EMTH210 Engineering Mathematics
Elements of probability and statistics

Prepared by Raazesh Sainudiin

during Canterbury-Oxford Exchange Fellowship, Michaelmas Term 2013, Dept. of Statistics, Univ. of Oxford, Oxford, UK

Contents

EMTH119-13S1(C) Engineering Mathematics 1B
Elements of probability and statistics

Probability Theory provides the mathematical **models** of phenomena governed by chance. Knowledge of probability is required in such diverse fields as quantum mechanics, control theory, the design of experiments and the interpretation of data. Operation analysis applies probability methods to questions in traffic control, allocation of equipment, and the theory of strategy. **Statistical Theory** provides the mathematical methods to gauge the accuracy of the probability models based on observations or data.

For a simple example we have already met, consider the leaky bucket from the differential equation topic. We assumed water was flowing into the bucket at a constant flow rate a. But real taps spurt and sputter due to fluctuations in the water pressure, so there are random variations in the flow rate. If these variations are small we can usually ignore them, but for many modelling problems they are significant. We can model them by including a random term in our differential equation; this leads to the idea of *stochastic differential equations*, well beyond the scope of this course, but these are the sorts of problems that several lecturers in our department are dealing with at a research level.

Probabilistic problems are not just for mathematicians and statisticians. Engineers routinely use probability models in their work. Here are some examples of applications of probability and statistics at UC.

Extreme events in civil engineering

Dr. Brendon Bradley, a lecturer in civil engineering here at UC (and an ex-student of this course), is interested in the applications of statistics and probability in civil engineering. This is what he has to say about his work:

> "Extreme events of nature, such as earthquakes, floods, volcanic eruptions, and hurricanes, by definition, occur infrequently and are very uncertain. In the design of buildings, bridges, etc., probabilistic models are used to determine the likelihood or probability of such an extreme event occurring in a given period of time, and also the consequent damage and economic cost which result if the event occurs."

Clinical trials in biomedical engineering

Hannah Farr, who is doing her PhD in Mechanical engineering, but with a medical emphasis, has the following to say:

> "Medical research can involve investigating new methods of diagnosing diseases. However, you need to think carefully about the probabilities involved to make sure you make the correct diagnosis, especially with rare diseases. The probabilities involved are conditional and the sort of questions you have to ask are: How accurate is your method? How often does your method give a false negative or a false positive? What percentage of the population suffer from that disease?"

Markov control problems in engineering

Instrumentation of modern machines, such as planes, rockets and cars allow the sensors in the machines to collect live data and dynamically take *decisions* and subsequent *actions* by executing algorithms to drive their devices in response to the data that is streaming into their sensors. For example, a rocket may have to adjust its boosters to compensate for the prevailing directional changes in wind in order to keep going up and launch a satellite. These types of decisions and actions, theorised by *controlled Markov processes*, typically arise in various fields of engineering, such as, aerospace, civil, electrical, mechanical, robotics, etc.
When making mathematical models of real-world phenomenon let us not forget the following wise words.

"*Essentially, all models are wrong, but some are useful.*" — George E. P. Box.

1 The rudiments of set theory

WHY SETS?

This topic is about probability so why worry about sets?

Well, how about I ask you for the probability of a rook. This makes no sense unless I explain the context. I have a bag of chess pieces. There are 32 pieces in total. There are 4 rooks and 28 pieces that are not rooks. I pick one piece at random. Now it is a little clearer. I might describe my bag as

$$\text{Bag} = \{R_1, R_2, R_3, R_4, O_1, O_2, \ldots, O_{28}\}$$

where R_1, R_2, R_3, R_4 are the 4 rooks and $O_1, O_2, \ldots, O_{28}$ are the 28 other pieces. When I ask for the probability of a rook I am asking for the probability of one of R_1, R_2, R_3, R_4. Clearly, collections of objects (sets) are useful in describing probability questions.

1. A **set** is a collection of *distinct* objects or elements. We enclose the elements by *curly braces*.

   For example, the collection of the two letters H and T is a set and we denote it by $\{\mathsf{H}, \mathsf{T}\}$.

   But the collection $\{\mathsf{H}, \mathsf{T}, \mathsf{T}\}$ is not a set (do you see why? think *distinct*!).

   Also, order is not important, e.g., $\{\mathsf{H}, \mathsf{T}\}$ is the same as $\{\mathsf{T}, \mathsf{H}\}$.

2. We give convenient **names to sets**.

   For example, we can give the set $\{\mathsf{H}, \mathsf{T}\}$ the name A and write $A = \{\mathsf{H}, \mathsf{T}\}$.

3. If a is **an element** of A, we write $a \in A$. For example, if $A = \{1, 2, 3\}$, then $1 \in A$.

4. If a is **not an element** of A, we write $a \notin A$. For example, if $A = \{1, 2, 3\}$, then $13 \notin A$.

5. We say a set A is a **subset** of a set B if every element of A is also an element of B, and write $A \subseteq B$. For example, $\{1, 2\} \subseteq \{1, 2, 3, 4\}$.

6. A set A is **not a subset** of a set B if at least one element of A is not an element of B, and write $A \not\subseteq B$. For example, $\{1, 2, 5\} \not\subseteq \{1, 2, 3, 4\}$.

7. We say a set A **is equal to** a set B and write $A = B$ if A and B have the same elements.

8. The **union** $A \cup B$ of A and B consists of elements that are in A or in B or in both A and B.

   For example, if $A = \{1, 2\}$ and $B = \{3, 2\}$ then $A \cup B = \{1, 2, 3\}$.

   Similarly the union of m sets

   $$\bigcup_{j=1}^{m} A_j \;=\; A_1 \cup A_2 \cup \cdots \cup A_m$$

   consists of elements that are in at least one of the m sets $A_1, A_2, \ldots, A_m$.

9. The **intersection** $A \cap B$ of A and B consists of elements that are in both A and B.

   For example, if $A = \{1, 2\}$ and $B = \{3, 2\}$ then $A \cap B = \{2\}$.

   Similarly, the intersection

   $$\bigcap_{j=1}^{m} A_j \;=\; A_1 \cap A_2 \cap \cdots \cap A_m$$

   of m sets consists of elements that are in each of the m sets.

   Sets that have no elements in common are called **disjoint**.

   For example, if $A = \{1, 2\}$ and $B = \{3, 4\}$ then A and B are disjoint.

10. The **empty set** contains no elements and is denoted by $\emptyset$ or by $\{\}$.

    Note that for any set A, $\emptyset \subseteq A$.

11. Given some set, Ω (the Greek letter Omega), and a subset $A \subseteq \Omega$, then the **complement** of A, denoted by A^c, is the set of all elements in Ω that are not in A.

    For example, if $\Omega = \{\mathsf{H}, \mathsf{T}\}$ and $A = \{\mathsf{H}\}$ then $A^c = \{\mathsf{T}\}$.

    Note that for any set $A \subseteq \Omega$: $A^c \cap A = \emptyset, \quad A \cup A^c = \Omega, \quad \Omega^c = \emptyset, \quad \emptyset^c = \Omega$.

Exercise 1.1

Let Ω be the universal set of students, lecturers and tutors involved in this course.
Now consider the following subsets:

- The set of 50 students, $S = \{S_1, S_2, S_2, \ldots S_{50}\}$.
- The set of 3 lecturers, $L = \{L_1, L_2, L_3\}$.
- The set of 4 tutors, $T = \{T_1, T_2, T_3, L_3\}$.

Note that one of the lecturers also tutors in the course. Find the following sets:

(a) $T \cap L$

(b) $T \cap S$

(c) $T \cup L$

(d) $T \cup L \cup S$

(e) S^c

(f) $S \cap L$

(g) $S^c \cap L$

(h) T^c

(i) $T^c \cap L$

(j) $T^c \cap T$

Solution

(a) $T \cap L = \{L_3\}$

(b) $T \cap S = \emptyset$

(c) $T \cup L = \{T_1, T_2, T_3, L_3, L_1, L_2\}$

(d) $T \cup L \cup S = \Omega$

(e) $S^c = \{T_1, T_2, T_3, L_3, L_1, L_2\}$

(f) $S \cap L = \emptyset$

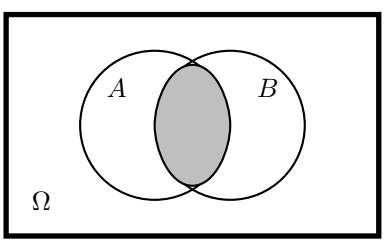
(g) $S^c \cap L = \{L_1, L_2, L_3\} = L$

(h) $T^c = \{L_1, L_2, S_1, S_2, S_2, \ldots S_{50}\}$

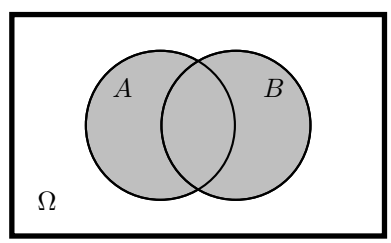
(i) $T^c \cap L = \{L_1, L_2\}$

(j) $T^c \cap T = \emptyset$

Venn diagrams are visual aids for set operations as in the diagrams below.



$A \cap B$



$A \cup B$

Exercise 1.2

Using the sets T and L in Exercise 1.1, draw Venn diagrams to illustrate the following intersections and unions.

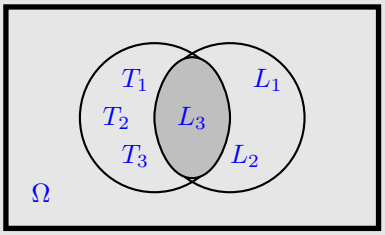
(a) $T \cap L$

(b) $T \cup L$

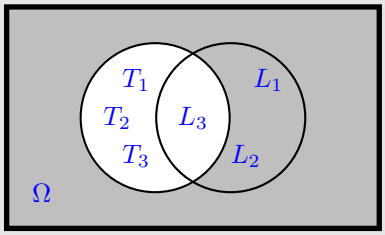
(c) T^c

(d) $T^c \cap L$

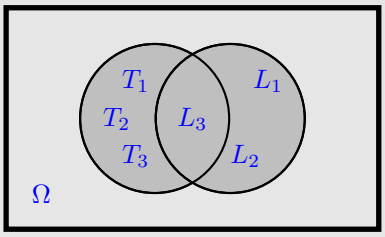
Solution



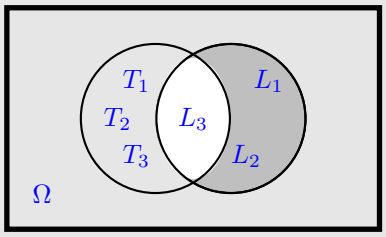
(a) $T \cap L$



(c) T^c



(b) $T \cup L$



(d) $T^c \cap L$

SET SUMMARY

$\{a_1, a_2, \ldots, a_n\}$	–	a set containing the elements, $a_1, a_2, \ldots, a_n$.
$a \in A$	–	a is an element of the set A.
$A \subseteq B$	–	the set A is a subset of B.
$A \cup B$	–	"union", meaning the set of all elements which are in A or B, or both.
$A \cap B$	–	"intersection", meaning the set of all elements in both A and B.
$\{\}$ or $\emptyset$	–	empty set.
Ω	–	universal set.
A^c	–	the complement of A, meaning the set of all elements in Ω, the universal set, which are not in A.

2 Experiments

Ideas about chance events and random behaviour arose out of thousands of years of game playing, long before any attempt was made to use mathematical reasoning about them. Board and dice games were well known in Egyptian times, and Augustus Caesar gambled with dice. Calculations of odds for gamblers were put on a proper theoretical basis by Fermat and Pascal in the early 17th century.

Definition 2.1 An **experiment** is an activity or procedure that produces distinct, well-defined possibilities called **outcomes**.

The set of all outcomes is called the **sample space**, and is denoted by Ω.

The subsets of Ω are called **events**.

A single outcome, ω, when seen as a subset of Ω, as in $\{\omega\}$, is called a **simple event**.

Events, E_1, $E_2 \ldots E_n$, that cannot occur at the same time are called **mutually exclusive** events, or **pair-wise disjoint** events. This means that $E_i \cap E_j = \emptyset$ where $i \neq j$.

Example 2.2 Some standard examples:

- $\Omega = \{\mathsf{Defective, Non\text{-}defective}\}$ if our experiment is to inspect a light bulb.

  There are only two outcomes here, so $\Omega = \{\omega_1, \omega_2\}$ where $\omega_1 = \mathsf{Defective}$ and $\omega_2 = \mathsf{Non\text{-}defective}$.

- $\Omega = \{\mathsf{Heads, Tails}\}$ if our experiment is to note the outcome of a coin toss.

  This time, $\Omega = \{\omega_1, \omega_2\}$ where $\omega_1 = \mathsf{Heads}$ and $\omega_2 = \mathsf{Tails}$.

- If our experiment is to roll a die then there are six outcomes corresponding to the number that shows on the top. For this experiment, $\Omega = \{\mathsf{1, 2, 3, 4, 5, 6}\}$.

  Some examples of events are the set of odd numbered outcomes $A = \{\mathsf{1, 3, 5}\}$, and the set of even numbered outcomes $B = \{\mathsf{2, 4, 6}\}$.

  The simple events of Ω are $\{\mathsf{1}\}$, $\{\mathsf{2}\}$, $\{\mathsf{3}\}$, $\{\mathsf{4}\}$, $\{\mathsf{5}\}$, and $\{\mathsf{6}\}$.

The outcome of a random experiment is uncertain until it is performed and observed. Note that sample spaces need to reflect the problem in hand. The example below is to convince you that an experiment's sample space is merely a collection of distinct elements called outcomes and these outcomes have to be *discernible in some well-specified sense* to the experimenter!

Example 2.3 Consider a generic die-tossing experiment by a human experimenter. Here $\Omega = \{\omega_1, \omega_2, \omega_3, \ldots, \omega_6\}$, but the experiment might correspond to rolling a die whose faces are:

1. sprayed with six different scents (nose!), or
2. studded with six distinctly flavoured candies (tongue!), or
3. contoured with six distinct bumps and pits (touch!), or
4. acoustically discernible at six different frequencies (ears!), or
5. painted with six different colours (eyes!), or

6. marked with six different numbers $\mathsf{1}, \mathsf{2}, \mathsf{3}, \mathsf{4}, \mathsf{5}, \mathsf{6}$ (eyes!), or , ...

These six experiments are equivalent as far as probability goes.

Definition 2.4 A **trial** is a single performance of an experiment and it results in an outcome.

Example 2.5 Some standard examples:

- A single roll of a die.
- A single toss of a coin.

An experimenter often performs more than one trial. Repeated trials of an experiment forms the basis of science and engineering as the experimenter learns about the phenomenon by repeatedly performing the same mother experiment with possibly different outcomes. This repetition of trials in fact provides the very motivation for the definition of probability.

Definition 2.6 An **n-product experiment** is obtained by repeatedly performing n trials of some experiment. The experiment that is repeated is called the "mother" experiment.

Exercise 2.7

Suppose we toss a coin twice (two trials here) and use the short-hand H and T to denote the outcome of Heads and Tails, respectively. Note that this is the 2-product experiment of the *coin toss* mother experiment.
Define the event A to be *at least one Head occurs*, and the event B to be *exactly one Head occurs*.

(a) Write down the sample space for this 2-product experiment in set notation.

(b) Write the sets A and B in set notation.

(c) Write B^c in set notation.

(d) Write $A \cup B$ in set notation.

(e) Write $A \cap B$ in set notation.

Solution

(a) The sample space is $\Omega = \{\mathsf{HH}, \mathsf{HT}, \mathsf{TH}, \mathsf{TT}\}$.

(b) The event that *at least one Head occurs* is $A = \{\mathsf{HH}, \mathsf{HT}, \mathsf{TH}\}$.
The event that *exactly one Head occurs* is $B = \{\mathsf{HT}, \mathsf{TH}\}$.

(c) $B^c = \{\mathsf{HH}, \mathsf{TT}\}$.

(d) $A \cup B = \{\mathsf{HH}, \mathsf{HT}, \mathsf{TH}\} = A$ since B is a subset of A.

(e) $A \cap B = \{\mathsf{HT}, \mathsf{TH}\} = B$ since B is a subset of A.

Remark 2.8 How big is a set?

Consider the set of students enrolled in this course. This is a **finite** set as we can count the number of elements in the set.

Loosely speaking, a set that can be tagged uniquely by natural numbers $\mathbb{N} = \{1, 2, 3, \ldots\}$ is said to be **countably infinite**. For example, it can be shown that the set of all integers $\mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, , 2, 3, \ldots\}$ is countably infinite.

The set of real numbers $\mathbb{R} = (-\infty, \infty)$, however, is **uncountably infinite**.

These ideas will become more important when we explore the concepts of discrete and continuous random variables.

EXPERIMENT SUMMARY

Experiment	–	an activity producing distinct outcomes.
Ω	–	set of all outcomes of the experiment.
ω	–	an individual outcome in Ω, called a simple event.
$A \subseteq \Omega$	–	a subset A of Ω is an event.
Trial	–	one performance of an experiment resulting in 1 outcome.

3 Background material: counting the possibilities

The most basic counting rule we use enables us to determine the number of distinct outcomes resulting from an experiment involving two or more steps, where each step has several different outcomes.

The multiplication principle: If a task can be performed in n_1 ways, a second task in n_2 ways, a third task in n_3 ways, etc., then the total number of distinct ways of performing all tasks together is

$$n_1 \times n_2 \times n_3 \times \ldots$$

Example 3.1 Suppose that a Personal Identification Number (PIN) is a six-symbol code word in which the first four entries are letters (lowercase) and the last two entries are digits. How many PINS are there? There are six selections to be made:

First letter: 26 possibilities	Fourth letter: 26 possibilities
Second letter: 26 possibilities	First digit: 10 possibilities
Third letter: 26 possibilities	Second digit: 10 possibilities

So in total, the total number of possible PINS is:

$$26 \times 26 \times 26 \times 26 \times 10 \times 10 \;=\; 26^4 \times 10^2 \;=\; 45,697,600\,.$$

Example 3.2 Suppose we now put restrictions on the letters and digits we use. For example, we might say that the first digit cannot be zero, and letters cannot be repeated. This time the the total number of possible PINS is:

$$26 \times 25 \times 24 \times 23 \times 9 \times 10 \;=\; 32,292,000\,.$$

When does order matter? In English we use the word "combination" loosely. If I say

"I have 17 probability texts on my bottom shelf"

then I don't care (usually) about what order they are in, but in the statement

"The combination of my PIN is math99"

I do care about order. A different order gives a different PIN.

So in mathematics, we use more precise language:

- A selection of objects in which the order is important is called a **permutation**.
- A selection of objects in which the order is *not* important is called a **combination**.

Permutations: There are basically two types of permutations:

1. Repetition is allowed, as in choosing the letters (unrestricted choice) in the PIN Example 3.1.

   More generally, when you have n objects to choose from, you have n choices each time, so when choosing r of them, the number of permutations are n^r.

2. No repetition is allowed, as in the restricted PIN Example 3.2. Here you have to reduce the number of choices. If we had a 26 letter PIN then the total permutations would be

$$26 \times 25 \times 24 \times 23 \times \ldots \times 3 \times 2 \times 1 = 26!$$

but since we want four letters only here, we have

$$\frac{26!}{22!} = 26 \times 25 \times 24 \times 23$$

choices.

The number of distinct **permutations** of n objects taking r at a time is given by

$$^{n}P_r = \frac{n!}{(n-r)!}$$

Combinations: There are also two types of combinations:

1. Repetition is allowed such as the coins in your pocket, say, ($10c, 50c$, $50c$, \$1, \$2, \$2).
2. No repetition is allowed as in the lottery numbers (2, 9, 11, 26, 29, 31). The numbers are drawn one at a time, and if you have the lucky numbers (no matter what order) you win! This is the type of combination we will need in this course.

Example 3.3 Suppose we need three students to be the class representatives in this course. In how many ways can we choose these three people from the class of 50 students? (Of course, everyone wants to be selected!) We start by assuming that order does matter, that is, we have a permutation, so that the number of ways we can select the three class representatives is

$$^{50}P_3 = \frac{50!}{(50-3)!} = \frac{50!}{47!}$$

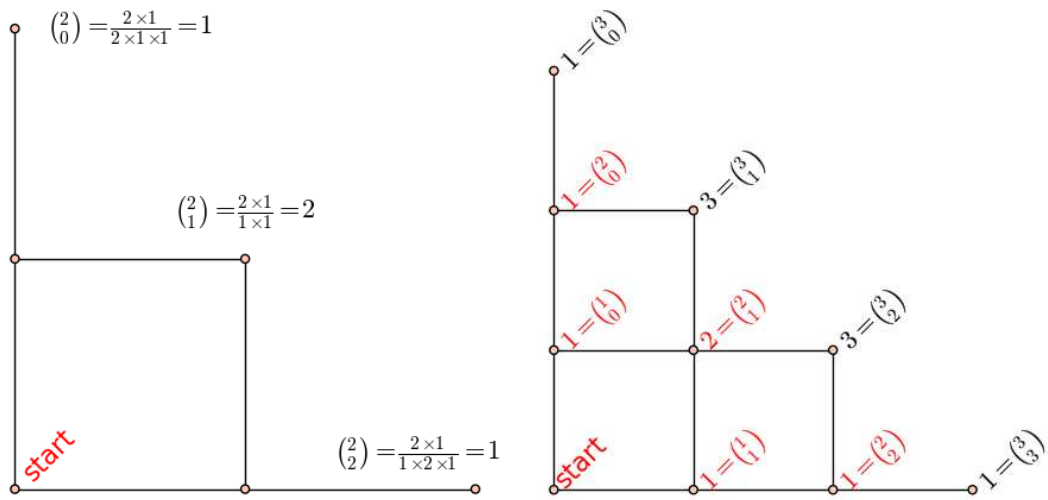
But, because order doesn't matter, all we have to do is to adjust our permutation formula by a factor representing the number of ways the objects could be in order. Here, three students can be placed in order 3! ways, so the required number of ways of choosing the class representatives is:

$$\frac{50!}{47!\,3!} = \frac{50\,.\,49\,.\,48}{3\,.\,2\,.\,1} = 19,600$$

The number of distinct **combinations** of n objects taking r at a time is given by

$$^{n}C_r = \binom{n}{r} = \frac{n!}{(n-r)!\,r!}$$

Example 3.4 Let us imagine being in the lower Manhattan in New York city with its perpendicular grid of streets and avenues. If you start at a given intersection and are asked to only proceed in a north-easterly direction then how may ways are there to reach another intersection by walking exactly two blocks or exactly three blocks?
Let us answer this question of combinations by drawing the following Figure.



(a) Walking two blocks north-easterly.

(b) Walking three blocks north-easterly.

Let us denote the number of easterly turns you take by r and the total number of blocks you are allowed to walk either easterly or northerly by n. From Figure (a) it is clear that the number of ways to reach each of the three intersections labeled by r is given by $\binom{n}{r}$, with $n = 2$ and $r \in \{0, 1, 2\}$. Similarly, from Figure (b) it is clear that the number of ways to reach each of the four intersections labeled by r is given by $\binom{n}{r}$, with $n = 3$ and $r \in \{0, 1, 2, 3\}$.

4 Probability

We all know the answer to the question "What is the probability of getting a head if I toss a coin once?", but is there a probability in the statement "I have a 50% chance of being late to my next lecture"? This section builds on your instinctive knowledge of probability derived from experience and from your previous mathematical studies to formalise the notions and language involved.

Definition 4.1 Probability is a function P that assigns real numbers to events, which satisfies the following four axioms:

Axiom (1): for any event A, $0 \leq P(A) \leq 1$

Axiom (2): if Ω is the sample space then $P(\Omega) = 1$

Axiom (3): if A and B are disjoint, i.e., $A \cap B = \emptyset$ then

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

Axiom (4): if $A_1, A_2, \ldots$ is an infinite sequence of pairwise disjoint, or mutually exclusive, events then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

where the infinite union

$$\bigcup_{j=1}^{\infty} A_j = A_1 \cup A_2 \cup \cdots \cup \cdots$$

is the set of elements that are in at least one of the sets $A_1, A_2, \ldots$.

Remark 4.2 There is nothing special about the scale $0 \leq P(A) \leq 1$ but it is traditional that probabilities are scaled to lie in this interval (or 0% to 100%) and that outcomes with probability 0 cannot occur. We also say that events with probability one or 100% are certain to occur.

These axioms are merely assumptions that are justified and motivated by the **frequency interpretation of probability in n-product experiments as n tends to infinity**, which states that if we repeat an experiment a large number of times then the fraction of times the event A occurs will be close to $P(A)$. [1] That is,

$$P(A) \;=\; \lim_{n\to\infty} \frac{N(A,n)}{n} \quad ,$$

where $N(A,n)$ is the number of times A occurs in the first n trials.

The Rules of Probability.

Theorem 4.3 Complementation Rule.

The probability of an event A and its complement A^c in a sample space Ω, satisfy

$$P(A^c) \;=\; 1 \;-\; P(A) \;. \tag{1}$$

Proof By the definition of complement, we have $\Omega = A \cup A^c$ and $A \cap A^c = \emptyset$. Hence by Axioms 2 and 3,

$$1 \;=\; P(\Omega) \;=\; P(A) \;+\; P(A^c), \text{ thus } P(A^c) = 1 - P(A).$$

[1] Probabilities obtained from theoretical models (Brendon's earthquakes) and from historical data (road statistics, the life insurance industry) are based on relative frequencies, but probabilities can be more subjective. For example, "I have a 50% chance of being late to my next lecture" is an assessment of what I believe will happen based on pooling all the relevant information (may be feeling sick, the car is leaking oil, etc.) and not on the idea of repeating something over and over again. The first few sections in "Chapter Four: Probabilities and Proportions" of the recommended text "Chance Encounters" discuss these ideas, and are well worth reading.

This formula is very useful in any situation where the probability of the complement of an event is easier to calculate than the probability of the event itself.

Theorem 4.4 Addition Rule for Mutually Exclusive Events.

For mutually exclusive or pair-wise disjoint events $A_1, \ldots, A_m$ in a sample space Ω,

$$P(A_1 \cup A_2 \cup A_3 \cup \cdots \cup A_m) \;=\; P(A_1) + P(A_2) + P(A_3) + \cdots + P(A_m) \;. \tag{2}$$

Proof: This is a consequence of applying Axiom (3) repeatedly:

$$\begin{aligned} P\left(A_1 \cup A_2 \cup A_3 \cup \cdots \cup A_m\right) &= P\left(A_1 \cup \left(A_2 \cup \cdots \cup A_m\right)\right) \\ &= P\left(A_1\right) + P\left(A_2 \cup \left(A_3 \cdots \cup A_m\right)\right) = P\left(A_1\right) + P\left(A_2\right) + P\left(A_3 \cdots \cup A_m\right) = \cdots \\ &= P(A_1) + P(A_2) + P(A_3) + \cdots + P(A_m) \;. \end{aligned}$$

Note: This can be more formally proved by induction.

Example 4.5 Let us observe the number on *the first ball that pops out in a New Zealand Lotto trial*. There are forty balls labelled 1 through 40 for this experiment and so the sample space is

$$\Omega \;=\; \{1, 2, 3, \ldots, 39, 40\}\,.$$

Because the balls are vigorously whirled around inside the Lotto machine before the first one pops out, we can model each ball to pop out first with the same probability. So, we assign each outcome $\omega \in \Omega$ the same probability of $\frac{1}{40}$, i.e., our probability model for this experiment is:

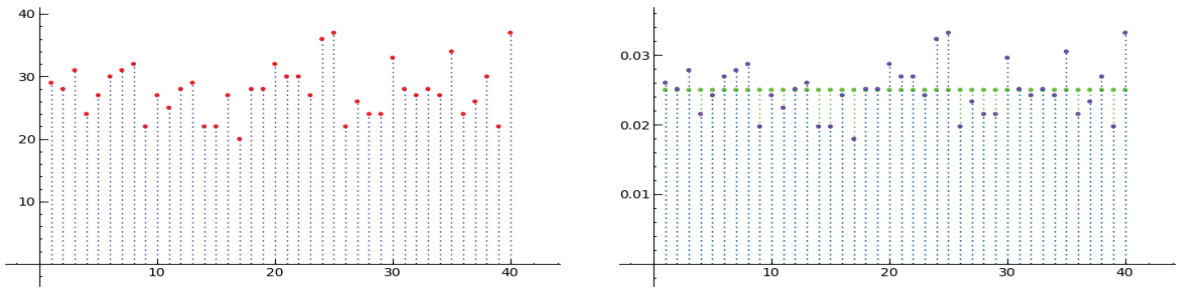
$$P(\omega) = \frac{1}{40}, \text{ for each } \;\omega \in \Omega = \{1, 2, 3, \ldots, 39, 40\} \;.$$

(Note: We sometimes abuse notation and write $P(\omega)$ instead of the more accurate but cumbersome $P(\{\omega\})$ when writing down probabilities of simple events.)
Figure 1 (a) shows the frequency of the first ball number in 1114 NZ Lotto draws. Figure 1 (b) shows the relative frequency, i.e., the frequency divided by 1114, the number of draws. Figure 1 (b) also shows the equal probabilities under our model.

Exercise 4.6

In the probability model of Example 4.5, show that for any event $E \subset \Omega$,

$$P(E) \;=\; \frac{1}{40} \;\times\; \text{number of elements in } E \;.$$



(a) Frequency of first Lotto ball.

(b) Relative frequency and probability of first Lotto ball.

Figure 1: First ball number in 1114 NZ Lotto draws from 1987 to 2008.

Solution Let $E = \{\omega_1, \omega_2, \ldots, \omega_k\}$ be an event with k outcomes (simple events). Then by Equation (2) of the addition rule for mutually exclusive events we get:

$$P(E) \;=\; P\left(\{\omega_1, \omega_2, \ldots, \omega_k\}\right) = P\left(\bigcup_{i=1}^{k}\{\omega_i\}\right) \;=\; \sum_{i=1}^{k} P\left(\{\omega_i\}\right) \;=\; \sum_{i=1}^{k}\frac{1}{40} \;=\; \frac{k}{40}\ .$$

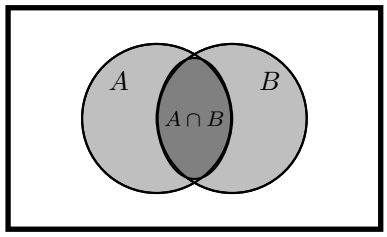
Recommended Activity 4.7 Explore the following web sites to learn more about NZ and British Lotto. The second link has animations of the British equivalent of NZ Lotto.
`http://lotto.nzpages.co.nz/`
`http://understandinguncertainty.org/node/39`

Theorem 4.8 Addition Rule for Two Arbitrary Events.

For events A and B in a sample space,

$$P(A \cup B) \;=\; P(A) \;+\; P(B) \;-\; P(A \cap B)\ . \tag{3}$$

Justification: First note that for two events A and B, the outcomes obtained when counting elements in A and B separately counts the outcomes in the "overlap" twice. This double counting must be corrected by subtracting the outcomes in the overlap – reference to the Venn digram below will help.



In probability terms, in adding $P(A)$ and $P(B)$, we add the probabilities relating to the outcomes in A *and* B twice so we must adjust by subtracting $P(A \cap B)$.

(Optional) Proof: Since $A \cup B = A \cup (B \cap A^c)$ and A and $B \cap A^c$ are mutually exclusive, we get

$$P(A \cup B) \;=\; P\left(A \cup (B \cap A^c)\right) \;=\; P(A) \;+\; P(B \cap A^c)$$

by Axiom (3). Now B may be written as the union of two mutually exclusive events as follows:

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

and so

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

which rearranges to give

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

Hence

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

Exercise 4.9

In English language text, the twenty six letters in the alphabet occur with the following frequencies:

E	13%	R	7.7%	A	7.3%	H	3.5%	F	2.8%	M	2.5%	W	1.6%	X	0.5%	J	0.2%
T	9.3%	O	7.4%	S	6.3%	L	3.5%	P	2.7%	Y	1.9%	V	1.3%	K	0.3%	Z	0.1%
N	7.8%	I	7.4%	D	4.4%	C	3%	U	2.7%	G	1.6%	B	0.9%	Q	0.3%		

Suppose you pick one letter at random from a randomly chosen English book from our central library with $\Omega = \{\mathsf{A}, \mathsf{B}, \mathsf{C}, \ldots, \mathsf{Z}\}$ (ignoring upper/lower cases), then what is the probability of these events?

(a) $P(\{\mathsf{Z}\})$

(b) $P($'picking any letter'$)$

(c) $P(\{\mathsf{E}, \mathsf{Z}\})$

(d) $P($'picking a vowel'$)$

(e) P('picking any letter in the word WAZZZUP')

(f) P('picking any letter in the word WAZZZUP or a vowel').

Solution

(a) $P(\{\mathsf{Z}\}) = 0.1\% = \frac{0.1}{100} = 0.001$

(b) P('picking any letter') $= P(\Omega) = 1$

(c) $P(\{\mathsf{E}, \mathsf{Z}\}) = P(\{\mathsf{E}\} \cup \{\mathsf{Z}\}) = P(\{\mathsf{E}\}) + P(\{\mathsf{Z}\}) = 0.13 + 0.001 = 0.131$, by Axiom (3)

(d) P('picking a vowel') $= P(\{\mathsf{A}, \mathsf{E}, \mathsf{I}, \mathsf{O}, \mathsf{U}\}) = (7.3\% + 13.0\% + 7.4\% + 7.4\% + 2.7\%) = 37.8\%$, by the addition rule for mutually exclusive events, rule (2).

(e) P('picking any letter in the word WAZZZUP') $= P(\{\mathsf{W}, \mathsf{A}, \mathsf{Z}, \mathsf{U}, \mathsf{P}\}) = 14.4\%$, by the addition rule for mutually exclusive events, rule (2).

(f) P('picking any letter in the word WAZZZUP or a vowel') $=$ $P(\{\mathsf{W}, \mathsf{A}, \mathsf{Z}, \mathsf{U}, \mathsf{P}\}) + P(\{\mathsf{A}, \mathsf{E}, \mathsf{I}, \mathsf{O}, \mathsf{U}\}) - P(\{\mathsf{A}, \mathsf{U}\}) = 14.4\% + 37.8\% - 10\% = 42.2\%$, by the addition rule for two arbitrary events, rule (3).

PROBABILITY SUMMARY

Axioms:

1. If $A \subseteq \Omega$ then $0 \leq P(A) \leq 1$ and $P(\Omega) = 1$.
2. If A, B are disjoint events, then $P(A \cup B) = P(A) + P(B)$.
   [This is true only when A and B are disjoint.]
3. If $A_1, A_2, \ldots$ are disjoint then $P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \ldots$

Rules:

$$P(A^c) = 1 - P(A)$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \qquad \text{[always true]}$$

5 Conditional probability

Conditional probabilities arise when we have partial information about the result of an experiment which restricts the sample space to a range of outcomes. For example, if there has been a lot of recent seismic activity in Christchurch, then the probability that an already damaged building will collapse tomorrow is clearly higher than if there had been no recent seismic activity.

Conditional probabilities are often expressed in English by phrases such as:

"If A happens, what is the probability that B happens?"

or

"What is the probability that A happens if B happens?"

or

" What is the probability that A occurs given that B occurs?"

Definition 5.1 The probability of an event B under the condition that an event A occurs is called the **conditional probability** of B given A, and is denoted by $P(B|A)$.
In this case A serves as a new (reduced) sample space, and the probability is the fraction of $P(A)$ which corresponds to $A \cap B$. That is,

$$P(B|A) = \frac{P(A \cap B)}{P(A)}, \qquad \text{if } \; P(A) \neq 0 \; . \tag{4}$$

Similarly, the conditional probability of A given B is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \qquad \text{if } \; P(B) \neq 0 \; . \tag{5}$$

Remember that conditional probabilities *are* probabilities, and so the axioms and rules of normal probabilities hold.

Axioms:

Axiom (1): For any event B, $0 \leq P(B|A) \leq 1$.

Axiom (2): $P(\Omega|A) = 1$.

Axiom (3): For any two disjoint events B_1 and B_2, $P(B_1 \cup B_2|A) = P(B_1|A) + P(B_2|A)$.

Axiom (4): For mutually exclusive or pairwise-disjoint events, $B_1, B_2, \ldots$,

$$P(B_1 \cup B_2 \cup \cdots |A) = P(B_1|A) + P(B_2|A) + \cdots \; .$$

Rules:

1. Complementation rule: $P(B|A) \; = \; 1 - P(B^c|A)$.
2. Addition rule for two arbitrary events B_1 and B_2:

$$P(B_1 \cup B_2|A) \; = \; P(B_1|A) + P(B_2|A) - P(B_1 \cap B_2|A) \, .$$

Solving for $P(A \cap B)$ with these definitions of conditional probability gives another rule:

Theorem 5.2 Multiplication Rule.

If A and B are events, and if $P(A) \neq 0$ and $P(B) \neq 0$, then

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

Exercise 5.3

Suppose the NZ All Blacks team is playing in a four team Rugby tournament. In the first round they have a tough opponent that they will beat 40% of the time but if they win that game they will play against an easy opponent where their probability of success is 0.8.

(a) What is the probability that they will win the tournament?

(b) What is the probability that the All Blacks will win the first game but lose the second?

Solution

(a) To win the tournament the All Blacks have to win in round one, W_1, *and* win in round two, W_2. We need to find $P(W_1 \cap W_2)$.

The probability that they win in round one is $P(W_1) = 0.4$, and the probability that they win in round two *given* that they have won in round one is the conditional probability, $P(W_2|W_1) = 0.8$.

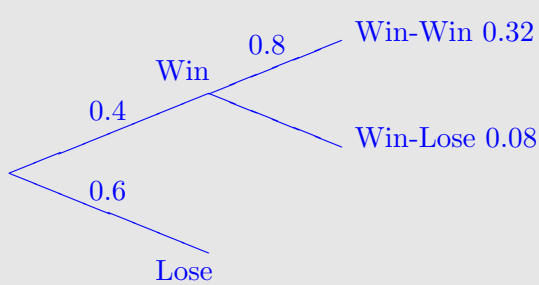
Therefore the probability that they will win the tournament is

$$P(W_1 \cap W_2) \;=\; P(W_1)\, P(W_2|W_1) \;=\; 0.4 \times 0.8 \;=\; 0.32\,.$$

(b) Similarly, the probability that the All Blacks win the first game and lose the second is

$$P(W_1 \cap L_2) \;=\; P(W_1)\, P(L_2|W_1) \;=\; 0.4 \times 0.2 \;=\; 0.08\,.$$

A **probability tree diagram** is a useful tool for tackling problems like this. The probability written beside each branch in the tree is the probability that the following event (at the right-hand end of the branch) occurs *given* the occurrence of all the events that have appeared along the path so far (reading from left to right). Because the probability information on a branch is conditional on what has gone before, the order of the tree branches should reflect the type of information that is available.



Multiplying along the first path gives the probability that the All Blacks will win the tournament, $0.4 \times 0.8 \;=\; 0.32$, and multiplying along the second path gives the probability that the All Blacks will win the first game but lose the second, $0.4 \times 0.2 \;=\; 0.08$.

Independent Events

In general, $P(A|B)$ and $P(A)$ are different, but sometimes the occurrence of B makes no difference, and gives no new information about the chances of A occurring. This is the idea behind independence. Events like "having blue eyes" and "having blond hair" are associated, but events like "my sister wins lotto" and "I win lotto" are not.

Definition 5.4 If two events A and B are such that

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

they are called **independent events**.

This means that $P(A|B) = P(A)$, and $P(B|A) = P(B)$, assuming that $P(A) \neq 0$, $P(B) \neq 0$, which justifies the term "independent" since the probability of A will not depend on the occurrence or nonoccurence of B, and conversely, the probability of B will not depend on the occurrence or nonoccurence of A.

Similarly, n events $A_1, \ldots, A_n$ are called **independent** if

$$P(A_1 \cap \cdots \cap A_n) \;=\; P(A_1)P(A_2)\cdots P(A_n) \;\; .$$

Example 5.5 Some Standard Examples

(a) Suppose you toss a fair coin twice such that the first toss is independent of the second. Then,

$$P(\mathsf{HT}) = P(\mathsf{Heads}\text{ on the first toss} \cap \mathsf{Tails}\text{ on the second toss}) = P(\mathsf{H})P(\mathsf{T}) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \;\; .$$

(b) Suppose you independently toss a fair die three times. Let E_i be the event that the outcome is an even number on the i-th trial. The probability of getting an even number in all three trials is:

$$\begin{aligned}
P(E_1 \cap E_2 \cap E_3) &= P(E_1)P(E_2)P(E_3) \\
&= (P(\{2,4,6\}))^3 \\
&= (P(\{2\} \cup \{4\} \cup \{6\}))^3 \\
&= (P(\{2\}) + P(\{4\}) + P(\{6\}))^3 \\
&= \left(\frac{1}{6} + \frac{1}{6} + \frac{1}{6}\right)^3 \\
&= \left(\frac{1}{2}\right)^3 \\
&= \frac{1}{8} \;\; .
\end{aligned}$$

This is an obvious answer but there is a lot of maths going on here!

(c) Suppose you toss a fair coin independently m times. Then each of the 2^m possible outcomes in the sample space Ω has equal probability of $\frac{1}{2^m}$ due to independence.

Theorem 5.6 Total probability theorem.

Suppose $B_1 \cup B_2 \cdots \cup B_n$ is a sequence of events with positive probability that partition the sample space, that is, $B_1 \cup B_2 \cdots \cup B_n = \Omega$ and $B_i \cap B_j = \emptyset$ for $i \neq j$, then

$$P(A) \;=\; \sum_{i=1}^{n} P(A \cap B_i) \;=\; \sum_{i=1}^{n} P(A|B_i)P(B_i) \tag{6}$$

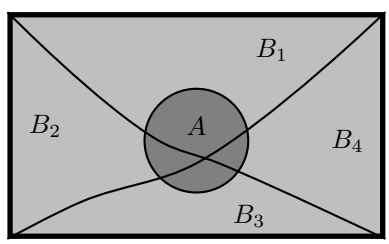
for some arbitrary event A.

Proof: The first equality is due to the addition rule for mutually exclusive events,

$$A \cap B_1, A \cap B_2, \ldots, A \cap B_n$$

and the second equality is due to the multiplication rule.

Reference to the Venn digram below will help you understand this idea for the four event case.

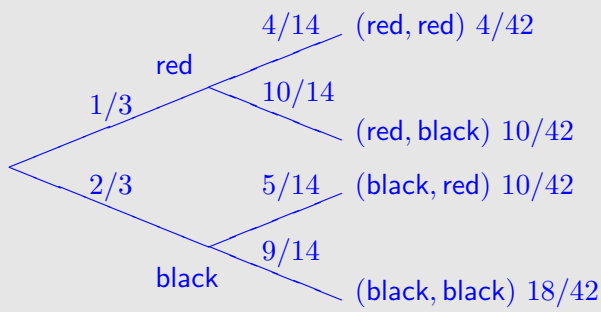


Exercise 5.7

A well-mixed urn contains five red and ten black balls. We draw two balls from the urn without replacement. What is the probability that the second ball drawn is red?

Solution This is easy to see if we draw a probability tree. The first split in the tree is based on the outcome of the first draw and the second on the outcome of the last draw. The outcome of the first draw dictates the probabilities for the second one since we are sampling without replacement. We multiply the probabilities on the edges to get probabilities of the four endpoints, and then sum the ones that correspond to red in the second draw, that is

$$P(\text{second ball is red}) \;=\; 4/42 + 10/42 \;=\; 1/3$$



Alternatively, use the total probability theorem to break the problem down into manageable pieces.

Let $R_1 = \{(\mathsf{red}, \mathsf{red}), (\mathsf{red}, \mathsf{black})\}$ and $R_2 = \{(\mathsf{red}, \mathsf{red}), (\mathsf{black}, \mathsf{red})\}$ be the events corresponding to a red ball in the 1st and 2nd draws, respectively, and let $B_1 = \{(\mathsf{black}, \mathsf{red}), (\mathsf{black}, \mathsf{black})\}$ be the event of a black ball on the first draw.

Now R_1 and B_1 partition Ω so we can write:

$$\begin{aligned} P(R_2) &= P(R_2 \cap R_1) + P(R_2 \cap B_1) \\ &= P(R_2|R_1)P(R_1) + P(R_2|B_1)P(B_1) \\ &= (4/14)(1/3) + (5/14)(2/3) \\ &= 1/3 \end{aligned}$$

Many problems involve reversing the order of conditional probabilities. Suppose we want to investigate some phenomenon A and have an observation B that is evidence about A: for example, A may be breast cancer and B may be a positive mammogram. Then Bayes' Theorem tells us how we should update our probability of A, given the new evidence B.

Or, put more simply, Bayes' Theorem is useful when you know $P(B|A)$ but want $P(A|B)$!

Theorem 5.8 Bayes' theorem.

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \quad . \tag{7}$$

Proof: From the definition of conditional probability and the multiplication rule we get:

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

Exercise 5.9

Approximately 1% of women aged 40–50 have breast cancer. A woman with breast cancer has a 90% chance of a positive test from a mammogram, while a woman without breast cancer has a 10% chance of a false positive result from the test.

What is the probability that a woman indeed has breast cancer given that she just had a positive test?

Solution Let $A =$"the woman has breast cancer", and $B =$"a positive test."

We want $P(A|B)$ but what we are given is $P(B|A) = 0.9$.

By the definition of conditional probability,

$$P(A|B) \;=\; P(A \cap B)/P(B)$$

To evaluate the numerator we use the multiplication rule

$$P(A \cap B) \;=\; P(A)P(B|A) \;=\; 0.01 \times 0.9 \;=\; 0.009$$

Similarly,

$$P(A^c \cap B) \;=\; P(A^c)P(B|A^c) \;=\; 0.99 \times 0.1 \;=\; 0.099$$

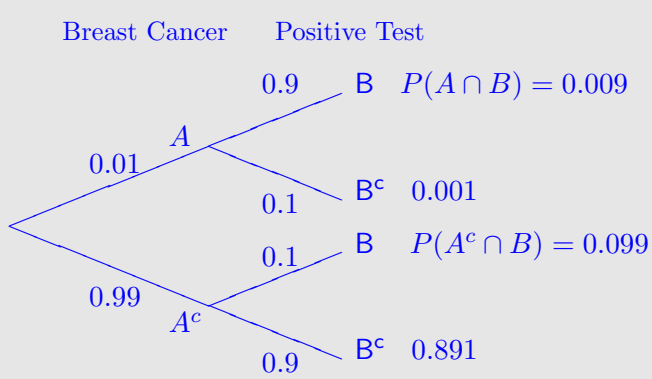
Now $P(B) \;=\; P(A \cap B) + P(A^c \cap B)$ so

$$P(A|B) \;=\; \frac{P(A \cap B)}{P(B)} \;=\; \frac{0.009}{0.009 + 0.099} \;=\; \frac{9}{108}$$

or a little less than 9%. This situation comes about because it is much easier to have a false positive for a healthy woman, which has probability 0.099, than to find a woman with breast cancer having a positive test, which has probability 0.009.

This answer is somewhat surprising. Indeed when ninety-five physicians were asked this question their average answer was 75%. The two statisticians who carried out this survey indicated that physicians were better able to see the answer when the data was presented in frequency format. 10 out of 1000 women have breast cancer. Of these 9 will have a positive mammogram. However of the remaining 990 women without breast cancer 99 will have a positive reaction, and again we arrive the answer $9/(9 + 99)$.

Alternative solution using a tree diagram:



So the probability that a woman has breast cancer given that she has just had a positive test is

$$P(A|B) \;=\; \frac{P(A\cap B)}{P(B)} \;=\; \frac{0.009}{0.009+0.099} \;=\; \frac{9}{108}$$

CONDITIONAL PROBABILITY SUMMARY

$P(A|B)$ means the probability that A occurs given that B has occurred.

$$P(A|B) \;=\; \frac{P(A\cap B)}{P(B)} \;=\; \frac{P(A)P(B|A)}{P(B)} \quad \text{if} \quad P(B)\neq 0$$

$$P(B|A) \;=\; \frac{P(A\cap B)}{P(A)} \;=\; \frac{P(B)P(A|B)}{P(A)} \quad \text{if} \quad P(A)\neq 0$$

Conditional probabilities obey the 4 axioms of probability.

6 Random variables

We are used to traditional variables such as x as an "unknown" in the equation: $x+3 \;=\; 7$.

We also use traditional variables to represent geometric objects such as a line:

$$y \;=\; 3x-2 \;,$$

where the variable y for the y-axis is determined by the value taken by the variable x, as x varies over the real line $\mathbb{R}=(-\infty,\infty)$.

Yet another example is the use of variables to represent sequences such as:

$$\{a_n\}_{n=1}^{\infty} \;=\; a_1, a_2, a_3, \ldots \;\; .$$

What these variables have in common is that they take a fixed or deterministic value when we can solve for them.

We need a new kind of variable to deal with real-world situations where the same variable may take different values in a non-deterministic manner. **Random variables** do this job for us. Random variables, unlike traditional deterministic variables can take a bunch of different values.

Definition 6.1 A **random variable** is a function from the sample space Ω to the set of real numbers $\mathbb{R}$, that is, $X:\Omega\to\mathbb{R}$.

Example 6.2 Suppose our experiment is to observe whether it will rain or not rain tomorrow. The sample space of this experiment is $\Omega = \{\mathsf{rain}, \mathsf{not\ rain}\}$. We can associate a random variable X with this experiment as follows:

$$X(\omega) = \begin{cases} 1, & \text{if } \omega = \mathsf{rain} \\ 0, & \text{if } \omega = \mathsf{not\ rain} \end{cases}$$

Thus, X is 1 if it will rain tomorrow and 0 otherwise. Note that another equally valid (though possibly not so useful) random variable , say Y, for this experiment is:

$$Y(\omega) = \begin{cases} \pi, & \text{if } \ \omega = \mathsf{rain} \\ \sqrt{2}, & \text{if } \ \omega = \mathsf{not\ rain} \end{cases}$$

Example 6.3 Suppose our experiment instead is to measure the volume of rain that falls into a large funnel stuck on top of a graduated cylinder that is placed at the centre of Ilam Field. Suppose the cylinder is graduated in millimeters then our random variable $X(\omega)$ can report a non-negative real number given by the lower miniscus of the water column, if any, in the cylinder tomorrow. Thus, $X(\omega)$ will measure the volume of rain in millilitres that will fall into our funnel tomorrow.

Example 6.4 Suppose ten seeds are planted. Perhaps fewer than ten will actually germinate. The number which do germinate, say X, must be one of the numbers

$$0,\ 1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8,\ 9,\ 10\,;$$

but until the seeds are actually planted and allowed to germinate it is impossible to say which number it is. The number of seeds which germinate is a variable, but it is not necessarily the same for each group of ten seeds planted, but takes values from the same set. As X is not known in advance it is called a **random variable**. Its value cannot be known until we actually experiment, and plant the seeds.

Certain things can be said about the value a random variable might take. In the case of these ten seeds we can be sure the number that germinate is less than eleven, and not less than zero! It may also be known that that the probability of seven seeds germinating is greater than the probability of one seed; or perhaps that the number of seeds germinating averages eight. These statements are based on probabilities unlike the sort of statements made about traditional variables.

Discrete versus continuous random variables.

A **discrete** random variable is one in which the set of possible values of the random variable is finite or at most countably infinite, whereas a **continuous** random variable may take on any value in some range, and its value may be any real value in that range (Think: uncountably infinite). Examples 6.2 and 6.4 are about discrete random variables and Example 6.3 is about a continuous random variable.

Discrete random variables are usually generated from experiments where things are "counted" rather than "measured" such as the seed planting experiment in Example 6.4. Continuous random variables appear in experiments in which we measure, such as the amount of rain, in millilitres in Example 6.3.

Random variables as functions.

In fact, random variables are actually functions! They take you from the "world of random processes and phenomena" to the world of real numbers. In other words, a random variable is a numerical value determined by the outcome of the experiment.

We said that a random variable can take one of many values, but we cannot be certain of which value it will take. However, *we can make probabilistic statements about the value x the random variable X will take.* A question like,

"What is the probability of it raining tomorrow?"

in the rain/not experiment of Example 6.2 becomes

"What is $P(\{\omega : X(\omega) = 1\})$?"

or, more simply,

"What is $P(X = 1)$?"

Definition 6.5 The **distribution function**, $F : \mathbb{R} \to [0, 1]$, of the random variable X is

$$F(x) \;=\; P(X \leq x) \;=\; P(\{\omega : X(\omega) \leq x\}) \text{, for any } x \in \mathbb{R} \;. \tag{8}$$

where the right-hand side represents the probability that the random variable X takes on a value less than or equal to x.

(The distribution function is sometimes called the cumulative distribution function.)

Remark 6.6 It is enough to understand the idea of random variables as functions, and work with random variables using simplified notation like

$$P(2 \leq X \leq 3)$$

rather than

$$P(\{\omega : 2 \leq X(\omega) \leq 3\})$$

but note that in advanced work this sample space notation is needed to clarify the true meaning of the simplified notation.

From the idea of a distribution function, we get:

Theorem 6.7 The probability that the random variable X takes a value x in the half-open interval $(a, b]$, i.e., $a < x \leq b$, is:

$$P(a < X \leq b) \;=\; F(b) - F(a) \;. \tag{9}$$

Proof Since $(X \le a)$ and $(a < X \le b)$ are disjoint events whose union is the event $(X \le b)$,

$$F(b) = P(X \le b) = P(X \le a) + P(a < X \le b) = F(a) + P(a < X \le b) \ .$$

Subtraction of $F(a)$ from both sides of the above equation yields Equation (9).

Exercise 6.8

Consider the *fair coin toss experiment* with $\Omega = \{\mathsf{H}, \mathsf{T}\}$ and $P(\mathsf{H}) = P(\mathsf{T}) = 1/2$.

We can associate a random variable X with this experiment as follows:

$$X(\omega) = \begin{cases} 1, & \text{if } \ \omega = \mathsf{H} \\ 0, & \text{if } \ \omega = \mathsf{T} \end{cases}$$

Find the distribution function for X.

Solution The probability that X takes on a specific value x is:

$$P(X = x) = P(\{\omega : X(\omega) = x\}) = \begin{cases} P(\emptyset) = 0, & \text{if } \ x \notin \{0, 1\} \\ P(\{\mathsf{T}\}) = \frac{1}{2}, & \text{if } \ x = 0 \\ P(\{\mathsf{H}\}) = \frac{1}{2}, & \text{if } \ x = 1 \end{cases}$$

or more simply,

$$P(X = x) = \begin{cases} \frac{1}{2} & \text{if } x = 0 \\ \frac{1}{2} & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases}$$

The distribution function for X is:

$$F(x) \ = \ P(X \le x) \ = \ P(\{\omega : X(\omega) \le x\}) \ = \ \begin{cases} P(\emptyset) = 0, & \text{if } \ -\infty < x < 0 \\ P(\{\mathsf{T}\}) = \frac{1}{2}, & \text{if } \ 0 \le x < 1 \\ P(\{\mathsf{H}, \mathsf{T}\}) = P(\Omega) = 1, & \text{if } \ 1 \le x < \infty \end{cases}$$

or more simply,

$$F(x) \ = \ P(X \le x) \ = \ \begin{cases} 0, & \text{if } \ -\infty < x < 0 \\ \frac{1}{2}, & \text{if } \ 0 \le x < 1 \\ 1, & \text{if } \ 1 \le x < \infty \end{cases}$$

7 Discrete random variables and their distributions

Definition 7.1 If X is a discrete random variable that assumes the values $x_1, x_2, x_3, \ldots$ with probabilities $p_1 = P(X = x_1)$, $p_2 = P(X = x_2)$ $p_3 = P(X = x_3)$ $\ldots$, then the **probability mass function** (or PMF) f of X is:

$$f(x) = P(X = x) = P\left(\{\omega : X(\omega) = x\}\right) = \begin{cases} p_i & \text{if } x = x_i, \text{ where } i = 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases} . \tag{10}$$

From this we get the values of the distribution function, $F(x)$ by simply taking sums,

$$\boxed{F(x) = \sum_{x_i \leq x} f(x_i) = \sum_{x_i \leq x} p_i} . \tag{11}$$

Two useful formulae for discrete distributions are readily obtained as follows. For the probability corresponding to intervals we have

$$\boxed{P(a < X \leq b) = F(b) - F(a) = \sum_{a < x_i \leq b} p_i} . \tag{12}$$

This is the sum of all probabilities p_i for which x_i satisfies $a < x_i \leq b$. From this and $P(\Omega) = 1$ we obtain the following formula that the sum of all probabilities is 1.

$$\boxed{\sum_i p_i = 1} . \tag{13}$$

DISCRETE RANDOM VARIABLES - SIMPLIFIED NOTATION

Notice that in equations (10) and (11), the use of the ω, Ω notation, where random variables are defined as functions, is much reduced. The reason is that in straightforward examples it is convenient to associate the possible values $x_1, x_2, \ldots$ with the outcomes $\omega_1, \omega_2, \ldots$ Hence, we can describe a discrete random variable by the table:

Possible values: x_i	x_1	x_2	x_3	$\ldots$
Probability: $P(X = x_i) = p_i$	p_1	p_2	p_3	$\ldots$

Table 1: Simplified notation for discrete random variables

Note that this table hides the more complex notation but it is still there, under the surface. In MATH100, you should be able to work with and manipulate discrete random variables using the simplified notation given in Table 1. The same comment applies to the continuous random variables discussed later.

Out of the class of discrete random variables we will define specific kinds as they arise often in applications. We classify discrete random variables into three types for convenience as follows:

- Discrete uniform random variables with finitely many possibilities
- Discrete non-uniform random variables with finitely many possibilities
- Discrete non-uniform random variables with (countably) infinitely many possibilities

Definition 7.2 Discrete Uniform Random Variable. We say that a discrete random variable X is uniformly distributed over k possible values $\{x_1, x_2, \ldots, x_k\}$ if its probability mass function is:

$$f(x) = \begin{cases} p_i = \frac{1}{k} & \text{if } x = x_i, \quad \text{where } i = 1, 2, \ldots, k \ , \\ 0 & \text{otherwise} \ . \end{cases} \tag{14}$$

The distribution function for the discrete uniform random variable X is:

$$F(x) = \sum_{x_i \le x} f(x_i) = \sum_{x_i \le x} p_i = \begin{cases} 0 & \text{if } -\infty < x < x_1 \ , \\ \frac{1}{k} & \text{if } x_1 \le x < x_2 \ , \\ \frac{2}{k} & \text{if } x_2 \le x < x_3 \ , \\ \vdots & \\ \frac{k-1}{k} & \text{if } x_{k-1} \le x < x_k \ , \\ 1 & \text{if } x_k \le x < \infty \ . \end{cases} \tag{15}$$

Exercise 7.3
The *fair coin toss experiment* of Exercise 6.8 is an example of a discrete uniform random variable with finitely many possibilities. Its probability mass function is given by

$$f(x) \ = \ P(X = x) \ = \ \begin{cases} \frac{1}{2} & \text{if } x = 0 \\ \frac{1}{2} & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases}$$

and its distribution function is given by

$$F(x) \ = \ P(X \le x) \ = \ \begin{cases} 0, & \text{if } \ -\infty < x < 0 \\ \frac{1}{2}, & \text{if } \ 0 \le x < 1 \\ 1, & \text{if } \ 1 \le x < \infty \end{cases}$$

Sketch the probability mass function and distribution function for X. (Notice the stair-like nature of the distribution function here.)

Solution

$f(x)$ and $F(x)$ of *the fair coin toss* random variable X

Exercise 7.4

Now consider the *toss a fair die* experiment and define X to be the number that shows up on the top face. Note that here Ω is the set of numerical symbols that label each face while each of these symbols are associated with the real number $x \in \{1, 2, 3, 4, 5, 6\}$. We can describe this random variable by the table

Possible values, x_i	1	2	3	4	5	6
Probability, p_i	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

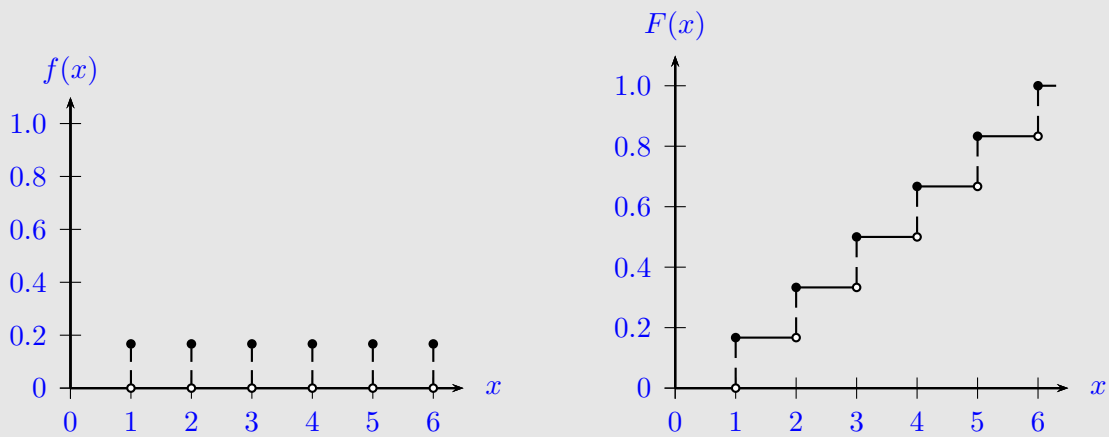
Find the probability mass function and distribution function for this random variable, and sketch their graphs.

Solution The probability mass function of this random variable is:

$$f(x) \;=\; P(X = x) \;=\; \begin{cases} \frac{1}{6} & \text{if } x = 1 \\ \frac{1}{6} & \text{if } x = 2 \\ \frac{1}{6} & \text{if } x = 3 \\ \frac{1}{6} & \text{if } x = 4 \\ \frac{1}{6} & \text{if } x = 5 \\ \frac{1}{6} & \text{if } x = 6 \\ 0 & \text{otherwise} \end{cases}$$

and the distribution function is:

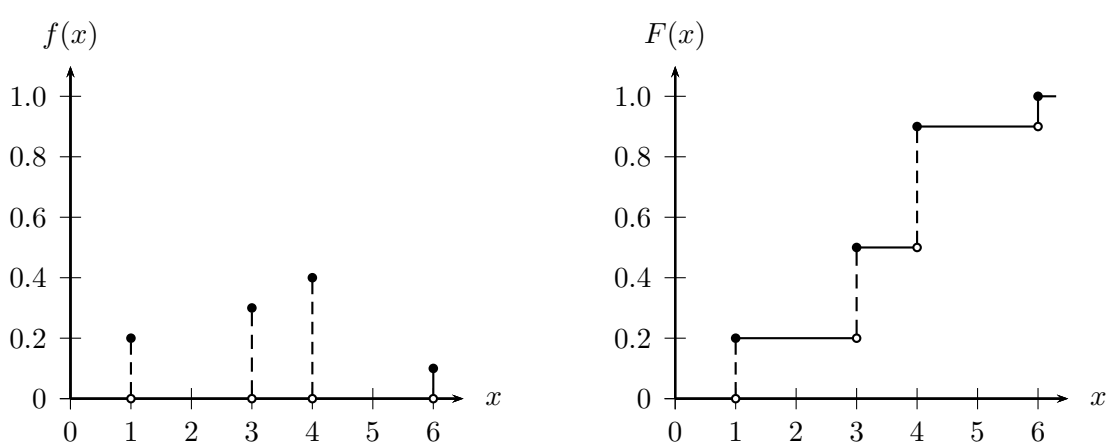
$$F(x) \;=\; P(X \leq x) \;=\; \begin{cases} 0, & \text{if } \; -\infty < x < 1 \\ \frac{1}{6}, & \text{if } \; 1 \leq x < 2 \\ \frac{1}{3}, & \text{if } \; 2 \leq x < 3 \\ \frac{1}{2}, & \text{if } \; 3 \leq x < 4 \\ \frac{2}{3}, & \text{if } \; 4 \leq x < 5 \\ \frac{5}{6}, & \text{if } \; 5 \leq x < 6 \\ 1, & \text{if } \; 6 \leq x < \infty \end{cases}$$



$f(x)$ and $F(x)$ of *the fair die toss* random variable X

Example 7.5 Astragali. Board games involving chance were known in Egypt, 3000 years before Christ. The element of chance needed for these games was at first provided by tossing astragali, the ankle bones of sheep. These bones could come to rest on only four sides, the other two sides being rounded. The upper side of the bone, broad and slightly convex counted four; the opposite side broad and slightly concave counted three; the lateral side flat and narrow, one, and the opposite narrow lateral side, which is slightly hollow, six. You may examine an astragali of a kiwi sheep (Ask at Maths & Stats Reception to access it).

This is an example of a discrete non-uniform random variable with finitely many possibilities. A surmised probability mass function with $f(4) = \frac{4}{10}$, $f(3) = \frac{3}{10}$, $f(1) = \frac{2}{10}$, $f(6) = \frac{1}{10}$ and distribution function are shown below.



$f(x)$ and $F(x)$ of an astragali toss random variable X

In many experiments there are only two outcomes. For instance:

- Flip a coin to see whether it is defective.
- Roll a die and determine whether it is a 6 or not.
- Determine whether there was flooding this year at the old Waimakariri bridge or not.

We call such an experiment a **Bernoulli trial**, and refer to the two outcomes – often arbitrarily – as success and failure.

Definition 7.6 Bernoulli(θ) Random Variable. Given a parameter $\theta \in (0, 1)$, the probability mass function and distribution function for the Bernoulli(θ) random variable X are:

$$f(x;\theta) = \begin{cases} \theta & \text{if } x = 1 \ , \\ 1-\theta & \text{if } x = 0 \ , \\ 0 & \text{otherwise} \ , \end{cases} \qquad F(x;\theta) = \begin{cases} 1 & \text{if } 1 \leq x \ , \\ 1-\theta & \text{if } 0 \leq x < 1 \ , \\ 0 & \text{otherwise} \end{cases}$$

We emphasise the dependence of the probabilities on the parameter θ by specifying it following the semicolon in the argument for f and F.

Random variables make sense for a series of experiments as well as just a single experiment. We now look at what happens when we perform a sequence of independent Bernoulli trials. For instance:

- Flip a coin 10 times; count the number of heads.
- Test 50 randomly selected circuits from an assembly line; count the number of defective circuits.
- Roll a die 100 times; count the number of sixes you throw.
- Provide a property near the Waimak bridge with flood insurance for 20 years; count the number of years, during the 20-year period, during which the property is flooded. Note: we assume that flooding is independent from year to year, and that the probability of flooding is the same each year.

Exercise 7.7

Suppose our experiment is to toss a fair coin independently and identically (that is, the same coin is tossed in essentially the same manner independent of the other tosses in each trial) as often as necessary until we have a head, H. Let the random variable X denote the *Number of trials until the first* H *appears.* Find the probability mass function of X.

Solution Now X can take on the values $\{1, 2, 3, \ldots\}$, so we have a non-uniform random variable with infinitely many possibilities. Since

$$\begin{aligned} f(1) &= P(X=1) = P(\mathsf{H}) = \frac{1}{2}\ , \\ f(2) &= P(X=2) = P(\mathsf{TH}) = \frac{1}{2}\cdot\frac{1}{2} = \left(\frac{1}{2}\right)^2\ , \\ f(3) &= P(X=3) = P(\mathsf{TTH}) = \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} = \left(\frac{1}{2}\right)^3\ , \qquad \text{etc.} \end{aligned}$$

the probability mass function of X is:

$$f(x) = P(X=x) = \left(\frac{1}{2}\right)^x, \quad x = 1, 2, \ldots\ .$$

In the previous Exercise, noting that we have independent trials here,

$$f(x) = P(X=x) = P(\underbrace{\mathsf{TT}\ldots\mathsf{T}}_{\mathsf{n-1}}\mathsf{H}) = P(\mathsf{T})^{x-1}\,P(\mathsf{H}) = \left(\frac{1}{2}\right)^{x-1}\frac{1}{2}\,.$$

More generally, let there be two possibilities, success (S) or failure (F), with $P(\mathsf{S}) = \theta$ and $P(\mathsf{F}) = 1-\theta$ so that:

$$P(X=x) = P(\underbrace{\mathsf{FF}\ldots\mathsf{F}}_{x-1}\mathsf{S}) = (1-\theta)^{x-1}\,\theta\,.$$

This is called a **geometric random variable** with "success probability" parameter θ. We can spot a geometric distribution because there will be *a sequence of independent trials with a constant probability of success. We are counting the number of trials until the first success appears.* Let us define this random variable formally next.

Definition 7.8 Geometric(θ) Random Variable. Given a parameter $\theta \in (0,1)$, the probability mass function for the Geometric(θ) random variable X is:

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

Example 7.9 Suppose we flip a coin 10 times and count the number of heads. Let's consider the probability of getting three heads, say. The probability that the first three flips are heads and the last seven flips are tails, *in order*, is

$$\underbrace{\frac{1}{2}\frac{1}{2}\frac{1}{2}}_{\text{3 successes}}\ \underbrace{\frac{1}{2}\frac{1}{2}\cdots\frac{1}{2}}_{\text{7 failures}}\,.$$

But there are

$$\binom{10}{3} = \frac{10!}{7!\,3!} = 120$$

ways of ordering three heads and seven tails, so the probability of getting three heads and seven tails *in any order*, is

$$P(3 \text{ heads}) = \binom{10}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^7 \approx 0.117$$

We can describe this sort of situation mathematically by considering a random variable X which counts the number of successes, as follows:

Definition 7.10 Binomial (n, θ) Random Variable. Suppose we have two parameters n and θ, and let the random variable X be the sum of n independent Bernoulli(θ) random variables, $X_1, X_2, \ldots, X_n$, that is:

$$X = \sum_{i=1}^{n} X_i, \ .$$

We call X the **Binomial (n, θ) random variable.** The probability mass function of X is:

$$f(x; n, \theta) = \begin{cases} \binom{n}{x}\theta^x(1-\theta)^{n-x} & \text{if } x \in \{0, 1, 2, 3, \ldots, n\} \ , \\ 0 & \text{otherwise} \end{cases}$$

Justification: The argument from Example 7.9 generalises as follows. Since the trials are independent and identical, the probability of x successes followed by $n - x$ failures, *in order*, is given by

$$\underbrace{\mathsf{SS}\ldots\mathsf{S}}_{x}\underbrace{\mathsf{FF}\ldots\mathsf{F}}_{n-x} = \theta^x(1-\theta)^{n-x}\,.$$

Since the n symbols $\mathsf{SS}\ldots\mathsf{S}\,\mathsf{FF}\ldots\mathsf{F}$ may be arranged in

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

ways, the probability of x successes and $n - x$ failures, *in any order*, is given by

$$\binom{n}{x}\theta^x(1-\theta)^{n-x}\,.$$

Exercise 7.11

Find the probability that seven of ten persons will recover from a tropical disease where the probability is identically 0.80 that any one of them will recover from the disease.

Solution We can assume independence here, so we have a binomial situation with $x = 7$, $n = 10$, and $\theta = 0.8$. Substituting these into the formula for the probability mass function for Binomial$(10, 0.8)$ random variable, we get:

$$\begin{aligned} f(7; 10, 0.8) &= \binom{10}{7} \times (0.8)^7 \times (1 - 0.8)^{10-7} \\ &= \frac{10!}{(10-7)!7!} \times (0.8)^7 \times (1 - 0.8)^{10-7} \\ &= 120 \times (0.8)^7 \times (1 - 0.8)^{10-7} \\ &\approx 0.20 \end{aligned}$$

Exercise 7.12

Compute the probability of obtaining *at least two* 6*'s* in rolling a fair die independently and identically four times.

Solution In any given toss let $\theta = P(\{6\}) = 1/6$, $1 - \theta = 5/6$, $n = 4$.

The event *at least two* 6*'s* occurs if we obtain two or three or four 6's. Hence the answer is:

$$\begin{aligned} P(\textit{at least two } 6\textit{'s}) &= f\left(2; 4, \frac{1}{6}\right) + f\left(3; 4, \frac{1}{6}\right) + f\left(4; 4, \frac{1}{6}\right) \\ &= \binom{4}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^{4-2} + \binom{4}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^{4-3} + \binom{4}{4}\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^{4-4} \\ &= \frac{1}{6^4}\left(6 \cdot 25 + 4 \cdot 5 + 1\right) \\ &\approx 0.132 \end{aligned}$$

We now consider the last of our discrete random variables, the **Poisson** case. A Poisson random variable counts the number of times an event occurs. We might, for example, ask:

- How many customers visit Cafe 101 each day?
- How many sixes are scored in a cricket season?
- How many bombs hit a city block in south London during World War II?

A Poisson experiment has the following characteristics:

- The average rate of an event occurring is known. This rate is constant.
- The probability that an event will occur during a short continuum is proportional to the size of the continuum.

- Events occur independently.

The number of events occurring in a Poisson experiment is referred to as a **Poisson random variable**.

Definition 7.13 Poisson(λ) Random Variable. Given a parameter $\lambda > 0$, the probability mass function of the Poisson(λ) random variable X is:

$$f(x;\lambda) = \frac{\lambda^x}{x!}\exp(-\lambda) \qquad \text{where} \quad x = 0, 1, \ldots \tag{16}$$

We interpret X as the number of times an event occurs during a specified continuum given that the average value in the continuum is λ.

Exercise 7.14

If on the average, 2 cars enter a certain parking lot per minute, what is the probability that during any given minute three cars or fewer will enter the lot?
Think: Why are the assumptions for a Poisson random variable likely to be correct here?
Note: Use calculators, or Excel or Maple, etc. In an exam you will be given suitable Poisson tables.

Solution Let the random variable X denote the number of cars arriving per minute. Note that the continuum is 1 minute here. Then X can be considered to have a Poisson distribution with $\lambda = 2$ because 2 cars enter on average.

The probability that three cars or fewer enter the lot is:

$$\begin{aligned} P(X \leq 3) &= f(0;2) + f(1;2) + f(2;2) + f(3;2) \\ &= e^{-2}\left(\frac{2^0}{0!} + \frac{2^1}{1!} + \frac{2^2}{2!} + \frac{2^3}{3!}\right) \\ &= 0.857 \quad (3 \text{ sig. fig.}) \end{aligned}$$

Exercise 7.15

The proprietor of a service station finds that, on average, 8 cars arrive *per hour* on Saturdays. What is the probability that during a randomly chosen 15 *minute period* on a Saturday:

(a) No cars arrive?

(b) At least three cars arrive?

Solution Let the random variable X denote the number of cars arriving in a 15 minute interval. The continuum is 15 minutes here so we need the average number of cars that arrive in a 15 minute period, or $\frac{1}{4}$ of an hour. We know that 8 cars arrive per hour, so X has a Poisson distribution with

$$\lambda = \frac{8}{4} = 2\,.$$

(a)

$$P(X=0) = f(0;2) = \frac{e^{-2}2^0}{0!} = 0.135 \quad \text{(3 sig. fig)}$$

(b)

$$\begin{aligned} P(X \geq 3) &= 1 - P(X<3) \\ &= 1 - P(X=0) - P(X=1) - P(X=2) \\ &= 1 - f(0;2) - f(1;2) - f(2;2) \\ &= 1 - 0.1353 - 0.2707 - 0.2707 \\ &= 0.323 \quad \text{(3 sig. fig.)} \end{aligned}$$

Remark 7.16 In the binomial case where θ is small and n is large, it can be shown that the binomial distribution with parameters n and θ is closely approximated by the Poisson distribution having $\lambda = n\theta$. The smaller the value of θ, the better the approximation.

Example 7.17 About 0.01% of babies are stillborn in a certain hospital. We find the probability that of the next 5000 babies born, there will be no more than 1 stillborn baby.
Let the random variable X denote the number of stillborn babies. Then X has a binomial distribution with parameters $n = 5000$ and $\theta = 0.0001$. Since θ is so small and n is large, this binomial distribution may be approximated by a Poisson distribution with parameter

$$\lambda = n\,\theta = 5000 \times 0.0001 = 0.5\,.$$

Hence

$$P(X \leq 1) = P(X=0) + P(X=1) = f(0;0.5) + f(1;0.5) = 0.910 \quad \text{(3 sig. fig.)}$$

THINKING POISSON

The Poisson distribution has been described as a limiting version of the Binomial. In particular, Exercise 7.14 thinks of a Poisson distribution as a model for the number of events (cars) that occur in a period of time (1 minute) when in each little chunk of time one car arrives with constant probability, independently of the other time intervals. This leads to the general view of the Poisson distribution as a good model when:

You count the number of events in a continuum when the events occur at constant rate, one at a time and are independent of one another.

DISCRETE RANDOM VARIABLE SUMMARY

Probability mass function

$$f(x) = P(X = x_i)$$

Distribution function

$$F(x) = \sum_{x_i \leq x} f(x_i)$$

Random Variable	**Possible Values**	**Probabilities**	**Modelled situations**
Discrete uniform	$\{x_1, x_2, \ldots, x_k\}$	$P(X = x_i) = \frac{1}{k}$	Situations with k equally likely values. Parameter: k.
Bernoulli(θ)	$\{0, 1\}$	$P(X = 0) = 1 - \theta$ $P(X = 1) = \theta$	Situations with only 2 outcomes, coded 1 for success and 0 for failure. Parameter: $\theta = P(\text{success}) \in (0, 1)$.
Geometric(θ)	$\{1, 2, 3, \ldots\}$	$P(X = x)$ $= (1 - \theta)^{x-1}\theta$	Situations where you count the number of trials until the first success in a sequence of independent trails with a constant probability of success. Parameter: $\theta = P(\text{success}) \in (0, 1)$.
Binomial(n, θ)	$\{0, 1, 2, \ldots, n\}$	$P(X = x)$ $= \binom{n}{x}\theta^x(1-\theta)^{n-x}$	Situations where you count the number of success in n trials where each trial is independent and there is a constant probability of success. Parameters: $n \in \{1, 2, \ldots\}$; $\theta = P(\text{success}) \in (0, 1)$.
Poisson(λ)	$\{0, 1, 2, \ldots\}$	$P(X = x)$ $= \frac{\lambda^x e^{-\lambda}}{x!}$	Situations where you count the number of events in a continuum where the events occur one at a time and are independent of one another. Parameter: λ= rate $\in (0, \infty)$.

8 Continuous random variables and distributions

If X is a measurement of a continuous quantity, such as,

- the maximum diameter in millimeters of a venus shell I picked up at New Brighton beach,
- the distance you drove to work today in kilometers,

- the volume of rain that fell on the roof of this building over the past 365 days in litres,
- etc.,

then X is a continuous random variable. Continuous random variables are based on measurements in a continuous scale of a given precision as opposed to discrete random variables that are based on counting.

Example 8.1 Suppose that X is the time, in minutes, before the next student leaves the lecture room. This is an example of a continuous random variable that takes one of (uncountably) infinitely many values. When a student leaves, X will take on the value x and this x could be 2.1 minutes, or 2.1000000001 minutes, or 2.9999999 minutes, etc. Finding $P(X = 2)$, for example, doesn't make sense because how can it ever be *exactly* $2.00000\ldots$ minutes? It is more sensible to consider probabilities like $P(X > x)$ or $P(X < x)$ rather than the discrete approach of trying to compute $P(X = x)$.

The characteristics of continuous random variables are:

- The outcomes are measured, not counted.
- Geometrically, *the probability of an outcome is equal to an area* under a mathematical curve.
- Each individual value has zero probability of occurring. Instead we find the probability that the value is between two endpoints.

We will consider continuous random variables, X, having the property that the distribution function $F(x)$ is of the form:

$$F(x) \;=\; P(X \leq x) \;=\; \int_{-\infty}^{x} f(v)\, dv \tag{17}$$

where f is a function. We assume that f is continuous, perhaps except for finitely many x-values. (We write v because x is needed as the upper limit of the integral.)

The integrand f, called the **probability density function** (PDF) of the distribution, is non-negative.

Differentiation gives the relation of f to F as

$$f(x) \;=\; F'(x) \tag{18}$$

for every x at which $f(x)$ is continuous.

From this we obtain the following two very important formulae.

The probability corresponding to an interval (this is an *area*):

$$P(a < X \leq b) \;=\; F(b) - F(a) \;=\; \int_{a}^{b} f(v)dv \;\; . \tag{19}$$

Since $P(\Omega) = 1$, we also have:

$$\int_{-\infty}^{\infty} f(v)dv \;=\; 1 \;.$$

Remark 8.2 Continuous random variables are simpler than discrete ones with respect to intervals. Indeed, in the continuous case the four probabilities $P(a < X \leq b)$, $P(a < X < b)$, $P(a \leq X < b)$, and $P(a \leq X \leq b)$ with any fixed a and b ($> a$) are all the same.

The next exercises illustrate notation and typical applications of our present formulae.

Exercise 8.3
Consider the continuous random variable, X, whose probability density function is:

$$f(x) \;=\; \begin{cases} 3x^2 & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

(a) Find the distribution function, $F(x)$.

(b) Find $P(\frac{1}{3} \leq X \leq \frac{2}{3})$.

Solution

(a) First note that if $x \leq 0$, then

$$F(x) \;=\; \int_{-\infty}^{x} 0dv \;=\; 0\,.$$

If $0 < x < 1$, then

$$\begin{aligned} F(x) &= \int_{-\infty}^{0} 0dv \;+\; \int_{0}^{x} 3v^2 dv \\ &= 0 \;+\; \left[v^3\right]_0^x \\ &= x^3 \end{aligned}$$

If $x \geq 1$, then

$$\begin{aligned} F(x) &= \int_{-\infty}^{0} 0dv \;+\; \int_{0}^{1} 3v^2 dv + \int_{1}^{x} 0dv \\ &= 0 \;+\; \left[v^3\right]_0^1 \;+\; 0 \\ &= 1 \end{aligned}$$

Hence

$$F(x) \;=\; \begin{cases} 0 & x \leq 0 \\ x^3 & 0 < x < 1 \\ 1 & x \geq 1 \end{cases}$$

(b)

$$\begin{aligned} P\left(\frac{1}{3} \leq X \leq \frac{2}{3}\right) &= F\left(\frac{2}{3}\right) - F\left(\frac{1}{3}\right) \\ &= \left(\frac{2}{3}\right)^3 - \left(\frac{1}{3}\right)^3 \\ &= \frac{7}{27} \end{aligned}$$

Example 8.4 Consider the continuous random variable, X, whose distribution function is:

$$F(x) = \begin{cases} 0 & x \leq 0 \\ \sin(x) & 0 < x < \frac{\pi}{2} \\ 1 & x \geq \frac{\pi}{2} \end{cases}.$$

(a) Find the probability density function, $f(x)$.

(b) Find $P\left(X > \frac{\pi}{4}\right)$

Solution

(a) The probability density function, $f(x)$ is given by

$$f(x) = F'(x) = \begin{cases} 0 & x < 0 \\ \cos x & 0 < x < \frac{\pi}{2} \\ 0 & x \geq \frac{\pi}{2} \end{cases}$$

(b)

$$P\left(X > \frac{\pi}{4}\right) = 1 - P\left(X \leq \frac{\pi}{4}\right) = 1 - F\left(\frac{\pi}{4}\right) = 1 - \sin\left(\frac{\pi}{4}\right) = 0.293 \text{ (3 sig. fig.)}$$

Note: $f(x)$ is not defined at $x = 0$ as $F(x)$ is not differentiable at $x = 0$. There is a "kink" in the distribution function at $x = 0$ causing this problem. It is standard to define $f(0) = 0$ in such situations, as $f(x) = 0$ for $x < 0$. This choice is arbitrary but it simplifies things and makes no difference to the important things in life like calculating probabilities!

Exercise 8.5
Let X have density function $f(x) = e^{-x}$, if $x \geq 0$, and zero otherwise.

(a) Find the distribution function.

(b) Find the probabilities, $P(\frac{1}{4} \leq X \leq 2)$ and $P\left(-\frac{1}{2} \leq X \leq \frac{1}{2}\right)$.

(c) Find x such that $P(X \leq x) = 0.95$.

Solution

(a)

$$F(x) \;=\; \int_0^x e^{-v}dv \;=\; -e^{-v}\Big]_0^x \;=\; -e^{-x}+1 \;=\; 1-e^{-x} \qquad \text{if } x \geq 0$$

Therefore,

$$F(x) \;=\; \begin{cases} 1-e^{-x} & \text{if } x \geq 0 \;, \\ 0 & \text{otherwise} \;. \end{cases}$$

(b)

$$P\left(\frac{1}{4} \leq X \leq 2\right) \;=\; F(2) - F\left(\frac{1}{4}\right) \;=\; 0.634 \;\; (3 \text{ sig. fig.})$$

$$P\left(-\frac{1}{2} \leq X \leq \frac{1}{2}\right) \;=\; F\left(\frac{1}{2}\right) - F\left(-\frac{1}{2}\right) \;=\; 0.394 \;\; (3 \text{ sig. fig.})$$

(c)

$$P(X \leq x) \;=\; F(x) \;=\; 1-e^{-x} \;=\; 0.95$$

Therefore,

$$x \;=\; -\log(1-0.95) \;=\; 3.00 \;\; (3 \text{ sig. fig.}) \;.$$

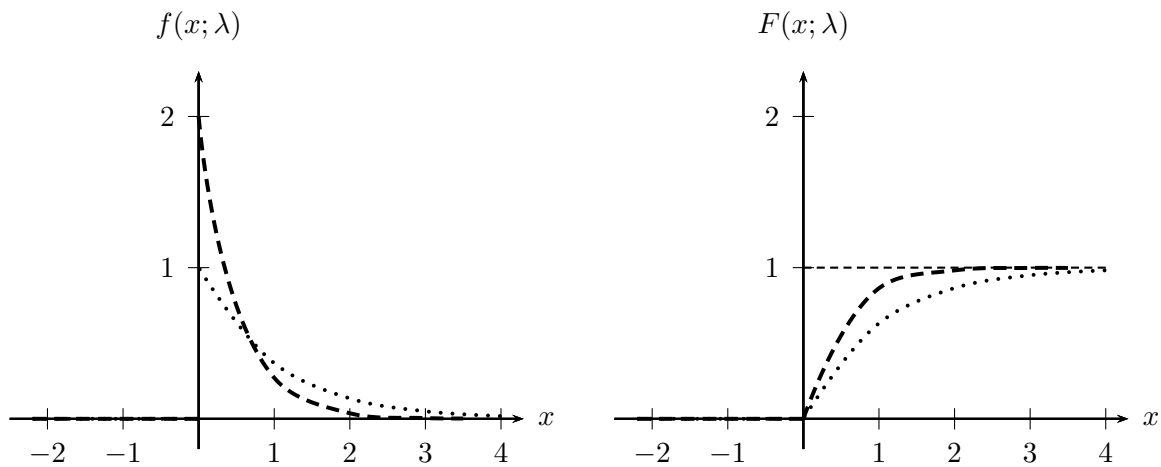
The previous example is a special case of the following parametric family of random variables.

Definition 8.6 Exponential(λ) Random Variable. Given a rate parameter $\lambda > 0$, the Exponential(λ) random variable X has probability density function given by:

$$f(x;\lambda) \;=\; \begin{cases} \lambda \exp(-\lambda x) & x > 0 \;, \\ 0 & \text{otherwise} \;, \end{cases}$$

and distribution function given by:

$$F(x;\lambda) \;=\; 1-\exp(-\lambda x) \;.$$



$f(x;\lambda)$ and $F(x;\lambda)$ of an exponential random variable where $\lambda = 1$ (dotted) and $\lambda = 2$ (dashed).

Exercise 8.7

At a certain location on a dark desert highway, the time in minutes between arrival of cars that exceed the speed limit is an Exponential($\lambda = 1/60$) random variable. If you just saw a car that exceeded the speed limit then what is the probability of waiting less than 5 minutes before seeing another car that will exceed the speed limit?

Solution The waiting time in minutes is simply given by the Exponential($\lambda = 1/60$) random variable. Thus, the desired probability is

$$P(0 \le X < 5) \;=\; \int_0^5 \frac{1}{60} e^{-\frac{1}{60}x} dx \;=\; -e^{-\frac{1}{60}x}\Big]_0^5 \;=\; -e^{-\frac{1}{12}} + 1 \;\approx\; 0.07996.$$

In exam you can stop at the expression $-e^{-\frac{1}{12}} + 1$ for full credit. You may need a calculator for the last step (with answer 0.07996).
Note: We could use the distribution function directly:

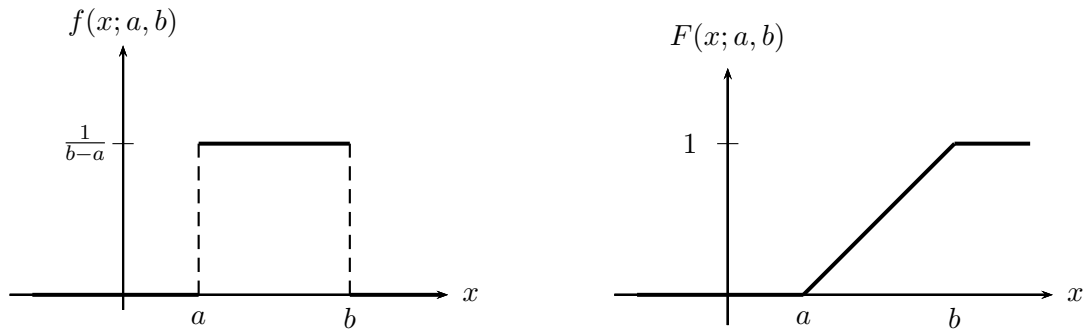
$$P(0 \le X < 5) \;=\; F\left(5; \frac{1}{60}\right) - F\left(0; \frac{1}{60}\right) \;=\; F\left(5; \frac{1}{60}\right) \;=\; 1 - e^{-\frac{1}{60}5} \;=\; 1 - -e^{-\frac{1}{12}} \;\approx\; 0.07996$$

Definition 8.8 Uniform(a, b) Random Variable. Given two real parameters a, b with $a < b$, the Uniform(a, b) random variable X has the following probability density function:

$$f(x; a, b) \;=\; \begin{cases} \dfrac{1}{b-a} & \text{if} \qquad a \le x \le b \;, \\ 0 & \text{otherwise} \end{cases}$$

X is also said to be **uniformly distributed** on the interval $[a, b]$. The distribution function of X is

$$F(x; a, b) \;=\; \begin{cases} 0 & x < a \;, \\ \dfrac{x-a}{b-a} & a \le x < b \;, \\ 1 & x \ge b \;. \end{cases}$$



$f(x; a, b)$ and $F(x; a, b)$ of Uniform(a, b) random variable X.

Exercise 8.9

Consider a random variable with a probability density function

$$f(x) \;=\; \begin{cases} k & \text{if} \quad 2 \le x \le 6 \;, \\ 0 & \text{otherwise} \end{cases}$$

(a) Find the value of k.

(b) Sketch the graphs of $f(x)$ and $F(x)$.

Solution

(a) Since $f(x)$ is a density function which integrates to one,

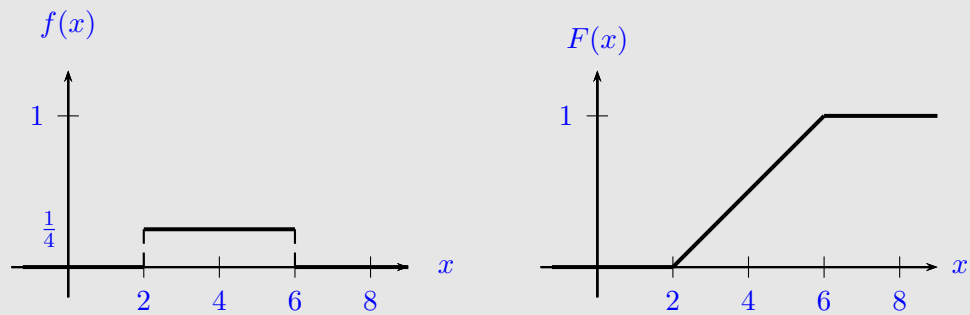
$$\begin{aligned} \int_2^6 f(x)\,dx \;&=\; \int_2^6 k\,dx \\ 1 \;&=\; kx]_2^6 \\ 1 \;&=\; 6k \;-\; 2k \\ 1 \;&=\; 4k \\ k \;&=\; \frac{1}{4} \end{aligned}$$

as expected!

(b) Now

$$F(x) = \begin{cases} 0 & x < 2 \\ \frac{1}{4}(x-2) & 2 \le x < 6 \\ 1 & x \ge 6 \ . \end{cases}$$

so the graphs are:



Graphs of $f(x)$ and $F(x)$.

The standard normal distribution is the most important continuous probability distribution. It was first described by De Moivre in 1733 and subsequently by the German mathematician C. F. Gauss (1777 - 1885). Many random variables have a normal distribution, or they are approximately normal, or can be transformed into normal random variables in a relatively simple fashion. Furthermore, the normal distribution is a useful approximation of more complicated distributions, and it also occurs in the proofs of various statistical tests.

Definition 8.10 A continuous random variable Z is called **standard normal** or **standard Gaussian** if its probability density function is

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \ . \tag{20}$$

An exercise in calculus yields the first two derivatives of φ as follows:

$$\frac{d\varphi}{dz} = -\frac{1}{\sqrt{2\pi}} z \exp\left(-\frac{z^2}{2}\right) = -z\varphi(z), \quad \frac{d^2\varphi}{dz^2} = \frac{1}{\sqrt{2\pi}}(z^2-1)\exp\left(-\frac{z^2}{2}\right) = (z^2-1)\varphi(z) \ .$$

Thus, φ has a global maximum at 0, it is concave down if $z \in (-1,1)$ and concave up if $z \in (-\infty,-1) \cup (1,\infty)$. This shows that the graph of φ is shaped like a smooth symmetric bell centred at the origin over the real line.

Exercise 8.11

From the above exercise in calculus let us draw the graph of φ by hand now!

Solution do it step by step: z^2, $-z^2$, $-z^2/2$, $\exp(-z^2/2)$, $\varphi(z) = \frac{1}{\sqrt{2\pi}}\exp(-z^2/2)$ live!

The distribution function of Z is given by

$$\Phi(z) \;=\; \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-v^2/2}\,dv \;. \tag{21}$$

Remark 8.12 This integral cannot be evaluated exactly by standard methods of calculus, but its values can be obtained numerically and tabulated. Values of $\Phi(z)$ are tabulated in the "Standard Normal Distribution Function Table" in ***S*** 20.

Exercise 8.13

Note that the curve of $\Phi(z)$ is S-shaped, increasing in a strictly monotone way from 0 at $-\infty$ to 1 at ∞, and intersects the vertical axis at 1/2. Draw this by hand too.

Solution do it!

Example 8.14 Find the probabilities, using normal tables, that a random variable having the standard normal distribution will take on a value:

(a) less that 1.72

(b) less than -0.88

(c) between 1.30 and 1.75

(d) between -0.25 and 0.45

(a)

$$P(Z < 1.72) \;=\; \Phi(1.72) \;=\; 0.9573$$

(b) First note that $P(Z < 0.88) \;=\; 0.8106$, so that

$$\begin{aligned} P(Z < -0.88) \;&=\; P(Z > 0.88) \\ &=\; 1 - P(Z < 0.88) \\ &=\; 1 - \Phi(0.88) \\ &=\; 1 - 0.8106 \;=\; 0.1894 \end{aligned}$$

(c) $P(1.30 < Z < 1.75) \;=\; \Phi(1.75) - \Phi(1.30) \;=\; 0.9599 - 0.9032 \;=\; 0.0567$

(d)

$$\begin{aligned} P(-0.25 < Z < 0.45) &= P(Z < 0.45) - P(Z < -0.25) \\ &= P(Z < 0.45) - (1 - P(Z < 0.25)) \\ &= \Phi(0.45) - (1 - \Phi(0.25)) \\ &= (0.6736) - (1 - 0.5987) \\ &= 0.2723 \end{aligned}$$

CONTINUOUS RANDOM VARIABLES: NOTATION

$f(x)$: Probability density function (PDF)

- $f(x) \geq 0$
- Areas underneath $f(x)$ measure probabilities.

$F(x)$: Distribution function (DF)

- $0 \leq F(x) \leq 1$
- $F(x) = P(X \leq x)$ is a probability
- $F'(x) = f(x)$ for every x where $f(x)$ is continuous
- $F(x) = \int_{-\infty}^{x} f(v)dv$
- $P(a < X \leq b) = F(b) - F(a) = \int_{a}^{b} f(v)dv$

9 Transformations of random variables

Suppose we know the distribution of a random variable X. How do we find the distribution of a transformation of X, say $g(X)$? Before we answer this question let us ask a motivational question. Why are we interested in functions of random variables?

Example 9.1 Consider a simple financial example where an individual sells X items per day, the profit per item is \$5 and the overhead costs are \$500 per day. The original random variable is X, but the random variable Y which gives the daily profit is of more interest, where

$$Y = 5X - 500 \ .$$

Example 9.2 In a cell-phone system a mobile signal may have a signal-to-noise-ratio of X, but engineers prefer to express such ratios in decibels, i.e.,

$$Y = 10 \log_{10}(X) \ .$$

Hence in a great many situations we are more interested in functions of random variables. Let us return to our original question of determining the distribution of a transformation or function of X. First note that this transformation of X is itself another random variable, say $Y = g(X)$, where g is a function from a subset $\mathbb{X}$ of $\mathbb{R}$ to a subset $\mathbb{Y}$ of $\mathbb{R}$, i.e., $g : \mathbb{X} \to \mathbb{Y}$, $\mathbb{X} \subset \mathbb{R}$ and $\mathbb{Y} \subset \mathbb{R}$. The **inverse image** of a set A is the set of all real numbers in $\mathbb{X}$ whose image is in A, i.e.,

$$g^{[-1]}(A) = \{x \in \mathbb{X} : g(x) \in A\} \ .$$

In other words,

$$x \in g^{[-1]}(A) \text{ if and only if } g(x) \in A \ .$$

For example,

- if $g(x) = 2x$ then $g^{[-1]}([4,6]) = [2,3]$
- if $g(x) = 2x + 1$ then $g^{[-1]}([5,7]) = [2,3]$
- if $g(x) = x^3$ then $g^{[-1]}([1,8]) = [1,2]$
- if $g(x) = x^2$ then $g^{[-1]}([1,4]) = [-2,-1] \cup [1,2]$
- if $g(x) = \sin(x)$ then $g^{[-1]}([-1,1]) = \mathbb{R}$
- if ...

For the singleton set $A = \{y\}$, we write $g^{[-1]}(y)$ instead of $g^{[-1]}(\{y\})$. For example,

- if $g(x) = 2x$ then $g^{[-1]}(4) = \{2\}$
- if $g(x) = 2x + 1$ then $g^{[-1]}(7) = \{3\}$
- if $g(x) = x^3$ then $g^{[-1]}(8) = \{2\}$
- if $g(x) = x^2$ then $g^{[-1]}(4) = \{-2, 2\}$
- if $g(x) = \sin(x)$ then $g^{[-1]}(0) = \{k\pi : k \in \mathbb{Z}\} = \{\ldots, -3\pi, -2\pi, -\pi, 0, \pi, 2\pi, 3\pi, \ldots\}$
- if ...

If $g : \mathbb{X} \to \mathbb{Y}$ is one-to-one (injective) and onto (surjective), then the inverse image of a singleton set is itself a singleton set. Thus, the inverse image of such a function g becomes itself a function and is called the **inverse function**. One can find the inverse function, if it exists by the following steps:

Step 1; write $y = g(x)$

Step 2; solve for x in terms of y

Step 3; set $g^{-1}(y)$ to be this solution

We write g^{-1} whenever the inverse image $g^{[-1]}$ exists as an inverse function of g. Thus, the inverse function g^{-1} is a specific type of inverse image $g^{[-1]}$. For example,

- if $g(x) = 2x$ then $g : \mathbb{R} \to \mathbb{R}$ is injective and surjective and therefore its inverse function is:
  Step 1; $y = 2x$, Step 2; $x = \frac{y}{2}$, Step 3; $g^{-1}(y) = \frac{y}{2}$
- if $g(x) = 2x+1$ then $g : \mathbb{R} \to \mathbb{R}$ is injective and surjective and therefore its inverse function is:
  Step 1; $y = 2x + 1$, Step 2; $x = \frac{y-1}{2}$, Step 3; $g^{-1}(y) = \frac{y-1}{2}$
- if $g(x) = x^3$ then $g : \mathbb{R} \to \mathbb{R}$ is injective and surjective and therefore its inverse function is:
  Step 1; $y = x^3$, Step 2; $x = y^{\frac{1}{3}}$, Step 3; $g^{-1}(y) = y^{\frac{1}{3}}$

However, you need to be careful by limiting the domain to obtain the inverse function for the following examples:

- if $g(x) = x^2$ and domain of g is $[0, +\infty)$ then its inverse function is $g^{-1}(y) = \sqrt{y}$, i.e., if $g(x) = x^2 : [0, +\infty) \to [0, +\infty)$ then the inverse image $g^{[-1]}(y)$ for $y \in [0, +\infty)$ is given by the inverse function $g^{-1}(y) = \sqrt{y} : [0, +\infty) \to [0, +\infty)$.
- if $g(x) = x^2$ and domain of g is $(-\infty, 0]$ then its inverse function is $g^{-1}(y) = -\sqrt{y}$, i.e., if $g(x) = x^2 : (-\infty, 0] \to [0, +\infty)$ then the inverse image $g^{[-1]}(y)$ for $y \in [0, +\infty)$ is given by the inverse function $g^{-1}(y) = -\sqrt{y} : [0, +\infty) \to (-\infty, 0]$.
- if $g(x) = \sin(x)$ and domain of g is $[0, \frac{\pi}{2}]$ then its inverse function $g^{-1}(y) = \arcsin(y)$, i.e., if $g(x) = \sin(x) : [0, \frac{\pi}{2}] \to [0, 1]$ then the inverse image $g^{[-1]}(y)$ for $y \in [0, 1]$ is given by the inverse function $g^{-1}(y) = \arcsin(y) : [0, 1] \to [0, \frac{\pi}{2}]$.
- if $g(x) = \sin(x)$ and domain of g is $[-\frac{\pi}{2}, \frac{\pi}{2}]$ then its inverse function $g^{-1}(y) = \arcsin(y)$, i.e., if $g(x) = \sin(x) : [-\frac{\pi}{2}, \frac{\pi}{2}] \to [-1, 1]$ then the inverse image $g^{[-1]}(y)$ for $y \in [-1, 1]$ is given by the inverse function $g^{-1}(y) = \arcsin(y) : [-1, 1] \to [-\frac{\pi}{2}, \frac{\pi}{2}]$.
- if ...

Now, let us return to our question of determining the distribution of the transformation $g(X)$. To answer this question we must first observe that the inverse image $g^{[-1]}$ satisfies the following properties:

- $g^{[-1]}(\mathbb{Y}) = \mathbb{X}$
- For any set A, $g^{[-1]}(A^c) = \left(g^{[-1]}(A)\right)^c$
- For any collection of sets $\{A_1, A_2, \ldots\}$,

$$g^{[-1]}\left(A_1 \cup A_2 \cup \cdots\right) = g^{[-1]}(A_1) \cup g^{[-1]}(A_2) \cup \cdots \quad .$$

Consequentially,

$$\boxed{P\left(g(X) \in A\right) = P\left(X \in g^{[-1]}(A)\right)} \tag{22}$$

satisfies the axioms of probability and gives the desired probability of the event A from the transformation $Y = g(X)$ in terms of the probability of the event given by the inverse image of A underpinned by the random variable X. It is crucial to understand this from the sample

space Ω of the underlying experiment in the sense that Equation (22) is just short-hand for its actual meaning:

$$P\left(\{\omega \in \Omega : g(X(\omega)) \in A\}\right) = P\left(\left\{\omega \in \Omega : X(\omega) \in g^{[-1]}(A)\right\}\right) \ .$$

Because we have more than one random variable to consider, namely, X and its transformation $Y = g(X)$ we will subscript the probability density or mass function and the distribution function by the random varaible itself. For example we denote the distribution function of X by $F_X(x)$ and that of Y by $F_Y(y)$.

9.1 Transformation of discrete random variables

For a discrete random variable X with probability mass function f_X we can obtain the probability mass function f_Y of $Y = g(X)$ using Equation (22) as follows:

$$\begin{aligned} f_Y(y) &= P(Y = y) = P(Y \in \{y\}) \\ &= P\left(g(X) \in \{y\}\right) = P\left(X \in g^{[-1]}(\{y\})\right) \\ &= P\left(X \in g^{[-1]}(y)\right) = \sum_{x \in g^{[-1]}(y)} f_X(x) = \sum_{x \in \{x : g(x) = y\}} f_X(x) \ . \end{aligned}$$

This gives the formula:

$$\boxed{f_Y(y) = P(Y = y) = \sum_{x \in g^{[-1]}(y)} f_X(x) = \sum_{x \in \{x : g(x) = y\}} f_X(x) \ .} \tag{23}$$

Example 9.3 Let X be the discrete random variable with probability mass function f_X as tabulated below:

x	-1	0	1
$f_X(x) = P(X = x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

If $Y = 2X$ then the transformation $g(X) = 2X$ has inverse image $g^{[-1]}(y) = \{y/2\}$. Then, by Equation (23) the probability mass function of Y is expressed in terms of the known probabilities of X as:

$$f_Y(y) = P(Y = y) = \sum_{x \in g^{[-1]}(y)} f_X(x) = \sum_{x \in \{y/2\}} f_X(x) = f_X(y/2) \ ,$$

and tabulated below:

y	-2	0	2
$f_Y(y)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Example 9.4 If X is the random variable in the previous Example then what is the probability mass function of $Y = 2X + 1$? Once again,

$$f_Y(y) = P(Y = y) = \sum_{x \in g^{[-1]}(y)} f_X(x) = \sum_{x \in \{(y-1)/2\}} f_X(x) = f_X((y-1)/2) \ ,$$

and tabulated below:

y	-1	1	3
$f_Y(y)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

In fact, obtaining the probability of a one-to-one transformation of a discrete random variable as in Examples 9.3 and 9.4 is merely a matter of looking up the probability at the image of the inverse function. This is because there is only one term in the sum that appears in Equation (23). When the transformation is not one-to-one the number of terms in the sum can be more than one as shown in the next Example.

Example 9.5 Reconsider the random variable X of the last two Examples and let $Y = X^2$. Recall that $g(x) = x^2$ does not have an inverse function unless the domain is restricted to the positive or the negative parts of the real line. Since our random variable X takes values on both sides of the real line, namely $\{-1, 0, 1\}$, let us note that the transformation $g(X) = X^2$ is no longer a one-to-one function. Then, by Equation (23) the probability mass function of Y is expressed in terms of the known probabilities of X as:

$$f_Y(y) = P(Y = y) = \sum_{x \in g^{[-1]}(y)} f_X(x) = \sum_{\{x : g(x) = y\}} f_X(x) = \sum_{\{x : x^2 = y\}} f_X(x) \ ,$$

computed for each $y \in \{0, 1\}$ as follows:

$$\begin{aligned} f_Y(0) &= \sum_{\{x : x^2 = 0\}} f_X(x) = f_X(0) = \frac{1}{2} \ , \\ f_Y(1) &= \sum_{\{x : x^2 = 1\}} f_X(x) = f_X(-1) + f_X(1) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \ , \end{aligned}$$

and finally tabulated below:

y	0	1
$f_Y(y)$	$\frac{1}{2}$	$\frac{1}{2}$

.

9.2 Transformation of continuous random variables

Suppose we know F_X and/or f_X of a continuous random variable X. Let $Y = g(X)$ be a transformation of X. Our objective is to obtain F_Y and/or f_Y of Y from F_X and/or f_X.

9.2.1 One-to-one transformations

The easiest case for transformations of continuous random variables is when g is **one-to-one and monotone**.

- First, let us consider the case when g is **monotone and increasing** on the range of the random variable X. In this case g^{-1} is also an increasing function and we can obtain the distribution function of $Y = g(X)$ in terms of the distribution function of X as

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P\left(X \leq g^{-1}(y)\right) = F_X(g^{-1}(y)) \ .$$

  Now, let us use a form of chainrule to compute the density of Y as follows:

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X\left(g^{-1}(y)\right) = f_X\left(g^{-1}(y)\right) \frac{d}{dy}\left(g^{-1}(y)\right) \ .$$

- Second, let us consider the case when g is **monotone and decreasing** on the range of the random variable X. In this case g^{-1} is also a decreasing function and we can obtain the distribution function of $Y = g(X)$ in terms of the distribution function of X as

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P\left(X \geq g^{-1}(y)\right) = 1 - F_X(g^{-1}(y)) \ ,$$

  and the density of Y as

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy}\left(1 - F_X\left(g^{-1}(y)\right)\right) = -f_X\left(g^{-1}(y)\right) \frac{d}{dy}\left(g^{-1}(y)\right) \ .$$

  For a monotonic and decreasing g, its inverse function g^{-1} is also decreasing and consequently the density f_Y is indeed positive because $\frac{d}{dy}\left(g^{-1}(y)\right)$ is negative.

We can combine the above two cases and obtain the following **change of variable formula** for the probability density of $Y = g(X)$ when g is one-to-one and monotone on the range of X.

$$\boxed{f_Y(y) = f_X\left(g^{-1}(y)\right) \left| \frac{d}{dy} g^{-1}(y) \right| \ .} \tag{24}$$

The steps involved in finding the density of $Y = g(X)$ for a one-to-one and monotone g are:

1. Write $y = g(x)$ for x in range of x and check that $g(x)$ is monotone over the required range to apply the change of variable formula.
2. Write $x = g^{-1}(y)$ for y in range of y.
3. Obtain $\left|\frac{d}{dy} g^{-1}(y)\right|$ for y in range of y.
4. Finally, from Equation (24) get $f_Y(y) = f_X\left(g^{-1}(y)\right) \left|\frac{d}{dy} g^{-1}(y)\right|$ for y in range of y.

Let us use these four steps to obtain the density of monotone transformations of continuous random variables.

Example 9.6 Let X be Uniform$(0,1)$ random variable and let $Y = g(X) = 1 - X$. We are interested in the density of the tranformed random variable Y. Let us follow the four steps and use the change of variable formula to obtain f_Y from f_X and g.

1. $y = g(x) = 1 - x$ is a monotone decreasing function over $0 \leq x \leq 1$, the range of X. So, we can apply the change of variable formula.

2. $x = g^{-1}(y) = 1 - y$ is a monotone decreasing function over $1 - 0 \geq 1 - x \geq 1 - 1$, i.e., $0 \leq y \leq 1$.

3. For $0 \leq y \leq 1$,
$$\left|\frac{d}{dy} g^{-1}(y)\right| = \left|\frac{d}{dy}(1-y)\right| = |-1| = 1 \ .$$

4. we can use Equation (24) to find the density of Y as follows:
$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{d}{dy} g^{-1}(y)\right| = f_X(1-y)\ 1 = 1 \ ,$$
for $0 \leq y \leq 1$

Thus, we have shown that if X is a Uniform$(0,1)$ random variable then $Y = 1 - X$ is also a Uniform$(0,1)$ random variable.

Example 9.7 Let X be a Uniform$(0,1)$ random variable and let $Y = g(X) = -\log(X)$. We are interested in the density of the tranformed random variable Y. Once again, since g is a one-to-one monotone function let us follow the four steps and use the change of variable formula to obtain f_Y from f_X and g.

1. $y = g(x) = -\log(x)$ is a monotone decreasing function over $0 < x < 1$, the range of X. So, we can apply the change of variable formula.

2. $x = g^{-1}(y) = \exp(-y)$ is a monotone decreasing function over $0 < y < \infty$.

3. For $0 < y < \infty$,
$$\left|\frac{d}{dy} g^{-1}(y)\right| = \left|\frac{d}{dy}(\exp(-y))\right| = |-\exp(-y)| = \exp(-y) \ .$$

4. We can use Equation (24) to find the density of Y as follows:
$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{d}{dy} g^{-1}(y)\right| = f_X(\exp(-y))\ \exp(-y) = 1\ \exp(-y) = \exp(-y) \ .$$
Note that $0 < \exp(-y) < 1$ for $0 < y < \infty$.

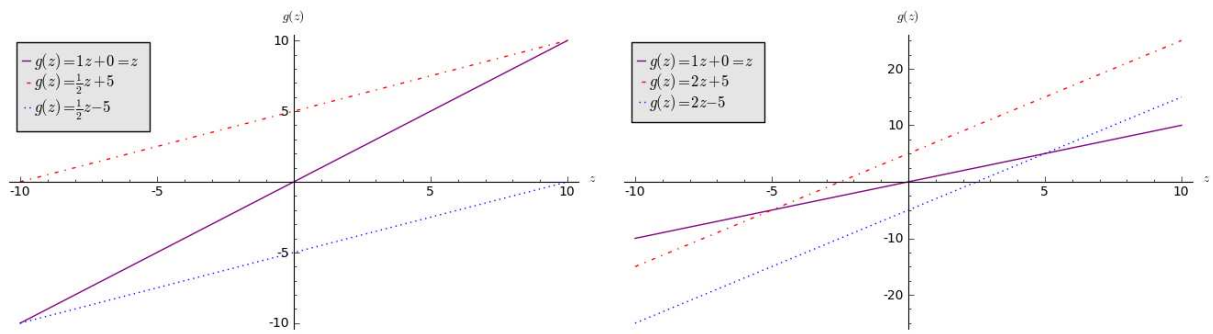
Thus, we have shown that if X is a Uniform$(0,1)$ random variable then $Y = -\log(X)$ is an Exponential(1) random variable.

The next example yields the 'location-scale' family of normal random variables via a family of linear transformations of the standard normal random variable.

Example 9.8 Let Z be the standard Gaussian or standard normal random variable with probability density function $\varphi(z)$ given by Equation (20). For real numbers $\sigma > 0$ and μ consider the linear transformation of Z given by

$$Y = g(Z) = \sigma Z + \mu \ .$$

Some graphs of such linear transformations of Z are shown in Figures (a) and (b).



(a) $g(z) = z$, $g(z) = \frac{1}{2}z + 5$ and $g(z) = \frac{1}{2}z - 5$.

(b) $g(z) = z$, $g(z) = (z + 5)/0.5$ and $g(z) = (z - 5)/0.5$.

We are interested in the density of the tranformed random variable $Y = g(Z) = \sigma Z + \mu$. Once again, since g is a one-to-one monotone function let us follow the four steps and use the change of variable formula to obtain f_Y from $f_Z = \varphi$ and g.

1. $y = g(z) = \sigma z + \mu$ is a monotone increasing function over $-\infty < z < \infty$, the range of Z. So, we can apply the change of variable formula.

2. $z = g^{-1}(y) = (y - \mu)/\sigma$ is a monotone increasing function over the range of y given by, $-\infty < y < \infty$.

3. For $-\infty < y < \infty$,
$$\left|\frac{d}{dy}g^{-1}(y)\right| = \left|\frac{d}{dy}\left(\frac{y-\mu}{\sigma}\right)\right| = \left|\frac{1}{\sigma}\right| = \frac{1}{\sigma} \ .$$

4. we can use Equation (24) and Equation (20) which gives
$$f_Z(z) = \varphi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \ ,$$
to find the density of Y as follows:
$$f_Y(y) = f_Z\left(g^{-1}(y)\right)\left|\frac{d}{dy}g^{-1}(y)\right| = \varphi\left(\frac{y-\mu}{\sigma}\right)\frac{1}{\sigma} = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right] \ ,$$
for $-\infty < y < \infty$.

Thus, we have obtained the expression for the probability density function of the linear transformation $\sigma Z + \mu$ of the standard normal random variable Z. This analysis leads to the following definition.

Definition 9.9 Given a location parameter $\mu \in (-\infty, +\infty)$ and a scale parameter $\sigma^2 > 0$, the **normal**(μ, σ^2) or **Gauss**(μ, σ^2) **random variable** X has probability density function:

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \qquad (\sigma > 0) \ . \tag{25}$$

This is simpler than it may at first look. $f(x; \mu, \sigma^2)$ has the following features.

- μ is the expected value or mean parameter and σ^2 is the variance parameter. These concepts, mean and variance, are described in more detail in the next section on expectations.
- $1/(\sigma\sqrt{2\pi})$ is a constant factor that makes the area under the curve of $f(x)$ from $-\infty$ to ∞ equal to 1, as it must be.
- The curve of $f(x)$ is symmetric with respect to $x = \mu$ because the exponent is quadratic. Hence for $\mu = 0$ it is symmetric with respect to the y-axis $x = 0$.
- The exponential function decays to zero very fast — the faster the decay, the smaller the value of σ.

The normal distribution has the **distribution function**

$$F(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[-\frac{1}{2}\left(\frac{v-\mu}{\sigma}\right)^2\right] dv \ . \tag{26}$$

Here we need x as the upper limit of integration and so we write v in the integrand.

9.2.2 Direct method

If the transformation g in $Y = g(X)$ is not necessarily one-to-one then special care is needed to obtain the distribution function or density of Y. For a continuous random variable X with a known distribution function F_X we can obtain the distribution function F_Y of $Y = g(X)$ using Equation (22) as follows:

$$\begin{aligned} F_Y(y) &= P(Y \le y) = P(Y \in (-\infty, y]) \\ &= P(g(X) \in (-\infty, y]) = P\left(X \in g^{[-1]}((-\infty, y])\right) = P(X \in \{x : g(x) \in (-\infty, y]\}) \ . \end{aligned}$$

In words, the above equalities just mean that the probability that $Y \le y$ is the probability that X takes a value x that satisfies $g(x) \le y$. We can use this approach if it is reasonably easy to find the set $g^{[-1]}((-\infty, y]) = \{x : g(x) = (-\infty, y]\}$.

Example 9.10 Let X be any random variable with distribution function F_X. Let $Y = g(X) = X^2$. Then we can find F_Y, the distribution function of Y from F_X as follows:

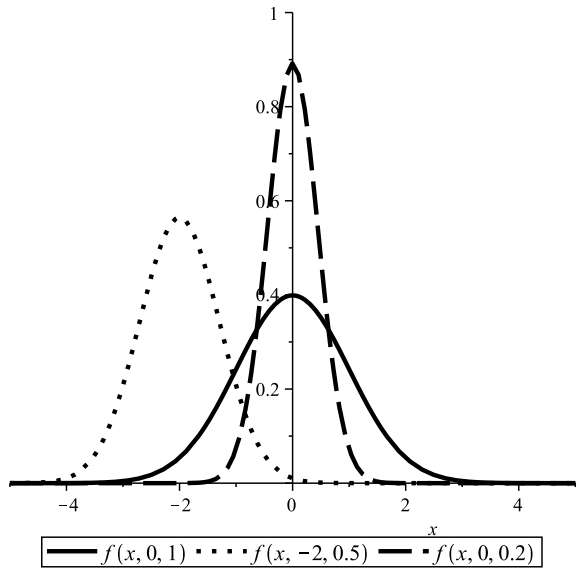


Figure 2: Probability density function of a Normal(μ, σ^2) RV for different values of μ and σ^2

- Since $Y = X^2 \geq 0$, if $y < 0$ then $F_Y(y) = P\left(X \in \{x : x^2 < y\}\right) = P(X \in \emptyset) = 0$.
- If $y \geq 0$ then

$$\begin{aligned} F_Y(y) = P\left(Y \leq y\right) &= P\left(X^2 \leq y\right) \\ &= P\left(-\sqrt{y} \leq X \leq \sqrt{y}\right) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}) \ . \end{aligned}$$

By differentiation we get:

- If $y < 0$ then $f_Y(y) = \frac{d}{dy}(F_Y(y)) = \frac{d}{dy}0 = 0$.
- If $y \geq 0$ then

$$\begin{aligned} f_Y(y) = \frac{d}{dy}\left(F_Y(y)\right) &= \frac{d}{dy}\left(F_X(\sqrt{y}) - F_X(-\sqrt{y})\right) \\ &= \frac{d}{dy}\left(F_X(\sqrt{y})\right) - \frac{d}{dy}\left(F_X(-\sqrt{y})\right) \\ &= \frac{1}{2}y^{-\frac{1}{2}} f_X(\sqrt{y}) - \left(-\frac{1}{2}y^{-\frac{1}{2}} f_X(-\sqrt{y})\right) \\ &= \frac{1}{2\sqrt{y}}\left(f_X(\sqrt{y}) + f_X(-\sqrt{y})\right) \ . \end{aligned}$$

Therefore, the distribution function of $Y = X^2$ is:

$$F_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ F_X(\sqrt{y}) - F_X(-\sqrt{y}) & \text{if } y \geq 0 \ . \end{cases} \tag{27}$$

and the probability density function of $Y = X^2$ is:

$$f_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ \frac{1}{2\sqrt{y}} \left(f_X(\sqrt{y}) + f_X(-\sqrt{y}) \right) & \text{if } y \geq 0 \end{cases} . \tag{28}$$

Example 9.11 If X is the standard normal random variable with density $f_X(x) = \varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-x^2/2\right)$ then by Equation (28) the density of $Y = X^2$ is:

$$f_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ \frac{1}{2\sqrt{y}} \left(f_X(\sqrt{y}) + f_X(-\sqrt{y}) \right) = \frac{1}{\sqrt{2\pi y}} \exp\left(-\frac{y}{2}\right) & \text{if } y \geq 0 \end{cases} .$$

Y is called the **chi-square** random variable with one degree of freedom.

Using the direct method, we can obtain the distribution function of the Normal(μ, σ^2) random variable from that of the tabulated distribution function of the Normal$(0, 1)$ or simply standard normal random variable.

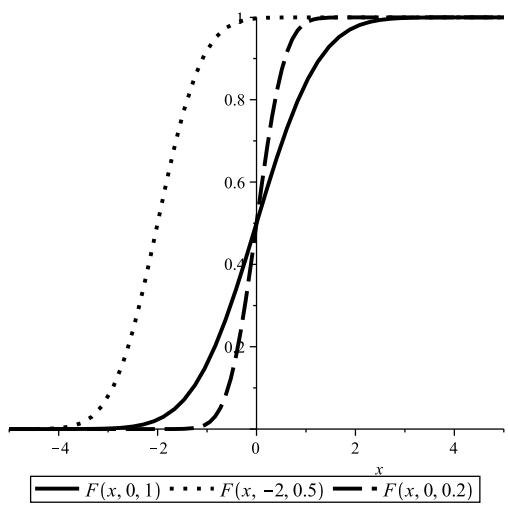


Figure 3: DF of a Normal(μ, σ^2) RV for different values of μ and σ^2

Theorem 9.12 The distribution function $F_X(x; \mu, \sigma^2)$ of the Normal(μ, σ^2) random variable X and the distribution function $F_Z(z) = \Phi(z)$ of the standard normal random variable Z are related by:

$$F_X(x; \mu, \sigma^2) \;=\; F_Z\left(\frac{x-\mu}{\sigma}\right) \;=\; \Phi\left(\frac{x-\mu}{\sigma}\right) \;.$$

Proof: Let Z be a Normal$(0,1)$ random variable with distribution function $\Phi(z) = P(Z \leq z)$. We know that if $X = g(Z) = \sigma Z + \mu$ then X is the Normal(μ, σ^2) random variable. Therefore,

$$\begin{aligned} F_X(x;\mu,\sigma^2) &= P(X \leq x) = P\left(g(Z) \leq x\right) = P(\sigma Z + \mu \leq x) = P\left(Z \leq \frac{x-\mu}{\sigma}\right) \\ &= F_Z\left(\frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right) \quad . \end{aligned}$$

Hence we often transform a general Normal(μ, σ^2) random variable, X, to a standardised Normal$(0,1)$ random variable, Z, by the substitution:

$$Z = \frac{X-\mu}{\sigma}.$$

Exercise 9.13 Suppose that the amount of cosmic radiation to which a person is exposed when flying by jet across the United States is a random variable, X, having a normal distribution with a mean of 4.35 mrem and a standard deviation of 0.59 mrem. What is the probability that a person will be exposed to more than 5.20 mrem of cosmic radiation on such a flight?

Solution:

$$\begin{aligned} P(X > 5.20) &= 1 - P(X \leq 5.20) \\ &= 1 - F(5.20) \\ &= 1 - \Phi\left(\frac{5.20 - 4.35}{0.59}\right) \\ &= 1 - \Phi(1.44) \\ &= 1 - 0.9251 \\ &= 0.0749 \end{aligned}$$

10 Expectations of functions of random variables

Expectation is one of the fundamental concepts in probability. The expected value of a real-valued random variable gives the population mean, a measure of the centre of the distribution of the variable in some sense. More importantly, by taking the expected value of various functions of a random variable, we can measure many interesting features of its distribution, including spread and correlation.

Definition 10.1 The **Expectation** of a function $g(X)$ of a random variable X is defined as:

$$E(g(X)) = \begin{cases} \sum_x g(x)f(x) & \text{if } X \text{ is a discrete RV} \\ \int_{-\infty}^{\infty} g(x)f(x)dx & \text{if } X \text{ is a continuous RV} \end{cases}$$

Definition 10.2 The **population mean** characterises the central location of the random variable X. It is the expectation of the function $g(x) = x$:

$$E(X) \;=\; \begin{cases} \displaystyle\sum_x x f(x) & \text{if } X \text{ is a discrete RV} \\ \displaystyle\int_{-\infty}^{\infty} x f(x) dx & \text{if } X \text{ is a continuous RV} \end{cases}$$

Often, population mean is denoted by μ.

Definition 10.3 Population variance characterises the spread or the variability of the random variable X. It is the expectation of the function $g(x) = (x - E(X))^2$:

$$V(X) \;=\; E\left((X - E(X))^2\right) \;=\; \begin{cases} \displaystyle\sum_x (x - E(X))^2 f(x) & \text{if } X \text{ is a discrete RV} \\ \displaystyle\int_{-\infty}^{\infty} (x - E(X))^2 f(x) dx & \text{if } X \text{ is a continuous RV} \end{cases}$$

Often, population variance is denoted by σ^2.

Definition 10.4 Population Standard Deviation is the square root of the variance, and it is often denoted by σ.

WHAT IS EXPECTATION?

Definition 10.1 gives expectation as a "weighted average" of the possible values. This is true but some intuitive idea of expectation is also helpful.

- Expectation is what you expect.

  Consider tossing a fair coin. If it is heads you lose \$10. If it is tails you win \$10. What do you expect to win? Nothing. If X is the amount you win then

  $$E(X) \;=\; -10 \times \frac{1}{2} + 10 \times \frac{1}{2} \;=\; 0\,.$$

  So what you expect (nothing) and the weighted average ($E(X) = 0$) agree.

- Expectation is a long run average.

  Suppose you are able to repeat an experiment independently, over and over again. Each experiment produces one value x of a random variable X. If you take the average of the x values for a large number of trials, then this average converges to $E(X)$ as the number of trials grows. In fact, this is called the **law of large numbers**.

Properties of expectations

The following results, where a is a constant, may easily be proved using the properties of summations and integrals:

$$\boxed{E(a) \;=\; a}$$

$$\boxed{E(a\,g(X)) \;=\; a\,E(g(X))}$$

$$\boxed{E(g(X)+h(X)) \;=\; E(g(X)) \;+\; E(h(X))}$$

Note that here $g(X)$ and $h(X)$ are functions of the random variable X: e.g. $g(X) = X^2$.
Using these results we can obtain the following useful formula for variance:

$$\begin{aligned}
V(X) &= E\left((X-E(X))^2\right) \\
&= E\left(X^2 - 2XE(X) + (E(X))^2\right) \\
&= E(X^2) - E\left(2XE(X)\right) + E\left((E(X))^2\right) \\
&= E(X^2) - 2E(X)E\left(X\right) + (E(X))^2 \\
&= E(X^2) - 2(E(X))^2 + (E(X))^2 \\
&= E(X^2) - (E(X))^2 \;\;.
\end{aligned}$$

That is,

$$\boxed{V(X) \;=\; E(X^2) \;-\; (E(X))^2}.$$

The above properties of expectations imply that for constants a and b,

$$\boxed{V(aX+b) = a^2V(X) \;\;.}$$

More generally, for random variables $X_1, X_2, \ldots, X_n$ and constants $a_1, a_2, \ldots, a_n$

- $\boxed{E\left(\sum_{i=1}^{n} a_iX_i\right) = \sum_{i=1}^{n} a_iE(X_i)}$

- $\boxed{V\left(\sum_{i=1}^{n} a_iX_i\right) = \sum_{i=1}^{n} a_i^2V(X_i)}$, provided $X_1, X_2, \ldots, X_n$ are independent

Exercise 10.5

The continuous random variable, X, has probability density function given by

$$f(x) \;=\; \begin{cases} 4x^3 & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

(a) Find $E(X)$.

(b) Find $V(X)$

Solution

(a)

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x\, f(x)\, dx \\ &= \int_0^1 x \cdot 4x^3\, dx \\ &= \int_0^1 4x^4\, dx \\ &= \frac{4}{5}x^5\Big]_0^1 \\ &= \frac{4}{5} \end{aligned}$$

(b) We use the formula

$$V(X) \;=\; E(X^2) \;-\; (E(X))^2$$

We already know $E(X) = \frac{4}{5}$ so we only need to find $E(X^2)$

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} x^2\, f(x)\, dx \\ &= \int_0^1 x^2\, 4x^3\, dx \\ &= \int_0^1 4x^5\, dx \\ &= \frac{4x^6}{6}\Big]_0^1 \\ &= \frac{2}{3} \end{aligned}$$

Therefore,

$$V(X) \;=\; \frac{4}{6} - \left(\frac{4}{5}\right)^2 \;=\; 0.0266$$

Note: $\sigma = \sqrt{V(X)} = 0.163$ is the standard deviation.

Exercise 10.6

A management consultant has been hired to evaluate the success of running an international tennis tournament in Christchurch. The consultant reaches the following conclusions.

Predicted Profit	Scenarios
\$2,000,000	1. Federer and Nadal both play
\$1,000,000	2. Nadal plays but not Federer
-\$500,000	3. Neither Federer nor Nadal plays

The consultant assesses the chance of scenarios 1, 2, 3 as $P(1) = 0.08$, $P(2) = 0.22$, $P(3) = 0.7$.

Should the tournament be run?

Solution Let X = profit in millions of dollars. Then we have the discrete random variable

x	$P(X = x)$
2	0.08
1	0.22
-0.5	0.7

$$\begin{aligned} E(X) &= \sum_x x f(x) \\ &= 2 \times 0.08 + 1 \times 0.22 + (-0.5) \times 0.7 \\ &= 0.03 \end{aligned}$$

So you expect to make a profit of 0.03 million, that is, \$30,000. Should you go ahead?

Look at the table! $P(X = -0.5) = 0.7$ so you have a 70% chance of losing half a million. The choice is yours. However, it would be nice to see Federer!

Exercise 10.7

Recall that the random variable X with density

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases}$$

is uniformly distributed on the interval $[a, b]$.

(a) Find $E(X)$.

(b) Find $V(X)$.

Solution

(a) From the definition of the expectation of a continuous random variable, we find that

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x\, f(x)\, dx \\ &= \int_a^b x \left(\frac{1}{b-a}\right) dx \\ &= \frac{1}{b-a} \int_a^b x\, dx \\ &= \frac{b^2-a^2}{2(b-a)} \\ &= \frac{(b+a)(b-a)}{2(b-a)} = \frac{a+b}{2}. \end{aligned}$$

(b) Since

$$\begin{aligned} E(X^2) &= \int_a^b x^2 f(x)\, dx \\ &= \frac{1}{b-a} \int_a^b x^2\, dx \\ &= \frac{1}{b-a} \left[\frac{1}{3}x^3\right]_a^b \\ &= \frac{1}{3(b-a)}(b^3-a^3) \\ &= \frac{(b-a)(b^2+ab+a^2)}{3(b-a)} \\ &= \frac{b^2+ab+a^2}{3}, \end{aligned}$$

the variance is

$$\begin{aligned} V(X) &= E(X^2) - (E(X))^2 \\ &= \frac{b^2+ab+a^2}{3} - \left(\frac{a+b}{2}\right)^2 \\ &= \frac{b^2+ab+a^2}{3} - \frac{a^2+2ab+b^2}{4} \\ &= \frac{4b^2+4ab+4a^2-3a^2-6ab-3b^2}{12} \\ &= \frac{b^2-2ab+a^2}{12} = \frac{(b-a)^2}{12}. \end{aligned}$$

Example 10.8 Let us obtain the expectation and variance of the Normal(μ, σ^2) random variable $X = \sigma Z + \mu$, where Z is Normal$(0, 1)$ with $E(Z) = 0$ and $V(Z) = 1$ as follows:

$$E(X) = E(\sigma Z + \mu) = E(\sigma Z) + E(\mu) = \sigma E(Z) + \mu = \sigma \times 0 + \mu = \mu \ ,$$

$$V(X) = V(\sigma Z + \mu) = \sigma^2 V(Z) + 0 = \sigma^2 \times 1 = \sigma^2 \ .$$

Thus, the population mean and variance of the Normal(μ, σ^2) random variable X are

$$\boxed{E(X) \ = \ \mu, \qquad V(X) \ = \ \sigma^2} \ .$$

Example 10.9 Let us obtain the population mean and variance of the Exponential(λ) random variable X.

$$E(X; \lambda) = \int_0^\infty x f_X(x; \lambda) dx = \int_0^\infty x\lambda \exp(-\lambda x) dx = (-x \exp(-\lambda x)]_0^\infty - \int_0^\infty -1 \exp(-\lambda x) dx \ .$$

The last equality above is a result of integration by parts(I.B.P.) with $u(x) = x$, $du(x) = 1dx$, $dv(x) = \lambda \exp(-\lambda x)dx$, $v(x) = \int dv(x) = -\exp(-\lambda x)$ and the formula $\int u(x)dv(x) = u(x)v(x) - \int v(x)du(x)$. Continuing on with the computation we get,

$$\begin{aligned}
E(X; \lambda) &= (-x \exp(-\lambda x)]_0^\infty + \int_0^\infty 1 \exp(-\lambda x) dx \\
&= (-x \exp(-\lambda x)]_0^\infty + \left(-\frac{1}{\lambda} \exp(-\lambda x)\right]_0^\infty \\
&= \lim_{x \to \infty} \frac{-x}{\exp(\lambda x)} - \left(\frac{-0}{\exp(\lambda \times 0)}\right) + \lim_{x \to \infty} \frac{-1}{\lambda \exp(\lambda x)} - \frac{-1}{\lambda \exp(\lambda \times 0)} \\
&= 0 + 0 - 0 + \frac{1}{\lambda} \\
&= \frac{1}{\lambda} \ .
\end{aligned}$$

Similarly, with $u(x) = x^2$, $du(x) = 2xdx$, $dv(x) = \lambda \exp(-\lambda x)dx$ and $v(x) = \int dv(x) = -\exp(-\lambda x)$ we can find that

$$\begin{aligned}
E(X^2; \lambda) &= \int_0^\infty x^2 f_X(x; \lambda) dx = \left(-x^2 \exp(-\lambda x)\right]_0^\infty - \int_0^\infty -2x \exp(-\lambda x) dx \\
&= 0 - 0 + 2 \int_0^\infty x \exp(-\lambda x) dx \\
&= 2 \left(\left(-\frac{x}{\lambda \exp(\lambda x)}\right]_0^\infty - \int_0^\infty \frac{-1}{\lambda} \exp(-\lambda x) dx \right) \ , \text{ I.B.P.; } u = x, dv = \exp(-\lambda x)dx \\
&= 2 \left(0 - 0 + \left(\frac{-1}{\lambda^2 \exp(\lambda x)}\right)_0^\infty \right) \\
&= \frac{2}{\lambda^2} \ .
\end{aligned}$$

Finally,

$$V(X;\lambda) = E(X^2) - (E(X))^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2} \quad .$$

Therefore, the population mean and variance of an Exponential(λ) random variable are $1/\lambda$ and $1/\lambda^2$, respectively.

END of EMTH119 Probability Content (with enhanced notation for EMTH210)

11 Recap of EMTH119 Material

11.1 What we covered in EMTH119

Sets, Experiments, Sample space, Events, Probability, Conditional Probability, Bayes' Rule, Random variables and Distribution Functions, Discrete RVs and Probability Mass Function, Continuous Random Variables and Probability Density function, Common RVs: (Bernoulli, Binomial, Geometric, Poisson, Uniform, Exponential, Normal), Expectations of RVs, Population Mean and Variance, Transformations of RVs.
Instead of doing a full review here, we will review each concept from EMTH119 as it is needed in EMTH210. Take notes in class for brief review of EMTH119 (if you have not taken EMTH119 then please study it ASAP!).

12 Approximate expectations of functions of random variables

In Section 10 we computed the exact expectation and variance of $Y = g(X)$ and in Section 9 we considered methods to obtain the exact distribution of $Y = g(X)$. In more general situations $E(Y)$ and $V(Y)$ may be very hard to compute directly using the exact approaches of the last two Sections. After all, not all integrals are easy! Hence, it is useful to develop approximations to $E(Y)$ and $V(Y)$. These approximations are often called the **delta method** and are based on the **Taylor series approximations** to $Y = g(X)$.
The Taylor series for $g(X)$ about a is:

$$\boxed{g(X) = \sum_{i=0}^{n} g^{(i)}(a)\frac{(X-a)^i}{i!} + \text{ remainder } .}$$

Taking the Taylor series about $a = \mu = E(X)$ and only considering the first two terms gives:

$$g(X) = g(\mu) + g^{(1)}(\mu)(X-\mu) + \text{ remainder } . \tag{29}$$

Taking the Taylor series about $a = \mu = E(X)$ and only considering the first three terms gives:

$$g(X) = g(\mu) + g^{(1)}(\mu)(X-\mu) + g^{(2)}(\mu)\frac{(X-\mu)^2}{2} + \text{ remainder } , \tag{30}$$

By ignoring the remainder term in Equations (29) and (30) we obtain the following **delta method approximations** to $g(X)$:

$$g(X) \approx g(\mu) + g^{(1)}(\mu)(X-\mu) , \tag{31}$$

$$g(X) \approx g(\mu) + g^{(1)}(\mu)(X-\mu) + g^{(2)}(\mu)\frac{(X-\mu)^2}{2} . \tag{32}$$

These approximations lead to reasonable results in situations where $\sigma = \sqrt{V(X)}$, the standard deviation of X, is small compared to $\mu = E(X)$, the mean of X.

Approximate expectation – approximating $\mathbf{E}(\mathbf{g}(\mathbf{X}))$

We can obtain a one-term approximation to $E(g(X))$ by taking expectation on both sides of Equation (31) as follows:

$$\begin{aligned} E(g(X)) &\approx E(g(\mu)) + E\left(g^{(1)}(\mu)(X-\mu)\right) = g(\mu) + g^{(1)}(\mu)\left(E\left(X-\mu\right)\right) \\ &= g(\mu) + g^{(1)}(\mu)\left(E(X) - E(\mu)\right) = g(\mu) + g^{(1)}(\mu)(\mu-\mu) = g(\mu) + 0 \\ &= g(\mu) \ , \end{aligned}$$

since μ, $g(\mu)$ and $g^{(1)}(\mu)$ are constants. Hence we have the one-term approximation:

$$\boxed{E(g(X)) \approx g(\mu) \ .} \tag{33}$$

We can obtain a two-term approximation to $E(g(X))$ by taking expectation on both sides of Equation (32) as follows:

$$\begin{aligned} E(g(X)) &\approx E(g(\mu)) + E\left(g^{(1)}(\mu)(X-\mu)\right) + E\left(g^{(2)}(\mu)\frac{(X-\mu)^2}{2}\right) \\ &= g(\mu) + 0 + g^{(2)}(\mu)\frac{1}{2}E\left((X-\mu)^2\right) \\ &= g(\mu) + \frac{V(X)}{2}g^{(2)}(\mu) \ . \end{aligned}$$

Hence we have the two-term approximation:

$$\boxed{E(g(X)) \approx g(\mu) + \frac{V(X)}{2}g^{(2)}(\mu) \ .} \tag{34}$$

Approximate variance – approximating $\mathbf{V}(\mathbf{g}(\mathbf{X}))$

We can obtain a one-term approximation to $V(g(X))$ by taking variance on both sides of Equation (31) as follows:

$$\begin{aligned} V(g(X)) &\approx V\left(g(\mu) + g^{(1)}(\mu)(X-\mu)\right) \\ &= V(g(\mu)) + \left(g^{(1)}(\mu)\right)^2\left(V\left(X-\mu\right)\right) \\ &= 0 + \left(g^{(1)}(\mu)\right)^2\left(V(X) - V(\mu)\right) = \left(g^{(1)}(\mu)\right)^2\left(V(X) - 0\right) \\ &= \left(g^{(1)}(\mu)\right)^2 V(X) \ , \end{aligned}$$

since $g(\mu)$ and $g^{(1)}(\mu)$ are constants and for constants a and b, $V(aX+b) = a^2V(X)$. Hence we have the one-term approximation:

$$\boxed{V(g(X)) \approx \left(g^{(1)}(\mu)\right)^2 V(X) \ .} \tag{35}$$

Let us now look at some examples next.

Example 12.1 Approximate $E(\sin(X))$ where X is a Uniform$(0,1)$ random variable. We know from Exercise 10.7 that $E(X) = \mu = \frac{1}{2}$ and $V(X) = \frac{1}{12}$. Hence, we have the one-term approximation

$$E\left(\sin(X)\right) \approx \sin(\mu) = \sin\left(\frac{1}{2}\right) \approx 0.48 \ ,$$

and the two-term approximation is

$$E\left(\sin(X)\right) \approx \sin(\mu) + \frac{\frac{1}{12}}{2}\left(-\sin\left(\frac{1}{2}\right)\right) \approx 0.46 \ .$$

The exact answer for this simple problem is given by

$$E(\sin(X)) = \int_0^1 \sin(x)\, 1\, dx = (-\cos(x)]_0^1 = -\cos(1) - (-\cos(0)) = 0.459697694131860 \ .$$

Example 12.2 Approximate $V\left(\frac{1}{1+X}\right)$ where X is the Exponential(λ) random variable. From Example 10.9 we know that $E(X) = \frac{1}{\lambda}$ and $V(X) = \frac{1}{\lambda^2}$. Since $g(x) = (x+1)^{-1}$ we get $g^{(1)}(x) = -(x+1)^{-2}$, hence

$$V\left(\frac{1}{1+X}\right) \approx \left(-\left(\frac{1}{\lambda}+1\right)^{-2}\right)^2 V(X) = \left(\frac{1+\lambda}{\lambda}\right)^{-4}\frac{1}{\lambda^2} = \frac{\lambda^2}{(1+\lambda)^4} \ .$$

Example 12.3 Find the mean and variance of $Y = 2X - 3$ in terms of the mean and variance of X. Note that we can use the basic properties of expectations here and don't have to use the delta method in every situation!

$$E(Y) = 2E(X) - 3, \qquad V(Y) = 4V(X) \ .$$

Exercise 12.4

Suppose X has density

$$f_X(x) = \begin{cases} 3x^2 & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \ . \end{cases}$$

Find the two-term approximation to $E(Y)$, where $Y = 10\log_{10}(X)$, i.e., X is on the decibel scale.

Solution Firstly, we need

$$E(X) = \int_0^1 x f_X(x)dx = \int_0^1 3x^3 dx = \left(\frac{3}{4}x^4\right]_0^1 = \frac{3}{4}(1-0) = \frac{3}{4} \ ,$$

and

$$E(X^2) = \int_0^1 x^2 f_X(x)dx = \int_0^1 3x^4 dx = \left(\frac{3}{5}x^5\right]_0^1 = \frac{3}{5}(1-0) = \frac{3}{5} \ ,$$

to obtain

$$V(X) = E(X^2) - (E(X))^2 = \frac{3}{5} - \left(\frac{3}{4}\right)^2 = \frac{3}{5} - \frac{9}{16} = \frac{3 \times 16 - 9 \times 5}{5 \times 16} = \frac{3}{80} \ .$$

Secondly, we need

$$g^{(1)}(\mu) = \left(\frac{d}{dx} 10 \log_{10}(x)\right]_{x=\mu} = \left(\frac{d}{dx} 10 \frac{\log(x)}{\log(10)}\right]_{x=\mu} = \left(\frac{10}{\log(10)} \frac{1}{x}\right]_{x=\mu} = \frac{10}{\log(10)} \frac{1}{\frac{3}{4}} = \frac{40}{3\log(10)}$$

and

$$g^{(2)}(\mu) = \left(\frac{d}{dx} \frac{10}{\log(10)} \frac{1}{x}\right]_{x=\mu} = \frac{-10}{\log(10)} \left(\frac{3}{4}\right)^{-2} = \frac{-10}{\log(10) \left(\frac{3}{4}\right)^2} \ .$$

Hence, by Equation (34)

$$E(10 \log_{10}(X)) \approx 10 \log_{10}\left(\frac{3}{4}\right) + \frac{\frac{3}{80}}{2} \left(\frac{-10}{\log(10) \left(\frac{3}{4}\right)^2}\right) = -1.394 \ .$$

The exact answer is -1.458.

13 Characteristic Functions

The characteristic function (CF) of a random variable gives another way to specify its distribution. Thus CF is a powerful tool for analytical results involving random variables (more).

Definition 13.1 Let X be a RV and $\imath = \sqrt{-1}$. The function $\phi_X(t) : \mathbb{R} \to \mathbb{C}$ defined by

$$\boxed{\phi_X(t) := E\left(\exp\left(\imath t X\right)\right) = \begin{cases} \sum_x \exp\left(\imath t x\right) f_X(x) & \text{if } X \text{ is discrete RV} \\ \int_{-\infty}^{\infty} \exp\left(\imath t x\right) f_X(x) dx & \text{if } X \text{ is continuous RV} \end{cases}} \tag{36}$$

is called the **characteristic function** of X.

Remark 13.2 $\phi_X(t)$ exists for any $t \in \mathbb{R}$, because

$$\begin{aligned} \phi_X(t) &= E\left(\exp\left(\imath t X\right)\right) \\ &= E\left(\cos(tX) + \imath \sin(tX)\right) \\ &= E\left(\cos(tX)\right) + \imath E\left(\sin(tX)\right) \end{aligned}$$

and the last two expected values are well-defined, because the sine and cosine functions are bounded by $[-1, 1]$.
For a continuous RV, $\int_{-\infty}^{\infty} \exp\left(-\imath t x\right) f_X(x) dx$ is called the *Fourier transform* of f_X. This is the CF but with t replaced by $-t$. You will also encounter Fourier transforms when solving differential equations.

Obtaining Moments from Characteristic Function

Recall that the k-th moment of X is $E(X^k)$ for any $k \in \mathbb{N} := \{1, 2, 3, \ldots\}$ is

$$E(X^k) = \begin{cases} \sum_x x^k f_X(x) & \text{if } X \text{ is a discrete RV} \\ \int_{-\infty}^{\infty} x^2 f_X(x) dx & \text{if } X \text{ is a continuous RV} \end{cases}$$

The characteristic function can be used to derive the moments of X due to the following nice relationship between the the k-th moment of X and the k-th derivative of the CF of X.

Theorem 13.3 (**Moment & CF.**) Let X be a random variable and $\phi_X(t)$ be its CF. If $E(X^k)$ exists and is finite, then $\phi_X(t)$ is k times continuously differentiable and

$$E(X^k) = \frac{1}{\imath^k} \left[\frac{d^k \phi_X(t)}{dt^k} \right]_{t=0} .$$

where $\left[\frac{d^k \phi_X(t)}{dt^k} \right]_{t=0}$ is the k-th derivative of $\phi_X(t)$ with respect to t, evaluated at the point $t = 0$.

Proof The proper proof is very messy so we just give a sketch of the ideas in the proof. Due to the linearity of the expectation (integral) and the derivative operators, we can change the order of operations:

$$\frac{d^k \phi_X(t)}{dt^k} = \frac{d^k}{dt^k} E(\exp(\imath t X)) = E\left(\frac{d^k}{dt^k} \exp(\imath t X) \right) = E\left((\imath X)^k \exp(\imath t X) \right) = \imath^k E\left(X^k \exp(\imath t X) \right)$$

The RHS evaluated at $t = 0$ is

$$\left[\frac{d^k \phi_X(t)}{dt^k} \right]_{t=0} = \left[\imath^k E\left(X^k \exp(\imath t X) \right) \right]_{t=0} = \imath^k E\left(X^k \right)$$

This completes the sketch of the proof.

The above Theorem gives us the relationship between the moments and the derivatives of the CF if we already know that the moment exists. When one wants to compute a moment of a random variable, what we need is the following Theorem.

Theorem 13.4 (**Moments from CF.**) Let X be a random variable and $\phi_X(t)$ be its CF. If $\phi_X(t)$ is k times differentiable at the point $t = 0$, then

1. if k is even, the n-th moment of X exists and is finite for any $0 \leq n \leq k$;
2. if k is odd, the n-th moment of X exists and is finite for any $0 \leq n \leq k-1$.

In both cases,

$$E(X^k) = \frac{1}{\imath^k}\left[\frac{d^k\phi_X(t)}{dt^k}\right]_{t=0} . \tag{37}$$

where $\left[\frac{d^k\phi_X(t)}{dt^k}\right]_{t=0}$ is the k-th derivative of $\phi_X(t)$ with respect to t, evaluated at the point $t = 0$.

Proof For proof see e.g., Ushakov, N. G. (1999) Selected topics in characteristic functions, VSP (p. 39).

Exercise 13.5

Let X be the Bernoulli(θ) RV. Find the CF of X. Then use CF to find $E(X)$, $E(X^2)$ and from this obtain the variance $V(X) = E(X^2) - (E(X))^2$.

Solution **Part 1**
Recall the PMF for this discrete RV with parameter $\theta \in (0, 1)$ is

$$f_X(x;\theta) = \begin{cases} \theta & \text{if } x = 1 \\ 1-\theta & \text{if } x = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Let's first find the CF of X

$$\begin{aligned} \phi_X(t) = E\left(\exp(\imath t X)\right) &= \sum_x \exp(\imath t x) f_X(x;\theta) \qquad \text{By Defn. in Equation (36)} \\ &= \exp(\imath t \times 0)(1-\theta) + \exp(\imath t \times 1)\theta = \exp(0)(1-\theta) + \exp(\imath t)\theta = 1 - \theta + \theta\exp(\imath t) \end{aligned}$$

Part 2:
Let's differentiate CF

$$\frac{d}{dt}\phi_X(t) = \frac{d}{dt}\left(1 - \theta + \theta e^{\imath t}\right) = \theta\imath\exp(\imath t)$$

We get $E(X)$ by evaluating $\frac{d}{dt}\phi_X(t)$ at $t = 0$ and dividing by $\imath$ according to Equation (37) as follows:

$$E(X) = \frac{1}{\imath}\left[\frac{d}{dt}\phi_X(t)\right]_{t=0} = \frac{1}{\imath}\left[\theta\imath\exp(\imath t)\right]_{t=0} = \frac{1}{\imath}\left(\theta\imath\exp(\imath 0)\right) = \theta \ .$$

Similarly from Equation (37) we can get $E(X^2)$ as follows:

$$E(X^2) = \frac{1}{\imath^2}\left[\frac{d^2}{dt^2}\phi_X(t)\right]_{t=0} = \frac{1}{\imath^2}\left[\frac{d}{dt}\frac{d}{dt}\phi_X(t)\right]_{t=0} = \frac{1}{\imath^2}\left[\frac{d}{dt}\theta\imath\exp(\imath t)\right]_{t=0}$$
$$= \frac{1}{\imath^2}\left[\theta\imath^2\exp(\imath t)\right]_{t=0} = \frac{1}{\imath^2}\left(\theta\imath^2\exp(\imath 0)\right) = \theta \ .$$

Finally, from the first and second moments we can get the variance as follows:

$$V(X) = E(X^2) - (E(X))^2 = \theta - \theta^2 = \theta(1-\theta) \ .$$

Let's check that this is what we have as variance for the Bernoulli(θ) RV if we directly computed it using weighted sums in the definition of expectations: $E(X) = 1\times\theta + 0\times(1-\theta) = \theta$, $E(X^2) = 1^2\times\theta + 0^2\times(1-\theta) = \theta$and thus giving the same $V(X) = E(X^2) - (E(X))^2 = \theta - \theta^2 = \theta(1-\theta)$.

Exercise 13.6

Let X be an Exponential(λ) RV. First show that its CF is $\lambda/(\lambda - \imath t)$. Then use CF to find $E(X)$, $E(X^2)$ and from this obtain the variance $V(X) = E(X^2) - (E(X))^2$.

Solution Recall that the PDF of an Exponential(λ) RV for a given parameter $\lambda \in (0,\infty)$ is $\lambda e^{-\lambda x}$ if $x \in [0,\infty)$ and 0 if $x \notin [0,\infty)$.
Part 1: Find the CF.
We will use the fact that

$$\int_0^\infty \alpha e^{-\alpha x}dx = \left[-e^{-\alpha x}\right]_0^\infty = 1$$

$$\phi_X(t) = E(\exp(\imath X t)) = E\left(e^{\imath X t}\right) = \int_{-\infty}^{\infty} e^{\imath x t}\lambda e^{-\lambda x}dx = \lambda\int_0^\infty e^{-(\lambda-\imath t)x}dx$$
$$= \frac{\lambda}{\lambda-\imath t}\int_0^\infty (\lambda-\imath t)e^{-(\lambda-\imath t)x}dx = \frac{\lambda}{\lambda-\imath t}\int_0^\infty \alpha e^{-\alpha x}dx = \frac{\lambda}{\lambda-\imath t} \ ,$$

where $\alpha = \lambda - \imath t$ with $\lambda > 0$.
Alternatively, you can use $e^{\imath t x} = \cos(tx) + \imath\sin(tx)$ and do integration by parts to arrive at the same answer starting from:

$$\phi_X(t) = \int_{-\infty}^{\infty} e^{\imath x t}\lambda e^{-\lambda x}dx = \int_{-\infty}^{\infty}\cos(tx)e^{-\lambda x}dx + \imath\int_{-\infty}^{\infty}\sin(tx)e^{-\lambda x}dx := \frac{\lambda}{\lambda-\imath t} \ .$$

Part 2:
Let us differentiate the CF to get moments using Equation (37) (CF has to be once and twice differentiable at $t = 0$ to get the first and second moments).

$$\frac{d}{dt}\phi_X(t) = \frac{d}{dt}\left(\frac{\lambda}{\lambda-\imath t}\right) = \lambda\left(-1\times(\lambda-\imath t)^{-2}\times\frac{d}{dt}(\lambda-\imath t)\right)$$
$$= \lambda\left(\frac{-1}{(\lambda-\imath t)^2}\times(-\imath)\right) = \frac{\lambda\imath}{(\lambda-\imath t)^2}$$

We get $E(X)$ by evaluating $\frac{d}{dt}\phi_X(t)$ at $t=0$ and dividing by $\imath$ according to Equation (37) as follows:

$$E(X) = \frac{1}{\imath}\left[\frac{d}{dt}\phi_X(t)\right]_{t=0} = \frac{1}{\imath}\left[\frac{\lambda\imath}{(\lambda-\imath t)^2}\right]_{t=0} = \frac{1}{\imath}\left(\frac{\lambda\imath}{\lambda^2}\right) = \frac{1}{\imath}\left(\frac{\imath}{\lambda}\right) = \frac{1}{\lambda}$$

Let's pause and see if this makes sense.... Yes, because the expected value of Exponential(λ) RV is indeed $1/\lambda$ (recall from when we introduced this RV).
Similarly from Equation (37) we can get $E(X^2)$ as follows:

$$\begin{aligned}
E(X^2) &= \frac{1}{\imath^2}\left[\frac{d^2}{dt^2}\phi_X(t)\right]_{t=0} = \frac{1}{\imath^2}\left[\frac{d}{dt}\frac{d}{dt}\phi_X(t)\right]_{t=0} = \frac{1}{\imath^2}\left[\frac{d}{dt}\frac{\lambda\imath}{(\lambda-\imath t)^2}\right]_{t=0} \\
&= \frac{1}{\imath^2}\left[\lambda\imath \times \frac{d}{dt}(\lambda-\imath t)^{-2}\right]_{t=0} = \frac{1}{\imath^2}\left[\lambda\imath\left(-2(\lambda-\imath t)^{-3}\frac{d}{dt}(\lambda-\imath t)\right)\right]_{t=0} \\
&= \frac{1}{\imath^2}\left[\lambda\imath\left(-2(\lambda-\imath t)^{-3}\times(-\imath)\right)\right]_{t=0} = \frac{1}{\imath^2}\left[\frac{2\lambda\imath^2}{(\lambda-\imath t)^3}\right]_{t=0} = \frac{1}{\imath^2}\left(\frac{2\lambda\imath^2}{\lambda^3}\right) = \frac{2}{\lambda^2}\ .
\end{aligned}$$

Finally, from the first and second moments we can get the variance as follows:

$$V(X) = E(X^2) - (E(X))^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{2-1}{\lambda^2} = \frac{1}{\lambda^2}\ .$$

Let's check that this is what we had as variance for the Exponential(λ) RV when we first introduced it and directly computed using integrals for definition of expectation.

Characteristic functions can be used to characterize the distribution of a random variable.
Two RVs X and Y have the same DFs , i.e., $F_X(x) = F_Y(x)$ for all $x \in \mathbb{R}$, if and only if they have the same characteristic functions, i.e. $\phi_X(t) = \phi_Y(t)$ for all $t \in \mathbb{R}$ (for proof see Resnick, S. I. (1999) A Probability Path, Birkhauser).
Thus, if we can show that two RVs have the same CF then we know they are the same. This can be much more challenging or impossible to do directly with their DFs.
Let Z be Normal$(0,1)$, the standard normal RV. We can find the CF for Z using couple of tricks as follows

$$\begin{aligned}
\phi_Z(t) &= E\left(e^{\imath tZ}\right) \\
&= \int_{-\infty}^{\infty} e^{\imath tz} f_Z(z)dz = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\imath tz}e^{-z^2/2}dz = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\imath tz - z^2/2}dz \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-(t^2+(z-\imath t)^2)/2}dz = e^{-t^2/2}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-(z-\imath t)^2/2}dz \\
&= e^{-t^2/2}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-y^2/2}dy \qquad \text{substituting } y = z-\imath t, dy = dz \\
&= e^{-t^2/2}\frac{1}{\sqrt{2\pi}}\sqrt{2\pi} \qquad \text{using the normalizing constant in PDF of Normal}(0,1)\text{ RV} \\
&= e^{-t^2/2}
\end{aligned}$$

Thus the CF of the standard normal RV Z is

$$\boxed{\phi_Z(t) = e^{-t^2/2}} \tag{38}$$

Let X be a RV with CF $\phi_X(t)$. Let Y be a linear transformation of X

$$Y = a + bX$$

where a and b are two constant real numbers and $b \neq 0$. Then the CF of Y is

$$\boxed{\phi_Y(t) = \exp(\imath a t)\phi_X(bt)} \tag{39}$$

Proof This is easy to prove using the definition of CF as follows:

$$\begin{aligned}\phi_Y(t) &= E\left(\exp(\imath t Y)\right) = E\left(\exp(\imath t(a + bX))\right) = E\left(\exp(\imath t a + \imath t b X)\right) \\ &= E\left(\exp(\imath t a)\exp(\imath t b X)\right) = \exp(\imath t a) E\left(\exp(\imath t b X)\right) = \exp(\imath t a)\phi_X(bt)\end{aligned}$$

Exercise 13.7

Let Y be a Normal(μ, σ^2) RV. Recall that Y is a linear transformation of Z, i.e., $Y = \mu + \sigma Z$ where Z is a Normal$(0, 1)$ RV. Using Equations (38) and (39) find the CF of Y.

Solution

$$\begin{aligned}\phi_Y(t) &= \exp(\imath \mu t)\phi_Z(\sigma t), && \text{since } Y = \mu + \sigma Z \\ &= e^{\imath \mu t} e^{(-\sigma^2 t^2)/2}, && \text{since } \phi_Z(t) = e^{-t^2/2} \\ &= e^{\imath \mu t - (\sigma^2 t^2)/2}\end{aligned}$$

A generalization of (39) is the following. If $X_1, X_2, \ldots, X_n$ are independent RVs and $a_1, a_2, \ldots, a_n$ are some constants, then the CF of the linear combination $Y = \sum_{i=1}^n a_i X_i$ is

$$\boxed{\phi_Y(t) = \phi_{X_1}(a_1 t) \times \phi_{X_2}(a_2 t) \times \cdots \times \phi_{X_n}(a_n t) = \prod_{i=1}^n \phi_{X_i}(a_i t) \ .} \tag{40}$$

Exercise 13.8

Using the following three facts:

- Eqn. (40)
- the Binomial(n, θ) RV Y is the sum of n independent Bernoulli(θ) RVs (from EMTH119)
- the CF of Bernoulli(θ) RV (from lecture notes for EMTH210)

find the CF of the Binomial(n, θ) RV Y.

Solution Let $X_1, X_2, \ldots, X_n$ be independent Bernoulli(θ) RVs with CF $\left(1-\theta+\theta e^{\imath t}\right)$ then $Y = \sum_{i=1}^{n} X_i$ is the Binomial(n, θ) RV and by Eqn. (40) with $a_1 = a_2 = \cdots = 1$, we get

$$\phi_Y(t) = \phi_{X_1}(t) \times \phi_{X_2}(t) \cdots \phi_{X_n}(t) = \prod_{i=1}^{n} \phi_{X_i}(t) = \prod_{i=1}^{n} \left(1-\theta+\theta e^{\imath t}\right) = \left(1-\theta+\theta e^{\imath t}\right)^n \quad .$$

Exercise 13.9

Let Z_1 and Z_2 be independent Normal$(0, 1)$ RVs.

1. Use Eqn. (40) to find the CF of $Z_1 + Z_2$.
2. From the CF of $Z_1 + Z_2$ identify what RV it is.
3. Use Eqn. (40) to find the CF of $2Z_1$.
4. From the CF of $2Z_1$ identify what RV it is.
5. Try to understand the difference between the distributions of $Z_1 + Z_2$ and $2Z_1$ inspite of Z_1 and Z_2 having the same distibution.

Hint: from lectures we know that $\phi_X(t) = e^{\imath\mu t - (\sigma^2 t^2)/2}$ for a Normal(μ, σ^2) RV X.

Solution

1. By Eqn. (40) we just multiply the characteristic functions of Z_1 and Z_2, both of which are $e^{-t^2/2}$,

$$\phi_{Z_1+Z_2}(t) = \phi_{Z_1}(t) \times \phi_{Z_2}(t) = e^{-t^2/2} \times e^{-t^2/2} = e^{-2t^2/2} = e^{-t^2} \quad .$$

2. The CF of $Z_1 + Z_2$ is that of the Normal(μ, σ^2) RV with $\mu = 0$ and $\sigma^2 = 2$. Thus $Z_1 + Z_2$ is the Normal$(0, 2)$ RV with mean parameter $\mu = 0$ and variance paramter $\sigma^2 = 2$.
3. We can again use Eqn. (40) to find the CF of $2Z_1$ as follows

$$\phi_{2Z_1} = \phi_{Z_1}(2t) = e^{-2^2 t^2/2} \quad .$$

4. The CF of $2Z_1$ is that of the Normal(μ, σ^2) RV with $\mu = 0$ and $\sigma^2 = 2^2 = 4$. Thus $2Z_1$ is the Normal$(0, 4)$ RV with mean parameter $\mu = 0$ and variance paramter $\sigma^2 = 4$.
5. $2Z_1$ has a bigger variance from multiplying the standard normal RV by 2 while Z_1+Z_2 has a smaller variance from adding two independent standard normal RVs. Thus, the result of adding the same RV twice does not have the same distribution as that of multiplying it by 2. In other words $2 \times Z$ is not equal to $Z + Z$ in terms of its probability distribution!

14 Random Vectors

Often, in experiments we are measuring two or more aspects simultaneously. For example, we may be measuring the diameters and lengths of cylindrical shafts manufactured in a plant or heights, weights and blood-sugar levels of individuals in a clinical trial. Thus, the underlying outcome $\omega \in \Omega$ needs to be mapped to measurements as realizations of random vectors in the real plane $\mathbb{R}^2 = (-\infty, \infty) \times (-\infty, \infty)$ or the real space $\mathbb{R}^3 = (-\infty, \infty) \times (-\infty, \infty) \times (-\infty, \infty)$:

$$\omega \mapsto (X(\omega), Y(\omega)) : \Omega \to \mathbb{R}^2 \qquad\qquad \omega \mapsto (X(\omega), Y(\omega), Z(\omega)) : \Omega \to \mathbb{R}^3$$

More generally, we may be interested in heights, weights, blood-sugar levels, family medical history, known allergies, etc. of individuals in the clinical trial and thus need to make m measurements of the outcome in $\mathbb{R}^m$ using a "measurable mapping" from $\Omega \to \mathbb{R}^m$. To deal with such multivariate measurements we need the notion of **random vectors** ($\vec{\text{RV}}$s), i.e. ordered pairs of random variables (X, Y), ordered triples of random variables (X, Y, Z), or more generally ordered m-tuples of random variables $(X_1, X_2, \ldots, X_m)$.

14.1 Bivariate Random Vectors

We first focus on understanding (X, Y), a bivariate $\vec{\text{RV}}$ that is obtained from a pair of discrete or continuous RVs. We then generalize to random vectors of length $m > 2$ in the next section.

Definition 14.1 The **joint distribution function (JDF)** or **joint cumulative distribution function (JCDF)**, $F_{X,Y}(x, y) : \mathbb{R}^2 \to [0, 1]$, of the bivariate random vector (X, Y) is

$$\begin{aligned} F_{X,Y}(x, y) &= P(X \leq x \cap Y \leq y) \;=\; P(X \leq x, Y \leq y) \\ &= P\left(\{\omega : X(\omega) \leq x, Y(\omega) \leq y\}\right), \text{ for any } (x, y) \in \mathbb{R}^2 \;, \end{aligned} \tag{41}$$

where the right-hand side represents the probability that the random vector (X, Y) takes on a value in $\{(x', y') : x' \leq x, y' \leq y\}$, the set of points in the plane that are south-west of the point (x, y).

The JDF $F_{X,Y}(x, y) : \mathbb{R}^2 \to \mathbb{R}$ satisfies the following properties:

1. $0 \leq F_{X,Y}(x, y) \leq 1$
2. $F_{X,Y}(x, y)$ is an non-decreasing function of both x and y
3. $F_{X,Y}(x, y) \to 1$ as $x \to \infty$ and $y \to \infty$
4. $F_{X,Y}(x, y) \to 0$ as $x \to -\infty$ and $y \to -\infty$

Definition 14.2 If (X, Y) is a **discrete random vector** that takes values in a discrete support set $\mathcal{S}_{X,Y} = \{(x_i, y_j) : i = 1, 2, \ldots, j = 1, 2, \ldots\} \subset \mathbb{R}^2$ with probabilities $p_{i,j} =$

$P(X = x_i, Y = y_j) > 0$, then its **joint probability mass function** (or JPMF) is:

$$f_{X,Y}(x_i, y_j) = P(X = x_i, Y = y_j) = \begin{cases} p_{i,j} & \text{if } (x_i, y_j) \in \mathcal{S}_{X,Y} \\ 0 & \text{otherwise} \end{cases} . \tag{42}$$

Since $P(\Omega) = 1$, $\sum_{(x_i,y_j)\in\mathcal{S}_{X,Y}} f_{X,Y}(x_i, y_j) = 1$.

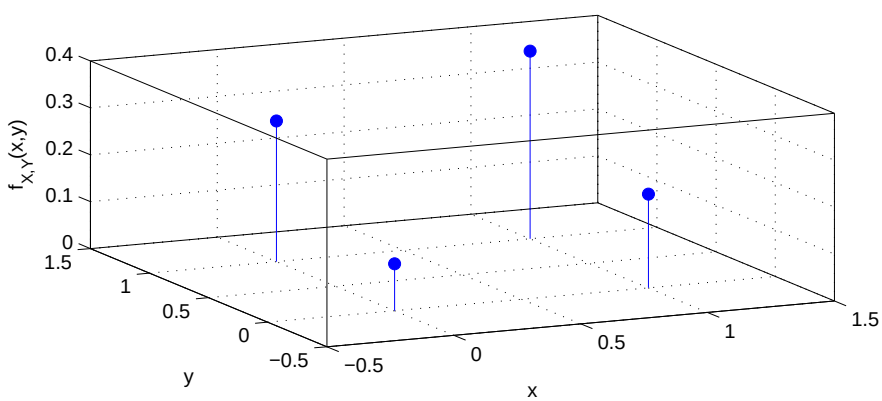
From JPMF $f_{X,Y}$ we can get the values of the JDF $F_{X,Y}(x, y)$ and the probability of any event B by simply taking sums,

$$\boxed{F_{X,Y}(x, y) \;=\; \sum_{x_i \le x, y_j \le y} f_{X,Y}(x_i, y_j)} \;, \qquad \boxed{P(B) \;=\; \sum_{(x_i,y_j)\in B\cap\mathcal{S}_{X,Y}} f_{X,Y}(x_i, y_j)} \;, \tag{43}$$

Exercise 14.3

Let (X, Y) be a discrete bivariate R$\vec{\text{V}}$with the following joint probability mass function (JPMF):

$$f_{X,Y}(x, y) := P(X = x, Y = y) = \begin{cases} 0.1 & \text{if } (x, y) = (0, 0) \\ 0.3 & \text{if } (x, y) = (0, 1) \\ 0.2 & \text{if } (x, y) = (1, 0) \\ 0.4 & \text{if } (x, y) = (1, 1) \\ 0.0 & \text{otherwise.} \end{cases}$$



It is helpful to write down the JPMF $f_{X,Y}(x, y)$ in a tabular form:

	$Y = 0$	$Y = 1$
$X = 0$	0.1	0.3
$X = 1$	0.2	0.4

From the above Table we can read for instance that the joint probability $f_{X,Y}(0, 0) = 0.1$.
Find $P(B)$ for the event $B = \{(0, 0), (1, 1)\}$, $F_{X,Y}(1/2, 1/2)$, $F_{X,Y}(3/2, 1/2)$, $F_{X,Y}(4, 5)$ and $F_{X,Y}(-4, -1)$.

Solution

1. $P(B) = \sum_{(x,y)\in\{(0,0),(1,1)\}} f_{X,Y}(x,y) = f_{X,Y}(0,0) + f_{X,Y}(1,1) = 0.1 + 0.4$
2. $F_{X,Y}(1/2, 1/2) = \sum_{\{(x,y):x\leq 1/2, y\leq 1/2\}} f_{X,Y}(x,y) = f_{X,Y}(0,0) = 0.1$
3. $F_{X,Y}(3/2, 1/2) = \sum_{\{(x,y):x\leq 3/2, y\leq 1/2\}} f_{X,Y}(x,y) = f_{X,Y}(0,0) + f_{X,Y}(1,0) = 0.1 + 0.2 = 0.3$
4. $F_{X,Y}(4,5) = \sum_{\{(x,y):x\leq 4, y\leq 5\}} f_{X,Y}(x,y) = f_{X,Y}(0,0) + f_{X,Y}(0,1) + f_{X,Y}(1,0) + f_{X,Y}(1,1) = 1$
5. $F_{X,Y}(-4,-1) = \sum_{\{(x,y):x\leq -4, y\leq -1\}} f_{X,Y}(x,y) = 0$

Definition 14.4 (X, Y) is a **continuous random vector** if its JDF $F_{X,Y}(x,y)$ is differentiable and the **joint probability density function (JPDF)** is given by:

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y) \ ,$$

From JPDF $f_{X,Y}$ we can compute the JDF $F_{X,Y}$ at any point $(x,y) \in \mathbb{R}^2$ and more generally we can compute the probability of any event B, that can be cast as a region in $\mathbb{R}^2$, by simply taking two-dimensional integrals:

$$\boxed{F_{X,Y}(x,y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(u,v) du dv} \ , \tag{44}$$

and

$$\boxed{P(B) \ = \ \int \int_{B} f_{X,Y}(x,y) dx dy} \ . \tag{45}$$

The JPDF satisfies the following two properties:

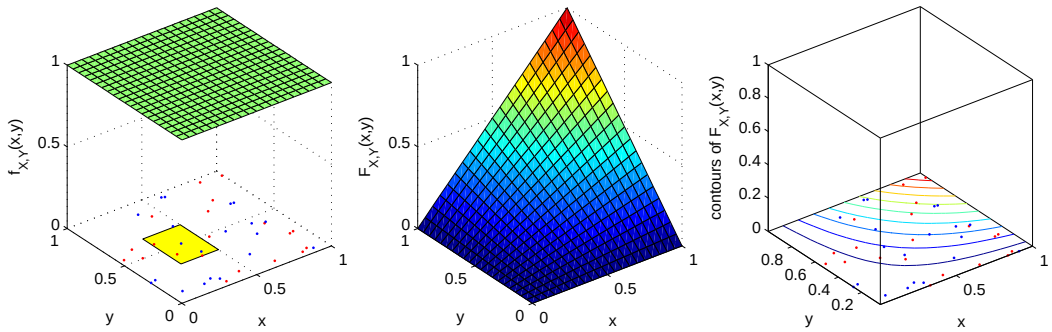
1. integrates to 1, i.e., $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx dy = 1$
2. is a non-negative function, i.e., $f_{X,Y}(x,y) \geq 0$i for every $(x,y) \in \mathbb{R}^2$.

Exercise 14.5

Let (X, Y) be a continuous R$\vec{\text{V}}$ that is uniformly distributed on the unit square $[0,1]^2 := [0,1] \times [0,1]$ with following JPDF:

$$f(x,y) = \begin{cases} 1 & \text{if } (x,y) \in [0,1]^2 \\ 0 & \text{otherwise.} \end{cases}$$

Find the following: (1) DF $F(x,y)$ for any $(x,y) \in [0,1]^2$, (2) $P(X \leq 1/3, Y \leq 1/2)$, (3) $P\left((X,Y) \in [1/4, 1/2] \times [1/3, 2/3]\right)$.



Solution

1. Let $(x, y) \in [0, 1]^2$ then by Equation (44):

$$F_{X,Y}(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(u, v)dudv = \int_0^y \int_0^x 1dudv = \int_0^y [u]_{u=0}^{x} dv = \int_0^y xdv \quad = [xv]_{v=0}^{y} = xy$$

2. We can obtain $P(X \leq 1/3, Y \leq 1/2)$ by evaluating $F_{X,Y}$ at $(1/3, 1/2)$:

$$P(X \leq 1/3, Y \leq 1/2) = F_{X,Y}(1/3, 1/2) = \frac{1}{3}\frac{1}{2} = \frac{1}{6}$$

   We can also find $P(X \leq 1/3, Y \leq 1/2)$ by integrating the JPDF over the rectangular event $A = \{X < 1/3, Y < 1/2\} \subset [0, 1]^2$ according to Equation (45). This amounts here to finding the area of A, we compute $P(A) = (1/3)(1/2) = 1/6$.

3. We can find $P((X, Y) \in [1/4, 1/2] \times [1/3, 2/3])$ by integrating the JPDF over the rectangular event $B = [1/4, 1/2] \times [1/3, 2/3]$ according to Equation (45):

$$\begin{aligned} P((X, Y) \in [1/4, 1/2] \times [1/3, 2/3]) &= \int\int_B f_{X,Y}(x, y)dxdy = \int_{1/3}^{2/3} \int_{1/4}^{1/2} 1dxdy \\ &= \int_{1/3}^{2/3} [x]_{1/4}^{1/2} dy = \int_{1/3}^{2/3} \left[\frac{1}{2} - \frac{1}{4}\right] dy = \left(\frac{1}{2} - \frac{1}{4}\right) [y]_{1/3}^{2/3} \\ &= \left(\frac{1}{2} - \frac{1}{4}\right)\left(\frac{2}{3} - \frac{1}{3}\right) = \frac{1}{4}\left(\frac{1}{3}\right) = \frac{1}{12} \end{aligned}$$

Remark 14.6 In general, for a bivariate uniform $\vec{\text{RV}}$ on the unit square the $P([a, b] \times [c, d]) = (b - a)(d - c)$ for any event given by the rectangular region $[a, b] \times [c, d]$ inside the unit square $[0, 1] \times [0, 1]$. Thus any two events with the same rectangular area have the same probability (imagine sliding a small rectangle inside the unit square... no matter where you slide this rectangle to while remaining in the unit square, the probability of $\omega \mapsto (X(\omega), Y(\omega)) = (x, y)$ falling inside this "slidable" rectangle is the same...).

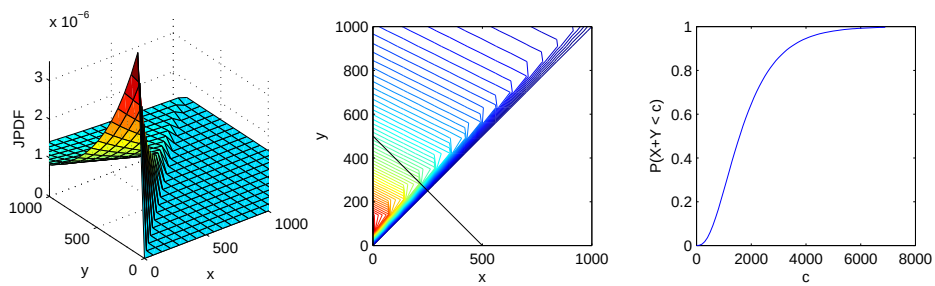
Exercise 14.7

Let the RV X denote the time until a web server connects to your computer, and let the RV Y denote the time until the server authorizes you as a valid user. Each of these RVs measures the waiting time from a common starting time (in milliseconds) and $X < Y$. From past response times of the web server we know that a good approximation for the JPDF of the $\vec{\text{RV}}$ (X, Y) is

$$f_{X,Y}(x,y) = \begin{cases} \frac{6}{10^6} \exp\left(-\frac{1}{1000}x - \frac{2}{1000}y\right) & \text{if } x > 0, y > 0, x < y \\ 0 & \text{otherwise.} \end{cases}$$

Answer the following:

1. identify the support of (X, Y), i.e., the region in the plane where $f_{X,Y}$ takes positive values
2. check that $f_{X,Y}$ indeed integrates to 1 as it should
3. Find $P(X \leq 400, Y \leq 800)$
4. It is known that humans prefer a response time of under 1/10 seconds (10^2 milliseconds) from the web server before they get impatient. What is $P(X + Y < 10^2)$?



Solution

1. The support is the intersection of the positive quadrant with the $y > x$ half-plane.

2.

$$\begin{aligned}
\int_{y=-\infty}^{\infty}\int_{x=-\infty}^{\infty} f_{X,Y}(x,y)dxdy &= \int_{x=0}^{\infty}\int_{y=x}^{\infty} f_{X,Y}(x,y)dydx \\
&= \int_{x=0}^{\infty}\int_{y=x}^{\infty} \frac{6}{10^6}\exp\left(-\frac{1}{1000}x-\frac{2}{1000}y\right)dydx \\
&= \frac{6}{10^6}\int_{x=0}^{\infty}\left(\int_{y=x}^{\infty}\exp\left(-\frac{2}{1000}y\right)dy\right)\exp\left(-\frac{1}{1000}x\right)dx \\
&= \frac{6}{10^6}\int_{x=0}^{\infty}\left[-\frac{1000}{2}\exp\left(-\frac{2}{1000}y\right)\right]_{y=x}^{\infty}\exp\left(-\frac{1}{1000}x\right)dx \\
&= \frac{6}{10^6}\int_{x=0}^{\infty}\left[0+\frac{1000}{2}\exp\left(-\frac{2}{1000}x\right)\right]\exp\left(-\frac{1}{1000}x\right)dx \\
&= \frac{6}{10^6}\int_{x=0}^{\infty}\frac{1000}{2}\exp\left(-\frac{2}{1000}x-\frac{1}{1000}x\right)dx \\
&= \frac{6}{10^6}\frac{1000}{2}\left[-\frac{1000}{3}\exp\left(-\frac{3}{1000}x\right)\right]_{x=0}^{\infty} \\
&= \frac{6}{10^6}\frac{1000}{2}\left[0+\frac{1000}{3}\right] \\
&= 1
\end{aligned}$$

3. First, identify the region with positive JPDF for the event $(X \le 400, Y \le 800)$

$$\begin{aligned}
&P(X \le 400, Y \le 800) \\
&= \int_{x=0}^{400}\int_{y=x}^{800} f_{X,Y}(x,y)dydx \\
&= \int_{x=0}^{400}\int_{y=x}^{800} \frac{6}{10^6}\exp\left(-\frac{1}{1000}x-\frac{2}{1000}y\right)dydx \\
&= \frac{6}{10^6}\int_{x=0}^{400}\left[-\frac{1000}{2}\exp\left(-\frac{2}{1000}y\right)\right]_{y=x}^{800}\exp\left(-\frac{1}{1000}x\right)dx \\
&= \frac{6}{10^6}\frac{1000}{2}\int_{x=0}^{400}\left(-\exp\left(-\frac{1600}{1000}\right)+\exp\left(-\frac{2}{1000}x\right)\right)\exp\left(-\frac{1}{1000}x\right)dx \\
&= \frac{6}{10^6}\frac{1000}{2}\int_{x=0}^{400}\left(\exp\left(-\frac{3}{1000}x\right)-e^{-8/5}\exp\left(-\frac{1}{1000}x\right)\right)dx \\
&= \frac{6}{10^6}\frac{1000}{2}\left(\left(-\frac{1000}{3}\exp\left(-\frac{3}{1000}x\right)\right)_{x=0}^{400}-e^{-8/5}\left(-1000\exp\left(-\frac{1}{1000}x\right)\right)_{x=0}^{400}\right) \\
&= \frac{6}{10^6}\frac{1000}{2}1000\left(\frac{1}{3}\left(1-e^{-6/5}\right)-e^{-8/5}\left(1-e^{-2/5}\right)\right) \\
&= 3\left(\frac{1}{3}\left(1-e^{-6/5}\right)-e^{-8/5}\left(1-e^{-2/5}\right)\right) \\
&\approx 0.499 \ .
\end{aligned}$$

4. First, identify the region with positive JPDF for the event $(X+Y \le c)$, say $c = 500$ (but generally c can be any positive number). This is the triangular region at the intersection

of the four half-planes: $x > 0$, $x < c$, $y > x$ and $y < c - x$. (*Draw picture here*) Let's integrate the JPDF over our triangular event as follows:

$$\begin{aligned}
P(X+Y \le c) &= \int_{x=0}^{c/2} \int_{y=x}^{c-x} f_{X,Y}(x,y)dydx \\
&= \int_{x=0}^{c/2} \int_{y=x}^{c-x} \frac{6}{10^6} \exp\left(-\frac{1}{1000}x - \frac{2}{1000}y\right) dydx \\
&= \frac{6}{10^6} \int_{x=0}^{c/2} \int_{y=x}^{c-x} \exp\left(-\frac{1}{1000}x - \frac{2}{1000}y\right) dydx \\
&= \frac{6}{10^6} \frac{1000}{2} \int_{x=0}^{c/2} \left[-\exp\left(-\frac{2}{1000}y\right)\right]_{y=x}^{c-x} \exp\left(-\frac{1}{1000}x\right) dx \\
&= \frac{3}{10^3} \int_{x=0}^{c/2} \left[-\exp\left(-\frac{2c-2x}{1000}\right) + \exp\left(-\frac{2x}{1000}\right)\right] \exp\left(-\frac{x}{1000}\right) dx \\
&= \frac{3}{10^3} \int_{x=0}^{c/2} \left(\exp\left(-\frac{3x}{1000}\right) - \exp\left(\frac{x-2c}{1000}\right)\right) dx \\
&= 3\left(\left[-\frac{1}{3}\exp\left(-\frac{3x}{1000}\right)\right]_{x=0}^{c/2} - \left[e^{-2c/1000}\exp\left(\frac{x}{1000}\right)\right]_{x=0}^{c/2}\right) \\
&= 3\left(\frac{1}{3}(1-e^{-3c/2000}) - e^{-2c/1000}(e^{c/2000}-1)\right) \\
&= 1 - e^{-3c/2000} + 3e^{-2c/1000} - 3e^{-3c/2000} \\
&= 1 - 4e^{-3c/2000} + 3e^{-c/500}
\end{aligned}$$

5. $P(X+Y<100) = 1 - 4e^{-300/2000} + 3e^{-100/500} \approxeq 0.134$. This means only about one in one hundred requests to this server will be processed within 100 milliseconds.

We can obtain $P(X+Y<c)$ for several values of c using MATLAB and note that about 96% of requests are processed in less than 3000 milliseconds or 3 seconds.

```
>> c = [100 1000 2000 3000 4000]
c =  100        1000        2000        3000        4000

>> p = 1 - 4 * exp(-3*c/2000) + 3 * exp(-c/500)

p =  0.0134     0.5135      0.8558      0.9630      0.9911
```

Definition 14.8 (Marginal PDF or PMF) *If the $R\vec{V}$ (X,Y) has $f_{X,Y}(x,y)$ as its joint PDF or joint PMF, then the* **marginal PDF or PMF** *of a random vector (X,Y) is defined by :*

$$f_X(x) = \begin{cases} \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy & \text{if } (X,Y) \text{ is a continuous } R\vec{V} \\ \sum_y f_{X,Y}(x,y) & \text{if } (X,Y) \text{ is a discrete } R\vec{V} \end{cases}$$

and the **marginal PDF or PMF** *of Y is defined by:*

$$f_Y(y) = \begin{cases} \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx & \text{if } (X,Y) \text{ is a continuous } R\vec{V} \\ \sum_x f_{X,Y}(x,y) & \text{if } (X,Y) \text{ is a discrete } R\vec{V} \end{cases}$$

Exercise 14.9

Obtain the marginal PMFs $f_Y(y)$ and $f_X(x)$ from the joint PMF $f_{X,Y}(x,y)$ of the discrete R$\vec{\text{V}}$ in Exercise 14.3.

Solution Just sum $f_{X,Y}(x,y)$ over x's and y's (reported in a tabular form):

	$Y=0$	$Y=1$
$X=0$	0.1	0.3
$X=1$	0.2	0.4

From the above Table we can find:

$$f_X(x) = P(X=x) = \sum_y f_{X,Y}(x,y)$$

$$= f_{X,Y}(x,0) + f_{X,Y}(x,1) = \begin{cases} f_{X,Y}(0,0) + f_{X,Y}(0,1) = 0.1 + 0.3 = 0.4 & \text{if } x = 0 \\ f_{X,Y}(1,0) + f_{X,Y}(1,1) = 0.2 + 0.4 = 0.6 & \text{if } x = 1 \end{cases}$$

Similarly,

$$f_Y(y) = P(Y=y) = \sum_x f_{X,Y}(x,y)$$

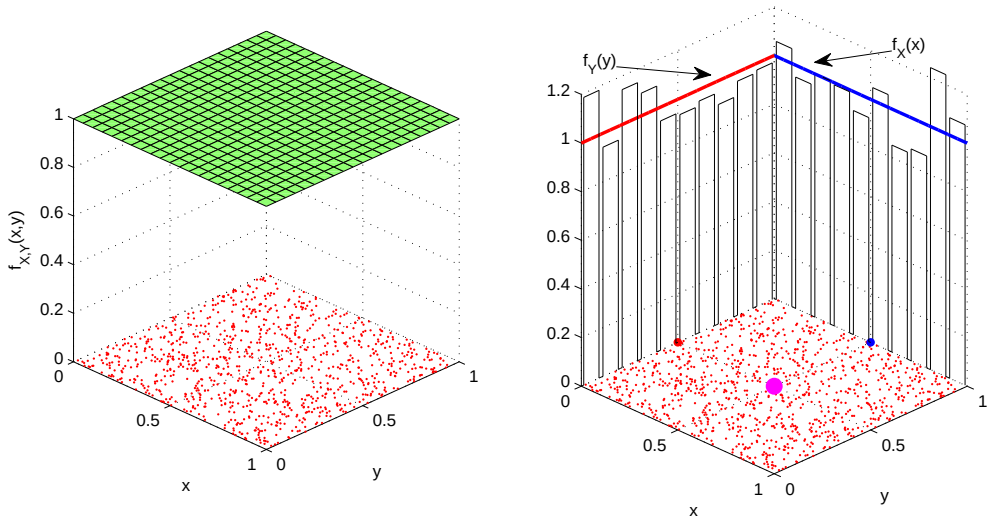
$$= f_{X,Y}(0,y) + f_{X,Y}(1,y) = \begin{cases} f_{X,Y}(0,0) + f_{X,Y}(1,0) = 0.1 + 0.2 = 0.3 & \text{if } y = 0 \\ f_{X,Y}(0,1) + f_{X,Y}(1,1) = 0.3 + 0.4 = 0.7 & \text{if } y = 1 \end{cases}$$

Just report the marginal probabilities as row and column sums of the JPDF table.

Thus marginal PMF gives us the probability of a specific RV, within a R$\vec{\text{V}}$, taking a value irrespective of the value taken by the other RV in this R$\vec{\text{V}}$.

Exercise 14.10

Obtain the marginal PMFs $f_Y(y)$ and $f_X(x)$ from the joint PDF $f_{X,Y}(x,y)$ of the continuous R$\vec{\text{V}}$ in Exercise 14.5 (the bivariate uniform R$\vec{\text{V}}$ on $[0,1]^2$).



Solution Let us suppose $(x, y) \in [0, 1]^2$ and note that $f_{X,Y} = 0$ if $(x, y) \notin [0, 1]^2$. We can obtain marginal PMFs $f_X(x)$ and $f_Y(y)$ by integrating the JPDF $f_{X,Y} = 1$ along y and x, respectively.

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy = \int_0^1 f_{X,Y}(x, y)dy = \int_0^1 1dy = [y]_0^1 = 1 - 0 = 1$$

Similarly,

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dx = \int_0^1 f_{X,Y}(x, y)dx = \int_0^1 1dx = [x]_0^1 = 1 - 0 = 1$$

We are seeing a histogram of the **marginal samples** and their marginal PDFs in the Figure.

Thus marginal PDF gives us the probability density of a specific RV in a R$\vec{\text{V}}$, irrespective of the value taken by the other RV in this R$\vec{\text{V}}$.

Exercise 14.11

Obtain the marginal PDF $f_Y(y)$ from the joint PDF $f_{X,Y}(x, y)$ of the continuous R$\vec{\text{V}}$ in Exercise 14.7 that gave the response times of a web server.

$$f_{X,Y}(x, y) = \begin{cases} \frac{6}{10^6} \exp\left(-\frac{1}{1000}x - \frac{2}{1000}y\right) & \text{if } x > 0, y > 0, x < y \\ 0 & \text{otherwise.} \end{cases}$$

Use $f_Y(y)$ to compute the probability that Y exceeds 2000 milliseconds.

Solution For $y > 0$,

$$\begin{aligned}
f_Y(y) &= \int_{x=-\infty}^{\infty} f_{X,Y}(x,y)dx \\
&= \int_{x=-\infty}^{\infty} 6 \times 10^{-6} e^{-0.001x-0.002y} dx \\
&= 6 \times 10^{-6} \int_{x=0}^{y} e^{-0.001x-0.002y} dx \\
&= 6 \times 10^{-6} e^{-0.002y} \int_{x=0}^{y} e^{-0.001x} dx \\
&= 6 \times 10^{-6} e^{-0.002y} \left[\frac{e^{-0.001x}}{-0.001} \right]_{x=0}^{x=y} \\
&= 6 \times 10^{-6} e^{-0.002y} \left(\frac{e^{-0.001y}}{-0.001} - \frac{e^{-0.001\times 0}}{-0.001} \right) \\
&= 6 \times 10^{-6} e^{-0.002y} \left(\frac{1 - e^{-0.001y}}{0.001} \right) \\
&= 6 \times 10^{-3} e^{-0.002y} \left(1 - e^{-0.001y}\right)
\end{aligned}$$

We have the marginal PDF of Y and from this we can obtain

$$\begin{aligned}
P(Y > 2000) &= \int_{2000}^{\infty} f_Y(y)dy \\
&= \int_{2000}^{\infty} 6 \times 10^{-3} e^{-0.002y} \left(1 - e^{-0.001y}\right) dy \\
&= 6 \times 10^{-3} \int_{2000}^{\infty} e^{-0.002y} dy - \int_{2000}^{\infty} e^{-0.003y} dy \\
&= 6 \times 10^{-3} \left(\left[\frac{e^{-0.002y}}{-0.002} \right]_{2000}^{\infty} - \left(\left[\frac{e^{-0.003y}}{-0.003} \right]_{2000}^{\infty} \right) \right) \\
&= 6 \times 10^{-3} \left(\frac{e^{-4}}{0.002} - \frac{e^{-6}}{0.003} \right) \\
&= 0.05
\end{aligned}$$

Alternatively, you can obtain $P(Y > 2000)$ by directly integrating the joint PDF $f_{X,Y}(x,y)$ over the appropriate region (but you may now have to integrate two pieces: rectangular infinite strip $(x,y) : 0 < x < 2000, y > 2000$ and a triangular infinite piece $\{(x,y) : y > x, y > 2000, x > 2000\}$)... more involved but we get the same answer.

$$\begin{aligned}
P(Y > 2000) = \int_{x=0}^{2000} \left(\int_{y=2000}^{\infty} 6 \times 10^{-6} e^{-0.001x-0.002y} dy \right) dx + \\
\int_{x=2000}^{\infty} \left(\int_{y=x}^{\infty} 6 \times 10^{-6} e^{-0.001x-0.002y} dy \right) dx
\end{aligned}$$

$\vdots$(try as a tutorial problem)

$$P(Y > 2000) = 0.0475 + 0.0025 = 0.05$$

14.2 Expectations of functions of bivariate random vectors

Expectation is one of the fundamental concepts in probability. In the case of a single random variable we saw that its expectation gives the population mean, a measure of the center of the distribution of the variable in some sense. Similarly, by taking the expected value of various functions of a random vector, we can measure many interesting features of its distribution.

Definition 14.12 The **Expectation** of a function $g(X,Y)$ of a random vector (X,Y) is defined as:

$$E(g(X,Y)) = \begin{cases} \sum_{(x,y)} g(x,y) f_{X,Y}(x,y) & \text{if } (X,Y) \text{ is a discrete R}\vec{\text{V}} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\,dxdy & \text{if } (X,Y) \text{ is a continuous R}\vec{\text{V}} \end{cases}$$

Some typical expectations for bivariate random vectors are:

1. Joint Moments
$$E(X^r Y^s)$$

2. We need a new notion for the variance of two RVs.
If $E(X^2) < \infty$ and $E(Y^2) < \infty$ then $E(|XY|) < \infty$ and $E(|(X-E(X))(Y-E(Y))|) < \infty$. This allows the definition of **covariance** of X and Y as
$$\mathbf{Cov}(X,Y) := E\left((X-E(X))(Y-E(Y))\right) = E(XY) - E(X)E(Y)$$

Two RVs X and Y are said to be **independent** if and only if for every (x,y)

$$F_{X,Y}(x,y) = F_X(x) \times F_Y(y) \qquad \text{or } f_{X,Y}(x,y) = f_X(x) \times f_Y(y)$$

Let (X,Y) be the Uniform$([0,1]^2)$ R$\vec{\text{V}}$ that is uniformly distributed on the unit square (introduced in Exercise 14.5). Are X and Y independent? Answer is Yes!

$$\begin{cases} 1 = f_{X,Y}(x,y) = f_X(x) \times f_Y(y) = 1 \times 1 = 1 & \text{if } (x,y) \in [0,1]^2 \\ 0 = f_{X,Y}(x,y) = f_X(x) \times f_Y(y) = 0 \times 0 = 0 & \text{if } (x,y) \notin [0,1]^2 \end{cases}$$

How about the server times R$\vec{\text{V}}$ from Exercise 14.7?
We can compute $f_X(x)$ and use the already computed $f_Y(y)$ to mechanically check if the JPDF is the product of the marginal PDFs. But intuitively, we know that these RVs (connection time and authentication time) are dependent – one is strictly greater than the other. Also the JPDF has zero density when $x > y$, but the product of the marginal densities won't.

Some useful properties of expectations

This is in addition to the ones from EMTH119 (see page 59).
If X and Y are independent then

- $$E(XY) = E(X)E(Y)$$

- $$E\left(g(X)h(Y)\right) = E(g(X))E(h(Y))$$

- $$E\left(aX + bY + c\right) = aE(X) + bE(Y) + c$$

- $$V\left(aX + bY + c\right) = a^2V(X) + b^2V(Y)$$

If X and Y are any two RVs, independent or dependent, then

- $$E\left(aX + bY + c\right) = aE(X) + bE(Y) + c$$

- $$V\left(aX + bY + c\right) = a^2V(X) + b^2V(Y) + 2ab\mathbf{Cov}(X, Y)$$

14.3 Multivariate Random Vectors

Consider the R$\vec{\text{V}}$ X whose components are the RVs $X_1, X_2, \ldots, X_m$, i.e., $X = (X_1, X_2, \ldots, X_m)$, where $m \geq 2$. A particular realization of this R$\vec{\text{V}}$ is a point $(x_1, x_2, \ldots, x_m)$ in $\mathbb{R}^m$. Now, let us extend the notions of JCDF, JPMF and JPDF to $\mathbb{R}^m$.

Definition 14.13 The **joint distribution function (JDF)** or **joint cumulative distribution function (JCDF)**, $F_{X_1,X_2,\ldots,X_m}(x_1, x_2, \ldots, x_m) : \mathbb{R}^m \to [0, 1]$, of the multivariate random vector $(X_1, X_2, \ldots, X_m)$ is

$$\begin{aligned} F_{X_1,X_2,\ldots,X_m}(x_1, x_2, \ldots, x_m) &= P(X \leq x_1 \cap X_2 \leq x_2 \cap \cdots \cap X_m \leq x_m) \\ &= P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_m \leq x_m) \\ &= P\left(\{\omega : X_1(\omega) \leq x_1, X_2(\omega) \leq x_2, \ldots, X_m(\omega) \leq x_m\}\right), \end{aligned} \tag{46}$$

for any $(x_1, x_2, \ldots, x_m) \in \mathbb{R}^m$, where the right-hand side represents the probability that the random vector $(X_1, X_2, \ldots, X_m)$ takes on a value in $\{(x'_1, x'_2, \ldots, x'_m) : x'_1 \leq x_1, x'_2 \leq x_2, \ldots, x'_m \leq x_m\}$, the set of points in $\mathbb{R}^m$ that are less than the point $(x_1, x_2, \ldots, x_m)$ in each coordinate $1, 2, \ldots, m$.

The JDF $F_{X_1,X_2,\ldots,X_m}(x_1, x_2, \ldots, x_m) : \mathbb{R}^m \to \mathbb{R}$ satisfies the following properties:

1. $0 \leq F_{X_1,X_2,\ldots,X_m}(x_1, x_2, \ldots, x_m) \leq 1$
2. $F_{X_1,X_2,\ldots,X_m}(x_1, x_2, \ldots, x_m)$ is an increasing function of x_1, x_2, ... and x_m
3. $F_{X_1,X_2,\ldots,X_m}(x_1, x_2, \ldots, x_m) \to 1$ as $x_1 \to \infty$, $x_2 \to \infty$, ... and $x_m \to \infty$
4. $F_{X_1,X_2,\ldots,X_m}(x_1, x_2, \ldots, x_m) \to 0$ as $x_1 \to -\infty$, $x_2 \to -\infty$, ... and $x_m \to -\infty$

Definition 14.14 If $(X_1, X_2, \ldots, X_m)$ is a **discrete random vector** that takes values in a discrete support set $\mathcal{S}_{X_1,X_2,\ldots,X_m}$, then its **joint probability mass function** (or JPMF) is:

$$f_{X_1,X_2,\ldots,X_m}(x_1, x_2, \ldots, x_m) = P(X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m) \ . \tag{47}$$

Since $P(\Omega) = 1$, $\sum_{(x_1,x_2,\ldots,x_m)\in\mathcal{S}_{X_1,X_2,\ldots,X_m}} f_{X_1,X_2,\ldots,X_m}(x_1, x_2, \ldots, x_m) = 1$.

From JPMF $f_{X_1,X_2,\ldots,X_m}$ we can get the JCDF $F_{X_1,X_2,\ldots,X_m}(x_1, x_2, \ldots, x_m)$ and the probability of any event B by simply taking sums as in Equation (43) but now over all m coordinates.

Definition 14.15 $(X_1, X_2, \ldots, X_m)$ is a **continuous random vector** if its JDF $F_{X_1,X_2,\ldots,X_m}(x_1, x_2, \ldots, x_m)$ is differentiable and the **joint probability density function (JPDF)** is given by:

$$f_{X_1,X_2,\ldots,X_m}(x_1, x_2, \ldots, x_m) = \frac{\partial^m}{\partial x_1 \partial x_2 \cdots \partial x_m} F_{X_1,X_2,\ldots,X_m}(x_1, x_2, \ldots, x_m) \ ,$$

From JPDF $f_{X_1,X_2,\ldots,X_m}$ we can compute the JDF $F_{X_1,X_2,\ldots,X_m}$ at any point $(x_1, x_2, \ldots, x_m) \in \mathbb{R}^m$ and more generally we can compute the probability of any event B, that can be cast as a region in $\mathbb{R}^m$, by "simply" taking m-dimensional integrals (you have done such iterated integrals when $m = 3$):

$$\boxed{F_{X_1,X_2,\ldots,X_m}(x_1, x_2, \ldots, x_m) = \int_{-\infty}^{x_m} \cdots \int_{-\infty}^{x_2} \int_{-\infty}^{x_1} f_{X_1,X_2,\ldots,X_m}(x_1, x_2, \ldots, x_m) dx_1 dx_2 \ldots dx_m} \ , \tag{48}$$

and

$$\boxed{P(B) \ = \ \int \cdots \int \int_B f_{X_1,X_2,\ldots,X_m}(x_1, x_2, \ldots, x_m) dx_1 dx_2 \ldots dx_m} \ . \tag{49}$$

The JPDF satisfies the following two properties:

1. integrates to 1, i.e., $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X_1,X_2,\ldots,X_m}(x_1, x_2, \ldots, x_m) dx_1 dx_2 \ldots dx_m = 1$
2. is a non-negative function, i.e., $f_{X_1,X_2,\ldots,X_m}(x_1, x_2, \ldots, x_m) \geq 0$.

The marginal PDF (marginal PMF) is obtained by integrating (summing) the JPDF (JPMF) over all other random variables. For example, the marginal PDF of X_1 is

$$f_{X_1}(x_1) = \int_{x_2=-\infty}^{\infty} \cdots \int_{x_m=-\infty}^{\infty} f_{X_1,X_2,\ldots,X_m}(x_1, x_2, \ldots, x_m) dx_2 \ldots dx_m$$

14.3.1 m Independent Random Variables

We say m **random variables** $X_1, X_2, \ldots, X_m$ **are jointly independent** or **mutually independent** if and only if for every $(x_1, x_2, \ldots, x_m) \in \mathbb{R}^m$

$$F_{X_1,X_2,\ldots,X_m}(x_1, x_2, \ldots, x_m) = F_{X_1}(x_1) \times F_{X_1}(x_2) \times \cdots \times F_{X_m}(x_m) \qquad \text{or}$$
$$f_{X_1,X_2,\ldots,X_m}(x_1, x_2, \ldots, x_m) = f_{X_1}(x_1) \times f_{X_1}(x_2) \times \cdots \times f_{X_m}(x_m)$$

Exercise 14.16

If X_1 and X_2 are independent random variables then what is their covariance $\mathbf{Cov}(X_1, X_2)$?

Solution We know for independent RVs from the properties of expectations that

$$E(X_1X_2) = E(X_1)E(X_2)$$

From the formula for covariance

$$\begin{aligned}\mathbf{Cov}(X_1, X_2) &= E(X_1X_2) - E(X_1)E(X_2) \\ &= E(X_1)E(X_2) - E(X_1)E(X_2) \qquad \text{due to independence} \\ &= 0\end{aligned}$$

Remark 14.17 The converse is not true: two random variables that have zero covariance are not necessarily independent.

14.3.2 Linear Combination of Independent Normal RVs is a Normal RV

We can get the following special property of normal RVs using Eqn. (40). If $X_1, X_2, \ldots, X_m$ be jointly independent RVs, where X_i is Normal(μ_i, σ_i^2), for $i = 1, 2, \ldots, m$ then $Y = c + \sum_{i=1}^m a_i X_i$ for some constants $c, a_1, a_2, \ldots, a_m$ is the Normal $\left(c + \sum_{i=1}^m a_i\mu_i, \sum_{i=1}^m a_i^2\sigma_i^2\right)$ RV.

Exercise 14.18

Let X be Normal$(2, 4)$, Y be Normal$(-1, 2)$ and Z be Normal$(0, 1)$ RVs that are jointly independent. Obtain the following:

1. $E(3X - 2Y + 4Z)$
2. $V(2Y - 3Z)$
3. the distribution of $6 - 2Z + X - Y$
4. the probability that $6 - 2Z + X - Y > 0$
5. $\mathbf{Cov}(X, W)$, where $W = X - Y$.

Solution

1. $$E(3X - 2Y + 4Z) = 3E(X) - 2E(Y) + 4(Z) = (3 \times 2) + (-2 \times (-1)) + 4 \times 0 = 6 + 2 + 0 = 8$$

2. $$V(2Y - 3Z) = 2^2V(Y) + (-3)^2V(Z) = (4 \times 2) + (9 \times 1) = 8 + 9 = 17$$

3. From the special property of normal RVs, the distribution of $6 - 2Z + X - Y$ is

$$\begin{aligned}&\text{Normal}\left(6 + (-2 \times 0) + (1 \times 2) + (-1 \times -1), ((-2)^2 \times 1) + (1^2 \times 4) + ((-1)^2 \times 2)\right)\\&= \text{Normal}\left(6 + 0 + 2 + 1, 4 + 4 + 2\right)\\&= \text{Normal}(9, 10)\end{aligned}$$

4. Let $U = 6 - 2Z + X - Y$ and we know U is Normal$(9, 10)$ RV.

$$\begin{aligned}P(6 - 2Z + X - Y > 0) = P(U > 0) = P(U - 9 > 0 - 9) &= P\left(\frac{U-9}{\sqrt{10}} > \frac{-9}{\sqrt{10}}\right)\\&= P\left(Z > \frac{-9}{\sqrt{10}}\right)\\&= P\left(Z < \frac{9}{\sqrt{10}}\right)\\&\approxeq P(Z < 2.85) = 0.9978\end{aligned}$$

5.

$$\begin{aligned}\mathbf{Cov}(X, W) &= E(XW) - E(X)E(W) = E(X(X - Y)) - E(X)E(X - Y)\\&= E(X^2 - XY) - E(X)(E(X) - E(Y)) = E(X^2) - E(XY) - 2 \times (2 - (-1))\\&= E(X^2) - E(X)E(Y) - 6 = E(X^2) - (2 \times (-1)) - 6\\&= (V(X) + (E(X))^2) + 2 - 6 = (4 + 2^2) - 4 = 4\end{aligned}$$

14.3.3 Independent Random Vectors

So far, we have treated our random vectors as random points in $\mathbb{R}^m$ and not been explicit about whether they are row or column vectors. We need to be more explicit now in order to perform arithmetic operations and transformations with them.
Let $X = (X_1, X_2, \ldots, X_{m_X})$ be a R$\vec{\text{V}}$ in $\mathbb{R}^{1 \times m_X}$, i.e., X is a random row vector with 1 row and m_X columns, with JCDF $F_{X_1, X_2, \ldots, X_{m_X}}$ and JPDF $f_{X_1, X_2, \ldots, X_{m_X}}$. Similarly, let $Y = (Y_1, Y_2, \ldots, Y_{m_Y})$ be a R$\vec{\text{V}}$ in $\mathbb{R}^{1 \times m_Y}$, i.e., Y is a random row vector with 1 row and m_Y columns, with JCDF $F_{Y_1, Y_2, \ldots, Y_{m_Y}}$ and JPDF $f_{Y_1, Y_2, \ldots, Y_{m_Y}}$. Let the JCDF of the random vectors X and Y together be $F_{X_1, X_2, \ldots, X_{m_X}, Y_1, Y_2, \ldots, Y_{m_Y}}$ and JPDF be $f_{X_1, X_2, \ldots, X_{m_X}, Y_1, Y_2, \ldots, Y_{m_Y}}$.
Two **random vectors are independent** if and only if for any $(x_1, x_2, \ldots, x_{m_X}) \in \mathbb{R}^{1 \times m_X}$ and any $(y_1, y_2, \ldots, y_{m_Y}) \in \mathbb{R}^{1 \times m_Y}$

$$\begin{aligned}&F_{X_1, X_2, \ldots, X_{m_X}, Y_1, Y_2, \ldots, Y_{m_Y}}(x_1, x_2, \ldots, x_{m_X}, y_1, y_2, \ldots, y_{m_Y})\\&\qquad = F_{X_1, X_2, \ldots, X_{m_X}}(x_1, x_2, \ldots, x_{m_X}) \times F_{Y_1, Y_2, \ldots, Y_{m_Y}}(y_1, y_2, \ldots, y_{m_Y})\end{aligned}$$

or, equivalently

$$\begin{aligned}&f_{X_1, X_2, \ldots, X_{m_X}, Y_1, Y_2, \ldots, Y_{m_Y}}(x_1, x_2, \ldots, x_{m_X}, y_1, y_2, \ldots, y_{m_Y})\\&\qquad = f_{X_1, X_2, \ldots, X_{m_X}}(x_1, x_2, \ldots, x_{m_X}) \times f_{Y_1, Y_2, \ldots, Y_{m_Y}}(y_1, y_2, \ldots, y_{m_Y})\end{aligned}$$

The notion of mutual independence or joint independence of n random vectors is obtained similarly from ensuring the independence of any subset of the n vectors in terms of their JCDFs (JPMFs or JPDFs) being equal to the product of their marginal CDFs (PMFs or PDFs).

Some Important Multivariate Random Vectors

Let us consider the natural two-dimensional analogue of the Bernoulli(θ) RV in the real plane $\mathbb{R}^2 := (-\infty, \infty)^2 := (-\infty, \infty) \times (-\infty, \infty)$. A natural possibility is to use the **ortho-normal basis vectors** in $\mathbb{R}^2$:

$$\boxed{e_1 := (1, 0), \qquad e_2 := (0, 1)}\ .$$

Recall that vector addition and subtraction are done component-wise, i.e. $(x_1, x_2) \pm (y_1, y_2) = (x_1 \pm y_1, x_2 \pm y_2)$. We introduce a useful function called the indicator function of a set, say A.

$$\mathbb{1}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise.} \end{cases}$$

$\mathbb{1}_A(x)$ returns 1 if x belongs to A and 0 otherwise.

Exercise 14.19

Let us recall the geometry and arithmetic of vector addition in the plane.

1. What is $(1, 0) + (1, 0)$, $(1, 0) + (0, 1)$, $(0, 1) + (0, 1)$?
2. What is the relationship between $(1, 0)$, $(0, 1)$ and $(1, 1)$ geometrically?
3. How does the diagonal of the parallelogram relate the its two sides in the geometry of addition in the plane?
4. What is $(1, 0) + (0, 1) + (1, 0)$?

Solution

1. addition is component-wise

$$\begin{aligned} (1, 0) + (1, 0) &= (1 + 1, 0 + 0) = (2, 0) \\ (1, 0) + (0, 1) &= (1 + 0, 0 + 1) = (1, 1) \\ (0, 1) + (0, 1) &= (0 + 0, 1 + 1) = (0, 2) \end{aligned}$$

2. $(1, 0)$ and $(0, 1)$ are vectors for the two sides of unit square and $(1, 1)$ is its diagonal.

3. Generally, the diagonal of the parallelogram is the resultant or sum of the vectors representing its two sides

4. $$(1,0)+(0,1)+(1,0)=(1+0+1,0+1+0)=(2,1)$$

Definition 14.20 [**Bernoulli**(θ) **R$\vec{\text{V}}$**] Given a parameter $\theta \in [0,1]$, we say that $X := (X_1, X_2)$ is a Bernoulli(θ) random vector (R$\vec{\text{V}}$) if it has only two possible outcomes in the set $\{e_1, e_2\} \subset \mathbb{R}^2$, i.e. $x := (x_1, x_2) \in \{(1,0),(0,1)\}$. The PMF of the R$\vec{\text{V}}$ $X := (X_1, X_2)$ with realization $x := (x_1, x_2)$ is:

$$f(x;\theta) := P(X = x) = \theta\, \mathbb{1}_{\{e_1\}}(x) + (1-\theta)\, \mathbb{1}_{\{e_2\}}(x) = \begin{cases} \theta & \text{if} \quad x = e_1 := (1,0) \\ 1-\theta & \text{if} \quad x = e_2 := (0,1) \\ 0 & \text{otherwise} \end{cases}$$

Exercise 14.21

What is the Expectation of Bernoulli(θ) R$\vec{\text{V}}$?

Solution

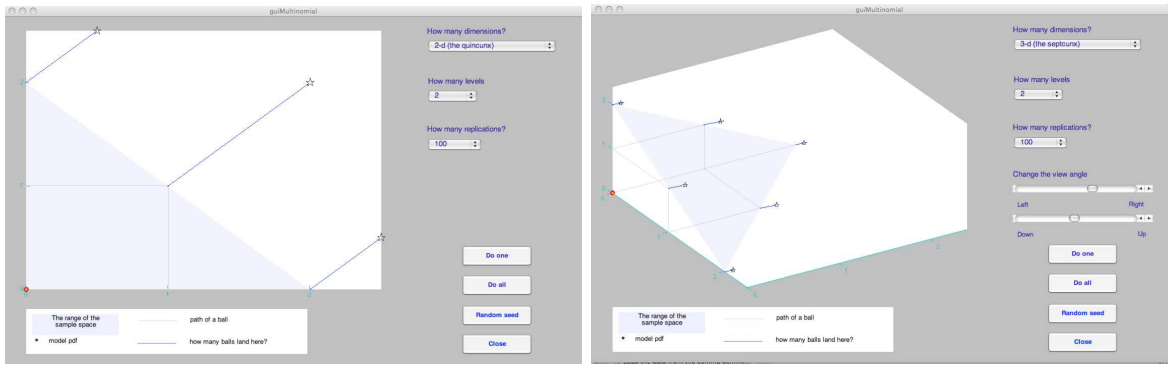
$$E_\theta(X) = E_\theta((X_1, X_2)) = \sum_{(x_1,x_2)\in\{e_1,e_2\}} (x_1, x_2) f((x_1, x_2);\theta) = (1,0)\theta + (0,1)(1-\theta) = (\theta, 1-\theta)\ .$$

We can write the Binomial(n,θ) RV Y as a Binomial(n,θ) R$\vec{\text{V}}$ $X := (Y, n-Y)$. In fact, this is the underlying model and the **bi** in the Binomial(n,θ) does refer to two in Latin. In the coin-tossing context this can be thought of keeping track of the number of Heads and Tails out of an IID sequence of n tosses of a coin with probability θ of observing Heads. In the Quincunx context, this amounts to keeping track of the number of right and left turns made by the ball as it drops through n levels of pegs where the probability of a right turn at each peg is independently and identically θ. In other words, the Binomial(n,θ) R$\vec{\text{V}}$ $(Y, n-Y)$ is the sum of n IID Bernoulli(θ) R$\vec{\text{V}}$s $X_1 := (X_{1,1}, X_{1,2}), X_2 := (X_{2,1}, X_{2,2}), \ldots, X_n := (X_{n,1}, X_{n,2})$:

$$(Y, n-Y) = X_1 + X_2 + \cdots + X_n = (X_{1,1}, X_{1,2}) + (X_{2,1}, X_{2,2}) + \cdots + (X_{n,1}, X_{n,2})$$

Labwork 14.22 (Quincunx Sampler Demo – Sum of n IID Bernoulli(1/2) **R$\vec{\text{V}}$s)** Let us understand the Quincunx construction of the Binomial(n, 1/2) R$\vec{\text{V}}X$ as the sum of n independent and identical Bernoulli(1/2) R$\vec{\text{V}}$s by calling the interactive visual cognitive tool as follows:

Figure 4: Visual Cognitive Tool GUI: Quincunx & Septcunx.




```
>> guiMultinomial
```

We are now ready to extend the Binomial(n, θ) RV or R$\vec{\text{V}}$ to its multivariate version called the Multinomial$(n, \theta_1, \theta_2, \ldots, \theta_k)$ R$\vec{\text{V}}$. We develop this R$\vec{\text{V}}$ as the sum of n IID de Moivre$(\theta_1, \theta_2, \ldots, \theta_k)$ R$\vec{\text{V}}$ that is defined next.

Definition 14.23 [**de Moivre**$(\theta_1, \theta_2, \ldots, \theta_k)$ **R$\vec{\text{V}}$**] The PMF of the de Moivre$(\theta_1, \theta_2, \ldots, \theta_k)$ R$\vec{\text{V}}$ $X := (X_1, X_2, \ldots, X_k)$ taking value $x := (x_1, x_2, \ldots, x_k) \in \{e_1, e_2, \ldots, e_k\}$, where the e_i's are ortho-normal basis vectors in $\mathbb{R}^k$ is:

$$f(x; \theta_1, \theta_2, \ldots, \theta_k) := P(X = x) = \sum_{i=1}^{k} \theta_i \mathbb{1}_{\{e_i\}}(x) = \begin{cases} \theta_1 & \text{if} \quad x = e_1 := (1, 0, \ldots, 0) \in \mathbb{R}^k \\ \theta_2 & \text{if} \quad x = e_2 := (0, 1, \ldots, 0) \in \mathbb{R}^k \\ \vdots & \\ \theta_k & \text{if} \quad x = e_k := (0, 0, \ldots, 1) \in \mathbb{R}^k \\ 0 & \text{otherwise} \end{cases}$$

Of course, $\sum_{i=1}^{k} \theta_i = 1$.

When we add n IID de Moivre$(\theta_1, \theta_2, \ldots, \theta_k)$ R$\vec{\text{V}}$ together, we get the Multinomial$(n, \theta_1, \theta_2, \ldots, \theta_k)$ R$\vec{\text{V}}$ as defined below.

Definition 14.24 [**Multinomial**$(n, \theta_1, \theta_2, \ldots, \theta_k)$ **R$\vec{\text{V}}$**] We say that a R$\vec{\text{V}}$ $Y := (Y_1, Y_2, \ldots, Y_k)$ obtained from the sum of n IID de Moivre$(\theta_1, \theta_2, \ldots, \theta_k)$ R$\vec{\text{V}}$s with realizations

$$y := (y_1, y_2, \ldots, y_k) \in \mathbb{Y} := \{(y_1, y_2, \ldots, y_k) \in \mathbb{Z}_+^k : \sum_{i=1}^{k} y_i = n\}$$

has the PMF given by:

$$f(y; n, \theta) := f(y; n, \theta_1, \theta_2, \ldots, \theta_k) := P(Y = y; n, \theta_1, \theta_2, \ldots, \theta_k) = \binom{n}{y_1, y_2, \ldots, y_k} \prod_{i=1}^{k} \theta_i^{y_i} \ ,$$

where, the multinomial coefficient:

$$\binom{n}{y_1, y_2, \ldots, y_k} := \frac{n!}{y_1! y_2! \cdots y_k!} \ .$$

Note that the marginal PMF of Y_j is Binomial(n, θ_j) for any $j = 1, 2, \ldots, k$.

We can visualize the Multinomial$(n, \theta_1, \theta_2, \theta_3)$ process as a sum of n IID de Moivre$(\theta_1, \theta_2, \theta_3)$ R$\vec{\text{V}}$s via a three dimensional extension of the Quincunx called the "Septcunx" and relate the number of paths that lead to a given trivariate sum (y_1, y_2, y_3) with $\sum_{i-1}^{3} y_i = n$ as the multinomial coefficient $\frac{n!}{y_1!y_2!y_3!}$. In the Septcunx, balls choose from one of three paths along e_1, e_2 and e_3 with probabilities θ_1, θ_2 and θ_3, respectively, in an IID manner at each of the n levels, before they collect at buckets placed at the integral points in the 3-simplex, $\mathbb{Y} = \{(y_1, y_2, y_3) \in \mathbb{Z}_+^3 : \sum_{i=1}^{3} y_i = n\}$. Once again, we can visualize that the sum of n IID de Moivre$(\theta_1, \theta_2, \theta_3)$ R$\vec{\text{V}}$s constitute the Multinomial$(n, \theta_1, \theta_2, \theta_3)$ R$\vec{\text{V}}$.

Labwork 14.25 (Septcunx Sampler Demo – Sum of n IID de Moivre$(1/3, 1/3, 1/3)$ **R$\vec{\text{V}}$s)** Let us understand the Septcunx construction of the Multinomial$(n, 1/3, 1/3, 1/3)$ R$\vec{\text{V}}X$ as the sum of n independent and identical de Moivre$(1/3, 1/3, 13/)$ R$\vec{\text{V}}$s by calling the interactive visual cognitive tool as follows:

```
>> guiMultinomial
```

Multinomial distributions are at the very foundations of various machine learning algorithms, including, filtering junk email, learning from large knowledge-based resources like www, Wikipedia, word-net, etc.

Definition 14.26 [**Normal**(μ, Σ) **R$\vec{\text{V}}$**] The univariate Normal(μ, σ^2) RV has two parameters, $\mu \in \mathbb{R}$ and $\sigma^2 \in (0, \infty)$. In the multivariate version $\mu \in \mathbb{R}^{m \times 1}$ is a column vector and σ^2 is replaced by a matrix Σ. To begin, let

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_m \end{pmatrix}$$

where, $Z_1, Z_2, \ldots, Z_m$ are jointly independent Normal$(0, 1)$ RVs. Then the JPDF of Z is

$$f_Z(z) = f_{Z_1, Z_2, \ldots, Z_m}(z_1, z_2, \ldots, z_m) = \frac{1}{(2\pi)^{m/2}} \exp\left(-\frac{1}{2}\sum_{j=1}^{m} z_j^2\right) = \frac{1}{(2\pi)^{m/2}} \exp\left(-\frac{1}{2} z^T z\right)$$

We say that Z has a standard multivariate normal distribution and write $Z \sim$ Normal$(0, I)$, where it is understood that 0 represents the vector of m zeros and I is the $m \times m$ identity matrix (with 1 along the diagonal entries and 0 on all off-diagonal entries).
More generally, a vector X has a multivariate normal distribution denoted by $X \sim$ Normal(μ, Σ), if it has joint probability density function

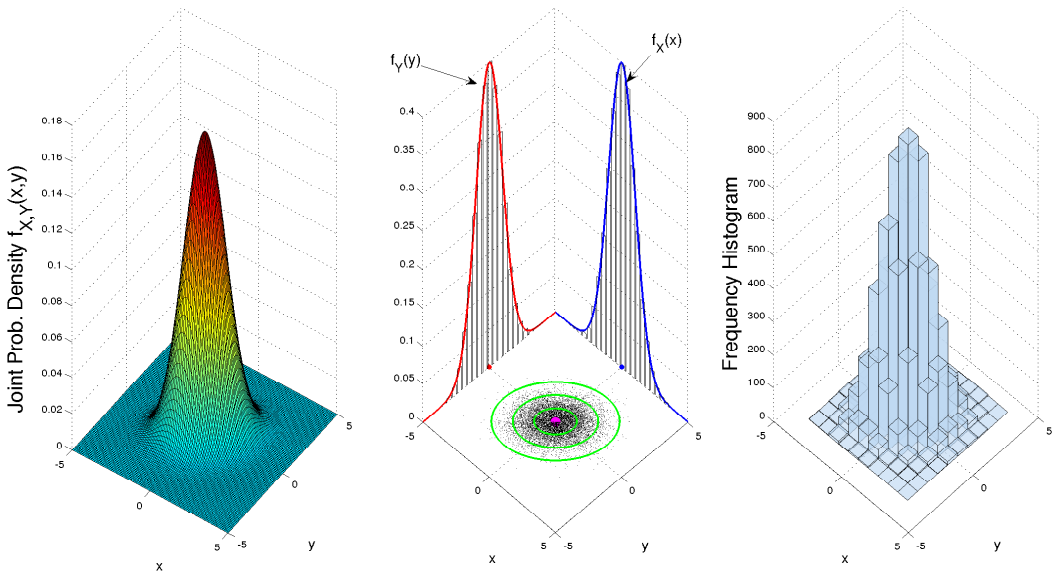
$$f_X(x; \mu, \Sigma) = f_{X_1, X_2, \ldots, X_m}(x_1, x_2, \ldots, x_m; \mu, \Sigma) = \frac{1}{(2\pi)^{m/2} |(\Sigma)|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right)$$

where $|\Sigma|$ denotes the determinant of Σ, μ is a vector of length m and Σ is a $m \times m$ symmetric, positive definite matrix. Setting $\mu = 0$ and $\Sigma = I$ gives back the standard multivariate normal R$\vec{\text{V}}$.

When we have a non-zero mean vector

$$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} = \begin{pmatrix} 6.49 \\ 5.07 \end{pmatrix}$$

Figure 5: JPDF, Marginal PDFs and Frequency Histogram of Bivariate Standard Normal R$\vec{\text{V}}$.



for the mean lengths and girths of cylindrical shafts from a manufacturing process with variance-covariance matrix

$$\Sigma = \begin{pmatrix} \mathbf{Cov}(X,X) & \mathbf{Cov}(X,Y) \\ \mathbf{Cov}(Y,X) & \mathbf{Cov}(Y,Y) \end{pmatrix} = \begin{pmatrix} V(X) & \mathbf{Cov}(X,Y) \\ \mathbf{Cov}(X,Y) & V(Y) \end{pmatrix} = \begin{pmatrix} 0.59 & 0.24 \\ 0.24 & 0.26 \end{pmatrix}$$

then the Normal(μ, Σ) R$\vec{\text{V}}$ has JPDF, marginal PDFs and samples with frequency histograms as shown in Figure 6.
We can use MATLAB to compute for instance the probability that a cylinder has length and girth below 6.0 cms as follows:

```
>> mvncdf([6.0 6.0],[6.49 5.07],[0.59  0.24; 0.24 0.26])
ans =    0.2615
```

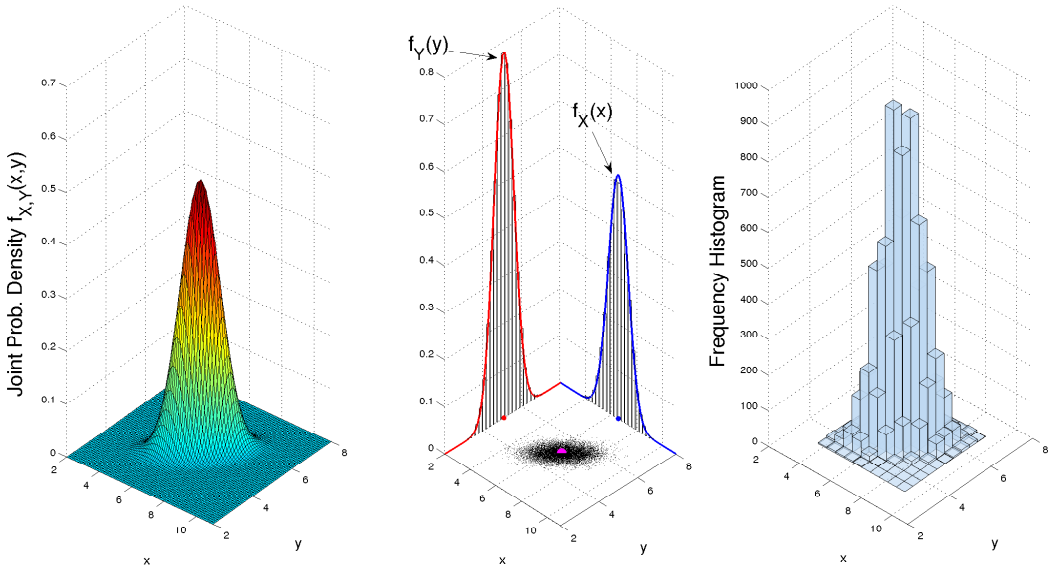
Or find the probability (with numerical error tolerance) that the cylinders are within the rectangular specifications of 6 ± 1.0 along x and y as follows:

```
>> [F err] = mvncdf([5.0 5.0], [7.0 7.0], [6.49 5.07],[0.59  0.24; 0.24 0.26])

F =    0.3352

err =   1.0000e-08
```

Figure 6: JPDF, Marginal PDFs and Frequency Histogram of a Bivariate Normal R$\vec{V}$ for lengths of girths of cylindrical shafts in a manufacturing process (in cm).



15 Conditional Densities and Conditional Expectations

Take notes in class!

16 From Observations to Laws – Linking Data to Models

What is the Engineering Method?

An Engineer solves important problems in society by applying scientific principles. Problems are solved by refining existing products and processes or designing novel ones. The *engineering method* is the approach to formulating and solving such problems. The block diagram shows the basic steps in the method.

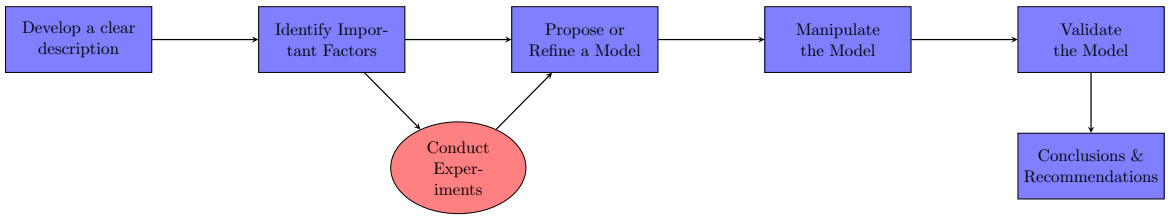


Figure 7: The Engineering Method (adapted from Montgomery & Runger, 2007)

A crucial aspect of the method is the interplay between the *mathematical model* from our minds and the *observed data* from experiments in the real-world to verify and improve the model. This crucial link between model and data, in the engineering method, is provided by probabilistic and statistical reasoning which allows us to handle randomness inherent in real-world measurements and go from the observation to the underlying law specified by the "true" parameter $\theta^* \in \Theta$, a parametric family of probability models. Such models could include independent and identically distributed observations from a family of RVs or $\vec{\text{RV}}$s, "noisy" ODEs, PDEs, etc.

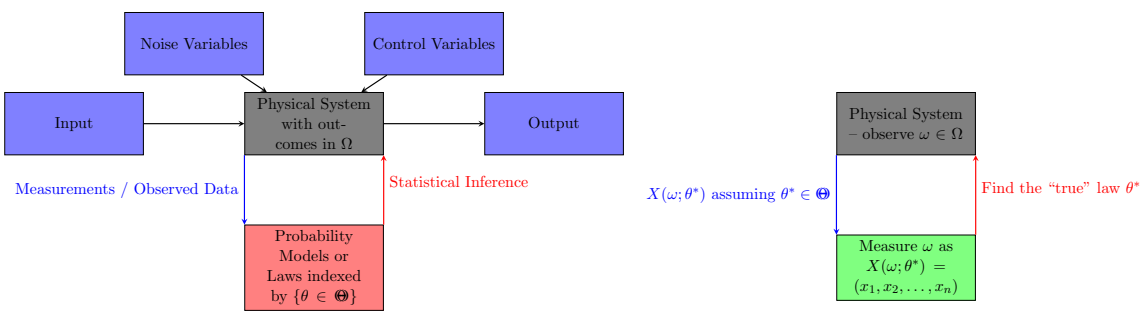


Figure 8: *Statistical inference* allows us to use observations to refine the laws governing them (specified by models)

16.1 Data and Statistics

Definition 16.1 (Data) The function X measures the outcome ω of an experiment with sample space Ω. Formally, X is a random variable or a random vector $X = (X_1, X_2, \ldots, X_n)$ taking values in the **data space** $\mathbb{X}$:

$$X(\omega) : \Omega \to \mathbb{X} \ .$$

The realization of the RV X when an experiment is performed is the observation or data $x \in \mathbb{X}$. That is, when the experiment is performed once and it yields a specific $\omega \in \Omega$, the data $X(\omega) = x \in \mathbb{X}$ is the corresponding realization of the RV X.

Figure 9: Sample Space, Random Variable, Realization, Data, and Data Space (copy figure from lectures).

Engineers have three basic settings for collecting data:

- A *retrospective study* using historical data to understand past system behavior
- An *observational study* to understand the system behavior without interference
- A *designed experiment* where deliberate changes are made to observe response.

With regard to mathematical models, there are fundamentally two types of models:

- **mechanistic models** that are built from our underlying knowledge of the laws governing the phenomenon of interest ("white box") and
- **descriptive models** that are derived from fitting the data on a purely descriptive basis using a flexible family of models ("black box").

The simplest way to collect data is the so-called **simple random sequence (SRS)**.

$$X_1, X_2, \ldots, X_n$$

An SRS is usually assumed to be a collection of n independent and identically distributed (IID) random variables from an n-product experiment (an experiment repeated in identical conditions n times). A particular realization of the SRS results in the observations:

$$x_1, x_2, \ldots, x_n$$

Example 16.2 Let us measure the current for a fixed input voltage through a given circuit with a simple resistor repeatedly many times throughout the day. Then we can get variations in these measurements due to uncontrollable factors including, changes in ambient temperature, impurities present in different locations along the wire, etc.
These variations in current can be modeled using our mechanistic understanding of the relationship between V, I and R:

$$I = v/r + \epsilon$$

where $\epsilon \sim \text{Normal}(0, \sigma^2)$ random variable. Assuming v/r is constant, we can think of the measurements of I as a SRS from a $\text{Normal}(v/r, \sigma^2)$ random variable.
For example, if applied voltage is 6.0 volts and resistance is 2 Ohms (with $\sigma^2 \approxeq 0.05$), then an SRS of current measurements in Amps is:

$$3.03, 3.09, 2.89, 3.04, 3.02, 2.94$$

And another SRS form this experiment is:

$$2.99, 3.02, 3.18, 3.14, 2.93, 3.15$$

Now, let us change the applied voltage to 3 volts but keep the same resistor with 2 Ohms, then an SRS of current measurements for this new voltage setting in Amps is:

$$1.4783, 1.5171, 1.6789, 1.6385, 1.4325, 1.6517$$

and another SRS for the same setting is:

$$1.5363, 1.4968, 1.5357, 1.4898, 1.4938, 1.5745$$

This is an example of a designed experiment because we can control the input voltage and resistance in the system and measure the response of the system to controllable factors (voltage and resistance).

We usually use descriptive models when we do not fully know the underlying laws governing the phenomenon and these are also called "black-box" models.

Example 16.3 Let us look a **descriptive model** obtained by fitting a straight line to CO_2 concentration and year. The "best fitted line" has slope of 1.46 and intersects $x = 1958$ at 307.96 (Figure 10). The data comes from a long-term data set gathered by US NOAA (like NZ's NIWA). The following three descriptive models in our **retrospective study** tells us that CO_2 concentration is increasing with time (in years). We have fit a polynomial (of order $k = 0, 1, 2$) to this data assuming Normally distributed errors. Let's see the Sage Worksheet now, shall we?

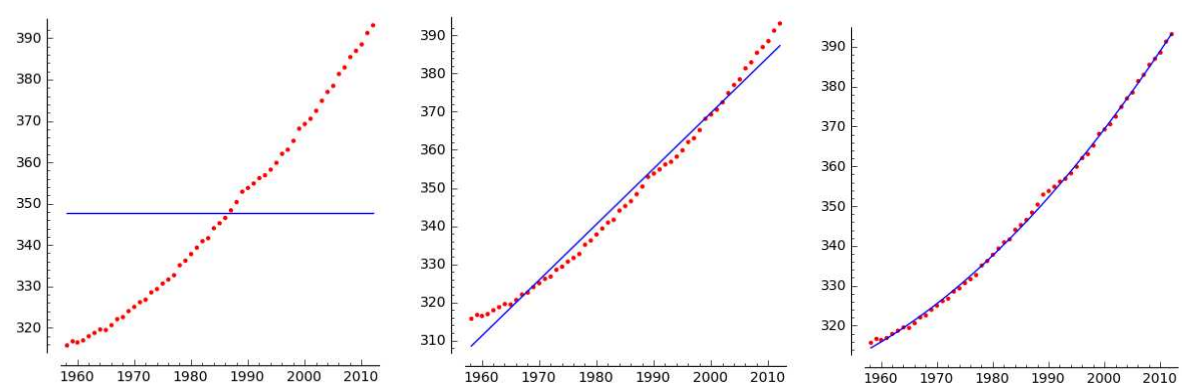


Figure 10: Fitting a flat line ($\hat{y} = 347.8$), straight line ($\hat{y} = 1.46x - 2554.9$, with slope = 1.46 ppm/year) and quadratic curve ($\hat{y} = 0.0125x^2 - 48.22x + 46757.73$) to CO_2 concentration (ppm) and year.

Example 16.4 The Coin-Toss Experiment of Persi Diaconis and his collaborators at Stanford's Statistics Department involved an electro-magenetically released iron ruler to flip an American quarter from its end in an independent and identical manner. Here we had a mechanistic ODE model, data from cameras of the flipping coin, and the descriptive Bernoulli(θ) RV with $\theta \in \Theta = [0, 1]$.

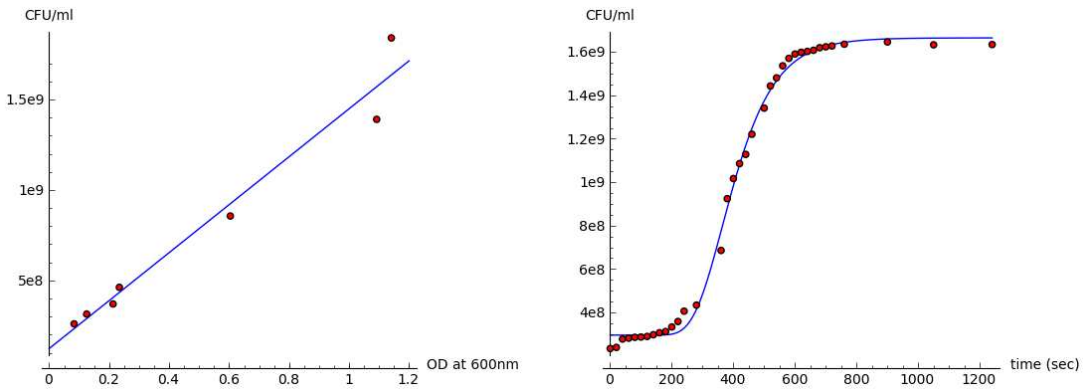


Figure 11: Descriptive model for Colony Forming Units (CFU) as a function of Optical Density (OD) is given by a best fitting line with intercept $= 1.237283e + 08$ and slope $= 1.324714e + 09$ (left subfigure). Mechanistic model for CFU as a function of time is given by the trajectory $y(t; k, l, y_0, g)$ of an ODE model with the best fitted parameters ($\widehat{y_0} = 3.051120e + 08, \widehat{k} = 1.563822e + 09, \widehat{l} = 2.950639e + 02, \widehat{g} = 6.606311e + 06$).

Most Problems are Black & White

Often, one needs to use both descriptive and mechanistic model in the same experiment. Let us consider a situation where a chemical process engineer wants to understand the population growth of a bacteria in a fermentation tank through time. However, measuring the exact number of bacteria in the tank through time is no easy task! Microbiologists use a measure called colony forming units per ml to count bacterial or fungal numbers. A much simpler measure is the optical density of the liquid in the tank (the liquid is more murky as more bacteria live in it). Optical density can be measured using a photo-spectrometer (basically measure how much light can pass through the liquid in a test-tube). OD is correlated to the concentration of bacteria in the liquid. To quantify this correlation, the engineer measured the optical density of a number of calibration standards of known concentration. She then fits a "standard curve" to the calibration data and uses it to determine the concentration of bacteria from OD readings alone.
For example, we can use ordinary differential equations to model the growth of bacteria in a fermentation chamber using the ecological laws of population dynamics in a limiting environment. This ODE-based model called *Gompertz's Growth Model* has four parameters that determine the trajectory of the population size through time:

$$y(t; k, l, y_0, g) = (k - y_0) \exp\left(-\exp\left(\frac{(l-t)ge}{k - y_0} + 1\right)\right) + y_0 + \epsilon$$

where, $\epsilon \sim \text{Normal}(0, \sigma)$ is Gaussian or white noise that we add to make the model stochastic (probabilistic). Let us do a break-down of this example step-by-step using the interactive sage worksheet with the following URL: `http://sage.math.canterbury.ac.nz/home/pub/265/`.

Definition 16.5 (Statistic) A **statistic** T is any function of the data:

$$T(x) : \mathbb{X} \to \mathbb{T} \ .$$

Thus, a statistic T is also an RV that takes values in the space $\mathbb{T}$. When $x \in \mathbb{X}$ is the realization of an experiment, we let $T(x) = t$ denote the corresponding realization of the statistic T.

Sometimes we use $T_n(X)$ and $\mathbb{T}_n$ to emphasize that X is an n-dimensional random vector, i.e. $\mathbb{X} \subset \mathbb{R}^n$.

Next, we define two important statistics called the **sample mean** and **sample variance**. Since they are obtained from the sample data, they are called **sample moments**, as opposed to the **population moments**. The corresponding population moments are $E(X_1)$ and $V(X_1)$, respectively.

Definition 16.6 (Sample Mean) From a given a sequence of RVs $X_1, X_2, \ldots, X_n$, we may obtain another RV called the n-samples mean or simply the sample mean:

$$T_n(\ (X_1, X_2, \ldots, X_n)\ ) = \overline{X}_n(\ (X_1, X_2, \ldots, X_n)\ ) := \frac{1}{n} \sum_{i=1}^{n} X_i\ . \tag{50}$$

We write

$$\overline{X}_n(\ (X_1, X_2, \ldots, X_n)\ ) \quad \text{as} \quad \overline{X}_n\ ,$$

and its realization

$$\overline{X}_n(\ (x_1, x_2, \ldots, x_n)\ ) \quad \text{as} \quad \overline{x}_n\ .$$

Note that the expectation of $\overline{X}_n$ is:

$$\begin{aligned} E(\overline{X}_n) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} E\left(X_i\right) \end{aligned}$$

Furthermore, if every X_i in the original sequence of RVs $X_1, X_2, \ldots$ is **identically** distributed with the same expectation, by convention $E(X_1)$, then:

$$E(\overline{X}_n) = \frac{1}{n}\sum_{i=1}^{n} E\left(X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(X_1\right) = \frac{1}{n}\ n\ E\left(X_1\right) = E\left(X_1\right)\ . \tag{51}$$

Similarly, we can show that the variance of $\overline{X}_n$ is:

$$\begin{aligned} V(\overline{X}_n) &= V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \left(\frac{1}{n}\right)^2 V\left(\sum_{i=1}^{n} X_i\right) \end{aligned}$$

Furthermore, if the original sequence of RVs $X_1, X_2, \ldots$ is **independently** distributed then:

$$V(\overline{X}_n) = \left(\frac{1}{n}\right)^2 V\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\ \sum_{i=1}^{n} V\left(X_i\right)$$

Finally, if the original sequence of RVs $X_1, X_2, \ldots$ is **independently and identically** distributed with the same variance ($V(X_1)$ by convention) then:

$$V(\overline{X}_n) = \frac{1}{n^2}\ \sum_{i=1}^{n} V\left(X_i\right) = \frac{1}{n^2}\ \sum_{i=1}^{n} V\left(X_1\right) = \frac{1}{n^2}\ n\ V\left(X_1\right) = \frac{1}{n}\ V\left(X_1\right)\ . \tag{52}$$

Remark 16.7 Note that the variance of the sample mean decreases with sample size!

Exercise 16.8
Let us compute the sample mean of the current measurements for each of the two samples of size $n = 6$ from Example 16.2 under an applied voltage of 6.0 volts and resistance of 2 Ohms:

$$3.03, 3.09, 2.89, 3.04, 3.02, 2.94 \quad \text{and} \quad 2.99, 3.02, 3.18, 3.14, 2.93, 3.15$$

Solution For the first SRS

$$\begin{aligned}\overline{x}_6 &= \frac{1}{6}(3.03 + 3.09 + 2.89 + 3.04 + 3.02 + 2.94) = \frac{1}{6} \times \frac{1}{100}(303 + 309 + 289 + 304 + 302 + 294) \\ &= \frac{1}{6} \times \frac{1}{100}(1600 + 170 + 31) = \frac{1}{6} \times \frac{1801}{100} = \frac{1}{6} \times 18.01 \\ &= 3.0017 \ .\end{aligned}$$

Similarly, for the second SRS

$$\overline{x}_6 = \frac{1}{6}(2.99 + 3.02 + 3.18 + 3.14 + 2.93 + 3.15) = 3.0683 \ .$$

We can use `mean` in MATLAB to obtain the sample mean $\overline{x}_n$ of n sample points $x_1, x_2, \ldots, x_n$ as follows:

```
>> data_n6_1 = [3.03 3.09 2.89 3.04 3.02 2.94];
>> data_n6_2 = [2.99 3.02 3.18 3.14 2.93 3.15];
>> [ mean(data_n6_1)   mean(data_n6_2) ]

ans =    3.0017    3.0683
```

Note how close the two sample means are to what we expect under Ohm's law: $I = v/r = 6/2 = 3$. If however, we only used samples of size $n = 2$ then we notice a bigger variance or spread in the six realizations of $\overline{X}_2$ in agreement with Remark 16.7:

```
>> data_n2_1 = [3.03 3.09]; data_n2_2 = [2.89 3.04]; data_n2_3 = [3.02 2.94];
>> data_n2_4 = [2.99 3.02]; data_n2_5 = [3.18 3.14]; data_n2_6 = [2.93 3.15];
>> [mean(data_n2_1) mean(data_n2_2) mean(data_n2_3) mean(data_n2_4) ...
                                     mean(data_n2_5) mean(data_n2_6)]

ans =    3.0600    2.9650    2.9800    3.0050    3.1600    3.0400
```

Definition 16.9 (Sample Variance & Standard Deviation) From a given a sequence of random variables $X_1, X_2, \ldots, X_n$, we may obtain another statistic called the **n-samples variance** or simply the **sample variance** :

$$T_n(\ (X_1, X_2, \ldots, X_n)\ ) = S_n^2(\ (X_1, X_2, \ldots, X_n)\ ) := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \ . \tag{53}$$

We write $S_n^2(\ (X_1, X_2, \ldots, X_n)\ )$ as S_n^2 and its realization $S_n^2(\ (x_1, x_2, \ldots, x_n)\ )$ as s_n^2.
Sample standard deviation is simply the square root of sample variance:

$$S_n(\ (X_1, X_2, \ldots, X_n)\ ) = \sqrt{S_n^2(\ (X_1, X_2, \ldots, X_n)\ )} \tag{54}$$

We write $S_n(\ (X_1, X_2, \ldots, X_n)\ )$ as S_n and its realization $S_n(\ (x_1, x_2, \ldots, x_n)\ )$ as s_n.

Once again, if $X_1, X_2, \ldots, X_n \overset{\text{IID}}{\sim} X_1$, i.e., all n RVs are independently and identically distributed, then the expectation of the sample variance is:

$$E(S_n^2) = V(X_1) \ .$$

Exercise 16.10

Let us compute the sample variance and sample standard deviation of the current measurements for the first sample of size $n = 6$ from Example 16.2: $3.03, 3.09, 2.89, 3.04, 3.02, 2.94$.

Solution From Exercise 16.8 we know $\overline{x}_6 = 3.0017$.

$$\begin{aligned}
s_6^2 &= \frac{1}{6-1}\left((3.03-3.0017)^2 + (3.09-3.0017)^2 + (2.89-3.0017)^2 + \right. \\
&= \qquad \left. (3.04-3.0017)^2 + (3.02-3.0017)^2 + (2.94-3.0017)^2\right) \\
&= \frac{1}{5}\left(0.0283^2 + 0.0883^2 + (-0.1117)^2 + 0.0383^2 + 0.0183^2 + (-0.0617)^2\right) \\
&\approxeq \frac{1}{5}\left(0.0008 + 0.0078 + 0.0125 + 0.0015 + 0.0003 + 0.0038\right) \\
&= 0.00534 \ .
\end{aligned}$$

And sample standard deviation is

$$s_6 = \sqrt{s_6^2} = \sqrt{0.00534} \approxeq 0.0731 \ .$$

We can compute the sample variance and sample standard deviation for both samples of size $n = 6$ using MATLAB 's functions `var` and `std`, respectively.

```
>> [ var(data_n6_1)   std(data_n6_1) ]
ans =    0.0053    0.0731

>> [ var(data_n6_2)   std(data_n6_2) ]
ans =    0.0104    0.1019
```

It is important to bear in mind that the statistics such as sample mean and sample variance are random variables and have an underlying distribution.

Definition 16.11 (Order Statistics) Suppose $X_1, X_2, \ldots, X_n \overset{\text{IID}}{\sim} F$, where F is the DF from the set of all DFs over the real line. Then, the n-sample **order statistics** $X_{([n])}$ is:

$$X_{([n])}(\ (X_1, X_2, \ldots, X_n)\ ) := \left(X_{(1)}, X_{(2)}, \ldots X_{(n)}\right), \text{ such that, } X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)} \ . \quad (55)$$

We write $X_{([n])}(\ (X_1, X_2, \ldots, X_n)\ )$ as $X_{([n])}$ and its realization $X_{([n])}(\ (x_1, x_2, \ldots, x_n)\ )$ as $x_{([n])} = (x_{(1)}, x_{(2)}, \ldots x_{(n)})$.

Exercise 16.12

Let us compute the order statistics of the current measurements for the first sample of size $n = 6$ from Example 16.2: $3.03, 3.09, 2.89, 3.04, 3.02, 2.94$.

Solution

$$x_{[6]} = \left(x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)}, x_{(5)}, x_{(6)}\right) = (2.89, 2.94, 3.02, 3.03, 3.04, 3.09) \ .$$

Without going into the details of how to sort the data in ascending order to obtain the order statistics (an elementary topic of an Introductory Computer Science course), we can use the `sort` function in MATLAB to obtain the order statistics for both samples of size $n = 6$ as follows:

```
>> sort(data_n6_1)
ans =    2.8900    2.9400    3.0200    3.0300    3.0400    3.0900

>> sort(data_n6_2)
ans =    2.9300    2.9900    3.0200    3.1400    3.1500    3.1800
```

Definition 16.13 (Empirical Distribution Function (EDF or ECDF)) Suppose we have n IID RVs, $X_1, X_2, \ldots, X_n \overset{\text{IID}}{\sim} F$, i.e., the n RVs are independently and identically distributed according to the distribution function F. Then, the n-sample empirical distribution function (EDF or ECDF) is the discrete distribution function $\widehat{F}_n$ that puts a probability mass of $1/n$ at each sample or data point x_i:

$$\widehat{F}_n(x) = \frac{\#\text{ of data points} \leq x}{n} = \frac{\sum_{i=1}^{n} \mathbb{1}(X_i \leq x)}{n}, \quad \text{where } \mathbb{1}(X_i \leq x) = \begin{cases} 1 & \text{if } X_i \leq x \\ 0 & \text{if } X_i > x \end{cases} \tag{56}$$

Roughly speaking, the EDF or ECDF is a staircase that jumps by $\frac{1}{n}$ at each data point.

Exercise 16.14

Plot the empirical distribution function of current measurements for the first sample of size $n = 6$ from Example 16.2:

$$3.03, 3.09, 2.89, 3.04, 3.02, 2.94$$

What is $\widehat{F}_n(2.91)$?

Solution (draw in class)

$\widehat{F}_n(2.91) = \frac{\#\text{ data points} \leq 2.9}{6} = \frac{1}{6}$

A **histogram** is a graphical representation of the frequency with which elements of a data array:

$$x = (x_1, x_2, \ldots, x_n) \ ,$$

of real numbers fall within each of the m intervals or **bins** of some **interval partition**:

$$b := (b_1, b_2, \ldots, b_m) := ([\underline{b}_1, \overline{b}_1), [\underline{b}_2, \overline{b}_2), \ldots, [\underline{b}_m, \overline{b}_m])$$

of the **data range** of x given by the closed interval:

$$\mathcal{R}(x) := [\min\{x_1, x_2, \ldots, x_n\}, \max\{x_1, x_2, \ldots, x_n\}] \ .$$

Elements of this partition b are called bins, their mid-points are called **bin centers**:

$$c := (c_1, c_2, \ldots, c_m) := ((\underline{b}_1 + \overline{b}_1)/2, (\underline{b}_2 + \overline{b}_2)/2, \ldots, (\underline{b}_m + \overline{b}_m)/2)$$

and their overlapping boundaries, i.e. $\overline{b}_i = \underline{b}_{i+1}$ for $1 \leq i < m$, are called **bin edges**:

$$d := (d_1, d_2, \ldots, d_{m+1}) := (\underline{b}_1, \underline{b}_2, \ldots, \underline{b}_{m-1}, \underline{b}_m, \overline{b}_m) \ .$$

For a given partition of the data range $\mathcal{R}(x)$ or some superset of $\mathcal{R}(x)$, three types of histograms are possible: frequency histogram, relative frequency histogram and density histogram. Typically, the partition b is assumed to be composed of m overlapping intervals of the same width $w = \overline{b}_i - \underline{b}_i$ for all $i = 1, 2, \ldots, m$. Thus, a histogram can be obtained by a set of bins along with their corresponding **heights**:

$$h = (h_1, h_2, \ldots, h_m) \ , \ \text{where } h_k := g(\#\{x_i : x_i \in b_k\})$$

Thus, h_k, the height of the k-th bin, is some function g of the number of data points that fall in the bin b_k Formally, a histogram is a sequence of ordered pairs:

$$((b_1, h_1), (b_2, h_2), \ldots, (b_m, h_m)) \ .$$

Given a partition b, a **frequency histogram** is the histogram:

$$((b_1, h_1), (b_2, h_2), \ldots, (b_m, h_m)) \ , \ \text{where } h_k := \#\{x_i : x_i \in b_k\} \ ,$$

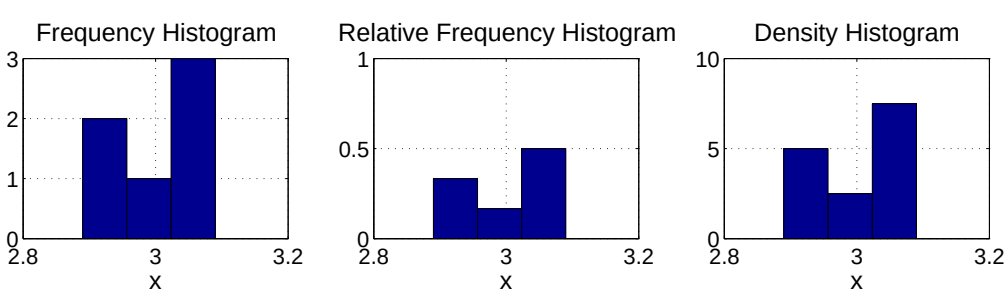
a **relative frequency histogram** or **empirical frequency histogram** is the histogram:

$$((b_1, h_1), (b_2, h_2), \ldots, (b_m, h_m)) \ , \ \text{where } h_k := n^{-1}\#\{x_i : x_i \in b_k\} \ ,$$

and a **density histogram** is the histogram:

$$((b_1, h_1), (b_2, h_2), \ldots, (b_m, h_m)) \ , \ \text{where } h_k := (w_k n)^{-1}\#\{x_i : x_i \in b_k\} \ , w_k := \overline{b}_k - \underline{b}_k \ .$$

Figure 12: Frequency, Relative Frequency and Density Histograms

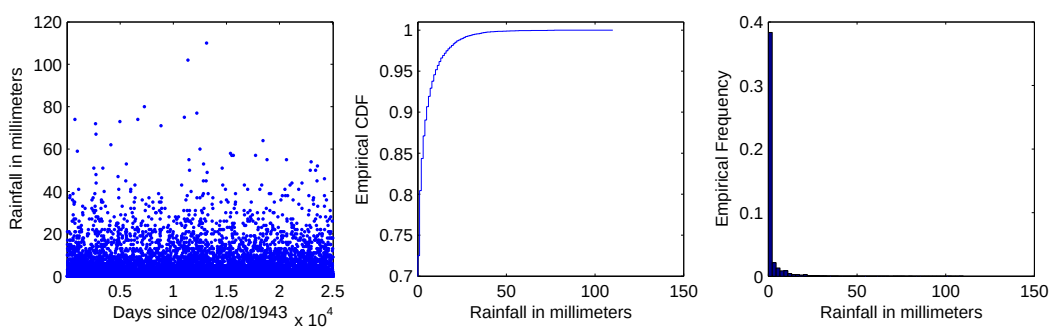


Labwork 16.15 (Histograms with specified number of bins for univariate data) Let us use samples from the current measurements and plot various histograms. First we use `hist` function (read `help hist`) to make a histogram with three bins. Then let's make the three types of histograms as shown in Figure 12 as follows:

```
>> data_n6_1 = [3.03 3.09 2.89 3.04 3.02 2.94];
>> hist(data_n6_1,3) % see what hist does with 3 bins in Figure Window
>> % Now let us look deeper into the last hist call
>> [Fs, Cs] = hist(data_n6_1,3) % Cs is the bin centers and Fs is the frequencies of data set
Fs =     2     1     3
Cs =    2.9233    2.9900    3.0567
>> % produce a histogram, the last argument 1 is the width value for immediately adjacent bars - help bar
>> bar(Cs,Fs,1) % create a frequency histogram
>> bar(Cs,Fs/6,1) % create a relative frequency histogram
>> bar(Cs,Fs/(0.0667*6),1) % create a density histogram (area of bars must sum to 1)
>> sum(Fs/(0.0667*6) .* ones(1,3)*0.0667) % checking if area does sum to 1
ans =     1
```

Example 16.16 Understanding the historical behavior of daily rainfall received by Christchurch can help the city's drainage engineers determine the limits of a newly designed earthquake-resistant drainage system. New Zealand's meteorological service NIWA provides weather data. Daily rainfall data (in millimetres) as measured from the CHCH aeroclub station available from `http://www.math.canterbury.ac.nz/php/lib/cliflo/rainfall.php`. The daily rainfall, its empirical cumulative distribution function (ECDF) and its relative frequency histogram are shown in Figure 13.

Figure 13: Daily rainfalls in Christchurch since February 8 1943



Example 16.17 In more complex experiments we do not just observe n realizations of IID random variables. In the case of observing the trajectory of a double pendulum we have a so-called **nonlinear time series**, i.e., a **dependent sequence of random variables** measuring the position of each arm through time. If we observe two or more independent realizations of such time series then we can estimate parameters in a probability model of the system – a "noisy" Euler-Lagrangian system of ODEs with parameters for friction, gravity, arm-length, center of mass, etc.

Sensors called *optical encoders* have been attached to the top end of each arm of a chaotic double pendulum in order to obtain the angular position of each arm through time as shown

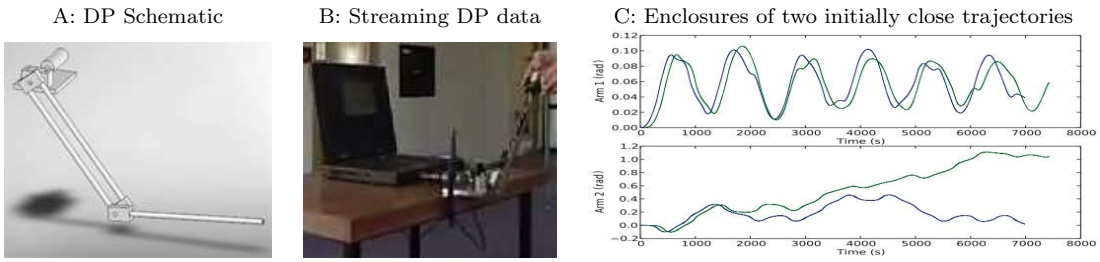
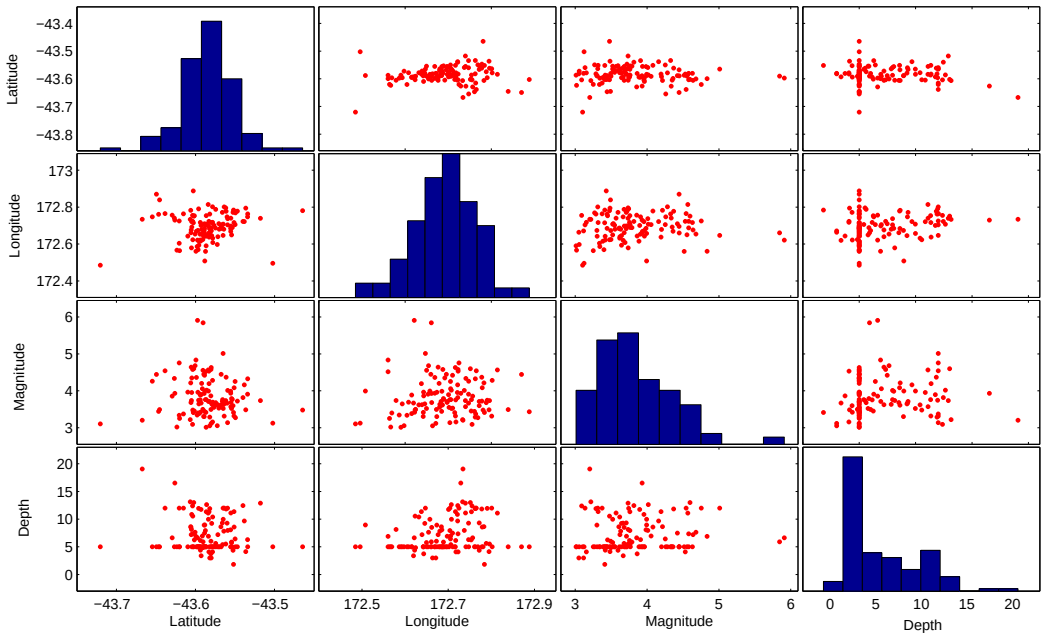


Figure 14: Double Pendulum

in Figure 14. Time series of the angular position of each arm for two trajectories that were initialized very similarly, say the angles of each arm of the double pendulum are almost the same at the initial time of release. Note how quickly the two trajectories diverge! System with such a sensitivity to initial conditions are said to be *chaotic*.

Example 16.18 Let us visualize the data of earth quakes, in terms of its latitude, longitude, magnitude and depth of each earthquake in Christchurch on February 22 2011. This data can be fetched from the URL `http://magma.geonet.org.nz/resources/quakesearch/`. Every pair of variables in the four tuple can be seen using a *matrix plot* shown in Figure 15.

Figure 15: Matrix of Scatter Plots of the latitude, longitude, magnitude and depth of the 22-02-2011 earth quakes in Christchurch, New Zealand.



17 Convergence of Random Variables

This important topic is concerned with the limiting behavior of sequences of RVs. We want to understand what it means for a sequence of random variables

$$\{X_i\}_{i=1}^{\infty} := X_1, X_2, \ldots$$

to converge to another random variable X? From a statistical or decision-making viewpoint $n \to \infty$ is associated with the amount of data or information $\to \infty$.
Let us first refresh ourselves with notions of convergence, limits and continuity in the real line (***S*** 17.1) before proceeding further.

17.1 Limits of Real Numbers – A Review

Definition 17.1 A sequence of real numbers $\{x_i\}_{i=1}^{\infty} := x_1, x_2, \ldots$ is said to converge to a limit $a \in \mathbb{R}$ and denoted by:

$$\lim_{i \to \infty} x_i = a \ ,$$

if for every natural number $m \in \mathbb{N}$, a natural number $N_m \in \mathbb{N}$ exists such that for every $j \geq N_m$, $|x_j - a| \leq \frac{1}{m}$.

In words, $\lim_{i \to \infty} x_i = a$ means the following: no matter how small you make $\frac{1}{m}$ by picking as large an m as you wish, I can find an N_m, that may depend on m, such that every number in the sequence beyond the N_m-th element is within distance $\frac{1}{m}$ of the limit a.

Example 17.2 Consider the boring sequence $\{x_i\}_{i=1}^{\infty} = 17, 17, 17, \ldots$. Show that $\lim_{i \to \infty} x_i = 17$ satisfies Definition 17.1.
For every $m \in \mathbb{N}$, we can take $N_m = 1$ and satisfy the Definition of the limit, i.e.:

$$\text{for every } j \geq N_m = 1, \ |x_j - 17| = |17 - 17| = 0 \leq \frac{1}{m} \ .$$

Example 17.3 Let $\{x_i\}_{i=1}^{\infty} = \frac{1}{1}, \frac{1}{2}, \frac{1}{3}, \ldots$, i.e. $x_i = \frac{1}{i}$. Show that $\lim_{i \to \infty} x_i = 0$ satisfies Definition 17.1.
For every $m \in \mathbb{N}$, we can take $N_m = m$ and satisfy the Definition of the limit, i.e.:

$$\text{for every } j \geq N_m = m, \ |x_j - 0| = \left| \frac{1}{j} - 0 \right| = \frac{1}{j} \leq \frac{1}{m} \ .$$

Consider the class of discrete RVs with distributions that place all probability mass on a single real number. This is the probability model for the deterministic real variable.

Definition 17.4 Point Mass(θ) Random Variable. Given a specific point $\theta \in \mathbb{R}$, we say an RV X has point mass at θ or is Point Mass(θ) distributed if the DF is:

$$F(x;\theta) = \begin{cases} 0 & \text{if } x < \theta \\ 1 & \text{if } x \geq \theta \end{cases} \tag{57}$$

and the PMF is:

$$f(x;\theta) = \begin{cases} 0 & \text{if } x \neq \theta \\ 1 & \text{if } x = \theta \end{cases} \tag{58}$$

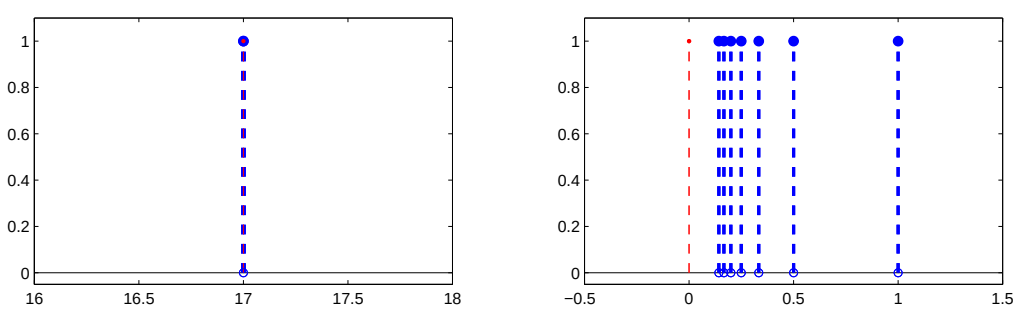
Thus, Point Mass(θ) RV X is deterministic in the sense that every realization of X is exactly equal to $\theta \in \mathbb{R}$. We will see that this distribution plays a central limiting role in asymptotic statistics.

Mean and variance of Point Mass(θ) **RV:** Let $X \sim$ Point Mass(θ). Then:

$$E(X) = \sum_x x f(x) = \theta \times 1 = \theta \ , \qquad V(X) = E(X^2) - (E(X))^2 = \theta^2 - \theta^2 = 0 \ .$$

Example 17.5 Can the sequences of $\{\text{Point Mass}(\theta_i = 17)\}_{i=1}^{\infty}$ and $\{\text{Point Mass}(\theta_i = 1/i)\}_{i=1}^{\infty}$ RVs be the same as the two sequences of real numbers $\{x_i\}_{i=1}^{\infty} = 17, 17, 17, \ldots$ and $\{x_i\}_{i=1}^{\infty} = \frac{1}{1}, \frac{1}{2}, \frac{1}{3}, \ldots$?
Yes why not – just move to space of distributions over the reals! See Figure 16.

Figure 16: Sequence of $\{\text{Point Mass}(17)\}_{i=1}^{\infty}$ RVs (left panel) and $\{\text{Point Mass}(1/i)\}_{i=1}^{\infty}$ RVs (only the first seven are shown on right panel) and their limiting RVs in red.



However, several other sequences also approach the limit 0. Some such sequences that approach the limit 0 from the right are:

$$\{x_i\}_{i=1}^{\infty} = \frac{1}{1}, \frac{1}{4}, \frac{1}{9}, \ldots \qquad \text{and} \qquad \{x_i\}_{i=1}^{\infty} = \frac{1}{1}, \frac{1}{8}, \frac{1}{27}, \ldots \ ,$$

and some that approach the limit 0 from the left are:

$$\{x_i\}_{i=1}^{\infty} = -\frac{1}{1}, -\frac{1}{2}, -\frac{1}{3}, \ldots \qquad \text{and} \qquad \{x_i\}_{i=1}^{\infty} = -\frac{1}{1}, -\frac{1}{4}, -\frac{1}{9}, \ldots \ ,$$

and finally some that approach 0 from either side are:

$$\{x_i\}_{i=1}^{\infty} = -\frac{1}{1}, +\frac{1}{2}, -\frac{1}{3}, \ldots \qquad \text{and} \qquad \{x_i\}_{i=1}^{\infty} = -\frac{1}{1}, +\frac{1}{4}, -\frac{1}{9} \ldots \ .$$

When we do not particularly care about the specifics of a sequence of real numbers $\{x_i\}_{i=1}^{\infty}$, in terms of the exact values it takes for each i, but we are only interested that it converges to a limit a we write:

$$x \to a$$

and say that x approaches a. If we are only interested in those sequences that converge to the limit a from the right or left, we write:

$$x \to a^+ \qquad \text{or} \qquad x \to a^-$$

and say x approaches a from the right or left, respectively.

Definition 17.6 (Limits of Functions) We say a function $f(x) : \mathbb{R} \to \mathbb{R}$ has a **limit** $L \in \mathbb{R}$ as x approaches a and write:

$$\lim_{x \to a} f(x) = L \ ,$$

provided $f(x)$ is arbitrarily close to L for all values of x that are sufficiently close to, but not equal to, a. We say that f has a **right limit** L_R or **left limit** L_L as x approaches a from the left or right, and write:

$$\lim_{x \to a^+} f(x) = L_R \qquad \text{or} \qquad \lim_{x \to a^-} f(x) = L_L \ ,$$

provided $f(x)$ is arbitrarily close to L_R or L_L for all values of x that are sufficiently close to but not equal to a from the right of a or the left of a, respectively. When the limit is not an element of $\mathbb{R}$ or when the left and right limits are distinct or when it is undefined, we say that the limit does not exist.

Example 17.7 Consider the function $f(x) = \frac{1}{x^2}$. Does $\lim_{x \to 1} f(x)$ exist? What about $\lim_{x \to 0} f(x)$?

$$\lim_{x \to 1} f(x) = \lim_{x \to 1} \frac{1}{x^2} = 1$$

exists since the limit $1 \in \mathbb{R}$, and the right and left limits are the same:

$$\lim_{x \to 1^+} f(x) = \lim_{x \to 1^+} \frac{1}{x^2} = 1 \qquad \text{and} \qquad \lim_{x \to 1^-} f(x) = \lim_{x \to 1^-} \frac{1}{x^2} = 1 \ .$$

However, the following limit does not exist:

$$\lim_{x \to 0} f(x) = \lim_{x \to 0} \frac{1}{x^2} = \infty$$

since $\infty \notin \mathbb{R}$. When $\lim_{x \to 0} f(x) = \pm\infty$ it is said to diverge.

Let us next look at some limits of functions that exist despite the function itself being undefined at the limit point.
For $f(x) = \frac{x^3-1}{x-1}$, the following limit exists:

$$\lim_{x\to 1} f(x) = \lim_{x\to 1} \frac{x^3-1}{x-1} = \lim_{x\to 1} \frac{(x-1)(x^2+x+1)}{(x-1)} = \lim_{x\to 1} x^2+x+1 = 3$$

despite the fact that $f(1) = \frac{1^3-1}{1-1} = \frac{0}{0}$ itself is undefined and does not exist.

Example 17.8 What is $f(x) = (1+x)^{\frac{1}{x}}$ as x approaches 0?
The limit of $\lim_{x\to 0}(1+x)^{\frac{1}{x}}$ exists and it is the Euler's constant e :

$$\begin{aligned}
\lim_{x\to 0} f(x) &= \lim_{x\to 0}(1+x)^{\frac{1}{x}} \\
&= \lim_{x\to 0}(x+1)^{(1/x)} \quad \text{Indeterminate form of type } 1^{\infty}. \\
&= \exp\left(\lim_{x\to 0}\log((x+1)^{(1/x)})\right) \quad \text{Transformed using } \exp(\lim_{x\to 0}\log((x+1)^{(1/x)})) \\
&= \exp\left(\lim_{x\to 0}(\log(x+1))/x\right) \quad \text{Indeterminate form of type } 0/0. \\
&= \exp\left(\lim_{x\to 0}\frac{d\log(x+1)/dx}{dx/dx}\right) \quad \text{Applying L'Hospital's rule} \\
&= \exp\left(\lim_{x\to 0} 1/(x+1)\right) \quad \text{limit of a quotient is the quotient of the limits} \\
&= \exp\left(1/(\lim_{x\to 0}(x+1))\right) \quad \text{The limit of } x+1 \text{ as } x \text{ approaches 0 is 1} \\
&= \exp(1) = e \approxeq 2.71828 \ .
\end{aligned}$$

Notice that the above limit exists despite the fact that $f(0) = (1+0)^{\frac{1}{0}}$ itself is undefined and does not exist.

Next we look at some Examples of limits at infinity.

Example 17.9 The limit of $f(n) = \left(1-\frac{\lambda}{n}\right)^n$ as n approaches ∞ exists and it is $e^{-\lambda}$:

$$\lim_{n\to\infty} f(n) = \lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda} \ .$$

Example 17.10 The limit of $f(n) = \left(1-\frac{\lambda}{n}\right)^{-\alpha}$, for some $\alpha > 0$, as n approaches ∞ exists and it is 1 :

$$\lim_{n\to\infty} f(n) = \lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-\alpha} = 1 \ .$$

Definition 17.11 (Continuity of a function) We say a real-valued function $f(x) : D \to \mathbb{R}$ with the domain $D \subset \mathbb{R}$ is **right continuous** or **left continuous** at a point $a \in D$, provided:

$$\lim_{x \to a^+} f(x) = f(a) \qquad \text{or} \qquad \lim_{x \to a^-} f(x) = f(a) \ ,$$

respectively. We say f is **continuous** at $a \in D$, provided:

$$\lim_{x \to a^+} f(x) = f(a) = \lim_{x \to a^-} f(x) \ .$$

Finally, f is said to be continuous if f is continuous at every $a \in D$.

Example 17.12 (Discontinuity of $f(x) = (1+x)^{\frac{1}{x}}$ at 0) Let us reconsider the function $f(x) = (1+x)^{\frac{1}{x}} : \mathbb{R} \to \mathbb{R}$. Clearly, $f(x)$ is continuous at 1, since:

$$\lim_{x \to 1} f(x) = \lim_{x \to 1} (1+x)^{\frac{1}{x}} = 2 = f(1) = (1+1)^{\frac{1}{1}} \ ,$$

but it is not continuous at 0, since:

$$\lim_{x \to 0} f(x) = \lim_{x \to 0} (1+x)^{\frac{1}{x}} = e \approxeq 2.71828 \neq f(0) = (1+0)^{\frac{1}{0}} \ .$$

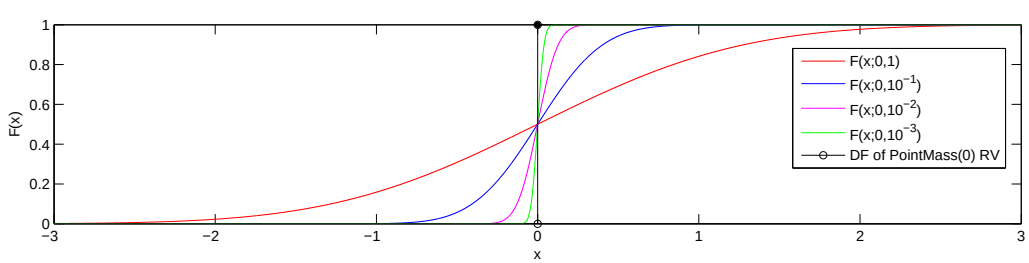
Thus, $f(x)$ is not a continuous function over $\mathbb{R}$.

This is the end of review of limits of sequences of real numbers. This section is a prequel to our main topic in the next section.

17.2 Limits of Random Variables

Example 17.13 (Convergence of $X_n \sim \text{Normal}(0, 1/n)$) Suppose you are given an independent sequence of RVs $\{X_n\}_{n=1}^{\infty}$, where $X_n \sim \text{Normal}(0, 1/n)$. How would you talk about the convergence of $X_n \sim \text{Normal}(0, 1/n)$ as n approaches ∞? Take a look at Figure 17 for insight. The probability mass of X_n increasingly concentrates about 0 as n approaches ∞ and the variance $1/n$ approaches 0, as depicted in Figure 17. Based on this observation, can we expect $\lim_{n \to \infty} X_n = X$, where the limiting RV $X \sim$ Point Mass(0)?
The answer is **no**. This is because $P(X_n = X) = 0$ for any n, since $X \sim$ Point Mass(0) is a discrete RV with exactly one outcome 0 and $X_n \sim \text{Normal}(0, 1/n)$ is a continuous RV for every n, however large. In other words, a continuous RV, such as X_n, has 0 probability of realizing any single real number in its support, such as 0.

Figure 17: Distribution functions of several Normal(μ, σ^2) RVs for $\sigma^2 = 1, \frac{1}{10}, \frac{1}{100}, \frac{1}{1000}$.



Thus, we need more sophisticated notions of convergence for sequences of RVs. One such notion is formalized next as it is a minimal prerequisite for a clear understanding of two basic Theorems in Probability and Statistics:

1. Law of Large Numbers and
2. Central Limit Theorem.

Definition 17.14 (Convergence in Distribution) Let $X_1, X_2, \ldots,$ be a sequence of RVs and let X be another RV. Let F_n denote the DF of X_n and F denote the DF of X. Then we say that X_n converges to X in distribution, and write:

$$X_n \rightsquigarrow X$$

if for any real number t at which F is continuous,

$$\lim_{n\to\infty} F_n(t) = F(t) \qquad \text{[in the sense of Definition 17.6].}$$

The above limit, by (8) in our Definition 6.5 of a DF, can be equivalently expressed as follows:

$$\lim_{n\to\infty} P(\ \{\omega : X_n(\omega) \le t\}\ ) = P(\ \{\omega : X(\omega) \le t\}\ ),$$
$$\text{i.e.} \quad P(\ \{\omega : X_n(\omega) \le t\}\ ) \to P(\ \{\omega : X(\omega) \le t\}\ ), \quad \text{as} \quad n \to \infty\ .$$

Let us revisit the problem of convergence in Example 17.13 armed with our new notion of convergence.

Example 17.15 (Example for Convergence in distribution) Suppose you are given an independent sequence of RVs $\{X_i\}_{i=1}^n$, where $X_i \sim \text{Normal}(0, 1/i)$ with DF F_n and let $X \sim$ Point Mass(0) with DF F. We can formalize our observation in Example 17.13 that X_n is concentrating about 0 as $n \to \infty$ by the statement:

$$X_n \text{ is converging in distribution to } X\text{, i.e.,} \qquad X_n \rightsquigarrow X\ .$$

Proof: To check that the above statement is true we need to verify that the Definition of convergence in distribution is satisfied for our sequence of RVs $X_1, X_2, \ldots$ and the limiting RV X. Thus, we need to verify that for any continuity point t of the Point Mass(0) DF F, $\lim_{n\to\infty} F_n(t) = F(t)$. First note that

$$X_n \sim \text{Normal}(0, 1/n) \implies Z := \sqrt{n}X_n \sim \text{Normal}(0, 1)\ ,$$

and thus

$$F_n(t) = P(X_n < t) = P(\sqrt{n}X_n < \sqrt{n}t) = P(Z < \sqrt{n}t)\ .$$

The only discontinuous point of F is 0 where F jumps from 0 to 1.
When $t < 0$, $F(t)$, being the constant 0 function over the interval $(-\infty, 0)$, is continuous at t. Since $\sqrt{n}t \to -\infty$, as $n \to \infty$,

$$\lim_{n\to\infty} F_n(t) = \lim_{n\to\infty} P(Z < \sqrt{n}t) = 0 = F(t)\ .$$

And, when $t > 0$, $F(t)$, being the constant 1 function over the interval $(0, \infty)$, is again continuous at t. Since $\sqrt{n}t \to \infty$, as $n \to \infty$,

$$\lim_{n\to\infty} F_n(t) = \lim_{n\to\infty} P(Z < \sqrt{n}t) = 1 = F(t)\ .$$

Thus, we have proved that $X_n \rightsquigarrow X$ by verifying that for any t at which the Point Mass(0) DF F is continuous, we also have the desired equality: $\lim_{n\to\infty} F_n(t) = F(t)$.

However, note that

$$F_n(0) = \frac{1}{2} \ne F(0) = 1\ ,$$

and so convergence fails at 0, i.e. $\lim_{n\to\infty} F_n(t) \ne F(t)$ at $t = 0$. But, $t = 0$ is not a continuity point of F and the Definition of convergence in distribution only requires the convergence to hold at continuity points of F.

18 Some Basic Limit Laws of Statistics

Theorem 18.1 (Law of Large Numbers (LLN)) If we are given a sequence of independently and identically distributed (IID) RVs, $X_1, X_2, \ldots \overset{\text{IID}}{\sim} X_1$ and if $E(X_1)$ exists, i.e. $E(\text{abs}(X)) < \infty$, then the sample mean $\overline{X}_n$ converges in distribution to the expectation of any one of the IID RVs, say Point Mass$(E(X_1))$ by convention. More formally, we write:

$$\text{If} \quad X_1, X_2, \ldots \overset{\text{IID}}{\sim} X_1 \text{ and if } E(X_1) \text{ exists, then } \overline{X}_n \rightsquigarrow \text{Point Mass}(E(X_1)) \text{ as } n \to \infty \ .$$

Proof: Our proof is based on the convergence of characteristic functions (CFs).
First, the CF of Point Mass$(E(X_1))$ is

$$E(e^{\imath t E(X_1)}) = e^{\imath t E(X_1)} \ ,$$

since $E(X_1)$ is just a constant, i.e., a Point Mass RV that puts all of its probability mass at $E(X_1)$.
Second, the CF of $\overline{X}_n$ is

$$E\left(e^{\imath t \overline{X}_n}\right) = E\left(e^{\imath t \frac{1}{n} \sum_{k=1}^n X_k}\right) = E\left(\prod_{k=1}^n e^{\imath t X_k/n}\right) = \prod_{k=1}^n E\left(e^{\imath t X_k/n}\right) = \prod_{k=1}^n \phi_{X_k}(t/n) = \prod_{k=1}^n \phi_{X_1}(t/n) = (\phi_{X_1}(t/n))^n \ .$$

Let us recall Landau's "small o" notation for the relation between two functions. We say, $f(x)$ is **small o** of $g(x)$ if f is dominated by g as $x \to \infty$, i.e., $\frac{|f(x)|}{|g(x)|} \to 0$ as $x \to \infty$. More formally, for every $\epsilon > 0$, there exists an x_ϵ such that for all $x > x_\epsilon$ $|f(x)| < \epsilon |g(x)|$. For example, $\log(x)$ is $o(x)$, x^2 is $o(x^3)$ and x^m is $o(x^{m+1})$ for $m \geq 1$.
Third, we can expand any CF whose expectation exists as a Taylor series with a remainder term that is $o(t)$ as follows:

$$\phi_X(t) = 1 + \imath t E(X) + o(t) \ .$$

Hence,

$$\phi_{X_1}(t/n) = 1 + \imath \frac{t}{n} E(X_1) + o\left(\frac{t}{n}\right)$$

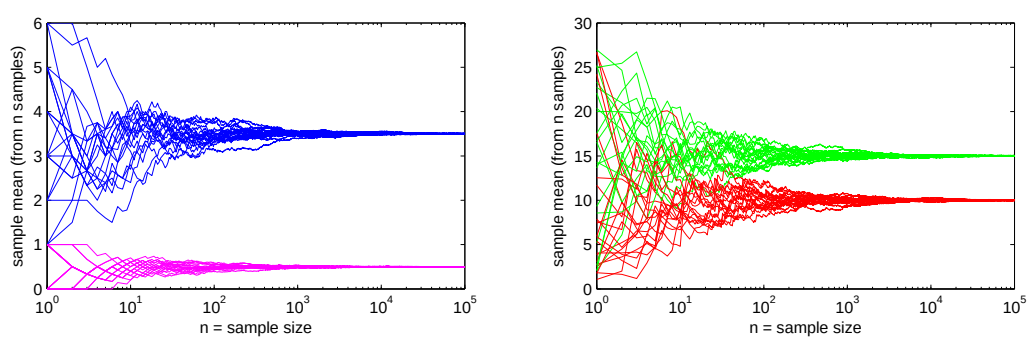
and

$$E\left(e^{\imath t \overline{X}_n}\right) = \left(1 + \imath \frac{t}{n} E(X_1) + o\left(\frac{t}{n}\right)\right)^n \to e^{\imath t E(X_1)} \text{ as } n \to \infty \ .$$

For the last limit we have used $\left(1 + \frac{x}{n}\right)^n \to e^x$ as $n \to \infty$.
Finally, we have shown that $E\left(e^{\imath t \overline{X}_n}\right)$, the CF of the n-sample mean RV $\overline{X}_n$, converges to $E(e^{\imath t E(X_1)}) = e^{\imath t E(X_1)}$, the CF of the Point Mass$(E(X_1))$ RV, as the sample size n tends to infinity.

Heuristic Interpretation of LLN

Figure 18: Sample mean $\overline{X}_n$ as a function of sample size n for 20 replications from independent realizations of a fair die (blue), fair coin (magenta), Uniform$(0, 30)$ RV (green) and Exponential(0.1) RV (red) with population means $(1+2+3+4+5+6)/6 = 21/6 = 3.5$, $(0+1)/2 = 0.5$, $(30-0)/2 = 15$ and $1/0.1 = 10$, respectively.



The distribution of the sample mean RV $\overline{X}_n$ obtained from an independent and identically distributed sequence of RVs $X_1, X_2, \ldots$ [i.e. all the RVs X_i's are independent of one another and have the same distribution function, and thereby the same expectation, variance and higher moments], concentrates around the expectation of any one of the RVs in the sequence, say that of the first one $E(X_1)$ [without loss of generality], as n approaches infinity. See Figure 18 for examples of 20 replicates of the sample mean of IID sequences from four RVs. All the sample mean trajectories converge to the corresponding population mean.

Application: Point Estimation of $E(X_1)$

LLN gives us a method to obtain a **point estimator** that gives "the single best guess" for the possibly unknown population mean $E(X_1)$ based on $\overline{X}_n$, the sample mean, of a simple random sequence (SRS) or independent and identically distributed (IID) sequence of n RVs $X_1, X_2, \ldots, X_n \stackrel{IID}{\sim} X_1$.

Example 18.2 Let $X_1, X_2, \ldots, X_n \stackrel{IID}{\sim} X_1$, where X_1 is an Exponential(λ^*) RV, i.e., let

$$X_1, X_2, \ldots, X_n \stackrel{IID}{\sim} \text{Exponential}(\lambda^*) \ .$$

Typically, we do not know the "true" parameter $\lambda^* \in \mathbb{\Lambda} = (0, \infty)$ or the population mean $E(X_1) = 1/\lambda^*$. But by LLN, we know that

$$\overline{X}_n \rightsquigarrow \text{Point Mass}(E(X_1)) \ ,$$

and therefore, we can use the sample mean $\overline{X}_n$ as a point estimator of $E(X_1) = 1/\lambda^*$.
Now, suppose you model seven waiting times in nearest minutes between Orbiter buses at Balgay street as follows:

$$X_1, X_2, \ldots, X_7 \stackrel{IID}{\sim} \text{Exponential}(\lambda^*) \ ,$$

and have the following realization as your observed data:

$$(x_1, x_2, \ldots, x_7) = (2, 12, 8, 9, 14, 15, 11) \ .$$

Then you can use the observed sample mean $\overline{x}_7 = (2+12+8+9+14+15+11)/7 = 71/7 \approxeq 10.14$ as a **point estimate** of the population mean $E(X_1) = 1/\lambda^*$. By the rearrangement $\lambda^* = 1/E(X_1)$, we can also obtain a point estimate of the "true" parameter λ^* from $1/\overline{x}_7 = 7/71 \approxeq 0.0986$.

Remark 18.3 Point estimates are realizations of the Point Estimator: We say the statistic $\overline{X}_n$, which is a random variable that depends on the data R$\vec{\text{V}}$ $(X_1, X_2, \ldots, X_n)$, is a **point estimator** of $E(X_1)$. But once we have a realization of the data R$\vec{\text{V}}$, i.e., our observed data vector $(x_1, x_2, \ldots, x_n)$ and its corresponding realization as observed sample mean $\overline{x}_n$, we say $\overline{x}_n$ is a **point estimate** of $E(X_1)$. In other words, the point estimate $\overline{x}_n$ is a realization of the the random variable $\overline{X}_n$ called the point estimator of $E(X_1)$. Therefore, when we observe a new data vector $(x'_1, x'_2, \ldots, x'_n)$ that is different from our first data vector $(x_1, x_2, \ldots, x_n)$, our point estimator of $E(X_1)$ is still $\overline{X}_n$ but the point estimate $n^{-1}\sum_{i=1}^n x'_i$ may be different from the first point estimate $n^{-1}\sum_{i=1}^n x_i$. The sample means from n samples for 20 replications (repeats of the experiment) are typically distinct especially for small n as shown in Figure 18.

Example 18.4 Let $X_1, X_2, \ldots, X_n \stackrel{IID}{\sim} X_1$, where X_1 is an Bernoulli(θ^*) RV, i.e., let

$$X_1, X_2, \ldots, X_n \stackrel{IID}{\sim} \text{Bernoulli}(\theta^*) \ .$$

Typically, we do not know the "true" parameter $\theta^* \in \Theta = [0, 1]$, which is the same as the population mean $E(X_1) = \theta^*$. But by LLN, we know that

$$\overline{X}_n \rightsquigarrow \text{Point Mass}(E(X_1)) \ ,$$

and therefore, we can use the sample mean $\overline{X}_n$ as a point estimator of $E(X_1) = \theta^*$.
Now, suppose you model seven coin tosses (encoding Heads as 1 with probability θ^* and Tails as 0 with probability $1 - \theta^*$) as follows:

$$X_1, X_2, \ldots, X_7 \stackrel{IID}{\sim} \text{Bernoulli}(\theta^*) \ ,$$

and have the following realization as your observed data:

$$(x_1, x_2, \ldots, x_7) = (0, 1, 1, 0, 0, 1, 0) \ .$$

Then you can use the observed sample mean $\overline{x}_7 = (0 + 1 + 1 + 0 + 0 + 1 + 0)/7 = 3/7 \approxeq 0.4286$ as a **point estimate** of the population mean $E(X_1) = \theta^*$. Thus, our "single best guess" for $E(X_1)$ which is the same as the probability of Heads is $\overline{x}_7 = 3/7$.

Of course, if we tossed the same coin in the same IID manner another seven times or if we observed another seven waiting times of orbiter buses at a different bus-stop or on a different day we may get a different point estimate for $E(X_1)$. See the intersection of the twenty magenta sample mean trajectories for simulated tosses of a fair coin from IID Bernoulli($\theta^* = 1/2$) RVs and the twenty red sample mean trajectories for simulated waiting times from IID Exponential($\lambda^* = 1/10$) RVs in Figure 18 with $n = 7$. Clearly, the point estimates for such a small sample size are fluctuating wildly! However, the fluctuations in the point estimates settles down for larger sample sizes.
The next natural question is how large should the sample size be in order to have a small interval of width, say 2ϵ, "contain" $E(X_1)$, the quantity of interest, with a high probability, say $1 - \alpha$? If we can answer this then we can make probability statements like the following:

$$P(\text{error} < \text{tolerance}) = P(|\overline{X}_n - E(X_1)| < \epsilon) = P(-\epsilon < \overline{X}_n - E(X_1) < \epsilon) = 1 - \alpha \ .$$

In order to ensure the $\text{error} = |\overline{X}_n - E(X_1)|$ in our estimate of $E(X_1)$ is within a required $\text{tolerance} = \epsilon$ we need to know the full distribution of $\overline{X}_n - E(X_1)$ itself. The Central Limit Theorem (CLT) helps us here.

Theorem 18.5 (Central Limit Theorem (CLT)) If we are given a sequence of independently and identically distributed (IID) RVs, $X_1, X_2, \ldots \stackrel{\text{IID}}{\sim} X_1$ and if $E(X) < \infty$ and $V(X_1) < \infty$, then the sample mean $\overline{X}_n$ converges in distribution to the Normal RV with mean given by any one of the IID RVs, say $E(X_1)$ by convention, and variance given by $\frac{1}{n}$ times the variance of any one of the IID RVs, say $V(X_1)$ by convention. More formally, we write:

$$\text{If} \quad X_1, X_2, \ldots \stackrel{\text{IID}}{\sim} X_1 \text{ and if } E(X_1) < \infty, V(X_1) < \infty$$
$$\text{then } \overline{X}_n \rightsquigarrow \text{Normal}\left(E(X_1), \frac{V(X_1)}{n}\right) \quad \text{as } n \to \infty \ , \quad (59)$$

or equivalently after standardization:

$$\text{If} \quad X_1, X_2, \ldots \overset{\text{IID}}{\sim} X_1 \text{ and if } E(X_1) < \infty, V(X_1) < \infty$$
$$\text{then} \quad \frac{\overline{X}_n - E(X_1)}{\sqrt{V(X_1)/n}} \rightsquigarrow Z \sim \text{Normal}\,(0,1) \quad \text{as } n \to \infty \ . \tag{60}$$

Proof: Our proof is based on the convergence of characteristic functions (CFs). We will prove the standardized form of the CLT in Equation (60) by showing that the CF of

$$U_n := \frac{\overline{X}_n - E(X_1)}{\sqrt{V(X_1)/n}}$$

converges to the CF of Z, the Normal$(0,1)$ RV. First, note from Equation (38) that the CF of $Z \sim$ Normal$(0,1)$ is:

$$\phi_Z(t) = E\left(e^{\imath t Z}\right) = e^{-t^2/2} \ .$$

Second,

$$U_n := \frac{\overline{X}_n - E(X_1)}{\sqrt{V(X_1)/n}} = \frac{\sum_{k=1}^n X_k - nE(X_1)}{\sqrt{nV(X_1)}} = \frac{1}{\sqrt{n}} \sum_{k=1}^n \left(\frac{X_k - E(X_1)}{\sqrt{V(X_1)}} \right) \ .$$

Therefore, the CF of U_n is

$$\begin{aligned}\phi_{U_n}(t) &= E\left(\exp\left(\imath t U_n\right)\right) = E\left(\exp\left(\imath \frac{t}{\sqrt{n}} \sum_{k=1}^n \frac{X_k - E(X_1)}{\sqrt{V(X_1)}}\right)\right) = \prod_{k=1}^n E\left(\exp\left(\imath \frac{t}{\sqrt{n}} \frac{X_k - E(X_1)}{\sqrt{V(X_1)}}\right)\right) \\ &= \left(E\left(\exp\left(\imath \frac{t}{\sqrt{n}} \frac{X_1 - E(X_1)}{\sqrt{V(X_1)}}\right)\right)\right)^n \ .\end{aligned}$$

Now, if we let

$$Y = \frac{X_1 - E(X_1)}{\sqrt{V(X_1)}}$$

then

$$E(Y) = 0\,, \quad E(Y^2) = 1\,, \text{and} \quad V(Y) = 1 \ .$$

So, the CF of U_n is

$$\phi_{U_n}(t) = \left(\phi_Y\left(\frac{t}{\sqrt{n}}\right)\right)^n \ ,$$

and since we can Taylor expand $\phi_Y(t)$ as follows:

$$\phi_Y(t) = 1 + \imath t E(Y) + \imath^2 \frac{t^2}{2} E(Y^2) + o(t^2) \ ,$$

which implies

$$\phi_Y\left(\frac{t}{\sqrt{n}}\right) = 1 + \frac{\imath t}{\sqrt{n}} E(Y) + \frac{\imath^2 t^2}{2n} E(Y^2) + o\left(\frac{t^2}{n}\right) \ ,$$

we finally get

$$\phi_{U_n}(t) = \left(\phi_Y\left(\frac{t}{\sqrt{n}}\right)\right)^n = \left(1 + \frac{\imath t}{\sqrt{n}} \times 0 + \frac{\imath^2 t^2}{2n} \times 1 + o\left(\frac{t^2}{n}\right)\right)^n = \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n \to e^{-t^2/2} = \phi_Z(t) \ .$$

For the last limit we have used $\left(1 + \frac{x}{n}\right)^n \to e^x$ as $n \to \infty$. Thus, we have proved Equation (60) which is equivalent to Equation (59) by a standardization argument that if $W \sim$ Normal(μ, σ^2) then $Z = \frac{W-\mu}{\sigma} \sim$ Normal$(0,1)$ through the linear transformation $W = \sigma Z + \mu$ of Example 9.8.

Application: Tolerating Errors in our estimate of $E(X_1)$

Recall that we wanted to ensure the $\mathsf{error} = |\overline{X}_n - E(X_1)|$ in our estimate of $E(X_1)$ is within a required $\mathsf{tolerance} = \epsilon$ and make the following probability statement:

$$P(\mathsf{error} < \mathsf{tolerance}) = P(|\overline{X}_n - E(X_1)| < \epsilon) = P(-\epsilon < \overline{X}_n - E(X_1) < \epsilon) = 1 - \alpha \ .$$

To be able to do this we needed to know the full distribution of $\overline{X}_n - E(X_1)$ itself.
Due to the Central Limit Theorem (CLT) we now know that (assuming n is large)

$$\begin{aligned} P(-\epsilon < \overline{X}_n - E(X_1) < \epsilon) &\approxeq P\left(-\frac{\epsilon}{\sqrt{V(X_1)/n}} < \frac{\overline{X}_n - E(X_1)}{\sqrt{V(X_1)/n}} < \frac{\epsilon}{\sqrt{V(X_1)/n}}\right) \\ &= P\left(-\frac{\epsilon}{\sqrt{V(X_1)/n}} < Z < \frac{\epsilon}{\sqrt{V(X_1)/n}}\right) \ , \end{aligned}$$

where $Z \sim \text{Normal}(0, 1)$.

Example 18.6 Suppose an IID sequence of observations $(x_1, x_2, \ldots, x_{80})$ was drawn from a distribution with variance $V(X_1) = 4$. What is the probability that the error in $\overline{x}_n$ used to estimate $E(X_1)$ is less than 0.1?
By CLT,

$$P(\mathsf{error} < 0.1) \approxeq P\left(-\frac{0.1}{\sqrt{4/80}} < Z < \frac{0.1}{\sqrt{4/80}}\right) = P(-0.447 < Z < 0.447) = 0.345 \ .$$

Suppose you want the error to be less than $\mathsf{tolerance} = \epsilon$ with a certain probability $1 - \alpha$. Then we can use CLT to do such **sample size calculations**. Recall the DF $\Phi(z) = P(Z < z)$ is tabulated in the standard normal table and now we want

$$P\left(-\frac{\epsilon}{\sqrt{V(X_1)/n}} < Z < \frac{\epsilon}{\sqrt{V(X_1)/n}}\right) = 1 - \alpha \ .$$

We know,

$$P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) = 1 - \alpha \ ,$$

make the picture here of $f_Z(z) = \varphi(z)$ to recall what $z_{\alpha/2}$, $z_{-\alpha/2}$, and the various areas below $f_Z(\cdot)$ in terms of $\Phi(\cdot)$ from the table really mean... (See Example 8.14).

where, $\Phi(z_{\alpha/2}) = 1 - \alpha/2$ and $\Phi(z_{-\alpha/2}) = 1 - \Phi(z_{\alpha/2}) = \alpha/2$. So, we set

$$\frac{\epsilon}{\sqrt{V(X_1)/n}} = z_{\alpha/2}$$

and rearrange to get

$$n = \left(\frac{\sqrt{V(X_1)} z_{\alpha/2}}{\epsilon} \right)^2 \tag{61}$$

for the needed sample size that will ensure that our error is less than our tolerance $= \epsilon$ with probability $1 - \alpha$. Of course, if n given by Equation (61) is not a natural number then we naturally round up to make it one!
A useful $z_{\alpha/2}$ value to remember: If $\alpha = 0.05$ when the probability of interest $1 - \alpha = 0.95$ then $z_{\alpha/2} = z_{0.025} = 1.96$.

Example 18.7 How large a sample size is needed to make the error in our estimate of the population mean $E(X_1)$ to be less than 0.1 with probability $1 - \alpha = 0.95$ if we are observing IID samples from a distribution with a population variance $V(X_1)$ of 4?
Using Equation (61) we see that the needed sample size is

$$n = \left(\frac{\sqrt{4} \times 1.96}{0.1} \right)^2 \approxeq 1537$$

Thus, it pays to check the sample size needed in advance of experimentation, provided you already know the population variance of the distribution whose population mean you are interested in estimating within a given tolerance and with a high probability.

Application: Set Estimation of $E(X_1)$

A useful byproduct of the CLT is the $(\mathbf{1} - \alpha)$ **confidence interval**, a random interval (or bivariate $\vec{\text{RV}}$) that contains $E(X_1)$, the quantity of interest, with probability $1 - \alpha$:

$$\left(\overline{X}_n \pm z_{\alpha/2} \sqrt{V(X_1)/n} \right) := \left(\overline{X}_n - z_{\alpha/2} \sqrt{V(X_1)/n} \, , \, \overline{X}_n + z_{\alpha/2} \sqrt{V(X_1)/n} \right) \quad . \tag{62}$$

We can easily see how Equation (62) is derived from CLT as follows:

$$\begin{aligned}
P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) &= 1 - \alpha \\
P\left(-z_{\alpha/2} < \frac{\overline{X}_n - E(X_1)}{\sqrt{V(X_1)/n}} < z_{\alpha/2}\right) &= 1 - \alpha \\
P\left(-\overline{X}_n - z_{\alpha/2}\sqrt{V(X_1)/n} < -E(X_1) < -\overline{X}_n + z_{\alpha/2}\sqrt{V(X_1)/n}\right) &= 1 - \alpha \\
P\left(\overline{X}_n + z_{\alpha/2}\sqrt{V(X_1)/n} > E(X_1) > \overline{X}_n - z_{\alpha/2}\sqrt{V(X_1)/n}\right) &= 1 - \alpha \\
P\left(\overline{X}_n - z_{\alpha/2}\sqrt{V(X_1)/n} < E(X_1) < \overline{X}_n + z_{\alpha/2}\sqrt{V(X_1)/n}\right) &= 1 - \alpha \\
P\left(E(X_1) \in \left(\overline{X}_n - z_{\alpha/2}\sqrt{V(X_1)/n}, \overline{X}_n + z_{\alpha/2}\sqrt{V(X_1)/n}\right)\right) &= 1 - \alpha \ .
\end{aligned}$$

Remark 18.8 Heuristic interpretation of the $(1 - \alpha)$ **confidence interval:** If we repeatedly produced samples of size n to contain $E(X_1)$ within a $\left(\overline{X}_n \pm z_{\alpha/2} \sqrt{V(X_1)/n} \right)$, say 100 times, then on average, $(1 - \alpha) \times 100$ repetitions will actually contain $E(X_1)$ within the random interval and $\alpha \times 100$ repetitions will fail to contain $E(X_1)$.

So far, we have assumed we know the population variance $V(X_1)$ in an IID experiment with n samples and tried to estimate the population mean $E(X_1)$. But in general, we will not know $V(X_1)$. We can still get a point estimate of $E(X_1)$ from the sample mean due to LLN but we won't be able to get a confidence interval for $E(X_1)$. Fortunately, a more elaborate form of the CLT tells us that even when we substitute the sample variance $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ for the population variance $V(X_1)$ the following $1-\alpha$ confidence interval for $E(X_1)$ works!

$$\left(\overline{X}_n \pm z_{\alpha/2} S_n/\sqrt{n}\right) := \left(\overline{X}_n - z_{\alpha/2} S_n/\sqrt{n}\,,\, \overline{X}_n + z_{\alpha/2} S_n/\sqrt{n}\right) \quad , \tag{63}$$

where, $S_n = \sqrt{S_n^2}$ is the sample standard deviation.
Let's return to our two examples again.

Example 18.9 We model the waiting times between Orbiter buses with unknown $E(X_1) = 1/\lambda^*$ as

$$X_1, X_2, \ldots, X_n \overset{IID}{\sim} \text{Exponential}(\lambda^*)$$

and observed the following data, sample mean, sample variance and sample standard deviation:

$$(x_1, x_2, \ldots, x_7) = (2, 12, 8, 9, 14, 15, 11),\ \overline{x}_7 = 10.143,\ s_7^2 = 19.143,\ s_7 = 4.375 \quad ,$$

respectively. Our point estimate and $1-\alpha = 95\%$ confidence interval for $E(X_1)$ are:

$$\overline{x}_7 = 10.143 \quad \text{and} \quad (\overline{x}_7 \pm z_{\alpha/2} s_7/\sqrt{7}) = (10.143 \pm 1.96 \times 4.375/\sqrt{7}) = (6.9016, 13.3841) \quad ,$$

respectively. So with 95% probability the true population mean $E(X_1) = 1/\lambda^*$ is contained in $(6.9016, 13.3841)$ and since the mean waiting time of 10 minutes promised by the Orbiter bus company is also within $(6.9016, 13.3841)$ we can be fairly certain that the company sticks to its promise.

Example 18.10 We model the tosses of a coin with unknown $E(X_1) = \theta^*$ as

$$X_1, X_2, \ldots, X_n \overset{IID}{\sim} \text{Bernoulli}(\theta^*)$$

and observed the following data, sample mean, sample variance and sample standard deviation:

$$(x_1, x_2, \ldots, x_7) = (0, 1, 1, 0, 0, 1, 0),\ \overline{x}_7 = 0.4286,\ s_7^2 = 0.2857,\ s_7 = 0.5345 \quad ,$$

respectively. Our point estimate and $1-\alpha = 95\%$ confidence interval for $E(X_1)$ are:

$$\overline{x}_7 = 0.4286 \quad \text{and} \quad (\overline{x}_7 \pm z_{\alpha/2} s_7/\sqrt{7}) = (0.4286 \pm 1.96 \times 0.5345/\sqrt{7}) = (0.0326, 0.8246) \quad ,$$

respectively. So with 95% probability the true population mean $E(X_1) = \theta^*$ is contained in $(0.0326, 0.8246)$ and since $1/2$ is contained in this interval of width 0.792 we cannot rule out that the flipped coin is not fair with $\theta^* = 1/2$.

Remark 18.11 The normal-based confidence interval for θ^* (as well as λ^* in the previous example) may not be a valid approximation here with just $n = 7$ samples. After all, the CLT only tells us that the point estimator $\widehat{\Theta}_n$ can be approximated by a normal distribution for large sample sizes. When the sample size n was increased from 7 to 100 by tossing the same coin another 93 times, a total of 57 trials landed as Heads. Thus the point estimate and confidence interval for $E(X_1) = \theta^*$ based on the sample mean and sample standard deviations are:

$$\widehat{\theta}_{100} = \frac{57}{100} = 0.57 \qquad \text{and} \qquad (0.57 \pm 1.96 \times 0.4975/\sqrt{100}) = (0.4725, 0.6675) \ .$$

Thus our confidence interval shrank considerably from a width of 0.792 to 0.195 after an additional 93 Bernoulli trials. Thus, we can make the width of the confidence interval as small as we want by making the number of observations or sample size n as large as we can.

19 Parameter Estimation and Likelihood

Now that we have been introduced to point and set estimation of the population mean using the notion of convergence in distribution for sequences of RVs, we can begin to appreciate the art of estimation in a more general setting. Parameter estimation is the basic problem in statistical inference and machine learning. We will formalize the general estimation problem here.
As we have already seen, when estimating the population mean, there are two basic types of estimators. In point estimation we are interested in estimating a particular point of interest that is supposed to belong to a set of points. In (confidence) set estimation, we are interested in estimating a set with a particular form that has a specified probability of "trapping" the particular point of interest from a set of points. Here, a point should be interpreted as an element of a collection of elements from some space.

19.1 Point and Set Estimation – in general

Point estimation is any statistical methodology that provides one with a "**single best guess**" of some specific quantity of interest. Traditionally, we denote this **quantity of interest as** θ^* and **its point estimate as** $\widehat{\theta}$ **or** $\widehat{\theta}_n$. The subscript n in the point estimate $\widehat{\theta}_n$ emphasizes that our estimate is based on n observations or data points from a given statistical experiment to estimate θ^*. This quantity of interest, which is usually unknown, can be:

- a **parameter** θ^* which is an element of the **parameter space** $\boldsymbol{\Theta}$, i.e. $\theta^* \in \boldsymbol{\Theta}$ such that θ^* specifies the "law" of the observations (realizations or samples) of the R$\vec{\text{V}}$ $(X_1, \ldots, X_n)$ modeled by JPDF or JPMF $f_{X_1,\ldots,X_n}(x_1, \ldots, x_n; \theta^*)$, or

- a **regression function** $\theta^* \in \boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is a class of regression functions in a regression experiment with model: $Y = \theta^*(X) + \epsilon$, such that $E(\epsilon) = 0$ and θ^* specifies the "law" of pairs of observations $\{(X_i, Y_i)\}_{i=1}^n$, for e.g., fitting parameters in noisy ODE or PDEs from observed data — one can always do a **prediction** in a regression experiment, i.e. when you want to estimate Y_i given X_i, or

- a **classifier** $\theta^* \in \boldsymbol{\Theta}$, i.e. a regression experiment with discrete $Y = \theta^*(X) + \epsilon$, for e.g. training an scrub-nurse robot to assist a human surgeon, or

- an **integral** $\theta^* := \int_A h(x)\, dx \in \boldsymbol{\Theta}$. If θ^* is finite, then $\boldsymbol{\Theta} = \mathbb{R}$, for e.g. θ^* could be the volume of a high-dimensional irregular polyhedron, a traffic congestion measure on a

network of roadways, the expected profit from a new brew of beer, or the probability of an extreme event such as the collapse of a dam in the Southern Alps in the next 150 years.

Set estimation is any statistical methodology that provides one with a "**best smallest set**", such as an interval, rectangle, ellipse, etc. that contains θ^* with a high probability $1-\alpha$. Recall that a statistic is a RV or R$\vec{\text{V}}$ $T(X)$ that maps every data point x in the data space $\mathbb{X}$ with $T(x) = t$ in its range $\mathbb{T}$, i.e. $T(x) : \mathbb{X} \to \mathbb{T}$ (Definition 16.5). Next, we look at a specific class of statistics whose range is the parameter space $\boldsymbol{\Theta}$.

Definition 19.1 (Point Estimator) A **point estimator** $\widehat{\Theta}$ of some **fixed and possibly unknown** $\theta^* \in \boldsymbol{\Theta}$ is a statistic that associates each data point $x \in \mathbb{X}$ with a **point estimate** $\widehat{\Theta}(x) = \widehat{\theta} \in \boldsymbol{\Theta}$,

$$\boxed{\widehat{\Theta} := \widehat{\Theta}(x) = \widehat{\theta} : \mathbb{X} \to \boldsymbol{\Theta}} \ .$$

If our data point $x := (x_1, x_2, \ldots, x_n)$ is an n-vector or a point in the n-dimensional real space, i.e. $x := (x_1, x_2, \ldots, x_n) \in \mathbb{X}_n \subset \mathbb{R}^n$, then we emphasize the dimension n in our point estimator $\widehat{\Theta}_n$ of $\theta^* \in \boldsymbol{\Theta}$.

$$\boxed{\widehat{\Theta}_n := \widehat{\Theta}_n(x := (x_1, x_2, \ldots, x_n)) = \widehat{\theta}_n : \mathbb{X}_n \to \boldsymbol{\Theta}, \quad \mathbb{X}_n \subset \mathbb{R}^n} \ .$$

The typical situation for us involves point **estimation** of $\theta^* \in \boldsymbol{\Theta}$ from $(x_1, x_2, \ldots, x_n)$, the **observed data** (realization or sample), based on the **model**

$$X = (X_1, X_2, \ldots, X_n) \sim f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n; \theta^*) \ .$$

Example 19.2 (Coin Tossing Experiment ($X_1, \ldots, X_n \overset{IID}{\sim}$ Bernoulli(θ^*))) I tossed a coin that has an unknown probability θ^* of landing Heads independently and identically 10 times in a row. Four of my outcomes were Heads and the remaining six were Tails, with the actual sequence of Bernoulli outcomes (Heads $\to$ 1 and Tails $\to$ 0) being $(1, 0, 0, 0, 1, 1, 0, 0, 1, 0)$. I would like to estimate the probability $\theta^* \in \boldsymbol{\Theta} = [0, 1]$ of observing Heads using the natural estimator $\widehat{\Theta}_n((X_1, X_2, \ldots, X_n))$ of θ^*:

$$\widehat{\Theta}_n((X_1, X_2, \ldots, X_n)) := \widehat{\Theta}_n = \frac{1}{n} \sum_{i=1}^{n} X_i =: \overline{X}_n$$

For the coin tossing experiment I just performed ($n = 10$ times), the point estimate of the unknown θ^* is:

$$\begin{aligned} \widehat{\theta}_{10} = \widehat{\Theta}_{10}((x_1, x_2, \ldots, x_{10})) &= \widehat{\Theta}_{10}((1, 0, 0, 0, 1, 1, 0, 0, 1, 0)) \\ &= \frac{1+0+0+0+1+1+0+0+1+0}{10} = \frac{4}{10} = 0.40 \ . \end{aligned}$$

19.2 Likelihood

We take a look at **likelihood** — one of the most fundamental concepts in Statistics.

Definition 19.3 (Likelihood Function) Suppose $(X_1, X_2, \ldots, X_n)$ is a R$\vec{\text{V}}$ with JPDF or JPMF $f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n; \theta)$ specified by parameter $\theta \in \boldsymbol{\Theta}$. Let the observed data be $(x_1, x_2, \ldots, x_n)$. Then the **likelihood** function given by $L_n(\theta)$ is merely the joint probability of the data, with the exception that we see it as a function of the parameter:

$$\boxed{L_n(\theta) := L_n(x_1, x_2, \ldots, x_n; \theta) = f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n; \theta) \ .} \tag{64}$$

The **log-likelihood** function is defined by:

$$\boxed{\ell_n(\theta) := \log(L_n(\theta))} \tag{65}$$

Example 19.4 (Likelihood of the IID Bernoulli(θ^*) **experiment)** Consider our IID Bernoulli experiment:

$$X_1, X_2, \ldots, X_n \overset{IID}{\sim} \text{Bernoulli}(\theta^*), \text{ with PDF } f_{X_i}(x_i;\theta) = \theta^{x_i}(1-\theta)^{1-x_i}\mathbb{1}_{\{0,1\}}(x_i), \text{ for } i \in \{1, 2, \ldots, n\} \ .$$

Let us understand the likelihood function for one observation first. There are two possibilities for the first observation.
If we only have one observation and it happens to be $x_1 = 1$, then our likelihood function is:

$$L_1(\theta) = L_1(x_1;\theta) = f_{X_1}(x_1;\theta) = \theta^1(1-\theta)^{1-1}\mathbb{1}_{\{0,1\}}(1) = \theta(1-\theta)^0 1 = \theta$$

If we only have one observation and it happens to be $x_1 = 0$, then our likelihood function is:

$$L_1(\theta) = L_1(x_1;\theta) = f_{X_1}(x_1;\theta) = \theta^0(1-\theta)^{1-0}\mathbb{1}_{\{0,1\}}(0) = 1(1-\theta)^1 1 = 1-\theta$$

If we have n observations $(x_1, x_2, \ldots, x_n)$, i.e. a vertex point of the unit hyper-cube $\{0,1\}^n$ (see top panel of Figure 19 when $n \in \{1, 2, 3\}$), then our likelihood function (see bottom panel of Figure 19) is obtained by multiplying the densities due to our IID assumption:

$$\begin{aligned} L_n(\theta) &:= L_n(x_1, x_2, \ldots, x_n;\theta) = f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n;\theta) \\ &= f_{X_1}(x_1;\theta)f_{X_2}(x_2;\theta)\cdots f_{X_n}(x_n;\theta) := \prod_{i=1}^{n} f_{X_i}(x_i;\theta) \\ &= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} \end{aligned} \tag{66}$$

Definition 19.5 (Maximum Likelihood Estimator (MLE)) Let the model for the data be

$$(X_1, \ldots, X_n) \sim f_{X_1,X_2,\ldots,X_n}(x_1, \ldots, x_n;\theta^*) \ .$$

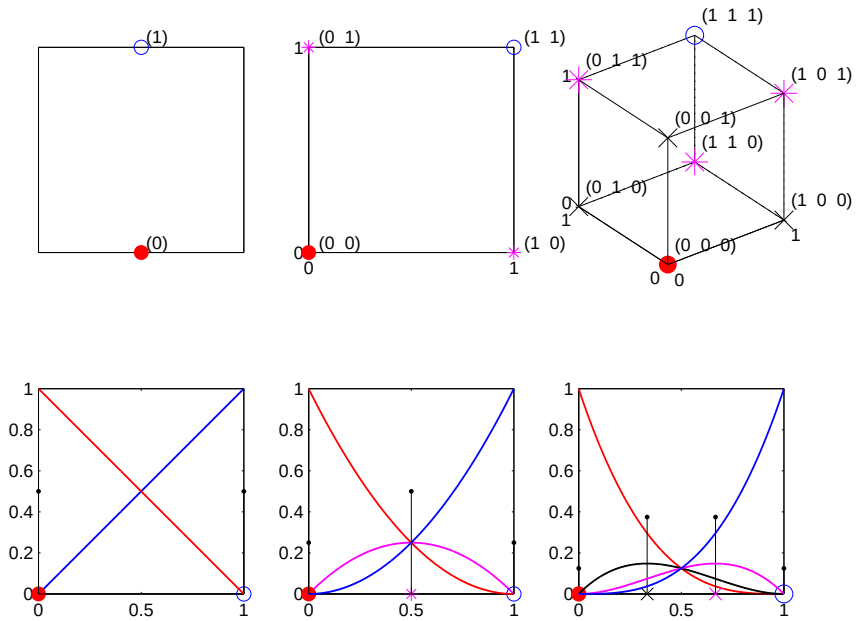
Then the maximum likelihood estimator (MLE) $\widehat{\Theta}_n$ of the fixed and possibly unknown parameter $\theta^* \in \boldsymbol{\Theta}$ is the value of θ that maximizes the likelihood function:

$$\boxed{\widehat{\Theta}_n := \widehat{\Theta}_n(X_1, X_2, \ldots, X_n) := \underset{\theta\in\boldsymbol{\Theta}}{\operatorname{argmax}}\, L_n(\theta) \ ,}$$

Equivalently, MLE is the value of θ that maximizes the log-likelihood function (since $\log = \log_e = \ln$ is a monotone increasing function):

$$\boxed{\widehat{\Theta}_n := \underset{\theta\in\boldsymbol{\Theta}}{\operatorname{argmax}}\, \ell_n(\theta) \ ,}$$

Figure 19: Data Spaces $\mathbb{X}_1 = \{0,1\}$, $\mathbb{X}_2 = \{0,1\}^2$ and $\mathbb{X}_3 = \{0,1\}^3$ for one, two and three IID Bernoulli trials, respectively and the corresponding likelihood functions.



Useful Properties of the Maximum Likelihood Estimator

1. The ML Estimator is *asymptotically consistent* (gives the "true" θ^* as sample size $n \to \infty$):

$$\boxed{\widehat{\Theta}_n \rightsquigarrow \text{Point Mass}(\theta^*)}$$

2. The ML Estimator is asymptotically normal (has a normal distribution concentrating on θ^* as $n \to \infty$):

$$\boxed{\widehat{\Theta}_n \rightsquigarrow \text{Normal}(\theta^*, (\widehat{\mathsf{se}}_n)^2)}$$

or equivalently:

$$\boxed{(\widehat{\Theta}_n - \theta^*)/\widehat{\mathsf{se}}_n \rightsquigarrow \text{Normal}(0,1)}$$

where $\widehat{\mathsf{se}}_n$ is the **estimated standard error**, i.e. the standard deviation of $\widehat{\Theta}_n$, and it is given by the square-root of the inverse negative curvature of $\ell_n(\theta)$ at $\widehat{\theta}_n$:

$$\boxed{\widehat{\mathsf{se}}_n = \sqrt{\left(\left[-\frac{d^2\ell_n(\theta)}{d\theta^2}\right]_{\theta=\widehat{\theta}_n}\right)^{-1}}}$$

3. Because of the previous two properties, the $1-\alpha$ confidence interval is:

$$\boxed{\widehat{\Theta}_n \pm z_{\alpha/2}\widehat{\mathsf{se}}_n}$$

MLE is a general methodology for parameter estimation in an essentially arbitrary parameter space $\boldsymbol{\Theta}$ that is defining or indexing the laws in a parametric family of models, although we are only seeing it in action when $\boldsymbol{\Theta} \subset \mathbb{R}$ for simplest parametric family of models involving IID product experiments here. When $\boldsymbol{\Theta} \subset \mathbb{R}^d$ with $2 \leq d < \infty$ then MLE $\widehat{\Theta}_n \rightsquigarrow$ Point Mass(θ^*), where $\theta^* = (\theta_1^*, \theta_2^*, \ldots, \theta_d^*)^T$ is a column vector in $\boldsymbol{\Theta} \subset \mathbb{R}^d$ and $\widehat{\Theta}_n \rightsquigarrow$ Normal $\left(\theta^*, \widehat{\Sigma(\mathsf{se})}_n\right)$, a multivariate Normal distribution with mean vector θ^* and variance-covariance matrix of standard errors given by the *Hessian* (a $d \times d$ matrix of mixed partial derivatives) of $\ell_n(\theta_1, \theta_2, \ldots, \theta_d)$. The ideas in the cased of dimension $d = 1$ naturally generalize to an arbitrary, but finite, dimension d.

Remark 19.6 In order to use MLE for parameter estimation we need to ensure that the following two conditions hold:

1. The *support* of the data, i.e. the set of possible values of $(X_1, X_2, \ldots, X_n)$ must not depend on θ for every $\theta \in \boldsymbol{\Theta}$ — of course the probabilities do depend on θ in an *identifiable* manner, i.e. for every θ and ϑ in $\boldsymbol{\Theta}$, if $\theta \neq \vartheta$ then $f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n; \theta) \neq f_{X_1,\ldots,X_n}(x_1, x_2, \ldots, x_n; \vartheta)$ at least for some $(x_1, x_2, \ldots, x_n) \in \mathbb{X}$.

2. If the parameter space $\boldsymbol{\Theta}$ is bounded then θ^* must not belong to the boundaries of $\boldsymbol{\Theta}$.

Maximum Likelihood Estimation Method in Six Easy Steps

Background: We have observed data:

$$(x_1, x_2, \ldots, x_n)$$

which is modeled as a sample or realization from the random vector:

$$(X_1, X_2, \ldots, X_n) \sim f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n; \theta^*), \quad \theta^* \in \boldsymbol{\Theta} \ .$$

Objective: We want to obtain an estimator $\widehat{\Theta}_n$ that will give:

1. the point estimate $\widehat{\theta}_n$ of the "true" parameter θ^* and

2. the $(1 - \alpha)$ confidence interval for θ^*.

Steps of MLE:

- Step 1: Find the expression for the log likelihood function:

$$\ell_n(\theta) = \log(L_n(\theta)) = \log\left(f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n; \theta)\right) \ .$$

  Note that if the model assumes that $(X_1, X_2, \ldots, X_n)$ is jointly independent, i.e. we have an independent and identically distributed (IID) experiment, then $\ell_n(\theta)$ simplifies further as follows:

$$\ell_n(\theta) = \log(L_n(\theta)) = \log\left(f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n; \theta)\right) = \log\left(\prod_{i=1}^{n} f_{X_i}(x_i; \theta)\right) \ .$$

- Step 2: Obtain the derivative of $\ell_n(\theta)$ with respect to θ:

$$\frac{d}{d\theta}\left(\ell_n(\theta)\right) \ .$$

- Step 3: Set the derivative equal to zero, solve for θ and let $\widehat{\theta}_n$ equal to this solution.
- Step 4: Check if this solution is indeed a maximum of $\ell_n(\theta)$ by checking if:

$$\frac{d^2}{d\theta^2}\left(\ell_n(\theta)\right) < 0 \ .$$

- Step 5: If $\frac{d^2}{d\theta^2}\left(\ell_n(\theta)\right) < 0$ then you have found the maximum likelihood estimate $\widehat{\theta}_n$.
- Step 6: If you also want the $(1-\alpha)$ confidence interval then get it from

$$\widehat{\theta}_n \pm z_{\alpha/2}\widehat{\mathsf{se}}_n \quad \text{, where} \quad \widehat{\mathsf{se}}_n = \sqrt{\left(\left[-\frac{d^2\ell_n(\theta)}{d\theta^2}\right]_{\theta=\widehat{\theta}_n}\right)^{-1}} \ .$$

Let us apply this method in some examples.

Example 19.7 Find (or derive) the maximum likelihood estimate $\widehat{\lambda}_n$ and the $(1-\alpha)$ confidence interval of the fixed and possibly unknown parameter λ^* for the IID experiment:

$$X_1, \ldots, X_n \overset{IID}{\sim} \text{Exponential}(\lambda^*), \qquad \lambda^* \in \mathbf{\Lambda} = (0, \infty) \ .$$

Note that $\mathbf{\Lambda}$ is the parameter space.
We first obtain the log-likelihood function $\ell_n(\theta)$ given data $(x_1, x_2, \ldots, x_n)$.

$$\begin{aligned}
\ell_n(\lambda) &:= \log(L(x_1, x_2, \ldots, x_n; \lambda)) = \log\left(\prod_{i=1}^n f_{X_i}(x_i;\lambda)\right) = \log\left(\prod_{i=1}^n \lambda e^{-\lambda x_i}\right) \\
&= \log\left(\lambda e^{-\lambda x_1} \cdot \lambda e^{-\lambda x_2} \cdots \lambda e^{-\lambda x_n}\right) = \log\left(\lambda^n e^{-\lambda \sum_{i=1}^n x_i}\right) = \log\left(\lambda^n\right) + \log\left(e^{-\lambda \sum_{i=1}^n x_i}\right) \\
&= \boxed{\log\left(\lambda^n\right) - \lambda \sum_{i=1}^n x_i}
\end{aligned}$$

Now, let us take the derivative with respect to λ,

$$\begin{aligned}
\frac{d}{d\lambda}\left(\ell_n(\lambda)\right) &:= \frac{d}{d\lambda}\left(\log\left(\lambda^n\right) - \lambda \sum_{i=1}^n x_i\right) = \frac{d}{d\lambda}\left(\log\left(\lambda^n\right)\right) - \frac{d}{d\lambda}\left(\lambda \sum_{i=1}^n x_i\right) = \frac{1}{\lambda^n}\frac{d}{d\lambda}\left(\lambda^n\right) - \sum_{i=1}^n x_i \\
&= \frac{1}{\lambda^n} n\lambda^{n-1} - \sum_{i=1}^n x_i = \boxed{\frac{n}{\lambda} - \sum_{i=1}^n x_i} \ .
\end{aligned}$$

Next, we set the derivative to 0, solve for λ, and let the solution equal to the ML estimate $\widehat{\lambda}_n$.

$$0 = \frac{d}{d\lambda}\left(\ell_n(\lambda)\right) \iff 0 = \frac{n}{\lambda} - \sum_{i=1}^n x_i \iff \sum_{i=1}^n x_i = \frac{n}{\lambda} \iff \lambda = \frac{n}{\sum_{i=1}^n x_i} \quad \text{and let } \boxed{\widehat{\lambda}_n = \frac{1}{\overline{x}_n}} \ .$$

Next, we find the second derivative and check if it is negative.

$$\frac{d^2}{d\lambda^2}\left(\ell_n(\lambda)\right) = \frac{d}{d\lambda}\left(\frac{d}{d\lambda}\left(\ell_n(\lambda)\right)\right) = \frac{d}{d\lambda}\left(\frac{n}{\lambda} - \sum_{i=1}^n x_i\right) = \boxed{-n\lambda^{-2}} \ .$$

Since $\lambda > 0$ and $n \in \mathbb{N}$, $\boxed{-n\lambda^{-2} = -n/\lambda^2 < 0}$, so we have found the maximum likelihood estimate:

$$\boxed{\widehat{\lambda}_n = \frac{1}{\overline{x}_n}} \ .$$

Now, let us find the estimated standard error:

$$\boxed{\widehat{\mathsf{se}}_n} = \sqrt{\left(\left[-\frac{d^2\ell_n(\lambda)}{d\lambda^2}\right]_{\lambda=\widehat{\lambda}_n}\right)^{-1}} = \sqrt{\left(\left[-\left(-\frac{n}{\lambda^2}\right)\right]_{\lambda=\widehat{\lambda}_n}\right)^{-1}} = \sqrt{\left(\frac{n}{\widehat{\lambda}_n^2}\right)^{-1}} = \sqrt{\frac{\widehat{\lambda}_n^2}{n}} = \frac{\widehat{\lambda}_n}{\sqrt{n}}$$
$$= \boxed{\frac{1}{\overline{x}_n\sqrt{n}}} \ .$$

And finally, the $(1-\alpha)$ confidence interval is

$$\widehat{\lambda}_n \pm z_{\alpha/2}\widehat{\mathsf{se}}_n = \boxed{\frac{1}{\overline{x}_n} \pm z_{\alpha/2}\frac{1}{\overline{x}_n\sqrt{n}}} \ .$$

Since we have worked "hard" to get the maximum likelihood estimate for a general IID model $X_1, X_2, \ldots, X_n \overset{IID}{\sim}$ Exponential(λ^*). Let us kill two birds with the same stone by applying it to two datasets:

1. Orbiter waiting times and
2. Time between measurable earthquakes in New Zealand over a few months.

Therefore, the ML estimate $\widehat{\lambda}_n$ of the unknown rate parameter $\lambda^* \in \mathbf{\Lambda}$ on the basis of n IID observations $x_1, x_2, \ldots, x_n \overset{IID}{\sim}$ Exponential(λ^*) is $1/\overline{x}_n$ and the ML estimator $\widehat{\Lambda}_n = 1/\overline{X}_n$. Let us apply this ML estimator of the rate parameter for the supposedly exponentially distributed waiting times at the on-campus Orbiter bus-stop.

Orbiter Waiting Times

Joshua Fenemore and Yiran Wang collected data on waiting times between buses at an Orbiter bus-stop close to campus. They collected a sample of size $n = 132$ with sample mean $\overline{x}_{132} = 9.0758$. From our work in Example 19.7 we can now easily obtain the maximum likelihood estimate of λ^* and the 95% confidence interval for it, under the assumption that the waiting times $X_1, \ldots, X_{132}$ are IID Exponential(λ^*) RVs as follows:

$$\widehat{\lambda}_{132} = 1/\overline{x}_{132} = 1/9.0758 = 0.1102 \quad (0.1102 \pm 1.96 \times 0.1102/\sqrt{132}) = (0.0914, 0.1290) \ ,$$

and thus the estimated mean waiting time is

$$1/\widehat{\lambda}_{132} = 9.0763 \text{ minutes.}$$

The estimated mean waiting time for a bus to arrive is well within the 10 minutes promised by the Orbiter bus company. This data and its maximum likelihood analysis is presented visually in Figure 20.

Notice how the exponential PDF $f(x; \widehat{\lambda}_{132} = 0.1102)$ and the DF $F(x; \widehat{\lambda}_{132} = 0.1102)$ based on the MLE fits with the histogram and the empirical DF, respectively.

Waiting Times between Earth Quakes in NZ:

Once again from our work in Example 19.7 we can now easily obtain the maximum likelihood estimate of λ^* and the 95% confidence interval for it, under the assumption that the waiting

Figure 20: Plot of $\log(L(\lambda))$ as a function of the parameter λ and the MLE $\widehat{\lambda}_{132}$ of 0.1102 for Fenemore-Wang Orbiter Waiting Times Experiment from STAT 218 S2 2007. The density or PDF and the DF at the MLE of 0.1102 are compared with a histogram and the empirical DF.

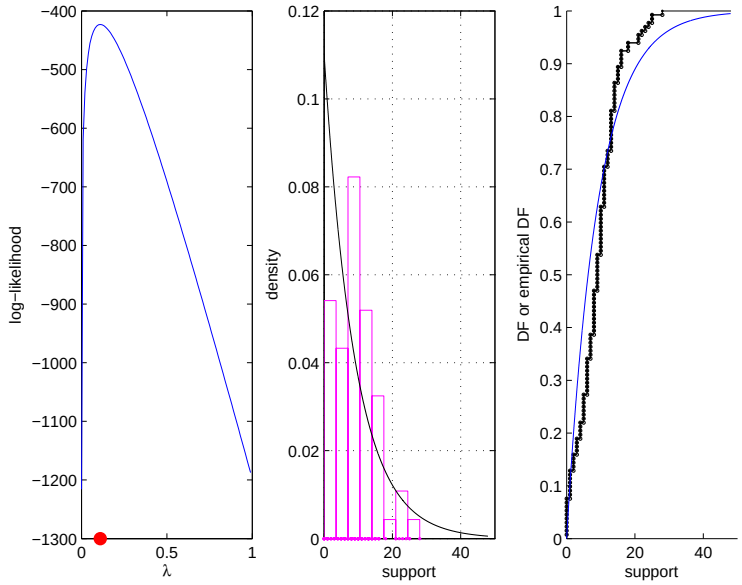
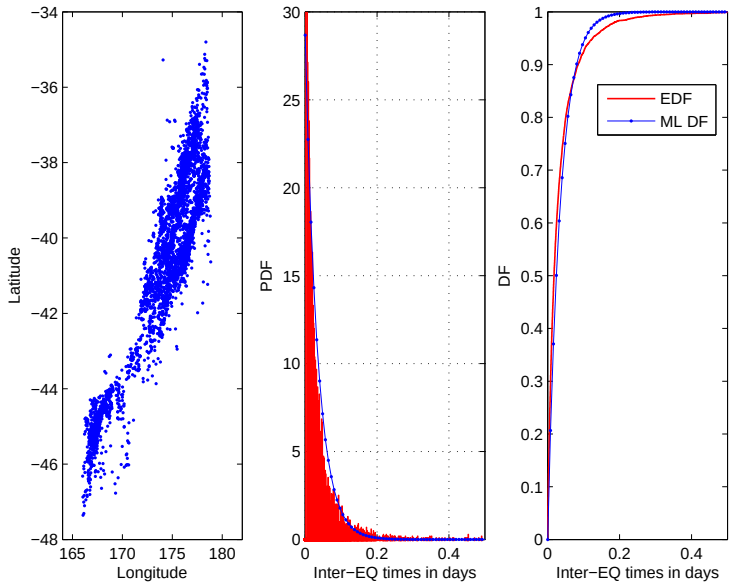


Figure 21: Comparing the Exponential($\widehat{\lambda}_{6128} = 28.6694$) PDF and DF with a histogram and empirical DF of the times (in units of days) between earth quakes in NZ. The epicenters of 6128 earth quakes are shown in left panel.



times (in days) between the 6128 measurable earth-quakes in NZ from 18-Jan-2008 02:23:44 to 18-Aug-2008 19:29:29 are IID Exponential(λ^*) RVs as follows:

$$\widehat{\lambda}_{6128} = 1/\overline{x}_{6128} = 1/0.0349 = 28.6694 \quad (28.6694 \pm 1.96 \times 28.6694/\sqrt{6128}) = (27.95, 29.39) \ ,$$

and thus the estimated mean time in days and minutes between earth quakes (somewhere in NZ over the first 8 months in 2008) is

$$1/\widehat{\lambda}_{6128} = \overline{x}_{6128} = 0.0349 \text{ days} \quad = \quad 0.0349 * 24 * 60 = 50.2560 \text{ minutes} \ .$$

This data and its maximum likelihood analysis is presented visually in Figure 21. The PDF and DF corresponding to the $\widehat{\lambda}_{6128}$ (blue curves in Figure 21) are the best fitting PDF and DF from the parametric family of PDFs in $\{\lambda e^{-\lambda x} : \lambda \in (0, \infty)\}$ and DFs in $\{1 - e^{-\lambda x} : \lambda \in (0, \infty)\}$ to the density histogram and the empirical distribution function given by the data, respectively. Clearly, there is room for improving beyond the model of IID Exponential(λ) RVs, but the fit with just one real-valued parameter is not too bad either. Finally, with the best fitting PDF $28.6694e^{-28.6694x}$ we can get probabilities of events and answer questions like: "what is the probability that there will be three earth quakes somewhere in NZ within the next hour?", etc.

Example 19.8 (ML Estimation for the IID Bernoulli(θ^*) **experiment)** Let us do maximum likelihood estimation for the coin-tossing experiment of Example 19.2 with likelihood derived in Example 19.4 to obtain the maximum likelihood estimate $\widehat{\theta}_n$ of the unknown parameter $\theta^* \in \boldsymbol{\Theta} = [0, 1]$ and the $(1 - \alpha)$ confidence interval for it.
From Equation (66) the log likelihood function is

$$\ell_n(\theta) = \log(L_n(\theta)) = \log\left(\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}\right) = \boxed{\left(\sum_{i=1}^n x_i\right)\log(\theta) + \left(n - \sum_{i=1}^n x_i\right)\log(1-\theta)} \ ,$$

Next, we take the derivative with respect to the parameter θ:

$$\frac{d}{d\theta}\left(\ell_n(\theta)\right) \quad = \quad \frac{d}{d\theta}\left(\left(\sum_{i=1}^n x_i\right)\log(\theta)\right) + \frac{d}{d\theta}\left(\left(n - \sum_{i=1}^n x_i\right)\log(1-\theta)\right) = \boxed{\frac{\sum_{i=1}^n x_i}{\theta} - \frac{n - \sum_{i=1}^n x_i}{1-\theta}} \ .$$

Now, set $\frac{d}{d\theta}\log(L_n(\theta)) = 0$, solve for θ and set the solution equal to $\widehat{\theta}_n$:

$$\frac{d}{d\theta}\left(\ell_n(\theta)\right) = 0 \iff \frac{\sum_{i=1}^n x_i}{\theta} = \frac{n - \sum_{i=1}^n x_i}{1-\theta} \iff \frac{1-\theta}{\theta} = \frac{n - \sum_{i=1}^n x_i}{\sum_{i=1}^n x_i}$$

$$\iff \frac{1}{\theta} - 1 = \frac{n}{\sum_{i=1}^n x_i} - 1 \quad \text{let } \boxed{\widehat{\theta}_n = \frac{\sum_{i=1}^n x_i}{n}}$$

Next, we find the second derivative and check if it is negative.

$$\frac{d^2}{d\theta^2}(\ell_n(\theta)) = \frac{d}{d\theta}\left(\frac{\sum_{i=1}^n x_i}{\theta} - \frac{n - \sum_{i=1}^n x_i}{1-\theta}\right) = \boxed{-\frac{\sum_{i=1}^n x_i}{\theta^2} - \frac{n - \sum_{i=1}^n x_i}{(1-\theta)^2}}$$

Since each term in the numerator and the denominator of the two fractions in the above box are non-negative, $\frac{d^2}{d\theta^2}(\ell_n(\theta)) < 0$ and therefore we have found the maximum likelihood estimate

$$\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^n x_i = \overline{x}_n$$

We already knew this to be a point estimate for $E(X_i) = \theta^*$ from LLN and CLT. But now we know that MLE also agrees. Now, let us find the estimated standard error:

$$\boxed{\widehat{\mathsf{se}}_n} = \sqrt{\left(\left[-\frac{d^2\ell_n(\theta)}{d\theta^2}\right]_{\theta=\widehat{\theta}_n}\right)^{-1}} = \sqrt{\left(\left[-\left(-\frac{\sum_{i=1}^n x_i}{\theta^2} - \frac{n-\sum_{i=1}^n x_i}{(1-\theta)^2}\right)\right]_{\theta=\widehat{\theta}_n}\right)^{-1}}$$

$$= \sqrt{\left(\frac{\sum_{i=1}^n x_i}{\widehat{\theta}_n^2} + \frac{n-\sum_{i=1}^n x_i}{(1-\widehat{\theta}_n)^2}\right)^{-1}} = \sqrt{\left(\frac{n\overline{x}_n}{\overline{x}_n^2} + \frac{n-n\overline{x}_n}{(1-\overline{x}_n)^2}\right)^{-1}} = \sqrt{\left(\frac{n}{\overline{x}_n} + \frac{n}{(1-\overline{x}_n)}\right)^{-1}}$$

$$= \sqrt{\left(\frac{n(1-\overline{x}_n)+n\overline{x}_n}{\overline{x}_n(1-\overline{x}_n)}\right)^{-1}} = \sqrt{\frac{\overline{x}_n(1-\overline{x}_n)}{n((1-\overline{x}_n)+\overline{x}_n}} = \boxed{\sqrt{\frac{\overline{x}_n(1-\overline{x}_n)}{n}}} \ .$$

.

And finally, the $(1-\alpha)$ confidence interval is

$$\widehat{\theta}_n \pm z_{\alpha/2}\widehat{\mathsf{se}}_n = \boxed{\overline{x}_n \pm z_{\alpha/2}\sqrt{\frac{\overline{x}_n(1-\overline{x}_n)}{n}}} \ .$$

For the coin tossing experiment that was performed ($n = 10$ times) in Example 19.2, the maximum likelihood estimate of θ^* and the 95% confidence interval for it, under the model that the tosses are IID Bernoulli(θ^*) RVs, are as follows:

$$\widehat{\theta}_{10} = \overline{x}_{10} = \frac{4}{10} = 0.40 \quad \text{and} \quad \left(0.4 \pm 1.96 \times \sqrt{\frac{0.4 \times 0.6}{10}}\right) = (0.0964, 0.7036) \ .$$

See Figures 22 and 23 to completely understand parameter estimation for IID Bernoulli experiments.

Figure 22: Plots of the log likelihood $\ell_n(\theta) = \log(L(1,0,0,0,1,1,0,0,1,0;\theta))$ as a function of the parameter θ over the parameter space $\Theta = [0,1]$ and the MLE $\widehat{\theta}_{10}$ of 0.4 for the coin-tossing experiment shown in standard scale (left panel) and log scale for x-axis (right panel).

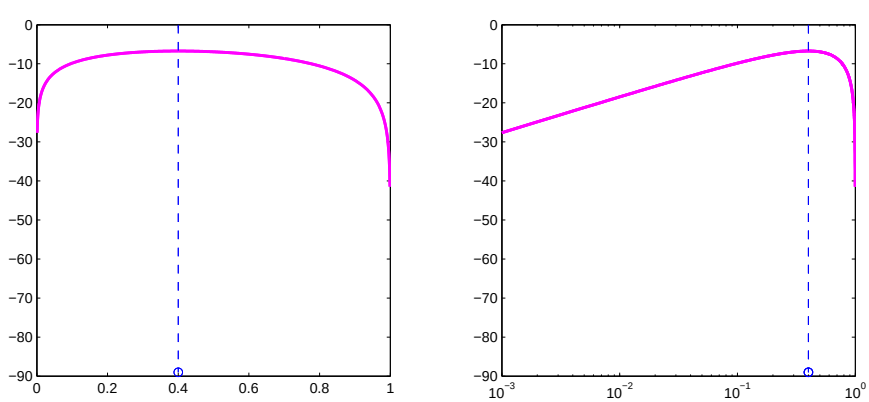
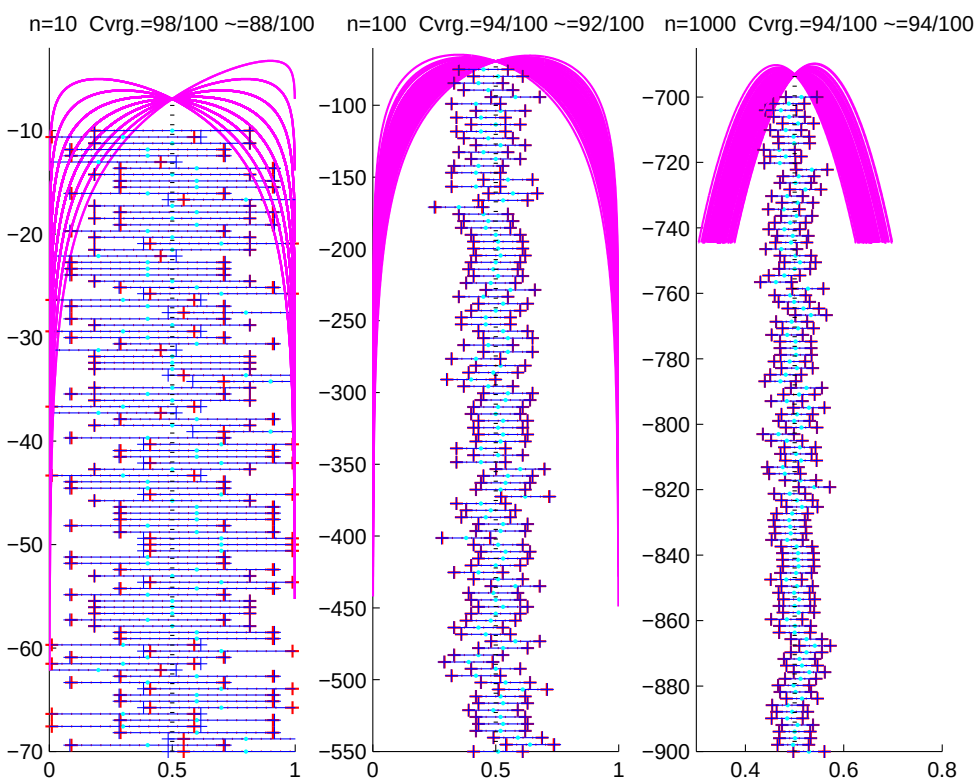


Figure 23: 100 realizations of 95% confidence intervals based on samples of size $n = 10$, 100 and 1000 simulated from IID Bernoulli($\theta^* = 0.5$) RVs. The MLE $\widehat{\theta}_n$ (cyan dot) and the log-likelihood function (magenta curve) for each of the 100 replications of the experiment for each sample size n are depicted. The approximate normal-based 95% confidence intervals with blue boundaries are based on the exact $\mathsf{se}_n = \sqrt{\theta^*(1-\theta^*)/n} = \sqrt{1/4}$, while those with red boundaries are based on the estimated $\widehat{\mathsf{se}}_n = \sqrt{\widehat{\theta}_n(1-\widehat{\theta}_n)/n} = \sqrt{\frac{\overline{x}_n(1-\overline{x}_n)}{n}}$. The fraction of times the true parameter $\theta^* = 0.5$ was contained by the exact and approximate confidence interval (known as *empirical coverage*) over the 100 replications of the simulation experiment for each of the three sample sizes are given by the numbers after `Cvrg.=` and ∼=, above each sub-plot, respectively.



20 Standard normal distribution function table

For any given value z, its cumulative probability $\Phi(z)$ was generated by `Excel` formula `NORMSDIST`, as `NORMSDIST`(z).

z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$
0.01	0.5040	0.51	0.6950	1.01	0.8438	1.51	0.9345	2.01	0.9778	2.51	0.9940
0.02	0.5080	0.52	0.6985	1.02	0.8461	1.52	0.9357	2.02	0.9783	2.52	0.9941
0.03	0.5120	0.53	0.7019	1.03	0.8485	1.53	0.9370	2.03	0.9788	2.53	0.9943
0.04	0.5160	0.54	0.7054	1.04	0.8508	1.54	0.9382	2.04	0.9793	2.54	0.9945
0.05	0.5199	0.55	0.7088	1.05	0.8531	1.55	0.9394	2.05	0.9798	2.55	0.9946
0.06	0.5239	0.56	0.7123	1.06	0.8554	1.56	0.9406	2.06	0.9803	2.56	0.9948
0.07	0.5279	0.57	0.7157	1.07	0.8577	1.57	0.9418	2.07	0.9808	2.57	0.9949
0.08	0.5319	0.58	0.7190	1.08	0.8599	1.58	0.9429	2.08	0.9812	2.58	0.9951
0.09	0.5359	0.59	0.7224	1.09	0.8621	1.59	0.9441	2.09	0.9817	2.59	0.9952
0.10	0.5398	0.60	0.7257	1.10	0.8643	1.60	0.9452	2.10	0.9821	2.60	0.9953
0.11	0.5438	0.61	0.7291	1.11	0.8665	1.61	0.9463	2.11	0.9826	2.61	0.9955
0.12	0.5478	0.62	0.7324	1.12	0.8686	1.62	0.9474	2.12	0.9830	2.62	0.9956
0.13	0.5517	0.63	0.7357	1.13	0.8708	1.63	0.9484	2.13	0.9834	2.63	0.9957
0.14	0.5557	0.64	0.7389	1.14	0.8729	1.64	0.9495	2.14	0.9838	2.64	0.9959
0.15	0.5596	0.65	0.7422	1.15	0.8749	1.65	0.9505	2.15	0.9842	2.65	0.9960
0.16	0.5636	0.66	0.7454	1.16	0.8770	1.66	0.9515	2.16	0.9846	2.66	0.9961
0.17	0.5675	0.67	0.7486	1.17	0.8790	1.67	0.9525	2.17	0.9850	2.67	0.9962
0.18	0.5714	0.68	0.7517	1.18	0.8810	1.68	0.9535	2.18	0.9854	2.68	0.9963
0.19	0.5753	0.69	0.7549	1.19	0.8830	1.69	0.9545	2.19	0.9857	2.69	0.9964
0.20	0.5793	0.70	0.7580	1.20	0.8849	1.70	0.9554	2.20	0.9861	2.70	0.9965
0.21	0.5832	0.71	0.7611	1.21	0.8869	1.71	0.9564	2.21	0.9864	2.71	0.9966
0.22	0.5871	0.72	0.7642	1.22	0.8888	1.72	0.9573	2.22	0.9868	2.72	0.9967
0.23	0.5910	0.73	0.7673	1.23	0.8907	1.73	0.9582	2.23	0.9871	2.73	0.9968
0.24	0.5948	0.74	0.7704	1.24	0.8925	1.74	0.9591	2.24	0.9875	2.74	0.9969
0.25	0.5987	0.75	0.7734	1.25	0.8944	1.75	0.9599	2.25	0.9878	2.75	0.9970
0.26	0.6026	0.76	0.7764	1.26	0.8962	1.76	0.9608	2.26	0.9881	2.76	0.9971
0.27	0.6064	0.77	0.7794	1.27	0.8980	1.77	0.9616	2.27	0.9884	2.77	0.9972
0.28	0.6103	0.78	0.7823	1.28	0.8997	1.78	0.9625	2.28	0.9887	2.78	0.9973
0.29	0.6141	0.79	0.7852	1.29	0.9015	1.79	0.9633	2.29	0.9890	2.79	0.9974
0.30	0.6179	0.80	0.7881	1.30	0.9032	1.80	0.9641	2.30	0.9893	2.80	0.9974
0.31	0.6217	0.81	0.7910	1.31	0.9049	1.81	0.9649	2.31	0.9896	2.81	0.9975
0.32	0.6255	0.82	0.7939	1.32	0.9066	1.82	0.9656	2.32	0.9898	2.82	0.9976
0.33	0.6293	0.83	0.7967	1.33	0.9082	1.83	0.9664	2.33	0.9901	2.83	0.9977
0.34	0.6331	0.84	0.7995	1.34	0.9099	1.84	0.9671	2.34	0.9904	2.84	0.9977
0.35	0.6368	0.85	0.8023	1.35	0.9115	1.85	0.9678	2.35	0.9906	2.85	0.9978
0.36	0.6406	0.86	0.8051	1.36	0.9131	1.86	0.9686	2.36	0.9909	2.86	0.9979
0.37	0.6443	0.87	0.8078	1.37	0.9147	1.87	0.9693	2.37	0.9911	2.87	0.9979
0.38	0.6480	0.88	0.8106	1.38	0.9162	1.88	0.9699	2.38	0.9913	2.88	0.9980
0.39	0.6517	0.89	0.8133	1.39	0.9177	1.89	0.9706	2.39	0.9916	2.89	0.9981
0.40	0.6554	0.90	0.8159	1.40	0.9192	1.90	0.9713	2.40	0.9918	2.90	0.9981
0.41	0.6591	0.91	0.8186	1.41	0.9207	1.91	0.9719	2.41	0.9920	2.91	0.9982
0.42	0.6628	0.92	0.8212	1.42	0.9222	1.92	0.9726	2.42	0.9922	2.92	0.9982
0.43	0.6664	0.93	0.8238	1.43	0.9236	1.93	0.9732	2.43	0.9925	2.93	0.9983
0.44	0.6700	0.94	0.8264	1.44	0.9251	1.94	0.9738	2.44	0.9927	2.94	0.9984
0.45	0.6736	0.95	0.8289	1.45	0.9265	1.95	0.9744	2.45	0.9929	2.95	0.9984
0.46	0.6772	0.96	0.8315	1.46	0.9279	1.96	0.9750	2.46	0.9931	2.96	0.9985
0.47	0.6808	0.97	0.8340	1.47	0.9292	1.97	0.9756	2.47	0.9932	2.97	0.9985
0.48	0.6844	0.98	0.8365	1.48	0.9306	1.98	0.9761	2.48	0.9934	2.98	0.9986
0.49	0.6879	0.99	0.8389	1.49	0.9319	1.99	0.9767	2.49	0.9936	2.99	0.9986
0.50	0.6915	1.00	0.8413	1.50	0.9332	2.00	0.9772	2.50	0.9938	3.00	0.9987

SET SUMMARY

$\{a_1, a_2, \ldots, a_n\}$	–	a set containing the elements, $a_1, a_2, \ldots, a_n$.
$a \in A$	–	a is an element of the set A.
$A \subseteq B$	–	the set A is a subset of B.
$A \cup B$	–	"union", meaning the set of all elements which are in A or B, or both.
$A \cap B$	–	"intersection", meaning the set of all elements in both A and B.
$\{\}$ or $\emptyset$	–	empty set.
Ω	–	universal set.
A^c	–	the complement of A, meaning the set of all elements in Ω, the universal set, which are not in A.

EXPERIMENT SUMMARY

Experiment	–	an activity producing distinct outcomes.
Ω	–	set of all outcomes of the experiment.
ω	–	an individual outcome in Ω, called a simple event.
$A \subseteq \Omega$	–	a subset A of Ω is an event.
Trial	–	one performance of an experiment resulting in 1 outcome.

PROBABILITY SUMMARY

Axioms:

1. If $A \subseteq \Omega$ then $0 \leq P(A) \leq 1$ and $P(\Omega) = 1$.
2. If A, B are disjoint events, then $P(A \cup B) = P(A) + P(B)$.
   [This is true only when A and B are disjoint.]
3. If $A_1, A_2, \ldots$ are disjoint then $P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \ldots$

Rules:

$$P(A^c) \;=\; 1 - P(A)$$

$$P(A \cup B) \;=\; P(A) + P(B) - P(A \cap B) \qquad \text{[always true]}$$

CONDITIONAL PROBABILITY SUMMARY

$P(A|B)$ means the probability that A occurs given that B has occurred.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B|A)}{P(B)} \quad \text{if} \quad P(B) \neq 0$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(B)P(A|B)}{P(A)} \quad \text{if} \quad P(A) \neq 0$$

Conditional probabilities obey the 4 axioms of probability.

DISCRETE RANDOM VARIABLE SUMMARY

Probability mass function

$$f(x) = P(X = x_i)$$

Distribution function

$$F(x) = \sum_{x_i \leq x} f(x_i)$$

Random Variable	**Possible Values**	**Probabilities**	**Modeled situations**
Discrete uniform	$\{x_1, x_2, \ldots, x_k\}$	$P(X = x_i) = \frac{1}{k}$	Situations with k equally likely values. Parameter: k.
Bernoulli(θ)	$\{0, 1\}$	$P(X = 0) = 1 - \theta$ $P(X = 1) = \theta$	Situations with only 2 outcomes, coded 1 for success and 0 for failure. Parameter: $\theta = P(\text{success}) \in (0, 1)$.
Geometric(θ)	$\{1, 2, 3, \ldots\}$	$P(X = x)$ $= (1 - \theta)^{x-1}\theta$	Situations where you count the number of trials until the first success in a sequence of independent trails with a constant probability of success. Parameter: $\theta = P(\text{success}) \in (0, 1)$.
Binomial(n, θ)	$\{0, 1, 2, \ldots, n\}$	$P(X = x)$ $= \binom{n}{x}\theta^x(1-\theta)^{n-x}$	Situations where you count the number of success in n trials where each trial is independent and there is a constant probability of success. Parameters: $n \in \{1, 2, \ldots\}$; $\theta = P(\text{success}) \in (0, 1)$.
Poisson(λ)	$\{0, 1, 2, \ldots\}$	$P(X = x)$ $= \frac{\lambda^x e^{-\lambda}}{x!}$	Situations where you count the number of events in a continuum where the events occur one at a time and are independent of one another. Parameter: λ= rate $\in (0, \infty)$.

CONTINUOUS RANDOM VARIABLES: NOTATION

$f(x)$: Probability density function (PDF)

- $f(x) \ \geq \ 0$
- Areas underneath $f(x)$ measure probabilities.

$F(x)$: Distribution function (DF)

- $0 \ \leq \ \ F(x) \ \leq \ 1$
- $F(x) \ = \ P(X \leq x)$ is a probability
- $F'(x) \ = \ f(x)$ for every x where $f(x)$ is continuous
- $F(x) \ = \ \displaystyle\int_{-\infty}^{x} f(v)dv$
- $P(a < X \leq b) \ = \ F(b) - F(a) \ = \ \displaystyle\int_{a}^{b} f(v)dv$

Expectation of a function $g(X)$ of a random variable X is defined as:

$$E(g(X)) \ = \ \begin{cases} \displaystyle\sum_{x} g(x)f(x) & \text{if } X \text{ is a discrete RV} \\ \displaystyle\int_{-\infty}^{\infty} g(x)f(x)dx & \text{if } X \text{ is a continuous RV} \end{cases}$$

Some Common Expectations

$g(x)$	definition	also known as
x	$E(X)$	Expectation, Population Mean or First Moment of X
$(x - E(X))^2$	$V(X) := E((X - E(X))^2)$ $= E(X^2) - (E(X))^2$	Variance or Population Variance of X
$e^{\imath tx}$	$\phi_X(t) := E\left(e^{\imath tX}\right)$	Characteristic Function (CF) of X
x^k	$E(X^k) = \frac{1}{\imath^k}\left[\frac{d^k\phi_X(t)}{dt^k}\right]_{t=0}$	k-th Moment of X

Symbol	Meaning
$\mathbb{1}_A(x)$	Indicator or set membership function that returns 1 if $x \in A$ and 0 otherwise
$\mathbb{R}^d := (-\infty, \infty)^d$	d-dimensional Real Space

Table 2: Symbol Table: Probability and Statistics

Model	PDF or PMF	Mean	Variance
Bernoulli(θ)	$\theta^x(1-\theta)^{1-x}\mathbb{1}_{\{0,1\}}(x)$	θ	$\theta(1-\theta)$
Binomial(n,θ)	$\binom{n}{\theta}\theta^x(1-\theta)^{n-x}\mathbb{1}_{\{0,1,\ldots,n\}}(x)$	$n\theta$	$n\theta(1-\theta)$
Geometric(θ)	$\theta(1-\theta)^x\mathbb{1}_{\mathbb{Z}_+}(x)$	$\frac{1}{\theta}-1$	$\frac{1-\theta}{\theta^2}$
Poisson(λ)	$\frac{\lambda^x e^{-\lambda}}{x!}\mathbb{1}_{\mathbb{Z}_+}(x)$	λ	λ
Uniform(θ_1,θ_2)	$\mathbb{1}_{[\theta_1,\theta_2]}(x)/(\theta_2-\theta_1)$	$\frac{\theta_1+\theta_2}{2}$	$\frac{(\theta_2-\theta_1)^2}{12}$
Exponential(λ)	$\lambda e^{-\lambda x}$	λ^{-1}	λ^{-2}
Normal(μ,σ^2)	$\frac{1}{\sigma\sqrt{2\pi}}e^{(x-\mu)^2/(2\sigma^2)}$	μ	σ^2

Table 3: Random Variables with PDF and PMF (using indicator function), Mean and Variance

Symbol	Meaning
$A=\{\star,\circ,\bullet\}$	A is a set containing the elements $\star$, $\circ$ and $\bullet$
$\circ\in A$	$\circ$ belongs to A or $\circ$ is an element of A
$A\ni\circ$	$\circ$ belongs to A or $\circ$ is an element of A
$\odot\notin A$	$\odot$ does not belong to A
$\#A$	Size of the set A, for e.g. $\#\{\star,\circ,\bullet,\odot\}=4$
$\mathbb{N}$	The set of natural numbers $\{1,2,3,\ldots\}$
$\mathbb{Z}$	The set of integers $\{\ldots,-3,-2,-1,0,1,2,3,\ldots\}$
$\mathbb{Z}_+$	The set of non-negative integers $\{0,1,2,3,\ldots\}$
$\emptyset$	Empty set or the collection of nothing or $\{\}$
$A\subset B$	A is a subset of B or A is contained by B, e.g. $A=\{\circ\},B=\{\bullet\}$
$A\supset B$	A is a superset of B or A contains B e.g. $A=\{\circ,\star,\bullet\},B=\{\circ,\bullet\}$
$A=B$	A equals B, i.e. $A\subset B$ and $B\subset A$
$Q\implies R$	Statement Q implies statement R or If Q then R
$Q\iff R$	$Q\implies R$ and $R\implies Q$
$\{x:x$ satisfies property $R\}$	The set of all x such that x satisfies property R
$A\cup B$	A union B, i.e. $\{x:x\in A$ or $x\in B\}$
$A\cap B$	A intersection B, i.e. $\{x:x\in A$ and $x\in B\}$
$A\setminus B$	A minus B, i.e. $\{x:x\in A$ and $x\notin B\}$
$A:=B$	A is equal to B by definition
$A=:B$	B is equal to A by definition
A^c	A complement, i.e. $\{x:x\in U,$ the universal set, but $x\notin A\}$
$A_1\times A_2\times\cdots\times A_m$	The m-product set $\{(a_1,a_2,\ldots,a_m):a_1\in A_1,a_2\in A_2,\ldots,a_m\in A_m\}$
A^m	The m-product set $\{(a_1,a_2,\ldots,a_m):a_1\in A,a_2\in A,\ldots,a_m\in A\}$
$f:=f(x)=y:\mathbb{X}\to\mathbb{Y}$	A function f from domain $\mathbb{X}$ to range $\mathbb{Y}$
$f^{[-1]}(y)$	Inverse image of y
$f^{[-1]}:=f^{[-1]}(y\in\mathbb{Y})=X\subset\mathbb{X}$	Inverse of f
$a<b$ or $a\leq b$	a is less than b or a is less than or equal to b
$a>b$ or $a\geq b$	a is greater than b or a is greater than or equal to b
$\mathbb{Q}$	Rational numbers
(x,y)	the open interval (x,y), i.e. $\{r:x<r<y\}$
$[x,y]$	the closed interval (x,y), i.e. $\{r:x\leq r\leq y\}$
$(x,y]$	the half-open interval $(x,y]$, i.e. $\{r:x<r\leq y\}$
$[x,y)$	the half-open interval $[x,y)$, i.e. $\{r:x\leq r<y\}$
$\mathbb{R}:=(-\infty,\infty)$	Real numbers, i.e. $\{r:-\infty<r<\infty\}$
$\mathbb{R}_+:=[0,\infty)$	Real numbers, i.e. $\{r:0\leq r<\infty\}$
$\mathbb{R}_{>0}:=(0,\infty)$	Real numbers, i.e. $\{r:0<r<\infty\}$

Table 4: Symbol Table: Sets and Numbers

Symbol	Meaning
$\mathbb{1}_A(x)$	Indicator or set membership function that returns 1 if $x \in A$ and 0 otherwise
$\mathbb{R}^d := (-\infty, \infty)^d$	d-dimensional Real Space
$\mathrm{R}\vec{\mathrm{V}}$	random vector
$F_{X,Y}(x,y)$	Joint distribution function (JDF) of the $\mathrm{R}\vec{\mathrm{V}}$ (X,Y)
$F_{X,Y}(x,y)$	Joint cumulative distribution function (JCDF) of the $\mathrm{R}\vec{\mathrm{V}}$ (X,Y) — same as JDF
$f_{X,Y}(x,y)$	Joint probability mass function (JPMF) of the discrete $\mathrm{R}\vec{\mathrm{V}}$ (X,Y)
$\mathcal{S}_{X,Y}$ $= \{(x_i, y_j) : f_{X,Y}(x_i, x_j) > 0\}$	The support set of the discrete $\mathrm{R}\vec{\mathrm{V}}$ (X,Y)
$f_{X,Y}(x,y)$	Joint probability density function (JPDF) of the continuous $\mathrm{R}\vec{\mathrm{V}}$ (X,Y)
$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$	Marginal probability density/mass function (MPDF/MPMF) of X
$f_Y(y)$ $= \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$	Marginal probability density/mass function (MPDF/MPMF) of Y
$E(g(X,Y))$ $= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f_{X,Y}(x,y)dxdy$	Expectation of a function $g(x,y)$ for continuous $\mathrm{R}\vec{\mathrm{V}}$
$E(g(X,Y))$ $= \sum_{(x,y)\in\mathcal{S}_{X,Y}} g(x,y)f_{X,Y}(x,y)$	Expectation of a function $g(x,y)$ for discrete $\mathrm{R}\vec{\mathrm{V}}$
$E(X^rY^s)$	Joint moment
$\mathbf{Cov}(X,Y) = E(XY) - E(X)E(Y)$	Covariance of X and Y, provided $E(X^2) < \infty$ and $E(Y^2) > \infty$
$F_{X,Y}(x,y) = F_X(x)F_Y(y)$, for every (x,y)	if and only if X and Y are said to be independent
$f_{X,Y}(x,y) = f_X(x)f_Y(y)$, for every (x,y)	if and only if X and Y are said to be independent
$F_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n)$	Joint (cumulative) distribution function (JDF/JCDF) of the discrete or continuous $\mathrm{R}\vec{\mathrm{V}}$ $(X_1,X_2,\ldots,X_n)$
$f_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n)$	Joint probability mass/density function (JPMF/JPDF) of the discrete/continuous $\mathrm{R}\vec{\mathrm{V}}$ $(X_1,X_2,\ldots,X_n)$
$f_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n)$ $= \prod_{i=1}^{n} f_{X_i}(x_i)$, for every $(x_1,x_2,\ldots,x_n)$	if and only if $X_1,X_2,\ldots,X_n$ are (mutually/jointly) independent

Table 5: Symbol Table: Probability and Statistics