# Probability

Lecture Notes

Adolfo J. Rumbos

April 23, 2008

# Contents

# Chapter 1

# Introduction

## 1.1 An example from statistical inference

I had two coins: a trick coin and a fair one. The fair coin has an equal chance of landing heads and tails after being tossed. The trick coin is rigged so that 40% of the time it comes up head. I lost one of the coins, and I don't know whether the coin I am left with is the trick coin, or the fair one. How do I determine whether I have the trick coin or the fair coin?

I believe that I have the trick coin, which has a probability of landing heads 40% of the time, $p = 0.40$. We can run an experiment to determine whether my belief is correct. For instance, we can toss the coin many times and determine the proportion of the tosses that the coin comes up head. If that proportion is very far off from 0.4, we might be led to believe that the coin is perhaps the fair one. On the other hand, even if the coin is fair, the outcome of the experiment might be close to 0.4; so that an outcome close to 0.4 should not be enough to give validity to my belief that I have the trick coin. What we need is a way to evaluate a given outcome of the experiment in the light of the assumption that the coin is fair.

| | Trick | Fair |
|---|---|---|
| Have Trick Coin | (1) | (2) |
| Have Fair Coin | (3) | (4) |

Table 1.1: Which Coin do I have?

Before we run the experiment, we set a decision criterion. For instance, suppose the experiment consists of tossing the coin 100 times and determining the number of heads, $N_H$, in the 100 tosses. If $35 \leqslant N_H \leqslant 45$ then I will conclude that I have the trick coin, otherwise I have the fair coin. There are four scenarios that may happen, and these are illustrated in Table 1.1. The first column shows the two possible decisions we can make: either we have the

fair coin, or we have the trick coin. Depending on the actual state of affairs, illustrated on the first row of the table (we actually have the fair coin or the trick coin), our decision might be in error. For instance, in scenarios (2) or (3), we'd have made an error. What are the chances of that happening? In this course we'll learn how to compute a measure the likelihood of outcomes (1) through (4). This notion of "measure of likelihood" is what is known as a *probability function.* It is a function that assigns a number between 0 and 1 (or 0% to 100%) to sets of outcomes of an experiment.

Once a measure of the likelihood of making an error in a decision is obtained, the next step is to minimize the probability of making the error. For instance, suppose that we actually have the fair coin; based on this assumption, we can compute the probability that the $N_H$ lies between 35 and 45. We will see how to do this later in the course. This would correspond to computing the probability of outcome (2) in Table 1.1. We get

$$\text{Probability of (2)} = \text{Prob}(35 \leqslant N_H \leqslant 45, \text{ given that } p = 0.5) = 18.3\%.$$

Thus, if we have the fair coin, and decide, according to our decision criterion, that we have the trick coin, then there is an 18.3% chance that we make a mistake.

Alternatively, if we have the trick coin, we could make the wrong decision if either $N_H > 45$ or $N_H < 35$. This corresponds to scenario (3) in Table 1.1. In this case we obtain

$$\text{Probability of (3)} = \text{Prob}(N_H < 35 \text{ or } N_H > 45, \text{ given that } p = 0.4) = 26.1\%.$$

Thus, we see that the chances of making the wrong decision are rather high. In order to bring those numbers down, we can modify the experiment in two ways:

- Increase the number of tosses
- Change the decision criterion

**Example 1.1.1** (Increasing the number of tosses)**.** Suppose we toss the coin 500 times. In this case, we will say that we have the trick coin if $175 \leqslant N_H \leqslant 225$. If we have the fair coin, then the probability of making the wrong decision is

$$\text{Probability of (2)} = \text{Prob}(175 \leqslant N_H \leqslant 225, \text{ given that } p = 0.5) = 1.4\%.$$

If we actually have the trick coin, the the probability of making the wrong decision is scenario (3) in Table 1.1. In this case we obtain

$$\text{Probability of (3)} = \text{Prob}(N_H < 175 \text{ or } N_H > 225, \text{ given that } p = 0.4) = 2.0\%.$$

**Example 1.1.2** (Change the decision criterion)**.** Toss the coin 100 times and suppose that we say that we have the trick coin if $38 \leqslant N_H \leqslant 44$. In this case,

$$\text{Probability of (2)} = \text{Prob}(38 \leqslant N_H \leqslant 44, \text{ given that } p = 0.5) = 6.1\%$$

and

$$\text{Probability of (3)} = \text{Prob}(N_H < 38 \text{ or } N_H > 42, \text{ given that } p = 0.4) = 48.0\%.$$

Observe that in case, the probability of making an error if we actually have the fair coin is decreased; however, if we do have the trick coin, then the probability of making an error is increased from that of the original setup.

Our first goal in this course is to define the notion of probability that allowed us to make the calculations presented in this example. Although we will continue to use the coin–tossing experiment as an example to illustrate various concepts and calculations that will be introduced, the notion of probability that we will develop will extends beyond the coin–tossing example presented in this section. In order to define a probability function, we will first need to develop the notion of a *Probability Space.*

# Chapter 2

# Probability Spaces

## 2.1 Sample Spaces and $\sigma$–fields

A *random experiment* is a process or observation, which can be repeated indefinitely under the same conditions, and whose outcomes cannot be predicted with certainty before the experiment is performed. For instance, if a coin is flipped 100 times, the number of heads that come up cannot be determined with certainty. The set of all possible outcomes of a random experiment is called the *sample space* of the experiment. In the case of 100 tosses of a coin, the sample spaces is the set of all possible sequences of Heads (H) and Tails (T) of length 100:

$$\begin{array}{cccccc} H & H & H & H & \dots & H \\ T & H & H & H & \dots & H \\ H & T & H & H & \dots & H \\ \vdots & & & & & \end{array}$$

Subsets of a sample space which satisfy the rules of a $\sigma$–algebra, or $\sigma$–field, are called *events*. These are subsets of the sample space for which we can compute probabilities.

**Definition 2.1.1** ($\sigma$-field)**.** A collection of subsets, $\mathcal{B}$, of a sample space, referred to as events, is called a $\sigma$–field if it satisfies the following properties:

1. $\emptyset \in \mathcal{B}$ ($\emptyset$ denotes the empty set)

2. If $E \in \mathcal{B}$, then its complement, $E^c$, is also an element of $\mathcal{B}$.

3. If $\{E_1, E_2, E_3 \ldots\}$ is a sequence of events, then

$$E_1 \cup E_2 \cup E_3 \cup \ldots = \bigcup_{k=1}^{\infty} E_k \in \mathcal{B}.$$

**Example 2.1.2.** Toss a coin three times in a row. The sample space, $\mathcal{C}$, for this experiment consists of all triples of heads ($H$) and tails ($T$):

$$\left.\begin{array}{l} HHH \\ HHT \\ HTH \\ HTT \\ THH \\ THT \\ TTH \\ TTT \end{array}\right\} \text{Sample Space}$$

A $\sigma$–field for this sample space consists of all possible subsets of the sample space. There are $2^8 = 64$ possible subsets of this sample space; these include the empty set $\emptyset$ and the entire sample space $\mathcal{C}$.

An example of an event,$E$, is the the set of outcomes that yield at least one head:

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

Its complement, $E^c$, is also an event:

$$E^c = \{TTT\}.$$

## 2.2 Some Set Algebra

Sets are collections of objects called elements. If $A$ denotes a set, and $a$ is an element of that set, we write $a \in A$.

**Example 2.2.1.** The sample space, $\mathcal{C}$, of all outcomes of tossing a coin three times in a row is a set. The outcome $HTH$ is an element of $\mathcal{C}$; that is, $HTH \in \mathcal{C}$.

If $A$ and $B$ are sets, and all elements in $A$ are also elements of $B$, we say that $A$ is a subset of $B$ and we write $A \subseteq B$. In symbols,

$$A \subseteq B \text{ if and only if } x \in A \Rightarrow x \in B.$$

**Example 2.2.2.** Let $\mathcal{C}$ denote the set of all possible outcomes of three consecutive tosses of a coin. Let $E$ denote the the event that exactly one of the tosses yields a head; that is,

$$E = \{HTT,\ THT,\ TTH\};$$

then, $E \subseteq \mathcal{C}$

Two sets $A$ and $B$ are said to be equal if and only if all elements in $A$ are also elements of $B$, and vice versa; i.e., $A \subseteq B$ and $B \subseteq A$. In symbols,

$$A = B \text{ if and only if } A \subseteq B \text{ and } B \subseteq A.$$

Let $E$ be a subset of a sample space $\mathcal{C}$. The complement of $E$, denoted $E^c$, is the set of elements of $\mathcal{C}$ which are not elements of $E$. We write,

$$E^c = \{x \in \mathcal{C} \mid x \notin E\}.$$

**Example 2.2.3.** If $E$ is the set of sequences of three tosses of a coin that yield exactly one head, then

$$E^c = \{HHH,\ HHT,\ HTH,\ THH,\ TTT\};$$

that is, $E^c$ is the event of seeing two or more heads, or no heads in three consecutive tosses of a coin.

If $A$ and $B$ are sets, then the set which contains all elements that are contained in either $A$ or in $B$ is called the union of $A$ and $B$. This union is denoted by $A \cup B$. In symbols,

$$A \cup B = \{x \mid x \in A \text{ or } x \in B\}.$$

**Example 2.2.4.** Let $A$ denote the event of seeing exactly one head in three consecutive tosses of a coin, and let $B$ be the event of seeing exactly one tail in three consecutive tosses. Then,

$$A = \{HTT,\ THT,\ TTH\},$$

$$B = \{THH,\ HTH,\ HHT\},$$

and

$$A \cup B = \{HTT,\ THT,\ TTH,\ THH,\ HTH,\ HHT\}.$$

Notice that $(A \cup B)^c = \{HHH,\ TTT\}$, i.e., $(A \cup B)^c$ is the set of sequences of three tosses that yield either heads or tails three times in a row.

If $A$ and $B$ are sets then the intersection of $A$ and $B$, denoted $A \cap B$, is the collection of elements that belong to both $A$ and $B$. We write,

$$A \cap B = \{x \mid x \in A \ \&\ x \in B\}.$$

Alternatively,

$$A \cap B = \{x \in A \mid x \in B\}$$

and

$$A \cap B = \{x \in B \mid x \in A\}.$$

We then see that

$$A \cap B \subseteq A \text{ and } A \cap B \subseteq B.$$

**Example 2.2.5.** Let $A$ and $B$ be as in the previous example (see Example 2.2.4). Then, $A \cap B = \emptyset$, the empty set, i.e., $A$ and $B$ have no elements in common.

**Definition 2.2.6.** *If $A$ and $B$ are sets, and $A \cap B = \emptyset$, we can say that $A$ and $B$ are disjoint.*

**Proposition 2.2.7** (De Morgan's Laws)**.** *Let $A$ and $B$ be sets.*

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

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

*Proof of (i).* Let $x \in (A \cap B)^c$. Then $x \notin A \cap B$. Thus, either $x \notin A$ or $x \notin B$; that is, $x \in A^c$ or $x \in B^c$. It then follows that $x \in A^c \cup B^c$. Consequently,

$$(A \cap B)^c \subseteq A^c \cup B^c. \tag{2.1}$$

Conversely, if $x \in A^c \cup B^c$, then $x \in A^c$ or $x \in B^c$. Thus, either $x \notin A$ or $x \notin B$; which shows that $x \notin A \cap B$; that is, $x \in (A \cap B)^c$. Hence,

$$A^c \cup B^c \subseteq (A \cap B)^c. \tag{2.2}$$

It therefore follows from (2.1) and (2.2) that

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

□

**Example 2.2.8.** Let $A$ and $B$ be as in Example 2.2.4. Then $(A \cap B)^C = \emptyset^C = \mathcal{C}$. On the other hand,

$$A^c = \{HHH,\ HHT,\ HTH,\ THH,\ TTT\},$$

and

$$B^c = \{HHH,\ HTT,\ THT,\ TTH,\ TTT\}.$$

Thus $A^c \cup B^c = C$. Observe that

$$A^c \cap B^c = \{HHH,\ TTT\}.$$

We can define unions and intersections of many (even infinitely many) sets. For example, if $E_1, E_2, E_3, \ldots$ is a sequence of sets, then

$$\bigcup_{k=1}^{\infty} E_k \quad = \quad \{x \mid x \text{ is in at least one of the sets in the sequence}\}$$

and

$$\bigcap_{k=1}^{\infty} E_k \quad = \quad \{x \mid x \text{ is in all of the sets in the sequence}\}.$$

**Example 2.2.9.** Let $E_k = \left\{ x \in \mathbb{R} \mid 0 \leqslant x < \frac{1}{k} \right\}$ for $k = 1, 2, 3, \ldots$; then,

$$\bigcup_{k=1}^{\infty} E_k = [0, 1) \quad \text{and} \quad \bigcap_{k=1}^{\infty} E^k = \{0\}.$$

Finally, if $A$ and $B$ are sets, then $A \backslash B$ denotes the set of elements in $A$ which are not in $B$; we write

$$A \backslash B = \{x \in A \mid x \notin B\}$$

**Example 2.2.10.** *Let $E$ be an event in a sample space $(C)$. Then, $\mathcal{C} \backslash E = E^c$.*

**Example 2.2.11.** Let $A$ and $B$ be sets. Then,

$$\begin{array}{rcl} x \in A \backslash B & \iff & x \in A \text{ and } x \notin B \\ & \iff & x \in A \text{ and } x \in B^c \\ & \iff & x \in A \cap B^c \end{array}$$

Thus $A \backslash B = A \cap B^c$.

## 2.3 More on $\sigma$–fields

**Proposition 2.3.1.** *Let $\mathcal{C}$ be a sample space, and $\mathcal{S}$ be a non-empty collection of subsets of $\mathcal{C}$. Then the intersection of all $\sigma$-fields which contain $\mathcal{S}$ is a $\sigma$-field. We denote it by $\mathcal{B}(\mathcal{S})$.*

*Proof.* Observe that every $\sigma$–field which contains $\mathcal{S}$ contains the empty set, $\emptyset$, by property (1) in Definition 2.1.1. Thus, $\emptyset$ is in every $\sigma$–field which contains $\mathcal{S}$. It then follows that $\emptyset \in \mathcal{B}(\mathcal{S})$.

Next, suppose $E \in \mathcal{B}(\mathcal{S})$, then $E$ is contained in every $\sigma$–field which contains $\mathcal{S}$. Thus, by (2) in Definition 2.1.1, $E^c$ is in every $\sigma$–field which contains $\mathcal{S}$. It then follows that $E^c \in \mathcal{B}(\mathcal{S})$.

Finally, let $\{E_1, E_2, E_3, \ldots\}$ be a sequence in $\mathcal{B}(\mathcal{S})$. Then, $\{E_1, E_2, E_3, \ldots\}$ is in every $\sigma$–field which contains $\mathcal{S}$. Thus, by (3) in Definition 2.1.1,

$$\bigcup_{k=1}^{\infty} E_k$$

is in every $\sigma$–field which contains $\mathcal{S}$. Consequently,

$$\bigcup_{k=1}^{\infty} E_k \in \mathcal{B}(\mathcal{S})$$

□

**Remark 2.3.2.** $\mathcal{B}(\mathcal{S})$ is the "smallest" $\sigma$–field which contains $\mathcal{S}$. In fact,

$$\mathcal{S} \subseteq \mathcal{B}(\mathcal{S}),$$

since $\mathcal{B}(\mathcal{S})$ is the intersection of all $\sigma$–fields which contain $\mathcal{S}$. By the same reason, if $\mathcal{E}$ is any $\sigma$–field which contains $\mathcal{S}$, then $\mathcal{B}(\mathcal{S}) \subseteq \mathcal{E}$.

**Definition 2.3.3.** $\mathcal{B}(\mathcal{S})$ called the $\sigma$-field generated by $\mathcal{S}$

**Example 2.3.4.** Let $\mathcal{C}$ denote the set of real numbers $\mathbb{R}$. Consider the collection, $\mathcal{S}$, of semi–infinite intervals of the form $(-\infty, b]$, where $b \in \mathbb{R}$; that is,

$$\mathcal{S} = \{(-\infty, b] \mid b \in \mathbb{R}\}.$$

Denote by $\mathcal{B}_o$ the $\sigma$–field generated by $\mathcal{S}$. This $\sigma$–field is called the **Borel** $\sigma$–field of the real line $\mathbb{R}$. In this example, we explore the different kinds of events in $\mathcal{B}_o$.

First, observe that since $\mathcal{B}_o$ is closed under the operation of complements, intervals of the form

$$(-\infty, b]^c = (b, +\infty), \qquad \text{for } b \in \mathbb{R},$$

are also in $\mathcal{B}_o$. It then follows that semi–infinite intervals of the form

$$(a, +\infty), \qquad \text{for } a \in \mathbb{R},$$

are also in the Borel $\sigma$–field $\mathcal{B}_o$.

Suppose that $a$ and $b$ are real numbers with $a < b$. Then, since

$$(a, b] = (-\infty, b] \cap (a, +\infty),$$

the half–open, half–closed, bounded intervals, $(a, b]$ for $a < b$, are also elements in $\mathcal{B}_o$.

Next, we show that open intervals $(a, b)$, for $a < b$, are also events in $\mathcal{B}_o$. To see why this is so, observe that

$$(a, b) = \bigcup_{k=1}^{\infty} \left(a, b - \frac{1}{k}\right]. \tag{2.3}$$

To see why this is so, observe that if

$$a < b - \frac{1}{k},$$

then

$$\left(a, b - \frac{1}{k}\right] \subseteq (a, b),$$

since $b - \dfrac{1}{k} < b$. On the other hand, if

$$a \geqslant b - \frac{1}{k},$$

then

$$\left(a, b - \frac{1}{k}\right] = \emptyset.$$

It then follows that

$$\bigcup_{k=1}^{\infty} \left(a, b - \frac{1}{k}\right] \subseteq (a, b). \tag{2.4}$$

Now, for any $x \in (a, b)$, we can find a $k \geqslant 1$ such that

$$\frac{1}{k} < b - x.$$

It then follows that

$$x < b - \frac{1}{k}$$

and therefore

$$x \in \left(a, b - \frac{1}{k}\right].$$

Thus,

$$x \in \bigcup_{k=1}^{\infty} \left(a, b - \frac{1}{k}\right].$$

Consequently,

$$(a, b) \subseteq \bigcup_{k=1}^{\infty} \left(a, b - \frac{1}{k}\right]. \tag{2.5}$$

Combining (2.4) and (2.5) yields (2.3).

## 2.4 Defining a Probability Function

Given a sample space, $\mathcal{C}$, and a $\sigma$–field, $\mathcal{B}$, defined in $\mathcal{C}$, we can now probability function on $\mathcal{B}$.

**Definition 2.4.1.** Let $\mathcal{C}$ be a sample space and $\mathcal{B}$ be a $\sigma$–field of subsets of $\mathcal{C}$. A probability function, Pr, defined on $\mathcal{B}$ is a real valued function

$$\Pr\colon \mathcal{B} \to [0, 1];$$

that is, Pr takes on values from 0 to 1, which satisfies:

(1) $\Pr(\mathcal{C}) = 1$

(2) If $\{E_1, E_2, E_3 \ldots\} \subseteq \mathcal{B}$ is a sequence of mutually disjoint subsets of $\mathcal{C}$ in $\mathcal{B}$, i.e., $E_i \cap E_j = \emptyset$ for $i \neq j$; then,

$$\Pr\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{k=1}^{\infty} \Pr(E_k) \tag{2.6}$$

**Remark 2.4.2.** The infinite sum on the right hand side of (2.6) is to be understood as

$$\sum_{k=1}^{\infty} \Pr(E_k) = \lim_{n\to\infty} \sum_{k=1}^{n} \Pr(E_k).$$

**Example 2.4.3.** Let $\mathcal{C} = R$ and $\mathcal{B}$ be the Borel $\sigma$–field, $\mathcal{B}_o$, in the real line. Given a nonnegative, integrable function, $f : \mathbb{R} \to \mathbb{R}$, satisfying

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

we define a probability function, Pr, on $\mathcal{B}_o$ as follows

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

for any bounded, open interval, $(a, b)$, of real numbers.

Since $\mathcal{B}_o$ is generated by all bounded open intervals, this definition allows us to define Pr on all Borel sets of the real line; in fact, we get

$$\Pr(E) = \int_E f(x)\ \mathrm{d}x$$

for all $E \in \mathcal{B}_o$.

We will see why this is a probability function in another example in this notes.

**Notation.** The triple $(\mathcal{C}, \mathcal{B}, \Pr)$ is known as a **Probability Space**.

### 2.4.1 Properties of Probability Spaces

Let $(\mathcal{C}, \mathcal{B}, \Pr)$ denote a probability space and $A$, $B$ and $E$ denote events in $\mathcal{B}$.

1. Since $E$ and $E^c$ are disjoint, by (2) in Definition 2.4.1, we get that

   $$\Pr(E \cup E^c) = \Pr(E) + \Pr(E^c)$$

   But $E \cup E^c = \mathcal{C}$ and $\Pr(\mathcal{C}) = 1$ by (1) in Definition 2.4.1. We therefore get that

   $$\Pr(E^c) = 1 - \Pr(E).$$

2. Since $\emptyset = \mathcal{C}^c$, $\Pr(\emptyset) = 1 - P(\mathcal{C}) = 1 - 1 = 0$, by (1) in Definition 2.4.1.

3. Suppose that $A \subset B$. Then,

   $$B = A \cup (B \backslash A)$$

   [Recall that $B \backslash A = B \cap A^C$, thus $B \backslash A \in \mathcal{B}$.]

Since $A \cap (B \backslash A) = \emptyset$,

$$\Pr(B) = \Pr(A) + \Pr(B \backslash A),$$

by (2) in in Definition 2.4.1.

Next, since $\Pr(B \backslash A) \geqslant 0$, by Definition 2.4.1,

$$\Pr(B) \geqslant \Pr(A).$$

We have therefore proved that

$$A \subseteq B \Rightarrow \Pr(A) \leqslant \Pr(B).$$

4. From (2) in Definition 2.4.1 we get that if $A$ and $B$ are disjoint, then $\Pr(A \cup B) = \Pr(A) + \Pr(B)$.

   On the other hand, if A & B are not disjoint, observe first that $A \subseteq A \cup B$ and so we can write,

   $$A \cup B = A \cup ((A \cup B) \backslash A)$$

   i.e., $A \cup B$ can be written as a disjoint union of A and $(A \cup B) \backslash A$, where

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

   Thus, by (2) in Definition 2.4.1,

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

   On the other hand, $A \cap B \subseteq B$ and so

   $$B = (A \cap B) \cup (B \backslash (A \cap B))$$

   where,

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

   Thus, $B$ is the disjoint union of $A \cap B$ and $B \cap A^c$. Thus,

   $$\Pr(B) = \Pr(A \cap B) + \Pr(B \cap A^c)$$

   by (2) in Definition 2.4.1. It then follows that

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

   Consequently,

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

**Proposition 2.4.4.** *Let $(\mathcal{C}, \mathcal{B}, \Pr)$ be a sample space. Suppose that $E_1, E_2, E_3, \ldots$ is a sequence of events in $\mathcal{B}$ satisfying*

$$E_1 \subseteq E_2 \subseteq E_3 \subseteq \cdots .$$

*Then,*

$$\lim_{n\to\infty} Pr\,(E_n) = Pr\left(\bigcup_{k=1}^{\infty} E_k\right).$$

*Proof:* Define the sequence of events $B_1, B_2, B_3, \ldots$ by

$$\begin{array}{rcl} B_1 & = & E_1 \\ B_2 & = & E_2 \setminus E_1 \\ B_3 & = & E_3 \setminus E_2 \\ & \vdots & \\ B_k & = & E_k \setminus E_{k-1} \\ & \vdots & \end{array}$$

The the events $B_1, B_2, B_3, \ldots$ are mutually disjoint and, therefore, by (2) in Definition 2.4.1,

$$\Pr\left(\bigcup_{k=1}^{\infty} B_k\right) = \sum_{k=1}^{\infty} \Pr(B_k),$$

where

$$\sum_{k=1}^{\infty} \Pr(B_k) = \lim_{n\to\infty} \sum_{k=1}^{n} \Pr(B_k).$$

Observe that

$$\bigcup_{k=1}^{\infty} B_k = \bigcup_{k=1}^{\infty} E_k. \tag{2.7}$$

(Why?)

Observe also that

$$\bigcup_{k=1}^{n} B_k = E_n,$$

ant therefore

$$\Pr(E_n) = \sum_{k=1}^{n} \Pr(B_k);$$

so that

$$\begin{aligned}\lim_{n\to\infty}\Pr(E_n) &= \lim_{n\to\infty}\sum_{k=1}^{n}\Pr(B_k)\\ &= \sum_{k=1}^{\infty}\Pr(B_k)\\ &= \Pr\left(\bigcup_{k=1}^{\infty}B_k\right)\\ &= \Pr\left(\bigcup_{k=1}^{\infty}E_k\right), \quad \text{by (2.7)},\end{aligned}$$

which we wanted to show. □

**Example 2.4.5.** As an example of an application of this result, consider the situation presented in Example 2.4.3. Given an integrable, non–negative function $f\colon \mathbb{R}\to\mathbb{R}$ satisfying

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

we define

$$\Pr\colon \mathcal{B}_o \to \mathbb{R}$$

by specifying what it does to generators of $\mathcal{B}_o$; for example, open, bounded intervals:

$$\Pr((a,b)) = \int_a^b f(x)\ \mathrm{d}x.$$

Then, since $\mathbb{R}$ is the union of the nested intervals $(-k,k)$, for $k = 1, 2, 3, \ldots$,

$$\Pr(\mathbb{R}) = \lim_{n\to\infty}\Pr((-n,n)) = \lim_{n\to\infty}\int_{-n}^{n} f(x)\ \mathrm{d}x = \int_{-\infty}^{\infty} f(x)\ \mathrm{d}x = 1.$$

It can also be shown (this is an exercise) that

**Proposition 2.4.6.** *Let $(\mathcal{C},\mathcal{B},\Pr)$ be a sample space. Suppose that $E_1, E_2, E_3, \ldots$ is a sequence of events in $\mathcal{B}$ satisfying*

$$E_1 \supseteq E_2 \supseteq E_3 \supseteq \cdots.$$

*Then,*

$$\lim_{n\to\infty} Pr\,(E_n) = Pr\,\left(\bigcap_{k=1}^{\infty} E_k\right).$$

**Example 2.4.7.** *[Continuation of Example 2.4.5]* Given $a \in \mathbb{R}$, observe that $\{a\}$ is the intersection of the nested intervals $\left(a - \frac{1}{k}, a + \frac{1}{k}\right)$, for $k = 1, 2, 3, \ldots$ Then,

$$\Pr(\{a\}) = \lim_{n\to\infty} \Pr\left(a - \frac{1}{n}, a + \frac{1}{n}\right) = \lim_{n\to\infty} \int_{a-1/n}^{a+1/n} f(x)\, \mathrm{d}x = \int_a^a f(x)\, \mathrm{d}x = 0.$$

### 2.4.2 Constructing Probability Functions

In Examples 2.4.3–2.4.7 we illustrated how to construct a probability function of the Borel $\sigma$–field of the real line. Essentially, we prescribed what the function does to the generators. When the sample space is finite, the construction of a probability function is more straight forward

**Example 2.4.8.** Three consecutive tosses of a fair coin yields the sample space

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

We take as our $\sigma$–field, $\mathcal{B}$, the collection of all possible subsets of $\mathcal{C}$.

We define a probability function, Pr, on $\mathcal{B}$ as follows. Assuming we have a fair coin, all the sequences making up $\mathcal{C}$ are equally likely. Hence, each element of $\mathcal{C}$ must have the same probability value, $p$. Thus,

$$\Pr(\{HHH\}) = \Pr(\{HHT\}) = \cdots = \Pr(\{TTT\}) = p.$$

Thus, by property (2) in Definition 2.4.1,

$$\Pr(\mathcal{C}) = 8p.$$

On the other hand, by the probability property (1) in Definition 2.4.1, $\Pr(\mathcal{C}) = 1$, so that

$$8p = 1 \Rightarrow p = \frac{1}{8}$$

**Example 2.4.9.** Let $E$ denote the event that three consecutive tosses of a fair coin yields exactly one head. Then,

$$\Pr(E) = \Pr(\{HTT, THT, TTH\}) = \frac{3}{8}.$$

**Example 2.4.10.** Let $A$ denote the event a head comes up in the first toss and $B$ denotes the event a head comes up on the second toss. Then,

$$A = \{HHH, HHT, HTH, HTT\}$$

and

$$B = \{HHH, HHT, THH, THT)\}.$$

Thus, $\Pr(A) = 1/2$ and $\Pr(B) = 1/2$. On the other hand,

$$A \cap B = \{HHH, HHT\}$$

and therefore

$$\Pr(A \cap B) = \frac{1}{4}.$$

Observe that $\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$. When this happens, we say that events $A$ and $B$ are independent, i.e., the outcome of the first toss does not influence the outcome of the second toss.

## 2.5 Independent Events

**Definition 2.5.1.** Let $(\mathcal{C}, \mathcal{B}, \Pr)$ be a probability space. Two events $A$ and $B$ in $\mathcal{B}$ are said to be independent if

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$$

**Example 2.5.2** (Two events that are not independent)**.** There are three chips in a bowl: one is red, the other two are blue. Suppose we draw two chips successively at random and without replacement. Let $E_1$ denote the event that the first draw is red, and $E_2$ denote the event that the second draw is blue. Then, the outcome of $E_1$ will influence the probability of $E_2$. For example, if the first draw is red, then $\mathrm{P}(E_2) = 1$; but, if the first draw is blue then $\mathrm{P}(E_2) = \frac{1}{2}$. Thus, $E_1$ & $E_2$ should not be independent. In fact, in this case we get $P(E_1) = \frac{1}{3}, P(E_2) = \frac{2}{3}$ and $P(E_1 \cap E_2) = \frac{1}{3} \neq \frac{1}{3} \cdot \frac{2}{3}$. To see this, observe that the outcomes of the experiment yield the sample space

$$\mathcal{C} = \{RB_1, RB_2, B_1R, B_1B_2, B_2R, B_2B_1\},$$

where $R$ denotes the red chip and $B_1$ and $B_2$ denote the two blue chips. Observe that by the nature of the random drawing, all of the outcomes in $\mathcal{C}$ are equally likely. Note that

$$\begin{aligned} E_1 &= \{RB_1, RB_2\}, \\ E_2 &= \{RB_1, RB_2, B_1B_2, B_2B_1\}; \end{aligned}$$

so that

$$\begin{aligned} \Pr(E_1) &= \tfrac{1}{6} + \tfrac{1}{6} = \tfrac{1}{3}, \\ \Pr(E_2) &= \tfrac{4}{6} = \tfrac{2}{3}. \end{aligned}$$

On the other hand,

$$E_1 \cap E_2 = \{RB_1, RB_2\}$$

so that,

$$P(E_1 \cap E_2) = \frac{2}{6} = \frac{1}{3}$$

**Proposition 2.5.3.** *Let $(\mathcal{C}, \mathcal{B}, Pr)$ be a probability space. If $E_1$ and $E_2$ are independent events in $\mathcal{B}$, then so are*

*(a)* $E_1^c$ *and* $E_2^c$

*(b)* $E_1^c$ *and* $E_2$

*(c)* $E_1$ *and* $E_2^c$

*Proof of (a):* Suppose $E_1$ and $E_2$ are independent. By De Morgan's Law

$$\Pr(E_1^c \cap E_2^c) = \Pr((E_1 \cap E_2)^c) = 1 - \Pr(E_1 \cup E_2)$$

Thus,

$$\Pr(E_1^c \cap E_2^c) = 1 - (\Pr(E_1) + \Pr(E_2) - \Pr(E_1 \cap E_2))$$

Hence, since $E_1$ and $E_2$ are independent,

$$\begin{aligned} \Pr(E_1^c \cap E_2^c) &= 1 - \Pr(E_1) - \Pr(E_2) + \Pr(E_1) \cdot \Pr(E_2) \\ &= (1 - \Pr(E_1)) \cdot (1 - \Pr(E_2) \\ &= P(E_1^c) \cdot P(E_2^c) \end{aligned}$$

Thus, $E_1^c$ and $E_2^c$ are independent. □

## 2.6 Conditional Probability

In Example 2.5.2 we saw how the occurrence on an event can have an effect on the probability of another event. In that example, the experiment consisting of drawing chips from a bowl without replacement. Of the three chips in a bowl, one was red, the others were blue. Two chips, one after the other, were drawn at random and without replacement from the bowl. We had two events:

$$\begin{aligned} E_1 &: \text{The first chip drawn is red,} \\ E_2 &: \text{The second chip drawn is blue.} \end{aligned}$$

We pointed out the fact that, since the sampling is drawn without replacement, the probability of $E_2$ is influenced by the outcome of the first draw. If a red chip is drawn in the first draw, then the probability of $E_2$ is 1. But, if a blue chip is drawn in the first draw, then the probability of $E_2$ is $\frac{1}{2}$. We would like to model the situation in which the outcome of the first draw is known.

Suppose we are given a probability space, $(\mathcal{C}, \mathcal{B}, \Pr)$, and we are told that an event, $B$, has occurred. Knowing that $B$ is true, changes the situation. This can be modeled by introducing a new probability space which we denote by $(B, \mathcal{B}_B, P_B)$. Thus, since we know $B$ has taken place, we take it to be our new sample space. We define new $\sigma$-fields and probabilities as follows.

$$\mathcal{B}_B = \{E \cap B | E \in \mathcal{B}\}$$

This is a $\sigma$-field. Why?

(i) Observe that $\emptyset = B^c \cap B \in \mathcal{B}_B$

(ii) If $E \cap B \in \mathcal{B}_B$, then its complement in $B$ is

$$\begin{aligned} B \backslash (E \cap B) &= B \cap (E \cap B)^c \\ &= B \cap (E^c \cup B^c) \\ &= (B \cap E^c) \cup (B \cap B^c) \\ &= (B \cap E^c) \cup \emptyset \\ &= E^c \cap B. \end{aligned}$$

Thus, the complement of $E \cap B$ in $B$ is in $\mathcal{B}_B$.

(iii) Let $\{E_1 \cap B, E_2 \cap B, E_3 \cap B, \ldots\}$ be a sequence of events in $\mathcal{B}_B$; then, by the distributive property,

$$\bigcup_{k=1}^{\infty} E_k \cap B = \left( \bigcup_{k=1}^{\infty} E_k \right) \cap B \in \mathcal{B}_B.$$

Next, we define a probability function on $\mathcal{B}_B$ as follows. Assume $P(B) > 0$ and define:

$$P_B(E \cap B) = \frac{\Pr(E \cap B)}{\Pr(B)}$$

for all $E \in \mathcal{B}$.

Observe that since

$$\emptyset \subseteq E \cap B \subseteq B \text{ for all } B \in \mathcal{B},$$

$$0 \leqslant \Pr(E \cap B) \leq \Pr(B) \quad \text{for all } E \in \mathcal{B}.$$

Thus, dividing by $\Pr(B)$ yields that

$$0 \leqslant P_B(E \cap B) \leqslant 1 \quad \text{for all } E \in \mathcal{B}.$$

Observe also that

$$P_B(B) = 1.$$

Finally, If $E_1, E_2, E_3, \ldots$ are mutually disjoint events, then so are $E_1 \cap B, E_2 \cap B, E_3 \cap B, \ldots$ It then follows that

$$\Pr\left( \bigcup_{k=1}^{\infty} (E_k \cap B) \right) = \sum_{k=1}^{\infty} \Pr(E_k \cap B).$$

Thus, dividing by $\Pr(B)$ yields that

$$P_B\left( \bigcup_{k=1}^{\infty} (E_k \cap B) \right) = \sum_{k=1}^{\infty} P_B(E_k \cap B).$$

Hence, $P_B \colon \mathcal{B}_B \to [0, 1]$ is indeed a probability function.

Notation: we write $P_B(E \cap B)$ as $P(E \mid B)$, which is read "probability of E given $B$" and we call this the conditional probability of E given $B$.

**Definition 2.6.1** (Conditional Probability)**.** For an event $B$ with $\Pr(B) > 0$, we define the conditional probability of any event $E$ given $B$ to be

$$\Pr(E \mid B) = \frac{\Pr(E \cap B)}{\Pr(B)}.$$

**Example 2.6.2** (Example 2.5.2 Revisited)**.** In the example of the three chips (one red, two blue) in a bowl, we had

$$\begin{array}{ll} E_1 & = \{RB_1, RB_2\} \\ E_2 & = \{RB_1, RB_2, B_1B_2, B_2B_1\} \\ E_1^c & = \{B_1R, B_1B_2, B_2R, B_2B_1\} \end{array}$$

Then,

$$\begin{array}{c} E_1 \cap E_2 = \{RB_1, RB_2\} \\ E_2 \cap E_1^c = \{B_1B_2, B_2B_1\} \end{array}$$

Then $\Pr(E_1) = 1/3$ and $\Pr(E_1^c) = 2/3$ and

$$\begin{array}{c} \Pr(E_1 \cap E_2) = 1/3 \\ \Pr(E_2 \cap E_1^c) = 1/3 \end{array}$$

Thus

$$\Pr(E_2 \mid E_1) = \frac{\Pr(E_2 \cap E_1)}{\Pr(E_1)} = \frac{1/3}{1/3} = 1$$

and

$$\Pr(E_2 \mid E_1^c) = \frac{\Pr(E_2 \cap E_1^C)}{\Pr(E_1^c)} = \frac{1/3}{2/3} = 1/2.$$

Some Properties of Conditional Probabilities

(i) For any events $E_1$ and $E_2$, $\Pr(E_1 \cap E_2) = \Pr(E_1) \cdot \Pr(E_2 \mid E_1)$.

*Proof:* If $\Pr(E_1) = 0$, then from $\emptyset \subseteq E_1 \cap E_2 \subseteq E_1$ we get that

$$0 \leqslant \Pr(E_1 \cap E_2) \leqslant \Pr(E_1) = 0.$$

Thus, $\Pr(E_1 \cap E_2) = 0$ and the result is true.

Next, if $\Pr(E_1) > 0$, from the definition of conditional probability we get that

$$\Pr(E_2 \mid E_1) = \frac{\Pr(E_2 \cap E_1)}{\Pr(E_1)},$$

and we therefore get that $\Pr(E_1 \cap E_2) = \Pr(E_1) \cdot \Pr(E_2 \mid E_1)$. □

(ii) Assume $\Pr(E_2) > 0$. Events $E_1$ and $E_2$ are independent iff

$$\Pr(E_1 \mid E_2) = \Pr(E_1).$$

*Proof.* Suppose first that $E_1$ and $E_2$ are independent. Then, $\Pr(E_1 \cap E_2) = \Pr(E_1) \cdot \Pr(E_2)$, and therefore

$$\Pr(E_1 \mid E_2) = \frac{\Pr(E_1 \cap E_2)}{\Pr(E_2)} = \frac{\Pr(E_1) \cdot \Pr(E_2)}{\Pr(E_2)} = \Pr(E_1).$$

Conversely, suppose that $\Pr(E_1 \mid E_2) = \Pr(E_1)$. Then, by Property 1.,

$$\begin{aligned} \Pr(E_1 \cap E_2) &= \Pr(E_2 \cap E_1) \\ &= \Pr(E_2) \cdot \Pr(E_1 \mid E_2) \\ &= \Pr(E_2) \cdot \Pr(E_1) \\ &= \Pr(E_1) \cdot \Pr(E_2), \end{aligned}$$

which shows that $E_1$ and $E_2$ are independent. □

(iii) Assume $\Pr(B) > 0$. Then, for any event $E$,

$$\Pr(E^c \mid B) = 1 - \Pr(E \mid B).$$

*Proof:* Recall that $P(E^c|B) = P_B(E^c \cap B)$, where $P_B$ defines a probability function of $\mathcal{B}_B$. Thus, since $E^c \cap B$ is the complement of $E \cap B$ in $B$, we obtain

$$P_B(E^c \cap B) = 1 - P_B(E \cap B).$$

Consequently, $\Pr(E^c \mid B) = 1 - \Pr(E \mid B)$. □

(iv) Suppose $E_1, E_2, E_3, \ldots, E_n$ are mutually exclusive events (i.e., $E_i \cap E_j = \emptyset$ for $i \neq j$) such that

$$\mathcal{C} = \bigcup_{k=1}^{n} E_k \text{ and } P(E_i) > 0 \text{ for i = 1,2,3....,n.}$$

Let B be another event in B. Then

$$B = B \cap \mathcal{C} = B \cap \bigcup_{k=1}^{n} E_k = \bigcup_{k=1}^{n} B \cap E_k$$

Since $\{B \cap E_1, B \cap E_2, \ldots, B \cap E_n\}$ are mutually exclusive (disjoint)

$$P(B) = \sum_{k=1}^{n} P(B \cap E_k)$$

or

$$P(B) = \sum_{k=1}^{n} P(E_k) \cdot P(B \mid E_k)$$

This is called the **Law of Total Probability**.

**Example 2.6.3** (An Application of the Law of Total Probability)**.** Medical tests can have false–positive results and false–negative results. That is, a person who does not have the disease might test positive, or a sick person might test negative, respectively. The proportion of false–positive results is called the **false–positive rate**, while that of false–negative results is called the **false–negative rate**. These rates can be expressed as conditional probabilities. For example, suppose the test is to determine if a person is HIV positive. Let $T_P$ denote the event that, in a given population, the person tests positive for HIV, and $T_N$ denote the event that the person tests negative. Let $H_P$ denote the event that the person tested is HIV positive and let $H_N$ denote the event that the person is HIV negative. The false positive rate is $P(T_P \mid H_N)$ and the false negative rate is $P(T_N \mid H_P)$. These rates are assumed to be known and ideally and very small. For instance, in the United States, according to a 2005 study presented in the *Annals of Internal Medicine*[1] suppose $P(T_P \mid H_N) = 1.5\%$ and $P(T_N \mid H_P) = 0.3\%$. That same year, the prevalence of HIV infections was estimated by the CDC to be about 0.6%, that is $P(H_P) = 0.006$

Suppose a person in this population is tested for HIV and that the test comes back positive, what is the probability that the person really is HIV positive? That is, what is $P(H_P \mid T_P)$?

***Solution**:*

$$\Pr(H_P|T_P) = \frac{\Pr(H_P \cap T_P)}{\Pr(T_P)} = \frac{\Pr(T_P \mid H_P)\Pr(H_P)}{\Pr(T_P)}$$

But, by the law of total probability, since $H_P \cup H_N = \mathcal{C}$,

$$\begin{aligned} \Pr(T_P) &= \Pr(H_P)\Pr(T_P|H_P) + \Pr(H_N)\Pr(T_P \mid H_N) \\ &= \Pr(H_P)[1 - \Pr(T_N \mid H_P)] + \Pr(H_N)\Pr(T_P \mid H_N) \end{aligned}$$

Hence,

$$\begin{aligned} \Pr(H_P \mid T_P) &= \frac{[1 - \Pr(T_N \mid H_P)]\Pr(H_P)}{\Pr(H_P)[1 - \Pr(T_N)]\Pr(H_P) + \Pr(H_N)\Pr(T_P \mid H_N)} \\ &= \frac{(1 - 0.003)(0.006)}{(0.006)(1 - 0.003) + (0.994)(0.015)}, \end{aligned}$$

which is about 0.286 or 28.6%. □

In general, if $\Pr(B) > 0$ and $\mathcal{C} = E_1 \cup E_2 \cup \cdots \cup E_n$ where $\{E_1, E_2, E_3, \ldots, E_n\}$ are mutually exclusive, then

$$\Pr(E_j \mid B) = \frac{\Pr(B \cap E_j)}{\Pr(B)} = \frac{\Pr(B \cap E_j)}{\sum_{k=1}^{k} \Pr(E_i)\Pr(B \mid E_k)}$$

[1] "Screening for HIV: A Review of the Evidence for the U.S. Preventive Services Task Force", Annals of Internal Medicine, Chou et. al, Volume 143 Issue 1, pp. 55-73

This result is known as **Baye's Theorem**, and is also be written as

$$\Pr(E_j \mid B) = \frac{\Pr(E_j)\Pr(B \mid E_j)}{\sum_{k=1}^{n} \Pr(E_k)\Pr(B \mid E_k)}$$

**Remark 2.6.4.** The specificity of a test is the proportion of healthy individuals that will test negative; this is the same as

$$\begin{aligned} 1 - \text{false positive rate} &= 1 - P(T_P|H_N) \\ &= P(T_P^c|H_N) \\ &= P(T_N|H_N) \end{aligned}$$

The proportion of sick people that will test positive is called the sensitivity of a test and is also obtained as

$$\begin{aligned} 1 - \text{false negative rate} &= 1 - P(T_N|H_P) \\ &= P(T_N^c|H_P) \\ &= P(T_P|H_P) \end{aligned}$$

**Example 2.6.5** (Sensitivity and specificity)**.** A test to detect prostate cancer has a sensitivity of 95% and a specificity of 80%. It is estimate on average that 1 in 1,439 men in the USA are afflicted by prostate cancer. If a man tests positive for prostate cancer, what is the probability that he actually has cancer? Let $T_N$ and $T_P$ represent testing negatively and testing positively, and let $C_N$ and $C_P$ denote being healthy and having prostate cancer, respectively. Then,

$$\begin{aligned} \Pr(T_P \mid C_P) &= 0.95, \\ \Pr(T_N \mid C_N) &= 0.80, \\ \Pr(C_P) &= \tfrac{1}{1439} \approx 0.0007, \end{aligned}$$

and

$$\Pr(C_P \mid T_P) = \frac{\Pr(C_P \cap T_P)}{\Pr(T_P)} = \frac{\Pr(C_P)\Pr(T_P \mid C_P)}{\Pr(T_P)} = \frac{(0.0007)(0.95)}{\Pr(T_P)},$$

where

$$\begin{aligned} \Pr(T_P) &= \Pr(C_P)\Pr(T_P \mid C_P) + \Pr(C_N)\Pr(T_P \mid C_N) \\ &= (0.0007)(0.95) + (0.9993)(.20) = 0.2005. \end{aligned}$$

Thus,

$$\Pr(C_P \mid T_P) = \frac{(0.0007)(0.95)}{0.2005} = 0.00392.$$

And so if a man tests positive for prostate cancer, there is a less than .4% probability of actually having prostate cancer.

# Chapter 3

# Random Variables

## 3.1 Definition of Random Variable

Suppose we toss a coin $N$ times in a row and that the probability of a head is $p$, where $0 < p < 1$; i.e., $\Pr(H) = p$. Then $\Pr(T) = 1 - p$. The sample space for this experiment is $\mathcal{C}$, the collection of all possible sequences of $H$'s and $T$'s of length $N$. Thus, $\mathcal{C}$ contains $2^N$ elements. The set of all subsets of $\mathcal{C}$, which contains $2^{2^N}$ elements, is the $\sigma$–field we'll be working with. Suppose we pick an element of $\mathcal{C}$, call it $c$. One thing we might be interested in is "How many heads ($H$'s) are in the sequence $c$?" This defines a function which we can call, $X$, from $\mathcal{C}$ to the set of integers. Thus

$$X(c) = \text{ the number of } H\text{'s in } c.$$

This is an example of a random variable.

More generally,

**Definition 3.1.1** (Random Variable)**.** Given a probability space $(\mathcal{C}, \mathcal{B}, \Pr)$, a random variable $X$ is a function $X : \mathcal{C} \to \mathbb{R}$ for which the set $\{c \in \mathcal{C} \mid X(c) \leqslant a\}$ is an element of the $\sigma$–field $\mathcal{B}$ for every $a \in \mathbb{R}$.

Thus we can compute the probability

$$\Pr[\{c \in C | X(c) \leq a\}]$$

for every $a \in \mathbb{R}$.

Notation. We denote the set $\{c \in C | X(c) \leq a\}$ by $(X \leq a)$.

**Example 3.1.2.** The probability that a coin comes up head is $p$, for $0 < p < 1$. Flip the coin $N$ times and count the number of heads that come up. Let $X$ be that number; then, $X$ is a random variable. We compute the following probabilities:

$$P(X \leq 0) = P(X = 0) = (1 - p)^N$$

$$P(X \leq 1) = P(X = 0) + P(X = 1) = (1 - p)^N + Np(1 - p)^{N-1}$$

There are two kinds of random variables:

(1) $X$ is discrete if $X$ takes on values in a finite set, or countable infinite set (that is, a set whose elements can be listed in a sequence $x_1, x_2, x_3, \ldots$)

(2) $X$ is continuous if $X$ can take on any value in interval of real numbers.

**Example 3.1.3** (Discrete Random Variable)**.** Flip a coin three times in a row and let $X$ denote the number of heads that come up. Then, the possible values of $X$ are 0, 1, 2 or 3. Hence, $X$ is discrete.

## 3.2 Distribution Functions

**Definition 3.2.1** (Probability Mass Function)**.** Given a discrete random variable $X$ defined on a probability space $(\mathcal{C}, \mathcal{B}, \Pr)$, the *probability mass function,* or pmf, of $X$ is define by

$$p(x) = \Pr(X = x)$$

for all $X \in \mathbb{R}$.

**Remark 3.2.2.** Here we adopt the convention that random variables will be denoted by capital letters ($X$, $Y$, $Z$, etc.) and there values are denoted by the corresponding lower case letters ($x$, $x$, $z$, etc.).

**Example 3.2.3** (Probability Mass Function)**.** Assume the coin in Example 3.1.3 is fair. Then, all the outcomes in then sample space

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

are equally likely. It then follows that

$$\Pr(X = 0) = \Pr(\{TTT\}) = \frac{1}{8},$$

$$\Pr(X = 1) = \Pr(\{HTT, THT, TTH\}) = \frac{3}{8},$$

$$\Pr(X = 2) = \Pr(\{HHT, HTH, THH\}) = \frac{3}{8},$$

and

$$\Pr(X = 3) = \Pr(\{HHH\}) = \frac{1}{8}.$$

We then have that the pmf for $X$ is

$$p(x) = \begin{cases} 1/8 & \text{if } x = 0, \\ 3/8 & \text{if } x = 1, \\ 3/8 & \text{if } x = 2, \\ 1/8 & \text{if } x = 0, \\ 0 & \text{otherwise.} \end{cases}$$

A graph of this function is picture in Figure 3.2.1.

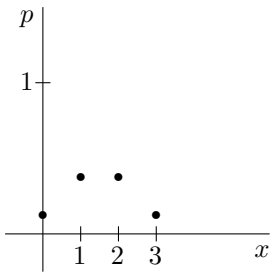



Figure 3.2.1: Probability Mass Function for $X$

**Remark 3.2.4** (Properties of probability mass functions)**.** In the previous example observe that the values of $p(x)$ are non–negative and add up to 1; that is, $p(x) \geqslant 0$ for all $x$ and

$$\sum_x p(x) = 1.$$

We usually just list the values of $X$ for which $p(x)$ is not 0:

$$p(x_1)\ p(x_2),\ \ldots,\ p(x_N),$$

in the case in which $X$ takes on a finite number of non–zero values, or

$$p(x_1),\ p(x_2),\ p(x_3),\ \ldots$$

if $X$ takes on countably many non–zero values. We then have that

$$\sum_{k=1}^{N} p(x_k) = 1$$

in the finite case, and

$$\sum_{k=1}^{\infty} p(x_k) = 1$$

in the countably infinite case.

**Definition 3.2.5** (Cumulative Distribution Function)**.** Given any random variable $X$ defined on a probability space $(\mathcal{C}, \mathcal{B}, \Pr)$, the *cumulative distribution function,* or cmf, of $X$ is define by

$$F_X(x) = \Pr(X \leqslant x)$$

for all $X \in \mathbb{R}$.

**Example 3.2.6** (Cumulative Distribution Function)**.** Let $(\mathcal{C}, \mathcal{B}, \Pr)$ and $X$ be as defined in Example 3.2.3. We compute $F_X$ as follows:

First observe that if $x < 0$, then $\Pr(X \leqslant x) = 0$; thus,

$$F_X(x) = 0 \quad \text{for all } x < 0.$$

Note that $p(x) = 0$ for $0 < x < 1$; it then follows that

$$F_X(x) = \Pr(X \leqslant x) = \Pr(X = 0) \quad \text{for } 0 < x < 1.$$

On the other hand, $\Pr(X \leqslant 1) = \Pr(X = 0) + \Pr(X = 1) = 1/8 + 3/8 = 1/2$; thus,

$$F_X(1) = 1/2.$$

Next, since $p(x) = 0$ for all $1 < x < 2$, we also get that

$$F_X(x) = 1/2 \quad \text{for } 1 < x < 2.$$

Continuing in this fashion we obtain the following formula for $F_X$:

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1/8 & \text{if } 0 \leqslant x < 1, \\ 1/2 & \text{if } 1 \leqslant x < 2, \\ 7/8 & \text{if } 2 \leqslant x < 3, \\ 1 & \text{if } x \geqslant 3. \\ 0 & \text{otherwise.} \end{cases}$$

Figure 3.2.2 shows the graph of $F_X$.

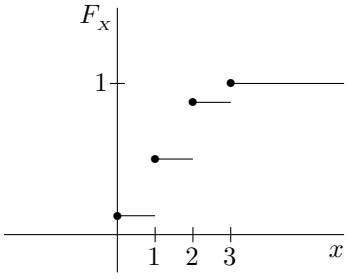



Figure 3.2.2: Cumulative Distribution Function for $X$

**Remark 3.2.7** (Properties of Cumulative Distribution Functions)**.** The graph in Figure 3.2.2 illustrates several properties of cumulative distribution functions which we will prove later in the course.

(1) $F_X$ is non–negative and non–degreasing; that is, $F_X(x) \geqslant 0$ for all $x \in \mathbb{R}$ and $F_X(a) \leqslant F_X(b)$ whenever $a < b$.

(2) $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to+\infty} F_X(x) = 1$.

(3) $F_X$ is right–continuous or upper semi–continuous; that is,

$$\lim_{x\to a^+} F_X(x) = F_X(a)$$

for all $a \in \mathbb{R}$. Observe that the limit is taken as $x$ approaches $a$ from the right.

**Example 3.2.8** (Service time at a checkout counter)**.** Suppose you sit by a checkout counter at a supermarket and measure the time, $T$, it takes for each customer to be served. This is a continuous random variable that takes on values in a time continuum. We would like to compute the cumulative distribution function $F_T(t) = \Pr(T \leqslant t)$, for all $t > 0$.

Let $N(t)$ denote the number of customers being served at a checkout counter (not in line) at time $t$. Then $N(t) = 1$ or $N(t) = 0$. Let $p(t) = P[N(t) = 1]$ and assume that $p(t)$ is a differentiable function of $t$. Assume also that $p(0) = 1$; that is, at the start of the observation, one person is being served.

Consider now $p(t + \Delta t)$, where $\Delta t$ is very small; i.e., the probability that a person is being served at time $t + \Delta t$. Suppose that the probability that service will be completed in the short time interval $[t, t + \Delta t]$ is proportional to $\Delta t$; say $\mu\Delta t$, where $\mu > 0$ is a proportionality constant. Then, the probability that service will not be completed at $t + \Delta t$ is $1 - \mu\Delta t$. This situation is illustrated in the *state diagram* pictured in Figure 3.2.3:

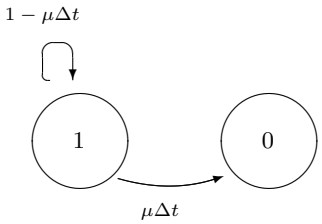



Figure 3.2.3: State diagram for $N(t)$

The circles in the state diagram represent the possible values of $N(t)$, or *states.* In this case, the states are 1 or 0, corresponding to one person being served and no person being served, respectively. The arrows represent *transition probabilities* from one state to another (or the same) in the interval from $t$ to $t + \Delta t$. Thus the probability of going from sate $N(t) = 1$ to state $N(t) = 0$ in that interval (that is, service is completed) is $\mu\Delta t$, while the probability that the person will still be at the counter at the end of the interval is $1 - \mu\Delta t$.

We therefore get that

$$p(t+\Delta t) = (\text{probability person is being served at t})(1-\mu\Delta t);$$

that is,

$$p(t+\Delta t) = p(t)(1-\mu\Delta t),$$

or

$$p(t+\Delta t) - p(t) = -\mu\Delta + p(t).$$

Dividing by $\Delta t \neq 0$ we therefore get that

$$\frac{p(t=\Delta t) - p(t)}{\Delta t} = -\mu p(t)$$

Thus, letting $\Delta t \to 0$ and using the the assumption that $p$ is differentiable, we get

$$\frac{dp}{dt} = -\mu p(t).$$

Since$p(0) = 1$, we get $p(t) = e^{-\mu t}$ for $t \geqslant 0$.

Recall that $T$ denotes the time it takes for service to be completed, or the service time at the checkout counter. Then, it is the case that

$$\begin{aligned} \Pr[T > t] &= p[N(t) = 1] \\ &= p(t) \\ &= e^{-\mu t} \end{aligned}$$

for all $t > 0$. Thus,

$$\Pr[T \leq t] = 1 - e^{-\mu t}$$

Thus, $T$ is a continuous random variable with *cdf*.

$$F_{_T}(t) = 1 - e^{-\mu t}, \quad t > 0.$$

A graph of this cdf is shown in Figure 3.2.4.

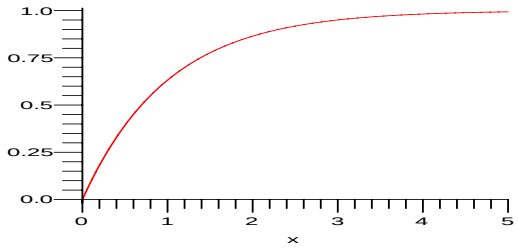



Figure 3.2.4: Cumulative Distribution Function for $T$

**Definition 3.2.9.** Let $X$ denote a continuous random variable such that $F_X(x)$ is differentiable. Then, the derivative $f_T(x) = F'_X(x)$ is called the **probability density function**, or pdf, of X.

**Example 3.2.10.** In the service time example, Example 3.2.8, if $T$ is the time that it takes for service to be completed at a checkout counter, then the cdf for $T$ is

$$F_T(t) = 1 - e^{-\mu t} \quad \text{for all } t \geqslant 0.$$

Thus,

$$f_T(t) = \mu e^{-\mu t}, \quad \text{for all } t > 0,$$

is the pdf for $T$, and we say that $T$ follows an **exponential distribution** with parameter $1/\mu$. We will see the significance of the parameter $\mu$ in the next chapter.

In general, given a function $f\colon \mathbb{R} \to \mathbb{R}$, which is non–negative and integrable with

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

$f$ defines the pdf for some continuous random variable $X$. In fact, the cdf for $X$ is defined by

$$F_X(x) = \int_{-\infty}^{x} f(t)\ \mathrm{d}t \quad \text{for all } x \in \mathbb{R}.$$

**Example 3.2.11.** Let $a$ and $b$ be real numbers with $a < b$. The function

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

defines a pdf since

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

and $f$ is non–negative.

**Definition 3.2.12** (Uniform Distribution)**.** A continuous random variable, $X$, having the pdf given in the previous example is said to be uniformly distributed on the interval $(a, b)$. We write

$$X \sim \text{Uniform}(a, b).$$

**Example 3.2.13** (Finding the distribution for the square of a random variable)**.** Let $X \sim \text{Uniform}(-1, 1)$ and $Y = X^2$ give the pdf for $Y$.

***Solution:*** Since $X \sim \text{Uniform}(-1, 1)$ its pdf is given by

$$f_X(x) = \begin{cases} \frac{1}{2} & \text{if } -1 < x < 1, \\ \\ 0 & \text{otherwise.} \end{cases}$$

We would like to compute $f_Y(y)$ for $0 < y < 1$. In order to do this, first we compute the cdf $F_Y(y)$ for $0 < y < 1$:

$$\begin{aligned} F_Y(y) &= \Pr(Y \leqslant y), \quad \text{for } 0 < y < 1, \\ &= \Pr(X^2 \leqslant y) \\ &= \Pr(|Y| \leqslant \sqrt{y}) \\ &= \Pr(-\sqrt{y} \leqslant X \leqslant \sqrt{y}) \\ &= \Pr(-\sqrt{y} < X \leqslant \sqrt{y}), \quad \text{since } X \text{ is continuous,} \\ &= \Pr(X \leqslant \sqrt{y}) - \Pr(X \leqslant -\sqrt{y}) \\ &= F_X(\sqrt{y}) - F_x(-\sqrt{y}). \end{aligned}$$

Differentiating with respect to $y$ we then obtain that

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) \\ \\ &= \frac{d}{dy} F_X(\sqrt{y}) - \frac{d}{dy} F_X(-\sqrt{y}) \\ \\ &= F'_X(\sqrt{y}) \cdot \frac{d}{dy}\sqrt{y} - F'_X(-\sqrt{y}) \frac{d}{dy}(-\sqrt{y}), \end{aligned}$$

by the Chain Rule, so that

$$\begin{aligned} f_Y(y) &= f_X(\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} + f'_X(-\sqrt{y}) \frac{1}{2\sqrt{y}} \\ \\ &= \frac{1}{2} \cdot \frac{1}{2\sqrt{y}} + \frac{1}{2} \cdot \frac{1}{2\sqrt{y}} \\ \\ &= \frac{1}{2\sqrt{y}} \end{aligned}$$

for $0 < y < 1$. We then have that

$$f_Y(y) = \begin{cases} \frac{1}{2\sqrt{y}} & \text{if } 0 < y < 1, \\ 0 & \text{otherwise.} \end{cases}$$

□

# Chapter 4

# Expectation of Random Variables

## 4.1 Expected Value of a Random Variable

**Definition 4.1.1** (Expected Value of a Continuous Random Variable)**.** Let $X$ be a continuous random variable with pdf $f_X$. If

$$\int_{-\infty}^{\infty} |x| f_X(x)\, \mathrm{d}x < \infty,$$

we define the **expected value of** $X$, denoted $E(X)$, by

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)\, \mathrm{d}x.$$

**Example 4.1.2** (Average Service Time)**.** In the service time example, Example 3.2.8, we show that the time, $T$, that it takes for service to be completed at checkout counter has an exponential distribution with pdf

$$f_T(t) = \begin{cases} \mu e^{-\mu t} & \text{for } t > 0, \\ 0 & \text{otherwise,} \end{cases}$$

where $\mu$ is a positive parameter.

Observe that

$$\begin{aligned}
\int_{-\infty}^{\infty} |t| f_{_T}(t)\ \mathrm{d}t &= \int_0^{\infty} t\mu e^{-\mu t}\ \mathrm{d}t \\
&= \lim_{b\to\infty} \int_0^b t\ \mu e^{-\mu t}\ \mathrm{d}t \\
&= \lim_{b\to\infty} \left[ -te^{-\mu t} - \frac{1}{\mu} e^{-\mu t} \right]_0^b \\
&= \lim_{b\to\infty} \left[ \frac{1}{\mu} - be^{-\mu b} - \frac{1}{\mu} e^{-\mu b} \right] \\
&= \frac{1}{\mu},
\end{aligned}$$

where we have used integration by parts and L'Hospital's rule. It then follows that

$$\int_{-\infty}^{\infty} |t| f_{_T}(t)\ \mathrm{d}t = \frac{1}{\mu} < \infty$$

and therefore the expected value of $T$ exists and

$$E(T) = \int_{-\infty}^{\infty} t f_{_T}(t)\ \mathrm{d}t = \int_0^{\infty} t\mu e^{-\mu t}\ \mathrm{d}t = \frac{1}{\mu}.$$

Thus, the parameter $\mu$ is the reciprocal of the expected service time, or **average service time**, at the checkout counter.

**Example 4.1.3.** Suppose the average service time, or **mean service time**, at a checkout counter is 5 minutes. Compute the probability that a given person will spend at least 6 minutes at the checkout counter.

***Solution:*** We assume that the service time, $T$, is exponentially distributed with pdf

$$f_{_T}(t) = \begin{cases} \mu e^{-\mu t} & \text{for } t > 0, \\ 0 & \text{otherwise,} \end{cases}$$

where $\mu = 1/5$. We then have that

$$\Pr(T \geqslant 6) = \int_6^{\infty} f_{_T}(t)\ \mathrm{d}t = \int_6^{\infty} \frac{1}{5} e^{-t/5}\ \mathrm{d}t = e^{-6/5} \approx 0.30.$$

Thus, there is a 30% chance that a person will spend 6 minutes or more at the checkout counter. □

**Definition 4.1.4** (Exponential Distribution)**.** A continuous random variable, $X$, is said to be exponentially distributed with parameter $\lambda > 0$, written

$$X \sim \text{Exponential}(\lambda),$$

if it has a pdf given by

$$f_X(x) = \begin{cases} \frac{1}{\lambda} e^{-x/\lambda} & \text{for } x > 0, \\ \\ 0 & \text{otherwise.} \end{cases}$$

The expected value of $X \sim \text{Exponential}(\lambda)$, for $\lambda > 0$, is $E(X) = \lambda$.

**Definition 4.1.5** (Expected Value of a Discrete Random Variable)**.** Let $X$ be a discrete random variable with pmf $p_X$. If

$$\sum_x |x| p_X(x)\ \mathrm{d}x < \infty,$$

we define the **expected value of** $X$, denoted $E(X)$, by

$$E(X) = \sum_x x\ p_X(x)\ \mathrm{d}x.$$

**Example 4.1.6.** Let $X$ denote the number on the top face of a balanced die. Compute $E(X)$.

> ***Solution:*** In this case the pmf of $X$ is $p_X(x) = 1/6$ for $x = 1, 2, 3, 4, 5, 6$, zero elsewhere. Then,
>
> $$E(X) = \sum_{k=1}^{6} k p_X(k) = \sum_{k=1}^{6} k \cdot \frac{1}{6} = \frac{7}{2} = 3.5.$$
>
> □

**Definition 4.1.7** (Bernoulli Trials)**.** A **Bernoulli Trial**, $X$, is a discrete random variable that takes on only the values of 0 and 1. The event $(X = 1)$ is called a "success", while $(X = 0)$ is called a "failure." The probability of a success is denoted by $p$, where $0 < p < 1$. We then have that the pmf of $X$ is

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

If a discrete random variable $X$ has this pmf, we write

$$X \sim \text{Bernoulli}(p),$$

and say that $X$ is a Bernoulli trial with parameter $p$.

**Example 4.1.8.** Let $X \sim \text{Bernoulli}(p)$. Compute $E(X)$.

***Solution:*** Compute

$$E(X) = 0 \cdot p_{_X}(0) + 1 \cdot p_{_X}(1) = p.$$

□

**Definition 4.1.9** (Independent Discrete Random Variable)**.** Two discrete random variables $X$ and $Y$ are said to independent if and only if

$$\Pr(X = x, Y = y) = \Pr(X = x) \cdot \Pr(Y = y)$$

for all values, $x$, of $X$ and all values, $y$, of $Y$.

**Note:** the event $(X = x, Y = y)$ denotes the event $(X = x) \cap (Y = y)$; that is, the events $(X = x)$ and $(Y = y)$ occur jointly.

**Example 4.1.10.** Suppose $X_1 \sim \text{Bernoulli}(p)$ and $X_2 \sim \text{Bernoulli}(p)$ are independent random variables with $0 < p < 1$. Define $Y_2 = X_1 + X_2$. Find the pmf for $Y_2$ and compute $E(Y_2)$.

***Solution:*** Observe that $Y_2$ takes on the values 0, 1 and 2. We compute

$$\begin{aligned}
\Pr(Y_2 = 0) &= \Pr(X_1 = 0, X_2 = 0) \\
&= \Pr(X_1 = 0) \cdot \Pr(X_2 = 0), \quad \text{by independence,} \\
&= (1 - p) \cdot (1 - p) \\
&= (1 - p)^2.
\end{aligned}$$

Next, since the event $(Y_2 = 1)$ consists of the disjoint union of the events $(X_1 = 1, X_2 = 0)$ and $(X_1 = 0, X_2 = 1)$,

$$\begin{aligned}
\Pr(Y_2 = 1) &= \Pr(X_1 = 1, X_2 = 0) + \Pr(X_1 = 0, X_2 = 1) \\
&= \Pr(X_1 = 1) \cdot \Pr(X_2 = 0) + \Pr(X_1 = 0) \cdot \Pr(X_2 = 1) \\
&= p(1 - p) + (1 - p)p \\
&= 2p(1 - p).
\end{aligned}$$

Finally,

$$\begin{aligned}
\Pr(Y_2 = 2) &= \Pr(X_1 = 1, X_2 = 1) \\
&= \Pr(X_1 = 1) \cdot \Pr(X_2 = 1) \\
&= p \cdot p \\
&= p^2.
\end{aligned}$$

We then have that the pmf of $Y_2$ is given by

$$p_{_{Y_2}}(y) = \begin{cases} (1 - p)^2 & \text{if } y = 0, \\ 2p(1 - p) & \text{if } y = 1, \\ p^2 & \text{if } y = 2. \end{cases}$$

To find $E(Y_2)$, compute

$$\begin{aligned} E(Y_2) &= 0 \cdot p_{Y_2}(0) + 1 \cdot p_{Y_2}(1) + 2 \cdot p_{Y_2}(2) \\ &= 2p(1-p) + 2p^2 \\ &= 2p[(1-p) + p] \\ &= 2p. \end{aligned}$$

□

We shall next consider the case in which we add three mutually independent Bernoulli trials. Before we present this example, we give a precise definition of mutual independence.

**Definition 4.1.11** (Mutual Independent Discrete Random Variable)**.** Three discrete random variables $X_1$, $X_2$ and $X_3$ are said to **mutually independent** if and only if

(i) they are *pair–wise independent;* that is,

$$\Pr(X_i = x_i, X_j = x_j) = \Pr(X_i = x_i) \cdot \Pr(X_j = x_j) \quad \text{for} \quad i \neq j,$$

for all values, $x_i$, of $X_i$ and all values, $x_j$, of $X_j$;

(ii) and

$$\Pr(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \Pr(X_1 = x_1) \cdot \Pr(X_2 = x_2) \cdot \Pr(X_3 = x_3).$$

**Lemma 4.1.12.** *Let $X_1$, $X_2$ and $X_3$ be mutually independent, discrete random variables and define $Y_2 = X_1 + X_2$. Then, $Y_2$ and $X_3$ are independent.*

*Proof:* Compute

$$\begin{aligned} \Pr(Y_2 = w, X_3 = z) &= \Pr(X_1 + X_2 = w, X_3 = z) \\ &= \Pr(X_1 = x, X_2 = w - x, X_3 = z) \\ &= \Pr(X_1 = x) \cdot \Pr(X_2 = w - x) \cdot \Pr(X_3 = z), \end{aligned}$$

where we have used (ii) in Definition 4.1.11. Thus, by pairwise independence, (i.e., (i) in Definition 4.1.11),

$$\begin{aligned} \Pr(Y_2 = w, X_3 = z) &= \Pr(X_1 = x) \cdot \Pr(X_2 = w - x) \cdot \Pr(X_3 = z) \\ &= \Pr(X_1 + X_2 = w) \cdot \Pr(X_3 = z) \\ &= \Pr(Y_2 = w) \cdot \Pr(X_3 = z), \end{aligned}$$

which shows the independence of $Y_2$ and $X_3$. □

**Example 4.1.13.** Suppose $X_1$, $X_2$ and $X_3$ be three mutually independent Bernoulli random variables with parameter $p$, where $0 < p < 1$. Define $Y_3 = X_1 + X_2 + X_3$. Find the pmf for $Y_3$ and compute $E(Y_3)$.

***Solution:*** Observe that $Y_3$ takes on the values 0, 1, 2 and 3, and that

$$Y_3 = Y_2 + X_3,$$

where the pmf and expected value of $Y_2$ were computed in Example 4.1.10.

We compute

$$\begin{aligned}
\Pr(Y_3 = 0) &= \Pr(Y_2 = 0, X_3 = 0) \\
&= \Pr(Y_2 = 0) \cdot \Pr(X_3 = 0), \text{ by independence (Lemma 4.1.12),} \\
&= (1-p)^2 \cdot (1-p) \\
&= (1-p)^3.
\end{aligned}$$

Next, since the event $(Y_3 = 1)$ consists of the disjoint union of the events $(Y_2 = 1, X_3 = 0)$ and $(Y_2 = 0, X_3 = 1)$,

$$\begin{aligned}
\Pr(Y_3 = 1) &= \Pr(Y_2 = 1, X_3 = 0) + \Pr(Y_2 = 0, X_3 = 1) \\
&= \Pr(Y_2 = 1) \cdot \Pr(X_3 = 0) + \Pr(Y_2 = 0) \cdot \Pr(X_3 = 1) \\
&= 2p(1-p)(1-p) + (1-p)^2 p \\
&= 3p(1-p)^2.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\Pr(Y_3 = 2) &= \Pr(Y_2 = 2, X_3 = 0) + \Pr(Y_2 = 1, X_3 = 1) \\
&= \Pr(Y_2 = 2) \cdot \Pr(X_3 = 0) + \Pr(Y_2 = 1) \cdot \Pr(X_3 = 1) \\
&= p^2(1-p) + 2p(1-p)p \\
&= 3p^2(1-p),
\end{aligned}$$

and

$$\begin{aligned}
\Pr(Y_3 = 3) &= \Pr(Y_2 = 2, X_3 = 1) \\
&= \Pr(Y_2 = 0) \cdot \Pr(X_3 = 0) \\
&= p^2 \cdot p \\
&= p^3.
\end{aligned}$$

We then have that the pmf of $Y_3$ is

$$p_{Y_3}(y) = \begin{cases} (1-p)^3 & \text{if } y = 0, \\ 3p(1-p)^2 & \text{if } y = 1, \\ 3p^2(1-p) & \text{if } y = 2 \\ p^3 & \text{if } y = 3. \end{cases}$$

To find $E(Y_2)$, compute

$$\begin{aligned}
E(Y_2) &= 0 \cdot p_{Y_3}(0) + 1 \cdot p_{Y_3}(1) + 2 \cdot p_{Y_3}(2) + 3 \cdot p_{Y_3}(3) \\
&= 3p(1-p)^2 + 2 \cdot 3p^2(1-p) + 3p^3 \\
&= 3p[(1-p)^2 + 2p(1-p) + p^2] \\
&= 3p[(1-p) + p]^2 \\
&= 3p.
\end{aligned}$$

□

If we go through the calculations in Examples 4.1.10 and 4.1.13 for the case of four mutually independent[1] Bernoulli trials with parameter $p$, where $0 < p < 1$, $X_1$, $X_2$, $X_3$ and $X_4$, we obtain that for $Y_4 = X_1 + X_2 + X_3 + X_4$,

$$p_{Y_4}(y) = \begin{cases} (1-p)^4 & \text{if } y = 0, \\ 4p(1-p)^3 & \text{if } y = 1, \\ 6p^2(1-p)^2 & \text{if } y = 2 \\ 4p^3(1-p) & \text{if } y = 3 \\ p^4 & \text{if } y = 4, \end{cases}$$

and

$$E(Y_4) = 4p.$$

Observe that the terms in the expressions for $p_{Y_2}(y)$, $p_{Y_3}(y)$ and $p_{Y_4}(y)$ are the terms in the expansion of $[(1-p)+p]^n$ for $n = 2, 3$ and 4, respectively. By the Binomial Expansion Theorem,

$$[(1-p)+p]^n = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k},$$

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}, \quad k = 0, 1, 2 \ldots, n,$$

are the called the **binomial coefficients**. This suggests that if

$$Y_n = X_1 + X_2 + \cdots + X_n,$$

where $X_1, X_2, \ldots, X_n$ are $n$ mutually independent Bernoulli trials with parameter $p$, for $0 < p < 1$, then

$$p_{Y_n}(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k = 0, 1, 2, \ldots, n.$$

Furthermore,

$$E(Y_n) = np.$$

We shall establish this as a the following Theorem:

**Theorem 4.1.14.** *Assume that $X_1, X_2, \ldots, X_n$ are mutually independent Bernoulli trials with parameter $p$, with $0 < p < 1$. Define*

$$Y_n = X_1 + X_2 + \cdots + X_n.$$

*Then the pmf of $Y_n$ is*

$$p_{Y_n}(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \textit{for } k = 0, 1, 2, \ldots, n,$$

[1]Here we do not only require that the random variable be pairwise independent, but also that for any group of $k \geq 2$ events $(X_j = x_j)$, the probability of their intersection is the product of their probabilities.

*and*

$$E(Y_n) = np.$$

*Proof:* We prove this result by induction on $n$. For $n = 1$ we have that $Y_1 = X_1$, and therefore

$$p_{Y_1}(0) = \Pr(X_1 = 0) = 1 - p$$

and

$$p_{Y_1}(1) = \Pr(X_1 = 1) = p.$$

Thus,

$$p_{Y_1}(k) = \begin{cases} 1-p & \text{if } k = 0, \\ p & \text{if } k = 1. \end{cases}$$

Observe that $\binom{1}{0} = \binom{1}{1} = 1$ and therefore the result holds true for $n = 1$.

Next, assume the theorem is true for $n$; that is, suppose that

$$p_{Y_n}(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k = 0, 1, 2, \ldots, n, \tag{4.1}$$

and that

$$E(Y_n) = np. \tag{4.2}$$

We show then show that the result also holds true for $n + 1$. In other words, we show that if $X_1, X_2, \ldots, X_n, X_{n+1}$ are mutually independent Bernoulli trials with parameter $p$, with $0 < p < 1$, and

$$Y_{n+1} = X_1 + X_2 + \cdots + X_n + X_{n+1}, \tag{4.3}$$

then, the pmf of $Y_{n+1}$ is

$$p_{Y_{n+1}}(k) = \binom{n+1}{k} p^k (1-p)^{n+1-k} \quad \text{for } k = 0, 1, 2, \ldots, n, n+1, \tag{4.4}$$

and

$$E(Y_{n+1}) = (n+1)p. \tag{4.5}$$

From (4.5) we see that

$$Y_{n+1} = Y_n + X_{n+1},$$

where $Y_n$ and $X_{n+1}$ are independent random variables, by an argument similar to the one in the proof of Lemma 4.1.12 since the $X_k$'s are mutually independent. Therefore, the following calculations are justified:

(i) for $k \leqslant n$,

$$\begin{aligned} \Pr(Y_{n+1} = k) &= \Pr(Y_n = k, X_{n+1} = 0) + \Pr(Y_n = k-1, X_{n+1} = 1) \\ &= \Pr(Y_n = k) \cdot \Pr(X_{n+1} = 0) + \Pr(Y_n = k-1) \cdot \Pr(X_{n-1} = 1) \\ &= \binom{n}{k} p^k (1-p)^{n-k}(1-p) + \binom{n}{k-1} p^{k-1}(1-p)^{n-k+1} p, \end{aligned}$$

where we have used the inductive hypothesis (4.1). Thus,

$$\Pr(Y_{n+1} = k) = \left[\binom{n}{k} + \binom{n}{k-1}\right] p^k (1-p)^{n+1-k}.$$

The expression in (4.4) will following from the fact that

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

which can be established by the following counting argument:

> Imagine $n+1$ balls in a bag, $n$ of which are blue and one is red. We consider the collection of all groups of $k$ balls that can be formed out of the $n+1$ balls in the bag. This collection is made up of two disjoint sub–collections: the ones with the red ball and the ones without the red ball. The number of elements in the collection with the one red ball is
>
> $$\binom{n}{k-1} \cdot \binom{1}{1} = \binom{n}{k-1},$$
>
> while the number of elements in the collection of groups without the red ball are
>
> $$\binom{n}{k}.$$
>
> Adding these two must yield $\binom{n+1}{k}$.

(ii) If $k = n+1$, then

$$\begin{aligned} \Pr(Y_{n+1} = k) &= \Pr(Y_n = n, X_{n+1} = 1) \\ &= \Pr(Y_n = n) \cdot \Pr(X_{n+1} = 1) \\ &= p^n p \\ &= p^{n+1} \\ \\ &= \binom{n+1}{k} p^k (1-p)^{n+1-k}, \end{aligned}$$

since $k = n+1$.

Finally, to establish (4.5) based on (4.2), use the result of problem 2 in Assignment 10 to show that, since $Y_n$ and $X_n$ are independent,

$$E(Y_{n+1}) = E(Y_n + X_{n+1}) = E(Y_n) + E(X_{n+1}) = np + p = (n+1)p.$$

□

**Definition 4.1.15** (Binomial Distribution)**.** Let $b$ be a natural number and $0 < p < 1$. A discrete random variable, $X$, having pmf

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k = 0, 1, 2, \ldots, n,$$

is said to have a **binomial distribution** with parameters $n$ and $p$.

We write $X \sim \text{Binomial}(n, p)$.

**Remark 4.1.16.** In Theorem 4.1.14 we showed that if $X \sim \text{Binomial}(n, p)$, then

$$E(X) = np.$$

We also showed in that theorem that the sum of $n$ mutually independent Bernoulli trials with parameter $p$, for $0 < p < 1$, follows a Binomial distribution with parameters $n$ and $p$.

**Definition 4.1.17** (Independent Identically Distributed Random Variables)**.** A set of random variables, $\{X_1, X_2, \ldots, X_n\}$, is said be **independent identically distributed**, or iid, if the random variables are mutually disjoint and if they all have the same distribution function.

If the random variables $X_1, X_2, \ldots, X_n$ are iid, then they form a **simple random sample** of size $n$.

**Example 4.1.18.** Let $X_1, X_2, \ldots, X_n$ be a simple random sample from a Bernoulli($p$) distribution, with $0 < p < 1$. Define the **sample mean** $\overline{X}$ by

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

Give the distribution function for $X$ and compute $E(\overline{X})$.

> ***Solution:*** Write $Y = n\overline{X} = X_1 + X_2 + \cdots + X_n$. Then, since $X_1, X_2, \ldots, X_n$ are iid Bernoulli($p$) random variables, Theorem 4.1.14 implies that $Y \sim \text{Binomial}(n, p)$. Consequently, the pmf of $Y$ is
>
> $$p_Y(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k = 0, 1, 2, \ldots, n,$$
>
> and $E(Y) = np$.
>
> Now, $\overline{X}$ may take on the values $0, \frac{1}{n}, \frac{2}{n}, \ldots \frac{n-1}{n}, 1$, and
>
> $$\Pr(\overline{X} = x) = \Pr(Y = nx) \quad \text{for } x = 0, \frac{1}{n}, \frac{2}{n}, \ldots \frac{n-1}{n}, 1,$$
>
> so that
>
> $$\Pr(\overline{X} = x) = \binom{n}{nx} p^{nx} (1-p)^{n-nx} \quad \text{for } x = 0, \frac{1}{n}, \frac{2}{n}, \ldots \frac{n-1}{n}, 1.$$

The expected value of $\overline{X}$ can be computed as follows

$$E(\overline{X}) = E\left(\frac{1}{n}Y\right) = \frac{1}{n}E(Y) = \frac{1}{n}(np) = p.$$

Observe that $\overline{X}$ is the proportion of successes in the simple random sample. It then follows that the expected proportion of successes in the random sample is $p$, the probability of a success. □

## 4.2 Law of the Unconscious Statistician

**Example 4.2.1.** Let $X$ denote a continuous random variable with pdf

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

Compute the expected value of $X^2$.

We show two ways to compute $E(X)$.

(i) *First Alternative.* Let $Y = X^2$ and compute the pdf of $Y$. To do this, first we compute the cdf:

$$\begin{aligned} F_Y(y) &= \Pr(Y \leqslant y), \quad \text{for } 0 < y < 1, \\ &= \Pr(X^2 \leqslant y) \\ &= \Pr(|X| \leqslant \sqrt{y}) \\ &= \Pr(-\sqrt{y} \leqslant |X| \leqslant \sqrt{y}) \\ &= \Pr(-\sqrt{y} < |X| \leqslant \sqrt{y}), \quad \text{since } X \text{ is continuous,} \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}) \\ &= F_X(\sqrt{y}), \end{aligned}$$

since $f_X$ is 0 for negative values.

It then follows that

$$\begin{aligned} f_Y(y) &= F'_X(\sqrt{y}) \cdot \frac{d}{dy}(\sqrt{y}) \\ &= f_X(\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} \\ &= 3(\sqrt{y})^2 \cdot \frac{1}{2\sqrt{y}} \\ &= \frac{3}{2}\sqrt{y} \end{aligned}$$

for $0 < y < 1$.

Consequently, the pdf for $Y$ is

$$f_Y(y) = \begin{cases} \frac{3}{2}\sqrt{y} & \text{if } 0 < y < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore,

$$\begin{aligned} E(X^2) &= E(Y) \\ &= \int_{-\infty}^{\infty} y f_Y(y)\ \mathrm{d}y \\ &= \int_0^1 y\ \frac{3}{2}\sqrt{y}\ \mathrm{d}y \\ &= \frac{3}{2}\int_0^1 y^{3/2}\ \mathrm{d}y \\ &= \frac{3}{2}\cdot\frac{2}{5} \\ &= \frac{3}{5}. \end{aligned}$$

(ii) *Second Alternative.* Alternatively, we could have compute $E(X^2)$ by evaluating

$$\begin{aligned} \int_{-\infty}^{\infty} x^2 f_X(x)\ \mathrm{d}x &= \int_0^1 x^2 \cdot 3x^2\ \mathrm{d}x \\ &= \frac{3}{5}. \end{aligned}$$

The fact that both ways of evaluating $E(X^2)$ presented in the previous example is a consequence of the so–called *Law of the Unconscious Statistician:*

**Theorem 4.2.2** (Law of the Unconscious Statistician, Continuous Case)**.** *Let $X$ be a continuous random variable and $g$ denote a continuous function defined on the range of $X$. Then, if*

$$\int_{-\infty}^{\infty} |g(x)| f_X(x)\ dx < \infty,$$

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

*Proof:* We prove this result for the special case in which $g$ is differentiable with $g'(x) > 0$ for all x in the range of $X$. In this case $g$ is strictly increasing and it

therefore has an inverse function $g^{-1}$ mapping onto the range of $X$ and which is also differentiable with derivative given by

$$\frac{d}{dy}\left[g^{-1}(y)\right] = \frac{1}{g'(x)} = \frac{1}{g'(g^{-1}(y))},$$

where we have set $y = g(x)$ for all $x$ is the range of $X$, or $Y = g(X)$. Assume also that the values of $X$ range from $-\infty$ to $\infty$ and those of $Y$ also range from $-\infty$ to $\infty$. Thus, using the Change of Variables Theorem, we have that

$$\int_{-\infty}^{\infty} g(x) f_X(x)\ \mathrm{d}x \quad = \quad \int_{-\infty}^{\infty} y f_X(g^{-1}(y)) \cdot \frac{1}{g\prime(g^{-1}(y))}\ \mathrm{d}y,$$

since $x = g^{-1}(y)$ and therefore

$$\mathrm{d}x = \frac{d}{dy}\left[g^{-1}(y)\right]\ \mathrm{d}y = \frac{1}{g'(g^{-1}(y))}\ \mathrm{d}y.$$

On the other hand,

$$\begin{aligned} F_Y(y) &= \Pr(Y \leqslant y) \\ &= \Pr(g(X) \leqslant y) \\ &= \Pr(X \leqslant g^{-1}(y)) \\ &= F_X(g^{-1}(y)), \end{aligned}$$

from which we obtain, by the Chain Rule, that

$$f_Y(y) = f_X(g^{-1}(y)\frac{1}{g'(g^{-1}(y))}.$$

Consequently,

$$\int_{-\infty}^{\infty} y f_Y(y)\ \mathrm{d}y = \int_{-\infty}^{\infty} g(x) f_X(x)\ \mathrm{d}x,$$

or

$$E(Y) = \int_{-\infty}^{\infty} g(x) f_X(x)\ \mathrm{d}x.$$

□

The law of the unconscious statistician also applies to functions of a discrete random variable. In this case we have

**Theorem 4.2.3** (Law of the Unconscious Statistician, Discrete Case)**.** *Let $X$ be a discrete random variable with pmf $p_X$, and let $g$ denote a function defined on the range of $X$. Then, if*

$$\sum_x |g(x)| p_X(x) < \infty,$$

$$E(g(X)) = \sum_x g(x) p_X(x).$$

## 4.3 Moments

The law of the unconscious statistician can be used to evaluate the expected values of powers of a random variable

$$E(X^m) = \int_{-\infty}^{\infty} x^m f_X(x)\, \mathrm{d}x, \quad m = 0, 1, 2, 3, \ldots,$$

in the continuous case, provided that

$$\int_{-\infty}^{\infty} |x|^m f_X(x)\, \mathrm{d}x < \infty, \quad m = 0, 1, 2, 3, \ldots.$$

In the discrete case we have

$$E(X^m) = \sum_x x^m p_X(x), \quad m = 0, 1, 2, 3, \ldots,$$

provided that

$$\sum_x |x|^m p_X(x) < \infty, \quad m = 0, 1, 2, 3, \ldots.$$

**Definition 4.3.1** ($m$th Moment of a Distribution)**.** $E(X^m)$, if it exists, is called the *$m$th moment of $X$* for $m = 0, 2, 3, \ldots$

Observe that the first moment of $X$ is its expectation.

**Example 4.3.2.** Let $X$ have a uniform distribution over the interval $(a, b)$ for $a < b$. Compute the second moment of $X$.

***Solution:*** Using the law of the unconscious statistician we get

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x)\, \mathrm{d}x,$$

where

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

Thus,

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

□

### 4.3.1 Moment Generating Function

Using the law of the unconscious statistician we can also evaluate $E(e^{tX})$ whenever this expectation is defined.

**Example 4.3.3.** Let $X$ have an exponential distribution with parameter $\lambda > 0$. Determine the values of $t \in \mathbb{R}$ for which $E(e^{tX})$ is defined and compute it.

***Solution:*** The pdf of $X$ is given by

$$f_X(x) = \begin{cases} \frac{1}{\lambda} e^{-x/\lambda} & \text{if } x > 0, \\ \\ 0 & \text{if } x \leqslant 0. \end{cases}$$

Then,

$$\begin{aligned} E(e^{tX}) &= \int_{-\infty}^{\infty} e^{tx} f_X(x)\, \mathrm{d}x \\ &= \frac{1}{\lambda} \int_0^{\infty} e^{tx} e^{-x/\lambda}\, \mathrm{d}x \\ &= \frac{1}{\lambda} \int_0^{\infty} e^{-[(1/\lambda)-t]x}\, \mathrm{d}x. \end{aligned}$$

We note that for the integral in the last equation to converge, we must require that

$$t < 1/\lambda.$$

For these values of $t$ we get that

$$\begin{aligned} E(e^{tX}) &= \frac{1}{\lambda} \cdot \frac{1}{(1/\lambda) - t} \\ &= \frac{1}{1 - \lambda t}. \end{aligned}$$

□

**Definition 4.3.4** (Moment Generating Function)**.** Given a random variable $X$, the expectation $E(e^{tX})$, for those values of $t$ for which it is defined, is called the **moment generating function**, or mgf, of $X$, and is denoted by $\psi_X(t)$. We then have that

$$\psi_X(t) = E(e^{tX}),$$

whenever the expectation is defined.

**Example 4.3.5.** If $X \sim \text{Exponential}(\lambda)$, for $\lambda > 0$, then Example 4.3.3 shows that the mgf of $X$ is given by

$$\psi_X(t) = \frac{1}{1 - \lambda t} \quad \text{for } t < \frac{1}{\lambda}.$$

**Example 4.3.6.** Let $X \sim \text{Binomial}(n, p)$, for $n \geqslant 1$ and $0 < p < 1$. Compute the mgf of $X$.

***Solution:*** The pmf of $X$ is

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k = 0, 1, 2, \ldots, n,$$

and therefore, by the law of the unconscious statistician,

$$\begin{aligned} \psi_X(t) &= E(e^{tX}) \\ &= \sum_{k=1}^{n} e^{tk} p_X(k) \\ &= \sum_{k=1}^{n} (e^t)^k \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=1}^{n} \binom{n}{k} (pe^t)^k (1-p)^{n-k} \\ &= (pe^t + 1 - p)^n, \end{aligned}$$

where we have used the Binomial Theorem.

We therefore have that if $X$ is binomially distributed with parameters $n$ and $p$, then its mgf is given by

$$\psi_X(t) = (1 - p + pe^t)^n \quad \text{for all } t \in \mathbb{R}.$$

□

### 4.3.2 Properties of Moment Generating Functions

First observe that for any random variable $X$,

$$\psi_X(0) = E(e^{0 \cdot X}) = E(1) = 1.$$

The importance of the moment generating function stems from the fact that, if $\psi_X(t)$ is defined over some interval around $t = 0$, then it is infinitely differentiable in that interval and its derivatives at $t = 0$ yield the moments of $X$. More precisely, in the continuous cases, the $m$–the order derivative of $\psi_X$ at $t$ is given by

$$\psi^{(m)}(t) = \int_{-\infty}^{\infty} x^m e^{tx} f_X(x)\, \mathrm{d}x, \quad \text{for all } m = 0, 1, 2, 3, \ldots,$$

and all $t$ in the interval around 0 where these derivatives are defined. It then follows that

$$\psi^{(m)}(0) = \int_{-\infty}^{\infty} x^m f_X(x)\, \mathrm{d}x = E(X^m), \quad \text{for all } m = 0, 1, 2, 3, \ldots.$$

**Example 4.3.7.** Let $X \sim \text{Exponential}(\lambda)$. Compute the second moment of $X$.

***Solution:*** From Example 4.3.5 we have that

$$\psi_X(t) = \frac{1}{1-\lambda t} \quad \text{for } t < \frac{1}{\lambda}.$$

Differentiating with respect to $t$ we obtain

$$\psi'_X(t) = \frac{\lambda}{(1-\lambda t)^2} \quad \text{for } t < \frac{1}{\lambda}$$

and

$$\psi''_X(t) = \frac{2\lambda^2}{(1-\lambda t)^3} \quad \text{for } t < \frac{1}{\lambda}.$$

It then follows that $E(X^2) = \psi''_X(0) = 2\lambda^2$. □

**Example 4.3.8.** Let $X \sim \text{Binomial}(n, p)$. Compute the second moment of $X$.

***Solution:*** From Example 4.3.6 we have that

$$\psi_X(t) = (1-p+pe^t)^n \quad \text{for all } t \in \mathbb{R}.$$

Differentiating with respect to $t$ we obtain

$$\psi'_X(t) = npe^t(1-p+pe^t)^{n-1}$$

and

$$\begin{aligned} \psi''_X(t) &= npe^t(1-p+pe^t)^{n-1} + npe^t(n-1)pe^t(1-p+pe^t)^{n-2} \\ &= npe^t(1-p+pe^t)^{n-1} + n(n-1)p^2e^{2t}(1-p+pe^t)^{n-2} \end{aligned}$$

It then follows that $E(X^2) = \psi''_X(0) = np + n(n-1)p^2$. □

## 4.4 Variance

Given a random variable $X$ for which an expectation exists, define

$$\mu = E(X).$$

$\mu$ is usually referred to as the **mean** of the random variable $X$. For given $m = 1, 2, 3, \ldots$, if $E(|X-\mu|^m) < \infty$,

$$E[(X-\mu)^m]$$

is called the $m$**th central moment** of $X$. The second central moment of $X$,

$$E[(X-\mu)^2],$$

if it exists, is called the **variance** of $X$ and is denoted by $\text{var}(X)$. We then have that

$$\text{var}(X) = E[(X-\mu)^2],$$

provided that the expectation is finite. The variance of $X$ measures the **average square deviation** of the distribution of $X$ from its mean. Thus,

$$\sqrt{E[(X-\mu)^2]}$$

is measure or deviation from the mean. It is usually denoted by $\sigma$ and is called the **standard deviation** form the mean. We then have

$$\text{var}(X) = \sigma^2.$$

We can compute the variance of a random variable $X$ as follows:

$$\begin{aligned}
\text{var}(X) &= E[(X-\mu)^2] \\
&= E(X^2 - 2\mu X + \mu^2) \\
&= E(X^2) - 2\mu E(X) + \mu^2 E(1)) \\
&= E(X^2) - 2\mu \cdot \mu + \mu^2 \\
&= E(X^2) - \mu^2.
\end{aligned}$$

**Example 4.4.1.** Let $X \sim \text{Exponential}(\lambda)$. Compute the variance of $X$.

***Solution:*** In this case $\mu = E(X) = \lambda$ and, from Example 4.3.7, $E(X^2)2\lambda^2$. Consequently,

$$\text{var}(X) = E(X^2) - \mu^2 = 2\lambda^2 - \lambda^2 = \lambda^2.$$

□

**Example 4.4.2.** Let $X \sim \text{Binomial}(n, p)$. Compute the variance of $X$.

***Solution:*** Here, $\mu = np$ and, from Example 4.3.8, $E(X^2) = np + n(n-1)p^2$. Thus,

$$\text{var}(X) = E(X^2) - \mu^2 = np + n(n-1)p^2 - n^2p^2 = np(1-p).$$

□

# Chapter 5

# Joint Distributions

When studying the outcomes of experiments more than one measurement (random variable) may be obtained. For example, when studying traffic flow, we might be interested in the number of cars that go past a point per minute, as well as the mean speed of the cars counted. We might also be interested in the spacing between consecutive cars, and the relative speeds of the two cars. For this reason, we are "interested" in probability statements concerning two or more random variables.

## 5.1 Definition of Joint Distribution

In order to deal with probabilities for events involving more than one random variable, we need to define their **joint distribution**. We begin with the case of two random variables $X$ and $Y$.

**Definition 5.1.1** (Joint cumulative distribution function)**.** Given random variables $X$ and $Y$, the **joint cumulative distribution function** of $X$ and $Y$ is defined by

$$F_{(X,Y)}(x,y) = \Pr(X \leqslant y, Y \leqslant y) \quad \text{for all } (x,y) \in \mathbb{R}^2.$$

When both $X$ and $Y$ are discrete random variables, it is more convenient to talk about the **joint probability mass function** of $X$ and $Y$:

$$p_{(X,Y)}(x,y) = \Pr(X = x, Y = y).$$

We have already talked about the joint distribution of two discrete random variables in the case in which they are independent. Here is an example in which $X$ and $Y$ are not independent:

**Example 5.1.2.** Suppose three chips are randomly selected from a bowl containing 3 red, 4 white, and 5 blue chips.

Let $X$ denote the number of red chips chosen, and $Y$ be the number of white chips chosen. We would like to compute the joint probability function of $X$ and $Y$; that is,

$$\Pr(X = x, Y = y),$$

were $x$ and $y$ range over 0, 1, 2 and 3.

For instance,

$$\begin{aligned}
\Pr(X = 0, Y = 0) &= \frac{\binom{5}{3}}{\binom{12}{3}} = \frac{10}{220} = \frac{1}{22}, \\
\Pr(X = 0, Y = 1) &= \frac{\binom{4}{1} \cdot \binom{5}{2}}{\binom{12}{3}} = \frac{40}{220} = \frac{2}{11}, \\
\Pr(X = 1, Y = 1) &= \frac{\binom{3}{1} \cdot \binom{4}{1} \cdot \binom{5}{1}}{\binom{12}{3}} = \frac{60}{220} = \frac{3}{11},
\end{aligned}$$

and so on for all 16 of the joint probabilities. These probabilities are more easily expressed in tabular form:

| $X \backslash Y$ | 0 | 1 | 2 | 3 | Row Sums |
|---|---|---|---|---|---|
| 0 | 1/22 | 2/11 | 3/22 | 2/110 | 21/55 |
| 1 | 3/22 | 3/11 | 9/110 | 0 | 27/55 |
| 2 | 3/42 | 3/55 | 0 | 0 | 27/220 |
| 3 | 1/220 | 0 | 0 | 0 | 1/220 |
| Column Sums | 14/55 | 28/55 | 12/55 | 1/55 | 1 |

Table 5.1: Joint Probability Distribution for $X$ and $Y$, $p_{(X,Y)}$

Notice that pmf's for the individual random variables $X$ and $Y$ can be obtained as follows:

$$p_{X}(i) = \sum_{j=0}^{3} P(i,j) \leftarrow \text{ adding up } i^{th} \text{ row}$$

for $i = 1, 2, 3$, and

$$p_{Y}(j) = \sum_{i=0}^{3} P(i,j) \leftarrow \text{ adding up } j^{th} \text{ column}$$

for $j = 1, 2, 3$.

These are expressed in Table 5.1 as "row sums" and "column sums," respectively, on the "margins" of the table. For this reason $p_X$ and $p_Y$ are usually called **marginal distributions**.

Observe that, for instance,

$$0 = \Pr(X = 3, Y = 1) \neq p_{X}(3) \cdot p_{Y}(2) = \frac{1}{220} \cdot \frac{28}{55},$$

and therefore $X$ and $Y$ are not independent.

**Definition 5.1.3** (Joint Probability Density Function)**.** An integrable function $f\colon \mathbb{R}^2 \to \mathbb{R}$ is said to be a **joint pdf** of the continuous random variables $X$ and $Y$ iff

(i) $f(x,y) \geqslant 0$ for all $(x,y) \in \mathbb{R}^2$, and

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

**Example 5.1.4.** Consider the disk of radius 1 centered at the origin,

$$D_1 = \{(x,y) \in \mathbb{R}^2 \mid x^2 + y^2 < 1\}.$$

The uniform joint distribution for the disc is given by the function

$$f(x,y) = \begin{cases} \dfrac{1}{\pi} & \text{if } \ x^2+y^2 < 1, \\ \\ 0 & \text{otherwise.} \end{cases}$$

This is the joint pdf for two random variables, $X$ and $Y$, since

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x\mathrm{d}y = \iint_{D_1} \frac{1}{\pi}\,\mathrm{d}x\mathrm{d}y = \frac{1}{\pi}\cdot \mathrm{area}(D_1) = 1,$$

since $\mathrm{area}(D_1) = \pi$.

This pdf models a situation in which the random variables $X$ and $Y$ denote the coordinates of a point on the unit disk chosen at random.

If $X$ and $Y$ are continuous random variables with joint pdf $f_{(X,Y)}$, then the joint cumulative distribution function, $F_{(X,Y)}$, is given by

$$F_{(X,Y)}(x,y) = \Pr(X \leq x, Y \leq y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{(X,Y)}(u,v)\,\mathrm{d}u\mathrm{d}v,$$

for all $(x,y) \in \mathbb{R}^2$.

Given the joint pdf, $f_{(X,Y)}$, of two random variables defined on the same sample space, $\mathcal{C}$, we can compute the probability of events of the form

$$((X,Y) \in A) = \{c \in \mathcal{C} \mid (X(c), Y(c)) \in A\},$$

where $A$ is a Borel subset[1] of $\mathbb{R}^2$, is compyted as follows

$$\Pr((X,Y) \in A) = \iint_A f_{(X,Y)}(x,y)\,\mathrm{d}x\mathrm{d}y.$$

[1]Borel sets in $\mathbb{R}^2$ are generated by bounded open disks; i.e, the sets

$$\{(x,y) \in \mathbb{R}^2 \mid (x-x_o)^2 + (y-y_o)^2 < r^2\},$$

where $(x_o, y_o) \in \mathbb{R}^2$ and $r > 0$.

**Example 5.1.5.** A point is chosen at random from the open unit disk $D_1$. Compute the probability that that the sum of its coordinates is bigger than 1.

***Solution:*** Let $(X, Y)$ denote the coordinates of a point drawn at random from the unit disk $D_1 = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 < 1\}$. Then the joint pdf of $X$ and $Y$ is the one given in Example 5.1.4; that is,

$$f_{(X,Y)}(x, y) = \begin{cases} \dfrac{1}{\pi} & \text{if } \ x^2 + y^2 < 1, \\ \\ 0 & \text{otherwise.} \end{cases}$$

We wish to compute $\Pr(X + Y > 1)$. This is given by

$$\Pr(X + Y > 1) = \iint_A f_{(X,Y)}(x, y)\ \mathrm{d}x\mathrm{d}y,$$

where

$$A = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 < 1,\ x + y > 1\}.$$

The set $A$ is sketched in Figure 5.1.1.

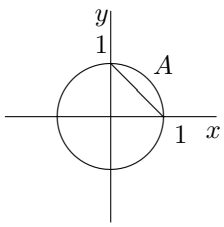



Figure 5.1.1: Sketch of region $A$

$$\begin{aligned} \Pr(X + Y > 1) &= \iint_A \frac{1}{\pi}\ \mathrm{d}x\mathrm{d}y \\ &= \frac{1}{\pi}\text{area}(A) \\ &= \frac{1}{\pi}\left(\frac{\pi}{4} - \frac{1}{2}\right) \\ &= \frac{1}{4} - \frac{1}{2\pi} \approx 0.09, \end{aligned}$$

Since the area of $A$ is the area of one quarter of that of the disk minus the are of the triangle with vertices $(0, 0)$, $(1, 0)$ and $(0, 1)$. □

## 5.2 Marginal Distributions

We have seen that if $X$ and $Y$ are discrete random variables with joint pmf $p_{(X,Y)}$, then the marginal distributions are given by

$$p_X(x) = \sum_y p_{(X,Y)}(x,y),$$

and

$$p_Y(y) = \sum_x p_{(X,Y)}(x,y),$$

where the sums are taken over all possible values of $y$ and $x$, respectively.

We can define marginal distributions for continuous random variables $X$ and $Y$ from their joint pdf in an analogous way:

$$f_X(x) = \int_{-\infty}^{\infty} f_{(X,Y)}(x,y)\ \mathrm{d}y$$

and

$$f_Y(y) = \int_{-\infty}^{\infty} f_{(X,Y)}(x,y)\mathrm{d}x.$$

**Example 5.2.1.** Let $X$ and $Y$ have joint pdf given by

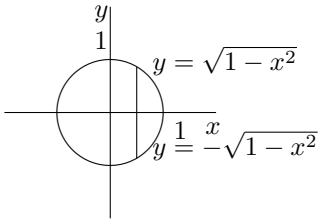
$$f_{(X,Y)}(x,y) = \begin{cases} \dfrac{1}{\pi} & \text{if } \ x^2+y^2<1, \\ \\ 0 & \text{otherwise.} \end{cases}$$

Then, the marginal distribution of $X$ is given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{(X,Y)}(x,y)\ \mathrm{d}y = \frac{1}{\pi}\int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}}\ \mathrm{d}y = \frac{2}{\pi}\sqrt{1-x^2} \text{ for } -1<x<1$$

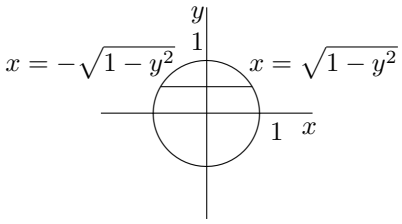
and

$$f_X(x) = 0 \text{ for } |x| \geqslant 1.$$




Similarly,

$$f_Y(y) = \begin{cases} \dfrac{2}{\pi}\sqrt{1-y^2} & \text{if } |y|<1 \\ \\ 0 & \text{if } |y| \geqslant 1. \end{cases}$$




Observe that, in the previous example,

$$f_{(X,Y)}(x,y) \neq f_X(x) \cdot f_Y(y),$$

since

$$\text{Since } \frac{1}{\pi} \neq \frac{4}{\pi^2}\sqrt{1-x^2}\sqrt{1-y^2},$$

for $(x, y)$ in the unit disk. We then say that $X$ and $Y$ are not independent.

**Example 5.2.2.** Let $X$ and $Y$ denote the coordinates of a point selected at random from the unit disc and set $Z = \sqrt{X^2 + Y^2}$

Compute the cdf for $Z$, $F_Z(z)$ for $0 < z < 1$ and the expectation of $Z$.

***Solution:*** The joint pdf of $X$ and $Y$ is given by

$$f_{(X,Y)}(x,y) = \begin{cases} \dfrac{1}{\pi} & \text{if } \ x^2 + y^2 < 1, \\ \\ 0 & \text{otherwise.} \end{cases}$$

$$\begin{aligned} \Pr(Z \leqslant z) &= \Pr(X^2 + Y^2 \leqslant z^2), \quad \text{for } 0 < z < 1, \\ &= \iint_{X^2+Y^2 \leqslant z^2} f_{(X,Y)}(x,y)\ \mathrm{d}y\ dx \\ &= \frac{1}{\pi}\int_0^{2\pi}\int_0^z r\ \mathrm{d}r\mathrm{d}\theta \\ &= \frac{1}{\pi}(2\pi)\frac{r^2}{2}\Big|_0^z \\ &= z^2, \end{aligned}$$

where we changed to polar coordinates. Thus,

$$F_Z(z) = z^2 \quad \text{for } 0 < z < 1.$$

Consequently,

$$f_Z(z) = \begin{cases} 2z & \text{if } 0 < z < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Thus, computing the expectation,

$$\begin{aligned} E(Z) &= \int_{-\infty}^{\infty} z f_{_Z}(z)\ \mathrm{d}z \\ &= \int_0^1 z(2z)\ \mathrm{d}z \\ &= \int_0^1 2z^2 \mathrm{d}z = \frac{2}{3}. \end{aligned}$$

□

Observe that we could have obtained the answer in the previous example by computing

$$\begin{aligned} E(D) &= E(\sqrt{X^2+Y^2}) \\ &= \iint_{\mathbb{R}^2} \sqrt{x^2+y^2}\ f_{_{(X,Y)}}(x,y)\ \mathrm{d}x\mathrm{d}y \\ &= \iint_A \sqrt{x^2+y^2}\ \frac{1}{\pi}\ \mathrm{d}x\mathrm{d}y, \end{aligned}$$

where $A = \{(x,y) \in \mathbb{R}^2 \mid x^2+y^2 < 1\}$. Thus, using polar coordinates again, we obtain

$$\begin{aligned} E(D) &= \frac{1}{\pi}\int_0^{2\pi}\int_0^1 r\ r\mathrm{d}r\mathrm{d}\theta \\ &= \frac{1}{\pi}\ 2\pi \int_0^1 r^2 \mathrm{d}r \\ &= \frac{2}{3}. \end{aligned}$$

This is, again, the "law of the unconscious statistician;" that is,

$$E[g(X,Y)] = \iint_{\mathbb{R}^2} g(x,y) f_{_{(X,Y)}}(x,y)\ \mathrm{d}x\mathrm{d}y,$$

for any integrable function $g$ of two variables for which

$$\iint_{\mathbb{R}^2} |g(x,y)| f_{_{(X,Y)}}(x,y)\ \mathrm{d}x\mathrm{d}y < \infty.$$

**Theorem 5.2.3.** *Let $X$ and $Y$ denote continuous random variable with joint pdf $f_{_{(X,Y)}}$. Then,*

$$E(X+Y) = E(X) + E(Y).$$

In this theorem,

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)\ \mathrm{d}x,$$

and

$$E(Y) = \int_{-\infty}^{\infty} y f_Y(x)\ \mathrm{d}y,$$

where $f_X$ and $f_Y$ are the marginal distributions of $X$ and $Y$, respectively.

*Proof of Theorem.*

$$\begin{aligned}
E(X+Y) &= \iint_{\mathbb{R}^2} (x+y) f_{(X,Y)}\ \mathrm{d}x\mathrm{d}y \\
&= \iint_{\mathbb{R}^2} x f_{(X,Y)}\ \mathrm{d}x\mathrm{d}y + \iint_{\mathbb{R}^2} y f_{(X,Y)}\ \mathrm{d}x\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{(X,Y)}\ \mathrm{d}y\mathrm{d}x + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f_{(X,Y)}\ \mathrm{d}x\mathrm{d}y \\
&= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{(X,Y)}\ \mathrm{d}y\mathrm{d}x + \int_{-\infty}^{\infty} y \int_{-\infty}^{\infty} f_{(X,Y)}\ \mathrm{d}x\mathrm{d}y \\
&= \int_{-\infty}^{\infty} x f_X(x)\ \mathrm{d}x + \int_{-\infty}^{\infty} y f_Y(y)\ \mathrm{d}y \\
&= E(X) + E(Y).
\end{aligned}$$

□

## 5.3 Independent Random Variables

Two random variables $X$ and $Y$ are said to be independent if for any events $A$ and $B$ in $\mathbb{R}$ (e.g., Borel sets),

$$\Pr((X,Y) \in A \times B) = \Pr(X \in A) \cdot \Pr(Y \in B),$$

where $A \times B$ denotes the Cartesian product of $A$ and $B$:

$$A \times B = \{(x,y) \in \mathbb{R}^2 \mid x \in A \text{ and } x \in B\}.$$

In terms of cumulative distribution functions, this translates into

$$F_{(X,Y)}(x,y) = F_X(x) \cdot F_Y(y) \quad \text{for all } (x,y) \in \mathbb{R}^2. \tag{5.1}$$

We have seen that, in the case in which $X$ and $Y$ are discrete, independence of $X$ and $Y$ is equivalent to

$$p_{(X,Y)}(x,y) = p_X(x) \cdot p_Y(y) \quad \text{for all } (x,y) \in \mathbb{R}^2.$$

For continuous random variables we have the analogous expression in terms of pdfs:

$$f_{(X,Y)}(x,y) = f_X(x) \cdot f_Y(y) \quad \text{for all } (x,y) \in \mathbb{R}^2. \tag{5.2}$$

This follows by taking second partial derivatives on the expression (5.1) for the cdfs since

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

for points $(x,y)$ at which $f_{(X,Y)}$ is continuous, by the Fundamental Theorem of Calculus.

Hence, knowing that $X$ and $Y$ are independent and knowing the corresponding marginal pdfs, in the case in which $X$ and $Y$ are continuous, allows us to compute their joint pdf by using Equation (5.2).

**Example 5.3.1.** A line segment of unit length is divided into three pieces by selecting two points at random and independently and then cutting the segment at the two selected pints. Compute the probability that the three pieces will form a triangle.

***Solution:*** Let $X$ and $Y$ denote the the coordinates of the selected points. We may assume that $X$ and $Y$ are uniformly distributed over $(0,1)$ and independent. We then have that

$$f_X(x) = \begin{cases} 1 & \text{if } 0 < x < 1, \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad f_Y(y) = \begin{cases} 1 & \text{if } 0 < y < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Consequently, the joint distribution of $X$ and $Y$ is

$$f_{(X,Y)} = f_X(x) \cdot f_Y(y) = \begin{cases} 1 & \text{if } 0 < x < 1, 0 < y < 1, \\ 0 & \text{otherwise.} \end{cases}$$

If $X < Y$, the three pieces of length $X$, $Y - X$ and $1 - Y$ will form a triangle if and only if the following three conditions derived from the triangle inequality hold (see Figure 5.3.2):

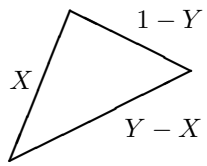



Figure 5.3.2: Random Triangle

$$X \leqslant Y - X + 1 - Y \Rightarrow X \leqslant 1/2,$$

$$Y - X \leqslant X + 1 - Y \Rightarrow Y \leqslant X + 1/2,$$

and

$$1 - Y \leqslant X + Y - X \Rightarrow Y \geqslant 1/2.$$

Similarly, if $X > Y$, a triangle is formed if

$$Y \leqslant 1/2,$$

$$Y \geqslant X - 1/2,$$

and

$$X \geqslant 1/2.$$

Thus, the event that a triangle is formed is the disjoint union of the events

$$A_1 = (X < Y, X \leqslant 1/2, Y \leqslant X + 1/2, Y \geqslant 1/2)$$

and

$$A_2 = (X > Y, Y \leqslant 1/2, Y \geqslant X - 1/2, X \geqslant 1/2).$$

These are pictured in Figure 5.3.3:

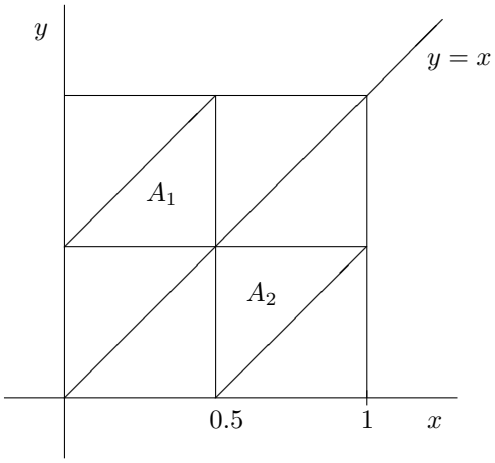



Figure 5.3.3: Sketch of events $A_1$ and $A_2$

We then have that

$$\begin{aligned}
\Pr(\text{triangle}) &= \Pr(A_1 \cup A_2) \\
&= \iint_{A_1 \cup A_2} f_{(X,Y)}(x,y)\ \mathrm{d}x\mathrm{d}y \\
&= \iint_{A_1 \cup A_2} \mathrm{d}x\mathrm{d}y \\
&= \text{area}(A_1) + \text{area}(A_2) \\
&= \frac{1}{4}.
\end{aligned}$$

Thus, there is a 25% chance that the three pieces will form a triangle. □

**Example 5.3.2** (Buffon's Needle Experiment)**.** An experiment consists of dropping a needle of unit length onto a table that has a grid of parallel lines one unit distance apart from one another (see Figure 5.3.4). Compute the probability that the needle will cross one of the lines.

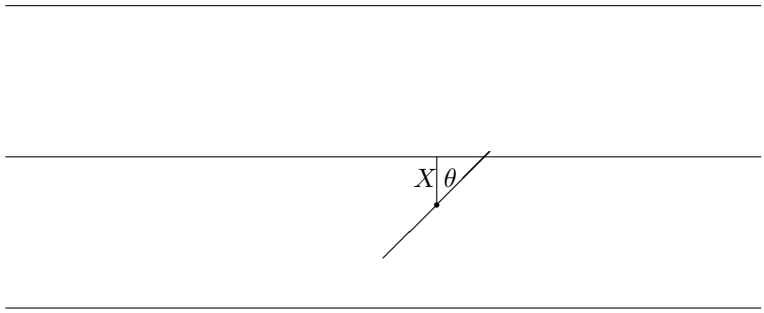



Figure 5.3.4: Buffon's Needle Experiment

***Solution:*** Let $X$ denote the distance from the mid–point of the needle to the closest line (see Figure 5.3.4). We assume that $X$ is uniformly distributed on $(0, 1/2)$; thus, the pdf of $X$ is:

$$f_X(x) = \begin{cases} 2 & \text{if } 0 < x < 1/2, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\Theta$ denote the acute angle that the needle makes with a line perpendicular to the parallel lines. We assume that $\Theta$ and $X$ are

independent and that $\Theta$ is uniformly distributed over $(0, \pi/2)$. Then, the pdf of $\Theta$ is given by

$$f_\Theta(\theta) = \begin{cases} \frac{2}{\pi} & \text{if } 0 < \theta < \pi/2, \\ \\ 0 & \text{otherwise,} \end{cases}$$

and the joint pdf of $X$ and $\Theta$ is

$$f_{(X,\Theta)}(x, \theta) = \begin{cases} \frac{4}{\pi} & \text{if } 0 < x < 1/2,\ 0 < \theta < \pi/2, \\ \\ 0 & \text{otherwise.} \end{cases}$$

When the needle meets a line, as shown in Figure 5.3.4, it makes a right triangle with the line it meets and the segment of length $X$ shown in the Figure. Then, $\dfrac{X}{\cos\theta}$ is the length of the hypothenuse of that triangle. We therefore have that the event that the needle will meet a line is equivalent to the event

$$\frac{X}{\cos\Theta} < \frac{1}{2},$$

or

$$A = \left\{ (x, \theta) \in (0, 2) \times (0, \pi/2) \;\middle|\; X < \frac{1}{2}\cos\theta \right\}.$$

We then have that the probability that the needle meets a line is

$$\begin{aligned} \Pr(A) &= \iint_A f_{(X,\Theta)}(x, \theta)\ \mathrm{d}x\mathrm{d}\theta \\ &= \int_0^{\pi/2} \int_0^{\cos(\theta)/2} \frac{4}{\pi}\ \mathrm{d}x\mathrm{d}\theta \\ &= \frac{2}{\pi} \int_0^{\pi/2} \cos\theta\ \mathrm{d}\theta \\ &= \frac{2}{\pi} \sin\theta \Big|_0^{\pi/2} \\ &= \frac{2}{\pi}. \end{aligned}$$

Thus, there is a $2/\pi$, or about 64%, chance that the needle will meet a line. □

**Example 5.3.3** (Infinite two–dimensional target)**.** Place the center of a target for a darts game at the origin of the $xy$–plane. If a dart lands at a point with

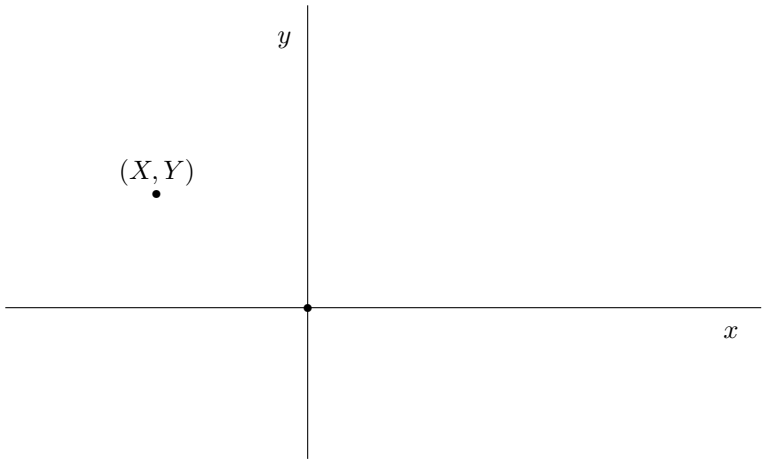



Figure 5.3.5: $xy$–target

coordinates $(X, Y)$, then the random variables $X$ and $Y$ measure the horizontal and vertical miss distance, respectively. For instance, if $X < 0$ and $Y > 0$ , then the dart landed to the left of the center at a distance $|X|$ from a vertical line going through the origin, and at a distance $Y$ above the horizontal line going through the origin (see Figure 5.3.5).

We assume that $X$ and $Y$ are independent, and we are interested in finding the marginal pdfs $f_X$ and $f_Y$ of $X$ and $Y$, respectively.

Assume further that the joint pdf of $X$ and $Y$ is given by the function $f(x, y)$, and that it depends only on the distance from $(X, Y)$ to the origin. More precisely, we suppose that

$$f(x, y) = g(x^2 + y^2) \quad \text{for all } (x, y) \in \mathbb{R}^2,$$

where $g$ is a differentiable function of a single variable. This implies that

$$g(t) \geqslant 0 \quad \text{for all } t \geqslant 0,$$

and

$$\int_0^\infty g(t)\ \mathrm{d}t = \frac{1}{\pi}. \tag{5.3}$$

This follows from the fact that $f$ is a pdf and therefore

$$\iint_{\mathbb{R}^2} f(x, y)\ \mathrm{d}x\mathrm{d}y = 1.$$

Thus, switching to polar coordinates,

$$1 = \int_0^{2\pi} \int_0^\infty g(r^2) r\ \mathrm{d}\mathrm{d}\theta.$$

Hence, after making the change of variables $t = r^2$, we get that

$$1 = 2\pi \int_0^\infty g(r^2)\ r\mathrm{d}r = \pi \int_0^\infty g(t)\ \mathrm{d}t,$$

from which (5.3) follows.

Now, since $X$ and $Y$ are independent, it follows that

$$f(x,y) = f_X(x) \cdot f_Y(y) \quad \text{for all } (x,y) \in \mathbb{R}^2.$$

Thus

$$f_X(x) \cdot f_Y(y) = g(x^2 + y^2).$$

Differentiating with respect to $x$, we get

$$f'_X(x) \cdot f_Y(y) = g'(x^2 + y^2) \cdot 2x.$$

Dividing this equation by $x \cdot f_X(x) \cdot f_Y(y) = x \cdot g(x^2 + y^2)$, we get

$$\frac{1}{x} \cdot \frac{f'_X(x)}{f_X(x)} = \frac{2g'(x^2 + y^2)}{g(x^2 + y^2)} \quad \text{for all } (x,y) \in \mathbb{R}^2$$

with $(x,y) \neq (0,0)$.

Observe that the left–hand side of the equation is a function of $x$, while the right hand side is a function of both $x$ and $y$. It follows that the left-hand-side must be constant. The reason for this is that, by symmetry,

$$2\frac{g'(x^2 + y^2}{g(x^2 + y^2)} = \frac{1}{y}\frac{f'_Y(y)}{f_Y(y)}$$

So that,

$$\frac{1}{x}\frac{f'_X(x)}{f_X(x)} = \frac{1}{y}\frac{f'_Y(y)}{f_Y(y)}$$

for all $(x,y) \in \mathbb{R}^2$, $(x,y) \neq (0,0.$ Thus, in particular

$$\frac{1}{x}\frac{f'_X(x)}{f_X(x)} = \frac{f'_Y(1)}{f_Y(1)} = a, \quad \text{for all } x \in \mathbb{R},$$

and some constant $a$. It then follows that

$$\frac{f'_X(x)}{f_X(x)} = ax$$

for all $x \in \mathbb{R}$. We therefore get that

$$\frac{d}{dx}\big[\ln(f_X(x))\big] = ax.$$

Integrating with respect to $x$, we then have that

$$\ln\big(f_X(x)\big) = a\frac{x^2}{2} + c_1,$$

for some constant of integration $c_1$. Thus, exponentiating on both sides of the equation we get that

$$f_X(x) = ce^{\frac{a}{2}x^2} \quad \text{for all } x \in \mathbb{R}.$$

Thus $f_X(x)$ must be a multiple of $e^{\frac{a}{2}x^2}$. Now, for this to be a *pdf*, we must have that

$$\int_{-\infty}^{\infty} f_X(x)\ \mathrm{d}x = 1.$$

Then, necessarily, it must be the case that $a < 0$ (Why?). Say, $a = -\delta^2$, then

$$f_X(x) = ce^{-\frac{\delta^2}{2}x^2}$$

To determine what the constant $c$ should be, we use the condition

$$\int_{-\infty}^{\infty} f_X(x)\ \mathrm{d}x = 1.$$

Let

$$I = \int_{-\infty}^{\infty} e^{-\delta^2 x^2/2}\ \mathrm{d}x.$$

Observe that

$$I = \int_{-\infty}^{\infty} e^{-\delta^2 y^2/2}\ \mathrm{d}y.$$

We then have that

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty} e^{-\delta^2 \frac{x^2}{2}}\ \mathrm{d}x \cdot \int_{-\infty}^{\infty} e^{-\delta^2 \frac{y^2}{2}}\ \mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{\frac{-\delta^2}{2}(x^2+y^2)}\ \mathrm{d}x\mathrm{d}y. \end{aligned}$$

Switching to polar coordinates we get

$$\begin{aligned} I^2 &= \int_0^{2\pi}\int_0^{\infty} e^{-\frac{\delta^2}{2}r^2}\ r\ \mathrm{d}r\mathrm{d}\theta \\ &= \frac{2\pi}{\delta^2}\int_0^{\infty} e^{-u}\ \mathrm{d}u \\ &= \frac{2\pi}{\delta^2}, \end{aligned}$$

where we have also made the change of variables $u = \dfrac{\delta^2}{2}r^2$. Thus, taking the square root,

$$I = \frac{\sqrt{2\pi}}{\delta}.$$

Hence, the condition

$$\int_{-\infty}^{\infty} f_X(x)dx = 1$$

implies that

$$c\frac{\sqrt{2\pi}}{\delta} = 1$$

from which we get that

$$c = \frac{\delta}{\sqrt{2\pi}}$$

Thus, the pdf for $X$ is

$$f_X(x) = \frac{\delta}{\sqrt{2\pi}} e^{-\frac{\delta^2 x^2}{2}} \quad \text{for all } x \in \mathbb{R}.$$

Similarly, or by using symmetry, we obtain that the pdf for $Y$ is

$$f_Y(y) = \frac{\delta}{\sqrt{2\pi}} e^{-\frac{\delta^2 y^2}{2}} \quad \text{for all } y \in \mathbb{R}.$$

That is, $X$ and $Y$ have the same distribution.

We next set out to compute the expected value and variance of $X$. In order to do this, we first compute the moment generating function, $\psi_X(t)$, of $X$:

$$\begin{aligned}
\psi_X(t) &= E\left(e^{tX}\right) \\
&= \frac{\delta}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{\frac{-\delta^2 x^2}{2}} \, \mathrm{d}x \\
&= \frac{\delta}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{\delta^2 x^2}{2} + tx} \, \mathrm{d}x \\
&= \frac{\delta}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{\delta^2}{2}\left(x^2 - \frac{2tx}{\delta^2}\right)} \, \mathrm{d}x
\end{aligned}$$

Complete the square on $x$ to get

$$x^2 - \frac{2t}{\delta^2}x = x^2 - \frac{2t}{\delta^2}x + \frac{t^2}{\delta^4} - \frac{t^2}{\delta^4} = \left(x - \frac{t}{\delta^2}\right)^2 - \frac{t^2}{\delta^4};$$

then,

$$\begin{aligned}
\psi_X(t) &= \frac{\delta}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{\delta^2}{2}\left(x - \frac{t}{\delta^2}\right)^2} e^{t^2/2\delta^2} \, \mathrm{d}x \\
&= e^{t^2/2\delta^2} \frac{\delta}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{\delta^2}{2}\left(x - \frac{t}{\delta^2}\right)} \, \mathrm{d}x.
\end{aligned}$$

Make the change of variables

$$u = x - \frac{t}{\delta^2}$$

then $\mathrm{d}u = \mathrm{d}x$ and

$$\begin{aligned}
\psi_X(t) &= e^{\frac{t^2}{2\delta^2}} \frac{\delta}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{-\delta^2}{2}u^2} \, \mathrm{d}u \\
&= e^{\frac{t^2}{2\delta^2}},
\end{aligned}$$

since, as we have previously seen in this example,

$$f(u) = \frac{\delta}{\sqrt{2\pi}} e^{\frac{-\delta^2}{2}u^2}, \quad \text{for } u \in \mathbb{R},$$

is a pdf and therefore

$$\int_{-\infty}^{\infty} \frac{\delta}{\sqrt{2\pi}} e^{\frac{-\delta^2}{2}u^2} \, \mathrm{d}u = 1.$$

Hence, the mgf of $X$ is

$$\psi_X(t) = e^{\frac{t^2}{2\delta^2}} \quad \text{for all } t \in \mathbb{R}.$$

Differentiating with respect to $t$ we obtain

$$\psi'_X(t) = \frac{t}{\delta^2} e^{\frac{t^2}{2\delta^2}}$$

and

$$\psi''_X(t) = \frac{1}{\delta^2}\left(1 + \frac{t^2}{\delta^2}\right) e^{t^2/2\delta^2}$$

for all $t \in \mathbb{R}$.

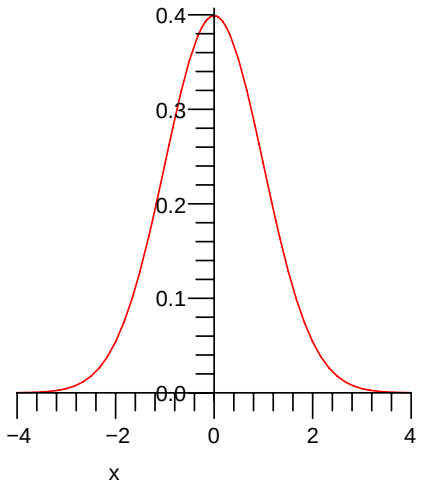



Figure 5.3.6: pdf for $X \sim \text{Normal}(0, 25/8\pi)$

Hence, the mean value of $X$ is

$$E(X) = \psi'_X(0) = 0$$

and the second moment of $X$ is

$$E(X^2) = \psi''_X(0) = \frac{1}{\delta^2}.$$

Thus, the variance of $X$ is

$$\text{var}(X) = \frac{1}{\delta^2}.$$

Next, set $\sigma^2 = \text{var}(X)$. We then have that

$$\sigma = \frac{1}{\delta},$$

and we therefore can write the pdf of $X$ as

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \quad -\infty < x < \infty.$$

A continuous random variable $X$ having the pdf

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}, \quad \text{for } -\infty < x < \infty,$$

is said to have a **normal** distribution with mean 0 and variance $\sigma^2$. We write $X \sim \text{Normal}(0, \sigma^2)$. Similarly, $Y \sim \text{Normal}(0, \sigma^2)$. A graph of the pdf for $X \sim \text{Normal}(0, \sigma^2)$, for $\sigma = 5/\sqrt{8\pi}$, is shown in Figure 5.3.6

# Chapter 6

# Some Special Distributions

In this notes and in the exercises we have encountered several special kinds of random variables and their respective distribution functions in the context of various examples and applications. For example, we have seen the uniform distribution over some interval $(a, b)$ in the context of medeling the selection of a point in $(a, b)$ at random, and the exponential distribution comes up when modeling the service time at a checkout counter. We have discussed Bernoulli trials and have seen that the sum of independent Bernoulli trials gives rise to the Binomial distribution. In the exercises we have seen the discrete uniform distribution, the geometric distribution and the hypergeometric distribution. More recently, in the last example of the previous section, we have seen the the miss horizontal distance from the $y$–axis of dart that lands in an infinite target follows a Normal$(0, \sigma^2)$ distribution (assuming that the miss horizontal and vertical distances are independent and that their joint pdf is radially symmetric). In the next section we discuss the normal distribution in more detail. In subsequent sections we discuss other distributions which come up frequently in application, such as the Poisson distribution.

## 6.1 The Normal Distribution

In Example 5.3.3 we saw that if $X$ denotes the horizontal miss distance from the $y$–axis of a dart that lands on an infinite two–dimensional target, the $X$ has a pdf given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}, \quad \text{for } -\infty < x < \infty,$$

for some constant $\sigma > 0$. In the derivation in Example 5.3.3 we assumed that the $X$ is independent from the $Y$ coordinate (or vertical miss distance) of the point where the dart lands, and that the joint distribution of $X$ and $Y$ depends only on the distance from that point to the origin. We say that $X$ has a normal distribution with mean 0 and variance $\sigma^2$. More generally,we have the following definition:

**Definition 6.1.1** (Normal Distribution)**.** A continuous random variable, $X$, with pdf

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad \text{for } -\infty < x < \infty,$$

and parameters $\mu$ and $\sigma^2$, where $\mu \in \mathbb{R}$ and $\sigma > 0$, is said to have a **normal** distribution with parameters $\mu$ and $\sigma^2$. We write $X \sim \text{Normal}(\mu, \sigma^2)$

The special random variable $Z \sim \text{Normal}(0, 1)$ has a distribution function known as the **standard normal distribution**:

$$f_Z(z) = \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2} \quad \text{for } -\infty < z < \infty. \tag{6.1}$$

From the calculations in Example 5.3.3 we get that the moment generating function for $Z$ is

$$\psi_Z(t) = e^{t^2/2} \quad \text{for } -\infty < t < \infty. \tag{6.2}$$

**Example 6.1.2.** Let $\mu \in \mathbb{R}$ and $\sigma > 0$ be given, and define

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

Show that $X \sim \text{Normal}(\mu, \sigma^2)$ and compute the moment generating function of $X$, its expectation and its variance.

***Solution:*** First we find the pdf of $X$ and show that it is the same as that given in Definition 6.1.1. In order to do this, we first compute the cdf of $X$:

$$\begin{aligned} F_X(x) &= \Pr(X \leqslant x) \\ &= \Pr(\sigma Z + \mu \leqslant x) \\ &= \Pr[Z \leqslant (x-\mu)/\sigma] \\ &= F_Z((x-\mu)/\sigma). \end{aligned}$$

Differentiating with respect to $x$, while remembering to use the Chain Rule, we get that

$$\begin{aligned} f_X(x) &= F_Z'((x-\mu)/\sigma) \cdot (1/\sigma) \\ &= \frac{1}{\sigma} f_Z\left(\frac{x-\mu}{\sigma}\right). \end{aligned}$$

Thus, using (6.1), we get

$$\begin{aligned} f_X(x) &= \frac{1}{\sigma} \cdot \frac{1}{\sqrt{2\pi}} e^{-[(x-\mu)/\sigma]^2/2} \\ &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \end{aligned}$$

for $-\infty < x < \infty$, which is the pdf for a Normal$(\mu, \sigma^2)$ random variable according to Definition 6.1.1.

Next, we compute the mgf of $X$:

$$\begin{aligned} \psi_X(t) &= E(e^{tX}) \\ &= E(e^{t(\sigma Z+\mu)}) \\ &= E(e^{t\sigma Z+t\mu}) \\ &= E(e^{\sigma tZ}e^{\mu t}) \\ &= e^{\mu t}E(e^{(\sigma t)Z}) \\ &= e^{\mu t}\psi_Z(\sigma t), \end{aligned}$$

for $-\infty < t < \infty$. It then follows from (6.2) that

$$\begin{aligned} \psi_X(t) &= e^{\mu t}e^{(\sigma t)^2/2} \\ &= e^{\mu t}e^{\sigma^2 t^2/2} \\ &= e^{\mu t+\sigma^2 t^2/2}, \end{aligned}$$

for $-\infty < t < \infty$.

Differentiating with respect to $t$, we obtain

$$\psi'_X(t) = (\mu + \sigma^2 t)e^{\mu t+\sigma^2 t^2/2}$$

and

$$\psi''_X(t) = \sigma^2 e^{\mu t+\sigma^2 t^2/2} + (\mu + \sigma^2 t)^2 e^{\mu t+\sigma^2 t^2/2}$$

for $-\infty < t < \infty$.

Consequently, the expected value of $X$ is

$$E(X) = \psi'_X(0) = \mu,$$

and the second moment of $X$ is

$$E(X^2) = \psi''_X(0) = \sigma^2 + \mu^2.$$

Thus, the variance of $X$ is

$$\mathrm{var}(X) = E(X^2) - \mu^2 = \sigma^2.$$

We have therefore shown that if $X \sim$ Normal$(\mu, \sigma^2)$, then the parameter $\mu$ is the mean of $X$ and $\sigma^2$ is the variance of $X$. □

**Example 6.1.3.** Suppose that many observations from a Normal$(\mu, \sigma^2)$ distribution are made. Estimate the proportion of observations that must lie within one standard deviation from the mean.

***Solution:*** We are interested in estimating $\Pr(|X - \mu| < \sigma)$, where $X \sim$ Normal$(\mu, \sigma^2)$.

Compute

$$\begin{aligned}\Pr(|X-\mu|<\sigma) &= \Pr(-\sigma<X-\mu<\sigma)\\ &= \Pr\left(-1<\frac{X-\mu}{\sigma}<1\right),\end{aligned}$$

where, according to Problem (1) in Assignment #16,

$$\frac{X-\mu}{\sigma}\sim \text{Normal}(0,1);$$

that is, it has a standard normal distribution. Thus, setting

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

we get that

$$\Pr(|X-\mu|<\sigma)=\Pr(-1<Z<1),$$

where $Z\sim\text{Normal}(0,1)$. Observe that

$$\begin{aligned}\Pr(|X-\mu|<\sigma) &= \Pr(-1<Z<1)\\ &= \Pr(-1<Z\leqslant 1)\\ &= F_Z(1)-F_Z(-1),\end{aligned}$$

where $F_Z$ is the cdf of $Z$. MS Excel has a built in function called `normdist` which returns the cumulative distribution function value at $x$, from a Normal$(\mu,\sigma^2)$ random variable, $X$, given $x$, the mean $\mu$, the standard deviation $\sigma$ and a "TRUE" tag as follows:

$$F_Z(z)=\texttt{normdist}(z,\mu,\sigma,\text{TRUE})=\texttt{normdist}(z,0,1,\text{TRUE}).$$

We therefore obtain the approximation

$$\Pr(|X-\mu|<\sigma)\approx 0.8413-0.1587=0.6826,\quad \text{or about } 68\%.$$

Thus, around 68% of observations from a normal distribution should lie within one standard deviation from the mean. □

**Example 6.1.4** (The Chi–Square Distribution with one degree of freedom)**.** Let $Z\sim\text{Normal}(0,1)$ and define $Y=Z^2$. Give the pdf of $Y$.

***Solution:*** The pdf of $Z$ is given by

$$f_Z(z)=\frac{1}{\sqrt{2\pi}}e^{-z^2/2},\quad \text{for } -\infty<z<\infty.$$

We compute the pdf for $Y$ by first determining its cdf:

$$\begin{aligned}P(Y\leq y) &= P(Z^2\leq y)\quad \text{for } y\geqslant 0\\ &= P(-\sqrt{y}\leq Z\leq\sqrt{y})\\ &= P(-\sqrt{y}<Z\leq\sqrt{y}),\quad \text{since Z is continuous.}\end{aligned}$$

Thus,

$$\begin{aligned} P(Y \leq y) &= P(Z \leq \sqrt{y}) - P(Z \leq -\sqrt{y}) \\ &= F_Z(\sqrt{y}) - F_Z(-\sqrt{y}) \quad \text{for } y > 0, \end{aligned}$$

since $Y$ is continuous.

We then have that the cdf of $Y$ is

$$F_Y(y) = F_Z(\sqrt{y}) - F_Z(-\sqrt{y}) \quad \text{for } y > 0,$$

from which we get, after differentiation with respect to $y$,

$$\begin{aligned} f_Y(y) &= F_Z'(\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} + F_Z'(-\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} \\ &= f_Z(\sqrt{y}) \frac{1}{2\sqrt{y}} + f_Z(-\sqrt{y}) \frac{1}{2\sqrt{y}} \\ &= \frac{1}{2\sqrt{y}} \left\{ \frac{1}{\sqrt{2\pi}}\, e^{-y/2} + \frac{1}{\sqrt{2\pi}}\, e^{-y/2} \right\} \\ &= \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{y}}\, e^{-y/2} \end{aligned}$$

for $y > 0$. □

**Definition 6.1.5.** A continuous random variable $Y$ having the pdf

$$f_Y(y) = \begin{cases} \dfrac{1}{\sqrt{2\pi}} \cdot \dfrac{1}{\sqrt{y}}\, e^{-y/2} & \text{if } y > 0 \\ \\ 0 & \text{otherwise,} \end{cases}$$

is said to have a Chi–Square distribution with one degree of freedom. We write

$$Y \sim \chi_1^2.$$

**Remark 6.1.6.** Observe that if $Y \sim \chi_1^2$, then its expected value is

$$E(Y) = E(Z^2) = 1.$$

To compute the second moment of $Y$, $E(Y^2) = E(Z^4)$, we need to compute the fourth moment of $Z$. Recall that the mgf of $Z$ is

$$\psi_Z(t) = e^{t^2/2} \quad \text{for all } t \in \mathbb{R}.$$

Its fourth derivative can be computed to be

$$\psi_Z^{(4)}(t) = (3 + 6t^2 + t^4)\, e^{t^2/2} \quad \text{for all } t \in \mathbb{R}.$$

Thus,

$$E(Z^4) = \psi_Z^{(4)}(0) = 3.$$

We then have that the variance of $Y$ is

$$\text{var}(Y) = E(Y^2) - 1 = E(Z^4) - 1 = 3 - 1 = 2.$$

## 6.2 The Poisson Distribution

**Example 6.2.1** (Bacterial Mutations)**.** Consider a colony of bacteria in a culture consisting of $N$ bacteria. This number, $N$, is typically very large (in the order of $10^6$ bacteria). We consider the question of how many of those bacteria will develop a specific mutation in their genetic material during one division cycle (e.g., a mutation that leads to resistance to certain antibiotic or virus). We assume that each bacterium has a very small, but positive, probability, $a$, of developing that mutation when it divides. This probability is usually referred to as the **mutation rate** for the particular mutation and organism. If we let $X_N$ denote the number of bacteria, out of the $N$ in the colony, that develop the mutation in a division cycle, we can model $X_N$ by a binomial random variable with parameters $a$ and $N$; that is,

$$X_N \sim \text{Binomial}(a, N).$$

This assertion is justified by assuming that the event that a given bacterium develops a mutation is independent of any other bacterium in the colony developing a mutation. We then have that

$$\Pr(X_N = k) = \binom{N}{k} a^k (1-a)^{N-k}, \quad \text{for } k = 0, 1, 2, \ldots, N. \tag{6.3}$$

Also,

$$E(X_N) = aN.$$

Thus, the average number of mutations in a division cycle is $aN$. We will denote this number by $\lambda$ and will assume that it remains constant; that is,

$$aN = \lambda \text{ (a constant.)}$$

Since $N$ is very large, it is reasonable to make the approximation

$$\Pr(X_N = k) \approx \lim_{N\to\infty} \binom{N}{k} a^k (1-a)^{N-k}, \quad \text{for } \ k = 0, 1, 2, 3, \ldots$$

provided that the limit on the right side of the expression exists. In this example we show that the limit in the expression above exists if $aN = \lambda$ is kept constant as $N$ tends to infinity. To see why this is the case, we first substitute $\lambda/N$ for $a$ in the expression for $\Pr(X_N = k)$ in equation (6.3). We then get that

$$\Pr(X_N = k) = \binom{N}{k} \left(\frac{\lambda}{N}\right)^k \left(1 - \frac{\lambda}{N}\right)^{N-k},$$

which can be re-written as

$$\Pr(X_N = k) = \frac{\lambda^k}{k!} \cdot \frac{N!}{(N-k)!} \cdot \frac{1}{N^k} \cdot \left(1 - \frac{\lambda}{N}\right)^N \cdot \left(1 - \frac{\lambda}{N}\right)^{-k}. \tag{6.4}$$

Observe that

$$\lim_{N\to\infty}\left(1-\frac{\lambda}{N}\right)^N = e^{-\lambda}, \tag{6.5}$$

since $\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n = e^x$ for all real numbers $x$. Also note that, since $k$ is fixed,

$$\lim_{N\to\infty}\left(1-\frac{\lambda}{N}\right)^{-k} = 1. \tag{6.6}$$

It remains to see then what happens to the term $\frac{N!}{(N-k)!}\cdot\frac{1}{N^k}$, in equation (6.4), as $N\to\infty$. To answer this question, we compute

$$\begin{aligned}\frac{N!}{(N-k)!}\cdot\frac{1}{N^k} &= \frac{N(N-1)(N-2)\cdots(N-(k-1))}{N^k}\\ &= \left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\cdots\left(1-\frac{k-1}{N}\right).\end{aligned}$$

Thus, since

$$\lim_{N\to\infty}\left(1-\frac{j}{N}\right) = 1 \quad \text{for all } j = 1,2,3,\ldots,k-1,$$

it follows that

$$\lim_{N\to\infty}\frac{N!}{(N-k)!}\cdot\frac{1}{N^k} = 1. \tag{6.7}$$

Hence, in view of equations (6.5), (6.6) and (6.7), we see from equation (6.4) that

$$\lim_{N\to\infty}\Pr(X_N = k) = \frac{\lambda^k}{k!}\,e^{-\lambda} \quad \text{for } k = 0,1,2,3,\ldots. \tag{6.8}$$

The limiting distribution obtained in equation (6.8) is known as the **Poisson Distribution** with parameter $\lambda$. We have therefore shown that the number of mutations occurring in a bacterial colony of size $N$, per division cycle, can be approximated by a Poisson random variable with parameter $\lambda = aN$, where $a$ is the mutation rate.

**Definition 6.2.2** (Poisson Distribution)**.** A discrete random variable, $X$, with pmf

$$p_X(k) = \frac{\lambda^k}{k!}\,e^{-\lambda} \quad \text{for } k = 0,1,2,3,\ldots;\ \text{zero elsewhere,}$$

where $\lambda > 0$, is said to have a **Poisson** distribution with parameter $\lambda$. We write $X \sim \text{Poisson}(\lambda)$.

To see that the expression for $p_X(k)$ in Definition 6.2.2 indeed defines a pmf, observe that

$$\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = e^{\lambda},$$

from which we get that

$$\sum_{k=0}^{\infty} p_X(k) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}\, e^{-\lambda} = e^{\lambda} \cdot e^{-\lambda} = 1.$$

In a similar way we can compute the mgf of $X \sim \text{Poisson}(\lambda)$ to obtain (see Problem 1 in Assignment #17):

$$\psi_X(t) = e^{\lambda(e^t - 1)} \quad \text{for all } t \in \mathbb{R}.$$

Using this mgf we can derive that the expected value of $X \sim \text{Poisson}(\lambda)$ is

$$E(X) = \lambda,$$

and its variance is

$$\text{var}(X) = \lambda$$

as well.

**Example 6.2.3** (Sum of Independent Poisson Random Variables)**.** Suppose that $X \sim \text{Poisson}(\lambda_1)$ and $Y \sim \text{Poisson}(\lambda_2)$ are independent random variables, where $\lambda_1$ and $\lambda_2$ are positive. Define $Z = X + Y$. Give the distribution of $Z$.

***Solution:*** Use the independence of $X$ and $Y$ to compute the mgf of $Z$:

$$\begin{aligned} \psi_Z(t) &= \psi_{X+Y}(t) \\ &= \psi_X(t) \cdot \psi_Y(t) \\ &= e^{\lambda_1(e^t-1)} \cdot e^{\lambda_2(e^t-1)} \\ &= e^{(\lambda_1+\lambda_2)(e^t-1)}, \end{aligned}$$

which is the mgf of a Poisson$(\lambda_1 + \lambda_2)$ distribution. It then follows that

$$Z \sim \text{Poisson}(\lambda_1 + \lambda_2)$$

and therefore

$$p_Z(k) = \frac{(\lambda_1 + \lambda_2)^k}{k!}\, e^{-\lambda_1 - \lambda_2} \quad \text{for } k = 0, 1, 2, 3, \ldots; \text{ zero elsewhere.}$$

□

**Example 6.2.4** (Estimating Mutation Rates in Bacterial Populations)**.** Luria and Delbrück[1] devised the following procedure (known as the *fluctuation test*) to estimate the *mutation rate, a*, for certain bacteria:
Imagine that you start with a single normal bacterium (with no mutations) and allow it to grow to produce several bacteria. Place each of these bacteria in test–tubes each with media conducive to growth. Suppose the bacteria in the

[1](1943) *Mutations of bacteria from virus sensitivity to virus resistance. Genetics,* **28**, 491–511

test–tubes are allowed to reproduce for $n$ division cycles. After the $n^{\text{th}}$ division cycle, the content of each test–tube is placed onto a agar plate containing a virus population which is lethal to the bacteria which have not developed resistance. Those bacteria which have mutated into resistant strains will continue to replicate, while those that are sensitive to the virus will die. After certain time, the resistant bacteria will develop visible colonies on the plates. The number of these colonies will then correspond to the number of resistant cells in each test tube at the time they were exposed to the virus. This number corresponds to the number of bacteria in the colony that developed a mutation which led to resistance. We denote this number by $X_N$, as we did in Example 6.2.1, where $N$ is the size of the colony after the $n^{\text{th}}$ division cycle. Assuming that the bacteria may develop mutation to resistance after exposure to the virus, the argument in Example 6.2.1 shows that, if $N$ is very large, the distribution of $X_N$ can be approximated by a Poisson distribution with parameter $\lambda = aN$, where $a$ is the mutation rate and $N$ is the size of the colony. It then follows that the probability of no mutations occurring in one division cycle is

$$\Pr(X_N = 0) \approx e^{-\lambda}. \tag{6.9}$$

This probability can also be estimated experimentally as Luria and nd Delbrück showed in their 1943 paper. In one of the experiments described in that paper, out of 87 cultures of $2.4\times10^8$ bacteria, 29 showed not resistant bacteria (i.e., none of the bacteria in the culture mutated to resistance and therefore all perished after exposure to the virus). We therefore have that

$$\Pr(X_N = 0) \approx \frac{29}{87}.$$

Comparing this to the expression in Equation (6.9), we obtain that

$$e^{-\lambda} \approx \frac{29}{87},$$

which can be solved for $\lambda$ to obtain

$$\lambda \approx -\ln\left(\frac{29}{87}\right)$$

or

$$\lambda \approx 1.12.$$

The mutation rate, $a$, can then be estimated from $\lambda = aN$:

$$a = \frac{\lambda}{N} \approx \frac{1.12}{2.4 \times 10^8} \approx 4.7 \times 10^{-9}.$$

# Chapter 7

# Convergence in Distribution

We have seen in Example 6.2.1 that if $X_n \sim \text{Binomial}(\lambda/n, n)$, for some constant $\lambda > 0$ and $n = 1, 2, 3, \ldots$, and if $Y \sim \text{Poisson}(\lambda)$, then

$$\lim_{n\to\infty} \Pr(X_n = k) = \Pr(Y = k) \quad \text{for all } k = 0, 1, 2, 3, \ldots$$

We then say that the sequence of Binomial random variables $(X_1, X_2, X_3, \ldots)$ **converges in distribution** to the Poisson random variable $Y$ with parameter $\lambda$. This concept will be made more general and precise in the following section.

## 7.1 Definition of Convergence in Distribution

**Definition 7.1.1** (Convergence in Distribution)**.** Let $(X_n)$ be a sequence of random variables with cumulative distribution functions $F_{X_n}$, for $n = 1, 2, 3, \ldots$, and $Y$ be a random variable with cdf $F_Y$. We say that the sequence $(X_n)$ converges to $Y$ in distribution, if

$$\lim_{n\to\infty} F_{X_n}(x) = F_Y(x)$$

for all $x$ where $F_Y$ is continuous.

We write

$$X_n \xrightarrow{D} Y \quad \text{as } n \to \infty.$$

Thus, if $(X_n)$ converges to $Y$ in distribution then

$$\lim_{n\to\infty} \Pr(X_n \leqslant x) = \Pr(Y \leqslant x) \quad \text{for } x \text{ at which } F_Y \text{ is continuous.}$$

If $Y$ is discrete, for instance taking on nonzero values at $k = 0, 1, 2, 3, \ldots$ (as is the case with the Poisson distribution), then $F_Y$ is continuous in between consecutive integers $k$ and $k+1$. We therefore get that for any $\varepsilon$ with $0 < \varepsilon < 1$,

$$\lim_{n\to\infty} \Pr(X_n \leqslant k + \varepsilon) = \Pr(Y \leqslant k + \varepsilon)$$

for all $k = 0, 1, 2, 3, \ldots$. Similarly,

$$\lim_{n\to\infty} \Pr(X_n \leqslant k - \varepsilon) = \Pr(Y \leqslant k - \varepsilon)$$

for all $k = 0, 1, 2, 3, \ldots$. We therefore get that

$$\lim_{n\to\infty} \Pr(k - \varepsilon < X_n \leqslant k + \varepsilon) = \Pr(k - \varepsilon < Y \leqslant k + \varepsilon)$$

for all $k = 0, 1, 2, 3, \ldots$. From this we get that

$$\lim_{n\to\infty} \Pr(X_n = k) = \Pr(Y = k)$$

for all $k = 0, 1, 2, 3, \ldots$, in the discrete case.

If $(X_n)$ converges to $Y$ in distribution and the moment generating functions $\psi_{X_n}(t)$ for $n = 1, 2, 3 \ldots$, and $\psi_Y(t)$ all exist on some common interval of values of $t$, then it might be the case, under certain technical conditions which may be found in a paper by Kozakiewicz,[1] that

$$\lim_{n\to\infty} \psi_{X_n}(t) = \psi_Y(t)$$

for all $t$ in the common interval of existence.

**Example 7.1.2.** Let $X_n \sim \text{Binomial}(\lambda/n, n)$ for $n = 1, 2, 3, \ldots$.

$$\psi_{X_n}(t) = \left(\frac{\lambda e^t}{n} + 1 - \frac{\lambda}{n}\right)^n \quad \text{for all } n \text{ and all } t.$$

Then,

$$\lim_{n\to\infty} \psi_{X_n}(t) = \lim_{n\to\infty} \left(1 + \frac{\lambda(e^t - 1)}{n}\right)^n = e^{\lambda(e^t - 1)},$$

which is the moment generating function of a Poisson($\lambda$) distribution. This is not surprising since we already know that $X_n$ converges to $Y \sim \text{Poisson}(\lambda)$ in distribution. What is surprising is the theorem discussed in the next section known as the **mgf Convergence Theorem**.

## 7.2 mgf Convergence Theorem

**Theorem 7.2.1** (mgf Convergence Theorem, Theorem 5.7.4 on page 289 in the text)**.** *Let $(X_n)$ be a sequence of random variables with moment generating functions $\psi_{X_n}(t)$ for $|t| < h$, $n = 1, 2, 3, \ldots$, and some positive number $h$. Suppose $Y$ has mgf $\psi_Y(t)$ which exists for $|t| < h$. Then, if*

$$\lim_{n\to\infty} \psi_{X_n}(t) = \psi_Y(t), \quad \textit{for } |t| < h,$$

*it follows that*

$$\lim_{n\to\infty} F_{X_n}(x) = F_Y(x)$$

*for all $x$ where $F_Y$ is continuous.*

[1](1947) *On the Convergence of Sequences of Moment Generating Functions.* Annals of Mathematical Statistics, Volume **28**, Number 1, pp. 61–69

**Notation.** If $(X_n)$ converges to $Y$ in distribution as $n$ tends to infinity, we write

$$X_n \xrightarrow{D} Y \quad \text{as} \quad n \to \infty.$$

The mgf Convergence Theorem then says that if the moment generating functions of a sequence of random variables, $(X_n)$, converges to the mgf of a ranfom variable $Y$ on some interval around 0, the $X_n$ converges to $Y$ in distribution as $n \to \infty$.

The mgf Theorem is usually ascribed to Lévy and Cramér (c.f. the paper by Curtiss[2])

**Example 7.2.2.** Let $X_1, X_2, X_3, \ldots$ denote independent, Poisson($\lambda$) random variables. For each $n = 1, 2, 3, \ldots$, define the random variable

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

$\overline{X}_n$ is called the **sample mean** of the **random sample of size** $n$, $\{X_1, X_2, \ldots, X_n\}$.

The expected value of the sample mean, $\overline{X}_n$, is obtained from

$$\begin{aligned} E(\overline{X}_n) &= E\left(\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right) \\ &= \frac{1}{n}\sum_{k=1}^{n} E(X_k) \\ &= \frac{1}{n}\sum_{k=1}^{n} \lambda \\ &= \frac{1}{n}(n\lambda) \\ &= \lambda. \end{aligned}$$

[2](1942) *A note on the Theory of Moment Generating Functions.* Annals of Mathematical Statistics, Volume **13**, Number 4, pp. 430–433

Since the $X_i$'s are independent, the variance of $\overline{X}_n$ can be computed as follows

$$\begin{aligned}
\text{var}(\overline{X}_n) &= \text{var}\left(\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right) \\
&= \frac{1}{n^2}\sum_{k=1}^{n}\text{var}(X_k) \\
&= \frac{1}{n^2}\sum_{k=1}^{n}\lambda \\
&= \frac{1}{n^2}(n\lambda) \\
&= \frac{\lambda}{n}.
\end{aligned}$$

We can also compute the mgf of $\overline{X}_n$ as follows:

$$\begin{aligned}
\psi_{\overline{X}_n}(t) &= E(e^{t\overline{X}_n}) \\
&= \psi_{X_1+X_2+\cdots+X_n}\left(\frac{t}{n}\right) \\
&= \psi_{X_1}\left(\frac{t}{n}\right)\psi_{X_2}\left(\frac{t}{n}\right)\cdots\psi_{X_n}\left(\frac{t}{n}\right),
\end{aligned}$$

since the $X_i$'s are linearly independent. Thus, given the $X_i$'s are identically distributed Poisson($\lambda$),

$$\begin{aligned}
\psi_{\overline{X}_n}(t) &= \left[\psi_{X_1}\left(\frac{t}{n}\right)\right]^n \\
&= \left[e^{\lambda(e^{t/n}-1)}\right]^n \\
&= e^{\lambda n(e^{t/n}-1)},
\end{aligned}$$

for all $t \in \mathbb{R}$.

Next we define the random variables

$$Z_n = \frac{\overline{X}_n - E(\overline{X}_n)}{\sqrt{\text{var}(\overline{X}_n)}} = \frac{\overline{X}_n - \lambda}{\sqrt{\lambda/n}}, \quad \text{for } n = 1, 2, 3, \ldots$$

We would like to see if the sequence of random variables $(Z_n)$ converges in distribution to a limiting random variable. To answer this question, we apply

the mgf Convergence Theorem (Theorem 7.2.1). Thus, we compute the mgf of $Z_n$ for $n = 1, 2, 3, \ldots$

$$\begin{aligned}
\psi_{Z_n}(t) &= E(e^{tZ_n}) \\
&= E\left(e^{\frac{t\sqrt{n}}{\sqrt{\lambda}}\overline{X}_n - t\sqrt{\lambda n}}\right) \\
&= e^{-t\sqrt{\lambda n}}\, E\left(e^{\frac{t\sqrt{n}}{\sqrt{\lambda}}\overline{X}_n}\right) \\
&= e^{-t\sqrt{\lambda n}}\, \psi_{\overline{X}_n}\left(\frac{t\sqrt{n}}{\sqrt{\lambda}}\right).
\end{aligned}$$

Next, use the fact that

$$\psi_{\overline{X}_n}(t) = e^{\lambda n(e^{t/n}-1)},$$

for all $t \in \mathbb{R}$, to get that

$$\begin{aligned}
\psi_{Z_n}(t) &= e^{-t\sqrt{\lambda n}}\, e^{\lambda n(e^{(t\sqrt{n}/\sqrt{\lambda})/n}-1)} \\
&= e^{-t\sqrt{\lambda n}}\, e^{\lambda n(e^{t/\sqrt{\lambda n}}-1)}.
\end{aligned}$$

Now, using the fact that

$$e^x = \sum_{k=0}^{n} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots,$$

we obtain that

$$e^{t/\sqrt{n\lambda}} = 1 + \frac{t}{\sqrt{n\lambda}} + \frac{1}{2}\,\frac{t^2}{n\lambda} + \frac{1}{3!}\,\frac{t^3}{n\lambda\sqrt{n\lambda}} + \frac{1}{4!}\,\frac{t^4}{(n\lambda)^2} + \cdots$$

so that

$$e^{t/\sqrt{n\lambda}} - 1 = \frac{t}{\sqrt{n\lambda}} + \frac{1}{2}\,\frac{t^2}{n\lambda} + \frac{1}{3!}\,\frac{t^3}{n\lambda\sqrt{n\lambda}} + \frac{1}{4!}\,\frac{t^4}{(n\lambda)^2} + \cdots$$

and

$$\lambda n(e^{t/\sqrt{n\lambda}} - 1) = \sqrt{n\lambda}t + \frac{1}{2}t^2 + \frac{1}{3!}\,\frac{t^3}{n\lambda} + \frac{1}{4!}\,\frac{t^4}{n\lambda} + \cdots$$

Consequently,

$$e^{\lambda n(e^{t/\sqrt{n\lambda}}-1)} = e^{\sqrt{n\lambda}t} \cdot e^{t^2/2} \cdot e^{\frac{1}{3!}\,\frac{t^3}{n\lambda} + \frac{1}{4!}\,\frac{t^4}{n\lambda} + \cdots}$$

and

$$e^{-\sqrt{n\lambda}t} e^{\lambda n(e^{t/\sqrt{n\lambda}}-1)} = e^{t^2/2} \cdot e^{\frac{1}{3!}\,\frac{t^3}{n\lambda} + \frac{1}{4!}\,\frac{t^4}{n\lambda} + \cdots}.$$

Observe that the exponent in the last exponential tends to 0 as $n \to \infty$. We therefore get that

$$\lim_{n\to\infty}\left[e^{-\sqrt{n\lambda}t}e^{\lambda n(e^{t/\sqrt{n\lambda}}-1)}\right] = e^{t^2/2}\cdot e^0 = e^{t^2/2}.$$

We have thus shown that

$$\lim_{n\to\infty}\psi_{Z_n}(t) = e^{t^2/2}, \quad \text{for all } t \in \mathbb{R},$$

where the right–hand side is the mgf of $Z \sim \text{Normal}(0,1)$. Hence, by the mgf Convergence Theorem, $Z_n$ converges in distribution to a Normal$(0,1)$ random variable.

**Example 7.2.3** (Estimating Proportions)**.** Suppose we sample from a given voting population by surveying a random group of $n$ people from that population and asking them whether or not they support certain candidate. In the population (which could consists of hundreds of thousands of voters) there is a proportion, $p$, of people who support the candidate, and which is not known with certainty. If we denote the responses of the surveyed $n$ people by $X_1, X_2, \ldots, X_n$, then we can model these responses as independent Bernoulli$(p)$ random variables: a responses of "yes" corresponds to $X_i = 1$, while $X_i = 0$ is a "no." The sample mean,

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n},$$

give the proportion of individuals from the sample who support the candidate. Since $E(X_i) = p$ for each $i$, the expected value of the sample mean is

$$E(\overline{X}_n) = p.$$

Thus, $\overline{X}_n$ can be used as an estimate for the true proportion, $p$, of people who support the candidate. How good can this estimate be? In this example we try to answer this question.

By independence of the $X_i$'s, we get that

$$\text{var}(\overline{X}_n) = \frac{1}{n}\text{var}(X_1) = \frac{p(1-p)}{n}.$$

Also, the mgf of $\overline{X}_n$ is given by

$$\begin{aligned}
\psi_{\overline{X}_n}(t) &= E(e^{t\overline{X}_n}) \\
&= \psi_{X_1+X_2+\cdots+X_n}\left(\frac{t}{n}\right) \\
&= \psi_{X_1}\left(\frac{t}{n}\right)\psi_{X_2}\left(\frac{t}{n}\right)\cdots\psi_{X_n}\left(\frac{t}{n}\right) \\
&= \left[\psi_{X_1}\left(\frac{t}{n}\right)\right]^n \\
&= \left(p\ e^{t/n}+1-p\right)^n.
\end{aligned}$$

As in the previous example, we define the random variables

$$Z_n = \frac{\overline{X}_n - E(\overline{X}_n)}{\sqrt{\text{var}(\overline{X}_n)}} = \frac{\overline{X}_n - p}{\sqrt{p(1-p)/n}}, \quad \text{for } n = 1, 2, 3, \ldots,$$

or

$$Z_n = \frac{\sqrt{n}}{\sqrt{p(1-p)}}\overline{X}_n - \sqrt{\frac{np}{1-p}}, \quad \text{for } n = 1, 2, 3, \ldots.$$

Thus, the mgf of $Z_n$ is

$$\begin{aligned}
\psi_{Z_n}(t) &= E(e^{tZ_n}) \\
&= E\left(e^{\frac{t\sqrt{n}}{\sqrt{p(1-p)}}\overline{X}_n - t\sqrt{\frac{np}{1-p}}}\right) \\
&= e^{-t\sqrt{\frac{np}{1-p}}}\ E\left(e^{\frac{t\sqrt{n}}{\sqrt{p(1-p)}}\overline{X}_n}\right) \\
&= e^{-t\sqrt{\frac{np}{1-p}}}\ \psi_{\overline{X}_n}\left(\frac{t\sqrt{n}}{\sqrt{p(1-p)}}\right) \\
&= e^{-t\sqrt{\frac{np}{1-p}}}\ \left(p\ e^{(t\sqrt{n}/\sqrt{p(1-p)}/n}+1-p\right)^n \\
&= e^{-t\sqrt{\frac{np}{1-p}}}\ \left(p\ e^{(t/\sqrt{np(1-p)}}+1-p\right)^n \\
&= \left[e^{-tp\sqrt{\frac{1}{np(1-p)}}}\left(p\ e^{(t/\sqrt{np(1-p)})}+1-p\right)\right]^n \\
&= \left(p\ e^{t(1-p)/\sqrt{np(1-p)})}+(1-p)e^{tp/\sqrt{np(1-p)})}\right)^n.
\end{aligned}$$

To compute the limit of $\psi_{Z_n}(t)$ as $n \to \infty$, we first compute the limit of $\ln\left(\psi_{Z_n}(t)\right)$ as $n \to \infty$, where

$$\begin{aligned}\ln\left(\psi_{Z_n}(t)\right) &= n \ln\left(p\ e^{t(1-p)/\sqrt{np(1-p)})} + (1-p)e^{-tp/\sqrt{np(1-p)})}\right) \\ &= n \ln\left(p\ e^{a/\sqrt{n}} + (1-p)\ e^{-b/\sqrt{n}}\right),\end{aligned}$$

where we have set

$$a = \frac{(1-p)t}{\sqrt{p(1-p)}} \quad \text{and} \quad b = \frac{p\ t}{\sqrt{p(1-p)}}.$$

Observe that

$$pa - (1-p)b = 0, \tag{7.1}$$

and

$$p\ a^2 + (1-p)b^2 = (1-p)t^2 + p\ t^2 = t^2. \tag{7.2}$$

Writing

$$\ln\left(\psi_{Z_n}(t)\right) = \frac{\ln\left(p\ e^{a/\sqrt{n}} + (1-p)\ e^{-b/\sqrt{n}}\right)}{\frac{1}{n}},$$

we see that L'Hospital's Rule can be applied to compute the limit of $\ln\left(\psi_{Z_n}(t)\right)$ as $n \to \infty$. We therefore get that

$$\begin{aligned}\lim_{n\to\infty} \ln\left(\psi_{Z_n}(t)\right) &= \lim_{n\to\infty} \frac{\frac{-\frac{a}{2}\ \frac{1}{n\sqrt{n}} p\ e^{a/\sqrt{n}} + \frac{b}{2}\ \frac{1}{n\sqrt{n}}(1-p)\ e^{-b/\sqrt{n}}}{p\ e^{a/\sqrt{n}} + (1-p)\ e^{-b/\sqrt{n}}}}{-\frac{1}{n^2}} \\ &= \lim_{n\to\infty} \frac{\frac{a}{2}\ \sqrt{n} p\ e^{a/\sqrt{n}} - \frac{b}{2}\ \sqrt{n}(1-p)\ e^{-b/\sqrt{n}}}{p\ e^{a/\sqrt{n}} + (1-p)\ e^{-b/\sqrt{n}}} \\ &= \frac{1}{2} \lim_{n\to\infty} \frac{\sqrt{n}\left(ap\ e^{a/\sqrt{n}} - b(1-p)\ e^{-b/\sqrt{n}}\right)}{p\ e^{a/\sqrt{n}} + (1-p)\ e^{-b/\sqrt{n}}}.\end{aligned}$$

Observe that the denominator in the last expression tends to 1 as $n \to \infty$. Thus, if we can prove that the limit of the numerator exists, then the limit of $\ln\left(\psi_{Z_n}(t)\right)$ will be 1/2 of that limit:

$$\lim_{n\to\infty} \ln\left(\psi_{Z_n}(t)\right) = \frac{1}{2} \lim_{n\to\infty} \left[\sqrt{n}\left(ap\ e^{a/\sqrt{n}} - b(1-p)\ e^{-b/\sqrt{n}}\right)\right]. \tag{7.3}$$

The limit of the right–hand side of Equation (7.3) can be computed by L'Hospital's rule by writing

$$\sqrt{n}\left(ap\ e^{a/\sqrt{n}} - b(1-p)\ e^{-b/\sqrt{n}}\right) = \frac{ap\ e^{a/\sqrt{n}} - b(1-p)\ e^{-b/\sqrt{n}}}{\frac{1}{\sqrt{n}}},$$

and observing that the numerator goes to 0 as $n$ tends to infinity by virtue of Equation (7.1). Thus, we may apply L'Hospital's Rule to obtain

$$\begin{aligned}
\lim_{n\to\infty} \sqrt{n}\left(ap\ e^{a/\sqrt{n}} - b(1-p)\ e^{-b/\sqrt{n}}\right) &= \lim_{n\to\infty} \frac{-\frac{a^2}{2n\sqrt{n}}p\ e^{a/\sqrt{n}} - \frac{b^2}{2n\sqrt{n}}(1-p)\ e^{-b/\sqrt{n}}}{-\frac{1}{2n\sqrt{n}}} \\
&= \lim_{n\to\infty} \left(a^2 p\ e^{a/\sqrt{n}} + b^2(1-p)\ e^{-b/\sqrt{n}}\right) \\
&= a^2p + b^2(1-p) \\
&= t^2,
\end{aligned}$$

by Equation (7.2). It then follows from Equation (7.3) that

$$\lim_{n\to\infty} \ln\left(\psi_{Z_n}(t)\right) = \frac{t^2}{2}.$$

Consequently, by continuity of the exponential function,

$$\lim_{n\to\infty} \psi_{Z_n}(t) = \lim_{n\to\infty} e^{\ln\left(\psi_{Z_n}(t)\right)} = e^{t^2/2},$$

which is the mgf of a Normal$(0,1)$ random variable. Thus, by the mgf Convergence Theorem, it follows that $Z_n$ converges in distribution to a standard normal random variable. Hence, for any $z \in \mathbb{R}$,

$$\lim_{n\to\infty} \Pr\left(\frac{\overline{X}_n - p}{\sqrt{p(1-p)/n}} \leqslant z\right) = \Pr\left(Z \leqslant z\right),$$

where $Z \sim$ Normal$(0,1)$. Similarly, with $-z$ instead of $z$,

$$\lim_{n\to\infty} \Pr\left(\frac{\overline{X}_n - p}{\sqrt{p(1-p)/n}} \leqslant -z\right) = \Pr\left(Z \leqslant -z\right).$$

It then follows that

$$\lim_{n\to\infty} \Pr\left(-z < \frac{\overline{X}_n - p}{\sqrt{p(1-p)/n}} \leqslant z\right) = \Pr\left(-z < Z \leqslant z\right).$$

In a later example, we shall see how this information can be used provide a good interval estimate for the true proposition $p$.

The last two examples show that if $X_1, X_2, X_3, \ldots$ are independent random variables which are distributed either Poisson$(\lambda)$ or Bernoulli$(p)$, then the random variables

$$\frac{\overline{X}_n - E(\overline{X}_n)}{\sqrt{\text{var}(\overline{X}_n)}},$$

where $\overline{X}_n$ denotes the sample mean of the random sample $\{X_1, X_2, \ldots, X_n\}$, for $n = 1, 2, 3, \ldots$, converge in distribution to a Normal$(0, 1)$ random variable; that is,

$$\frac{\overline{X}_n - E(\overline{X}_n)}{\sqrt{\text{var}(\overline{X}_n)}} \xrightarrow{D} Z \quad \text{as } n \to \infty,$$

where $Z \sim \text{Normal}(0, 1)$. It is not a coincidence that in both cases we obtain a standard normal random variable as the limiting distribution. The examples in this section are instances of general result known as the **Central Limit Theorem** to be discussed in the next section.

## 7.3 Central Limit Theorem

**Theorem 7.3.1** (Central Limit Theorem)**.** *Suppose $X_1, X_2, X_3 \ldots$ are independent, identically distributed random variables with $E(X_i) = \mu$ and finite variance $var(X_i) = \sigma^2$, for all $i$. Then*

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} Z \sim \textit{Normal}(0, 1)$$

*Thus, for large values of $n$, the distribution function for $\dfrac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$ can be approximated by the standard normal distribution.*

*Proof.* We shall prove the theorem for the special case in which the mgf of $X_1$ exists is some interval around 0. This will allow us to use the mgf Convergence Theorem (Theorem 7.2.1).

We shall first prove the theorem for the case $\mu = 0$. We then have that

$$\psi'_{X_1}(0) = 0$$

and

$$\psi''_{X_1}(0) = \sigma^2.$$

Let

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}}{\sigma}\overline{X}_n \quad \text{for } n = 1, 2, 3, \ldots,$$

since $\mu = 0$. Then, the mgf of $Z_n$ is

$$\psi_{Z_n}(t) = \psi_{\overline{X}_n}\left(\frac{t\sqrt{n}}{\sigma}\right),$$

where the mgf of $\overline{X}_n$ is

$$\psi_{\overline{X}_n}(t) = \left[\psi_{X_1}\left(\frac{t}{n}\right)\right]^n,$$

for all $n = 1, 2, 3, \ldots$ Consequently,

$$\psi_{Z_n}(t) = \left[\psi_{X_1}\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n, \quad \text{for } n = 1, 2, 3, \ldots$$

To determine if the limit of $\psi_{Z_n}(t)$ as $n \to \infty$ exists, we first consider $\ln(\psi_{Z_n}(t))$:

$$\ln\left(\psi_{Z_n}(t)\right) = n \ln \psi_{X_1}\left(\frac{t}{\sigma\sqrt{n}}\right).$$

Set $a = t/\sigma$ and introduce the new variable $u = 1/\sqrt{n}$. We then have that

$$\ln\left(\psi_{Z_n}(t)\right) = \frac{1}{u^2} \ln \psi_{X_1}(au).$$

Thus, since $u \to 0$ as $n \to \infty$, we have that, if $\lim_{u\to 0} \frac{1}{u^2} \ln \psi_{X_1}(au)$ exists, it will be the limit of $\ln\left(\psi_{Z_n}(t)\right)$ as $n \to \infty$. Since $\psi_{X_1}(0) = 1$, we can apply L'Hospital's Rule to get

$$\begin{aligned} \lim_{u\to 0} \frac{\ln(\psi_{X_1}(au))}{u^2} &= \lim_{u\to 0} \frac{a\frac{\psi'_{X_1}(au)}{\psi_{X_1}(au)}}{2u} \\ &= \frac{a}{2} \lim_{u\to 0} \left(\frac{1}{\psi_{X_1}(au)} \cdot \frac{\psi'_{X_1}(au)}{u}\right). \end{aligned} \tag{7.4}$$

Now, since $\psi'_{X_1}(0) = 0$, we can apply L'Hospital's Rule to compute the limit

$$\lim_{u\to 0} \frac{\psi'_{X_1}(au)}{u} = \lim_{u\to 0} \frac{a\psi''_{X_1}(au)}{1} = a\psi''_{X_1}(0) = a\sigma^2.$$

It then follows from Equation 7.4 that

$$\lim_{u\to 0} \frac{\ln(\psi_{X_1}(au))}{u^2} = \frac{a}{2} \cdot a\sigma^2 = \frac{t^2}{2},$$

since $a = t/\sigma$. Consequently,

$$\lim_{n\to\infty} \ln\left(\psi_{Z_n}(t)\right) = \frac{t^2}{2},$$

which implies that

$$\lim_{n\to\infty} \psi_{Z_n}(t) = e^{t^2/2},$$

the mgf of a Normal$(0, 1)$ random variable. It then follows from the mgf Convergence Theorem that

$$Z_n \xrightarrow{D} Z \sim \text{Normal}(0, 1) \quad \text{as} \quad n \to \infty.$$

Next, suppose that $\mu \neq 0$ and define $Y_k = X_k - \mu$ for each $k = 1, 2, 3 \ldots$ Then $E(Y_k) = 0$ for all $k$ and $\text{var}(Y_k) = E((X_k - \mu)^2) = \sigma^2$. Thus, by the preceding,

$$\frac{\sum_{k=1}^{n} Y_k/n}{\sigma/\sqrt{n}} \xrightarrow{D} Z \sim \text{Normal}(0, 1),$$

or

$$\frac{\sum_{k=1}^{n} (X_k - \mu)/n}{\sigma/\sqrt{n}} \xrightarrow{D} Z \sim \text{Normal}(0, 1),$$

or

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} Z \sim \text{Normal}(0, 1),$$

which we wanted to prove. □

**Example 7.3.2** (Trick coin example, revisited)**.** Recall the first example we did in this course. The one about determining whether we had a trick coin or not. Suppose we toss a coin 500 times and we get 225 heads. Is that enough evidence to conclude that the coin is not fair?

Suppose the coin was fair. What is the likelihood that we will see 225 heads or fewer in 500 tosses? Let $Y$ denote the number of heads in 500 tosses. Then, if the coin is fair,

$$Y \sim \text{Binomial}\left(\frac{1}{2}, 500\right)$$

Thus,

$$P(X \leq 225) = \sum_{k=0}^{225} \binom{1}{2k} \left(\frac{1}{2}\right)^{500}.$$

We can do this calculation or approximate it as follows:

By the Central Limit Theorem, we know that

$$\frac{Y_n - np}{\sqrt{np(1-p)}} \approx Z \sim \text{Normal}(0, 1)$$

if $Y_n \sim$Binomial$(n, p)$ and $n$ is large. In this case $n = 500$, $p = 1/2$. Thus, $np = 250$ and $\sqrt{np(1-p)} = \sqrt{250(\frac{1}{2})} \doteq 11.2$. We then have that

$$\begin{aligned} \Pr(X \leq 225) &\doteq \Pr\left(\frac{X - 250}{11.2} \leqslant 22.23\right) \\ &\doteq \Pr(Z \leqslant -2.23) \\ &\doteq \int_{-\infty}^{-2.23} \frac{1}{\sqrt{2\pi}}\, e^{-t^2/2}\, \mathrm{d}z. \end{aligned}$$

The value of $\Pr(Z \leqslant -2.23)$ can be obtained in at least two ways:

1. Using the `normdist` function in MS Excel:

$$\Pr(Z \leqslant -2.23) = \texttt{normdist}(-2.23, 0, 1, \texttt{true}) \approx 0.013.$$

2. Using the table of the standard normal distribution function on page 778 in the text. This table lists values of the cdf, $F_Z(z)$, of $Z \sim \text{Normal}(0, 1)$ with an accuracy of four decimal palces, for positive values of $z$, where $z$ is listed with up to two decimal places.

   Values of $F_Z(z)$ are not given for negative values of $z$ in the table on page 778. However, these can be computed using the formula

$$F_Z(-z) = 1 - F_Z(z)$$

   for $z \geqslant 0$, where $Z \sim \text{Normal}(0, 1)$.

   We then get

$$F_Z(-2.23) = 1 - F_Z(2.23) \approx 1 - 0.9871 = 0.0129$$

We then get that $\Pr(X \leq 225) \approx 0.013$, or about 1.3%. This is a very small probability. So, it is very likely that the coin we have is the trick coin.

# Chapter 8

# Introduction to Estimation

In this chapter we see how the ideas introduced in the previous chapter can be used to estimate the proportion of a voters in a population that support a candidate based on the sample mean of a random sample.

## 8.1 Point Estimation

We have seen that if $X_1, X_2, \ldots, X_n$ is a random sample from a distribution of mean $\mu$, then the expected value of the sample mean $\overline{X}_n$ is

$$E(\overline{X}_n) = \mu.$$

We say that $\overline{X}_n$ is an **unbiased** estimator for the mean $\mu$.

**Example 8.1.1** (Unbiased Estimation of the Variance)**.** Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution of mean $\mu$ and variance $\sigma^2$. Consider

$$\begin{aligned}
\sum_{k=1}^{n}(X_k - \mu)^2 &= \sum_{k=1}^{n}\left[X_k^2 - 2\mu X_k + \mu^2\right] \\
&= \sum_{k=1}^{n} X_k^2 - 2\mu\sum_{k=1}^{n} X_k + n\mu^2 \\
&= \sum_{k=1}^{n} X_k^2 - 2\mu n\overline{X}_n + n\mu^2.
\end{aligned}$$

On the other hand,

$$\begin{aligned}\sum_{k=1}^{n}(X_k-\overline{X}_n)^2 &= \sum_{k=1}^{n}\left[X_k^2-2\overline{X}_nX_k+\overline{X}_n^2\right]\\ &= \sum_{k=1}^{n}X_k^2-2\overline{X}_n\sum_{k=1}^{n}X_k+n\overline{X}_n^2\\ &= \sum_{k=1}^{n}X_k^2-2n\overline{X}_n\overline{X}_n+n\overline{X}_n^2\\ &= \sum_{k=1}^{n}X_k^2-n\overline{X}_n^2.\end{aligned}$$

Consequently,

$$\sum_{k=1}^{n}(X_k-\mu)^2-\sum_{k=1}^{n}(X_k-\overline{X}_n)^2 = n\overline{X}_n^2-2\mu n\overline{X}_n+n\mu^2=n(\overline{X}_n-\mu)^2.$$

It then follows that

$$\sum_{k=1}^{n}(X_k-\overline{X}_n)^2=\sum_{k=1}^{n}(X_k-\mu)^2-n(\overline{X}_n-\mu)^2.$$

Taking expectations on both sides, we get

$$\begin{aligned}E\left(\sum_{k=1}^{n}(X_k-\overline{X}_n)^2\right) &= \sum_{k=1}^{n}E\left[(X_k-\mu)^2\right]-nE\left[(\overline{X}_n-\mu)^2\right]\\ &= \sum_{k=1}^{n}\sigma^2-n\mathrm{var}(\overline{X}_n)\\ &= n\sigma^2-n\frac{\sigma^2}{n}\\ &= (n-1)\sigma^2.\end{aligned}$$

Thus, dividing by $n-1$,

$$E\left(\frac{1}{n-1}\sum_{k=1}^{n}(X_k-\overline{X}_n)^2\right)=\sigma^2.$$

Hence, the random variable

$$S^2=\frac{1}{n-1}\sum_{k=1}^{n}(X_k-\overline{X}_n)^2,$$

called the **sample variance**, is an unbiased estimator of the variance.

## 8.2 Estimating the Mean

In Problem 1 of Assignment #20 you were asked to show that the sample means, $\overline{X}_n$, converge in distribution to a limiting distribution with pmf

$$p(x) = \begin{cases} 1 & \text{if } x = \mu; \\ 0 & \text{elsewhere.} \end{cases}$$

It then follows that, for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr(\overline{X}_n \leqslant \mu + \varepsilon) = \Pr(X_\mu \leqslant \mu + \varepsilon) = 1,$$

while

$$\lim_{n\to\infty} \Pr(\overline{X}_n \leqslant \mu - \varepsilon) = \Pr(X_\mu) \leqslant \mu - \varepsilon) = 0.$$

Consequently,

$$\lim_{n\to\infty} \Pr(\mu - \varepsilon < \overline{X}_n \leqslant \mu + \varepsilon) = 1$$

for all $\varepsilon > 0$. Thus, with probability 1, the sample mean will be within and arbitrarily small distance from the mean of the distribution as the sample size increases to infinity.

This conclusion was attained through the use of the mgf Convergence Theorem, which assumes that the mgf of $X_1$ exists. However, the result is true more generally; for example, we only need to assume that the variance of $X_1$ exists. This follows from the following inequality

**Theorem 8.2.1** (Chebyshev Inequality)**.** *Let $X$ be a random variable with mean $\mu$ and variance var$(X)$. Then, for every $\varepsilon > 0$,*

$$Pr(|X - \mu| \geqslant \varepsilon) \leqslant \frac{var(X)}{\varepsilon^2}.$$

*Proof:* We shall prove this inequality for the case in which $X$ is continuous with pdf $f_X$.

Observe that $\text{var}(X) = E[(X-\mu)^2] = \displaystyle\int_{-\infty}^{\infty} |x-\mu|^2 f_X(x)\ \mathrm{d}x$. Thus,

$$\text{var}(X) \geqslant \int_{A_\varepsilon} |x-\mu|^2 f_X(x)\ \mathrm{d}x,$$

where $A_\varepsilon = \{x \in \mathbb{R} \mid |x-\mu| \geqslant \varepsilon\}$. Consequently,

$$\text{var}(X) \geqslant \varepsilon^2 \int_{A_\varepsilon} f_X(x)\ \mathrm{d}x = \varepsilon^2 \Pr(A_\varepsilon).$$

we therefore get that

$$\Pr(A_\varepsilon) \leqslant \frac{\text{var}(X)}{\varepsilon^2},$$

or

$$\Pr(|X-\mu| \geqslant \varepsilon) \leqslant \frac{\text{var}(X)}{\varepsilon^2}.$$

□

Applying Chebyshev Inequality to the case in which $X$ is the sample mean, $\overline{X}_n$, we get

$$\Pr(|\overline{X}_n - \mu| \geqslant \varepsilon) \leqslant \frac{\text{var}(\overline{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}.$$

We therefore obtain that

$$\Pr(|\overline{X}_n - \mu| < \varepsilon) \geqslant 1 - \frac{\sigma^2}{n\varepsilon^2}.$$

Thus, letting $n \to \infty$, we get that, for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr(|\overline{X}_n - \mu| < \varepsilon) = 1.$$

We then say that $\overline{X}_n$ **converges to $\mu$ in probability** and write

$$\overline{X}_n \xrightarrow{\Pr} \mu \quad \text{as } n \to \infty.$$

This is known as the weak **Law of Large Numbers**.

**Definition 8.2.2** (Convergence in Probability)**.** A sequence, $(Y_n)$, of random variables is said to converge in probability to $b \in \mathbb{R}$, if for every $\varepsilon > 0$

$$\lim_{n\to\infty} \Pr(|Y_n - b| < \varepsilon) = 1.$$

We write

$$Y_n \xrightarrow{\Pr} b \quad \text{as } n \to \infty.$$

**Theorem 8.2.3** (Slutsky's Theorem)**.** *Suppose that $(Y_n)$ converges in probability to $b$ and that $g$ is a function which is continuous at $b$ as $n \to \infty$. Then, $(g(Y_n))$ converges in probability to $g(b)$ as $n \to \infty$.*

*Proof:* Let $\varepsilon > 0$ be given. Since $g$ is continuous at $b$, there exists $\delta > 0$ such that

$$|y - b| < \delta \Rightarrow |g(y) - g(b)| < \varepsilon.$$

It then follows that the event $A_\delta = \{y \mid |y - b| < \delta\}$ is a subset the event $B_\varepsilon = \{y \mid |g(y) - g(b)| < \varepsilon\}$. Consequently,

$$\Pr(A_\delta) \leqslant \Pr(B_\varepsilon).$$

It then follows that

$$\Pr(|Y_n - b| < \delta) \leqslant \Pr(|g(Y_n) - g(b)| < \varepsilon) \leqslant 1. \tag{8.1}$$

Now, since $Y_n \xrightarrow{\Pr} b$ as $n \to \infty$,

$$\lim_{n\to\infty} \Pr(|Y_n - b| < \delta) = 1.$$

It then follows from Equation (8.1) and the Squeeze or Sandwich Theorem that

$$\lim_{n\to\infty} \Pr(|g(Y_n) - g(b)| < \varepsilon) = 1.$$

□

Since the sample mean, $\overline{X}_n$, converges in probability to the mean, $\mu$, of sampled distribution, by the weak Law of Large Numbers, we say that $\overline{X}_n$ is a **consistent** estimator for $\mu$.

## 8.3 Estimating Proportions

**Example 8.3.1** (Estimating Proportions, Revisited)**.** Let $X_1, X_2, X_3, \ldots$ denote independent identically distributed (iid) Bernoulli($p$) random variables. Then the sample mean, $\overline{X}_n$, is an unbiased and consistent estimator for $p$. Denoting $\overline{X}_n$ by $\widehat{p}_n$, we then have that

$$E(\widehat{p}_n) = p \quad \text{for all } n = 1, 2, 3, \ldots,$$

and

$$\widehat{p}_n \xrightarrow{\Pr} p \quad \text{as } n \to \infty;$$

that is, for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr(|\widehat{p}_n - p| < \varepsilon) = 1.$$

By Slutsky's Theorem, we also have that

$$\sqrt{\widehat{p}_n(1-\widehat{p}_n)} \xrightarrow{\Pr} \sqrt{p(1-p)} \quad \text{as } n \to \infty.$$

Thus, the statistic $\sqrt{\widehat{p}_n(1-\widehat{p}_n)}$ is a consistent estimator of the standard deviation $\sigma = \sqrt{p(1-p)}$ of the Bernoulli($p$) trials $X_1, X_2, X_3, \ldots$

Now, by the Central Limit Theorem, we have that

$$\lim_{n\to\infty} \Pr\left(\frac{\widehat{p}_n - p}{\sigma/\sqrt{n}} \leqslant z\right) = \Pr(Z \leqslant z),$$

where $Z \sim \text{Normal}(0, 1)$, for all $z \in \mathbb{R}$. Hence, since $\sqrt{\widehat{p}_n(1-\widehat{p}_n)}$ is a consistent estimator for $\sigma$, we have that, for large values of $n$,

$$\Pr\left(\frac{\widehat{p}_n - p}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}/\sqrt{n}} \leqslant z\right) \approx \Pr(Z \leqslant z),$$

for all $z \in \mathbb{R}$. Similarly, for large values of $n$,

$$\Pr\left(\frac{\widehat{p}_n - p}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}/\sqrt{n}} \leqslant -z\right) \approx \Pr(Z \leqslant -z).$$

subtracting this from the previous expression we get

$$\Pr\left(-z < \frac{\widehat{p}_n - p}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}/\sqrt{n}} \leqslant z\right) \approx \Pr(-z < Z \leqslant z)$$

for large values of $n$, or

$$\Pr\left(-z \leqslant \frac{p-\widehat{p}_n}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}/\sqrt{n}} < z\right) \approx \Pr(-z < Z \leqslant z)$$

for large values of $n$.

Now, suppose that $z > 0$ is such that $\Pr(-z < Z \leqslant z) \geqslant 0.95$. Then, for that value of $z$, we get that, approximately, for large values of $n$,

$$\Pr\left(\widehat{p}_n - z\frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}} \leqslant p < \widehat{p}_n + z\frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}}\right) \geqslant 0.95$$

Thus, for large values of $n$, the intervals

$$\left[\widehat{p}_n - z\frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}},\ \widehat{p}_n + z\frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}}\right)$$

have the property that the probability that the true proportion $p$ lies in them is at least 95%. For this reason, the interval

$$\left[\widehat{p}_n - z\frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}},\ \widehat{p}_n + z\frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}}\right)$$

is called the 95% **confidence interval estimate for the proportion** $p$. To find the value of $z$ that yields the 95% confidence interval for $p$, observe that

$$\Pr(-z < Z \leqslant z) = F_Z(z) - F_Z(-z) = F_Z(z) - (1 - F_Z(z)) = 2F_Z(z) - 1.$$

Thus, we need to solve for $z$ in the inequality

$$2F_Z(z) - 1 \geqslant 0.95$$

or

$$F_Z(z) \geqslant 0.975.$$

This yields $z = 1.96$. We then get that the **approximate** 95% confidence interval estimate for the proportion $p$ is

$$\left[\widehat{p}_n - 1.96\frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}},\ \widehat{p}_n + 1.96\frac{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}}{\sqrt{n}}\right)$$

**Example 8.3.2.** A random sample of 600 voters in a certain population yields 53% of the voters supporting certain candidate. Is this enough evidence to to conclude that a simple majority of the voters support the candidate?

> ***Solution:*** Here we have $n = 600$ and $\widehat{p}_n = 0.53$. An approximate 95% confidence interval estimate for the proportion of voters, $p$, who support the candidate is then
>
> $$\left[0.53 - 1.96\frac{\sqrt{0.53(0.47)}}{\sqrt{600}},\ 0.53 + 1.96\frac{\sqrt{0.53(0.47)}}{\sqrt{600}}\right),$$

or about $[0.49, 0.57)$. Thus, there is a 95% chance that the true proportion is below 50%. Hence, the data do not provide enough evidence to support that assertion that a majority of the voters support the given candidate. □

**Example 8.3.3.** Assume that, in the previous example, the sample size is 1900. What do you conclude?

***Solution:*** In this case the approximate 95% confidence interval for $p$ is about $[0.507, 0.553)$. Thus, in this case we are "95% confident" that the data support the conclusion that a majority of the voters support the candidate. □