obtained is called the conditional probability that A occurs given that B has occurred and is denoted by

$$P(A|B).$$

Suppose now that A and B are given events in a sample space Ω with probability measure P. Suppose we run an experiment and the event B occurs. Then, in order for A to occur it is necessary that the outcome ω be a point in both A and B. That is, it must be in $A \cap B$. Now, as we know that B has occurred, it follows that B becomes our new or "reduced" sample space; hence the probability that the event $A \cap B$ occurs will equal the probability of $A \cap B$ *relative* to the probability of B.

Definition 61. For any two events A and B, with $P(B) \neq 0$, we define the *conditional probability of A given B* by

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

We read $P(A|B)$ as "the probability of A, given B." We can rearrange the definition to give the so-called *multiplication rule*:

$$P(A \cap B) = P(A|B)P(B).$$

The multiplication rule says that in order to find the probability of both events $A \cap B$, we need to find the probability of B and then the probability of A given B. The multiplication rule also holds when $P(B) = 0$, since that implies that $P(A \cap B) = 0$ as well.

Example 62. An experiment consists of rolling a fair die once. Let A be the event that the outcome is 6, and let B be the event that the outcome is strictly bigger than 4. Suppose that the die is rolled and we are told that the event B has occurred. This leaves only two possible outcomes: 5 and 6. In the absence of any other information, we would still regard these outcomes to be equally likely, so the probability of A becomes $1/2$, making $P(A|B) = 1/2$. Alternatively, the formula for the conditional probability gives us

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{1/3} = 1/2.$$

Example 63. Mary is undecided as to whether to take a French course or a chemistry course. She estimates that her probability of receiving an A grade would be $1/2$ in a French course, and $2/3$ in a chemistry course. If Mary decides to base her decision on the flip of a fair coin, what is the probability that she gets an A in chemistry?

Solution. Let A be the event that Mary takes chemistry and let B be the event that she receives an A. Then

$$P(A \cap B) = P(A)P(B|A) = \frac{1}{2}\frac{2}{3} = \frac{1}{3}.$$

Conditional probabilities behave just like probabilities. In fact $P(\cdot|B)$ is a probability measure on the set B. Compare the following proposition with the three properties in Definition 14.

Proposition 64. The conditional probability $P(\cdot|B)$ satisfies the following properties.

(i) $P(B|B) = 1$;

(ii) For every event $A \subset B$, $P(A|B) \geq 0$;

(iii) For every sequence $A_1, A_2, A_3, \ldots$ of disjoint events

$$P\Big(\bigcup_{i=1}^{\infty} A_i \Big| B\Big) = \sum_{i=1}^{\infty} P(A_i|B).$$

Prove the proposition as an exercise. For the third property use the fact that $\Big(\bigcup_{i=1}^{\infty} A_i\Big) \cap B = \bigcup_{i=1}^{\infty}(A_i \cap B)$.

Hence, and this is very important, conditional probabilities satisfy the properties of probabilities, given in Section 4.1. For example, for any three events A, B, C we have

$$P(A \cup B|C) \leq P(A|C) + P(B|C)$$

but if A and B are disjoint, then there is equality above.

The multiplication rule can be generalized to n events.

Theorem 65 (Multiplication rule)**.** For any events $A_1, A_2, \ldots, A_n$, we have

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}).$$

Proof. Apply the definition of conditional probability to the following identity

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)\frac{P(A_2 \cap A_1)}{P(A_1)}\frac{P(A_3 \cap A_1 \cap A_2)}{P(A_1 \cap A_2)} \cdots \frac{P(A_n \cap A_1 \cap A_2 \cap \cdots \cap A_{n-1})}{P(A_1 \cap A_2 \cap \cdots \cap A_{n-1})}.$$

□

6.2 Independence

The previous examples show that $P(A|B)$ is not generally equal to $P(A)$. In other words, knowing that B has occurred generally changes the chances that A occurs. In the special cases where $P(A|B)$ is equal to $P(A)$, we say that A is independent of B. Two events A and B are *independent* if $P(A|B) = P(A)$. Otherwise, they are called *dependent.* Note that if A and B are independent then

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

and so $P(A \cap B) = P(A)P(B)$. Often the last equality is used as a definition of independence to avoid worrying about the case when $P(B) = 0$.

Definition 66. Two events A and B are *independent* if $P(A \cap B) = P(A)P(B)$. If the last equality does not hold then A and B are *dependent*.

Note that if A and B are independent, then, $P(B) = \frac{P(A \cap B)}{P(A)} = \frac{P(B \cap A)}{P(A)} = P(B|A)$.

Example 67. We toss two fair dice. Let B denote the event that the first die equals 4.
a) Let A denote the event that the sum of the dice is 6.
b) Let A be the event that the sum of the dice is 7.
Are A and B independent events?

Solution. a) On the one hand we have $P(A \cap B) = P(\{(4, 2)\}) = 1/36$, while on the other hand $P(A)P(B) = (5/36)(1/6) = 5/216$. So the events A and B are dependent.
b) On the one hand we have $P(A \cap B) = P(\{(4, 3)\}) = 1/36$, while on the other hand $P(A)P(B) = (6/36)(1/6) = (1/36)$. So the events A and B are independent. □

Lemma 68. If A and B are independent then so are A and B^c.

Proof. Suppose A and B are independent. Since $A = (A \cap B) \cup (A \cap B^c)$ is a disjoint union, we have

$$P(A) = P(A \cap B) + P(A \cap B^c) = P(A)P(B) + P(A \cap B^c).$$

Solving the last equality for $P(A \cap B^c)$, we get

$$P(A \cap B^c) = P(A)(1 - P(B)) = P(A)P(B^c).$$

This shows that A and B^c are independent. □

Corollary 69. If A and B are independent then so are
a) A and B^c;
b) A^c and B;
c) A^c and B^c.

Example 70. Two fair dice are thrown. Let A denote the event that the sum of the dice is 7. Let B denote the event that the first die equals 4 and let C be the event that the second die equals 3. From Example 67 we know that A is independent of B, and the same reasoning shows that A is independent of C. But surprisingly, we have that A is not independent of $B \cap C$, since $P(A|B \cap C) = 1$ which is not equal to $P(A) = 1/6$.

When are more than two events independent? Intuitively, we want information about any of them occurring not to change the probability that the rest occur.

Definition 71. The events $A_1, A_2, \dots, A_n$ are *independent* if

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_m}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_m}),$$

for any subset of events $\{A_{i_1}, A_{i_2}, \dots, A_{i_m}\}$ of the events $\{A_1, A_2, \dots, A_n\}$.

For example, three events A, B, and C are independent if

$$\begin{aligned} P(A \cap B \cap C) &= P(A)P(B)P(C); \\ P(A \cap B) &= P(A)P(B); \\ P(A \cap C) &= P(A)P(C); \text{ and} \\ P(B \cap C) &= P(B)P(C). \end{aligned}$$

The definition says that if the events $A_1, A_2, \ldots, A_n$ are independent then any number of those events are also independent.

The following example describes three events A, B, and C such that A and B are independent; A and C are independent; and B and C are independent; **but** the three events $A, B,$ and C are **not** independent. The problem is that information about $B \cap C$ occurring changes the probability of A. This example is of great importance and explains why, when we have more than 2 events, assuming pair-wise independence is not enough to conclude that they **all** are independent.

Example 72. Consider three fair coins that are tossed and the events

A := the second and the third coin show the same outcome;
B := the first and the third coin show the same outcome;
C := the first and the second coin show the same outcome;

Then, we have

$$\begin{aligned} P(A) &= P(\{\text{HHH, HTT, THH, TTT}\}) = \frac{1}{2}, \\ P(B) &= P(\{\text{HHH, THT, HTH, TTT}\}) = \frac{1}{2}, \\ P(C) &= P(\{\text{HHH, TTH, HHT, TTT}\}) = \frac{1}{2}. \end{aligned}$$

On the other hand, we have

$$\begin{aligned} P(A \cap B) &= P(\{\text{HHH, TTT}\}) = \frac{1}{4} = P(A)P(B); \\ P(A \cap C) &= P(\{\text{HHH, TTT}\}) = \frac{1}{4} = P(A)P(C); \\ P(B \cap C) &= P(\{\text{HHH, TTT}\}) = \frac{1}{4} = P(B)P(C); \end{aligned}$$

showing that any pair of events is independent. But now note that

$$P(A|B \cap C) = \frac{P(A \cap B \cap C)}{P(B \cap C)} = \frac{P(\{\text{HHH, TTT}\})}{P(\{\text{HHH, TTT}\})} = 1 \neq \frac{1}{2} = P(A).$$

Thus, information that B and C occurred, changes the probability of A. In that sense we cannot consider the three events $A, B,$ and C to be independent together as a triple. Note that we do not have equality above, precisely because we do not have equality in

$$P(A \cap B \cap C) = P(\{\text{HHH, TTT}\}) = \frac{1}{4} \neq \frac{1}{8} = P(A)P(B)P(C).$$

That is why in Definition 71 we have to require that this equality holds.

Here is a useful fact about independent events that we state without proof. The next lemma generalizes Lemma 68.

Lemma 73. If the events $A_1, A_2, \ldots, A_n$ are independent, then
a) the events $A_1^c, A_2, \ldots, A_n$ are independent;
b) the events $A_1^c, A_2^c, \ldots, A_n$ are independent; and so on.

Knowing any combination of independent events gives no information about the probability of any other combination of events, as the next lemma shows.

Lemma 74. Suppose the events $A_1, A_2, \ldots, A_n$ are independent, then

$$P(A_1 \cap A_2 \cap \cdots \cap A_m | A_{m+1} \cap A_{m+2} \cap \cdots \cap A_n) = P(A_1 \cap A_2 \cap \cdots \cap A_m).$$

Proof. By the definition of conditional probability, we have

$$\begin{aligned} P(A_1 \cap A_2 \cap \cdots \cap A_m | A_{m+1} \cap A_{m+2} \cap \cdots \cap A_n) &= \frac{P(A_1 \cap A_2 \cap \cdots \cap A_m \cap A_{m+1} \cap A_{m+2} \cap \cdots \cap A_n)}{P(A_{m+1} \cap A_{m+2} \cap \cdots \cap A_n)} \\ &= \frac{P(A_1)P(A_2)\cdots P(A_m)P(A_{m+1})P(A_{m+2})\cdots P(A_n)}{P(A_{m+1})P(A_{m+2})\cdots P(A_n)} \\ &= P(A_1)P(A_2)\cdots P(A_m) \\ &= P(A_1 \cap A_2 \cap \cdots \cap A_m), \end{aligned}$$

□

The People of the State of California vs. Collins was a 1968 jury trial in California, USA that made notorious forensic use of mathematics and probability.[5]

Example 75 (People vs. Collins)**.** Bystanders to a robbery in Los Angeles testified that the perpetrators had been a black male, with a beard and moustache, and a caucasian female with blonde hair tied in a ponytail. They had escaped in a yellow motor car.

After testimony from an "instructor in mathematics" about the multiplication rule for probability, the prosecutor invited the jury to consider the probability that the accused pair, who fitted the description of the witnesses, were not the robbers. Even though the "instructor" had not discussed conditional probability, the prosecutor suggested that the jury would be safe in estimating:

Black man with beard	1 in 10
Man with moustache	1 in 4
White woman with pony tail	1 in 10
White woman with blonde hair	1 in 3
Yellow motor car	1 in 10
Interracial couple in a car	1 in 1000

The jury returned a verdict of guilty. Was the jury correct?

Solution. First, we find the probability that a randomly chosen couple in Los Angeles matches the description of the witnesses? Assuming that the events: "Black man with beard", "Man with moustache", ..., "Interracial couple in a car" are independent, then

$$P(\text{A random couple matches the description}) = \frac{1}{10}\frac{1}{4}\frac{1}{10}\frac{1}{3}\frac{1}{10}\frac{1}{1000} = \frac{1}{12,000,000}.$$

Denote this probability by p, that is, $p := 1/12,000,000$. At the time of the trial there were approximately $n := 8,000,000$ couples in Los Angeles area who could have possibly committed the crime. (So, p is the probability that a randomly chosen couple from the population matches the description.)

So, a priori, it seems that the jury was right to reach a guilty verdict since the probability of a couple matching the description is so small. To go into the issue deeper, we will calculate the probability that at least two couples in a large population have the characteristics, given that at least one has the characteristics. Let

A := the event that at least one couple matches the description,
B := the event that at least two couples match the description,

We are going to calculate the probability $P(B|A)$ that there is another couple that matches the description, given that there is at least one couple that matches the description. Using that $B \subset A$, we have

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

Calculating $P(A)$ is easier, so we start with it. We imagine n binomial trials with with probability of success p, and we want to know the probability of exactly one success. That is, we examine, the n possible couples and we know that with probability p any one of them may match the description. The probability of A is

$$P(A) = 1 - P(A^c) = 1 - P(\text{no couple matches the descr}) = 1 - b(n, p, 0) = 1 - (1-p)^n.$$

The probability of B is

$$\begin{aligned} P(B) &= 1 - P(B^c) = 1 - P(\text{0 or 1 couples match descr.}) \\ &= 1 - P(\text{exactly 0 couple matches descr.}) - P(\text{exactly 1 couples match descr.}) \\ &= 1 - b(n, p, 0) - b(n, p, 1) \\ &= 1 - (1-p)^n - np(1-p)^{n-1}. \end{aligned}$$

Substituting into the conditional probability we obtain

$$P(B|A) = \frac{P(B)}{P(A)} = \frac{1 - (1-p)^n - np(1-p)^{n-1}}{1 - (1-p)^n} = 1 - \frac{np(1-p)^{n-1}}{1 - (1-p)^n} \approx 0.2961.$$

Our results show that there is almost 30% chance that there is another couple with the given characteristics, if there is one. Note that the a priori probability of B is $P(B) = 0.1440765$ almost half of the conditional (posterior) probability $P(B|A)$. So, the fact that there is one couple matching the description on the stand, increases the probability that there is at least one more couple by over two times The jury had to be more careful handing out the verdict. □

The last example illustrates the so-called Prosecutor's Fallacy. You may find more information about it in the appendix.

6.3 Bayes' theorem

Definition 76. If $A_1, A_2, \ldots, A_n$ are disjoint events such that $\bigcup_{i=1}^{n} A_i = \Omega$ then say say that the events $A_1, A_2, \ldots, A_n$ *partition* Ω.

Lemma 77 (The law of total probability)**.** Let the events $A_1, A_2, \ldots, A_n$ *partition* Ω and let B be any other event. Then

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_n)P(A_n).$$

Proof. Since the events $A_1, A_2, \ldots, A_n$ are disjoint, then so are the events $B \cap A_1, B \cap A_2, \ldots, B \cap A_n$ and since $B = \bigcup_{i=1}^{n} B \cap A_i$, from the axioms of probability measure, we have

$$\begin{aligned} P(B) &= P(B \cap A_1) + P(B \cap A_2) + \cdots + P(B \cap A_n) \\ &= P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_n)P(A_n). \end{aligned}$$

□

The next example is analogous to Example 59. In fact, it can be solved in exactly the same way but now we will give an alternative solution based on conditional probability.

Example 78. Suppose we perform an experiment in a sequence, independently, and suppose A and B are two mutually exclusive (disjoint) events. What is the probability that event A occurs before event B?

Solution. The events A and B are subsets of some sample space Ω with probability measure P. Let $C := \Omega \setminus (A \cup B)$, that is when the experiment is performed exactly one of A, B, or C happens. Performing the experiment in a sequence is another, let us call it, *big* experiment. The set of all possible outcomes of the big experiment are all sequences of letters A, B, or C. This is the sample space of the big experiment, call it Ω^*. Let $H \subseteq \Omega^*$ be the event that A occurs before B, that is, the set of all sequences in Ω^* in which the letter A occurs before B. Let $E \subseteq \Omega^*$ be the event that A occurs on the first trial of the experiment, that is, the set of all sequences in Ω^* that start with the letter A. Let $F \subseteq \Omega^*$ be the event that B occurs on the first trial of the experiment, that is, the set of all sequences in Ω^* that start with the letter B. Let $G \subseteq \Omega^*$ be the event that neither A nor B occurs on the first trial of the experiment, that is, the set of all sequences in Ω^* that start with the letter C. Note that the events E, F, G partition the sample space Ω^* since they are disjoint and exactly one of A,B, or C must always occur on the first trial of the experiment. Let P^* be the probability measure on Ω^* that we know how to compute on certain events, as described in the examples in Subsection 5.6.1 Then by the law of total probability, on the one hand we have

$$P^*(H) = P^*(H|E)P^*(E) + P^*(H|F)P^*(F) + P^*(H|G)P^*(G).$$

On the other hand we have $P^*(E) = P(A)$, $P^*(F) = P(B)$, $P^*(G) = 1 - P(A) - P(B)$ and

$$\begin{aligned} P^*(H|E) &= 1, \\ P^*(H|F) &= 0, \\ P^*(H|G) &= P^*(H). \end{aligned}$$

The first two conditional probabilities above are obvious. For the third note that if the first experimental outcome is neither in A nor in B, then at that point the situation is exactly as when the problem first started; namely, the experimenter will continue to perform the experiment until either A or B occurs. The trials are independent, therefore, the outcome of the first trial will have no effect on subsequent ones. Thus, $P^*(H|G) = P^*(H \cap G)/P^*(G) = P^*(H)P^*(G)/P^*(G) = P^*(H)$. Substituting everything into the law of total probability, we get

$$P^*(H) = P(A) + P^*(H)(1 - P(A) - P(B)).$$

Solving for $P^*(H)$ gives

$$P^*(H) = \frac{P(A)}{P(A) + P(B)}.$$

Theorem 79 (Bayes theorem)**.** Let $A_1, A_2, \ldots, A_n$ be a partition of Ω with $P(A_i) > 0$, $i = 1, 2, \ldots, n$. Let B be any event with $P(B) > 0$. Then

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B|A_1)P(A_1) + \cdots + P(B|A_n)P(A_n)}.$$

Proof. By the definition of conditional probability, we have

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{P(B|A_1)P(A_1) + \cdots + P(B|A_n)P(A_n)},$$

where for the last equality, we used the law of total probability to express the denominator. □

The mutually exclusive sets A_1, $A_2, \ldots,$ A_n forming the partition of Ω are called **states of nature**. One of them must occur after every experiment (since their union is Ω). The probabilities $P(A_1)$, $P(A_2), \ldots,$ $P(A_n)$ are called **prior** probabilities and it is supposed that $P(A_1)$, $P(A_2), \ldots,$ $P(A_n)$ are known.

Then, based on new information (obtained from an experiment telling us that B occurred) we revise the prior probabilities to what are called **posterior** probabilities. The Bayes' theorem tells us how this revision can be done.

In other words, let B be an outcome of an experiment designed to help determine which is the true state of nature (that is, which A_i has occurred), suppose $P(B|A_i)$ is known, $i = 1, ..., n$. Then the **posterior probability** of a state of nature, say A_i, given the experiment B is $P(A_i|B)$.

Example 80. The AIDS incidence rate is 6 cases per $1,000$ Americans. Suppose that a person selected randomly for testing, tests positive for AIDS. The test is known to be highly accurate (99.9% for people with AIDS and 99% for people who do not). The AIDS incidence rate in the general population is 6 cases in 1000 people. What is the probability that the person actually has AIDS?

Soluton. Denote by $AIDS$ the event that the person has AIDS and by $no\, AIDS$ the complement of this event. We are given that

$$P(AIDS) = 0.006 \quad \text{hence} \quad P(no\, AIDS) = 0.994.$$

The two states of nature here are $AIDS$ and $no\,AIDS$. We want to recalculate the probabilities of these events, in view of the additional information that a test came up positive. Denote by POS the event that the test is positive and by $no\,POS$ the complement of this event, that is, the test is negative. The test accuracy information gives:

$$P(POS|AIDS) = 0.999 \qquad P(POS|no\,AIDS) = 0.01$$

We are primarily concerned with recomputing the new probability of the state of nature $AIDS$.

$$P(AIDS|POS) = \frac{P(AIDS \cap POS)}{P(POS)} = \frac{P(AIDS \cap POS)}{P(AIDS \cap POS) + P(no\,AIDS \cap POS)}$$

$$= \frac{P(AIDS)P(POS|AIDS)}{P(AIDS)P(POS|AIDS) + P(no\,AIDS)P(POS|no\,AIDS)} \qquad \text{(Bayes' theorem)}$$

$$= \frac{0.006 \times 0.999}{0.006 \times 0.999 + 0.994 \times 0.01} = \frac{0.005994}{0.005994 + 0.00994} = 0.38.$$

Thus, the probability that a person has AIDS given that he tested positive is 38%, a shockingly low number in view of the perceived high accuracy of the test. □

The conditional probability $P(POS|no\,AIDS) = 0.01$ is called **false positive**, the test is positive but a conclusion based on it would be false. The conditional probability

$$P(not\,POS|AIDS) = 1 - P(POS|AIDS) = 0.001$$

is called **false negative**, the test is negative but a conclusion based on it would be false.

Example 81. Suppose we have a box that contains one fair coin and one coin with a head on each side. A coin is selected at random and is tossed. Heads is obtained. Find the probability that we flipped the fair coin.

Solution. Let S_1 be the event that the fair coin is tossed and let S_2 be the probability that the unfair coin was tossed. The events S_1 and S_2 form a partition of the sample space (whatever it is) since the box does not contain anything else and exactly one coin is selected. Let E be the event that heads is obtained when the coin is tossed. We need to find $P(S_1|E)$. By the Bayes' theorem we have

$$\begin{aligned} P(S_1|E) &= \frac{P(S_1 \cap E)}{P(E)} = \frac{P(S_1)P(E|S_1)}{P(S_1 \cap E) + P(S_2 \cap E)} \\ &= \frac{P(S_1)P(E|S_1)}{P(S_1)P(E|S_1) + P(S_2)P(E|S_2)} = \frac{(1/2)(1/2)}{(1/2)(1/2) + (1/2)(1)} = 1/3. \end{aligned}$$

Example 82. Suppose we have a box that contains one fair coin and one coin with a head on each side. A coin is selected at random and is tossed *twice*. Both times heads were obtained. Find the probability that we flipped the fair coin.

Solution. Let S_1 be the event that the fair coin is tossed and let S_2 be the probability that the unfair coin was tossed. The events S_1 and S_2 form a partition of the sample space (whatever it is) since the box does not contain anything else and exactly one coin is selected. Let E be the event that heads are obtained when the coin is tossed twice. We need to find $P(S_1|E)$. By the Bayes' theorem we have

$$\begin{aligned} P(S_1|E) &= \frac{P(S_1 \cap E)}{P(E)} = \frac{P(S_1)P(E|S_1)}{P(S_1 \cap E) + P(S_2 \cap E)} \\ &= \frac{P(S_1)P(E|S_1)}{P(S_1)P(E|S_1) + P(S_2)P(E|S_2)} = \frac{(1/2)(1/4)}{(1/2)(1/4) + (1/2)(1)} = 1/5. \end{aligned}$$

As more heads are observed, the probability that the coin is fair keeps decreasing. That makes sense. If we keep observing heads, then it would seem more likely that we indeed have the two-headed coin. One may verify as an exercise that if three consecutive heads are observed then the probability of having selected the fair coin would be 1/9.

Example 83 (Monty Hall problem)**.** Suppose you're on Monty Hall's *Let's Make a Deal!* You are given the choice of three doors, behind one door is a car, behind the others—goats. You pick a door, say 1, Monty opens another door, say 3, which has a goat. Monty says to you "Do you want to pick door 2?" Is it to your advantage to switch your choice of doors?

Solution. We will calculate the probability that you pick a door and you stay with it and then the probability that you pick a door and switch. In the first case you pick a door at random from three possible ones, so the probability to win the car is 1/3. Let E_i be the event that the car is behind door $i = 1, 2, 3$. It is natural to assume that $P(E_1) = P(E_2) = P(E_3) = 1/3$. Let F_i be the event that the host reveals door $i = 1, 2, 3$. Note that the events E_1, E_2, E_3 partition the sample space. Since the prices are randomly permuted behind the doors, without loss of generality, suppose the player chooses door 1. Then, $P(F_1) = 0$ since the host cannot reveal the door that the contestant chose; $P(E_2 \cap F_2) = P(E_3 \cap F_3) = 0$ since the host will not open the door with the car behind. We assume that if the host has a choice between two doors to open, he will choose any one of them at random with probability 1/2, that is, we assume $P(F_2|E_1) = P(F_3|E_1) = 1/2$.

We now calculate the probability that the player wins if he switches.

$$\begin{aligned} P(E_3|F_2) &= \frac{P(E_3 \cap F_2)}{P(F_2)} = \frac{P(F_2|E_3)P(E_3)}{P(F_2)} = \frac{P(F_2|E_3)P(E_3)}{P(F_2|E_1)P(E_1) + P(F_2|E_2)P(E_2) + P(F_2|E_3)P(E_3)} \\ &= \frac{1(1/3)}{(1/2)(1/3) + 0(1/3) + 1(1/3)} = \frac{1/3}{1/2} = \frac{2}{3}. \end{aligned}$$

So the probability of winning doubles if the player switches. □

Exercise 84. Flip a fair coin repeatedly. What is the probability that the first sequence of heads is exactly two heads long?

7 Random variables

7.1 Density function

Suppose you are given a sample space Ω with a probability measure P. A random variable is a function $X : \Omega \to \mathbb{R}$.

Definition 85. A random variable that takes on finite or at most a countable number of possible values is said to be *discrete*. The *probability mass function* $p(a)$ of X is defined by

$$p(a) := P(X = a) := P(\{\omega \in \Omega : X(\omega) = a\}).$$

So, $p(a)$ is just the probability that X takes the value a. If X takes only the values $x_1, x_2, \ldots$, then $p(x_i) \geq 0$ for all $i = 1, 2, \ldots$ and $p(x) = 0$ for all other x's. Since X must take on one of the values x_i, we have

$$\sum_{i=1}^{\infty} p(x_i) = 1.$$

Definition 86. If the random variable X takes uncountably many different values, for example, if it can take any value in an interval (a, b), then X is called *continuous* random variable.

If X is a continuous random variable, then things are a little bit more complicated and that is why we will consider only a special class of continuous random variables.

Definition 87. A random variable X is *absolutely continuous* if there is a non-negative function $f(x)$ defined on $\mathbb{R}$ such that for any interval $[a, b]$ of values of X the following equality is true

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx. \tag{11}$$

The function $f(x)$ is called *probability density of* X, or *p.d.f. of* X, or just *density of* X.

The essential properties of a density function are

1.) $f(x) \geq 0$ for all x, and

2.) $\int_{-\infty}^{\infty} f(x)\,dx = P(X \in (-\infty, \infty)) = 1$.

If we let $a = b$ in Equation (11), we obtain

$$P(X = a) = \int_a^a f(x)\,dx = 0.$$

This equation states that the probability a continuous random variable assumes any fixed value is zero. Isn't this a paradox? This does not mean that X can never assume value a, just that if you pre-specify a you will never see that *exact* value, maybe due to measuring errors (you can only measure with finite precision) or some other problem. In practice continuous measurements are usually rounded. When I say that I weigh 75kgs, I really mean that I weigh between 74.99 and 75.01 depending on the precision of my scales. Then

$$P(X \text{ is about } 75) = P(74.99 \leq X \leq 75.01) = \int_{74.99}^{75.01} f(x)\, dx$$

and the last integral is perhaps different from zero. Another oddity is that the pdf function is not unique. You can change a pdf at a discrete (for example) set of points without changing the integral. For example the following three functions give the same integrals

$$f(x) = \begin{cases} e^{-x} & x > 0, \\ 0 & x \leq 0; \end{cases} \qquad f(x) = \begin{cases} e^{-x} & x \geq 0, \\ 0 & x < 0; \end{cases} \qquad f(x) = \begin{cases} e^{-x} & x \geq 0, x \neq 4, \\ 10 & x = 4, \\ 0 & x < 0; \end{cases}$$

Definition 88. We say that a continuous random variable X is *uniformly distributed* in the interval $[a, b]$ if its pdf is

$$f(x) = \begin{cases} 1/(b-a) & \text{if } x \in [a, b], \\ 0 & \text{otherwise.} \end{cases}$$

Let $[c, d] \subset [a, b]$, from the definition, we compute

$$P(c \leq X \leq d) = \int_c^d \frac{1}{b-a}\, dx = \frac{d-c}{b-a}.$$

Uniformly distributed random variables are used to model experiments whose outcomes are values in an interval $[a, b]$ and the probability that the outcome is in a subinterval $[c, d] \subset [a, b]$ depends only on the length of $[c, d]$.

Notice that if $a = 0$ and $b = 1/2$, then $f(x) = 1/(1/2 - 0) = 2$ for all $x \in [0, 1/2]$. But that is a number bigger than 1 and probabilities are less than 1! This reinforces the fact that density are not probabilities, the integral of the density gives probabilities.

It is important to remember that continuous random variables are two types: those that have a probability density function and those that do not.

Example 89. Consider a random variable X whose values are the outcomes of the following experiment. Throw a fair coin. If the outcome is H then pick a random number uniformly distributed in $[0, 1]$. If the outcome is T then pick a random number from the set $\{2, 3\}$ with probabilities $2/3, 1/3$, respectively. This random variable is continuous since it can take uncountably many values, namely any value in $[0, 1] \cup \{2, 3\}$ but it is not absolutely continuous as we will see in the next section.

7.2 Cumulative distribution function

Definition 90. Let X be a random variable (discrete or continuous). The *cumulative distribution function of* X is

$$F(x) = P(X \leq x) \text{ for } -\infty < x < \infty.$$

For short we will call $F(x)$ just the *distribution of* X or the *cdf of* X. It has the following properties

1.) $F(x)$ is increasing function: if $x_1 \leq x_2$ then $F(x_1) \leq F(x_2)$. Indeed, the event $\{X \leq x_1\}$ is a subset of the event $\{X \leq x_2\}$ so the probability of the first is less-than-or-equal than the probability of the second.

2.) $\lim_{x\to\infty} F(x) = 1$.

3.) $\lim_{x\to-\infty} F(x) = 0$.

4.) $F(x)$ is continuous from the right. This means that if y approaches x with bigger values (on the right), then $F(y)$ approaches $F(x)$. We write this as $\lim_{y\to x^+} F(y) = F(x)$. In other words, if $y_1 \geq y_2 \geq y_3 \geq \cdots$ is a decreasing sequence converging to x (note that necessarily $x \leq y_n$.) then $F(y_n)$ is a decreasing sequence converging to $F(x)$.

Example 91. If X is a discrete random variable having a probability mass function $p(1) = 1/6$, $p(2) = 1/6$, $p(3) = 1/4$, $p(4) = 5/12$, then its cdf is

$$F(x) = \begin{cases} 0 & \text{if } x < 1, \\ 1/6 & \text{if } 1 \leq x < 2, \\ 1/3 & \text{if } 2 \leq x < 3, \\ 7/12 & \text{if } 3 \leq x < 4, \\ 1 & \text{if } 4 \leq x. \end{cases}$$

Note that if is an absolutely continuous random variable with pdf $f(x)$, then

$$F(x) = \int_{-\infty}^{x} f(x)\,dx. \tag{12}$$

Example 92. If X is uniformly distributed in $[a, b]$, then

$$F(x) = \begin{cases} 0 & \text{if } x < a, \\ (x-a)/(b-a) & \text{if } a \leq x < b, \\ 1 & \text{if } b \leq x. \end{cases}$$

The four properties of the cumulative distribution function completely characterize it.

Theorem 93. If a function $F(x)$ satisfies the four properties after Definition 90 then there is a random variable X with cumulative distribution function equal to $F(x)$.

It is important to remember that not every random variable X has a probability density function, for example the discrete random variables do not have probability density function, they have probability mass function. By our definition, if X is continuous, then it has to be absolutely continuous to have probability density function. But, every random variable has cumulative distribution function $F(x)$ because by definition $F(x) = P(X \leq x)$ and the latter probability is a well-defined number. The next rule for recognizing when a random variable has a probability density function is very important.

If a random variable X has density, then by (12), its cumulative distribution function $F(x)$ is continuous. So, looking at the cumulative distribution function, if it has jumps, then the random variable cannot be absolutely continuous.

Let us look at the cdf of the random variable X described in Exercise 89.

Example 94. Consider the random variable X described in Exercise 89. Let Y be a random variable representing the coin flip, that is Y takes values H and T with probabilities $1/2$, $1/2$ respectively. Let Z be a random variable uniformly distributed in $[0,1]$ and let T be a random variable taking values $\{2,3\}$ with probabilities $2/3, 1/3$, respectively. We consider several cases.

1) Let $x < 0$, then $F(x) = P(X \leq x) = 0$ since the r.v. X never takes values smaller than 0.

2) Let $0 \leq x \leq 1$, then $\{X \leq x\} = \{Y = H\} \cap \{Z \leq x\}$ where the last two events are independent. Hence

$$F(x) = P(X \leq x) = P(\{Y = H\} \cap \{Z \leq x\}) = P(Y = H)P(Z \leq x) = \frac{1}{2}\frac{x-0}{1-0} = \frac{x}{2}.$$

3) Let $1 \leq x < 2$, then $\{X \leq x\} = \{X \leq 1\}$ and $F(x) = P(X \leq x) = P(X \leq 1) = F(1) = 1/2$, where $F(1)$ was computed in case 1.

4) Let $2 \leq x < 3$, then

$$\{X \leq x\} = \{X \leq 2\} = \{X \leq 1\} \cup \{1 < X \leq 2\} = \{X \leq 1\} \cup \{X = 2\},$$

where the union is disjoint. Hence, $F(x) = P(\{X \leq 1\} \cup \{X = 2\}) = P(\{X \leq 1\}) + P(\{X = 2\}) = 1/2 + P(\{X = 2\})$. Now, $\{X = 2\} = \{Y = T\} \cap \{T = 2\}$, where the last two events are independent, implying that $P(X = 2) = P(Y = T)P(T = 2) = (1/2)(2/3) = 1/3$. Thus, in this case $F(x) = 1/2 + 1/3 = 5/6$.

5) Let $3 \leq x$, then

$$\{X \leq x\} = \{X \leq 3\} = \{X \leq 1\} \cup \{1 < X \leq 2\} \cup \{2 < X \leq 3\} = \{X \leq 1\} \cup \{X = 2\} \cup \{X = 3\}.$$

But we do not need the last union at all. Since X can takes only values that are always less-than-or-equal to 3, we have $F(x) = P(X \leq 3) = 1$.

We summarize all cases in

$$F(x) = \begin{cases} 0 & \text{if } x < 0, \\ x/2 & \text{if } 0 \leq x < 1, \\ 1/2 & \text{if } 1 \leq x < 2, \\ 5/6 & \text{if } 2 \leq x < 3, \\ 1 & \text{if } 3 \leq x. \end{cases}$$

Since $F(x)$ is not continuous (note that it always has to be right-continuous) we conclude that X does not have a probability density function. So this is an example of a continuous random variable that is not absolutely continuous.

As we know a cdf is always right continuous but it may not be left continuous. Since F is increasing function $F(x) \leq F(a)$ for all $x \leq a$. Hence taking the limit as x approaches a from the left we obtain that

$$\lim_{x \to a^-} F(x) \leq F(a).$$

The limit of $F(x)$ as x approaches a number a from the left is denoted by $F(a-)$, that is, we define the notation

$$F(a-) := \lim_{x \to a^-} F(x)$$

and the above observation may be written as

$$F(a-) \leq F(a) \text{ for all } a \in \mathbb{R}.$$

The cumulative distribution function can answer all questions about a random variable X.

Lemma 95. If X is a random variable with cdf $F(x)$ then

(i) $P(X < a) = F(a-)$;

(ii) $P(X = a) = F(a) - F(a-)$.

(iii) $P(a < X \leq b) = F(b) - F(a)$;

(iv) $P(a < X < b) = F(b-) - F(a)$;

(v) $P(a \leq X < b) = F(b-) - F(a-)$;

(vi) $P(a < X) = 1 - F(a)$.

Proof. (i) We are going to take this property on faith, without proof and will use it to derive the rest of the properties.

(ii) Since $\{X \leq a\} = \{X = a\} \cup \{X < a\}$ with the union being disjoint, taking probabilities gives

$$F(a) = P(X \leq a) = P(X = a) + P(X < a) = P(X = a) + F(a-).$$

Solve for $P(X = a)$ to get the result.

(iii) Since $\{X \leq b\} = \{a < X \leq b\} \cup \{X \leq a\}$ with the union being disjoint, taking probabilities one concludes as in the previous case.

The reset of the items are left as an exercise. □

If X is an absolutely continuous random variable, then X has a probability density function and the cdf $F(x)$ is continuous. Then, $F(a-) = F(a)$ for all $a \in \mathbb{R}$. From the lemma we see that

- $P(X \leq a) = P(X < a) = F(a)$;
- $P(X = a) = 0$;
- $P(a < X \leq b) = P(a \leq X < b) = P(a < X < b) = P(a \leq X \leq b) = F(b) - F(a)$.

Henceforth, all continuous random variables that we consider will be absolutely continuous random variables. That is why we will omit the word ‘absolutely’ and just call them continuous.

Suppose X is absolutely continuous random variable, that is, X has a probability density function $f(x)$. So how can we obtain $f(x)$ if we know $F_X(x)$? The answer lies in the integral representation (12) of $F_X(x)$ and the fundamental theorem of calculus.

Theorem 96 (The fundamental theorem of calculus)**.** If $f(x)$ is continuous and integrable function, then the function

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

is differentiable and $F'(x) = f(x)$ for all x.

When we want to emphasize that a p.d.f. $f(x)$ and a c.d.f. $F(x)$ are those of a random variable X, we write $f_X(x)$ and $F_X(x)$.

Example 97. Let X be a continuous random variable with p.d.f. $f(x)$ and c.d.f. $F(x)$. Find the p.d.f. and the c.d.f. of the random variable $Y := |X|$.

Solution. Note that Y takes only positive values, so $F_Y(y) = P(Y \le y) = 0$ for $y < 0$. When $0 \le y$ we compute

$$F_Y(y) = P(Y \le y) = P(|X| \le y) = P(-y \le X \le y) = F_X(y) - F_X(-y).$$

To find $f_Y(y)$ differentiate both sides of the above with respect to y, using that $\frac{d}{dy}F_X(y) = f_X(y)$:

$$f_Y(y) = f_X(y) + f_X(-y).$$

Example 98. Let X be a continuous random variable with p.d.f. $f(x)$ and c.d.f. $F(x)$. Find the p.d.f. and the c.d.f. of the random variable $Y := X^2$.

Solution. Note that Y takes only positive values, so $F_Y(y) = P(Y \le y) = 0$ for $y < 0$. When $0 \le y$ we compute

$$F_Y(y) = P(Y \le y) = P(X^2 \le y) = P(-\sqrt{y} \le X \le \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y}).$$

To find $f_Y(y)$ differentiate both sides of the above with respect to y, using that $\frac{d}{dy}F_X(\sqrt{y}) = f_X(\sqrt{y})\frac{1}{2\sqrt{y}}$:

$$f_Y(y) = \frac{1}{2\sqrt{y}}\big(f_X(\sqrt{y}) + f_X(-\sqrt{y})\big).$$

We conclude this section with an interpretation of the probability density function. As we said, the value $f(x)$ is not the probability that X will take value x. Instead we have the following

$$F(x+h) - F(x) = P(x \le X \le x+h)$$

Dividing both sides by h, and taking limit as h goes to 0, we obtain

$$f(x) = F'(x) = \lim_{h\to 0} \frac{F(x+h) - F(x)}{h} = \lim_{h\to 0} \frac{P(x \le X \le x+h)}{h},$$

where we used the fundamental theorem of calculus, Theorem 96, to assertain the first equality. Hence, for values of h close to 0

$$f(x)h \approx P(x \le X \le x+h). \tag{13}$$

This is the intuitive meaning behind the p.d.f.

8 Expected value & variance

One of the most important concepts in probability theory is that of the expectation of a random variable.

Definition 99 (Expected value)**.** Let X be a random variable. The *expectation* or the *expected value*, or the *mean* of X, denoted by $E[X]$, is defined as follows.

- If X is a discrete random variable with values $\{x_1, x_2, \ldots\}$ and probability mass function $p(x)$
$$E[X] := \sum_{i=1}^{\infty} x_i p(x_i).$$
- If X is a continuous random variable with values in $\mathbb{R}$ and probability density function $f(x)$
$$E[X] := \int_{-\infty}^{\infty} x f(x)\, dx.$$

Often the expected value is denoted by μ, that is $\mu := E[X]$.

Note the similarities between the discrete and the continuous case: the sum becomes an integral and the probability mass function becomes the probability density function. The expected value of X is just the *average* of all the values that X takes. It is actually a weighted average with weights determined by the probability mass function $p(x)$ or the probability density function $f(x)$.

Assume that the relative frequency interpretation of probabilities holds. That is, if we run an experiment an infinite number of times, then for any event E, the proportion of time that E occurs will be $P(E)$. Now, consider a random variable X that takes values $\{x_1, x_2, \ldots, x_n\}$ with probabilities $\{p(x_1), p(x_2), \ldots, p(x_n)\}$. Think of X as representing our winings in a single game. That is, with probability $p(x_i)$ we win x_i units $i = 1, 2, \ldots, x_n$. It follows that if we continually play this game, then the proportion of time that we win x_i will be $p(x_i)$. As this is true for all $i = 1, 2, \ldots, n$, it follows that our average winnings per game will be $\sum_{i=1}^{n} x_i p(x_i) = E[X]$.

Example 100. Find $E[X]$ where X is the outcome when we roll a fair die. The random variable X takes values $\{1, 2, 3, 4, 5, 6\}$ with probabilities $p(1) = p(2) = p(3) = p(4) = p(5) = p(6) = 1/6$. Thus
$$E[X] = 1(1/6) + 2(1/6) + \cdots + 6(1/6) = \frac{1+2+3+4+5+6}{6} = \frac{7}{2}.$$

Example 101. Let $A \subset \Omega$ be an event and consider the random variable
$$X(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \notin A. \end{cases}$$

The random variable X is discrete, it takes only two values $\{0, 1\}$, and is called the *indicator function of A*. Its pdf is $p(1) = P(X = 1) = P(A)$ and $p(0) = P(X = 0) = P(A^c)$. Hence, the expected value of X is
$$E[X] = 0p(0) + 1p(1) = P(A).$$

Example 102. Let X be a binomial random variable, that is, it takes values $\{0, 1, 2, \ldots, n\}$ with probabilities $\{\binom{n}{0}p^0q^n, \binom{n}{1}p^1q^{n-1}, \binom{n}{2}p^2q^{n-2}, \ldots, \binom{n}{n}p^nq^0\}$, where $p, q \in [0, 1]$ with $p + q = 1$. Its expected value is

$$E[X] = \sum_{i=0}^{n} i\binom{n}{i}p^iq^{n-i} = \sum_{i=1}^{n} i\binom{n}{i}p^iq^{n-i} = \sum_{i=1}^{n} n\binom{n-1}{i-1}p^iq^{n-i},$$

where we used the easy to verify fact that $i\binom{n}{i} = n\binom{n-1}{i-1}$ valid for $i \geq 1$. We continue by factoring a p out and changing the summation index to $j := i - 1$, that is $i = j + 1$:

$$E[X] = np\sum_{i=1}^{n}\binom{n-1}{i-1}p^{i-1}q^{n-i} = np\sum_{j=0}^{n-1}\binom{n-1}{j}p^jq^{(n-1)-j} = np(p+q)^{n-1} = np.$$

Example 103. Let X be a uniformly distributed random variable in the interval $[a, b]$. Then

$$E[X] = \int_{-\infty}^{\infty} xf(x)\,dx = \int_a^b \frac{x}{b-a}\,dx = \frac{1}{2}\frac{1}{b-a}\left(x^2\Big|_{x=a}^{b}\right) = \frac{1}{2}\frac{1}{b-a}(b^2 - a^2) = \frac{a+b}{2}.$$

Suppose that X is a random variable and g is a function defined on $\mathbb{R}$. (We require that the domain of g includes the range of X.) The composition $(g \circ X)(\omega) = g(X(\omega))$ is a function on Ω, hence a random variable. Often one needs to find the expected value of $g(X)$. The next theorem explains how this is done.

Theorem 104. Suppose that X is a random variable and g is a function defined on $\mathbb{R}$. The *expected value* of $g(X)$, denoted by $E[g(X)]$, is calculated as follows.

- If X is a discrete random variable with values $\{x_1, x_2, \ldots\}$ and probability mass function $p(x)$

$$E[g(X)] := \sum_{i=1}^{\infty} g(x_i)p(x_i).$$

- If X is a continuous random variable with values in $\mathbb{R}$ and probability density function $f(x)$

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

Proof. The proof given is only in the case of a discrete random variable X. The proof proceeds by grouping together all the terms in $\sum_{i=1}^{\infty} g(x_i)p(x_i)$ having the same value of $g(x_i)$. Suppose that $\{y_1, y_2, \ldots\}$ represent the different values of $g(x_i)$, $i = 1, 2, \ldots$ Grouping all the $g(x_i)$ having the same value gives

$$\begin{aligned}\sum_{i=1}^{\infty} g(x_i)p(x_i) &= \sum_{j=1}^{\infty} \sum_{\substack{\text{over all } i \text{ s.t.} \\ g(x_i)=y_j}} g(x_i)p(x_i) = \sum_{j=1}^{\infty} y_j \sum_{\substack{\text{over all } i \text{ s.t.} \\ g(x_i)=y_j}} p(x_i) \\ &= \sum_{j=1}^{\infty} y_jP(g(X) = y_j) = E[g(X)].\end{aligned}$$

□

Corollary 105. If a and b are constants, then $E[aX+b] = aE[X]+b$.

Proof. The proof in the discrete case is left as an exercise. In the continuous case, let the function $g(x) := ax+b$. By the last theorem we have

$$\begin{aligned} E[aX+b] = E[g(X)] &= \int_{-\infty}^{\infty} g(x)f(x)\,dx = \int_{-\infty}^{\infty} (ax+b)f(x)\,dx = \int_{-\infty}^{\infty} (axf(x)+bf(x))\,dx \\ &= a\int_{-\infty}^{\infty} xf(x)\,dx + b\int_{-\infty}^{\infty} f(x)\,dx = aE[X]+b. \end{aligned}$$

□

The expected value of a random variable X, $E[X]$, is also referred to as the *first moment* of X. The quantity $E[X^k]$, for $k \geq 1$, is called the *k-th moment of X*. Applying Theorem 104 with function $g(x) = x^k$, we obtain

- If X is a discrete random variable with values $\{x_1, x_2, \ldots\}$ and probability mass function $p(x)$
$$E[X^k] = \sum_{i=1}^{\infty} x_i^k p(x_i).$$
- If X is a continuous random variable with values in $\mathbb{R}$ and probability density function $f(x)$
$$E[X^k] = \int_{-\infty}^{\infty} x^k f(x)\,dx.$$

Let us compute the second moment of the random variables in Examples 100, 101, 102, and 103.

Example 106. Find $E[X^2]$ where X is the outcome when we roll a fair die. The random variable X takes values $\{1,2,3,4,5,6\}$ with probabilities $p(1) = p(2) = p(3) = p(4) = p(5) = p(6) = 1/6$. Thus

$$E[X^2] = 1^2(1/6) + 2^2(1/6) + \cdots + 6^2(1/6) = \frac{1^2+2^2+3^2+4^2+5^2+6^2}{6} = \frac{91}{6}.$$

Example 107. Let $A \subset \Omega$ be an event and consider the random variable

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

The random variable X is discrete, it takes only two values $\{0,1\}$, and is called the *indicator function of A*. Its pdf is $p(1) = P(X=1) = P(A)$ and $p(0) = P(X=0) = P(A^c)$. Hence, the second moment of X is

$$E[X^2] = 0^2 p(0) + 1^2 p(1) = P(A).$$

Example 108. Let X be a binomial random variable, that is, it takes values $\{0, 1, 2, \ldots, n\}$ with probabilities $\{\binom{n}{0}p^0q^n, \binom{n}{1}p^1q^{n-1}, \binom{n}{2}p^2q^{n-2}, \ldots, \binom{n}{n}p^nq^0\}$, where $p, q \in [0, 1]$ with $p + q = 1$. Its second moment is

$$E[X^2] = \sum_{i=0}^{n} i^2 \binom{n}{i} p^i q^{n-i} = \sum_{i=1}^{n} i^2 \binom{n}{i} p^i q^{n-i} = \sum_{i=1}^{n} ni \binom{n-1}{i-1} p^i q^{n-i},$$

where we used the easy to verify fact that $i\binom{n}{i} = n\binom{n-1}{i-1}$ valid for $i \geq 1$. We continue by factoring a p out and changing the summation index to $j := i - 1$, that is $i = j + 1$:

$$\begin{aligned} E[X^2] &= np \sum_{i=1}^{n} i \binom{n-1}{i-1} p^{i-1} q^{n-i} = np \sum_{j=0}^{n-1} (j+1) \binom{n-1}{j} p^j q^{(n-1)-j} \\ &= \left(np \sum_{j=0}^{n-1} j \binom{n-1}{j} p^j q^{(n-1)-j} \right) + \left(np \sum_{j=0}^{n-1} \binom{n-1}{j} p^j q^{(n-1)-j} \right) \\ &= npE[Y] + np(p+q)^{n-1} = npE[Y] + np, \end{aligned}$$

where Y is a binomial random variable taking values $\{0, 1, 2, \ldots, n-1\}$ with probabilities $P(Y = k) = \binom{n-1}{j} p^j q^{(n-1)-j}$, for $k = 0, 1, 2, \ldots, n-1$. But we know what its expected value is $E[Y] = (n-1)p$ and substituting above, we finally obtain

$$E[X^2] = np((n-1)p + 1).$$

Example 109. Let X be a uniformly distributed random variable in the interval $[a, b]$. Then

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f(x)\, dx = \int_a^b \frac{x^2}{b-a}\, dx = \frac{1}{3} \frac{1}{b-a} \left(x^3 \Big|_{x=a}^{b} \right) = \frac{1}{3} \frac{1}{b-a} (b^3 - a^3) = \frac{b^2 + ba + a^2}{3}.$$

Let X be a random variable with expected value $\mu := E[X]$. Since the expected value of X is its average value, X takes values that are both smaller and larger than $E[X]$. A reasonable way of measuring the possible variation of X would be to look at how far apart X would be from its mean on the average. One possible way to measure this is to consider $E[|X - \mu|]$. This quantity is the average distance between X and μ. The absolute value makes it mathematically inconvenient to deal with this quantity. More tractable quantity is the average squared difference between X and its mean. Note that $|X - \mu|^2 = (X - \mu)^2$. Consider the function $g(x) := (x - \mu)^2$. The expected value of $g(X)$, calculate it using Theorem 104, is called the *variance of* X.

Definition 110. Let X be a random variable with mean μ. The *variance* of X, denoted $\mathrm{Var}\,(X)$ is defined by

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

Proposition 111 (Properties of variance)**.** The variance of a random variable X has the following properties.

(i) $\mathrm{Var}\,(X) \geq 0$;

(ii) $\mathrm{Var}\,(X) = E[X^2] - (E[X])^2$;

(iii) For any constants a and b we have $\mathrm{Var}\,(aX+b) = a^2\mathrm{Var}\,(X)$.

Proof. We present the proof only for discrete random variable. The situation for continuous random variable is completely analogous. 1) $\mathrm{Var}\,(X) = E[(X-\mu)^2] = \sum_{i=1}^{\infty}(x_i-\mu)^2 p(x_i) \geq 0$;
2) Continuing the development in part 1), we have

$$\begin{aligned}\mathrm{Var}\,(X) = E[(X-\mu)^2] &= \sum_{i=1}^{\infty}(x_i-\mu)^2 p(x_i) = \sum_{i=1}^{\infty}(x_i^2 - 2x_i\mu + \mu^2)p(x_i)\\ &= \sum_{i=1}^{\infty} x_i^2 p(x_i) - 2\mu\sum_{i=1}^{\infty} x_i p(x_i) + \mu^2\sum_{i=1}^{\infty} p(x_i)\\ &= E[X^2] - 2\mu^2 + \mu^2\\ &= E[X^2] - \mu^2.\end{aligned}$$

3) First recall that $E[aX+b] = aE[X]+b = a\mu+b$. By the definition of the variance, considering the random variable $aX+b$ we have

$$\begin{aligned}\mathrm{Var}\,(aX+b) &= E\big[\big((aX+b) - E[aX+b]\big)^2\big] = E\big[\big((aX+b) - (a\mu+b)\big)^2\big]\\ &= E\big[a^2(X-\mu)^2\big] = a^2E\big[(X-\mu)^2\big] = a^2\mathrm{Var}\,(X).\end{aligned}$$

□

Using the second property in Proposition 111 we calculate the variance for the random variables in Examples 100, 101, 102, and 103.

Example 112. Find $E[X^2]$ where X is the outcome when we roll a fair die. The random variable X takes values $\{1,2,3,4,5,6\}$ with probabilities $p(1) = p(2) = p(3) = p(4) = p(5) = p(6) = 1/6$. The variance of X is

$$\mathrm{Var}\,(X) = E[X^2] - (E[X])^2 = \left(\frac{91}{6}\right) - \left(\frac{7}{2}\right)^2 = \frac{35}{12}.$$

Example 113. Let $A \subset \Omega$ be an event and consider the random variable

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

The random variable X is discrete, it takes only two values $\{0,1\}$, and is called the *indicator function of* A. Its pdf is $p(1) = P(X=1) = P(A)$ and $p(0) = P(X=0) = P(A^c)$. The variance of X is

$$\mathrm{Var}\,(X) = E[X^2] - (E[X])^2 = P(A) - P(A)^2 = P(A)(1-P(A)) = P(A)P(A^c).$$

Example 114. Let X be a binomial random variable, that is, it takes values $\{0,1,2,\ldots,n\}$ with probabilities $\{\binom{n}{0}p^0q^n, \binom{n}{1}p^1q^{n-1}, \binom{n}{2}p^2q^{n-2}, \ldots, \binom{n}{n}p^nq^0\}$, where $p, q \in [0,1]$ with $p+q=1$. The variance of X is

$$\mathrm{Var}\,(X) = E[X^2] - (E[X])^2 = np((n-1)p+1) - (np)^2 = np(1-p) = npq.$$

Example 115. Let X be a uniformly distributed random variable in the interval $[a, b]$. The variance of X is

$$\mathrm{Var}\,(X) = E[X^2] - (E[X])^2 = \frac{b^2 + ab + a^2}{3} - \Big(\frac{a+b}{2}\Big)^2 = \frac{(a-b)^2}{12}.$$

Definition 116. The square root of the variance, that is $\sqrt{\mathrm{Var}\,(X)}$, is called the *standard deviation of* X.

We conclude this section with general formulas that may be used for computing the expected value and the variance of a random variable X.

If the random variable X has probability density function $f(x)$, then

$$E[X] = \int_{-\infty}^{\infty} x f(x)\, dx,$$
$$E[X^2] = \int_{-\infty}^{\infty} x^2 f(x)\, dx.$$

Hence for the variance of X we get

$$\mathrm{Var}\,(X) = E[X^2] - (E[X])^2 = \int_{-\infty}^{\infty} x^2 f(x)\, dx - \Big(\int_{-\infty}^{\infty} x f(x)\, dx\Big)^2.$$

But what if X does not have a probability density function $f(x)$? We said that the cumulative distribution function $F(x)$ of a random variable X can answer all questions about X. Hence it is not surprising that we have the formulas, summarized in Proposition 117 and Proposition 118. We are not going to prove them and we are not going to use these formulas anywhere else in the course, but they are worth mentioning nonetheless.

Proposition 117. For a non-negative random variable X with cumulative distribution function $F(x)$, and any $k > 0$, we have

$$E[X^k] = \int_0^{\infty} k x^{k-1} (1 - F(x))\, dx.$$

In particular, for $k = 1$ and $k = 2$ we have

$$E[X] = \int_0^{\infty} (1 - F(x))\, dx \quad \text{and} \quad E[X^2] = 2\int_0^{\infty} x(1 - F(x))\, dx.$$

Proposition 118. For any random variable X with cumulative distribution function $F(x)$, we have

$$\begin{aligned} E[X] &= \int_0^{\infty} P(X > x)\, dx - \int_0^{\infty} P(X < -x)\, dx \\ &= \int_0^{\infty} (1 - F(x))\, dx - \int_0^{\infty} F(-x-)\, dx. \end{aligned}$$

In particular, if X is absolutely continuous random variable ($F(x)$ is a continuous function), then

$$E[X] = \int_0^{\infty} (1 - F(x))\, dx - \int_0^{\infty} F(-x)\, dx.$$

Definition 119. The *moment generation function* $M(t)$ of a random variable X is defined by

$$M(t) = E[e^{tX}], \text{ for all } t \in \mathbb{R}.$$

To calculate the moment generating function of a random variable X, we use Theorem 104 with $g(x) = e^{tx}$ where we treat t as a fixed parameter. Thus, if X is a discrete random variable with values $\{x_1, x_2, \ldots\}$ and probability mass function $p(x)$ then

$$M(t) = \sum_{i=1}^{\infty} e^{tx_i} p(x_i);$$

and if X is a continuous random variable with values in $\mathbb{R}$ and probability density function $f(x)$

$$M(t) = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx.$$

The function $M(t)$ is called moment generating because the k-th moment of X can be obtained by differentiating $M(t)$ k-times and then evaluating the result at $t = 0$.

Assumption 120. Assume that we may interchange the differentiation and expectation operators. That is, assume that, when X is a discrete random variable

$$\frac{d}{dt}\Big(\sum_{i=1}^{\infty} e^{tx_i} p(x_i)\Big) = \sum_{i=1}^{\infty} \frac{d}{dt} e^{tx_i} p(x_i)$$

and when X is continuous

$$\frac{d}{dt}\Big(\int_{-\infty}^{\infty} e^{tx} f(x)\, dx\Big) = \int_{-\infty}^{\infty} \frac{d}{dt} e^{tx} f(x)\, dx.$$

This assumption, basically means that when differentiating $M(t)$ we can do it as follows

$$M'(t) = \frac{d}{dt} M(t) = \frac{d}{dt} E[e^{tX}] = E\Big[\frac{d}{dt} e^{tX}\Big] = E[Xe^{tX}].$$

Evaluating the result at $t = 0$, we get the first moment of X, or its expected value:

$$M'(0) = E[X].$$

Differentiating $M'(t)$ we have

$$M''(t) = \frac{d}{dt} M'(t) = \frac{d}{dt} E[Xe^{tX}] = E\Big[\frac{d}{dt} Xe^{tX}\Big] = E[X^2 e^{tX}].$$

Evaluating the result at $t = 0$, we get the second moment of X:

$$M''(0) = E[X^2].$$

We may continue like that, to obtain

$$M^{(k)}(t) = E[X^k e^{tX}] \text{ and } M^{(k)}(0) = E[X^k].$$

Example 121. Let X be a binomial random variable, that is, it takes values $\{0,1,2,\dots,n\}$ with probabilities $\{\binom{n}{0}p^0q^n, \binom{n}{1}p^1q^{n-1}, \binom{n}{2}p^2q^{n-2}, \dots, \binom{n}{n}p^nq^0\}$, where $p,q \in [0,1]$ with $p+q=1$. The moment generating function of X is

$$M(t) = E[e^{tX}] = \sum_{k=0}^{n} e^{tk}\binom{n}{k}p^kq^{n-k} = \sum_{k=0}^{n}\binom{n}{k}(e^tp)^kq^{n-k} = (pe^t+q)^n.$$

Example 122. Let X be a uniformly distributed random variable in the interval $[a,b]$. The moment generating function of X is

$$M(t) = E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx}f(x)\,dx = \int_a^b \frac{e^{tx}}{b-a}\,dx = \frac{e^{tx}}{t(b-a)}\Big|_{x=a}^{b} = \frac{e^{tb}-e^{ta}}{t(b-a)}.$$

Given a random variable X, its moment generating function is sometimes denoted by $M_X(t)$ and its cumulative distribution function by $F_X(x)$. We know already that $F_X(x)$ contains all information about X, that is, we can answer all probability questions about X using $F_X(x)$, see Lemma 95. The surprising fact is that $M_X(x)$ also contains all the information about X. That is, the moment generating function determines uniquely the cumulative distribution function.

Theorem 123 (Uniqueness theorem)**.** Suppose that random variables X and Y have moment generating functions $M_X(t)$ and $M_Y(t)$ respectively. If $M_X(t) = M_Y(t)$ for all values of t, then X and Y have the same cumulative distribution functions, that is $F_X(x) = F_Y(x)$ for all values of x.

The proof of the Uniqueness Theorem is difficult and will be omitted.

9 Important distributions

9.1 The Poisson Distribution

Definition 124. A random variable X taking the values $\{0,1,2,3,\dots\}$ is called a *Poisson random variable with parameter* $\lambda > 0$ if

$$p(i) := P(X=i) = e^{-\lambda}\frac{\lambda^i}{i!}, \text{ for } i = 0,1,2,\dots.$$

The above definition indeed defines a probability mass function since

$$\sum_{i=0}^{\infty} p(i) = \sum_{i=0}^{\infty} e^{-\lambda}\frac{\lambda^i}{i!} = e^{-\lambda}\sum_{i=0}^{\infty}\frac{\lambda^i}{i!} = e^{-\lambda}e^{\lambda} = 1.$$

The Poisson random variable has a tremendous range of applications in diverse areas because it may be used as an approximation for a binomial random variable with parameters (n,p) when n is large and p is small enough so that np is a moderate size. We will prove that shortly. The Poisson random variable is very frequently used to count data when there is no upper limit on the count. For example, it is used to model the number of customers arriving each day; the number of decays in a radioactive material; the number of cases of a rare disease in a large population; the number of vacancies occurring during a year in the Supreme Court; the number of typos on a page from a book; and so on.

Theorem 125. Let X be a Poisson random variable with parameter λ, then its moment generating function is $M(t) = e^{\lambda(e^t-1)}$. Consequently, we have $E[X] = \lambda$ and $\text{Var}\,[X] = \lambda$.

Proof. For the moment generating function, we have

$$M(t) = E[e^{tX}] = \sum_{k=0}^{\infty} e^{tk} e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{(e^t\lambda)^k}{k!} = e^{-\lambda}e^{\lambda e^t} = e^{\lambda(e^t-1)}.$$

Differentiate $M(t)$ two times and after each differentiation set $t = 0$ to find the first two moments of X:

$$M'(t) = e^{\lambda(e^t-1)}\lambda e^t \text{ and } M''(t) = e^{\lambda(e^t-1)}(\lambda e^t)^2 + e^{\lambda(e^t-1)}\lambda e^t.$$

Hence $E[X] = M'(0) = \lambda$ and $E[X^2] = M''(0) = \lambda^2 + \lambda$. Now we can compute the variance

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

□

There are two ways of looking at a Poisson random variable and these two ways are tightly connected with each other. The first way is that a Poisson random variable is the limit of a sequence of binomial random variables whose parameters satisfy a condition. Imagine we run a binomial experiment with n trials and probability of success on each trial is p_n. Let X be the number of successes in these n trials, and say we are interested in the probability that there are k success, $P(X = k) = \binom{n}{k}p_n^k(1-p_n)^{n-k}$. When the number of trials n in the binomial experiment gets large it will become more and more difficult to calculate the probability $P(X = k)$. Fortunately, if at the same time the probability of success p_n declines so that the product np_n converges to a a finite number λ, then the probability $P(X = k)$ also converges to the nice formula of the Poisson distribution.

We need the following lemma which is a standard fact from Calculus, we include it without a proof.

Lemma 126. If $\lim_{n\to\infty} a_n = a \in \mathbb{R}$ then $\lim_{n\to\infty}\left(1 + \frac{a_n}{n}\right)^n = e^a$.

The next theorem shows that a Poisson random variable with parameter λ may be used as an approximation for a binomial random variable with parameters (n, p) when n is large and p is small enough and so that np is approximately equal to λ. This is our first limiting theorem. At the end of this course we will encounter more limiting theorems.

Theorem 127 (Poisson's theorem)**.** *Let $\lambda > 0$ and suppose that the product np_n converges to λ as n approaches infinity. Then*

$$\lim_{n\to\infty}\binom{n}{k}p_n^k(1-p_n)^{n-k} = e^{-\lambda}\frac{\lambda^k}{k!} \quad \textit{for all } k = 0, 1, 2, ...$$

Proof. Fix an integer $k \geq 0$. Then

$$\binom{n}{k}p_n^k(1-p_n)^{n-k} = \frac{n(n-1)\cdots(n-k+1)}{k!}p_n^k(1-p_n)^{n-k}$$

$$
\begin{aligned}
&= \frac{1}{k!}\frac{n(n-1)\cdots(n-k+1)}{n^k}(np_n)^k(1-p_n)^{n-k} \\
(14) \qquad &= \frac{1}{k!}\Big[(1)\Big(1-\frac{1}{n}\Big)\cdots\Big(1-\frac{k-1}{n}\Big)\Big](np_n)^k\Big(1-\frac{np_n}{n}\Big)^n(1-p_n)^{-k}.
\end{aligned}
$$

Taking the limit at n goes to infinity, note that $\lim_{n\to\infty}(np_n)^k = \lambda^k$ and that

$$\lim_{n\to\infty}(1)\Big(1-\frac{1}{n}\Big)\cdots\Big(1-\frac{k-1}{n}\Big)\Big(1-\frac{\lambda}{n}\Big)^{-k} = 1.$$

Next, use Lemma 126 with $a_n := -np_n$ to conclude that

$$\lim_{n\to\infty}\Big(1-\frac{np_n}{n}\Big)^n = e^{-\lambda}.$$

Finally, since np_n converges to the finite number λ as n approaches infinity, we conclude that p_n converges to 0 as n approaches infinity. Hence

$$\lim_{n\to\infty}(1-p_n)^n = 1.$$

Taking the limit as n approaches infinity in (14) and substituting the last four limits into (14) concludes the proof. □

The above theorem explains why the Poisson distribution is frequently used. Suppose that there is a small probability p that each letter typed on a page will be misprinted and let there be n letters per page. Hence the number of misprints on a page will have a binomial distribution, but since p is small and n is large by the Poisson theorem, the binomial distribution will be approximately Poisson with parameter $\lambda = np$.

As another example, suppose each person in a city has small probability of reaching age 100. Also, each person entering a store may be thought of as having some small probability of buying a package of dog biscuits, and so on.

The condition "np_n converges to λ as n approaches infinity" is just the mean, np_n, of the binomial distribution converging to the mean, λ, of the Poisson distribution.

Example 128. Consider an experiment that consists of counting the number of α-particles given off in a 1-second interval by 1 gram of radioactive material. If we know from past experience that, on the average, 3.2 such α-particles are given off, what is a good approximation to the probability that no more than 2 α-particles will appear?

Solution. Think of the gram of radioactive material as consisting of a large number n of atoms, each of which has equal small probability of disintegrating and sending off an α-particle in one second, then we see that, to a very close approximation, the number of α-particles given off will be a Poisson random variable with parameter $\lambda = 3.2$. Hence the desired probability is

$$P(X \le 2) = e^{-3.2} + 3.2e^{-3.2} + e^{-3.2}\frac{(3.2)^2}{2} \approx 0.3799.$$

We now describe the second way of looking at a Poisson random variable.

Definition 129 (Poisson process)**.** Suppose that events are occurring at random points of time (or space), and assume that for a constant $\lambda > 0$ the following hold true:

(i) For any interval $(a, a+h)$ of length h

$$\lim_{h\to 0} \frac{P(\text{exactly 1 event occurs in } (a, a+h))}{h} = \lambda. \tag{15}$$

(ii) Events occur independently of each other in non-overlapping time (or space) intervals.

In that case, we say that the events occur according to a *Poisson process* with *rate* λ.

The first condition in the definition of a Poisson process says that when h is close to 0, we have $P(\text{exactly 1 event occurs in } (a, a+h)) \approx \lambda h$. More precisely, the limit (15) says that

$$P(\text{exactly 1 event occurs in } (a, a+h)) = \lambda h + o(h), \tag{16}$$

where the function $o(h)$ is such that $o(h)/h$ approaches 0 as h approaches 0. That is, all that we know about the function $o(h)$ is that is makes equality (16) hold, and that it approaches 0 much faster than h.

Theorem 130. Suppose an event occurs according to a Poisson process. Let X be the number of times that the event occurred in the time interval $[0, t]$, then

$$P(X = k) = e^{-\lambda t}\frac{(\lambda t)^k}{k!}.$$

Intuitive explanation. To find the probability of k events occurring in a time interval $[0, t]$, divide the interval $[0, t]$ into n non-overlapping subintervals each of length t/n. Then

$$P(\text{exactly 1 event occurs in one of the subintervals}) = \lambda(t/n) + o(t/n).$$

By the second condition, the events occur independently in the different subintervals so, letting $p_n := \lambda(t/n) + o(t/n)$, we have

$$P(\text{exactly } k \text{ events occurs in } [0, t] \text{ with at most 1 event in each subinterbal}) = b(n, p_n, k). \tag{17}$$

Thus,

$$P(\text{exactly } k \text{ events occurs in } [0, t]) = \lim_{n\to\infty} b(n, p_n, k). \tag{18}$$

Now

$$np_n = n\big(\lambda(t/n) + o(t/n)\big) = \lambda t + no(t/n) = \lambda t + t\frac{o(t/n)}{t/n}.$$

The property of the $o(h)$ function implies that $\frac{o(t/n)}{t/n}$ approaches 0 as n goes to infinity. (Note that in that case t/n approaches 0.) Thus, np_n approaches λt as n goes to infinity. Using the Poisson theorem, we take the limit in (18) as n approaches infinity, to conclude that

$$P(\text{exactly } k \text{ events occurs in } [0, t]) = \lim_{n\to\infty} b(n, p_n, k) = e^{-\lambda t}\frac{(\lambda t)^k}{k!}. \tag{19}$$

This concludes the intuitive explanation of the theorem. □

We see that in a unit time interval, $t = 1$, the number of events has Poisson distribution with average λ, while in a time interval of length t, the number of events has Poisson distribution with average λt.

Example 131. Suppose that we are told that the number of potholes on Richmond Street follows a Poisson process with a rate of 3 potholes per km.
a) What is the probability that we will see at most 1 pothole on a 3 km stretch of Richmond Street?
b) What is the probability that we will see at least 3 pothole on a 3 km stretch of Richmond Street?

Solution. Let X represent the number of potholes on a 3 km stretch of Richmond Street. The random variable X follows a Poisson distribution with a mean, $\lambda = 3 \times 3 = 9$. Remember that we expected to see 3 for every km, so if we travel 3 km, then we would expect to see 9.

$$P(X = x) = p(x) = \frac{e^{-9}9^x}{x!}, \quad x = 0, 1, 2, \ldots$$

a)

$$\begin{aligned} P(X \le 1) = P(X = 0) + P(X = 1) &= \frac{e^{-9}9^0}{0!} + \frac{e^{-9}9^1}{1!} \\ &= e^{-9}(1 + 9) = 0.0012341 \end{aligned}$$

b)

$$\begin{aligned} P(X \ge 3) &= 1 - \big(P(X = 0) + P(X = 1) + P(X = 2)\big) \\ &= 1 - \frac{e^{-9}9^0}{0!} - \frac{e^{-9}9^1}{1!} - \frac{e^{-9}9^2}{2!} \\ &= 1 - e^{-9}(1 + 9 + 9^2/2) = 0.99377 \end{aligned}$$

Example 132. Suppose that earthquake occurrences in Canada are a Poisson process with average of 2 earthquakes per week.

(a) Find the probability that 3 earthquakes occur during the next 2 weeks.

(b) Find the probability that at least 3 earthquakes occur during the next 2 weeks.

Solution. First we are going to offer the naive approach to solving part a) of this problem.

a) The probability that k earthquakes occur in any week is $p(k) = e^{-2}2^k/k!$. The event that that 3 earthquakes occur during the next 2 weeks is the disjoint union:

$$\begin{aligned} &\{3 \text{ quakes 1st week } \textbf{and } \ 0 \text{ in 2nd week}\} \cup \{2 \text{ quakes 1st week } \textbf{and } \ 1 \text{ in 2nd week}\} \\ &\cup \{1 \text{ quakes 1st week } \textbf{and } \ 2 \text{ in 2nd week}\} \cup \{0 \text{ quakes 1st week } \textbf{and } \ 3 \text{ in 2nd week}\}. \end{aligned}$$

By part (ii) in Definition 129, taking probabilities we get

$$P(\text{ 3 quakes occur in next 2 weeks}) = e^{-2}\frac{2^3}{3!}e^{-2}\frac{2^0}{0!} + e^{-2}\frac{2^2}{2!}e^{-2}\frac{2^1}{1!} + e^{-2}\frac{2^1}{1!}e^{-2}\frac{2^2}{2!} + e^{-2}\frac{2^0}{0!}e^{-2}\frac{2^3}{3!}$$

$$= 0.19536681.$$

The last computation can easily become huge and much more cumbersome if the complexity of the problems increases slightly. So, an alternative approach is to use Theorem 130, to observe that since the earthquakes follow a Poisson process, the average occurrences in a two week period is $2 \times 2 = 4$. Thus,

$$P(\text{ 3 quakes occur in next 2 weeks}) = e^{-4}\frac{4^3}{3!} = 0.19536681.$$

Note that both approaches give the same answer (why?).

b) Let X be the number of earthquakes in the next two weeks. X has a Poisson distribution with average $2\lambda = 4.$. Hence, $P(X \geq 3) = 1 - P(X=0) - P(X=1) - P(X=2) = 1 - e^{-4} - 4e^{-4} - \frac{4^2}{2}e^{-4} = 1 - 13e^{-4}$.

Example 133. Suppose that earthquake occurrences in Canada are a Poisson process with average of λ earthquakes per week. Let Y be a random variable measuring the time starting from now, until the next earthquake. Find the cumulative distribution function of Y.

Solution. Let Y denote the amount of time in weeks until the next earthquake. Note that, Y is greater than t if and only if no earthquake occurs in the next t weeks. The number of earthquakes in the next t weeks, denote that number by Z, has a Poisson distribution with average λt. Thus

$$P(Y > t) = P(Z = 0) = e^{-\lambda t},$$

and

$$P(Y \leq t) = 1 - P(Y > t) = 1 - e^{-\lambda t}.$$

Example 133 is very important. It shows that if an event occurs according to a Poisson process with parameter λ, then the time Y between two consecutive occurrences of the event is a random variable with cdf $F(t) = 1 - e^{-\lambda t}$. The next subsection is devoted to those random variables.

The final issue that we want to address is how good is the approximation in the Poisson theorem. An estimate of the error of the approximation is given without proof in the next result.

Theorem 134. The following bound holds

$$\Big|\binom{n}{k}p_n^k(1-p_n)^{n-k} - e^{-np_n}\frac{(np_n)^k}{k!}\Big| \leq np_n^2.$$

The last theorem says that if np_n is close to λ then $e^{-np_n}\frac{(np_n)^k}{k!}$ will be close to $e^{-\lambda}\frac{\lambda^k}{k!}$. That is, the last expression will be a good approximation for $\binom{n}{k}p_n^k(1-p_n)^{n-k}$ with the error being about $np_n^2 \approx \lambda^2/n$. In short, the Poisson approximation is good, when np_n is not a large number, but n is.

If np_n is a large number, then the binomial distribution is best approximated by a normal distribution, as we will see in Theorem 156.

9.2 The exponential distribution

Definition 135. A continuous random variable X is called *exponential* if its probability density function is given by

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

for some positive constant λ. We say that X is *exponentially distributed with parameter* λ.

First note that the integral of the probability density function is 1. Indeed

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{0}^{\infty} \lambda e^{-\lambda x}\,dx = -e^{-\lambda x}\Big|_{x=0}^{\infty} = 1.$$

The cumulative distribution function of an exponential random variable is

$$F(t) = P(X \leq t) = \int_{0}^{t} \lambda e^{-\lambda x}\,dx = -e^{-\lambda x}\Big|_{x=0}^{t} = 1 - e^{-\lambda t},$$

if $t \geq 0$, and $F(t) = 0$ if $t < 0$.

The moment generating function of an exponential random variable is

$$M(t) = E[e^{tX}] = \int_{0}^{\infty} e^{tx}\lambda e^{-\lambda x}\,dx = \frac{\lambda}{t-\lambda} e^{(t-\lambda)x}\Big|_{x=0}^{\infty} = \begin{cases} -\frac{\lambda}{t-\lambda} & \text{if } t < \lambda \\ \infty & \text{if } t \geq \lambda. \end{cases}$$

The first two derivatives of $M(t)$ are

$$M'(t) = \frac{\lambda}{(t-\lambda)^2} \text{ and } M''(t) = -\frac{2\lambda}{(t-\lambda)^3}.$$

Hence, we have

$$M'(0) = \frac{1}{\lambda} \text{ and } M''(0) = \frac{2}{\lambda^2}$$

and from here

$$E[X] = M'(0) = \frac{1}{\lambda} \text{ and } \mathrm{Var}\,[X] = E[X^2] - (E[X])^2 = M''(0) - (M'(0))^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}.$$

As shown in Example 132, the exponential distribution is the distribution of the amount of time until a specific event in a Poisson process occurs. For exmple, the amount of time (starting from now) until an earthquake occurs, or until a new war breaks out, or until a telephone call you receive turns out to be a wrong number are all random variables that tend in practice to have exponential distributions. The examples are all about an amount of time, but the variable may also measure distance, area, and so on. For example, the length of road until the next pothole may be a exponentially distributed random variable.

Example 136. The length of a phone call in minutes is an exponential random variable with parameter $\lambda = 0.1$. Suppose someone arrives immediately ahead of you at a public telephone booth. What is the probability that you will have to wait (a) more than 10 minutes; (b) between 10 and 20 minutes.

Solution. Let X be the amount of time it takes the person in the booth to finish their phone call. (a) $P(X > 10) = 1 - P(X \leq 10) = e^{-0.1 \cdot 10} \approx 0.37$. (b) $P(10 < X < 20) = F(20) - F(10) = e^{-1} - e^{-2} = 0.23$.

Theorem 137. The exponential distribution is the only one non-negative continuous distribution with the following *memoryless* property

$$P(X > x + y | X > y) = P(X > x) \text{ for any } x, y \geq 0.$$

Proof. Using the definition of conditional probability we calculate

$$\begin{aligned} P(X > x+y|X > y) &= \frac{P(\{X > x+y\} \cap \{X > y\})}{P(X > y)} = \frac{P(X > x+y)}{P(X > y)} = \frac{1 - F(x+y)}{1 - F(y)} \\ &= \frac{e^{\lambda(x+y)}}{e^{\lambda y}} = e^{\lambda x} = 1 - F(x) = P(X > x). \end{aligned}$$

The only thing left to prove is that the exponential distribution is the only non-negative one with this property. We omit that part of the proof. □

Suppose that the lifetime of a randomly selected light bulb is exponentially distributed random variable X. The memoryless property says that the probability that the light bulb works for at least $x + y$ years, given that it has survived y years, is the same as the initial probability that it survives for at least x years. In other words, if the light bulb is working after y years, the distribution of the remaining amount of time that it will work is the same as the original lifetime distribution (that is, it is as if the light bulb does not remember that it has already been in use for a time y).

Example 138. The number of miles that a car can run before its battery wears out is exponentially distributed with an average value of $10,000$ miles. If a person desires to take a $5,000$-mile trip, what is the probability that he or she will be able to complete the trip without having to replace the battery?

Solution. By the memoryless property of the exponential distribution, the remaining lifetime (in thousands of miles) of the battery is exponential with parameter $\lambda = 1/10$, no matter how many miles the car has on it now with the current battery. The desired probability is $P(\text{remaining lifetime } > 5) = 1 - F(5) = e^{-5\lambda} \approx 0.6$. □

9.3 The geometric distribution

Definition 139. Let $p \in (0, 1)$ be a fixed number. A discrete random variable X taking values $\{1, 2, 3, \ldots\}$ with probabilities $P(X = k) = (1-p)^{k-1}p$ is called a geometric random variable with parameter p.

Note that the probability mass function sums up to 1. Indeed

$$\sum_{k=1}^{\infty} P(X = k) = \sum_{k=1}^{\infty} (1-p)^{k-1}p = p\sum_{k=1}^{\infty}(1-p)^{k-1} = \frac{p}{1-(1-p)} = 1.$$

Geometric random variables occur most often when performing independent trials, each having a probability p, $p \in (0,1)$, of being a success. Let X equal the number of trials required until a success occurs. Then X is a geometric random variable with parameter p. (Why?) The geometric distribution is the discrete equivalent of the exponential distribution.

The moment generating function is

$$M(t) = E[e^{tX}] = \sum_{k=1}^{\infty} e^{tk}(1-p)^{k-1}p = pe^t \sum_{k=1}^{\infty} \left(e^t(1-p)\right)^{k-1} = \begin{cases} \frac{pe^t}{1-e^t(1-p)} & \text{if } e^t(1-p) < 1 \\ \infty & \text{if } e^t(1-p) \geq 1. \end{cases}$$

Note that, taking logarithm on both sides of the condition $e^t(1-p) < 1$ makes it into $t < -\log(1-p)$. This means that we can differentiate $M(t)$ for values of t close to 0:

$$M'(t) = \frac{e^t p}{(1-e^t(1-p))^2} \text{ and } M''(t) = \frac{e^t p(1+e^t(1-p))}{(1-e^t(1-p))^3}.$$

From here we get

$$E[X] = M'(0) = \frac{1}{p} \text{ and } \mathrm{Var}\,[X] = E[X^2] - (E[X])^2 = M''(0) - (M'(0))^2 = \frac{1-p}{p^2}.$$

Example 140. An urn contains n white and m black balls. Balls are randomly selected, one at a time with replacement, until a black one is obtained. Then the number of draws is a geometric random variable with parameter $p = m/(n+m)$.

The geometric distribution also has the memoryless property.

Lemma 141. The geometric distribution has the *memoryless* property

$$P(X = k+\ell | X > k) = P(X = \ell) \text{ for any non-negative integers } k, \ell \geq 1.$$

Proof. We only need to calculate the conditional probability:

$$\begin{aligned} P(X = k+\ell|X > k) &= \frac{P(\{X = k+\ell\} \cap \{X > k\})}{P(X > k)} = \frac{P(X = k+\ell)}{P(X > k)} = \frac{(1-p)^{k+\ell-1}p}{\sum_{i=k+1}^{\infty}(1-p)^{i-1}p} \\ &= \frac{(1-p)^{k+\ell-1}p}{(1-p)^k p \sum_{i=0}^{\infty}(1-p)^i} = \frac{(1-p)^{k+\ell-1}p}{(1-p)^k p[1/(1-(1-p))]} = \frac{(1-p)^{k+\ell-1}p}{(1-p)^k} \\ &= (1-p)^{\ell-1}p = P(X = \ell). \end{aligned}$$

□

Corollary 142. The geometric distribution has the *memoryless* property

$$P(X > k+m|X > k) = P(X > m) \text{ for any } k, \ell \geq 1.$$

Proof. Sum the equalities $P(X = k+\ell|X > k) = P(X = \ell)$ for all $\ell = m+1, m+2, \ldots$ to obtain:

$$P(X > k+m|X > k) = \sum_{\ell=m+1}^{\infty} P(X = k+\ell|X > k) = \sum_{\ell=m+1}^{\infty} P(X = \ell) = P(X > m).$$

□

One should compare the last corollary with Theorem 137. It says that if you are wayting for the first success, it does not matter how long you have already been waiting.

9.4 The negative binomial distribution

Definition 143. Let $p \in (0,1)$ be a fixed number and let r be a fixed positive integer. A discrete random variable X taking values $\{r, r+1, r+2, r+3, \ldots\}$ with probabilities $P(X=k) = \binom{k-1}{r-1}(1-p)^{k-r}p^r$ is called a negative binomial random variable with parameters r and p.

Negative binomial random variables occur most often when performing independent trials, each having a probability p, $p \in (0,1)$, of being a success. Here is what I mean. Let X equal the number of trials required until r successes occur. Then X is a negative binomial random variable with parameters r and p. (Why?)

Note that when $r = 1$, the negative binomial distribution becomes geometric distribution.

The number of trials needed to obtain r successes can be expressed as $Y_1 + Y_2 + \cdots + Y_r$ where Y_1 equals the number of trials required for the first success, Y_2 the number of additional trials after the first success until the second success occurs, Y_3 the number of additional trials until the third success, and so on. As the trials are independent and all have the same probability of a success, it follows that Y_1, $Y_2, \ldots, Y_r$ are all geometric random variables and *independent*. (We will talk about independent random variables later, but at the moment we will use two of their properties to shorten the presentation.)

Theorem 144. a) If Y_1, $Y_2, \ldots, Y_r$ are independent random variables and $g : \mathbb{R} \to \mathbb{R}$ is a function, then $g(Y_1)$, $g(Y_2), \ldots, g(Y_r)$ are also independent random variables.

b) If Z_1, $Z_2, \ldots, Z_r$ are independent random variables then $E[Z_1Z_2\cdots Z_r] = E[Z_1]E[Z_2]\cdots E[Z_r]$.

Let X be a negative binomial random variable. Then $X = Y_1 + Y_2 + \cdots + Y_r$, where the latter are independent geometric random variables. The moment generating function of X is

$$\begin{aligned} M(t) &= E[e^{tX}] = E[e^{t(Y_1+Y_2+\cdots+Y_r)}] = E[e^{tY_1}e^{tY_2}\cdots e^{tY_r}] = E[e^{tY_1}]E[e^{tY_2}]\cdots E[e^{tY_r}] \\ &= \begin{cases} \left(\frac{pe^t}{1-e^t(1-p)}\right)^r & \text{if } e^t(1-p) < 1 \\ \infty & \text{if } e^t(1-p) \geq 1, \end{cases} \end{aligned}$$

where we used Theorem 144, part a) with $g(x) = e^{tx}$, then part b), and the formula for the moment generating function of a geometric random variable.

Now, notice that

$$M(0) = E[1] = \sum_{k=r}^{\infty} P(X=k) = \sum_{k=r}^{\infty} \binom{k-1}{r-1}(1-p)^{k-r}p^r.$$

Since, from the formula for $M(t)$, we have $M(0) = 1$, we proved that the sum of the probabilities of a negative binomial random variable is 1.

Next, we differentiate $M(t)$ for values of t close to 0:

$$M'(t) = \Big(\frac{e^tp}{1-e^t(1-p)}\Big)^r \frac{r}{1-e^t(1-p)} \text{ and } M''(t) = \Big(\frac{e^tp}{1-e^t(1-p)}\Big)^r \frac{r(r+e^t(1-p))}{(1-e^t(1-p))^2}.$$

From here we get

$$E[X] = M'(0) = \frac{r}{p},$$

$$
\begin{aligned}
E[X^2] = M''(0) &= \frac{r(r+1-p)}{p^2}, \text{ and} \\
\operatorname{Var}[X] = E[X^2] - (E[X])^2 &= \frac{r(1-p)}{p^2}.
\end{aligned}
$$

Example 145 (The Banach match problem)**.** A chain-smoking man carries 2 matchboxes, one in his left-hand pocket and one in his right-hand pocket. Each matchbox contains N matches. Every time he lights up a cigarette he takes a match from a random pocket. Consider the moment when the man first discovers that one of his matchboxes is empty. What is the probability that there are exactly ℓ matches in the other box, $\ell = 0, 1, \ldots, N$?

Solution. Each pocket is equally likely to be selected. Let Y be the number of matches left in one pocket when the other one is discovered to be empty. The problem asks us to find $P(Y = \ell)$. The event $\{Y = \ell\}$ is the union of two disjoint events.

$$\{Y = \ell\} = \{Y = \ell \text{ \& the right pocket is discovered empty}\} \cup \{Y = \ell \text{ \& the left pocket is discovered empty}\}.$$

Let us look at the first event first. The event $\{Y = \ell \text{ \& the right pocket is empty}\}$ occurs precisely when the left pocket was selected $N - \ell$ times (that many matches were removed from it) and the right pocket was selected $N+1$ times (N matches are removed from it and then on the $N+1$ time it was discovered that it was empty). Selecting a pocket is a binomial trial and let us say that a success is when the right pocket is selected. Thus, the event $\{Y = \ell \text{ \& the right pocket is empty}\}$ occurs exactly when the $(N+1)$-th success occurs on the $(N-\ell)+(N+1) = 2N-\ell+1$ trial. Let X be a negative binomial random variable with parameters $p = 1/2$ and $r = N+1$. Then,

$$P(Y = \ell \text{ \& the right pocket is discovered empty}) = P(X = 2N-\ell+1) = \binom{2N-\ell}{N}\left(\frac{1}{2}\right)^{2N-\ell+1}.$$

Clearly, by symmetry

$$P(Y = \ell \text{ \& the right pocket is discovered empty}) = P(Y = \ell \text{ \& the left pocket is discovered empty}).$$

So,

$$P(Y = \ell) = 2P(Y = \ell \text{ \& the right pocket is discovered empty}) = \binom{2N-\ell}{N}\left(\frac{1}{2}\right)^{2N-\ell}.$$

9.5 The hypergeometric distribution

Example 146. You have an urn with 3 white and 4 black balls. You pick two balls at random (**with** replacement). What is the probability that you picked 2 white balls?

The probability that the first ball is white is 3/7. The probability that the second ball is white is 3/7. So the answer is: $\left(\frac{3}{7}\right)\left(\frac{3}{7}\right) = \frac{9}{49}$. If you consider that the two trials (selection of balls) are independent, then the probability of success in each trial is the same. This is a binomial experiment. □

Example 147. You have an urn with 3 white and 4 black balls. You pick two balls at random (**without** replacement). What is the probability that you picked 2 white balls?

The probability that the first ball is white is 3/7. The probability that the second ball is white is 2/6. So the answer is: $\left(\frac{3}{7}\right)\left(\frac{2}{6}\right) = \frac{1}{7}$. The probability of success in each trial is different. This is a hypergeometric experiment! □

Example 148. You have an urn with 3 white and 4 black balls. You pick two balls at random (**without** replacement). What is the probability that you picked exactly one white balls?

We have to compute $p(wb) + p(bw) =$? The probability that the first ball is white is 3/7. The probability that the second ball is black is 4/6. So $p(wb) = \left(\frac{3}{7}\right)\left(\frac{4}{6}\right) = \frac{2}{7}$.

The probability that the first ball is black is 4/7. The probability that the second ball is white is 3/6. So $p(bw) = \left(\frac{4}{7}\right)\left(\frac{3}{6}\right) = \frac{2}{7}$.

Answer: $p(wb) + p(bw) = \frac{2}{7} + \frac{2}{7} = \frac{4}{7}$. □

Consider the experiment of choosing a random sample of size n (without replacement) from an urn containing N balls, of which m are white and $N - m$ are black. Let X equal the number of white balls selected. Clearly, we must require that

$$1 \le n \le N \text{ and } 0 \le m \le N.$$

What are the possible values of X? X cannot be bigger than n or m, that is,

$$X \le \min\{n, m\}.$$

Next, if X is the number of white balls in the sample, then $X \ge 0$, and the black balls are $n - X$. Hence $n - X \le N - m$, that is $n + m - N \le X$. Putting the lower bounds together we have

$$\max\{0, n + m - N\} \le X.$$

Now, let $k \in \{\max\{0, n + m - N\}, \ldots, \min\{n, m\}\}$, the probability that there are k white balls in the sample is (why?)

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

Extending the definition of the binomial coefficients

$$\binom{a}{b} := 0 \text{ if } a < b \text{ or if } b < 0,$$

we have

$$P(X = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} \text{ for all } k = 0, 1, \ldots, n.$$

(Indeed, if $k < \max\{0, n+m-N\}$ then $\binom{m}{k}\binom{N-m}{n-k} = 0$ and hence $P(X = k) = 0$; and if $\min\{n, m\} < k$ then again $\binom{m}{k}\binom{N-m}{n-k} = 0$ and hence $P(X = k) = 0$ again.) Since, at the end of the experiment, X will take precisely one value in $\{0, 1, \ldots, n\}$ with certainty, we have

$$\sum_{k=0}^{n} \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} = 1.$$

We make the following definition.

Definition 149. Let n, m, N be positive integers satisfying $0 \le m \le N$ and $1 \le n \le N$. A discrete random variable X taking values $\{0, 1, \dots, n\}$ with probabilities

$$P(X = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} \text{ for all } k = 0, 1, \dots, n.$$

is called a *hypergeometric* random variable with parameters n, N, and m.

The mean of the hypergeometric distribution is computed later in Example 172. Again both the mean and the variance are computed in (36) on page 106. For the moment, we state them without a proof:

$$E[X] = \frac{nm}{N} \text{ and } \mathrm{Var}\,[X] = \frac{N-n}{N-1}\frac{nm}{N}\frac{N-m}{N}. \tag{20}$$

9.6 The normal distribution

The normal distribution, also known as the Bell Curve, is the most frequently used distribution in statistics because of the central limit theorem, that we will study later. The central limit theorem, is one of the two most important results in probability theory. It gives a theoretical base to the often noted empirical observation that, in practice, many random phenomena obey, at least approximately, a normal probability distribution. Some examples of this behaviour are the height of a man, the velocity in any direction of a molecule in gas, and the error made in measuring a physical quantity.

Definition 150. A continuous random variable X is *normal*, or simply that X is *normally distributed*, with parameters μ and σ^2 if the probability density function of X is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \text{ for all } x \in \mathbb{R}. \tag{21}$$

To denote that X is a normal random variable with parameters μ and σ^2, we write $X \sim N(\mu, \sigma^2)$.

This density function is a bell-shaped curve that is symmetric about μ, that is, $f(\mu+t) = f(\mu-t)$ for all $t \in \mathbb{R}$. Before we establish that $f(x)$ is indeed a probability density function. We take care of a technical lemma.

Lemma 151. *The following identity holds.*

$$\int_{-\infty}^{\infty} e^{-\frac{y^2}{2}}\, dy = \sqrt{2\pi}.$$

Proof. Denote the integral by I:

$$I := \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}}\, dy.$$

We perform a trick to calculate the value of I, defined by the last integral.

$$I^2 = \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}}\, dy \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}}\, dz = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{y^2+z^2}{2}}\, dydz.$$

Evaluate the double integral by means of a change of variables to polar coordinates. That is, let $y = r\sin\theta$ and $z = r\cos\theta$. The Jacobian of this transformation is r, so $dydz = rd\theta dr$. The region over which we integrate $(-\infty,\infty)\times(-\infty,\infty)$ is transformed into $(0,\infty)\times[0,2\pi)$. We continue

$$I^2 = \int_0^\infty \int_0^{2\pi} e^{-\frac{r^2}{2}} r\, d\theta dr = 2\pi \int_0^\infty e^{-\frac{r^2}{2}} r\, dr = 2\pi\Big(- e^{-\frac{r^2}{2}}\Big|_{r=0}^{\infty}\Big) = 2\pi.$$

Since $I \geq 0$, we get that $I = \sqrt{2\pi}$. □

Lemma 152. Function (21) is indeed a probability density function.

Proof. Clearly $f(x) \geq 0$ for all x. We only need to show that $\int_{-\infty}^{\infty} f(x)\, dx = 1$, that is, that

$$\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx = 1.$$

After the change of variables $y = (x-\mu)/\sigma$ (implying that $dx = \sigma dy$), and using Lemma 151, we get

$$\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{y^2}{2}}\, dy = 1.$$

□

The cumulative distribution function is

$$F(x) := \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx \tag{22}$$

and there is no closed form formula for this integral in the case of the normal density function. Its values are typically obtained from pre-calculated tables or with the use of computers.

The moment generating function of a normal random variable with parameters μ and σ^2 is

$$M(t) = e^{\left(\mu t + \frac{\sigma^2 t^2}{2}\right)}.$$

Its first two derivatives are

$$M'(t) = e^{\left(\mu t + \frac{\sigma^2 t^2}{2}\right)}(\mu + \sigma^2 t),$$
$$M''(t) = e^{\left(\mu t + \frac{\sigma^2 t^2}{2}\right)}\big((\mu + \sigma^2 t)^2 + \sigma^2\big).$$

Using them, we can find the expected value and the variance.

$$E[X] = M'(0) = \mu,$$
$$E[X^2] = M''(0) = \mu^2 + \sigma^2,$$
$$\operatorname{Var}[X] = E[X^2] - (E[X])^2 = \sigma^2.$$

Lemma 153. If X is normally distributed with parameters μ and σ^2, then $Y := aX + b$ is normally distributed with parameters $a\mu + b$ and $a^2\sigma^2$, provided that $a \neq 0$.

Proof. We only deal with the case $a > 0$, since the case $a < 0$ is similar. Let $F_X(x)$ and $F_Y(x)$ be the c.d.f.'s of X and Y respectively. Then

$$F_Y(x) = P(Y \le x) = P(aX + b \le x) = P\Big(X \le \frac{x-b}{a}\Big) = F_X\Big(\frac{x-b}{a}\Big).$$

To obtain the probability density of Y, we differentiate both sides with respect to x, using Theorem 96:

$$f_Y(x) = \frac{1}{a} f_X\Big(\frac{x-b}{a}\Big) = \frac{1}{\sqrt{2\pi}a\sigma} e^{-\frac{(\frac{x-b}{a}-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}a\sigma} e^{-\frac{(x-(a\mu+b))^2}{2(a\sigma)^2}}.$$

This shows that Y is normally distributed with parameters $a\mu + b$ and $a^2\sigma^2$. □

Important corollary of the last lemma is that if X is normally distributed with parameters μ and σ^2, then $Z := (X - \mu)/\sigma$ is normally distributed with parameters 0 and 1. And vice versa, if Z is normally distributed with parameters 0 and 1, then $\sigma Z + \mu$ is normally distributed with parameters μ and σ^2. A normally distributed random variable Z with parameters 0 and 1 is called *standard normal*. Its probability density function is

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

Note that $f_Z(x) = f_Z(-x)$ showing that the standard normal p.d.f. is symmetric with respect to 0 (its mean). Letting $X := \sigma Z + \mu$ be normally distributed with parameters μ and σ^2 the relationship

$$f_X(x) = \frac{1}{\sigma} f_Z\Big(\frac{x-\mu}{\sigma}\Big)$$

(see the proof of Lemma 153) shows that the p.d.f. of X is obtained from that of Z by shifting the latter μ units to the right (if $\mu > 0$ or to the left if $\mu < 0$) and "squishing" it by a factor σ, so the larger σ is the larger the squishing is. In general, the rules are depicted on Figure 2.

Traditionally, the cumulative distribution function of a standard normal random variable is denoted by $\Phi(z)$. That is, we define

$$\Phi(z) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2}\, dx. \tag{23}$$

the values of $\Phi(z)$ can be obtained from tables or using a computer software. There a two types of tables available: *standard normal table* and *cumulative normal table*. The cumulative normal table, see Figure 3, lists the values of the probabilities of the kind $\Phi(z) = P(Z \le z)$ for different values of z. Sometimes, for brevity, cumulative normal tables list only the values of $\Phi(z) = P(Z \le z)$ for positive z. In that case, the values of $\Phi(z)$ for negative z can be obtained using the following identity.

Lemma 154. For any $z \in \mathbb{R}$, we have $\Phi(-z) = 1 - \Phi(z)$.

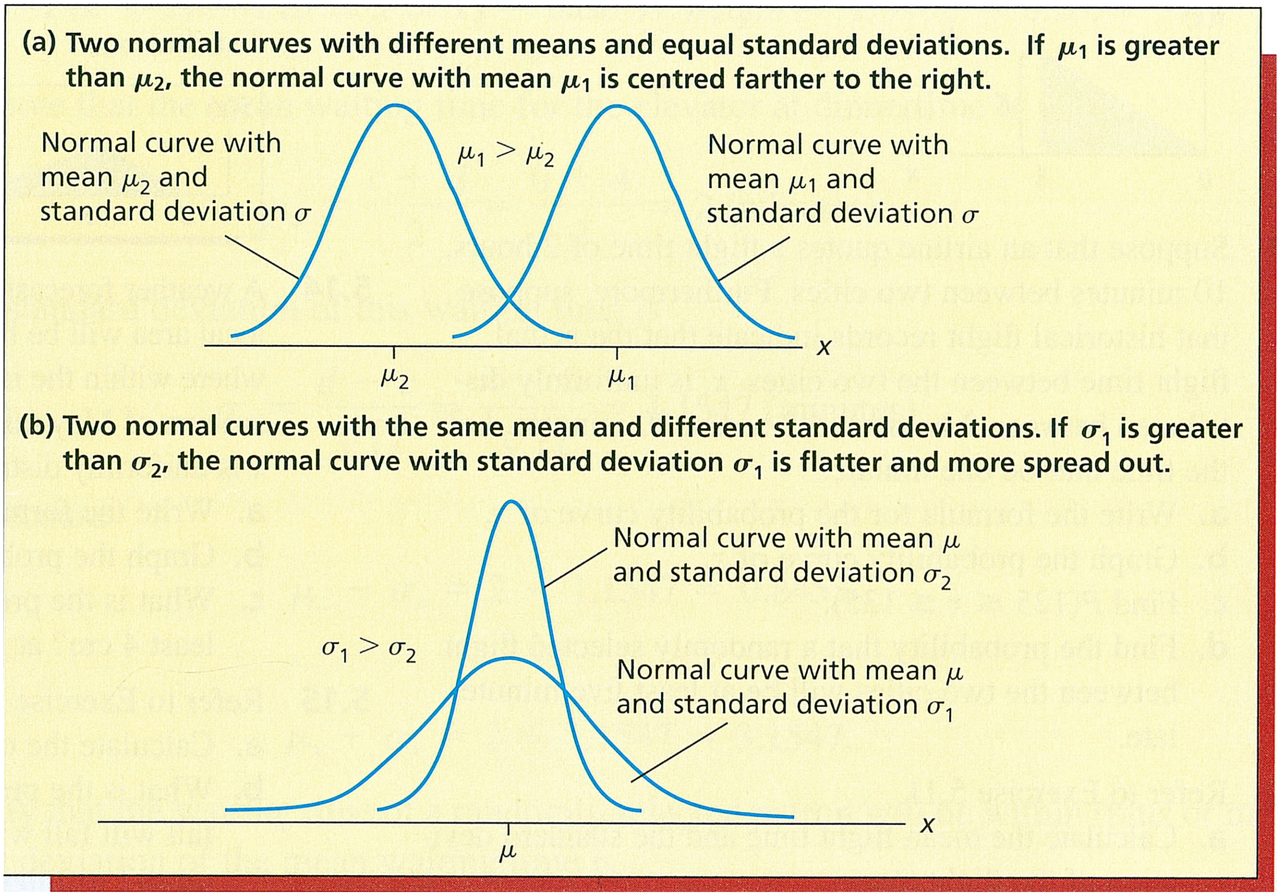
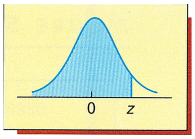


Figure 2: The shape of the normal distribution



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
−3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
−3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
−3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

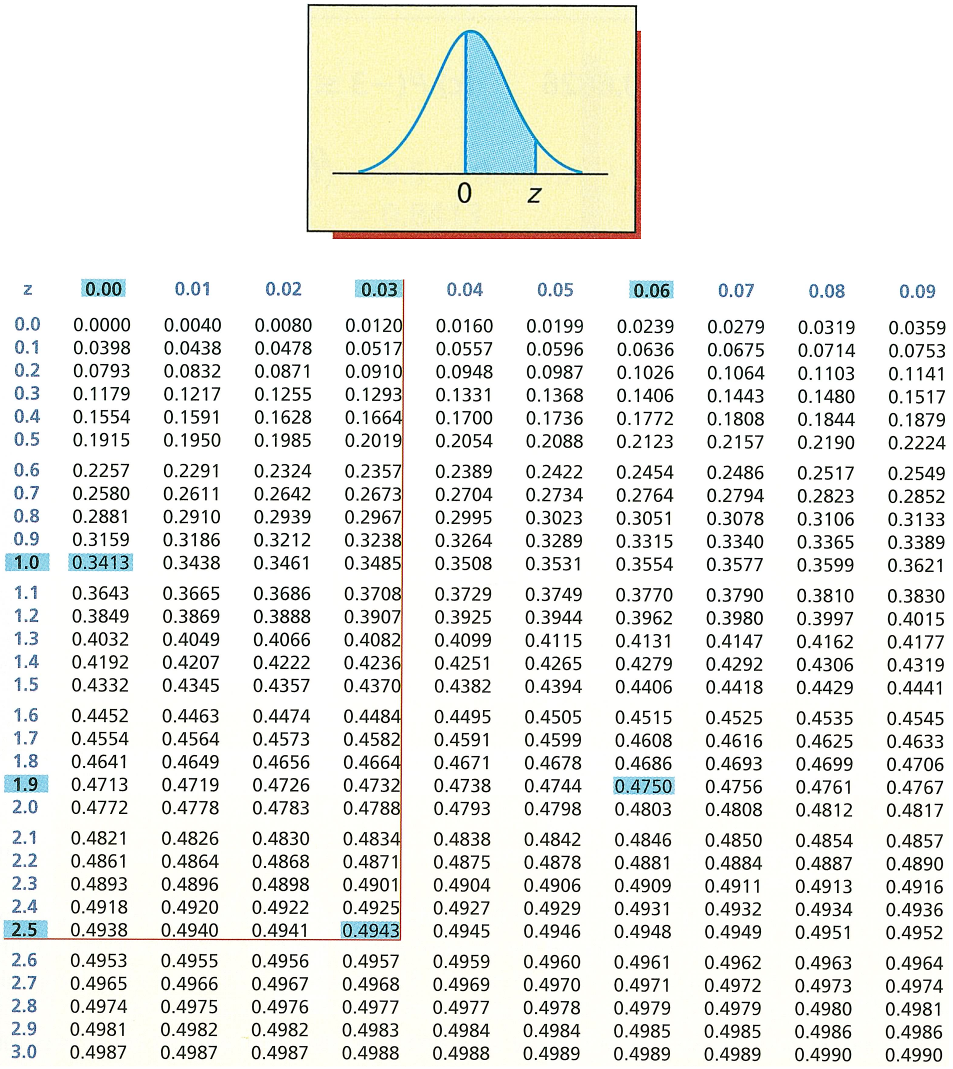
Figure 3: The cumulative normal table

Proof. Using the fact that the p.d.f. of Z is symmetric with respect to 0, we get

$$\Phi(-z) = P(Z \leq -z) = P(Z \geq z) = 1 - P(Z \leq z) = 1 - \Phi(z).$$

□

The Standard normal table, see Figure 4, lists the values of the probabilities of the kind $P(0 \leq Z \leq z)$ for positive values of z. The values of z in the table range from 0.00 to 3.09 in increments of 0.01. z values accurate to tenths are listed in the far left column. The hundredths digit of z is listed across the top of the table. The areas under the normal curve between 0 and z are given in the body of the table.



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Figure 4: The standard normal table

For example, the standard normal table gives us that $P(0 \leq Z \leq 1) = 0.3413$ and that $P(0 \leq Z \leq 1.96) = 0.4750$. Using the symmetry of the distribution function of Z and the fact that

$P(Z \leq 0) = P(0 \leq Z) = 0.5$, Figure 5 shows how to find the probabilities

$$\begin{aligned} P(-1 \leq Z \leq 1) &= P(-1 \leq Z \leq 0) + P(0 \leq Z \leq 1) = 2P(0 \leq Z \leq 1), \\ P(-1 \leq Z) &= P(-1 \leq Z \leq 0) + P(0 \leq Z) = P(0 \leq Z \leq 1) + 0.5 = 0.3413 + 0.5 = 0.8413, \\ P(1 \leq Z) &= P(0 \leq Z) - P(0 \leq Z \leq 1) = 0.5 - 0.3413 = 0.1587. \end{aligned}$$

How about computing probabilities for general normally distributed random variable X with parameters μ and σ^2? We use the fact that then $Z := (X - \mu)/\sigma$ is standard normal random variable.

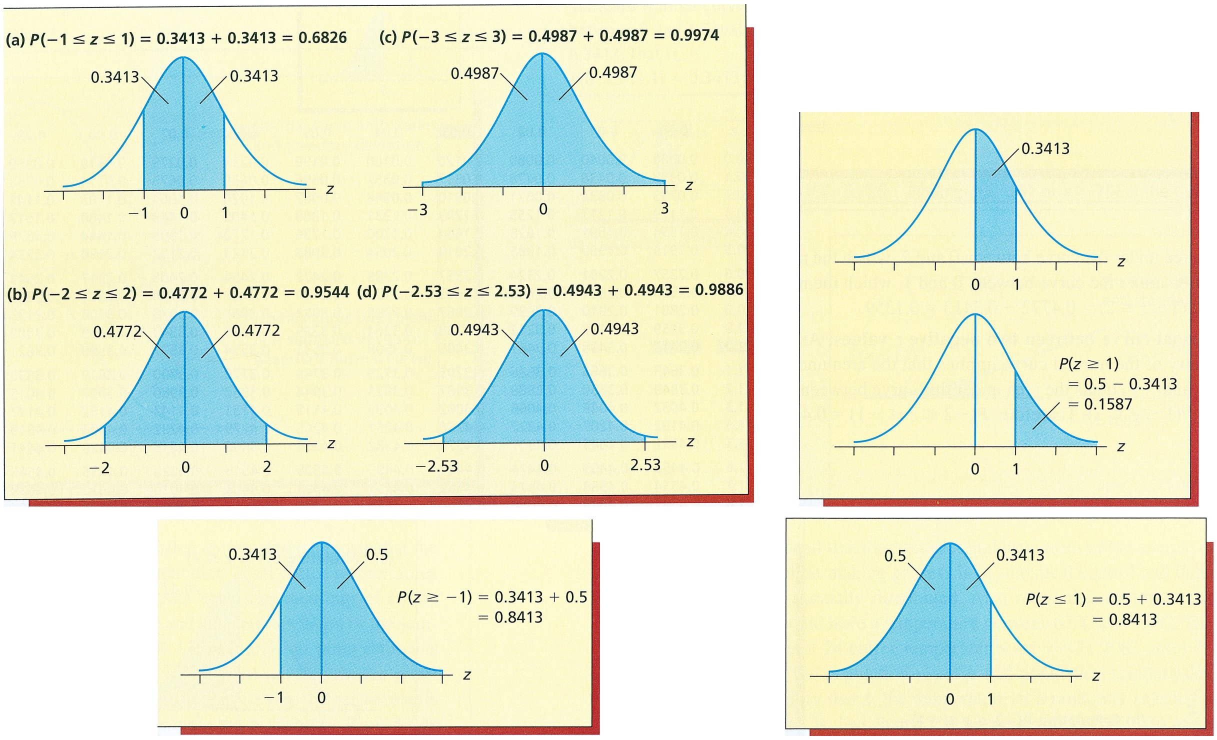


Figure 5: Examples of standard normal probabilities

Example 155. If X is a normal random variable with parameters $\mu = 2$ and $\sigma^2 = 9$, find $P(|X-3| > 5)$.

Solution. $P(|X - 3| > 5) = P(\{X - 3 > 5\} \cup \{X - 3 < -5\}) = P(X - 3 > 5) + P(X - 3 < -5) = P(X > 8) + P(X < -2)$ since the two events are disjoint. Next

$$P(X > 8) = P\Big(\frac{X-2}{3} > \frac{8-2}{3}\Big) = P(Z > 2) = 1 - \Phi(2) = 1 - 0.9772 = 0.0228;$$

$$P(X < -2) = P\Big(\frac{X-2}{3} < \frac{-2-2}{3}\Big) = P(Z < -1.333) = \Phi(-1.333) = 1 - \Phi(1.333) = 1 - 0.9082 = 0.0918.$$

Where at the end in both cases we used the cumulative normal table. Putting everything together, we obtain

$$P(|X - 3| > 5) = P(X > 8) + P(X < -2) = 0.0228 + 0.0918 = 0.1146.$$

Next, we present our second limit theorem.

Theorem 156 (DeMoivre-Laplace limit theorem)**.** Let X be the number of successes in a binomial experiment with n trials and probability of success p. Then, for any $x \in \mathbb{R}$ we have

$$P\Big(\frac{X-np}{\sqrt{npq}} \le x\Big) \to P(Z \le x),$$

as n approaches ∞, where Z is the standard normal random variable and $q := 1-p$.

The theorem is a special case of the central limit theorem, and we will not present a proof.

Now, we have two approximations to binomial probabilities: the Poisson theorem, which yields a good approximation when n is large and np is moderate, and the DeMoivre-Laplace theorem. A good rule of thumb is that the approximation in the DeMoivre-Laplace theorem is good when $npq \ge 10$. See Figure 6.

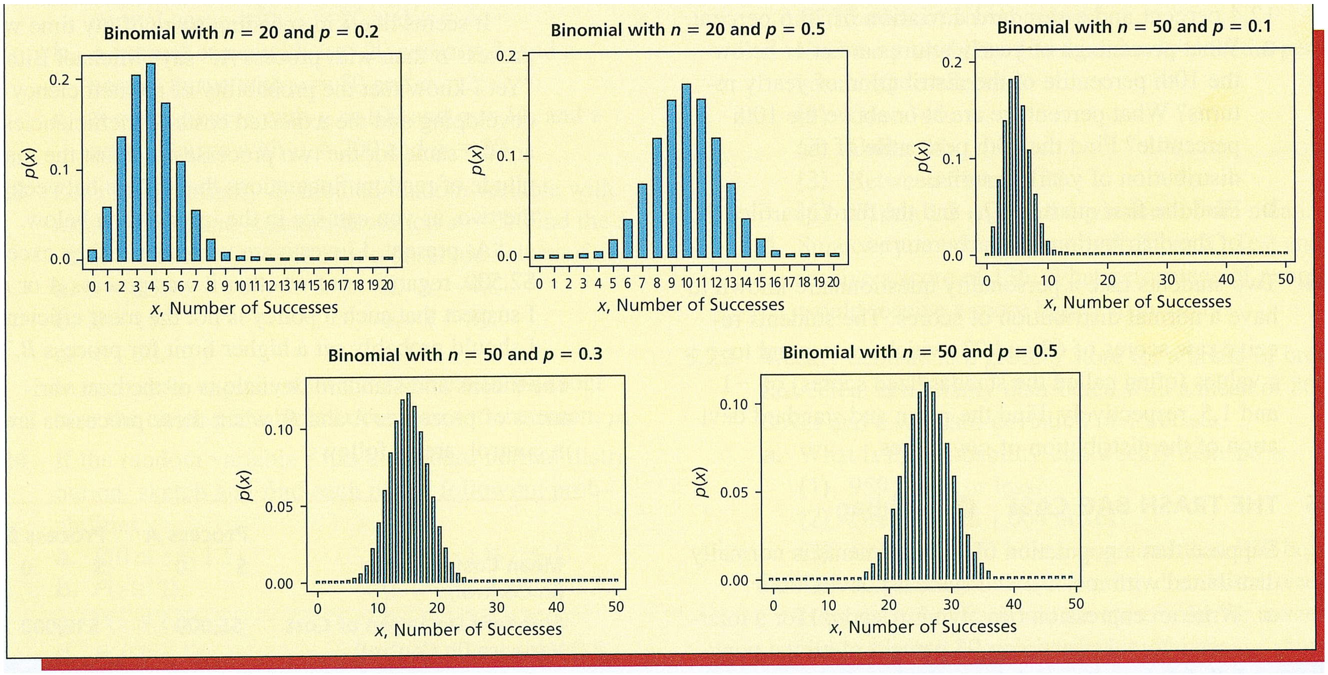


Figure 6: Normal approximation of the binomial

Example 157. Let X be the number of times that a fair coin, flipped 50 times, lands heads. Find the probability that $X = 23$. Use the normal approximation and then compare it to the exact solution.

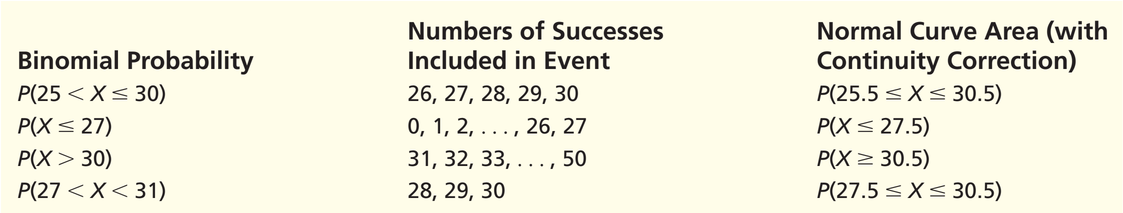
Solution. First check that $npq = 50(1/2)(1/2) = 12.5 \ge 10$, so that the normal approximation to the binomial will be good. Next, we need to calculate $P(X = 23)$ but instead we will apply the the DeMoivre-Laplace theorem to calculate $P(22.5 \le X \le 23.5)$. These, two probabilities are equal since X takes only integer values. But it will make a difference in the approximation. This is step called *continuity correction*, since we use a continuous distribution to approximate a discrete one. Next $np = 25$ and $\sqrt{npq} = 3.5355$, we apply the DeMoivre-Laplace theorem:

$$P(X = 23) = P(22.5 \le X \le 23.5) = P\Big(\frac{22.5-25}{3.5355} \le \frac{X-25}{3.5355} \le \frac{23.5-25}{3.5355}\Big)$$

$$\approx P(-0.71 \leq Z \leq -0.42) = 0.3372 - 0.2389 = 0.0983.$$

The actual answer, using the binomial probability distribution with $n = 50$, $p = 0.5$, and $x = 23$ is 0.0959617. Pretty good approximation! □

Figure 7 gives several examples how to apply the continuity correction in other cases.



Binomial Probability	Numbers of Successes Included in Event	Normal Curve Area (with Continuity Correction)
$P(25 < X \leq 30)$	26, 27, 28, 29, 30	$P(25.5 \leq X \leq 30.5)$
$P(X \leq 27)$	0, 1, 2, . . . , 26, 27	$P(X \leq 27.5)$
$P(X > 30)$	31, 32, 33, . . . , 50	$P(X \geq 30.5)$
$P(27 < X < 31)$	28, 29, 30	$P(27.5 \leq X \leq 30.5)$

Figure 7: Continuity correction

In summary, there are three steps that one needs to perform when approximating binomial distribution with normal: 1) check that $npq \geq 10$; 2) apply continuity correction to the event whose probability one computes; 3) apply the DeMoivre-Laplace theorem.

9.7 Gamma distribution

The *gamma function* is defined by

$$\Gamma(t) = \int_0^\infty e^{-x} x^{t-1}\, dx.$$

This integral is hard to evaluate for all values of t, but for some we can do it. Integration by parts of the integral in the definition of $\Gamma(t)$ yields

$$\Gamma(t) = -e^{-x}x^{t-1}\Big|_{x=0}^{\infty} + (t-1)\int_0^\infty e^{-x}x^{t-2}\, dx = (t-1)\int_0^\infty e^{-x}x^{t-2}\, dx = (t-1)\Gamma(t-1).$$

Thus, for integer values of t, say $t = n$, applying the above argument repeatedly, we obtain

$$\Gamma(n) = (n-1)\Gamma(n-1) = (n-1)(n-2)\Gamma(n-2) = \cdots = (n-1)(n-2)\cdots(3)(2)\Gamma(1).$$

Since $\Gamma(1) = \int_0^\infty e^{-x}\, dx = 1$ it follows that for all integer n

$$\Gamma(n) = (n-1)!$$

Hence, the gamma function extends factorials to all real numbers except the negative integers and zero. At those values the gamma function blows up to infinity. See Figure 8 for the graph of the gamma function.

It is curious to mention that

$$\Gamma(1/2) = \sqrt{\pi}.$$

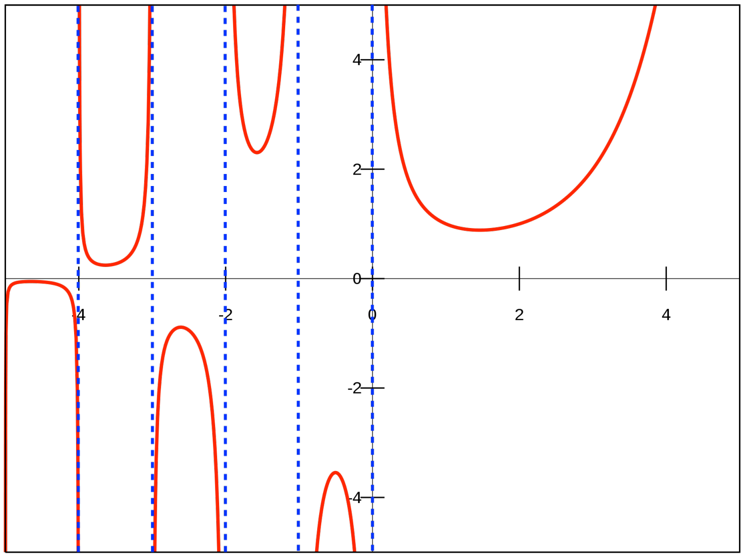


Figure 8: The graph of the gamma function

Definition 158. A continuous random variable has *gamma distribution* with parameters $k > 0$ and $\lambda > 0$, if its probability density function is

$$f(x) = \begin{cases} \frac{\lambda e^{-\lambda x}(\lambda x)^{k-1}}{\Gamma(k)} & \text{if } x > 0, \\ 0 & \text{if } x \leq 0. \end{cases}$$

To see that $f(x)$ is indeed a probability density function, note that $f(x) \geq 0$ and

$$\frac{1}{\Gamma(k)}\int_0^\infty \lambda e^{-\lambda x}(\lambda x)^{k-1}\,dx = \frac{1}{\Gamma(k)}\int_0^\infty e^{-y}y^{k-1}\,dy = \frac{1}{\Gamma(k)}\Gamma(k) = 1.$$

where we performed the change of variables $y = \lambda x$.

Note that when $k = 1$ the density of the gamma distribution reduces to that of the exponential distribution. So the gamma distribution is a generalization of the exponential one. Now recall Definition 129 of a Poisson process with rate λ together with Example 133. The next example is an extension of Example 133 and gives an interpretation of the gamma distribution when the parameter k is equal to an integer n.

Example 159. Given a Poisson process with rate λ, let X be the the amount of time one has to wait until a total of n events have occurred. Show that X has a gamma distribution with parameters n and λ.

Solution. Let T_n denote the time at which the n-th event occurs and note that T_n is less than or equal to x if and only if the number of events that have occurred by time x is at least n. That is, with $N(x)$ equal to the number of events in $[0, x]$, we have

$$P(T_n \leq x) = P(N(x) \geq n) = \sum_{j=n}^{\infty} P(N(x) = j) = \sum_{j=n}^{\infty} \frac{e^{-\lambda x}(\lambda x)^j}{j!},$$

where the last equality follows by the fact that the number of events in $[0, x]$ has a Poisson distribution with parameter λx. To obtain the density function of T_n, we need to differentiate $P(T_n \leq x)$ with respect to x.

$$\begin{aligned} f(x) &= \sum_{j=n}^{\infty} \frac{e^{-\lambda x} j(\lambda x)^{j-1}\lambda}{j!} - \sum_{j=n}^{\infty} \frac{\lambda e^{-\lambda x}(\lambda x)^j}{j!} \\ &= \sum_{j=n}^{\infty} \frac{\lambda e^{-\lambda x}(\lambda x)^{j-1}}{(j-1)!} - \sum_{j=n}^{\infty} \frac{\lambda e^{-\lambda x}(\lambda x)^j}{j!} \\ &= \frac{\lambda e^{-\lambda x}(\lambda x)^{n-1}}{(n-1)!}. \end{aligned}$$

Hence T_n is the gamma distribution with parameters n and λ. □

The gamma distribution with $k = n$, an integer, is also called *n-Erlang distribution*. The gamma distribution with $\lambda = 1/2$ and $k = n/2$ (n being a positive integer) is called the χ_n^2 (read "*chi-squared*") distribution with n degrees of freedom. As we will see later, the chi-squared distribution arises in practice as being the distribution of the error involved in attempting to hit a target in n-dimensional space when each coordinate error is normally distributed.

Note that when $k = 1$ the gamma distribution reduces to exponential.

The moment generating function of a gamma random variable is

$$M(t) = \frac{1}{(1 - t/\lambda)^k} \text{ for } t < \lambda. \tag{24}$$

The first two derivatives of $M(t)$ are

$$M'(t) = \frac{k}{\lambda}\frac{1}{(1 - t/\lambda)^{k+1}} \text{ and } M''(t) = \frac{k(k+1)}{\lambda^2}\frac{1}{(1 - t/\lambda)^{k+2}}.$$

Hence, we have

$$M'(0) = \frac{k}{\lambda} \text{ and } M''(0) = \frac{k(k+1)}{\lambda^2}$$

and from here

$$E[X] = M'(0) = \frac{k}{\lambda} \text{ and } \mathrm{Var}\,[X] = E[X^2] - (E[X])^2 = M''(0) - (M'(0))^2 = \frac{k}{\lambda^2}.$$

9.8 Beta function

The *beta function* is defined for $s > 0$, $t > 0$ by

$$B(s, t) = \int_0^1 z^{t-1}(1 - z)^{s-1}\, dz \tag{25}$$

The property that we are going to need is the identity

$$B(s, t) = \frac{\Gamma(s)\Gamma(t)}{\Gamma(s + t)} \tag{26}$$

holding for all $s > 0$, $t > 0$.

10 Jointly distributed random variables

So far we only dealt with one random variable at a time. But often two or more random variables are at play at the same time.

Definition 160. Let X and Y be random variables defined on the same sample space Ω. The *joint cumulative distribution function of X and Y* is defined by

$$F(x, y) := P(X \leq x, Y \leq y).$$

Here, the event $\{X \leq x, Y \leq y\}$ is $\{\omega \in \Omega : X(\omega) \leq x \text{ and } Y(\omega) \leq y\}$. Before we move on to the properties of the joint c.d.f. we need a lemma about a property of any probability measure P on a sample space Ω.

A sequence $A_1, A_2, A_3, \ldots$ of events is called *increasing sequence* if

$$A_1 \subset A_2 \subset A_3 \subset \cdots \subset A_n \subset A_{n+1} \subset \cdots$$

conversely, it is called *decreasing sequence* if

$$A_1 \supset A_2 \supset A_3 \supset \cdots \supset A_n \supset A_{n+1} \supset \cdots$$

If a sequence $A_1, A_2, A_3, \ldots$ of events is increasing, then the union of all events in the sequence

$$\bigcup_{n=1}^{\infty} A_n$$

is the smallest event that contains every A_i from the sequence. Conversely, if a sequence $A_1, A_2, A_3, \ldots$ of events is decreasing, then the intersection of all events in the sequence

$$\bigcap_{n=1}^{\infty} A_n$$

is the largest event that is contained in every A_i from the sequence.

Proposition 161 (Probability as a continuous set function)**.** Let P be a probability measure on a sample space Ω. If a sequence $A_1, A_2, A_3, \ldots$ of events in Ω is increasing, then the union of all events in the sequence has probability

$$P\Big(\bigcup_{k=1}^{\infty} A_k\Big) = \lim_{k\to\infty} P(A_k).$$

If a sequence $A_1, A_2, A_3, \ldots$ of events in Ω is decreasing, then the intersection of all events in the sequence has probability

$$P\Big(\bigcap_{k=1}^{\infty} A_k\Big) = \lim_{k\to\infty} P(A_k).$$

Proof. Suppose the sequence $A_1, A_2, A_3, \ldots$ of events in Ω is increasing. In order to evaluate $P\Big(\bigcup_{k=1}^{\infty} A_k\Big)$, we are going to partition the event $\bigcup_{k=1}^{\infty} A_k$ (that is, express is as a disjoint union) into events whose probability we know. Define the events

$$B_1 := A_1,\ B_2 := A_2 \setminus A_1,\ B_3 := A_3 \setminus A_2, \ldots \text{ and so on.}$$

That is $B_k = A_k \setminus A_{k-1}$ contains those points in A_k that are not in any of the previous sets $A_1, A_2, \ldots, A_{k-1}$. It is easy to see that the events $B_1, B_2, B_3, \ldots$ are disjoint and

$$\bigcup_{k=1}^{\infty} A_k = \bigcup_{k=1}^{\infty} B_k \text{ and } A_n = \bigcup_{k=1}^{n} B_k \text{ for all } n \geq 1.$$

We are ready to calculate

$$\begin{aligned} P\Big(\bigcup_{k=1}^{\infty} A_k\Big) &= P\Big(\bigcup_{k=1}^{\infty} B_k\Big) = \sum_{k=1}^{\infty} P(B_k) = \lim_{n\to\infty} \sum_{k=1}^{n} P(B_k) = \lim_{n\to\infty} P\Big(\bigcup_{k=1}^{n} B_k\Big) \\ &= \lim_{n\to\infty} P(A_n). \end{aligned}$$

For the second equality we used property (iii) in Definition 14, while for the last equality we used that $A_n = \bigcup_{k=1}^{n} A_k$.

The proof of the second part is left as an exercise. □

The proposition allows us to prove one of the characteristic properties of the cumulative distribution function, see property 4) on page 51.

Corollary 162. Let $F_X(x)$ be the c.d.f. of a random variable X, then

$$\lim_{x\to\infty} F_X(x) = 1 \text{ and } \lim_{x\to-\infty} F_X(x) = 0.$$

Proof. Let $x_1, x_2, x_3, \ldots$ be a sequence converging to infinity. Define the events $A_k := \{X \leq x_k\}$, for $k = 1, 2, 3, \ldots$, they form an increasing sequence of events, and moreover $\bigcup_{k=1}^{\infty} A_k = \Omega$. Thus, by the proposition

$$1 = P(\Omega) = P\Big(\bigcup_{k=1}^{\infty} A_k\Big) = \lim_{k\to\infty} P(A_k) = \lim_{k\to\infty} P(X \leq x_k) = \lim_{k\to\infty} F_X(x_k).$$

This proves that $\lim_{x\to\infty} F_X(x) = 1$. The other statement is left as an exercise. □

The cumulative distribution function of X can be obtained from $F(x, y)$. Let $y_1, y_2, y_3, \ldots$ be a increasing sequence converging to infinity. By Proposition 161, we have

$$F_X(x) = P(X \leq x) = P(X \leq x, Y \leq \infty) = P(\bigcup_{k=1}^{\infty}\{X \leq x, Y \leq y_k\}) = \lim_{k\to\infty} P(\{X \leq x, Y \leq y_k\}).$$

This shows that

$$F_X(x) = \lim_{y\to\infty} F(x, y)$$

Analogously, we can recover $F_Y(y)$:

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

The distribution functions F_X and F_Y are referred to as the *marginal distributions* of X and Y. All probability statements about X and Y can be answered in terms of the joint distribution function. For example, to compute the probability that X is greater than x and Y is greater than y, we argue as follows

$$\begin{aligned} P(X > x, Y > y) &= 1 - P(\{X > x, Y > y\}^c) = 1 - P\big((\{X > x\} \cap \{Y > y\})^c\big) \\ &= 1 - P\big(\{X > x\}^c \cup \{Y > y\}^c\big) = 1 - P\big(\{X \le x\} \cup \{Y \le y\}\big) \\ &= 1 - \big(P(X \le x) + P(Y \le y) - P\big(\{X \le x\} \cap \{Y \le y\}\big)\big) \\ &= 1 - \big(P(X \le x) + P(Y \le y) - P(X \le x, Y \le y)\big) \\ &= 1 - F_X(x) - F_Y(y) + F(x,y). \end{aligned}$$

Exercise 163. Show that for any $x_1 < x_2$ and $y_1 < y_2$ one has

$$P(x_1 < X \le x_2, y_1 < Y \le y_2) = F(x_2,y_2) + F(x_1,y_1) - F(x_1,y_2) - F(x_2,y_1).$$

We defined the joint c.d.f. of two random variables, but what about their joint probability density (reps. mass) function?

Definition 164 (Discrete joint p.m.f.)**.** Let X and Y be two discrete random variables. The function

$$p(x,y) := P(X = x, Y = y)$$

is called their *joint probability mass function.*

Suppose X is discrete and takes values $\{x_1, x_2, x_3, \ldots\}$. Note that the events $\{X = x_1\}, \{X = x_2\}, \{X = x_3\}, \ldots$ are disjoint and have union Ω, that is they partition the sample space. Similarly, if Y is discrete and takes values $\{y_1, y_2, y_3, \ldots\}$ the events $\{Y = y_1\}, \{Y = y_2\}, \{Y = y_3\}, \ldots$ partition the sample space. In addition, any event, say $\{X = x\}$ can be expressed as a disjoint union

$$\{X = x\} = \bigcup_{k=1}^{\infty} \big(\{X = x\} \cap \{Y = y_k\}\big)$$

Taking probability of both sides shows how the (*marginal*) probability mass function of X can be recovered from $p(x,y)$:

$$p_X(x) = P(X = x) = \sum_{k=1}^{\infty} p(x, y_k).$$

Similarly, the (*marginal*) probability mass function of Y can be recovered from $p(x,y)$

$$p_Y(y) = P(Y = y) = \sum_{k=1}^{\infty} p(x_k, y).$$

Example 165 (The multinomial distribution)**.** A sequence of n independent and identical experiments is performed. Suppose that each experiment can result in any one of r possible outcomes, with probabilities $p_1, p_2, \ldots, p_r$ respectively, $p_1 + p_2 + \cdots + p_r = 1$. An outcome of this experiment may be viewed as a sequence of length n whose elements are the integers from 1 to r. For example, the sequence $(1, 4, 3, \ldots, 5)$ indicates that the first experiment resulted in outcome 1, the second in outcome 4, the third in outcome 3, and so on, the n-th in outcome 5. The sample space Ω is the set of all such sequences and ω will denote such a sequence of length n.

Note that when $r = 2$ this set up reduces to the familiar binomial experiment, where we counted the number of successes in n experiments. There were two possible outcomes of each experiment 'success' (outcome 1) and 'failure' (outcome 2).

What is the probability measure on Ω? Let ω be a sequence of length n in which 1 appears n_1 times, 2 appears n_2 times, and so on, r appears p_r times. Since the experiments are independent, the natural way to define the probability measure is

$$P(\omega) := p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r}. \tag{27}$$

Let X_k be the number of experiments that resulted in outcome k, for $k = 1, 2, \ldots, r$. That is given a sequence $\omega \in \Omega$, $X_k(\omega)$ is equal to the number of times k appears in the sequence. This defines r random variables and note that for any $\omega \in \Omega$ we have

$$X_1(\omega) + X_2(\omega) + \cdots + X_r(\omega) = n.$$

Each of these random variables by itself, say X_k, is a binomial random variable with parameters n and p_k. So, we know that

$$P(X_k = m) = \binom{n}{m} p_k^m (1 - p_k)^{n-m} \text{ for all } m = 1, 2, \ldots, n.$$

Now we want to investigate how all random variables $X_1, X_2, \ldots, X_r$ interact with each other.

Let $n_1, n_2, \ldots, n_r$ be non-negative integers. We want to find the probability $P(X_1 = n_1, X_2 = n_2, \ldots, X_r = n_r)$, that is, the probability of the event that the outcome k occurred n_k times, for all $k = 1, 2, \ldots, r$. First of all, the event

$$\{X_1 = n_1, X_2 = n_2, \ldots, X_r = n_r\}$$

is empty if $n_1 + n_2 + \cdots + n_r \neq n$. So, suppose that $n_1 + n_2 + \cdots + n_r = n$. The event is made up of all sequences of length n in which 1 appears n_1 times, 2 appears n_2 times and so on r appears n_r times. Each such sequence is nothing else but dividing the integers $\{1, 2, \ldots, n\}$ (the position in the sequence) into r distinct groups: in the first group we have all positions where 1 appears in the sequence; in the second group we have all positions where 2 appears in the sequence; and so on. Thus, in the first group there are n_1 elements , in the second group we have n_2 elements and so on. Thus, from Example 38 we know that there are $\binom{n}{n_1, n_2, \ldots, n_r}$ such sequences and each one has probability (27). Multiplying the number of such sequences by that probability we find

$$P(X_1 = n_1, X_2 = n_2, \ldots, X_r = n_r) = \binom{n}{n_1, n_2, \ldots, n_r} p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r}$$

$$= \frac{n!}{n_1!n_2!\cdots n_r!} p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r}.$$

What we found is the joint probability mass function of the random variables $X_1, X_2, \ldots, X_r$.

□

Exercise 166. Define $n_1 := n - n_2 - \cdots - n_r$. Calculate the sum

$$\sum_{\substack{n_2,\ldots,n_r \geq 0 \\ n_2+\cdots+n_r \leq n}} \binom{n}{n_1, n_2, \ldots, n_r} p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r}.$$

Definition 167 (Continuous joint p.d.f.)**.** Let X and Y be two continuous random variables. We say that X and Y are *jointly (absolutely) continuous* if there is a function $f(x,y) \geq 0$ defined for all $(x,y) \in \mathbb{R}^2$ and such that for each set $C \subseteq \mathbb{R}^2$ satisfies

$$P((X,Y) \in C) = \int\int_{(x,y)\in C} f(x,y)\, dxdy.$$

The function $f(x,y)$ is called the *joint probability density function.*

Remember that it is possible to have two continuous random variables that are *not* jointly continuous!

Some sets in $\mathbb{R}^2$ are nice, these are the rectangles: if A and B are subsets of $\mathbb{R}$, then the set $C = \{(x,y) \in \mathbb{R}^2 : x \in A, y \in B\}$ is a *rectangle.* The probability of the pair (X,Y) to be in the rectangle C is

$$P((X,Y) \in C) = P(X \in A, Y \in B) = \int_B \int_A f(x,y)\, dxdy,$$

in particular, the joint cumulative distribution function of X and Y is given by

$$F(x,y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(s,t)\, dtds,$$

and from the fundamental theorem of calculus we obtain

$$f(x,y) = \frac{\partial^2}{\partial x \partial y} F(x,y). \tag{28}$$

Now, if you know the joint density $f(x,y)$ of X and Y, how to find the density of, say, X? Let A be an event and compute

$$\begin{aligned} P(X \in A) &= P(X \in A, Y \in (-\infty,\infty)) \\ &= \int_{-\infty}^{\infty} \int_A f(x,y)\, dxdy = \int_A \int_{-\infty}^{\infty} f(x,y)\, dydx \\ &= \int_A \Big(\int_{-\infty}^{\infty} f(x,y)\, dy \Big) dx. \end{aligned}$$

This shows that the function in the big parenthesis is *a* p.d.f. of X, that is

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

Analogously, we have

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

Finally, note that since $\Omega = \{X \in (-\infty,\infty), Y \in (-\infty,\infty)\}$, taking probability on both sides gives

$$1 = P(\Omega) = P(X \in (-\infty,\infty), Y \in (-\infty,\infty)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dxdy.$$

Example 168. Let the joint p.d.f. of X and Y be given by

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

Find a) $P(X>1, Y<1)$; b) $P(X<Y)$; c) $P(X<5)$.

Solution. a) Define the sets $A = \{x \in \mathbb{R} : x > 1\}$ and $B = \{y \in \mathbb{R} : y < 1\}$. Then the event $\{X > 1, Y < 1\}$ is the same as the event $\{X \in A, Y \in B\}$. Then,

$$\begin{aligned} P(X>1,Y<1) = P(X \in A, Y \in B) &= \int_{-\infty}^{1}\int_{1}^{\infty} f(x,y)\,dxdy = \int_{0}^{1}\int_{1}^{\infty} 2e^{-x}e^{-2y}\,dxdy \\ &= 2\int_1^\infty e^{-x}\,dx \int_0^1 e^{-2y}\,dy = 2\Big(-e^{-x}\Big|_{x=1}^{\infty}\Big)\Big((-1/2)e^{-2y}\Big|_{y=0}^{1}\Big) = e^{-1}(1-e^{-2}). \end{aligned}$$

b) Define the set $C = \{(x,y) \in \mathbb{R}^2 : x < y\} = \{(x,y) \in \mathbb{R}^2 : -\infty < y < \infty, -\infty < x < y\}$. Then the event $\{X < Y\}$ is the same as the event $\{(X,Y) \in C\}$. Then

$$\begin{aligned} P(X<Y) = P((X,Y) \in C) &= \int\int_{(x,y)\in C} f(x,y)\,dxdy = \int_0^\infty \int_0^y 2e^{-x}e^{-2y}\,dxdy \\ &= 2\int_0^\infty e^{-2y}\Big(\int_0^y e^{-x}\,dx\Big)dy = 2\int_0^\infty e^{-2y}(1-e^{-y})\,dy \\ &= 2\int_0^\infty e^{-2y}\,dy - 2\int_0^\infty e^{-3y}\,dy = 1 - 2/3 = 1/3. \end{aligned}$$

c) In this case

$$\begin{aligned} P(X<5) = P(X \in (-\infty,5), Y \in (-\infty,\infty)\} &= \int_0^5\int_0^\infty 2e^{-x}e^{-2y}\,dydx = 2\int_0^5 e^{-x}\,dx \int_0^\infty e^{-2y}\,dy \\ &= (1-e^{-5})(1-0) = 1-e^{-5}. \end{aligned}$$

The next theorem is should be compared with Theorem 104.

Theorem 169. Suppose that X and Y are random variables and $g(x,y)$ is a function defined on $\mathbb{R}^2$. The *expected value* of $g(X,Y)$, denoted by $E[g(X,Y)]$, is calculated as follows.

- If X and Y are discrete random variables taking values $\{x_1, x_2, \ldots\}$ and $\{y_1, y_2, \ldots\}$ respectively, and having joint probability mass function $p(x, y)$, then

$$E[g(X,Y)] := \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} g(x_i, y_j) p(x_i, y_j).$$

- If X and Y are continuous random variables with values in $\mathbb{R}$ and joint probability density function $f(x, y)$, then

$$E[g(X,Y)] := \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f(x,y)\, dxdy.$$

Corollary 170. Let X and Y be random variables. Suppose that $E[X]$ and $E[Y]$ are both finite numbers. Then

$$E[X+Y] = E[X] + E[Y].$$

Proof. Consider the function $g(x, y) = x + y$. Then

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

□

It is straightforward to generalize the last corollary to n random variables

$$E[X_1 + X_2 + \cdots + X_n] = E[X_1] + E[X_2] + \cdots + E[X_n]. \tag{29}$$

Corollary 171. If the random variables X and Y satisfy $X \geq Y$, then $E[X] \geq E[Y]$.

Proof. Since $X - Y \geq 0$, it follows that $E[X - Y] \geq 0$ or $0 \leq E[X - Y] = E[X] + E[-Y] = E[X] - E[Y]$. □

In general, it is not true that $E[XY] = E[X]E[Y]$, but we will see in the next section that $E[XY] = E[X]E[Y]$ whenever X and Y are independent random variables.

Example 172 (Mean of a hypergeometric random variable)**.** n balls are randomly selected from an urn containing N balls of which m are white and the rest are black. Find the expected number of white balls selected.

Solution. Let X denote the number of white balls selected. Enumerate the white balls from 1 to m. Define the variables X_i to be 1 if i-th white ball is in the sample of n balls; and 0 otherwise. Then, $X = X_1 + X_2 + \cdots + X_m$. By Example 101 we have that

$$E[X_i] = P(X_i = 1) = P(\text{the } i\text{-th white ball is selected}) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}.$$

Thus,

$$E[X] = E[X_1] + E[X_2] + \cdots + E[X_m] = \frac{nm}{N}.$$

We now give an intuitive interpretation of the joint probability density function. Using Exercise 163 and formula (28), an argument similar to the one leading to (13) shows that

$$f(x, y) = \lim_{h \to 0} \frac{P(x \le X \le x + h, y \le Y \le y + h)}{h^2}$$

implying that for values of h close to 0, we have

$$f(x, y)h^2 \approx P(x \le X \le x + h, y \le Y \le y + h).$$

We omit the details. Finally, we list the important properties of the joint cumulative distribution function.

Proposition 173. Let $F(x, y)$ be the joint c.d.f. of two continuous random variables. Then, $F(x, y)$ is a continuous function on $\mathbb{R}^2$ and

(i) $\lim_{x,y \to -\infty} F(x, y) = 0, \quad \lim_{x,y \to \infty} F(x, y) = 1;$

(ii) $F(x_1, y) \le F(x_2, y)$ if $x_1 \le x_2$, $\quad F(x, y_1) \le F(x, y_2)$ if $y_1 \le y_2$;

(iii) $\lim_{x \to \infty} F(x, y) = F_Y(y), \quad \lim_{y \to \infty} F(x, y) = F_X(x),$

11 Independent random variables

To simplify the notation the event $\{X \in A\} \cap \{Y \in B\}$ will be denoted simply by $\{X \in A, Y \in B\}$.

Definition 174. Two random variables X and Y are called *independent* if for any two sets A, and B of real numbers, the events $\{X \in A\}$ and $\{Y \in B\}$ are independent. We write this with the formula

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B) \text{ for any } A, B \subset \mathbb{R}.$$

In general, n random variables $X_1, X_2, \ldots, X_n$, are said to be *independent* if, for any sets of real numbers $A_1, A_2, \ldots, A_n$, the events $\{X_1 \in A_1\}$, $\{X_2 \in A_2\}, \ldots, \{X_n \in A_n\}$ are independent. That is

$$P(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) = P(X_1 \in A_1)P(X_2 \in A_2) \cdots P(X_n \in A_n)$$

holds for any sets of real numbers $A_1, A_2, \ldots, A_n$

If X and Y are independent, then their joint c.d.f. has the following property

$$F(x,y) = P(X \le x, Y \le y) = P(X \le x)P(Y \le y) = F_X(x)F_Y(y).$$

It can be shown (proof omitted) that, the opposite statement also holds. That is, if $F(x,y) = F_X(x)F_Y(y)$ for every x and y, then X and Y are independent. If X and Y are discrete random variables, then the definition of independence simply says that the joint probability mass function satisfies

$$p(x,y) = p_X(x)p_Y(y).$$

The same is true for continuous random variables and the next theorem tells us how to recognize if two random variables are independent by only looking at their joint probability density function—if one can "separate" the variables in the joint p.d.f. then X and Y are independent.

Theorem 175. Suppose X and Y are jointly continuous with joint p.d.f. $f(x,y)$. Then, X and Y are independent if and only if the joint p.d.f. has the form $f(x,y) = h(x)g(y)$ for any x, y.

In the theorem, the functions $h(x)$ and $g(y)$ are not necessarily the marginal p.d.f.'s of X and Y, but are close to them in the sense that there are constants C_1 and C_2 with $C_1C_2 = 1$ such that

$$f_X(x) = C_1h(x) \text{ and } f_Y(y) = C_2g(y).$$

Corollary 176. Suppose X and Y are jointly continuous with joint p.d.f. $f(x,y)$. Then, X and Y are independent if and only if $f(x,y) = f_X(x)f_Y(y)$.

Corollary 177. Let X and Y be independent random variables. Then, for any functions g and h, we have

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)].$$

Proof. Consider the function $p(x,y) = g(x)h(y)$. Then, by Theorem 169, we have

$$\begin{aligned} E[g(X)h(Y)] = E[p(X,Y)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)f(x,y)\,dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f_X(x)f_Y(y)\,dxdy \\ &= \Big(\int_{-\infty}^{\infty} g(x)f_X(x)\,dx\Big)\Big(\int_{-\infty}^{\infty} h(y)f_Y(y)\,dy\Big) = E[g(X)]E[h(Y)], \end{aligned}$$

where at the end we used Theorem 104. □

Take $g(x) = x$ and $h(y) = y$ in the last corollary, to obtain that if X and Y are independent random variables, then $E[XY] = E[X]E[Y]$. These observations can be generalized and we do it without a proof.

Theorem 178. If $X_1, X_2, \ldots, X_n$ are independent random variables, then

$$E[X_1X_2\cdots X_n] = E[X_1]E[X_2]\cdots E[X_n].$$

The converse of the theorem is not true. That is, if $E[XY] = E[X]E[Y]$ we cannot conclude that X and Y are independent as in the following example.

Example 179. Let X be a random variable such that

$$P(X=0)=P(X=1)=P(X=-1)=\frac{1}{3}.$$

Let Y be a random variable such that

$$Y=\begin{cases} 0 & \text{if } X\neq 0,\\ 1 & \text{if } X=0.\end{cases}$$

Since $XY=0$, we have $E[XY]=0$. We also have

$$E[X]=0P(X=0)+1P(X=1)+(-1)P(X=-1)=0.$$

Hence $E[XY]=E[X]E[Y]$. But X and Y are dependent. Indeed

$$P(X=1,Y=0)=P(X=1)P(Y=0|X=1)=(1/3)(1)=1/3$$

and

$$P(X=1)=1/3 \text{ and } P(Y=0)=P(X\neq 0)=P(X=1)+P(X=-1)=2/3$$

implying that

$$P(X=1,Y=0)=1/3\neq(1/3)(2/3)=P(X=1)P(Y=0).$$

Example 180. A man and a woman decide to meet at a certain location. If each person independently arrives at a time uniformly distributed between 12 noon and 1 pm, find the probability that the first to arrive has to wait longer than 10 minutes.

Solution. Let X and Y denote the time of arrival of the man and the woman, respectively, measured in minutes starting from 12 noon. We need to calculate the probability $P(X+10<Y)+P(Y+10<X)$ and since X and Y are completely interchangeable, that sum is just $2P(X+10<Y)$. Next, the p.d.f. of X is $f_X(x)=1/60$ for $x\in[0,60]$ and 0 otherwise. Similarly, the p.d.f. of Y is $f_Y(y)=1/60$ for $y\in[0,60]$ and 0 otherwise. Since X and Y are independent, $f(x,y)=1/60^2$ for $x\in[0,60]$ and $y\in[0,60]$, and 0 otherwise. So we calculate

$$2P(X+10<Y)=2\int\int_{x+10<y}f(x,y)\,dxdy=2\int_{10}^{60}\int_0^{y-10}1/60^2\,dxdy=\frac{2}{60^2}\int_{10}^{60}(y-10)\,dy=\frac{25}{36}.$$

Example 181. Let X, Y, Z be independent and uniformly distributed over $[0,1]$. Find that probability that X is not-smaller than the product of Y and Z.

Solution. Each random variable has a p.d.f. $f(x)=1$ for $x\in[0,1]$ and 0 otherwise. Since they are independent their joint p.d.f. is $f(x,y,z)=f(x)f(y)f(z)=1$ for $x,y,z\in[0,1]$. Thus

$$\begin{aligned}P(X\geq YZ)&=\int\int\int_{x\geq yz}f(x,y,z)\,dxdydz=\int_0^1\int_0^1\int_{yz}^1 1\,dxdydz\\&=\int_0^1\int_0^1(1-yz)\,dydz=\int_0^1\left(1-\frac{z}{2}\right)=\frac{3}{4}.\end{aligned}$$

We conclude this part with a useful fact about independence. We are not going to prove it.

Theorem 182. If $X_1,X_2,\ldots,X_n$ are independent random variables and $g_1,g_2,\ldots,g_n$ are functions on $\mathbb{R}$, then the random variables $g_1(X_1),g_2(X_2),\ldots,g_n(X_n)$ are also independent.

11.1 Sums of independent random variables

This subsection deals with the following problem. If you know the distribution of X and Y and they are independent, find the distribution of $X+Y$.

11.1.1 Discrete random variables

Example 183. Let X and Y be independent binomial random variables with parameters (n,p) and (m,p) respectively. Calculate the distribution of $X+Y$.

Solution. X represents the number of successes in n independent trials, each of which results in a success with probability p. Similarly, Y represents the number of successes in m independent trials, each trial being a success with probability p. Hence, as X and Y are assumed independent, it follows that $X+Y$ represents the number of successes in $n+m$ independent trials when each trial has a probability p of being a success. Therefore, $X+Y$ is a binomial random variable with parameters $(n+m,p)$. □

Example 184. Let X and Y be independent Poisson random variables with parameters λ and μ respectively. Calculate the distribution of $X+Y$.

Solution. Each X and Y take values $0,1,2,\ldots$, hence $X+Y$ takes values $0,1,2,\ldots$ as well. We need to calculate $P(X+Y=n)$ for $n=0,1,2,\ldots$. First, note that the event $\{X+Y=n\}$ is a disjoint union of the events $\{X=k,Y=n-k\}$, where $k=0,1,2,\ldots,n$. Second, taking probabilities, we have

$$P(X+Y=n)=\sum_{k=0}^{n}P(X=k,Y=n-k)=\sum_{k=0}^{n}P(X=k)P(Y=n-k),$$

using that X and Y are independent. Continue with the formula for the p.m.f. of the Poisson random variable.

$$\begin{aligned}P(X+Y=n)&=\sum_{k=0}^{n}e^{-\lambda}\frac{\lambda^k}{k!}e^{-\mu}\frac{\mu^{n-k}}{(n-k)!}=e^{-\lambda-\mu}\sum_{k=0}^{n}\frac{\lambda^k\mu^{n-k}}{k!(n-k)!}\\&=\frac{e^{-\lambda-\mu}}{n!}\sum_{k=0}^{n}\lambda^k\mu^{n-k}\frac{n!}{k!(n-k)!}=\frac{e^{-\lambda-\mu}}{n!}(\lambda+\mu)^n.\end{aligned}$$

That is, the p.m.f. of $X+Y$ is the same as the one of a Poisson random variable with parameter $\lambda+\mu$. □

11.1.2 Continuous random variables

If X and Y are (absolutely) continuous random variables, then the c.d.f. of $X+Y$ is

$$\begin{aligned}F_{X+Y}(t)=P(X+Y\le t)&=\int\int_{x+y\le t}f(x,y)\,dxdy=\int\int_{x+y\le t}f_X(x)f_Y(y)\,dxdy\\&=\int_{-\infty}^{\infty}\int_{-\infty}^{t-y}f_X(x)f_Y(y)\,dxdy=\int_{-\infty}^{\infty}f_Y(y)\Big(\int_{-\infty}^{t-y}f_X(x)\,dx\Big)\,dy\end{aligned}$$

$$= \int_{-\infty}^{\infty} f_Y(y) F_X(t-y)\, dy.$$

To obtain the p.d.f. of $X+Y$ we need to differentiate $F_{X+Y}(t)$ with respect to t. Using the above calculation, we obtain

$$\begin{aligned} f_{X+Y}(t) &= \frac{d}{dt} F_{X+Y}(t) = \frac{d}{dt} \int_{-\infty}^{\infty} f_Y(y) F_X(t-y)\, dy = \int_{-\infty}^{\infty} \frac{d}{dt} \big(f_Y(y) F_X(t-y) \big)\, dy \\ &= \int_{-\infty}^{\infty} f_Y(y) f_X(t-y)\, dy. \qquad (30) \end{aligned}$$

The above two formulas show how, knowing the p.d.f. and the c.d.f. of X and Y we can find the p.d.f. and the c.d.f. of $X+Y$. This covers the case when both X and Y are (absolutely) continuous random variables. The case when X and Y are both discrete is more straightforward as the previous examples show.

Example 185. If X and Y are independent random variables, both uniformly distributed in $[0,1]$, find the p.d.f. of $X+Y$.

Solution. First of all, note that $X+Y$ takes values in $[0,2]$, so its p.d.f. may be non-zero there. Using the formula above, we calculate.

$$f_{X+Y}(t) = \int_{-\infty}^{\infty} f_Y(y) f_X(t-y)\, dy = \int_0^1 f_X(t-y)\, dy,$$

where we used that $f_Y(y) = 1$ if $y \in [0,1]$ and 0 otherwise. Continuing, we need to consider two cases. If $t \in [0,1]$ then

$$\int_0^1 f_X(t-y)\, dy = \int_0^t 1\, dy = t.$$

If $t \in [1,2]$ then

$$\int_0^1 f_X(t-y)\, dy = \int_{t-1}^1 1\, dy = 2-t.$$

Putting it all together gives

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

Graph this function to see why the distribution of $X+Y$ is called *triangular distribution.* □

Example 186. Suppose X and Y are independent random variables. If X has gamma distribution with parameters s, λ and Y has gamma distribution with parameters t, λ, show that $X+Y$ has gamma distribution with parameters $s+t, \lambda$.

Solution. It is given that

$$f_X(x) = \begin{cases} \frac{1}{\Gamma(s)}\lambda e^{-\lambda x}(\lambda x)^{s-1} & \text{if } 0 < x < \infty, \\ 0 & \text{otherwise,} \end{cases} \quad \text{and}$$
$$f_Y(y) = \begin{cases} \frac{1}{\Gamma(t)}\lambda e^{-\lambda y}(\lambda y)^{t-1} & \text{if } 0 < y < \infty, \\ 0 & \text{otherwise.} \end{cases}$$

Using the formula for the density of $X+Y$, we get (note that if the argument y in $f_X(u-y)$ is bigger than u, then $f_X(u-y) = 0$, in addition, if $y \leq 0$ then $f_Y(y) = 0$)

$$\begin{aligned} f_{X+Y}(u) &= \int_{-\infty}^{\infty} f_Y(y) f_X(u-y)\, dy = \frac{1}{\Gamma(s)\Gamma(t)} \int_0^u \lambda e^{-\lambda y}(\lambda y)^{t-1} \lambda e^{-\lambda(u-y)}(\lambda(u-y))^{s-1}\, dy \\ &= \frac{e^{-\lambda u}\lambda^{s+t}}{\Gamma(s)\Gamma(t)} \int_0^u y^{t-1}(u-y)^{s-1}\, dy \\ &= \frac{e^{-\lambda u}\lambda^{s+t}u^{s+t-1}}{\Gamma(s)\Gamma(t)} \int_0^1 z^{t-1}(1-z)^{s-1}\, dz, \end{aligned}$$

where the last integral is obtained after the change of variables $z = y/u$ with $dz = dy/u$. Now, recalling the definition of the beta function (25) and its property (26), we continue

$$f_{X+Y}(u) = \frac{e^{-\lambda u}\lambda^{s+t}u^{s+t-1}}{\Gamma(s)\Gamma(t)} B(s,t) = \frac{e^{-\lambda u}\lambda^{s+t}u^{s+t-1}}{\Gamma(s+t)} = \frac{1}{\Gamma(s+t)} \lambda e^{-\lambda u}(\lambda u)^{s+t-1}.$$

This is precisely the p.d.f. of a gamma random variable with parameters $s+t$ and λ. □

Using the previous example, it is a simple inductive argument to see that if X_i, for $i = 1, 2, 3, \ldots, n$, are independent gamma random variables with parameters t_i and λ, then $X_1 + X_2 + \cdots + X_n$ is a gamma distributed random variable with parameters $t_1 + t_2 + \cdots + t_n$ and λ.

Example 187. Let $X_1, X_2, \ldots, X_n$ be independent exponential random variables with parameter λ. Recall that an exponential random variable with parameter λ is the same as a gamma random variable with parameters 1 and λ. Thus, by the previous example $X_1 + X_2 + \cdots + X_n$ is a gamma random variable with parameters n and λ. Note that this is also the essence of Example 159.

Example 188. If X and Y are independent, normally distributed, random variables, with parameters μ_X, σ_X^2 and μ_Y, σ_Y^2, then $X+Y$ is normally distributed with parameters $\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2$.

Solution. We consider first a simpler case. Suppose that $\mu_X = 0$ and $\sigma_X^2 = \sigma^2$, while $\mu_Y = 0$ and $\sigma_Y^2 = 1$. Then

$$\begin{aligned} f_{X+Y}(t) &= \int_{-\infty}^{\infty} f_Y(y) f_X(t-y)\, dy = \int_{-\infty}^{\infty} \Big(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-y)^2}{2\sigma^2}}\Big)\Big(\frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\Big)\, dy \\ &= \frac{1}{2\pi\sigma} e^{-\frac{t^2}{2\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{y^2(1+\sigma^2)-2ty}{2\sigma^2}}\, dy = \frac{1}{2\pi\sigma} e^{-\frac{t^2}{2\sigma^2}} e^{\frac{t^2}{2\sigma^2(1+\sigma^2)}} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2\sigma^2(1+\sigma^2)}} e^{-\frac{y^2(1+\sigma^2)-2ty}{2\sigma^2}}\, dy \\ &= \frac{1}{2\pi\sigma} e^{-\frac{t^2}{2(1+\sigma^2)}} \int_{-\infty}^{\infty} e^{-\frac{y^2(1+\sigma^2)^2 - 2ty(1+\sigma^2) + t^2}{2\sigma^2(1+\sigma^2)}}\, dy \end{aligned}$$

$$= \frac{1}{2\pi\sigma} e^{-\frac{t^2}{2(1+\sigma^2)}} \int_{-\infty}^{\infty} e^{-\frac{(y(1+\sigma^2)-t)^2}{2\sigma^2(1+\sigma^2)}} \, dy.$$

Change the variable in the integral by letting $x := \frac{y(1+\sigma^2)-t}{\sqrt{\sigma^2(1+\sigma^2)}}$ leading to $dx = \frac{1+\sigma^2}{\sqrt{\sigma^2(1+\sigma^2)}}dy$ or $dy = \frac{\sigma}{\sqrt{1+\sigma^2}}dx$. Then,

$$f_{X+Y}(t) = \frac{1}{2\pi\sqrt{1+\sigma^2}} e^{-\frac{t^2}{2(1+\sigma^2)}} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} \, dx = \frac{1}{\sqrt{2\pi(1+\sigma^2)}} e^{-\frac{t^2}{2(1+\sigma^2)}},$$

where the last integral was calculated in Lemma 151. We recognize the last function as the p.d.f. of a normal random variable with mean 0 and variance $(1+\sigma^2)$.

We now consider the general case. Consider the representation

$$X + Y = \sigma_Y\Big(\frac{X-\mu_X}{\sigma_Y} + \frac{Y-\mu_Y}{\sigma_Y}\Big) + \mu_X + \mu_Y. \tag{31}$$

Now $(X-\mu_X)/\sigma_Y$ is normal with mean 0 and variance σ_X^2/σ_Y^2, while $(Y-\mu_Y)/\sigma_Y$ is normal with mean 0 and variance 1. Hence, by the particular case above

$$\frac{X-\mu_X}{\sigma_Y} + \frac{Y-\mu_Y}{\sigma_Y}$$

is normal with mean 0 and variance $1 + \sigma_X^2/\sigma_Y^2$. Thus, $X+Y$ is normal with mean $\mu_X + \mu_Y$ and variance $\sigma_Y^2(1+\sigma_X^2/\sigma_Y^2) = \sigma_X^2 + \sigma_Y^2$. □

Using the last example, it is straightforward to see that the sum of n normal random variables is normal with mean equal to the sum of the means and variance equal to the sum of the variances.

Example 189. Let Z be a standard normal random variable. Find the probability density function of Z^2.

Solution. Recall that

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

We use the result from Example 98. Let $Y := Z^2$, then for $y \geq 0$ we have

$$f_Y(y) = \frac{1}{2\sqrt{y}}\big(f_Z(\sqrt{y}) + f_Z(-\sqrt{y})\big) = \frac{1}{2\sqrt{y}}\frac{2}{\sqrt{2\pi}}e^{-\frac{y}{2}} = \frac{\frac{1}{2}e^{-\frac{y}{2}}\big(\frac{1}{2}y\big)^{\frac{1}{2}-1}}{\sqrt{\pi}}.$$

Since $Y \geq 0$ we find that its density function $f_Y(y)$ must be equal to zero for $y < 0$. The reason we wrote the last expression in such a fancy form is so that we can recognize the d.p.f. of the gamma distribution with parameters $\big(\frac{1}{2}, \frac{1}{2}\big)$. □

Suppose now that $Z_1, Z_2, \ldots, Z_n$ are independent standard normal random variables. By Example 189, each one of the random variables $Z_1^2, Z_2^2, \ldots, Z_n^2$ has gamma distribution with parameters $\big(\frac{1}{2}, \frac{1}{2}\big)$. Next, applying Theorem 182 to $Z_1, Z_2, \ldots, Z_n$ and the function $g_1(z) = g_2(z) = \cdots = g_n(z) = z^2$, we see that the random variables $Z_1^2, Z_2^2, \ldots, Z_n^2$ are independent. And finally, using Example 186, we see that the sum

$$Y := Z_1^2 + Z_2^2 + \cdots + Z_n^2$$

has gamma distribution with parameters $\left(\frac{n}{2},\frac{1}{2}\right)$ and hence its p.d.f. is

$$f_Y(y)=\begin{cases}\dfrac{\frac{1}{2}e^{-\frac{y}{2}}\left(\frac{1}{2}y\right)^{\frac{n}{2}-1}}{\Gamma\left(\frac{n}{2}\right)} & \text{if } 0<y,\\ 0 & \text{otherwise.}\end{cases}$$

This distribution is so important in statistics that it was given a name.

Definition 190 (Chi-quared)**.** If $Z_1, Z_2, \ldots, Z_n$ are independent standard normal random variables, then the distribution of

$$Y := Z_1^2 + Z_2^2 + \cdots + Z_n^2$$

is called *chi-squared with n degrees of freedom.* Often the chi-squared distribution is denoted by χ^2.

Using (24) we obtain the moment generating function of the chi-squared distribution with n degrees of freedom

$$M_{\chi^2}(t)=\frac{1}{(1-2t)^{n/2}} \text{ for } t<\frac{1}{2}. \tag{32}$$

11.1.3 Using moment generation functions

Recall the notion of moment generating function.

Theorem 191. Let $X_1, X_2, \ldots, X_n$ be independent random variables, and let $Y := X_1+X_2+\cdots+X_n$. Then

$$M_Y(t)=M_{X_1}(t)M_{X_2}(t)\cdots M_{X_n}(t).$$

Proof. By the definition of a moment generating function

$$M_Y(t)=E[e^{tY}]=E[e^{tX_1}e^{tX_2}\cdots e^{tX_n}].$$

Applying Theorem 182 with $g(x)=e^{tx}$ we conclude that the random variables e^{tX_1}, $e^{tX_2}, \ldots,$ e^{tX_n} are independent. Thus by Theorem 178 we conclude

$$M_Y(t)=E[e^{tX_1}]E[e^{tX_2}]\cdots E[e^{tX_n}]=M_{X_1}(t)M_{X_2}(t)\cdots M_{X_n}(t).$$

□

Most of the examples of this section, say Example 188, about the distribution of a sum of two independent random variables, could be done using the above theorem and our knowledge of the moment generating functions of the most important distributions. One would also need to use Theorem 123. But the approach presented is more general and applies in more situations, since not every random variable has moment generating function. Try to redo the examples using moment generating functions. Here is an instance of how the method is applied by redoing Example 188.

Example 192. If X and Y are independent, normally distributed, random variables, with parameters μ_X, σ_X^2 and μ_Y, σ_Y^2, then $X+Y$ is normally distributed with parameters $\mu_X+\mu_Y, \sigma_X^2+\sigma_Y^2$.

Solution. Recall the formula for the m.g.f. of a normal random variable with mean μ and variance σ^2, given in Subsection 9.6:

$$M(t) = e^{\left(\mu t + \frac{\sigma^2 t^2}{2}\right)}.$$

According to Theorem 191

$$\begin{aligned} M_{X+Y}(t) = M_X(t)M_Y(t) &= e^{\left(\mu_X t + \frac{\sigma_Y^2 t^2}{2}\right)} e^{\left(\mu_Y t + \frac{\sigma_Y^2 t^2}{2}\right)} \\ &= e^{\left((\mu_X + \mu_Y)t + \frac{(\sigma_X^2 + \sigma_Y^2)t^2}{2}\right)}. \end{aligned}$$

Since this is the m.g.f. of a normal random variable with mean $\mu_X + \mu_Y$ and variance $\sigma_X^2 + \sigma_Y^2$, by Theorem 123 it follows that $X + Y$ is a normal r.v. with those parameters. □

This method appears much simpler than the one used in Example 188 but that is an illusion. The calculations done in Example 188 are of the same difficulty as those that one needs to do in order to calculate the m.g.f. of the normal r.v., something that we did not in Subsection 9.6.

12 Covariance and correlation

12.1 Covariance

Definition 193. The *covariance* between X and Y, denoted $\mathrm{Cov}\,[X, Y]$, is defined by

$$\mathrm{Cov}\,[X, Y] = E[(X - E[X])(Y - E[Y])].$$

The definition, as stated, emphasized the intuition behind the notion of covariance. It is the average joint deviation of X and Y from their respective means. But, by expanding the right-hand side of the definition we can simplify it.

$$\begin{aligned} \mathrm{Cov}\,[X, Y] &= E[(X - E[X])(Y - E[Y])] = E[XY - XE[Y] - YE[X] + E[X]E[Y]] \\ &= E[XY] - E[X]E[Y] - E[Y]E[X] + E[X]E[Y] = E[XY] - E[X]E[Y]. \end{aligned}$$

This shows that if X and Y are independent random variables, then $E[XY] = E[X]E[Y]$ implying that $\mathrm{Cov}\,[X, Y] = 0$. The converse is not true. Take the random variables considered in Example 179. They satisfy $E[XY] = E[X]E[Y]$, equivalently $\mathrm{Cov}\,[X, Y] = 0$, but they are dependent. Below are the properties of covariance.

Proposition 194. (i) $\mathrm{Cov}\,[X, Y] = \mathrm{Cov}\,[Y, X]$;

(ii) $\mathrm{Cov}\,[X, X] = \mathrm{Var}\,[X]$;

(iii) $\mathrm{Cov}\,[aX, Y] = a\mathrm{Cov}\,[X, Y]$;

(iv) $\mathrm{Cov}\,[X, aY] = a\mathrm{Cov}\,[X, Y]$;

(v) $\mathrm{Cov}\,[\sum_{i=1}^n X_i, \sum_{j=1}^m Y_j] = \sum_{i=1}^n \sum_{j=1}^m \mathrm{Cov}\,[X_i, Y_j]$;

Proof. The first three parts are easy exercises. To simplify the notation for the fourth part, let $\mu_i := \mathbb{E}[X_i]$ and $\nu_i := E[Y_i]$. Then

$$\begin{aligned}\mathrm{Cov}\Big[\sum_{i=1}^n X_i, \sum_{j=1}^m Y_j\Big] &= E\Big[\Big(\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i\Big)\Big(\sum_{j=1}^m Y_j - \sum_{j=1}^m \nu_i\Big)\Big] = E\Big[\sum_{i=1}^n (X_i-\mu_i)\sum_{j=1}^m (Y_j-\nu_i)\Big] \\ &= E\Big[\sum_{i=1}^n\sum_{j=1}^m (X_i-\mu_i)(Y_j-\nu_i)\Big] = \sum_{i=1}^n\sum_{j=1}^m E[(X_i-\mu_i)(Y_j-\nu_i)],\end{aligned}$$

where the last equality follows from (29). □

With the notion of covariance, we can enhance our understanding of the variance.

Corollary 195. (i) For any random variables $X_1, X_2, \dots, X_n$

$$\mathrm{Var}\Big[\sum_{i=1}^n X_i\Big] = \sum_{i=1}^n \mathrm{Var}\,[X_i] + \sum_{i=1}^n \sum_{\substack{j=1 \\ j\neq i}}^n \mathrm{Cov}\,[X_i, X_j].$$

(ii) For any independent random variables $X_1, X_2, \dots, X_n$

$$\mathrm{Var}\Big[\sum_{i=1}^n X_i\Big] = \sum_{i=1}^n \mathrm{Var}\,[X_i].$$

Proof. We use the fact that $\mathrm{Var}\,[X] = \mathrm{Cov}\,[X, X]$ and then the last part of the proposition.

$$\mathrm{Var}\Big[\sum_{i=1}^n X_i\Big] = \mathrm{Cov}\Big[\sum_{i=1}^n X_i, \sum_{i=1}^n X_i\Big] = \sum_{i=1}^n\sum_{j=1}^m \mathrm{Cov}\,[X_i, X_j] = \sum_{i=1}^n \mathrm{Var}\,[X_i] + \sum_{i=1}^n \sum_{\substack{j=1 \\ j\neq i}}^n \mathrm{Cov}\,[X_i, X_j].$$

For the second part, just recall that if X_i and X_j are independent, then $\mathrm{Cov}\,[X_i, X_j] = 0$. □

Combining the second part of Corollary 195 with part (iii) of Proposition 111 we get that for independent r.v.s $X_1, X_2, \dots, X_n$ and any constants $a_1, a_2, \dots, a_n$

$$\mathrm{Var}\Big[\sum_{i=1}^n a_i X_i\Big] = \sum_{i=1}^n a_i^2 \mathrm{Var}\,[X_i]. \tag{33}$$

Let $X_1, X_2, \dots, X_n$ be independent random variables *having the same* cumulative distribution function $F(x)$. In particular, they have the same mean and variance. Such a sequence of random variables is called a *sample from the distribution F*. The *sample mean* is defined by

$$\bar{X} := \frac{\sum_{i=1}^n X_i}{n}.$$

Let the common expected value be $\mu := E[X_i]$, $i = 1, 2, 3 \dots$. Note that the expected value is a constant, while the sample mean is a random variable! The *sample variance* is defined by

$$S^2 := \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}.$$

Let the common variance be $\sigma^2 := \mathrm{Var}\,[X_i]$, $i = 1, 2, 3 \dots$. Note that the variance of X_i is a constant, while the sample variance is a random variable. Note also that the denominator in the definition of S^2 is $n-1$ not n as one might have expected. We will see why in a moment.

12.1.1 Examples and applications

Example 196 (Sample mean and sample variance)**.** Calculate the mean and the variance of the sample mean and the mean of the sample variance.

Solution. We have to do three things. a) The mean of $\bar{X}$. Using (29) we get

$$E[\bar{X}] = \sum_{i=1}^{n} E\left[\frac{X_i}{n}\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \mu.$$

b) The variance of $\bar{X}$. Using (33) with constants $a_i = 1/n$, $i = 1, 2, \ldots$, we get

$$\text{Var}\,[\bar{X}] = \sum_{i=1}^{n} \text{Var}\left[\frac{X_i}{n}\right] = \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}\,[X_i] = \frac{\sigma^2}{n}.$$

c) The mean of S^2. We start by adding and subtracting μ in each square in the definition of S^2, after that we expand the square.

$$\begin{aligned}(n-1)S^2 &= \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}\big((X_i-\mu) - (\bar{X}-\mu)\big)^2 \\ &= \sum_{i=1}^{n}(X_i-\mu)^2 + \sum_{i=1}^{n}(\bar{X}-\mu)^2 - 2(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\mu) \\ &= \sum_{i=1}^{n}(X_i-\mu)^2 + n(\bar{X}-\mu)^2 - 2(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\mu)\end{aligned}$$

Note that $\sum_{i=1}^{n}(X_i - \mu) = n(\bar{X} - \mu)$ and substitute it above.

$$(34) \qquad (n-1)S^2 = \sum_{i=1}^{n}(X_i-\mu)^2 + n(\bar{X}-\mu)^2 - 2n(\bar{X}-\mu)^2 = \sum_{i=1}^{n}(X_i-\mu)^2 - n(\bar{X}-\mu)^2.$$

Now take expectation from both sides to obtain the following.

$$(n-1)E[S^2] = \sum_{i=1}^{n} E[(X_i-\mu)^2] - nE[(\bar{X}-\mu)^2] = n\sigma^2 - n\text{Var}\,[\bar{X}] = n\sigma^2 - \sigma^2 = (n-1)\sigma^2.$$

Thus, we conclude that

$$E[S^2] = \sigma^2.$$

The above example is a fundamental first step in statistics. Event in nature that appears to be random has unknown to us probability distribution. If we want to estimate the average value and the variance of the event we draw a random sample and measure it. That is represented by the independent random variables $X_1, X_2, \ldots, X_n$ having the same cumulative distribution function $F(x)$. That is, our sample is $X_1(\omega), X_2(\omega), \ldots, X_n(\omega)$. We calculate the sample mean $\bar{X}(\omega)$ and the sample variance $S^2(\omega)$ and we would like to say that these values are close to the real μ and σ^2. Part a) of the example says that on average our sample mean will be equal to μ. Part b) of the example says how far away on average will the sample mean be from μ. Moreover, part b) of the example says that the larger the sample size, n, the smaller the variance of $\bar{X}$. Hence, samples with larger size will have sample means that are more likely to be close to the real value μ. Part c) says something similar to part a) but for the unknown parameter σ^2. Part c) says that the sample variance will on average be equal to σ^2. We say that $\bar{X}$ is an *estimator* for μ and S^2 is an *estimator* for σ^2. If the mean of an estimator is equal to the constant that it is supposed to estimate, then we say that it is an *unbiased estimator*. In the above examle, both $\bar{X}$ and S^2 are unbiased estimators. That is precisely the reason why in the definition of S^2 the denominator is $n-1$ and not n.

Sample mean and sample variance of normal distribution. Let $X_1, X_2, \ldots, X_n$ be independent, normal random variables, each with mean μ and variance σ^2. From a problem on the assignment we know that $(1/n)X_k$ is normal with mean μ/n and variance σ^2/n^2. Using Example 192 we conclude that the sample mean

$$\bar{X} := \frac{\sum_{i=1}^n X_i}{n}$$

is a normal random variable with mean μ and variance σ^2/n. From here we can go a step further to conclude something that we will need below, namely that

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

is a standard normal random variable, that is, it is a normal random variable with mean 0 and variance 1.

Our next goal is to determine the distribution of the sample variance S^2. For that purpose we need the following important theorem which we state without a proof.

Theorem 197. Let $X_1, X_2, \ldots, X_n$ be independent, normal random variables, each with mean μ and variance σ^2. Then, the sample mean $\bar{X}$ and the sample variance S^2 are independent random variables.

Exercise 198. Use Theorem 197, under the same assumptions, and Theorem 182, to show that
a) $\left(\frac{X_i-\mu}{\sigma}\right)^2$ for $i = 1, 2, \ldots, n$ are independent random variables, and
b) $(n-1)S^2/\sigma^2$ and $\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2$ are independent random variables.

The left-hand side and the right-hand side of (34) give an important identity

$$(n-1)S^2 = \sum_{i=1}^n (X_i-\mu)^2 - n(\bar{X}-\mu)^2.$$

Dividing both sides by σ^2 and reordering we get

$$\frac{(n-1)S^2}{\sigma^2} + \Big(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\Big)^2 = \sum_{i=1}^n \Big(\frac{X_i-\mu}{\sigma}\Big)^2 \tag{35}$$

Now let $Y := \frac{(n-1)S^2}{\sigma^2}$, we want to find the distribution of Y. For that goal we use moment generating functions. Since $\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2$ is the square of a standard normal random variable, by Definition 190 it has a chi-squared distribution with 1 degree of freedom. Thus, by (32) its moment generating function is $(1-2t)^{-1/2}$ for $t < 1/2$. By Exercise 198, Y is independent of $\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2$ so by Theorem 191 the moment generating function of the left-hand side of (35) is $M_Y(t)(1-2t)^{-1/2}$. On the right-hand side of (35) we have a sum of squares of n standard normal random variables. By Definition 190 it has a chi-squared distribution with n degree of freedom, and by (32) the moment generating function of the right-hand side is $(1-2t)^{-n/2}$ for $t < 1/2$. Thus, we get the equation

$$M_Y(t)(1-2t)^{-1/2} = (1-2t)^{-n/2}.$$

Solving it for $M_Y(t)$ we get $M_Y(t) = (1-2t)^{-(n-1)/2}$. This is the moment generating function of a chi-squared random variable with $(n-1)$-degrees of freedom. By the uniqueness theorem, Theorem 123, the distribution of Y is chi-squared with $(n-1)$-degrees of freedom. We summarize everything in the following result.

Theorem 199. Let $X_1, X_2, \ldots, X_n$ be independent, normal random variables, each with mean μ and variance σ^2. Let $\bar{X}$ be the sample mean and S^2 be the sample variance. Then, $\bar{X}$ and S^2 are independent random variables. The sample mean $\bar{X}$ is normal with mean μ and variance σ^2/n; while $(n-1)S^2/\sigma^2$ is a chi-squared random variable with $(n-1)$ degrees of freedom.

Sampling from a finite population. Suppose we want to find out the proportion p of people in the population who are in favour of a particular political candidate. The number p is called *population proportion.* Suppose the population has N individuals, they have fixed opinion, and let v_k be 1 if the k-th person in the population is in favour; and 0 otherwise. (The v_k's are constants.) That is, we have $p = \left(\sum_{k=1}^{N} v_k\right)/N$. Of course, we know the value on N but not the values of the v_k's and thus we do not know p. If we could interview every single individual in the population there would be nothing more to do. Since that is practically infeasible, we want to estimate the unknown proportion p. For that purpose we interview n randomly selected individuals and count how many of them are in favour and divide the result by n. This ratio, called *sample proportion*, is our estimate for the population proportion p. We want to find out how far, on average, will the sample proportion be from the true value of the population proportion. Thus, our sample space Ω is the set of all possible subsets of people numbered $\{1, 2, \ldots, N\}$ of size n. The probability that a particular sample of size n is picked is $1/\binom{N}{n}$. Let ω be a subset of $\{1, 2, \ldots, N\}$ of size n and let $S(\omega)$ be the number of people in ω who are in favour. Then the sample proportion $S(\omega)/n$ is the point estimate for the value of p. We want to find out what is the expectation and the variance of S/n. (Thus, the population proportion is a constant, while the sample proportion is a random variable.) For that purpose, we represent S as a sum of simpler random variables.

Let X_i be a random variable that equals 1 if the i-th person is in the sample and 0 otherwise. Then,

$$S = \sum_{k=1}^{N} v_k X_k.$$

Let us examine the variables X_k separately first. We have

$$E[X_k] = 1P(X_k = 1) + 0P(X_k = 0) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N};$$
$$E[X_k^2] = 1^2P(X_k = 1) + 0^2P(X_k = 0) = \frac{n}{N}.$$

In addition, or any $k \neq \ell$, we have

$$\begin{aligned} E[X_kX_\ell] &= 1P(X_k = 1, X_\ell = 1) + 0P(X_k = 1, X_\ell = 0) + 0P(X_k = 0, X_\ell = 1) + 0P(X_k = 0, X_\ell = 0) \\ &= \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}. \end{aligned}$$

These allow us to calculate the next quantities

$$\mathrm{Var}\,[X_k] = E[X_k^2] - (E[X_k])^2 = \frac{n}{N}\Big(1 - \frac{n}{N}\Big) = \frac{n(N-n)}{N^2};$$
$$\mathrm{Cov}\,[X_k, X_\ell] = E[X_kX_\ell] - E[X_k]E[X_\ell] = \frac{n(n-1)}{N(N-1)} - \frac{n}{N}\frac{n}{N} = -\frac{n(N-n)}{N^2(N-1)}.$$

Thus, we find

$$E[S] = \sum_{k=1}^N E[v_kX_k] = \sum_{k=1}^N v_kE[X_k] = \frac{n}{N}\sum_{k=1}^N v_k = np,$$

and using part (i) of Corollary 195, we get

$$\begin{aligned} \mathrm{Var}\,[S] &= \sum_{k=1}^N \mathrm{Var}\,[v_kX_k] + \sum_{k=1}^N \sum_{\substack{\ell=1\\ \ell\neq k}}^N \mathrm{Cov}\,[v_kX_k, v_\ell X_\ell] \\ &= \sum_{k=1}^N v_k^2\mathrm{Var}\,[X_k] + \sum_{k=1}^N \sum_{\substack{\ell=1\\ \ell\neq k}}^N v_kv_\ell\mathrm{Cov}\,[X_k, X_\ell] \\ &= \frac{n(N-n)}{N^2}\sum_{k=1}^N v_k^2 - \frac{n(N-n)}{N^2(N-1)}\sum_{k=1}^N \sum_{\substack{\ell=1\\ \ell\neq k}}^N v_kv_\ell \\ &= \frac{n(N-n)}{N^2(N-1)}\Big((N-1)\sum_{k=1}^N v_k^2 - \sum_{k=1}^N \sum_{\substack{\ell=1\\ \ell\neq k}}^N v_kv_\ell\Big) \\ &= \frac{n(N-n)}{N^2(N-1)}\Big(N\sum_{k=1}^N v_k^2 - \Big(\sum_{k=1}^N v_k^2 + \sum_{k=1}^N \sum_{\substack{\ell=1\\ \ell\neq k}}^N v_kv_\ell\Big)\Big) \\ &= \frac{n(N-n)}{N^2(N-1)}\Big(N\sum_{k=1}^N v_k^2 - \Big(\sum_{k=1}^N v_k\Big)^2\Big), \end{aligned}$$

where for the last equality, we used (9). Recall now that $\sum_{k=1}^{N} v_k = Np$ and since v_k is equal to 0 or 1, we have $v_k^2 = v_k$, that is $\sum_{k=1}^{N} v_k^2 = Np$. Substituting above, we continue

$$\mathrm{Var}\,[S] = \frac{n(N-n)}{N^2(N-1)}\left(N^2p - N^2p^2\right) = \frac{n(N-n)}{(N-1)}p(1-p).$$

Finally, we have

$$\begin{aligned} E\left[\frac{S}{n}\right] &= \frac{1}{n}E[S] = p, \\ \mathrm{Var}\left[\frac{S}{n}\right] &= \frac{1}{n^2}\mathrm{Var}\,[S] = \frac{(N-n)}{n(N-1)}p(1-p). \end{aligned}$$

In conclusion, we see that S/n is an unbiased estimator for p, and the larger the sample size the smaller the variance of the sample proportion. That is, there is a bigger chance that S/n will be close to p. In particular, if $n = N$, then the variance is 0, that is, $S/n = p$, which should be the case.

Suppose in the population proportion example, there are m people in the population who are in favour, that is $p = m/N$. We now rename a few objects. The 'people' will be called 'balls'. Those 'in favour' are the 'white balls'. Those 'against' are the 'black balls'. We select n balls at random and S denotes the number of white balls. Then S has hypergeometric distribution and according to the above its mean and variance are

$$\begin{aligned} E[S] &= np = \frac{nm}{N}; \\ \mathrm{Var}\,[S] &= n^2\frac{(N-n)}{n(N-1)}p(1-p) = \frac{(N-n)}{(N-1)}\frac{nm}{N}\frac{N-m}{N}. \end{aligned} \tag{36}$$

12.2 Correlation

Suppose X and Y are two random variables with strictly positive variance: $\mathrm{Var}\,[X] > 0$ and $\mathrm{Var}\,[Y] > 0$. Then the correlation between X and Y, denoted by $\rho[X,Y]$ is defined by

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

Proposition 200. For any two random variables X and Y with strictly positive variance: $\mathrm{Var}\,[X] > 0$ and $\mathrm{Var}\,[Y] > 0$, we have

$$-1 \leq \rho[X,Y] \leq 1.$$

Proof. Let σ_X^2 and σ_Y^2 be the variances of X and Y respectively. Then

$$0 \leq \mathrm{Var}\left[\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right] = \frac{\mathrm{Var}\,[X]}{\sigma_X^2} + \frac{\mathrm{Var}\,[Y]}{\sigma_Y^2} + 2\frac{\mathrm{Cov}\,[X,Y]}{\sigma_X\sigma_Y} = 2(1+\rho[X,Y]).$$

From here we conclude that $\rho[X,Y] \geq -1$. Analogously, we have

$$0 \leq \mathrm{Var}\left[\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right] = \frac{\mathrm{Var}\,[X]}{\sigma_X^2} + \frac{\mathrm{Var}\,[Y]}{\sigma_Y^2} - 2\frac{\mathrm{Cov}\,[X,Y]}{\sigma_X\sigma_Y} = 2(1-\rho[X,Y]).$$

From here we conclude that $\rho[X,Y] \leq 1$. □

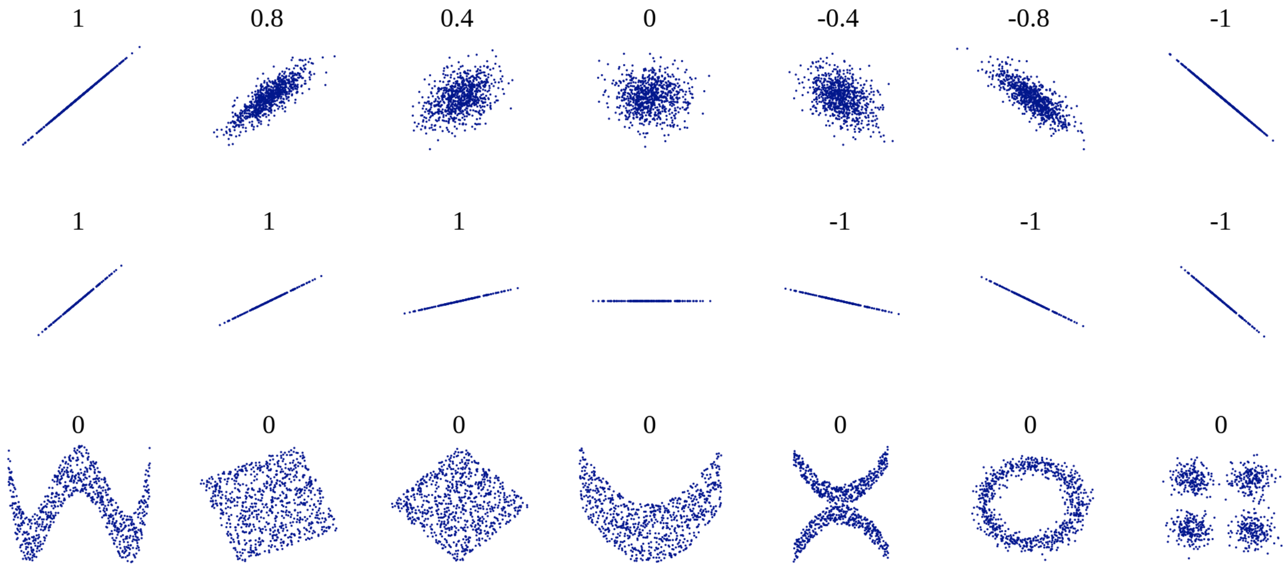


Figure 9: Different correlation cases between X and Y

In fact, since $\operatorname{Var}[Z] = 0$ implies that Z is constant with probability 1 (this intuitive fact will be rigorously proved later), we see from the proof of the proposition that $\rho(X, Y) = 1$ implies that $Y = a + bX$, where $b = \sigma_Y / \sigma_X > 0$ and $\rho(X, Y) = -1$ implies that $Y = a + bX$, where $b = -\sigma_Y / \sigma_X < 0$.

Exercise 201. Show that if $Y = a + bX$ then $\rho(X, Y)$ is either 1 or -1 depending on the sign of b.

Thus, the correlation coefficient is a measure of the degree of linearity between X and Y. A value of $\rho(X, Y)$ near 1 or -1 indicates a high degree of linearity between X and Y, whereas a value near 0 indicates a lack of such linearity. A positive value of $\rho(X, Y)$ indicates that Y tends to increase when X does, whereas a negative value indicates that Y tends to decrease when X increases. If $\rho(X, Y) = 0$, then X and Y are said to be uncorrelated.

Figure 9 depicts the points $(X(\omega), Y(\omega))$ in the plane for all possible values of $\omega \in \Omega$ and different pairs of random variables X and Y. For each data set, the correlation is calculated and given above it. Note that the correlation reflects the noisiness and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). The figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of Y is zero.

Exercise 202. Let $X_1, X_2, \cdots, X_n$ be independent and identically distributed random variables having variance σ^2. Show that $\operatorname{Cov}[X_i - \bar{X}, \bar{X}] = 0$.

13 Limit theorems

Note that from the definition of expected value, we have that if a random variable X is non-negative, then $E[X] \geq 0$. This implies that if $X \geq Y$ then $X - Y \geq 0$, and taking expectation we have $E[X] - E[Y] = E[X - Y] \geq 0$, that is $E[X] \geq E[Y]$.

Proposition 203 (Markov's inequality)**.** If $X \geq 0$ is a random variable, then for any number $t \geq 0$ we have

$$P(X \geq t) \leq \frac{E[X]}{t}.$$

Proof. Fix t, and let the random variable I be the indicator function of the event $\{X \geq t\}$, that is

$$I(\omega) = \begin{cases} 1 & \text{if } X(\omega) \geq t, \\ 0 & \text{otherwise.} \end{cases}$$

Since $X \geq 0$, we have $I \leq X/t$. Taking expectations of both sides, we get $E[I] \leq E[X/t] = E[X]/t$. Using Example 101, we have $E[I] = P(X \geq t)$, and we are done. □

Corollary 204 (Chebyshev's inequality I)**.** If X is a random variable with mean μ and variance σ^2, then for any value of $k > 0$, we have

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

Proof. Apply Markov's inequality to the non-negative random variable $(X - \mu)^2$ and $t = k^2$:

$$P((X - \mu)^2 \geq k^2) \leq \frac{E[(X - \mu)^2]}{k^2}.$$

It remains to note that the inequality $(X - \mu)^2 \geq k^2$ is equivalent to $|X - \mu| \geq k$ and that $E[(X - \mu)^2] = \sigma^2$. □

Corollary 205 (Chebyshev's inequality II)**.** If X is a random variable with mean μ and variance σ^2, then for any value of $n > 0$, we have

$$P(X \in [\mu - n\sigma, \mu + n\sigma]) \geq 1 - \frac{1}{n^2}.$$

Proof. One can rewrite Chebyshev's inequality as follows. First let $k := n\sigma$, where $n > 0$ to obtain

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

Second, observe that

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

The inequality follows from here. □

For example, let X be the grade a student from this class receives on the midterm exam. The average score on the mideterm is 72% and the standard deviation is 4.7. Thus, with $n = 2$, the second Chebyshev's inequality says that at least $1 - 1/n^2 = 75\%$ of the students have a score in the interval $[\mu - n\sigma, \mu + n\sigma] = [62.6\%, 81.4\%]$. Interval like that, that contains a specified percentage of the population values (the values that a random variable may take) is called a *tolerance interval.*

Chebyshev's theorem applies to any random variable X no matter what is its distribution. That is why it is a very conservative estimate. This means that $1 - 1/n^2$ does not increase very fast compared to the length of the interval $[\mu - n\sigma, \mu + n\sigma]$. If we have additional information, for example, we know that X is normally distributed, then we have the following much better tolerance interval.

Proposition 206 (The empirical rule)**.** If X is a normal random variable with mean μ and variance σ^2, then for any value of $n > 0$ we have

$$P(X \in [\mu - n\sigma, \mu + n\sigma]) \geq 1 - \sqrt{\frac{2}{\pi}} \frac{1}{ne^{n^2/2}}.$$

Proof. Recall that $Z := (X - \mu)/\sigma$ is a standard normal random variable, and recall the definition of the cumulative distribution function $\Phi(x)$ of Z, see (23). we have

$$\begin{aligned} P(X \in [\mu - n\sigma, \mu + n\sigma]) &= P\Big(-n \leq \frac{X - \mu}{\sigma} \leq n \Big) = P(-n \leq Z \leq n) = P(Z \leq n) - P(Z \leq -n) \\ &= \Phi(n) - \Phi(-n) = \Phi(n) - (1 - \Phi(n)) = 1 - 2(1 - \Phi(n)) \\ &\geq 1 - 2\frac{1}{\sqrt{2\pi}} \frac{1}{n} e^{-n^2/2}, \end{aligned}$$

where we used (without proof) the inequality

$$1 - \Phi(n) \leq \frac{1}{\sqrt{2\pi}} \frac{1}{n} e^{-n^2/2},$$

holding for all $n > 0$. □

For example, let X be the grade a student from this class receives on the midterm exam. Suppose that X has a normal (or approximately normal) distribution. The average score on the mideterm is 72% and the standard deviation is 4.7. Thus, with $n = 2$, the empirical rule says that at least $1 - \sqrt{\frac{2}{\pi}} \frac{1}{ne^{n^2/2}} = 94.60\%$ of the students have a score in the interval $[\mu - n\sigma, \mu + n\sigma] = [62.6\%, 81.4\%]$.

So, is X normally distributed? Figure 10 represents the *relative frequency histogram* of the midterm test scores. For example, a bit over 11% of the students scored in the $(70\%, 75\%]$ range; and a bit over 8% of the students scored in the $(80\%, 85\%]$ range; and so on. The relative frequency histogram indicates that the distribution of X is not normal. There is a peak at 60% and a peak at 80% and the distribution does not appear to be symmetric. So the second Chebyshev's inequality should be used rather than the empirical rule.

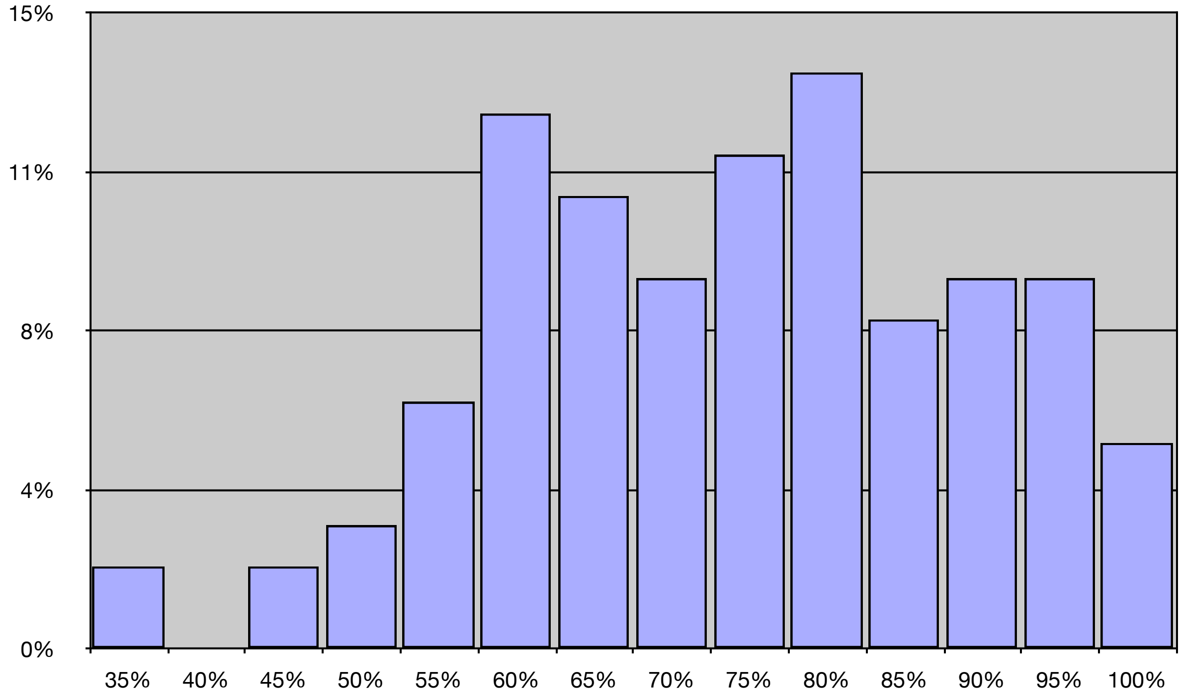


Figure 10: Relative frequency histogram of the midterm test scores

The next corollary says that if the variance of a random variable is zero, then without loss of generality this random variable may be treated as a constant.

Corollary 207. Let X be a random variable with mean μ. If $\mathrm{Var}\,[X] = 0$ then $P(X = \mu) = 1$.

Proof. By Chebyshev's inequality we have, for any $n \geq 1$, that

$$P(|X - \mu| \geq 1/n) \leq \sigma^2 n^2 = 0,$$

that is $P(|X - \mu| \geq 1/n) = 0$ for all $n \geq 1$. Define the events $A_n := \{|X - \mu| \geq 1/n\}$ and note that $A_n \subseteq A_{n+1}$. Hence these events form an increasing sequence of events. Their union is

$$\bigcup_{n=1}^{\infty} A_n = \{|X - \mu| > 0\} = \{X \neq \mu\}.$$

Thus, by Proposition 161, we have

$$0 = \lim_{n\to\infty} P(|X - \mu| \geq 1/n) = \lim_{n\to\infty} P(A_n) = P\Big(\bigcup_{n=1}^{\infty} A_n\Big) = P(X \neq \mu).$$

Alternatively, this is the same as $P(X = \mu) = 1$. □

A particular case of the next theorem can also be considered a corollary of the Chebyshev's inequality.

Theorem 208 (The weak law of large numbers)**.** Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables, each with mean $E[X_i] = \mu$. Then, for any number $\epsilon > 0$, we have

$$P\Big(\Big|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\Big| \geq \epsilon\Big) \to 0 \tag{37}$$

as n approaches infinity.

Proof. We only proof the result when the variance, σ^2, of X_i is also finite. Since

$$E\Big[\frac{X_1 + X_2 + \cdots + X_n}{n}\Big] = \mu \text{ and } \mathrm{Var}\Big[\frac{X_1 + X_2 + \cdots + X_n}{n}\Big] = \frac{\sigma^2}{n},$$

applying the Chebyshev's inequality, gives

$$0 \leq P\Big(\Big|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\Big| \geq \epsilon\Big) = \frac{\sigma^2}{n\epsilon^2}.$$

Letting n approach infinity, the last probability is sandwiched between 0 and something that approaches zero. So, it must approach zero as well. □

Note that (37) may also be written as: for any number $\epsilon > 0$, we have

$$P\Big(\Big|\frac{X_1 + X_2 + \cdots + X_n - n\mu}{n}\Big| \leq \epsilon\Big) \to 1 \tag{38}$$

as n approaches infinity.

Theorem 209 (Strong law of large numbers)**.** Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables, each with mean $E[X_i] = \mu$. There is an event A with probability $P(A) = 1$ such that for every $\omega \in A$ we have

$$\frac{X_1(\omega) + X_2(\omega) + \cdots + X_n(\omega)}{n} \to \mu$$

as n approaches infinity.

In other words, we say that $(X_1 + X_2 + \cdots + X_n)/n$ converges to μ with probability 1 as n approaches infinity.

Suppose that a sequence of independent trials of some experiment is performed. Let E be an event and denote by $P(E)$ the probability that E occurs on any particular trial. Define the random variable X_i to be 1 if E occurs on the i-th trial and to be 0 otherwise. Then $(X_1+X_2+\cdots+X_n)/n$ is the proportion of times that E occurs, and the strong law of large numbers says that in the limit the proportion of times that E occurs is $E[X_i] = P(E)$.

13.1 The central limit theorem

The central limit theorem is one of the most remarkable results in probability theory. It states roughly that the sum of a large number of independent random variables has a distribution that is approximately normal. That not only provides a simple method for computing approximate probabilities for sums of independent random variables, but also helps explain the remarkable fact that the empirical frequencies of so many natural populations exhibit bell-shaped (that is, normal) curves.

Theorem 210 (The central limit theorem)**.** Let $X_1, X_2, \dots$ be a sequence of independent identically distributed random variables, each with mean $E[X_i] = \mu$ and variance $\mathrm{Var}\,[X_i] = \sigma^2$. Then, for any number $\epsilon > 0$, we have

$$P\Big(\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leq \epsilon\Big) \to \Phi(\epsilon),$$

as n approaches infinity, where the function Φ was defined in (23).

For example, let X_i be a equal to 1 with probability $p \in [0, 1]$; and 0 with probability $q := 1-p$. Such a random variable is called *Bernoulli random variable.* Suppose $X_1, X_2, \dots$ are independent. Then, the sum $X_1 + X_2 + \cdots + X_n$ is the number of successes in n independent trials, each resulting in success with probability p and failure with probability $1 - p$. Thus, $X := X_1 + X_2 + \cdots + X_n$ is a binomial random variable with parameters n and p. In addition, we have $E[X_i] = p$ and $\mathrm{Var}\,[X_i] = pq$. Thus, by the central limit theorem we have

$$P\Big(\frac{X_1 + X_2 + \cdots + X_n - np}{\sqrt{pq}\sqrt{n}} \leq \epsilon\Big) \to \Phi(\epsilon),$$

as n approaches infinity; or equivalently

$$P\Big(\frac{X - np}{\sqrt{npq}} \leq \epsilon\Big) \to \Phi(\epsilon),$$

as n approaches infinity. This is precisely the statement of Theorem 156.

Example 211. An astronomer is interested in measuring, in light years, the distance from his observatory to a distant star. Although the astronomer has a measuring technique, he knows that, because of changing atmospheric conditions and normal error each time a measurement is made it will not yield the exact distance but merely an estimate. As a result the astronomer plans to make a series of measurements and then use the average value of these measurements as his estimated value of the actual distance. If the astronomer believes that the values of the measurements are independent and identically distributed random variables having a common mean d (the actual distance) and a common standard deviation of 2 (light years), how many measurements need he make to be at least 95% certain, that his estimated distance is accurate to within ±0.5 light year?

Solution. Let $X_1, X_2, \dots, X_n$ be the n observations that the astronomer decides to make. To be 95% certain means that if the astronomer takes 100 samples of size n (that is $100 \times n$ measurements in total) roughly 95 samples of size n will have an average

$$\frac{X_1 + X_2 + \cdots + X_n}{n}$$

that is within 0.5 light years of d. Thus, we a looking for the value of n such that

$$\begin{aligned}
0.95 \leq P\Big(-0.5 \leq \frac{X_1+X_2+\cdots+X_n}{n} - d \leq 0.5\Big) &= P\Big(-0.5 \leq \frac{X_1+X_2+\cdots+X_n-nd}{2\sqrt{n}}\frac{2}{\sqrt{n}} \leq 0.5\Big)\\
&= P\Big(-\frac{\sqrt{n}}{4} \leq \frac{X_1+X_2+\cdots+X_n-nd}{2\sqrt{n}} \leq \frac{\sqrt{n}}{4}\Big)\\
&\approx P\Big(-\frac{\sqrt{n}}{4} \leq Z \leq \frac{\sqrt{n}}{4}\Big)\\
&= \Phi\Big(\frac{\sqrt{n}}{4}\Big) - \Phi\Big(-\frac{\sqrt{n}}{4}\Big)\\
&= 2\Phi\Big(\frac{\sqrt{n}}{4}\Big) - 1.
\end{aligned}$$

Thus, $\Phi\Big(\frac{\sqrt{n}}{4}\Big) \geq 0.975$ and from the cumulative normal table, we find that $\frac{\sqrt{n}}{4} \geq 1.96$ or $n \geq 61.47$. The astronomer needs to make 62 observations.

Of course, the central limit theorem does not tell us how large should n be so that the approximation above is good. In fact the speed of convergence in the central limit theorem depends on the distribution of X_i, which in this case is unknown. One way to insure ourselves is to overshoot and make way more measurements than it is probably necessary, as shown by a use of Chebyshev's theorem. Since

$$E\Big[\frac{X_1+X_2+\cdots+X_n}{n}\Big] = d \text{ and } \mathrm{Var}\Big[\frac{X_1+X_2+\cdots+X_n}{n}\Big] = \frac{4}{n}$$

the Chebyshev's inequality gives

$$P\Big(\Big|\frac{X_1+X_2+\cdots+X_n}{n} - d\Big| \geq 0.5\Big) \leq \frac{(4/n)}{(0.5)^2}.$$

This time we want to make the probability on the left-hand side be smaller than 0.05 (why?). The Chebyshev's inequality tells us that this would be the case when

$$\frac{(4/n)}{(0.5)^2} \leq 0.05.$$

Solving for n, gives $n \geq 320$. □

Example 212. The number of students that enroll in a psychology course is a Poisson random variable with mean 100. The professor in charge of the course has decided that if the number enrolling is 120 or more he will teach the course in two separate sections, whereas if fewer than 120 students enroll he will teach all of the students together in a single section. What is the probability that the professor will have to teach two sections?

Solution. The exact solution $e^{-100}\sum_{i=120}^{\infty} 100^i/i!$ does not readily yield a numerical answer. However, by recalling that a Poisson random variable with mean 100 is the sum of 100 independent Poisson random variables with mean 1, we can use the central limit theorem to find an approximate solution. Let X denote the number of students that enroll in the course. Using the fact that the mean and the variance of a Poisson random variable with parameter 1 are 1, we get

$$P(X \geq 120) = P\Big(\frac{X - 100(1)}{\sqrt{100(1)}} \geq \frac{120-100}{\sqrt{100}}\Big) \approx P(Z \geq 2) = 1 - \Phi(2) = 0.0228.$$

14 Conditional distributions

Recall that for any two events A and B the conditional probability of A given B is defined by

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

provided that $P(B) > 0$. If $P(B) = 0$ then event B practically never occurs so it does not make sense to talk about the probability of A given that B has occurred.

Let X and Y be discrete random variables with joint p.m.f. $p(x, y)$ and marginal p.m.f. $p_X(x)$ and $p_Y(y)$. If $p_Y(y) > 0$, then we define the *conditional probability mass function of X given that $Y = y$* by

$$p_{X|Y}(x|y) := P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{p(x, y)}{p_Y(y)}$$

Note that if X and Y are independent then $p(x, y) = p_X(x)p_Y(y)$ implying that $p_{X|Y}(x|y) = p_X(x)$ for every y.

If X and Y are continuous random variables with joint probability density function $f(x, y)$ and marginal p.d.f. $f_X(x)$ and $f_Y(y)$ then define the conditional probability density function of X, given that $Y = y$ (provided that $f_Y(y) > 0$), by

$$f_{X|Y}(x|y) := \frac{f(x, y)}{f_Y(y)}.$$

Note again that if X and Y are independent then $f(x, y) = f_X(x)f_Y(y)$ implying that $f_{X|Y}(x|y) = f_X(x)$ for every y.

What we actually did was defined a new random variable Z, often denoted as '$X|Y = y$', with p.d.f. $f_Z(x) := f_{X|Y}(x|y)$. The random variable Z takes the same values as X but with re-weighted probabilities. That the function $f_{X|Y}(x|y)$, in the argument x, is indeed a p.d.f. follows from the fact that it is non-negative and

$$\int_{-\infty}^{\infty} f_{X|Y}(x|y)\, dx = \int_{-\infty}^{\infty} \frac{f(x, y)}{f_Y(y)}\, dx = \frac{1}{f_Y(y)} \int_{-\infty}^{\infty} f(x, y)\, dx = \frac{1}{f_Y(y)} f_Y(y) = 1.$$

The situation in the jointly discrete case is similar.

Example 213. Suppose X and Y are independent random variables that are binomial(m, p) and binomial$(N - m, p)$, respectively. Calculate the probability mass function of X, given that $X + Y = n$.

Solution. We know from class that $X + Y$ is binomial(N, p). Since $\mathbb{P}(X = k, X + Y = n) = \mathbb{P}(X = k, Y = n - k) = \mathbb{P}(X = k)\mathbb{P}(Y = n - k)$, we have

$$\begin{aligned}\mathbb{P}(X = k|X + Y = n) &= \frac{\mathbb{P}(X = k, X + Y = n)}{\mathbb{P}(X + Y = n)} = \frac{\mathbb{P}(X = k)\mathbb{P}(Y = n - k)}{\mathbb{P}(X + Y = n)} \\ &= \frac{\binom{m}{k} p^k (1 - p)^{m-k} \binom{N-m}{n-k} p^{n-k} (1 - p)^{(N-m)-(n-k)}}{\binom{N}{n} p^n (1 - p)^{N-n}}\end{aligned}$$

$$= \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}.$$

This answer is already striking since it does not depend on p, but more shockingly, this is the hypergeometric distribution. This is the conditional distribution of X given that $X+Y=n$ is the same as the number of white balls that we get when we draw n balls (without replacement) from an urn, containing N balls, of which m are white and the rest $N-m$ are black.

All this means that the conditional p.m.f. of X, given that $X+Y=n$ is hypergeometric with parameters n, N, and m. □

Example 214. Suppose that the joint density of X and Y is given by

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

Find $P(X>1|Y=y)$.

Solution. The probability $P(X>1|Y=y)$ should be understood as $P(\text{`}X|Y=y\text{'} > 1)$, so we need to find the density function of '$X|Y=y$', that is the function $f_{X|Y}(x|y)$. So, first we need to find the marginal density of Y:

$$f_Y(y) = \int_0^\infty \frac{e^{-x/y}e^{-y}}{y}\,dx = e^{-y}\int_0^\infty \frac{e^{-x/y}}{y}\,dx = e^{-y}\Big(-e^{-x/y}\Big|_{x=0}^\infty\Big) = e^{-y}(-0+1) = e^{-y}.$$

Thus, the conditional density of X given that $Y=y$:

$$f_{X|Y}(x|y) := \frac{f(x,y)}{f_Y(y)} = \frac{e^{-x/y}e^{-y}/y}{e^{-y}} = \frac{e^{-x/y}}{y}.$$

We should keep in mind that this formula is valid for $x>0$ and $y>0$, otherwise $f_{X|Y}(x|y)=0$. (Note incidentally, that $Z := (X|Y=y)$ is an exponential random variable with parameter $1/y$.) We can now compute the desired probability

$$P(X>1|Y=y) = \int_1^\infty f_{X|Y}(x|y)\,dx = \int_1^\infty \frac{e^{-x/y}}{y}\,dx = -e^{-x/y}\Big|_{x=1}^\infty = e^{-1/y}. \quad \square$$

Note in the last example, that the expression $P(X>1|Y=y)$ is to be understood as the probability that the random variable $X|Y=y$ is bigger than 1.

On the other hand, if X and Y are (jointly) continuous random variables, then $P(X=1|Y=y)=0$ since this is the probability that the (continuous) random variable $X|Y=y$ is equal to 1.

Finally, the expression $P(X>1|Y\geq y)$ is to be understood as the probability of the event $\{X>1\}$ given that the event $\{Y\geq y\}$ occurred. That is, for $y>0$ we have

$$P(X>1|Y\geq y) = \frac{P(X>1, Y\geq y)}{P(Y\geq y)} = \frac{\int_y^\infty\int_1^\infty f(x,t)\,dxdt}{\int_y^\infty f_Y(t)\,dt} = \frac{\int_y^\infty\int_1^\infty \frac{e^{-x/t}e^{-t}}{t}\,dxdt}{\int_y^\infty e^{-t}\,dt}$$

$$= \frac{\int_y^\infty e^{-t}\Big(\int_1^\infty \frac{e^{-x/t}}{t}\,dx\Big)dt}{e^{-y}} = \frac{\int_y^\infty e^{-t}e^{-1/t}dt}{e^{-y}}.$$

The last integral is not easy to evaluate exactly. Using a computer, one can evaluate the last integral for particular values of y. For example

$$P(X > 1|Y \geq 1) = 0.207.$$

15 Appendix A: The Prosecutor's Fallacy

The following example is taken from [10].

Suppose a city has a population of $1,000,000$ people. Suppose one of them commits a crime. Suppose eye-witness testimony provides police with information that leads to 10 suspects and eventually one of them is charged with the crime and brought to trial. The prosecutor makes the following argument: If this person is innocent, the probability that he matches the eye-witness description is very very small. It is therefore unlikely that this person is actually innocent. In other words, assuming the defendant was actually innocent, the chance of his matching the eye-witness description is just too small to believe he is not guilty. Therefore, he must be guilty. Using the language of conditional probability, the prosecutor is saying the following: Let $M :=$ the event the defendant matches the eye-witness description and let $I :=$ the event that the defendant is innocent. Now the prosecutor says, because $P(M|I)$ is small, we should not believe his is actually innocent.

With the numbers in this example, the prosecutor would argue, $P(M|I) = 9/999,999 = 0.000009$, because, assuming there are $999,999$ innocent people, there are only 9 that are innocent and match the description. This is a very small probability. There is only a 0.0009% probability that, if the defendant were innocent, he would match the eye-witness description. Therefore, if the defendant were innocent, the chance of matching the description is too small to believe he is innocent. Therefore, he must be guilty.

The fallacy is in a misunderstanding of conditional probability. A skilled prosecutor could easily persuade an uneducated jury with such an argument, arriving at a conviction based on rhetorical skills and a mathematical trick.

The jury must actually decide the following: "What is the probability the defendant is innocent given that he matches the eye-witness description? Thats different from the question "What is the probability the defendant matches the description given that he is innocent? A good lawyer could easily make the two sound the same. The jury must actually consider the conditional probability $P(I|M)$. That is, what is the probability of innocence given that the defendant matches the description? In this case, $P(I|M) = 9/10 = 0.90 = 90\%$. That is, assuming the defendant matches the eye-witness evidence, there is still a 90% chance that he is innocent. That completely changes the argument. Reversing order in which we write the probability, our poor defendant has gone from looking almost certainly guilty to almost certainly innocent!

The lesson is that we need to be careful with conditional probability and that being sloppy can have some really serious consequences.

To see some real life examples where this has really happened read about the Sally Clark case in Britain (1998) , the OJ Simpson case (1995) and People vs. Collins (1968).

Acknowledgements. I am thankful to Robert James Lee, Melissa Mccorriston, and Liuyan Li for pointing out numerous typos and inconsistencies.

References

[1] H. Gordon, *Discrete Probability*, Springer 1997.

[2] C. Grinstead, J.-L. Snell, *Introduction to Probability, 2nd ed.*, 1997, http://www.math.dartmouth.edu/ prob/prob/prob.pdf.

[3] D. Murdoch, *Private communications*, September 2012.

[4] S. Ross, *A First Course in Probability, 6th ed.*, Prentice Hall, 2002.

[5] http://en.wikipedia.org/wiki/People_v._Collins

[6] http://en.wikipedia.org/wiki/Gamma_function

[7] http://en.wikipedia.org/wiki/Correlation_and_dependence

[8] R. Durrett, *Probability Theory and Examples, 2nd ed.*, Duxbury Press, 1996.

[9] http://www.youtube.com/watch?v=mhlc7peGlGg&feature=related

[10] http://faculty.mc3.edu/cvaughen/

[11] B. Bowerman, R. O'Connell, J. Aitken, J. Adcock, *Business Statistics in Practice, 2nd Canadian ed.*, 2012.