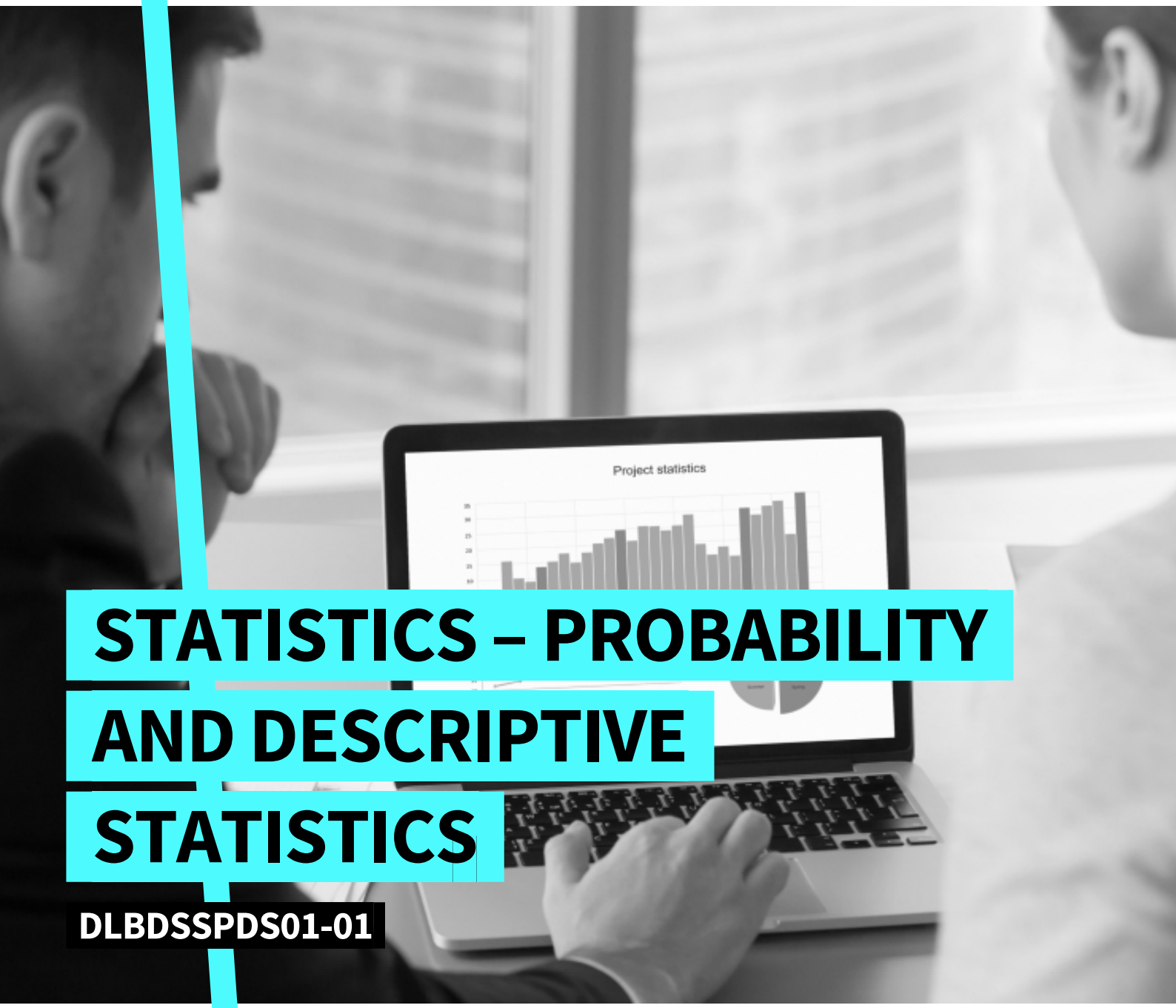
Course Book



STATISTICS – PROBABILITY AND DESCRIPTIVE STATISTICS

DLBDSSPDS01-01

iu
INTERNATIONAL
UNIVERSITY OF
APPLIED SCIENCES

STATISTICS – PROBABILITY AND DESCRIPTIVE STATISTICS

MASTHEAD

Publisher:
IU Internationale Hochschule GmbH
IU International University of Applied Sciences
Juri-Gagarin-Ring 152
D-99084 Erfurt

Mailing address:
Albert-Proeller-Straße 15-19
D-86675 Buchdorf
media@iu.org
www.iu.de

DLBDSSPDS01-01
Version No.: 001-2023-0306
N.N.

© 2023 IU Internationale Hochschule GmbH
This course book is protected by copyright. All rights reserved.
This course book may not be reproduced and/or electronically edited, duplicated, or distributed in any kind of form without written permission by the IU Internationale Hochschule GmbH.
The authors/publishers have identified the authors and sources of all graphics to the best of their abilities. However, if any erroneous information has been provided, please notify us accordingly.

MODULE DIRECTOR

PETER ANKLAM

Mr. Anklam is an academic teacher of business administration and marketing with a focus on fashion management at IU International University of Applied Sciences.

Parallel to his business studies at the University of Nuremberg, he completed several semesters of business education and wrote his diploma thesis on controlling in the textile trade. Guest lectureships at the Academy for Fashion Management LDT Nagold, DHBW Heilbronn, DHBW Stuttgart, and Nürtingen University of Applied Sciences followed.

He is a management consultant and data protection officer for various companies in fashion, textile, footwear, and sports sectors as well as a partner at MAX Fashion GbR. He was also chairman of city marketing in Nagold.

His research focuses on controlling and market research in retail as well as location strategies for city centers.

TABLE OF CONTENTS

STATISTICS – PROBABILITY AND DESCRIPTIVE STATISTICS

INTRODUCTION

WELCOME

SIGNPOSTS THROUGHOUT THE COURSE BOOK

This course book contains the core content for this course. Additional learning materials can be found on the learning platform, but this course book should form the basis for your learning.

The content of this course book is divided into units, which are divided further into sections. Each section contains only one new key concept to allow you to quickly and efficiently add new learning material to your existing knowledge.

At the end of each section of the digital course book, you will find self-check questions. These questions are designed to help you check whether you have understood the concepts in each section.

For all modules with a final exam, you must complete the knowledge tests on the learning platform. You will pass the knowledge test for each unit when you answer at least 80% of the questions correctly.

When you have passed the knowledge tests for all the units, the course is considered finished and you will be able to register for the final assessment. Please ensure that you complete the evaluation prior to registering for the assessment.

Good luck!

BASIC READING

Downey, A.B. (2014). *Think stats* (2nd ed.). O'Reilly. http://search.ebscohost.com.pxz.iubh.de:8080/login.aspx?direct=true&db=cat05114a&AN=ihb.28838&site=eds-live&scope=site

Kim, A. (2019). Exponential Distribution - Intuition, Derivation, and Applications. (Available online)

Rohatgi, V. K., & Saleh, A. K. E. (2015). *An introduction to probability and statistics*. John Wiley & Sons, Incorporated. http://search.ebscohost.com.pxz.iubh.de:8080/login.aspx?direct=true&db=cat05114a&AN=ihb.45506&site=eds-live&scope=site

Wagaman, A.S & Dobrow, R.P. (2021). *Probability: With applications and R*. Wiley. http://search.ebscohost.com.pxz.iubh.de:8080/login.aspx?direct=true&db=edsebk&AN=2947734&site=eds-live&scope=site

Triola , M.F. (2013). *Elementary statistics*. Pearson Education. http://search.ebscohost.com.pxz.iubh.de:8080/login.aspx?direct=true&db=cat05114a&AN=ihb.45501&site=eds-live&scope=site

FURTHER READING

UNIT 1

Downey, A.B. (2014). *Think stats* (2nd ed.). O'Reilly. http://search.ebscohost.com.pxz.iubh.de:8080/login.aspx?direct=true&db=cat05114a&AN=ihb.28838&site=eds-live&scope=site

Wagaman, A.S & Dobrow, R.P. (2021). *Probability: With applications and R*. Wiley http://search.ebscohost.com.pxz.iubh.de:8080/login.aspx?direct=true&db=edsebk&AN=2947734&site=eds-live&scope=site

Rohatgi, V. K., & Saleh, A. K. E. (2015). *An introduction to probability and statistics*. John Wiley & Sons, Incorporated. (Chapter 1). http://search.ebscohost.com.pxz.iubh.de:8080/login.aspx?direct=true&db=cat05114a&AN=ihb.45506&site=eds-live&scope=site

UNIT 2

Downey, A.B. (2014). *Think Bayes*. Sebastopol, CA: O'Reilly. (Chapters 3, 4, 5, and 6) http://search.ebscohost.com.pxz.iubh.de:8080/login.aspx?direct=true&db=cat05114a&AN=ihb.28839&site=eds-live&scope=site

Wagaman, A.S & Dobrow, R.P. (2021). *Probability: With applications and R*. Wiley (Chapter 3,6-8). http://search.ebscohost.com.pxz.iubh.de:8080/login.aspx?direct=true&db=edsebk&AN=2947734&site=eds-live&scope=site

Rohatgi, V. K., & Saleh, A. K. E. (2015). *An introduction to probability and statistics*. John Wiley & Sons, Incorporated. (Chapter 5). http://search.ebscohost.com.pxz.iubh.de:8080/login.aspx?direct=true&db=cat05114a&AN=ihb.45506&site=eds-live&scope=site

UNIT 3

Downey, A.B. (2014). *Think Bayes*. Sebastopol, CA: O'Reilly. (Chapter 7) http://search.ebscohost.com.pxz.iubh.de:8080/login.aspx?direct=true&db=cat05114a&AN=ihb.28839&site=eds-live&scope=site

Rohatgi, V. K., & Saleh, A. K. E. (2015). *An introduction to probability and statistics*. John Wiley & Sons, Incorporated. (Chapter 4). http://search.ebscohost.com.pxz.iubh.de:8080/login.aspx?direct=true&db=cat05114a&AN=ihb.45506&site=eds-live&scope=site

UNIT 4

Wagaman, A.S. & Dobrow, R.P. (2021). *Probability: With applications and R*. Wiley (Chapter 9). http://search.ebscohost.com.pxz.iubh.de:8080/login.aspx?direct=true&db=edsebk&AN=2947734&site=eds-live&scope=site

Triola, M. F. (2013). *Elementary statistics*. Pearson Education. (Chapter 11). http://search.ebscohost.com.pxz.iubh.de:8080/login.aspx?direct=true&db=cat05114a&AN=ihb.45501&site=eds-live&scope=site

Rohatgi, V. K., & Saleh, A. K. E. (2015). *An introduction to probability and statistics*. John Wiley & Sons, Incorporated. (Chapter 7). http://search.ebscohost.com.pxz.iubh.de:8080/login.aspx?direct=true&db=cat05114a&AN=ihb.45506&site=eds-live&scope=site

UNIT 5

Amy S. Wagaman, & Robert P. Dobrow. (2021). *Probability: With applications and R*. Wiley (Chapter 10). http://search.ebscohost.com.pxz.iubh.de:8080/login.aspx?direct=true&db=edsebk&AN=2947734&site=eds-live&scope=site

Triola, M. F. (2013). *Elementary statistics*. Pearson Education. (Chapter 6). http://search.ebscohost.com.pxz.iubh.de:8080/login.aspx?direct=true&db=cat05114a&AN=ihb.45501&site=eds-live&scope=site

LEARNING OBJECTIVES

Welcome to **Statistics - Probability and Descriptive Statistics**! This course will provide you with a foundation in mathematical probability, preparing you for further courses in statistical inference and data science. The statistical tools that you will learn in this course will enable you to review, analyze, and draw conclusions from data.

You will learn the key terms and concepts that are at the core of probability theory, including random experiments, sample spaces, events, and the axioms of probability. You will learn to classify events as mutually exclusive and independent, and how to compute the probability of unions and joint events. You will also learn how to interpret and use conditional probability and apply Bayes' theorem to selected applications.

A random variable is a numerical description of the outcome of a statistical experiment. As a mathematical formalization it quantifies random events. When studying a given data set, we generally consider the data points as an observation of a random occurrence, which can be described by the underlying distribution of a random variable.

You will learn to define a random variable and express and interpret its distribution using Probability Mass Functions (PMFs), Probability Density Functions (PDFs), and Cumulative Distribution Functions (CDFs). You will learn about important probability distributions, their characteristics, and how they are used to model real-world experiments.

Sometimes data comes in the form of pairs of triples or random variables. The variables in these tuples may be independent or dependent. You will learn how to express the randomness of these tuples using joint distributions, PMFs and PDFs. Marginal and conditional distributions play a key role in isolating the distribution of one variable from the tuple in different ways. You will be provided with examples that will help you to learn how to compute and interpret such distributions.

The average and standard deviation are the most popular summaries we can compute from numerical data. These ideas are extended using general notions of the expected value of a random variable as well as other expectation quantities. You will learn how to compute means, variances, general moments, and central moments. More importantly, you will be able to describe certain characteristics of distributions, such as skewness and kurtosis, using these quantities.

Finally, you will be introduced to important inequalities and limit theorems. These inequalities and theorems are at the very foundation of the methods of statistical inference, providing a sound framework for drawing conclusions about scientific truths from data. Furthermore, they will be used to define and evaluate performance metrics of learning algorithms in your further studies.

NOTE

Given the main focus of this course (on fundamental theories and applications of statistics), it would be preferable for students to have some prior knowledge of basic topics of mathematical analysis, i.e., integral and differential calculus, as well as properties of functions. However, for the sake of completeness, the tools of analysis that are most important for this course will be briefly introduced and discussed at relevant points throughout the course book.

UNIT 1

PROBABILITY

STUDY GOALS

On completion of this unit, you will be able to ...

- understand the key terms outcome, event, and sample space and how these terms are used to define and compute probabilities.
- identify the three fundamental axioms of probability measures.
- compute and interpret probabilities involving mutually exclusive events.
- compute and interpret probabilities of two independent events and conditional probabilities.
- compute probabilities of two events that are not necessarily independent.
- compute probabilities of two events that are not necessarily mutually exclusive.
- understand the concept of partitioning a sample space and how it frames the statement of the total law of probability.
- apply Bayes' theorem to real-world examples.

1. PROBABILITY

Introduction

Probability is the primary tool we use when we are dealing with random experiments-that is, experiments where the outcome cannot be determined with complete certainty (see Wackerly, Mendenhall & Schaeffer, 2008; Wasserman, 2004). Consider rolling a pair of fair 6-sided dice. The outcome of any such roll cannot be determined with absolute certainty. How many possible outcomes are there? Is a sum of five or eight more likely? What is the most likely sum? What is the least likely sum? The tools we discuss in this unit will help address these and other questions.

Perhaps you have heard of the phrase "lucky number 7". The origin of this statement lies in the fact that when a pair of fair dice are rolled, seven is the most likely sum. On completion of this unit, you will be able to quantify this fact. Furthermore, you will be able to develop the relevant concepts much further in order to answer more complex questions.

1.1 Definitions

Although we cannot predict the outcome of random experiments with absolute certainty, we can write down all the possible outcomes the experiment could have. For the coin toss random experiment, the possible outcomes, also called elements (see Klenke, 2014), are H (heads) or T (tails). The set containing all the possible outcomes is called the **sample space** of the experiment. We say that an outcome a is an element of Ω and write

Sample space
This is a set containing all possible outcomes of a random experiment. It is usually denoted by Ω or S.

$$a \in \Omega\,.$$

Now consider the experiment of tossing two coins. One possible outcome could be to observe heads on the first coin and tails on the second coin. We can summarize this outcome as HT. Using this notation, the sample space can thus be written as

$$\Omega = \{HH, HT, TH, TT\}$$

Outcome
This is a single result from a trial of a random experiment.

Each element is an **outcome** of the random experiment. In general, we can denote the outcome of an experiment by ω_i, where $i \in \mathbb{N}$ is just the index of the outcome. In this notation, the sample space can be denoted as $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ for $\mathrm{n} \in \mathbb{N}$ and for a finite sample space and $\Omega = \{\omega_1, \omega_2, \ldots\}$ for a countably infinite sample space.

In some applications we are interested in a single outcome and want to calculate the probability of that outcome, but sometimes we are interested in a group of outcomes. Therefore, the next term we will define is an event. An **event** of a random experiment is a set of outcomes. The following notation is used to denote an event A, which is contained in Ω,

$$A \subseteq \Omega\,.$$

An event is also called a set (see Klenke, 2014). The following notation

$$A \subset \Omega,$$

means that the event A is contained in Ω and at least one outcome exists which is not contained in A, but in Ω.

Event
This is a collection of zero or more outcomes of a random experiment. Events are usually denoted using capital letters: *A*, *B*, *C*, …

For the two-coin toss experiment, perhaps we are interested in the outcomes where the result for the two coins match. In this case, we are talking about the event $A = \{HH, TT\}$. Note that the order of the elements in a set does not matter, so $\{HH, TT\} = \{TT, HH\}$. Finally, we can have an event that contains a single outcome: $B = \{HT\}$.

Finally, we will introduce two fundamental operations for any events in the sample space Ω. For two events $A, B \subseteq \Omega$, the **union** of A and B, which is denoted by

$$A \cup B = \{x \in \Omega \,|\, x \in A \text{ or } x \in B\}$$

Union
The union of the events A and B is also an event containing all outcomes of A and all outcomes of B.

is the event of all outcomes contained in A or in B.

In addition, the **intersection** of A and B, denoted by

$$A \cap B = \{x \in \Omega \,|\, x \in A \text{ and } x \in B\},$$

Intersection
The intersection of the events A and B is also an n event containing all outcomes of A, which are also contained in B. We can also say the intersection of A and B is the event, which contains all outcomes of B, which are also contained in A.

is the event in which all outcomes are common to both A and B.

Figure 1: A Venn Diagram of Three Events

A and B are events of a given sample space.
On the left-hand side you can see the intersection of A and B.
The diagram on the right-hand side depicts the union of A and B.

Source: George Dekermenjian (2019).

Special Events

There are two special events that require a mention here. At one extreme, an event may contain nothing, in which case we have the null event or the empty set: $\emptyset = \{\}$. At the other extreme, we have the whole sample space itself which, of course, contains all the possible outcomes.

Axioms of Probability

Now that we have an understanding of the fundamental terms, we are ready to talk about probability. The probability of an event measures the likelihood of observing that an event when, for example, a random experiment is performed. For a given sample space Ω every **probability measure** P, which maps an event of Ω to a real number, has to satisfy the following three axioms of probability:

Probability measure
This is used to assign probabilities to events of a given sample space.

1. $P(\emptyset) = 0$,
2. For any event $A \subseteq \Omega$ it holds that $P(A) \geq 0$,
3. For mutually exclusive events A_1,A_2,A_3,... $\subseteq \Omega$,

   $$P(A_1 \cup A_2 \cup A_3 \cup \quad ...) = \sum_{i=1}^{\infty} P(A_i),$$

4. $P(\Omega) = 1$.

Mutually exclusive
Two events are called mutually exclusive if their intersection yields an empty set.

Two events (sets) are **mutually exclusive** if they have no common outcomes (elements).

For non-mutually exclusive events we can deduce, according to the axioms of probability, that

$$P(A \ \cap \ B) + P(A \ \cup \ B) = P(A) + P(B) \text{ for any events } A,\, B \subseteq \Omega\,.$$

Example 1.1

Consider the random experiment of tossing two coins. We will assume that the probability of each outcome is equally likely so that singleton events (events with only one outcome) have equal probability. Since there are four outcomes, the probability of each singleton event is $\frac{1}{4}$.

$$P(\{HH)\} = P(\{HT)\} = P(\{TH)\} = P(\{TT)\} = 14$$

In practice, if an event contains one element, we can just write $P(HT) = \frac{1}{4}$, excluding the brackets.

Classical Probability

There are two approaches to defining probability: the classical (frequentist) approach and the Bayesian approach. We will discuss the classical approach first and then move on to a discussion of the Bayesian approach.

Consider a random experiment with $n \ \in \mathbb{N}$ equally likely outcomes. In other words, the sample space contains n outcomes

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

The probability of an event $A = \left\{\omega_{i_1}, \omega_{i_2}, ..., \omega_{i_m}\right\}$ for $m, \mathrm{i}_m \in \mathbb{N}$ of this experiment is the ratio of the number of outcomes in A to the size of the sample space. We will denote the number of outcomes in A by $|A|$ so that $|A| = m$.

$$P(A) = \frac{|A|}{|\Omega|} = \frac{m}{n}\,.$$

Suppose a bag contains seven red marbles denoted by $r_1, r_2, \ ..., r_7$ and three blue marbles denoted by $b_1, b_2, \ ..., b_3$. We will draw one marble out of this bag at random. The sample space for this experiment is $\Omega = \left\{r_1, r_2, r_3, r_4, r_5, r_6, r_7, b_1, b_2, b_3\right\}$. We are interested in computing the probability that the marble drawn is blue. The event corresponding to drawing a blue marble is $A = \left\{b_1, b_2, b_3\right\}$. The event contains $|A| = 3$ outcomes and the sample space contains $|\Omega| = 10$ outcomes. Therefore, the probability of drawing a blue marble is $P(A) = \frac{|A|}{|\Omega|} = \frac{3}{10}$.

Let us now verify that this formulation is a valid probability measure. In other words, we need to verify that the axioms of probability are satisfied.

1. $P(\emptyset) = \frac{|\emptyset|}{|\Omega|} = \frac{0}{n} = 0$ and $P(\Omega) = \frac{|\Omega|}{|\Omega|} = \frac{n}{n} = 1$.
2. If A is an event, then $0 \leq |A| \leq n$. Dividing by $|\Omega| = n$ gives $0 \leq \frac{|A|}{|\Omega|} \leq 1$. In other words, we have $0 \leq P(A) \leq 1$ as required.
3. Now suppose that A and B are mutually exclusive events. Then the number of elements in the event A or B is the union $A \ \cup \ B$. Since they are mutually exclusive, it must hold that

   $$|A \ \cup \ B| = |A| + |B|,$$

   because a marble cannot be in A and B simultaneously. Dividing by $|\Omega|$ we obtain $\frac{|A \ \cup \ B|}{|\Omega|} = \frac{|A|}{|\Omega|} + \frac{|B|}{|\Omega|}$. In other words, it holds that

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

   as required.

We do not have to deal with the case of infinitely mutually exclusive events since our sample space is finite i.e., it consists of 10 marbles. That means if we assume mutually exclusive events such that $\mathrm{A}_1, \mathrm{A}_2, \mathrm{A}_3, ... \subseteq \Omega$ then only finite events can contain at least one marble. The rest of the sets must be empty sets. Thus, we reduced the problem to finite mutually disjoint events, which can be discussed in the same way as in the case of two mutually disjoint events.

Since the classical definition of probability satisfies all probability axioms, it is a valid probability measure.

Example 1.2

Consider the random experiment of tossing three coins. Find the probability of observing at least one .

Solution 1.2

Recall that the sample space is

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

The event of observing at least one H is exactly the event $A = \{TTH, THH, HTH, THT, HTT, HHT, HHH\}$. This event contains $|A| = 7$ outcomes. Furthermore, the sample space contains $|\Omega| = 8$ outcomes. Therefore, the probability of observing at least one H is

$$P(\text{at least on } H) = P(A) = \frac{|A|}{|\Omega|} = \frac{7}{8} = 0{,}875\,.$$

Example 1.3

Consider the experiment of rolling a 6-sided die.

a) Write down the sample space.
b) Write down the event of observing an even number.
c) Calculate the probability of observing an even number.

Solution 1.3

a) $\Omega = \{1,2,3,4,5,6\}$.
b) $A = \{2,4,6\}$.
c) $P(A) = \frac{|A|}{|\Omega|} = \frac{3}{6} = \frac{1}{2} = 0{,}5 = 50\ \%$.

Consider the experiment of rolling a pair of 6-sided dice. For each die, we can observe a number from 1 to 6. If we paired the observations from each die, we would have a single observation from the pair. For example, if the first die lands on 2 and the second lands on 5, we can write down this outcome as (2,5). The sample space S of this experiment is shown in the table below.

Table 1: Sample Space of Rolling a Pair of 6-Sided Dice

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)

(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Source: George Dekermenjian (2019).

The sample space consists of $|\Omega| = 36$ outcomes. Using this information, let us explore some questions related to this experiment.

Example 1.4

Using the information provided in the table above:

a) Write down the event of observing the same number on both dice.
b) Write down the event of observing numbers that sum to 4.
c) Calculate the probability of each of these events.

Solution 1.4

a) $A = \{(1,1),\ (2,2),\ (3,3),\ (4,4),\ (5,5),\ (6,6)\}$.
b) $B = \{(1,3),\ (2,2),\ (3,1)\}$.
c) $P(A) = \frac{|A|}{|\Omega|} = \frac{6}{36} = \frac{1}{6}$.
d) $P(B) = \frac{|B|}{|\Omega|} = \frac{3}{36} = \frac{1}{12}$.

Do we have to write down the outcomes? The formula for probability that we are using only makes use of the number of outcomes. As you can imagine, for more complex experiments, the size of the sample space can become very large, and it would not be wise to write down all the possible outcomes. However, to compute the probability we still need to be able to count the number of outcomes, whether it is in the sample space or for another event. To this end, we will take a short departure from our main topic and review some basic counting techniques that will be useful in answering probability questions.

Counting

All of the formulas we will discuss here are based on one simple principle: the multiplication principle of counting. If there are N_1 ways of performing task 1 and N_2 ways of performing task 2, then there are $N_1 \cdot N_2$ ways of performing both tasks. This principle is easily extended to more than two tasks.

Suppose a pizza parlor offers its patrons the option of customizing their own pizzas. They are offered three types of crusts, two types of sauces, and can also choose one from a selection of five toppings. To count the number of different pizzas one can order at this pizza parlor, we can break down making a pizza into three tasks: (i) there are $N_1 = 3$ ways of choosing a crust, (ii) there are $N_2 = 2$ ways of choosing a sauce, and (iii) $N_3 = 5$ ways of choosing a topping. Therefore, there are $N_1 \cdot N_2 \cdot N_3 = 3 \cdot 2 \cdot 5 = 30$ ways of making a pizza.

Permutations

Suppose you have four different books, and you want to arrange them on a shelf. We want to count the total number of arrangements possible. There are four tasks. At first there are four books to set in place. After placing the first book, there are three books to set in place, then two books, and, finally, after placing these, there is one book left to place on the shelf. Therefore, using the multiplication principle, there are $4 \cdot 3 \cdot 2 \cdot 1 = 24$ ways of arranging the four books on the shelf. This is an example of a permutation. Using factorial notation, we can write this computation as $4! = 4 \cdot 3 \cdot 2 \cdot 1$. In general, if there are $n \in \mathbb{N}$ different objects to arrange, there are

$$n! = n(n-1)(n-2) \cdot ... \cdot 3 \cdot 2 \cdot 1$$

permutations (arrangements) possible.

Now suppose that we have $n = 10$ objects, but we want to select and arrange $k = 3$ of them. We can choose the first objects (10 choices), then the second objects (9 choices), and, finally, the third object (8 choices). Therefore, the total number of arrangements is $10 \cdot 9 \cdot 8 = 720$. In general, if we have n distinct objects, the number of permutations of k of these objects is

$$\frac{n!}{(n-k)!}$$

Combinations

Suppose there are 10 people at a dinner party and each person shakes the hands of every other person. We want to work out how many handshakes there would be. Using the multiplication rule, we can argue that observing the event of a handshake involves two tasks: (i) the first person in the handshake (10 people available) and (ii) the second person in the handshake (9 people available). So far, we have $10 \cdot 9$. However, the order of the people does not matter. If John shakes hands with Mary and Mary shakes hands with John, the handshake is the same. Therefore, we do not need to count these handshakes twice. We divide the expression to get $\frac{10 \cdot 9}{2} = 45$ handshakes.

This is an example of a combination, which is similar to a permutation, but order does not matter. In general, if we have $n \in \mathbb{N}$ distinct objects, the number of ways of choosing $k \in \mathbb{N}$ of them is given by

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

The expression $\binom{n}{k}$ is read as “n choose k”. For the handshake example, this formula indeed gives the correct answer:

$$\binom{10}{2} = \frac{10!}{(10-2)!2!} = \frac{10!}{8!2!} = 45.$$

Now that we have some efficient tools for counting, we are equipped to tackle some more probability questions. Below is one such example.

Example 1.5

Suppose there are five women and four men. We will randomly choose three people.

a) Calculate the size of the sample space for this experiment.
b) How many ways are there of choosing two women and one man?
c) What is the probability of choosing two women and one man?

Solution 1.5

a) The sample space consists of all possible groups of three people from nine different people. The order does not matter here. The number of ways is "9 choose 3" or

$$|\Omega| = \binom{9}{3} = \frac{9!}{(9-3)!3!} = \frac{9!}{6!3!} = \frac{9 \cdot 8 \cdot 7}{3 \cdot 2 \cdot 1} = 84$$

b) Choosing two women and one man is actually two tasks. We will count the number of ways of performing each task and then multiply them (using the multiplication principle).

Task 1: Choosing two women from five. There are $\binom{5}{2} = 10$ ways.

Task 2: Choosing one man from four. There are $\binom{4}{1} = 4$ ways.

According to the multiplication rule, there are $10 \cdot 4 = 40$ ways of choosing two women and one man.

c) Let us call the event of choosing two women and one man A. We found that $|A| = 40$. Therefore, the probability of this event is

$$P(A) = \frac{40}{84} \approx 0.4762 = 47.62\ \%\,.$$

Complementary Events

The complement of an event, just like the complement of a set, is the event of not observing the included outcomes. For example, in a dice roll experiment we have the sample space $\Omega = \{1,2,3,4,5,6\}$. If A is the event of observation 1 or 2 $(A = \{1,2\})$, then the complement of A is $A^c = \{3,4,5,6\}$.

The probability of A is $P(A) = \frac{2}{6}$, and the probability of its complement is $P(A^c) = \frac{4}{6}$. Indeed, we have $P(A) + P(A^c) = \frac{2}{6} + \frac{4}{6} = 1$.

This means that for a given sample Ω it holds that

$$P(A) + P(A^c) = 1 \text{ for any } \quad A \subseteq \Omega\,.$$

1.2 Independent Events

Consider the experiment of tossing a fair coin and then rolling a fair 6-sided die. The probability of observing the joint event$(H, 2)$ is $\frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12}$. That is, we multiply the probabilities. This is because the tossing of a fair coin does not influence the result of rolling a die. The two events are independent. More formally, for a given sample space Ω, two events, $A \subseteq \Omega$ and $B \subseteq \Omega$, are said to be **independent** if

Independence of two events
Two events are independent if the probability of their intersection yields the same as the product of each probability.

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

Example 1.6

Suppose we draw two cards at random with replacement from a standard deck of 52 cards. That is, we draw the first card, place it back in the deck, and then draw another card. What is the probability that both cards are spades?

Solution 1.6

The event of the first card being a spade is independent from the second card being a spade. Therefore, the probability of both being spades is

$$\frac{13}{52} \cdot \frac{13}{52} = \frac{1}{16} = 0{,}0625 = 6.25\ \%\,.$$

Suppose the two events A and B with $P(A),\ P(B) > 0$ are disjoint (mutually exclusive). Can they be independent? If the two events are mutually exclusive, then they cannot both occur at the same time, so the probability of the joint event is $P(A \cap B) = 0$. Therefore, the two events are not mutually exclusive, since

$$0 = P(A \cap B) \neq P(A) \cdot P(B) > 0\,.$$

Example 1.7

Suppose a fair coin is tossed five times. What is the probability of observing at least one tail?

Solution 1.7

Note that each of the tosses is independent. Furthermore, it is easier to work with the complement of this event. Let A be the event of observing at least one tail. Then, the complement event A^c is the event of observing no tails (that is, observing heads on each of the five tosses). Let H_i denote the event of observing heads on the i^{th}toss for $i \in \mathbb{N}$, and then use the formula for the probability of complements. We then have

$$\begin{aligned} P(A) &= 1 - P(A^c) \\ &= 1 - P(H_1 H_2 H_3 H_4 H_5) \\ &= 1 - P(H_1)P(H_2)P(H_3)P(H_4)P(H_5) \\ &= 1 - \left(\frac{1}{2}\right)^5 = \frac{31}{32} \approx 0.9688 = 96.88\ \%. \end{aligned}$$

We have used independence in the third equality. In the fourth equality, we used the fact that the probability of observing heads in any toss is $\frac{1}{2}$.

Example 1.8

A bag contains three red marbles and five blue marbles. Two marbles are drawn, one after the other, without replacement. Is the event of observing a red marble on the first draw and a blue marble on the second draw independent? Why or why not?

Solution 1.8

The two events are not independent. The result of the first event will change the number of available marbles in the bag, since there is one marble missing for the second draw.

We will see how to calculate the probability of joint events that are dependent in the following section.

1.3 Conditional Probability

Conditional probability is a way of calculating the probability of an event using prior information. The notation $P(A \,|\, B)$ is read as the "probability of A given that we have already observed B". In other words, while $P(A)$ is the (unconditional) probability of observing A, $P(A \,|\, B)$ is the conditional probability of A conditioned on B. Suppose we have three red marbles and five blue marbles in a bag. We draw two marbles at random without replacement. Let A denote the event of observing a red marble. Let B denote the event of observing a blue marble. The probability of B given that we have already observed A is written as $P(B \,|\, A)$. After observing A, a red marble, there are only seven marbles left in the bag: two red and five blue. Therefore, the probability $P(B \,|\, A) = \frac{5}{7}$. In contrast, $P(B) = \frac{5}{8}$. Thus for a given sample space Ω, we say that the conditional probability of $A \subseteq \Omega$ conditioned on $B \subseteq \Omega$ with $P(B) > 0$, is defined by

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

Example 1.9

Suppose that the probability of a randomly chosen person having cancer is 1%, and that if a person has cancer, a medical test will yield a positive result with a probability of 98%. What is the probability that the person has cancer and the medical test result shown is positive?

Solution 1.9

Let A denote the event that a person has cancer. We know that $P(A) = 0.01$. Now let B denote the event that the medical test yields a positive result. We want to find $P(A \cap B)$. We know the conditional probability $P(B\,|\,A) = 0.98$. Using the formula for conditional probability we have

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

Rewriting this formula, we have

$$P(A \cap B) = P(B\,|\,A) \cdot P(A) = 0.98 \cdot 0.01 = 0.0098 = 0.98\ \%\,.$$

If the two events A and B are independent, then observing one of the events gives us no information about the other event. In other words, $P(A\,|\,B) = P(A)$. Indeed, we can show this using the result of independent events and the formula for conditional probability as follows:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \cdot P(B)}{P(B)} = P(A) \quad \text{for} \quad P(B) > 0\,.$$

Let us revisit the experiment of drawing two cards.

Example 1.10

Suppose two cards are drawn out of a deck of 52 cards, one after the other, without replacement. What is the probability that both are spades?

Solution 1.10

Let S_1 denote the event that the first card is a spade and let S_2 denote the event that the second card is a spade. Since these two events are dependent, think about why this is the case. We can use the conditional probability formula in the form

$$P(S_1 \cap S_2) = P(S_2 \mid S_1) \cdot P(S_1)\,.$$

The left-hand side is the probability that we observe a spade on both draws. The first factor on the right denotes the probability that the second card is a spade given that the first card was a spade. The last factor is the probability that the first card is a spade. Since there are 13 spades out of 52 cards, we have $P(S_1) = \frac{13}{52}$. After having observed a spade, there are only 12 spades left in the deck of a total of 51 cards.

Therefore, $P(S_1 \,|\, S_2) = \frac{12}{51}$.

Therefore,

$$P(S_1 \cap S_2) = \tfrac{12}{51} \cdot \tfrac{13}{52} = \tfrac{1}{17} \approx 0.0588 = 5.88\ \%\,.$$

Compare this answer with Example 1.6. Does the result surprise you?

1.4 Bayesian Statistics

In contrast to classical statistics, Bayesian statistics is all about modifying conditional probabilities – it uses prior distributions for unknown quantities which it then updates to posterior distributions using the laws of probability.

Let us revisit the medical cancer test in Example 1.9. Let us say a randomly chosen person tests positive for cancer. What is the probability that they actually have cancer? Biomedical tests are never perfect; there is typically a small false positive and false negative rate. In the setting of the example, recall that A represents the event that a person has cancer and B represents the event that the medical test returns a positive result. We were given the prevalence of the disease in the general population, which was 1%, so that $P(A) = 0.01$. The test is 98% accurate for people who actually have the disease—that is, $P(B|A) = 0.98$. Finally, suppose the test gives a false positive 20% of the time. We are now interested in finding out $P(A \,|\, B)$. This is the subject of Bayes' theorem. Before discussing Bayes' theorem, let us first write down a preliminary result.

To motivate the result, suppose that we **partition** the sample space Ω into disjoint events A_1, A_2, and A_3. That is, these events are mutually exclusive, and together they contain all the outcomes in the sample space, meaning

Partition of an event
Let A be an event. When the union of two or more mutually exclusive events is A, the group of events is called a partition of A.

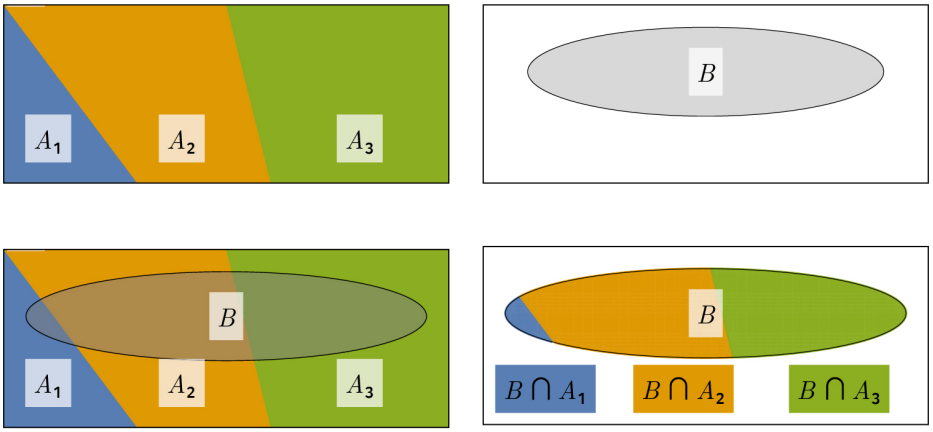
$$A_1 \cup A_2 \cup A_3 = \Omega.$$

Now consider another event . Then it holds

$$B = (A_1 \cap\ B) \cup (A_2 \cap\ B) \cup\ (A_3 \cap\ B)$$

meaning that the events $A_1 \cap B$, $A_2 \cap B$, and $A_3 \cap B$ partition the event B. In other words, these events are mutually disjoint and together they contain all of B. See the figure below for an illustration.

Figure 2: Partitions



Top Left: A partition of the sample space with three sets
Top Right: The event B in the sample space
Bottom Left: The event B and the partition of the sample space
Bottom Right: The induced partition of the event B

Source: George Dekermenjian (2019).

Theorem: The Law of Total Probability

Let A_1, A_2, A_3, … be a countably infinite collection that partitions the sample space Ω. In other words, the events A_1, A_2, A_3, … are pairwise mutually exclusive and

$$\bigcup_{i=1}^{\infty} A_i \,.$$

Let $\mathrm{B} \subseteq \Omega$ be another event. Then it follows that

$$P(B) = \sum_{i=1}^{\infty} P(A_i \cap B) \,.$$

or, equivalently,

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

We are now ready to state one of the most important theorems in modern probability theory.

Theorem: Bayes' theorem

Let A_1, A_2, A_3, … $\subseteq \Omega$ be a countably infinite set of a partition of a sample space Ω such that $P\big(A_j\big) > 0$ for all $j \in \mathbb{N}$. Then for fixed A_i with $i \in \mathbb{N}$ and $B \subseteq \Omega$ such that $P(B) > 0$, it holds that

$$P\big(A_j \big| B\big) = \frac{P\big(B\big|A_j\big)P\big(A_j\big)}{\sum_{j=1}^{\infty} P\big(B\big|A_j\big)P\big(A_j\big)}.$$

Proof

We know for the conditional probability formula yields

$$P\big(A_j\big|B\big)P(B) = P\big(B\big|A_j\big)P\big(A_j\big).$$

Dividing by $P(B)$ and we use the Law of Total Probability for the event $B \subseteq \Omega$ yielding the result.

Note that as a special case of Bayes' theorem, we can apply the results with just two events, A and B, and use the two events A and A^c as the partition. In this case, the result is reduced to

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

Example 1.11

Suppose that the probability of a randomly chosen person having developed cancer is 1% given that if a person has cancer, a medical test will yield a positive result with a probability of 98%. Also, given that if a person does not have cancer, the test will yield a negative result with a probability of 0.80. Now, suppose a randomly chosen person tests positive. What is the probability they have actually developed cancer?

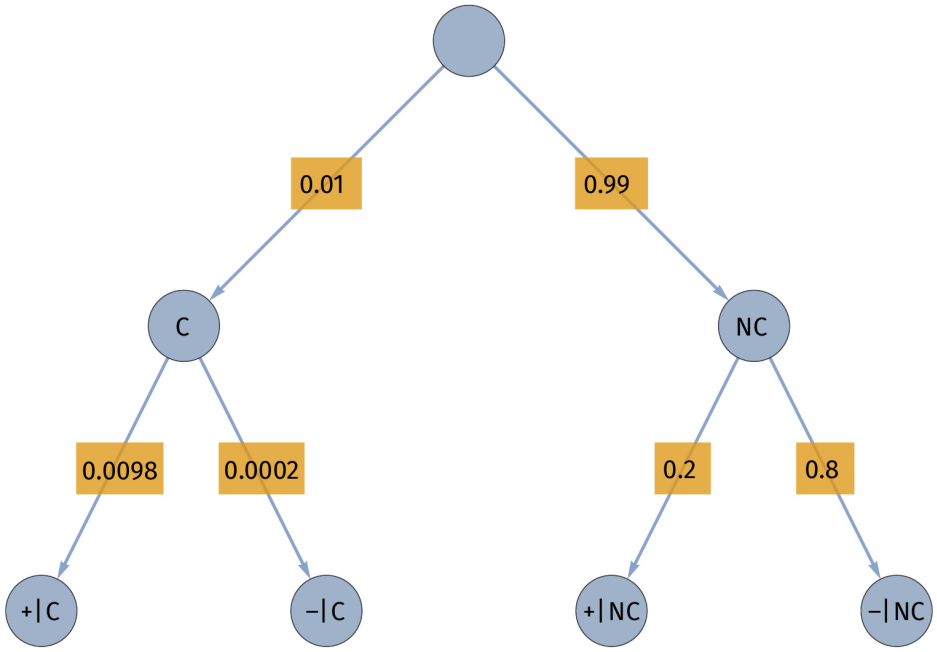
Solution 1.11

Let A denote the event that a person has cancer. We know that $P(A) = 0.01$ and $P(A^c) = 0.99$. Let B denote the event that the test returns a positive result. We know that $P(B\,|\,A) = 0.98$ and $P\big(B^c\big|\,A^c\big) = 0.80$. We want to find $P(A\,|\,B)$. Note that $P\big(B\big|\,A^c\big) = 1 - 0.80 = 0.20$. Now, using Bayes' theorem, we have

$$\begin{aligned} P(A\,|\,B) &= \frac{P(B|\,A)P(A)}{P(B|\,A)P(A) + P\big(B\big|A^c\big)P\big(A^c\big)} = \frac{0.98 \cdot 0.01}{0.98 \cdot 0.01 + 0.20 \cdot 0.99} \\ &= \frac{0.0098}{0.0098 + 0.198} \approx 0.0472 = 4.72\%. \end{aligned}$$

Note that the result in Solution 1.11 is very low and applies to biomedical tests with significant false positive and false negative rates, making population-wide screening programs of relatively rare diseases with such tests pointless. Tree diagrams and two-way tables help us understand how the total law of probability, Bayes' theorem, and applications such as the one in Example 1.11 work. Below is an example of such a probability tree together with the associated two-way table.

Figure 3: The Probability Tree Diagram from Example 1.11



Source: George Dekermenjian (2019).

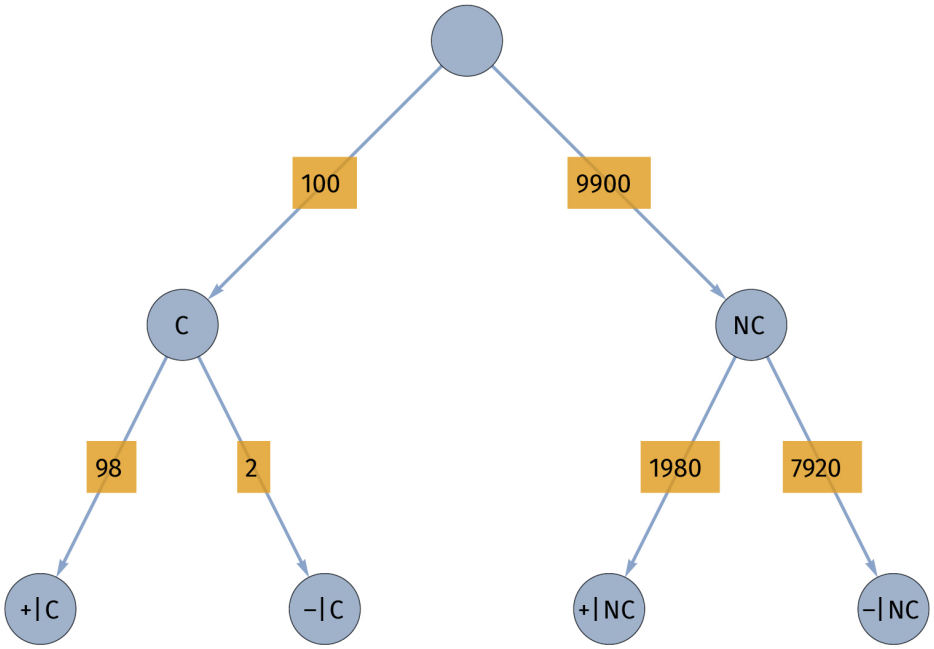
Table 2: Table of Probabilities from Example 1.11

		True Diagnosis		Total
		Cancer	No Cancer	
Medical test result	Positive	0.0098	0.1980	0.2078
	Negative	0.0002	0.7920	0.7922
Total		0.01	0.99	1

Source: George Dekermenjian (2019).

Now consider a sample with a size of 10,000. The natural frequencies corresponding to the probabilities help us get a “feel” for how these types of probabilities impact a real-world data set. Below is a tree diagram with natural frequencies followed by the corresponding two-way table.

Figure 4: The Tree Diagram of Natural Frequencies from Example 1.11



Source: George Dekermenjian (2019).

Table 3: Table of Natural Frequencies for a Sample Size of 10,000 from Example 1.11

		True Diagnosis		Total
		Cancer	No Cancer	
Medical test result	Positive	98	1980	2078
	Negative	2	7920	7922
Total		100	9900	10,000

Source: George Dekermenjian (2019).

In Bayes' theorem, $P(A)$ is interpreted as the prior probability while $P(A \mid B)$ is the posterior probability. So, for the example above, before knowing the test result, we could say with 1% probability that the person has cancer, but after getting the result of the test, we could say that, based on the new information, the probability that the person has cancer is almost 5%.

SUMMARY

Several fundamental concepts were introduced in this unit, including random experiment, outcome, event, sample space, probability axioms, and counting techniques.

We used these concepts to compute probabilities of certain events for simple experiments. Mutually exclusive events and the sum of probabilities axiom were used to compute probabilities of unions of events:

$P(A \cup B) = P(A) + P(B)$ for mutually exclusive events $A, B \subseteq \Omega$.

When events are not mutually exclusive, a general sum of probabilities rule gives

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ for any events $A, B \subseteq \Omega$.

The joint event $A \cap B$ led to a discussion of independent events in which case we have the product of probabilities rule

$P(A \cap B) = P(A) \cdot P(B)$ for any independent events $A, B \subseteq \Omega$.

When two events A and B are not independent, $P(A)$ and $P(A \mid B)$ are not the same. Therefore, we introduced the conditional probability of $A \subseteq \Omega$ conditioned on $B \subseteq \Omega$ by

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \quad \text{where} \quad P(B) > 0 .$$

This definition can be interpreted as a general product of probabilities for events that are not necessarily independent.

Bayes' rule is central to understanding Bayesian probability. We discussed instances where a collection of events partitions a sample space and how such a collection induces a partition of any event. These ideas led to an important theorem known as the law of total probability. Finally, building on this theorem, we introduced Bayes' theorem and discussed a number of applications.

UNIT 2

RANDOM VARIABLES

STUDY GOALS

On completion of this unit, you will be able to ...

- describe and compare the properties of discrete and continuous random variables.
- understand the roles of PMFs and CDFs for discrete distributions and their properties.
- understand the roles of PDFs and CDFs for continuous distributions and their properties.
- apply PMFs, PDFs, and CDFs to answer probability questions.
- identify important discrete distributions and important continuous distributions.

2. RANDOM VARIABLES

Introduction

Random variable
This is a rule (function) which assigns outcomes of a given sample space to a real number. The sample space is equipped with a probability measure such that the outcomes or events have a defined likelihood.

In real-world applications of data analysis and statistics we work with numerical data. In order to describe the occurrence of data points, a mathematical model or formalization, called a **random variable**, is necessary. From a scientific point of view, we assume that the data points are realizations of random variables. Each random variable has a specific sample space, probability measure and therefore, distribution, which describes the frequency of occurrence of our data points. A random variable is different from traditional variables in terms of the value it takes. It is a function which performs the mapping of the outcomes of a random process to a numeric value. Given their importance, the main subject of this unit will be random variables (see Wackerly, Mendenhall & Schaeffer, 2008) and their mathematical properties. Random variables have many real-world applications and are used, for example, to model stock charts, the temperature, customer numbers, and the number of traffic accidents that occur in a given timeframe or location.

2.1 Random Variables

Informally, a random variable is a rule that assigns a real number to each outcome of the sample space (see Wasserman, 2004). We usually denote random variables using the capital letters X, Y, Z. When appropriate, we sometimes also use subscripts: X_1, X_2 and so on.

Consider the random experiment of tossing a fair coin four times. Let X be the random variable that counts the number of heads. For the outcome $HHTH$, we have $X(HHTH) = 3$ and for another outcome $TTTH$, we have $X(TTTH) = 1$. Now consider the random experiment of rolling two fair 6-sided dice. Let Y denote the random variable that adds the numbers observed from each of the dice. For example, $Y((1,2)) = 3$ and $Y((4,4)) = 8$. Finally, the same random experiment can have many different random variables. For example, for the experiment of rolling two 6-sided dice, let M denote the random variable that gives the maximum of the numbers from the dice. For example, $M((1,2)) = 2$ and $M((5,2)) = 5$. Since the values of a random variable depend on the outcome, which is random, we know that the values of a random variable are random numbers.

So far, we have made the connection between the value of a random variable and an outcome of the random experiment. Before moving onto events, we will make this relationship more formal.

For a given sample space Ω, equipped with a probability measure P, a random variable X is a mapping from the sample space Ω to the set of real numbers $\mathbb{R}$ that assigns for each outcome $\omega \in \Omega$ a real number $x \in \mathrm{R}$. In standard notation, this is written as

$$\begin{aligned} X \quad &: \Omega \to \mathbb{R}, \\ &\omega \to x = X(\omega)\,. \end{aligned}$$

This is the most abstract definition of a random variable we can encounter. For our purposes we will restrict ourselves to discrete and piecewise continuous random variables and describe a wide range of random variables.

For random variables with a finite sample space, we can write down the possible values of a given random variable. Consider the random experiment of tossing three coins; let X denote the random variable which counts the number of tails. The table below gives the values of each of the outcomes.

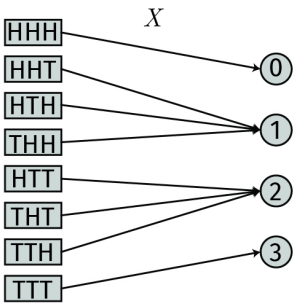
Table 4: Values of the Random Variable Counting Tails When a Coin is Tossed Three Times

ω	$X(\omega)$
HHH	0
HHT	1
HTH	1
THH	1
HTT	2
THT	2
TTH	2
TTT	3

Source: George Dekermenjian (2019).

Now we want to establish the connection of random variables with events. Consider the equation $X(\omega) = 1$ from the table above. There are three outcomes ω that fit this equation. If we put these three outcomes in a set, it becomes an event. More formally, the event $[X(\omega) = 1]$ corresponds to the event $\{HHT, HTH, THH\}$. We can also write this relationship as $X^{-1}(1) = \{HHT, HTH, THH\}$. Here, X^{-1} denotes the inverse relation. It takes a value (from $\mathbb{R}$) to an event (in Ω).

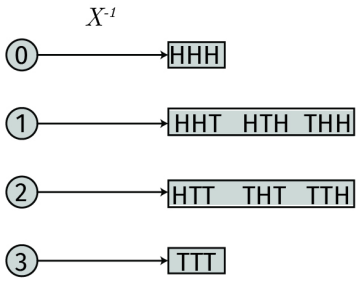
Figure 5: The Random Variable as a Mapping from the Sample Space to Real Numbers



The random variable counts the number of tails in three tosses of a coin.

Source: George Dekermenjian (2019).

Figure 6: The Inverse Mapping



Mapping the values of the random variable back to events in the sample space

Source: George Dekermenjian (2019).

It is standard practice to use shorthand notation when describing events using random variables. Formally, in the example above, the event $[X(\omega) = 1]$ describes the event $\{\omega \in \Omega | X(\omega) = 1\}$. However, in practice we usually write this event as $[X = 1]$, meaning that we have the following notation for the event that the random variable X equals one

$$[X(\omega) = 1] = \{\omega \in \Omega | X(\omega)\} = 1 = [X = 1]\,.$$

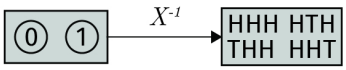
Since our sample space Ω is equipped with a probability measure P, we can ask ourselves how likely this event is. Consequently, the symbolic form of writing a probability such as "the probability of observing one tail in a sequence of three tosses of a coin" would be written as $P(X = 1)$. When we talk about the probability of all such (single value) events, we are describing a probability mass function. We will look at these functions in the next two sections.

Sometimes we are interested in events corresponding to multiple values of a random variable. The event of observing 0 or 1 tail can be written as

$$[0 \leq X(\omega) \leq 1] = X - 1(0,1) = [\omega \in \Omega | 0 \leq X(\omega) \leq 1],$$

which is written in shorthand as $[0 \leq X \leq 1]$.

Figure 7: The Inverse Mapping of a Set of Values



Mapping a set of values of the random variable back to an event in the sample space

Source: George Dekermenjian (2019).

An important range of values that comes up in the study of probability distributions is the range of values up to and including a specified number, such as $[X \leq 1]$ or $[X \leq 2]$. For our example above, the former is equivalent to $[0 \leq X \leq 1]$ and the latter is equivalent to $[0 \leq X \leq 2]$. When we speak about the probability of such events, we are describing a distribution function. This is the subject of the next section.

2.2 Probability Mass Functions and Distribution Functions

In the previous section, we defined a random variable as a rule that connects each outcome of a sample space to a real number. In this section, we will continue building the connection by taking the values of a random variable and connecting them to probabilities.

Probability Mass Functions

For a given sample space Ω and its corresponding probability measure P, we consider a random variable $X: \Omega \rightarrow \{x_1, x_2, x_3, ...\}$, where $x_1, x_2, \ldots$ are real numbers. This random variable is called a **discrete random variable**, because the set

Discrete random variable
A random variable which takes only finite or countably infinite values.

of possible values is countable infinite or finite. Now we consider a function $f: \{x_1, x_2, x_3, ...\} \rightarrow [0,1]$. If the function f satisfies the following properties

- $P(\omega \in \Omega \,|\, X(\omega) = \mathrm{i}) = P(X = i) = f(x_i)$ for all $i \in \mathbb{N}$,
- $\sum_{i=1}^{\infty} f(x_i) = 1$,

Probability mass function
A series which determines the likelihood that a discrete random variable takes a value.

f is then called a **probability mass function (PMF)**. In that case the support of the PMF consists of all $x_i \in \mathbb{R}$ such that

$$P(\omega \in \Omega \,|\, X(\omega) = i) = P(X = i) = f(x_i) > 0\,.$$

When we are working with multiple random variables, we can write f_x instead of f to specify the random variable to which the PMF refers.

Example 2.1

Consider the experiment of tossing a fair two-sided coin three times. Let X denote the random variable that counts the number of tails. Write down the PMF f of X defined by $f(x) = P(X = x)$.

Solution 2.1

The possible values of X are $\{0, 1, 2, 3\}$. The table below summarizes the PMF.

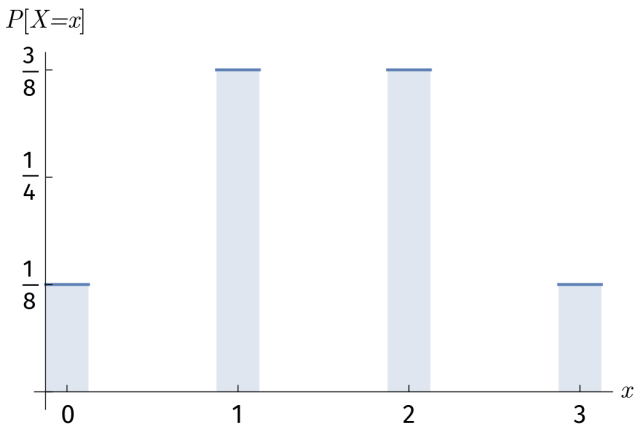
Table 5: Values, Events, and PMF of Tossing a Fair Coin Three Times

x	$[X = x]$	$f(x) = P(X = x)$
0	$\{HHH\}$	1/8
1	$\{HHT, HTH, THH\}$	3/8
2	$\{HTT, THT, TTH\}$	3/8
3	$\{TTT\}$	1/8

Source: George Dekermenjian (2019).

Note that each value $f(x)$ is non-negative and $f(0) + f(1) + f(2) + f(3) = 1$. Therefore, f is indeed a valid PMF.

Figure 8: A Plot of the PMF from Example 2.1



The PMF of the number of tails when tossing a fair coin three times

Source: George Dekermenjian (2019).

Example 2.2

Suppose that f is a PMF defined using the table below. What is $f(3)$?

Table 6: A Discrete PMF with a Missing Value

x	$f(x)$
1	0.2
2	0.05
3	?
4	0.39
5	0.01
6	0.05
7	0.05
8	0.10

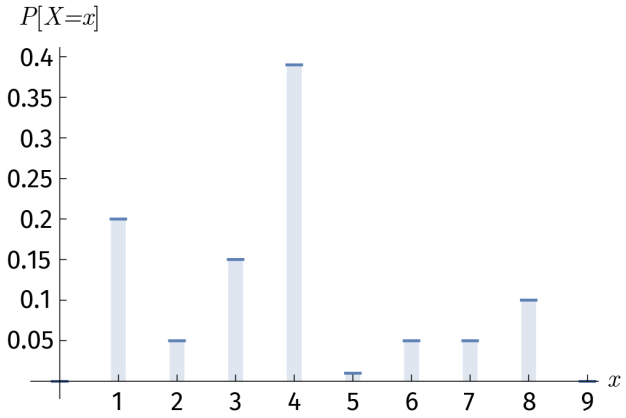
Source: George Dekermenjian (2019).

Solution 2.2

Since we are told that this is a probability mass function, we know that $f(3) \geq 0$ and $f(1) + f(2) + f(3) + ... + f(8) = 1$. Therefore, the second equation reduces to $0.85 + f(3) = 1$ which gives $f(3) = 0.15$.

Probability mass functions can be represented graphically as point plots with the horizontal axis containing the values of the random variable and the vertical axis containing the values of f. Below is a plot of the PMF for Example 2.2.

Figure 9: A Plot of the PMF from Example 2.2



Source: George Dekermenjian (2019).

Cumulative Distribution Function

Cumulative distribution function
A CDF of a random variable X is a function which measures the probability that X will take a value less or equal to x for fixed x.

In this section, we consider events corresponding to values of the random variable in the form $[X \leq x]$. Given a random variable X, the **cumulative distribution function** (CDF) is defined by

$$F_X(x) = P(X \leq x) \text{ for any } \mathrm{x} \in \mathrm{R}\,.$$

Formally, a CFD is a function F_X such that F_X: $\mathbb{R} \rightarrow [0,1]$. We also write F instead of F_X if the random variable is clear from the context. In addition, we can prove that for any CDF the following three properties must hold:

- F is normalized:

$$\lim_{x \rightarrow -\infty} F(x) = 0$$

and

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

- F is non-decreasing:

$$F(x_1) \leq F(x_2) \text{ for } x_1 < x_2$$

- F is right-continuous:

$$F(x) = \lim_{t = x, t > x} F(t) .$$

We can verify the three properties in an intuitive way. For the first property, when x tends to negative infinity, then the set of outcomes that are in the event $[X \leq x]$ become the empty set. Also, as x tends to positive infinity, the set of outcomes in the event become the whole sample space, in which case the probability is 1. For the second point, notice that if an outcome is in the event $[X \leq x_1]$ and $x_1 \leq x_2$, then automatically, this same outcome must be in $[X \leq x_2]$, which basically means

$$[X \leq x_1] \subseteq \quad [X \leq x_2] \quad \text{for} \quad x_1 \leq x_2 .$$

Therefore, the former event is a subset of the latter one. Hence, the former event is, at most, as probable as the latter. For the final property, take $t > x$, then we have

$$F(t) - F(x) = P(X \leq t) - P(X \leq x) = P(x < X \leq t) .$$

The event $[x < X \leq t]$ becomes the empty set as t approaches x. Therefore, the probability of this event tends to zero.

Example 2.3

Suppose X is a random variable with values in $\Omega = \{0,1,2,3,4\}$ and with PMF $f(x) = 1/5$ for all $x \in \Omega$. Such a random variable is said to have a discrete uniform distribution. Write down the CDF $F(x)$ of X and sketch its graph.

Solution 2.3

If $x < 0$, then $F(x) = P(X \leq x) = 0$ since X cannot take on values outside the given set. If $0 \leq x < 1$, then $F(x) = P(X \leq x) = P(X) = 0 = f(0) = 1/5$. For $1 \leq x < 2 F(x) = P(X \leq x) = P(X) = 0 + P(X) = 1 = f(0) + f(1) = 2/5$. Continuing in this way, we obtain

$$F(x) = \begin{cases} 0 & x < 0, \\ \frac{1}{5} & 0 \leq x < 1, \\ \frac{2}{5} & 1 \leq x < 2, \\ \frac{3}{5} & 2 \leq x < 3, \\ \frac{4}{5} & 3 \leq x < 4, \\ 1 & x \geq 4\,. \end{cases}$$

given that either the PMF or the CDF of a random variable completely describes everything we may want to know about the random variable.

2.3 Important Discrete Random Variables

In this section, we will discuss important discrete random variables, their probability mass functions, and cumulative distribution functions.

The Discrete Uniform Distribution

A random variable X, which takes on a finite number of integer values, such as $\{1,2,...,K\}$ for $K \in \mathbb{N}$, with each value being equally likely, is said to follow a discrete uniform distribution written f as $X \sim$ Uniform $(\{1,2,3,...,K\})$. The probability at each of these integers is

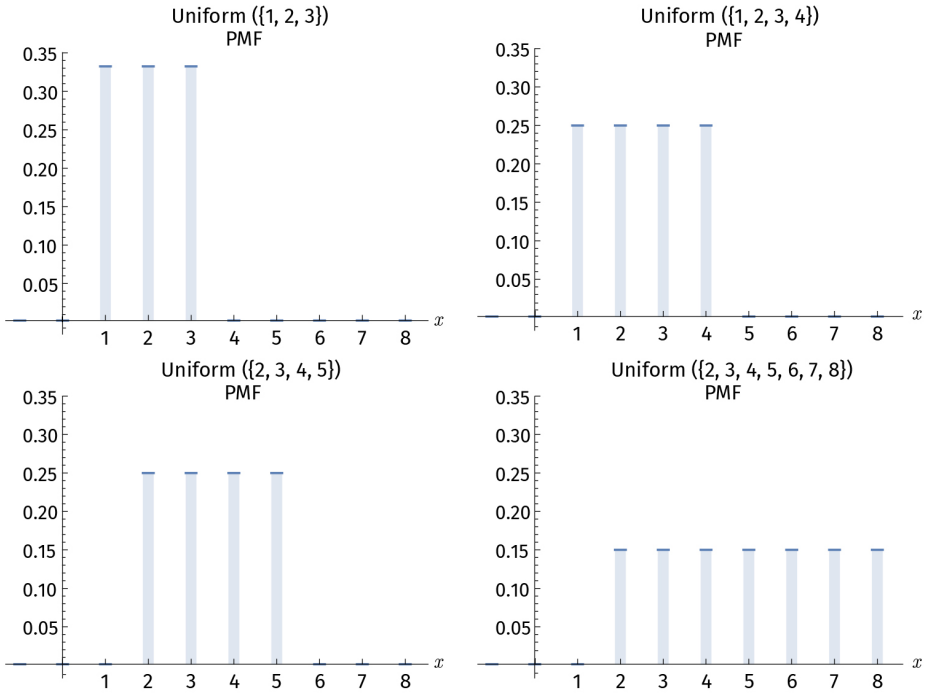
$$P(X = k) = 1/K \quad \text{for} \quad k \in 1,2,...,K,$$

and otherwise it is zero. Thus the PMF is

$$f(x) = \begin{cases} \frac{1}{K} & x = 1,2,\ldots,K, \\ 0 & \text{otherwise.} \end{cases}$$

Below are some PMF graphs for different discrete uniform distributions.

Figure 10: Plots of Various Discrete Uniform PMFs



Top Left: Uniform ({1, 2, 3})
Top Right: Uniform ({1, 2, 3, 4})
Bottom Left: Uniform ({2, 3, 4, 5})
Bottom Right: Uniform ({2, 3, 4, 5, 6, 7, 8})

Source: George Dekermenjian (2019).

The distribution function $F(x) = P(X \leq x)$ is given by

$$F(x) = \begin{cases} 0 & x < 1, \\ \frac{\lfloor x \rfloor}{K} & 1 \leq x < K, \\ 1 & x \geq K\,. \end{cases}$$

Note that the distribution function is non-decreasing, right continuous, and $F(x) \longrightarrow 1$ as $x \longrightarrow \infty$.

Example 2.4

Let X represent the face value of the roll of a fair six-sided die. Write down the PMF and CDF of X. Sketch a graph of both the PMF and CDF.

Solution 2.4

The possible values that X can take are $\{1,2,3,4,5,6\}$. Furthermore, since this is a fair die, each of the values is equally likely. Therefore, $X\sim$ Uniform $(\{1,2,3,4,5,6\})$. Its PMF is given by

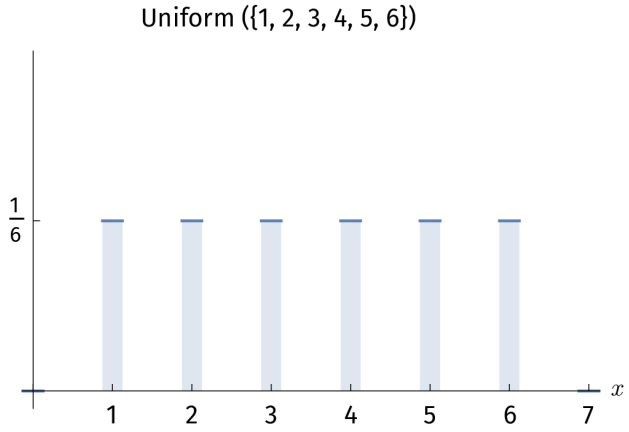
$$\mathrm{f}(x) = \begin{cases} \frac{1}{6} & \text{for } x \in \{1,2,3,4,5,6\}, \\ 0 & \text{otherwise.} \end{cases}$$

The CDF of X is given by

$$F(x) = \begin{cases} 0 & \text{for} \quad x < 1, \\ \frac{\lfloor x \rfloor}{6} & \text{for} \quad 1 \leq x < 6 \\ 1 & \text{for} \quad 6 \geq x\,. \end{cases}$$

Below is a graph of the PMF for Example 2.4.

Figure 11: A Plot of the PMF of Example 2.4



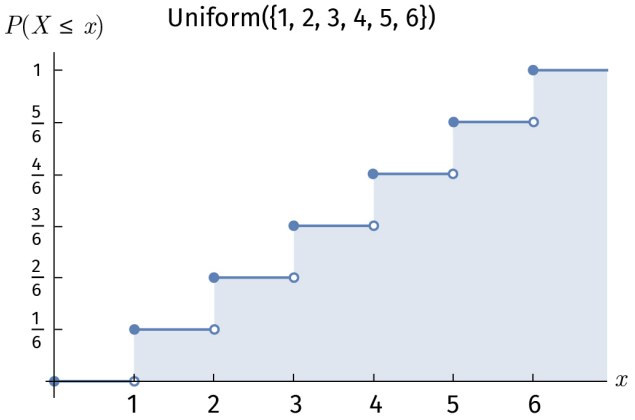
Source: George Dekermenjian (2019).

We can further simplify this expression to get

$$F(x) = \begin{cases} 0 & x < 1, \\ \frac{1}{6} & 1 \leq x < 2, \\ \frac{2}{6} & 2 \leq x < 3, \\ \frac{3}{6} & 3 \leq x < 4, \\ \frac{4}{6} & 4 \leq x < 5, \\ \frac{5}{6} & 5 \leq x < 6, \\ 1 & x \geq 6\,. \end{cases}$$

Below is a graph of the CDF for Example 2.4.

Figure 12: A Plot of the CDF for Example 2.4



Source: George Dekermenjian (2019).

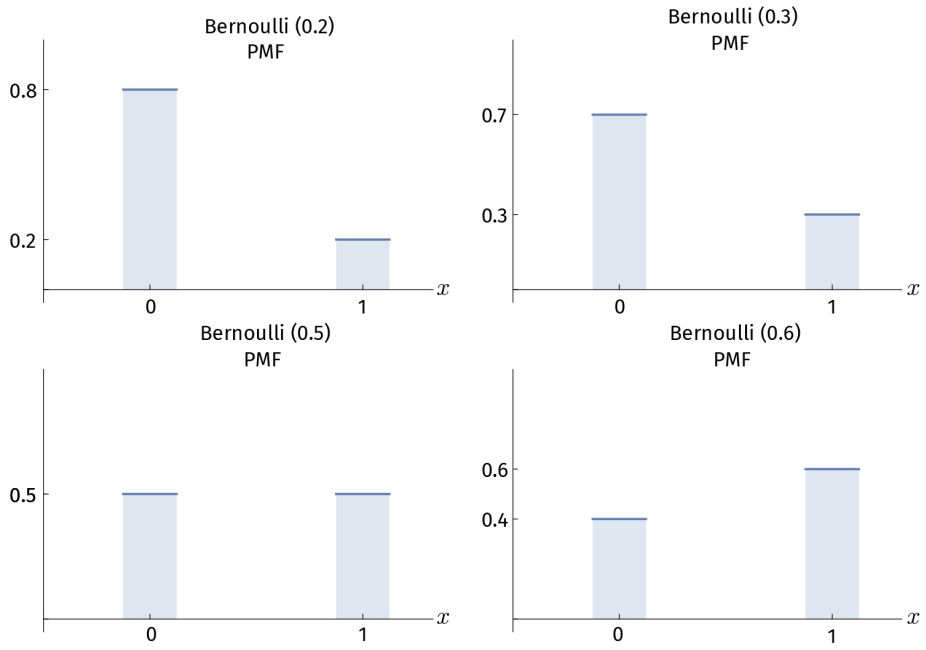
The next few discrete distributions are based on the so-called Bernoulli trial. The Bernoulli trial (or experiment) is a fundamental building block for some of the distributions we will consider in this section. As such, we begin our discussion with this distribution, which is arguably the simplest discrete distribution.

Bernoulli Trial

A Bernoulli trial, or a Bernoulli experiment, is an experiment that has exactly two possible outcomes. These outcomes are typically labeled "success" and "failure". Suppose the probability of "success" is p for $0 \leq p \leq 1$ and, consequently, the probability of "failure" is $1 - p$. Now, consider a random variable X defined on this sample space such that $X(\text{success}) = 1$ and $X(\text{failure}) = 0$. The PMF of this random variable is given by $f(1) = p$ and $f(0) = 1 - p$. Such a random variable is called a Bernoulli random variable with parameter p, written as $X \sim$ Bernoulli(p). Formally, the PMF of the Bernoulli distribution is given by

$$f(x) = \begin{cases} 1-p & \text{for } \ x=0, \\ p & \text{for } x=1, \\ 0 & \text{otherwise.} \end{cases}$$

Figure 13: A Plot of Various Bernoulli PMFs



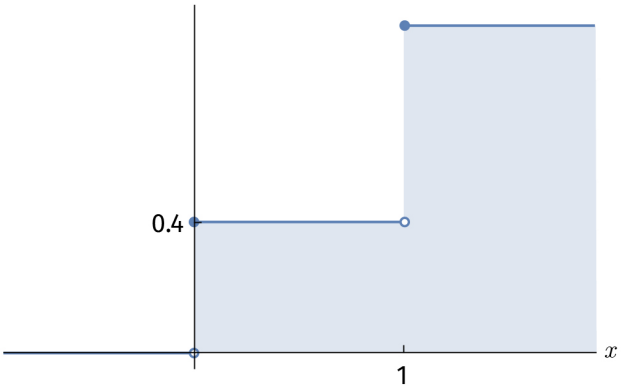
Top Left: Bernoulli (0.2)
Top Right: Bernoulli (0.3)
Bottom Left: Bernoulli (0.5)
Bottom Right: Bernoulli (0.6)

Source: George Dekermenjian (2019).

We notice that the PMF of that random variable is $\{0,1\}$. That means it is non-zero on that set. The distribution function is given by

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

Figure 14: A Plot of the CDF for Bernoulli (0.6)



Source: George Dekermenjian (2019).

The Binomial Distribution

Suppose we repeat a Bernoulli trial five times and each trial is independent and has the same probability of success p. Let X be the random variable which counts the total number of successes for these five trials. For example, we can have an outcome such as *"FFFFF"*, no successes, in which case $X = 0$. We could have *"SSSSS"*, five successes, in which case $X = 5$. These are the two limiting cases. In particular, X can take on values from the set $\{0,1,2,3,4,5\}$. Using independence, we can compute the PMF value at 0 as

$$\begin{aligned} f(0) &= P(X = 0) = P(FFFFF) \\ &= (1-p) \cdot (1-p) \cdot (1-p) \cdot (1-p) \cdot (1-p) = (1-p)^5 . \end{aligned}$$

Note that we have used the independence assumption in the third equality. Similarly, the PMF value at 5 is $f(5) = P(X = 5) = P(SSSSS) = p^5$.

In an effort to compute the value of the PMF for the other values, let us first consider an outcome that has two successes:

$$P(SFSFF) = p \cdot (1-p) \cdot p \cdot (1-p) \cdot (1-p) = p^2(1-p)^3 .$$

This is not the only outcome with two successes. There are some others: *"SSFFF"*, *"FSSFF"*, *"FFSSF"*, and so on. How many such outcomes are there? We answer this using the combinations formula: counting the number of ways the two successes can occur is like choosing two positions out of five where the successes will appear, and order does not matter. This is given by $\binom{5}{2}$. Therefore, the probability of observing two successes is

$$f(2) = P(X = 2) = \binom{5}{2} p^2(1-p)^3 .$$

Let us examine this expression in more detail:

$$\underbrace{\binom{5}{2}}_{\#\text{ outcomes with 2 successes}} \qquad \underbrace{p^2}_{\text{probability of 2 successes}} \qquad \underbrace{(1-p)^3}_{\text{probability of 3 failures}}$$

Similarly, the value of the PMF at 4 would be $f(4) = \binom{5}{4} p^4 (1-p)$.

With this mindset, we are ready to define the binomial distribution. The general binomial distribution has two parameters: the number of independent and identical trials (n) and the probability of observing a success on any of these trials (p). Remember that this probability does not change from one trial to the next. A random variable which has a binomial distribution with the parameters n and p is written as $X \sim \text{Binomial}(n, p)$. The PMF of X is then given by

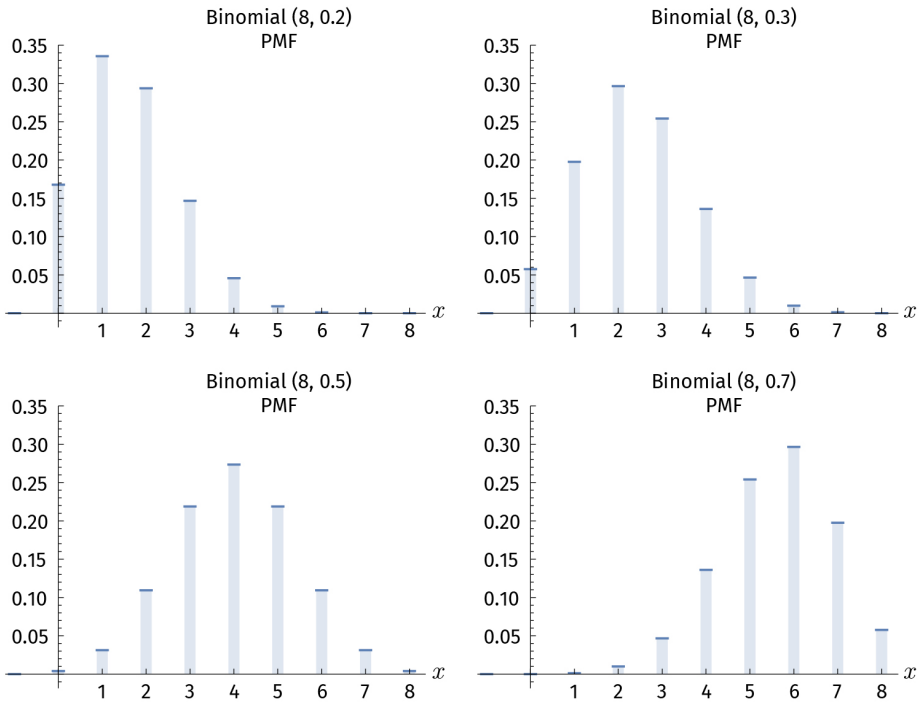
$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{if } x = 0,1,2,\ldots,n, \\ 0 & \text{otherwise.} \end{cases}$$

Note that this PMF satisfies $f(x) \geq 0$ for every x and, furthermore, by a simple application of the binomial theorem, one can see that

$$1 = (p + (1-p))^n = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} = \sum_{x=0}^{n} f(x)\,.$$

Therefore, f is indeed a valid PMF.

Figure 15: A Plot of the PMFs for Various Binomial Distributions



Top Left: Binomial (8, 0.2)
Top Right: Binomial (8, 0.3)
Bottom Left: Binomial (8, 0.5)
Bottom Right: Binomial (8, 0.7)

Source: George Dekermenjian (2019).

Example 2.5

If a fair coin is tossed ten times, what is the probability of observing exactly four tails? Interpret your answer.

Solution 2.5

Although we can solve this problem with standard tools of counting and probability, it is more efficient to model this problem using the binomial distribution. We note that each toss of the coin is independent of other tosses, and since we are tossing the same (fair) coin, the probability of success, $p = 0.5$, stays the same from one toss to the next. Finally, we have a fixed number of tosses (trials), $n = 10$. Let X be the number of tails ("successes"), then $X\tilde{}$ Binomial(10,0.5). The PMF of X is given by

$$f(x) \ = \binom{10}{x}(0.5)^x(1-0.5)^{10-x} = \binom{10}{x}0.5^{10}, \ x = 0,1,2,\ldots,10.$$

To answer the question, the probability of observing exactly four tails is given by

$$P(X = 4) = f(4) = \binom{10}{4} 0.5^{10} \approx 0.2051 \,.\, (20.51\ \%).$$

This means that if this experiment (of tossing a coin ten times) is performed many, many times, about 20% of them will end up observing exactly four tails in the long run. More concretely, if 100 different people toss a fair coin ten times, about 20 of them are expected to see exactly four tails.

Example 2.6

Continuing on from Example 2.5, calculate the probability of observing at least one tail.

Solution 2.6

In the context of Example 2.5, this is the probability of observing at least one success. In other words,

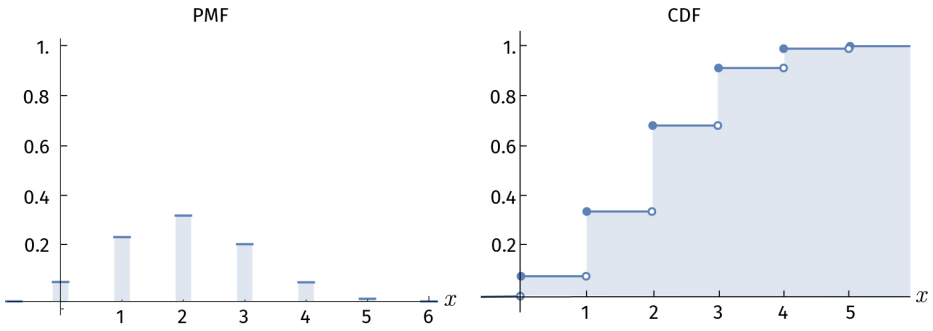
$$\begin{aligned} P(X \geq 1) &= P(X = 1) + P(X = 2) + \ldots \quad P(X = 9) + P(X = 10) \\ &= f(1) + f(2) + \ldots \quad f(9) + f(10) \,. \end{aligned}$$

This requires computing the PMF ten times! We approach this problem by using the complement event $P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 0) = 1 - f(0)$.

$$f(0) = \binom{10}{0} 0.5^{10} \approx 0.0010$$

Therefore, the probability of observing at least one tail is approximately $1 - 0.0010 = 0.9990$, which is about 99.9 %.

Figure 16: Plots of the PMF and CDF of Binomial (5, 0.4)



Left: The PMF of Binomial (5, 0.4)
Right: The CDF of Binomial (5, 0.4)

Source: George Dekermenjian (2019).

MNEMONIC

When the number of trials n is large, obtaining the exact binomial PMF is computationally very expensive if defined as above. Almost all data science libraries have more efficient implementations of the binomial PMF that you should use instead of coding it yourself. The computational expense is mainly due to the factorial calculation in the binomial coefficient.

The Geometric Distribution

For our next distribution, we will consider performing independent Bernoulli trials where the probability of success, p, is fixed from one trial to the next once again. However, unlike the binomial distribution, we do not fix the number of trials. Instead, we perform the trials until we observe the first success.

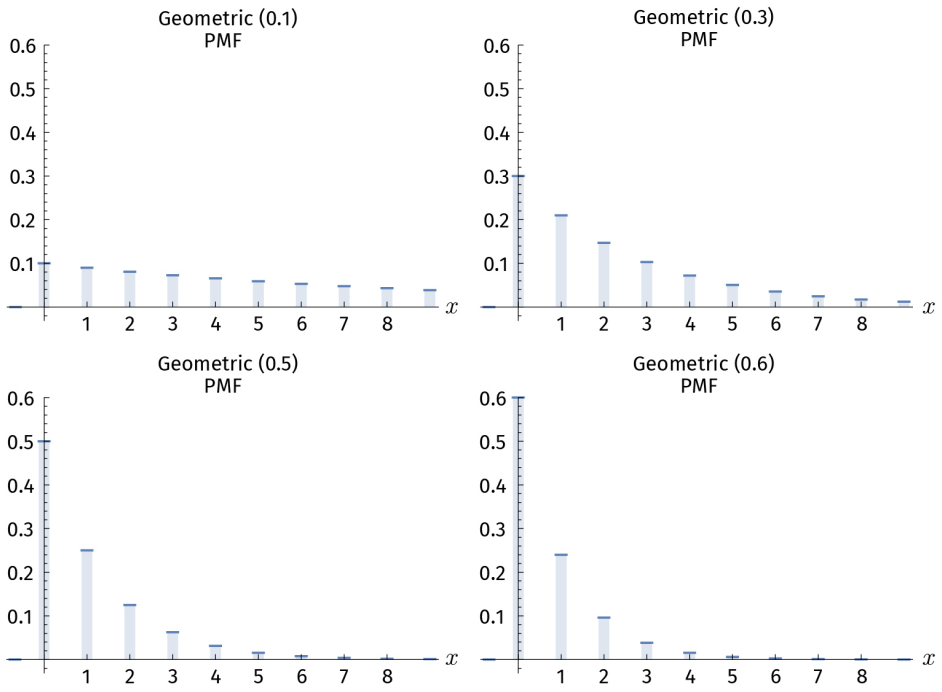
Let X count the failures until the first success occurs. The possible values of X are $0,1,2,3,$ …. We may observe a success on the first trial, in which case $X = 0$ (no failures). As there is no upper bound on the number of trials, we may get quite unlucky! In this setup, X is said to have a geometric distribution, with the probability of success p written as X~Geometric(p). Remember that we can only observe outcomes with one success and this success occurs on the last trial. Therefore, possible outcomes look like *"S"*, *"FS"*, *"FFS"*, *"FFFS"*, *"FFFFS"*, etc. Using independence, values of the PMF, f, for X are:

- $f(0) = P(S) = p$,
- $f(1) = P(FS) = (1 - p)p$,
- $f(2) = P(FFS) = (1 - p)^2 p$,
- $f(3) = P(FFFS) = (1 - p)^3 p$.

Continuing in this way, we see that the PMF of X is given by

$$f(x) = \begin{cases} (1-p)^x p & \text{if } x = 0,1,2,3,\ldots, \\ 0 & \text{otherwise.} \end{cases}$$

Figure 17: A Plot of the PMFs for Various Geometric Distributions



Top Left: Geometric (0.1)
Top Right: Geometric (0.3)
Bottom Left: Geometric (0.5)
Bottom Right: Geometric (0.6)

Source: George Dekermenjian (2019).

We immediately have $f(x) \geq 0$ for all x. From the formula of a geometric series, we also have

$$\begin{aligned}\sum_{x=0}^{\infty} f(x) &= \sum_{x=0}^{\infty} (1-p)^x p = p \sum_{x=0}^{\infty} (1-p)^x = p \cdot \frac{1}{1-(1-p)} \\ &= p \cdot \frac{1}{p} = 1.\end{aligned}$$

Therefore, f is indeed a valid PMF.

The CDF of X, denoted by $F(x) = P(X \leq x)$, is given by

$$F(x) = \begin{cases} 0 & x < 0, \\ \sum_{k=0}^{\lfloor x \rfloor} (1-p)^k p & x \geq 0\,. \end{cases}$$

Further simplifying this expression using the formula for the partial sum of a geometric series with the ratio $(1 - p)$, we see that

$$F(x) = \begin{cases} 0 & x < 0, \\ 1 - (1-p)^{\lfloor x \rfloor + 1} & x \geq 0\,. \end{cases}$$

Notice that $F(x) \geq 0$, non-decreasing, and $F(x) \to 1$ as $x \to \infty$.

Example 2.7

We roll a fair six-sided die until we observe five. What is the probability that we roll the die three times? Interpret your answer.

Solution 2.7

Each roll of the die can result in one of two outcomes: rolling a 5 ("success") and not rolling a 5 ("failure"). We repeat the trial until we see a success. If X is the number of failures (not rolling a 5) before the first success (rolling a 5), then its probability can be modeled using a geometric distribution. The parameter of this distribution is the probability of success (rolling a 5). Since the die is fair, this probability is $p = 1/6$. Therefore, we have $X \sim \text{Geometric } (1/6)$. Finally, the probability that we roll the die three times is the probability that we see two failures, which is given by

$$P(X = 2) = f(2) = \left(1 - \tfrac{1}{6}\right)^2 \cdot \tfrac{1}{6} = \tfrac{25}{216} \approx 0.1157$$

or 11.57 %.

Example 2.8

Suppose that 15 % of males in Germany are daily smokers. If we approach males at random in Germany to check whether they are daily smokers until we find a smoker, what is the probability that we approach three people?

Solution 2.8

Let X denote the number of non-smokers ("failures") before we speak to a daily smoker ("success"). The probability of "success" here is 0.15 and the probability of "failure" is $1 - 0.15 = 0.85$. Therefore, $X \sim \text{Geometric}(0.15)$. The PMF of X is given by

$$f(x) = \begin{cases} (0.85)^x (0.15) & \text{if } x = 0,1,2,3,\ldots, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, the probability of speaking to three people is $f(2) = (0.85)^2 (0.15) \approx 0.1084$ or 10.84 %.

The Negative Binomial Distribution

The negative binomial distribution is a generalization of the geometric distribution. While the Bernoulli trials associated with the geometric distribution stop after seeing the first success, for the negative binomial distribution, the Bernoulli trials stop after we observe the first $n \in \mathbb{N}$ successes. If we put $n = 1$, then we get to the geometric distribution.

Let us suppose that $n = 2$ and X the number of failures before seeing the $n^{th} = 2^{nd}$ success. We say that X follows a negative binomial distribution with parameters 2 and p, written as $X \sim \text{Negative} - \text{Binomial}(2, p)$. Write down all the outcomes for $X = 4$—that is, four Bernoulli trials: *"FFFFSS"*, *"FFFSFS"*, *"FFSFFS"*, *"SFFFFS"*, and *"FSFFFS"*. Note that we stop as soon as we see the second success; thus, every outcome must end with a success. As before, let p denote the probability of success so that $(1 - p)$ denotes the probability of observing a failure. The probability of each of the outcomes listed above is the sample: $P(FFFFSS) = \cdots = P(SFFFFS) = (1 - p)^4 p^2$. Since these events are mutually exclusive, we have $P(X = 4) = 5(1 - p)^4 p^2$.

Now suppose that $n = 3$ and $X \sim \text{Negative} - \text{Binomial}(3, p)$. How many outcomes does the event $[X = 4]$ have? We have to write down all the outcomes that have four failures and three successes where the third success is on the last trial.

Figure 18: Outcomes Corresponding to $[X = 4]$ where $X \sim \text{Neg} - \text{Binomial}(3,\ p)$

FFFFSSS	FFFSFSS	FFFSSFS	FFSFFSS	FFSFSFS
FFSSFFS	FSFFFSS	FSFFSFS	FSFSFFS	FSSFFFS
SFFFFSS	SFFFSFS	SFFSFFS	SFSFFFS	SSFFFFS

Source: George Dekermenjian (2019).

There are 15 such outcomes. The probability of each of these outcomes is $(1 - p)^4 p^3$, and since they are mutually exclusive, $P(X = 4) = 15(1 - p)^4 p^3$.

The factor 15 comes from counting the number of ways of observing four failures and three successes where the third success is on the last trial. Since the outcome of the last trial is always a success, this number just counts the number of ways of choosing where the remaining two successes can be observed out of the six (four failures and two successes) possible positions. This is exactly,

$$\binom{4+3-1}{3-1} = \binom{6}{2} = 15.$$

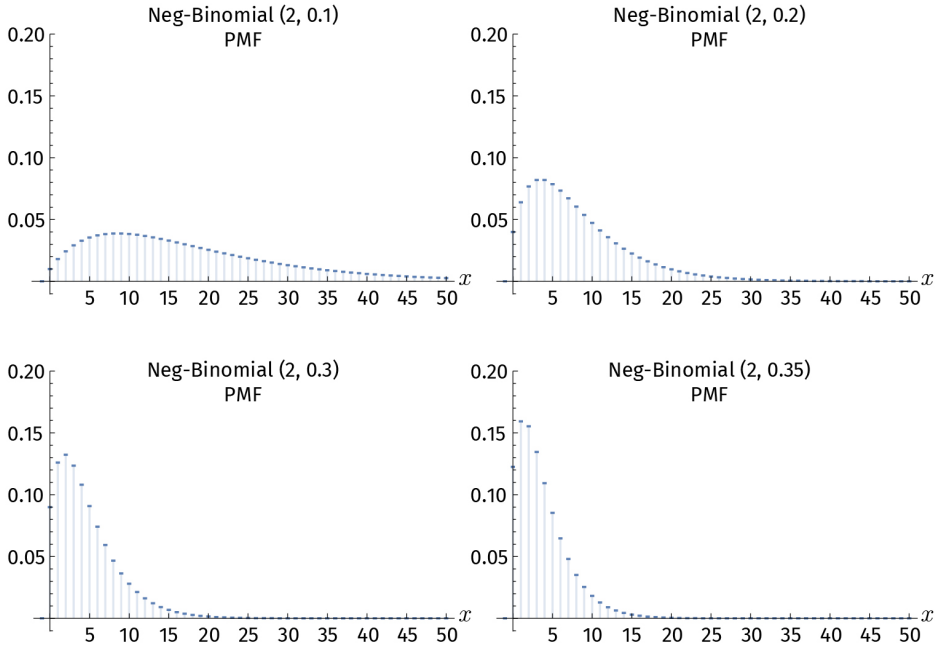
The next factor, $(1-p)^4$, comes from the four failures and the final factor; p^3 comes from the three successes.

We are now ready to write down the PMF for any random variable that has a negative binomial distribution with parameters n and p—that is, for $X \sim \text{Negative} - \text{Binomial}(n, p)$:

$$f(x) = \begin{cases} \binom{x+n-1}{n-1}(1-p)^x p^n & x = 0,1,2,\ldots, \\ 0 & \text{otherwise.} \end{cases}$$

Here are some graphs of various random variables that have a negative binomial distribution.

Figure 19: A Plot of the PMFs for Various Neg-Bionomial Distributions



Top Left: Neg-Binomial (2, 0.1)
Top Right: Neg-Binomial (2, 0.2)
Bottom Left: Neg-Binomial (2, 0.3)
Bottom Right: Neg-Binomial (2, 0.35)

Source: George Dekermenjian (2019).

The CDF $F(x) = \sum_{k=0}^{\lfloor x \rfloor} f(x)$ does not have a simple closed form equation, but most programming languages have a stable implementation.

The Poisson Distribution

Suppose that, on average, four car accidents per week occur at a certain intersection. Suppose these accidents also occur independently. In other words, if we observe an accident, it does not affect the probability of observing a second accident. Finally, suppose that only one accident can occur at a given moment. In other words, there is a small interval of time during which either one accident occurs, or no accident occurs. If X denotes the random variable that counts the number of accidents in a week, then we say that it follows a Poisson distribution with a rate $\lambda = 4$ (see Kim, 2019c). This is written as $X \sim \text{Poisson}(4)$ or, in general, $X \sim \text{Poisson}(\lambda)$.

The PMF of $X \sim \text{Poisson}(\lambda)$ is given by

$$f(x) = \begin{cases} \frac{\lambda^x}{x!}\text{e}^{-\lambda} & x = 0,\ 1,\ 2,\ \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

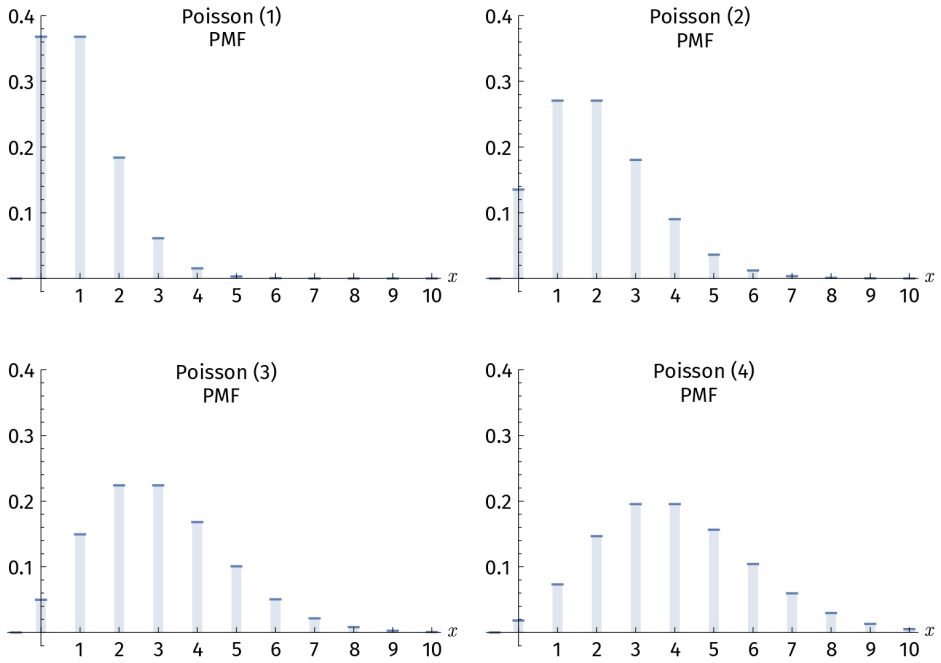
In this example, the probability of observing two accidents in a given week is then given by

$$P(X = 2) = f(2) = \frac{4^2}{2!}\text{e}^{-4} \approx 0.1465$$

or 14.65%.

The following graphs show Poisson PMFs:

Figure 20: A Plot of the PMFs for Various Poisson Distributions

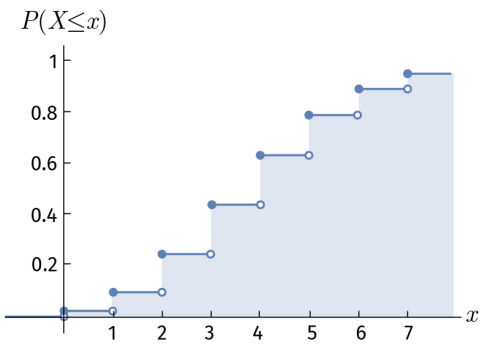


Top Left: Poisson (1)
Top Right: Poisson (2)
Bottom Left: Poisson (3)
Bottom Right: Poisson (4)

Source: George Dekermenjian (2019).

As for the negative binomial distribution, the CDF of the Poisson distribution does not have a simple closed form equation. The graph below shows the CDF of Poisson (4).

Figure 21: A Plot of the CDF for Poisson (4)



Source: George Dekermenjian (2019).

Example 2.9

The customer support center of company ABC receives calls according to a Poisson distribution at an average of 10 calls per hour. Within a given hour, what is the probability that

a) no more than two calls are received.
b) at least three calls are received.

Solution 2.9

Let X denote the number of calls received in a given hour. We know that X follows a Poisson distribution with parameter $\lambda = 10$, which means $X \sim \text{Poisson}(10)$. The PMF is given by

$$f(x) = \begin{cases} \frac{10^x}{x!}\mathrm{e}^{-10} & x = 0,\ 1,\ 2,\ 3,\ \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

a) The probability that no more than two calls are received is $P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = f(0) + f(1) + f(2)$. We calculate each of the summands separately:

$$f(0) = \frac{10^0}{0!}e^{-10} = \mathrm{e}^{-10} \approx 0.00005,$$

$$f(1) = \frac{10^1}{1!}\mathrm{e}^{-10} = 10\mathrm{e}^{-10} \approx 0.00045,$$

$$f(2) = \frac{10^2}{2!}\mathrm{e}^{-10} = 50\mathrm{e}^{-10} \approx 0.00227\,.$$

Therefore, $P(X \leq 2) = 61e - 10 \approx 0.0028$ or 0.28%.

b) The probability that at least three calls are received is $P(X \geq 3) = P(X = 3) + P(X + 4) + P(X = 5) + ... = f(3) + f(4) + f(5) + ...$, which is a very tedious calculation. However, from basic probability we can use the complement event to greatly simplify things. In particular, $P(X \geq 3) = 1 - P(X < 3) = 1 - P(X \leq 2) = 1 - 61e - 10 \approx 1 - 0.0028 = 0.9972$ or 99.72%.

2.4 Important Continuous Random Variables

Many random variables that are observed in the real world fall into this category. For example, as the number of births in a given day is discrete (0,1,2, ...), the mass (in grams) of a baby at birth can be any non-negative real number. The distribution function of a random variable tells us everything we might want to know about a random variable.

Let X be a random variable. Recall that the distribution function or CDF is given by $F_X(x) = P(X \leq x)$, which is normalized, non-decreasing, and right-continuous. If, in addition, the CDF is also left-continuous (or just continuous) and has a derivative that is continuous everywhere (except possibly at a countable number of points), then the associated random variable is said to be continuous. The CDFs we will be studying in this unit will satisfy this requirement.

All the distribution functions we discussed for discrete random variables had discontinuities at exactly the points that had a positive probability. Everywhere else (where the PMF was zero), the distribution function was continuous. Since we want the distribution function of continuous random variables to be continuous everywhere, this suggests that $P(X = x) = 0$ for every x.

The range of a continuous random variable X is, for example, the interval $[0,1]$. This means that X can take on any value within this interval. However, as mentioned above, we cannot assign a positive probability for any one value. Instead, we can assign a probability for a sub-interval of values, as for example $P(0.2 \leq X \leq 0.3)$. Since this interval can be made arbitrarily small, say $P(x \leq X \leq x + h)$ for small h, we can consider the density of the random variable at x by examining the rate of change of the probability at x.

Since we require the CDF to be continuous, this reduces to examining the expression

$$\frac{F(x+h) - F(x)}{h}$$

for small values of h. As you know from calculus, as $h \to 0$, we have

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

the derivative of the distribution function.

This derivative, $f(x)$ is what we call the **probability density function (PDF)** of X. To make it especially clear, we will sometimes write down $f_X(x)$. It is extremely important to note that the PDF does not give us probabilities but rather the density, which needs to be summed to get a probability. Although we define the PDF based on the CDF, it is common in practice to instead define the distribution of a random variable by writing down the PDF and then using integration (the reverse process of differentiation) to write down the CDF.

Probability Density Function (PDF)
A PDF, in cases where it exists, is a function which computes the cumulative distribution. function of a random variable.

For a given sample space and its corresponding probability measure P, we consider a random variable $X : \Omega \to \mathbb{R}$. This random variable has a continuous density function (CDF), if there exists a piecewise continuous function $f : \mathbb{R} \to [0,1]$, satisfying the following properties

- $P(\omega \in \Omega \,|\, X(\omega) \leq x) = P(X \leq\ x) = \int_{-\infty}^{x} f(t)\mathrm{d}t$ for all $x \in \mathbb{R}$,
- $\int_{-\infty}^{\infty} f(t)dt = 1$.

We also say in that case that X is a (piecewise) continuous random variable. The support of the PDF consists of all $x \in \mathbb{R}$ such that

$$f(x) > 0\,.$$

The Uniform Distribution

We have already looked at the discrete uniform distribution; the (continuous) uniform distribution is its continuous analog. For such that $a < b$ random variable X is said to follow a uniform distribution over the interval $[a, b]$ if its PDF is given by

$$f(x) = \begin{cases} 0 & a \leq x \leq b, \\ \frac{1}{b-a} & \text{otherwise}\,. \end{cases}$$

In other words, the density is constant (uniform) across the range $[a, b]$. Consequently, the CDF has a constant slope on this interval. We can write down the CDF from a PDF as follows:

$$F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t$$

This particular PDF is 0 for $x < a$. For $a \leq x < b$ we have

$$F(x) = \int_a^x \frac{1}{b-a} dt = \frac{x-a}{b-a}$$

and for $x \geq b$, it holds $F(x) = 1$. Altogether, the CDF of $X \sim \mathrm{Uniform}(\{a, b\})$ is given by

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

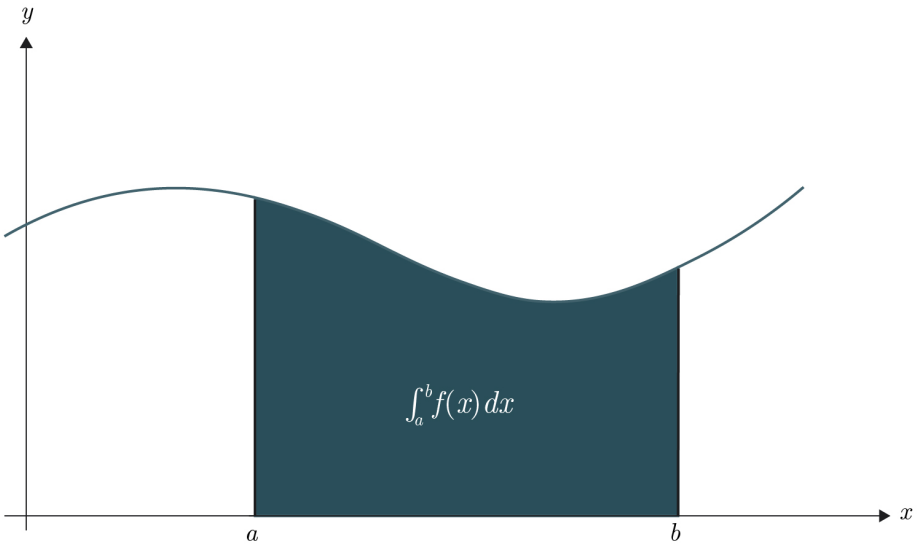
A Short Introduction to Integration

For fixed $a, b \in \mathbb{R}$ such that $a < b$ and a given (piecewise) continuous function $f\colon [a, b] \to \mathbb{R}$ the integral of f on the interval $[a, b]$, which is denoted by

$$\int_a^b f(x)\mathrm{d}x,$$

is, in its most intuitive form, the blue-colored area of the following picture.

Figure 22: The Meaning of an Integral



The blue-colored area denotes the integral

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

Source: Bartosch Ruszkowski (2022).

In the picture, the function f is assumed to be non-negative for all $x \in \mathbb{R}$, which is sufficient for our purposes, hence we consider PDFs of random variables, which satisfy the non-negativity.

We recall that for a given PDF $f\colon \mathbb{R} \longrightarrow \mathbb{R}$ the integral must satisfy

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1,$$

meaning that the area is one, which is the same as saying that the total probability is one.

We recall the **Fundamental Theorem of Calculus**, which is one of the most powerful algebraic tools to compute integrals. For fixed $a, b \in \mathbb{R}$ such that $a < b$ and a given (piecewise) continuous function $f\colon [a,b] \longrightarrow \mathbb{R}$ the function $F\colon [a,b] \longrightarrow \mathbb{R}$ is called primitive function of f, if it satisfies

$$F'(x) = f(x) \text{ for all } x \in [a,b],$$

then, it holds

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

As an example, let us consider the function

$$f(x) = 3x^2 \quad \text{for} \quad x \in [0,1]\,.$$

The function $F(x) = x^3$ for $x \in [0,1]$ satisfies

$$F'(x) = f(x) \text{ for all } x \in [0,1],$$

which means that the following integral is algebraic computable in the following sense

$$\int_0^1 3x^2\, dx = 1^3 - 0^3 = 1\,.$$

We see in this example that we constructed a PDF, since the integral yields one. In detail, this example shows that the function

$$\begin{cases} 3x^2 \text{ for } x \in [0,1], \\ \quad 0 \text{ otherwise}, \end{cases}$$

is a PDF and its CDF is given by

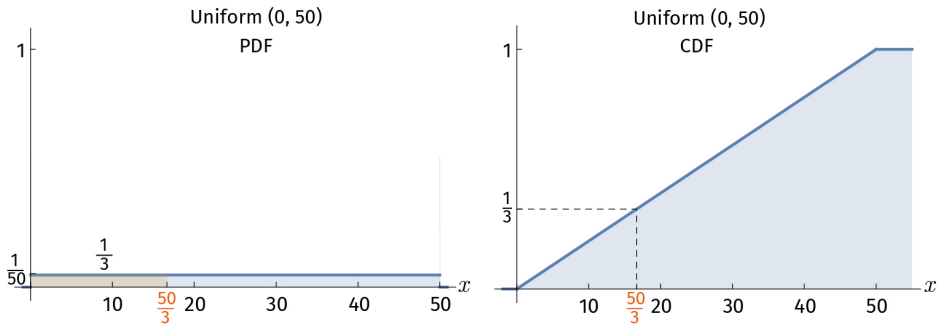
$$F(x) = P[X \leq x] = \begin{cases} 0 & \text{for } x < 0, \\ x^3 & \text{for } x \in [0,1], \\ 1 & \text{for } x > 1\,. \end{cases}$$

A parachutist lands at a random point between two markers (A and B) that are 50 meters apart. Find the probability that the distance to marker A is less than 1/3 of the distance to marker B. We can model the parachutist's landing point as X. We have $X \sim \text{Uniform}([0,50])$, where 0 is the location of marker A and 50 is the location of marker B. The PDF is given by $f(x) = \frac{1}{50-0} = \frac{1}{50}$ for $0 \leq x \leq 50$ and 0 otherwise. The CDF is given by

$$F(x) = \begin{cases} 0 & x < 0, \\ \frac{x}{50} & 0 \leq x < 50, \\ 1 & x \geq 50\,. \end{cases}$$

Below are the graphs of the PDF and CDF respectively.

Figure 23: Plots of the PDF and CDF of a Uniform Distribution (x=50/3)



Left: The PDF of Uniform (0, 50) showing the area under the curve between 0 and $\frac{50}{3}$
Right: The CDF of Uniform (0, 50) highlighting the point at $x = \frac{50}{3}$

Source: George Dekermenjian (2019).

The probability that the parachutist lands less at a point that is less than 1/3 the distance to marker A than marker B is

$$P\left(X < \frac{50}{3}\right) = F\left(\frac{50}{3}\right) = \frac{\frac{50}{3}}{50} = \frac{1}{3}$$

Now let us calculate the probability that the parachutist lands at least 20 meters from marker A and at least 10 meters from marker B. This corresponds to the event $[20 \leq X \leq 50]$ and so $P(20 \leq X \leq 40) = P(X \leq 40) - P(X < 20) = F(40) - F(20)$.

Furthermore, from the fundamental theorem of calculus, we know that

$$F(40) - F(20) = \int_{20}^{40} f(t)\mathrm{d}t$$

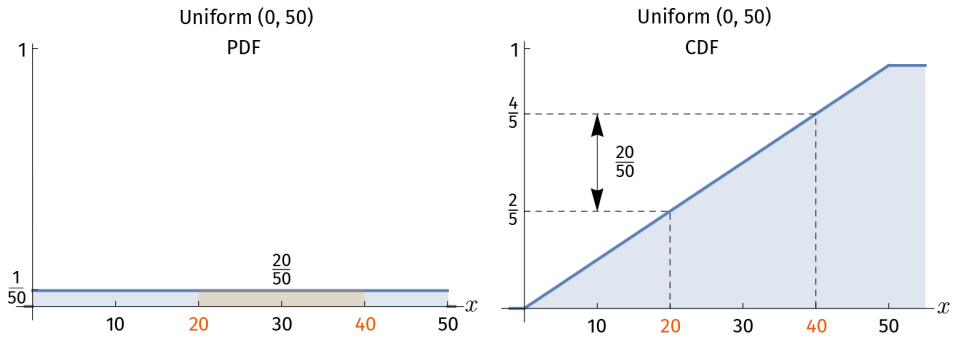
Therefore, we have two ways of computing probabilities. If we have the CDF in simple form, we can plug in the relevant numbers as follows:

$$P(40 \leq X \leq 20) = F(40) - F(20) = \frac{40}{50} - \frac{20}{50} = \frac{20}{50} = 0.4 \text{ or } 40\ \%.$$

In some cases, we will not be able to write down the CDF in simple form. For such distributions, it will be necessary to write down the integral expression using the PDF and, if necessary, approximate this integral as follows:

$$P(20 \leq X \leq 40) = \int_{20}^{40} f(t)\mathrm{d}t = \int_{20}^{40} \frac{1}{50}\mathrm{d}t = \frac{20}{50}.$$

Figure 24: Plots of the PDF and CDF of a Uniform Distribution (x=20 and x=40)



Left: The PDF of Uniform (0, 50) showing the area under the curve between x = 20 and x = 40
Right: The CDF of Uniform (0, 50) highlighting the points at x = 20 and x = 40

Source: George Dekermenjian (2019).

If you examine the figure above, you will find that this probability is the change of the y-coordinate in the CDF graph from $x = 20$ to $x = 40$. Additionally, it is also the area under the PDF curve between these two x values. The mathematical relationship between these two computations and the graphs is crucial to understanding the nature of PDFs and CDFs and to using them correctly in relevant applications.

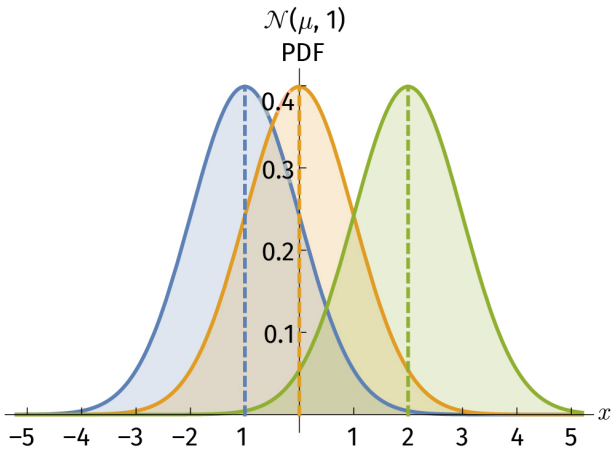
The Normal Distribution

Arguably the most widely known distribution both among academics and non-academics is the bell-shaped distribution, known as the Normal (Gaussian) Distribution. Indeed, many natural quantities have a shape that is approximately bell-shaped. A random variable X following a normal distribution with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$ is written as $X \sim \mathcal{N}(\mu, \sigma^2)$ and has the PDF

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2 \cdot \sigma^2}}$$

The location parameter μ is called the mean and the scale parameter σ is called the standard deviation. It is easier to work with the quantity σ^2, which we call the variance. Below is a graph of several PDFs that have a unit variance $\sigma^2 = 1$ and various means μ.

Figure 25: Plots of PDFs of Normal Distributions with Various μ



The three plots show the PDFs of $\mathcal{N}(\mu, 1)$ for μ = -1 (blue), μ = 0 (orange), and μ = 2 (green).

Source: George Dekermenjian (2019).

For each of these PDFs, the density peaks at the mean (center) and quickly vanishes away from it. This means that for a normally distributed random variable, the values most likely to be observed are near the center. Values much larger or much lower than the center value are also less likely.

Consider one of these distributions, say $X \sim \mathcal{N}(0,1)$. We examine the probability of unit length intervals in the table below. Notice that as the interval moves away from the center (zero) the probability decreases significantly.

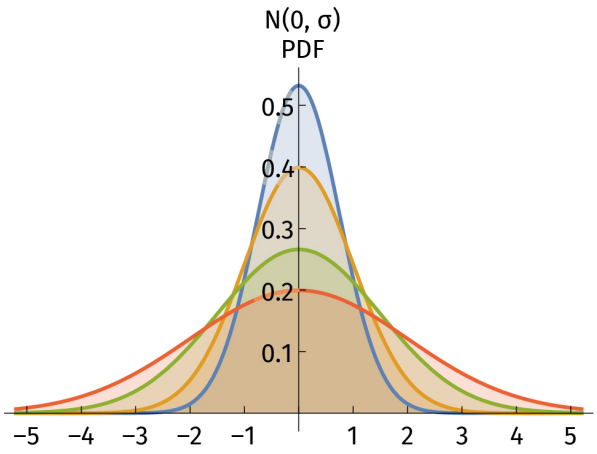
Table 7: Standard Normal Probabilities of Various Unit Intervals

Unit-length interval (a, b)	$P(a \leq X \leq b)$
(0,1)	0.3413
(0.5,1.5)	0.2417
(1,2)	0.1359
(1.5,2.5)	0.0606

Source: George Dekermenjian (2019).

Next, let us take a look at graphs of PDFs that have the same mean (center) at $\mu = 0$ but different scales σ.

Figure 26: Plots of PDFs of Normal Distributions with Various σ



The four plots show the PDFs of N(0, σ) for σ = 0.75 (blue), σ = 1 (orange), σ = 1.5 (green), and σ = 2 (red).

Source: George Dekermenjian (2019).

These PDFs show that the density is spread over a wider range of values around the center for larger values of σ and vice-versa for smaller values (a narrower range of values). In other words, if we consider the same interval across various scales, the distributions that are spread out more will give a smaller probability. To explore this idea, we evaluate the probability $P(0 \leq X_i \leq 1)$ for $X_i \sim \mathcal{N}(0, \sigma_i^2)$ for ${\sigma_1}^2 = 1$, ${\sigma_2}^2 = 1.5$, and ${\sigma_3}^2 = 2$.

Table 8: Normal Probabilities on [0,1] with Various σ

$X \sim \mathcal{N}(0, \sigma_i^2)$	$P(0 \leq X \leq 1)$
$X \sim \mathcal{N}(0,1)$	0.3413
$X \sim \mathcal{N}(0,1.5^2)$	0.2475
$X \sim \mathcal{N}(0,2^2)$	0.1915

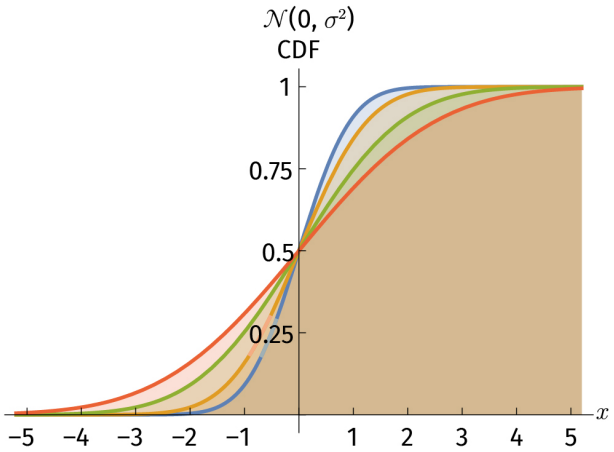
Source: George Dekermenjian (2019).

The CDF of a normally distributed random variable cannot be written in closed form, indeed the integral

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\mathrm{d}t = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\mathrm{x}} \mathrm{e}^{-\frac{(\mathrm{t}-\mu)^2}{2 \cdot \sigma^2}} \mathrm{d}t$$

is the best we can do. However, almost all programming packages you will encounter will have a built-in function to evaluate the CDF of normally distributed random variables. As such, it is not necessary to try to evaluate this integral directly. Below are graphs of CDFs of normally distributed random variables with mean 0 and varying σ.

Figure 27: Plot of CDFs of Normal Distributions with Various σ



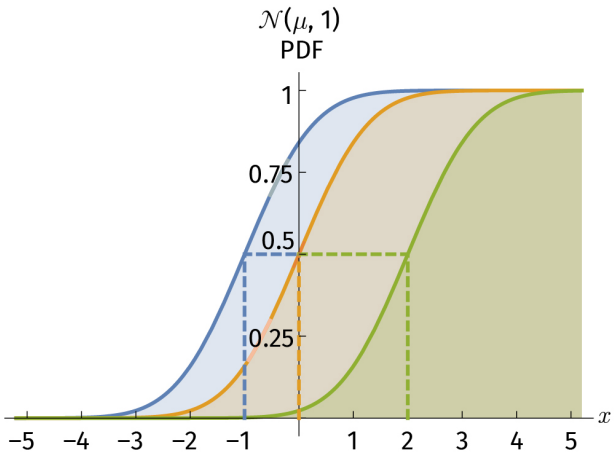
The four curves show the CDFs of $\mathcal{N}(0, \sigma^2)$ for σ^2 = 0.75 (blue), σ^2 = 1 (orange), σ^2 = 1.5 (green), and σ^2 = 2 (red).

Source: George Dekermenjian (2019).

Notice that for distributions with larger variance, it takes more time to accumulate the probability than with ones with smaller variance.

The following graph shows CDFs of normally distributed random variables with unit variance and various means μ.

Figure 28: Plot of CDFs of Normal Distributions with Various μ



The four curves show the CDFs of $\mathcal{N}(\mu, 1)$ for μ = –2 (blue), μ = 0 (orange), and μ = 1 (green).

Source: George Dekermenjian (2019).

For the CDFs with the same variance, notice that the shapes are identical, and the center just shifts the graph. Additionally, notice how each of the distributions accumulate 50% of the probability up to their respective means. This means that the mean of a normally distributed random variable is also its median. This is typical of symmetric distributions.

Among all the different normal distributions, one deserves special attention: the standard normal distribution. A random variable Z has a standard normal distribution if it has a normal distribution with mean 0 and standard deviation 1, which means $Z \sim \mathcal{N}(0,1)$. The PDF of Z is given by

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

and the CDF is given by

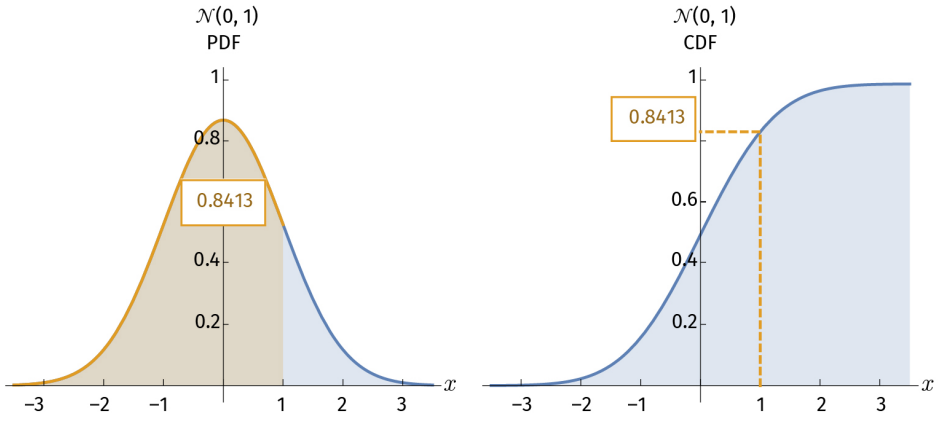
$$F_Z(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} \mathrm{d}t \,.$$

The PDF of the standard normal distribution is symmetric with respect to the vertical axis. This means that the area to the left of the center, zero, is exactly 1/2 , and the area to the right of zero is also 1/2 . We can exploit this symmetry to compute certain probabilities.

Probabilities can be computed by finding the area under the curve. For example,

$$P(Z < 1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{1} e^{-\frac{t^2}{2}} \mathrm{d}t = \Psi(1) \approx 0.8413.$$

Figure 29: Plots of the PDF and CDF of a Standard Normal Distribution (1)



Left: PDF showing the area shaded under the curve to the left of 1
Right: CDF showing the corresponding point on the graph

Source: George Dekermenjian (2019).

As shown in the figure above, the area up to 1 under the PDF curve is the same as the vertical coordinate of the point on the CDF graph.

We can also compute probabilities between two finite numbers. For example,

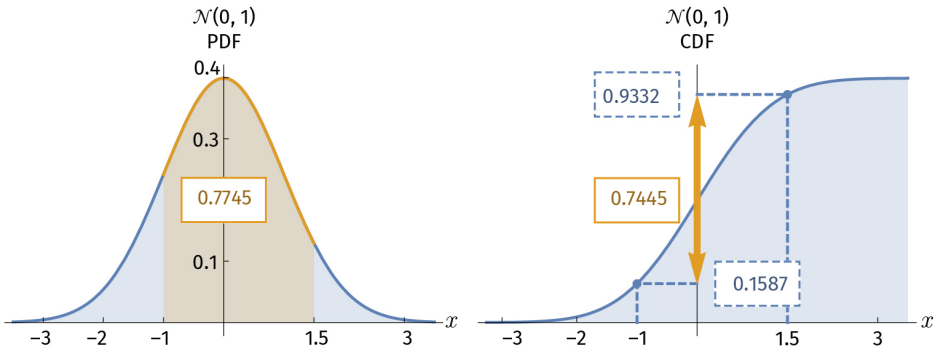
$$\begin{aligned} P(-1 \leq Z \leq 1.5) &= \frac{1}{\sqrt{2\pi}} \int_{-1}^{1.5} e^{-\frac{t^2}{2}} \mathrm{dt} = P(Z \leq 1.5) - P(-1 \leq Z) \\ &= P(Z \leq 1.5) - (1 - P(Z \leq 1)) \approx 0.9332 - 0.1587 = 0.7745. \end{aligned}$$

We note that the following integral property was used

$$\begin{aligned} P(a \leq Z) &= \frac{1}{\sqrt{2\pi}} \int_{a}^{\infty} e^{-\frac{t^2}{2}} \mathrm{dt} \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} \mathrm{dt} - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{t^2}{2}} \mathrm{dt} = 1 - P(Z \leq a), \end{aligned}$$

which holds for any $a \in \mathbb{R}$.

Figure 30: Plots of the PDF and CDF of a Standard Normal Distribution (2)



Left: PDF showing the area shaded under the curve to the left between −1 and 1.5
Right: CDF showing the corresponding points on the graph

Source: George Dekermenjian (2019).

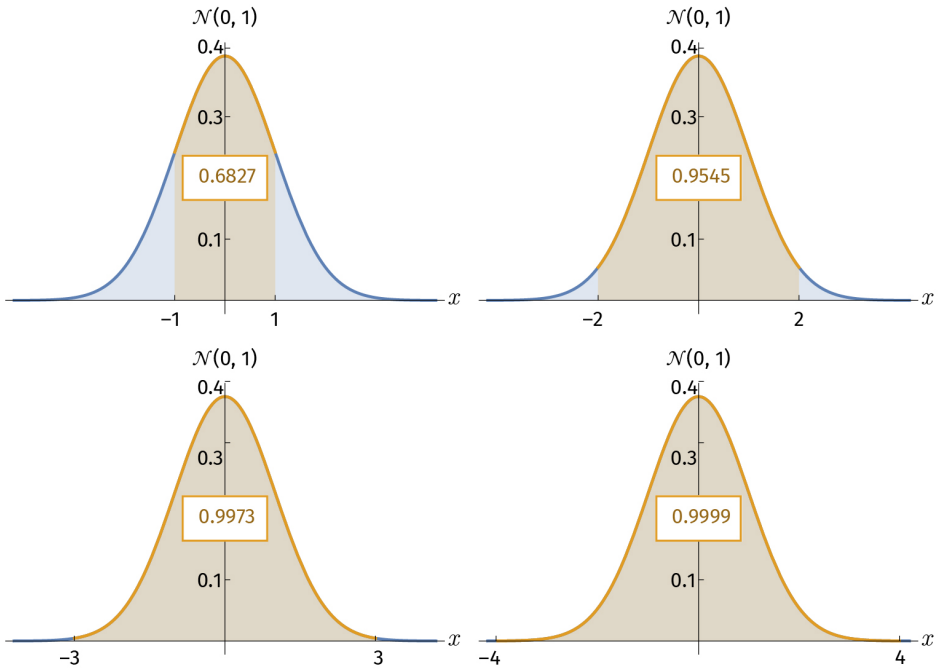
Let us now summarize probabilities on intervals symmetric around zero, the center (mean) of the distribution.

Table 9: Standard Normal Probabilities on Symmetric Intervals

$\boldsymbol{P(-1 \leq Z \leq 1) \approx 0.6827}$
$P(-2 \leq Z \leq 2) \approx 0.9545$
$P(-3 \leq Z \leq 3) \approx 0.9973$
$P(-4 \leq Z \leq 4) \approx 0.9999$

Source: George Dekermenjian (2019).

Figure 31: Areas Under the PDF of the Standard Normal Distribution



Top Left: The area between −1 and 1 is about 68.3%.
Top Right: The area between −2 and 2 is about 95.5%.
Bottom Left: The area between −3 and 3 is about 99.7%.
Bottom Right: The area between −4 and 4 is about 99.9%.

Source: George Dekermenjian (2019).

Example 2.10

Calculate the following probabilities.

a) $P(0 \leq Z \leq 2)$.
b) $P(Z \leq 2)$.
c) $P(Z > 2)$.
d) $P(Z < -2)$.

Solution 2.10

a) Using the previously mentioned fact that $P(-2 \leq Z \leq 2) \approx 0.9545$, together with the symmetry, we know that

$$P(0 \leq Z \leq 2) = \frac{1}{2}P(-2 \leq Z \leq 2) \approx \frac{0.9545}{2} \approx 0.4773$$

b) The event $Z \leq 2$ is the disjoint union of the events $Z < 0$ and $0 \leq Z \leq 2$. Therefore, by the sum property of probabilities, we have

$$P(Z \leq 2) = P(Z < 0) + P(0 \leq Z \leq 2) \approx 0.5 + 0.4773 \approx 0.9773.$$

c) The event $Z > 2$ is the complement of the event $Z \leq 2$. Therefore, by the complement rule we have

$$P(Z > 2) = 1 - P(Z \leq 2) \approx 1 - 0.9773 \approx 0.0227.$$

d) Using symmetry, we have

$$P(Z < -2) = P(Z > 2) \approx 0.0227.$$

The probabilities in parts c. and d. of Example 2.10 are called tail probabilities. Such quantities come up quite often in statistical inference. As such, here are some tail probabilities.

Table 10: Standard Normal Tail Probabilities

$P(Z > 2.576) \approx 0.0050$
$P(Z > 2.3264) \approx 0.0100$
$P(Z > 1.6449) \approx 0.0500$
$P(Z > 1.2816) \approx 0.1000$

Source: George Dekermenjian (2019).

We have devoted a substantial amount of time to discussing the standard normal distribution. What about all the other normal distributions? It turns out that computing with the standard normal distribution is enough because of the following fact: if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$.

In other words, if we want to compute probabilities with non-standard normal distributions, we can work with the related standard normal distribution by using the above transformation.

Example 2.11

It is believed that IQ is normally distributed with a mean of 100 and a standard deviation of 15. Compute the probability that a randomly chosen person has an IQ of between 85 and 115.

Solution 2.11

Let X denote the IQ of a randomly selected person. We are given that $X \sim \mathcal{N}\left(100, 15^2\right)$. Using the transformation above we have

$$P(85 \leq X \leq 115) = P\left(\frac{85-100}{15} \leq \frac{X-100}{15} \leq \frac{115-100}{15}\right),$$

if we set $Z = \frac{X-100}{15}$, then $Z \sim \mathcal{N}(0,1)$. Therefore, the above probability is the same as

$$P(-1 \leq Z \leq 1) \approx 0.6827\,.$$

The reverse transformation also works. Suppose $Z \sim \mathcal{N}(0,1)$, then the transformed random variable $X = \mu + \sigma Z$ is a normal distribution with mean μ and standard deviation σ, i.e., $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$.

Example 2.12

Continuing from the previous example, find the IQ score that separates the top 5% from the rest. (Hint: use the fact that $P(Z > 1.64485) \approx 0.0500$.)

Solution 2.12

Let X denote the IQ score of a randomly selected person. We want to find x_{95} such that $P\left(X \leq x_{95}\right) = 0.9500$. Following the hint, together with the complementary event, we know that

$$P(Z \leq 1.64485) = 1 - P(Z > 1.64485) \approx 0.9500\,.$$

Furthermore, the transformed random variable $100 + 15Z$ follows a normal distribution with a mean of 100 and a standard deviation of 15. Therefore, $X = 100 + 15Z$. Therefore,

$$0.9500 \approx P(100 + 15Z \leq 100 + 15(1.64485)) = P(X \leq 124.673)\,.$$

In other words, the 95th percentile of X is $x^{95} \approx 124.673$. In our context, an IQ score of 124.673 would be higher than 95% of the population (since IQ is defined as an integer, this number would be rounded up to 125).

Student's T Distribution

The Student's T distribution comes up in statistical inference when the sample size is not sufficiently large. It behaves very much like the standard normal distribution, but the tails are “heavier” or “thicker” than the standard normal distribution. The mean (center) of the T distribution is always 0, but the standard deviation changes with the degrees of freedom parameter, $v > 0$. If X follows a T distribution with v degrees of freedom, then we write $X \sim T(v)$. The PDF of X is given by

$$f_X(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \text{ for any } x \in \mathbb{R}$$

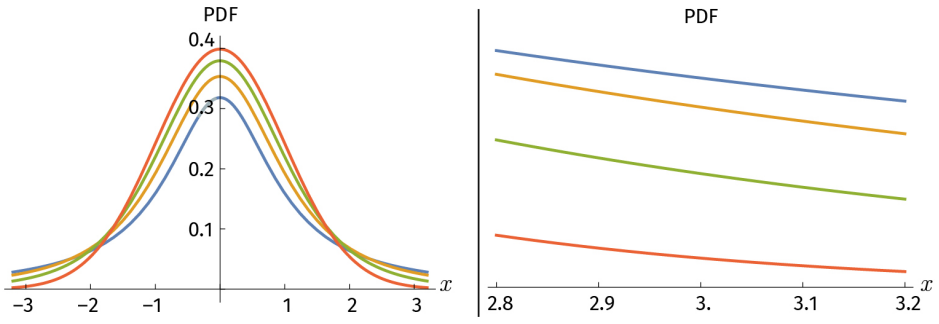
Here, $\Gamma(y)$ is called the gamma function; it is a generalization of the factorial for all real numbers. In fact, if $n \geq 0$ is an integer, then $n! = \Gamma(n+1)$.

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} \mathrm{d}t \text{ for any } x > 0\,.$$

(We will discuss the gamma distribution, which is closely related to the gamma function, later on.)

The graphs below show PDFs of random variables that follow the T distribution with different degrees of freedom. Additionally, the graph of the standard normal PDF is included as a reference for the comparison of the thickness in the tails.

Figure 32: Plots of PDFs of Various Student T Distributions Together with the Standard Normal Distribution

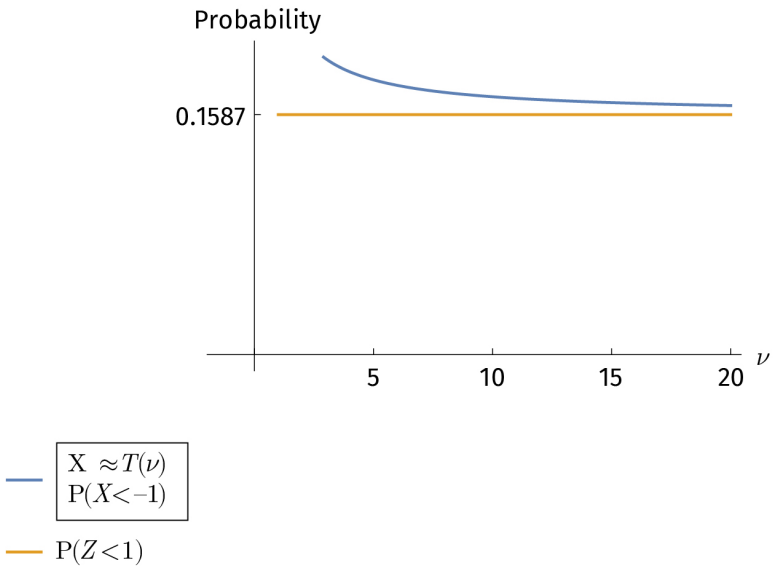


Left: The Plot of the PDFs from –3 to 3
Right: The Plots of the PDFs from 2.8 to 3.2 showing the tail behaviors for various ν

Source: George Dekermenjian (2019).

As you may have noticed, the larger the degrees of freedom, the closer the PDF is to the standard normal distribution. In statistical inference, the degrees of freedom are tied to the sample size. Therefore, if the sample size is large enough, the standard normal distribution is substituted in the calculations without much loss of accuracy. As a matter of fact, it can be shown in the limit that $T(v)$ for $v \rightarrow \infty$ yields $\mathcal{N}(0,1)$. To illustrate this fact, we compare the probability of $P(X <= -1)$ and $P(Z < -1)$ where $X \sim T(v)$ and $Z \sim \mathcal{N}(0,1)$. Notice that as the degrees of freedom get larger, the two probabilities get very close to one another.

Figure 33: Convergence of Probability from Student T to Standard Normal



Source: George Dekermenjian (2019).

The CDF of the T distribution does not have a simple closed form. However, most statistical software packages have an implementation of this CDF.

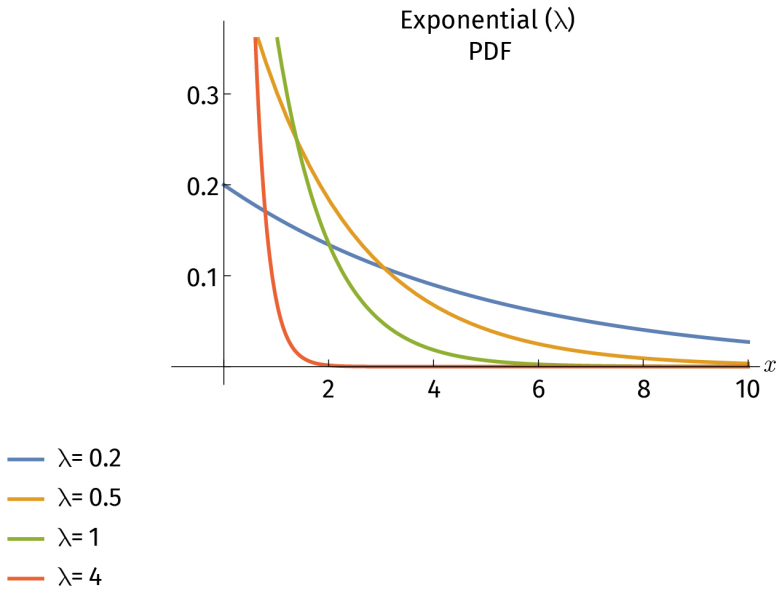
The Exponential Distribution

Exponential distributions are used, among other things, to model the interarrival times between two successive events where the number of events follows a Poisson distribution (see Kim, 2019a). If X is exponentially distributed at a rate of $\lambda > 0$, we write $X \sim \text{Exponential}(\lambda)$. Notice this distribution only has one parameter. We will see in the next section that λ is related to the mean of the distribution. The PDF is given by

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

The following graphic shows plots of PDFs of different exponential distributions:

Figure 34: Plots of PDFs of Various Exponential Distributions



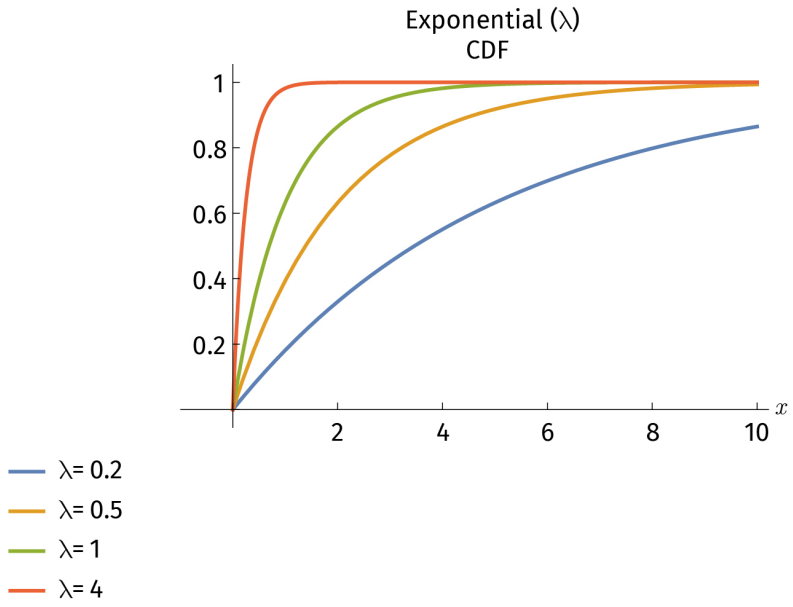
Source: George Dekermenjian (2019).

The CDF of an exponential distribution can be derived by simple integration:

$$F_x(x) = \int_{-\infty}^{x} f_{\mathrm{x}}(t)\mathrm{d}t = \int_{0}^{\mathrm{x}} \lambda e^{-\lambda \mathrm{t}} \mathrm{d}t = 1 - e^{-\lambda x} \text{ for any } x \geq 0,$$

and for $x < 0$ we have $F_{\mathrm{x}}(x) = 0$.

Figure 35: Plots of CDFs of Various Exponential Distributions



Source: George Dekermenjian (2019).

Example 2.13

A battery has a lifespan in hours that is exponentially distributed with a parameter rate of $\lambda = 1/2500$. Find the probability that a randomly selected battery will die out before 3000 hours.

Solution 2.13

Let X be the lifespan of a randomly selected battery. We know that $X \sim \text{Exponential}(1/2500)$. The probability that this battery dies out before 3000 hours is the same as saying that the lifespan of the battery is less than 3000. Therefore, we compute

$$P(X < 3000) = F_X(3000) = 1 - e^{-\frac{3000}{2500}} \approx 0.6988$$

Therefore, the probability is about 69.88%.

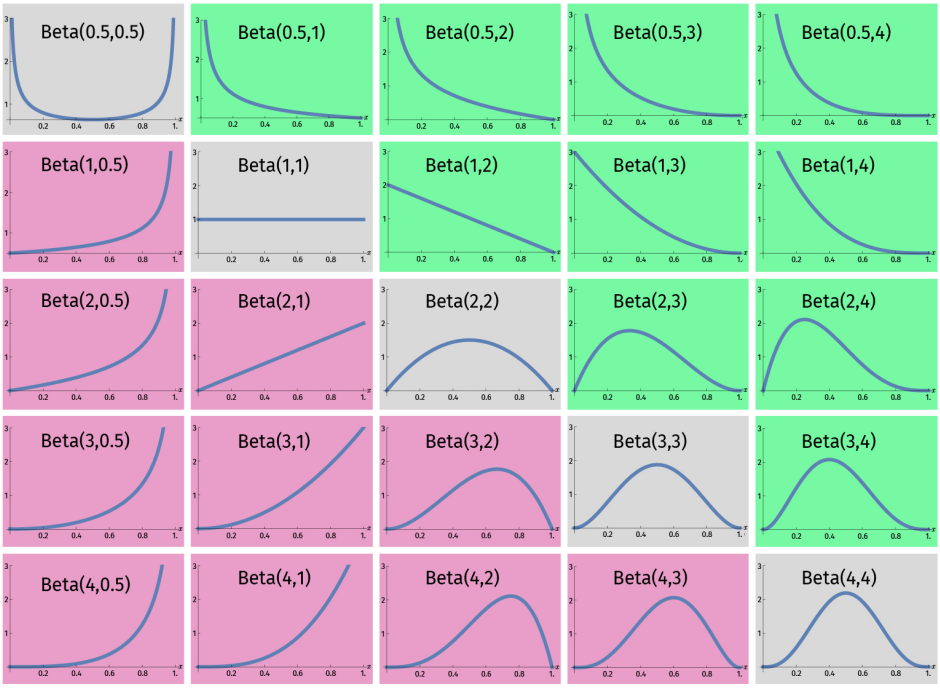
The Beta Distribution

The beta distribution can be used to model the behavior of a random variable whose range is a finite interval. The PDF of any beta distribution is supported on the closed interval $[0,1]$. In data science, the beta distribution comes up in Bayesian inference when we want to incorporate prior knowledge from data into the modeling of unknown parameter(s) of Bernoulli, binomial, geometric, and negative binomial distributions. Historically,

the reason for its popularity comes from the fact that the posterior is in the same distribution family as the prior when using the beta distribution. A well-known application of the beta distribution in education is to model the true test score for students, see (Sinharay, 2010). The beta distribution has two parameters, $\alpha > 0$ and $\beta > 0$. Both of these parameters are interpreted as shape parameters. If the random variable X follows a beta distribution with parameters α and β, we write $X \sim \mathrm{Beta}(\alpha, \beta)$. The PDF is given by

$$f_X(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} & \text{for } 0 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Figure 36: Plots of PDFs of Various Beta Distributions



Below Diagonal (Pink) have α > β and are left skewed. Diagonal (Gray) have α = β and are symmetric. Above Diagonal (Green) have α < β and are right skewed.

Source: George Dekermenjian (2019).

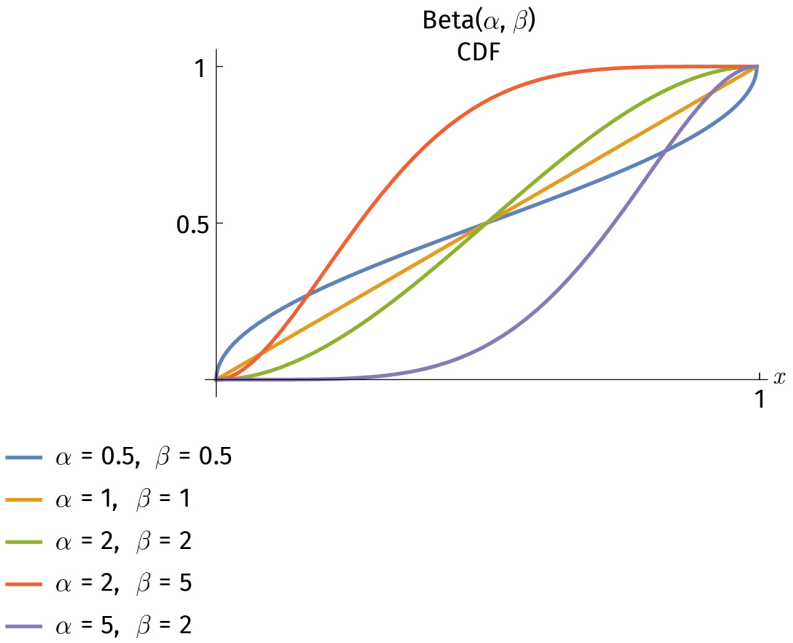
As you can see in the figure above, the beta family of distributions is quite diverse. One member of the family even reduces to the uniform distribution (with parameters $\alpha = \beta = 1$). Some of the PDFs have a maximum, while others have a minimum. Some PDFs are symmetric with respect to $x = 1/2$, while others are not.

Consider the following example, a coin, possibly biased, is tossed four times and the result is heads on all four tosses. We want to compute the probability that the next toss will result in tails. With the Bayesian approach, one of the ways we can encode our observa-

tions into modeling the probability of heads is as a Beta $(4+1,0+1)$ distribution, where the 4 and 0 correspond to 4 heads and 0 tails respectively. Our prior expectation would look similar to the graphs on the lower left panel in the above figure.

The CDF of the beta distribution does not have a simple closed form, but it is implemented in most statistical software packages. Below are some graphs of CDFs for various parameters.

Figure 37: Plots of CDFs of Various Beta Distributions



Source: George Dekermenjian (2019).

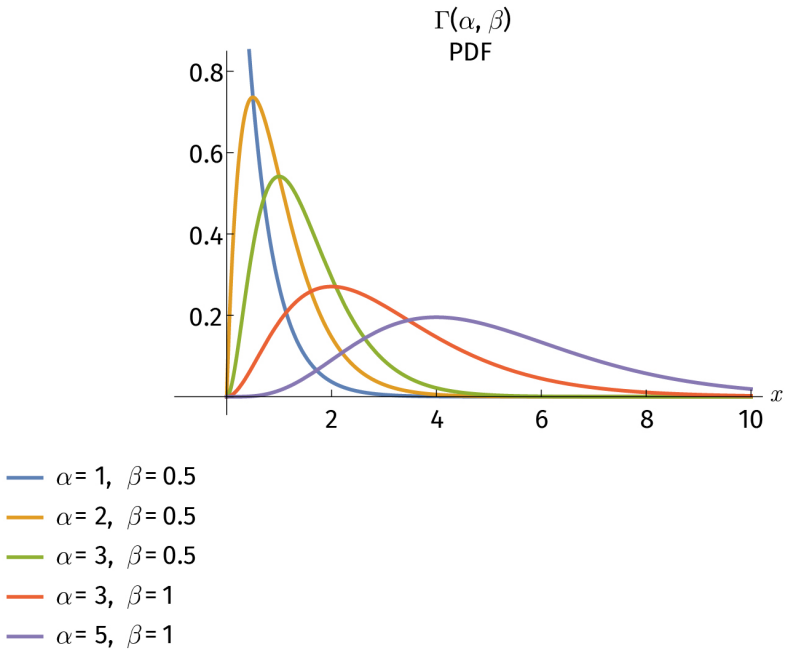
The Gamma Distribution

The gamma distribution can be used in data science to estimate the precision (inverse variance) of a normal distribution by incorporating prior knowledge. It is also used to estimate the rate parameter of an exponential distribution (see Kim, 2019b). The gamma distribution can be applied in order to model the size of insurance claims or to predict the amount of rainfall. The gamma distribution has two parameters, a shape parameter $\alpha > 0$, and an (inverse) scale parameter $\beta > 0$. If a random variable X follows the gamma distribution with these two parameters, we write $X \sim \text{Gamma}(\alpha, \beta) = \Gamma(\alpha, \beta)$. The PDF of such an X is given by

$$f_X(x) = \begin{cases} \frac{\beta^a}{\Gamma(\alpha)} x^{a-1} e^{-\beta x} & x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Notice that when $\alpha = 1$, the gamma distribution reduces to the exponential distribution with rate $\lambda = \beta$. Therefore, the exponential distribution is a special case of the gamma distribution.

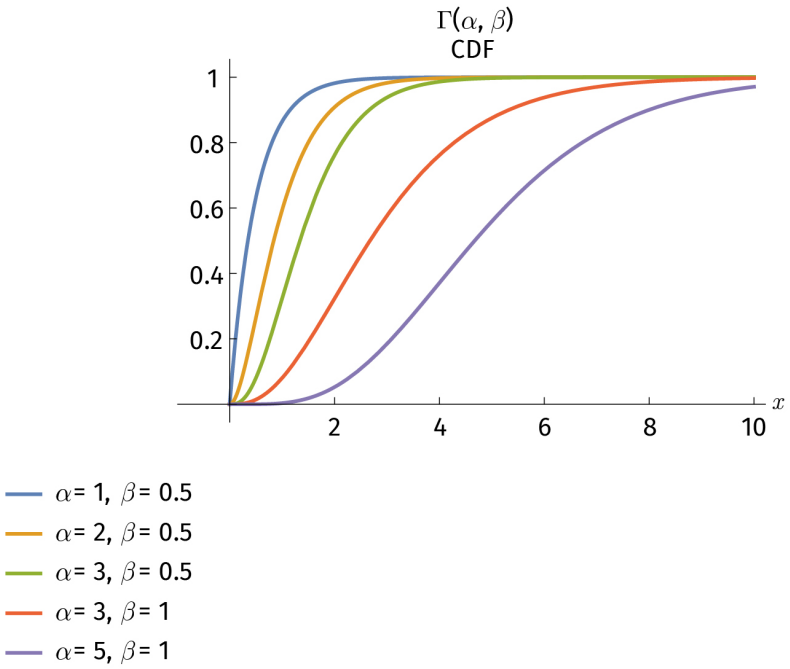
Figure 38: Plots of PDFs of Various Gamma Distributions



Source: George Dekermenjian (2019).

The CDF of the gamma distribution does not have a simple closed form, but like the CDF of the beta distribution it is implemented in most statistical software packages. Here are some graphs with various parameters, which shows that for, increasing α and β, the tail of the CDF tends to the right-hand side, meaning that outcomes far from the origin have a higher likelihood to occur than for smaller α and β.

Figure 39: Plots of CDFs of Various Gamma Distributions



Source: George Dekermenjian (2019).

Weibull Distribution

In Example 2.13, we modeled the time to failure of a battery as an exponential distribution. In that setting, the failure rate was constant. In some applications, it makes sense to assume that the failure rate may decrease or increase over time. In such cases, modeling the time to failure with an exponential distribution is no longer valid. Therein lies the need for another distribution-the Weibull distribution. Among other things, the Weibull distribution is well suited to modeling the failure time for a system when the rate of failure may vary over time. The Weibull distribution is parameterized by two parameters: the shape parameter, $k > 0$, and the scale parameter, $\lambda > 0$. We say the random variable X follows a Weibull distribution with these parameters and write $X \sim \text{Weibull}(k, \lambda)$ if its PDF is given by

$$f_x(x) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} \mathrm{e}^{-\left(\frac{x}{\lambda}\right)} & x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

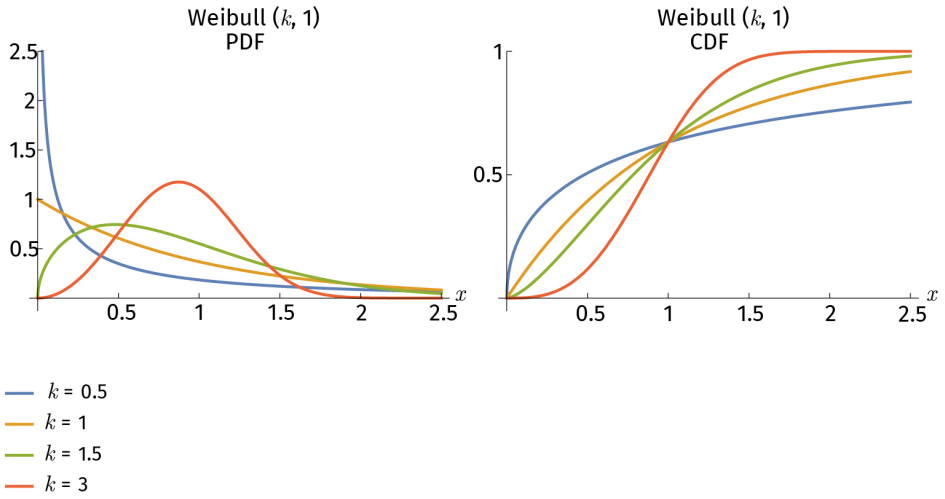
If $k = 1$, then the PDF for $x \geq 0$ reduces to which is exactly the PDF of the exponential distribution with rate $1/\lambda$. Thus, as discussed above, the Weibull distribution is a generalization of the exponential distribution.

The corresponding CDF is

$$F_X(x) = \begin{cases} 1 - \mathrm{e}^{-\left(\frac{x}{\lambda}\right)^k} & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Let us examine the cases of $k > 1$ and $k < 1$ with graphs of PDFs and CDFs.

Figure 40: Plots of PDFs and CDFs of Various Weibull Distributions



Left: PDFs of Weibull distribution with scale 1 and various shape parameters
Right: Corresponding CDFs

Source: George Dekermenjian (2019).

Example 2.14

The lifetime of an electric motor in years has a Weibull distribution of $\lambda = 3$ and $k = 5$. Find a guarantee time, g, such that 95% of the motors last more than g years.

Solution 2.14

In other words, we need to solve the following equation

$$P(X > g) = 0.95,$$

where X is the random variable that describes the lifetime of the electric motor. We use the PDF with $\lambda = 3$ and $k = \quad 5$ and obtain the following problem

$$\int_{\mathrm{g}}^{\infty} \frac{5}{3}\left(\frac{\mathrm{x}}{3}\right)^4 \mathrm{e}^{-\left(\frac{\mathrm{x}}{3}\right)^5} \, \mathrm{d}x > 0.95\,.$$

We compute the integral in with the chain rule and obtain

$$\int_g^\infty \tfrac{5}{3}\left(\tfrac{x}{3}\right)^4 \mathrm{e}^{-\left(\frac{x}{3}\right)^5}\, dx = \left[-\mathrm{e}^{-\left(\frac{x}{3}\right)^5}\right]_g^\infty = \mathrm{e}^{-\left(\frac{g}{3}\right)^5} > 0.95\,.$$

First of all, we use the reciprocal to obtain

$$\mathrm{e}^{\left(\frac{g}{3}\right)^5} < \tfrac{1}{0.95}\,.$$

Then we use the natural logarithm and obtain

$$\left(\tfrac{g}{3}\right)^5 < ln\left(\tfrac{1}{0.95}\right).$$

Taking the fifth root yields

$$g < 3 \cdot \sqrt[5]{ln\left(\tfrac{1}{0.95}\right)} \approx 1{,}65628$$

This means that there is a likelihood of at least 95% that the electric motor will still work after 1.65 years.

SUMMARY

In this unit, we introduced the notion of a random variable and discussed the transition of probability questions from the events of a sample space to that of random variables. A random variable X is a function that maps outcomes of a sample space to values on the real line.

For discrete random variables, the image of a random variable is finite or a countable infinite set. For continuous random variables, the image is a finite or an infinite interval of the real line; in some cases, it is the whole real line. We then further developed this connection, learning that, to compute probabilities related to random variables, it is instructive to look at the inverse mapping. This takes us from a value (or values) of a random variable back to the sample space as events.

We defined probability mass functions (PMFs) for discrete random variables, probability density functions (PDFs) for continuous random variables and discussed their relationship to the (cumulative) distribution function $F_X(x) = P(X \leq x)$.

The important discrete probability distributions considered in this unit were the Bernoulli, binomial, geometric, negative binomial and Poisson distribution.

The important continuous probability distributions discussed were the uniform, normal, Student's T, exponential, beta, gamma, and Weibull distributions. We defined the PDFs for all of these distributions and, for some, gave an explicit formula for the CDF.

We looked at the PMF and CDF of these distributions and saw some examples of how to apply some of them to real-world problems. We discussed how some of these distributions have one parameter (Bernoulli, geometric, and Poisson) while others had two parameters (binomial and negative binomial).

For the normal distribution, we introduced the special standard normal distribution $\mathcal{N}(0,1)$ and explained how any normally distributed random variables can be transformed into the standard normal distribution, meaning that if $X \sim \mathcal{N}(\mu, \sigma^2)$ for $\mu \in \mathbb{R}$ and $\sigma > 0$, then for the random variable $\frac{X-\mu}{\sigma} = Z$ holds $Z \sim \mathcal{N}(0,1)$. Vice versa if $X \sim \mathcal{N}(0,1)$ the for the random variable $Z \mapsto \mu + \sigma Z$ holds $Z \sim \mathcal{N}(\mu, \sigma^2)$.

It is important to note that when writing down the PDF of any continuous distribution, we must take great care to mention the support, which basically consists of all positive values of the PDF. For example, the support of the PDF of the beta distribution is the interval $[0,1]$, and the support for the gamma distribution is $(0, \infty)$.

UNIT 3

JOINT DISTRIBUTIONS

STUDY GOALS

On completion of this unit, you will be able to ...

- define and work with joint PMFs and PDFs.
- understand the meaning of independent random variables.
- determine whether or not random variables are independent based on their joint PMF or PDF.
- understand the meaning of marginal distributions.
- derive marginal PMFs (PDFs) from the joint PMF (PDF).
- derive the conditional PMF (PDF) given the joint PMF (PDF) and the appropriate marginal.

3. JOINT DISTRIBUTIONS

Introduction

Joint probability mass (density) function
This is a function of several variables that gives the probability (density) of values of tuples of random variables.

More often than not, real-world data has more than one dimension. To analyze such data, we need a way to model the probability of multidimensional data. To that end, this unit introduces joint probability distributions. You will study these distributions through **joint Probability Mass (Density) Functions** for discrete random variables and joint PDFs for continuous random variables. We can also quantify relationships between variables using other tools such as marginal and conditional densities. The concepts of expectation and variance are used to quantify the accumulation of the density (probability) and the dispersion around this central value respectively. In this unit, these concepts will be extended to multidimensional distributions. Additionally, the concept of covariance will be introduced. This will allow you to quantify how changes in one variable affect those in another variable.

3.1 Joint Distributions

Joint probability measure
This is a probability measure for a tuple (pair, triple, etc.) of random variables.

Joint Cumulative Distribution Function (CDF)
This is a function of several variables that gives the accumulated probability across a range of values for tuples of random variables.

Given two random variables X and Y, their **joint probability measure** gives the probability that the values of (X, Y) are in a set of values. In the case of discrete random variables, these sets of values are a countably infinite (or finite) set, and in the case of continuous random variables, the set of values is a continuous range of real numbers. As with univariate distributions, there are two ways to specify a joint probability distribution (i) by the joint PMF or (ii) by the **joint Cumulative Distribution Function** (CDF).

First of all, we need the cross-product of two given sample spaces Ω_1 and Ω_2, which is defined as

$$\Omega_1 \times \Omega_2 = \{(x, y) \,|\, x \in \Omega_1 \text{ and } y \in \Omega_2\},$$

which consists of all pairs of points of the form (x, y) such that $x \in \Omega_1$ and $y \in \Omega_2$. It must be stressed that we can define probability measures on $\Omega_1 x \Omega_2$ as well as on Ω_1 or on Ω_2. If we consider $\Omega_1 = \{1,2,3\}$ and $\Omega_2 = \{a, b\}$, for fixed letters a and b, the cross-product is then

$$\Omega_1 \times \Omega_2 = \ \{(1, a), (1, b),\ (2, a), (2, b), (3, a), (3, b)\},$$

and as an example we consider the following probability measure P, with $P(x) = \frac{1}{6}$ for each $x \in \Omega_1 \times \Omega_2$. (Please note: x is a pair of values, since it is contained in $\Omega_1 \times \Omega_2$)

Now we can introduce the discrete case of a PMF for two random variables with common PMFs. For the sample spaces Ω_1, Ω_2, equipped with the probability measure P on $\Omega_1 \times \Omega_2$, we consider the discrete random variables $X: \Omega_1 \to \{x_1, x_2, x_3, \dots .\} \subseteq \mathbb{R}$ and $Y: \Omega_2 \to \{y_1, y_2, y_3, \dots .\} \subseteq \mathbb{R}$. We consider a function

$f: \{x_1, x_2, x_3, ...\} \times \{y_1, y_2, y_3, \dots .\} . \to [0,1]$.

If the function f satisfies the following properties

- $P((\omega_1, \omega_2) \in \Omega_1 \times \Omega_2 | X(\omega_1) = i, Y(\omega_2) = k\ ) = P(X = i,\ Y = k) = f\big(x_{i,k}\big)$ for all $i, k \in \mathbb{N}$,
- $\sum_{i,k=1}^{\infty} f\big(x_{i,k}\big) = 1$,

then f, which may be also denote by $f_{X,Y}$ is called a joint PMF of X and Y**.** The support of f consists of all points such that

$$f\big(x_{i,k}\big) > 0 \text{ for } i, k \in \ \mathbb{N}\,.$$

The corresponding CDF $F_{X,Y}$ is then given by

$$F_{X,Y}(x, y) = \sum_{\{j | y_j \le \lfloor y \rfloor\}} \sum_{\{i | x_i \le \lfloor x \rfloor\}} f_{X,Y}\big(x_i, y_j\big)$$

for fixed $x \in \Omega_1,\ y \in \Omega_2$.

For the (piecewise) continuous case, we consider the sample spaces Ω_1, Ω_2, equipped with the probability measure P on $\Omega_1 \times \Omega_2$ and the (piecewise) continuous random variables $X: \Omega_1 \to \mathbb{R}$ and $Y: \Omega_2 \to \mathbb{R}$. If a piecewise continuous function f exists : $\mathbb{R}x\mathbb{R} \to [0,1]$, satisfying the following properties

- $P((\omega_1, \omega_2) \in \Omega_1 \times \Omega_2 | X(\omega_1) \le x, Y(\omega_2) \le y) = P(X \le x,\ Y \le y)$
  $= \int_{-\infty}^{x} \int_{-\infty}^{y} f(t, s) \mathrm{d}s\ \mathrm{d}t = \int_{-\infty}^{y} \int_{-\infty}^{x} f(t, s) \mathrm{d}t \quad \mathrm{d}s = \text{for all } x, y \in \mathbb{R},$
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(t, s) \mathrm{d}s \quad dt = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(t, s) \mathrm{d}t \quad \mathrm{d}s = 1$,

then the function f, which may be also denoted by $f_{X,Y}$,is called a joint PDF of X and Y. The support of f consists of all points such that

$$f(x, y) > 0 \quad x, y \in \mathbb{R}\,.$$

It is obvious that the joint CDF $F_{X,Y}$ is given by

$$F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(t, s) \mathrm{ds}\ \mathrm{dt} = \int_{-\infty}^{y} \int_{-\infty}^{x} f(t, s) \mathrm{dt}\ \mathrm{ds} = y \text{ for all x,} y \in \mathbb{R}.$$

We recall that the one-dimensional integral measures the area below a given non-negative function and the x-axis. In this case the integral

$$\int_{-\infty}^{x} \int_{-\infty}^{y} f(t, s) \mathrm{d}s\ \mathrm{d}t = \int_{-\infty}^{y} \int_{-\infty}^{x} f(t, s) \mathrm{d}t\ \mathrm{d}s$$

measures the volume below the function f and the the square $(-\infty, x]x(-\infty, y]$ for fixed $x, y \in \mathbb{R}$. The integration order does not matter for the value of the integral, which can be proven with technical mathematical tools (for example, Fubini's Theorem).

Example 3.1

A fair die is rolled, with the random variable $X = 1$ for the event $\{1,2\}$ and zero otherwise and the random variable $Y = 1$ for the event of an odd result $\{1,3,5\}$ and zero otherwise.

a) Write down the table of all possible values of the pairs $(X, Y = x, y)$ that come from mapping each outcome of the roll.
b) Write down the PMF as a two-way table.

Solution 3.1

a) The outcome of the die is from the sample space $\Omega_1 = \Omega_2 = \{1,2,3,4,5,6\}$. Outcome 1 makes both X and Y map to 1. Outcome 2 puts $X = 1$ and $Y = 0$. Continuing in this way, we arrive at the following table that maps each outcome of the die to the values of X and Y.

Table 11: The Mapping of Outcomes to Values of (*X*,*Y*) from Rolling a Die

Outcome	(x, y)
1	(0,0)
2	(1,0)
3	(0,1)
4	(1,1)

Source: George Dekermenjian (2019).

b) To write down the PMF, it is helpful to first write down the inverse mapping from the value of (X, Y) to the corresponding event. We do this in the following table.

Table 12: The Inverse Mapping from Values of (*X*,*Y*) to Events from Rolling a Die

(x, y)	Event
(0,0)	$\{3,4,5,6\} \times \{2,4,6\}$
(0,1)	$\{3,4,5,6\} \times \{1,3,5\}$
(1,0)	$\{1,2\} \times \{2,4,6\}$
(1,1)	$\{1,2\} \times \{1,3,5\}$

Source: George Dekermenjian (2019).

Now we are ready to compute the probabilities and write down the PMF as a table.

Table 13: The Joint PMF of (X,Y) from Rolling a Die

(x, y)	$f_{X,Y}(x, y)$
(0,0)	$(4 \cdot 3))/36 = 1/3$
(0,1)	$(4 \cdot 3)/36 = 1/3$
(1,0)	$(2 \cdot 3)/36 = 1/6$
(1,1)	$(2 \cdot 3)/36 = 1/6$

Source: George Dekermenjian (2019).

Sometimes a discrete joint PMF is given as a two-way table. This form of the PMF will be useful when we want to calculate the marginal probability mass function, which will be the subject of the next section.

Table 14: The Joint PMF of (X, Y) from Rolling a Die as a Two-Way Table

		Y	
		0	1
X	0	1/3	1/3
	1	1/6	1/6

Source: George Dekermenjian (2019).

Example 3.2

An urn contains three green (G) and six blue (B) marbles. Two marbles are drawn, one after the other, without replacement. Let the random variables X and Y represent the number of green and blue marbles respectively.

a) Write down all possible value pairs (x, y) that (X, Y) can take by mapping the sample space of outcomes.
b) Write down the joint PMF of (X, Y) by first writing down the inverse mapping from all possible pairs of values (x, y) from (X, Y) to events in the sample space.

Solution 3.2

a) We begin by writing the sample space of this experiment $\Omega = \{GG, GB, BG, BB\}$.
The outcome GG contains 2 green and 0 blue; therefore, the value of (X, Y) corresponding to this outcome is (2,0). Similarly, we compute all possible values of (X, Y) and summarize them in the following table:

Table 15: The Mapping of Outcomes to Values of (X,Y) for Drawing Marbles

Outcome	(x, y)
GG	(2,0)
GB	(1,1)
BG	(1,1)
BB	(0,2)

Source: George Dekermenjian (2019).

b) Using the result from part a., we can write down the events corresponding to each of the three distinct value pairs. The following table summarizes the inverse mapping.

Table 16: The Inverse Mapping from Values of (X,Y) to Events for Drawing Marbles

(x, y)	Event
(2,0)	$\{GG\}$
(1,1)	$\{GB, BG\}$
(0,2)	$\{BB\}$

Source: George Dekermenjian (2019).

Unlike in the previous example, the outcomes of the sample space are not equally likely. Therefore, we need to calculate the probabilities of each outcome using the rules of probability. Note that since we are sampling without replacement, the two draws are not independent. Therefore, we will use conditional probabilities to compute the probabilities.

$$
\begin{aligned}
P(GG) &= P(\text{1st } G \text{ and 2nd } G) = P(\text{1st } G) \cdot P(\text{2nd G} \,|\, \text{1st}) \; G = \frac{3}{9} \cdot \frac{2}{8} = \frac{6}{72}, \\
P(GB) &= P(\text{1st } G \text{ and 2nd } B) = P(\text{1st } G) \cdot P(\text{2nd } B \,|\, \text{1st } G) = \frac{3}{9} \cdot \frac{6}{8} = \frac{18}{72}, \\
P(BG) &= P(\text{1st } B \text{ and 2nd } G) = P(\text{1st } B) \cdot P(\text{2nd } G \,|\, \text{1st } B) = \frac{6}{9} \cdot \frac{3}{8} = \frac{18}{72}, \\
P(BB) &= P(\text{1st } B \text{ and 2nd } B) = P(\text{1st } B) \cdot P(\text{2nd } B \,|\, \text{1st } B) = \frac{6}{9} \cdot \frac{5}{8} = \frac{30}{72}.
\end{aligned}
$$

Finally, $P(\{GB,\ BG\}) = P(GB) = \frac{36}{72}$.

We are now ready to write down the joint PMF:

Table 17: The Joint PMF of (*X,Y*) from Drawing Marbles

(x, y)	$f_{X,Y}(x, y)$
(2,0)	1/12

(x, y)	$f_{X,Y}(x, y)$
(1,1)	½
(0,2)	5/12

Source: George Dekermenjian (2019).

Here is the same PMF as a two-way table.

Table 18: The Joint PMF of (X,Y) from Drawing Marbles as a Two-Way Table

		X		
		0	1	2
Y	0	0	0	1/12
	1	0	1/2	0
	2	5/12	0	0

Source: George Dekermenjian (2019).

The Multivariate Hyper-Geometric Distribution

The distribution from Example 3.2 is a special case of the multivariate hyper-geometric distribution. In that example we had two random variables, X and Y, corresponding to two colors. Now we will look at the general scenario that leads us to the multivariate hyper-geometric distribution.

Suppose an urn contains marbles that are one of K colors ($K \geq 2$ is an integer). Let n_k denote the number of marbles of the k^{th} color where $k = 1,2, ..., K$. We draw n marbles at random, one after the other, without replacement. Let X_k count the number of marbles of the k^{th} color in the draw for $k = 1,2, ..., K$. In other words, X_1 is the number of marbles that have the first color and X_2 is the number of marbles in the draw of the second color, etc. We have K random variables.

We say that the joint distribution of $(X_1, X_2,..., X_K)$ follows a multivariate hyper-geometric distribution with parameters n and $(n_1, n_2, ..., n_k)$. Note that the only possible k-**tuples** are $(x_1, x_2, ..., x_K)$ such that $x_1 + x_2 + ... \quad x_K = n$ and$0 \leq x_k \leq min \quad (n, n \quad k)$ for $k = 1,2, ... \quad K$. The joint PMF is given by

Tuples
A 2-tuple is a pair like (x_1, x_2). A 3-tuple is a triple like (x_1, x_2, x_3).

$$f_{X_1, \ \ldots, \ X_K}(x_1, \ \ \ldots, \ x_K) = \frac{\binom{n_1}{x_1} \cdots \binom{n_K}{x_K}}{\binom{n_1 + \ \ldots \ + n_K}{n}}$$

if $x_1 + \quad ... \quad + x_K = n$ and $0 \leq x_K \leq n$ and 0 otherwise.

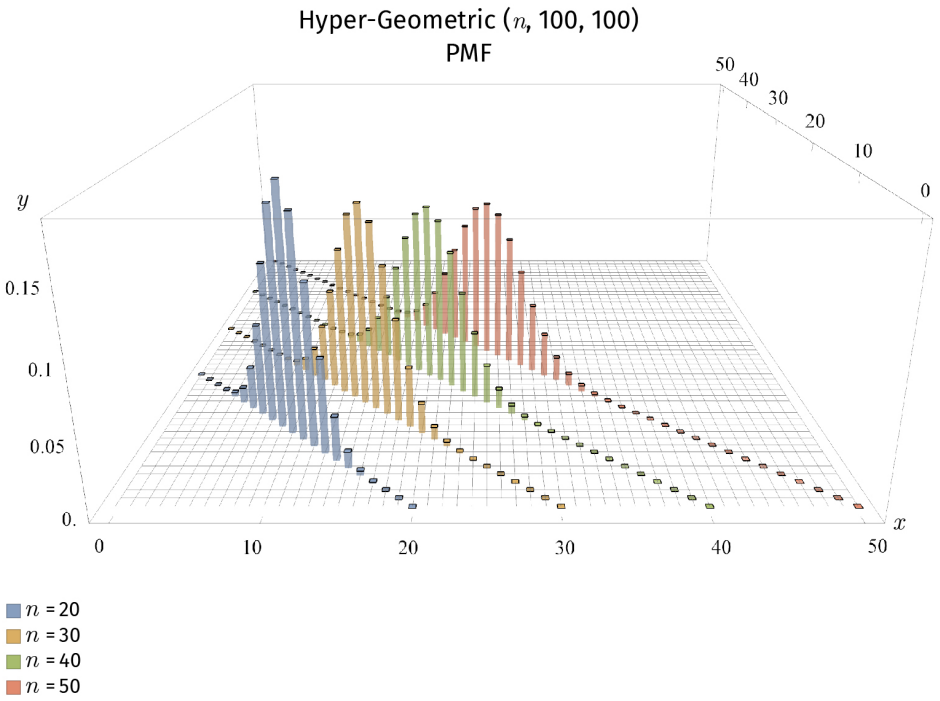
Let us explore this PMF by revisiting Example 3.2. Recall that in that example we had two colors (1 green and 2 blue); in other words, we had $K = 2$. We drew two marbles, so we set $n = 2$, and there were $n_1 = 3$ green and $n_2 = 6$ blue marbles. Therefore, the parameters of the hyper-geometric distribution were $n = 2$ and $(n_1, n_2) = (3,6)$. The PMF is therefore given by

$$f_{X_1,\ \ X_2}(x_1,\ \ x_2) = \frac{\binom{3}{x_1}\binom{6}{x_2}}{\binom{9}{2}}$$

for $x_1 + x_2 = 2$ and $0 \leq x_1, x_2 \leq 2$. The possible pairs are therefore (2,0), (1,1), and (0,2). Evaluating the PMF at these pairs will give you exactly the same values that we found in Solution 3.2.

The figure below shows the PMFs of various multivariate hyper-geometric distributions.

Figure 41: Plots of PMFs of Various Multivariate Hyper-Geometric Distributions



Source: George Dekermenjian (2019).

The Uniform Distribution

The support of the univariate distribution is a closed interval $[a, b]$ for fixed $a, b \in \mathbb{R}$ and the PDF assumes a constant value on this support equal to the reciprocal of the length of this interval: $1/b - a$. The multivariate uniform distribution is analogously defined over

generalizations of intervals. We will study the bivariate case. If the random pair (X, Y) is uniformly distributed over the rectangle $[a, b] \cdot [c, d]$ for fixed $a, b, c, d \in \mathbb{R}$, then its joint PDF is given by

$$f_{X,Y}(x, y) = \frac{1}{(b-a)(d-c)} \text{ if } (x, y) \in [a, b] \cdot [c, d]$$

and zero otherwise. In comparison to the univariate case, the value of the PDF over its support is the reciprocal of the measure of the support. This measure is the length of the interval $[a, b]$ and the area of the rectangle in the bivariate case. In the case of three variables, the support would be a box, with its measure just the volume of this box.

Indeed, the PDF of the bivariate case given above satisfies the properties of PDFs. In particular,

$$\int_c^d \int_a^b \frac{1}{(b-a)(d-c)} \mathrm{d}x \mathrm{d}y = \frac{1}{(b-a)(d-c)} \cdot (d-c)(b-a) = 1 \,.$$

The joint PDF can now be used to compute probabilities that the values of (X, Y) are in a region within this support. The simplest case is when the subregion is a rectangular region. Suppose the random pair (X, Y) is jointly uniform on the rectangle $[0,4] \times [0,2]$, meaning that

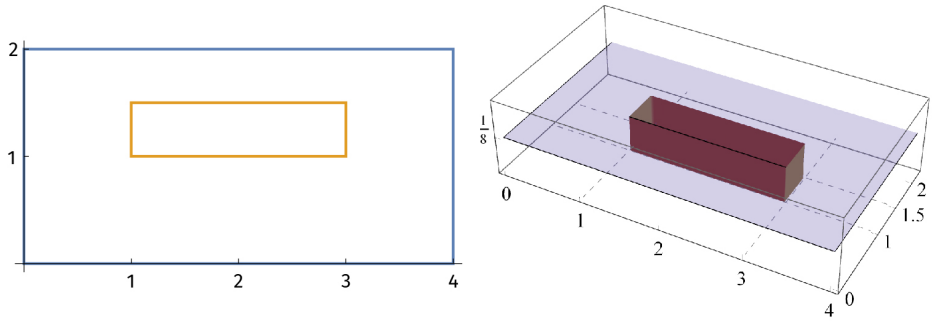
$$f_{X,Y}(x, y) = \frac{1}{8} \text{ for } (x, y) \in [0,4] \times [0,2]$$

and zero otherwise. We will compute the probability that (X, Y) falls in the rectangle $S = [1,3] \cdot [1,1.5]$, which is given by

$$P(1 \leq X \leq 3,\ 1 \leq Y \leq 1.5) = \int_1^3 \int_1^{1.5} \frac{1}{8} \mathrm{d}y \mathrm{d}x = 0.125$$

In practice, it is helpful to sketch the support of the PDF together with the region of interest (see the figure below).

Figure 42: Support and PDF of a Bivariate Uniform Distribution Over a Rectangular Region

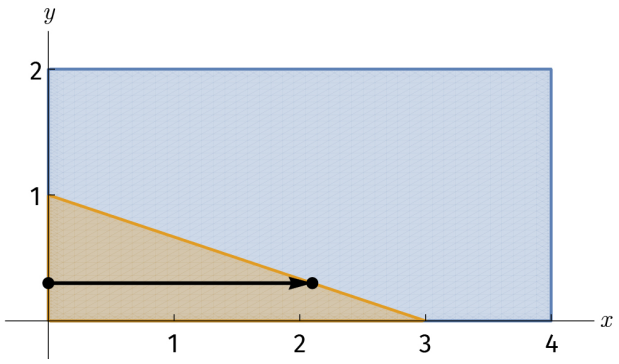


Left: The support of the PDF (blue rectangle) and the region of interest (orange rectangle)
Right: The surface plot of the PDF and the volume of the area under the graph over the region of interest

Source: George Dekermenjian (2019).

Sometimes we need to compute probabilities that (X, Y) falls in a non-rectangular region. Consider the region $S = \{(x, y) \,|\, x + 3y \leq 3\}$. A quick sketch reveals that this is a triangular region with vertices at the points (0,0), (0,1), and (3,0). As this is a simple region, one can parametrize either of the variables x or y. We choose to parametrize x so that $0 \leq x \leq 3 - 3y$, while we let *y* vary in $0 \leq y \leq 1$.

Figure 43: Rectangular Support and Triangular Region of Interest



The support of the PDF (blue rectangle) and the region of interest (orange rectangle)

Source: George Dekermenjian (2019).

$$P(X+3Y\leq 3)=\int_0^1\int_0^{3-3y}\frac{1}{8}\mathrm{d}x\mathrm{d}y=\int_0^1\frac{1}{8}(3-3y)\mathrm{d}y$$
$$=\left[\frac{3}{8}\left(y-\frac{y^2}{2}\right)\right]_0^1=\frac{3}{8}\left[\left(1-\frac{1}{2}\right)-\left(0-\frac{0}{2}\right)\right]=\frac{3}{8}\cdot\frac{1}{2}=\frac{3}{16}.$$

The Multivariate Normal Distribution

The univariate normal distribution is the most widely used distribution in modeling univariate data. The most widely used distribution for modeling multivariate data is the multivariate normal distribution. We will restrict our discussion to the bivariate case, but the discussion is easily generalized to three or more variables.

A pair of random variables (X,Y) is said to follow a bivariate normal distribution, which means, $\mu=(\mu_1,\mu_2)$ for fixed $\mu_1,\mu_2\in\mathbb{R}$ and the covariance matrix is given by

$$\Sigma=\begin{bmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{bmatrix}\quad\text{for fixed}\quad\sigma_1,\sigma_2,\sigma_{12}\in\mathbb{R}.$$

We define the determinant of Σ by

$$\det(\Sigma)=\sigma_1^2\sigma_2^2-\sigma_{12}\sigma_{12},$$

and the joint PDF is then

$$f_{X,Y}(x,y)=\frac{1}{2\pi\sqrt{\det(\Sigma)}}\,\mathrm{e}^{-\frac{1}{2}\left(\binom{x}{y}-\binom{\mu_1}{\mu_2}\right)^T\cdot\Sigma^{-1}\cdot\left(\binom{x}{y}-\binom{\mu_1}{\mu_2}\right)}\quad\text{for } x,y\in\mathbb{R},$$

where Σ^{-1} the inverse oft the matrix Σ, which is

$$\Sigma^{-1}=\frac{1}{\det(\Sigma)}\begin{bmatrix}\sigma_2^2 & -\sigma_{12}\\ -\sigma_{12} & \sigma_1^2\end{bmatrix}.$$

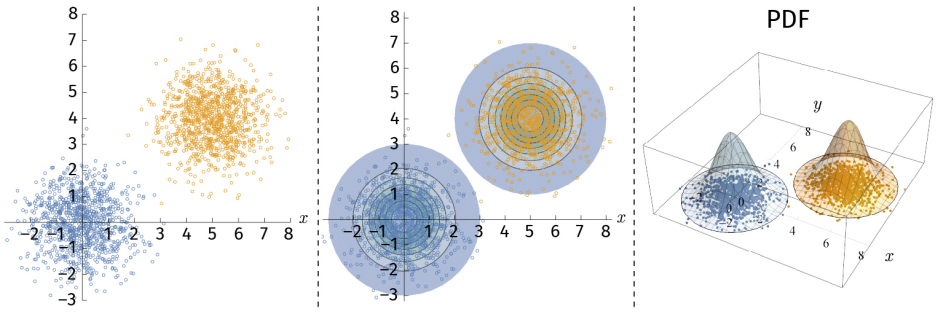
We can extend this definition to a multivariate distribution where $X=\left(X_1,X_2,\quad ...,X_p\right)$ by analogously extending μ and Σ. It is necessary that $\det(\Sigma)>0$. To get a feeling for this distribution, we will begin by exploring a few bivariate data sets generated from different bivariate normal distributions.

In the figure below, we generate two samples containing 1000 data points each. One sample (in blue) is from a bivariate normal distribution with mean $\mu=(0,0)$ and covariance matrix

$$\Sigma=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$$

The other sample, in orange, has mean $\mu = (5,4)$ and the same covariance matrix. The mean is where the points are most dense. They become sparser as they move away from this point. The diagonal elements of the covariance matrix quantify the variation in the two variables. The zeros on the off-diagonal indicate that there is no relationship between the two variables.

Figure 44: Bivariate Normal Distribution: Random Samples and PDFs - (Identity Covariance)



The above two samples were generated from bivariate normal distributions with an identity covariance matrix. The points in blue have a mean at (0,0), and the orange points have a mean at (5,4).

Left: Scatter plot showing the two samples
Middle: Scatter plots overlaid with the contours of the respective PDFs
Right: Scatter plots in 3D and surfaces of the joint PDFs

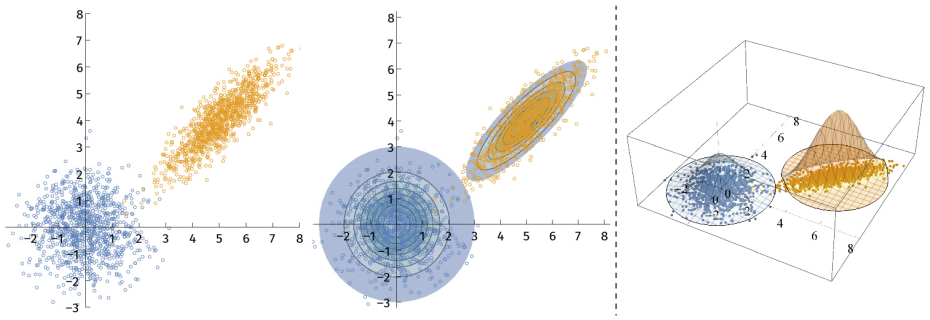
Source: George Dekermenjian (2019).

In this next figure, the blue points are as before, but the orange sample is generated from a bivariate normal distribution with a mean of $\mu = (5,4)$ and a covariance matrix of

$$\Sigma = \begin{bmatrix} 1 & 0.85 \\ 0.85 & 1 \end{bmatrix}$$

The off-diagonal entries (always equal) indicate that there is a positive relationship between the two variables. If one variable has a higher value, it is likely that the other variable has a higher value as well. Notice how the shape of the contours as well as the surface of the joint PDF of the orange distribution are elliptical, with the major axis having a positive slope.

Figure 45: Bivariate Normal Distribution: Random Samples and PDFs - (Positive Covariance)



The above two samples were generated from bivariate normal distributions.
The points in blue have a mean at (0,0) and an identity covariance matrix.
The orange points have a mean at (5,4) and a dense covariance matrix
(the ones on the main diagonal and 0.85 for the off-diagonal entries).

Left: Scatter plot showing the two samples
Middle: Scatter plots overlaid with the contours of the respective PDFs
Right: Scatter plots in 3D and surfaces of the joint PDFs

Source: George Dekermenjian (2019).

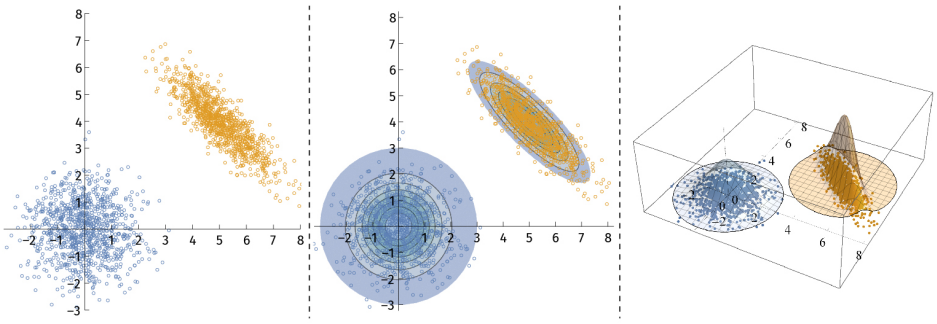
Finally, in the figure below, the blue points are again as before, but the orange sample is generated from a bivariate normal distribution with $\mu = (5,4)$ mean and covariance matrix

$$\Sigma = \begin{bmatrix} 1 & -0.85 \\ -0.85 & 1 \end{bmatrix}$$

The off-diagonal entries indicate that there is a negative relationship between the two variables. If one variable has a higher value, it is likely that the other variable has a lower value. Notice how the shape of the contours as well as the surface of the joint PDF of the orange distribution are elliptical, with the major axis having a negative slope.

The off-diagonal elements of the covariance matrix can be computed by studying a special kind of expectation.

Figure 46: Bivariate Normal Distribution: Random Samples and PDFs - (Negative Covariance)



The above two samples were generated from bivariate normal distributions. The points in blue have a mean at (0,0) and an identity covariance matrix. The orange points have a mean at (5,4) and a dense covariance matrix (the ones on the main diagonal and −0.85 for the off-diagonal entries).

Left: A scatter plot showing the two samples
Middle: Scatter plots overlaid with the contours of the respective PDFs
Right: Scatter plots in 3D and surfaces of the joint PDFs

Source: George Dekermenjian (2019).

3.2 Marginal Distributions

A company sells products from two categories, A and B. Some of its customers purchase products from both categories. These are the customers we are interested in. Category A has three products priced at \$100, \$300, and \$400. Category B has two products priced at \$100 and \$250. For a given customer, let the random variables X and Y be the prices of the products purchased from category A and B, respectively. The following table gives the joint distribution of the customers who purchased a product from each of the two categories.

Table 19: Joint PMF of Product Purchases

		X		
		100	300	400
Y	100	1/12	5/12	2/12
	250	2/12	1/12	1/12

Source: George Dekermenjian (2019).

The probability that a randomly chosen customer purchases the $100 product from category A and the $250 product from category B is $P(X = 100, \quad Y = 250) = \frac{2}{12}$. What is the probability that a randomly chosen customer purchases the $100 product from category A? In this question, we have no mention of category B. In other words, we want the event where $X = 100$ and Y is "free". This translates to

$$P(X = 100) = P(X = 100 \quad \text{and} \quad (Y = 100 \quad \text{or} \quad Y = 250)),$$

or equivalently

$$P((X = 100, Y = 100) \quad \text{or} \quad (X = 100, \quad Y = 250)) .$$

Since the events are mutually exclusive, we have

$$P(X = 100) = P(X = 100,\ Y = 100) + P(X = 100,\ Y = 250) = \frac{3}{12} = \frac{1}{4}$$

Going back to the PMF given in the table, this is just the sum of the values under $X = 100$. Let us add the values from the other columns and summarize them in a new table:

Table 20: Joint PMF of Product Purchases and the Marginal of *X*

		X		
		100	300	400
Y	100	1/12	5/12	2/12
	250	2/12	1/12	1/12
	f_X	**1/4**	**1/2**	**1/4**

Source: George Dekermenjian (2019).

This yields values for the function f_X, which is obviously a PMF. We will define this PMF as marginal PMF of X, which will be discussed later in this unit. Indeed, the values are between zero and one and their sum is one. Therefore, these values satisfy the properties of a (univariate) PMF. We can now add the rows to get the marginal probability mass function of Y.

Table 21: Joint PMF of Product Purchases and Marginals of *X* and *Y*

		X			
		100	300	400	f_Y
Y	100	1/12	5/12	2/12	**2/3**
	250	2/12	1/12	1/12	**1/3**

		X			
		100	300	400	f_Y
	f_X	**1/4**	**1/2**	**¼**	

Source: George Dekermenjian (2019).

Now that we have the marginal PMF, it is easy to answer questions like: "what is the probability that a randomly chosen customer purchased the $250 dollar item from category B?". The answer is:

$$P(Y = 250) = \frac{1}{3}$$

Marginal distributions for jointly discrete random variables

Suppose X and Y are discrete random variables with the joint PMF $f_{X,Y} \quad (x, y)$ with $x \in \{x_1, x_2...\} \subseteq \mathbb{R}$ and $y \in \{y_1, y_2, ..., \} \subseteq \mathbb{R}$. Then the marginal PMF of X is given by

$$f_X(x) = \sum_{j=1}^{\infty} f_{X,Y}\big(x, y_j\big) \text{ for all } x \in \{x_1, x_2...\},$$

and the marginal PMF of Y

$$f_Y(y) = \sum_{j=1}^{\infty} f_{X,Y}\big(x_j, y\big) \quad \text{for all} \quad y \in \{y_1, y_2...\},$$

In other words, the marginal distribution of X is the sum of the joint distribution over all the possible values of Y and vice versa for the marginal distribution of Y. The marginal distributions of X and Y are then

$$P[X \leq x] = \sum_{i=1}^{\lfloor x \rfloor} f_X(x_i) \text{ for any } x \in \mathbb{R},$$
$$P[Y \leq y] = \sum_{i=1}^{\lfloor y \rfloor} f_Y(y_i) \text{ for any } y \in \mathbb{R},$$

where P denotes the probability measure of the cross-product of the sample spaces of X and Y.

Marginal distributions for jointly continuous random variables

Given continuous random variables X and Y with the joint PDF $f_{X,Y}(x, y)$ for $x, y \in \mathbb{R}$, we can define the marginal PDF of X by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy \text{ for all } \quad x \in \mathbb{R},$$

and the marginal PDF of Y by

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dx \text{ for all } y \in \mathbb{R}.$$

The marginal cumulative distribution functions of X and Y are for the continuous case

$$P[X \leq x] = \int_{-\infty}^{x} f_X(t)\mathrm{d}t \text{ for any } x \in \mathbb{R},$$
$$P[Y \leq y] = \int_{-\infty}^{y} f_Y(t)\mathrm{d}t \text{ for any } y \in \mathbb{R},$$

where P denotes the probability measure of $\mathbb{R}\mathrm{x}\mathbb{R}$ ..

Example 3.3

Let X and Y be continuous random variables with the joint the PDF given by $f_{X,Y}(x,y) = \mathrm{e}^{-(x+y)}$ if $x, y > 0$ and zero otherwise. Find the marginal PMF $f_X(x)$ and $f_Y(y)$.

Solution 3.3

Following the discussion above, we integrate to find the marginals. Since the support of the joint density is only for positive values, the limits of integration will reflect this.

$$f_x(x) = \int_0^{\infty} \mathrm{e}^{-(x+y)}\mathrm{d}y = \mathrm{e}^{-x}\int_0^{\infty} e^{-y}\mathrm{d}y = \mathrm{e}^{-x} \cdot 1 = \mathrm{e}^{-x},\ x > 0$$

Otherwise, the function f_X is zero. Similarly, we have $f_Y(y) = e^{-y}$ for $y > 0$ and zero otherwise. Both of the marginals are exponential distributions: $X \sim \mathrm{Exponential}(1)$ and $Y \sim \mathrm{Exponential}(1)$.

Example 3.4

Let X and Y be continuous random variables with the joint PDF given by

$$f_{X,Y}(x,y) = \frac{x+y}{3} \text{ if } (x,y) \in [0,1] \times [0,2]$$

and zero otherwise. Find the marginal densities $f_X(x)$ and $f_Y(y)$. Sketch the graphs of the joint density as well as the graphs of the marginals.

Solution 3.4

To find the f_X we integrate out y. Remember that integrating with respect to y means treating x as a constant in the integral:

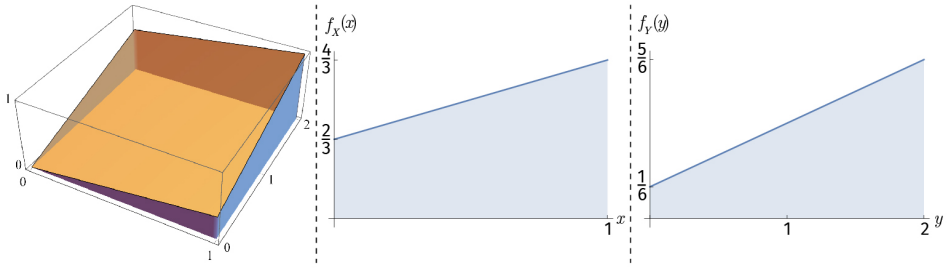
$$f_X(x) = \int_0^2 \frac{x+y}{3} dy = \left[\frac{x}{3} \cdot y + \frac{1}{3} \cdot \frac{y^2}{2}\right]_{y=0}^{y=2} = \left(\frac{2x}{3} + \frac{2}{3}\right) - (0+0) = \frac{2}{3}(x+1),\ 0 \leq x \leq 1\,.$$

Otherwise, the function f_X is zero. To find $f_Y(y)$ we integrate with respect to the variable x and obtain

$$f_Y(y) = \int_0^1 \frac{x+y}{3} dx = \frac{1}{6}(2y+1),\ 0 \leq y \leq 2\,.$$

Otherwise, the function f_Y is zero.

Figure 47: Joint PDF and Marginals for Example 3.4



Left: Surface plot of the joint PDF
Middle: Plot of the marginal PDF of X
Right: Plot of the marginal PDF of Y

Source: George Dekermenjian (2019).

3.3 Independent Random Variables

We recall that for a given sample space Ω two events $A \subseteq \Omega$ and $B \subseteq \Omega$ are independent if the probability of the joint event is the product of the individual events, which means

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

We extend the independence of events to random variables. For motivation let us consider two discrete random variables X and Y. We put $A = [X = 1]$ and $B = [Y = 2]$. Then it holds by definition

$$P(A \cap B) = P(X = 1, Y = 2) = f_{X,Y}(1,2)\,.$$

If we assume that A and B are independent, then we obtain

$$f_{X,Y}(1,2) = P(A \cap B) = P(A)P(B) = P(X = 1)P(Y = 2) = f_X(1)f_Y(2)$$

Here we have an example of the independence for a joint PMF of two random variables. Therefore, we define the following: for the sample spaces Ω_1, Ω_2, equipped with the probability measure P on $\Omega_1 \times \Omega_2$, we consider the discrete random variables $X : \Omega_1 \to \{x_1, x_2, x_3, \dots\} \subseteq \mathbb{R}$ and $Y : \Omega_2 \to \{y_1, y_2, y_3, \dots\} \subseteq \mathbb{R}$. We say that X and Y are independent random variables if for the joint PMF $f_{X,Y}$ holds

$$f_{X,Y}(x, y) = f_X(x)f_Y(y) \text{ for all } x \in \{x_1, x_2, x_3, \dots\} \text{ and } y \in \{y_1, y_2, y_3, \dots\},$$

where f_X denotes the marginal probability mass function of X and f_Y denotes the marginal probability mass function of Y.

Without focusing on the mathematical details, we can see that two discrete random variables are independent if their joint PMF is the same as the product of each marginal probability mass function. The same is true for joint PDFs, which will be stated below for the sake of completeness.

For the sample spaces Ω_1, Ω_2, equipped with the probability measure P on $\Omega_1 \times \Omega_2$, we consider the (piecewise) continuous random variables $X: \Omega_1 \to \mathbb{R}$ and $Y: \Omega_2 \to \mathbb{R}$. We say that X and Y are independent random variables if for the joint PDF $f_{X,Y}$ holds

$$f_{X,Y}(x,y) = f_X(x) f_Y(y) \quad \text{for all} \quad x, y \in \mathbb{R},$$

where f_X denotes the marginal PDF of X and f_Y denotes the marginal PDF of Y.

Example 3.5

Suppose X and Y are independent discrete random variables with the marginals of the PMFs summarized in the table below. Write down the joint PDF as a two-way table.

Table 22: Marginal PDF of X and Y for Example 3.5

½$_X$ (x)	¼/2	1/4	1/8	1/8
$f_Y(y)$	1/3	1/6	1/6	1/3

Source: George Dekermenjian (2019).

Solution 3.5

Since the random variables are independent, we know that the joint PMF is the product of the marginals. For example,

$$f_{X,Y}(0,0) = f_X(0) f_Y(0) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$$

and

$$f_{X,Y}(1,3) = f_X(1) f_Y(3) = \frac{1}{4} \cdot \frac{1}{3} = \frac{1}{12}$$

Continuing in this way, we can compute all the values of $f_{X,Y}(x,y)$:

Table 23: Joint PMF of (X,Y) for Example 3.5

		X			
		0	1	2	3
Y	0	1/6	1/12	1/24	1/24
	1	1/12	1/24	1/48	1/48
	2	1/12	1/24	1/48	1/48
	3	1/6	1/12	1/24	1/24

Source: George Dekermenjian (2019).

Example 3.6

Suppose $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(2,1)$ are independent. Write down the joint PDF $f_{X,Y}(x,y)$.

Solution 3.6

We know that the marginals are

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

and

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{(y-2)^2}{2}}$$

Since the variables are independent, the joint PDF is just the product of the marginals:

$$f_{X,Y}(x,y) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{(y-2)^2}{2}} = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2+(y-2)^2}{2}}.$$

Note that that is exactly the multivariate normal distribution with mean vector

$$\mu = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$$

and the covariance matrix

$$\sum = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

In fact, if the covariance matrix is diagonal (zeros on the off diagonal), then the random variables in the multivariate normal distribution are independent and their marginals have the univariate normal PDF.

Example 3.7

Consider the joint PDF and the marginals given by the two-way table below. Are the random variables X and Y independent? Explain.

Table 24: Joint PMF and Marginals for Example 3.7

		X			
		100	300	400	f_Y
Y	100	1/12	5/12	2/12	2/3
	250	2/12	1/12	1/12	1/3
	f_X	1/4	1/2	1/4	

Source: George Dekermenjian (2019).

Solution 3.7

If X and Y are independent, the joint PMF must be the product of the marginals, meaning

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) \text{ for all } x \in \{100,300,400\} \text{ and } y \in \{100,250\},$$

but we immediately see that

$$f_{X,Y}(100,100) = \frac{1}{12}$$

is not the same as

$$f_X(100)f_Y(100) = \frac{1}{4} \cdot \frac{2}{3} = \frac{2}{12}.$$

So, if one pair of values fails to hold, it means that the two random variables fail to be independent.

Note that to prove that two random variables are independent, we are required to confirm that the product holds for every possible pair. To prove that they are dependent (not independent), the product needs to fail for just one pair. These requirements are for theoretical models (distributions). However, when dealing with real-world data the departure from independence can often tolerate some error.

Through the following example we can see that the property of independence for real-world data generally allows for some error toleration.

A survey was conducted which asked 500 participants about their level of education and whether or not they believe that global warming is a real fact. The results are summarized in the table below, with the cells showing the numbers counted in each category.

Table 25: Two-Way Table of Observed Frequencies for Education Versus Belief in Global Warming

		Education (*X*)			Total
		None	Undergraduate degree	Graduate degree	
Believe in global warming (*Y*)	No	50	100	50	**200**
	Yes	70	100	130	**300**
Total		**120**	**200**	**180**	**500**

Source: George Dekermenjian (2019).

We are interested in testing the hypothesis that level of education and belief in global warming are independent. To this end, we define two random variables, X (level of education) and Y (belief in global warming). The distributions of X and Y can be modeled as so-called discrete multi-noulli, which is a natural probability model for a small data set.

In the case of X we assume three probabilities $0 < p_1, p_2, p_3$ such that

$$\begin{aligned} p_1 + p_2 + p_3 &= 1 \text{ and} \\ f_X(\text{None}) &= p_1, f_X(\text{Undergraduate Degree}) \\ &= p_2, f_X(\text{Graduate Degree}) = p_3, \end{aligned}$$

where f_X denotes the marginal PMF of X. Hence if we compute p_1, p_2, p_3 we have a PMF for the marginal PMF of X. Note, that the input variables of f_X are not numbers. In general, we would label the three events None, Undergraduate Degree and Graduate Degree with fixed numbers, but for our purposes this is not necessary and would distract us from the aim of computing f_X.

In the same way we model the marginal PMF of Y, which means that we assume two probabilities $0 < q_1, q_2$ such that

$$\begin{aligned} q_1 + q_2 &= 1 \quad \text{and} \\ f_Y(\text{No}) &= q_1, \quad f_Y(\text{Yes}) = q_2, \end{aligned}$$

where f_Y denotes the marginal PMF of Y. Under the assumption that the two variables are independent, the joint PMF is given by

Table 26: Joint PMF of Education Versus Belief in Global Warming (Symbolic Probabilities)

		None	Undergraduate degree	Graduate degree	Marginal f_Y
Y	No	p_1q_1	p_2q_1	p_3q_1	q_1
	Yes	p_1q_2	p_2q_2	p_3q_2	q_2
Marginal f_X		p_1	p_2	p_3	1

Source: George Dekermenjian (2019).

Therefore, using the count data, we can obtain

$$p_1=\frac{120}{500},$$
$$p_2=\frac{200}{500},$$
$$p_3=\frac{180}{500}.$$

for the marginals, and

$$q_1=\frac{200}{500},$$
$$q_2=\frac{300}{500}.$$

The table above now reads:

Table 27: Joint PMF of Education Versus Belief in Global Warming (Numerical Probabilities)

		None	Undergraduate degree	Graduate degree	Marginal
Y	No	0.096	0.16	0.144	0.4
	Yes	0.144	0.24	0.216	0.6
Marginal		0.24	0.4	0.36	1

Source: George Dekermenjian (2019).

Based on this model, which assumes independence, if we sample 500 participants, the expected counts are given by multiplying each probability by 500, which yields the following theoretical count data. We merge the actual counts with the expected counts appearing in parentheses:

Table 28: Observed and Expected Frequencies of Education Versus Belief in Global Warming

		Education			Total
		None	Undergraduate degree	Graduate degree	
Believe in global warming	No	50 (48)	100 (80)	50 (72)	**200**
	Yes	70 (72)	100 (120)	130 (108)	**300**
Total		**120**	**200**	**180**	**500**

Source: George Dekermenjian (2019).

Examining the values, we compare the observed counts from the survey to the expected counts from the assumption of independence, and although the two sets of counts do not match exactly, we can perhaps argue that there is not much evidence to conclude that the two variables are not independent. However, some of the counts are quite close, while others are not that close, which means that we cannot prove for certain whether or not we have an independence. In statistical inference, this departure from independence is modeled quantitatively using theχ^2 (chi-square) distribution.

3.4 Conditional Distributions

Let $A \subseteq \Omega$ and $B \subseteq \Omega$ be two events from some random experiment. We recall that the conditional probability of A conditioned on B is defined as

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \text{ for } P(B) > 0\,.$$

We can easily extend this definition to the case of PDFs and PMFs considering the joint distributions of two given random variables.

For the sample spaces Ω_1, Ω_2, equipped with the probability measure P on $\Omega_1 \times \Omega_2$, we consider the discrete random variables $X : \Omega_1 \to \{x_1, x_2, x_3, \dots\} \subseteq \mathbb{R}$ and $Y : \Omega_2 \to \{y_1, y_2, y_3, \dots\} \subseteq \mathbb{R}$. The conditional PMF of X conditioned on $Y = y$ for any fixed $y \in \{y_1, y_2, y_3, \dots\}$ such that $f_Y(y) > 0$ is defined as

$$f_{X\,|\,Y}(x\,|\,y) = \frac{f_{X,\,Y}(x, y)}{f_Y(y)} \quad \text{for all} \quad x \in \{x_1, x_2, x_3, \dots\},$$

where $f_{X,\,Y}$ is the joint PMF of X and Y. In the same way we define the conditional distribution for continuous random variables.

For the sample spaces Ω_1, Ω_2, equipped with the probability measure P on $\Omega_1 x \Omega_2$, we consider the (piecewise) continuous random variables $X : \Omega_1 \to \mathbb{R}$ and $Y : \Omega_2 \to \mathbb{R}$. The conditional PDF of X conditioned on $Y = y$ for any fixed $y \in \mathbb{R}$ such that $f_Y(y) > 0$ is defined as

$$f_{X\,|\,Y}(x\,|\,y) = \frac{f_{X,\,Y}(x,\,y)}{f_Y(y)} \quad \text{for all} \quad x \in \{x_1, x_2, x_3, \dots\},$$

where $f_{X,\,Y}$ is the joint probability density function of X and Y.

Example 3.8

Let X and Y be discrete random variables, with the joint PMF and the marginals defined by the following two-way table.

Table 29: Joint PMF and Marginals for Example 3.8

		X			
		100	300	400	f_Y
Y	100	1/12	5/12	2/12	2/3
	250	2/12	1/12	1/12	1/3
		¼1	½1	¼1/4	

Source: George Dekermenjian (2019).

a) Write down the conditional probability mass function for Y given $X = 100$.
b) Write down the conditional probability mass function for Y given $X = 300$.

Solution 3.8

a) Following the definition, we have

$$f_{Y\,|\,X}(y\,|\,100) = \frac{f_{X,\,Y}(100,\,y)}{f_X(100)} = \frac{f_{X,\,Y}(100,\,y)}{\frac{1}{4}} = 4 f_{X,\,Y}(100,\,y)\,.$$

Therefore,

$$f_{Y\,|\,X}(100\,|\,100) = 4 \cdot \tfrac{1}{12} = \tfrac{1}{3}.$$

Similarly,

$$f_{Y\,|\,X}(250\,|\,100) = 4 \cdot \tfrac{2}{12} = \tfrac{2}{3}.$$

We can summarize the conditional PMF in tabular form:

Table 30: Conditional PDF of $Y \mid X = 100$ for Example 3.8

y	100	250
$f_{Y \mid X}(y \mid 100)$	1/3	2/3

Source: George Dekermenjian (2019).

b) Similar to a) we have

$$f_{Y \mid X}(y \mid 300) = \frac{f_{X,Y}(300, y)}{f_X(300)} = \frac{f_{X,Y}(300, y)}{\frac{1}{2}} = 2 f_{X,Y}(300, y).$$

Table 31: Conditional PDF of $Y \mid X = 300$ for Example 3.8

y	100	250
$f_{Y \mid X}(y \mid 300)$	5/6	1/6

Source: George Dekermenjian (2019).

As we have seen from the formula and the example above, we can get the marginal PMF (PDF) from the joint PMF (PDF) if we have the appropriate marginal. What about the other way around? Can we get the joint PMF (PDF) if we have the conditional PMF (PDF) and the appropriate marginal? This is the subject of the next example.

Example 3.9

Suppose that $X \sim \text{Uniform}([1,2])$ and $Y \mid X = x \sim \text{Uniform}([1, x])$ for 1 and otherwise zero.

a) Find the joint PDF of X and Y.
b) Find the marginal density function for Y.

Solution 3.9

a) We will start by writing the PDFs of X, which is

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

and the PDF of $Y \mid X = x$ for fixed $1 \leq x \leq 2$, which is

$$f_{Y|X}(y|x) = \begin{cases} \frac{1}{x-1} & \text{for } 1 \leq y \leq x, \\ 0 & \text{otherwise.} \end{cases}$$

b) Rewriting the definition of the conditional density function, we obtain

$$f_{X,Y}(x,y) = f_{Y|X}(y|x) f_X(x) = \begin{cases} \frac{1}{x-1} & \text{for } 1 \leq x \leq 2 \text{ and } 1 \leq y \leq x, \\ 0 & \text{otherwise.} \end{cases}.$$

Example 3.10

An ecologist is interested in quantifying the effects of eagle predation on the behavior of rabbits. The ecologist sees an eagle with a probability of 0.2% during an hour of observation, a jack rabbit with a probability of 0.5%, and both with a probability of 0.05%. Find the joint PMF, the marginal PMFs, and the conditional PMFs.

Solution 3.10

From the exercise we know the following:

Table 32: Joint PMF and Marginals for Example 3.10

		Observing an Eagle		
		$E = 0$	$E = 1$	f_R
Observing a rabbit	$R = 0$			
	$R = 1$		0.05	0.5
	f_E		0.2	

Source: Bartosch Ruszkowski (2022).

Here $E = 1$ denotes the event to observe an eagle and $E = 0$ for not observing one. The same holds for $R = 0$ and $R = 1$ replacing the eagle with the rabbit. First of all, the marginals need to sum up to one, yielding:

Table 33: Joint PMF and Marginals for Example 3.10

		Observing an Eagle		
		$E = 0$	$E = 1$	f_R
Observing a rabbit	$R = 0$			0.5
	$R = 1$		0.05	0.5
	f_E	0.8	0.2	

Source: Bartosch Ruszkowski (2022).

We take into consideration that the second row and column have to sum up to the corresponding probability of the margin and obtain:

Table 34: Joint PMF and Marginals for Example 3.10

		Observing an Eagle		
		$E = 0$	$E = 1$	f_R
Observing a rabbit	$R = 0$		0.15	0.5
	$R = 1$	0.45	0.05	0.5
	f_E	0.8	0.2	

Source: Bartosch Ruszkowski (2022).

We find the last missing value in the same way to obtain the complete PMF:

Table 35: Joint PMF and Marginals for Example 3.10

		Observing an Eagle		
		$E = 0$	$E = 1$	f_R
Observing a rabbit	$R = 0$	0.35	0.15	0.5
	$R = 1$	0.45	0.05	0.5
	f_E	0.8	0.2	

Source: Bartosch Ruszkowski (2022).

We denote the joint PMF by $f_{E,\,R}$. Then the conditional PMFs for the eagles are

$$f_{E\,|\,R}(E = 0\,|\,R = 0) = \frac{f_{E,\,R}(E = 0, R = 0)}{f_R\,(R = 0)} = \frac{0.35}{0.5} = 0.7,$$
$$f_{E\,|\,R}(E = 1\,|\,R = 0) = \frac{f_{E,\,R}(E = 1, R = 0)}{f_R(R = 0)} = \frac{0.15}{0.5} = 0.3,$$

and

$$f_{E\,|\,R}(E = 0\,|\,R = 1) = \frac{f_{E,\,R}(E = 0, R = 1)}{f_R\,(R = 1)} = \frac{0.45}{0.5} = 0.9,$$
$$f_{E\,|\,R}(E = 1\,|\,R = 1) = \frac{f_{E,\,R}(E = 1, R = 1)}{f_R(R = 1)} = \frac{0.05}{0.5} = 0.1\,.$$

Now we change the roles of E and R and compute the conditional PMFs for the rabbits

$$f_{R|E}(R=0\,|\,E=0)=\frac{f_{E,R}(E=0,R=0)}{f_E\,(E=0)}=\frac{0.35}{0.8}=0.4375,$$
$$f_{R|E}(R=1\,|\,E=0)=\frac{f_{E,R}(E=0,R=1)}{f_E\,(E=0)}=\frac{0.45}{0.8}=0.5625,$$

and

$$f_{R|E}(R=0\,|\,E=1)=\frac{f_{E,R}(E=1,R=0)}{f_E\,(E=1)}=\frac{0.15}{0.2}=0.75,$$
$$f_{R|E}(R=1\,|\,E=1)=\frac{f_{E,R}(E=1,R=1)}{f_E\,(E=1)}=\frac{0.05}{0.2}=0.25\,.$$

SUMMARY

In this unit, we discussed joint distributions for pairs of random variables that are both either discrete or continuous. Just as for univariate distributions, we discussed how the joint distributions (PMF or PDFs) must satisfy the analogous two properties. The simplest example of joint PMFs comes from finite discrete random variables for which we can write down all the values.

In addition, we introduced the multivariate normal distribution, which comes up very often in data science. We also introduced the multivariate hyper-geometric distribution (which models a general urn problem). The two-dimensional uniform distribution was also discussed.

In section 3.2, we introduced the marginal. We saw that for two discrete random variables X and Y and its joint PMF the marginal PMF of X is also a PMF, since it is defined as

$$f_X(x) = \sum_{y=1}^{\infty} f_{X,Y}(x,y),$$

where $f_{X,Y}$ is the joint PMF and x is a fixed outcome of the sample space of X. For two piecewise continous random variables X and Y and its joint PDF the marginal PDF of X is also a PDF , since it is defined as

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\mathrm{d}y,$$

where $\mathrm{f_{X,Y}}$ *is the joint PDF and* $x \in \mathbb{R}$.

In section 3.3, we defined the independence of two given random variables and realized that the definition for PDF and PMF are basically the same. Namely, the random variables X and Y are independent if their joint PMF (PDF) can be factored into a product of their marginals such that the following equality holds

$$f_{X,Y}(x,y) = f_X(x) f_Y(y)$$

for all x in the sample space of X and y in the sample space of Y. If we think the two variables are dependent (not independent), we have to show that this product fails to hold for only one pair (x, y).

In section 3.4 we defined the conditional PMF and PDF. For a joint PMF (PDF) $f_{X,Y}$ of two random variables X and Y the conditional PMF (PDF) $f_{X|Y}$ of X on $Y = y$ is

$$f_{X|Y}(x\,|\,y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

for all x in the sample space of X and fixed y in the sample space of Y such that $f_Y(y) > 0$. If we know the conditional PMF (PDF) and the appropriate marginal we can always obtain the joint PMF(PDF) by multiplying that equality with $f_Y(y)$.

UNIT 4

EXPECTATION AND VARIANCE

STUDY GOALS

On completion of this unit, you will be able to ...

- compute and interpret conditional expectation and the expectation of a random variable.
- understand variance and covariance.
- compute and interpret the expectation and variance of important discrete and continuous probability distributions.
- work with central moments of random variables.
- apply moment generating functions in order to identify a distribution and to compute moments.

4. EXPECTATION AND VARIANCE

Introduction

Consider the following experiment: we toss a coin, and if we get heads, we must pay $1. If we get tails, we will win $1.20. Suppose we play 10 rounds of this game and have the following outcomes:

Table 36: A Sequence of Coin Tosses and Corresponding Returns

T	$1.20
H	–$1.00
H	–$1.00
T	$1.20
H	–$1.00
T	$1.20
T	$1.20
H	–$1.00
T	$1.20
T	$1.20

Source: George Dekermenjian (2019).

We want to compute the average gain. Based on the results, the average gain is just the arithmetic mean of the dollar values:

$$\frac{\$1.20 + (-\$1.00) + ... + (\$1.20)}{10} = \$0.32\,.$$

Note that we cannot hope to gain $0.32 on any one toss, but in the long run, according to the data, we gain $0.32 per toss. This is the empirical expected value or the sample mean gain of the game. We will discuss this topic in detail in this section.

Another topic of interest is how much, on average, the data vary from this mean. There are a number of ways to measure variation. One important measure is the average square distance of the data points from the mean:

$$\sqrt{\frac{(\$1.20 - \$0.32)^2 + (-\$1.00 - \$0.32)^2 + ... + (\$1.20 - \$0.32)^2}{10 - 1}} \approx \$1.14\,.$$

In other words, each data point is expected to be, on average, \$1.14 away from the mean value of \$0.32. Of course, any one toss will never be exactly \$1.14 away. In fact, if we get tails, the distance from the mean is $1.20 - 0.32 = \$0.88$, and if we get tails, the distance from the mean is $-1.00 - 0.32 = -\$1.32$. So this number, called the sample standard deviation of the gain, is just a long-term average over many realizations of the game. For mathematical ease, it is common to work with the quantity inside the square-root instead:

$$\frac{(\$1.20 - \$0.32)^2 + (-\$1.00 - \$0.32)^2 + ... + (\$1.20 - \$0.32)^2}{10 - 1} \approx \$1.29\,.$$

This is called the sample variance of the gains. The benefit of stating the standard deviation is that it has the same units as the data and so is easily interpretable. On the other hand, the variance is simpler to work with.

For paired (bivariate) data, we might be interested in finding out if it is correlated or not. For example, consider the following data showing the birth mass in grams of a child together with the mother's age in years at the time of birth.

Table 37: A Bivariate Sample of Birthweights and the Mother's Age at Birth

Mother's age (years)	Birthweight (grams)
24	1621
33	2337
25	2474
24	3098
28	3218
24	3405
26	3419
34	3602
40	3642
32	4230

Source: George Dekermenjian (2019).

Table 38: Sample Mean and Standard Deviation of Birthweight and the Mother's Age at Birth

	Sample mean	Sample standard deviation
Mother's age (years)	29	5.50
Birthweight (grams)	3104.6	759.6

Source: George Dekermenjian (2019).

The sample covariance is given by the average of the products of the differences between each of the data points and their respective means:

$$\frac{(24-29)(1621-3104.6)+...+(32-29)(4230-3104.6)}{10-1} \approx 1791$$

We can standardize this value to be between –1 and 1 by dividing it by the product of the sample standard deviations:

$$\frac{1791}{5.50 \cdot 759.6} \approx 0.4\,.$$

This is called the sample correlation.

In this unit, we will discuss these concepts of mean, variance, covariance, and other quantities for random variables, and set up the theoretical framework for understanding different properties of discrete and continuous random variables as well as for some important distributions.

4.1 Expectation of a Random Variable

Expected Value
The long-term average of a random variable based on the values it takes and its corresponding probabilities.

For a given sample space and its corresponding probability measure P, we consider a discrete random variable $X\colon\Omega \to \{x_1, x_2, x_3, ...\}$, where $x_1, x_2, ...$ are real numbers and its PMF is denoted by f_x. Then the **expected value** (also called expectation value) of X, written as $\mathrm{E}[X]$, and defined by

$$\mathrm{E}[X] = \textstyle\sum_{i=1}^{\infty} x_i \cdot f_X(x_i)\,.$$

Similarly, if $X\colon\Omega \to \mathbb{R}$ is a (piecewise) continuous random variable with PDF f_x, then its expected value is

$$\mathrm{E}[X] = \textstyle\int_{-\infty}^{\infty} x f_X(x) dx\,.$$

The following are examples of expectations with discrete random variables.

Example 4.1

Let X be a discrete random variable with the PMF defined by the values in the table below. Calculate the expected value of X.

Table 39: Discrete PMF for Example 4.1

x	$f(x)$
1	0.2
2	0.5

x	$f(x)$
3	0.1
4	0.2

Source: George Dekermenjian (2019).

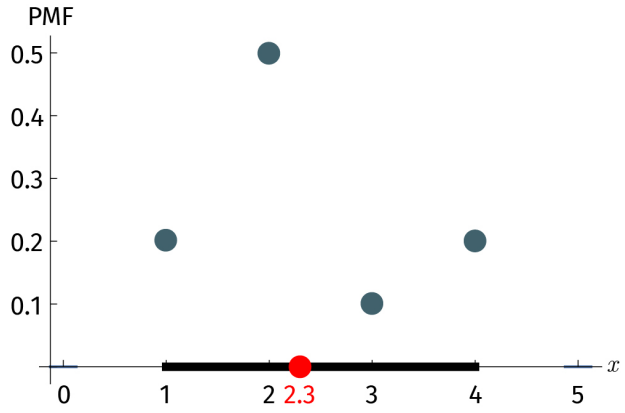
Solution 4.1

Following the definition, we have

$$\mathrm{E}[X] = 1 \cdot 0.2 + 2 \cdot 0.5 + 3 \cdot 0.1 + 4 \cdot 0.2 = 0.2 + 1 + 0.3 + 0.8 = 2.3$$

Visually, you can think of the supports (x-values) as locations on a straight rod and the probabilities as weights on those locations. The expected value is the location at which the rod will balance. The diagram below shows a graph of the PMF together with the rod as a thick line segment and the mean as the red point.

Figure 48: Plot of PDF for Example 4.1 Illustrating the Position of the Mean



Source: George Dekermenjian (2019).

Example 4.2

We own one share of a stock that is currently valued at $100. We believe that by tomorrow, the price of this stock will either increase by $5 with a probability of 40% or decrease by $10. What is the expected value of our share?

Solution 4.2

Let X be tomorrow's stock price. The possible values of X are $105 and $90 with probabilities of 40% and 60% respectively. Therefore, the PMF of X is given by

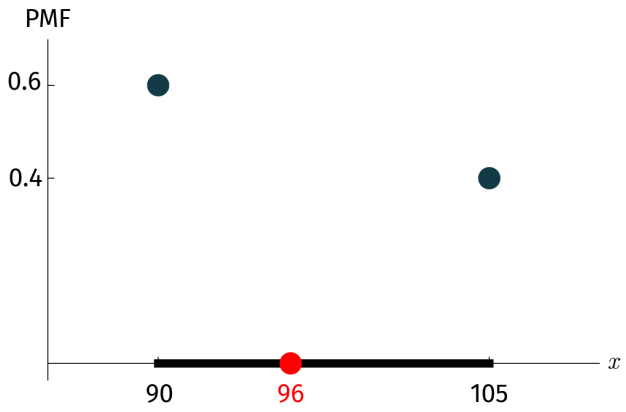
Table 40: Discrete PMF for Example 4.2

x	$f_x(x)$
90	0.60
105	0.40

Source: George Dekermenjian (2019).

The expected value of our share is $\mathrm{E}[X] = 90 \cdot 0.60 + 105 \cdot 0.40 = 96$.

Figure 49: Plot of PDF for Example 4.2 Illustrating the Position of the Mean



Source: George Dekermenjian (2019).

Let us now take a look at some examples of expectations for continuous random variables.

Example 4.3

Let X be a uniformly distributed random variable on the interval $[1,5]$, meaning $X \sim \mathrm{Uniform}([1,5])$. Compute the expected value of X.

Solution 4.3

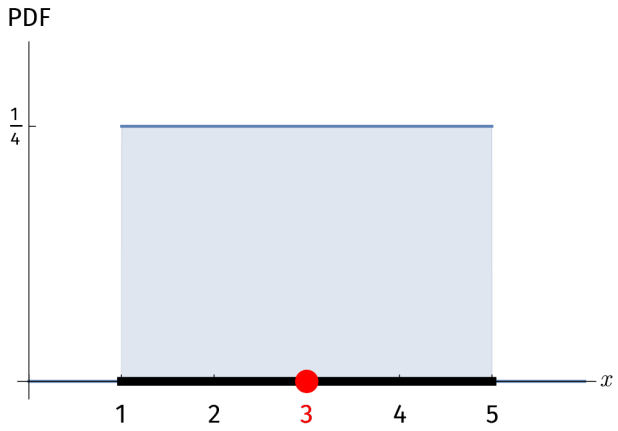
We start by writing down the PDF of X:

$$f_x(x) = \begin{cases} \frac{1}{4} \; 1 \leq x \leq 5, \\ 0 \; \text{otherwise}\,. \end{cases}$$

Following the definition of the expected value for a continuous random variable, we have

$$\mathrm{E}[x] = \int_{-\infty}^{\infty} x f_x(x) \mathrm{d}x = \int_1^5 \frac{x}{4} \mathrm{d}x = \left[\frac{1}{4} \cdot \frac{1}{2} x^2\right]_1^5 = \left[\frac{x^2}{8}\right]_1^5 = \frac{25-1}{8} = 3$$

Figure 50: Plot of PDF for Example 4.3 Illustrating the Position of the Mean



Source: George Dekermenjian (2019).

Example 4.4

Let X be a random variable with a PDF $f_x(x)$ defined by

$$f_x(x) = \begin{cases} \frac{2x}{15} & 1 \leq x \leq 4, \\ 0 & \text{otherwise.} \end{cases}$$

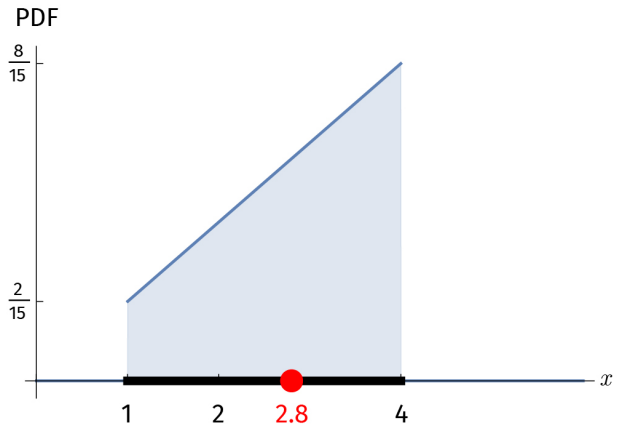
Find the expected value of X.

Solution 4.4

Following the definition, the expected value of X is computed by

$$\mathrm{E}[X] = \int_1^4 \frac{2x^2}{15} \mathrm{d}x = \left[\frac{2}{15} \cdot \frac{1}{3} x^3\right]_1^4 = \left[\frac{2x^3}{45}\right]_1^4 = \frac{14}{5} = 2.8$$

Figure 51: Plot of PDF for Example 4.4 Illustrating the Position of the Mean



Source: George Dekermenjian (2019).

Sometimes we are interested in computing the expected value of a function of the random variable $\mathrm{E}[g(X)]$. For example, we may want to calculate $\mathrm{E}\left[X^2\right]$; here $g(X) = X^2$. We compute the expected value of $g(X)$ in the following sense: for a given sample space Ω and the discrete random variable $X{:}\Omega \to \{x_1, x_2, x_3, \ldots\}$, where $x_1, x_2, \ldots$ are real numbers and its PMF is denoted by f_x, the **expected value** of $g(X)$ for $g{:}\ \{x_1, x_2, x_3, \ldots\} \to \mathbb{R}$ is given by

Expected value
The long-term average of a random variable based on its values it takes and its corresponding probabilities.

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

For continuous random variables, we require that $g{:}\ \mathbb{R} \to \mathbb{R}$ has piecewise continuous functions, allowing the integral to exist, in which case we have

$$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx,$$

where f_x denotes the PDF of X.

Example 4.5

Suppose for the random variable $X \sim \mathrm{Uniform}([-2,6])$. Compute the expected values of the following functions of X:

a) $g_1(X) = X$.
b) $g_1(X) = X2$.
c) $g_2(X) = 2X + 3$.

Solution 4.5

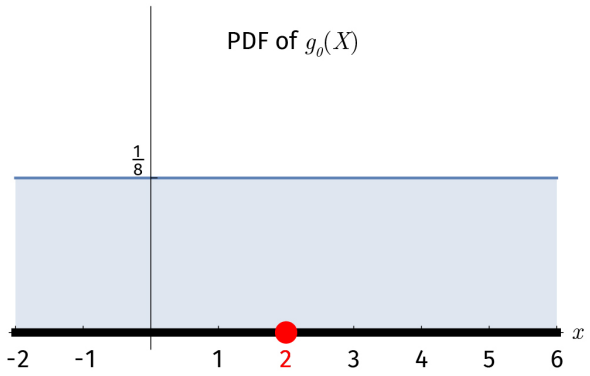
The PDF of X is given by

$$f_x(x) = \begin{cases} \frac{1}{8} & -2 \leq x \leq 6, \\ 0 & \text{otherwise.} \end{cases}$$

We proceed by the definition

a) $\mathrm{E}[g_0(X)] = \mathrm{E}[X] = \int_{-2}^{6} \frac{x}{8} \mathrm{d}x = 2.$

Figure 52: Distribution of $g_0(X)$ and its Expectation for Example 4.5 (a)



Source: George Dekermenjian (2019).

b) $E[g_1(X)] = E\left[X^2\right] = \int_{-2}^{6} \frac{x^2}{8} dx = \frac{1}{24}\left(6^3 - (-2)^3\right) = \frac{224}{24} = \frac{28}{3}$ ..

c) $E[g_2(X)] = E[2X+3] = \int_{-2}^{6} \frac{2x+3}{8} dx = \frac{6^2+3\cdot 6}{8} - \frac{(-2)^2+3\cdot(-2)}{8} = 7.$

Two important properties of expectations

The expected value of a non-random quantity is itself. For example, if $c \in \mathbb{R}$ is a constant, then $\mathrm{E}[c] = c$. This follows directly from the fact that the PMF f_x sums to one

$$\mathrm{E}[c] = \sum_{i=1}^{\infty} c \cdot f_X(x_i) = \mathrm{E}[X] = c \sum_{i=1}^{\infty} f_X(x_i) = c\,.$$

For a PDF f_x we have

$$\mathrm{E}[c] = \int_{-\infty}^{\infty} c \cdot f_X(x) = c \int_{-\infty}^{\infty} f_X(x) = c \cdot 1 = c\,.$$

A useful property of the expectation value is its linearity. For instance, in part c. of Example 4.5 we compute

$$\begin{aligned} \mathrm{E}[2X+3] &= \int_{-2}^{6} \frac{2x+3}{8} \mathrm{d}x = 2\int_{-2}^{6} \frac{x}{8} \mathrm{d}x + \int_{-2}^{6} \frac{3}{8} \mathrm{d}x = 2\mathrm{E}[X] + \mathrm{E}[3] \\ &= 2\mathrm{E}[X] + 3\,. \end{aligned}$$

In general, for non-random quantities $a, b \in \mathbb{R}$, for any random variable X, we have,

$$\mathrm{E}[aX+b] = a\mathrm{E}[X] + b$$

For instance, if we consider the Example 4.4 we obtain

$$\mathrm{E}[5X-4] = 5\mathrm{E}[X] - \ 4 = 5 \cdot \tfrac{14}{5} - 4 = 10\,.$$

Example 4.6

a) You are about to make an investment which gives you a 30% chance of making \$60,000 and 70% chance of losing \$30000. Should you invest? Explain.
b) A game involves rolling a Korean die (4 faces). If a one, two, or three is rolled, the player receives the face value of the die in dollars, but if a four is rolled, the player is obligated to pay \$4. What is the expected value of the game? Would you play it?
c) A game involves drawing a single card from a standard deck (52 cards). One receives 60 cents for an ace, and 10 cents for a red card that is not an ace. If the cost of each draw is 5 cents, should you play?

Solution 4.6

a) The PMF for the random variable X of gaining or losing money is

$$fx(x) = \begin{cases} 0.3 & \text{for } x = 60000, \\ 0.7 & \text{for } x = -30000, \\ 0 & \text{otherwise.} \end{cases}$$

The expected value is then

$$\mathrm{E}[X] = 0.3 \cdot 60000 + 0.7 \cdot (-30000) = -3000,$$

which means that we should avoid this investment since we will have a mean loss of \$3000.
b) The PMF for the random variable X, which describes face of the die, is

$$fx(x) = \begin{cases} 0.25 & \text{for } x \in \{1,2,3,-4\}, \\ 0 & \text{otherwise.} \end{cases}$$

The expected value is then

$$\mathrm{E}[X] = 0.25 \cdot (1+2+3-4) = 0.5$$

which means that we will have a mean gain of \$0.50 and will benefit from participating.

c) The PMF for the random variable X, which describes the net winnings or the loss in cents, is described by

$$fx(x) = \begin{cases} \frac{4}{52} & \text{for } x = 55, \\ \frac{12}{52} & \text{for } x = 5, \\ \frac{52-12-4}{52} & \text{for } x = -5, \end{cases}$$

otherwise, it is zero. Note that we took into account here that there are 4 aces in a card deck and 13 red cards, where we must subtract the red ace to compute the probabilities of the PMF. Then the expected value is

$$\mathrm{E}[X] = \frac{4}{52} \cdot 55 + \frac{12}{52} \cdot 5 - \frac{36}{52} \cdot 5 = \frac{100}{52} \approx 1.92,$$

which means that we should definitely play that game, since it will result in a mean gain.

Expectations of Joint Distributions

For given sample spaces Ω_1 and Ω_2 and discrete random variables $X\colon \Omega_1 \to \{x_1, x_2, x_3, ...\} \subseteq \mathbb{R}$ and $Y\colon$ where $\Omega_2 \to \{y_1, y_2, y_3, ...\} \subseteq \mathbb{R}$ and its joint PMF are denoted by f_{xy}. The expected value of X (or Y) is then

$$\begin{aligned} \mathrm{E}[X] &= \sum_{k=1}^{\infty} x_k \sum_{i=1}^{\infty} f_{X,Y}(x_k, y_i) = \sum_{i=1}^{\infty} \sum_{k=1}^{\infty} x_k f_{X,Y}(x_k, y_i) \\ &= \sum_{k=1}^{\infty} x_k f_X(x_k), \\ \mathrm{E}[Y] &= \sum_{i=1}^{\infty} y_i \sum_{k=1}^{\infty} f_{X,Y}(x_k, y_i) = \sum_{k=1}^{\infty} \sum_{i=1}^{\infty} y_i f_{X,Y}(x_k, y_i) \\ &= \sum_{i=1}^{\infty} y_i f_Y(y_i), \end{aligned}$$

where f_y denotes the marginal PMF of Y and f_x denotes the marginal PMF of X.

Consider two (piecewise) continuous random variables X and Y, whose joint density is given by $f_{X,Y}$. Then the expected value of X is given by

$$\begin{aligned} \mathrm{E}(X) &= \int_{\mathbb{R}} x \Big(\int_{\mathbb{R}} f_{X,Y}(x, y) dy \Big) dx = \int_{\mathbb{R}} \Big(\int_{\mathbb{R}} x f_{X,Y}(x, y) dx \Big) dy \\ &= \int_{\mathbb{R}} x f_X(x) dx, \end{aligned}$$

Where f_x denotes the marginal probability mass function of X. The expected value of Y is given by

$$\begin{aligned}\mathrm{E}(Y) &= \int_{\mathbb{R}} y\Big(\int_{\mathbb{R}} f_{X,Y}(x,y)dx\Big)dy = \int_{\mathbb{R}}\Big(\int_{\mathbb{R}} y f_{X,Y}(x,y)dy\Big)dx \\ &= \int_{\mathbb{R}} y f_Y(y)dy,\end{aligned}$$

where f_Y denotes the marginal probability density function of Y.

Example 4.7

Suppose the random variables X and Y have a joint PDF given by

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{3}(4x+2y) & \text{for } 0 \le x \le 1 \text{ and } 0 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Compute the expected values of X and Y.

Solution 4.7

Following the definition, we have

$$\begin{aligned}\mathrm{E}[X] &= \int_0^1 \int_0^1 \frac{x}{3}(4x+2y)\mathrm{d}y\mathrm{d}x = \int_0^1 \Big(\frac{4}{3}x^2 + \frac{1}{3}x\Big)\mathrm{d}x \\ &= \frac{4}{9}\cdot 1^3 + \frac{1}{6}\cdot 1 - \Big(\frac{4}{9}\cdot 0^3 + \frac{1}{6}\cdot 0\Big) = \frac{11}{18}.\end{aligned}$$

We can compute $\mathrm{E}[Y]$ in an alternative way. We first compute the marginal f_Y

$$f_Y(y) = \int_0^1 \frac{1}{3}(4x+2y)\mathrm{d}x = \frac{2}{3}1^2 + \frac{2}{3}y \ = \frac{2}{3}(y+1) \text{ for } \quad 0 \le y \le 1$$

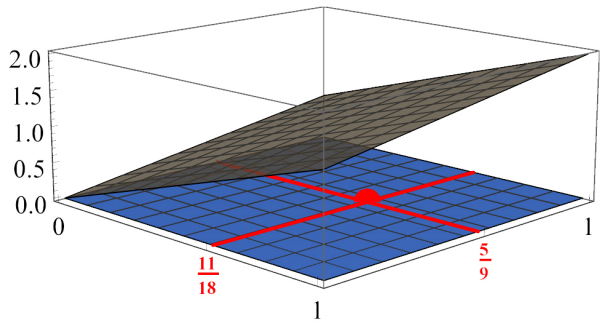
and otherwise zero, and use the marginal

$$\mathrm{E}[Y] = \int_0^1 y f_Y(y)\mathrm{d}y = \int_0^1 \frac{2y}{3}(y+1)\mathrm{d}y = \ \frac{2}{9} + \frac{2}{6} = \ \frac{2}{9} + \frac{1}{3} \ = \frac{5}{9}.$$

The diagram below shows the surface plot of the joint PDF along with the expected values of X and Y as lines in the plane. These lines intersect at the point

$$(\mathrm{E}[X], \mathrm{E}[Y]) = \Big(\frac{11}{18}, \frac{5}{9}\Big)$$

Figure 53: Plot of PDF for Example 4.7 Illustrating the Position of the Mean



Source: George Dekermenjian (2019).

Another way we could compute the expected values is by first finding the marginal distributions of X and Y respectively and computing the univariate version of the expectations.

Just as for the univariate case, we can also compute the expected value of a function of the random variables. For example, for the discrete random variables X and Y, we consider the joint PMF $f_{X,Y}$. Then for an integrable $g : \mathbb{R}^2 \longrightarrow \mathbb{R}$ the expected value of $g(X,Y)$ is

$$\begin{aligned}\mathrm{E}[g(X,Y)] &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} g(x_i,y_i) f_{X,Y}\left(x_i,y_j\right) \\ &= \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} g(x_i,y_i) f_{X,Y}\left(x_i,y_j\right),\end{aligned}$$

where $\{x_1, x_2, \ldots\} \cdot \{y_1, y_2, \ldots\} \subseteq \mathbb{R}\ x\ \mathbb{R}$ is the support of $f_{X,Y}$, which means that $f_{X,Y}$ is positive and non-zero on the support.

Similarly for the joint PDF of two (piecewise) continuous random variables X and Y, denoted by $f_{X,Y}$, and for an integrable $g : \mathbb{R}^2 \longrightarrow \mathbb{R}$ the expected value of $g(X,Y)$ is

$$\mathrm{E}[g(X,Y)] = \int_{\mathbb{R}}\int_{\mathbb{R}} g(x,y) f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y.$$

For example, suppose the random variables have the PMF given by the table below:

Table 41: A PMF of Discrete Random Variables

		X	
		1	2
Y	1	¼	3/8
	2	1/8	¼

Source: George Dekermenjian (2019).

We consider the function $g(x,y) = x \cdot y$ for $x, y \in \mathbb{R}$. The expected value of $\mathrm{E}[XY]$ is therefore computed as

$$\begin{aligned}\mathrm{E}(XY) &= \mathrm{E}(g(X,Y)) \\ &= (1\cdot 1)\left(\tfrac{1}{4}\right) + (1\cdot 2)\left(\tfrac{1}{8}\right) + (2\cdot 1)\left(\tfrac{3}{8}\right) + (2\cdot 2)\left(\tfrac{1}{4}\right) = \tfrac{9}{4}\,.\end{aligned}$$

The linearity of the expectation still holds, meaning that for $a,\ b \in \mathbb{R}$ holds

$$\mathrm{E}[aX + bY] = a\mathrm{E}[X] + b\mathrm{E}[Y]\,.$$

So, for the random variables in Example 4.7, we computed $\mathrm{E}(X) = \frac{11}{18}$ and $\mathrm{E}(Y) = \frac{5}{9}$, which yields

$$\mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y] = \tfrac{11}{18} + \tfrac{5}{9} = \tfrac{21}{18} = \tfrac{7}{6}\,.$$

Similarly, we obtain

$$\mathrm{E}[2X+3Y] = 2\mathrm{E}[X] + 3\mathrm{E}[Y] = 2\cdot\tfrac{11}{18} + 3\cdot\tfrac{5}{9} = \tfrac{22}{18} + \tfrac{15}{9} = \tfrac{42}{18} = \tfrac{7}{3}\,.$$

Independent random variables

We recall that, if X and Y are independent, then their joint PMF(PDF) is the product of each marginal. This allows us to present an important property for expectations that is frequently used when we work with independent random variables X and Y, which is to say

$$\mathrm{E}(X\cdot Y) = \mathrm{E}(X)\cdot\mathrm{E}(Y)$$

The discrete random variables in the table above are not independent. We can see this in two different ways: (i) the joint PMF cannot be written as a product of the marginals and (ii) and we compute

$$\begin{aligned}\mathrm{E}(X) &= (1)\left(\tfrac{1}{4}\right) + (1)\left(\tfrac{1}{8}\right) + (2)\left(\tfrac{3}{8}\right) + (2)\left(\tfrac{1}{4}\right) = \tfrac{13}{8}, \\ \mathrm{E}(Y) &= (1)\left(\tfrac{1}{4}\right) + (2)\left(\tfrac{1}{8}\right) + (1)\left(\tfrac{3}{8}\right) + (2)\left(\tfrac{1}{4}\right) = \tfrac{11}{8},\end{aligned}$$

and see that that

$$\mathrm{E}[XY] = \tfrac{9}{4} \neq \tfrac{13}{8}\cdot\tfrac{11}{8} = \mathrm{E}[X]\cdot\mathrm{E}[Y]$$

Conditional expectations

For the sake of simplicity, we will only discuss the conditional expectation for (piecewise) continuous random variables, although this theory is extendable to PMFs as well.

Given the joint distribution PDF of $f_{X,Y}(x,y)$ of two (piecewise) continuous random variables X and Y, we define the **conditional expectation** of X conditioned on $Y = y$ for $y \in \mathbb{R}$ as

Conditional expectation The expectation value of a random variable X conditioned to the event $Y = y$ describes the long-term event conditioned to the event that Y may take the value y.

$$\mathrm{E}[X \,|\, Y = y] = \int_{\mathbb{R}} x f_{X\,|\,Y}(x \,|\, y)\ \mathrm{d}x = \int_{\mathbb{R}} x \frac{f_{X,Y}(x,y)}{f_Y(y)}\ \mathrm{d}x$$
$$\text{where} \quad f_Y(y) > 0$$

If X and Y are independent, we have $\mathrm{E}(X \,|\, Y) = \mathrm{E}(X)$, which can be easily proved. We mention, that if we don't fix the parameter y, we see that $\mathrm{E}[X \,|\, Y]$ is a random variable with the sample space of Y as domain.

Example 4.8

Consider the random variables from Example 4.7. Compute $\mathrm{E}[X \,|\, Y]$ in terms of y and then use this to evaluate $\mathrm{E}\big[X|Y = \frac{1}{3}\big]$.

Solution 4.8

We start by computing the conditional PDF:

$$f_{X|Y}(x \,|\, y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{\frac{1}{3}(4x+2y)}{\frac{2}{3}(y+1)} = \frac{4x+2y}{2(y+1)} = \frac{2x+y}{y+1},\ 0 \le y \le 1.$$

Therefore,

$$\mathrm{E}[X|Y] = \int_0^1 x \cdot \frac{2x+y}{y+1} \mathrm{d}x = \frac{3y+4}{6(y+1)},\ 0 \le y \le 1.$$

Finally, we have

$$\mathrm{E}\big[X|Y = \tfrac{1}{3}\big] = \Big[\frac{3y+4}{6(y+1)}\Big] y = \frac{1}{3} = \frac{5}{8}.$$

4.2 Variance and Covariance

For univariate data, the (sample) variance measures the variation of the values comparing each data point to the mean. The (theoretical) variance of a random variable is similar. For any random variable X, the variance of X is defined by

$$\mathrm{Var}[X] = \mathrm{E}\big[X - \mathrm{E}[X])^2\big]$$

It is the expected value of the square difference between the random variable and its expected value (mean). For computational ease, it is often useful to work out a simpler expression:

$$\begin{aligned}\mathrm{Var}[X] &= \mathrm{E}\Big[(X-\mathrm{E}[X])^2\Big] = \mathrm{E}\Big[x^2 - 2\mathrm{E}[X]X + \mathrm{E}[X]^2\Big] \\ &= \mathrm{E}\Big[X^2\Big] - 2\mathrm{E}[X]\mathrm{E}[X] + \mathrm{E}[X]^2 = \mathrm{E}\Big[X^2\Big] - \mathrm{E}[X]^2.\end{aligned}$$

Example 4.9

Suppose X is a discrete random variable with the PMF defined by the table below.

Table 42: Discrete PMF for Example 4.9

x	$f_X(x)$
1	0.2
2	0.5
3	0.3

Source: George Dekermenjian (2019).

Compute:

a) $\mathrm{E}[X]$.
b) $\mathrm{E}[X]^2$.
c) $\mathrm{Var}[X]$.

Solution 4.9

a) Following the definition, we have $\mathrm{E}[X] = 1 \cdot 0.2 + 2 \cdot 0.5 + 3 \cdot 0.3 = 2.1$.
b) Similarly, $\mathrm{E}\Big[X^2\Big] = 1^2 \cdot 0.2 + 2^2 \cdot 0.5 + 3^2 \cdot 0.3 = 4.9$.
c) Following the simpler expression for the variance, we have

$$\mathrm{Var}[X] = \mathrm{E}\Big[X^2\Big] - \mathrm{E}[X]^2 = 4.9 - (2.1)^2 = 0.49$$

Sometimes we want to report the standard deviation, which is denoted by σ_{X}, of a random variable X. This is just the square root of the variance.

For Example 4.9, we have $\sigma_{\mathrm{X}} = \sqrt{0.49} = 0.7$.

The variance of a constant $c \in \mathbb{R}$ (non-random) quantity is zero since it has no variation

$$\mathrm{Var}[c] = \mathrm{E}\Big[c^2\Big] - \mathrm{E}[c]^2 = c^2\mathrm{E}(1) - c \cdot c \cdot \mathrm{E}(1) \cdot \mathrm{E}(1) = c^2 - c \cdot c = 0$$

Also, adding a constant $c \in \mathbb{R}$ to a random variable does not affect its variance, since we use the linearity of the expectation value

$$\begin{aligned}\mathrm{Var}[X+c] &= \mathrm{E}\Big[(X+c)^2\Big] - \mathrm{E}[X+c]^2 \\ &= \mathrm{E}\Big(X^2 + 2\cdot X\cdot c + c^2\Big) - (\mathrm{E}(X)+c)^2 \\ &= \mathrm{E}\Big(X^2\Big) + 2\mathrm{E}(X)\cdot c + c^2 - \Big(\mathrm{E}(X)^2 + 2\cdot c\cdot \mathrm{E}(X) + c^2\Big) \\ &= \mathrm{Var}(X)\,.\end{aligned}$$

If we multiply a random variable by a constant, then its variance may change. Let $a \in \mathbb{R}$ be a constant, then

$$\begin{aligned}\mathrm{Var}[aX] &= \mathrm{E}\Big[(aX)^2\Big] - \mathrm{E}[aX]^2 \\ &= \mathrm{E}\Big[a^2X^2\Big] - (a\mathrm{E}[X])^2 \\ &= a^2\mathrm{E}\Big[X^2\Big] - a^2\mathrm{E}[X]^2 \\ &= a^2\Big(\mathrm{E}\Big[X^2\Big] - \mathrm{E}[X]^2\Big) \\ &= a^2\mathrm{Var}[X]\end{aligned}$$

Putting these two properties together, we have for constants $a, c \in \mathbb{R}$

$$\mathrm{Var}[aX+c] = a^2\mathrm{Var}[x]\,.$$

Consider the random variable X from Example 4.9, then

$$\mathrm{Var}[3X+10] = 3^2\cdot\mathrm{Var}[X] = 9\cdot 0.49 = 4.41.$$

The computations above are special cases for a general idea of computing the variance of a function of the random variable. In general, for a suitable function g we have

$$\mathrm{Var}[g(X)] = \mathrm{E}\Big[(g(X) - \mathrm{E}[g(X)])^2\Big] = \mathrm{E}\Big[g(X)^2 - \mathrm{E}\Big[g(X)\Big]^2$$

Example 4.10

Let X be a continuous random variable with the density given by

$$f_X(x) = \begin{cases}\frac{x}{2} & \text{for } 0 \le x \le 2, \\ 0 & \text{otherwise.}\end{cases}$$

Compute the following:

a) $\mathrm{Var}[X]$.
b) $\mathrm{Var}[5X]$.
c) $\mathrm{Var}\left[\frac{1}{\left(X^2+1\right)^2}\right]$.

Solution 4.10

a) We start by computing the two expected values we need to compute the variance:

$$\mathrm{E}[X] = \int_0^2 \mathrm{x} \cdot \tfrac{\mathrm{x}}{2} \mathrm{d}x = \tfrac{4}{3}$$

and

$$\mathrm{E}\big[X^2\big] = \int_0^2 x^2 \cdot \tfrac{x}{2} \mathrm{d}x = 2.$$

Therefore,

$$\mathrm{Var}[X] = \mathrm{E}[X]^2 - \mathrm{E}[X]^2 = 2 - \left(\tfrac{4}{3}\right)^2 = \tfrac{2}{9}$$

b) From the property of variance, we have

$$\mathrm{Var}[5X] = 5^2 \mathrm{Var}[X] = 25 \cdot \tfrac{2}{9} = \tfrac{50}{9}$$

c) We will start by commuting the expected values we need for the variance

$$\begin{aligned}\mathrm{E}\left[\frac{1}{\left(x^2+1\right)^2}\right] &= \int_0^2 \frac{x}{2\left(x^2+1\right)^2} \mathrm{d}x = \left[-\tfrac{1}{4}\left(x^2+1\right)^{-1}\right]_0^2 = -\tfrac{1}{4}\left(2^2+1\right)^{-1} + \tfrac{1}{4}\left(0^2+1\right)^{-1} \\ &= -\tfrac{1}{20} + \tfrac{1}{4} = \tfrac{1}{5}\,.\end{aligned}$$

Remember that we used the chain rule to compute the primitive function. We use this rule again to obtain

$$\begin{aligned}\mathrm{E}\left[\frac{1}{\left(x^2+1\right)^4}\right] &= \int_0^2 \frac{x}{2\left(x^2+1\right)^4} \mathrm{d}x = \left[-\tfrac{1}{12}\left(x^2+1\right)^{-3}\right]_0^2 \\ &= -\tfrac{1}{12}\left(2^2+1\right)^{-3} + \tfrac{1}{12}\left(0^2+1\right)^{-3} = = -\tfrac{1}{1500} + \tfrac{1}{12} = \tfrac{31}{375}\end{aligned}$$

Finally, we have

$$\mathrm{Var}\left[\frac{1}{\left(\mathrm{x}^2+1\right)^2}\right] = \tfrac{1}{5} - \left(\tfrac{31}{375}\right)^2 = \tfrac{27164}{140625}.$$

Note that the in the last example we computed the variance of a random variable X and its PDF (which also works in the same way for PMFs). If we replace X by two given random variables and its joint PDF (or PMF) we can also compute the variance in the same way. We only have to replace the PDF of X with the joint PDF of the two given random variables in the computation of the expected values.

In addition, we can also compute for a suitable function $g: \mathbb{R}^2 \to \mathbb{R}$ the variance of the expected value of $g(X, Y)$, where X and Y are two random variables. Then we

$$\begin{aligned}\mathrm{Var}[g(X,Y)] &= \mathrm{E}\Big[\big(g(X,Y) - \mathrm{E}[g(X,Y)]\big)^2\Big] \\ &= \mathrm{E}\Big[g(X,Y)^2\Big] - \mathrm{E}[g(X,Y)]^2.\end{aligned}$$

We will start with the easiest case when $g(X, Y) = X + Y$ and assume that X and Y are independent, then we can simplify the variance in the following form

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y].$$

In the above equality, we worked on the basis that $E(X)E(Y) = E(XY)$ holds for the independent random variables X and Y. Now suppose that $\mathrm{a}, \mathrm{b} \in \mathbb{R}$ are constants and X and Y are still independent. Then we have

$$\mathrm{Var}[aX + bY] = a^2\mathrm{Var}[X] + b^2\mathrm{Var}[Y].$$

If X and Y are dependent, then we can quantify their dependence by computing the covariance of X and Y. The covariance measures the average product of the differences in X and Y from their respective expected values. The covariance is defined by

$$\mathrm{Cov}(X, Y) = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])]$$

for given random variables X and Y.

This definition of the covariance is quite cumbersome to use in practice so we will introduce a different expression:

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

This expression gives an alternative interpretation of the covariance. We can say that the covariance measures the difference between the expected product and the product of the expectations. What happens to the covariance when X and Y are independent? In this case, we already know that $\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y]$; therefore, $\mathrm{Cov}(X, Y) = 0$ for independent random variables X and Y.

It is important to note that the other direction does not hold if $\mathrm{Cov}(X, Y) = 0$ does not imply that X and Y are independent.

Example 4.11

The table below gives the joint PMF of (X, Y). Show that the covariance $\mathrm{Cov}(X, Y) = 0$, but X and Y are not independent.

Table 43: The Joint PMF for Example 4.11

		X			f_Y
		0	1	2	
Y	0	1/3	0	1/3	2/3
	1	0	1/3	0	1/3
f_X		1/3	1/3	1/3	1

Source: George Dekermenjian (2019).

Solution 4.11

We first compute the expected values:

$$\begin{aligned}\mathrm{E}[X]&=0\left(\tfrac{1}{3}\right)+1\left(\tfrac{1}{3}\right)+2\left(\tfrac{1}{3}\right)=1\\ \mathrm{E}[Y]&=0\left(\tfrac{2}{3}\right)+1\left(\tfrac{1}{3}\right)=\tfrac{1}{3}\\ \mathrm{E}[XY]&=0\left(\tfrac{1}{3}+0+0+\tfrac{1}{3}\right)+1\left(\tfrac{1}{3}\right)+2(0)=\tfrac{1}{3}\end{aligned}$$

Therefore,

$$\mathrm{Cov}(X,Y)=\mathrm{E}[XY]-\mathrm{E}[X]\mathrm{E}[Y]=\tfrac{1}{3}-1\cdot\tfrac{1}{3}=0$$

To show that X and Y are not independent, observe that

$$P(X=0,Y=0)=\tfrac{1}{3}\neq P(X=0)\cdot P(Y=0)=\tfrac{2}{3}\cdot\tfrac{1}{3}=\tfrac{2}{9}$$

Example 4.12

Consider the random variables from Example 4.7 whose PDF was given by

$$f_{X,Y}(x,y)=\begin{cases}\frac{1}{3}(4x+2y) & \text{for } 0\leq x\leq 1 \text{ and } \quad 0\leq y\leq 1,\\ 0 & \text{otherwise.}\end{cases}$$

Compute the covariance $\mathrm{Cov}(X,Y)$.

Solution 4.12

From the computation above, $\mathrm{Cov}(X,Y)=\mathrm{E}[XY]-\mathrm{E}[X]\mathrm{E}[Y]$. Luckily, we have already computed all three of these expectations in the example 4.7, which were $\mathrm{E}(X)=\frac{11}{18}$ and $\mathrm{E}(Y)=\frac{5}{9}$, so we only have to compute

$$\begin{aligned}\mathrm{E}(XY) &= \int_0^1\int_0^1 \tfrac{xy}{3}(4x+2y)\mathrm{d}x\mathrm{d}y = \int_0^1\int_0^1 \tfrac{4x^2y}{3} + \tfrac{2xy^2}{3}\mathrm{d}x\mathrm{d}y \\ &= \int_0^1 \tfrac{4y}{9} + \tfrac{2y^2}{6}\mathrm{d}y = \tfrac{4}{18} + \tfrac{2}{18} = \tfrac{1}{3},\end{aligned}$$

and obtain

$$\mathrm{Cov}(X,Y) = \tfrac{1}{3} - \tfrac{11}{18}\cdot\tfrac{5}{9} = -\tfrac{1}{162}$$

We are now ready to consider the quantity $\mathrm{Var}[X+Y]$ for any two random variables X and Y not necessarily independent:

$$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}(X,Y)$$

Now let $\mathrm{a}, \mathrm{b} \in \mathbb{R}$ be constants. We compute $\mathrm{Cov}(aX, bY) = ab\mathrm{Cov}(X,Y)$. From this property, together with the result immediately above, we have a useful property formula for the variance of linear combinations of random variables:

$$\mathrm{Var}(aX+bY) = a^2\mathrm{Var}[X] + b^2\mathrm{Var}[Y] + 2ab\mathrm{Cov}(X,Y)$$

Example 4.13

Consider the random variables from Examples 4.7. Compute the following

a) $\mathrm{Var}[X+Y]$.
b) $\mathrm{Var}[2X+3Y]$.

Solution 4.13

a) In example 4.12 we already computed $\mathrm{Cov}(X,Y) = -\frac{1}{162}$. So all we need are the variances. To that end, we first compute

$$\begin{aligned}\mathrm{E}\big[X^2\big] &= \tfrac{1}{3}\int_0^1\int_0^1 (4x^3+2x^2y)\mathrm{d}x\mathrm{d}y = \tfrac{1}{3}\int_0^1 \big[x^4 + \tfrac{2}{3}x^3y\big]_{x=0}^{x=1}\mathrm{d}y \\ &= \tfrac{1}{3}\int_0^1 \big[\big(1+\tfrac{2}{3}y\big) - (0)\big]\mathrm{d}y \\ &= \tfrac{1}{3}\int_0^1 \big(1+\tfrac{2}{3}y\big)\mathrm{d}y = \tfrac{1}{3}\big[y + \tfrac{1}{3}y^2\big]_{y=0}^{y=1} = \tfrac{1}{3}\big(1+\tfrac{1}{3}\big) = \tfrac{4}{9}.\end{aligned}$$

and in a similar way we obtain

$$\begin{aligned}\mathrm{E}\big[Y^2\big] &= \int_0^1\int_0^1 \tfrac{y^2}{3}(4x+2y)\mathrm{d}x\mathrm{d}y = \int_0^1\int_0^1 \tfrac{4xy^2}{3} + \tfrac{2y^3}{3}\mathrm{d}x\mathrm{d}y \\ &= \int_0^1 \tfrac{4y^2}{6} + \tfrac{2y^3}{3}\mathrm{d}y \quad = \tfrac{4}{18} + \tfrac{2}{12} = \tfrac{7}{18}\end{aligned}$$

From Solution 4.7, we know that $E[X] = \frac{11}{18}$ and $E[Y] = \frac{5}{9}$.

Therefore we obtain

$$\mathrm{Var}[X] = \tfrac{4}{9} - \left(\tfrac{11}{18}\right)^2 = \tfrac{23}{324}$$

and

$$\mathrm{Var}[Y] = \tfrac{7}{18} - \left(\tfrac{5}{9}\right)^2 = \tfrac{13}{162}.$$

We now have all the quantities we need to compute the desired variance.

$$\mathrm{Var}[X+Y] = \tfrac{23}{324} + \tfrac{13}{162} + 2\left(-\tfrac{1}{162}\right) = \tfrac{5}{36}$$

b) The formula for the variance of a linear combination of random variables gives

$$\begin{aligned}\mathrm{Var}[2X+3Y] &= 2^2\mathrm{Var}[X] + 3^2\mathrm{Var}[Y] + 2(2)(3)\mathrm{Cov}(X,Y)\\ &= 4\left(\tfrac{23}{324}\right) + 9\left(\tfrac{13}{162}\right) + 12\left(-\tfrac{1}{162}\right) = \tfrac{151}{162}\,.\end{aligned}$$

4.3 Expectations and Variances of Important Probability Distributions

In this section, we will apply the definitions of expectation and variance to important univariate and multivariate distributions.

Discrete Probability Distributions

The Bernouilli distribution

Let $X \sim \mathrm{Bernoulli}(p)$ for $0 < p < 1$, then its PMF is given by

$$f_x(x) = \begin{cases} 1-p & x = 0, \\ p & x = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Then the expected value is $\mathrm{E}(X) = p$ and the variance is $\mathrm{Var}[X] = p(1-p)$.

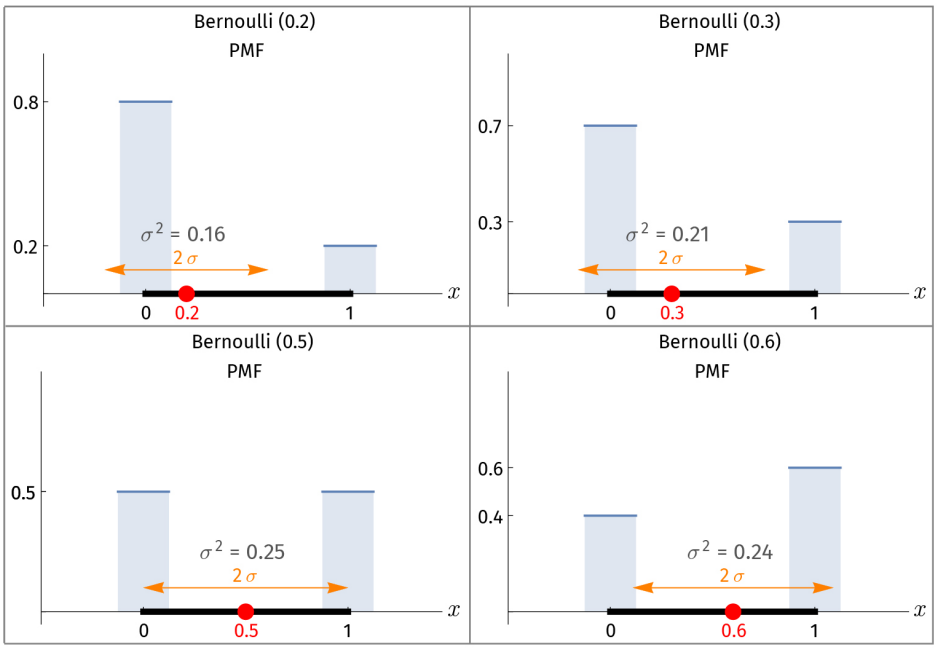
Example 4.14

Suppose that in one day the price of a share will increase by $1 with a probability of 0.2 or stay the same with a probability of 0.8. What is the expected change in the stock price over one day?

Solution 4.14

Let X denote the change in the stock price in dollars over one day. Then $X \sim \text{Bernoulli}(0.2)$; therefore, $\text{E}[X] = 0.2$.

Figure 54: Various Bernoulli PMFs Together with Their Mean and Variance



Source: George Dekermenjian (2019).

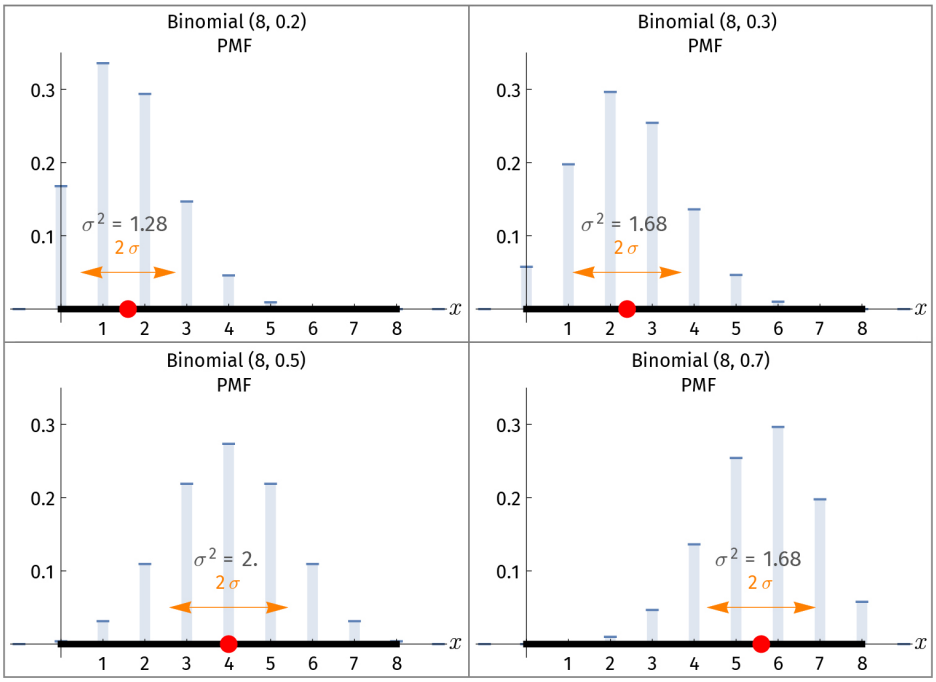
The binomial distribution

Let $X \sim \text{Binomial}(n, p)$ for $n \in \mathbb{N}$ and $0 < p < 1$, so that its PMF is given by

$$f_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0,1, \ldots, n, \\ 0 & \text{otherwise.} \end{cases}$$

The expected value is $\text{E}[X] = np$ and the variance is $\text{Var}[X] = np(1-p)$.

Figure 55: Various Binomial PMFs Together with Their Mean and Variance



Source: George Dekermenjian (2019).

Example 4.15

Suppose a multiple-choice exam has 10 questions and each question has only one correct answer, but three other incorrect choices are provided. If a student guesses on every question, what is the expected number of questions they will guess correctly? Interpret the answer.

Solution 4.15

Let X be the number of questions the student will guess correctly. Since they will guess at random for each question, the probability that they get any one question correct is $p = \frac{1}{4}$ and there are $n = 10$ questions. Therefore, $X \sim \text{Binomial}\left(10, \frac{1}{4}\right)$.

We know that $\text{E}[X] = 10 \cdot \frac{1}{4} = \frac{5}{2} = 2.5$. Therefore, we expect the student to get 2.5 questions correct. The interpretation of this number is that if many students all participate in this test and all guess at random, the average number of questions they will answer correctly will be about 2.5.

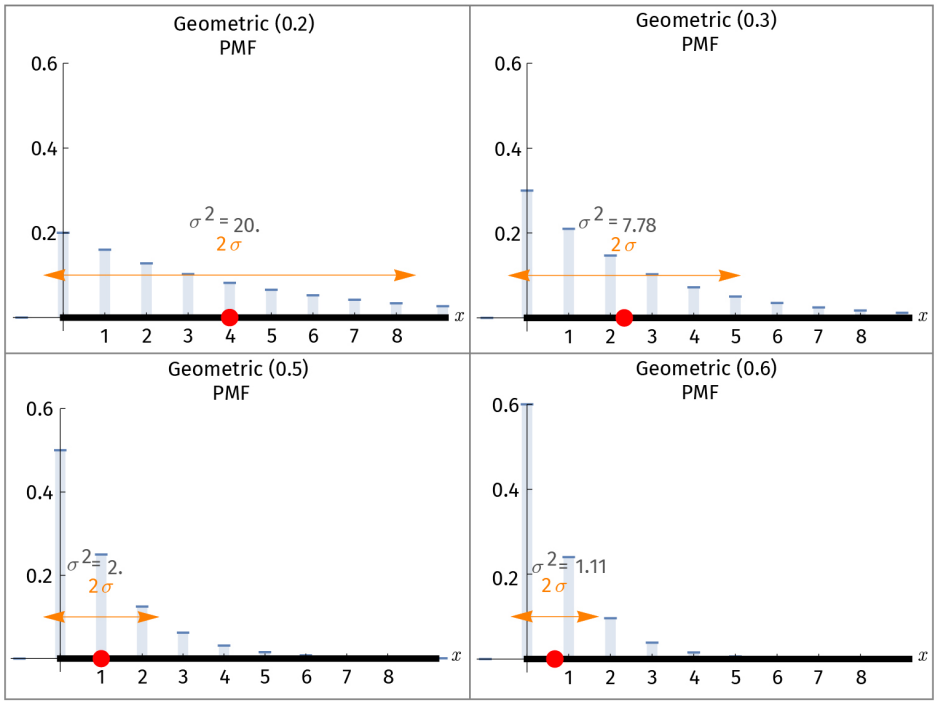
The geometric distribution

Let $X \sim \text{Geometric}(p)$ for $0 < p < 1$, then its PMF is given by

$$f_X(x) = \begin{cases} (1-p)^x p & x = 0,1,2,\ldots, \\ 0 & \text{otherwise.} \end{cases}$$

The expected value is $\mathrm{E}[X]=\frac{1-p}{p}$ and the variance is $\mathrm{Var}[X] = \frac{1-p}{p^2}$.

Figure 56: Various Geomatric PMFs Together with Their Mean and Variance



Source: George Dekermenjian (2019).

Example 4.16

Suppose that 20% of applicants for a certain data science job have advanced knowledge of statistics. Applicants are interviewed in random order, one after the other. What is the expected number of applicants interviewed before the interviewer encounters an applicant with an advanced knowledge of statistics?

Solution 4.16

Let X denote the number of applicants the interviewer talks to before encountering the first applicant with advanced knowledge of statistics. Then, $X \sim \mathrm{Geometric}(0.20)$ where $p = 0.2$ is the probability of "success". The expected value of X is

$$\mathrm{E}[X] = \frac{1-0.2}{0.2} = 4$$

The Poisson distribution

Let $X \sim \text{Poisson}(\lambda)$ for $\lambda > 0$ so that its PMF is

$$f_X(x) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!} & x = 0,1,2,\ldots, \\ 0 & \text{otherwise.} \end{cases}$$

The expected value is $E[X] = \lambda$ and the variance is $Var[X] = \lambda$.

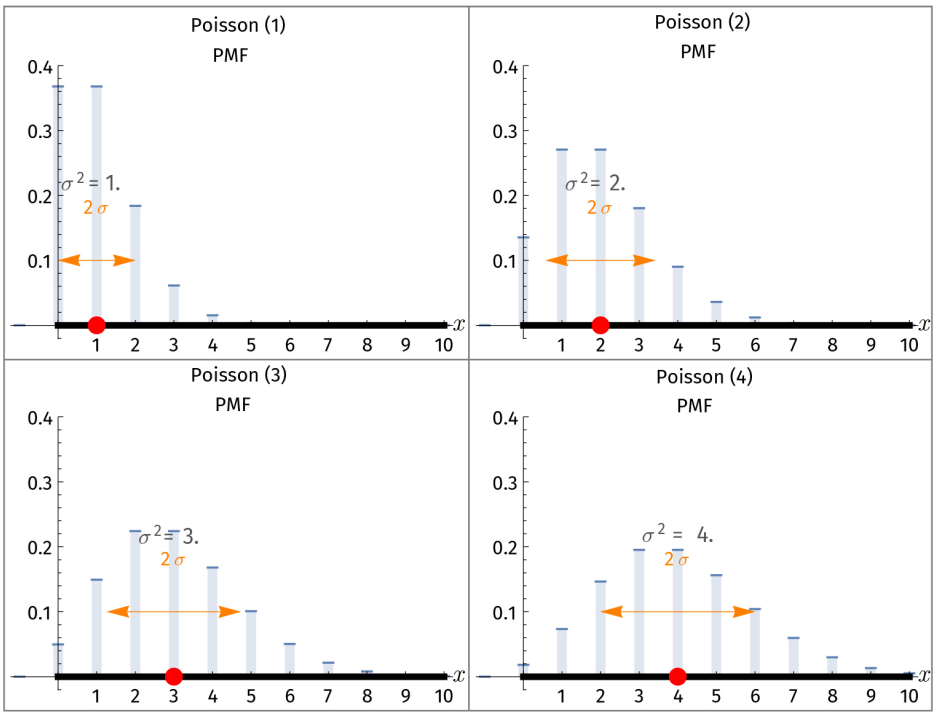
Example 4.17

Suppose that the number of accidents over a period of one day follows a Poisson distribution with a standard deviation of 3. What is the average number of accidents over one day?

Solution 4.17

We know that for the Poisson distribution, the mean is the same as the variance. If $X \sim \text{Poisson}(\lambda),\ \lambda = \text{E}[X] = \text{Var(X)} = 32 = 9$. Therefore, the average number of accidents over one day is 9.

Figure 57: Various Poisson PMFs Together with Their Mean and Variance



Source: George Dekermenjian (2019).

Continuous Probability Distributions

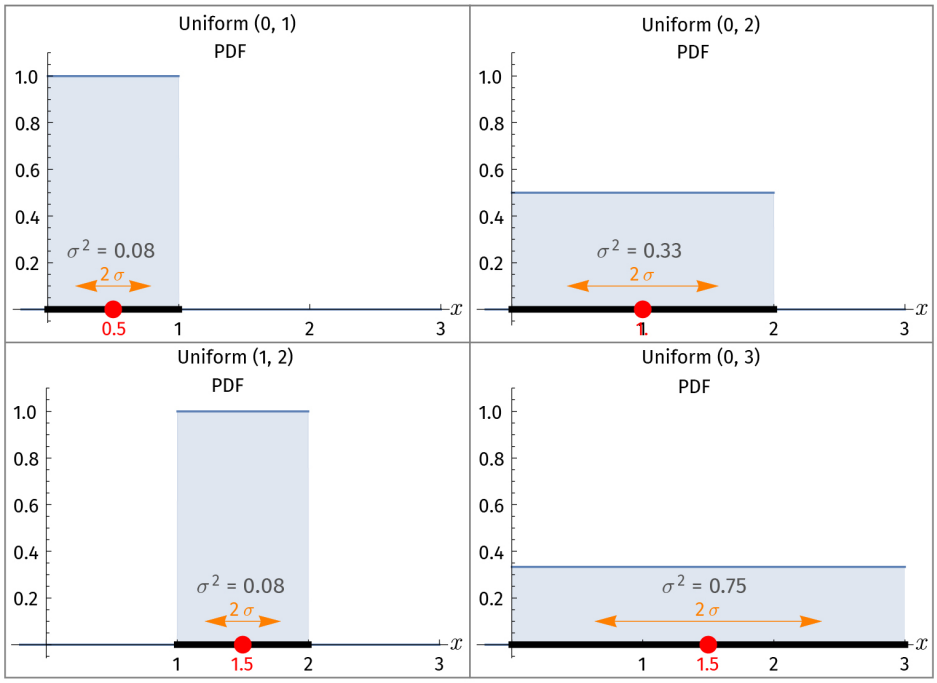
The uniform distribution

Let $X \sim \text{Uniform}([a,b])$ for $a,b \in \mathbb{R}$ such that $a < b$, then its PDF is given by

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b, \\ 0 & \text{otherwise.} \end{cases}$$

The expected value is $\text{E}[X] = \frac{a+b}{2}$ and the variance is $\text{Var}[X] = \frac{(a-b)^2}{12}$.

Figure 58: Various Uniform PDFs Together with Their Mean and Variance



Source: George Dekermenjian (2019).

Example 4.18

What is the expected area of a square whose side length is uniformly distributed on the interval $[1,2]$?

Solution 4.18

Let S be the side of the square. Then $X = S^2$ is the area of this square. We are told that $S \sim \text{Uniform}[1, 2]$. Therefore, the expected area is $\text{E}[X] = \text{E}\left[S^2\right] = \text{Var}[S] + \text{E}[S]^2$. From above, we know that $\text{Var}[S] = \frac{1}{12}$ and $\text{E}[S]^2 = \frac{3}{2}$, so that $\text{E}[S]^2 = \frac{9}{4}$.

Therefore, the expected area (in square units) is

$$\text{E}[X] = \frac{1}{12} + \frac{9}{4} = \frac{7}{3}\,.$$

The exponential distribution

Let $X \sim \text{Exponential}(\lambda)$ for $\lambda > 0$, then its PDF is given by

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

The expectation is $\text{E}[X] = \frac{1}{\lambda}$ and the variance is $\text{Var}[X] = \frac{1}{\lambda^2}$.

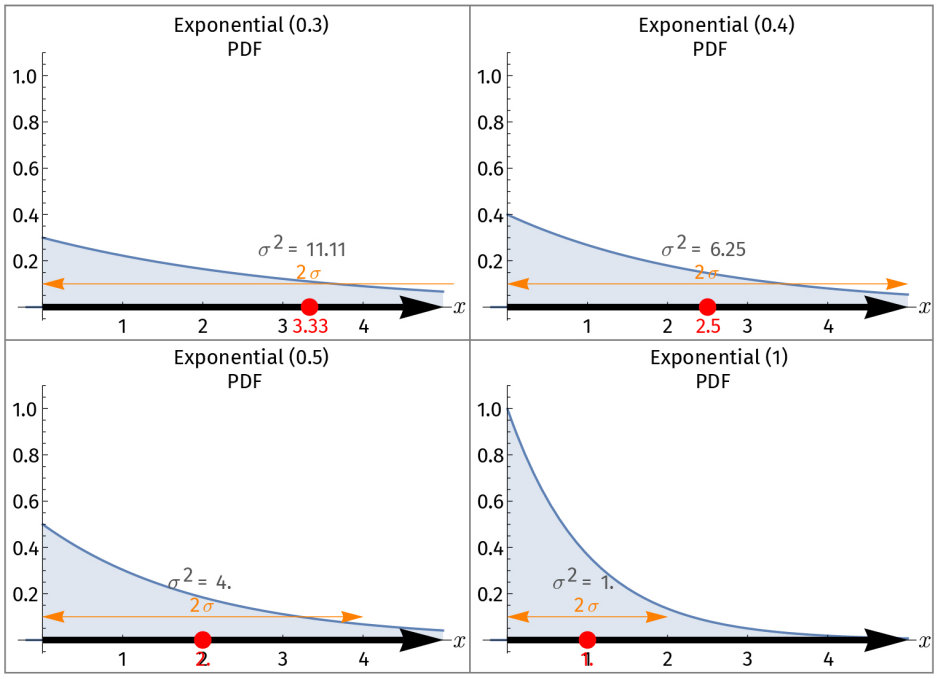
Example 4.19

Suppose that the lifespan of a certain electronic component follows an exponential distribution with a mean of three years. What is the parameter λ of this distribution? What is its interpretation?

Solution 4.19

Let $X \sim \text{Exponential}(\lambda)$. We know that $3 = \text{E}[X] = \frac{1}{\lambda}$. Therefore, $\lambda = \frac{1}{3}$. This means that the rate of failure of this component is 1/3 per year.

Figure 59: Various Exponential PDFs Together with Their Mean and Variance



Source: George Dekermenjian (2019).

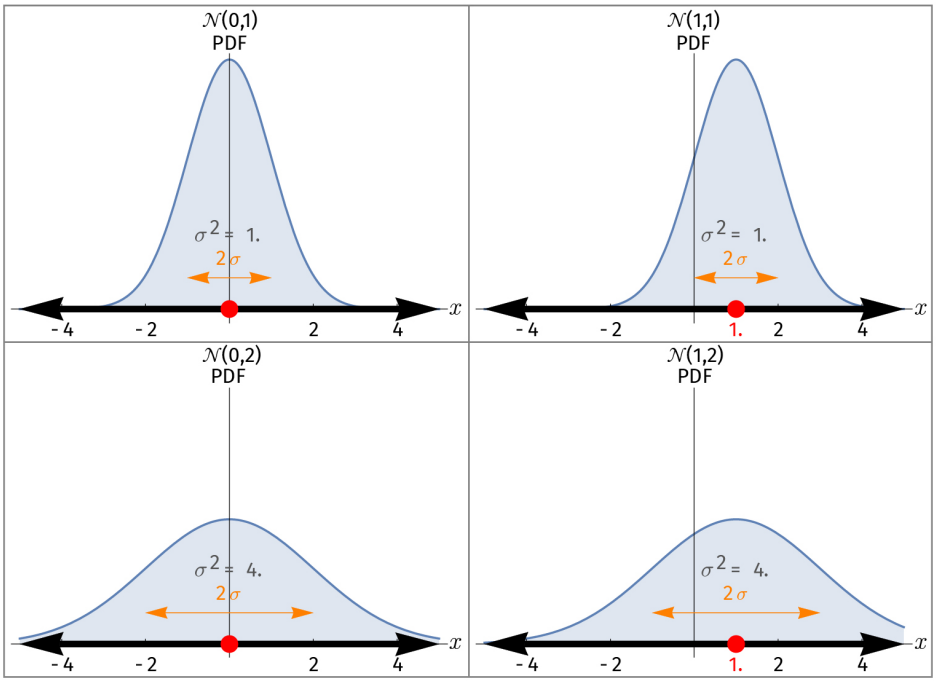
The normal distribution

Let $X \sim \mathcal{N}(\mu, \sigma^2)$, then its PDF is given by

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{for} \quad x \in \mathbb{R},$$

where the expectation is $\mathrm{E}[X] = \mu \in \mathbb{R}$ and the variance is $\mathrm{Var}[X] = \sigma^2$ and $\sigma > 0$.

Figure 60: Various Gaussian PDFs Together with Their Mean and Variance



Source: George Dekermenjian (2019).

Example 4.20

The radius of a circle is normally distributed with mean $\mu = 3$ cm with a standard deviation of $\sigma = 0.1$ cm. What is the expected area of this circle?

Solution 4.20

Let R be the radius of the circle, then the area is $X = \pi R^2$. We are told that $\mathrm{R} \sim \mathcal{N}\left(3, 0.1^2\right)$. The expected area is $\mathrm{E}[X] = \pi \mathrm{E}\left[R^2\right] = \pi\left(\mathrm{Var}[R] + \mathrm{E}[R]^2\right) = \pi\left(0.1^2 + 3^2\right) = 9.01\pi \ cm^2$ or about $28.31 \ \mathrm{cm}^2$.

The Student's T distribution

Let $X \sim T(v)$ where the freedom of degree is $v > 0$. Its PDF is therefore given by

$$f_x(x) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{\pi v}\Gamma\left(\frac{v}{2}\right)}\left(1 + \frac{x^2}{v}\right)^{-\frac{v+1}{2}} \quad \text{for } x \in \mathbb{R},$$

where the gamma function $\Gamma(\cdot)$ is defined on the positive numbers $z > 0$ by

$$\Gamma(z) = \int_0^{\infty} x^{z-1} \mathrm{e}^{-x} \mathrm{d}x$$

The expectation is $\mathrm{E}[X] = 0$ for $v > 1$, otherwise is not defined and the variance is $\mathrm{Var}[X] = \frac{v}{v-2}$ for $v > 2$, otherwise it is undefined.

Example 4.21

The standardized sample mean exam score of a group of 12 students follows a Student's T distribution with 11 degrees of freedom. What is the standard deviation of the standardized sample mean exam score?

Solution 4.21

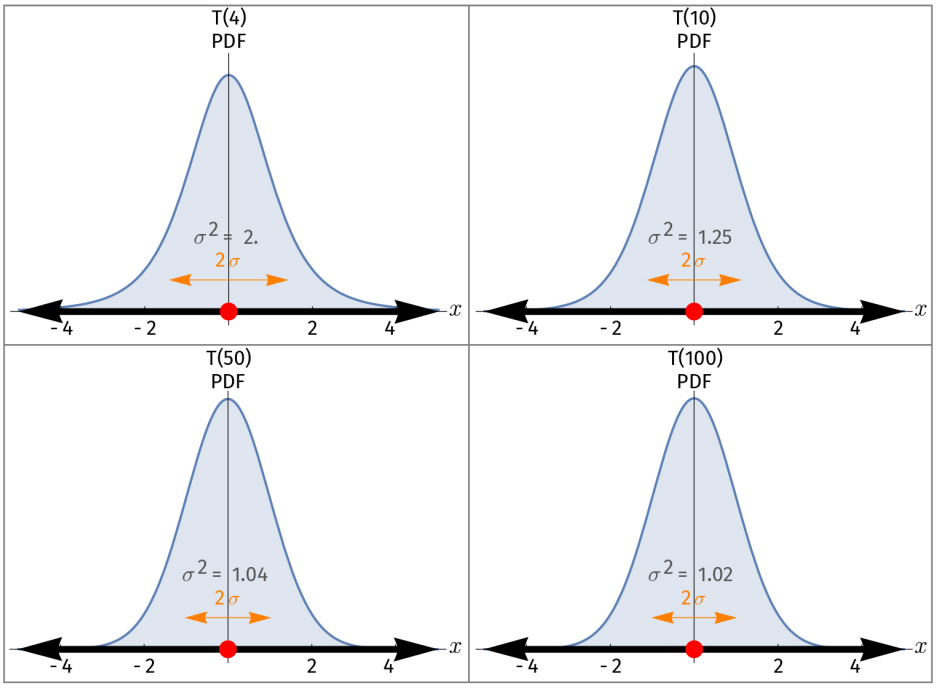
Let Y denote the standardized sample mean exam score of this group of students. We are told that $Y \sim T(11)$ so that

$$\mathrm{Var}[X] = \frac{11}{11-2} = \frac{11}{9}.$$

Therefore, the standard deviation is

$$\sqrt{\frac{11}{9}} = \frac{\sqrt{11}}{3} \approx 1.11.$$

Figure 61: Various Student's T PDFs Together with Their Mean and Variance



Source: George Dekermenjian (2019).

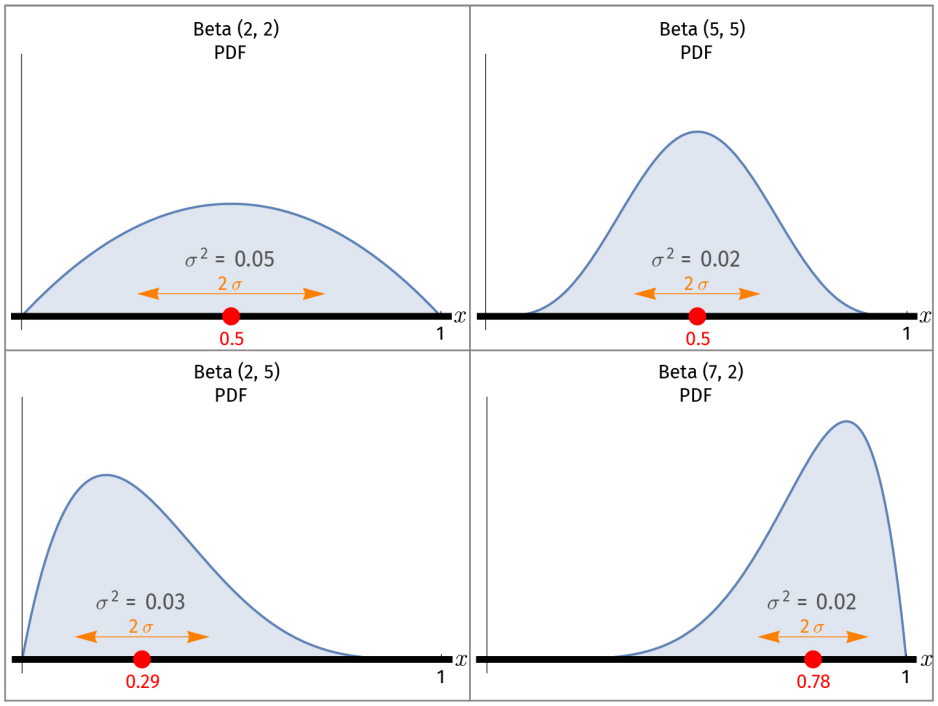
Beta distribution

Let $X \sim \mathrm{Beta}(\alpha, \beta)$ for $\alpha > 0$, and $\beta > 0$, then its PDF is given by

$$f_X(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} & \text{for } 0 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

The expectation is $\mathrm{E}(X) = \frac{\alpha}{\alpha+\beta}$ and the variance is $\mathrm{Var}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

Figure 62: Various Beta PDFs Together with Their Mean and Variance



Source: George Dekermenjian (2019).

Example 4.22

In an automated manufacturing facility that operates 24 hours a day, the proportion of time that a certain machine is operational T is modeled by the Beta(5,2) distribution. The revenue associated with this machine for the 24-hour period is given by $X = 240T + 2T^2$. What is the expected revenue associated with this machine for a 24-hour period?

Solution 4.22

We are asked to find $\mathrm{E}[X]$ where $X = 240T + 2T^2$ with $T \sim \mathrm{Beta}(5,2)$. From the formula above, we know that

$$\mathrm{E}[T] = \frac{5}{5+2} = \frac{5}{7}.$$

In addition, we have

$$\mathrm{Var}[T] = \frac{5 \cdot 2}{(5+2)^2(5+2+1)} = \frac{10}{49 \cdot 8} = \frac{5}{196}.$$

Also,

$$\mathrm{E}\Big[T^2\Big] = \mathrm{Var}[T] + \mathrm{E}[T]^2 = \tfrac{5}{196} + \tfrac{25}{49} = \tfrac{15}{28}.$$

Finally, the expected revenue is given by

$$\mathrm{E}[X] = 240\mathrm{E}[T] + 2\mathrm{E}\Big[T^2\Big] = 240 \cdot \tfrac{5}{7} + 2 \cdot \tfrac{15}{28} = \tfrac{345}{2} = 172.5.$$

4.4 Central Moments

Let X be a random variable. Recall that the variance of X is defined by

$$Var[X] = E\Big[(X - E[X])^2\Big]$$

This quantity is the average squared distance between the random variable and its mean. This is called the second central moment. We say it is the second because of the exponent of 2. This is probably the most widely used central moment because it gives the variance. However, other central moments describe other characteristics of a random variable (distribution). The central moments of orders 0 and 1 are as follows:

- Central Moment of Order 0:$E\Big[(X - E[X])^0\Big] = E[1] = 1$
- Central Moment of Order 1: $E\Big[(X - E[X])^1\Big] = E[X] - E[X] = 0$

In general, the $^{r^{th}}$ central moment for $r > 0$ of a random variable X is given by

$$\mathrm{E}\big[(X \ - \mathrm{E}[X])^r\big] = \mathrm{E}\big[(X - \mu)^r\big],$$

where $\mu = \mathrm{E}(\mathrm{X})$. The third central moment of the $X \sim \mathrm{Binomial}(n, p)$ for $\mathrm{n} \in \mathbb{N}$ and $0 < p < 1$ is given by

$$np(1-p)(1-2p).$$

Skewness

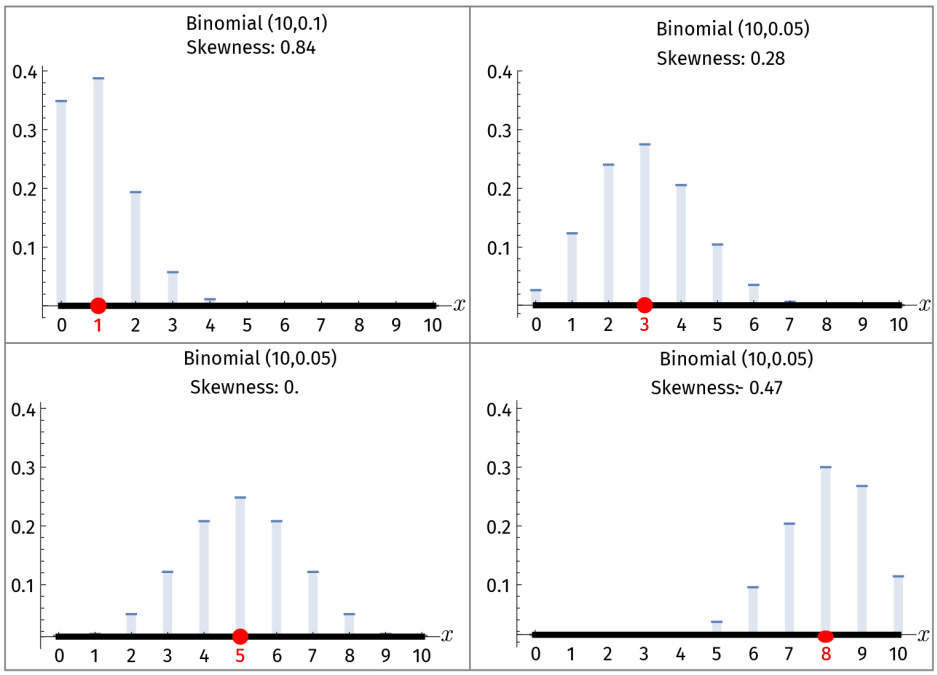
The third central moment $\mathrm{E}\Big[(X - \mathrm{E}(X))^3\Big]$ is related to the skewness of the distribution. As a matter of fact, the skewness of a random variable (or a distribution) is the standardized third central moment given by

$$\mathrm{E}\Big[\Big(\tfrac{X-\mu}{\sigma}\Big)^3\Big]$$

where $\sigma^2 = \mathrm{Var}[X] > 0$ and $\mu = \mathrm{E}(\mathrm{X})$.

We will explore this quantity using the binomial distribution with $n = 10$ and varying values of p. In other words, suppose that $X \sim \text{Binomial}(10, p)$ for $0 < p < 1$. We plot the PMF of the binomial distribution for $p = 0.1, 0.3, 0.5, 0.8$ and compute the corresponding values of the third moment.

Figure 63: Various Binomial PMFs Together with Their Mean and Skewness



Source: George Dekermenjian (2019).

It looks like the third central moment is zero for a symmetric PMF (when $p = 0.5$). When the PMF has a positive skew, the tail on the right is longer and the third moment is positive. When the PMF has a negative skew, the tail is longer on the left, and the third moment is negative. This is exactly what the third moment measures—the departure from symmetry of the random variable (distribution). For example, the normal distribution is symmetric and so we expect the third (as well as any odd) central moment to be zero.

Using the fact that the variance of a binomial distribution is $\sigma^2 = np(1-p)$, we have $\sigma^3 = np(1-p)\sqrt{np(1-p)}$ and that for the third central moment of $X \sim \text{Binomial}(n, p)$ for $n \in \mathbb{N}$ and $0 < p < 1$ we know

$$np(1-p)(1-2p)$$

Thus, we obtain the following formula for the third central moment

$$\text{Skewness} = \frac{\text{E}(X-\mu)^3}{\sigma^3} = \frac{n(1-p)(1-2p)}{np(1-p)\sqrt{np(1-p)}} = \frac{1-2p}{\sqrt{np(1-p)}}$$

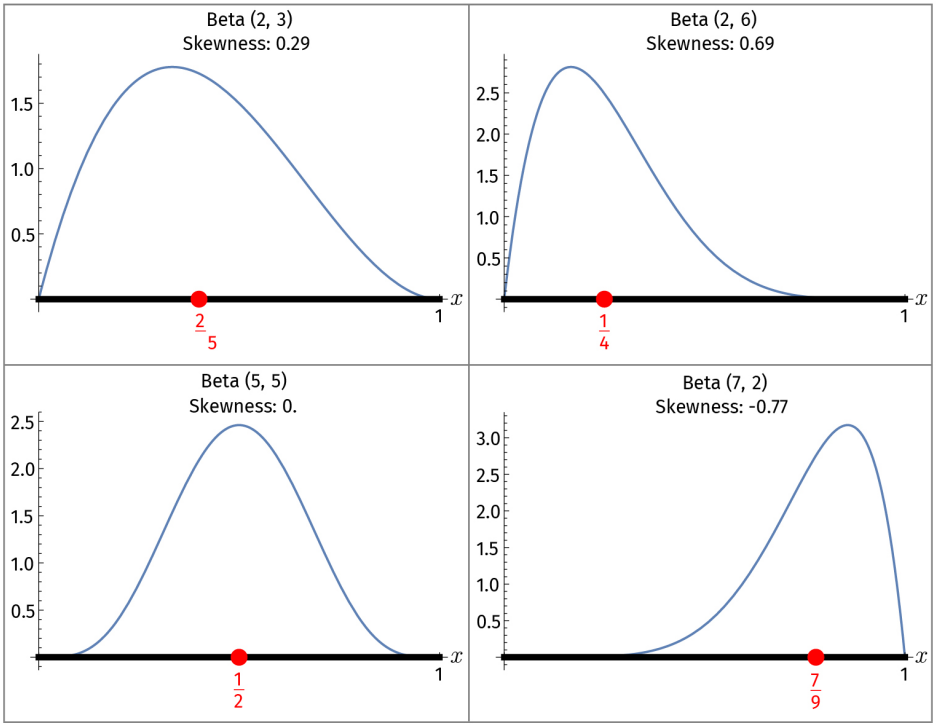
For $n = 10$ and $p = \frac{1}{10}$, this expression evaluates as

$$\text{Skewness} = \frac{1 - \frac{2}{10}}{10\left(\frac{1}{10}\right)\sqrt{\left(\frac{9}{10}\right)}} = \frac{\frac{8}{10}}{\sqrt{\frac{9}{10}}} = \frac{8}{3\sqrt{10}} \approx 0.84\,.$$

For practice, you can verify the skewness of the other binomial distributions shown in the figure above using this formula.

The following graphs show beta distributions along with calculations of skewness.

Figure 64: Various Beta PDFs Together with Their Mean and Skewness

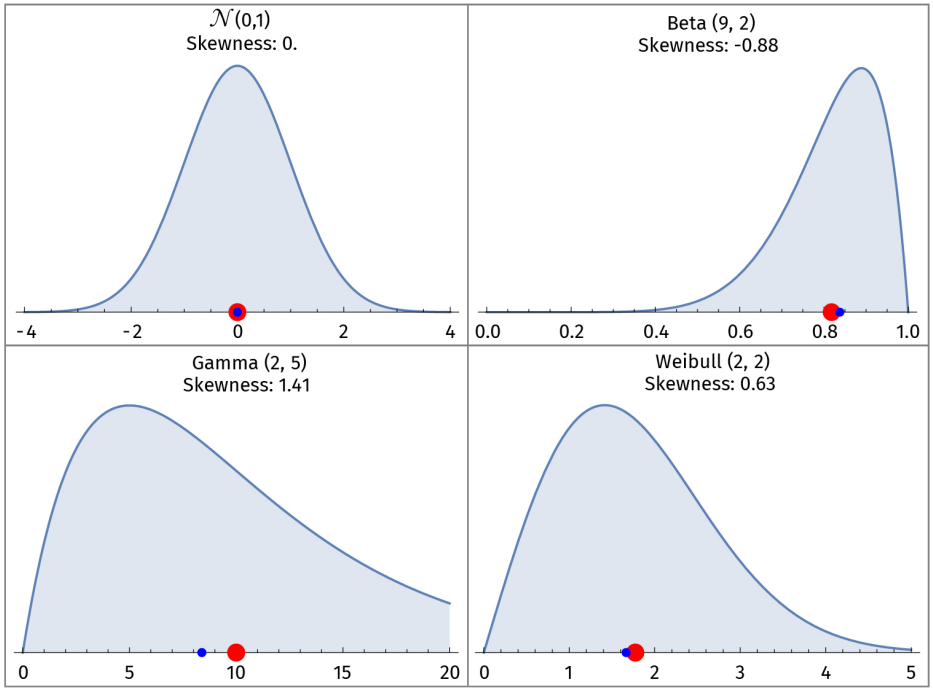


Source: George Dekermenjian (2019).

For specific random variables the relative position of the median to the mean of a distribution can be identified by the skewness or vice versa. If the skewness is positive, this means that it is more likely to observe a value above the mean than below the mean. Therefore, the median is more than the mean for these distributions. If the skewness is negative, this means that it is more likely that values will be observed below the mean than above the mean. This means that the median for these distributions is less than the mean. Finally, as you might have deduced, if the skewness is zero, then the likelihood of observing data above the mean is the same as the values below the mean. Therefore, the median for these distributions coincides with the mean.

The following graphic shows plots of various distributions, with the mean shown in red and the median in blue.

Figure 65: Various Continuous PDFs Together with Their Mean, Median and Skewness



Source: George Dekermenjian (2019).

Kurtosis

The fourth central moment $\mathrm{E}\left[(X-\mu)^4\right]$ is related to another important characteristic of a distribution. This characteristic is the extent to which we are likely to observe extreme values. More concretely, the standardized version of the fourth central moment, kurtosis, is given by

$$\mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right],$$

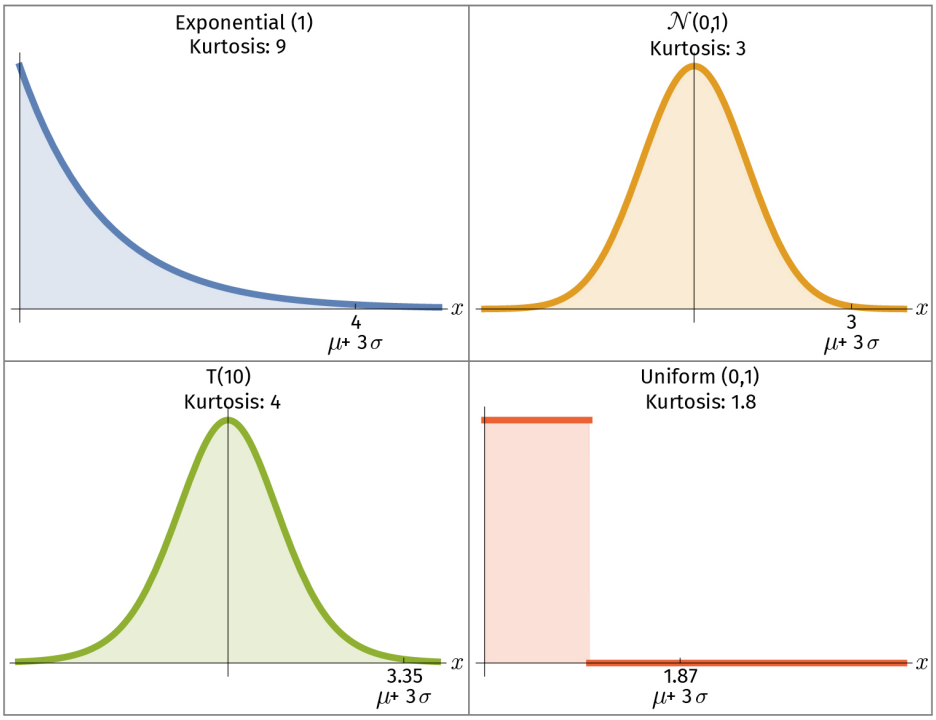
where $\sigma^2 = \mathrm{Var}[X] > 0$ and $\mathrm{E}(\mathrm{X}) = \mu \in \mathbb{R}$. If we look at the values of X within less than one standard deviation of the mean, then the quantity

$$\frac{X-\mu}{\sigma}$$

is less than one in absolute value. Raising this quantity to the fourth power makes it quite small. Therefore, the quantities that contribute most to this expectation are those values of X that are far away from the mean (more than one standard deviation). For instance, the values of X more than three standard deviations from the mean raised to the fourth

power give more than nine. If we have many such values, then the expected value is affected correspondingly. The normal distribution has a kurtosis equal to three. This is the usual baseline to which we compare the kurtosis of other distributions.

Figure 66: Various Continuous PDFs Together with Their Kurtosis



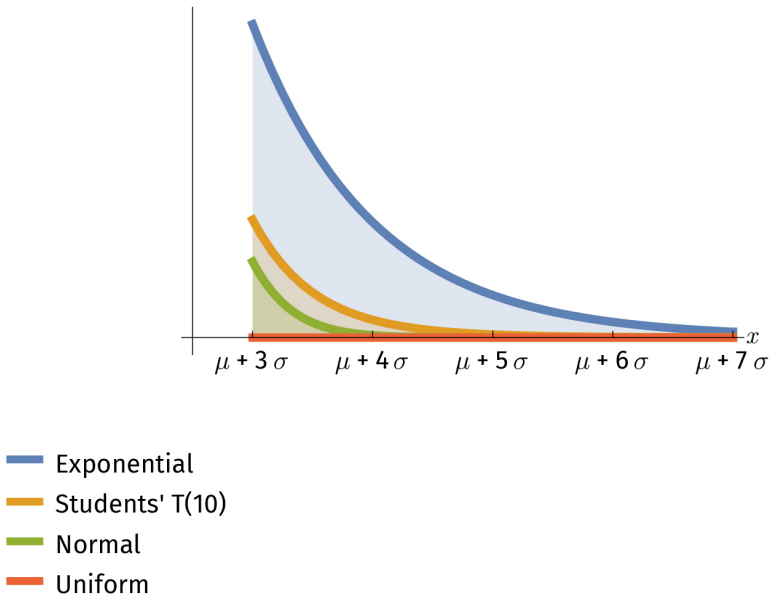
Source: George Dekermenjian (2019).

On the one extreme, we have the exponential distribution, which has a kurtosis equal to nine. This distribution has a very long right tail. This means that it is more likely to observe outliers than with the normal distribution. Therefore, it is no surprise that its kurtosis is larger than three. Let us make this a little more concrete. The probability of observing values that are more than three standard deviations away for the normal distribution is about 0.14%, but for the exponential distribution this probability is about 1.83%. In other words, it is about 14 times more likely to observe outliers from the exponential distribution than from the normal distribution!

On the other extreme, we have the uniform distribution. Consider that the standard deviation of $X \sim \text{Uniform}[0,1]$ is about 0.29. To observe values that are more than three standard deviations larger than the mean of 0.5 would be to observe values more than about 1.37, which is impossible. Therefore, it is unlikely that we can observe outliers from a uniform distribution, and so it is no surprise that its kurtosis is less than three.

Here are some plots of the same four distributions showing the area beyond three standard deviations from the mean. The distributions are adjusted so that they all have the same mean and standard deviation.

Figure 67: Tail Behavior of Various Continuous PDFs with Equal Mean and Variance



Source: George Dekermenjian (2019).

4.5 Moment Generating Functions

The mean and variance of a distribution are two important properties of a distribution. However, as we saw in the previous section, they are not enough to describe the whole distribution. For instance, two distributions can have the same mean and variance but different kurtosis. The mean is $\mathrm{E}[X]$. The variance is related to $\mathrm{E}\left[X^2\right]$ and $\mathrm{E}[X]$. The skewness is related to $\mathrm{E}\left[X^3\right]$, $\mathrm{E}\left[X^2\right]$, and $\mathrm{E}[X]$. Kurtosis is related to skewness and $\mathrm{E}\left[X^4\right]$. These expected values of powers of X are called the moments of the random variable (or corresponding distribution). We can compute these directly from the definition of the expected value, which is quite cumbersome, to say the least. Wouldn't it be wonderful if there were a function that gave us these moments without requiring summation or integration? Just such a function is the main subject of this section.

The **moment generating function** (mgf) of a random variable X is

$$\mathrm{mgf}(t) = \mathrm{E}\left[et^X\right] \text{ for } t \in \mathbb{R}$$

Moment generating function
A random variable, which gives an alternative route for the computation of

expectation values, variances, probability measures and so on.

The moment generating function of a random variable, where it exists, completely identifies the distribution of the random variable. So, in a way, it can take the place of the PDF (PMF) or CDF to derive the characteristics of a random variable. As the name suggests, it is very useful for computing the moments of a random variable-for a given random variable X the first moment is denoted by

$$\mu^{(1)} = \mathrm{E}[X],$$

the second moment is

$$\mu^{(2)} = \mathrm{E}\big[X^2\big],$$

and so on. The r-th moment is

$$\mu^{(r)} = \mathrm{E}[X^r] \text{ for } \ r > 0$$

We know from calculus that for $\mathrm{z} \in \mathbb{R}$

$$\mathrm{e}^{\mathrm{z}} = 1 + \mathrm{z} + \frac{\mathrm{z}^2}{2!} + \frac{\mathrm{z}^3}{3!} + \ \ldots$$

Therefore for $\mathrm{t} \in \mathbb{R}$ and a given random variable X it holds

$$\mathrm{e}^{tX} = 1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + \ \ldots$$

Taking expectations, we have

$$\mathrm{mgf}(t) = 1 + tE[X] + \frac{t^2}{2!}\mathrm{E}\big[X^2\big] + \frac{t^3}{3!}\mathrm{E}\big[X^3\big] + ...$$

or

$$\mathrm{mgf}(t) = 1 + t\mu^{(1)} + \frac{t^2}{2!}\mu^{(2)} + \frac{t^3}{3!}\mu^{(3)} + \ \ldots$$

Differentiating once with respect to t yields

$$\frac{\mathrm{d}}{dt}\mathrm{mgf}(t) = \mu^{(1)} + \frac{t}{1!}\mu^{(3)} + \ \ldots$$

and evaluating this derivative at $t = 0$ gives

$$\Big[\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{mgf}(t)\Big]_{t\,=\,0} = \mu^{(1)} = \mathrm{E}[X]$$

which is the first moment of X. Taking the second derivative gives

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathrm{mgf}(t) = \mu^{(2)} + \frac{t}{1!}\mu^{(3)} + \ \ldots$$

Evaluating again at $t = 0$ gives the second moment

$$\left[\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathrm{mgf}(t)\right]_{t\,=\,0} = \mu^{(2)} = \mathrm{E}\big[X^2\big]\,.$$

We define the r^{th} derivative of the moment generating function of X by $\mathrm{mgf}^{(r)}(t)$ for $r \in \mathbb{N}$.

Then, we obtain

$$\mathrm{mgf}^{(r)}(0) = \mu^{(r)} = \mathrm{E}[X^r]$$

where $E[X^r]$ is called the r^{th} moment of X. In what follows, we compute the moment generating functions for some of the most important distributions.

Example 4.23

The PMF of X is given by $f(0) = 0.1$, $f(1) = 0.3$, $f(2) = 0.6$, and zero otherwise.

a) Find the moment generating function of X.
b) Find the first moment, $\mu^{(1)} = \mathrm{E}[X] \text{ of } X$, using the moment generating function.
c) Compute $E[X]$ directly from the definition. Verify that your answer agrees with part b.

Solution 4.23

a) $\mathrm{mgf}(t) = \mathrm{E}\big[\mathrm{e}^{tX}\big] = 0.1\mathrm{e}^{t\,\cdot\,0} + 0.3\mathrm{e}^{t\,\cdot\,1} + 0.6\mathrm{e}^{t\,\cdot\,2} = 0.1 + 0.3\mathrm{e}^{t} + 0.6\mathrm{e}^{2t}$.
b) $\mathrm{mgf}'(t) = 0.3\mathrm{e}^{t} + 2 \cdot 0.6\mathrm{e}^{2t} = 0.3\mathrm{e}^{t} + 1.2\mathrm{e}^{2t}\ ,\ \mathrm{mgf}'(0) = 0.3 + 1.2 = 1.5.$
c) $\mathrm{E}[X] = 0.1 \cdot 0 + 0.3 \cdot 1 + 0.6 \cdot 2 = 1.5.$

Example 4.24

The moment generating function of a discrete random variable X is given by

$$\mathrm{mgf}(t) = 0\ .125\ \big(1 + \mathrm{e}^{t}\big)^3 \text{ for } t \in \mathbb{R}\,.$$

a) Compute $\mathrm{E}[X]$.
b) Compute $\mu^{(2)} = \mathrm{E}\big[X^2\big]$.

Solution 4.24

a) First of all, we compute the first derivative obtaining

$$\mathrm{mgf}'(t) = 3 \cdot 0.125\big(1 + \mathrm{e}^{t}\big)^2\mathrm{e}^{t} = 0.375\big(1 + \mathrm{e}^{t}\big)^2\mathrm{e}^{t}\,.$$

Thus we get

$$\mathrm{E}[X] = \mathrm{mgf}'(0) = 0.375(1+1)^2 = 0.375 \cdot 4 = 1.5\,.$$

b) Now let us compute the second derivative

$$\begin{aligned}\mathrm{mgf}''(t) &= 2 \cdot 0.375\left(1+\mathrm{e}^t\right)\mathrm{e}^t + 0.375\left(1+\mathrm{e}^t\right)^2\mathrm{e}^t = \\ &= 0.375(1+\mathrm{e}^t)\mathrm{e}^t(2\mathrm{e}^t + 1 + \mathrm{e}^t) = 0.375(1+\mathrm{e}^t)\mathrm{e}^t(1+3\mathrm{e}^t)\,.\end{aligned}$$

Now we get

$$\mu^{(2)} = \mathrm{E}\left[X^2\right] = \mathrm{mgf}''(0) = 0.375(2)(1)(4) = 3.$$

Example 4.25

Suppose that X has a moment generating function given by $\mathrm{mgf}(t) = \mathrm{e}^{2t^2}$. What is the variance $\mathrm{Var}[X]$?

Solution 4.25

We need to compute the first and second moments so that we can compute the following

$$\mathrm{Var}[X] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 = \mu^{(2)} - \mu^{(1)^2}.$$

Therefore, we need the first and the second derivative of the moment generating function

$$\begin{aligned}\mathrm{mgf}'(t) &= 4t\mathrm{e}^{2t^2},\ \text{so } \mu^{(1)} = \mathrm{mgf}'(0) = 0,\\ \mathrm{mgf}''(t) &= 4\mathrm{e}^{2t^2} + 16t^2\mathrm{e}^{2t^2},\ \text{so } \mu^{(2)} = \mathrm{mgf}''(0) = 4\,.\end{aligned}$$

Hence, we obtain $Var[X] = 4 - 0^2 = 4$.

Table 44

Distribution	Moment generating function
X~Binomial(n, p)	$\mathrm{mgf}_X(t) = E\left[e^{tX}\right] = \left(pe^t + 1 - p\right)^n$
X~Geometric(p)	$\mathrm{mgf}_X(t) = \frac{pe^t}{1 - e^t(1-p)}$ provided $e^t(1-p) < 1$
X~Poisson(λ)	$\mathrm{mgf}_X(t) = e^{-\lambda} exp\left(\lambda e^t\right)$ for $t \in \mathbb{R}$
X~Uniform($[a, b]$)	$\mathrm{mgf}_X(t) = \frac{e^{tb} - e^{ta}}{t(b-a)}$ for $t \neq 0$
X~Exponential(λ)	$\mathrm{mgf}_X(t) = \frac{\lambda}{\lambda - t}$ for $t < \lambda$

Distribution	Moment generating function
$X \sim \mathcal{N}(\mu, \sigma^2)$	$\mathrm{mgf}_X(t) = e^{\frac{\sigma^2 t^2}{2} + t\mu}$ for $t \in \mathbb{R}$
$X \sim \mathrm{Gamma}(\alpha, \beta)$	$\mathrm{mgf}_X(t) = \left(1 - \frac{t}{\beta}\right)^{-\alpha}$ for $t < \beta$

Source: George Dekermenjian (2019).

Example 4.26

Use the moment generating function of the gamma distribution to show that $\mathrm{E}[X] = \frac{\alpha}{\beta}$ for $X \sim \mathrm{Gamma}(\alpha, \beta)$ with $\alpha, \beta > 0$.

Solution 4.26

The moment generating function from the table above is

$$\mathrm{mgf}_X(t) = \left(1 - \frac{t}{\beta}\right)^{-\alpha} \text{for } t < \beta\,.$$

We compute the first derivative:

$$\mathrm{mgf}_X(t) = (-\alpha)\left(-\frac{1}{\beta}\right)\left(1 - \frac{t}{\beta}\right)^{-\alpha - 1} = \left(\frac{\alpha}{\beta}\right)\left(1 - \frac{t}{\beta}\right)^{-\alpha - 1}.$$

Next, the expected value is $\mathrm{E}[X] = \mathrm{mgf}'(0) = \frac{\alpha}{\beta}$.

It is useful to become familiar with the moment generating functions of these important distributions, because one of the most important properties of the moment generating function is, if two random variables have the same moment generating functions, then they must have the same distribution. We can say that the moment generating function determines the distribution. For example, given the moment generating function of a random variable X

$$\mathrm{mgf}_X(t) = \frac{2}{2 - t},$$

then we see that the function $\frac{\lambda}{\lambda - t}$, which is the moment generating function for the exponential, coincides for $\lambda = 2$. Therefore, we have $X \sim \mathrm{Exponential}(2)$. Similarly, if Y is a random variable with moment generating function

$$\mathrm{mgf}_X(t) = \mathrm{e}^{\frac{9t^2}{2}} = \mathrm{e}^{\frac{\sigma^2 t^2 + \mu t}{2}} \text{ for } \mu = 0 \text{ and } \sigma = 3,$$

we see that $Y \sim \mathcal{N}(0, 3^2)$.

One of the most powerful features of the moment generating functions is that they help us work with sums of independent random variables. In particular, let us say that $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are independent. What is the distribution of $Y = X_1 + X_2$?

We will suggest two ways of tackling this problem.

First method

Write down the joint density, which since they x_1 and x_2 independent, is just the product of the marginals. Next, determine the CDF of Y by

$$F_Y(y) = P(X_1 + X_2 < y) = \int_{-\infty}^{\infty} \int_{-\infty}^{y - x_2} f_{x_1, x_2}\left(x_1, x_2\right) \mathrm{d}x_1 \mathrm{d}x_2$$

Once we have the CDF, we can differentiate with respect to y to get the PDF:

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y)$$

Let us just say that this method is not entirely user-friendly!

Second method

Using the fact that a moment generating function entirely identifies the distribution of a random variable, we proceed to find the moment generating function of Y as follows:

$$\begin{aligned} \mathrm{mgf}_Y(t) &= \mathrm{E}\left[\mathrm{e}^{tY}\right] = \mathrm{E}\left[\mathrm{e}^{t(X_1 + X_2)}\right] = \mathrm{E}\left[\mathrm{e}^{tX_1}\mathrm{e}^{tX_2}\right] = \mathrm{E}\left[\mathrm{e}^{tX_1}\right]\mathrm{E}\left[\mathrm{e}^{tX_2}\right] \\ &= \mathrm{mgf}_{X_1}(t)\mathrm{mgf}_{X_2}(t) \end{aligned}$$

where we used the independence of X_1 and X_2 in the last equality. Therefore, we have

$$\mathrm{mgf}_Y(t) = \mathrm{e}^{\left(\frac{\sigma_1^2 t^2}{2} + \mu_1 t\right)} \, \mathrm{e}^{\left(\frac{\sigma_2^2 t^2}{2} + \mu_2 t\right)} = \mathrm{e}^{\frac{(\sigma_1{}^2 + \sigma_1{}^2) t^2}{2} + (\mu_1 + \mu_2) t}.$$

Finally, we see that $\mathrm{Y} \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ must hold. As a matter of fact, we can extend this argument for a sum of an arbitrary number of independent normal random variables.

If $X_1, X_2, \ldots, X_n$ are independent and normally distributed for all $n \in \mathbb{N}$ with means $\mu_1, \mu_2, \ldots, \mu_n \in \mathbb{R}$ and standard deviations $0 < \sigma_1, \sigma_2, \ldots, \sigma_n$**,** respectively, then $S_n = \sum_{i=1}^{n} X_i$ is also normally distributed with mean

$$\mu_S = \sum_{i=1}^{n} \mu_i$$

and standard deviation

$$\sigma_S = \left(\sum_{i=1}^{n} \sigma_i^2\right)^{\frac{1}{2}}.$$

One of the cornerstones of probability theory is the **central limit theorem**. This theorem makes a statement about the sum of independent and identically distributed random variables. This is similar to what we had above, except that all the random variables $X_1, X_2, \ ..., X_n$ would have the same distribution. In the case that they are all normally distributed with $\mathrm{X_i} \sim \mathcal{N}\left(\mu, \sigma^2\right)$ for $i = 1,2, \ ..., n$, the above computation reduces to $\mathrm{S_n} \sim \mathcal{N}\left(\mathrm{n}\mu, \mathrm{n}\sigma^2\right)$.

Central limit theorem
The distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the distribution.

The next related and equally important quantity is that of the sample mean of a set of independent and identically distributed normal variables. The sample mean is

$$\overline{X} = \frac{S_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

From the property of expectation, we know that $\mathrm{E}\left[\overline{X}\right] = \frac{1}{n}\mathrm{E}\left[S_n\right] = \mu$ and from the property of variance we know that $\mathrm{Var}\left[\overline{X}\right] = \frac{1}{n^2}\mathrm{Var}\left[S_n\right] = \frac{\sigma}{n}$. But what about the distribution? Once again, the moment generating function can be used to establish the distribution of a constant multiple of a random variable. Suppose the moment generating function of a random variable is $\mathrm{mgf}_y(t)$, then the moment generating function of cY where $c \in \mathbb{R}$ is a constant is

$$\mathrm{mgf_{cY}}(t) = \mathrm{E}\left[\mathrm{e}^{t(cY)}\right] = \mathrm{E}\left[\mathrm{e}^{(tc)Y}\right] = \mathrm{mgf_Y}(ct)$$

Therefore, if $\mathrm{Y} \sim \mathcal{N}\left(\mu_\mathrm{Y}, \sigma_\mathrm{Y}^2\right)$, its moment generating function is

$$\mathrm{mgf}_Y(t) = \exp\left(\frac{\sigma_Y^2 t^2}{2} + \mu_Y t\right)$$

and so

$$\mathrm{mgf}_{cY}(t) = \exp\left(\frac{\left(c^2\sigma_Y^2\right)t^2}{2} + c\mu_Y\right)$$

In other words, $(\mathrm{cY}) \sim \mathcal{N}\left(\mathrm{c}\mu_Y, \left(\mathrm{c}\sigma_Y\right)^2\right)$. Now going back to the sample mean of independent and identically distributed (i.i.d) normal random variables, we have

$$\overline{\mathrm{X}} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{\mathrm{n}}\right).$$

Actually, it turns out that even if the independent and identically distributed variables are not normally distributed, the sample mean is still approximately normally distributed for large $n \in \mathbb{N}$, and this approximation gets closer to the mean as n tends to infinity. This is essentially what the central limit theorem is about.

Sums of independent random variables, whether they are from the same distribution or not, play an important role in data science. Therefore, it is worth remembering that if $X_1, X_2, \ldots, X_n$ are independent random variables, then the moment generating function of their sum is the product of the individual moment generating functions:

$$\mathrm{mgf}_{X_1+X_2+\cdots+X_n}(t) = \prod_{i=1}^{n} \mathrm{mgf}_{X_i}(t) \text{ for all } t \in \mathbb{R}.$$

Example 4.27

Let $X_1 \sim \mathrm{Exponential}(2)$ and $X_2 \sim \mathrm{Exponential}(5)$ be independent. Write down the moment generating function $Y = X_1 + X_2$.

Solution 4.27

We write down the moment generating functions of X_1 and X_2:

$$\mathrm{mgf}_{X_1}(t) = \frac{2}{2-t} \text{ and } \mathrm{mgf}_{X_2}(t) = \frac{5}{5-t}.$$

Therefore, the mgf of their sum is the product of these two functions

$$\mathrm{mgf}_{Y}(t) = \frac{10}{(2-t)(5-t)}.$$

Example 4.28

Let $X_1 \sim \mathrm{Poisson}(\lambda_1)$ and $X_2 \sim \mathrm{Poisson}(\lambda_2)$ be independent. Use the moment generating function to determine the distribution of $Y = X_1 + X_2$.

Solution 4.28

The moment generating functions of X_1 and X_2 are

$$\mathrm{mgf}_{X_1}(t) = \mathrm{e}^{\lambda_1 + \lambda_2 t} \text{ and } \mathrm{mgf}_{X_2}(t) = \mathrm{e}^{\lambda_2 + \lambda_2 t}$$

Therefore, the moment generating function of their sum is

$$\begin{aligned} \mathrm{mgf}_{Y}(t) &= \mathrm{mgf}_{X_1}(t)\mathrm{mgf}_{X_2}(t) = \mathrm{e}^{\lambda_1 + \lambda_1 t + \lambda_2 + \lambda_2 t} \\ &= \mathrm{e}^{(\lambda_1 + \lambda_2) + (\lambda_1 + \lambda_2)t} \end{aligned}$$

which is the moment generating function of the Poisson distribution with mean $\lambda_1 + \lambda_2$ or $Y \sim \text{Poisson}(\lambda_1 + \lambda_2)$. In other words, this says that the sum of two independent Poisson random variables is again Poisson.

SUMMARY

In this unit, we defined the expectation of a random variable, extended this definition to expectations of powers of a random variable (called the moments), and formalized a measure of dispersion called the variance. We discussed key formulas and properties of expectations and variances and saw several practical examples of how these formulas and properties are used. We defined the covariance of two random variables as a quantification of their relationship or dependence. In the examples, we computed the covariances of several pairs of random variables both in the continuous and discrete type.

Although the expected value and variance of random variables are very important characteristics, they are not enough to describe the complete distribution (behavior) of a random variable. The central moments and, more importantly, their standard versions are important expectations that quantify other characteristics of a random variable. Two key quantities, the skewness and the kurtosis of random variables, were discussed and we saw the important roles they play in quantifying the departure from symmetry and the likelihood of observing extreme values respectively.

One of the most powerful concepts in the theory of random variables was also introduced in this section: the moment generating function. We defined this function, explained its role in finding the moments of a random variable, and outlined how it can completely describe a distribution. Therefore, becoming familiar with the specific forms of moment generating functions for the important distributions we have discussed in this course can be quite useful in practice.

UNIT 5

INEQUALITIES AND LIMIT THEOREMS

STUDY GOALS

On completion of this unit, you will be able to ...

- understand Markov's inequality and apply it to find upper bounds for tail probabilities of non-negative random variables.
- use Chebyshev's inequality and apply it in order to find the upper bounds of the probability.
- understand Hoeffding's inequality and be able to use it to find upper bounds of certain probabilities relating to the sample mean.
- understand the Cauchy-Schwartz inequality and apply it to find the upper bounds for the absolute expectation of a product of random variables.
- identify concave and convex functions and differentiate between them based on their graphs.
- apply Jensen's inequality and the central limit theorem as well as the weak law of large numbers as a convergence of probability.

5. INEQUALITIES AND LIMIT THEOREMS

Introduction

Much of statistics and, in particular, data science involves estimating distributions or the parameters of the "true" population from which the data derives. In general, it can be said that the "true" distribution of an observation (which is described by a given data set) can only be computed in rare situations. Therefore, it is very useful to be able to estimate the unknown distribution using a well-studied distribution, which has better known mathematical properties. When dealing with an unknown distribution, we cannot be sure of the margin of error associated with our estimates. In this unit we will introduce some tools that can be used to compute the probability that the estimates we obtain deviate from the true population parameters.

We will also deal with the weak law of large numbers. In basic terms this law means that if the same experiment is repeated independently a large number of times, the average of the results of the trials must be close to the expected value. In addition, we will prove this law for special cases using the inequalities introduced above.

Finally, we will discuss the central limit theorem, which says that, for a given sufficiently large sample size, the sampling distribution of the mean for a variable will approximate a normal distribution regardless of that variable's distribution in the population.

5.1 Probability Inequalities

Markov's Inequality

Probabilities of the form $P(X > t)$ or $P(|X| > t)$ are called tail probabilities for $t > 0$. They measure the likelihood of something extreme happening. To this end, Markov's inequality is a way of bounding that probability based on the mean.

Markov's Inequality

If X is a non-negative random variable with $\mathrm{E}[X] < \infty$ and $t > 0$, then

$$P(X > t) \leq \frac{\mathrm{E}[X]}{t}$$

In other words, the probability that X exceeds t is bounded by the mean X of divided by t.

All of the discrete distributions that we will discuss are non-negative: binomial, Poisson, and geometric, etc. Some of the continuous distributions—exponential, gamma, and beta —are also non-negative.

Let us now focus on the exponential distribution and verify Markov's upper bound. Suppose that the time to failure of a certain product follows an exponential distribution with a rate of 1/2 per year. Suppose $X \sim \text{Exponential}\left(\frac{1}{2}\right)$ so that its mean is $\mathrm{E}[X] = 2$. Recall that the CDF is given by

$$F(x) = 1 - \mathrm{e}^{-\frac{x}{2}}$$

for $x \geq 0$.

Let $t > 0$, as required by Markov's inequality, then

$$P(X > t) \leq \frac{2}{t}.$$

For $t = 6$, this gives

$$P(X > t) \leq \frac{2}{6} = \frac{1}{3}.$$

Direct computation of the probability gives

$$P(X > t) = 1 - F(t) = \mathrm{e}^{-\frac{t}{2}}.$$

so, for $t = 6$ we have

$$P(X > 6) = \frac{1}{\mathrm{e}^3} \approx 0.05 < \frac{1}{3}.$$

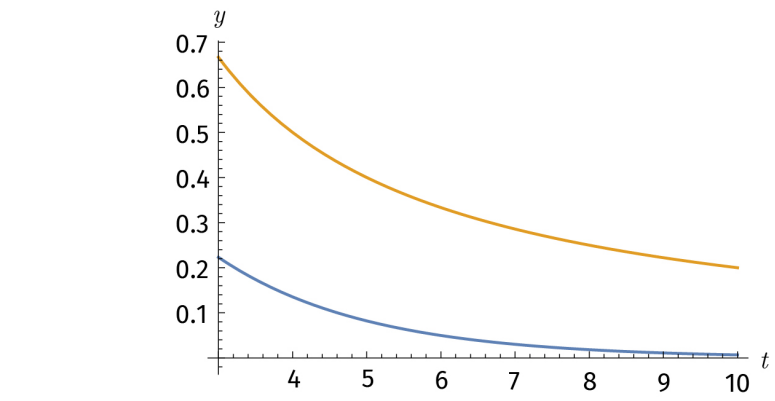
Markov's inequality is verified for this value. Here is a table showing Markov's bounds as well as the direct computation of the relevant probability.

Table 45: Verifying Markov's Bound with X ⊠ Exponential (1/2)

T	Markov's upper bound	Direct computation
3	0.67	0.22
4	0.5	0.14
5	0.4	0.08
6	0.33	0.05
7	0.29	0.03

Source: George Dekermenjian (2019).

Figure 68: Comparing Actual Probability with Markov's Bound



Here $X \sim$ Exponential (1/2)

— $P(X>t)=y$
— $Markov's\ Bound$

Source: George Dekermenjian (2019).

One of the strengths of Markov's inequality is that we do not need to know anything about the distribution of X except its mean (and that the X is non-negative). On the other hand, as we can see in the table above, Markov's inequality in its current form does not provide a very good upper bound. Let us adapt it to suit all random variables, not only non-negative ones. To that end, we will replace the random variable by its absolute value. Hence, if X is any random variable with $\mathrm{E}[|X|] < \infty$, we have for any $t > 0$ the following

$$P(|X| > t) \leq \frac{\mathrm{E}[|X|]}{t}.$$

As an example, suppose that $X \sim \mathcal{N}(1,1)$. Then $\mathrm{E}[|X|] \approx 1.17$. The new version of Markov's inequality reads:

$$P(|X| > t) \leq \frac{1.17}{t}.$$

Suppose $t = 3$, then numerical integration gives $P(|X| > 3) \approx 0.02$ and the upper bound from the inequality above gives $\frac{1.17}{3} \approx 0.39$. In other words, the probability that the random variable takes values more than three units from zero is less than about 39%.

Continuing our discussion of how to use Markov's inequality with any random variable (not necessarily non-negative), we recall that $x \longrightarrow e^x$ produces a non-negative value for any $x \in \mathbb{R}$. Furthermore, since this function is strictly increasing (one-to-one, meaning bijective or invertible), the event $[X > t]$ is equivalent to the event $[e^x > e^t]$. To this end, we have

$$P[X > t] = P[\mathrm{e}^x > \mathrm{e}^t] \leq \frac{\mathrm{E}[\mathrm{e}^X]}{\mathrm{e}^t} \quad \text{for any} \quad t \in \mathbb{R},$$

where we used Markov's inequality in the last inequality.

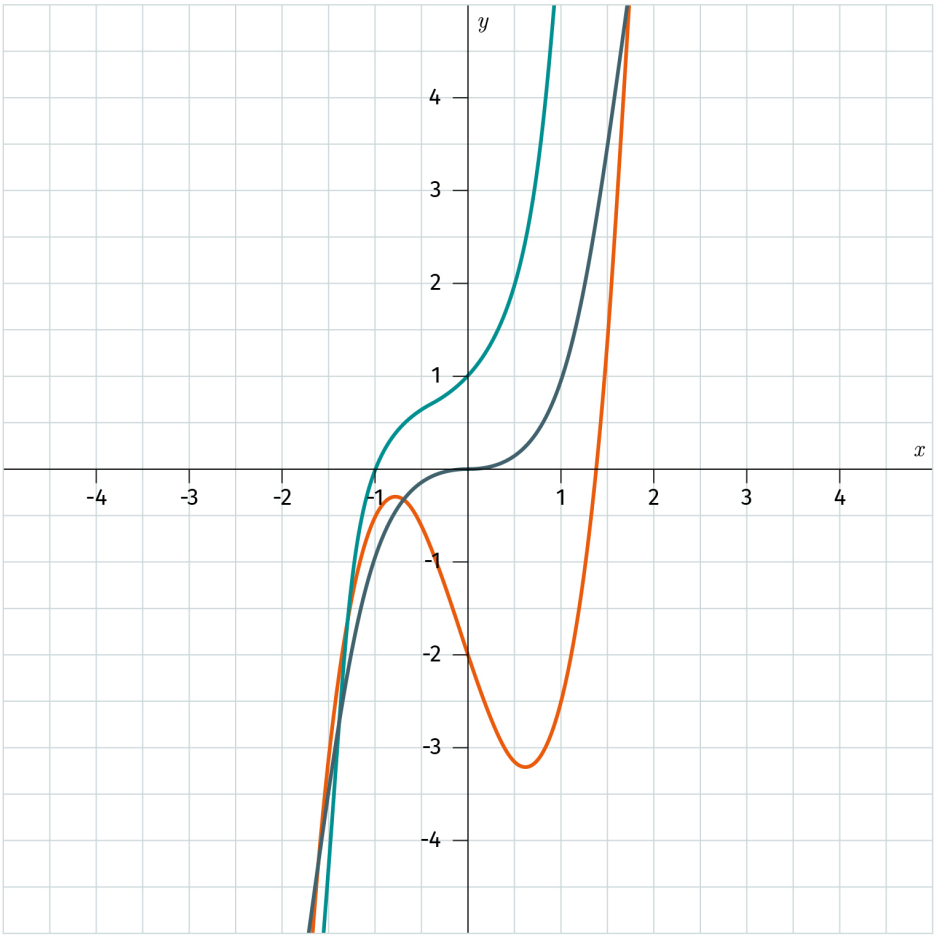
A short introduction to increasing functions

For a function $f\colon\ A \subseteq \mathbb{R} \to B \subseteq \mathbb{R}$, where A is called domain and B range, we define the following properties of f:

- f is strictly increasing if for any $x_1 < x_2$ follows $f(x_1) < f(x_2)$ for any $\mathrm{x}_1, \mathrm{x}_2 \in \mathrm{A}$,
- f is strictly decreasing if for any $x_1 < x_2$ follows $f(x_1) > f(x_2)$ for any $x_1, x_2 \in A$.

If we consider the following example of polynomials, we see that the blue ones are strictly increasing, since the function increases from left to the right. The orange one is neither increasing nor decreasing since it goes up and down.

Figure 69: Examples of Polynomial Functions



Source: George Dekermenjian (2019).

A short introduction to invertible (one-to-one) functions

For a function $f\colon\ A \subseteq \mathbb{R} \to B \subseteq \mathbb{R}$, where A is called domain and B, range we say that f is invertible, also called one-to-one or bijective, if there is a function $g\colon\ B \to A$ such that

$$f \circ g(y) = y \quad \text{for all} \quad y \in B \text{ and } \quad g \circ f(x) = x \quad \text{for all} \quad x \in A\,.$$

In that case is the inverse function of and we write

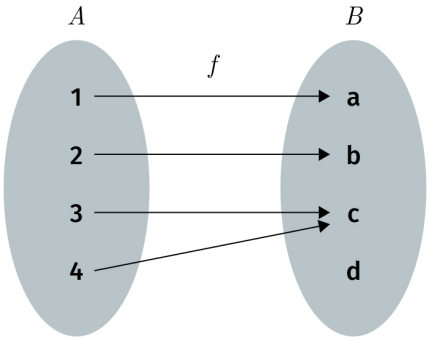
$$g(y) = f^{-1}(y) \text{ for all } \quad y \in B$$

Let us consider the following example, with domain $A = \{1,2,3,4\}$, the range $\mathrm{B} = \{\mathrm{a},\mathrm{b},\mathrm{c},\mathrm{d}\}$ and

$$f(1) = a,\ f(2) = b,\ f(3) = c,\ f(4) = c\,.$$

In the following graphic we can see that, starting from $4 \in A$ we have two arrows pointing to $c \in B$. Therefore, f is not one-to-one and cannot be invertible.

Figure 70: The Domain and Range of a Function



Source: George Dekermenjian (2019).

Recall that each strictly increasing or decreasing function is invertible.

Now consider a random variable with mean zero $\mathrm{E}[X] = 0$ and $\mathrm{Var}[X] = \mathrm{E}\left[X^2\right] = \sigma^2 < \infty$. Note that the mapping $z \mapsto z^2$ is one-to-one (strictly increasing) for $z > 0$. So, the event $[|X| > t]$ is equivalent to $\left[|X|^2 > t^2\right]$; therefore, for any $t > 0$ we have

$$P(|X| \geq t) = P\left(|X|^2 > t^2\right) < \frac{\mathrm{E}\left[X^2\right]}{t^2} = \frac{\sigma^2}{t^2}.$$

This extension of Markov's inequality gives bounds for random variables , which are not necessarily positive as well. It relates the probability to the variation of the random variable, which coincides with the second moment.

As an example, consider $X \sim \mathcal{N}(0,1)$. This version of Markov's inequality gives the following bounds:

$$P(|X| \geq 2) < \frac{1}{4} \text{ and } P(|X| \geq 3) < \frac{1}{9}.$$

Chebyshev's Inequality

Markov's inequality provides a way for us to discuss probabilities of extreme values. Sometimes it is useful to state bounds for the distance of probabilities from the mean. For random variables whose mean is zero, we saw how to find a bound using the **second moment.** Now we will deal with a more general version of that inequality.

Second moment
Recall that the second moment of a random variable X is $\mathrm{E}\left(X^2\right)$.

Recall that the mean is the long-term average of repeated experiments. We are often interested in the probability that any of those experiments produces a value that is far from the mean. As you will see, this will depend on the variation of the random variable and thus the inequality will use the variance of the random variable.

Chebyshev's inequality

Let X be a random variable with $\mu = \mathrm{E}[X] < \infty$ and $0 < \sigma^2 = \mathrm{Var}[X]$. Then for any $t > 0$ it holds that.

$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}$$

What this inequality says is that the probability that the distance of the random variable from the mean exceeding a given value t is inversely proportional to $t > 0$ and directly proportional to the variance of the random variable.

When we replace t by $\sigma^2 t$ for any $t \ > 0$ it holds that

$$P(|X - \mu| \geq \sigma^2 t) < \frac{1}{t^2}$$

Therefore, we know that for any random variable (with finite mean and variance), $Z = \frac{|X-\mu|}{\sigma}$ has a zero mean and unit variance. This quantity plays a role in standardizing random variables when working with real data. It centers the data to zero and scales the variance. When we are working with multiple variables, it is often useful to level the playing field for all variables before using the data in a data science setting.

Consider $X \sim \mathrm{Beta}(\alpha, \beta)$ for any $\alpha, \beta > 0$: we know that

$$\mathrm{E}[X] = \frac{\alpha}{\alpha + \beta} \text{and } \sigma^2 = \mathrm{Var}[X] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

Thus for $X \sim \mathrm{Beta}(2,3)$, we have $\mathrm{E}[X] = \frac{2}{5}$ and $\sigma^2 = \frac{1}{25}$. Therefore, Chebyshev's inequality gives

$$P\Big(\Big|x - \frac{2}{5}\Big| \geq t\Big) \leq \frac{1}{25t^2}.$$

We will now compare the upper bound from Chebyshev with the direct computation value for the probability.

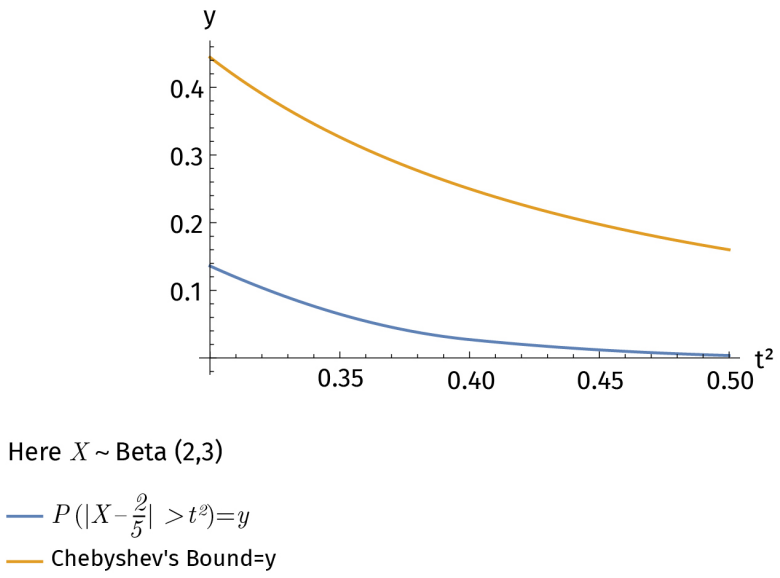
Table 46: Verifying Chebyshev's Bound with X ~ "Beta"(2,3)

t	Chebyshev's upper bound	Direct computation
0.3	0.4444	0.1360
0.4	0.2500	0.0272

t	Chebyshev's upper bound	Direct computation
0.5	0.1600	0.0037

Source: George Dekermenjian (2019).

Figure 71: Comparing Actual Probability with Chebyshev's Bound



Source: George Dekermenjian (2019)

Suppose that we want to estimate the true but unknown probability of success of Binomial(5,p). To this end, for $n \in \mathbb{N}$ we observe the realizations of $X_1, X_2, \ldots, X_n$ and estimate the mean via the sample mean

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

How likely is it that our estimate falls within 0.2 of the true mean? We know that

$$\mu = 5p \text{ and } \sigma^2 = \mathrm{Var}[\overline{X}] = \frac{Var[X]}{n} = \frac{5p(1-p)}{n}.$$

Note that $p(1-p) \leq \frac{1}{4}$ and, therefore we get

$$P(|\overline{X} - \mu| < 0.2) \geq 1 - \frac{125}{4n}.$$

If our sample has 2500 data points, the probability that our sample mean is within 0.2 of the true (unknown) mean is at least 98.75%.

Hoeffding's Inequality

The above example of estimating the parameter involved finding a probability bound of the mean of independent random variables. Hoeffding's inequality deals with such bounds. In much of data science, the data we work with comes from independent realizations of the same random process. Furthermore, we are often concerned with estimating an unknown parameter of the random process using these independent realizations. To this end, we frequently come across sums of independent and identically distributed random variables. Although Chebyshev's inequality (via Markov's inequality) gives us upper bounds for the relevant probabilities, Hoeffding's inequality gives us tighter (and thus more accurate) bounds, which is a result of the additional assumption that the range of the random variables are bounded.

Hoeffding's Inequality

Suppose that $X_1, X_2, \ldots, X_n$ for $n \in \mathbb{N}$ are independent observations from the same process. In other words, they are independent and identically distributed (often abbreviated to i.i.e). Suppose further that with $\mu = \mathrm{E}[X_i] < \infty$ and $a \leq X_i \leq b$ for all $i = 1,2, \ ..., n$ and $a, b \in \mathbb{R}$. In other words, they have a finite mean and are bounded with values between a and b. Then, for any $t > 0$, we have

$$P(\overline{X} - \mu \geq t) \leq \ \mathrm{e}^{-\frac{2nt^2}{(b-a)^2}}.$$

We can use Hoeffding's inequality twice to get a result for the absolute value as well, which is

$$P(|\overline{X} - \mu| \geq t) \leq 2\ \mathrm{e}^{-\frac{2nt^2}{(b-a)^2}},$$

where we recall the definition of the sample mean as follows

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

There is a lot going on here. Let us see this in the context of the previous problem where we were trying to estimate the probability of success of Binomial(5,p) for $0 < p < 1$.

Suppose, as we did before, we have a random sample of size n. We know that $\mu = 5p$. What is the probability that the sample mean is within 0.2 of the true mean? We will apply Hoeffding's inequality with $t = 0.2$, $a = 0$, and $b = 5$.

$$P(|\overline{X} - \mu| < 0.2) = 1 - P(|\overline{X} - \mu| \geq 0\ .2\ ) > 1 - 2\ \mathrm{e}^{-\frac{2n \cdot 0.2^2}{5^2}} = 1 - 2\mathrm{e}^{-0.0032n}$$

So, if our sample has 2500 data points, the probability that our sample mean is within 0.2 of the true (unknown) mean is at least $1 - 2e^{-0.0032 \cdot 2500} \approx 0.9993 = 99.93\%$.

Let us pause for a moment and interpret what we are saying here. If a large number of people each collect samples of 2500 from the sample population and compute the sample mean, then 99.93% of these people will produce answers within 0.2 of the true mean.

A special case of Hoeffding's inequality for Bernoulli random variables is worth pointing out. Machine learning algorithms that predict the class of a binary target variable such as "disease" or "no disease" are evaluated based on how accurately they can predict the class of a case. This accuracy is measured as the average of the correctly predicted cases. To this end, we introduce the following inequality.

Hoeffding's Inequality of Bernoulli random variables

Suppose that $X_1, X_2, \ldots, X_n \sim \text{Bernoulli}(p)$ for $0 < p < 1$ and $n \in \mathbb{N}$. Because $a = 0$ and $b = 1$ Heoffding's inequality yields for any $t > 0$,

$$P(|\overline{X} - p| > t) \leq 2\mathrm{e}^{-2nt^2},$$

where

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is the sample mean.

Suppose that a coin, possibly biased, is tossed many times and $X = 1$ is assigned to heads. In this setting, the sample mean $\overline{X}$ estimates the probability of getting heads. Our goal is to make sure the probability that our estimate differs from the true probability of heads, p, more than t, is less than some small number 2α, where $\alpha > 0$. According to Hoeffding's inequality, we are able to establish estimates for the event of getting heads. Our goal is to make sure the probability that our estimate differs from the true probability of heads, p, more than t, is less than some small number 2α. According to Hoeffding's inequality, we define

$$\alpha = \mathrm{e}^{-2nt^2}$$

as an upper bound, and solve for $n \in \mathbb{N}$ to get

$$n = -\frac{\ln\alpha}{2t^2}$$

For instance, our estimate will differ from the true proportion of heads by more than $t = 0.01$, with a probability of $2\alpha = 0.02$, as long as we sample at least

$$n = -\frac{\ln 0.01}{2 \cdot 0.01^2} \approx 23{,}026$$

data points.

Example 5.1

A machine learning model is trained on a sample of 1000 examples to predict whether a given transaction is fraudulent or not. Use Hoeffding's inequality to find an upper bound of the probability that the observed accuracy differs from the actual accuracy by more than 1%. Assume that the data the model is based on comes from the same distribution as the training data.

Solution 5.1

Let $0 < p < 1$ be the actual accuracy of the model (of the complete data set). The outputs of the training examples are $X_1, X_2, \ ..., X_{1000}$~$\mathrm{Bernoullli}(p)$ where $X_i = 1$ is correctly classified and $X_i = 0$ is incorrectly classified. For $t = 0.01$ we use Hoeffding's inequality to obtain an upper bound for the difference between the actual accuracy and the observed accuracy

$$P(|\overline{X} - p| > 0.01) \leq 2\ \mathrm{e}^{-2 \cdot 1000 \cdot (0.01)^2} = 2\mathrm{e}^{-0.2}.$$

5.2 Inequalities and Expectations

In this section, we will introduce some important inequalities related to expectations. In this unit we will present the Cauchy-Schwartz inequality and Jensen's inequality. These inequalities are very useful if quantities such as the excepted value and variance are well known but the data is not. They can also be used to estimate the parameters of an assumed distribution, which should describe a random variable.

Cauchy-Schwartz Inequality

The motivation for the Cauchy-Schwartz inequality comes from the inequality from linear algebra with the same name. We will discuss it for two-dimensional vectors

$$v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \quad \text{and} \quad v = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \quad \text{where} \quad v_1, v_2, u_1, u_2 \in \mathbb{R}\,.$$

The scalar product in two dimensions is then defined by

$$|< \mathrm{u,v} >| = u_1 \cdot v_1 + u_2 \cdot v_2$$

and the associated norm, which measures the length of the vectors is

$$|u| = \sqrt{u_1^2 + u_2^2}\,.$$

Recall that given two vectors u and v, we have that

$$|< u, v >| \leq |u||v|,$$

which is called the Cauchy-Schwartz inequality in the two-dimensional space. It turns out that we can define an inner product $< \cdot, \cdot >$, which is the equivalent to the scalar product for vectors, of two random variables X and Y as $< X, Y > = \mathrm{E}[XY]$.

Cauchy-Schwartz inequality

Let X and Y be two random variables such that $\mathrm{Var}[X] < \infty$ and $Var\ [Y] < \infty$, then it holds that

$$|\mathrm{E}[XY]| \leq \sqrt{\mathrm{E}\left[X^2\right]\mathrm{E}\left[Y^2\right]}.$$

This inequality gives an upper bound on the absolute value of the expected value of the product of two random variables.

Example 5.2

Suppose that for the random variables X and Y, $\mathrm{E}[X] = 1$, $\mathrm{E}[Y] = 2$ and $\mathrm{Var}[X] = \mathrm{Var}[Y] = 1$. Give an estimate for $|\mathrm{E}[XY]$ using the Cauchy-Schwartz inequality.

Solution 5.2

We are working on the basis that $\mathrm{E}\left[X^2\right] = \mathrm{Var}[X] + \mathrm{E}[X]^2 = 1 + 1 = 2$ and $\mathrm{E}\left[Y^2\right] = \mathrm{Var}[Y] + \mathrm{E}[Y]^2 = 1 + 4 = 5$. Then, according to Cauchy-Schwartz, we have

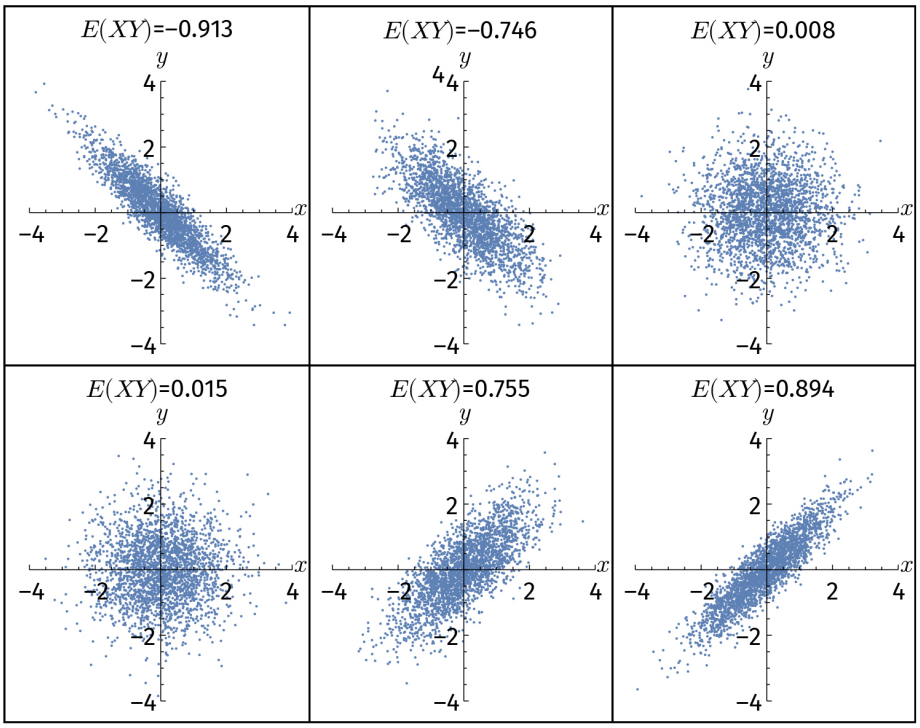
$$|\mathrm{E}[XY]| \leq \sqrt{2 \cdot 5} = \sqrt{10}.$$

Let us consider the case where $\mathrm{E}[X] = \mathrm{E}[Y] = 0 \text{ and } \mathrm{E}\left[X^2\right] = \mathrm{E}\left[Y^2\right] = 1$. Cauchy-Schwartz tells us that $|\mathrm{E}[XY]| \leq 1$. The covariance of X and Y is

$$\mathrm{Cov}(X, Y) = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y] = \mathrm{E}[XY] - 0 = \mathrm{E}[XY]$$

Therefore, $|\mathrm{Cov}(X, Y)| \leq 1$. On the one hand, if X and Y are strongly correlated, we would have $|\mathrm{Cov}(X, Y)|$, which is close to one. On the other hand, if X and Y are almost independent (weak or no correlation), then $|\mathrm{Cov}(X, Y)|$ is close to zero. This kind of interpretation of the Cauchy-Schwartz inequality vis-à-vis the interpretation of the inner products is analogous to the interpretations with vectors, which we will also deal with in this section. The following scatter plots show random points of observations of (X, Y) with $\mathrm{E}[X] = \mathrm{E}[Y] = 0$, $\mathrm{E}\left[X^2\right] = \mathrm{E}\left[Y^2\right] = 1$, $\mathrm{Var}[X] = \mathrm{Var}[Y] = 1$ and computed the correlation, which is in that case $\mathrm{Cov}(X, Y) = \mathrm{E}(XY)$.

Figure 72: Scatter Plots of Observations Drawn from Distributions with Various Values of Correlation

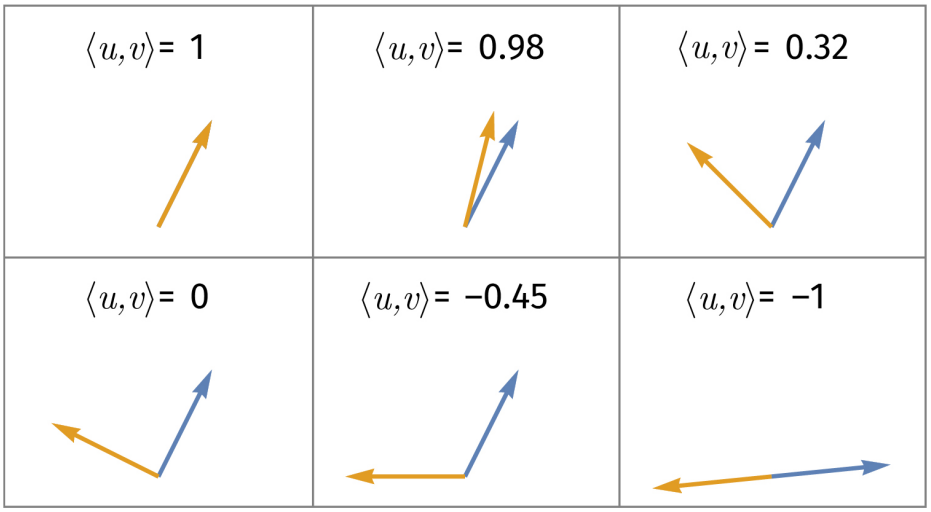


The correlation is computed for the data points in each heading of a picture and we see in which direction the data points direct with respect to the covariance.

Source: George Dekermenjian (2019).

For unit vectors $|u| = |v| = 1$, recall that the scalar product measures the similarity of their directions, as shown in the following illustrations.

Figure 73: Pairs of Unit Vectors and Their Inner Products



Source: George Dekermenjian (2019).

When two vectors are directed more or less in the same or opposite directions, their inner product is large (positive and negative, respectively). When two vectors are close to perpendicular to one another, their inner product is small. Similarly, when the two random variables (zero mean and unit variance) are dependent on one another (large covariance), the expected value of their product is relatively large, and when they are almost independent, the expected value is small.

Jensen's Inequality

For a given random variable X it is sometimes more convenient to consider the random variable $g(X)$ for suitable function $g: \mathbb{R} \to \mathbb{R}$. For instance, if we consider $ln(X)$ instead of X it is much easier to measure the percentage of variation in a stock chart or changes in climatic conditions. In some cases, we are also interested in the tails of our distribution or in the asymptotic behavior for high values of our random variable, which can be estimated using Jensen's inequality.

Perhaps you are wondering how we can derive a bound for $g(X)$ if we only know something about the expectation value of X. This is where the following inequality comes into play.

Jensen's inequality

Let X be a random variable with $\mathrm{E}[X] < \infty$ and let $g: \mathbb{R} \to \mathbb{R}$ be a function. If g is convex, then

$$\mathrm{E}[g(X)] \geq g(\mathrm{E}[X])$$

If g is concave, then

$$\mathrm{E}[g(X)] \leq g(\mathrm{E}[X])$$

A short introduction to convex and concave functions

Consider a function $g: \mathbb{R} \longrightarrow \mathbb{R}$. We say that g is a convex function if, for any two points $x, y \in \mathbb{R}$ and $0 \leq \alpha \leq 1$, we have

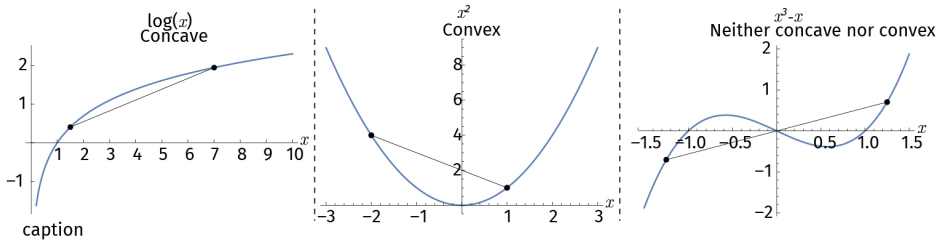
$$g(\alpha x + (1-\alpha)y) \leq \alpha g(x) + (1-\alpha)g(y)\,.$$

We say that g is concave if

$$g(\alpha x + (1-\alpha)y) \geq \alpha g(x) + (1-\alpha)g(y)\,.$$

In geometrical terms, a function is convex if, when you pick any two points on the graph of the function and draw a line segment between the two points, the entire segment lies above the graph. On the other hand, if the line segment always lies below the graph, the function is said to be concave.

Figure 74: Plots of Log(x), x^2, x^3-x



Source: George Dekermenjian (2019).

In the above example we see that $g(x) = \log x$ is a concave function. An application of that theorem yields $\mathrm{E}[\log(X)] \leq \log(\mathrm{E}[X])$. If we put $g(x) = X^2$, which is a convex function, we have that $\mathrm{E}\left[X^2\right] \geq \mathrm{E}[X]^2$. This last inequality proves the well-known inequality

$$\mathrm{Var}[X] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 \geq 0,$$

which is basically a result of the positivity of the variance.

For a smooth enough function g we can prove whether a function is convex or concave. If a function g is twice differentiable, then it is convex if and only if $g''(x) \geq 0$ holds for any $x \in \mathbb{R}$, and it is concave if and only if $g''(x) \leq 0$ for all $x \in \mathbb{R}$.

From the examples above, if $g(x) = \log x$, then $g'(x) = \frac{1}{x}$ and $g''(x) = -\frac{1}{x^2} < 0$, which confirms that $\log x$ is concave. If $g(x) = x^2$, then $g'(x) = 2x$ and $g''(x) = 2 > 0$, which confirms that x^2 is convex. You can use this fact to check that e^x, $-\log x$, and x^4 are all convex.

5.3 The Law of Large Numbers

Generally, the more data we have, the better we can estimate parameters of the theoretical distributions. This is the general idea behind the **law of large numbers** when the parameter of interest is the average. The law of large numbers is a limit theorem. It makes a statement about the limiting behavior of the sample mean as the sample size gets larger without bound with respect to probability. To demonstrate the idea, we will take observations of increasing sample sizes and look at the behavior of the sample mean.

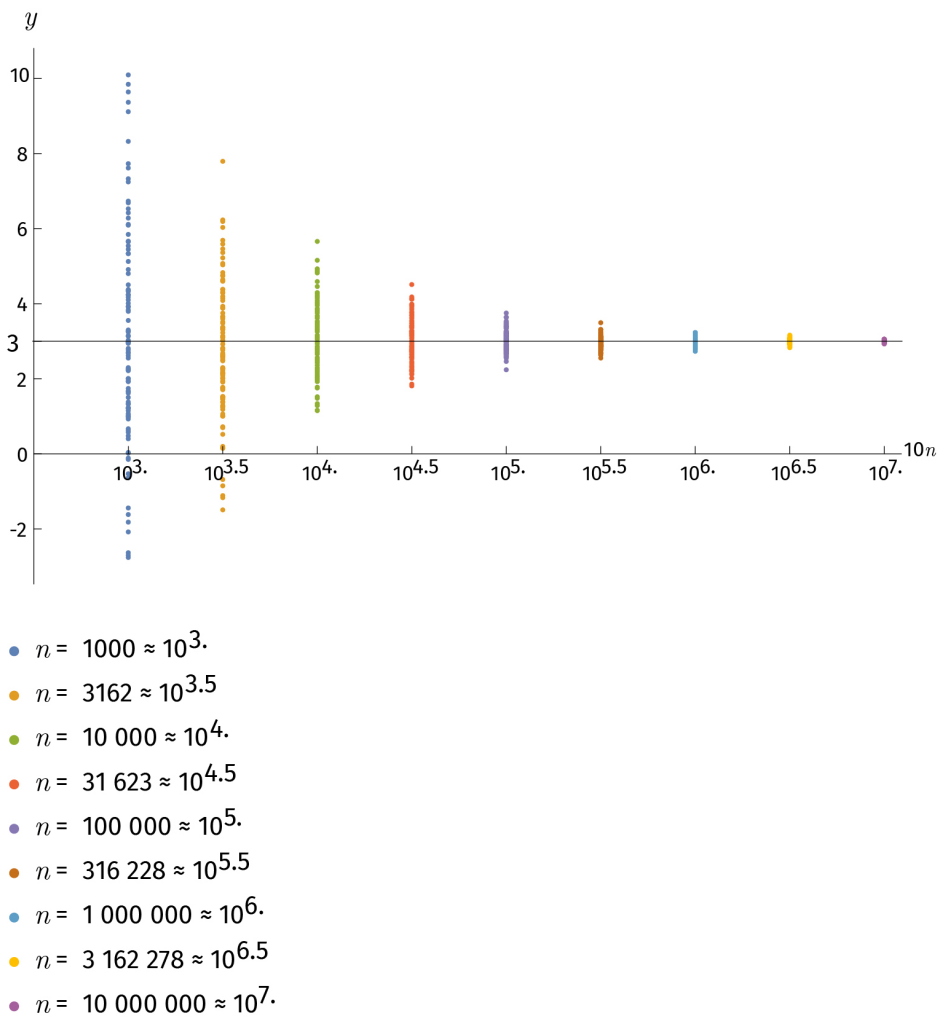
Weak law of large numbers
The weak law of large numbers states that the average of a sufficiently large number of observations will be close to the expected value.

Let us consider the following random variables $X_1, X_2, ..., X_n$, where $X_j \sim \mathcal{N}\left(3, 100^2\right)$ for all $1 \leq j \leq n$ for $n \in \mathbb{N}$. We create data points $X_1, X_2, ..., X_n$ by simulating the random variables $X_1, X_2, ..., X_n$, and consider for fixed $n \in \mathbb{N}$ the sample mean

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} x_i \,.$$

We will use sample sizes ranging from $n = 10^3$ to $n = 10^7$ and compute $\overline{X}$ 100 times. The scatter plot below shows the sample means (on the vertical axis) versus the sample sizes (on the horizontal axis). Each color corresponds to a different sample size and each point corresponds to a single sample mean.

Figure 75: Sample Means from 100 Samples of Various Sizes



Source: George Dekermenjian (2019).

Now, for each sample size, let us see how many of the 100 sample means are within 0.1 of the target mean $\mu = E[X] = 3$. In other words, we want to compute

$$P(|\overline{X} - \mu| > 0.1)\,.$$

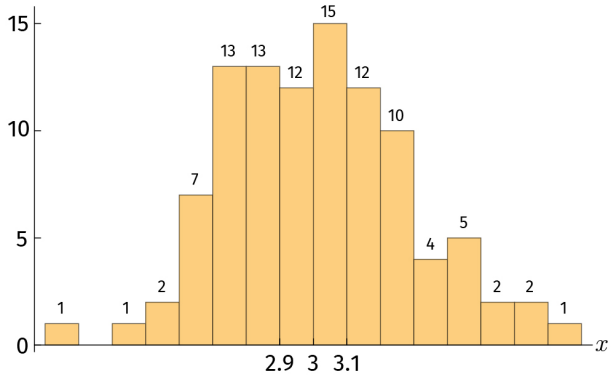
Histograms

A histogram is an approximate representation (similar to a piecewise continuous PDF) of the distribution of a given data. In our example we have 100 data points, which were constructed by the sample mean

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_I,$$

where x_I were realization of the random variables such that $X_j \sim \mathcal{N}(3,100^2)$ for fixed $n \in \mathbb{N}$ and $j = 1,2,3,...,n$. Thus we have 100 data points denoted by $\overline{X}_1, \overline{X}_2, \ldots \overline{X}_{100}$, We construct an interval, starting with the smallest value of $\overline{X}_1, \overline{X}_2, \ldots \overline{X}_{100}$ and ending with the largest value of $\overline{X}_1, \overline{X}_2, \ldots \overline{X}_{100}$. This interval is then divided into a series of intervals of the same width, called bins. For each bin, we define a rectangle with a height proportional to the frequency of the data points that lie within it. Here is a histogram for the sample means for $n = 10^5$.

Figure 76: Histogramm of 100 Sample Means for n=10^5



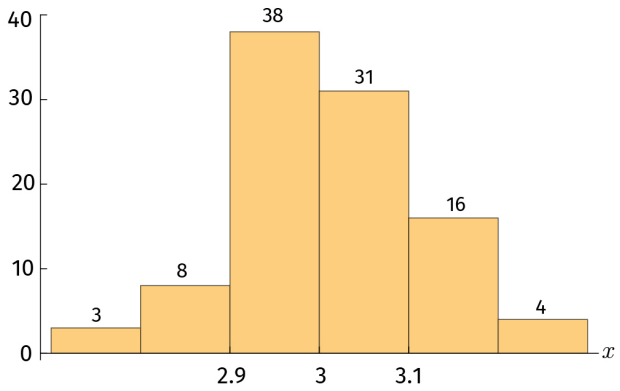
Note: Each sample has a size of 100,000.

Source: George Dekermenjian (2019).

We can see from the histogram that $(12 + 15) = 27$ of the 100 sample means are within 0.1 of the true mean and 73 are not. In other words, our estimate of $P(|\overline{X} - \mu| > 0.1)$ is 0.73.

Here is a histogram of the sample means for $n = 10^7$.

Figure 77: Histogramm of 100 Sample Means for n=10^7



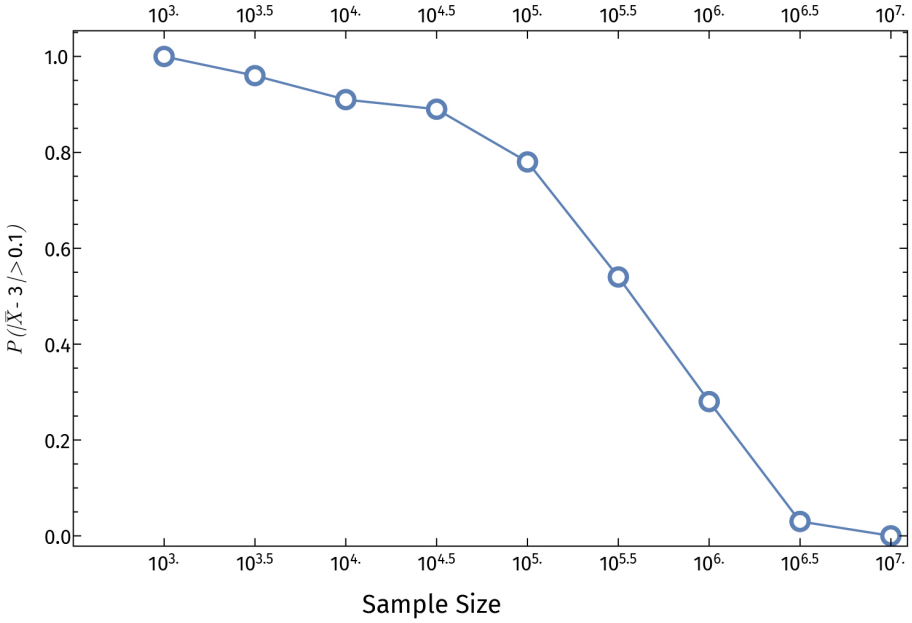
Note: Each sample has a size of 10,000,000.

Source: George Dekermenjian (2019).

From the histogram, we see that $(38 + 31) = 69$ of the 100 sample means are within 0.1 of the true mean and 31 are not. In other words, our estimate of $P(|\overline{X} - \mu| > 0.1)$ is 0.31.

We continue in this way and compute the probability for all the sample sizes. The following plot shows the probability versus the sample size.

Figure 78: Probability That the Sample Mean Differs from the True Mean by More Than 0.1



Source: George Dekermenjian (2019).

As you can see from the plot here, the probability $P(|\overline{X} - \mu| > 0.1)$ goes to zero as the sample size increases. We are now ready to write down the formal statement of the weak law of large numbers.

The Weak Law of Large Numbers

Let $X_1, X_2, \ldots, X_n$ be random variables such that $\mathrm{E}[X_i] = \mu < \infty$ for $I = 1,2, \ldots n \in \mathbb{N}$. Then, for any $t > 0$ (no matter how small)

$$\lim_{n \to \infty} P(|\overline{X} - \mu| > t) = 0,$$

where the sample mean is defined as

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

In other words, as the sample size increases without bound, the probability that the sample mean differs more than any fixed amount approaches zero. This is also called convergence in probability literature (see, for example, Klenke, 2014). When it comes to convergence types, this is the weakest. Hence, the name “weak” in the title of the theorem.

This also follows directly from Chebyshev’s inequality. Recall that

$$P(|X-\mu|>t)\leq\frac{\sigma^2}{t^2}$$

If we want to apply this with $\overline{X}$ in place of X, we need to first write down the variance, $\mathrm{Var}|\overline{\mathrm{X}}|=\frac{\sigma^2}{n}$. Recall that $\mathrm{E}[\overline{X}]=\mathrm{E}[X]=\mu$. Therefore, Chebyshev's inequality gives

$$P(|\overline{X}-\mu|>t)\leq\frac{\sigma^2}{nt}$$

For fixed t, if $n\to\infty$, the right-hand side goes to zero, as does the probability.

5.4 The Central Limit Theorem

Since statistical inference deals with the problem of analyzing properties of an underlying distribution, describing a random process, the central limit theorem gives in an insight about the behavior of that process in case it is repeated often enough. It is perhaps the most widely known theorem in all of statistics, and it is the main reason the standard normal distribution is called "normal". Before we state the central limit theorem (CLT), we will conduct some experiments and visualize the result of the CLT using histograms and QQ-plots.

We will draw 1000 samples each of size $n\in\mathbb{N}$ from the $X\sim\mathrm{Uniform}[0,1]$. The first sample is $X_{1,1},X_{1,2},\ldots,X_{1,n}$, the second sample is $X_{2,1},X_{2,2},\ldots,X_{2,n}$, and the 1000th sample is $X_{1000,1},X_{1000,2},\ldots,X_{1000,n}$. We calculate the sample mean of each sample as

$$\overline{X_i}=\frac{1}{n}\sum_{j=1}^{n}X_{i,j}$$

for $i=1,2,\ ...,1000$.

Recall that $\mu=\mathrm{E}[\overline{X}_i]=\mathrm{E}[X]=\frac{1}{2}$ for all $i=1,2,\ ...,1000$. Furthermore, the variance of each sample mean is

$$\sigma_n^2=\mathrm{Var}[\overline{X}_i]=\frac{\mathrm{Var}[X]}{n}=\frac{1}{12n}$$

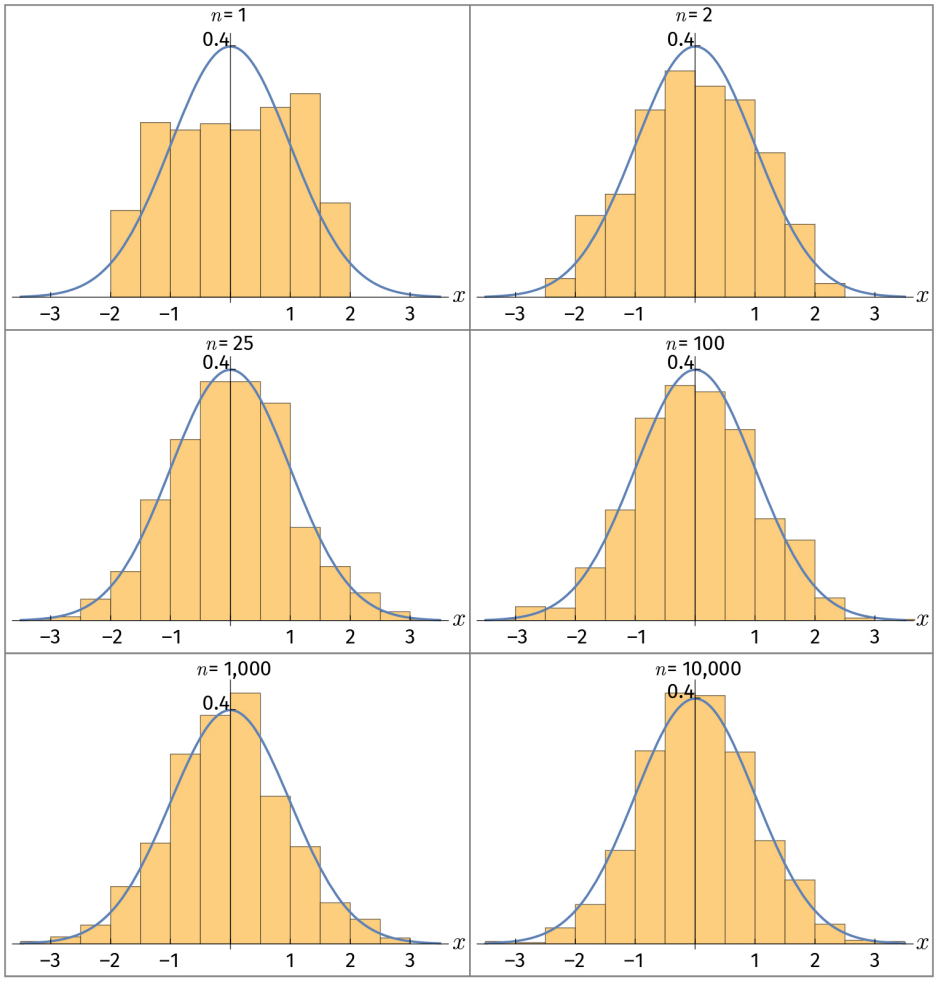
Finally, we will center and scale these sample means by

$$Z_{i,n}=\frac{\overline{X}_I-\mu}{\sigma_n}=\frac{\overline{X}_i-\frac{1}{2}}{\frac{1}{\sqrt{12n}}}$$

We know that $\mathrm{E}\big[Z_{i,n}\big]=0$ and $\mathrm{Var}\big[Z_{i,n}\big]=1$ for $i=1,2,\ ...,1000$.

We are interested in the distribution of the $Z_{i,n}$ for various values of $n\in\mathbb{N}$.

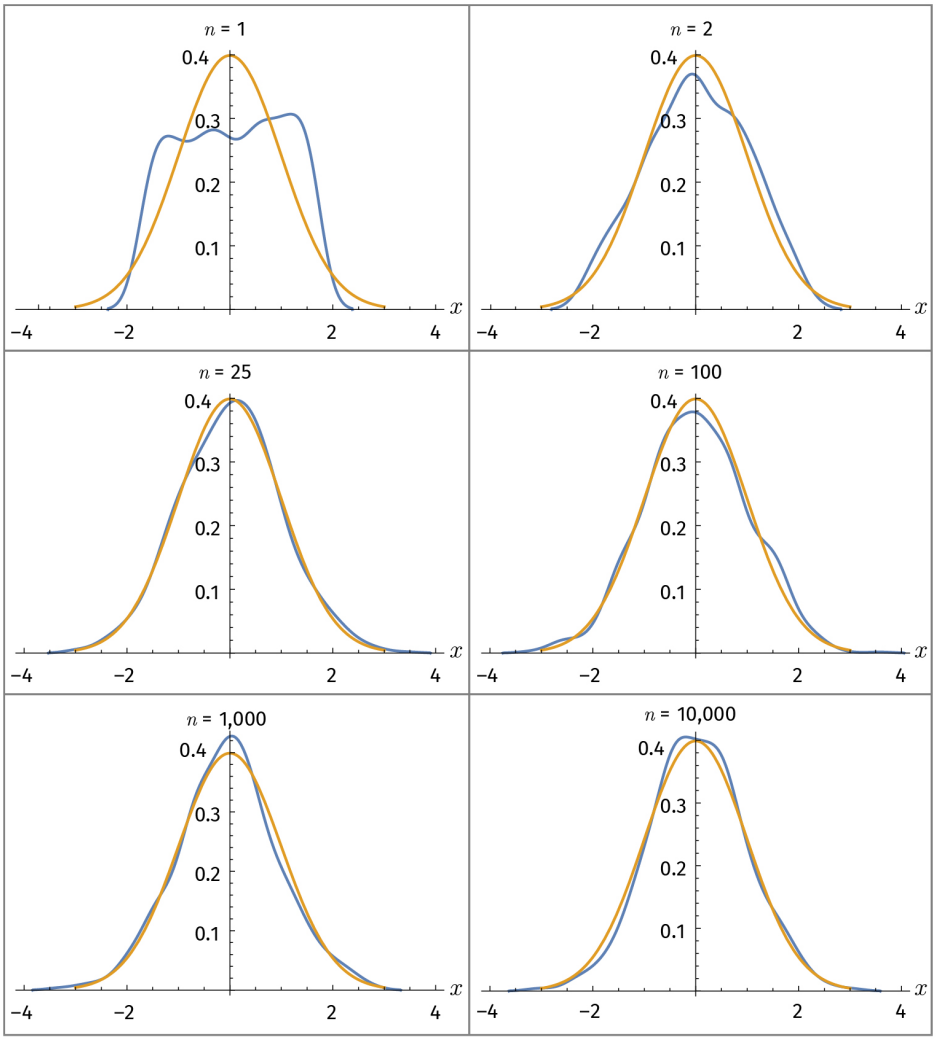
Figure 79: Demonstrating the Central Limit Theorem via Histograms: Samples Drawn from Uniform $[0, 1]$



The function depicted in blue is the PDF of the standard normal distribution.

Source: George Dekermenjian (2019).

Figure 80: Demonstrating the Central Limit Theorem via Smooth Histograms: Samples Drawn from Uniform $[0, 1]$



Orange: Empirical PDF of the sample data

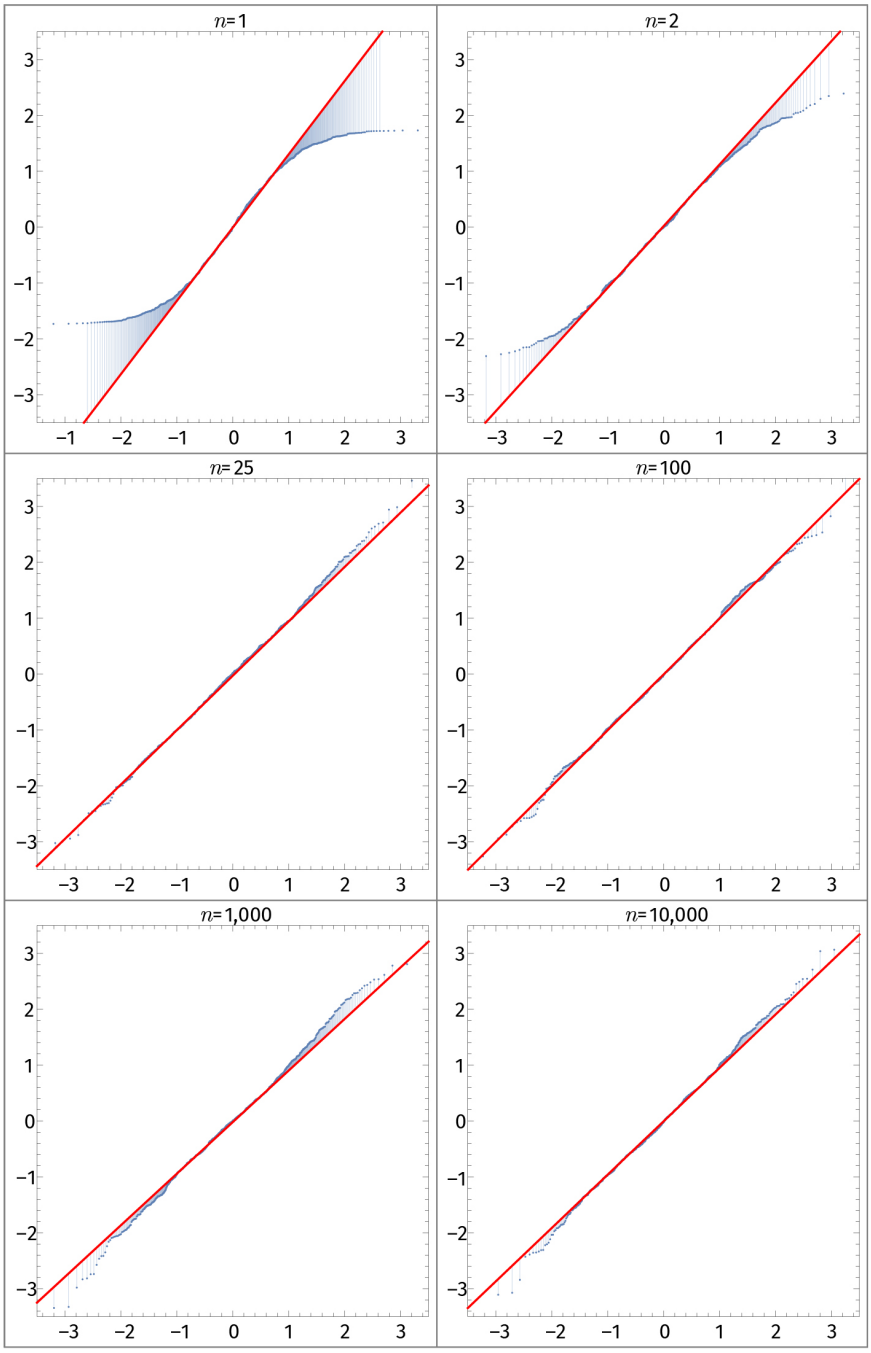
Blue: Standard Normal PDF

Source: George Dekermenjian (2019).

QQ-plot

A QQ scatter plot is a plot of points where the horizontal coordinate is a theoretical quantile (in this case of the normal distribution), and the vertical coordinate is the empirical quantile of the data. A quantile is also called a percentile. For instance, the tenth percentile is the point that separates the bottom 10% of the data from the rest. When the points follow a straight line, the distribution of the data is close to the theoretical distribution.

Figure 81: Demonstrating the Central Limit Theorem via QQ-Plots: Samples Drawn from Uniform (0,1)



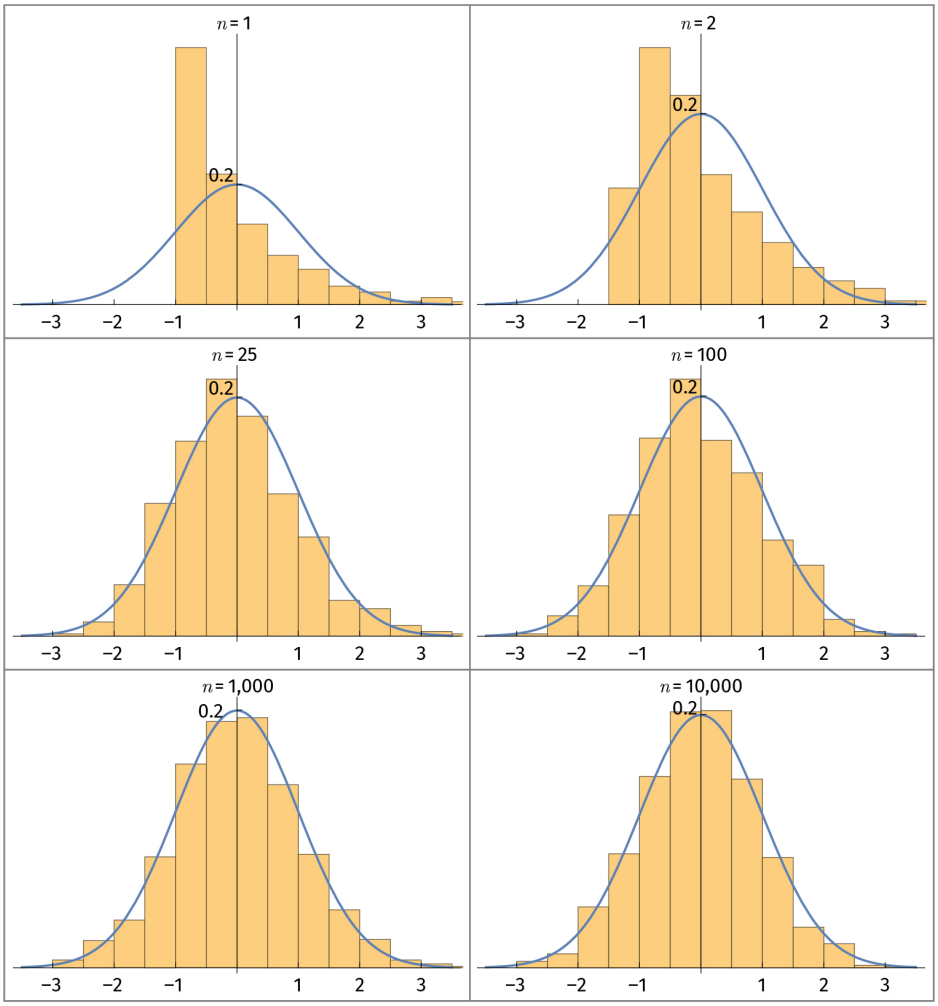
The horizontal coordinates describe the theoretical quantiles and the vertical ones the empirical quantile of the data.

Source: George Dekermenjian (2019).

The first and second figures show that, as we increase the sample size $n \in \mathbb{N}$, the distribution of the standardized sample mean becomes more and more symmetric and, in fact, matches up with the standard normal distribution quite closely. The QQ-plot, which plots the empirical quantiles against the theoretical quantiles, is used to see how well a data set follows the normal distribution. These plots also confirm that as $n \in \mathbb{N}$ increases, the distribution of the standardized sample means indeed become more and more like the standard normal distribution.

This result is not unique to the uniform distribution. As a matter of fact, we can perform the same experiment by using the exponential distribution instead. In the following figures, the parent distribution comes from $X \sim \mathrm{Exponential}(1)$. The rest of the computation is the same as before.

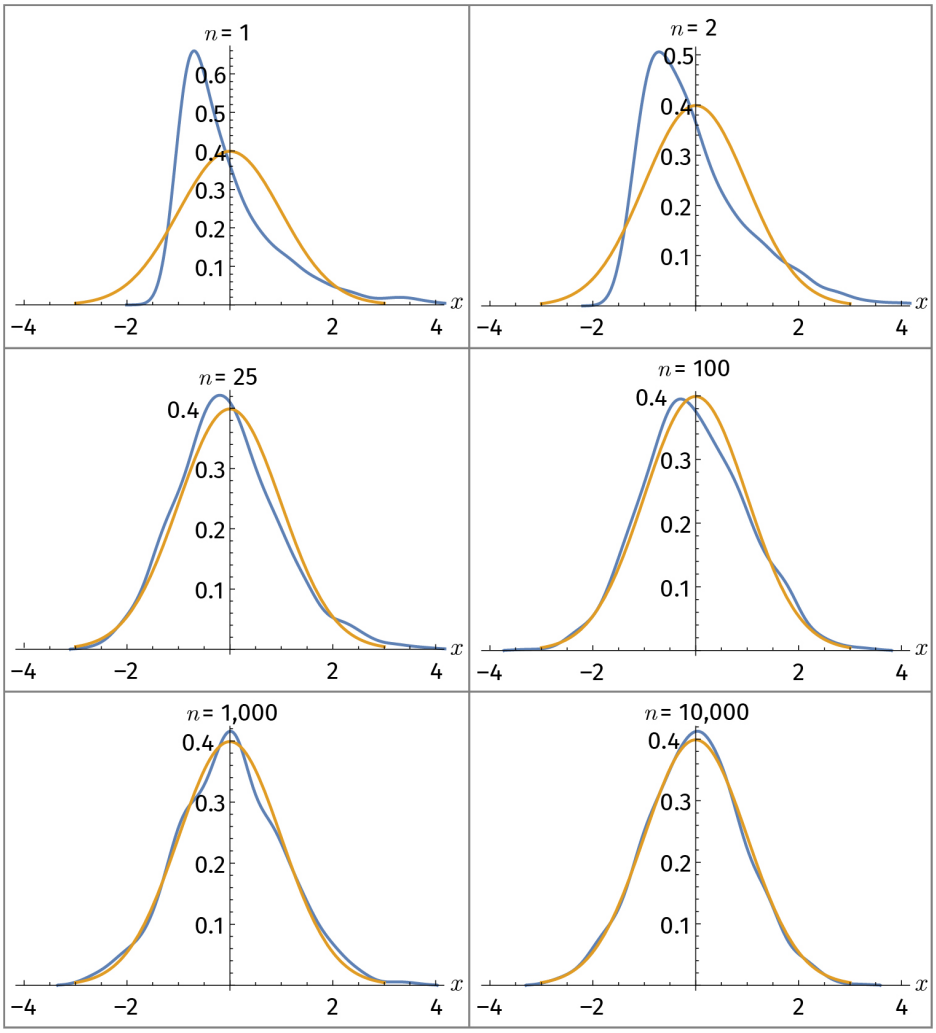
Figure 82: Demonstrating the Central Limit Theorem via Histograms: Samples Drawn from Exponential (1/2)



Please note that a random variable X can be standardized by subtracting its mean and then dividing it by its standard deviation.

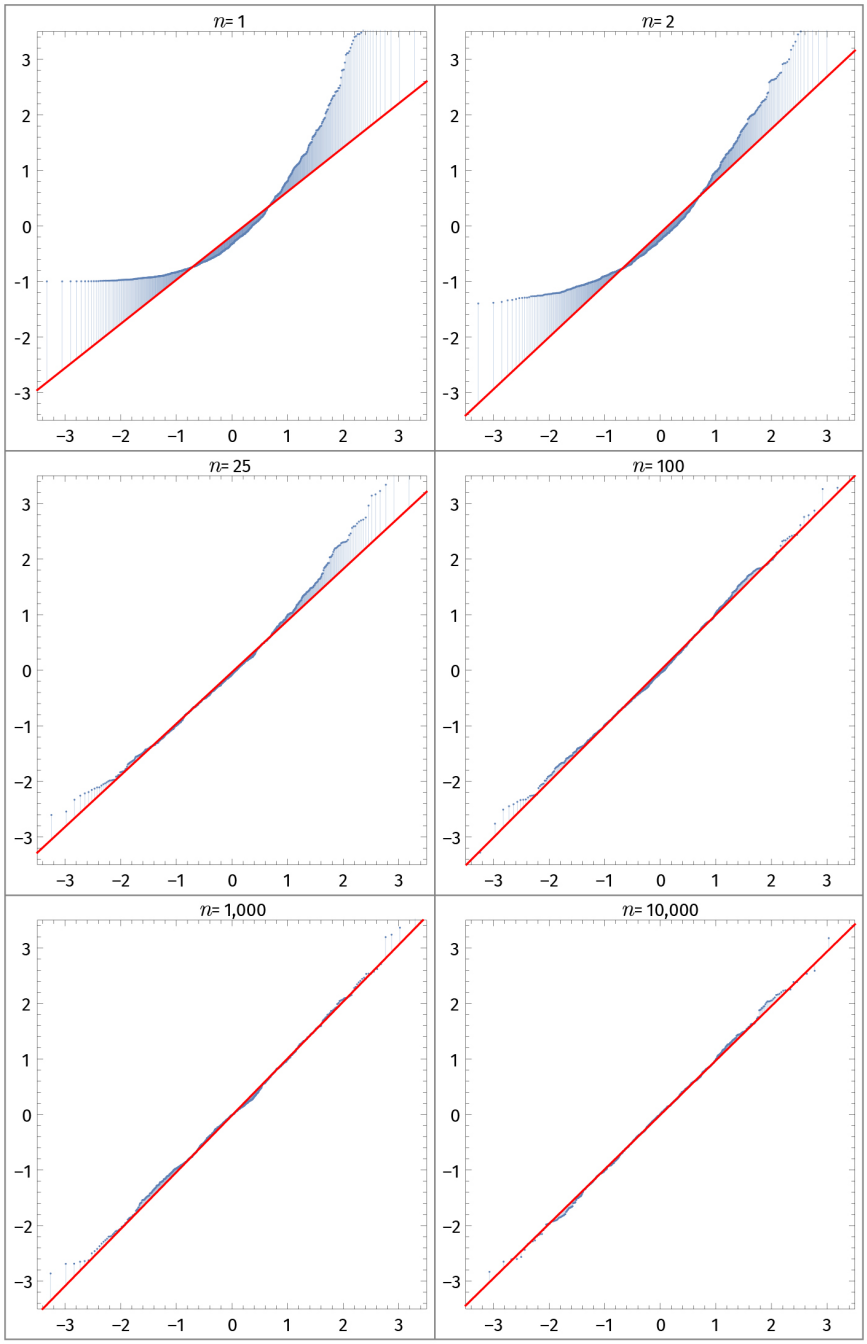
Source: George Dekermenjian (2019).

Figure 83: Demonstrating the Central Limit Theorem via Smooth Histograms: Samples Drawn from Exponential (1/2)



Source: George Dekermenjian (2019).

Figure 84: Demonstrating the Central Limit Theorem via QQ-Plots: Samples Drawn from Exponential (1/2)



The horizontal coordinates describe the theoretical quantiles and the vertical ones the empirical quantile of the data.

Source: George Dekermenjian (2019).

As you can see, although this parent distribution requires a relatively large $n \in \mathbb{N}$ before the fit with the standard normal occurs, it still, nevertheless, occurs. This is indeed fantastic news! It means that if we have a random sample with a large enough sample size, then even if we do not know the distribution of the parent population, we can still make inferences about the mean of the parent population via the standardized sample mean of the data we have. The fact that the distribution of the sample mean is always approximately normally distributed is one of the most celebrated results in statistical theory. This will be illustrated with examples later on in this section.

Now that you understand the importance of this key statistical theorem, you are ready to move on to the statement of the CLT.

The Central Limit Theorem

For $n \in \mathbb{N}$ let $X_1, X_2, \ldots X_n$ be independent and identically distributed (random sample) with mean $\mu = \mathrm{E}[X_i] < \infty$ and variance $0 < \mathrm{Var}[X_i] = \sigma^2 < \infty$ for $i = 1,2, \ ..., n$. Let $\overline{X}_n$ be the sample mean as defined by

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \, .$$

Then, the distribution of the standardized sample mean

$$\overline{Z}_n := \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}},$$

is close to the standard normal distribution, meaning that for all $z \in \mathbb{R}$ it holds that

$$\lim_{n \to \infty} P(\overline{Z}_n \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \underline{\mathrm{dt.}}$$

Equivalently, the sample mean is approximately normally distributed, meaning that for any $z \in \mathbb{R}$ it holds that

$$\lim_{n \to \infty} P(\overline{X}_n \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} \underline{\mathrm{d}t}$$

Example 5.3

Suppose a sample of 50 observations is drawn from Beta(2,3). What is the approximate distribution of the sample mean?

Solution 5.3

From the setting of the CLT, we have

$$X_1, X_2, X_{50} \overset{i.i.d}{\sim} \mathrm{Beta}(2,3)$$

so that

$$\mathrm{E}[X_i] = \frac{2}{5} \text{ and } \sigma^2 = \mathrm{Var}[X_i] = \frac{1}{25}$$

for $i = 1,2, \ ..., 50$. Therefore, for the sample mean

$$\overline{X} = \frac{1}{50}\sum_{i=1}^{50} X_i$$

holds approximately $\overline{X} \sim \mathcal{N}\left(\frac{2}{5}, \frac{1}{25 \cdot 50}\right)$, because of

$$\begin{aligned}\mathrm{E}[\overline{X}] &= \frac{2}{5}, \\ \mathrm{Var}[\overline{X}] &= \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{50} X_i\right) = \frac{1}{50^2}\sum_{i=1}^{50}\mathrm{Var}(X_i) = \frac{1}{50^2}\sum_{i=1}^{50}\frac{1}{25} = \frac{1}{25 \cdot 50}.\end{aligned}$$

It is important to note that because of the independence of X_i for all $i = 1,2, ..., 50$ we could change the summation with the computation of the variance.

Example 5.4

Suppose that the worldwide average IQ is 100 with a standard deviation of 15.

a) Assuming that IQ is normally distributed within the given parameters, find the probability that a randomly chosen person has an IQ of more than 103?
b) We draw 64 people at random. What is the probability that the average IQ of this sample is more than 103?

Solution 5.4

a) If X represents the IQ of a randomly selected person, we are told that $\mathrm{X} \sim \mathcal{N}\left(100, 15^2\right)$. Therefore,

$$\begin{aligned}P(X > 103) &= P\left(\frac{X - 100}{15} > \frac{103 - 100}{15}\right) \\ &= P(Z > 0.2) \approx 0.4207\end{aligned}$$

or approximately 42.07%.

b) From the CLT, we know that the distribution of the sample mean is approximately normal with mean 100 and

$$\begin{aligned}\mathrm{Var}[\overline{X}] &= \frac{1}{64^2}\mathrm{Var}\left(\sum_{i=1}^{64} X_i\right) = \frac{1}{64^2}\sum_{i=1}^{64}\mathrm{Var}(X_i) \\ &= \frac{1}{64^2}\sum_{i=1}^{64} 15^2 = \frac{15^2}{64}.\end{aligned}$$

Thus, the standard deviation is $\frac{15}{\sqrt{64}} = \frac{15}{8}$. Therefore, we have

$$\begin{aligned} P(\overline{X} > 103) &= P\left(\frac{\overline{X} - 100}{\frac{15}{8}} > \frac{103 - 100}{\frac{15}{8}}\right) \\ &\approx P(Z > 1.6) \approx 0.0548\,. \end{aligned}$$

or about 5.48%. Note that in the last equality we used the fact that Z is normally distributed with a mean of 0 and a standard deviation of 1.

Example 5.5

Suppose there is a referendum in a country, meaning that there is a vote in which all the people in a country are asked to decide whether or not to pass a law. In this country, 50 million people are eligible to vote. 5000 of them will vote to pass the law and the rest of the voters are either undecided or indifferent towards the outcome. In order for the law to be passed, 50% of the voters must vote in favor. Assume that all people will vote. Use the CLT to compute the likelihood of the law being passed. Discuss the result.

Solution 5.5

First of all, we have 50 million voters, and we know that 5000 of them will vote for the law. Since the rest of the voters are undecided or indifferent, we can assume that the probability of these voters voting for the law or against it is 50%, meaning that the distribution of each undecided voter is $X \sim \text{Bernoulli}(0\ .5\ )$. We want to compute the probability that the following random variable

$$\sum_{i=1}^{50.000.000 - 5000} X_i + 5000,$$

where $X_i \sim \text{Bernoulli}(0.5)$ for all $i = 1,2,\ldots$ 49.995.000 is bigger than 25.000.000. Let us take a look at

$$\begin{aligned} &P\Big(\sum_{i=1}^{50.000.000 - 5000} X_i + 5000 > 25.000.000\Big) \\ &= P\Big(\sum_{i=1}^{49.995.000} X_i > 24.995.000\Big) \\ &= P\Big(\frac{1}{49.995.000}\sum_{i=1}^{49.995.000} X_i > \frac{24.995.000}{49.995.000}\Big)\,. \end{aligned}$$

We compute

$$\begin{aligned} &\text{E}\Big(\frac{1}{49.995.000}\sum_{i=1}^{49.995.000} X_i\Big) = 0.5 = \mu, \\ &\text{Var}\Big[\frac{1}{49.995.000}\sum_{i=1}^{49.995.000} X_i\Big] \\ &= \frac{1}{49.995.000^{\,2}}\sum_{i=1}^{49.995.000} \text{Var}[X_i] = \frac{0.25}{49.995.000} = \sigma^2 \end{aligned}$$

and use the CLT to obtain

$$
\begin{aligned}
& P\left[\frac{1}{49.995.000}\sum_{i=1}^{49.995.000} X_i > \frac{24.995.000}{49.995.000}\right] \\
= & P\left[\frac{\frac{1}{49.995.000}\sum_{i=1}^{49.995.000} X_i - \mu}{\sigma} > \frac{\frac{24.995.000}{49.995.000} - \mu}{\sigma}\right] \\
\approx & P(Z > -0{,}7071) = P(Z \leq 0{,}7071) \approx 0{,}758,
\end{aligned}
$$

where Z is normally distributed with a mean of 0 and a standard deviation of 1. Hence the law will be passed with a likelihood of 75,8%. Although only 5000 of 50 million people were clearly in favor of the law to begin with, the probability is high that it will be established. It is almost as if the undecided voters were voting as if they had flipped a coin. From this we can infer that a small group of voters can have a high degree of influence on the results of a referendum even if they are ambivalent towards the outcome.

COMPUTING NORMAL PROBABILITY WITH SOFTWARE

To compute $P(Z > 0.2)$

In R use

pnorm(0.2,mean=0,sd=1,lower.tail=FALSE)

In Python use

from scipy import stats

1-stats.norm.cdf(0.2,loc=0,scale=1)

In Microsoft Excel use

=1-NORM.DIST(0.2,0,1,TRUE)

In Matlab use

normcdf(0.2,0,1,'upper')

SUMMARY

In this unit, we introduced Markov's inequality, a fundamental inequality from which many other inequalities follow. This inequality gives a bound for the tail probability of a non-negative random variable X only based on its mean.

A closely related inequality is that of Chebyshev, which works for any random variable as long as it has a finite variance. The third inequality, which rounds up our set of probability inequalities, is Hoeffding's inequality. This inequality makes a statement about the probability that the sample mean deviates from the true mean based on the bounds of the random variable and the sample size.

In section 5.2, we introduced two expectation inequalities. The Cauchy-Schwartz inequality gives a bound for the expectation of a product of two random variables. This inequality is closely related to the inequality for vectors with the same name. We also introduced the concepts of convex and concave functions and discussed Jensen's inequality. The weak law of large numbers showed us that the probability that the sample mean deviates from the true mean by, at most, a fixed amount can be made as small as possible by increasing the sample size.

The central limit theorem is a cornerstone of probability. It tells us that the sample mean of a random sample is approximately normally distributed if the sample size is large enough and the random variables are independent and identically distributed, even if we do not know the underlying distribution of these random variables. This is indeed a powerful tool, as we can make inferences about the true population mean using the normal distribution.

BACKMATTER

LIST OF REFERENCES

Kim, A. (2019a, August 6). *Exponential Distribution—Intuition, Derivation, and Applications*. Towards Data Science. https://link.medium.com/nyPAVfx1u1

Kim, A. (2019b, October 12). *Gamma Distribution—Intuition, Derivation and Examples.* Towards Data Science. https://medium.com/@aerinykim/gamma-distribution-intuition-derivation-and-examples-55f407423840

Kim, A. (2019c, June 1). *Poisson Distribution — Intuition, Examples, and Derivation.* Towards Data Science. https://towardsdatascience.com/poisson-distribution-intuition-and-derivation-1059aeab90d

Klenke, Achim (2014). *Probability theory: A comprehensive course* (2nd ed.). Springer.

Wackerly, D.D., Mendenhall, W., & Scheaffer, R.L. (2008). *Mathematical statistics with applications* (7**th** ed.). Thomson Brooks.

Wasserman, L. (2010). *All of statistics: A concise course in statistical inference*. Springer.

Rohatgi, V. K., & Saleh, A. K. E. (2015). *An introduction to probability and statistics*. John Wiley & Sons.

Amy S. Wagaman, & Robert P. Dobrow. (2021). *Probability: With applications and R.* Wiley.

Triola, M. F. (2013). *Elementary statistics*. Pearson Education.

LIST OF TABLES AND FIGURES

IU Internationale Hochschule GmbH
IU International University of Applied Sciences
Juri-Gagarin-Ring 152
D-99084 Erfurt

Mailing Address
Albert-Proeller-Straße 15-19
D-86675 Buchdorf

media@iu.org
www.iu.org

Help & Contacts (FAQ)
On myCampus you can always find answers to questions concerning your studies.