Chapter 1

Introduction

This class is a calculus based introduction to the theory of probability. Our goal is to cover topics such as random variables, random vectors, law of large numbers, central limit theorem, point estimation and confidence intervals, hypothesis testing and P-value, as well as linear regression.

1.1 Sample Space

Traditionally, a **sample space** is denoted by Ω and a generic element in Ω is denoted by ω.

In every investigation of probability and statistics, there is *always* an underlying sample space, explicitly or implicitly. In some advanced and very specific topics, sample space can be an important component. However, by and large, in most studies we will pay no attention to the details of the underlying sample space, because it simply does not matter. We nearly always work with *random variables* defined on the sample space, not the sample space itself.

In many introductory textbooks, sample space is "defined" as the collection of all possible outcomes when one studies a phenomenon or perform experiments, while each element $\omega \in \Omega$ represents one possible outcome. For example, if a fair coin is tossed once, then one can define a sample space with only two elements, namely $\Omega = \{H, T\}$. However, this is only one choice of a sample space, a choice for convenience and clarity. You can make your sample space so complicated that it contains possible outcomes from (say) future baseball games, stock market movements, seismic activities around the globe, and so on. And yet, in the midst of all, you are tossing a fair coin, once. No choice of a sample space will yield a different

probabilistic prediction of your toss — it is always 50% heads and 50% tails.

In short, the choice of sample space does not matter, what matters is the probabilistic property of the outcome of interest. In this class, we will occasionally use sample spaces to explain concepts and facilitate computations.

1.2 Random Variables

Typically a **random variable** is denoted by capital letters such as X, Y, Z, while lower case letters such as a, b, c are reserved for constants.

Definition 1.1. *Any function $\Omega \to \mathbb{R}$ is said to be a* **random variable**.

That is, a random variable X in probability theory is merely a function on the sample space. It maps an element $\omega \in \Omega$ to a real number $X(\omega)$. In the setup of every study in probability, it is assumed, explicitly or implicitly, that there is a single sample space and all relevant random variables are defined on this common sample space. This allows us to perform algebraic operations among random variables.

For example, let X and Y be random variables. Here we have already assumed that they are both defined on some common sample space Ω. Then algebraic operations such as X^2, $X + Y$, $\max(X, Y)$, and so on, generate new random variables. More precisely,

$$\begin{aligned} X^2: &\quad \omega \mapsto X^2(\omega) \\ X+Y: &\quad \omega \mapsto X(\omega) + Y(\omega) \\ \max(X,Y): &\quad \omega \mapsto \max\{X(\omega), Y(\omega)\}. \end{aligned}$$

The study of random variables, their properties/relations and interpretations, is in the center of probability theory.

Example 1.1. Tossing a coin repeatedly. A sample space Ω can be defined as the collection of all elements ω of form

$$\omega = \omega_1 \omega_2 \cdots \omega_n \cdots$$

where $\omega_n = H$ or T denotes the outcome of the n-th toss. Define a sequence of random variables $X_1, X_2, \ldots$ such that

$$X_n(\omega) \doteq \begin{cases} 1 & \text{if } \omega_n = H \\ 0 & \text{if } \omega_n = T \end{cases}.$$

For example, if $\omega = HHTHT\ldots$, then $X_1(\omega) = X_2(\omega) = X_4(\omega) = 1$ and $X_3(\omega) = X_5(\omega) = 0$.

X_n represents the outcome of the n-th toss: if $X_n = 1$, it means the n-th toss is heads; if $X_n = 0$, it means the n-th toss is tails. These X_n's can be used to create more random variables. For example

$$X_1 + X_2 + \cdots + X_n \quad \text{and} \quad n - (X_1 + X_2 + \cdots + X_n)$$

represent the total number of heads and tails in the first n tosses, respectively. The random variable

$$T = \min\{n \geq 1 : \ X_n = 1\}$$

represents the number of tosses until the first heads.

1.3 Events

An **event** is simply a subset of the sample space. For example, to study the event "getting even number of heads in two tosses of a coin", one can define a sample space

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

and the event of interest is the subset

$$A = \{HH, TT\}.$$

Events can also be defined through random variables. For example, given a random variable X and two constants $a \leq b$, the event $A \doteq \{a \leq X \leq b\}$ is the shorthanded notation for

$$A = \{\omega \in \Omega : \ a \leq X(\omega) \leq b\}.$$

If an outcome $\omega \in A$, then "event A happens". If not, then "event A does not happen".

Given events A and B, the common set notation and their interpretation are as follows:

- $A \subseteq B$: *when event A happens, B must happen.*
- $A \cap B$ or AB: *Events A and B both happen.*
- $A \cup B$: *Events A or B happens.*
- A^c or $\bar{A}$ (complement): *Event A does not happen.*
- A and B are **disjoint** or **mutually exclusive** ($A \cap B = \emptyset$): *Events A and B cannot happen at the same time.*

1.4 Probabilities of Events

Given an event $A \subseteq \Omega$, *the probability that event A happens* is simply denoted by $P(A)$. Probability means chance. It means that if the same experiment is performed many many times, then roughly $P(A)$ fraction of the time the outcome (ω) belongs to the subset A (event A happens).

Example 1.2. Toss a fair coin three times, what is the probability that there are even number of heads?

Solution: The calculation is rather straightforward. There are four possible outcomes with even number of heads:

$$HHT,\ HTH,\ THH,\ TTT.$$

Each one of these outcomes has probability

$$\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}.$$

There are in total four such outcomes, thus the probability of even number of heads is

$$4 \times \frac{1}{8} = \frac{1}{2}.$$

Another way to calculate this probability is that there are in total $2^3 = 8$ possible outcomes (sample space), and 4 outcomes with even number of heads. Because it is a fair coin, **each outcome is equally likely**. Therefore the probability is

$$\frac{4}{8} = \frac{1}{2}.$$

The technique in the above computation is **counting**. Basically one counts the number of outcomes that make the event happen, and then either multiply it with the probability of each outcome, or divide it by the total number of possible outcomes. The idea is very simple. But the key condition for this approach to work is that *each outcome is equally likely*.

Example 1.3. (*Is each outcome equally likely?*) In general, whether each outcome is equally likely is a very easy question to answer. Take coin toss as an example. Each possible outcome is equally likely only if the coin is fair. In some problems, however, this equal-probability-condition becomes less obvious.

The classical brain teaser of *Monty Hall Problem* is a deceptively simple, yet misleading, probabilistic puzzle. You are in a game show and presented

with three closed door. The host knows what is behind each door, but you don't. All you know is that there is nothing behind two of the doors, and a brand new car that is worth tens of thousands of dollars behind one of the doors. You are to open one of the doors and claim what is behind.

The game plays like this. At the beginning you are asked to pick a door. Then the host will open one of the remaining two doors to show you that there is nothing behind it. Now you have to make a choice.

1. Stick to your original choice of the door.

2. Pay $10 and switch to the other closed door.

There is no "pay $10" for switching in the original version of the problem. It is added here as a tie-breaker. What should you do?

One argument is for Choice 1. Think about it. Now there are two closed doors left. One door has a car behind it and the other has nothing. Thus the chance of each door having a car behind it is 50%. Paying $10 to switch to the other equal-probability door is a losing play.

The mistake in this argument is to assume implicitly that each door has an "equal probability" to have the car behind it. Actually, when you have selected a door at the beginning of the show, the probability that it has a car behind it is one third. This probability will *not* change regardless of what has happened afterwards. Thus, the probability that the other door has a car behind it is two thirds. Paying $10 to have an extra one third chance to claim a new car is a winning play. The correct action is Choice 2.

Example 1.4. A standard 52-card deck has 26 red cards and 26 black cards. For a well shuffled deck,

1. what is the probability that the first card is red?

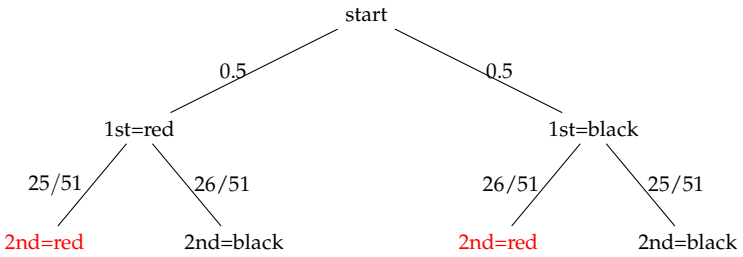
2. what is the probability that the second card is red?

Solution: The answer to the first question is rather trivial. The correct probability is

$$\frac{26}{52} = \frac{1}{2}$$

because every card is equally likely to be the first card.

What is more interesting is the second question. Because the way this question is phrased, many of us will get busy since it is very natural to think that the outcome of the second card should be affected by the first card. This reasoning can be summarized in the following tree.



Therefore, the probability for the second card to be red is the sum of all the probabilities of the branches (paths) with red leaves:

$$\sum_{\text{path with red leaf}} P(\text{path}) = \frac{1}{2} \cdot \frac{25}{51} + \frac{1}{2} \cdot \frac{26}{51} = \frac{1}{2}.$$

This approach is entirely correct. It is actually based on our intuitive understanding of the conditional probability and the so-called **law of total probability**, a powerful result to be established in full generality later.

Incidentally, there is a simpler way to solve the second question. Ignore the first card. The total number of possible outcomes for the second card is 52, each one equally likely, half of them red. Thus the probability of the second card being red is one half. By the same token, the probability of the third card (fourth, fifth, ...) being red is also one half.

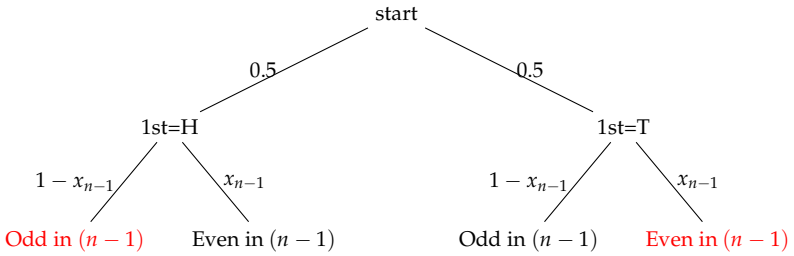
Example 1.5. Toss a fair coin n times, what is the probability that there are even number of heads? This is an extension of Example 1.2 where $n = 3$. If you try a few more cases with small n, you will convince yourself that the probability is most likely 50% for all n. But how do we justify it?

The most natural method is probably counting. After all, every outcome is equally likely. However, you will soon realize that counting the number of outcomes with even number of heads is not easy. Yes, if you get creative, you *can* solve this counting problem in a very elegant way. But this is not the approach we wish to investigate, for a couple of reasons: (1) the trick of counting for this problem lacks generality; (2) we wish to develop a method that works even if the condition of equal probability fails (thus we cannot use the counting method).

The idea is to study the behavior the probability with n as a parameter. We would like to know how this probability changes according to n. To this end, define

$$x_n = P(\text{even number of heads in } n \text{ tosses}),$$

which is exactly the quantity of interest. We emphasize that the index n is important, since our goal is to find out how x_n evolves as n changes. For general n, consider the following tree.



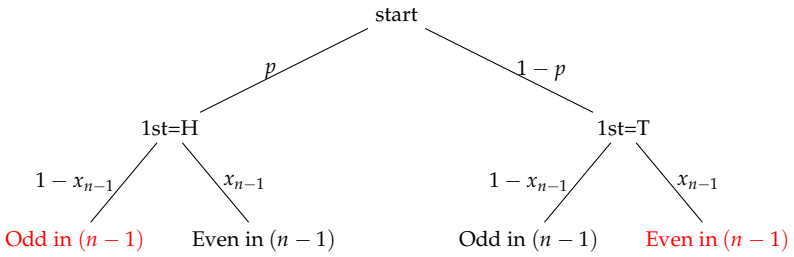
It follows that

$$x_n = \sum_{\text{path with red leaf}} P(\text{path}) = \frac{1}{2}(1 - x_{n-1}) + \frac{1}{2}x_{n-1} = \frac{1}{2}.$$

This technique is called *first step analysis*, very useful in probability and optimization. The similarity to Example 1.4 is obvious. Thus it is not surprising that this technique is also based on the law of total probability. The term *first step* does not necessarily mean the first action in a sequence. For example, we can define the first layer of the tree according to the outcome from the last toss instead.

Example 1.6. Let us look at a generalization of Example 1.5, where we drop the assumption of a fair coin and assume $P(H) = p$ for some $0 < p < 1$.

Method of counting is not useful because the outcomes are no longer equally likely. However, first step analysis still works. Using the same terminologies and notation, we have the same tree but with different probabilities.



It follows that

$$x_n = p(1 - x_{n-1}) + (1 - p)x_{n-1} \;=\; p + (1 - 2p)x_{n-1},$$

with the trivial initial condition $x_1 = P(T) = 1 - p$. This gives us a recursive way to compute each x_n. For such a simple difference equation, one can actually solve it explicitly. To this end, observe that the difference equation amounts to

$$x_n - \frac{1}{2} = (1-2p)\left(x_{n-1} - \frac{1}{2}\right)$$

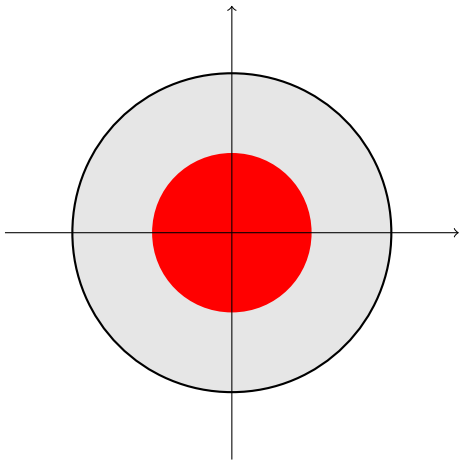
Repeatedly applying this relation, we have

$$\begin{aligned} x_n - \frac{1}{2} &= (1-2p)\left(x_{n-1} - \frac{1}{2}\right) = (1-2p)^2\left(x_{n-2} - \frac{1}{2}\right) \\ &= \cdots = (1-2p)^{n-1}\left(x_1 - \frac{1}{2}\right) \\ &= (1-2p)^{n-1}\left(1 - p - \frac{1}{2}\right) = \frac{1}{2}(1-2p)^n. \end{aligned}$$

Therefore,

$$x_n = \frac{1}{2}\left[1 + (1-2p)^n\right].$$

Example 1.7. Randomly select a point from a disc. What is the probability that this point is less than half of the radius to the center?



The radius of the red disc is half of the gray one. The question is asking the probability that a random selected point from the gray disc falls into the red disc. It is quite intuitive that the probability should be

$$\frac{\text{area of green disc}}{\text{area of red disc}} = \left(\frac{1}{2}\right)^2 = \frac{1}{4}.$$

Our intuition is based on the implicit understanding that each point in the gray disc is *equally likely* if a point is randomly chosen from the disc, and somehow we are *counting* the number of points by area. But now that there are infinitely many points in the disc, what does *equally likely* mean in this scenario? And how come area can be used as a way to count points? Lastly, if points can be chosen with different likelihood, what model should we use to describe this difference? How should we compute probabilities then?

This example raises more questions than answers. We will answer them when we introduce probability density functions.

1.5 Axioms of Probability

So far we have seen a few examples of computing probabilities without even formally introducing the rules of probability. It is surprising that the whole theory of probability is based on three simple axioms.

Let Ω be a sample space. The probability of an arbitrary event $A \subseteq \Omega$ is denote by $P(A)$. It satisfies the following three axioms.

- **Axiom 1**: $P(\Omega) = 1$.
- **Axiom 2**: $0 \leq P(A) \leq 1$ for every event $A \subseteq \Omega$.
- **Axiom 3** *(countable additivity)*: If $\{A_1, A_2, A_3, \ldots\}$ is a sequence of disjoint events, then

$$P\left(A_1 \cup A_2 \cup A_3 \cup \cdots\right) = \sum_{n=1}^{\infty} P(A_n).$$

Introduced by mathematician Andrey Kolmogorov in 1933, these three axioms become the foundation of the entire theory of probability and statistics. Their significance cannot be overstated.

1.6 Rules of Probabilities

With the three axioms, we can derive all the lemmas, propositions, and theorems in probability and statistics. In this section we collect a few very basic results.

Lemma 1.1. *Let Ω denote the sample space. Let A and B be arbitrary events.*

- $P(\emptyset) = 0$;
- $P(A) + P(A^c) = 1$;
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$;
- *(finite additivity) Let $n \geq 1$ be arbitrary. Let $\{A_1, A_2, \ldots, A_n\}$ be any finite sequence of disjoint events. Then*

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

Proof. The proof will be based on the three axioms. We first prove $P(\emptyset) = 0$. Define a sequence of events $\{\emptyset, \emptyset, \emptyset, \ldots\}$. By definition, this is a sequence of disjoint events, whose union is again $\emptyset$. Thus

$$P(\emptyset) = P(\emptyset \cup \emptyset \cup \cdots) = P(\emptyset) + P(\emptyset) + \cdots.$$

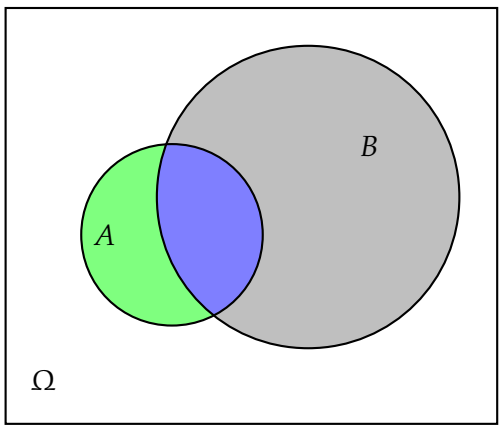
This identity cannot be satisfied unless $P(\emptyset) = 0$. This completes the proof of the first claim.

Now we should prove the last claim of finite additivity. Define a sequence of events $\{A_1, A_2, \ldots, A_n, \emptyset, \emptyset, \ldots\}$. This is a disjoint sequence. Therefore, by the axioms, we have

$$\begin{aligned} & P(A_1 \cup \cdots \cup A_n \cup \emptyset \cup \emptyset \cup \cdots) \\ = \quad & P(A_1) + \cdots + P(A_n) + P(\emptyset) + P(\emptyset) + \cdots. \end{aligned}$$

Since $P(\emptyset) = 0$, the finite additivity follows readily.

The second claim follows readily from the finite additivity and the axiom of $P(\Omega) = 1$ since $A \cap A^c = \emptyset$ and $A \cup A^c = \Omega$. As for the third claim, we have the following *Venn diagram*:



Define events $A_1 \doteq A \cap B^c$ (green), $A_2 \doteq A \cap B$ (blue), $A_3 \doteq A^c \cap B$ (gray). It is not hard to verify that these three events are disjoint and satisfy

$$A_1 \cup A_2 = A, \quad A_3 \cup A_2 = B, \quad A_1 \cup A_2 \cup A_3 = A \cup B.$$

Therefore, by finite additivity, we have

$$\begin{aligned} P(A) &= P(A_1) + P(A_2) \\ P(B) &= P(A_3) + P(A_2) \\ P(A \cup B) &= P(A_1) + P(A_2) + P(A_3). \end{aligned}$$

It follows readily that

$$P(A \cup B) = P(A) + P(B) - P(A_2) = P(A) + P(B) - P(A \cap B).$$

We complete the proof. ■

1.7 Conditional Probability

We have seen conditional probability in Example 1.4. For example, given that the first card in a well shuffled deck is red, the second card has a probability

$$\frac{25}{51}$$

to be red. In general, knowing an event A has happened may change our view about another event B. An extreme scenario is when events A and B are disjoint. Then knowing event A has occurred, we know event B cannot happen!

Definition 1.2. *Let A and B be two arbitrary events with $P(B) > 0$. The* **conditional probability of A given** *B, denoted by $P(A|B)$, is defined as*

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

If $P(B) = 0$, this conditional probability is undefined.

Example 1.8. Your probability textbook is missing. You recall this morning you went to two lectures: APMA 1655 and MUSC 1610. You estimate a 60% chance that the book is still in the lecture room of APMA 1655; a 30% chance in the room of MUSC 1610; and a 10% chance in a place you have

no idea of. You rush to the lecture room of APMA 1655. The book is not there. What is now the probability that the book is in the lecture room of MUSC 1610?

Solution: This question is to find the conditional probability $P(B|A^c)$, where the events A and B are simply:

$$\begin{aligned} A &= \{\text{textbook is in the room of APMA 1655}\}, \\ B &= \{\text{textbook is in the room of MUSC 1610}\}. \end{aligned}$$

Observing that $B \subseteq A^c$, we have

$$P(B|A^c) = \frac{P(B \cap A^c)}{P(A^c)} = \frac{P(B)}{P(A^c)} = \frac{0.3}{1-0.6} = \frac{3}{4}.$$

Example 1.9. Toss a possibly unfair coin n times. Given that there is in total one heads, what is the probability that the first toss is heads?

Solution: Assume that $P(H) = p$. Again, this question is to find the conditional probability $P(B|A)$, where

$$\begin{aligned} A &= \{\text{there is in total one heads in } n \text{ tosses}\}, \\ B &= \{\text{first toss is heads}\}. \end{aligned}$$

The event $B \cap A$ is simply that the first toss is heads and the rest $(n-1)$ tosses are all tails. Therefore,

$$P(B \cap A) = p(1-p)^{n-1}.$$

In order to compute $P(A)$, we observe that there are n possible ways A can happen: the heads appears in the k-th toss and the rest are all tails, for $k = 1, 2, \ldots, n$. Each of these outcomes has probability $p(1-p)^{n-1}$ to happen. Thus

$$P(A) = np(1-p)^{n-1}.$$

It follows that

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

Actually, given that there is only one heads, it is equally likely for the heads to appear in any toss.

Example 1.10. You are playing poker with your friend. You are dealt a pair of kings. A few raises and re-raises later, your friend goes all-in. You know your friend will only go all-in with four types of hands: a pair of aces, kings, queens, or ace-king. What is the probability that you are facing a pair of aces, the only hand that is ahead of yours?

Solution: Let us use A for ace, K for king, and Q for queen. This question is asking the conditional probability $P(E|F)$, where the events E and F are given by

$$\begin{aligned} E &= \{\text{your friend has AA}\}, \\ F &= \{\text{your friend has AA, KK, QQ, or AK}\}. \end{aligned}$$

Since $E \subseteq F$, we have

$$P(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{P(E)}{P(F)}.$$

Because you have two kings in your hand, there are 50 cards out there (52 cards minus two kings). Your friend can be dealt with any one of the

$$\binom{50}{2}$$

combinations of two cards, all equally likely. Among them, the number of combinations for AA, KK, QQ, AK are respectively,

$$\binom{4}{2}, \quad \binom{2}{2}, \quad \binom{4}{2}, \quad 4 \times 2.$$

Therefore,

$$P(E) = \frac{\binom{4}{2}}{\binom{50}{2}}, \quad P(F) = \frac{\binom{4}{2} + \binom{2}{2} + \binom{4}{2} + 4 \times 2}{\binom{50}{2}}$$

and

$$P(E|F) = \frac{P(E)}{P(F)} = \frac{\binom{4}{2}}{\binom{4}{2} + \binom{2}{2} + \binom{4}{2} + 4 \times 2} = \frac{2}{7}.$$

Interested students may also want to do the following exercises: if you call your friend's all-in, what is your overall chance of winning/tying/losing the hand? Here is the rough winning probability for KK:

$$\text{KK vs AA: 20\%}, \quad \text{KK vs QQ: 80\%}, \quad \text{KK vs AK: 65\%}$$

Exercise 1.11. The purpose of this exercise is to explain that conditional probability behaves just like a regular probability. More precisely, fix an arbitrary event C with $P(C) > 0$. For any event $A \subseteq \Omega$, define

$$Q(A) \doteq P(A|C).$$

1. Show that Q satisfies the three axioms of probability.

2. For any events $A, B \subseteq \Omega$ with $Q(B) > 0$, we can similarly define the conditional probability for Q:

$$Q(A|B) \doteq \frac{Q(A \cap B)}{Q(B)}.$$

   Show that $Q(A|B) = P(A|B, C)$. Here $P(A|B, C)$ or $P(A|BC)$ is an alternative notation for $P(A|B \cap C)$.

1.8 Multiplication Rule

The definition of conditional probability directly leads to the so-called **multiplication rule** of probability: for any two events A and B,

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

This rule holds even if $P(A) = 0$ or $P(B) = 0$. It says "*the probability that A and B both happen equals the probability that A happens, multiplied by the (conditional) probability that B happens, given that A has happened*." This rule agrees with our intuition quite well and has been used in our calculations implicitly. For example, randomly select two cards from a deck, the probability that they are both red is

$$\frac{26}{52} \times \frac{25}{51},$$

where the first fraction is the probability that the first card is red and the second fraction is the conditional probability that the second card is red, given that the first card is red.

This multiplication rule can be easily generalized to a sequence of events. That is, for any events $\{A_1, A_2, \ldots, A_n\}$, we have

$$P(A_1A_2\cdots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1A_2)\cdots P(A_n|A_1A_2\cdots A_{n-1}).$$

We will leave its proof as an exercise to interested students. Please note that here we have used notation $AB = A \cap B$.

Example 1.12. Randomly put 8 checker pieces onto an 8×8 checkerboard. What is the probability that no two pieces are in the same row or same column?

Solution: The first piece can be put anywhere on the checker board. Once it is put down, there are $(8^2 - 1)$ ways to put the second piece, among which $(8-1)^2$ of them will be in a different row and column than the first piece. Repeat this discussion for the third, fourth, ..., eighth piece. The probability of interest, by the multiplication rule, is

$$\frac{(8-1)^2}{8^2-1} \cdot \frac{(8-2)^2}{8^2-2} \cdot \cdots \cdot \frac{(8-7)^2}{8^2-7} = \frac{(7!)^2}{63 \cdot 62 \cdot \cdots \cdot 57}$$

1.9 Independent Events

Two events A and B are said to be **independent** if $P(A \cap B) = P(A)P(B)$. Intuitively speaking, if two events A and B are independent, then whether event A happens or not has no impact on event B, and vice versa.

We have been using independence implicitly already. For example, toss a coin with $P(H) = p$ three times. What is the probability that the outcome is THT? The probability of the first (or the third) toss being T is $(1-p)$, that of the second toss being H is p. Therefore, the probability that the outcome is THT is simply the product of these three individual probabilities:

$$(1-p) \times p \times (1-p) = p(1-p)^2.$$

Our implicit assumption in the computation is that the coin tosses are independent of each other. This is a very natural assumption since the outcome of one toss should not affect the outcome of the other.

Many people believe that randomness eventually evens out. For example, if someone tosses a coin five times and gets five heads, then he would think the next toss is more likely to be tails since heads and tails should

even out and tails is overdue. This is commonly referred to as *the gambler's fallacy*. Since coin tosses are independent, the chance that the next toss is tails is always 50% (assuming the coin is fair), regardless how many heads you get in a row previously. Of course, if you get twenty heads in a row, you should worry that it may not be a fair coin (or a two-headed coin), but not about the independence of tosses.

Lemma 1.2. *Let A and B be two arbitrary events. Then A and B are independent if and only if $P(B) = 0$ or $P(A|B) = P(A)$.*

Lemma 1.3. *Let A and B be two independent events. Then A and B^c are independent, so are A^c and B, A^c and B^c.*

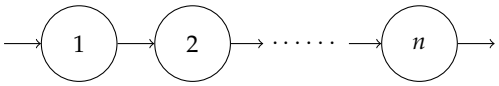
The proof of these two lemmas is very straightforward from definition, and thus omitted.

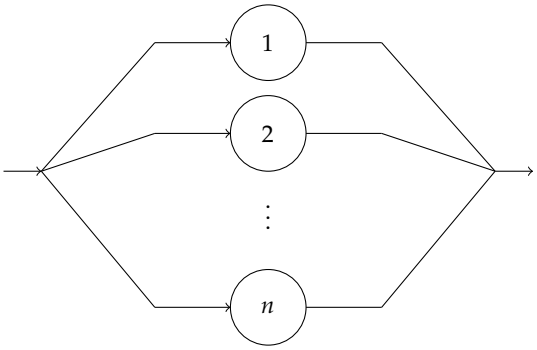
The definition of independence for multiple events $\{A_1, A_2, \ldots, A_n\}$ is slightly more complicated. We say they are **independent** if any subcollection $\{A_{j_1}, A_{j_2}, \ldots, A_{j_k}\}$ satisfies

$$P(A_{j_1} \cap A_{j_2} \cap \cdots \cap A_{j_k}) = P(A_{j_1})P(A_{j_2}) \cdots P(A_{j_k})$$

for any k and any $1 \leq j_1 < j_2 < \cdots < j_k \leq n$. In practice, independence usually arises naturally or is assumed within reason. It is rare that one has to *verify* the definition of independence.

Example 1.13. *(serial system vs. parallel system)* There are n independent components, each having probability p to function properly. What are the probabilities that these systems function properly?





Solution: The serial system functions properly if and only if every component does so. Therefore, by independence,

$$P(\text{serial system functions}) = p^n.$$

In contrast, the parallel system functions properly if and only if at least one component does so. In other words, the parallel system fails if and only if every components fails. Therefore, by independence,

$$\begin{aligned} P(\text{parallel system functions}) &= 1 - P(\text{parallel system fails}) \\ &= P(\text{every component fails}) = 1 - (1-p)^n. \end{aligned}$$

Example 1.14. *(independent events vs. disjoint events)* Independence should *not* be confused with disjointness. A common misconception is that two disjoint events are independent. But think about it, if events A and B are disjoint, then event A happens means event B cannot happen, and vice versa; if event A does not happen, then event B has a higher chance to happen, and vice versa. Therefore two disjoint events are very much *dependent* (of course, they can be independent, but only in the trivial cases when one of the events has probability zero).

Example 1.15. Here are a couple of curious cases of independence. They illustrate the distinction between the mathematically defined "independence" and the common sense "independence".

1. If A and B are independent and A and C are independent, are A and $B \cup C$ independent?
2. Three events A, B, C are pairwise independent. That is, A and B are independent, B and C are independent, and A and C are independent. Does this imply that the events $\{A, B, C\}$ are independent?

The answer to either is "No". Consider the following example. Toss a fair coin twice. Define

$$\begin{aligned} A &= \text{the first toss is heads} \\ B &= \text{the second toss is heads} \\ C &= \text{there is exactly one heads in two tosses.} \end{aligned}$$

It is not hard to compute that

$$P(A) = P(B) = \frac{1}{2}, \;\; P(C) = P(HT) + P(TH) = \frac{1}{2}$$

$$P(AB) = \frac{1}{4}, \;\; P(AC) = P(HT) = \frac{1}{4}, \;\; P(BC) = P(TH) = \frac{1}{4}.$$

It follows that the events A, B, C are pairwise independent, which satisfies the conditions for both Questions 1 and 2. However,

$$B \cup C = \{HH, HT, TH\}, \quad A \cap (B \cup C) = \{HH, HT\}, \quad A \cap B \cap C = \emptyset.$$

From here, one can easily verify that A and $B \cup C$ are not independent, neither are $\{A, B, C\}$. This example shows that mathematically defined "independence" is weaker than the common sense "independence".

Exercise 1.16. Let A and B be independent with $P(A) = P(B) = 1/2$. Let $C \doteq AB^c \cup A^cB$. Show that $\{A, B, C\}$ are pairwise independent, but A and $B \cup C$ are not independent, neither are $\{A, B, C\}$. This is a generalization of the counterexample in Example 1.15. *Hint: use Lemma 1.3.*

1.10 Law of Total Probability and Bayes' Rule

We have used *law of total probability* in Examples 1.4, 1.5, and 1.6. There is nothing mysterious about it. The law can be easily visualized as "summing up branches in a tree".

Here is a general version. Let $\{B_1, B_2, \ldots, B_n\}$ be a *partition* of the sample space Ω. By partition, we mean that all these B_i's are disjoint and $B_1 \cup B_2 \cup \cdots \cup B_n = \Omega$. In other words, one and only one event in the collection $\{B_1, B_2, \ldots, B_n\}$ must happen. Let A be an arbitrary event.

Law of Total Probability.

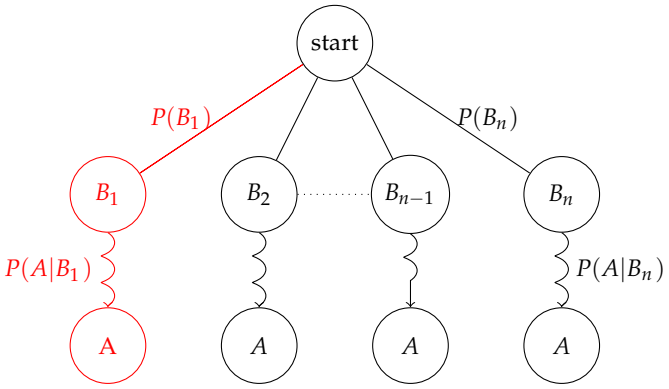
$$P(A) = \sum_{k=1}^{n} P(A \cap B_k) = \sum_{k=1}^{n} P(A \mid B_k) P(B_k).$$

Bayes' Rule. For all $i = 1, 2, \ldots, n$,

$$P(B_i \mid A) = \frac{P(A \cap B_i)}{P(A)} = \frac{P(A \mid B_i) P(B_i)}{\sum_{k=1}^{n} P(A \mid B_k) P(B_k)}.$$

The proof of these two results are trivial. In particular, Bayes' rule follows directly from the definition of conditional probability and law of total probability. As for law of total probability, it is simply the finite additivity of probability, when one realizes that $\{A \cap B_k : \; k = 1, 2, \ldots, n\}$ are disjoint and their union is A.

It is much more intuitive and convenient to represent law of total probability and Bayes' rule in a tree diagram. Each path or branch represents $B_i \cap A$ for some i, and its probability is the product of the probabilities along the branches (multiplication rule). When we sum up the probabilities of all paths or branches, it gives us the probability of A (law of total probability). If we are trying to compute the conditional probability (say) $P(B_1|A)$, it will be the probability of the red path divided by the sum of all the paths (Bayes' rule). We encourage you to figure out what partition $\{B_1, B_2, \ldots, B_n\}$ is used in Examples 1.4, 1.5, and 1.6, respectively.



Example 1.17. A drunkard randomly removes two letters in the message "HAPPY HOUR" that is attached on a billboard in a pub. His drunken friend puts the two letters back in a random order. What is the probability that "HAPPY HOUR" appears again?

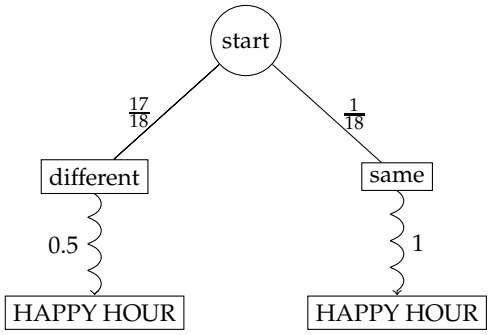
Solution: This can be done by law of total probability. We will divide the sample space up according to if the letters removed are the same or not.

$$\begin{aligned} B_1 &= \{\text{drunkard removes two different letters}\}, \\ B_2 &= \{\text{drunkard removes two identical letters}\}. \end{aligned}$$

Given that B_1 happens, his friend has 50% chance to put them back correctly. Given that B_2 happens, the chance is 100%. Note that only letter "H" and "P" appear twice in the message. Thus

$$P(B_2) = \frac{2}{\binom{9}{2}} = \frac{1}{18}, \quad P(B_1) = 1 - P(B_2) = \frac{17}{18}.$$

We have the following tree.



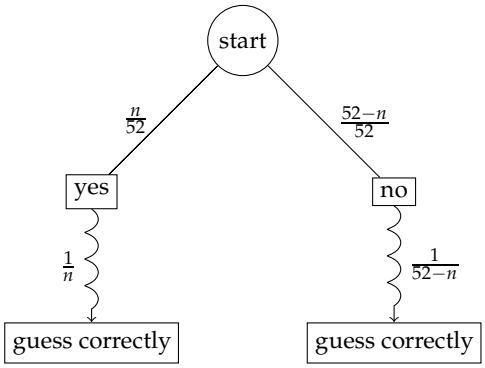
By law of total probability (summing up the branches), the probability of interest equals

$$\frac{17}{18} \cdot \frac{1}{2} + \frac{1}{18} \cdot 1 = \frac{19}{36}.$$

Example 1.18. Your friend has chosen at random a card from a standard deck of 52 cards. You have to guess the card. Before doing so, you are allowed to ask your friend a "yes/no" question, and your friend must answer truthfully. For example, you can ask questions such as "is the card red", or "is this card one of ace of spades, queen of hearts, and two of clubs". What question should you ask to maximize your chance of guessing the correct card?

Solution: The correct answer is that any reasonable question will do. By "reasonable" we mean questions that are relevant to cards. Not questions such as "is the weather good today", which yield no information on the card whatsoever. The key to solving this problem is to realize that any "yes/no" question essentially amounts to a partition of the sample space of 52 cards into two pieces, one piece corresponding to the answer "yes", the other "no".

Now let us assume that your question will leave n cards in the "yes" and $52 - n$ cards in the "no". We will have the following tree.



By law of total probability, your overall chance of guessing the card correctly is

$$\frac{n}{52} \cdot \frac{1}{n} + \frac{52-n}{52} \cdot \frac{1}{52-n} = \frac{1}{26}$$

regardless of the value of n. That is, all reasonable questions will double your chance from $1/52$ to $1/26$.

Example 1.19. This is a variant of the Monty Hall Problem. There are three identical cards, one red on both faces (RR), one green on both faces (GG), and one red on one face and green on the other (GR). A card is randomly drawn and you are shown at random one face of the card. Given that you see the color red, what is the probability that the other face of the card is also red?

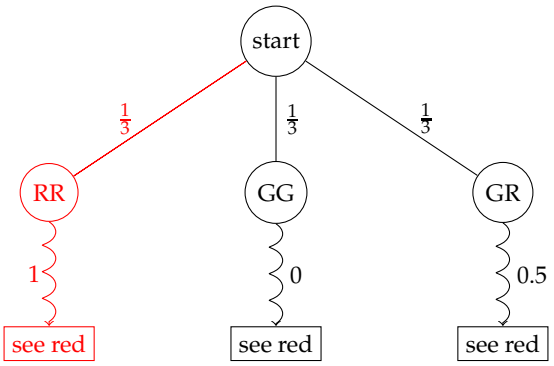
Solution: The question is asking the conditional probability of the card being RR given that you see a red face of the card. Once you see red, you know there are only two possibilities: the card is either RR or GR. Thus it is rather tempting (but wrong) to say the probability is 50% for the card to be RR. Let us use Bayes' rule to find the correct conditional probability. Define

$$\begin{aligned} B_1 &= \{\text{card is RR}\} \\ B_2 &= \{\text{card is GG}\} \\ B_3 &= \{\text{card is GR}\} \\ A &= \{\text{you see a red face}\}. \end{aligned}$$

Then question is asking $P(B_1|A)$. See the following tree. By Bayes' rule, the

conditional probability is

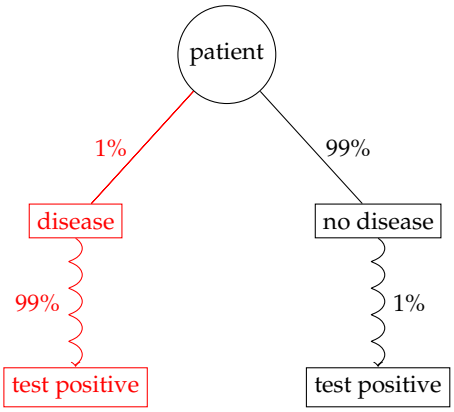
$$\begin{aligned} P(B_1|A) &= \frac{P(B_1 \cap A)}{P(A)} \\ &= \frac{P(B_1)P(A|B_1)}{P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + P(B_3)P(A|B_3)} \\ &= \frac{\frac{1}{3} \cdot 1}{\frac{1}{3} \cdot 1 + \frac{1}{3} \cdot 0 + \frac{1}{3} \cdot 0.5} = \frac{2}{3}. \end{aligned}$$



Example 1.20. Bayes' rule can explain why many medical tests have a high "false positive" rate. For example, suppose 1% of a population are contracted with a certain type of disease. There is a medical test for this disease that is 99% accurate. That is, 99% of the people who have the disease will test positive and 99% of the people who do not have the disease will test negative. If a patient is tested positive, what is the probability that the patient is actually contracted with the disease?

Solution: Define $B = \{\text{patient has disease}\}$ and $A = \{\text{patient tested positive}\}$. The question is asking for $P(B|A)$. By Bayes' rule,

$$\begin{aligned} P(B|A) &= \frac{P(A \cap B)}{P(A)} = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)} \\ &= \frac{0.99 \times 0.01}{0.99 \times 0.01 + 0.01 \times 0.99} = 50\%. \end{aligned}$$



To explain the high false positive rate, imagine that we test everyone in the population for the disease. Some of those who have the disease will test positive, and some of those who do not have the disease will also test positive. Even though for the latter the chance of testing positive is rather low, the population without the disease is much larger compared to the population with the disease (in our example, it is a whopping 99:1 ratio). Consequently, lots of the positives will actually come from those who do not have the disease, especially if the disease is rare. Thus, if someone is tested positive, it very likely comes from a patient who does not have the disease.

Chapter 2

Preface on Random Variables

Before we start a systemic study of random variables, let us give a brief overview.

Let Ω be a sample space. A **random variable** is a function defined on Ω. A random variable $X : \Omega \to \mathbb{R}$ associates each possible outcome or sample point $\omega \in \Omega$ with a numerical value $X(\omega) \in \mathbb{R}$. When there are multiple random variables involved, it is *always* assumed that they are defined on the *same* sample space Ω. With the introduction of random variables, one can focus on important aspects of the outcomes, without worrying about the details of the sample space. For example, in stochastic modeling it is often the case that assumptions and conditions are imposed upon relevant random variables directly, instead of the underlying sample space.

In the dealing of random variables, *the dependence on ω is often suppressed* for notational convenience. We list a few examples of abbreviations.

1. $\{X \leq 5\}$: this denotes the event $\{\omega : X(\omega) \leq 5\}$.
2. $P(X \geq 1)$: this denotes the probability $P(\{\omega : X(\omega) \geq 1\})$.
3. $X + 3Y$: when there are multiple random variables (say) X and Y, algebraic operation such as $X + 3Y$ defines a new random variable
$$X + 3Y : \ \omega \mapsto X(\omega) + 3Y(\omega).$$
4. For any function $h : \mathbb{R} \to \mathbb{R}$, $h(X)$ denotes the composition $h \circ X : \Omega \to \mathbb{R}$, which is also a random variable. That is,
$$h(X) : \omega \mapsto (h \circ X)(\omega) = h(X(\omega)).$$
For instance, if $h(x) = x^2$, then $h(X) = X^2$; if $h(x) = (x - a)^+$, then $h(X) = (X - a)^+$, and so on.

2.1 Distributions of Random Variables

In probability and statistics, you will often hear people say things such as "X is a *standard normal* random variable", or "Y is *Poisson* with parameter ...", and so on. What these names stand for is the *distribution* of a random variable. So what is distribution?

The **distribution** of a random variable is a common terminology referring to the detailed probabilistic properties of the random variable. Knowing the distribution amounts to knowing the probability of this random variable taking values in *any* range. This explanation still seems a bit vague. Let us use some examples to illustrate this concept.

Example 2.1. If we are told that a random variable X has *Poisson* distribution with parameter 3, then we know precisely what the distribution of X is, that is, X can only take values of nonnegative integers $\{0, 1, 2, \ldots\}$ with the *probability mass function* (pmf) given by

$$p_X(x) \doteq P(X = x) = e^{-3} \cdot \frac{3^x}{x!}, \quad x = 0, 1, 2, \ldots$$

With the probability mass function $p_X(x)$, we can compute the probability of X falling into any range. For instance, take an arbitrary interval (say) $[1.2, 3]$, the probability that X falls into that interval is simply

$$P(1.2 \leq X \leq 3) = \sum_{1.2 \leq x \leq 3} p_X(x),$$

which equals $P(X = 2) + P(X = 3)$. With the probability mass function, we can also compute important quantities associated with this random variable, such as *mean* and *variance*.

Example 2.2. If X is said to be a *standard normal* or have the the standard normal distribution, then we know precisely what the distribution of X is, that is, X can take any real values, with the *probability density function* (pdf) given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad -\infty < x < \infty.$$

As of now, we can intuitively regard the probability density function as a curve of "relative likelihood". Therefore, it is more likely for X to be closer to 0 than away from it. Again, with the probability density function $f_X(x)$, we can compute the probability of X falling into any range. For instance,

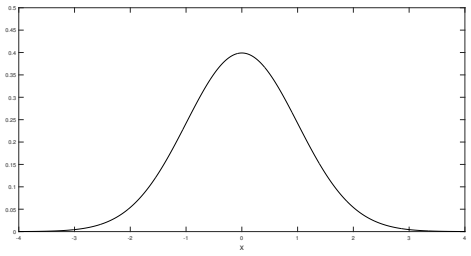


Figure 2.1: Probability density function of $N(0, 1)$.

take an arbitrary interval (say) $[1.2, 3]$, the probability that X falls into that interval is simply

$$P(1.2 \leq X \leq 3) = \int_{1.2}^{3} f_X(x)\, dx.$$

With the probability density function, we can also compute important quantities associated with this random variable, such as *mean* and *variance*. ■

The Poisson and normal distributions in Examples 2.1 and 2.2 belong to the family of special distributions. But most random variables and their distributions do not belong to this family. This is not a problem at all. We can still use probability mass function or probability density function to describe the distribution of a random variable. In this sense, one can also equate the distribution of a random variable with its probability mass function or probability density function.

2.2 Different Types of Random Variables

The Poisson random variable in Example 2.1 is very different than the standard normal random variable in Example 2.2. The former can only take discrete values, while the latter can take a continuum of possible values. For this reason, random variables such as Poisson are said to be **discrete**, while random variables such as standard normal are said to be **continuous**. This distinction is also the reason that we use probability mass functions for the former and probability density functions for the latter, to describe their distributions, respectively.

There is a universal measure-theoretic approach to treat all random variables and study their numerical properties simultaneously, be it discrete or continuous. However, due to its higher technical requirements,

such an approach is often replaced by a more pragmatic treatment, where we have to treat discrete and continuous random variables separately. It is inconvenient and incomplete, but we will have to make do.

Remark 2.1. We should mention that there are random variables that are neither discrete or continuous. Consider the following example. You toss a fair coin once. If heads comes up, you receive 40 cents. If tails comes up, a number is chosen at random from interval $[0, 1]$ and you receive the dollar amount equal to that number. Let X be the amount you receive (in dollars). Then X has a 50% chance to be 0.4 and 50% chance to be a random number on interval $[0, 1]$. This random variable is sort of in the middle. It is not discrete because it can take a continuum of values. Neither is it continuous because it can take a specific value with positive probability, namely, $P(X = 0.4) = 0.5$. One cannot use a probability mass function nor a probability density function to describe its distribution.

2.3 Cumulative Distribution Function

Before we start the discussion on discrete and continuous random variables from the next chapter, we would like to introduce a universal concept that works for all random variables. Given a random variable $X : \Omega \to \mathbb{R}$, its **cumulative distribution function (cdf)** is defined to be a function $F : \mathbb{R} \to [0, 1]$ such that for $x \in \mathbb{R}$,

$$F(x) \doteq P(X \leq x).$$

There are a few properties regarding cumulative distribution functions:

1. F is nondecreasing with $0 \leq F(x) \leq 1$.
2. $F(-\infty) = 0$ and $F(\infty) = 1$.
3. F is continuous from the right, that is, $\lim_{x_n \downarrow x} F(x_n) = F(x)$ for all x.

Property 1 and 2 are rather obvious. As for Property 3, intuitively it should be correct since events $\{X \leq x_n\} \downarrow \{X \leq x\}$ when $x_n \downarrow x$. But if one wants to be meticulous, a proof is necessary. We leave it as an exercise to interested students; see Exercise 2.4.

Cumulative distribution functions are important in a couple of ways. First of all, it can be defined for any random variables, be it discrete, continuous, or of any other kinds. Secondly, it determines the probability of

X within any range. That is, *cumulative distribution function completely characterizes the distribution of a random variable*, which in turn determines the important numerical characteristics of a random variable such as expected value, variance, and so on. In this sense, one can also equate the distribution of any random variable with its cumulative distribution function.

This also implies that there must be a one-to-one correspondence between cumulative distribution functions and probability mass (resp. density) functions for discrete (resp. continuous) random variables. We should discuss their relations in details in future chapters.

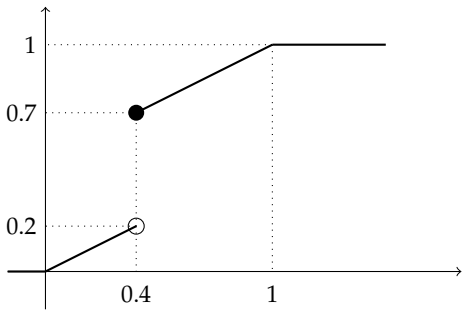
Example 2.3. Let us revisit the random variable X in Remark 2.1 and determine its cumulative distribution function. Recall that you toss a fair coin. If it is heads, then X takes the value 0.4; if it is tails, X takes the value of a random number from interval $[0, 1]$.

Note that X can only take values in $[0, 1]$ regardless. Therefore, it is trivial that $F(x) = P(X \leq x)$ should equal 0 if $x < 0$ and 1 if $x > 1$. Now take any $x \in [0, 1]$. By the law of total probability, we have

$$\begin{aligned} P(X \leq x) &= P(X \leq x \mid \text{heads})P(\text{heads}) + P(X \leq x \mid \text{tails})P(\text{tails}) \\ &= 0.5P(X \leq x \mid \text{heads}) + 0.5P(X \leq x \mid \text{tails}). \end{aligned}$$

The first conditional probability on the right-hand-side is 1 if $x \geq 0.4$ and 0 if $x < 0.4$, because $X = 0.4$ given heads. The second conditional probability should be x (see also Example 1.7). Therefore the cumulative distribution function is

$$F(x) = P(X \leq x) = \begin{cases} 0 & \text{if } x \leq 0 \\ 0.5x & \text{if } 0 \leq x < 0.4 \\ 0.5 + 0.5x & \text{if } 0.4 \leq x \leq 1 \\ 1 & \text{if } x \geq 1 \end{cases}.$$



This function is continuous at all points except at $x = 0.4$, where it makes a jump of size $P(X = 0.4) = P(\text{heads}) = 0.5$. You can verify that $F(x)$ satisfies all three properties mentioned at the beginning of this section: nondecreasing, right-continuous, and always take values in $[0, 1]$. ■

Exercise 2.4. *(continuity from above)* Let $\{B_1, B_2, \ldots\}$ be a decreasing sequence of events. That is, $B_1 \supseteq B_2 \supseteq \cdots$. Let $B = B_1 \cap B_2 \cap \cdots$. Use the axioms of probability to show that

$$P(B) = \lim_{n \to \infty} P(B_n).$$

Exercise 2.5. *(continuity from below)* Let $\{B_1, B_2, \ldots\}$ be an increasing sequence of events. That is, $B_1 \subseteq B_2 \subseteq \cdots$. Let $B = B_1 \cup B_2 \cup \cdots$. Use the axioms of probability to show that

$$P(B) = \lim_{n \to \infty} P(B_n).$$

Chapter 3

Discrete Random Variables

A random variable X is said to be **discrete** if X takes values in a finite or infinite sequence $\{x_1, x_2, \ldots\} \subseteq \mathbb{R}$. Without loss of generality, we assume all these x_i's are distinct. The distribution of X is completely determined by its **probability mass function (pmf)** $p : \{x_1, x_2, \ldots\} \to [0, 1]$ with

$$p(x_i) = P(X = x_i),$$

which must satisfy

$$p(x_i) \geq 0, \quad \sum_{x_i} p(x_i) = 1.$$

The **cumulative distribution function (cdf)** $F : \mathbb{R} \to [0, 1]$ of X, defined by

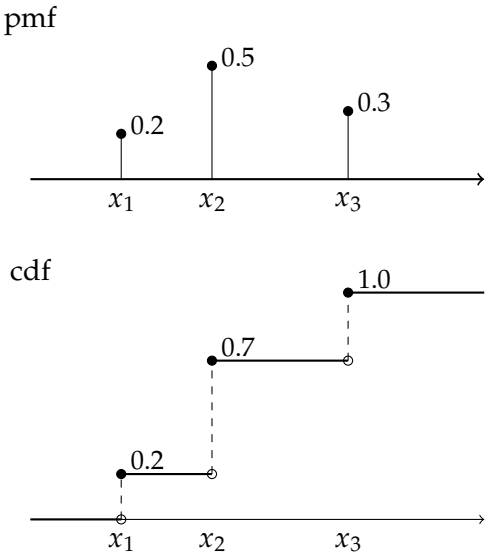
$$F(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i) = \sum_{x_i \leq x} p(x_i)$$

is a step function that makes jumps at points $\{x_i\}$ with jump sizes $\{p(x_i)\}$, respectively. It is not difficult to see that the probability mass function and cumulative distribution function uniquely determine each other.

Example 3.1. Consider a discrete random variable that take three possible values $\{x_1, x_2, x_3\}$ in ascending order, with the respective probability $\{0.2, 0.5, 0.3\}$. The corresponding probability mass function and cumulative distribution function are given in the following figure. The latter is a step function, right-continuous, where jumps occur at x_1, x_2, x_3 with jump size $p(x_1)$, $p(x_2)$, $p(x_3)$, respectively. We would also like to take this chance to introduce a useful notation:

$$F(x) = 0.2 \cdot 1_{(x_1, x_2]}(x) + 0.7 \cdot 1_{(x_2, x_3]}(x) + 1_{[x_3, \infty)}(x)$$

where those **indicator functions** $1_E(x)$ take value 1 if $x \in E$ and 0 otherwise.



Definition 3.1. Let X be a discrete random variable that takes values in $\{x_1, x_2, \ldots\}$. Then its **expected value** (or **expectation**, **mean**), **variance**, and **standard deviation** are defined as

$$\begin{aligned} E[X] &\doteq \sum_{x_i} x_i P(X = x_i) \\ \mathrm{Var}[X] &\doteq E[(X - E[X])^2] = \sum_{x_i} (x_i - E[X])^2 P(X = x_i) \\ \mathrm{Std}[X] &\doteq \sqrt{\mathrm{Var}[X]} \end{aligned}$$

The expected value or mean of a random variable is the average of all possible values weighted by their respective probabilities. Variance measures the "spread" or "variation" of a random variable. The larger the variance, the more variation a random variable can have. When variance is close to zero, then the random variable will mostly take values in a tight range and stay close to its own mean. It is the opposite when the variance gets larger. For example, comparing the two probabilities mass functions in the following figure. The one on the right has a smaller variance than the one on the left, since its range is visibly tighter.

There are other ways to measure the variation of a random variable. For example, $E|X - E[X]|$ measures the average absolution deviation from the mean, which is another very natural way to define variation. However, there are at least two important reasons that people have adopted the current definition of variance: (1) It is mathematically much easier to work with squares than other functions such as absolute values; (2) Arguably the most important result in probability theory, *central limit theorem*, needs two parameters: expected values and variance (from this definition).

Lemma 3.1. *Let X and Y be any discrete random variables. Let a, b be any constants and $h : \mathbb{R} \to \mathbb{R}$ a given function. Then*

1. $E[aX + b] = aE[X] + b$
2. $E[aX + bY] = aE[X] + bE[Y]$
3. $\mathrm{Var}[X] = E[X^2] - E^2[X]$
4. $E[h(X)] = \sum_{x_i} h(x_i)P(X = x_i)$, *if X takes values in $\{x_1, x_2, \ldots\}$*
5. $\mathrm{Var}[aX + b] = a^2\mathrm{Var}[X]$
6. $\mathrm{Var}[X] = 0$ *if and only if $X = c$ for some constant c.*

Proof. Throughout the proof we assume that X takes values in $\{x_1, x_2, \ldots\}$ and Y takes values in $\{y_1, y_2, \ldots\}$.

1. The identity is trivial when $a = 0$. Assume now $a \neq 0$. Then $aX + b$ is a discrete random variable that takes values in $\{ax_1 + b, ax_2 + b, \ldots\}$, with respective probability $P(aX + b = ax_i + b) = P(X = x_i)$. Therefore, by definition

$$\begin{aligned} E[aX + b] &= \sum_{x_i} (ax_i + b)P(X = x_i) \\ &= a\sum_{x_i} x_i P(X = x_i) + b\sum_{x_i} P(X = x_i) \\ &= aE[X] + b. \end{aligned}$$

2. Define $Z = aX + bY$. Then Z is a discrete random variables that takes values in $\{z_1, z_2, \ldots\}$ where each z_k is equal to $ax_i + by_j$ for some i and j, and

$$P(Z = z_k) = \sum_{(x_i, y_j):\, ax_i + by_j = z_k} P(X = x_i,\, Y = y_j)$$

It follows that (by exchanging the order of summation in some steps)

$$\begin{aligned}
E[Z] &= \sum_{z_k} z_k P(Z = z_k) \\
&= \sum_{z_k} z_k \sum_{(x_i,y_j):\, ax_i+by_j=z_k} P(X = x_i,\, Y = y_j) \\
&= \sum_{z_k} \sum_{(x_i,y_j):\, ax_i+by_j=z_k} (ax_i + by_j) P(X = x_i,\, Y = y_j) \\
&= \sum_{(x_i,y_j)} (ax_i + by_j) P(X = x_i,\, Y = y_j) \\
&= \sum_{x_i} \sum_{y_j} ax_i P(X = x_i,\, Y = y_j) + \sum_{y_j} \sum_{x_i} by_j P(X = x_i,\, Y = y_j) \\
&= \sum_{x_i} ax_i P(X = x_i) + \sum_{y_j} by_j P(Y = y_j) \\
&= aE[X] + bE[Y].
\end{aligned}$$

The first three equalities in last display is to overcome the little hassle that $ax_i + bx_j$ may take the same value for different (i, j)'s.

3. Note that $(X - E[X])^2 = X^2 - 2E[X]X + E^2[X]$. Then by definition of variance and Parts 1 and 2, we have

$$\mathrm{Var}[X] = E[X^2] - 2E[X]E[X] + E^2[X] = E[X^2] - E^2[X].$$

4. The proof is similar to that of Part 2. Once we observe that $Z = h(X)$ is a discrete random variable that takes values in $\{z_1, z_2, \ldots\}$ with each $z_k = h(x_i)$ for some i, the rest of the proof is very similar and thus omitted.

5. Denote $Z = aX + b$. Then $E[Z] = aE[X] + b$ from Part 1, and $Z - E[Z] = a(X - E[X])$. The claim follows from the definition of variance and Part 1.

6. If $X = c$ for some constant c, then it is trivial from definition of variance that $\mathrm{Var}[X] = 0$. For the other direction, note that by definition $\mathrm{Var}[X] = 0$ if and only if $(x_i - E[X])^2 = 0$ for all i such that $P(X = x_i) > 0$. Therefore, $x_i = E[X]$ for all such i. We complete the proof. ■

Example 3.2. An insurance policy pays \$100 per day for up to two days of hospitalization and \$50 per day for each day of hospitalization thereafter.

The number of days of hospitalization, X, is a discrete random variable with probability mass function

$$p_X(x) = \frac{5-x}{10} \cdot 1_{x\in\{1,2,3,4\}}$$

Determine the mean and standard deviation of the payment of hospitalization under the insurance policy.

Solution: For this problem, it is probably easiest to express the probability mass function, as well as the payment using a table. Note that the payment Y is indeed a function of X.

x	1	2	3	4
$p_X(x)$	0.4	0.3	0.2	0.1
Payment y	100	200	250	300

$$E[Y] = 0.4 \cdot 100 + 0.3 \cdot 200 + 0.2 \cdot 250 + 0.1 \cdot 300 = 180$$
$$E[Y^2] = 0.4 \cdot 100^2 + 0.3 \cdot 200^2 + 0.2 \cdot 250^2 + 0.1 \cdot 300^2 = 37500$$
$$\text{Var}[Y] = E[Y^2] - E^2[Y] = 37500 - (180)^2 = 5100.$$

In other words, the expectation and standard deviation of the payment are \$180 and $\$\sqrt{5100} \approx \$71.4$, respectively. ■

Example 3.3. A company has a rule regarding holidays. Every employee of the company can take a day off if it happens to be the birthday of an employee. There are no other holidays. Every employee works 8 hours on every working day. How many employees should this company hire to maximize the expected total number of working hours from all employees during a whole year? For simplicity we assume that there are 365 days in a year every year.

Solution: Suppose that the company will hire n employees. We will use *indicator random variables* to analyze the number of holidays. Define for each $1 \le k \le 365$

$$X_k = \begin{cases} 1 & \text{if Day } k \text{ is } \textit{not} \text{ the birthday of any employee} \\ 0 & \text{otherwise} \end{cases}.$$

The total number of working days in a year is

$$X = \sum_{k=1}^{365} X_k,$$

and the total number of working hours in a year is $8nX$ (n employees, 8 hours a day, X working days). Furthermore,

$$\begin{aligned} P(X_k = 1) &= P(\text{none of } n \text{ employees have birthday on Day } k) \\ &= \prod_{i=1}^{n} P(\text{Employee } i\text{'s birthday is not Day } k) \\ &= \left(\frac{364}{365}\right)^n. \end{aligned}$$

Therefore, the expected total number of working hours (denoted by t_n) is

$$t_n = E[8nX] = 8n \sum_{k=1}^{n} E[X_k] = 8n \sum_{k=1}^{n} P(X_k = 1) = 8n \left(\frac{364}{365}\right)^n.$$

In order to find the number n that maximizes the expected total number of working hours T_n, we observe that

$$\frac{t_{n+1}}{t_n} = \frac{n+1}{n} \times \frac{364}{365},$$

which implies that

$$t_1 < t_2 < \cdots < t_{364} = t_{365} > t_{366} > t_{367} > \cdots.$$

Therefore, to maximize the expected total number of working hours the company should hire either 364 or 365 employees.

Example 3.4. Suppose X is a nonnegative random variable taking values on nonnegative integers. Show that

$$E[X] = \sum_{n=1}^{\infty} P(X \geq n).$$

Proof. We introduce *indicator random variable* of form 1_A for any event $A \subseteq \Omega$. It is defined such that $1_A(\omega)$ takes value 1 if $\omega \in A$ and 0 otherwise. It is straightforward by definition that $E[1_A] = P(A)$. Now we observe that

$$X = \sum_{n=1}^{\infty} 1_{\{X \geq n\}}.$$

Indeed, when $X = k$ for some nonnegative integer k, the right-hand-side equals

$$\sum_{n=1}^{k} 1 = k = X.$$

It follows that

$$E[X] = E\left[\sum_{n=1}^{\infty} 1_{\{X \geq n\}}\right] = \sum_{n=1}^{\infty} E\left[1_{\{X \geq n\}}\right] = \sum_{n=1}^{\infty} P(X \geq n).$$

We complete the proof. ■

Exercise 3.5. *(standardization)* Let X be a random variable with mean μ and variance σ^2. Define

$$Y = \frac{X - \mu}{\sigma}$$

Show that $E[Y] = 0$ and $\mathrm{Var}[Y] = 1$. This random variable Y is said to be the **standardization** of X.

Exercise 3.6. Given a random variable X, define a *square loss function* $L(a) \doteq E[(X-a)^2]$ for every $a \in \mathbb{R}$. Show that L is minimized at $a = E[X]$.

3.1 Common Discrete Random Variables

In this section we collect some very commonly used discrete random variables.

Definition 3.2. Throughout the following definitions, $0 < p < 1$ is assumed to be a given constant.

1. A random variable X is said to be **Bernoulli with parameter** p if X takes values in $\{0, 1\}$ with

$$P(X = 1) = p, \quad P(X = 0) = 1 - p.$$

2. A random variable X is said to be **binomial with parameters** (n, p), denoted by $B(n, p)$, if X takes values in $\{0, 1, \ldots, n\}$ with

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \ldots, n.$$

   Bernoulli distribution is a special case of binomial with $n = 1$.

3. A random variable X is said to be **geometric with parameter** p, if X takes values in $\{1, 2, 3, \ldots\}$ with

$$P(X = n) = (1-p)^{n-1} p, \quad n = 1, 2, 3, \ldots.$$

4. A random variable X is said to be **Poisson with parameter** $\lambda > 0$, if X takes values in $\{0, 1, 2, \ldots\}$ with

$$P(X = n) = e^{-\lambda}\frac{\lambda^n}{n!}, \quad n = 0, 1, 2, \ldots.$$

Motivation for several of these random variables involves the so called *binomial experiments*. These are identical and independent trials that have only two outcomes: "success" and "failure". Assume that the probability of "success" for each trial is p.

1. The total number of "success" in a single trial is Bernoulli with parameter p; this is trivial.

2. The total number of "success" in n trials is binomial with parameters (n, p), i.e., $B(n, p)$; this is because there are in total

$$\binom{n}{k}$$

ways for the k "success" to appear in the n trials. Each one these outcome has the identical probability $p^k(1-p)^{n-k}$.

3. The number of trials until the first "success" is geometric with parameter p; this is because the first "success" happens in the n-th trial if and only if the first $(n-1)$ trials are "failure" and the n-th trial is a "success". The probability for such a sequence is $(1-p)^{n-1}p$.

Example 3.7. *(Poisson as a limit of Binomial)* Suppose that we are interested in the total number of emails we receive during an hour, a day, or any amount of time. To model these numbers, which are obviously random, we impose two assumptions:

1. numbers of emails received in non-overlapping time intervals are independent;

2. number of emails received in a *very short* time interval of length Δt is approximately Bernoulli with parameter $\lambda\Delta t$ for some positive constant λ.

Let X be the number of emails received during time interval $[0, 1]$. To find out the distribution of X, we can divide the time interval $[0, 1]$ into n subintervals of equal length. Since $\Delta t = 1/n$, the number of emails on each

subinterval is one ("success") with probability $p_n = \lambda\Delta t = \lambda/n$ and zero ("failure") with probability $1 - p_n = 1 - \lambda/n$, approximately. The total number of emails on interval $[0, 1]$ can be viewed as the total number of "success" in n trials (subintervals). Therefore, X should be approximately $B(n, p_n)$ where $p_n = \lambda\Delta t = \lambda/n$, and the approximation should be more accurate as n gets larger. This implies that for any $k = 0, 1, 2, \ldots$,

$$P(X = k) = \lim_{n\to\infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} = e^{-\lambda}\frac{\lambda^k}{k!},$$

here we have omitted details of the algebra and recalled the classical formula for Euler's number e:

$$\lim_{x\to\infty} \left(1 + \frac{1}{x}\right)^x = e,$$

This argument explains why Poisson can be viewed as a limit of binomial. It is also for this reason that the parameter λ is referred to as the "arrival rate" in many counting processes. ■

Lemma 3.2. *Here is the table of expected values and variances for these commonly used random variables.*

X	$E[X]$	$\mathrm{Var}[X]$
Bernoulli with parameter p	p	$p(1-p)$
Binomial with parameters (n, p)	np	$np(1-p)$
Geometric with parameter p	$1/p$	$(1-p)/p^2$
Poisson with parameter λ	λ	λ

Proof. The argument for Bernoulli random variables is trivial. However, We defer the complete proof to Appendix A, using *moment generating functions* for a comprehensive treatment. ■

Example 3.8. Toss a fair die n times. Let X denote the total number of 6's. What is the distribution of X? What is $P(X \geq 1)$?

Solution: It suffices to define that getting a six is "success". Then clearly X is $B(n, 1/6)$. Thus

$$P(X \geq 1) = 1 - P(X = 0) = 1 - \left(1 - \frac{1}{6}\right)^n = 1 - \left(\frac{5}{6}\right)^n.$$

Example 3.9. For each setup, specify the distributions of the relevant random variables.

1. Consider a binomial experiment with n independent and identical trials and assume that the probability of "success" for each trial is p. Let X denote the total number of "success" and Y that of "failure" from these n trials.

2. Two fair dice (labeled 1 and 2) are tossed together n times. Let X denote the total number of tosses where the face value from Dice 1 is less than that from Dice 2.

Solution: For Part 1, it is immediate that X is $B(n, p)$. In order to find the distribution of $Y = n - X$, it is useful to observe that the definition of "success" and "failure" is relative. We can define the old "failure" as the *new* "success" and the old "success" as the *new* "failure". Therefore, Y is $B(n, 1-p)$.

As for Part 2, we can define each trial as a toss of the two dice, and define the outcome as a "success" if the face value of Dice 1 is less than that of Dice 2. There are in total $6 \times 6 = 36$ outcomes when two dices are tossed, among which $1 + 2 + 3 + 4 + 5 = 15$ of them are "success" (number of pairs i and j with $1 \leq i < j \leq 6$). Thus $P(\text{success}) = 15/36 = 5/12$, and consequently X is $B(n, 5/12)$. ■

Example 3.10. *(Binomial or not Binomial)* Consider the following two scenarios in random sampling.

1. 35% of teenagers in USA consider Instagram to be the most important social network. Randomly select 5 teenagers. Let X denote the number of those that regard Instagram as most important. What is the distribution of X?

2. 35% of the members of a large of family of size 20 consider Instagram to be the most important social network. Randomly select 5 members from this family. Let Y denote the number of those that regard Instagram as most important. What is the distribution of Y?

Solution: For the first question, the answer is $B(5, 0.35)$. One can regard each teenager represents a "trial": if he/she regards Instagram as most important, then the outcome is a "success"; otherwise it is a "failure". Each trial has a chance 0.35 to be a success. Thus X, the total number of "success", should be binomial $B(5, 0.35)$.

It seems that if we apply the same argument to the second question, then we should also conclude that Y is $B(5, 0.35)$. But this is incorrect. The key observation is that these "trials" are *not* independent. The outcomes of trials can affect one another. For example, if the first trial is a success

(regards Instagram as most important), the probability of the second trial being a success is not 0.35, but

$$\frac{7-1}{20-1},$$

since there are in total $20 \times 35\% = 7$ members in the family that regard Instagram as most important.

But now you will ask how come in the first question, we can regard X as binomial? Wouldn't the same argument also imply that X should not be binomial? In the most strict sense, this statement of X not being binomial is actually not wrong. Suppose the population size of teenagers in America is N. Then given the first trial is a success, the second trial is a success with probability

$$\frac{0.35N-1}{N-1}.$$

But because N is so large, this probability, in all practicality, is essentially 0.35 again. In other words, it is safe to say all the trials in the first question are practically independent. Thus X is binomial.

This example is to show that in practice, samples from random sampling are assumed to be independent, because the population size is magnitudes larger than the sample size. If this condition fails, then the independence assumption cannot be employed. ■

Example 3.11. The trial in the definition of binomial experiments can be quite complicated. For example, can we use a coin, whose probability of heads is unknown to us, to generate a $B(n, 0.5)$ random variable?

Solution: The key for solving this interview question is to build a trial with 50% probability of "success". If we can do that, then we simply perform this trial n times and count the total number of "success".

Here is one way to build such a trial. A *single* trial consists of tossing the coin repeatedly, twice at a time. If the outcome is HH or TT, we continue. We only stop when the outcome is either HT or TH. If the outcome is HT, then the trial ends with a "success". If the outcome is TH, then the trial ends with a "failure". For instance, a single trial could look like this

$$(HH),\ (TT),\ (TT),\ (HT),$$

which stops at the 8th toss and ends with a "success". In contrast, it may also look like this

$$(TT),\ (HH),\ (TH),$$

where the trial ends in the 6th toss as a "failure".

The probability of "success" for such a trial equals the *conditional probability* that, given the outcome of two tosses is HT or TH, it is indeed HT. Note that $P(HT) = P(TH) = P(H)P(T)$. Therefore,

$$P(\text{success}) = P(HT \,|\, HT \text{ or } TH) = \frac{P(HT)}{P(HT) + P(TH)} = \frac{1}{2}.$$

Example 3.12. Repeatedly toss two coins, each with $P(H) = p$, simultaneously. Let X be the number of tosses until one coin shows heads and the other shows tails. What is the distribution of X? What is $P(X > n)$?

Solution: Regard each toss as a trial, with success defined as HT or TH. Then X is the number of trials until the first success. Since

$$P(\text{success}) = P(HT) + P(TH) = 2p(1-p),$$

X is geometric with parameter $2p(1-p)$. The event $\{X > n\}$ amounts to that the first n trials are all failures. Thus

$$P(X > n) = [1 - P(\text{success})]^n = [1 - 2p(1-p)]^n.$$

Example 3.13. Suppose that the number of lobsters a lobster fishing boat can catch in a day is Poisson distributed with parameter λ. However, some of these lobsters must be put back into the water because of regulation, e.g., lobsters that are too big or too small, or female lobsters that are bearing eggs, and so on. Assume that for each lobster caught it has probability p to be put back into the water (independent of all other lobsters). What is the distribution of the number of lobsters that a lobster fishing boat can catch and keep in a day?

Solution: Let Y be the number of lobsters that a boat can catch in a day, and X that a boat can catch and keep in a day. By assumption, Y is Poisson with parameter λ, and given $Y = k$, X is $B(k, 1-p)$. Therefore, for any $n \geq 0$,

$$\begin{aligned} P(X = n) &= \sum_{k=0}^{\infty} P(X = n, Y = k) = \sum_{k=n}^{\infty} P(X = n, Y = k) \\ &= \sum_{k=n}^{\infty} P(X = n \,|\, Y = k) P(Y = k) \\ &= \sum_{k=n}^{\infty} \binom{k}{n} (1-p)^n p^{k-n} \cdot e^{-\lambda} \frac{\lambda^k}{k!}. \end{aligned}$$

Plugging in the formula for binomial coefficients and reorganizing terms, we have

$$\begin{aligned} P(X=n) &= \frac{e^{-\lambda(1-p)}(1-p)^n\lambda^n}{n!}\sum_{k=n}^{\infty}\frac{e^{-\lambda p}(\lambda p)^{k-n}}{(k-n)!} \\ &= \frac{e^{-\lambda(1-p)}(1-p)^n\lambda^n}{n!}\sum_{m=0}^{\infty}\frac{e^{-\lambda p}(\lambda p)^{m}}{m!} \end{aligned}$$

Note that the infinite sum is actually the sum of the probability mass function of the Poisson distribution with parameter λp, which must equal 1. Therefore,

$$P(X=n) = e^{-\lambda(1-p)}\frac{[\lambda(1-p)]^n}{n!}.$$

That is, the number of lobster a fishing boat and catch and keep in a day is Poisson with parameter $\lambda(1-p)$. ■

Exercise 3.14. (hard) Let X be a discrete random variable taking values on positive integers $\{1, 2, \ldots\}$. Show that X satisfies the memoryless property:

$$P(X > k+n \mid X > k) = P(X > n)$$

for all $n, k \geq 0$ if and only if X is a geometric random variable.

Exercise 3.15. (Putnam 2002, hard) Shanille OKeal shoots free throws on a basketball court. She hits the first and misses the second, and thereafter the probability that she hits the next shot is equal to the proportion of shots she has hit so far. What is the probability she hits exactly 50 of her first 100 shots?

Chapter 4

Continuous Random Variables

Continuous random variables and discrete random variables are on the opposite end of the spectrum. While discrete random variables can only take values in a finite or infinite sequence, continuous random variables can takes values in a continuum, for example, intervals like $[a, b]$ or the whole real line. In basic probability theory, the distributions of continuous random variables are described by probability density functions. More precisely, we say $X : \Omega \to \mathbb{R}$ is a **continuous random variable** if there exists a **probability density function (pdf)**, or simply **density function**, $f : \mathbb{R} \to [0, \infty)$ such that for any set $B \subseteq \mathbb{R}$,

$$P(X \in B) = \int_B f(x)\, dx,$$

and

$$f(x) \geq 0, \qquad \int_{\mathbb{R}} f(x)\, dx = 1.$$

It is immediate that for any continuous random variable the probability of taking a specific value is zero, that is, $P(X = x) = 0$ for any $x \in \mathbb{R}$. This is very different from discrete random variables.

Note that by definition, the probability of X taking values in an interval equals the **area under the probability density function** on that interval. However, the values of a probability density function itself, which can often take values greater than one, *cannot* be directly regarded as probability. There are a few ways to understand the meaning of a probability density function of a continuous random variable in general.

1. *Probability on a small interval.* Consider any point $x \in \mathbb{R}$ and a very small interval around x with length Δx, say $[a, b]$, where $x \in [a, b]$

with $b - a = \Delta x$ very small. Then

$$P(X \in [a, b]) \approx f(x)(b - a) = f(x)\Delta x.$$

This is because the region under the density function f over a small interval $[a, b]$ around x is approximately a rectangle with height $f(x)$ and width $(b - a) = \Delta x$.

2. *Density as relative probability weight.* Even though the values of a probability density function do not directly relate to probabilities, their ratios at different points do give the relative probability weight. To be more precise, consider two points $x, y \in \mathbb{R}$, and two small intervals around x and y with the same length, say $x \in [a, b]$, $y \in [c, d]$ with $\Delta x = b - a = d - c = \Delta y$ very small. By Part 1, we have

$$\frac{P(X \in [a, b])}{P(X \in [c, d])} \approx \frac{f(x)(b - a)}{f(y)(d - c)} = \frac{f(x)}{f(y)}.$$

For example, if $f(x) = 2f(y)$, then the random variable X is twice as likely to take values around x as around y.

3. *Approximation by discrete random variables.* Continuous random variables can be approximated by discrete random variables. Consider a very fine partition $-\infty < x_1 < x_2 < \cdots < x_n < \infty$, where n is large, $\Delta x_i \doteq x_{i+1} - x_i$ very small for each i, x_1 a very large negative number, and x_n a very large positive number. Then a discrete random variable Y with

$$P(Y = x_i) = f(x_i)\Delta x_i$$

is a good approximation of X. Even though $\sum_i f(x_i)\Delta x_i$ may not equal 1, it should be close. This approximation is very helpful in understanding the intuition behind many definitions and "proofs" of this section. Furthermore, this approximation implies that many results for discrete random variables should carry over to continuous random variables.

Lemma 4.1. *Let X be a continuous random variable with probability density function f. Denote by F the cumulative distribution function of X, that is, $F(x) = P(X \leq x)$ for every $x \in \mathbb{R}$. Then*

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

Proof. For an arbitrary $x \in \mathbb{R}$, take $B = (-\infty, x]$. By the definition of probability density function, we have

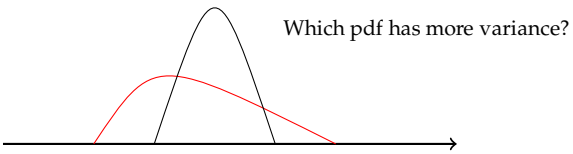
$$F(x) = P(X \leq x) = P(X \in B) = \int_B f(y)\,dy = \int_{-\infty}^{x} f(y)\,dy.$$

By the fundamental theorem of calculus, $f(x) = F'(x)$. ■

Definition 4.1. Let X be a continuous random variable with probability density function $f : \mathbb{R} \to [0, \infty)$. Then its **expected value** (or **expectation**, **mean**), **variance**, and **standard deviation** are defined as

$$\begin{aligned} E[X] &\doteq \int_{\mathbb{R}} x f(x)\,dx \\ \mathrm{Var}[X] &\doteq E[(X - E[X])^2] = \int_{\mathbb{R}} (x - E[X])^2 f(x)\,dx \\ \mathrm{Std}[X] &\doteq \sqrt{\mathrm{Var}[X]}. \end{aligned}$$

The meaning of expectation, variance, and standard deviation of a continuous random variable is completely parallel to that of a discrete random variable. If we view (correctly so) integrals as summations, then the similarity between these definitions are obvious. Thus, expected value and variance represent the average and variation of a random variable, respectively. Comparing the two probability density functions in the following figure, the red one has a larger variance or more variation as it is not as tight as the black one.



Lemma 4.2. *Let X be a continuous random variable with probability density function f, and $h : \mathbb{R} \to \mathbb{R}$ a given function. Then*

$$E[h(X)] = \int_{\mathbb{R}} h(x) f(x)\,dx.$$

Proof. This lemma is the version of Part 4 of Lemma 3.1 for continuous random variables. We will present two intuitive arguments. One is to approximate h by piecewise constant functions, the other is to approximate X by discrete random variables.

- *Argument by approximating h.* Suppose that h is a piecewise constant function with
$$h(x) = \sum_{i=1}^{n} c_i 1_{[a_i,b_i]}(x),$$
where $[a_i, b_i]$'s are non-overlapping intervals and c_i's are constants. For this type of functions h, $Y = h(X)$ is a discrete random variable taking values in $\{c_1, c_2, \ldots, c_n\}$ with
$$P(Y = c_i) = P(X \in [a_i, b_i]) = \int_{a_i}^{b_i} f(x)\,dx.$$
Therefore,
$$\begin{aligned} E[h(X)] &= E[Y] = \sum_{i=1}^{n} c_i P(Y = c_i) = \sum_{i=1}^{n} c_i \int_{a_i}^{b_i} f(x)\,dx \\ &= \sum_{i=1}^{n} c_i \int_{\mathbb{R}} 1_{[a_i,b_i]}(x) f(x)\,dx = \int_{\mathbb{R}} h(x) f(x)\,dx. \end{aligned}$$
The claim holds when h is piecewise constant function. For an arbitrary function h, one can use a sequence of piecewise constant functions to $\{h_n\}$ approximate it, i.e., $h(x) = \lim_n h_n(x)$. Thus
$$E[h(X)] = \lim_n E[h_n(X)] = \lim_n \int_{\mathbb{R}} h_n(x) f(x)\,dx = \int_{\mathbb{R}} h(x) f(x)\,dx.$$

- *Argument by approximating X.* Let $-\infty < x_1 < x_2 < \cdots < x_n < \infty$ be a very fine partition of the real line, where n is large, $\Delta x_i \doteq x_{i+1} - x_i$ very small for $1 \le i < n$, x_1 a very large negative number, and x_n a very large positive number. Define a discrete random variable X_n with
$$P(X_n = x_i) = \frac{1}{c_n} f(x_i)\Delta x_i, \quad i = 1, 2, \ldots, n-1$$
where c_n is the normalizing constant with $c_n = \sum_i f(x_i)\Delta x_i$. When n is large, X_n is a good approximation of X and c_n is approximately 1. Thanks to Lemma 3.1,
$$E[h(X_n)] = \sum_i h(x_i) P(X_n = x_i) = \frac{1}{c_n} \sum_i h(x_i) f(x_i)\Delta x_i.$$
Letting $n \to \infty$, we have $X_n \to X$ and $c_n \to 1$, which implies that
$$\begin{aligned} E[h(X)] &= \lim_n E[h(X_n)] = \lim_n \frac{1}{c_n} \sum_i h(x_i) f(x_i)\Delta x_i \\ &= \int_{\mathbb{R}} h(x) f(x)\,dx. \end{aligned}$$

We have to admit that the preceding "proof" is not entirely rigorous, especially regarding the construction of approximation and taking limit. An entirely rigorous proof will require measure theory, which is out of the scope of this class. We should mostly view these proofs as a way to convey the main ideas. ■

Since continuous random variables can be approximated by discrete random variables, results such as Lemma 3.1 should also hold for continuous random variables. Risking repetition, we should collect all these results in the following lemma. We also drop the restriction that the random variables are either discrete or continuous — the lemma holds for general random variables. The proofs are omitted, as they are either a simple application of Lemma 4.2 or similar to the approximation argument in Lemma 4.2.

Lemma 4.3. *Let X, Y and Z be arbitrary random variables. Let a and b be constants. Then*

1. $E[aX+b] = aE[X]+b$
2. $E[aX+bY] = aE[X]+bE[Y]$
3. $\mathrm{Var}[X] = E[X^2] - E^2[X]$
4. $\mathrm{Var}[aX+b] = a^2\mathrm{Var}[X]$
5. $\mathrm{Var}[X] = 0$ *if and only if* $X = c$ *for some constant* c.

Example 4.1. Let X be a continuous random variable whose probability density function takes the form

$$f(x) = cx1_{[0,1]}(x) = \begin{cases} cx & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

where c is some constant. Determine c, the cumulative distribution function of X, $E[X]$, $\mathrm{Var}[X]$, and $P(X \ge 0.5)$ and $P(X > 0.5)$.

Solution: To determine the value of c, we recall the property that the integral of a probability density function on the whole real line must equal one. That is,

$$1 = \int_{-\infty}^{\infty} f(x)\,dx = \int_0^1 cx\,dx = \frac{c}{2}.$$

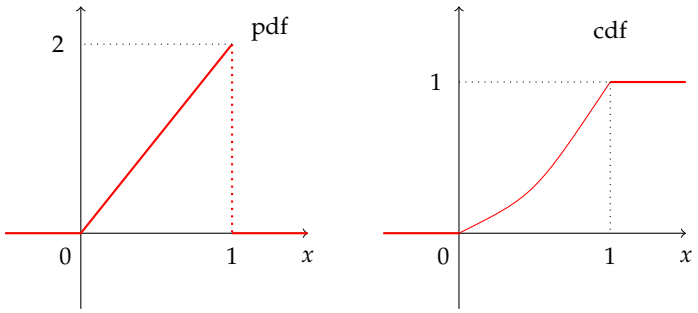
Therefore, $c = 2$. The cumulative distribution function F is defined as

$$F(x) = P(X \le x)$$

for all $x \in \mathbb{R}$. Since X can only take values in $[0,1]$, we have $F(x) = 0$ if $x < 0$ and $F(x) = 1$ if $X > 1$. It remains to consider all those $x \in [0,1]$. For such x,

$$F(x) = \int_{-\infty}^{x} f(y)\,dy = \int_{0}^{x} 2y\,dy = x^2.$$

To summarize,



Furthermore,

$$E[X] = \int_{\mathbb{R}} x\,f(x)\,dx = \int_{0}^{1} x \cdot (2x)\,dx = \frac{2}{3}.$$

$$E[X^2] = \int_{\mathbb{R}} x^2 f(x)\,dx = \int_{0}^{1} x^2 \cdot (2x)\,dx = \frac{1}{2}.$$

$$\mathrm{Var}[X] = E[X^2] - E^2[X] = \frac{1}{2} - \left(\frac{2}{3}\right)^2 = \frac{1}{18}$$

$$\mathrm{Std}[X] = \sqrt{\mathrm{Var}[X]} = \sqrt{\frac{1}{18}}.$$

Finally, we note that the probability for a continuous random variable to take a specific value is zero. In particular, $P(X = 0.5) = 0$. Therefore

$$P(X \geq 0.5) = P(X > 0.5) = 1 - P(X \leq 0.5) = 1 - F(0.5) = 0.75.$$

4.1 Common Continuous Random Variables

In this section, we collect some commonly used continuous random variables, and discuss their properties.

Definition 4.2. 1. A random variable X is said to be **uniform on interval** $[a, b]$ if X takes values in $[a, b]$ with probability density function

$$f(x) = \frac{1}{b-a} 1_{[a,b]}(x).$$

2. A random variable X is said to be **exponential with rate** $\lambda > 0$ if X is nonnegative with probability density function

$$f(x) = \lambda e^{-\lambda x} 1_{[0,\infty)}(x).$$

3. A random variable X is said to be **standard normal**, denoted by $N(0, 1)$, if X takes values in $\mathbb{R}$ with probability density function

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

4. A random variable X is said to be **normal with mean** μ **and variance** σ^2, denoted by $N(\mu, \sigma^2)$, if X takes values in $\mathbb{R}$ with probability density function

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

Standard normal is simply normal with mean $\mu = 0$ and variance $\sigma^2 = 1$. In general, $\mu \in \mathbb{R}$ and $\sigma > 0$.

4.1.1 Uniform distribution on $[a, b]$

A uniform random variable X takes values on some interval $[a, b]$ with constant density. In other words, every point on interval $[a, b]$ is equally likely. In practice, phrases such as "a number is chosen at random from interval $[a, b]$" mean implicitly that the random number is uniform on $[a, b]$. If X is uniform on $[a, b]$, it is not hard to verify that

$$E[X] = \frac{a+b}{2}, \quad \text{Var}[X] = \frac{1}{12}(b-a)^2.$$

A very interesting and very useful (in Monte Carlo simulation) result regarding uniform distribution is the following lemma.

Lemma 4.4. *Let X be a continuous random variable with cumulative distribution function F. Then $U = F(X)$ is a uniform random variable on $[0, 1]$.*

Proof. It is obvious that U takes values on interval $[0,1]$. To find out the distribution of U, we compute its cumulative distribution function. Fix an arbitrary $u \in (0,1)$.

$$G(u) \doteq P(U \leq u) = P(F(X) \leq u) = P(X \leq x^*) = F(x^*)$$

where $x^* \doteq \max\{x : F(x) = u\}$. Note that $F(x^*) = u$ by the continuity of F and thus $G(u) = u$. It follows that the probability density function for U is

$$G'(u) = 1.$$

Thus U is uniform on $[0,1]$. ■

Example 4.2. *(linear transform of uniform is still uniform)* Let U be a random variable uniformly distributed on $[0,1]$. Let $a > 0$ and b be constants and define $X = aU + b$. What is the distribution of X?

Solution: In order to determine the probability density function of X, we need to compute its cumulative distribution function F first and then take derivative. To this end, we observe that X takes values in $[b, a+b]$ because $a > 0$. Therefore, $F(x) = 0$ for $x < b$ and 1 for $x > a + b$. Now fix any $x \in [b, a+b]$. Then

$$F(x) = P(X \leq x) = P(aU + b \leq x) = P\left(U \leq \frac{x-b}{a}\right) = \frac{x-b}{a}.$$

It follows that the probability density function for X is

$$f(x) = F'(x) = \frac{1}{a} 1_{[b,a+b]}(x).$$

That is, X is uniformly distributed on interval $[b, a+b]$. It is not hard to mimic this argument to show that if U is uniform on an arbitrary interval, then X will also be uniform on some interval. ■

Example 4.3. The radius of a disc is uniformly distributed on $[0,1]$. What is the probability density function for the area of the disc?

Solution: Denote by R the radius. By assumption, R is uniformly distributed on $[0,1]$. Let X be the area of the disc. Then $X = \pi R^2$. Analogous to Example 4.2, we would compute the cumulative distribution function F of X first and then take derivative. Note that X takes values in $[0, \pi]$. Thus $F(x) = 0$ for $x < 0$ and $F(x) = 1$ for $x > \pi$. For $x \in [0, \pi]$, we have

$$F(x) = P(X \leq x) = P(\pi R^2 \leq x) = P\left(0 \leq R \leq \sqrt{\frac{x}{\pi}}\right) = \sqrt{\frac{x}{\pi}}.$$

Therefore, the probability density function for the area of the disc is

$$f(x) = F'(x) = \frac{1}{2\sqrt{\pi x}} 1_{[0,\pi]}(x).$$

Example 4.4. John and Emily play a game. At the start of the game, they agree upon a fixed number $0 < a < 1$. Then a number is chosen from interval $[0, 1]$ at random, say X. If $X < a$, then John loses \$1 to Emily. If $X \geq a$, then Emily loses \$$X$ to John. Determine the value of a so that this is a fair game to John (and thus to Emily as well), that is, John's expected winning is zero.

Solution: We can write John's winning as $h(X)$ where h is defined to be

$$h(x) = -1_{[0,a)}(x) + x1_{[a,1]}(x) = \begin{cases} -1 & \text{if } x < a \\ x & \text{if } x \geq a \end{cases}.$$

Since X is uniform on $[0, 1]$, John's expected winning is

$$\begin{aligned} E[h(X)] &= \int_{\mathbb{R}} h(x) f_X(x)\, dx = \int_0^1 h(x)\, dx \\ &= \int_0^a (-1)\, dx + \int_a^1 x\, dx = -a + \frac{1}{2}(1 - a^2). \end{aligned}$$

To make this a fair game, we need $E[h(X)] = 0$, which yields (note $0 < a < 1$)

$$a = \sqrt{2} - 1.$$

4.1.2 Exponential distributions

An exponential random variable X with rate $\lambda > 0$ takes nonnegative values. Its cumulative distribution function, expected value, and variance are

$$F(x) = \int_{-\infty}^{x} \lambda e^{-\lambda y} 1_{[0,\infty)}(y)\, dy = \int_0^x \lambda e^{-\lambda y}\, dy = 1 - e^{-\lambda x}, \quad \forall x > 0.$$

$$E[X] = \frac{1}{\lambda}, \quad \text{Var}[X] = \frac{1}{\lambda^2}$$

respectively. Here we leave the verification of the expected value and variance as an exercise to interested students. Exponential distributions have a very special **memoryless property**, which states that for all $s, t \geq 0$

$$P(X > t + s \mid X > s) = P(X > t).$$

Indeed, observe that $P(X > x) = 1 - P(X \leq x) = 1 - F(x) = e^{-\lambda x}$ for all $x \geq 0$. Therefore,

$$\begin{aligned} P(X > t+s \mid X > s) &= \frac{P(X > t+s,\, X > s)}{P(X > s)} = \frac{P(X > t+s)}{P(X > s)} \\ &= \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} = e^{-\lambda t} = P(X > t). \end{aligned}$$

What the memoryless property says is that the conditional distribution of $(X - s)$ given $X > s$, is the same as the distribution of X itself, when X is exponentially distributed! The memoryless property can be intuitively explained in the following example. Suppose that in a bus stop, the arrival times between buses are exponentially distributed with mean 20 minutes. When you arrive at the bus stop, there are a few passengers waiting for the bus as well. They tell you that the last bus left about 15 minutes ago. How much longer do you need to wait for the bus to arrive? The information "15 minutes ago" actually does not make any difference — due to the memoryless property, the time you need to wait is still exponentially distributed with mean 20 minutes, regardless of departure time of the last bus!

The memoryless property leads to the classical **inspection paradox**. Take the previous example of bus waiting. It can be derived that whenever you arrive at a bus stop, the average time between the departure of the last bus and the arrival of the next bus is always larger than 20 minutes, the average interarrival time between buses! Another inspection paradox is with respect to the lifetime of a light bulb. Assume that in a classroom a certain brand of light bulb is used for lighting. The lifetime for this brand of light bulbs is exponentially distributed with mean 6 months. Anytime you walk into the classroom and find the light bulb working, this light bulb has an average life time larger than 6 months! Actually, this light bulb can last, from the time you walk into the classroom, an exponentially distributed time with mean 6 months due to the memoryless property. Now add in the age of this light bulb, the average lifetime of this light bulb must be larger than 6 months.

Inspection paradox is a form of *selection bias*. It does not even require memoryless property or exponential distribution. The only reason the exponential distribution is used to illustrate this paradox is because memoryless property allows one to state the phenomenon in a much transparent fashion. Selection bias is everywhere. Take our example of bus waiting (or light bulb, they are very similar). Imagine a time axis marked by the arrival time of each bus. These marks will divide the time axis into intervals with varying length equal to the interarrival times. When you arrive at the bus

stop, you are more likely to arrive at a longer interval than a small interval. Therefore, the average interarrival time you observe at your arrival will be longer than the overall average of interarrivals times.

Take another example. When you ask the students about the average class size, you may get an average of class size around 100. But if you ask the college for average class size, you will find out that the average class size is around 50. Why such a large discrepancy? That is because larger (smaller) classes are more (less) likely to be sampled when you ask the students about class size. Your estimate from students' response will be biased high.

Example 4.5. *(Inspection paradox, Selection bias)* Let $0 = a_0 < a_1 < a_2 < \cdots < a_{n-1} < a_n = 1$ be a partition of the interval $[0, 1]$. It divides the interval into n subintervals with length $b_i = a_{i+1} - a_i$ for $i = 0, 1, 2, \ldots, n-1$, respectively. What is the average length of these subinterval? Now suppose we choose a point at random from interval $[0, 1]$. What is the average length of the subinterval that covers this random point? Which average is larger?

Solution: The total length of these n subintervals is of course 1, and thus the average length is $1/n$. Now let X be a point chosen at random from $[0, 1]$, or X is uniform on $[0, 1]$. Define $h(X)$ to be the length of the subinterval that covers X. In other words, if $a_i \le X < a_{i+1}$ then the subinterval covers X is $[a_i, a_{i+1})$, whose length is $b_i = a_{i+1} - a_i$, and thus $h(X) = b_i$. There are two ways to compute $E[h(X)]$.

1. Regard $h(X)$ as a discrete random variable by itself. It takes values in $\{b_1, b_2, \ldots, b_n\}$ with $P(h(X) = b_i) = P(a_i \le X < a_{i+1}) = b_i$. Therefore,
$$E[h(X)] = \sum_{i=1}^{n} b_i P(h(X) = b_i) = \sum_{i=1}^{n} b_i^2$$
There is a slight hole in the argument that b_i's may not be all distinct. But this is easy to fix by combining the probabilities of those subintervals of same length.

2. Regard $h(X)$ as a function of the continuous (uniform) random variable X and use Lemma 4.2. Note that
$$h(X) = \sum_{i=1}^{n} b_i 1_{[a_i, a_{i+1})}(X).$$

Since X is uniform on $[0,1]$, we have

$$E[h(X)] = \int_0^1 h(x)\,dx = \sum_{i=1}^{n} b_i \int_{a_i}^{a_{i+1}} dx = \sum_{i=1}^{n} b_i^2.$$

Note that since a random point is more likely to fall into a longer subinterval, the average length of the subintervals that cover the random point will be larger than the average length of all subintervals, which is $1/n$. In other words,

$$\sum_{i=1}^{n} b_i^2 \geq \frac{1}{n}.$$

This is indeed a special case of the celebrated *Cauchy-Schwarz inequality*. Equality holds if and only if $b_1 = b_2 = \cdots b_n = 1/n$, in which case every subinterval has equal chance to cover the random point X. ■

Lemma 4.5. *Let X be a nonnegative random variable with continuous cumulative distribution function F. Assume that X satisfies the memoryless property:*

$$P(X > t + s \mid X > s) = P(X > t)$$

for all $s, t \geq 0$. Then X is an exponential random variable.

Proof. The memoryless property amounts $\bar{F}(s)\bar{F}(t) = \bar{F}(s+t)$ for all $s, t \geq 0$, where $\bar{F}$ is the tail probability function with

$$\bar{F}(x) \doteq P(X > x) = 1 - F(x).$$

The first observation we make is that $\bar{F}(x) > 0$ for all $x \geq 0$. Actually if $\bar{F}(x) = 0$ for some $x \geq 0$, then by assumption $\bar{F}^n(x/n) = \bar{F}(x)$ implies $\bar{F}(x/n) = 0$ for all n. Now letting n tend to ∞, the continuity of $\bar{F}$ implies that $\bar{F}(0) = P(X > 0) = 0$. This, along with the assumption that X is nonnegative, leads to $X = 0$ with probability one. This is a contradiction since we have assumed that F is continuous. The second observation we make is that $\bar{F}(x) < 1$ for all $x > 0$. Otherwise (say) $\bar{F}(a) = 1$ for some $a > 0$. Then $\bar{F}(na) = \bar{F}^n(a) = 1$ for all n. Letting n tend to ∞, we have $\bar{F}(\infty) = 1$, a contradiction.

With $0 < \bar{F}(x) < 1$ for all $x > 0$, we can define $H(x) = -\log \bar{F}(x)$, which is continuous and strictly positive for all $x > 0$. Memoryless property yields that for all $s, t > 0$,

$$H(s) + H(t) = H(s+t).$$

In particular, $H(nt) = nH(t)$ for all $t > 0$ and positive integer n. Define $\lambda \doteq H(1)$. Set $t = 1$, then $H(n) = nH(1) = n\lambda$ for all positive integer n. Now for any positive integers n, m, let $t = m/n$ and we have $H(m) = nH(m/n)$, which leads to $H(m/n) = H(m)/n = m\lambda/n$. In other words, $H(t) = t\lambda$ for all positive rational numbers t. Now by the continuity of H, $H(x) = x\lambda$ for all $x > 0$, since any real number can be approximated by rational numbers.

It follows that for any $x > 0$, $\bar{F}(x) = e^{-H(x)} = e^{-\lambda x}$. The cumulative distribution function is $F(x) = 1 - \bar{F}(x) = 1 - e^{-\lambda x}$. Therefore, for $x > 0$, the probability density function is

$$f(x) = F'(x) = \lambda e^{-\lambda x}.$$

That is, X is exponential with rate $\lambda > 0$. We complete the proof. ■

4.1.3 Standard normal distribution

The ubiquitous normal distributions are arguably *the most important* distributions in probability and statistics. Normal random variables can take any positive or negative values. Among all the normal distributions, the one with density

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

is said to be the **standard normal distribution**, denoted by $N(0, 1)$. But before we discuss the properties of the standard normal distribution, we first verify that φ really defines a probability density function, with mean 0 and variance 1.

Lemma 4.6. *The function φ defines a true probability density function, that is, $\varphi \geq 0$ and*

$$\int_{\mathbb{R}} \varphi(x)\, dx = 1.$$

Furthermore,

$$\int_{\mathbb{R}} x\varphi(x)\, dx = 0, \qquad \int_{\mathbb{R}} x^2\varphi(x)\, dx = 1.$$

Proof. φ is obviously nonnegative. To calculate the first integral, we will utilize double integral and polar coordinates. Observe that

$$\left(\int_{\mathbb{R}} \varphi(x)\, dx\right)^2 = \int_{\mathbb{R}} \varphi(x)\, dx \cdot \int_{\mathbb{R}} \varphi(y)\, dy = \iint_{\mathbb{R}^2} \varphi(x)\varphi(y)\, dxdy.$$

Using polar coordinates: $x = r\cos\theta$, $y = r\sin\theta, r \in [0,\infty), \theta \in [0, 2\pi)$, we have $dxdy = r\,drd\theta$, $x^2 + y^2 = r^2$, and

$$\begin{aligned}\left(\int_{\mathbb{R}} \varphi(x)\,dx\right)^2 &= \int_0^\infty \int_0^{2\pi} \frac{1}{2\pi} e^{-r^2/2} r\,d\theta dr = \int_0^\infty e^{-r^2/2} r\,dr \\ &= \int_0^\infty e^{-r^2/2}\,d\left(\frac{r^2}{2}\right) = 1.\end{aligned}$$

Therefore,

$$\int_{\mathbb{R}} \varphi(x)\,dx = 1.$$

The equation

$$\int_{\mathbb{R}} x\varphi(x)\,dx = 0$$

is trivial by symmetry. Finally, observe that $\varphi'(x) = -x\varphi(x)$. It follows from integration by parts,

$$\int_{\mathbb{R}} x^2\varphi(x)\,dx = -\int_{\mathbb{R}} x\varphi'(x)\,dx = -x\varphi(x)\Big|_0^\infty + \int_{\mathbb{R}} \varphi(x)\,dx = 0 + 1 = 1.$$

We complete the proof. ■

Lemma 4.6 justifies that φ indeed defines a probability density function, and that standard normal has mean 0 and variance 1. Throughout the lectures, we also denote by Φ the **cumulative distribution function of the standard normal distribution**:

$$\Phi(x) = \int_{-\infty}^x \varphi(y)\,dy = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\,dy = P(N(0,1) \le x).$$

It is straightforward from the symmetry of the density function φ that for any $x \in \mathbb{R}$,

$$\Phi(x) + \Phi(-x) = 1.$$

4.1.4 General normal distributions

The probability density function of a general normal distribution with mean μ and variance σ^2, denoted by $N(\mu, \sigma^2)$, is given by

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

Lemma 4.6, along with a simple change of variable, can verify that f defines a probability density function with mean μ and variance σ^2 (exercise).

Not only that, the next lemma will also argue that any linear transform of normal is still normal, and any normal is indeed a linear transform of the standard normal.

Lemma 4.7. *Suppose X is a normal random variable with mean μ and variance σ^2. Define $Y = aX + b$ where $a \neq 0$ and b are constants. Then Y is also normal, with mean $a\mu + b$ and variance $a^2\sigma^2$. In particular,*

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

is standard normal (standardization).

Proof. Without loss of generality, we assume $a > 0$. The proof for the case of $a < 0$ is similar and thus omitted. To compute the density of Y, we will first compute its cumulative distribution function F and then take derivative to obtain the density. Fix any $y \in \mathbb{R}$. We have

$$\begin{aligned} F(y) &= P(Y \leq y) = P(aX + b \leq y) = P\left(X \leq \frac{y-b}{a}\right) \\ &= \int_{-\infty}^{(y-b)/a} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}\, dx. \end{aligned}$$

Define a change of variable with $z = ax + b$. We have $dz = a\, dx$ and

$$F(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}a\sigma} e^{-(z-a\mu-b)^2/(2a^2\sigma^2)}\, dz.$$

Therefore, the probability density function of Y is (by the fundamental theorem of calculus)

$$F'(y) = \frac{1}{\sqrt{2\pi}a\sigma} e^{-(z-a\mu-b)^2/(2a^2\sigma^2)},$$

which is the probability density function for the normal distribution with mean $a\mu + b$ and variance $a^2\sigma^2$. We complete the proof. ∎

Example 4.6. Let X be a normal random variable with mean $\mu = 8$ and variance $\sigma^2 = 4$. Express $P(6 < X < 12)$ in terms of Φ.

Solution: The idea for this type of computation is standardization. Also note that the probability will not change if both the "$<$" are replaced by "$\leq$", since the probability of X taking a specific value is 0. Therefore,

$$P(6 < X < 12) = P\left(\frac{6-8}{\sqrt{4}} < \frac{X-8}{\sqrt{4}} < \frac{12-8}{\sqrt{4}}\right) = P(-1 < Z < 2).$$

where Z is a standard normal, thanks to Lemma 4.7. It follows that

$$P(30 < X < 60) = \Phi(2) - \Phi(-1) = \Phi(2) + \Phi(1) - 1.$$

If you check the normal distribution table in the textbook, this probability is approximately 0.8185.

Example 4.7. Assume that X is a normal random variable with distribution $N(\mu, \sigma^2)$. Show that

$$P(|X - \mu| \leq k\sigma) = 1 - 2\Phi(-k) = 2\Phi(k) - 1.$$

for any $k > 0$.

Proof. The argument is again based on standardization. Let $Z = (X - \mu)/\sigma$. Thanks to Lemma 4.7, Z is standard normal or $N(0, 1)$. Therefore,

$$P(|X - \mu| \leq k\sigma) = P(|Z| \leq k) = \Phi(k) - \Phi(-k).$$

The claim now follows readily since $\Phi(x) + \Phi(-x) = 1$ for all x.

When $k = 2$, this probability is roughly 95%. That is, "*With 95% probability, a normal random variable is within two standard deviations away from the mean*". This observation is quite often used in the construction of confidence intervals and hypothesis testing, topics we will cover later in the class. ■

Example 4.8. The price S of a stock is often modeled by *lognormal* distributions. That is, $S = e^X$ where X is $N(\mu, \sigma^2)$. Compute $P(S > a)$ and $E[S]$, where a is a given positive constant.

Solution: Again, standardization can be quite convenient in the computation of such quantities. Let $Z = (X - \mu)/\sigma$, which is a standard normal. We can write $X = \mu + \sigma Z$. Then

$$P(S > a) = P(\log S > \log a) = P(X > \log a) = P\left(Z > \frac{\log a - \mu}{\sigma}\right).$$

Therefore,

$$P(S > a) = 1 - \Phi\left(\frac{\log a - \mu}{\sigma}\right) = \Phi\left(-\frac{\log a - \mu}{\sigma}\right).$$

In order to compute $E[S]$, we first establish a useful identity, that is, for any $\theta \in \mathbb{R}$, we have

$$E[e^{\theta Z}] = e^{\theta^2/2}.$$

This claim actually gives the moment generating function of the standard normal distribution. The justification is straightforward:

$$E[e^{\theta Z}] = \int_{-\infty}^{\infty} e^{\theta z} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz = e^{\theta^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(z-\theta)^2/2}\, dz.$$

Observe that the last integrand is indeed the probability density function for $N(\theta, 1)$, whose integral must be 1. The claim follows. Now

$$E[S] = E[e^X] = E[e^{\mu+\sigma Z}] = e^{\mu} E[e^{\sigma Z}] = e^{\mu+\sigma^2/2}.$$

Exercise 4.9. Let S be the same as in Example 4.8. Given a constant $K > 0$, compute

$$E[(S-K)^+],$$

where $x^+ \doteq \max(x, 0)$ for all x. Express your answer in terms of Φ.

Chapter 5

Multivariate Probability Distributions

So far, our dealings are mostly with respect to a single random variable. However, more often than not, a stochastic model or statistical application will involve more than one random variable. Therefore, it is essential to study multiple random variables and their relations.

A **random vector** is simply a collection of random variables all defined on the same sample space, say $(X_1, X_2, \ldots, X_d)$. In a lot of literature, such a random vector is also called a random variable. But in general, whether a random variable means a vector or not is quite clear within context.

Most of our definitions and theorems in this chapter will be presented for the case of two random variables. The generalization to any finite collection of random variables is straightforward though. We will need the following terminology. Let A and B be two arbitrary sets. The **(Cartesian) product set** of A and B, denoted by $A \times B$, is defined to be the set

$$A \times B = \{(a, b) : a \in A, b \in B\}.$$

For example, the classical space $\mathbb{R}^2$ is indeed the product set $\mathbb{R} \times \mathbb{R}$. This definition can be easily extended to the product of finitely many sets.

5.1 Basics of Multivariate Probability Distributions

Definition 5.1. Let X, Y be two discrete random variables. Assume X takes values in $\{x_1, x_2, \ldots\}$ and Y in $\{y_1, y_2, \ldots\}$. Then the **joint probability mass function (pmf)** of the random vector (X, Y) is defined by

$$p(x_i, y_j) = P(X = x_i, Y = y_j).$$

The **(marginal) probability mass functions** for X and for Y are defined by, respectively,

$$\begin{aligned} p_X(x_i) &= \sum_{y_j} p(x_i, y_j) = \sum_{y_j} P(X = x_i, Y = y_j) = P(X = x_i) \\ p_Y(y_j) &= \sum_{x_i} p(x_i, y_j) = \sum_{x_i} P(X = x_i, Y = y_j) = P(Y = y_j). \end{aligned}$$

Definition 5.2. Let X, Y be two random variables. We say (X, Y) are **jointly continuous random variables**, if there exists a **joint probability density function (pdf)** $f : \mathbb{R}^2 \to [0, \infty)$ such that for any $B \subseteq \mathbb{R}^2$,

$$P((X, Y) \in B) = \iint_B f(x, y)\, dxdy,$$

and

$$f(x, y) \geq 0, \quad \iint_{\mathbb{R}^2} f(x, y)\, dxdy = 1.$$

The **(marginal) probability density functions** for X and for Y are defined by, respectively,

$$f_X(x) = \int_{\mathbb{R}} f(x, y)\, dy, \quad f_Y(y) = \int_{\mathbb{R}} f(x, y)\, dx.$$

Lemma 5.1. *Suppose that (X, Y) are discrete random variables with joint probability mass function $p(x_i, y_j) = P(X = x_i, Y = y_j)$.*

1. *The marginal probability mass functions p_X and p_Y are the probability mass functions for X and Y, respectively.*
2. *For any function $h : \mathbb{R}^2 \to \mathbb{R}$,*

$$E[h(X, Y)] = \sum_{x_i} \sum_{y_j} h(x_i, y_j) p(x_i, y_j).$$

Suppose that (X, Y) are continuous random variables with joint probability density function $f : \mathbb{R}^2 \to [0, \infty)$.

1. *The marginal probability density functions f_X and f_Y are the probability density functions for X and Y, respectively.*
2. *For any function $h : \mathbb{R}^2 \to \mathbb{R}$,*

$$E[h(X, Y)] = \iint_{\mathbb{R}^2} h(x, y) f(x, y)\, dxdy$$

Proof. When (X, Y) is a discrete random vector, Part 1 is trivial from definition. The proof for Part 2 is omitted since it is similar to that of Lemma 3.1 (Parts 2 and 4). When (X, Y) is a continuous random vector, the proof for Part 2 is similar to that of Lemma 4.2, and thus omitted as well. For Part 1, we should only prove that f_X is the probability density function for X. The proof for f_Y is almost verbatim.

Let $A \subseteq \mathbb{R}$ be arbitrary. Then $\{X \in A\} = \{(X, Y) \in B\}$ where $B = A \times \mathbb{R}$. Therefore, by definition

$$P(X \in A) = \iint_B f(x, y)\, dxdy = \int_A \int_{\mathbb{R}} f(x, y)\, dy\, dx = \int_A f_X(x)\, dx,$$

which implies that f_X is the probability density function for X. We complete the proof. ■

Lemma 5.1 shows that the "*marginal* probability mass function" for X means exactly the same as the "probability mass function" for X. The appearance of the word "marginal" is simply to emphasize that X is a component of a higher-dimensional random vector or its relevance to the more detailed joint probability mass function. The word "marginal" is frequently omitted.

Lemma 5.1 also reinforces the linearity of expectation (Part 2 of Lemma 3.1 and Lemma 4.3). In other words, if we take $h(x, y) = ax + by$ for some constants a and b, we have

$$E[aX + bY] = aE[X] + bE[Y].$$

for all discrete or continuous random vectors (X, Y). We would like to emphasize that this linearity actually holds for *all* X and Y, regardless if they are discrete, continuous, or otherwise.

Remark 5.1. Analogous to the cumulative distribution function for a random variable, one can define the (joint) cumulative distribution function F for a random vector (X, Y) as follows. Define $F : \mathbb{R}^2 \to [0, 1]$ by

$$F(x, y) \doteq P(X \leq x,\, Y \leq y)$$

for any $x, y \in \mathbb{R}$. When (X, Y) is continuous with joint probability density function f, then

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(s, t)\, dtds.$$

Similar to the relation between pdf and cdf for continuous random variables, we have the following relation

$$f(x,y) = \frac{\partial^2 F}{\partial x \partial y}(x,y).$$

5.2 Covariance and Correlation Coefficient

Definition 5.3. For two random variables X and Y, their **covariance** and **correlation coefficient** are defined by, respectively,

$$\begin{aligned} Cov(X,Y) &\doteq E[(X - E[X])(Y - E[Y])] \\ Corr(X,Y) &\doteq \frac{Cov(X,Y)}{\sqrt{\mathrm{Var}[X]}\sqrt{\mathrm{Var}[Y]}}, \quad \text{if } \mathrm{Var}[X] \neq 0 \text{ and } \mathrm{Var}[Y] \neq 0. \end{aligned}$$

The covariance measures how X and Y change together. Loosely speaking, when the covariance is positive, X and Y change more or less in the same direction — when X increases (decreases), Y tends to increase (decrease) as well. On the other hand, when the covariance is negative, X and Y change more or less in the opposite directions. The correlation coefficient inherits the sign of the covariance, but always takes values in interval $[-1, 1]$. It also indicates how close the relation between X and Y is to a linear relationship — when it equals ± 1, X and Y are linearly related, i.e., $Y = \alpha X + \beta$ for some constants α and β (see Lemma 5.2).

Lemma 5.2. *Let X, Y, Z be arbitrary random variables, and a, b arbitrary constants. Then*

1. $Cov(X, a) = Cov(a, X) = 0$
2. $Cov(X, Y) = Cov(Y, X)$
3. $Cov(aX + bY, Z) = aCov(X, Z) + bCov(Y, Z)$
4. $Cov(X, aY + bZ) = aCov(X, Y) + bCov(X, Z)$
5. $\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2Cov(X, Y)$
6. $-1 \leq Corr(X, Y) \leq 1$
7. $Corr(X, Y) = \pm 1$ *if and only if* $Y = \alpha X + \beta$ *for some constants* α *and* β.

Proof. The proof for Parts 1-5 follows directly from definition and straightforward algebraic calculation, and thus omitted. We should only give a proof for Parts 6 and 7. In the proof we will implicitly assume that X and Y are not constants, lest the correlation coefficient is not defined. We will apply the *Cauchy-Schwarz inequality* (Remark 5.3) to the random variables $X - E[X]$ and $Y - E[Y]$ to obtain

$$\mathrm{Cov}^2(X, Y) \leq \mathrm{Var}[X] \cdot \mathrm{Var}[Y],$$

This implies Part 6 immediately. As for Part 7, note that $\mathrm{Corr}(X, Y) = \pm 1$ if and only if the preceding inequality is indeed an equality. Therefore, thanks to Remark 5.3, $\mathrm{Corr}(X, Y) = \pm 1$ if and only if $a(X - E[X]) + b(Y - E[Y]) = 0$ for some constants a and b, which amounts to $Y = \alpha X + \beta$ for some constants α and β. We complete the proof. ■

Example 5.1. A fair coin is tossed three times. Let X be the number of heads among the first two tosses and Y the number of heads among the last two tosses. Find the joint probability mass function for (X, Y), the marginal probability mass functions for X and Y, and compute $\mathrm{Cov}(X, Y)$.

Solution: Both X and Y can only take values in $\{0, 1, 2\}$. It is probably easiest to represent the probability mass function by a table. For example, $P(X = 1, Y = 1) = P(HTH) + P(THT) = 1/4$.

		Y			
		0	1	2	p_X
X	0	1/8	1/8	0	1/4
	1	1/8	1/4	1/8	1/2
	2	0	1/8	1/8	1/4
	p_Y	1/4	1/2	1/4	1

The marginal probability mass functions for X and for Y can be obtained by summing up the corresponding rows or columns, respectively. To compute the covariance, we have

$$E[X] = E[Y] = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1$$

and (we omit the terms where $XY = 0$)

$$E[XY] = (1 \cdot 1) \cdot \frac{1}{4} + (1 \cdot 2)\frac{1}{8} + (2 \cdot 1)\frac{1}{8} + (2 \cdot 2)\frac{1}{8} = \frac{5}{4}.$$

Therefore, $\mathrm{Cov}(X, Y) = 0.25$.

Example 5.2. Let (X, Y) be a continuous random vector with joint probability density function

$$f(x,y) = \begin{cases} a & \text{if } x, y \geq 0,\ x+y \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Find the value of a, the marginal probability density function for X and compute $P(X \leq 0.5)$.

Solution: Let $D = \{(x,y) : x, y \geq 0,\ x+y \leq 1\}$. The joint probability density function is only nonzero on D. Since

$$1 = \iint_{\mathbb{R}^2} f(x,y)\,dxdy = \iint_D a\,dxdy = a \cdot \text{Area}(D),$$

and $\text{Area}(D) = 0.5$, we have $a = 2$. To compute the marginal probability density function for X, we first observe that X can only take values on $[0,1]$. Thus the density will be 0 outside this interval. Now take any $x \in [0,1]$. We have

$$f_X(x) = \int_{\mathbb{R}} f(x,y)\,dy = \int_0^{1-x} a\,dy = 2(1-x).$$

Finally, to compute $P(X \leq 0.5)$, we can do it in two equivalent ways — either compute the probability from the marginal probability density function $f_X(x)$, or directly from the joint probability density function f:

1. *from marginal pdf:*

$$P(X \leq 0.5) = \int_{-\infty}^{0.5} f_X(x)\,dx = \int_0^{0.5} 2(1-x)\,dx = 0.75.$$

2. *from joint pdf:*

$$P(X \leq 0.5) = \int_{-\infty}^{0.5}\int_{-\infty}^{\infty} f(x,y)\,dydx = \int_0^{0.5}\int_0^{1-x} 2\,dydx = 0.75.$$

Remark 5.2. The formula for the variance of sum of random variables in Part 5 of lemma 5.2 can be generalized to multiple random variables: Let $X_1, X_2, \ldots, X_d$ be a collection of random variables. Then

$$\text{Var}\left[\sum_{i=1}^{d} X_i\right] = \sum_{i=1}^{d} \text{Var}[X_i] + 2\sum_{1 \leq i < j \leq d} \text{Cov}(X_i, X_j)$$

This can be shown by repeatedly applying Part 6 of Lemma 5.2. We will leave this exercise to interested students.

Remark 5.3. *(Cauchy-Schwarz inequality)* For any random variables X and Y, we have $E^2[XY] \leq E[X^2]E[Y^2]$ with equality if and only if $aX + bY = 0$ for some constants a and b that are not both zero. We will defer the proof to Appendix A.

Exercise 5.3. Suppose that the joint probability density function for (X, Y) is

$$f(x, y) = \begin{cases} axy & \text{if } 0 \leq x \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Determine constant a and $\mathrm{Cov}(X, Y)$. (*Solution:* $8, 4/225$)

Exercise 5.4. Let X and Y be two random variables and $\mathrm{Var}[Y] > 0$. Which constant β minimizes $\mathrm{Var}[X + \beta Y]$?

5.3 Independence

The joint distribution of a random vector is completely characterized by its joint probability mass/density function, which also completely determines the relation between the components of the random vector. However, it is sometimes more convenient to specify the relation between these components directly and use it to study the random vector. The most basic relation between random variables is *independence*.

Definition 5.4. A collection of random variables $\{X_1, X_2, \ldots, X_d\}$ are said to be **independent** if for any subsets $B_i \subseteq \mathbb{R}$ where $i = 1, 2, \ldots, d$, we have

$$P(X_1 \in B_1,\ X_2 \in B_2, \ldots,\ X_d \in B_d) = \prod_{i=1}^{d} P(X_i \in B_i).$$

Lemma 5.3. *Let $\{X_1, X_2, \ldots, X_d\}$ be a collection of independent random variables. Let $h_i : \mathbb{R} \to \mathbb{R}$, $i = 1, 2, \ldots, d$, be a collection of functions. Then $\{h_1(X_1), h_2(X_2), \ldots, h_d(X_d)\}$ are also independent.*

Proof. All we need to observe is that for any $B_i \subseteq \mathbb{R}$, $h_i(X_i) \in B_i$ if and only if $X_i \in h_i^{-1}(B_i) \subseteq \mathbb{R}$ where $h_i^{-1}(B_i)$ is the **preimage** of B_i:

$$h_i^{-1}(B_i) \doteq \{x \in \mathbb{R} :\ h_i(x) \in B_i\}.$$

The rest is directly by definition. ■

Lemma 5.4. *This lemma will be stated for two random variables, but its extension to any finite collection of random variables is straightforward. Let X, Y be two random variables.*

1. *Assume that X and Y are discrete. Then the following statements are equivalent.*

   (a) *X, Y are independent;*

   (b) *The joint probability mass function is the product of marginal probability mass functions: $p(x_i, y_j) = p_X(x_i)p_Y(y_j)$;*

   (c) *The joint probability mass function can be written in the form: $p(x_i, y_j) = h(x_i)g(y_j)$ for some nonnegative function h and g.*

2. *Assume that X and Y are continuous. Then the following statements are equivalent.*

   (a) *X, Y are independent;*

   (b) *The joint probability density function is the product of marginal probability density functions: $f(x, y) = f_X(x)f_Y(y)$;*

   (c) *The joint probability density function can be written in the form: $f(x, y) = h(x)g(y)$ for some nonnegative function h and g.*

Proof. We will only present the proof for the case when X and Y are continuous, and leave the similar proof for the discrete case to interested students. We first argue that $(a) \Rightarrow (b)$. Suppose X and Y are independent. Then the joint cumulative distribution function for (X, Y) is, by definition of independence,

$$\begin{aligned} F(x, y) &= P(X \le x, Y \le y) = P(X \le x)P(Y \le y) \\ &= \int_{-\infty}^{x} f_X(s)\,ds \int_{-\infty}^{y} f_Y(t)\,dt = \int_{-\infty}^{x}\int_{-\infty}^{y} f_X(s)f_Y(t)\,dtds. \end{aligned}$$

for any $x, y \in \mathbb{R}$. Therefore, the joint probability density function for (X, Y) is

$$f(x, y) = \frac{\partial^2 F}{\partial x \partial y}(x, y) = f_X(x)f_Y(y).$$

Now let us prove the direction $(b) \Rightarrow (a)$. Let $A, B \subseteq \mathbb{R}$ be any subsets. Then by assumption,

$$\begin{aligned} P(X \in A, Y \in B) &= \int_A \int_B f(x, y)\,dydx = \int_A \int_B f_X(x)f_Y(y)\,dydx \\ &= \int_A f_X(x)\,dx \int_B f_Y(y)\,dy \\ &= P(X \in A)P(Y \in B). \end{aligned}$$

Therefore, X and Y are independent. Lastly, $(b) \Rightarrow (c)$ is trivial. It remains to prove $(c) \Rightarrow (b)$. Suppose $f(x,y) = h(x)g(y)$. Define

$$a = \int_{\mathbb{R}} h(x)\,dx, \quad b = \int_{\mathbb{R}} g(y)\,dy.$$

It follows that, for any $x \in \mathbb{R}$,

$$\begin{aligned} P(X \le x) &= P((X,Y) \in (-\infty, x] \times \mathbb{R}) = \iint_{(-\infty,x]\times\mathbb{R}} f(s,y)\,dsdy \\ &= \int_{-\infty}^{x}\int_{\mathbb{R}} h(s)g(y)\,dyds = b\int_{-\infty}^{x} h(s)\,ds. \end{aligned}$$

Therefore, the marginal probability density function for X is $f_X(x) = bh(x)$. Similarly, the marginal probability density function for Y can be shown to be $f_Y(y) = ag(y)$. But we also observe that $ab = 1$ since

$$\begin{aligned} 1 &= \iint_{\mathbb{R}^2} f(x,y)\,dxdy = \iint_{\mathbb{R}^2} h(x)g(y)\,dxdy \\ &= \int_{\mathbb{R}} h(x)\,dx \int_{\mathbb{R}} g(y)\,dy = ab. \end{aligned}$$

Thus $f(x,y) = h(x)g(y) = f_X(x)f_Y(y)$. We complete the proof. ■

Example 5.5. Given the following cases of joint probability density functions for (X,Y), in which cases are X and Y independent?

1. $$f(x,y) = \begin{cases} x+y & \text{if } 0 \le x,y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

2. $$f(x,y) = \begin{cases} 6x^2y & \text{if } 0 \le x,y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

3. $$f(x,y) = \begin{cases} 8xy & \text{if } 0 \le x \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Solution: X and Y are not independent in the first case because $f(x,y)$ cannot be written in form of $h(x)g(y)$. They are, however, independent in the second case since we can take

$$h(x) = 6x^2 1_{[0,1]}(x), \quad g(y) = y1_{[0,1]}(y).$$

In contrast, X and Y are *not* independent in the third case. Even though it seems that $8xy$ is of the desired form, but the domain $\{0 \leq x \leq y \leq 1\}$ will not allow us to find appropriate indicator functions (like in the second case) to split $f(x, y)$. ■

Example 5.6. Assume that X and Y are independent binomial random variables, where X is $B(n, p)$ and Y is $B(m, p)$. What is the distribution of $X + Y$?

Solution: One can, of course, use brutal force to compute the probability mass function for $X + Y$. However, it is easier to observe the following. X is the total number of "success" from n independent trials with probability of success p. Y is the total number of "success" from m independent trials with the same probability of success p. Since X and Y are independent, we can regard $X + Y$ as the total number of "success" from the combined $n + m$ trials. Therefore, $X + Y$ is indeed $B(n + m, p)$. ■

Example 5.7. Let X and Y be independent Poisson random variables with parameter λ and μ, respectively. What is the distribution of $X + Y$?

Solution: First of all, $X + Y$ takes values in all nonnegative integers. Fix any $n \geq 0$. We have, by the independence of X and Y,

$$P(X+Y=n) = \sum_{k=0}^{n} P(X=k, Y=n-k) = \sum_{k=0}^{n} P(X=k)P(Y=n-k)$$

By assumption,

$$P(X=k)P(Y=n-k) = e^{-\lambda}\frac{\lambda^k}{k!} \cdot e^{-\mu}\frac{\mu^{n-k}}{(n-k)!} = e^{-(\lambda+\mu)}\frac{1}{n!}\binom{n}{k}\lambda^k\mu^{n-k}.$$

It follows from binomial expansion that

$$P(X+Y=n) = \sum_{k=0}^{n} e^{-(\lambda+\mu)}\frac{1}{n!}\binom{n}{k}\lambda^k\mu^{n-k} = e^{-(\lambda+\mu)}\frac{(\lambda+\mu)^n}{n!}.$$

That is, $X + Y$ is Poisson with parameter $\lambda + \mu$. ■

Example 5.8. Let X and Y be independent exponential random variables with rate λ and μ, respectively. What is the distribution of $\min(X, Y)$?

Solution: In order to compute the probability density function, we should first compute the cumulative distribution function. Observe that $\min(X, Y)$

takes nonnegative values. Fix any $t \geq 0$. By independence and the form of the tail probability for exponential distributions (see Section 4.1.2), we have

$$\begin{aligned} F(t) &= P(\min(X,Y) \leq t) \;=\; 1 - P(\min(X,Y) > t) \\ &= 1 - P(X > t, Y > t) \;=\; 1 - P(X > t)P(Y > t) \\ &= 1 - e^{-\lambda t} e^{-\mu t} \;=\; 1 - e^{-(\lambda+\mu)t}. \end{aligned}$$

Taking derivative with respect to t, we obtain the density function

$$f(t) = F'(t) = (\lambda + \mu) e^{-(\lambda+\mu)t} 1_{[0,\infty)}(t)$$

That is, $\min(X,Y)$ is exponential with rate $(\lambda + \mu)$. ■

Example 5.9. Let X and Y be independent exponential random variables with rate λ and μ, respectively. Compute $P(X < Y)$.

Solution: By independence, the joint probability density function of (X,Y) is

$$f(x,y) = f_X(x) f_Y(y) = \lambda e^{-\lambda x} 1_{[0,\infty)}(x) \cdot \mu e^{-\mu y} 1_{[0,\infty)}(y).$$

Therefore,

$$\begin{aligned} P(X < Y) &= \iint_{\{x<y\}} f(x,y)\, dxdy \;=\; \int_0^\infty \int_x^\infty \lambda e^{-\lambda x} \mu e^{-\mu y} dy dx \\ &= \int_0^\infty \lambda e^{-\lambda x} e^{-\mu x}\, dx \;=\; \frac{\lambda}{\lambda + \mu}. \end{aligned}$$

Exercise 5.10. Let X and Y be independent exponential random variables with rates λ and μ, respectively.

1. Show that the discrete random variable $1_{\{X<Y\}}$ is *independent* of the continuous random variable $\min(X,Y)$.
2. Can you compute $E[\max(X,Y)]$ using Examples 5.8, 5.9, Part 1 and the memoryless property of exponential distributions?
3. Verify your answer for Part 2 by direct integration.

Lemma 5.5. (convolution) *Suppose X and Y are independent continuous random variables with probability density functions $f_X(x)$ and $f_Y(y)$, respectively. Then $X + Y$ is a continuous random variable with density*

$$g(x) = (f_X * f_Y)(x) \doteq \int_{\mathbb{R}} f_X(s) f_Y(x-s)\, ds = \int_{\mathbb{R}} f_X(x-t) f_Y(t)\, dt$$

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

Proof. Denote by G the cumulative distribution function of $X+Y$. For any $x \in \mathbb{R}$,

$$G(x) = P(X+Y \le x) = \iint_{\{s+t\le x\}} f(s,t)\, dsdt,$$

where f is the joint probability density function for (X,Y). By independence, we have $f(s,t) = f_X(s)f_Y(t)$, thanks to Lemma 5.4. Therefore, by integrating against t first and then s,

$$G(x) = \iint_{\{s+t\le x\}} f_X(s)f_Y(t)\, dtds = \int_{-\infty}^{\infty} f_X(s) \int_{-\infty}^{x-s} f_Y(t)\, dtds.$$

Using a change of variable $t = u - s$ and afterwards exchanging the order of integral, we have

$$G(x) = \int_{-\infty}^{\infty} f_X(s) \int_{-\infty}^{x} f_Y(u-s)\, duds = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f_X(s)f_Y(u-s)\, ds\, du.$$

Taking derivative with respect to x, we obtain the density for $X+Y$:

$$G'(x) = \int_{-\infty}^{\infty} f_X(s)f_Y(x-s)\, ds.$$

The other equality is similar — all we need to do is to integrate against s first and then t when we evaluate $G(x)$. We complete the proof. ∎

Theorem 5.6. *Suppose $\{X_1, X_2, \ldots, X_d\}$ is a collection of independent random variables. Then*

$$E\left[\prod_{i=1}^{d} X_i\right] = \prod_{i=1}^{d} E[X_i].$$

In particular, $\mathrm{Cov}(X_i, X_j) = 0$ *for all* $1 \le i \ne j \le d$, *and*

$$\mathrm{Var}\left[\sum_{i=1}^{d} X_i\right] = \sum_{i=1}^{d} \mathrm{Var}[X_i].$$

Proof. Without loss of generality we assume $d = 2$ and that (X_1, X_2) is continuous. The case of general d and discrete random variables is very similar. Denote by f_1 and f_2 the probability density functions of X_1 and X_2, respectively. Then by independence, the joint probability density function of (X_1, X_2) is

$$f(x_1, x_2) = f_1(x_1)f_2(x_2).$$

Therefore,

$$\begin{aligned} E[X_1X_2] &= \iint_{\mathbb{R}^2} x_1x_2f(x_1,x_2)\,dx_1dx_2 = \iint_{\mathbb{R}^2} x_1x_2f_1(x_1)f_2(x_2)\,dx_1dx_2 \\ &= \int_{\mathbb{R}} x_1f_1(x_1)\,dx_1 \int_{\mathbb{R}} x_2f_2(x_2)\,dx_2 = E[X_1]E[X_2]. \end{aligned}$$

Furthermore, $\mathrm{Cov}(X_1, X_2) = E[X_1X_2] - E[X_1]E[X_2] = 0$ and the variance formula is simply Remark 5.2 with zero covariances. ■

Definition 5.5. A sequence of random variables $\{X_1, X_2, \ldots\}$ are said to be **independent and identically distributed**, denoted by **iid**, if all the random variables X_i's are independent and have the same distribution (which is equivalent to having the same probability mass function, or probability density function, or cumulative distribution function)

Example 5.11. Let $\{X_1, X_2, \ldots\}$ be a sequence of iid random variables with mean $E[X_1] = \mu$ and variance $\mathrm{Var}[X_1] = \sigma^2$. Define

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

What are the mean and variance of $\overline{X}_n$?

Solution: Since X_i's have the same distribution, they all have the same expected value and variance. That is, $E[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2$ for all i. Therefore, by linearity of expectation,

$$E[\overline{X}_n] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

and by independence

$$\begin{aligned} \mathrm{Var}[\overline{X}_n] &= \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}[X_i] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n}\sigma^2. \end{aligned}$$

These formulae will be used quite frequently when we study statistics.

Example 5.12. *(alternative way to compute the mean/variance of binomial)* Let X be a binomial random variable with distribution $B(n, p)$. Find $E[X]$ and $\mathrm{Var}[X]$.

Solution: One can view X as the total number of "success" in n independent and identical Bernoulli trials. Denote by X_i the outcome of the i-th trial, with $X_i = 1$ if the outcome is a "success" and 0 otherwise. Then

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

Note that all these X_i's are iid Bernoulli random variables with parameter p, which implies that $E[X_i] = p$ and $\mathrm{Var}[X_i] = p(1-p)$. It follows that

$$E[X] = E[X_1] + E[X_2] + \cdots + E[X_n] = np$$

$$\mathrm{Var}[X] = \mathrm{Var}[X_1] + \mathrm{Var}[X_2] + \cdots + \mathrm{Var}[X_n] = np(1-p).$$

Here the independence of X_i's plays a crucial role in the formula of variance (Theorem 5.6). ■

Example 5.13. *(covariance and independence)* Covariance between independent random variables is zero, thanks to Theorem 5.6. Is the converse true? That is, if $\mathrm{Cov}(X, Y) = 0$, can we claim that X and Y are independent?

Solution: The answer is NO in general. There are plenty of counterexamples. Take, for example, a point selected at random from the unit circle centered at the origin. Let (X, Y) denote its coordinates. Then by symmetry $E[XY] = E[X] = E[Y] = 0$ and thus $\mathrm{Cov}(X, Y) = 0$. But X and Y cannot be independent because $X^2 + Y^2 = 1$.

Exercise 5.14. Let X and Y be independent standard normal random variables. What is the distribution of $X^2 + Y^2$? *(Hint: use polar coordinates to evaluate the cumulative distribution)*

5.4 Sum of Independent Normals

Lemma 4.7 shows that a linear transform of a single normal random variable is still normal. Here we establish another very important property that linear combinations of independent normal random variables are still normal.

Theorem 5.7. *Suppose that $X_1, X_2, \ldots, X_d$ are independent normal random variables where X_i is $N(\mu_i, \sigma_i^2)$ for $i = 1, 2, \ldots, d$. Let $c_1, c_2, \ldots, c_d$ be a collection of reals that are not all zero. Then $c_1X_1 + c_2X_2 + \cdots + c_dX_d$ is normal with mean μ and variance σ^2 with*

$$\mu = \sum_{i=1}^{n} c_i\mu_i, \qquad \sigma^2 = \sum_{i=1}^{n} c_i^2\sigma_i^2.$$

Proof. Without loss of generality, we can assume that $c_i = 1$ for every i, since $c_i X_i$ is also normal with mean $c_i\mu_i$ and variance $c_i^2\sigma_i^2$, thanks to Lemma 4.7. It suffices to show for $d = 2$. The general case of d is simply an iteration of $d = 2$. Moreover, we can assume $\mu_1 = \mu_2 = 0$, thanks to Lemma 4.7. By Lemma 5.5, the probability density function for $X_1 + X_2$ is

$$f(x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_1} e^{-s^2/(2\sigma_1^2)} \cdot \frac{1}{\sqrt{2\pi}\sigma_2} e^{-(x-s)^2/(2\sigma_2^2)}\, ds.$$

Observe that

$$\frac{s^2}{2\sigma_1^2} + \frac{(x-s)^2}{2\sigma_2^2} = \frac{1}{2b^2}(s-a)^2 + \frac{x^2}{2\sigma^2}, \quad a = \frac{\sigma_1^2 x}{\sigma^2},\ b = \frac{\sigma_1\sigma_2}{\sigma}$$

and

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}b} e^{-(s-a)^2/(2b^2)}\, ds = 1,$$

since the integrand is the probability density function of $N(a, b^2)$. It follows that

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma_1} \cdot \frac{1}{\sqrt{2\pi}\sigma_2} \cdot e^{-x^2/(2\sigma^2)} \cdot \sqrt{2\pi}b = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/(2\sigma^2)},$$

which is the density function for $N(0, \sigma^2)$. This completes the proof. ■

5.5 Uniform Distributions in High Dimensions

Geometric probability deals with random geometric objects. For example, what is the probability that a fallen needle intersects with any of the equally spaced parallel lines on a floor (*Buffon's needle*)? Or, what is the average distance between two points randomly selected within a disc? Many of these questions involve uniform distributions on domains of finite volume in $\mathbb{R}^d$. The uniform distribution on interval $[a, b]$ is a special case of such distributions with $d = 1$.

From now on, when we say the **"volume"** of a domain $D \subseteq \mathbb{R}^d$, denoted by $\mathrm{Vol}(D)$, we mean the d-dimensional volume. For $d = 1$, 2, and 3, it amounts to length, area, and usual volume, respectively.

Definition 5.6. Let D be a domain in $\mathbb{R}^d$ with finite volume. We say a random vector X is **uniformly distributed on domain** D if X takes values in D and the joint probability density function of X is

$$f(x) = \frac{1}{\mathrm{Vol}(D)} 1_D(x), \quad \forall x \in \mathbb{R}^d.$$

That is, the probability density function is constant on the region D. One of the characteristics of the uniform distribution on domain $D \subseteq \mathbb{R}^d$ is that

$$P(X \in B) = \frac{\text{Vol}(B)}{\text{Vol}(D)}, \quad \forall B \subseteq D$$

which is straightforward from definition. When we say a number of points are chosen from domain D *at random*, it means that these points are independent and uniformly distributed on D, unless otherwise specified.

Lemma 5.8. *Let $D = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_d, b_d]$ be a rectangle in $\mathbb{R}^d$. Then $X = (X_1, X_2, \ldots, X_d)$ is uniform on domain D if and only if*

1. *X_i is uniform on $[a_i, b_i]$ for all $i = 1, 2, \ldots, d$;*
2. *$X_1, X_2, \ldots, X_d$ are independent.*

Proof. The proof is an application of Lemma 5.4. It suffices to observe that

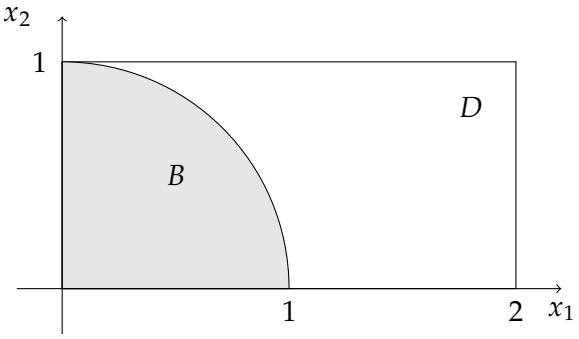
$$\prod_{i=1}^{d} \frac{1}{b_i - a_i} 1_{[a_i, b_i]}(x_i) = \frac{1}{\text{Vol}(D)} 1_D(x),$$

for all $x = (x_1, x_2, \ldots, x_d)$. The claim follows readily. ∎

Example 5.15. We randomly and independently select a point X_1 from $[0, 2]$ and a point X_2 from $[0, 1]$. What is the probability that $X_1^2 + X_2^2 \leq 1$?

Solution: Thanks to Lemma 5.8, the random vector $X = (X_1, X_2)$ is uniform on the rectangle $D = [0, 2] \times [0, 1]$. Denote by B the part of the unit disc in the first quadrant, centered at the origin. Then

$$P(X_1^2 + X_2^2 \leq 1) = P(X \in B) = \frac{\text{Area}(B)}{\text{Area}(D)} = \frac{\pi}{8}$$

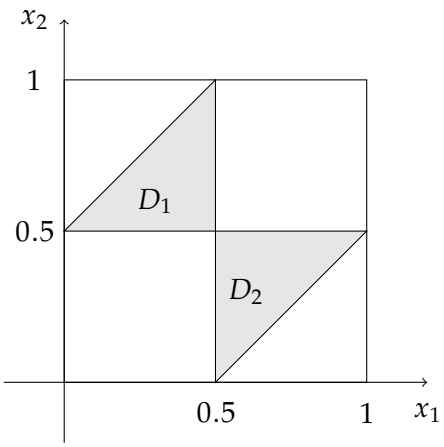


Example 5.16. Choose two points from $[0,1]$ at random. They will divide the interval into three pieces. What is the probability that these three pieces can form a triangle?

Solution: Denote these two points by X_1 and X_2. Then X_1 and X_2 are iid uniform on $[0,1]$. Thanks to lemma 5.8, $X = (X_1, X_2)$ is uniform on the square $D = [0,1] \times [0,1]$. Note that three line segments of length a, b, c can form a triangle if and only if $a < b + c$, $b < a + c$, and $c < a + b$.

1. If $X_1 \leq X_2$, the three line segments have length X_1, $X_2 - X_1$, $1 - X_2$, respectively. It is not hard to verify that these three pieces can form a triangle if and only if $(X_1, X_2) \in D_1$, where $D_1 = \{0 < x_1 < 0.5 < x_2 < 1, x_2 - x_1 < 0.5\}$.

2. If $X_1 > X_2$, the three line segments have length X_2, $X_1 - X_2$, $1 - X_1$, respectively. It is not hard to verify that these three pieces can form a triangle if and only if $(X_1, X_2) \in D_2$, where $D_2 = \{0 < x_2 < 0.5 < x_1 < 1, x_1 - x_2 < 0.5\}$.



In other words, the three pieces can form a triangle if and only if $X \in D_1 \cup D_2$. It follows that the probability of interest is

$$\frac{\text{Area}(D_1 \cup D_2)}{\text{Area}(D)} = \frac{\text{Area}(D_1) + \text{Area}(D_2)}{\text{Area}(D)} = \frac{1}{8} + \frac{1}{8} = \frac{1}{4}.$$

Example 5.17. Randomly pick a point within the d-dimensional unit ball $D = \{x = (x_1, \ldots, x_d) : x_1^2 + \cdots + x_d^2 \leq 1\} \subseteq \mathbb{R}^d$. What is the average distance from this point to the origin?

Solution: Denote the point by X. Then X is uniformly distributed on the unit ball D. Its distance to the origin is $\|X\|$, which takes values in $[0,1]$. Note that for any $r \in [0,1]$, $\|X\| \le r$ if and only if X is within the ball of radius r centered at the origin (denoted by D_r). Therefore, the cumulative distribution function for $\|X\|$ is

$$F(r) = P(\|X\| \le r) = \frac{\mathrm{Vol}(D_r)}{\mathrm{Vol}(D)} = \frac{c_d r^d}{c_d} = r^d, \quad \forall r \in [0,1].$$

Here c_d is some constant (it is the volume of unit ball in $\mathbb{R}^d$). Therefore, the probability density function of $\|X\|$ is

$$f(r) = F'(r) = d r^{d-1} 1_{[0,1]}(r), \quad \forall r \in \mathbb{R},$$

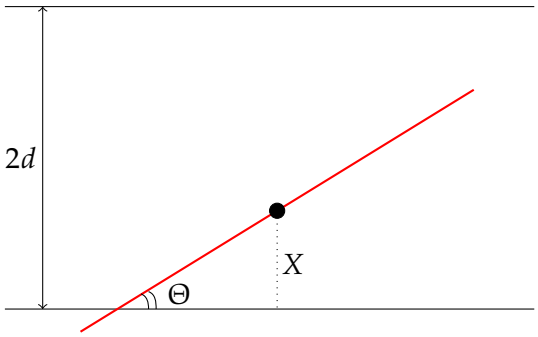
and

$$E[\|X\|] = \int_{\mathbb{R}} r f(r)\, dr = \int_0^1 r \cdot d r^{d-1}\, dr = \frac{d}{d+1}.$$

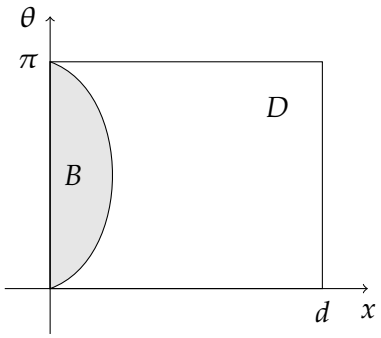
When the dimension d is large, the average distance to the origin is very close to 1. This is because for a ball in the high dimensional Euclidean space, almost all the volume is concentrated near the sphere of the ball. ■

Example 5.18. *(Buffon's needle, 1777)* Imagine a large floor that is marked with parallel lines. These parallel lines are equally spaced with distance $2d$. Drop a needle of length 2ℓ at random onto the floor. Assume $\ell < d$. What is the probability that the needle will intersect with one of the parallel lines?

Solution: First we observe that if the needle intersects with these parallel lines, it can intersect with one and only one of these lines since $\ell < d$. Secondly, the position of the needle is characterized by two quantities: (1) the distance of the middle point of the needle to the closet parallel line (the only parallel line it can intersect, if it ever intersects with any), denoted by X; (2) the angle it forms with the parallel lines, denoted by Θ.



We can assume that X is uniform on $[0, d]$, Θ uniform on $[0, \pi]$, and X and Θ are independent. Thanks to Lemma 5.8, the random vector (X, Θ) is uniform on the rectangle $D = [0, d] \times [0, \pi]$. Note that the needle intersects with the parallel line if and only if $\ell \sin\Theta \geq X$ if and only if $(X, \Theta) \in B$ where $B = \{(x, \theta) \in D : \ x \leq \ell \sin\theta\}$.



Since

$$\text{Area}(B) = \int_0^\pi \ell \sin\theta \, d\theta = 2\ell, \quad \text{Area}(D) = \pi d,$$

it follows that

$$P(\text{intersection}) = \frac{\text{Area}(B)}{\text{Area}(D)} = \frac{2\ell}{\pi d}.$$

Exercise 5.19. *(Buffon's noodle)* Imagine a large floor that is marked with parallel lines. These parallel lines are equally spaced with distance $2d$. There is a rigid "noodle" (a plane curve) of length 2ℓ. But ℓ may or may not be less than d. The noodle can be of any shape. What is the average number of intersections that this noodle makes with the parallel lines when the noodle is dropped at random onto the floor? (*Hint: divide the noodle into small pieces and use the linearity of expectation.*)

5.6 Indicator random variables

Given an arbitrary event $A \subseteq \Omega$, one can define an **indicator random variable** of the form $1_A : \Omega \to \{0, 1\}$ such that

$$1_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise} \end{cases}.$$

For example, a Bernoulli random variable is a type of indicator random variable. It is immediate that

$$E[1_A] = P(A), \quad \text{Var}[1_A] = P(A) - P^2(A).$$

Example 5.20. There are n couples ($2n$ people in total) standing in a circle in random order. What is the average number of couples that are standing next to each other? To clarify, suppose there are 5 couples (labeled A, B, C, D, E) standing like this in a circle.

$$A1,\ C2,\ D1,\ D2,\ E1,\ B2,\ B1,\ C1,\ E2,\ A2$$

Then three couples (A, D, B) are standing next to each other ($A1$ and $A2$ are next to each other because it is indeed a circle).

Solution: To solve this problem, it is the easiest to utilize indicator random variables. Let X be the total number of couples that are standing next to each other. Label these n couples $1, 2, \ldots, n$. Define indicator random variables

$$X_k = \begin{cases} 1 & \text{if the } k\text{-th couple stand next to each other} \\ 0 & \text{otherwise} \end{cases}$$

for every $k = 1, 2, \ldots, n$. It follows that $X = X_1 + X_2 + \cdots + X_n$ and thus

$$E[X] = \sum_{k=1}^{n} E[X_k] = \sum_{k=1}^{n} P(X_k = 1).$$

But consider the event $\{X_k = 1\}$, or equivalently, the event that the k-th couple stand next to each other. No matter where $k1$ stands, as long as $k1$ takes its position, $k2$ will have $(2n - 1)$ choices for its position, 2 out of which will be next to $k1$. Therefore, for any k,

$$P(X_k = 1) = \frac{2}{2n - 1}.$$

It follows that

$$E[X] = \sum_{k=1}^{n} P(X_k = 1) = \frac{2n}{2n - 1}.$$

We would like to remark that in this calculation, one basically breaks X into the sum of much simpler indicator random variables. These indicators are *not* independent and their relations are complicated. However, the linearity of expected value (Part 2 of Lemma 3.1) does not require independence and holds for *any* random variables.

Example 5.21. There are 3 mice and a box with 10 pieces of cheese. You will randomly give two pieces of cheese to each mouse. However, the expiration date on the cheese has passed, and among these 10 pieces of cheese,

4 pieces are spoiled, which will cause food poisoning. What is the average number of mice that will get food poisoning?

Solution: This is a perfect setup for using the indicator random variables. Define

$$X_k = \begin{cases} 1 & \text{if the } k\text{-th mouse gets food poisoning} \\ 0 & \text{otherwise} \end{cases}.$$

Then the total number of mice that get food poisoning is $X = X_1 + X_2 + X_3$. Thus

$$E[X] = \sum_{k=1}^{3} E[X_k] = \sum_{k=1}^{3} P(X_k = 1).$$

Note that for each k, the k-th mouse does *not* get food poisoning if and only if the two pieces of cheese it receives are both unspoiled. The chance that the 1st piece it receives is unspoiled is $6/10$. Given that the 1st piece is unspoiled, the probability that the 2nd piece it receives is also unspoiled is $5/9$. Therefore, by the product rule, $P(X_k = 0) = 6/10 \times 5/9 = 1/3$, and $P(X_k = 1) = 1 - P(X_k = 0) = 2/3$. It follows that

$$E[X] = \sum_{k=1}^{3} P(X_k = 1) = 3 \times \frac{2}{3} = 2.$$

Example 5.22. An absent minded professor signed n letters of recommendation for one of his students and put them randomly into n pre-addressed envelopes. Let X be the number of letters that were put into the right envelope. Find the expected value and variance of X.

Solution: We can write $X = X_1 + X_2 + \cdots + X_n$, where X_i's are indicator random variables defined by

$$X_i \doteq \begin{cases} 1 & \text{if the } i\text{-th letter is put into the right envelope} \\ 0 & \text{otherwise} \end{cases}.$$

Clearly, for each $1 \leq i \leq n$,

$$P(X_i = 1) = \frac{1}{n}, \;\; P(X_i = 0) = \frac{n-1}{n}, \;\; E[X_i] = \frac{1}{n}, \;\; \text{Var}[X_i] = \frac{n-1}{n^2},$$

and for $1 \leq i \neq j \leq n$,

$$\begin{aligned} E[X_i X_j] &= P(X_i = X_j = 1) = \frac{1}{n(n-1)} \\ \text{Cov}(X_i, X_j) &= E[X_i X_j] - E[X_i]E[X_j] = \frac{1}{n^2(n-1)} \end{aligned}$$

It follows from the linearity of expected value and the general variance formula (Remark 5.2) that

$$\begin{aligned} E[X] &= \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} \frac{1}{n} = 1 \\ \mathrm{Var}[X] &= \sum_{i=1}^{n} \mathrm{Var}[X_i] + 2 \sum_{1 \le i < j \le n} \mathrm{Cov}(X_i, X_j) \\ &= n \times \frac{n-1}{n^2} + 2\binom{n}{2} \times \frac{1}{n^2(n-1)} = \frac{n-1}{n} + \frac{1}{n} = 1. \end{aligned}$$

Exercise 5.23. Randomly draw five cards from a deck of 52 cards. Find the expected number of Aces in these five cards. (*Solution:* 5/13).

Exercise 5.24. (hard) n cars are driving along an infinitely long highway with different but constant speeds. Since the highway has only one lane, the cars cannot pass each other. If a faster car catches up with a slower car, it must slow down to the speed of the slower car in the front. With this rule, eventually the cars will form clusters. Find the expected number of clusters of cars. (*Solution:* $1 + 1/2 + \cdots + 1/n$)

5.7 Conditional Distributions

Independence between random variables means that one random variable has no impact on the other – regardless of the value of one random variable, the distribution of the other random variable should remain the same. We cannot always expect such a nice and simple relation. The general relation between random variables are described by conditional distributions.

Recall that the distribution of a random variable X describes $P(X \in B)$ for any subset $B \subseteq \mathbb{R}$. Knowing the distribution of X allows us to compute quantities related to the random variable X, such as the expected values and variances of functions of X. In this sense, distribution is all we need to know about a random variable when it comes to such quantities. The conditional distribution of a random variable X is exactly the same — it describes the distribution of X in the presence of some additional information. For example, suppose you are asked to guess a number that I have chosen at random from interval $[0, 1]$. What would you guess? Of course, $1/2$ is the most natural answer. But what if I tell you that the random number I have chosen is larger than $1/2$, would you still guess $1/2$? Of course not. This extra piece of information alters your conception of the distribution of the random number.

More precisely, the conditional distribution of X, given that some event E happens (additional information), describes $P(X \in B \mid E)$ for any $B \subseteq \mathbb{R}$. From the conditional distribution, we can similarly compute quantities such as the conditional expectations and variances of functions of X, given that E happens. In this sense, conditional distribution can be treated just as a "usual" distribution.

We should explain this via two random variables. The extension to a larger collection of random variables is straightforward. Consider two random variables (X, Y). We will discuss the conditional distribution separately for discrete and continuous cases.

5.7.1 Conditional distribution for discrete random variables

Suppose X and Y take values in $\{x_1, x_2, \ldots\}$ and $\{y_1, y_2, \ldots\}$, respectively. Denote the joint probability mass function by $p(x_i, y_j)$, and the two marginal probability mass functions by $p_X(x_i)$ and $p_Y(y_j)$. Fix any $y \in \{y_1, y_2, \ldots\}$. The **conditional probability mass function of X, given $Y = y$**, is defined by

$$\begin{aligned} p_{X|Y}(x|y) &\doteq P(X = x \mid Y = y) \\ &= \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{p(x, y)}{p_Y(y)}, \qquad \forall x = x_1, x_2, \ldots. \end{aligned}$$

Of course, just like in the case of conditional probability, the conditional probability mass function is undefined when $p_Y(y) = 0$. From now on, we assume $p_Y(y) > 0$.

Lemma 5.9. *The conditional probability mass function $p_{X|Y}(\cdot \mid y)$ defines a true probability mass function on $\{x_1, x_2, \ldots\}$. That is,*

$$p_{X|Y}(x_i \mid y) \geq 0, \qquad \sum_i p_{X|Y}(x_i \mid y) = 1.$$

The proof of this lemma is trivial from the definition of marginal probability mass function. However, the implication is important. It says that conditional distribution is a true probability distribution. All the definitions such as expected value and variance, therefore, can be extended to their respective conditional version without much trouble. For example, the conditional expected value of $h(X)$ given $Y = y$, denoted by $E[h(X) \mid Y = y]$, is simply

$$\sum_i h(x_i) P(X = x_i \mid Y = y) = \sum_i h(x_i) p_{X|Y}(x_i \mid y)$$

for any function $h : \mathbb{R} \to \mathbb{R}$. For another example, the conditional variance of X given $Y = y$, denoted by $\mathrm{Var}[X \mid Y = y]$, is simply

$$\mathrm{Var}[X \mid Y = y] = E[X^2 \mid Y = y] - E^2[X \mid Y = y],$$

where the conditional expectations on the right-hand-side correspond to $h(x) = x^2$ and $h(x) = x$, respectively.

In the above definition of conditional probability mass function, it is conditional on the event of form $\{Y = y\}$. This is not necessary at all. In general, let E be an arbitrary event with $P(E) > 0$. The **conditional probability mass function of X given event E happens**, can be defined in an analogous way:

$$p_{X|E}(x|E) \doteq P(X = x \mid E) = \frac{P(\{X = x\} \cap E)}{P(E)}, \qquad \forall x = x_1, x_2, \ldots.$$

Lemma 5.9 still holds for this general version of conditional probability mass functions.

Example 5.25. Two dice are tossed and the sum of the face values is 8. What is the distribution of the smaller face value (defined as the common face value if the two face values are equal) and its expectation?

Solution: Denote by X and Y the face values from the first and the second dice, respectively. The question is asking for the probability mass function of $Z = \min(X, Y)$ given that the event $E = \{X + Y = 8\}$ happens. Note that Z can take values in $\{1, 2, \ldots, 6\}$. It is not difficult to see that

$$P(E) = \sum_{x=1}^{6} P(X = x, Y = 8 - x) = \sum_{x=2}^{6} \frac{1}{36} = \frac{5}{36}.$$

and

$$\begin{aligned} p_{Z|E}(z \mid E) &= P(Z = z \mid E) = \frac{\{\min(X, Y) = z\} \cap E}{P(E)} \\ &= \begin{cases} 0.4 & \text{if } z = 2 \\ 0.4 & \text{if } z = 3 \\ 0.2 & \text{if } z = 4 \\ 0 & \text{if } z = 1, 5, 6 \end{cases}. \end{aligned}$$

The conditional expected value of $\min(X, Y)$ given that $X + Y = 8$ is

$$E[Z \mid E] = \sum_{z=1}^{6} zP(Z = z \mid E) = 2 \times 0.4 + 3 \times 0.4 + 4 \times 0.2 = 2.8.$$

Example 5.26. Let (X, Y) be a discrete random vector with joint probability mass function

$$P(X = n, Y = k) = \binom{n}{k} \left(\frac{1}{6}\right)^k \left(\frac{1}{3}\right)^{n-k}$$

for every $n = 1, 2, \ldots$ and $k = 0, 1, \ldots, n$. What is the conditional probability mass function of Y given $X = n$?

Solution: Thanks to the binomial expansion, we have

$$\begin{aligned} P(X = n) &= \sum_{k=0}^{n} P(X = n, Y = k) = \sum_{k=0}^{n} \binom{n}{k} \left(\frac{1}{6}\right)^k \left(\frac{1}{3}\right)^{n-k} \\ &= \left(\frac{1}{6} + \frac{1}{3}\right)^n = \left(\frac{1}{2}\right)^n. \end{aligned}$$

Thus

$$P(Y = k | X = n) = \frac{P(Y = k, X = n)}{P(X = n)} = \binom{n}{k} \left(\frac{1}{3}\right)^k \left(\frac{2}{3}\right)^{n-k}.$$

In other words, given $X = n$, Y is Binomial $B(n, 1/3)$. ■

Remark 5.4. Two discrete random variables X and Y are independent if and only if the conditional probability mass function $p_{X|Y}(x \mid y) = p_X(x)$ for every $x \in \{x_1, x_2, \ldots\}$ and $y \in \{y_1, y_2, \ldots\}$, thanks to Lemma 5.4.

Exercise 5.27. Let X be a geometric random variable with parameter p. Let n be a given positive integer. Find the conditional probability mass function of X given $X > n$ and $E[X|X > n]$.

5.7.2 Conditional distribution for continuous random variables

The conditional distribution of a continuous random variable X is defined in a very similar way. Let E be an event with $P(E) > 0$. Then the **conditional cumulative distribution function of X, given E happens**, can be defined by

$$F_{X|E}(x \mid E) = P(X \leq x \mid E) = \frac{P(\{X \leq x\} \cap E)}{P(E)}, \quad \forall x \in \mathbb{R}$$

and the **conditional probability density function of X, given E happens**, is

$$f_{X|E}(x \mid E) = \frac{d}{dx} F_{X|E}(x \mid E).$$

Lemma 5.10. *The conditional probability density function $f_{X|E}(\cdot \mid E)$ defines a true probability density function on $\mathbb{R}$. That is,*

$$f_{X|E}(x \mid E) \geq 0, \qquad \int_{\mathbb{R}} f_{X|E}(x \mid E)\, dx = 1.$$

The proof is again trivial and the implication of the lemma is the same — one can regard the conditional distribution as a true probability distribution.

Example 5.28. Let X be a continuous random variable with probability density function f. Let a be a given constant with $P(X > a) > 0$. What are the conditional density of X given $X > a$ and the tail expectation $E[X \mid X > a]$?

Solution: Denote by F the cumulative distribution function of X. Then the conditional cumulative distribution function of X given $E = \{X > a\}$ happens is

$$F_{X|E}(x \mid E) = P(X \leq x \mid E) = \frac{P(\{X \leq x\} \cap \{X > a\})}{P(X > a)}.$$

Clearly $F_{X|E}(x \mid E) = 0$ if $x \leq a$, and for $x > a$,

$$F_{X|E}(x \mid E) = \frac{P(a < X \leq x)}{P(X > a)} = \frac{F(x) - F(a)}{1 - F(a)}.$$

Therefore, the conditional probability density function of X given $E = \{X > a\}$ happens is

$$f_{X|E}(x \mid E) = \frac{d}{dx} F_{X|E}(x \mid E) = \frac{f(x)}{1 - F(a)} 1_{[a,\infty)}(x),$$

and

$$E[X \mid X > a] = \int_{\mathbb{R}} x f_{X|E}(x \mid E)\, dx = \frac{1}{1 - F(a)} \int_a^\infty x f(x)\, dx.$$

An interesting exercise for the students is to compute this tail expectation for exponential random variables and explain the results via memoryless property. ■

So far, the discussion on conditional probability density functions for continuous random variables is based on the assumption that $P(E) > 0$.

However, recall that for discrete random variables we can define the conditional probability mass function or conditional distribution of X given $Y = y$, via the joint probability mass function. Can we do the same for continuous random variables? To fix ideas, let (X, Y) be a continuous random vector with joint probability density function $f(x, y)$. Similar to the discrete case, we wish to define the conditional density of X, given that $Y = y$. The immediate difficulty is that, as a continuous random variable, $P(Y = y) = 0$. Thus the conditional cumulative distribution function $P(X \leq x \mid Y = y)$ cannot be defined in the classical sense. Is there any way to make sense of it though? If there is, then we can derive the conditional probability density function and consequently all the related quantities.

There are, indeed, a few ways to make sense of such conditional distributions. The most intuitive approach is to use discrete approximation, while the rigorous approach is to consider $P(X \leq x \mid Y = y)$ *simultaneously for all* $y \in \mathbb{R}$ under the measure-theoretic framework. Here we will use the discrete approximation to make sense of this conditional distribution.

Fix any y. Let $[a, b]$ be a very small interval that contains y and denote $\Delta y \doteq b - a$. We will define

$$\begin{aligned} P(X \leq x \mid Y = y) &= \lim_{\Delta y \to 0} P(X \leq x \mid Y \in [a, b]) \\ &= \lim_{\Delta y \to 0} \frac{P(X \leq x, Y \in [a, b])}{P(Y \in [a, b])}. \end{aligned}$$

Observe that

$$P(Y \in [a, b]) = \int_a^b f_Y(t)\, dt \approx f_Y(y)(b - a) = f_Y(y)\Delta y$$

and

$$\begin{aligned} P(X \leq x, Y \in [a, b]) &= \int_{-\infty}^{x} \int_a^b f(s, t)\, dtds = \int_a^b \int_{-\infty}^{x} f(s, t)\, dsdt \\ &\approx \int_a^b \int_{-\infty}^{x} f(s, y) dsdt = (b - a) \int_{-\infty}^{x} f(s, y) ds \\ &= \Delta y \cdot \int_{-\infty}^{x} f(s, y) ds. \end{aligned}$$

It follows that

$$P(X \leq x \mid Y = y) = \lim_{\Delta y \to 0} \frac{P(X \leq x, Y \in [a, b])}{P(Y \in [a, b])} = \int_{-\infty}^{x} \frac{f(s, y)}{f_Y(y)}\, ds.$$

Therefore, the **conditional probability density function of X given $Y = y$** is define to be

$$f_{X|Y}(x \mid y) = \frac{d}{dx} \int_{-\infty}^{x} \frac{f(s,y)}{f_Y(y)}\, ds = \frac{f(x,y)}{f_Y(y)}.$$

Example 5.29. Let (X, Y) be a continuous random vector with joint probability density function

$$f(x,y) = \begin{cases} 2 & \text{if } x \geq 0,\, y \geq 0,\, x+y \leq 1 \\ 0 & \text{otherwise} \end{cases}.$$

For any $y \in (0,1)$, compute $E[X \mid Y = y]$.

Solution: We first compute the conditional density function. For any $y \in (0,1)$, the marginal probability density function for Y is

$$f_Y(y) = \int_{\mathbb{R}} f(x,y)\, dx = \int_0^{1-y} 2\, dx = 2(1-y)$$

and the conditional probability density function of X given $Y = y$ is by definition

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

That is, given $Y = y$, X is uniformly distributed on interval $[0, 1-y]$. Therefore,

$$E[X \mid Y = y] = \frac{1}{2}(1-y).$$

Interested students can also verify this by direct integration. ■

Remark 5.5. Two continuous random variables X and Y are independent if and only if the conditional probability density function $f_{X|Y}(x \mid y) = f_X(x)$ for every x and y, thanks to Lemma 5.4.

Remark 5.6. The conditional distribution of X given that some event E happens or given $Y = y$, be it conditional probability mass function or conditional probability density function, can be extended to the cases where X and/or Y are random vectors in a straightforward fashion. Furthermore, properties such as Remarks 5.4 and 5.5 hold.

Exercise 5.30. Let (X, Y) be a continuous random vector with joint probability density function

$$f(x, y) = e^{-x} 1_{\{0<y<x\}}$$

Fix arbitrarily $x, y > 0$.

1. Find the conditional distribution of Y given $X = x$?
2. Find $E[Y|X = x]$ and $\text{Var}[Y|X = x]$
3. Find $E[X|Y = y]$.
4. Find $P(X \geq 2Y|Y = y)$ and $P(X \geq 2Y|Y \geq y)$.

5.7.3 Law of total probability

The classical law of total probability claims that, given a partition of the sample space Ω, say $\{B_1, B_2, \ldots, B_n\}$,

$$P(A) = \sum_{k=1}^{n} P(A \cap B_k) = \sum_{k=1}^{n} P(A \mid B_k) P(B_k)$$

for any event $A \subseteq \Omega$. When this partition is induced by a random variable, we will have a different version of law of total probability.

Theorem 5.11. (Law of total probability) *Let $A \subseteq \Omega$ be an arbitrary event and X any random variable.*

1. *If X is a discrete random variable taking values in $\{x_1, x_2, \ldots\}$ with probability mass function $p(x_i)$, then*

$$P(A) = \sum_{k} P(A \mid X = x_k) P(X = x_k) = \sum_{k} P(A \mid X = x_k) p(x_k).$$

2. *If X is a continuous random variable with probability density function f, then*

$$P(A) = \int_{\mathbb{R}} P(A \mid X = x) f(x)\, dx.$$

Proof. For Part 1, define $B_k = \{X = x_k\}$ for every k. Then $\{B_1, B_2, \ldots\}$ is a partition of the sample space Ω. By countable additivity (one of the three axioms of probability),

$$P(A) = \sum_{k} P(A \cap B_k) = \sum_{k} P(A \mid B_k) P(B_k).$$

The claim follows readily. Now we show Part 2. Assume that X is continuous with density f. Note that $P(A \mid X = x)$ should be defined in a way similar to conditional probability density function. That is, let $x \in [a, b]$ and $\Delta x = b - a$. Then

$$P(A \mid X = x) = \lim_{\Delta x \to 0} P(A \mid X \in [a, b]).$$

Fix a large integer n. Mark the real line with $x_k = k/n$ where $k = 0, \pm 1, \pm 2, \ldots$. Then $\{x_k\}$ satisfy $\cdots < x_{-2} < x_{-1} < x_0 < x_1 < x_2 < \ldots$ and divide the real line into small pieces of equal length $1/n$. It follows that

$$\begin{aligned}
P(A) &= \sum_{k=-\infty}^{\infty} P(A \cap \{X \in [x_k, x_{k+1})\}) \\
&= \sum_{k=-\infty}^{\infty} P(A \mid X \in [x_k, x_{k+1})) P(X \in [x_k, x_{k+1})) \\
&= \sum_{k=-\infty}^{\infty} \int_{x_k}^{x_{k+1}} P(A \mid X \in [x_k, x_{k+1})) f(x)\, dx \\
&\xrightarrow{n \uparrow \infty} \sum_{k=-\infty}^{\infty} \int_{x_k}^{x_{k+1}} P(A \mid X = x) f(x)\, dx \\
&= \int_{\mathbb{R}} P(A \mid X = x) f(x)\, dx.
\end{aligned}$$

We hasten to add that this proof is only intuitive and by no means rigorous. Everything can be made rigorous under the measure-theoretic framework though. ■

Example 5.31. Toss a coin with $P(H) = \theta$ repeatedly n times, where θ is drawn at random from $[0, 1]$. Let X be the total number of heads. Find $P(X = 0)$ and $P(X = 1)$.

Solution: First of all, X takes values in $\{0, 1, 2, \ldots, n\}$. By law of total probability,

$$P(X = k) = \int_{\mathbb{R}} P(X = k \mid \theta = x) f(x)\, dx, \quad 0 \le k \le n$$

where f is the probability density function of θ, which is uniform on $[0, 1]$. It follows that

$$P(X = k) = \int_0^1 \binom{n}{k} x^k (1 - x)^{n-k} \cdot 1\, dx.$$

In particular, by a change of variable $z = 1 - x$, we have

$$\begin{aligned} P(X=0) &= \int_0^1 (1-x)^n \, dx = \int_0^1 z^n \, dz = \frac{1}{n+1} \\ P(X=1) &= \int_0^1 nx(1-x)^{n-1} dx = \int_0^1 n(z^{n-1} - z^n) \, dz = \frac{1}{n+1} \end{aligned}$$

Interested students may take it as an exercise to prove that

$$P(X=k) = \frac{1}{n+1}, \quad \forall\, 0 \le k \le n.$$

That is, X is (discrete) uniform on $\{0, 1, 2, \ldots, n\}$.

Example 5.32. (*Gambler' ruin*) In each round of betting, a gambler has 50% chance to win \$1 and 50% chance to lose \$1. Suppose the gambler starts with \$$n$. He will leave the game if his total wealth reaches 0 (ruin) or reaches a target amount N, whichever comes first. Assume that $0 < n < N$ are both positive integers, and all the betting rounds are independent. Find the probability of ruin.

Solution: Define x_n to be the probability of ruin, given that the gambler has \$$n$ at present, for all $n = 0, 1, \ldots, N$. By definition it is immediate that $x_0 = 1$ and $x_N = 0$. Given $0 < n < N$, after one betting round, gambler's wealth will become $n \pm 1$ with probability 50% each. Therefore, by law of total probability,

$$x_n = \frac{1}{2} x_{n+1} + \frac{1}{2} x_{n-1}.$$

This implies that $x_{n+1} - x_n = x_n - x_{n-1}$. In other words, $\{x_n\}$ is a classical arithmetic sequence. Given the boundary condition $x_0 = 1$ and $x_N = 0$, it is not hard to see that

$$x_n = x_0 + \frac{x_N - x_0}{N} n = 1 - \frac{n}{N}.$$

5.7.4 Conditional expectation and tower property

Law of total probability is indeed a special case of the so called *law of total expectation* or *tower property*. It is the mathematical justification for the first step analysis we have been using for some of the previous examples such as Examples 1.5 and 1.6. To state the tower property, we need to extend the definition of conditional expectation. Let X and Y be two random variables.

We have defined the conditional expectation $E[X \mid Y = y]$ through conditional probability mass or density function. For example, when (X, Y) are continuous, we have defined the conditional density $f_{X|Y}$ and thus

$$E[X \mid Y = y] = \int_{\mathbb{R}} x f_{X|Y}(x \mid y)\, dx.$$

Regardless, $E[X \mid Y = y]$ is a function of y. Denote it by $h(y) \doteq E[X \mid Y = y]$. Now we define the **conditional expectation of** X **given** Y, denote by $E[X \mid Y]$, to be $h(Y)$. In other words,

1. $E[X \mid Y] = h(Y)$ is a *random variable* and a function of Y.
2. $E[X \mid Y]$ takes value $h(y) = E[X \mid Y = y]$ on the set $\{Y = y\}$.

In short, we pretend that Y is a constant when we work with $E[X|Y]$. The definition of conditional expectation extends naturally to the case where Y is an arbitrary random vector.

Lemma 5.12. *Let X and Z be any random variables, Y any random vector of dimension $d \geq 1$, a and b any constants, and $\ell : \mathbb{R}^d \to \mathbb{R}$ any function. Then*

1. $E[\ell(Y)X \mid Y] = \ell(Y)E[X \mid Y]$.
2. $E[X \mid Y] = E[X]$ *if X and Y are independent.*
3. $E[aX + bZ \mid Y] = aE[X \mid Y] + bE[Z \mid Y]$.

Proof. Note that Part 1 follows directly from the definition of conditional expectation and that for any $y \in \mathbb{R}^d$,

$$E[\ell(Y)X \mid Y = y] = E[\ell(y)X \mid Y = y] = \ell(y)E[X \mid Y = y].$$

As for Part 2, it suffices to observe that the conditional probability mass (density) function of X given $Y = y$ is the same as the probability mass (density) function of X itself, for any $y \in \mathbb{R}^d$; see Remarks 5.6. As for Part 3, it is straightforward from the linearity of classical expectations since conditional probability distributions are true probability distributions. We complete the proof. ∎

Theorem 5.13. (Law of Total Expectation, Tower Property) *Let X be an arbitrary random variable and Y any random vector. Then $E[E[X \mid Y]] = E[X]$.*

Proof. We will give a proof under the assumption that Y is a random variable and (X, Y) is discrete with joint probability mass function $p(x_i, y_j)$. By definition,

$$h(y_j) \doteq E[X \mid Y = y_j] = \sum_{x_i} x_i P(X = x_i | Y = y_j).$$

Therefore, $E[X \mid Y] = h(Y)$ and

$$\begin{aligned} E[E[X \mid Y]] \;&=\; E[h(Y)] \;=\; \sum_{y_j} h(y_j) P(Y = y_j) \\ &= \sum_{y_j} \sum_{x_i} x_i P(X = x_i | Y = y_j) P(Y = y_j) \\ &= \sum_{y_j} \sum_{x_i} x_i P(X = x_i, Y = y_j) \;=\; \sum_{x_i} \sum_{y_j} x_i P(X = x_i, Y = y_j) \\ &= \sum_{x_i} x_i P(X = x_i) \;=\; E[X]. \end{aligned}$$

The proof is similar when (X, Y) is a continuous random vector since all we need to do is to replace sum by integral and probability mass function by probability density function. We complete the proof. ∎

Example 5.33. Suppose the number of automobile accidents in a certain intersection in one week is Poisson distributed with parameter Λ. Further assume that Λ itself is random and varies week from week. If $E[\Lambda] = a$, what is the average number of accidents in a week.

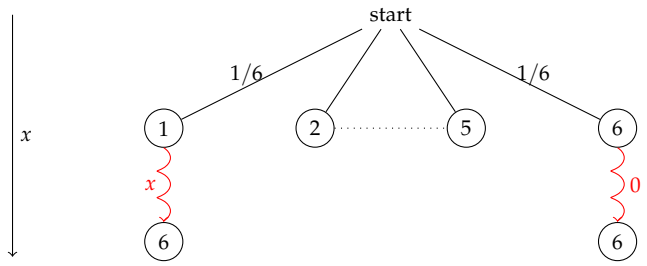
Solution: Let X be the number of accidents in a week. By assumption, given $\Lambda = \lambda$, X is Poisson with parameter λ, whose average is λ. Therefore,

$$E[X | \Lambda = \lambda] = \lambda.$$

It follows that $E[X|\Lambda] = \Lambda$. By law of total expectation, $E[X] = E[E[X|\Lambda]] = E[\Lambda] = a$. ∎

Example 5.34. A fair die (six-sided, each side equally likely) is tossed repeatedly. What is the average cumulative sum of face values until a six is tossed? For example, if the sequence of tosses is $\{3, 5, 4, 4, 2, 1, 2, 6\}$, then the cumulative sum until a six is 27.

Solution: Let x be the average cumulative sum of face values until a six is tossed. Consider the following tree:



Each branch represents a possible outcome from the first toss. For example, if the first toss is a one, then by definition, from the second toss on, the average cumulative sum until a six is tossed is x. Therefore, taking into consideration of the first toss (which is a one), the average cumulative sum of face values is $1 + x$ until the first six. On the other hand, if the first toss is a six, then the cumulative sum of face values until the first six is simply 6. Since each branch is of probability $1/6$, we have

$$x = \frac{1}{6}(1+x) + \frac{1}{6}(2+x) + \cdots + \frac{1}{6}\cdot(6+0) = \frac{7}{2} + \frac{5}{6}x,$$

or $x = 21$. This calculation is indeed the law of total expectation, conditioning on the outcome from the first toss. ■

Example 5.35. Let $X_0, X_1, X_2, \ldots$ be a sequence of iid random variables uniformly distributed on $[0,1]$. Define T to be the first time X_n exceeds X_0. That is,

$$T \doteq \min\{n \geq 1 : \ X_n > X_0\}.$$

Find the probability mass function of T and compute $E[T]$.

Solution: The random variable T takes values in $\{1, 2, 3, \ldots\}$. For $n \geq 1$, by law of total probability,

$$P(T = n) = \int_{-\infty}^{\infty} P(T = n \mid X_0 = x) f(x)\, dx,$$

where f is the probability density function of X_0, which is uniform on $[0,1]$. Thus

$$\begin{aligned} P(T = n) &= \int_0^1 P(T = n \mid X_0 = x)\, dx \\ &= \int_0^1 x^{n-1}(1-x)\, dx = \frac{1}{n} - \frac{1}{n+1}. \end{aligned}$$

Note that in this case

$$P(T < \infty) = \sum_{n=1}^{\infty} P(T = n) = \sum_{n=1}^{\infty} \left(\frac{1}{n} - \frac{1}{n+1} \right) = 1.$$

That is, with probability one, X_0 will be exceeded by some X_n. However,

$$E[T] = \sum_{n=1}^{\infty} nP(T = n) = \sum_{n=1}^{\infty} n \left(\frac{1}{n} - \frac{1}{n+1} \right) = \sum_{n=1}^{\infty} \frac{1}{n+1} = \infty,$$

which means the average time until exceedance of X_0 is infinity!

Exercise 5.36. In Example 5.35, what is $E[T|X_0]$? Use this and law of total expectation to find another proof of $E[T] = \infty$.

Exercise 5.37. Suppose X and Y are two random variables such that, given $X = x$, Y is normally distributed with mean x and variance 1. Suppose $E[X] = \mu$ and $\text{Var}[X] = \sigma^2$. Find $\text{Cov}(X, Y)$ and $\text{Var}[Y]$.

Exercise 5.38. (Variance Decomposition) Let X be any random variable and Y any random vector. Show that

$$\text{Var}[X] = \text{Var}(E[X \mid Y]) + E[\text{Var}(X \mid Y)].$$

5.8 Multivariate Normal Distributions

In this section, we briefly discuss multivariate normal distributions and jointly normal random vectors. We will omit most of the proofs, since they are either too technical or require advanced knowledge in probability and linear algebra. We will adopt the convention that vectors are in column form by default, and A^t denotes the transpose of A.

Definition 5.7. Let $X = (X_1, X_2, \ldots, X_d)^t$ be a d-dimensional random vector. We say X has a **multivariate normal distribution** or **jointly normal distribution**, or X_i's are **jointly normal random variables** if either one of the following equivalent conditions holds:

1. Every nontrivial linear combination of the components is a univariate normal random variable, that is, $c_1X_1 + \cdots + c_dX_d$ is a normal random variable for every collection of constants $\{c_1, c_2, \ldots, c_d\} \subseteq \mathbb{R}$ that are not all zero.

2. There exist a $d \times d$ invertible square matrix A, a $d \times 1$ vector μ, and a family of iid standard normal random variables $\{Z_1, \ldots, Z_d\}$, such that $X = AZ + \mu$, where $Z = (Z_1, Z_2, \ldots, Z_d)^t$.

3. There exist a $d \times 1$ vector μ and a $d \times d$ symmetric positive definite matrix Σ such that the probability density function of X is

$$f(x) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left\{-\frac{1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu)\right\}$$

for all $x \in \mathbb{R}^d$. We denote this distribution by $N(\mu, \Sigma)$.

Theorem 5.14. *Assume that $X = (X_1, X_2, \ldots, X_d)^t$ is a multivariate normal random vector with distribution $N(\mu, \Sigma)$. Then μ and Σ are the mean vector and covariance matrix of X, respectively. That is,*

$$\mu = \begin{bmatrix} EX_1 \\ EX_2 \\ EX_3 \\ \vdots \\ EX_d \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \mathrm{Cov}(X_1, X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_d) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_2, X_2) & \cdots & \mathrm{Cov}(X_2, X_d) \\ \mathrm{Cov}(X_3, X_1) & \mathrm{Cov}(X_3, X_2) & \cdots & \mathrm{Cov}(X_3, X_d) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_d, X_1) & \mathrm{Cov}(X_d, X_2) & \cdots & \mathrm{Cov}(X_d, X_d) \end{bmatrix}$$

Exercise 5.39. Consider any random vector $X = (X_1, X_2, \ldots, X_d)^t$ and let Σ denote its covariance matrix. Show that Σ is symmetric and positive semidefinite, that is, $v^t \Sigma v \geq 0$ for any nonzero vector v.

Theorem 5.15. *Let $X = (X_1, X_2, \ldots, X_d)^t$ be a jointly normal random vector with mean vector $\mu = [\mu_i]$ and covariance matrix $\Sigma = [\Sigma_{ij}]$.*

1. *X_i is normal with mean μ_i and variance Σ_{ii}.*

2. *X_i and X_j are independent if and only if $\mathrm{Cov}(X_i, X_j) = \Sigma_{ij} = 0$.*

3. ***(linear combination of jointly normal random variables is normal)** For any nonzero vector $c = [c_i]_{d \times 1}$ and constant b, the random variable $c^t X + b$ is normal with mean $c^t \mu + b$ and variance $c^t \Sigma c$.*

4. ***(linear combinations of jointly normal random variables are jointly normal)** Let A be an $n \times d$ matrix with $n \leq d$ and $\mathrm{rank}(A) = n$. Let b be an $n \times 1$ vector. Then $Y = AX + b$ is multivariate normal with mean $A\mu + b$ and covariance matrix $A\Sigma A^t$.*

Example 5.40. Let $X = (X_1, X_2, \ldots, X_d)^t$ be a random vector. Show that $X_1, X_2, \ldots, X_d$ are iid standard normal random variables if and only if X is multivariate normal with distribution $N(0, I_d)$, where I_d stands for the $d \times d$ identity matrix.

Proof. Let φ be the probability density function for standard normal distribution. Then $X_1, X_2, \ldots, X_d$ are iid standard normal if and only if the joint probability density function of $X_1, X_2, \ldots, X_d$ is

$$f(x) = \varphi(x_1)\varphi(x_2)\cdots\varphi(x_d)$$

if and only if f takes the form of the joint probability density function in Definition 5.7 with $\mu = 0$ and $\Sigma = I_d$. We complete the proof. ■

Example 5.41. Let Z_1, Z_2 be iid standard normal random variables. Let $X_1 = Z_1 + 2Z_2$ and $X_2 = 3Z_1 + 4Z_2 + 1$. What is the distribution of $X = (X_1, X_2)^t$?

Solution: Thanks to Example 5.40, the random vector $(Z_1, Z_2)^t$ is jointly normal. Therefore, as linear combinations of Z_1 and Z_2, X is also jointly normal, thanks to Theorem 5.15. To figure out its distribution, it suffices to figure out the mean and covariance matrix of X. It is not hard to see that

$$\mu_1 = E[X_1] = 0, \;\; \mu_2 = E[X_2] = 1$$

$$\Sigma_{11} = \mathrm{Var}[X_1] = 1^2 + 2^2 = 5, \quad \Sigma_{22} = \mathrm{Var}[X_2] = 3^2 + 4^2 = 25$$

$$\Sigma_{12} = \Sigma_{21} = \mathrm{Cov}(X_1, X_2) = 1 \cdot 3 + 2 \cdot 4 = 11.$$

Therefore, X is jointly normal with distribution $N(\mu, \Sigma)$ where

$$\mu = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 5 & 11 \\ 11 & 25 \end{bmatrix}.$$

Exercise 5.42. Let $X = (X_1, X_2)^t$ be a bivariate normal random vector with distribution $N(\mu, \Sigma)$ with

$$\mu = \begin{bmatrix} 3 \\ 4 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix},$$

where $-1 < \rho < 1$. Find the distribution of $X_1 - X_2$.

Exercise 5.43. *(Cholesky factorization)* Let (X, Y) be a bivariate normal random vector with distribution $N(\mu, \Sigma)$, where

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}, \quad \rho \in (-1, 1).$$

Show that there exist independent standard normal random variables Z_1, Z_2 such that

$$X = \sigma_1 Z_1, \quad Y = \sigma_2(\rho Z_1 + \sqrt{1-\rho^2} Z_2).$$

Exercise 5.44. Let X be a standard normal random variable and Y an independent discrete random variable with

$$P(Y = \pm 1) = \frac{1}{2}.$$

Show that $Z = XY$ is a standard normal random variable and $\mathrm{Cov}(X, Z) = 0$. Are Z and X independent? Is the random vector $(X, Z)^t$ a bivariate normal random vector?

Exercise 5.45. Consider any random vector $X = (X_1, X_2, \ldots, X_d)^t$ and let Σ denote its covariance matrix. Let C be a $m \times d$ matrix of constants and define $Y = CX$. Then Y is a $m \times 1$ random vector. Let Π be the covariance matrix for Y. Show that $E[Y] = CE[X]$ and $\Pi = C\Sigma C^t$.

Chapter 6

Limit Theorems

Two most fundamental theorems in probability are unquestionably the *Strong Law of Large Numbers* (SLLN) and *Central Limit Theorem* (CLT). Both of them are concerned with

$$\overline{X}_n \doteq \frac{X_1 + X_2 + \cdots + X_n}{n}$$

where $\{X_1, X_2, \ldots\}$ is a sequence independent identically distributed (iid) random variables. Our interest lies in the behavior of $\overline{X}_n$ as $n \to \infty$.

6.1 Strong Law of Large Numbers

The limit of $\overline{X}_n$ is easy to guess. Denote $\mu = E[X_1]$ and $\sigma^2 = \mathrm{Var}[X_1]$. Recall the mean and variance formula in Example 5.11:

$$E[\overline{X}_n] = \mu, \quad \mathrm{Var}[\overline{X}_n] = \frac{\sigma^2}{n}.$$

Clearly as $n \to \infty$, $\mathrm{Var}[\overline{X}_n] \to 0$. That is, $\overline{X}_n$ is a random variable with less and less variation as n gets larger and larger. Thus, as $n \to \infty$, we expect that $\overline{X}_n$ is centered around μ with overwhelming probability. The strong law of large numbers makes this precise.

Theorem 6.1. (Strong Law of Large Numbers, SLLN) *Let $\{X_1, X_2, \ldots\}$ be a sequence of iid random variables with $E[X_1] = \mu$. Then with probability one,*

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

The "*with probability one*" convergence in SLLN is not that different from the usual pointwise convergence. To be precise, let Ω be the underlying sample space for all those $\{X_1, X_2, \ldots\}$. Then "$\overline{X}_n \to \mu$ *with probability one*" means

$$P\left(\left\{\omega \in \Omega : \overline{X}_n(\omega) = \frac{X_1(\omega) + X_2(\omega) + \cdots + X_n(\omega)}{n} \to \mu\right\}\right) = 1.$$

In other words, the set of those ω's on which $\overline{X}_n(\omega)$ converges to μ has probability one. Here are a few simulated paths of $\overline{X}_n$ when X_i's are iid exponential random variables with mean 1. One can see that as n gets larger, $\overline{X}_n$ appears to be converging to the mean 1.

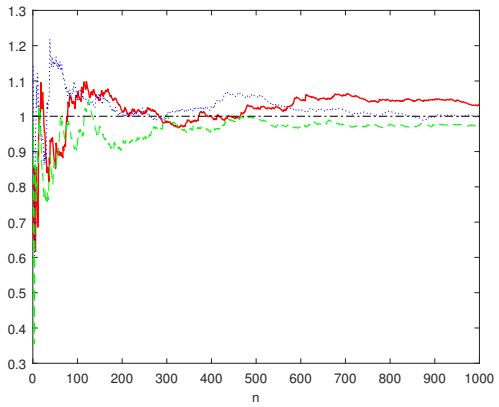


Figure 6.1: Simulated Path for SLLN

Even though we are not well equipped to prove the SLLN, it still helps to get a rough idea of how close $\overline{X}_n$ is to the mean μ. To this end, we will introduce a very basic, but highly effective, inequality.

Lemma 6.2. (**Chebychev Inequality**) *Let X be any nonnegative random variable. Then for any $a > 0$, we have*

$$P(X \geq a) \leq \frac{E[X]}{a}.$$

Proof. Define a new random variable $Y = a1_{\{X \geq a\}}$. Then $X \geq Y$, and thus $E[X] \geq E[Y]$. But $E[Y] = aP(X \geq a)$. We complete the proof. ■

Applying the Chebychev inequality to the random variable $(\overline{X}_n - \mu)^2$, we arrive at

$$P\left(|\overline{X}_n - \mu| \geq \varepsilon\right) = P\left((\overline{X}_n - \mu)^2 \geq \varepsilon^2\right) \leq \frac{\mathrm{Var}[\overline{X}_n]}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2},$$

for any given constant $\varepsilon > 0$. In particular, it yields that $P(|\overline{X}_n - \mu| \geq \varepsilon) \to 0$ as $n \to \infty$. Of course this convergence is not exactly strong law of large numbers (interestingly, it is called the *weak law of large numbers*). But it gives us at least some rough idea.

6.2 Central Limit Theorem

There have been many versions of central limit theorem in the history. But the earliest work was widely attributed to the French mathematician Abraham de Moivre in 1733, for binomial distributions.

Theorem 6.3. (Central Limit Theorem, CLT) *Let $\{X_1, X_2, \ldots\}$ be a sequence of iid random variables with $E[X_1] = \mu$ and $\mathrm{Var}[X_1] = \sigma^2$. Then*

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sqrt{n}\sigma} \to N(0,1).$$

The convergence is in the sense of distribution. That is, for every $x \in \mathbb{R}$,

$$P\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leq x\right) \to \Phi(x)$$

where Φ is the cumulative distribution function for the standard normal $N(0,1)$.

The strong law of large numbers shows us the limit of $\overline{X}_n$ to be μ. But how fast is this convergence? The answer is provided by the central limit theorem. There are two conclusions we can make from the CLT:

1. $\overline{X}_n$ converges to μ in the order of $1/\sqrt{n}$;
2. The deviation of $\overline{X}_n$ from mean μ is approximately normal.

You will often see the central limit theorem expressed in a slightly different form:

$$\overline{X}_n \approx \mu + \frac{\sigma}{\sqrt{n}} N(0,1) = N\left(\mu, \frac{\sigma^2}{n}\right).$$

Example 6.1. (*simulation*) We simulate $\overline{X}_n$ when X_i's are iid exponential random variables with mean $\mu = 1$ (hence variance $\sigma^2 = 1$ as well). We take $n = 100$ and $n = 1000$ respectively. For each case of n, we have generated 1000 samples of $\overline{X}_n$, where each sample of $\overline{X}_n$ is the mean of n iid samples of exponential with mean 1. We then plot the histogram of these 1000 samples of $\overline{X}_n$ and superimpose the probability density function of $N(\mu, \sigma^2/n)$ on top it.

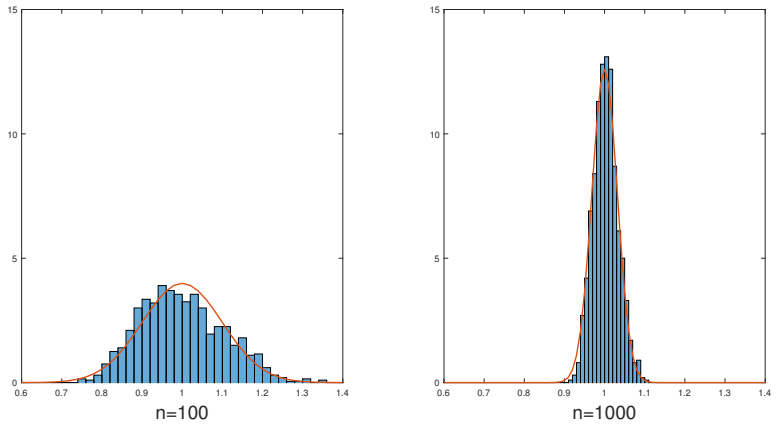


Figure 6.2: CLT simulation from exponentials

Example 6.2. A fair die is tossed repeatedly 100 times. What is the probability that the average face value exceeds 3?

Solution: Let X_i be the face value from the i-th toss. Then X_i's are iid uniform on $\{1, 2, \ldots, 6\}$ with mean $\mu = E[X_i] = 3.5$ and variance $\sigma^2 = 35/12$. The average face value of 100 tosses is

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

with $n = 100$. By the central limit theorem,

$$\overline{X}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right) = N(3.5, 0.17^2).$$

Therefore, using the standardization of normal distributions (Lemma 4.7) ,

we have

$$\begin{aligned} P(\overline{X}_n > 3) &\approx P\left(N(3.5, 0.17^2) > 3\right) = P\left(N(0,1) > \frac{3-3.5}{0.17}\right) \\ &= 1 - \Phi(-2.94) = 1 - 0.0016 = 0.9984. \end{aligned}$$

6.3 Normal Approximation for Binomials

A binomial random variable can be written as the sum of iid Bernoulli random variables. Let X be binomial with distribution $B(n,p)$. If we denote by Y_i the outcome of the i-th trial (1 for "success" and 0 for "failure" as usual), then Y_i's are iid with $E[Y_i] = p$, $\mathrm{Var}[Y_i] = pq$ with $q \doteq 1-p$, and

$$X = Y_1 + Y_2 + \cdots + Y_n.$$

By the central limit theorem, *when n is large*, we have the normal approximation:

$$\frac{X}{n} \approx N\left(p, \frac{pq}{n}\right) \quad \text{or} \quad X \approx N(np, npq).$$

Thus, the computation of binomial distributions can be converted into that of normal distributions, which is in general much more succinct.

Example 6.3. Toss a fair coin 100 times. Let X be the total number of heads. Compute the following probabilities: $P(X \le 50)$, $P(X = 50)$, and $P(X \ge 60)$.

Solution: X is $B(100, 0.5)$ and thus approximately $N(50, 5^2)$ by the central limit theorem. It follows that

$$\begin{aligned} P(X \le 50) &\approx P(N(50,5^2) \le 50) = 0.5 \\ P(X = 50) &\approx P(N(50,5^2) = 50) = 0 \\ P(X \ge 60) &\approx P(N(50,5^2) \ge 60) = P(N(0,1) \ge 2) = 2.3\%. \end{aligned}$$

Remark 6.1. There are a couple of scenarios where normal approximations will not work well. One scenario is when n is not large enough. A rule of thumb for n to be considered large is when

$$n > 9 \cdot \frac{\max(p,q)}{\min(p,q)}.$$

Another scenario where we have to be careful is when the probability is about extreme events (for example, events that X is close to 0 or n). Normal approximation usually does not work well for such events.

Chapter 7

Basics of Point Estimation

Statistical inference is closely related to the theory of probability. In the latter, we study the probabilistic behavior of the outcomes of stochastic systems based upon various assumptions and conditions. In contrast, the goal of statistical inference is kind of opposite in the sense that, *after observing the outcomes of a stochastic system*, what can we infer about the assumptions and conditions we have imposed on the system.

There are two large schools of philosophically different approaches toward statistical inference. On one side, the *frequentist statistics*; on the other, the *Bayesian statistics*. Both follow the identical rules of probability. They differ, however, in their interpretation of *uncertainty*. Take, for example, coin tossing. A coin with unknown probability of heads (say θ) is tossed n times and we observe x heads. How do we estimate θ?

- *Frequentist Point of View:* OK, let us *ignore the data* (observing x heads) at the moment. The probability of heads θ is unknown but fixed. If we toss the coin n times, the number of heads X should be binomial $B(n, \theta)$. A good estimate for θ should be

  $$\hat{\theta} = \frac{X}{n},$$

  because we can prove all sorts of nice properties about it. We can also characterize the error associated with this estimate (bias and variance, mean square error, and so on), tossing in a few confidence intervals into the mix. *All of these derivations are based on X, not x.* This is important because it means that all these estimates/confidence intervals are functions of X, thus random variables themselves!

Once all this is done, we pick up the data x and plug in $X = x$ to report estimates and confidence intervals, while keeping in mind that the specific values in our report are a realization or instance of our true estimates and confidence intervals, which are all random variables.

- *Bayesian Point of View:* OK, we have a coin with unknown but fixed probability of heads θ. But before we do any analysis, let us quantify what we mean by "unknown". For someone (Person A) who is clueless about θ, he may perceive that θ could be of any value between 0 and 1, and no value can be more likely or less likely than the other. For another person (Person B) who thinks that the coin is more likely to be fair than otherwise, he may perceive that θ could be any value between 0 and 1, but it is more likely to be around 0.5.

  For Person A, one can depict his belief of the unknown parameter θ by a uniform distribution on $[0, 1]$ — no point is more or less likely than another. For Person B, his belief of θ can be represented by a different probability density function that has higher values around 0.5. In other words, a person's belief or uncertainty of the unknown parameter θ can be summarized by some *probability distribution* on θ. We should emphasize that this "probability distribution" is not that the true value of θ is random, but represents our "*subjective degree of belief*" on the unknown (but fixed) parameter θ. It can be viewed as our prior knowledge about θ, and is said to be the **prior distribution**. For this reason, θ will be treated *as if* it were a random variable in the consequent analysis. The randomness is not from the true value of θ, but represents our perception of θ. Also for this reason, the probability in Bayesian statistics is said to be *subjective*.

  Now we can take into consideration of the data. The coin has been tossed n times and we have observed that $X = x$ heads. This of course will change, or more precisely, update our perception of the parameter θ. Our new degree of belief of θ can be summarized by $f(\theta|X = x)$, the conditional probability density function of θ given $X = x$, which is said to be the **posterior distribution** of θ. We can now report our estimate of θ with this posterior distribution and/or any of its summaries such as the posterior mean of θ, posterior mode, credible intervals (Bayesian version of confidence intervals), and so on.

  Some people criticize the Bayesian approach, because of its subjec-

tive nature in philosophy and that the inference or the posterior distribution is affected by the prior distribution (which could differ by different investigators), besides the data. But it can be proved that, with more and more data, the impact of prior becomes weaker and eventually vanishes.

Our introduction to statistical inference will be frequentist. We will cover point estimation, confidence intervals, hypothesis testing, and linear regression.

7.1 Basic Setup and Terminologies

In classical point estimation, the observations or the data are often assumed to be a realization or instance of a sequence of **samples** $\{X_1, X_2, \ldots, X_n\}$, which are iid random variables from a common probability distribution $f_\theta(x)$. The number n is said to be the **sample size**. The probability distribution $f_\theta(x)$, which can be in the form of a probability mass function, probability density function, or cumulative distribution function, or as such, is said to be the **population distribution**. The parameter of primary interest, θ, unknown but fixed, and often the subject of estimation and/or testing, is said to be the **target parameter** or **population parameter**.

Definition 7.1. An **estimator** for a population parameter θ is a function of samples, designed for the purpose of estimating θ. It is often denoted by

$$\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n).$$

It is a *random variable* by definition.

1. We say $\hat{\theta}$ is **consistent** if $\hat{\theta} \to \theta$ with probability one, as the sample size $n \to \infty$.
2. The **bias** of $\hat{\theta}$ is defined to be $E[\hat{\theta}] - \theta$.
3. We say $\hat{\theta}$ is **unbiased** if $E[\hat{\theta}] = \theta$.
4. The **mean square error** of $\hat{\theta}$ is defined to be $E[(\hat{\theta} - \theta)^2]$.

Lemma 7.1. *Denote by $B(\hat{\theta})$ the bias of $\hat{\theta}$. Then the mean square error of $\hat{\theta}$, denoted by $MSE(\hat{\theta})$, satisfies*

$$MSE(\hat{\theta}) = [B(\hat{\theta})]^2 + \mathrm{Var}[\hat{\theta}].$$

Proof. Denote $a = E[\hat{\theta}]$. Then $B(\hat{\theta}) = a - \theta$. By definition, $\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$. Therefore,

$$\begin{aligned} \text{MSE}(\hat{\theta}) &= E[(\hat{\theta} - a + a - \theta)^2] \\ &= E[(\hat{\theta} - a)^2] + 2(a - \theta)E[\hat{\theta} - a] + (a - \theta)^2. \end{aligned}$$

The first summand on the right-hand-side is $\text{Var}[\hat{\theta}]$; the second summand is 0 since $E[\hat{\theta}] = a$; the last summand is $[B(\hat{\theta})]^2$ by definition. We complete the proof. ■

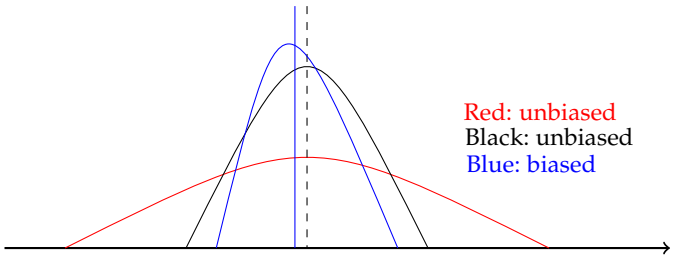
7.2 Comparisons of Estimators

There can be various estimators for a single population parameter. Therefore, the comparison of these estimators becomes an important topic. Definition 7.1 has introduced a few concepts or criteria in that direction. Consistency, bias, and mean square error all measure the error of an estimator from the target parameter θ, in one way or another. Among them, consistency is a rather weak criterion.

The more interesting discussion is between bias and mean square error. The error of an estimator is largely determined by two quantities: bias and variance, because

$$\hat{\theta} - \theta = (\hat{\theta} - E[\hat{\theta}]) + \text{Bias}(\hat{\theta}).$$

In that sense, mean square error takes both into consideration. In practice, an estimator with smaller mean square error is more desirable. In particular, if two estimators are both unbiased, then the one with the smaller variance is preferred.



Example 7.1. We wish to estimate the probability of heads $P(H) = \theta$ of a coin. To this end, we toss the coin n times. Denote by X_i the outcome of the

i-th toss, with 1 for heads and 0 for tails. Thus all those $\{X_1, X_2, \ldots, X_n\}$ are iid Bernoilli random variables with parameter θ. That is,

$$P(X_i = 1) = \theta, \quad P(X_i = 0) = 1 - \theta.$$

For the following estimators of θ, determine if they are unbiased and consistent, and evaluate their mean square errors.

$$\hat{\theta}_1 \doteq \frac{X_1 + X_2 + \cdots + X_n}{n}, \quad \hat{\theta}_2 \doteq X_n, \quad \hat{\theta}_3 \doteq \frac{X_1 + X_2 + \cdots + X_n}{n+1}.$$

Solution: The consistency of $\hat{\theta}_1$ and $\hat{\theta}_3$ is implied by the strong law of large numbers. $\hat{\theta}_2$ is not consistent obviously. Let $B(\hat{\theta})$ denote the bias of an estimator. Then

$$B(\hat{\theta}_1) = E[\hat{\theta}_1] - \theta = 0, \quad B(\hat{\theta}_2) = E[\hat{\theta}_2] - \theta = 0,$$

$$B(\hat{\theta}_3) = E[\hat{\theta}_3] - \theta = \frac{n\theta}{n+1} - \theta = \frac{-\theta}{n+1}.$$

Thus $\hat{\theta}_1$ and $\hat{\theta}_2$ are unbiased, but $\hat{\theta}_3$ is not. We will summarize the results in the following table.

	consistent	not consistent
biased	$\hat{\theta}_3$	so many
unbiased	$\hat{\theta}_1$	$\hat{\theta}_2$

The take-away message from this table is that bias and consistency has no definitive relation. Finally, thanks to Lemma 7.1,

$$\mathrm{MSE}(\hat{\theta}_1) = [B(\hat{\theta}_1)]^2 + \mathrm{Var}[\hat{\theta}_1] = \mathrm{Var}[\hat{\theta}_1] = \frac{\theta(1-\theta)}{n}$$

$$\mathrm{MSE}(\hat{\theta}_2) = [B(\hat{\theta}_2)]^2 + \mathrm{Var}[\hat{\theta}_2] = \mathrm{Var}[\hat{\theta}_2] = \theta(1-\theta)$$

$$\mathrm{MSE}(\hat{\theta}_3) = [B(\hat{\theta}_3)]^2 + \mathrm{Var}[\hat{\theta}_3] = \frac{\theta^2}{(n+1)^2} + \mathrm{Var}[\hat{\theta}_3] = \frac{\theta^2 + n\theta(1-\theta)}{(n+1)^2}.$$

An interesting exercise is to compare the mean square errors of $\hat{\theta}_1$ and $\hat{\theta}_3$. It will turn out that none is always better than the other. ■

Example 7.2. Let $X_1, X_2, \ldots, X_n$ be iid samples uniformly distributed on $[0, \theta]$. We wish to estimate θ. Consider two estimators:

$$\hat{\theta}_1 = \frac{2(X_1 + X_2 + \cdots + X_n)}{n}, \quad \hat{\theta}_2 = \frac{n+1}{n} \max(X_1, X_2, \ldots, X_n).$$

Show that both estimators are unbiased but $\mathrm{MSE}(\hat{\theta}_2) \leq \mathrm{MSE}(\hat{\theta}_1)$. Find an estimator whose mean square error is less than $\hat{\theta}_2$'s.

Solution: Recall that $E[X_i] = \theta/2$ and $\mathrm{Var}[X_i] = \theta^2/12$. Therefore $\hat{\theta}_1$ is unbiased with

$$\mathrm{Var}[\hat{\theta}_1] = \frac{4}{n^2} \cdot \frac{n\theta^2}{12} = \frac{\theta^2}{3n}.$$

To analyze $\hat{\theta}_2$, denote $L = \max(X_1, X_2, \ldots, X_n)$. We first find the probability density function of L. Note that L takes values in $[0, \theta]$. Fix any $x \in [0, \theta]$. We have

$$P(L \leq x) = P(X_1 \leq x, X_2 \leq x, \ldots, X_n \leq x) = \prod_{i=1}^{n} P(X_i \leq x) = \left(\frac{x}{\theta}\right)^n.$$

Taking derivative with respect to x, we arrive at the probability density function of L:

$$f(x) = \frac{nx^{n-1}}{\theta^n} 1_{[0,\theta]}(x),$$

from which it follows that

$$E[L] = \int_0^\theta x f(x)\, dx = \frac{n\theta}{n+1}, \quad E[L^2] = \int_0^\theta x^2 f(x)\, dx = \frac{n\theta^2}{n+2},$$

and thus

$$E[\hat{\theta}_2] = \frac{n+1}{n} E[L] = \theta, \quad \mathrm{Var}[\hat{\theta}_2] = \frac{(n+1)^2}{n^2} \mathrm{Var}[L] = \frac{\theta^2}{n(n+2)}.$$

Therefore $\hat{\theta}_2$ is unbiased and $\mathrm{MSE}[\hat{\theta}_2] = \mathrm{Var}[\hat{\theta}_2] \leq \mathrm{Var}[\hat{\theta}_1] = \mathrm{MSE}[\hat{\theta}_1]$.

It remains to construct an estimator $\hat{\theta}$ with a smaller mean square error than $\hat{\theta}_2$. An interesting fact, though it has no impact in our construction, is that $\hat{\theta}_2$ has the smallest variance among all unbiased estimators for θ. Thus $\hat{\theta}$ will necessarily be biased! Take $\hat{\theta} = c\hat{\theta}_2$ for some constant c to be determined. Then

$$B(\hat{\theta}) = E[\hat{\theta}] - \theta = E[c\hat{\theta}_2] - \theta = (c-1)\theta.$$

and

$$\mathrm{Var}[\hat{\theta}] = \mathrm{Var}[c\hat{\theta}_2] = \frac{c^2\theta^2}{n(n+2)}.$$

Therefore,

$$\mathrm{MSE}(\hat{\theta}) = [B(\hat{\theta})]^2 + \mathrm{Var}[\hat{\theta}] = \theta^2\left[(c-1)^2 + \frac{c^2}{n(n+2)}\right].$$

Note that we recover the mean square error of $\hat{\theta}_2$ if we take $c = 1$. Thus all we need to do is to find c that minimizes the right-hand-side (a quadratic function of c), which will automatically yield an estimator with a smaller mean square error than $\hat{\theta}_2$. This is rather simple. Taking derivative with respect to c and setting it to 0, we have

$$2(c-1) + \frac{2c}{n(n+2)} = 0 \quad \text{or} \quad c = \frac{n(n+2)}{(n+1)^2}.$$

The corresponding estimator is

$$\hat{\theta} = c\hat{\theta}_2 = \frac{n+2}{n+1}\max(X_1, X_2, \ldots, X_n).$$

7.3 Estimators for Population Mean and Variance

It is a common task in statistics to estimate population mean and variance. Let $\{X_1, X_2, \ldots, X_n\}$ be iid samples from some population distribution with mean μ and variance σ^2. The **sample mean** and **sample variance** are defined by

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}, \quad S^2 \doteq \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2,$$

respectively.

Lemma 7.2. *Let $\{X_1, X_2, \ldots\}$ be iid random variables with $E[X_1] = \mu$ and $\mathrm{Var}[X_1] = \sigma^2$. Then $\overline{X}_n$ and S^2 are unbiased estimators for population mean μ and population variance σ^2, respectively. Furthermore, $\overline{X}_n$ and S^2 are both consistent. That is,*

$$\overline{X}_n \to \mu, \quad S^2 \to \sigma^2$$

with probability one, as $n \to \infty$.

Proof. Observe that $E[\overline{X}_n] = \mu$ trivially and $\overline{X}_n \to \mu$ by SLLN. Thus $\overline{X}_n$ is an unbiased and consistent estimator for μ. For the sample variance S^2, we have

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 &= \sum_{i=1}^{n} X_i^2 - 2\overline{X}_n \sum_{i=1}^{n} X_i + n\overline{X}_n^2 \\ &= \sum_{i=1}^{n} X_i^2 - 2n\overline{X}_n^2 + n\overline{X}_n^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}_n^2.\end{aligned}$$

Furthermore,

$$E[X_i^2] = E^2[X_i] + \mathrm{Var}[X_i] = \mu^2 + \sigma^2$$

and

$$E[\overline{X}_n^2] = E^2[\overline{X}_n] + \mathrm{Var}[\overline{X}_n] = \mu^2 + \frac{1}{n}\sigma^2.$$

Therefore,

$$E[S^2] = \frac{1}{n-1}\left(\sum_{i=1}^{n} E[X_i^2] - nE[\overline{X}_n^2]\right) = \sigma^2.$$

In other words, S^2 is an unbiased estimator for the population variance σ^2. As for the consistency of S^2, we note that by SLLN,

$$S^2 = \frac{n}{n-1} \cdot \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2\right) \to 1 \cdot (E[X_1^2] - \mu^2) = \sigma^2.$$

We complete the proof. ■

The sample mean $\overline{X}_n$ is an unbiased estimate for the population mean μ. Its standard deviation is

$$\sigma_{\overline{X}_n} = \sqrt{\mathrm{Var}[\overline{X}_n]} = \frac{\sigma}{\sqrt{n}}$$

which is a measurement of the error of $\overline{X}_n$ as an estimate of μ. Even though σ is unknown, we can use S as an approximation. This leads to the **standard error** of $\overline{X}_n$:

$$\text{S.E.}(\overline{X}_n) = \frac{S}{\sqrt{n}}.$$

Many times, we will use the term "standard error" for similar quantities in statistics.

Our discussion on estimators for population mean and variance can easily extend to the case of estimating the difference of population means

and its variance. For example, suppose $\{X_1, X_2, \ldots, X_n\}$ are iid samples from Population 1 with mean μ_X and variance σ_X^2, and $\{Y_1, Y_2, \ldots, Y_m\}$ are iid samples from Population 2 with mean μ_Y and variance σ_Y^2. Our interest is to estimate $\theta = \mu_X - \mu_Y$. A natural estimator would be

$$\hat{\theta} \doteq \overline{X}_n - \overline{Y}_m.$$

Clearly , this estimator is unbiased and consistent (as $n, m \to \infty$). Its variance is, by independence,

$$\mathrm{Var}[\hat{\theta}] = \mathrm{Var}[\overline{X}_n] + \mathrm{Var}[\overline{Y}_m] = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}.$$

We can use the sample variances to approximate the population variances, which leads to the standard error of $\hat{\theta}$:

$$\text{S.E.}(\hat{\theta}) = \sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}.$$

Remark 7.1. In point estimation, it is standard to report both the estimate and the associated standard error (or at least some rough idea of its magnitude). The latter is important because an estimate without a sense of error is meaningless.

7.4 Estimator for Population Proportion

Estimating population proportion is a special case of estimating population mean. More precisely, suppose $\{X_1, X_2, \ldots, X_n\}$ are iid Bernoulli random variables with

$$P(X_i = 1) = p, \quad P(X_i = 0) = 1 - p.$$

Our parameter of interest is p, the **population proportion**. It is also the population mean since $E[X_i] = p$. The estimator is of course the **sample proportion** (which is also the sample mean)

$$\hat{p} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

Its variance is

$$\mathrm{Var}[\hat{p}] = \frac{1}{n}\mathrm{Var}[X_i] = \frac{p(1-p)}{n}.$$

When we use $\hat{p}$ to approximate p in the formula of variance, we obtain the **standard error** of sample proportion $\hat{p}$:

$$\text{S.E.}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The extension to the case of estimating the difference of population proportions is straightforward. More precisely, suppose $\{X_1, X_2, \ldots, X_n\}$ are iid Bernoulli samples from Population 1 with $P(X_i = 1) = p_X$ and $\{Y_1, Y_2, \ldots, Y_m\}$ iid Bernoulli samples from Population 2 with $P(Y_i = 1) = p_Y$. The estimator for the difference $\theta \doteq p_X - p_Y$ and its standard error are

$$\hat{\theta} = \hat{p}_X - \hat{p}_Y, \quad \text{S.E.}(\hat{\theta}) = \sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{m}}.$$

Exercise 7.3. You may ask why we are not using sample variance to estimate the variance of $\hat{p}$, just like what we have done in Section 7.3, since $\hat{p}$ is also the population mean. The answer is that it is unnecessary — even if you do so, it will be nearly identical when n is large. Show that the sample variance when $\{X_1, X_2, \ldots, X_n\}$ are Bernoulli random variables is

$$S^2 = \frac{n\hat{p}(1-\hat{p})}{n-1},$$

which is essentially $\hat{p}(1-\hat{p})$ when n is large. (*Hint:* $X_i^2 = X_i$ *for all* i)

Chapter 8

Confidence Interval

In point estimation, it is common to report both the estimate and the associated standard error. It is also quite common to report both the estimate and the associated *confidence intervals*, which is another type of measurement of error and more meaningful when the distribution of the estimates are not normal or approximately normal.

Definition 8.1. Throughout the notes, for any $\alpha \in (0,1)$, we define z_α to be $\Phi^{-1}(1-\alpha)$, or equivalently, the unique number such that

$$P(N(0,1) \geq z_\alpha) = \alpha$$

α	0.25	0.05	0.025	0.005
z_α	0.67	1.64	$1.96 \approx 2$	2.58

Definition 8.2. A **confidence interval with confidence level** $(1-\alpha)$, or a $(1-\alpha)$ **confidence interval**, for a population parameter θ is a *random interval* $[L, R]$ such that

1. L and R are both functions of samples $\{X_1, X_2, \ldots, X_n\}$;
2. $P(\theta \in [L, R]) = 1 - \alpha$.

Example 8.1. Let $\{X_1, X_2, \ldots, X_n\}$ be iid samples from the population distribution $N(\theta, \sigma^2)$. The target parameter is the unknown population mean θ. We assume, at the moment, that the population variance σ^2 is *known* to us. This is unrealistic, but we adopt this assumption to illustrate the idea of confidence interval. The estimator for θ is the sample mean

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

Since X_i's are iid normal random variables, *$\hat{\theta}$ is normal with mean θ and variance σ^2/n*. It follows that, for any $\alpha \in (0,1)$,

$$P\left(-z_{\alpha/2} \leq \frac{\hat{\theta}-\theta}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1-\alpha,$$

This amounts to $P(L \leq \theta \leq R) = 1-\alpha$, where

$$L = \hat{\theta} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \qquad R = \hat{\theta} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

In other words, $[L, R]$ defines a confidence interval of θ with confidence level $1-\alpha$.

In practice, the population variance σ^2 is unknown. But we can use the sample variance S^2 to approximate σ^2. Also, recall that

$$\text{S.E.}(\hat{\theta}) = \frac{S}{\sqrt{n}}.$$

We can find an approximate confidence interval with confidence level $1-\alpha$ in the following form

$$\left[\hat{\theta} - z_{\alpha/2}\text{S.E.}(\hat{\theta}),\ \ \hat{\theta} + z_{\alpha/2}\text{S.E.}(\hat{\theta})\right]$$

For instance, suppose we have $n = 100$ samples from a normal distribution with mean θ. Assume that the sample mean and sample variance are (say)

$$\overline{x} = 2.3, \quad s^2 = 4$$

respectively. Then a confidence interval with confidence level 95% is (we take $z_{0.025} = 2$)

$$\left[2.3 - 2\cdot\frac{\sqrt{4}}{\sqrt{100}},\ \ 2.3 + 2\cdot\frac{\sqrt{4}}{\sqrt{100}}\right] = [1.9,\ \ 2.7].$$

We would like to remark that we have used lower case $\overline{x}$ and s^2 to represent the values of $\overline{X}_n$ and S^2 from a specific set of samples. ■

Remark 8.1. (*understanding confidence interval*) In Example 8.1, we end up with a 95% confidence interval $[1.9, 2.7]$ for the population parameter θ. *Does this imply that the interval $[1.9, 2.7]$ has a 95% probability to cover the true value θ?* The answer is NO. The true parameter θ is an unknown but *fixed*

number. The interval $[1.9, 2.7]$ will either cover it or not. There is no middle ground. So what does this 95% confidence level mean?

The true 95% confidence interval is a *random interval*. When we apply the formula of this random interval to a specific collection of samples, we will produce a specific realization or instance of the random interval such as $[1.9, 2.7]$. We cannot tell for sure whether this specific interval covers the true value θ. However, if we repeat this procedure 100 times and apply the formula to 100 different collections of samples, we can be confident that roughly 95 of the 100 resulting confidence intervals will cover the true value. So in that sense, this 95% confidence level is referring to the procedure of generating the confidence interval.

Remark 8.2. The tighter the confidence interval, the more accurate the estimate. To reduce the size of a confidence interval by half, we need to quadruple the sample size, everything else being equal.

Remark 8.3. Confidence interval is not unique, given the same confidence level. One approach is to pick the tightest one, that is, the one with the shortest length $(R - L)$. This is probably the most ideal choice, but difficult to implement in practice. A commonly adopted approach is to choose L and R such that

$$P(\theta < L) = P(\theta > R) = \frac{\alpha}{2}.$$

This is what we have done in Example 8.1. When the error $\hat{\theta} - \theta$ is symmetrically distributed with a single mode (like normal), these two approaches coincide. Thus the confidence interval in Example 8.1 is also the tightest.

Remark 8.4. Confidence intervals such as $[\hat{\theta} - x,\ \hat{\theta} + x]$ are often denoted by

$$\hat{\theta} \pm x$$

for notational simplicity.

Exercise 8.2. Each of 100 students is asked to independently draw samples from a population distribution and form a 95% confidence interval of the population mean. What is the probability that at least one of the intervals do not cover the true population mean?

8.1 Large Sample Confidence Interval

Example 8.1 derives the form of confidence intervals under the assumption that the population distribution is normal. This assumption is not essential.

The key step in the derivation is that "$\hat{\theta}$ is normally distributed" (the red text in the example). In general, if an estimator such as $\hat{\theta}$ is approximately normally distributed, the same derivation will work and a confidence interval for θ with confidence level $(1-\alpha)$ is approximately

$$\hat{\theta} \pm z_{\alpha/2}\text{S.E.}(\hat{\theta}).$$

In many studies, even if the population distribution is not normal, an estimate $\hat{\theta}$ will be approximately normal. In particular, if $\hat{\theta}$ is

- sample mean,
- sample proportion,
- difference of sample means,
- difference of sample proportions,

then $\hat{\theta}$ is approximately normal when the sample sizes are large, thanks to the central limit theorem.

Example 8.3. Samples of life span of nonsmokers and smokers are collected and the summary of the data is shown in the table. Find the 95% confidence intervals for the average life span of nonsmokers and smokers, respectively, and their difference.

	sample size	sample mean	sample std
Nonsmokers	$n = 36$	$\overline{x} = 72$	$s_1 = 9$
Smokers	$m = 44$	$\overline{y} = 62$	$s_2 = 11$

Solution: Let μ_1 and μ_2 denote the population average life span of nonsmokers and smokers, respectively. Let $\theta = \mu_1 - \mu_2$ be the difference.

1. *95% confidence interval for* μ_1: The estimate for μ_1 is $\overline{x}$, and the 95% confidence interval is

$$\overline{x} \pm 2\frac{s_1}{\sqrt{n}} = 72 \pm 2\frac{9}{\sqrt{36}} = 72 \pm 3$$

2. *95% confidence interval for* μ_2: The estimate for μ_2 is $\overline{y}$, and the 95% confidence interval is

$$\overline{y} \pm 2\frac{s_2}{\sqrt{m}} = 62 \pm 2\frac{11}{\sqrt{44}} = 62 \pm 3.3$$

3. *95% confidence interval for θ*: The estimate for θ is $\hat{\theta} = \overline{x} - \overline{y} = 10$. The standard error associated with $\hat{\theta}$ is

$$\text{S.E.}(\hat{\theta}) = \sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}} = \sqrt{\frac{9^2}{36} + \frac{11^2}{44}} = 2.24.$$

The 95% confidence interval is

$$\hat{\theta} \pm 2\text{S.E.}(\hat{\theta}) = 10 \pm 4.48.$$

Example 8.4. A coin is tossed n times and X is the total number of heads. Find the 95% confidence interval for p, the probability of heads.

Solution: The estimate for p is sample proportion $\hat{p} = X/n$. The 95% confidence interval is

$$\hat{p} \pm 2\text{S.E.}(\hat{p}) = \hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Example 8.5. In order to compare the proportions of supporters among male and female voters for a candidate in an upcoming election, a random sample of 100 male and 100 female voters are selected.

	Sample size	Support
Male	100	80
Female	100	60

Let p_m and p_f denote the population proportion of supporters for the candidate among male and female voters, respectively. Find a 95% confidence interval for $p_m - p_f$.

Solution: The estimates for p_m and p_f are respectively $\hat{p}_m = 80/100 = 0.8$ and $\hat{p}_f = 60/100 = 0.6$. Thus the estimate for $p_m - p_f$ is

$$\hat{p}_m - \hat{p}_f = 0.8 - 0.6 = 0.2$$

and the confidence interval is

$$0.2 \pm 2\sqrt{\frac{0.8(1-0.8)}{100} + \frac{0.6(1-0.6)}{100}} = 0.2 \pm 0.13$$

Chapter 9

Maximum Likelihood Estimate

In this chapter we discuss one of the most commonly used estimators in statistical inference, the *maximum likelihood estimate* (MLE), and its asymptotic properties.

Definition 9.1. Let $\{X_1, X_2, \ldots, X_n\}$ be iid samples from a population with distribution $f_\theta(x)$. Here f_θ is the probability mass function if X_i's are discrete and the probability density function if X_i's are continuous. The **Likelihood function** is defined to be the joint distribution of $\{X_1, X_2, \ldots, X_n\}$. That is,

$$L_\theta(x_1, x_2, \ldots, x_n) = f_\theta(x_1) f_\theta(x_2) \cdots f_\theta(x_n).$$

The **maximum likelihood estimate** $\hat{\theta}$ is defined to be the θ that maximizes the likelihood $L_\theta(X_1, X_2, \ldots, X_n)$.

Example 9.1. Let $X_1, X_2, \ldots, X_n$ be iid samples from a Bernoulli distribution with parameter p. Find the maximum likelihood estimate for p.

Solution: For a Bernoulli random variable X_i with $P(X_i = 1) = p$ and $P(X_i = 0) = 1 - p$, we can express its probability mass function as

$$f_\theta(x_i) = P(X_i = x_i) = p^{x_i}(1-p)^{1-x_i}, \quad x_i = 0, 1.$$

Therefore, the likelihood function is

$$L_p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^x(1-p)^{n-x}, \quad x = \sum_{i=1}^{n} x_i.$$

To maximize the likelihood is equivalent to maximizing the log-likelihood:

$$\log L_p(x_1, x_2, \ldots, x_n) = x \log p + (n - x) \log(1 - p).$$

The maximizer p^* satisfies

$$\frac{x}{p^*} - \frac{n-x}{1-p^*} = 0 \quad \text{or} \quad p^* = \frac{x}{n}.$$

Therefore, the maximum likelihood estimate of p is (replacing x_i with X_i)

$$\hat{p} = \frac{X}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n},$$

which is the sample proportion. ■

Example 9.2. Let $X_1, X_2, \ldots, X_n$ be iid samples from a Poisson distribution with parameter λ. Find the maximum likelihood estimate for λ.

Solution: The likelihood function is

$$L_\lambda(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda} \frac{\lambda^{x_1+x_2+\cdots+x_n}}{x_1! x_2! \cdots x_n!}$$

To maximize the likelihood is equivalent to maximizing the log-likelihood:

$$\log L_\lambda = -n\lambda + (x_1 + x_2 + \cdots + x_n) \log \lambda - \log(x_1! x_2! \cdots x_n!).$$

The maximizer λ^* satisfies

$$-n + (x_1 + x_2 + \cdots + x_n)\frac{1}{\lambda^*} = 0 \quad \text{or} \quad \lambda^* = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Therefore, the maximum likelihood estimate of λ is (replacing x_i with X_i)

$$\hat{\lambda} = \frac{X_1 + X_2 + \cdots + X_n}{n},$$

which is the sample mean. ■

Example 9.3. Let $X_1, X_2, \ldots, X_n$ be iid samples from a normal distribution with mean μ and variance σ^2. Find the maximum likelihood estimates for μ and σ.

Solution: In this problem, the population parameter $\theta = (\mu, \sigma)$ consists of two unknowns. The likelihood function is

$$L_\theta(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i-\mu)^2/(2\sigma^2)}$$

and the log-likelihood function is

$$\log L_\theta(x_1, x_2, \ldots, x_n) = -n \log \sqrt{2\pi} - n \log \sigma - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}.$$

Taking derivatives with respect to μ and σ and setting them to zero, we arrive at the equations for the maximizer $\theta^* = (\mu^*, \sigma^*)$:

$$-\sum_{i=1}^{n} \frac{\mu^* - x_i}{(\sigma^*)^2} = 0, \quad -\frac{n}{\sigma^*} + \sum_{i=1}^{n} \frac{(x_i - \mu^*)^2}{(\sigma^*)^3} = 0.$$

Solving them, we have

$$\mu^* = \overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}, \quad (\sigma^*)^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

Therefore, the maximum likelihood estimates of μ and σ are respectively

$$\hat{\mu} = \overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2.$$

Note that $\hat{\mu}$ is indeed the sample mean. But $\hat{\sigma}^2$ is not exactly the sample variance, though the difference is quite negligible when n is large, neither is it an unbiased estimator for the population variance σ^2. ■

Example 9.4. Let $X_1, X_2, \ldots, X_n$ be iid samples from an exponential distribution with rate λ. Find the maximum likelihood estimates for λ.

Solution: The likelihood function is

$$L_\lambda(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda(x_1 + x_2 + \cdots + x_n)}.$$

Here we have omitted the indicator function $1_{\{x_i \geq 0\}}$ in the probability density function for exponentials, because the samples are all automatically nonnegative. To maximize the likelihood is equivalent to maximizing the log-likelihood:

$$\log L_\lambda = n \log \lambda - \lambda(x_1 + x_2 + \cdots + x_n)$$

The maximizer λ^* satisfies

$$\frac{n}{\lambda^*} + (x_1 + x_2 + \cdots + x_n) = 0 \quad \text{or} \quad \lambda^* = \frac{n}{x_1 + x_2 + \cdots + x_n}.$$

Therefore, the maximum likelihood estimate of λ is (replacing x_i with X_i)

$$\hat{\lambda} = \frac{n}{X_1 + X_2 + \cdots + X_n} = \frac{1}{\overline{X}_n}.$$

This estimator is *not* an unbiased estimate for λ. ■

Example 9.5. Let $X_1, X_2, \ldots, X_n$ be iid samples from a uniform distribution on $[0, \theta]$. Find the maximum likelihood estimates for θ.

Solution: The likelihood function is

$$L_\theta(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\theta} = \frac{1}{\theta^n}.$$

Here we have omitted the indicator function $1_{[0,\theta]}(x_i)$ in the probability density function for uniforms, because the samples are all automatically within the interval $[0, \theta]$. However, since $\theta \geq x_{(n)} \doteq \max(x_1, x_2, \ldots, x_n)$, the maximizer for the likelihood function is

$$\theta^* = x_{(n)}.$$

Therefore, the maximum likelihood estimate of θ is (replacing x_i with X_i)

$$\hat{\theta} = X_{(n)} \doteq \max(X_1, X_2, \ldots, X_n).$$

This estimator is *not* an unbiased estimate for θ — it always underestimates the true value of θ. ■

9.1 Asymptotic Properties of MLE

We are primarily interested in the properties of maximum likelihood estimators when the sample size n is large. However, since a rigorous treatment will require a number of carefully stated technical conditions, we should restrict ourselves to informal arguments.

For notational simplicity, we only consider the case where the population distribution is continuous with probability density function $f_\theta(x)$. There are a few basic conditions we would like to impose throughout in order to facilitate our analysis: (1) The probability density function $f_\theta(x)$ is different for different θ; (2) θ consists of a single unknown parameter; (3) $f_\theta(x)$ is differentiable with respect to θ to any orders; (4) There exists a unique maximum likelihood estimate. We also let $\hat{\theta}$ and θ^* denote the maximum likelihood estimate and the true value of the population parameter θ, respectively.

Lemma 9.1. (consistency) *Under some regularity conditions, the maximum likelihood estimate is consistent. That is,* $\hat{\theta} \to \theta^*$ *as* $n \to \infty$.

Proof. This is the sketch of a proof. The maximum likelihood estimate $\hat{\theta}$ maximizes the likelihood function, or equivalently, the log-likelihood function (scaled by n)

$$\ell_\theta(X_1, X_2, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} \log f_\theta(X_i).$$

Thanks to the strong law of large numbers and the assumption that X_i's are iid with common probability density function $f_{\theta^*}(x)$, we have

$$\ell_\theta(X_1, X_2, \ldots, X_n) \to E[\log f_\theta(X)] = \int_{\mathbb{R}} \log f_\theta(x) \cdot f_{\theta^*}(x)\, dx := H(\theta)$$

with probability one. It follows heuristically that the maximum likelihood estimate $\hat{\theta}$, which maximizes the left-hand-side, should converge to the maximizer of the right-hand-side. Therefore, it suffices for us to verify that θ^* is the unique maximizer of $H(\theta)$.

To this end, we introduce an inequality: $\log(x) \le x - 1$ for all $x \ge 0$, with equality if and only if $x = 1$. The proof of this inequality is left to the interested students, which serves as a little exercise in calculus. With this inequality, we have

$$H(\theta) - H(\theta^*) = \int_{\mathbb{R}} \log \frac{f_\theta(x)}{f_{\theta^*}(x)} \cdot f_{\theta^*}(x)\, dx \le \int_{\mathbb{R}} \left(\frac{f_\theta(x)}{f_{\theta^*}(x)} - 1 \right) f_{\theta^*}(x)\, dx.$$

But because $f_\theta(x)$ and $f_{\theta^*}(x)$ are probability density functions, they both integrate to 1. Therefore, the integral on the right-hand-side is 0 and

$$H(\theta) - H(\theta^*) \le 0$$

with equality if and only if $f_\theta(x) = f_{\theta^*}(x)$ for all x, which amounts to $\theta = \theta^*$. That is, θ^* is the unique maximizer for $H(\theta)$. We complete the proof. ■

Our next result is to establish the *asymptotic normality* of the maximum likelihood estimates. For that, we need the following technical lemma.

Lemma 9.2. *If* X *is a random variable with probability density function* $f_\theta(x)$, *then*

$$E\left[\frac{\partial}{\partial \theta} \log f_\theta(X) \right] = 0, \ \ \mathrm{Var}\left[\frac{\partial}{\partial \theta} \log f_\theta(X) \right] = E\left[-\frac{\partial^2}{\partial \theta^2} \log f_\theta(X) \right].$$

Proof. The key observation is that, since $f_\theta(x)$ is a probability density function for every θ, we have

$$\int_{\mathbb{R}} f_\theta(x)dx = 1,$$

which implies that

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta}\int_{\mathbb{R}} f_\theta(x)dx = \int_{\mathbb{R}} \frac{\partial f_\theta}{\partial\theta}(x)dx \\ 0 &= \frac{\partial^2}{\partial\theta^2}\int_{\mathbb{R}} f_\theta(x)dx = \int_{\mathbb{R}} \frac{\partial^2 f_\theta}{\partial\theta^2}(x)dx. \end{aligned}$$

Denote

$$\begin{aligned} g_\theta(x) &\doteq \frac{\partial}{\partial\theta}\log f_\theta(x) = \frac{1}{f_\theta(x)}\frac{\partial f_\theta}{\partial\theta}(x) \\ \frac{\partial^2}{\partial\theta^2}\log f_\theta(x) &= \frac{1}{f_\theta(x)}\frac{\partial^2 f_\theta}{\partial\theta^2}(x) - \frac{1}{f_\theta^2(x)}\left[\frac{\partial f_\theta}{\partial\theta}(x)\right]^2 \\ &= \frac{1}{f_\theta(x)}\frac{\partial^2 f_\theta}{\partial\theta^2}(x) - [g_\theta(x)]^2 \end{aligned}$$

It follows that

$$\frac{\partial}{\partial\theta}\log f_\theta(X) = g_\theta(X).$$

Observe that

$$E[g_\theta(X)] = \int_{\mathbb{R}} g_\theta(x) f_\theta(x)dx = \int_{\mathbb{R}} \frac{\partial f_\theta}{\partial\theta}(x)dx = 0,$$

$$\mathrm{Var}[g_\theta(X)] = E[g_\theta^2(X)] = E\left[\frac{1}{f_\theta(X)}\frac{\partial^2 f_\theta}{\partial\theta^2}(X) - \frac{\partial^2}{\partial\theta^2}\log f_\theta(X)\right].$$

However, since

$$E\left[\frac{1}{f_\theta(X)}\frac{\partial^2 f_\theta}{\partial\theta^2}(X)\right] = \int_{\mathbb{R}} \frac{\partial^2 f_\theta}{\partial\theta^2}(x)dx = 0,$$

we complete the proof. ■

Theorem 9.3. (asymptotic normality) *Under some regularity conditions, the maximum likelihood estimate is asymptotically normal. More precisely,*

$$\sqrt{n}(\hat{\theta} - \theta^*) \to N\left(0, \frac{1}{I(\theta^*)}\right), \quad \textit{where } I(\theta^*) \doteq E\left[-\frac{\partial^2}{\partial\theta^2}\log f_\theta(X)\Big|_{\theta=\theta^*}\right]$$

as $n \to \infty$. Here $I(\theta^*)$ is said to be the **Fisher information**.

Proof. This is the sketch of a proof. The maximum likelihood estimate $\hat{\theta}$ maximizes the likelihood function, or equivalently, the log-likelihood function (scaled by n)

$$\ell(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log f_\theta(X_i).$$

We have suppressed the dependence of ℓ on $\{X_1, X_2, \ldots, X_n\}$for notational simplicity. The maximum likelihood estimate $\hat{\theta}$ is the solution to the equation

$$\frac{\partial}{\partial \theta} \ell(\theta) = 0.$$

However, by the classical mean value theorem, we have (the derivatives of ℓ are always with respect to θ)

$$0 = \ell'(\hat{\theta}) = \ell'(\theta^*) + \ell''(\overline{\theta})(\hat{\theta} - \theta^*).$$

for some $\overline{\theta}$ that is between θ^* and $\hat{\theta}$. However, thanks to the consistency of the maximum likelihood estimates, $\hat{\theta} \to \theta^*$. Therefore, $\overline{\theta} \to \theta^*$ as well. This and the strong law of large numbers imply

$$\ell''(\overline{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log f_\theta(X_i) \bigg|_{\theta=\overline{\theta}} \to -I(\theta^*)$$

with probability one. On the other hand, thanks to Lemma 9.2 and the central limit theorem, we have

$$\sqrt{n}\ell'(\theta^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f_\theta(X_i) \bigg|_{\theta=\theta^*} \to N(0, I(\theta^*)).$$

It follows that

$$\sqrt{n}(\hat{\theta} - \theta^*) = -\frac{\sqrt{n}\ell'(\theta^*)}{\ell''(\overline{\theta})} \to \frac{N(0, I(\theta^*))}{I(\theta^*)} = N\left(0, \frac{1}{I(\theta^*)}\right).$$

We complete the proof. ∎

Example 9.6. We should directly verify the asymptotic normality of the maximum likelihood estimate for the Bernoulli distribution; see Example 9.1. Let $X_1, X_2, \ldots, X_n$ be iid samples from a Bernoulli distribution with parameter p. The maximum likelihood estimate is $\hat{p} = \overline{X}_n$, the sample proportion. Let $q \doteq 1 - p$. It follows directly from the central limit theorem that

$$\sqrt{n}(\hat{p} - p) \to N(0, pq).$$

The Fisher information is

$$I(p) = E\left[-\frac{\partial^2}{\partial p^2}\log f_p(X)\right],$$

where X is Bernoulli with parameter p and $f_p(x)$ its probability mass function. In other words,

$$f_p(x) = p^x q^{1-x}, \quad \log f_p(x) = x\log p + (1-x)\log(1-p),$$

$$-\frac{\partial^2}{\partial p^2}\log f_p(x) = \frac{x}{p^2} + \frac{1-x}{q^2}.$$

Therefore,

$$I(p) = E\left[\frac{X}{p^2} + \frac{1-X}{q^2}\right] = \frac{1}{p} + \frac{1}{q} = \frac{1}{pq}.$$

Theorem 9.3 holds for Bernoulli distributions. ■

Example 9.7. We should directly verify the asymptotic normality of the maximum likelihood estimate for exponential distributions; see Example 9.4. Let $X_1, X_2, \ldots, X_n$ be iid samples from an exponential distribution with rate λ. The maximum likelihood estimates for λ is

$$\hat{\lambda} = \frac{1}{\overline{X}_n} = \frac{n}{X_1 + X_2 + \cdots + X_n}.$$

Thanks to the central limit theorem, $E[X_i] = 1/\lambda$ and $\mathrm{Var}[X_i] = 1/\lambda^2$, we have

$$\sqrt{n}\frac{1}{\lambda\hat{\lambda}}(\lambda - \hat{\lambda}) = \sqrt{n}\left(\frac{1}{\hat{\lambda}} - \frac{1}{\lambda}\right) = \sqrt{n}\left(\overline{X}_n - \frac{1}{\lambda}\right) \to N\left(0, \frac{1}{\lambda^2}\right).$$

However, by the strong law of large numbers, $\overline{X}_n \to 1/\lambda$ and hence $\hat{\lambda} \to \lambda$. It follows that

$$\sqrt{n}(\hat{\lambda} - \lambda) \to -\lambda^2 N\left(0, \frac{1}{\lambda^2}\right) = N(0, \lambda^2).$$

The Fisher information is

$$I(\lambda) = E\left[-\frac{\partial^2}{\partial \lambda^2}\log f_\lambda(X)\right],$$

where X is exponential with rate λ and $f_\lambda(x)$ its probability density function, that is (we can assume $x \geq 0$ because X is nonnegative),

$$f_\lambda(x) = \lambda e^{-\lambda x}, \quad \log f_\lambda(x) = \log\lambda - x\lambda, \quad -\frac{\partial^2}{\partial \lambda^2}\log f_\lambda(x) = \frac{1}{\lambda^2}.$$

Therefore, $I(\lambda) = 1/\lambda^2$, and Theorem 9.3 holds for exponential distributions. ■

Exercise 9.8. Recall Example 9.5. Let $X_1, X_2, \ldots, X_n$ be iid sample from uniform distribution on $[0, \theta]$. The maximum likelihood estimate for θ is

$$\hat{\theta} = X_{(n)} \doteq \max(X_1, X_2, \ldots, X_n)$$

1. Use the fact that $X_{(n)}$ is non-decreasing with respect to n to show that $X_{(n)}$ is consistent.

2. Explain that Theorem 9.3 does *not* hold for this maximum likelihood estimate.

3. In Example 7.2, we have derived the probability density function for $X_{(n)}$. Use it and the Chebychev inequality to show that for an arbitrary constant $\varepsilon > 0$,

$$P(\theta - \hat{\theta} \geq \varepsilon) \leq \frac{2\theta^2}{n^2\varepsilon^2}.$$

   This gives an idea of how close $\hat{\theta}$ is to the true θ.

This is an example where the maximum likelihood estimate is consistent but not asymptotically normal.

9.2 Efficiency of MLE

We have shown some asymptotic properties of the maximum likelihood estimate. But how does it compare to other estimates? It turns out that in general the maximum likelihood estimate is *asymptotically* as efficient as it can be, in terms of the mean square error. In this section, we will partially prove this claim by comparing maximum likelihood estimates with unbiased estimates.

Theorem 9.4. (Cramér -Rao Lower Bound) *Let $\{X_1, X_2, \ldots, X_n\}$ be iid samples from a population distribution $f_\theta(x)$. Then for any unbiased estimator $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$, we have*

$$\mathrm{Var}[\hat{\theta}] \geq \frac{1}{nI(\theta)}.$$

Proof. Without loss of generality, we will assume the distribution to be continuous and $f_\theta(x)$ is the probability density function for X_i. Define

$$g_\theta(x) \doteq \frac{\partial}{\partial \theta} \log f_\theta(x) = \frac{1}{f_\theta(x)} \frac{\partial f_\theta}{\partial \theta}(x)$$

$$G_\theta(x_1, x_2, \ldots, x_n) \doteq \sum_{i=1}^{n} g_\theta(x_i), \quad Y \doteq G_\theta(X_1, X_2, \ldots, X_n).$$

Thanks to Lemma 5.2, the correlation between Y and $\hat{\theta}$ must be between -1 and 1. In particular,

$$\mathrm{Cov}^2(Y, \hat{\theta}) \leq \mathrm{Var}[Y]\mathrm{Var}[\hat{\theta}].$$

Thanks to Lemma 9.2, we have $E[g_\theta(X_i)] = 0$ and $\mathrm{Var}[g_\theta(X_i)] = I(\theta)$. Thus,

$$E[Y] = 0, \quad \mathrm{Var}[Y] = nI(\theta), \quad \mathrm{Cov}(Y, \hat{\theta}) = E[Y\hat{\theta}],$$

and

$$E^2[Y\hat{\theta}] = \mathrm{Cov}^2(Y, \hat{\theta}) \leq nI(\theta)\mathrm{Var}[\hat{\theta}].$$

It remains to show that $E[Y\hat{\theta}] = 1$. To this end, note that $E[\hat{\theta}] = \theta$ since $\hat{\theta}$ is unbiased. That is,

$$\theta = \int_{\mathbb{R}^n} \hat{\theta}(x_1, \cdots, x_n) f_\theta(x_1) \cdots f_\theta(x_n) dx_1 \cdots dx_n.$$

Taking derivative with respect to θ on both sides and observing that

$$\frac{\partial}{\partial \theta}[f_\theta(x_1) \cdots f_\theta(x_n)] = G_\theta(x_1, \ldots, x_n) f_\theta(x_1) \cdots f_\theta(x_n),$$

we have

$$1 = \int_{\mathbb{R}^n} \hat{\theta}(x_1, \ldots, x_n) G_\theta(x_1, \ldots, x_n) f_\theta(x_1) \cdots f_\theta(x_n) dx_1 \cdots dx_n$$

Note that the integral on the right-hand-side is exactly $E[\hat{\theta}Y]$. We complete the proof. ∎

The Cramér-Rao lower bound and Theorem 9.3 imply that the maximum likelihood estimate is asymptotically efficient, at least in the sense that no unbiased estimates can do better. A similar result can be proven even for biased estimates in terms of mean square error (barring possibly a small set of θ's). This justifies why maximum likelihood estimate is so commonly used in practice.

9.3 Sufficient Statistics

In this section we briefly discuss the *sufficient statistics* in statistical inference. It is the key tool in the study of unbiased estimates with minimum variance. It also somewhat shows that the maximum likelihood estimate is not easy to be improved upon.

As usual, let $\{X_1, X_2, \ldots, X_n\}$ be iid samples from a population distribution $f_\theta(x)$. Let $T = T(X_1, X_2, \ldots, X_n)$ be a single or a family of random variables that are functions of the samples (any such random variable is said to be a **statistic**). For instance, $T = X_n$ or $T = (X_1 + X_2, X_n)$ are both valid examples of statistics, but $T = \theta(X_1 - X_2)$ is not since T cannot depend on the unknown parameter.

Now consider an arbitrary estimator $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ for the population parameter θ. The variance decomposition formula (Exercise 5.38) provides an interesting insight. Define

$$\overline{\theta} \doteq E[\hat{\theta}|T].$$

Then by the law of total expectation and the variance decomposition formula

$$E[\overline{\theta}] = E[\hat{\theta}], \quad \mathrm{Var}[\overline{\theta}] \leq \mathrm{Var}[\hat{\theta}].$$

This seems to suggest that $\overline{\theta}$ is at least as good an "estimator" as $\hat{\theta}$ — they have the same bias, but $\overline{\theta}$ always has a smaller (at least the same) variance. However, the problem with this technique is that $\overline{\theta}$ may not be an estimator after all. The calculation of the conditional expectation $E[\hat{\theta}|T]$ will involve θ, which means $\overline{\theta}$ may depend on θ.

That said, what if $E[\hat{\theta}|T]$ does not depend on θ? If this is the case, then $\overline{\theta}$ is a better estimator than the original estimator $\hat{\theta}$ (at least as good). For this to happen, we would need *the conditional distribution of* $(X_1, X_2, \ldots, X_n)$ *given* T *does not depend on* θ *at all*. What this condition says is that all the information about θ within the data $\{X_1, X_2, \ldots, X_n\}$ is contained in the statistic T. For this reason, any statistic that satisfies this condition is said to be a **sufficient statistic**. There is a very simple criterion for a statistic to be sufficient. Well, at least the statement is quite simple. But the proof is a different story. Interested students can try to prove the theorem for discrete distributions, which is quite doable.

Theorem 9.5. (factorization criterion) *A statistic* $T = T(x_1, x_2, \ldots, x_n)$ *is sufficient if and only if*

$$f_\theta(x_1) f_\theta(x_2) \cdots f_\theta(x_n) = h(x_1, x_2, \ldots, x_n) \cdot G_\theta(T(x_1, x_2, \ldots, x_n))$$

for some function h that does not depend on θ and some function $G_\theta(t)$ that only depends on θ and t.

A sufficient statistic T sometimes consists of a family of random variables all of which are functions of samples. It does not have to be unique. For example, an n-component statistic $T(X_1, X_2, \ldots, X_n) = (X_1, X_2, \ldots, X_n)$ is automatically sufficient for any population distribution. But it is useless since $E[\hat{\theta}|T] = \hat{\theta}$ for all estimators $\hat{\theta}$. In general, we seek sufficient statistics with as few components as possible.

The preceding discussion asserts that we only need to consider those estimators that are functions of sufficient statistics, since otherwise we can improve them by conditioning on sufficient statistics. So can we improve the maximum likelihood estimator in that direction? Unfortunately the answer is "no". It is rather straightforward from the factorization criterion Theorem 9.5 that a maximum likelihood estimate is always a function of any sufficient statistic. Thus conditioning does not provide any improvement.

Exercise 9.9. Consider the estimation problems in Examples 9.1 through 9.5. Show that the sample mean $\overline{X}_n$ is a sufficient statistic for Bernoulli, Poisson, and exponential distributions; $X_{(n)}$ is a sufficient statistic for uniform distribution; and $(\overline{X}_n, S^2)$ is a sufficient statistic for normal with unknown mean and variance.

Remark 9.1. Sufficient statistics is an important component in the study of *minimum variance unbiased estimators*, whose goal is to find among all unbiased estimators the one with the minimal variance. The idea is to find a statistic, say T, such that

1. T is sufficient;
2. T is *complete*, which means that $E[h(T)] = 0$ for any possible values of population parameter θ if and only if $h(t) = 0$ for all t.

Now find an arbitrary unbiased estimator say $\hat{\theta}$. Define $\hat{\theta}^* = E[\hat{\theta}|T]$. The completeness of T implies that regardless of what $\hat{\theta}$ is chosen, $\hat{\theta}^*$ is always identical. Then $\hat{\theta}^*$ must be the minimum variance unbiased estimator.

Chapter 10

Hypothesis Testing

A large part of statistical inference is about hypothesis testing. Many questions such as the effectiveness of a new drug, the fairness of a coin, or is smoking bad for lung but good for Parkinson disease, and so on, can be formulated into problems of hypothesis testing.

Hypothesis testing is very much analogous to NFL's protocol for the challenges of on-field calls. A coach is allowed two opportunities to challenge on-field calls.

1. A coach throws a red-flag to signal a challenge.
2. A referee must see *incontrovertible visual evidence* to overturn a call.
3. After reviewing the instant replay, a referee will say "the ruling on the field stands" if no such visual evidence is observed.
4. The hypothesis is that the previous call should stand. A referee tries to find strong evidence against this hypothesis during the replay.
5. A referee is very careful not to say "the ruling on the field was correct" — what he means is that he could not find strong visual evidence to disprove the previous call.

Statistical hypothesis testing is the same. The goal is to decide if the data provides sufficient evidence to reject or disprove a particular hypothesis. Try not to treat the conclusion about a hypothesis from a test as right or wrong. When a hypothesis is not rejected, it does *not* mean the hypothesis is correct. It only means that the data does not contain enough evidence for us to dispute the hypothesis. On the other hand, when a hypothesis is rejected, it does *not* mean the hypothesis is wrong. It only means that the

data contains sufficient evidence for us to believe that the hypothesis is not very plausible.

10.1 Elements of Hypothesis Testing

We will use a simple example to illustrate the essential elements of hypothesis testing. In Spring 2009 two Berkeley undergraduates, Priscilla Ku and Janet Larwood, undertook a task to perform 40,000 coin tosses to see if the coin was fair. They took turn to toss the coin, which "only" cost them about one hour per day for one semester. The result was 20,217 heads and 19,783 tails. Was the coin fair? Let p be the probability of heads. We wish to test if the coin is fair or if $p = 0.5$. Our empirical data is that in $n = 40,000$ tosses, we have obtained $X = 20,217$ heads.

1. *Set up a null hypothesis H_0 and an alternative hypothesis H_a.* Null hypothesis is the claim to be tested (against). In general, hypothesis testing evaluates the strength of the empirical evidence against null hypothesis. Without strong evidence, we will not reject the null hypothesis. Often the null hypothesis is the one claiming "no difference" or conforming to conventional wisdom. In this example, we take

$$H_0 : \quad p = 0.5, \qquad H_a : \quad p \neq 0.5.$$

2. *Set up a test statistic.* A test statistic is a function of the samples, just like an estimator. In many cases, a test statistic is related to the estimator that one would use to estimate the parameter in question. In this example, an obvious choice is

$$\hat{p} = \frac{X}{n}.$$

3. *Quantify by P-value the strength of the empirical evidence against the null hypothesis.* **P-value** is defined to be the probability of observing the test statistic as extreme as or more extreme than what is actually observed, *assuming that the null hypothesis is true*. This means, in the present example, we would study the probability of

$$|\hat{p} - 0.5| \geq \left| \frac{20217}{40000} - 0.5 \right| = 0.005425,$$

under the assumption that the coin is fair (H_0). Under H_0, $\hat{p}$ is approximately $N(0.5, 0.0025^2)$. Therefore,

$$P\text{-value} = P\left(|\hat{p} - 0.5| \geq 0.005424\right) = 0.03.$$

This test is said to be a *two-sided test* because in the alternative hypothesis p can be on either side of 0.5.

4. *Reporting your result:* Many hypothesis testing problems have a prefixed *level of significance* α. A very commonly used level is $\alpha = 0.05$. This is the threshold to divide what we deem plausible or implausible under the null hypothesis.

   (a) When the P-value is less than or equal to α, we conclude that something implausible has been observed if the null hypothesis is true. Thus we can say that the empirical evidence is statistically significant for us to reject the null hypothesis. Again, rejection of the null hypothesis does not mean the null hypothesis is incorrect. It only means that the data seems to suggest otherwise.

   (b) When the P-value is larger than α, we conclude that our observation is plausible under the null hypothesis. Thus we can say that the empirical evidence is not statistically significant for us to reject the null hypothesis. Again, not rejecting the null hypothesis does not mean the null hypothesis is correct. It only means that the data seems to be compatible with the null hypothesis.

   There are a couple of comments we would like to make. One is that "statistical significance" changes depending on the significance level α. For this example, if $\alpha = 0.05$, then the data is statistically significant for us to reject the null hypothesis. That is, we have sufficient evidence to believe that the coin is not fair. On the other hand, if $\alpha = 0.01$, then the data is not statistically significant for us to reject the null hypothesis. In other words, we don't have enough evidence to say that it is an unfair coin. *Both interpretations are valid*. It shows that we cannot use conclusions such as "correct/incorrect" or "right/wrong" regarding the hypotheses. The second comment is that "statistical significance" is not "practical significance". The empirical proportion of heads is roughly 50.54%. A small practical discrepancy can be statistically very significant, especially with a large data set.

Example 10.1. *(Testing a population mean, one-sided test)* Are mutual funds better than index? Mutual funds often compare their performance with a benchmark index. The Vanguard International Growth Fund benchmarks its performance with EAFE index (Europe, Australasia, Far East). The following table is the performance difference (fund return - index return) each year from 1984 to 2010, in percentage.

-8.40	0.78	-12.73	-12.15	-16.66	14.22	11.40	-7.39	6.38
12.18	-7.02	3.68	8.60	2.34	-3.07	-0.62	5.57	2.5
-1.85	-4.14	-1.30	1.46	-0.42	4.81	-1.56	9.85	7.53

Solution: We first set up the null and alternative hypotheses. Let θ be the population mean of the difference (fund return - index return). Mutual funds perform better than the index amounts to $\theta > 0$. Thus the hypotheses are

$$H_0 : \theta = 0, \quad H_a : \theta > 0.$$

We would like to find strong evidence against the null hypothesis and to support the alternative hypothesis. This is a *one-sided test* because in the alternative hypothesis θ can only be on one side of 0 (value in the null hypothesis).

You may ask why we don't postulate the null hypothesis as $H_0 : \theta \leq 0$. The reason is that it is unnecessary. If we cannot find evidence to reject $\theta = 0$, we cannot reject $\theta \leq 0$ either. Besides, the P-value, even though dependent on the true value of θ, will attain its maximum on the boundary case $\theta = 0$. Therefore, the P-value will be identical if we write $H_0 : \theta \leq 0$. In many such one-sided tests, you will often see the null hypothesis H_0 expressed in this simplified form.

The test statistic obviously will be based on the sample mean $\overline{X}_n$. Under the null hypothesis, the central limit theorem (the sample standard deviation S is a good approximation to the population standard deviation σ) implies that

$$Z = \frac{\overline{X}_n - \theta_0}{S/\sqrt{n}}$$

is approximately standard normal, where $\theta_0 = 0$ is the value of θ in the null hypothesis.

The data consists of $n = 27$ samples with sample mean $\overline{x}_n = 0.52$ and sample standard deviation $s = 7.89$ (both in percentage). Therefore, the

observed value of the test statistic is

$$z = \frac{\overline{x}_n - \theta_0}{s/\sqrt{n}} = \frac{0.52 - 0}{7.89/\sqrt{27}} = 0.3425.$$

In this one-sided test, being extreme means the sample mean and thus the test statistic takes a large value. Therefore, the P-value is

$$P(Z \geq z) = P(N(0,1) \geq 0.3425) = \Phi(-0.3425) = 0.366.$$

If the significance level $\alpha = 0.05$ as usual, then P-value $> \alpha$. Therefore, we conclude that the data is not statistically significant to reject the null hypothesis. That is, there is no strong evidence to suggest that mutual funds perform better than the benchmark index. ■

Example 10.2. *(Comparing two population means)* Do indoor cats live longer than wild cats?

Cats	Sample size	Mean Life Span	Sample Std
Indoor	64	14	4
Wild	36	10	5

Solution: Let μ_1 and μ_2 be the population mean life span for indoor cats and outdoor cats, respectively. Our hypotheses are

$$H_0 : \mu_1 = \mu_2, \quad H_a : \mu_1 > \mu_2.$$

The test statistic is the difference in sample means:

$$\overline{D} = \overline{X}_1 - \overline{X}_2$$

where $\overline{X}_1$ is the sample mean life span of indoor cats and $\overline{X}_2$ is that of outdoor cats. The observed value of the test statistic is

$$\overline{d} = \overline{x}_1 - \overline{x}_2 = 14 - 10 = 4.$$

Under the null hypothesis, $\overline{D}$ is approximately normally distributed with mean 0 and standard deviation

$$\sigma_{\overline{D}} = \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{4^2}{64} + \frac{5^2}{36}} = 0.97.$$

In this one-sided test, being extreme means that the difference $\overline{D}$ takes a large value. Therefore, the P-value is

$$P(\overline{D} \geq \overline{d}) = P(N(0, 0.97^2) \geq 4) < 0.00003$$

This P-value is less than the usual significance level $\alpha = 0.05$. Therefore, we reject the null hypothesis. That is, there is statistically significant evidence in data that suggests indoor cats live longer than outdoor cats. ■

Example 10.3. *(Comparing two population proportions)* In order to test if there is any difference in opinions on abortion between males and females, random samples of 100 males and 150 females were taken.

Gender	Sample size	Favor	Oppose
Male	100	52	48
Female	150	95	55

Solution: Let p_1 and p_2 be the population proportions of males and females that support abortion, respectively. The hypotheses are

$$H_0 : p_1 = p_2, \quad H_a : p_1 \neq p_2.$$

The test statistic is the difference in the sample proportions, that is,

$$\overline{D} = \hat{p}_1 - \hat{p}_2.$$

The observed value of the test statistic $\overline{D}$ is

$$\overline{d} = \frac{52}{100} - \frac{95}{150} = -0.113.$$

For this two-sided test, the larger $|\overline{D}|$ is, the more extreme the outcome becomes. Therefore, the P-value equals $P(|\overline{D}| \geq |\overline{d}|)$, which should be computed assuming the null hypothesis.

Denote $p = p_1 = p_2$ under the null hypothesis. The central limit theorem implies that $\overline{D}$ is approximately normal with mean $p_1 - p_2 = 0$ and standard deviation

$$\sigma_{\overline{D}} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

Note that under the null hypothesis, p can be estimated using a *pooled estimate*:

$$\hat{p} = \frac{52+95}{100+150} = 0.588,$$

Therefore, the standard deviation of $\overline{D}$ is approximately (under the null hypothesis)

$$\begin{aligned}\sigma_{\overline{D}} &\approx \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \\ &= \sqrt{0.588(1-0.588)\left(\frac{1}{100}+\frac{1}{150}\right)} = 0.06354.\end{aligned}$$

That is, the distribution of $\overline{D}$ under the null hypothesis is approximately $N(0, 0.06354^2)$. Thus the P-value is

$$P(|\overline{D}| \geq |\overline{d}|) = 2P(\overline{D} \geq 0.113) = 0.075.$$

If the significance level is the usual $\alpha = 0.05$, then we cannot reject the null hypothesis. There is no statistically significant evidence that suggests males and females have different opinions on abortion. ■

Exercise 10.4. Let $\{X_1, X_2, \ldots, X_n\}$ be iid samples from a population with mean θ. Consider a two-sided hypothesis testing problem:

$$H_0 : \theta = \theta_0, \quad H_a : \theta \neq \theta_0,$$

where θ_0 is a given constant. For any $\alpha \in (0, 1)$, show that the following two statements are equivalent:

1. The null hypothesis is rejected at significance level α.
2. The $(1-\alpha)$ confidence interval of θ does not contain θ_0.

You can assume that the sample size is large enough so that the central limit theorem can be applied.

10.2 Errors and Power

Hypothesis testing can only provide answers about the plausibility of the hypotheses. Due to the randomness of samples, it is inevitable that errors will be made: rejecting a hypothesis when it is true or not rejecting (accepting) a hypothesis when it is not true. We can categorize these errors in the following table:

	Reject H_0	Accept H_0
H_0 is true	type I error	$\surd$
H_a is true	$\surd$	type II error

In other words, **type I error** is *rejecting H_0 when H_0 is true*, and **type II error** is *not rejecting H_0 when H_a is true*. Let us consider a concrete example in order to study the properties of these errors and their relations. This study does *not* need any data — it only depends on the intrinsic structure of hypothesis testing.

Example 10.5. Let $\{X_1, X_2, \ldots, X_n\}$ be iid samples from $N(\theta, 1)$ with an unknown mean θ. Consider the one-sided test:

$$H_0 : \theta = 0, \quad H_a : \theta > 0.$$

Given a significance level α, what are the probabilities of type I error and type II error, respectively?

Solution: The test statistic is the sample mean $\overline{X}_n$. For this test, the larger $\overline{X}_n$ is, the more extreme the observations are. Therefore, if we have a specific set of data with sample mean $\overline{x}$, then

$$P\text{-value} = P(\overline{X}_n \geq \overline{x}) = P\left(N\left(0, \frac{1}{n}\right) \geq \overline{x}\right)$$

because under null hypothesis $\overline{X}_n$ is normally distributed with mean 0 and variance $1/n$. We reject the null hypothesis if and only if P-value $\leq \alpha$. In other words, we reject the null hypothesis if and only if $\{\overline{x} \geq x^*\}$ where x^* is determined by

$$P\left(N\left(0, \frac{1}{n}\right) \geq x^*\right) = \alpha.$$

The region $\{\overline{x} \geq x^*\}$ is said to be the *rejection region (RR)*. To summarize, the null hypothesis will be rejected if and only if $\overline{X}_n$ falls into the rejection region RR.

1. *Type I error.* This error happens when the null hypothesis is true but the test statistic $\overline{X}_n$ falls into the rejection region. Therefore, the probability of type I error is exactly, by the definition of rejection region, the significance level α. That is,

$$P(\text{type I error}) = P(RR|H_0) = \alpha.$$

2. *Type II error.* This error happens when the null hypothesis is not true but the test statistic $\overline{X}_n$ does not fall into the rejection region (thus we accept H_0). That is,

$$\beta \doteq P(\text{type II error}) = P(RR^c|H_a).$$

This probability depends on the true value of θ. For instance, if the true value of θ is $\theta_0 > 0$, then

$$\beta = P(\overline{X}_n < x^* \text{when } \theta = \theta_0) \;=\; P\left(N\left(\theta_0, \frac{1}{n}\right) < x^*\right).$$

When θ_0 becomes larger (moves further away from the null hypothesis), the probability of type II error becomes smaller. This is intuitive in the sense that when the true distribution is further away from the null hypothesis, it is more likely for the observations to fall into the rejection region (observations that are deemed extreme under H_0), and less likely to fall outside the rejection region. ■

The discussion in Example 10.5 can be extended to general hypothesis testing problems. That is,

$$\begin{aligned} P(\text{type I error}) &= P(RR|H_0) \;=\; \alpha \\ P(\text{type II error}) &= P(RR^c|H_a) \;=\; \beta. \end{aligned}$$

Is it possible to make both probabilities of type I error and type II error small, everything else being equal? The answer is NO. Both probabilities are related to the rejection region. If one wishes to decrease the probability of type I error, one needs to reduce the size of the rejection region. This will, however, make the size of its complement larger, and hence make the probability of type II error larger as well.

Power of a test is defined to be the probability of rejecting the null hypothesis when the null hypothesis is not true (making the correct conclusion). By definition,

$$\text{Power} = 1 - P(\text{type II error}) = P(RR|H_a) = 1 - \beta.$$

Analogous to the probability of type II error, power depends on the true value of the population parameter. When the sample size increases or the true value is further away from the null hypothesis, the power grows in general.

Exercise 10.6. In Example 10.5, evaluate x^* and use it to compute the probability of type II error and the power, assuming that the true value of θ is $\theta_0 > 0$. Show that when the sample size n increases or when θ_0 becomes larger, the power grows.

10.3 Neyman-Pearson Lemma

According to the discussion on type I error, type II error, and power, given a significance level α, the rejection region RR satisfies

$$P(RR|H_0) = \alpha, \quad P(RR|H_a) = 1 - P(\text{type II error}) = \text{power}.$$

This brings up an interesting question. Among all possible rejection regions RR that satisfy

$$P(RR|H_0) = \alpha,$$

which one gives the minimal probability of type II error or the maximal power? The optimal rejection region is characterized by the *Neyman-Pearson Lemma* below, whose proof is deferred to Appendix A.

Lemma 10.1. (Neyman-Pearson) *Let $\{X_1, X_2, \ldots, X_n\}$ be iid samples from a population distribution with probability density function $f(x)$. Consider the simple hypothesis testing problem*

$$H_0 : f(x) = f_0(x), \quad H_a : f(x) = f_a(x).$$

Let $\alpha \in (0, 1)$ be an arbitrarily given significance level. For each $b > 0$, define a rejection region $R_b \subseteq \mathbb{R}^n$ by

$$R_b \doteq \left\{ (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n : \frac{f_a(x_1) f_a(x_2) \cdots f_a(x_n)}{f_0(x_1) f_0(x_2) \cdots f_0(x_n)} \geq b \right\},$$

Let b^* be such that $P(R_{b^*}|H_0) = \alpha$. Then R_{b^*} yields the maximum power among all rejection regions whose probability of type I error is bounded by α.

Example 10.7. Let X be a single sample from a population distribution with the probability density function $f(x)$. Suppose the hypotheses are

$$H_0 : f(x) = 1_{[0,1]}(x), \quad H_a : f(x) = 2x 1_{[0,1]}(x).$$

Find the most powerful test given a significance level $\alpha \in (0, 1)$.

Solution: Thanks to Neyman-Pearson Lemma, we only need to consider rejection regions of form

$$RR_b = \left\{ x \in [0, 1] : \frac{f_a(x)}{f_0(x)} = \frac{2x}{1} \geq b \right\} = \left\{ x \in [0, 1] : x \geq \frac{b}{2} \right\}.$$

In other words, the most powerful test has the rejection region given by (denote the constant $b/2$ by a)

$$RR^* = \{x \in [0,1] : \ x \geq a\}$$

for some constant a. All we need to do is to find a constant a such that the probability of type I error associated with RR^* is α, or equivalently,

$$\alpha = P(RR^*|H_0) = P(X \geq a|H_0) = 1 - a.$$

This means $a = 1 - \alpha$. Therefore, the rejection region for the most powerful test is

$$RR^* = \{x \in [0,1] : \ x \geq 1 - \alpha\}.$$

Interested students may work out the power associated with this rejection region. ■

Example 10.8. Let $\{X_1, X_2, \ldots, X_n\}$ be iid samples from a normal distribution with unknown mean θ but known variance σ^2. Consider the hypotheses

$$H_0 : \theta = \theta_0, \quad H_a : \theta = \theta_a,$$

where θ_0 and θ_a are two distinct constants. Given an arbitrary significance level $\alpha \in (0,1)$, find the most powerful test.

Solution: We first consider the case where $\theta_a > \theta_0$. Let $\vec{x} = (x_1, x_2, \ldots, x_n)$ and

$$L_\theta(\vec{x}) \doteq f_\theta(x_1) f_\theta(x_2) \cdots f_\theta(x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i-\theta)^2/(2\sigma^2)}$$

be the joint probability density function. Thanks to Neyman-Pearson Lemma 10.1, the rejection region of the most powerful test takes the form R_{b^*}, where

$$R_b \doteq \left\{ \vec{x} \in \mathbb{R}^n : \ \frac{L_{\theta_a}(\vec{x})}{L_{\theta_0}(\vec{x})} \geq b \right\}$$

and b^* is chosen such that $P(R_{b^*}|H_0) = \alpha$. Plugging in the form of L_θ, we have

$$\left\{ \frac{L_{\theta_a}(\vec{x})}{L_{\theta_0}(\vec{x})} \geq b \right\} = \left\{ (\theta_a - \theta_0) \sum_{i=1}^{n} x_i \geq a \right\}$$

for some constant a that depends on $\sigma, \theta_a, \theta_0, n, b$ — the formula for this constant a is not important at all, the most important thing is the form of the rejection region R_b. Since $\theta_a - \theta_0 > 0$, we can write

$$R_b \doteq \{\vec{x} \in \mathbb{R}^n : \ \overline{x}_n \geq c\}$$

where $\overline{x}_n$ is defined to be, as usual, the average of $(x_1, x_2, \ldots, x_n)$. Therefore, the optimal rejection region should be

$$R^* = \{\vec{x} \in \mathbb{R}^n : \overline{x}_n \geq c^*\}$$

where c^* is determined by the equation

$$\alpha = P(\overline{X}_n \geq c^* | H_0) = P\left(N\left(\theta_0, \frac{\sigma^2}{n}\right) \geq c^*\right),$$

or

$$c^* = \theta_0 + z_\alpha \frac{\sigma}{\sqrt{n}}.$$

This explicitly determines the optimal rejection region. If you compare this with our Example 10.1 of hypothesis testing, you will see that this is exactly the rejection region we have used there.

When $\theta_a < \theta_0$, the analysis is similar, except that the optimal rejection region will take the form

$$R^* = \{\vec{x} \in \mathbb{R}^n : \overline{x}_n \leq c^*\}, \quad c^* = \theta_0 - z_\alpha \frac{\sigma}{\sqrt{n}}.$$

We will leave the verification to the interested students. ■

Remark 10.1. An interesting observation from Example 10.8 is that the optimal rejection region R^* does *not* depend on the value of θ_a, except whether $\theta_a > \theta_0$ or $\theta_a < \theta_0$. This implies that R^* indeed leads to the uniformly most powerful test for one-sided test such as

$$H_0 : \theta = \theta_0, \quad H_a : \theta > \theta_0 \text{ or } H_a : \theta < \theta_0.$$

However, it also shows that there is no uniformly most powerful test for a two-sided test such as

$$H_0 : \theta = \theta_0, \quad H_a : \theta \neq \theta_0.$$

Chapter 11

Regression

So far our study of statistical inference has focused on iid samples from a population distribution. For example, if we wish to predict the final exam score for a randomly selected student, we may simply use the average final exam score from previous semesters to make our prediction. In other words, we regard the student as a random sample from the population of all students and use the previous sample mean to estimate the population mean and then use it to predict the score of the said student. There is nothing wrong about this. However, what if we know the midterm score from this student? Will this information help us make a better prediction? Even better, what if we have information such as the student's performance in other similar classes? How do we incorporate such information in our prediction? In statistical inference, a lot of such studies involve *regression* models. Among them, the most prominent family is without a doubt the *linear regression models*.

11.1 Introduction to Linear Regression

In a linear regression model, there are two types of variables. One is the **dependent variable** or **response variable**, which is usually denoted by Y. The other is the **independent variable** or **explanatory variable**, which is often denoted by x. If there are multiple independent variables, we sometimes denote them by a vector $\vec{x}$, where each component of $\vec{x}$ corresponds to one independent variable.

In general, we would like to use the explanatory variables $\vec{x}$ to help us predict the response variable Y, or help us explore the relation between the explanatory variables and the response variable. Please note that the "de-

pendent variable" or "independent variables" in regression does not mean that they are dependent random variables or independent random variables, as we have been using in our previous discussion on probability and statistical inference. They are dependent/independent variables in the classical mathematical sense, where we use them to describe functions. In our previous example of predicting the final exam score of a randomly selected student, the response variable Y is the final exam score of the student and the explanatory variables may include the student's midterm score, GPA, and so on.

The relation between the response variable Y and the explanatory variables in **linear regression** is simply, linear. More precisely, suppose the explanatory variables $\vec{x} = (x_1, x_2, \ldots, x_d)$. Then the linear regression model assumes that

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_d x_d + \varepsilon,$$

where $\beta_0, \beta_1, \ldots, \beta_d$ are constants and ε is a random variable with $E[\varepsilon] = 0$ and $\mathrm{Var}[\varepsilon] = \sigma^2$ that represents the "noise". Note that parameters such as $(\beta_0, \beta_1, \ldots, \beta_d)$ *and* σ^2 *are population parameters and unknown*. Sometimes we abuse notation and express the linear regression model by

$$y = E[Y|\vec{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_d x_d.$$

Even though the linear regression models seem to deal with linear relations exclusively, many interesting nonlinear models can be obtained via linear regression and nonlinear transforms. Here are but a few examples of polynomial, exponential, and power relations:

1. $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta x^d$: linear with alternative independent variables $\vec{x} = (x, x^2, \ldots, x^d)$;
2. $y = ce^{\beta x}$: linear with alternative dependent variable $\log(y)$;
3. $y = cx^{\alpha}$: linear with alternative dependent variable $\log(y)$ and independent variable $\log(x)$.

11.2 Least Square Estimators

The classical procedure for estimating the parameters in a linear regression model is the *method of least square*. Suppose there are n observations, denoted by $(Y_i, \vec{x}_i)$ for $i = 1, 2, \ldots, n$, where $\vec{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})$ has d components. Then

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_d x_{id} + \varepsilon_i$$

for every i. We will assume throughout the chapter that *ε_i's are iid with mean 0 and variance σ^2*. For this linear regression model, the unknown populations parameters are $(\beta_0, \beta_1, \ldots, \beta_d)$ and σ^2.

11.2.1 Matrix notation

We will use matrix notation extensively in the subsequent analysis. Recall that for a matrix A, its transpose is denoted by A^t. We also use $\mathrm{Var}[X]$ to denote the covariance matrix of a random vector X. Define

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1d} \\ 1 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nd} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{bmatrix},$$

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Then

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = E[\mathbf{Y}|\mathbf{X}] + \varepsilon, \qquad \mathrm{Var}[\mathbf{Y}] = \mathrm{Var}[\boldsymbol{\varepsilon}] = \sigma^2 \mathbf{I}_n$$

where $\mathbf{I}_n$ stands for the identity matrix with dimension $n \times n$, since ε_i's are assumed to be iid with mean 0 and variance σ^2.

11.2.2 Method of least square

The method of least square finds $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_d)$ that minimizes the **sum of squared errors** defined by

$$\mathrm{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_d x_{id})^2,$$

and uses the minimizer as an estimate for $(\beta_0, \beta_1, \ldots, \beta_d)$. Here

$$\hat{Y}_i \doteq \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_d x_{id}$$

can be viewed as the prediction. Therefore, sometimes one can write the sum of squared errors as

$$\mathrm{SSE} = \sum_{i=1}^{n} (\text{observed value} - \text{predicted value})^2.$$

In order to find the minimizer we observe that the sum of squared errors can be written in the matrix form

$$\begin{aligned} \text{SSE} &= (\mathbf{Y}-\hat{\mathbf{Y}})^t(\mathbf{Y}-\hat{\mathbf{Y}}) = (\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})^t(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{Y}^t\mathbf{Y} - 2\hat{\boldsymbol{\beta}}^t\mathbf{X}^t\mathbf{Y} + \hat{\boldsymbol{\beta}}^t\mathbf{X}^t\mathbf{X}\hat{\boldsymbol{\beta}}, \end{aligned}$$

where

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_d \end{bmatrix}, \quad \hat{\mathbf{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} \doteq \mathbf{X}\hat{\boldsymbol{\beta}}.$$

Taking derivatives with respect to each component of $\hat{\boldsymbol{\beta}}$ and setting them to 0, we have a system of equations

$$\mathbf{X}^t\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^t\mathbf{Y}.$$

Therefore, the solution to the method of least square, which is our **least-square estimator** for $\boldsymbol{\beta}$, is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}.$$

11.2.3 Properties of the least-square estimators

The derivation of the least square estimator $\hat{\boldsymbol{\beta}}$ does not seem to use any probabilistic properties of the model. But it enjoys some very nice properties. We will only list a few basic ones in this section.

Lemma 11.1. *The least-square estimator $\hat{\boldsymbol{\beta}}$ is an unbiased estimator for $\boldsymbol{\beta}$ with covariance matrix $\text{Var}[\hat{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$. Furthermore, $S^2 \doteq SSE/(n-d-1)$ is an unbiased estimator for σ^2, where SSE is the sum of squared errors.*

Proof. By the model assumptions, $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Therefore, $E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$ and consequently

$$E[\hat{\boldsymbol{\beta}}] = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^tE[\mathbf{Y}] = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

That is, $\boldsymbol{\beta}$ is unbiased. Thanks to Exercise 5.45 and that $\text{Var}[\mathbf{Y}] = \sigma^2\mathbf{I}_n$, we have

$$\begin{aligned} \text{Var}[\hat{\boldsymbol{\beta}}] &= \text{Var}[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}] \\ &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\text{Var}[\mathbf{Y}]((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t)^t \\ &= \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}. \end{aligned}$$

Plugging in the formula of $\hat{\boldsymbol{\beta}}$ and $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, we can verify that the sum of squared errors SSE equals (we omit the tedious but straightforward algebraic details)

$$\begin{aligned} \text{SSE} &= (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^t(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{Y}^t\mathbf{Y} - \mathbf{Y}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y} \\ &= \boldsymbol{\varepsilon}^t\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\boldsymbol{\varepsilon}. \end{aligned}$$

Since ε_i's are assumed to be iid with mean 0 and variance σ^2, it is not hard to verify that $\text{Var}[\boldsymbol{\varepsilon}^t A\boldsymbol{\varepsilon}] = \text{trace}(A) \cdot \sigma^2$ for any $n \times n$ constant matrix A. We also recall from linear algebra that $\text{trace}(AB) = \text{trace}(BA)$ for any $m \times n$ matrix A and $n \times m$ matrix B, which implies that

$$\text{trace}[\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t] = \text{trace}[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}] = \text{trace}(\mathbf{I}_{d+1}) = d + 1.$$

Therefore,

$$\text{SSE} = \sigma^2[\text{trace}(\mathbf{I}_n) - \text{trace}(\mathbf{I}_{d+1})] = \sigma^2(n - d - 1).$$

We complete the proof. ■

Lemma 11.1 shows that the least square estimator $\hat{\boldsymbol{\beta}}$ is unbiased. But how does it compare to other possible estimators? Without any assumptions on the underlying distributions of the noise $\boldsymbol{\varepsilon}$, it is very hard to make a general comparison. Even so, we have the following *Gauss-Markov theorem*, which states that $\hat{\boldsymbol{\beta}}$ is indeed optimal (in the sense of variance) among all linear unbiased estimators. Its proof is deferred to Appendix A.

Theorem 11.2. (Gauss-Markov) *The least square estimator $\hat{\boldsymbol{\beta}}$ has the smallest variance among all linear (with respect to $\mathbf{Y}$) unbiased estimators for $\boldsymbol{\beta}$.*

Exercise 11.1. Show that the least-square estimator $\hat{\boldsymbol{\beta}}$ is indeed the maximum likelihood estimator for $\boldsymbol{\beta}$ when ε_i's are iid normally distributed with mean 0 and variance σ^2 (in this case, it can be shown that $\hat{\boldsymbol{\beta}}$ has the smallest variance among *all* unbiased estimators).

Exercise 11.2. To better understand the method of least square, let us consider a very special linear regression model without any explanatory variables. That is, we assume

$$Y_i = \beta_0 + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

where ε_i's are iid with mean 0 and variance σ^2. This assumption is equivalent to saying that Y_i's are iid with mean β_0 and variance σ^2. Therefore,

estimating β_0 amounts to estimating the population mean from iid samples $\{Y_1, Y_2, \ldots, Y_n\}$. The least square estimator $\hat{\beta}_0$ minimizes the sum of square of errors

$$\text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0)^2.$$

Show that $\hat{\beta}_0$ equals the sample mean of $\{Y_1, Y_2, \ldots, Y_n\}$, and its variance is the smallest among all linear unbiased estimators (Gauss-Markov).

Exercise 11.3. Let $\overline{Y}$ and $\overline{\mathbf{x}} = (\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_d)$ denote the average response and average explanatory variables, respectively. That is,

$$\overline{Y} = \frac{1}{n} \sum_{j=1}^{n} Y_j, \quad \overline{x}_i = \frac{1}{n} \sum_{j=1}^{n} x_{ij}.$$

One can think of the point $(\overline{\mathbf{x}}, \overline{Y})$ as the center of the data. Show that the regression line always passes through it, i.e.,

$$\overline{Y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{x}_1 + \hat{\beta}_2 \overline{x}_2 + \cdots + \hat{\beta}_d \overline{x}_d.$$

Exercise 11.4. When there is only one explanatory variable ($d = 1$, *simple linear regression*), we can denote the observations by (x_i, Y_i), $i = 1, 2, \ldots, n$. Show that the least square estimators are

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = r\sqrt{\frac{S_{yy}}{S_{xx}}}, \quad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x}, \quad \text{SSE} = (1 - r^2) S_{yy}$$

$$\text{Var}[\hat{\beta}_0] = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} \right), \quad \text{Var}[\hat{\beta}_1] = \frac{\sigma^2}{S_{xx}}, \quad \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\overline{x}\sigma^2}{S_{xx}}$$

where

$$S_{xx} \doteq \sum_{i=1}^{n} (x_i - \overline{x})^2, \quad S_{yy} \doteq \sum_{i=1}^{n} (Y_i - \overline{Y})^2, \quad S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(Y_i - \overline{Y})$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}, \quad \overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

Remark 11.1. The *consistency* of the least square estimator $\hat{\boldsymbol{\beta}}$ is dependent on the explanatory variables. This is obvious from Lemma 11.1 since $\text{Var}[\hat{\boldsymbol{\beta}}]$ depends on $(\mathbf{X}^t\mathbf{X})^{-1}$. One can develop very mild sufficient conditions for consistency. For example, when there is only one explanatory variable, it can be shown that $\hat{\boldsymbol{\beta}}$ is consistent if the observed explanatory variables $\{x_1, x_2, \ldots, x_n, \ldots\}$ do not converge as $n \to \infty$.

11.3 Inference on Least Square Estimators

Throughout this section, we will conduct the inference assuming that the sample size n is large. Under this assumption, the least square estimator $\hat{\boldsymbol{\beta}}$ and many of the relevant test statistics are approximately normally distributed. Actually,

1. if ε_i's are normally distributed, then $\hat{\boldsymbol{\beta}}$, as linear combinations of these independent normal random variables, is automatically (jointly) normally distributed, regardless of the sample size n;

2. if ε_i's are not normally distributed, then stronger versions of the central limit theorem can show that $\hat{\boldsymbol{\beta}}$ is approximately jointly normal under mild conditions when the sample size n is large.

This assumption does not prevent us from conveying the main idea of inference on least square estimators. For example, if n is not large, then many of the test statistics are not normal, but have a Student's t-distribution (with appropriate degrees of freedom) under the classical normality assumption on ε_i's. The only modification one needs to make in the inference, such as confidence interval or P-value, is to replace the normal distribution and critical values such as z_α by the appropriate t-distribution and the corresponding critical values t_α.

Thanks to the assumption of large sample size and Lemma 11.1, $\hat{\boldsymbol{\beta}}$ is normal with mean $\boldsymbol{\beta}$ and covariance matrix $\text{Var}[\hat{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$. The population variance parameter σ^2 can be approximated by

$$\hat{\sigma}^2 = \text{SSE}/(n-d-1)$$

Therefore, $\hat{\boldsymbol{\beta}}$ is approximately $N(\boldsymbol{\beta},\ \hat{\sigma}^2(\mathbf{X}^t\mathbf{X})^{-1})$. This allows us to construct confidence intervals and perform hypothesis testing on the population parameter $\boldsymbol{\beta}$.

Example 11.5. The winning speed (mph) for Indianapolis 500 auto races from 1962 to 1971 were as follows.

Year x	1962	1963	1964	1965	1966
Speed Y	140.3	143.1	147.4	151.4	144.3
Year x	1967	1968	1969	1970	1971
Speed Y	151.2	152.9	156.9	155.7	157.7

We will omit the computational details and simply report the results (all these calculations can be easily done on any software such as MATLAB, R, or Python). The regression line is given by

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x = -3469 + 1.84x.$$

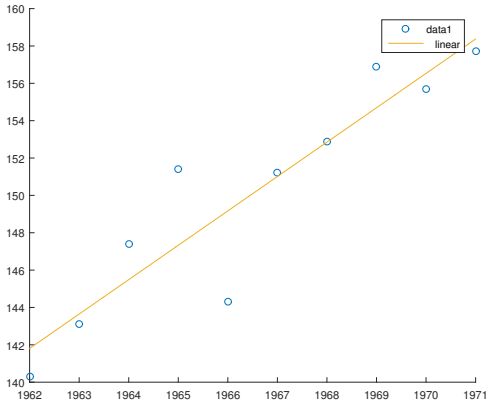


Figure 11.1: Indianapolis 500 auto races

with

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n-2} = 6.57, \quad \text{Var}[\hat{\boldsymbol{\beta}}] = \hat{\sigma}^2(\mathbf{X}^t\mathbf{X})^{-1} = \begin{bmatrix} 555^2 & -157 \\ -157 & 0.28^2 \end{bmatrix}.$$

Estimate the change of winning speed per year and give a 95% confidence interval. Is there statistically significant evidence that the winning speed is increasing every year? What's your prediction for the winning speed at 1974?

Solution: The change of winning speed per year is the parameter β_1. The estimate is $\hat{\beta}_1 = 1.84$ and its 95% confidence interval is

$$\hat{\beta}_1 \pm z_{0.025}\text{S.E.}(\hat{\beta}_1) = 1.84 \pm 2 \cdot 0.28 = 1.84 \pm 0.56.$$

In order to decide if there is statistically significant evidence that the winning speed is increasing every year, we can perform a test with hypotheses

$$H_0 : \beta_1 = 0, \quad H_a : \beta_1 > 0.$$

Under the null hypothesis, $\hat{\beta}_1$ is approximately $N(0, 0.28^2)$. Therefore, the P-value is

$$P(N(0, 0.28^2) \geq 1.84) = P(N(0,1) \geq 6.57) = 2.5 \times 10^{-11}.$$

We conclude that there is statistically significant evidence to support that the winning speed is increasing every year. Finally, the prediction of the winning speed in 1974 is

$$\hat{Y} = -3469 + 1.84 \cdot 1974 = 163.2 \,\text{mph}.$$

Here we would like to remark that if you think the sample size n is not large enough to use normal approximation, the classical approach is to use Student's t-distribution under the normality assumption on ε_i's. All one needs to do is to replace $z_{0.025}$ by $t_{0.025}(n-2) = t_{0.025}(8) = 2.3$ in the construction of the confidence interval and replace $N(0,1)$ by t-distribution with degrees of freedom 8 in the calculation of P-value. This will lead to a slightly wider confidence interval and a larger P-value 8.7×10^{-5}. ■

11.3.1 Estimation vs. Prediction

In the preceding Example 11.5, there is one piece of important detail missing. When we were predicting the winning speed in 1974, we did not submit the error associated with our prediction. It is not that the derivation of its error is impossible. On the contrary, it is quite simple. The important point is to understand what kind of error we are dealing with.

It is probably better to use a different example to explain this issue. Suppose you have built a linear regression model to predict weight Y from height x. You model is

$$Y = \beta_0 + \beta_1 x + \varepsilon.$$

You have collected a bunch of samples, made your estimation $(\hat{\beta}_0, \hat{\beta}_1)$, and calculated all relevant quantities such as SSE, $\text{Var}[\hat{\boldsymbol{\beta}}]$ and so on. You are now facing two problems.

1. What is your estimate for the average weight of a person with height $x = x^*$? What is the error associated with your estimate?

2. Your friend Alan has a friend Ben, whom you have never met. But Alan told you Ben's height is x^*. What is you guess of Ben's weight? What is the error associated with your guess?

Your answers to these two questions would be the same, $\hat{\beta}_0 + \hat{\beta}_1 x^*$, but the interpretations are very different.

The first question is to estimate an unknown but fixed population parameter $\theta \doteq \beta_0 + \beta_1 x^* = E[Y|x^*]$. The second question is to predict the value of a particular response with $x = x^*$, which is a *random variable*. In other words,

1. For the first question, we use $\hat{\theta} = \beta_0 + \beta_1 x^*$ as an estimate to the unknown but fixed population parameter $\theta = \beta_0 + \beta_1 x^*$.

2. For the second question, we use $\hat{Y} = \beta_0 + \beta_1 x^*$ as a prediction to the value of the response variable $Y = \beta_0 + \beta_1 x^* + \varepsilon$.

Even though $\hat{\theta}$ and $\hat{Y}$ have exactly the same formula, their interpretations are very different, so are the associated errors.

The error associated with $\hat{\theta}$ is standard. It is an unbiased estimate for θ with variance (thanks to Exercise 5.45 and Lemma 11.1)

$$\mathrm{Var}[\hat{\theta}] = \mathrm{Var}[\hat{\beta}_0 + \hat{\beta}_1 x^*] = \mathrm{Var}[\mathbf{a}\hat{\boldsymbol{\beta}}] = \mathbf{a}\mathrm{Var}[\hat{\boldsymbol{\beta}}]\mathbf{a}^t = \sigma^2 \mathbf{a}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{a}^t$$

$$\mathbf{a} = [1,\ x^*]$$

The error associated with $\hat{Y}$ is

$$Y - \hat{Y} = (\beta_0 + \beta_1 x^* + \varepsilon) - (\hat{\beta}_0 + \hat{\beta}_1 x^*) = \mathbf{a}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \varepsilon.$$

Since $\hat{\boldsymbol{\beta}}$ is an unbiased estimator for $\boldsymbol{\beta}$ and ε is *independent* of $\hat{\boldsymbol{\beta}}$ because ε is not within the samples that are used to estimate $\hat{\boldsymbol{\beta}}$, we have $E[Y - \hat{Y}] = 0$ and

$$\begin{aligned} \mathrm{Var}[Y - \hat{Y}] &= \mathrm{Var}[\mathbf{a}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})] + \mathrm{Var}[\varepsilon] = \mathbf{a}\mathrm{Var}[\hat{\boldsymbol{\beta}}]\mathbf{a}^t + \sigma^2 \\ &= \sigma^2[\mathbf{a}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{a}^t + 1]. \end{aligned}$$

Naturally, the error associated with the predictor $\hat{Y}$ is always larger than that of the estimator $\hat{\theta}$. The unknown variance parameter σ^2 can be approximated as in Lemma 11.1.

Even though our discussion has been for the model with a single explanatory variable, the extension to multiple explanatory variables is straightforward. The formulae are exactly the same, except that

$$\mathbf{a} = [1,\ x_1^*,\ x_2^*, \ldots,\ x_d^*]$$

when we estimate or predict with explanatory variables $(x_1^*, x_2^*, \ldots, x_d^*)$.

Finally, as a side note, the standard error associated with the predictor $\hat{Y}$ in Example 11.5 is

$$\text{S.E.}(\hat{Y}) = \hat{\sigma}\sqrt{\mathbf{a}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{a}^t + 1} = 3.4, \quad \mathbf{a} = [1,\ 1974];$$

and the 95% *prediction interval* is

$$\hat{Y} \pm z_{0.025}\text{S.E.}(\hat{Y}) = 163.2 \pm 2 \cdot 3.4 = 163.2 \pm 6.8.$$

You can use $t_{0.025}(8) = 2.3$ in place of $z_{0.025}$ as we have discussed in the example. The actual winning speed was 158.6mph in 1974.

11.4 Goodness of Fit

How can we decide if the linear regression model is a good fit of a given data set? This is a tricky question without a perfect answer. There are, however, ways to help us make our decisions. For example, it is always helpful to make various plots of the data set. Visual evidence can be very powerful and sometimes suggest better linear regression models with nonlinear transforms.

In this section, we will focus on a very specific measure of goodness of fit through *analysis of variance*. We consider three types of variations:

1. *Total variation of Y*. The total variation of the response variable Y is defined by
$$\text{SST} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$
You can regard $\overline{Y}$ as our prediction for each observation in a linear regression model without any explanatory variable.

2. *Variation accounted for by regression*. The variation accounted for or explained by the regression model is defined to be
$$\text{SSM} = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2.$$
Once one introduces the regression model into the analysis of the data, the variation of the predicted values for those observations is quantified by SSM.

3. *Variation NOT accounted for by regression*. This is our sum of squared errors:
$$\text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$
The least square estimator $\hat{\boldsymbol{\beta}}$ minimizes this variation.

Lemma 11.3. *In a linear regression model, SST = SSM + SSE. The "variation explained by the regression model" is defined to be*
$$R^2 = \frac{SSM}{SST}.$$

Proof. Recall the matrix notation and the formula of $\hat{\boldsymbol{\beta}}$ in Section 11.2. We have
$$(\mathbf{Y} - \hat{\mathbf{Y}})^t \mathbf{X} = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^t \mathbf{X} = \mathbf{Y}^t\mathbf{X} - \mathbf{Y}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}(\mathbf{X}^t\mathbf{X}) = \mathbf{Y}^t\mathbf{X} - \mathbf{Y}^t\mathbf{X} = \mathbf{0}.$$
In particular,
$$(\mathbf{Y} - \hat{\mathbf{Y}})^t \hat{\mathbf{Y}} = (\mathbf{Y} - \hat{\mathbf{Y}})^t \mathbf{X}\hat{\boldsymbol{\beta}} = 0, \quad (\mathbf{Y} - \hat{\mathbf{Y}})^t \mathbf{X}_1 = 0$$
where $\mathbf{X}_1$ denotes the first column of $\mathbf{X}$, that is, $\mathbf{X}_1$ is $d \times 1$ column vector with every component 1. These two equalities amount to
$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)\hat{Y}_i = 0, \quad \sum_{i=1}^{n} (Y_i - \hat{Y}_i) = 0.$$
Write $Y_i - \overline{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \overline{Y})$. It follows that
$$\begin{aligned} \text{SST} &= \text{SSM} + \text{SSE} + 2\sum_{i=1}^{n} (Y_i - \hat{Y}_i)(\hat{Y}_i - \overline{Y}) \\ &= \text{SSM} + \text{SSE} + 2\sum_{i=1}^{n} (Y_i - \hat{Y}_i)\hat{Y}_i - 2\overline{Y}\sum_{i=1}^{n} (Y_i - \hat{Y}_i) \\ &= \text{SSM} + \text{SSE}. \end{aligned}$$
We complete the proof. ∎

When there is a single explanatory variable, one can show that $R^2 = r^2$, where r is akin to the "correlation coefficient" between the explanatory variable and response variable; see Exercise 11.4. In general, $0 \leq R^2 \leq 1$. In the "ideal" case where the predictions coincide with the observations perfectly (all observations lie on the regression line), then $R^2 = 1$, the variation

in the data is completely explained by the regression model. This leads to the rule of thumb that a regression with a large R^2 is considered a good fit in general. But how large is large depends on the context and there is no absolute answer to this question. The R^2 in Example 11.5 is about 84%. Also taken into consideration of the visual evidence, the linear regression model seems to be a good fit.

There is an important remark we wish to make. Using R^2 as a singular measure for the goodness of fit is NOT suggested. Indeed, it can be detrimental. A good fit should not only explain a given data set, but also provide reasonable power in predicting future observations. One can easily fit the data in Example 11.5 *perfectly* with a polynomial of degree $n - 1 = 9$ (which is a linear regression model with explanatory variables $x_1 = x$, $x_2 = x^2$, ... ,$x_9 = x^9$). This leads to $R^2 = 1$. But such a regression model grossly overfits the data and has little prediction power. R^2 should be combined with other evidence in the study of goodness of fit.

11.5 Words of Caution

Linear regression, as one of most popular models in statistical inference, does have its own pitfalls. Here we should list but a few of them.

- *Sensitivity to outliers*. Outliers in a data set can easily skew the regression line, due to the squared errors. In the following figure, there

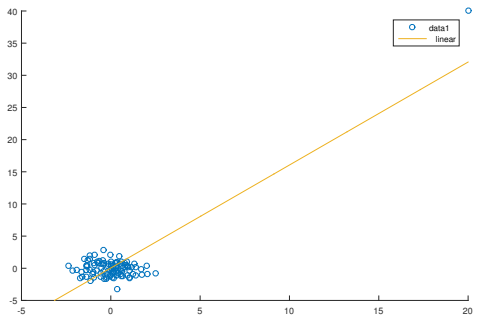


Figure 11.2: Outliers in linear regression

are 100 points whose coordinates are all generated from independent standard normals. Then an artificial outlier $(20, 40)$ is added (the upper right corner). This lifts the original $R^2 = 0.4\%$ (without outlier, as

it should be) to $R^2 = 74\%$ (with outlier). The regression line, which should be roughly horizontal, now has a slop $\hat{\beta}_1 = 1.6$ with a standard error of "merely" 0.1.

- *Danger of overfitting:* We have mentioned the danger of overfitting in the discussion of R^2. In general, the symptom of overfitting is that the model has a very good fit to the given data set, but performs very poorly in predicting new observations. This usually happens when too many explanatory variables are included into the model. One can use more advanced analysis of variance (ANOVA) to perform hypothesis testing on whether a group of independent variables are contributing to explaining the variations in the data. Alternatively, one can also use methods such as cross-validation to test the predictive power of the model. What cross-validation does is to split the data into two parts. The model's parameters are estimated from one part of the data, and its predictive power is tested on the other.

- *Simpson Paradox:* This is not exactly a problem or a paradox only relevant to linear regression. It is actually a general phenomenon in statistical inference that illustrates the danger of *lurking variables*. In linear regression, it may lead to completely erroneous conclusions. It is probably best explained by the following figure. Should X and Y be positively correlated or negatively correlated?

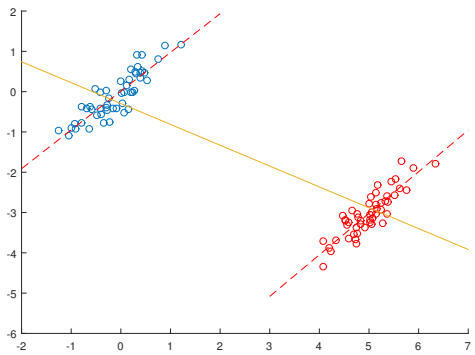


Figure 11.3: Simpson Paradox

Appendix A

Collection of Proofs

- **Proof of Lemma 3.2**. The proof for the case of Bernoulli is trivial. We only need to prove for the cases of binomial, geometric and Poisson random variables. Our approach is via *moment generating function*. Given a random variable X, its **moment generating function** is define to be a function $M : \mathbb{R} \to [0, \infty]$ with

$$M(\theta) \doteq E[e^{\theta X}].$$

Taking the n-th derivative with respect to θ on both sides and then letting $\theta = 0$, we have

$$E[X^n] = \frac{d^n}{d\theta^n} M(\theta)\Big|_{\theta=0}.$$

It remains to calculate the moment generating functions. Denote $q = 1 - p$. Suppose X is $B(n, p)$. Its moment generating function is

$$\begin{aligned} M(\theta) &= E[e^{\theta X}] = \sum_{k=0}^{n} e^{\theta k} P(X = k) \\ &= \sum_{k=0}^{n} \binom{n}{k} \left(e^{\theta} p\right)^k q^{n-k} = (e^{\theta} p + q)^n. \end{aligned}$$

Here the last equality follows from binomial expansion. Therefore

$$M'(0) = npe^{\theta}(e^{\theta} p + q)^{n-1}\Big|_{\theta=0} = np$$

and

$$\begin{aligned} M''(0) &= n(n-1)p^2 e^{2\theta}(e^{\theta} p + q)^{n-2} + npe^{\theta}(e^{\theta} p + q)^{n-1}\Big|_{\theta=0} \\ &= n(n-1)p^2 + np. \end{aligned}$$

It follows that $E[X] = M'(0) = np$ and $\text{Var}[X] = E[X^2] - E^2[X] = M''(0) - (np)^2 = np(1-p)$.

Suppose now X is geometric with parameter p. Then its moment generating function is

$$M(\theta) = \sum_{n=1}^{\infty} e^{\theta n} P(X = n) = \sum_{n=1}^{\infty} e^{\theta n} q^{n-1} p = \frac{p}{q} \sum_{n=1}^{\infty} \left(e^{\theta} q\right)^n$$

Note that for θ such that $e^{\theta} q < 1$ or $\theta < -\log(q)$, the infinite sum on the right-hand-side of the preceding display is finite and

$$M(\theta) = \frac{p e^{\theta}}{1 - e^{\theta} q}.$$

It follows that

$$M'(\theta) = \frac{p e^{\theta}}{(1 - e^{\theta} q)^2}, \quad M''(\theta) = \frac{p e^{\theta}}{(1 - e^{\theta} q)^2} + \frac{2 p q e^{2\theta}}{(1 - e^{\theta} q)^3}$$

and $E[X] = M'(0) = 1/p$, and $\text{Var}[X] = E[X^2] - E^2[X] = M''(0) - (1/p)^2 = q/p^2$.

Finally suppose X is a Poisson random variable with parameter λ. Its moment generating function is

$$M(\theta) = \sum_{n=0}^{\infty} e^{\theta n} P(X = n) = \sum_{n=0}^{\infty} e^{\theta n} e^{-\lambda} \frac{\lambda^n}{n!} = e^{\lambda (e^{\theta} - 1)},$$

where we have used the Taylor series

$$e^x = \sum_{n=0}^{n} \frac{x^n}{n!}.$$

It follows that

$$M'(\theta) = \lambda e^{\theta} e^{\lambda (e^{\theta} - 1)}, \quad M''(\theta) = \lambda e^{\theta} e^{\lambda (e^{\theta} - 1)} + \lambda^2 e^{2\theta} e^{\lambda (e^{\theta} - 1)}$$

and $E[X] = M'(0) = \lambda$, $\text{Var}[X] = E[X^2] - E^2[X] = M''(0) - \lambda^2 = \lambda$. We complete the proof. ■

- **Proof of Remark 5.3 (Cauchy-Schwarz inequality).** If $Y = 0$ then the inequality is trivial. Assume now $Y \neq 0$. For any constant t, we have

$$0 \leq h(t) \doteq E[(X - tY)^2] = E[X^2] - 2tE[XY] + t^2 E[Y^2].$$

In particular, if we take $t^* = E[XY]/E[Y^2]$, then

$$0 \leq h(t^*) = E[X^2] - \frac{E^2[XY]}{E[Y^2]}.$$

The inequality follows readily. It is an equality if and only if $Y = 0$ or $h(t^*) = 0$, if and only if $Y = 0$ or $X - t^*Y = 0$, if and only if $aX + bY = 0$ for some constants a and b that are not both zero. We complete the proof. ■

- **Proof of Lemma 10.1.** Let $R \subseteq \mathbb{R}^n$ be an arbitrary rejection region whose probability of type I error is bounded by α, that is, $P(R|H_0) \leq \alpha$. It suffices to show that $P(R_{b^*}|H_a) \geq P(R|H_a)$.

  To simplify notation, we denote $x = (x_1, \ldots, x_n)$, $dx = dx_1 \cdots dx_n$, $L_0(x) = f_0(x_1) \cdots f_0(x_n)$, and $L_a(x) = f_a(x_1) \cdots f_a(x_n)$. Then

$$\begin{aligned} P(R_{b^*}|H_a) - P(R|H_a) &= \int_{R_{b^*}} L_a(x)dx - \int_R L_a(x)dx \\ &= \int_{R_{b^*} \setminus R} L_a(x)dx - \int_{R \setminus R_{b^*}} L_a(x)dx. \end{aligned}$$

  Note that by the definition of R_{b^*}, $L_a(x) \geq b^* L_0(x)$ on the set $R_{b^*} \setminus R$ and $L_a(x) < b^* L_0(x)$ on the set $R \setminus R_{b^*}$. It follows that

$$\begin{aligned} P(R_{b^*}|H_a) - P(R|H_a) &\geq b^* \int_{R_{b^*} \setminus R} L_0(x)dx - b^* \int_{R \setminus R_{b^*}} L_0(x)dx \\ &= b^* \int_{R_{b^*}} L_0(x)dx - b^* \int_R L_0(x)dx \end{aligned}$$

  The first integral on the right-hand-side equals $P(R_{b^*}|H_0) = \alpha$ and the second integral equals $P(R|H_0) \leq \alpha$. We complete the proof. ■

- **Proof of Theorem 11.2.** Consider any linear unbiased estimator $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\beta}$. Since $\hat{\boldsymbol{\theta}}$ is linear, we can write $\hat{\boldsymbol{\theta}} = \mathbf{cY}$ for some $(d+1) \times n$ matrix $\mathbf{c}$ of constants. The unbiasedness of $\hat{\theta}$ implies that

$$\boldsymbol{\beta} = E[\hat{\boldsymbol{\theta}}] = E[\mathbf{cY}] = \mathbf{cX}\boldsymbol{\beta}$$

  or $\mathbf{cX} = \mathbf{I}_{d+1}$, where $\mathbf{I}_{d+1}$ represents the identity matrix with dimension $(d+1) \times (d+1)$. The covariance matrix for $\hat{\boldsymbol{\theta}}$ is

$$\text{Var}[\hat{\boldsymbol{\theta}}] = \mathbf{c}\text{Var}[\mathbf{Y}]\mathbf{c}^t = \mathbf{c}\sigma^2\mathbf{I}_n c^t = \sigma^2 \mathbf{c}\mathbf{c}^t.$$

We would like to show that $\mathrm{Var}[\hat{\boldsymbol{\theta}}] - \mathrm{Var}[\hat{\boldsymbol{\beta}}]$ or equivalently $\mathbf{c}\mathbf{c}^t - (\mathbf{X}^t\mathbf{X})^{-1}$ is a *positive-semidefinite* matrix. To this end, define

$$\boldsymbol{\delta} \doteq \mathbf{c} - (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t.$$

It follows that

$$\boldsymbol{\delta}\mathbf{X} = [\mathbf{c} - (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t]\mathbf{X} = \mathbf{c}\mathbf{X} - \mathbf{I}_{d+1} = \mathbf{0},$$

and thus

$$\begin{aligned} \mathbf{c}\mathbf{c}^t - (\mathbf{X}^t\mathbf{X})^{-1} &= [(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t + \boldsymbol{\delta}][(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t + \boldsymbol{\delta}]^t - (\mathbf{X}^t\mathbf{X})^{-1} \\ &= [(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t + \boldsymbol{\delta}][\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1} + \boldsymbol{\delta}^t] - (\mathbf{X}^t\mathbf{X})^{-1} \\ &= \boldsymbol{\delta}\boldsymbol{\delta}^t, \end{aligned}$$

which is automatically a positive-semidefinite matrix. We complete the proof. ■