Lectures on Probability Theory and Mathematical Statistics Second Edition

Marco Taboga

Contents

Preface

This book is a collection of lectures on probability theory and mathematical statistics that I have been publishing on the website StatLect.com since 2010. Visitors to the website have been constantly increasing and, while I have received positive feedback from many of them, some have suggested that the lectures would be easier to study if they were collected and printed as a traditional paper textbook. I followed their suggestion and started to work on this book. The painful editing work needed to convert the webpages to book chapters took a significant portion of my spare time for almost one year. I hope the result is vaguely satisfactory, but I am sure that not all typos and mistakes have been eliminated. For this, I humbly ask the forgiveness of my readers.

There were two main reasons why I started writing these lectures. First of all, I thought it was difficult to find a thorough yet accessible treatment of the basics of probability theory and mathematical statistics. While there are many excellent textbooks on these subjects, the easier ones often do not touch on many important topics, while the more complete ones frequently require a level of mathematical sophistication not possessed by most people. In these lectures I tried to give an accessible introduction to topics that are not usually found in elementary books. Secondly, I tried to collect in these lectures results and proofs (especially on probability distributions) that are hard to find in standard references and are scattered here and there in more specialistic books. I hope this will help my readers save some precious time in their study of probability and statistics.

The plan of the book is as follows: Part 1 is a review of some elementary mathematical tools that are needed to understand the lectures; Part 2 introduces the fundamentals of probability theory; Part 3 presents additional topics in probability theory; Part 4 deals with special probability distributions; Part 5 contains more details about the normal distribution; Part 6 discusses the basics of asymptotic theory (sequences of random variables and their convergence); Part 7 is an introduction to mathematical statistics.

Preface to second edition

Besides some minor editing, the main changes I introduced in the second edition are as follows: I have expanded the lectures on the Gamma and Beta functions, and on the multinomial distribution; I have entirely rewritten the lecture on the binomial distribution; I have added solved exercises to several chapters that did not have any.

Dedication

This book is dedicated to Emanuela and Anna.

Part I

Mathematical tools

Chapter 1

Set theory

This lecture introduces the basics of set theory.

1.1 Sets

A set is a collection of objects. Sets are usually denoted by a letter and the objects (or elements) belonging to a set are usually listed within curly brackets.

Example 1 *Denote by the letter S the set of the natural numbers less than or equal to 5. Then, we can write*

$$S = \{1, 2, 3, 4, 5\}$$

Example 2 *Denote by the letter A the set of the first five letters of the alphabet. Then, we can write*

$$A = \{a, b, c, d, e\}$$

Note that a set is an unordered collection of objects, i.e., the order in which the elements of a set are listed does not matter.

Example 3 *The two sets*

$$\{a, b, c, d, e\}$$

and

$$\{b, d, a, c, e\}$$

are considered identical.

Sometimes a set is defined in terms of one or more properties satisfied by its elements. For example, the set

$$S = \{1, 2, 3, 4, 5\}$$

could be equivalently defined as

$$S = \{n \in \mathbb{N} : n \leq 5\}$$

which reads as follows: "S is the set of all natural numbers n such that n is less than or equal to 5", where the colon symbol (:) means "such that" and precedes a list of conditions that the elements of the set need to satisfy.

Example 4 *The set*

$$S = \left\{ n \in \mathbb{N} : \frac{n}{4} \in \mathbb{N} \right\}$$

is the set of all natural numbers n such that n divided by 4 is also a natural number, i.e.,

$$S = \{4, 8, 12, \ldots\}$$

1.2 Set membership

When an element a belongs to a set A, we write

$$a \in A$$

which reads "a belongs to A" or "a is a member of A".

On the contrary, when an element a does not belong to a set A, we write

$$a \notin A$$

which reads "a does not belong to A" or "a is not a member of A".

Example 5 *Let the set S be defined as follows:*

$$A = \{2, 4, 6, 8, 10\}$$

Then, for example,

$$4 \in A$$

and

$$7 \notin A$$

1.3 Set inclusion

If A and B are two sets, and if every element of A also belongs to B, then we write

$$A \subseteq B$$

which reads "A is included in B", or

$$B \supseteq A$$

and we read "B includes A". We also say that A is a subset of B.

Example 6 *The set*

$$A = \{2, 3\}$$

is included in the set

$$B = \{1, 2, 3, 4\}$$

because all the elements of A also belong to B. Thus, we can write

$$A \subseteq B$$

When $A \subseteq B$ but A is not the same as B, i.e., there are elements of B that do not belong to A, then we write

$$A \subset B$$

which reads "A is strictly included in B", or

$$B \supset A$$

We also say that A is a proper subset of B.

Example 7 *Given the sets*

$$\begin{aligned} A &= \{2,3\} \\ B &= \{1,2,3,4\} \\ C &= \{2,3\} \end{aligned}$$

we have that

$$\begin{aligned} A &\subset B \\ A &\subseteq C \end{aligned}$$

but we cannot write

$$A \subset C$$

1.4 Union

The union of two sets A and B is the set of all elements that belong to at least one of them, and it is denoted by

$$A \cup B$$

Example 8 *Define two sets A and B as follows:*

$$\begin{aligned} A &= \{a,b,c,d\} \\ B &= \{c,d,e,f\} \end{aligned}$$

Their union is

$$A \cup B = \{a,b,c,d,e,f\}$$

If A_1, A_2, ..., A_n are n sets, their union is the set of all elements that belong to at least one of them, and it is denoted by

$$\bigcup_{i=1}^{n} A_i = A_1 \cup A_2 \cup \ldots \cup A_n$$

Example 9 *Define three sets A_1, A_2 and A_3 as follows:*

$$\begin{aligned} A_1 &= \{a,b,c,d\} \\ A_2 &= \{c,d,e,f\} \\ A_3 &= \{c,f,g\} \end{aligned}$$

Their union is

$$\bigcup_{i=1}^{3} A_i = A_1 \cup A_2 \cup A_3 = \{a,b,c,d,e,f,g\}$$

1.5 Intersection

The intersection of two sets A and B is the set of all elements that belong to both of them, and it is denoted by

$$A \cap B$$

Example 10 *Define two sets A and B as follows:*

$$\begin{aligned} A &= \{a, b, c, d\} \\ B &= \{c, d, e, f\} \end{aligned}$$

Their intersection is

$$A \cap B = \{c, d\}$$

If A_1, A_2, ..., A_n are n sets, their intersection is the set of all elements that belong to all of them, and it is denoted by

$$\bigcap_{i=1}^{n} A_i = A_1 \cap A_2 \cap \ldots \cap A_n$$

Example 11 *Define three sets A_1, A_2 and A_3 as follows:*

$$\begin{aligned} A_1 &= \{a, b, c, d\} \\ A_2 &= \{c, d, e, f\} \\ A_3 &= \{c, f, g\} \end{aligned}$$

Their intersection is

$$\bigcap_{i=1}^{3} A_i = A_1 \cap A_2 \cap A_3 = \{c\}$$

1.6 Complement

Suppose that our attention is confined to sets that are all included in a larger set Ω, called universal set. Let A be one of these sets. The complement of A is the set of all elements of Ω that do not belong to A and it is indicated by

$$A^c$$

Example 12 *Define the universal set Ω as follows:*

$$\Omega = \{a, b, c, d, e, f, g, h\}$$

and the two sets

$$\begin{aligned} A &= \{b, c, d\} \\ B &= \{c, d, e\} \end{aligned}$$

The complements of A and B are

$$\begin{aligned} A^c &= \{a, e, f, g, h\} \\ B^c &= \{a, b, f, g, h\} \end{aligned}$$

1.7 De Morgan's Laws

De Morgan's Laws are

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

and can be extended to collections of more than two sets as follows:

$$\begin{aligned}\left(\bigcup_{i=1}^{n} A_i\right)^c &= \bigcap_{i=1}^{n} A_i^c \\ \left(\bigcap_{i=1}^{n} A_i\right)^c &= \bigcup_{i=1}^{n} A_i^c\end{aligned}$$

1.8 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Define the following sets

$$\begin{aligned}A_1 &= \{a, b, c\} \\ A_2 &= \{b, c, d, e, f\} \\ A_3 &= \{b, f\} \\ A_4 &= \{a, b, d\}\end{aligned}$$

List all the elements belonging to the set

$$A = \bigcup_{i=2}^{4} A_i$$

Solution

The union can be written as

$$A = A_2 \cup A_3 \cup A_4$$

The union of the three sets A_2, A_3 and A_4 is the set of all elements that belong to at least one of them:

$$\begin{aligned}A &= A_2 \cup A_3 \cup A_4 \\ &= \{a, b, c, d, e, f\}\end{aligned}$$

Exercise 2

Given the sets defined in the previous exercise, list all the elements belonging to the set

$$A = \bigcap_{i=1}^{4} A_i$$

Solution

The intersection can be written as

$$A = A_1 \cap A_2 \cap A_3 \cap A_4$$

The intersection of the four sets A_1, A_2, A_3 and A_4 is the set of elements that are members of all the four sets:

$$\begin{aligned} A &= A_1 \cap A_2 \cap A_3 \cap A_4 \\ &= \{b\} \end{aligned}$$

Exercise 3

Suppose that A and B are two subsets of a universal set Ω, and that

$$\begin{aligned} A^c &= \{a, b, c\} \\ B^c &= \{b, c, d\} \end{aligned}$$

List all the elements belonging to the set

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

Solution

Using De Morgan's laws, we obtain

$$\begin{aligned} (A \cup B)^c &= A^c \cap B^c \\ &= \{a, b, c\} \cap \{b, c, d\} \\ &= \{b, c\} \end{aligned}$$

Chapter 2

Permutations

This lecture introduces permutations, one of the most important concepts in combinatorial analysis.

We first deal with permutations without repetition, also called simple permutations, and then with permutations with repetition.

2.1 Permutations without repetition

A permutation without repetition of n objects is one of the possible ways of ordering the n objects.

A permutation without repetition is also simply called a permutation.

The following subsections give a slightly more formal definition of permutation and deal with the problem of counting the number of possible permutations of n objects.

2.1.1 Definition of permutation without repetition

Let a_1, a_2,..., a_n be n objects. Let s_1, s_2, ..., s_n be n slots to which the n objects can be assigned. A **permutation** (or **permutation without repetition**, or **simple permutation**) of a_1, a_2,..., a_n is one of the possible ways to fill each of the n slots with one and only one of the n objects, with the proviso that each object can be assigned to only one slot.

Example 13 *Consider three objects a_1, a_2 and a_3. There are three slots, s_1, s_2 and s_3, to which we can assign the three objects. There are six possible permutations of the three objects, that is, six possible ways to fill the three slots with the three objects:*

Slots	s_1	s_2	s_3
Permutation 1	a_1	a_2	a_3
Permutation 2	a_1	a_3	a_2
Permutation 3	a_2	a_1	a_3
Permutation 4	a_2	a_3	a_1
Permutation 5	a_3	a_2	a_1
Permutation 6	a_3	a_1	a_2

2.1.2 Number of permutations without repetition

Denote by P_n the number of possible permutations of n objects. How much is P_n in general? In other words, how do we count the number of possible permutations of n objects?

We can derive a general formula for P_n by using a sequential argument:

1. First, we assign an object to the first slot. There are n objects that can be assigned to the first slot, so there are

   $$n \text{ possible ways to fill the first slot}$$

2. Then, we assign an object to the second slot. There were n objects, but one has already been assigned to a slot. So, we are left with $n-1$ objects that can be assigned to the second slot. Thus, there are

   $$n-1 \text{ possible ways to fill the second slot}$$

   and

   $$n \cdot (n-1) \text{ possible ways to fill the first two slots}$$

3. Then, we assign an object to the third slot. There were n objects, but two have already been assigned to a slot. So, we are left with $n-2$ objects that can be assigned to the third slot. Thus, there are

   $$n-2 \text{ possible ways to fill the third slot}$$

   and

   $$n \cdot (n-1) \cdot (n-2) \text{ possible ways to fill the first three slots}$$

4. An so on, until only one object and one free slot remain.

5. Finally, when only one free slot remains, we assign the remaining object to it. There is only one way to do this. Thus, there is

   $$1 \text{ possible way to fill the last slot}$$

   and

   $$n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 2 \cdot 1 \text{ possible ways to fill all the } n \text{ slots}$$

Therefore, by the above sequential argument, the total **number of possible permutations** of n objects is

$$P_n = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 2 \cdot 1$$

The number P_n is usually indicated as follows:

$$P_n = n!$$

where $n!$ is read "n **factorial**", with the convention that

$$0! = 1$$

Example 14 *The number of possible permutations of* 5 *objects is*

$$P_5 = 5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$$

2.2 Permutations with repetition

A permutation with repetition of n objects is one of the possible ways of selecting another set of n objects from the original one. The selection rules are:

1. each object can be selected more than once;
2. the order of selection matters (the same n objects selected in different orders are regarded as different permutations).

Thus, the difference between simple permutations and permutations with repetition is that objects can be selected only once in the former, while they can be selected more than once in the latter.

The following subsections give a slightly more formal definition of permutation with repetition and deal with the problem of counting the number of possible permutations with repetition.

2.2.1 Definition of permutation with repetition

Let a_1, a_2,..., a_n be n objects. Let s_1, s_2, ..., s_n be n slots to which the n objects can be assigned. A **permutation with repetition** of a_1, a_2,..., a_n is one of the possible ways to fill each of the n slots with one and only one of the n objects, with the proviso that an object can be assigned to more than one slot.

Example 15 *Consider two objects, a_1 and a_2. There are two slots to fill, s_1 and s_2. There are four possible permutations with repetition of the two objects, that is, four possible ways to assign an object to each slot, being allowed to assign the same object to more than one slot:*

Slots	s_1	s_2
Permutation 1	a_1	a_1
Permutation 2	a_1	a_2
Permutation 3	a_2	a_1
Permutation 4	a_2	a_2

2.2.2 Number of permutations with repetition

Denote by P'_n the number of possible permutations with repetition of n objects. How much is P'_n in general? In other words, how do we count the number of possible permutations with repetition of n objects?

We can derive a general formula for P'_n by using a sequential argument:

1. First, we assign an object to the first slot. There are n objects that can be assigned to the first slot, so there are

   $$n \text{ possible ways to fill the first slot}$$

2. Then, we assign an object to the second slot. Even if one object has been assigned to a slot in the previous step, we can still choose among n objects, because we are allowed to choose an object more than once. So, there are n objects that can be assigned to the second slot and

   $$n \text{ possible ways to fill the second slot}$$

and

$$n \cdot n \text{ possible ways to fill the first two slots}$$

3. Then, we assign an object to the third slot. Even if two objects have been assigned to a slot in the previous two steps, we can still choose among n objects, because we are allowed to choose an object more than once. So, there are n objects that can be assigned to the third slot and

$$n \text{ possible ways to fill the third slot}$$

and

$$n \cdot n \cdot n \text{ possible ways to fill the first three slots}$$

4. An so on, until we are left with only one free slot (the n-th).

5. When only one free slot remains, we assign one of the n objects to it. Thus, there are

$$n \text{ possible ways to fill the last slot}$$

and

$$\underbrace{n \cdot n \cdot \ldots \cdot n}_{n \text{ times}} \text{ possible ways to fill the } n \text{ available slots}$$

Therefore, by the above sequential argument, the total **number of possible permutations with repetition** of n objects is

$$P'_n = n^n$$

Example 16 *The number of possible permutations with repetition of* 3 *objects is*

$$P'_3 = 3^3 = 27$$

2.3 Solved exercises

This exercise set contains some solved exercises on permutations.

Exercise 1

There are 5 seats around a table and 5 people to be seated at the table. In how many different ways can they seat themselves?

Solution

Sitting 5 people at the table is a sequential problem. We need to assign a person to the first chair. There are 5 possible ways to do this. Then we need to assign a person to the second chair. There are 4 possible ways to do this, because one person has already been assigned. An so on, until there remain one free chair and one person to be seated. Therefore, the number of ways to seat the 5 people at the table is equal to the number of permutations of 5 objects (without repetition). If we denote it by P_5, then

$$P_5 = 5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$$

Exercise 2

Bob, John, Luke and Tim play a tennis tournament. The rules of the tournament are such that at the end of the tournament a ranking will be made and there will be no ties. How many different rankings can there be?

Solution

Ranking 4 people is a sequential problem. We need to assign a person to the first place. There are 4 possible ways to do this. Then we need to assign a person to the second place. There are 3 possible ways to do this, because one person has already been assigned. An so on, until there remains one person to be assigned. Therefore, the number of ways to rank the 4 people participating in the tournament is equal to the number of permutations of 4 objects (without repetition). If we denote it by P_4, then

$$P_4 = 4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$$

Exercise 3

A byte is a number consisting of 8 digits that can be equal either to 0 or to 1. How many different bytes are there?

Solution

To answer this question we need to follow a line of reasoning similar to the one we followed when we derived the number of permutations with repetition. There are 2 possible ways to choose the first digit and 2 possible ways to choose the second digit. So, there are 4 possible ways to choose the first two digits. There are 2 possible ways two choose the third digit and 4 possible ways to choose the first two. Thus, there are 8 possible ways to choose the first three digits. An so on, until we have chosen all digits. Therefore, the number of ways to choose the 8 digits is equal to

$$\underbrace{2 \cdot \ldots \cdot 2}_{8 \text{ times}} = 2^8 = 256$$

Chapter 3

k-permutations

This lecture introduces the concept of k-permutation, which is a slight generalization of the concept of permutation[1].

We first deal with k-permutations without repetition and then with k-permutations with repetition.

3.1 k-permutations without repetition

A k-permutation without repetition of n objects is a way of selecting k objects from a list of n. The selection rules are:

1. the order of selection matters (the same k objects selected in different orders are regarded as different k-permutations);

2. each object can be selected only once.

A k-permutation without repetition is also simply called a k-permutation.

The following subsections give a slightly more formal definition of k-permutation and deal with the problem of counting the number of possible k-permutations.

3.1.1 Definition of k-permutation without repetition

Let a_1, a_2,..., a_n be n objects. Let s_1, s_2, ..., s_k be k ($k \leq n$) slots to which k of the n objects can be assigned. A **k-permutation** (or **k-permutation without repetition** or **simple k-permutation**) of n objects from a_1, a_2,..., a_n is one of the possible ways to choose k of the n objects and fill each of the k slots with one and only one object. Each object can be chosen only once.

Example 17 *Consider three objects, a_1, a_2 and a_3. There are two slots, s_1 and s_2, to which we can assign two of the three objects. There are six possible* 2*-permutations of the three objects, that is, six possible ways to choose two objects*

[1] See p. 9.

and fill the two slots with the two objects:

Slots	s_1	s_2
2-*permutation 1*	a_1	a_2
2-*permutation 2*	a_1	a_3
2-*permutation 3*	a_2	a_1
2-*permutation 4*	a_2	a_3
2-*permutation 5*	a_3	a_1
2-*permutation 6*	a_3	a_2

3.1.2 Number of k-permutations without repetition

Denote by $P_{n,k}$ the number of possible k-permutations of n objects. How much is $P_{n,k}$ in general? In other words, how do we count the number of possible k-permutations of n objects?

We can derive a general formula for $P_{n,k}$ by using a sequential argument:

1. First, we assign an object to the first slot. There are n objects that can be assigned to the first slot, so there are

   $$n \text{ possible ways to fill the first slot}$$

2. Then, we assign an object to the second slot. There were n objects, but one has already been assigned to a slot. So, we are left with $n-1$ objects that can be assigned to the second slot. Thus, there are

   $$n-1 \text{ possible ways to fill the second slot}$$

   and

   $$n \cdot (n-1) \text{ possible ways to fill the first two slots}$$

3. Then, we assign an object to the third slot. There were n objects, but two have already been assigned to a slot. So, we are left with $n-2$ objects that can be assigned to the third slot. Thus, there are

   $$n-2 \text{ possible ways to fill the third slot}$$

   and

   $$n \cdot (n-1) \cdot (n-2) \text{ possible ways to fill the first three slots}$$

4. An so on, until we are left with $n-k+1$ objects and only one free slot (the k-th).

5. Finally, when only one free slot remains, we assign one of the remaining $n-k+1$ objects to it. Thus, there are

   $$n-k+1 \text{ possible ways to fill the last slot}$$

   and

   $$n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-k+1) \text{ possible ways to fill the } k \text{ available slots}$$

Therefore, by the above sequential argument, the total **number of possible k-permutations** of n objects is

$$P_{n,k} = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-k+1)$$

$P_{n,k}$ can be written as

$$P_{n,k} = \frac{n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-k+1) \cdot (n-k) \cdot (n-k-1) \cdot \ldots \cdot 2 \cdot 1}{(n-k) \cdot (n-k-1) \cdot \ldots \cdot 2 \cdot 1}$$

Remembering the definition of factorial[2], we can see that the numerator of the above ratio is $n!$ while the denominator is $(n-k)!$, so the number of possible k-permutations of n objects is

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

The number $P_{n,k}$ is usually indicated as follows:

$$P_{n,k} = n^{\underline{k}}$$

Example 18 *The number of possible* 3*-permutations of* 5 *objects is*

$$P_{5,3} = \frac{5!}{2!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1} = 5 \cdot 4 \cdot 3 = 60$$

3.2 k-permutations with repetition

A k-permutation with repetition of n objects is a way of selecting k objects from a list of n. The selection rules are:

1. the order of selection matters (the same k objects selected in different orders are regarded as different k-permutations);

2. each object can be selected more than once.

Thus, the difference between k-permutations without repetition and k-permutations with repetition is that objects can be selected more than once in the latter, while they can be selected only once in the former.

The following subsections give a slightly more formal definition of k-permutation with repetition and deal with the problem of counting the number of possible k-permutations with repetition.

3.2.1 Definition of k-permutation with repetition

Let a_1, a_2,..., a_n be n objects. Let s_1, s_2, ..., s_k be k ($k \leq n$) slots to which k of the n objects can be assigned. A **k-permutation with repetition** of n objects from a_1, a_2,..., a_n is one of the possible ways to choose k of the n objects and fill each of the k slots with one and only one object. Each object can be chosen more than once.

[2] See p. 10.

Example 19 *Consider three objects a_1, a_2 and a_3 and two slots, s_1 and s_2. There are nine possible 2-permutations with repetition of the three objects, that is, nine possible ways to choose two objects and fill the two slots with the two objects, being allowed to pick the same object more than once:*

Slots	s_1	s_2
2-*permutation 1*	a_1	a_1
2-*permutation 2*	a_1	a_2
2-*permutation 3*	a_1	a_3
2-*permutation 4*	a_2	a_1
2-*permutation 5*	a_2	a_2
2-*permutation 6*	a_2	a_3
2-*permutation 7*	a_3	a_1
2-*permutation 8*	a_3	a_2
2-*permutation 9*	a_3	a_3

3.2.2 Number of k-permutations with repetition

Denote by $P'_{n,k}$ the number of possible k-permutations with repetition of n objects. How much is $P'_{n,k}$ in general? In other words, how do we count the number of possible k-permutations with repetition of n objects?

We can derive a general formula for $P'_{n,k}$ by using a sequential argument:

1. First, we assign an object to the first slot. There are n objects that can be assigned to the first slot, so there are

   $$n \text{ possible ways to fill the first slot}$$

2. Then, we assign an object to the second slot. Even if one object has been assigned to a slot in the previous step, we can still choose among n objects, because we are allowed to choose an object more than once. So, there are n objects that can be assigned to the second slot and

   $$n \text{ possible ways to fill the second slot}$$

   and

   $$n \cdot n \text{ possible ways to fill the first two slots}$$

3. Then, we assign an object to the third slot. Even if two objects have been assigned to a slot in the previous two steps, we can still choose among n objects, because we are allowed to choose an object more than once. So, there are n objects that can be assigned to the second slot and

   $$n \text{ possible ways to fill the third slot}$$

   and

   $$n \cdot n \cdot n \text{ possible ways to fill the first three slots}$$

4. An so on, until we are left with only one free slot (the k-th).

5. When only one free slot remains, we assign one of the n objects to it. Thus, there are

   $$n \text{ possible ways to fill the last slot}$$

and

$$\underbrace{n \cdot n \cdot \ldots \cdot n}_{k \text{ times}} \text{ possible ways to fill the } k \text{ available slots}$$

Therefore, by the above sequential argument, the total **number of possible k-permutations with repetition** of n objects is

$$P'_{n,k} = n^k$$

Example 20 *The number of possible* 2*-permutations of* 4 *objects is*

$$P'_{4,2} = 4^2 = 16$$

3.3 Solved exercises

This exercise set contains some solved exercises on k-permutations.

Exercise 1

There is a basket of fruit containing an apple, a banana and an orange and there are five girls who want to eat one fruit. How many ways are there to give three of the five girls one fruit each and leave two of them without a fruit to eat?

Solution

Giving the three fruits to three of the five girls is a sequential problem. We first give the apple to one of the girls. There are 5 possible ways to do this. Then we give the banana to one of the remaining girls. There are 4 possible ways to do this, because one girl has already been given a fruit. Finally, we give the orange to one of the remaining girls. There are 3 possible ways to do this, because two girls have already been given a fruit. Summing up, the number of ways to assign the three fruits is equal to the number of 3-permutations of 5 objects (without repetition). If we denote it by $P_{5,3}$, then

$$\begin{aligned} P_{5,3} &= \frac{5!}{(5-3)!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1} \\ &= 5 \cdot 4 \cdot 3 = 60 \end{aligned}$$

Exercise 2

An hexadecimal number is a number whose digits can take sixteen different values: either one of the ten numbers from 0 to 9, or one of the six letters from A to F. How many different 8-digit hexadecimal numbers are there, if an hexadecimal number is allowed to begin with any number of zeros?

Solution

Choosing the 8 digits of the hexadecimal number is a sequential problem. There are 16 possible ways to choose the first digit and 16 possible ways to choose the second digit. So, there are 16×16 possible ways to choose the first two digits. There are 16 possible ways two choose the third digit and 16×16 possible ways to

choose the first two. Thus, there are $16 \times 16 \times 16$ possible ways to choose the first three digits. An so on, until we have chosen all digits. Therefore, the number of ways to choose the 8 digits is equal to the number of 8-permutations with repetition of 16 objects:

$$P'_{16,8} = 16^8$$

Exercise 3

An urn contains ten balls, each representing one of the ten numbers from 0 to 9. Three balls are drawn at random from the urn and the corresponding numbers are written down to form a 3-digit number, writing down the digits from left to right in the order in which they have been extracted. When a ball is drawn from the urn it is set aside, so that it cannot be extracted again. If one were to write down all the 3-digit numbers that could possibly be formed, how many would they be?

Solution

The 3 balls are drawn sequentially. At the first draw there are 10 balls, hence 10 possible values for the first digit of our 3-digit number. At the second draw there are 9 balls left, hence 9 possible values for the second digit of our 3-digit number. At the third and last draw there are 8 balls left, hence 8 possible values for the third digit of our 3-digit number. Summing up, the number of possible 3-digit numbers is equal to the number of 3-permutations of 10 objects (without repetition). If we denote it by $P_{10,3}$, then

$$\begin{aligned} P_{10,3} &= \frac{10!}{(10-3)!} = \frac{10 \cdot 9 \cdot \ldots \cdot 2 \cdot 1}{7 \cdot 6 \cdot \ldots \cdot 2 \cdot 1} \\ &= 10 \cdot 9 \cdot 8 = 720 \end{aligned}$$

Chapter 4

Combinations

This lecture introduces combinations, one of the most important concepts in combinatorial analysis. Before reading this lecture, you should be familiar with the concept of permutation[1].

We first deal with combinations without repetition and then with combinations with repetition.

4.1 Combinations without repetition

A combination without repetition of k objects from n is a way of selecting k objects from a list of n. The selection rules are:

1. the order of selection does not matter (the same objects selected in different orders are regarded as the same combination);
2. each object can be selected only once.

A combination without repetition is also called a simple combination or, simply, a combination.

The following subsections give a slightly more formal definition of combination and deal with the problem of counting the number of possible combinations.

4.1.1 Definition of combination without repetition

Let a_1, a_2,..., a_n be n objects. A **simple combination** (or **combination without repetition**) of k objects from the n objects is one of the possible ways to form a set containing k of the n objects. To form a valid set, any object can be chosen only once. Furthermore, the order in which the objects are chosen does not matter.

Example 21 *Consider three objects, a_1, a_2 and a_3. There are three possible combinations of two objects from a_1, a_2 and a_3, that is, three possible ways to choose two objects from this set of three:*

Combination 1	a_1 *and* a_2
Combination 2	a_1 *and* a_3
Combination 3	a_2 *and* a_3

[1] See the lecture entitled *Permutations* (p. 9).

Other combinations are not possible, because, for example, $\{a_2, a_1\}$ *is the same as* $\{a_1, a_2\}$.

4.1.2 Number of combinations without repetition

Denote by $C_{n,k}$ the number of possible combinations of k objects from n. How much is $C_{n,k}$ in general? In other words, how do we count the number of possible combinations of k objects from n?

To answer this question, we need to recall the concepts of permutation and k-permutation introduced in previous lectures[2].

Like a combination, a k-permutation of n objects is one of the possible ways of choosing k of the n objects. However, in a k-permutation the order of selection matters: two k-permutations are regarded as different if the same k objects are chosen, but they are chosen in a different order. On the contrary, in the case of combinations, the order in which the k objects are chosen does not matter: two combinations that contain the same objects are regarded as equal.

Despite this difference between k-permutations and combinations, it is very easy to derive the number of possible combinations ($C_{n,k}$) from the number of possible k-permutations ($P_{n,k}$). Consider a combination of k objects from n. This combination will be repeated many times in the set of all possible k-permutations. It will be repeated one time for each possible way of ordering the k objects. So, it will be repeated $P_k = k!$ times[3]. Therefore, if each combination is repeated P_k times in the set of all possible k-permutations, dividing the total number of k-permutations ($P_{n,k}$) by P_k, we obtain the **number of possible combinations**:

$$C_{n,k} = \frac{P_{n,k}}{P_k} = \frac{n!}{(n-k)!k!}$$

The number of possible combinations is often denoted by

$$C_{n,k} = \binom{n}{k}$$

and $\binom{n}{k}$ is called **binomial coefficient**.

Example 22 *The number of possible combinations of* 3 *objects from* 5 *is*

$$C_{5,3} = \frac{5!}{2!3!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{(2 \cdot 1)(3 \cdot 2 \cdot 1)} = \frac{5 \cdot 4}{2 \cdot 1} = 10$$

4.2 Combinations with repetition

A combination with repetition of k objects from n is a way of selecting k objects from a list of n. The selection rules are:

1. the order of selection does not matter (the same objects selected in different orders are regarded as the same combination);

2. each object can be selected more than once.

[2] See the lectures entitled *Permutations* (p. 9) and *k-permutations* (p. 15).

[3] P_k is the number of all possible ways to order the k objects - the number of permutations of k objects.

Thus, the difference between simple combinations and combinations with repetition is that objects can be selected only once in the former, while they can be selected more than once in the latter.

The following subsections give a slightly more formal definition of combination with repetition and deal with the problem of counting the number of possible combinations with repetition.

4.2.1 Definition of combination with repetition

A more rigorous definition of combination with repetition involves the concept of multiset, which is a generalization of the notion of set[4]. Roughly speaking, the difference between a multiset and a set is the following: the same object is allowed to appear more than once in the list of members of a **multiset**, while the same object is allowed to appear only once in the list of members of an ordinary **set**. Thus, for example, the collection of objects

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

is a valid multiset, but not a valid set, because the letter a appears more than once. Like sets, multisets are unordered collections of objects, i.e. the order in which the elements of a multiset are listed does not matter.

Let a_1, a_2,..., a_n be n objects. A **combination with repetition** of k objects from the n objects is one of the possible ways to form a multiset containing k objects taken from the set $\{a_1, a_2, \ldots, a_n\}$.

Example 23 *Consider three objects, a_1, a_2 and a_3. There are six possible combinations with repetition of two objects from a_1, a_2 and a_3, that is, six possible ways to choose two objects from this set of three, allowing for repetitions:*

Combination 1	a_1 *and* a_2
Combination 2	a_1 *and* a_3
Combination 3	a_2 *and* a_3
Combination 4	a_1 *and* a_1
Combination 5	a_2 *and* a_2
Combination 6	a_3 *and* a_3

Other combinations are not possible, because, for example, $\{a_2, a_1\}$ is the same as $\{a_1, a_2\}$.

4.2.2 Number of combinations with repetition

Denote by $C'_{n,k}$ the number of possible combinations with repetition of k objects from n. How much is $C'_{n,k}$ in general? In other words, how do we count the number of possible combinations with repetition of k objects from n?

To answer this question, we need to use a slightly unusual procedure, which is introduced by the next example.

Example 24 *We need to order two scoops of ice cream, choosing among four flavours: chocolate, pistachio, strawberry and vanilla. It is possible to order two scoops of the same flavour. How many different combinations can we order? The*

[4] See the lecture entitled *Set theory* (p. 3).

number of different combinations we can order is equal to the number of possible combinations with repetition of 2 *objects from* 4*. Let us represent an order as a string of crosses (*$\times$*) and vertical bars (*$|$*), where a vertical bar delimits two adjacent flavours and a cross denotes a scoop of a given flavour. For example,*

$\times \mid \mid \mid \times$	1 *chocolate,* 1 *vanilla*
$\mid \mid \times \mid \times$	1 *strawberry,* 1 *vanilla*
$\times \times \mid \mid \mid$	2 *chocolate*
$\mid \mid \times \times \mid$	2 *strawberry*

where the first vertical bar (the leftmost one) delimits chocolate and pistachio, the second one delimits pistachio and strawberry and the third one delimits strawberry and vanilla. Each string contains three vertical bars, one less than the number of flavours, and two crosses, one for each scoop. Therefore, each string contains a total of five symbols. Making an order is equivalent to choosing which two of the five symbols will be a cross (the remaining will be vertical bars). So, to make an order, we need to choose 2 *objects from* 5*. The number of possible ways to choose* 2 *objects from* 5 *is equal to the number of possible combinations without repetition*[5] *of* 2 *objects from* 5*. Therefore, there are*

$$\binom{5}{2} = \frac{5!}{(5-2)!2!} = 10$$

different orders we can make.

In general, choosing k objects from n with repetition is equivalent to writing a string with $n+k-1$ symbols, of which $n-1$ are vertical bars ($|$) and k are crosses ($\times$). In turn, this is equivalent to choose the k positions in the string (among the available $n+k-1$) that will contain a cross (the remaining ones will contain vertical bars). But choosing k positions from $n+k-1$ is like choosing a combination without repetition of k objects from $n+k-1$. Therefore, the **number of possible combinations with repetition** is

$$\begin{aligned} C'_{n,k} &= C_{n+k-1,k} = \binom{n+k-1}{k} \\ &= \frac{(n+k-1)!}{(n+k-1-k)!k!} = \frac{(n+k-1)!}{(n-1)!k!} \end{aligned}$$

The number of possible combinations with repetition is often denoted by

$$C'_{n,k} = \left(\binom{n}{k}\right)$$

and $\left(\binom{n}{k}\right)$ is called a **multiset coefficient**.

Example 25 *The number of possible combinations with repetition of* 3 *objects from* 5 *is*

$$\begin{aligned} C'_{5,3} &= \frac{(5+3-1)!}{(5-1)!3!} = \frac{7!}{4!3!} \\ &= \frac{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{(4 \cdot 3 \cdot 2 \cdot 1)(3 \cdot 2 \cdot 1)} \\ &= \frac{7 \cdot 6 \cdot 5}{3 \cdot 2 \cdot 1} = 7 \cdot 5 = 35 \end{aligned}$$

[5] See p. 21.

4.3 More details

4.3.1 Binomial coefficients and binomial expansions

The binomial coefficient is so called because it appears in the **binomial expansion**:

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

where $n \in \mathbb{N}$.

4.3.2 Recursive formula for binomial coefficients

The following is a useful recursive formula for computing binomial coefficients:

$$\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1}$$

Proof. It is proved as follows:

$$\begin{aligned}
& \binom{n}{k} + \binom{n}{k-1} = \frac{n!}{k!\,(n-k)!} + \frac{n!}{(k-1)!\,(n+1-k)!} \\
= \; & \frac{n!}{(k-1)!\,(n-k)!k} + \frac{n!}{(k-1)!\,(n-k)!\,(n+1-k)} \\
= \; & \frac{n!\,(n+1-k+k)}{(k-1)!\,(n-k)!k\,(n+1-k)} \\
= \; & \frac{n!\,(n+1)}{k!\,(n+1-k)!} = \frac{(n+1)!}{k!\,(n+1-k)!} = \binom{n+1}{k}
\end{aligned}$$

■

4.4 Solved exercises

This exercise set contains some solved exercises on combinations.

Exercise 1

3 cards are drawn from a standard deck of 52 cards. How many different 3-card hands can possibly be drawn?

Solution

First of all, the order in which the 3 cards are drawn does not matter, that is, the same cards drawn in different orders are regarded as the same 3-card hand. Furthermore, each card can be drawn only once. Therefore the number of different 3-card hands that can possibly be drawn is equal to the number of possible combinations without repetition of 3 objects from 52. If we denote it by $C_{52,3}$, then

$$\begin{aligned}
C_{52,3} &= \binom{52}{3} = \frac{52!}{(52-3)!3!} = \frac{52!}{49!3!} \\
&= \frac{52 \cdot 51 \cdot 50}{3!} = \frac{52 \cdot 51 \cdot 50}{3 \cdot 2 \cdot 1} = 22100
\end{aligned}$$

Exercise 2

John has got one dollar, with which he can buy green, red and yellow candies. Each candy costs 50 cents. John will spend all the money he has on candies. How many different combinations of green, red and yellow candies can he buy?

Solution

First of all, the order in which the 3 different colors are chosen does not matter. Furthermore, each color can be chosen more than once. Therefore, the number of different combinations of colored candies John can choose is equal to the number of possible combinations with repetition of 2 objects from 3. If we denote it by $C'_{3,2}$, then

$$\begin{aligned} C'_{3,2} &= \left(\binom{3}{2}\right) = \binom{3+2-1}{2} = \binom{4}{2} \\ &= \frac{4!}{(4-2)!2!} = \frac{4!}{2!2!} = \frac{4 \cdot 3}{2!} = \frac{4 \cdot 3}{2 \cdot 1} = 6 \end{aligned}$$

Exercise 3

The board of directors of a corporation comprises 10 members. An executive board, formed by 4 directors, needs to be elected. How many possible ways are there to form the executive board?

Solution

First of all, the order in which the 4 directors are selected does not matter. Furthermore, each director can be elected to the executive board only once. Therefore, the number of different ways to form the executive board is equal to the number of possible combinations without repetition of 4 objects from 10. If we denote it by $C_{10,4}$, then

$$\begin{aligned} C_{10,4} &= \binom{10}{4} = \frac{10!}{(10-4)!4!} = \frac{10!}{6!4!} \\ &= \frac{10 \cdot 9 \cdot 8 \cdot 7}{4!} = \frac{10 \cdot 9 \cdot 8 \cdot 7}{4 \cdot 3 \cdot 2 \cdot 1} = 210 \end{aligned}$$

Chapter 5

Partitions into groups

This lecture introduces partitions into groups. Before reading this lecture, you should read the lectures entitled *Permutations* (p. 9) and *Combinations* (p. 21).

A partition of n objects into k groups is one of the possible ways of subdividing the n objects into k groups ($k \leq n$). The rules are:

1. the order in which objects are assigned to a group does not matter;
2. each object can be assigned to only one group.

The following subsections give a slightly more formal definition of partition into groups and deal with the problem of counting the number of possible partitions into groups.

5.1 Definition of partition into groups

Let a_1, a_2,..., a_n be n objects. Let g_1, g_2, ..., g_k be k (with $k \leq n$) groups to which we can assign the n objects. Moreover, n_1 objects need to be assigned to group g_1, n_2 objects need to be assigned to group g_2, and so on. The numbers n_1, n_2, ..., n_k are such that

$$n_1 + n_2 + \ldots + n_k = n$$

A **partition** of a_1, a_2,..., a_n **into** the k **groups** g_1, g_2, ..., g_k is one of the possible ways to assign the n objects to the k groups.

Example 26 *Consider three objects, a_1, a_2 and a_3, and two groups, g_1 and g_2, with*

$$\begin{aligned} n_1 &= 2 \\ n_2 &= 1 \end{aligned}$$

There are three possible partitions of the three objects into the two groups:

Groups	g_1	g_2
Partition 1	$\{a_1, a_2\}$	a_3
Partition 2	$\{a_1, a_3\}$	a_2
Partition 3	$\{a_2, a_3\}$	a_1

Note that the order of objects belonging to a group does not matter, so, for example, $\{a_1, a_2\}$ in Partition 1 is the same as $\{a_2, a_1\}$.

5.2 Number of partitions into groups

Denote by $P_{n_1,n_2,\ldots,n_k}$ the number of possible partitions into the k groups (where group i contains n_i objects). How much is $P_{n_1,n_2,\ldots,n_k}$ in general?

The number $P_{n_1,n_2,\ldots,n_k}$ can be derived using a sequential argument:

1. First, we assign n_1 objects to the first group. There is a total of n objects to choose from. The number of possible ways to choose n_1 of the n objects is equal to the number of combinations[1] of n_1 elements from n. So there are

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

possible ways to form the first group.

2. Then, we assign n_2 objects to the second group. There were n objects, but n_1 have already been assigned to the first group. So, there are $n-n_1$ objects left, that can be assigned to the second group. The number of possible ways to choose n_2 of the remaining $n-n_1$ objects is equal to the number of combinations of n_2 elements from $n-n_1$. So there are

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

possible ways to form the second group and

$$\begin{aligned}\binom{n}{n_1}\binom{n-n_1}{n_2} &= \frac{n!}{n_1!\,(n-n_1)!}\frac{(n-n_1)!}{n_2!\,(n-n_1-n_2)!} \\ &= \frac{n!}{n_1!n_2!\,(n-n_1-n_2)!}\end{aligned}$$

possible ways to form the first two groups.

3. Then, we assign n_3 objects to the third group. There were n objects, but n_1+n_2 have already been assigned to the first two groups. So, there are $n-n_1-n_2$ objects left, that can be assigned to the third group. The number of possible ways to choose n_3 of the remaining $n-n_1-n_2$ objects is equal to the number of combinations of n_3 elements from $n-n_1-n_2$. So there are

$$\binom{n-n_1-n_2}{n_3} = \frac{(n-n_1-n_2)!}{n_3!\,(n-n_1-n_2-n_3)!}$$

possible ways to form the third group and

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

possible ways to form the first three groups.

[1] See the lecture entitled *Combinations* (p. 21).

4. An so on, until we are left with n_k objects and the last group. There is only one way to form the last group, which can also be written as:

$$\binom{n-n_1-n_2-\ldots-n_{k-1}}{n_k} = \frac{(n-n_1-n_2-\ldots-n_{k-1})!}{n_k!\,(n-n_1-n_2-\ldots-n_k)!}$$

As a consequence, there are

$$\begin{aligned}
& \binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdot\ldots\cdot\binom{n-n_1-n_2-\ldots-n_{k-1}}{n_k} \\
= & \frac{n!}{n_1!n_2!\cdot\ldots\cdot n_{k-1}!\,(n-n_1-n_2-\ldots-n_{k-1})!} \\
& \cdot\frac{(n-n_1-n_2-\ldots-n_{k-1})!}{n_k!\,(n-n_1-n_2-\ldots-n_k)!} \\
= & \frac{n!}{n_1!n_2!\cdot\ldots\cdot n_k!\,(n-n_1-n_2-\ldots-n_k)!} \\
= & \frac{n!}{n_1!n_2!\cdot\ldots\cdot n_k!0!} \\
= & \frac{n!}{n_1!n_2!\cdot\ldots\cdot n_k!}
\end{aligned}$$

possible ways to form all the groups.

Therefore, by the above sequential argument, the total **number of possible partitions** into the k groups is

$$P_{n_1,n_2,\ldots,n_k} = \frac{n!}{n_1!n_2!\cdot\ldots\cdot n_k!}$$

The number $P_{n_1,n_2,\ldots,n_k}$ is often indicated as follows:

$$P_{n_1,n_2,\ldots,n_k} = \binom{n}{n_1,n_2,\ldots,n_k}$$

and $\binom{n}{n_1,n_2,\ldots,n_k}$ is called a **multinomial coefficient**.

Sometimes the following notation is also used:

$$P_{n_1,n_2,\ldots,n_k} = (n_1,n_2,\ldots,n_k)!$$

Example 27 *The number of possible partitions of* 4 *objects into* 2 *groups of* 2 *objects is*

$$P_{2,2} = \binom{4}{2,2} = \frac{4!}{2!2!} = \frac{4\cdot3\cdot2\cdot1}{(2\cdot1)\,(2\cdot1)} = 6$$

5.3 More details

5.3.1 Multinomial expansions

The multinomial coefficient is so called because it appears in the **multinomial expansion**:

$$(x_1+x_2+\ldots+x_k)^n = \sum\binom{n}{n_1,n_2,\ldots,n_k}x_1^{n_1}x_2^{n_2}\cdot\ldots\cdot x_k^{n_k}$$

where $n\in\mathbb{N}$ and the summation is over all the k-tuples $n_1,n_2,\ldots,n_k$ such that

$$n_1+n_2+\ldots+n_k = n$$

5.4 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

John has a basket of fruit containing one apple, one banana, one orange and one kiwi. He wants to give one fruit to each of his two little sisters and two fruits to his big brother. In how many different ways can he do this?

Solution

John needs to decide how to partition 4 objects into 3 groups. The first two groups will contain one object and the third one will contain two objects. The total number of partitions is

$$\begin{aligned} P_{1,1,2} &= \binom{4}{1,1,2} = \frac{4!}{1!1!2!} \\ &= \frac{4 \cdot 3 \cdot 2 \cdot 1}{1 \cdot 1 \cdot 2 \cdot 1} = \frac{24}{2} = 12 \end{aligned}$$

Exercise 2

Ten friends want to play basketball. They need to divide into two teams of five players. In how many different ways can they do this?

Solution

They need to decide how to partition 10 objects into 2 groups. Each group will contain 5 objects. The total number of partitions is

$$\begin{aligned} P_{5,5} &= \binom{10}{5,5} = \frac{10!}{5!5!} \\ &= \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{(5 \cdot 4 \cdot 3 \cdot 2 \cdot 1)\,(5 \cdot 4 \cdot 3 \cdot 2 \cdot 1)} \\ &= \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = \frac{9 \cdot 8 \cdot 7 \cdot 6}{4 \cdot 3} \\ &= 9 \cdot 2 \cdot 7 \cdot 2 = 252 \end{aligned}$$

Chapter 6

Sequences and limits

This lecture discusses the concepts of sequence and limit of a sequence.

6.1 Definition of sequence

Let A be a set of objects, for example, real numbers. A **sequence** of elements of A is a function from the set of natural numbers $\mathbb{N}$ to the set A, that is, a correspondence that associates one and only one element of A to each natural number $n \in \mathbb{N}$. In other words, a sequence of elements of A is an ordered list of elements of A, where the ordering is provided by the natural numbers.

A sequence is usually indicated by enclosing a generic element of the sequence in curly brackets:

$$\{a_n\}$$

where a_n is the n-th element of the sequence. Alternative notations are

$$\{a_n\}_{n=1}^{\infty}$$
$$\{a_1, a_2, \ldots, a_n, \ldots\}$$
$$a_n \ , \ n \in \mathbb{N}$$
$$a_1, a_2, \ldots, a_n, \ldots$$

Thus, if $\{a_n\}$ is a sequence, a_1 is its first element, a_2 is its second element, a_n is its n-th element, and so on.

Example 28 *Define a sequence $\{a_n\}$ by characterizing its n-th element a_n as follows:*

$$a_n = \frac{1}{n}$$

$\{a_n\}$ is a sequence of rational numbers. The elements of the sequence are $a_1 = 1$, $a_2 = \frac{1}{2}$, $a_3 = \frac{1}{3}$, $a_4 = \frac{1}{4}$, and so on.

Example 29 *Define a sequence $\{a_n\}$ by characterizing its n-th element a_n as follows:*

$$a_n = \begin{cases} 1 \text{ if } n \text{ is even} \\ 0 \text{ if } n \text{ is odd} \end{cases}$$

$\{a_n\}$ is a sequence of 0 and 1. The elements of the sequence are $a_1 = 0$, $a_2 = 1$, $a_3 = 0$, $a_4 = 1$, and so on.

Example 30 *Define a sequence* $\{a_n\}$ *by characterizing its n-th element* a_n *as follows:*

$$a_n = \left[\frac{1}{n+1}, \frac{1}{n}\right]$$

$\{a_n\}$ *is a sequence of closed subintervals of the interval* $[0, 1]$*. The elements of the sequence are* $a_1 = \left[\frac{1}{2}, 1\right]$, $a_2 = \left[\frac{1}{3}, \frac{1}{2}\right]$, $a_3 = \left[\frac{1}{4}, \frac{1}{3}\right]$, $a_4 = \left[\frac{1}{5}, \frac{1}{4}\right]$, *and so on.*

6.2 Countable and uncountable sets

A set of objects A is a **countable set** if all its elements can be arranged into a sequence, that is, if there exists a sequence $\{a_n\}$ such that

$$\forall a \in A, \exists n \in \mathbb{N} : a_n = a$$

In other words, A is a countable set if there exists at least one sequence $\{a_n\}$ such that every element of A belongs to the sequence. A is an **uncountable set** if such a sequence does not exist. The most important example of an uncountable set is the set of real numbers $\mathbb{R}$.

6.3 Limit of a sequence

This section introduces the notion of limit of a sequence $\{a_n\}$. We start from the simple case in which $\{a_n\}$ is a sequence of real numbers, then we deal with the general case in which $\{a_n\}$ is a sequence of objects that are not necessarily real numbers.

6.3.1 The limit of a sequence of real numbers

We give first an informal and then a more formal definition of the limit of a sequence of real numbers.

Informal definition of limit - Sequences of real numbers

Let $\{a_n\}$ be a sequence of real numbers. Let $n_0 \in \mathbb{N}$. Denote by $\{a_n\}_{n>n_0}$ a subsequence of $\{a_n\}$ obtained by dropping the first n_0 terms of $\{a_n\}$, i.e.,

$$\{a_n\}_{n>n_0} = \{a_{n_0+1}, a_{n_0+2}, a_{n_0+3}, \ldots\}$$

The following is an intuitive definition of limit of a sequence.

Definition 31 (informal) *We say that a real number* a *is a* ***limit of a sequence*** $\{a_n\}$ *of real numbers, if, by appropriately choosing* n_0*, the distance between* a *and any term of the subsequence* $\{a_n\}_{n>n_0}$ *can be made as close to zero as we like. If* a *is a limit of the sequence* $\{a_n\}$*, we say that the sequence* $\{a_n\}$ *is a* ***convergent sequence*** *and that it* ***converges*** *to* a*. We indicate the fact that* a *is a limit of* $\{a_n\}$ *by*

$$a = \lim_{n\to\infty} a_n$$

Thus, a is a limit of $\{a_n\}$ if, by dropping a sufficiently high number of initial terms of $\{a_n\}$, we can make the remaining terms of $\{a_n\}$ as close to a as we like. Intuitively, a is a limit of $\{a_n\}$ if a_n becomes closer and closer to a by letting n go to infinity.

Formal definition of limit - Sequences of real numbers

The distance between two real numbers is the absolute value of their difference. For example, if $a \in \mathbb{R}$ and a_n is a term of a sequence $\{a_n\}$, the distance between a_n and a, denoted by $d(a_n, a)$, is

$$d(a_n, a) = |a_n - a|$$

Using the concept of distance, the above informal definition can be made rigorous.

Definition 32 (formal) *We say that $a \in \mathbb{R}$ is a **limit of a sequence** $\{a_n\}$ of real numbers if*

$$\forall \varepsilon > 0, \exists n_0 \in \mathbb{N} : d(a_n, a) < \varepsilon, \forall n > n_0$$

*If a is a limit of the sequence $\{a_n\}$, we say that the sequence $\{a_n\}$ is a **convergent sequence** and that it **converges** to a. We indicate the fact that a is a limit of $\{a_n\}$ by*

$$a = \lim_{n \to \infty} a_n$$

For those unfamiliar with the universal quantifiers $\forall$ (any) and $\exists$ (exists), the notation

$$\forall \varepsilon > 0, \exists n_0 \in \mathbb{N} : d(a_n, a) < \varepsilon, \forall n > n_0$$

reads as follows: "For any arbitrarily small number ε, there exists a natural number n_0 such that the distance between a_n and a is less than ε for all the terms a_n with $n > n_0$", which can also be restated as "For any arbitrarily small number ε, you can find a subsequence $\{a_n\}_{n>n_0}$ such that the distance between any term of the subsequence and a is less than ε", or as "By dropping a sufficiently high number of initial terms of $\{a_n\}$, you can make the remaining terms as close to a as you wish".

It is possible to prove that a convergent sequence has a unique limit, that is, if $\{a_n\}$ has a limit a, then a is the unique limit of $\{a_n\}$.

Example 33 *Define a sequence $\{a_n\}$ by characterizing its n-th element a_n as follows:*

$$a_n = \frac{1}{n}$$

The elements of the sequence are $a_1 = 1$, $a_2 = \frac{1}{2}$, $a_3 = \frac{1}{3}$, $a_4 = \frac{1}{4}$, and so on. The higher n is, the smaller a_n is and the closer it gets to 0. Therefore, intuitively, the limit of the sequence should be

$$\lim_{n \to \infty} a_n = 0$$

It is straightforward to prove that 0 is indeed a limit of $\{a_n\}$ by using Definition 32. Choose any $\varepsilon > 0$. We need to find an $n_0 \in \mathbb{N}$ such that all terms of the subsequence $\{a_n\}_{n>n_0}$ have distance from zero less than ε:

$$d(a_n, 0) < \varepsilon, \forall n > n_0 \tag{6.1}$$

The distance between a generic term of the sequence a_n and 0 is

$$d(a_n, 0) = |a_n - 0| = |a_n| = a_n$$

where the last equality holds because all the terms of the sequence are positive and hence equal to their absolute values. Therefore, we need to find an $n_0 \in \mathbb{N}$ *such that all the terms of the subsequence* $\{a_n\}_{n>n_0}$ *satisfy*

$$a_n < \varepsilon, \forall n > n_0 \tag{6.2}$$

Since

$$a_n < a_{n_0}, \forall n > n_0$$

condition (6.2) is satisfied if $a_{n_0} < \varepsilon$*, which is equivalent to* $\frac{1}{n_0} < \varepsilon$*. As a consequence, it suffices to pick any* n_0 *such that* $n_0 > \frac{1}{\varepsilon}$ *to satisfy condition (6.1). Summing up, we have just shown that, for any* ε*, we are able to find* $n_0 \in \mathbb{N}$ *such that all terms of the subsequence* $\{a_n\}_{n>n_0}$ *have distance from zero less than* ε*, which implies that* 0 *is the limit of the sequence* $\{a_n\}$*.*

6.3.2 The limit of a sequence in general

We now deal with the more general case in which the terms of the sequence $\{a_n\}$ are not necessarily real numbers. As before, we first give an informal definition, and then a more formal one.

Informal definition of limit - The general case

Let A be a set of objects and let $\{a_n\}$ be a sequence of elements of A. The limit of $\{a_n\}$ is defined as follows.

Definition 34 (informal) *Let* $a \in A$*. We say that* a *is a* ***limit of a sequence*** $\{a_n\}$ *of elements of* A*, if, by appropriately choosing* n_0*, the distance between* a *and any term of the subsequence* $\{a_n\}_{n>n_0}$ *can be made as close to zero as we like. If* a *is a limit of the sequence* $\{a_n\}$*, we say that the sequence* $\{a_n\}$ *is a* ***convergent sequence*** *and that it* ***converges*** *to* a*. We indicate the fact that* a *is a limit of* $\{a_n\}$ *by*

$$a = \lim_{n\to\infty} a_n$$

The definition is the same given in Definition 31, except for the fact that now both a and the terms of the sequence $\{a_n\}$ belong to a generic set of objects A.

Metrics and the definition of distance

In Definition 34 we have implicitly assumed that the concept of distance between elements of A is well-defined. Thus, for this definition to make any sense, we need to properly define distance.

We need a function $d : A \times A \to \mathbb{R}$ that associates to any couple of elements of A a real number measuring how far these two elements are. For example, if a and a' are two elements of A, $d(a, a')$ needs to be a real number measuring the distance between a and a'.

A function $d : A \times A \to \mathbb{R}$ is considered a valid distance function if it satisfies the properties listed in the following definition.

Definition 35 *Let* A *be a set of objects. Let* $d : A \times A \to \mathbb{R}$*.* d *is considered a valid distance function, in which case it is called a* ***metric*** *on* A*, if the following conditions are satisfied for any* a*,* a' *and* a'' *belonging to* A*:*

1. ***non-negativity:*** $d(a, a') \geq 0$;
2. ***identity of indiscernibles:*** $d(a, a') = 0$ *if and only if* $a = a'$;
3. ***symmetry:*** $d(a, a') = d(a', a)$;
4. ***triangle inequality:*** $d(a, a') + d(a', a'') \geq d(a, a'')$.

All four properties are very intuitive: property 1) says that the distance between two points cannot be a negative number; property 2) says that the distance between two points is zero if and only if the two points coincide; property 3) says that the distance from a to a' is the same as the distance from a' to a; property 4) says that the distance you cover when you go from a to a'' directly is less than or equal to the distance you cover when you go from a to a'' passing from a third point a' (in other words, if a' is not on the way from a to a'', you are increasing the distance covered).

Example 36 (Euclidean distance) *Consider the set of K-dimensional real vectors $\mathbb{R}^K$. The metric usually employed to measure the distance between elements of $\mathbb{R}^K$ is the so-called Euclidean distance. If a and b are two vectors belonging to $\mathbb{R}^K$, then their Euclidean distance is*

$$d(a, b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots + (a_K - b_K)^2}$$

where $a_1, \ldots, a_K$ are the K components of a and $b_1, \ldots, b_K$ are the K components of b. It is possible to prove that the Euclidean distance satisfies all the four properties that a metric needs to satisfy. Furthermore, when $K = 1$, it becomes

$$d(a, b) = \sqrt{(a - b)^2} = |a - b|$$

which coincides with the definition of distance between real numbers already given above.

Whenever we are faced with a sequence of objects and we want to assess whether it is convergent, we need to define a distance function on the set of objects to which the terms of the sequence belong, and verify that the proposed distance function satisfies all the properties of a proper distance function (a metric). For example, in probability theory and statistics we often deal with sequences of random variables. To assess whether these sequences are convergent, we need to define a metric to measure the distance between two random variables. As we will see in the lecture entitled *Sequences of random variables* (see p. 491), there are several ways of defining the concept of distance between two random variables. All these ways are legitimate and are useful in different situations.

Formal definition of limit - The general case

Having defined the concept of a metric, we are now ready to state the formal definition of a limit of a sequence.

Definition 37 (formal) *Let A be a set of objects. Let $d : A \times A \to \mathbb{R}$ be a metric on A. We say that $a \in A$ is a* ***limit of a sequence*** *$\{a_n\}$ of objects belonging to A if*

$$\forall \varepsilon > 0, \exists n_0 \in \mathbb{N} : d(a_n, a) < \varepsilon, \forall n > n_0$$

If a is a limit of the sequence $\{a_n\}$, we say that the sequence $\{a_n\}$ is a ***convergent sequence*** *and that it* ***converges*** *to a. We indicate the fact that a is a limit of $\{a_n\}$ by*

$$a = \lim_{n\to\infty} a_n$$

Also in this case, it is possible to prove that a convergent sequence has a unique limit.

Proposition 38 *If $\{a_n\}$ has a limit a, then a is the unique limit of $\{a_n\}$.*

Proof. The proof is by contradiction. Suppose that a and a' are two limits of a sequence $\{a_n\}$ and $a \neq a'$. By combining property 1) and 2) of a metric (see above), we obtain

$$d(a, a') > 0$$

i.e., $d(a, a') = \overline{d}$, where $\overline{d}$ is a strictly positive constant. Pick any term a_n of the sequence. By property 4) of a metric (the triangle inequality), we have

$$d(a, a_n) + d(a_n, a') \geq d(a, a')$$

Since $d(a, a') = \overline{d}$, the previous inequality becomes

$$d(a, a_n) + d(a', a_n) \geq \overline{d} > 0$$

Now, take any $\varepsilon < \overline{d}$. Since a is a limit of the sequence, we can find n_0 such that $d(a, a_n) < \varepsilon, \forall n > n_0$, which means that

$$\varepsilon + d(a', a_n) \geq d(a, a_n) + d(a', a_n) \geq \overline{d} > 0, \forall n > n_0$$

and

$$d(a', a_n) \geq \overline{d} - \varepsilon > 0, \forall n > n_0$$

Therefore, $d(a', a_n)$ can not be made smaller than $\overline{d} - \varepsilon$, and, as a consequence, a' cannot be a limit of the sequence. ■

Convergence criterion

In practice, it is usually difficult to assess the convergence of a sequence using Definition 37. Instead, convergence can be assessed using the following criterion.

Proposition 39 *Let A be a set of objects. Let $d : A \times A \to \mathbb{R}$ be a metric on A. Let $\{a_n\}$ be a sequence of objects belonging to A, and $a \in A$. The sequence $\{a_n\}$ converges to a if and only if*

$$\lim_{n\to\infty} d(a_n, a) = 0$$

Proof. This is easily proved by defining a sequence of real numbers $\{d_n\}$ whose generic term is

$$d_n = d(a_n, a)$$

and noting that the definition of convergence of $\{a_n\}$ to a, which is

$$\forall \varepsilon > 0, \exists n_0 \in \mathbb{N} : d(a_n, a) < \varepsilon, \forall n > n_0$$

can be written as

$$\forall \varepsilon > 0, \exists n_0 \in \mathbb{N} : |d_n - 0| < \varepsilon, \forall n > n_0$$

which is the definition of convergence of $\{d_n\}$ to 0. ■

So, in practice, the problem of assessing the convergence of a generic sequence of objects is simplified as follows:

1. find a metric $d(a_n, a)$ to measure the distance between the terms of the sequence a_n and the candidate limit a;
2. define a new sequence $\{d_n\}$, where $d_n = d(a_n, a)$;
3. study the convergence of the sequence $\{d_n\}$, which is a simple problem, because $\{d_n\}$ is a sequence of real numbers.

Chapter 7

Review of differentiation rules

This lecture contains a summary of differentiation rules, i.e. of rules for computing the derivative of a function. This review is neither detailed nor rigorous and it is not meant to be a substitute for a proper lecture on differentiation. Its only purpose it to serve as a quick review of differentiation rules.

In what follows, $f(x)$ will denote a function of one variable and $\frac{d}{dx}f(x)$ will denote its first derivative.

7.1 Derivative of a constant function

If $f(x)$ is a constant function:

$$f(x) = c$$

where $c \in \mathbb{R}$, then its first derivative is

$$\frac{d}{dx}f(x) = 0$$

7.2 Derivative of a power function

If $f(x)$ is a power function:

$$f(x) = x^n$$

where $n \in \mathbb{R}$, then its first derivative is

$$\frac{d}{dx}f(x) = nx^{n-1}$$

Example 40 *Define*

$$f(x) = x^5$$

The derivative of $f(x)$ *is*

$$\frac{d}{dx}f(x) = 5x^{5-1} = 5x^4$$

Example 41 *Define*

$$f(x) = \sqrt[3]{x^4}$$

The derivative of $f(x)$ *is*

$$\frac{d}{dx}f(x) = \frac{d}{dx}\left(\sqrt[3]{x^4}\right) = \frac{d}{dx}\left(x^{4/3}\right) = \frac{4}{3}x^{4/3-1} = \frac{4}{3}x^{1/3}$$

7.3 Derivative of a logarithmic function

If $f(x)$ is the natural logarithm of x:

$$f(x) = \ln(x)$$

then its first derivative is

$$\frac{d}{dx}f(x) = \frac{1}{x}$$

If $f(x)$ is the logarithm to base b of x:

$$f(x) = \log_b(x)$$

then its first derivative is[1]

$$\frac{d}{dx}f(x) = \frac{1}{x\ln(b)}$$

Example 42 *Define*

$$f(x) = \log_2(x)$$

The derivative of $f(x)$ *is*

$$\frac{d}{dx}f(x) = \frac{1}{x\ln(2)}$$

7.4 Derivative of an exponential function

If $f(x)$ is the exponential function

$$f(x) = \exp(x)$$

then its first derivative is

$$\frac{d}{dx}f(x) = \exp(x)$$

If the exponential function $f(x)$ does not have the natural base e, but another positive base b:

$$f(x) = b^x$$

then its first derivative is[2]

$$\frac{d}{dx}f(x) = \ln(b)\, b^x$$

Example 43 *Define*

$$f(x) = 5^x$$

The derivative of $f(x)$ *is*

$$\frac{d}{dx}f(x) = \ln(5)\, 5^x$$

[1] Remember that $\log_b(x) = \frac{\ln(x)}{\ln(b)}$.

[2] Remember that $b^x = \exp(x\ln(b))$.

7.5 Derivative of a linear combination

If $f_1(x)$ and $f_2(x)$ are two functions and $c_1, c_2 \in \mathbb{R}$ are two constants, then

$$\frac{d}{dx}\left(c_1 f_1(x) + c_2 f_2(x)\right) = c_1 \frac{d}{dx} f_1(x) + c_2 \frac{d}{dx} f_2(x)$$

In other words, the derivative of a linear combination is equal to the linear combinations of the derivatives. This property is called "linearity of the derivative".

Two special cases of this rule are

1. Multiplication by a constant
$$\frac{d}{dx}\left(c_1 f_1(x)\right) = c_1 \frac{d}{dx} f_1(x)$$
2. Addition
$$\frac{d}{dx}\left(f_1(x) + f_2(x)\right) = \frac{d}{dx} f_1(x) + \frac{d}{dx} f_2(x)$$

Example 44 *Define*

$$f(x) = 2 + \exp(x)$$

The derivative of $f(x)$ *is*

$$\frac{d}{dx} f(x) = \frac{d}{dx}\left(2 + \exp(x)\right) = \frac{d}{dx}(2) + \frac{d}{dx}\left(\exp(x)\right)$$

The first summand is

$$\frac{d}{dx}(2) = 0$$

because the derivative of a constant is 0*. The second summand is*

$$\frac{d}{dx}\left(\exp(x)\right) = \exp(x)$$

by the rule for differentiating exponentials. Therefore

$$\frac{d}{dx} f(x) = \frac{d}{dx}(2) + \frac{d}{dx}\left(\exp(x)\right) = 0 + \exp(x) = \exp(x)$$

7.6 Derivative of a product of functions

If $f_1(x)$ and $f_2(x)$ are two functions, then the derivative of their product is

$$\frac{d}{dx}\left(f_1(x) f_2(x)\right) = \left(\frac{d}{dx} f_1(x)\right) f_2(x) + f_1(x)\left(\frac{d}{dx} f_2(x)\right)$$

Example 45 *Define*

$$f(x) = x \ln(x)$$

The derivative of $f(x)$ *is*

$$\begin{aligned} \frac{d}{dx} f(x) &= \frac{d}{dx}\left(x \ln(x)\right) = \frac{d}{dx}(x) \cdot \ln(x) + x \cdot \frac{d}{dx}\left(\ln(x)\right) \\ &= 1 \cdot \ln(x) + x \cdot \frac{1}{x} = \ln(x) + 1 \end{aligned}$$

7.7 Derivative of a composition of functions

If $g(y)$ and $h(x)$ are two functions, then the derivative of their composition is

$$\frac{d}{dx}\left(g\left(h\left(x\right)\right)\right)=\left(\frac{d}{dy}g\left(y\right)\Big|_{y=h(x)}\right)\frac{d}{dx}h\left(x\right)$$

What does this chain rule mean in practice? It means that first you need to compute the derivative of $g(y)$:

$$\frac{d}{dy}g\left(y\right)$$

Then, you substitute y with $h(x)$:

$$\frac{d}{dy}g\left(y\right)\Big|_{y=h(x)}$$

Finally, you multiply it by the derivative of $h(x)$:

$$\frac{d}{dx}h\left(x\right)$$

Example 46 *Define*

$$f\left(x\right)=\ln\left(x^{2}\right)$$

The function $f(x)$ is a composite function:

$$f\left(x\right)=g\left(h\left(x\right)\right)$$

where

$$g\left(y\right)=\ln\left(y\right)$$

and

$$h\left(x\right)=x^{2}$$

The derivative of $h(x)$ is

$$\frac{d}{dx}h\left(x\right)=\frac{d}{dx}\left(x^{2}\right)=2x$$

The derivative of $g(y)$ is

$$\frac{d}{dy}g\left(y\right)=\frac{d}{dy}\left(\ln\left(y\right)\right)=\frac{1}{y}$$

which, evaluated at $y=h(x)=x^{2}$, gives

$$\frac{d}{dy}g\left(y\right)\Big|_{y=h(x)}=\frac{1}{h\left(x\right)}=\frac{1}{x^{2}}$$

Therefore

$$\frac{d}{dx}\left(g\left(h\left(x\right)\right)\right)=\left(\frac{d}{dy}g\left(y\right)\Big|_{y=h(x)}\right)\frac{d}{dx}h\left(x\right)=\frac{1}{x^{2}}\cdot 2x=\frac{2}{x}$$

7.8 Derivatives of trigonometric functions

The trigonometric functions have the following derivatives:

$$\begin{aligned}\frac{d}{dx}\sin(x) &= \cos(x)\\ \frac{d}{dx}\cos(x) &= -\sin(x)\\ \frac{d}{dx}\tan(x) &= \frac{1}{\cos^2(x)}\end{aligned}$$

while the inverse trigonometric functions have the following derivatives:

$$\begin{aligned}\frac{d}{dx}\arcsin(x) &= \frac{1}{\sqrt{1-x^2}}\\ \frac{d}{dx}\arccos(x) &= -\frac{1}{\sqrt{1-x^2}}\\ \frac{d}{dx}\arctan(x) &= \frac{1}{1+x^2}\end{aligned}$$

Example 47 *Define*

$$f(x) = \cos\left(x^2\right)$$

We need to use the chain rule for the derivative of a composite function:

$$\frac{d}{dx}(g(h(x))) = \left(\frac{d}{dy}g(y)\bigg|_{y=h(x)}\right)\frac{d}{dx}h(x)$$

The derivative of $h(x)$ is

$$\frac{d}{dx}h(x) = \frac{d}{dx}\left(x^2\right) = 2x$$

The derivative of $g(y)$ is

$$\frac{d}{dy}g(y) = \frac{d}{dy}(\cos(y)) = -\sin(y)$$

which, evaluated at $y = h(x) = x^2$, gives

$$\frac{d}{dy}g(y)\bigg|_{y=h(x)} = -\sin(h(x)) = -\sin\left(x^2\right)$$

Therefore

$$\frac{d}{dx}(g(h(x))) = \left(\frac{d}{dy}g(y)\bigg|_{y=h(x)}\right)\frac{d}{dx}h(x) = -\sin\left(x^2\right)\cdot 2x$$

7.9 Derivative of an inverse function

If $y = f(x)$ is a function with derivative

$$\frac{d}{dx}f(x)$$

then its inverse $x = f^{-1}(y)$ has derivative

$$\frac{d}{dy} f^{-1}(y) = \left(\frac{d}{dx} f(x) \Big|_{x=f^{-1}(y)} \right)^{-1}$$

Example 48 *Define*

$$f(x) = \exp(3x)$$

Its inverse is

$$f^{-1}(y) = \frac{1}{3} \ln(y)$$

The derivative of $f(x)$ *is*

$$\frac{d}{dx} f(x) = 3 \exp(3x)$$

As a consequence

$$\frac{d}{dx} f(x) \Big|_{x=f^{-1}(y)} = 3 \exp(3x)|_{x=\frac{1}{3}\ln(y)} = 3 \exp\left(3 \cdot \frac{1}{3} \ln(y) \right) = 3y$$

and

$$\frac{d}{dy} f^{-1}(y) = \left(\frac{d}{dx} f(x) \Big|_{x=f^{-1}(y)} \right)^{-1} = (3y)^{-1}$$

Chapter 8

Review of integration rules

This lecture contains a summary of integration rules, i.e. of rules for computing definite and indefinite integrals of a function. This review is neither detailed nor rigorous and it is not meant to be a substitute for a proper lecture on integration. Its only purpose it to serve as a quick review of integration rules.

In what follows, $f(x)$ will denote a function of one variable and $\frac{d}{dx}f(x)$ will denote its first derivative.

8.1 Indefinite integrals

If $f(x)$ is a function of one variable, an **indefinite integral** of $f(x)$ is a function $F(x)$ whose first derivative is equal to $f(x)$:

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

An indefinite integral $F(x)$ is denoted by

$$F(x) = \int f(x)\,dx$$

Indefinite integrals are also called **antiderivatives** or **primitives**.

Example 49 *Let*

$$f(x) = x^3$$

The function

$$F(x) = \frac{1}{4}x^4$$

is an indefinite integral of $f(x)$ *because*

$$\frac{d}{dx}F(x) = \frac{d}{dx}\left(\frac{1}{4}x^4\right) = \frac{1}{4}\frac{d}{dx}\left(x^4\right) = \frac{1}{4}\cdot 4x^3 = x^3$$

Also the function

$$G(x) = \frac{1}{2} + \frac{1}{4}x^4$$

is an indefinite integral of $f(x)$ *because*

$$\begin{aligned}\frac{d}{dx}G(x) &= \frac{d}{dx}\left(\frac{1}{2}+\frac{1}{4}x^4\right) = \frac{d}{dx}\left(\frac{1}{2}\right)+\frac{1}{4}\frac{d}{dx}\left(x^4\right)\\ &= 0+\frac{1}{4}\cdot 4x^3 = x^3\end{aligned}$$

Note that if a function $F(x)$ is an indefinite integral of $f(x)$ then also the function

$$G(x) = F(x) + c$$

is an indefinite integral of $f(x)$ for any constant $c \in \mathbb{R}$, because

$$\begin{aligned}\frac{d}{dx}G(x) &= \frac{d}{dx}(F(x)+c) = \frac{d}{dx}(F(x)) + \frac{d}{dx}(c)\\ &= f(x)+0 = f(x)\end{aligned}$$

This is also the reason why the adjective indefinite is used: because indefinite integrals are defined only up to a constant.

The following subsections contain some rules for computing the indefinite integrals of functions that are frequently encountered in probability theory and statistics. In all these subsections, c will denote a constant and the integration rules will be reported without a proof. Proofs are trivial and can be easily performed by the reader: it suffices to compute the first derivative of $F(x)$ and verify that it equals $f(x)$.

8.1.1 Indefinite integral of a constant function

If $f(x)$ is a constant function:

$$f(x) = a$$

where $a \in \mathbb{R}$, then an indefinite integral of $f(x)$ is

$$F(x) = ax + c$$

8.1.2 Indefinite integral of a power function

If $f(x)$ is a power function:

$$f(x) = x^n$$

then an indefinite integral of $f(x)$ is

$$F(x) = \frac{1}{n+1}x^{n+1} + c$$

when $n \neq -1$. When $n = -1$, i.e. when

$$f(x) = \frac{1}{x}$$

the integral is

$$F(x) = \ln(x) + c$$

8.1.3 Indefinite integral of a logarithmic function

If $f(x)$ is the natural logarithm of x:

$$f(x) = \ln(x)$$

then its indefinite integral is

$$F(x) = x\ln(x) - x + c$$

If $f(x)$ is the logarithm to base b of x:

$$f(x) = \log_b(x)$$

then its indefinite integral is[1]

$$F(x) = \frac{1}{\ln(b)}(x\ln(x) - x) + c$$

8.1.4 Indefinite integral of an exponential function

If $f(x)$ is the exponential function:

$$f(x) = \exp(x)$$

then its indefinite integral is

$$F(x) = \exp(x) + c$$

If the exponential function $f(x)$ does not have the natural base e, but another positive base b:

$$f(x) = b^x$$

then its indefinite integral is[2]

$$F(x) = \frac{1}{\ln(b)}b^x + c$$

8.1.5 Indefinite integral of a linear combination of functions

If $f_1(x)$ and $f_2(x)$ are two functions and $c_1, c_2 \in \mathbb{R}$ are two constants, then:

$$\int (c_1 f_1(x) + c_2 f_2(x))\,dx = c_1 \int f_1(x)\,dx + c_2 \int f_2(x)\,dx$$

In other words, the integral of a linear combination is equal to the linear combinations of the integrals. This property is called "linearity of the integral".

Two special cases of this rule are:

1. Multiplication by a constant:

$$\int c_1 f_1(x)\,dx = c_1 \int f_1(x)\,dx$$

2. Addition:

$$\int (f_1(x) + f_2(x))\,dx = \int f_1(x)\,dx + \int f_2(x)\,dx$$

[1] Remember that $\log_b(x) = \frac{\ln(x)}{\ln(b)}$.

[2] Remember that $b^x = \exp(x\ln(b))$.

8.1.6 Indefinite integrals of trigonometric functions

The trigonometric functions have the following indefinite integrals:

$$\begin{aligned}\int \sin(x)\,dx &= -\cos(x) + c \\ \int \cos(x)\,dx &= \sin(x) + c \\ \int \tan(x)\,dx &= \ln\left(\left|\frac{1}{\cos(x)}\right|\right) + c\end{aligned}$$

8.2 Definite integrals

Let $f(x)$ be a function of one variable and $[a, b]$ an interval of real numbers. The **definite integral** (or, simply, the **integral**) from a to b of $f(x)$ is the area of the region in the xy-plane bounded by the graph of $f(x)$, the x-axis and the vertical lines $x = a$ and $x = b$, where regions below the x-axis have negative sign and regions above the x-axis have positive sign.

The integral from a to b of $f(x)$ is denoted by

$$\int_a^b f(x)\,dx$$

$f(x)$ is called the **integrand function** and a and b are called the **upper bound of integration** and the **lower bound of integration**.

The following subsections contain some properties of definite integrals, which are also often utilized to actually compute definite integrals.

8.2.1 Fundamental theorem of calculus

The fundamental theorem of calculus provides the link between definite and indefinite integrals. It has two parts.

On the one hand, if you define

$$F(x) = \int_a^x f(t)\,dt$$

then the first derivative of $F(x)$ is equal to $f(x)$, i.e.

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

In other words, if you differentiate a definite integral with respect to its upper bound of integration, then you obtain the integrand function.

Example 50 *Define*

$$F(x) = \int_a^x \exp(2t)\,dt$$

Then:

$$\frac{d}{dx}F(x) = \exp(2x)$$

On the other hand, if $F(x)$ is an indefinite integral (an antiderivative) of $f(x)$, then

$$\int_a^b f(x)\,dx = F(b) - F(a)$$

In other words, you can use the indefinite integral to compute the definite integral.

The following notation is often used:

$$\int_a^b f(x)\,dx = [F(x)]_a^b$$

where

$$[F(x)]_a^b = F(b) - F(a)$$

Sometimes the variable of integration x is explicitly specified and we write

$$[F(x)]_{x=a}^{x=b}$$

Example 51 *Consider the definite integral*

$$\int_0^1 x^2 dx$$

The integrand function is

$$f(x) = x^2$$

An indefinite integral of $f(x)$ *is*

$$F(x) = \frac{1}{3}x^3$$

Therefore, the definite integral from 0 *to* 1 *can be computed as follows:*

$$\int_0^1 x^2 dx = \left[\frac{1}{3}x^3\right]_0^1 = \frac{1}{3}\cdot 1^3 - \frac{1}{3}\cdot 0^3 = \frac{1}{3}$$

8.2.2 Definite integral of a linear combination of functions

Like indefinite integrals, also definite integrals are linear. If $f_1(x)$ and $f_2(x)$ are two functions and $c_1, c_2 \in \mathbb{R}$ are two constants, then:

$$\int_a^b (c_1 f_1(x) + c_2 f_2(x))\,dx = c_1 \int_a^b f_1(x)\,dx + c_2 \int_a^b f_2(x)\,dx$$

with the two special cases:

1. Multiplication by a constant:

$$\int_a^b c_1 f_1(x)\,dx = c_1 \int_a^b f_1(x)\,dx$$

2. Addition:

$$\int_a^b (f_1(x) + f_2(x))\,dx = \int_a^b f_1(x)\,dx + \int_a^b f_2(x)\,dx$$

Example 52 *For example:*

$$\int_0^1 \left(3x + 2x^2\right) dx = 3\int_0^1 x dx + 2\int_0^1 x^2 dx$$

8.2.3 Change of variable

If $f(x)$ and $g(x)$ are two functions, then the following integral

$$\int_a^b f(g(x)) \left(\frac{d}{dx} g(x) \right) dx$$

can be computed by a change of variable, using the variable

$$t = g(x)$$

The change of variable is performed in the following steps:

1. Differentiate the change of variable formula

   $$t = g(x)$$

   and obtain

   $$dt = \frac{d}{dx} g(x)\, dx$$

2. Recompute the bounds of integration:

   $$\begin{aligned} x &= a \Rightarrow t = g(x) = g(a) \\ x &= b \Rightarrow t = g(x) = g(b) \end{aligned}$$

3. Substitute $g(x)$ and $\frac{d}{dx} g(x)\, dx$ in the integral:

   $$\int_a^b f(g(x)) \left(\frac{d}{dx} g(x) \right) dx = \int_{g(a)}^{g(b)} f(t)\, dt$$

Example 53 *The integral*

$$\int_1^2 \frac{\ln(x)}{x} dx$$

can be computed performing the change of variable

$$t = \ln(x)$$

Differentiating the change of variable formula, we obtain

$$dt = \frac{d}{dx} \ln(x)\, dx = \frac{1}{x} dx$$

The new bounds of integration are

$$\begin{aligned} x &= 1 \Rightarrow t = \ln(1) = 0 \\ x &= 2 \Rightarrow t = \ln(2) \end{aligned}$$

Therefore, the integral can be written as follows:

$$\int_1^2 \frac{\ln(x)}{x} dx = \int_0^{\ln(2)} t dt$$

8.2.4 Integration by parts

Let $f(x)$ and $g(x)$ be two functions and $F(x)$ and $G(x)$ their indefinite integrals. The following integration by parts formula holds:

$$\int_a^b f(x) G(x) dx = [F(x) G(x)]_a^b - \int_a^b F(x) g(x) dx$$

Example 54 *The integral*

$$\int_0^1 \exp(x) x dx$$

can be integrated by parts, by setting

$$\begin{aligned} f(x) &= \exp(x) \\ G(x) &= x \end{aligned}$$

An indefinite integral of $f(x)$ *is*

$$F(x) = \exp(x)$$

and $G(x)$ *is an indefinite integral of*

$$g(x) = 1$$

or, said differently, $g(x) = 1$ *is the derivative of* $G(x) = x$*. Therefore:*

$$\begin{aligned} \int_0^1 \exp(x) x dx &= [\exp(x) x]_0^1 - \int_0^1 \exp(x) dx \\ &= \exp(1) - 0 - \int_0^1 \exp(x) dx \\ &= \exp(1) - [\exp(x)]_0^1 \\ &= \exp(1) - [\exp(1) - 1] = 1 \end{aligned}$$

8.2.5 Exchanging the bounds of integration

Given the integral

$$\int_a^b f(x) dx$$

exchanging its bounds of integration is equivalent to changing its sign:

$$\int_b^a f(x) dx = -\int_a^b f(x) dx$$

8.2.6 Subdividing the integral

Given the two bounds of integration a and b, with $a \leq b$, and a third point m such that $a \leq m \leq b$, then

$$\int_a^b f(x) dx = \int_a^m f(x) dx + \int_m^b f(x) dx$$

8.2.7 Leibniz integral rule

Given a function of two variables $f(x,y)$ and the integral

$$I(y) = \int_{a(y)}^{b(y)} f(x,y)\,dx$$

where both the lower bound of integration a and the upper bound of integration b may depend on y, under appropriate technical conditions (not discussed here) the first derivative of the function $I(y)$ with respect to y can be computed as follows:

$$\begin{aligned}\frac{d}{dy}I(y) &= \left(\frac{d}{dy}b(y)\right) f(b(y),y) - \left(\frac{d}{dy}a(y)\right) f(a(y),y) \\ &\quad + \int_{a(y)}^{b(y)} \frac{\partial}{\partial y} f(x,y)\,dx\end{aligned}$$

where $\frac{\partial}{\partial y} f(x,y)$ is the first partial derivative of $f(x,y)$ with respect to y.

Example 55 *The derivative of the integral*

$$I(y) = \int_{y^2}^{y^2+1} \exp(xy)\,dx$$

is

$$\begin{aligned}\frac{d}{dy}I(y) &= \left(\frac{d}{dy}\left(y^2+1\right)\right) \exp\left(\left(y^2+1\right)y\right) \\ &\quad - \left(\frac{d}{dy}y^2\right) \exp\left(y^2 y\right) + \int_{y^2}^{y^2+1} \frac{\partial}{\partial y}\left(\exp(xy)\right) dx \\ &= 2y\exp\left(y^3+y\right) - 2y\exp\left(y^3\right) + \int_{y^2}^{y^2+1} x\exp(xy)\,dx\end{aligned}$$

8.3 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Compute the following integral:

$$\int_0^{\infty} \cos(x)\exp(-x)\,dx$$

Hint: perform two integrations by parts.

Solution

Performing two integrations by parts, we obtain:

$$\int_0^{\infty} \cos(x)\exp(-x)\,dx$$

$$\begin{aligned}
\boxed{\text{A}} &= \left[\sin(x)\exp(-x)\right]_0^\infty - \int_0^\infty \sin(x)\left(-\exp(-x)\right)dx \\
&= 0 - 0 + \int_0^\infty \sin(x)\exp(-x)\,dx \\
\boxed{\text{B}} &= \left[-\cos(x)\exp(-x)\right]_0^\infty - \int_0^\infty \left(-\cos(x)\right)\left(-\exp(-x)\right)dx \\
&= 0 - (-1) - \int_0^\infty \cos(x)\exp(-x)\,dx \\
&= 1 - \int_0^\infty \cos(x)\exp(-x)\,dx
\end{aligned}$$

where integration by parts has been performed in steps A and B. Therefore

$$\int_0^\infty \cos(x)\exp(-x)\,dx = 1 - \int_0^\infty \cos(x)\exp(-x)\,dx$$

which can be rearranged to yield

$$2\int_0^\infty \cos(x)\exp(-x)\,dx = 1$$

or

$$\int_0^\infty \cos(x)\exp(-x)\,dx = \frac{1}{2}$$

Exercise 2

Use Leibniz integral rule to compute the derivative with respect to y of the following integral:

$$I(y) = \int_0^{y^2} \exp(-xy)\,dx$$

Solution

Leibniz integral rule is

$$\begin{aligned}
\frac{d}{dy}\int_{a(y)}^{b(y)} f(x,y)\,dx &= \left(\frac{d}{dy}b(y)\right) f\left(b(y),y\right) - \left(\frac{d}{dy}a(y)\right) f\left(a(y),y\right) \\
&\quad + \int_{a(y)}^{b(y)} \frac{\partial}{\partial y} f(x,y)\,dx
\end{aligned}$$

We can apply it as follows:

$$\begin{aligned}
\frac{d}{dy}I(y) &= \frac{d}{dy}\int_0^{y^2} \exp(-xy)\,dx \\
&= \left(\frac{d}{dy}y^2\right)\exp\left(-y^2 y\right) + \int_0^{y^2} \frac{\partial}{\partial y}\exp(-xy)\,dx \\
&= 2y\exp\left(-y^3\right) - \int_0^{y^2} x\exp(-xy)\,dx
\end{aligned}$$

$$\begin{aligned}
\boxed{\text{A}} &= 2y\exp\left(-y^3\right) - \left\{\left[x\left(-\frac{1}{y}\exp\left(-xy\right)\right)\right]_0^{y^2} + \frac{1}{y}\int_0^{y^2}\exp\left(-xy\right)dx\right\} \\
&= 2y\exp\left(-y^3\right) + y\exp\left(-y^3\right) - \frac{1}{y}\left[-\frac{1}{y}\exp\left(-xy\right)\right]_0^{y^2} \\
&= 3y\exp\left(-y^3\right) + \frac{1}{y^2}\exp\left(-y^3\right) - \frac{1}{y^2}
\end{aligned}$$

where: in step $\boxed{\text{A}}$ we have performed an integration by parts.

Exercise 3

Compute the following integral:

$$\int_0^1 x\left(1+x^2\right)^{-2}dx$$

Solution

This integral can be solved using the change of variable technique:

$$\begin{aligned}
\int_0^1 x\left(1+x^2\right)^{-2}dx &= \int_0^1 \frac{1}{2}\left(1+t\right)^{-2}dt \\
= \left[-\frac{1}{2}\left(1+t\right)^{-1}\right]_0^1 &= -\frac{1}{2}\cdot\frac{1}{2} + \frac{1}{2}\cdot 1 = \frac{1}{4}
\end{aligned}$$

where the change of variable is

$$t = x^2$$

and the differential is

$$dt = 2xdx$$

so that we can substitute xdx with $\frac{1}{2}dt$.

Chapter 9

Special functions

This chapter briefly introduces some special functions that are frequently used in probability and statistics.

9.1 Gamma function

The Gamma function is a generalization of the factorial function[1] to non-integer numbers.

Recall that if $n \in \mathbb{N}$, its factorial $n!$ is

$$n! = 1 \cdot 2 \cdot \ldots \cdot (n-1) \cdot n$$

so that $n!$ satisfies the recursion

$$n! = (n-1)! \cdot n$$

The Gamma function $\Gamma(z)$ satisfies a similar recursion:

$$\Gamma(z) = \Gamma(z-1) \cdot (z-1)$$

but it is defined also when z is not an integer.

9.1.1 Definition

The following is a possible definition of the Gamma function.

Definition 56 *The Gamma function* Γ *is a function* $\Gamma : \mathbb{R}_{++} \to \mathbb{R}_{++}$ *satisfying the equation*

$$\Gamma(z) = \int_0^\infty x^{z-1} \exp(-x)\, dx$$

While the domain of definition of the Gamma function can be extended beyond the set $\mathbb{R}_{++}$ of strictly positive real numbers, for example, to complex numbers, the somewhat restrictive definition given above is more than sufficient to address all the problems involving the Gamma function that are found in these lectures.

[1] See p. 10.

9.1.2 Recursion

The next proposition states a recursive property that is used to derive several other properties of the Gamma function.

Proposition 57 *The Gamma function satisfies the recursion*

$$\Gamma(z) = \Gamma(z-1) \cdot (z-1) \tag{9.1}$$

Proof. By integrating by parts[2], we get

$$\begin{aligned}
\Gamma(z) &= \int_0^\infty x^{z-1} \exp(-x)\, dx \\
&= \left[-x^{z-1} \exp(-x)\right]_0^\infty + \int_0^\infty (z-1)\, x^{z-2} \exp(-x)\, dx \\
&= (0-0) + (z-1) \int_0^\infty x^{(z-1)-1} \exp(-x)\, dx \\
&= (z-1)\, \Gamma(z-1)
\end{aligned}$$

■

9.1.3 Relation to the factorial function

The next proposition states the relation between the Gamma and factorial functions.

Proposition 58 *When the argument of the Gamma function is an integer $n \in \mathbb{N}$, then its value is equal to the factorial of $n-1$:*

$$\Gamma(n) = (n-1)!$$

Proof. First of all, we need to compute a starting value:

$$\Gamma(1) = \int_0^\infty x^{1-1} \exp(-x)\, dx = \int_0^\infty \exp(-x)\, dx = \left[-\exp(-x)\right]_0^\infty = 1$$

By using the recursion (9.1), we obtain

$$\begin{aligned}
\Gamma(1) &= 1 = 0! \\
\Gamma(2) &= \Gamma(2-1) \cdot (2-1) = \Gamma(1) \cdot 1 = 1 = 1! \\
\Gamma(3) &= \Gamma(3-1) \cdot (3-1) = \Gamma(2) \cdot 2 = 1 \cdot 2 = 2! \\
\Gamma(4) &= \Gamma(4-1) \cdot (4-1) = \Gamma(3) \cdot 3 = 1 \cdot 2 \cdot 3 = 3! \\
&\vdots \\
\Gamma(n) &= \Gamma(n-1) \cdot (n-1) = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot (n-1) = (n-1)!
\end{aligned}$$

■

[2] See p.51.

9.1.4 Values of the Gamma function

The next proposition states a well-known fact, which is often used in probability theory and statistics.

Proposition 59 *The Gamma function is such that*

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

Proof. Using the definition of Gamma function and performing a change of variable, we obtain

$$\begin{aligned}
\Gamma\left(\frac{1}{2}\right) &= \int_0^\infty x^{1/2-1}\exp\left(-x\right)dx \\
&= \int_0^\infty x^{-1/2}\exp\left(-x\right)dx \\
\boxed{\text{A}} &= 2\int_0^\infty \exp\left(-t^2\right)dt \\
&= 2\left(\int_0^\infty \exp\left(-t^2\right)dt\int_0^\infty \exp\left(-t^2\right)dt\right)^{1/2} \\
&= 2\left(\int_0^\infty \exp\left(-t^2\right)dt\int_0^\infty \exp\left(-s^2\right)ds\right)^{1/2} \\
&= 2\left(\int_0^\infty\int_0^\infty \exp\left(-t^2-s^2\right)dtds\right)^{1/2} \\
\boxed{\text{B}} &= 2\left(\int_0^\infty\int_0^\infty \exp\left(-s^2u^2-s^2\right)sduds\right)^{1/2} \\
&= 2\left(\int_0^\infty\int_0^\infty \exp\left(-\left(1+u^2\right)s^2\right)sdsdu\right)^{1/2} \\
&= 2\left(\int_0^\infty\left[-\frac{1}{2\left(1+u^2\right)}\exp\left(-\left(1+u^2\right)s^2\right)\right]_0^\infty du\right)^{1/2} \\
&= 2\left(\int_0^\infty\left[0+\frac{1}{2\left(1+u^2\right)}\right]du\right)^{1/2} \\
&= 2^{1/2}\left(\int_0^\infty\frac{1}{1+u^2}du\right)^{1/2} \\
&= 2^{1/2}\left(\left[\arctan\left(u\right)\right]_0^\infty\right)^{1/2} \\
&= 2^{1/2}\left(\arctan\left(\infty\right)-\arctan\left(0\right)\right)^{1/2} \\
&= 2^{1/2}\left(\frac{\pi}{2}-0\right)^{1/2} = \pi^{1/2}
\end{aligned}$$

where: in step A we have made the change of variable $t = x^{1/2}$; in step B we have performed the change of variable $t = su$. ■

By using the result stated in the previous proposition, it is possible to derive other values of the Gamma function.

Proposition 60 *The Gamma function is such that*

$$\Gamma\left(n+\frac{1}{2}\right)=\sqrt{\pi}\prod_{j=0}^{n-1}\left(j+\frac{1}{2}\right)$$

for $n\in\mathbb{N}$.

Proof. The result is obtained by iterating the recursion formula (9.1):

$$\begin{aligned}
&\Gamma\left(n+\frac{1}{2}\right)\\
=&\left(n-1+\frac{1}{2}\right)\Gamma\left(n-1+\frac{1}{2}\right)\\
=&\left(n-1+\frac{1}{2}\right)\left(n-2+\frac{1}{2}\right)\Gamma\left(n-2+\frac{1}{2}\right)\\
&\vdots\\
=&\left(n-1+\frac{1}{2}\right)\left(n-2+\frac{1}{2}\right)\ldots\left(n-n+\frac{1}{2}\right)\Gamma\left(n-n+\frac{1}{2}\right)\\
=&\left(n-1+\frac{1}{2}\right)\left(n-2+\frac{1}{2}\right)\ldots\left(\frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right)\\
=&\sqrt{\pi}\prod_{j=0}^{n-1}\left(j+\frac{1}{2}\right)
\end{aligned}$$

■

There are also other special cases in which the value of the Gamma function can be derived analytically, but it is not possible to express $\Gamma(z)$ in terms of elementary functions for every z. As a consequence, one often needs to resort to numerical algorithms to compute $\Gamma(z)$. For example, the Matlab command

```
gamma(z)
```

returns the value of the Gamma function at the point `z`.

For a thorough discussion of a number of algorithms that can be employed to compute numerical approximations of $\Gamma(z)$ see Abramowitz and Stegun[3] (1965).

9.1.5 Lower incomplete Gamma function

The definition of the Gamma function

$$\Gamma(z)=\int_0^\infty x^{z-1}\exp(-x)\,dx$$

can be generalized by substituting the upper bound of integration, equal to infinity, with a variable y:

$$\gamma(z,y)=\int_0^y x^{z-1}\exp(-x)\,dx$$

The function $\gamma(z,y)$ thus obtained is called lower incomplete Gamma function.

[3] Abramowitz, M. and I. A. Stegun (1965) *Hanbook of mathematical functions: with formulas, graphs, and mathematical tables*, Courier Dover Publications.

9.2 Beta function

The Beta function is a function of two variables that is often found in probability theory and mathematical statistics. We report here some basic facts about the Beta function.

9.2.1 Definition

The following is a possible definition of the Beta function.

Definition 61 *The **Beta function** is a function $B : \mathbb{R}^2_{++} \to \mathbb{R}_{++}$ defined by*

$$B(x,y) = \frac{\Gamma(x)\,\Gamma(y)}{\Gamma(x+y)}$$

where $\Gamma()$ is the Gamma function.

While the domain of definition of the Beta function can be extended beyond the set $\mathbb{R}^2_{++}$ of couples of strictly positive real numbers, for example, to couples of complex numbers, the somewhat restrictive definition given above is more than sufficient to address all the problems involving the Beta function that are found in these lectures.

9.2.2 Integral representations

The Beta function has several integral representations, which are sometimes also used as a definition of the Beta function in place of the definition we have given above. We report here two often used representations.

Integral between zero and infinity

The first representation involves an integral from zero to infinity.

Proposition 62 *The Beta function has the integral representation*

$$B(x,y) = \int_0^\infty t^{x-1}(1+t)^{-x-y}\,dt \tag{9.2}$$

Proof. Given the definition of the Beta function as a ratio of Gamma functions, the equality holds if and only if

$$\int_0^\infty t^{x-1}(1+t)^{-x-y}\,dt = \frac{\Gamma(x)\,\Gamma(y)}{\Gamma(x+y)}$$

or

$$\Gamma(x+y)\int_0^\infty t^{x-1}(1+t)^{-x-y}\,dt = \Gamma(x)\,\Gamma(y)$$

That the latter equality indeed holds is proved as follows:

$$\begin{aligned} & & \Gamma(x)\,\Gamma(y) \\ \boxed{\text{A}} & = & \int_0^\infty u^{x-1}\exp(-u)\,du \int_0^\infty v^{y-1}\exp(-v)\,dv \end{aligned}$$

$$
\begin{aligned}
& = \int_0^\infty v^{y-1}\exp(-v)\int_0^\infty u^{x-1}\exp(-u)\,dudv \\
\boxed{\text{B}} \quad & = \int_0^\infty v^{y-1}\exp(-v)\int_0^\infty (vt)^{x-1}\exp(-vt)\,vdtdv \\
& = \int_0^\infty v^{y-1}\exp(-v)\int_0^\infty v^x t^{x-1}\exp(-vt)\,dtdv \\
& = \int_0^\infty v^{x+y-1}\exp(-v)\int_0^\infty t^{x-1}\exp(-vt)\,dtdv \\
& = \int_0^\infty\int_0^\infty v^{x+y-1}t^{x-1}\exp(-(1+t)v)\,dtdv \\
& = \int_0^\infty t^{x-1}\int_0^\infty v^{x+y-1}\exp(-(1+t)v)\,dvdt \\
\boxed{\text{C}} \quad & = \int_0^\infty t^{x-1}\int_0^\infty \left(\frac{s}{1+t}\right)^{x+y-1}\exp(-s)\,\frac{1}{1+t}dsdt \\
& = \int_0^\infty t^{x-1}(1+t)^{-x-y}\int_0^\infty s^{x+y-1}\exp(-s)\,dsdt \\
\boxed{\text{D}} \quad & = \int_0^\infty t^{x-1}(1+t)^{-x-y}\Gamma(x+y)\,dt \\
& = \Gamma(x+y)\int_0^\infty t^{x-1}(1+t)^{-x-y}\,dt
\end{aligned}
$$

where: in step A we have used the definition of Gamma function; in step B we have performed the change of variable $u = vt$; in step C we have made the change of variable $s = (1+t)v$; in step D we have again used the definition of Gamma function. ■

Integral between zero and one

Another representation involves an integral from zero to one.

Proposition 63 *The Beta function has the integral representation*

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\,dt \tag{9.3}$$

Proof. This can be obtained from the previous integral representation:

$$B(x,y) = \int_0^\infty t^{x-1}(1+t)^{-x-y}\,dt$$

by performing a change of variable. The change of variable is

$$s = \frac{t}{1+t} = 1 - \frac{1}{1+t}$$

Before performing it, note that

$$\lim_{t\to\infty}\frac{t}{1+t} = 1$$

and that

$$t = \frac{1}{1-s} - 1 = \frac{s}{1-s}$$

Furthermore, by differentiating the previous expression, we obtain

$$dt = \left(\frac{1}{1-s}\right)^2 ds$$

We are now ready to perform the change of variable:

$$\begin{aligned} B(x,y) &= \int_0^\infty t^{x-1}(1+t)^{-x-y}\,dt \\ &= \int_0^1 \left(\frac{s}{1-s}\right)^{x-1}\left(1+\frac{s}{1-s}\right)^{-x-y}\left(\frac{1}{1-s}\right)^2 ds \\ &= \int_0^1 \left(\frac{s}{1-s}\right)^{x-1}\left(\frac{1}{1-s}\right)^{-x-y}\left(\frac{1}{1-s}\right)^2 ds \\ &= \int_0^1 s^{x-1}\left(\frac{1}{1-s}\right)^{x-1-x-y+2} ds \\ &= \int_0^1 s^{x-1}\left(\frac{1}{1-s}\right)^{1-y} ds \\ &= \int_0^1 s^{x-1}(1-s)^{y-1}\,ds \end{aligned}$$

■

Note that the two representations (9.2) and (9.3) involve improper integrals that converge if $x > 0$ and $y > 0$. This might help you to see why the arguments of the Beta function are required to be strictly positive in Definition 61.

9.3 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Compute the ratio

$$\frac{\Gamma\left(\frac{16}{3}\right)}{\Gamma\left(\frac{10}{3}\right)}$$

Solution

We need to repeatedly apply the recursive formula

$$\Gamma(z) = (z-1)\,\Gamma(z-1)$$

to the numerator of the ratio, as follows:

$$\begin{aligned} \frac{\Gamma\left(\frac{16}{3}\right)}{\Gamma\left(\frac{10}{3}\right)} &= \frac{\left(\frac{16}{3}-1\right)\Gamma\left(\frac{16}{3}-1\right)}{\Gamma\left(\frac{10}{3}\right)} = \frac{13}{3}\frac{\Gamma\left(\frac{13}{3}\right)}{\Gamma\left(\frac{10}{3}\right)} \\ &= \frac{13}{3}\frac{\left(\frac{13}{3}-1\right)\Gamma\left(\frac{13}{3}-1\right)}{\Gamma\left(\frac{10}{3}\right)} = \frac{13}{3}\frac{10}{3}\frac{\Gamma\left(\frac{10}{3}\right)}{\Gamma\left(\frac{10}{3}\right)} = \frac{130}{9} \end{aligned}$$

Exercise 2

Compute

$$\Gamma(5)$$

Solution

We need to use the relation of the Gamma function to the factorial function:

$$\Gamma(n) = (n-1)!$$

which for $n = 5$ becomes

$$\Gamma(5) = (5-1)! = 4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$$

Exercise 3

Express the integral

$$\int_0^\infty x^{9/2} \exp\left(-\frac{1}{2}x\right) dx$$

in terms of the Gamma function.

Solution

This is accomplished as follows:

$$\begin{aligned} & \int_0^\infty x^{9/2} \exp\left(-\frac{1}{2}x\right) dx \\ \boxed{\text{A}} \quad &= \int_0^\infty (2t)^{9/2} \exp(-t)\, 2dt \\ &= 2^{11/2} \int_0^\infty t^{11/2-1} \exp(-t)\, dt \\ \boxed{\text{B}} \quad &= 2^{11/2} \Gamma(11/2) \end{aligned}$$

where: in step A we have performed a change of variable ($t = \frac{1}{2}x$); in step B we have used the definition of Gamma function.

Exercise 4

Compute the product

$$\Gamma\left(\frac{5}{2}\right) B\left(\frac{3}{2}, 1\right)$$

where $\Gamma()$ is the Gamma function and $B()$ is the Beta function.

Solution

We need to write the Beta function in terms of Gamma functions:

$$\Gamma\left(\frac{5}{2}\right) B\left(\frac{3}{2}, 1\right) = \Gamma\left(\frac{5}{2}\right) \frac{\Gamma\left(\frac{3}{2}\right)\Gamma(1)}{\Gamma\left(\frac{3}{2}+1\right)}$$

$$\begin{aligned}
&= \Gamma\left(\frac{5}{2}\right)\frac{\Gamma\left(\frac{3}{2}\right)\Gamma\left(1\right)}{\Gamma\left(\frac{5}{2}\right)} \\
&= \Gamma\left(\frac{3}{2}\right)\Gamma\left(1\right) \\
\boxed{\text{A}} \quad &= \Gamma\left(\frac{3}{2}\right) \\
\boxed{\text{B}} \quad &= \left(\frac{3}{2}-1\right)\Gamma\left(\frac{3}{2}-1\right) \\
&= \frac{1}{2}\Gamma\left(\frac{1}{2}\right) \\
\boxed{\text{C}} \quad &= \frac{1}{2}\sqrt{\pi}
\end{aligned}$$

where: in step A we have used the fact that $\Gamma\left(1\right)=1$; in step B we have used the recursive formula for the Gamma function; in step C we have used the fact that

$$\Gamma\left(\frac{1}{2}\right)=\sqrt{\pi}$$

Exercise 5

Compute the ratio

$$\frac{B\left(\frac{7}{2},\frac{9}{2}\right)}{B\left(\frac{5}{2},\frac{11}{2}\right)}$$

where $B\left(\right)$ is the Beta function.

Solution

This is achieved by rewriting the numerator of the ratio in terms of Gamma functions:

$$\begin{aligned}
&\frac{B\left(\frac{7}{2},\frac{9}{2}\right)}{B\left(\frac{5}{2},\frac{11}{2}\right)} \\
&= \frac{1}{B\left(\frac{5}{2},\frac{11}{2}\right)}\frac{\Gamma\left(\frac{7}{2}\right)\Gamma\left(\frac{9}{2}\right)}{\Gamma\left(\frac{7}{2}+\frac{9}{2}\right)} \\
\boxed{\text{A}} \quad &= \frac{1}{B\left(\frac{5}{2},\frac{11}{2}\right)}\frac{\left(\frac{7}{2}-1\right)\Gamma\left(\frac{7}{2}-1\right)\Gamma\left(\frac{9}{2}\right)}{\Gamma\left(\frac{7}{2}+\frac{9}{2}\right)} \\
\boxed{\text{B}} \quad &= \frac{1}{B\left(\frac{5}{2},\frac{11}{2}\right)}\frac{\frac{5}{2}\Gamma\left(\frac{5}{2}\right)\left[\Gamma\left(\frac{11}{2}\right)/\left(\frac{11}{2}-1\right)\right]}{\Gamma\left(\frac{7}{2}-1+\frac{9}{2}+1\right)} \\
&= \frac{5}{2}\frac{2}{9}\frac{1}{B\left(\frac{5}{2},\frac{11}{2}\right)}\frac{\Gamma\left(\frac{5}{2}\right)\Gamma\left(\frac{11}{2}\right)}{\Gamma\left(\frac{5}{2}+\frac{11}{2}\right)} \\
\boxed{\text{C}} \quad &= \frac{5}{9}\frac{1}{B\left(\frac{5}{2},\frac{11}{2}\right)}B\left(\frac{5}{2},\frac{11}{2}\right)=\frac{5}{9}
\end{aligned}$$

where: in steps A and B we have used the the recursive formula for the Gamma function; in step C we have used the definition of Beta function.

Exercise 6

Compute the integral

$$\int_0^\infty x^{3/2}\left(1+2x\right)^{-5}dx$$

Solution

We first express the integral in terms of the Beta function:

$$\begin{aligned}
& \int_0^\infty x^{3/2}\left(1+2x\right)^{-5}dx \\
\boxed{\text{A}} \quad &= \int_0^\infty \left(\frac{1}{2}t\right)^{3/2}\left(1+t\right)^{-5}\frac{1}{2}dt \\
&= \left(\frac{1}{2}\right)^{5/2}\int_0^\infty t^{3/2}\left(1+t\right)^{-5}dt \\
&= \left(\frac{1}{2}\right)^{5/2}\int_0^\infty t^{5/2-1}\left(1+t\right)^{-5/2-5/2}dt \\
\boxed{\text{B}} \quad &= \left(\frac{1}{2}\right)^{5/2}B\left(\frac{5}{2},\frac{5}{2}\right)
\end{aligned}$$

where: in step $\boxed{\text{A}}$ we have performed a change of variable ($t=2x$); in step $\boxed{\text{B}}$ we have used the integral representation of Beta function.

Now, write the Beta function in terms of Gamma functions:

$$\begin{aligned}
B\left(\frac{5}{2},\frac{5}{2}\right) &= \frac{\Gamma\left(\frac{5}{2}\right)\Gamma\left(\frac{5}{2}\right)}{\Gamma\left(\frac{5}{2}+\frac{5}{2}\right)} \\
&= \frac{\left[\Gamma\left(\frac{5}{2}\right)\right]^2}{\Gamma\left(5\right)} \\
\boxed{\text{A}} \quad &= \frac{\left[\frac{3}{2}\frac{1}{2}\Gamma\left(\frac{1}{2}\right)\right]^2}{4\cdot 3\cdot 2\cdot 1} \\
&= \frac{1}{24}\left(\frac{3}{4}\right)^2\left[\Gamma\left(\frac{1}{2}\right)\right]^2 \\
\boxed{\text{B}} \quad &= \frac{9}{384}\pi
\end{aligned}$$

where: in step $\boxed{\text{A}}$ we have used the recursive formula for the Gamma function; in step $\boxed{\text{B}}$ we have used the fact that

$$\Gamma\left(\frac{1}{2}\right)=\sqrt{\pi}$$

Substituting the above number into the previous expression for the integral, we obtain

$$\begin{aligned}
\int_0^\infty x^{3/2}\left(1+2x\right)^{-5}dx &= \left(\frac{1}{2}\right)^{5/2}B\left(\frac{5}{2},\frac{5}{2}\right) = \left(\frac{1}{2}\right)^{5/2}\frac{9}{384}\pi \\
= \quad \frac{1}{8}\sqrt{2}\frac{9}{384}\pi &= \frac{9}{3072}\sqrt{2}\pi = \frac{3}{1024}\sqrt{2}\pi
\end{aligned}$$

If you wish, you can check the above result using the MATLAB commands

```
syms x
f = (x^(3/2)) * ((1 + 2 * x)^ - 5)
int(f, 0, Inf)
```

Part II

Fundamentals of probability

Chapter 10

Probability

Probability is used to quantify the likelihood of things that can happen, when it is not yet known whether they will happen. Sometimes probability is also used to quantify the likelihood of things that could have happened in the past, when it is not yet known whether they actually happened.

Since we usually speak of the "probability of an event", the next section introduces a formal definition of the concept of event. We then discuss the properties that probability needs to satisfy. Finally, we discuss some possible interpretations of the concept of probability.

10.1 Sample space, sample points and events

Let Ω be a set of things that can happen[1]. We say that Ω is a **sample space**, or **space of all possible outcomes**, if it satisfies the following two properties:

1. **Mutually exclusive outcomes.** Only one of the things in Ω will happen. That is, when we learn that $\overline{\omega} \in \Omega$ has happened, then we also know that none of the things in the set $\{\omega \in \Omega : \omega \neq \overline{\omega}\}$ has happened.

2. **Exhaustive outcomes.** At least one of the things in Ω will happen.

An element $\omega \in \Omega$ is called a **sample point**, or a **possible outcome**.

When (and if) we learn that $\overline{\omega} \in \Omega$ has happened, $\overline{\omega}$ is called the **realized outcome**.

A subset $E \subseteq \Omega$ is called an **event** (you will see below[2] that not every subset of Ω is, strictly speaking, an event; however, on a first reading you can be happy with this definition).

Note that Ω itself is an event, because every set is a subset of itself, and the empty set $\emptyset$ is also an event, because it can be considered a subset of Ω.

Example 64 *Suppose that we toss a die. Six numbers, from* 1 *to* 6*, can appear face up, but we do not yet know which one of them will appear. The sample space is*

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

[1] In this lecture we are going to use the Greek letter Ω (Omega), which is often used in probability theory. Ω is upper case, while ω is lower case.

[2] P. 75.

Each of the six numbers is a sample point. The outcomes are mutually exclusive, because only one number at a time can appear face up. The outcomes are also exhaustive, because at least one of the six numbers in Ω *will appear face up, after we toss the die. Define*

$$E = \{1, 3, 5\}$$

E is an event (a subset of Ω*). In words, the event E can be described as "an odd number appears face up". Now, define*

$$F = \{6\}$$

Also F is an event (a subset of Ω*). In words, the event F can be described as "the number* 6 *appears face up".*

10.2 Probability

The **probability** of an event is a real number, attached to the event, that tells us how likely that event is. Suppose E is an event. We denote the probability of E by $\mathrm{P}(E)$.

Probability needs to satisfy the following properties:

1. **Range.** For any event E, $0 \leq \mathrm{P}(E) \leq 1$.
2. **Sure thing.** $\mathrm{P}(\Omega) = 1$.
3. **Sigma-additivity** (or countable additivity)**.** Let $\{E_1, E_2, \ldots, E_n, \ldots\}$ be a sequence[3] of events. Let all the events in the sequence be mutually exclusive, i.e., $E_i \cap E_j = \emptyset$ if $i \neq j$. Then,

$$\mathrm{P}\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} \mathrm{P}(E_n)$$

The first property is self-explanatory. It just means that the probability of an event is a real number between 0 and 1.

The second property is also intuitive. It says that with probability 1 at least one of the possible outcomes will happen.

The third property is a bit more cumbersome. It can be proved (see p. 72) that if sigma-additivity holds, then also the following holds:

$$\text{If } E \text{ and } F \text{ are two events and } E \cap F = \emptyset, \text{ then } \mathrm{P}(E \cup F) = \mathrm{P}(E) + \mathrm{P}(F) \quad (10.1)$$

This property, called **finite additivity**, while very similar to sigma-additivity, is easier to interpret. It says that if two events are disjoint, then the probability that either one or the other happens is equal to the sum of their individual probabilities.

Example 65 *Suppose that we flip a coin. The possible outcomes are either tail (T) or head (H), i.e.,*

$$\Omega = \{T, H\}$$

[3] See p. 31.

There are a total of four subsets of Ω *(events):* Ω *itself, the empty set* $\emptyset$*, the event* $\{T\}$ *and the event* $\{H\}$*. The following assignment of probabilities satisfies the properties enumerated above:*

$$\mathrm{P}(\Omega) = 1,\ \mathrm{P}(\emptyset) = 0,\ \mathrm{P}(\{T\}) = \frac{1}{2},\ \mathrm{P}(\{H\}) = \frac{1}{2}$$

All these probabilities are between 0 *and* 1*, so the range property is satisfied.* $\mathrm{P}(\Omega) = 1$*, so the sure thing property is satisfied. Also sigma-additivity is satisfied, because*

$$\begin{aligned}
\mathrm{P}(\{T\} \cup \{H\}) &= \mathrm{P}(\Omega) = 1 = \frac{1}{2} + \frac{1}{2} = \mathrm{P}(\{T\}) + \mathrm{P}(\{H\}) \\
\mathrm{P}(\{\Omega\} \cup \{\emptyset\}) &= \mathrm{P}(\Omega) = 1 = 1 + 0 = \mathrm{P}(\{\Omega\}) + \mathrm{P}(\{\emptyset\}) \\
\mathrm{P}(\{T\} \cup \{\emptyset\}) &= \mathrm{P}(T) = \frac{1}{2} = \frac{1}{2} + 0 = \mathrm{P}(\{T\}) + \mathrm{P}(\{\emptyset\}) \\
\mathrm{P}(\{H\} \cup \{\emptyset\}) &= \mathrm{P}(H) = \frac{1}{2} = \frac{1}{2} + 0 = \mathrm{P}(\{H\}) + \mathrm{P}(\{\emptyset\})
\end{aligned}$$

and the four couples (T, H)*,* $(\Omega, \emptyset)$*,* $(T, \emptyset)$*,* $(H, \emptyset)$ *are the only four possible couples of disjoint sets.*

Before ending this section, two remarks are in order. First, we have not discussed the interpretations of probability, but below you can find a brief discussion of the interpretations of probability. Second, we have been somewhat sloppy in defining events and probability, but you can find a more rigorous definition of probability below.

10.3 Properties of probability

The following subsections discuss some of the properties enjoyed by probability.

10.3.1 Probability of the empty set

The empty set has zero probability:

$$\mathrm{P}(\emptyset) = 0$$

Proof. Define a sequence of events as follows: $E_1 = \Omega$, $E_2 = \emptyset$, ..., $E_n = \emptyset$, ... The sequence is a sequence of disjoint events, because the empty set is disjoint from any other set. Then,

$$\begin{aligned}
1 &= \mathrm{P}(\Omega) \\
&= \mathrm{P}\left(\bigcup_{n=1}^{\infty} E_n\right) \\
&= \sum_{n=1}^{\infty} \mathrm{P}(E_n) \\
&= \mathrm{P}(\Omega) + \sum_{n=2}^{\infty} \mathrm{P}(\emptyset)
\end{aligned}$$

that is,

$$\mathrm{P}(\Omega) + \sum_{n=2}^{\infty} \mathrm{P}(\emptyset) = 1$$

Since $\mathrm{P}(\Omega) = 1$, we have that

$$\sum_{n=2}^{\infty} \mathrm{P}(\emptyset) = 1 - 1 = 0$$

which implies $\mathrm{P}(\emptyset) = 0$. ■

10.3.2 Additivity and sigma-additivity

A sigma-additive function is also additive (see 10.1).

Proof. Let E and F be two events and $E \cap F = \emptyset$. Define a sequence of events as follows: $E_1 = E$, $E_2 = F$, $E_3 = \emptyset$, ..., $E_n = \emptyset$, ... The sequence is a sequence of disjoint events, because the empty set is disjoint from any other set. Then,

$$\begin{aligned} \mathrm{P}(E \cup F) &= \mathrm{P}\left(\bigcup_{n=1}^{\infty} E_n\right) \\ &= \sum_{n=1}^{\infty} \mathrm{P}(E_n) \\ &= \mathrm{P}(E) + \mathrm{P}(F) + \sum_{n=3}^{\infty} \mathrm{P}(\emptyset) \\ &= \mathrm{P}(E) + \mathrm{P}(F) \end{aligned}$$

since $\mathrm{P}(\emptyset) = 0$. ■

10.3.3 Probability of the complement

Let E be an event and E^c its complement (i.e., the set of all elements of Ω that do not belong to E). Then,

$$\mathrm{P}(E^c) = 1 - \mathrm{P}(E)$$

In words, the probability that an event does not occur ($\mathrm{P}(E^c)$) is equal to one minus the probability that it occurs.

Proof. Note that

$$\Omega = E \cup E^c \tag{10.2}$$

and that E and E^c are disjoint sets. Then, using the sure thing property and finite additivity, we obtain

$$1 = \mathrm{P}(\Omega) = \mathrm{P}(E \cup E^c) = \mathrm{P}(E) + \mathrm{P}(E^c)$$

By rearranging the terms of this equality, we obtain 10.2. ■

10.3.4 Probability of a union

Let E and F be two events. We have already seen how to compute $\mathrm{P}(E \cup F)$ in the special case in which E and F are disjoint. In the more general case (E and F are not necessarily disjoint), the formula is

$$\mathrm{P}(E \cup F) = \mathrm{P}(E) + \mathrm{P}(F) - \mathrm{P}(E \cap F)$$

Proof. First, note that

$$\begin{aligned} \mathrm{P}(E) &= \mathrm{P}(E \cap \Omega) = \mathrm{P}(E \cap (F \cup F^c)) \\ &= \mathrm{P}((E \cap F) \cup (E \cap F^c)) = \mathrm{P}(E \cap F) + \mathrm{P}(E \cap F^c) \end{aligned}$$

and

$$\begin{aligned} \mathrm{P}(F) &= \mathrm{P}(F \cap \Omega) = \mathrm{P}(F \cap (E \cup E^c)) \\ &= \mathrm{P}((F \cap E) \cup (F \cap E^c)) = \mathrm{P}(F \cap E) + \mathrm{P}(F \cap E^c) \end{aligned}$$

so that

$$\begin{aligned} \mathrm{P}(E \cap F^c) &= \mathrm{P}(E) - \mathrm{P}(E \cap F) \\ \mathrm{P}(F \cap E^c) &= \mathrm{P}(F) - \mathrm{P}(E \cap F) \end{aligned}$$

Furthermore, the event $E \cup F$ can be written as follows:

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

and the three events on the right hand side are disjoint. Thus,

$$\begin{aligned} \mathrm{P}(E \cup F) &= \mathrm{P}((E \cap F) \cup (E \cap F^c) \cup (F \cap E^c)) \\ &= \mathrm{P}(E \cap F) + \mathrm{P}(E \cap F^c) + \mathrm{P}(F \cap E^c) \\ &= \mathrm{P}(E \cap F) + \mathrm{P}(E) - \mathrm{P}(E \cap F) + \mathrm{P}(F) - \mathrm{P}(E \cap F) \\ &= \mathrm{P}(E) + \mathrm{P}(F) - \mathrm{P}(E \cap F) \end{aligned}$$

■

10.3.5 Monotonicity of probability

If two events E and F are such that $E \subseteq F$, then

$$\mathrm{P}(E) \leq \mathrm{P}(F)$$

In other words, if E occurs less often than F, because F contemplates more occurrences, then the probability of E must be less than the probability of F.
Proof. This is easily proved by using additivity:

$$\begin{aligned} \mathrm{P}(F) &= \mathrm{P}((F \cap E) \cup (F \cap E^c)) \\ &= \mathrm{P}(F \cap E) + \mathrm{P}(F \cap E^c) \\ &= \mathrm{P}(E) + \mathrm{P}(F \cap E^c) \end{aligned}$$

Since, by the range property, $\mathrm{P}(F \cap E^c) \geq 0$, it follows that

$$\mathrm{P}(F) = \mathrm{P}(E) + \mathrm{P}(F \cap E^c) \geq \mathrm{P}(E)$$

■

10.4 Interpretations of probability

This subsection briefly discusses some common interpretations of probability. Although none of these interpretations is sufficient per se to clarify the meaning of probability, they all touch upon important aspects of probability.

10.4.1 Classical interpretation of probability

According to the classical definition of probability, when all the possible outcomes of an experiment are equally likely, the probability of an event is the ratio between the number of outcomes that are favorable to the event and the total number of possible outcomes. While intuitive, this definition has two main drawbacks:

1. it is circular, because it uses the concept of probability to define probability: it is based on the assumption of "equally likely" outcomes, where equally likely means "having the same probability";

2. it is limited in scope, because it does not allow to define probability when the possible outcomes are not all equally likely.

10.4.2 Frequentist interpretation of probability

According to the frequentist definition of probability, the probability of an event is the relative frequency of the event itself, observed over a large number of repetitions of the same experiment. In other words, it is the limit to which the ratio

$$\frac{\text{number of occurrences of the event}}{\text{total number of repetitions of the experiment}}$$

converges when the number of repetitions of the experiment tends to infinity. Despite its intuitive appeal, also this definition of probability has some important drawbacks:

1. it assumes that all probabilistic experiments can be repeated many times, which is false;

2. it is also somewhat circular, because it implicitly relies on a Law of Large Numbers[4], which can be derived only after having defined probability.

10.4.3 Subjectivist interpretation of probability

According to the subjectivist definition of probability, the probability of an event is related to the willingness of an individual to accept bets on that event. Suppose a lottery ticket pays off 1 dollar in case the event occurs and 0 in case the event does not occur. An individual is asked to set a price for this lottery ticket, at which she must be indifferent between being a buyer or a seller of the ticket. The subjective probability of the event is defined to be equal to the price thus set by the individual. Also this definition of probability has some drawbacks:

1. different individuals can set different prices, therefore preventing an objective assessment of probabilities;

[4] See the lecture entitled *Laws of Large Numbers* (p. 535).

2. the price an individual is willing to pay to participate in a lottery can be influenced by other factors that have nothing to do with probability; for example, an individual's betting behavior can be influenced by her preferences.

10.5 More rigorous definitions

10.5.1 A more rigorous definition of event

The definition of event given above is not entirely rigorous. Often, statisticians work with probability models where some subsets of Ω are not considered events. This happens mainly for the following two reasons:

1. sometimes, Ω is a really complicated set; to make things simpler, attention is restricted to only some subsets of Ω;

2. sometimes, it is possible to assign probabilities only to some subsets of Ω; in these cases, only the subsets to which probabilities can be assigned are considered events.

Denote by $\mathcal{F}$ the set of subsets of Ω which are considered events. $\mathcal{F}$ is called the **space of events**. In rigorous probability theory, $\mathcal{F}$ is required to be a **sigma-algebra** on Ω. $\mathcal{F}$ is a sigma-algebra on Ω if it is a set of subsets of Ω satisfying the following three properties:

1. **Whole set.** $\Omega \in \mathcal{F}$.

2. **Closure under complementation.** If $E \in \mathcal{F}$ then also $E^c \in \mathcal{F}$ (E^c, the complement of E with respect to Ω, is the set of all elements of Ω that do not belong to E).

3. **Closure under countable unions.** If E_1, E_2, ..., E_n,... are a sequence of subsets of Ω belonging to $\mathcal{F}$, then

$$\left(\bigcup_{n=1}^{\infty} E_n \right) \in \mathcal{F}$$

Why is a space of events required to satisfy these properties? Besides a number of mathematical reasons, it seems pretty intuitive that they must be satisfied. Property a) means that the space of events must include the event "*something will happen*", quite a trivial requirement! Property b) means that if "*one of the things in the set E will happen*" is considered an event, then also "*none of the things in the set E will happen*" is considered an event. This is quite natural: if you are considering the possibility that an event will happen, then, by necessity, you must also be simultaneously considering the possibility that the same event will not happen. Property c) is a bit more complex. However, the following property, implied by c), is probably easier to interpret:

$$\text{If } E \in \mathcal{F} \text{ and } F \in \mathcal{F}, \text{ then } (E \cup F) \in \mathcal{F}$$

It means that if *"one of the things in E will happen"* and *"one of the things in F will happen"* are considered two events, then also "*one of the things in E or one of the things in F will happen*" must be considered an event. This simply means

that if you are able to separately assess the possibility of two events E and F happening, then, of course, you must be able to assess the possibility of one or the other happening. Property c) simply extends this intuitive property to countable[5] collection of events: the extension is needed for mathematical reasons, to derive certain continuity properties of probability measures.

10.5.2 A more rigorous definition of probability

The definition of probability given above was not entirely rigorous. Now that we have defined sigma-algebras and spaces of events, we can make it completely rigorous. Let Ω be a sample space. Let $\mathcal{F}$ be a sigma-algebra on Ω (a space of events). A function $\mathrm{P} : \mathcal{F} \to [0,1]$ is a **probability measure** if and only if it satisfies the following two properties:

1. **Sure thing.** $\mathrm{P}(\Omega) = 1$.
2. **Sigma-additivity.** Let $\{E_1, E_2, \ldots, E_n, \ldots\}$ be any sequence of elements of $\mathcal{F}$ such that $i \neq j$ implies $E_i \cap E_j = \emptyset$. Then,

$$\mathrm{P}\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} \mathrm{P}(E_n)$$

Nothing new has been added to the definition given above. This definition just clarifies that a probability measure is a function defined on a sigma-algebra of events. Hence, it is not possible to properly speak of probability for subsets of Ω that do not belong to the sigma-algebra.

A triple $(\Omega, \mathcal{F}, \mathrm{P})$ is called a **probability space** and the sets belonging to the sigma-algebra $\mathcal{F}$ are called **measurable sets**.

10.6 Solved exercises

This exercise set contains some solved exercises on probability and events.

Exercise 1

A ball is drawn at random from an urn containing colored balls. The balls can be either red or blue (no other colors are possible). The probability of drawing a blue ball is 1/3. What is the probability of drawing a red ball?

Solution

The sample space Ω can be represented as the union of two disjoint events E and F:

$$\Omega = E \cup F$$

where the event E can be described as "a red ball is drawn" and the event F can be described as "a blue ball is drawn". Note that E is the complement of F:

$$E = F^c$$

[5] See p. 32.

We know $\mathrm{P}(F)$, the probability of drawing a a blue ball:

$$\mathrm{P}(F) = \frac{1}{3}$$

We need to find $\mathrm{P}(E)$, the probability of drawing a a red ball. Using the formula for the probability of a complement:

$$\mathrm{P}(E) = \mathrm{P}(F^c) = 1 - \mathrm{P}(F) = 1 - \frac{1}{3} = \frac{2}{3}$$

Exercise 2

Consider a sample space Ω comprising three possible outcomes:

$$\Omega = \{\omega_1, \omega_2, \omega_3\}$$

Suppose the probabilities assigned to the three possible outcomes are

$$\mathrm{P}(\{\omega_1\}) = 1/4 \quad \mathrm{P}(\{\omega_2\}) = 1/4 \quad \mathrm{P}(\{\omega_3\}) = 1/2$$

Can you find an event whose probability is 3/4?

Solution

There are two events whose probability is 3/4.

The first one is

$$E = \{\omega_1, \omega_3\}$$

By using the formula for the probability of a union of disjoint events, we get

$$\begin{aligned} \mathrm{P}(E) &= \mathrm{P}(\{\omega_1\} \cup \{\omega_3\}) = \mathrm{P}(\{\omega_1\}) + \mathrm{P}(\{\omega_3\}) \\ &= \frac{1}{4} + \frac{1}{2} = \frac{3}{4} \end{aligned}$$

The second one is

$$E = \{\omega_2, \omega_3\}$$

By using the formula for the probability of a union of disjoint events, we obtain:

$$\begin{aligned} \mathrm{P}(E) &= \mathrm{P}(\{\omega_2\} \cup \{\omega_3\}) = \mathrm{P}(\{\omega_2\}) + \mathrm{P}(\{\omega_3\}) \\ &= \frac{1}{4} + \frac{1}{2} = \frac{3}{4} \end{aligned}$$

Exercise 3

Consider a sample space Ω comprising four possible outcomes:

$$\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$$

Consider the three events E, F and G defined as follows:

$$\begin{aligned} E &= \{\omega_1\} \\ F &= \{\omega_1, \omega_2\} \\ G &= \{\omega_1, \omega_2, \omega_3\} \end{aligned}$$

Suppose their probabilities are

$$\mathrm{P}(E) = \tfrac{1}{10} \quad \mathrm{P}(F) = \tfrac{5}{10} \quad \mathrm{P}(G) = \tfrac{7}{10}$$

Now, consider a fourth event H defined as follows:

$$H = \{\omega_2, \omega_4\}$$

Find $\mathrm{P}(H)$.

Solution

First note that, by additivity,

$$\mathrm{P}(H) = \mathrm{P}(\{\omega_2\} \cup \{\omega_4\}) = \mathrm{P}(\{\omega_2\}) + \mathrm{P}(\{\omega_4\})$$

Therefore, in order to compute $\mathrm{P}(H)$, we need to compute $\mathrm{P}(\{\omega_2\})$ and $\mathrm{P}(\{\omega_4\})$. $\mathrm{P}(\{\omega_2\})$ is found using additivity on F:

$$\begin{aligned} \frac{5}{10} &= \mathrm{P}(F) = \mathrm{P}(\{\omega_1\} \cup \{\omega_2\}) = \mathrm{P}(\{\omega_1\}) + \mathrm{P}(\{\omega_2\}) \\ &= \mathrm{P}(E) + \mathrm{P}(\{\omega_2\}) = \frac{1}{10} + \mathrm{P}(\{\omega_2\}) \end{aligned}$$

so that

$$\mathrm{P}(\{\omega_2\}) = \frac{5}{10} - \frac{1}{10} = \frac{4}{10}$$

$\mathrm{P}(\{\omega_4\})$ is found using the fact that one minus the probability of an event is equal to the probability of its complement and the fact that $\{\omega_4\} = G^c$:

$$\mathrm{P}(\{\omega_4\}) = \mathrm{P}(G^c) = 1 - \mathrm{P}(G) = 1 - \frac{7}{10} = \frac{3}{10}$$

As a consequence,

$$\mathrm{P}(H) = \mathrm{P}(\{\omega_2\}) + \mathrm{P}(\{\omega_4\}) = \frac{4}{10} + \frac{3}{10} = \frac{7}{10}$$

Chapter 11

Zero-probability events

The notion of a zero-probability event plays a special role in probability theory and statistics, because it underpins the important concepts of almost sure property and almost sure event. In this lecture, we define zero-probability events and discuss some counterintuitive aspects of their apparently simple definition, in particular the fact that a zero-probability event is not an event that never happens: there are common probabilistic settings where zero-probability events do happen all the time! After discussing this matter, we introduce the concepts of almost sure property and almost sure event.

11.1 Definition and discussion

Tautologically, zero-probability events are events whose probability is equal to zero.

Definition 66 *Let E be an event*[1] *and denote its probability by $\mathrm{P}(E)$. We say that E is a* ***zero-probability event*** *if and only if*

$$\mathrm{P}(E) = 0$$

Despite the simplicity of this definition, there are some features of zero-probability events that might seem paradoxical. We illustrate these features with the following example.

Example 67 *Consider a probabilistic experiment whose set of possible outcomes, called sample space and denoted by Ω, is the unit interval*

$$\Omega = [0, 1]$$

It is possible to assign probabilities in such a way that each sub-interval has probability equal to its length:

$$[a, b] \subseteq [0, 1] \Rightarrow \mathrm{P}([a, b]) = b - a$$

The proof that such an assignment of probabilities can be consistently performed is beyond the scope of this example, but you can find it in any elementary measure

[1] See the lecture entitled *Probability* (p. 69) for a definition of sample space and event.

theory book (e.g., Williams[2] - 1991). As a direct consequence of this assignment, all the possible outcomes $\omega \in \Omega$ have zero probability:

$$\forall \omega \in \Omega,\ \mathrm{P}(\{\omega\}) = \mathrm{P}([\omega, \omega]) = \omega - \omega = 0$$

Stated differently, every possible outcome is a zero-probability event. This might seem counterintuitive. In everyday language, a zero-probability event is an event that never happens. However, this example illustrates that a zero-probability event can indeed happen. Since the sample space provides an exhaustive description of the possible outcomes, one and only one of the sample points[3] $\omega \in \Omega$ will be the realized outcome[4]. But we have just demonstrated that all the sample points are zero-probability events; as a consequence, the realized outcome can only be a zero-probability event. Another apparently paradoxical aspect of this probability model is that the sample space Ω can be obtained as the union of disjoint zero-probability events:

$$\Omega = \bigcup_{\omega \in \Omega} \{\omega\}$$

where each $\omega \in \Omega$ is a zero-probability event and all events in the union are disjoint. If we forgot that the additivity property of probability applies only to countable collections of subsets, we would mistakenly deduce that

$$\mathrm{P}(\Omega) = \mathrm{P}\left(\bigcup_{\omega \in \Omega} \{\omega\}\right) = \sum_{\omega \in \Omega} \mathrm{P}(\omega) = 0$$

and we would come to a contradiction: $\mathrm{P}(\Omega) = 0$, when, by the properties of probability[5], it should be $\mathrm{P}(\Omega) = 1$. Of course, the fallacy in such an argument is that Ω is not a countable set, and hence the additivity property cannot be used.

The main lesson to be taken from this example is that a zero-probability event is not an event that never happens: in some probability models, where the sample space is not countable, zero-probability events do happen all the time!

11.2 Almost sure and almost surely

Zero-probability events are of paramount importance in probability and statistics. Often, we want to prove that some property is almost always satisfied, or something happens almost always. "Almost always" means that the property is satisfied for all sample points, except possibly for a negligible set of sample points. The concept of zero-probability event is used to determine which sets are negligible: if a set is included in a zero-probability event, then it is negligible.

Definition 68 *Let Φ be some property that a sample point $\omega \in \Omega$ can either satisfy or not satisfy. Let F be the set of all sample points that satisfy the property:*

$$F = \{\omega \in \Omega : \omega \text{ satisfies property } \Phi\}$$

[2] Williams, D. (1991) *Probability with martingales*, Cambridge University Press.
[3] See p. 69.
[4] See p. 69.
[5] See p. 70.

Denote its complement, that is, the set of all points not satisfying property Φ, by F^c. Property Φ is said to be ***almost sure*** *if there exists a zero-probability event E such that*[6] $F^c \subseteq E$.

An almost sure property is said to hold **almost surely** (often abbreviated as **a.s.**). Sometimes, an almost sure property is also said to hold **with probability one** (abbreviated as **w.p.1**).

11.3 Almost sure events

Remember[7] that some subsets of the sample space may not be considered events. The above definition of almost sure property allows us to consider also sets F that are not, strictly speaking, events. However, in the case in which F is an event, F is called an **almost sure event** and we say that F happens almost surely. Furthermore, since there exists an event E such that $F^c \subseteq E$ and $\mathrm{P}(E) = 0$, we can apply the monotonicity of probability[8]:

$$F^c \subseteq E \Rightarrow \mathrm{P}(F^c) \leq \mathrm{P}(E)$$

which in turn implies $\mathrm{P}(F^c) = 0$. Finally, recalling the formula for the probability of a complement[9], we obtain

$$\mathrm{P}(F) = 1 - \mathrm{P}(F^c) = 1 - 0 = 1$$

Thus, an almost sure event is an event that happens with probability 1.

Example 69 *Consider the sample space $\Omega = [0, 1]$ and the assignment of probabilities introduced in the previous example:*

$$[a, b] \subseteq [0, 1] \Rightarrow \mathrm{P}([a, b]) = b - a$$

We want to prove that the event

$$E = \{\omega \in \Omega : \omega \text{ is a rational number}\}$$

is a zero-probability event. Since the set of rational numbers is countable[10] *and E is a subset of the set of rational numbers, E is countable. This implies that the elements of E can be arranged into a sequence:*

$$E = \{\omega_1, \ldots, \omega_n, \ldots\}$$

Furthermore, E can be written as a countable union:

$$E = \bigcup_{n=1}^{\infty} \{\omega_n\}$$

[6] In other words, the set F^c of all points that do not satisfy the property is included in a zero-probability event.

[7] See the lecture entitled *Probability* (p. 69).

[8] See p. 73

[9] See p. 72.

[10] See p. 32.

Applying the countable additivity property of probability[11]*, we obtain*

$$\mathrm{P}(E) = \mathrm{P}\left(\bigcup_{n=1}^{\infty} \{\omega_n\}\right) = \sum_{n=1}^{\infty} \mathrm{P}(\{\omega_n\}) = 0$$

since $\mathrm{P}(\{\omega_n\}) = 0$ *for every* n*. Therefore,* E *is a zero-probability event. This might seem surprising: in this probability model there are also zero-probability events comprising infinitely many sample points! It can also easily be proved that the event*

$$F = \{\omega \in \Omega : \omega \text{ is an irrational number}\}$$

is an almost sure event. In fact

$$F = E^c$$

and applying the formula for the probability of a complement, we get

$$\mathrm{P}(F) = \mathrm{P}(E^c) = 1 - \mathrm{P}(E) = 1 - 0 = 1$$

11.4 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Let E and F be two events. Let E^c be a zero-probability event and $\mathrm{P}(F) = \frac{2}{3}$. Compute $\mathrm{P}(E \cup F)$.

Solution

E^c is a zero-probability event, which means that

$$\mathrm{P}(E^c) = 0$$

Furthermore, using the formula for the probability of a complement, we obtain

$$\mathrm{P}(E) = 1 - \mathrm{P}(E^c) = 1 - 0 = 1$$

Since $(E \cup F) \supseteq E$, by monotonicity we obtain

$$\mathrm{P}(E \cup F) \geq \mathrm{P}(E)$$

Since $\mathrm{P}(E) = 1$ and probabilities cannot be greater than 1, this implies

$$\mathrm{P}(E \cup F) = 1$$

Exercise 2

Let E and F be two events. Let E^c be a zero-probability event and $\mathrm{P}(F) = \frac{1}{2}$. Compute $\mathrm{P}(E \cap F)$.

[11] See p. 72.

Solution

E^c is a zero-probability event, which means that

$$\mathrm{P}\left(E^c\right) = 0$$

Furthermore, using the formula for the probability of a complement, we obtain

$$\mathrm{P}\left(E\right) = 1 - \mathrm{P}\left(E^c\right) = 1 - 0 = 1$$

It is also true that

$$\begin{aligned} \mathrm{P}\left(E \cap F\right) &= \mathrm{P}\left(E\right) + \mathrm{P}\left(F\right) - \mathrm{P}\left(E \cup F\right) \\ &= 1 + \frac{1}{2} - \mathrm{P}\left(E \cup F\right) = \frac{3}{2} - \mathrm{P}\left(E \cup F\right) \end{aligned}$$

Since $(E \cup F) \supseteq E$, by monotonicity, we obtain

$$\mathrm{P}\left(E \cup F\right) \geq \mathrm{P}\left(E\right)$$

Since $\mathrm{P}\left(E\right) = 1$ and probabilities cannot be greater than 1, this implies

$$\mathrm{P}\left(E \cup F\right) = 1$$

Thus, putting pieces together, we get

$$\mathrm{P}\left(E \cap F\right) = \frac{3}{2} - \mathrm{P}\left(E \cup F\right) = \frac{3}{2} - 1 = \frac{1}{2}$$

Chapter 12

Conditional probability

This lecture introduces the concept of conditional probability. A more sophisticated treatment of conditional probability can be found in the lecture entitled *Conditional probability as a random variable* (p. 201).

Before reading this lecture, make sure you are familiar with the concepts of sample space, sample point, event, possible outcome, realized outcome and probability (see the lecture entitled *Probability* - p. 69).

12.1 Introduction

Let Ω be a sample space and let $\mathrm{P}(E)$ denote the probability assigned to the events $E \subseteq \Omega$. Suppose that, after assigning probabilites $\mathrm{P}(E)$ to the events in Ω, we receive new information about the things that will happen (the possible outcomes). In particular, suppose that we are told that the realized outcome will belong to a set $I \subseteq \Omega$. How should we revise the probabilities assigned to the events in Ω, to properly take the new information into account?

Denote by $\mathrm{P}(E\,|\,I)$ the revised probability assigned to an event $E \subseteq \Omega$ after learning that the realized outcome will be an element of I. $\mathrm{P}(E\,|\,I)$ is called the **conditional probability of E given I**.

Despite being an intuitive concept, conditional probability is quite difficult to define in a rigorous way. We take a gradual approach in this lecture. We first discuss conditional probability for the very special case in which all the sample points are equally likely. We then give a more general definition. Finally, we refer the reader to other lectures where conditional probability is defined in even more abstract ways.

12.2 The case of equally likely sample points

Suppose a sample space Ω has a finite number n of sample points $\omega_1, \ldots, \omega_n$:

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

Suppose also that each sample point is assigned the same probability:

$$\mathrm{P}(\{\omega_1\}) = \ldots = \mathrm{P}(\{\omega_n\}) = \frac{1}{n}$$

In such a simple space, the probability of a generic event E is obtained as

$$\mathrm{P}(E) = \frac{\mathrm{card}(E)}{\mathrm{card}(\Omega)}$$

where card denotes the cardinality of a set, i.e. the number of its elements. In other words, the probability of an event E is obtained in two steps:

1. counting the number of "cases that are favorable to the event E", i.e. the number of elements ω_i belonging to E;
2. dividing the number thus obtained by the number of "all possible cases", i.e. the number of elements ω_i belonging to Ω.

For example, if $E = \{\omega_1, \omega_2\}$ then

$$\mathrm{P}(E) = \frac{\mathrm{card}(\{\omega_1, \omega_2\})}{\mathrm{card}(\Omega)} = \frac{2}{n}$$

When we learn that the realized outcome will belong to a set $I \subseteq \Omega$, we still apply the rule

$$\text{probability of an event} = \frac{\text{number of cases that are favorable to the event}}{\text{number of all possible cases}}$$

However, the number of all possible cases is now equal to the number of elements of I, because only the outcomes beloning to I are still possible. Furthermore, the number of favorable cases is now equal to the number of elements of $E \cap I$, because the outcomes in $E \cap I^c$ are no longer possible. As a consequence:

$$\mathrm{P}(E \,|\, I) = \frac{\mathrm{card}(E \cap I)}{\mathrm{card}(I)}$$

Dividing numerator and denominator by $\mathrm{card}(\Omega)$ one obtains

$$\mathrm{P}(E \,|\, I) = \frac{\mathrm{card}(E \cap I)/\mathrm{card}(\Omega)}{\mathrm{card}(I)/\mathrm{card}(\Omega)} = \frac{\mathrm{P}(E \cap I)}{\mathrm{P}(I)}$$

Therefore, when all sample points are equally likely, conditional probabilities are computed as

$$\mathrm{P}(E \,|\, I) = \frac{\mathrm{P}(E \cap I)}{\mathrm{P}(I)}$$

Example 70 *Suppose that we toss a die. Six numbers (from* 1 *to* 6*) can appear face up, but we do not yet know which one of them will appear. The sample space is*

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

Each of the six numbers is a sample point and is assigned probability $1/6$*. Define the event* E *as follows:*

$$E = \{1, 3, 5\}$$

where the event E *could be described as "an odd number appears face up". Define the event* I *as follows:*

$$I = \{4, 5, 6\}$$

where the event I could be described as "a number greater than 3 appears face up". The probability of I is

$$\mathrm{P}(I) = \mathrm{P}(\{4\}) + \mathrm{P}(\{5\}) + \mathrm{P}(\{6\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$$

Suppose we are told that the realized outcome will belong to I. How do we have to revise our assessment of the probability of the event E, according to the rules of conditional probability? First of all, we need to compute the probability of the event $E \cap I$:

$$\mathrm{P}(E \cap I) = \mathrm{P}(\{5\}) = \frac{1}{6}$$

Then, the conditional probability of E given I is

$$\mathrm{P}(E\,|\,I) = \frac{\mathrm{P}(E \cap I)}{\mathrm{P}(I)} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{2}{6} = \frac{1}{3}$$

In the next section, we will show that the **conditional probability formula**

$$\mathrm{P}(E\,|\,I) = \frac{\mathrm{P}(E \cap I)}{\mathrm{P}(I)}$$

is valid also for more general cases (i.e. when the sample points are not all equally likely). However, this formula already allows us to understand why defining conditional probability is a challenging task. In the conditional probability formula, a division by $\mathrm{P}(I)$ is performed. This division is impossible when I is a zero-probability event[1]. If we want to be able to define $\mathrm{P}(E\,|\,I)$ also when $\mathrm{P}(I) = 0$, then we need to give a more complicated definition of conditional probability. We will return to this point later.

12.3 A more general approach

In this section we give a more general definition of conditional probability, by taking an axiomatic approach. First, we list the properties that we would like conditional probability to satisfy. Then, we prove that the conditional probability formula introduced above satisfies these properties. The discussion of the case in which the conditional probability formula cannot be used because $\mathrm{P}(I) = 0$ is postponed to the next section.

The conditional probability $\mathrm{P}(E\,|\,I)$ is required to satisfy the following properties:

1. **Probability measure.** $\mathrm{P}(E\,|\,I)$ has to satisfy all the properties of a probability measure[2].
2. **Sure thing.** $\mathrm{P}(I\,|\,I) = 1$.
3. **Impossible events.** If $E \subseteq I^c$ [3], then $\mathrm{P}(E\,|\,I) = 0$.

[1] I.e. $\mathrm{P}(I) = 0$; see p. 79.

[2] See p. 70.

[3] I^c, the complement of I with respect to Ω, is the set of all elements of Ω that do not belong to I.

4. **Constant likelihood ratios on I.** If $E \subseteq I$, $F \subseteq I$ and $\mathrm{P}(E) > 0$, then:

$$\frac{\mathrm{P}(F\,|I)}{\mathrm{P}(E\,|I)} = \frac{\mathrm{P}(F)}{\mathrm{P}(E)}$$

These properties are very intutitve:

1. **Probability measure.** This property requires that also conditional probability measures satisfy the fundamental properties that any other probability measure needs to satisfy.
2. **Sure thing.** This property says that the probability of a sure thing must be 1: since we know that only things belonging to the set I can happen, then the probability of I must be 1.
3. **Impossible events.** This property says that the probability of an impossible thing must be 0: since we know that things not belonging to the set I will not happen, then the probability of the events that are disjoint from I must be 0.
4. **Constant likelihood ratios on I.** This property is a bit more complex: it says that if $F \subseteq I$ is - say - two times more likely than $E \subseteq I$ before receiving the information I, then F remains two times more likely than E, also after reiceiving the information, because all the things in E and F remain possible (can still happen) and, hence, there is no reason to expect that the ratio of their likelihoods changes.

It is possible to prove that:

Proposition 71 *Whenever* $\mathrm{P}(I) > 0$, $\mathrm{P}(E\,|I)$ *satisfies the four above properties if and only if*

$$\mathrm{P}(E\,|I) = \frac{\mathrm{P}(E \cap I)}{\mathrm{P}(I)}$$

Proof. We first show that

$$\mathrm{P}(E\,|I) = \frac{\mathrm{P}(E \cap I)}{\mathrm{P}(I)}$$

satisfies the four properties whenever $\mathrm{P}(I) > 0$. As far as property 1) is concerned, we have to check that the three requirements for a probabilitiy measure are satisfied. The first requirement for a probability measure is that $0 \leq \mathrm{P}(E\,|I) \leq 1$. Since $(E \cap I) \subseteq I$, by the monotonicity of probability[4] we have that:

$$\mathrm{P}(E \cap I) \leq \mathrm{P}(I)$$

Hence:

$$\frac{\mathrm{P}(E \cap I)}{\mathrm{P}(I)} \leq 1$$

Furthermore, since $\mathrm{P}(E \cap I) \geq 0$ and $\mathrm{P}(I) \geq 0$, also

$$\frac{\mathrm{P}(E \cap I)}{\mathrm{P}(I)} \geq 0$$

[4] See p. 73.

The second requirement for a probability measure is that $\mathrm{P}(\Omega\,|I) = 1$. This is satisfied because

$$\mathrm{P}(\Omega\,|I) = \frac{\mathrm{P}(\Omega \cap I)}{\mathrm{P}(I)} = \frac{\mathrm{P}(I)}{\mathrm{P}(I)} = 1$$

The third requirement for a probability measure is that for any sequence of disjoint sets $\{E_1, E_2, \ldots, E_n, \ldots\}$ the following holds:

$$\mathrm{P}\left(\bigcup_{n=1}^{\infty} E_n\,|I\right) = \sum_{n=1}^{\infty} \mathrm{P}(E_n\,|I)$$

But

$$\begin{aligned}
\mathrm{P}\left(\bigcup_{n=1}^{\infty} E_n\,|I\right) &= \frac{\mathrm{P}\left(\left(\bigcup_{n=1}^{\infty} E_n\right) \cap I\right)}{\mathrm{P}(I)} \\
&= \frac{\mathrm{P}\left(\bigcup_{n=1}^{\infty} (E_n \cap I)\right)}{\mathrm{P}(I)} \\
&= \frac{\sum_{n=1}^{\infty} \mathrm{P}(E_n \cap I)}{\mathrm{P}(I)} \\
&= \sum_{n=1}^{\infty} \frac{\mathrm{P}(E_n \cap I)}{\mathrm{P}(I)} \\
&= \sum_{n=1}^{\infty} \mathrm{P}(E_n\,|I)
\end{aligned}$$

so that also the third requirement is satisfied. Property 2) is trivially satisfied:

$$\mathrm{P}(I\,|I) = \frac{\mathrm{P}(I \cap I)}{\mathrm{P}(I)} = \frac{\mathrm{P}(I)}{\mathrm{P}(I)} = 1$$

Property 3) is verified because, if $E \subseteq I^c$, then

$$\mathrm{P}(E\,|I) = \frac{\mathrm{P}(E \cap I)}{\mathrm{P}(I)} = \frac{\mathrm{P}(\emptyset)}{\mathrm{P}(I)} = 0$$

Property 4) is verified because, if $E \subseteq I$, $F \subseteq I$ and $\mathrm{P}(E) > 0$, then

$$\begin{aligned}
\frac{\mathrm{P}(F\,|I)}{\mathrm{P}(E\,|I)} &= \frac{\mathrm{P}(F \cap I)}{\mathrm{P}(I)} \frac{\mathrm{P}(I)}{\mathrm{P}(E \cap I)} \\
&= \frac{\mathrm{P}(F \cap I)}{\mathrm{P}(E \cap I)} = \frac{\mathrm{P}(F)}{\mathrm{P}(E)}
\end{aligned}$$

So, the "if" part has been proved. Now we prove the "only if" part. We prove it by contradiction. Suppose there exist another conditional probability $\overline{\mathrm{P}}$ that satisfies the four properties. Then, there exists an event E, such that

$$\overline{\mathrm{P}}(E\,|I) \neq \mathrm{P}(E\,|I)$$

It can not be that $E \subseteq I$, otherwise we would have

$$\frac{\overline{\mathrm{P}}(E\,|I)}{\overline{\mathrm{P}}(I\,|I)} = \frac{\overline{\mathrm{P}}(E\,|I)}{1} \neq \frac{\mathrm{P}(E\,|I)}{1} = \frac{\mathrm{P}(E \cap I)}{\mathrm{P}(I)} = \frac{\mathrm{P}(E)}{\mathrm{P}(I)}$$

which would be a contradiction, since if $\overline{\mathrm{P}}$ was a conditional probability it would satisfy

$$\frac{\overline{\mathrm{P}}\left(E\left|I\right.\right)}{\overline{\mathrm{P}}\left(I\left|I\right.\right)} = \frac{\mathrm{P}\left(E\right)}{\mathrm{P}\left(I\right)}$$

If E is not a subset of I then $\overline{\mathrm{P}}\left(E\left|I\right.\right) \neq \mathrm{P}\left(E\left|I\right.\right)$ implies also

$$\overline{\mathrm{P}}\left(E \cap I\left|I\right.\right) \neq \mathrm{P}\left(E \cap I\left|I\right.\right)$$

because

$$\begin{aligned}\overline{\mathrm{P}}\left(E\left|I\right.\right) &= \overline{\mathrm{P}}\left(\left(E \cap I\right) \cup \left(E \cap I^{c}\right)\left|I\right.\right) \\ &= \overline{\mathrm{P}}\left(E \cap I\left|I\right.\right) + \overline{\mathrm{P}}\left(E \cap I^{c}\left|I\right.\right) \\ &= \overline{\mathrm{P}}\left(E \cap I\left|I\right.\right)\end{aligned}$$

and

$$\begin{aligned}\mathrm{P}\left(E\left|I\right.\right) &= \mathrm{P}\left(\left(E \cap I\right) \cup \left(E \cap I^{c}\right)\left|I\right.\right) \\ &= \mathrm{P}\left(E \cap I\left|I\right.\right) + \mathrm{P}\left(E \cap I^{c}\left|I\right.\right) \\ &= \mathrm{P}\left(E \cap I\left|I\right.\right)\end{aligned}$$

but this would also lead to a contradiction, because $\left(E \cap I\right) \subseteq I$. ■

12.4 Tackling division by zero

In the previous section we have generalized the concept of conditional probability. However, we have not been able to define the conditional probability $\mathrm{P}\left(E\left|I\right.\right)$ for the case in which $\mathrm{P}\left(I\right) = 0$. This case is discussed in the lectures entitled *Conditional probability as a random variable* (p. 201) and *Conditional probability distributions* (p. 209).

12.5 More details

12.5.1 The law of total probability

Let $I_1, \ldots, I_n$ be n events having the following characteristics:

1. they are mutually disjoint: $I_j \cap I_k = \emptyset$ whenever $j \neq k$;
2. they cover all the sample space:
$$\Omega = \bigcup_{j=1}^{n} I_j$$
3. they have strictly positive probability: $\mathrm{P}\left(I_j\right) > 0$ for any j.

$I_1, \ldots, I_n$ is a **partition** of Ω.
The **law of total probability** states that, for any event E, the following holds:

$$\mathrm{P}\left(E\right) = \mathrm{P}\left(E\left|I_1\right.\right)\mathrm{P}\left(I_1\right) + \ldots + \mathrm{P}\left(E\left|I_n\right.\right)\mathrm{P}\left(I_n\right)$$

which can, of course, also be written as

$$\mathrm{P}(E) = \sum_{j=1}^{n} \mathrm{P}(E\,|I_j)\,\mathrm{P}(I_j)$$

Proof. The law of total probability is proved as follows:

$$\begin{aligned}
\mathrm{P}(E) &= \mathrm{P}(E \cap \Omega) \\
&= \mathrm{P}\left(E \cap \left(\bigcup_{j=1}^{n} I_j\right)\right) \\
&= \mathrm{P}\left(\bigcup_{j=1}^{n} (E \cap I_j)\right) \\
\boxed{\mathrm{A}} &= \sum_{j=1}^{n} \mathrm{P}(E \cap I_j) \\
\boxed{\mathrm{B}} &= \sum_{j=1}^{n} \mathrm{P}(E\,|I_j)\,\mathrm{P}(I_j)
\end{aligned}$$

where: in step $\boxed{\mathrm{A}}$ we have used the additivity of probability; in step $\boxed{\mathrm{B}}$ we have used the conditional probability formula. ■

12.6 Solved exercises

Some solved exercises on conditional probability can be found below.

Exercise 1

Consider a sample space Ω comprising three possible outcomes ω_1, ω_2, ω_3:

$$\Omega = \{\omega_1, \omega_2, \omega_3\}$$

Suppose the three possible outcomes are assigned the following probabilities:

$$\begin{aligned}
\mathrm{P}(\omega_1) &= \frac{1}{5} \\
\mathrm{P}(\omega_2) &= \frac{2}{5} \\
\mathrm{P}(\omega_3) &= \frac{2}{5}
\end{aligned}$$

Define the events

$$\begin{aligned}
E &= \{\omega_1, \omega_2\} \\
F &= \{\omega_1, \omega_3\}
\end{aligned}$$

and denote by E^c the complement of E.
Compute $\mathrm{P}(F\,|E^c)$, the conditional probability of F given E^c.

Solution

We need to use the conditional probability formula

$$\mathrm{P}\left(F\,|E^c\right) = \frac{\mathrm{P}\left(F \cap E^c\right)}{\mathrm{P}\left(E^c\right)}$$

The numerator is

$$\mathrm{P}\left(F \cap E^c\right) = \mathrm{P}\left(\{\omega_1, \omega_3\} \cap \{\omega_3\}\right) = \mathrm{P}\left(\{\omega_3\}\right) = \frac{2}{5}$$

and the denominator is

$$\mathrm{P}\left(E^c\right) = \mathrm{P}\left(\{\omega_3\}\right) = \frac{2}{5}$$

As a consequence:

$$\mathrm{P}\left(F\,|E^c\right) = \frac{\mathrm{P}\left(F \cap E^c\right)}{\mathrm{P}\left(E^c\right)} = \frac{2/5}{2/5} = 1$$

Exercise 2

Consider a sample space Ω comprising four possible outcomes ω_1, ω_2, ω_3, ω_4:

$$\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$$

Suppose the four possible outcomes are assigned the following probabilities:

$$\begin{aligned} \mathrm{P}\left(\omega_1\right) &= \frac{1}{10} \\ \mathrm{P}\left(\omega_2\right) &= \frac{4}{10} \\ \mathrm{P}\left(\omega_3\right) &= \frac{3}{10} \\ \mathrm{P}\left(\omega_4\right) &= \frac{2}{10} \end{aligned}$$

Define two events

$$\begin{aligned} E &= \{\omega_1, \omega_2\} \\ F &= \{\omega_2, \omega_3\} \end{aligned}$$

Compute $\mathrm{P}\left(E\,|F\right)$, the conditional probability of E given F.

Solution

We need to use the formula

$$\mathrm{P}\left(E\,|F\right) = \frac{\mathrm{P}\left(E \cap F\right)}{\mathrm{P}\left(F\right)}$$

But

$$\mathrm{P}\left(E \cap F\right) = \mathrm{P}\left(\{\omega_2\}\right) = \frac{4}{10}$$

while, using additivity:

$$\mathrm{P}\left(F\right) = \mathrm{P}\left(\{\omega_2, \omega_3\}\right) = \mathrm{P}\left(\{\omega_2\}\right) + \mathrm{P}\left(\{\omega_3\}\right) = \frac{4}{10} + \frac{3}{10} = \frac{7}{10}$$

Therefore:

$$\mathrm{P}\left(E\,|F\right) = \frac{\mathrm{P}\left(E \cap F\right)}{\mathrm{P}\left(F\right)} = \frac{4/10}{7/10} = \frac{4}{7}$$

Exercise 3

The Census Bureau has estimated the following survival probabilities for men:

1. probability that a man lives at least 70 years: 80%;
2. probability that a man lives at least 80 years: 50%.

What is the conditional probability that a man lives at least 80 years given that he has just celebrated his 70th birthday?

Solution

Given an hypothetical sample space Ω, define the two events:

$$\begin{aligned} E &= \{\omega \in \Omega : \text{ man lives at least 70 years}\} \\ F &= \{\omega \in \Omega : \text{ man lives at least 80 years}\} \end{aligned}$$

We need to find the following conditional probability

$$\mathrm{P}(F \,|E\,) = \frac{\mathrm{P}(F \cap E)}{\mathrm{P}(E)}$$

The denominator is known:

$$\mathrm{P}(E) = 80\% = \frac{4}{5}$$

As far as the numerator is concerned, note that $F \subseteq E$ (if you live at least 80 years then you also live at least 70 years). But $F \subseteq E$ implies

$$F \cap E = F$$

Therefore:

$$\mathrm{P}(F \cap E) = \mathrm{P}(F) = 50\% = \frac{1}{2}$$

Thus:

$$\mathrm{P}(F \,|E\,) = \frac{\mathrm{P}(F \cap E)}{\mathrm{P}(E)} = \frac{1/2}{4/5} = \frac{5}{8}$$

Chapter 13

Bayes' rule

This lecture introduces Bayes' rule. Before reading this lecture, make sure you are familiar with the concept of conditional probability (p. 85).

13.1 Statement of Bayes' rule

Bayes' rule, named after the English mathematician Thomas Bayes, is a rule for computing conditional probabilities.

Let A and B be two events. Denote their probabilities by $\mathrm{P}(A)$ and $\mathrm{P}(B)$ and suppose that both $\mathrm{P}(A) > 0$ and $\mathrm{P}(B) > 0$. Denote by $\mathrm{P}(A \,|B)$ the conditional probability of A given B and by $\mathrm{P}(B \,|A)$ the conditional probability of B given A.

Bayes'rule states that:

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

Proof. Take the conditional probability formulae

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

and

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

Re-arrange the second formula:

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

and plug it into the first formula:

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

■

The following example shows how Bayes' rule can be applied in a practical situation.

Example 72 *An HIV test gives a positive result with probability 98% when the patient is indeed affected by HIV, while it gives a negative result with 99% probability when the patient is not affected by HIV. If a patient is drawn at random from a population in which 0,1% of individuals are affected by HIV and he is found positive, what is the probability that he is indeed affected by HIV? In probabilistic terms, what we know about this problem can be formalized as follows:*

$$\begin{aligned} \mathrm{P}\left(positive|HIV\right) &= 0.98 \\ \mathrm{P}\left(positive|NO\ HIV\right) &= 1 - 0.99 = 0.01 \\ \mathrm{P}\left(HIV\right) &= 0.001 \\ \mathrm{P}\left(NO\ HIV\right) &= 1 - 0.001 = 0.999 \end{aligned}$$

The unconditional probability of being found positive can be derived using the law of total probability[1]*:*

$$\begin{aligned} \mathrm{P}\left(positive\right) &= \mathrm{P}\left(positive|HIV\right)\mathrm{P}\left(HIV\right) + \mathrm{P}\left(positive|NO\ HIV\right)\mathrm{P}\left(NO\ HIV\right) \\ &= 0.98 \cdot 0.001 + 0.01 \cdot 0.999 = 0.00098 + 0.00999 = 0.01097 \end{aligned}$$

Using Bayes' rule:

$$\begin{aligned} \mathrm{P}\left(HIV|positive\right) &= \frac{\mathrm{P}\left(positive|HIV\right)\mathrm{P}\left(HIV\right)}{\mathrm{P}\left(positive\right)} \\ &= \frac{0.98 \cdot 0.001}{0.01097} = \frac{0.00098}{0.01097} \\ &\simeq 0.08933 \end{aligned}$$

Therefore, even if the test is conditionally very accurate, the unconditional probability of being affected by HIV when found positive is less than 10 per cent!

13.2 Terminology

The quantities involved in Bayes' rule

$$\mathrm{P}\left(A|B\right) = \frac{\mathrm{P}\left(B|A\right)\mathrm{P}\left(A\right)}{\mathrm{P}\left(B\right)}$$

often take the following names:

1. $\mathrm{P}\left(A\right)$ is called **prior probability** or, simply, **prior**;
2. $\mathrm{P}\left(B|A\right)$ is called **conditional probability** or **likelihood**;
3. $\mathrm{P}\left(B\right)$ is called **marginal probability**;
4. $\mathrm{P}\left(A|B\right)$ is called **posterior probability** or, simply, **posterior**.

13.3 Solved exercises

Below you can find some exercises with explained solutions.

[1] See p. 90.

Exercise 1

There are two urns containing colored balls. The first urn contains 50 red balls and 50 blue balls. The second urn contains 30 red balls and 70 blue balls. One of the two urns is randomly chosen (both urns have probability 50% of being chosen) and then a ball is drawn at random from one of the two urns. If a red ball is drawn, what is the probability that it comes from the first urn?

Solution

In probabilistic terms, what we know about this problem can be formalized as follows:

$$\begin{aligned}
\mathrm{P}(\text{red}\,|\text{urn 1}) &= \frac{1}{2} \\
\mathrm{P}(\text{red}\,|\text{urn 2}) &= \frac{3}{10} \\
\mathrm{P}(\text{urn1}) &= \frac{1}{2} \\
\mathrm{P}(\text{urn 2}) &= \frac{1}{2}
\end{aligned}$$

The unconditional probability of drawing a red ball can be derived using the law of total probability:

$$\begin{aligned}
\mathrm{P}(\text{red}) &= \mathrm{P}(\text{red}\,|\text{urn 1})\,\mathrm{P}(\text{urn 1}) + \mathrm{P}(\text{red}\,|\text{urn 2})\,\mathrm{P}(\text{urn 2}) \\
&= \frac{1}{2}\cdot\frac{1}{2} + \frac{3}{10}\cdot\frac{1}{2} = \frac{1}{4} + \frac{3}{20} = \frac{5+3}{20} = \frac{2}{5}
\end{aligned}$$

Using Bayes' rule we obtain:

$$\begin{aligned}
\mathrm{P}(\text{urn 1}\,|\text{red}) &= \frac{\mathrm{P}(\text{red}\,|\text{urn 1})\,\mathrm{P}(\text{urn 1})}{\mathrm{P}(\text{red})} \\
&= \frac{\frac{1}{2}\cdot\frac{1}{2}}{\frac{2}{5}} = \frac{1}{4}\cdot\frac{5}{2} = \frac{5}{8}
\end{aligned}$$

Exercise 2

An economics consulting firm has created a model to predict recessions. The model predicts a recession with probability 80% when a recession is indeed coming and with probability 10% when no recession is coming. The unconditional probability of falling into a recession is 20%. If the model predicts a recession, what is the probability that a recession will indeed come?

Solution

What we know about this problem can be formalized as follows:

$$\begin{aligned}
\mathrm{P}(\text{rec. pred.}\,|\text{rec. coming}) &= \frac{8}{10} \\
\mathrm{P}(\text{rec. pred.}\,|\text{rec. not coming}) &= \frac{1}{10} \\
\mathrm{P}(\text{rec. coming}) &= \frac{2}{10}
\end{aligned}$$

$$\mathrm{P}\left(\text{rec. not coming}\right) = 1 - \mathrm{P}\left(\text{rec. coming}\right) = 1 - \frac{2}{10} = \frac{8}{10}$$

The unconditional probability of predicting a recession can be derived using the law of total probability:

$$\begin{aligned}\mathrm{P}\left(\text{rec. pred.}\right) &= \mathrm{P}\left(\text{rec. pred.}\,|\text{rec. coming}\right)\mathrm{P}\left(\text{rec. coming}\right)\\ &\quad +\mathrm{P}\left(\text{rec. pred.}\,|\text{rec. not coming}\right)\mathrm{P}\left(\text{rec. not coming}\right)\\ &= \frac{8}{10}\cdot\frac{2}{10}+\frac{1}{10}\cdot\frac{8}{10}=\frac{24}{100}\end{aligned}$$

Using Bayes' rule we obtain:

$$\begin{aligned}\mathrm{P}\left(\text{rec. coming}\,|\text{rec. pred.}\right) &= \frac{\mathrm{P}\left(\text{rec. pred.}\,|\text{rec. coming}\right)\mathrm{P}\left(\text{rec. coming}\right)}{\mathrm{P}\left(\text{rec. pred.}\right)}\\ &= \frac{\frac{8}{10}\cdot\frac{2}{10}}{\frac{24}{100}}=\frac{16}{100}\cdot\frac{100}{24}=\frac{2}{3}\end{aligned}$$

Exercise 3

Alice has two coins in her pocket, a fair coin (head on one side and tail on the other side) and a two-headed coin. She picks one at random from her pocket, tosses it and obtains head. What is the probability that she flipped the fair coin?

Solution

What we know about this problem can be formalized as follows:

$$\begin{aligned}&\mathrm{P}\left(\text{head}\,|\text{fair coin}\right)=\frac{1}{2}\\ &\mathrm{P}\left(\text{head}\,|\text{unfair coin}\right)=1\\ &\mathrm{P}\left(\text{fair coin}\right)=\frac{1}{2}\\ &\mathrm{P}\left(\text{unfair coin}\right)=\frac{1}{2}\end{aligned}$$

The unconditional probability of obtaining head can be derived using the law of total probability:

$$\begin{aligned}\mathrm{P}\left(\text{head}\right) &= \mathrm{P}\left(\text{head}\,|\text{fair coin}\right)\mathrm{P}\left(\text{fair coin}\right)\\ &\quad +\mathrm{P}\left(\text{head}\,|\text{unfair coin}\right)\mathrm{P}\left(\text{unfair coin}\right)\\ &= \frac{1}{2}\cdot\frac{1}{2}+1\cdot\frac{1}{2}=\frac{1}{4}+\frac{2}{4}=\frac{3}{4}\end{aligned}$$

Using Bayes' rule we obtain:

$$\begin{aligned}\mathrm{P}\left(\text{fair coin}\,|\text{head}\right) &= \frac{\mathrm{P}\left(\text{head}\,|\text{fair coin}\right)\mathrm{P}\left(\text{fair coin}\right)}{\mathrm{P}\left(\text{head}\right)}\\ &= \frac{\frac{1}{2}\cdot\frac{1}{2}}{\frac{3}{4}}=\frac{1}{4}\cdot\frac{4}{3}=\frac{1}{3}\end{aligned}$$

Chapter 14

Independent events

This lecture introduces the notion of independent event. Before reading this lecture, make sure you are familiar with the concept of conditional probability[1].

14.1 Definition of independent event

Two events E and F are said to be independent if the occurrence of E makes it neither more nor less probable that F occurs and, conversely, if the occurrence of F makes it neither more nor less probable that E occurs. In other words, after receiving the information that E will happen, we revise our assessment of the probability that F will happen, computing the conditional probability of F given E; if F is independent of E, the probability of F remains the same as it was before receiving the information:

$$\mathrm{P}\left(F\,|E\right) = \mathrm{P}\left(F\right) \tag{14.1}$$

Conversely,

$$\mathrm{P}\left(E\,|F\right) = \mathrm{P}\left(E\right) \tag{14.2}$$

In standard probability theory, rather than characterizing independence by the above two properties, independence is characterized in a more compact way.

Definition 73 *Two events E and F are **independent** if and only if*

$$\mathrm{P}\left(E \cap F\right) = \mathrm{P}\left(E\right)\mathrm{P}\left(F\right)$$

This definition implies properties (14.1) and (14.2) above: if E and F are independent, and (say) $\mathrm{P}\left(E\right) > 0$, then

$$\mathrm{P}\left(F\,|E\right) = \frac{\mathrm{P}\left(E \cap F\right)}{\mathrm{P}\left(E\right)} = \frac{\mathrm{P}\left(E\right)\mathrm{P}\left(F\right)}{\mathrm{P}\left(E\right)} = \mathrm{P}\left(F\right)$$

Example 74 *An urn contains four balls B_1, B_2, B_3 and B_4. We draw one of them at random. The sample space is*

$$\Omega = \{B_1, B_2, B_3, B_4\}$$

[1] See p. 85.

Each of the four balls has the same probability of being drawn, equal to $\frac{1}{4}$, i.e.,

$$\mathrm{P}(\{B_1\}) = \mathrm{P}(\{B_2\}) = \mathrm{P}(\{B_3\}) = \mathrm{P}(\{B_4\}) = \frac{1}{4}$$

Define the events E and F as follows:

$$\begin{aligned} E &= \{B_1, B_2\} \\ F &= \{B_2, B_3\} \end{aligned}$$

Their respective probabilities are

$$\begin{aligned} \mathrm{P}(E) &= \mathrm{P}(\{B_1\} \cup \{B_2\}) = \mathrm{P}(\{B_1\}) + \mathrm{P}(\{B_2\}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \\ \mathrm{P}(F) &= \mathrm{P}(\{B_2\} \cup \{B_3\}) = \mathrm{P}(\{B_2\}) + \mathrm{P}(\{B_3\}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \end{aligned}$$

The probability of the event $E \cap F$ is

$$\mathrm{P}(E \cap F) = \mathrm{P}(\{B_1, B_2\} \cap \{B_2, B_3\}) = \mathrm{P}(\{B_2\}) = \frac{1}{4}$$

Hence,

$$\mathrm{P}(E)\,\mathrm{P}(F) = \frac{1}{2}\frac{1}{2} = \frac{1}{4} = \mathrm{P}(E \cap F)$$

As a consequence, E and F are independent.

14.2 Mutually independent events

The definition of independence can be extended also to collections of more than two events.

Definition 75 *Let $E_1, \ldots, E_n$ be n events. $E_1, \ldots, E_n$ are **jointly independent** (or **mutually independent**) if and only if, for any sub-collection of k events ($k \leq n$) $E_{i_1}, \ldots, E_{i_k}$, we have that*

$$\mathrm{P}\left(\bigcap_{j=1}^{k} E_{i_j}\right) = \prod_{j=1}^{k} \mathrm{P}\left(E_{i_j}\right)$$

Let $E_1, \ldots, E_n$ be a collection of n events. It is important to note that even if all the possible couples of events are independent (i.e., E_i is independent of E_j for any $j \neq i$), this does not imply that the events $E_1, \ldots, E_n$ are jointly independent. This is proved with a simple counter-example:

Example 76 *Consider the experiment presented in the previous example (extracting a ball from an urn that contains four balls). Define the events E, F and G as follows:*

$$\begin{aligned} E &= \{B_1, B_2\} \\ F &= \{B_2, B_3\} \\ G &= \{B_2, B_4\} \end{aligned}$$

It is immediate to show that

$$\begin{aligned} \mathrm{P}(E \cap F) &= \frac{1}{4} = \mathrm{P}(E)\,\mathrm{P}(F) \Longrightarrow E \text{ and } F \text{ are independent} \\ \mathrm{P}(E \cap G) &= \frac{1}{4} = \mathrm{P}(E)\,\mathrm{P}(G) \Longrightarrow E \text{ and } G \text{ are independent} \\ \mathrm{P}(F \cap G) &= \frac{1}{4} = \mathrm{P}(F)\,\mathrm{P}(G) \Longrightarrow F \text{ and } G \text{ are independent} \end{aligned}$$

Thus, all the possible couple of events in the collection E, F, G are independent. However, the three events are not jointly independent because

$$\mathrm{P}(E \cap F \cap G) = \mathrm{P}(\{B_2\}) = \frac{1}{4} \neq \frac{1}{8} = \mathrm{P}(E)\,\mathrm{P}(F)\,\mathrm{P}(G)$$

On the contrary, it is obviously true that if $E_1, \ldots, E_n$ are jointly independent, then E_i is independent of E_j for any $j \neq i$.

14.3 Zero-probability events and independence

Proposition 77 *If E is a zero-probability event*[2]*, then E is independent of any other event F.*

Proof. Note that

$$(E \cap F) \subseteq E$$

As a consequence, by the monotonicity of probability[3], we have that

$$\mathrm{P}(E \cap F) \leq \mathrm{P}(E)$$

But $\mathrm{P}(E) = 0$, so $\mathrm{P}(E \cap F) \leq 0$. Since probabilities cannot be negative, it must be $\mathrm{P}(E \cap F) = 0$. The latter fact implies independence:

$$\mathrm{P}(E \cap F) = 0 = 0 \cdot \mathrm{P}(F) = \mathrm{P}(E) \cdot \mathrm{P}(F)$$

■

14.4 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Suppose that we toss a die. Six numbers (from 1 to 6) can appear face up, but we do not yet know which one of them will appear. The sample space is

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

Each of the six numbers is a sample point and is assigned probability $\frac{1}{6}$. Define the events E and F as follows:

$$\begin{aligned} E &= \{1, 3, 4\} \\ F &= \{3, 4, 5, 6\} \end{aligned}$$

Prove that E and F are independent.

[2] See p. 79.
[3] See p. 73.

Solution

The probability of E is

$$\begin{aligned} \mathrm{P}(E) &= \mathrm{P}(\{1\}) + \mathrm{P}(\{3\}) + \mathrm{P}(\{4\}) \\ &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2} \end{aligned}$$

The probability of F is

$$\begin{aligned} \mathrm{P}(F) &= \mathrm{P}(\{3\}) + \mathrm{P}(\{4\}) + \mathrm{P}(\{5\}) + \mathrm{P}(\{6\}) \\ &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{4}{6} = \frac{2}{3} \end{aligned}$$

The probability of $E \cap F$ is

$$\begin{aligned} \mathrm{P}(E \cap F) &= \mathrm{P}(\{1,3,4\} \cap \{3,4,5,6\}) = \mathrm{P}(\{3,4\}) \\ &= \mathrm{P}(\{3\}) + \mathrm{P}(\{4\}) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3} \end{aligned}$$

The events E and F are independent because

$$\mathrm{P}(E \cap F) = \frac{1}{3} = \frac{1}{2} \cdot \frac{2}{3} = \mathrm{P}(E) \cdot \mathrm{P}(F)$$

Exercise 2

A firm undertakes two projects, A and B. The probabilities of having a successful outcome are $\frac{3}{4}$ for project A and $\frac{1}{2}$ for project B. The probability that both projects will have a successful outcome is $\frac{7}{16}$. Are the outcomes of the two projects independent?

Solution

Denote by E the event "project A is successful", by F the event "project B is successful" and by G the event "both projects are successful". The event G can be expressed as

$$G = E \cap F$$

If E and F are independent, it must be that

$$\begin{aligned} \mathrm{P}(G) &= \mathrm{P}(E \cap F) = \mathrm{P}(E)\,\mathrm{P}(F) \\ &= \frac{3}{4} \cdot \frac{1}{2} = \frac{3}{8} \neq \frac{7}{16} \end{aligned}$$

Therefore, the outcomes of the two projects are not independent.

Exercise 3

A firm undertakes two projects, A and B. The probabilities of having a successful outcome are $\frac{2}{3}$ for project A and $\frac{4}{5}$ for project B. What is the probability that neither of the two projects will have a successful outcome if their outcomes are independent?

Solution

Denote by E the event "project A is successful", by F the event "project B is successful" and by G the event "neither of the two projects is successful". The event G can be expressed as

$$G = E^c \cap F^c$$

where E^c and F^c are the complements of E and F. Thus, the probability that neither of the two projects will have a successful outcome is

$$\begin{aligned}
\mathrm{P}(G) &= \mathrm{P}(E^c \cap F^c) \\
\boxed{\text{A}} &= \mathrm{P}((E \cup F)^c) \\
\boxed{\text{B}} &= 1 - \mathrm{P}(E \cup F) \\
\boxed{\text{C}} &= 1 - (\mathrm{P}(E) + \mathrm{P}(F) - \mathrm{P}(E \cap F)) \\
&= 1 - \mathrm{P}(E) - \mathrm{P}(F) + \mathrm{P}(E \cap F) \\
\boxed{\text{D}} &= 1 - \mathrm{P}(E) - \mathrm{P}(F) + \mathrm{P}(E)\,\mathrm{P}(F) \\
&= 1 - \frac{2}{3} - \frac{4}{5} + \frac{2}{3} \cdot \frac{4}{5} \\
&= \frac{15 - 10 - 12 + 8}{15} = \frac{1}{15}
\end{aligned}$$

where: in step A we have used De Morgan's law[4]; in step B we have used the formula for the probability of a complement[5]; in step C we have used the formula for the probability of a union[6]; in step D we have used the fact that E and F are independent.

[4] See p. 7.
[5] See p. 72.
[6] See p. 73.

Chapter 15

Random variables

This lecture introduces the concept of random variable. Before reading this lecture, make sure you are familiar with the concepts of sample space, sample point, event, possible outcome, realized outcome, and probability (see the lecture entitled *Probability* - p. 69).

15.1 Definition of random variable

A random variable is a variable whose value depends on the outcome of a probabilistic experiment. Its value is a priori unknown, but it becomes known once the outcome of the experiment is realized.

Denote by Ω the sample space, that is, the set of all possible outcomes of the experiment. A random variable associates a real number to each element of Ω, as stated by the following definition.

Definition 78 *A **random variable** X is a function from the sample space Ω to the set of real numbers $\mathbb{R}$:*

$$X : \Omega \to \mathbb{R}$$

In rigorous (measure-theoretic) probability theory, the function X is also required to be measurable[1].

The real number $X(\omega)$ associated to a sample point $\omega \in \Omega$ is called a **realization** of the random variable. The set of all possible realizations is called **support** and it is denoted by R_X.

Some remarks on notation are in order:

1. the dependence of X on ω is often omitted, i.e., we simply write X instead of $X(\omega)$;

2. if $A \subseteq \mathbb{R}$, the exact meaning of the notation $\mathrm{P}(X \in A)$ is the following:

$$\mathrm{P}(X \in A) = \mathrm{P}(\{\omega \in \Omega : X(\omega) \in A\})$$

3. if $A \subseteq \mathbb{R}$, we sometimes use the notation $\mathrm{P}_X(A)$ with the following meaning:

$$\mathrm{P}_X(A) = \mathrm{P}(X \in A) = \mathrm{P}(\{\omega \in \Omega : X(\omega) \in A\})$$

[1] See below the subsection entitled *A more rigorous definition of random variable* (p. 109).

In this case, P_X is to be interpreted as a probability measure on the set of real numbers, induced by the random variable X. Often, statisticians construct probabilistic models where a random variable X is defined by directly specifying P_X, without specifying the sample space Ω.

Example 79 *Suppose that we flip a coin. The possible outcomes are either tail (T) or head (H), i.e.,*

$$\Omega = \{T, H\}$$

The two outcomes are assigned equal probabilities:

$$\mathrm{P}(\{T\}) = \mathrm{P}(\{H\}) = \frac{1}{2}$$

If tail (T) is the outcome, we win one dollar, if head (H) is the outcome we lose one dollar. The amount X we win (or lose) is a random variable, defined as

$$X(\omega) = \begin{cases} 1 & \text{if } \omega = T \\ -1 & \text{if } \omega = H \end{cases}$$

The probability of winning one dollar is

$$\mathrm{P}(X = 1) = \mathrm{P}(\{\omega \in \Omega : X(\omega) = 1\}) = \mathrm{P}(\{T\}) = \frac{1}{2}$$

The probability of losing one dollar is

$$\mathrm{P}(X = -1) = \mathrm{P}(\{\omega \in \Omega : X(\omega) = -1\}) = \mathrm{P}(\{H\}) = \frac{1}{2}$$

The probability of losing two dollars is

$$\mathrm{P}(X = -2) = \mathrm{P}(\{\omega \in \Omega : X(\omega) = -2\}) = \mathrm{P}(\emptyset) = 0$$

Most of the time, statisticians deal with two special kinds of random variables:

1. discrete random variables;
2. absolutely continuous random variables.

We define these two kinds of random variables below.

15.2 Discrete random variables

Discrete random variables are defined as follows.

Definition 80 *A random variable X is **discrete** if*

1. *its support R_X is a countable set[2];*
2. *there is a function $p_X : \mathbb{R} \to [0, 1]$, called the **probability mass function** (or pmf or probability function) of X, such that, for any $x \in \mathbb{R}$:*

$$p_X(x) = \begin{cases} \mathrm{P}(X = x) & \text{if } x \in R_X \\ 0 & \text{if } x \notin R_X \end{cases}$$

[2] See the lecture entitled *Sequences and Limits* (p. 32) for a definition of countable set.

The following is an example of a discrete random variable.

Example 81 *Let X be a discrete random variable that can take only two values: 1 with probability q and 0 with probability $1-q$, where $0 \leq q \leq 1$. Its support is*

$$R_X = \{0, 1\}$$

Its probability mass function is

$$p_X(x) = \begin{cases} q & \text{if } x = 1 \\ 1-q & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases}$$

The properties of probability mass functions are discussed in more detail in the lecture entitled *Legitimate probability mass functions* (p. 247). We anticipate here that probability mass functions are characterized by two fundamental properties:

1. **non-negativity:** $p_X(x) \geq 0$ for any $x \in \mathbb{R}$;
2. **sum over the support equals 1:** $\sum_{x \in R_X} p_X(x) = 1$.

It turns out not only that any probability mass function must satisfy these two properties, but also that any function satisfying these two properties is a legitimate probability mass function. You can find a detailed discussion of this fact in the aforementioned lecture.

15.3 Absolutely continuous random variables

Absolutely continuous random variables are defined as follows.

Definition 82 *A random variable X is **absolutely continuous** if*

1. *its support R_X is not countable;*
2. *there is a function $f_X : \mathbb{R} \to [0, 1]$, called the **probability density function** (or pdf or density function) of X, such that, for any interval $[a, b] \subseteq \mathbb{R}$,*

$$\mathrm{P}(X \in [a, b]) = \int_a^b f_X(x)\, dx$$

Absolutely continuous random variables are often called **continuous random variables**, omitting the adverb absolutely.

The following is an example of an absolutely continuous random variable.

Example 83 *Let X be an absolutely continuous random variable that can take any value in the interval $[0, 1]$. All sub-intervals of equal length are equally likely. Its support is*

$$R_X = [0, 1]$$

Its probability density function is

$$f_X(x) = \begin{cases} 1 & \text{if } x \in [0, 1] \\ 0 & \text{otherwise} \end{cases}$$

The probability that the realization of X belongs, for example, to the interval $\left[\frac{1}{4}, \frac{3}{4}\right]$ is

$$\begin{aligned} \mathrm{P}\left(X \in [a,b]\right) &= \int_{1/4}^{3/4} f_X(x)\,dx = \int_{1/4}^{3/4} dx \\ &= [x]_{1/4}^{3/4} = \frac{3}{4} - \frac{1}{4} = \frac{1}{2} \end{aligned}$$

The properties of probability density functions are discussed in more detail in the lecture entitled *Legitimate probability density functions* (p. 251). We anticipate here that probability density functions are characterized by two fundamental properties:

1. **non-negativity:** $f_X(x) \geq 0$ for any $x \in \mathbb{R}$;
2. **integral over $\mathbb{R}$ equals 1:** $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$.

It turns out not only that any probability density function must satisfy these two properties, but also that any function satisfying these two properties is a legitimate probability density function. You can find a detailed discussion of this fact in the aforementioned lecture.

15.4 Random variables in general

Random variables, also those that are neither discrete nor absolutely continuous, are often characterized in terms of their distribution function.

Definition 84 *Let X be a random variable. The **distribution function** (or cumulative distribution function or cdf) of X is a function $F_X : \mathbb{R} \to [0,1]$ such that*

$$F_X(x) = \mathrm{P}\left(X \leq x\right), \ \forall x \in \mathbb{R}$$

If we know the distribution function of a random variable X, then we can easily compute the probability that X belongs to an interval $(a,b] \subseteq \mathbb{R}$, as

$$\mathrm{P}\left(a < X \leq b\right) = F_X(b) - F_X(a)$$

Proof. Note that

$$(-\infty, b] = (-\infty, a] \cup (a, b]$$

where the two sets on the right hand side are disjoint. Hence, by additivity, we get

$$\begin{aligned} F_X(b) &= \mathrm{P}\left(X \in (-\infty, b]\right) \\ &= \mathrm{P}_X\left((-\infty, b]\right) \\ &= \mathrm{P}_X\left((-\infty, a] \cup (a, b]\right) \\ &= \mathrm{P}_X\left((-\infty, a]\right) + \mathrm{P}_X\left((a, b]\right) \\ &= \mathrm{P}\left(X \leq a\right) + \mathrm{P}\left(a < X \leq x\right) \\ &= F_X(a) + \mathrm{P}\left(a < X \leq x\right) \end{aligned}$$

Rearranging terms, we obtain

$$\mathrm{P}\left(a < X \leq x\right) = F_X(b) - F_X(a)$$

■

15.5 More details

15.5.1 Derivative of the distribution function

If X is an absolutely continuous random variable, then its distribution function can be written as

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt$$

Hence, by taking the derivative with respect to x of both sides of the above equation, we obtain

$$\frac{dF_X(x)}{dx} = f_X(x)$$

15.5.2 Continuous variables and zero-probability events

If X is an absolutely continuous random variable, then the probability that X takes on any specific value $x \in R_X$ is equal to zero:

$$\mathrm{P}(X = x) = \int_{x}^{x} f_X(t)\,dt = 0$$

Thus, the event $\{\omega : X(\omega) = x\}$ is a zero-probability event for any $x \in R_X$. The lecture entitled *Zero-probability events* (p. 79) contains a thorough discussion of this apparently paradoxical fact: although it can happen that $X(\omega) = x$, the event $\{\omega : X(\omega) = x\}$ has zero probability of happening.

15.5.3 A more rigorous definition of random variable

Random variables can be defined in a more rigorous manner using the terminology of measure theory.

Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space[3]. Let X be a function $X : \Omega \to \mathbb{R}$. Let $\mathfrak{B}(\mathbb{R})$ be the Borel sigma-algebra of $\mathbb{R}$, i.e. the smallest sigma-algebra containing all the open subsets of $\mathbb{R}$. If, for any $B \in \mathfrak{B}(\mathbb{R})$,

$$\{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}$$

then X is a random variable on Ω.

If X satisfies this property, then it is possible to define

$$\mathrm{P}(X \in B) := \mathrm{P}(\{\omega \in \Omega : X(\omega) \in B\})$$

for any $B \in \mathfrak{B}(\mathbb{R})$ and the probability on the right hand side is well-defined because the set

$$\{\omega \in \Omega : X(\omega) \in B\}$$

is measurable by the very definition of random variable.

15.6 Solved exercises

Below you can find some exercises with explained solutions.

[3] See p. 76.

Exercise 1

Let X be a discrete random variable. Let its support R_X be

$$R_X = \{0, 1, 2, 3, 4\}$$

Let its probability mass function $p_X(x)$ be

$$p_X(x) = \begin{cases} 1/5 & \text{if } x \in R_X \\ 0 & \text{if } x \notin R_X \end{cases}$$

Compute

$$\mathrm{P}(1 \leq X < 4)$$

Solution

By using the additivity of probability, we have

$$\begin{aligned} \mathrm{P}(1 \leq X < 4) &= \mathrm{P}(\{X = 1\} \cup \{X = 2\} \cup \{X = 3\}) \\ &= \mathrm{P}(\{X = 1\}) + \mathrm{P}(\{X = 2\}) + \mathrm{P}(\{X = 3\}) \\ &= p_X(1) + p_X(2) + p_X(3) = \frac{1}{5} + \frac{1}{5} + \frac{1}{5} = \frac{3}{5} \end{aligned}$$

Exercise 2

Let X be a discrete random variable. Let its support R_X be the set of the first 20 natural numbers:

$$R_X = \{1, 2, \ldots, 19, 20\}$$

Let its probability mass function $p_X(x)$ be

$$p_X(x) = \begin{cases} x/210 & \text{if } x \in R_X \\ 0 & \text{if } x \notin R_X \end{cases}$$

Compute the probability

$$\mathrm{P}(X > 17)$$

Solution

By the additivity of probability, we have

$$\begin{aligned} \mathrm{P}(X > 17) &= \mathrm{P}(\{X = 18\} \cup \{X = 19\} \cup \{X = 20\}) \\ &= \mathrm{P}(\{X = 18\}) + \mathrm{P}(\{X = 19\}) + \mathrm{P}(\{X = 20\}) \\ &= p_X(18) + p_X(19) + p_X(20) \\ &= \frac{18}{210} + \frac{19}{210} + \frac{20}{210} = \frac{57}{210} = \frac{19}{70} \end{aligned}$$

Exercise 3

Let X be a discrete random variable. Let its support R_X be

$$R_X = \{0, 1, 2, 3\}$$

Let its probability mass function $p_X(x)$ be

$$p_X(x) = \begin{cases} \binom{3}{x}\left(\frac{1}{4}\right)^x\left(\frac{3}{4}\right)^{3-x} & \text{if } x \in R_X \\ 0 & \text{if } x \notin R_X \end{cases}$$

where

$$\binom{3}{x} = \frac{3!}{x!\,(3-x)!}$$

is a binomial coefficient[4].

Calculate the probability

$$\mathrm{P}(X < 3)$$

Solution

First note that, by additivity, we have

$$\begin{aligned} \mathrm{P}(X<3) &= \mathrm{P}(\{X=0\} \cup \{X=1\} \cup \{X=2\}) \\ &= \mathrm{P}(\{X=0\}) + \mathrm{P}(\{X=1\}) + \mathrm{P}(\{X=2\}) \\ &= p_X(0) + p_X(1) + p_X(2) \end{aligned}$$

Therefore, in order to compute $\mathrm{P}(X<3)$, we need to evaluate the probability mass function at the three points $x=0$, $x=1$ and $x=2$:

$$p_X(0) = \binom{3}{0}\left(\frac{1}{4}\right)^0\left(\frac{3}{4}\right)^{3-0} = \frac{3!}{0!3!}\cdot 1 \cdot \frac{27}{64} = \frac{27}{64}$$

$$\begin{aligned} p_X(1) &= \binom{3}{1}\left(\frac{1}{4}\right)^1\left(\frac{3}{4}\right)^{3-1} = \frac{3!}{1!2!}\cdot\frac{1}{4}\cdot\frac{9}{16} \\ &= 3\cdot\frac{1}{4}\cdot\frac{9}{16} = \frac{27}{64} \end{aligned}$$

$$p_X(2) = \binom{3}{2}\left(\frac{1}{4}\right)^2\left(\frac{3}{4}\right)^{3-2} = \frac{3!}{2!1!}\cdot\frac{1}{16}\cdot\frac{3}{4} = \frac{9}{64}$$

Finally,

$$\begin{aligned} \mathrm{P}(X<3) &= p_X(0) + p_X(1) + p_X(2) \\ &= \frac{27}{64} + \frac{27}{64} + \frac{9}{64} = \frac{63}{64} \end{aligned}$$

Exercise 4

Let X be an absolutely continuous random variable. Let its support R_X be

$$R_X = [0,1]$$

Let its probability density function $f_X(x)$ be

$$f_X(x) = \begin{cases} 1 & \text{if } x \in R_X \\ 0 & \text{if } x \notin R_X \end{cases}$$

Compute

$$\mathrm{P}\left(\frac{1}{2} \leq X \leq 2\right)$$

[4] See p. 22.

Solution

The probability that an absolutely continuous random variable takes a value in a given interval is equal to the integral of the probability density function over that interval:

$$\begin{aligned} \mathrm{P}\left(\frac{1}{2} \leq X \leq 2\right) &= \mathrm{P}\left(X \in \left[\frac{1}{2}, 2\right]\right) = \int_{1/2}^{2} f_X(x)\, dx \\ &= \int_{1/2}^{1} dx = [x]_{1/2}^{1} = 1 - \frac{1}{2} = \frac{1}{2} \end{aligned}$$

Exercise 5

Let X be an absolutely continuous random variable. Let its support R_X be

$$R_X = [0, 1]$$

Let its probability density function $f_X(x)$ be

$$f_X(x) = \begin{cases} 2x & \text{if } x \in R_X \\ 0 & \text{if } x \notin R_X \end{cases}$$

Compute

$$\mathrm{P}\left(\frac{1}{4} \leq X \leq \frac{1}{2}\right)$$

Solution

The probability that an absolutely continuous random variable takes a value in a given interval is equal to the integral of the probability density function over that interval:

$$\begin{aligned} \mathrm{P}\left(\frac{1}{4} \leq X \leq \frac{1}{2}\right) &= \mathrm{P}\left(X \in \left[\frac{1}{4}, \frac{1}{2}\right]\right) = \int_{1/4}^{1/2} f_X(x)\, dx \\ &= \int_{1/4}^{1/2} 2x dx = \left[x^2\right]_{1/4}^{1/2} = \frac{1}{4} - \frac{1}{16} = \frac{3}{16} \end{aligned}$$

Exercise 6

Let X be an absolutely continuous random variable. Let its support R_X be

$$R_X = [0, \infty)$$

Let its probability density function $f_X(x)$ be

$$f_X(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{if } x \in R_X \\ 0 & \text{if } x \notin R_X \end{cases}$$

where $\lambda > 0$.

Compute

$$\mathrm{P}(X \geq 1)$$

Solution

The probability that an absolutely continuous random variable takes a value in a given interval is equal to the integral of the probability density function over that interval:

$$\begin{aligned} \mathrm{P}(X \geq 1) &= \mathrm{P}(X \in [1, \infty)) = \int_1^{\infty} f_X(x)\, dx \\ &= \int_1^{\infty} \lambda \exp(-\lambda x)\, dx = [-\exp(-\lambda x)]_1^{\infty} \\ &= 0 - (-\exp(-\lambda)) = \exp(-\lambda) \end{aligned}$$

Chapter 16

Random vectors

This lecture introduces the concept of random vector, which is a multidimensional generalization of the concept of random variable. Before reading this lecture, make sure you are familiar with the concepts of sample space, sample point, event, possible outcome, realized outcome and probability (see the lecture entitled *Probability* - p. 69) and with the concept of random variable (see the lecture entitled *Random variables* - p. 105).

16.1 Definition of random vector

Suppose that we conduct a probabilistic experiment and that the possible outcomes of the experiment are described by a sample space Ω. A random vector is a vector whose value depends on the outcome of the experiment, as stated by the following definition.

Definition 85 *Let Ω be a sample space. A **random vector** X is a function from the sample space Ω to the set of K-dimensional real vectors $\mathbb{R}^K$:*

$$X : \Omega \to \mathbb{R}^K$$

In rigorous probability theory, the function X is also required to be measurable[1].

The real vector $X(\omega)$ associated to a sample point $\omega \in \Omega$ is called a **realization** of the random vector. The set of all possible realizations is called **support** and it is denoted by R_X.

Denote by $\mathrm{P}(E)$ the probability of an event $E \subseteq \Omega$. When dealing with random vectors, the following conventions are used:

- If $A \subseteq \mathbb{R}^K$, we often write $\mathrm{P}(X \in A)$ with the meaning

$$\mathrm{P}(X \in A) = \mathrm{P}(\{\omega \in \Omega : X(\omega) \in A\})$$

- If $A \subseteq \mathbb{R}^K$, we sometimes use the notation $\mathrm{P}_X(A)$ with the meaning

$$\mathrm{P}_X(A) = \mathrm{P}(X \in A) = \mathrm{P}(\{\omega \in \Omega : X(\omega) \in A\})$$

  In applied work, it is very commonplace to build statistical models where a random vector X is defined by directly specifying P_X, omitting the specification of the sample space Ω altogether.

[1] See below the subsection entitled *A more rigorous definition of random vector* (p. 121).

- We often write X instead of $X(\omega)$, omitting the dependence on ω.

Example 86 *Two coins are tossed. The possible outcomes of each toss can be either tail (T) or head (H). The sample space is*

$$\Omega = \{TT, TH, HT, HH\}$$

The four possible outcomes are assigned equal probabilities:

$$\mathrm{P}(\{TT\}) = \mathrm{P}(\{TH\}) = \mathrm{P}(\{HT\}) = \mathrm{P}(\{HH\}) = \frac{1}{4}$$

If tail (T) is the outcome, we win one dollar, if head (H) is the outcome we lose one dollar. A 2-dimensional random vector X indicates the amount we win (or lose) on each toss:

$$X(\omega) = \begin{cases} \begin{bmatrix} 1 & 1 \end{bmatrix} & \text{if } \omega = TT \\ \begin{bmatrix} 1 & -1 \end{bmatrix} & \text{if } \omega = TH \\ \begin{bmatrix} -1 & 1 \end{bmatrix} & \text{if } \omega = HT \\ \begin{bmatrix} -1 & -1 \end{bmatrix} & \text{if } \omega = HH \end{cases}$$

The probability of winning one dollar on both tosses is

$$\begin{aligned} \mathrm{P}\left(X = \begin{bmatrix} 1 & 1 \end{bmatrix}\right) &= \mathrm{P}\left(\left\{\omega \in \Omega : X(\omega) = \begin{bmatrix} 1 & 1 \end{bmatrix}\right\}\right) \\ &= \mathrm{P}(\{TT\}) = \frac{1}{4} \end{aligned}$$

The probability of losing one dollar on the second toss is

$$\begin{aligned} \mathrm{P}(X_2 = -1) &= \mathrm{P}(\{\omega \in \Omega : X_2(\omega) = -1\}) \\ &= \mathrm{P}(\{TH, HH\}) = \mathrm{P}(\{TH\}) + \mathrm{P}(\{HH\}) = \frac{1}{2} \end{aligned}$$

The next sections deal with discrete random vectors and absolutely continuous random vectors, two kinds of random vectors that have special properties and are often found in applications.

16.2 Discrete random vectors

Discrete random vectors are defined as follows.

Definition 87 *A random vector X is **discrete** if:*

1. *its support R_X is a countable set[2];*
2. *there is a function $p_X : \mathbb{R}^K \to [0,1]$, called the **joint probability mass function** (or joint pmf, or joint probability function) of X, such that, for any $x \in \mathbb{R}^K$:*

$$p_X(x) = \begin{cases} \mathrm{P}(X = x) & \text{if } x \in R_X \\ 0 & \text{if } x \notin R_X \end{cases}$$

[2] See the lecture entitled *Sequences and Limits* (p. 32) for a definition of countable set.

The following notations are used interchangeably to indicate the joint probability mass function:

$$p_X(x) = p_X(x_1, \ldots, x_K) = p_{X_1,\ldots,X_K}(x_1, \ldots, x_K)$$

In the second and third notation the K components of the random vector X are explicitly indicated.

Example 88 *Suppose X is a 2-dimensional random vector whose components X_1 and X_2 can take only two values: 1 or 0. Furthermore, the four possible combinations of 0 and 1 are all equally likely. X is an example of a discrete random vector. Its support is*

$$R_X = \left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\}$$

Its probability mass function is

$$p_X(x) = \begin{cases} 1/4 & \text{if } x = \begin{bmatrix} 1 & 1 \end{bmatrix}^\top \\ 1/4 & \text{if } x = \begin{bmatrix} 1 & 0 \end{bmatrix}^\top \\ 1/4 & \text{if } x = \begin{bmatrix} 0 & 1 \end{bmatrix}^\top \\ 1/4 & \text{if } x = \begin{bmatrix} 0 & 0 \end{bmatrix}^\top \\ 0 & \text{otherwise} \end{cases}$$

16.3 Absolutely continuous random vectors

Absolutely continuous random vectors are defined as follows.

Definition 89 *A random vector X is **absolutely continuous** (or, simply, continuous) if:*

1. *its support R_X is not countable;*
2. *there is a function $f_X : \mathbb{R}^K \to [0, 1]$, called the **joint probability density function** (or joint pdf or joint density function) of X, such that, for any set $A \subseteq \mathbb{R}^K$ where*

   $$A = [a_1, b_1] \times \ldots \times [a_K, b_K]$$

   the probability that X belongs to A can be calculated as follows:

   $$\mathrm{P}(X \in A) = \int_{a_1}^{b_1} \ldots \int_{a_K}^{b_K} f_X(x_1, \ldots, x_K)\, dx_K \ldots dx_1$$

   provided the above multiple integral is well defined.

The following notations are used interchangeably to indicate the joint probability density function:

$$f_X(x) = f_X(x_1, \ldots, x_K) = f_{X_1,\ldots,X_K}(x_1, \ldots, x_K)$$

In the second and third notation the K components of the random vector X are explicitly indicated.

Example 90 *Suppose X is a 2-dimensional random variable whose components X_1 and X_2 are independent uniform[3] random variables on the interval $[0,1]$. X is an example of an absolutely continuous random variable. Its support is*

$$R_X = [0,1] \times [0,1]$$

Its probability density function is

$$f_X(x) = \begin{cases} 1 & \text{if } x \in [0,1] \times [0,1] \\ 0 & \text{otherwise} \end{cases}$$

The probability that the realization of X falls in the rectangle $[0,1/2] \times [0,1/2]$ is

$$\begin{aligned} \mathrm{P}\left(X \in \left[0,\frac{1}{2}\right] \times \left[0,\frac{1}{2}\right]\right) &= \int_0^{1/2}\int_0^{1/2} f_X(x_1,x_2)\,dx_2 dx_1 \\ &= \int_0^{1/2}\int_0^{1/2} dx_2 dx_1 = \int_0^{1/2} [x_2]_0^{1/2}\,dx_1 \\ &= \int_0^{1/2} \frac{1}{2} dx_1 = \frac{1}{2}\int_0^{1/2} dx_1 = \frac{1}{2}[x_1]_0^{1/2} \\ &= \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4} \end{aligned}$$

16.4 Random vectors in general

Random vectors, also those that are neither discrete nor absolutely continuous, are often characterized in terms of their joint distribution function:

Definition 91 *Let X be a random vector. The **joint distribution function** (or joint df, or joint cumulative distribution function, or joint cdf) of X is a function $F_X : \mathbb{R}^K \to [0,1]$ such that*

$$F_X(x) = \mathrm{P}(X_1 \leq x_1, \ldots, X_K \leq x_K),\ \forall x \in \mathbb{R}^K$$

where the components of X and x are denoted by X_k and x_k respectively, for $k = 1, \ldots, K$.

The following notations are used interchangeably to indicate the joint distribution function:

$$F_X(x) = F_X(x_1, \ldots, x_K) = F_{X_1,\ldots,X_K}(x_1, \ldots, x_K)$$

In the second and third notation the K components of the random vector X are explicitly indicated.

Sometimes, we talk about the **joint distribution** of a random vector, without specifying whether we are referring to the joint distribution function, or to the joint probability mass function (in the case of discrete random vectors), or to the joint probability density function (in the case of absolutely continuous random vectors). This ambiguity is legitimate, since:

1. the joint probability mass function completely determines (and is completely determined by) the joint distribution function of a discrete random vector;

[3] See p. 359.

2. the joint probability density function completely determines (and is completely determined by) the joint distribution function of an absolutely continuous random vector.

In the remainder of this lecture, we use the term joint distribution when we are making statements that apply both to the distribution function and to the probability mass (or density) function of a random vector.

16.5 More details

16.5.1 Random matrices

A random matrix is a matrix whose entries are random variables. It is not necessary to develop a separate theory for random matrices, because a random matrix can always be written as a random vector. Given a $K \times L$ random matrix A, its vectorization, denoted by $\operatorname{vec}(A)$, is the $KL \times 1$ random vector obtained by stacking the columns of A on top of each other.

Example 92 *Let A be the following 2×2 random matrix:*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

The vectorization of A is the following 4×1 random vector:

$$\operatorname{vec}(A) = \begin{bmatrix} A_{11} \\ A_{21} \\ A_{12} \\ A_{22} \end{bmatrix}$$

When $\operatorname{vec}(A)$ is a discrete random vector, then we say that A is a discrete random matrix and the joint probability mass function of A is just the joint probability mass function of $\operatorname{vec}(A)$. By the same token, when $\operatorname{vec}(A)$ is an absolutely continuous random vector, then we say that A is an absolutely continuous random matrix and the joint probability density function of A is just the joint probability density function of $\operatorname{vec}(A)$.

16.5.2 Marginal distribution of a random vector

Let X_i be the i-th component of a K-dimensional random vector X. The distribution function $F_{X_i}(x)$ of X_i is called **marginal distribution function** of X_i. If X is discrete, then X_i is a discrete random variable[4] and its probability mass function $p_{X_i}(x)$ is called **marginal probability mass function** of X_i. If X is absolutely continuous, then X_i is an absolutely continuous random variable[5] and its probability density function $f_{X_i}(x)$ is called **marginal probability density function** of X_i.

[4] See p. 106.
[5] See p. 107.

16.5.3 Marginalization of a joint distribution

The process of deriving the distribution of a component X_i of a random vector X from the joint distribution of X is known as **marginalization**. Marginalization can also have a broader meaning: it can refer to the act of deriving the joint distribution of a subset of the set of components of X from the joint distribution of X. For example, if X is a random vector having three components X_1, X_2 and X_3, we can marginalize the joint distribution of X_1, X_2 and X_3 to find the joint distribution of X_1 and X_2; in this case we say that X_3 is marginalized out of the joint distribution of X_1, X_2 and X_3.

16.5.4 Marginal distribution of a discrete random vector

Let X_i be the i-th component of a K-dimensional discrete random vector X. The **marginal probability mass function** of X_i can be derived from the joint probability mass function of X as follows:

$$p_{X_i}(x) = \sum_{(x_1,\ldots,x_K)\in R_X : x_i = x} p_X(x_1,\ldots,x_K)$$

In other words, the probability that $X_i = x$ is obtained summing the probabilities of all the vectors of R_X whose i-th component is equal to x.

16.5.5 Marginalization of a discrete distribution

Let X_i be the i-th component of a discrete random vector X. Marginalizing X_i out of the joint distribution of X, we obtain the joint distribution of the remaining components of X, i.e. we obtain the joint distribution of the random vector X_{-i} defined as follows:

$$X_{-i} = [X_1 \ \ldots \ X_{i-1} \ \ X_{i+1} \ \ldots \ X_K]$$

The joint probability mass function of X_{-i} is

$$p_{X_{-i}}(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_K) = \sum_{x_i \in R_{X_i}} p_X(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_K)$$

In other words, the joint probability mass function of X_{-i} is obtained summing the joint probability mass function of X over all values x_i that belong to the support of X_i.

16.5.6 Marginal distribution of a continuous random vector

Let X_i be the i-th component of a K-dimensional absolutely continuous random vector X. The **marginal probability density function** of X_i can be derived from the joint probability density function of X as follows:

$$f_{X_i}(x) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f_X(x_1,\ldots,x_{i-1},x,x_{i+1},\ldots,x_K)\, dx_K \ldots dx_{i+1} dx_{i-1} \ldots dx_1$$

In other words, the joint probability density function, evaluated at $x_i = x$, is integrated with respect to all variables except x_i (so it is integrated a total of $K-1$ times).

16.5.7 Marginalization of a continuous distribution

Let X_i be the i-th component of an absolutely continuous random vector X. Marginalizing X_i out of the joint distribution of X, we obtain the joint distribution of the remaining components of X, i.e. we obtain the joint distribution of the random vector X_{-i} defined as follows:

$$X_{-i} = [X_1 \ \ldots \ X_{i-1} \ \ X_{i+1} \ \ldots \ X_K]$$

The joint probability density function of X_{-i} is

$$f_{X_{-i}}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_K) = \int_{-\infty}^{\infty} f_X(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_K)\, dx_i$$

In other words, the joint probability density function of X_{-i} is obtained integrating the joint probability density function of X with respect to x_i.

16.5.8 Partial derivative of the distribution function

Note that, if X is absolutely continuous, then

$$F_X(x) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_K} f_X(t_1, \ldots, t_K)\, dt_K \ldots dt_1$$

Hence, by taking the K^{th}-order cross-partial derivative with respect to $x_1, \ldots, x_K$ of both sides of the above equation, we obtain

$$\frac{\partial^K F_X(x)}{\partial x_1 \ldots \partial x_K} = f_X(x)$$

16.5.9 A more rigorous definition of random vector

Random vectors can be defined in a more rigorous manner using the terminology of measure theory:

Definition 93 *Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space*[6]*. Let X be a function $X : \Omega \to \mathbb{R}^K$. Let $\mathfrak{B}(\mathbb{R}^K)$ be the Borel σ-algebra of $\mathbb{R}^K$ (i.e. the smallest σ-algebra containing all open hyper-rectangles in $\mathbb{R}^K$). If*

$$\{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}$$

for any $B \in \mathfrak{B}(\mathbb{R}^K)$, then X is a random vector on Ω.

Thus, if X satisfies this property, we are allowed to define

$$\mathrm{P}(X \in B) := \mathrm{P}(\{\omega \in \Omega : X(\omega) \in B\}) \ , \forall B \in \mathfrak{B}(\mathbb{R}^K)$$

because the set $\{\omega \in \Omega : X(\omega) \in B\}$ is measurable by the very definition of random vector.

16.6 Solved exercises

Some solved exercises on random vectors can be found below.

[6] See p. 76 for a definition of probability space and measurable sets.

Exercise 1

Let X be a 2×1 discrete random vector and denote its components by X_1 and X_2. Let the support of X be the set of all 2×1 vectors such that their entries belong to the set of the first three natural numbers, i.e.,

$$R_X = \left\{ x = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^\top : x_1 \in N_3 \text{ and } x_2 \in N_3 \right\}$$

where

$$N_3 = \{1, 2, 3\}$$

Let the joint probability mass function of X be

$$p_X(x_1, x_2) = \begin{cases} \frac{1}{36} x_1 x_2 & \text{if } \begin{bmatrix} x_1 & x_2 \end{bmatrix}^\top \in R_X \\ 0 & \text{if } \begin{bmatrix} x_1 & x_2 \end{bmatrix}^\top \notin R_X \end{cases}$$

Find $\mathrm{P}(X_1 = 2 \text{ and } X_2 = 3)$.

Solution

Trivially, we need to evaluate the joint probability mass function at the point $(2, 3)$, i.e.,

$$\mathrm{P}(X_1 = 2 \text{ and } X_2 = 3) = p_X(2, 3) = \frac{1}{36} \cdot 2 \cdot 3 = \frac{6}{36} = \frac{1}{6}$$

Exercise 2

Let X be a 2×1 discrete random vector and denote its components by X_1 and X_2. Let the support of X be the set of all 2×1 vectors such that their entries belong to the set of the first three natural numbers, i.e.,

$$R_X = \left\{ x = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^\top : x_1 \in N_3 \text{ and } x_2 \in N_3 \right\}$$

where

$$N_3 = \{1, 2, 3\}$$

Let the joint probability mass function of X be

$$p_X(x_1, x_2) = \begin{cases} \frac{1}{36} (x_1 + x_2) & \text{if } \begin{bmatrix} x_1 & x_2 \end{bmatrix}^\top \in R_X \\ 0 & \text{if } \begin{bmatrix} x_1 & x_2 \end{bmatrix}^\top \notin R_X \end{cases}$$

Find $\mathrm{P}(X_1 + X_2 = 3)$.

Solution

There are only two possible cases that give rise to the occurrence $X_1 + X_2 = 3$. These cases are

$$X = \begin{bmatrix} 1 & 2 \end{bmatrix}^\top$$

and

$$X = \begin{bmatrix} 2 & 1 \end{bmatrix}^\top$$

Since these two cases are disjoint events, we can use the additivity property of probability[7]:

$$\begin{aligned} \mathrm{P}\left(X_1 + X_2 = 3\right) &= \mathrm{P}\left(\left\{X = [1 \;\; 2]^\top\right\} \cup \left\{X = [2 \;\; 1]^\top\right\}\right) \\ &= \mathrm{P}\left(\left\{X = [1 \;\; 2]^\top\right\}\right) + \mathrm{P}\left(\left\{X = [2 \;\; 1]^\top\right\}\right) \\ &= p_X\left(1,2\right) + p_X\left(2,1\right) \\ &= \frac{1}{36}\left(1+2\right) + \frac{1}{36}\left(2+1\right) = \frac{6}{36} = \frac{1}{6} \end{aligned}$$

Exercise 3

Let X be a 2×1 discrete random vector and denote its components by X_1 and X_2. Let the support of X be

$$R_X = \left\{[1 \;\; 1]^\top, [2 \;\; 0]^\top, [0 \;\; 0]^\top\right\}$$

and its joint probability mass function be

$$p_X\left(x\right) = \begin{cases} 1/3 & \text{if } x = [1 \;\; 1]^\top \\ 1/3 & \text{if } x = [2 \;\; 0]^\top \\ 1/3 & \text{if } x = [0 \;\; 0]^\top \\ 0 & \text{otherwise} \end{cases}$$

Derive the marginal probability mass functions of X_1 and X_2.

Solution

The support of X_1 is

$$R_{X_1} = \{0, 1, 2\}$$

We need to compute the probability of each element of the support of X_1:

$$\begin{aligned} p_{X_1}\left(0\right) &= \sum_{\{(x_1,x_2) \in R_X : x_1 = 0\}} p_X\left(x_1, x_2\right) = p_X\left(0,0\right) = \frac{1}{3} \\ p_{X_1}\left(1\right) &= \sum_{\{(x_1,x_2) \in R_X : x_1 = 1\}} p_X\left(x_1, x_2\right) = p_X\left(1,1\right) = \frac{1}{3} \\ p_{X_1}\left(2\right) &= \sum_{\{(x_1,x_2) \in R_X : x_1 = 2\}} p_X\left(x_1, x_2\right) = p_X\left(2,0\right) = \frac{1}{3} \end{aligned}$$

Thus, the probability mass function of X_1 is

$$p_{X_1}\left(x\right) = \sum_{\{(x_1,x_2) \in R_X : x_1 = x\}} p_X\left(x_1, x_2\right) = \begin{cases} 1/3 & \text{if } x = 0 \\ 1/3 & \text{if } x = 1 \\ 1/3 & \text{if } x = 2 \\ 0 & \text{otherwise} \end{cases}$$

The support of X_2 is

$$R_{X_2} = \{0, 1\}$$

[7] See p. 72.

We need to compute the probability of each element of the support of X_2:

$$\begin{aligned} p_{X_2}(0) &= \sum_{\{(x_1,x_2)\in R_X : x_2=0\}} p_X(x_1,x_2) = p_X(2,0) + p_X(0,0) = \frac{2}{3} \\ p_{X_2}(1) &= \sum_{\{(x_1,x_2)\in R_X : x_2=1\}} p_X(x_1,x_2) = p_X(1,1) = \frac{1}{3} \end{aligned}$$

Thus, the probability mass function of X_2 is

$$p_{X_2}(x) = \sum_{\{(x_1,x_2)\in R_X : x_2=x\}} p_X(x_1,x_2) = \begin{cases} 2/3 & \text{if } x = 0 \\ 1/3 & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases}$$

Exercise 4

Let X be a 2×1 absolutely continuous random vector and denote its components by X_1 and X_2. Let the support of X be

$$R_X = [0,2] \times [0,3]$$

i.e. the set of all 2×1 vectors such that the first component belongs to the interval $[0,2]$ and the second component belongs to the interval $[0,3]$. Let the joint probability density function of X be

$$f_X(x) = \begin{cases} 1/6 & \text{if } x \in R_X \\ 0 & \text{otherwise} \end{cases}$$

Compute $\mathrm{P}(1 \leq X_1 \leq 3, -1 \leq X_2 \leq 1)$.

Solution

By the very definition of joint probability density function:

$$\begin{aligned} & \mathrm{P}(1 \leq X_1 \leq 3, -1 \leq X_2 \leq 1) \\ = & \int_1^3 \int_{-1}^1 f_X(x_1,x_2)\, dx_2 dx_1 \\ = & \int_1^2 \int_0^1 \frac{1}{6} dx_2 dx_1 = \frac{1}{6} \int_1^2 [x_2]_0^1 \, dx_1 \\ = & \frac{1}{6} \int_1^2 1 dx_1 = \frac{1}{6} [x_1]_1^2 = \frac{1}{6} \end{aligned}$$

Exercise 5

Let X be a 2×1 absolutely continuous random vector and denote its components by X_1 and X_2. Let the support of X be

$$R_X = [0,\infty) \times [0,2]$$

i.e. the set of all 2×1 vectors such that the first component belongs to the interval $[0,\infty)$ and the second component belongs to the interval $[0,2]$. Let the joint probability density function of X be

$$f_X(x) = f_X(x_1,x_2) = \begin{cases} \exp(-2x_1) & \text{if } x \in R_X \\ 0 & \text{otherwise} \end{cases}$$

Compute $\mathrm{P}(X_1 + X_2 \leq 3)$.

Solution

First of all note that $X_1 + X_2 \leq 3$ if and only if $X_2 \leq 3 - X_1$. Using the definition of joint probability density function, we obtain

$$\begin{aligned} \mathrm{P}\left(X_1 + X_2 \leq 3\right) &= \mathrm{P}\left(\left\{\left[x_1 \;\; x_2\right]^\top : x_1 \in \mathbb{R}, x_2 \in (-\infty, 3 - x_1]\right\}\right) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{3-x_1} f_X\left(x_1, x_2\right) dx_2 dx_1 \end{aligned}$$

When $x_1 \in [0, \infty)$, the inner integral is

$$\begin{aligned} & \int_{-\infty}^{3-x_1} f_X\left(x_1, x_2\right) dx_2 \\ = & \begin{cases} 0 & \text{if } 3 - x_1 < 0, \text{ i.e. if } x_1 > 3 \\ \int_0^{3-x_1} \exp\left(-2x_1\right) dx_2 & \text{if } 0 \leq 3 - x_1 \leq 2, \text{ i.e. if } 1 \leq x_1 \leq 3 \\ \int_0^{2} \exp\left(-2x_1\right) dx_2 & \text{if } 3 - x_1 > 2, \text{ i.e. if } x_1 < 1 \end{cases} \end{aligned}$$

Therefore,

$$\begin{aligned} & \mathrm{P}\left(X_1 + X_2 \leq 3\right) \\ = & \int_{-\infty}^{\infty} \int_{-\infty}^{3-x_1} f_X\left(x_1, x_2\right) dx_2 dx_1 \\ = & \int_0^1 \int_0^2 \exp\left(-2x_1\right) dx_2 dx_1 + \int_1^3 \int_0^{3-x_1} \exp\left(-2x_1\right) dx_2 dx_1 \\ = & \int_0^1 \exp\left(-2x_1\right) \int_0^2 dx_2 dx_1 + \int_1^3 \exp\left(-2x_1\right) \int_0^{3-x_1} dx_2 dx_1 \\ = & \int_0^1 \exp\left(-2x_1\right) 2 dx_1 + \int_1^3 \exp\left(-2x_1\right) \left(3 - x_1\right) dx_1 \\ = & \left[-\exp\left(-2x_1\right)\right]_0^1 + 3 \int_1^3 \exp\left(-2x_1\right) dx_1 - \int_1^3 x_1 \exp\left(-2x_1\right) dx_1 \\ = & -\exp\left(-2\right) + 1 + 3\left[-\frac{1}{2}\exp\left(-2x_1\right)\right]_1^3 \\ & - \left\{\left[x_1\left(-\frac{1}{2}\exp\left(-2x_1\right)\right)\right]_1^3 - \int_1^3 \left(-\frac{1}{2}\exp\left(-2x_1\right)\right) dx_1\right\} \\ = & 1 - \exp\left(-2\right) - \frac{3}{2}\exp\left(-6\right) + \frac{3}{2}\exp\left(-2\right) + \frac{3}{2}\exp\left(-6\right) \\ & -\frac{1}{2}\exp\left(-2\right) + \left[\frac{1}{4}\exp\left(-2x_1\right)\right]_1^3 \\ = & 1 + \frac{1}{4}\exp\left(-6\right) - \frac{1}{4}\exp\left(-2\right) \end{aligned}$$

Exercise 6

Let X be a 2×1 absolutely continuous random vector and denote its components by X_1 and X_2. Let the support of X be

$$R_X = \mathbb{R}_+^2$$

i.e., the set of all 2-dimensional vectors with positive entries. Let its joint probability density function be

$$f_X(x) = f_X(x_1, x_2) = \begin{cases} \exp(-x_1 - x_2) & \text{if } x_1 \geq 0 \text{ and } x_2 \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Derive the marginal probability density functions of X_1 and X_2.

Solution The support of X_1 is

$$R_{X_1} = \mathbb{R}_+$$

We can find the marginal density by integrating the joint density with respect to x_2:

$$f_{X_1}(x) = \int_{-\infty}^{\infty} f_X(x, x_2)\, dx_2$$

When $x < 0$, then $f_X(x, x_2) = 0$ and the above integral is trivially equal to 0. Thus, when $x < 0$, then $f_{X_1}(x) = 0$.

When $x > 0$, then

$$f_{X_1}(x) = \int_{-\infty}^{\infty} f_X(x, x_2)\, dx_2 = \int_{-\infty}^{0} f_X(x, x_2)\, dx_2 + \int_{0}^{\infty} f_X(x, x_2)\, dx_2$$

but the first of the two integrals is zero since $f_X(x, x_2) = 0$ when $x_2 < 0$; as a consequence,

$$\begin{aligned} f_{X_1}(x) &= \int_{-\infty}^{0} f_X(x, x_2)\, dx_2 + \int_{0}^{\infty} f_X(x, x_2)\, dx_2 \\ &= \int_{0}^{\infty} f_X(x, x_2)\, dx_2 = \int_{0}^{\infty} \exp(-x - x_2)\, dx_2 \\ &= \int_{0}^{\infty} \exp(-x)\exp(-x_2)\, dx_2 = \exp(-x)\int_{0}^{\infty} \exp(-x_2)\, dx_2 \\ &= \exp(-x)\left[-\exp(-x_2)\right]_0^{\infty} = \exp(-x)(-0-(-1)) = \exp(-x) \end{aligned}$$

So, putting pieces together, the marginal density function of X_1 is

$$f_{X_1}(x) = \begin{cases} \exp(-x) & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Obviously, by symmetry, the marginal density function of X_2 is

$$f_{X_2}(x) = \begin{cases} \exp(-x) & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Chapter 17

Expected value

The concept of expected value of a random variable[1] is one of the most important concepts in probability theory. It was first devised in the 17th century to analyze gambling games and answer questions such as: how much do I gain - or lose - on average, if I repeatedly play a given gambling game? how much can I expect to gain - or lose - by performing a certain bet? If the possible outcomes of the game (or the bet) and their associated probabilities are described by a random variable, then these questions can be answered by computing its expected value, which is equal to a weighted average of the outcomes, in which each outcome is weighted by its probability. For example, if you play a game where you gain 2$ with probability 1/2 and you lose 1$ with probability 1/2, then the expected value of the game is half a dollar:

$$2\$ \cdot (1/2) + (-1\$) \cdot (1/2) = 1/2\$$$

What does this mean? Roughly speaking, it means that if you play this game many times and the number of times each of the two possible outcomes occurs is proportional to its probability, then, on average you gain 1/2$ each time you play the game. For instance, if you play the game 100 times, win 50 times and lose the remaining 50, then your average winning is equal to the expected value:

$$\left(2\$ \cdot 50 + (-1\$) \cdot 50\right) / 100 = 1/2\$$$

In general, giving a rigorous definition of expected value requires quite a heavy mathematical apparatus. To keep things simple, we provide an informal definition of expected value and we discuss its computation in this lecture, while we relegate a more rigorous definition to the (optional) lecture entitled *Expected value and the Lebesgue integral* (p. 141).

17.1 Definition of expected value

The following is an informal definition of expected value:

Definition 94 (informal) *The **expected value** of a random variable X is the weighted average of the values that X can take on, where each possible value is weighted by its respective probability.*

[1] See p. 105.

The expected value of a random variable X is denoted by $\mathrm{E}[X]$ and it is often called the **expectation** of X or the **mean** of X.

The following sections discuss how the expected value of a random variable is computed.

17.2 Discrete random variables

When X is a discrete random variable having support R_X and probability mass function $p_X(x)$, the formula for computing its expected value is a straightforward implementation of the informal definition given above: the expected value of X is the weighted average of the values that X can take on (the elements of R_X), where each possible value $x \in R_X$ is weighted by its respective probability $p_X(x)$.

Definition 95 *Let X be a discrete random variable with support R_X and probability mass function $p_X(x)$. The expected value of X is:*

$$\mathrm{E}[X] = \sum_{x \in R_X} x p_X(x)$$

provided that:

$$\sum_{x \in R_X} |x| p_X(x) < \infty$$

The symbol

$$\sum_{x \in R_X}$$

indicates summation over all the elements of the support R_X. So, for example, if

$$R_X = \{1, 2, 3\}$$

then:

$$\sum_{x \in R_X} x p_X(x) = 1 \cdot p_X(1) + 2 \cdot p_X(2) + 3 \cdot p_X(3)$$

The requirement that

$$\sum_{x \in R_X} |x| p_X(x) < \infty$$

is called **absolute summability** and ensures that the summation

$$\sum_{x \in R_X} x p_X(x) \tag{17.1}$$

is well-defined also when the support R_X contains infinitely many elements. When summing infinitely many terms, the order in which you sum them can change the result of the sum. However, if the terms are absolutely summable, then the order in which you sum becomes irrelevant. In the above definition of expected value, the order of the sum

$$\sum_{x \in R_X} x p_X(x)$$

is not specified, therefore the requirement of absolute summability is introduced in order to ensure that the expected value is well-defined.

When the absolute summability condition is not satisfied, we say that the expected value of X is not well-defined or that it does not exist.

Example 96 *Let X be a random variable with support*

$$R_X = \{0, 1\}$$

and probability mass function:

$$p_X(x) = \begin{cases} 1/2 & \text{if } x = 1 \\ 1/2 & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases}$$

Its expected value is:

$$\begin{aligned} \mathrm{E}[X] &= \sum_{x \in R_X} x p_X(x) = 1 \cdot p_X(1) + 0 \cdot p_X(0) \\ &= 1 \cdot \frac{1}{2} + 0 \cdot \frac{1}{2} = \frac{1}{2} \end{aligned}$$

17.3 Continuous random variables

When X is an absolutely continuous random variable with probability density function $f_X(x)$, the formula for computing its expected value involves an integral:

Definition 97 *Let X be an absolutely continuous random variable with probability density function $f_X(x)$. The expected value of X is:*

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx$$

provided that:

$$\int_{-\infty}^{\infty} |x| f_X(x)\, dx < \infty$$

This integral can be thought of as the limiting case of the sum (17.1) found in the discrete case. Here $p_X(x)$ is replaced by $f_X(x)\, dx$ (the infinitesimal probability of x) and the integral sign $\int_{-\infty}^{\infty}$ replaces the summation sign $\sum_{x \in R_X}$.

The requirement that

$$\int_{-\infty}^{\infty} |x| f_X(x)\, dx < \infty$$

is called **absolute integrability** and ensures that the improper integral

$$\int_{-\infty}^{\infty} x f_X(x)\, dx$$

is well-defined. This improper integral is a shorthand for:

$$\lim_{t \to -\infty} \int_t^0 x f_X(x)\, dx + \lim_{t \to \infty} \int_0^t x f_X(x)\, dx$$

and it is well-defined only if both limits are finite. Absolute integrability guarantees that the latter condition is met and that the expected value is well-defined.

When the absolute integrability condition is not satisfied, we say that the expected value of X is not well-defined or that it does not exist.

Example 98 *Let X be an absolutely continuous random variable with support*

$$R_X = [0, \infty)$$

and probability density function:

$$f_X(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{if } x \in [0, \infty) \\ 0 & \text{otherwise} \end{cases}$$

where $\lambda > 0$. Its expected value is:

$$\begin{aligned} \mathrm{E}[X] &= \int_{-\infty}^{\infty} x f_X(x)\, dx \\ &= \int_{0}^{\infty} x\lambda \exp(-\lambda x)\, dx \\ \boxed{A} &= \frac{1}{\lambda}\int_{0}^{\infty} t \exp(-t)\, dt \\ \boxed{B} &= \frac{1}{\lambda}\left\{[-t\exp(-t)]_0^{\infty} + \int_{0}^{\infty} \exp(-t)\, dt\right\} \\ &= \frac{1}{\lambda}\{0 - 0 + [-\exp(-t)]_0^{\infty}\} \\ &= \frac{1}{\lambda}\{0 + 1\} = \frac{1}{\lambda} \end{aligned}$$

where: in step $\boxed{A}$ we have made a change of variable ($t = \lambda x$); in step $\boxed{B}$ we have integrated by parts.

17.4 The Riemann-Stieltjes integral

This section introduces a general formula for computing the expected value of a random variable X. The formula, which does not require X to be discrete or absolutely continuous and is applicable to any random variable, involves an integral called Riemann-Stieltjes integral (see below for an introduction). While we briefly discuss this formula for the sake of completeness, no deep understanding of this formula or of the Riemann-Stieltjes integral is required to understand the other lectures.

Definition 99 *Let X be a random variable having distribution function[2] $F_X(x)$. The expected value of X is:*

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x dF_X(x)$$

where the integral is a Riemann-Stieltjes integral and the expected value exists and is well-defined only as long as the integral is well-defined.

Also this integral is the limiting case of formula (17.1) for the expected value of a discrete random variable. Here $dF_X(x)$ replaces $p_X(x)$ (the probability of x) and the integral sign $\int_{-\infty}^{\infty}$ replaces the summation sign $\sum_{x \in R_X}$.

[2] See p. 108.

The following section contains a brief and informal introduction to the Riemann-Stieltjes integral and an explanation of the above formula. Less technically oriented readers can safely skip it: when they encounter a Riemann-Stieltjes integral, they can just think of it as a formal notation which allows a unified treatment of discrete and absolutely continuous random variables and can be treated as a sum in one case and as an ordinary Riemann integral in the other.

17.4.1 Intuition

As we have already seen above, the expected value of a discrete random variable is straightforward to compute: the expected value of a discrete variable X is the weighted average of the values that X can take on (the elements of the support R_X), where each possible value x is weighted by its respective probability $p_X(x)$:

$$\mathrm{E}[X] = \sum_{x \in R_X} x p_X(x)$$

or, written in a slightly different fashion:

$$\mathrm{E}[X] = \sum_{x \in R_X} x \mathrm{P}(X = x)$$

When X is not discrete the above summation does not make any sense. However, there is a workaround that allows to extend the formula to random variables that are not discrete. The workaround entails approximating X with discrete variables that can take on only finitely many values.

Let $x_0, x_1, \ldots, x_n$ be $n+1$ real numbers ($n \in \mathbb{N}$) such that:

$$x_0 < x_1 < \ldots < x_n$$

Define a new random variable X_n (function of X) as follows:

$$X_n = \begin{cases} x_1 & \text{when } x_0 < X \leq x_1 \\ x_2 & \text{when } x_1 < X \leq x_2 \\ \vdots & \vdots \\ x_n & \text{when } x_{n-1} < X \leq x_n \end{cases}$$

As the number n of points increases and the points become closer and closer (the maximum distance between two successive points tends to zero), X_n becomes a very good approximation of X, until, in the limit, it is indistinguishable from X. The expected value of X_n is easy to compute:

$$\begin{aligned} \mathrm{E}[X_n] &= \sum_{i=1}^{n} x_i \mathrm{P}(X_n = x_i) \\ &= \sum_{i=1}^{n} x_i \mathrm{P}(X \in (x_{i-1}, x_i]) \\ &= \sum_{i=1}^{n} x_i [F_X(x_i) - F_X(x_{i-1})] \end{aligned}$$

where $F_X(x)$ is the distribution function of X.

The expected value of X is then defined as the limit of $\mathrm{E}[X_n]$ when n tends to infinity (i.e. when the approximation becomes better and better):

$$\mathrm{E}[X] = \lim_{n\to\infty} \mathrm{E}[X_n] = \lim_{n\to\infty} \sum_{i=1}^{n} x_i \left[F_X(x_i) - F_X(x_{i-1})\right]$$

When the latter limit exists and is well-defined, it is called the Riemann-Stieltjes integral of x with respect to $F_X(x)$ and it is indicated as follows:

$$\int_{-\infty}^{\infty} x dF_X(x) = \lim_{n\to\infty} \sum_{i=1}^{n} x_i \left[F_X(x_i) - F_X(x_{i-1})\right]$$

Roughly speaking, the integral notation $\int_{-\infty}^{\infty}$ can be thought of as a shorthand for $\lim_{n\to\infty} \sum_{i=1}^{n}$ and the differential notation $dF_X(x)$ can be thought of as a shorthand for $[F_X(x_i) - F_X(x_{i-1})]$.

17.4.2 Some rules

We present here some rules for computing the Riemann-Stieltjes integral when the integrator function is the distribution function of a random variable X, i.e. we limit attention to integrals of the kind:

$$\int_a^b g(x)\, dF_X(x)$$

where $F_X(x)$ is the distribution function of a random variable X and $g : \mathbb{R} \to \mathbb{R}$. Before stating the rules, note that the above integral does not necessarily exist or is not necessarily well-defined. Roughly speaking, for the integral to exist the integrand function g must be well-behaved. For example, if g is continuous on $[a, b]$, then the integral exists and is well-defined.

That said, we are ready to present the calculation rules:

1. $F_X(x)$ **is continuously differentiable on** $[a, b]$**.** If $F_X(x)$ is continuously differentiable on $[a, b]$ and $f_X(x)$ is its first derivative, then:

$$\int_a^b g(x)\, dF_X(x) = \int_a^b g(x) f_X(x)\, dx$$

2. $F_X(x)$ **is continuously differentiable on** $[a, b]$ **except at a finite number of points.** Suppose $F_X(x)$ is continuously differentiable on $[a, b]$ except at a finite number of points $c_1, \ldots, c_n$ such that:

$$a < c_1 < c_2 < \ldots < c_n \leq b$$

Denote the derivative of $F_X(x)$ (where it exists) by $f_X(x)$. Then:

$$\begin{aligned} & \int_a^b g(x)\, dF_X(x) \\ = & \int_a^{c_1} g(x) f_X(x)\, dx + g(c_1) \left[F_X(c_1) - \lim_{\substack{x\to c_1 \\ x<c_1}} F_X(x)\right] \end{aligned}$$

$$
\begin{aligned}
&+\int_{c_1}^{c_2} g(x) f_X(x)\,dx + g(c_2)\left[F_X(c_2) - \lim_{\substack{x\to c_2\\ x<c_2}} F_X(x)\right]\\
&+\dots\\
&+\int_{c_{n-1}}^{c_n} g(x) f_X(x)\,dx + g(c_n)\left[F_X(c_n) - \lim_{\substack{x\to c_n\\ x<c_n}} F_X(x)\right]\\
&+\int_{c_n}^{b} g(x) f_X(x)\,dx
\end{aligned}
$$

Example 100 *Let X be a random variable with support*

$$R_X = [0,1]$$

and distribution function:

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ x/2 & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

Its expected value is:

$$
\begin{aligned}
\mathrm{E}[X] &= \int_{-\infty}^{\infty} x dF_X(x)\\
&= \int_0^1 x \frac{d}{dx}\left(\frac{1}{2}x\right) dx + 1\cdot\left[F_X(1) - \lim_{\substack{x\to 1\\ x<1}} F_X(x)\right]\\
&= \int_0^1 \frac{1}{2} x dx + 1\cdot\left[1 - \frac{1}{2}\right]\\
&= \left[\frac{1}{4}x^2\right]_0^1 + \frac{1}{2} = \frac{1}{4} + \frac{1}{2} = \frac{3}{4}
\end{aligned}
$$

17.5 The Lebesgue integral

A completely general and rigorous definition of expected value is based on the Lebesgue integral. We report it below without further comments. Less technically inclined readers can safely skip it, while interested readers can read more about it in the lecture entitled *Expected value and the Lebesgue integral* (p. 141).

Definition 101 *Let Ω be a sample space*[3]*, P a probability measure defined on the events of Ω and X a random variable defined on Ω. The expected value of X is:*

$$\mathrm{E}[X] = \int X d\mathrm{P}$$

provided $\int X d\mathrm{P}$ (the Lebesgue integral of X with respect to P) exists and is well-defined.

[3] See p. 69.

17.6 More details

17.6.1 The transformation theorem

An important property of the expected value, known as transformation theorem, allows to easily compute the expected value of a function of a random variable.

Let X be a random variable. Let $g : \mathbb{R} \to \mathbb{R}$ be a real function. Define a new random variable Y as follows:

$$Y = g(X)$$

Then:

$$\mathrm{E}[Y] = \int_{-\infty}^{\infty} g(x)\, dF_X(x)$$

provided the above integral exists.

This is an important property. It says that, if you need to compute the expected value of $Y = g(X)$, you do not need to know the support of Y and its distribution function $F_Y(y)$: you can compute it just by replacing x with $g(x)$ in the formula for the expected value of X.

For discrete random variables the formula becomes:

$$\mathrm{E}[Y] = \sum_{x \in R_X} g(x)\, p_X(x)$$

while for absolutely continuous random variables it is:

$$\mathrm{E}[Y] = \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx$$

It is possible (albeit non-trivial) to prove that the above two formulae hold also when X is a K-dimensional random vector, $g : \mathbb{R}^K \to \mathbb{R}$ is a real function of K variables and $Y = g(X)$. When X is a discrete random vector and $p_X(x)$ is its joint probability mass function, then:

$$\mathrm{E}[Y] = \sum_{x \in R_X} g(x)\, p_X(x)$$

When X is an absolutely continuous random vector and $f_X(x)$ is its joint probability density function, then:

$$\mathrm{E}[Y] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \ldots, x_K)\, f_X(x_1, \ldots, x_K)\, dx_1 \ldots dx_K$$

17.6.2 Linearity of the expected value

If X is a random variable and Y is another random variable such that:

$$Y = a + bX$$

where $a \in \mathbb{R}$ and $b \in \mathbb{R}$ are two constants, then the following holds:

$$\mathrm{E}[Y] = a + b\mathrm{E}[X]$$

Proof. For discrete random variables this is proved as follows:

$$\mathrm{E}[Y]$$

$$\begin{aligned}
\boxed{\text{A}} \quad &= \sum_{x \in R_X} (a + bx)\, p_X(x) \\
&= \sum_{x \in R_X} a p_X(x) + \sum_{x \in R_X} b x p_X(x) \\
&= a \sum_{x \in R_X} p_X(x) + b \sum_{x \in R_X} x p_X(x) \\
\boxed{\text{B}} \quad &= a + b \sum_{x \in R_X} x p_X(x) \\
\boxed{\text{C}} \quad &= a + b\mathrm{E}[X]
\end{aligned}$$

where: in step A we have used the transformation theorem; in step B we have used the fact that probabilities sum up to 1; in step C we have used the definition of $\mathrm{E}[X]$. For absolutely continuous random variables the proof is:

$$\begin{aligned}
&\mathrm{E}[Y] \\
\boxed{\text{A}} \quad &= \int_{-\infty}^{\infty} (a + bx)\, f_X(x)\, dx \\
&= \int_{-\infty}^{\infty} a f_X(x)\, dx + \int_{-\infty}^{\infty} b x f_X(x)\, dx \\
&= a \int_{-\infty}^{\infty} f_X(x)\, dx + b \int_{-\infty}^{\infty} x f_X(x)\, dx \\
\boxed{\text{B}} \quad &= a + b \int_{-\infty}^{\infty} x f_X(x)\, dx \\
\boxed{\text{C}} \quad &= a + b\mathrm{E}[X]
\end{aligned}$$

where: in step A we have used the transformation theorem; in step B we have used the fact that probability densities integrate to 1; in step C we have used the definition of $\mathrm{E}[X]$. In general, the linearity property is a consequence of the transofrmation theorem and of the fact that the Riemann-Stieltjes integral is a linear operator:

$$\begin{aligned}
\mathrm{E}[Y] &= \int_{-\infty}^{\infty} (a + bx)\, dF_X(x) \\
&= a \int_{-\infty}^{\infty} dF_X(x) + b \int_{-\infty}^{\infty} x dF_X(x) \\
&= a + b\mathrm{E}[X]
\end{aligned}$$

■

A stronger linearity property holds, which involves two or more random variables. The property can be proved only using the Lebesgue integral[4]. The property is as follows: let X_1 and X_2 be two random variables and let $c_1 \in \mathbb{R}$ and $c_2 \in \mathbb{R}$ be two constants; then:

$$\mathrm{E}[c_1 X_1 + c_2 X_2] = c_1 \mathrm{E}[X_1] + c_2 \mathrm{E}[X_2]$$

[4] See the lecture entitled *Expected value and the Lebesgue integral* (p. 141).

17.6.3 Expected value of random vectors

Let X be a K-dimensional random vector and denote its components by $X_1, \ldots, X_K$. The expected value of X, denoted by $\mathrm{E}[X]$, is just the vector of the expected values of the K components of X. Suppose, for example, that X is a row vector; then:

$$\mathrm{E}[X] = \begin{bmatrix} \mathrm{E}[X_1] & \ldots & \mathrm{E}[X_K] \end{bmatrix}$$

17.6.4 Expected value of random matrices

Let Σ be a $K \times L$ random matrix, i.e. a $K \times L$ matrix whose entries are random variables. Denote its (i,j)-th entry by Σ_{ij}. The expected value of Σ, denoted by $\mathrm{E}[\Sigma]$, is just the matrix of the expected values of the entries of Σ:

$$\mathrm{E}[\Sigma] = \begin{bmatrix} \mathrm{E}[\Sigma_{11}] & \ldots & \mathrm{E}[\Sigma_{1L}] \\ \vdots & \ddots & \vdots \\ \mathrm{E}[\Sigma_{K1}] & \ldots & \mathrm{E}[\Sigma_{KL}] \end{bmatrix}$$

17.6.5 Integrability

Denote the absolute value of a random variable X by $|X|$. If $\mathrm{E}[|X|]$ exists and is finite, we say that X is an **integrable random variable**, or just that X is **integrable**.

17.6.6 L^p spaces

Let $1 \leq p < \infty$. The space of all random variables X such that $\mathrm{E}[|X|^p]$ exists and is finite is denoted by L^p or $L^p(\Omega, \mathcal{F}, \mathrm{P})$, where the triple $(\Omega, \mathcal{F}, \mathrm{P})$ makes the dependence on the underlying probability space[5] explicit. If X belongs to L^p, we write $X \in L^p(\Omega, \mathcal{F}, \mathrm{P})$. Hence, if X is integrable, we write $X \in L^1(\Omega, \mathcal{F}, \mathrm{P})$.

17.6.7 Other properties of the expected value

Other properties of the expected value are discussed in the lecture entitled *Properties of the expected value* (p. 147).

17.7 Solved exercises

Some solved exercises on the expected value can be found below.

Exercise 1

Let X be a discrete random variable. Let its support be

$$R_X = \{0, 1, 2, 3, 4\}$$

Let its probability mass function $p_X(x)$ be

$$p_X(x) = \begin{cases} 1/5 & \text{if } x \in R_X \\ 0 & \text{if } x \notin R_X \end{cases}$$

Compute the expected value of X.

[5] See p. 76.

Solution

Since X is discrete, its expected value is computed as a sum over the support of X:

$$\begin{aligned} \mathrm{E}[X] &= \sum_{x \in R_X} x p_X(x) \\ &= 0 \cdot p_X(0) + 1 \cdot p_X(1) + 2 \cdot p_X(2) + 3 \cdot p_X(3) + 4 \cdot p_X(4) \\ &= 0 \cdot \frac{1}{5} + 1 \cdot \frac{1}{5} + 2 \cdot \frac{1}{5} + 3 \cdot \frac{1}{5} + 4 \cdot \frac{1}{5} \\ &= \frac{1+2+3+4}{5} = \frac{10}{5} = 2 \end{aligned}$$

Exercise 2

Let X be a discrete random variable. Let its support be

$$R_X = \{1, 2, 3\}$$

Let its probability mass function $p_X(x)$ be

$$p_X(x) = \begin{cases} x/6 & \text{if } x \in R_X \\ 0 & \text{if } x \notin R_X \end{cases}$$

Compute the expected value of X.

Solution

Since X is discrete, its expected value is computed as a sum over the support of X:

$$\begin{aligned} \mathrm{E}[X] &= \sum_{x \in R_X} x p_X(x) = 1 \cdot p_X(1) + 2 \cdot p_X(2) + 3 \cdot p_X(3) \\ &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{2}{6} + 3 \cdot \frac{3}{6} = \frac{1}{6} + \frac{4}{6} + \frac{9}{6} = \frac{14}{6} = \frac{7}{3} \end{aligned}$$

Exercise 3

Let X be a discrete random variable. Let its support be

$$R_X = \{2, 4\}$$

Let its probability mass function $p_X(x)$ be

$$p_X(x) = \begin{cases} x^2/20 & \text{if } x \in R_X \\ 0 & \text{if } x \notin R_X \end{cases}$$

Compute the expected value of X.

Solution Since X is discrete, its expected value is computed as a sum over the support of X:

$$\begin{aligned} \mathrm{E}[X] &= \sum_{x \in R_X} x p_X(x) = 2 \cdot p_X(2) + 4 \cdot p_X(4) \\ &= 2 \cdot \frac{4}{20} + 4 \cdot \frac{16}{20} = \frac{8}{20} + \frac{64}{20} = \frac{72}{20} = \frac{18}{5} \end{aligned}$$

Exercise 4

Let X be an absolutely continuous random variable with uniform distribution on the interval $[1, 3]$.

Its support is

$$R_X = [1, 3]$$

Its probability density function is

$$f_X(x) = \begin{cases} 1/2 & \text{if } x \in [1, 3] \\ 0 & \text{otherwise} \end{cases}$$

Compute the expected value of X.

Solution

Since X is absolutely continuous, its expected value can be computed as an integral:

$$\begin{aligned} \mathrm{E}[X] &= \int_{-\infty}^{\infty} x f_X(x)\, dx \\ &= \int_{-\infty}^{1} x f_X(x)\, dx + \int_{1}^{3} x f_X(x)\, dx + \int_{3}^{\infty} x f_X(x)\, dx \\ &= \int_{-\infty}^{1} x 0 dx + \int_{1}^{3} x \frac{1}{2} dx + \int_{3}^{\infty} x 0 dx \\ &= 0 + \left[\frac{1}{4}x^2\right]_1^3 + 0 = \frac{1}{4}3^2 - \frac{1}{4}1^2 = \frac{9-1}{4} = \frac{8}{4} = 2 \end{aligned}$$

Note that the trick is to: 1) subdivide the interval of integration to isolate the sub-intervals where the density is zero; 2) split up the integral among the sub-intervals thus identified.

Exercise 5

Let X be an absolutely continuous random variable. Its support is

$$R_X = \mathbb{R}$$

Its probability density function is

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$$

Compute the expected value of X.

Solution

Since X is absolutely continuous, its expected value can be computed as an integral:

$$\begin{aligned} \mathrm{E}[X] &= \int_{-\infty}^{\infty} x f_X(x)\, dx = (2\pi)^{-1/2} \int_{-\infty}^{\infty} x \exp\left(-\frac{1}{2}x^2\right) dx \\ &= (2\pi)^{-1/2} \int_{-\infty}^{0} x \exp\left(-\frac{1}{2}x^2\right) dx + (2\pi)^{-1/2} \int_{0}^{\infty} x \exp\left(-\frac{1}{2}x^2\right) dx \end{aligned}$$

$$\begin{aligned} &= (2\pi)^{-1/2}\left[-\exp\left(-\frac{1}{2}x^2\right)\right]_{-\infty}^{0} + (2\pi)^{-1/2}\left[-\exp\left(-\frac{1}{2}x^2\right)\right]_{0}^{\infty} \\ &= (2\pi)^{-1/2}[-1+0] + (2\pi)^{-1/2}[0+1] = -(2\pi)^{-1/2} + (2\pi)^{-1/2} = 0 \end{aligned}$$

Exercise 6

Let X be an absolutely continuous random variable. Its support is

$$R_X = [0,1]$$

Its probability density function is

$$f_X(x) = 3x^2$$

Compute the expected value of X.

Solution

Since X is absolutely continuous, its expected value can be computed as an integral:

$$\begin{aligned} \mathrm{E}[X] &= \int_{-\infty}^{\infty} x f_X(x)\,dx = \int_0^1 x 3x^2 dx \\ &= 3\int_0^1 x^3 dx = 3\left[\frac{1}{4}x^4\right]_0^1 = 3\left[\frac{1}{4}-0\right] = \frac{3}{4} \end{aligned}$$

Exercise 7

Let $F_X(x)$ be defined as follows:

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

where $\lambda > 0$.

Compute the following integral:

$$\int_1^2 x dF_X(x)$$

Solution

$F_X(x)$ is continuously differentiable on the interval $[1,2]$. Its derivative $f_X(x)$ is:

$$f_X(x) = \lambda \exp(-\lambda x)$$

As a consequence, the integral becomes:

$$\begin{aligned} &\int_1^2 x dF_X(x) = \int_1^2 x f_X(x)\,dx \\ = &\int_1^2 \lambda x \exp(-\lambda x)\,dx = \frac{1}{\lambda}\int_\lambda^{2\lambda} t\exp(-t)\,dt \\ = &\frac{1}{\lambda}\left\{[-t\exp(-t)]_\lambda^{2\lambda} + \int_\lambda^{2\lambda}\exp(-t)\,dt\right\} \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\lambda}\left\{-2\lambda\exp\left(-2\lambda\right)+\lambda\exp\left(-\lambda\right)+\left[-\exp\left(-t\right)\right]_{\lambda}^{2\lambda}\right\}\\
&= \frac{1}{\lambda}\left\{-2\lambda\exp\left(-2\lambda\right)+\lambda\exp\left(-\lambda\right)-\exp\left(-2\lambda\right)+\exp\left(-\lambda\right)\right\}\\
&= \frac{1}{\lambda}\left\{\left(-2\lambda-1\right)\exp\left(-2\lambda\right)+\left(\lambda+1\right)\exp\left(-\lambda\right)\right\}
\end{aligned}
$$

Chapter 18

Expected value and the Lebesgue integral

The Lebesgue integral is used to give a completely general definition of expected value. This lecture introduces the Lebesgue integral, first in an intuitive manner and then in a more rigorous manner. Understanding the material presented in this lecture is not necessary to understand the material presented in subsequent lectures.

18.1 Intuition

Let us recall the informal definition of expected value we have given in the lecure entitled *Expected Value* (p. 127):

Definition 102 *The **expected value** of a random variable X is the weighted average of the values that X can take on, where each possible value is weighted by its respective probability.*

When X is discrete and can take on only finitely many values, it is straightforward to compute the expected value of X, by just applying the above definition. Denote by $x_1, \ldots, x_n$ the n values that X can take on (the n elements of its support). Let Ω be the sample space on which X is defined. Also define the following events:

$$\begin{aligned} E_1 &= \{\omega \in \Omega : X(\omega) = x_1\} \\ &\vdots \\ E_n &= \{\omega \in \Omega : X(\omega) = x_n\} \end{aligned}$$

i.e. when the event E_i happens, then X equals x_i.

We can write the expected value of X as:

$$\mathrm{E}[X] = \sum_{i=1}^{n} x_i \mathrm{P}(E_i)$$

i.e. the expected value of X is the weighted average of the values that X can take on, where each possible value x_i is weighted by its respective probability $\mathrm{P}(E_i)$.

Note that this is a way of expressing the expected value that uses neither $F_X(x)$, the distribution function[1] of X, nor its probability mass function[2] $p_X(x)$. Instead, the above way of expressing the expected value uses only the probability $\mathrm{P}(E)$ defined on the events $E \subseteq \Omega$. In many applications, it turns out that this is a very convenient way of expressing (and calculating) the expected value: for example, when the distribution function $F_X(x)$ is not directly known and it is difficult to derive, it is sometimes easier to directly compute the probabilities $\mathrm{P}(E)$ defined on the events $E \subseteq \Omega$. Below, this will be illustrated with an example.

When X is discrete, but can take on infinitely many values, in a similar fashion we can write:

$$\mathrm{E}[X] = \sum_{i=1}^{\infty} x_i \mathrm{P}(E_i) \tag{18.1}$$

In this case, however, there is a possibility that $\mathrm{E}[X]$ is not well-defined: this happens when the infinite series above does not converge, i.e. when the limit

$$\lim_{n \to \infty} \sum_{i=1}^{n} x_i \mathrm{P}(E_i)$$

does not exist. In the next section we will show how to take care of this possibility.

In the case in which X in not discrete (its support has the power of the continuum), things are much more complicated. In this case, the summation in (18.1) does not make any sense, because the support of X cannot be arranged into a sequence[3] and so there is no sequence over which we can sum. Thus, we have to find a workaround. The workaround is similar to the one we have discussed in the presentation of the Stieltjes integral[4]: first, we build a simpler random variable Y that is a good approximation of X and whose expected value can easily be computed; then, we make the approximation better and better; finally, we define the expected value of X to be equal to the expected value of Y when the approximation tends to become perfect.

How does the approximation work, intuitively? We illustrate it in three steps:

1. in the first step, we partition the sample space Ω into n events $E_1, \ldots, E_n$, such that $E_i \cap E_j = \emptyset$ for $i \neq j$ and

$$\Omega = \bigcup_{i=1}^{n} E_i$$

2. in the second step we find, for each event E_i, the smallest value that X can take on when the event E_i happens:

$$y_i = \inf \{X(\omega) : \omega \in E_i\}$$

3. in the third step, we define the random variable Y (which approximates X) as follows:

$$Y(\omega) = \begin{cases} y_1 & \text{when } \omega \in E_1 \\ \vdots & \\ y_n & \text{when } \omega \in E_n \end{cases}$$

[1] See p. 108.
[2] See p. 106.
[3] See p. 31.
[4] See p. 130.

In this way, we have built a random variable Y such that $Y(\omega) \leq X(\omega)$ for any ω. The finer the partition $E_1, \ldots, E_n$ is, the better the approximation is: intuitively, when the sets E_i become smaller, then y_i becomes on average closer to the values that X can take on when E_i happens.

The expected value of Y is, of course, easy to compute:

$$\mathrm{E}[Y] = \sum_{i=1}^{n} y_i \mathrm{P}(E_i)$$

The expected value of X is defined as follows:

$$\mathrm{E}[X] = \lim_{Y \to X} \mathrm{E}[Y] = \lim_{Y \to X} \sum_{i=1}^{n} y_i \mathrm{P}(E_i)$$

where the notation $Y \to X$ means that Y becomes a better approximation of X (because the partition $E_1, \ldots, E_n$ is made finer).

Several equivalent integral notations are used to denote the above limit:

$$\mathrm{E}[X] = \int X d\mathrm{P} = \int X(\omega) d\mathrm{P}(\omega) = \int X(\omega) \mathrm{P}(d\omega)$$

and the integral is called the Lebesgue integral of X with respect to the probability measure P. The notation $d\mathrm{P}$ (or $d\omega$) indicates that the sets E_i become very small by improving the approximation (making the partition $E_1, \ldots, E_n$ finer); the integral notation $\int$ can be thought of as a shorthand for $\lim_{Y \to X} \sum_{i=1}^{n}$; X appears in place of Y in the integral, because the two tend to coincide when the approximation becomes better and better.

18.2 Linearity of the Lebesgue integral

An important property enjoyed by the Lebesgue integral is **linearity**:

Proposition 103 *Let X_1 and X_2 be two random variables defined on a sample space Ω and let $c_1, c_2 \in \mathbb{R}$ be two constants. Then:*

$$\int (c_1 X_1 + c_2 X_2) d\mathrm{P} = c_1 \int X_1 d\mathrm{P} + c_2 \int X_2 d\mathrm{P}$$

The next example shows an important application of the linearity of the Lebesgue integral. The example also shows how the Lebesgue integral can, in certain situations, be much simpler to use than the Stieltjes integral when computing the expected value of a random variable.

Example 104 *Let X_1 and X_2 be two random variables defined on a sample space Ω. We want to define (and compute) the expected value of the sum $X_1 + X_2$. Define a new random variable*

$$Z = X_1 + X_2$$

Using the Stieltjes integral[5], the expected value is defined as follows:

$$\mathrm{E}[Z] = \int_{-\infty}^{\infty} z dF_Z(z)$$

[5] See p. 130.

where $F_Z(z)$ is the distribution function of Z. Hence, to compute the above integral, we first need to know the distribution function of Z (which might be extremely difficult to derive). Using the Lebesgue integral, the expected value is defined as follows:

$$\mathrm{E}[Z] = \int Z d\mathrm{P}$$

However, by linearity of the Lebesgue integral, we obtain:

$$\mathrm{E}[Z] = \int Z d\mathrm{P} = \int (X_1 + X_2) d\mathrm{P} = \int X_1 d\mathrm{P} + \int X_2 d\mathrm{P} = \mathrm{E}[X_1] + \mathrm{E}[X_2]$$

Thus, to compute the expected value of Z, we do not need to know the distribution function of Z, but we only need to know the expected values of X_1 and X_2.

This is just an example of how linearity of the Lebesgue integral translates into **linearity of the expected value**. The more general property is summarized by the following:

Proposition 105 *Let X_1 and X_2 be two random variables defined on a sample space Ω and let $c_1, c_2 \in \mathbb{R}$ be two constants. Then:*

$$\mathrm{E}[c_1 X_1 + c_2 X_2] = c_1 \mathrm{E}[X_1] + c_2 \mathrm{E}[X_2]$$

18.3 A more rigorous definition

A more rigorous definition of the Lebesgue integral requires that we introduce the notion of a **simple random variable**.

Definition 106 *A random variable Y is called simple if and only if it takes on finitely many positive values. In this case:*

- *there exist n events $E_1, \ldots, E_n$ such that $E_i \cap E_j = \emptyset$ for $i \neq j$ and the sample space can be written as*

$$\Omega = \bigcup_{i=1}^{n} E_i$$

- *Y can be written as*

$$Y(\omega) = \begin{cases} y_1 & \text{when } \omega \in E_1 \\ \vdots & \\ y_n & \text{when } \omega \in E_n \end{cases}$$

- *$y_i \geq 0$ for all i.*

Note that a simple random variable is also a discrete random variable. Hence, the expected value of a simple random variable is easy to compute, because it is just the weighted sum of the elements of its support.

The Lebesgue integral of a simple random variable Y is defined to be equal to its expected value:

$$\int Y d\mathrm{P} = \sum_{i=1}^{n} y_i \mathrm{P}(E_i)$$

Let X be the random variable whose integral we want to compute. Let X^+ and X^- be the positive and negative part of X respectively:

$$\begin{aligned} X^+(\omega) &= \max(X(\omega), 0) \text{ for any } \omega \\ X^-(\omega) &= -\min(X(\omega), 0) \text{ for any } \omega \end{aligned}$$

Note that $X^+(\omega) \geq 0$, $X^-(\omega) \geq 0$ for any ω and:

$$X = X^+ - X^-$$

The Lebesgue integral of X^+ is defined as follows:

$$\int X^+ d\mathrm{P} = \sup\left\{ \int Y d\mathrm{P} : Y \text{ is simple and } Y(\omega) \leq X^+(\omega) \text{ for all } \omega \in \Omega \right\}$$

In words, the Lebesgue integral of X^+ is obtained by taking the supremum over the Lebesgue integrals of all the simple functions Y that are less than X^+.

The Lebesgue integral of X^- is defined as follows:

$$\int X^- d\mathrm{P} = \sup\left\{ \int Y d\mathrm{P} : Y \text{ is simple and } Y(\omega) \leq X^-(\omega) \text{ for all } \omega \in \Omega \right\}$$

Finally, the Lebesgue integral of X is defined as the difference between the integrals of its positive and negative parts:

$$\int X d\mathrm{P} = \int X^+ d\mathrm{P} - \int X^- d\mathrm{P}$$

provided the difference makes sense; in case both $\int X^+ d\mathrm{P}$ and $\int X^- d\mathrm{P}$ are equal to infinity, then the difference is not well-defined and we say that X is not integrable.

Chapter 19

Properties of the expected value

This lecture discusses some fundamental properties of the expected value operator. Although most of these properties can be understood and proved using the material presented in previous lectures, some properties are gathered here for convenience, but can be proved and understood only after reading the material presented in subsequent lectures.

19.1 Linearity of the expected value

The following properties are related to the linearity of the expected value.

19.1.1 Scalar multiplication of a random variable

If X is a random variable and $a \in \mathbb{R}$ is a constant, then

$$\mathrm{E}\left[aX\right] = a\mathrm{E}\left[X\right]$$

This property has already been discussed in the lecture entitled *Expected value* (p. 134).

Example 107 *Let X be a random variable with expected value*

$$\mathrm{E}\left[X\right] = 3$$

and Y be a random variable defined as follows:

$$Y = 2X$$

Then:

$$\mathrm{E}\left[Y\right] = \mathrm{E}\left[2X\right] = 2\mathrm{E}\left[X\right] = 2 \cdot 3 = 6$$

19.1.2 Sums of random variables

If X_1, X_2, ..., X_K are K random variables, then:

$$\mathrm{E}\left[X_1 + X_2 + \ldots + X_K\right] = \mathrm{E}\left[X_1\right] + \mathrm{E}\left[X_2\right] + \ldots + \mathrm{E}\left[X_K\right]$$

Also this property has already been discussed in the lecture entitled *Expected value* (p. 134).

Example 108 *Let X and Y be two random variables with expected values*

$$\begin{aligned} \mathrm{E}[X] &= 2 \\ \mathrm{E}[Y] &= 5 \end{aligned}$$

and Z be a random variable defined as follows:

$$Z = X + Y$$

Then

$$\mathrm{E}[Z] = \mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y] = 2 + 5 = 7$$

19.1.3 Linear combinations of random variables

If $X_1, X_2, \ldots, X_K$ are K random variables and $a_1, a_2, \ldots, a_K \in \mathbb{R}$ are K constants, then

$$\mathrm{E}[a_1 X_1 + a_2 X_2 + \ldots + a_K X_K] = a_1 \mathrm{E}[X_1] + a_2 \mathrm{E}[X_2] + \ldots + a_K \mathrm{E}[X_K]$$

This can be trivially obtained combining the two properties above (scalar multiplication and sum). Considering $a_1, a_2, \ldots, a_K$ as the K entries of a $1 \times K$ vector a and X_1, X_2, ..., X_K as the K entries of a $K \times 1$ random vector X, the above property can be written as

$$\mathrm{E}[aX] = a\mathrm{E}[X]$$

which is a multivariate generalization of the scalar multiplication property above.

Example 109 *Let X and Y be two random variables with expected values*

$$\begin{aligned} \mathrm{E}[X] &= 1 \\ \mathrm{E}[Y] &= 4 \end{aligned}$$

and Z be a random variable defined as follows:

$$Z = X + 3Y$$

Then

$$\begin{aligned} \mathrm{E}[Z] &= \mathrm{E}[X + 3Y] = \mathrm{E}[X] + 3\mathrm{E}[Y] \\ &= 1 + 3 \cdot 4 = 1 + 12 = 13 \end{aligned}$$

19.1.4 Addition of a constant and a random matrix

Let Σ be a $K \times L$ random matrix[1], i.e. a $K \times L$ matrix whose entries are random variables. If A is a $K \times L$ matrix of constants, then

$$\mathrm{E}[A + \Sigma] = A + \mathrm{E}[\Sigma]$$

This is easily proved by applying the linearity properties above to each entry of the random matrix $A + \Sigma$.

[1] See p. 119

Note that a **random vector is just a particular instance of a random matrix**. So, if X is a $K \times 1$ random vector and a is a $K \times 1$ vector of constants, then

$$\mathrm{E}[a + X] = a + \mathrm{E}[X]$$

Example 110 *Let X be a 2×1 random vector such that its two entries X_1 and X_2 have expected values*

$$\begin{aligned} \mathrm{E}[X_1] &= 0 \\ \mathrm{E}[X_2] &= 2 \end{aligned}$$

Let A be the following 2×1 constant vector:

$$A = \begin{bmatrix} 1 \\ 7 \end{bmatrix}$$

Let the random vector Y be defined as follows:

$$Y = A + X$$

Then

$$\begin{aligned} \mathrm{E}[Y] &= \mathrm{E}[A + X] = A + \mathrm{E}[X] \\ &= \begin{bmatrix} 1 \\ 7 \end{bmatrix} + \begin{bmatrix} \mathrm{E}[X_1] \\ \mathrm{E}[X_2] \end{bmatrix} \\ &= \begin{bmatrix} 1 \\ 7 \end{bmatrix} + \begin{bmatrix} 0 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ 9 \end{bmatrix} \end{aligned}$$

19.1.5 Multiplication of a constant and a random matrix

Let Σ be a $K \times L$ random matrix, i.e. a $K \times L$ matrix whose entries are random variables. If B is a $M \times K$ matrix of constants, then

$$\mathrm{E}[B\Sigma] = B\mathrm{E}[\Sigma]$$

If C is a a $L \times N$ matrix of constants, then

$$\mathrm{E}[\Sigma C] = \mathrm{E}[\Sigma] C$$

These are immediate consequences of the linearity properties above.

By iteratively applying this property, if B is a $M \times K$ matrix of constants and C is a a $L \times N$ matrix of constants, we obtain

$$\mathrm{E}[B\Sigma C] = \mathrm{E}[B(\Sigma C)] = B\mathrm{E}[\Sigma C] = B\mathrm{E}[\Sigma] C$$

Example 111 *Let X be a 1×2 random vector such that*

$$\mathrm{E}[X_1] = \mathrm{E}[X_2] = 3$$

where X_1 and X_2 are the two components of X. Let A be the following 2×2 matrix of constants:

$$A = \begin{bmatrix} 2 & 0 \\ 3 & 1 \end{bmatrix}$$

Let the random vector Y be defined as follows:

$$Y = XA$$

Then

$$\begin{aligned} \mathrm{E}[Y] &= \mathrm{E}[XA] = \mathrm{E}[X]A \\ &= \begin{bmatrix} \mathrm{E}[X_1] & \mathrm{E}[X_2] \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 3 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 3 & 3 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 3 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 3 \cdot 2 + 3 \cdot 3 & 3 \cdot 0 + 3 \cdot 1 \end{bmatrix} \\ &= \begin{bmatrix} 15 & 3 \end{bmatrix} \end{aligned}$$

19.2 Other properties

The following properties of the expected value are also very important.

19.2.1 Expectation of a positive random variable

Let X be an integrable[2] random variable defined on a sample space Ω. Let X be a positive random variable, i.e.

$$X(\omega) \geq 0 \ , \forall \omega \in \Omega$$

Then

$$\mathrm{E}[X] \geq 0$$

Proof. Intuitively, this is obvious: the expected value of X is a weighted average of the values that X can take on; X can take on only positive values; therefore, also its expected value must be positive. Formally, the expected value is the Lebesgue integral[3] of X; X can be approximated to any degree of accuracy by positive simple random variables whose Lebesgue integral is obviously positive; therefore, also the Lebesgue integral of X must be positive. ■

19.2.2 Preservation of almost sure inequalities

Let X and Y be two integrable random variables defined on a sample space Ω. Let X and Y be such that $X \leq Y$ almost surely[4]. Then

$$\mathrm{E}[X] \leq \mathrm{E}[Y]$$

Proof. Let E be a zero-probability event such that

$$\{\omega \in \Omega : X(\omega) > Y(\omega)\} \subseteq E$$

[2] See p. 136.

[3] See p. 141.

[4] In other words, there exists a zero-probability event E such that $\{\omega \in \Omega : X(\omega) > Y(\omega)\} \subseteq E$. See the lecture entitled *Zero-probability events* (p. 79) for a definition of zero-probability event and of almost sure property.

First note that

$$1_E + 1_{E^c} = 1$$

where 1_E is the indicator[5] of the event E and 1_{E^c} is the indicator of the complement of E. We can write

$$\begin{aligned} \mathrm{E}[Y - X] &= \mathrm{E}[(Y - X) \cdot 1] \\ &= \mathrm{E}[(Y - X) \cdot 1_E] + \mathrm{E}[(Y - X) \cdot 1_{E^c}] \\ &= \mathrm{E}[(Y - X) \cdot 1_{E^c}] \end{aligned} \tag{19.1}$$

because

$$\mathrm{E}[(Y - X) \cdot 1_E] = 0$$

by the properties of indicators of zero-probability events. If $\omega \in E^c$, then

$$(Y - X) \geq 0$$

and

$$(Y - X) \cdot 1_{E^c} \geq 0$$

On the contrary, if $\omega \in E$, then

$$1_{E^c} = 0$$

and

$$(Y - X) \cdot 1_{E^c} = 0$$

Therefore:

$$(Y - X) \cdot 1_{E^c} \geq 0 \ ,\forall \omega \in \Omega$$

which means that $(Y - X) \cdot 1_{E^c}$ is a positive random variable. Thus

$$\mathrm{E}[(Y - X) \cdot 1_{E^c}] \geq 0 \tag{19.2}$$

because the expectation of a positive random variable is positive (see 19.2.1). Putting together (19.1) and (19.2), we obtain

$$\mathrm{E}[Y - X] \geq 0$$

By linearity of the expected value:

$$\mathrm{E}[Y - X] = \mathrm{E}[Y] - \mathrm{E}[X] \geq 0$$

Therefore:

$$\mathrm{E}[X] \leq \mathrm{E}[Y]$$

■

19.3 Solved exercises

Below you can find some exercises with explained solutions.

[5] See p. 197.

Exercise 1

Let X and Y be two random variables, having expected values

$$\begin{aligned} \mathrm{E}[X] &= \sqrt{2} \\ \mathrm{E}[Y] &= 1 \end{aligned}$$

Compute the expected value of the random variable Z defined as follows:

$$Z = \sqrt{2}X + Y$$

Solution

Using the linearity of the expected value operator, we obtain

$$\begin{aligned} \mathrm{E}[Z] &= \mathrm{E}\left[\sqrt{2}X + Y\right] = \sqrt{2}\mathrm{E}[X] + \mathrm{E}[Y] \\ &= \sqrt{2}\sqrt{2} + 1 = 2 + 1 = 3 \end{aligned}$$

Exercise 2

Let X be a 2×1 random vector such that its two entries X_1 and X_2 have expected values:

$$\begin{aligned} \mathrm{E}[X_1] &= 2 \\ \mathrm{E}[X_2] &= 3 \end{aligned}$$

Let A be the following 2×2 matrix of constants:

$$A = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$$

Compute the expected value of the random vector Y defined as follows:

$$Y = AX$$

Solution

The linearity property of the expected value applies also to the multiplication of a constant matrix and a random vector:

$$\begin{aligned} \mathrm{E}[Y] &= \mathrm{E}[AX] = A\mathrm{E}[X] \\ &= \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mathrm{E}[X_1] \\ \mathrm{E}[X_2] \end{bmatrix} \\ &= \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \end{bmatrix} \\ &= \begin{bmatrix} 1 \cdot 2 + 2 \cdot 3 \\ 0 \cdot 2 + 1 \cdot 3 \end{bmatrix} = \begin{bmatrix} 8 \\ 3 \end{bmatrix} \end{aligned}$$

Exercise 3

Let Σ be a 2×2 matrix with random entries, such that all its entries have expected value equal to 1. Let A be the following 1×2 constant vector:

$$A = \begin{bmatrix} 2 & 3 \end{bmatrix}$$

Compute the expected value of the random vector Y defined as follows:

$$Y = A\Sigma$$

Solution

The linearity property of the expected value applies also to the multiplication of a constant vector and a matrix with random entries:

$$\begin{aligned} \mathrm{E}[Y] &= \mathrm{E}[A\Sigma] = A\mathrm{E}[\Sigma] \\ &= \begin{bmatrix} 2 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 2 \cdot 1 + 3 \cdot 1 & 2 \cdot 1 + 3 \cdot 1 \end{bmatrix} \\ &= \begin{bmatrix} 5 & 5 \end{bmatrix} \end{aligned}$$

Chapter 20

Variance

This lecture introduces the concept of variance. Before reading this lecture, make sure you are familiar with the concept of random variable (see the lecture entitled *Random variables* - p. 105) and with the concept of expected value (see the lecture entitled *Expected value* - p. 127).

20.1 Definition of variance

The following is a definition of variance.

Definition 112 *Let X be a random variable. The* ***variance*** *of X, denoted by* $\operatorname{Var}[X]$*, is defined as*

$$\operatorname{Var}[X] = \operatorname{E}\left[(X - \operatorname{E}[X])^2\right] \tag{20.1}$$

provided the expected values in (20.1) exist and are well-defined.

The variance of X is also called the **second central moment** of X.

20.2 Interpretation of variance

Variance is a measure of the dispersion of a random variable around its mean. Being the expected value of a squared number, variance is always positive. When a random variable X is constant (whatever happens, it always takes on the same value), then its variance is zero, because X is always equal to its expected value $\operatorname{E}[X]$. On the contrary, the larger are the possible deviations of X from its expected value $\operatorname{E}[X]$, the larger the variance of X is.

20.3 Computation of variance

To better understand how variance is computed, you can break up its computation in several steps:

1. compute $\operatorname{E}[X]$, the expected value of X;

2. construct a random variable Y that measures how much the realizations of X deviate from their expected value:

$$Y = X - \mathrm{E}[X]$$

3. take the square of Y, so that positive and negative deviations from the mean having the same magnitude yield the same measure of distance from $\mathrm{E}[X]$;

4. finally, compute the expected value of the squared deviation Y^2 to know how much on average X deviates from $\mathrm{E}[X]$:

$$\mathrm{Var}[X] = \mathrm{E}\left[Y^2\right] = \mathrm{E}\left[(X - \mathrm{E}[X])^2\right]$$

20.4 Variance formula

The following is a very important formula for computing variance.

Proposition 113 *The variance of a random variable X can be expressed as*

$$\mathrm{Var}[X] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2$$

Proof. The variance formula is derived as follows. First, expand the square:

$$\mathrm{Var}[X] = \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] = \mathrm{E}\left[X^2 + \mathrm{E}[X]^2 - 2\mathrm{E}[X]X\right]$$

Then, by exploiting the linearity of the expected value[1], you obtain

$$\begin{aligned}\mathrm{Var}[X] &= \mathrm{E}\left[X^2 + \mathrm{E}[X]^2 - 2\mathrm{E}[X]X\right] \\ &= \mathrm{E}\left[X^2\right] + \mathrm{E}[X]^2 - 2\mathrm{E}[X]\mathrm{E}[X] \\ &= \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2\end{aligned}$$

■

The above variance formula also makes clear that variance exists and is well-defined only as long as $\mathrm{E}[X]$ and $\mathrm{E}\left[X^2\right]$ exist and are well-defined.

20.5 Example

The following example shows how to compute the variance of a discrete random variable[2] using both the definition and the variance formula above.

Example 114 *Let X be a discrete random variable with support*

$$R_X = \{0, 1\}$$

and probability mass function

$$p_X(x) = \begin{cases} q & \text{if } x = 1 \\ 1 - q & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases}$$

[1] See p. 134.
[2] See p. 106.

where $0 \leq q \leq 1$. *Its expected value is*

$$\mathrm{E}[X] = 1 \cdot p_X(1) + 0 \cdot p_X(0) = 1 \cdot q + 0 \cdot (1-q) = q$$

The expected value of its square is

$$\mathrm{E}\left[X^2\right] = 1^2 \cdot p_X(1) + 0^2 \cdot p_X(0) = 1 \cdot q + 0 \cdot (1-q) = q$$

Its variance is

$$\mathrm{Var}[X] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 = q - q^2 = q(1-q)$$

Alternatively, we can compute the variance of X *using the definition. Define a new random variable, the squared deviation of* X *from* $\mathrm{E}[X]$, *as*

$$Z = (X - \mathrm{E}[X])^2$$

The support of Z *is*

$$R_Z = \left\{(1-q)^2, q^2\right\}$$

and its probability mass function is

$$p_Z(z) = \begin{cases} q & \text{if } z = (1-q)^2 \\ 1-q & \text{if } z = q^2 \\ 0 & \text{otherwise} \end{cases}$$

The variance of X *equals the expected value of* Z*:*

$$\begin{aligned} \mathrm{Var}[X] &= \mathrm{E}[Z] = (1-q)^2 \cdot p_Z\left((1-q)^2\right) + q^2 \cdot p_Z\left(q^2\right) \\ &= (1-q)^2 \cdot q + q^2 \cdot (1-q) \\ &= (1-q)\, q\, [(1-q) + q] = (1-q)\, q \end{aligned}$$

The exercises at the end of this lecture provide more examples of how variance can be computed.

20.6 More details

The following subsections contain more details on variance.

20.6.1 Variance and standard deviation

The square root of variance is called **standard deviation**. The standard deviation of a random variable X is usually denoted by $\mathrm{std}[X]$ or by $\mathrm{stdev}[X]$:

$$\mathrm{stdev}[X] = \sqrt{\mathrm{Var}[X]}$$

20.6.2 Addition to a constant

Adding a constant to a random variable does not change its variance.

Proposition 115 *Let* $a \in \mathbb{R}$ *be a constant and let* X *be a random variable. Then,*

$$\mathrm{Var}[a + X] = \mathrm{Var}[X]$$

Proof. This is proved as follows:

$$\begin{aligned}\operatorname{Var}[a+X] &= \operatorname{E}\left[(a+X-\operatorname{E}[a+X])^2\right] \\ &= \operatorname{E}\left[(a+X-a-\operatorname{E}[X])^2\right] \\ &= \operatorname{E}\left[(X-\operatorname{E}[X])^2\right] \\ &= \operatorname{Var}[X]\end{aligned}$$

where we have used the fact that, by linearity of the expected value, it holds that

$$\operatorname{E}[a+X] = a + \operatorname{E}[X]$$

■

20.6.3 Multiplication by a constant

When a random variable is multiplied by a constant, its variance is multiplied by the square of that constant.

Proposition 116 *Let $b \in \mathbb{R}$ be a constant and let X be a random variable. Then,*

$$\operatorname{Var}[bX] = b^2 \operatorname{Var}[X]$$

Proof. This is proved as follows:

$$\begin{aligned}\operatorname{Var}[bX] &= \operatorname{E}\left[(bX-\operatorname{E}[bX])^2\right] \\ &= \operatorname{E}\left[(bX-b\operatorname{E}[X])^2\right] \\ &= \operatorname{E}\left[b^2(X-\operatorname{E}[X])^2\right] \\ &= b^2\operatorname{E}\left[(X-\operatorname{E}[X])^2\right] \\ &= b^2\operatorname{Var}[X]\end{aligned}$$

where we have used the fact that, by linearity of the expected value, it holds that

$$\operatorname{E}[bX] = b\operatorname{E}[X]$$

■

20.6.4 Linear transformations

By combining the previous two properties, we obtain the following proposition.

Proposition 117 *Let $a, b \in \mathbb{R}$ be two constants and let X be a random variable. Then,*

$$\operatorname{Var}[a+bX] = b^2 \operatorname{Var}[X]$$

20.6.5 Square integrability

If $\mathrm{E}\left[X^2\right]$ exists and is finite, we say that X is a **square integrable random variable**, or just that X is **square integrable**. It can easily be proved that, if X is square integrable then X is also integrable[3], that is, $\mathrm{E}\left[|X|\right]$ exists and is finite.

Thus, if X is square integrable, also its variance

$$\mathrm{Var}\left[X\right] = \mathrm{E}\left[X^2\right] - \mathrm{E}\left[X\right]^2$$

exists and is finite.

20.7 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Let X be a discrete random variable with support

$$R_X = \{0, 1, 2, 3\}$$

and probability mass function

$$p_X\left(x\right) = \begin{cases} 1/4 & \text{if } x \in R_X \\ 0 & \text{otherwise} \end{cases}$$

Compute its variance.

Solution

The expected value of X is

$$\begin{aligned} \mathrm{E}\left[X\right] &= \sum_{x \in R_X} x p_X\left(x\right) \\ &= 0 \cdot p_X\left(0\right) + 1 \cdot p_X\left(1\right) + 2 \cdot p_X\left(2\right) + 3 \cdot p_X\left(3\right) \\ &= 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{4} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{4} = \frac{6}{4} = \frac{3}{2} \end{aligned}$$

The expected value of X^2 is

$$\begin{aligned} \mathrm{E}\left[X^2\right] &= \sum_{x \in R_X} x^2 p_X\left(x\right) \\ &= 0^2 \cdot p_X\left(0\right) + 1^2 \cdot p_X\left(1\right) + 2^2 \cdot p_X\left(2\right) + 3^2 \cdot p_X\left(3\right) \\ &= 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{4} + 4 \cdot \frac{1}{4} + 9 \cdot \frac{1}{4} = \frac{14}{4} = \frac{7}{2} \end{aligned}$$

The variance of X is

$$\begin{aligned} \mathrm{Var}\left[X\right] &= \mathrm{E}\left[X^2\right] - \mathrm{E}\left[X\right]^2 = \frac{7}{2} - \left(\frac{3}{2}\right)^2 \\ &= \frac{14}{4} - \frac{9}{4} = \frac{5}{4} \end{aligned}$$

[3] See p. 136.

Exercise 2

Let X be a discrete random variable with support

$$R_X = \{1, 2, 3, 4\}$$

and probability mass function

$$p_X(x) = \begin{cases} \frac{1}{30}x^2 & \text{if } x \in R_X \\ 0 & \text{otherwise} \end{cases}$$

Compute its variance.

Solution

The expected value of X is

$$\begin{aligned} \mathrm{E}[X] &= \sum_{x \in R_X} x p_X(x) \\ &= 1 \cdot p_X(1) + 2 \cdot p_X(2) + 3 \cdot p_X(3) + 4 \cdot p_X(4) \\ &= 1 \cdot \frac{1}{30} + 2 \cdot \frac{4}{30} + 3 \cdot \frac{9}{30} + 4 \cdot \frac{16}{30} \\ &= \frac{1}{30}(1 + 8 + 27 + 64) = \frac{100}{30} = \frac{10}{3} \end{aligned}$$

The expected value of X^2 is

$$\begin{aligned} \mathrm{E}\left[X^2\right] &= \sum_{x \in R_X} x^2 p_X(x) \\ &= 1^2 \cdot p_X(1) + 2^2 \cdot p_X(2) + 3^2 \cdot p_X(3) + 4^2 \cdot p_X(4) \\ &= 1 \cdot \frac{1}{30} + 4 \cdot \frac{4}{30} + 9 \cdot \frac{9}{30} + 16 \cdot \frac{16}{30} \\ &= \frac{1}{30}(1 + 16 + 81 + 256) = \frac{354}{30} = \frac{118}{10} = \frac{59}{5} \end{aligned}$$

The variance of X is

$$\begin{aligned} \mathrm{Var}[X] &= \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 = \frac{59}{5} - \left(\frac{10}{3}\right)^2 \\ &= \frac{59 \cdot 9 - 100 \cdot 5}{45} = \frac{531 - 500}{45} = \frac{31}{45} \end{aligned}$$

Exercise 3

Read and try to understand how the variance of a Poisson random variable is derived in the lecture entitled *Poisson distribution* (p. 349).

Exercise 4

Let X be an absolutely continuous random variable[4] with support

$$R_X = [0, 1]$$

[4] See p. 107.

and probability density function

$$f_X(x) = \begin{cases} 1 & \text{if } x \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

Compute its variance.

Solution

The expected value of X is

$$\begin{aligned} \mathrm{E}[X] &= \int_{-\infty}^{\infty} x f_X(x)\, dx = \int_0^1 x dx \\ &= \left[\frac{1}{2}x^2\right]_0^1 = \frac{1}{2} - 0 = \frac{1}{2} \end{aligned}$$

The expected value of X^2 is

$$\begin{aligned} \mathrm{E}\left[X^2\right] &= \int_{-\infty}^{\infty} x^2 f_X(x)\, dx = \int_0^1 x^2 dx \\ &= \left[\frac{1}{3}x^3\right]_0^1 = \frac{1}{3} - 0 = \frac{1}{3} \end{aligned}$$

The variance of X is

$$\begin{aligned} \mathrm{Var}[X] &= \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 = \frac{1}{3} - \left(\frac{1}{2}\right)^2 \\ &= \frac{4}{12} - \frac{3}{12} = \frac{1}{12} \end{aligned}$$

Exercise 5

Let X be an absolutely continuous random variable with support

$$R_X = [0,1]$$

and probability density function

$$f_X(x) = \begin{cases} 3x^2 & \text{if } x \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

Compute its variance.

Solution

The expected value of X is

$$\begin{aligned} \mathrm{E}[X] &= \int_{-\infty}^{\infty} x f_X(x)\, dx = \int_0^1 x 3x^2 dx = \int_0^1 3x^3 dx \\ &= \left[\frac{3}{4}x^4\right]_0^1 = \frac{3}{4} - 0 = \frac{3}{4} \end{aligned}$$

The expected value of X^2 is

$$\begin{aligned} \mathrm{E}\left[X^2\right] &= \int_{-\infty}^{\infty} x^2 f_X(x)\,dx = \int_0^1 x^2 3x^2 dx = \int_0^1 3x^4 dx \\ &= \left[\frac{3}{5}x^5\right]_0^1 = \frac{3}{5} - 0 = \frac{3}{5} \end{aligned}$$

The variance of X is

$$\begin{aligned} \mathrm{Var}[X] &= \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 = \frac{3}{5} - \left(\frac{3}{4}\right)^2 \\ &= \frac{3}{5} - \frac{9}{16} = \frac{48-45}{80} = \frac{3}{80} \end{aligned}$$

Exercise 6

Read and try to understand how the variance of a Chi-square random variable is derived in the lecture entitled *Chi-square distribution* (p. 387).

Chapter 21

Covariance

This lecture introduces the concept of covariance. Before reading this lecture, make sure you are familiar with the concepts of random variable (p. 105), expected value (p. 127) and variance (p. 155).

21.1 Definition of covariance

Let X and Y be two random variables. The **covariance** between X and Y, denoted by $\operatorname{Cov}[X,Y]$, is defined as follows:

$$\operatorname{Cov}[X,Y] = \operatorname{E}[(X - \operatorname{E}[X])(Y - \operatorname{E}[Y])]$$

provided the above expected value exists and is well-defined.

21.2 Interpretation of covariance

Covariance is a measure of association between two random variables.

To understand the meaning of covariance, let us analyze how it is constructed. It is the expected value of the product $\overline{X}\overline{Y}$, where $\overline{X}$ and $\overline{Y}$ are defined as follows:

$$\begin{aligned} \overline{X} &= X - \operatorname{E}[X] \\ \overline{Y} &= Y - \operatorname{E}[Y] \end{aligned}$$

In other words, $\overline{X}$ and $\overline{Y}$ are the deviations of X and Y from their respective means. When the product $\overline{X}\overline{Y}$ is **positive**, it means that:

- either X and Y are both above their respective means;
- or X and Y are both below their respective means.

On the contrary, when $\overline{X}\overline{Y}$ is **negative**, it means that:

- either X is above its mean and Y is below its mean;
- or X is below its mean and Y is above its mean.

In other words, when $\overline{XY}$ is positive, X and Y are **concordant** (their deviations from the mean have the same sign); when $\overline{XY}$ is negative, X and Y are **discordant** (their deviations from the mean have opposite signs). Since

$$\operatorname{Cov}[X,Y] = \mathrm{E}\left[\overline{XY}\right]$$

a positive covariance means that on average X and Y are concordant; on the contrary, a negative covariance means that on average X and Y are discordant.

Thus, the covariance of X and Y provides a measure of the degree to which X and Y tend to "move together": a positive covariance indicates that the deviations of X and Y from their respective means tend to have the same sign; a negative covariance indicates that deviations of X and Y from their respective means tend to have opposite signs. Intuitively, we could express the concept as follows:

- $\operatorname{Cov}[X,Y] > 0$ implies that X tends to be high when Y is high and low when Y is low;
- $\operatorname{Cov}[X,Y] < 0$ implies that X tends to be high when Y is low and vice versa.

When $\operatorname{Cov}[X,Y] = 0$, X and Y do not display any of the above two tendencies.

21.3 Covariance formula

The following covariance formula is often used to compute the covariance between two random variables.

Proposition 118 *The covariance between two random variables X and Y can be expressed as*

$$\operatorname{Cov}[X,Y] = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]$$

Proof. The formula is derived as follows:

$$\begin{aligned}
\operatorname{Cov}[X,Y] &= \mathrm{E}\left[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\right] \\
&= \mathrm{E}\left[XY - X\mathrm{E}[Y] - \mathrm{E}[X]Y + \mathrm{E}[X]\,\mathrm{E}[Y]\right] \\
\boxed{\mathrm{A}} &= \mathrm{E}[XY] - \mathrm{E}[Y]\,\mathrm{E}[X] - \mathrm{E}[X]\,\mathrm{E}[Y] + \mathrm{E}[X]\,\mathrm{E}[Y] \\
&= \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]
\end{aligned}$$

where: in step $\boxed{\mathrm{A}}$ we have used the linearity of the expected value[1]. ■

This formula also makes clear that $\operatorname{Cov}[X,Y]$ exists and is well-defined only as long as $\mathrm{E}[X]$, $\mathrm{E}[Y]$ and $\mathrm{E}[XY]$ exist and are well-defined.

21.4 Example

The following example shows how to compute the covariance between two discrete random variables.

[1] See p. 134.

Example 119 *Let X be a 2×1 discrete random vector[2] and denote its components by X_1 and X_2. Let the support of X be*

$$R_X = \left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\}$$

and its joint probability mass function be

$$p_X(x) = \begin{cases} 1/3 & \text{if } x = \begin{bmatrix} 1 & 1 \end{bmatrix}^\top \\ 1/3 & \text{if } x = \begin{bmatrix} 2 & 0 \end{bmatrix}^\top \\ 1/3 & \text{if } x = \begin{bmatrix} 0 & 0 \end{bmatrix}^\top \\ 0 & \text{otherwise} \end{cases}$$

The support of X_1 is

$$R_{X_1} = \{0, 1, 2\}$$

and its marginal probability mass function[3] is

$$p_{X_1}(x) = \sum_{\{(x_1,x_2)\in R_X : x_1 = x\}} p_X(x_1, x_2) = \begin{cases} 1/3 & \text{if } x = 0 \\ 1/3 & \text{if } x = 1 \\ 1/3 & \text{if } x = 2 \\ 0 & \text{otherwise} \end{cases}$$

The expected value of X_1 is

$$\mathrm{E}[X_1] = \sum_{x\in R_{X_1}} x p_{X_1}(x) = \frac{1}{3}\cdot 0 + \frac{1}{3}\cdot 1 + \frac{1}{3}\cdot 2 = 1$$

The support of X_2 is

$$R_{X_2} = \{0, 1\}$$

and its marginal probability mass function is

$$p_{X_2}(x) = \sum_{\{(x_1,x_2)\in R_X : x_2 = x\}} p_X(x_1, x_2) = \begin{cases} 2/3 & \text{if } x = 0 \\ 1/3 & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases}$$

The expected value of X_2 is

$$\mathrm{E}[X_2] = \sum_{x\in R_{X_2}} x p_{X_2}(x) = \frac{2}{3}\cdot 0 + \frac{1}{3}\cdot 1 = \frac{1}{3}$$

By using the transformation theorem[4], we can compute the expected value of X_1X_2:

$$\begin{aligned} \mathrm{E}[X_1X_2] &= \sum_{x\in R_X} x_1 x_2 p_X(x_1, x_2) \\ &= (1\cdot 1)\cdot\frac{1}{3} + (2\cdot 0)\cdot\frac{1}{3} + (0\cdot 0)\cdot\frac{1}{3} = \frac{1}{3} \end{aligned}$$

Hence, the covariance between X_1 and X_2 is

$$\mathrm{Cov}[X_1, X_2] = \mathrm{E}[X_1X_2] - \mathrm{E}[X_1]\,\mathrm{E}[X_2] = \frac{1}{3} - 1\cdot\frac{1}{3} = 0$$

[2] See p. 116.
[3] See p. 119.
[4] See p. 134.

21.5 More details

The following subsections contain more details on covariance.

21.5.1 Covariance of a random variable with itself

The covariance of a random variable with itself is equal to its variance.

Proposition 120 *Let* $\mathrm{Cov}[X,X]$ *be the covariance of a random variable with itself. Then,*

$$\mathrm{Cov}[X,X] = \mathrm{Var}[X]$$

Proof. This is proved as follows:

$$\begin{aligned}\mathrm{Cov}[X,X] &= \mathrm{E}[(X-\mathrm{E}[X])(X-\mathrm{E}[X])] \\ &= \mathrm{E}\left[(X-\mathrm{E}[X])^2\right] \\ &= \mathrm{Var}[X]\end{aligned}$$

where in the last step we have used the very definition of variance. ■

21.5.2 Symmetry

The covariance operator is symmetric.

Proposition 121 *Let* $\mathrm{Cov}[X,Y]$ *be the covariance between two random variables* X *and* Y*. Then,*

$$\mathrm{Cov}[X,Y] = \mathrm{Cov}[Y,X]$$

Proof. This is proved as follows:

$$\begin{aligned}\mathrm{Cov}[X,Y] &= \mathrm{E}[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])] \\ &= \mathrm{E}[(Y-\mathrm{E}[Y])(X-\mathrm{E}[X])] \\ &= \mathrm{Cov}[Y,X]\end{aligned}$$

■

21.5.3 Bilinearity

The covariance operator is linear in both of its arguments.

Proposition 122 *Let* a_1 *and* a_2 *be two constants. Let* X_1*,* X_2 *and* Y *be three random variables such that* $\mathrm{Cov}[X_1,Y]$ *and* $\mathrm{Cov}[X_2,Y]$ *exist and are well-defined. Then,*

$$\mathrm{Cov}[a_1X_1+a_2X_2,Y] = a_1\mathrm{Cov}[X_1,Y] + a_2\mathrm{Cov}[X_2,Y]$$

and

$$\mathrm{Cov}[Y,a_1X_1+a_2X_2] = a_1\mathrm{Cov}[Y,X_1] + a_2\mathrm{Cov}[Y,X_2]$$

Proof. That the first argument is linear is proved by using the linearity of the expected value:

$$\begin{aligned} & \mathrm{Cov}\left[a_1X_1 + a_2X_2, Y\right] \\ = \; & \mathrm{E}\left[\left(a_1X_1 + a_2X_2 - \mathrm{E}\left[a_1X_1 + a_2X_2\right]\right)\left(Y - \mathrm{E}\left[Y\right]\right)\right] \\ = \; & \mathrm{E}\left[\left(a_1X_1 - \mathrm{E}\left[a_1X_1\right]\right)\left(Y - \mathrm{E}\left[Y\right]\right) + \left(a_2X_2 - \mathrm{E}\left[a_2X_2\right]\right)\left(Y - \mathrm{E}\left[Y\right]\right)\right] \\ = \; & a_1\mathrm{E}\left[\left(X_1 - \mathrm{E}\left[X_1\right]\right)\left(Y - \mathrm{E}\left[Y\right]\right)\right] + a_2\mathrm{E}\left[\left(X_2 - \mathrm{E}\left[X_2\right]\right)\left(Y - \mathrm{E}\left[Y\right]\right)\right] \\ = \; & a_1\mathrm{Cov}\left[X_1, Y\right] + a_2\mathrm{Cov}\left[X_2, Y\right] \end{aligned}$$

By symmetry (see 21.5.2), also the second argument is linear. ■

Linearity in both the first and second argument is called **bilinearity**.

By iteratively applying the above arguments, one can prove that bilinearity holds also for linear combinations of more than two variables:

$$\begin{aligned} \mathrm{Cov}\left[\sum_{i=1}^{n} a_iX_i, Y\right] &= \sum_{i=1}^{n} a_i\mathrm{Cov}\left[X_i, Y\right] \\ \mathrm{Cov}\left[Y, \sum_{i=1}^{n} a_iX_i\right] &= \sum_{i=1}^{n} a_i\mathrm{Cov}\left[Y, X_i\right] \end{aligned}$$

21.5.4 Variance of the sum of two random variables

The next proposition provides a formula that links the variance of a sum of two random variables to their covariance.

Proposition 123 *Let X_1 and X_2 be two random variables. If the variance of their sum exists and is well-defined, then*

$$\mathrm{Var}\left[X_1 + X_2\right] = \mathrm{Var}\left[X_1\right] + \mathrm{Var}\left[X_2\right] + 2\mathrm{Cov}\left[X_1, X_2\right]$$

Proof. The above formula is derived as follows:

$$\begin{aligned} & \mathrm{Var}\left[X_1 + X_2\right] \\ = \; & \mathrm{E}\left[\left(X_1 + X_2 - \mathrm{E}\left[X_1 + X_2\right]\right)^2\right] \\ = \; & \mathrm{E}\left[\left(\left(X_1 - \mathrm{E}\left[X_1\right]\right) + \left(X_2 - \mathrm{E}\left[X_2\right]\right)\right)^2\right] \\ = \; & \mathrm{E}\left[\left(X_1 - \mathrm{E}\left[X_1\right]\right)^2 + \left(X_2 - \mathrm{E}\left[X_2\right]\right)^2 + 2\left(X_1 - \mathrm{E}\left[X_1\right]\right)\left(X_2 - \mathrm{E}\left[X_2\right]\right)\right] \\ = \; & \mathrm{E}\left[\left(X_1 - \mathrm{E}\left[X_1\right]\right)^2\right] + \mathrm{E}\left[\left(X_2 - \mathrm{E}\left[X_2\right]\right)^2\right] \\ & +2\mathrm{E}\left[\left(X_1 - \mathrm{E}\left[X_1\right]\right)\left(X_2 - \mathrm{E}\left[X_2\right]\right)\right] \\ = \; & \mathrm{Var}\left[X_1\right] + \mathrm{Var}\left[X_2\right] + 2\mathrm{Cov}\left[X_1, X_2\right] \end{aligned}$$

■

Thus, to compute the variance of the sum of two random variables we need to know their covariance.

Obviously then, the formula

$$\mathrm{Var}\left[X_1 + X_2\right] = \mathrm{Var}\left[X_1\right] + \mathrm{Var}\left[X_2\right]$$

holds only when X_1 and X_2 have zero covariance.

21.5.5 Variance of the sum of n random variables

The previous proposition can be generalized as follows.

Proposition 124 *The variance of the sum of n random variables $X_1, \ldots, X_n$, if it exists, can be expressed as*

$$\operatorname{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \operatorname{Var}\left[X_i\right] + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} \operatorname{Cov}\left[X_i, X_j\right] \tag{21.1}$$

Proof. The formula is proved by using the bilinearity of the covariance operator (see 21.5.3):

$$\begin{aligned}
\operatorname{Var}\left[\sum_{i=1}^{n} X_i\right] &= \operatorname{Cov}\left[\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i\right] \\
&= \operatorname{Cov}\left[\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right] \\
&= \sum_{i=1}^{n} \operatorname{Cov}\left[X_i, \sum_{j=1}^{n} X_j\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \operatorname{Cov}\left[X_i, X_j\right] \\
&= \sum_{i=1}^{n} \operatorname{Cov}\left[X_i, X_i\right] + \sum_{i=2}^{n}\sum_{j=1}^{i-1} \operatorname{Cov}\left[X_i, X_j\right] \\
&\quad + \sum_{i=2}^{n}\sum_{j=i+1}^{n} \operatorname{Cov}\left[X_i, X_j\right] \\
&= \sum_{i=1}^{n} \operatorname{Var}\left[X_i\right] + \sum_{i=2}^{n}\sum_{j=1}^{i-1} \operatorname{Cov}\left[X_i, X_j\right] \\
&\quad + \sum_{j=2}^{n}\sum_{i=1}^{j-1} \operatorname{Cov}\left[X_j, X_i\right] \\
&= \sum_{i=1}^{n} \operatorname{Var}\left[X_i\right] + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} \operatorname{Cov}\left[X_i, X_j\right]
\end{aligned}$$

■

Formula (21.1) implies that when all the random variables in the sum have zero covariance with each other, then the variance of the sum is just the sum of the variances:

$$\operatorname{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \operatorname{Var}\left[X_i\right]$$

This is true, for example, when the random variables in the sum are mutually independent[5], because independence implies zero covariance[6].

[5] See p. 233.
[6] See p. 234.

21.6 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Let X be a 2×1 discrete random vector and denote its components by X_1 and X_2. Let the support of X be

$$R_X = \left\{ \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \end{bmatrix} \right\}$$

and its joint probability mass function be

$$p_X(x) = \begin{cases} 2/3 & \text{if } x = \begin{bmatrix} 1 & 3 \end{bmatrix}^\top \\ 1/3 & \text{if } x = \begin{bmatrix} 2 & 1 \end{bmatrix}^\top \\ 0 & \text{otherwise} \end{cases}$$

Compute the covariance between X_1 and X_2.

Solution

The support of X_1 is

$$R_{X_1} = \{1, 2\}$$

and its marginal probability mass function[7] is

$$p_{X_1}(x) = \sum_{\{(x_1,x_2) \in R_X : x_1 = x\}} p_X(x_1, x_2) = \begin{cases} 2/3 & \text{if } x = 1 \\ 1/3 & \text{if } x = 2 \\ 0 & \text{otherwise} \end{cases}$$

The expected value of X_1 is

$$\mathrm{E}[X_1] = \sum_{x \in R_{X_1}} x p_{X_1}(x) = \frac{2}{3} \cdot 1 + \frac{1}{3} \cdot 2 = \frac{4}{3}$$

The support of X_2 is

$$R_{X_2} = \{1, 3\}$$

and its marginal probability mass function is

$$p_{X_2}(x) = \sum_{\{(x_1,x_2) \in R_X : x_2 = x\}} p_X(x_1, x_2) = \begin{cases} 1/3 & \text{if } x = 1 \\ 2/3 & \text{if } x = 3 \\ 0 & \text{otherwise} \end{cases}$$

The expected value of X_2 is

$$\mathrm{E}[X_2] = \sum_{x \in R_{X_2}} x p_{X_2}(x) = \frac{1}{3} \cdot 1 + \frac{2}{3} \cdot 3 = \frac{7}{3}$$

[7] See p. 120.

By using the transformation theorem[8], we can compute the expected value of X_1X_2:

$$\begin{aligned} \mathrm{E}\left[X_1X_2\right] &= \sum_{x \in R_X} x_1 x_2 p_X\left(x_1, x_2\right) = \\ &= (1 \cdot 3) \cdot p_X\left(1,3\right) + (2 \cdot 1) \cdot p_X\left(2,1\right) \\ &= 3 \cdot \frac{2}{3} + 2 \cdot \frac{1}{3} = \frac{8}{3} \end{aligned}$$

Hence, the covariance between X_1 and X_2 is

$$\begin{aligned} \mathrm{Cov}\left[X_1, X_2\right] &= \mathrm{E}\left[X_1X_2\right] - \mathrm{E}\left[X_1\right]\mathrm{E}\left[X_2\right] \\ &= \frac{8}{3} - \frac{4}{3} \cdot \frac{7}{3} = \frac{24-28}{9} = -\frac{4}{9} \end{aligned}$$

Exercise 2

Let X be a 2×1 discrete random vector and denote its entries by X_1 and X_2. Let the support of X be

$$R_X = \left\{ \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 4 \\ 4 \end{bmatrix} \right\}$$

and its joint probability mass function be

$$p_X\left(x\right) = \begin{cases} 1/3 & \text{if } x = \begin{bmatrix} 1 & 2 \end{bmatrix}^\top \\ 1/3 & \text{if } x = \begin{bmatrix} 2 & 1 \end{bmatrix}^\top \\ 1/3 & \text{if } x = \begin{bmatrix} 4 & 4 \end{bmatrix}^\top \\ 0 & \text{otherwise} \end{cases}$$

Compute the covariance between X_1 and X_2.

Solution

The support of X_1 is

$$R_{X_1} = \{1, 2, 4\}$$

and its marginal probability mass function is

$$p_{X_1}\left(x\right) = \sum_{\{(x_1,x_2) \in R_X : x_1 = x\}} p_X\left(x_1, x_2\right) = \begin{cases} 1/3 & \text{if } x = 1 \\ 1/3 & \text{if } x = 2 \\ 1/3 & \text{if } x = 4 \\ 0 & \text{otherwise} \end{cases}$$

The mean of X_1 is

$$\mathrm{E}\left[X_1\right] = \sum_{x \in R_{X_1}} x p_{X_1}\left(x\right) = \frac{1}{3} \cdot 1 + \frac{1}{3} \cdot 2 + \frac{1}{3} \cdot 4 = \frac{7}{3}$$

The support of X_2 is

$$R_{X_2} = \{1, 2, 4\}$$

[8] See p. 134.

and its probability mass function is

$$p_{X_2}(x) = \sum_{\{(x_1,x_2)\in R_X : x_2 = x\}} p_X(x_1, x_2) = \begin{cases} 1/3 & \text{if } x = 1 \\ 1/3 & \text{if } x = 2 \\ 1/3 & \text{if } x = 4 \\ 0 & \text{otherwise} \end{cases}$$

The mean of X_2 is

$$\mathrm{E}[X_2] = \sum_{x\in R_{X_2}} x p_{X_2}(x) = \frac{1}{3}\cdot 1 + \frac{1}{3}\cdot 2 + \frac{1}{3}\cdot 4 = \frac{7}{3}$$

The expected value of the product X_1X_2 can be derived thanks to the transformation theorem:

$$\begin{aligned} \mathrm{E}[X_1X_2] &= \sum_{x\in R_X} x_1x_2p_X(x_1,x_2) = \\ &= (1\cdot 2)\cdot p_X(1,2) + (2\cdot 1)\cdot p_X(2,1) + (4\cdot 4)\cdot p_X(4,4) \\ &= 2\cdot\frac{1}{3} + 2\cdot\frac{1}{3} + 16\cdot\frac{1}{3} = \frac{20}{3} \end{aligned}$$

By putting pieces together, we obtain the covariance between X_1 and X_2:

$$\begin{aligned} \mathrm{Cov}[X_1,X_2] &= \mathrm{E}[X_1X_2] - \mathrm{E}[X_1]\,\mathrm{E}[X_2] \\ &= \frac{20}{3} - \frac{7}{3}\cdot\frac{7}{3} = \frac{60-49}{9} = \frac{11}{9} \end{aligned}$$

Exercise 3

Let X and Y be two random variables such that

$$\begin{aligned} \mathrm{Var}[X] &= 2 \\ \mathrm{Cov}[X,Y] &= 1 \end{aligned}$$

Compute the covariance

$$\mathrm{Cov}[5X, 2X+3Y]$$

Solution

By exploiting the bilinearity of the covariance operator, we obtain

$$\begin{aligned} \mathrm{Cov}[5X, 2X+3Y] &= 5\mathrm{Cov}[X, 2X+3Y] = 10\mathrm{Cov}[X,X] + 15\mathrm{Cov}[X,Y] \\ &= 10\mathrm{Var}[X] + 15\mathrm{Cov}[X,Y] = 10\cdot 2 + 15\cdot 1 = 35 \end{aligned}$$

Exercise 4

Let $[X \;\; Y]$ be an absolutely continuous random vector[9] with support

$$R_{XY} = \{(x,y) : 0 \leq x \leq y \leq 2\}$$

In other words, the support R_{XY} is the set of all couples (x,y) such that $0 \leq y \leq 2$ and $0 \leq x \leq y$. Let the joint probability density function of $[X \;\; Y]$ be

$$f_{XY}(x,y) = \begin{cases} \frac{3}{8}y & \text{if } (x,y)\in R_{XY} \\ 0 & \text{otherwise} \end{cases}$$

Compute the covariance between X and Y.

[9] See p. 117.

Solution

The support of X is:

$$R_X = [0, 2]$$

thus, when $x \notin [0, 2]$, the marginal probability density function[10] of X is 0, while, when $x \in [0, 2]$, the marginal probability density function of X is

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dy = \int_x^2 \frac{3}{8} y dy = \left[\frac{3}{16} y^2\right]_x^2 = \frac{3}{4} - \frac{3}{16} x^2$$

Therefore, the marginal probability density function of X is

$$f_X(x) = \begin{cases} \frac{3}{4} - \frac{3}{16} x^2 & \text{if } x \in [0, 2] \\ 0 & \text{otherwise} \end{cases}$$

The expected value of X is

$$\begin{aligned} \mathrm{E}[X] &= \int_{-\infty}^{\infty} x f_X(x)\, dx = \int_0^2 x \left(\frac{3}{4} - \frac{3}{16} x^2\right) dx \\ &= \frac{3}{4} \int_0^2 x dx - \frac{3}{16} \int_0^2 x^3 dx \\ &= \frac{3}{4} \left[\frac{1}{2} x^2\right]_0^2 - \frac{3}{16} \left[\frac{1}{4} x^4\right]_0^2 dx \\ &= \frac{3}{4} \cdot (2 - 0) - \frac{3}{16} \cdot (4 - 0) \\ &= \frac{6}{4} - \frac{12}{16} = \frac{6}{4} - \frac{3}{4} = \frac{3}{4} \end{aligned}$$

The support of Y is

$$R_Y = [0, 2]$$

When $y \notin [0, 2]$, the marginal probability density function of Y is 0, while, when $y \in [0, 2]$, the marginal probability density function of Y is

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{XY}(x, y)\, dx = \int_0^y \frac{3}{8} y dx \\ &= \frac{3}{8} y \int_0^y dx = \frac{3}{8} y^2 \end{aligned}$$

Therefore, the marginal probability density function of Y is

$$f_Y(y) = \begin{cases} \frac{3}{8} y^2 & \text{if } y \in [0, 2] \\ 0 & \text{otherwise} \end{cases}$$

The expected value of Y is

$$\begin{aligned} \mathrm{E}[Y] &= \int_{-\infty}^{\infty} y f_Y(y)\, dy = \int_0^2 y \frac{3}{8} y^2 dy \\ &= \frac{3}{8} \int_0^2 y^3 dy = \frac{3}{8} \left[\frac{1}{4} y^4\right]_0^2 = \frac{3}{8} \cdot 4 = \frac{3}{2} \end{aligned}$$

[10] See p. 120.

The expected value of the product XY can be computed by using the transformation theorem:

$$\begin{aligned}
\mathrm{E}[XY] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_{XY}(xy)\,dydx = \int_0^2 \left(\int_0^y xy\frac{3}{8}ydx\right)dy \\
&= \int_0^2 \frac{3}{8}y^2\left(\int_0^y xdx\right)dy = \int_0^2 \frac{3}{8}y^2\left(\left[\frac{1}{2}x^2\right]_0^y\right)dy \\
&= \int_0^2 \frac{3}{8}y^2\frac{1}{2}y^2dy = \int_0^2 \frac{3}{16}y^4dy \\
&= \frac{3}{16}\left[\frac{1}{5}y^5\right]_0^2 = \frac{3}{16}\frac{32}{5} = \frac{6}{5}
\end{aligned}$$

Hence, by using the covariance formula, the covariance between X and Y can be computed as

$$\begin{aligned}
\mathrm{Cov}[X,Y] &= \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y] = \frac{6}{5} - \frac{3}{4}\cdot\frac{3}{2} \\
&= \frac{6}{5} - \frac{9}{8} = \frac{48-45}{40} = \frac{3}{40}
\end{aligned}$$

Exercise 5

Let $[X \;\; Y]$ be an absolutely continuous random vector with support

$$R_{XY} = [0,\infty) \times [1,4]$$

and its joint probability density function be

$$f_{XY}(x,y) = \begin{cases} \frac{1}{3}y\exp(-xy) & \text{if } x \in [0,\infty) \text{ and } y \in [1,4] \\ 0 & \text{otherwise} \end{cases}$$

Compute the covariance between X and Y.

Solution

The support of Y is

$$R_Y = [1,4]$$

When $y \notin [1,4]$, the marginal probability density function of Y is 0, while, when $y \in [1,4]$, the marginal probability density function of Y is

$$\begin{aligned}
f_Y(y) &= \int_{-\infty}^{\infty} f_{XY}(x,y)\,dx = \int_0^{\infty} \frac{1}{3}y\exp(-xy)\,dx \\
&= \frac{1}{3}\left[-\exp(-xy)\right]_0^{\infty} = \frac{1}{3}\left[0-(-1)\right] = \frac{1}{3}
\end{aligned}$$

Thus, the marginal probability density function of Y is

$$f_Y(y) = \begin{cases} 1/3 & \text{if } y \in [1,4] \\ 0 & \text{otherwise} \end{cases}$$

The expected value of Y is

$$\mathrm{E}[Y] = \int_{-\infty}^{\infty} yf_Y(y)\,dy = \int_1^4 y\frac{1}{3}dy$$

$$= \left[\frac{1}{6}y^2\right]_1^4 = \frac{1}{6}\cdot 16 - \frac{1}{6} = \frac{15}{6} = \frac{5}{2}$$

The support of X is

$$R_X = [0, \infty)$$

When $x \notin [0, \infty)$, the marginal probability density function of X is 0, while, when $x \in [0, \infty)$, the marginal probability density function of X is

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dy = \int_1^4 \frac{1}{3} y \exp(-xy)\, dy$$

We do not explicitly compute the integral, but we write the marginal probability density function of X as follows:

$$f_X(x) = \begin{cases} \int_1^4 \frac{1}{3} y \exp(-xy)\, dy & \text{if } x \in [0, \infty) \\ 0 & \text{otherwise} \end{cases}$$

The expected value of X is

$$\begin{aligned}
\mathrm{E}[X] &= \int_{-\infty}^{\infty} x f_X(x)\, dx \\
&= \int_0^{\infty} x \left(\int_1^4 \frac{1}{3} y \exp(-xy)\, dy \right) dx \\
\boxed{\text{A}} &= \frac{1}{3} \int_1^4 \left(\int_0^{\infty} xy \exp(-xy)\, dx \right) dy \\
\boxed{\text{B}} &= \frac{1}{3} \int_1^4 \left(\frac{1}{y} \int_0^{\infty} t \exp(-t)\, dt \right) dy \\
\boxed{\text{C}} &= \frac{1}{3} \int_1^4 \frac{1}{y} \left([-t \exp(-t)]_0^{\infty} + \int_0^{\infty} \exp(-t)\, dt \right) dy \\
&= \frac{1}{3} \int_1^4 \frac{1}{y} \left(0 + [-\exp(-t)]_0^{\infty}\right) dy \\
&= \frac{1}{3} \int_1^4 \frac{1}{y} dy = \frac{1}{3} [\ln(y)]_1^4 = \frac{1}{3} \ln(4)
\end{aligned}$$

where: in step A we have exchanged the order of integration; in step B we have changed variable in the inner integral ($t = xy$); in step C we have integrated by parts.

The expected value of the product XY can be computed by using the transformation theorem:

$$\begin{aligned}
\mathrm{E}[XY] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_{XY}(xy)\, dy dx \\
&= \int_0^{\infty} \left(\int_1^4 xy \frac{1}{3} y \exp(-xy)\, dy \right) dx \\
\boxed{\text{A}} &= \frac{1}{3} \int_1^4 y \left(\int_0^{\infty} xy \exp(-xy)\, dx \right) dy
\end{aligned}$$

$$\begin{aligned}
\boxed{\text{B}} &= \frac{1}{3}\int_1^4 y\left(\frac{1}{y}\int_0^\infty t\exp\left(-t\right)dt\right)dy \\
\boxed{\text{C}} &= \frac{1}{3}\int_1^4 \left(\left[-t\exp\left(-t\right)\right]_0^\infty + \int_0^\infty \exp\left(-t\right)dt\right)dy \\
&= \frac{1}{3}\int_1^4 \left(0 + \left[-\exp\left(-t\right)\right]_0^\infty\right)dy \\
&= \frac{1}{3}\int_1^4 dy = \frac{1}{3}\cdot 3 = 1
\end{aligned}$$

where: in step $\boxed{\text{A}}$ we have exchanged the order of integration; in step $\boxed{\text{B}}$ we have changed variable in the inner integral ($t = xy$); in step $\boxed{\text{C}}$ we have integrated by parts.

Hence, using the covariance formula, the covariance between X and Y is

$$\begin{aligned}
\text{Cov}\left[X,Y\right] &= \text{E}\left[XY\right] - \text{E}\left[X\right]\text{E}\left[Y\right] \\
&= 1 - \frac{1}{3}\ln\left(4\right)\cdot\frac{5}{2} = 1 - \frac{5}{6}\ln\left(4\right)
\end{aligned}$$

Exercise 6

Let X and Y be two random variables such that

$$\begin{aligned}
&\text{Var}\left[X\right] = 4 \\
&\text{Cov}\left[X,Y\right] = 2
\end{aligned}$$

Compute the following covariance:

$$\text{Cov}\left[3X, X+3Y\right]$$

Solution

The bilinearity of the covariance operator implies that

$$\begin{aligned}
\text{Cov}\left[3X, X+3Y\right] &= 3\text{Cov}\left[X, X+3Y\right] = 3\text{Cov}\left[X,X\right] + 9\text{Cov}\left[X,Y\right] \\
&= 3\text{Var}\left[X\right] + 9\text{Cov}\left[X,Y\right] = 3\cdot 4 + 9\cdot 2 = 30
\end{aligned}$$

Chapter 22

Linear correlation

This lecture introduces the linear correlation coefficient. Before reading this lecture, make sure you are familiar with the concept of covariance (p. 163).

22.1 Definition of linear correlation coefficient

Let X and Y be two random variables. The **linear correlation coefficient** (or Pearson's correlation coefficient) between X and Y, denoted by $\operatorname{Corr}[X,Y]$ or by ρ_{XY}, is defined as follows:

$$\operatorname{Corr}[X,Y] = \frac{\operatorname{Cov}[X,Y]}{\operatorname{stdev}[X]\operatorname{stdev}[Y]}$$

where $\operatorname{Cov}[X,Y]$ is the covariance between X and Y and $\operatorname{stdev}[X]$ and $\operatorname{stdev}[Y]$ are the standard deviations[1] of X and Y.

Of course, the linear correlation coefficient is well-defined only as long as the three quantities $\operatorname{Cov}[X,Y]$, $\operatorname{stdev}[X]$ and $\operatorname{stdev}[Y]$ exist and are well-defined.

Moreover, while the ratio is well-defined only if $\operatorname{stdev}[X]$ and $\operatorname{stdev}[Y]$ are strictly greater than zero, it is often assumed that $\operatorname{Corr}[X,Y] = 0$ when one of the two standard deviations is zero. This is equivalent to assuming that $0/0 = 0$, because $\operatorname{Cov}[X,Y] = 0$ when one of the two standard deviations is zero.

22.2 Interpretation

Linear correlation is a measure of dependence, or association, between two random variables. Its interpretation is similar to the interpretation of covariance[2].

The correlation between X and Y provides a measure of the degree to which X and Y tend to "move together": $\operatorname{Corr}[X,Y] > 0$ indicates that deviations of X and Y from their respective means tend to have the same sign; $\operatorname{Corr}[X,Y] < 0$ indicates that deviations of X and Y from their respective means tend to have opposite signs; when $\operatorname{Corr}[X,Y] = 0$, X and Y do not display any of these two tendencies.

[1] See p. 157.

[2] See the lecture entitled *Covariance* (p. 163) for a detailed explanation.

Linear correlation has the property of being bounded between -1 and 1:

$$-1 \leq \text{Corr}\left[X,Y\right] \leq 1$$

Thanks to this property, correlation allows to easily understand the intensity of the linear dependence between two random variables: the closer the correlation is to 1, the stronger the positive linear dependence between X and Y is, and the closer it is to -1, the stronger the negative linear dependence between X and Y is.

22.3 Terminology

The following terminology is often used:

1. If $\text{Corr}\left[X,Y\right] > 0$ then X and Y are said to be **positively linearly correlated** (or simply **positively correlated**).
2. If $\text{Corr}\left[X,Y\right] < 0$ then X and Y are said to be **negatively linearly correlated** (or simply **negatively correlated**).
3. If $\text{Corr}\left[X,Y\right] \neq 0$ then X and Y are said to be **linearly correlated** (or simply **correlated**).
4. If $\text{Corr}\left[X,Y\right] = 0$ then X and Y are said to be **uncorrelated**. Also note that $\text{Cov}\left[X,Y\right] = 0 \Rightarrow \text{Corr}\left[X,Y\right] = 0$, therefore two random variables X and Y are uncorrelated whenever $\text{Cov}\left[X,Y\right] = 0$.

22.4 Example

The following example shows how to compute the coefficient of linear correlation between two discrete random variables.

Example 125 *Let X be a 2×1 random vector and denote its components by X_1 and X_2. Let the support of X be*

$$R_X = \left\{ \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right\}$$

and its joint probability mass function[3] *be*

$$p_X\left(x\right) = \begin{cases} 1/3 & \text{if } x = \begin{bmatrix} -1 & 1 \end{bmatrix}^\top \\ 1/3 & \text{if } x = \begin{bmatrix} -1 & -1 \end{bmatrix}^\top \\ 1/3 & \text{if } x = \begin{bmatrix} 1 & 1 \end{bmatrix}^\top \\ 0 & \text{otherwise} \end{cases}$$

The support of X_1 is

$$R_{X_1} = \{-1, 1\}$$

and its probability mass function is

$$p_{X_1}\left(x\right) = \sum_{\{(x_1,x_2) \in R_X : x_1 = x\}} p_X\left(x_1, x_2\right) = \begin{cases} 2/3 & \text{if } x = -1 \\ 1/3 & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases}$$

[3] See p. 116.

The expected value of X_1 *is*

$$\mathrm{E}\left[X_1\right] = \sum_{x \in R_{X_1}} x p_{X_1}\left(x\right) = \frac{2}{3} \cdot \left(-1\right) + \frac{1}{3} \cdot 1 = -\frac{1}{3}$$

The expected value of X_1^2 *is*

$$\mathrm{E}\left[X_1^2\right] = \sum_{x \in R_{X_1}} x^2 p_{X_1}\left(x\right) = \frac{2}{3} \cdot \left(-1\right)^2 + \frac{1}{3} \cdot 1^2 = 1$$

The variance of X_1 *is*

$$\mathrm{Var}\left[X_1\right] = \mathrm{E}\left[X_1^2\right] - \mathrm{E}\left[X_1\right]^2 = 1 - \left(-\frac{1}{3}\right)^2 = \frac{8}{9}$$

The standard deviation of X_1 *is*

$$\mathrm{stdev}\left[X_1\right] = \sqrt{\mathrm{Var}\left[X_1\right]} = \sqrt{\frac{8}{9}}$$

The support of X_2 *is*

$$R_{X_2} = \{-1, 1\}$$

and its probability mass function is

$$p_{X_2}\left(x\right) = \sum_{\{(x_1,x_2) \in R_X : x_2 = x\}} p_X\left(x_1, x_2\right) = \begin{cases} 1/3 & \text{if } x = -1 \\ 2/3 & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases}$$

The expected value of X_2 *is*

$$\mathrm{E}\left[X_2\right] = \sum_{x \in R_{X_2}} x p_{X_2}\left(x\right) = \frac{1}{3} \cdot \left(-1\right) + \frac{2}{3} \cdot 1 = \frac{1}{3}$$

The expected value of X_2^2 *is*

$$\mathrm{E}\left[X_2^2\right] = \sum_{x \in R_{X_2}} x^2 p_{X_2}\left(x\right) = \frac{1}{3} \cdot \left(-1\right)^2 + \frac{2}{3} \cdot 1^2 = 1$$

The variance of X_2 *is*

$$\mathrm{Var}\left[X_2\right] = \mathrm{E}\left[X_2^2\right] - \mathrm{E}\left[X_2\right]^2 = 1 - \left(\frac{1}{3}\right)^2 = \frac{8}{9}$$

The standard deviation of X_2 *is*

$$\mathrm{stdev}\left[X_2\right] = \sqrt{\mathrm{Var}\left[X_2\right]} = \sqrt{\frac{8}{9}}$$

By using the transformation theorem[4]*, we can compute the expected value of the product* $X_1 X_2$*:*

$$\mathrm{E}\left[X_1 X_2\right] = \sum_{x \in R_X} x_1 x_2 p_X\left(x_1, x_2\right)$$

[4] See p. 134.

$$= \frac{1}{3}\cdot(-1\cdot 1)+\frac{1}{3}\cdot((-1)\cdot(-1))+\frac{1}{3}\cdot(1\cdot 1)=\frac{1}{3}$$

Hence, the covariance between X_1 and X_2 is

$$\operatorname{Cov}[X_1,X_2]=\operatorname{E}[X_1X_2]-\operatorname{E}[X_1]\operatorname{E}[X_2]=\frac{1}{3}-\left(-\frac{1}{3}\right)\cdot\frac{1}{3}=\frac{4}{9}$$

and the linear correlation coefficient is

$$\operatorname{Corr}[X_1,X_2]=\frac{\operatorname{Cov}[X_1,X_2]}{\operatorname{stdev}[X_1]\operatorname{stdev}[X_2]}=\frac{4/9}{\sqrt{8/9}\sqrt{8/9}}=\frac{4/9}{8/9}=\frac{1}{2}$$

22.5 More details

22.5.1 Correlation of a random variable with itself

The correlation of a random variable with itself is equal to 1.

Proposition 126 *If the correlation coefficient of a random variable with itself exists and is well-defined, then*

$$\operatorname{Corr}[X,X]=1$$

Proof. This is proved as follows:

$$\begin{aligned}\operatorname{Corr}[X,X] &= \frac{\operatorname{Cov}[X,X]}{\operatorname{stdev}[X]\operatorname{stdev}[X]}\\ &= \frac{\operatorname{Var}[X]}{\sqrt{\operatorname{Var}[X]}\sqrt{\operatorname{Var}[X]}}\\ &= \frac{\operatorname{Var}[X]}{\operatorname{Var}[X]}=1\end{aligned}$$

where we have used the fact that[5]

$$\operatorname{Cov}[X,X]=\operatorname{Var}[X]$$

■

22.5.2 Symmetry

The linear correlation coefficient is symmetric.

Proposition 127 *If the correlation coefficient between two random variables exists and is well-defined, then it satisfies*

$$\operatorname{Corr}[X,Y]=\operatorname{Corr}[Y,X]$$

Proof. This is proved as follows:

$$\operatorname{Corr}[X,Y] = \frac{\operatorname{Cov}[X,Y]}{\operatorname{stdev}[X]\operatorname{stdev}[Y]}$$

[5] See p. 166.

$$\begin{aligned} &= \frac{\operatorname{Cov}[Y,X]}{\operatorname{stdev}[X]\operatorname{stdev}[Y]} \\ &= \frac{\operatorname{Cov}[Y,X]}{\operatorname{stdev}[Y]\operatorname{stdev}[X]} \\ &= \operatorname{Corr}[Y,X] \end{aligned}$$

where we have used the fact that covariance is symmetric[6]:

$$\operatorname{Cov}[X,Y] = \operatorname{Cov}[Y,X]$$

■

22.6 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Let X be a 2×1 discrete random vector and denote its components by X_1 and X_2. Let the support of X be

$$R_X = \left\{ \begin{bmatrix} 1 \\ 5 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \end{bmatrix} \right\}$$

and its joint probability mass function be

$$p_X(x) = \begin{cases} 4/5 & \text{if } x = \begin{bmatrix} 1 & 5 \end{bmatrix}^\top \\ 1/5 & \text{if } x = \begin{bmatrix} 2 & 1 \end{bmatrix}^\top \\ 0 & \text{otherwise} \end{cases}$$

Compute the coefficient of linear correlation between X_1 and X_2.

Solution

The support of X_1 is

$$R_{X_1} = \{1, 2\}$$

and its marginal probability mass function[7] is

$$p_{X_1}(x) = \sum_{\{(x_1,x_2) \in R_X : x_1 = x\}} p_X(x_1, x_2) = \begin{cases} 4/5 & \text{if } x = 1 \\ 1/5 & \text{if } x = 2 \\ 0 & \text{otherwise} \end{cases}$$

The expected value of X_1 is

$$\operatorname{E}[X_1] = \sum_{x \in R_{X_1}} x p_{X_1}(x) = 1 \cdot \frac{4}{5} + 2 \cdot \frac{1}{5} = \frac{6}{5}$$

The expected value of X_1^2 is

$$\operatorname{E}\left[X_1^2\right] = \sum_{x \in R_{X_1}} x^2 p_{X_1}(x) = 1^2 \cdot \frac{4}{5} + 2^2 \cdot \frac{1}{5} = 1 \cdot \frac{4}{5} + 4 \cdot \frac{1}{5} = \frac{8}{5}$$

[6] See p. 166.
[7] See p. 120.

The variance of X_1 is

$$\text{Var}[X_1] = \text{E}\left[X_1^2\right] - \text{E}[X_1]^2 = \frac{8}{5} - \left(\frac{6}{5}\right)^2 = \frac{40-36}{25} = \frac{4}{25}$$

The standard deviation of X_1 is

$$\text{stdev}[X_1] = \sqrt{\frac{4}{25}} = \frac{2}{5}$$

The support of X_2 is

$$R_{X_2} = \{1, 5\}$$

and its marginal probability mass function is

$$p_{X_2}(x) = \sum_{\{(x_1,x_2)\in R_X : x_2 = x\}} p_X(x_1, x_2) = \begin{cases} 1/5 & \text{if } x = 1 \\ 4/5 & \text{if } x = 5 \\ 0 & \text{otherwise} \end{cases}$$

The expected value of X_2 is

$$\text{E}[X_2] = \sum_{x\in R_{X_2}} x p_{X_2}(x) = 1 \cdot \frac{1}{5} + 5 \cdot \frac{4}{5} = \frac{21}{5}$$

The expected value of X_2^2 is

$$\text{E}\left[X_2^2\right] = \sum_{x\in R_{X_2}} x^2 p_{X_2}(x) = 1^2 \cdot \frac{1}{5} + 5^2 \cdot \frac{4}{5} = 1 \cdot \frac{1}{5} + 25 \cdot \frac{4}{5} = \frac{101}{5}$$

The variance of X_2 is

$$\text{Var}[X_2] = \text{E}\left[X_2^2\right] - \text{E}[X_2]^2 = \frac{101}{5} - \left(\frac{21}{5}\right)^2 = \frac{505-441}{25} = \frac{64}{25}$$

The standard deviation of X_1 is

$$\text{stdev}[X_2] = \sqrt{\frac{64}{25}} = \frac{8}{5}$$

By using the transformation theorem, we can compute the expected value of X_1X_2:

$$\begin{aligned} \text{E}[X_1X_2] &= \sum_{x\in R_X} x_1 x_2 p_X(x_1, x_2) = (1 \cdot 5) \cdot p_X(1, 5) + (2 \cdot 1) \cdot p_X(2, 1) \\ &= 5 \cdot \frac{4}{5} + 2 \cdot \frac{1}{5} = \frac{22}{5} \end{aligned}$$

Hence, the covariance between X_1 and X_2 is

$$\begin{aligned} \text{Cov}[X_1, X_2] &= \text{E}[X_1X_2] - \text{E}[X_1]\,\text{E}[X_2] = \frac{22}{5} - \frac{6}{5} \cdot \frac{21}{5} \\ &= \frac{110-126}{25} = -\frac{16}{25} \end{aligned}$$

and the coefficient of linear correlation between X_1 and X_2 is

$$\text{Corr}[X_1, X_2] = \frac{\text{Cov}[X_1, X_2]}{\text{stdev}[X_1]\,\text{stdev}[X_2]} = \frac{-\frac{16}{25}}{\frac{2}{5} \cdot \frac{8}{5}} = -1$$

Exercise 2

Let X be a 2×1 discrete random vector and denote its entries by X_1 and X_2. Let the support of X be

$$R_X = \left\{ \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 3 \\ 3 \end{bmatrix} \right\}$$

and its joint probability mass function be

$$p_X(x) = \begin{cases} 1/3 & \text{if } x = \begin{bmatrix} 1 & 2 \end{bmatrix}^\top \\ 1/3 & \text{if } x = \begin{bmatrix} 2 & 1 \end{bmatrix}^\top \\ 1/3 & \text{if } x = \begin{bmatrix} 3 & 3 \end{bmatrix}^\top \\ 0 & \text{otherwise} \end{cases}$$

Compute the covariance between X_1 and X_2.

Solution

The support of X_1 is

$$R_{X_1} = \{1, 2, 3\}$$

and its marginal probability mass function is

$$p_{X_1}(x) = \sum_{\{(x_1, x_2) \in R_X : x_1 = x\}} p_X(x_1, x_2) = \begin{cases} 1/3 & \text{if } x = 1 \\ 1/3 & \text{if } x = 2 \\ 1/3 & \text{if } x = 3 \\ 0 & \text{otherwise} \end{cases}$$

The mean of X_1 is

$$\mathrm{E}[X_1] = \sum_{x \in R_{X_1}} x p_{X_1}(x) = 1 \cdot \frac{1}{3} + 2 \cdot \frac{1}{3} + 3 \cdot \frac{1}{3} = \frac{6}{3} = 2$$

The expected value of X_1^2 is

$$\begin{aligned} \mathrm{E}\left[X_1^2\right] &= \sum_{x \in R_{X_1}} x^2 p_{X_1}(x) = 1^2 \cdot \frac{1}{3} + 2^2 \cdot \frac{1}{3} + 3^2 \cdot \frac{1}{3} \\ &= 1 \cdot \frac{1}{3} + 4 \cdot \frac{1}{3} + 9 \cdot \frac{1}{3} = \frac{14}{3} \end{aligned}$$

The variance of X_1 is

$$\mathrm{Var}[X_1] = \mathrm{E}\left[X_1^2\right] - \mathrm{E}[X_1]^2 = \frac{14}{3} - 2^2 = \frac{14 - 12}{3} = \frac{2}{3}$$

The standard deviation of X_1 is

$$\mathrm{stdev}[X_1] = \sqrt{\frac{2}{3}}$$

The support of X_2 is

$$R_{X_2} = \{1, 2, 3\}$$

and its probability mass function is

$$p_{X_2}(x) = \sum_{\{(x_1,x_2)\in R_X : x_2 = x\}} p_X(x_1,x_2) = \begin{cases} 1/3 & \text{if } x = 1 \\ 1/3 & \text{if } x = 2 \\ 1/3 & \text{if } x = 3 \\ 0 & \text{otherwise} \end{cases}$$

The mean of X_2 is

$$\mathrm{E}[X_2] = \sum_{x\in R_{X_2}} x p_{X_2}(x) = \frac{1}{3}\cdot 1 + \frac{1}{3}\cdot 2 + \frac{1}{3}\cdot 3 = \frac{6}{3} = 2$$

The expected value of X_2^2 is

$$\begin{aligned} \mathrm{E}\left[X_2^2\right] &= \sum_{x\in R_{X_2}} x^2 p_{X_2}(x) = 1^2\cdot\frac{1}{3} + 2^2\cdot\frac{1}{3} + 3^2\cdot\frac{1}{3} \\ &= 1\cdot\frac{1}{3} + 4\cdot\frac{1}{3} + 9\cdot\frac{1}{3} = \frac{14}{3} \end{aligned}$$

The variance of X_2 is

$$\mathrm{Var}[X_2] = \mathrm{E}\left[X_2^2\right] - \mathrm{E}[X_2]^2 = \frac{14}{3} - 2^2 = \frac{14-12}{3} = \frac{2}{3}$$

The standard deviation of X_2 is

$$\mathrm{stdev}[X_2] = \sqrt{\frac{2}{3}}$$

The expected value of the product X_1X_2 can be derived thanks to the transformation theorem:

$$\begin{aligned} \mathrm{E}[X_1X_2] &= \sum_{x\in R_X} x_1x_2 p_X(x_1,x_2) \\ &= (1\cdot 2)\cdot p_X(1,2) + (2\cdot 1)\cdot p_X(2,1) + (3\cdot 3)\cdot p_X(3,3) \\ &= 2\cdot\frac{1}{3} + 2\cdot\frac{1}{3} + 9\cdot\frac{1}{3} = \frac{13}{3} \end{aligned}$$

Therefore, putting pieces together, the covariance between X_1 and X_2 is

$$\mathrm{Cov}[X_1,X_2] = \mathrm{E}[X_1X_2] - \mathrm{E}[X_1]\,\mathrm{E}[X_2] = \frac{13}{3} - 2\cdot 2 = \frac{13-12}{3} = \frac{1}{3}$$

and the coefficient of linear correlation between X_1 and X_2 is

$$\mathrm{Corr}[X_1,X_2] = \frac{\mathrm{Cov}[X_1,X_2]}{\mathrm{stdev}[X_1]\,\mathrm{stdev}[X_2]} = \frac{\frac{1}{3}}{\sqrt{\frac{2}{3}}\cdot\sqrt{\frac{2}{3}}} = \frac{1}{2}$$

Exercise 3

Let $[X \;\; Y]$ be an absolutely continuous random vector with support

$$R_{XY} = [0,\infty)\times[1,2]$$

and let its joint probability density function[8] be

$$f_{XY}(x,y) = \begin{cases} 2y\exp(-2xy) & \text{if } x\in[0,\infty) \text{ and } y\in[1,2] \\ 0 & \text{otherwise} \end{cases}$$

Compute the covariance between X and Y.

[8] See p. 117.

Solution

The support of Y is

$$R_Y = [1, 2]$$

When $y \notin R_Y$, the marginal probability density function[9] of Y is 0, while, when $y \in R_Y$, the marginal probability density function of Y can be obtained by integrating x out of the joint probability density as follows:

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{XY}(x, y)\, dx = \int_0^{\infty} 2y \exp(-2xy)\, dx \\ &= [-\exp(-2xy)]_0^{\infty} = [0 - (-1)] = 1 \end{aligned}$$

Thus, the marginal probability density function of Y is

$$f_Y(y) = \begin{cases} 1 & \text{if } y \in [1, 2] \\ 0 & \text{otherwise} \end{cases}$$

The expected value of Y is

$$\mathrm{E}[Y] = \int_{-\infty}^{\infty} y f_Y(y)\, dy = \int_1^2 y dy = \left[\frac{1}{2}y^2\right]_1^2 = \frac{1}{2} \cdot 4 - \frac{1}{2} \cdot 1 = \frac{3}{2}$$

The expected value of Y^2 is:

$$\mathrm{E}\left[Y^2\right] = \int_{-\infty}^{\infty} y^2 f_Y(y)\, dy = \int_1^2 y^2 dy = \left[\frac{1}{3}y^3\right]_1^2 = \frac{1}{3} \cdot 8 - \frac{1}{3} \cdot 1 = \frac{7}{3}$$

The variance of Y is

$$\mathrm{Var}[Y] = \mathrm{E}\left[Y^2\right] - \mathrm{E}[Y]^2 = \frac{7}{3} - \left(\frac{3}{2}\right)^2 = \frac{28 - 27}{12} = \frac{1}{12}$$

The standard deviation of Y is

$$\mathrm{stdev}[Y] = \sqrt{\frac{1}{12}} = \frac{1}{2}\sqrt{\frac{1}{3}}$$

The support of X is

$$R_X = [0, \infty)$$

When $x \notin R_X$, the marginal probability density function of X is 0, while, when $x \in R_X$, the marginal probability density function of X can be obtained by integrating y out of the joint probability density as follows:

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dy = \int_1^2 2y \exp(-2xy)\, dy$$

We do not explicitly compute the integral, but we write the marginal probability density function of X as follows:

$$f_X(x) = \begin{cases} \int_1^2 2y \exp(-2xy)\, dy & \text{if } x \in [0, \infty) \\ 0 & \text{otherwise} \end{cases}$$

[9] See p. 120.

The expected value of X is

$$\begin{aligned}
\mathrm{E}[X] &= \int_{-\infty}^{\infty} x f_X(x)\,dx \\
&= \int_0^{\infty} x \left(\int_1^2 2y \exp(-2xy)\,dy \right) dx \\
\boxed{\text{A}} &= \int_1^2 \left(\int_0^{\infty} 2xy \exp(-2xy)\,dx \right) dy \\
\boxed{\text{B}} &= \int_1^2 \left(\frac{1}{2y} \int_0^{\infty} t \exp(-t)\,dt \right) dy \\
\boxed{\text{C}} &= \frac{1}{2} \int_1^2 \frac{1}{y} \left([-t \exp(-t)]_0^{\infty} + \int_0^{\infty} \exp(-t)\,dt \right) dy \\
&= \frac{1}{2} \int_1^2 \frac{1}{y} \left(0 + [-\exp(-t)]_0^{\infty} \right) dy \\
&= \frac{1}{2} \int_1^2 \frac{1}{y} dy = \frac{1}{2} [\ln(y)]_1^2 = \frac{1}{2} \ln(2)
\end{aligned}$$

where: in step $\boxed{\text{A}}$ we have exchanged the order of integration; in step $\boxed{\text{B}}$ we have made a change of variable in the inner integral ($t = 2xy$); in step $\boxed{\text{C}}$ we have performed an integration by parts.

The expected value of X^2 is

$$\begin{aligned}
\mathrm{E}\left[X^2\right] &= \int_{-\infty}^{\infty} x^2 f_X(x)\,dx \\
&= \int_0^{\infty} x^2 \left(\int_1^2 2y \exp(-2xy)\,dy \right) dx \\
\boxed{\text{A}} &= \int_1^2 \left(\int_0^{\infty} 2yx^2 \exp(-2yx)\,dx \right) dy \\
\boxed{\text{B}} &= \int_1^2 \left(\left[-x^2 \exp(-2yx)\right]_0^{\infty} + \int_0^{\infty} 2x \exp(-2yx)\,dx \right) dy \\
&= \int_1^2 \left(\int_0^{\infty} 2x \exp(-2yx)\,dx \right) dy \\
&= \int_1^2 \frac{1}{y} \left(\int_0^{\infty} 2yx \exp(-2yx)\,dx \right) dy \\
\boxed{\text{C}} &= \int_1^2 \frac{1}{y} \left([-x \exp(-2yx)]_0^{\infty} + \int_0^{\infty} \exp(-2yx)\,dx \right) dy \\
&= \int_1^2 \frac{1}{y} \left(\int_0^{\infty} \exp(-2yx)\,dx \right) dy \\
&= \int_1^2 \frac{1}{y} \left(\left[-\frac{1}{2y} \exp(-2yx) \right]_0^{\infty} \right) dy \\
&= \int_1^2 \frac{1}{y} \left(\frac{1}{2y} \right) dy = \frac{1}{2} \int_1^2 y^{-2} dy \\
&= \frac{1}{2} \left[-y^{-1} \right]_1^2 = \frac{1}{2} \left[-\frac{1}{2} + 1 \right] = \frac{1}{4}
\end{aligned}$$

where: in step A we have exchanged the order of integration; in step B and C we have performed an integration by parts.

The variance of X is

$$\operatorname{Var}[X] = \operatorname{E}\left[X^2\right] - \operatorname{E}[X]^2 = \frac{1}{4} - \left(\frac{1}{2}\ln(2)\right)^2 = \frac{1}{4}\left[1 - (\ln(2))^2\right]$$

The standard deviation of X is

$$\operatorname{stdev}[X] = \sqrt{\frac{1}{4}\left[1 - (\ln(2))^2\right]} = \frac{1}{2}\sqrt{1 - (\ln(2))^2}$$

The expected value of the product XY can be computed thanks to the transformation theorem:

$$\begin{aligned}
\operatorname{E}[XY] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{XY}(xy)\,dydx \\
&= \int_0^{\infty}\left(\int_1^2 xy2y\exp(-2xy)\,dy\right)dx \\
\boxed{\text{A}} &= \int_1^2 y\left(\int_0^{\infty} 2xy\exp(-2xy)\,dx\right)dy \\
\boxed{\text{B}} &= \int_1^2 y\left([-x\exp(-2xy)]_0^{\infty} + \int_0^{\infty}\exp(-2xy)\,dx\right)dy \\
&= \int_1^2 y\left(0 + \left[-\frac{1}{2y}\exp(-2xy)\right]_0^{\infty}\right)dy \\
&= \int_1^2 y\frac{1}{2y}dy = \frac{1}{2}\int_1^2 dy = \frac{1}{2}
\end{aligned}$$

where: in step A we have exchanged the order of integration; in step B we have performed an integration by parts.

Hence, by using the covariance formula, we obtain the covariance between X and Y as follows:

$$\operatorname{Cov}[X,Y] = \operatorname{E}[XY] - \operatorname{E}[X]\operatorname{E}[Y] = \frac{1}{2} - \frac{1}{2}\ln(2)\cdot\frac{3}{2} = \frac{1}{2} - \frac{3}{4}\ln(2)$$

Thus, the coefficient of linear correlation between X and Y is

$$\begin{aligned}
\operatorname{Corr}[X,Y] &= \frac{\operatorname{Cov}[X,Y]}{\operatorname{stdev}[X]\operatorname{stdev}[Y]} = \frac{\frac{1}{2} - \frac{3}{4}\ln(2)}{\frac{1}{2}\sqrt{1-(\ln(2))^2}\cdot\frac{1}{2}\sqrt{\frac{1}{3}}} \\
&= \frac{2 - 3\ln(2)}{\sqrt{1-(\ln(2))^2}\cdot\sqrt{\frac{1}{3}}}
\end{aligned}$$

Chapter 23

Covariance matrix

This lecture introduces the covariance matrix of a random vector, which is a multivariate generalization of the concept of variance of a random variable. Before reading this lecture, make sure you are familiar with the concepts of variance (p. 155) and covariance (p. 163).

23.1 Definition

Let X be a $K \times 1$ random vector. The **covariance matrix** of X, or **variance-covariance** matrix of X, denoted by $\operatorname{Var}[X]$, is defined as follows:

$$\operatorname{Var}[X] = \operatorname{E}\left[(X - \operatorname{E}[X])(X - \operatorname{E}[X])^{\top}\right]$$

provided the above expected value exists and is well-defined. It is a multivariate generalization of the definition of variance of a random variable Y:

$$\operatorname{Var}[Y] = \operatorname{E}\left[(Y - \operatorname{E}[Y])^2\right] = \operatorname{E}[(Y - \operatorname{E}[Y])(Y - \operatorname{E}[Y])]$$

23.2 Structure of the covariance matrix

Let $X_1, \ldots, X_K$ denote the K components of the vector X. From the definition of $\operatorname{Var}[X]$, it can easily be seen that $\operatorname{Var}[X]$ is a $K \times K$ matrix with the following structure:

$$\operatorname{Var}[X] = \begin{bmatrix} V_{11} & V_{12} & \ldots & V_{1K} \\ V_{21} & V_{22} & \ldots & V_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ V_{K1} & V_{K2} & \ldots & V_{KK} \end{bmatrix}$$

where the (i,j)-th entry of the matrix is equal to the covariance between X_i and X_j:

$$V_{ij} = \operatorname{E}[(X_i - \operatorname{E}[X_i])(X_j - \operatorname{E}[X_j])] = \operatorname{Cov}[X_i, X_j]$$

Since the covariance between X_i and X_j is equal to the variance of X_i when $i = j$ (i.e., $\operatorname{Cov}[X_i, X_i] = \operatorname{Var}[X_i]$), the diagonal entries of the covariance matrix are equal to the variances of the individual components of X.

Example 128 *Suppose X is a 2×1 random vector with components X_1 and X_2. Let*

$$\begin{aligned} \operatorname{Var}[X_1] &= 2 \\ \operatorname{Var}[X_2] &= 4 \\ \operatorname{Cov}[X_1, X_2] &= 1 \end{aligned}$$

By the symmetry of covariance[1]*, it must also be that*

$$\operatorname{Cov}[X_2, X_1] = \operatorname{Cov}[X_1, X_2] = 1$$

Therefore, the covariance matrix of X is

$$\operatorname{Var}[X] = \begin{bmatrix} \operatorname{Var}[X_1] & \operatorname{Cov}[X_1, X_2] \\ \operatorname{Cov}[X_2, X_1] & \operatorname{Var}[X_2] \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}$$

23.3 Covariance matrix formula

The following result is often used to compute the covariance matrix.

Proposition 129 *The covariance matrix of a $K \times 1$ random vector X, if it exists, can be computed as follows:*

$$\operatorname{Var}[X] = \operatorname{E}\left[XX^\top\right] - \operatorname{E}[X]\operatorname{E}[X]^\top$$

Proof. The above formula can be derived as follows:

$$\begin{aligned} \operatorname{Var}[X] &= \operatorname{E}\left[(X - \operatorname{E}[X])(X - \operatorname{E}[X])^\top\right] \\ &= \operatorname{E}\left[XX^\top - X\operatorname{E}[X]^\top - \operatorname{E}[X]X^\top + \operatorname{E}[X]\operatorname{E}[X]^\top\right] \\ &= \operatorname{E}\left[XX^\top - X\operatorname{E}[X]^\top - \left(X\operatorname{E}[X]^\top\right)^\top + \operatorname{E}[X]\operatorname{E}[X]^\top\right] \\ \boxed{\text{A}} \quad &= \operatorname{E}\left[XX^\top - X\operatorname{E}[X]^\top - X\operatorname{E}[X]^\top + \operatorname{E}[X]\operatorname{E}[X]^\top\right] \\ &= \operatorname{E}\left[XX^\top - 2X\operatorname{E}[X]^\top + \operatorname{E}[X]\operatorname{E}[X]^\top\right] \\ \boxed{\text{B}} \quad &= \operatorname{E}\left[XX^\top\right] - 2\operatorname{E}[X]\operatorname{E}[X]^\top + \operatorname{E}[X]\operatorname{E}[X]^\top \\ &= \operatorname{E}\left[XX^\top\right] - \operatorname{E}[X]\operatorname{E}[X]^\top \end{aligned}$$

where: in step A we have used the fact that a scalar is equal to its transpose; in step B we have used the linearity of the expected value[2]. ■

This formula also makes clear that the covariance matrix exists and is well-defined only as long as the vector of expected values $\operatorname{E}[X]$ and the matrix of second cross-moments[3] $\operatorname{E}\left[XX^\top\right]$ exist and are well-defined.

23.4 More details

The following subsections contain more details about the covariance matrix.

[1] See p. 166.
[2] See p. 134.
[3] See p. 285.

23.4.1 Addition to a constant vector

Adding a constant to a random vector does not change its covariance matrix.

Proposition 130 *Let $a \in \mathbb{R}^K$ be a constant $K \times 1$ vector and let X be a $K \times 1$ random vector having covariance matrix* $\operatorname{Var}[X]$*. Then,*

$$\operatorname{Var}[a + X] = \operatorname{Var}[X]$$

Proof. This is a consequence of the fact that[4] $\operatorname{E}[a + X] = a + \operatorname{E}[X]$:

$$\begin{aligned}
\operatorname{Var}[a + X] &= \operatorname{E}\left[(a + X - \operatorname{E}[a + X])(a + X - \operatorname{E}[a + X])^\top\right] \\
&= \operatorname{E}\left[(a + X - a - \operatorname{E}[X])(a + X - a - \operatorname{E}[X])^\top\right] \\
&= \operatorname{E}\left[(X - \operatorname{E}[X])(X - \operatorname{E}[X])^\top\right] \\
&= \operatorname{Var}[X]
\end{aligned}$$

■

23.4.2 Multiplication by a constant matrix

If a random vector is pre-multiplied by a constant matrix b, then its covariance matrix is pre-multiplied by b and post-multiplied by the transpose of b.

Proposition 131 *Let b be a constant $M \times K$ matrix and let X be a $K \times 1$ random vector having covariance matrix* $\operatorname{Var}[X]$*. Then,*

$$\operatorname{Var}[bX] = b\operatorname{Var}[X]\, b^\top$$

Proof. This is easily proved using the fact that[5] $\operatorname{E}[bX] = b\operatorname{E}[X]$:

$$\begin{aligned}
\operatorname{Var}[bX] &= \operatorname{E}\left[(bX - \operatorname{E}[bX])(bX - \operatorname{E}[bX])^\top\right] \\
&= \operatorname{E}\left[(bX - b\operatorname{E}[X])(bX - b\operatorname{E}[X])^\top\right] \\
&= \operatorname{E}\left[b(X - \operatorname{E}[X])(X - \operatorname{E}[X])^\top b^\top\right] \\
&= b\operatorname{E}\left[(X - \operatorname{E}[X])(X - \operatorname{E}[X])^\top\right] b^\top \\
&= b\operatorname{Var}[X]\, b^\top
\end{aligned}$$

■

23.4.3 Linear transformations

By combining the two previous properties, one obtains the following proposition.

Proposition 132 *Let $a \in \mathbb{R}^K$ be a constant $K \times 1$ vector, b be a constant $M \times K$ matrix and X a $K \times 1$ random vector having covariance matrix* $\operatorname{Var}[X]$*. Then,*

$$\operatorname{Var}[a + bX] = b\operatorname{Var}[X]\, b^\top$$

[4] See the property of the expected value at p. 148.
[5] See the property of the expected value at p. 149.

23.4.4 Symmetry

The covariance matrix is a symmetric matrix, i.e., it is equal to its transpose.

Proposition 133 *The covariance matrix of a random vector X satisfies*

$$\operatorname{Var}[X]^{\top} = \operatorname{Var}[X]$$

Proof. This is proved as follows:

$$\begin{aligned}
\operatorname{Var}[X]^{\top} &= \operatorname{E}\left[(X - \operatorname{E}[X])(X - \operatorname{E}[X])^{\top}\right]^{\top} \\
&= \operatorname{E}\left[\left((X - \operatorname{E}[X])(X - \operatorname{E}[X])^{\top}\right)^{\top}\right] \\
&= \operatorname{E}\left[(X - \operatorname{E}[X])(X - \operatorname{E}[X])^{\top}\right] \\
&= \operatorname{Var}[X]
\end{aligned}$$

■

23.4.5 Semi-positive definiteness

Proposition 134 *The covariance matrix of a random vector X is a positive-semidefinite matrix, i.e., it holds that*

$$a\operatorname{Var}[X]\,a^{\top} \geq 0$$

for any $1 \times K$ vector $a \in \mathbb{R}^K$.

Proof. This is easily proved using property 23.4.2 above:

$$a\operatorname{Var}[X]\,a^{\top} = \operatorname{Var}[aX] \geq 0$$

where the last inequality follows from the fact that variance is always positive. ■

23.4.6 Covariance between linear transformations

Proposition 135 *Let a and b be two constant $1 \times K$ vectors and X a $K \times 1$ random vector. Then, the covariance between the two linear transformations aX and bX can be expressed as a function of the covariance matrix:*

$$\operatorname{Cov}[aX, bX] = a\operatorname{Var}[X]\,b^{\top}$$

Proof. This can be proved as follows:

$$\begin{aligned}
& \operatorname{Cov}[aX, bX] \\
\boxed{\text{A}} \quad &= \operatorname{E}\left[(aX - \operatorname{E}[aX])(bX - \operatorname{E}[bX])\right] \\
&= \operatorname{E}\left[a(X - \operatorname{E}[X])\,b(X - \operatorname{E}[X])\right] \\
\boxed{\text{B}} \quad &= \operatorname{E}\left[a(X - \operatorname{E}[X])(b(X - \operatorname{E}[X]))^{\top}\right] \\
\boxed{\text{C}} \quad &= \operatorname{E}\left[a(X - \operatorname{E}[X])(X - \operatorname{E}[X])^{\top} b^{\top}\right]
\end{aligned}$$

$$\begin{aligned} \boxed{\text{D}} &= a\mathrm{E}\left[(X-\mathrm{E}[X])(X-\mathrm{E}[X])^{\top}\right]b^{\top} \\ \boxed{\text{E}} &= a\mathrm{Var}[X]\,b^{\top} \end{aligned}$$

where: in step A we have used the definition of covariance; in step B we have used the fact that the transpose of a scalar is equal to the scalar itself; in step C we have used the formula for the transpose of a product; in step D we have used the linearity of the expected value; in step E we have used the definition of covariance matrix. ■

23.4.7 Cross-covariance

The term covariance matrix is sometimes also referred to the matrix of covariances between the elements of two vectors. Let X be a $K \times 1$ random vector and Y be a $L \times 1$ random vector. The **covariance matrix** between X and Y, or **cross-covariance** between X and Y, denoted by $\mathrm{Cov}[X,Y]$, is defined as follows:

$$\mathrm{Cov}[X,Y] = \mathrm{E}\left[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])^{\top}\right]$$

provided the above expected value exists and is well-defined. It is a multivariate generalization of the definition of covariance between two scalar random variables. Let $X_1, \ldots, X_K$ denote the K components of the vector X and $Y_1, \ldots, Y_L$ denote the L components of the vector Y . From the definition of $\mathrm{Cov}[X,Y]$, it can easily be seen that $\mathrm{Cov}[X,Y]$ is a $K \times L$ matrix with the following structure:

$$\mathrm{Cov}[X,Y] = \begin{bmatrix} V_{11} & V_{12} & \ldots & V_{1L} \\ V_{21} & V_{22} & \ldots & V_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ V_{K1} & V_{K2} & \ldots & V_{KL} \end{bmatrix}$$

where the (i,j)-th entry of the matrix is equal to the covariance between X_i and Y_j:

$$V_{ij} = \mathrm{E}\left[(X_i-\mathrm{E}[X_i])(Y_j-\mathrm{E}[Y_j])\right] = \mathrm{Cov}[X_i,Y_j]$$

Note that $\mathrm{Cov}[X,Y]$ is not the same as $\mathrm{Cov}[Y,X]$. In fact, $\mathrm{Cov}[Y,X]$ is a $L \times K$ matrix equal to the transpose of $\mathrm{Cov}[X,Y]$:

$$\begin{aligned} \mathrm{Cov}[Y,X] &= \mathrm{E}\left[(Y-\mathrm{E}[Y])(X-\mathrm{E}[X])^{\top}\right] \\ &= \mathrm{E}\left[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])^{\top}\right]^{\top} \\ &= \mathrm{Cov}[X,Y]^{\top} \end{aligned}$$

23.5 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Let X be a 2×1 random vector and denote its components by X_1 and X_2. The covariance matrix of X is

$$\text{Var}[X] = \begin{bmatrix} 4 & 1 \\ 1 & 2 \end{bmatrix}$$

Compute the variance of the random variable Y defined as

$$Y = 3X_1 + 4X_2$$

Solution

Using a matrix notation, Y can be written as

$$Y = \begin{bmatrix} 3 & 4 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = bX$$

where we have defined

$$b = \begin{bmatrix} 3 & 4 \end{bmatrix}$$

Therefore, the variance of Y can be computed by using the formula for the covariance matrix of a linear transformation:

$$\begin{aligned} \text{Var}[Y] &= \text{Var}[bX] = b\text{Var}[X]\, b^\top \\ &= \begin{bmatrix} 3 & 4 \end{bmatrix} \begin{bmatrix} 4 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ 4 \end{bmatrix} \\ &= \begin{bmatrix} 3 & 4 \end{bmatrix} \begin{bmatrix} 4 \cdot 3 + 1 \cdot 4 \\ 1 \cdot 3 + 2 \cdot 4 \end{bmatrix} \\ &= \begin{bmatrix} 3 & 4 \end{bmatrix} \begin{bmatrix} 16 \\ 11 \end{bmatrix} \\ &= 3 \cdot 16 + 4 \cdot 11 = 92 \end{aligned}$$

Exercise 2

Let X be a 3×1 random vector and denote its components by X_1, X_2 and X_3. The covariance matrix of X is

$$\text{Var}[X] = \begin{bmatrix} 3 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Compute the following covariance:

$$\text{Cov}[X_1 + 2X_3, 3X_2]$$

Solution

Using the bilinearity of the covariance operator[6], we obtain

$$\begin{aligned} & \text{Cov}[X_1 + 2X_3, 3X_2] \\ = \; & \text{Cov}[X_1, 3X_2] + 2\text{Cov}[X_3, 3X_2] \end{aligned}$$

[6] See p. 166.

$$\begin{aligned} &= 3\text{Cov}\left[X_1, X_2\right] + 6\text{Cov}\left[X_3, X_2\right] \\ &= 3 \cdot 1 + 6 \cdot 0 = 3 \end{aligned}$$

The same result can be obtained using the formula for the covariance between two linear transformations. Let us define

$$\begin{aligned} a &= \begin{bmatrix} 1 & 0 & 2 \end{bmatrix} \\ b &= \begin{bmatrix} 0 & 3 & 0 \end{bmatrix} \end{aligned}$$

Then, we have

$$\begin{aligned} &\text{Cov}\left[X_1 + 2X_3, 3X_2\right] = \text{Cov}\left[aX, bX\right] = a\text{Var}\left[X\right]b^\top \\ = &\begin{bmatrix} 1 & 0 & 2 \end{bmatrix} \begin{bmatrix} 3 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 3 \\ 0 \end{bmatrix} \\ = &\begin{bmatrix} 1 & 0 & 2 \end{bmatrix} \begin{bmatrix} 3 \cdot 0 + 1 \cdot 3 + 0 \cdot 0 \\ 1 \cdot 0 + 2 \cdot 3 + 0 \cdot 0 \\ 0 \cdot 0 + 0 \cdot 3 + 1 \cdot 0 \end{bmatrix} \\ = &\begin{bmatrix} 1 & 0 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ 6 \\ 0 \end{bmatrix} = 1 \cdot 3 + 0 \cdot 6 + 2 \cdot 0 = 3 \end{aligned}$$

Exercise 3

Let X be a $K \times 1$ random vector whose covariance matrix is equal to the identity matrix:

$$\text{Var}\left[X\right] = I$$

Define a new random vector Y as follows:

$$Y = AX$$

where A is a $K \times K$ matrix of constants such that

$$AA^\top = I$$

Derive the covariance matrix of Y.

Solution

By using the formula for the covariance matrix of a linear transformation, we obtain

$$\begin{aligned} \text{Var}\left[Y\right] &= \text{Var}\left[AX\right] = A\text{Var}\left[X\right]A^\top \\ &= AIA^\top = AA^\top = I \end{aligned}$$

Chapter 24

Indicator function

This lecture introduces the concept of indicator function. Before reading this lecture, make sure you are familiar with the concepts of random variable (p. 105) and expected value (p. 127).

24.1 Definition

Let Ω be a sample space[1], let $E \subseteq \Omega$ be an event and denote by $\mathrm{P}(E)$ the probability assigned to the event E. The **indicator function** of the event E (or **indicator random variable** of the event E), denoted by 1_E, is a random variable defined as follows:

$$1_E(\omega) = \begin{cases} 1 & \text{if } \omega \in E \\ 0 & \text{if } \omega \notin E \end{cases}$$

In other words, the indicator function of the event E is a random variable that takes value 1 when the event E happens and value 0 when the event E does not happen.

Example 136 *We toss a die and one of the six numbers from* 1 *to* 6 *can appear face up. The sample space is:*

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

Define the event

$$E = \{1, 3, 5\}$$

i.e. E is the event "An odd number appears face up". A random variable that takes value 1 *when an odd number appears face up and value* 0 *otherwise is an indicator of the event E.*

From the above definition, it can easily be seen that 1_E is a discrete random variable[2] with support $R_{1_E} = \{0, 1\}$ and probability mass function:

$$p_{1_E}(x) = \begin{cases} \mathrm{P}(E) & \text{if } x = 1 \\ \mathrm{P}(E^c) = 1 - \mathrm{P}(E) & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases}$$

Indicator functions are heavily used in probability theory to simplify notation and to prove theorems.

[1] See p. 69.
[2] See p. 106.

24.2 Properties of the indicator function

Indicator functions enjoy the following properties.

24.2.1 Powers

The n-th power of 1_E is equal to 1_E:

$$\left(1_E(\omega)\right)^n = 1_E(\omega), \forall n, \omega$$

because 1_E can be either 0 or 1 and

$$\begin{aligned} 0^n &= 0 \\ 1^n &= 1 \end{aligned}$$

24.2.2 Expected value

The expected value of 1_E is equal to $\mathrm{P}(E)$:

$$\begin{aligned} \mathrm{E}[1_E] &= \sum_{x \in R_{1_E}} x p_{1_E}(x) \\ &= 1 \cdot p_{1_E}(1) + 0 \cdot p_{1_E}(0) \\ &= 1 \cdot \mathrm{P}(E) + 0 \cdot \mathrm{P}(E^c) \\ &= \mathrm{P}(E) \end{aligned}$$

24.2.3 Variance

The variance of 1_E is equal to $\mathrm{P}(E) \cdot (1 - \mathrm{P}(E))$. Using the powers property above and the formula for computing the variance[3]:

$$\begin{aligned} \mathrm{Var}[1_E] &= \mathrm{E}\left[(1_E)^2\right] - \mathrm{E}[1_E]^2 \\ &= \mathrm{E}[1_E] - \mathrm{E}[1_E]^2 \\ &= \mathrm{P}(E) - \mathrm{P}(E)^2 \\ &= \mathrm{P}(E) \cdot (1 - \mathrm{P}(E)) \end{aligned}$$

24.2.4 Intersections

If E and F are two events, then:

$$1_{E \cap F} = 1_E 1_F$$

In fact:

1. if $\omega \in E \cap F$, then

$$1_{E \cap F}(\omega) = 1$$

and

$$\begin{aligned} & \omega \in E, \omega \in F \\ \Longrightarrow \quad & 1_E(\omega) = 1, 1_F(\omega) = 1 \\ \Longrightarrow \quad & 1_E(\omega)\, 1_F(\omega) = 1 \end{aligned}$$

[3] See p. 156.

2. if $\omega \notin E \cap F$, then

$$1_{E \cap F}(\omega) = 0$$

and

$$\begin{aligned} & \text{either } \omega \notin E \text{ or } \omega \notin F \text{ or both} \\ \Longrightarrow \quad & \text{either } 1_E(\omega) = 0 \text{ or } 1_F(\omega) = 0 \text{ or both} \\ \Longrightarrow \quad & 1_E(\omega)\, 1_F(\omega) = 0 \end{aligned}$$

24.2.5 Indicators of zero-probability events

Let E be a zero-probability event[4] and X an integrable random variable[5]. Then:

$$\mathrm{E}[X 1_E] = 0$$

While a rigorous proof of this fact is beyond the scope of this introductory exposition, this property should be intuitive. The random variable $(X 1_E)(\omega)$ is equal to zero for all sample points ω except possibly for the points $\omega \in E$. The expected value is a weighted average of the values $X 1_E$ can take on, where each value is weighted by its respective probability. The non-zero values $X 1_E$ can take on are weighted by zero probabilities, so $\mathrm{E}[X 1_E]$ must be zero.

24.3 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Consider a random variable X and another random variable Y defined as a function of X.

$$Y = \begin{cases} 2 & \text{if } X < 2 \\ X & \text{if } X \geq 2 \end{cases}$$

Express Y using the indicator functions of the events $\{X < 2\}$ and $\{X \geq 2\}$.

Solution

Denote by $1_{\{X<2\}}$ the indicator of the event $\{X < 2\}$ and denote by $1_{\{X \geq 2\}}$ the indicator of the event $\{X \geq 2\}$. We can write Y as:

$$Y = 2 \cdot 1_{\{X<2\}} + X \cdot 1_{\{X \geq 2\}}$$

Exercise 2

Let X be a positive random variable, i.e. a random variable that can take on only positive values. Let c be a constant. Prove that

$$\mathrm{E}[X] \geq \mathrm{E}\left[X 1_{\{X \geq c\}}\right]$$

where $1_{\{X \geq c\}}$ is the indicator of the event $\{X \geq c\}$.

[4] See p. 79.
[5] See p. 136.

Solution

First note that the sum of the indicators $1_{\{X \geq c\}}$ and $1_{\{X<c\}}$ is always equal to 1:

$$1_{\{X \geq c\}} + 1_{\{X<c\}} = 1$$

As a consequence, we can write:

$$\begin{aligned} \mathrm{E}[X] &= \mathrm{E}[X \cdot 1] \\ &= \mathrm{E}\left[X \cdot \left(1_{\{X \geq c\}} + 1_{\{X<c\}}\right)\right] \\ &= \mathrm{E}\left[X 1_{\{X \geq c\}}\right] + \mathrm{E}\left[X 1_{\{X<c\}}\right] \end{aligned}$$

Now, note that $X 1_{\{X<c\}}$ is a positive random variable and that the expected value of a positive random variable is positive[6]:

$$\mathrm{E}\left[X 1_{\{X<c\}}\right] \geq 0$$

Thus:

$$\mathrm{E}[X] = \mathrm{E}\left[X 1_{\{X \geq c\}}\right] + \mathrm{E}\left[X 1_{\{X<c\}}\right] \geq \mathrm{E}\left[X 1_{\{X \geq c\}}\right]$$

Exercise 3

Let E be an event and denote its indicator function by 1_E. Let E^c be the complement of E and denote its indicator function by 1_{E^c}. Can you express 1_{E^c} as a function of 1_E?

Solution

The sum of the two indicators is always equal to 1:

$$1_E + 1_{E^c} = 1$$

Therefore:

$$1_{E^c} = 1 - 1_E$$

[6] See p. 150.

Chapter 25

Conditional probability as a random variable

In the lecture entitled *Conditional probability* (p. 85) we have stated a number of properties that conditional probabilities shoud satisfy to be rational in some sense. We have proved that, whenever $\mathrm{P}(G) > 0$, these properties are satisfied if and only if

$$\mathrm{P}(E\,|G) = \frac{\mathrm{P}(E \cap G)}{\mathrm{P}(G)}$$

but we have not been able to derive a formula for probabilities conditional on zero-probability events[1], i.e. we have not been able to find a way to compute $\mathrm{P}(E\,|G)$ when $\mathrm{P}(G) = 0$.

Thus, we have concluded that the above elementary formula cannot be taken as a general definition of conditional probability, because it does not cover zero-probability events.

In this lecture we discuss a completely general definition of conditional probability, which covers also the case in which $\mathrm{P}(G) = 0$.

The plan of the lecture is as follows.

1. We define the concept of a partition of events.

2. We show that, given a partition of events, conditional probability can be regarded as a random variable (probability conditional on a partition).

3. We show that, when no zero-probability events are involved, probabilities conditional on a partition satisfy a certain property (the fundamental property of conditional probability).

4. We require that the fundamental property of conditional probability be satisfied also when zero-probability events are involved and we show that this requirement is sufficient to unambiguously pin down probabilities conditional on a partition. This requirement can therefore be used to give a completely general defintion of conditional probability.

[1] See p. 79.

25.1 Partitions of events

Let Ω be a sample space[2] and let $\mathrm{P}(E)$ denote the probability assigned to the events $E \subseteq \Omega$.

Define a partition of events of Ω as follows:

Definition 137 *Let $\mathcal{G}$ be a collection of non-empty subsets of Ω. $\mathcal{G}$ is called a **partition of events of** Ω if*

1. *all subsets $G \in \mathcal{G}$ are events;*
2. *if $G, F \in \mathcal{G}$ then either $G = F$ or $G \cap F = \emptyset$;*
3. $\Omega = \bigcup_{G \in \mathcal{G}} G$.

In other words, $\mathcal{G}$ is a partition of events of Ω if it is a division of Ω into non-overlapping and non-empty events that cover all of Ω.

A partition of events of Ω is said to be **finite** if there are a finite number of sets G in the partition; it is said to be **infinite** if there are an infinite number of sets G in the partition; it is said to be **countable** if the sets G in the partition are countable; it is said to be **arbitrary** if the number of sets G in the partition is not necessarily finite or countable (i.e. it can be uncountable).

Example 138 *Suppose that we toss a die. Six numbers (from 1 to 6) can appear face up, but we do not yet know which one of them will appear. The sample space is*

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

Let any subset of Ω be considered an event. Define the two events:

$$\begin{aligned} G_1 &= \{1, 2, 3\} \\ G_2 &= \{4, 5, 6\} \end{aligned}$$

Then, $\mathcal{G} = \{G_1, G_2\}$ is a partition of events of Ω:

$$\begin{aligned} \bigcup_{G \in \mathcal{G}} G &= G_1 \cup G_2 = \{1, 2, 3\} \cup \{4, 5, 6\} \\ &= \{1, 2, 3, 4, 5, 6\} = \Omega \end{aligned}$$

and $G_1 \cap G_2 = \emptyset$. Now define the the three events:

$$\begin{aligned} F_1 &= \{1, 2\} \\ F_2 &= \{3, 4, 5\} \\ F_3 &= \{5, 6\} \end{aligned}$$

and the collection $\mathcal{F} = \{F_1, F_2, F_3\}$. $\mathcal{F}$ is not a partition of events of Ω: while condition 1 and 3 in the definition above are satisfied, condition 2 is not, because

$$F_2 \cap F_3 = \{5\} \neq \emptyset \text{ and } F_2 \neq F_3$$

[2] See p. 69.

25.2 Probabilities conditional on a partition

Suppose we are given a finite partition

$$\mathcal{G} = \{G_1, G_2, \ldots, G_n\}$$

of events of Ω, such that $\mathrm{P}(G_i) > 0$ for every i.

Suppose that we are interested in the probability of a specific event $E \subseteq \Omega$ and that at a certain time in the future we will receive some information about the realized outcome[3] $\overline{\omega}$. In particular, we will be told to which one of the n subsets $G_1, G_2, \ldots, G_n$ the realized outcome $\overline{\omega}$ belongs. When we receive the information that the realized outcome belongs to the set G_i, we will update our assessment of the probability of the event E, computing its conditional probability:

$$\mathrm{P}(E \,|G_i) = \frac{\mathrm{P}(E \cap G_i)}{\mathrm{P}(G_i)}$$

Before receiving the information, this conditional probability is unknown and can be regarded as a random variable, denoted by $\mathrm{P}(E \,|\mathcal{G})$ and defined as follows:

$$\mathrm{P}(E \,|\mathcal{G})(\omega) = \begin{cases} \mathrm{P}(E \,|G_1) & \text{if } \omega \in G_1 \\ \vdots & \\ \mathrm{P}(E \,|G_n) & \text{if } \omega \in G_n \end{cases} \tag{25.1}$$

$\mathrm{P}(E \,|\mathcal{G})$ is the **probability of E conditional on the partition $\mathcal{G}$.**

Example 139 *Let us continue with Example 138, where the sample space is*

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

and $\mathcal{G} = \{G_1, G_2\}$ is a partition of events of Ω with

$$\begin{aligned} G_1 &= \{1, 2, 3\} \\ G_2 &= \{4, 5, 6\} \end{aligned}$$

Let us assign equal probability to all the outcomes:

$$\mathrm{P}(\omega) = \frac{1}{6} \quad , \forall \omega \in \Omega$$

Let us now analyze the conditional probability of the event

$$E = \{2, 3, 4\}$$

We have:

$$\begin{aligned} \mathrm{P}(E \,|G_1) &= \frac{\mathrm{P}(E \cap G_1)}{\mathrm{P}(G_1)} = \frac{\mathrm{P}(\{2,3\})}{\mathrm{P}(\{1,2,3\})} = \frac{2/6}{3/6} = \frac{2}{3} \\ \mathrm{P}(E \,|G_2) &= \frac{\mathrm{P}(E \cap G_2)}{\mathrm{P}(G_2)} = \frac{\mathrm{P}(\{4\})}{\mathrm{P}(\{4,5,6\})} = \frac{1/6}{3/6} = \frac{1}{3} \end{aligned}$$

The conditional probability $\mathrm{P}(E \,|\mathcal{G})$ is a random variable defined as follows:

$$\mathrm{P}(E \,|\mathcal{G})(\omega) = \begin{cases} 2/3 & \text{if } \omega \in \{1, 2, 3\} \\ 1/3 & \text{if } \omega \in \{4, 5, 6\} \end{cases}$$

[3] See p. 69.

Since $\mathrm{P}(G_1) = \mathrm{P}(G_2) = \frac{1}{2}$, *the probability mass function of* $\mathrm{P}(E|\mathcal{G})$ *is*

$$p_{\mathrm{P}(E|\mathcal{G})}(x) = \begin{cases} 1/2 & \text{if } x = 2/3 \\ 1/2 & \text{if } x = 1/3 \\ 0 & \text{otherwise} \end{cases} \tag{25.2}$$

25.3 The fundamental property

A fundamental property of $\mathrm{P}(E|\mathcal{G})$ is that its expected value equals the unconditional probability $\mathrm{P}(E)$:

$$\mathrm{E}[\mathrm{P}(E|\mathcal{G})] = \mathrm{P}(E)$$

Proof. This is proved as follows:

$$\begin{aligned}
\mathrm{E}[\mathrm{P}(E|\mathcal{G})] &= \mathrm{P}(E|G_1)\cdot\mathrm{P}(\{\omega\in\Omega : \mathrm{P}(E|\mathcal{G})(\omega) = \mathrm{P}(E|G_1)\}) \\
&\quad + \ldots + \mathrm{P}(E|G_n)\cdot\mathrm{P}(\{\omega\in\Omega : \mathrm{P}(E|\mathcal{G})(\omega) = \mathrm{P}(E|G_n)\}) \\
&= \mathrm{P}(E|G_1)\cdot\mathrm{P}(G_1) + \ldots + \mathrm{P}(E|G_n)\cdot\mathrm{P}(G_n) \\
&= \frac{\mathrm{P}(E\cap G_1)}{\mathrm{P}(G_1)}\cdot\mathrm{P}(G_1) + \ldots + \frac{\mathrm{P}(E\cap G_n)}{\mathrm{P}(G_n)}\cdot\mathrm{P}(G_n) \\
&= \mathrm{P}(E\cap G_1) + \ldots + \mathrm{P}(E\cap G_n) \\
&= \sum_{i=1}^{n}\mathrm{P}(E\cap G_i) = \mathrm{P}\left(\bigcup_{i=1}^{n}(E\cap G_i)\right) \\
&= \mathrm{P}\left(E\cap\left(\bigcup_{i=1}^{n}G_i\right)\right) = \mathrm{P}(E\cap\Omega) = \mathrm{P}(E)
\end{aligned}$$

■

Example 140 *In Example 139, where the probability mass function of* $\mathrm{P}(E|\mathcal{G})$ *is (25.2), it is easy to verify the above property:*

$$\begin{aligned}
\mathrm{E}[\mathrm{P}(E|\mathcal{G})] &= \frac{2}{3}\cdot p_{\mathrm{P}(E|\mathcal{G})}\left(\frac{2}{3}\right) + \frac{1}{3}\cdot p_{\mathrm{P}(E|\mathcal{G})}\left(\frac{1}{3}\right) \\
&= \frac{2}{3}\cdot\frac{1}{2} + \frac{1}{3}\cdot\frac{1}{2} = \frac{1}{2} = \mathrm{P}(E)
\end{aligned}$$

The above property can be generalized as follows:

Proposition 141 (Fundamental property) *Let*

$$\mathcal{G} = \{G_1, G_2, \ldots, G_n\}$$

be a finite partition of events of Ω *such that* $\mathrm{P}(G_i) > 0$ *for every* i. *Let* H *be any event obtained as a union of events* $G_i \in \mathcal{G}$. *Let* 1_H *be the indicator function*[4] *of* H. *Let* $\mathrm{P}(E|\mathcal{G})$ *be defined as in (25.1). Then:*

$$\mathrm{E}[1_H\mathrm{P}(E|\mathcal{G})] = \mathrm{P}(E\cap H) \tag{25.3}$$

[4] See p. 197.

Proof. Without loss of generality, we can assume that H is obtained as the union of the first k ($k \leq n$) sets of the partition $\mathcal{G}$ (we can always re-arrange the sets G_i by changing their indices):

$$H = \bigcup_{i=1}^{k} G_i$$

First note that:

$$\left(1_H \mathrm{P}\left(E \,|\mathcal{G}\right)\right)(\omega) = \begin{cases} \mathrm{P}\left(E \,|G_1\right) & \text{if } \omega \in G_1 \\ \vdots & \\ \mathrm{P}\left(E \,|G_k\right) & \text{if } \omega \in G_k \\ 0 & \text{if } \omega \in G_{k+1} \\ \vdots & \\ 0 & \text{if } \omega \in G_n \end{cases}$$

The property is proved as follows:

$$\begin{aligned} \mathrm{E}\left[1_H \mathrm{P}\left(E \,|\mathcal{G}\right)\right] &= \mathrm{P}\left(E \,|G_1\right) \cdot \mathrm{P}\left(G_1\right) + \ldots + \mathrm{P}\left(E \,|G_k\right) \cdot \mathrm{P}\left(G_k\right) \\ &= \frac{\mathrm{P}\left(E \cap G_1\right)}{\mathrm{P}\left(G_1\right)} \cdot \mathrm{P}\left(G_1\right) + \ldots + \frac{\mathrm{P}\left(E \cap G_k\right)}{\mathrm{P}\left(G_k\right)} \cdot \mathrm{P}\left(G_k\right) \\ &= \mathrm{P}\left(E \cap G_1\right) + \ldots + \mathrm{P}\left(E \cap G_k\right) \\ &= \sum_{i=1}^{k} \mathrm{P}\left(E \cap G_i\right) = \mathrm{P}\left(\bigcup_{i=1}^{k} \left(E \cap G_i\right)\right) \\ &= \mathrm{P}\left(E \cap \left(\bigcup_{i=1}^{k} G_i\right)\right) = \mathrm{P}\left(E \cap H\right) \end{aligned}$$

■

25.4 The fundamental property as a definition

Suppose we are not able to explicitly define $\mathrm{P}\left(E \,|\mathcal{G}\right)$ as in (25.1). This can happen, for example, because $\mathcal{G}$ contains a zero-probability event G and, therefore, we cannot use the formula

$$\mathrm{P}\left(E \,|G\right) = \frac{\mathrm{P}\left(E \cap G\right)}{\mathrm{P}\left(G\right)}$$

to define $\mathrm{P}\left(E \,|\mathcal{G}\right)(\omega)$ for $\omega \in G$.

Although we are not able to explicitly define $\mathrm{P}\left(E \,|\mathcal{G}\right)$, we require, by analogy with the cases in which we are instead able to define it, that $\mathrm{P}\left(E \,|\mathcal{G}\right)$ satisfies the fundamental property (25.1). How can we be sure that there exists a random variable $\mathrm{P}\left(E \,|\mathcal{G}\right)$ satisfying this property? Existence is guaranteed by the following important theorem, that we state without providing a proof:

Proposition 142 (Existence of conditional probability) *Let $\mathcal{G}$ be an arbitrary partition of events of Ω. Let $E \in \Omega$ be an event. Then there exists at least one random variable Y that satisfies the property:*

$$\mathrm{E}\left[1_H Y\right] = \mathrm{P}\left(E \cap H\right)$$

for all events H obtainable as unions of events $G \in \mathcal{G}$. Furthermore, if two random variables Y_1 and Y_2 satisfy this property, i.e.:

$$\begin{aligned} \mathrm{E}\left[1_H Y_1\right] &= \mathrm{P}\left(E \cap H\right) \\ \mathrm{E}\left[1_H Y_2\right] &= \mathrm{P}\left(E \cap H\right) \end{aligned}$$

for all H, then Y_1 and Y_2 are almost surely equal[5]*.*

According to this proposition, a random variable Y satisfying the fundamental property of conditional probability exists and is unique, up to almost sure equality. As a consequence, we can indirectly define the probability of an event E conditional on the partition $\mathcal{G}$ as $\mathrm{P}\left(E\left|\mathcal{G}\right.\right) = Y$. This indirect way of defining conditional probability is summarized in the following:

Definition 143 (Probability conditional on a partition) *Let $\mathcal{G}$ be a partition of events of Ω. Let $E \in \Omega$ be an event. We say that a random variable $\mathrm{P}\left(E\left|\mathcal{G}\right.\right)$ is a **probability of E conditional on the partition $\mathcal{G}$** if*

$$\mathrm{E}\left[1_H \mathrm{P}\left(E\left|\mathcal{G}\right.\right)\right] = \mathrm{P}\left(E \cap H\right)$$

for all events H obtainable as unions of events $G \in \mathcal{G}$.

As we have seen above, such a random variable is guaranteed to exist and is unique up to almost sure equality.

This apparently abstract definition of conditional probability is extremely useful. One of its most important applications is the derivation of conditional probability density functions for absolutely continuous random vectors (see the lecture entitled *Conditional probability distributions* - p. 209).

25.5 More details

25.5.1 Conditioning with respect to sigma-algebras

In rigorous probability theory, when conditional probability is regarded as a random variable, it is defined with respect to sigma-algebras[6], rather than with respect to partitions. Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space[7]. Let $\mathcal{I}$ be a sub-σ-algebra of $\mathcal{F}$, i.e. $\mathcal{I}$ is a σ-algebra and $\mathcal{I} \subseteq \mathcal{F}$. Let $E \in \mathcal{F}$ be an event. We say that a random variable $\mathrm{P}\left(E\left|\mathcal{I}\right.\right)$ is a **conditional probability of E with respect to the σ-algebra $\mathcal{I}$** if

$$\begin{aligned} &\mathrm{P}\left(E\left|\mathcal{I}\right.\right) \text{ is } \mathcal{I}\text{-measurable} \\ &\mathrm{E}\left[1_H \mathrm{P}\left(E\left|\mathcal{I}\right.\right)\right] = \mathrm{P}\left(E \cap H\right) \text{ for any } H \in \mathcal{I} \end{aligned}$$

It can be shown that this definition is completely equivalent to our definition above, provided $\mathcal{I}$ is the smallest σ-algebra containing all the events $H \in \mathcal{F}$ obtainable as unions of events $G \in \mathcal{G}$ (where $\mathcal{G}$ is a partition of events of Ω).

[5] In other words, there exists a zero-probability event F such that:

$$\{\omega \in \Omega : Y_1(\omega) \neq Y_2(\omega)\} \subseteq F$$

See the lecture entitled *Zero-probability events* (p. 79) for a definition of almost sure events and zero-probability events.

[6] See p. 75.

[7] See p. 76.

25.5.2 Regular conditional probabilities

Let $\mathrm{P}(E\,|\mathcal{G})$ denote the probability of a generic event E conditional on the partition $\mathcal{G}$. We say that the probability space $(\Omega, \mathcal{F}, \mathrm{P})$ admits a **regular probability conditional on the partition** $\mathcal{G}$ if, for any fixed ω, $\mathrm{P}(E\,|\mathcal{G})(\omega)$ is a probability measure on the events E, i.e. for any ω, $(\Omega, \mathcal{F}, \mathrm{P}(E\,|\mathcal{G})(\omega))$ is a probability space.

Chapter 26

Conditional probability distributions

To understand conditional probability distributions, you need to be familiar with the concept of conditional probability, which has been introduced in the lecture entitled *Conditional probability* (p. 85).

We discuss here how to update the probability distribution of a random variable X after observing the realization of another random variable Y, i.e. after receiving the information that another random variable Y has taken a specific value y. The updated probability distribution of X will be called the conditional probability distribution of X given $Y = y$.

The two random variables X and Y, considered together, form a random vector $[X \;\; Y]$. Depending on the characteristics of the random vector $[X \;\; Y]$, different procedures need to be adopted in order to compute the conditional probability distribution of X given $Y = y$. In the remainder of this lecture, these procedures are presented in the following order:

1. first, we tackle the case in which the random vector $[X \;\; Y]$ is a discrete random vector[1];

2. then, we tackle the case in which $[X \;\; Y]$ is an absolutely continuous random vector[2];

3. finally, we briefly discuss the case in which $[X \;\; Y]$ is neither discrete nor absolutely continuous.

Important: note that if we are able to update the probability distribution of X when we observe the realization of Y (i.e. when we receive the information that $Y = y$), then we are also able to update the probability distribution of X when we receive the information that a generic event E has happened: it suffices to set $Y = 1_E$, where 1_E is the indicator function[3] of the event E, and update the distribution of X based on the information $Y = 1_E = 1$.

[1] See p. 116.
[2] See p. 117.
[3] See p. 197.

26.1 Conditional probability mass function

In the case in which $[X \;\; Y]$ is a discrete random vector (as a consequence X is a discrete random variable), the probability mass function[4] of X conditional on the information that $Y = y$ is called conditional probability mass function:

Definition 144 *Let $[X \;\; Y]$ be a discrete random vector. We say that a function $p_{X|Y=y} : \mathbb{R} \to [0,1]$ is the **conditional probability mass function** of X given $Y = y$ if, for any $x \in \mathbb{R}$:*

$$p_{X|Y=y}(x) = \mathrm{P}(X = x \,|Y = y)$$

where $\mathrm{P}(X = x \,|Y = y)$ is the conditional probability that $X = x$ given that $Y = y$.

How do we derive the conditional probability mass function from the joint probability mass function[5] $p_{XY}(x, y)$? The following proposition provides an answer to this question:

Proposition 145 *Let $[X \;\; Y]$ be a discrete random vector. Let $p_{XY}(x, y)$ be its joint probability mass function and let $p_Y(y)$ be the marginal probability mass function[6] of Y. The conditional probability mass function of X given $Y = y$ is*

$$p_{X|Y=y}(x) = \frac{p_{XY}(x, y)}{p_Y(y)}$$

provided $p_Y(y) > 0$.

Proof. This is just the usual formula for computing conditional probabilities (conditional probability equals joint probability divided by marginal probability):

$$\begin{aligned} p_{X|Y=y}(x) &= \mathrm{P}(X = x \,|Y = y) \\ &= \frac{\mathrm{P}(X = x \text{ and } Y = y)}{\mathrm{P}(Y = y)} \\ &= \frac{p_{XY}(x, y)}{p_Y(y)} \end{aligned}$$

■

Note that the above proposition assumes knowledge of the marginal probability mass function $p_Y(y)$, which can be derived from the joint probability mass function $p_{XY}(x, y)$ by marginalization[7].

Example 146 *Let the support of $[X \;\; Y]$ be*

$$R_{XY} = \{[1 \;\; 1], [2 \;\; 0], [0 \;\; 0]\}$$

and its joint probability mass function be

$$p_{XY}(x, y) = \begin{cases} 1/3 & \text{if } x = 1 \text{ and } y = 1 \\ 1/3 & \text{if } x = 2 \text{ and } y = 0 \\ 1/3 & \text{if } x = 0 \text{ and } y = 0 \\ 0 & \text{otherwise} \end{cases}$$

[4] See p. 106.
[5] See p. 116
[6] See p. 120
[7] See p. 120

Let us compute the conditional probability mass function of X given $Y = 0$. The support of Y is

$$R_Y = \{0, 1\}$$

The marginal probability mass function of Y evaluated at $y = 0$ is

$$\begin{aligned} p_Y(0) &= \sum_{\{(x,y) \in R_{XY} : y=0\}} p_{XY}(x, y) \\ &= p_{XY}(2, 0) + p_{XY}(0, 0) = \frac{2}{3} \end{aligned}$$

The support of X is

$$R_X = \{0, 1, 2\}$$

Thus, the conditional probability mass function of X given $Y = 0$ is

$$p_{X|Y=0}(x) = \begin{cases} \frac{p_{XY}(0,0)}{p_Y(0)} = \frac{1/3}{2/3} = \frac{1}{2} & \text{if } x = 0 \\ \frac{p_{XY}(1,0)}{p_Y(0)} = \frac{0}{2/3} = 0 & \text{if } x = 1 \\ \frac{p_{XY}(2,0)}{p_Y(0)} = \frac{1/3}{2/3} = \frac{1}{2} & \text{if } x = 2 \\ \frac{p_{XY}(x,0)}{p_Y(0)} = \frac{0}{2/3} = 0 & \text{if } x \notin R_X \end{cases}$$

In the case in which $p_Y(y) = 0$, there is, in general, no way to unambiguously derive the conditional probability mass function of X, as we will show below with an example. The impossibility of deriving the conditional probability mass function unambiguously in this case (called by some authors the Borel-Kolmogorov paradox) is not particularly worrying, as this case is seldom relevant in applications. The following is an example of a case in which the conditional probability mass function cannot be derived unambiguously (the example is a bit involved; the reader might safely skip it on a first reading).

Example 147 *Suppose we are given the following sample space:*

$$\Omega = [0, 1]$$

i.e. the sample space Ω is the set of all real numbers between 0 and 1. It is possible to build a probability measure P on Ω, such that P assigns to each sub-interval of $[0, 1]$ a probability equal to its length, i.e.:

$$\text{if } 0 \leq a \leq b \leq 1 \text{ and } E = [a, b], \text{ then } \mathrm{P}(E) = b - a$$

This is the same sample space discussed in the lecture on zero-probability events[8]*. Define a random variable X as follows:*

$$X(\omega) = \begin{cases} 1 & \text{if } \omega = 0 \\ 0 & \text{otherwise} \end{cases}$$

and another random variable Y as follows:

$$Y(\omega) = \begin{cases} 1 & \text{if } \omega = 0 \\ 1 & \text{if } \omega = 1 \\ 0 & \text{otherwise} \end{cases}$$

[8] See p. 79.

Both X and Y are discrete random variables and, considered together, they constitute a discrete random vector $[X \;\; Y]$. Suppose we want to compute the conditional probability mass function of X conditional on $Y = 1$. It is easy to see that $p_Y(1) = 0$. As a consequence, we cannot use the formula:

$$p_{X|Y=1}(x) = \frac{p_{XY}(x,1)}{p_Y(1)}$$

because division by zero is not possible. It turns out that also the technique of implicitly deriving a conditional probability as a realization of a random variable satisfying the definition of a conditional probability with respect to a partition (see the lecture entitled Conditional probability as a random variable - p. 201) does not allow to unambiguously derive $p_{X|Y=1}(x)$. In this case, the partition of interest is $\mathcal{G} = \{G_1, G_2\}$, where:

$$\begin{aligned} G_1 &= \{\omega \in \Omega : Y(\omega) = 1\} = \{0,1\} \\ G_2 &= \{\omega \in \Omega : Y(\omega) = 0\} = (0,1) \end{aligned}$$

and $p_{X|Y=1}(x)$ can be viewed as the realization of the conditional probability

$$\mathrm{P}(X = x \,|\mathcal{G}\,)(\omega)$$

when $\omega \in G_1$. The fundamental property of conditional probability[9] is satisfied in this case if and only if, for a given x, the following system of equations is satisfied:

$$\begin{cases} p_{X|Y=1}(x)\, p_Y(1) = p_{XY}(x,1) \\ p_{X|Y=0}(x)\, p_Y(0) = p_{XY}(x,0) \end{cases}$$

which implies:

$$\begin{cases} p_{X|Y=1}(x) \cdot 0 = 0 \\ p_{X|Y=0}(x) = p_{XY}(x,0) \end{cases}$$

The second equation does not help determining $p_{X|Y=1}(x)$. So, from the first equation it is evident that $p_{X|Y=1}(x)$ is undetermined (any number, when multiplied by zero, gives zero). One can show that also the requirement that

$$\mathrm{P}(X = x \,|\mathcal{G}\,)$$

be a regular conditional probability[10] does not help to pin down

$$p_{X|Y=1}(x) \tag{26.1}$$

What does it mean that (26.1) is undetermined? It means that any choice of (26.1) is legitimate, provided the requirement

$$0 \leq p_{X|Y=1}(x) \leq 1$$

is satisfied. Is this really a paradox? No, because conditional probability with respect to a partition is defined up to almost sure equality, G_1 is a zero-probability event, so the value that $\mathrm{P}(E \,|\mathcal{G}\,)$ takes on G_1 does not matter (roughly speaking, we do not really need to care about zero-probability events, provided there is only a countable number of them).

[9] $\mathrm{E}[1_H \mathrm{P}(E \,|\mathcal{G}\,)] = \mathrm{P}(E \cap H)$ - see p. 204.

[10] See p. 207.

26.2 Conditional probability density function

In the case in which $[X \;\; Y]$ is an absolutely continuous random vector (as a consequence X is an absolutely continuous random variable), the probability density function[11] of X conditional on the information that $Y = y$ is called conditional probability density function:

Definition 148 *Let $[X \;\; Y]$ be an absolutely continuous random vector. We say that a function $f_{X|Y=y} : \mathbb{R} \to [0,1]$ is the **conditional probability density function** of X given $Y = y$ if, for any interval $[a,b] \subseteq \mathbb{R}$:*

$$\mathrm{P}\left(X \in [a,b] \,|Y = y\right) = \int_a^b f_{X|Y=y}\left(x\right) dx$$

and $f_{X|Y=y}$ is such that the above integral is well defined.

How do we derive the conditional probability mass function from the joint probability density function[12] $f_{XY}(x,y)$?

Deriving the conditional distribution of X given $Y = y$ is far from obvious: whatever value of y we choose, we are conditioning on a zero-probability event ($\mathrm{P}(Y = y) = 0$ - see p. 109 for an explanation); therefore, the standard formula (conditional probability equals joint probability divided by marginal probability) cannot be used. However, it turns out that the definition of conditional probability with respect to a partition[13] can be fruitfully applied in this case to derive the conditional probability density function of X given $Y = y$:

Proposition 149 *Let $[X \;\; Y]$ be an absolutely continuous random vector. Let $f_{XY}(x,y)$ be its joint probability density function, and $f_Y(y)$ be the marginal probability density function of Y. The conditional probability density function of X given $Y = y$ is*

$$f_{X|Y=y}\left(x\right) = \frac{f_{XY}\left(x,y\right)}{f_Y\left(y\right)}$$

provided $f_Y(y) > 0$.

Proof. To prove that

$$f_{X|Y=y}\left(x\right) = \frac{f_{XY}\left(x,y\right)}{f_Y\left(y\right)}$$

is a legitimate choice, we need to prove that conditional probabilities calculated using the above conditional density function satisfy the fundamental property of conditional probability:

$$\mathrm{E}\left[1_H \mathrm{P}\left(E\,|\mathcal{G}\right)\right] = \mathrm{P}\left(E \cap H\right)$$

for any H and E. Thanks to some basic results in measure theory, we can confine our attention to the events H and E that can be written as follows:

$$\begin{aligned} H &= \left\{\omega \in \Omega : Y \in [y_1, y_2], [y_1, y_2] \subseteq R_Y\right\} \\ E &= \left\{\omega \in \Omega : X \in [x_1, x_2], [x_1, x_2] \subseteq R_X\right\} \end{aligned}$$

[11] See p. 107.
[12] See p. 117.
[13] See p. 206.

For these events, it is immediate to verify that the fundamental property of conditional probability holds. First, by the very definition of a conditional probability density function:

$$\mathrm{P}(E\,|\mathcal{G}) = \int_{x_1}^{x_2} f_{X|Y=y}(x)\,dx$$

Furthermore, $1_H = 1_{\{Y\in[y_1,y_2]\}}$ is also a function of Y. Therefore, the product $1_H \mathrm{P}(E\,|\mathcal{G})$ is a function of Y , so we can use the transformation theorem[14] to compute its expected value:

$$\begin{aligned}
\mathrm{E}\left[1_H \mathrm{P}(E\,|\mathcal{G})\right] &= \int_{-\infty}^{\infty} \left(1_{\{y\in[y_1,y_2]\}} \int_{x_1}^{x_2} f_{X|Y=y}(x)\,dx\right) f_Y(y)\,dy \\
&\quad \int_{y_1}^{y_2} \left(\int_{x_1}^{x_2} f_{X|Y=y}(x)\,dx\right) f_Y(y)\,dy \\
&= \int_{y_1}^{y_2}\int_{x_1}^{x_2} f_{X|Y=y}(x)\, f_Y(y)\,dxdy \\
&= \int_{y_1}^{y_2}\int_{x_1}^{x_2} \frac{f_{XY}(x,y)}{f_Y(y)} f_Y(y)\,dxdy \\
&= \int_{y_1}^{y_2}\int_{x_1}^{x_2} f_{XY}(x,y)\,dxdy \\
&= \mathrm{P}(X\in[x_1,x_2], Y\in[y_1,y_2]) \\
&= \mathrm{P}(E\cap H)
\end{aligned}$$

The last equality proves the proposition. ■

Example 150 *Let the support of* $[X \;\; Y]$ *be*

$$R_{XY} = [0,\infty)\times[1,5]$$

and its joint probability density function be

$$f_{XY}(x,y) = \begin{cases} \frac{1}{4} y \exp(-xy) & \text{if } x\in[0,\infty) \text{ and } y\in[1,5] \\ 0 & \text{otherwise} \end{cases}$$

Let us compute the conditional probability density function of X *given* $Y=1$. *The support of* Y *is*

$$R_Y = [1,5]$$

When $y\notin[1,5]$, *the marginal probability density function of* Y *is* 0; *when* $y\in[1,5]$, *the marginal probability density function is*

$$\begin{aligned}
f_Y(y) &= \int_{-\infty}^{\infty} f_{XY}(x,y)\,dx = \int_0^{\infty} \frac{1}{4} y \exp(-xy)\,dx \\
&= \frac{1}{4}\left[-\exp(-xy)\right]_0^{\infty} = \frac{1}{4}\left[0-(-1)\right] = \frac{1}{4}
\end{aligned}$$

Thus, the marginal probability density function of Y *is*

$$f_Y(y) = \begin{cases} 1/4 & \text{if } y\in[1,5] \\ 0 & \text{otherwise} \end{cases}$$

[14] See p. 134.

When evaluated at $y = 1$*, it is*

$$f_Y(1) = \frac{1}{4}$$

The support of X *is*

$$R_X = [0, \infty)$$

Thus, the conditional probability density function of X *given* $Y = 1$ *is*

$$f_{X|Y=1}(x) = \frac{f_{XY}(x,1)}{f_Y(1)} = \begin{cases} \exp(-x) & \text{if } x \in [0,\infty) \\ 0 & \text{otherwise} \end{cases}$$

26.3 Conditional distribution function

In general, when $[X \;\; Y]$ is neither discrete nor absolutely continuous, we can characterize the distribution function[15] of X conditional on the information that $Y = y$. We define the conditional distribution function of X given $Y = y$ as follows:

Definition 151 *We say that a function* $F_{X|Y=y} : \mathbb{R} \to [0,1]$ *is the* ***conditional distribution function*** *of* X *given* $Y = y$ *if*

$$F_{X|Y=y}(x) = \mathrm{P}(X \leq x \,|Y = y), \;\; \forall x \in \mathbb{R}$$

where $\mathrm{P}(X \leq x \,|Y = y)$ *is the conditional probability that* $X \leq x$ *given that* $Y = y$*.*

There is no immediate way of deriving the conditional distribution of X given $Y = y$. However, we can characterize it using the concept of conditional probability with respect to a partition[16], as follows.

Define the events G_y as follows:

$$G_y = \{\omega \in \Omega : Y = y\}$$

and a partition $\mathcal{G}$ of events as:

$$\mathcal{G} = \{G_y : y \in R_Y\}$$

where, as usual, R_Y is the support of Y.

Then, for any $\omega \in G_y$ we have:

$$F_{X|Y=y}(x) = \mathrm{P}(X \leq x \,|\mathcal{G})(\omega)$$

where $\mathrm{P}(X \leq x \,|\mathcal{G})$ is the probability that $X \leq x$ conditional on the partition $\mathcal{G}$. As we know, $\mathrm{P}(X \leq x \,|\mathcal{G})$ is guaranteed to exist and is unique up to almost sure equality. Of course, this does not mean that we are able to compute it. Nonetheless, this characterization is extremely useful, because it allows us to speak of the conditional distribution of X given $Y = y$ in general, without a need to specify whether X and Y are discrete or continuous.

[15] See p. 108.
[16] See p. 206.

26.4 More details

26.4.1 Conditional distribution of a random vector

We have discussed how to update the probability distribution of a random variable X after observing the realization of another random variable Y, i.e. after receiving the information that $Y = y$. What happens when X and Y are random vectors rather than random variables? Basically, everything we said above still applies with straightforward modifications.

Thus, if X and Y are discrete random vectors, then the conditional probability mass function of X given $Y = y$ is (provided $p_Y(y) \neq 0$):

$$p_{X|Y=y}(x) = \frac{p_{XY}(x,y)}{p_Y(y)}$$

If X and Y are absolutely continuous random vectors then the conditional probability density function of X given $Y = y$ is (provided $f_Y(y) \neq 0$):

$$f_{X|Y=y}(x) = \frac{f_{XY}(x,y)}{f_Y(y)}$$

In general, the conditional distribution function of X given $Y = y$ is

$$F_{X|Y=y}(x) = \mathrm{P}(X \leq x \,|Y = y)$$

26.4.2 Joint equals marginal times conditional

As we have explained above, the joint distribution of X and Y can be used to derive the marginal distribution of Y and the conditional distribution of X given $Y = y$. This process can also go in the reverse direction: if we know the marginal distribution of Y and the conditional distribution of X given $Y = y$, then we can derive the joint distribution of X and Y. For discrete random variables:

$$p_{XY}(x,y) = p_{X|Y=y}(x)\, p_Y(y)$$

For absolutely continuous random variables:

$$f_{XY}(x,y) = f_{X|Y=y}(x)\, f_Y(y)$$

26.5 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Let $[X \;\; Y]$ be a discrete random vector with support:

$$R_{XY} = \{[1 \;\; 0], [2 \;\; 0], [1 \;\; 1], [1 \;\; 2]\}$$

and joint probability mass function:

$$p_{XY}(x,y) = \begin{cases} 1/4 & \text{if } x = 1 \text{ and } y = 0 \\ 1/4 & \text{if } x = 2 \text{ and } y = 0 \\ 1/4 & \text{if } x = 1 \text{ and } y = 1 \\ 1/4 & \text{if } x = 1 \text{ and } y = 2 \\ 0 & \text{otherwise} \end{cases}$$

Compute the conditional probability mass function of X given $Y = 0$.

Solution

The marginal probability mass function of Y evaluated at $y = 0$ is

$$\begin{aligned} p_Y(0) &= \sum_{\{(x,y)\in R_{XY}:y=0\}} p_{XY}(x,y) \\ &= p_{XY}(1,0) + p_{XY}(2,0) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \end{aligned}$$

The support of X is:

$$R_X = \{1, 2\}$$

Thus, the conditional probability mass function of X given $Y = 0$ is

$$p_{X|Y=0}(x) = \begin{cases} \frac{p_{XY}(1,0)}{p_Y(0)} = \frac{1/4}{1/2} = \frac{1}{2} & \text{if } x = 1 \\ \frac{p_{XY}(2,0)}{p_Y(0)} = \frac{1/4}{1/2} = \frac{1}{2} & \text{if } x = 2 \\ \frac{p_{XY}(x,0)}{p_Y(0)} = \frac{0}{1/2} = 0 & \text{if } x \notin R_X \end{cases}$$

Exercise 2

Let $[X \;\; Y]$ be an absolutely continuous random vector with support:

$$R_{XY} = [0, \infty) \times [1, 2]$$

and let its joint probability density function be:

$$f_{XY}(x,y) = \begin{cases} \frac{2}{3}y^2 \exp(-xy) & \text{if } x \in [0,\infty) \text{ and } y \in [1,2] \\ 0 & \text{otherwise} \end{cases}$$

Compute the conditional probability density function of X given $Y = 2$.

Solution

The support of Y is:

$$R_Y = [1, 2]$$

When $y \notin [1, 2]$, the marginal probability density function of Y is $f_Y(y) = 0$; when $y \in [1, 2]$, the marginal probability density function of Y is

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{XY}(x,y)\,dx = \int_0^{\infty} \frac{2}{3}y^2 \exp(-xy)\,dx \\ &= \left[-\frac{2}{3}y \exp(-xy)\right]_0^{\infty} = \frac{2}{3}y \end{aligned}$$

Thus, the marginal probability density function of Y is

$$f_Y(y) = \begin{cases} \frac{2}{3}y & \text{if } y \in [1,2] \\ 0 & \text{otherwise} \end{cases}$$

When evaluated at the point $y = 2$, it becomes

$$f_Y(2) = \frac{4}{3}$$

The support of X is

$$R_X = [0, \infty)$$

Thus, the conditional probability density function of X given $Y = 2$ is

$$\begin{aligned} f_{X|Y=2}(x) &= \frac{f_{XY}(x,2)}{f_Y(2)} \\ &= \begin{cases} \frac{(2/3)\cdot 2^2 \exp(-2x)}{4/3} = 2\exp(-2x) & \text{if } x \in [0,\infty) \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Exercise 3

Let X be an absolutely continuous random variable with support

$$R_X = [0, 1]$$

and probability density function

$$f_X(x) = \begin{cases} 1 & \text{if } x \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

Let Y be another absolutely continuous random variable with support

$$R_Y = [0, 1]$$

and conditional probability density function

$$f_{Y|X=x}(y) = \begin{cases} (1+2xy)/(1+x) & \text{if } y \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

Find the marginal probability density function function of Y.

Solution

The support of the vector $[X \;\; Y]$ is

$$R_{XY} = [0,1] \times [0,1]$$

and the joint probability function of X and Y is

$$\begin{aligned} f_{XY}(x,y) &= f_{Y|X=x}(y) f_X(x) \\ &= \begin{cases} (1+2xy)/(1+x) & \text{if } x \in [0,1]\times[0,1] \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

The marginal probability density function of Y is obtained by marginalization, integrating x out of the joint probability density function:

$$f_Y(y) = \int_0^1 f_{XY}(x,y)\,dx$$

Thus, for $y \notin [0,1]$ we trivially have $f_Y(y) = 0$ (because $f_{XY}(x,y) = 0$), while for $y \in [0,1]$ we have:

$$f_Y(y) = \int_0^1 f_{XY}(x,y)\,dx$$

$$
\begin{aligned}
& = \int_0^1 \frac{1+2xy}{1+x}dx \\
\boxed{\text{A}} \quad & = \int_0^1 \frac{1+x-x+2yx}{1+x}dx \\
& = \int_0^1 \left(1+\frac{(2y-1)\,x}{1+x}\right)dx \\
\boxed{\text{B}} \quad & = \int_0^1 dx + (2y-1)\int_0^1 \frac{x}{1+x}dx \\
\boxed{\text{C}} \quad & = [x]_0^1 + (2y-1)\int_0^1 \frac{1+x-1}{1+x}dx \\
& = 1 + (2y-1)\int_0^1 \left(1-\frac{1}{1+x}\right)dx \\
\boxed{\text{D}} \quad & = 1 + (2y-1)\left[\int_0^1 dx - \int_0^1 \frac{1}{1+x}dx\right] \\
& = 1 + (2y-1)\left[[x]_0^1 - [\ln(1+x)]_0^1\right] \\
& = 1 + (2y-1)\left[1-\ln(2)\right] \\
& = 1 + 2\left[1-\ln(2)\right]y - 1 + \ln(2) \\
& = \ln(2) + 2\left[1-\ln(2)\right]y
\end{aligned}
$$

where: in step A we have added and subtracted x from the numerator; in step B we have used the linearity of the integral; in step C we have added and subtracted 1 from the numerator; in step D we have used the linearity of the integral. Thus, the marginal probability density function of Y is

$$
f_Y(y) = \begin{cases} \ln(2) + 2\left[1-\ln(2)\right]y & \text{if } y \in [0,1] \\ 0 & \text{otherwise} \end{cases}
$$

Chapter 27

Conditional expectation

The conditional expectation (or conditional mean, or conditional expected value) of a random variable is the expected value of the random variable itself, computed with respect to its conditional probability distribution[1].

As in the case of the expected value, giving a completely rigorous definition of conditional expected value requires a complicated mathematical apparatus. To make things simpler, we do not give a completely rigorous definition in this lecture. We rather give an informal definition and we show how conditional expectation can be computed. In particular, we discuss how to compute the expected value of a random variable X when we observe the realization of another random variable Y, that is, when we receive the information that $Y = y$. The expected value of X conditional on the information that $Y = y$ is called conditional expectation of X given $Y = y$.

27.1 Definition

The following informal definition is very similar to the definition of expected value we have given in the lecture entitled *Expected value* (p. 127).

Definition 152 (informal) *Let X and Y be two random variables. The **conditional expectation** of X given $Y = y$ is the weighted average of the values that X can take on, where each possible value is weighted by its respective conditional probability (conditional on the information that $Y = y$).*

The expectation of a random variable X conditional on $Y = y$ is denoted by

$$\mathrm{E}\left[X \,|Y = y\right]$$

27.2 Discrete random variables

In the case in which X and Y are two discrete random variables and, considered together, they form a discrete random vector[2], the formula for computing the conditional expectation of X given $Y = y$ is a straightforward implementation of

[1] See p. 209.
[2] See p. 116.

Definition 152: the weights of the average are given by the conditional probability mass function[3] of X.

Definition 153 *Let X and Y be two discrete random variables. Let R_X be the support of X and let $p_{X|Y=y}(x)$ be the conditional probability mass function of X given $Y=y$. The conditional expectation of X given $Y=y$ is*

$$\mathrm{E}[X|Y=y] = \sum_{x \in R_X} x p_{X|Y=y}(x)$$

provided that

$$\sum_{x \in R_X} |x| \, p_{X|Y=y}(x) < \infty$$

If you do not understand the symbol $\sum_{x \in R_X}$ and the finiteness condition above (absolute summability) go back to the lecture entitled *Expected value* (p. 127), where they are explained.

Example 154 *Let the support of the random vector $[X \;\; Y]$ be*

$$R_{XY} = \{[1 \;\; 3], [2 \;\; 0], [0 \;\; 0]\}$$

and its joint probability mass function be

$$p_{XY}(x,y) = \begin{cases} 1/3 & \text{if } x=1 \text{ and } y=3 \\ 1/3 & \text{if } x=2 \text{ and } y=0 \\ 1/3 & \text{if } x=0 \text{ and } y=0 \\ 0 & \text{otherwise} \end{cases}$$

Let us compute the conditional probability mass function of X given $Y=0$. The marginal probability mass function of Y evaluated at $y=0$ is

$$p_Y(0) = \sum_{\{(x,y) \in R_{XY} : y=0\}} p_{XY}(x,y) = p_{XY}(2,0) + p_{XY}(0,0) = \frac{2}{3}$$

The support of X is

$$R_X = \{0,1,2\}$$

Thus, the conditional probability mass function of X given $Y=0$ is

$$p_{X|Y=0}(x) = \begin{cases} \frac{p_{XY}(0,0)}{p_Y(0)} = \frac{1/3}{2/3} = 1/2 & \text{if } x=0 \\ \frac{p_{XY}(1,0)}{p_Y(0)} = \frac{0}{2/3} = 0 & \text{if } x=1 \\ \frac{p_{XY}(2,0)}{p_Y(0)} = \frac{1/3}{2/3} = 1/2 & \text{if } x=2 \\ 0 & \text{if } x \notin R_X \end{cases}$$

The conditional expectation of X given $Y=0$ is

$$\begin{aligned} \mathrm{E}[X|Y=0] &= 0 \cdot p_{X|Y=0}(0) + 1 \cdot p_{X|Y=0}(1) + 2 \cdot p_{X|Y=0}(2) \\ &= 0 \cdot \frac{1}{2} + 1 \cdot 0 + 2 \cdot \frac{1}{2} = 1 \end{aligned}$$

[3] See p. 210.

27.3 Absolutely continuous random variables

When X and Y are absolutely continuous random variables, forming an absolutely continuous random vector[4], the formula for computing the conditional expectation of X given $Y = y$ involves an integral, which can be thought of as the limiting case of the summation $\sum_{x \in R_X} x p_{X|Y=y}(x)$ found in the discrete case above.

Definition 155 *Let X and Y be two absolutely continuous random variables. Let R_X be the support of X and let $f_{X|Y=y}(x)$ be the conditional probability density function[5] of X given $Y = y$. The conditional expectation of X given $Y = y$ is*

$$\mathrm{E}[X|Y=y] = \int_{-\infty}^{\infty} x f_{X|Y=y}(x)\,dx$$

provided that

$$\int_{-\infty}^{\infty} |x| f_{X|Y=y}(x)\,dx < \infty$$

If you do not understand why an integration is required and why the finiteness condition above (absolute integrability) is imposed, you can find an explanation in the lecture entitled *Expected value* (p. 127).

Example 156 *Let the support of the random vector $[X \;\; Y]$ be*

$$R_{XY} = [0,\infty) \times [2,4]$$

and its joint probability density function be

$$f_{XY}(x,y) = \begin{cases} \frac{1}{2} y \exp(-xy) & \text{if } x \in [0,\infty) \text{ and } y \in [2,4] \\ 0 & \text{otherwise} \end{cases}$$

Let us compute the conditional probability density function of X given $Y = 2$. The support of Y is

$$R_Y = [2,4]$$

When $y \notin [2,4]$, the marginal probability density function of Y is 0; when $y \in [2,4]$, the marginal probability density function is

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{XY}(x,y)\,dx = \int_0^{\infty} \frac{1}{2} y \exp(-xy)\,dx \\ &= \frac{1}{2}\left[-\exp(-xy)\right]_0^{\infty} = \frac{1}{2}[0-(-1)] = \frac{1}{2} \end{aligned}$$

Thus, the marginal probability density function of Y is

$$f_Y(y) = \begin{cases} 1/2 & \text{if } y \in [2,4] \\ 0 & \text{otherwise} \end{cases}$$

When evaluated at $y = 2$, it is

$$f_Y(2) = \frac{1}{2}$$

[4] See p. 117.
[5] See p. 213.

The support of X is

$$R_X = [0, \infty)$$

Thus, the conditional probability density function of X given $Y = 2$ is

$$f_{X|Y=2}(x) = \frac{f_{XY}(x, 2)}{f_Y(2)} = \begin{cases} 2\exp(-2x) & \text{if } x \in [0, \infty) \\ 0 & \text{otherwise} \end{cases}$$

The conditional expectation of X given $Y = 2$ is

$$\begin{aligned} & \mathrm{E}[X|Y=2] \\ = & \int_{-\infty}^{\infty} x f_{X|Y=2}(x)\, dx \\ = & \int_{0}^{\infty} x 2\exp(-2x)\, dx \\ \boxed{A} = & \frac{1}{2}\int_{0}^{\infty} t\exp(-t)\, dt \\ \boxed{B} = & \frac{1}{2}\left\{[-t\exp(-t)]_0^{\infty} + \int_{0}^{\infty}\exp(-t)\, dt\right\} \\ = & \frac{1}{2}\{0 - 0 + [-\exp(-t)]_0^{\infty}\} \\ = & \frac{1}{2}\{0 + 1\} = \frac{1}{2} \end{aligned}$$

where: in step $\boxed{A}$ we have performed a change of variable ($t = 2x$); in step $\boxed{B}$ we have performed an integration by parts.

27.4 Conditional expectation in general

The general formula for computing the conditional expectation of X given $Y = y$ does not require that the two variables form a discrete or an absolutely continuous random vector, but it is applicable to any random vector.

Definition 157 *Let $F_{X|Y=y}(x)$ be the conditional distribution function[6] of X given $Y = y$. The conditional expectation of X given $Y = y$ is*

$$\mathrm{E}[X|Y=y] = \int_{-\infty}^{\infty} x dF_{X|Y=y}(x)$$

where the integral is a Riemann-Stieltjes integral and the expected value exists and is well-defined only as long as the integral is well-defined.

The above formula follows the same logic of the formula for the expected value:

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x dF_X(x)$$

with the only difference that the unconditional distribution function $F_X(x)$ has now been replaced with the conditional distribution function $F_{X|Y=y}(x)$. The reader who feels unfamiliar with this formula can go back to the lecture entitled *Expected value* (p. 127) and read an intuitive introduction to the Riemann-Stieltjes integral and its use in probability theory.

[6] See p. 215.

27.5 More details

27.5.1 Properties of conditional expectation

From the above sections, it should be clear that the conditional expectation is computed exactly as the expected value, with the only difference that probabilities and probability densities are replaced by conditional probabilities and conditional probability densities. Therefore, the properties enjoyed by the expected value, such as linearity, are also enjoyed by the conditional expectation. For an exposition of the properties of the expected value, you can go to the lecture entitled *Properties of the expected value* (p. 147).

27.5.2 Law of iterated expectations

Quite obviously, before knowing the realization of Y, the conditional expectation of X given Y is unknown and can itself be regarded as a random variable. We denote it by $\mathrm{E}[X|Y]$. In other words, $\mathrm{E}[X|Y]$ is a random variable such that its realization equals $\mathrm{E}[X|Y=y]$ when y is the realization of Y.

This random variable satisfies a very important property, known as **law of iterated expectations**:

$$\mathrm{E}[\mathrm{E}[X|Y]] = \mathrm{E}[X]$$

Proof. For discrete random variables this is proved as follows:

$$\begin{aligned}
& \mathrm{E}[\mathrm{E}[X|Y]] \\
\boxed{\text{A}} \quad &= \sum_{y \in R_Y} \mathrm{E}[X|Y=y]\, p_Y(y) \\
\boxed{\text{B}} \quad &= \sum_{y \in R_Y} \sum_{x \in R_X} x p_{X|Y=y}(x)\, p_Y(y) \\
\boxed{\text{C}} \quad &= \sum_{y \in R_Y} \sum_{x \in R_X} x p_{XY}(x,y) \\
&= \sum_{x \in R_X} x \sum_{y \in R_Y} p_{XY}(x,y) \\
\boxed{\text{D}} \quad &= \sum_{x \in R_X} x p_X(x) \\
\boxed{\text{E}} \quad &= \mathrm{E}[X]
\end{aligned}$$

where: in step A we have used the definition of expected value; in step B we have used the definition of conditional expectation; in step C we have used the fact that the joint probability mass function $p_{XY}(x,y)$ is equal to the product of the marginal and conditional probability mass functions; in step D we have performed a marginalization of the joint probability mass function[7]; in step E we have used the definition of expected value. For absolutely continuous random variables the proof is analogous:

$$\mathrm{E}[\mathrm{E}[X|Y]]$$

[7] See p. 120.

$$\begin{aligned}
\boxed{\text{A}} \quad &= \int_{-\infty}^{\infty} \mathrm{E}\left[X \,|Y = y\right] f_Y(y)\, dy \\
\boxed{\text{B}} \quad &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X|Y=y}(x)\, dx f_Y(y)\, dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X|Y=y}(x)\, f_Y(y)\, dx dy \\
\boxed{\text{C}} \quad &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{XY}(x,y)\, dx dy \\
&= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{XY}(x,y)\, dy dx \\
\boxed{\text{D}} \quad &= \int_{-\infty}^{\infty} x f_X(x)\, dx \\
\boxed{\text{E}} \quad &= \mathrm{E}\left[X\right]
\end{aligned}$$

where: in step A we have used the definition of expected value; in step B we have used the definition of conditional expectation; in step C we have used the fact that the joint probability density function $f_{XY}(x,y)$ is equal to the product of the marginal and conditional probability density functions; in step D we have performed a marginalization of the joint probability density function[8]; in step E we have used the definition of expected value. ■

27.6 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Let $\begin{bmatrix} X & Y \end{bmatrix}$ be a random vector with support

$$R_{XY} = \left\{\begin{bmatrix} 2 & 2 \end{bmatrix}, \begin{bmatrix} 2 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 2 \end{bmatrix}, \begin{bmatrix} 0 & 2 \end{bmatrix}\right\}$$

and joint probability mass function

$$p_{XY}(x,y) = \begin{cases} 1/4 & \text{if } x = 2 \text{ and } y = 2 \\ 1/4 & \text{if } x = 2 \text{ and } y = 0 \\ 1/4 & \text{if } x = 1 \text{ and } y = 2 \\ 1/4 & \text{if } x = 0 \text{ and } y = 2 \\ 0 & \text{otherwise} \end{cases}$$

What is the conditional expectation of X given $Y = 2$?

Solution

Let us compute the conditional probability mass function of X given $Y = 2$. The marginal probability mass function of Y evaluated at $y = 2$ is

$$p_Y(2) = \sum_{\{(x,y) \in R_{XY} : y = 2\}} p_{XY}(x,y) = p_{XY}(2,2) + p_{XY}(1,2) + p_{XY}(0,2) = \frac{3}{4}$$

[8] See p. 120.

The support of X is

$$R_X = \{0, 1, 2\}$$

Thus, the conditional probability mass function of X given $Y = 2$ is

$$p_{X|Y=2}(x) = \begin{cases} \frac{p_{XY}(0,2)}{p_Y(2)} = \frac{1/4}{3/4} = 1/3 & \text{if } x = 0 \\ \frac{p_{XY}(1,2)}{p_Y(2)} = \frac{1/4}{3/4} = 1/3 & \text{if } x = 1 \\ \frac{p_{XY}(2,2)}{p_Y(2)} = \frac{1/4}{3/4} = 1/3 & \text{if } x = 2 \\ 0 & \text{if } x \notin R_X \end{cases}$$

The conditional expectation of X given $Y = 2$ is

$$\begin{aligned} \mathrm{E}[X|Y=2] &= 0 \cdot p_{X|Y=2}(0) + 1 \cdot p_{X|Y=2}(1) + 2 \cdot p_{X|Y=2}(2) \\ &= 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} + 2 \cdot \frac{1}{3} = 1 \end{aligned}$$

Exercise 2

Let X and Y be two random variables. Remember that the variance of X can be computed as

$$\mathrm{Var}[X] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 \tag{27.1}$$

In a similar manner, the conditional variance of X, given $Y = y$, can be defined as

$$\mathrm{Var}[X|Y=y] = \mathrm{E}\left[X^2|Y=y\right] - \mathrm{E}[X|Y=y]^2 \tag{27.2}$$

Use the law of iterated expectations to prove that

$$\mathrm{Var}[X] = \mathrm{E}[\mathrm{Var}[X|Y=y]] + \mathrm{Var}[\mathrm{E}[X|Y=y]]$$

Solution

This is proved as follows:

$$\begin{aligned} & \mathrm{Var}[X] \\ = \; & \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 \\ \boxed{\text{A}} \; = \; & \mathrm{E}\left[\mathrm{E}\left[X^2|Y=y\right]\right] - \mathrm{E}[\mathrm{E}[X|Y=y]]^2 \\ \boxed{\text{B}} \; = \; & \mathrm{E}\left[\mathrm{Var}[X|Y=y] + \mathrm{E}[X|Y=y]^2\right] - \mathrm{E}[\mathrm{E}[X|Y=y]]^2 \\ \boxed{\text{C}} \; = \; & \mathrm{E}[\mathrm{Var}[X|Y=y]] + \mathrm{E}\left[\mathrm{E}[X|Y=y]^2\right] - \mathrm{E}[\mathrm{E}[X|Y=y]]^2 \\ \boxed{\text{D}} \; = \; & \mathrm{E}[\mathrm{Var}[X|Y=y]] + \mathrm{Var}[\mathrm{E}[X|Y=y]] \end{aligned}$$

where: in step A we have used the law of iterated expectations; in step B we have used equation (27.2); in step C we have used the linearity of the expected value; in step D we have used equation (27.1).

Chapter 28

Independent random variables

Two random variables are independent if they convey no information about each other and, as a consequence, receiving information about one of the two does not change our assessment of the probability distribution of the other.

This lecture provides a formal definition of independence and discusses how to verify whether two or more random variables are independent.

28.1 Definition

Recall (see the lecture entitled *Independent events* - p. 99) that two events A and B are independent if and only if

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

This definition is extended to random variables as follows.

Definition 158 *Two random variables X and Y are said to be **independent** if and only if*

$$\mathrm{P}(\{X \in A\} \cap \{Y \in B\}) = \mathrm{P}(\{X \in A\})\,\mathrm{P}(\{Y \in B\})$$

for any couple of events $\{X \in A\}$ and $\{Y \in B\}$, where $A \subseteq \mathbb{R}$ and $B \subseteq \mathbb{R}$.

In other words, two random variables are independent if and only if the events related to those random variables are independent events.

The independence between two random variables is also called statistical independence.

28.2 Independence criterion

Checking the independence of all possible couples of events related to two random variables can be very difficult. This is the reason why the above definition is seldom used to verify whether two random variables are independent. The following criterion is more often used instead.

Proposition 159 *Two random variables X and Y are independent if and only if*

$$F_{XY}(x,y) = F_X(x) F_Y(y) \qquad \forall x, y \in \mathbb{R}$$

where $F_{X,Y}(x,y)$ is their joint distribution function[1] *and $F_X(x)$ and $F_Y(y)$ are their marginal distribution functions*[2]*.*

Proof. Using some facts from measure theory (not proved here), it is possible to demonstrate that, when checking for the condition

$$\mathrm{P}(\{X \in A\} \cap \{Y \in B\}) = \mathrm{P}(\{X \in A\}) \mathrm{P}(\{Y \in B\})$$

it is sufficient to confine attention to sets A and B taking the form

$$\begin{aligned} A &= (-\infty, x] \\ B &= (-\infty, y] \end{aligned}$$

Thus, two random variables are independent if and only if

$$\mathrm{P}(\{X \in (-\infty,x]\} \cap \{Y \in (-\infty,y]\}) = \mathrm{P}(\{X \in (-\infty,x]\}) \mathrm{P}(\{Y \in (-\infty,y]\})$$

for any $x, y \in \mathbb{R}$. Using the definitions of joint and marginal distribution function, this condition can be written as

$$F_{XY}(x,y) = F_X(x) F_Y(y) \qquad \forall x, y \in \mathbb{R}$$

■

Example 160 *Let X and Y be two random variables with marginal distribution functions*

$$\begin{aligned} F_X(x) &= \begin{cases} 0 & \text{if } x < 0 \\ 1 - \exp(-x) & \text{if } x \geq 0 \end{cases} \\ F_Y(y) &= \begin{cases} 0 & \text{if } y < 0 \\ 1 - \exp(-y) & \text{if } y \geq 0 \end{cases} \end{aligned}$$

and joint distribution function

$$F_{X,Y}(x,y) = \begin{cases} 0 & \text{if } x < 0 \text{ or } y < 0 \\ 1 - \exp(-x) - \exp(-y) + \exp(-x-y) & \text{if } x \geq 0 \text{ and } y \geq 0 \end{cases}$$

X and Y are independent if and only if

$$F_{XY}(x,y) = F_X(x) F_Y(y)$$

which is straightforward to verify. When $x < 0$ or $y < 0$, then

$$F_X(x) F_Y(y) = 0 = F_{X,Y}(x,y)$$

When $x \geq 0$ and $y \geq 0$, then:

$$\begin{aligned} F_X(x) F_Y(y) &= [1 - \exp(-x)][1 - \exp(-y)] \\ &= 1 - \exp(-x) - \exp(-y) + \exp(-x)\exp(-y) \\ &= 1 - \exp(-x) - \exp(-y) + \exp(-x-y) \\ &= F_{X,Y}(x,y) \end{aligned}$$

[1] See p. 118.
[2] See p. 119.

28.3 Independence between discrete variables

When the two variables, taken together, form a discrete random vector, independence can also be verified using the following proposition:

Proposition 161 *Two random variables X and Y, forming a discrete random vector, are independent if and only if*

$$p_{XY}(x,y) = p_X(x)\, p_Y(y) \qquad \forall x, y \in \mathbb{R}$$

where $p_{X,Y}(x,y)$ is their joint probability mass function[3] and $p_X(x)$ and $p_Y(y)$ are their marginal probability mass functions[4].

The following example illustrates how this criterion can be used.

Example 162 *Let $[X \;\; Y]$ be a discrete random vector with support*

$$R_{XY} = \{[1 \;\; 1], [2 \;\; 0], [0 \;\; 0]\}$$

Let its joint probability mass function be

$$p_{X,Y}(x,y) = \begin{cases} 1/3 & \text{if } [x \;\; y] = [1 \;\; 1] \\ 1/3 & \text{if } [x \;\; y] = [2 \;\; 0] \\ 1/3 & \text{if } [x \;\; y] = [0 \;\; 0] \\ 0 & \text{otherwise} \end{cases}$$

In order to verify whether X and Y are independent, we first need to derive the marginal probability mass functions of X and Y. The support of X is

$$R_X = \{0, 1, 2\}$$

and the support of Y is

$$R_Y = \{0, 1\}$$

We need to compute the probability of each element of the support of X:

$$\begin{aligned} p_X(0) &= \sum_{y \in R_Y} p_{XY}(0,y) = p_{XY}(0,0) + p_{XY}(0,1) = \frac{1}{3} + 0 = \frac{1}{3} \\ p_X(1) &= \sum_{y \in R_Y} p_{XY}(1,y) = p_{XY}(1,0) + p_{XY}(1,1) = 0 + \frac{1}{3} = \frac{1}{3} \\ p_X(2) &= \sum_{y \in R_Y} p_{XY}(2,y) = p_{XY}(2,0) + p_{XY}(2,1) = \frac{1}{3} + 0 = \frac{1}{3} \end{aligned}$$

Thus, the marginal probability mass function of X is

$$p_X(x) = \sum_{y \in R_Y} p_{XY}(x,y) = \begin{cases} 1/3 & \text{if } x = 0 \\ 1/3 & \text{if } x = 1 \\ 1/3 & \text{if } x = 2 \\ 0 & \text{otherwise} \end{cases}$$

[3] See p. 116.
[4] See p. 119.

We need to compute the probability of each element of the support of Y:

$$\begin{aligned} p_Y(0) &= \sum_{x \in R_X} p_{XY}(x,0) = p_{XY}(0,0) + p_{XY}(1,0) + p_{XY}(2,0) \\ &= \frac{1}{3} + 0 + \frac{1}{3} = \frac{2}{3} \\ p_Y(1) &= \sum_{x \in R_X} p_{XY}(x,1) = p_{XY}(0,1) + p_{XY}(1,1) + p_{XY}(2,1) \\ &= 0 + \frac{1}{3} + 0 = \frac{1}{3} \end{aligned}$$

Thus, the marginal probability mass function of Y is

$$p_Y(y) = \sum_{y \in R_Y} p_{XY}(x,y) = \begin{cases} 2/3 & \text{if } y = 0 \\ 1/3 & \text{if } y = 1 \\ 0 & \text{otherwise} \end{cases}$$

The product of the marginal probability mass functions is

$$p_X(x)\, p_Y(y) = \begin{cases} (1/3)\cdot(2/3) = 2/9 & \text{if } [x \;\; y] = [0 \;\; 0] \\ (1/3)\cdot(2/3) = 2/9 & \text{if } [x \;\; y] = [1 \;\; 0] \\ (1/3)\cdot(2/3) = 2/9 & \text{if } [x \;\; y] = [2 \;\; 0] \\ (1/3)\cdot(1/3) = 1/9 & \text{if } [x \;\; y] = [0 \;\; 1] \\ (1/3)\cdot(1/3) = 1/9 & \text{if } [x \;\; y] = [1 \;\; 1] \\ (1/3)\cdot(1/3) = 1/9 & \text{if } [x \;\; y] = [2 \;\; 1] \\ 0 & \text{otherwise} \end{cases}$$

which is obviously different from $p_{XY}(x,y)$. Therefore, X and Y are not independent.

28.4 Independence between continuous variables

When the two variables, taken together, form an absolutely continuous random vector, independence can also be verified using the following proposition:

Proposition 163 *Two random variables X and Y, forming an absolutely continuous random vector, are independent if and only if*

$$f_{XY}(x,y) = f_X(x)\, f_Y(y) \qquad \forall x, y \in \mathbb{R}$$

where $f_{X,Y}(x,y)$ is their joint probability density function[5] *and $f_X(x)$ and $f_Y(y)$ are their marginal probability density functions*[6]*.*

The following example illustrates how this criterion can be used.

Example 164 *Let the joint probability density function of X and Y be*

$$f_{X,Y}(x,y) = \begin{cases} 0 & \text{if } x \notin [0,1] \text{ or } y \notin [0,1] \\ 1 & \text{if } x \in [0,1] \text{ and } y \in [0,1] \end{cases}$$

[5] See p. 117.
[6] See p. 119.

Its marginals are

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy \\ &= \int_{-\infty}^{0} f_{X,Y}(x,y)\,dy + \int_{0}^{1} f_{X,Y}(x,y)\,dy + \int_{1}^{\infty} f_{X,Y}(x,y)\,dy \\ &= 0 + \int_{0}^{1} f_{X,Y}(x,y)\,dy + 0 \\ &= \begin{cases} 0 & \text{if } x \notin [0,1] \\ 1 & \text{if } x \in [0,1] \end{cases} \end{aligned}$$

and

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx \\ &= \int_{-\infty}^{0} f_{X,Y}(x,y)\,dx + \int_{0}^{1} f_{X,Y}(x,y)\,dx + \int_{1}^{\infty} f_{X,Y}(x,y)\,dx \\ &= 0 + \int_{0}^{1} f_{X,Y}(x,y)\,dx + 0 \\ &= \begin{cases} 0 & \text{if } y \notin [0,1] \\ 1 & \text{if } y \in [0,1] \end{cases} \end{aligned}$$

Verifying that

$$f_{XY}(x,y) = f_X(x)\, f_Y(y)$$

is straightforward. When $x \notin [0,1]$ or $y \notin [0,1]$, then

$$f_X(x)\, f_Y(y) = 0 = f_{X,Y}(x,y)$$

When $x \in [0,1]$ and $y \in [0,1]$, then:

$$f_X(x)\, f_Y(y) = 1 \cdot 1 = 1 = f_{X,Y}(x,y)$$

28.5 More details

The following subsections contain more details about statistical independence.

28.5.1 Mutually independent random variables

The definition of mutually independent random variables extends the definition of mutually independent events to random variables.

Definition 165 *We say that n random variables $X_1, \ldots, X_n$ are **mutually independent** (or jointly independent) if and only if*

$$\mathrm{P}\left(\bigcap_{j=1}^{k} \{X_{i_j} \in A_j\}\right) = \prod_{j=1}^{k} \mathrm{P}\left(\{X_{i_j} \in A_j\}\right)$$

for any sub-collection of k random variables $X_{i_1}, \ldots, X_{i_k}$ (where $k \leq n$) and for any collection of events $\{X_{i_1} \in A_1\}, \ldots, \{X_{i_k} \in A_k\}$, where $A_1, \ldots, A_k \subseteq \mathbb{R}$.

In other words, n random variables are mutually independent if the events related to those random variables are mutually independent events[7].

Denote by X a random vector whose components are $X_1, \ldots, X_n$. The above condition for mutual independence can be replaced:

1. in general, by a condition on the joint distribution function of X:
$$F_X(x_1, \ldots, x_n) = \prod_{j=1}^{n} F_{X_j}(x_j)$$

2. for discrete random variables, by a condition on the joint probability mass function of X:
$$p_X(x_1, \ldots, x_n) = \prod_{j=1}^{n} p_{X_j}(x_j)$$

3. for absolutely continuous random variables, by a condition on the joint probability density function of X:
$$f_X(x_1, \ldots, x_n) = \prod_{j=1}^{n} f_{X_j}(x_j)$$

28.5.2 Mutual independence via expectations

It can be proved that n random variables $X_1, \ldots, X_n$ are mutually independent if and only if:
$$\mathrm{E}\left[\prod_{j=1}^{n} g_j(X_j)\right] = \prod_{j=1}^{n} \mathrm{E}\left[g_j(X_j)\right]$$
for any n functions $g_1, \ldots, g_n$ such that the above expected values exist and are well-defined.

28.5.3 Independence and zero covariance

If two random variables X_1 and X_2 are independent, then their covariance is zero:
$$\mathrm{Cov}[X_1, X_2] = 0$$

Proof. This is an immediate consequence of the fact that, if X_1 and X_2 are independent then:
$$\mathrm{E}\left[g_1(X_1)\, g_2(X_2)\right] = \mathrm{E}\left[g_1(X_1)\right] \mathrm{E}\left[g_2(X_2)\right]$$
(see the *Mutual independence via expectations* property above). When g_1 and g_2 are identity functions ($g_1(X_1) = X_1$ and $g_2(X_2) = X_2$), then:
$$\mathrm{E}[X_1 X_2] = \mathrm{E}[X_1] \mathrm{E}[X_2]$$
Therefore, by the covariance formula[8]:
$$\mathrm{Cov}[X_1, X_2] = \mathrm{E}[X_1 X_2] - \mathrm{E}[X_1] \mathrm{E}[X_2]$$

[7] See p. 100.
[8] See p. 164.

$$\begin{aligned} &= \mathrm{E}[X_1]\,\mathrm{E}[X_2] - \mathrm{E}[X_1]\,\mathrm{E}[X_2] \\ &= 0 \end{aligned}$$

■

The converse is not true: two random variables that have zero covariance are not necessarily independent.

28.5.4 Independent random vectors

The above notions are easily generalized to the case in which X and Y are two random vectors, having dimensions $K_X \times 1$ and $K_Y \times 1$ respectively. Denote their joint distribution functions by $F_X(X)$ and $F_Y(Y)$ and the joint distribution function of X and Y together by

$$F_{XY}(X,Y) = F_{X_1,X_2,\ldots,X_{K_X},Y_1,Y_2,\ldots,Y_{K_Y}}(x_1,x_2,\ldots,x_{K_X},y_1,y_2,\ldots,y_{K_Y})$$

Also, if the two vectors are discrete or absolutely continuous replace F with p or f to denote the corresponding probability mass or density functions.

Definition 166 *Two random vectors X and Y are independent if and only if one of the following equivalent conditions is satisfied:*

1. *Condition 1:*

$$\mathrm{P}(\{X \in A\} \cap \{Y \in B\}) = \mathrm{P}(\{X \in A\})\,\mathrm{P}(\{Y \in B\})$$

for any couple of events $\{X \in A\}$ and $\{Y \in B\}$, where $A \subseteq \mathbb{R}^{K_X}$ and $B \subseteq \mathbb{R}^{K_Y}$.

2. *Condition 2:*

$$F_{XY}(x,y) = F_X(x)\,F_Y(y)$$

for any $x \in \mathbb{R}^{K_X}$ and $y \in \mathbb{R}^{K_Y}$ (replace F with p or f when the distributions are discrete or absolutely continuous).

3. *Condition 3:*

$$\mathrm{E}[g_1(X)\,g_2(Y)] = \mathrm{E}[g_1(X)]\,\mathrm{E}[g_2(Y)]$$

for any functions $g_1 : \mathbb{R}^{K_X} \to \mathbb{R}$ and $g_2 : \mathbb{R}^{K_Y} \to \mathbb{R}$ such that the above expected values exist and are well-defined.

28.5.5 Mutually independent random vectors

Also the definition of mutual independence extends in a straightforward manner to random vectors:

Definition 167 *We say that n random vectors $X_1, \ldots, X_n$ are **mutually independent** (or jointly independent) if and only if*

$$\mathrm{P}\left(\bigcap_{j=1}^{k}\{X_{i_j} \in A_j\}\right) = \prod_{j=1}^{k}\mathrm{P}\left(\{X_{i_j} \in A_j\}\right)$$

for any sub-collection of k random vectors $X_{i_1}, \ldots, X_{i_k}$ (where $k \leq n$) and for any collection of events $\{X_{i_1} \in A_1\}, \ldots, \{X_{i_k} \in A_k\}$.

All the equivalent conditions for the joint independence of a set of random variables (see above) apply with obvious modifications also to random vectors.

28.6 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Consider two random variables X and Y having marginal distribution functions

$$F_X(x) = \begin{cases} 0 & \text{if } x < 1 \\ 1/2 & \text{if } 1 \leq x < 2 \\ 1 & \text{if } x \geq 2 \end{cases}$$

$$F_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ 1 - \frac{1}{2}\exp(-y) - \frac{1}{2}\exp(-2y) & \text{if } y \geq 0 \end{cases}$$

If X and Y are independent, what is their joint distribution function?

Solution

For X and Y to be independent, their joint distribution function must be equal to the product of their marginal distribution functions:

$$F_{X,Y}(x,y) = \begin{cases} 0 & \text{if } x < 1 \text{ or } y < 0 \\ \frac{1}{2} - \frac{1}{4}\exp(-y) - \frac{1}{4}\exp(-2y) & \text{if } 1 \leq x < 2 \text{ and } y \geq 0 \\ 1 - \frac{1}{2}\exp(-y) - \frac{1}{2}\exp(-2y) & \text{if } x \geq 2 \text{ and } y \geq 0 \end{cases}$$

Exercise 2

Let $[X \;\; Y]$ be a discrete random vector with support

$$R_{XY} = \{[0 \;\; 0], [0 \;\; 1], [1 \;\; 0], [1 \;\; 1]\}$$

Let its joint probability mass function be

$$p_{X,Y}(x,y) = \begin{cases} 1/4 & \text{if } [x \;\; y] = [0 \;\; 0] \\ 1/4 & \text{if } [x \;\; y] = [0 \;\; 1] \\ 1/4 & \text{if } [x \;\; y] = [1 \;\; 0] \\ 1/4 & \text{if } [x \;\; y] = [1 \;\; 1] \\ 0 & \text{otherwise} \end{cases}$$

Are X and Y independent?

Solution

In order to verify whether X and Y are independent, we first need to derive the marginal probability mass functions of X and Y. The support of X is

$$R_X = \{0, 1\}$$

and the support of Y is

$$R_Y = \{0, 1\}$$

We need to compute the probability of each element of the support of X:

$$p_X(0) = \sum_{y \in R_Y} p_{XY}(0,y) = p_{XY}(0,0) + p_{XY}(0,1) = \frac{1}{2}$$

$$p_X(1) = \sum_{y \in R_Y} p_{XY}(1,y) = p_{XY}(1,0) + p_{XY}(1,1) = \frac{1}{2}$$

Thus, the marginal probability mass function of X is

$$p_X(x) = \sum_{y \in R_Y} p_{XY}(x,y) = \begin{cases} 1/2 & \text{if } x = 0 \\ 1/2 & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases}$$

We need to compute the probability of each element of the support of Y:

$$p_Y(0) = \sum_{x \in R_X} p_{XY}(x,0) = p_{XY}(0,0) + p_{XY}(1,0) = \frac{1}{2}$$

$$p_Y(1) = \sum_{x \in R_X} p_{XY}(x,1) = p_{XY}(0,1) + p_{XY}(1,1) = \frac{1}{2}$$

Thus, the marginal probability mass function of Y is

$$p_Y(y) = \sum_{x \in R_X} p_{XY}(x,y) = \begin{cases} 1/2 & \text{if } y = 0 \\ 1/2 & \text{if } y = 1 \\ 0 & \text{otherwise} \end{cases}$$

The product of the marginal probability mass functions is

$$p_X(x) p_Y(y) = \begin{cases} (1/2) \cdot (1/2) = 1/4 & \text{if } [x \;\; y] = [0 \;\; 0] \\ (1/2) \cdot (1/2) = 1/4 & \text{if } [x \;\; y] = [1 \;\; 0] \\ (1/2) \cdot (1/2) = 1/4 & \text{if } [x \;\; y] = [0 \;\; 1] \\ (1/2) \cdot (1/2) = 1/4 & \text{if } [x \;\; y] = [1 \;\; 1] \\ 0 & \text{otherwise} \end{cases}$$

which is equal to $p_{XY}(x,y)$. Therefore, X and Y are independent.

Exercise 3

Let $[X \;\; Y]$ be an absolutely continuous random vector with support

$$R_{XY} = [0, \infty) \times [2,3]$$

and let its joint probability density function be

$$f_{XY}(x,y) = \begin{cases} \exp(-x) & \text{if } x \in [0,\infty) \text{ and } y \in [2,3] \\ 0 & \text{otherwise} \end{cases}$$

Are X and Y independent?

Solution

The support of Y is

$$R_Y = [2,3]$$

When $y \notin [2,3]$, the marginal probability density function of Y is 0, while, when $y \in [2,3]$, the marginal probability density function of Y is

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)\,dx = \int_0^{\infty} \exp(-x)\,dx$$

$$= \quad [-\exp(-x)]_0^{\infty} = [0-(-1)] = 1$$

Summing up, the marginal probability density function of Y is

$$f_Y(y) = \begin{cases} 1 & \text{if } y \in [2,3] \\ 0 & \text{otherwise} \end{cases}$$

The support of X is

$$R_X = [0, \infty)$$

When $x \notin [0,\infty)$, the marginal probability density function of X is 0, while, when $x \in [0,\infty)$, the marginal probability density function of X is

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{XY}(x,y)\,dy = \int_2^3 \exp(-x)\,dy \\ &= \exp(-x)\int_2^3 dy = \exp(-x) \end{aligned}$$

The marginal probability density function of X is

$$f_X(x) = \begin{cases} \exp(-x) & \text{if } x \in [0,\infty) \\ 0 & \text{otherwise} \end{cases}$$

Verifying that

$$f_{XY}(x,y) = f_X(x)\, f_Y(y)$$

is straightforward. When $x \notin [0,\infty)$ or $y \notin [2,3]$, then

$$f_X(x)\, f_Y(y) = 0 = f_{X,Y}(x,y)$$

When $x \in [0,\infty)$ and $y \in [2,3]$, then:

$$f_X(x)\, f_Y(y) = \exp(-x)\cdot 1 = \exp(-x) = f_{X,Y}(x,y)$$

Therefore, X and Y are independent.

Part III

Additional topics in probability theory

Chapter 29

Probabilistic inequalities

This lecture introduces some probabilistic inequalities that are used in the proofs of several important theorems in probability theory.

29.1 Markov's inequality

Proposition 168 *Let X be an integrable*[1] *random variable defined on a sample space Ω. Let X be a positive random variable*[2]*. Let $c \in \mathbb{R}_{++}$. Then, the following inequality, called Markov's inequality, holds:*

$$\mathrm{P}(X \geq c) \leq \frac{\mathrm{E}[X]}{c} \tag{29.1}$$

Reading and understanding the proof of Markov's inequality is highly recommended, because it is an interesting application of many elementary properties of the expected value.

Proof. First note that

$$1_{\{X \geq c\}} + 1_{\{X < c\}} = 1$$

where $1_{\{X \geq c\}}$ is the indicator[3] of the event $\{X \geq c\}$, and $1_{\{X < c\}}$ is the indicator of the event $\{X < c\}$. As a consequence, we can write

$$\begin{aligned} \mathrm{E}[X] &= \mathrm{E}[X \cdot 1] \\ &= \mathrm{E}\left[X \cdot \left(1_{\{X \geq c\}} + 1_{\{X < c\}}\right)\right] \\ &= \mathrm{E}\left[X 1_{\{X \geq c\}}\right] + \mathrm{E}\left[X 1_{\{X < c\}}\right] \end{aligned}$$

Now, note that $X 1_{\{X < c\}}$ is a positive random variable and that the expected value of a positive random variable is positive[4]:

$$\mathrm{E}\left[X 1_{\{X < c\}}\right] \geq 0$$

Therefore,

$$\mathrm{E}[X] = \mathrm{E}\left[X 1_{\{X \geq c\}}\right] + \mathrm{E}\left[X 1_{\{X < c\}}\right] \geq \mathrm{E}\left[X 1_{\{X \geq c\}}\right]$$

[1] See p. 136.
[2] In other words, $X(\omega) \geq 0$ for all $\omega \in \Omega$.
[3] See p. 197.
[4] See p. 150.

The random variable $c \cdot 1_{\{X \geq c\}}$ is less than or equal to the random variable $X \cdot 1_{\{X \geq c\}}$ for any $\omega \in \Omega$:

$$c \cdot 1_{\{X \geq c\}} \leq X \cdot 1_{\{X \geq c\}}$$

because c is always smaller than X when the indicator $1_{\{X \geq c\}}$ is not zero. Since the expected value operator preserves inequalities[5], we have

$$c \cdot 1_{\{X \geq c\}} \leq X \cdot 1_{\{X \geq c\}} \Longrightarrow \mathrm{E}\left[c \cdot 1_{\{X \geq c\}}\right] \leq \mathrm{E}\left[X \cdot 1_{\{X \geq c\}}\right]$$

Furthermore, by using the linearity of the expected value[6] and the fact that the expected value of an indicator is equal to the probability of the event it indicates[7], we obtain

$$\mathrm{E}\left[c \cdot 1_{\{X \geq c\}}\right] = c\mathrm{E}\left[1_{\{X \geq c\}}\right] = c\mathrm{P}\left(X \geq c\right) \Longrightarrow c\mathrm{P}\left(X \geq c\right) \leq \mathrm{E}\left[X 1_{\{X \geq c\}}\right]$$

The above inequalities can be put together:

$$\begin{array}{l} \mathrm{E}\left[X\right] \geq \mathrm{E}\left[X 1_{\{X \geq c\}}\right] \\ \mathrm{E}\left[X 1_{\{X \geq c\}}\right] \geq c\mathrm{P}\left(X \geq c\right) \end{array} \Longrightarrow \mathrm{E}\left[X\right] \geq c\mathrm{P}\left(X \geq c\right)$$

Finally, since c is strictly positive, we can divide both sides of the right-hand inequality to obtain Markov's inequality:

$$\mathrm{P}\left(X \geq c\right) \leq \frac{\mathrm{E}\left[X\right]}{c}$$

■

This property also holds when $X \geq 0$ almost surely[8].

29.2 Chebyshev's inequality

Proposition 169 *Let X be a square integrable[9] random variable defined on a sample space Ω. Let μ and σ^2 denote the mean and variance of X respectively. Let $k \in \mathbb{R}_{++}$. Then, the following inequality, called Chebyshev's inequality, holds:*

$$\mathrm{P}\left(|X - \mu| \geq k\right) \leq \frac{\sigma^2}{k^2}$$

Proof. The proof is a straightforward application of Markov's inequality (29.1). Since $(X - \mu)^2$ is a positive random variable, we can apply Markov's inequality with $c = k^2$ to obtain

$$\mathrm{P}\left((X - \mu)^2 \geq k^2\right) \leq \frac{\mathrm{E}\left[(X - \mu)^2\right]}{k^2}$$

But $(X - \mu)^2 \geq k^2$ if and only if $|X - \mu| \geq k$; so, we can write

$$\mathrm{P}\left(|X - \mu| \geq k\right) \leq \frac{\mathrm{E}\left[(X - \mu)^2\right]}{k^2}$$

[5] See p. 150.
[6] See p. 134.
[7] See p. 198.
[8] See the lecture entitled *Zero-probability events* (p. 79) for a definition of almost sure property.
[9] See p. 159.

Furthermore, by the very definition of variance, we have

$$\mathrm{E}\left[(X-\mu)^2\right] = \mathrm{Var}\left[X\right] = \sigma^2$$

Therefore,

$$\mathrm{P}\left(|X-\mu| \geq k\right) \leq \frac{\sigma^2}{k^2}$$

■

29.3 Jensens's inequality

Proposition 170 *Let X be an integrable random variable. Let $g : \mathbb{R} \to \mathbb{R}$ be a convex function such that*

$$Y = g\left(X\right)$$

is also integrable. Then, the following inequality, called Jensen's inequality, holds:

$$\mathrm{E}\left[g\left(X\right)\right] \geq g\left(\mathrm{E}\left[X\right]\right)$$

Proof. A function g is convex if, for any point x_0, the graph of g lies entirely above its tangent at the point x_0:

$$g\left(x\right) \geq g\left(x_0\right) + b\left(x - x_0\right) \quad , \forall x$$

where b is the slope of the tangent. By setting $x = X$ and $x_0 = \mathrm{E}\left[X\right]$, the inequality becomes

$$g\left(X\right) \geq g\left(\mathrm{E}\left[X\right]\right) + b\left(X - \mathrm{E}\left[X\right]\right)$$

By taking the expected value of both sides of the inequality, and by using the fact that the expected value operator preserves inequalities[10], we obtain

$$\begin{array}{rcl} \mathrm{E}\left[g\left(X\right)\right] & \geq & \mathrm{E}\left[g\left(\mathrm{E}\left[X\right]\right) + b\left(X - \mathrm{E}\left[X\right]\right)\right] \\ \boxed{\mathrm{A}} & = & g\left(\mathrm{E}\left[X\right]\right) + b\left(\mathrm{E}\left[X\right] - \mathrm{E}\left[X\right]\right) \\ & = & g\left(\mathrm{E}\left[X\right]\right) \end{array}$$

where: in step $\boxed{\mathrm{A}}$ we have used the linearity of the expected value. ■

If the function g is strictly convex and X is not almost surely constant, then we have a strict inequality:

$$\mathrm{E}\left[g\left(X\right)\right] > g\left(\mathrm{E}\left[X\right]\right)$$

Proof. A function g is strictly convex if, for any point x_0, the graph of g lies entirely above its tangent at the point x_0 (and strictly so for points different from x_0):

$$g\left(x\right) > g\left(x_0\right) + b\left(x - x_0\right) \quad , \forall x \neq x_0$$

where b is the slope of the tangent. By setting $x = X$ and $x_0 = \mathrm{E}\left[X\right]$, the inequality becomes

$$g\left(X\right) > g\left(\mathrm{E}\left[X\right]\right) + b\left(X - \mathrm{E}\left[X\right]\right) \quad , \forall X \neq \mathrm{E}\left[X\right]$$

[10] See p. 150.

and, of course, $g(X) = g(\mathrm{E}[X])$ when $X = \mathrm{E}[X]$. By taking the expected value of both sides of the inequality, and by using the fact that the expected value operator preserves inequalities, we obtain

$$\begin{aligned} \mathrm{E}[g(X)] &> \mathrm{E}[g(\mathrm{E}[X]) + b(X - \mathrm{E}[X])] \\ &= g(\mathrm{E}[X]) + b(\mathrm{E}[X] - \mathrm{E}[X]) \\ &= g(\mathrm{E}[X]) \end{aligned}$$

where the first inequality is strict because we have assumed that X is not almost surely[11] constant, and, as a consequence, the event

$$\{g(X) = g(\mathrm{E}[X])\}$$

does not have probability 1. ■

If the function g is concave, then

$$\mathrm{E}[g(X)] \leq g(\mathrm{E}[X])$$

Proof. If g is concave, then $-g$ is convex, and by Jensen's inequality we have

$$\mathrm{E}[-g(X)] \geq -g(\mathrm{E}[X])$$

By multiplying both sides by -1, and by using the linearity of the expected value, we obtain the result. ■

If the function g is strictly concave and X is not almost surely constant, then

$$\mathrm{E}[g(X)] < g(\mathrm{E}[X])$$

Proof. Similar to previous proof. ■

29.4 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Let X be a positive random variable whose expected value is

$$\mathrm{E}[X] = 10$$

Find a lower bound to the probability

$$\mathrm{P}(X < 20)$$

Solution

First of all, we need to use the formula for the probability of a complement:

$$\mathrm{P}(X < 20) = 1 - \mathrm{P}(X \geq 20)$$

[11] See p. 79.

Now, we can use Markov's inequality:

$$\mathrm{P}\left(X \geq 20\right) \leq \frac{\mathrm{E}\left[X\right]}{20} = \frac{10}{20} = \frac{1}{2}$$

By multiplying both sides of the inequality by -1, we obtain

$$-\mathrm{P}\left(X \geq 20\right) \geq -\frac{1}{2}$$

By adding 1 to both sides of the inequality, we obtain

$$1 - \mathrm{P}\left(X \geq 20\right) \geq 1 - \frac{1}{2} = \frac{1}{2}$$

Thus, the lower bound is

$$\mathrm{P}\left(X < 20\right) \geq \frac{1}{2}$$

Exercise 2

Let X be a random variable such that

$$\begin{aligned}\mathrm{E}\left[X\right] &= 0\\ \mathrm{P}\left(-3 < X < 2\right) &= \frac{1}{2}\end{aligned}$$

Find a lower bound to its variance.

Solution

The lower bound can be derived thanks to Chebyshev's inequality:

$$\begin{aligned}
& & \mathrm{Var}\left[X\right] & \\
\boxed{\mathrm{A}} & \geq & 2^2 \cdot \mathrm{P}\left(\left|X - \mathrm{E}\left[X\right]\right| \geq 2\right) & \\
\boxed{\mathrm{B}} & = & 4 \cdot \mathrm{P}\left(\left|X\right| \geq 2\right) & \\
\boxed{\mathrm{C}} & = & 4 \cdot \left[1 - \mathrm{P}\left(-2 < X < 2\right)\right] & \\
\boxed{\mathrm{D}} & \geq & 4 \cdot \left[1 - \mathrm{P}\left(-3 < X < 2\right)\right] & \\
& = & 4 \cdot \left[1 - \frac{1}{2}\right] = 2 &
\end{aligned}$$

where: in step $\boxed{\mathrm{A}}$ we have used Chebyshev's inequality; in step $\boxed{\mathrm{B}}$ we have used the fact that $\mathrm{E}\left[X\right] = 0$; in step $\boxed{\mathrm{C}}$ we have used the formula for the probability of a complement; in step $\boxed{\mathrm{D}}$ we have used the fact that, by the monotonicity of probability,

$$\mathrm{P}\left(-3 < X < 2\right) \geq \mathrm{P}\left(-2 < X < 2\right)$$

Thus, the lower bound is

$$\mathrm{Var}\left[X\right] \geq 2$$

Exercise 3

Let X be a strictly positive random variable, such that

$$\begin{aligned} \mathrm{E}[X] &= \frac{1}{2} \\ \mathrm{Var}[X] &= 1 \end{aligned}$$

What can you infer, using Jensen's inequality, about the expected value $\mathrm{E}[\ln(2X)]$?

Solution

The function

$$g(x) = \ln(2x)$$

has first derivative

$$\frac{d}{dx} g(x) = \frac{1}{2x} \cdot 2 = \frac{1}{x}$$

and second derivative

$$\frac{d^2}{dx^2} g(x) = -\frac{1}{x^2}$$

The second derivative is strictly negative on the domain of definition of the function. Therefore, the function is strictly concave. Furthermore, X is not almost surely constant because it has strictly positive variance. Hence, by Jensen's inequality, we obtain

$$\mathrm{E}[\ln(2X)] < \ln(2\mathrm{E}[X]) = \ln(1) = 0$$

Chapter 30

Legitimate probability mass functions

In this lecture we analyze two properties of probability mass functions[1]. We prove not only that any probability mass function satisfies these two properties, but also that any function satisfying these two properties is a legitimate probability mass function.

30.1 Properties of probability mass functions

Any probability mass function satisfies two basic properties, stated by the following proposition.

Proposition 171 *Let X be a discrete random variable. Its probability mass function $p_X(x)$ satisfies the following two properties:*

1. ***non-negativity:***
$$p_X(x) \geq 0, \ \forall x \in \mathbb{R} \tag{30.1}$$

2. ***sum over the support equals*** 1***:***
$$\sum_{x \in R_X} p_X(x) = 1 \tag{30.2}$$

where R_X is the support of X.

Proof. Remember that, by the definition of a probability mass function, $p_X(x)$ is such that
$$p_X(x) = \begin{cases} \mathrm{P}(X = x) & \text{if } x \in R_X \\ 0 & \text{if } x \notin R_X \end{cases}$$
Probabilities cannot be negative, therefore $\mathrm{P}(X = x) \geq 0$ and, as a consequence, $p_X(x) \geq 0$. This proves property (30.1). Furthermore, the probability of a sure

[1] See p. 106.

thing[2] must be equal to 1. Since, by the very definition of support, the event $\{X \in R_X\}$ is a sure thing, then

$$1 = \mathrm{P}\left(X \in R_X\right) = \sum_{x \in R_X} p_X\left(x\right)$$

which proves property (30.2). ■

30.2 Identification of a legitimate pmf

Any probability mass function must satisfy properties (30.1) and (30.2). Using some standard results from measure theory (omitted here), it is possible to prove that the converse is also true, i.e., any function $p_X\left(x\right)$ satisfying the two properties above is a probability mass function.

Proposition 172 *Let $p_X\left(x\right)$ be a function satisfying properties (30.1) and (30.2). Then, there exists a discrete random variable X whose probability mass function is $p_X\left(x\right)$.*

This proposition gives us a powerful method for constructing probability mass functions. Take a subset of the set of real numbers $R_X \subseteq \mathbb{R}$. Take any function $g\left(x\right)$ that is non-negative on R_X (non-negative means that $g\left(x\right) \geq 0$ for any $x \in R_X$). If the sum

$$S = \sum_{x \in R_X} g\left(x\right)$$

is well-defined and is finite and strictly positive, then define

$$p_X\left(x\right) = \frac{1}{S} g\left(x\right)$$

S is strictly positive, thus $p_X\left(x\right)$ is non-negative and it satisfies property (30.1). It also satisfies property (30.2), because

$$\begin{aligned} \sum_{x \in R_X} p_X\left(x\right) &= \sum_{x \in R_X} \frac{1}{S} g\left(x\right) \\ &= \frac{1}{S} \sum_{x \in R_X} g\left(x\right) \\ &= \frac{1}{S} S = 1 \end{aligned}$$

Therefore, any function $g\left(x\right)$ that is non-negative on R_X (R_X is chosen arbitrarily) can be used to construct a probability mass function if its sum over R_X is well-defined and is finite and strictly positive.

Example 173 *Define*

$$R_X = \{1, 2, 3, 4, 5\}$$

and a function

$$g\left(x\right) = \begin{cases} x^2 & \text{if } x \in R_X \\ 0 & \text{otherwise} \end{cases}$$

[2] See the properties of probability (p. 70).

Can we use $g(x)$ to build a probability mass function? First of all, we have to check that $g(x)$ is non-negative. This is obviously true, because x^2 is always non-negative. Then, we have to check that the sum of $g(x)$ over R_X exists and is finite and strictly positive:

$$\begin{aligned} S &= \sum_{x \in R_X} g(x) = g(1) + g(2) + g(3) + g(4) + g(5) \\ &= 1 + 4 + 9 + 16 + 25 = 55 \end{aligned}$$

Since S exists and is finite and strictly positive, we can define

$$p_X(x) = \frac{1}{S} g(x) = \begin{cases} \frac{1}{55} x^2 & \text{if } x \in R_X \\ 0 & \text{otherwise} \end{cases}$$

By the above proposition, $p_X(x)$ is a legitimate probability mass function.

30.3 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Consider the following function:

$$p_X(x) = \begin{cases} \frac{1}{10} x & \text{if } x \in \{1, 2, 3, 4\} \\ 0 & \text{otherwise} \end{cases}$$

Prove that $p_X(x)$ is a legitimate probability mass function.

Solution

For $x \in \{1, 2, 3, 4\}$ we have

$$p_X(x) = \frac{1}{10} x > 0$$

while for $x \notin \{1, 2, 3, 4\}$ we have

$$p_X(x) = 0$$

Therefore, $p_X(x) \geq 0$ for any $x \in \mathbb{R}$, which implies that property (30.1) is satisfied. Also property (30.2) is satisfied, because

$$\begin{aligned} \sum_{x \in R_X} p_X(x) &= p_X(1) + p_X(2) + p_X(3) + p_X(4) \\ &= \frac{1}{10} \cdot 1 + \frac{1}{10} \cdot 2 + \frac{1}{10} \cdot 3 + \frac{1}{10} \cdot 4 = \frac{10}{10} = 1 \end{aligned}$$

Exercise 2

Consider the function

$$p_X(x) = \begin{cases} \frac{1}{14} x^2 & \text{if } x \in \{1, 2, 3\} \\ 0 & \text{otherwise} \end{cases}$$

Prove that $p_X(x)$ is a legitimate probability mass function.

Solution

For $x \in \{1, 2, 3\}$ we have

$$p_X(x) = \frac{1}{14}x^2 > 0$$

while for $x \notin \{1, 2, 3\}$ we have

$$p_X(x) = 0$$

Therefore, $p_X(x) \geq 0$ for any $x \in \mathbb{R}$, which implies that property (30.1) is satisfied. Also property (30.2) is satisfied, because

$$\begin{aligned}\sum_{x \in R_X} p_X(x) &= p_X(1) + p_X(2) + p_X(3) \\ &= \frac{1}{14} \cdot 1^2 + \frac{1}{14} \cdot 2^2 + \frac{1}{14} \cdot 3^2 \\ &= \frac{1}{14} \cdot 1 + \frac{1}{14} \cdot 4 + \frac{1}{14} \cdot 9 = \frac{14}{14} = 1\end{aligned}$$

Exercise 3

Consider the function

$$p_X(x) = \begin{cases} \frac{3}{4} \cdot 4^{1-x} & \text{if } x \in \mathbb{N} \\ 0 & \text{otherwise} \end{cases}$$

Prove that $p_X(x)$ is a legitimate probability mass function.

Solution

For $x \in \mathbb{N}$ we have

$$p_X(x) = \frac{3}{4} \cdot 4^{1-x} > 0$$

because 4^{1-x} is strictly positive. For $x \notin \mathbb{N}$ we have

$$p_X(x) = 0$$

Therefore, $p_X(x) \geq 0$ for any $x \in \mathbb{R}$, which implies that property (30.1) is satisfied. Also property (30.2) is satisfied, because

$$\begin{aligned}\sum_{x \in R_X} p_X(x) &= \sum_{x=1}^{\infty} p_X(x) = \sum_{x=1}^{\infty} \frac{3}{4} \cdot 4^{1-x} \\ &= \frac{3}{4} \sum_{x=1}^{\infty} \left(\frac{1}{4}\right)^{x-1} = \frac{3}{4}\left[1 + \frac{1}{4} + \left(\frac{1}{4}\right)^2 + \left(\frac{1}{4}\right)^3 + \ldots\right] \\ &= \frac{3}{4} \cdot \frac{1}{1 - \frac{1}{4}} = \frac{3}{4} \cdot \frac{1}{\frac{3}{4}} = 1\end{aligned}$$

Chapter 31

Legitimate probability density functions

In this lecture we analyze two properties of probability density functions[1]. We prove not only that any probability density function satisfies these two properties, but also that any function satisfying these two properties is a legitimate probability density function.

31.1 Properties of probability density functions

Any probability density function satisfies two basic properties, stated by the following proposition.

Proposition 174 *Let X be an absolutely continuous random variable. Its probability density function $f_X(x)$ satisfies the following two properties:*

1. ***non-negativity:***
$$f_X(x) \geq 0, \ \forall x \in \mathbb{R} \tag{31.1}$$

2. ***integral over $\mathbb{R}$ equals 1:***
$$\int_{-\infty}^{\infty} f_X(x)\,dx = 1 \tag{31.2}$$

Proof. Remember that, by the definition of a probability density function, $f_X(x)$ is such that
$$\int_a^b f_X(x)\,dx = \mathrm{P}(X \in [a,b])$$
for any interval $[a,b]$. Probabilities cannot be negative; therefore, $\mathrm{P}(X \in [a,b]) \geq 0$ and
$$\int_a^b f_X(x)\,dx \geq 0$$
for any interval $[a,b]$. But the above integral can be non-negative for all intervals $[a,b]$ only if the integrand function itself is non-negative, i.e. if $f_X(x) \geq 0$ for all

[1] See p. 107.

x. This proves property (31.1). Furthermore, the probability of a sure thing[2] must be equal to 1. Since $\{X \in (-\infty, \infty)\}$ is a sure thing, then

$$1 = \mathrm{P}\left(X \in (-\infty, \infty)\right) = \int_{-\infty}^{\infty} f_X(x)\, dx$$

which proves property (31.2). ■

31.2 Identification of a legitimate pdf

Any probability density function must satisfy properties (31.1) and (31.2) above. Using some standard results from measure theory (omitted here), it is possible to prove that the converse is also true, i.e., any function $f_X(x)$ satisfying the two properties above is a probability density function.

Proposition 175 *Let $f_X(x)$ be a function satisfying properties (31.1) and (31.2). Then, there exists an absolutely continuous random variable X whose probability density function is $f_X(x)$.*

This proposition gives us a powerful method for constructing probability density functions. Take any non-negative[3] function $g(x)$. If the integral

$$I = \int_{-\infty}^{\infty} g(x)\, dx$$

exists and is finite and strictly positive, then define

$$f_X(x) = \frac{1}{I} g(x)$$

I is strictly positive, thus $f_X(x)$ is non-negative and it satisfies property (31.1). It also satisfies Property (31.2), because

$$\begin{aligned} \int_{-\infty}^{\infty} f_X(x)\, dx &= \int_{-\infty}^{\infty} \frac{1}{I} g(x)\, dx \\ &= \frac{1}{I} \int_{-\infty}^{\infty} g(x)\, dx \\ &= \frac{1}{I} I = 1 \end{aligned}$$

Therefore, any non-negative function $g(x)$ can be used to construct a probability density function if its integral over $\mathbb{R}$ exists and is finite and strictly positive.

Example 176 *Define a function $g(x)$ as*

$$g(x) = \begin{cases} x^2 & \text{if } x \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

[2] See the properties of probability (p. 70).

[3] Non-negative means that $g(x) \geq 0$ for any $x \in \mathbb{R}$.

Can we use $g(x)$ to build a probability density function? First of all, we have to check that $g(x)$ is non-negative. This is obviously true, because x^2 is always non-negative. Then, we have to check that the integral of $g(x)$ over $\mathbb{R}$ exists and is finite and strictly positive:

$$\begin{aligned} I &= \int_{-\infty}^{\infty} g(x)\,dx = \int_0^1 x^2 dx \\ &= \left[\frac{1}{3}x^3\right]_0^1 = \frac{1}{3} - 0 = \frac{1}{3} \end{aligned}$$

Since I exists and is finite and strictly positive, we can define

$$f_X(x) = \frac{1}{I} g(x) = \begin{cases} 3x^2 & \text{if } x \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

By Proposition 175, $f_X(x)$ is a legitimate probability density function.

31.3 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Consider the function

$$f_X(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{if } x \in [0,\infty) \\ 0 & \text{if } x \in (-\infty, 0) \end{cases}$$

where $\lambda \in (0,\infty)$. Prove that $f_X(x)$ is a legitimate probability density function.

Solution

Since $\lambda > 0$ and the exponential function is strictly positive, $f_X(x) \geq 0$ for any $x \in \mathbb{R}$, so the non-negativity property is satisfied. The integral property is also satisfied, because

$$\begin{aligned} \int_{-\infty}^{\infty} f_X(x)\,dx &= \int_0^{\infty} \lambda \exp(-\lambda x)\,dx \\ &= [-\exp(-\lambda x)]_0^{\infty} \\ &= -0 - (-1) = 1 \end{aligned}$$

Exercise 2

Consider the function

$$f_X(x) = \begin{cases} \frac{1}{u-l} & \text{if } x \in [l,u] \\ 0 & \text{if } x \notin [l,u] \end{cases}$$

where $l, u \in \mathbb{R}$ and $l < u$. Prove that $f_X(x)$ is a legitimate probability density function.

Solution

$l < u$ implies $\frac{1}{u-l} > 0$, so $f_X(x) \geq 0$ for any $x \in \mathbb{R}$ and the non-negativity property is satisfied. The integral property is also satisfied, because

$$\begin{aligned}\int_{-\infty}^{\infty} f_X(x)\,dx &= \int_l^u \frac{1}{u-l}dx \\ &= \frac{1}{u-l}\int_l^u dx \\ &= \frac{1}{u-l}[x]_l^u \\ &= \frac{1}{u-l}(u-l) = 1\end{aligned}$$

Exercise 3

Consider the function

$$f_X(x) = \begin{cases} 2^{-n/2}\left(\Gamma(n/2)\right)^{-1} x^{n/2-1}\exp\left(-\frac{1}{2}x\right) & \text{if } x \in [0,\infty) \\ 0 & \text{if } x \notin [0,\infty)\end{cases}$$

where $n \in \mathbb{N}$ and $\Gamma()$ is the Gamma function[4]. Prove that $f_X(x)$ is a legitimate probability density function.

Solution

Remember the definition of Gamma function:

$$\Gamma(z) = \int_0^{\infty} x^{z-1}\exp(-x)\,dx$$

$\Gamma(z)$ is obviously strictly positive for any z, since $\exp(-x)$ is strictly positive and x^{z-1} is strictly positive on the interval of integration (except at 0, where it is 0). Therefore, $f_X(x)$ satisfies the non-negativity property, because the four factors in the product

$$2^{-n/2}\cdot\left(\Gamma(n/2)\right)^{-1}\cdot x^{n/2-1}\cdot\exp\left(-\frac{1}{2}x\right)$$

are all non-negative on the interval $[0,\infty)$.

The integral property is also satisfied, because

$$\begin{aligned}\int_{-\infty}^{\infty} f_X(x)\,dx &= \int_0^{\infty} 2^{-n/2}\left(\Gamma(n/2)\right)^{-1} x^{n/2-1}\exp\left(-\frac{1}{2}x\right)dx \\ &= 2^{-n/2}\left(\Gamma(n/2)\right)^{-1}\int_0^{\infty} x^{n/2-1}\exp\left(-\frac{1}{2}x\right)dx \\ \boxed{\text{A}} &= 2^{-n/2}\left(\Gamma(n/2)\right)^{-1}\int_0^{\infty}(2t)^{n/2-1}\exp\left(-\frac{1}{2}2t\right)2dt \\ &= 2^{-n/2}\left(\Gamma(n/2)\right)^{-1}(2)^{n/2-1}\,2\int_0^{\infty} t^{n/2-1}\exp(-t)\,dt \\ &= \left(\Gamma(n/2)\right)^{-1}\int_0^{\infty} t^{n/2-1}\exp(-t)\,dt\end{aligned}$$

[4] See p. 55.

$$\begin{aligned} \boxed{\text{B}} \quad &= \left(\Gamma\left(n/2\right)\right)^{-1}\Gamma\left(n/2\right) \\ &= 1 \end{aligned}$$

where: in step $\boxed{\text{A}}$ we have made a change of variable ($x = 2t$); in step $\boxed{\text{B}}$ we have used the definition of Gamma function.

Chapter 32

Factorization of joint probability mass functions

This lecture discusses how to factorize the joint probability mass function[1] of two discrete random variables X and Y into two factors:

1. the conditional probability mass function[2] of X given $Y = y$;
2. the marginal probability mass function[3] of Y.

32.1 The factorization

The factorization, which has already been discussed in the lecture entitled *Conditional probability distributions* (p. 209), is formally stated in the following proposition.

Proposition 177 (factorization) *Let $[X \;\; Y]$ be a discrete random vector with support R_{XY} and joint probability mass function $p_{XY}(x,y)$. Denote by $p_{X|Y=y}(x)$ the conditional probability mass function of X given $Y = y$ and by $p_Y(y)$ the marginal probability mass function of Y. Then:*

$$p_{XY}(x,y) = p_{X|Y=y}(x)\, p_Y(y)$$

for any x and y.

32.2 A factorization method

When we know the joint probability mass function $p_{XY}(x,y)$ and we need to factorize it into the conditional probability mass function $p_{X|Y=y}(x)$ and the marginal probability mass function $p_Y(y)$, we usually proceed in two steps:

1. marginalize $p_{XY}(x,y)$ by summing it over all possible values of x and obtain the marginal probability mass function $p_Y(y)$;

[1] See p. 116.
[2] See p. 210.
[3] See p. 120.

2. divide $p_{XY}(x,y)$ by $p_Y(y)$ and obtain the conditional probability mass function $p_{X|Y=y}(x)$ (of course this step makes sense only when $p_Y(y) > 0$).

In some cases, the first step (marginalization) can be difficult to perform. In these cases, it is possible to avoid the marginalization step, by making a guess about the factorization of $p_{XY}(x,y)$ and verifying whether the guess is correct with the help of the following proposition:

Proposition 178 (factorization method) *Suppose there are two functions $h(y)$ and $g(x,y)$ such that:*

1. *for any x and y, the following holds:*

$$p_{XY}(x,y) = g(x,y)\,h(y)$$

2. *for any fixed y, $g(x,y)$, considered as a function of x, is a probability mass function with the same support of X (i.e. R_X).*

Then:

$$\begin{aligned} p_{X|Y=y}(x) &= g(x,y) \\ p_Y(y) &= h(y) \end{aligned}$$

Proof. The marginal probability mass function of Y satisfies:

$$p_Y(y) = \sum_{x \in R_X} p_{XY}(x,y)$$

Therefore, by property 1:

$$\begin{aligned} p_Y(y) &= \sum_{x \in R_X} f_{XY}(x,y) \\ &= \sum_{x \in R_X} g(x,y)\,h(y) \\ \boxed{\text{A}} &= h(y) \sum_{x \in R_X} g(x,y) \\ \boxed{\text{B}} &= h(y) \end{aligned}$$

where: in step A we have used the fact that $h(y)$ does not depend on x; in step B we have used the fact that, for any fixed y, $g(x,y)$, considered as a function of x, is a probability mass function and the sum[4] of a probability mass function over its support equals 1. Therefore,

$$p_{XY}(x,y) = g(x,y)\,h(y) = g(x,y)\,p_Y(y), \ \forall (x,y)$$

Since we also have that

$$p_{XY}(x,y) = p_{X|Y=y}(x)\,p_Y(y), \ \forall (x,y)$$

[4] See p. 247.

then, by necessity, it must be that:

$$g(x,y) = p_{X|Y=y}(x)$$

■

Thus, whenever we are given a formula for the joint probability mass function $p_{XY}(x,y)$ and we want to find the marginal and the conditional functions, we have to manipulate the formula and express it as the product of:

1. a function of x and y that is a probability mass function in x for all values of y;
2. a function of y that does not depend on x.

Example 179 *Let X be a 3×1 random vector having a multinomial distribution*[5] *with parameters p_1, p_2 and p_3 (the probabilities of the three possible outcomes of each trial) and n (the number of trials). The probabilities are strictly positive numbers such that:*

$$p_1 + p_2 + p_3 = 1$$

The support of X is

$$R_X = \left\{x \in \mathbb{Z}_+^3 : x_1 + x_2 + x_3 = n\right\}$$

where x_1, x_2, x_3 denote the components of the vector x. The joint probability mass function of X is

$$p_X(x_1,x_2,x_3) = \begin{cases} \frac{n!}{x_1!x_2!x_3!}p_1^{x_1}p_2^{x_2}p_2^{x_3} & \text{if } (x_1,x_2,x_3) \in R_X \\ 0 & \text{otherwise} \end{cases}$$

Note that:

$$\begin{aligned}
& \frac{n!}{x_1!x_2!x_3!}p^{x_1}p^{x_2}p^{x_3} \\
= \; & \frac{(n-x_3)!}{x_1!x_2!}\frac{n!}{(n-x_3)!x_3!}\frac{p_1^{x_1}p_2^{x_2}}{p_3^{n-x_3}}p_3^{x_3}p_3^{n-x_3} \\
= \; & \left(\frac{(n-x_3)!}{x_1!x_2!}\frac{p_1^{x_1}p_2^{x_2}}{p_3^{n-x_3}}\right)\left(\frac{n!}{(n-x_3)!x_3!}p_3^{x_3}p_3^{n-x_3}\right) \\
\boxed{A} \; = \; & \left(\frac{(n-x_3)!}{x_1!x_2!}\frac{p_1^{x_1}p_2^{x_2}}{p_3^{x_1+x_2}}\right)\left(\frac{n!}{(n-x_3)!x_3!}p_3^{x_3}p_3^{n-x_3}\right) \\
= \; & \left(\frac{(n-x_3)!}{x_1!x_2!}(p_1/p_3)^{x_1}(p_2/p_3)^{x_2}\right)\left(\frac{n!}{(n-x_3)!x_3!}p_3^{x_3}p_3^{n-x_3}\right)
\end{aligned}$$

where in step $\boxed{A}$ we have used the fact that

$$x_1 + x_2 + x_3 = n$$

Therefore, the joint probability mass function can be factorized as:

$$p_X(x_1,x_2,x_3) = g(x_1,x_2,x_3)\,h(x_3)$$

[5] See p. 431.

where

$$g(x_1, x_2, x_3) = \begin{cases} \frac{(n-x_3)!}{x_1!x_2!}(p_1/p_3)^{x_1}(p_2/p_3)^{x_2} & \text{if } (x_1, x_2) \in \mathbb{Z}_+^2 \\ & \text{and } x_1 + x_2 = n - x_3 \\ 0 & \text{otherwise} \end{cases}$$

and:

$$h(x_3) = \begin{cases} \frac{n!}{(n-x_3)!x_3!} p_3^{x_3} p_3^{n-x_3} & \text{if } x_3 \in \mathbb{Z}_+ \text{ and } x_3 \leq n \\ 0 & \text{otherwise} \end{cases}$$

For any $x_3 \leq n$, $g(x_1, x_2, x_3)$ is the probability mass function of a multinomial distribution with parameters p_1/p_3, p_2/p_3 and $n - x_3$. Therefore:

$$\begin{aligned} p_{X_1, X_2 | X_3 = x_3}(x_1, x_2) &= g(x_1, x_2, x_3) \\ p_{X_3}(x_3) &= h(x_3) \end{aligned}$$

Chapter 33

Factorization of joint probability density functions

This lecture discusses how to factorize the joint probability density function[1] of two absolutely continuous random variables (or random vectors) X and Y into two factors:

1. the conditional probability density function[2] of X given $Y = y$;
2. the marginal probability density function[3] of Y.

33.1 The factorization

The factorization, which has already been discussed in the lecture entitled *Conditional probability distributions* (p. 209), is formally stated in the following proposition.

Proposition 180 (factorization) *Let $[X \;\; Y]$ be an absolutely continuous random vector with support R_{XY} and joint probability density function $f_{XY}(x, y)$. Denote by $f_{X|Y=y}(x)$ the conditional probability density function of X given $Y = y$ and by $f_Y(y)$ the marginal probability density function of Y. Then*

$$f_{XY}(x, y) = f_{X|Y=y}(x) f_Y(y)$$

for any x and y.

33.2 A factorization method

When we know the joint probability density function $f_{XY}(x, y)$ and we need to factorize it into the conditional probability density function $f_{X|Y=y}(x)$ and the marginal probability density function $f_Y(y)$, we usually proceed in two steps:

1. marginalize $f_{XY}(x, y)$ by integrating it with respect to x and obtain the marginal probability density function $f_Y(y)$;

[1] See p. 116.
[2] See p. 213.
[3] See p. 120.

2. divide $f_{XY}(x,y)$ by $f_Y(y)$ and obtain the conditional probability density function $f_{X|Y=y}(x)$ (of course this step makes sense only when $f_Y(y) > 0$).

In some cases, the first step (marginalization) can be difficult to perform. In these cases, it is possible to avoid the marginalization step, by making a guess about the factorization of $f_{XY}(x,y)$ and verifying whether the guess is correct with the help of the following proposition:

Proposition 181 (factorization method) *Suppose there are two functions $h(y)$ and $g(x,y)$ such that:*

1. *for any x and y, the following holds:*

$$f_{XY}(x,y) = g(x,y)\,h(y)$$

2. *for any fixed y, $g(x,y)$, considered as a function of x, is a probability density function.*

Then:

$$\begin{aligned} f_{X|Y=y}(x) &= g(x,y) \\ f_Y(y) &= h(y) \end{aligned}$$

Proof. The marginal probability density of Y satisfies

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

Thus, by property 1 above:

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{XY}(x,y)\,dx \\ &= \int_{-\infty}^{\infty} g(x,y)\,h(y)\,dx \\ \boxed{\text{A}} &= h(y)\int_{-\infty}^{\infty} g(x,y)\,dx \\ \boxed{\text{B}} &= h(y) \end{aligned}$$

where: in step A we have used the fact that $h(y)$ does not depend on x; in step B we have used the fact that, for any fixed y, $g(x,y)$, considered as a function of x, is a probability density function and the integral of a probability density function over $\mathbb{R}$ equals 1 (see p. 251). Therefore

$$f_{XY}(x,y) = g(x,y)\,h(y) = g(x,y)\,f_Y(y)$$

which, in turn, implies

$$g(x,y) = \frac{f_{XY}(x,y)}{f_Y(y)} = f_{X|Y=y}(x)$$

■

Thus, whenever we are given a formula for the joint density function $f_{XY}(x,y)$ and we want to find the marginal and the conditional functions, we have to manipulate the formula and express it as the product of:

1. a function of x and y that is a probability density function in x for all values of y;
2. a function of y that does not depend on x.

Example 182 *Let the joint density function of X and Y be*

$$f_{XY}(x,y) = \begin{cases} \frac{1}{2} y \exp(-yx) & \text{if } x \in [0,\infty) \text{ and } y \in [1,3] \\ 0 & \text{otherwise} \end{cases}$$

The joint density can be factorized as follows:

$$f_{XY}(x,y) = g(x,y)\, h(y)$$

where

$$g(x,y) = \begin{cases} y \exp(-yx) & \text{if } x \in [0,\infty) \\ 0 & \text{otherwise} \end{cases}$$

and

$$h(y) = \begin{cases} \frac{1}{2} & \text{if } y \in [1,3] \\ 0 & \text{otherwise} \end{cases}$$

Note that $g(x,y)$ is a probability density function in x for any fixed y (it is the probability density function of an exponential random variable[4] with parameter y). Therefore:

$$\begin{aligned} f_{X|Y=y}(x) &= g(x,y) \\ f_Y(y) &= h(y) \end{aligned}$$

[4] See p. 365.

Chapter 34

Functions of random variables and their distribution

Let X be a random variable with known distribution. Let another random variable Y be a function of X:

$$Y = g(X)$$

where $g : \mathbb{R} \to \mathbb{R}$. How do we derive the distribution of Y from the distribution of X?

There is no general answer to this question. However, there are several special cases in which it is easy to derive the distribution of Y. We discuss these cases below.

34.1 Strictly increasing functions

When the function g is strictly increasing on the support of X, i.e.

$$\forall x_1, x_2 \in R_X, x_1 > x_2 \Rightarrow g(x_1) > g(x_2)$$

then g admits an inverse defined on the support of Y, i.e. a function $g^{-1}(y)$ such that:

$$X = g^{-1}(Y)$$

Furthermore $g^{-1}(y)$ is itself strictly increasing.

The distribution function of a strictly increasing function of a random variable can be computed as follows:

Proposition 183 (cdf of an increasing function) *Let X be a random variable with support R_X and distribution function[1] $F_X(x)$. Let $g : \mathbb{R} \to \mathbb{R}$ be strictly increasing on the support of X. Then, the support of $Y = g(X)$ is*

$$R_Y = \{y = g(x) : x \in R_X\}$$

[1] See p. 108.

and the distribution function of Y is

$$F_Y(y) = \begin{cases} 0 & \text{if } y < \chi, \forall \chi \in R_Y \\ F_X\left(g^{-1}(y)\right) & \text{if } y \in R_Y \\ 1 & \text{if } y > \chi, \forall \chi \in R_Y \end{cases}$$

Proof. Of course, the support R_Y is determined by $g(x)$ and by all the values X can take. The distribution function of Y can be derived as follows:

- if y is lower than than the lowest value Y can take on, then $\mathrm{P}(Y \leq y) = 0$, so:
$$F_Y(y) = 0 \text{ if } y < \chi, \forall \chi \in R_Y$$

- if y belongs to the support of Y, then $F_Y(y)$ can be derived as follows:
$$\begin{aligned} & F_Y(y) \\ \boxed{\text{A}} \quad &= \mathrm{P}(Y \leq y) \\ \boxed{\text{B}} \quad &= \mathrm{P}(g(X) \leq y) \\ \boxed{\text{C}} \quad &= \mathrm{P}\left(g^{-1}(g(X)) \leq g^{-1}(y)\right) \\ &= \mathrm{P}\left(X \leq g^{-1}(y)\right) \\ \boxed{\text{D}} \quad &= F_X\left(g^{-1}(y)\right) \end{aligned}$$
where: in step A we have used the definition of distribution function of Y; in step B we have used the definition of Y; in step C we have used the fact that g^{-1} exists and is strictly increasing on the support of Y; in step D we have used the definition of distribution function of X;

- if y is higher than than the highest value Y can take on, then $\mathrm{P}(Y \leq y) = 1$, so:
$$F_Y(y) = 1 \text{ if } y > \chi, \forall \chi \in R_Y$$

■

Therefore, in the case of an increasing function, knowledge of g^{-1} and of the upper and lower bounds of the support of Y is all we need to derive the distribution function of Y from the distribution function of X.

Example 184 *Let X be a random variable with support*

$$R_X = [1, 2]$$

and distribution function

$$F_X(x) = \begin{cases} 0 & \text{if } x < 1 \\ \frac{1}{2}x & \text{if } 1 \leq x \leq 2 \\ 1 & \text{if } x > 2 \end{cases}$$

Let

$$Y = X^2$$

The function $g(x) = x^2$ *is strictly increasing and it admits an inverse on the support of* X*:*

$$g^{-1}(y) = \sqrt{y}$$

The support of Y *is* $R_Y = [1, 4]$*. The distribution function of* Y *is*

$$F_Y(y) = \begin{cases} 0 & \text{if } y < \chi, \forall \chi \in R_Y, \text{ i.e. if } y < 1 \\ F_X\left(g^{-1}(y)\right) = \frac{1}{2}\sqrt{y} & \text{if } y \in R_Y, \text{ i.e. if } 1 \leq y \leq 4 \\ 1 & \text{if } y > \chi, \forall \chi \in R_Y, \text{ i.e. if } y > 4 \end{cases}$$

In the cases in which X is either discrete or absolutely continuous there are specialized formulae for the probability mass and probability density functions, which are reported below.

34.1.1 Strictly increasing functions of a discrete variable

When X is a discrete random variable, the probability mass function of $Y = g(X)$ can be computed as follows:

Proposition 185 (pmf of an increasing function) *Let* X *be a discrete random variable with support* R_X *and probability mass function* $p_X(x)$*. Let* $g : \mathbb{R} \to \mathbb{R}$ *be strictly increasing on the support of* X*. Then, the support of* $Y = g(X)$ *is*

$$R_Y = \{y = g(x) : x \in R_X\}$$

and its probability mass function is

$$p_Y(y) = \begin{cases} p_X\left(g^{-1}(y)\right) & \text{if } y \in R_Y \\ 0 & \text{if } y \notin R_Y \end{cases}$$

Proof. This proposition is a trivial consequence of the fact that a strictly increasing function is invertible:

$$\begin{aligned} p_Y(y) &= \mathrm{P}(Y = y) \\ &= \mathrm{P}(g(X) = y) \\ &= \mathrm{P}\left(X = g^{-1}(y)\right) \\ &= p_X\left(g^{-1}(y)\right) \end{aligned}$$

■

Example 186 *Let* X *be a discrete random variable with support*

$$R_X = \{1, 2, 3\}$$

and probability mass function

$$p_X(x) = \begin{cases} \frac{1}{6}x & \text{if } x \in R_X \\ 0 & \text{if } x \notin R_X \end{cases}$$

Let

$$Y = g(X) = 3 + X^2$$

The support of Y *is*

$$R_Y = \{4, 7, 12\}$$

The function g is strictly increasing and its inverse is

$$g^{-1}(y) = \sqrt{y-3}$$

The probability mass function of Y is:

$$p_Y(y) = \begin{cases} \frac{1}{6}\sqrt{y-3} & \text{if } y \in R_Y \\ 0 & \text{if } y \notin R_Y \end{cases}$$

34.1.2 Strictly increasing functions of a continuous variable

When X is an absolutely continuous random variable and g is differentiable, then also Y is absolutely continuous and its probability density function can be easily computed as follows:

Proposition 187 (pdf of an increasing function) *Let X be an absolutely continuous random variable with support R_X and probability density function $f_X(x)$. Let $g : \mathbb{R} \to \mathbb{R}$ be strictly increasing and differentiable on the support of X. Then, the support of $Y = g(X)$ is*

$$R_Y = \{y = g(x) : x \in R_X\}$$

and its probability density function is

$$f_Y(y) = \begin{cases} f_X\left(g^{-1}(y)\right)\frac{dg^{-1}(y)}{dy} & \text{if } y \in R_Y \\ 0 & \text{if } y \notin R_Y \end{cases}$$

Proof. This proposition is a trivial consequence of the fact that the density function is the first derivative of the distribution function[2]: it can be obtained by differentiating the expression for the distribution function $F_Y(y)$ found above. ■

Example 188 *Let X be an absolutely continuous random variable with support*

$$R_X = (0, 1]$$

and probability density function

$$f_X(x) = \begin{cases} 2x & \text{if } x \in R_X \\ 0 & \text{if } x \notin R_X \end{cases}$$

Let

$$Y = g(X) = \ln(X)$$

The support of Y is

$$R_Y = (-\infty, 0]$$

The function g is strictly increasing and its inverse is

$$g^{-1}(y) = \exp(y)$$

with derivative

$$\frac{dg^{-1}(y)}{dy} = \exp(y)$$

The probability density function of Y is

$$f_Y(y) = \begin{cases} 2\exp(y)\exp(y) = 2\exp(2y) & \text{if } y \in R_Y \\ 0 & \text{if } y \notin R_Y \end{cases}$$

[2] See p. 109.

34.2 Strictly decreasing functions

When the function g is strictly decreasing on the support of X, i.e.

$$\forall x_1, x_2 \in R_X, x_1 > x_2 \Rightarrow g(x_1) < g(x_2)$$

then g admits an inverse defined on the support of Y, i.e. a function $g^{-1}(y)$ such that

$$X = g^{-1}(Y)$$

Furthermore $g^{-1}(y)$ is itself strictly decreasing.

The distribution function of a strictly decreasing function of a random variable can be computed as follows:

Proposition 189 (cdf of a decreasing function) *Let X be a random variable with support R_X and distribution function $F_X(x)$. Let $g : \mathbb{R} \to \mathbb{R}$ be strictly decreasing on the support of X. Then, the support of $Y = g(X)$ is:*

$$R_Y = \{y = g(x) : x \in R_X\}$$

and the distribution function of Y is:

$$F_Y(y) = \begin{cases} 0 & \text{if } y < \chi, \forall \chi \in R_Y \\ 1 - F_X\left(g^{-1}(y)\right) + \mathrm{P}\left(X = g^{-1}(y)\right) & \text{if } y \in R_Y \\ 1 & \text{if } y > \chi, \forall \chi \in R_Y \end{cases}$$

Proof. Of course, the support R_Y is determined by $g(x)$ and by all the values X can take. The distribution function of y can be derived as follows:

- if y is lower than than the lowest value Y can take on, then $\mathrm{P}(Y \leq y) = 0$, so:
$$F_Y(y) = 0 \text{ if } y < \chi, \forall \chi \in R_Y$$
- if y belongs to the support of Y, then $F_Y(y)$ can be derived as follows:

$$\begin{aligned}
& & F_Y(y) & \\
& \boxed{\text{A}} & = & \mathrm{P}(Y \leq y) \\
& & = & 1 - \mathrm{P}(Y > y) \\
& \boxed{\text{B}} & = & 1 - \mathrm{P}(g(X) > y) \\
& \boxed{\text{C}} & = & 1 - \mathrm{P}\left(g^{-1}(g(X)) < g^{-1}(y)\right) \\
& & = & 1 - \mathrm{P}\left(X < g^{-1}(y)\right) \\
& & = & 1 - \mathrm{P}\left(X < g^{-1}(y)\right) - \mathrm{P}\left(X = g^{-1}(y)\right) + \mathrm{P}\left(X = g^{-1}(y)\right) \\
& & = & 1 - \mathrm{P}\left(X \leq g^{-1}(y)\right) + \mathrm{P}\left(X = g^{-1}(y)\right) \\
& \boxed{\text{D}} & = & 1 - F_X\left(g^{-1}(y)\right) + \mathrm{P}\left(X = g^{-1}(y)\right)
\end{aligned}$$

where: in step A we have used the definition of distribution function of Y; in step B we have used the definition of Y; in step C we have used the fact that g^{-1} exists and is strictly decreasing; in step D we have used the definition of distribution function of X;

- if y is higher than than the highest value Y can take on, then $\mathrm{P}(Y \leq y) = 1$, so:
$$F_Y(y) = 1 \text{ if } y > \chi, \forall \chi \in R_Y$$

■

Therefore, also in the case of a decreasing function, knowledge of g^{-1} and of the upper and lower bounds of the support of Y is all we need to derive the distribution function of Y from the distribution function of X.

Example 190 *Let X be a random variable with support*
$$R_X = [1, 2]$$
and distribution function
$$F_X(x) = \begin{cases} 0 & \text{if } x < 1 \\ \frac{1}{2}x & \text{if } 1 \leq x \leq 2 \\ 1 & \text{if } x > 2 \end{cases}$$
Let
$$Y = -X^2$$
The function $g(x) = -x^2$ is strictly decreasing and it admits an inverse on the support of X:
$$g^{-1}(y) = \sqrt{-y}$$
The support of Y is $R_Y = [-4, -1]$. The distribution function of Y is
$$F_Y(y) = \begin{cases} 0 & \text{if } y < \chi, \forall \chi \in R_Y, \\ & \text{i.e. if } y < -4 \\ 1 - F_X\left(g^{-1}(y)\right) + \mathrm{P}\left(X = g^{-1}(y)\right) & \text{if } y \in R_Y, \\ = 1 - \frac{1}{2}\sqrt{-y} + \frac{1}{2}1_{\{y=-1\}} & \text{i.e. if } -4 \leq y \leq -1 \\ 1 & \text{if } y > \chi, \forall \chi \in R_Y, \\ & \text{i.e. if } y > -1 \end{cases}$$
where $1_{\{y=-1\}}$ equals 1 when $y = -1$ and 0 otherwise (because $\mathrm{P}\left(X = g^{-1}(y)\right)$ is always zero except when $y = -1$ and $g^{-1}(y) = 1$).

We report below the formulae for the special cases in which X is either discrete or absolutely continuous.

34.2.1 Strictly decreasing functions of a discrete variable

When X is a discrete random variable, the probability mass function of $Y = g(X)$ can be computed as follows:

Proposition 191 (pmf of a decreasing function) *Let X be a discrete random variable with support R_X and probability mass function $p_X(x)$. Let $g : \mathbb{R} \to \mathbb{R}$ be strictly decreasing on the support of X. Then, the support of $Y = g(X)$ is*
$$R_Y = \{y = g(x) : x \in R_X\}$$
and its probability mass function is
$$p_Y(y) = \begin{cases} p_X\left(g^{-1}(y)\right) & \text{if } y \in R_Y \\ 0 & \text{if } y \notin R_Y \end{cases}$$

Proof. The proof of this proposition is identical to the proof of the proposition for strictly increasing functions. In fact, the only property that matters is that a strictly decreasing function is invertible:

$$\begin{aligned} p_Y(y) &= \mathrm{P}(Y = y) \\ &= \mathrm{P}(g(X) = y) \\ &= \mathrm{P}\left(X = g^{-1}(y)\right) \\ &= p_X\left(g^{-1}(y)\right) \end{aligned}$$

■

Example 192 *Let X be a discrete random variable with support*

$$R_X = \{1, 2, 3\}$$

and probability mass function

$$p_X(x) = \begin{cases} \frac{1}{14}x^2 & \text{if } x \in R_X \\ 0 & \text{if } x \notin R_X \end{cases}$$

Let

$$Y = g(X) = 1 - 2X$$

The support of Y is

$$R_Y = \{-5, -3, -1\}$$

The function g is strictly decreasing and its inverse is

$$g^{-1}(y) = \frac{1}{2} - \frac{1}{2}y$$

The probability mass function of Y is

$$p_Y(y) = \begin{cases} \frac{1}{14}\left(\frac{1}{2} - \frac{1}{2}y\right)^2 & \text{if } y \in R_Y \\ 0 & \text{if } y \notin R_Y \end{cases}$$

34.2.2 Strictly decreasing functions of a continuous variable

When X is an absolutely continuous random variable and g is differentiable, then also Y is absolutely continuous and its probability density function is derived as follows:

Proposition 193 (pdf of a decreasing function) *Let X be an absolutely continuous random variable with support R_X and probability density function $f_X(x)$. Let $g : \mathbb{R} \to \mathbb{R}$ be strictly decreasing and differentiable on the support of X. Then, the support of $Y = g(X)$ is*

$$R_Y = \{y = g(x) : x \in R_X\}$$

and its probability density function is

$$f_Y(y) = \begin{cases} -f_X\left(g^{-1}(y)\right)\frac{dg^{-1}(y)}{dy} & \text{if } y \in R_Y \\ 0 & \text{if } y \notin R_Y \end{cases}$$

Proof. This proposition is easily derived: 1) remembering that the probability that an absolutely continuous random variable takes on any specific value[3] is 0 and, as a consequence, $\mathrm{P}\left(X = g^{-1}(y)\right) = 0$ for any y; 2) using the fact that the density function is the first derivative of the distribution function; 3) differentiating the expression for the distribution function $F_Y(y)$ found above. ■

Example 194 *Let X be a uniform random variable[4] on the interval $[0,1]$, i.e. an absolutely continuous random variable with support*

$$R_X = [0,1]$$

and probability density function

$$f_X(x) = \begin{cases} 1 & \text{if } x \in R_X \\ 0 & \text{if } x \notin R_X \end{cases}$$

Let

$$Y = g(X) = -\frac{1}{\lambda}\ln(X)$$

where $\lambda \in \mathbb{R}_{++}$ is a constant. The support of Y is

$$R_Y = [0,\infty)$$

where we can safely ignore the fact that $g(0) = \infty$, because $\{X = 0\}$ is a zero-probability event[5]. The function g is strictly decreasing and its inverse is

$$g^{-1}(y) = \exp(-\lambda y)$$

with derivative

$$\frac{dg^{-1}(y)}{dy} = -\lambda\exp(-\lambda y)$$

The probability density function of Y is

$$f_Y(y) = \begin{cases} \lambda\exp(-\lambda y) & \text{if } y \in R_Y \\ 0 & \text{if } y \notin R_Y \end{cases}$$

Therefore, Y has an exponential distribution with parameter λ (see the lecture entitled Exponential distribution - p. 365).

34.3 Invertible functions

In the case in which the function $g(x)$ is neither strictly increasing nor strictly decreasing, the formulae given in the previous sections for discrete and absolutely continuous random variables are still applicable, provided $g(x)$ is one-to-one and hence invertible. We report these formulae below.

[3] See p. 109.
[4] See p. 359.
[5] See p. 109.

34.3.1 One-to-one functions of a discrete variable

When X is a discrete random variable the probability mass function of $Y = g(X)$ is given by the following:

Proposition 195 (pmf of a one-to-one function) *Let X be a discrete random variable with support R_X and probability mass function $p_X(x)$. Let $g : \mathbb{R} \to \mathbb{R}$ be one-to-one on the support of X. Then, the support of $Y = g(X)$ is*

$$R_Y = \{y = g(x) : x \in R_X\}$$

and its probability mass function is

$$p_Y(y) = \begin{cases} p_X\left(g^{-1}(y)\right) & \text{if } y \in R_Y \\ 0 & \text{if } y \notin R_Y \end{cases}$$

Proof. The proof of this proposition is identical to the proof of the propositions for strictly increasing and stricly decreasing functions found above:

$$\begin{aligned} p_Y(y) &= \mathrm{P}(Y = y) \\ &= \mathrm{P}(g(X) = y) \\ &= \mathrm{P}\left(X = g^{-1}(y)\right) \\ &= p_X\left(g^{-1}(y)\right) \end{aligned}$$

■

34.3.2 One-to-one functions of a continuous variable

When X is an absolutely continuous random variable and g is differentiable, then also Y is absolutely continuous and its probability density function is given by the following:

Proposition 196 (pdf of a one-to-one function) *Let X be an absolutely continuous random variable with support R_X and probability density function $f_X(x)$. Let $g : \mathbb{R} \to \mathbb{R}$ be one-to-one and differentiable on the support of X. Then, the support of $Y = g(X)$ is*

$$R_Y = \{y = g(x) : x \in R_X\}$$

If

$$\frac{d}{dy} g^{-1}(y) \neq 0 \quad , \forall y \in R_Y$$

then the probability density function of Y is

$$f_Y(y) = \begin{cases} f_X\left(g^{-1}(y)\right) \left|\frac{dg^{-1}(y)}{dy}\right| & \text{if } y \in R_Y \\ 0 & \text{if } y \notin R_Y \end{cases}$$

Proof. For a proof of this proposition see: Poirier[6] (1995). ■

[6] Poirier, D. J. (1995) *Intermediate statistics and econometrics: a comparative approach*, MIT Press.

34.4 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Let X be an absolutely continuous random variable with support

$$R_X = [0, 2]$$

and probability density function:

$$f_X(x) = \begin{cases} \frac{3}{8}x^2 & \text{if } x \in R_X \\ 0 & \text{if } x \notin R_X \end{cases}$$

Let

$$Y = g(X) = \sqrt{X+1}$$

Find the probability density function of Y.

Solution

The support of Y is

$$R_Y = \left[\sqrt{1}, \sqrt{3}\right]$$

The function g is strictly increasing and its inverse is

$$g^{-1}(y) = y^2 - 1$$

with derivative

$$\frac{dg^{-1}(y)}{dy} = 2y$$

The probability density function of Y is

$$f_Y(y) = \begin{cases} \frac{3}{8}\left(y^2-1\right)^2 2y & \text{if } y \in R_Y \\ 0 & \text{if } y \notin R_Y \end{cases}$$

Exercise 2

Let X be an absolutely continuous random variable with support

$$R_X = [0, 2]$$

and probability density function

$$f_X(x) = \begin{cases} \frac{1}{2} & \text{if } x \in R_X \\ 0 & \text{if } x \notin R_X \end{cases}$$

Let

$$Y = g(X) = -X^2$$

Find the probability density function of Y.

Solution

The support of Y is:

$$R_Y = [-4, 0]$$

The function g is strictly decreasing and its inverse is

$$g^{-1}(y) = (-y)^{1/2}$$

with derivative

$$\frac{dg^{-1}(y)}{dy} = -\frac{1}{2}(-y)^{-1/2}$$

The probability density function of Y is

$$f_Y(y) = \begin{cases} \frac{1}{2}\frac{1}{2}(-y)^{-1/2} = \frac{1}{4}(-y)^{-1/2} & \text{if } y \in R_Y \\ 0 & \text{if } y \notin R_Y \end{cases}$$

Exercise 3

Let X be a discrete random variable with support

$$R_X = \{1, 2, 3, 4\}$$

and probability mass function

$$p_X(x) = \begin{cases} \frac{1}{30}x^2 & \text{if } x \in R_X \\ 0 & \text{if } x \notin R_X \end{cases}$$

Let

$$Y = g(X) = X - 1$$

Find the probability mass function of Y.

Solution

The support of Y is

$$R_Y = \{0, 1, 2, 3\}$$

The function g is strictly increasing and its inverse is

$$g^{-1}(y) = y + 1$$

The probability mass function of Y is

$$p_Y(y) = \begin{cases} \frac{1}{30}(y+1)^2 & \text{if } y \in R_Y \\ 0 & \text{if } y \notin R_Y \end{cases}$$

Chapter 35

Functions of random vectors and their distribution

Let X be a $K \times 1$ random vector with known distribution, and let a $L \times 1$ random vector Y be a function of X:

$$Y = g(X)$$

where $g : \mathbb{R}^K \to \mathbb{R}^L$. How do we derive the distribution of Y from the distribution of X?

Although there is no general answer to this question, there are some special cases in which the distribution of Y can be easily derived from the distribution of X. This lecture discusses some of these special cases.

35.1 One-to-one functions

When the function $g(x)$ is one-to-one and hence invertible, and the random vector X is either discrete or absolutely continuous, there are readily applicable formulae for the distribution of Y, which we report below.

35.1.1 One-to-one function of a discrete vector

When X is a discrete random vector, the joint probability mass function[1] of $Y = g(X)$ is given by the following proposition.

Proposition 197 *Let X be a $K \times 1$ discrete random vector with support R_X and joint probability mass function $p_X(x)$. Let $g : \mathbb{R}^K \to \mathbb{R}^K$ be one-to-one on the support of X. Then, the random vector $Y = g(X)$ has support*

$$R_Y = \{y = g(x) : x \in R_X\}$$

and probability mass function

$$p_Y(y) = \begin{cases} p_X\left(g^{-1}(y)\right) & \text{if } y \in R_Y \\ 0 & \text{if } y \notin R_Y \end{cases}$$

[1] See p. 116.

Proof. If $y \in R_Y$, then

$$p_Y(y) = \mathrm{P}(Y = y) = \mathrm{P}(g(X) = y) = \mathrm{P}\left(X = g^{-1}(y)\right) = p_X\left(g^{-1}(y)\right)$$

where we have used the fact that g is one-to-one on the support of Y, and hence it possesses an inverse $g^{-1}(y)$. If $y \notin R_Y$, then, trivially, $p_Y(y) = 0$. ■

Example 198 *Let X be a 2×1 discrete random vector and denote its components by X_1 and X_2. Let the support of X be*

$$R_X = \left\{[1 \quad 1]^\top, [2 \quad 0]^\top\right\}$$

and its joint probability mass function be

$$p_X(x) = \begin{cases} 1/3 & \text{if } x = [1 \quad 1]^\top \\ 2/3 & \text{if } x = [2 \quad 0]^\top \\ 0 & \text{otherwise} \end{cases}$$

Let

$$Y = g(X) = 2X$$

The support of Y is

$$R_Y = \left\{[2 \quad 2]^\top, [4 \quad 0]^\top\right\}$$

The inverse function is

$$x = g^{-1}(y) = \frac{1}{2}y$$

The joint probability mass function of Y is

$$\begin{aligned} p_Y(y) &= \begin{cases} p_X\left(\frac{1}{2}y\right) & \text{if } y = [2 \quad 2]^\top \\ p_X\left(\frac{1}{2}y\right) & \text{if } y = [4 \quad 0]^\top \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} p_X(1,1) & \text{if } y = [2 \quad 2]^\top \\ p_X(2,0) & \text{if } y = [4 \quad 0]^\top \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} 1/3 & \text{if } y = [2 \quad 2]^\top \\ 2/3 & \text{if } y = [4 \quad 0]^\top \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

35.1.2 One-to-one function of a continuous vector

When X is an absolutely continuous random vector and g is differentiable, then also Y is absolutely continuous, and its joint probability density function[2] is given by the following proposition.

Proposition 199 *Let X be a $K \times 1$ absolutely continuous random vector with support R_X and joint probability density function $f_X(x)$. Let $g : \mathbb{R}^K \to \mathbb{R}^K$ be*

[2] See p. 117.

one-to-one and differentiable on the support of X. Denote by $J_{g^{-1}}(y)$ the Jacobian matrix of $g^{-1}(y)$, i.e.,

$$J_{g^{-1}}(y) = \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_K} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_K} \\ \vdots & & \ddots & \vdots \\ \frac{\partial x_K}{\partial y_1} & \frac{\partial x_K}{\partial y_2} & \cdots & \frac{\partial x_K}{\partial y_K} \end{bmatrix}$$

where y_i is the i-th component of y, and x_i is the i-th component of $x = g^{-1}(y)$. Then, the support of $Y = g(X)$ is

$$R_Y = \{y = g(x) : x \in R_X\}$$

If the determinant of the Jacobian matrix satisfies

$$\det\left(J_{g^{-1}}(y)\right) \neq 0 \quad , \forall y \in R_Y$$

then the joint probability density function of Y is

$$f_Y(y) = \begin{cases} f_X\left(g^{-1}(y)\right)\left|\det\left(J_{g^{-1}}(y)\right)\right| & \text{if } y \in R_Y \\ 0 & \text{if } y \notin R_Y \end{cases}$$

Proof. See e.g. Poirier[3] (1995). ■

A special case of the above proposition obtains when the function g is a linear one-to-one mapping.

Proposition 200 *Let X be a $K \times 1$ absolutely continuous random vector with joint probability density $f_X(x)$. Let Y be a $K \times 1$ random vector such that*

$$Y = \mu + \Sigma X$$

where μ is a constant $K \times 1$ vector and Σ is a constant $K \times K$ invertible matrix. Then, Y is an absolutely continuous random vector whose probability density function $f_Y(y)$ satisfies

$$f_Y(y) = \begin{cases} \frac{1}{|\det(\Sigma)|} f_X\left(\Sigma^{-1}(y-\mu)\right) & \text{if } y \in R_Y \\ 0 & \text{if } y \notin R_Y \end{cases}$$

where $\det(\Sigma)$ is the determinant of Σ.

Proof. In this case the inverse function is

$$g^{-1}(y) = \Sigma^{-1}(y - \mu)$$

The Jacobian matrix is

$$J_{g^{-1}}(y) = \Sigma^{-1}$$

When $y \in R_Y$, the joint density of Y is

$$\begin{aligned} f_X\left(g^{-1}(y)\right)\left|\det\left(J_{g^{-1}}(y)\right)\right| &= f_X\left(\Sigma^{-1}(y-\mu)\right)\left|\det\left(\Sigma^{-1}\right)\right| \\ &= f_X\left(\Sigma^{-1}(y-\mu)\right)\left|\det(\Sigma)\right|^{-1} \end{aligned}$$

■

[3] Poirier, D. J. (1995) *Intermediate statistics and econometrics: a comparative approach*, MIT Press.

Example 201 *Let X be a 2×1 random vector with support*

$$R_X = [1,2] \times [0,\infty)$$

and joint probability density function

$$f_X(x) = \begin{cases} x_1 \exp(-x_1 x_2) & \text{if } x_1 \in [1,2] \text{ and } x_2 \in [0,\infty) \\ 0 & \text{otherwise} \end{cases}$$

where x_1 and x_2 are the two components of x. Define a 2×1 random vector $Y = g(X)$ with components Y_1 and Y_2 as follows:

$$\begin{aligned} Y_1 &= 3X_1 \\ Y_2 &= -X_2 \end{aligned}$$

The inverse function $g^{-1}(y)$ is defined by

$$\begin{aligned} x_1 &= y_1/3 \\ x_2 &= -y_2 \end{aligned}$$

The Jacobian matrix of $g^{-1}(y)$ is

$$J_{g^{-1}}(y) = \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{bmatrix} = \begin{bmatrix} 1/3 & 0 \\ 0 & -1 \end{bmatrix}$$

Its determinant is

$$\det\left(J_{g^{-1}}(y)\right) = \frac{1}{3} \cdot (-1) - 0 \cdot 0 = -\frac{1}{3}$$

The support of Y is

$$\begin{aligned} R_Y &= \left\{ \begin{bmatrix} y_1 & y_2 \end{bmatrix}^\top : y_1 = 3x_1, y_2 = -x_2, x_1 \in [1,2], x_2 \in [0,\infty) \right\} \\ &= [3,6] \times (-\infty, 0] \end{aligned}$$

For $y \in R_Y$, the joint probability density function of Y is

$$\begin{aligned} f_Y(y) &= f_X\left(g^{-1}(y)\right) \left|\det\left(J_{g^{-1}}(y)\right)\right| \\ &= (y_1/3) \exp\left(-(y_1/3)(-y_2)\right) \left|-\frac{1}{3}\right| \\ &= \frac{1}{9} y_1 \exp(y_1 y_2/3) \end{aligned}$$

while, for $y \notin R_Y$, the joint probability density function is $f_Y(y) = 0$.

35.2 Independent sums

When the components of X are independent and

$$g(x) = x_1 + \ldots + x_K$$

then the distribution of $Y = g(X)$ can be derived by using the convolution formulae illustrated in the lecture entitled *Sums of independent random variables* (p. 323).

35.3 Known moment generating function

The joint moment generating function[4] of $Y = g(X)$, provided it exists, can be computed as

$$M_Y(t) = \mathrm{E}\left[\exp\left(t^\top Y\right)\right] = \mathrm{E}\left[\exp\left(t^\top g(X)\right)\right]$$

by using the transformation theorem[5]. If $M_Y(t)$ is recognized as the joint moment generating function of a known distribution, then such a distribution is the distribution of Y, because two random vectors have the same distribution if and only if they have the same joint moment generating function, provided the latter exists.

35.4 Known characteristic function

The joint characteristic function[6] of $Y = g(X)$ can be computed as

$$\varphi_Y(t) = \mathrm{E}\left[\exp\left(it^\top Y\right)\right] = \mathrm{E}\left[\exp\left(it^\top g(X)\right)\right]$$

by using the transformation theorem. If $\varphi_Y(t)$ is recognized as the joint characteristic function of a known distribution, then such a distribution is the distribution of Y, because two random vectors have the same distribution if and only if they have the same joint characteristic function.

35.5 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Let X_1 be a uniform random variable[7] with support

$$R_{X_1} = [1, 2]$$

and probability density function

$$f_{X_1}(x_1) = \begin{cases} 1 & \text{if } x_1 \in R_{X_1} \\ 0 & \text{if } x_1 \notin R_{X_1} \end{cases}$$

Let X_2 be an absolutely continuous random variable, independent of X_1, with support

$$R_{X_2} = [0, 2]$$

and probability density function

$$f_{X_2}(x_2) = \begin{cases} \frac{3}{8}x_2^2 & \text{if } x_2 \in R_{X_2} \\ 0 & \text{if } x_2 \notin R_{X_2} \end{cases}$$

Let

$$Y_1 = X_1^2$$

[4] See p. 297.
[5] See p. 134.
[6] See p. 315.
[7] See p. 359.

$$Y_2 = X_1 + X_2$$

Find the joint probability density function of the random vector

$$Y = [Y_1 \;\; Y_2]^\top$$

Solution

Since X_1 and X_2 are independent, their joint probability density function is equal to the product of their marginal density functions:

$$f_X(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2) = \begin{cases} \frac{3}{8}x_2^2 & \text{if } x_1 \in [1,2] \text{ and } x_2 \in [0,2] \\ 0 & \text{otherwise} \end{cases}$$

The support of Y is

$$\begin{aligned} R_Y &= \left\{ [y_1 \;\; y_2]^\top : y_1 = x_1^2, y_2 = x_1 + x_2, x_1 \in [1,2], x_2 \in [0,2] \right\} \\ &= \left\{ [y_1 \;\; y_2]^\top : y_1 \in [1,4], y_2 \in [\sqrt{y_1}, 4] \right\} \end{aligned}$$

The function $y = g(x)$ is one-to-one on R_Y and its inverse $g^{-1}(y)$ is defined by

$$\begin{aligned} x_1 &= \sqrt{y_1} \\ x_2 &= y_2 - \sqrt{y_1} \end{aligned}$$

with Jacobian matrix

$$J_{g^{-1}}(y) = \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{bmatrix} = \begin{bmatrix} \frac{1}{2}y_1^{-1/2} & 0 \\ -\frac{1}{2}y_1^{-1/2} & 1 \end{bmatrix}$$

The determinant of the Jacobian matrix is

$$\det\left(J_{g^{-1}}(y)\right) = \frac{1}{2}y_1^{-1/2} \cdot 1 - 0 \cdot \left(-\frac{1}{2}y_1^{-1/2}\right) = \frac{1}{2}y_1^{-1/2}$$

which is different from zero for any y belonging to R_Y. For $y \in R_Y$, the joint probability density function of Y is

$$\begin{aligned} f_Y(y) &= f_X\left(g^{-1}(y)\right) \left|\det\left(J_{g^{-1}}(y)\right)\right| = f_X\left(\sqrt{y_1}, y_2 - \sqrt{y_1}\right) \frac{1}{2}y_1^{-1/2} \\ &= \frac{3}{8}\left(y_2 - \sqrt{y_1}\right)^2 \frac{1}{2}y_1^{-1/2} = \frac{3}{16}y_1^{-1/2}\left(y_2 - \sqrt{y_1}\right)^2 \end{aligned}$$

while, for $y \notin R_Y$, the joint probability density function is $f_Y(y) = 0$.

Exercise 2

Let X be a 2×1 random vector with support

$$R_X = [0, \infty) \times [0, \infty)$$

and joint probability density function

$$f_X(x) = \begin{cases} \exp(-x_1 - x_2) & \text{if } x_1 \in [0, \infty) \text{ and } x_2 \in [0, \infty) \\ 0 & \text{otherwise} \end{cases}$$

where x_1 and x_2 are the two components of x. Define a 2×1 random vector $Y = g(X)$ with components Y_1 and Y_2 as follows:

$$\begin{aligned} Y_1 &= 2X_1 \\ Y_2 &= X_1 + X_2 \end{aligned}$$

Find the joint probability density function of the random vector Y.

Solution

The inverse function $g^{-1}(y)$ is defined by

$$\begin{aligned} x_1 &= y_1/2 \\ x_2 &= y_2 - y_1/2 \end{aligned}$$

The Jacobian matrix of $g^{-1}(y)$ is

$$J_{g^{-1}}(y) = \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{bmatrix} = \begin{bmatrix} 1/2 & 0 \\ -1/2 & 1 \end{bmatrix}$$

Its determinant is

$$\det\left(J_{g^{-1}}(y)\right) = \frac{1}{2} \cdot 1 - 0 \cdot \left(-\frac{1}{2}\right) = \frac{1}{2}$$

The support of Y is

$$\begin{aligned} R_Y &= \left\{ \begin{bmatrix} y_1 & y_2 \end{bmatrix}^\top : y_1 = 2x_1, y_2 = x_1 + x_2, x_1 \in [0,\infty), x_2 \in [0,\infty) \right\} \\ &= \left\{ \begin{bmatrix} y_1 & y_2 \end{bmatrix}^\top : y_1 \in [0,\infty), y_2 \in [y_1/2,\infty) \right\} \end{aligned}$$

For $y \in R_Y$, the joint probability density function of Y is

$$\begin{aligned} f_Y(y) &= f_X\left(g^{-1}(y)\right) \left|\det\left(J_{g^{-1}}(y)\right)\right| \\ &= f_X(y_1/2, y_2 - y_1/2) \cdot \frac{1}{2} \\ &= \frac{1}{2} \exp(-y_1/2 - y_2 + y_1/2) = \frac{1}{2} \exp(-y_2) \end{aligned}$$

while, for $y \notin R_Y$, the joint probability density function is $f_Y(y) = 0$.

Chapter 36

Moments and cross-moments

This lecture introduces the notions of moment of a random variable and cross-moment of a random vector.

36.1 Moments

36.1.1 Definition of moment

The n-th moment of a random variable is the expected value of its n-th power:

Definition 202 *Let X be a random variable. Let $n \in \mathbb{N}$. If*

$$\mu_X(n) = \mathrm{E}[X^n]$$

*exists and is finite, then X is said to possess a **finite n-th moment** and $\mu_X(n)$ is called the **n-th moment of** X. If $\mathrm{E}[X^n]$ is not well-defined, then we say that X does not possess the n-th moment.*

36.1.2 Definition of central moment

The n-th central moment of a random variable X is the expected value of the n-th power of the deviation of X from its expected value:

Definition 203 *Let X be a random variable. Let $n \in \mathbb{N}$. If*

$$\overline{\mu}_X(n) = \mathrm{E}[(X - \mathrm{E}[X])^n]$$

*exists and is finite, then X is said to possess a **finite n-th central moment** and $\overline{\mu}_X(n)$ is called the **n-th central moment of** X.*

36.2 Cross-moments

36.2.1 Definition of cross-moment

Let X be a $K \times 1$ random vector. A cross-moment of X is the expected value of the product of integer powers of the entries of X:

$$\mathrm{E}[X_1^{n_1} \cdot X_2^{n_2} \cdot \ldots \cdot X_K^{n_K}]$$

where X_i is the i-th entry of X and $n_1, n_2, \ldots, n_K \in \mathbb{Z}_+$ are non-negative integers.

The following is a formal definition of cross-moment:

Definition 204 *Let X be a $K \times 1$ random vector. Let n_1, n_2, ..., n_K be K non-negative integers and*

$$n = \sum_{k=1}^{K} n_k \tag{36.1}$$

If

$$\mu_X(n_1, n_2, \ldots, n_K) = \mathrm{E}\left[X_1^{n_1} \cdot X_2^{n_2} \cdot \ldots \cdot X_K^{n_K}\right] \tag{36.2}$$

exists and is finite, then it is called a ***cross-moment*** *of X of order n. If all cross-moments of order n exist and are finite, i.e. if (36.2) exists and is finite for all K-tuples of non-negative integers n_1, n_2, ..., n_K such that condition (36.1) is satisfied, then X is said to possess* ***finite cross-moments*** *of order n.*

The following example shows how to compute a cross-moment of a discrete random vector:

Example 205 *Let X be a 3×1 discrete random vector and denote its components by X_1, X_2 and X_3. Let the support of X be:*

$$R_X = \left\{ \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}, \begin{bmatrix} 3 \\ 3 \\ 2 \end{bmatrix} \right\}$$

and its joint probability mass function[1] be:

$$p_X(x) = \begin{cases} \frac{1}{3} & \text{if } x = \begin{bmatrix} 1 & 2 & 1 \end{bmatrix}^\top \\ \frac{1}{3} & \text{if } x = \begin{bmatrix} 2 & 1 & 3 \end{bmatrix}^\top \\ \frac{1}{3} & \text{if } x = \begin{bmatrix} 3 & 3 & 2 \end{bmatrix}^\top \\ 0 & \text{otherwise} \end{cases}$$

The following is a cross-moment of X of order 4:

$$\mu_X(1,2,1) = \mathrm{E}\left[X_1 \cdot X_2^2 \cdot X_3\right]$$

which can be computed using the transformation theorem[2]:

$$\begin{aligned} \mu_X(1,2,1) &= \mathrm{E}\left[X_1 \cdot X_2^2 \cdot X_3\right] \\ &= \sum_{(x_1,x_2,x_3) \in R_X} x_1 \cdot x_2^2 \cdot x_3 \cdot p_X(x_1, x_2, x_3) \\ &= 1 \cdot 2^2 \cdot 1 \cdot p_X(1,2,1) + 2 \cdot 1^2 \cdot 3 \cdot p_X(2,1,3) \\ &\quad +3 \cdot 3^2 \cdot 2 \cdot p_X(3,3,2) \\ &= 4 \cdot \frac{1}{3} + 6 \cdot \frac{1}{3} + 54 \cdot \frac{1}{3} = \frac{64}{3} \end{aligned}$$

[1] See p. 117.
[2] See p. 134.

36.2.2 Definition of central cross-moment

The central cross-moments of a random vector X are just the cross-moments of the random vector of deviations $X - \mathrm{E}[X]$:

Definition 206 *Let X be a $K \times 1$ random vector. Let n_1, n_2, ..., n_K be K non-negative integers and*

$$n = \sum_{k=1}^{K} n_k \tag{36.3}$$

If

$$\overline{\mu}_X(n_1, n_2, \ldots, n_K) = \mathrm{E}\left[\prod_{k=1}^{K} (X_k - \mathrm{E}[X_k])^{n_k}\right] \tag{36.4}$$

*exists and is finite, then it is called a **central cross-moment** of X of order n. If all central cross-moments of order n exist and are finite, i.e. if (36.4) exists and is finite for all K-tuples of non-negative integers n_1, n_2, ..., n_K such that condition (36.3) is satisfied, then X is said to possess **finite central cross-moments** of order n.*

The following example shows how to compute a central cross-moment of a discrete random vector:

Example 207 *Let X be a 3×1 discrete random vector and denote its components by X_1, X_2 and X_3. Let the support of X be:*

$$R_X = \left\{ \begin{bmatrix} 4 \\ 2 \\ 4 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix} \right\}$$

and its joint probability mass function be:

$$p_X(x) = \begin{cases} \frac{1}{5} & \text{if } x = \begin{bmatrix} 4 & 2 & 4 \end{bmatrix}^\top \\ \frac{2}{5} & \text{if } x = \begin{bmatrix} 2 & 1 & 1 \end{bmatrix}^\top \\ \frac{2}{5} & \text{if } x = \begin{bmatrix} 1 & 3 & 2 \end{bmatrix}^\top \\ 0 & \text{otherwise} \end{cases}$$

The expected values of the three components of X are:

$$\begin{aligned} \mathrm{E}[X_1] &= 4 \cdot \frac{1}{5} + 2 \cdot \frac{2}{5} + 1 \cdot \frac{2}{5} = 2 \\ \mathrm{E}[X_2] &= 2 \cdot \frac{1}{5} + 1 \cdot \frac{2}{5} + 3 \cdot \frac{2}{5} = 2 \\ \mathrm{E}[X_3] &= 4 \cdot \frac{1}{5} + 1 \cdot \frac{2}{5} + 2 \cdot \frac{2}{5} = 2 \end{aligned}$$

The following is a central cross-moment of X of order 3:

$$\mu_X(2,1,0) = \mathrm{E}\left[(X_1 - \mathrm{E}[X_1])^2 \cdot (X_2 - \mathrm{E}[X_2])\right]$$

which can be computed using the transformation theorem:

$$\mu_X(2,1,0) = \mathrm{E}\left[(X_1 - \mathrm{E}[X_1])^2 \cdot (X_2 - \mathrm{E}[X_2])\right]$$

$$
\begin{aligned}
&= \sum_{(x_1,x_2,x_3)\in R_X} (x_1-2)^2 \cdot (x_2-2) \cdot p_X(x_1,x_2,x_3) \\
&= (4-2)^2 \cdot (2-2) \cdot p_X(4,2,4) \\
&\quad + (2-2)^2 \cdot (1-2) \cdot p_X(2,1,1) \\
&\quad + (1-2)^2 \cdot (3-2) \cdot p_X(1,3,2) \\
&= (1-2)^2 \cdot (3-2) \cdot \frac{2}{5} = \frac{2}{5}
\end{aligned}
$$

Chapter 37

Moment generating function of a random variable

The distribution of a random variable is often characterized in terms of its moment generating function (mgf), a real function whose derivatives at zero are equal to the moments[1] of the random variable. Mgfs have great practical relevance not only because they can be used to easily derive moments, but also because a probability distribution is uniquely determined by its mgf, a fact that, coupled with the analytical tractability of mgfs, makes them a handy tool to solve several problems, such as deriving the distribution of a sum of two or more random variables.

It must be mentioned that not all random variables possess an mgf. However, all random variables possess a characteristic function[2], another transform that enjoys properties similar to those enjoyed by the mgf.

37.1 Definition

We start this lecture by giving a definition of mgf.

Definition 208 *Let X be a random variable. If the expected value*

$$\mathrm{E}\left[\exp\left(tX\right)\right]$$

exists and is finite for all real numbers t belonging to a closed interval $[-h,h] \subseteq \mathbb{R}$, with $h > 0$, then we say that X ***possesses a moment generating function*** *and the function $M_X : [-h,h] \to \mathbb{R}$ defined by*

$$M_X\left(t\right) = \mathrm{E}\left[\exp\left(tX\right)\right]$$

is called the ***moment generating function*** *of X.*

The following example shows how the mgf of an exponential random variable is derived.

[1] See p. 285.
[2] See p. 307.

Example 209 *Let X be an exponential random variable[3] with parameter $\lambda \in \mathbb{R}_{++}$. Its support is the set of positive real numbers*

$$R_X = [0, \infty)$$

and its probability density function is

$$f_X(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{if } x \in R_X \\ 0 & \text{if } x \notin R_X \end{cases}$$

Its mgf is computed as follows:

$$\begin{aligned} \mathrm{E}[\exp(tX)] &= \int_{-\infty}^{\infty} \exp(tx) f_X(x)\, dx \\ &= \int_{0}^{\infty} \exp(tx) \lambda \exp(-\lambda x)\, dx \\ \boxed{A} &= \lambda \int_{0}^{\infty} \exp((t-\lambda)x)\, dx \\ &= \lambda \left[\frac{1}{t-\lambda} \exp((t-\lambda)x) \right]_0^{\infty} \\ &= \lambda \left[0 - \frac{1}{t-\lambda} \right] = \frac{\lambda}{\lambda - t} \end{aligned}$$

where: in step $\boxed{A}$ we have assumed that $t < \lambda$, which is necessary for the integral to be finite. Therefore, the expected value exists and is finite for $t \in [-h, h]$ if h is such that $0 < h < \lambda$, and X possesses an mgf

$$M_X(t) = \frac{\lambda}{\lambda - t}$$

37.2 Moments and mgfs

The mgf takes its name by the fact that it can be used to derive the moments of X, as stated in the following proposition.

Proposition 210 *If a random variable X possesses an mgf $M_X(t)$, then, for any $n \in \mathbb{N}$, the n-th moment of X, denoted by $\mu_X(n)$, exists and is finite. Furthermore,*

$$\mu_X(n) = \mathrm{E}[X^n] = \left. \frac{d^n M_X(t)}{dt^n} \right|_{t=0}$$

where $\left. \frac{d^n M_X(t)}{dt^n} \right|_{t=0}$ is the n-th derivative of $M_X(t)$ with respect to t, evaluated at the point $t = 0$.

Proof. Proving the above proposition is quite complicated, because a lot of analytical details must be taken care of (see, e.g., Pfeiffer[4] - 1978). The intuition,

[3] See p. 365.

[4] Pfeiffer, P. E. (1978) *Concepts of probability theory*, Courier Dover Publications.

however, is straightforward: since the expected value is a linear operator and differentiation is a linear operation, under appropriate conditions we can differentiate through the expected value, as follows:

$$\frac{d^n M_X(t)}{dt^n} = \frac{d^n}{dt^n}\mathrm{E}\left[\exp(tX)\right] = \mathrm{E}\left[\frac{d^n}{dt^n}\exp(tX)\right] = \mathrm{E}\left[X^n \exp(tX)\right]$$

which, evaluated at the point $t = 0$, yields

$$\left.\frac{d^n M_X(t)}{dt^n}\right|_{t=0} = \mathrm{E}\left[X^n \exp(0 \cdot X)\right] = \mathrm{E}\left[X^n\right] = \mu_X(n)$$

■

The following example shows how this proposition can be applied.

Example 211 *In Example 209 we have demonstrated that the mgf of an exponential random variable is*

$$M_X(t) = \frac{\lambda}{\lambda - t}$$

The expected value of X can be computed by taking the first derivative of the mgf:

$$\frac{dM_X(t)}{dt} = \frac{\lambda}{(\lambda - t)^2}$$

and evaluating it at $t = 0$:

$$\mathrm{E}[X] = \left.\frac{dM_X(t)}{dt}\right|_{t=0} = \frac{\lambda}{(\lambda - 0)^2} = \frac{1}{\lambda}$$

The second moment of X can be computed by taking the second derivative of the mgf:

$$\frac{d^2 M_X(t)}{dt^2} = \frac{2\lambda}{(\lambda - t)^3}$$

and evaluating it at $t = 0$:

$$\mathrm{E}\left[X^2\right] = \left.\frac{d^2 M_X(t)}{dt^2}\right|_{t=0} = \frac{2\lambda}{(\lambda - 0)^3} = \frac{2}{\lambda^2}$$

And so on for the higher moments.

37.3 Distributions and mgfs

The following proposition states the most important property of the mgf.

Proposition 212 (equality of distributions) *Let X and Y be two random variables. Denote by $F_X(x)$ and $F_Y(y)$ their distribution functions*[5]*, and by $M_X(t)$ and $M_Y(t)$ their mgfs. X and Y have the same distribution, i.e., $F_X(x) = F_Y(x)$ for any x, if and only if they have the same mgfs, i.e., $M_X(t) = M_Y(t)$ for any t.*

[5] See p. 108.

Proof. For a fully general proof of this proposition see, e.g., Feller[6] (2008). We just give an informal proof for the special case in which X and Y are discrete random variables taking only finitely many values. The "only if" part is trivial. If X and Y have the same distribution, then

$$M_X(t) = \mathrm{E}[\exp(tX)] = \mathrm{E}[\exp(tY)] = M_Y(t)$$

The "if" part is proved as follows. Denote by R_X and R_Y the supports of X and Y, and by $p_X(x)$ and $p_Y(y)$ their probability mass functions[7]. Denote by A the union of the two supports:

$$A = R_X \cup R_Y$$

and by $a_1, \ldots, a_n$ the elements of A. The mgf of X can be written as

$$\begin{aligned} & M_X(t) \\ = \; & \mathrm{E}[\exp(tX)] \\ \boxed{\text{A}} \; = \; & \sum_{x \in R_X} \exp(tx)\, p_X(x) \\ \boxed{\text{B}} \; = \; & \sum_{i=1}^{n} \exp(ta_i)\, p_X(a_i) \end{aligned}$$

where: in step A we have used the definition of expected value; in step B we have used the fact that $p_X(a_i) = 0$ if $a_i \notin R_X$. By the same token, the mgf of Y can be written as

$$M_Y(t) = \sum_{i=1}^{n} \exp(ta_i)\, p_Y(a_i)$$

If X and Y have the same mgf, then, for any t belonging to a closed neighborhood of zero,

$$M_X(t) = M_Y(t)$$

and

$$\sum_{i=1}^{n} \exp(ta_i)\, p_X(a_i) = \sum_{i=1}^{n} \exp(ta_i)\, p_Y(a_i)$$

By rearranging terms, we obtain

$$\sum_{i=1}^{n} \exp(ta_i)\, [p_X(a_i) - p_Y(a_i)] = 0$$

This can be true for any t belonging to a closed neighborhood of zero only if

$$p_X(a_i) - p_Y(a_i) = 0$$

for every i. It follows that that the probability mass functions of X and Y are equal. As a consequence, also their distribution functions are equal. ■

It must be stressed that this proposition is extremely important and relevant from a practical viewpoint: in many cases where we need to prove that two distributions are equal, it is much easier to prove equality of the mgfs than to prove equality of the distribution functions.

[6] Feller, W. (2008) *An introduction to probability theory and its applications*, Volume 2, Wiley.
[7] See p. 106.

Also note that equality of the distribution functions can be replaced in the proposition above by equality of the probability mass functions[8] if X and Y are discrete random variables, or by equality of the probability density functions[9] if X and Y are absolutely continuous random variables.

37.4 More details

37.4.1 Mgf of a linear transformation

The next proposition gives a formula for the mgf of a linear transformation.

Proposition 213 *Let X be a random variable possessing an mgf $M_X(t)$. Define*

$$Y = a + bX$$

where $a, b \in \mathbb{R}$ are two constants and $b \neq 0$. Then, the random variable Y possesses an mgf $M_Y(t)$ and

$$M_Y(t) = \exp(at) M_X(bt)$$

Proof. Using the definition of mgf, we obtain

$$\begin{aligned} M_Y(t) &= \mathrm{E}[\exp(tY)] = \mathrm{E}[\exp(at + btX)] \\ &= \mathrm{E}[\exp(at)\exp(btX)] = \exp(at)\,\mathrm{E}[\exp(btX)] \\ &= \exp(at) M_X(bt) \end{aligned}$$

If $M_X(t)$ is defined on a closed interval $[-h, h]$, then $M_Y(t)$ is defined on the interval $\left[-\frac{h}{b}, \frac{h}{b}\right]$. ■

37.4.2 Mgf of a sum

The next proposition shows how to derive the mgf of a sum of independent random variables.

Proposition 214 *Let $X_1, \ldots, X_n$ be n mutually independent[10] random variables. Let Z be their sum:*

$$Z = \sum_{i=1}^{n} X_i$$

Then, the mgf of Z is the product of the mgfs of $X_1, \ldots, X_n$:

$$M_Z(t) = \prod_{i=1}^{n} M_{X_i}(t)$$

provided the latter exist.

Proof. This is proved as follows:

$$M_Z(t) = \mathrm{E}[\exp(tZ)]$$

[8] See p. 106.
[9] See p. 107.
[10] See p. 233.

$$\begin{aligned}
&= \mathrm{E}\left[\exp\left(t\sum_{i=1}^{n} X_i\right)\right] \\
&= \mathrm{E}\left[\exp\left(\sum_{i=1}^{n} tX_i\right)\right] \\
&= \mathrm{E}\left[\prod_{i=1}^{n} \exp\left(tX_i\right)\right] \\
\boxed{\text{A}} \quad &= \prod_{i=1}^{n} \mathrm{E}\left[\exp\left(tX_i\right)\right] \\
\boxed{\text{B}} \quad &= \prod_{i=1}^{n} M_{X_i}\left(t\right)
\end{aligned}$$

where: in step A we have used the properties of mutually independent variables[11]; in step B we have used the definition of mgf. ■

37.5 Solved exercises

Some solved exercises on mgfs can be found below.

Exercise 1

Let X be a discrete random variable having a Bernoulli distribution[12]. Its support is

$$R_X = \{0, 1\}$$

and its probability mass function[13] is

$$p_X\left(x\right) = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \\ 0 & \text{if } x \notin R_X \end{cases}$$

where $p \in (0, 1)$ is a constant. Derive the mgf of X, if it exists.

Solution

Using the definition of mgf, we get

$$\begin{aligned}
M_X\left(t\right) &= \mathrm{E}\left[\exp\left(tX\right)\right] = \sum_{x \in R_X} \exp\left(tx\right) p_X\left(x\right) \\
&= \exp\left(t \cdot 1\right) \cdot p_X\left(1\right) + \exp\left(t \cdot 0\right) \cdot p_X\left(0\right) \\
&= \exp\left(t\right) \cdot p + 1 \cdot \left(1 - p\right) = 1 - p + p\exp\left(t\right)
\end{aligned}$$

The mgf exists and it is well-defined because the above expected value exists for any $t \in \mathbb{R}$.

[11] See p. 234.
[12] See p. 335.
[13] See p. 106.

Exercise 2

Let X be a random variable with mgf

$$M_X(t) = \frac{1}{2}(1 + \exp(t))$$

Derive the variance of X.

Solution

We can use the following formula for computing the variance[14]:

$$\operatorname{Var}[X] = \operatorname{E}\left[X^2\right] - \operatorname{E}[X]^2$$

The expected value of X is computed by taking the first derivative of the mgf:

$$\frac{dM_X(t)}{dt} = \frac{1}{2}\exp(t)$$

and evaluating it at $t = 0$:

$$\operatorname{E}[X] = \left.\frac{dM_X(t)}{dt}\right|_{t=0} = \frac{1}{2}\exp(0) = \frac{1}{2}$$

The second moment of X is computed by taking the second derivative of the mgf:

$$\frac{d^2M_X(t)}{dt^2} = \frac{1}{2}\exp(t)$$

and evaluating it at $t = 0$:

$$\operatorname{E}\left[X^2\right] = \left.\frac{d^2M_X(t)}{dt^2}\right|_{t=0} = \frac{1}{2}\exp(0) = \frac{1}{2}$$

Therefore,

$$\begin{aligned}\operatorname{Var}[X] &= \operatorname{E}\left[X^2\right] - \operatorname{E}[X]^2 = \frac{1}{2} - \left(\frac{1}{2}\right)^2 \\ &= \frac{1}{2} - \frac{1}{4} = \frac{1}{4}\end{aligned}$$

Exercise 3

A random variable X is said to have a Chi-square distribution[15] with n degrees of freedom if its mgf is defined for any $t < \frac{1}{2}$ and it is equal to

$$M_X(t) = (1 - 2t)^{-n/2}$$

Define

$$Y = X_1 + X_2$$

where X_1 and X_2 are two independent random variables having Chi-square distributions with n_1 and n_2 degrees of freedom respectively. Prove that Y has a Chi-square distribution with $n_1 + n_2$ degrees of freedom.

[14] See p. 156.
[15] See p. 387.

Solution

The mgfs of X_1 and X_2 are

$$\begin{aligned} M_{X_1}(t) &= (1-2t)^{-n_1/2} \\ M_{X_2}(t) &= (1-2t)^{-n_2/2} \end{aligned}$$

The mgf of a sum of independent random variables is the product of the mgfs of the summands:

$$M_Y(t) = (1-2t)^{-n_1/2}(1-2t)^{-n_2/2} = (1-2t)^{-(n_1+n_2)/2}$$

Therefore, $M_Y(t)$ is the mgf of a Chi-square random variable with n_1+n_2 degrees of freedom. As a consequence, Y has a Chi-square distribution with n_1+n_2 degrees of freedom.

Chapter 38

Moment generating function of a random vector

The concept of joint moment generating function (joint mgf) is a multivariate generalization of the concept of moment generating function (mgf). Similarly to the univariate case, the joint mgf uniquely determines the joint distribution of its associated random vector, and it can be used to derive the cross-moments[1] of the distribution by partial differentiation.

If you are not familiar with the univariate concept, you are advised to first read the lecture entitled *Moment generating functions* (p. 289).

38.1 Definition

Let us start with a formal definition.

Definition 215 *Let X be a $K \times 1$ random vector. If the expected value*

$$\mathrm{E}\left[\exp\left(t^\top X\right)\right] = \mathrm{E}\left[\exp\left(t_1 X_1 + t_2 X_2 + \ldots + t_K X_K\right)\right]$$

exists and is finite for all $K \times 1$ real vectors t belonging to a closed rectangle H such that

$$H = [-h_1, h_1] \times [-h_2, h_2] \times \ldots \times [-h_K, h_K] \subseteq \mathbb{R}^K$$

*with $h_i > 0$ for all $i = 1, \ldots, K$, then we say that X **possesses a joint moment generating function** and the function $M_X : H \to \mathbb{R}$ defined by*

$$M_X(t) = \mathrm{E}\left[\exp\left(t^\top X\right)\right]$$

*is called the **joint moment generating function** of X.*

As an example, we derive the joint mgf of a standard multivariate normal random vector.

Example 216 *Let X be a $K \times 1$ standard multivariate normal random vector[2]. Its support is*

$$R_X = \mathbb{R}^K$$

[1] See p. 285.
[2] See p. 439.

and its joint probability density function[3] *is*

$$f_X(x) = (2\pi)^{-K/2} \exp\left(-\frac{1}{2}x^\intercal x\right)$$

As explained in the lecture entitled Multivariate normal distribution (p. 439), the K components of X are K mutually independent[4] *standard normal random variables, because the joint probability density function of X can be written as*

$$f_X(x) = f(x_1) \cdot f(x_2) \cdot \ldots \cdot f(x_K)$$

where x_i is the i-th entry of x, and $f(x_i)$ is the probability density function of a standard normal random variable:

$$f(x_i) = (2\pi)^{-1/2} \cdot \exp\left(-\frac{1}{2}x_i^2\right)$$

The joint mgf of X can be derived as follows:

$$\begin{aligned}
M_X(t) &= \mathrm{E}\left[\exp\left(t^\top X\right)\right] \\
&= \mathrm{E}\left[\exp\left(t_1X_1 + t_2X_2 + \ldots + t_KX_K\right)\right] \\
&= \mathrm{E}\left[\prod_{i=1}^{K} \exp\left(t_iX_i\right)\right] \\
\boxed{A} &= \prod_{i=1}^{K} \mathrm{E}\left[\exp\left(t_iX_i\right)\right] \\
\boxed{B} &= \prod_{i=1}^{K} M_{X_i}(t_i)
\end{aligned}$$

where: in step $\boxed{A}$ we have used the fact that the entries of X are mutually independent[5]*; in step $\boxed{B}$ we have used the definition of mgf of a random variable*[6]*. Since the mgf of a standard normal random variable is*[7]

$$M_{X_i}(t_i) = \exp\left(\frac{1}{2}t_i^2\right)$$

the joint mgf of X is

$$\begin{aligned}
M_X(t) &= \prod_{i=1}^{K} M_{X_i}(t_i) = \prod_{i=1}^{K} \exp\left(\frac{1}{2}t_i^2\right) \\
&= \exp\left(\frac{1}{2}\sum_{i=1}^{K} t_i^2\right) = \exp\left(\frac{1}{2}t^\top t\right)
\end{aligned}$$

Note that the mgf $M_{X_i}(t_i)$ of a standard normal random variable is defined for any $t_i \in \mathbb{R}$. As a consequence, the joint mgf of X is defined for any $t \in \mathbb{R}^K$.

[3] See p. 117.
[4] See p. 233.
[5] See p. 234.
[6] See p. 289.
[7] See p. 378.

38.2 Cross-moments and joint mgfs

The next proposition shows how the joint mgf can be used to derive the cross-moments of a random vector.

Proposition 217 *If a $K \times 1$ random vector X possesses a joint mgf $M_X(t)$, then it possesses finite cross-moments of order n for any $n \in \mathbb{N}$. Furthermore, if you define a cross-moment of order n as*

$$\mu_X(n_1, n_2, \ldots, n_K) = \mathrm{E}[X_1^{n_1} \cdot X_2^{n_2} \cdot \ldots \cdot X_K^{n_K}]$$

where $n_1, n_2, \ldots, n_K \in \mathbb{Z}_+$ and $n = \sum_{k=1}^{K} n_k$, then

$$\mu_X(n_1, n_2, \ldots, n_K) = \left. \frac{\partial^{n_1+n_2+\ldots+n_K} M_X(t_1, t_2, \ldots, t_K)}{\partial t_1^{n_1} \partial t_2^{n_2} \ldots \partial t_K^{n_K}} \right|_{t_1=0, t_2=0, \ldots, t_K=0}$$

where the derivative on the right-hand side is an n-th order cross-partial derivative of $M_X(t)$ evaluated at the point $t_1 = 0$, $t_2 = 0$, ..., $t_K = 0$.

Proof. We do not provide a rigorous proof of this proposition, but see, e.g., Pfeiffer[8] (1978) and DasGupta[9] (2010). The intuition of the proof, however, is straightforward: since the expected value is a linear operator and differentiation is a linear operation, under appropriate conditions one can differentiate through the expected value, as follows:

$$\begin{aligned}
& \frac{\partial^{n_1+n_2+\ldots+n_K} M_X(t_1, t_2, \ldots, t_K)}{\partial t_1^{n_1} \partial t_2^{n_2} \ldots \partial t_K^{n_K}} \\
= & \frac{\partial^{n_1+n_2+\ldots+n_K}}{\partial t_1^{n_1} \partial t_2^{n_2} \ldots \partial t_K^{n_K}} \mathrm{E}[\exp(t_1 X_1 + t_2 X_2 + \ldots + t_K X_K)] \\
= & \mathrm{E}\left[\frac{\partial^{n_1+n_2+\ldots+n_K}}{\partial t_1^{n_1} \partial t_2^{n_2} \ldots \partial t_K^{n_K}} \exp(t_1 X_1 + t_2 X_2 + \ldots + t_K X_K) \right] \\
= & \mathrm{E}[X_1^{n_1} \cdot X_2^{n_2} \cdot \ldots \cdot X_K^{n_K} \exp(t_1 X_1 + t_2 X_2 + \ldots + t_K X_K)]
\end{aligned}$$

which, evaluated at the point $t_1 = 0, t_2 = 0, \ldots, t_K = 0$, yields

$$\begin{aligned}
& \left. \frac{\partial^{n_1+n_2+\ldots+n_K} M_X(t_1, t_2, \ldots, t_K)}{\partial t_1^{n_1} \partial t_2^{n_2} \ldots \partial t_K^{n_K}} \right|_{t_1=0, t_2=0, \ldots, t_K=0} \\
= & \mathrm{E}[X_1^{n_1} \cdot X_2^{n_2} \cdot \ldots \cdot X_K^{n_K} \exp(0 \cdot X_1 + 0 \cdot X_2 + \ldots + 0 \cdot X_K)] \\
= & \mathrm{E}[X_1^{n_1} \cdot X_2^{n_2} \cdot \ldots \cdot X_K^{n_K}] \\
= & \mu_X(n_1, n_2, \ldots, n_K)
\end{aligned}$$

■

The following example shows how the above proposition can be applied.

Example 218 *Let us continue with the previous example. The joint mgf of a 2×1 standard normal random vector X is*

$$M_X(t) = \exp\left(\frac{1}{2} t^\top t\right) = \exp\left(\frac{1}{2} t_1^2 + \frac{1}{2} t_2^2\right)$$

[8] Pfeiffer, P. E. (1978) *Concepts of probability theory*, Courier Dover Publications.
[9] DasGupta, A. (2010) *Fundamentals of probability: a first course*, Springer.

The second cross-moment of X can be computed by taking the second cross-partial derivative of the joint mgf:

$$\begin{aligned}\mu_X(1,1) &= \mathrm{E}[X_1 \cdot X_2] \\ &= \frac{\partial^2}{\partial t_1 \partial t_2} \exp\left(\frac{1}{2}t_1^2 + \frac{1}{2}t_2^2\right)\Big|_{t_1=0,t_2=0} \\ &= \frac{\partial}{\partial t_1}\left(\frac{\partial}{\partial t_2} \exp\left(\frac{1}{2}t_1^2 + \frac{1}{2}t_2^2\right)\right)\Big|_{t_1=0,t_2=0} \\ &= \frac{\partial}{\partial t_1}\left(t_2 \exp\left(\frac{1}{2}t_1^2 + \frac{1}{2}t_2^2\right)\right)\Big|_{t_1=0,t_2=0} \\ &= t_1 t_2 \exp\left(\frac{1}{2}t_1^2 + \frac{1}{2}t_2^2\right)\Big|_{t_1=0,t_2=0} = 0\end{aligned}$$

38.3 Joint distributions and joint mgfs

One of the most important properties of the joint mgf is that it completely characterizes the joint distribution of a random vector.

Proposition 219 (equality of distributions) *Let X and Y be two $K \times 1$ random vectors, possessing joint mgfs $M_X(t)$ and $M_Y(t)$. Denote by $F_X(x)$ and $F_Y(y)$ their joint distribution functions*[10]*. X and Y have the same distribution, i.e., $F_X(x) = F_Y(x)$ for any $x \in \mathbb{R}^K$, if and only if they have the same mgfs, i.e., $M_X(t) = M_Y(t)$ for any $t \in H \subseteq \mathbb{R}^K$.*

Proof. The reader may refer to Feller[11] (2008) for a rigorous proof. The informal proof given here is almost identical to that given for the univariate case[12]. We confine our attention to the case in which X and Y are discrete random vectors taking only finitely many values. As far as the left-to-right direction of the implication is concerned, it suffices to note that if X and Y have the same distribution, then

$$M_X(t) = \mathrm{E}\left[\exp\left(t^\top X\right)\right] = \mathrm{E}\left[\exp\left(t^\top Y\right)\right] = M_Y(t)$$

The right-to-left direction of the implication is proved as follows. Denote by R_X and R_Y the supports of X and Y, and by $p_X(x)$ and $p_Y(y)$ their joint probability mass functions[13]. Define the union of the two supports:

$$A = R_X \cup R_Y$$

and denote its members by $a_1, \ldots, a_n$. The joint mgf of X can be written as

$$\begin{aligned}M_X(t) &= \mathrm{E}\left[\exp\left(t^\top X\right)\right] \\ \boxed{\text{A}} &= \sum_{x \in R_X} \exp\left(t^\top x\right) p_X(x) \\ \boxed{\text{B}} &= \sum_{i=1}^{n} \exp\left(t^\top a_i\right) p_X(a_i)\end{aligned}$$

[10] See p. 118.
[11] Feller, W. (2008) *An introduction to probability theory and its applications*, Volume 2, Wiley.
[12] See p. 291.
[13] See p. 116.

where: in step A we have used the definition of expected value; in step B we have used the fact that $p_X(a_i) = 0$ if $a_i \notin R_X$. By the same line of reasoning, the joint mgf of Y can be written as

$$M_Y(t) = \sum_{i=1}^{n} \exp\left(t^\top a_i\right) p_Y(a_i)$$

If X and Y have the same joint mgf, then

$$M_X(t) = M_Y(t)$$

for any t belonging to a closed rectangle where the two mgfs are well-defined, and

$$\sum_{i=1}^{n} \exp\left(t^\top a_i\right) p_X(a_i) = \sum_{i=1}^{n} \exp\left(t^\top a_i\right) p_Y(a_i)$$

By rearranging terms, we obtain

$$\sum_{i=1}^{n} \exp\left(t^\top a_i\right) \left[p_X(a_i) - p_Y(a_i)\right] = 0$$

This equality can be verified for every t only if

$$p_X(a_i) - p_Y(a_i) = 0$$

for every i. As a consequence, the joint probability mass functions of X and Y are equal, which implies that also their joint distribution functions are equal. ■

This proposition is used very often in applications where one needs to demonstrate that two joint distributions are equal. In such applications, proving equality of the joint mgfs is often much easier than proving equality of the joint distribution functions (see also the comments to Proposition 212).

38.4 More details

38.4.1 Joint mgf of a linear transformation

The next proposition gives a formula for the joint mgf of a linear transformation.

Proposition 220 *Let X be a $K \times 1$ random vector possessing a joint mgf $M_X(t)$. Define*

$$Y = A + BX$$

where A is a $L \times 1$ constant vector and B is an $L \times K$ constant matrix. Then, the $L \times 1$ random vector Y possesses a joint mgf $M_Y(t)$, and

$$M_Y(t) = \exp\left(t^\top A\right) M_X\left(B^\top t\right)$$

Proof. Using the definition of joint mgf, we obtain

$$\begin{aligned} M_Y(t) &= \mathrm{E}\left[\exp\left(t^\top Y\right)\right] \\ &= \mathrm{E}\left[\exp\left(t^\top A + t^\top BX\right)\right] \\ &= \mathrm{E}\left[\exp\left(t^\top A\right) \exp\left(t^\top BX\right)\right] \end{aligned}$$

$$\begin{aligned} &= \exp\left(t^\top A\right) \mathrm{E}\left[\exp\left(t^\top BX\right)\right] \\ &= \exp\left(t^\top A\right) \mathrm{E}\left[\exp\left(\left(B^\top t\right)^\top X\right)\right] \\ &= \exp\left(t^\top A\right) M_X\left(B^\top t\right) \end{aligned}$$

If $M_X(t)$ is defined on a closed rectangle H, then $M_Y(t)$ is defined on another closed rectangle whose shape and location depend on A and B. ■

38.4.2 Joint mgf of a vector with independent entries

The next proposition shows how to derive the joint mgf of a vector whose components are independent random variables.

Proposition 221 *Let X be a $K \times 1$ random vector. Let its entries $X_1, \ldots, X_K$ be K mutually independent random variables possessing an mgf. Denote the mgf of the i-th entry of X by $M_{X_i}(t_i)$. Then, the joint mgf of X is*

$$M_X(t_1, \ldots, t_K) = \prod_{i=1}^{K} M_{X_i}(t_i)$$

Proof. This is proved as follows:

$$\begin{aligned} M_X(t) &= \mathrm{E}\left[\exp\left(t^\top X\right)\right] \\ &= \mathrm{E}\left[\exp\left(\sum_{i=1}^{K} t_i X_i\right)\right] \\ &= \mathrm{E}\left[\prod_{i=1}^{K} \exp\left(t_i X_i\right)\right] \\ \boxed{\text{A}} &= \prod_{i=1}^{K} \mathrm{E}\left[\exp\left(t_i X_i\right)\right] \\ \boxed{\text{B}} &= \prod_{i=1}^{K} M_{X_i}(t_i) \end{aligned}$$

where: in step A we have used the fact that the entries of X are mutually independent; in step B we have used the definition of mgf of a random variable. ■

38.4.3 Joint mgf of a sum

The next proposition shows how to derive the joint mgf of a sum of independent random vectors.

Proposition 222 *Let $X_1, \ldots, X_n$ be n mutually independent random vectors*[14]*, all of dimension $K \times 1$. Let Z be their sum:*

$$Z = \sum_{i=1}^{n} X_i$$

[14] See p. 235.

Then, the joint mgf of Z is the product of the joint mgfs of X_1, ..., X_n:

$$M_Z(t) = \prod_{i=1}^{n} M_{X_i}(t)$$

provided the latter exist.

Proof. This is proved as follows:

$$\begin{aligned}
M_Z(t) &= \mathrm{E}\left[\exp\left(t^\top Z\right)\right] \\
&= \mathrm{E}\left[\exp\left(t^\top \sum_{i=1}^{n} X_i\right)\right] \\
&= \mathrm{E}\left[\exp\left(\sum_{i=1}^{n} t^\top X_i\right)\right] \\
&= \mathrm{E}\left[\prod_{i=1}^{n} \exp\left(t^\top X_i\right)\right] \\
\boxed{\text{A}} &= \prod_{i=1}^{n} \mathrm{E}\left[\exp\left(t^\top X_i\right)\right] \\
\boxed{\text{B}} &= \prod_{i=1}^{n} M_{X_i}(t)
\end{aligned}$$

where: in step A we have used the fact that the vectors X_i are mutually independent; in step B we have used the definition of joint mgf. ■

38.5 Solved exercises

Some solved exercises on joint mgfs can be found below.

Exercise 1

Let X be a 2×1 discrete random vector and denote its components by X_1 and X_2. Let the support of X be

$$R_X = \left\{[1 \;\; 1]^\top, [2 \;\; 0]^\top, [0 \;\; 0]^\top\right\}$$

and its joint probability mass function be

$$p_X(x) = \begin{cases} 1/3 & \text{if } x = [1 \;\; 1]^\top \\ 1/3 & \text{if } x = [2 \;\; 0]^\top \\ 1/3 & \text{if } x = [0 \;\; 0]^\top \\ 0 & \text{otherwise} \end{cases}$$

Derive the joint mgf of X, if it exists.

Solution

By using the definition of joint mgf, we get

$$\begin{aligned} M_X(t) &= \mathrm{E}\left[\exp\left(t^\top X\right)\right] = \mathrm{E}\left[\exp\left(t_1X_1 + t_2X_2\right)\right] \\ &= \sum_{(x_1,x_2)\in R_X} \exp\left(t_1x_1 + t_2x_2\right) p_X(x_1,x_2) \\ &= \exp\left(t_1\cdot 1 + t_2\cdot 1\right)\cdot p_X(1,1) + \exp\left(t_1\cdot 2 + t_2\cdot 0\right)\cdot p_X(2,0) \\ &\quad + \exp\left(t_1\cdot 0 + t_2\cdot 0\right)\cdot p_X(0,0) \\ &= \exp\left(t_1+t_2\right)\cdot\frac{1}{3} + \exp\left(2t_1\right)\cdot\frac{1}{3} + \exp\left(0\right)\cdot\frac{1}{3} \\ &= \frac{1}{3}\left[1 + \exp\left(2t_1\right) + \exp\left(t_1+t_2\right)\right] \end{aligned}$$

Obviously, the joint mgf exists and it is well-defined because the above expected value exists for any $t \in \mathbb{R}^2$.

Exercise 2

Let

$$X = \left[X_1 \;\; X_2\right]^\top$$

be a 2×1 random vector with joint mgf

$$M_{X_1,X_2}(t_1,t_2) = \frac{1}{3} + \frac{2}{3}\exp\left(t_1 + 2t_2\right)$$

Derive the expected value of X_1.

Solution

The mgf of X_1 is

$$\begin{aligned} M_{X_1}(t_1) &= \mathrm{E}\left[\exp\left(t_1X_1\right)\right] = \mathrm{E}\left[\exp\left(t_1X_1 + 0\cdot X_2\right)\right] \\ &= M_{X_1,X_2}(t_1,0) = \frac{1}{3} + \frac{2}{3}\exp\left(t_1 + 2\cdot 0\right) \\ &= \frac{1}{3} + \frac{2}{3}\exp\left(t_1\right) \end{aligned}$$

The expected value of X_1 is obtained by taking the first derivative of its mgf:

$$\frac{dM_{X_1}(t_1)}{dt_1} = \frac{2}{3}\exp\left(t_1\right)$$

and evaluating it at $t_1 = 0$:

$$\mathrm{E}\left[X_1\right] = \left.\frac{dM_{X_1}(t_1)}{dt_1}\right|_{t_1=0} = \frac{2}{3}\exp\left(0\right) = \frac{2}{3}$$

Exercise 3

Let

$$X = \left[X_1 \;\; X_2\right]^\top$$

be a 2×1 random vector with joint mgf

$$M_{X_1,X_2}(t_1,t_2) = \frac{1}{3}\left[1 + \exp(t_1 + 2t_2) + \exp(2t_1 + t_2)\right]$$

Derive the covariance between X_1 and X_2.

Solution

We can use the following covariance formula:

$$\operatorname{Cov}[X_1, X_2] = \operatorname{E}[X_1 X_2] - \operatorname{E}[X_1]\operatorname{E}[X_2]$$

The mgf of X_1 is

$$\begin{aligned} M_{X_1}(t_1) &= \operatorname{E}[\exp(t_1 X_1)] = \operatorname{E}[\exp(t_1 X_1 + 0 \cdot X_2)] \\ &= M_{X_1,X_2}(t_1, 0) = \frac{1}{3}\left[1 + \exp(t_1 + 2 \cdot 0) + \exp(2t_1 + 0)\right] \\ &= \frac{1}{3}\left[1 + \exp(t_1) + \exp(2t_1)\right] \end{aligned}$$

The expected value of X_1 is obtained by taking the first derivative of its mgf:

$$\frac{dM_{X_1}(t_1)}{dt_1} = \frac{1}{3}\left[\exp(t_1) + 2\exp(2t_1)\right]$$

and evaluating it at $t_1 = 0$:

$$\operatorname{E}[X_1] = \left.\frac{dM_{X_1}(t_1)}{dt_1}\right|_{t_1=0} = \frac{1}{3}\left[\exp(0) + 2\exp(0)\right] = 1$$

The mgf of X_2 is

$$\begin{aligned} M_{X_2}(t_2) &= \operatorname{E}[\exp(t_2 X_2)] = \operatorname{E}[\exp(0 \cdot X_1 + t_2 X_2)] \\ &= M_{X_1,X_2}(0, t_2) = \frac{1}{3}\left[1 + \exp(0 + 2t_2) + \exp(2 \cdot 0 + t_2)\right] \\ &= \frac{1}{3}\left[1 + \exp(2t_2) + \exp(t_2)\right] \end{aligned}$$

To compute the expected value of X_2 we take the first derivative of its mgf:

$$\frac{dM_{X_2}(t_2)}{dt_2} = \frac{1}{3}\left[2\exp(2t_2) + \exp(t_2)\right]$$

and we evaluate it at $t_2 = 0$:

$$\operatorname{E}[X_2] = \left.\frac{dM_{X_2}(t_2)}{dt_2}\right|_{t_2=0} = \frac{1}{3}\left[2\exp(0) + \exp(0)\right] = 1$$

The second cross-moment of X is computed by taking the second cross-partial derivative of the joint mgf:

$$\begin{aligned} \frac{\partial^2 M_{X_1,X_2}(t_1,t_2)}{\partial t_1 \partial t_2} &= \frac{\partial}{\partial t_1}\left(\frac{\partial}{\partial t_2}\left(\frac{1}{3}\left[1 + \exp(t_1 + 2t_2) + \exp(2t_1 + t_2)\right]\right)\right) \\ &= \frac{\partial}{\partial t_1}\left(\frac{1}{3}\left[2\exp(t_1 + 2t_2) + \exp(2t_1 + t_2)\right]\right) \end{aligned}$$

$$= \frac{1}{3}\left[2\exp\left(t_1+2t_2\right)+2\exp\left(2t_1+t_2\right)\right]$$

and evaluating it at $(t_1, t_2) = (0, 0)$:

$$\begin{aligned}\mathrm{E}\left[X_1X_2\right] &= \left.\frac{\partial^2 M_{X_1,X_2}\left(t_1,t_2\right)}{\partial t_1 \partial t_2}\right|_{t_1=0,t_2=0} \\ &= \frac{1}{3}\left[2\exp\left(0\right)+2\exp\left(0\right)\right] = \frac{4}{3}\end{aligned}$$

Therefore,

$$\begin{aligned}\mathrm{Cov}\left[X_1,X_2\right] &= \mathrm{E}\left[X_1X_2\right]-\mathrm{E}\left[X_1\right]\mathrm{E}\left[X_2\right] \\ &= \frac{4}{3}-1\cdot 1 = \frac{1}{3}\end{aligned}$$

Chapter 39

Characteristic function of a random variable

In the lecture entitled *Moment generating function* (p. 289), we have explained that the distribution of a random variable can be characterized in terms of its moment generating function, a real function that enjoys two important properties: it uniquely determines its associated probability distribution, and its derivatives at zero are equal to the moments of the random variable. We have also explained that not all random variables possess a moment generating function.

The characteristic function (cf) enjoys properties that are almost identical to those enjoyed by the moment generating function, but it has an important advantage: all random variables possess a characteristic function.

39.1 Definition

We start this lecture by giving a definition of characteristic function.

Definition 223 *Let X be a random variable. Let $i = \sqrt{-1}$ be the imaginary unit. The function $\varphi : \mathbb{R} \to \mathbb{C}$ defined by*

$$\varphi_X(t) = \mathrm{E}\left[\exp(itX)\right]$$

is called the ***characteristic function*** *of X.*

The first thing to be noted is that the characteristic function $\varphi_X(t)$ exists for any t. This can be proved as follows:

$$\begin{aligned}\varphi_X(t) &= \mathrm{E}\left[\exp(itX)\right] \\ &= \mathrm{E}\left[\cos(tX) + i\sin(tX)\right] \\ &= \mathrm{E}\left[\cos(tX)\right] + i\mathrm{E}\left[\sin(tX)\right]\end{aligned}$$

and the last two expected values are well-defined, because the sine and cosine functions are bounded in the interval $[-1, 1]$.

39.2 Moments and cfs

Like the moment generating function of a random variable, the characteristic function can be used to derive the moments of X, as stated in the following proposition.

Proposition 224 *Let X be a random variable and $\varphi_X(t)$ its characteristic function. Let $n \in \mathbb{N}$. If the n-th moment of X, denoted by $\mu_X(n)$, exists and is finite, then $\varphi_X(t)$ is n times continuously differentiable and*

$$\mu_X(n) = \mathrm{E}[X^n] = \frac{1}{i^n} \left. \frac{d^n \varphi_X(t)}{dt^n} \right|_{t=0}$$

where $\left. \frac{d^n \varphi_X(t)}{dt^n} \right|_{t=0}$ is the n-th derivative of $\varphi_X(t)$ with respect to t, evaluated at the point $t = 0$.

Proof. The proof of the above proposition is quite complex (see, e.g., Resnick[1] - 1999). The intuition, however, is straightforward: since the expected value is a linear operator and differentiation is a linear operation, under appropriate conditions one can differentiate through the expected value, as follows:

$$\begin{aligned} \frac{d^n \varphi_X(t)}{dt^n} &= \frac{d^n}{dt^n} \mathrm{E}[\exp(itX)] \\ &= \mathrm{E}\left[\frac{d^n}{dt^n} \exp(itX) \right] \\ &= \mathrm{E}[(iX)^n \exp(itX)] \\ &= i^n \mathrm{E}[X^n \exp(itX)] \end{aligned}$$

which, evaluated at the point $t = 0$, yields

$$\left. \frac{d^n \varphi_X(t)}{dt^n} \right|_{t=0} = i^n \mathrm{E}[X^n \exp(0 \cdot iX)] = i^n \mathrm{E}[X^n] = i^n \mu_X(n)$$

■

In practice, the proposition above is not very useful when one wants to compute a moment of a random variable, because it requires to know in advance whether the moment exists or not. A much more useful statement is provided by the next proposition.

Proposition 225 *Let X be a random variable and $\varphi_X(t)$ its characteristic function. If $\varphi_X(t)$ is n times differentiable at the point $t = 0$, then*

1. *if n is **even**, the k-th moment of X exists and is finite for any $k \leq n$;*
2. *if n is **odd**, the k-th moment of X exists and is finite for any $k < n$.*

In both cases, the following holds:

$$\mu_X(k) = \mathrm{E}\left[X^k\right] = \frac{1}{i^k} \left. \frac{d^k \varphi_X(t)}{dt^k} \right|_{t=0}$$

[1] Resnick, S. I. (1999) *A Probability Path*, Birkhauser.

Proof. See e.g. Ushakov[2] (1999). ■

The following example shows how this proposition can be used to compute the second moment of an exponential random variable.

Example 226 *Let X be an exponential random variable with parameter $\lambda \in \mathbb{R}_{++}$. Its support is*

$$R_X = [0, \infty)$$

and its probability density function is

$$f_X(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{if } x \in R_X \\ 0 & \text{if } x \notin R_X \end{cases}$$

Its characteristic function is

$$\varphi_X(t) = \mathrm{E}[\exp(itX)] = \frac{\lambda}{\lambda - it}$$

which is proved in the lecture entitled Exponential distribution (p. 365). Note that dividing by $(\lambda - it)$ does not pose any division-by-zero problem, because $\lambda > 0$ and the denominator is different from 0 also when $t = 0$. The first derivative of the characteristic function is

$$\frac{d\varphi_X(t)}{dt} = \frac{i\lambda}{(\lambda - it)^2}$$

The second derivative of the characteristic function is

$$\frac{d^2\varphi_X(t)}{dt^2} = -\frac{2\lambda}{(\lambda - it)^3}$$

Evaluating it at $t = 0$, we obtain

$$\left.\frac{d^2\varphi_X(t)}{dt^2}\right|_{t=0} = -\frac{2}{\lambda^2}$$

Therefore, the second moment of X exists and is finite. Furthermore, it can be computed as

$$\mathrm{E}\left[X^2\right] = \frac{1}{i^2}\left.\frac{d^2\varphi_X(t)}{dt^2}\right|_{t=0} = \frac{1}{i^2}\left(-\frac{2}{\lambda^2}\right) = \frac{2}{\lambda^2}$$

39.3 Distributions and cfs

Characteristic functions, like moment generating functions, can also be used to characterize the distribution of a random variable.

Proposition 227 (equality of distributions) *Let X and Y be two random variables. Denote by $F_X(x)$ and $F_Y(y)$ their distribution functions[3] and by $\varphi_X(t)$ and $\varphi_Y(t)$ their characteristic functions. X and Y have the same distribution, i.e., $F_X(x) = F_Y(x)$ for any x, if and only if they have the same characteristic function, i.e., $\varphi_X(t) = \varphi_Y(t)$ for any t.*

[2] Ushakov, N. G. (1999) *Selected topics in characteristic functions*, VSP.

[3] See p. 108.

Proof. For a formal proof, see, e.g., Resnick[4] (1999). An informal proof for the special case in which X and Y have a finite support can be provided along the same lines of the proof of Proposition 212, which concerns the moment generating function. This is left as an exercise (just replace $\exp(tX)$ and $\exp(tY)$ in that proof with $\exp(itX)$ and $\exp(itY)$). ■

This property is analogous to the property of joint moment generating functions stated in Proposition 212. The same comments we made about that proposition also apply to this one.

39.4 More details

39.4.1 Cf of a linear transformation

The next proposition gives a formula for the characteristic function of a linear transformation.

Proposition 228 *Let X be a random variable with characteristic function $\varphi_X(t)$. Define*

$$Y = a + bX$$

where $a, b \in \mathbb{R}$ are two constants and $b \neq 0$. Then, the characteristic function of Y is

$$\varphi_Y(t) = \exp(iat)\,\varphi_X(bt)$$

Proof. Using the definition of characteristic function, we get

$$\begin{aligned}
\varphi_Y(t) &= \mathrm{E}[\exp(itY)] \\
&= \mathrm{E}[\exp(iat + ibtX)] \\
&= \mathrm{E}[\exp(iat)\exp(ibtX)] \\
&= \exp(iat)\,\mathrm{E}[\exp(ibtX)] \\
&= \exp(iat)\,\varphi_X(bt)
\end{aligned}$$

■

39.4.2 Cf of a sum

The next proposition shows how to derive the characteristic function of a sum of independent random variables.

Proposition 229 *Let $X_1, \ldots, X_n$ be n mutually independent random variables[5]. Let Z be their sum:*

$$Z = \sum_{j=1}^{n} X_j$$

Then, the characteristic function of Z is the product of the characteristic functions of $X_1, \ldots, X_n$:

$$\varphi_Z(t) = \prod_{j=1}^{n} \varphi_{X_j}(t)$$

[4] Resnick, S. I. (1999) *A Probability Path*, Birkhauser.

[5] See p. 233.

Proof. This is proved as follows:

$$\begin{aligned}
\varphi_Z(t) &= \mathrm{E}\left[\exp(itZ)\right] \\
&= \mathrm{E}\left[\exp\left(it\sum_{j=1}^{n} X_j\right)\right] \\
&= \mathrm{E}\left[\exp\left(\sum_{j=1}^{n} itX_j\right)\right] \\
&= \mathrm{E}\left[\prod_{j=1}^{n} \exp(itX_j)\right] \\
\boxed{\text{A}} &= \prod_{j=1}^{n} \mathrm{E}\left[\exp(itX_j)\right] \\
\boxed{\text{B}} &= \prod_{j=1}^{n} \varphi_{X_j}(t)
\end{aligned}$$

where: in step A we have used the properties of mutually independent variables[6]; in step B we have used the definition of characteristic function. ■

39.4.3 Computation of the characteristic function

When X is a discrete random variable with support R_X and probability mass function $p_X(x)$, its characteristic function is

$$\varphi_X(t) = \mathrm{E}\left[\exp(itX)\right] = \sum_{x \in R_X} \exp(itx)\, p_X(x)$$

Thus, the computation of the characteristic function is pretty straightforward: all we need to do is to sum the complex numbers $\exp(itx)\, p_X(x)$ over all values of x belonging to the support of X.

When X is an absolutely continuous random variable with probability density function $f_X(x)$, its characteristic function is

$$\varphi_X(t) = \mathrm{E}\left[\exp(itX)\right] = \int_{-\infty}^{\infty} \exp(itx)\, f_X(x)\, dx$$

The right-hand side integral is a contour integral of a complex function along the real axis. As people reading these lecture notes are usually not familiar with contour integration (a topic in complex analysis), we avoid it altogether in the rest of this book. We instead exploit the fact that

$$\exp(itx) = \cos(tx) + i\sin(tx)$$

to rewrite the contour integral as the complex sum of two ordinary integrals:

$$\int_{-\infty}^{\infty} \exp(itx)\, f_X(x)\, dx = \int_{-\infty}^{\infty} \cos(tx) f_X(x)\, dx + i\int_{-\infty}^{\infty} \sin(tx) f_X(x)\, dx$$

and to compute the two integrals separately.

[6] See p. 234.

39.5 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Let X be a discrete random variable having support

$$R_X = \{0, 1, 2\}$$

and probability mass function

$$p_X(x) = \begin{cases} 1/3 & \text{if } x = 0 \\ 1/3 & \text{if } x = 1 \\ 1/3 & \text{if } x = 2 \\ 0 & \text{if } x \notin R_X \end{cases}$$

Derive the characteristic function of X.

Solution

By using the definition of characteristic function, we obtain

$$\begin{aligned} \varphi_X(t) &= \mathrm{E}\left[\exp(itX)\right] = \sum_{x \in R_X} \exp(itx)\, p_X(x) \\ &= \exp(it \cdot 0) \cdot p_X(0) + \exp(it \cdot 1) \cdot p_X(1) + \exp(it \cdot 2) \cdot p_X(2) \\ &= \frac{1}{3} + \frac{1}{3}\exp(it) + \frac{1}{3}\exp(2it) = \frac{1}{3}\left[1 + \exp(it) + \exp(2it)\right] \end{aligned}$$

Exercise 2

Use the characteristic function found in the previous exercise to derive the variance of X.

Solution

We can use the following formula for computing the variance:

$$\mathrm{Var}[X] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2$$

The expected value of X is computed by taking the first derivative of the characteristic function:

$$\frac{d\varphi_X(t)}{dt} = \frac{1}{3}\left[i\exp(it) + 2i\exp(2it)\right]$$

evaluating it at $t = 0$, and dividing it by i:

$$\mathrm{E}[X] = \frac{1}{i}\left.\frac{d\varphi_X(t)}{dt}\right|_{t=0} = \frac{1}{i}\frac{1}{3}\left[i\exp(i \cdot 0) + 2i\exp(2i \cdot 0)\right] = 1$$

The second moment of X is computed by taking the second derivative of the characteristic function:

$$\frac{d^2\varphi_X(t)}{dt^2} = \frac{1}{3}\left[i^2\exp(it) + 4i^2\exp(2it)\right]$$

evaluating it at $t = 0$, and dividing it by i^2:

$$\mathrm{E}\left[X^2\right] = \frac{1}{i^2}\left.\frac{d^2\varphi_X(t)}{dt^2}\right|_{t=0} = \frac{1}{i^2}\frac{1}{3}\left[i^2\exp(i\cdot 0) + 4i^2\exp(2i\cdot 0)\right] = \frac{5}{3}$$

Therefore,

$$\mathrm{Var}[X] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 = \frac{5}{3} - 1^2 = \frac{2}{3}$$

Exercise 3

Read and try to understand how the characteristic functions of the uniform and exponential distributions are derived in the lectures entitled *Uniform distribution* (p. 359) and *Exponential distribution* (p. 365).

Chapter 40

Characteristic function of a random vector

This lecture introduces the notion of joint characteristic function (joint cf) of a random vector, which is a multivariate generalization of the concept of characteristic function of a random variable. Before reading this lecture, you are advised to first read the lecture entitled *Characteristic function* (p. 307).

40.1 Definition

Let us start this lecture with a definition.

Definition 230 *Let X be a $K \times 1$ random vector. Let $i = \sqrt{-1}$ be the imaginary unit. The function $\varphi : \mathbb{R}^K \to \mathbb{C}$ defined by*

$$\varphi_X(t) = \mathrm{E}\left[\exp\left(it^\top X\right)\right] = \mathrm{E}\left[\exp\left(i\sum_{j=1}^{K} t_j X_j\right)\right]$$

is called the ***joint characteristic function*** *of X.*

The first thing to be noted is that the joint characteristic function $\varphi_X(t)$ exists for any $t \in \mathbb{R}^K$. This can be proved as follows:

$$\begin{aligned}\varphi_X(t) &= \mathrm{E}\left[\exp\left(it^\top X\right)\right] \\ &= \mathrm{E}\left[\cos\left(t^\top X\right) + i\sin\left(t^\top X\right)\right] \\ &= \mathrm{E}\left[\cos\left(t^\top X\right)\right] + i\mathrm{E}\left[\sin\left(t^\top X\right)\right]\end{aligned}$$

and the last two expected values are well-defined, because the sine and cosine functions are bounded in the interval $[-1, 1]$.

40.2 Cross-moments and joint cfs

Like the joint moment generating function[1] of a random vector, the joint characteristic function can be used to derive the cross-moments[2] of X, as stated in the

[1] See p. 297.
[2] See p. 285.

following proposition.

Proposition 231 *Let X be a random vector and $\varphi_X(t)$ its joint characteristic function. Let $n \in \mathbb{N}$. Define a cross-moment of order n as follows:*

$$\mu_X(n_1, n_2, \ldots, n_K) = \mathrm{E}[X_1^{n_1} \cdot X_2^{n_2} \cdot \ldots \cdot X_K^{n_K}]$$

where $n_1, n_2, \ldots, n_K \in \mathbb{Z}_+$ and

$$n = \sum_{k=1}^{K} n_k$$

If all cross-moments of order n exist and are finite, then all the n-th order partial derivatives of $\varphi_X(t)$ exist and

$$\mu_X(n_1, n_2, \ldots, n_K) = \frac{1}{i^n} \left. \frac{\partial^{n_1+n_2+\ldots+n_K} \varphi_X(t_1, t_2, \ldots, t_K)}{\partial t_1^{n_1} \partial t_2^{n_2} \ldots \partial t_K^{n_K}} \right|_{t_1=0, t_2=0, \ldots, t_K=0}$$

where the partial derivative on the right-hand side of the equation is evaluated at the point $t_1 = 0$, $t_2 = 0$, $\ldots$, $t_K = 0$.

Proof. See Ushakov[3] (1999). ■

In practice, the proposition above is not very useful when one wants to compute a cross-moment of a random vector, because the proposition requires to know in advance whether the cross-moment exists or not. A much more useful proposition is the following.

Proposition 232 *Let X be a random vector and $\varphi_X(t)$ its joint characteristic function. If all the n-th order partial derivatives of $\varphi_X(t)$ exist, then:*

1. *if n is **even**, for any*

$$m = \sum_{k=1}^{K} m_k \leq n$$

all m-th cross-moments of X exist and are finite;

2. *if n is **odd**, for any*

$$m = \sum_{k=1}^{K} m_k < n$$

all m-th cross-moments of X exist and are finite.

In both cases,

$$\mu_X(m_1, m_2, \ldots, m_K) = \frac{1}{i^n} \left. \frac{\partial^{m_1+m_2+\ldots+m_K} \varphi_X(t_1, t_2, \ldots, t_K)}{\partial t_1^{m_1} \partial t_2^{m_2} \ldots \partial t_K^{m_K}} \right|_{t_1=0, t_2=0, \ldots, t_K=0}$$

Proof. See Ushakov (1999). ■

[3] Ushakov, N. G. (1999) *Selected topics in characteristic functions*, VSP.

40.3 Joint distributions and joint cfs

The next proposition states the most important property of the joint characteristic function.

Proposition 233 (equality of distributions) *Let X and Y be two $K \times 1$ random vectors. Denote by $F_X(x)$ and $F_Y(y)$ their joint distribution functions*[4] *and by $\varphi_X(t)$ and $\varphi_Y(t)$ their joint characteristic functions. X and Y have the same distribution, i.e., $F_X(x) = F_Y(x)$ for any $x \in \mathbb{R}^K$, if and only if they have the same characteristic functions, i.e., $\varphi_X(t) = \varphi_Y(t)$ for any $t \in \mathbb{R}^K$.*

Proof. See Ushakov (1999). An informal proof for the special case in which X and Y have a finite support can be provided along the same lines of the proof of Proposition 219, which concerns the joint moment generating function. This is left as an exercise (just replace $\exp\left(t^\top X\right)$ and $\exp\left(t^\top Y\right)$ in that proof with $\exp\left(it^\top X\right)$ and $\exp\left(it^\top Y\right)$). ■

This property is analogous to the property of joint moment generating functions stated in Proposition 219. The same comments we made about that proposition also apply to this one.

40.4 More details

40.4.1 Joint cf of a linear transformation

The next proposition gives a formula for the joint characteristic function of a linear transformation.

Proposition 234 *Let X be a $K \times 1$ random vector with characteristic function $\varphi_X(t)$. Define*

$$Y = A + BX$$

where A is a $L \times 1$ constant vector and B is a $L \times K$ constant matrix. Then, the characteristic function of Y is

$$\varphi_Y(t) = \exp\left(it^\top A\right)\varphi_X\left(B^\top t\right)$$

Proof. By using the definition of characteristic function, we obtain

$$\begin{aligned}
\varphi_Y(t) &= \mathrm{E}\left[\exp\left(it^\top Y\right)\right] \\
&= \mathrm{E}\left[\exp\left(it^\top A + it^\top BX\right)\right] \\
&= \mathrm{E}\left[\exp\left(it^\top A\right)\exp\left(it^\top BX\right)\right] \\
&= \exp\left(it^\top A\right)\mathrm{E}\left[\exp\left(it^\top BX\right)\right] \\
&= \exp\left(it^\top A\right)\mathrm{E}\left[\exp\left(i\left(B^\top t\right)^\top X\right)\right] \\
&= \exp\left(it^\top A\right)\varphi_X\left(B^\top t\right)
\end{aligned}$$

■

[4] See p. 118.

40.4.2 Joint cf of a random vector with independent entries

The next proposition shows how to derive the joint characteristic function of a vector whose components are independent random variables.

Proposition 235 *Let X be a $K \times 1$ random vector. Let its entries $X_1, \ldots, X_K$ be K mutually independent random variables. Denote the characteristic function of the j-th entry of X by $\varphi_{X_j}(t_j)$. Then, the joint characteristic function of X is*

$$\varphi_X(t_1, \ldots, t_K) = \prod_{j=1}^{K} \varphi_{X_j}(t_j)$$

Proof. This is proved as follows:

$$\begin{aligned}
\varphi_X(t) &= \mathrm{E}\left[\exp\left(it^\top X\right)\right] \\
&= \mathrm{E}\left[\exp\left(i\sum_{j=1}^{K} t_j X_j\right)\right] \\
&= \mathrm{E}\left[\prod_{j=1}^{K} \exp\left(it_j X_j\right)\right] \\
\boxed{\text{A}} &= \prod_{j=1}^{K} \mathrm{E}\left[\exp\left(it_j X_j\right)\right] \\
\boxed{\text{B}} &= \prod_{j=1}^{K} \varphi_{X_j}(t_j)
\end{aligned}$$

where: in step A we have used the fact that the entries of X are mutually independent[5]; in step B we have used the definition of characteristic function of a random variable[6]. ■

40.4.3 Joint cf of a sum

The next proposition shows how to derive the joint characteristic function of a sum of independent random vectors.

Proposition 236 *Let $X_1, \ldots, X_n$ be n mutually independent random vectors. Let Z be their sum:*

$$Z = \sum_{j=1}^{n} X_j$$

Then, the joint characteristic function of Z is the product of the joint characteristic functions of $X_1, \ldots, X_n$:

$$\varphi_Z(t) = \prod_{j=1}^{n} \varphi_{X_j}(t)$$

[5] In particular, see the *mutual independence via expectations* property (p. 234).
[6] See p. 307.

Proof. This is proved as follows:

$$\begin{aligned}
\varphi_Z(t) &= \mathrm{E}\left[\exp\left(it^\top Z\right)\right] \\
&= \mathrm{E}\left[\exp\left(it^\top \sum_{j=1}^{n} X_j\right)\right] \\
&= \mathrm{E}\left[\exp\left(\sum_{j=1}^{n} it^\top X_j\right)\right] \\
&= \mathrm{E}\left[\prod_{j=1}^{n} \exp\left(it^\top X_j\right)\right] \\
\boxed{\text{A}} &= \prod_{j=1}^{n} \mathrm{E}\left[\exp\left(it^\top X_j\right)\right] \\
\boxed{\text{B}} &= \prod_{j=1}^{n} \varphi_{X_j}(t)
\end{aligned}$$

where: in step A we have used the fact that the vectors X_j are mutually independent; in step B we have used the definition of joint characteristic function of a random vector given above. ■

40.5 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Let Z_1 and Z_2 be two independent standard normal random variables[7]. Let X be a 2×1 random vector whose components are defined as follows:

$$\begin{aligned}
X_1 &= Z_1^2 \\
X_2 &= Z_1^2 + Z_2^2
\end{aligned}$$

Derive the joint characteristic function of X.

Hint: use the fact that Z_1^2 and Z_2^2 are two independent Chi-square random variables[8] having characteristic function

$$\varphi_{Z_1^2}(t) = \varphi_{Z_2^2}(t) = (1 - 2it)^{-1/2}$$

Solution

By using the definition of characteristic function, we get

$$\begin{aligned}
\varphi_X(t) &= \mathrm{E}\left[\exp\left(it^\top X\right)\right] \\
&= \mathrm{E}\left[\exp\left(it_1 X_1 + it_2 X_2\right)\right]
\end{aligned}$$

[7] See p. 376.
[8] See p. 387.

$$
\begin{aligned}
&= \mathrm{E}\left[\exp\left(it_1 Z_1^2 + it_2\left(Z_1^2 + Z_2^2\right)\right)\right] \\
&= \mathrm{E}\left[\exp\left(i\left(t_1+t_2\right)Z_1^2 + it_2 Z_2^2\right)\right] \\
&= \mathrm{E}\left[\exp\left(i\left(t_1+t_2\right)Z_1^2\right)\exp\left(it_2 Z_2^2\right)\right] \\
\boxed{\text{A}} \quad &= \mathrm{E}\left[\exp\left(i\left(t_1+t_2\right)Z_1^2\right)\right]\mathrm{E}\left[\exp\left(it_2 Z_2^2\right)\right] \\
\boxed{\text{B}} \quad &= \varphi_{Z_1^2}\left(t_1+t_2\right)\varphi_{Z_2^2}\left(t_2\right) \\
&= \left(1-2it_1-2it_2\right)^{-1/2}\left(1-2it_2\right)^{-1/2} \\
&= \left[\left(1-2it_1-2it_2\right)\left(1-2it_2\right)\right]^{-1/2} \\
&= \left[1-2it_1-2it_2-2it_2\cdot 1-2it_2\cdot(-2it_1)-2it_2\cdot(-2it_2)\right]^{-1/2} \\
&= \left[1-2it_1-4it_2-4t_1t_2-4t_2^2\right]^{-1/2}
\end{aligned}
$$

where: in step A we have used the fact that Z_1 and Z_2 are independent; in step B we have used the definition of characteristic function.

Exercise 2

Use the joint characteristic function found in the previous exercise to derive the expected value and the covariance matrix of X.

Solution

We need to compute the partial derivatives of the joint characteristic function:

$$
\begin{aligned}
\frac{\partial\varphi}{\partial t_1} &= -\frac{1}{2}\left[1-2it_1-4it_2-4t_1t_2-4t_2^2\right]^{-3/2}\left(-2i-4t_2\right) \\
\frac{\partial\varphi}{\partial t_2} &= -\frac{1}{2}\left[1-2it_1-4it_2-4t_1t_2-4t_2^2\right]^{-3/2}\left(-4i-4t_1-8t_2\right) \\
\frac{\partial^2\varphi}{\partial t_1^2} &= \frac{3}{4}\left[1-2it_1-4it_2-4t_1t_2-4t_2^2\right]^{-5/2}\left(-2i-4t_2\right)^2 \\
\frac{\partial^2\varphi}{\partial t_2^2} &= \frac{3}{4}\left[1-2it_1-4it_2-4t_1t_2-4t_2^2\right]^{-5/2}\left(-4i-4t_1-8t_2\right)^2 \\
&\quad +4\cdot\left[1-2it_1-4it_2-4t_1t_2-4t_2^2\right]^{-3/2} \\
\frac{\partial^2\varphi}{\partial t_1\partial t_2} &= \frac{3}{4}\left[1-2it_1-4it_2-4t_1t_2-4t_2^2\right]^{-5/2}\left(-2i-4t_2\right)\left(-4i-4t_1-8t_2\right) \\
&\quad +2\cdot\left[1-2it_1-4it_2-4t_1t_2-4t_2^2\right]^{-3/2}
\end{aligned}
$$

All partial derivatives up to the second order exist and are well defined. As a consequence, all cross-moments up to the second order exist and are finite and they can be computed from the above partial derivatives:

$$
\begin{aligned}
\mathrm{E}\left[X_1\right] &= \frac{1}{i}\left.\frac{\partial\varphi}{\partial t_1}\right|_{t_1=0,t_2=0} = \frac{1}{i}\cdot i = 1 \\
\mathrm{E}\left[X_2\right] &= \frac{1}{i}\left.\frac{\partial\varphi}{\partial t_2}\right|_{t_1=0,t_2=0} = \frac{1}{i}\cdot 2i = 2
\end{aligned}
$$

$$\begin{aligned}
\mathrm{E}\left[X_1^2\right] &= \frac{1}{i^2}\left.\frac{\partial^2\varphi}{\partial t_1^2}\right|_{t_1=0,t_2=0} = \frac{1}{i^2}\cdot 3i^2 = 3 \\
\mathrm{E}\left[X_2^2\right] &= \frac{1}{i^2}\left.\frac{\partial^2\varphi}{\partial t_2^2}\right|_{t_1=0,t_2=0} = \frac{1}{i^2}\cdot\left(12i^2+4\right) = 8 \\
\mathrm{E}\left[X_1X_2\right] &= \frac{1}{i^2}\left.\frac{\partial^2\varphi}{\partial t_1\partial t_2}\right|_{t_1=0,t_2=0} = \frac{1}{i^2}\cdot\left(6i^2+2\right) = 4
\end{aligned}$$

The covariances are derived as follows:

$$\begin{aligned}
\mathrm{Var}\left[X_1\right] &= \mathrm{E}\left[X_1^2\right] - \mathrm{E}\left[X_1\right]^2 = 3-1 = 2 \\
\mathrm{Var}\left[X_2\right] &= \mathrm{E}\left[X_2^2\right] - \mathrm{E}\left[X_2\right]^2 = 8-4 = 4 \\
\mathrm{Cov}\left[X_1,X_2\right] &= \mathrm{E}\left[X_1X_2\right] - \mathrm{E}\left[X_1\right]\mathrm{E}\left[X_2\right] = 4-2 = 2
\end{aligned}$$

Summing up, we have

$$\begin{aligned}
\mathrm{E}\left[X\right] &= \begin{bmatrix} \mathrm{E}\left[X_1\right] & \mathrm{E}\left[X_2\right] \end{bmatrix}^\top = \begin{bmatrix} 1 & 2 \end{bmatrix}^\top \\
\mathrm{Var}\left[X\right] &= \begin{bmatrix} \mathrm{Var}\left[X_1\right] & \mathrm{Cov}\left[X_1,X_2\right] \\ \mathrm{Cov}\left[X_1,X_2\right] & \mathrm{Var}\left[X_2\right] \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 2 & 4 \end{bmatrix}
\end{aligned}$$

Exercise 3

Read and try to understand how the joint characteristic function of the multinomial distribution is derived in the lecture entitled *Multinomial distribution* (p. 431).

Chapter 41

Sums of independent random variables

This lecture discusses how to derive the distribution of the sum of two independent random variables[1]. We explain first how to derive the distribution function[2] of the sum and then how to derive its probability mass function[3] (if the summands are discrete) or its probability density function[4] (if the summands are continuous).

41.1 Distribution function of a sum

The following proposition characterizes the distribution function of the sum in terms of the distribution functions of the two summands:

Proposition 237 *Let X and Y be two independent random variables and denote by $F_X(x)$ and $F_Y(y)$ their respective distribution functions. Let*

$$Z = X + Y$$

and denote the distribution function of Z by $F_Z(z)$. The following holds:

$$F_Z(z) = \mathrm{E}\left[F_X(z - Y)\right]$$

or

$$F_Z(z) = \mathrm{E}\left[F_Y(z - X)\right]$$

Proof. The first formula is derived as follows:

$$\begin{aligned} & F_Z(z) \\ \boxed{\text{A}} \quad = \quad & \mathrm{P}(Z \leq z) \\ = \quad & \mathrm{P}(X + Y \leq z) \\ = \quad & \mathrm{P}(X \leq z - Y) \end{aligned}$$

[1] See p. 229.
[2] See p. 108.
[3] See p. 106.
[4] See p. 107.

$$\begin{aligned} \boxed{\text{B}} &= \mathrm{E}\left[\mathrm{P}\left(X \leq z - Y \,|Y = y\right)\right] \\ \boxed{\text{C}} &= \mathrm{E}\left[F_X\left(z - Y\right)\right] \end{aligned}$$

where: in step $\boxed{\text{A}}$ we have used the definition of distribution function; in step $\boxed{\text{B}}$ we have used the law of iterated expectations; in step $\boxed{\text{C}}$ we have used the fact that X and Y are independent. The second formula is symmetric to the first. ■

The following example illustrates how the above proposition can be used.

Example 238 *Let X be a uniform random variable[5] with support*

$$R_X = [0, 1]$$

and probability density function

$$f_X(x) = \begin{cases} 1 & \text{if } x \in R_X \\ 0 & \text{otherwise} \end{cases}$$

and Y another uniform random variable, independent of X, with support

$$R_Y = [0, 1]$$

and probability density function

$$f_Y(y) = \begin{cases} 1 & \text{if } y \in R_Y \\ 0 & \text{otherwise} \end{cases}$$

The distribution function of X is

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt = \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } 0 < x \leq 1 \\ 1 & \text{if } x > 1 \end{cases}$$

The distribution function of $Z = X + Y$ is

$$\begin{aligned} F_Z(z) &= \mathrm{E}\left[F_X\left(z - Y\right)\right] \\ &= \int_{-\infty}^{\infty} F_X\left(z - y\right) f_Y(y)\,dy \\ &= \int_{0}^{1} F_X\left(z - y\right) dy \\ \boxed{A} &= -\int_{z}^{z-1} F_X(t)\,dt \\ \boxed{B} &= \int_{z-1}^{z} F_X(t)\,dt \end{aligned}$$

where: in step $\boxed{A}$ we have made a change of variable[6] ($t = z - y$); in step $\boxed{B}$ we have exchanged the bounds of integration[7]. There are four cases to consider:

[5] See p. 45.
[6] See p. 50.
[7] See p. 51.

1. *If* $z \leq 0$, *then*

$$F_Z(z) = \int_{z-1}^{z} F_X(t)\,dt = \int_{z-1}^{z} 0dt = 0$$

2. *If* $0 < z \leq 1$, *then*

$$\begin{aligned} F_Z(z) &= \int_{z-1}^{z} F_X(t)\,dt = \int_{z-1}^{0} F_X(t)\,dt + \int_{0}^{z} F_X(t)\,dt \\ &= \int_{z-1}^{0} 0dt + \int_{0}^{z} tdt = 0 + \left[\frac{1}{2}t^2\right]_0^z = \frac{1}{2}z^2 \end{aligned}$$

3. *If* $1 < z \leq 2$, *then*

$$\begin{aligned} F_Z(z) &= \int_{z-1}^{z} F_X(t)\,dt = \int_{z-1}^{1} F_X(t)\,dt + \int_{1}^{z} F_X(t)\,dt \\ &= \int_{z-1}^{1} tdt + \int_{1}^{z} 1dt = \left[\frac{1}{2}t^2\right]_{z-1}^{1} + [t]_1^z \\ &= \frac{1}{2}1^2 - \frac{1}{2}(z-1)^2 + z - 1 \\ &= \frac{1}{2} - \frac{1}{2}z^2 + \frac{1}{2}2z - \frac{1}{2}1^2 + z - 1 \\ &= -\frac{1}{2}z^2 + 2z - 1 \end{aligned}$$

4. *If* $z > 2$, *then*

$$\begin{aligned} F_Z(z) &= \int_{z-1}^{z} F_X(t)\,dt = \int_{z-1}^{z} 1dt \\ &= [t]_{z-1}^{z} = z - (z-1) = 1 \end{aligned}$$

Therefore, combining these four possible cases, we obtain

$$F_Z(z) = \begin{cases} 0 & \text{if } z \leq 0 \\ \frac{1}{2}z^2 & \text{if } 0 < z \leq 1 \\ -\frac{1}{2}z^2 + 2z - 1 & \text{if } 1 < z \leq 2 \\ 1 & \text{if } z > 2 \end{cases}$$

41.2 Probability mass function of a sum

When the two summands are discrete random variables, the probability mass function of their sum can be derived as follows:

Proposition 239 *Let* X *and* Y *be two independent discrete random variables and denote by* $p_X(x)$ *and* $p_Y(y)$ *their respective probability mass functions and by* R_X *and* R_Y *their supports. Let*

$$Z = X + Y$$

and denote the probability mass function of Z *by* $p_Z(z)$. *The following holds:*

$$p_Z(z) = \sum_{y \in R_Y} p_X(z-y)\,p_Y(y)$$

or

$$p_Z(z) = \sum_{x \in R_X} p_Y(z-x)\, p_X(x)$$

Proof. The first formula is derived as follows:

$$\begin{aligned}
& p_Z(z) \\
\boxed{\text{A}} \quad &= \mathrm{P}(Z = z) \\
&= \mathrm{P}(X + Y = z) \\
&= \mathrm{P}(X = z - Y) \\
\boxed{\text{B}} \quad &= \mathrm{E}\left[\mathrm{P}(X = z - Y \,|Y = y)\right] \\
\boxed{\text{C}} \quad &= \mathrm{E}\left[p_X(z - Y)\right] \\
\boxed{\text{D}} \quad &= \sum_{y \in R_Y} p_X(z-y)\, p_Y(y)
\end{aligned}$$

where: in step A we have used the definition of probability mass function; in step B we have used the law of iterated expectations; in step C we have used the fact that X and Y are independent; in step D we have used the definition of expected value. The second formula is symmetric to the first. ■

The two summations above are called **convolutions** (of two probability mass functions).

Example 240 *Let X be a discrete random variable with support*

$$R_X = \{0, 1\}$$

and probability mass function

$$p_X(x) = \begin{cases} 1/2 & \text{if } x \in R_X \\ 0 & \text{otherwise} \end{cases}$$

and Y another discrete random variable, independent of X, with support

$$R_Y = \{0, 1\}$$

and probability mass function

$$p_Y(y) = \begin{cases} 1/2 & \text{if } y \in R_Y \\ 0 & \text{otherwise} \end{cases}$$

Define

$$Z = X + Y$$

Its support is

$$R_Y = \{0, 1, 2\}$$

The probability mass function of Z, evaluated at $z = 0$ is

$$p_Z(0) = \sum_{y \in R_Y} p_X(0-y)\, p_Y(y) = p_X(0-0)\, p_Y(0) + p_X(0-1)\, p_Y(1)$$

$$= \frac{1}{2}\cdot\frac{1}{2}+0\cdot\frac{1}{2}=\frac{1}{4}$$

Evaluated at $z=1$, it is

$$\begin{aligned} p_Z(1) &= \sum_{y\in R_Y} p_X(1-y)\,p_Y(y) = p_X(1-0)\,p_Y(0)+p_X(1-1)\,p_Y(1) \\ &= \frac{1}{2}\cdot\frac{1}{2}+\frac{1}{2}\cdot\frac{1}{2}=\frac{1}{2} \end{aligned}$$

Evaluated at $z=2$, it is

$$\begin{aligned} p_Z(2) &= \sum_{y\in R_Y} p_X(2-y)\,p_Y(y) = p_X(2-0)\,p_Y(0)+p_X(2-1)\,p_Y(1) \\ &= 0\cdot\frac{1}{2}+\frac{1}{2}\cdot\frac{1}{2}=\frac{1}{4} \end{aligned}$$

Therefore, the probability mass function of Z is

$$p_Z(z)=\begin{cases} 1/4 & \text{if } z=0 \\ 1/2 & \text{if } z=1 \\ 1/4 & \text{if } z=2 \\ 0 & \text{otherwise} \end{cases}$$

41.3 Probability density function of a sum

When the two summands are absolutely continuous random variables, the probability density function of their sum can be derived as follows:

Proposition 241 *Let X and Y be two independent absolutely continuous random variables and denote by $f_X(x)$ and $f_Y(y)$ their respective probability density functions. Let*

$$Z=X+Y$$

and denote the probability density function of Z by $f_Z(z)$. The following holds:

$$f_Z(z)=\int_{-\infty}^{\infty} f_X(z-y)\,f_Y(y)\,dy$$

or

$$f_Z(z)=\int_{-\infty}^{\infty} f_Y(z-x)\,f_X(x)\,dx$$

Proof. As stated in Proposition 237, the distribution function of a sum of independent variables is

$$F_Z(z)=\mathrm{E}\left[F_X(z-Y)\right]$$

Differentiating both sides and using the fact that the density function is the derivative of the distribution function[8], we obtain

$$\begin{aligned} & f_Z(z) \\ = \; & \frac{d}{dz}\mathrm{E}\left[F_X(z-Y)\right] \end{aligned}$$

[8] See p. 109.

$$\begin{aligned} \boxed{\text{A}} \quad &= \mathrm{E}\left[\frac{d}{dz}F_X\left(z-Y\right)\right] \\ &= \mathrm{E}\left[f_X\left(z-Y\right)\right] \\ \boxed{\text{B}} \quad &= \int_{-\infty}^{\infty} f_X\left(z-y\right) f_Y\left(y\right) dy \end{aligned}$$

where: in step A we have interchanged differentiation and expectation; in step B we have used the definition of expected value. The second formula is symmetric to the first. ■

The two integrals above are called **convolutions** (of two probability density functions).

Example 242 *Let X be an exponential random variable*[9] *with support*

$$R_X = [0, \infty)$$

and probability density function

$$f_X\left(x\right) = \begin{cases} \exp\left(-x\right) & \text{if } x \in R_X \\ 0 & \text{otherwise} \end{cases}$$

and Y another exponential random variable, independent of X, with support

$$R_Y = [0, \infty)$$

and probability density function

$$f_Y\left(y\right) = \begin{cases} \exp\left(-y\right) & \text{if } y \in R_Y \\ 0 & \text{otherwise} \end{cases}$$

Define

$$Z = X + Y$$

The support of Z is

$$R_Z = [0, \infty)$$

When $z \in R_Z$, the probability density function of Z is

$$\begin{aligned} f_Z\left(z\right) &= \int_{-\infty}^{\infty} f_X\left(z-y\right) f_Y\left(y\right) dy = \int_{0}^{\infty} f_X\left(z-y\right) \exp\left(-y\right) dy \\ &= \int_{0}^{\infty} \exp\left(-\left(z-y\right)\right) 1_{\{z-y\geq 0\}} \exp\left(-y\right) dy \\ &= \int_{0}^{\infty} \exp\left(-z+y\right) 1_{\{y\leq z\}} \exp\left(-y\right) dy \\ &= \int_{0}^{z} \exp\left(-z+y\right) \exp\left(-y\right) dy = \exp(-z) \int_{0}^{z} dy = z\exp\left(-z\right) \end{aligned}$$

Therefore, the probability density function of Z is

$$f_Z\left(z\right) = \begin{cases} z\exp\left(-z\right) & \text{if } z \in R_Z \\ 0 & \text{otherwise} \end{cases}$$

[9] See p. 365.

41.4 More details

41.4.1 Sum of n independent random variables

We have discussed above how to derive the distribution of the sum of two independent random variables. How do we derive the distribution of the sum of more than two mutually independent[10] random variables? Suppose X_1, X_2, ..., X_n are n mutually independent random variables and let Z be their sum:

$$Z = X_1 + \ldots + X_n$$

The distribution of Z can be derived recursively, using the results for sums of two random variables given above:

1. first, define
$$Y_2 = X_1 + X_2$$
and compute the distribution of Y_2;
2. then, define
$$Y_3 = Y_2 + X_3$$
and compute the distribution of Y_3;
3. and so on, until the distribution of Z can be computed from
$$Z = Y_n = Y_{n-1} + X_n$$

41.5 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Let X be a uniform random variable with support

$$R_X = [0, 1]$$

and probability density function

$$f_X(x) = \begin{cases} 1 & \text{if } x \in R_X \\ 0 & \text{otherwise} \end{cases}$$

and Y an exponential random variable, independent of X, with support

$$R_Y = [0, \infty)$$

and probability density function

$$f_Y(y) = \begin{cases} \exp(-y) & \text{if } y \in R_Y \\ 0 & \text{otherwise} \end{cases}$$

Derive the probability density function of the sum

$$Z = X + Y$$

[10] See p. 233.

Solution

The support of Z is

$$R_Z = [0, \infty)$$

When $z \in R_Z$, the probability density function of Z is

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} f_X(z-y) f_Y(y)\, dy = \int_0^{\infty} f_X(z-y) \exp(-y)\, dy \\ &= \int_0^{\infty} 1_{\{0 \leq z-y \leq 1\}} \exp(-y)\, dy = \int_0^{\infty} 1_{\{-1 \leq y-z \leq 0\}} \exp(-y)\, dy \\ &= \int_0^{\infty} 1_{\{z-1 \leq y \leq z\}} \exp(-y)\, dy = \int_{z-1}^{z} \exp(-y)\, dy \\ &= [-\exp(-y)]_{z-1}^{z} = -\exp(-z) + \exp(1-z) \end{aligned}$$

Therefore, the probability density function of Z is

$$f_Z(z) = \begin{cases} -\exp(-z) + \exp(1-z) & \text{if } z \in R_Z \\ 0 & \text{otherwise} \end{cases}$$

Exercise 2

Let X be a discrete random variable with support

$$R_X = \{0, 1, 2\}$$

and probability mass function:

$$p_X(x) = \begin{cases} 1/3 & \text{if } x \in R_X \\ 0 & \text{otherwise} \end{cases}$$

and Y another discrete random variable, independent of X, with support

$$R_Y = \{1, 2\}$$

and probability mass function:

$$p_Y(y) = \begin{cases} y/3 & \text{if } y \in R_Y \\ 0 & \text{otherwise} \end{cases}$$

Derive the probability mass function of the sum

$$Z = X + Y$$

Solution

The support of Z is:

$$R_Z = \{1, 2, 3, 4\}$$

The probability mass function of Z, evaluated at $z = 1$ is:

$$\begin{aligned} p_Z(1) &= \sum_{y \in R_Y} p_X(1-y)\, p_Y(y) = p_X(1-1)\, p_Y(1) + p_X(1-2)\, p_Y(2) \\ &= \frac{1}{3} \cdot \frac{1}{3} + 0 \cdot \frac{2}{3} = \frac{1}{9} \end{aligned}$$

Evaluated at $z = 2$, it is:

$$\begin{aligned} p_Z(2) &= \sum_{y \in R_Y} p_X(2-y)\, p_Y(y) = p_X(2-1)\, p_Y(1) + p_X(2-2)\, p_Y(2) \\ &= \frac{1}{3} \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{2}{3} = \frac{1}{3} \end{aligned}$$

Evaluated at $z = 3$, it is:

$$\begin{aligned} p_Z(3) &= \sum_{y \in R_Y} p_X(3-y)\, p_Y(y) = p_X(3-1)\, p_Y(1) + p_X(3-2)\, p_Y(2) \\ &= \frac{1}{3} \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{2}{3} = \frac{1}{3} \end{aligned}$$

Evaluated at $z = 4$, it is:

$$\begin{aligned} p_Z(4) &= \sum_{y \in R_Y} p_X(4-y)\, p_Y(y) = p_X(4-1)\, p_Y(1) + p_X(4-2)\, p_Y(2) \\ &= 0 \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{2}{3} = \frac{2}{9} \end{aligned}$$

Therefore, the probability mass function of Z is:

$$p_Z(z) = \begin{cases} 1/9 & \text{if } z = 1 \\ 1/3 & \text{if } z = 2 \\ 1/3 & \text{if } z = 3 \\ 2/9 & \text{if } z = 4 \\ 0 & \text{otherwise} \end{cases}$$

Part IV

Probability distributions

Chapter 42

Bernoulli distribution

Suppose you perform an experiment with two possible outcomes: either success or failure. Success happens with probability p, while failure happens with probability $1-p$. A random variable that takes value 1 in case of success and 0 in case of failure is called a Bernoulli random variable (alternatively, it is said to have a Bernoulli distribution).

42.1 Definition

Bernoulli random variables are characterized as follows:

Definition 243 *Let X be a discrete random variable. Let its support be*

$$R_X = \{0, 1\}$$

*Let $p \in (0, 1)$. We say that X has a **Bernoulli distribution** with parameter p if its probability mass function*[1] *is*

$$p_X(x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \\ 0 & \text{if } x \notin R_X \end{cases}$$

Note that, by the above definition, any indicator function[2] is a Bernoulli random variable.

The following is a proof that $p_X(x)$ is a legitimate probability mass function[3]:
Proof. Non-negativity is obvious. We need to prove that the sum of $p_X(x)$ over its support equals 1. This is proved as follows:

$$\begin{aligned} \sum_{x \in R_X} p_X(x) &= p_X(1) + p_X(0) \\ &= p + (1 - p) = 1 \end{aligned}$$

■

[1] See p. 106.
[2] See p. 197.
[3] See p. 247.

42.2 Expected value

The expected value of a Bernoulli random variable X is

$$\mathrm{E}[X] = p$$

Proof. It can be derived as follows:

$$\begin{aligned} \mathrm{E}[X] &= \sum_{x \in R_X} x p_X(x) \\ &= 1 \cdot p_X(1) + 0 \cdot p_X(0) \\ &= 1 \cdot p + 0 \cdot (1-p) = p \end{aligned}$$

■

42.3 Variance

The variance of a Bernoulli random variable X is

$$\mathrm{Var}[X] = p(1-p)$$

Proof. It can be derived thanks to the usual formula for computing the variance[4]:

$$\begin{aligned} \mathrm{E}\left[X^2\right] &= \sum_{x \in R_X} x^2 p_X(x) \\ &= 1^2 \cdot p_X(1) + 0^2 \cdot p_X(0) \\ &= 1 \cdot p + 0 \cdot (1-p) = p \\ \mathrm{E}[X]^2 &= p^2 \\ \mathrm{Var}[X] &= \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 = p - p^2 = p(1-p) \end{aligned}$$

■

42.4 Moment generating function

The moment generating function of a Bernoulli random variable X is defined for any $t \in \mathbb{R}$:

$$M_X(t) = 1 - p + p\exp(t)$$

Proof. Using the definition of moment generating function:

$$\begin{aligned} M_X(t) &= \mathrm{E}[\exp(tX)] \\ &= \sum_{x \in R_X} \exp(tx) p_X(x) \\ &= \exp(t \cdot 1) \cdot p_X(1) + \exp(t \cdot 0) \cdot p_X(0) \\ &= \exp(t) \cdot p + 1 \cdot (1-p) \\ &= 1 - p + p\exp(t) \end{aligned}$$

Obviously, the above expected value exists for any $t \in \mathbb{R}$. ■

[4] $\mathrm{Var}[X] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2$. See p. 156.

42.5 Characteristic function

The characteristic function of a Bernoulli random variable X is

$$\varphi_X(t) = 1 - p + p\exp(it)$$

Proof. Using the definition of characteristic function:

$$\begin{aligned}\varphi_X(t) &= \mathrm{E}[\exp(itX)] \\ &= \sum_{x \in R_X} \exp(itx)\, p_X(x) \\ &= \exp(it \cdot 1) \cdot p_X(1) + \exp(it \cdot 0) \cdot p_X(0) \\ &= \exp(it) \cdot p + 1 \cdot (1 - p) \\ &= 1 - p + p\exp(it)\end{aligned}$$

■

42.6 Distribution function

The distribution function of a Bernoulli random variable X is

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - p & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

Proof. Remember the definition of distribution function:

$$F_X(x) = \mathrm{P}(X \leq x)$$

and the fact that X can take either value 0 or value 1. If $x < 0$, then $\mathrm{P}(X \leq x) = 0$, because X can not take values strictly smaller than 0. If $0 \leq x < 1$, then $\mathrm{P}(X \leq x) = 1 - p$, because 0 is the only value strictly smaller than 1 that X can take. Finally, if $x \geq 1$, then $\mathrm{P}(X \leq x) = 1$, because all values X can take are smaller than or equal to 1. ■

42.7 More details

In the following subsections you can find more details about the Bernoulli distribution.

42.7.1 Relation to the binomial distribution

A sum of independent Bernoulli random variables is a binomial random variable. This is discussed and proved in the lecture entitled *Binomial distribution* (p. 341).

42.8 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Let X and Y be two independent Bernoulli random variables with parameter p. Derive the probability mass function of their sum:

$$Z = X + Y$$

Solution

The probability mass function of X is

$$p_X(x) = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases}$$

The probability mass function of Y is

$$p_Y(y) = \begin{cases} p & \text{if } y = 1 \\ 1-p & \text{if } y = 0 \\ 0 & \text{otherwise} \end{cases}$$

The support of Z (the set of values Z can take) is

$$R_Y = \{0, 1, 2\}$$

The formula for the probability mass function of a sum of two independent variables is[5]

$$p_Z(z) = \sum_{y \in R_Y} p_X(z-y)\, p_Y(y)$$

where R_Y is the support of Y. When $z = 0$, the formula gives:

$$\begin{aligned} p_Z(0) &= \sum_{y \in R_Y} p_X(-y)\, p_Y(y) \\ &= p_X(-0)\, p_Y(0) + p_X(-1)\, p_Y(1) \\ &= (1-p)(1-p) + 0 \cdot p = (1-p)^2 \end{aligned}$$

When $z = 1$, the formula gives:

$$\begin{aligned} p_Z(1) &= \sum_{y \in R_Y} p_X(1-y)\, p_Y(y) \\ &= p_X(1-0)\, p_Y(0) + p_X(1-1)\, p_Y(1) \\ &= p \cdot (1-p) + (1-p) \cdot p = 2p(1-p) \end{aligned}$$

When $z = 2$, the formula gives:

$$\begin{aligned} p_Z(2) &= \sum_{y \in R_Y} p_X(2-y)\, p_Y(y) \\ &= p_X(2-0)\, p_Y(0) + p_X(2-1)\, p_Y(1) \\ &= 0 \cdot (1-p) + p \cdot p = p^2 \end{aligned}$$

Therefore, the probability mass function of Z is

$$p_Z(z) = \begin{cases} (1-p)^2 & \text{if } z = 0 \\ 2p(1-p) & \text{if } z = 1 \\ p^2 & \text{if } z = 2 \\ 0 & \text{otherwise} \end{cases}$$

[5] See p. 325.

Exercise 2

Let X be a Bernoulli random variable with parameter $p = 1/2$. Find its tenth moment.

Solution

The moment generating function of X is

$$M_X(t) = \frac{1}{2} + \frac{1}{2}\exp(t)$$

The tenth moment of X is equal to the tenth derivative of its moment generating function[6], evaluated at $t = 0$:

$$\mu_X(10) = \mathrm{E}\left[X^{10}\right] = \left.\frac{d^{10}M_X(t)}{dt^{10}}\right|_{t=0}$$

But

$$\begin{aligned}
\frac{dM_X(t)}{dt} &= \frac{1}{2}\exp(t) \\
\frac{d^2M_X(t)}{dt^2} &= \frac{1}{2}\exp(t) \\
&\vdots \\
\frac{d^{10}M_X(t)}{dt^{10}} &= \frac{1}{2}\exp(t)
\end{aligned}$$

so that:

$$\begin{aligned}
\mu_X(10) &= \left.\frac{d^{10}M_X(t)}{dt^{10}}\right|_{t=0} \\
&= \frac{1}{2}\exp(0) = \frac{1}{2}
\end{aligned}$$

[6] See p. 290.

Chapter 43

Binomial distribution

Consider an experiment having two possible outcomes: either success or failure. Suppose the experiment is repeated several times and the repetitions are independent of each other. The total number of experiments where the outcome turns out to be a success is a random variable whose distribution is called binomial distribution. The distribution has two parameters: the number n of repetitions of the experiment, and the probability p of success of an individual experiment.

A binomial distribution can be seen as a sum of mutually independent Bernoulli random variables[1] that take value 1 in case of success of the experiment and value 0 otherwise. This connection between the binomial and Bernoulli distributions will be illustrated in detail in the remainder of this lecture and will be used to prove several properties of the binomial distribution.

43.1 Definition

The binomial distribution is characterized as follows.

Definition 244 *Let X be a discrete random variable. Let $n \in \mathbb{N}$ and $p \in (0,1)$. Let the support of X be*[2]

$$R_X = \{0, 1, \ldots, n\}$$

*We say that X has a **binomial distribution** with parameters n and p if its probability mass function*[3] *is*

$$p_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{if } x \in R_X \\ 0 & \text{if } x \notin R_X \end{cases}$$

where $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ is the binomial coefficient[4].

The two parameters of the distribution are the **number of experiments** n and the **probability of success** p of an individual experiment.

The following is a proof that $p_X(x)$ is a legitimate probability mass function[5].

[1] See p. 335.
[2] In other words, R_X is the set of the first n natural numbers and 0.
[3] See p. 106.
[4] See p. 22.
[5] See p. 247.

Proof. Non-negativity is obvious. We need to prove that the sum of $p_X(x)$ over the support of X equals 1. This is proved as follows:

$$\sum_{x \in R_X} p_X(x) = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} = [p + (1-p)]^n = 1^n = 1$$

where we have used the formula for binomial expansions[6]

$$(a+b)^n = \sum_{x=0}^{n} \binom{n}{x} a^x b^{n-x}$$

■

43.2 Relation to the Bernoulli distribution

The binomial distribution is intimately related to the Bernoulli distribution. The following propositions show how.

Proposition 245 *A random variable has a binomial distribution with parameters n and p, with $n = 1$, if and only if it has a Bernoulli distribution with parameter p.*

Proof. We demonstrate that the two distributions are equivalent by showing that they have the same probability mass function. The probability mass function of a binomial distribution with parameters n and p, with $n = 1$, is

$$p_X(x) = \begin{cases} \binom{1}{x} p^x (1-p)^{1-x} & \text{if } x \in \{0,1\} \\ 0 & \text{if } x \notin \{0,1\} \end{cases}$$

but

$$p_X(0) = \binom{1}{0} p^0 (1-p)^{1-0} = \frac{1!}{0!1!}(1-p) = 1-p$$

and

$$p_X(1) = \binom{1}{1} p^1 (1-p)^{1-1} = \frac{1!}{1!0!} p = p$$

Therefore, the probability mass function can be written as

$$p_X(x) = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases}$$

which is the probability mass function of a Bernoulli random variable. ■

Proposition 246 *A random variable has a binomial distribution with parameters n and p if and only if it can be written as a sum of n jointly independent Bernoulli random variables with parameter p.*

[6] See p. 25.

Proof. We prove it by induction. So, we have to prove that it is true for $n = 1$ and for a generic n, given that it is true for $n - 1$. For $n = 1$, it has been proved in Proposition 245. Now, suppose the claim is true for a generic $n - 1$. We have to verify that Y_n is a binomial random variable, where

$$Y_n = X_1 + X_2 + \ldots + X_n$$

and X_1, X_2, ..., X_n are independent Bernoulli random variables. Since the claim is true for $n - 1$, this is tantamount to verifying that

$$Y_n = Y_{n-1} + X_n$$

is a binomial random variable, where Y_{n-1} has a binomial distribution with parameters $n-1$ and p. By performing a convolution[7], we can compute the probability mass function of Y_n:

$$\begin{aligned}
& p_{Y_n}(y_n) \\
= & \sum_{y_{n-1} \in R_{Y_{n-1}}} p_{X_n}(y_n - y_{n-1})\, p_{Y_{n-1}}(y_{n-1}) \\
= & \sum_{y_{n-1} \in R_{Y_{n-1}}} I\left((y_n - y_{n-1}) \in \{0,1\}\right) p^{y_n - y_{n-1}} (1-p)^{1-y_n+y_{n-1}} \\
& \cdot \binom{n-1}{y_{n-1}} p^{y_{n-1}} (1-p)^{n-1-y_{n-1}} \\
= & \sum_{y_{n-1} \in R_{Y_{n-1}}} I\left(y_n \in \{y_{n-1}, y_{n-1}+1\}\right) \\
& \cdot \binom{n-1}{y_{n-1}} p^{y_n - y_{n-1} + y_{n-1}} (1-p)^{1-y_n+y_{n-1}+n-1-y_{n-1}} \\
= & \sum_{y_{n-1} \in R_{Y_{n-1}}} I\left(y_n \in \{y_{n-1}, y_{n-1}+1\}\right) \binom{n-1}{y_{n-1}} p^{y_n} (1-p)^{n-y_n} \\
= & \; p^{y_n} (1-p)^{n-y_n} \sum_{y_{n-1} \in R_{Y_{n-1}}} I\left(y_n \in \{y_{n-1}, y_{n-1}+1\}\right) \binom{n-1}{y_{n-1}}
\end{aligned}$$

If $1 \leq y_n \leq n-1$, then

$$\sum_{y_{n-1} \in R_{Y_{n-1}}} I\left(y_n \in \{y_{n-1}, y_{n-1}+1\}\right) \binom{n-1}{y_{n-1}} = \binom{n-1}{y_n} + \binom{n-1}{y_n - 1} = \binom{n}{y_n}$$

where the last equality is the recursive formula for binomial coefficients[8]. If $y_n = 0$, then

$$\begin{aligned}
& \sum_{y_{n-1} \in R_{Y_{n-1}}} I\left(y_n \in \{y_{n-1}, y_{n-1}+1\}\right) \binom{n-1}{y_{n-1}} \\
= & \; \binom{n-1}{y_n} = \binom{n-1}{0} = 1 = \binom{n}{0} = \binom{n}{y_n}
\end{aligned}$$

[7] See p. 326.
[8] See p. 25.

Finally, if $y_n = n$, then

$$\begin{aligned} & \sum_{y_{n-1} \in R_{Y_{n-1}}} I\left(y_n \in \{y_{n-1}, y_{n-1}+1\}\right) \binom{n-1}{y_{n-1}} \\ = & \binom{n-1}{y_n - 1} = \binom{n-1}{n-1} = 1 = \binom{n}{n} = \binom{n}{y_n} \end{aligned}$$

Therefore, for $y_n \in R_{Y_n}$, we have

$$\sum_{y_{n-1} \in R_{Y_{n-1}}} I\left(y_n \in \{y_{n-1}, y_{n-1}+1\}\right) \binom{n-1}{y_{n-1}} = \binom{n}{y_n}$$

and

$$p_{Y_n}(y_n) = \begin{cases} \binom{n}{y_n} p^{y_n} (1-p)^{n-y_n} & \text{if } y_n \in R_{Y_n} \\ 0 & \text{otherwise} \end{cases}$$

which is the probability mass function of a binomial random variable with parameters n and p. This completes the proof. ■

43.3 Expected value

The expected value of a binomial random variable X is

$$\mathrm{E}[X] = np$$

Proof. It can be derived as follows:

$$\begin{aligned} & & \mathrm{E}[X] \\ \boxed{\text{A}} & = & \mathrm{E}\left[\sum_{i=1}^{n} Y_i\right] \\ \boxed{\text{B}} & = & \sum_{i=1}^{n} \mathrm{E}[Y_i] \\ \boxed{\text{C}} & = & \sum_{i=1}^{n} p \\ & = & np \end{aligned}$$

where: in step A we have used the fact that X can be represented as a sum of n independent Bernoulli random variables $Y_1, \ldots, Y_n$; in step B we have used the linearity of the expected value; in step C we have used the formula for the expected value of a Bernoulli random variable[9]. ■

43.4 Variance

The variance of a binomial random variable X is

$$\mathrm{Var}[X] = np(1-p)$$

[9] See p. 336.

Proof. It can be derived as follows:

$$\begin{aligned} & & \operatorname{Var}[X] \\ \boxed{\text{A}} & = & \operatorname{Var}\left[\sum_{i=1}^{n} Y_i\right] \\ \boxed{\text{B}} & = & \sum_{i=1}^{n} \operatorname{Var}[Y_i] \\ \boxed{\text{C}} & = & \sum_{i=1}^{n} p(1-p) \\ & = & np(1-p) \end{aligned}$$

where: in step A we have used the fact that X can be represented as a sum of n independent Bernoulli random variables $Y_1, \ldots, Y_n$; in step B we have used the formula for the variance of the sum of jointly independent random variables; in step C we have used the formula for the variance of a Bernoulli random variable[10]. ■

43.5 Moment generating function

The moment generating function of a binomial random variable X is defined for any $t \in \mathbb{R}$:

$$M_X(t) = (1 - p + p\exp(t))^n$$

Proof. This is proved as follows:

$$\begin{aligned} & & M_X(t) \\ \boxed{\text{A}} & = & \mathrm{E}[\exp(tX)] \\ \boxed{\text{B}} & = & \mathrm{E}[\exp(t(Y_1 + \ldots + Y_n))] \\ & = & \mathrm{E}[\exp(tY_1) \cdot \ldots \cdot \exp(tY_n)] \\ \boxed{\text{C}} & = & \mathrm{E}[\exp(tY_1)] \cdot \ldots \cdot \mathrm{E}[\exp(tY_n)] \\ \boxed{\text{D}} & = & M_{Y_1}(t) \cdot \ldots \cdot M_{Y_n}(t) \\ \boxed{\text{E}} & = & (1 - p + p\exp(t)) \cdot \ldots \cdot (1 - p + p\exp(t)) \\ & = & (1 - p + p\exp(t))^n \end{aligned}$$

where: in step A we have used the definition of moment generating function; in step B we have used the fact that X can be represented as a sum of n independent Bernoulli random variables $Y_1, \ldots, Y_n$; in step C we have used the fact that $Y_1, \ldots, Y_n$ are jointly independent; in step D we have used the definition of moment generating function of $Y_1, \ldots, Y_n$; in step E we have used the formula for the moment generating function of a Bernoulli random variable[11]. Since the moment generating function of a Bernoulli random variable exists for any $t \in \mathbb{R}$, also the moment generating function of a binomial random variable exists for any $t \in \mathbb{R}$. ■

[10] See p. 336.
[11] See p. 336.

43.6 Characteristic function

The characteristic function of a binomial random variable X is

$$\varphi_X(t) = (1 - p + p\exp(it))^n$$

Proof. Similar to the previous proof:

$$\begin{aligned}
& \varphi_X(t) \\
\boxed{\text{A}} \quad &= \mathrm{E}[\exp(itX)] \\
\boxed{\text{B}} \quad &= \mathrm{E}[\exp(it(Y_1 + \ldots + Y_n))] \\
&= \mathrm{E}[\exp(itY_1) \cdot \ldots \cdot \exp(itY_n)] \\
\boxed{\text{C}} \quad &= \mathrm{E}[\exp(itY_1)] \cdot \ldots \cdot \mathrm{E}[\exp(itY_n)] \\
\boxed{\text{D}} \quad &= \varphi_{Y_1}(t) \cdot \ldots \cdot \varphi_{Y_n}(t) \\
\boxed{\text{E}} \quad &= (1 - p + p\exp(it)) \cdot \ldots \cdot (1 - p + p\exp(it)) \\
&= (1 - p + p\exp(it))^n
\end{aligned}$$

where: in step A we have used the definition of characteristic function; in step B we have used the fact that X can be represented as a sum of n independent Bernoulli random variables $Y_1, \ldots, Y_n$; in step C we have used the fact that $Y_1, \ldots, Y_n$ are jointly independent; in step D we have used the definition of characteristic function of $Y_1, \ldots, Y_n$; in step E we have used the formula for the characteristic function of a Bernoulli random variable[12]. ■

43.7 Distribution function

The distribution function of a binomial random variable X is

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ \sum_{s=0}^{\lfloor x \rfloor} \binom{n}{s} p^s (1-p)^{n-s} & \text{if } 0 \leq x \leq n \\ 1 & \text{if } x > n \end{cases}$$

where $\lfloor x \rfloor$ is the floor of x, i.e. the largest integer not greater than x.

Proof. For $x < 0$, $F_X(x) = 0$, because X cannot be smaller than 0. For $x > n$, $F_X(x) = 1$, because X is always smaller than or equal to n. For $0 \leq x \leq n$:

$$\begin{aligned}
& F_X(x) \\
\boxed{\text{A}} \quad &= \mathrm{P}(X \leq x) \\
\boxed{\text{B}} \quad &= \sum_{s=0}^{\lfloor x \rfloor} \mathrm{P}(X = s) \\
\boxed{\text{C}} \quad &= \sum_{s=0}^{\lfloor x \rfloor} p_X(s)
\end{aligned}$$

[12] See p. 337.

$$= \sum_{s=0}^{\lfloor x \rfloor} \binom{n}{s} p^s (1-p)^{n-s}$$

where: in step A we have used the definition of distribution function; in step B we have used the fact that X can take only integer values; in step C we have used the definition of probability mass function of X. ■

Values of $F_X(x)$ are usually computed by means of computer algorithms. For example, the MATLAB command

```
binocdf(x,n,p)
```

returns the value of the distribution function at the point `x` when the parameters of the distribution are `n` and `p`.

43.8 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Suppose you independently flip a coin 4 times and the outcome of each toss can be either head (with probability 1/2) or tails (also with probability 1/2). What is the probability of obtaining exactly 2 tails?

Solution

Denote by X the number of times the outcome is tails (out of the 4 tosses). X has a binomial distribution with parameters $n = 4$ and $p = 1/2$. The probability of obtaining exactly 2 tails can be computed from the probability mass function of X as follows:

$$\begin{aligned} p_X(2) &= \binom{n}{2} p^2 (1-p)^{n-2} = \binom{4}{2} \left(\frac{1}{2}\right)^2 \left(1-\frac{1}{2}\right)^{4-2} \\ &= \frac{4!}{2!2!} \frac{1}{4} \frac{1}{4} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 2 \cdot 1} \frac{1}{16} = \frac{6}{16} = \frac{3}{8} \end{aligned}$$

Exercise 2

Suppose you independently throw a dart 10 times. Each time you throw a dart, the probability of hitting the target is 3/4. What is the probability of hitting the target less than 5 times (out of the 10 total times you throw a dart)?

Solution

Denote by X the number of times you hit the target. X has a binomial distribution with parameters $n = 10$ and $p = 3/4$. The probability of hitting the target less than 5 times can be computed from the distribution function of X as follows:

$$\mathrm{P}(X < 5) = \mathrm{P}(X \leq 4) = F_X(4)$$

$$\begin{aligned} &= \sum_{s=0}^{4} \binom{n}{s} p^s (1-p)^{n-s} \\ &= \sum_{s=0}^{4} \binom{10}{s} \left(\frac{3}{4}\right)^s \left(\frac{1}{4}\right)^{10-s} \simeq 0.0197 \end{aligned}$$

where F_X is the distribution function of X and the value of $F_X(4)$ can be calculated with a computer algorithm, for example with the MATLAB command

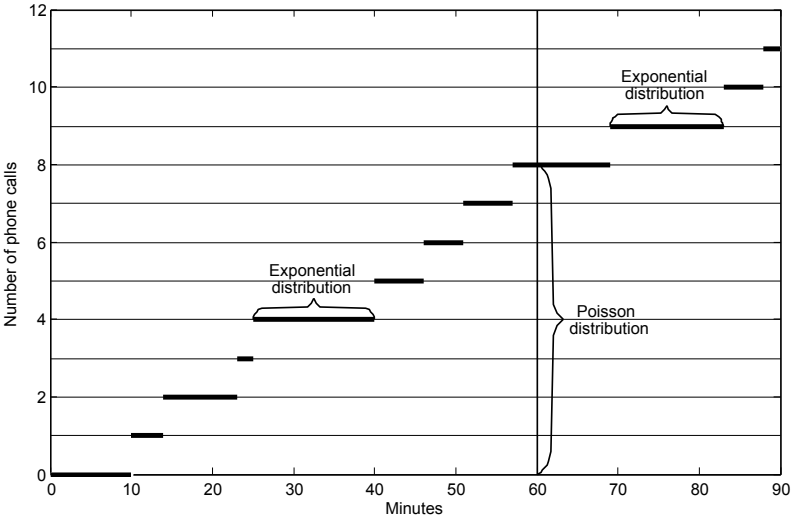
```
binocdf(4,10,3/4)
```

Chapter 44

Poisson distribution

The Poisson distribution is related to the exponential distribution[1]. Suppose an event can occur several times within a given unit of time. When the total number of occurrences of the event is unknown, we can think of it as a random variable. This random variable has a Poisson distribution if and only if the time elapsed between two successive occurrences of the event has an exponential distribution and it is independent of previous occurrences.

A classical example of a random variable having a Poisson distribution is the number of phone calls received by a call center. If the time elapsed between two successive phone calls has an exponential distribution and it is independent of the time of arrival of the previous calls, then the total number of calls received in one hour has a Poisson distribution.



The concept is illustrated by the plot above, where the number of phone calls received is plotted as a function of time. The graph of the function makes an upward jump each time a phone call arrives. The time elapsed between two successive phone calls is equal to the length of each horizontal segment and it has an

[1] See p. 365.

exponential distribution. The number of calls received in 60 minutes is equal to the length of the segment highlighted by the vertical curly brace and it has a Poisson distribution.

The following sections provide a more formal treatment of the main characteristics of the Poisson distribution.

44.1 Definition

The Poisson distribution is characterized as follows:

Definition 247 *Let X be a discrete random variable. Let its support be the set of non-negative integer numbers:*

$$R_X = \mathbb{Z}_+$$

*Let $\lambda \in (0, \infty)$. We say that X has a **Poisson distribution** with parameter λ if its probability mass function[2] is*

$$p_X(x) = \begin{cases} \exp(-\lambda)\frac{1}{x!}\lambda^x & \text{if } x \in R_X \\ 0 & \text{if } x \notin R_X \end{cases}$$

where $x!$ is the factorial[3] of x.

44.2 Relation to the exponential distribution

The relation between the Poisson distribution and the exponential distribution is summarized by the following proposition:

Proposition 248 *The number of occurrences of an event within a unit of time has a Poisson distribution with parameter λ if and only if the time elapsed between two successive occurrences of the event has an exponential distribution with parameter λ and it is independent of previous occurrences.*

Proof. Denote by:

τ_1 time elapsed before first occurrence
τ_2 time elapsed between first and second occurrence
$\vdots$
τ_n time elapsed between $(n-1)$-th and n-th occurrence
$\vdots$

and by X the number of occurrences of the event. Since $X \geq x$ if and only if $\tau_1 + \ldots + \tau_x \leq 1$ (convince yourself of this fact), the proposition is true if and only if:

$$\mathrm{P}(X \geq x) = \mathrm{P}(\tau_1 + \ldots + \tau_x \leq 1)$$

for any $x \in R_X$. To verify that the equality holds, we need to separately compute the two probabilities. We start with:

$$\mathrm{P}(\tau_1 + \ldots + \tau_x \leq 1)$$

[2] See p. 106.
[3] See p. 10.

Denote by Z the sum of waiting times:

$$Z = \tau_1 + \ldots + \tau_x$$

Since the sum of independent exponential random variables[4] with common parameter λ is a Gamma random variable[5] with parameters $2x$ and $\frac{x}{\lambda}$, then Z is a Gamma random variable with parameters $2x$ and $\frac{x}{\lambda}$, i.e. its probability density function is

$$f_Z(z) = \begin{cases} cz^{x-1}\exp(-\lambda z) & \text{if } z \in [0,\infty) \\ 0 & \text{if } z \notin [0,\infty) \end{cases}$$

where

$$c = \frac{\lambda^x}{\Gamma(x)} = \frac{\lambda^x}{(x-1)!}$$

and the last equality stems from the fact that we are considering only integer values of x. We need to integrate the density function to compute the probability that Z is less than 1:

$$\begin{aligned} \mathrm{P}(\tau_1 + \ldots + \tau_x \leq 1) &= \mathrm{P}(Z \leq 1) \\ &= \int_{-\infty}^{1} f_Z(z)\,dz \\ &= \int_0^1 cz^{x-1}\exp(-\lambda z)\,dz \\ &= c\int_0^1 z^{x-1}\exp(-\lambda z)\,dz \end{aligned}$$

The last integral can be computed integrating by parts $x-1$ times:

$$\begin{aligned} & \int_0^1 z^{x-1}\exp(-\lambda z)\,dz \\ = & \left[-\frac{1}{\lambda}z^{x-1}\exp(-\lambda z)\right]_0^1 + \int_0^1 (x-1)\,z^{x-2}\frac{1}{\lambda}\exp(-\lambda z)\,dz \\ = & -\frac{1}{\lambda}\exp(-\lambda) + (x-1)\frac{1}{\lambda}\int_0^1 z^{x-2}\exp(-\lambda z)\,dz \\ = & -\frac{1}{\lambda}\exp(-\lambda) + (x-1)\frac{1}{\lambda}\left\{\left[-\frac{1}{\lambda}z^{x-2}\exp(-\lambda z)\right]_0^1\right. \\ & \left. + \int_0^1 (x-2)\,z^{x-3}\frac{1}{\lambda}\exp(-\lambda z)\,dz\right\} \\ = & -\frac{1}{\lambda}\exp(-\lambda) - (x-1)\frac{1}{\lambda^2}\exp(-\lambda) \\ & + (x-1)(x-2)\frac{1}{\lambda^2}\int_0^1 z^{x-3}\exp(-\lambda z)\,dz \\ = & \ldots \\ = & -\sum_{i=1}^{x-1}\frac{(x-1)!}{(x-i)!}\frac{1}{\lambda^i}\exp(-\lambda) + \frac{(x-1)!}{1}\frac{1}{\lambda^{x-1}}\int_0^1 \exp(-\lambda z)\,dz \end{aligned}$$

[4] See p. 372.
[5] See p. 397.

$$
\begin{aligned}
&= -\sum_{i=1}^{x-1} \frac{(x-1)!}{(x-i)!} \frac{1}{\lambda^i} \exp(-\lambda) + \frac{(x-1)!}{\lambda^{x-1}} \left[-\frac{1}{\lambda} \exp(-\lambda z) \right]_0^1 \\
&= -\sum_{i=1}^{x-1} \frac{(x-1)!}{(x-i)!} \frac{1}{\lambda^i} \exp(-\lambda) - \frac{(x-1)!}{\lambda^x} \exp(-\lambda) + \frac{(x-1)!}{\lambda^x}
\end{aligned}
$$

Multiplying by c, we obtain:

$$
\begin{aligned}
& c \int_0^1 z^{x-1} \exp(-\lambda z)\, dz \\
= & \frac{\lambda^x}{(x-1)!} \int_0^1 z^{x-1} \exp(-\lambda z)\, dz \\
= & -\sum_{i=1}^{x-1} \frac{\lambda^{x-i}}{(x-i)!} \exp(-\lambda) - \exp(-\lambda) + 1 \\
= & 1 - \sum_{i=1}^{x} \frac{\lambda^{x-i}}{(x-i)!} \exp(-\lambda) \\
= & 1 - \sum_{j=0}^{x-1} \frac{\lambda^j}{j!} \exp(-\lambda)
\end{aligned}
$$

Thus, we have obtained:

$$
\mathrm{P}(\tau_1 + \ldots + \tau_x \leq 1) = 1 - \sum_{j=0}^{x-1} \frac{\lambda^j}{j!} \exp(-\lambda)
$$

Now, we need to compute the probability that X is greater than or equal to x:

$$
\begin{aligned}
\mathrm{P}(X \geq x) &= 1 - \mathrm{P}(X < x) \\
&= 1 - \mathrm{P}(X \leq x - 1) \\
&= 1 - \sum_{j=0}^{x-1} \mathrm{P}(X = j) \\
&= 1 - \sum_{j=0}^{x-1} p_X(j) \\
&= 1 - \sum_{j=0}^{x-1} \frac{\lambda^j}{j!} \exp(-\lambda) \\
&= \mathrm{P}(\tau_1 + \ldots + \tau_x \leq 1)
\end{aligned}
$$

which is exactly what we needed to prove. ■

44.3 Expected value

The expected value of a Poisson random variable X is:

$$
\mathrm{E}[X] = \lambda
$$

Proof. It can be derived as follows:

$$\begin{aligned}
\mathrm{E}[X] &= \sum_{x \in R_X} x p_X(x) \\
&= \sum_{x=0}^{\infty} x \exp(-\lambda) \frac{1}{x!} \lambda^x \\
\boxed{\mathrm{A}} &= 0 + \sum_{x=1}^{\infty} x \exp(-\lambda) \frac{1}{x!} \lambda^x \\
\boxed{\mathrm{B}} &= \sum_{y=0}^{\infty} (y+1) \exp(-\lambda) \frac{1}{(y+1)!} \lambda^{y+1} \\
\boxed{\mathrm{C}} &= \sum_{y=0}^{\infty} (y+1) \exp(-\lambda) \frac{1}{(y+1)\, y!} \lambda \lambda^y \\
&= \lambda \sum_{y=0}^{\infty} \exp(-\lambda) \frac{1}{y!} \lambda^y \\
\boxed{\mathrm{D}} &= \lambda \sum_{y=0}^{\infty} p_Y(y) \\
\boxed{\mathrm{E}} &= \lambda
\end{aligned}$$

where: in step $\boxed{\mathrm{A}}$ we have used the fact that the first term of the sum is zero, because $x = 0$; in step $\boxed{\mathrm{B}}$ we have made a change of variable $y = x - 1$; in step $\boxed{\mathrm{C}}$ we have used the fact that $(y+1)! = (y+1)\, y!$; in step $\boxed{\mathrm{D}}$ we have defined

$$p_Y(y) = \exp(-\lambda) \frac{1}{y!} \lambda^y$$

where $p_Y(y)$ is the probability mass function of a Poisson random variable with parameter λ; in step $\boxed{\mathrm{E}}$ we have used the fact that the sum of a probability mass function over its support equals 1. ■

44.4 Variance

The variance of a Poisson random variable X is:

$$\mathrm{Var}[X] = \lambda$$

Proof. It can be derived thanks to the usual formula for computing the variance[6]:

$$\begin{aligned}
\mathrm{E}\left[X^2\right] &= \sum_{x \in R_X} x^2 p_X(x) \\
&= \sum_{x=0}^{\infty} x^2 \exp(-\lambda) \frac{1}{x!} \lambda^x \\
\boxed{\mathrm{A}} &= 0 + \sum_{x=1}^{\infty} x^2 \exp(-\lambda) \frac{1}{x!} \lambda^x
\end{aligned}$$

[6] $\mathrm{Var}[X] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2$. See p. 156.

$$\begin{aligned}
\boxed{\text{B}} &= \sum_{y=0}^{\infty} (y+1)^2 \exp(-\lambda) \frac{1}{(y+1)!} \lambda^{y+1} \\
\boxed{\text{C}} &= \sum_{y=0}^{\infty} (y+1)^2 \exp(-\lambda) \frac{1}{(y+1)\, y!} \lambda \lambda^{y} \\
&= \lambda \sum_{y=0}^{\infty} (y+1) \exp(-\lambda) \frac{1}{y!} \lambda^{y} \\
\boxed{\text{D}} &= \lambda \sum_{y=0}^{\infty} (y+1)\, p_Y(y) \\
&= \lambda \left\{ \sum_{y=0}^{\infty} y p_Y(y) + \sum_{y=0}^{\infty} p_Y(y) \right\} \\
\boxed{\text{E}} &= \lambda \{ \mathrm{E}[Y] + 1 \} \\
\boxed{\text{F}} &= \lambda \{ \lambda + 1 \} \\
&= \lambda^2 + \lambda
\end{aligned}$$

where: in step A we have used the fact that the first term of the sum is zero, because $x = 0$; in step B we have made a change of variable $y = x - 1$; in step C we have used the fact that $(y+1)! = (y+1)\, y!$; in step D we have defined

$$p_Y(y) = \exp(-\lambda) \frac{1}{y!} \lambda^{y}$$

where $p_Y(y)$ is the probability mass function of a Poisson random variable with parameter λ; in step E we have used the fact that the sum of a probability mass function over its support equals 1; in step F we have used the fact that the expected value of a Poisson random variable with parameter λ is λ. Finally,

$$\mathrm{E}[X]^2 = \lambda^2$$

and

$$\begin{aligned}
\mathrm{Var}[X] &= \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 \\
&= \lambda^2 + \lambda - \lambda^2 = \lambda
\end{aligned}$$

■

44.5 Moment generating function

The moment generating function of a Poisson random variable X is defined for any $t \in \mathbb{R}$:

$$M_X(t) = \exp(\lambda [\exp(t) - 1])$$

Proof. Using the definition of moment generating function:

$$M_X(t) = \mathrm{E}[\exp(tX)]$$

$$\begin{aligned} &= \sum_{x \in R_X} \exp(tx)\, p_X(x) \\ &= \sum_{x=0}^{\infty} [\exp(t)]^x \exp(-\lambda) \frac{1}{x!} \lambda^x \\ &= \exp(-\lambda) \sum_{x=0}^{\infty} \frac{(\lambda \exp(t))^x}{x!} \\ &= \exp(-\lambda) \exp(\lambda \exp(t)) \\ &= \exp(\lambda [\exp(t) - 1]) \end{aligned}$$

where

$$\exp(\lambda \exp(t)) = \sum_{x=0}^{\infty} \frac{(\lambda \exp(t))^x}{x!}$$

is the usual Taylor series expansion of the exponential function. Furthermore, since the series converges for any value of t, the moment generating function of a Poisson random variable exists for any $t \in \mathbb{R}$. ■

44.6 Characteristic function

The characteristic function of a Poisson random variable X is:

$$\varphi_X(t) = \exp(\lambda [\exp(it) - 1])$$

Proof. Using the definition of characteristic function:

$$\begin{aligned} \varphi_X(t) &= \mathrm{E}[\exp(itX)] \\ &= \sum_{x \in R_X} \exp(itx)\, p_X(x) \\ &= \sum_{x \in R_X} [\exp(it)]^x \exp(-\lambda) \frac{1}{x!} \lambda^x \\ &= \exp(-\lambda) \sum_{x=0}^{\infty} \frac{(\lambda \exp(it))^x}{x!} \\ &= \exp(-\lambda) \exp(\lambda \exp(it)) \\ &= \exp(\lambda [\exp(it) - 1]) \end{aligned}$$

where:

$$\exp(\lambda \exp(it)) = \sum_{x=0}^{\infty} \frac{(\lambda \exp(it))^x}{x!}$$

is the usual Taylor series expansion of the exponential function (note that the series converges for any value of t). ■

44.7 Distribution function

The distribution function of a Poisson random variable X is:

$$F_X(x) = \begin{cases} \exp(-\lambda) \sum_{s=0}^{\lfloor x \rfloor} \frac{1}{s!} \lambda^s & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\lfloor x \rfloor$ is the floor of x, i.e. the largest integer not greater than x.

Proof. Using the definition of distribution function:

$$\begin{aligned} & F_X(x) \\ \boxed{\text{A}} \quad &= \mathrm{P}(X \leq x) \\ \boxed{\text{B}} \quad &= \sum_{s=0}^{\lfloor x \rfloor} \mathrm{P}(X = s) \\ \boxed{\text{C}} \quad &= \sum_{s=0}^{\lfloor x \rfloor} p_X(s) \\ &= \sum_{s=0}^{\lfloor x \rfloor} \exp(-\lambda) \frac{1}{s!} \lambda^s \\ &= \exp(-\lambda) \sum_{s=0}^{\lfloor x \rfloor} \frac{1}{s!} \lambda^s \end{aligned}$$

where: in step A we have used the definition of distribution function; in step B we have used the fact that X can take only positive integer values; in step C we have used the definition of probability mass function of X. ■

Values of $F_X(x)$ are usually computed by computer algorithms. For example, the MATLAB command:

```
poisscdf(x,lambda)
```

returns the value of the distribution function at the point `x` when the parameter of the distribution is equal to `lambda`.

44.8 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

The time elapsed between the arrival of a customer at a shop and the arrival of the next customer has an exponential distribution with expected value equal to 15 minutes. Furthermore, it is independent of previous arrivals. What is the probability that more than 6 customers will arrive at the shop during the next hour?

Solution

If a random variable has an exponential distribution with parameter λ, then its expected value is equal to $1/\lambda$. Here

$$\frac{1}{\lambda} = 0.25 \text{ hours}$$

Therefore, $\lambda = 4$. If inter-arrival times are independent exponential random variables with parameter λ, then the number of arrivals during a unit of time has a Poisson distribution with parameter λ. Thus, the number of customers that will

arrive at the shop during the next hour (denote it by X) is a Poisson random variable with parameter $\lambda = 4$. The probability that more than 6 customers arrive at the shop during the next hour is:

$$\begin{aligned} \mathrm{P}(X > 6) &= 1 - \mathrm{P}(X \leq 6) = 1 - F_X(6) \\ &= 1 - \exp(-4) \sum_{s=0}^{6} \frac{4^s}{s!} \simeq 0.1107 \end{aligned}$$

The value of $F_X(6)$ can be calculated with a computer algorithm, for example with the MATLAB command:

```
poisscdf(6,4)
```

Exercise 2

At a call center, the time elapsed between the arrival of a phone call and the arrival of the next phone call has an exponential distribution with expected value equal to 15 seconds. Furthermore, it is independent of previous arrivals. What is the probability that less than 50 phone calls arrive during the next 15 minutes?

Solution

If a random variable has an exponential distribution with parameter λ, then its expected value is equal to $1/\lambda$. Here

$$\frac{1}{\lambda} = \frac{1}{4} \text{ minutes} = \frac{1}{60} \text{ quarters of hour}$$

where, in the last equality, we have taken 15 minutes as the unit of time. Therefore, $\lambda = 60$. If inter-arrival times are independent exponential random variables with parameter λ, then the number of arrivals during a unit of time has a Poisson distribution with parameter λ. Thus, the number of phone calls that will arrive during the next 15 minutes (denote it by X) is a Poisson random variable with parameter $\lambda = 60$. The probability that less than 50 phone calls arrive during the next 15 minutes is:

$$\begin{aligned} \mathrm{P}(X < 50) &= \mathrm{P}(X \leq 49) = F_X(49) \\ &= \exp(-60) \sum_{s=0}^{49} \frac{60^s}{s!} \simeq 0.0844 \end{aligned}$$

The value of $F_X(49)$ can be calculated with a computer algorithm, for example with the MATLAB command:

```
poisscdf(49,60)
```

Chapter 45

Uniform distribution

A continuous random variable has a uniform distribution if all the values belonging to its support have the same probability density.

45.1 Definition

The uniform distribution is characterized as follows:

Definition 249 *Let X be an absolutely continuous random variable. Let its support be a closed interval of real numbers:*

$$R_X = [l, u]$$

*We say that X has a **uniform distribution** on $[l, u]$ if its probability density function*[1] *is*

$$f_X(x) = \begin{cases} \frac{1}{u-l} & \text{if } x \in R_X \\ 0 & \text{if } x \notin R_X \end{cases}$$

Sometimes, we also say that X has a **rectangular distribution** or that X is a **rectangular random variable**.

45.2 Expected value

The expected value of a uniform random variable X is

$$\mathrm{E}[X] = \frac{u+l}{2}$$

Proof. It can be derived as follows:

$$\begin{aligned} \mathrm{E}[X] &= \int_{-\infty}^{\infty} x f_X(x)\, dx \\ &= \int_{l}^{u} x \frac{1}{u-l} dx \\ &= \frac{1}{u-l} \int_{l}^{u} x dx \end{aligned}$$

[1] See p. 107.

$$\begin{aligned} &= \frac{1}{u-l}\left[\frac{1}{2}x^2\right]_l^u \\ &= \frac{1}{u-l}\frac{1}{2}\left[u^2-l^2\right] \\ &= \frac{(u-l)(u+l)}{2(u-l)} = \frac{u+l}{2} \end{aligned}$$

■

45.3 Variance

The variance of a uniform random variable X is

$$\operatorname{Var}[X] = \frac{(u-l)^2}{12}$$

Proof. It can be derived thanks to the usual formula for computing the variance[2]:

$$\begin{aligned} \mathrm{E}\left[X^2\right] &= \int_{-\infty}^{\infty} x^2 f_X(x)\,dx \\ &= \int_l^u x^2 \frac{1}{u-l}dx \\ &= \frac{1}{u-l}\int_l^u x^2 dx \\ &= \frac{1}{u-l}\left[\frac{1}{3}x^3\right]_l^u \\ &= \frac{1}{u-l}\frac{1}{3}\left[u^3-l^3\right] \\ &= \frac{(u-l)\left(u^2+ul+l^2\right)}{3(u-l)} \\ &= \frac{u^2+ul+l^2}{3} \end{aligned}$$

$$\mathrm{E}[X]^2 = \left(\frac{u+l}{2}\right)^2 = \frac{u^2+2ul+l^2}{4}$$

$$\begin{aligned} \operatorname{Var}[X] &= \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 \\ &= \frac{u^2+ul+l^2}{3} - \frac{u^2+2ul+l^2}{4} \\ &= \frac{4u^2+4ul+4l^2-3u^2-6ul-3l^2}{12} \\ &= \frac{(4-3)u^2+(4-6)ul+(4-3)l^2}{12} \\ &= \frac{u^2-2ul+l^2}{12} = \frac{(u-l)^2}{12} \end{aligned}$$

■

[2] $\operatorname{Var}[X] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2$. See p. 156.

45.4 Moment generating function

The moment generating function of a uniform random variable X is defined for any $t \in \mathbb{R}$:

$$M_X(t) = \begin{cases} \frac{1}{(u-l)t}\left[\exp(tu) - \exp(tl)\right] & \text{if } t \neq 0 \\ 1 & \text{if } t = 0 \end{cases}$$

Proof. Using the definition of moment generating function:

$$\begin{aligned} M_X(t) &= \mathrm{E}\left[\exp(tX)\right] \\ &= \int_{-\infty}^{\infty} \exp(tx) f_X(x)\, dx \\ &= \int_{l}^{u} \exp(tx) \frac{1}{u-l} dx \\ &= \frac{1}{u-l}\left[\frac{1}{t}\exp(tx)\right]_{l}^{u} \\ &= \frac{\exp(tu) - \exp(tl)}{(u-l)\,t} \end{aligned}$$

Note that the above derivation is valid only when $t \neq 0$. However, when $t = 0$:

$$M_X(0) = \mathrm{E}\left[\exp(0 \cdot X)\right] = \mathrm{E}[1] = 1$$

Furthermore, it is easy to verify that

$$\lim_{t \to 0} M_X(t) = M_X(0)$$

When $t \neq 0$, the integral above is well-defined and finite for any $t \in \mathbb{R}$. Thus, the moment generating function of a uniform random variable exists for any $t \in \mathbb{R}$. ■

45.5 Characteristic function

The characteristic function of a uniform random variable X is

$$\varphi_X(t) = \begin{cases} \frac{1}{(u-l)it}\left[\exp(itu) - \exp(itl)\right] & \text{if } t \neq 0 \\ 1 & \text{if } t = 0 \end{cases}$$

Proof. Using the definition of characteristic function:

$$\begin{aligned} \varphi_X(t) &= \mathrm{E}\left[\exp(itX)\right] \\ &= \mathrm{E}\left[\cos(tX)\right] + i\mathrm{E}\left[\sin(tX)\right] \\ &= \int_{-\infty}^{\infty} \cos(tx) f_X(x)\, dx + i\int_{-\infty}^{\infty} \sin(tx) f_X(x)\, dx \\ &= \int_{l}^{u} \cos(tx) \frac{1}{u-l} dx + i\int_{l}^{u} \sin(tx) \frac{1}{u-l} dx \\ &= \frac{1}{u-l}\left\{\int_{l}^{u} \cos(tx)\, dx + i\int_{l}^{u} \sin(tx)\, dx\right\} \\ &= \frac{1}{u-l}\left\{\left[\frac{1}{t}\sin(tx)\right]_{l}^{u} + i\left[-\frac{1}{t}\cos(tx)\right]_{l}^{u}\right\} \end{aligned}$$

$$\begin{aligned} &= \frac{1}{(u-l)\,t}\left\{\sin(tu)-\sin(tl)-i\cos(tu)+i\cos(tl)\right\} \\ &= \frac{1}{(u-l)\,it}\left\{i\sin(tu)-i\sin(tl)+\cos(tu)-\cos(tl)\right\} \\ &= \frac{1}{(u-l)\,it}\left\{[\cos(tu)+i\sin(tu)]-[\cos(tl)+i\sin(tl)]\right\} \\ &= \frac{\exp(itu)-\exp(itl)}{(u-l)\,it} \end{aligned}$$

Note that the above derivation is valid only when $t \neq 0$. However, when $t = 0$:

$$\varphi_X(0) = \mathrm{E}[\exp(i \cdot 0 \cdot X)] = \mathrm{E}[1] = 1$$

Furthermore, it is easy to verify that

$$\lim_{t \to 0} \varphi_X(t) = \varphi_X(0)$$

■

45.6 Distribution function

The distribution function of a uniform random variable X is

$$F_X(x) = \begin{cases} 0 & \text{if } x < l \\ (x-l)/(u-l) & \text{if } l \leq x \leq u \\ 1 & \text{if } x > u \end{cases}$$

Proof. If $x < l$, then:

$$F_X(x) = \mathrm{P}(X \leq x) = 0$$

because X can not take on values smaller than l. If $l \leq x \leq u$, then:

$$\begin{aligned} F_X(x) &= \mathrm{P}(X \leq x) \\ &= \int_{-\infty}^{x} f_X(t)\,dt \\ &= \int_{l}^{x} \frac{1}{u-l}dt \\ &= \frac{1}{u-l}\left[t\right]_l^x \\ &= (x-l)/(u-l) \end{aligned}$$

If $x > u$, then:

$$F_X(x) = \mathrm{P}(X \leq x) = 1$$

because X can not take on values greater than u. ■

45.7 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Let X be a uniform random variable with support

$$R_X = [5, 10]$$

Compute the following probability:

$$\mathrm{P}\left(7 \leq X \leq 9\right)$$

Solution

We can compute this probability either using the probability density function or the distribution function of X. Using the probability density function:

$$\begin{aligned} \mathrm{P}\left(7 \leq X \leq 9\right) &= \int_7^9 f_X\left(x\right) dx = \int_7^9 \frac{1}{10-5} dx \\ &= \frac{1}{5}\left[x\right]_7^9 = \frac{1}{5}\left(9-7\right) = \frac{2}{5} \end{aligned}$$

Using the distribution function:

$$\begin{aligned} \mathrm{P}\left(7 \leq X \leq 9\right) &= F_X\left(9\right) - F_X\left(7\right) = \frac{9-5}{10-5} - \frac{7-5}{10-5} \\ &= \frac{4}{5} - \frac{2}{5} = \frac{2}{5} \end{aligned}$$

Exercise 2

Suppose the random variable X has a uniform distribution on the interval $[-2, 4]$. Compute the following probability:

$$\mathrm{P}\left(X > 2\right)$$

Solution

This probability can be easily computed using the distribution function of X:

$$\begin{aligned} \mathrm{P}\left(X > 2\right) &= 1 - \mathrm{P}\left(X \leq 2\right) = 1 - F_X\left(2\right) \\ &= 1 - \frac{2-(-2)}{4-(-2)} = 1 - \frac{4}{6} = \frac{1}{3} \end{aligned}$$

Exercise 3

Suppose the random variable X has a uniform distribution on the interval $[0, 1]$. Compute the third moment[3] of X, i.e.:

$$\mu_X\left(3\right) = \mathrm{E}\left[X^3\right]$$

[3] See p. 36.

Solution

We can compute the third moment of X using the transformation theorem[4]:

$$\begin{aligned} \mathrm{E}\left[X^3\right] &= \int_{-\infty}^{\infty} x^3 f_X(x)\, dx = \int_0^1 x^3 dx \\ &= \left[\frac{1}{4}x^4\right]_0^1 = \frac{1}{4} \end{aligned}$$

[4] See p. 134.

Chapter 46

Exponential distribution

How much time will elapse before an earthquake occurs in a given region? How long do we need to wait before a customer enters our shop? How long will it take before a call center receives the next phone call? How long will a piece of machinery work without breaking down?

Questions such as these are often answered in probabilistic terms using the exponential distribution.

All these questions concern the time we need to wait before a certain event occurs. If this waiting time is unknown, it is often appropriate to think of it as a random variable having an exponential distribution. Roughly speaking, the time X we need to wait before an event occurs has an exponential distribution if the probability that the event occurs during a certain time interval is proportional to the length of that time interval. More precisely, X has an exponential distribution if the conditional probability

$$\mathrm{P}\left(t < X \leq t + \Delta t \,|\, X > t\right)$$

is approximately proportional to the length Δt of the time interval comprised between the times t and $t + \Delta t$, for any time instant t. In most practical situations this property is very realistic and this is the reason why the exponential distribution is so widely used to model waiting times.

The exponential distribution is also related to the Poisson distribution. When the event can occur more than once and the time elapsed between two successive occurrences is exponentially distributed and independent of previous occurrences, the number of occurrences of the event within a given unit of time has a Poisson distribution. See the lecture entitled *Poisson distribution* (p. 349) for a more detailed explanation and an intuitive graphical representation of this fact.

46.1 Definition

The exponential distribution is characterized as follows.

Definition 250 *Let X be an absolutely continuous random variable. Let its support be the set of positive real numbers:*

$$R_X = [0, \infty)$$

*Let $\lambda \in \mathbb{R}_{++}$. We say that X has an **exponential distribution** with parameter λ if its probability density function*[1] *is*

$$f_X(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{if } x \in R_X \\ 0 & \text{if } x \notin R_X \end{cases}$$

*The parameter λ is called **rate parameter**.*

The following is a proof that $f_X(x)$ is a legitimate probability density function[2].
Proof. Non-negativity is obvious. We need to prove that the integral of $f_X(x)$ over $\mathbb{R}$ equals 1. This is proved as follows:

$$\int_{-\infty}^{\infty} f_X(x)\,dx = \int_0^{\infty} \lambda \exp(-\lambda x)\,dx = \left[-\exp(-\lambda x)\right]_0^{\infty} = 0 - (-1) = 1$$

■

46.2 The rate parameter and its interpretation

We have mentioned that the probability that the event occurs between two dates t and $t + \Delta t$ is proportional to Δt, conditional on the information that it has not occurred before t. The rate parameter λ is the constant of proportionality:

$$\mathrm{P}(t < X \leq t + \Delta t \,|\, X > t) = \lambda \Delta t + o(\Delta t)$$

where $o(\Delta t)$ is an infinitesimal of higher order than Δt, i.e., a function of Δt that goes to zero more quickly than Δt does.

The above proportionality condition is also sufficient to completely characterize the exponential distribution.

Proposition 251 *The proportionality condition*

$$\mathrm{P}(t < X \leq t + \Delta t \,|\, X > t) = \lambda \Delta t + o(\Delta t)$$

is satisfied only if X has an exponential distribution.

Proof. The conditional probability $\mathrm{P}(t < X \leq t + \Delta t \,|\, X > t)$ can be written as

$$\mathrm{P}(t < X \leq t + \Delta t \,|\, X > t) = \frac{\mathrm{P}(t < X \leq t + \Delta t, X > t)}{\mathrm{P}(X > t)} = \frac{\mathrm{P}(t < X \leq t + \Delta t)}{\mathrm{P}(X > t)}$$

Denote by $F_X(x)$ the distribution function[3] of X:

$$F_X(x) = \mathrm{P}(X \leq x)$$

and by $S_X(x)$ its survival function:

$$S_X(x) = 1 - F_X(x) = \mathrm{P}(X > x)$$

Then,

$$\frac{\mathrm{P}(t < X \leq t + \Delta t)}{\mathrm{P}(X > t)} = \frac{F_X(t + \Delta t) - F_X(t)}{1 - F_X(t)} = -\frac{S_X(t + \Delta t) - S_X(t)}{S_X(t)}$$

[1] See p. 107.
[2] See p. 251.
[3] See p. 108.

$$= \quad \lambda\Delta t + o(\Delta t)$$

Dividing both sides by $-\Delta t$, we obtain

$$\frac{S_X(t+\Delta t) - S_X(t)}{\Delta t}\frac{1}{S_X(t)} = -\lambda + o\left(-\frac{\Delta t}{\Delta t}\right) = -\lambda + o(1)$$

where $o(1)$ is a quantity that tends to 0 when Δt tends to 0. Taking limits on both sides, we obtain

$$\lim_{\Delta t \to 0} \frac{S_X(t+\Delta t) - S_X(t)}{\Delta t}\frac{1}{S_X(t)} = -\lambda$$

or, by the definition of derivative:

$$\frac{dS_X(t)}{dt}\frac{1}{S_X(t)} = -\lambda$$

This differential equation is easily solved using the chain rule:

$$\frac{dS_X(t)}{dt}\frac{1}{S_X(t)} = \frac{d\ln(S_X(t))}{dt} = -\lambda$$

Taking the integral from 0 to x of both sides

$$\int_0^x \frac{d\ln(S_X(t))}{dt}dt = -\int_0^x \lambda dt$$

we obtain

$$[\ln(S_X(t))]_0^x = -[\lambda t]_0^x$$

or

$$\ln(S_X(x)) = \ln(S_X(0)) - \lambda x$$

But X cannot take negative values. So

$$S_X(0) = 1 - F_X(0) = 1$$

which implies

$$\ln(S_X(x)) = -\lambda x$$

Exponentiating both sides, we get

$$S_X(x) = \exp(-\lambda x)$$

Therefore,

$$1 - F_X(x) = \exp(-\lambda x)$$

or

$$F_X(x) = 1 - \exp(-\lambda x)$$

Since the density function is the first derivative of the distribution function[4], we obtain

$$f_X(x) = \frac{dF_X(x)}{dx} = \lambda\exp(-\lambda x)$$

which is the density of an exponential random variable. Therefore, the proportionality condition is satisfied only if X is an exponential random variable. ■

[4] See p. 109.

46.3 Expected value

The expected value of an exponential random variable X is

$$\mathrm{E}[X] = \frac{1}{\lambda}$$

Proof. It can be derived as follows:

$$\begin{aligned}
\mathrm{E}[X] &= \int_0^\infty x f_X(x)\,dx \\
&= \int_0^\infty x\lambda \exp(-\lambda x)\,dx \\
\boxed{\text{A}} &= [-x\exp(-\lambda x)]_0^\infty + \int_0^\infty \exp(-\lambda x)\,dx \\
&= (0-0) + \left[-\frac{1}{\lambda}\exp(-\lambda x)\right]_0^\infty \\
&= 0 + \left(0 + \frac{1}{\lambda}\right) = \frac{1}{\lambda}
\end{aligned}$$

where: in step A we have performed an integration by parts[5]. ■

46.4 Variance

The variance of an exponential random variable X is

$$\mathrm{Var}[X] = \frac{1}{\lambda^2}$$

Proof. The second moment of X is

$$\begin{aligned}
\mathrm{E}\left[X^2\right] &= \int_0^\infty x^2\lambda \exp(-\lambda x)\,dx \\
\boxed{\text{A}} &= \left[-x^2\exp(-\lambda x)\right]_0^\infty + \int_0^\infty 2x\exp(-\lambda x)\,dx \\
\boxed{\text{B}} &= (0-0) + \left[-\frac{2}{\lambda}x\exp(-\lambda x)\right]_0^\infty + \frac{2}{\lambda}\int_0^\infty \exp(-\lambda x)\,dx \\
&= (0-0) + \frac{2}{\lambda}\left[-\frac{1}{\lambda}\exp(-\lambda x)\right]_0^\infty = \frac{2}{\lambda^2}
\end{aligned}$$

where: in step A and B we have performed two integrations by parts. Furthermore,

$$\mathrm{E}[X]^2 = \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$$

The usual formula for computing the variance[6] gives

$$\mathrm{Var}[X] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

■

[5] See p. 51.
[6] See p. 156.

46.5 Moment generating function

The moment generating function of an exponential random variable X is defined for any $t < \lambda$ and it is equal to

$$M_X(t) = \frac{\lambda}{\lambda - t}$$

Proof. Using the definition of moment generating function, we obtain

$$\begin{aligned} M_X(t) &= \mathrm{E}[\exp(tX)] = \int_{-\infty}^{\infty} \exp(tx) f_X(x)\, dx \\ &= \int_0^{\infty} \exp(tx)\, \lambda \exp(-\lambda x)\, dx = \lambda \int_0^{\infty} \exp((t-\lambda)x)\, dx \\ &= \lambda \left[\frac{1}{t-\lambda} \exp((t-\lambda)x) \right]_0^{\infty} = \frac{\lambda}{\lambda - t} \end{aligned}$$

Of course, the above integrals converge only if $(t - \lambda) < 0$, i.e., only if $t < \lambda$. Therefore, the moment generating function of an exponential random variable exists for all $t < \lambda$. ■

46.6 Characteristic function

The characteristic function of an exponential random variable X is

$$\varphi_X(t) = \frac{\lambda}{\lambda - it}$$

Proof. Using the definition of characteristic function and the fact that

$$\exp(itx) = \cos(tx) + i \sin(tx)$$

we can write

$$\begin{aligned} \varphi_X(t) &= \mathrm{E}[\exp(itX)] = \int_{-\infty}^{\infty} \exp(itx) f_X(x)\, dx \\ &= \int_0^{\infty} \exp(itx)\, \lambda \exp(-\lambda x)\, dx \\ &= \lambda \int_0^{\infty} \cos(tx) \exp(-\lambda x)\, dx + i\lambda \int_0^{\infty} \sin(tx) \exp(-\lambda x)\, dx \end{aligned}$$

We now compute separately the two integrals. The first integral is

$$\begin{aligned} & \int_0^{\infty} \cos(tx) \exp(-\lambda x)\, dx \\ \boxed{\text{A}} \quad &= \left[\frac{1}{t} \sin(tx) \exp(-\lambda x) \right]_0^{\infty} \\ & \quad - \int_0^{\infty} \frac{1}{t} \sin(tx) (-\lambda \exp(-\lambda x))\, dx \\ &= \frac{\lambda}{t} \int_0^{\infty} \sin(tx) \exp(-\lambda x)\, dx \end{aligned}$$

$$\begin{aligned}
\boxed{\text{B}} \quad &= \frac{\lambda}{t}\left\{\left[-\frac{1}{t}\cos(tx)\exp(-\lambda x)\right]_0^\infty \right. \\
&\quad \left. -\int_0^\infty \left(-\frac{1}{t}\cos(tx)\right)(-\lambda\exp(-\lambda x))\,dx\right\} \\
&= \frac{\lambda}{t}\left\{\frac{1}{t}-\frac{\lambda}{t}\int_0^\infty \cos(tx)\exp(-\lambda x)\,dx\right\} \\
&= \frac{\lambda}{t^2}-\frac{\lambda^2}{t^2}\int_0^\infty \cos(tx)\exp(-\lambda x)\,dx
\end{aligned}$$

where in step $\boxed{\text{A}}$ and $\boxed{\text{B}}$ we have performed two integrations by parts. Therefore,

$$\int_0^\infty \cos(tx)\exp(-\lambda x)\,dx = \frac{\lambda}{t^2}-\frac{\lambda^2}{t^2}\int_0^\infty \cos(tx)\exp(-\lambda x)\,dx$$

which can be rearranged to yield

$$\left(1+\frac{\lambda^2}{t^2}\right)\int_0^\infty \cos(tx)\exp(-\lambda x)\,dx = \frac{\lambda}{t^2}$$

or

$$\int_0^\infty \cos(tx)\exp(-\lambda x)\,dx = \frac{\lambda}{t^2}\left(1+\frac{\lambda^2}{t^2}\right)^{-1} = \frac{\lambda}{t^2}\frac{t^2}{t^2+\lambda^2} = \frac{\lambda}{t^2+\lambda^2}$$

The second integral is

$$\begin{aligned}
&\int_0^\infty \sin(tx)\exp(-\lambda x)\,dx \\
\boxed{\text{A}} \quad &= \left[-\frac{1}{t}\cos(tx)\exp(-\lambda x)\right]_0^\infty \\
&\quad -\int_0^\infty \left(-\frac{1}{t}\cos(tx)\right)(-\lambda\exp(-\lambda x))\,dx \\
&= \frac{1}{t}-\frac{\lambda}{t}\int_0^\infty \cos(tx)\exp(-\lambda x)\,dx \\
\boxed{\text{B}} \quad &= \frac{1}{t}-\frac{\lambda}{t}\left\{\left[\frac{1}{t}\sin(tx)\exp(-\lambda x)\right]_0^\infty \right. \\
&\quad \left. -\int_0^\infty \frac{1}{t}\sin(tx)(-\lambda\exp(-\lambda x))\,dx\right\} \\
&= \frac{1}{t}-\frac{\lambda}{t}\left\{\frac{\lambda}{t}\int_0^\infty \sin(tx)\exp(-\lambda x)\,dx\right\} \\
&= \frac{1}{t}-\frac{\lambda^2}{t^2}\int_0^\infty \sin(tx)\exp(-\lambda x)\,dx
\end{aligned}$$

where in step $\boxed{\text{A}}$ and $\boxed{\text{B}}$ we have performed two integrations by parts. Therefore,

$$\int_0^\infty \sin(tx)\exp(-\lambda x)\,dx = \frac{1}{t}-\frac{\lambda^2}{t^2}\int_0^\infty \sin(tx)\exp(-\lambda x)\,dx$$

which can be rearranged to yield

$$\left(1+\frac{\lambda^2}{t^2}\right)\int_0^\infty \sin(tx)\exp(-\lambda x)\,dx = \frac{1}{t}$$

or

$$\int_0^\infty \sin(tx)\exp(-\lambda x)\,dx = \frac{1}{t}\left(1+\frac{\lambda^2}{t^2}\right)^{-1} = \frac{t}{t^2+\lambda^2}$$

Putting pieces together, we get

$$\begin{aligned}\varphi_X(t) &= \lambda\int_0^\infty \cos(tx)\exp(-\lambda x)\,dx + i\lambda\int_0^\infty \sin(tx)\exp(-\lambda x)\,dx \\ &= \lambda\frac{\lambda}{t^2+\lambda^2} + i\lambda\frac{t}{t^2+\lambda^2} = \lambda\frac{\lambda+it}{t^2+\lambda^2} \\ &= \lambda\frac{\lambda+it}{t^2+\lambda^2}\frac{\lambda-it}{\lambda-it} = \frac{\lambda}{\lambda-it}\end{aligned}$$

■

46.7 Distribution function

The distribution function of an exponential random variable X is

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

Proof. If $x<0$, then

$$F_X(x) = \mathrm{P}(X\leq x) = 0$$

because X can not take on negative values. If $x\geq 0$, then

$$\begin{aligned}F_X(x) &= \mathrm{P}(X\leq x) = \int_{-\infty}^x f_X(t)\,dt = \int_0^x \lambda\exp(-\lambda t)\,dt \\ &= [-\exp(-\lambda t)]_0^x = -\exp(-\lambda x)+1\end{aligned}$$

■

46.8 More details

In the following subsections you can find more details about the exponential distribution.

46.8.1 Memoryless property

One of the most important properties of the exponential distribution is the **memoryless property**:

$$\mathrm{P}(X\leq x+y\,|X>x) = \mathrm{P}(X\leq y)$$

for any $x\geq 0$.
Proof. This is proved as follows:

$$\begin{aligned}\mathrm{P}(X\leq x+y\,|X>x) &= \frac{\mathrm{P}(X\leq x+y \text{ and } X>x)}{\mathrm{P}(X>x)} \\ &= \frac{\mathrm{P}(x<X\leq x+y)}{\mathrm{P}(X>x)}\end{aligned}$$

$$\begin{aligned} & = \frac{F_X(x+y) - F_X(x)}{1 - F_X(x)} \\ & = \frac{1 - \exp(-\lambda(x+y)) - (1 - \exp(-\lambda x))}{\exp(-\lambda x)} \\ & = \frac{\exp(-\lambda x) - \exp(-\lambda(x+y))}{\exp(-\lambda x)} \\ & = 1 - \exp(-\lambda y) = F_X(y) = \mathrm{P}(X \leq y) \end{aligned}$$

■

Remember that X is the time we need to wait before a certain event occurs. The memoryless property states that the probability that the event happens during a time interval of length y is independent of how much time x has already elapsed without the event happening.

46.8.2 Sums of exponential random variables

If $X_1, X_2, \ldots, X_n$ are n mutually independent[7] random variables having an exponential distribution with parameter λ, then the sum

$$Z = \sum_{i=1}^{n} X_i$$

has a Gamma distribution[8] with parameters $2n$ and n/λ.

Proof. This is proved using moment generating functions[9]:

$$\begin{aligned} M_Z(t) & = \prod_{i=1}^{n} M_{X_i}(t) = \prod_{i=1}^{n} \frac{\lambda}{\lambda - t} = \left(\frac{\lambda}{\lambda - t}\right)^n \\ & = \left(1 - \frac{1}{\lambda}t\right)^{-n} = \left(1 - \frac{2(n/\lambda)}{(2n)}t\right)^{-(2n)/2} \end{aligned}$$

The latter is the moment generating function of a Gamma distribution[10] with parameters $2n$ and n/λ. So Z has a Gamma distribution, because two random variables have the same distribution when they have the same moment generating function. ■

The random variable Z is also sometimes said to have an **Erlang distribution**. The Erlang distribution is just a special case of the Gamma distribution: a Gamma random variable is also an Erlang random variable when it can be written as a sum of exponential random variables.

46.9 Solved exercises

Below you can find some exercises with explained solutions.

[7] See p. 233.

[8] See p. 397.

[9] Remember that the moment generating function of a sum of mutually independent random variables is just the product of their moment generating functions (see p. 293).

[10] See p. 399.

Exercise 1

The probability that a new customer enters a shop during a given minute is approximately 1%, irrespective of how many customers have entered the shop during the previous minutes. Assume that the total time we need to wait before a new customer enters the shop (denote it by X) has an exponential distribution. What is the probability that no customer enters the shop during the next hour?

Solution

Time is measured in minutes. Therefore, the probability that no customer enters the shop during the next hour is

$$\mathrm{P}(X > 60) = 1 - \mathrm{P}(X \leq 60) = 1 - F_X(60)$$

where $F_X(x)$ is the distribution function of X. Since X is an exponential random variable with rate parameter 1%, its distribution function is

$$F_X(x) = 1 - \exp(-0.01 \cdot x)$$

Therefore,

$$\begin{aligned} \mathrm{P}(X > 60) &= 1 - F_X(60) = 1 - (1 - \exp(-0.01 \cdot 60)) \\ &= \exp(-0.01 \cdot 60) = \exp(-0.6) \end{aligned}$$

Exercise 2

Let X be an exponential random variable with parameter $\lambda = \ln(3)$. Compute the probability

$$\mathrm{P}(2 \leq X \leq 4)$$

Solution

First of all we can write the probability as

$$\begin{aligned} \mathrm{P}(2 \leq X \leq 4) &= \mathrm{P}(\{X = 2\} \cup \{2 < X \leq 4\}) \\ &= \mathrm{P}(X = 2) + \mathrm{P}(2 < X \leq 4) = \mathrm{P}(2 < X \leq 4) \end{aligned}$$

where we have used the fact that the probability that an absolutely continuous random variable takes on any specific value is equal to zero[11]. Now, the probability can be written in terms of the distribution function of X as

$$\begin{aligned} \mathrm{P}(2 \leq X \leq 4) &= \mathrm{P}(2 < X \leq 4) = F_X(4) - F_X(2) \\ &= [1 - \exp(-\ln(3) \cdot 4)] - [1 - \exp(-\ln(3) \cdot 2)] \\ &= \exp(-\ln(3) \cdot 2) - \exp(-\ln(3) \cdot 4) = 3^{-2} - 3^{-4} \end{aligned}$$

Exercise 3

Suppose the random variable X has an exponential distribution with parameter $\lambda = 1$. Compute the probability

$$\mathrm{P}(X > 2)$$

[11] See p. 109.

Solution

The above probability can be easily computed using the distribution function of X:

$$\mathrm{P}\left(X > 2\right) = 1 - \mathrm{P}\left(X \leq 2\right) = 1 - F_X\left(2\right) = 1 - \left[1 - \exp\left(-2\right)\right] = \exp\left(-2\right)$$

Exercise 4

What is the probability that a random variable X is less than its expected value, if X has an exponential distribution with parameter λ?

Solution

The expected value of an exponential random variable with parameter λ is

$$\mathrm{E}\left[X\right] = \frac{1}{\lambda}$$

The probability above can be computed using the distribution function of X:

$$\begin{aligned} \mathrm{P}\left(X \leq \mathrm{E}\left[X\right]\right) &= \mathrm{P}\left(X \leq \frac{1}{\lambda}\right) = F_X\left(\frac{1}{\lambda}\right) \\ &= 1 - \exp\left(-\lambda\frac{1}{\lambda}\right) = 1 - \exp\left(-1\right) \end{aligned}$$

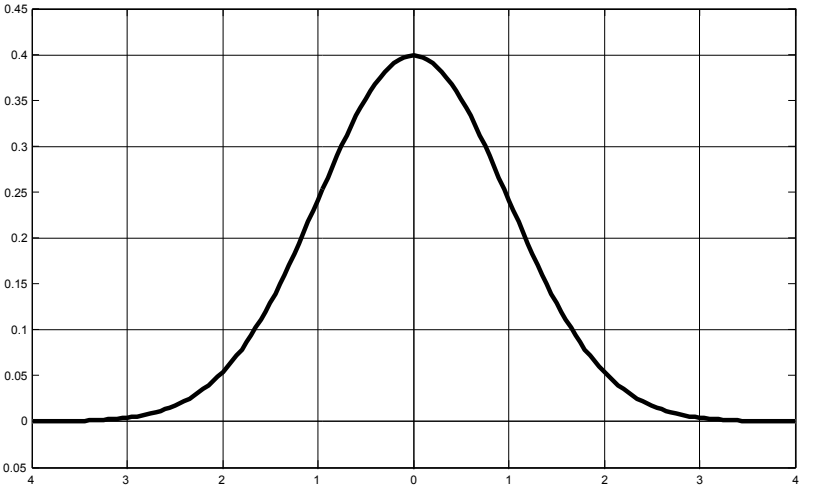
Chapter 47

Normal distribution

The normal distribution is one of the cornerstones of probability theory and statistics, because of the role it plays in the Central Limit Theorem[1], because of its analytical tractability and because many real-world phenomena involve random quantities that are approximately normal (e.g. errors in scientific measurement).

It is often called Gaussian distribution, in honor of Carl Friedrich Gauss (1777-1855), an eminent German mathematician who gave important contributions towards a better understanding of the normal distribution.

Sometimes it is also referred to as "bell-shaped distribution", because the graph of its probability density function resembles the shape of a bell.



As you can see from the above plot of the density of a normal distribution, the density is symmetric around the mean (indicated by the vertical line at zero). As a consequence, deviations from the mean having the same magnitude, but different signs, have the same probability. The density is also very concentrated around the mean and becomes very small by moving from the center to the left or to the right of the distribution (the so called "tails" of the distribution). This means that

[1] See p. 545.

the further a value is from the center of the distribution, the less probable it is to observe that value.

The remainder of this lecture gives a formal presentation of the main characteristics of the normal distribution, dealing first with the special case in which the distribution has zero mean and unit variance, then with the general case, in which mean and variance can take any value.

47.1 The standard normal distribution

The adjective "standard" indicates the special case in which the mean is equal to zero and the variance is equal to one.

47.1.1 Definition

The standard normal distribution is characterized as follows:

Definition 252 *Let X be an absolutely continuous random variable. Let its support be the whole set of real numbers:*

$$R_X = \mathbb{R}$$

*We say that X has a **standard normal distribution** if its probability density function[2] is:*

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$$

The following is a proof that $f_X(x)$ is indeed a legitimate probability density function[3].

Proof. $f_X(x)$ is a legitimate probability density function if it is non-negative and if its integral over the support equals 1. The former property is obvious, while the latter can be proved as follows:

$$\begin{aligned}
& \int_{-\infty}^{\infty} f_X(x)\,dx \\
= \; & (2\pi)^{-1/2} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}x^2\right) dx \\
\boxed{\text{A}} \; = \; & (2\pi)^{-1/2}\, 2 \int_{0}^{\infty} \exp\left(-\frac{1}{2}x^2\right) dx \\
= \; & (2\pi)^{-1/2}\, 2 \left(\int_{0}^{\infty} \exp\left(-\frac{1}{2}x^2\right) dx \int_{0}^{\infty} \exp\left(-\frac{1}{2}x^2\right) dx\right)^{1/2} \\
= \; & (2\pi)^{-1/2}\, 2 \left(\int_{0}^{\infty} \exp\left(-\frac{1}{2}x^2\right) dx \int_{0}^{\infty} \exp\left(-\frac{1}{2}y^2\right) dy\right)^{1/2} \\
= \; & (2\pi)^{-1/2}\, 2 \left(\int_{0}^{\infty} \int_{0}^{\infty} \exp\left(-\frac{1}{2}\left(x^2+y^2\right)\right) dydx\right)^{1/2} \\
\boxed{\text{B}} \; = \; & (2\pi)^{-1/2}\, 2 \left(\int_{0}^{\infty} \int_{0}^{\infty} \exp\left(-\frac{1}{2}\left(x^2+s^2x^2\right)\right) xdsdx\right)^{1/2}
\end{aligned}$$

[2] See p. 107.
[3] See p. 251.

$$
\begin{aligned}
&= (2\pi)^{-1/2} 2 \left(\int_0^{\infty} \int_0^{\infty} \exp\left(-\frac{1}{2} x^2 \left(1 + s^2 \right) \right) x dx ds \right)^{1/2} \\
&= (2\pi)^{-1/2} 2 \left(\int_0^{\infty} \left[-\frac{1}{1+s^2} \exp\left(-\frac{1}{2} x^2 \left(1 + s^2 \right) \right) \right]_0^{\infty} ds \right)^{1/2} \\
&= (2\pi)^{-1/2} 2 \left(\int_0^{\infty} \left(0 + \frac{1}{1+s^2} \right) ds \right)^{1/2} \\
&= (2\pi)^{-1/2} 2 \left(\int_0^{\infty} \frac{1}{1+s^2} ds \right)^{1/2} \\
&= (2\pi)^{-1/2} 2 \left(\left[\arctan(s) \right]_0^{\infty} \right)^{1/2} \\
&= (2\pi)^{-1/2} 2 \left(\arctan(\infty) - \arctan(0) \right)^{1/2} \\
&= (2\pi)^{-1/2} 2 \left(\frac{\pi}{2} - 0 \right)^{1/2} = 2^{-1/2} \pi^{-1/2} 2 \pi^{1/2} 2^{-1/2} = 1
\end{aligned}
$$

where: in step A we have used the fact that the integrand is even; in step B we have made a change of variable ($y = xs$). ■

47.1.2 Expected value

The expected value of a standard normal random variable X is:

$$\mathrm{E}[X] = 0$$

Proof. It can be derived as follows:

$$
\begin{aligned}
\mathrm{E}[X] &= \int_{-\infty}^{\infty} x f_X(x) dx \\
&= (2\pi)^{-1/2} \int_{-\infty}^{\infty} x \exp\left(-\frac{1}{2} x^2 \right) dx \\
&= (2\pi)^{-1/2} \int_{-\infty}^{0} x \exp\left(-\frac{1}{2} x^2 \right) dx \\
&\quad + (2\pi)^{-1/2} \int_{0}^{\infty} x \exp\left(-\frac{1}{2} x^2 \right) dx \\
&= (2\pi)^{-1/2} \left[-\exp\left(-\frac{1}{2} x^2 \right) \right]_{-\infty}^{0} \\
&\quad + (2\pi)^{-1/2} \left[-\exp\left(-\frac{1}{2} x^2 \right) \right]_{0}^{\infty} \\
&= (2\pi)^{-1/2} \left[-1 + 0 \right] + (2\pi)^{-1/2} \left[0 + 1 \right] \\
&= (2\pi)^{-1/2} - (2\pi)^{-1/2} = 0
\end{aligned}
$$

■

47.1.3 Variance

The variance of a standard normal random variable X is:

$$\mathrm{Var}[X] = 1$$

Proof. It can be proved with the usual formula for computing the variance[4]:

$$\begin{aligned}
& \mathrm{E}\left[X^2\right] \\
= & \int_{-\infty}^{\infty} x^2 f_X(x)\,dx \\
= & (2\pi)^{-1/2} \int_{-\infty}^{\infty} x^2 \exp\left(-\frac{1}{2}x^2\right) dx \\
= & (2\pi)^{-1/2} \left\{ \int_{-\infty}^{0} x \left(x \exp\left(-\frac{1}{2}x^2\right)\right) dx \right. \\
& \left. + \int_{0}^{\infty} x \left(x \exp\left(-\frac{1}{2}x^2\right)\right) dx \right\} \\
\boxed{\text{A}} = & (2\pi)^{-1/2} \left\{ \left[-x \exp\left(-\frac{1}{2}x^2\right)\right]_{-\infty}^{0} + \int_{-\infty}^{0} \exp\left(-\frac{1}{2}x^2\right) dx \right. \\
& \left. + \left[-x \exp\left(-\frac{1}{2}x^2\right)\right]_{0}^{\infty} + \int_{0}^{\infty} \exp\left(-\frac{1}{2}x^2\right) dx \right\} \\
= & (2\pi)^{-1/2} \left\{ (0-0) + (0-0) + \int_{-\infty}^{0} \exp\left(-\frac{1}{2}x^2\right) dx \right. \\
& \left. + \int_{0}^{\infty} \exp\left(-\frac{1}{2}x^2\right) dx \right\} \\
= & (2\pi)^{-1/2} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}x^2\right) dx \\
\boxed{\text{B}} = & \int_{-\infty}^{\infty} f_X(x)\,dx = 1
\end{aligned}$$

where: in step A we have performed an integration by parts[5]; in step B we have used the fact that the integral of a probability density function over its support is equal to 1. Finally:

$$\mathrm{E}[X]^2 = 0^2 = 0$$

and

$$\mathrm{Var}[X] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 = 1 - 0 = 1$$

■

47.1.4 Moment generating function

The moment generating function of a standard normal random variable X is defined for any $t \in \mathbb{R}$:

$$M_X(t) = \exp\left(\frac{1}{2}t^2\right)$$

Proof. Using the definition of moment generating function:

$$\begin{aligned}
M_X(t) &= \mathrm{E}[\exp(tX)] \\
&= \int_{-\infty}^{\infty} \exp(tx) f_X(x)\,dx
\end{aligned}$$

[4] $\mathrm{Var}[X] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2$. See p. 156.
[5] See p. 51.

$$
\begin{aligned}
&= (2\pi)^{-1/2} \int_{-\infty}^{\infty} \exp(tx) \exp\left(-\frac{1}{2}x^2\right) dx \\
&= (2\pi)^{-1/2} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\left(x^2 - 2tx\right)\right) dx \\
&= (2\pi)^{-1/2} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\left(x^2 - 2tx + t^2 - t^2\right)\right) dx \\
&= (2\pi)^{-1/2} \int_{-\infty}^{\infty} \exp\left(\frac{1}{2}t^2\right) \exp\left(-\frac{1}{2}\left(x^2 - 2tx + t^2\right)\right) dx \\
&= \exp\left(\frac{1}{2}t^2\right) (2\pi)^{-1/2} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(x-t)^2\right) dx \\
\boxed{\text{A}} \quad &= \exp\left(\frac{1}{2}t^2\right) (2\pi)^{-1/2} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}z^2\right) dz \\
\boxed{\text{B}} \quad &= \exp\left(\frac{1}{2}t^2\right) \int_{-\infty}^{\infty} f_Z(z)\, dz \\
\boxed{\text{C}} \quad &= \exp\left(\frac{1}{2}t^2\right)
\end{aligned}
$$

where: in step A we have performed a change of variable ($z = x - t$); in step B we have defined

$$f_Z(z) = \exp\left(-\frac{1}{2}z^2\right)$$

where $f_Z(z)$ is the probability density function of a standard normal random variable; in step C we have used the fact that the integral of a probability density function over its support is equal to 1. The integral above is well-defined and finite for any $t \in \mathbb{R}$. Thus, the moment generating function of a standard normal random variable exists for any $t \in \mathbb{R}$. ■

47.1.5 Characteristic function

The characteristic function of a standard normal random variable X is:

$$\varphi_X(t) = \exp\left(-\frac{1}{2}t^2\right)$$

Proof. Using the definition of characteristic function:

$$
\begin{aligned}
\varphi_X(t) &= \mathrm{E}\left[\exp(itX)\right] \\
&= \mathrm{E}\left[\cos(tX)\right] + i\mathrm{E}\left[\sin(tX)\right] \\
&= \int_{-\infty}^{\infty} \cos(tx) f_X(x)\, dx + i \int_{-\infty}^{\infty} \sin(tx) f_X(x)\, dx \\
\boxed{\text{A}} \quad &= \int_{-\infty}^{\infty} \cos(tx) f_X(x)\, dx
\end{aligned}
$$

where: in step A we have used the fact that $\sin(tx) f_X(x)$ is an odd function of x. Now, take the derivative with respect to t of the characteristic function:

$$\frac{d}{dt}\varphi_X(t) = \frac{d}{dt}\mathrm{E}\left[\exp(itX)\right]$$

$$\begin{aligned}
&= \mathrm{E}\left[\frac{d}{dt}\exp(itX)\right] \\
&= \mathrm{E}\left[iX\exp(itX)\right] \\
&= i\mathrm{E}\left[X\cos(tX)\right] - \mathrm{E}\left[X\sin(tX)\right] \\
&= i\int_{-\infty}^{\infty} x\cos(tx)\, f_X(x)\, dx - \int_{-\infty}^{\infty} x\sin(tx)\, f_X(x)\, dx \\
\boxed{\text{A}} &= -\int_{-\infty}^{\infty} x\sin(tx)\, f_X(x)\, dx \\
&= -\int_{-\infty}^{\infty} x\sin(tx)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right) dx \\
&= \int_{-\infty}^{\infty} \sin(tx)\frac{d}{dx}\left[\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right)\right] dx \\
&= \int_{-\infty}^{\infty} \sin(tx)\frac{d}{dx} f_X(x)\, dx \\
\boxed{\text{B}} &= \left[\sin(tx)\, f_X(x)\right]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} t\cos(tx)\, f_X(x)\, dx \\
&= -t\int_{-\infty}^{\infty} \cos(tx)\, f_X(x)\, dx
\end{aligned}$$

where: in step A we have used the fact that $x\cos(tx)\, f_X(x)$ is an odd function of x; in step B we have performed an integration by parts. Putting together the previous two results, we obtain:

$$\frac{d}{dt}\varphi_X(t) = -t\varphi_X(t)$$

The only function that satisfies this ordinary differential equation (subject to the condition $\varphi_X(0) = \mathrm{E}\left[\exp(i\cdot 0\cdot X)\right] = 1$) is:

$$\varphi_X(t) = \exp\left(-\frac{1}{2}t^2\right)$$

■

47.1.6 Distribution function

There is no simple formula for the distribution function $F_X(x)$ of a standard normal random variable X, because the integral

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt$$

cannot be expressed in terms of elementary functions. Therefore, it is necessary to resort to computer algorithms to compute the values of the distribution function of a standard normal random variable. For example, the MATLAB command:

```
normcdf(x)
```

returns the value of the distribution function at the point `x`.

Some values of the distribution function of X are used very frequently and people usually learn them by heart:

$$\begin{array}{ll} F_X(-2.576) = 0.005 & F_X(2.576) = 0.995 \\ F_X(-2.326) = 0.01 & F_X(2.326) = 0.99 \\ F_X(-1.96) = 0.025 & F_X(1.96) = 0.975 \\ F_X(-1.645) = 0.05 & F_X(1.645) = 0.95 \end{array}$$

Note also that:

$$F_X(-x) = 1 - F_X(x)$$

which is due to the symmetry around 0 of the standard normal density and is often used in calculations.

In the past, when computers were not widely available, people used to look up the values of $F_X(x)$ in normal distribution tables. A **normal distribution table** is a table where $F_X(x)$ is tabulated for several values of x. For values of x that are not tabulated, approximations of $F_X(x)$ can be computed by interpolating the two tabulated values that are closest to x. For example, if x is not tabulated, x_1 is the greatest tabulated number smaller than x and x_2 is the smallest tabulated number greater than x, the approximation is as follows:

$$F_X(x) = F_X(x_1) + \frac{F_X(x_2) - F_X(x_1)}{x_2 - x_1}(x - x_1)$$

47.2 The normal distribution in general

While in the previous section we restricted our attention to the normal distribution with zero mean and unit variance, we now deal with the general case.

47.2.1 Definition

The normal distribution with mean μ and variance σ^2 is characterized as follows:

Definition 253 *Let X be an absolutely continuous random variable. Let its support be the whole set of real numbers:*

$$R_X = \mathbb{R}$$

*Let $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}_{++}$. We say that X has a **normal distribution** with mean μ and variance σ^2, if its probability density function is*

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

We often indicate that X has a normal distribution with mean μ and variance σ^2 by:

$$X \sim N\left(\mu, \sigma^2\right)$$

47.2.2 Relation to the standard normal distribution

A random variable X having a normal distribution with mean μ and variance σ^2 is just a linear function of a standard normal random variable:

Proposition 254 *If X has a normal distribution with mean μ and variance σ^2, then:*

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

where Z is a random variable having a standard normal distribution.

Proof. This can be easily proved using the formula for the density of a function[6] of an absolutely continuous variable:

$$\begin{aligned} f_X(x) &= f_Z\left(g^{-1}(x)\right)\frac{dg^{-1}(x)}{dx} \\ &= f_Z\left(\frac{x-\mu}{\sigma}\right)\frac{1}{\sigma} \\ &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)\frac{1}{\sigma} \end{aligned}$$

■

Obviously, then, a standard normal distribution is just a normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$.

47.2.3 Expected value

The expected value of a normal random variable X is:

$$\mathrm{E}[X] = \mu$$

Proof. It is an immediate consequence of the fact that $X = \mu + \sigma Z$ (where Z has a standard normal distribution) and the linearity of the expected value[7]:

$$\mathrm{E}[X] = \mathrm{E}[\mu + \sigma Z] = \mu + \sigma \mathrm{E}[Z] = \mu + \sigma \cdot 0 = \mu$$

■

47.2.4 Variance

The variance of a normal random variable X is:

$$\mathrm{Var}[X] = \sigma^2$$

Proof. It can be derived using the formula for the variance of linear transformations[8] on $X = \mu + \sigma Z$ (where Z has a standard normal distribution):

$$\mathrm{Var}[X] = \mathrm{Var}[\mu + \sigma Z] = \sigma^2 \mathrm{Var}[Z] = \sigma^2$$

■

[6] See p. 265. Note that $X = g(Z) = \mu + \sigma Z$ is a strictly increasing function of Z, since σ is strictly positive.

[7] See p. 134.

[8] See p. 158.

47.2.5 Moment generating function

The moment generating function of a normal random variable X is defined for any $t \in \mathbb{R}$:

$$M_X(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$$

Proof. Recall that $X = \mu + \sigma Z$ (where Z has a standard normal distribution) and that the moment generating function of a standard normal random variable is:

$$M_Z(t) = \exp\left(\frac{1}{2}t^2\right)$$

We can use the formula for the moment generating function of a linear transformation[9]:

$$\begin{aligned} M_X(t) &= \exp(\mu t) M_Z(\sigma t) \\ &= \exp(\mu t) \exp\left(\frac{1}{2}(\sigma t)^2\right) \\ &= \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right) \end{aligned}$$

It is defined for any $t \in \mathbb{R}$ because the moment generating function of Z is defined for any $t \in \mathbb{R}$. ■

47.2.6 Characteristic function

The characteristic function of a normal random variable X is:

$$\varphi_X(t) = \exp\left(i\mu t - \frac{1}{2}\sigma^2 t^2\right)$$

Proof. Recall that $X = \mu + \sigma Z$ (where Z has a standard normal distribution) and that the characteristic function of a standard normal random variable is:

$$\varphi_Z(t) = \exp\left(-\frac{1}{2}t^2\right)$$

We can use the formula for the characteristic function of a linear transformation[10]:

$$\begin{aligned} \varphi_X(t) &= \exp(i\mu t)\varphi_Z(\sigma t) \\ &= \exp(i\mu t)\exp\left(-\frac{1}{2}(\sigma t)^2\right) \\ &= \exp\left(i\mu t - \frac{1}{2}\sigma^2 t^2\right) \end{aligned}$$

■

[9] See p. 293.
[10] See p. 310.

47.2.7 Distribution function

The distribution function $F_X(x)$ of a normal random variable X can be written as:

$$F_X(x) = F_Z\left(\frac{x-\mu}{\sigma}\right)$$

where $F_Z(z)$ is the distribution function of a standard normal random variable Z.

Proof. Remember that any normal random variable X with mean μ and variance σ^2 can be written as:

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

where Z is a standard normal random variable. Using this fact, we obtain the following relation between the distribution function of Z and the distribution function of X:

$$\begin{aligned} F_X(x) &= \mathrm{P}(X \leq x) \\ &= \mathrm{P}(\mu + \sigma Z \leq x) \\ &= \mathrm{P}\left(Z \leq \frac{x-\mu}{\sigma}\right) \\ &= F_Z\left(\frac{x-\mu}{\sigma}\right) \end{aligned}$$

■

Therefore, if we know how to compute the values of the distribution function of a standard normal distribution (see above), we also know how to compute the values of the distribution function of a normal distribution with mean μ and variance σ^2.

Example 255 *If we need to compute the value $F_X\left(\frac{1}{2}\right)$ of a normal random variable X with mean $\mu = 1$ and variance $\sigma^2 = 1$, we can compute it using the distribution function of a standard normal random variable Z:*

$$F_X\left(\frac{1}{2}\right) = F_Z\left(\frac{1/2-\mu}{\sigma}\right) = F_Z\left(\frac{1/2-1}{1}\right) = F_Z\left(-\frac{1}{2}\right)$$

47.3 More details

More details about the normal distribution can be found in the following subsections.

47.3.1 Multivariate normal distribution

A multivariate generalization of the normal distribution is introduced in the lecture entitled *Multivariate normal distribution* (p. 439).

47.3.2 Linear combinations of normal random variables

The lecture entitled *Linear combinations of normals* (p. 469) explains and proves one of the most important facts about the normal distribution: the linear combination of jointly normal random variables also has a normal distribution.

47.3.3 Quadratic forms involving normal random variables

The lecture entitled *Quadratic forms in normal vectors* (p. 481) discusses quadratic forms involving normal random variables.

47.4 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Let X be a normal random variable with mean $\mu = 3$ and variance $\sigma^2 = 4$. Compute the following probability:

$$\mathrm{P}\left(-0.92 \leq X \leq 6.92\right)$$

Solution

First of all, we need to express the above probability in terms of the distribution function of X:

$$\begin{aligned} & \mathrm{P}\left(-0.92 \leq X \leq 6.92\right) \\ = \ & \mathrm{P}\left(X \leq 6.92\right) - \mathrm{P}\left(X < -0.92\right) \\ \boxed{\mathrm{A}} \ = \ & \mathrm{P}\left(X \leq 6.92\right) - \mathrm{P}\left(X \leq -0.92\right) \\ = \ & F_X\left(6.92\right) - F_X\left(-0.92\right) \end{aligned}$$

where: in step $\boxed{\mathrm{A}}$ we have used the fact that the probability that an absolutely continuous random variable takes on any specific value is equal to zero[11].

Then, we need to express the distribution function of X in terms of the distribution function of a standard normal random variable Z:

$$F_X\left(x\right) = F_Z\left(\frac{x-\mu}{\sigma}\right) = F_Z\left(\frac{x-3}{2}\right)$$

Therefore, the above probability can be expressed as:

$$\begin{aligned} & \mathrm{P}\left(-0.92 \leq X \leq 6.92\right) = F_X\left(6.92\right) - F_X\left(-0.92\right) \\ = \ & F_Z\left(\frac{6.92-3}{2}\right) - F_Z\left(\frac{-0.92-3}{2}\right) \\ = \ & F_Z\left(1.96\right) - F_Z\left(-1.96\right) = 0.975 - 0.025 = 0.95 \end{aligned}$$

where we have used the fact that

$$F_Z\left(1.96\right) = 1 - F_Z\left(-1.96\right) = 0.975$$

which has been discussed above.

[11] See p. 109.

Exercise 2

Let X be a random variable having a normal distribution with mean $\mu = 1$ and variance $\sigma^2 = 16$. Compute the following probability:

$$\mathrm{P}(X > 9)$$

Solution

We need to use the same technique used in the previous exercise and express the probability in terms of the distribution function of a standard normal random variable:

$$\begin{aligned} \mathrm{P}(X > 9) &= 1 - \mathrm{P}(X \leq 9) = 1 - F_X(9) \\ &= 1 - F_Z\left(\frac{9-1}{\sqrt{16}}\right) = 1 - F_Z(2) \\ &= 1 - 0.9772 = 0.0228 \end{aligned}$$

where the value $F_Z(2)$ can be found with a computer algorithm, for example with the MATLAB command

```
normcdf(2)
```

Exercise 3

Suppose the random variable X has a normal distribution with mean $\mu = 1$ and variance $\sigma^2 = 1$. Define the random variable Y as follows:

$$Y = \exp(2 + 3X)$$

Compute the expected value of Y.

Solution

The moment generating function of X is:

$$\begin{aligned} M_X(t) &= \mathrm{E}[\exp(tX)] = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right) \\ &= \exp\left(t + \frac{1}{2}t^2\right) \end{aligned}$$

Using the linearity of the expected value, we obtain:

$$\begin{aligned} \mathrm{E}[Y] &= \mathrm{E}[\exp(2 + 3X)] = \mathrm{E}[\exp(2)\exp(3X)] \\ &= \exp(2)\,\mathrm{E}[\exp(3X)] = \exp(2)\,M_X(3) \\ &= \exp(2)\exp\left(3 + \frac{1}{2}\cdot 9\right) = \exp\left(\frac{19}{2}\right) \end{aligned}$$

Chapter 48

Chi-square distribution

A random variable X has a Chi-square distribution if it can be written as a sum of squares:

$$X = Y_1^2 + \ldots + Y_n^2$$

where $Y_1, \ldots, Y_n$ are n mutually independent[1] standard normal random variables[2].

The importance of the Chi-square distribution stems from the fact that sums of this kind are encountered very often in statistics, especially in the estimation of variance and in hypothesis testing.

48.1 Definition

Chi-square random variables are characterized as follows.

Definition 256 *Let X be an absolutely continuous random variable. Let its support be the set of positive real numbers:*

$$R_X = [0, \infty)$$

*Let $n \in \mathbb{N}$. We say that X has a **Chi-square distribution** with n degrees of freedom if its probability density function[3] is*

$$f_X(x) = \begin{cases} cx^{n/2-1} \exp\left(-\frac{1}{2}x\right) & \text{if } x \in R_X \\ 0 & \text{if } x \notin R_X \end{cases}$$

where c is a constant:

$$c = \frac{1}{2^{n/2}\Gamma(n/2)}$$

and $\Gamma()$ is the Gamma function[4].

The following notation is often employed to indicate that a random variable X has a Chi-square distribution with n degrees of freedom:

$$X \sim \chi^2(n)$$

where the symbol $\sim$ means "is distributed as" and $\chi^2(n)$ indicates a Chi-square distribution with n degrees of freedom.

[1] See p. 233.
[2] See p. 376.
[3] See p. 107.
[4] See p. 55.

48.2 Expected value

The expected value of a Chi-square random variable X is

$$\mathrm{E}[X] = n$$

Proof. It can be derived as follows:

$$\begin{aligned}
\mathrm{E}[X] &= \int_0^\infty x f_X(x)\,dx \\
&= \int_0^\infty x c x^{n/2-1} \exp\left(-\frac{1}{2}x\right) dx \\
&= c\int_0^\infty x^{n/2} \exp\left(-\frac{1}{2}x\right) dx \\
\boxed{\text{A}} &= c\left\{\left[-x^{n/2} 2\exp\left(-\frac{1}{2}x\right)\right]_0^\infty + \int_0^\infty \frac{n}{2} x^{n/2-1} 2\exp\left(-\frac{1}{2}x\right) dx\right\} \\
&= c\left\{(0-0) + n\int_0^\infty x^{n/2-1} \exp\left(-\frac{1}{2}x\right) dx\right\} \\
&= n\int_0^\infty c x^{n/2-1} \exp\left(-\frac{1}{2}x\right) dx \\
&= n\int_0^\infty f_X(x)\,dx \\
\boxed{\text{B}} &= n
\end{aligned}$$

where: in step $\boxed{\text{A}}$ we have performed an integration by parts[5]; in step $\boxed{\text{B}}$ we have used the fact that the integral of a probability density function over its support is equal to 1. ■

48.3 Variance

The variance of a Chi-square random variable X is

$$\mathrm{Var}[X] = 2n$$

Proof. The second moment of X is

$$\begin{aligned}
\mathrm{E}\left[X^2\right] &= \int_0^\infty x^2 f_X(x)\,dx \\
&= \int_0^\infty x^2 c x^{n/2-1} \exp\left(-\frac{1}{2}x\right) dx \\
&= c\int_0^\infty x^{n/2+1} \exp\left(-\frac{1}{2}x\right) dx \\
\boxed{\text{A}} &= c\left\{\left[-x^{n/2+1} 2\exp\left(-\frac{1}{2}x\right)\right]_0^\infty \right. \\
&\quad \left. + \int_0^\infty \left(\frac{n}{2}+1\right) x^{n/2} 2\exp\left(-\frac{1}{2}x\right) dx\right\}
\end{aligned}$$

[5] See p. 51.

$$\begin{aligned}
&= c\left\{(0-0)+(n+2)\int_0^\infty x^{n/2}\exp\left(-\frac{1}{2}x\right)dx\right\} \\
&= c\,(n+2)\left\{\int_0^\infty x^{n/2}\exp\left(-\frac{1}{2}x\right)dx\right\} \\
\boxed{\text{B}} &= c\,(n+2)\left\{\left[-x^{n/2}2\exp\left(-\frac{1}{2}x\right)\right]_0^\infty \right. \\
&\quad \left. +\int_0^\infty \frac{n}{2}x^{n/2-1}2\exp\left(-\frac{1}{2}x\right)dx\right\} \\
&= c\,(n+2)\left\{(0-0)+n\int_0^\infty x^{n/2-1}\exp\left(-\frac{1}{2}x\right)dx\right\} \\
&= (n+2)\,n\int_0^\infty cx^{n/2-1}\exp\left(-\frac{1}{2}x\right)dx \\
&= (n+2)\,n\int_0^\infty f_X(x)\,dx \\
\boxed{\text{C}} &= (n+2)\,n
\end{aligned}$$

where: in step A and B we have performed an integration by parts; in step C we have used the fact that the integral of a probability density function over its support is equal to 1. Furthermore,

$$\mathrm{E}[X]^2 = n^2$$

By employing the usual formula for computing the variance[6], we obtain

$$\begin{aligned}
\mathrm{Var}[X] &= \mathrm{E}\left[X^2\right]-\mathrm{E}[X]^2 \\
&= (n+2)\,n\ -n^2 = n\,(n+2-n) = 2n
\end{aligned}$$

■

48.4 Moment generating function

The moment generating function of a Chi-square random variable X is defined for any $t < \frac{1}{2}$:

$$M_X(t) = (1-2t)^{-n/2}$$

Proof. By using the definition of moment generating function, we get

$$\begin{aligned}
M_X(t) &= \mathrm{E}[\exp(tX)] \\
&= \int_{-\infty}^\infty \exp(tx)\,f_X(x)\,dx \\
&= c\int_0^\infty \exp(tx)\,x^{n/2-1}\exp\left(-\frac{1}{2}x\right)dx \\
&= c\int_0^\infty x^{n/2-1}\exp\left(-\left(\frac{1}{2}-t\right)x\right)dx \\
\boxed{\text{A}} &= c\int_0^\infty \left(\frac{2}{1-2t}y\right)^{n/2-1}\exp(-y)\,\frac{2}{1-2t}dy
\end{aligned}$$

[6] See p. 156.

$$\begin{aligned}
&= c\int_0^\infty \left(\frac{2}{1-2t}\right)^{n/2} y^{n/2-1}\exp(-y)\,dy \\
&= c\left(\frac{2}{1-2t}\right)^{n/2}\int_0^\infty y^{n/2-1}\exp(-y)\,dy \\
\boxed{\text{B}} &= c\left(\frac{2}{1-2t}\right)^{n/2}\Gamma(n/2) \\
\boxed{\text{C}} &= \frac{1}{2^{n/2}\Gamma(n/2)}\left(\frac{2}{1-2t}\right)^{n/2}\Gamma(n/2) \\
&= \frac{1}{2^{n/2}}\frac{2^{n/2}}{(1-2t)^{n/2}} \\
&= (1-2t)^{-n/2}
\end{aligned}$$

where: in step A we have performed the change of variable

$$y = \left(\frac{1}{2} - t\right)x$$

in step B we have used the definition of Gamma function[7]; in step C we have used the definition of c. The integral above is well-defined and finite only when $\frac{1}{2} - t > 0$, i.e., when $t < \frac{1}{2}$. Thus, the moment generating function of a Chi-square random variable exists for any $t < \frac{1}{2}$. ■

48.5 Characteristic function

The characteristic function of a Chi-square random variable X is

$$\varphi_X(t) = (1-2it)^{-n/2}$$

Proof. Using the definition of characteristic function, we get

$$\begin{aligned}
\varphi_X(t) &= \mathrm{E}\left[\exp(itX)\right] \\
&= \int_{-\infty}^{\infty}\exp(itx)\,f_X(x)\,dx \\
&= c\int_0^\infty \exp(itx)\,x^{n/2-1}\exp\left(-\frac{1}{2}x\right)dx \\
\boxed{\text{A}} &= c\int_0^\infty \left(\sum_{k=0}^{\infty}\frac{1}{k!}(itx)^k\right)x^{n/2-1}\exp\left(-\frac{1}{2}x\right)dx \\
&= c\sum_{k=0}^{\infty}\frac{1}{k!}(it)^k\int_0^\infty x^k x^{n/2-1}\exp\left(-\frac{1}{2}x\right)dx \\
&= c\sum_{k=0}^{\infty}\frac{1}{k!}(it)^k\int_0^\infty x^{k+n/2-1}\exp\left(-\frac{1}{2}x\right)dx \\
&= c\sum_{k=0}^{\infty}\frac{1}{k!}(it)^k\,2^{k+n/2}\Gamma(k+n/2)
\end{aligned}$$

[7] See p. 55.

$$\cdot \int_0^\infty \frac{1}{2^{k+n/2}\Gamma\left(k+n/2\right)} x^{k+n/2-1} \exp\left(-\frac{1}{2}x\right) dx$$

$$\boxed{\text{B}} = c\sum_{k=0}^{\infty} \frac{1}{k!} \left(it\right)^k 2^{k+n/2}\Gamma\left(k+n/2\right) \int_0^\infty f_k\left(x\right) dx$$

$$\boxed{\text{C}} = c\sum_{k=0}^{\infty} \frac{1}{k!} \left(it\right)^k 2^{k+n/2}\Gamma\left(k+n/2\right)$$

$$\boxed{\text{D}} = \frac{1}{2^{n/2}\Gamma\left(n/2\right)} \sum_{k=0}^{\infty} \frac{1}{k!} \left(it\right)^k 2^{k+n/2}\Gamma\left(k+n/2\right)$$

$$= \sum_{k=0}^{\infty} \frac{1}{k!} \left(it\right)^k 2^k \frac{\Gamma\left(k+n/2\right)}{\Gamma\left(n/2\right)}$$

$$= \sum_{k=0}^{\infty} \frac{1}{k!} \left(2it\right)^k \frac{\Gamma\left(k+n/2\right)}{\Gamma\left(n/2\right)}$$

$$= 1 + \sum_{k=1}^{\infty} \frac{1}{k!} \left(2it\right)^k \prod_{j=0}^{k-1} \left(\frac{n}{2}+j\right)$$

$$\boxed{\text{E}} = \left(1-2it\right)^{-n/2}$$

where: in step A we have substituted the Taylor series expansion of $\exp\left(itx\right)$; in step B we have defined

$$f_k\left(x\right) = \frac{1}{2^{k+n/2}\Gamma\left(k+n/2\right)} x^{k+n/2-1} \exp\left(-\frac{1}{2}x\right)$$

where $f_k\left(x\right)$ is the probability density function of a Chi-square random variable with $2k+n$ degrees of freedom; in step C we have used the fact that the integral of a probability density function over its support is equal to 1; in step D we have used the definition of c; in step E we have used the fact that

$$1 + \sum_{k=1}^{\infty} \frac{1}{k!} \left(2it\right)^k \prod_{j=0}^{k-1} \left(\frac{n}{2}+j\right)$$

is the Taylor series expansion of $\left(1-2it\right)^{-n/2}$, which you can verify by computing the expansion yourself. ■

48.6 Distribution function

The distribution function of a Chi-square random variable is:

$$F_X\left(x\right) = \frac{\gamma\left(n/2, x/2\right)}{\Gamma\left(n/2\right)}$$

where the function

$$\gamma\left(z,y\right) = \int_{-\infty}^{y} s^{z-1} \exp\left(-s\right) ds$$

is called lower incomplete Gamma function[8] and is usually computed by means of specialized computer algorithms.

Proof. This is proved as follows:

$$\begin{aligned}
F_X(x) &= \int_{-\infty}^{x} f_X(t)\,dt \\
&= \int_{-\infty}^{x} ct^{n/2-1}\exp\left(-\frac{1}{2}t\right)dt \\
\boxed{\text{A}} &= c\int_{-\infty}^{x/2}(2s)^{n/2-1}\exp(-s)\,2ds \\
&= c2^{n/2}\int_{-\infty}^{x/2}s^{n/2-1}\exp(-s)\,ds \\
\boxed{\text{B}} &= \frac{1}{2^{n/2}\Gamma(n/2)}2^{n/2}\int_{-\infty}^{x/2}s^{n/2-1}\exp(-s)\,ds \\
&= \frac{1}{\Gamma(n/2)}\int_{-\infty}^{x/2}s^{n/2-1}\exp(-s)\,ds \\
&= \frac{\gamma(n/2,x/2)}{\Gamma(n/2)}
\end{aligned}$$

where: in step A we have performed a change of variable ($s = t/2$); in step B we have used the definition of c. ■

Usually, it is possible to resort to computer algorithms that directly compute the values of $F_X(x)$. For example, the MATLAB command

```
chi2cdf(x,n)
```

returns the value at the point `x` of the distribution function of a Chi-square random variable with `n` degrees of freedom.

In the past, when computers were not widely available, people used to look up the values of $F_X(x)$ in Chi-square distribution tables. A **Chi-square distribution table** is a table where $F_X(x)$ is tabulated for several values of x and n. For values of x that are not tabulated, approximations of $F_X(x)$ can be computed by interpolation, with the same procedure described for the normal distribution (p. 380).

48.7 More details

In the following subsections you can find more details about the Chi-square distribution.

48.7.1 Sums of independent Chi-square random variables

Let X_1 be a Chi-square random variable with n_1 degrees of freedom and X_2 another Chi-square random variable with n_2 degrees of freedom. If X_1 and X_2 are independent, then their sum has a Chi-square distribution with $n_1 + n_2$ degrees of

[8] See p. 58.

freedom:

$$\begin{array}{l} X_1 \sim \chi^2(n_1),\ X_2 \sim \chi^2(n_2) \\ X_1 \text{ and } X_2 \text{ are independent} \end{array} \Longrightarrow X_1 + X_2 \sim \chi^2(n_1 + n_2)$$

This can be generalized to sums of more than two Chi-square random variables, provided they are mutually independent:

$$\begin{array}{l} X_i \sim \chi^2(n_i) \text{ for } i = 1, \ldots, k \\ X_1, X_2, \ldots, X_k \text{ are mutually independent} \end{array} \Longrightarrow \sum_{i=1}^{k} X_i \sim \chi^2\left(\sum_{i=1}^{k} n_i\right)$$

Proof. This can be easily proved using moment generating functions. The moment generating function of X_i is

$$M_{X_i}(t) = (1-2t)^{-n_i/2}$$

Define

$$X = \sum_{i=1}^{k} X_i$$

The moment generating function of a sum of mutually independent random variables is just the product of their moment generating functions[9]:

$$\begin{aligned} M_X(t) &= \prod_{i=1}^{k} M_{X_i}(t) \\ &= \prod_{i=1}^{k} (1-2t)^{-n_i/2} \\ &= (1-2t)^{-\sum_{i=1}^{k} n_i/2} \\ &= (1-2t)^{-n/2} \end{aligned}$$

where

$$n = \sum_{i=1}^{k} n_i$$

Therefore, the moment generating function of X is the moment generating function of a Chi-square random variable with n degrees of freedom. As a consequence, X is a Chi-square random variable with n degrees of freedom. ■

48.7.2 Relation to the standard normal distribution

Let Z be a standard normal random variable[10] and let X be its square:

$$X = Z^2$$

Then X is a Chi-square random variable with 1 degree of freedom.
Proof. For $x \geq 0$, the distribution function of X is

$$F_X(x)$$

[9] See p. 293.
[10] See p. 376.

$$\begin{aligned}
\boxed{\text{A}} &= \mathrm{P}\left(X \leq x\right) \\
\boxed{\text{B}} &= \mathrm{P}\left(Z^2 \leq x\right) \\
\boxed{\text{C}} &= \mathrm{P}\left(-x^{1/2} \leq Z \leq x^{1/2}\right) \\
\boxed{\text{D}} &= \int_{-x^{1/2}}^{x^{1/2}} f_Z\left(z\right) dz
\end{aligned}$$

where: in step $\boxed{\text{A}}$ we have used the definition of distribution function; in step $\boxed{\text{B}}$ we have used the definition of X; in step $\boxed{\text{C}}$ we have taken the square root on both sides of the inequality; in step $\boxed{\text{D}}$ $f_Z\left(z\right)$ is the probability density function of a standard normal random variable:

$$f_Z\left(z\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right)$$

For $x < 0$, $F_X\left(x\right) = 0$, because X, being a square, cannot be negative. By using Leibniz integral rule[11] and the fact that the density function is the derivative of the distribution function[12], the probability density function of X, denoted by $f_X\left(x\right)$, is obtained as follows (for $x \geq 0$):

$$\begin{aligned}
f_X\left(x\right) &= \frac{dF_X\left(x\right)}{dx} \\
&= \frac{d}{dx}\int_{-x^{1/2}}^{x^{1/2}} f_Z\left(z\right) dz \\
&= f_Z\left(x^{1/2}\right)\frac{d\left(x^{1/2}\right)}{dx} - f_Z\left(-x^{1/2}\right)\frac{d\left(-x^{1/2}\right)}{dx} \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(x^{1/2}\right)^2\right)\frac{1}{2}x^{-1/2} \\
&\quad -\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(-x^{1/2}\right)^2\right)\left(-\frac{1}{2}x^{-1/2}\right) \\
&= \frac{1}{\sqrt{2\pi}}\frac{1}{2}x^{-1/2}\exp\left(-\frac{1}{2}x\right) + \frac{1}{\sqrt{2\pi}}\frac{1}{2}x^{-1/2}\exp\left(-\frac{1}{2}x\right) \\
&= \frac{1}{\sqrt{2\pi}}x^{-1/2}\exp\left(-\frac{1}{2}x\right) \\
&= \frac{1}{2^{1/2}\Gamma\left(1/2\right)}x^{1/2-1}\exp\left(-\frac{1}{2}x\right)
\end{aligned}$$

where in the last step we have used the fact that[13] $\Gamma\left(1/2\right) = \sqrt{\pi}$. For $x < 0$, it trivially holds that $f_X\left(x\right) = 0$. Therefore,

$$f_X\left(x\right) = \begin{cases} \frac{1}{2^{1/2}\Gamma(1/2)}x^{1/2-1}\exp\left(-\frac{1}{2}x\right) & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

which is the probability density function of a Chi-square random variable with 1 degree of freedom. ■

[11] See p. 52.
[12] See p. 109.
[13] See p. 57.

48.7.3 Relation to the standard normal distribution (2)

Combining the results obtained in the previous two subsections, we obtain that the sum of squares of n independent standard normal random variables is a Chi-square random variable with n degrees of freedom.

48.8 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Let X be a Chi-square random variable with 3 degrees of freedom. Compute the probability

$$\mathrm{P}\left(0.35 \leq X \leq 7.81\right)$$

Solution

First of all, we need to express the above probability in terms of the distribution function of X:

$$\begin{aligned} \mathrm{P}\left(0.35 \leq X \leq 7.81\right) &= \mathrm{P}\left(X \leq 7.81\right) - \mathrm{P}\left(X < 0.35\right) \\ \boxed{\text{A}} \quad &= \mathrm{P}\left(X \leq 7.81\right) - \mathrm{P}\left(X \leq 0.35\right) \\ &= F_X\left(7.81\right) - F_X\left(0.35\right) \\ \boxed{\text{B}} \quad &= 0.95 - 0.05 = 0.90 \end{aligned}$$

where: in step A we have used the fact that the probability that an absolutely continuous random variable takes on any specific value is equal to zero[14]; in step B the values

$$\begin{aligned} F_X\left(7.81\right) &= 0.95 \\ F_X\left(0.35\right) &= 0.05 \end{aligned}$$

can be computed with a computer algorithm or found in a Chi-square distribution table.

Exercise 2

Let X_1 and X_2 be two independent normal random variables having mean $\mu = 0$ and variance $\sigma^2 = 16$. Compute the probability

$$\mathrm{P}\left(X_1^2 + X_2^2 > 8\right)$$

Solution

First of all, the two variables X_1 and X_2 can be written as

$$X_1 = 4Z_1$$

[14] See p. 109.

$$X_2 = 4Z_2$$

where Z_1 and Z_2 are two standard normal random variables. Thus, we can write

$$\begin{aligned} \mathrm{P}\left(X_1^2 + X_2^2 > 8\right) &= \mathrm{P}\left(16Z_1^2 + 16Z_2^2 > 8\right) \\ &= \mathrm{P}\left(Z_1^2 + Z_2^2 > \frac{8}{16}\right) \\ &= \mathrm{P}\left(Z_1^2 + Z_2^2 > \frac{1}{2}\right) \end{aligned}$$

but the sum $Z_1^2 + Z_2^2$ has a Chi-square distribution with 2 degrees of freedom. Therefore,

$$\begin{aligned} \mathrm{P}\left(X_1^2 + X_2^2 > 8\right) &= \mathrm{P}\left(Z_1^2 + Z_2^2 > \frac{1}{2}\right) \\ &= 1 - \mathrm{P}\left(Z_1^2 + Z_2^2 \leq \frac{1}{2}\right) \\ &= 1 - F_Y\left(\frac{1}{2}\right) \end{aligned}$$

where $F_Y\left(1/2\right)$ is the distribution function of a Chi-square random variable Y with 2 degrees of freedom, evaluated at the point $y = \frac{1}{2}$. Using any computer program, we can find

$$F_Y\left(\frac{1}{2}\right) = 0.2212$$

Exercise 3

Suppose the random variable X has a Chi-square distribution with 5 degrees of freedom. Define the random variable Y as follows:

$$Y = \exp\left(1 - X\right)$$

Compute the expected value of Y.

Solution

The expected value of Y can be easily calculated using the moment generating function of X:

$$M_X\left(t\right) = \mathrm{E}\left[\exp\left(tX\right)\right] = \left(1 - 2t\right)^{-5/2}$$

Now, exploiting the linearity of the expected value, we obtain

$$\begin{aligned} \mathrm{E}\left[Y\right] &= \mathrm{E}\left[\exp\left(1 - X\right)\right] = \mathrm{E}\left[\exp\left(1\right)\exp\left(-X\right)\right] \\ &= \exp\left(1\right)\mathrm{E}\left[\exp\left(-X\right)\right] = \exp\left(1\right)M_X\left(-1\right) \\ &= \exp\left(1\right)3^{-5/2} \end{aligned}$$

Chapter 49

Gamma distribution

The Gamma distribution can be thought of as a generalization of the Chi-square distribution[1]. If a random variable Z has a Chi-square distribution with n degrees of freedom and h is a strictly positive constant, then the random variable X defined as

$$X = \frac{h}{n} Z$$

has a Gamma distribution with parameters n and h.

49.1 Definition

Gamma random variables are characterized as follows:

Definition 257 *Let X be an absolutely continuous random variable. Let its support be the set of positive real numbers:*

$$R_X = [0, \infty)$$

*Let $n, h \in \mathbb{R}_{++}$. We say that X has a **Gamma distribution** with parameters n and h if its probability density function[2] is*

$$f_X(x) = \begin{cases} cx^{n/2-1} \exp\left(-\frac{n}{h}\frac{1}{2}x\right) & \text{if } x \in R_X \\ 0 & \text{if } x \notin R_X \end{cases}$$

where c is a constant:

$$c = \frac{(n/h)^{n/2}}{2^{n/2}\Gamma(n/2)}$$

and $\Gamma()$ is the Gamma function[3].

A random variable having a Gamma distribution is also called a Gamma random variable.

[1] See p. 387.
[2] See p. 107.
[3] See p. 55.

49.2 Expected value

The expected value of a Gamma random variable X is

$$\mathrm{E}[X] = h$$

Proof. It can be derived as follows:

$$\begin{aligned}
\mathrm{E}[X] &= \int_0^\infty x f_X(x)\,dx \\
&= \int_0^\infty x c x^{n/2-1} \exp\left(-\frac{n}{h}\frac{1}{2}x\right) dx \\
&= c\int_0^\infty x^{n/2} \exp\left(-\frac{n}{h}\frac{1}{2}x\right) dx \\
\boxed{\text{A}} &= c\left\{\left[-x^{n/2}2\frac{h}{n}\exp\left(-\frac{n}{h}\frac{1}{2}x\right)\right]_0^\infty \right. \\
&\quad \left. + \int_0^\infty \frac{n}{2}x^{n/2-1}2\frac{h}{n}\exp\left(-\frac{n}{h}\frac{1}{2}x\right) dx\right\} \\
&= c\left\{(0-0) + h\int_0^\infty x^{n/2-1}\exp\left(-\frac{n}{h}\frac{1}{2}x\right) dx\right\} \\
&= h\int_0^\infty c x^{n/2-1}\exp\left(-\frac{n}{h}\frac{1}{2}x\right) dx \\
&= h\int_0^\infty f_X(x)\,dx \\
\boxed{\text{B}} &= h
\end{aligned}$$

where: in step A we have performed an integration by parts[4]; in step B we have used the fact that the integral of a probability density function over its support is equal to 1. ■

49.3 Variance

The variance of a Gamma random variable X is

$$\mathrm{Var}[X] = 2\frac{h^2}{n}$$

Proof. It can be derived thanks to the usual formula for computing the variance[5]:

$$\begin{aligned}
\mathrm{E}\left[X^2\right] &= \int_0^\infty x^2 f_X(x)\,dx \\
&= \int_0^\infty x^2 c x^{n/2-1}\exp\left(-\frac{n}{h}\frac{1}{2}x\right) dx \\
&= c\int_0^\infty x^{n/2+1}\exp\left(-\frac{n}{h}\frac{1}{2}x\right) dx
\end{aligned}$$

[4] See p. 51.

[5] $\mathrm{Var}[X] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2$. See p. 156.

$$\begin{aligned}
\boxed{\text{A}} &= c\left\{\left[-x^{n/2+1}2\frac{h}{n}\exp\left(-\frac{n}{h}\frac{1}{2}x\right)\right]_0^\infty \right. \\
&\quad \left. +\int_0^\infty \left(\frac{n}{2}+1\right)x^{n/2}2\frac{h}{n}\exp\left(-\frac{n}{h}\frac{1}{2}x\right)dx\right\} \\
&= c\left\{(0-0)+(n+2)\frac{h}{n}\int_0^\infty x^{n/2}\exp\left(-\frac{n}{h}\frac{1}{2}x\right)dx\right\} \\
&= c\,(n+2)\frac{h}{n}\left\{\int_0^\infty x^{n/2}\exp\left(-\frac{n}{h}\frac{1}{2}x\right)dx\right\} \\
\boxed{\text{B}} &= c\,(n+2)\frac{h}{n}\left\{\left[-x^{n/2}2\frac{h}{n}\exp\left(-\frac{n}{h}\frac{1}{2}x\right)\right]_0^\infty \right. \\
&\quad \left. +\int_0^\infty \frac{n}{2}x^{n/2-1}2\frac{h}{n}\exp\left(-\frac{n}{h}\frac{1}{2}x\right)dx\right\} \\
&= c\,(n+2)\frac{h}{n}\left\{(0-0)+h\int_0^\infty x^{n/2-1}\exp\left(-\frac{n}{h}\frac{1}{2}x\right)dx\right\} \\
&= (n+2)\frac{h^2}{n}\int_0^\infty cx^{n/2-1}\exp\left(-\frac{n}{h}\frac{1}{2}x\right)dx \\
&= (n+2)\frac{h^2}{n}\int_0^\infty f_X(x)\,dx \\
\boxed{\text{C}} &= (n+2)\frac{h^2}{n}
\end{aligned}$$

where: in step A and B we have performed an integration by parts; in step C we have used the fact that the integral of a probability density function over its support is equal to 1. Finally:

$$\mathrm{E}[X]^2 = h^2$$

and

$$\begin{aligned}
\mathrm{Var}[X] &= \mathrm{E}\left[X^2\right]-\mathrm{E}[X]^2 = (n+2)\frac{h^2}{n} - h^2 \\
&= (n+2-n)\frac{h^2}{n} = 2\frac{h^2}{n}
\end{aligned}$$

■

49.4 Moment generating function

The moment generating function of a Gamma random variable X is defined for any $t < \frac{n}{2h}$:

$$M_X(t) = \left(1-\frac{2h}{n}t\right)^{-n/2}$$

Proof. Using the definition of moment generating function:

$$\begin{aligned}
M_X(t) &= \mathrm{E}[\exp(tX)] \\
&= \int_{-\infty}^{\infty}\exp(tx)\,f_X(x)\,dx \\
&= \int_0^\infty \exp(tx)\,\frac{(n/h)^{n/2}}{2^{n/2}\Gamma(n/2)}x^{n/2-1}\exp\left(-\frac{n}{h}\frac{1}{2}x\right)dx
\end{aligned}$$

$$
\begin{aligned}
&= \int_0^\infty \frac{(n/h)^{n/2}}{2^{n/2}\Gamma(n/2)} x^{n/2-1} \exp\left(-\frac{n}{h}\frac{1}{2}x + tx\right) dx \\
&= \int_0^\infty \frac{(n/h)^{n/2}}{2^{n/2}\Gamma(n/2)} x^{n/2-1} \exp\left(-\frac{1}{h}\frac{n}{2}x - \frac{-2t}{n}\frac{n}{2}x\right) dx \\
&= \int_0^\infty \frac{(1/h)^{n/2} n^{n/2}}{2^{n/2}\Gamma(n/2)} x^{n/2-1} \exp\left(-\left(\frac{1}{h} - \frac{2t}{n}\right)\frac{n}{2}x\right) dx \\
&= (1/h)^{n/2}\left(\frac{1}{h} - \frac{2t}{n}\right)^{-n/2} \\
&\quad \cdot \int_0^\infty \frac{\left(\frac{1}{h} - \frac{2t}{n}\right)^{n/2} n^{n/2}}{2^{n/2}\Gamma(n/2)} x^{n/2-1} \exp\left(-\left(\frac{1}{h} - \frac{2t}{n}\right)\frac{n}{2}x\right) dx \\
&= (1/h)^{n/2}\left(\frac{1}{h} - \frac{2t}{n}\right)^{-n/2}
\end{aligned}
$$

where the integral equals 1 because it is the integral of the probability density function of a Gamma random variable with parameters n and $\left(\frac{1}{h} - \frac{2t}{n}\right)^{-1}$. Thus:

$$
\begin{aligned}
M_X(t) &= (1/h)^{n/2}\left(\frac{1}{h} - \frac{2t}{n}\right)^{-n/2} \\
&= (h)^{-n/2}\left(\frac{1}{h} - \frac{2t}{n}\right)^{-n/2} \\
&= \left(1 - \frac{2h}{n}t\right)^{-n/2}
\end{aligned}
$$

Of course, the above integrals converge only if $\left(\frac{1}{h} - \frac{2t}{n}\right) > 0$, i.e. only if $t < \frac{n}{2h}$. Therefore, the moment generating function of a Gamma random variable exists for all $t < \frac{n}{2h}$. ■

49.5 Characteristic function

The characteristic function of a Gamma random variable X is:

$$
\varphi_X(t) = \left(1 - \frac{2h}{n}it\right)^{-n/2}
$$

Proof. Using the definition of characteristic function:

$$
\begin{aligned}
& \varphi_X(t) \\
&= \mathrm{E}\left[\exp(itX)\right] \\
&= \int_{-\infty}^{\infty} \exp(itx) f_X(x)\, dx \\
&= c\int_0^\infty \exp(itx)\, x^{n/2-1} \exp\left(-\frac{n}{h}\frac{1}{2}x\right) dx \\
\boxed{\text{A}} &= c\int_0^\infty \left(\sum_{k=0}^{\infty} \frac{1}{k!}(itx)^k\right) x^{n/2-1} \exp\left(-\frac{n}{h}\frac{1}{2}x\right) dx
\end{aligned}
$$

$$
\begin{aligned}
&= c\sum_{k=0}^{\infty}\frac{1}{k!}(it)^k\int_0^{\infty}x^k x^{n/2-1}\exp\left(-\frac{1}{2}\frac{n}{h}x\right)dx \\
&= c\sum_{k=0}^{\infty}\frac{1}{k!}(it)^k\int_0^{\infty}x^{k+n/2-1}\exp\left(-\frac{1}{2}\frac{2k+n}{h\frac{2k+n}{n}}x\right)dx \\
&= c\sum_{k=0}^{\infty}\frac{1}{k!}(it)^k 2^{k+n/2}\Gamma(k+n/2)(n/h)^{-k-n/2} \\
&\quad \cdot\int_0^{\infty}\frac{(n/h)^{k+n/2}}{2^{k+n/2}\Gamma(k+n/2)}x^{k+n/2-1}\exp\left(-\frac{1}{2}\frac{2k+n}{h\frac{2k+n}{n}}x\right)dx \\
\boxed{\text{B}} &= c\sum_{k=0}^{\infty}\frac{1}{k!}(it)^k 2^{k+n/2}\Gamma(k+n/2)(n/h)^{-k-n/2}\int_0^{\infty}f_k(x)\,dx \\
\boxed{\text{C}} &= c\sum_{k=0}^{\infty}\frac{1}{k!}(it)^k 2^{k+n/2}\Gamma(k+n/2)(n/h)^{-k-n/2} \\
\boxed{\text{D}} &= \frac{(n/h)^{n/2}}{2^{n/2}\Gamma(n/2)}\sum_{k=0}^{\infty}\frac{1}{k!}(it)^k 2^{k+n/2}\Gamma(k+n/2)(n/h)^{-k-n/2} \\
&= \frac{1}{\Gamma(n/2)}\sum_{k=0}^{\infty}\frac{1}{k!}(it)^k 2^k\Gamma(k+n/2)(n/h)^{-k} \\
&= \sum_{k=0}^{\infty}\frac{1}{k!}(it)^k\left(\frac{2h}{n}\right)^k\frac{\Gamma(k+n/2)}{\Gamma(n/2)} \\
&= \sum_{k=0}^{\infty}\frac{1}{k!}\left(\frac{2h}{n}it\right)^k\frac{\Gamma(k+n/2)}{\Gamma(n/2)} \\
&= 1+\sum_{k=1}^{\infty}\frac{1}{k!}\left(\frac{2h}{n}it\right)^k\prod_{j=0}^{k-1}\left(\frac{n}{2}+j\right) \\
\boxed{\text{E}} &= \left(1-\frac{2h}{n}it\right)^{-n/2}
\end{aligned}
$$

where: in step A we have substituted $\exp(itx)$ with its Taylor series expansion; in step B we have have defined

$$f_k(x)=\frac{(n/h)^{k+n/2}}{2^{k+n/2}\Gamma(k+n/2)}x^{k+n/2-1}\exp\left(-\frac{1}{2}\frac{2k+n}{h\frac{2k+n}{n}}x\right)$$

where $f_k(x)$ is the probability density function of a Gamma random variable with parameters $2k+n$ and $h\frac{2k+n}{n}$; in step C we have used the fact that probability density functions integrate to 1; in step D we have used the definition of c; in step E we have recognized that

$$1+\sum_{k=1}^{\infty}\frac{1}{k!}\left(\frac{2h}{n}it\right)^k\prod_{j=0}^{k-1}\left(\frac{n}{2}+j\right)$$

is the Taylor series expansion of $\left(1-\frac{2h}{n}it\right)^{-n/2}$. ■

49.6 Distribution function

The distribution function of a Gamma random variable is:

$$F_X(x) = \frac{\gamma(n/2, nx/2h)}{\Gamma(n/2)}$$

where the function

$$\gamma(z, y) = \int_{-\infty}^{y} s^{z-1} \exp(-s)\, ds$$

is called lower incomplete Gamma function[6] and is usually evaluated using specialized computer algorithms.

Proof. This is proved as follows:

$$\begin{aligned}
F_X(x) &= \int_{-\infty}^{x} f_X(t)\, dt \\
&= \int_{-\infty}^{x} c t^{n/2-1} \exp\left(-\frac{n}{h}\frac{1}{2}t\right) dt \\
\boxed{\text{A}} &= c\int_{-\infty}^{nx/2h} \left(\frac{2h}{n}s\right)^{n/2-1} \exp(-s)\frac{2h}{n} ds \\
&= c\left(\frac{2h}{n}\right)^{n/2} \int_{-\infty}^{nx/2h} s^{n/2-1}\exp(-s)\, ds \\
\boxed{\text{B}} &= \frac{(n/h)^{n/2}}{2^{n/2}\Gamma(n/2)}\left(\frac{2h}{n}\right)^{n/2} \int_{-\infty}^{nx/2h} s^{n/2-1}\exp(-s)\, ds \\
&= \frac{1}{\Gamma(n/2)}\int_{-\infty}^{nx/2h} s^{n/2-1}\exp(-s)\, ds \\
&= \frac{\gamma(n/2, nx/2h)}{\Gamma(n/2)}
\end{aligned}$$

where: in step $\boxed{\text{A}}$ we have performed a change of variable ($s = \frac{n}{2h}t$); in step $\boxed{\text{B}}$ we have used the definition of c. ■

49.7 More details

In the following subsections you can find more details about the Gamma distribution.

49.7.1 Relation to the Chi-square distribution

The Gamma distribution is a scaled Chi-square distribution.

Proposition 258 *If a variable X has a Gamma distribution with parameters n and h, then:*

$$X = \frac{h}{n}Z$$

where Z has a Chi-square distribution[7] with n degrees of freedom.

[6] See p. 58.
[7] See p. 387.

Proof. This can be easily proved using the formula for the density of a function[8] of an absolutely continuous variable:

$$\begin{aligned} f_X(x) &= f_Z\left(g^{-1}(x)\right)\frac{dg^{-1}(x)}{dx} \\ &= f_Z\left(\frac{n}{h}x\right)\frac{n}{h} \end{aligned}$$

The density function of a Chi-square random variable with n degrees of freedom is

$$f_Z(z) = \begin{cases} kz^{n/2-1}\exp\left(-\frac{1}{2}z\right) & \text{if } x \in [0,\infty) \\ 0 & \text{otherwise} \end{cases}$$

where

$$k = \frac{1}{2^{n/2}\Gamma(n/2)}$$

Therefore:

$$\begin{aligned} f_X(x) &= f_Z\left(\frac{n}{h}x\right)\frac{n}{h} \\ &= \begin{cases} k\left(\frac{n}{h}\right)^{n/2}x^{n/2-1}\exp\left(-\frac{1}{2}\frac{n}{h}x\right) & \text{if } x \in [0,\infty) \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

which is the density of a Gamma distribution with parameters n and h. ■

Thus, the Chi-square distribution is a special case of the Gamma distribution, because, when $h = n$, we have:

$$X = \frac{h}{n}Z = \frac{n}{n}Z = Z$$

In other words, a Gamma distribution with parameters n and $h = n$ is just a Chi square distribution with n degrees of freedom.

49.7.2 Multiplication by a constant

Multiplying a Gamma random variable by a strictly positive constant one obtains another Gamma random variable.

Proposition 259 *If X is a Gamma random variable with parameters n and h, then the random variable Y defined as*

$$Y = cX \qquad (c \in \mathbb{R}_{++})$$

has a Gamma distribution with parameters n and ch.

Proof. This can be easily seen using the result from the previous subsection:

$$X = \frac{h}{n}Z$$

where Z has a Chi-square distribution with n degrees of freedom. Therefore:

$$Y = cX = c\left(\frac{h}{n}Z\right) = \frac{ch}{n}Z$$

In other words, Y is equal to a Chi-square random variable with n degrees of freedom, divided by n and multiplied by ch. Therefore, it has a Gamma distribution with parameters n and ch. ■

[8] See p. 265. Note that $X = g(Z) = \frac{h}{n}Z$ is a strictly increasing function of Z, since $\frac{h}{n}$ is strictly positive

49.7.3 Relation to the normal distribution

In the lecture entitled *Chi-square distribution* (p. 387) we have explained that a Chi-square random variable Z with $n \in \mathbb{N}$ degrees of freedom can be written as a sum of squares of n independent normal random variables $W_1, \ldots, W_n$ having mean 0 and variance 1:

$$Z = W_1^2 + \ldots + W_n^2$$

In the previous subsections we have seen that a variable X having a Gamma distribution with parameters n and h can be written as

$$X = \frac{h}{n} Z$$

where Z has a Chi-square distribution with n degrees of freedom.

Putting these two things together, we obtain

$$\begin{aligned} X &= \frac{h}{n} Z = \frac{h}{n} \left(W_1^2 + \ldots + W_n^2 \right) \\ &= \left(\sqrt{\frac{h}{n}} W_1 \right)^2 + \ldots + \left(\sqrt{\frac{h}{n}} W_n \right)^2 \\ &= Y_1^2 + \ldots + Y_n^2 \end{aligned}$$

where we have defined

$$Y_i = \sqrt{\frac{h}{n}} W_i \quad , i = 1, \ldots, n$$

But the variables Y_i are normal random variables with mean 0 and variance $\frac{h}{n}$. Therefore, a Gamma random variable with parameters n and h can be seen as a sum of squares of n independent normal random variables having mean 0 and variance h/n.

49.8 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Let X_1 and X_2 be two independent Chi-square random variables having 3 and 5 degrees of freedom respectively. Consider the following random variables:

$$\begin{aligned} Y_1 &= 2X_1 \\ Y_2 &= \frac{1}{3} X_2 \\ Y_3 &= 3X_1 + 3X_2 \end{aligned}$$

What distribution do they have?

Solution

Being multiples of Chi-square random variables, the variables Y_1, Y_2 and Y_3 all have a Gamma distribution. The random variable X_1 has $n = 3$ degrees of freedom and the random variable Y_1 can be written as

$$Y_1 = \frac{h}{n} X_1$$

where $h = 6$. Therefore Y_1 has a Gamma distribution with parameters $n = 3$ and $h = 6$. The random variable X_2 has $n = 5$ degrees of freedom and the random variable Y_2 can be written as

$$Y_2 = \frac{h}{n} X_2$$

where $h = 5/3$. Therefore Y_2 has a Gamma distribution with parameters $n = 5$ and $h = 5/3$. The random variable $X_1 + X_2$ has a Chi-square distribution with $n = 3 + 5 = 8$ degrees of freedom, because X_1 and X_2 are independent[9], and the random variable Y_3 can be written as

$$Y_3 = \frac{h}{n} (X_1 + X_2)$$

where $h = 24$. Therefore Y_3 has a Gamma distribution with parameters $n = 8$ and $h = 24$.

Exercise 2

Let X be a random variable having a Gamma distribution with parameters $n = 4$ and $h = 2$. Define the following random variables:

$$\begin{aligned} Y_1 &= \frac{1}{2} X \\ Y_2 &= 5X \\ Y_3 &= 2X \end{aligned}$$

What distribution do these variables have?

Solution

Multiplying a Gamma random variable by a strictly positive constant one still obtains a Gamma random variable. In particular, the random variable Y_1 is a Gamma random variable with parameters $n = 4$ and

$$h = 2 \cdot \frac{1}{2} = 1$$

The random variable Y_2 is a Gamma random variable with parameters $n = 4$ and

$$h = 2 \cdot 5 = 10$$

The random variable Y_3 is a Gamma random variable with parameters $n = 4$ and

$$h = 2 \cdot 2 = 4$$

The random variable Y_3 is also a Chi-square random variable with 4 degrees of freedom (remember that a Gamma random variable with parameters n and h is also a Chi-square random variable when $n = h$).

[9] See the lecture entitled *Chi-square distribution* (p. 387).

Exercise 3

Let X_1, X_2 and X_3 be mutually independent normal random variables having mean $\mu = 0$ and variance $\sigma^2 = 3$. Consider the random variable

$$X = 2X_1^2 + 2X_2^2 + 2X_3^2$$

What distribution does X have?

Solution

The random variable X can be written as

$$\begin{aligned} X &= 2\left(\left(\sqrt{3}Z_1\right)^2 + \left(\sqrt{3}Z_2\right)^2 + \left(\sqrt{3}Z_3\right)^2\right) \\ &= \frac{18}{3}\left(Z_1^2 + Z_2^2 + Z_3^2\right) \end{aligned}$$

where Z_1, Z_2 and Z_3 are mutually independent standard normal random variables. The sum $Z_1^2 + Z_2^2 + Z_3^2$ has a Chi-square distribution with 3 degrees of freedom. Therefore X has a Gamma distribution with parameters $n = 3$ and $h = 18$.

Chapter 50

Student's t distribution

A random variable X has a **standard Student's t distribution** with n degrees of freedom if it can be written as a ratio:

$$X = \frac{Y}{\sqrt{Z}}$$

between a standard normal random variable[1] Y and the square root of a Gamma random variable[2] Z with parameters n and $h = 1$, independent of Y.

Equivalently, we can write

$$X = \frac{Y}{\sqrt{\chi^2_n / n}}$$

where χ^2_n is a Chi-square random variable[3] with n degrees of freedom (dividing by n a Chi-square random variable with n degrees of freedom, one obtains a Gamma random variable with parameters n and $h = 1$ - see the lecture entilled *Gamma distribution* - p. 397).

A random variable X has a **non-standard Student's t distribution** with mean μ, scale σ^2 and n degrees of freedom if it can be written as a linear transformation of a standard Student's t random variable:

$$X = \mu + \sigma \frac{Y}{\sqrt{Z}}$$

where Y and Z are defined as before.

The importance of Student's t distribution stems from the fact that ratios and linearly transformed ratios of this kind are encountered very often in statistics (see e.g. the lecture entitled *Hypothesis tests about the mean* - p. 619).

We first introduce the standard Student's t distribution. We then deal with the non-standard Student's t distribution.

50.1 The standard Student's t distribution

The standard Student's t distribution is a special case of Student's t distribution. By first explaining this special case, the exposition of the more general case is greatly facilitated.

[1] See p. 375.
[2] See p. 397.
[3] See p. 387.

50.1.1 Definition

The standard Student's t distribution is characterized as follows:

Definition 260 *Let X be an absolutely continuous random variable. Let its support be the whole set of real numbers:*

$$R_X = \mathbb{R}$$

*Let $n \in \mathbb{R}_{++}$. We say that X has a **standard Student's t distribution** with n degrees of freedom if its probability density function[4] is*

$$f_X(x) = c\left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}$$

where c is a constant:

$$c = \frac{1}{\sqrt{n}} \frac{1}{B\left(\frac{n}{2}, \frac{1}{2}\right)}$$

and $B()$ is the Beta function[5].

Usually the number of degrees of freedom is integer ($n \in \mathbb{N}$), but it can also be real ($n \in \mathbb{R}_{++}$).

50.1.2 Relation to the normal and Gamma distributions

A standard Student's t random variable can be written as a normal random variable whose variance is equal to the reciprocal of a Gamma random variable, as shown by the following proposition:

Proposition 261 (Integral representation) *The probability density function of X can be written as*

$$f_X(x) = \int_0^\infty f_{X|Z=z}(x) f_Z(z) dz$$

where:

1. *$f_{X|Z=z}(x)$ is the probability density function of a normal distribution with mean 0 and variance $\sigma^2 = \frac{1}{z}$:*

$$\begin{aligned} f_{X|Z=z}(x) &= \left(2\pi\sigma^2\right)^{-1/2} \exp\left(-\frac{1}{2}\frac{x^2}{\sigma^2}\right) \\ &= (2\pi)^{-1/2} z^{1/2} \exp\left(-\frac{1}{2} z x^2\right) \end{aligned}$$

2. *$f_Z(z)$ is the probability density function of a Gamma random variable with parameters n and $h = 1$:*

$$f_Z(z) = c z^{n/2-1} \exp\left(-n\frac{1}{2}z\right)$$

where

$$c = \frac{n^{n/2}}{2^{n/2}\Gamma(n/2)}$$

[4] See p. 107.
[5] See p. 59.

Proof. We need to prove that

$$f_X(x) = \int_0^\infty f_{X|Z=z}(x) f_Z(z) dz$$

where

$$f_{X|Z=z}(x) = (2\pi)^{-1/2} z^{1/2} \exp\left(-\frac{1}{2} z x^2\right)$$

and

$$f_Z(z) = c z^{n/2-1} \exp\left(-n\frac{1}{2} z\right)$$

Let us start from the integrand function:

$$\begin{aligned}
& f_{X|Z=z}(x) f_Z(z) \\
= \; & (2\pi)^{-1/2} z^{1/2} \exp\left(-\frac{1}{2} z x^2\right) c z^{n/2-1} \exp\left(-n\frac{1}{2} z\right) \\
= \; & (2\pi)^{-1/2} c z^{(n+1)/2-1} \exp\left(-\left(x^2+n\right)\frac{1}{2} z\right) \\
= \; & (2\pi)^{-1/2} c z^{(n+1)/2-1} \exp\left(-\frac{n+1}{\left(\frac{n+1}{x^2+n}\right)}\frac{1}{2} z\right) \\
= \; & (2\pi)^{-1/2} c\frac{1}{c_2} c_2 z^{(n+1)/2-1} \exp\left(-\frac{n+1}{\left(\frac{n+1}{x^2+n}\right)}\frac{1}{2} z\right) \\
= \; & (2\pi)^{-1/2} c\frac{1}{c_2} f_{Z|X=x}(z)
\end{aligned}$$

where

$$\begin{aligned}
c_2 &= \frac{\left(\left(n+1\right)/\left(\frac{n+1}{x^2+n}\right)\right)^{(n+1)/2}}{2^{(n+1)/2}\Gamma\left(\left(n+1\right)/2\right)} \\
&= \frac{\left(x^2+n\right)^{(n+1)/2}}{2^{n/2}2^{1/2}\Gamma\left(\frac{n}{2}+\frac{1}{2}\right)}
\end{aligned}$$

and $f_{Z|X=x}(z)$ is the probability density function of a random variable having a Gamma distribution with parameters $n+1$ and $\frac{n+1}{x^2+n}$. Therefore,

$$\begin{aligned}
& \int_0^\infty f_{X|Z=z}(x) f_Z(z) dz \\
= \; & \int_0^\infty (2\pi)^{-1/2} c\frac{1}{c_2} f_{Z|X=x}(z) dz \\
\boxed{\text{A}} \; = \; & (2\pi)^{-1/2} c\frac{1}{c_2}\int_0^\infty f_{Z|X=x}(z) dz \\
\boxed{\text{B}} \; = \; & (2\pi)^{-1/2} c\frac{1}{c_2} \\
= \; & (2\pi)^{-1/2} \frac{n^{n/2}}{2^{n/2}\Gamma(n/2)} 2^{n/2}2^{1/2}\Gamma\left(\frac{n}{2}+\frac{1}{2}\right)\left(x^2+n\right)^{-(n+1)/2}
\end{aligned}$$

$$
\begin{aligned}
&= (2\pi)^{-1/2} \frac{n^{n/2}}{\Gamma(n/2)} 2^{1/2} \Gamma\left(\frac{n}{2}+\frac{1}{2}\right) \left(n\left(1+\frac{1}{n}x^2\right)\right)^{-(n+1)/2} \\
&= (2\pi)^{-1/2} \frac{n^{n/2}}{\Gamma(n/2)} \left(\frac{1}{2}\right)^{-1/2} \Gamma\left(\frac{n}{2}+\frac{1}{2}\right) \\
&\quad \cdot n^{-n/2-1/2} \left(1+\frac{1}{n}x^2\right)^{-(n+1)/2} \\
&= (2\pi)^{-1/2} \frac{\Gamma\left(\frac{n}{2}+\frac{1}{2}\right)}{\Gamma(n/2)} \left(\frac{1}{2}\right)^{-1/2} n^{-1/2} \left(1+\frac{1}{n}x^2\right)^{-(n+1)/2} \\
&= \left(2\pi\frac{1}{2}n\right)^{-1/2} \frac{\Gamma\left(\frac{n}{2}+\frac{1}{2}\right)}{\Gamma(n/2)} \left(1+\frac{1}{n}x^2\right)^{-(n+1)/2} \\
&= n^{-1/2} \frac{\Gamma\left(\frac{n}{2}+\frac{1}{2}\right)}{\sqrt{\pi}\Gamma(n/2)} \left(1+\frac{1}{n}x^2\right)^{-(n+1)/2} \\
\boxed{\text{C}} \quad &= n^{-1/2} \frac{\Gamma\left(\frac{n}{2}+\frac{1}{2}\right)}{\Gamma(1/2)\Gamma(n/2)} \left(1+\frac{1}{n}x^2\right)^{-(n+1)/2} \\
\boxed{\text{D}} \quad &= n^{-1/2} \frac{1}{B(n/2,1/2)} \left(1+\frac{1}{n}x^2\right)^{-(n+1)/2} \\
&= f_X(x)
\end{aligned}
$$

where: in step A we have used the fact that c and c_2 do not depend on z; in step B we have used the fact that the integral of a density function over its support is equal to 1; in step C we have used the fact that $\sqrt{\pi} = \Gamma(1/2)$; in step D we have used the definition of Beta function. ■

Since X is a zero-mean normal random variable with variance $1/z$, conditional on $Z = z$, then we can also think of it as a ratio

$$X = \frac{Y}{\sqrt{Z}}$$

where Y has a standard normal distribution, Z has a Gamma distribution and Y and Z are independent.

50.1.3 Expected value

The expected value of a standard Student's t random variable X is well-defined only for $n > 1$ and it is equal to

$$\mathrm{E}[X] = 0$$

Proof. It follows from the fact that the density function is symmetric around 0:

$$
\begin{aligned}
\mathrm{E}[X] &= \int_{-\infty}^{\infty} x f_X(x)\, dx \\
&= \int_{-\infty}^{0} x f_X(x)\, dx + \int_{0}^{\infty} x f_X(x)\, dx \\
\boxed{\text{A}} \quad &= -\int_{\infty}^{0} (-t) f_X(-t)\, dt + \int_{0}^{\infty} x f_X(x)\, dx
\end{aligned}
$$

$$\begin{aligned}
& = \int_{\infty}^{0} t f_X(-t)\,dt + \int_{0}^{\infty} x f_X(x)\,dx \\
\boxed{\text{B}} \quad & = -\int_{0}^{\infty} t f_X(-t)\,dt + \int_{0}^{\infty} x f_X(x)\,dx \\
& = -\int_{0}^{\infty} x f_X(-x)\,dx + \int_{0}^{\infty} x f_X(x)\,dx \\
\boxed{\text{C}} \quad & = -\int_{0}^{\infty} x f_X(x)\,dx + \int_{0}^{\infty} x f_X(x)\,dx \\
& = 0
\end{aligned}$$

where: in step A we have performed a change of variable in the first integral ($t = -x$); in step B we have exchanged the bounds of integration; in step C we have used the fact that

$$f_X(-x) = f_X(x)$$

The above integrals are finite (and so the expected value is well-defined) only if $n > 1$, because

$$\begin{aligned}
& \int_{0}^{\infty} x f_X(x)\,dx \\
= & \lim_{u\to\infty} \int_{0}^{u} xc\left(1+\frac{x^2}{n}\right)^{-\frac{1}{2}(n+1)} dx \\
= & \; c \lim_{u\to\infty} \left[-\frac{n}{n-1}\left(1+\frac{x^2}{n}\right)^{-\frac{1}{2}(n-1)}\right]_0^u \\
= & -\frac{cn}{n-1}\left\{\lim_{u\to\infty}\left(1+\frac{u^2}{n}\right)^{-\frac{1}{2}(n-1)} - 1\right\}
\end{aligned}$$

and the above limit is finite only if $n > 1$. ■

50.1.4 Variance

The variance of a standard Student's t random variable X is well-defined only for $n > 2$ and it is equal to

$$\operatorname{Var}[X] = \frac{n}{n-2}$$

Proof. It can be derived thanks to the usual formula for computing the variance[6] and to the integral representation of the Beta function:

$$\begin{aligned}
\mathrm{E}\left[X^2\right] & = \int_{-\infty}^{\infty} x^2 f_X(x)\,dx \\
& = \int_{-\infty}^{0} x^2 f_X(x)\,dx + \int_{0}^{\infty} x^2 f_X(x)\,dx \\
\boxed{\text{A}} \quad & = -\int_{\infty}^{0} t^2 f_X(-t)\,dt + \int_{0}^{\infty} x^2 f_X(x)\,dx \\
\boxed{\text{B}} \quad & = \int_{0}^{\infty} t^2 f_X(-t)\,dt + \int_{0}^{\infty} x^2 f_X(x)\,dx
\end{aligned}$$

[6] $\operatorname{Var}[X] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2$. See p. 156.

$$
\begin{aligned}
\boxed{\text{C}} \quad &= \int_0^\infty t^2 f_X(t)\,dt + \int_0^\infty x^2 f_X(x)\,dx \\
&= 2\int_0^\infty x^2 f_X(x)\,dx \\
&= 2c\int_0^\infty x^2 \left(1+\frac{x^2}{n}\right)^{-\frac{1}{2}(n+1)} dx \\
\boxed{\text{D}} \quad &= 2c\int_0^\infty nt\,(1+t)^{-n/2-1/2}\,\frac{\sqrt{n}}{2}\frac{1}{\sqrt{t}}dt \\
&= cn^{3/2}\int_0^\infty t^{3/2-1}\,(1+t)^{-3/2-(n/2-1)}\,dt \\
\boxed{\text{E}} \quad &= cn^{3/2}B\left(\frac{3}{2},\frac{n}{2}-1\right) \\
\boxed{\text{F}} \quad &= \frac{1}{\sqrt{n}}\frac{1}{B\left(\frac{n}{2},\frac{1}{2}\right)}\,n^{3/2}B\left(\frac{1}{2}+1,\frac{n}{2}-1\right) \\
\boxed{\text{G}} \quad &= n\frac{\Gamma\left(\frac{n}{2}+\frac{1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{1}{2}\right)}\frac{\Gamma\left(\frac{1}{2}+1\right)\Gamma\left(\frac{n}{2}-1\right)}{\Gamma\left(\frac{n}{2}+\frac{1}{2}\right)} \\
&= n\frac{\Gamma\left(\frac{1}{2}+1\right)\Gamma\left(\frac{n}{2}-1\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{1}{2}\right)} \\
\boxed{\text{H}} \quad &= n\frac{\Gamma\left(\frac{1}{2}\right)\frac{1}{2}\Gamma\left(\frac{n}{2}\right)\frac{2}{n-2}}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{1}{2}\right)} \\
&= \frac{n}{n-2}
\end{aligned}
$$

where: in step A we have performed a change of variable in the first integral ($t=-x$); in step B we have exchanged the bounds of integration; in step C we have used the fact that

$$f_X(-t) = f_X(t)$$

in step D we have performed a change of variable ($t=\frac{x^2}{n}$); in step E we have used the integral representation of the Beta function; in step F we have used the definition of c; in step G we have used the definition of Beta function; in step H we have used the fact that

$$\Gamma(z) = \Gamma(z-1)(z-1)$$

Finally:

$$\mathrm{E}[X]^2 = 0$$

and:

$$\mathrm{Var}[X] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 = \frac{n}{n-2}$$

From the above derivation, it should be clear that the variance is well-defined only when $n > 2$. Otherwise, if $n \leq 2$, the above improper integrals do not converge (and the Beta function is not well-defined). ■

50.1.5 Higher moments

The k-th moment of a standard Student's t random variable X is well-defined only for $k < n$ and it is equal to

$$\mu_X(k) = \begin{cases} n^{k/2}\Gamma\left(\frac{k+1}{2}\right)\Gamma\left(\frac{n-k}{2}\right)/\left(\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{1}{2}\right)\right) & \text{if } k \text{ is even} \\ 0 & \text{if } k \text{ is odd} \end{cases}$$

Proof. Using the definition of moment:

$$\begin{aligned} \mu_X(k) &= \mathrm{E}\left[X^k\right] \\ &= \int_{-\infty}^{\infty} x^k f_X(x)\,dx \\ &= \int_{-\infty}^{0} x^k f_X(x)\,dx + \int_{0}^{\infty} x^k f_X(x)\,dx \\ \boxed{\text{A}} &= -\int_{\infty}^{0} (-t)^k f_X(-t)\,dt + \int_{0}^{\infty} x^k f_X(x)\,dx \\ \boxed{\text{B}} &= (-1)^k \int_{0}^{\infty} t^k f_X(-t)\,dt + \int_{0}^{\infty} x^k f_X(x)\,dx \\ \boxed{\text{C}} &= (-1)^k \int_{0}^{\infty} t^k f_X(t)\,dt + \int_{0}^{\infty} x^k f_X(x)\,dx \\ &= \left(1+(-1)^k\right)\int_{0}^{\infty} x^k f_X(x)\,dx \end{aligned}$$

where: in step A we have performed a change of variable in the first integral ($t = -x$); in step B we have exchanged the bounds of integration; in step C we have used the fact that

$$f_X(-t) = f_X(t)$$

Therefore, to compute the k-th moment and to verify whether it exists and is finite, we need to study the following integral:

$$\begin{aligned} &\int_{0}^{\infty} x^k f_X(x)\,dx \\ &= c\int_{0}^{\infty} x^k \left(1+\frac{x^2}{n}\right)^{-\frac{1}{2}(n+1)} dx \\ \boxed{\text{A}} &= c\int_{0}^{\infty} (nt)^{k/2}(1+t)^{-n/2-1/2}\frac{\sqrt{n}}{2}\frac{1}{\sqrt{t}}dt \\ &= c\frac{1}{2}n^{(k+1)/2}\int_{0}^{\infty} t^{k/2-1/2}(1+t)^{-n/2-1/2}\,dt \\ &= c\frac{1}{2}n^{(k+1)/2}\int_{0}^{\infty} t^{(k+1)/2-1}(1+t)^{-(k+1)/2-(n-k)/2}\,dt \\ \boxed{\text{B}} &= c\frac{1}{2}n^{(k+1)/2}B\left(\frac{k+1}{2},\frac{n-k}{2}\right) \\ \boxed{\text{C}} &= \frac{1}{2}\frac{1}{\sqrt{n}}\frac{1}{B\left(\frac{n}{2},\frac{1}{2}\right)}n^{(k+1)/2}B\left(\frac{k+1}{2},\frac{n-k}{2}\right) \\ \boxed{\text{D}} &= \frac{1}{2}n^{k/2}\left[\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{1}{2}\right)\Big/\Gamma\left(\frac{n+1}{2}\right)\right]^{-1} \end{aligned}$$

$$\cdot \Gamma\left(\frac{k+1}{2}\right)\Gamma\left(\frac{n-k}{2}\right)\Big/\Gamma\left(\frac{n+1}{2}\right)$$
$$= \frac{1}{2}n^{k/2}\frac{\Gamma\left(\frac{k+1}{2}\right)\Gamma\left(\frac{n-k}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{1}{2}\right)}$$

where: in step A we have performed a change of variable in the first integral ($t = \frac{x^2}{n}$); in step B we have used the integral representation of the Beta function; in step C we have used the definition of c; in step D we have used the definition of Beta function. From the above derivation, it should be clear that the k-th moment is well-defined only when $n > k$. Otherwise, if $n \leq k$, the above improper integrals do not converge (the integrals involve the Beta function, which is well-defined and converges only when its arguments are strictly positive - in this case only if $\frac{n-k}{2} > 0$). Therefore, the k-th moment of X is:

$$\begin{aligned}\mu_X(k) &= \left(1+(-1)^k\right)\int_0^\infty x^k f_X(x)\,dx \\ &= \left(1+(-1)^k\right)\frac{1}{2}n^{k/2}\frac{\Gamma\left(\frac{k+1}{2}\right)\Gamma\left(\frac{n-k}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{1}{2}\right)} \\ &= \begin{cases} n^{k/2}\Gamma\left(\frac{k+1}{2}\right)\Gamma\left(\frac{n-k}{2}\right)/\left(\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{1}{2}\right)\right) & \text{if } k \text{ is even} \\ 0 & \text{if } k \text{ is odd}\end{cases}\end{aligned}$$

■

50.1.6 Moment generating function

A standard Student's t random variable X does not possess a moment generating function.

Proof. When a random variable X possesses a moment generating function, then the k-th moment of X exists and is finite for any $k \in \mathbb{N}$. But we have proved above that the k-th moment of X exists only for $k < n$. Therefore, X can not have a moment generating function. ■

50.1.7 Characteristic function

There is no simple expression for the characteristic function of the standard Student's t distribution. It can be expressed in terms of a Modified Bessel function of the second kind (a solution of a certain differential equation, called modified Bessel's differential equation). The interested reader can consult Sutradhar[7] (1986).

50.1.8 Distribution function

There is no simple formula for the distribution function $F_X(x)$ of a standard Student's t random variable X, because the integral

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt$$

[7] Sutradhar, B. C. (1986) On the characteristic function of multivariate Student t-distribution, *Canadian Journal of Statistics*, 14, 329-337.

cannot be expressed in terms of elementary functions. Therefore, it is usually necessary to resort to computer algorithms to compute the values of $F_X(x)$. For example, the MATLAB command:

```
tcdf(x,n)
```

returns the value of the distribution function at the point `x` when the degrees of freedom parameter is equal to `n`.

50.2 The Student's t distribution in general

While in the previous section we restricted our attention to the Student's t distribution with zero mean and unit scale, we now deal with the general case.

50.2.1 Definition

The Student's t distribution is characterized as follows:

Definition 262 *Let X be an absolutely continuous random variable. Let its support be the whole set of real numbers:*

$$R_X = \mathbb{R}$$

*Let $\mu \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}_{++}$ and $n \in \mathbb{R}_{++}$. We say that X has a **Student's t distribution** with mean μ, scale σ^2 and n degrees of freedom if its probability density function is:*

$$f_X(x) = c\frac{1}{\sigma}\left(1 + \frac{(x-\mu)^2}{n\sigma^2}\right)^{-\frac{1}{2}(n+1)}$$

where c is a constant:

$$c = \frac{1}{\sqrt{n}}\frac{1}{B\left(\frac{n}{2}, \frac{1}{2}\right)}$$

and $B()$ is the Beta function.

We indicate that X has a t distribution with mean μ, scale σ^2 and n degrees of freedom by:

$$X \sim T\left(\mu, \sigma^2, n\right)$$

50.2.2 Relation to the standard Student's t distribution

A random variable X which has a t distribution with mean μ, scale σ^2 and n degrees of freedom is just a linear function of a standard Student's t random variable[8]:

Proposition 263 *If $X \sim T\left(\mu, \sigma^2, n\right)$, then:*

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

where Z is a random variable having a standard t distribution.

[8] See p. 407.

Proof. This can be easily proved using the formula for the density of a function[9] of an absolutely continuous variable:

$$\begin{aligned} f_X(x) &= f_Z\left(g^{-1}(x)\right)\frac{dg^{-1}(x)}{dx} \\ &= f_Z\left(\frac{x-\mu}{\sigma}\right)\frac{1}{\sigma} \\ &= c\frac{1}{\sigma}\left(1+\frac{(x-\mu)^2}{n\sigma^2}\right)^{-\frac{1}{2}(n+1)} \end{aligned}$$

■

Obviously, then, a standard t distribution is just a normal distribution with mean $\mu = 0$ and scale $\sigma^2 = 1$.

50.2.3 Expected value

The expected value of a Student's t random variable X is well-defined only for $n > 1$ and it is equal to

$$\mathrm{E}[X] = \mu$$

Proof. It is an immediate consequence of the fact that $X = \mu + \sigma Z$ (where Z has a standard t distribution) and the linearity of the expected value[10]:

$$\mathrm{E}[X] = \mathrm{E}[\mu + \sigma Z] = \mu + \sigma \mathrm{E}[Z] = \mu + \sigma \cdot 0 = \mu$$

As we have seen above, $\mathrm{E}[Z]$ is well-defined only for $n > 1$ and, as a consequence, also $\mathrm{E}[X]$ is well-defined only for $n > 1$. ■

50.2.4 Variance

The variance of a Student's t random variable X is well-defined only for $n > 2$ and it is equal to

$$\mathrm{Var}[X] = \frac{n}{n-2}\sigma^2$$

Proof. It can be derived using the formula for the variance of linear transformations[11] on $X = \mu + \sigma Z$ (where Z has a standard t distribution):

$$\mathrm{Var}[X] = \mathrm{Var}[\mu + \sigma Z] = \sigma^2 \mathrm{Var}[Z] = \sigma^2 \frac{n}{n-2}$$

As we have seen above, $\mathrm{Var}[Z]$ is well-defined only for $n > 2$ and, as a consequence, also $\mathrm{Var}[X]$ is well-defined only for $n > 2$. ■

50.2.5 Moment generating function

A Student's t random variable X does not possess a moment generating function.
Proof. It is a consequence of the fact that $X = \mu + \sigma Z$ (where Z has a standard t distribution) and of the fact that a standard Student's t random variable does not possess a moment generating function (see above). ■

[9] See p. 265. Note that $X = g(Z) = \mu + \sigma Z$ is a strictly increasing function of Z, since σ is strictly positive.
[10] See p. 134.
[11] See p. 158.

50.2.6 Characteristic function

There is no simple expression for the characteristic function of the Student's t distribution (see the comments above, for the standard case).

50.2.7 Distribution function

As for the standard t distribution (see above), there is no simple formula for the distribution function $F_X(x)$ of a Student's t random variable X and it is usually necessary to resort to computer algorithms to compute the values of $F_X(x)$. Most computer programs provide only routines for the computation of the standard t distribution function (denote it by $F_Z(z)$). In these cases we need to make a conversion, as follows:

$$\begin{aligned} F_X(x) &= \mathrm{P}(X \leq x) \\ &= \mathrm{P}(\mu + \sigma Z \leq x) \\ &= \mathrm{P}\left(Z \leq \frac{x-\mu}{\sigma}\right) \\ &= F_Z\left(\frac{x-\mu}{\sigma}\right) \end{aligned}$$

For example, the MATLAB command:

```
tcdf((x-mu)/sigma,n)
```

returns the value at the point `x` of the distribution function of a Student's t random variable with mean `mu`, scale `sigma` and `n` degrees of freedom.

50.3 More details

50.3.1 Convergence to the normal distribution

A Student's t distribution with mean μ, scale σ^2 and n degrees of freedom converges in distribution[12] to a normal distribution with mean μ and variance σ^2 when the number of degrees of freedom n becomes large (converges to infinity).

Proof. As explained before, if X_n has a t distribution, it can be written as:

$$X_n = \mu + \sigma \frac{Y}{\sqrt{\chi_n^2/n}}$$

where Y is a standard normal random variable, and χ_n^2 is a Chi-square random variable with n degrees of freedom, independent of Y. Moreover, as explained in the lecture entitled *Chi-square distribution* (p. 387), χ_n^2 can be written as a sum of squares of n independent standard normal random variables $Z_1, \ldots, Z_n$:

$$\chi_n^2 = \sum_{i=1}^{n} Z_i^2$$

When n tends to infinity, the ratio

$$\frac{\chi_n^2}{n} = \frac{1}{n} \sum_{i=1}^{n} Z_i^2$$

[12] See p. 527.

converges in probability to $\mathrm{E}\left(Z_i^2\right) = 1$, by the Law of Large Numbers[13]. As a consequence, by Slutski's theorem[14], X_n converges in distribution to

$$X = \mu + \sigma Y$$

which is a normal random variable with mean μ and variance σ^2. ■

50.3.2 Non-central t distribution

As discussed above, if Y has a standard normal distribution, Z has a Gamma distribution with parameters n and $h = 1$ and Y and Z are independent, then the random variable X defined as

$$X = \frac{Y}{\sqrt{Z}}$$

has a standard Student's t distribution with n degrees of freedom.

Given the same assumptions on Y and Z, define a random variable W as follows:

$$W = \frac{Y + c}{\sqrt{Z}}$$

where $c \in \mathbb{R}$ is a constant. W is said to have a **non-central standard Student's t distribution** with n **degrees of freedom** and **non-centrality parameter** c. We do not discuss the details of this distribution here, but be aware that this distribution is sometimes used in statistical theory (also in elementary problems) and that routines to compute its moments and its distribution function can be found in most statistical software packages.

50.4 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Let X_1 be a normal random variable with mean $\mu = 0$ and variance $\sigma^2 = 4$. Let X_2 be a Gamma random variable with parameters $n = 10$ and $h = 3$, independent of X_1. Find the distribution of the ratio

$$X = \frac{X_1}{\sqrt{X_2}}$$

Solution

We can write:

$$X = \frac{X_1}{\sqrt{X_2}} = \frac{2}{\sqrt{3}} \frac{Y}{\sqrt{Z}}$$

where $Y = X_1/2$ has a standard normal distribution and $Z = X_2/3$ has a Gamma distribution with parameters $n = 10$ and $h = 1$. Therefore, the ratio

$$\frac{Y}{\sqrt{Z}}$$

[13] See p. 535.
[14] See p. 557.

has a standard Student's t distribution with $n = 10$ degrees of freedom and X has a Student's t distribution with mean $\mu = 0$, scale $\sigma^2 = 4/3$ and $n = 10$ degrees of freedom.

Exercise 2

Let X_1 be a normal random variable with mean $\mu = 3$ and variance $\sigma^2 = 1$. Let X_2 be a Gamma random variable with parameters $n = 15$ and $h = 2$, independent of X_1. Find the distribution of the random variable

$$X = \sqrt{\frac{2}{X_2}}\left(X_1 - 3\right)$$

Solution

We can write:

$$X = \sqrt{\frac{2}{X_2}}\left(X_1 - 3\right) = \frac{Y}{\sqrt{Z}}$$

where $Y = X_1 - 3$ has a standard normal distribution and $Z = X_2/2$ has a Gamma distribution with parameters $n = 15$ and $h = 1$. Therefore, the ratio

$$\frac{Y}{\sqrt{Z}}$$

has a standard Stutent's t distribution with $n = 15$ degrees of freedom.

Exercise 3

Let X be a Student's t random variable with mean $\mu = 1$, scale $\sigma^2 = 4$ and $n = 6$ degrees of freedom. Compute:

$$\mathrm{P}\left(0 \leq X \leq 1\right)$$

Solution

First of all, we need to write the probability in terms of the distribution function of X:

$$\begin{aligned} & \mathrm{P}\left(0 \leq X \leq 1\right) \\ = \; & \mathrm{P}\left(X \leq 1\right) - \mathrm{P}\left(X < 0\right) \\ \boxed{\text{A}} \quad = \; & \mathrm{P}\left(X \leq 1\right) - \mathrm{P}\left(X \leq 0\right) \\ \boxed{\text{B}} \quad = \; & F_X\left(1\right) - F_X\left(0\right) \end{aligned}$$

where: in step $\boxed{\text{A}}$ we have used the fact that any specific value of X has probability zero; in step $\boxed{\text{B}}$ we have used the definition of distribution function. Then, we express the distribution function $F_X\left(x\right)$ in terms of the distribution function of a standard Student's t random variable Z with $n = 6$ degrees of freedom:

$$F_X\left(x\right) = F_Z\left(\frac{x-1}{2}\right)$$

so that:

$$\mathrm{P}\left(0 \leq X \leq 1\right) = F_X\left(1\right) - F_X\left(0\right) = F_Z\left(0\right) - F_Z\left(-\frac{1}{2}\right) = 0.1826$$

where the difference $F_Z\left(0\right) - F_Z\left(-1/2\right)$ can be computed with a computer algorithm, for example using the MATLAB command

```
tcdf(0,6)-tcdf(-1/2,6)
```

Chapter 51

F distribution

A random variable X has an F distribution if it can be written as a ratio:

$$X = \frac{Y_1/n_1}{Y_2/n_2}$$

between a Chi-square random variable[1] Y_1 with n_1 degrees of freedom and a Chi-square random variable Y_2, independent of Y_1, with n_2 degrees of freedom (where each of the two random variables has been divided by its degrees of freedom). The importance of the F distribution stems from the fact that ratios of this kind are encountered very often in statistics.

51.1 Definition

The F distribution is characterized as follows:

Definition 264 *Let X be an absolutely continuous random variable. Let its support be the set of positive real numbers:*

$$R_X = [0, \infty)$$

*Let $n_1, n_2 \in \mathbb{N}$. We say that X has an **F distribution** with n_1 and n_2 degrees of freedom if its probability density function[2] is*

$$f_X(x) = cx^{n_1/2-1}\left(1 + \frac{n_1}{n_2}x\right)^{-(n_1+n_2)/2}$$

where c is a constant:

$$c = \left(\frac{n_1}{n_2}\right)^{n_1/2} \frac{1}{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)}$$

and $B()$ is the Beta function[3].

[1] See p. 387.
[2] See p. 107.
[3] See p. 59.

51.2 Relation to the Gamma distribution

An F random variable can be written as a Gamma random variable[4] with parameters n_1 and h_1, where the parameter h_1 is equal to the reciprocal of another Gamma random variable, independent of the first one, with parameters n_2 and $h_2 = 1$:

Proposition 265 (Integral representation) *The probability density function of X can be written as*

$$f_X(x) = \int_0^\infty f_{X|Z=z}(x) f_Z(z) dz$$

where:

1. *$f_{X|Z=z}(x)$ is the probability density function of a Gamma random variable with parameters n_1 and $h_1 = \frac{1}{z}$:*

$$\begin{aligned} f_{X|Z=z}(x) &= \frac{(n_1/h_1)^{n_1/2}}{2^{n_1/2}\Gamma(n_1/2)} x^{n_1/2-1} \exp\left(-\frac{n_1}{h_1}\frac{1}{2}x\right) \\ &= \frac{(n_1 z)^{n_1/2}}{2^{n_1/2}\Gamma(n_1/2)} x^{n_1/2-1} \exp\left(-n_1 z\frac{1}{2}x\right) \end{aligned}$$

2. *$f_Z(z)$ is the probability density function of a Gamma random variable with parameters n_2 and $h_2 = 1$:*

$$f_Z(z) = \frac{(n_2)^{n_2/2}}{2^{n_2/2}\Gamma(n_2/2)} z^{n_2/2-1} \exp\left(-n_2\frac{1}{2}z\right)$$

Proof. We need to prove that

$$f_X(x) = \int_0^\infty f_{X|Z=z}(x) f_Z(z) dz$$

where

$$f_{X|Z=z}(x) = \frac{(n_1 z)^{n_1/2}}{2^{n_1/2}\Gamma(n_1/2)} x^{n_1/2-1} \exp\left(-n_1 z\frac{1}{2}x\right)$$

and

$$f_Z(z) = \frac{n_2^{n_2/2}}{2^{n_2/2}\Gamma(n_2/2)} z^{n_2/2-1} \exp\left(-n_2\frac{1}{2}z\right)$$

Let us start from the integrand function:

$$\begin{aligned} & f_{X|Z=z}(x) f_Z(z) \\ = & \frac{(n_1 z)^{n_1/2}}{2^{n_1/2}\Gamma(n_1/2)} x^{n_1/2-1} \exp\left(-n_1 z\frac{1}{2}x\right) \\ & \cdot \frac{n_2^{n_2/2}}{2^{n_2/2}\Gamma(n_2/2)} z^{n_2/2-1} \exp\left(-n_2\frac{1}{2}z\right) \end{aligned}$$

[4] See p. 397.

$$\begin{aligned}
&= \frac{(n_1 z)^{n_1/2}}{2^{n_1/2}\Gamma(n_1/2)} \frac{n_2^{n_2/2}}{2^{n_2/2}\Gamma(n_2/2)} \\
&\quad \cdot x^{n_1/2-1} z^{n_2/2-1} \exp\left(-(n_1 x + n_2)\frac{1}{2} z\right) \\
&= \frac{n_1^{n_1/2} n_2^{n_2/2}}{2^{(n_1+n_2)/2}\Gamma(n_1/2)\Gamma(n_2/2)} \\
&\quad \cdot x^{n_1/2-1} z^{(n_1+n_2)/2-1} \exp\left(-(n_1 x + n_2)\frac{1}{2} z\right) \\
&= \frac{n_1^{n_1/2} n_2^{n_2/2}}{2^{(n_1+n_2)/2}\Gamma(n_1/2)\Gamma(n_2/2)} x^{n_1/2-1} z^{(n_1+n_2)/2-1} \\
&\quad \cdot \exp\left(-\left(\frac{n_1+n_2}{n_1 x + n_2}\right)^{-1}(n_1+n_2)\frac{1}{2} z\right) \\
&= \frac{n_1^{n_1/2} n_2^{n_2/2}}{2^{(n_1+n_2)/2}\Gamma(n_1/2)\Gamma(n_2/2)} x^{n_1/2-1} \frac{1}{c} f_{Z|X=x}(z)
\end{aligned}$$

where

$$c = \frac{(n_1 x + n_2)^{(n_1+n_2)/2}}{2^{(n_1+n_2)/2}\Gamma((n_1+n_2)/2)}$$

and $f_{Z|X=x}(z)$ is the probability density function of a random variable having a Gamma distribution with parameters

$$n_1 + n_2$$

and

$$\frac{n_1+n_2}{n_1 x + n_2}$$

Therefore:

$$\begin{aligned}
&\int_0^\infty f_{X|Z=z}(x) f_Z(z)\, dz \\
&= \int_0^\infty \frac{n_1^{n_1/2} n_2^{n_2/2}}{2^{(n_1+n_2)/2}\Gamma(n_1/2)\Gamma(n_2/2)} x^{n_1/2-1} \frac{1}{c} f_{Z|X=x}(z)\, dz \\
\boxed{\text{A}} &= \frac{n_1^{n_1/2} n_2^{n_2/2}}{2^{(n_1+n_2)/2}\Gamma(n_1/2)\Gamma(n_2/2)} x^{n_1/2-1} \frac{1}{c} \int_0^\infty f_{Z|X=x}(z)\, dz \\
\boxed{\text{B}} &= \frac{n_1^{n_1/2} n_2^{n_2/2}}{2^{(n_1+n_2)/2}\Gamma(n_1/2)\Gamma(n_2/2)} x^{n_1/2-1} \frac{1}{c} \\
&= \frac{n_1^{n_1/2} n_2^{n_2/2}}{2^{(n_1+n_2)/2}\Gamma(n_1/2)\Gamma(n_2/2)} x^{n_1/2-1} \frac{2^{(n_1+n_2)/2}\Gamma((n_1+n_2)/2)}{(n_1 x + n_2)^{(n_1+n_2)/2}} \\
&= \frac{\Gamma((n_1+n_2)/2)}{\Gamma(n_1/2)\Gamma(n_2/2)} n_1^{n_1/2} n_2^{n_2/2} x^{n_1/2-1} (n_1 x + n_2)^{-(n_1+n_2)/2} \\
&= \left(\frac{\Gamma(n_1/2)\Gamma(n_2/2)}{\Gamma((n_1+n_2)/2)}\right)^{-1} n_1^{n_1/2} n_2^{n_2/2} x^{n_1/2-1} n_2^{-(n_1+n_2)/2} \\
&\quad \cdot \left(1 + \frac{n_1}{n_2} x\right)^{-(n_1+n_2)/2}
\end{aligned}$$

$$
\begin{aligned}
&= \left(\frac{\Gamma(n_1/2)\,\Gamma(n_2/2)}{\Gamma((n_1+n_2)/2)}\right)^{-1} n_1^{n_1/2} n_2^{-n_1/2} x^{n_1/2-1}\left(1+\frac{n_1}{n_2}x\right)^{-(n_1+n_2)/2} \\
&= \left(\frac{\Gamma(n_1/2)\,\Gamma(n_2/2)}{\Gamma((n_1+n_2)/2)}\right)^{-1} \left(\frac{n_1}{n_2}\right)^{n_1/2} x^{n_1/2-1}\left(1+\frac{n_1}{n_2}x\right)^{-(n_1+n_2)/2} \\
\boxed{\text{C}} \quad &= \frac{1}{B(n_1/2, n_2/2)} \left(\frac{n_1}{n_2}\right)^{n_1/2} x^{n_1/2-1}\left(1+\frac{n_1}{n_2}x\right)^{-(n_1+n_2)/2} \\
&= f_X(x)
\end{aligned}
$$

where: in step A we have used the fact that c does not depend on z; in step B we have used the fact that the integral of a density function over its support is equal to 1; in step C we have used the definition of Beta function. ■

51.3 Relation to the Chi-square distribution

In the introduction, we have stated (without a proof) that a random variable X has an F distribution with n_1 and n_2 degrees of freedom if it can be written as a ratio:

$$X = \frac{Y_1/n_1}{Y_2/n_2}$$

where:

1. Y_1 is a Chi-square random variable with n_1 degrees of freedom;
2. Y_2 is a Chi-square random variable, independent of Y_1, with n_2 degrees of freedom.

The statement can be proved as follows.

Proof. It is a consequence of Proposition 265 above: X can be thought of as a Gamma random variable with parameters n_1 and h_1, where the parameter h_1 is equal to the reciprocal of another Gamma random variable Z, independent of the first one, with parameters n_2 and $h_2 = 1$. The equivalence can be proved as follows. Since a Gamma random variable with parameters n_1 and h_1 is just the product between the ratio h_1/n_1 and a Chi-square random variable with n_1 degrees of freedom (see the lecture entitled *Gamma distribution* - p. 397), we can write:

$$X = \frac{h_1}{n_1} Y_1$$

where Y_1 is a Chi-square random variable with n_1 degrees of freedom. Now, we know that h_1 is equal to the reciprocal of another Gamma random variable Z, independent of Y_1, with parameters n_2 and $h_2 = 1$. Therefore:

$$X = \frac{Y_1/n_1}{Z}$$

But a Gamma random variable with parameters n_2 and $h_2 = 1$ is just the product between the ratio $1/n_2$ and a Chi-square random variable with n_2 degrees of freedom (see the lecture entitled *Gamma distribution* - p. 397). Therefore, we can write

$$X = \frac{Y_1/n_1}{Y_2/n_2}$$

■

51.4 Expected value

The expected value of an F random variable X is well-defined only for $n_2 > 2$ and it is equal to

$$\mathrm{E}[X] = \frac{n_2}{n_2 - 2}$$

Proof. It can be derived thanks to the integral representation of the Beta function:

$$\begin{aligned}
\mathrm{E}[X] &= \int_{-\infty}^{\infty} x f_X(x)\, dx \\
&= \int_0^{\infty} x c x^{n_1/2-1} \left(1 + \frac{n_1}{n_2} x\right)^{-(n_1+n_2)/2} dx \\
&= c \int_0^{\infty} x^{n_1/2} \left(1 + \frac{n_1}{n_2} x\right)^{-(n_1+n_2)/2} dx \\
\boxed{\text{A}} &= c \int_0^{\infty} \left(\frac{n_2}{n_1} t\right)^{n_1/2} (1+t)^{-(n_1+n_2)/2} \frac{n_2}{n_1} dt \\
&= c \left(\frac{n_2}{n_1}\right)^{n_1/2+1} \int_0^{\infty} t^{n_1/2} (1+t)^{-n_1/2-n_2/2}\, dt \\
&= c \left(\frac{n_2}{n_1}\right)^{n_1/2+1} \int_0^{\infty} t^{(n_1/2+1)-1} (1+t)^{-(n_1/2+1)-(n_2/2-1)}\, dt \\
\boxed{\text{B}} &= c \left(\frac{n_2}{n_1}\right)^{n_1/2+1} B\left(\frac{n_1}{2} + 1, \frac{n_2}{2} - 1\right) \\
\boxed{\text{C}} &= \left(\frac{n_1}{n_2}\right)^{n_1/2} \frac{1}{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \left(\frac{n_2}{n_1}\right)^{n_1/2+1} B\left(\frac{n_1}{2} + 1, \frac{n_2}{2} - 1\right) \\
&= \frac{n_2}{n_1} \frac{1}{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} B\left(\frac{n_1}{2} + 1, \frac{n_2}{2} - 1\right) \\
\boxed{\text{D}} &= \frac{n_2}{n_1} \frac{\Gamma(n_1/2 + n_2/2)}{\Gamma(n_1/2)\,\Gamma(n_2/2)} \frac{\Gamma(n_1/2+1)\,\Gamma(n_2/2-1)}{\Gamma(n_1/2+1+n_2/2-1)} \\
&= \frac{n_2}{n_1} \frac{\Gamma(n_1/2 + n_2/2)}{\Gamma(n_1/2)\,\Gamma(n_2/2)} \frac{\Gamma(n_1/2+1)\,\Gamma(n_2/2-1)}{\Gamma(n_1/2+n_2/2)} \\
&= \frac{n_2}{n_1} \frac{\Gamma(n_1/2+1)}{\Gamma(n_1/2)} \frac{\Gamma(n_2/2-1)}{\Gamma(n_2/2)} \\
\boxed{\text{E}} &= \frac{n_2}{n_1} (n_1/2) \frac{1}{n_2/2 - 1} \\
&= \frac{n_2}{n_2 - 2}
\end{aligned}$$

where: in step A we have performed a change of variable ($t = \frac{n_1}{n_2} x$); in step B we have used the integral representation of the Beta function; in step C we have used the definition of c; in step D we have used the definition of Beta function; in step E we have used the following property of the Gamma function:

$$\Gamma(z) = \Gamma(z-1)(z-1)$$

It is also clear that the expected value is well-defined only when $n_2 > 2$: when $n_2 \leq 2$, the above improper integrals do not converge (both arguments of the Beta

function must be strictly positive). ■

51.5 Variance

The variance of an F random variable X is well-defined only for $n_2 > 4$ and it is equal to

$$\operatorname{Var}[X] = \frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}$$

Proof. It can be derived thanks to the usual formula for computing the variance[5] and to the integral representation of the Beta function:

$$\begin{aligned}
\mathrm{E}\left[X^2\right] &= \int_{-\infty}^{\infty} x^2 f_X(x)\,dx \\
&= \int_0^{\infty} x^2 c x^{n_1/2-1}\left(1+\frac{n_1}{n_2}x\right)^{-(n_1+n_2)/2} dx \\
&= c\int_0^{\infty} x^{n_1/2+1}\left(1+\frac{n_1}{n_2}x\right)^{-(n_1+n_2)/2} dx \\
\boxed{\text{A}} &= c\int_0^{\infty}\left(\frac{n_2}{n_1}t\right)^{n_1/2+1}(1+t)^{-(n_1+n_2)/2}\frac{n_2}{n_1}dt \\
&= c\left(\frac{n_2}{n_1}\right)^{n_1/2+2}\int_0^{\infty} t^{n_1/2+1}(1+t)^{-n_1/2-n_2/2}\,dt \\
&= c\left(\frac{n_2}{n_1}\right)^{n_1/2+2}\int_0^{\infty} t^{(n_1/2+2)-1}(1+t)^{-(n_1/2+2)-(n_2/2-2)}\,dt \\
\boxed{\text{B}} &= c\left(\frac{n_2}{n_1}\right)^{n_1/2+2} B\left(\frac{n_1}{2}+2,\frac{n_2}{2}-2\right) \\
\boxed{\text{C}} &= \left(\frac{n_1}{n_2}\right)^{n_1/2}\frac{1}{B\left(\frac{n_1}{2},\frac{n_2}{2}\right)}\left(\frac{n_2}{n_1}\right)^{n_1/2+2} B\left(\frac{n_1}{2}+2,\frac{n_2}{2}-2\right) \\
\boxed{\text{D}} &= \left(\frac{n_2}{n_1}\right)^{2}\frac{1}{B\left(\frac{n_1}{2},\frac{n_2}{2}\right)} B\left(\frac{n_1}{2}+2,\frac{n_2}{2}-2\right) \\
&= \left(\frac{n_2}{n_1}\right)^{2}\frac{\Gamma(n_1/2+n_2/2)}{\Gamma(n_1/2)\Gamma(n_2/2)}\frac{\Gamma(n_1/2+2)\Gamma(n_2/2-2)}{\Gamma(n_1/2+2+n_2/2-2)} \\
&= \left(\frac{n_2}{n_1}\right)^{2}\frac{\Gamma(n_1/2+n_2/2)}{\Gamma(n_1/2)\Gamma(n_2/2)}\frac{\Gamma(n_1/2+2)\Gamma(n_2/2-2)}{\Gamma(n_1/2+n_2/2)} \\
&= \left(\frac{n_2}{n_1}\right)^{2}\frac{\Gamma(n_1/2+2)}{\Gamma(n_1/2)}\frac{\Gamma(n_2/2-2)}{\Gamma(n_2/2)} \\
\boxed{\text{E}} &= \left(\frac{n_2}{n_1}\right)^{2}(n_1/2+1)(n_1/2)\frac{1}{(n_2/2-1)(n_2/2-2)} \\
&= \frac{n_2^2(n_1+2)n_1}{n_1^2(n_2-2)(n_2-4)} \\
&= \frac{n_2^2(n_1+2)}{n_1(n_2-2)(n_2-4)}
\end{aligned}$$

[5] $\operatorname{Var}[X] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2$. See p. 156.

where: in step A we have performed a change of variable ($t = \frac{n_1}{n_2}x$); in step B we have used the integral representation of the Beta function; in step C we have used the definition of c; in step D we have used the definition of Beta function; in step E we have used the following property of the Gamma function:

$$\Gamma(z) = \Gamma(z-1)(z-1)$$

Finally:

$$\mathrm{E}[X]^2 = \left(\frac{n_2}{n_2-2}\right)^2$$

and:

$$\begin{aligned}
\mathrm{Var}[X] &= \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 \\
&= \frac{n_2^2(n_1+2)}{n_1(n_2-2)(n_2-4)} - \frac{n_2^2}{(n_2-2)^2} \\
&= \frac{n_2^2((n_1+2)(n_2-2) - n_1(n_2-4))}{n_1(n_2-2)^2(n_2-4)} \\
&= \frac{n_2^2(n_1n_2 - 2n_1 + 2n_2 - 4 - n_1n_2 + 4n_1)}{n_1(n_2-2)^2(n_2-4)} \\
&= \frac{n_2^2(2n_1+2n_2-4)}{n_1(n_2-2)^2(n_2-4)} = \frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}
\end{aligned}$$

It is also clear that the expected value is well-defined only when $n_2 > 4$: when $n_2 \leq 4$, the above improper integrals do not converge (both arguments of the Beta function must be strictly positive). ■

51.6 Higher moments

The k-th moment of an F random variable X is well-defined only for $n_2 > 2k$ and it is equal to

$$\mu_X(k) = \left(\frac{n_2}{n_1}\right)^k \frac{\Gamma(n_1/2+k)}{\Gamma(n_1/2)} \frac{\Gamma(n_2/2-k)}{\Gamma(n_2/2)}$$

Proof. Using the definition of moment:

$$\begin{aligned}
\mu_X(k) &= \mathrm{E}\left[X^k\right] = \int_{-\infty}^{\infty} x^k f_X(x)\,dx \\
&= \int_0^{\infty} x^k c x^{n_1/2-1}\left(1+\frac{n_1}{n_2}x\right)^{-(n_1+n_2)/2} dx \\
&= c\int_0^{\infty} x^{n_1/2+k-1}\left(1+\frac{n_1}{n_2}x\right)^{-(n_1+n_2)/2} dx \\
\boxed{\text{A}} &= c\int_0^{\infty} \left(\frac{n_2}{n_1}t\right)^{n_1/2+k-1}(1+t)^{-(n_1+n_2)/2}\frac{n_2}{n_1}dt \\
&= c\left(\frac{n_2}{n_1}\right)^{n_1/2+k}\int_0^{\infty} t^{n_1/2+k-1}(1+t)^{-n_1/2-n_2/2}\,dt
\end{aligned}$$

$$
\begin{aligned}
&= c\left(\frac{n_2}{n_1}\right)^{n_1/2+k}\int_0^\infty t^{(n_1/2+k)-1}\left(1+t\right)^{-(n_1/2+k)-(n_2/2-k)}dt \\
\boxed{\text{B}} \quad &= c\left(\frac{n_2}{n_1}\right)^{n_1/2+k} B\left(\frac{n_1}{2}+k,\frac{n_2}{2}-k\right) \\
\boxed{\text{C}} \quad &= \left(\frac{n_1}{n_2}\right)^{n_1/2}\frac{1}{B\left(\frac{n_1}{2},\frac{n_2}{2}\right)}\left(\frac{n_2}{n_1}\right)^{n_1/2+k} B\left(\frac{n_1}{2}+k,\frac{n_2}{2}-k\right) \\
&= \left(\frac{n_2}{n_1}\right)^{k}\frac{1}{B\left(\frac{n_1}{2},\frac{n_2}{2}\right)} B\left(\frac{n_1}{2}+k,\frac{n_2}{2}-k\right) \\
\boxed{\text{D}} \quad &= \left(\frac{n_2}{n_1}\right)^{k}\frac{\Gamma\left(n_1/2+n_2/2\right)}{\Gamma\left(n_1/2\right)\Gamma\left(n_2/2\right)}\frac{\Gamma\left(n_1/2+k\right)\Gamma\left(n_2/2-k\right)}{\Gamma\left(n_1/2+k+n_2/2-k\right)} \\
&= \left(\frac{n_2}{n_1}\right)^{k}\frac{\Gamma\left(n_1/2+n_2/2\right)}{\Gamma\left(n_1/2\right)\Gamma\left(n_2/2\right)}\frac{\Gamma\left(n_1/2+k\right)\Gamma\left(n_2/2-k\right)}{\Gamma\left(n_1/2+n_2/2\right)} \\
&= \left(\frac{n_2}{n_1}\right)^{k}\frac{\Gamma\left(n_1/2+k\right)}{\Gamma\left(n_1/2\right)}\frac{\Gamma\left(n_2/2-k\right)}{\Gamma\left(n_2/2\right)}
\end{aligned}
$$

where: in step $\boxed{\text{A}}$ we have performed a change of variable ($t = \frac{n_1}{n_2}x$); in step $\boxed{\text{B}}$ we have used the integral representation of the Beta function; in step $\boxed{\text{C}}$ we have used the definition of c; in step $\boxed{\text{D}}$ we have used the definition of Beta function. It is also clear that the expected value is well-defined only when $n_2 > 2k$: when $n_2 \leq 2k$, the above improper integrals do not converge (both arguments of the Beta function must be strictly positive). ■

51.7 Moment generating function

An F random variable X does not possess a moment generating function.

Proof. When a random variable X possesses a moment generating function, then the k-th moment of X exists and is finite for any $k \in \mathbb{N}$. But we have proved above that the k-th moment of X exists only for $k < n_2/2$. Therefore, X can not have a moment generating function. ■

51.8 Characteristic function

There is no simple expression for the characteristic function of the F distribution. It can be expressed in terms of the confluent hypergeometric function of the second kind (a solution of a certain differential equation, called confluent hypergeometric differential equation). The interested reader can consult Phillips[6] (1982).

51.9 Distribution function

The distribution function of an F random variable is:

$$F_X\left(x\right) = \frac{1}{B\left(\frac{n_1}{2},\frac{n_2}{2}\right)}\int_{-\infty}^{n_1 x/n_2} s^{n_1/2-1}\left(1+s\right)^{-n_1/2-n_2/2}ds$$

[6] Phillips, P. C. B. (1982) "The true characteristic function of the F distribution", Biometrika, 69, 261-264.

where the integral

$$\int_{-\infty}^{n_1 x/n_2} s^{n_1/2-1}\left(1+s\right)^{-n_1/2-n_2/2} ds$$

is known as incomplete Beta function and is usually computed numerically with the help of a computer algorithm.

Proof. This is proved as follows:

$$\begin{aligned}
& F_X(x) \\
= & \int_{-\infty}^{x} f_X(t)\, dt \\
= & \int_{-\infty}^{x} c t^{n_1/2-1}\left(1+\frac{n_1}{n_2}t\right)^{-(n_1+n_2)/2} dt \\
\boxed{\text{A}} = & c\int_{-\infty}^{n_1 x/n_2}\left(\frac{n_2}{n_1}s\right)^{n_1/2-1}\left(1+s\right)^{-n_1/2-n_2/2}\frac{n_2}{n_1} ds \\
= & c\left(\frac{n_2}{n_1}\right)^{n_1/2}\int_{-\infty}^{n_1 x/n_2} s^{n_1/2-1}\left(1+s\right)^{-n_1/2-n_2/2} ds \\
\boxed{\text{B}} = & \frac{\left(n_1/n_2\right)^{n_1/2}}{B\left(\frac{n_1}{2},\frac{n_2}{2}\right)}\left(\frac{n_2}{n_1}\right)^{n_1/2}\int_{-\infty}^{n_1 x/n_2} s^{n_1/2-1}\left(1+s\right)^{-n_1/2-n_2/2} ds \\
= & \frac{1}{B\left(\frac{n_1}{2},\frac{n_2}{2}\right)}\int_{-\infty}^{n_1 x/n_2} s^{n_1/2-1}\left(1+s\right)^{-n_1/2-n_2/2} ds
\end{aligned}$$

where: in step $\boxed{\text{A}}$ we have performed a change of variable ($s=\frac{n_1}{n_2}t$); in step $\boxed{\text{B}}$ we have used the definition of c. ■

51.10 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Let X_1 be a Gamma random variable with parameters $n_1=3$ and $h_1=2$. Let X_2 be another Gamma random variable, independent of X_1, with parameters $n_2=5$ and $h_1=6$. Find the expected value of the ratio:

$$\frac{X_1}{X_2}$$

Solution

We can write:

$$\begin{aligned}
X_1 &= 2Z_1 \\
X_2 &= 6Z_2
\end{aligned}$$

where Z_1 and Z_2 are two independent Gamma random variables, the parameters of Z_1 are $\overline{n}_1=3$ and $\overline{h}_1=1$ and the parameters of Z_2 are $\overline{n}_2=5$ and $\overline{h}_2=1$

(see the lecture entitled *Gamma distribution* - p. 397). Using this fact, the ratio becomes:

$$\frac{X_1}{X_2} = \frac{2}{6}\frac{Z_1}{Z_2} = \frac{1}{3}\frac{Z_1}{Z_2}$$

where Z_1/Z_2 has an F distribution with parameters $n_1 = 3$ and $n_2 = 5$. Therefore:

$$\begin{aligned} \mathrm{E}\left[\frac{X_1}{X_2}\right] &= \mathrm{E}\left[\frac{1}{3}\frac{Z_1}{Z_2}\right] = \frac{1}{3}\mathrm{E}\left[\frac{Z_1}{Z_2}\right] \\ &= \frac{1}{3}\frac{n_2}{n_2-2} = \frac{1}{3}\frac{5}{5-2} = \frac{5}{9} \end{aligned}$$

Exercise 2

Find the third moment of an F random variable with parameters $n_1 = 6$ and $n_2 = 18$.

Solution

We need to use the formula for the k-th moment of an F random variable:

$$\mu_X(k) = \left(\frac{n_2}{n_1}\right)^k \frac{\Gamma(n_1/2+k)}{\Gamma(n_1/2)}\frac{\Gamma(n_2/2-k)}{\Gamma(n_2/2)}$$

Plugging in the parameter values, we obtain:

$$\begin{aligned} \mu_X(3) &= \left(\frac{18}{6}\right)^3 \frac{\Gamma(3+3)}{\Gamma(3)}\frac{\Gamma(9-3)}{\Gamma(9)} = 3^3\frac{\Gamma(6)}{\Gamma(3)}\frac{\Gamma(6)}{\Gamma(9)} \\ &= 27\cdot\frac{5!}{2!}\cdot\frac{5!}{8!} = 27\cdot(5\cdot 4\cdot 3)\cdot\frac{1}{8\cdot 7\cdot 6} \\ &= 27\cdot 5\cdot\frac{1}{2\cdot 7\cdot 2} = \frac{135}{28} \end{aligned}$$

where we have used the relation between the Gamma function[7] and the factorial function.

[7] See p. 55.

Chapter 52

Multinomial distribution

The multinomial distribution is a generalization of the binomial distribution[1].

If you perform n times an experiment that can have only two outcomes (either success or failure), then the number of times you obtain one of the two outcomes (success) is a binomial random variable.

If you perform n times an experiment that can have K outcomes (K can be any natural number) and you denote by X_i the number of times that you obtain the i-th outcome, then the random vector

$$X = \begin{bmatrix} X_1 & X_2 & \dots & X_K \end{bmatrix}^\top$$

is a multinomial random vector.

In this lecture we will first present the special case in which there is only one experiment ($n = 1$), and we will then employ the results obtained for this simple special case to discuss the more general case of many experiments ($n \geq 1$).

52.1 The special case of one experiment

In this case, one experiment is performed, having K possible outcomes with probabilities $p_1, \dots, p_K$. When the i-th outcome is obtained, the i-th entry of the multinomial random vector X takes value 1, while all other entries take value 0.

52.1.1 Definition

The distribution is characterized as follows.

Definition 266 *Let X be a $K \times 1$ discrete random vector. Let the support of X be the set of $K \times 1$ vectors having one entry equal to 1 and all other entries equal to 0:*

$$R_X = \left\{ x \in \{0,1\}^K : \sum_{j=1}^{K} x_j = 1 \right\}$$

Let $p_1, \dots, p_K$ be K strictly positive numbers such that

$$\sum_{j=1}^{K} p_j = 1$$

[1] See p. 341.

We say that X has a ***multinomial distribution*** *with probabilities $p_1, \ldots, p_K$ and number of trials $n = 1$ if its joint probability mass function[2] is*

$$p_X(x_1, \ldots, x_K) = \begin{cases} \prod_{j=1}^{K} p_j^{x_j} & \text{if } (x_1, \ldots, x_K) \in R_X \\ 0 & \text{otherwise} \end{cases}$$

If you are puzzled by the above definition of the joint pmf, note that when $(x_1, \ldots, x_K) \in R_X$ and x_i is equal to 1, because the i-th outcome has been obtained, then all other entries are equal to 0 and

$$\begin{aligned} \prod_{j=1}^{K} p_j^{x_j} &= p_1^{x_1} \cdot \ldots \cdot p_{i-1}^{x_{i-1}} \cdot p_i^{x_i} \cdot p_{i+1}^{x_{i+1}} \cdot \ldots \cdot p_K^{x_K} \\ &= p_1^0 \cdot \ldots \cdot p_{i-1}^0 \cdot p_i^1 \cdot p_{i+1}^0 \cdot \ldots \cdot p_K^0 \\ &= 1 \cdot \ldots \cdot 1 \cdot p_i^1 \cdot 1 \cdot \ldots \cdot 1 = p_i \end{aligned}$$

52.1.2 Expected value

The expected value of X is

$$\mathrm{E}[X] = p \tag{52.1}$$

where the $K \times 1$ vector p is defined as follows:

$$p = [p_1 \;\; p_2 \;\; \ldots \;\; p_K]^\intercal$$

Proof. The i-th entry of X, denoted by X_i, is an indicator function[3] of the event "the i-th outcome has happened". Therefore, its expected value is equal to the probability of the event it indicates:

$$\mathrm{E}[X_i] = p_i$$

■

52.1.3 Covariance matrix

The covariance matrix of X is

$$\mathrm{Var}[X] = \Sigma$$

where Σ is a $K \times K$ matrix whose generic entry is

$$\Sigma_{ij} = \begin{cases} p_i(1 - p_i) & \text{if } j = i \\ -p_i p_j & \text{if } j \neq i \end{cases} \tag{52.2}$$

Proof. We need to use the formula[4]

$$\Sigma_{ij} = \mathrm{Cov}[X_i, X_j] = \mathrm{E}[X_i X_j] - \mathrm{E}[X_i]\,\mathrm{E}[X_j]$$

If $j = i$, then

$$\Sigma_{ii} = \mathrm{E}[X_i X_i] - \mathrm{E}[X_i]\,\mathrm{E}[X_i] = \mathrm{E}[X_i] - \mathrm{E}[X_i]^2$$

[2] See p. 116.
[3] See p. 197.
[4] See the lecture entitled *Covariance matrix* - p. 189.

$$= p_i - p_i^2 = p_i (1 - p_i)$$

where we have used the fact that $X_i^2 = X_i$ because X_i can take only values 0 and 1. If $j \neq i$, then

$$\begin{aligned} \Sigma_{ij} &= \mathrm{E}[X_i X_j] - \mathrm{E}[X_i]\,\mathrm{E}[X_j] \\ &= -\mathrm{E}[X_i]\,\mathrm{E}[X_j] = -p_i p_j \end{aligned}$$

where we have used the fact that $X_i X_j = 0$, because X_i and X_j cannot be both equal to 1 at the same time. ■

52.1.4 Joint moment generating function

The joint moment generating function[5] of X is defined for any $t \in \mathbb{R}^K$:

$$M_X(t) = \sum_{j=1}^{K} p_j \exp(t_j) \tag{52.3}$$

Proof. If the j-th outcome is obtained, then $X_i = 0$ for $i \neq j$ and $X_i = 1$ for $i = j$. As a consequence

$$\exp(t_1 X_1 + \ldots + t_K X_K) = \exp(t_j)$$

and the joint moment generating function is

$$\begin{aligned} M_X(t) &= \mathrm{E}\left[\exp\left(t^\top X\right)\right] = \mathrm{E}\left[\exp\left(t_1 X_1 + t_2 X_2 + \ldots + t_K X_K\right)\right] \\ &= \sum_{j=1}^{K} p_j \exp(t_j) \end{aligned}$$

■

52.1.5 Joint characteristic function

The joint characteristic function[6] of X is

$$\varphi_X(t) = \sum_{j=1}^{K} p_j \exp(it_j) \tag{52.4}$$

Proof. If the j-th outcome is obtained, then $X_i = 0$ for $i \neq j$ and $X_i = 1$ for $i = j$. As a consequence

$$\exp(it_1 X_1 + it_2 X_2 + \ldots + it_K X_K) = \exp(it_j)$$

and the joint characteristic function is

$$\begin{aligned} \varphi_X(t) &= \mathrm{E}\left[\exp\left(it^\top X\right)\right] = \mathrm{E}\left[\exp\left(it_1 X_1 + it_2 X_2 + \ldots + it_K X_K\right)\right] \\ &= \sum_{j=1}^{K} p_j \exp(it_j) \end{aligned}$$

■

[5] See p. 297.
[6] See p. 315.

52.2 Multinomial distribution in general

We now deal with the general case, in which the number of experiments can take any value $n \geq 1$.

52.2.1 Definition

Multinomial random vectors are characterized as follows.

Definition 267 *Let X be a $K \times 1$ discrete random vector. Let $n \in \mathbb{N}$. Let the support of X be the set of $K \times 1$ vectors having non-negative integer entries summing up to n:*

$$R_X = \left\{ x \in \{0, 1, 2, \ldots, n\}^K : \sum_{i=1}^{K} x_i = n \right\}$$

Let $p_1, \ldots, p_K$ be K strictly positive numbers such that

$$\sum_{i=1}^{K} p_i = 1$$

*We say that X has a **multinomial distribution** with probabilities $p_1, \ldots, p_K$ and number of trials n, if its joint probability mass function is*

$$p_X(x_1, \ldots, x_K) = \begin{cases} \binom{n}{x_1, x_2, \ldots, x_K} \prod_{i=1}^{K} p_i^{x_i} & \text{if } (x_1, \ldots, x_K) \in R_X \\ 0 & \text{otherwise} \end{cases}$$

where $\binom{n}{x_1, x_2, \ldots, x_K}$ is the multinomial coefficient[7].

52.2.2 Representation as a sum of simpler multinomials

The connection between the general case ($n \geq 1$) and the simpler case illustrated above ($n = 1$) is given by the following proposition.

Proposition 268 *A random vector X having a multinomial distribution with parameters $p_1, \ldots, p_K$ and n can be written as*

$$X = Y_1 + \ldots + Y_n$$

where $Y_1, \ldots, Y_n$ are n independent random vectors all having a multinomial distribution with parameters $p_1, \ldots, p_K$ and 1.

Proof. The sum $Y_1 + \ldots + Y_n$ is equal to the vector $[x_1, \ldots, x_K]^\top$ when: 1) x_1 terms of the sum have their first entry equal to 1 and all other entries equal to 0; 2) x_2 terms of the sum have their second entry equal to 1 and all other entries equal to 0; ... ; K) x_K terms of the sum have their K-th entry equal to 1 and all other entries equal to 0. Provided $x_i \geq 0$ for each i and $\sum_{i=1}^{K} x_i = n$, there are several different realizations of the matrix

$$[Y_1 \quad \ldots \quad Y_n]$$

[7] See p. 29.

satisfying these conditions. Since $Y_1, \ldots, Y_n$ are independent, each of these realizations has probability

$$p_1^{x_1} \cdot \ldots \cdot p_K^{x_K}$$

Furthermore, their number is equal to the number of partitions of n objects into K groups[8] having numerosities $x_1, \ldots, x_K$, which in turn is equal to the multinomial coefficient

$$\binom{n}{x_1, x_2, \ldots, x_K}$$

Therefore

$$\begin{aligned} & \mathrm{P}\left(Y_1 + \ldots + Y_n = \begin{bmatrix} x_1 & \ldots & x_K \end{bmatrix}^\top\right) \\ = & \binom{n}{x_1, x_2, \ldots, x_K} p_1^{x_1} \cdot \ldots \cdot p_K^{x_K} = p_X\left(x_1, \ldots, x_K\right) \end{aligned}$$

which proves that X and $Y_1 + \ldots + Y_n$ have the same distribution. ■

52.2.3 Expected value

The expected value of a multinomial random vector X is

$$\mathrm{E}\left[X\right] = np$$

where the $K \times 1$ vector p is defined as follows:

$$p = \begin{bmatrix} p_1 & p_2 & \ldots & p_K \end{bmatrix}^\intercal$$

Proof. Using the fact that X can be written as a sum of n multinomials with parameters $p_1, \ldots, p_K$ and 1, we obtain

$$\begin{aligned} \mathrm{E}\left[X\right] &= \mathrm{E}\left[Y_1 + \ldots + Y_n\right] = \mathrm{E}\left[Y_1\right] + \ldots + \mathrm{E}\left[Y_n\right] \\ &= p + \ldots + p = np \end{aligned}$$

where the result $\mathrm{E}\left[Y_j\right] = p$ has been derived in the previous section (formula 52.1). ■

52.2.4 Covariance matrix

The covariance matrix of a multinomial random vector X is

$$\mathrm{Var}\left[X\right] = n\Sigma$$

where Σ is a $K \times K$ matrix whose generic entry is

$$\Sigma_{ij} = \begin{cases} p_i\left(1 - p_i\right) & \text{if } j = i \\ -p_i p_j & \text{if } j \neq i \end{cases}$$

Proof. Since X can be represented as a sum of n independent multinomials with parameters $p_1, \ldots, p_K$ and 1, we obtain

$$\mathrm{Var}\left[X\right]$$

[8] See the lecture entitled *Partitions* - p. 27.

$$\begin{aligned} & = \operatorname{Var}\left[Y_1 + \ldots + Y_n\right] \\ \boxed{\text{A}} & = \operatorname{Var}\left[Y_1\right] + \ldots + \operatorname{Var}\left[Y_n\right] \\ \boxed{\text{B}} & = n\operatorname{Var}\left[Y_1\right] \\ \boxed{\text{C}} & = n\Sigma \end{aligned}$$

where: in step A we have used the fact that $Y_1, \ldots, Y_n$ are mutually independent; in step B we have used the fact that $Y_1, \ldots, Y_n$ have the same distribution; in step C we have used formula (52.2) for the covariance matrix of Y_1. ■

52.2.5 Joint moment generating function

The joint moment generating function of a multinomial random vector X is defined for any $t \in \mathbb{R}^K$:

$$M_X(t) = \left(\sum_{j=1}^{K} p_j \exp(t_j)\right)^n$$

Proof. Writing X as a sum of n independent multinomial random vectors with parameters $p_1, \ldots, p_K$ and 1, the joint moment generating function of X is derived from that of the summands:

$$\begin{aligned} M_X(t) & = \operatorname{E}\left[\exp\left(t^\top X\right)\right] \\ & = \operatorname{E}\left[\exp\left(t^\top (Y_1 + \ldots + Y_n)\right)\right] \\ & = \operatorname{E}\left[\exp\left(t^\top Y_1 + \ldots + t^\top Y_n\right)\right] \\ & = \operatorname{E}\left[\prod_{l=1}^{n} \exp\left(t^\top Y_l\right)\right] \\ \boxed{\text{A}} & = \prod_{l=1}^{n} \operatorname{E}\left[\exp\left(t^\top Y_l\right)\right] \\ \boxed{\text{B}} & = \prod_{l=1}^{n} M_{Y_l}(t) \\ \boxed{\text{C}} & = \left(\sum_{j=1}^{K} p_j \exp(t_j)\right)^n \end{aligned}$$

where: in step A we have used the fact that $Y_1, \ldots, Y_n$ are mutually independent; in step B we have used the definition of moment generating function of Y_l; in step C we have used formula (52.3) for the moment generating function of Y_1. ■

52.2.6 Joint characteristic function

The joint characteristic function of X is

$$\varphi_X(t) = \left(\sum_{j=1}^{K} p_j \exp(it_j)\right)^n$$

Proof. The derivation is similar to the derivation of the joint moment generating function:

$$\begin{aligned}
\varphi_X(t) &= \mathrm{E}\left[\exp\left(it^\top X\right)\right] \\
&= \mathrm{E}\left[\exp\left(it^\top (Y_1 + \ldots + Y_n)\right)\right] \\
&= \mathrm{E}\left[\exp\left(it^\top Y_1 + \ldots + it^\top Y_n\right)\right] \\
&= \mathrm{E}\left[\prod_{l=1}^{n} \exp\left(it^\top Y_l\right)\right] \\
\boxed{\text{A}} &= \prod_{l=1}^{n} \mathrm{E}\left[\exp\left(it^\top Y_l\right)\right] \\
\boxed{\text{B}} &= \prod_{l=1}^{n} \varphi_{Y_l}(t) \\
\boxed{\text{C}} &= \left(\sum_{j=1}^{K} p_j \exp(it_j)\right)^n
\end{aligned}$$

where: in step A we have used the fact that $Y_1, \ldots, Y_n$ are mutually independent; in step B we have used the definition of characteristic function of Y_l; in step C we have used formula (52.4) for the characteristic function of Y_1. ■

52.3 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

A shop selling two items, labeled A and B, needs to construct a probabilistic model of the sales that will be generated by its next 10 customers. Each time a customer arrives, only three outcomes are possible: 1) nothing is sold; 2) one unit of item A is sold; 3) one unit of item B is sold. It has been estimated that the probabilities of these three outcomes are 0.50, 0.25 and 0.25 respectively. Furthermore, the shopping behavior of a customer is independent of the shopping behavior of all other customers. Denote by X a 3×1 vector whose entries X_1, X_2 and X_3 are equal to the number of times each of the three outcomes occurs. Derive the expected value and the covariance matrix of X.

Solution

The vector X has a multinomial distribution with parameters

$$p = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}^\top$$

and $n = 10$. Therefore, its expected value is

$$\mathrm{E}[X] = np = 10 \cdot \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}^\top = \begin{bmatrix} 5 & \frac{5}{2} & \frac{5}{2} \end{bmatrix}^\top$$

and its covariance matrix is

$$\begin{aligned}\operatorname{Var}[X] &= n \cdot \begin{bmatrix} p_1(1-p_1) & -p_1p_2 & -p_1p_3 \\ -p_1p_2 & p_2(1-p_2) & -p_2p_3 \\ -p_1p_3 & -p_2p_3 & p_3(1-p_3) \end{bmatrix} \\ &= 10 \cdot \begin{bmatrix} \frac{1}{2}\cdot\frac{1}{2} & -\frac{1}{2}\cdot\frac{1}{4} & -\frac{1}{2}\cdot\frac{1}{4} \\ -\frac{1}{2}\cdot\frac{1}{4} & \frac{1}{4}\cdot\frac{3}{4} & -\frac{1}{4}\cdot\frac{1}{4} \\ -\frac{1}{2}\cdot\frac{1}{4} & -\frac{1}{4}\cdot\frac{1}{4} & \frac{1}{4}\cdot\frac{3}{4} \end{bmatrix} \\ &= 10 \cdot \begin{bmatrix} \frac{1}{4} & -\frac{1}{8} & -\frac{1}{8} \\ -\frac{1}{8} & \frac{3}{16} & -\frac{1}{16} \\ -\frac{1}{8} & -\frac{1}{16} & \frac{3}{16} \end{bmatrix} = \begin{bmatrix} \frac{5}{2} & -\frac{5}{4} & -\frac{5}{4} \\ -\frac{5}{4} & \frac{15}{8} & -\frac{5}{8} \\ -\frac{5}{4} & -\frac{5}{8} & \frac{15}{8} \end{bmatrix}\end{aligned}$$

Exercise 2

Given the assumptions made in the previous exercise, suppose that item A costs \$1,000 and item B costs \$2,000. Derive the expected value and the variance of the total revenue generated by the 10 customers.

Solution

The total revenue Y can be written as a linear transformation of the vector X:

$$Y = AX$$

where

$$A = [0 \quad 1,000 \quad 2,000]$$

By the linearity of the expected value operator, we obtain

$$\begin{aligned}\operatorname{E}[Y] &= \operatorname{E}[AX] = A\operatorname{E}[X] = [0 \quad 1,000 \quad 2,000] \begin{bmatrix} 5 \\ 5/2 \\ 5/2 \end{bmatrix} \\ &= 0 \cdot 5 + 1,000 \cdot 5/2 + 2,000 \cdot 5/2 = 7,500\end{aligned}$$

By using the formula for the covariance matrix of a linear transformation, we obtain

$$\begin{aligned}\operatorname{Var}[Y] &= \operatorname{Var}[AX] = A\operatorname{Var}[X]A^\top \\ &= [0 \quad 1,000 \quad 2,000] \begin{bmatrix} \frac{5}{2} & -\frac{5}{4} & -\frac{5}{4} \\ -\frac{5}{4} & \frac{15}{8} & -\frac{5}{8} \\ -\frac{5}{4} & -\frac{5}{8} & \frac{15}{8} \end{bmatrix} \begin{bmatrix} 0 \\ 1,000 \\ 2,000 \end{bmatrix} \\ &= [0 \quad 1,000 \quad 2,000] \begin{bmatrix} (5/2)\cdot 0 - (5/4)\cdot 1,000 - (5/4)\cdot 2,000 \\ -(5/4)\cdot 0 + (15/8)\cdot 1,000 - (5/8)\cdot 2,000 \\ -(5/4)\cdot 0 - (5/8)\cdot 1,000 + (15/8)\cdot 2,000 \end{bmatrix} \\ &= [0 \quad 1,000 \quad 2,000] \begin{bmatrix} -1,250 - 2,500 \\ 1,875 - 1,250 \\ -1,250 + 3,750 \end{bmatrix} \\ &= [0 \quad 1,000 \quad 2,000] \begin{bmatrix} -3,750 \\ 625 \\ 2,500 \end{bmatrix} \\ &= 0 \cdot (-3,750) + 1,000 \cdot 625 + 2,000 \cdot 2,500 = 5,625,000\end{aligned}$$

Chapter 53

Multivariate normal distribution

The multivariate normal (MV-N) distribution is a multivariate generalization of the one-dimensional normal distribution[1]. In its simplest form, which is called the "standard" MV-N distribution, it describes the joint distribution of a random vector whose entries are mutually independent univariate normal random variables, all having zero mean and unit variance. In its general form, it describes the joint distribution of a random vector that can be represented as a linear transformation of a standard MV-N vector.

It is a common mistake to think that any set of normal random variables, when considered together, form a multivariate normal distribution. This is not the case. In fact, it is possible to construct random vectors that are not MV-N, but whose individual elements have normal distributions. The latter fact is very well-known in the theory of copulae (a theory which allows to specify the distribution of a random vector by first specifying the distribution of its components and then linking the univariate distributions through a function called copula).

The remainder of this lecture illustrates the main characteristics of the multivariate normal distribution, dealing first with the "standard" case and then with the more general case.

53.1 The standard MV-N distribution

The adjective "standard" is used to indicate that the mean of the distribution is equal to zero and its covariance matrix is equal to the identity matrix.

53.1.1 Definition

Standard MV-N normal random vectors are characterized as follows.

Definition 269 *Let X be a $K \times 1$ absolutely continuous random vector. Let its support be the set of K-dimensional real vectors:*

$$R_X = \mathbb{R}^K$$

[1] See p. 375.

We say that X has a ***standard multivariate normal distribution*** *if its joint probability density function*[2] *is*

$$f_X(x) = (2\pi)^{-K/2} \exp\left(-\frac{1}{2}x^\intercal x\right)$$

53.1.2 Relation to the univariate normal distribution

Denote the i-th component of x by x_i. The joint probability density function can be written as

$$\begin{aligned} f_X(x) &= (2\pi)^{-K/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{K} x_i^2\right) \\ &= \prod_{i=1}^{K} (2\pi)^{-1/2} \exp\left(-\frac{1}{2}x_i^2\right) \\ &= \prod_{i=1}^{K} f(x_i) \end{aligned}$$

where

$$f(x_i) = (2\pi)^{-1/2} \exp\left(-\frac{1}{2}x_i^2\right)$$

is the probability density function of a standard normal random variable[3].

Therefore, the K components of X are K mutually independent[4] standard normal random variables. A more detailed proof follows.

Proof. As we have seen, the joint probability density function can be written as

$$f_X(x) = \prod_{i=1}^{K} f(x_i)$$

where $f(x_i)$ is the probability density function of a standard normal random variable. But $f(x_i)$ is also the marginal probability density function[5] of the i-th component of X:

$$\begin{aligned} & f_{X_i}(x_i) \\ = & \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_X(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_K)\, dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_K \\ = & \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{j=1}^{K} f(x_j)\, dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_K \\ = & f(x_i) \int_{-\infty}^{\infty} f(x_1)\, dx_1 \cdot \ldots \cdot \int_{-\infty}^{\infty} f(x_{i-1})\, dx_{i-1} \\ & \cdot \int_{-\infty}^{\infty} f(x_{i+1})\, dx_{i+1} \cdot \ldots \cdot \int_{-\infty}^{\infty} f(x_K)\, dx_K \\ = & f(x_i) \end{aligned}$$

[2] See p. 117.
[3] See p. 376.
[4] See p. 233.
[5] See p. 120.

where, in the last step, we have used the fact that all the integrals are equal to 1, because they are integrals of probability density functions over their respective supports. Therefore, the joint probability density function of X is equal to the product of its marginals, which implies that the components of X are mutually independent. ■

53.1.3 Expected value

The expected value of a standard MV-N random vector X is

$$\mathrm{E}[X] = 0$$

Proof. All the components of X are standard normal random variables and a standard normal random variable has mean 0. ■

53.1.4 Covariance matrix

The covariance matrix of a standard MV-N random vector X is

$$\mathrm{Var}[X] = I$$

where I is the $K \times K$ identity matrix, i.e., a $K \times K$ matrix whose diagonal entries are equal to 1 and whose off-diagonal entries are equal to 0.

Proof. This is proved using the structure of the covariance matrix[6]:

$$\mathrm{Var}[X] = \begin{bmatrix} \mathrm{Var}[X_1] & \mathrm{Cov}[X_1, X_2] & \ldots & \mathrm{Cov}[X_1, X_K] \\ \mathrm{Cov}[X_1, X_2] & \mathrm{Var}[X_2] & \ldots & \mathrm{Cov}[X_2, X_K] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[X_1, X_K] & \mathrm{Cov}[X_2, X_K] & \ldots & \mathrm{Var}[X_K] \end{bmatrix}$$

where X_i is the i-th component of X. Since the components of X are all standard normal random variables, their variances are all equal to 1:

$$\mathrm{Var}[X_1] = \ldots = \mathrm{Var}[X_K] = 1$$

Furthermore, since the components of X are mutually independent and independence implies zero-covariance[7], all the covariances are equal to 0:

$$\mathrm{Cov}[X_i, X_j] = 0 \quad \text{if } i \neq j$$

Therefore,

$$\begin{aligned} \mathrm{Var}[X] &= \begin{bmatrix} \mathrm{Var}[X_1] & \mathrm{Cov}[X_1, X_2] & \ldots & \mathrm{Cov}[X_1, X_K] \\ \mathrm{Cov}[X_1, X_2] & \mathrm{Var}[X_2] & \ldots & \mathrm{Cov}[X_2, X_K] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[X_1, X_K] & \mathrm{Cov}[X_2, X_K] & \ldots & \mathrm{Var}[X_K] \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{bmatrix} = I \end{aligned}$$

■

[6] See p. 189.
[7] See p. 234.

53.1.5 Joint moment generating function

The joint moment generating function of a standard MV-N random vector X is defined for any $t \in \mathbb{R}^K$:

$$M_X(t) = \exp\left(\frac{1}{2}t^\top t\right)$$

Proof. The K components of X are K mutually independent standard normal random variables (see 53.1.2). As a consequence, the joint mgf of X can be derived as follows:

$$\begin{aligned} M_X(t) &= \mathrm{E}\left[\exp\left(t^\top X\right)\right] \\ &= \mathrm{E}\left[\exp\left(t_1X_1 + t_2X_2 + \ldots + t_KX_K\right)\right] \\ &= \mathrm{E}\left[\prod_{j=1}^{K}\exp\left(t_jX_j\right)\right] \\ \boxed{\text{A}} &= \prod_{j=1}^{K}\mathrm{E}\left[\exp\left(t_jX_j\right)\right] \\ \boxed{\text{B}} &= \prod_{j=1}^{K}M_{X_j}(t_j) \end{aligned}$$

where: in step A we have used the fact that the components of X are mutually independent[8]; in step B we have used the definition of moment generating function[9]. The moment generating function of a standard normal random variable[10] is

$$M_{X_j}(t_j) = \exp\left(\frac{1}{2}t_j^2\right)$$

which implies that the joint mgf of X is

$$\begin{aligned} M_X(t) &= \prod_{j=1}^{K}M_{X_j}(t_j) = \prod_{j=1}^{K}\exp\left(\frac{1}{2}t_j^2\right) \\ &= \exp\left(\frac{1}{2}\sum_{j=1}^{K}t_j^2\right) = \exp\left(\frac{1}{2}t^\top t\right) \end{aligned}$$

The mgf $M_{X_j}(t_j)$ of a standard normal random variable is defined for any $t_j \in \mathbb{R}$. As a consequence, the joint mgf of X is defined for any $t \in \mathbb{R}^K$. ■

53.1.6 Joint characteristic function

The joint characteristic function of a standard MV-N random vector X is

$$\varphi_X(t) = \exp\left(-\frac{1}{2}t^\top t\right)$$

[8] See *Mutual independence via expectations* (p. 234).
[9] See p. 297.
[10] See p. 378.

Proof. The K components of X are K mutually independent standard normal random variables (see 53.1.2). As a consequence, the joint characteristic function of X can be derived as follows:

$$\begin{aligned}
\varphi_X(t) &= \mathrm{E}\left[\exp\left(it^\top X\right)\right] \\
&= \mathrm{E}\left[\exp\left(it_1X_1 + it_2X_2 + \ldots + it_KX_K\right)\right] \\
&= \mathrm{E}\left[\prod_{j=1}^{K}\exp\left(it_jX_j\right)\right] \\
\boxed{\text{A}} &= \prod_{j=1}^{K}\mathrm{E}\left[\exp\left(it_jX_j\right)\right] \\
\boxed{\text{B}} &= \prod_{j=1}^{K}\varphi_{X_j}(t_j)
\end{aligned}$$

where: in step A we have used the fact that the components of X are mutually independent; in step B we have used the definition of joint characteristic function[11]. The characteristic function of a standard normal random variable is[12]

$$\varphi_{X_j}(t_j) = \exp\left(-\frac{1}{2}t_j^2\right)$$

which implies that the joint characteristic function of X is

$$\begin{aligned}
\varphi_X(t) &= \prod_{j=1}^{K}\varphi_{X_j}(t_j) = \prod_{j=1}^{K}\exp\left(-\frac{1}{2}t_j^2\right) \\
&= \exp\left(-\frac{1}{2}\sum_{j=1}^{K}t_j^2\right) = \exp\left(-\frac{1}{2}t^\top t\right)
\end{aligned}$$

■

53.2 The MV-N distribution in general

While in the previous section we restricted our attention to the multivariate normal distribution with zero mean and unit variance, we now deal with the general case.

53.2.1 Definition

MV-N random vectors are characterized as follows.

Definition 270 *Let X be a $K \times 1$ absolutely continuous random vector. Let its support be the set of K-dimensional real vectors:*

$$R_X = \mathbb{R}^K$$

[11] See p. 315.
[12] See p. 379.

Let μ be a $K \times 1$ constant vector and V a $K \times K$ symmetric and positive definite matrix. We say that X has a ***multivariate normal distribution*** *with mean μ and covariance V if its joint probability density function is*

$$f_X(x) = (2\pi)^{-K/2} |\det(V)|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^\intercal V^{-1}(x-\mu)\right)$$

We indicate that X has a multivariate normal distribution with mean μ and covariance V by

$$X \sim N(\mu, V)$$

The K random variables $X_1 \ldots, X_K$ constituting the vector X are said to be **jointly normal**.

53.2.2 Relation to the standard MV-N distribution

The next proposition states that a multivariate normal random vector with arbitrary mean and covariance is just a linear transformation of a standard MV-N vector.

Proposition 271 *Let X be a $K \times 1$ random vector having a multivariate normal distribution with mean μ and covariance V. Then,*

$$X = \mu + \Sigma Z \tag{53.1}$$

where Z is a standard MV-N $K \times 1$ vector and Σ is a $K \times K$ invertible matrix such that $V = \Sigma\Sigma^\intercal = \Sigma^\intercal\Sigma$.

Proof. This is proved using the formula for the joint density of a linear function[13] of an absolutely continuous random vector:

$$\begin{aligned}
& f_X(x) \\
= \; & \frac{1}{|\det(\Sigma)|} f_Z\left(\Sigma^{-1}(x-\mu)\right) \\
= \; & \frac{1}{|\det(\Sigma)|} (2\pi)^{-K/2} \exp\left(-\frac{1}{2}\left(\Sigma^{-1}(x-\mu)\right)^\intercal \left(\Sigma^{-1}(x-\mu)\right)\right) \\
= \; & |\det(\Sigma)|^{-\frac{1}{2}} |\det(\Sigma)|^{-\frac{1}{2}} (2\pi)^{-K/2} \exp\left(-\frac{1}{2}(x-\mu)^\intercal \left(\Sigma^{-1}\right)^\intercal \Sigma^{-1}(x-\mu)\right) \\
= \; & (2\pi)^{-K/2} |\det(\Sigma)\det(\Sigma)|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^\intercal \left(\Sigma^\intercal\right)^{-1} \Sigma^{-1}(x-\mu)\right) \\
= \; & (2\pi)^{-K/2} |\det(\Sigma)\det(\Sigma^\intercal)|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^\intercal \left(\Sigma\Sigma^\intercal\right)^{-1}(x-\mu)\right) \\
= \; & (2\pi)^{-K/2} |\det(\Sigma\Sigma^\intercal)|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^\intercal \left(\Sigma\Sigma^\intercal\right)^{-1}(x-\mu)\right) \\
= \; & (2\pi)^{-K/2} |\det(V)|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^\intercal V^{-1}(x-\mu)\right)
\end{aligned}$$

The existence of a matrix Σ satisfying $V = \Sigma\Sigma^\intercal = \Sigma^\intercal\Sigma$ is guaranteed by the fact that V is symmetric and positive definite. ■

[13] See p. 279. Note that $X = g(Z) = \mu + \Sigma X$ is a linear one-to-one mapping because Σ is invertible.

53.2.3 Expected value

The expected value of a MV-N random vector X is

$$\mathrm{E}[X] = \mu$$

Proof. This is an immediate consequence of (53.1) and of the linearity of the expected value[14]:

$$\mathrm{E}[X] = \mathrm{E}[\mu + \Sigma Z] = \mu + \Sigma \mathrm{E}[Z] = \mu + \Sigma 0 = \mu$$

■

53.2.4 Covariance matrix

The covariance matrix of a MV-N random vector X is

$$\mathrm{Var}[X] = V$$

Proof. This is an immediate consequence of (53.1) and of the properties of covariance matrices[15]:

$$\begin{aligned} \mathrm{Var}[X] &= \mathrm{Var}[\mu + \Sigma Z] = \Sigma \mathrm{Var}[Z] \Sigma^{\intercal} \\ &= \Sigma I \Sigma^{\intercal} = \Sigma \Sigma^{\intercal} = V \end{aligned}$$

■

53.2.5 Joint moment generating function

The joint moment generating function of a MV-N random vector X is defined for any $t \in \mathbb{R}^K$:

$$M_X(t) = \exp\left(t^\top \mu + \frac{1}{2} t^\top V t\right)$$

Proof. This is an immediate consequence of (53.1), of the fact that Σ is a $K \times K$ invertible matrix such that $V = \Sigma\Sigma^{\intercal}$, and of the rule for deriving the joint mgf of a linear transformation[16]:

$$\begin{aligned} M_X(t) &= \exp\left(t^\top \mu\right) M_Z\left(\Sigma^\top t\right) \\ &= \exp\left(t^\top \mu\right) \exp\left(\frac{1}{2} t^\top \Sigma \Sigma^\top t\right) \\ &= \exp\left(t^\top \mu + \frac{1}{2} t^\top V t\right) \end{aligned}$$

where

$$M_Z(t) = \exp\left(\frac{1}{2} t^\top t\right)$$

■

[14] See, in particular the *Addition to constant matrices* (p. 148) and *Multiplication by constant matrices* (p. 149) properties of the expected value of a random vector.

[15] See, in particular, the *Addition to constant vectors* (p. 191) and *Multiplication by constant matrices* (p. 191) properties.

[16] See p. 301.

53.2.6 Joint characteristic function

The joint characteristic function of a MV-N random vector X is

$$\varphi_X(t) = \exp\left(it^\top \mu - \frac{1}{2}t^\top V t\right)$$

Proof. This is an immediate consequence of (53.1), of the fact that Σ is a $K \times K$ invertible matrix such that $V = \Sigma\Sigma^\intercal$, and of the rule for deriving the joint characteristic function of a linear transformation[17]:

$$\begin{aligned}\varphi_X(t) &= \exp\left(it^\top \mu\right)\varphi_Z\left(\Sigma^\top t\right) \\ &= \exp\left(it^\top \mu\right)\exp\left(-\frac{1}{2}t^\top \Sigma\Sigma^\top t\right) \\ &= \exp\left(it^\top \mu - \frac{1}{2}t^\top V t\right)\end{aligned}$$

where

$$\varphi_Z(t) = \exp\left(-\frac{1}{2}t^\top t\right)$$

■

53.3 More details

53.3.1 The univariate normal as a special case

The univariate normal distribution[18] is just a special case of the multivariate normal distribution: setting $K = 1$ in the joint density function of the multivariate normal distribution one obtains the density function of the univariate normal distribution[19].

53.3.2 Mutual independence and joint normality

Let $X_1, \ldots, X_K$ be K mutually independent random variables all having a normal distribution. Denote by μ_i the mean of X_i and by σ_i^2 its variance. Then the $K \times 1$ random vector X defined as

$$X = [X_1 \ \ldots \ X_K]^\intercal$$

has a multivariate normal distribution with mean

$$\mu = [\mu_1 \ \ldots \ \mu_K]^\intercal$$

and covariance matrix

$$V = \begin{bmatrix} \sigma_1^2 & 0 & \ldots & 0 \\ 0 & \sigma_2^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_K^2 \end{bmatrix}$$

[17] See p. 317.
[18] See p. 375.
[19] Remember that the determinant and the transpose of a scalar are equal to the scalar itself.

In other words, mutually independent normal random variables are also jointly normal.

Proof. This can be proved by showing that the product of the probability density functions of $X_1, \ldots, X_K$ is equal to the joint probability density function of X (this is left as an exercise). ■

53.3.3 Linear combinations and transformations

Linear transformations and combinations of multivariate normal random vectors are also multivariate normal. This is explained and proved in the lecture entitled *Linear combinations of normals* (p. 469).

53.3.4 Quadratic forms

The lecture entitled *Quadratic forms in normal vectors* (p. 481) discusses quadratic forms involving standard normal random vectors.

53.3.5 Partitioned vectors

The lecture entitled *Partitioned multivariate normal vectors* (p. 477) discusses partitions of normal random vectors into sub-vectors.

53.4 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Let $X = \begin{bmatrix} X_1 & X_2 \end{bmatrix}^\top$ be a multivariate normal random vector with mean

$$\mu = \begin{bmatrix} 1 & 2 \end{bmatrix}^\top$$

and covariance matrix

$$V = \begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix}$$

Prove that the random variable

$$Y = X_1 + X_2$$

has a normal distribution with mean equal to 3 and variance equal to 7.

Hint: use the joint moment generating function of X and its properties.

Solution

The random variable Y can be written as

$$Y = BX$$

where

$$B = \begin{bmatrix} 1 & 1 \end{bmatrix}$$

By using the formula for the joint moment generating function of a linear transformation of a random vector[20]

$$M_Y(t) = M_X\left(B^\top t\right)$$

and the fact that the mgf of a multivariate normal vector X is

$$M_X(t) = \exp\left(t^\top \mu + \frac{1}{2} t^\top V t\right)$$

we obtain

$$\begin{aligned} M_Y(t) &= M_X\left(B^\top t\right) = \exp\left(t^\top B\mu + \frac{1}{2} t^\top BVB^\intercal t\right) \\ &= \exp\left(B\mu t + \frac{1}{2} BVB^\intercal t^2\right) \end{aligned}$$

where, in the last step, we have also used the fact that t is a scalar, because Y is unidimensional. Now,

$$B\mu = [1 \quad 1] \begin{bmatrix} 1 \\ 2 \end{bmatrix} = 1 \cdot 1 + 1 \cdot 2 = 3$$

and

$$\begin{aligned} BVB^\intercal &= \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 3 \cdot 1 + 1 \cdot 1 \\ 1 \cdot 1 + 2 \cdot 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 4 \\ 3 \end{bmatrix} = 1 \cdot 4 + 1 \cdot 3 = 7 \end{aligned}$$

Plugging the values just obtained into the formula for the mgf of Y, we get

$$M_Y(t) = \exp\left(B\mu t + \frac{1}{2} BVB^\intercal t^2\right) = \exp\left(3t + \frac{7}{2} t^2\right)$$

But this is the moment generating function of a normal random variable with mean equal to 3 and variance equal to 7 (see the lecture entitled *Normal distribution* - p. 375). Therefore, Y is a normal random variable with mean equal to 3 and variance equal to 7 (remember that a distribution is completely characterized by its moment generating function).

Exercise 2

Let $X = [X_1 \quad X_2]^\top$ be a multivariate normal random vector with mean

$$\mu = [2 \quad 3]^\top$$

and covariance matrix

$$V = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

Using the joint moment generating function of X, derive the cross-moment[21]

$$\mathrm{E}\left[X_1^2 X_2\right]$$

[20] See p. 301.
[21] See p. 285.

Solution

The joint mgf of X is

$$\begin{aligned} M_X(t) &= \exp\left(t^\top \mu + \frac{1}{2} t^\top V t\right) \\ &= \exp\left(2t_1 + 3t_2 + \frac{1}{2}\left(2t_1^2 + 2t_2^2 + 2t_1 t_2\right)\right) \\ &= \exp\left(2t_1 + 3t_2 + t_1^2 + t_2^2 + t_1 t_2\right) \end{aligned}$$

The third-order cross-moment we want to compute is equal to a third partial derivative of the mgf, evaluated at zero:

$$\mathrm{E}\left[X_1^2 X_2\right] = \left.\frac{\partial^3 M_X(t_1, t_2)}{\partial t_1^2 \partial t_2}\right|_{t_1=0, t_2=0}$$

The partial derivatives are

$$\begin{aligned} \frac{\partial M_X(t_1, t_2)}{\partial t_1} &= (2 + 2t_1 + t_2) \exp\left(2t_1 + 3t_2 + t_1^2 + t_2^2 + t_1 t_2\right) \\ \frac{\partial^2 M_X(t_1, t_2)}{\partial t_1^2} &= \frac{\partial}{\partial t_1}\left(\frac{\partial M_X(t_1, t_2)}{\partial t_1}\right) \\ &= 2 \exp\left(2t_1 + 3t_2 + t_1^2 + t_2^2 + t_1 t_2\right) \\ &\quad + (2 + 2t_1 + t_2)^2 \exp\left(2t_1 + 3t_2 + t_1^2 + t_2^2 + t_1 t_2\right) \\ \frac{\partial^3 M_X(t_1, t_2)}{\partial t_1^2 \partial t_2} &= \frac{\partial}{\partial t_2}\left(\frac{\partial^2 M_X(t_1, t_2)}{\partial t_1^2}\right) \\ &= 2(3 + 2t_2 + t_1) \exp\left(2t_1 + 3t_2 + t_1^2 + t_2^2 + t_1 t_2\right) \\ &\quad + 2(2 + 2t_1 + t_2) \exp\left(2t_1 + 3t_2 + t_1^2 + t_2^2 + t_1 t_2\right) \\ &\quad + (2 + 2t_1 + t_2)^2 (3 + 2t_2 + t_1) \exp\left(2t_1 + 3t_2 + t_1^2 + t_2^2 + t_1 t_2\right) \end{aligned}$$

Thus,

$$\begin{aligned} \mathrm{E}\left[X_1^2 X_2\right] &= \left.\frac{\partial^3 M_X(t_1, t_2)}{\partial t_1^2 \partial t_2}\right|_{t_1=0, t_2=0} = 2 \cdot 3 \cdot 1 + 2 \cdot 2 \cdot 1 + 2^2 \cdot 3 \cdot 1 \\ &= 6 + 4 + 12 = 22 \end{aligned}$$

Chapter 54

Multivariate Student's t distribution

This lecture deals with the multivariate (MV) Student's t distribution. We first introduce the special case in which the mean is equal to zero and the scale matrix is equal to the identity matrix. We then deal with the more general case.

54.1 The standard MV Student's t distribution

The adjective "standard" is used for a multivariate Student's t distribution having zero mean and unit scale matrix.

54.1.1 Definition

Standard multivariate Student's t random vectors are characterized as follows.

Definition 272 *Let X be a $K \times 1$ absolutely continuous random vector. Let its support be the set of K-dimensional real vectors:*

$$R_X = \mathbb{R}^K$$

Let $n \in \mathbb{R}_{++}$. We say that X has a ***standard multivariate Student's t distribution*** *with n degrees of freedom if its joint probability density function[1] is*

$$f_X(x) = c\left(1 + \frac{1}{n}x^\intercal x\right)^{-(n+K)/2}$$

where

$$c = (n\pi)^{-K/2}\frac{\Gamma(n/2 + K/2)}{\Gamma(n/2)}$$

and $\Gamma()$ is the Gamma function[2].

[1] See p. 117.
[2] See p. 55.

54.1.2 Relation to the univariate Student's t distribution

When $K = 1$, the definition of the standard multivariate Student's t distribution coincides with the definition of the standard univariate Student's t distribution[3]:

$$\begin{aligned} f_X(x) &= (n\pi)^{-K/2} \frac{\Gamma(n/2+K/2)}{\Gamma(n/2)} \left(1+\frac{1}{n}x^\intercal x\right)^{-(n+K)/2} \\ &= n^{-1/2}\pi^{-1/2}\frac{\Gamma(n/2+1/2)}{\Gamma(n/2)} \left(1+\frac{x^2}{n}\right)^{-(n+1)/2} \\ \boxed{\text{A}} &= n^{-1/2}\frac{\Gamma(n/2+1/2)}{\Gamma(1/2)\Gamma(n/2)} \left(1+\frac{x^2}{n}\right)^{-(n+1)/2} \\ &= \frac{1}{\sqrt{n}B(n/2,1/2)} \left(1+\frac{x^2}{n}\right)^{-(n+1)/2} \end{aligned}$$

where: in step A we have used the fact that[4] $\Gamma(1/2) = \pi^{1/2}$.

54.1.3 Relation to the Gamma and MV normal distributions

A standard multivariate Student's t random vector can be written as a multivariate normal random vector[5] whose covariance matrix is scaled by the reciprocal of a Gamma random variable[6], as shown by the following proposition.

Proposition 273 (Integral representation) *The joint probability density function of X can be written as*

$$f_X(x) = \int_0^\infty f_{X|Z=z}(x)\, f_Z(z)\, dz$$

where:

1. *$f_{X|Z=z}(x)$ is the joint probability density function of a multivariate normal distribution with mean 0 and covariance $V = \frac{1}{z}I$ (where I is the $K \times K$ identity matrix):*

$$\begin{aligned} f_{X|Z=z}(x) &= c_1 \left|\det(V)\right|^{-1/2} \exp\left(-\frac{1}{2}x^\intercal V^{-1}x\right) \\ &= c_1 \left|\left(\frac{1}{z}\right)^K \det(I)\right|^{-1/2} \exp\left(-\frac{1}{2}x^\intercal \left(\frac{1}{z}\right)^{-1} I^{-1}x\right) \\ &= c_1 z^{K/2} \exp\left(-z\frac{1}{2}x^\intercal x\right) \end{aligned}$$

where

$$c_1 = (2\pi)^{-K/2}$$

[3] See p. 407.
[4] See p. 57.
[5] See p. 439.
[6] See p. 397.

2. *$f_Z(z)$ is the probability density function of a Gamma random variable with parameters n and $h = 1$:*

$$f_Z(z) = c_2 z^{n/2-1} \exp\left(-n\frac{1}{2}z\right)$$

where

$$c_2 = \frac{n^{n/2}}{2^{n/2}\Gamma(n/2)}$$

Proof. We need to prove that

$$f_X(x) = \int_0^\infty f_{X|Z=z}(x) f_Z(z)\, dz$$

where

$$f_{X|Z=z}(x) = c_1 z^{K/2} \exp\left(-z\frac{1}{2}x^\intercal x\right)$$

and

$$f_Z(z) = c_2 z^{n/2-1} \exp\left(-n\frac{1}{2}z\right)$$

We start from the integrand function:

$$\begin{aligned}
f_{X|Z=z}(x) f_Z(z) &= c_1 z^{K/2} \exp\left(-z\frac{1}{2}x^\intercal x\right) c_2 z^{n/2-1} \exp\left(-n\frac{1}{2}z\right) \\
&= c_1 c_2 z^{(n+K)/2-1} \exp\left(-\left(x^\intercal x + n\right)\frac{1}{2}z\right) \\
&= c_1 c_2 z^{(n+K)/2-1} \exp\left(-\frac{n+K}{\left(\frac{n+K}{x^\intercal x+n}\right)}\frac{1}{2}z\right) \\
&= c_1 c_2 \frac{1}{c_3} c_3 z^{(n+K)/2-1} \exp\left(-\frac{n+K}{\left(\frac{n+K}{x^\intercal x+n}\right)}\frac{1}{2}z\right) \\
&= c_1 c_2 \frac{1}{c_3} f_{Z|X=x}(z)
\end{aligned}$$

where

$$c_3 = \frac{\left((n+K)/\left(\frac{n+K}{x^\intercal x+n}\right)\right)^{(n+K)/2}}{2^{(n+K)/2}\Gamma\left((n+K)/2\right)} = \frac{\left(x^\intercal x+n\right)^{(n+K)/2}}{2^{n/2}2^{K/2}\Gamma\left(\frac{n}{2}+\frac{K}{2}\right)}$$

and $f_{Z|X=x}(z)$ is the probability density function of a random variable having a Gamma distribution with parameters $n+K$ and $\frac{n+K}{x^\intercal x+n}$. Therefore,

$$\begin{aligned}
&\int_0^\infty f_{X|Z=z}(x) f_Z(z)\, dz \\
&= \int_0^\infty c_1 c_2 \frac{1}{c_3} f_{Z|X=x}(z)\, dz \\
\boxed{\text{A}} \quad &= c_1 c_2 \frac{1}{c_3} \int_0^\infty f_{Z|X=x}(z)\, dz
\end{aligned}$$

$$
\begin{aligned}
\boxed{\text{B}} &= c_1 c_2 \frac{1}{c_3} \\
&= (2\pi)^{-K/2} \frac{n^{n/2}}{2^{n/2}\Gamma(n/2)} 2^{n/2} 2^{K/2} \Gamma\left(\frac{n}{2}+\frac{K}{2}\right) \left(x^\intercal x + n\right)^{-(n+K)/2} \\
&= (2\pi)^{-K/2} \frac{n^{n/2}}{\Gamma(n/2)} 2^{K/2} \Gamma\left(\frac{n}{2}+\frac{K}{2}\right) \left(n\left(1+\frac{1}{n}x^\intercal x\right)\right)^{-(n+K)/2} \\
&= (2\pi)^{-K/2} \frac{n^{n/2}}{\Gamma(n/2)} \left(\frac{1}{2}\right)^{-K/2} \Gamma\left(\frac{n}{2}+\frac{K}{2}\right) \\
&\quad \cdot n^{-n/2-K/2} \left(1+\frac{1}{n}x^\intercal x\right)^{-(n+K)/2} \\
&= (2\pi)^{-K/2} \frac{\Gamma\left(\frac{n}{2}+\frac{K}{2}\right)}{\Gamma(n/2)} \left(\frac{1}{2}\right)^{-K/2} n^{-K/2} \left(1+\frac{1}{n}x^\intercal x\right)^{-(n+K)/2} \\
&= \left(2\pi\frac{1}{2}n\right)^{-K/2} \frac{\Gamma\left(\frac{n}{2}+\frac{K}{2}\right)}{\Gamma(n/2)} \left(1+\frac{1}{n}x^\intercal x\right)^{-(n+K)/2} \\
&= (n\pi)^{-K/2} \frac{\Gamma\left(\frac{n}{2}+\frac{K}{2}\right)}{\Gamma(n/2)} \left(1+\frac{1}{n}x^\intercal x\right)^{-(n+K)/2} \\
&= f_X(x)
\end{aligned}
$$

where: in step $\boxed{\text{A}}$ we have used the fact that c_1, c_2 and c_3 do not depend on z; in step $\boxed{\text{B}}$ we have used the fact that the integral of a probability density function over its support is 1. ■

Since X has a multivariate normal distribution with mean 0 and covariance $V = \frac{1}{z}I$, conditional on $Z = z$, then we can also think of it as a ratio

$$X = \frac{1}{\sqrt{Z}} Y$$

where Y has a standard multivariate normal distribution, Z has a Gamma distribution and Y and Z are independent.

54.1.4 Marginals

The marginal distribution of the i-th component of X (denote it by X_i) is a standard Student's t distribution with n degrees of freedom. It suffices to note that the marginal probability density function[7] of X_i can be written as

$$f_{X_i}(x_i) = \int_0^\infty f_{X_i|Z=z}(x_i) f_Z(z)\, dz$$

where $f_{X_i|Z=z}(x_i)$ is the marginal density of $X_i\,|Z = z$, i.e., the density of a normal random variable[8] with mean 0 and variance $\frac{1}{z}$:

$$f_{X_i|Z=z}(x_i) = (2\pi)^{-1/2} z^{1/2} \exp\left(-z\frac{1}{2}x_i^2\right)$$

[7] See p. 120.
[8] See p. 375.

Proof. This is obtained by exchanging the order of integration:

$$\begin{aligned}
& f_{X_i}(x_i) \\
= & \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_X(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_K)\, dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_K \\
= & \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_0^{\infty} f_{X|Z=z}(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_K)\, f_Z(z)\, dz \\
& \cdot dx_1 \ldots dx_{i-1} x_{i+1} \ldots dx_K \\
= & \int_0^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X|Z=z}(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_K) \\
& \cdot dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_K f_Z(z)\, dz \\
= & \int_0^{\infty} f_{X_i|Z=z}(x_i)\, f_Z(z)\, dz
\end{aligned}$$

But, by Proposition 273, the fact that

$$f_{X_i}(x_i) = \int_0^{\infty} f_{X_i|Z=z}(x_i)\, f_Z(z)\, dz$$

implies that X_i has a standard multivariate Student's t distribution with n degrees of freedom (hence a standard univariate Student's t distribution with n degrees of freedom, because the two are the same thing when $K = 1$). ■

54.1.5 Expected value

The expected value of a standard multivariate Student's t random vector X is well-defined only when $n > 1$ and it is

$$\mathrm{E}[X] = 0 \tag{54.1}$$

Proof. $\mathrm{E}[X] = 0$ if $\mathrm{E}[X_i] = 0$ for all K components X_i. But the marginal distribution of X_i is a standard Student's t distribution with n degrees of freedom. Therefore,

$$\mathrm{E}[X_i] = 0$$

provided $n > 1$. ■

54.1.6 Covariance matrix

The covariance matrix of a standard multivariate Student's t random vector X is well-defined only when $n > 2$ and it is

$$\mathrm{Var}[X] = \frac{n}{n-2} I$$

where I is the $K \times K$ identity matrix.

Proof. First of all, we can rewrite the covariance matrix as follows:

$$\begin{aligned}
& \mathrm{Var}[X] \\
\boxed{\text{A}} \quad = & \quad \mathrm{E}\left[XX^\top\right] - \mathrm{E}[X]\,\mathrm{E}[X]^\top \\
\boxed{\text{B}} \quad = & \quad \mathrm{E}\left[XX^\top\right]
\end{aligned}$$

$$\begin{aligned}
\boxed{\text{C}} &= \mathrm{E}\left[\mathrm{E}\left[XX^{\top}\,|Z=z\right]\right] \\
\boxed{\text{D}} &= \mathrm{E}\left[\mathrm{E}\left[XX^{\top}\,|Z=z\right]-\mathrm{E}\left[X\,|Z=z\right]\mathrm{E}\left[X\,|Z=z\right]^{\top}\right] \\
\boxed{\text{E}} &= \mathrm{E}\left[\mathrm{Var}\left[X\,|Z=z\right]\right] \\
&= \mathrm{E}\left[\frac{1}{z}I\right] \\
&= \mathrm{E}\left[\frac{1}{z}\right]I
\end{aligned}$$

where: in step $\boxed{\text{A}}$ we have used the formula for computing the covariance matrix[9]; in step $\boxed{\text{B}}$ we have used the fact that $\mathrm{E}[X]=0$ (eq. 54.1); in step $\boxed{\text{C}}$ we have used the Law of Iterated Expectations[10]; in step $\boxed{\text{D}}$ we have used the fact that

$$\mathrm{E}\left[X\,|Z=z\right]=0$$

because X has a normal distribution with zero mean, conditional on $Z=z$; in step $\boxed{\text{E}}$ we have used the fact that, by the definition of variance,

$$\mathrm{Var}\left[X\,|Z=z\right]=\mathrm{E}\left[XX^{\top}\,|Z=z\right]-\mathrm{E}\left[X\,|Z=z\right]\mathrm{E}\left[X\,|Z=z\right]^{\top}$$

But

$$\begin{aligned}
&\mathrm{E}\left[\frac{1}{z}\right] \\
&= \int_0^{\infty}\frac{1}{z}f_Z\left(z\right)dz \\
&= \int_0^{\infty}\frac{1}{z}\frac{n^{n/2}}{2^{n/2}\Gamma\left(n/2\right)}z^{n/2-1}\exp\left(-n\frac{1}{2}z\right)dz \\
&= \frac{n^{n/2}}{2^{n/2}\Gamma\left(n/2\right)}\int_0^{\infty}z^{(n-2)/2-1}\exp\left(-n\frac{1}{2}z\right)dz \\
&= \frac{n^{n/2}}{2^{n/2}\Gamma\left(n/2\right)}\frac{2^{(n-2)/2}\cdot\Gamma\left(\left(n-2\right)/2\right)}{n^{(n-2)/2}} \\
&\quad\cdot\int_0^{\infty}\frac{n^{(n-2)/2}}{2^{(n-2)/2}\cdot\Gamma\left(\left(n-2\right)/2\right)}z^{(n-2)/2-1}\exp\left(-\frac{n-2}{\frac{n-2}{n}}\frac{1}{2}z\right)dz \\
\boxed{\text{A}} &= \frac{2^{(n-2)/2}}{2^{n/2}}\frac{n^{n/2}}{n^{(n-2)/2}}\frac{\Gamma\left(\left(n-2\right)/2\right)}{\Gamma\left(n/2\right)}\int_0^{\infty}\varphi\left(z\right)dz \\
\boxed{\text{B}} &= \frac{2^{(n-2)/2}}{2^{n/2}}\frac{n^{n/2}}{n^{(n-2)/2}}\frac{\Gamma\left(\left(n-2\right)/2\right)}{\Gamma\left(n/2\right)} \\
\boxed{\text{C}} &= \frac{1}{2}\cdot n\cdot\frac{1}{\left(n-2\right)/2} \\
&= \frac{n}{n-2}
\end{aligned}$$

[9] See p. 190.
[10] See p. 225.

where: in step A we have defined

$$\varphi(z) = \frac{n^{(n-2)/2}}{2^{(n-2)/2} \cdot \Gamma\left((n-2)/2\right)} z^{(n-2)/2-1} \exp\left(-\frac{n-2}{\frac{n-2}{n}} \frac{1}{2} z\right)$$

which is the density of a Gamma random variable with parameters $n-2$ and $\frac{n-2}{n}$; in step B we have used the fact that the integral of a probability density function over its support is equal to 1; in step C we have used the properties of the Gamma function. As a consequence,

$$\operatorname{Var}[X] = \mathrm{E}\left[\frac{1}{z}\right] I = \frac{n}{n-2} I$$

■

54.2 The MV Student's t distribution in general

While in the previous section we restricted our attention to the multivariate Student's t distribution with zero mean and unit scale matrix, we now deal with the general case.

54.2.1 Definition

Multivariate Student's t random vectors are characterized as follows.

Definition 274 *Let X be a $K \times 1$ absolutely continuous random vector. Let its support be the set of K-dimensional real vectors:*

$$R_X = \mathbb{R}^K$$

Let $n \in \mathbb{R}_{++}$, let μ be a $K \times 1$ vector and let V be a $K \times K$ symmetric and positive definite matrix. We say that X has a ***multivariate Student's t distribution*** *with mean μ, scale matrix V and n degrees of freedom if its joint probability density function is*

$$f_X(x) = c\left(1 + \frac{1}{n}(x-\mu)^\intercal V^{-1}(x-\mu)\right)^{-(n+K)/2}$$

where

$$c = (n\pi)^{-K/2} \frac{\Gamma\left(n/2 + K/2\right)}{\Gamma\left(n/2\right)} \left|\det(V)\right|^{-1/2}$$

We indicate that X has a multivariate Student's t distribution with mean μ, scale matrix V and n degrees of freedom by

$$X \sim T(\mu, V, n)$$

54.2.2 Relation to the standard MV Student's t distribution

If $X \sim T(\mu, V, n)$, then X is a linear function of a standard Student's t random vector.

Proposition 275 *Let $X \sim T(\mu, V, n)$. Then,*

$$X = \mu + \Sigma Z \tag{54.2}$$

where Z is a $K \times 1$ vector having a standard multivariate Student's t distribution with n degrees of freedom and Σ is a $K \times K$ invertible matrix such that $V = \Sigma\Sigma^\intercal = \Sigma^\intercal\Sigma$.

Proof. This is proved using the formula for the joint density of a linear function[11] of an absolutely continuous random vector:

$$\begin{aligned}
f_X(x) &= \frac{1}{|\det(\Sigma)|} f_Z\left(\Sigma^{-1}(x-\mu)\right) \\
&= \frac{1}{|\det(\Sigma)|} (n\pi)^{-K/2} \frac{\Gamma(n/2+K/2)}{\Gamma(n/2)} \\
&\quad \cdot \left(1 + \frac{1}{n}\left(\Sigma^{-1}(x-\mu)\right)^\intercal \left(\Sigma^{-1}(x-\mu)\right)\right)^{-(n+K)/2} \\
&= (n\pi)^{-K/2} \frac{\Gamma(n/2+K/2)}{\Gamma(n/2)} |\det(\Sigma)|^{-\frac{1}{2}} |\det(\Sigma)|^{-\frac{1}{2}} \\
&\quad \cdot \left(1 + \frac{1}{n}(x-\mu)^\intercal \left(\Sigma^{-1}\right)^\intercal \Sigma^{-1}(x-\mu)\right)^{-(n+K)/2} \\
&= (n\pi)^{-K/2} \frac{\Gamma(n/2+K/2)}{\Gamma(n/2)} |\det(\Sigma)\det(\Sigma)|^{-\frac{1}{2}} \\
&\quad \cdot \left(1 + \frac{1}{n}(x-\mu)^\intercal \left(\Sigma^\intercal\right)^{-1} \Sigma^{-1}(x-\mu)\right)^{-(n+K)/2} \\
&= (n\pi)^{-K/2} \frac{\Gamma(n/2+K/2)}{\Gamma(n/2)} |\det(\Sigma)\det(\Sigma^\intercal)|^{-\frac{1}{2}} \\
&\quad \cdot \left(1 + \frac{1}{n}(x-\mu)^\intercal \left(\Sigma\Sigma^\intercal\right)^{-1}(x-\mu)\right)^{-(n+K)/2} \\
&= (n\pi)^{-K/2} \frac{\Gamma(n/2+K/2)}{\Gamma(n/2)} |\det(\Sigma\Sigma^\intercal)|^{-\frac{1}{2}} \\
&\quad \cdot \left(1 + \frac{1}{n}(x-\mu)^\intercal \left(\Sigma\Sigma^\intercal\right)^{-1}(x-\mu)\right)^{-(n+K)/2} \\
&= (n\pi)^{-K/2} \frac{\Gamma(n/2+K/2)}{\Gamma(n/2)} |\det(V)|^{-\frac{1}{2}} \\
&\quad \cdot \left(1 + \frac{1}{n}(x-\mu)^\intercal V^{-1}(x-\mu)\right)^{-(n+K)/2}
\end{aligned}$$

The existence of a matrix Σ satisfying $V = \Sigma\Sigma^\intercal = \Sigma^\intercal\Sigma$ is guaranteed by the fact that V is symmetric and positive definite. ■

54.2.3 Expected value

The expected value of a multivariate Student's t random vector X is

$$\mathrm{E}[X] = \mu$$

[11] See p. 279. Note that $X = g(Z) = \mu + \Sigma X$ is a linear one-to-one mapping since Σ is invertible.

Proof. This is an immediate consequence of (54.2) and of the linearity of the expected value[12]:

$$\mathrm{E}[X] = \mathrm{E}[\mu + \Sigma Z] = \mu + \Sigma \mathrm{E}[Z] = \mu + \Sigma 0 = \mu$$

■

54.2.4 Covariance matrix

The covariance matrix of a multivariate Student's t random vector X is

$$\mathrm{Var}[X] = \frac{n}{n-2} V$$

Proof. This is an immediate consequence of (54.2) and of the properties of covariance matrices[13]:

$$\begin{aligned} \mathrm{Var}[X] &= \mathrm{Var}[\mu + \Sigma Z] \\ &= \Sigma \mathrm{Var}[Z] \Sigma^{\intercal} \\ &= \Sigma \frac{n}{n-2} I \Sigma^{\intercal} \\ &= \frac{n}{n-2} \Sigma \Sigma^{\intercal} \\ &= \frac{n}{n-2} V \end{aligned}$$

■

54.3 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Let X be a multivariate normal random vector with mean $\mu = 0$ and covariance matrix V. Let $Z_1, \ldots, Z_n$ be n normal random variables having zero mean and variance σ^2. Suppose that $Z_1, \ldots, Z_n$ are mutually independent, and also independent of X. Find the distribution of the random vector Y defined as

$$Y = \frac{2}{\sqrt{Z_1^2 + \ldots + Z_n^2}} X$$

Solution

We can write

$$Y = \frac{2}{\sqrt{Z_1^2 + \ldots + Z_n^2}} X$$

[12] See, in particular the *Addition to constant matrices* (p. 148) and *Multiplication by constant matrices* (p. 149) properties of the expected value of a random vector.

[13] See, in particular, the *Addition to constant vectors* (p. 191) and *Multiplication by constant matrices* (p. 191) properties.

$$
\begin{aligned}
&= \frac{2}{\sqrt{n}\sigma\left(\sqrt{\left(W_1^2+\ldots+W_n^2\right)/n}\right)}X \\
&= \frac{2}{\sqrt{n}\sigma}\Sigma\frac{1}{\sqrt{\left(W_1^2+\ldots+W_n^2\right)/n}}Q
\end{aligned}
$$

where $W_1,\ldots,W_n$ are standard normal random variables, Q has a standard multivariate normal distribution and $\Sigma\Sigma^\top = V$ (if you are wondering about the standardizations $Z_i^2 = \sigma^2 W_i^2$ and $X = \Sigma Q$, revise the lectures entitled *Normal distribution* - p. 375, and *Multivariate normal distribution* - p. 439). Now, the sum $W_1^2+\ldots+W_n^2$ has a Chi-square distribution with n degrees of freedom and the ratio

$$\frac{W_1^2+\ldots+W_n^2}{n}$$

has a Gamma distribution with parameters n and $h = 1$ (see the lectures entitled *Chi-square distribution* - p. 387, and *Gamma distribution* - p. 397). As a consequence, by the results in Subsection 54.1.3, the ratio

$$\frac{1}{\sqrt{\left(W_1^2+\ldots+W_n^2\right)/n}}Q$$

has a standard multivariate Student's t distribution with n degrees of freedom. Finally, by equation (54.2), Y has a multivariate Student's t distribution with mean 0 and scale matrix

$$\left(\frac{2}{\sqrt{n}\sigma}\Sigma\right)\left(\frac{2}{\sqrt{n}\sigma}\Sigma\right)^\top = \frac{4}{n\sigma^2}\Sigma\Sigma^\top = \frac{4}{n\sigma^2}V$$

Chapter 55

Wishart distribution

This lecture provides a brief introduction to the Wishart distribution, which is a multivariate generalization of the Gamma distribution[1].

In previous lectures we have explained that:

1. a Chi-square random variable[2] with n degrees of freedom can be seen as a sum of squares of n independent normal random variables having mean 0 and variance 1;

2. a Gamma random variable with parameters n and h can be seen as a sum of squares of n independent normal random variables having mean 0 and variance h/n.

A Wishart random matrix[3] with parameters n and H can be seen as a sum of outer products of n independent multivariate normal random vectors[4] having mean 0 and covariance matrix $\frac{1}{n}H$. In this sense, the Wishart distribution can be considered a generalization of the Gamma distribution (take point 2 above and substitute normal random variables with multivariate normal random vectors, squares with outer products and the variance h/n with the covariance matrix $\frac{1}{n}H$).

At the end of this lecture you can find a brief review of some basic concepts in matrix algebra that will be helpful in understanding the remainder of this lecture.

55.1 Definition

Wishart random matrices are characterized as follows:

Definition 276 *Let W be a $K \times K$ absolutely continuous random matrix. Let its support be the set of all $K \times K$ symmetric and positive definite real matrices:*

$$R_W = \left\{ w \in \mathbb{R}^{K \times K} : w \text{ is symmetric and positive definite} \right\}$$

Let n be a constant such that $n > K - 1$ and let H be a symmetric and positive definite matrix. We say that W has a ***Wishart distribution*** *with parameters n*

[1] See p. 397.
[2] See p. 387.
[3] See p. 119.
[4] See p. 439.

and H if its joint probability density function[5] *is*

$$f_W(w) = c\left[\det(w)\right]^{n/2-(K+1)/2} \exp\left(-\frac{n}{2}\operatorname{tr}\left(H^{-1}w\right)\right)$$

where

$$c = \frac{n^{n/2}}{2^{nK/2}\left[\det(H)\right]^{n/2}\pi^{K(K-1)/4}\prod_{j=1}^{K}\Gamma\left(n/2+(1-j)/2\right)}$$

and $\Gamma()$ is the Gamma function[6]*.*

The parameter n needs not be an integer, but, when n is not an integer, W can no longer be interpreted as a sum of outer products of multivariate normal random vectors.

55.2 Relation to the MV normal distribution

The following proposition provides the link between the multivariate normal distribution and the Wishart distribution:

Proposition 277 *Let $X_1, \ldots, X_n$ be n independent $K \times 1$ random vectors all having a multivariate normal distribution with mean 0 and covariance matrix $\frac{1}{n}H$. Let $K \leq n$. Define:*

$$W = \sum_{i=1}^{n} X_i X_i^\top$$

Then W has a Wishart distribution with parameters n and H.

Proof. The proof of this proposition is quite lengthy and complicated. The interested reader might have a look at Ghosh and Sinha[7] (2002). ■

55.3 Expected value

The expected value of a Wishart random matrix W is

$$\mathrm{E}[W] = H$$

Proof. We do not provide a fully general proof, but we prove this result only for the special case in which n is integer and W can be written as

$$W = \sum_{i=1}^{n} X_i X_i^\top$$

(see subsection 55.2 above). In this case:

$$\mathrm{E}[W] = \mathrm{E}\left[\sum_{i=1}^{n} X_i X_i^\top\right]$$

[5] See p. 117.
[6] See p. 55.
[7] Ghosh, M. and Sinha, B. K. (2002) "A simple derivation of the Wishart distribution", *The American Statistician*, 56, 100-101.

$$\begin{aligned} & = \sum_{i=1}^{n} \mathrm{E}\left[X_i X_i^\top\right] \\ \boxed{\text{A}} \quad & = \sum_{i=1}^{n} \left\{\mathrm{Var}\left[X_i\right] + \mathrm{E}\left[X_i\right]\mathrm{E}\left[X_i^\top\right]\right\} \\ \boxed{\text{B}} \quad & = \sum_{i=1}^{n} \mathrm{Var}\left[X_i\right] \\ & = \sum_{i=1}^{n} \frac{1}{n} H = n\frac{1}{n} H = H \end{aligned}$$

where: in step A we have used the fact that the covariance matrix of X can be written as[8]

$$\mathrm{Var}\left[X_i\right] = \mathrm{E}\left[X_i X_i^\top\right] - \mathrm{E}\left[X_i\right]\mathrm{E}\left[X_i^\top\right]$$

and in step B we have used the fact that $\mathrm{E}\left[X_i\right] = 0$. ■

55.4 Covariance matrix

The concept of covariance matrix is well-defined only for random vectors. However, when dealing with a random matrix, one might want to compute the covariance matrix of its associated vectorization (if you are not familiar with the concept of vectorization, see the review of matrix algebra below for a definition). Therefore, in the case of a Wishart random matrix W, we might want to compute the following covariance matrix:

$$\mathrm{Var}\left[\mathrm{vec}\left(W\right)\right]$$

Since $\mathrm{vec}\left(W\right)$, the vectorization of W, is a $K^2 \times 1$ random vector, V is a $K^2 \times K^2$ matrix.

It is possible to prove that:

$$\mathrm{Var}\left[\mathrm{vec}\left(W\right)\right] = \frac{1}{n}\left(I + P_{\mathrm{vec}(W)}\right)\left(H \otimes H\right)$$

where $\otimes$ denotes the Kronecker product and $P_{\mathrm{vec}(W)}$ is the transposition-permutation matrix associated to $\mathrm{vec}\left(W\right)$.

Proof. The proof of this formula can be found in Muirhead[9] (2005). ■

There is a simpler expression for the covariances between the diagonal entries of W:

$$\mathrm{Cov}\left[W_{ii}, W_{jj}\right] = \frac{2}{n} H_{ij}^2$$

Proof. Again, we do not provide a fully general proof, but we prove this result only for the special case in which n is integer and W can be written as:

$$W = \sum_{i=1}^{n} X_i X_i^\top$$

(see above). To compute this covariance, we first need to compute the following fourth cross-moment:

$$\mathrm{E}\left[X_{si}^2 X_{sj}^2\right]$$

[8] See p. 190.

[9] Muirhead, R.J. (2005) *Aspects of multivariate statistical theory*, Wiley.

where X_{si} denotes the i-th component ($i = 1, \ldots, K$) of the random vector X_s ($s = 1, \ldots, n$). This cross-moment can be computed by taking the fourth cross-partial derivative of the joint moment generating function[10] of X_{si} and X_{sj} and evaluating it at zero. While this is not complicated, the algebra is quite tedious. I recommend doing it with computer algebra, for example utilizing the Matlab Symbolic Toolbox and the following four commands:

```
syms t1 t2 s1 s2 s12;
f=exp(0.5*(s1^2)*(t1^2)+0.5*(s2^2)*(t2^2)+s12*t1*t2);
d4f=diff(diff(f,t1,2),t2,2);
subs(d4f,{t1,t2},{0,0})
```

The result of the computations is

$$\begin{aligned} \mathrm{E}\left[X_{si}^2 X_{sj}^2\right] &= \mathrm{Var}\left[X_{si}\right]\mathrm{Var}\left[X_{sj}\right] + 2\left(\mathrm{Cov}\left[X_{si}, X_{sj}\right]\right)^2 \\ &= \left(\frac{1}{n}H_{ii}\right)\left(\frac{1}{n}H_{jj}\right) + 2\left(\frac{1}{n}H_{ij}\right)^2 \end{aligned}$$

Using this result, the covariance between W_{ii} and W_{jj} is derived as follows:

$$\begin{aligned} & \mathrm{Cov}\left[W_{ii}, W_{jj}\right] \\ = \; & \mathrm{Cov}\left[\sum_{s=1}^{n} X_{si}^2, \sum_{t=1}^{n} X_{tj}^2\right] \\ \boxed{\text{A}} \quad = \; & \sum_{s=1}^{n}\sum_{t=1}^{n} \mathrm{Cov}\left[X_{si}^2, X_{tj}^2\right] \\ \boxed{\text{B}} \quad = \; & \sum_{s=1}^{n} \mathrm{Cov}\left[X_{si}^2, X_{sj}^2\right] \\ \boxed{\text{C}} \quad = \; & \sum_{s=1}^{n} \left\{\mathrm{E}\left[X_{si}^2 X_{sj}^2\right] - \mathrm{E}\left[X_{si}^2\right]\mathrm{E}\left[X_{sj}^2\right]\right\} \\ = \; & \sum_{s=1}^{n} \mathrm{E}\left[X_{si}^2 X_{sj}^2\right] - \sum_{s=1}^{n} \mathrm{E}\left[X_{si}^2\right]\mathrm{E}\left[X_{sj}^2\right] \\ = \; & \sum_{s=1}^{n} \left[\left(\frac{1}{n}H_{ii}\right)\left(\frac{1}{n}H_{jj}\right) + 2\left(\frac{1}{n}H_{ij}\right)^2\right] - \sum_{s=1}^{n}\left(\frac{1}{n}H_{ii}\right)\left(\frac{1}{n}H_{jj}\right) \\ = \; & \sum_{s=1}^{n}\left(\frac{1}{n^2}H_{ii}H_{jj} + \frac{1}{n^2}2H_{ij}^2\right) - \sum_{s=1}^{n}\frac{1}{n^2}H_{ii}H_{jj} \\ = \; & n\left(\frac{1}{n^2}H_{ii}H_{jj} + \frac{1}{n^2}2H_{ij}^2\right) - n\left(\frac{1}{n^2}H_{ii}H_{jj}\right) \\ = \; & \frac{1}{n}\left(H_{ii}H_{jj} + 2H_{ij}^2 - H_{ii}H_{jj}\right) \\ = \; & \frac{2}{n}H_{ij}^2 \end{aligned}$$

where: in step A we have used the bilinearity of covariance; in step B we have used the fact that

$$\mathrm{Cov}\left[X_{si}^2, X_{tj}^2\right] = 0 \text{ for } s \neq t$$

[10] See p. 297.

in step C we have used the usual covariance formula[11]. ■

55.5 Review of some facts in matrix algebra

55.5.1 Outer products

As the Wishart distribution involves outer products of multivariate normal random vectors, we briefly review here the concept of outer product.

If X is a $K \times 1$ column vector, the **outer product** of X with itself is the $K \times K$ matrix A obtained from the multiplication of X with its transpose:

$$A = XX^\top$$

Example 278 *If X is the 2×1 random vector*

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

then its outer product $XX^\top$ is the 2×2 random matrix

$$\begin{aligned} XX^\top &= \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \begin{bmatrix} X_1 & X_2 \end{bmatrix} \\ &= \begin{bmatrix} X_1^2 & X_1 X_2 \\ X_2 X_1 & X_2^2 \end{bmatrix} \end{aligned}$$

55.5.2 Symmetric matrices

A $K \times K$ matrix A is **symmetric** if and only if

$$A = A^\top$$

i.e. if and only if A equals its transpose.

55.5.3 Positive definite matrices

A $K \times K$ matrix A is said to be **positive definite** if and only if

$$x^\top Ax > 0$$

for any $K \times 1$ real vector x such that $x \neq 0$.

All positive definite matrices are also invertible.

Proof. The proof is by contradiction. Suppose a positive definite matrix A were not invertible. Then A would not be full rank, i.e. there would be a vector $x \neq 0$ such that

$$Ax = 0$$

which, premultiplied by $x^\top$, would yield

$$x^\top Ax = x^\top 0 = 0$$

But this is a contradiction. ■

[11] See p. 164.

55.5.4 Trace of a matrix

Let A be a $K \times K$ matrix and denote by A_{ij} the (i,j)-th entry of A (i.e. the entry at the intersection of the i-th row and the j-th column). The **trace** of A, denoted by $\operatorname{tr}(A)$, is the sum of all the diagonal entries of A:

$$\operatorname{tr}(A) = \sum_{i=1}^{K} A_{ii}$$

55.5.5 Vectorization of a matrix

Given a $K \times L$ matrix A, its **vectorization**, denoted by $\operatorname{vec}(A)$, is the $KL \times 1$ vector obtained by stacking the columns of A on top of each other.

Example 279 *If A is a 2×2 matrix:*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

the vectorization of A is the 4×1 random vector:

$$\operatorname{vec}(A) = \begin{bmatrix} A_{11} \\ A_{21} \\ A_{12} \\ A_{22} \end{bmatrix}$$

For a given matrix A, the vectorization of A will in general be different from the vectorization of its transpose $A^\top$. The **transposition permutation matrix** associated to $\operatorname{vec}(A)$ is the $KL \times KL$ matrix $P_{\operatorname{vec}(A)}$ such that:

$$\operatorname{vec}\left(A^\top\right) = P_{\operatorname{vec}(A)} \operatorname{vec}(A)$$

55.5.6 Kronecker product

Given a $K \times L$ matrix A and a $M \times N$ matrix B, the **Kronecker product** of A and B, denoted by $A \otimes B$, is a $KM \times LN$ matrix having the following structure:

$$A \otimes B = \begin{bmatrix} A_{11}B & A_{12}B & \dots & A_{1N}B \\ A_{21}B & A_{22}B & \dots & A_{2N}B \\ \vdots & & \ddots & \vdots \\ A_{M1}B & A_{M2}B & \dots & A_{MN}B \end{bmatrix}$$

where A_{ij} is the (i,j)-th entry of A.

Part V

More about normal distributions

Chapter 56

Linear combinations of normals

One property that makes the normal distribution extremely tractable from an analytical viewpoint is its closure under linear combinations: the linear combination of two independent random variables having a normal distribution also has a normal distribution. This lecture presents a multivariate generalization of this elementary property and then discusses some special cases.

56.1 Linear transformation of a MV-N vector

A linear transformation of a multivariate normal random vector[1] also has a multivariate normal distribution, as illustrated by the following:

Proposition 280 *Let X be a $K \times 1$ multivariate normal random vector with mean μ and covariance matrix V. Let A be an $L \times 1$ real vector and B an $L \times K$ full-rank real matrix. Then the $L \times 1$ random vector Y defined by*

$$Y = A + BX$$

has a multivariate normal distribution with mean

$$\mathrm{E}[Y] = A + B\mu$$

and covariance matrix

$$\mathrm{Var}[Y] = BVB^{\intercal}$$

Proof. This is proved using the formula for the joint moment generating function of the linear transformation of a random vector[2]. The joint moment generating function of X is

$$M_X(t) = \exp\left(t^\top \mu + \frac{1}{2} t^\top V t\right)$$

Therefore, the joint moment generating function of Y is

$$M_Y(t) = \exp\left(t^\top A\right) M_X\left(B^\top t\right)$$

[1] See p. 439.
[2] See p. 301.

$$\begin{aligned} &= \exp\left(t^\top A\right) \exp\left(t^\top B\mu + \frac{1}{2} t^\top BVB^\intercal t\right) \\ &= \exp\left(t^\top \left(A + B\mu\right) + \frac{1}{2} t^\top BVB^\intercal t\right) \end{aligned}$$

which is the moment generating function of a multivariate normal distribution with mean $A + B\mu$ and covariance matrix $BVB^\intercal$. Note that $BVB^\intercal$ needs to be positive definite in order to be the covariance matrix of a proper multivariate normal distribution, but this is implied by the assumption that B is full-rank. Therefore, Y has a multivariate normal distribution with mean $A + B\mu$ and covariance matrix $BVB^\intercal$, because two random vectors have the same distribution when they have the same joint moment generating function. ■

The following subsections present some important special cases of the above property.

56.1.1 Sum of two independent variables

The sum of two independent normal random variables has a normal distribution, as stated in the following:

Proposition 281 *Let X_1 be a random variable having a normal distribution with mean μ_1 and variance σ_1^2. Let X_2 be a random variable, independent of X_1, having a normal distribution with mean μ_2 and variance σ_2^2. Then, the random variable Y defined as:*

$$Y = X_1 + X_2$$

has a normal distribution with mean

$$\mathrm{E}\left[Y\right] = \mu_1 + \mu_2$$

and variance

$$\mathrm{Var}\left[Y\right] = \sigma_1^2 + \sigma_2^2$$

Proof. First of all, we need to use the fact that mutually independent normal random variables are jointly normal[3]: the 2×1 random vector X defined as

$$X = \begin{bmatrix} X_1 & X_2 \end{bmatrix}^\intercal$$

has a multivariate normal distribution with mean

$$\mathrm{E}\left[X\right] = \begin{bmatrix} \mu_1 & \mu_2 \end{bmatrix}^\intercal$$

and covariance matrix

$$\mathrm{Var}\left[X\right] = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

We can write:

$$Y = X_1 + X_2 = BX$$

where

$$B = \begin{bmatrix} 1 & 1 \end{bmatrix}$$

[3] See p. 446.

Therefore, according to Proposition 280, Y has a normal distribution with mean

$$\mathrm{E}[Y] = B\mathrm{E}[X] = \mu_1 + \mu_2$$

and variance

$$\mathrm{Var}[Y] = B\mathrm{Var}[X]B^{\intercal} = \sigma_1^2 + \sigma_2^2$$

■

56.1.2 Sum of more than two independent variables

The sum of more than two independent normal random variables also has a normal distribution, as shown in the following:

Proposition 282 *Let $X_1, \ldots, X_K$ be K mutually independent normal random variables, having means $\mu_1, \ldots, \mu_K$ and variances $\sigma_1^2, \ldots, \sigma_K^2$. Then, the random variable Y defined as:*

$$Y = \sum_{i=1}^{K} X_i$$

has a normal distribution with mean

$$\mathrm{E}[Y] = \sum_{i=1}^{K} \mu_i$$

and variance

$$\mathrm{Var}[Y] = \sum_{i=1}^{K} \sigma_i^2$$

Proof. This can be obtained, either generalizing the proof of Proposition 281, or using Proposition 281 recursively (starting from the first two components of X, then adding the third one and so on). ■

56.1.3 Linear combinations of independent variables

The properties illustrated in the previous two examples can be further generalized to linear combinations of K mutually independent normal random variables:

Proposition 283 *Let $X_1, \ldots, X_K$ be K mutually independent normal random variables, having means $\mu_1, \ldots, \mu_K$ and variances $\sigma_1^2, \ldots, \sigma_K^2$. Let $b_1, \ldots, b_K$ be K constants. Then, the random variable Y defined as:*

$$Y = \sum_{i=1}^{K} b_i X_i$$

has a normal distribution with mean

$$\mathrm{E}[Y] = \sum_{i=1}^{K} b_i \mu_i$$

and variance

$$\mathrm{Var}[Y] = \sum_{i=1}^{K} b_i^2 \sigma_i^2$$

Proof. First of all, we need to use the fact that mutually independent normal random variables are jointly normal: the $K \times 1$ random vector X defined as

$$X = [X_1 \quad \dots \quad X_K]^\intercal$$

has a multivariate normal distribution with mean

$$\mathrm{E}[X] = [\mu_1 \quad \dots \quad \mu_K]^\intercal$$

and covariance matrix

$$\mathrm{Var}[X] = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_K^2 \end{bmatrix}$$

We can write:

$$Y = \sum_{i=1}^{K} b_i X_i = BX$$

where

$$B = [b_1 \quad \dots \quad b_K]$$

Therefore, according to Proposition 280, Y has a (multivariate) normal distribution with mean:

$$\mathrm{E}[Y] = B\mathrm{E}[X] = \sum_{i=1}^{K} b_i \mu_i$$

and variance:

$$\mathrm{Var}[Y] = B\mathrm{Var}[X]\, B^\intercal = \sum_{i=1}^{K} b_i^2 \sigma_i^2$$

■

56.1.4 Linear transformation of a variable

A special case of Proposition 280 obtains when X has dimension 1×1 (i.e. it is a random variable):

Proposition 284 *Let X be a normal random variable with mean μ and variance σ^2. Let a and b be two constants (with $b \neq 0$). Then the random variable Y defined by:*

$$Y = a + bX$$

has a normal distribution with mean

$$\mathrm{E}[Y] = a + b\mu$$

and variance

$$\mathrm{Var}[Y] = b^2 \sigma^2$$

Proof. This is just a special case (for $K = 1$) of Proposition 280. ■

56.1.5 Linear combinations of independent vectors

The property illustrated in Example 3 can be generalized to linear combinations of mutually independent normal random vectors.

Proposition 285 *Let $X_1, \ldots, X_n$ be n mutually independent $K \times 1$ normal random vectors, having means $\mu_1, \ldots, \mu_n$ and covariance matrices $V_1, \ldots, V_n$. Let $B_1, \ldots, B_n$ be n real $L \times K$ full-rank matrices. Then, the $L \times 1$ random vector Y defined as:*

$$Y = \sum_{i=1}^{n} B_i X_i$$

has a normal distribution with mean

$$\mathrm{E}[Y] = \sum_{i=1}^{n} B_i \mu_i$$

and covariance matrix

$$\mathrm{Var}[Y] = \sum_{i=1}^{n} B_i V_i B_i^{\intercal}$$

Proof. This is a consequence of the fact that mutually independent normal random vectors are jointly normal: the $Kn \times 1$ random vector X defined as

$$X = [X_1^{\intercal} \quad \ldots \quad X_n^{\intercal}]^{\intercal}$$

has a multivariate normal distribution with mean

$$\mathrm{E}[X] = [\mu_1^{\intercal} \quad \ldots \quad \mu_n^{\intercal}]^{\intercal}$$

and covariance matrix

$$\mathrm{Var}[X] = \begin{bmatrix} V_1 & 0 & \ldots & 0 \\ 0 & V_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & V_n \end{bmatrix}$$

Therefore, we can apply Proposition 280 to the vector X. ■

56.2 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Let

$$X = [X_1 \quad X_2]^{\top}$$

be a 2×1 normal random vector with mean

$$\mu = [1 \quad 3]^{\top}$$

and covariance matrix

$$V = \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}$$

Find the distribution of the random variable Z defined as

$$Z = X_1 + 2X_2$$

Solution

We can write

$$Z = BX$$

where

$$B = [1 \quad 2]$$

Being a linear transformation of a multivariate normal random vector, Z is also multivariate normal. Actually, it is univariate normal, because it is a scalar. Its mean is

$$\mathrm{E}[Z] = B\mu = 1 \cdot 1 + 2 \cdot 3 = 7$$

and its variance is

$$\begin{aligned} \mathrm{Var}[Z] &= BVB^{\top} = [1 \quad 2] \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} \\ &= [1 \quad 2] \begin{bmatrix} 2 \cdot 1 + 1 \cdot 2 \\ 1 \cdot 1 + 3 \cdot 2 \end{bmatrix} = [1 \quad 2] \begin{bmatrix} 4 \\ 7 \end{bmatrix} \\ &= 1 \cdot 4 + 2 \cdot 7 = 18 \end{aligned}$$

Exercise 2

Let $X_1, \ldots, X_n$ be n mutually independent standard normal random variables. Let $b \in (0, 1)$ be a constant. Find the distribution of the random variable Y defined as

$$Y = \sum_{i=1}^{n} b^i X_i$$

Solution

Being a linear combination of mutually independent normal random variables, Y has a normal distribution with mean

$$\mathrm{E}[Y] = \sum_{i=1}^{n} b^i \mathrm{E}[X_i] = \sum_{i=1}^{n} b^i \cdot 0 = 0$$

and variance

$$\begin{aligned} \mathrm{Var}[Y] &= \sum_{i=1}^{n} \left(b^i\right)^2 \mathrm{Var}[X_i] \\ &= \sum_{i=1}^{n} \left(b^2\right)^i \cdot 1 \\ \boxed{\text{A}} &= \sum_{i=1}^{n} c^i \\ &= c + c^2 + \ldots + c^n \\ &= c\left(1 + c + \ldots + c^{n-1}\right) \\ &= c\left(1 + c + \ldots + c^{n-1}\right) \frac{1-c}{1-c} \\ &= \frac{c}{1-c}\left(1 - c^n\right) \end{aligned}$$

$$\begin{aligned} &= \frac{c - c^{n+1}}{1 - c} \\ &= \frac{b^2 - b^{2n+2}}{1 - b^2} \end{aligned}$$

where: in step $\boxed{\text{A}}$ we have defined $c = b^2$.

Chapter 57

Partitioned multivariate normal vectors

Let X be a $K \times 1$ multivariate normal random vector[1] with mean μ and covariance matrix V. In this lecture we present some useful facts about partitionings of X, that is, about subdivisions of X into two sub-vectors X_a and X_b such that

$$X = \begin{bmatrix} X_a \\ X_b \end{bmatrix}$$

where X_a and X_b have dimensions $K_a \times 1$ and $K_b \times 1$ respectively and $K_a + K_b = K$.

57.1 Notation

In what follows, we will denote by

- $\mu_a = \mathrm{E}[X_a]$ the mean of X_a;
- $\mu_b = \mathrm{E}[X_b]$ the mean of X_b;
- $V_a = \mathrm{Var}[X_a]$ the covariance matrix of X_a;
- $V_b = \mathrm{Var}[X_b]$ the covariance matrix of X_b;
- $V_{ab} = \mathrm{Cov}[X_a, X_b]$ the cross-covariance[2] between X_a and X_b.

Given this notation, it follows that

$$\mu = \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}$$

and

$$V = \begin{bmatrix} V_a & V_{ab}^{\intercal} \\ V_{ab} & V_b \end{bmatrix}$$

[1] See p. 439.
[2] See p. 193.

57.2 Normality of the sub-vectors

The following proposition states a first fact about the two sub-vectors.

Proposition 286 *Both X_a and X_b have a multivariate normal distribution, i.e.,*

$$\begin{aligned} X_a &\sim N(\mu_a, V_a) \\ X_b &\sim N(\mu_b, V_b) \end{aligned}$$

Proof. The random vector X_a can be written as a linear transformation of X:

$$X_a = AX$$

where A is a $K_a \times K$ matrix whose entries are either zero or one. Thus, X_a has a multivariate normal distribution, because it is a linear transformation of the multivariate normal random vector X, and multivariate normality is preserved[3] by linear transformations. By the same token, also X_b has a multivariate normal distribution, because it can be written as a linear transformation of X:

$$X_b = BX$$

where B is a $K_b \times K$ matrix whose entries are either zero or one. ■

This, of course, implies that any sub-vector of X is multivariate normal when X is multivariate normal.

57.3 Independence of the sub-vectors

The following proposition states a necessary and sufficient condition for the independence of the two sub-vectors.

Proposition 287 *X_a and X_b are independent if and only if $V_{ab} = 0$.*

Proof. X_a and X_b are independent if and only if their joint moment generating function is equal to the product of their individual moment generating functions[4]. Since X_a is multivariate normal, its joint moment generating function is

$$M_{X_a}(t_a) = \exp\left(t_a^\top \mu_a + \frac{1}{2} t_a^\top V_a t_a\right)$$

The joint moment generating function of X_b is

$$M_{X_b}(t_b) = \exp\left(t_b^\top \mu_b + \frac{1}{2} t_b^\top V_b t_b\right)$$

The joint moment generating function of X_a and X_b, which is just the joint moment generating function of X, is

$$\begin{aligned} & M_{X_a, X_b}(t_a, t_b) \\ = \; & M_X(t) \end{aligned}$$

[3] See p. 469.
[4] See p. 297.

$$
\begin{aligned}
&= \exp\left(t^\top \mu + \frac{1}{2} t^\top V t\right) \\
&= \exp\left(\begin{bmatrix} t_a^\top & t_b^\top \end{bmatrix} \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix} + \frac{1}{2} \begin{bmatrix} t_a^\top & t_b^\top \end{bmatrix} \begin{bmatrix} V_a & V_{ab}^\intercal \\ V_{ab} & V_b \end{bmatrix} \begin{bmatrix} t_a & t_b \end{bmatrix}\right) \\
&= \exp\left(t_a^\top \mu_a + t_b^\top \mu_b + \frac{1}{2} t_a^\top V_a t_a + \frac{1}{2} t_b^\top V_b t_b + \frac{1}{2} t_b^\top V_{ab} t_a + \frac{1}{2} t_a^\top V_{ab}^\top t_b\right) \\
&= \exp\left(t_a^\top \mu_a + \frac{1}{2} t_a^\top V_a t_a + t_b^\top \mu_b + \frac{1}{2} t_b^\top V_b t_b + t_b^\top V_{ab} t_a\right) \\
&= \exp\left(t_a^\top \mu_a + \frac{1}{2} t_a^\top V_a t_a\right) \exp\left(t_b^\top \mu_b + \frac{1}{2} t_b^\top V_b t_b\right) \exp\left(t_b^\top V_{ab} t_a\right) \\
&= M_{X_a}(t_a)\, M_{X_b}(t_b) \exp\left(t_b^\top V_{ab} t_a\right)
\end{aligned}
$$

from which it is obvious that

$$M_{X_a, X_b}(t_a, t_b) = M_{X_a}(t_a)\, M_{X_b}(t_b)$$

if and only if $V_{ab} = 0$. ■

Chapter 58

Quadratic forms in normal vectors

Let X be a $K \times 1$ multivariate normal random vector[1] with mean μ and covariance matrix V. In this lecture we present some useful facts about quadratic forms in X, i.e. about forms of the kind

$$Q = X^\top AX$$

where A is a $K \times K$ matrix and $\top$ denotes transposition.

Before discussing quadratic forms in X, we review some facts about matrix algebra that are needed to understand this lecture.

58.1 Review of relevant facts in matrix algebra

58.1.1 Orthogonal matrices

A $K \times K$ real matrix A is orthogonal if

$$A^\top = A^{-1}$$

which also implies

$$A^\top A = AA^\top = I$$

where I is the identity matrix. Of course, if A is orthogonal also $A^\top$ is orthogonal.

An important property of orthogonal matrices is the following:

Proposition 288 *Le X be a $K \times 1$ standard multivariate normal random vector, i.e. $X \sim N(0, I)$. Let A be an orthogonal $K \times K$ real matrix. Define*

$$Y = AX$$

Then also Y has a standard multivariate normal distribution, i.e. $Y \sim N(0, 1)$.

Proof. The random vector Y has a multivariate normal distribution, because it is a linear transformation of another multivariate normal random vector (see the

[1] See p. 439.

lecture entitled *Linear combinations of normal random variables* - p. 469). Y is standard normal because its expected value is

$$\mathrm{E}[Y] = \mathrm{E}[AX] = A\mathrm{E}[X] = A0 = 0$$

and its covariance matrix is

$$\begin{aligned} \mathrm{Var}[Y] &= \mathrm{Var}[AX] = A\mathrm{Var}[X]A^{\intercal} \\ &= AIA^{\intercal} = AA^{\intercal} = I \end{aligned}$$

where the definition of orthogonal matrix has been used in the last step. ■

58.1.2 Symmetric matrices

A $K \times K$ real matrix A is symmetric if

$$A = A^{\top}$$

i.e. A equals its transpose.

Real symmetric matrices have the property that they can be decomposed as

$$A = PDP^{\top}$$

where P is an othogonal matrix and D is a diagonal matrix (i.e. a matrix whose off-diagonal entries are zero). The diagonal elements of D, which are all real, are the eigenvalues of A and the columns of P are the eigenvectors of A.

58.1.3 Idempotent matrices

A $K \times K$ real matrix A is idempotent if

$$A^2 = A$$

which implies

$$A^n = A$$

for any $n \in \mathbb{N}$.

58.1.4 Symmetric idempotent matrices

If a matrix A is both symmetric and idempotent then its eigenvalues are either zero or one. In other words, the diagonal entries of the diagonal matrix D in the decomposition

$$A = PDP^{\top}$$

are either zero or one.

Proof. This can be easily seen as follows:

$$\begin{aligned} PDP^{\top} &= A = A^n = A \cdot \ldots \cdot A \\ &= PDP^{\top} \cdot \ldots \cdot PDP^{\top} \\ &= PD(P^{\top}P) \cdot \ldots \cdot (P^{\top}P)DP^{\top} \\ &= PD^nP^{\top} \end{aligned}$$

which implies

$$D = D^n \quad \forall n \in \mathbb{N}$$

But this is possible only if the diagonal entries of D are either zero or one. ■

58.1.5 Trace of a matrix

Let A be a $K \times K$ real matrix and denote by A_{ij} the (i,j)-th entry of A (i.e. the entry at the intersection of the i-th row and the j-th column). The trace of A, denoted by $\operatorname{tr}(A)$ is

$$\operatorname{tr}(A) = \sum_{i=1}^{K} A_{ii}$$

In other words, the trace is equal to the sum of all the diagonal entries of A. The trace of A enjoys the following important property:

$$\operatorname{tr}(A) = \sum_{i=1}^{K} \lambda_i$$

where $\lambda_1, \ldots, \lambda_K$ are the K eigenvalues of A.

58.2 Quadratic forms in normal vectors

The following proposition shows that certain quadratic forms in standard normal random vectors have a Chi-square distribution[2].

Proposition 289 *Le X be a $K \times 1$ standard multivariate normal random vector, i.e. $X \sim N(0, I)$. Let A be a symmetric and idempotent matrix. Let $\operatorname{tr}(A)$ be the trace of A. Define:*

$$Q = X^\top A X$$

Then Q has a Chi-square distribution with $\operatorname{tr}(A)$ degrees of freedom.

Proof. Since A is symmetric, it can be decomposed as

$$A = PDP^\top$$

where P is orthogonal and D is diagonal. The quadratic form can be written as

$$\begin{aligned} Q &= X^\top A X = X^\top P D P^\top X \\ &= \left(P^\top X\right)^\top D \left(P^\top X\right) = Y^\top D Y \end{aligned}$$

where we have defined

$$Y = P^\top X$$

By the above theorem on orthogonal transformations of standard multivariate normal random vectors, the orthogonality of $P^\top$ implies that $Y \sim N(0,1)$. Since D is diagonal, we can write the quadratic form as

$$Q = Y^\top D Y = \sum_{j=1}^{K} D_{jj} Y_j^2$$

where Y_j is the j-th component of Y and D_{jj} is the j-th diagonal entry of D. Since A is symmetric and idempotent, the diagonal entries of D are either zero or one. Denote by J the set

$$J = \{j \leq K : D_{jj} = 1\}$$

[2] See p. 387.

and by r its cardinality, i.e. the number of diagonal entries of D that are equal to 1. Since $D_{ij} \neq 1 \Rightarrow D_{ij} = 0$, we can write

$$Q = \sum_{j=1}^{K} D_{jj} Y_j^2 = \sum_{j \in J} D_{jj} Y_j^2 = \sum_{j \in J} Y_j^2$$

But the components of a standard normal random vector are mutually independent standard normal random variables. Therefore, Q is the sum of the squares of r independent standard normal random variables. Hence, it has a Chi-square distribution with r degrees of freedom[3]. Finally, by the properties of idempotent matrices and of the trace of a matrix (see above), r is not only the sum of the number of diagonal entries of D that are equal to 1, but it is also the sum of the eigenvalues of A. Since the trace of a matrix is equal to the sum of its eigenvalues, then $r = \operatorname{tr}(A)$. ■

58.3 Independence of quadratic forms

We start this section with a proposition on independence between linear transformations:

Proposition 290 *Le X be a $K \times 1$ standard multivariate normal random vector, i.e. $X \sim N(0, I)$. Let A be a $L_A \times K$ matrix and B be a $L_B \times K$ matrix. Define:*

$$\begin{aligned} T_1 &= AX \\ T_2 &= BX \end{aligned}$$

Then T_1 and T_2 are two independent random vectors if and only if $AB^\intercal = 0$.

Proof. First of all, note that T_1 and T_2 are linear transformations of the same multivariate normal random vector X. Therefore, they are jointly normal (see the lecture entitled *Linear combinations of normal random variables* - p. 469). Their cross-covariance[4] is

$$\begin{aligned} \operatorname{Cov}[X, Y] &= \operatorname{E}\left[(T_1 - \operatorname{E}[T_1])(T_2 - \operatorname{E}[T_2])^\top\right] \\ &= \operatorname{E}\left[(AX - \operatorname{E}[AX])(BX - \operatorname{E}[BX])^\top\right] \\ \boxed{\text{A}} &= A\operatorname{E}\left[(X - \operatorname{E}[X])(X - \operatorname{E}[X])^\top\right] B^\intercal \\ \boxed{\text{B}} &= A\operatorname{Var}[X] B^\intercal \\ \boxed{\text{C}} &= AB^\intercal \end{aligned}$$

where: in step A we have used the linearity of the expected value; in step B we have used the definition of covariance matrix; in step C we have used the fact that $\operatorname{Var}[X] = I$. But, as we explained in the lecture entitled *Partitioned multivariate normal vectors* (p. 477), two jointly normal random vectors are independent if and only if their cross-covariance is equal to 0. In our case, the cross-covariance is equal to zero if and only if $AB^\intercal = 0$, which proves the proposition. ■

[3] See p. 395.
[4] See p. 193.

The following proposition gives a necessary and sufficient condition for the independence of two quadratic forms in the same standard multivariate normal random vector.

Proposition 291 *Le X be a $K \times 1$ standard multivariate normal random vector, i.e. $X \sim N(0, I)$. Let A and B be two $K \times K$ symmetric and idempotent matrices. Define*

$$\begin{aligned} Q_1 &= X^\top A X \\ Q_2 &= X^\top B X \end{aligned}$$

Then Q_1 and Q_2 are two independent random variables if and only if $AB = 0$.

Proof. Since A and B are symmetric and idempotent, we can write

$$\begin{aligned} Q_1 &= (AX)^\top (AX) \\ Q_2 &= (BX)^\top (BX) \end{aligned}$$

from which it is apparent that Q_1 and Q_2 can be independent only as long as AX and BX are independent. But, by the above proposition on the independence between linear transformations of jointly normal random vectors, AX and BX are independent if and only if $AB^\intercal = 0$. Since B is symmetric, this is the same as $AB = 0$. ■

The following proposition gives a necessary and sufficient condition for the independence between a quadratic form and a linear transformation involving the same standard multivariate normal random vector.

Proposition 292 *Le X be a $K \times 1$ standard multivariate normal random vector, i.e. $X \sim N(0, I)$. Let A be a $L \times K$ vector and B a $K \times K$ symmetric and idempotent matrix. Define*

$$\begin{aligned} T &= AX \\ Q &= X^\top B X \end{aligned}$$

Then T and Q are independent if and only if $AB = 0$.

Proof. Since B is symmetric and idempotent, we can write

$$\begin{aligned} T &= AX \\ Q &= (BX)^\top (BX) \end{aligned}$$

from which it is apparent that T and Q can be independent only as long as AX and BX are independent. But, by the above proposition on the independence between linear transformations of jointly normal random vectors, AX and AB are independent if and only if $AB^\intercal = 0$. Since B is symmetric, this is the same as $AB = 0$. ■

58.4 Examples

We discuss here some quadratic forms that are commonly found in statistics.

Sample variance as a quadratic form

Let $X_1, \ldots, X_n$ be n independent random variables, all having a normal distribution with mean μ and variance σ^2. Let their sample mean[5] $\overline{X}_n$ be defined as

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and their adjusted sample variance[6] be defined as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$$

Define the following matrix:

$$M = I - \frac{1}{n}\imath\imath^\top$$

where I is the n-dimensional identity matrix and $\imath$ is a $n \times 1$ vector of ones. In other words, M has the following structure:

$$M = \begin{bmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & \cdots & -\frac{1}{n} \\ & & \ddots & \\ -\frac{1}{n} & -\frac{1}{n} & \cdots & 1-\frac{1}{n} \end{bmatrix}$$

M is a symmetric matrix. By computing the product $M \cdot M$, it can also be easily verified that M is idempotent.

Denote by X the $n \times 1$ random vector whose i-th entry is equal to X_i and note that X has a multivariate normal distribution with mean $\mu\imath$ and covariance matrix $\sigma^2 I$.

The matrix M can be used to write the sample variance as

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 \\ &= \frac{1}{n-1}\left(MX\right)^\top MX \\ &= \frac{1}{n-1}X^\top M^\top MX \\ \boxed{\text{A}} &= \frac{1}{n-1}X^\top MMX \\ \boxed{\text{B}} &= \frac{1}{n-1}X^\top MX \end{aligned}$$

where: in step A we have used the fact that M is symmetric; in step B we have used the fact that M is idempotent.

Now define a new random vector

$$Z = \frac{1}{\sigma}\left(X - \mu\imath\right)$$

[5] See p. 573.
[6] See p. 583.

and note that Z has a standard (mean zero and covariance I) multivariate normal distribution (see the lecture entitled *Linear combinations of normal random variables* - p. 469).

The sample variance can be written as

$$\begin{aligned} s^2 &= \frac{1}{n-1} X^\top M X \\ &= \frac{1}{n-1} \left(X - \mu\imath + \mu\imath\right)^\top M \left(X - \mu\imath + \mu\imath\right) \\ &= \frac{\sigma^2}{n-1} \left(\frac{X - \mu\imath}{\sigma} + \frac{\mu}{\sigma}\imath\right)^\top M \left(\frac{X - \mu\imath}{\sigma} + \frac{\mu}{\sigma}\imath\right) \\ &= \frac{\sigma^2}{n-1} \left(Z + \frac{\mu}{\sigma}\imath\right)^\top M \left(Z + \frac{\mu}{\sigma}\imath\right) \\ &= \frac{\sigma^2}{n-1} \left(Z^\top M Z + \frac{\mu}{\sigma} Z^\top M\imath + \frac{\mu}{\sigma}\imath^\top M Z + \left(\frac{\mu}{\sigma}\right)^2 \imath^\top M\imath\right) \end{aligned}$$

The last three terms in the sum are equal to zero, because

$$M\imath = 0$$

which can be verified by directly performing the multiplication of M and $\imath$.

Therefore, the sample variance

$$s^2 = \frac{\sigma^2}{n-1} Z^\top M Z$$

is proportional to a quadratic form in a standard normal random vector ($Z^\top M Z$) and the quadratic form is obtained from a symmetric and idempotent matrix (M). Thanks to the propositions above, we know that the quadratic form $Z^\top M Z$ has a Chi-square distribution with $\operatorname{tr}(M)$ degrees of freedom, where $\operatorname{tr}(M)$ is the trace of M. But the trace of M is

$$\begin{aligned} \operatorname{tr}(M) &= \sum_{i=1}^{n} M_{ii} = \sum_{i=1}^{n} \left(1 - \frac{1}{n}\right) \\ &= n\left(1 - \frac{1}{n}\right) = n - 1 \end{aligned}$$

So, the quadratic form $Z^\top M Z$ has a Chi-square distribution with $n-1$ degrees of freedom. Multiplying a Chi-square random variable with $n-1$ degrees of freedom by $\frac{\sigma^2}{n-1}$ one obtains a Gamma random variable with parameters $n-1$ and σ^2 (see the lecture entitled *Gamma distribution* - p. 403 - for more details).

So, summing up, the adjusted sample variance s^2 has a Gamma distribution with parameters $n-1$ and σ^2.

Futhermore, the adjusted sample variance s^2 is independent of the sample mean $\overline{X}_n$, which is proved as follows. The sample mean can be written as

$$\overline{X}_n = \frac{1}{n}\imath^\top X$$

and the sample variance can be written as

$$s^2 = \frac{1}{n-1} X^\top M X$$

Using Proposition 292, verifying the independence of $\overline{X}_n$ and s^2 boils down to verifying that

$$\left(\frac{1}{n}\imath^\top\right)\left(\frac{1}{n-1}M\right) = 0$$

which can be easily checked by directly performing the multiplication of $\imath^\top$ and M.

Part VI

Asymptotic theory

Chapter 59

Sequences of random variables

One of the central topics in probability theory and statistics is the study of sequences of random variables, i.e. of sequences[1] $\{X_n\}$ whose generic element X_n is a random variable.

There are several reasons why sequences of random variables are important:

1. Often, we need to analyze a random variable X, but for some reasons X is too complex to analyze directly. What we usually do in this case is to approximate X by simpler random variables X_n that are easier to study; these approximating random variables are arranged into a sequence $\{X_n\}$ and they become better and better approximations of X as n increases. This is exactly what we did when we introduced the Lebesgue integral[2].

2. In statistical theory, X_n is often an estimate of an unknown quantity whose value and whose properties depend on the sample size n (the sample size is the number of observations used to compute the estimate). Usually, we are able to analyze the properties of X_n only asymptotically (i.e. when n tends to infinity). In this case, $\{X_n\}$ is a sequence of estimates and we analyze the properties of the limit of $\{X_n\}$, in the hope that a large sample (the one we observe) and an infinite sample (the one we analyze by taking the limit of X_n) have a similar behavior.

3. In many applications a random variable is observed repeatedly through time (for example, the price of a stock is observed every day). In this case $\{X_n\}$ is the sequence of observations on the random variable and n is a time-index (in the stock price example, X_n is the price observed in the n-th period).

59.1 Terminology

In this lecture, we introduce some terminology related to sequences of random variables.

[1] See p. 31.
[2] See p. 141.

59.1.1 Realization of a sequence

Let $\{x_n\}$ be a sequence of real numbers and $\{X_n\}$ a sequence of random variables. If the real number x_n is a realization[3] of the random variable X_n for every n, then we say that the sequence of real numbers $\{x_n\}$ is a **realization of the sequence** of random variables $\{X_n\}$ and we write

$$\{X_n\} = \{x_n\}$$

59.1.2 Sequences on a sample space

Let Ω be a sample space[4]. Let $\{X_n\}$ be a sequence of random variables. We say that $\{X_n\}$ is a **sequence of random variables defined on the sample space** Ω if all the random variables X_n belonging to the sequence $\{X_n\}$ are functions from Ω to $\mathbb{R}$.

59.1.3 Independent sequences

Let $\{X_n\}$ be a sequence of random variables defined on a sample space Ω. We say that $\{X_n\}$ is an **independent sequence of random variables** (or a **sequence of independent random variables**) if every finite subset of $\{X_n\}$ (i.e. every finite set of random variables belonging to the sequence) is a set of mutually independent random variables[5].

59.1.4 Identically distributed sequences

Let $\{X_n\}$ be a sequence of random variables. Denote by $F_n(x)$ the distribution function[6] of a generic element of the sequence X_n. We say that $\{X_n\}$ is a **sequence of identically distributed random variables** if any two elements of the sequence have the same distribution function:

$$\forall x \in \mathbb{R}, \forall i, j \in \mathbb{N}, F_i(x) = F_j(x)$$

59.1.5 IID sequences

Let $\{X_n\}$ be a sequence of random variables defined on a sample space Ω. We say that $\{X_n\}$ is a **sequence of independent and identically distributed random variables** (or an **IID sequence of random variables**), if $\{X_n\}$ is both a sequence of independent random variables (see 59.1.3) and a sequence of identically distributed random variables (see 59.1.4).

59.1.6 Stationary sequences

Let $\{X_n\}$ be a sequence of random variables defined on a sample space Ω. Take a first group of q successive terms of the sequence $X_{n+1}, \ldots, X_{n+q}$. Now take a

[3] See p. 105.
[4] See p. 69.
[5] See p. 233.
[6] See p. 118.

second group of q successive terms of the sequence $X_{n+k+1}, \ldots, X_{n+k+q}$. The second group is located k positions after the first group. Denote the joint distribution function[7] of the first group of terms by

$$F_{n+1,\ldots,n+q}(x_1, \ldots, x_q)$$

and the joint distribution function of the second group of terms by

$$F_{n+k+1,\ldots,n+k+q}(x_1, \ldots, x_q)$$

The sequence $\{X_n\}$ is said to be **stationary** (or **strictly stationary**) if and only if

$$F_{n+1,\ldots,n+q}(x_1, \ldots, x_q) = F_{n+k+1,\ldots,n+k+q}(x_1, \ldots, x_q)$$

for any $n, k, q \in \mathbb{N}$ and for any vector $(x_1, \ldots, x_q) \in \mathbb{R}^q$.

In other words, a sequence is strictly stationary if and only if the two random vectors

$$[X_{n+1} \quad \ldots \quad X_{n+q}]$$

and

$$[X_{n+k+1} \quad \ldots \quad X_{n+k+q}]$$

have the same distribution for any n, k and q. Requiring strict stationarity is weaker than requiring that a sequence be IID (see 59.1.5): if $\{X_n\}$ is an IID sequence, then it is also strictly stationary, while the converse is not necessarily true.

59.1.7 Weakly stationary sequences

Let $\{X_n\}$ be a sequence of random variables defined on a sample space Ω. We say that $\{X_n\}$ is a **covariance stationary sequence** (or **weakly stationary sequence**) if

$$\exists \mu \in \mathbb{R} : \mathrm{E}[X_n] = \mu, \forall n > 0 \tag{1}$$
$$\forall j \geq 0, \exists \gamma_j \in \mathbb{R} : \mathrm{Cov}[X_n, X_{n-j}] = \gamma_j, \forall n > j \tag{2}$$

where n and j are, of course, integers.

Property (1) means that all the random variables belonging to the sequence $\{X_n\}$ have the same mean.

Property (2) means that the covariance between a term X_n of the sequence and the term that is located j positions before it (X_{n-j}) is always the same, irrespective of how X_n has been chosen. In other words, $\mathrm{Cov}[X_n, X_{n-j}]$ depends only on j and not on n. Property (2) also implies that all the random variables in the sequence have the same variance[8]:

$$\exists \gamma_0 \in \mathbb{R} : \mathrm{Var}[X_n] = \gamma_0, \forall n \in \mathbb{N}$$

Strictly stationarity (see 59.1.6) implies weak stationarity only if the mean $\mathrm{E}[X_n]$ and all the covariances $\mathrm{Cov}[X_n, X_{n-j}]$ exist and are finite.

Covariance stationarity does not imply strict stationarity, because the former imposes restrictions only on first and second moments, while the latter imposes restrictions on the whole distribution.

[7] See p. 118.

[8] Remember that $\mathrm{Cov}[X_n, X_n] = \mathrm{Var}[X_n]$.

59.1.8 Mixing sequences

Let $\{X_n\}$ be a sequence of random variables defined on a sample space Ω. Intuitively, $\{X_n\}$ is a mixing sequence if any two groups of terms of the sequence that are far apart from each other are approximately independent (and the further the closer to being independent).

Take a first group of q successive terms of the sequence $X_{n+1}, \ldots, X_{n+q}$. Now take a second group of q successive terms of the sequence $X_{n+k+1}, \ldots, X_{n+k+q}$. The second group is located k positions after the first group. The two groups of terms are independent if and only if

$$\begin{aligned} & \mathrm{E}\left[f\left(X_{n+1},\ldots,X_{n+q}\right) g\left(X_{n+k+1},\ldots,X_{n+k+q}\right)\right] \\ = \; & \mathrm{E}\left[f\left(X_{n+1},\ldots,X_{n+q}\right)\right] \mathrm{E}\left[g\left(X_{n+k+1},\ldots,X_{n+k+q}\right)\right] \end{aligned} \tag{59.1}$$

for any two functions f and g. This is just the definition of independence[9] between the two random vectors

$$\left[X_{n+1} \quad \ldots \quad X_{n+q}\right] \tag{59.2}$$

and

$$\left[X_{n+k+1} \quad \ldots \quad X_{n+k+q}\right] \tag{59.3}$$

Trivially, condition (59.1) can be written as

$$\begin{aligned} & \mathrm{E}\left[f\left(X_{n+1},\ldots,X_{n+q}\right) g\left(X_{n+k+1},\ldots,X_{n+k+q}\right)\right] \\ & -\mathrm{E}\left[f\left(X_{n+1},\ldots,X_{n+q}\right)\right] \mathrm{E}\left[g\left(X_{n+k+1},\ldots,X_{n+k+q}\right)\right] = 0 \end{aligned}$$

If this condition is true asymptotically (i.e. when $k \to \infty$), then we say that the sequence $\{X_n\}$ is mixing:

Definition 293 *We say that a sequence of random variables $\{X_n\}$ is **mixing** (or **strongly mixing**) if and only if*

$$\begin{aligned} \lim_{k\to\infty} \; & \{\mathrm{E}\left[f\left(X_{n+1},\ldots,X_{n+q}\right) g\left(X_{n+k+1},\ldots,X_{n+k+q}\right)\right] \\ & - \mathrm{E}\left[f\left(X_{n+1},\ldots,X_{n+q}\right)\right] \mathrm{E}\left[g\left(X_{n+k+1},\ldots,X_{n+k+q}\right)\right]\} = 0 \end{aligned}$$

for any two functions f and g and for any n and q.

In other words, a sequence is strongly mixing if and only if the two random vectors (59.2) and (59.3) tend to become more and more independent by increasing k (for any n and q). This is a milder requirement than the requirement of independence: if $\{X_n\}$ is an independent sequence (see 59.1.3), all its terms are independent from one another; if $\{X_n\}$ is a mixing sequence, its terms can be dependent, but they become less and less dependent as the distance between their locations in the sequence increases. Of course, an independent sequence is also a mixing sequence, while the converse is not necessarily true.

59.1.9 Ergodic sequences

In this section we discuss ergodicity. Roughly speaking, ergodicity is a weak concept of independence for sequences of random variables.

In the subsections above we have discussed other two concepts of independence for sequences of random variables:

[9] See p. 235.

1. independent sequences are sequences of random variables whose terms are mutually independent;

2. mixing sequences are sequences of random variables whose terms can be dependent but become less and less dependent as their distance increases (by distance we mean how far apart they are located in the sequence).

Requiring that a sequence be mixing is weaker than requiring that a sequence be independent: in fact, an independent sequence is also mixing, but the converse is not true.

Requiring that a sequence be ergodic is even weaker than requiring that a sequence be mixing (mixing implies ergodicity but not vice versa). This is probably all you need to know if you are not studying asymptotic theory at an advanced level, because ergodicity is quite a complicated topic and the definition of ergodicity is fairly abstract. Nevertheless, we give here a quick definition of ergodicity for the sake of completeness.

Denote by $\mathbb{R}^{\mathbb{N}}$ the set of all possible sequences of real numbers. When $\{x_n\}$ is a sequence of real numbers, denote by $\{x_n\}_{n>1}$ the subsequence obtained by dropping the first term of $\{x_n\}$, i.e.

$$\{x_n\}_{n>1} = \{x_2, x_3, \ldots\}$$

We say that a subset $A \subseteq \mathbb{R}^{\mathbb{N}}$ is a **shift invariant** set if and only if $\{x_n\}_{n>1}$ belongs to A whenever $\{x_n\}$ belongs to A. In symbols:

Definition 294 *A set $A \subseteq \mathbb{R}^{\mathbb{N}}$ is shift invariant if and only if*

$$\{x_n\} \in A \Longrightarrow \{x_n\}_{n>1} \in A$$

Shift invariance is used to define ergodicity:

Definition 295 *A sequence of random variables $\{X_n\}$ is said to be an **ergodic sequence** if an olny if*

$$\textit{either } \mathrm{P}(\{X_n\} \in A) = 0 \textit{ or } \mathrm{P}(\{X_n\} \in A) = 1$$

whenever A is a shift invariant set.

59.2 Limit of a sequence of random variables

As we explained in the lecture entitled *Sequences and limits* (p. 31), whenever we want to assess whether a sequence is convergent (and find its limit), we need to define a distance function (or metric) to measure the distance between the terms of the sequence. Intuitively, a sequence converges to a limit if, by dropping a sufficiently high number of initial terms of the sequence, the remaining terms can be made as close to each other as we wish. The problem is how to define "close to each other". As we have explained, the concept of "close to each other" can be made fully rigorous using the notion of a metric. Therefore, discussing convergence of a sequence of random variables boils down to discussing what metrics can be used to measure the distance between two random variables.

In the following lectures, we introduce several different notions of convergence of a sequence of random variables: to each different notion corresponds a different way of measuring the distance between two random variables.

The notions of convergence (also called **modes of convergence**) introduced in the following lectures are:

1. Pointwise convergence (p. 501).
2. Almost sure convergence (p. 505).
3. Convergence in probability (p. 511).
4. Mean-square convergence (p. 519).
5. Convergence in distribution (p. 527).

Chapter 60

Sequences of random vectors

In this lecture, we generalize the concepts introduced in the lecture entitled *Sequences of random variables* (p. 491). We no longer consider sequences whose elements are random variables, but we now consider sequences $\{X_n\}$ whose generic element X_n is a $K \times 1$ random vector. The generalization is straightforward, as the terminology and the basic concepts are almost the same used for sequences of random variables.

60.1 Terminology

60.1.1 Realization of a sequence

Let $\{x_n\}$ be a sequence of $K \times 1$ real vectors and $\{X_n\}$ a sequence of $K \times 1$ random vectors. If the real vector x_n is a realization[1] of the random vector X_n for every n, then we say that the sequence of real vectors $\{x_n\}$ is a **realization of the sequence** of random vectors $\{X_n\}$ and we write

$$\{X_n\} = \{x_n\}$$

60.1.2 Sequences on a sample space

Let Ω be a sample space[2]. Let $\{X_n\}$ be a sequence of random vectors. We say that $\{X_n\}$ is a **sequence of random vectors defined on the sample space** Ω if all the random vectors X_n belonging to the sequence $\{X_n\}$ are functions from Ω to $\mathbb{R}^K$.

60.1.3 Independent sequences

Let $\{X_n\}$ be a sequence of random vectors defined on a sample space Ω. We say that $\{X_n\}$ is an **independent sequence of random vectors** (or a **sequence of independent random vectors**) if every finite subset of $\{X_n\}$ (i.e. every finite subset of random vectors belonging to the sequence) is a set of mutually independent random vectors[3].

[1] See p. 105.
[2] See p. 69.
[3] See p. 235.

60.1.4 Identically distributed sequences

Let $\{X_n\}$ be a sequence of random vectors. Denote by $F_n(x)$ the joint distribution function[4] of a generic element of the sequence X_n. We say that $\{X_n\}$ is a **sequence of identically distributed random vectors** if any two elements of the sequence have the same joint distribution function:

$$\forall x \in \mathbb{R}^K, \forall i, j \in \mathbb{N},\ F_i(x) = F_j(x)$$

60.1.5 IID sequences

Let $\{X_n\}$ be a sequence of random vectors defined on a sample space Ω. We say that $\{X_n\}$ is a **sequence of independent and identically distributed random vectors** (or an **IID sequence of random vectors**), if $\{X_n\}$ is both a sequence of independent random vectors (see above) and a sequence of identically distributed random vectors (see above).

60.1.6 Stationary sequences

Let $\{X_n\}$ be a sequence of random vectors defined on a sample space Ω. Take a first group of q successive terms of the sequence $X_{n+1}, \ldots, X_{n+q}$. Now take a second group of q successive terms of the sequence $X_{n+k+1}, \ldots, X_{n+k+q}$. The second group is located k positions after the first group. Denote the joint distribution function of the first group of terms by

$$F_{n+1,\ldots,n+q}(x_1, \ldots, x_q)$$

and the joint distribution function of the second group of terms by

$$F_{n+k+1,\ldots,n+k+q}(x_1, \ldots, x_q)$$

The sequence $\{X_n\}$ is said to be **stationary** (or **strictly stationary**) if and only if

$$F_{n+1,\ldots,n+q}(x_1, \ldots, x_q) = F_{n+k+1,\ldots,n+k+q}(x_1, \ldots, x_q)$$

for any $n, k, q \in \mathbb{N}$ and for any vector $\begin{bmatrix} x_1^\top & \ldots & x_q^\top \end{bmatrix}^\top \in \mathbb{R}^{Kq}$.

In other words, a sequence is strictly stationary if and only if the two random vectors

$$\begin{bmatrix} X_{n+1}^\top & \cdots & X_{n+q}^\top \end{bmatrix}^\top$$

and

$$\begin{bmatrix} X_{n+k+1}^\top & \cdots & X_{n+k+q}^\top \end{bmatrix}^\top$$

have the same distribution (for any n, k and q). Requiring strict stationarity is weaker than requiring that a sequence be IID (see the subsection *IID sequences* above): if $\{X_n\}$ is an IID sequence, then it is also strictly stationary, while the converse is not necessarily true.

[4] See p. 118.

60.1.7 Weakly stationary sequences

Let $\{X_n\}$ be a sequence of random vectors defined on a sample space Ω. We say that $\{X_n\}$ is a **covariance stationary sequence** (or **weakly stationary sequence**) if

$$\exists \mu \in \mathbb{R}^K : \mathrm{E}[X_n] = \mu, \forall n \in \mathbb{N} \quad (1)$$
$$\forall j \geq 0, \exists \Gamma_j \in \mathbb{R}^{K \times K} : \mathrm{Cov}[X_n, X_{n-j}] = \Gamma_j, \forall n > j \quad (2)$$

where n and j are, of course, integers. Property (1) means that all the random vectors belonging to the sequence $\{X_n\}$ have the same mean. Property (2) means that the cross-covariance[5] between a term X_n of the sequence and the term that is located j positions before it (X_{n-j}) is always the same, irrespective of how X_n has been chosen. In other words, $\mathrm{Cov}[X_n, X_{n-j}]$ depends only on j and not on n. Note also that property (2) implies that all the random vectors in the sequence have the same covariance matrix (because $\mathrm{Cov}[X_n, X_n] = \mathrm{Var}[X_n]$):

$$\exists \Gamma_0 \in \mathbb{R}^{K \times K} : \mathrm{Var}[X_n] = \Gamma_0, \forall n \in \mathbb{N}$$

60.1.8 Mixing sequences

The definition of mixing sequence of random vectors is a straightforward generalization of the definition of mixing sequence of random variables, which has been discussed in the lecture entitled *Sequences of random variables* (p. 491). Therefore, we report here the definition of mixing sequence of random vectors without further comments and we refer the reader to the aforementioned lecture for an explanation of the concept of mixing sequence.

Definition 296 *We say that a sequence of random vectors $\{X_n\}$ is **mixing** (or **strongly mixing**) if and only if*

$$\lim_{k \to \infty} \{\mathrm{E}[f(X_{n+1}, \ldots, X_{n+q}) g(X_{n+k+1}, \ldots, X_{n+k+q})] - \mathrm{E}[f(X_{n+1}, \ldots, X_{n+q})] \mathrm{E}[g(X_{n+k+1}, \ldots, X_{n+k+q})]\} = 0$$

for any two functions f and g and for any n and q.

60.1.9 Ergodic sequences

As in the previous section, we report here a definition of ergodic sequence of random vectors, which is a straightforward generalization of the definition of ergodic sequence of random variables, and we refer the reader to the lecture entitled *Sequences of random variables* (p. 491) for explanations of the concept of ergodicity.

Denote by $\left(\mathbb{R}^K\right)^{\mathbb{N}}$ the set of all possible sequences of real $K \times 1$ vectors. When $\{x_n\}$ is a sequence of real vectors, denote by $\{x_n\}_{n>1}$ the subsequence obtained by dropping the first term of $\{x_n\}$, i.e.:

$$\{x_n\}_{n>1} = \{x_2, x_3, \ldots\}$$

We say that a subset $A \subseteq \left(\mathbb{R}^K\right)^{\mathbb{N}}$ is a **shift invariant** set if and only if $\{x_n\}_{n>1}$ belongs to A whenever $\{x_n\}$ belongs to A. In symbols:

[5] See p. 193.

Definition 297 *A set* $A \subseteq \left(\mathbb{R}^K\right)^{\mathbb{N}}$ *is shift invariant if and only if*

$$\{x_n\} \in A \Longrightarrow \{x_n\}_{n>1} \in A$$

Shift invariance is used to define ergodicity.

Definition 298 *A sequence of random vectors* $\{X_n\}$ *is said to be an* ***ergodic sequence*** *if an olny if*

$$\textit{either } \mathrm{P}\left(\{X_n\} \in A\right) = 0 \textit{ or } \mathrm{P}\left(\{X_n\} \in A\right) = 1$$

whenever A *is a shift invariant set.*

60.2 Limit of a sequence of random vectors

Similarly to what happens for sequences of random variables, there are several different notions of convergence also for sequences of random vectors. In particular, all the modes of convergence found for random variables can be generalized to random vectors:

1. Pointwise convergence (p. 501).
2. Almost sure convergence (p. 505).
3. Convergence in probability (p. 511).
4. Mean-square convergence (p. 519).
5. Convergence in distribution (p. 527).

Chapter 61

Pointwise convergence

This lecture discusses pointwise convergence. We deal first with pointwise convergence of sequences of random variables and then with pointwise convergence of sequences of random vectors.

61.1 Sequences of random variables

Let $\{X_n\}$ be a sequence of random variables defined on a sample space[1] Ω. Let us consider a single sample point[2] $\omega \in \Omega$ and a generic random variable X_n belonging to the sequence. X_n is a function $X_n : \Omega \to \mathbb{R}$. However, once we fix ω, the realization $X_n(\omega)$ associated to the sample point ω is just a real number. By the same token, once we fix ω, the sequence $\{X_n(\omega)\}$ is just a sequence of real numbers[3]. Therefore, for a fixed ω, it is very easy to assess whether the sequence $\{X_n(\omega)\}$ is convergent; this is done employing the usual definition of convergence of a sequence of real numbers. If, for a fixed ω, the sequence $\{X_n(\omega)\}$ is convergent, we denote its limit by $X(\omega)$, to underline that the limit depends on the specific ω we have fixed. A sequence of random variables is said to be pointwise convergent if and only if the sequence $\{X_n(\omega)\}$ is convergent for any choice of ω:

Definition 299 *Let $\{X_n\}$ be a sequence of random variables defined on a sample space Ω. We say that $\{X_n\}$ is **pointwise convergent** to a random variable X defined on Ω if and only if $\{X_n(\omega)\}$ converges to $X(\omega)$ for all $\omega \in \Omega$. X is called the **pointwise limit** of the sequence and convergence is indicated by:*

$$X_n \to X \text{ pointwise}$$

Roughly speaking, using pointwise convergence we somehow circumvent the problem of defining the concept of distance between random variables: by fixing ω, we reduce ourselves to the familiar problem of measuring distance between two real numbers, so that we can employ the usual notion of convergence of sequences of real numbers.

Example 300 *Let $\Omega = \{\omega_1, \omega_2\}$ be a sample space with two sample points (ω_1 and ω_2). Let $\{X_n\}$ be a sequence of random variables such that a generic term X_n of*

[1] See p. 492.
[2] See p. 69.
[3] See p. 33.

the sequence satisfies:

$$X_n(\omega) = \begin{cases} \frac{1}{n} & \text{if } \omega = \omega_1 \\ 1 + \frac{2}{n} & \text{if } \omega = \omega_2 \end{cases}$$

We need to check the convergence of the sequences $\{X_n(\omega)\}$ *for all* $\omega \in \Omega$*, i.e. for* $\omega = \omega_1$ *and for* $\omega = \omega_2$*: (1) the sequence* $\{X_n(\omega_1)\}$*, whose generic term is*

$$X_n(\omega_1) = \frac{1}{n}$$

is a sequence of real numbers converging to 0*; (2) the sequence* $\{X_n(\omega_2)\}$*, whose generic term is*

$$X_n(\omega_2) = 1 + \frac{2}{n}$$

is a sequence of real numbers converging to 1*. Therefore, the sequence of random variables* $\{X_n\}$ *converges pointwise to the random variable* X*, where* X *is defined as follows:*

$$X(\omega) = \begin{cases} 0 & \text{if } \omega = \omega_1 \\ 1 & \text{if } \omega = \omega_2 \end{cases}$$

61.2 Sequences of random vectors

The above notion of convergence generalizes to sequences of random vectors in a straightforward manner.

Let $\{X_n\}$ be a sequence of random vectors defined on a sample space[4] Ω, where each random vector X_n has dimension $K \times 1$. If we fix a single sample point $\omega \in \Omega$, the sequence $\{X_n(\omega)\}$ is a sequence of real $K \times 1$ vectors. By the standard criterion for convergence[5], the sequence of real vectors $\{X_n(\omega)\}$ is convergent to a vector $X(\omega)$ if

$$\lim_{n \to \infty} d(X_n(\omega), X(\omega)) = 0$$

where $d(X_n(\omega), X(\omega))$ is the distance between a generic term of the sequence $X_n(\omega)$ and the limit $X(\omega)$. The distance between $X_n(\omega)$ and $X(\omega)$ is defined to be equal to the Euclidean norm of their difference:

$$\begin{aligned} & d(X_n(\omega), X(\omega)) \\ = \; & \|X_n(\omega) - X(\omega)\| \\ = \; & \sqrt{[X_{n,1}(\omega) - X_{\bullet,1}(\omega)]^2 + \ldots + [X_{n,K}(\omega) - X_{\bullet,K}(\omega)]^2} \end{aligned}$$

where the second subscript is used to indicate the individual components of the vectors $X_n(\omega)$ and $X(\omega)$. Thus, for a fixed ω, the sequence of real vectors $\{X_n(\omega)\}$ is convergent to a vector $X(\omega)$ if

$$\lim_{n \to \infty} \|X_n(\omega) - X(\omega)\| = 0$$

A sequence of random vectors $\{X_n\}$ is said to be pointwise convergent if and only if the sequence $\{X_n(\omega)\}$ is convergent for any choice of ω:

[4] See p. 497.
[5] See p. 36.

Definition 301 *Let $\{X_n\}$ be a sequence of random vectors defined on a sample space Ω. We say that $\{X_n\}$ is **pointwise convergent** to a random vector X defined on Ω if and only if $\{X_n(\omega)\}$ converges to $X(\omega)$ for all $\omega \in \Omega$, i.e.*

$$\forall \omega, \lim_{n\to\infty} \|X_n(\omega) - X(\omega)\| = 0$$

*X is called the **pointwise limit** of the sequence and convergence is indicated by:*

$$X_n \to X \text{ pointwise}$$

Now, denote by $\{X_{n,i}\}$ the sequence of the i-th components of the vectors X_n. It can be proved that the sequence of random vectors $\{X_n\}$ is pointwise convergent if and only if all the K sequences of random variables $\{X_{n,i}\}$ are pointwise convergent:

Proposition 302 *Let $\{X_n\}$ be a sequence of random vectors defined on a sample space Ω. Denote by $\{X_{n,i}\}$ the sequence of random variables obtained by taking the i-th component of each random vector X_n. The sequence $\{X_n\}$ converges pointwise to the random vector X if and only if $\{X_{n,i}\}$ converges pointwise to the random variable $X_{\bullet,i}$ (the i-th component of X) for each $i = 1, \ldots, K$.*

61.3 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Let the sample space Ω be[6]:

$$\Omega = [0, 1]$$

Define a sequence of random variables $\{X_n\}$ as follows:

$$X_n(\omega) = \frac{\omega}{2n} \qquad \forall \omega \in \Omega$$

Find the pointwise limit of the sequence $\{X_n\}$.

Solution

For a fixed sample point ω, the sequence of real numbers $\{X_n(\omega)\}$ has limit:

$$\lim_{n\to\infty} X_n(\omega) = \lim_{n\to\infty} \frac{\omega}{2n} = 0$$

Therefore, the sequence of random variables $\{X_n\}$ converges pointwise to the random variable X defined as follows:

$$X(\omega) = 0 \qquad \forall \omega \in \Omega$$

[6] In other words, the sample space Ω is the set of all real numbers between 0 and 1.

Exercise 2

Suppose the sample space Ω is as in the previous exercise:

$$\Omega = [0,1]$$

Define a sequence of random variables $\{X_n\}$ as follows:

$$X_n(\omega) = \left(1 + \frac{\omega}{n}\right)^n \qquad \forall \omega \in \Omega$$

Find the pointwise limit of the sequence $\{X_n\}$.

Solution

For a given sample point ω, the sequence of real numbers $\{X_n(\omega)\}$ has limit[7]:

$$\lim_{n\to\infty} X_n(\omega) = \lim_{n\to\infty} \left(1 + \frac{\omega}{n}\right)^n = \exp(\omega)$$

Thus, the sequence of random variables $\{X_n\}$ converges pointwise to the random variable X defined as follows:

$$X(\omega) = \exp(\omega) \qquad \forall \omega \in \Omega$$

Exercise 3

Suppose the sample space Ω is as in the previous exercises:

$$\Omega = [0,1]$$

Define a sequence of random variables $\{X_n\}$ as follows:

$$X_n(\omega) = \omega^n \qquad \forall \omega \in \Omega$$

Define a random variable X as follows:

$$X(\omega) = 0 \qquad \forall \omega \in \Omega$$

Does the sequence $\{X_n\}$ converge pointwise to the random variable X?

Solution

For $\omega \in [0,1)$, the sequence of real numbers $\{X_n(\omega)\}$ has limit:

$$\lim_{n\to\infty} X_n(\omega) = \lim_{n\to\infty} \omega^n = 0$$

However, for $\omega = 1$, the sequence of real numbers $\{X_n(\omega)\}$ has limit:

$$\lim_{n\to\infty} X_n(\omega) = \lim_{n\to\infty} 1^n = 1$$

Thus, the sequence of random variables $\{X_n\}$ does not converge pointwise to the random variable X, but it converges pointwise to the random variable Y defined as follows:

$$Y(\omega) = \begin{cases} 0 & \text{if } \omega \in [0,1) \\ 1 & \text{if } \omega = 1 \end{cases}$$

[7] Note that this limit is encountered very frequently and you can find a proof of it in most calculus textbooks.

Chapter 62

Almost sure convergence

This lecture introduces the concept of almost sure convergence. In order to understand this lecture, you should first understand the concepts of almost sure property and almost sure event[1] and the concept of pointwise convergence of a sequence of random variables[2].

We deal first with almost sure convergence of sequences of random variables and then with almost sure convergence of sequences of random vectors.

62.1 Sequences of random variables

Let $\{X_n\}$ be a sequence of random variables defined on a sample space[3] Ω. The concept of **almost sure convergence** (or **a.s. convergence**) is a slight variation of the concept of pointwise convergence. As we have seen, a sequence of random variables $\{X_n\}$ is pointwise convergent if and only if the sequence of real numbers $\{X_n(\omega)\}$ is convergent for all $\omega \in \Omega$. Achieving convergence for all $\omega \in \Omega$ is a very stringent requirement. Therefore, this requirement is usually weakened, by requiring the convergence of $\{X_n(\omega)\}$ for a large enough subset of Ω, and not necessarily for all $\omega \in \Omega$. In particular, $\{X_n(\omega)\}$ is usually required to be a convergent sequence almost surely: if F is the set of all sample points ω for which the sequence $\{X_n(\omega)\}$ is convergent, its complement F^c must be included in a zero-probability event:

$$F = \{\omega \in \Omega : \{X_n(\omega)\} \text{ is a convergent sequence}\}$$
$$E \text{ is a zero-probability event}$$
$$F^c \subseteq E$$

In other words, almost sure convergence requires that the sequences $\{X_n(\omega)\}$ converge for all sample points $\omega \in \Omega$, except, possibly, for a very small set F^c of sample points (F^c must be included in a zero-probability event).

Definition 303 *Let $\{X_n\}$ be a sequence of random variables defined on a sample space Ω. We say that $\{X_n\}$ is* ***almost surely convergent (a.s. convergent)*** *to a random variable X defined on Ω if and only if the sequence of real numbers*

[1] See the lecture entitled *Zero-probability events* (p. 79).
[2] See p. 501.
[3] See p. 492.

$\{X_n(\omega)\}$ converges to $X(\omega)$ almost surely, i.e. if there exists a zero-probability event E such that:

$$\{\omega \in \Omega : \{X_n(\omega)\} \text{ does not converge to } X(\omega)\} \subseteq E$$

*X is called the **almost sure limit** of the sequence and convergence is indicated by:*

$$X_n \overset{a.s.}{\to} X$$

The following is an example of a sequence that converges almost surely:

Example 304 *Suppose the sample space Ω is:*

$$\Omega = [0,1]$$

It is possible to build a probability measure P on Ω, such that P assigns to each sub-interval of $[0,1]$ a probability equal to its length[4]:

$$\text{if } 0 \leq a \leq b \leq 1 \text{ and } E = [a,b], \text{ then } \mathrm{P}(E) = b - a$$

Remember that in this probability model all the sample points $\omega \in \Omega$ are assigned zero probability (each sample point, when considered as an event, is a zero-probability event):

$$\forall \omega \in \Omega,\ \mathrm{P}(\{\omega\}) = \mathrm{P}([\omega, \omega]) = \omega - \omega = 0$$

Now, consider a sequence of random variables $\{X_n\}$ defined as follows:

$$X_n(\omega) = \begin{cases} 1 & \text{if } \omega = 0 \\ \frac{1}{n} & \text{if } \omega \neq 0 \end{cases}$$

When $\omega \in (0,1]$, the sequence of real numbers $\{X_n(\omega)\}$ converges to 0, because:

$$\lim_{n\to\infty} X_n(\omega) = \lim_{n\to\infty} \frac{1}{n} = 0$$

However, when $\omega = 0$, the sequence of real numbers $\{X_n(\omega)\}$ is not convergent to 0, because:

$$\lim_{n\to\infty} X_n(\omega) = \lim_{n\to\infty} 1 = 1$$

Define a constant random variable X:

$$X(\omega) = 0, \forall \omega \in [0,1]$$

We have that:

$$\{\omega \in \Omega : \{X_n(\omega)\} \text{ does not converge to } X(\omega)\} = \{0\}$$

But $\mathrm{P}(\{0\}) = 0$ because:

$$\mathrm{P}(\{0\}) = \mathrm{P}([0,\ 0]) = 0 - 0 = 0$$

which means that the event

$$\{\omega \in \Omega : \{X_n(\omega)\} \text{ does not converge to } X(\omega)\}$$

is a zero-probability event. Therefore, the sequence $\{X_n\}$ converges to X almost surely, but it does not converge pointwise to X because $\{X_n(\omega)\}$ does not converge to $X(\omega)$ for all $\omega \in \Omega$.

[4] See the lecture entitled *Zero-probability events* (p. 79).

62.2 Sequences of random vectors

The above notion of convergence generalizes to sequences of random vectors in a straightforward manner.

Let $\{X_n\}$ be a sequence of random vectors defined on a sample space[5] Ω, where each random vector X_n has dimension $K \times 1$. Also in the case of random vectors, the concept of almost sure convergence is obtained from the concept of pointwise convergence by relaxing the assumption that the sequence $\{X_n(\omega)\}$ converges for all $\omega \in \Omega$ (remember that the sequence of real vectors $\{X_n(\omega)\}$ converges to a real vector $X(\omega)$ if and only if $\lim_{n\to\infty} \|X_n(\omega) - X(\omega)\| = 0$). Instead, it is required that the sequence $\{X_n(\omega)\}$ converges for almost all ω (i.e. almost surely).

Definition 305 *Let $\{X_n\}$ be a sequence of random vectors defined on a sample space Ω. We say that $\{X_n\}$ is **almost surely convergent** to a random vector X defined on Ω if and only if the sequence of real vectors $\{X_n(\omega)\}$ converges to the real vector $X(\omega)$ almost surely, i.e. if there exists a zero-probability event E such that:*

$$\{\omega \in \Omega : \{X_n(\omega)\} \text{ does not converge to } X(\omega)\} \subseteq E$$

*X is called the **almost sure limit** of the sequence and convergence is indicated by:*

$$X_n \overset{a.s.}{\to} X$$

Now, denote by $\{X_{n,i}\}$ the sequence of the i-th components of the vectors X_n. It can be proved that the sequence of random vectors $\{X_n\}$ is almost surely convergent if and only if all the K sequences of random variables $\{X_{n,i}\}$ are almost surely convergent.

Proposition 306 *Let $\{X_n\}$ be a sequence of random vectors defined on a sample space Ω. Denote by $\{X_{n,i}\}$ the sequence of random variables obtained by taking the i-th component of each random vector X_n. The sequence $\{X_n\}$ converges almost surely to the random vector X if and only if $\{X_{n,i}\}$ converges almost surely to the random variable $X_{\bullet,i}$ (the i-th component of X) for each $i = 1, \ldots, K$.*

62.3 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Let the sample space Ω be:

$$\Omega = [0, 1]$$

Sub-intervals of $[0, 1]$ are assigned a probability equal to their length:

$$\text{if } 0 \leq a \leq b \leq 1 \text{ and } E = [a, b], \text{ then } \mathrm{P}(E) = b - a$$

Define a sequence of random variables $\{X_n\}$ as follows:

$$X_n(\omega) = \omega^n \qquad \forall \omega \in \Omega$$

[5] See p. 497.

Define a random variable X as follows:

$$X(\omega) = 0 \qquad \forall \omega \in \Omega$$

Does the sequence $\{X_n\}$ converge almost surely to X?

Solution

For a fixed sample point $\omega \in [0,1)$, the sequence of real numbers $\{X_n(\omega)\}$ has limit:

$$\lim_{n\to\infty} X_n(\omega) = \lim_{n\to\infty} \omega^n = 0$$

For $\omega = 1$, the sequence of real numbers $\{X_n(\omega)\}$ has limit:

$$\lim_{n\to\infty} X_n(\omega) = \lim_{n\to\infty} \omega^n = \lim_{n\to\infty} 1^n = 1$$

Therefore, the sequence of random variables $\{X_n\}$ does not converge pointwise to X, because

$$\lim_{n\to\infty} X_n(\omega) \neq X(\omega)$$

for $\omega = 1$. However, the set of sample points ω such that $\{X_n(\omega)\}$ does not converge to $X(\omega)$ is a zero-probability event:

$$\mathrm{P}(\{\omega \in \Omega : \{X_n(\omega)\} \text{ does not converge to } X(\omega)\}) = \mathrm{P}(\{1\}) = 0$$

Therefore, the sequence $\{X_n\}$ converges almost surely to X.

Exercise 2

Let $\{X_n\}$ and $\{Y_n\}$ be two sequences of random variables defined on a sample space Ω. Let X and Y be two random variables defined on Ω such that:

$$X_n \overset{a.s.}{\to} X$$
$$Y_n \overset{a.s.}{\to} Y$$

Prove that

$$X_n + Y_n \overset{a.s.}{\to} X + Y$$

Solution

Denote by F_X the set of sample points for which $\{X_n(\omega)\}$ converges to $X(\omega)$:

$$F_X = \{\omega \in \Omega : \{X_n(\omega)\} \text{ converges to } X(\omega)\}$$

The fact that $\{X_n\}$ converges almost surely to X implies that

$$F_X^c \subseteq E_X$$

where $\mathrm{P}(E_X) = 0$.

Denote by F_Y the set of sample points for which $\{Y_n(\omega)\}$ converges to $Y(\omega)$:

$$F_Y = \{\omega \in \Omega : \{Y_n(\omega)\} \text{ converges to } Y(\omega)\}$$

The fact that $\{Y_n\}$ converges almost surely to Y implies that

$$F_Y^c \subseteq E_Y$$

where $\mathrm{P}(E_Y) = 0$.

Now, denote by F_{XY} the set of sample points for which $\{X_n(\omega) + Y_n(\omega)\}$ converges to $X(\omega) + Y(\omega)$:

$$F_{XY} = \{\omega \in \Omega : \{X_n(\omega) + Y_n(\omega)\} \text{ converges to } X(\omega) + Y(\omega)\}$$

Observe that if $\omega \in F_X \cap F_Y$ then $\{X_n(\omega) + Y_n(\omega)\}$ converges to $X(\omega) + Y(\omega)$, because the sum of two sequences of real numbers is convergent if the two sequences are convergent. Therefore

$$F_X \cap F_Y \subseteq F_{XY}$$

Taking the complement of both sides, we obtain

$$\begin{array}{rcl} F_{XY}^c & \subseteq & (F_X \cap F_Y)^c \\ \boxed{\text{A}} & = & F_X^c \cup F_Y^c \\ \boxed{\text{B}} & \subseteq & E_X \cup E_Y \end{array}$$

where: in step $\boxed{\text{A}}$ we have used De Morgan's law; in step $\boxed{\text{B}}$ we have used the fact that $F_X^c \subseteq E_X$ and $F_Y^c \subseteq E_Y$. But

$$\begin{array}{rcl} \mathrm{P}(E_X \cup E_Y) & = & \mathrm{P}(E_X) + \mathrm{P}(E_Y) - \mathrm{P}(E_X \cap E_Y) \\ & \leq & \mathrm{P}(E_X) + \mathrm{P}(E_Y) = 0 + 0 = 0 \end{array}$$

and, as a consequence, $\mathrm{P}(E_X \cup E_Y) = 0$. Thus, the set F_{XY}^c of sample points ω such that $\{X_n(\omega) + Y_n(\omega)\}$ does not converge to $X(\omega) + Y(\omega)$ is included in the zero-probability event $E_X \cup E_Y$, which means that

$$X_n + Y_n \overset{a.s.}{\to} X + Y$$

Exercise 3

Let the sample space Ω be:

$$\Omega = [0, 1]$$

Sub-intervals of $[0, 1]$ are assigned a probability equal to their length, as in Exercise 1 above.

Define a sequence of random variables $\{X_n\}$ as follows:

$$X_n(\omega) = \begin{cases} 1 & \text{if } \omega \in \left(0, 1 - \frac{1}{n}\right) \\ n & \text{otherwise} \end{cases}$$

Find an almost sure limit of the sequence.

Solution

If $\omega = 0$ or $\omega = 1$, then the sequence of real numbers $\{X_n(\omega)\}$ is not convergent:

$$\lim_{n \to \infty} X_n(\omega) = \lim_{n \to \infty} n = \infty$$

For $\omega \in (0,1)$, the sequence of real numbers $\{X_n(\omega)\}$ has limit:

$$\lim_{n\to\infty} X_n(\omega) = 1$$

because for any ω we can find n_0 such that $\omega \in \left(0, 1-\frac{1}{n}\right)$ for any $n \geq n_0$ (as a consequence $X_n(\omega) = 1$ for any $n \geq n_0$).

Thus, the sequence of random variables $\{X_n\}$ converges almost surely to the random variable X defined as:

$$X(\omega) = 1 \quad \forall \omega \in \Omega$$

because the set of sample points ω such that $\{X_n(\omega)\}$ does not converge to $X(\omega)$ is a zero-probability event:

$$\begin{aligned} & \mathrm{P}\left(\{\omega \in \Omega : \{X_n(\omega)\} \text{ does not converge to } X(\omega)\}\right) \\ = \; & \mathrm{P}(\{0,1\}) = \mathrm{P}(\{0\}) + \mathrm{P}(\{1\}) = 0 + 0 = 0 \end{aligned}$$

Chapter 63

Convergence in probability

This lecture discusses convergence in probability. We deal first with convergence in probability of sequences of random variables and then with convergence in probability of sequences of random vectors.

63.1 Sequences of random variables

As we have discussed in the lecture entitled *Sequences of random variables* (p. 495), different concepts of convergence are based on different ways of measuring the distance between two random variables (how "close to each other" two random variables are). The concept of **convergence in probability** is based on the following intuition: two random variables are "close to each other" if there is a high probability that their difference is very small.

Let $\{X_n\}$ be a sequence of random variables defined on a sample space[1] Ω. Let X be a random variable and ε a strictly positive number. Consider the following probability:

$$\mathrm{P}\left(|X_n - X| > \varepsilon\right) \tag{63.1}$$

Intuitively, X_n is considered far from X when $|X_n - X| > \varepsilon$; therefore, (63.1) is the probability that X_n is far from X. If $\{X_n\}$ converges to X, then (63.1) should become smaller and smaller as n increases. In other words, the probability of X_n being far from X should go to zero when n increases. Formally, we should have:

$$\lim_{n\to\infty} \mathrm{P}\left(|X_n - X| > \varepsilon\right) = 0 \tag{63.2}$$

Note that the sequence

$$\{\mathrm{P}\left(|X_n - X| > \varepsilon\right)\}$$

is a sequence of real numbers, therefore the limit in (63.2) is the usual limit of a sequence of real numbers.

Furthermore, condition (63.2) should be satisfied for any ε (also for very small ε, which means that we are very restrictive on our criterion for deciding whether X_n is far from X). This leads us to the following definition of convergence:

[1] See p. 492.

Definition 307 *Let $\{X_n\}$ be a sequence of random variables defined on a sample space Ω. We say that $\{X_n\}$ is **convergent in probability** to a random variable X defined on Ω if and only if*

$$\lim_{n\to\infty} \mathrm{P}\left(|X_n - X| > \varepsilon\right) = 0$$

*for any $\varepsilon > 0$. X is called the **probability limit** of the sequence and convergence is indicated by*

$$X_n \xrightarrow{\mathrm{P}} X$$

or by

$$\operatorname*{plim}_{n\to\infty} X_n = X$$

The following example illustrates the concept of convergence in probability:

Example 308 *Let X be a discrete random variable with support*

$$R_X = \{0, 1\}$$

and probability mass function[2]:

$$p_X(x) = \begin{cases} 1/3 & \text{if } x = 1 \\ 2/3 & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases}$$

Consider a sequence of random variables $\{X_n\}$ whose generic term is:

$$X_n = \left(1 + \frac{1}{n}\right) X$$

We want to prove that $\{X_n\}$ converges in probability to X. Take any $\varepsilon > 0$. Note that:

$$\begin{aligned} & |X_n - X| \\ \boxed{A} \quad &= \left|\left(1 + \frac{1}{n}\right) X - X\right| \\ &= \left|\frac{1}{n} X\right| \\ \boxed{B} \quad &= \frac{1}{n} X \end{aligned}$$

where: in step $\boxed{A}$ we have used the defintion of X_n; in step $\boxed{B}$ we have used the fact that X cannot be negative. When $X = 0$, which happens with probability $\frac{2}{3}$, we have that:

$$|X_n - X| = \frac{1}{n} X = 0$$

and, of course, $|X_n - X| \leq \varepsilon$. When $X = 1$, which happens with probability $\frac{1}{3}$, we have that

$$|X_n - X| = \frac{1}{n} X = \frac{1}{n}$$

[2] See p. 106.

and $|X_n - X| \leq \varepsilon$ *if and only if* $\frac{1}{n} \leq \varepsilon$ *(or* $n \geq \frac{1}{\varepsilon}$*). Therefore:*

$$\mathrm{P}\left(|X_n - X| \leq \varepsilon\right) = \begin{cases} 2/3 & \text{if } n < \frac{1}{\varepsilon} \\ 1 & \text{if } n \geq \frac{1}{\varepsilon} \end{cases}$$

and

$$\mathrm{P}\left(|X_n - X| > \varepsilon\right) = 1 - \mathrm{P}\left(|X_n - X| \leq \varepsilon\right) = \begin{cases} 1/3 & \text{if } n < \frac{1}{\varepsilon} \\ 0 & \text{if } n \geq \frac{1}{\varepsilon} \end{cases}$$

Thus, $\mathrm{P}\left(|X_n - X| > \varepsilon\right)$ *trivially converges to* 0*, because it is identically equal to* 0 *for all* n *such that* $n \geq \frac{1}{\varepsilon}$*. Since* ε *was arbitrary, we have obtained the desired result:*

$$\lim_{n \to \infty} \mathrm{P}\left(|X_n - X| > \varepsilon\right) = 0$$

for any $\varepsilon > 0$*.*

63.2 Sequences of random vectors

The above notion of convergence generalizes to sequences of random vectors in a straightforward manner.

Let $\{X_n\}$ be a sequence of random vectors defined on a sample space[3] Ω, where each random vector X_n has dimension $K \times 1$.

We have explained above that a sequence of random variables $\{X_n\}$ converges in probability if and only if

$$\lim_{n \to \infty} \mathrm{P}\left(d\left(X_n, X\right) > \varepsilon\right) = 0$$

for any $\varepsilon > 0$, where

$$d\left(X_n, X\right) = |X_n - X|$$

is the distance[4] of X_n from X.

In the case of random vectors, the definition of convergence in probability remains the same, but distance is measured by the Euclidean norm of the difference between the two vectors:

$$\begin{aligned} d\left(X_n, X\right) &= \|X_n - X\| \\ &= \sqrt{\left[X_{n,1}\left(\omega\right) - X_{\bullet,1}\left(\omega\right)\right]^2 + \ldots + \left[X_{n,K}\left(\omega\right) - X_{\bullet,K}\left(\omega\right)\right]^2} \end{aligned}$$

where the second subscript is used to indicate the individual components of the vectors $X_n\left(\omega\right)$ and $X\left(\omega\right)$.

The following is a formal definition:

Definition 309 *Let* $\{X_n\}$ *be a sequence of random vectors defined on a sample space* Ω*. We say that* $\{X_n\}$ *is* ***convergent in probability*** *to a random vector* X *defined on* Ω *if and only if*

$$\lim_{n \to \infty} \mathrm{P}\left(\|X_n - X\| > \varepsilon\right) = 0$$

for any $\varepsilon > 0$*.* X *is called the* ***probability limit*** *of the sequence and convergence is indicated by*

$$X_n \xrightarrow{\mathrm{P}} X$$

[3] See p. 497.
[4] See p. 34.

or by

$$\operatorname*{plim}_{n\to\infty} X_n = X$$

Now, denote by $\{X_{n,i}\}$ the sequence of the i-th components of the vectors X_n. It can be proved that the sequence of random vectors $\{X_n\}$ is convergent in probability if and only if all the K sequences of random variables $\{X_{n,i}\}$ are convergent in probability:

Proposition 310 *Let $\{X_n\}$ be a sequence of random vectors defined on a sample space Ω. Denote by $\{X_{n,i}\}$ the sequence of random variables obtained by taking the i-th component of each random vector X_n. The sequence $\{X_n\}$ converges in probability to the random vector X if and only if the sequence $\{X_{n,i}\}$ converges in probability to the random variable $X_{\bullet,i}$ (the i-th component of X) for each $i = 1, \ldots, K$.*

63.3 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Let U be a random variable having a uniform distribution[5] on the interval $[0, 1]$. In other words, U is an absolutely continuous random variable with support

$$R_U = [0, 1]$$

and probability density function

$$f_U(u) = \begin{cases} 1 & \text{if } u \in [0,1] \\ 0 & \text{if } u \notin [0,1] \end{cases}$$

Now, define a sequence of random variables $\{X_n\}$ as follows:

$$\begin{array}{rcl}
X_1 & = & 1_{\{U\in[0,1]\}} \\
X_2 & = & 1_{\{U\in[0,1/2]\}} \quad X_3 = 1_{\{U\in[1/2,1]\}} \\
X_4 & = & 1_{\{U\in[0,1/4]\}} \quad X_5 = 1_{\{U\in[1/4,2/4]\}} \quad X_6 = 1_{\{U\in[2/4,3/4]\}} \quad X_7 = 1_{\{U\in[3/4,1]\}} \\
X_8 & = & 1_{\{U\in[0,1/8]\}} \quad X_9 = 1_{\{U\in[1/8,2/8]\}} \quad X_{10} = 1_{\{U\in[2/8,3/8]\}} \quad \cdots \\
X_{16} & = & 1_{\{U\in[0,1/16]\}} \quad X_{17} = 1_{\{U\in[1/16,2/16]\}} \quad X_{18} = 1_{\{U\in[2/16,3/16]\}} \quad \cdots \\
 & \vdots &
\end{array}$$

where $1_{\{U\in[a,b]\}}$ is the indicator function[6] of the event $\{U \in [a, b]\}$.

Find the probability limit (if it exists) of the sequence $\{X_n\}$.

[5] See p. 359.
[6] See p. 197.

Solution

A generic term X_n of the sequence, being an indicator function, can take only two values:

- it can take value 1 with probability:

$$\mathrm{P}\left(X_n = 1\right) = \mathrm{P}\left(U \in \left[\frac{j}{m}, \frac{j+1}{m}\right]\right) = \frac{1}{m}$$

  where m is an integer satisfying

$$\frac{n}{2} < m \leq n \tag{63.3}$$

  and j is an integer satisfying

$$n = m + j$$

- it can take value 0 with probability:

$$\mathrm{P}\left(X_n = 0\right) = 1 - \mathrm{P}\left(X_n = 1\right) = 1 - \frac{1}{m}$$

By (515), m goes to infinity as n goes to infinity and

$$\lim_{n\to\infty} \mathrm{P}\left(X_n = 0\right) = \lim_{m\to\infty} \left(1 - \frac{1}{m}\right) = 1$$

Therefore, the probability that X_n is equal to zero converges to 1 as n goes to infinity. So, obviously, $\{X_n\}$ converges in probability to the constant random variable

$$X\left(\omega\right) = 0, \ \forall \omega \in \Omega$$

because for any $\varepsilon > 0$,

$$\begin{aligned} \lim_{n\to\infty} \mathrm{P}\left(\left|X_n - X\right| > \varepsilon\right) &= \lim_{n\to\infty} \mathrm{P}\left(\left|X_n - 0\right| > \varepsilon\right) \\ \boxed{\text{A}} &= \lim_{n\to\infty} \mathrm{P}\left(X_n > \varepsilon\right) \\ \boxed{\text{B}} &= \lim_{n\to\infty} \mathrm{P}\left(X_n = 1\right) \\ &= \lim_{m\to\infty} \frac{1}{m} = 0 \end{aligned}$$

where: in step A we have used the fact that X_n is positive; in step B we have used the fact that X_n can take only value 0 or value 1.

Exercise 2

Does the sequence in the previous exercise also converge almost surely[7]?

[7] See p. 505.

Solution

We can identify the sample space[8] Ω with the support of U:

$$\Omega = R_U = [0,1]$$

and the sample points $\omega \in \Omega$ with the realizations of U: when the realization is $U = u$, then $\omega = u$. Almost sure convergence requires that

$$\left\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}^c \subseteq E$$

where E is a zero-probability event[9] and the superscript c denotes the complement of a set. In other words, the set of sample points ω for which the sequence $\{X_n(\omega)\}$ does not converge to $X(\omega)$ must be included in a zero-probability event E. In our case, it is easy to see that, for any fixed sample point $\omega = u \in [0,1]$, the sequence $\{X_n(\omega)\}$ does not converge to $X(\omega) = 0$, because infinitely many terms in the sequence are equal to 1. Therefore:

$$\mathrm{P}\left(\left\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}^c\right) = 1$$

and, trivially, there does not exist a zero-probability event including the set

$$\left\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}^c$$

Thus, the sequence does not converge almost surely to X.

Exercise 3

Let $\{X_n\}$ be an IID sequence[10] of continuous random variables having a uniform distribution with support

$$R_{X_n} = \left[-\frac{1}{n}, \frac{1}{n}\right]$$

and probability density function

$$f_{X_n}(x) = \begin{cases} \frac{n}{2} & \text{if } x \in \left[-\frac{1}{n}, \frac{1}{n}\right] \\ 0 & \text{if } x \notin \left[-\frac{1}{n}, \frac{1}{n}\right] \end{cases}$$

Find the probability limit (if it exists) of the sequence $\{X_n\}$.

Solution

As n tends to infinity, the probability density tends to become concentrated around the point $x = 0$. Therefore, it seems reasonable to conjecture that the sequence $\{X_n\}$ converges in probability to the constant random variable

$$X(\omega) = 0, \ \forall \omega \in \Omega$$

To rigorously verify this claim we need to use the formal definition of convergence in probability. For any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathrm{P}(|X_n - X| > \varepsilon) \quad = \quad \lim_{n\to\infty} \mathrm{P}(|X_n - 0| > \varepsilon)$$

[8] See p. 69.
[9] See p. 79.
[10] See p. 492.

$$
\begin{aligned}
&= \lim_{n\to\infty} \left[1 - \mathrm{P}\left(-\varepsilon \leq X_n \leq \varepsilon\right)\right] \\
&= 1 - \lim_{n\to\infty} \int_{-\varepsilon}^{\varepsilon} f_{X_n}(x)\, dx \\
&= 1 - \lim_{n\to\infty} \int_{\max(-\varepsilon,-1/n)}^{\min(\varepsilon,1/n)} \frac{n}{2} dx \\
\boxed{\text{A}} \quad &= 1 - \lim_{n\to\infty} \int_{-1/n}^{1/n} \frac{n}{2} dx \\
&= 1 - \lim_{n\to\infty} 1 \\
&= 0
\end{aligned}
$$

where: in step $\boxed{\text{A}}$ we have used the fact that $\frac{1}{n} < \varepsilon$ when n becomes large.

Chapter 64

Mean-square convergence

This lecture discusses mean-square convergence. We deal first with mean-square convergence of sequences of random variables and then with mean-square convergence of sequences of random vectors.

64.1 Sequences of random variables

In the lecture entitled *Sequences of random variables* (p. 495) we have stressed the fact that different concepts of convergence are based on different ways of measuring the distance between two random variables (how "close to each other" two random variables are). The concept of **mean-square convergence**, or **convergence in mean-square**, is based on the following intuition: two random variables are "close to each other" if the square of their difference is on average small.

Let $\{X_n\}$ be a sequence of random variables defined on a sample space[1] Ω. Let X be a random variable. The sequence $\{X_n\}$ is said to converge to X in mean-square if $\{X_n\}$ converges to X according to the metric[2] $d(X_n, X)$ defined as follows:

$$d(X_n, X) = \mathrm{E}\left[(X_n - X)^2\right] \tag{64.1}$$

Note that $d(X_n, X)$ is well-defined only if the expected value on the right hand side exists. Usually, X_n and X are required to be square integrable[3], which ensures that (64.1) is well-defined and finite.

Intuitively, for a fixed sample point[4] ω, the squared difference

$$(X_n(\omega) - X(\omega))^2$$

between the two realizations of X_n and X provides a measure of how different those two realizations are. The mean squared difference (64.1) provides a measure of how different those two realizations are on average (as ω varies): if it becomes smaller and smaller by increasing n, then the sequence $\{X_n\}$ converges to X.

We summarize the concept of mean-square convergence in the following:

[1] See p. 492.

[2] If you do not understand what it means to "converge according to a metric", you need to revise the material discussed in the lecture entitled *Sequences and limits* (p. 34).

[3] See p. 159.

[4] See p. 69.

Definition 311 *Let $\{X_n\}$ be a sequence of square integrable random variables defined on a sample space Ω. We say that $\{X_n\}$ is **mean-square convergent**, or **convergent in mean-square**, if and only if there exists a square integrable random variable X such that $\{X_n\}$ converges to X, according to the metric*

$$d(X_n, X) = \mathrm{E}\left[(X_n - X)^2\right]$$

i.e.

$$\lim_{n\to\infty} \mathrm{E}\left[(X_n - X)^2\right] = 0 \tag{64.2}$$

*X is called the **mean-square limit** of the sequence and convergence is indicated by*

$$X_n \overset{m.s.}{\to} X$$

or by

$$X_n \overset{L^2}{\to} X$$

Note that (64.2) is just the usual criterion for convergence[5], while $X_n \overset{L^2}{\to} X$ indicates that convergence is in the L^p space[6] L^2, because both $\{X_n\}$ and X have been required to be square integrable.

The following example illustrates the concept of mean-square convergence:

Example 312 *Let $\{X_n\}$ be a covariance stationary[7] sequence of random variables such that all the random variables in the sequence have the same expected value μ, the same variance σ^2 and zero covariance with each other. Define the sample mean $\overline{X}_n$ as follows:*

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and define a constant random variable $X = \mu$. The distance between a generic term of the sequence $\left\{\overline{X}_n\right\}$ and X is

$$d\left(\overline{X}_n, X\right) = \mathrm{E}\left[\left(\overline{X}_n - X\right)^2\right] = \mathrm{E}\left[\left(\overline{X}_n - \mu\right)^2\right]$$

But μ is equal to the expected value of $\overline{X}_n$, because

$$\begin{aligned}\mathrm{E}\left[\overline{X}_n\right] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[X_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}n\mu = \mu\end{aligned}$$

Therefore

$$\begin{aligned}d\left(\overline{X}_n, X\right) &= \mathrm{E}\left[\left(\overline{X}_n - X\right)^2\right] \\ &= \mathrm{E}\left[\left(\overline{X}_n - \mu\right)^2\right]\end{aligned}$$

[5] See p. 36.
[6] See p. 136.
[7] See p. 493.

$$\begin{aligned} &= \mathrm{E}\left[\left(\overline{X}_n - \mathrm{E}\left[\overline{X}_n\right]\right)^2\right] \\ &= \mathrm{Var}\left[\overline{X}_n\right] \end{aligned}$$

by the very definition of variance. In turn, the variance of $\overline{X}_n$ *is*

$$\begin{aligned} \mathrm{Var}\left[\overline{X}_n\right] &= \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\ \boxed{A} &= \frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] \\ \boxed{B} &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left[X_i\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n} \end{aligned}$$

where: in step $\boxed{A}$ *we have used the properties of variance*[8]*; in step* $\boxed{B}$ *we have used the fact that the variance of a sum is equal to the sum of the variances when the random variables in the sum have zero covariance with each other*[9]*. Thus:*

$$d\left(\overline{X}_n, X\right) = \mathrm{E}\left[\left(\overline{X}_n - X\right)^2\right] = \mathrm{Var}\left[\overline{X}_n\right] = \frac{\sigma^2}{n}$$

and

$$\lim_{n\to\infty}\mathrm{E}\left[\left(\overline{X}_n - X\right)^2\right] = \lim_{n\to\infty}\frac{\sigma^2}{n} = 0$$

But this is just the definition of mean square convergence of $\overline{X}_n$ *to* X*. Therefore, the sequence* $\left\{\overline{X}_n\right\}$ *converges in mean-square to the constant random variable* $X = \mu$*.*

64.2 Sequences of random vectors

The above notion of convergence generalizes to sequences of random vectors in a straightforward manner.

Let $\{X_n\}$ be a sequence of random vectors defined on a sample space[10] Ω, where each random vector X_n has dimension $K \times 1$. The sequence of random vectors $\{X_n\}$ is said to converge to a random vector X in mean-square if $\{X_n\}$ converges to X according to the metric[11] $d(X_n, X)$ defined as follows:

$$\begin{aligned} d(X_n, X) &= \mathrm{E}\left[\|X_n - X\|^2\right] \\ &= \mathrm{E}\left[(X_{n,1} - X_{\bullet,1})^2 + \ldots + (X_{n,K} - X_{\bullet,K})^2\right] \end{aligned} \tag{64.3}$$

where $\|X_n - X\|$ is the Euclidean norm of the difference between X_n and X and the second subscript is used to indicate the individual components of the vectors X_n and X.

[8] See, in particular, the property *Multiplication by a constant* (p. 158).
[9] See p. 168.
[10] See p. 497.
[11] See p. 34.

Of course, $d(X_n, X)$ is well-defined only if the expected value on the right hand side exists. A sufficient condition for (64.3) to be well-defined is that all the components of X_n and X be square integrable random variables.

Intuitively, for a fixed sample point ω, the square of the Euclidean norm

$$\|X_n(\omega) - X(\omega)\|^2$$

of the difference between the two realizations of X_n and X provides a measure of how different those two realizations are. The mean of the square of the Euclidean norm (formula 64.3 above) provides a measure of how different those two realizations are on average (as ω varies): if it becomes smaller and smaller by increasing n, then the sequence of random vectors $\{X_n\}$ converges to the vector X.

The following is a formal definition of mean-square convergence for random vectors:

Definition 313 *Let $\{X_n\}$ be a sequence of random vectors defined on a sample space Ω, whose components are square integrable random variables. We say that $\{X_n\}$ is **mean-square convergent**, or **convergent in mean-square**, if and only if there exists a random vector X with square integrable components such that $\{X_n\}$ converges to X, according to the metric*

$$d(X_n, X) = \mathrm{E}\left[\|X_n - X\|^2\right]$$

i.e.

$$\lim_{n\to\infty} \mathrm{E}\left[\|X_n - X\|^2\right] = 0 \tag{64.4}$$

*X is called the **mean-square limit** of the sequence and convergence is indicated by*

$$X_n \overset{m.s.}{\to} X$$

or by:

$$X_n \overset{L^2}{\to} X$$

Note that (64.4) is just the usual criterion for convergence, while $X_n \overset{L^2}{\to} X$ indicates that convergence is in the L^p space L^2, because both $\{X_n\}$ and X have been required to have square integrable components.

Now, denote by $\{X_{n,i}\}$ the sequence of the i-th components of the vectors X_n. It can be proved that the sequence of random vectors $\{X_n\}$ is convergent in mean-square if and only if all the K sequences of random variables $\{X_{n,i}\}$ are convergent in mean-square:

Proposition 314 *Let $\{X_n\}$ be a sequence of random vectors defined on a sample space Ω, such that their components are square integrable random variables. Denote by $\{X_{n,i}\}$ the sequence of random variables obtained by taking the i-th component of each random vector X_n. The sequence $\{X_n\}$ converges in mean-square to the random vector X if and only if $\{X_{n,i}\}$ converges in mean-square to the random variable $X_{\bullet,i}$ (the i-th component of X) for each $i = 1, \ldots, K$.*

64.3 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Let U be a random variable having a uniform distribution[12] on the interval $[1,2]$. In other words, U is an absolutely continuous random variable with support

$$R_U = [1,2]$$

and probability density function

$$f_U(u) = \begin{cases} 1 & \text{if } u \in [1,2] \\ 0 & \text{if } u \notin [1,2] \end{cases}$$

Consider a sequence of random variables $\{X_n\}$ whose generic term is

$$X_n = 1_{\{U \in [1,2-1/n]\}}$$

where $1_{\{U \in [1,2-1/n]\}}$ is the indicator function[13] of the event $\{U \in [1,2-1/n]\}$.

Find the mean-square limit (if it exists) of the sequence $\{X_n\}$.

Solution

When n tends to infinity, the interval $[1,2-1/n]$ becomes similar to the interval $[1,2]$, because

$$\lim_{n\to\infty}\left(2-\frac{1}{n}\right) = 2$$

Therefore, we conjecture that the indicators $1_{\{U \in [1,2-1/n]\}}$ converge in mean-square to the indicator $1_{\{U \in [1,2]\}}$. But $1_{\{U \in [1,2]\}}$ is always equal to 1, so our conjecture is that the sequence $\{X_n\}$ converges in mean square to 1. To verify our conjecture, we need to verify that

$$\lim_{n\to\infty} \mathrm{E}\left[(X_n-1)^2\right] = 0$$

The expected value can be computed as follows:

$$\begin{aligned}
\mathrm{E}\left[(X_n-1)^2\right] &= \int_{-\infty}^{\infty} \left(1_{\{u \in [1,2-1/n]\}} - 1\right)^2 f_U(u)\, du \\
&= \int_1^2 \left(1_{\{u \in [1,2-1/n]\}} - 1\right)^2 du \\
&= \int_1^2 \left(1^2_{\{u \in [1,2-1/n]\}} + 1^2 - 2 \cdot 1_{\{u \in [1,2-1/n]\}} 1\right) du \\
&= \int_1^2 \left(1_{\{u \in [1,2-1/n]\}} + 1 - 2 \cdot 1_{\{u \in [1,2-1/n]\}}\right) du \\
&= \int_1^2 1_{\{u \in [1,2-1/n]\}} du + \int_1^2 du - 2\int_1^2 1_{\{u \in [1,2-1/n]\}} du \\
&= \int_1^{2-1/n} du + \int_1^2 du - 2\int_1^{2-1/n} du \\
&= [u]_1^{2-1/n} + [u]_1^2 - 2\,[u]_1^{2-1/n}
\end{aligned}$$

[12] See p. 359.
[13] See p. 197.

$$= \quad 2 - \frac{1}{n} - 1 + 2 - 1 - 2\left(2 - \frac{1}{n} - 1\right) = \frac{1}{n}$$

Thus, the sequence $\{X_n\}$ converges in mean-square to 1 because

$$\lim_{n\to\infty} \mathrm{E}\left[(X_n - 1)^2\right] = \lim_{n\to\infty} \frac{1}{n} = 0$$

Exercise 2

Let $\{X_n\}$ be a sequence of discrete random variables. Let the probability mass function of a generic term of the sequence X_n be

$$p_{X_n}(x_n) = \begin{cases} 1/n & \text{if } x_n = n \\ 1 - 1/n & \text{if } x_n = 0 \\ 0 & \text{otherwise} \end{cases}$$

Find the mean-square limit (if it exists) of the sequence $\{X_n\}$.

Solution

Note that

$$\lim_{n\to\infty} \mathrm{P}(X_n = 0) = \lim_{n\to\infty} \left(1 - \frac{1}{n}\right) = 1$$

Therefore, one would expect that the sequence $\{X_n\}$ converges to the constant random variable $X = 0$. However, the sequence $\{X_n\}$ does not converge in mean-square to 0. The distance of a generic term of the sequence from 0 is

$$\mathrm{E}\left[(X_n - 0)^2\right] = \mathrm{E}\left[X_n^2\right] = n^2 \cdot \frac{1}{n} + 0 \cdot \left(1 - \frac{1}{n}\right) = n$$

Thus

$$\lim_{n\to\infty} \mathrm{E}\left[(X_n - 0)^2\right] = \infty$$

while, if $\{X_n\}$ was convergent, we would have

$$\lim_{n\to\infty} \mathrm{E}\left[(X_n - 0)^2\right] = 0$$

Exercise 3

Does the sequence in the previous exercise converge in probability?

Solution

The sequence $\{X_n\}$ converges in probability to the constant random variable $X = 0$ because for any $\varepsilon > 0$

$$\begin{aligned} & \lim_{n\to\infty} \mathrm{P}(|X_n - X| > \varepsilon) \\ = \quad & \lim_{n\to\infty} \mathrm{P}(|X_n - 0| > \varepsilon) \\ \boxed{\mathrm{A}} \quad = \quad & \lim_{n\to\infty} \mathrm{P}(X_n > \varepsilon) \end{aligned}$$

$$\begin{aligned} \boxed{\text{B}} \quad &= \lim_{n\to\infty} \mathrm{P}\left(X_n = n\right) \\ &= \lim_{n\to\infty} \frac{1}{n} = 0 \end{aligned}$$

where: in step $\boxed{\text{A}}$ we have used the fact that X_n is positive; in step $\boxed{\text{B}}$ we have used the fact that X_n can take only value 0 or value n.

Chapter 65

Convergence in distribution

This lecture discusses convergence in distribution. We deal first with convergence in distribution of sequences of random variables and then with convergence in distribution of sequences of random vectors.

65.1 Sequences of random variables

In the lecture entitled *Limit of a sequence of random variables* (p. 495) we explained that different concepts of convergence are based on different ways of measuring the distance between two random variables (how "close to each other" two random variables are). The concept of **convergence in distribution** is based on the following intuition: two random variables are "close to each other" if their distribution functions[1] are "close to each other" .

Let $\{X_n\}$ be a sequence of random variables[2]. Let us consider a generic random variable X_n belonging to the sequence. Denote by $F_n(x)$ its distribution function. $F_n(x)$ is a function $F_n : \mathbb{R} \to [0,1]$. Once we fix x, the value $F_n(x)$ associated to the point x is a real number. By the same token, once we fix x, the sequence $\{F_n(x)\}$ is a sequence of real numbers. Therefore, for a fixed x, it is very easy to assess whether the sequence $\{F_n(x)\}$ is convergent; this is done employing the usual definition of convergence of sequences of real numbers[3]. If, for a fixed x, the sequence $\{F_n(x)\}$ is convergent, we denote its limit by $F_X(x)$ (note that the limit depends on the specific x we have fixed). A sequence of random variables $\{X_n\}$ is said to be convergent in distribution if and only if the sequence $\{F_n(x)\}$ is convergent for any choice of x (except, possibly, for some "special values" of x where $F_X(x)$ is not continuous in x):

Definition 315 *Let $\{X_n\}$ be a sequence of random variables. Denote by $F_n(x)$ the distribution function of X_n. We say that $\{X_n\}$ is* ***convergent in distribution*** *(or convergent in law) if and only if there exists a distribution function $F_X(x)$ such that the sequence $\{F_n(x)\}$ converges to $F_X(x)$ for all points $x \in \mathbb{R}$ where $F_X(x)$ is continuous. If a random variable X has distribution function $F_X(x)$, then X is called the* ***limit in distribution*** *(or limit in law) of the sequence and convergence*

[1] See p. 108.
[2] See p. 491.
[3] See p. 33.

is indicated by

$$X_n \xrightarrow{d} X$$

Note that convergence in distribution only involves the distribution functions of the random variables belonging to the sequence $\{X_n\}$ and that these random variables need not be defined on the same sample space[4]. On the contrary, the modes of convergence we have discussed in previous lectures (pointwise convergence, almost sure convergence, convergence in probability, mean-square convergence) require that all the variables in the sequence be defined on the same sample space.

Example 316 *Let $\{X_n\}$ be a sequence of IID[5] random variables all having a uniform distribution[6] on the interval $[0,1]$, i.e. the distribution function of X_n is*

$$F_{X_n}(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

Define:

$$Y_n = n\left(1 - \max_{1 \leq i \leq n} X_i\right)$$

The distribution function of Y_n is

$$\begin{aligned}
F_{Y_n}(y) &= \mathrm{P}(Y_n \leq y) \\
&= \mathrm{P}\left(n\left(1 - \max_{1 \leq i \leq n} X_i\right) \leq y\right) \\
&= \mathrm{P}\left(\max_{1 \leq i \leq n} X_i \geq 1 - \frac{y}{n}\right) \\
&= 1 - \mathrm{P}\left(\max_{1 \leq i \leq n} X_i < 1 - \frac{y}{n}\right) \\
&= 1 - \mathrm{P}\left(X_1 < 1 - \frac{y}{n}, X_2 < 1 - \frac{y}{n}, \ldots, X_n < 1 - \frac{y}{n}\right) \\
\boxed{A} &= 1 - \mathrm{P}\left(X_1 < 1 - \frac{y}{n}\right) \cdot \mathrm{P}\left(X_2 < 1 - \frac{y}{n}\right) \cdot \ldots \cdot \mathrm{P}\left(X_n < 1 - \frac{y}{n}\right) \\
\boxed{B} &= 1 - \mathrm{P}\left(X_1 \leq 1 - \frac{y}{n}\right) \cdot \mathrm{P}\left(X_2 \leq 1 - \frac{y}{n}\right) \cdot \ldots \cdot \mathrm{P}\left(X_n \leq 1 - \frac{y}{n}\right) \\
\boxed{C} &= 1 - F_{X_1}\left(1 - \frac{y}{n}\right) \cdot F_{X_2}\left(1 - \frac{y}{n}\right) \cdot \ldots \cdot F_{X_n}\left(1 - \frac{y}{n}\right) \\
\boxed{D} &= 1 - \left[F_{X_n}\left(1 - \frac{y}{n}\right)\right]^n
\end{aligned}$$

where: in step $\boxed{A}$ we have used the fact that the variables X_i are mutually independent; in step $\boxed{B}$ we have used the fact that the variables X_i are absolutely continuous; in step $\boxed{C}$ we have used the definition of distribution function; in step $\boxed{D}$ we have used the fact that the variables X_i have identical distributions. Thus:

$$F_{Y_n}(y) = \begin{cases} 0 & \text{if } y < 0 \\ 1 - \left(1 - \frac{y}{n}\right)^n & \text{if } 0 \leq y < n \\ 1 & \text{if } y \geq n \end{cases}$$

[4] See p. 69.
[5] See p. 492.
[6] See p. 359.

Since

$$\lim_{n\to\infty}\left(1-\frac{y}{n}\right)^{n}=\exp\left(-y\right)$$

we have

$$\lim_{n\to\infty}F_{Y_n}\left(y\right)=F_Y\left(y\right)=\begin{cases}0 & \text{if } y<0\\ 1-\exp\left(-y\right) & \text{if } y\geq 0\end{cases}$$

where $F_Y(y)$ is the distribution function of an exponential random variable[7]*. Therefore, the sequence $\{Y_n\}$ converges in law to an exponential distribution.*

65.2 Sequences of random vectors

The definition of convergence in distribution of a sequence of random vectors is almost identical; we just need to replace distribution functions in the above definition with joint distribution functions[8]:

Definition 317 *Let $\{X_n\}$ be a sequence of $K\times 1$ random vectors. Denote by $F_n(x)$ the joint distribution function of X_n. We say that $\{X_n\}$ is* ***convergent in distribution*** *(or convergent in law) if and only if there exists a joint distribution function $F_X(x)$ such that the sequence $\{F_n(x)\}$ converges to $F_X(x)$ for all points $x\in\mathbb{R}^K$ where $F_X(x)$ is continuous. If a random vector X has joint distribution function $F_X(x)$, then X is called the* ***limit in distribution*** *(or limit in law) of the sequence and convergence is indicated by*

$$X_n\xrightarrow{d}X$$

65.3 More details

65.3.1 Proper distribution functions

Let $\{X_n\}$ be a sequence of random variables and denote by $F_n(x)$ the distribution function of X_n. Suppose that we find a function $F_X(x)$ such that

$$F_X\left(x\right)=\lim_{n\to\infty}F_n\left(x\right)$$

for all $x\in\mathbb{R}$ where $F_X(x)$ is continuous. How do we check that $F_X(x)$ is a proper distribution function, so that we can say that the sequence $\{X_n\}$ converges in distribution?

$F_X(x)$ is a proper distribution function if it satisfies the following four properties:

1. **Increasing**. $F_X(x)$ is increasing, i.e. $F_X(x_1)\leq F_X(x_2)$ if $x_1<x_2$.

2. **Right-continuous**. $F_X(x)$ is right-continuous, i.e.

$$\lim_{\substack{t\to x\\ t\geq x}}F_X\left(t\right)=F_X\left(x\right)$$

   for any $x\in\mathbb{R}$.

[7] See p. 365.
[8] See p. 118.

3. **Limit at minus infinity**. $F_X(x)$ satisfies

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

4. **Limit at plus infinity**. $F_X(x)$ satisfies

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

65.4 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Let $\{X_n\}$ be a sequence of random variables having distribution functions

$$F_n(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \frac{n}{2n+1}x + \frac{1}{4n+2}x^2 & \text{if } 0 < x \leq 1 \\ \frac{n}{2n+1}x - \frac{1}{4n+2}\left(x^2 - 4x + 2\right) & \text{if } 1 < x \leq 2 \\ 1 & \text{if } x > 2 \end{cases}$$

Find the limit in distribution (if it exists) of the sequence $\{X_n\}$.

Solution

If $0 < x \leq 1$, then

$$\begin{aligned} \lim_{n\to\infty} F_n(x) &= \lim_{n\to\infty}\left[\frac{n}{2n+1}x + \frac{1}{4n+2}x^2\right] \\ &= x\cdot\lim_{n\to\infty}\left(\frac{n}{2n+1}\right) + x^2\cdot\lim_{n\to\infty}\left(\frac{1}{4n+2}\right) \\ &= x\cdot\frac{1}{2} + x^2\cdot 0 = \frac{1}{2}x \end{aligned}$$

If $1 < x \leq 2$, then

$$\begin{aligned} \lim_{n\to\infty} F_n(x) &= \lim_{n\to\infty}\left[\frac{n}{2n+1}x - \frac{1}{4n+2}\left(x^2 - 4x + 2\right)\right] \\ &= x\cdot\lim_{n\to\infty}\left(\frac{n}{2n+1}\right) + \left(x^2 - 4x + 2\right)\cdot\lim_{n\to\infty}\left(\frac{1}{4n+2}\right) \\ &= x\cdot\frac{1}{2} + \left(x^2 - 4x + 2\right)\cdot 0 = \frac{1}{2}x \end{aligned}$$

We now need to verify that the function

$$F_X(x) = \lim_{n\to\infty} F_n(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \frac{1}{2}x & \text{if } 0 < x \leq 2 \\ 1 & \text{if } x > 2 \end{cases}$$

is a proper distribution function. The function is increasing, continuous, its limit at minus infinity is 0 and its limit at plus infinity is 1, hence it satisfies the four properties that a proper distribution function needs to satisfy. This implies that $\{X_n\}$ converges in distribution to a random variable X having distribution function $F_X(x)$.

Exercise 2

Let $\{X_n\}$ be a sequence of random variables having distribution functions

$$F_n(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1-(1-x)^n & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x > 1 \end{cases}$$

Find the limit in distribution (if it exists) of the sequence $\{X_n\}$.

Solution

If $x = 0$, then

$$\begin{aligned} \lim_{n\to\infty} F_n(x) &= \lim_{n\to\infty} [1-(1-x)^n] = 1 - \lim_{n\to\infty} (1-0)^n \\ &= 1-1=0 \end{aligned}$$

If $0 < x \leq 1$, then

$$\begin{aligned} \lim_{n\to\infty} F_n(x) &= \lim_{n\to\infty} [1-(1-x)^n] = 1 - \lim_{n\to\infty} (1-x)^n \\ &= 1-0=1 \end{aligned}$$

Therefore, the distribution functions $F_n(x)$ converge to the function

$$G_X(x) = \lim_{n\to\infty} F_n(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 & \text{if } x > 0 \end{cases}$$

which is not a proper distribution function, because it is not right-continuous at the point $x = 0$. However, note that the function

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases}$$

is a proper distribution function and it is equal to $\lim_{n\to\infty} F_n(x)$ at all points except at the point $x = 0$. But this is a point of discontinuity of $F_X(x)$. As a consequence, the sequence $\{X_n\}$ converges in distribution to a random variable X having distribution function $F_X(x)$.

Exercise 3

Let $\{X_n\}$ be a sequence of random variables having distribution functions

$$F_n(x) = \begin{cases} 0 & \text{if } x < 0 \\ nx & \text{if } 0 \leq x \leq 1/n \\ 1 & \text{if } x > 1/n \end{cases}$$

Find the limit in distribution (if it exists) of the sequence $\{X_n\}$.

Solution

The distribution functions $F_n(x)$ converge to the function

$$G_X(x) = \lim_{n\to\infty} F_n(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 & \text{if } x > 0 \end{cases}$$

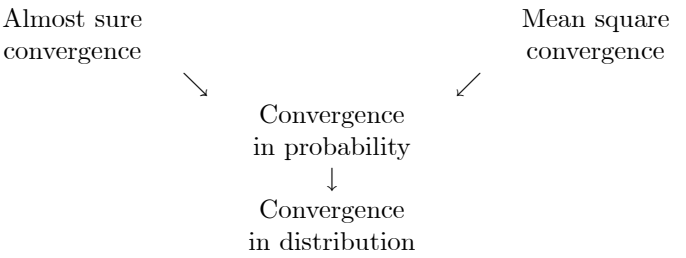
This is the same limiting function found in the previous exercise. As a consequence, the sequence $\{X_n\}$ converges in distribution to a random variable X having distribution function

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases}$$

Chapter 66

Relations between modes of convergence

In the previous lectures, we have introduced several notions of convergence of a sequence of random variables (also called **modes of convergence**). There are several relations between the various modes of convergence, which are discussed in the following subsections and are summarized by the following diagram (an arrow denotes implication in the arrow's direction):



66.1 Almost sure ⇒ Probability

Proposition 318 *If a sequence of random variables $\{X_n\}$ converges almost surely to a random variable X, then $\{X_n\}$ also converges in probability to X.*

Proof. See e.g. Resnick[1] (1999). ■

66.2 Probability ⇒ Distribution

Proposition 319 *If a sequence of random variables $\{X_n\}$ converges in probability to a random variable X, then $\{X_n\}$ also converges in distribution to X.*

Proof. See e.g. Resnick (1999). ■

[1] Resnick, S.I. (1999) "A Probability Path", Birkhauser.

66.3 Almost sure ⇒ Distribution

Proposition 320 *If a sequence of random variables $\{X_n\}$ converges almost surely to a random variable X, then $\{X_n\}$ also converges in distribution to X.*

Proof. This is obtained putting together Propositions (318) and (319) above. ■

66.4 Mean square ⇒ Probability

Proposition 321 *If a sequence of random variables $\{X_n\}$ converges in mean square to a random variable X, then $\{X_n\}$ also converges in probability to X.*

Proof. We can apply Markov's inequality[2] to a generic term of the sequence $\left\{(X_n - X)^2\right\}$:

$$\mathrm{P}\left((X_n - X)^2 \geq c^2\right) \leq \frac{\mathrm{E}\left[(X_n - X)^2\right]}{c^2}$$

for any strictly positive real number c. Taking the square root of both sides of the left-hand inequality, we obtain

$$\mathrm{P}\left(|X_n - X| \geq c\right) \leq \frac{\mathrm{E}\left[(X_n - X)^2\right]}{c^2}$$

Taking limits on both sides, we get

$$\lim_{n\to\infty} \mathrm{P}\left(|X_n - X| \geq c\right) \leq \lim_{n\to\infty} \frac{\mathrm{E}\left[(X_n - X)^2\right]}{c^2} = \frac{\lim_{n\to\infty} \mathrm{E}\left[(X_n - X)^2\right]}{c^2} = 0$$

where we have used the fact that, by the very definition of convergence in mean square:

$$\lim_{n\to\infty} \mathrm{E}\left[(X_n - X)^2\right] = 0$$

Since, by the very definition of probability, it must be that

$$\mathrm{P}\left(|X_n - X| \geq c\right) \geq 0$$

then it must be that also

$$\lim_{n\to\infty} \mathrm{P}\left(|X_n - X| \geq c\right) = 0$$

Note that this holds for any arbitrarily small c. By the definition of convergence in probability, this means that X_n converges in probability to X (if you are wondering about strict and weak inequalities here and in the definition of convergence in probability, note that $|X_n - X| \geq c$ implies $|X_n - X| > \varepsilon$ for any strictly positive $\varepsilon < c$). ■

66.5 Mean square ⇒ Distribution

Proposition 322 *If a sequence of random variables $\{X_n\}$ converges in mean square to a random variable X, then $\{X_n\}$ also converges in distribution to X.*

Proof. This is obtained putting together Propositions (321) and (319) above. ■

[2] See p. 241.

Chapter 67

Laws of Large Numbers

Let $\{X_n\}$ be a sequence of random variables[1]. Let $\overline{X}_n$ be the sample mean of the first n terms of the sequence:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

A **Law of Large Numbers** (LLN) is a proposition stating a set of conditions that are sufficient to guarantee the convergence of the sample mean $\overline{X}_n$ to a constant, as the sample size n increases. It is called a **Weak** Law of Large Numbers (WLLN) if the sequence $\left\{\overline{X}_n\right\}$ converges in probability[2] and a **Strong** Law of Large Numbers (SLLN) if the sequence $\left\{\overline{X}_n\right\}$ converges almost surely[3].

There are literally dozens of Laws of Large Numbers. We report some examples below.

67.1 Weak Laws of Large Numbers

67.1.1 Chebyshev's WLLN

Probably, the best known Law of Large Numbers is Chebyshev's:

Proposition 323 (Chebyshev's WLLN) *Let $\{X_n\}$ be an uncorrelated and covariance stationary sequence of random variables[4]:*

$$\begin{array}{l}\exists \mu \in \mathbb{R} : \mathrm{E}\left[X_n\right] = \mu, \forall n \in \mathbb{N} \\ \exists \sigma^2 \in \mathbb{R}_+ : \mathrm{Var}\left[X_n\right] = \sigma^2, \forall n \in \mathbb{N} \\ \mathrm{Cov}\left[X_n, X_{n+k}\right] = 0, \forall n, k \in \mathbb{N}\end{array}$$

Then, a Weak Law of Large Numbers applies to the sample mean:

$$\operatorname*{plim}_{n\to\infty} \overline{X}_n = \mu$$

[1] See p. 491.
[2] See p. 511.
[3] See p. 505.
[4] In other words, all the random variables in the sequence have the same mean μ, the same variance σ^2 and zero covariance with each other. See p. 493 for a definition of covariance stationary sequence.

where plim *denotes a probability limit*[5].

Proof. The expected value of the sample mean $\overline{X}_n$ is:

$$\begin{aligned} \mathrm{E}\left[\overline{X}_n\right] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[X_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}n\mu = \mu \end{aligned}$$

The variance of the sample mean $\overline{X}_n$ is:

$$\begin{aligned} \mathrm{Var}\left[\overline{X}_n\right] &= \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\ \boxed{\text{A}} &= \frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] \\ \boxed{\text{B}} &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left[X_i\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n} \end{aligned}$$

where: in step $\boxed{\text{A}}$ we have used the properties of variance[6]; in step $\boxed{\text{B}}$ we have used the fact that the variance of a sum is equal to the sum of the variances when the random variables in the sum have zero covariance with each other[7]. Now we can apply Chebyshev's inequality[8] to the sample mean $\overline{X}_n$:

$$\mathrm{P}\left(\left|\overline{X}_n - \mathrm{E}\left[\overline{X}_n\right]\right| \geq k\right) \leq \frac{\mathrm{Var}\left[\overline{X}_n\right]}{k^2}$$

for any strictly positive real number k. Plugging in the values for the expected value and the variance derived above, we obtain:

$$\mathrm{P}\left(\left|\overline{X}_n - \mu\right| \geq k\right) \leq \frac{\sigma^2}{nk^2}$$

Since

$$\lim_{n\to\infty}\frac{\sigma^2}{nk^2} = 0$$

and

$$\mathrm{P}\left(\left|\overline{X}_n - \mu\right| \geq k\right) \geq 0$$

then it must also be that:

$$\lim_{n\to\infty}\mathrm{P}\left(\left|\overline{X}_n - \mu\right| \geq k\right) = 0$$

Note that this holds for any arbitrarily small k. By the very definition of convergence in probability, this means that $\overline{X}_n$ converges in probability to μ (if you are

[5] See p. 511.
[6] See, in particular, the *Multiplication by a constant* property (p. 158).
[7] See p. 168.
[8] See p. 242.

wondering about strict and weak inequalities here and in the definition of convergence in probability, note that $\left|\overline{X}_n - \mu\right| \geq k$ implies $\left|\overline{X}_n - \mu\right| > \varepsilon$ for any strictly positive $\varepsilon < k$). ■

Note that it is customary to state Chebyshev's Weak Law of Large Numbers as a result on the convergence in probability of the sample mean:

$$\operatorname*{plim}_{n\to\infty} \overline{X}_n = \mu$$

However, the conditions of the above theorem guarantee the mean square convergence[9] of the sample mean to μ:

$$\overline{X}_n \overset{m.s.}{\to} \mu$$

Proof. In the above proof of Chebyshev's Weak Law of Large Numbers, it is proved that:

$$\operatorname{Var}\left[\overline{X}_n\right] = \frac{\sigma^2}{n}$$

and that

$$\operatorname{E}\left[\overline{X}_n\right] = \mu$$

This implies that

$$\operatorname{E}\left[\left(\overline{X}_n - \mu\right)^2\right] = \operatorname{E}\left[\left(\overline{X}_n - \operatorname{E}\left[\overline{X}_n\right]\right)^2\right] = \operatorname{Var}\left[\overline{X}_n\right] = \frac{\sigma^2}{n}$$

As a consequence:

$$\lim_{n\to\infty} \operatorname{E}\left[\left(\overline{X}_n - \mu\right)^2\right] = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0$$

but this is just the definition of mean square convergence of $\overline{X}_n$ to μ. ■

Hence, in Chebyshev's Weak Law of Large Numbers, convergence in probability is just a consequence of the fact that convergence in mean square implies convergence in probability[10].

67.1.2 Chebyshev's WLLN for correlated sequences

Chebyshev's Weak Law of Large Numbers (see above) sets forth the requirement that the terms of the sequence $\{X_n\}$ have zero covariance with each other. By relaxing this requirement and allowing for some correlation between the terms of the sequence $\{X_n\}$, a more general version of Chebyshev's Weak Law of Large Numbers can be obtained:

Proposition 324 (Chebyshev's WLLN for correlated sequences) *Let $\{X_n\}$ be a covariance stationary sequence of random variables[11]:*

$$\begin{array}{l} \exists \mu \in \mathbb{R} : \operatorname{E}\left[X_n\right] = \mu, \forall n > 0 \\ \forall j \geq 0, \exists \gamma_j \in \mathbb{R} : \operatorname{Cov}\left[X_n, X_{n-j}\right] = \gamma_j, \forall n > j \end{array}$$

[9] See p. 519.

[10] See p. 534.

[11] In other words, all the random variables in the sequence have the same mean μ, the same variance γ_0 and the covariance between a term X_n of the sequence and the term that is located j positions before it (X_{n-j}) is always the same (γ_j), irrespective of how X_n has been chosen.

*If **covariances tend to be zero on average**, i.e. if*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n}\gamma_i = 0 \tag{67.1}$$

then a Weak Law of Large Numbers applies to the sample mean:

$$\operatorname*{plim}_{n\to\infty}\overline{X}_n = \mu$$

Proof. For a full proof see e.g. Karlin and Taylor[12] (1975). We give here a proof based on the assumption that covariances are absolutely summable:

$$\sum_{j=0}^{\infty}\left|\gamma_j\right| < \infty$$

which is stronger than (67.1). The expected value of the sample mean $\overline{X}_n$ is

$$\begin{aligned}\mathrm{E}\left[\overline{X}_n\right] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[X_i\right]\\ &= \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}n\mu = \mu\end{aligned}$$

The variance of the sample mean $\overline{X}_n$ is:

$$\begin{aligned}\mathrm{Var}\left[\overline{X}_n\right] &= \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n}X_i\right]\\ \boxed{\text{A}} &= \frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^{n}X_i\right]\\ \boxed{\text{B}} &= \frac{1}{n^2}\left\{\sum_{i=1}^{n}\mathrm{Var}\left[X_i\right] + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1}\mathrm{Cov}\left[X_i, X_j\right]\right\}\\ &= \frac{1}{n^2}\left\{\sum_{i=1}^{n}\gamma_0 + 2\gamma_1 + 2\left(\gamma_1+\gamma_2\right)\right.\\ &\quad \left.+\ldots+2\left(\gamma_1+\gamma_2+\ldots\gamma_{n-1}\right)\right\}\\ &= \frac{1}{n^2}\left\{n\gamma_0 + 2\left(n-1\right)\gamma_1 + 2\left(n-2\right)\gamma_2 + \ldots + 2\gamma_{n-1}\right\}\\ &= \frac{1}{n}\left\{\gamma_0 + 2\sum_{i=1}^{n-1}\frac{(n-i)}{n}\gamma_i\right\}\end{aligned}$$

where: in step $\boxed{\text{A}}$ we have used the properties of variance[13]; in step $\boxed{\text{B}}$ we have used the formula for the variance of a sum[14]. Note that:

$$\mathrm{Var}\left[\overline{X}_n\right] = \frac{1}{n}\left\{\gamma_0 + 2\sum_{i=1}^{n-1}\frac{(n-i)}{n}\gamma_i\right\}$$

[12] Karlin, S., Taylor, H. M. (1975) "A first course in stochastic processes", Academic Press.
[13] See, in particular, the *Multiplication by a constant* property (p. 158).
[14] See p. 168.

$$\begin{aligned}
&\leq \frac{1}{n}\left\{\gamma_0 + 2\sum_{i=1}^{n-1}\left|\frac{(n-i)}{n}\gamma_i\right|\right\} \\
&= \frac{1}{n}\left\{\gamma_0 + 2\sum_{i=1}^{n-1}\left|\frac{(n-i)}{n}\right||\gamma_i|\right\} \\
\boxed{\text{A}} \quad &\leq \frac{1}{n}\left\{\gamma_0 + 2\sum_{i=1}^{n-1}|\gamma_i|\right\} \\
&\leq \frac{1}{n}\left\{\gamma_0 + 2\sum_{i=1}^{\infty}|\gamma_i|\right\}
\end{aligned}$$

where: in step A we have used the fact that

$$\left|\frac{(n-i)}{n}\right| < 1$$

But the covariances are absolutely summable, so that:

$$\gamma_0 + 2\sum_{i=1}^{\infty}|\gamma_i| = \overline{\gamma}$$

where $\overline{\gamma}$ is a finite constant. Therefore:

$$\text{Var}\left[\overline{X}_n\right] \leq \frac{\overline{\gamma}}{n}$$

Now we can apply Chebyshev's inequality to the sample mean $\overline{X}_n$:

$$\text{P}\left(\left|\overline{X}_n - \text{E}\left[\overline{X}_n\right]\right| \geq k\right) \leq \frac{\text{Var}\left[\overline{X}_n\right]}{k^2}$$

for any strictly positive real number k. Plugging in the values for the expected value and the variance derived above, we obtain:

$$\text{P}\left(\left|\overline{X}_n - \mu\right| \geq k\right) \leq \frac{\text{Var}\left[\overline{X}_n\right]}{k^2} \leq \frac{\overline{\gamma}}{nk^2}$$

Since

$$\lim_{n\to\infty}\frac{\overline{\gamma}}{nk^2} = 0$$

and

$$\text{P}\left(\left|\overline{X}_n - \mu\right| \geq k\right) \geq 0$$

then it must also be that:

$$\lim_{n\to\infty}\text{P}\left(\left|\overline{X}_n - \mu\right| \geq k\right) = 0$$

Note that this holds for any arbitrarily small k. By the very definition of convergence in probability, this means that $\overline{X}_n$ converges in probability to μ (if you are wondering about strict and weak inequalities here and in the definition of convergence in probability, note that $\left|\overline{X}_n - \mu\right| \geq k$ implies $\left|\overline{X}_n - \mu\right| > \varepsilon$ for any strictly positive $\varepsilon < k$). ■

Also Chebyshev's Weak Law of Large Numbers for correlated sequences has been stated as a result on the convergence in probability of the sample mean:

$$\operatorname*{plim}_{n\to\infty} \overline{X}_n = \mu$$

However, the conditions of the above theorem also guarantee the mean square convergence of the sample mean to μ:

$$\overline{X}_n \overset{m.s.}{\to} \mu$$

Proof. In the above proof of Chebyshev's Weak Law of Large Numbers for correlated sequences, we proved that

$$\operatorname{Var}\left[\overline{X}_n\right] \leq \frac{\overline{\gamma}}{n}$$

and that

$$\operatorname{E}\left[\overline{X}_n\right] = \mu$$

This implies:

$$\operatorname{E}\left[\left(\overline{X}_n - \mu\right)^2\right] = \operatorname{E}\left[\left(\overline{X}_n - \operatorname{E}\left[\overline{X}_n\right]\right)^2\right] = \operatorname{Var}\left[\overline{X}_n\right] \leq \frac{\overline{\gamma}}{n}$$

Thus, taking limits on both sides, we obtain:

$$\lim_{n\to\infty} \operatorname{E}\left[\left(\overline{X}_n - \mu\right)^2\right] \leq \lim_{n\to\infty} \frac{\overline{\gamma}}{n} = 0$$

But

$$\operatorname{E}\left[\left(\overline{X}_n - \mu\right)^2\right] \geq 0$$

so it must be:

$$\lim_{n\to\infty} \operatorname{E}\left[\left(\overline{X}_n - \mu\right)^2\right] = 0$$

This is just the definition of mean square convergence of $\overline{X}_n$ to μ. ■

Hence, also in Chebyshev's Weak Law of Large Numbers for correlated sequences, convergence in probability descends from the fact that convergence in mean square implies convergence in probability.

67.2 Strong Laws of Large numbers

67.2.1 Kolmogorov's SLLN

Among the Strong Laws of Large Numbers, Kolmogorov's is probably the best known:

Proposition 325 (Kolmogorov's SLLN) *Let $\{X_n\}$ be an IID sequence[15] of random variables having finite mean:*

$$\operatorname{E}[X_n] = \mu < \infty, \forall n \in \mathbb{N}$$

Then, a Strong Law of Large Numbers applies to the sample mean:

$$\overline{X}_n \overset{a.s.}{\to} \mu$$

where $\overset{a.s.}{\to}$ denotes almost sure convergence[16].

[15] See p. 492.
[16] See p. 505.

Proof. See, for example, Resnick[17] (1999) and Williams[18] (1991). ■

67.2.2 Ergodic theorem

In Kolmogorov's Strong Law of Large Numbers, the sequence $\{X_n\}$ is required to be an IID sequence. This requirement can be weakened, by requiring $\{X_n\}$ to be stationary[19] and ergodic[20].

Proposition 326 (Ergodic Theorem) *Let $\{X_n\}$ be a stationary and ergodic sequence of random variables having finite mean:*

$$\mathrm{E}[X_n] = \mu < \infty, \forall n \in \mathbb{N}$$

Then, a Strong Law of Large Numbers applies to the sample mean:

$$\overline{X}_n \xrightarrow{a.s.} \mu$$

Proof. See, for example, Karlin and Taylor[21] (1975) and White[22] (2001). ■

67.3 Laws of Large numbers for random vectors

The Laws of Large Numbers we have just presented concern sequences of random variables. However, they can be extended in a straightforward manner to sequences of random vectors:

Proposition 327 *Let $\{X_n\}$ be a sequence of $K \times 1$ random vectors, let $\mathrm{E}[X_n] = \mu$ be their common expected value and*

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

their sample mean. Denote the j-th component of X_n by $X_{n,j}$ and the j-th component of $\overline{X}_n$ by $\overline{X}_{n,j}$. Then:

- *a Weak Law of Large Numbers applies to the sample mean $\overline{X}_n$ if and only if a Weak Law of Large numbers applies to each of the components of the vector $\overline{X}_n$, i.e. if and only if*

$$\operatorname*{plim}_{n\to\infty} \overline{X}_{n,j} = \mu_j, \ j = 1, \ldots, K$$

- *a Strong Law of Large Numbers applies to the sample mean $\overline{X}_n$ if and only if a Strong Law of Large numbers applies to each of the components of the vector $\overline{X}_n$, i.e. if and only if*

$$\overline{X}_{n,j} \xrightarrow{a.s.} \mu_j, \ j = 1, \ldots, K$$

Proof. This is a consequence of the fact that a vector converges in probability (almost surely) if and only if all of its components converge in probability (almost surely). See the lectures entitled *Convergence in probability* (p. 511) and *Almost sure convergence* (p. 505). ■

[17] Resnick, S.I. (1999) "A Probability Path", Birkhauser.
[18] Williams, D. (1991) "Probability with martingales", Cambridge University Press.
[19] See p. 492.
[20] See p. 494.
[21] Karlin, S., Taylor, H. M. (1975) "A first course in stochastic processes", Academic Press.
[22] White, H. (2001) "Asymptotic theory for econometricians", Academic Press.

67.4 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Let $\{\varepsilon_n\}$ be an IID sequence. A generic term of the sequence has mean μ and variance σ^2. Let $\{X_n\}$ be a covariance stationary sequence such that a generic term of the sequence satisfies

$$X_n = \rho X_{n-1} + \varepsilon_n$$

where $-1 < \rho < 1$. Denote by

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_n$$

the sample mean of the sequence. Verify whether the sequence $\left\{\overline{X}_n\right\}$ satisfies the conditions that are required by Chebyshev's Weak Law of Large Numbers. In affirmative case, find its probability limit.

Solution

By assumption the sequence $\{X_n\}$ is covariance stationary. So all the terms of the sequence have the same expected value. Taking the expected value of both sides of the equation

$$X_n = \rho X_{n-1} + \varepsilon_n$$

we obtain:

$$\begin{aligned} \mathrm{E}\left[X_n\right] &= \mathrm{E}\left[\rho X_{n-1} + \varepsilon_n\right] \\ &= \rho \mathrm{E}\left[X_{n-1}\right] + \mathrm{E}\left[\varepsilon_n\right] \\ &= \rho \mathrm{E}\left[X_n\right] + \mu \end{aligned}$$

Solving for $\mathrm{E}\left[X_n\right]$ we obtain:

$$\mathrm{E}\left[X_n\right] = \frac{\mu}{1-\rho}$$

By the same token, the variance can be derived from:

$$\begin{aligned} \mathrm{Var}\left[X_n\right] &= \mathrm{Var}\left[\rho X_{n-1} + \varepsilon_n\right] \\ \boxed{\text{A}} \quad &= \rho^2 \mathrm{Var}\left[X_{n-1}\right] + \mathrm{Var}\left[\varepsilon_n\right] \\ &= \rho^2 \mathrm{Var}\left[X_n\right] + \sigma^2 \end{aligned}$$

where: in step $\boxed{\text{A}}$ we have used the fact that X_{n-1} is independent of ε_n because $\{\varepsilon_n\}$ is IID. Solving for $\mathrm{Var}\left[X_n\right]$, we obtain

$$\mathrm{Var}\left[X_n\right] = \frac{\sigma^2}{1-\rho^2}$$

Now, we need to derive $\mathrm{Cov}\left[X_n, X_{n+j}\right]$. Note that:

$$X_{n+1} = \rho X_n + \varepsilon_{n+1}$$

$$\begin{aligned}
X_{n+2} &= \rho X_{n+1} + \varepsilon_{n+2} = \rho^2 X_n + \varepsilon_{n+2} + \rho\varepsilon_{n+1} \\
X_{n+3} &= \rho X_{n+2} + \varepsilon_{n+3} = \rho^3 X_n + \varepsilon_{n+3} + \rho\varepsilon_{n+2} + \rho^2\varepsilon_{n+1} \\
&\vdots \\
X_{n+j} &= \rho X_{n+j-1} + \varepsilon_{n+j} = \rho^j X_n + \sum_{s=0}^{j-1} \rho^s \varepsilon_{n+j-s}
\end{aligned}$$

The covariance between two terms of the sequence is:

$$\begin{aligned}
\gamma_j &= \operatorname{Cov}[X_n, X_{n+j}] \\
&= \operatorname{Cov}\left[X_n, \rho^j X_n + \sum_{s=0}^{j-1} \rho^s \varepsilon_{n+j-s}\right] \\
\boxed{\text{A}} &= \rho^j \operatorname{Cov}[X_n, X_n] + \sum_{s=0}^{j-1} \rho^s \operatorname{Cov}[X_n, \varepsilon_{n+j-s}] \\
\boxed{\text{B}} &= \rho^j \operatorname{Cov}[X_n, X_n] \\
&= \rho^j \operatorname{Var}[X_n] \\
&= \rho^j \frac{\sigma^2}{1-\rho^2}
\end{aligned}$$

where: in step A we have used the bilinearity of covariance; in step B we have used the fact that X_n is independent of ε_{n+j-s} because $\{\varepsilon_n\}$ is IID. The sum of the covariances is:

$$\begin{aligned}
\sum_{j=0}^{n} \gamma_j &= \sum_{j=0}^{n} \rho^j \frac{\sigma^2}{1-\rho^2} \\
&= \frac{\sigma^2}{1-\rho^2} \sum_{j=0}^{n} \rho^j \\
&= \frac{\sigma^2}{1-\rho^2} \frac{1-\rho}{1-\rho} \sum_{j=0}^{n} \rho^j \\
&= \frac{\sigma^2}{1-\rho^2} \frac{1}{1-\rho} \left(1 - \rho + \rho - \rho^2 + \ldots + \rho^n - \rho^{n+1}\right) \\
&= \frac{\sigma^2}{1-\rho^2} \frac{1-\rho^{n+1}}{1-\rho}
\end{aligned}$$

Thus, covariances tend to be zero on average:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n} \gamma_j = \lim_{n\to\infty} \frac{1}{n} \frac{\sigma^2}{1-\rho^2} \frac{1-\rho^{n+1}}{1-\rho} = 0$$

and the conditions of Chebyshev's Weak Law of Large Numbers are satisfied. Therefore, the sample mean converges in probability to the population mean:

$$\operatorname*{plim}_{n\to\infty} \overline{X}_n = \operatorname{E}[X_n] = \frac{\mu}{1-\rho}$$

Chapter 68

Central Limit Theorems

Let $\{X_n\}$ be a sequence of random variables[1]. Let $\overline{X}_n$ be the sample mean of the first n terms of the sequence:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

A **Central Limit Theorem** (**CLT**) is a proposition giving a set of conditions that are sufficient to guarantee the convergence of the sample mean $\overline{X}_n$ to a normal distribution, as the sample size n increases.

More precisely, a Central Limit Theorem is a proposition giving a set of conditions that are sufficient to guarantee that

$$\sqrt{n}\left(\frac{\overline{X}_n - \mu}{\sigma}\right) \xrightarrow{d} Z$$

where Z is a standard normal random variable[2], μ and σ are two constants and $\xrightarrow{d}$ indicates convergence in distribution[3].

Why is the expression $\left(\overline{X}_n - \mu\right)/\sigma$ multiplied by the square root of n? If we do not multiply it by $\sqrt{n}$, then $\left(\overline{X}_n - \mu\right)/\sigma$ converges to a constant, provided that the conditions of a Law of Large Numbers[4] apply. On the contrary, multiplying it by $\sqrt{n}$, we obtain a sequence that converges to a proper random variable (i.e. a random variable that is not constant). When the conditions of a Central Limit Theorem apply, this variable has a normal distribution.

In practice, the CLT is used as follows:

1. we observe a sample consisting of n observations X_1, X_2, ..., X_n;
2. if n is large enough, then a standard normal distribution is a good approximation of the distribution of $\sqrt{n}\left(\overline{X}_n - \mu\right)/\sigma$;
3. therefore, we pretend that

$$\sqrt{n}\left(\overline{X}_n - \mu\right)/\sigma \sim N(0,1)$$

[1] See p. 491.

[2] Remember that a standard normal random variable is a normal random variable with zero mean and unit variance (p. 376).

[3] See p. 527.

[4] See p. 535.

where $\sim N$ indicates the normal distribution;

4. as a consequence, the distribution of the sample mean $\overline{X}_n$ is

$$\overline{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

There are several Central Limit Theorems. We report some examples below.

68.1 Examples of Central Limit Theorems

68.1.1 Lindeberg-Lévy CLT

The best known Central Limit Theorem is probably Lindeberg-Lévy CLT:

Proposition 328 (Lindeberg-Lévy CLT) *Let $\{X_n\}$ be an IID sequence[5] of random variables such that*

$$\begin{aligned} &\mathrm{E}\left[X_n\right] = \mu < \infty, \forall n \in \mathbb{N} \\ &\mathrm{Var}\left[X_n\right] = \sigma^2 < \infty, \forall n \in \mathbb{N} \end{aligned}$$

where $\sigma^2 > 0$. Then, a Central Limit Theorem applies to the sample mean $\overline{X}_n$:

$$\sqrt{n}\left(\frac{\overline{X}_n - \mu}{\sigma}\right) \xrightarrow{d} Z$$

where Z is a standard normal random variable and $\xrightarrow{d}$ denotes convergence in distribution.

Proof. We will just sketch a proof. For a detailed and rigorous proof see, for example, Resnick[6] (1999) and Williams[7] (1991). First of all, denote by $\{Z_n\}$ the sequence whose generic term is

$$Z_n = \sqrt{n}\left(\frac{\overline{X}_n - \mu}{\sigma}\right)$$

The characteristic function[8] of Z_n is

$$\begin{aligned} \varphi_{Z_n}(t) &= \mathrm{E}\left[\exp\left(itZ_n\right)\right] \\ &= \mathrm{E}\left[\exp\left(it\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma}\right)\right] \\ &= \mathrm{E}\left[\exp\left(it\sqrt{n}\frac{1}{\sigma}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right)\right)\right] \\ &= \mathrm{E}\left[\exp\left(i\frac{t}{\sqrt{n}}\sum_{i=1}^{n}\frac{X_i - \mu}{\sigma}\right)\right] \end{aligned}$$

[5] See p. 492.
[6] Resnick, S.I. (1999) "A Probability Path", Birkhauser.
[7] Williams, D. (1991) "Probability with martingales", Cambridge University Press.
[8] See p. 307.

$$\begin{aligned}
&= \mathrm{E}\left[\prod_{i=1}^{n}\exp\left(i\frac{t}{\sqrt{n}}\frac{X_i-\mu}{\sigma}\right)\right] \\
\boxed{\mathrm{A}} \quad &= \prod_{i=1}^{n}\mathrm{E}\left[\exp\left(i\frac{t}{\sqrt{n}}\frac{X_i-\mu}{\sigma}\right)\right] \\
\boxed{\mathrm{B}} \quad &= \prod_{i=1}^{n}\mathrm{E}\left[\exp\left(i\frac{t}{\sqrt{n}}Y_i\right)\right] \\
\boxed{\mathrm{C}} \quad &= \prod_{i=1}^{n}\varphi_{Y_i}\left(\frac{t}{\sqrt{n}}\right) \\
\boxed{\mathrm{D}} \quad &= \left[\varphi_{Y_1}\left(\frac{t}{\sqrt{n}}\right)\right]^n
\end{aligned}$$

where: in step $\boxed{\mathrm{A}}$ we have used the fact that the random variables X_i are mutually independent[9]; in step $\boxed{\mathrm{B}}$ we have defined

$$Y_i = \frac{X_i-\mu}{\sigma}$$

in step $\boxed{\mathrm{C}}$ we have used the definition of characteristic function and we have denoted the characteristic function of Y_i by $\varphi_{Y_i}(t)$; in step $\boxed{\mathrm{D}}$ we have used the fact that all the variables Y_i have the same distribution and hence the same characteristic function. Now take a second order Taylor series expansion of $\varphi_{Y_1}(s)$ around the point $s=0$:

$$\begin{aligned}
\varphi_{Y_1}(s) &= \mathrm{E}\left[\exp\left(isY_1\right)\right] \\
&= \mathrm{E}\left[\exp\left(isY_1\right)\right]\big|_{s=0} + \frac{d}{ds}\left(\mathrm{E}\left[\exp\left(isY_1\right)\right]\right)\Big|_{s=0} s \\
&\quad + \frac{1}{2}\frac{d^2}{ds^2}\left(\mathrm{E}\left[\exp\left(isY_1\right)\right]\right)\Big|_{s=0} s^2 + o\left(s^2\right) \\
&= \mathrm{E}\left[\exp\left(isY_1\right)\right]\big|_{s=0} + \left(\mathrm{E}\left[\frac{d}{ds}\exp\left(isY_1\right)\right]\right)\Big|_{s=0} s \\
&\quad + \frac{1}{2}\left(\mathrm{E}\left[\frac{d^2}{ds^2}\exp\left(isY_1\right)\right]\right)\Big|_{s=0} s^2 + o\left(s^2\right) \\
&= \mathrm{E}\left[\exp\left(isY_1\right)\right]\big|_{s=0} + \left(\mathrm{E}\left[iY_1\exp\left(isY_1\right)\right]\right)\big|_{s=0} s \\
&\quad + \frac{1}{2}\left(\mathrm{E}\left[-Y_1^2\exp\left(isY_1\right)\right]\right)\big|_{s=0} s^2 + o\left(s^2\right) \\
&= 1 + i\mathrm{E}\left[Y_1\right]s - \frac{1}{2}\mathrm{E}\left[Y_1^2\right]s^2 + o\left(s^2\right) \\
\boxed{\mathrm{A}} \quad &= 1 - \frac{1}{2}\mathrm{Var}\left[Y_1\right]s^2 + o\left(s^2\right) \\
\boxed{\mathrm{B}} \quad &= 1 - \frac{1}{2}s^2 + o\left(s^2\right)
\end{aligned}$$

where: $o\left(s^2\right)$ is an infinitesimal of higher order than s^2, i.e. a quantity that converges to 0 faster than s^2 does; in step $\boxed{\mathrm{A}}$ we have used the fact that

$$\mathrm{E}\left[Y_1\right] = 0$$

[9] In particular, see the *Mutual independence via expectations* property (p. 234).

in step B we have used the fact that

$$\operatorname{Var}[Y_1] = 1$$

Therefore:

$$\begin{aligned}
\lim_{n\to\infty} \varphi_{Z_n}(t) &= \lim_{n\to\infty} \left[\varphi_{Y_1}\left(\frac{t}{\sqrt{n}}\right)\right]^n \\
&= \lim_{n\to\infty} \left[1 - \frac{1}{2}\left(\frac{t}{\sqrt{n}}\right)^2 + o\left(\left(\frac{t}{\sqrt{n}}\right)^2\right)\right]^n \\
&= \lim_{n\to\infty} \left[1 - \frac{1}{2}\frac{t^2}{n} + o\left(\frac{t^2}{n}\right)\right]^n \\
&= \exp\left(-\frac{1}{2}t^2\right) = \varphi_Z(t)
\end{aligned}$$

So, we have that:

$$\lim_{n\to\infty} \varphi_{Z_n}(t) = \varphi_Z(t)$$

where

$$\varphi_Z(t) = \exp\left(-\frac{1}{2}t^2\right)$$

is the characteristic function of a standard normal random variable Z (see the lecture entitled *Normal distribution* - p. 379). A theorem, called Lévy continuity theorem, which we do not cover in these lectures, states that if a sequence of random variables $\{Z_n\}$ is such that their characteristic functions $\varphi_{Z_n}(t)$ converge to the characteristic function $\varphi_Z(t)$ of a random variable Z, then the sequence $\{Z_n\}$ converges in distribution to Z. Therefore, in our case the sequence $\{Z_n\}$ converges in distribution to a standard normal distribution. ■

So, roughly speaking, under the stated assumptions, the distribution of the sample mean $\overline{X}_n$ can be approximated by a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$ (provided n is large enough).

Also note that the conditions for the validity of Lindeberg-Lévy Central Limit Theorem resemble the conditions for the validity of Kolmogorov's Strong Law of Large Numbers[10]. The only difference is the additional requirement that

$$\operatorname{Var}[X_n] = \sigma^2 < \infty, \forall n \in \mathbb{N}$$

68.1.2 A CLT for correlated sequences

In the Lindeberg-Lévy CLT (see above), the sequence $\{X_n\}$ is required to be an IID sequence. The assumption of independence can be weakened as follows:

Proposition 329 (CLT for correlated sequences) *Let $\{X_n\}$ be a stationary*[11] *and mixing*[12] *sequence of random variables satisfying a CLT technical condition (defined in the proof below) and such that*

$$\mathrm{E}[X_n] = \mu < \infty, \forall n \in \mathbb{N}$$

[10] See p. 540.
[11] See p. 492.
[12] See p. 494.

$$\begin{aligned}&\text{Var}\left[X_n\right]=\sigma^2<\infty,\forall n\in\mathbb{N}\\&\lim_{n\to\infty}n\text{Var}\left[\overline{X}_n\right]=\sigma^2+2\sum_{i=2}^{\infty}\text{Cov}\left[X_1,X_i\right]=V<\infty\end{aligned}$$

where $V>0$. Then, a Central Limit Theorem applies to the sample mean $\overline{X}_n$:

$$\sqrt{n}\left(\frac{\overline{X}_n-\mu}{\sqrt{V}}\right)\xrightarrow{d}Z$$

where Z is a standard normal random variable and $\xrightarrow{d}$ denotes convergence in distribution.

Proof. Several different technical conditions (beyond those explicitly stated in the above proposition) are imposed in the literature in order to derive Central Limit Theorems for correlated sequences. These conditions are usually very mild and differ from author to author. We do not mention these technical conditions here and just refer to them as **CLT technical conditions**. For a proof see, for example, Durrett[13] (2010) and White[14] (2001). ■

So, roughly speaking, under the stated assumptions, the distribution of the sample mean $\overline{X}_n$ can be approximated by a normal distribution with mean μ and variance $\frac{V}{n}$ (provided n is large enough).

Also note that the conditions for the validity of the Central Limit Theorem for correlated sequences resemble the conditions for the validity of the ergodic theorem[15]. The main differences (beyond some technical conditions that are not explicitly stated in the above proposition) are the additional requirements that

$$\begin{aligned}&\text{Var}\left[X_n\right]=\sigma^2<\infty,\forall n\in\mathbb{N}\\&V=\lim_{n\to\infty}n\text{Var}\left[\overline{X}_n\right]=\sigma^2+2\sum_{i=2}^{\infty}\text{Cov}\left[X_1,X_i\right]<\infty\end{aligned}$$

and the fact that ergodicity is replaced by the stronger condition of mixing.

Finally, let us mention that the variance V in the above proposition, which is defined as

$$V=\lim_{n\to\infty}n\text{Var}\left[\overline{X}_n\right]$$

is called the **long-run variance** of $\overline{X}_n$.

68.2 Multivariate generalizations

The results illustrated above for sequences of random variables extend in a straightforward manner to sequences of random vectors. For example, the multivariate version of the Lindeberg-Lévy CLT is:

Proposition 330 (Multivariate Lindeberg-Lévy CLT) *Let $\{X_n\}$ be an IID sequence of $K\times 1$ random vectors such that*

$$\text{E}\left[X_n\right]=\mu\in\mathbb{R}^K,\forall n\in\mathbb{N}$$

[13] Durrett, R. (2010) "Probability: Theory and Examples", Cambridge University Press.
[14] White, H. (2001) "Asymptotic theory for econometricians", Academic Press.
[15] See p. 541.

$$\text{Var}\left[X_n\right] = \Sigma \in \mathbb{R}^{K\times K}, \forall n \in \mathbb{N}$$

where Σ is a positive definite matrix. Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the vector of sample means. Then:

$$\sqrt{n}\Sigma^{-1}\left(\overline{X}_n - \mu\right) \xrightarrow{d} Z$$

where Z is a standard multivariate normal random vector[16] *and $\xrightarrow{d}$ denotes convergence in distribution.*

Proof. For a proof see, for example, Basu[17] (2004), DasGupta[18] (2008) and McCabe and Tremayne[19] (1993). ■

In a similar manner, the CLT for correlated sequences generalizes to random vectors (V becomes a matrix, called long-run covariance matrix).

68.3 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Let $\{X_n\}$ be a sequence of independent Bernoulli random variables[20] with parameter $p = \frac{1}{2}$, i.e. a generic term X_n of the sequence has support

$$R_{X_n} = \{0, 1\}$$

and probability mass function

$$p_{X_n}(x) = \begin{cases} 1/2 & \text{if } x = 1 \\ 1/2 & \text{if } x = 0 \\ 0 & \text{if } x \notin R_{X_n} \end{cases}$$

Use a Central Limit Theorem to derive an approximate distribution for the mean of the first 100 terms of the sequence.

Solution

The sequence $\{X_n\}$ is and IID sequence. The mean of a generic term of the sequence is

$$\begin{aligned} \text{E}\left[X_n\right] &= \sum_{x\in R_{X_n}} x p_{X_n}(x) = 1 \cdot p_{X_n}(1) + 0 \cdot p_{X_n}(0) \\ &= 1 \cdot \frac{1}{2} + 0 \cdot \left(1 - \frac{1}{2}\right) = \frac{1}{2} < \infty \end{aligned}$$

[16] See p. 439.
[17] Basu, A. K. (2004) Measure theory and probability, PHI Learning PVT.
[18] DasGupta, A. (2008) Asymptotic theory of statistics and probability, Springer.
[19] McCabe, B. and A. Tremayne (1993) Elements of modern asymptotic theory with statistical applications, Manchester University Press.
[20] See p. 335.

The variance of a generic term of the sequence can be derived thanks to the usual formula for computing the variance[21]:

$$\begin{aligned}
\mathrm{E}\left[X_n^2\right] &= \sum_{x \in R_{X_n}} x^2 p_X(x) = 1^2 \cdot p_{X_n}(1) + 0^2 \cdot p_{X_n}(0) \\
&= 1 \cdot \frac{1}{2} + 0 \cdot \frac{1}{2} = \frac{1}{2} \\
\mathrm{E}\left[X_n\right]^2 &= \frac{1}{4} \\
\mathrm{Var}\left[X_n\right] &= \mathrm{E}\left[X_n^2\right] - \mathrm{E}\left[X_n\right]^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4} < \infty
\end{aligned}$$

Therefore, the sequence $\{X_n\}$ satisfies the conditions of Lindeberg-Lévy Central Limit Theorem (IID, finite mean, finite variance). The mean of the first 100 terms of the sequence is:

$$\overline{X}_{100} = \frac{1}{100} \sum_{i=1}^{100} X_i$$

Using the Central Limit Theorem to approximate its distribution, we obtain:

$$\overline{X}_n \sim N\left(\mathrm{E}\left[X_n\right], \frac{\mathrm{Var}\left[X_n\right]}{n}\right)$$

or

$$\overline{X}_{100} \sim N\left(\frac{1}{2}, \frac{1}{400}\right)$$

Exercise 2

Let $\{X_n\}$ be a sequence of independent Bernoulli random variables with parameter $p = \frac{1}{2}$, as in the previous exercise. Let $\{Y_n\}$ be another sequence of random variables such that

$$Y_n = X_{n+1} - \frac{1}{2} X_n \quad , \forall n$$

Suppose $\{Y_n\}$ satisfies the conditions of a Central Limit Theorem for correlated sequences. Derive an approximate distribution for the mean of the first n terms of the sequence $\{Y_n\}$.

Solution

The sequence $\{X_n\}$ is and IID sequence. The mean of a generic term of the sequence is

$$\begin{aligned}
\mathrm{E}\left[Y_n\right] &= \mathrm{E}\left[X_{n+1} - \frac{1}{2} X_n\right] = \mathrm{E}\left[X_{n+1}\right] - \frac{1}{2}\mathrm{E}\left[X_n\right] \\
&= \frac{1}{2} - \frac{1}{2}\frac{1}{2} = \frac{1}{4}
\end{aligned}$$

The variance of a generic term of the sequence is

$$\mathrm{Var}\left[Y_n\right] = \mathrm{Var}\left[X_{n+1} - \frac{1}{2} X_n\right]$$

[21] $\mathrm{Var}[X] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2$. See p. 156.

$$\begin{aligned} & = \operatorname{Var}[X_{n+1}] + \frac{1}{4}\operatorname{Var}[X_n] - 2\frac{1}{2}\operatorname{Cov}[X_{n+1}, X_n] \\ \boxed{\text{A}} \quad & = \operatorname{Var}[X_{n+1}] + \frac{1}{4}\operatorname{Var}[X_n] \\ & = \frac{1}{4} + \frac{1}{4}\frac{1}{4} = \frac{5}{16} \end{aligned}$$

where: in step A we have used the fact that X_n and X_{n+1} are independent. The covariance between two successive terms of the sequence is

$$\begin{aligned} & \operatorname{Cov}[Y_{n+1}, Y_n] \\ & = \operatorname{Cov}\left[X_{n+2} - \frac{1}{2}X_{n+1}, X_{n+1} - \frac{1}{2}X_n\right] \\ \boxed{\text{A}} \quad & = \operatorname{Cov}[X_{n+2}, X_{n+1}] - \frac{1}{2}\operatorname{Cov}[X_{n+2}, X_n] \\ & \quad -\frac{1}{2}\operatorname{Cov}[X_{n+1}, X_{n+1}] + \frac{1}{4}\operatorname{Cov}[X_{n+1}, X_n] \\ \boxed{\text{B}} \quad & = -\frac{1}{2}\operatorname{Cov}[X_{n+1}, X_{n+1}] \\ \boxed{\text{C}} \quad & = -\frac{1}{2}\operatorname{Var}[X_{n+1}] \\ & = -\frac{1}{2}\frac{1}{4} = -\frac{1}{8} \end{aligned}$$

where: in step A we have used the bilinearity of covariance[22]; in step B we have used the fact that the terms of $\{X_n\}$ are independent; in step C we have used the fact that the covariance of a random variable with itself is equal to its variance. The covariance between two terms that are not adjacent (Y_n and Y_{n+j}, with $j > 1$) is

$$\begin{aligned} & \operatorname{Cov}[Y_{n+j}, Y_n] \\ & = \operatorname{Cov}\left[X_{n+j+1} - \frac{1}{2}X_{n+j}, X_{n+1} - \frac{1}{2}X_n\right] \\ \boxed{\text{A}} \quad & = \operatorname{Cov}[X_{n+j+1}, X_{n+1}] - \frac{1}{2}\operatorname{Cov}[X_{n+j+1}, X_n] \\ & \quad -\frac{1}{2}\operatorname{Cov}[X_{n+j}, X_{n+1}] + \frac{1}{4}\operatorname{Cov}[X_{n+j}, X_n] \\ \boxed{\text{B}} \quad & = 0 \end{aligned}$$

where: in step A we have used the bilinearity of covariance; in step B we have used the fact that the terms of $\{X_n\}$ are independent. The long-run variance is

$$\begin{aligned} V & = \operatorname{Var}[Y_1] + 2\sum_{j=2}^{\infty}\operatorname{Cov}[X_j, X_1] \\ & = \operatorname{Var}[Y_1] + 2\operatorname{Cov}[X_2, X_1] \\ & = \frac{5}{16} - \frac{2}{8} = \frac{1}{16} \end{aligned}$$

[22] See p. 166.

The mean of the first n terms of the sequence $\{Y_n\}$ is

$$\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Using the Central Limit Theorem for correlated sequences to approximate its distribution, we obtain

$$\overline{Y}_n \sim N\left(\mathrm{E}\left[Y_n\right], \frac{V}{n}\right)$$

or

$$\overline{Y}_n \sim N\left(\frac{1}{4}, \frac{1}{16 \cdot n}\right)$$

Exercise 3

Let Y be a binomial random variable with parameters $n = 100$ and $p = \frac{1}{2}$ (you need to read the lecture entitled *Binomial distribution*[23] in order to be able to solve this exercise). Using the Central Limit Theorem, show that a normal random variable X with mean $\mu = 50$ and variance $\sigma^2 = 25$ can be used as an approximation of Y.

Solution

A binomial random variable Y with parameters $n = 100$ and $p = \frac{1}{2}$ can be written as

$$Y = \sum_{i=1}^{100} X_i$$

where $X_1, \ldots, X_{100}$ are mutually independent Bernoulli random variables with parameter $p = \frac{1}{2}$. Thus:

$$\begin{aligned} Y &= 100\left(\frac{1}{100}\sum_{i=1}^{100} X_i\right) \\ &= 100 \cdot \overline{X}_{100} \end{aligned}$$

In the first exercise, we have shown that the distribution of $\overline{X}_{100}$ can be approximated by a normal distribution:

$$\overline{X}_{100} \sim N\left(\frac{1}{2}, \frac{1}{400}\right)$$

Therefore, the distribution of Y can be approximated by

$$Y \sim N\left(\frac{1}{2} \cdot 100, \frac{1}{400} \cdot 100^2\right)$$

Thus, Y can be approximated by a normal distribution with mean $\mu = 50$ and variance $\sigma^2 = 25$.

[23] See p. 341.

Chapter 69

Convergence of transformations

This lecture discusses some well-known results on the convergence of transformed sequences of random vectors.

69.1 Continuous mapping theorem

Suppose a sequence of random vectors $\{X_n\}$ converges to a random vector X (either in probability, in distribution or almost surely). Now, take a transformed sequence $\{g(X_n)\}$, where g is a function. Under what conditions is $\{g(X_n)\}$ also a convergent sequence?

The following proposition, known as Continuous Mapping Theorem, states that convergence is preserved by continuous transformations.

Proposition 331 *Let $\{X_n\}$ be a sequence of K-dimensional random vectors. Let $g : \mathbb{R}^K \to \mathbb{R}^L$ be a continuous function. Then:*

$$X_n \xrightarrow{\text{P}} X \Longrightarrow g(X_n) \xrightarrow{\text{P}} g(X)$$
$$X_n \xrightarrow{a.s.} X \Longrightarrow g(X_n) \xrightarrow{a.s.} g(X)$$
$$X_n \xrightarrow{d} X \Longrightarrow g(X_n) \xrightarrow{d} g(X)$$

where $\xrightarrow{\text{P}}$ denotes convergence in probability, $\xrightarrow{a.s.}$ denotes almost sure convergence and $\xrightarrow{d}$ denotes convergence in distribution.

Proof. See e.g. Shao[1] (2003). ■

The following subsections present some important applications of the continuous mapping theorem.

69.1.1 Convergence in probability of sums and products

An important implication of the continuous mapping theorem is that arithmetic operations preserve convergence in probability.

[1] Shao, J. (2003) *Mathematical statistics*, Springer.

Proposition 332 *If* $X_n \xrightarrow{\text{P}} X$ *and* $Y_n \xrightarrow{\text{P}} Y$*. Then:*

$$\begin{aligned} X_n + Y_n &\xrightarrow{\text{P}} X + Y \\ X_n Y_n &\xrightarrow{\text{P}} XY \end{aligned}$$

Proof. First of all, note that convergence in probability of $\{X_n\}$ and of $\{Y_n\}$ implies their joint convergence in probability[2], i.e. their convergence as a vector:

$$[X_n \;\; Y_n] \xrightarrow{\text{P}} [X \;\; Y]$$

Now, the sum and the product are continuous functions of the operands. Thus, for example:

$$X + Y = g\left([X \;\; Y]\right)$$

where g is a continuous function, and, using the continuous mapping theorem:

$$\begin{aligned} \operatorname*{plim}_{n\to\infty} (X_n + Y_n) &= \operatorname*{plim}_{n\to\infty} g\left([X_n \;\; Y_n]\right) \\ &= g\left(\operatorname*{plim}_{n\to\infty} [X_n \;\; Y_n]\right) \\ &= g\left([X \;\; Y]\right) \\ &= X + Y \end{aligned}$$

where plim denotes a limit in probability. ■

69.1.2 Almost sure convergence of sums and products

Everything that was said in the previous subsection applies, with obvious modifications, also to almost surely convergent sequences:

Proposition 333 *If* $X_n \xrightarrow{a.s.} X$ *and* $Y_n \xrightarrow{a.s.} Y$*, then:*

$$\begin{aligned} X_n + Y_n &\xrightarrow{a.s.} X + Y \\ X_n Y_n &\xrightarrow{a.s.} XY \end{aligned}$$

Proof. Similar to previous proof. Just replace convergence in probability with almost sure convergence. ■

69.1.3 Convergence in distribution of sums and products

For convergence almost surely and convergence in probability, the convergence of $\{X_n\}$ and $\{Y_n\}$ individually implies their joint convergence as a vector (see the previous two proofs), but this is not the case for convergence in distribution. Therefore, to obtain preservation of convergence in distribution under arithmetic operations, we need the stronger assumption of joint convergence in distribution:

Proposition 334 *If*

$$[X_n \;\; Y_n] \xrightarrow{d} [X \;\; Y]$$

then:

$$\begin{aligned} X_n + Y_n &\xrightarrow{d} X + Y \\ X_n Y_n &\xrightarrow{d} XY \end{aligned}$$

Proof. Again, similar to the proof for convergence in probability, but this time joint convergence is already in the assumptions. ■

[2] See the lecture entitled *Convergence in probability* (p. 511).

69.2 Slutski's Theorem

Slutski's theorem concerns the convergence in distribution of the product of two sequences, one converging in distribution and the other converging in probability to a constant:

Proposition 335 *If $X_n \xrightarrow{d.} X$ and $Y_n \xrightarrow{P.} c$, where c is a constant, then:*

$$\begin{aligned} X_n + Y_n &\xrightarrow{d} X + c \\ X_n Y_n &\xrightarrow{d} cX \end{aligned}$$

Proof. It is possible to prove (see e.g. van der Vaart[3] - 2000) that $X_n \xrightarrow{d.} X$ and $Y_n \xrightarrow{P.} c$ imply

$$[X_n \quad Y_n] \xrightarrow{d} [X \quad c]$$

but this in turn implies convergence under arithmetic operations (see above). ■

69.3 More details

69.3.1 Convergence of ratios

As a byproduct of the propositions stated above, we also have:

Proposition 336 *If a sequence of random variables $\{X_n\}$ converges to X, then*

$$1/X_n \to 1/X$$

provided X is almost surely different from 0 (we did not specify the kind of convergence, which can be in probability, almost surely or in distribution).

Proof. This is a consequence of the Continuous mapping theorem and of the fact that

$$g(x) = 1/x$$

is a continuous function for $x \neq 0$. ■

As a consequence:

Proposition 337 *If two sequences of random variables $\{X_n\}$ and $\{Y_n\}$ converge to X and Y respectively, then*

$$X_n/Y_n \to X/Y$$

provided Y is almost surely different from 0. Convergence can be in probability, almost surely or in distribution (but the latter requires joint convergence in distribution of $\{X_n\}$ and $\{Y_n\}$).

Proof. This is a consequence of the fact that the ratio can be written as a product

$$X_n/Y_n = X_n \cdot (1/Y_n)$$

The first operand of the product converges by assumption. The second converges because of the previous proposition. Therefore, their product converges because convergence is preserved under products. ■

[3] A. W. van der Vaart (2000) *Asymptotic Statistics*, Cambridge University Press.

69.3.2 Random matrices

The continuous mapping theorem and Slutski's theorem apply also to random matrices[4], because random matrices are just random vectors whose elements have been arranged into columns.

In particular:

- if two sequences or random matrices are convergent, then also the sum and the product of their terms are convergent (provided their dimensions are such that they can be summed or multiplied);
- if a sequence of square random matrices $\{X_n\}$ converges to a random matrix X, then the sequence of inverse matrices $\left\{X_n^{-1}\right\}$ converges to the random matrix X^{-1} (provided the matrices are invertible). This is a consequence of the fact that matrix inversion is a continuous transformation.

69.4 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Let $\{X_n\}$ be a sequence of $K \times 1$ random vectors such that

$$X_n \xrightarrow{d} X$$

where X is a normal random vector with mean μ and invertible covariance matrix V.

Let $\{A_n\}$ be a sequence of $L \times K$ random matrices such that:

$$A_n \xrightarrow{\mathrm{P}} A$$

where A is a constant matrix. Find the limit in distribution of the sequence of products $\{A_n X_n\}$.

Solution

By Slutski's theorem

$$A_n X_n \xrightarrow{d} Y$$

where

$$Y = AX$$

The random vector Y has a multivariate normal distribution, because it is a linear transformation of a multivariate normal random vector[5]. The expected value of Y is

$$\mathrm{E}[Y] = \mathrm{E}[AX] = A\mathrm{E}[X] = A\mu$$

and its covariance matrix is

$$\mathrm{Var}[Y] = \mathrm{Var}[AX] = A\mathrm{Var}[X]A^{\intercal} = AVA^{\intercal}$$

Therefore, the sequence of products $\{A_n X_n\}$ converges in distribution to a multivariate normal random vector with mean $A\mu$ and covariance matrix $AVA^{\intercal}$.

[4] See p. 119.
[5] See p. 469.

Exercise 2

Let $\{X_n\}$ be a sequence of $K \times 1$ random vectors such that

$$X_n \xrightarrow{d} X$$

where X is a normal random vector with mean 0 and invertible covariance matrix V.

Let $\{V_n\}$ be a sequence of $K \times K$ random matrices such that

$$V_n \xrightarrow{\mathrm{P}} V$$

Find the limit in distribution of the sequence

$$\left\{X_n^\intercal V_n^{-1} X_n\right\}$$

Solution

By the continuous mapping theorem

$$V_n^{-1} \xrightarrow{\mathrm{P}} V^{-1}$$

Therefore, by Slutski's theorem

$$V_n^{-1} X_n \xrightarrow{d} V^{-1} X$$

Using the continuous mapping theorem again:

$$X_n^\intercal V_n^{-1} X_n = X_n^\intercal \left(V_n^{-1} X_n\right) \xrightarrow{d} X^\intercal \left(V^{-1} X\right) = X^\intercal V^{-1} X$$

Since V is an invertible covariance matrix, there exists an invertible matrix Σ such that

$$V = \Sigma\Sigma^\intercal$$

Therefore

$$\begin{aligned} X^\intercal V^{-1} X &= X^\intercal \left(\Sigma\Sigma^\intercal\right)^{-1} X \\ &= X^\intercal \left(\Sigma^\intercal\right)^{-1} \Sigma^{-1} X \\ &= \left(\Sigma^{-1} X\right)^\intercal \Sigma^{-1} X \\ &= Z^\intercal Z \end{aligned}$$

where we have defined

$$Z = \Sigma^{-1} X$$

The random vector Z has a multivariate normal distribution, because it is a linear transformation of a multivariate normal random vector. The expected value of Z is

$$\mathrm{E}\left[Z\right] = \mathrm{E}\left[\Sigma^{-1} X\right] = \Sigma^{-1} \mathrm{E}\left[X\right] = \Sigma^{-1} 0 = 0$$

and its covariance matrix is

$$\begin{aligned} \mathrm{Var}\left[Z\right] &= \mathrm{Var}\left[\Sigma^{-1} X\right] \\ &= \Sigma^{-1} \mathrm{Var}\left[X\right] \left(\Sigma^{-1}\right)^\intercal \end{aligned}$$

$$\begin{aligned} &= \Sigma^{-1}\Sigma\Sigma^{\intercal}\left(\Sigma^{-1}\right)^{\intercal} \\ &= \Sigma^{\intercal}\left(\Sigma^{-1}\right)^{\intercal} \\ &= \Sigma^{\intercal}\left(\Sigma^{\intercal}\right)^{-1} = I \end{aligned}$$

Thus, Z has a standard multivariate normal distribution (mean 0 and variance I) and

$$X^{\intercal}V^{-1}X = Z^{\intercal}Z = Z^{\intercal}IZ$$

is a quadratic form in a standard normal random vector[6]. So, $X^{\intercal}V^{-1}X$ has a Chi-square distribution with $n = \operatorname{tr}(I)$ degrees of freedom. Summing up, the sequence $\left\{X_n^{\intercal}V_n^{-1}X_n\right\}$ converges in distribution to a Chi-square distribution with n degrees of freedom.

Exercise 3

Let everything be as in the previous exercise, except for the fact that now X has mean μ. Find the limit in distribution of the sequence

$$\left\{(X_n - \mu_n)^{\intercal}V_n^{-1}(X_n - \mu_n)\right\}$$

where $\{\mu_n\}$ is a sequence of $K \times 1$ random vectors converging in probability to μ.

Solution

Define

$$Y_n = X_n - \mu_n$$

By Slutski's theorem

$$Y_n \xrightarrow{d} Y$$

where

$$Y = X - \mu$$

is a multivariate normal random variable with mean 0 and variance V. Thus, we can use the results of the previous exercise on the sequence

$$\left\{Y_n^{\intercal}V_n^{-1}Y_n\right\}$$

which is the same as

$$\left\{(X_n - \mu_n)^{\intercal}V_n^{-1}(X_n - \mu_n)\right\}$$

and we find that it converges in distribution to a Chi-square distribution with n degrees of freedom.

[6] See p. 481.

Part VII

Fundamentals of statistics

Chapter 70

Statistical inference

Statistical inference is the act of using **observed data** to infer unknown properties and characteristics of the probability distribution from which the observed data have been generated. The set of data that is used to make inferences is called **sample**.

70.1 Samples

In the simplest possible case, we observe the realizations[1] $x_1, \ldots, x_n$ of n independent random variables[2] $X_1, \ldots, X_n$ having a common distribution function[3] $F_X(x)$ and we use the observed realizations to **infer** some characteristics of $F_X(x)$. With a slight abuse of language, we sometimes say "n **independent realizations** of a random variable X" instead of saying "the realizations of n independent random variables $X_1, \ldots, X_n$ having a common distribution function $F_X(x)$".

Example 338 *The lifetime of a certain type of electronic device is a random variable X, whose distribution function $F_X(x)$ is unknown. Suppose we independently observe the lifetimes of* 10 *components. Denote these realizations by x_1, x_2, ..., x_{10}. We are interested in the expected value of X, which is an unknown characteristic of $F_X(x)$. We infer $\mathrm{E}[X]$ from the data, estimating $\mathrm{E}[X]$ with the sample mean*

$$\overline{x} = \frac{1}{10}\sum_{i=1}^{10} x_i$$

In this simple example the observed data x_1, x_2, ..., x_{10} constitute our sample and $\mathrm{E}[X]$ is the quantity about which we are making a statistical inference.

While in the simplest case $X_1, \ldots, X_n$ are independent random variables, more complicated cases are possible. For example:

1. $X_1, \ldots, X_n$ are not independent;
2. $X_1, \ldots, X_n$ are random vectors having a common joint distribution function[4] $F_X(x)$;

[1] See p. 105.
[2] See p. 229.
[3] See p. 108.
[4] See p. 118.

3. $X_1, \ldots, X_n$ do not have a common probability distribution.

Is there a definition of sample that generalizes all of the above special cases? Fortunately, there is one and it is extremely simple:

Definition 339 *A **sample** ξ is the realization of a random vector Ξ.*

The distribution function of Ξ, denoted by $F_{\Xi}(\xi)$, is the **unknown distribution function** that constitutes the object of inference.

Therefore, "sample" is just a synonym of "realization of a random vector". The following examples show how this general definition accommodates the special cases mentioned above.

Example 340 *When we observe n realizations $x_1, \ldots, x_n$ of n independent random variables $X_1, \ldots, X_n$ having a common distribution function $F_X(x)$, the sample is the n-dimensional vector*

$$\xi = [x_1 \quad \ldots \quad x_n] \tag{70.1}$$

which is a realization of the random vector

$$\Xi = [X_1 \quad \ldots \quad X_n] \tag{70.2}$$

The joint distribution function of Ξ is:

$$F_{\Xi}(\xi) = F_X(x_1) \cdot \ldots \cdot F_X(x_n)$$

Example 341 *When we observe n realizations $x_1, \ldots, x_n$ of n random variables $X_1, \ldots, X_n$ that are not independent but have a common distribution function $F_X(x)$, the sample is again the n-dimensional vector (70.1), which is a realization of the random vector (70.2). However, in this case the joint distribution function $F_{\Xi}(\xi)$ can no longer be written as the product of the distribution functions of $X_1, \ldots, X_n$.*

Example 342 *When we observe n realizations $x_1, \ldots, x_n$ of n independent $K \times 1$ random vectors $X_1, \ldots, X_n$ having a common joint distribution function $F_X(x)$, the sample is the $nK \times 1$ vector*

$$\xi = [x_1^{\intercal} \quad \ldots \quad x_n^{\intercal}]^{\intercal} \tag{70.3}$$

which is a realization of the random vector

$$\Xi = [X_1^{\intercal} \quad \ldots \quad X_n^{\intercal}] \tag{70.4}$$

The joint distribution function of Ξ is:

$$F_{\Xi}(\xi) = F_X(x_1) \cdot \ldots \cdot F_X(x_n)$$

Example 343 *When we observe n realizations $x_1, \ldots, x_n$ of n independent $K \times 1$ random vectors $X_1, \ldots, X_n$ having different joint distribution functions $F_{X_1}(x_1), \ldots, F_{X_n}(x_n)$, the sample is the $nK \times 1$ vector (70.3), which is a realization of the random vector (70.4). The joint distribution function of Ξ is:*

$$F_{\Xi}(\xi) = F_{X_1}(x_1) \cdot \ldots \cdot F_{X_n}(x_n)$$

When the sample is made of the realizations of n random variables (or n random vectors), as in (70.1) and (70.3), then we say that the sample has **size** n (or that the **sample size** is n). An individual realization x_i is referred to as an **observation** from the sample.

70.2 Statistical models

In the previous section we have defined a sample ξ as a realization of a random vector Ξ having joint distribution function $F_{\Xi}(\xi)$. The sample ξ is used to infer some characteristics of $F_{\Xi}(\xi)$ that are not fully known by the statistician. The properties and the characteristics of $F_{\Xi}(\xi)$ that are already known (or are assumed to be known) before observing the sample are called a model for Ξ. In mathematical terms, a model for Ξ is a set of joint distribution functions to which $F_{\Xi}(\xi)$ is assumed to belong.

Definition 344 (Statistical model) *Let the sample ξ be a realization of a l-dimensional random vector Ξ having joint distribution function $F_{\Xi}(\xi)$. Let Ψ be the set of all l-dimensional joint distribution functions:*

$$\Psi = \left\{F : \mathbb{R}^l \to \mathbb{R}_+ \text{ such that } F \text{ is a joint distribution function}\right\}$$

*A subset $\Phi \subseteq \Psi$ is called a **statistical model** (or a **model specification** or, simply, a **model**) for Ξ. If $F_{\Xi} \in \Phi$ the model is said to be **correctly specified** (or well-specified). Otherwise, if $F_{\Xi} \notin \Phi$ the model is said to be **mis-specified**.*

Continuing the examples of the previous section:

Example 345 *Suppose our sample is made of n realizations $x_1, \ldots, x_n$ of n random variables $X_1, \ldots, X_n$. Assume that the n random variables are mutually independent and that they have a common distribution function $F_X(x)$. The sample is the n-dimensional vector*

$$\xi = \begin{bmatrix} x_1 & \ldots & x_n \end{bmatrix}$$

Ψ is the set of all possible distribution functions of the random vector

$$\Xi = \begin{bmatrix} X_1 & \ldots & X_n \end{bmatrix}$$

Recalling the definition of marginal distribution function[5] and the characterization of mutual independence[6], the statistical model Φ is defined as follows:

$$\Phi = \left\{ F \in \Psi : \begin{array}{l} \text{all the marginals of } F \text{ are equal and} \\ F \text{ is equal to the product of its marginals} \end{array} \right\}$$

Example 346 *Take the example above and drop the assumption that the n random variables $X_1, \ldots, X_n$ are mutually independent. The statistical model Φ is now:*

$$\Phi = \left\{F \in \Psi : \text{ all the marginals of } F \text{ are equal}\right\}$$

The next subsections introduce some terminology related to model specification.

70.2.1 Parametric models

A model Φ for Ξ is called a parametric model if the joint distribution functions belonging to Φ are put into correspondence with a set Θ of real vectors.

[5] See p. 119.
[6] See p. 235.

Definition 347 (Parametric model) *Let Φ be a model for Ξ. Let $\Theta \subseteq \mathbb{R}^p$ be a set of p-dimensional real vectors. Let $\gamma(\theta)$ be a correspondence that associates a subset of Φ to each $\theta \in \Theta$. The triple (Φ, Θ, γ) is a **parametric model** if and only if*

$$\Phi = \bigcup_{\theta \in \Theta} \gamma(\theta)$$

*The set Θ is called **parameter space**. A vector $\theta \in \Theta$ is called a **parameter**.*

Therefore, in a parametric model every element of Φ is put into correspondence with at least one parameter θ.

When $\gamma(\theta)$ associates to each parameter a unique joint distribution function (i.e. when $\gamma(\theta)$ is a function) the parametric model is called a parametric family.

Definition 348 (Parametric family) *Let (Φ, Θ, γ) be a parametric model. If γ is a function from Θ to Φ, then the parametric model is called a **parametric family**. In this case, the joint distribution function associated to a parameter θ is denoted by $F_\Xi(\xi; \theta)$.*

When each distribution function is associated with only one parameter, the parametric family is said to be identifiable.

Definition 349 (Identifiable parametric family) *Let (Φ, Θ, γ) be a parametric family. If γ is one-to-one (i.e. each distribution function F is associated with only one parameter), then the parametric family is said to be **identifiable**.*

70.3 Statistical inferences

A **statistical inference** is a statement about the unknown distribution function $F_\Xi(\xi)$, based on the observed sample ξ and the statistical model Φ. Statistical inferences are often chosen among a set of possible inferences and take the form of model restrictions. Given a subset of the original model $\Phi_R \subset \Phi$, a **model restriction** can be either an **inclusion restriction**:

$$F_\Xi \in \Phi_R$$

or an **exclusion restriction**:

$$F_\Xi \notin \Phi_R$$

The following are common kinds of statistical inferences:

1. In **hypothesis testing**, a restriction $F_\Xi \in \Phi_R$ is proposed and the choice is between two possible statements:
   (a) reject the restriction;
   (b) do not reject the restriction.
2. In **estimation**, a restriction $F_\Xi \in \Phi_R$ must be chosen among a set of possible restrictions.
3. In **Bayesian inference**, the observed sample ξ is used to update the subjective probability that a restriction $F_\Xi \in \Phi_R$ is true.

70.4 Decision theory

The choice of the statement (the statistical inference) to make based on the observed data can often be formalized as a decision problem where:

1. making a statistical inference is regarded as an **action**;
2. each action can have different **consequences**, depending on which distribution function $F_{\Xi}(\xi)$ is the true one;
3. a **preference ordering** over possible consequences needs to be elicited;
4. an **optimal course of action** needs to be taken, coherently with elicited preferences.

There are several different ways of formalizing such a decision problem. The branch of statistics that analyzes these decision problems is called **statistical decision theory**.

Chapter 71

Point estimation

In the lecture entitled *Statistical inference* (p. 563) we have defined statistical inference as the act of using a sample ξ to make statements about the probability distribution that generated the sample. The sample ξ is regarded as the realization of a random vector Ξ, whose unknown joint distribution function[1], denoted by $F_{\Xi}(\xi)$, is assumed to belong to a set of distribution functions Φ, called statistical model.

When the model Φ is put into correspondence with a set $\Theta \subseteq \mathbb{R}^p$ of real vectors, then we have a parametric model. Θ is called the parameter space and its elements are called parameters. Denote by θ_0 the parameter that is associated with the unknown distribution function $F_{\Xi}(\xi)$ and assume that θ_0 is unique. θ_0 is called the **true parameter**, because it is associated to the distribution that actually generated the sample. This lecture introduces a type of inference about the true parameter called point estimation.

71.1 Estimate and estimator

Roughly speaking, point estimation is the act of choosing a parameter $\widehat{\theta} \in \Theta$ that is our best guess of of the true (and unknown) parameter θ_0. Our best guess $\widehat{\theta}$ is called an **estimate** of θ_0.

When the estimate $\widehat{\theta}$ is produced using a predefined rule (a function) that associates a parameter estimate $\widehat{\theta}$ to each ξ in the support of Ξ, we can write:

$$\widehat{\theta} = \widehat{\theta}(\xi)$$

The function $\widehat{\theta}(\xi)$ is called an **estimator**. Often, the symbol $\widehat{\theta}$ is used to denote both the estimate and the estimator. The meaning is usually clear from the context.

71.2 Estimation error, loss and risk

Using the decision-theoretic terminology introduced in the lecture entitled *Statistical inference*[2], making an estimate $\widehat{\theta}$ is an act that produces some consequences.

[1] See p. 118.

[2] See, in particular, the section on Decision Theory (p. 567).

Among the consequences that are usually considered in a parametric decision problem, the most relevant one is the estimation error. The **estimation error** e is the difference between the estimate $\widehat{\theta}$ and the true parameter θ_0:

$$e = \widehat{\theta} - \theta_0$$

Of course, the statistician's goal is to commit the smallest possible estimation error. This preference can be formalized using loss functions. A **loss function** $L\left(\widehat{\theta}, \theta_0\right)$, mapping $\Theta \times \Theta$ into $\mathbb{R}$, quantifies the loss incurred by estimating θ_0 with $\widehat{\theta}$.

Frequently used loss functions are:

1. the **absolute error**:
$$L\left(\widehat{\theta}, \theta_0\right) = \left\|\widehat{\theta} - \theta_0\right\|$$
where $\|\|$ is the Euclidean norm (it coincides with the absolute value when $\Theta \subseteq \mathbb{R}$).

2. the **squared error**:
$$L\left(\widehat{\theta}, \theta_0\right) = \left\|\widehat{\theta} - \theta_0\right\|^2$$

When the estimate $\widehat{\theta}$ is obtained from an estimator (a function of the sample ξ, which in turn is a realization of the random vector Ξ), then the loss

$$L\left(\widehat{\theta}\left(\Xi\right), \theta_0\right)$$

can be thought of as a random variable. Its expected value is called the **statistical risk** (or, simply, the **risk**) of the estimator $\widehat{\theta}$ and it is denoted by $R\left(\widehat{\theta}\right)$:

$$R\left(\widehat{\theta}\right) = \mathrm{E}\left[L\left(\widehat{\theta}\left(\Xi\right), \theta_0\right)\right]$$

where the expected value is computed with respect to the true distribution function $F_\Xi\left(\xi\right)$. Thus, the risk $R\left(\widehat{\theta}\right)$ depends both on the true parameter θ_0 and on the distribution function of Ξ. In practice, these quantities are unknown, so also the risk needs to be estimated. For example, we can compute an estimate $\widehat{R}$ of the risk by pretending that the estimate $\widehat{\theta}$ were the true parameter, denoting by $\widetilde{\theta}$ the estimator of $\widehat{\theta}$ and computing the estimated risk as:

$$R\left(\widetilde{\theta}\right) = \mathrm{E}\left[L\left(\widetilde{\theta}\left(\Xi\right), \widehat{\theta}\right)\right]$$

where the expected value is with respect to the estimated distribution function $F_\Xi\left(\xi; \widehat{\theta}\right)$.

Even if the risk is unknown, the notion of risk is often used to derive theoretical properties of estimators. In any case, parameter estimation is always guided, at least ideally, by the principle of risk minimization, i.e. by the search for estimators $\widehat{\theta}$ that minimize the risk $R\left(\widehat{\theta}\right)$.

Depending on the specific loss function we use, the statistical risk of an estimator can take different names:

1. when the absolute error is used as a loss function, then the risk

$$R\left(\widehat{\theta}\right) = \mathrm{E}\left[\left\|\widehat{\theta} - \theta_0\right\|\right]$$

is called the **mean absolute error** of the estimator.

2. when the squared error is used as a loss function, then the risk

$$R\left(\widehat{\theta}\right) = \mathrm{E}\left[\left\|\widehat{\theta} - \theta_0\right\|^2\right]$$

is called **mean squared error** (**MSE**). The square root of the mean squared error is called **root mean squared error** (**RMSE**).

71.3 Other criteria to evaluate estimators

In this section we discuss other criteria that are commonly used to evaluate estimators.

71.3.1 Unbiasedness

If an estimator produces parameter estimates that are on average correct, then it is said to be unbiased. The following is a formal definition:

Definition 350 (unbiasedness) *Let θ_0 be the true parameter and let $\widehat{\theta}$ be an estimator of θ_0. $\widehat{\theta}$ is an **unbiased estimator** of θ_0 if and only if:*

$$\mathrm{E}\left[\widehat{\theta}\right] = \theta_0$$

*If an estimator is not unbiased, then it is called a **biased estimator**.*

Note that in the above definition of unbiasedness $\mathrm{E}\left[\widehat{\theta}\right]$ is a shorthand for:

$$\mathrm{E}\left[\widehat{\theta}\left(\Xi\right)\right]$$

where Ξ is the random vector of which the sample ξ is a realization and the expected value is computed with respect to the true distribution function $F_{\Xi}\left(\xi\right)$.

Also note that if an estimator is unbiased, this implies that the estimation error is on average zero:

$$\mathrm{E}\left[e\right] = \mathrm{E}\left[\widehat{\theta} - \theta_0\right] = \mathrm{E}\left[\widehat{\theta}\right] - \theta_0 = \theta_0 - \theta_0 = 0$$

71.3.2 Consistency

If an estimator produces parameter estimates that converge to the true value when the sample size increases, then it is said to be consistent. The following is a formal definition:

Definition 351 (consistency) *Let $\{\xi_n\}$ be a sequence of samples such that all the distribution functions $F_{\Xi_n}(\xi_n)$ are put into correspondence with the same parameter θ_0. A sequence of estimators $\left\{\widehat{\theta}_n(\xi_n)\right\}$ is said to be* ***consistent*** *(or weakly consistent) if and only if:*

$$\operatorname*{plim}_{n\to\infty}\widehat{\theta}_n = \theta_0$$

where plim *indicates convergence in probability*[3]*. The sequence of estimators is said to be* ***strongly consistent*** *if and only if:*

$$\widehat{\theta}_n \overset{a.s.}{\to} \theta_0$$

where $\overset{a.s.}{\to}$ indicates almost sure convergence[4]*. A sequence of estimators which is not consistent is called* ***inconsistent****.*

When the sequence of estimators is obtained using the same predefined rule for every sample ξ_n, we often say, with a slight abuse of language, "consistent estimator" instead of saying "consistent sequence of estimators". In such cases, what we mean is that the predefined rule produces a consistent sequence of estimators.

71.4 Examples

You can find examples of point estimation in the lectures entitled *Point estimation of the mean* (p. 573) and *Point estimation of the variance* (p. 579).

[3] See p. 511.
[4] See p. 505.

Chapter 72

Point estimation of the mean

This lecture presents some examples of point estimation problems, focusing on **mean estimation**, i.e. on using a sample to produce a point estimate of the mean of an unknown distribution. Before reading this lecture you need to be familiar with the material presented in the lecture entitled *Point estimation* (p. 569).

72.1 Normal IID samples

72.1.1 The sample

In this example, which is probably the most important in the history of statistics, the sample ξ_n is made of n independent draws from a normal distribution having unknown mean μ and variance σ^2. Specifically, we observe n realizations $x_1, \ldots, x_n$ of n independent random variables $X_1, \ldots, X_n$, all having a normal distribution with mean μ and variance σ^2. The sample[1] is the n-dimensional vector

$$\xi_n = \begin{bmatrix} x_1 & \ldots & x_n \end{bmatrix}$$

which is a realization of the random vector

$$\Xi_n = \begin{bmatrix} X_1 & \ldots & X_n \end{bmatrix}$$

72.1.2 The estimator

As an estimator[2] of the mean μ, we use the **sample mean** $\overline{X}_n$:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

72.1.3 Expected value of the estimator

The expected value of the estimator $\overline{X}_n$ is equal to the true mean μ:

$$\mathrm{E}\left[\overline{X}_n\right] = \mu$$

[1] See p. 564.
[2] See p. 569.

Proof. This can be proved using linearity of the expected value:

$$\begin{aligned} \mathrm{E}\left[\overline{X}_n\right] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[X_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu \end{aligned}$$

■

Therefore, the estimator $\overline{X}_n$ is unbiased[3].

72.1.4 Variance of the estimator

The variance of the estimator $\overline{X}_n$ is:

$$\mathrm{Var}\left[\overline{X}_n\right] = \frac{\sigma^2}{n}$$

Proof. This can be proved using the formula for the variance of an independent sum[4]:

$$\begin{aligned} \mathrm{Var}\left[\overline{X}_n\right] &= \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}\left[X_i\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 \\ &= \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n} \end{aligned}$$

■

Therefore, the variance of the estimator tends to zero as the sample size n tends to infinity.

72.1.5 Distribution of the estimator

The estimator $\overline{X}_n$ has a normal distribution:

$$\overline{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Proof. Note that the sample mean $\overline{X}_n$ is a linear combination of the normal and independent random variables $X_1, \ldots, X_n$ (all the coefficients of the linear

[3] See p. 571.
[4] See p. 168.

combination are equal to $\frac{1}{n}$). Therefore, $\overline{X}_n$ is normal because a linear combination of independent normal random variables is normal[5]. The mean and the variance of the distribution have already been derived above. ■

72.1.6 Risk of the estimator

The mean squared error[6] of the estimator is:

$$\operatorname{MSE}\left(\overline{X}_n\right) = \frac{\sigma^2}{n}$$

Proof. This is proved as follows:

$$\begin{aligned} \operatorname{MSE}\left(\overline{X}_n\right) &= \operatorname{E}\left[\left\|\overline{X}_n - \mu\right\|^2\right] \\ \boxed{\text{A}} &= \operatorname{E}\left[\left|\overline{X}_n - \mu\right|^2\right] \\ &= \operatorname{E}\left[\left(\overline{X}_n - \mu\right)^2\right] \\ \boxed{\text{B}} &= \operatorname{Var}\left[\overline{X}_n\right] \\ &= \frac{\sigma^2}{n} \end{aligned}$$

where: in step A we have used the fact that in one dimension the Euclidean norm is the same as the absolute value; in step B we have used the definition of variance and the fact that $\operatorname{E}\left[\overline{X}_n\right] = \mu$ (see above). ■

72.1.7 Consistency of the estimator

The sequence $\{X_n\}$ satisfies the conditions of Kolmogorov's Strong Law of Large Numbers[7] ($\{X_n\}$ is an IID sequence with finite mean). Therefore, the sample mean $\overline{X}_n$ converges almost surely to the true mean μ:

$$\overline{X}_n \xrightarrow{a.s.} \mu$$

i.e. the estimator $\overline{X}_n$ is strongly consistent[8]. Of course, the estimator is also weakly consistent, because almost sure convergence implies convergence in probability[9]:

$$\operatorname*{plim}_{n\to\infty} \overline{X}_n = \mu$$

72.2 IID samples

72.2.1 The sample

In this example, the sample ξ_n is made of n independent draws from a probability distribution having unknown mean μ and variance σ^2. Specifically, we observe

[5] See p. 471.
[6] See p. 571.
[7] See p. 540.
[8] See p. 571.
[9] See p. 533.

n realizations $x_1, \ldots, x_n$ of n independent random variables $X_1, \ldots, X_n$, all having the same distribution with mean μ and variance σ^2. The sample is the n-dimensional vector

$$\xi_n = [x_1 \quad \ldots \quad x_n]$$

which is a realization of the random vector

$$\Xi_n = [X_1 \quad \ldots \quad X_n]$$

The difference with respect to the previous example is that now we are no longer assuming that the sample points come from a normal distribution.

72.2.2 The estimator

Again, the estimator of the mean μ is the **sample mean** $\overline{X}_n$:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

72.2.3 Expected value of the estimator

The expected value of the estimator $\overline{X}_n$ is equal to the true mean μ and is therefore unbiased:

$$\mathrm{E}\left[\overline{X}_n\right] = \mu$$

The proof is the same found in the previous example.

72.2.4 Variance of the estimator

The variance of the estimator $\overline{X}_n$ is:

$$\mathrm{Var}\left[\overline{X}_n\right] = \frac{\sigma^2}{n}$$

Also in this case the proof is the same found in the previous example.

72.2.5 Distribution of the estimator

Unlike in the previous example, the estimator $\overline{X}_n$ does not necessarily have a normal distribution (its distribution depends on the distribution of the terms of the sequence $\{X_n\}$). However, we will see below that $\overline{X}_n$ has a normal distribution asymptotically (i.e. it converges to a normal distribution when n becomes large).

72.2.6 Risk of the estimator

The mean squared error of the estimator is:

$$\mathrm{MSE}\left(\overline{X}_n\right) = \mathrm{Var}\left[\overline{X}_n\right] = \frac{\sigma^2}{n}$$

The proof is the same found in the previous example.

72.2.7 Consistency of the estimator

The sequence $\{X_n\}$ satisfies the conditions of Kolmogorov's Strong Law of Large Numbers ($\{X_n\}$ is an IID sequence with finite mean). Therefore, the estimator $\overline{X}_n$ is both strongly consistent and weakly consistent (see example above).

72.2.8 Asymptotic normality

The sequence $\{X_n\}$ satisfies the conditions of Lindeberg-Lévy Central Limit Theorem[10] ($\{X_n\}$ is an IID sequence with finite mean and variance). Therefore, the sample mean $\overline{X}_n$ is asymptotically normal:

$$\sqrt{n}\left(\frac{\overline{X}_n - \mu}{\sigma}\right) \xrightarrow{d} Z$$

where Z is a standard normal random variable[11] and $\xrightarrow{d}$ denotes convergence in distribution. In other words, the sample mean $\overline{X}_n$ converges in distribution to a normal random variable with mean μ and variance $\frac{\sigma^2}{n}$.

72.3 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Consider an experiment that can have only two outcomes: either success, with probability p, or failure, with probability $1 - p$. The probability of success is unknown, but we know that

$$p \in \left[\frac{1}{10}, \frac{1}{5}\right]$$

Suppose we can independently repeat the experiment as many times as we wish and use the ratio

$$\frac{\text{Successes obtained}}{\text{Total experiments performed}}$$

as an estimator of p. What is the minimum number of experiments needed in order to be sure that the standard deviation of the estimator is less than 1/100?

Solution

Denote by $\widehat{p}$ the estimator of p. It can be written as

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

where n is the number of repetitions of the experiment and $X_1, \ldots, X_n$ are n independent random variables having a Bernoulli distribution[12] with parameter p.

[10] See p. 546.
[11] A normal random variable with zero mean and unit variance (see p. 376).
[12] See p. 335.

Therefore, $\widehat{p}$ is the sample mean of n independent Bernoulli random variables with expected value p and

$$\begin{aligned} & \operatorname{Var}[\widehat{p}] \\ \boxed{\text{A}} \quad &= \frac{\operatorname{Var}[X_i]}{n} \\ \boxed{\text{B}} \quad &= \frac{p(1-p)}{n} \\ &\leq \max_{p \in \left[\frac{1}{10}, \frac{1}{5}\right]} \frac{p(1-p)}{n} \\ &= \frac{\frac{1}{5}\left(1-\frac{1}{5}\right)}{n} \\ &= \frac{4}{25n} \end{aligned}$$

where: in step $\boxed{\text{A}}$ we have used the formula for the variance of the sample mean; in step $\boxed{\text{B}}$ we have used the formula for the variance of a Bernoulli random variable. Thus

$$\operatorname{Var}[\widehat{p}] \leq \frac{4}{25n}$$

We need to ensure that

$$\operatorname{Std}[\widehat{p}] = \sqrt{\operatorname{Var}[\widehat{p}]} \leq \frac{1}{100}$$

or

$$\operatorname{Var}[\widehat{p}] \leq \frac{1}{10000}$$

which is certainly verified if

$$\frac{4}{25n} = \frac{1}{10000}$$

or

$$n = \frac{40000}{25} = 1600$$

Exercise 2

Suppose you observe a sample of 100 independent draws from a distribution having unknown mean μ and known variance $\sigma^2 = 1$. How can you approximate the distribution of their sample mean?

Solution

We can approximate the distribution of the sample mean with its asymptotic distribution. So the distribution of the sample mean can be approximated by a normal distribution with mean μ and variance

$$\frac{\sigma^2}{n} = \frac{1}{100}$$

Chapter 73

Point estimation of the variance

This lecture presents some examples of point estimation problems, focusing on **variance estimation**, i.e. on using a sample to produce a point estimate of the variance of an unknown distribution. Before reading this lecture you need to be familiar with the material presented in the lecture entitled *Point estimation* (p. 569).

73.1 Normal IID samples - Known mean

In this example we make assumptions that are similar to those we made in the example of mean estimation entitled *Normal IID samples* (p. 573). The reader is strongly advised to read that example before reading this one.

73.1.1 The sample

The sample ξ_n is made of n independent draws from a normal distribution having known mean μ and unknown variance σ^2. Specifically, we observe n realizations $x_1, \ldots, x_n$ of n independent random variables $X_1, \ldots, X_n$, all having a normal distribution with known mean μ and unknown variance σ^2. The sample[1] is the n-dimensional vector

$$\xi_n = \begin{bmatrix} x_1 & \ldots & x_n \end{bmatrix}$$

which is a realization of the random vector

$$\Xi_n = \begin{bmatrix} X_1 & \ldots & X_n \end{bmatrix}$$

73.1.2 The estimator

We use the following estimator[2] of variance:

$$\widehat{\sigma_n^2} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$$

[1] See p. 564.
[2] See p. 569.

73.1.3 Expected value of the estimator

The expected value of the estimator $\widehat{\sigma_n^2}$ is equal to the true variance σ^2:

$$\mathrm{E}\left[\widehat{\sigma_n^2}\right] = \sigma^2 \tag{73.1}$$

Proof. This can be proved using the linearity of the expected value:

$$\begin{aligned}
\mathrm{E}\left[\widehat{\sigma_n^2}\right] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu\right)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\left(X_i-\mu\right)^2\right] \\
\boxed{\text{A}} &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}\left[X_i\right] \\
&= \frac{1}{n}n\sigma^2 = \sigma^2
\end{aligned}$$

where: in step $\boxed{\text{A}}$ we have used the definition of variance. ■

Therefore, the estimator $\widehat{\sigma_n^2}$ is unbiased[3].

73.1.4 Variance of the estimator

The variance of the estimator $\widehat{\sigma_n^2}$ is:

$$\mathrm{Var}\left[\widehat{\sigma_n^2}\right] = \frac{2\sigma^4}{n}$$

Proof. This can be proved as follows:

$$\begin{aligned}
\mathrm{Var}\left[\widehat{\sigma_n^2}\right] &= \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu\right)^2\right] \\
&= \frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^{n}\left(X_i-\mu\right)^2\right] \\
\boxed{\text{A}} &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left[\left(X_i-\mu\right)^2\right] \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathrm{E}\left[\left(X_i-\mu\right)^4\right]-\sum_{i=1}^{n}\mathrm{E}\left[\left(X_i-\mu\right)^2\right]^2\right) \\
\boxed{\text{B}} &= \frac{1}{n^2}\left(\sum_{i=1}^{n}3\sigma^4-\sum_{i=1}^{n}\left(\sigma^2\right)^2\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}2\sigma^4\right) \\
&= \frac{1}{n^2}n2\sigma^4 = \frac{2\sigma^4}{n}
\end{aligned}$$

[3] See p. 571.

where: in step [A] we have used the formula for the variance of an independent sum[4]; in step [B] we have used the fact that for a normal distribution

$$\mathrm{E}\left[(X_i-\mu)^4\right]=3\sigma^4$$

■

Therefore, the variance of the estimator tends to zero as the sample size n tends to infinity.

73.1.5 Distribution of the estimator

The estimator $\widehat{\sigma_n^2}$ has a Gamma distribution[5] with parameters n and σ^2.

Proof. The estimator $\widehat{\sigma_n^2}$ can be written as:

$$\begin{aligned}\widehat{\sigma_n^2} &= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 \\ &= \frac{\sigma^2}{n}\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \\ &= \frac{\sigma^2}{n}\sum_{i=1}^{n}Z_i^2=\frac{\sigma^2}{n}W\end{aligned}$$

where the variables Z_i are independent standard normal random variables[6] and W, being a sum of squares of n independent standard normal random variables, has a Chi-square distribution with n degrees of freedom[7]. Multiplying a Chi-square random variable with n degrees of freedom by σ^2/n one obtains[8] a Gamma random variable with parameters n and σ^2. ■

73.1.6 Risk of the estimator

The mean squared error[9] of the estimator is:

$$\begin{aligned}\mathrm{MSE}\left(\widehat{\sigma_n^2}\right) &= \mathrm{E}\left[\left\|\widehat{\sigma_n^2}-\sigma^2\right\|^2\right] \\ \boxed{\mathrm{A}} &= \mathrm{E}\left[\left|\widehat{\sigma_n^2}-\sigma^2\right|^2\right] \\ &= \mathrm{E}\left[\left(\widehat{\sigma_n^2}-\sigma^2\right)^2\right] \\ \boxed{\mathrm{B}} &= \mathrm{Var}\left[\widehat{\sigma_n^2}\right]=\frac{2\sigma^4}{n}\end{aligned}$$

where: in step [A] we have used the fact that the Euclidean norm in one dimension is the same as the absolute value; in step [B] we have used the definition of variance and (73.1).

[4] See p. 168.
[5] See p. 397.
[6] See p. 375.
[7] See p. 395.
[8] See p. 402.
[9] See p. 571.

73.1.7 Consistency of the estimator

The estimator

$$\widehat{\sigma_n^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

can be viewed as the sample mean of a sequence $\{Y_n\}$ where the generic term of the sequence is

$$Y_n = (X_n - \mu)^2$$

The sequence $\{Y_n\}$ satisfies the conditions of Kolmogorov's Strong Law of Large Numbers[10] ($\{X_n\}$ is an IID sequence with finite mean). Therefore, the sample mean of Y_n converges almost surely to the true mean $\mathrm{E}[Y_n]$:

$$\widehat{\sigma^2} \overset{a.s.}{\to} \mathrm{E}[Y_n] = \sigma^2$$

Therefore the estimator $\widehat{\sigma^2}$ is strongly consistent. It is also weakly consistent[11]:

$$\operatorname*{plim}_{n\to\infty} \widehat{\sigma^2} = \sigma^2$$

because almost sure convergence implies convergence in probability[12].

73.2 Normal IID samples - Unknown mean

This example is similar to the previous one. The only difference is that we relax the assumption that the mean of the distribution is known.

73.2.1 The sample

The sample ξ_n is made of n independent draws from a normal distribution having unknown mean μ and unknown variance σ^2. Specifically, we observe n realizations $x_1, \ldots, x_n$ of n independent random variables $X_1, \ldots, X_n$, all having a normal distribution with unknown mean μ and unknown variance σ^2. The sample is the n-dimensional vector

$$\xi_n = [x_1 \quad \ldots \quad x_n]$$

which is a realization of the random vector

$$\Xi_n = [X_1 \quad \ldots \quad X_n]$$

73.2.2 The estimator

In this example also the mean of the distribution, being unknown, needs to be estimated. It is estimated with the sample mean $\overline{X}_n$:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

We use the following estimators of variance:

[10] See p. 540.
[11] See p. 571.
[12] See p. 533.

1. the unadjusted **sample variance**:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$$

2. the adjusted **sample variance**:

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$$

73.2.3 Expected value of the estimator

The expected value of the unadjusted sample variance is:

$$\mathrm{E}\left[S_n^2\right] = \frac{n-1}{n}\sigma^2$$

Proof. This can be proved as follows:

$$\begin{aligned}
\mathrm{E}\left[S_n^2\right] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\left(X_i - \mu - \left(\overline{X}_n - \mu\right)\right)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\left(X_i - \mu\right)^2 + \left(\overline{X}_n - \mu\right)^2 - 2\left(X_i - \mu\right)\left(\overline{X}_n - \mu\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\left(X_i - \mu\right)^2\right] + \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\left(\overline{X}_n - \mu\right)^2\right] \\
&\quad -\frac{2}{n}\sum_{i=1}^{n}\mathrm{E}\left[\left(X_i - \mu\right)\left(\overline{X}_n - \mu\right)\right] \\
\boxed{\text{A}} &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}\left[X_i\right] + \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}\left[\overline{X}_n\right] \\
&\quad -\frac{2}{n}\sum_{i=1}^{n}\mathrm{E}\left[\left(X_i - \mu\right)\left(\overline{X}_n - \mu\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\sigma^2 + \frac{1}{n}\sum_{i=1}^{n}\frac{\sigma^2}{n} \\
&\quad -\frac{2}{n}\sum_{i=1}^{n}\mathrm{E}\left[\left(X_i - \mu\right)\left(\frac{1}{n}\sum_{j=1}^{n}X_j - \mu\right)\right] \\
&= \frac{1}{n}n\sigma^2 + \frac{1}{n}n\frac{\sigma^2}{n} \\
&\quad -\frac{2}{n}\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\left(X_i - \mu\right)\left(\sum_{j=1}^{n}X_j - n\mu\right)\right] \\
&= \sigma^2 + \frac{\sigma^2}{n} - \frac{2}{n^2}\sum_{i=1}^{n}\mathrm{E}\left[\left(X_i - \mu\right)\left(\sum_{j=1}^{n}\left(X_j - \mu\right)\right)\right]
\end{aligned}$$

$$= \frac{n+1}{n}\sigma^2 - \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{E}\left[\left(X_i-\mu\right)\left(X_j-\mu\right)\right]$$

where: in step $\boxed{\text{A}}$ we have used the fact that[13]

$$\mathrm{Var}\left[\overline{X}_n\right] = \frac{\sigma^2}{n}$$

But

$$\mathrm{E}\left[\left(X_i-\mu\right)\left(X_j-\mu\right)\right] = 0$$

when $i \neq j$ (because X_i and X_j are independent when $i \neq j$). Therefore:

$$\begin{aligned}
\mathrm{E}\left[S_n^2\right] &= \frac{n+1}{n}\sigma^2 - \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{E}\left[\left(X_i-\mu\right)\left(X_j-\mu\right)\right] \\
&= \frac{n+1}{n}\sigma^2 - \frac{2}{n^2}\sum_{i=1}^{n}\mathrm{E}\left[\left(X_i-\mu\right)\left(X_i-\mu\right)\right] \\
&= \frac{n+1}{n}\sigma^2 - \frac{2}{n^2}\sum_{i=1}^{n}\mathrm{E}\left[\left(X_i-\mu\right)^2\right] \\
&= \frac{n+1}{n}\sigma^2 - \frac{2}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left[X_i\right] \\
&= \frac{n+1}{n}\sigma^2 - \frac{2}{n^2}n\sigma^2 \\
&= \frac{n+1}{n}\sigma^2 - \frac{2}{n}\sigma^2 = \frac{n-1}{n}\sigma^2
\end{aligned}$$

■

Therefore, the unadjusted sample variance S_n^2 is a biased[14] estimator of the true variance σ^2.

The adjusted sample variance s_n^2, on the contrary, is an unbiased estimator of variance:

$$\mathrm{E}\left[s_n^2\right] = \sigma^2$$

Proof. This can be proved as follows:

$$\begin{aligned}
\mathrm{E}\left[s_n^2\right] &= \mathrm{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}_n\right)^2\right] \\
&= \mathrm{E}\left[\frac{n}{n-1}\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}_n\right)^2\right] \\
&= \frac{n}{n-1}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}_n\right)^2\right] \\
&= \frac{n}{n-1}\mathrm{E}\left[S_n^2\right] \\
&= \frac{n}{n-1}\frac{n-1}{n}\sigma^2 = \sigma^2
\end{aligned}$$

[13] See p. 574.
[14] See p. 571.

■

Thus, when also the mean μ is being estimated, we need to divide by $n-1$ rather than by n to obtain an unbiased estimator. Intuitively, by considering squared deviations from the sample mean rather than squared deviations from the true mean, we are underestimating the true variability of the data. In fact, the sum of squared deviations from the true mean is always larger than the sum of squared deviations from the sample mean. Dividing by $n-1$ rather than by n exactly corrects this bias. The number $n-1$ by which we divide is called the **number of degrees of freedom** and it is equal to the number n of sample points minus the number of other parameters to be estimated (in our case 1, the true mean μ).

The factor by which we need to multiply the biased estimatot S_n^2 to obtain the unbiased estimator s_n^2 is, of course:

$$\frac{n}{n-1}$$

This factor is known as **degrees of freedom adjustment**, which explains why S_n^2 is called unadjusted sample variance and s_n^2 is called adjusted sample variance.

73.2.4 Variance of the estimator

The variance of the unadjusted sample variance is:

$$\operatorname{Var}\left[S_n^2\right] = \frac{n-1}{n}\frac{2\sigma^4}{n}$$

Proof. This is proved in subsection 73.2.5. ■

The variance of the adjusted sample variance is:

$$\operatorname{Var}\left[s_n^2\right] = \frac{2\sigma^4}{n-1}$$

Proof. This is also proved in subsection 73.2.5. ■

Therefore, both the variance of S_n^2 and the variance of s_n^2 converge to zero as the sample size n tends to infinity. Also note that the unadjusted sample variance S_n^2, despite being biased, has a smaller variance than the adjusted sample variance s_n^2, which is instead unbiased.

73.2.5 Distribution of the estimator

The unadjusted sample variance S_n^2 has a Gamma distribution with parameters $n-1$ and $\frac{(n-1)\sigma^2}{n}$.

Proof. To prove this result, we need to use some facts on quadratic forms involving normal random variables, which have been introduced in the lecture entitled *Quadratic forms in normal vectors* (p. 481). To understand this proof, you need to first read that lecture, in particular the section entitled *Sample variance as a quadratic form* (p. 486). Define the matrix:

$$M = I - \frac{1}{n}\imath\imath^\top$$

where I is an $n \times n$ identity matrix and $\imath$ is a $n \times 1$ vector of ones. M is symmetric and idempotent. Denote by X the $n \times 1$ random vector whose i-th entry is equal

to X_i. The random vector X has a multivariate normal distribution with mean $\mu\imath$ and covariance matrix $\sigma^2 I$. Using the fact that the matrix M is symmetric and idempotent, the unadjusted sample variance can be written as:

$$S_n^2 = \frac{1}{n} X^\top M X$$

Using the fact that the random vector

$$Z = \frac{1}{\sigma}(X - \mu\imath)$$

has a standard multivariate normal distribution[15] and the fact that $M\imath = 0$, we can rewrite:

$$S_n^2 = \frac{\sigma^2}{n} Z^\top M Z$$

In other words, S_n^2 is proportional to a quadratic form in a standard normal random vector ($Z^\top M Z$) and the quadratic form involves a symmetric and idempotent matrix whose trace is equal to $n-1$. Therefore, the quadratic form $Z^\top M Z$ has a Chi-square distribution with $n-1$ degrees of freedom. Finally, we can write:

$$S_n^2 = \frac{(n-1)\sigma^2}{n}\left(\frac{Z^\top M Z}{n-1}\right)$$

i.e. S_n^2 is a Chi-square random variable divided by its number of degrees of freedom and multiplied by $\frac{(n-1)\sigma^2}{n}$. Thus[16], S_n^2 is a Gamma random variable with parameters $n-1$ and $\frac{(n-1)\sigma^2}{n}$. Also, by the properties of Gamma random variables, its expected value is:

$$\mathrm{E}\left[S_n^2\right] = \frac{(n-1)\sigma^2}{n}$$

and its variance is:

$$\begin{aligned}\mathrm{Var}\left[S_n^2\right] &= \frac{2}{n-1}\left(\frac{(n-1)\sigma^2}{n}\right)^2 \\ &= \frac{n-1}{n}\frac{2\sigma^4}{n}\end{aligned}$$

■

The adjusted sample variance s_n^2 has a Gamma distribution with parameters $n-1$ and σ^2.

Proof. The proof of this result is similar to the proof for unadjusted sample variance found above. It can also be found in the lecture entitled *Quadratic forms in normal vectors* (p. 481). Here, we just notice that s_n^2, being a Gamma random variable with parameters $n-1$ and σ^2, has expected value

$$\mathrm{E}\left[s_n^2\right] = \sigma^2$$

and variance

$$\mathrm{Var}\left[s_n^2\right] = \frac{2\sigma^4}{n-1}$$

■

[15] See p. 439.
[16] See p. 397.

73.2.6 Risk of the estimator

The mean squared error of the unadjusted sample variance is:

$$\operatorname{MSE}\left(S_n^2\right) = \frac{2n-1}{n^2}\sigma^4$$

Proof. It can be proved as follows:

$$\begin{aligned}
\operatorname{MSE}\left(S_n^2\right) &= \operatorname{E}\left[\left\|S_n^2 - \sigma^2\right\|^2\right] \\
\boxed{\text{A}} &= \operatorname{E}\left[\left|S_n^2 - \sigma^2\right|^2\right] \\
&= \operatorname{E}\left[\left(S_n^2 - \sigma^2\right)^2\right] \\
&= \operatorname{E}\left[\left(S_n^2 - \operatorname{E}\left[S_n^2\right] + \operatorname{E}\left[S_n^2\right] - \sigma^2\right)^2\right] \\
&= \operatorname{E}\left[\left(S_n^2 - \operatorname{E}\left[S_n^2\right]\right)^2\right] + \operatorname{E}\left[\left(\operatorname{E}\left[S_n^2\right] - \sigma^2\right)^2\right] \\
&\quad +2\operatorname{E}\left[\left(S_n^2 - \operatorname{E}\left[S_n^2\right]\right)\left(\operatorname{E}\left[S_n^2\right] - \sigma^2\right)\right] \\
&= \operatorname{Var}\left[S_n^2\right] + \left(\operatorname{E}\left[S_n^2\right] - \sigma^2\right)^2 + 2\left(\operatorname{E}\left[S_n^2\right] - \sigma^2\right)\operatorname{E}\left[S_n^2 - \operatorname{E}\left[S_n^2\right]\right] \\
&= \frac{n-1}{n}\frac{2\sigma^4}{n} + \left(\frac{(n-1)\sigma^2}{n} - \sigma^2\right)^2 + 2\left(\operatorname{E}\left[S_n^2\right] - \sigma^2\right)\cdot 0 \\
&= \frac{2n-2}{n^2}\sigma^4 + \left(\frac{(n-1-n)\sigma^2}{n}\right)^2 \\
&= \frac{2n-1}{n^2}\sigma^4
\end{aligned}$$

where: in step $\boxed{\text{A}}$ we have used the fact that the Euclidean norm in one dimension is the same as the absolute value ■

The mean squared error of the adjusted sample variance is:

$$\operatorname{MSE}\left(s_n^2\right) = \frac{2}{n-1}\sigma^4$$

Proof. It can be proved as follows:

$$\begin{aligned}
\operatorname{MSE}\left(s_n^2\right) &= \operatorname{E}\left[\left\|s_n^2 - \sigma^2\right\|^2\right] \\
&= \operatorname{E}\left[\left|s_n^2 - \sigma^2\right|^2\right] \\
&= \operatorname{E}\left[\left(s_n^2 - \sigma^2\right)^2\right] \\
&= \operatorname{E}\left[\left(s_n^2 - \operatorname{E}\left[s_n^2\right]\right)^2\right] \\
&= \operatorname{Var}\left[s_n^2\right] = \frac{2\sigma^4}{n-1}
\end{aligned}$$

■

Therefore, the mean squared error of the unadjusted sample variance is always smaller than the mean squared error of the adjusted sample variance:

$$\operatorname{MSE}\left(S_n^2\right) = \frac{2n-1}{n^2}\sigma^4$$

$$\begin{aligned} &= \left(\frac{2n}{n^2} - \frac{1}{n^2}\right)\sigma^4 \\ &= \left(\frac{2}{n} - \frac{1}{n^2}\right)\sigma^4 \\ &< \frac{2}{n}\sigma^4 < \frac{2}{n-1}\sigma^4 = \mathrm{MSE}\left(s_n^2\right) \end{aligned}$$

73.2.7 Consistency of the estimator

The unadjusted sample variance can be written as:

$$\begin{aligned} S_n^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}X_i^2 + \frac{1}{n}\sum_{i=1}^{n}\overline{X}_n^2 - \frac{2}{n}\sum_{i=1}^{n}X_i\overline{X}_n \\ &= \frac{1}{n}\sum_{i=1}^{n}X_i^2 + \frac{1}{n}n\overline{X}_n^2 - 2\overline{X}_n\frac{1}{n}\sum_{i=1}^{n}X_i \\ \boxed{\text{A}} &= \frac{1}{n}\sum_{i=1}^{n}X_i^2 + \overline{X}_n^2 - 2\overline{X}_n\overline{X}_n \\ &= \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \overline{X}_n^2 \\ \boxed{\text{B}} &= W_n - \overline{X}_n^2 \end{aligned}$$

where: in step A we have used the fact that

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n}X_i$$

in step B we have defined:

$$W_n = \frac{1}{n}\sum_{i=1}^{n}X_i^2$$

The two sequences $\overline{X}_n$ and W_n are the sample means of X_n and X_n^2 respectively. The latter both satisfy the conditions of Kolmogorov's Strong Law of Large Numbers (they form IID sequences with finite means), which implies that their sample means $\overline{X}_n$ and W_n converge almost surely to their true means:

$$\begin{aligned} \overline{X}_n &\overset{a.s.}{\to} \mathrm{E}\left[X_n\right] \\ W_n &\overset{a.s.}{\to} \mathrm{E}\left[X_n^2\right] \end{aligned}$$

Since the function

$$g\left(\overline{X}_n, W_n\right) = W_n - \overline{X}_n^2$$

is continuous and almost sure convergence is preserved by continuous transformations[17], we obtain:

$$S_n^2 \overset{a.s.}{\to} g\left(\mathrm{E}\left[X_n\right], \mathrm{E}\left[X_n^2\right]\right) = \mathrm{E}\left[X_n^2\right] - \mathrm{E}\left[X_n\right]^2 = \mathrm{Var}\left[X_n\right] = \sigma^2$$

[17] See p. 555.

Therefore the estimator S_n^2 is strongly consistent. It is also weakly consistent, because almost sure convergence implies convergence in probability:

$$\operatorname*{plim}_{n\to\infty} S_n^2 = \sigma^2$$

The adjusted sample variance s_n^2 can be written as:

$$s_n^2 = \frac{n}{n-1} S_n^2$$

The ratio $\frac{n}{n-1}$ can be thought of as a constant random variable Z_n defined as follows:

$$Z_n = \frac{n}{n-1}$$

which converges almost surely to 1. Therefore

$$s_n^2 = Z_n S_n^2$$

where both Z_n and S_n^2 are almost surely convergent. Since the product is a continuous function and almost sure convergence is preserved by continuous transformation, we have:

$$s_n^2 \overset{a.s.}{\to} 1 \cdot \sigma^2 = \sigma^2$$

Thus, also s_n^2 is strongly consistent.

73.3 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

You observe three independent draws from a normal distribution having unknown mean μ and unknown variance σ^2. Their values are 50, 100 and 150. Use these values to produce an unbiased estimate of the variance of the distribution.

Solution

The sample mean is

$$\overline{X}_n = \frac{50 + 100 + 150}{3} = 100$$

An unbiased estimate of the variance is provided by the adjusted sample variance:

$$\begin{aligned} s_n^2 &= \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2 \\ &= \frac{1}{3-1} \left[(50-100)^2 + (100-100)^2 + (150-100)^2\right] \\ &= \frac{1}{2} \left[2500 + 0 + 2500\right] = \frac{5000}{2} = 2500 \end{aligned}$$

Exercise 2

A machine (a laser rangefinder) is used to measure the distance between the machine itself and a given object. When measuring the distance to an object located 10 meters apart, measurement errors committed by the machine are normally and independently distributed and are on average equal to zero. The variance of the measurement errors is less than 1 squared centimeter, but its exact value is unknown and needs to be estimated. To estimate it, we repeatedly take the same measurement and we compute the sample variance of the measurement errors (which we are also able to compute, because we know the true distance). How many measurements do we need to take to obtain an estimator of variance having a standard deviation less than 0.1 squared centimeters?

Solution

Denote the measurement errors by $X_1, \ldots, X_n$. The following estimator of variance is used:

$$\widehat{\sigma_n^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

The variance of this estimator is:

$$\begin{aligned} & \operatorname{Var}\left[\widehat{\sigma_n^2}\right] \\ \boxed{\text{A}} \quad = \quad & \frac{2\left(\operatorname{Var}\left[X_i\right]\right)^2}{n} \\ \boxed{\text{B}} \quad \leq \quad & \frac{2\left(1cm^2\right)^2}{n} \\ = \quad & \frac{2}{n}cm^4 \end{aligned}$$

where: in step $\boxed{\text{A}}$ we have used the formula for the variance of the sample variance; in step $\boxed{\text{B}}$ we have used the upper bound stated in the problem (variance less than 1 squared centimeter). Thus

$$\operatorname{Var}\left[\widehat{\sigma_n^2}\right] \leq \frac{2}{n}cm^4$$

We need to ensure that

$$\operatorname{Std}\left[\widehat{\sigma_n^2}\right] = \sqrt{\operatorname{Var}\left[\widehat{\sigma_n^2}\right]} \leq \frac{1}{10}cm^2$$

or

$$\operatorname{Var}\left[\widehat{\sigma_n^2}\right] \leq \frac{1}{100}cm^4$$

which is certainly verified if

$$\frac{2}{n} = \frac{1}{100}$$

or

$$n = 200$$

Chapter 74

Set estimation

In statistical inference, a sample[1] ξ is employed to make statements about the probability distribution from which the sample has been generated (see the lecture entitled *Statistical inference* - p. 563). The sample ξ can be regarded as a realization of a random vector Ξ, whose joint distribution function, denoted by $F_{\Xi}(\xi)$, is unknown, but is assumed to belong to a set of distribution functions Φ, called statistical model.

In a parametric model, the set Φ is put into correspondence with a set $\Theta \subseteq \mathbb{R}^p$ of p-dimensional real vectors. Θ is called the parameter space and its elements are called parameters. Denote by θ_0 the parameter that is associated with the unknown distribution function $F_{\Xi}(\xi)$ and assume that θ_0 is unique. θ_0 is called the **true parameter**, because it is associated to the distribution that actually generated the sample. This lecture discusses a kind of inference about the true parameter θ_0 called set estimation.

74.1 Confidence set

Roughly speaking, **set estimation** is the act of choosing a subset T of the parameter space ($T \subseteq \Theta$) in such a way that T has a high probability of containing the true (and unknown) parameter θ_0. The chosen subset T is called a **set estimate** of θ_0 or a **confidence set** for θ_0.

When the parameter space Θ is a subset of the set of real numbers $\mathbb{R}$ and the subset T is chosen among the intervals of $\mathbb{R}$ (e.g. intervals of the kind $[a, b]$), we speak about **interval estimation** (instead of set estimation), **interval estimate** (instead of set estimate) and **confidence interval** (instead of confidence set).

When the set estimate T is produced using a predefined rule (a function) that associates a set estimate T to each ξ in the support of Ξ, we can write:

$$T = T(\xi)$$

The function $T(\xi)$ is called a **set estimator**. Often, the symbol T is used to denote both the set estimate and the set estimator. The meaning is usually clear from the context.

[1] See p. 564.

74.2 Coverage probability - confidence coefficient

As we already said, set estimation is the act of choosing a subset T of the parameter space in such a way that T has a high probability of containing the true parameter θ_0. The probability that T contains the true parameter is called **coverage probability** and it is usually chosen by the statistician. Intuitively, before observing the data the statistician makes a statement:

$$\theta_0 \in \Theta$$

where Θ is the parameter space, containing all the parameters that are deemed plausible. The statistician believes the statement to be true, but the statement is not very informative, because Θ is a very large set. After observing the data, she makes a more informative statement:

$$\theta_0 \in T$$

This statement is more informative, because T is smaller than Θ, but it has a positive probability of being wrong (which is the complement to 1 of the coverage probability). In controlling this probability, the statistician faces a trade-off: if she decreases the probability of being wrong, then her statements become less informative; on the contrary, if she increases the probability of being wrong, then her statements become more informative.

In formal terms, the coverage probability of a set estimator is defined as follows:

$$\mathrm{C}\left(T, \theta_0\right) = \mathrm{P}_{\theta_0}\left(\theta_0 \in T\left(\Xi\right)\right)$$

where the notation P_{θ_0} is used to indicate the fact that the probability is calculated using the distribution function $F_\Xi\left(\xi; \theta_0\right)$ associated to the true parameter θ_0. It is important to note that in the above definition of coverage probability the random quantity is the interval $T\left(\Xi\right)$, while the parameter is fixed.

In practice, the coverage probability is seldom known, because it depends on the unknown parameter θ_0 (although in some cases it is equal for all parameters belonging to the parameter space). When the coverage probability is not known, it is customary to compute the **confidence coefficient** $\mathrm{c}\left(T\right)$, which is defined as follows:

$$\mathrm{c}\left(T\right) = \inf_{\theta \in \Theta} \mathrm{C}\left(T, \theta\right)$$

In other words, the confidence coefficient $\mathrm{c}\left(T\right)$ is equal to the smallest possible coverage probability. The confidence coefficient is also often called **level of confidence**.

74.3 Size of a confidence set

We already mentioned that there is a trade-off in the construction and choice of a set estimator. On the one hand, we want our set estimator T to have a high coverage probability, i.e. we want the set T to include the true parameter with a high probability. On the other hand, we want the size of T to be as small as possible, so as to make our interval estimate more precise. What do we mean by size of T? If the parameter space Θ is unidimensional and T is an interval estimate, then the size of T is just its length. If the space Θ is multidimensional,

then the size of T is its volume. The **size of a confidence set** is also called **measure of a confidence set** (for those who have a grasp of measure theory, the name stems from the fact that Lebesgue measure is the generalization of volume in multidimensional spaces). If we denote by $\lambda(T)$ the size of a confidence set, then we can also define the **expected size of a set estimator** T:

$$\mathrm{E}_{\theta_0}\left[\lambda\left(T\left(\Xi\right)\right)\right]$$

where the notation E_{θ_0} is used to indicate the fact that the expected value is calculated using the distribution function $F_{\Xi}\left(\xi;\theta_0\right)$ associated to the true parameter θ_0. Like the coverage probability, also the expected size of a set estimator depends on the unknown parameter θ_0. Hence, unless it is a constant function of θ_0, one needs to somehow estimate it or to take the infimum over all possible values of the parameter, as we did above for coverage probabilities.

74.4 Other criteria to evaluate set estimators

Although size is probably the simplest criterion to evaluate and select set estimators, there are several other criteria. We do not discuss them here, but we refer the reader to the very nice exposition in Berger and Casella[2] (2002).

74.5 Examples

You can find examples of set estimation in the lectures entitled *Set estimation of the mean* (p. 595) and *Set estimation of the variance* (p. 607).

[2]Berger, R. L. and G. Casella (2002) "Statistical inference", Duxbury Advanced Series.

Chapter 75

Set estimation of the mean

This lecture presents some examples of set estimation[1] problems, focusing on **set estimation of the mean**, i.e. on using a sample to produce a set estimate of the mean of an unknown distribution.

75.1 Normal IID samples - Known variance

In this example we make assumptions that are similar to those we made in the example of point estimation of the mean entitled *Normal IID samples* (p. 573). It would be beneficial to read that example before reading this one.

75.1.1 The sample

In this example, the sample ξ_n is made of n independent draws from a normal distribution having unknown mean μ and known variance σ^2. Specifically, we observe n realizations $x_1, \ldots, x_n$ of n independent random variables $X_1, \ldots, X_n$, all having a normal distribution with unknown mean μ and known variance σ^2. The sample is the n-dimensional vector

$$\xi_n = \begin{bmatrix} x_1 & \ldots & x_n \end{bmatrix}$$

which is a realization of the random vector

$$\Xi_n = \begin{bmatrix} X_1 & \ldots & X_n \end{bmatrix}$$

75.1.2 The interval estimator

To construct an interval estimator[2] of the mean μ, we use the sample mean

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

The interval estimator is

$$T_n = \left[\overline{X}_n - \sqrt{\frac{\sigma^2}{n}}z, \overline{X}_n + \sqrt{\frac{\sigma^2}{n}}z\right]$$

[1] See p. 591.
[2] See p. 591.

where $z \in \mathbb{R}_{++}$ is a strictly positive constant.

75.1.3 Coverage probability

The coverage probability[3] of the interval estimator T_n is

$$\mathrm{C}(T_n; \mu) = \mathrm{P}(\mu \in T_n) = \mathrm{P}(-z \leq Z \leq z)$$

where Z is a standard normal random variable[4].

Proof. The coverage probability can be written as

$$\begin{aligned}
\mathrm{P}(\mu \in T_n) &= \mathrm{P}\left(\overline{X}_n - \sqrt{\frac{\sigma^2}{n}} z \leq \mu \leq \overline{X}_n + \sqrt{\frac{\sigma^2}{n}} z\right) \\
&= \mathrm{P}\left(\left\{\overline{X}_n - \sqrt{\frac{\sigma^2}{n}} z \leq \mu\right\} \cap \left\{\mu \leq \overline{X}_n + \sqrt{\frac{\sigma^2}{n}} z\right\}\right) \\
&= \mathrm{P}\left(\left\{\overline{X}_n - \mu \leq \sqrt{\frac{\sigma^2}{n}} z\right\} \cap \left\{\overline{X}_n - \mu \geq -\sqrt{\frac{\sigma^2}{n}} z\right\}\right) \\
&= \mathrm{P}\left(\left\{\frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} \leq z\right\} \cap \left\{\frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} \geq -z\right\}\right) \\
&= \mathrm{P}\left(-z \leq \frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} \leq z\right) \\
&= \mathrm{P}(-z \leq Z \leq z)
\end{aligned}$$

where we have defined

$$Z = \frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}}$$

In the lecture entitled *Point estimation of the mean* (p. 573), we have demonstrated that, given the assumptions on the sample ξ_n made above, the sample mean $\overline{X}_n$ has a normal distribution with mean μ and variance σ^2/n. Subtracting the mean of a normal random variable from the random variable itself and dividing it by the square root of its variance, one obtains a standard normal random variable. Therefore, the variable Z has a standard normal distribution. ■

75.1.4 Confidence coefficient

Note that the coverage probability does not depend on the unknown parameter μ. Therefore, the confidence coefficient[5] of the interval estimator T_n coincides with its coverage probability:

$$\mathrm{c}(T_n) = \inf_{\mu \in \mathbb{R}} \mathrm{C}(T_n, \mu) = \mathrm{P}(-z \leq Z \leq z)$$

where Z is a standard normal random variable.

[3] See p. 592.
[4] See p. 375.
[5] See p. 592.

75.1.5 Size

The size[6] of the interval estimator T_n is

$$\begin{aligned}\lambda(T_n) &= \lambda\left(\left[\overline{X}_n - \sqrt{\frac{\sigma^2}{n}}z, \overline{X}_n + \sqrt{\frac{\sigma^2}{n}}z\right]\right) \\ &= \overline{X}_n + \sqrt{\frac{\sigma^2}{n}}z - \overline{X}_n + \sqrt{\frac{\sigma^2}{n}}z = 2\sqrt{\frac{\sigma^2}{n}}z\end{aligned}$$

75.1.6 Expected size

Note that the size does not depend on the sample ξ_n. Therefore, the expected size is

$$\mathrm{E}\left[\lambda(T_n)\right] = \mathrm{E}\left[2\sqrt{\frac{\sigma^2}{n}}z\right] = 2\sqrt{\frac{\sigma^2}{n}}z$$

75.2 Normal IID samples - Unknown variance

This example is similar to the previous one. The only difference is that we now relax the assumption that the variance of the distribution is known.

75.2.1 The sample

In this example, the sample ξ_n is made of n independent draws from a normal distribution having unknown mean μ and unknown variance σ^2. Specifically, we observe n realizations $x_1, \ldots, x_n$ of n independent random variables $X_1, \ldots, X_n$, all having a normal distribution with unknown mean μ and unknown variance σ^2. The sample is the n-dimensional vector

$$\xi_n = \begin{bmatrix} x_1 & \ldots & x_n \end{bmatrix}$$

which is a realization of the random vector

$$\Xi_n = \begin{bmatrix} X_1 & \ldots & X_n \end{bmatrix}$$

75.2.2 The interval estimator

To construct interval estimators of the mean μ, we use the sample mean

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and either the unadjusted sample variance[7]

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$$

[6] See p. 592.
[7] See p. 583.

or the adjusted sample variance[8]

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$$

We consider two interval estimators of the mean:

$$\begin{aligned} T_n^u &= \left[\overline{X}_n - \sqrt{\frac{S_n^2}{n}}z, \overline{X}_n + \sqrt{\frac{S_n^2}{n}}z\right] \\ T_n^a &= \left[\overline{X}_n - \sqrt{\frac{s_n^2}{n}}z, \overline{X}_n + \sqrt{\frac{s_n^2}{n}}z\right] \end{aligned}$$

where $z \in \mathbb{R}_{++}$ is a strictly positive constant and the superscripts u and a indicate whether the estimator is based on the unadjusted or the adjusted sample variance.

75.2.3 Coverage probability

The coverage probability of the interval estimator T_n^u is

$$\mathrm{C}\left(T_n^u; \mu, \sigma^2\right) = \mathrm{P}\left(\mu \in T_n^u\right) = \mathrm{P}\left(-\sqrt{\frac{n-1}{n}}z \leq Z_{n-1} \leq \sqrt{\frac{n-1}{n}}z\right)$$

where Z_{n-1} is a standard Student's t random variable[9] with $n-1$ degrees of freedom.

Proof. The coverage probability can be written as

$$\begin{aligned} \mathrm{P}\left(\mu \in T_n^u\right) &= \mathrm{P}\left(\overline{X}_n - \sqrt{\frac{S_n^2}{n}}z \leq \mu \leq \overline{X}_n + \sqrt{\frac{S_n^2}{n}}z\right) \\ &= \mathrm{P}\left(\left\{\overline{X}_n - \sqrt{\frac{S_n^2}{n}}z \leq \mu\right\} \cap \left\{\mu \leq \overline{X}_n + \sqrt{\frac{S_n^2}{n}}z\right\}\right) \\ &= \mathrm{P}\left(\left\{\overline{X}_n - \mu \leq \sqrt{\frac{S_n^2}{n}}z\right\} \cap \left\{\overline{X}_n - \mu \geq -\sqrt{\frac{S_n^2}{n}}z\right\}\right) \\ &= \mathrm{P}\left(\left\{\frac{\overline{X}_n - \mu}{\sqrt{S_n^2/n}} \leq z\right\} \cap \left\{\frac{\overline{X}_n - \mu}{\sqrt{S_n^2/n}} \geq -z\right\}\right) \\ &= \mathrm{P}\left(-z \leq \frac{\overline{X}_n - \mu}{\sqrt{S_n^2/n}} \leq z\right) \\ &= \mathrm{P}\left(-\sqrt{\frac{n-1}{n}}z \leq \sqrt{\frac{n-1}{n}}\frac{\overline{X}_n - \mu}{\sqrt{S_n^2/n}} \leq \sqrt{\frac{n-1}{n}}z\right) \\ &= \mathrm{P}\left(-\sqrt{\frac{n-1}{n}}z \leq Z_{n-1} \leq \sqrt{\frac{n-1}{n}}z\right) \end{aligned}$$

where we have defined

$$Z_{n-1} = \sqrt{\frac{n-1}{n}}\frac{\overline{X}_n - \mu}{\sqrt{S_n^2/n}}$$

[8] See p. 583.
[9] See p. 407.

Now, rewrite Z_{n-1} as

$$\begin{aligned} Z_{n-1} &= \sqrt{\frac{n-1}{n}}\frac{\overline{X}_n - \mu}{\sqrt{S_n^2/n}} \\ &= \sqrt{\frac{n-1}{n}}\frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} \cdot \frac{\sqrt{\sigma^2/n}}{\sqrt{S_n^2/n}} \\ &= \sqrt{\frac{n-1}{n}}\frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} \cdot \frac{1}{\sqrt{S_n^2/\sigma^2}} \\ &= \sqrt{\frac{n-1}{n}}\frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} \cdot \frac{1}{\sqrt{\frac{n-1}{n}s_n^2/\sigma^2}} \\ &= \frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} \cdot \frac{1}{\sqrt{s_n^2/\sigma^2}} = \frac{Y}{\sqrt{W}} \end{aligned}$$

where we have defined

$$\begin{aligned} Y &= \frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} \\ W &= s_n^2/\sigma^2 \end{aligned}$$

and we have used the fact that the unadjusted sample variance can be expressed as a function of the adjusted sample variance as follows:

$$S_n^2 = \frac{n-1}{n}s_n^2$$

In the lecture entitled *Point estimation of the variance* (p. 579), we have demonstrated that, given the assumptions on the sample ξ_n made above, the adjusted sample variance s_n^2 has a Gamma distribution[10] with parameters $n-1$ and σ^2. Therefore, the random variable W has a Gamma distribution with parameters $n-1$ and 1. Moreover, the random variable Y has a standard normal distribution (see the previous section). Hence, Z_{n-1} is the ratio between a standard normal random variable and the square root of a Gamma random variable with parameters $n-1$ and 1. As a consequence, Z_{n-1} has a standard Student's t distribution with $n-1$ degrees of freedom[11]. ■

The coverage probability of the interval estimator T_n^a is

$$\mathrm{C}\left(T_n^a; \mu, \sigma^2\right) = \mathrm{P}\left(\mu \in T_n^a\right) = \mathrm{P}\left(-z \leq Z_{n-1} \leq z\right)$$

where Z_{n-1} is a standard Student's t random variable with $n-1$ degrees of freedom.

Proof. The coverage probability can be written as

$$\begin{aligned} \mathrm{P}\left(\mu \in T_n^a\right) &= \mathrm{P}\left(\overline{X}_n - \sqrt{\frac{s_n^2}{n}}z \leq \mu \leq \overline{X}_n + \sqrt{\frac{s_n^2}{n}}z\right) \\ &= \mathrm{P}\left(\left\{\overline{X}_n - \sqrt{\frac{s_n^2}{n}}z \leq \mu\right\} \cap \left\{\mu \leq \overline{X}_n + \sqrt{\frac{s_n^2}{n}}z\right\}\right) \end{aligned}$$

[10] See p. 397.

[11] See the lecture entitled *Student's t distribution* (p. 407) for a proof of this fact.

$$\begin{aligned}
&= \mathrm{P}\left(\left\{\overline{X}_n - \mu \leq \sqrt{\frac{s_n^2}{n}}z\right\} \cap \left\{\overline{X}_n - \mu \geq -\sqrt{\frac{s_n^2}{n}}z\right\}\right) \\
&= \mathrm{P}\left(\left\{\frac{\overline{X}_n - \mu}{\sqrt{s_n^2/n}} \leq z\right\} \cap \left\{\frac{\overline{X}_n - \mu}{\sqrt{s_n^2/n}} \geq -z\right\}\right) \\
&= \mathrm{P}\left(-z \leq \frac{\overline{X}_n - \mu}{\sqrt{s_n^2/n}} \leq z\right) \\
&= \mathrm{P}\left(-z \leq Z_{n-1} \leq z\right)
\end{aligned}$$

where we have defined

$$Z_{n-1} = \frac{\overline{X}_n - \mu}{\sqrt{s_n^2/n}}$$

Now, rewrite Z_{n-1} as

$$\begin{aligned}
Z_{n-1} &= \frac{\overline{X}_n - \mu}{\sqrt{s_n^2/n}} = \frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} \cdot \frac{\sqrt{\sigma^2/n}}{\sqrt{s_n^2/n}} \\
&= \frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} \cdot \frac{1}{\sqrt{s_n^2/\sigma^2}} = \frac{Y}{\sqrt{W}}
\end{aligned}$$

where we have defined

$$\begin{aligned}
Y &= \frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} \\
W &= s_n^2/\sigma^2
\end{aligned}$$

In the lecture entitled *Point estimation of the variance* (p. 579), we have demonstrated that, given the assumptions on the sample ξ_n made above, the adjusted sample variance s_n^2 has a Gamma distribution with parameters $n-1$ and σ^2. Therefore, the random variable W has a Gamma distribution with parameters $n-1$ and 1. Moreover, the random variable Y has a standard normal distribution (see the previous section). Hence, Z_{n-1} is the ratio between a standard normal random variable and the square root of a Gamma random variable with parameters $n-1$ and 1. As a consequence, Z_{n-1} has a standard Student's t distribution with $n-1$ degrees of freedom (see also the previous proof). ■

Note that the coverage probability of the confidence interval based on the unadjusted sample variance S_n^2 is lower than the coverage probability of the confidence interval based on the adjusted sample variance s_n^2, because

$$\sqrt{\frac{n-1}{n}}z < z$$

and, as a consequence

$$\begin{aligned}
\mathrm{C}\left(T_n^u; \mu, \sigma^2\right) &= \mathrm{P}\left(-\sqrt{\frac{n-1}{n}}z \leq Z_{n-1} \leq \sqrt{\frac{n-1}{n}}z\right) \\
&< \mathrm{P}\left(-z \leq Z_{n-1} \leq z\right) = \mathrm{C}\left(T_n^a; \mu, \sigma^2\right)
\end{aligned}$$

75.2.4 Confidence coefficient

Note that the coverage probability of both T_n^u and T_n^a does not depend on the unknown parameters μ and σ^2. Therefore, the confidence coefficients of the two confidence intervals coincide with the respective coverage probabilities:

$$\begin{aligned} \mathrm{c}\left(T_n^u\right) &= \inf_{\mu\in\mathbb{R},\sigma^2\in\mathbb{R}_{++}} \mathrm{C}\left(T_n^u;\mu,\sigma^2\right) = \mathrm{P}\left(-\sqrt{\frac{n-1}{n}}z \leq Z_{n-1} \leq \sqrt{\frac{n-1}{n}}z\right) \\ \mathrm{c}\left(T_n^a\right) &= \inf_{\mu\in\mathbb{R},\sigma^2\in\mathbb{R}_{++}} \mathrm{C}\left(T_n^a;\mu,\sigma^2\right) = \mathrm{P}\left(-z \leq Z_{n-1} \leq z\right) \end{aligned}$$

where Z_{n-1} is a standard Student's t distribution with $n-1$ degrees of freedom.

75.2.5 Size

The size of the confidence interval T_n^u is

$$\begin{aligned} \lambda\left(T_n^u\right) &= \lambda\left(\left[\overline{X}_n - \sqrt{\frac{S_n^2}{n}}z, \overline{X}_n + \sqrt{\frac{S_n^2}{n}}z\right]\right) \\ &= \overline{X}_n + \sqrt{\frac{S_n^2}{n}}z - \overline{X}_n + \sqrt{\frac{S_n^2}{n}}z = 2\sqrt{\frac{S_n^2}{n}}z \end{aligned}$$

while the size of the confidence interval T_n^a is

$$\begin{aligned} \lambda\left(T_n^a\right) &= \lambda\left(\left[\overline{X}_n - \sqrt{\frac{s_n^2}{n}}z, \overline{X}_n + \sqrt{\frac{s_n^2}{n}}z\right]\right) \\ &= \overline{X}_n + \sqrt{\frac{s_n^2}{n}}z - \overline{X}_n + \sqrt{\frac{s_n^2}{n}}z = 2\sqrt{\frac{s_n^2}{n}}z \end{aligned}$$

Note that the size of the confidence interval based on the unadjusted sample variance S_n^2 is smaller than the size of the confidence interval based on the adjusted sample variance s_n^2, because

$$S_n^2 < s_n^2$$

and, as a consequence

$$\lambda\left(T_n^u\right) = 2\sqrt{\frac{S_n^2}{n}}z < 2\sqrt{\frac{s_n^2}{n}}z = \lambda\left(T_n^a\right)$$

Thus, the confidence interval based on the unadjusted sample variance has a smaller size and a smaller coverage probability. As we have explained in the lecture entitled *Set estimation* (p. 591), the choice of set estimators is often inspired by the principle of achieving the highest possible coverage probability for a given size or the smallest possible size for a given coverage probability. Following this principle, there is no clear ranking between the estimator based on the unadjusted sample variance and the estimator based on the adjusted sample variance, because the former has smaller size, but the latter has higher coverage probability.

75.2.6 Expected size

The expected size of T_n^u is

$$\mathrm{E}\left[\lambda\left(T_n^u\right)\right] = \mathrm{E}\left[2\sqrt{\frac{S_n^2}{n}}z\right] = \left[\sqrt{\frac{2}{n}}\frac{\Gamma\left(n/2\right)}{\Gamma\left(\left(n-1\right)/2\right)}\right]2\sqrt{\frac{\sigma^2}{n}}z$$

where $\Gamma\left(\right)$ is the Gamma function[12].

Proof. We need to use the fact that S_n^2 has a Gamma distribution with parameters $n-1$ and $\frac{n-1}{n}\sigma^2$. To simplify the notation, set

$$X = S_n^2$$

The probability density function of X is:

$$f_X\left(x\right) = \begin{cases} cx^{(n-1)/2-1}\exp\left(-\frac{n}{\sigma^2}\frac{1}{2}x\right) & \text{if } x \in [0,\infty) \\ 0 & \text{otherwise} \end{cases}$$

where c is a constant:

$$c = \frac{\left(n/\sigma^2\right)^{(n-1)/2}}{2^{(n-1)/2}\Gamma\left(\left(n-1\right)/2\right)}$$

and $\Gamma\left(\right)$ is the Gamma function. Therefore:

$$\begin{aligned}
\mathrm{E}\left[\lambda\left(T_n^u\right)\right] &= \mathrm{E}\left[2\sqrt{\frac{S_n^2}{n}}z\right] = 2z\sqrt{\frac{1}{n}}\mathrm{E}\left[\sqrt{S_n^2}\right] \\
&= 2z\sqrt{\frac{1}{n}}\mathrm{E}\left[X^{1/2}\right] \\
&= 2zn^{-1/2}\int_0^\infty x^{1/2}cx^{(n-1)/2-1}\exp\left(-\frac{n}{\sigma^2}\frac{1}{2}x\right)dx \\
&= 2zn^{-1/2}c\int_0^\infty x^{n/2-1}\exp\left(-\frac{n}{\sigma^2}\frac{1}{2}x\right)dx \\
&= 2zn^{-1/2}c\frac{1}{c_1}\int_0^\infty c_1x^{n/2-1}\exp\left(-\frac{n}{\sigma^2}\frac{1}{2}x\right)dx \\
&= 2zn^{-1/2}c\frac{1}{c_1}
\end{aligned}$$

where we have defined

$$c_1 = \frac{\left(n/\sigma^2\right)^{n/2}}{2^{n/2}\Gamma\left(n/2\right)}$$

and we have used the fact that

$$\int_0^\infty c_1x^{n/2-1}\exp\left(-\frac{n}{\sigma^2}\frac{1}{2}x\right)dx = 1$$

because it is the integral of the density of a Gamma random variable with parameters n and σ^2 over its support and probability densities integrate to 1. Thus:

$$\mathrm{E}\left[\lambda\left(T_n^u\right)\right] = 2zn^{-1/2}c\frac{1}{c_1}$$

[12] See p. 55.

$$\begin{aligned}
&= 2zn^{-1/2}\frac{\left(n/\sigma^2\right)^{(n-1)/2}}{2^{(n-1)/2}\Gamma\left((n-1)/2\right)}\frac{2^{n/2}\Gamma\left(n/2\right)}{\left(n/\sigma^2\right)^{n/2}} \\
&= 2zn^{-1/2}\frac{2^{n/2}}{2^{(n-1)/2}}\frac{\left(n/\sigma^2\right)^{(n-1)/2}}{\left(n/\sigma^2\right)^{n/2}}\frac{\Gamma\left(n/2\right)}{\Gamma\left((n-1)/2\right)} \\
&= 2zn^{-1/2}2^{1/2}\left(n/\sigma^2\right)^{-1/2}\frac{\Gamma\left(n/2\right)}{\Gamma\left((n-1)/2\right)} \\
&= 2z\sqrt{\frac{2}{n}}\sqrt{\frac{\sigma^2}{n}}\frac{\Gamma\left(n/2\right)}{\Gamma\left((n-1)/2\right)} \\
&= \left[\sqrt{\frac{2}{n}}\frac{\Gamma\left(n/2\right)}{\Gamma\left((n-1)/2\right)}\right]2\sqrt{\frac{\sigma^2}{n}}z
\end{aligned}$$

■

The expected size of T_n^a is

$$\mathrm{E}\left[\lambda\left(T_n^a\right)\right] = \mathrm{E}\left[2\sqrt{\frac{s_n^2}{n}}z\right] = \left[\sqrt{\frac{2}{n-1}}\frac{\Gamma\left(n/2\right)}{\Gamma\left((n-1)/2\right)}\right]2\sqrt{\frac{\sigma^2}{n}}z$$

where $\Gamma\left(\right)$ is the Gamma function.

Proof. Using the fact that

$$s_n^2 = \frac{n}{n-1}S_n^2$$

we obtain

$$\begin{aligned}
\mathrm{E}\left[\lambda\left(T_n^a\right)\right] &= \mathrm{E}\left[2\sqrt{\frac{s_n^2}{n}}z\right] = \mathrm{E}\left[2\sqrt{\frac{\frac{n}{n-1}S_n^2}{n}}z\right] \\
&= \sqrt{\frac{n}{n-1}}\mathrm{E}\left[2\sqrt{\frac{S_n^2}{n}}z\right] = \sqrt{\frac{n}{n-1}}\mathrm{E}\left[\lambda\left(T_n^u\right)\right] \\
&= \sqrt{\frac{n}{n-1}}\left[\sqrt{\frac{2}{n}}\frac{\Gamma\left(n/2\right)}{\Gamma\left((n-1)/2\right)}\right]2\sqrt{\frac{\sigma^2}{n}}z \\
&= \left[\sqrt{\frac{2}{n-1}}\frac{\Gamma\left(n/2\right)}{\Gamma\left((n-1)/2\right)}\right]2\sqrt{\frac{\sigma^2}{n}}z
\end{aligned}$$

■

75.3 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Suppose you observe a sample of 100 independent draws from a normal distribution having unknown mean μ and known variance $\sigma^2 = 1$. Denote the 100 draws by $X_1, \ldots, X_{100}$. Suppose their sample mean $\overline{X}_{100}$ is equal to 1, i.e.:

$$\overline{X}_{100} = \frac{1}{100}\sum_{i=1}^{100} X_i = 1$$

Find a confidence interval for μ, using a set estimator of μ having 90% coverage probability.

Solution

For a given sample size n, the interval estimator

$$T_n = \left[\overline{X}_n - \sqrt{\frac{\sigma^2}{n}}z, \overline{X}_n + \sqrt{\frac{\sigma^2}{n}}z\right]$$

has coverage probability

$$\mathrm{C}\left(T_n; \mu\right) = \mathrm{P}\left(\mu \in T_n\right) = \mathrm{P}\left(-z \leq Z \leq z\right)$$

where Z is a standard normal random variable and $z \in \mathbb{R}_{++}$ is a strictly positive constant. Thus, we need to find z such that

$$\mathrm{P}\left(-z \leq Z \leq z\right) = 90\%$$

But

$$\begin{aligned} \mathrm{P}\left(-z \leq Z \leq z\right) &= \mathrm{P}\left(Z \leq z\right) - \mathrm{P}\left(Z < -z\right) \\ &= 1 - \mathrm{P}\left(Z > z\right) - \mathrm{P}\left(Z < -z\right) \\ &= 1 - 2\mathrm{P}\left(Z > z\right) \end{aligned}$$

where the last equality stems from the fact that the standard normal distribution is symmetric around zero. Therefore z must be such that

$$1 - 2\mathrm{P}\left(Z > z\right) = 0.9$$

or

$$\mathrm{P}\left(Z > z\right) = 0.05$$

Using normal distribution tables or a computer program to find the value of z, we obtain

$$z = 1.645$$

Thus, the confidence interval for μ is

$$\begin{aligned} T_{100} &= \left[\overline{X}_{100} - \sqrt{\frac{\sigma^2}{100}}z, \overline{X}_n + \sqrt{\frac{\sigma^2}{100}}z\right] \\ &= \left[1 - \sqrt{\frac{1}{100}} \cdot 1.645, 1 + \sqrt{\frac{1}{100}} \cdot 1.645\right] \\ &= \left[1 - 0.1645, 1 + 0.1645\right] \\ &= \left[0.8355, 1.1645\right] \end{aligned}$$

Exercise 2

Suppose you observe a sample of 100 independent draws from a normal distribution having unknown mean μ and unknown variance σ^2. Denote the 100 draws by X_1, ..., X_{100}. Suppose their sample mean $\overline{X}_{100}$ is equal to 1, i.e.:

$$\overline{X}_{100} = \frac{1}{100}\sum_{i=1}^{100} X_i = 1$$

and their adjusted sample variance s_{100}^2 is equal to 4, i.e.:

$$s_{100}^2 = \frac{1}{99}\sum_{i=1}^{100}\left(X_i - \overline{X}_{100}\right)^2 = 4$$

Find a confidence interval for μ, using a set estimator of μ having 99% coverage probability.

Solution

For a given sample size n, the interval estimator

$$T_n^a = \left[\overline{X}_n - \sqrt{\frac{s_n^2}{n}}z, \overline{X}_n + \sqrt{\frac{s_n^2}{n}}z\right]$$

has coverage probability

$$\mathrm{C}\left(T_n^a; \mu, \sigma^2\right) = \mathrm{P}\left(\mu \in T_n^a\right) = \mathrm{P}\left(-z \leq Z_{n-1} \leq z\right)$$

where Z_{n-1} is a standard Student's t random variable with $n-1$ degrees of freedom and $z \in \mathbb{R}_{++}$ is a strictly positive constant. Thus, we need to find z such that

$$\mathrm{P}\left(-z \leq Z_{n-1} \leq z\right) = 99\%$$

But

$$\begin{aligned}\mathrm{P}\left(-z \leq Z \leq z\right) &= \mathrm{P}\left(Z \leq z\right) - \mathrm{P}\left(Z < -z\right) \\ &= 1 - \mathrm{P}\left(Z > z\right) - \mathrm{P}\left(Z < -z\right) \\ &= 1 - 2\mathrm{P}\left(Z > z\right)\end{aligned}$$

where the last equality stems from the fact that the standard Student's t distribution is symmetric around zero. Therefore z must be such that

$$1 - 2\mathrm{P}\left(Z > z\right) = 0.99$$

or

$$\mathrm{P}\left(Z > z\right) = 0.005$$

Using a computer program to find the value of z (for example, with the MATLAB command `tinv(0.995,99)`), we obtain

$$z = 2.6264$$

Thus, the confidence interval for μ is

$$\begin{aligned}T_{100} &= \left[\overline{X}_{100} - \sqrt{\frac{s_{100}^2}{100}}z, \overline{X}_n + \sqrt{\frac{s_{100}^2}{100}}z\right] \\ &= \left[1 - \sqrt{\frac{4}{100}}\cdot 2.6264, 1 + \sqrt{\frac{4}{100}}\cdot 2.6264\right] \\ &= \left[1 - 0.5253, 1 + 0.5253\right] \\ &= \left[0.4747, 1.5253\right]\end{aligned}$$

Chapter 76

Set estimation of the variance

This lecture presents some examples of set estimation[1] problems, focusing on **set estimation of the variance**, i.e. on using a sample to produce a set estimate of the variance of an unknown distribution.

76.1 Normal IID samples - Known mean

In this example we make the same assumptions we made in the example of point estimation of the variance entitled *Normal IID samples - Known mean* (p. 579). The reader is strongly advised to read that example before reading this one.

76.1.1 The sample

The sample ξ_n is made of n independent draws from a normal distribution having known mean μ and unknown variance σ^2. Specifically, we observe n realizations $x_1, \ldots, x_n$ of n independent random variables $X_1, \ldots, X_n$, all having a normal distribution with known mean μ and unknown variance σ^2. The sample[2] is the n-dimensional vector

$$\xi_n = \begin{bmatrix} x_1 & \ldots & x_n \end{bmatrix}$$

which is a realization of the random vector

$$\Xi_n = \begin{bmatrix} X_1 & \ldots & X_n \end{bmatrix}$$

76.1.2 The interval estimator

The interval estimator[3] of the variance σ^2, is based on the following point estimator of the variance:

$$\widehat{\sigma_n^2} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$$

[1] See p. 591.
[2] See p. 564.
[3] See p. 591.

The interval estimator is

$$T_n = \left[\frac{n}{z_2} \widehat{\sigma_n^2}, \frac{n}{z_1} \widehat{\sigma_n^2} \right]$$

where z_1 and z_2 are strictly positive constants and $z_1 < z_2$.

76.1.3 Coverage probability

The coverage probability[4] of the interval estimator T_n is

$$\mathrm{C}\left(T_n; \sigma^2\right) = \mathrm{P}\left(\sigma^2 \in T_n\right) = \mathrm{P}\left(z_1 \leq Z \leq z_2\right)$$

where Z is a Chi-square random variable[5] with n degrees of freedom.

Proof. The coverage probability can be written as

$$\begin{aligned}
\mathrm{P}\left(\sigma^2 \in T_n\right) &= \mathrm{P}\left(\frac{n}{z_2} \widehat{\sigma_n^2} \leq \sigma^2 \leq \frac{n}{z_1} \widehat{\sigma_n^2} \right) \\
&= \mathrm{P}\left(\left\{ \frac{n}{z_2} \widehat{\sigma_n^2} \leq \sigma^2 \right\} \cap \left\{ \sigma^2 \leq \frac{n}{z_1} \widehat{\sigma_n^2} \right\} \right) \\
&= \mathrm{P}\left(\left\{ \frac{n}{\sigma^2} \widehat{\sigma_n^2} \leq z_2 \right\} \cap \left\{ z_1 \leq \frac{n}{\sigma^2} \widehat{\sigma_n^2} \right\} \right) \\
&= \mathrm{P}\left(z_1 \leq \frac{n}{\sigma^2} \widehat{\sigma_n^2} \leq z_2 \right) \\
&= \mathrm{P}\left(z_1 \leq Z \leq z_2\right)
\end{aligned}$$

where we have defined

$$Z = \frac{n}{\sigma^2} \widehat{\sigma_n^2}$$

In the lecture entitled *Point estimation of the variance* (p. 579), we have demonstrated that, given the assumptions on the sample ξ_n made above, the estimator $\widehat{\sigma_n^2}$ has a Gamma distribution[6] with parameters n and σ^2. Multiplying a Gamma random variable with parameters n and σ^2 by $\frac{n}{\sigma^2}$ one obtains a Chi-square random variable with n degrees of freedom. Therefore, the variable Z has a Chi-square distribution with n degrees of freedom. ■

76.1.4 Confidence coefficient

Note that the coverage probability does not depend on the unknown parameter σ^2. Therefore, the confidence coefficient[7] of the interval estimator T_n coincides with its coverage probability:

$$\mathrm{c}\left(T_n\right) = \inf_{\sigma^2 \in \mathbb{R}_{++}} \mathrm{C}\left(T_n, \sigma^2\right) = \mathrm{P}\left(z_1 \leq Z \leq z_2\right)$$

where Z is a Chi-square random variable with n degrees of freedom.

[4] See p. 592.
[5] See p. 387.
[6] See p. 397.
[7] See p. 592.

76.1.5 Size

The size[8] of the interval estimator T_n is

$$\begin{aligned}\lambda(T_n) &= \lambda\left(\left[\frac{n}{z_2}\widehat{\sigma_n^2}, \frac{n}{z_1}\widehat{\sigma_n^2}\right]\right) \\ &= \frac{n}{z_1}\widehat{\sigma_n^2} - \frac{n}{z_2}\widehat{\sigma_n^2} \\ &= n\left(\frac{1}{z_1} - \frac{1}{z_2}\right)\widehat{\sigma_n^2}\end{aligned}$$

76.1.6 Expected size

Note that the size depends on $\widehat{\sigma_n^2}$ and hence on the sample ξ_n. The expected size of the interval estimator T_n is

$$\begin{aligned}\mathrm{E}[\lambda(T_n)] &= \mathrm{E}\left[n\left(\frac{1}{z_1} - \frac{1}{z_2}\right)\widehat{\sigma_n^2}\right] \\ &= n\left(\frac{1}{z_1} - \frac{1}{z_2}\right)\mathrm{E}\left[\widehat{\sigma_n^2}\right] \\ &= n\left(\frac{1}{z_1} - \frac{1}{z_2}\right)\sigma^2\end{aligned}$$

where we have used the fact that $\widehat{\sigma_n^2}$ is an unbiased estimator of σ^2 (i.e. $\mathrm{E}\left[\widehat{\sigma_n^2}\right] = \sigma^2$, see p. 580).

76.2 Normal IID samples - Unknown mean

This example is similar to the previous one. The only difference is that we now relax the assumption that the mean of the distribution is known.

76.2.1 The sample

In this example, the sample ξ_n is made of n independent draws from a normal distribution having unknown mean μ and unknown variance σ^2. Specifically, we observe n realizations $x_1, \ldots, x_n$ of n independent random variables $X_1, \ldots, X_n$, all having a normal distribution with unknown mean μ and unknown variance σ^2. The sample is the n-dimensional vector

$$\xi_n = [x_1 \quad \ldots \quad x_n]$$

which is a realization of the random vector

$$\Xi_n = [X_1 \quad \ldots \quad X_n]$$

[8] See p. 592.

76.2.2 The interval estimator

To construct interval estimators of the variance σ^2, we use the sample mean $\overline{X}_n$:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and either the unadjusted sample variance[9]:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$$

or the adjusted sample variance:

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$$

We consider the following interval estimator of the variance:

$$T_n = \left[\frac{n}{z_2}S_n^2, \frac{n}{z_1}S_n^2\right] = \left[\frac{n-1}{z_2}s_n^2, \frac{n-1}{z_1}s_n^2\right]$$

where z_1 and z_2 are strictly positive constants and $z_1 < z_2$.

76.2.3 Coverage probability

The coverage probability of the interval estimator T_n is

$$\mathrm{C}\left(T_n; \mu, \sigma^2\right) = \mathrm{P}\left(\sigma^2 \in T_n\right) = \mathrm{P}\left(z_1 \leq Z_{n-1} \leq z_2\right)$$

where Z_{n-1} is a Chi-square random variable with $n-1$ degrees of freedom.

Proof. The coverage probability can be written as

$$\begin{aligned}
\mathrm{P}\left(\sigma^2 \in T_n\right) &= \mathrm{P}\left(\frac{n}{z_2}S_n^2 \leq \sigma^2 \leq \frac{n}{z_1}S_n^2\right) \\
&= \mathrm{P}\left(\left\{\frac{n}{z_2}S_n^2 \leq \sigma^2\right\} \cap \left\{\sigma^2 \leq \frac{n}{z_1}S_n^2\right\}\right) \\
&= \mathrm{P}\left(\left\{\frac{n}{\sigma^2}S_n^2 \leq z_2\right\} \cap \left\{z_1 \leq \frac{n}{\sigma^2}S_n^2\right\}\right) \\
&= \mathrm{P}\left(z_1 \leq \frac{n}{\sigma^2}S_n^2 \leq z_2\right) \\
&= \mathrm{P}\left(z_1 \leq Z_{n-1} \leq z_2\right)
\end{aligned}$$

where we have defined

$$Z_{n-1} = \frac{n}{\sigma^2}S_n^2$$

In the lecture entitled *Point estimation of the variance* (p. 579), we have demonstrated that, given the assumptions on the sample ξ_n made above, the unadjusted sample variance S_n^2 has a Gamma distribution with parameters $n-1$ and $\frac{n-1}{n}\sigma^2$.

[9] See p. 583 for a definition and a discussion of adjusted and unadjusted sample variance.

Therefore, the random variable Z_{n-1} has a Gamma distribution with parameters $n-1$ and h where

$$h = \left(\frac{n}{\sigma^2}\right)\left(\frac{n-1}{n}\sigma^2\right) = n-1$$

But a Gamma distribution with parameters $n-1$ and $n-1$ is a Chi-square distribution with $n-1$ degrees of freedom. Therefore, Z_{n-1} has a Chi-square distribution with $n-1$ degrees of freedom. ■

76.2.4 Confidence coefficient

Note that the coverage probability of T_n does not depend on the unknown parameters μ and σ^2. Therefore, the confidence coefficient of the confidence interval coincides with its coverage probability:

$$\mathrm{c}\left(T_n\right) = \inf_{\mu\in\mathbb{R},\sigma^2\in\mathbb{R}_{++}} \mathrm{C}\left(T_n;\mu,\sigma^2\right) = \mathrm{P}\left(z_1 \leq Z_{n-1} \leq z_2\right)$$

where Z_{n-1} is a Chi-square distribution with $n-1$ degrees of freedom.

76.2.5 Size

The size of the confidence interval T_n is

$$\begin{aligned}\lambda\left(T_n\right) &= \lambda\left(\left[\frac{n}{z_2}S_n^2, \frac{n}{z_1}S_n^2\right]\right) \\ &= n\left(\frac{1}{z_1}-\frac{1}{z_2}\right)S_n^2\end{aligned}$$

76.2.6 Expected size

The expected size of T_n is

$$\begin{aligned}\mathrm{E}\left[\lambda\left(T_n\right)\right] &= \mathrm{E}\left[n\left(\frac{1}{z_1}-\frac{1}{z_2}\right)S_n^2\right] \\ &= n\left(\frac{1}{z_1}-\frac{1}{z_2}\right)\mathrm{E}\left[S_n^2\right] \\ &= n\left(\frac{1}{z_1}-\frac{1}{z_2}\right)\frac{n-1}{n}\sigma^2 \\ &= (n-1)\left(\frac{1}{z_1}-\frac{1}{z_2}\right)\sigma^2\end{aligned}$$

where in the penultimate step we have used the fact (proved in the lecture entitled *Point estimation of the variance* - p. 579) that

$$\mathrm{E}\left[S_n^2\right] = \frac{n-1}{n}\sigma^2$$

76.3 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Suppose you observe a sample of 100 independent draws from a normal distribution having known mean $\mu = 0$ and unknown variance σ^2. Denote the 100 draws by $X_1, \ldots, X_{100}$. Suppose that

$$\widehat{\sigma^2_{100}} = \frac{1}{100}\sum_{i=1}^{n} X_i^2 = 1$$

Find a confidence interval for σ^2, using a set estimator of σ^2 having 90% coverage probability.

Hint: a Chi-square random variable Z with 100 degrees of freedom has a distribution function $F_Z(z)$ such that

$$\begin{aligned} F_Z(77.9295) &= 0.05 \\ F_Z(124.3421) &= 0.95 \end{aligned}$$

Solution

For a given sample size n, the interval estimator

$$T_n = \left[\frac{n}{z_2}\widehat{\sigma^2_n}, \frac{n}{z_1}\widehat{\sigma^2_n}\right]$$

has coverage probability

$$\mathrm{C}\left(T_n; \sigma^2\right) = \mathrm{P}\left(\sigma^2 \in T_n\right) = \mathrm{P}\left(z_1 \leq Z \leq z_2\right)$$

where Z is a Chi-square random variable with n degrees of freedom and $z_1, z_2 \in \mathbb{R}_{++}$ are strictly positive constants. Thus, if we set

$$\begin{aligned} z_1 &= 77.9295 \\ z_2 &= 124.3421 \end{aligned}$$

then

$$\begin{aligned} \mathrm{P}\left(z_1 \leq Z \leq z_2\right) &= \mathrm{P}\left(Z \leq z_2\right) - \mathrm{P}\left(Z < z_1\right) \\ \boxed{\text{A}} &= \mathrm{P}\left(Z \leq z_2\right) - \mathrm{P}\left(Z \leq z_1\right) \\ \boxed{\text{B}} &= F_Z(z_2) - F_Z(z_1) \\ &= F_Z(124.3421) - F_Z(77.9295) \\ &= 0.95 - 0.05 = 0.9 \end{aligned}$$

which is equal to our desired coverage probability (in step A we have used the fact that any specific realization of an absolutely continuous random variable has zero probability; in step B we have used the definition of distribution function). Thus, the confidence interval for σ^2 is

$$\begin{aligned} T_{100} &= \left[\frac{100}{z_2}\widehat{\sigma^2_{100}}, \frac{100}{z_1}\widehat{\sigma^2_{100}}\right] \\ &= \left[\frac{100}{124.3421}, \frac{100}{77.9295}\right] \\ &= [0.8042, 1.2832] \end{aligned}$$

Exercise 2

Suppose you observe a sample of 100 independent draws from a normal distribution having unknown mean μ and unknown variance σ^2. Denote the 100 draws by X_1, ..., X_{100}. Suppose that their adjusted sample variance s^2_{100} is equal to 5, i.e.:

$$s^2_{100} = \frac{1}{99}\sum_{i=1}^{100}\left(X_i - \overline{X}_{100}\right)^2 = 5$$

Find a confidence interval for σ^2, using a set estimator of σ^2 having 99% coverage probability.

Hint: a Chi-square random variable Z with 99 degrees of freedom has a distribution function $F_Z(z)$ such that

$$\begin{aligned} F_Z(66.5101) &= 0.005 \\ F_Z(138.9868) &= 0.995 \end{aligned}$$

Solution

For a given sample size n, the interval estimator

$$T_n = \left[\frac{n-1}{z_2}s^2_n, \frac{n-1}{z_1}s^2_n\right]$$

has coverage probability

$$\mathrm{C}\left(T_n; \mu, \sigma^2\right) = \mathrm{P}\left(\sigma^2 \in T_n\right) = \mathrm{P}\left(z_1 \leq Z \leq z_2\right)$$

where Z is a Chi-square random variable with $n-1$ degrees of freedom and $z_1, z_2 \in \mathbb{R}_{++}$ are strictly positive constants. Thus, if we set

$$\begin{aligned} z_1 &= 66.5101 \\ z_2 &= 138.9868 \end{aligned}$$

then

$$\begin{aligned} \mathrm{P}\left(z_1 \leq Z \leq z_2\right) &= \mathrm{P}\left(Z \leq z_2\right) - \mathrm{P}\left(Z < z_1\right) \\ \boxed{\text{A}} \quad &= \mathrm{P}\left(Z \leq z_2\right) - \mathrm{P}\left(Z \leq z_1\right) \\ \boxed{\text{B}} \quad &= F_Z(z_2) - F_Z(z_1) \\ &= F_Z(138.9868) - F_Z(66.5101) \\ &= 0.995 - 0.005 = 0.99 \end{aligned}$$

which is equal to our desired coverage probability (in step $\boxed{\text{A}}$ we have used the fact that any specific realization of an absolutely continuous random variable has zero probability; in step $\boxed{\text{B}}$ we have used the definition of distribution function). Thus, the confidence interval for σ^2 is

$$\begin{aligned} T_{100} &= \left[\frac{99}{z_2}s^2_{100}, \frac{99}{z_1}s^2_{100}\right] \\ &= \left[\frac{99}{138.9868}\cdot 5, \frac{99}{66.5101}\cdot 5\right] \\ &= [3.5615, 7.4425] \end{aligned}$$

Chapter 77

Hypothesis testing

Hypothesis testing is a method of making statistical inferences.

As we have discussed in the lecture entitled *Statistical inference* (p. 563), a statistical inference is a statement about the probability distribution from which a sample ξ has been drawn. The sample ξ can be regarded as a realization of a random vector Ξ, whose unknown joint distribution function[1] $F_{\Xi}(\xi)$ is assumed to belong to a set of distribution functions Φ, called statistical model.

In **hypothesis testing** we make a statement about a model restriction involving a subset $\Phi_R \subset \Phi$ of the original model. The statement we make is chosen between two possible statements:

1. reject the restriction $F_{\Xi} \in \Phi_R$;
2. do not reject the restriction $F_{\Xi} \in \Phi_R$.

Roughly speaking, we start from a large set Φ of distributions that might possibly have generated the sample ξ and we would like to restrict our attention to a smaller set Φ_R. In a test of hypothesis, we use the sample ξ to decide whether or not to restrict our attention to the smaller set Φ_R.

If we have a parametric model, we can also carry out parametric tests of hypothesis.

Remember that in a parametric model the set of distribution functions Φ is put into correspondence with a set $\Theta \subseteq \mathbb{R}^p$ of p-dimensional real vectors called the parameter space. The elements of Θ are called parameters. Denote by θ_0 the parameter that is associated with the unknown distribution function $F_{\Xi}(\xi)$ and assume that θ_0 is unique. θ_0 is called the **true parameter**, because it is associated to the distribution that actually generated the sample.

In parametric hypothesis testing we have a restriction $\Theta_R \subset \Theta$ on the parameter space and we choose one of the following two statements about the restriction:

1. reject the restriction $\theta_0 \in \Theta_R$;
2. do not reject the restriction $\theta_0 \in \Theta_R$.

For concreteness, we will focus on parametric hypothesis testing in this lecture, but most of the things we will say apply with straightforward modifications to hypothesis testing in general.

[1] See p. 118.

77.1 Null hypothesis

The hypothesis that the restriction is true is called **null hypothesis** and it is usually denoted by H_0:

$$H_0 : \theta_0 \in \Theta_R$$

77.2 Alternative hypothesis

The restriction $\theta_0 \in \Theta_R^c$ (where Θ_R^c is the complement of Θ_R) is often called **alternative hypothesis** and it is denoted by H_1:

$$H_1 : \theta_0 \in \Theta_R^c$$

For some authors, "rejecting the null hypothesis H_0" and "accepting the alternative hypothesis H_1" are synonyms. For other authors, however, "rejecting the null hypothesis H_0" does not necessarily imply "accepting the alternative hypothesis H_1". Although this is mostly a matter of language, it is possible to envision situations in which, after rejecting H_0, a second test of hypothesis is performed whereby H_1 becomes the new null hypothesis and it is rejected (this may happen for example if the model is mis-specified[2]). In these situations, if "rejecting the null hypothesis H_0" and "accepting the alternative hypothesis H_1" are treated as synonyms, then some confusion arises, because the first test leads to "accept H_1" and the second test leads to "reject H_1".

Also note that some statisticians sometimes take into consideration as an alternative hypothesis a set smaller than Θ_R^c. In these cases, the null hypothesis and the alternative hypothesis do not cover all the possibilities contemplated by the parameter space Θ.

77.3 Types of errors

When we decide whether to reject a restriction or not to reject it, we can incur in two types of errors:

1. reject the restriction $\theta_0 \in \Theta_R$ when the restriction is true; this is called an **error of the first kind** or a **Type I error**;

2. do not reject the restriction $\theta_0 \in \Theta_R$ when the restriction is false; this is called an **error of the second kind** or a **Type II error**.

77.4 Critical region

Remember that the sample ξ is regarded as a realization of a random vector Ξ having support R_Ξ.

A test of hypothesis is usually carried out by explicitly or implicitly subdividing the support R_Ξ into two disjoint subsets. One of the two subsets, denoted by C_Ξ

[2] See p. 565.

is called the **critical region** (or **rejection region**) and it is the set of all values of ξ for which the null hypothesis is rejected:

$$C_{\Xi} = \{\xi \in R_{\Xi} : H_0 \text{ is rejected whenever } \xi \text{ is observed}\}$$

The other subset is just the complement of the critical region:

$$C_{\Xi}^{c} = \{\xi \in R_{\Xi} : H_0 \text{ is not rejected whenever } \xi \text{ is observed}\}$$

and it is, of course, such that

$$C_{\Xi} \cup C_{\Xi}^{c} = R_{\Xi}$$

77.5 Test statistic

The critical region is often implicitly defined in terms of a test statistic and a critical region for the test statistic. A **test statistic** is a random variable S whose realization is a function of the sample ξ. In symbols:

$$S = s(\Xi)$$

A critical region for S is a subset $C_S \subset \mathbb{R}$ of the set of real numbers and the test is performed based on the test statistic, as follows:

$$\begin{aligned} s(\xi) &\in C_S \Longrightarrow \xi \in C_{\Xi} \Longrightarrow H_0 \text{ is rejected} \\ s(\xi) &\notin C_S \Longrightarrow \xi \notin C_{\Xi} \Longrightarrow H_0 \text{ is not rejected} \end{aligned}$$

If the complement of the critical region for S is an interval, i.e.

$$C_S^{c} = [s_l, s_u]$$

the lower bound of the interval s_l and the upper bound s_u are called **critical values** of the test.

77.6 Power function

The **power function** of a test of hypothesis is the function that associates the probability of rejecting H_0 to each parameter $\theta \in \Theta$. Denoting the critical region by C_{Ξ}, the power function $\pi(\theta)$ is defined as follows:

$$\pi(\theta) = \mathrm{P}_{\theta}(\Xi \in C_{\Xi})$$

where the notation P_{θ} is used to indicate the fact that the probability is calculated using the distribution function $F_{\Xi}(\xi; \theta)$ associated to the parameter θ.

77.7 Size of a test

When $\theta \in \Theta_R$, the power function $\pi(\theta)$ gives us the probability of committing a Type I error, i.e. the probability of rejecting the null hypothesis when the null hypothesis is true.

The maximum probability of committing a Type I error is

$$\alpha = \sup_{\theta \in \Theta_R} \pi(\theta)$$

and it is called the **size of the test**. The size of the test is also called by some authors the **level of significance of the test**. However, according to other authors, who assign a slightly different meaning to the term, the level of significance of a test is an upper bound on the size of the test, i.e. a constant α that, to the statistician's knowledge, satisfies:

$$\alpha \geq \sup_{\theta \in \Theta_R} \pi(\theta)$$

77.8 Criteria to evaluate tests

Tests of hypothesis are most commonly evaluated based on their size and power. An ideal test should have size equal to 0 (i.e. the probability of rejecting the null hypothesis when the null hypothesis is true should be 0) and power equal to 1 when $\theta_0 \notin \Theta_R$ (i.e. the probability of rejecting the null hypothesis when the null hypothesis is false should be 1). Of course, such an ideal test is never found in practice, but the best we can hope for is a test with a very small size and a very high probability of rejecting a false hypothesis. Nevertheless, this ideal is routinely used to choose among different tests: for example, when choosing between two tests having the same size, we will always utilize the test that has the higher power when $\theta_0 \notin \Theta_R$; also, when choosing between two tests that have the same power when $\theta_0 \notin \Theta_R$, we will always utilize the test that has the smaller size.

Several other criteria, beyond power and size, are used to evaluate tests of hypothesis. We do not discuss them here, but we refer the reader to the very nice exposition in Berger and Casella[3] (2002).

77.9 Examples

You can find examples of hypothesis testing in the lectures entitled *Hypothesis tests about the mean* (p. 619) and *Hypothesis tests about the variance* (p. 619).

[3]Berger, R. L. and G. Casella (2002) "Statistical inference", Duxbury Advanced Series.

Chapter 78

Hypothesis tests about the mean

This lecture presents some examples of hypothesis testing[1], focusing on **tests of hypothesis about the mean**, that is, on using a sample to perform tests of hypothesis about the mean of an unknown distribution.

78.1 Normal IID samples - Known variance

In this example we make the same assumptions we made in the example of set estimation of the mean entitled *Normal IID samples - Known variance* (p. 595). The reader is strongly advised to read that example before reading this one.

78.1.1 The sample

The sample ξ_n is made of n independent draws from a normal distribution having unknown mean μ and known variance σ^2. Specifically, we observe n realizations $x_1, \ldots, x_n$ of n independent random variables $X_1, \ldots, X_n$, all having a normal distribution with unknown mean μ and known variance σ^2. The sample is the n-dimensional vector

$$\xi_n = \begin{bmatrix} x_1 & \ldots & x_n \end{bmatrix}$$

which is a realization of the random vector

$$\Xi_n = \begin{bmatrix} X_1 & \ldots & X_n \end{bmatrix}$$

78.1.2 The null hypothesis

We test the null hypothesis[2] that the mean μ is equal to a specific value μ_0:

$$H_0 : \mu = \mu_0$$

[1] See p. 615.
[2] See p. 616.

78.1.3 The alternative hypothesis

We assume that the parameter space is the whole real line, i.e., $\mu \in \mathbb{R}$. Therefore, the alternative hypothesis[3] is

$$H_1 : \mu \neq \mu_0$$

78.1.4 The test statistic

To construct a test statistic[4], we use the sample mean

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

The test statistic is

$$Z_n = \frac{\overline{X}_n - \mu_0}{\sqrt{\sigma^2/n}}$$

This test statistic is often called **z-statistic** or **normal z-statistic**, and a test of hypothesis based on this statistic is called **z-test** or **normal z-test**.

78.1.5 The critical region

Let $z \in \mathbb{R}_{++}$. We reject the null hypothesis H_0 if $Z_n > z$ or if $Z_n < -z$. In other words, the critical region[5] is

$$C_{Z_n} = (-\infty, -z) \cup (z, \infty)$$

Thus, the critical values of the test are $-z$ and z.

78.1.6 The power function

The power function[6] of the test is

$$\pi(\mu) = \mathrm{P}_\mu(Z_n \notin [-z, z]) = 1 - \mathrm{P}\left(-z + \frac{\mu_0 - \mu}{\sqrt{\sigma^2/n}} \leq Z \leq z + \frac{\mu_0 - \mu}{\sqrt{\sigma^2/n}}\right) \quad (78.1)$$

where Z is a standard normal random variable, and the notation P_μ is used to indicate the fact that the probability of rejecting the null hypothesis is computed under the hypothesis that the true mean is equal to μ.

Proof. The power function can be written as

$$\begin{aligned}
\pi(\mu) &= \mathrm{P}_\mu(Z_n \notin [-z, z]) \\
&= 1 - \mathrm{P}_\mu(Z_n \in [-z, z]) \\
&= 1 - \mathrm{P}_\mu\left(-z \leq \frac{\overline{X}_n - \mu_0}{\sqrt{\sigma^2/n}} \leq z\right) \\
&= 1 - \mathrm{P}_\mu\left(-z + \frac{\mu_0}{\sqrt{\sigma^2/n}} \leq \frac{\overline{X}_n}{\sqrt{\sigma^2/n}} \leq z + \frac{\mu_0}{\sqrt{\sigma^2/n}}\right)
\end{aligned}$$

[3] See p. 616.
[4] See p. 617.
[5] See p. 616.
[6] See p. 617.

$$\begin{aligned} &= 1 - \mathrm{P}_\mu\left(-z + \frac{\mu_0 - \mu}{\sqrt{\sigma^2/n}} \leq \frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} \leq z + \frac{\mu_0 - \mu}{\sqrt{\sigma^2/n}}\right) \\ &= 1 - \mathrm{P}_\mu\left(-z + \frac{\mu_0 - \mu}{\sqrt{\sigma^2/n}} \leq Z \leq z + \frac{\mu_0 - \mu}{\sqrt{\sigma^2/n}}\right) \end{aligned}$$

where we have defined

$$Z = \frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}}$$

As demonstrated in the lecture entitled *Point estimation of the mean* (p. 573), the sample mean $\overline{X}_n$ has a normal distribution with mean μ and variance σ^2/n, given the assumptions on the sample ξ_n we made above. By subtracting the mean of a normal random variable from the random variable itself, and dividing it by the square root of its variance, one obtains a standard normal random variable. Therefore, the variable Z has a standard normal distribution. ■

78.1.7 The size of the test

When evaluated at the point $\mu = \mu_0$, the power function is equal to the probability of committing a Type I error, that is, the probability of rejecting the null hypothesis when the null hypothesis is true. This probability is called the size of the test[7] and it is equal to

$$\pi(\mu_0) = \mathrm{P}_{\mu_0}(Z_n \notin [-z, z]) = 1 - \mathrm{P}(-z \leq Z \leq z)$$

where Z is a standard normal random variable.

Proof. This is trivially obtained by substituting μ with μ_0 in formula (78.1). ■

78.2 Normal IID samples - Unknown variance

This example is similar to the previous one. The only difference is that we now relax the assumption that the variance of the distribution is known.

78.2.1 The sample

In this example, the sample ξ_n is made of n independent draws from a normal distribution having unknown mean μ and unknown variance σ^2. Specifically, we observe n realizations $x_1, \ldots, x_n$ of n independent random variables $X_1, \ldots, X_n$, all having a normal distribution with unknown mean μ and unknown variance σ^2. The sample is the n-dimensional vector

$$\xi_n = [x_1 \quad \ldots \quad x_n]$$

which is a realization of the random vector

$$\Xi_n = [X_1 \quad \ldots \quad X_n]$$

[7] See p. 617.

78.2.2 The null hypothesis

We test the null hypothesis that the mean μ is equal to a specific value μ_0:

$$H_0 : \mu = \mu_0$$

78.2.3 The alternative hypothesis

We assume that the parameter space is the whole real line, i.e., $\mu \in \mathbb{R}$. Therefore, the alternative hypothesis is

$$H_1 : \mu \neq \mu_0$$

78.2.4 The test statistic

We construct two test statistics, using the sample mean

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

and either the unadjusted sample variance[8]

$$S_n^2 = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X}_n \right)^2$$

or the adjusted sample variance

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}_n \right)^2$$

The two test statistics are

$$\begin{aligned} Z_n^u &= \frac{\overline{X}_n - \mu_0}{\sqrt{S_n^2/n}} \\ Z_n^a &= \frac{\overline{X}_n - \mu_0}{\sqrt{s_n^2/n}} \end{aligned}$$

where the superscripts u and a indicate whether the test statistic is based on the unadjusted or the adjusted sample variance. These two test statistics are often called **t-statistics** or **Student's t-statistics**, and tests of hypothesis based on these statistics are called **t-tests** or **Student's t-tests**.

78.2.5 The critical region

Let $z \in \mathbb{R}_{++}$. We reject the null hypothesis H_0 if $Z_n^i > z$ or if $Z_n^i < -z$ (for $i = u$ or $i = a$). In other words, the critical region is

$$C_{Z_n^i} = (-\infty, -z) \cup (z, \infty)$$

Thus, the critical values of the test are $-z$ and z.

[8] See p. 583.

78.2.6 The power function

The power function of the test based on the unadjusted sample variance is

$$\pi^u(\mu) = \mathrm{P}_\mu\left(Z_n^u \notin [-z,z]\right) = 1 - \mathrm{P}\left(-\sqrt{\frac{n-1}{n}}\cdot z \leq W_{n-1} \leq \sqrt{\frac{n-1}{n}}\cdot z\right) \quad (78.2)$$

where the notation P_μ is used to indicate the fact that the probability of rejecting the null hypothesis is computed under the hypothesis that the true mean is equal to μ, and W_{n-1} is a non-central standard Student's t distribution[9] with $n-1$ degrees of freedom and non-centrality parameter equal to

$$\frac{\mu - \mu_0}{\sqrt{\sigma^2/n}}$$

Proof. The power function can be written as

$$\begin{aligned}
\pi^u(\mu) &= \mathrm{P}_\mu\left(Z_n^u \notin [-z,z]\right) \\
&= 1 - \mathrm{P}_\mu\left(Z_n^u \in [-z,z]\right) \\
&= 1 - \mathrm{P}_\mu\left(-z \leq \frac{\overline{X}_n - \mu_0}{\sqrt{S_n^2/n}} \leq z\right) \\
&= 1 - \mathrm{P}_\mu\left(-\sqrt{\frac{n-1}{n}}\cdot z \leq \frac{\left(\overline{X}_n - \mu\right) + \left(\mu - \mu_0\right)}{\sqrt{S_n^2/\left(n-1\right)}} \leq \sqrt{\frac{n-1}{n}}\cdot z\right) \\
&= 1 - \mathrm{P}_\mu\left(-\sqrt{\frac{n-1}{n}}\cdot z \leq \frac{\frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} + \frac{\mu - \mu_0}{\sqrt{\sigma^2/n}}}{\sqrt{S_n^2 \cdot \frac{n}{(n-1)\sigma^2}}} \leq \sqrt{\frac{n-1}{n}}\cdot z\right) \\
&= 1 - \mathrm{P}\left(-\sqrt{\frac{n-1}{n}}\cdot z \leq W_{n-1} \leq \sqrt{\frac{n-1}{n}}\cdot z\right)
\end{aligned}$$

where we have defined

$$W_{n-1} = \frac{\left(\overline{X}_n - \mu\right)\Big/\sqrt{\sigma^2/n} + \left(\mu - \mu_0\right)\Big/\sqrt{\sigma^2/n}}{\sqrt{S_n^2 \cdot \frac{n}{(n-1)\sigma^2}}}$$

Given the assumptions on the sample ξ_n we made above, the sample mean $\overline{X}_n$ has a normal distribution with[10] mean μ and variance σ^2/n, so that the random variable

$$\left(\overline{X}_n - \mu\right)\Big/\sqrt{\sigma^2/n}$$

has a standard normal distribution. Furthermore, the unadjusted sample variance S_n^2 has a Gamma distribution with[11] parameters $n-1$ and $\frac{n-1}{n}\sigma^2$, so that the random variable

$$S_n^2 \cdot \frac{n}{(n-1)\,\sigma^2}$$

has a Gamma distribution with parameters $n-1$ and 1. By adding a constant c to a standard normal distribution and dividing the sum thus obtained by the square

[9] See p. 418.
[10] See p. 574.
[11] See p. 585.

root of a Gamma random variable with parameters $n-1$ and 1, one obtains a non-central standard Student's t distribution with $n-1$ degrees of freedom and non-centrality parameter c. Therefore, the random variable W_{n-1} has a non-central standard Student's t distribution with $n-1$ degrees of freedom and non-centrality parameter

$$(\mu - \mu_0) \Big/ \sqrt{\sigma^2/n}$$

■

The power function of the test based on the adjusted sample variance is

$$\pi^a(\mu) = \mathrm{P}_\mu\left(Z_n^a \notin [-z, z]\right) = 1 - \mathrm{P}\left(-z \leq W_{n-1} \leq z\right)$$

where the notation P_μ is used to indicate the fact that the probability of rejecting the null hypothesis is computed under the hypothesis that the true mean is equal to μ, and W_{n-1} is a non-central standard Student's t distribution with $n-1$ degrees of freedom and non-centrality parameter equal to

$$\frac{\mu - \mu_0}{\sqrt{\sigma^2/n}}$$

Proof. The power function can be written as

$$\begin{aligned}
\pi^a(\mu) &= \mathrm{P}_\mu\left(Z_n^a \notin [-z, z]\right) \\
&= 1 - \mathrm{P}_\mu\left(Z_n^a \in [-z, z]\right) \\
&= 1 - \mathrm{P}_\mu\left(-z \leq \frac{\overline{X}_n - \mu_0}{\sqrt{s_n^2/n}} \leq z\right) \\
&= 1 - \mathrm{P}_\mu\left(-z \leq \frac{\left(\overline{X}_n - \mu\right) + \left(\mu - \mu_0\right)}{\sqrt{s_n^2/n}} \leq z\right) \\
&= 1 - \mathrm{P}_\mu\left(-z \leq \frac{\left(\overline{X}_n - \mu\right) \Big/ \sqrt{\sigma^2/n} + \left(\mu - \mu_0\right) \Big/ \sqrt{\sigma^2/n}}{\sqrt{s_n^2/n} \Big/ \sqrt{\sigma^2/n}} \leq z\right) \\
&= 1 - \mathrm{P}_\mu\left(-z \leq \frac{\left(\overline{X}_n - \mu\right) \Big/ \sqrt{\sigma^2/n} + \left(\mu - \mu_0\right) \Big/ \sqrt{\sigma^2/n}}{\sqrt{s_n^2/\sigma^2}} \leq z\right) \\
&= 1 - \mathrm{P}\left(-z \leq W_{n-1} \leq z\right)
\end{aligned}$$

where we have defined

$$W_{n-1} = \frac{\left(\overline{X}_n - \mu\right) \Big/ \sqrt{\sigma^2/n} + \left(\mu - \mu_0\right) \Big/ \sqrt{\sigma^2/n}}{\sqrt{s_n^2/\sigma^2}}$$

Given the assumptions on the sample ξ_n we made above, the sample mean $\overline{X}_n$ has a normal distribution with mean μ and variance σ^2/n, so that the random variable

$$\left(\overline{X}_n - \mu\right) \Big/ \sqrt{\sigma^2/n}$$

has a standard normal distribution. Furthermore, the adjusted sample variance s_n^2 has a Gamma distribution with parameters $n-1$ and σ^2, so that the random variable

$$s_n^2/\sigma^2$$

has a Gamma distribution with parameters $n-1$ and 1. By adding a constant c to a standard normal distribution and dividing the sum thus obtained by the square root of a Gamma random variable with parameters $n-1$ and 1, one obtains a non-central standard Student's t distribution with $n-1$ degrees of freedom and non-centrality parameter c. Therefore, the random variable W_{n-1} has a non-central standard Student's t distribution with $n-1$ degrees of freedom and non-centrality parameter

$$(\mu - \mu_0) \Big/ \sqrt{\sigma^2/n}$$

■

Note that, for a fixed z, the test based on the unadjusted sample variance is more powerful than the test based on the adjusted sample variance:

$$\begin{aligned} \pi^u(\mu) &= 1 - \mathrm{P}\left(-\sqrt{\frac{n-1}{n}} \cdot z \leq W_{n-1} \leq \sqrt{\frac{n-1}{n}} \cdot z\right) \\ &> 1 - \mathrm{P}\left(-z \leq W_{n-1} \leq z\right) = \pi^a(\mu) \end{aligned}$$

because

$$\sqrt{\frac{n-1}{n}} < 1$$

and, as a consequence

$$\mathrm{P}\left(-\sqrt{\frac{n-1}{n}} \cdot z \leq W_{n-1} \leq \sqrt{\frac{n-1}{n}} \cdot z\right) < \mathrm{P}\left(-z \leq W_{n-1} \leq z\right)$$

78.2.7 The size of the test

The size of the test based on the unadjusted sample variance is equal to

$$\pi^u(\mu_0) = 1 - \mathrm{P}\left(-\sqrt{\frac{n-1}{n}} \cdot z \leq W_{n-1} \leq \sqrt{\frac{n-1}{n}} \cdot z\right)$$

where W_{n-1} is a standard Student's t distribution[12] with $n-1$ degrees of freedom.

Proof. When evaluated at the point $\mu = \mu_0$, the power function is equal to the size of the test, that is, the probability of committing a Type I error. The power function evaluated at μ_0 is

$$\pi^u(\mu_0) = 1 - \mathrm{P}\left(-\sqrt{\frac{n-1}{n}} \cdot z \leq W_{n-1} \leq \sqrt{\frac{n-1}{n}} \cdot z\right)$$

where W_{n-1} is a non-central standard Student's t distribution with $n-1$ degrees of freedom and non-centrality parameter equal to

$$\frac{\mu_0 - \mu_0}{\sqrt{\sigma^2/n}}$$

Therefore, when $\mu = \mu_0$, the non-centrality parameter is equal to 0 and W_{n-1} is just a standard Student's t distribution. ■

[12] See p. 407.

The size of the test based on the adjusted sample variance is equal to

$$\pi^a(\mu_0) = 1 - \mathrm{P}(-z \leq W_{n-1} \leq z)$$

where W_{n-1} is a standard Student's t distribution with $n-1$ degrees of freedom.

Proof. When evaluated at the point $\mu = \mu_0$, the power function is equal to the size of the test, that is, the probability of committing a Type I error. The power function evaluated at μ_0 is

$$\pi^u(\mu_0) = 1 - \mathrm{P}(-z \leq W_{n-1} \leq z)$$

where W_{n-1} is a non-central standard Student's t distribution with $n-1$ degrees of freedom and non-centrality parameter equal to

$$\frac{\mu_0 - \mu_0}{\sqrt{\sigma^2/n}}$$

Therefore, when $\mu = \mu_0$, the non-centrality parameter is equal to 0 and W_{n-1} is just a standard Student's t distribution. ■

Note that, for a fixed z, the test based on the unadjusted sample variance has a greater size than the test based on the adjusted sample variance, because, as demonstrated above, the former also has a greater power than the latter for any value of the true parameter μ.

78.3 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Denote by $F_n(x;k)$ the distribution function of a non-central standard Student's t distribution with n degrees of freedom and non-centrality parameter equal to k. Suppose a statistician observes 100 independent realizations of a normal random variable. The mean and the variance of the random variable, which the statistician does not know, are equal to 1 and 4 respectively. What is the probability, expressed in terms of $F_n(x;k)$, that the statistician will reject the null hypothesis that the mean is equal to zero if she runs a t-test based on the 100 observed realizations, setting $z = 2$ as the critical value, and using the adjusted sample variance to compute the t-statistic?

Solution

The probability of rejecting the null hypothesis $\mu_0 = 0$ is obtained by evaluating the power function of the test at $\mu = 1$:

$$\pi^a(\mu) = \pi^a(1) = \mathrm{P}_\mu(Z_n^a \notin [-z, z]) = 1 - \mathrm{P}(-2 \leq W_{99} \leq 2)$$

where the notation P_μ is used to indicate the fact that the probability of rejecting the null hypothesis is computed under the hypothesis that the true mean is equal to $\mu = 1$, and W_{99} is a non-central standard Student's t distribution with 99 degrees of freedom and non-centrality parameter

$$k = \frac{\mu - \mu_0}{\sqrt{\sigma^2/n}} = \frac{1-0}{\sqrt{4/100}} = \frac{10}{2} = 5$$

Thus, the probability of rejecting the null hypothesis is equal to

$$\begin{aligned}\pi^{a}(1) &= 1-\mathrm{P}(-2\leq W_{99}\leq 2)\\ &= 1-\left[F_{99}(2;5)-F_{99}(-2;5)\right]\\ &= 1+F_{99}(-2;5)-F_{99}(2;5)\end{aligned}$$

Exercise 2

Denote by $F_n(x)$ the distribution function of a standard Student's t distribution with n degrees of freedom, and by $F_n^{-1}(p)$ its inverse. Suppose that a statistician observes 100 independent realizations of a normal random variable, and she performs a t-test of the null hypothesis that the mean of the variable is equal to zero, based on the 100 observed realizations, and using the unadjusted sample variance to compute the t-statistic. What critical value should she use in order to incur into a Type I error with 10% probability? Express it in terms of $F_n(x)$.

Solution

A Type I error is committed when the null hypothesis is true, but it is rejected. The probability of rejecting the null hypothesis $\mu_0 = 0$ is

$$\begin{aligned}\pi^{u}(\mu_0) &= \pi^{u}(0) = 1-\mathrm{P}\left(-\sqrt{\frac{n-1}{n}}\cdot z\leq W_{n-1}\leq\sqrt{\frac{n-1}{n}}\cdot z\right)\\ &= 1-\mathrm{P}\left(-\sqrt{\frac{99}{100}}\cdot z\leq W_{99}\leq\sqrt{\frac{99}{100}}\cdot z\right)\end{aligned}$$

where z is the critical value, and W_{99} is a standard Student's t distribution with 99 degrees of freedom. This probability can be expressed as

$$\begin{aligned}&1-\mathrm{P}\left(-\sqrt{\frac{99}{100}}\cdot z\leq W_{99}\leq\sqrt{\frac{99}{100}}\cdot z\right)\\ =\;&1-\left[F_{99}\left(\sqrt{\frac{99}{100}}\cdot z\right)-F_{99}\left(-\sqrt{\frac{99}{100}}\cdot z\right)\right]\\ =\;&1-F_{99}\left(\sqrt{\frac{99}{100}}\cdot z\right)+F_{99}\left(-\sqrt{\frac{99}{100}}\cdot z\right)\\ \boxed{\text{A}}\;=\;&1-\left[1-F_{99}\left(-\sqrt{\frac{99}{100}}\cdot z\right)\right]+F_{99}\left(-\sqrt{\frac{99}{100}}\cdot z\right)\\ =\;&2\cdot F_{99}\left(-\sqrt{\frac{99}{100}}\cdot z\right)\end{aligned}$$

where: in step A we have used the fact that the density of a standard Student's t distribution is symmetric around zero. Thus, we need to set z in such a way that

$$2\cdot F_{99}\left(-\sqrt{\frac{99}{100}}\cdot z\right)=\frac{1}{10}$$

This is accomplished by

$$z = -\sqrt{\frac{100}{99}} F_{99}^{-1}\left(\frac{1}{20}\right)$$

Chapter 79

Hypothesis tests about the variance

This lecture presents some examples of hypothesis testing[1], focusing on **tests of hypothesis about the variance**, that is, on using a sample to perform tests of hypothesis about the variance of an unknown distribution.

79.1 Normal IID samples - Known mean

In this example we make the same assumptions we made in the example of set estimation of the variance entitled *Normal IID samples - Known mean* (p. 607). The reader is strongly advised to read that example before reading this one.

79.1.1 The sample

The sample ξ_n is made of n independent draws from a normal distribution having known mean μ and unknown variance σ^2. Specifically, we observe n realizations $x_1, \ldots, x_n$ of n independent random variables $X_1, \ldots, X_n$, all having a normal distribution with known mean μ and unknown variance σ^2. The sample is the n-dimensional vector

$$\xi_n = \begin{bmatrix} x_1 & \ldots & x_n \end{bmatrix}$$

which is a realization of the random vector

$$\Xi_n = \begin{bmatrix} X_1 & \ldots & X_n \end{bmatrix}$$

79.1.2 The null hypothesis

We test the null hypothesis[2] that the variance σ^2 is equal to a specific value $\sigma_0^2 > 0$:

$$H_0 : \sigma^2 = \sigma_0^2$$

[1] See p. 615.
[2] See p. 616.

79.1.3 The alternative hypothesis

We assume that the parameter space is the set of strictly positive real numbers, i.e., $\sigma^2 \in \mathbb{R}_{++}$. Therefore, the alternative hypothesis[3] is

$$H_1 : \sigma^2 > 0, \sigma^2 \neq \sigma_0^2$$

79.1.4 The test statistic

To construct a test statistic[4], we use the following point estimator of the variance

$$\widehat{\sigma_n^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

The test statistic is

$$\chi_n^2 = \frac{n}{\sigma_0^2}\widehat{\sigma_n^2}$$

This test statistic is often called **Chi-square statistic** (also written as **χ^2-statistic**) and a test of hypothesis based on this statistic is called **Chi-square test** (also written as **χ^2-test**).

79.1.5 The critical region

Let $z_1, z_2 \in \mathbb{R}_+$ and $z_1 < z_2$. We reject the null hypothesis H_0 if $\chi_n^2 < z_1$ or if $\chi_n^2 > z_2$. In other words, the critical region[5] is

$$C_{\chi_n^2} = [0, z_1) \cup (z_2, \infty)$$

Thus, the critical values of the test are z_1 and z_2.

79.1.6 The power function

The power function[6] of the test is

$$\pi\left(\sigma^2\right) = \mathrm{P}_{\sigma^2}\left(\chi_n^2 \notin [z_1, z_2]\right) = 1 - \mathrm{P}\left(\frac{\sigma_0^2}{\sigma^2}z_1 \leq \kappa_n \leq \frac{\sigma_0^2}{\sigma^2}z_2\right) \tag{79.1}$$

where κ_n is a Chi-square random variable[7] with n degrees of freedom and the notation P_{σ^2} is used to indicate the fact that the probability of rejecting the null hypothesis is computed under the hypothesis that the true variance is equal to σ^2.

Proof. The power function can be written as

$$\begin{aligned}\pi\left(\sigma^2\right) &= \mathrm{P}_{\sigma^2}\left(\chi_n^2 \notin [z_1, z_2]\right) \\ &= 1 - \mathrm{P}_{\sigma^2}\left(\chi_n^2 \in [z_1, z_2]\right) \\ &= 1 - \mathrm{P}_{\sigma^2}\left(z_1 \leq \frac{n}{\sigma_0^2}\widehat{\sigma_n^2} \leq z_2\right)\end{aligned}$$

[3] See p. 616.
[4] See p. 617.
[5] See p. 616.
[6] See p. 617.
[7] See p. 387.

$$= 1 - \mathrm{P}_{\sigma^2}\left(\frac{\sigma_0^2}{\sigma^2} z_1 \leq \frac{n}{\sigma^2}\widehat{\sigma_n^2} \leq \frac{\sigma_0^2}{\sigma^2} z_2\right)$$
$$= 1 - \mathrm{P}\left(\frac{\sigma_0^2}{\sigma^2} z_1 \leq \kappa_n \leq \frac{\sigma_0^2}{\sigma^2} z_2\right)$$

where we have defined

$$\kappa_n = \frac{n}{\sigma^2}\widehat{\sigma_n^2}$$

As demonstrated in the lecture entitled *Point estimation of the variance* (p. 579), the estimator $\widehat{\sigma_n^2}$ has a Gamma distribution[8] with parameters n and σ^2, given the assumptions on the sample ξ_n we made above. Multiplying a Gamma random variable with parameters n and σ^2 by n/σ^2 one obtains a Chi-square random variable with n degrees of freedom. Therefore, the variable κ_n has a Chi-square distribution with n degrees of freedom. ■

79.1.7 The size of the test

When evaluated at the point $\sigma^2 = \sigma_0^2$, the power function is equal to the probability of committing a Type I error, that is, the probability of rejecting the null hypothesis when the null hypothesis is true. This probability is called the size of the test[9] and it is equal to

$$\pi\left(\sigma_0^2\right) = \mathrm{P}_{\sigma_0^2}\left(\chi_n^2 \notin [z_1, z_2]\right) = 1 - \mathrm{P}\left(z_1 \leq \kappa_n \leq z_2\right)$$

where κ_n is a Chi-square random variable with n degrees of freedom.
Proof. This is obtained by substituting σ^2 with σ_0^2 in formula (79.1). ■

79.2 Normal IID samples - Unknown mean

This example is similar to the previous one. The only difference is that we now relax the assumption that the mean of the distribution is known.

79.2.1 The sample

In this example, the sample ξ_n is made of n independent draws from a normal distribution having unknown mean μ and unknown variance σ^2. Specifically, we observe n realizations $x_1, \ldots, x_n$ of n independent random variables $X_1, \ldots, X_n$, all having a normal distribution with unknown mean μ and unknown variance σ^2. The sample is the n-dimensional vector

$$\xi_n = [x_1 \quad \ldots \quad x_n]$$

which is a realization of the random vector

$$\Xi_n = [X_1 \quad \ldots \quad X_n]$$

79.2.2 The null hypothesis

We test the null hypothesis that the variance σ^2 is equal to a specific value $\sigma_0^2 > 0$:

$$H_0 : \sigma^2 = \sigma_0^2$$

[8] See p. 397.
[9] See p. 617.

79.2.3 The alternative hypothesis

We assume that the parameter space is the set of strictly positive real numbers, i.e., $\sigma^2 \in \mathbb{R}_{++}$. Therefore, the alternative hypothesis is

$$H_1 : \sigma^2 > 0, \sigma^2 \neq \sigma_0^2$$

79.2.4 The test statistic

We construct a test statistic, using the sample mean

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and either the unadjusted sample variance[10]

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$$

or the adjusted sample variance

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$$

The test statistic is

$$\chi_n^2 = \frac{n}{\sigma_0^2} S_n^2 = \frac{n-1}{\sigma_0^2} s_n^2$$

This test statistic is often called **Chi-square statistic** (also written as $\boldsymbol{\chi}^2$**-statistic**) and a test of hypothesis based on this statistic is called **Chi-square test** (also written as $\boldsymbol{\chi}^2$**-test**).

79.2.5 The critical region

Let $z_1, z_2 \in \mathbb{R}_+$ and $z_1 < z_2$. We reject the null hypothesis H_0 if $\chi_n^2 < z_1$ or if $\chi_n^2 > z_2$. In other words, the critical region is

$$C_{\chi_n^2} = [0, z_1) \cup (z_2, \infty)$$

Thus, the critical values of the test are z_1 and z_2.

79.2.6 The power function

The power function of the test is

$$\pi\left(\sigma^2\right) = \mathrm{P}_{\sigma^2}\left(\chi_n^2 \notin [z_1, z_2]\right) = 1 - \mathrm{P}\left(\frac{\sigma_0^2}{\sigma^2} z_1 \leq \kappa_n \leq \frac{\sigma_0^2}{\sigma^2} z_2\right) \tag{79.2}$$

where the notation P_{σ^2} is used to indicate the fact that the probability of rejecting the null hypothesis is computed under the hypothesis that the true variance is equal to σ^2 and κ_n has a Chi-square distribution with $n-1$ degrees of freedom.

[10] See p. 583.

Proof. The power function can be written as

$$\begin{aligned}\pi\left(\sigma^2\right) &= \mathrm{P}_{\sigma^2}\left(\chi_n^2 \notin [z_1, z_2]\right) \\ &= 1 - \mathrm{P}_{\sigma^2}\left(\chi_n^2 \in [z_1, z_2]\right) \\ &= 1 - \mathrm{P}_{\sigma^2}\left(z_1 \leq \frac{n}{\sigma_0^2} S_n^2 \leq z_2\right) \\ &= 1 - \mathrm{P}_{\sigma^2}\left(\frac{\sigma_0^2}{\sigma^2} z_1 \leq \frac{n}{\sigma^2} S_n^2 \leq \frac{\sigma_0^2}{\sigma^2} z_2\right) \\ &= 1 - \mathrm{P}\left(\frac{\sigma_0^2}{\sigma^2} z_1 \leq \kappa_n \leq \frac{\sigma_0^2}{\sigma^2} z_2\right)\end{aligned}$$

where we have defined

$$\kappa_n = \frac{n}{\sigma^2} S_n^2$$

Given the assumptions on the sample ξ_n we made above, the unadjusted sample variance S_n^2 has a Gamma distribution with parameters[11] $n-1$ and $\frac{n-1}{n}\sigma^2$, so that the random variable

$$\frac{n-1}{\frac{n-1}{n}\sigma^2} S_n^2 = \frac{n}{\sigma^2} S_n^2$$

has a Chi-square distribution with $n-1$ degrees of freedom. ■

79.2.7 The size of the test

The size of the test is equal to

$$\pi\left(\sigma_0^2\right) = \mathrm{P}_{\sigma_0^2}\left(\chi_n^2 \notin [z_1, z_2]\right) = 1 - \mathrm{P}\left(z_1 \leq \kappa_n \leq z_2\right)$$

where κ_n has a Chi-square distribution with $n-1$ degrees of freedom.
Proof. This is obtained by substituting σ^2 with σ_0^2 in formula (79.2). ■

79.3 Solved exercises

Below you can find some exercises with explained solutions.

Exercise 1

Denote by $F_n(x)$ the distribution function of a Chi-square random variable with n degrees of freedom. Suppose you observe 40 independent realizations of a normal random variable. What is the probability, expressed in terms of $F_n(x)$, that you will commit a Type I error if you run a Chi-square test of the null hypothesis that the variance is equal to 1, based on the 40 observed realizations, and choosing $z_1 = 0.8$ and $z_2 = 1.2$ as the critical values?

Solution

The probability of committing a Type I error is equal to the size of the test:

$$\pi\left(\sigma_0^2\right) = \pi(1) = 1 - \mathrm{P}\left(z_1 \leq \kappa_{40} \leq z_2\right)$$

[11] See p. 579.

where κ_{40} has a Chi-square distribution with 39 degrees of freedom. But

$$\begin{aligned} \mathrm{P}\left(z_1 \leq \kappa_{40} \leq z_2\right) &= F_{39}\left(z_2\right) - F_{39}\left(z_1\right) \\ &= F_{39}\left(1.2\right) - F_{39}\left(0.8\right) \end{aligned}$$

Thus

$$\pi\left(\sigma_0^2\right) = \pi\left(1\right) = 1 - \mathrm{P}\left(z_1 \leq \kappa_{40} \leq z_2\right) = 1 - F_{39}\left(1.2\right) + F_{39}\left(0.8\right)$$

Exercise 2

Make the same assumptions of the previous exercise and denote by $F_n^{-1}\left(p\right)$ the inverse of $F_n\left(x\right)$. Change the critical value z_1 in such a way that the size of the test becomes exactly equal to 5%.

Solution

Replace 0.8 with z_1 in the formula for the size of the test:

$$\pi\left(\sigma_0^2\right) = 1 - F_{39}\left(1.2\right) + F_{39}\left(z_1\right)$$

You need to set z_1 in such a way that $\pi\left(\sigma_0^2\right) = 0.05$. In other words, you need to solve

$$0.05 = 1 - F_{39}\left(1.2\right) + F_{39}\left(z_1\right)$$

which is equivalent to

$$F_{39}\left(z_1\right) = 0.05 - 1 + F_{39}\left(1.2\right)$$

Provided the right-hand side of the equation is positive, this is solved by

$$z_1 = F_{39}^{-1}\left(-0.95 + F_{39}\left(1.2\right)\right)$$