STAT 230 PROBABILITY

Course Notes by Chris Springer
Revised by Jerry Lawless, Don McLeish, Cyntha Struthers, and Steve Drekic

Spring 2025 Edition

Contents

Chapter 1

Introduction to Probability

1.1 Definitions of Probability

Randomness, in some intuitive sense, is all around us. We experience apparently random events everywhere in our daily lives, and even in the classroom. Understanding randomness is essential for modern life, as it forms the basis of our decision-making process under situations involving uncertainty. Much of the real world, from the biological sciences to the business marketplace, involves uncertainty and variability. For example, it is uncertain whether it will rain tomorrow; the price of a given stock a week from today is uncertain; the number of claims that a car insurance policy holder will make over a one-year period is uncertain. Uncertainty or "randomness" (i.e., variability of results) is usually due to some mixture of at least two factors including: (1) *variability in populations* consisting of animate or inanimate objects (e.g., people vary in height, weight, blood type, etc.), and (2) *variability in processes* or phenomena (e.g., the random selection of 6 numbers from 49 in a lottery draw can lead to a very large number of different outcomes).

Variability and uncertainty in a system make it more difficult to plan or to make decisions without suitable tools. We cannot eliminate uncertainty but it is usually possible to describe, quantify, and deal with variability and uncertainty using the theory of probability. This course develops both the mathematical theory and some of the applications of probability. The applications of this methodology are far-reaching, from finance to the life-sciences, from the analysis of computer algorithms to the simulation of queues and networks or the spread of epidemics. We do not have the time in this course to develop these applications in detail, but some of the end-of-section problems will give a hint of the extraordinary range of application of the mathematical theory of probability and statistics.

It seems logical to begin by defining probability. People have attempted to do this by giving

definitions that reflect the uncertainty whether some specified outcome or "event" will occur in a given setting. The setting is often termed an "experiment" or "process" for the sake of discussion. We often consider simple examples: it is uncertain whether the number 2 will turn up when a six-sided die is rolled. It is similarly uncertain whether the Canadian dollar will be higher tomorrow, relative to the U.S. dollar, than it is today. One step in defining probability requires envisioning a random experiment with a number of possible outcomes. We refer to the set of all possible distinct outcomes to a random experiment as the **sample space** (usually denoted by S). Groups or sets of outcomes of possible interest (i.e., subsets of the sample space), we will call events. Then we might define probability in three different ways:

(1) The **classical** definition: The probability of some event is

$$\frac{\text{number of ways the event can occur}}{\text{number of outcomes in } S},$$

provided all points in the sample space S are *equally likely*. For example, when a die is rolled, the probability of getting a 2 is $\frac{1}{6}$ because one of the six faces is a 2.

(2) The **relative frequency** definition: The probability of an event is the (limiting) proportion (or fraction) of times the event occurs in a very long series of repetitions of an experiment or process. For example, this definition could be used to argue that the probability of getting a 2 from a rolled die is $\frac{1}{6}$.

(3) The **subjective probability** definition: The probability of an event is a measure of how sure the person making the statement is that the event will happen. For example, after considering all available data, a weather forecaster might say that the probability of rain today is 30% or 0.3.

Unfortunately, all three of these definitions have serious limitations, which we address below:

Classical Definition: What does "equally likely" mean? This appears to use the concept of probability while trying to define it! We could remove the phrase "provided all outcomes are equally likely", but then the definition would clearly be unusable in many settings where the outcomes in S did not tend to occur equally often.

Relative Frequency Definition: Since we can never repeat an experiment or process indefinitely, we can never know the probability of any event from the relative frequency definition. In many cases, we cannot even obtain a long series of repetitions due to time, cost, or other limitations. For example, the probability of rain today cannot really be obtained by the relative frequency definition since today cannot be repeated again under identical conditions. Intuitively, however, if a probability is correct,

we expect it to be close to relative frequency, when the experiment is repeated many times.

Subjective Probability Definition: This definition gives no rational basis for people to agree on a right answer, and thus would disqualify probability as an objective science. Are everyone's opinions equally valid or should we only consult "experts"? There is some controversy about when, if ever, to use subjective probability except for personal decision-making but it does play a part in a branch of statistics that is often called "Bayesian Statistics". This type of statistics will not be discussed in these course notes, but it is a common and useful method for updating subjective probabilities with objective experimental results.

The difficulties in producing a satisfactory definition can be overcome by treating probability as a mathematical system defined by a set of axioms. We do not worry about the numerical values of probabilities until we consider a specific application. This is consistent with the way that other branches of mathematics are defined and then used in specific applications (e.g., the way calculus and real-valued functions are used to model and describe the physics of gravity and motion).

The mathematical approach that we will develop is based on the following description of a **probability model**:

- a sample space of all possible outcomes of a random experiment is defined,
- a set of events, subsets of the sample space to which we can assign probabilities, is defined,
- a mechanism for assigning probabilities (i.e., numbers between 0 and 1) to events is specified.

In order to understand the material in these course notes, you may need to review your understanding of basic counting arguments, elementary set theory, as well as some of the important series and integrals that you have encountered in calculus that provide a basis for some of the probability distributions discussed in these course notes. In the next section, we begin a more mathematical description of probability theory.

Section 1.1 Problems

1.1.1 Try to think of examples of probabilities you have encountered which might have been obtained by each of the three "definitions".

1.1.2 Which definitions do you think could be used for obtaining the following probabilities?

(a) You have a claim on your car insurance in the next year.

(b) There is a meltdown at a nuclear power plant during the next 5 years.

(c) A person's birthday is in April.

1.1.3 Give examples of how probability applies to each of the following areas:

(a) Lottery draws.

(b) Auditing of expense items in a financial statement.

(c) Disease transmission (e.g., measles, tuberculosis, STDs).

(d) Public opinion polls.

1.1.4 Which of the following can be accurately described by a "deterministic" model (i.e., a model which does not require any concept of probability)?

(a) The position of a small particle in space.

(b) The velocity of an object dropped from the Leaning Tower of Pisa.

(c) The value of a stock which you purchased for \$20 one month ago.

(d) The number of earthquakes in California in one year.

1.2 Mathematical Probability Models

Consider some phenomenon or process that is repeatable, at least in theory, and suppose that certain outcomes from observing the phenomenon are defined. We will often term the phenomenon or process an "experiment" and refer to a single repetition of the experiment as a "trial". We use the following terminology and notation to describe the collection of potential outcomes of the experiment.

Definition 1.2.1. *A **sample space** S is a set of distinct outcomes for an experiment or process, with the property that in a single trial, one and only one of these outcomes occurs.*

The outcomes that make up the sample space may sometimes be called "sample points" or just "points" on occasion. A sample space is defined as part of the probability model in a given setting but it is not necessarily uniquely defined. For example, if we roll a six-sided die and define the outcomes

$$a_i = \text{top face is } i \text{ for } i = 1, 2, 3, 4, 5, 6,$$

then we could take the sample space as $S = \{a_1, a_2, a_3, a_4, a_5, a_6\}$. Instead of using this definition of the sample space, however, we could instead define the outcomes

$$\begin{aligned} E &= \text{an even number turns up, and} \\ O &= \text{an odd number turns up.} \end{aligned}$$

Then we could take the sample space as $S = \{E, O\}$. Note that both sample spaces satisfy Definition 1.2.1. Which one we use depends on what we wanted to use the probability model for. If we expect **never** to have to consider events like "a number less than 3 turns up", then the space $S = \{E, O\}$ will suffice, but in most cases, if possible, we choose sample points that are the smallest possible or "indivisible". Thus, the first sample space is likely the preferred one in this example.

Sample spaces may be either **discrete** or **non-discrete**. S is discrete if it consists of a finite or countably infinite set of sample points. Recall that a countably infinite sequence is one that can be put into a one-to-one correspondence with a subset of the positive integers, so for example $\{\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \dots\}$ is countably infinite as is the set of all rational numbers. The two sample spaces in the preceding example are clearly discrete. A sample space $S = \{1, 2, 3, \dots\}$ consisting of all the positive integers is discrete, but a sample space $S = \{x : x > 0\}$ consisting of all positive real numbers is not. We will initially focus most of our attention on discrete sample spaces. For discrete sample spaces, it is easier to specify the class of events to which we may wish to assign probabilities. In particular, we will allow **all possible subsets** of the sample space. For example, if $S = \{a_1, a_2, a_3\}$ is the sample space, then

$$\mathcal{S} = \{\emptyset, \{a_1\}, \{a_2\}, \{a_3\}, \{a_1, a_2\}, \{a_1, a_3\}, \{a_2, a_3\}, S\}$$

represents the set of all events derived from the sample space S where $\emptyset$ denotes the *empty event* or the so-called *null set*.

Definition 1.2.2. *An event is a subset $A \subseteq S$. If the event is indivisible so it contains only one point (e.g., $A_1 = \{a_1\}$), we call it a **simple event**. An event A made up of two or more simple events is called a **compound event.***

Remark: Our notation will often not distinguish between the point a_i and the simple event $A_i = \{a_i\}$ which has this point as its only element, although they differ as mathematical objects.

We are interested in defining the probability of observing events resulting from an experiment. In order to begin to describe how this is done, let us start with this: the probability of an event A, denoted by $P(A)$, will be a number between 0 and 1. For example, when an experiment is repeatable and we wish to define probability via the relative frequency definition, we might define $P(A)$ as the limiting proportion of experimental trials in which we observe an outcome belonging to the event A. Since we wish to define this value for any potential event A, it is natural to view probability as being a *function*. A probability function, which in more advanced courses is usually termed a "probability measure", takes as its input an event, and outputs a number between 0 and 1 (i.e., $P(\cdot)$ is a function such that $P : \mathcal{S} \mapsto [0, 1]$). We often also refer to the probability function $P(\cdot)$ as a "probability model", since it provides a model for the experiment by encoding at what frequency or "chance" we expect to

observe events or outcomes.

For probability to be a useful mathematical concept, it should possess some other properties. For instance, if our experiment consists of tossing a coin with two sides, "Head" and "Tail", then we might wish to consider the two events A_1 = "Head turns up" and A_2 = "Tail turns up". It does not make much sense to allow for probability assignments $P(A_1) = 0.6$ and $P(A_2) = 0.6$. Why? Intuitively, we expect to have that $P($ "Head or Tail turns up"$) = 1$, since a coin flip (unless it can land on its edge!) is expected to always land with one of its sides turned up. On the other hand, since the coin cannot simultaneously land with its "Head" and "Tail" sides turned up, we expect to have that $P($ "Head or Tail turns up"$) = P(A_1) + P(A_2)$. Clearly, these properties are not compatible if also $P(A_1) = P(A_2) = 0.6$.

In order to rule out such "intuition-violating" definitions of probability, we will assume that the probability function $P(\cdot)$ satisfies certain fundamental rules. Throughout these course notes, we often restrict our attention to the case when the sample space is discrete. In this case, there is a relatively simple recipe for defining a proper probability model: we start by specifying the probabilities of all of the simple events, and then define the probability of a compound event as the sum of the probabilities of the simple events that make it up. In what follows, when we wish to write the probability of the simple event $A_1 = \{a_1\}$, we should write $P(A_1)$ or $P(\{a_1\})$, but the latter is often shortened to $P(a_i)$ for the sake of convenience.

Definition 1.2.3. *Let $S = \{a_1, a_2, a_3, \dots\}$ be a discrete sample space. Assign numbers (i.e., **probabilities**) $P(a_i)$, $i = 1, 2, 3, \dots$, to the a_i's such that the following two conditions hold:*

(1) $0 \le P(a_i) \le 1$,

(2) $\sum\limits_{\text{all } i} P(a_i) = \sum\limits_{\text{all } a_i} P(a_i) = 1.$

*The set of probabilities $\{P(a_i), i = 1, 2, 3, \dots\}$ is called a **probability distribution on** S.*

Note that $P(\cdot)$ is a function whose domain may be thought of as the sample space S in this case. The condition $\sum\limits_{\text{all } i} P(a_i) = 1$ above reflects the idea that when the process or experiment happens, one or other of the simple events in S must occur (recall that the sample space includes all possible outcomes). The probability of a more general event A (not necessarily a simple event) is then defined as follows:

Definition 1.2.4. *The probability $P(A)$ of an event A is the sum of the probabilities for all the simple events that make up A, or simply $P(A) = \sum\limits_{a \in A} P(a)$.*

For example, the probability of the compound event $A = \{a_1, a_2, a_3\}$ is $P(a_1) + P(a_2) + P(a_3)$. We note that Definition 1.2.3 does not specify what particular probabilities to assign to the simple events for a given application, only those properties guaranteeing mathematical consistency. In an actual application of a probability model, we try to specify numerical values of the probabilities that are more or less consistent with the frequencies of events when the experiment is repeated. In other words, we try to specify probabilities that are consistent with the real world. There is nothing mathematically wrong with a probability model for a toss of a coin that specifies that the probability of heads is zero, except that it likely won't agree with the frequencies we obtain when the experiment is repeated with a balanced (i.e., fair) coin.

Along these same lines, suppose that a six-sided die is rolled and let the sample space be represented by $S = \{1, 2, 3, 4, 5, 6\}$, where 1 represents the simple event that the number of dots showing on the upturned face is 1, and so on. If the die is a fair one, we would likely define probabilities as

$$P(i) = \frac{1}{6} \text{ for } i = 1, 2, 3, 4, 5, 6, \tag{1.2.1}$$

because if the die were tossed repeatedly by a fair roller (as in some games or gambling situations), then each number would occur close to $\frac{1}{6}$ of the time. However, if the die were weighted in some way, or if the roller were able to manipulate the die so that 1 is more likely, these numerical values would not be so useful. To have a useful mathematical model, some degree of compromise or approximation is usually required. Is it likely that the die or the roller are perfectly "fair"? Given (1.2.1), if we wish to consider some compound event of interest, the probability is easily obtained. For example, if $A =$ "number of dots showing on upturned face is even", then because $A = \{2, 4, 6\}$, we get

$$P(A) = P(2) + P(4) + P(6) = \frac{3}{6} = \frac{1}{2}.$$

We now consider some additional examples, starting with some simple problems involving cards, coins, and dice. Once again, to calculate probability for discrete sample spaces, we usually approach a given problem using three steps:

(1) Specify a sample space S.

(2) Assign a probability distribution to the simple events in S.

(3) For any compound event A, calculate $P(A)$ by adding the probabilities of all the simple events that make up A.

When S has only a few points, one of the easiest methods for finding the probability of an event is to list all outcomes. However, we will soon discover that having a detailed specification or list of the

elements of the sample space may prove difficult to obtain. Indeed in many cases the sample space is so large that at best we can describe it in words. For the moment, we will solve problems that are stated as "find the probability that..." by carrying out step (2) above, assigning probabilities that we expect should reflect the long run relative frequencies of the simple events in repeated trials, and then summing these probabilities to obtain $P(A)$. In many problems, a sample space S with equally probable simple events can be used, and the first few examples are of this type.

Example 1.2.1. Draw one card from a standard well-shuffled deck of cards, comprised of 13 cards (i.e., $2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A$) in each of 4 distinct suits: diamonds (♦), hearts (♥), spades (♠), and clubs (♣). Find the probability that the card drawn is a club.

Solution 1: Let $S = \{\diamondsuit, \heartsuit, \spadesuit, \clubsuit\}$. Then S has 4 points, with 1 of them being "club", so $P(\clubsuit) = \frac{1}{4}$. ■

Solution 2: Consider the sample space

$$S = \{2\diamondsuit, 3\diamondsuit, \ldots, A\diamondsuit, 2\heartsuit, 3\heartsuit, \ldots, A\heartsuit, 2\spadesuit, 3\spadesuit, \ldots, A\spadesuit, 2\clubsuit, 3\clubsuit, \ldots, A\clubsuit\}.$$

Then each of the 52 cards in S has probability $\frac{1}{52}$. If A denotes the event of interest, then

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

and this event has 13 simple outcomes in it all with the same probability $\frac{1}{52}$. Therefore,

$$P(A) = \underbrace{\frac{1}{52} + \frac{1}{52} + \cdots + \frac{1}{52}}_{\text{13 terms}} = \frac{13}{52} = \frac{1}{4}.$$

■

Remarks:

(1) A sample space is not necessarily unique, as mentioned earlier. The two solutions above illustrate this. Note that in Solution 1, the event A = "the card is a club" is a simple event because of the way the sample space was defined, but in Solution 2 it is a compound event.

(2) In calculating the probability, we have assumed that each simple event in S is equally probable. For example, in Solution 1, each simple event has probability $\frac{1}{4}$. This seems to be the only sensible choice of numerical value in this setting, but you will encounter problems later on where it is not obvious whether outcomes are all equally probable.

The term "odds" is sometimes used in describing probabilities. In Example 1.2.1, the odds in favour of clubs are 1 to 3. We could also say the odds against clubs are 3 to 1. In general, we have the following definition:

Definition 1.2.5. *The **odds in favour of an event** A is the probability the event occurs divided by the probability it does not occur, or simply* $\frac{P(A)}{1-P(A)}$. *The **odds against the event** is the reciprocal of this quantity, or simply* $\frac{1-P(A)}{P(A)}$.

For example, if the odds against a given horse winning a race are 20 to 1 (or simply written as 20:1), what is the probability of the event A = "the horse will win the race"? According to the definition above, $\frac{1-P(A)}{P(A)} = 20$, which yields $P(A) = \frac{1}{21}$. Keep in mind, of course, that these odds are derived from the bettor's collective opinion and are therefore subjective.

Example 1.2.2. Toss a coin twice. Find the probability of getting one head. (Generally speaking, "one head" is taken to mean exactly one head. If we meant "at least one head", we would say so.)

Solution 1: Let $S = \{HH, HT, TH, TT\}$ and assume the simple events each have probability $\frac{1}{4}$. (Here, the notation HT means head on the 1st toss and tail on the 2nd toss.) Since one head occurs for simple events HT and TH, the event of interest is $A = \{HT, TH\}$ and we get $P(A) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$. ■

Solution 2: Let $S = \{0 \text{ heads}, 1 \text{ head}, 2 \text{ heads}\}$ and assume the simple events each have probability $\frac{1}{3}$. Then, in this case, $P(1 \text{ head}) = \frac{1}{3}$. ■

Which of these two solutions is right? Both are mathematically "correct" in the sense that they are both consequences of probability models. However, we want a solution that reflects the relative frequency of occurrence in repeated trials in the real world, not just one that agrees with some mathematical model. In that respect, the points in Solution 2 are **not** equally likely. The simple event {1 head} occurs more often than either {0 heads} or {2 heads} in actual repeated trials.

Figure 1.2.1: Ten tosses of two coins.

You can experiment to verify this. For example, of the 10 replications of the experiment in Figure 1.2.1, 2 heads occurred 2 of the 10 times, 1 head occurred 7 of the 10 times, and 0 heads occurred 1 of the 10 times. For more certainty, you should replicate this experiment many times. So we say Solution 2 is incorrect for ordinary fair coins because it is based on an incorrect model. If we were determined to use the sample space in Solution 2, we could do it by assigning appropriate probabilities to each of the three simple events but then 0 heads would need to have a probability of $\frac{1}{4}$, 1 head a probability of

$\frac{1}{2}$, and 2 heads a probability of $\frac{1}{4}$. We do not usually do this because there seems little point in using a sample space whose points are not equally probable when one with equally probable points is readily available.

Example 1.2.3. Roll a red die and a green die. Find the probability of the event A = "the total number of dots showing on the upturned faces is 5".

Solution: Let (x, y) represent getting x on the red die and y on the green die. Then, with these as simple events, the sample space S is comprised of:

$$\begin{array}{ccccc} (1,1) & (1,2) & (1,3) & \cdots & (1,6) \\ (2,1) & (2,2) & (2,3) & \cdots & (2,6) \\ (3,1) & (3,2) & (3,3) & \cdots & (3,6) \\ \vdots & \vdots & \vdots & & \vdots \\ (6,1) & (6,2) & (6,3) & \cdots & (6,6). \end{array}$$

Each simple event, for example $\{(1,1)\}$, is assigned a probability of $\frac{1}{36}$. For the event of interest, $A = \{(1,4),(2,3),(3,2),(4,1)\}$ and therefore $P(A) = \frac{4}{36} = \frac{1}{9}$. ■

As a follow-up to the above example, suppose instead the 2 dice were identical in colour. Since we can no longer distinguish between (x, y) and (y, x), the only distinguishable points in S are:

$$\begin{array}{ccccc} (1,1) & (1,2) & (1,3) & \cdots & (1,6) \\ & (2,2) & (2,3) & \cdots & (2,6) \\ & & (3,3) & \cdots & (3,6) \\ & & & \ddots & \vdots \\ & & & & (6,6). \end{array}$$

Using this sample space, the event A is represented as $A = \{(1,4),(2,3)\}$. If we assign equal probabilities of $\frac{1}{21}$ to each simple event, then we get $P(A) = \frac{2}{21}$.

At this point, you should be suspicious since $\frac{2}{21} \neq \frac{1}{9}$. The colour of the dice should not have any effect on what total we get. The universe does not change the frequency of real physical events depending on whether the dice are identical or not, so one answer must be wrong! The problem is that the 21 points in S here are not equally likely. There was nothing theoretically wrong with the probability model except that if this experiment is repeated in the real world, the point $(1, 2)$ occurs about twice as often in the long run as the point $(1, 1)$. Therefore, the only sensible way to use this sample space so it is consistent with the real world is to assign probabilities of $\frac{1}{36}$ to the points of the form (x, x) and $\frac{2}{36}$ to the points (x, y) for $x \neq y$. We can compare these probabilities with experimental evidence. For example, when we threw virtual dice (using computer software) up to 10,000 times

and recorded the results, there were 121 occasions when the total on the dice was 5, indicating the probability of the event A should be close to $\frac{121}{1000}$ or 0.121. This compares more favourably with the earlier probability $P(A) = \frac{1}{9} = 0.111$ from Example 1.2.3.

For a more straightforward solution to the above problem, let us pretend the dice can be distinguished even though they cannot. (Imagine, for example, that we put a tiny mark on one die, or label one of them differently.) We then get the same 36 sample points as in Example 1.2.3. The fact that one die has a tiny mark cannot change the probabilities, so that $P(A) = \frac{4}{36} = \frac{1}{9}$. The laws determining the probabilities associated with these two dice do not, of course, know whether your eyesight is so keen that you can or cannot distinguish the dice. These probabilities must be the same in either case. In many problems when objects are indistinguishable and we are interested in calculating a probability, you will discover that the calculation is made easier by pretending the objects can be distinguished. This illustrates a common pitfall. When treating objects in an experiment as distinguishable leads to a different answer from treating them as identical, the points in the sample space for identical objects are usually not "equally likely" in terms of their long run relative frequencies. It is generally safer to pretend objects can be distinguished even when they cannot be, in order to get equally likely sample points.

While the method of finding a probability by listing all the points in S can be useful, it is not practical when there are a lot of points to write out (e.g., if 3 dice were tossed, there would be $6^3 = 216$ points in S). We need to have a more systematic procedure for determining the number of outcomes in S or in a compound event without having to list them all. The next section considers ways to do this.

Section 1.2 Problems

1.2.1 Students in a particular program have the same 4 math professors. Two students in the program each independently ask one of their math professors for a letter of reference. Assume each student is equally likely to ask any of the math professors.

(a) List a sample space for this "experiment".

(b) Use this sample space to determine the odds in favour of both students asking the same professor for a letter of reference.

1.2.2 Toss a fair coin 3 times.

(a) List a sample space for this "experiment".

(b) What are the odds against getting exactly 2 tails?

1.2.3 You wish to choose 2 different numbers without replacement (so the same number cannot be chosen twice) from the set $\{1, 2, 3, 4, 5\}$. List all possible pairs you could obtain, assuming all pairs are equally probable, and find the probability the numbers chosen differ by 1 (i.e., the two numbers are consecutive).

1.2.4 Four letters addressed to individuals W, X, Y, and Z are randomly placed in four addressed envelopes, one letter in each envelope.

(a) List a 24-point sample space for this experiment. Be sure to explain your notation.

(b) List the sample points belonging to each of the following events:
$A =$ "W's letter goes into the correct envelope",
$B =$ "no letters go into the correct envelopes",
$C =$ "exactly two letters go into the correct envelopes", and
$D =$ "exactly three letters go into the correct envelopes".

(c) Assuming that the 24 sample points are equally probable, find the probabilities of the four events in part (b).

1.2.5 (a) Three balls are placed at random into three boxes, with no restriction on the number of balls per box. List the 27 possible outcomes of this experiment. Be sure to explain your notation. Assuming that the outcomes are all equally probable, find the probability of each of the following events:
$A =$ "the first box is empty",
$B =$ "the first two boxes are empty", and
$C =$ "no box contains more than one ball".

(b) Find the probabilities of events A, B, and C when three balls are placed at random into n boxes where $n \geq 3$.

(c) Find the probabilities of events A, B, and C when k balls are placed at random into n boxes where $n \geq k$.

1.2.6 **Diagnostic Tests**. Suppose that in a large population, some persons have a specific disease at a given point in time. A person can be tested for the disease, but inexpensive tests are often imperfect, and may give either a "false positive" result (i.e., the person does not have the disease but the test says they do) or a "false negative" result (i.e., the person has the disease but the test says they do not).

In a random sample of 1000 people, individuals with the disease were identified according to a completely accurate but expensive test, and also according to a less accurate but inexpensive test. The results for the less accurate test were as follows:

- 920 persons without the disease tested negative,
- 60 persons without the disease tested positive,
- 18 persons with the disease tested positive, and
- 2 persons with the disease tested negative.

(a) Estimate the fraction of the population that has the disease and tests positive using the inexpensive test.

(b) Estimate the fraction of the population that has the disease.

(c) Suppose that someone randomly selected from the same population as those tested above was administered the inexpensive test and it indicated positive. Based on the above information, how would you estimate the probability that they actually have the disease?

1.2.7 **Machine Recognition of Handwritten Digits**. Suppose that you have an optical scanner and associated software for determining which of the digits $0, 1, \ldots, 9$ an individual has written in a square box. The system may of course be wrong sometimes, depending on the legibility of the handwritten number.

(a) Describe a sample space S that includes points (x, y), where x stands for the number actually written, and y stands for the number that the machine identifies.

(b) Suppose that the machine is asked to identify very large numbers of digits, of which $0, 1, \ldots, 9$ occur equally often, and suppose that the following probabilities apply to the points in your sample space:

$P((0,6)) = P((6,0)) = 0.004$; $P((0,0)) = P((6,6)) = 0.096$;
$P((5,9)) = P((9,5)) = 0.005$; $P((5,5)) = P((9,9)) = 0.095$;
$P((4,7)) = P((7,4)) = 0.002$; $P((4,4)) = P((7,7)) = 0.098$;
$P((y,y)) = 0.1$ for $y = 1, 2, 3, 8$.

Construct a table with probabilities for each point (x, y) in S. What fraction of numbers is correctly identified?

1.3 Counting in Uniform Probability Models

We have now seen how a number of probability problems can be solved by specifying a sample space $S = \{a_1, a_2, \ldots, a_n\}$ in which each outcome (i.e., simple event) of S has probability $\frac{1}{n}$ (i.e., is "equally likely"). This is referred to as a *uniform distribution* over the set $\{a_1, a_2, \ldots, a_n\}$. If a compound event A contains m equally likely outcomes, then $P(A) = \frac{m}{n}$ from Definition 1.2.4. In other words, we need to be able to count the number of outcomes in S as well as those in A. We begin by looking at some formal counting methods to help calculate probabilities in uniform models.

There are two fundamental rules for counting, phrased in terms of "jobs" which are to be done. For convenience, we state these rules in the case of two jobs only, but they naturally extend to more than two jobs.

(1) The **Addition Rule:** *Suppose we can do job 1 in p ways and job 2 in q ways. Then we can do either job 1 **OR** job 2 (but not both) in $p + q$ ways.*

For example, suppose a lecture has 30 undergraduate students and 40 graduate students in attendance. There are $30 + 40 = 70$ ways the instructor can pick one student to answer a question. If there are 5 vowels and 20 consonants on a list and we must pick one letter, this can be done in $5 + 20 = 25$ ways.

(2) The **Multiplication Rule:** *Suppose we can do job 1 in p ways and, **for each of these ways**, we can do job 2 in q ways. Then we can do both job 1 **AND** job 2 in $p \times q$ ways.*

Figure 1.3.1 summarizes the possible ways in which both job 1 and job 2 can be done. For example, if there are 5 vowels and 20 consonants and we must choose one consonant followed by one vowel for a two-letter "word", this can be done in 20×5 ways (i.e., there are 100 such words). To ride a bike, you must have the chain on both a front sprocket and a rear sprocket. For a 21-speed bike, there are 3 ways to select the front sprocket and 7 ways to select the rear sprocket (i.e., $3 \times 7 = 21$ such combinations).

This interpretation of "OR" as addition and "AND" as multiplication evident in the addition and multiplication rules above will occur throughout probability, so it is helpful to make this association in your mind. Of course, questions do not always have an AND or an OR in them and you may have to play around with re-wording the question to discover implied AND's or OR's. The next example illustrates this particular point.

Example 1.3.1. Consider an experiment in which we select two digits from the set $\{1, 2, 3, 4, 5\}$ *with replacement* (meaning that after the first digit is selected, it is "replaced" in the set of digits, so

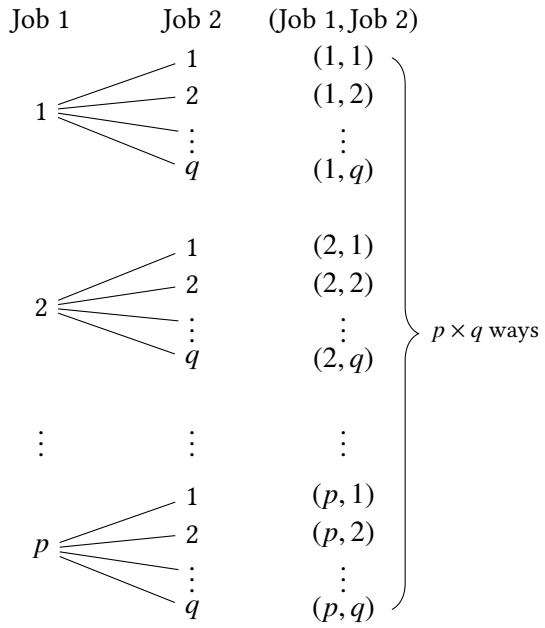


Figure 1.3.1: Visual depiction of the multiplication rule for 2 jobs

it could be selected again as the second digit). Assuming a uniform distribution on the sample space (i.e., every pair of digits has the same probability of being chosen), find the probability that one digit is even.

Solution: Note that the event of interest can be re-worded as: "The first digit is even AND the second is odd (this can be done in 2×3 ways) OR the first digit is odd AND the second is even (done in 3×2 ways)". Since these are connected with the word OR, we combine them using the addition rule to calculate that there are $(2 \times 3) + (3 \times 2) = 12$ ways for this event to occur. Since the first digit can be chosen in 5 ways AND the second digit also in 5 ways, S contains $5 \times 5 = 25$ outcomes (via the multiplication rule) and since each outcome has the same probability, they all have probability $\frac{1}{25}$. Therefore, it immediately follows that

$$P(\text{one digit is even}) = \frac{12}{25}.$$

■

Remarks:

(1) When objects are selected and replaced after each draw, the addition and multiplication rules are generally sufficient to find probabilities. When objects are drawn *without replacement* (meaning

that after an object is selected, it is "not replaced" in the set of objects, so it could not be picked again in subsequent selections), some special rules may simplify the solution, as we will see shortly.

(2) The phrases *at random* or *uniformly* again signify that all of the outcomes in the sample space are to be treated as equally likely, so that in Example 1.3.1, every possible pair of numbers chosen from this set has the same probability of $\frac{1}{25}$.

In many problems, the sample space is a set of ordered arrangements or sequences. These are classically called *permutations*. A key step in your understanding of a problem is to be clear on what it is you are counting. In this regard, it is often helpful to invent a notation for what the outcomes in the sample space look like as well as in the events you wish to consider (since these are the outcomes you would want to count).

Example 1.3.2. Suppose the letters a, b, c, d, e, and f are arranged at random to form a six-letter *word* (i.e., an ordered arrangement), and we must use each letter once only. Find the probability that the second letter in the word is e or f.

Solution: Considering all possible six-letter words that can be formulated, the sample space for this experiment would look like

$$S = \{abcdef, abcdfe, \ldots, fedcba\},$$

which clearly has a large number of points. Moreover, because we are forming the word "at random", we assign the same probability to each point in S. To determine the number of points in S, we count the number of ways we can construct such an arrangement, as each way corresponds to a unique word.

Consider filling the boxes $\boxed{}\boxed{}\boxed{}\boxed{}\boxed{}\boxed{}$ corresponding to the six positions in the arrangement. We can fill the first box in 6 ways with any one of the letters. For each of these choices, we can fill the second box in 5 ways with any one of the remaining letters. By the multiplication rule, there are $6 \times 5 = 30$ ways to fill the first two boxes. (If you are not convinced by this argument, list all the possible ways that the first two boxes can be filled.) For each of these 30 choices, we can fill the third box in 4 ways using any one of the remaining letters, so that there are $6 \times 5 \times 4 = 120$ ways to fill the first three boxes. Applying the same reasoning, we ultimately see that there are $6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$ ways to fill the 6 boxes, and hence 720 equally probable words in S.

Now, consider the event $A =$ "the second letter in the word is e or f". In this case, A is of the form

$$A = \{aebcdf, afbcde, \ldots, fedcba\}.$$

We can count the number of outcomes in A using a similar argument if we start with the second box. In particular, we can fill the second box in 2 ways (i.e., with an e or f). For each of these choices, we can then fill the first box in 5 ways, meaning that we can now fill the first two boxes in $2 \times 5 = 10$ ways. For each of these choices, we can fill the remaining four boxes in $4 \times 3 \times 2 \times 1 = 24$ ways, and so the number of outcomes in A is $10 \times 24 = 240$. Since we have a uniform probability model, it follows that

$$P(A) = \frac{240}{720} = \frac{1}{3}. \tag{1.3.1}$$

■

Remarks:

(1) In determining the number of outcomes in A in Example 1.3.2, it is important that we started with the second box. Suppose, instead, we had begun by saying there are 6 ways to fill the first box. Now, the number of ways of filling the second box depends on what happened in the first box. If we had used e or f in the first box, there is only one way to fill the second box. However, if we had used a, b, c, or d for the first box, there are 2 ways of filling the second box. We avoid this complication by starting with the second box.

(2) Since each six-letter word has the same probability of being chosen, it stands to reason that it would be equally likely that either of a, b, c, d, e, or f would be the second letter of the word. Therefore, we could have used the "simpler" sample space

$$S = \{2^{\text{nd}} \text{ letter is } a, 2^{\text{nd}} \text{ letter is } b, 2^{\text{nd}} \text{ letter is } c, 2^{\text{nd}} \text{ letter is } d, 2^{\text{nd}} \text{ letter is } e, 2^{\text{nd}} \text{ letter is } f\}.$$

Using this alternative sample space with only 6 points, we immediately obtain $P(A) = \frac{2}{6} = \frac{1}{3}$, which agrees with (1.3.1).

We can generalize the principles used in Example 1.3.2 in several ways. In each case which follows, we count the number of ordered arrangements by counting the number of ways we can fill the positions in the arrangement. Assuming that we begin with n distinct symbols, we can make:

(1) $n \times (n-1) \times \cdots \times 1$ ordered arrangements of length n using each symbol once and only once. This product is denoted by $n!$ (read "n factorial"). By mathematical convention, we define $0! = 1$. As a result, $n! = n \times (n-1)!$ for $n \geq 1$.

(2) $n \times (n-1) \times \cdots \times (n-k+1)$ ordered arrangements of length $k \leq n$ using each symbol at most once. This product is denoted by $n^{(k)}$ (read "n to k factors"). Note that $n^{(k)} = \frac{n!}{(n-k)!}$ and $n^{(n)} = n!$.

(3) $\underbrace{n \times n \times \cdots \times n}_{k \text{ terms}} = n^k$ ordered arrangements of length k using each symbol as often as we wish.

Remark: It is worth noting that the quantity $n!$ grows at an extraordinary rate as a function of n as evidenced in the following table:

n	1	2	3	4	5	6	7	8	9	10
$n!$	1	2	6	24	120	720	5040	40320	362880	3628800

Since calculating $n!$ is computationally expensive, there is an approximation to $n!$ called *Stirling's formula* which is often used for large values of n. First of all, what would it mean for two sequences of numbers which are growing very quickly to be asymptotically equal? Suppose we wish to approximate one sequence $\{a_n\}$ with another sequence $\{b_n\}$ and we want the percentage error of the approximation to approach zero as n grows larger. This is equivalent to saying $a_n/b_n \to 1$ as $n \to \infty$, and under these circumstances, we will call the two sequences *asymptotically equivalent*. Stirling's approximation says that $n!$ is *asymptotically equivalent* to $n^n e^{-n}\sqrt{2\pi n}$. The error in Stirling's approximation is less than 1% if $n \geq 8$ and becomes very small quite quickly as n increases.

Example 1.3.3. A password of length 4 is formed by randomly selecting *with replacement* 4 digits from the set $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$. Find the probability of the following events:

A = "the password has only even digits",

B = "all of the digits in the password are unique", and

C = "the password contains at least one 2".

Solution: Since the selection process is with replacement, the outcomes in the sample space can have repeated digits. In other words, the sample space is clearly represented by

$$S = \{0000, 0001, \ldots, 9999\},$$

which has 10^4 equally probable points. First of all, consider the event A which we can express as

$$A = \{0000, 0002, \ldots, 8888\}.$$

To count the number of outcomes in A, note that we can select the first digit of the password in 5 ways, as there are 5 even digits (i.e., 0, 2, 4, 6, and 8) to pick from. For each of these choices, the second digit can also be selected in 5 ways (since repeated digits are permitted), and so on. As a result, there are 5^4 outcomes in A and this leads to

$$P(A) = \frac{5^4}{10^4} = \left(\frac{1}{2}\right)^4 = \frac{1}{16}.$$

Consider next the event $B = \{0123, 0124, \ldots, 9876\}$. To count the number of outcomes in B, note that we can select the first digit in 10 ways, and for each of these choices, the second digit in 9

ways (since we do not want the first digit chosen to be repeated), and so on. Specifically, there are $10 \times 9 \times 8 \times 7 = 10^{(4)}$ outcomes in B and so

$$P(B) = \frac{10^{(4)}}{10^4} = \frac{63}{125}.$$

Finally, consider the event $C = \{0002, 0022, \ldots, 2222\}$. To count the number of outcomes in C, let us instead look at the *complement* of C, the set of all outcomes in S but not in C. If we denote this event by $\bar{C}$, then we see that $\bar{C} = \{0000, 0001, \ldots, 9999\}$. It is often easier to count outcomes in the complement rather than in the event itself. In particular, since each of the 4 digits in the password cannot be a 2, there would be $9 \times 9 \times 9 \times 9 = 9^4$ outcomes in $\bar{C}$, implying that there are $10^4 - 9^4$ outcomes in C. This results in

$$P(C) = \frac{10^4 - 9^4}{10^4} = 1 - \left(\frac{9}{10}\right)^4 = \frac{3439}{10000}.$$

■

Example 1.3.4. Suppose instead that a password of length 4 is formed by randomly selecting *without replacement* 4 digits from the set $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$. Find the probability of the following events:

A = "the password has only even digits",
B = "the password begins or ends with a 1", and
C = "the password contains a 2".

Solution: With selection process now being without replacement, the sample space is represented by

$$S = \{0123, 0132, \ldots, 9876\},$$

which has $10^{(4)}$ equally probable points. Consider the first event A which looks like

$$A = \{0246, 0248, \ldots, 8642\}.$$

The number of passwords in A is determined by taking $5^{(4)}$ arrangements of length 4 using only the even digits $\{0, 2, 4, 6, 8\}$, so that

$$P(A) = \frac{5^{(4)}}{10^{(4)}} = \frac{5 \times 4 \times 3 \times 2}{10 \times 9 \times 8 \times 7} = \frac{1}{42}.$$

Consider next the event $B = \{1023, 0231, \ldots, 9871\}$. To count the number of outcomes in B, note that there are two positions for the digit 1 to appear (i.e., first or last). For each of these choices, we can fill the remaining three positions in $9^{(3)}$ ways. Applying the multiplication rule, it follows that

$$P(B) = \frac{2 \times 9^{(3)}}{10^{(4)}} = \frac{1}{5}.$$

Finally, consider the event $C = \{1234, 1324, \ldots, 9872\}$. To count the number of outcomes in C, we can once again use the complement and count the number of passwords that do not contain a 2. Removing the digit 2 from consideration, there are $9^{(4)}$ passwords that do not contain a 2, implying that there are $10^{(4)} - 9^{(4)}$ passwords that do contain a 2. Therefore, we obtain

$$P(C) = \frac{10^{(4)} - 9^{(4)}}{10^{(4)}} = 1 - \frac{9^{(4)}}{10^{(4)}} = 1 - \frac{9 \times 8 \times 7 \times 6}{10 \times 9 \times 8 \times 7} = \frac{2}{5}.$$

■

In some problems, the outcomes in the sample space are subsets of a fixed size, and we are interested in counting such subsets, or as they classically called, *combinations*. For example, consider an experiment in which we randomly select a subset of 3 *distinct* digits from the set $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ so that the sample space is represented by

$$S = \{\{0, 1, 2\}, \{0, 1, 3\}, \ldots, \{7, 8, 9\}\}.$$

In contrast to permutations, the order of the elements in a subset is not relevant (i.e., the subsets $\{0, 1, 2\}$ and $\{1, 0, 2\}$ are considered to be the same and must be counted only once). Thus, if we attempted to count the number of outcomes in S using $10^{(3)}$ arrangements, we would be over-counting by a factor $3! = 6$. That is because we can form $3!$ arrangements of length 3 of any subset, but each subset must be counted only once. (For instance, the subset $\{0, 1, 2\}$ generates the $3! = 6$ arrangements $012, 021, 102, 120, 201$, and 210 but only one of them, 012, gets counted). The number of outcomes in S is therefore given by

$$\frac{10^{(3)}}{3!} = 120.$$

Since we selected the subset at random, each of the 120 subsets has the same probability $\frac{1}{120}$ of being chosen.

By an argument similar to that above, the number of subsets of size k that can be formed from n distinct objects where $k \leq n$ is

$$\frac{n^{(k)}}{k!}.$$

We will use the combinatorial symbol $\binom{n}{k}$ (read "n choose k") to denote the number of subsets of size k that can be formed from a set of n distinct objects. In other words, we define

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

Note that $\binom{n}{k} = \binom{n}{n-k}$ for $k = 0, 1, \ldots, n$ with $\binom{n}{0} = \binom{n}{n} = 1$.

Example 1.3.5. Consider a group of six third-year undergraduate students and seven fourth-year undergraduate students. Suppose that among the group of third-year students is one student whose first name is Roger. A committee of size five is randomly formed. Find the probability of the following events:

A = "the third-year student named Roger is included on the committee",

B = "the committee is comprised only of fourth-year students", and

C = "at most four of the committee members are from third year".

Solution: Let us represent the 6 third-year students by $T_1, T_2, T_3, T_4, T_5, T_6$ (with T_1 denoting Roger) and the 7 fourth-year students by $F_1, F_2, F_3, F_4, F_5, F_6, F_7$. We consider a uniform probability model in which all committees (i.e., subsets) of size 5 are equally probable. As the order of students within a committee does not matter, this leads to the following form of the sample space:

$$S = \{\{T_1, T_2, T_3, T_4, T_5\}, \{T_1, T_2, T_3, T_4, T_6\}, \ldots, \{F_3, F_4, F_5, F_6, F_7\}\}.$$

In this case, the number of outcomes in S is given by $\binom{13}{5} = 1287$.

To count the number of outcomes in the event A, it is clear that we must have T_1 in the subset, thereby leaving the other four spots on the committee to be populated from the remaining $13-1 = 12$ students in $\binom{12}{4}$ ways. As a result, we obtain

$$P(A) = \frac{\binom{12}{4}}{\binom{13}{5}} = \frac{12!}{4!8!} \cdot \frac{5!8!}{13!} = \frac{5}{13}.$$

Consider next the event $B = \{\{F_1, F_2, F_3, F_4, F_5\}, \{F_1, F_2, F_3, F_4, F_6\}, \ldots, \{F_3, F_4, F_5, F_6, F_7\}\}$. We can form the outcomes in B by choosing 5 students from the 7 available fourth-year students in $\binom{7}{5}$ ways, which leads to

$$P(B) = \frac{\binom{7}{5}}{\binom{13}{5}} = \frac{21}{1287} = \frac{7}{429}.$$

Finally, consider the event C which has the form

$$C = \{\{F_1, F_2, F_3, F_4, F_5\}, \{F_1, F_2, F_3, F_4, F_6\}, \ldots, \{T_3, T_4, T_5, T_6, F_7\}\}.$$

Once more, it is convenient to consider the complement of C in which all 5 committee members are from third year. This leads to the following 6 outcomes which comprise $\bar{C}$:

$$\{T_1, T_2, T_3, T_4, T_5\}, \{T_1, T_2, T_3, T_4, T_6\}, \{T_1, T_2, T_3, T_5, T_6\},$$
$$\{T_1, T_2, T_4, T_5, T_6\}, \{T_1, T_3, T_4, T_5, T_6\}, \{T_2, T_3, T_4, T_5, T_6\}.$$

Note that the above subsets are formed by selecting a subset of size 5 from the 6 third-year students to choose from. This can be done in $\binom{6}{5} = 6$ ways. Therefore, the number of outcomes in C is $\binom{13}{5} - \binom{6}{5}$ and its probability is

$$P(C) = \frac{\binom{13}{5} - \binom{6}{5}}{\binom{13}{5}} = 1 - \frac{\binom{6}{5}}{\binom{13}{5}} = \frac{1281}{1287} = \frac{427}{429}.$$

■

Example 1.3.6. Suppose a box contains ten balls of which three are red, four are white, and three are green. A sample of four balls is selected at random without replacement. Find the probability of the following events:

A = "the sample contains 2 red balls",

B = "the sample contains 2 red balls, 1 white ball, and 1 green ball", and

C = "the sample contains 2 or more red balls".

Solution: Imagine that we label the balls from 1 to 10 with labels $1, 2, 3$ representing red, labels $4, 5, 6, 7$ representing white, and labels $8, 9, 10$ representing green. We consider a uniform probability model in which all subsets of size 4 are equally probable. In this case, the sample space would be of the form

$$S = \{\{1, 2, 3, 4\}, \{1, 2, 3, 5\}, \ldots, \{7, 8, 9, 10\}\},$$

which consists of $\binom{10}{4} = 210$ outcomes. To determine the number of outcomes in A, we wish to count those subsets in S which have exactly two red balls. To do so, we first choose 2 red balls from the 3 available red balls in $\binom{3}{2}$ ways. For each of these choices, we select the other 2 balls from the 7 non-red balls in $\binom{7}{2}$ ways, so that there are $\binom{3}{2} \times \binom{7}{2}$ outcomes in A. Therefore, we get

$$P(A) = \frac{\binom{3}{2}\binom{7}{2}}{\binom{10}{4}} = \frac{63}{210} = \frac{3}{10}.$$

To count the number of outcomes in B, note that we can select the two red balls in $\binom{3}{2}$ ways, followed by the one white ball in $\binom{4}{1}$ ways, and finally the single green ball in $\binom{3}{1}$ ways. Applying the multiplication rule, we obtain

$$P(B) = \frac{\binom{3}{2}\binom{4}{1}\binom{3}{1}}{\binom{10}{4}} = \frac{36}{210} = \frac{6}{35}.$$

Finally, consider the event C which looks like

$$C = \{\{1, 2, 3, 4\}, \{1, 2, 3, 5\}, \ldots, \{2, 3, 9, 10\}\}.$$

Note that C has outcomes with both two and three red balls. We need to count these separately. In particular, there are $\binom{3}{2} \times \binom{7}{2}$ outcomes with exactly two red balls and $\binom{3}{3} \times \binom{7}{1}$ outcomes with exactly

three red balls. Applying the addition rule, we obtain

$$P(C) = \frac{\binom{3}{2}\binom{7}{2} + \binom{3}{3}\binom{7}{1}}{\binom{10}{4}} = \frac{63 + 7}{210} = \frac{1}{3}.$$

■

Remark: A common mistake counting the outcomes of C in Example 1.3.6 is to use the following argument: There are $\binom{3}{2}$ ways to select two red balls and then for each of these choices, we can select the next two balls from the remaining eight balls in $\binom{8}{2}$ ways, resulting in the number of outcomes in C being $\binom{3}{2} \times \binom{8}{2}$. You can easily verify that this number is greater than $\binom{3}{2}\binom{7}{2} + \binom{3}{3}\binom{7}{1}$. The reason for the error is that some of the outcomes in C have been counted more than once. For example, you might pick red balls $\{1, 2\}$ and then other balls $\{3, 4\}$ to get the subset $\{1, 2, 3, 4\}$. On the other hand, you may pick red balls $\{1, 3\}$ and then other balls $\{2, 4\}$ to get the subset $\{1, 3, 2, 4\}$. These are counted as two separate outcomes but they are in fact the same subset. To avoid this kind of counting error, whenever you are asked about events defined in terms such as "at most…", "more than…", " fewer than…", etc., break the events into pieces where each piece has outcomes with specific values (e.g., exactly two red balls and exactly three red balls).

Section 1.3 Problems

1.3.1 (a) A course has four sections with no limit on how many can enrol in each section. Three students each pick a section at random.

(i) Find the probability that all three students end up in the same section.

(ii) Find the probability that all three students end up in different sections.

(iii) Find the probability that no one picks section 1.

(b) Repeat part (a) in the case when there are n sections and s students where $n \geq s$.

1.3.2 Canadian postal codes consist of 3 letters (chosen from the 26 possible letters available) alternated with 3 digits (chosen from the 10 possible digits available), starting with a letter in the first position (e.g., N2L 3G1). Assume no other restrictions on the construction of postal codes. For a postal code chosen at random, what is the probability that

(a) all 3 letters are the same?

(b) the digits are all even or all odd? Treat 0 as being an even digit.

1.3.3 A binary sequence is an ordered arrangement of zeros and ones. Suppose we have a uniform probability model on the sample space of all binary sequences of length 10. What is the probability that the sequence has exactly 5 zeros?

1.3.4 Suppose that a 4-digit number is formed by randomly selecting four digits from the set $\{1, 2, 3, 4, 5, 6, 7\}$ without replacement. Find the probability the number formed is

(a) even.

(b) over 3000.

(c) an even number over 3000.

1.3.5 A factory parking lot has 160 cars in it, of which 35 have faulty emission controls. An air quality inspector does spot checks on 8 random cars on the lot. Give an expression for the probability that at least 3 of these 8 cars will have faulty emission controls.

1.3.6 Suppose r individuals get on an elevator at the basement floor of a building. There are n floors above (numbered $1, 2, \ldots, n$) where individuals may get off. Assuming that each individual is equally likely to get off at any floor, find the probability that

(a) no individual gets off at floor 1.

(b) individuals all get off at different floors (assuming that $n \geq r$).

1.3.7 There are 5 stops left on a subway line and 4 passengers on a train. Assume they are each equally likely to get off at any stop. What is the probability that two passengers get off at the same stop and the other two passengers get off at another same stop?

1.3.8 Three digits are randomly selected from the set $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$. Find the probability that the digits are drawn in *increasing order* (i.e., the first selected digit is less than the second selected digit which is less than the third selected digit) if

(a) draws are made without replacement.

(b) draws are made with replacement.

1.3.9 The 10,000 tickets for a raffle are numbered 0000 to 9999. A 4-digit winning number is drawn and a prize is paid on each ticket whose 4-digit number is *any arrangement* of the number drawn. For instance, if winning number 0011 is drawn, prizes are paid on tickets numbered 0011, 0101, 0110, 1001, 1010, and 1100. An operator charges \$1 to buy a ticket and each prize is \$500.

(a) What is the probability of winning a prize with ticket number 7337? What is the probability of winning a prize with ticket number 7235? What advice would you give to someone buying a ticket for this raffle?

(b) Assuming that all tickets are sold, what is the probability that the raffle operator will lose money?

1.3.10 **Lotto 6/49.** In Lotto 6/49, you purchase a lottery ticket with 6 different numbers, selected from the set of digits $\{1, 2, \ldots, 49\}$. In the draw, six (different) numbers are randomly selected. Find the probability that your ticket matches exactly x of the 6 numbers drawn, $x = 0, 1, \ldots, 6$.

1.3.11 **The Birthday Problem.** Suppose there are r persons in a room. Ignoring February 29 and assuming that every person is equally likely to have been born on any of the 365 other days in a year, find the probability that no two persons in the room have the same birthday. Find the numerical value of this probability for $r = 20, 40,$ and 60.

Chapter 2

Probability Rules and Conditional Probability

2.1 Use of Sets

Recall that a probability model consists of a sample space S, a set of events of the sample space to which we can assign probabilities, and a mechanism for assigning these probabilities. From Definition 1.2.4, the probability of an arbitrary event A can be determined by summing the probabilities for all the simple events that make up A. This leads to the following three rules which are immediately established:

Rule 1 (Normalization): $P(S) = 1$.

Proof: Note that $P(S) = \sum_{a \in S} P(a) = \sum_{\text{all } a} P(a) = 1$ from Definition 1.2.3. ■

Rule 2 (Boundedness): For any event A, $0 \leq P(A) \leq 1$.

Proof: Note that $P(A) = \sum_{a \in A} P(a) \leq \sum_{a \in S} P(a) = P(S) = 1$ from Rule 1. Since each $P(a) \geq 0$ from Definition 1.2.3, we clearly have $0 \leq P(A) \leq 1$. ■

Rule 3 (Monotonicity): If A and B are two events with $A \subseteq B$, then $P(A) \leq P(B)$.

Proof: If $A \subseteq B$, then all of the simple events in A are also in B. Therefore,

$$P(A) = \sum_{a \in A} P(a) \leq \sum_{a \in B} P(a) = P(B).$$

■

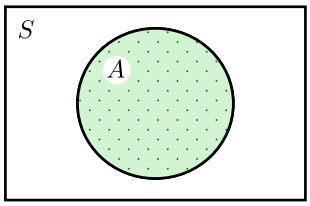


Figure 2.1.1: Event A in the sample space S

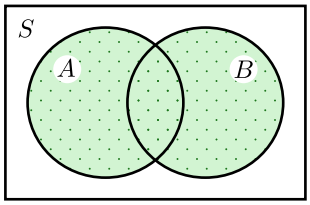


Figure 2.1.2: The union of two events A and B

Probability makes extensive use of set operations, so let us introduce at the outset the relevant notation and terminology which will facilitate a set-theoretic description of a probability model. First of all, we begin by asking a fundamental question: *What do sets have to do with the occurrence of events?* Suppose a random experiment having sample space S is conducted. When would we say an event $A \subseteq S$ occurs? For example, if a six-sided die is thrown so that $S = \{1, 2, 3, 4, 5, 6\}$, when would we say an event such as $A = \{2, 4, 6\}$ occurs? In this case, the occurrence of the event A means that the number of dots showing on the upturned face is even. In general, we would say that the occurrence of an event A means that *one of the simple events in A occurred.*

We often illustrate the relationship among sets using *Venn diagrams.* In the Venn diagrams we present, imagine S as consisting of all of the points in a rectangle of area one[1]. To illustrate the event A, we can draw a region within the rectangle with area roughly proportional to the probability of the event A. In fact, we might think of the random experiment as throwing a dart at the rectangle in Figure 2.1.1, and say the event A occurs if the dart lands within the shaded region of A.

Consider two events A and B. What if we combine these events by including all of the outcomes in either A or B or both? This gives rise to the *union* of the two events, denoted by $A \cup B$, as illustrated by the shaded region in the Venn diagram of Figure 2.1.2. The union of the two events occurs if one of the simple events in either A or B or both occurs. We refer to this as the event "A or B" with the

[1] As you may know, however, the number of points in a rectangle is not countable, so technically speaking this is not a discrete sample space. Nevertheless, this representation of S is commonly used to illustrate various combinations of sets.

understanding that in this course we will use the word "or" inclusively to also permit both. Another way of expressing a union is to say $A \cup B$ occurs means "at least one of A or B occurs". In a similar fashion, if we have three events A, B, and C, the event $A \cup B \cup C$ occurs would mean "at least one of A, B, or C occurs". In general, for a sequence of events $\{A_n\}_{n=1}^{\infty}$, we have that

$$\bigcup_{n=1}^{\infty} A_n = A_1 \cup A_2 \cup \cdots = \{a \in S : a \in A_n \text{ for some } n\}.$$

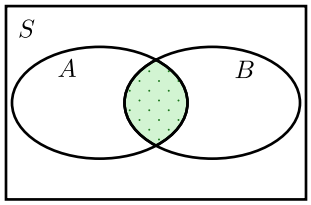


Figure 2.1.3: The intersection of two events A and B

Next, consider the *intersection* of two events A and B, denoted by $A \cap B$, representing the set of all simple events in S that are in both A and B. This is illustrated by the shaded region in the Venn diagram of Figure 2.1.3. The event $A \cap B$ occurs if and only if a simple event in the intersection of A and B occurs, meaning that both A and B occur. It is a common convention to shorten the notation for the intersection of events, so that AB means $A \cap B$ and ABC means $A \cap B \cap C$. We will use both of these notation choices interchangeably throughout the remainder of this chapter. Moreover, for a sequence of events $\{A_n\}_{n=1}^{\infty}$, we have that

$$\bigcap_{n=1}^{\infty} A_n = A_1 \cap A_2 \cap \cdots = \{a \in S : a \in A_n \text{ for all } n\}.$$

As we first introduced in Example 1.3.3, the complement of an event A, denoted by $\bar{A}$, is the set of all simple events which are in S but not in A. This is illustrated by the shaded region in the Venn diagram of Figure 2.1.4. Finally, recall the empty event $\emptyset$. This is an event made up of no simple events, and so $P(\emptyset) = 0$ from Definition 1.2.4 (since an empty sum yields a value of 0). Moreover, we clearly have that $\emptyset = \bar{S}$.

Since probability theory is built from the application of such set-theoretic operations, it is often helpful to employ Venn diagrams in visualizing relationships among events. To this point, there are two particularly useful rules governing taking the complements of unions and intersections that can easily be verified with the aid of Venn diagrams. We state these rules for two events only, but they naturally extend beyond the two event case.

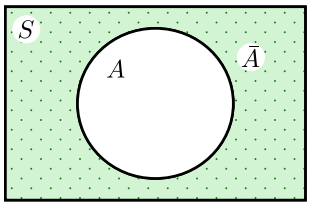


Figure 2.1.4: The complement of the event A

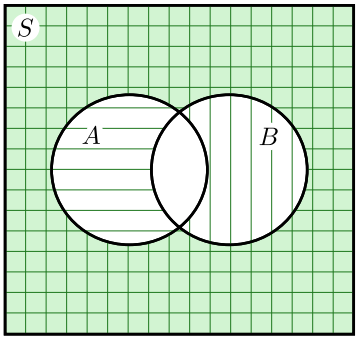


Figure 2.1.5: Illustration of De Morgan's Law using a Venn diagram

De Morgan's Laws:

(a) $\overline{A \cup B} = \bar{A} \cap \bar{B}$.

(b) $\overline{A \cap B} = \bar{A} \cup \bar{B}$.

Proof: To prove (a), we use a Venn diagram (see Figure 2.1.5) in which $\bar{A}$ is indicated with vertical lines and $\bar{B}$ with horizontal lines. Therefore, $\bar{A} \cap \bar{B}$ is the region with both vertical and horizontal lines. Note that this agrees with the shaded region representing $\overline{A \cup B}$. Hence, both regions are identical and (a) is proven. As an alternative verification, we prove (b) using the definitions of sets. Specifically, when a point a is in the set $\overline{A \cap B}$, this means that $a \in S$ but a is not in $A \cap B$. This, in turn, implies that either a is not in A or a is not in B. In other words, $a \in \bar{A}$ or $a \in \bar{B}$, or equivalently $a \in \bar{A} \cup \bar{B}$. This proves that $\overline{A \cap B} \subseteq \bar{A} \cup \bar{B}$. Since the converse inclusion is established by simply reversing the above argument, (b) subsequently follows. ■

Section 2.1 Problems

2.1.1 Consider tossing a coin four times. Let A be the event of getting exactly 2 heads. Let B be the event of getting tails on either the 3$^{\text{rd}}$ or 4$^{\text{th}}$ toss. Determine the events on both sides of the two

De Morgan's laws and verify that the equalities hold true.

2.1.2 Let A and B be two events.

(a) Show that

$$\bar{A} = (\bar{A} \cap B) \cup (\bar{A} \cap \bar{B}).$$

(b) Show that

$$\overline{A \cap B} = (\bar{A} \cap B) \cup (\bar{A} \cap \bar{B}) \cup (A \cap \bar{B}).$$

(c) Consider rolling a six-sided die twice. Let A be event where at least one of the upturned faces is an odd number. Let B be the event where the total of the two rolls is greater than 8. Determine the events on both sides of the equality in part (b) and verify that the equality holds true.

2.1.3 (a) Prove the following identities:

(i) $A \cup (\bigcap_{i=1}^{\infty} B_i) = \bigcap_{i=1}^{\infty} (A \cup B_i)$.

(ii) $A \cap (\bigcup_{i=1}^{\infty} B_i) = \bigcup_{i=1}^{\infty} (A \cap B_i)$.

(b) Use the results of parts (i) and (ii) to conclude that $A \cup (\bigcap_{i=1}^{n} B_i) = \bigcap_{i=1}^{n} (A \cup B_i)$ and $A \cap (\bigcup_{i=1}^{n} B_i) = \bigcup_{i=1}^{n} (A \cap B_i)$ for $n \in \{1, 2, 3, \ldots\}$.

2.2 Addition Rules for Unions of Events

In addition to the three basic rules specified in the previous section, we can establish several rules governing the calculation of probabilities for unions of events. We begin with a general rule involving two events.

Rule 4a (Probability of the Union of Two Events):

$$P(A \cup B) = P(A) + P(B) - P(AB). \tag{2.2.1}$$

Proof: First of all, note that $A\bar{B}$ consists of simple events which are in A but not in B. It immediately follows that $A = (A\bar{B}) \cup (AB)$. Moreover, if $a \in A$, then either $a \in A\bar{B}$ or $a \in AB$ but a cannot possibly belong to both $A\bar{B}$ and AB. A similar description would apply to the representation of B as

$B = (\bar{A}B) \cup (AB)$. From Definition 1.2.4, we therefore obtain

$$\begin{aligned} P(A) + P(B) &= \sum_{a \in A} P(a) + \sum_{a \in B} P(a) \\ &= \left(\sum_{a \in A\bar{B}} P(a) + \sum_{a \in AB} P(a) \right) + \left(\sum_{a \in \bar{A}B} P(a) + \sum_{a \in AB} P(a) \right) \\ &= \left(\sum_{a \in A\bar{B}} P(a) + \sum_{a \in AB} P(a) + \sum_{a \in \bar{A}B} P(a) \right) + \sum_{a \in AB} P(a) \\ &= \sum_{a \in A \cup B} P(a) + \sum_{a \in AB} P(a) \\ &= P(A \cup B) + P(AB), \end{aligned} \tag{2.2.2}$$

where we used the fact (easily verified with a Venn diagram) that $A \cup B = (A\bar{B}) \cup (AB) \cup (\bar{A}B)$. Rearranging (2.2.2) immediately yields (2.2.1). ■

Remark: Rule 4a can be justified solely with the use a Venn diagram. Each outcome in $A \cup B$ must be counted once. In the expression $P(A) + P(B)$, however, outcomes in AB have their probability counted twice – once in $P(A)$ and once in $P(B)$. As a result, $P(AB)$ needs to be subtracted from this sum to compensate for this double counting.

Armed with the result in the case of two events, we now examine the situation involving the probability of the union of three events.

Rule 4b (Probability of the Union of Three Events):

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC). \tag{2.2.3}$$

Proof: Consider the Venn diagram depicted in Figure 2.2.1. In the sum $P(A) + P(B) + P(C)$, those outcomes in the regions labelled D, H, and J in Figure 2.2.1 lie in only one of the events and their probabilities are added only once. However, outcomes in the regions labelled G, E, and I lie in two of the events. We can compensate for this double counting by subtracting these probabilities and using the revised formula

$$P(A) + P(B) + P(C) - (P(AB) + P(AC) + P(BC)).$$

However, now those outcomes in all three events (i.e., those outcomes in $F = ABC$) have their probabilities added in three times and then subtracted three times, and so they are not included at all. We must therefore account for its non-inclusion by inserting $P(ABC)$ into the above formula, thereby giving rise to (2.2.3). ■

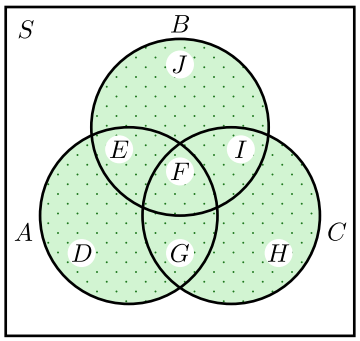


Figure 2.2.1: The union of three events A, B, and C

Remarks:

(1) An alternative justification of Rule 4b could be obtained by using Rule 4a in conjunction with the *distributive property* of events (see Problem 2.1.3). Specifically, note that

$$\begin{aligned} P(A \cup B \cup C) &= P((A \cup B) \cup C) \\ &= P(A \cup B) + P(C) - P((A \cup B) \cap C) \\ &= P(A) + P(B) - P(AB) + P(C) - P((AC) \cup (BC)) \\ &= P(A) + P(B) + P(C) - P(AB) - (P(AC) + P(BC) - P((AC) \cap (BC))) \\ &= P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC), \end{aligned}$$

where we used the fact that $(AC) \cap (BC) = ABC$.

(2) There is a natural generalization of Rules 4a and 4b to the case of n events $A_1, A_2, \ldots, A_n$. This rule is often referred to as the *inclusion-exclusion principle* because of the process outlined in the proof of Rule 4b for constructing its formula. Using mathematical induction, it can be shown that

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{\text{all } i<j} P(A_i A_j) + \sum_{\text{all } i<j<k} P(A_i A_j A_k) - \cdots + (-1)^{n-1} P\left(\bigcap_{i=1}^{n} A_i\right), \quad (2.2.4)$$

where the distinct summation indices are chosen from the set $\{1, 2, \ldots, n\}$ for $n \geq 2$.

Example 2.2.1. In a standard deck of 52 cards (as described in Example 1.2.1), two of the suits (i.e., diamonds and hearts) are coloured red and the other two suits (i.e., spades and clubs) are coloured black. In addition, the cards J, Q, and K are generally referred to as "face" cards. If one card is randomly drawn from this deck, what is the probability that it is a red card or a face card?

Solution: Let A be the event of drawing a red card and B the event of drawing a face card. Using the sample space of 52 cards

$$S = \{2\blacklozenge, 3\blacklozenge, \ldots, A\blacklozenge, 2\heartsuit, 3\heartsuit, \ldots, A\heartsuit, 2\spadesuit, 3\spadesuit, \ldots, A\spadesuit, 2\clubsuit, 3\clubsuit, \ldots, A\clubsuit\},$$

note that

$$P(A) = P(\{2\blacklozenge, 3\blacklozenge, \ldots, A\blacklozenge, 2\heartsuit, 3\heartsuit, \ldots, A\heartsuit\}) = \frac{26}{52} = \frac{1}{2},$$

$$P(B) = P(\{J\blacklozenge, Q\blacklozenge, K\blacklozenge, J\heartsuit, Q\heartsuit, K\heartsuit, J\spadesuit, Q\spadesuit, K\spadesuit, J\clubsuit, Q\clubsuit, K\clubsuit\}) = \frac{12}{52} = \frac{3}{13},$$

and

$$P(AB) = P(\{J\blacklozenge, Q\blacklozenge, K\blacklozenge, J\heartsuit, Q\heartsuit, K\heartsuit\}) = \frac{6}{52} = \frac{3}{26}.$$

Using Rule 4a, we obtain

$$P(\text{red card or face card}) = P(A \cup B) = P(A) + P(B) - P(AB) = \frac{1}{2} + \frac{3}{13} - \frac{3}{26} = \frac{8}{13}.$$

■

In set theory, two sets are said to be disjoint if they share no common points. This concept extends to events as the following definition specifies.

Definition 2.2.1. *Events A and B are **mutually exclusive** if $AB = \emptyset$.*

Since mutually exclusive events A and B have no outcomes in common, $P(AB) = P(\emptyset) = 0$. More generally, events $A_1, A_2, \ldots, A_n$ are mutually exclusive if $A_iA_j = \emptyset$ for all $i, j \in \{1, 2, \ldots, n\}$ such that $i < j$. This means that there is no chance of two or more of these events occurring together, implying that we either have exactly one of the events occur or none. For example, if a six-sided die is rolled twice, the events A = "the number 2 occurs on the 1st roll" and B = "the total of the two rolls equals 10" are mutually exclusive events. Similarly, the events $A_2, A_3, \ldots, A_{12}$, where A_j is the event that the total of the two rolls equals j, are mutually exclusive. In the case of mutually exclusive events, (2.2.4) simplifies to give the following rule.

Rule 5 (Probability of the Union of n Mutually Exclusive Events): If $A_1, A_2, \ldots, A_n$ are mutually exclusive events, then

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i). \tag{2.2.5}$$

Proof: Since $A_1, A_2, \ldots, A_n$ are mutually exclusive, all intersections of two or more of these events will yield the empty event, which has a probability of 0. Therefore, all intersection probabilities in (2.2.4) disappear, simply resulting in the formula given by (2.2.5). ■

Remark: Rule 5 applies more generally to a countably infinite collection of mutually exclusive events. In particular, if $\{A_i\}_{i=1}^{\infty}$ is a sequence of mutually exclusive events (i.e., $A_iA_j = \emptyset$ when $i \neq j$), then we have from Definition 1.2.4 that

$$\begin{aligned} P\left(\bigcup_{i=1}^{\infty} A_i\right) &= \sum_{a \in \bigcup_{i=1}^{\infty} A_i} P(a) \\ &= \sum_{a \in A_1} P(a) + \sum_{a \in A_2} P(a) + \sum_{a \in A_3} P(a) + \cdots \\ &= P(A_1) + P(A_2) + P(A_3) + \cdots \\ &= \sum_{i=1}^{\infty} P(A_i). \end{aligned} \tag{2.2.6}$$

For this reason, (2.2.6) is often referred to as the rule of *countable additivity*.

As we witnessed in several examples in Section 1.3, it can be easier to work with the complement of an event rather than the actual event itself. This leads to the following familiar rule, which is based off of the idea that two parts make a whole.

Rule 6 (Probability of the Complement of an Event):

$$P(A) = 1 - P(\bar{A}). \tag{2.2.7}$$

Proof: Note that A and $\bar{A}$ are mutually exclusive events such that $A \cup \bar{A} = S$. By Rules 1 and 5, we get

$$1 = P(S) = P(A \cup \bar{A}) = P(A) + P(\bar{A}).$$

Rearranging the above equation, we immediately obtain (2.2.7). ■

Remark: A similar approach used in the above proof can be applied to find a formula for $P(A\bar{B})$, sometimes referred to as the probability of the *event difference* between A and B. In particular, we observed in the proof of Rule 4a that A can be represented as $A = (A\bar{B}) \cup (AB)$. However, $A\bar{B}$ and AB are mutually exclusive events, and so Rule 5 gives us $P(A) = P(A\bar{B}) + P(AB)$, or equivalently, $P(A\bar{B}) = P(A) - P(AB)$. In the case of three events A, B, and C, this approach would yield the formula $P(AB\bar{C}) = P(AB) - P(ABC)$.

Example 2.2.2. An elementary school is offering three language classes: one in Russian, one in French, and one in German. These classes are open to any of the 100 students in the school. There are 26 students in the Russian class, 29 in the French class, and 17 in the German class. There are 12 students that are in both Russian and French, 6 that are in both Russian and German, and 4 that are

in both French and German. In addition, there are 2 students taking all three classes. What is the probability that a randomly chosen student is taking no language classes?

Solution: Consider the events R, F, and G, denoting whether a randomly chosen student is enrolled in the Russian, French, and German class, respectively. In terms of these events, we are given the following information:

$$\begin{array}{ccc} P(R) = 0.26, & P(F) = 0.29, & P(G) = 0.17, \\ P(RF) = 0.12, & P(RG) = 0.06, & P(FG) = 0.04, \\ & P(RFG) = 0.02. & \end{array}$$

Using this information, we can calculate (via Rule 4b)

$$\begin{aligned} P(R \cup F \cup G) &= P(R) + P(F) + P(G) - P(RF) - P(RG) - P(FG) + P(RFG) \\ &= 0.26 + 0.29 + 0.17 - 0.12 - 0.06 - 0.04 + 0.02 \\ &= 0.52. \end{aligned}$$

Therefore,

$$\begin{aligned} P(\text{student takes no language classes}) &= P(\bar{R} \cap \bar{F} \cap \bar{G}) \\ &= P(\overline{R \cup F \cup G}) \text{ by De Morgan's Law (for 3 events)} \\ &= 1 - P(R \cup F \cup G) \text{ by Rule 6} \\ &= 1 - 0.52 \\ &= 0.48. \end{aligned}$$

■

Example 2.2.3. Three fair six-sided dice are rolled. Calculate the probability that at least one of the three dice turns up a 6.

Solution 1: Suppose that the sample space for this experiment is represented by

$$S = \{(1,1,1), (1,1,2), \ldots, (6,6,6)\},$$

which has $6 \times 6 \times 6 = 216$ simple events. Let us define the following events:

$$\begin{aligned} A_1 &= \text{“6 occurs on the } 1^{\text{st}} \text{ die”}, \\ A_2 &= \text{“6 occurs on the } 2^{\text{nd}} \text{ die”, and} \\ A_3 &= \text{“6 occurs on the } 3^{\text{rd}} \text{ die”}. \end{aligned}$$

Note that

$$A_1A_2 = \{(6,6,1),(6,6,2),(6,6,3),(6,6,4),(6,6,5),(6,6,6)\},$$

$$A_1A_3 = \{(6,1,6),(6,2,6),(6,3,6),(6,4,6),(6,5,6),(6,6,6)\},$$

$$A_2A_3 = \{(1,6,6),(2,6,6),(3,6,6),(4,6,6),(5,6,6),(6,6,6)\},$$

and

$$A_1A_2A_3 = \{(6,6,6)\}.$$

Each of A_1A_2, A_1A_3, and A_2A_3 has 6 outcomes and $A_1A_2A_3$ has only 1 outcome. Applying Rule 4b, we obtain

$$\begin{aligned} P(\text{at least one of the 3 dice turns up a 6}) &= P(A_1 \cup A_2 \cup A_3) \\ &= P(A_1) + P(A_2) + P(A_3) - P(A_1A_2) - P(A_1A_3) - P(A_2A_3) + P(A_1A_2A_3) \\ &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} - \frac{6}{216} - \frac{6}{216} - \frac{6}{216} + \frac{1}{216} \\ &= \frac{91}{216}. \end{aligned}$$

■

Solution 2: This is a situation where it is perhaps easier to consider the complement of the event $A_1 \cup A_2 \cup A_3$, since the complement is the event that there is no 6 turning up on any of the three dice. As such, note that

$$\overline{A_1 \cup A_2 \cup A_3} = \{(1,1,1),(1,1,2),\ldots,(5,5,5)\},$$

consisting of $5 \times 5 \times 5 = 125$ outcomes. Applying Rule 6, we simply obtain

$$P(A_1 \cup A_2 \cup A_3) = 1 - P(\overline{A_1 \cup A_2 \cup A_3}) = 1 - \frac{125}{216} = \frac{91}{216}.$$

■

The rules introduced in this section link the concepts of addition of probabilities with unions of events. Over the next two sections, we will see how intersections of events are closely related to the multiplication of probabilities. Establishing these connections will make problem solving easier and the construction of probability models more amenable.

Section 2.2 Problems

2.2.1 According to a survey of people on the last Ontario voters list, 55% are female, 55% are politically to the right, and 15% are male and politically to the left. What percentage of voters are female and politically to the right? Assume voter attitudes are classified simply as being either left or right.

2.2.2 If A and B are mutually exclusive events with $P(A) = 0.25$ and $P(B) = 0.4$, find the probability of each of the following events:

(a) $\bar{A}$.

(b) $\bar{B}$.

(c) $A \cup B$.

(d) $A \cap B$.

(e) $\bar{A} \cup \bar{B}$.

(f) $\bar{A} \cap \bar{B}$.

(g) $\overline{A \cap B}$.

2.2.3 Prove that $P(A \cup B) = 1 - P(\bar{A}\,\bar{B})$ for arbitrary events A and B in S.

2.2.4 One of the most important consequences of the countable additivity condition given by (2.2.6) is that it implies a certain "continuity property" of the probability function $P(\cdot)$. In particular, show that for any sequence of events $\{A_i\}_{i=1}^{\infty}$ that are *nested* so that $A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$,

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{n\to\infty} P(A_n).$$

Can you explain in what way this is similar to the standard definition of continuity of a real-valued function?

2.2.5 Let $\{A_i\}_{i=1}^{\infty}$ be a sequence of events of a sample space S.

(a) Prove *Boole's inequality*:

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

(b) Use the result of part (a) to show that

$$P\left(\bigcap_{i=1}^{\infty} A_i\right) \geq 1 - \sum_{i=1}^{\infty} P(\bar{A}_i).$$

2.2.6 Let A, B, and C be events for which $P(A) = 0.2, P(B) = 0.5, P(C) = 0.3$, and $P(AB) = 0.1$.

(a) Find the largest possible value for $P(A \cup B \cup C)$.

(b) For this largest value to occur, are the events A and C mutually exclusive, not mutually exclusive, or can this not be determined?

2.2.7 For students finishing second year Math at UW, suppose that 22% have a math average greater than 80%, 24% have a STAT 230 mark greater than 80%, 20% have an overall average greater than 80%, 14% have both a math average and STAT 230 mark greater than 80%, 13% have both an overall average and STAT 230 mark greater than 80%, 10% have all 3 of these components greater than 80%, and 67% have none of these 3 components greater than 80%. Find the probability a randomly chosen student finishing second year Math at UW has math and overall averages both greater than 80% and a STAT 230 mark less than or equal to 80%.

2.3 Dependent and Independent Events

One of the main objectives of a probability model is to determine how likely it is that an event A will occur when a particular random experiment is performed. In numerous situations, however, the probability assigned to A will be affected by knowledge of the occurrence or non-occurrence of another event B. For example, consider the events A = "airplane engine fails in flight" and B = "airplane reaches its destination safely". Do we normally consider these events as related or dependent in some way? Certainly, if a blown fuse or a tripped breaker in the engine should occur, this would affect the probability that the airplane safely reaches its destination. In other words, it would affect the probability that should be assigned to the event B. As a second example, suppose we toss a fair coin twice. Consider the events A = "a head is obtained on the 1^{st} toss" and B = "a head is obtained on both tosses". Intuitively, there again appears to be some dependence at play here. On the other hand, if we instead consider the event C = "a head is obtained on the 2^{nd} toss", we do not think that the occurrence of A affects the chances that C will occur. If we need to reassess the probability of an event should we have knowledge that another event has occurred, then we call such a pair of events *dependent*, and otherwise we call them *independent*. We formalize this concept in the following mathematical definition.

Definition 2.3.1. *Events A and B are **independent events** if and only if $P(AB) = P(A)P(B)$. If the events are not independent, we refer to them as **dependent**.*

When Venn diagrams are used, let us imagine that the probability of events are roughly proportional to their area. This is justified in part because area and probability are two examples of "measures" in mathematics and share much of the same properties. We continue with this tradition, so that if two events (each with non-zero probability) are independent, then the "size" of their intersection, as represented by the corresponding area in a Venn diagram, must equal the product of their individual probabilities. This means, of course, that the intersection must be non-empty, and therefore the events are not mutually exclusive. For example, in the Venn diagram depicted in Figure 2.3.1, the area of (rectangular) region A is $0.4 \times 0.5 = 0.2$, the area of region B is $0.6 \times 0.5 = 0.3$, and the area

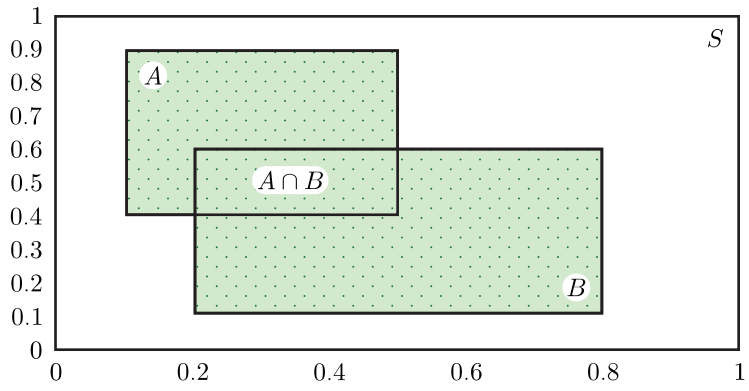


Figure 2.3.1: A Venn diagram illustrating independent events

of region AB is $0.3 \times 0.2 = 0.06$. Since $P(A)P(B) = (0.2)(0.3) = 0.06 = P(AB)$, the events A and B are independent. In fact, if you were to hold the rectangle A in place and only move the rectangle B to the right, the probability of the intersection (as represented by its area) would decrease and the events would become dependent.

Remark: A common misconception is that two events are independent if they are mutually exclusive. In fact, however, the opposite is true: mutually exclusive events A and B with $P(A) > 0$ and $P(B) > 0$ are never independent, since their intersection AB is empty and has a probability of 0.

Let us now revisit the coin tossing experiment from the opening paragraph of this section in light of Definition 2.3.1.

Example 2.3.1. Suppose that a fair coin is tossed twice. Define the following events:

A = "a head is obtained on the 1^{st} toss",

B = "a head is obtained on both tosses", and

C = "a head is obtained on the 2^{nd} toss".

Verify mathematically that A and B are dependent events, whereas A and C are independent events.

Solution: Using the sample space $S = \{HH, HT, TH, TT\}$, we clearly have that

$$P(A) = P(C) = \frac{2}{4} = \frac{1}{2},\ P(B) = \frac{1}{4},\ \text{ and }\ P(AB) = P(AC) = P(\{HH\}) = \frac{1}{4}.$$

Since $P(A)P(B) = \left(\frac{1}{2}\right)\left(\frac{1}{4}\right) = \frac{1}{8} \neq P(AB)$, A and B are dependent events by Definition 2.3.1. On the other hand, note that $P(A)P(C) = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{1}{4} = P(AC)$, implying that A and C are independent events. ■

Example 2.3.2. A fair die is rolled once. Define the events A = "the number rolled is even" and B = "the number rolled is greater than 3". Are A and B dependent or independent events?

Solution: Using the sample space $S = \{1, 2, 3, 4, 5, 6\}$, it immediately follows that

$$P(A) = P(B) = \frac{3}{6} = \frac{1}{2} \text{ and } P(AB) = P(\{4, 6\}) = \frac{2}{6} = \frac{1}{3}.$$

Note that $P(A)P(B) = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{1}{4} \neq P(AB)$. Therefore, A and B are dependent events. ■

Remark: An intuitive way to explain the dependence between A and B in Example 2.3.2 is to realize that B, on its own, only happens half the time. However, if A occurs, then we know that the number rolled is either a 2, 4, or 6. In other words, B now occurs $\frac{2}{3}$ of the time when A occurs. Thus, the occurrence of A does affect the chances of B occurring, inferring that A and B are <u>not</u> independent events.

Example 2.3.3. If A and B are independent events, show that $\bar{A}$ and B are independent events.

Solution: First of all, note that AB and $\bar{A}B$ are mutually exclusive events. Since $B = (AB) \cup (\bar{A}B)$, we have from Rule 5 that

$$P(B) = P(AB) + P(\bar{A}B).$$

Therefore,

$$\begin{aligned} P(\bar{A}B) &= P(B) - P(AB) \\ &= P(B) - P(A)P(B) \text{ since } A \text{ and } B \text{ are independent events} \\ &= (1 - P(A))\, P(B) \\ &= P(\bar{A})P(B). \end{aligned}$$

By Definition 2.3.1, $\bar{A}$ and B are independent events. ■

Remark: If A and B are independent events, the same approach used in the solution of Example 2.3.3 can be extended to show that (i) A and $\bar{B}$ are independent events, and (ii) $\bar{A}$ and $\bar{B}$ are independent events.

When there are two or more events under consideration, Definition 2.3.1 generalizes in the following way.

Definition 2.3.2. *The events $A_1, A_2, \ldots, A_n$, $n \geq 2$, are **mutually independent** if and only if*

$$P(A_{i_1}A_{i_2} \cdots A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k}),$$

for all distinct subscripts $i_1, i_2, \ldots, i_k$ chosen from the set $\{1, 2, \ldots, n\}$.

For example, in order for events A_1, A_2, and A_3 to be mutually independent, we would require

$$P(A_1A_2) = P(A_1)P(A_2),$$
$$P(A_1A_3) = P(A_1)P(A_3),$$
$$P(A_2A_3) = P(A_2)P(A_3),$$

and

$$P(A_1A_2A_3) = P(A_1)P(A_2)P(A_3).$$

Remarks:

(1) We will often shorten "mutually independent" to "independent" in order to reduce confusion with the term "mutually exclusive".

(2) The definition of independence can be applied in two different ways. First of all, if we can determine $P(A)$, $P(B)$, and $P(AB)$, then we can check whether A and B are actually independent events. On the other hand, if we happen to know (or assume) that, say A, B, and C, are independent events, then we can use Definition 2.3.2 as a rule of probability to calculate, for instance, a quantity such as $P(ABC)$.

(3) In the real world, careful scientific study is needed to determine if two events are independent. For example, are the events A and B independent if, for a random child living in a country, the events are defined as A = "the child lives within 5 kilometers of a nuclear power plant" and B = "the child has leukemia"? Determining whether such events are dependent, and if so the extent of the dependence, are problems of substantial importance, and can be handled by methods discussed in later statistics courses.

Example 2.3.4. A pseudo random number generator on a computer can give a sequence of independent random digits chosen from the set $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$. This means that (i) digit i, $i = 0, 1, \ldots, 9$, has probability $\frac{1}{10}$ of being chosen, and (ii) events determined by the different "trials" (i.e., selections) are independent of one another. As a result, we often refer to this as an "experiment with independent trials". Determine the probability that

(a) in a sequence of 5 trials, all the digits generated are odd.

(b) the number 9 occurs for the first time on trial 10.

Solution: (a) For $i = 1, 2, 3, 4, 5$, let A_i represent the event that the chosen digit from trial i is odd. Clearly, $P(A_i) = \frac{5}{10} = \frac{1}{2}$. Since the events A_1, A_2, A_3, A_4, and A_5 are independent, we have that

$$P(\text{all digits are odd}) = P(A_1A_2A_3A_4A_5) = P(A_1)P(A_2)P(A_3)P(A_4)P(A_5) = \frac{1}{2^5} = \frac{1}{32}.$$

(b) For $i = 1, 2, \ldots, 10$, let B_i represent the event that the digit 9 occurs on trial i. Note that $P(B_i) = 1 - P(\bar{B}_i) = \frac{1}{10} = 0.1$. Once again, the events $B_1, B_2, \ldots, B_{10}$ are independent, and this leads to

$$\begin{aligned} P(\text{9 occurs for the first time on trial 10}) &= P(\bar{B}_1 \bar{B}_2 \cdots \bar{B}_9 B_{10}) \\ &= P(\bar{B}_1)P(\bar{B}_2) \cdots P(\bar{B}_9)P(B_{10}) \\ &= (0.9)^9(0.1). \end{aligned}$$

■

Remark: We implicitly assumed independence of trials in some of our earlier probability calculations from Chapter 1. For example, suppose a fair coin is tossed three times and consider the sample space

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

Assuming that the outcomes of the three tosses are independent of one another, we get that

$$P(\{HHH\}) = P(\text{heads on } 1^{\text{st}} \text{ toss})P(\text{heads on } 2^{\text{nd}} \text{ toss})P(\text{heads on } 3^{\text{rd}} \text{ toss}) = \left(\frac{1}{2}\right)^3 = \frac{1}{8}.$$

In a similar fashion, all the other simple events in S have probability $\frac{1}{8}$. In our earlier calculations, we implicitly assumed this was true by assigning the same probability of $\frac{1}{8}$ to all possible outcomes without thinking directly about independence. However, it is clear that if somehow the three tosses were not independent, then it might be a bad idea to assume each outcome has probability $\frac{1}{8}$. For example, instead of heads and tails, suppose H stands for "rain" and T stands for "no rain" on a given day. If you were to consider the weather over three consecutive days, would you really want to assign a probability of $\frac{1}{8}$ to each of the 8 simple events in S, even if this were in a season when the probability of rain on a given day was $\frac{1}{2}$?

Example 2.3.5. Suppose that a fair die is rolled twice. Define the following events:

A = "the number 3 occurs on the 1^{st} roll",

B = "the total of the two rolls equals 7", and

C = "the total of the two rolls equals 8".

Are A, B, and C independent events?

Solution: Using the sample space $S = \{(1,1), (1,2), \ldots, (6,6)\}$ with 36 simple events, we clearly have that

$$A = \{(3,1), (3,2), (3,3), (3,4), (3,5), (3,6)\},$$

$$B = \{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\},$$

and

$$C = \{(2,6),(3,5),(4,4),(5,3),(6,2)\}.$$

Note that $P(A) = P(B) = \frac{6}{36} = \frac{1}{6}$ and $P(C) = \frac{5}{36}$. Since

$$P(AB) = P(\{(3,4)\}) = \frac{1}{36} = \left(\frac{1}{6}\right)\left(\frac{1}{6}\right) = P(A)P(B),$$

A and B are independent events. On the other hand, we have that

$$P(AC) = P(\{(3,5)\}) = \frac{1}{36} \text{ and } P(A)P(C) = \left(\frac{1}{6}\right)\left(\frac{5}{36}\right) = \frac{5}{216}.$$

Since $P(AC) \neq P(A)P(C)$, A and C are dependent events. Therefore, the events A, B, and C are not independent. ■

Remark: At first glance, the findings in Example 2.3.5 can be a bit puzzling. In particular, if we look at the relationship between A and B and that between A and C, why are the pair of events independent if the total is 7 but dependent if the total is 8? The key to understanding is that regardless of the first roll, there is always one number on the second roll which makes the total equal to 7. Since the probability of getting a total of 7 started off being $\frac{1}{6}$, the outcome of the first roll does not affect these chances. However, for any total other than 7, the outcome of the first roll does affect the chances of getting that total (e.g., a first roll of 1 guarantees that the total cannot possibly be 8). This line of reasoning involves something called "conditional probability", which is closely related to independence and will be treated in the next section.

Section 2.3 Problems

2.3.1 If events A and B are independent with $P(A) = 0.3$ and $P(B) = 0.2$, determine $P(A \cup B)$.

2.3.2 If events A, B, and C are independent, prove that A is independent of $\bar{B} \cup C$.

2.3.3 Let E and F be independent events with $E = A \cup B$ and $F = AB$. Prove that either $P(AB) = 0$ or $P\left(\bar{A}\,\bar{B}\right) = 0$.

2.3.4 Three digits are chosen at random with replacement from the set $\{0,1,2,3,4,5,6,7,8,9\}$. Find the probability of each of the following events:

(a) all three digits are identical.

(b) all three digits exceed 4.

(c) all three digits are different.

(d) all three digits have the same parity (i.e., all odd or all even).

(e) at least two of the digits are strictly positive.

2.3.5 A six-sided die is weighted to give the following probabilities on a single throw:

number	1	2	3	4	5	6
probability	0.3	0.1	0.15	0.15	0.15	0.15

Suppose that this die is thrown five times. Assuming that events determined by different throws of the die are independent, find the probability that

(a) the number 1 does not occur.

(b) the number 2 does not occur.

(c) neither the number 1 nor the number 2 occurs.

(d) the numbers 1 and 2 both occur.

2.3.6 Customers at a store independently decide whether to pay by credit card or with cash. Suppose that the probability is 70% that a customer pays by credit card. Find the probability that

(a) 3 out of 5 customers pay by credit card.

(b) the 5^{th} customer is the 3^{rd} one to pay by credit card.

2.3.7 Two baseball teams play a best-of-seven series, in which the series ends as soon as one team wins four games. The first two games are to be played on team A's field, the next three games on team B's field, and the last two games on team A's field. The probability that A wins a game is 0.7 at home and 0.5 away from home. Assuming that the results of the games are independent of each other, find the probability that

(a) team A wins the series in 4 games.

(b) team B wins the series in 5 games.

(c) the series does not go to 6 games.

2.3.8 Consider a population consisting of F females and M males. This population includes f female smokers and m male smokers. An individual is chosen at random from the population. If A is the event that this individual is female and B is the event that he or she is a smoker, find necessary and sufficient conditions on f, m, F, and M so that A and B are independent events.

2.4 Conditional Probability and Product Rules for Intersections of Events

In many situations, we may want to determine the probability of some event A, while knowing that some other event B has already occurred. To this point, let $P(A|B)$ represent the probability that event A occurs, when we know that event B occurs. We call this the "conditional probability of A given B". Before providing a formal definition of $P(A|B)$, let us first revisit an example from the previous section to get a sense as to how $P(A|B)$ is to be defined.

In Example 2.3.2, we considered rolling a fair die once which resulted in the sample space $S = \{1, 2, 3, 4, 5, 6\}$. Recall the events A = "the number rolled is even" and B = "the number rolled is greater than 3". If we know that B occurs, this tells us that we rolled either a 4, 5, or 6. Of the times when B occurs, we have an even number $\frac{2}{3}$ of the time. Therefore, it follows that $P(A|B) = \frac{2}{3}$. More formally, note that we could obtain this same result by calculating $\frac{P(AB)}{P(B)}$, since $P(AB) = P(\{4, 6\}) = \frac{2}{6}$ and $P(B) = P(\{4, 5, 6\}) = \frac{3}{6}$. This is no coincidence, and it leads to the following definition.

Definition 2.4.1. *The **conditional probability of event** A, **given event** B, is*

$$P(A|B) = \frac{P(AB)}{P(B)} \text{ provided that } P(B) \neq 0.$$

Remark: If A and B are independent events, then $P(AB) = P(A)P(B)$. In this case, the definition of conditional probability simplifies to become

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A) \text{ provided that } P(B) \neq 0.$$

Therefore, we can say that two events A and B such that $P(A) > 0$ and $P(B) > 0$ are independent if and only if either of the following statements is true:

$$P(A|B) = P(A) \text{ or } P(B|A) = P(B).$$

This could have been taken as the definition of independence, as in some sense it is more intuitive than Definition 2.3.1. However, the above formulation fails to hold in the case when $P(A) = 0$ or $P(B) = 0$ whereas Definition 2.3.1 does continue to hold.

Example 2.4.1. Suppose that a fair coin is tossed three times. Find the probability that if at least one head occurs, then exactly one head occurs.

Solution: The sample space for this experiment is given by

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

Let us define the events A = "exactly one head occurs" and B = "at least one head occurs". Note that $A \subseteq B$. We wish to calculate $P(A|B)$. Clearly, we have

$$P(B) = 1 - P(\bar{B}) = 1 - P(\{TTT\}) = 1 - \frac{1}{8} = \frac{7}{8}$$

and

$$P(AB) = P(A) = P(\{HTT, THT, TTH\}) = \frac{3}{8}.$$

Therefore, we immediately obtain

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{\frac{3}{8}}{\frac{7}{8}} = \frac{3}{7}.$$

■

Example 2.4.2. A local design team and an international design team are each tasked with creating an advertisement for a new product. From past experience, the local team is successful 60% of the time and the international team is successful 50% of the time. The probability that at least one team is successful is 0.75. If exactly one successful design is produced, what is the probability that it was designed by the local team?

Solution: Define the events A = "local team has successful design" and B = "exactly one design is successful". Let us represent the underlying sample space as follows:

$$S = \{ss, sf, fs, ff\},$$

where ss denotes the outcome where both teams succeed, sf denotes the outcome where the local team succeeds and the international team fails, fs denotes the outcome where the international team succeeds and the local team fails, and ff denotes the outcome where both teams fail. Based on the given information, we have that

$$\begin{aligned} P(ss) + P(sf) &= 0.6, \\ P(ss) + P(fs) &= 0.5, \text{ and} \\ P(ss) + P(sf) + P(fs) &= 0.75. \end{aligned}$$

From the above equations, we readily obtain

$$\begin{aligned} P(fs) &= 0.75 - 0.6 = 0.15, \\ P(ss) &= 0.5 - 0.15 = 0.35, \text{ and} \\ P(sf) &= 0.6 - 0.35 = 0.25. \end{aligned}$$

Therefore, we wish to calculate

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(sf)}{P(\{sf, fs\})} = \frac{0.25}{0.25 + 0.15} = 0.625.$$

■

Example 2.4.3. The probability a randomly selected male is colour blind is 0.05, whereas the probability a female is colour blind is only 0.0025. If the population is 50% male, what fraction of the population is colour blind?

Solution: Let C be the event that the person selected is colour blind, M the event that the person selected is male, and $F = \bar{M}$ the event that the person selected is female. We wish to calculate $P(C)$ and are provided with the following information:

$$\begin{aligned} P(C|M) &= 0.05, \\ P(C|F) &= 0.0025, \text{ and} \\ P(M) &= P(F) = 0.5. \end{aligned}$$

From the definition of conditional probability, note that

$$P(C|M)P(M) = \frac{P(CM)}{P(M)}P(M) = P(CM).$$

In a similar fashion, we also have $P(C|F)P(F) = P(CF)$. In order to calculate $P(C)$, we use the fact that

$$C = CM \cup CF,$$

and since the events CM and CF are clearly mutually exclusive, we readily obtain

$$\begin{aligned} P(C) &= P(CM) + P(CF) \\ &= P(C|M)P(M) + P(C|F)P(F) \\ &= (0.05)(0.5) + (0.0025)(0.5) \\ &= 0.02625. \end{aligned}$$

■

The solution procedure described in Example 2.4.3 highlights a useful rule concerning the calculation of intersection probabilities. We begin with general rules pertaining to two and three events, respectively.

Rule 7a (Probability of the Intersection of Two Events):

$$P(AB) = P(A)P(B|A) \text{ provided that } P(A) > 0. \tag{2.4.1}$$

Proof: Assuming that $P(A) > 0$, (2.4.1) comes directly from the definition of $P(B|A)$ since

$$P(A)P(B|A) = P(A)\frac{P(BA)}{P(A)} = P(AB).$$

■

Rule 7b (Probability of the Intersection of Three Events):

$$P(ABC) = P(A)P(B|A)P(C|AB) \text{ provided that } P(A) > 0 \text{ and } P(AB) > 0. \tag{2.4.2}$$

Proof: Assuming that $P(A) > 0$ and $P(AB) > 0$, note that

$$P(A)P(B|A)P(C|AB) = P(A)\frac{P(AB)}{P(A)} \cdot \frac{P(CAB)}{P(AB)} = P(ABC),$$

which yields (2.4.2). ■

Remarks:

(1) With the aid of mathematical induction, a natural extension of Rules 7a and 7b can be made to the case of n events $A_1, A_2, \ldots, A_n$. In particular, we have

$$P(A_1A_2 \cdots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1A_2) \cdots P(A_n|A_1A_2 \cdots A_{n-1}). \tag{2.4.3}$$

For obvious reasons, (2.4.3) is generally referred to as the *product rule* for the n-fold intersection probability. Note that in the case when $A_1, A_2, \ldots, A_n$ are independent events (see Definition 2.3.2), (2.4.3) simplifies to become

$$\begin{aligned} P(A_1A_2 \cdots A_n) &= P(A_1)\frac{P(A_1A_2)}{P(A_1)} \cdot \frac{P(A_1A_2A_3)}{P(A_1A_2)} \cdots \frac{P(A_1A_2 \cdots A_n)}{P(A_1A_2 \cdots A_{n-1})} \\ &= P(A_1)\frac{P(A_1)P(A_2)}{P(A_1)} \cdot \frac{P(A_1)P(A_2)P(A_3)}{P(A_1)P(A_2)} \cdots \frac{P(A_1)P(A_2) \cdots P(A_n)}{P(A_1)P(A_2) \cdots P(A_{n-1})} \\ &= P(A_1)P(A_2) \cdots P(A_n). \end{aligned}$$

(2) In order to remember the above product rules, it helps to imagine that the events unfold in some chronological order, even if they do not. For example, the formula

$$P(ABCD) = P(A)P(B|A)P(C|AB)P(D|ABC)$$

could be interpreted as the probability that "A occurs" (first) and then "given A occurs, that B occurs" (next), and so on.

One of the most important rules in probability is based on the idea of breaking down an event of interest into smaller mutually exclusive pieces.

Rule 8 (Law of Total Probability): Let $\{A_i\}_{i=1}^n$ be a sequence of events which forms a partition of the sample space S into n mutually exclusive events, namely

$$\bigcup_{i=1}^{n} A_i = S \text{ and } A_iA_j = \emptyset \text{ for all } i \neq j.$$

If B is an arbitrary event in S and $P(A_i) > 0$ for $i = 1, 2, \ldots, n$, then

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

Proof: First of all, since the events $A_1, A_2, \ldots, A_n$ are mutually exclusive, it immediately follows that the events $BA_1, BA_2, \ldots, BA_n$ are also mutually exclusive. Making use of the distributive property of events (see Problem 2.1.3), we see that

$$B = B \cap S = B \cap \left(\bigcup_{i=1}^{n} A_i \right) = \bigcup_{i=1}^{n} BA_i.$$

By Rule 5, we therefore have that

$$P(B) = P\left(\bigcup_{i=1}^{n} BA_i \right) = \sum_{i=1}^{n} P(BA_i). \tag{2.4.4}$$

By Rule 7a, $P(BA_i) = P(A_i)P(B|A_i)$, and so (2.4.4) becomes

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

■

Example 2.4.4. In an insurance portfolio, 10% of the policyholders are in class 1 (high risk), 40% are in class 2 (medium risk), and 50% are in class 3 (low risk). Suppose that the probability there is a claim on a class 1 policy in a given year is 0.15. Similar claim probabilities for classes 2 and 3 are 0.05 and 0.02, respectively. Find the probability that a claim is made in a given year.

Solution: For a randomly selected policy, define the events A_i = "policy is of class i", $i = 1, 2, 3$, and B = "policy has a claim made". We are asked to find $P(B)$. Since A_1, A_2, and A_3 forms a partition

of the sample space, we apply the Law of Total Probability to obtain

$$\begin{aligned} P(B) &= P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3) \\ &= (0.1)(0.15) + (0.4)(0.05) + (0.5)(0.02) \\ &= 0.015 + 0.02 + 0.01 \\ &= 0.045. \end{aligned}$$

■

Remarks:

(1) Tree diagrams can be a practical tool for keeping track of conditional probabilities when using rules such as the Law of Total Probability. The idea is to draw a tree where each path represents a sequence of events. On any given branch of the tree, we write the conditional probability of that event given all the events on the branches leading up to it. The probability at any node of the tree is obtained by multiplying the probabilities on the branches connecting to the node, and equals the probability of the intersection of the events leading to it. For instance, the information given in Example 2.4.4 could be represented by the tree diagram in Figure 2.4.1. Note that the probabilities on the terminal nodes must always add up to 1.

(2) Due to the rule of countable additivity given by (2.2.6), the Law of Total Probability can be expressed more generally as follows: If $\{A_i\}_{i=1}^{\infty}$ is a sequence of mutually exclusive events such that $S = \bigcup_{i=1}^{\infty} A_i$ and $P(A_i) > 0$ for all $i \geq 1$, then for any event B in S,

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

The Law of Total Probability is often used in conjunction with another important rule in probability, enabling us to express conditional probabilities in terms of similar conditional probabilities but with the order of conditioning reversed.

Rule 9 (Bayes' Rule): Let $\{A_i\}_{i=1}^{n}$ be a sequence of events which forms a partition of the sample space S into n mutually exclusive events with $P(A_i) > 0$ for $i = 1, 2, \ldots, n$. If B is an arbitrary event in S and $j \in \{1, 2, \ldots, n\}$, then

$$P(A_j|B) = \frac{P(A_j)P(B|A_j)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)}.$$

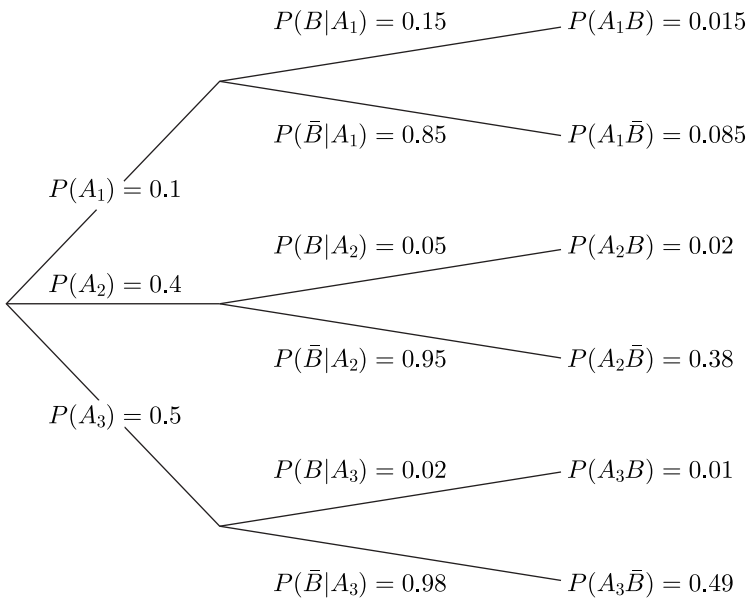


Figure 2.4.1: Tree diagram for Example 2.4.4

Proof: Note that

$$\begin{aligned} P(A_j|B) &= \frac{P(A_jB)}{P(B)} \\ &= \frac{P(A_j)P(B|A_j)}{P(B)} \quad \text{by Rule 7a} \\ &= \frac{P(A_j)P(B|A_j)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)} \quad \text{by the Law of Total Probability.} \end{aligned}$$

■

Example 2.4.5. Tests used to diagnose medical conditions are often imperfect, and give *false positive* or *false negative* results (as described in Problem 1.2.6 of Chapter 2). A relatively inexpensive blood test for HIV (Human Immunodeficiency Virus) that causes AIDS (Acquired Immune Deficiency Syndrome) has the following characteristics: the false negative rate is 2% and the false positive rate is 0.5%. It is assumed that 0.04% of Canadian males are infected with HIV. Find the probability that if a male tests positive for HIV, he actually has HIV.

Solution: Consider a Canadian male who is randomly selected from the population. Define the events A = "selected male has HIV" and B = "blood test is positive". We are interested in finding

$P(A|B)$. From the information provided, we know that

$$\begin{aligned} P(B|A) &= 0.98, \\ P(B|\bar{A}) &= 0.005, \text{ and} \\ P(A) &= 1 - P(\bar{A}) = 0.0004. \end{aligned}$$

By Bayes' Rule, it follows that

$$\begin{aligned} P(A|B) &= \frac{P(A)P(B|A)}{P(A)P(B|A) + P(\bar{A})P(B|\bar{A})} \\ &= \frac{(0.0004)(0.98)}{(0.0004)(0.98) + (0.9996)(0.005)} \\ &= \frac{0.000392}{0.00539} \\ &= \frac{196}{2695}. \end{aligned}$$

In other words, if a randomly selected male tests positive, there is still only a small chance (about 7.27%) that they actually have HIV! ■

Remark: It is common to think of Bayes' Rule in terms of updating our belief about a particular event A_j that takes into account new evidence in the form of event B. Specifically, our *posterior* belief $P(A_j|B)$ is calculated by multiplying our *prior* belief $P(A_j)$ by the likelihood $P(B|A_j)$ that B will occur if A_j is true. It is a relatively simple rule, but it has inspired approaches to problems in statistics and other areas including machine learning as well as classification and pattern recognition. In these areas, the term "Bayesian methods" is often used. The result is named after (Rev) Thomas Bayes, an English mathematician who proved it in the 1700's.

Example 2.4.6. Consider a box which contains three red balls, four black balls, and two white balls. Suppose that a fair six-sided die is tossed. If the result on the die is a 1, then two balls are drawn from the box without replacement. If the result on the die is a 2 or 3, then three balls are drawn from the box without replacement. If the result on the die is a 4 or 5, then four balls are drawn from the box without replacement. If the result on the die is a 6, then two balls are drawn from the box with replacement. If exactly one of the balls drawn is black, what is the probability that the balls were drawn from the box with replacement?

Solution: We begin by observing that the box consists of 4 black balls and 5 non-black balls. Let A_1 be the event that the die toss is a 1. Let A_2 be the event that the die toss is a 2 or 3. Let A_3 be the event that the die toss is a 4 or 5. Finally, let A_4 be the event that the die toss is a 6. Let B be the event that exactly one of the balls drawn is black.

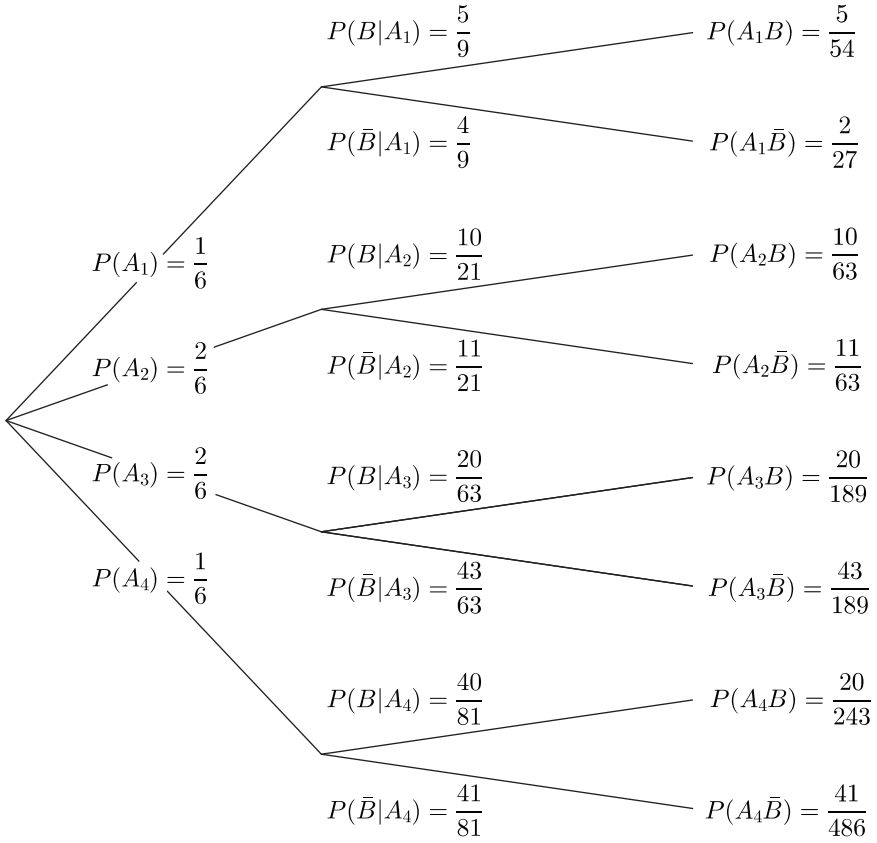


Figure 2.4.2: Tree diagram for Example 2.4.6

Using the Law of Total Probability, we have that

$$\begin{aligned} P(B) &= P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3) + P(A_4)P(B|A_4) \\ &= \frac{1}{6} \cdot \frac{\binom{4}{1}\binom{5}{1}}{\binom{9}{2}} + \frac{2}{6} \cdot \frac{\binom{4}{1}\binom{5}{2}}{\binom{9}{3}} + \frac{2}{6} \cdot \frac{\binom{4}{1}\binom{5}{3}}{\binom{9}{4}} + \frac{1}{6} \cdot \frac{4 \times 5 \times 2}{9^2} \\ &= \frac{5}{54} + \frac{10}{63} + \frac{20}{189} + \frac{20}{243} \\ &= \frac{1495}{3402}. \end{aligned}$$

Note that the event A_4 corresponds to the selection process being with replacement. As a result, if we apply Bayes' Rule, we immediately obtain

$$P(A_4|B) = \frac{P(A_4)P(B|A_4)}{P(B)} = \frac{\frac{1}{6} \cdot \frac{4\times5\times2}{9^2}}{\frac{1495}{3402}} = \frac{20}{243} \cdot \frac{3402}{1495} = \frac{56}{299}.$$

For completeness, the above information is nicely summarized by the tree diagram in Figure 2.4.2. ■

Section 2.4 Problems

2.4.1 Let A and B be events defined on a sample space with $P(A) = 0.3$, $P(B) = 0.4$, and $P(A|B) = 0.5$. Given that event B does not occur, what is the probability of event A?

2.4.2 In a typical year, 20% of the days have a temperature $> 22^oC$. On 40% of these days, there is no rain. During the rest of the year, when the temperature is $\leq 22^oC$, 70% of the days have no rain. What percentage of days in the year have rain and a temperature $\leq 22^oC$?

2.4.3 If you take a bus to work in the morning, there is a 20% chance you will arrive late. When you go by bicycle, there is a 10% chance you will arrive late. You go by bicycle 70% of the time and by bus 30% of the time. Given that you arrived late to work one morning, what is the probability you took the bus?

2.4.4 A box contains 4 coins – 3 fair coins and 1 biased coin for which the probability of tossing a head is 0.8. A coin is picked at random from the box and tossed 6 times. It shows 5 heads. Find the probability that this coin is fair.

2.4.5 At a police spot check, 10% of cars stopped have defective headlights and a faulty muffler. 15% have defective headlights and a muffler which is satisfactory. If a car which is stopped has defective headlights, what is the probability that the muffler is also faulty?

2.4.6 Among the very large population of UW students, suppose that 15% speak French and 45% are women. Suppose also that 20% of the women speak French. A committee of ten students is formed by randomly drawing from the population of UW students. What is the probability that there will be at least 1 woman and at least 1 French speaking student on the committee? (Because the population is very large, whether we draw students with or without replacement will make little difference. Therefore, assume in your calculations that the selection process is done *with replacement* so that the ten draws are independent of one another.)

2.4.7 In a very large population, people are classified as one of 3 genetic types: 30% are type A, 60% type are B, and 10% are type C. The probability a person carries another gene making them susceptible for a disease is 0.05 for A, 0.04 for B, and 0.02 for C. If nine unrelated persons are independently selected from this population, what is the probability that at least one of them is susceptible for the disease?

2.4.8 Abby, Barry, and Cindy are students who each independently answer a question on a test. The probability of getting the correct answer is 0.9 for Abby, 0.7 for Barry, and 0.4 for Cindy. If two of them get the correct answer, what is the probability that Cindy was the one with the incorrect answer?

2.4.9 A researcher wishes to estimate the proportion p of university students who have cheated on an examination. The researcher prepares a box containing 100 cards, 20 of which contain Question A and 80 of which contain Question B:

Question A: Were you born in July or August?

Question B: Have you ever cheated on an examination?

Each student who is interviewed draws a card at random with replacement from the box and answers the question it contains (either "yes" or "no"). Since only the student knows which question he or she is answering, confidentiality is assured and so the researcher hopes that the answers will be truthful. It is known that one-sixth of birthdays fall in July or August.

(a) What is the probability that a student answers "yes"?

(b) If x of n students answer "yes", estimate p.

(c) What proportion of the students who answer "yes" are responding to Question B?

2.4.10 Standard slot machines have three wheels, each marked with some number of symbols at equally spaced positions around the wheel. Suppose that there are 10 positions on each wheel, with three different types of symbols being used: flower, dog, and house. The three wheels spin independently and each has probability 0.1 of landing at any position. Each of the symbols (flower, dog, and house) is used in a total of 10 positions across the three wheels. A payout occurs whenever all three symbols showing are the same.

(a) If wheels 1, 2, and 3 have 2, 6, and 2 flowers, respectively, what is the probability all three positions show a flower?

(b) In order to minimize the probability of all three positions showing a flower, what number of flowers should go on wheels 1, 2, and 3? Assume that each wheel must have at least one flower.

2.4.11 **Spam Detection 1.** Many methods of spam detection are based on words or features that appear much more frequently in spam than in regular email. Conditional probability methods are then used to decide whether an email message is spam or not. For example, suppose we define the following events associated with a random email message:
Spam = "message is spam",

Not Spam = "message is not spam (i.e., regular)", and

A = "message contains the word Viagra".

From a study of email messages coming into a certain system, it is estimated that $P(\text{Spam}) = 0.5$, $P(A|\text{Spam}) = 0.2$, and $P(A|\text{Not Spam}) = 0.001$.

(a) What is the probability that a random email message contains the word Viagra?

(b) Calculate $P(\text{Spam}|A)$ and $P(\text{Not Spam}|A)$.

(c) If you declare any email message containing the word Viagra as spam, what fraction of spam emails would you detect?

2.4.12 **Spam Detection 2.** To increase the probability of detecting spam, we can use a larger set of email "features". These could be words or other features of a message which tend to occur with much different probabilities in spam and in regular email. (From your experience, what might be some useful features?) Suppose we identify 3 binary features, so that A_i denotes the event that feature i appears in a message, $i = 1, 2, 3$. Assume that A_1, A_2, and A_3 are independent events, given that a message is spam, and that they are also independent events, given that a message is regular. From a study of email messages coming into a certain system, it is estimated that

$$P(\text{Spam}) = 0.5, \quad \begin{array}{ll} P(A_1|\text{Spam}) = 0.2, & P(A_1|\text{Not Spam}) = 0.005, \\ P(A_2|\text{Spam}) = 0.1, & P(A_2|\text{Not Spam}) = 0.004, \\ P(A_3|\text{Spam}) = 0.1, & P(A_3|\text{Not Spam}) = 0.005. \end{array}$$

(a) What is the probability that a random email message has all three features?

(b) Calculate $P(\text{Spam}|A_1A_2A_3)$.

(c) Suppose that an email message has features 1 and 2 present, but feature 3 is not present. Calculate $P(\text{Spam}\,|A_1A_2\bar{A}_3)$.

(d) If you declare any email message with at least one of the features present as spam, what fraction of spam emails would you detect?

(e) Given that an email message is declared as spam according to the rule in part (d), what is the probability that the message is actually spam?

(f) Given that an email message is declared as spam according to the rule in part (d), what is the probability that feature 1 is present?

Chapter 3

Univariate Discrete Probability Distributions

3.1 Discrete Random Variables

As we have seen over the first two chapters, probability models are used to describe outcomes associated with random processes. So far, we have used simple events in sample spaces to describe such outcomes. In this chapter, we introduce numerical-valued variables, known as **random variables**, to describe outcomes. This formulation allows probability models to be manipulated more readily using ideas from algebra, calculus, and even geometry.

Simply put, a random variable is a numerical-valued variable that represents outcomes in an experiment or random process. For example, suppose an experiment consists of tossing a coin three times. If we define X to be the number of heads that occur over the three tosses, then X would be a random variable. Associated with any random variable is a **range** A, which is the set of possible values for the random variable. In our coin-tossing experiment, the random variable X has range $A = \{0, 1, 2, 3\}$.

Random variables are defined for every outcome of a random experiment (i.e., for every outcome $a \in S$). For each possible value x of a random variable X, there is a corresponding set of outcomes in the sample space S which results in this value of x (meaning "$X = x$" occurs). In rigorous mathematical treatments of probability, a random variable is defined as a function on a sample space in the following manner:

Definition 3.1.1. *A **random variable** X is a function (i.e., $X : S \mapsto \mathbb{R}$) that assigns a real number to each point in a sample space S.*

To understand this definition, consider again the experiment in which a coin is tossed three times. As before, let X be the number of heads that occur so that the range of X is $A = \{0, 1, 2, 3\}$. Suppose that we use the sample space

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

to describe the outcomes of this experiment. For each outcome a in S, the value of the function $X(a)$ is obtained by counting the number of heads corresponding to the outcome a. As a result, each of the outcomes "$X = x$" represents an event (either simple or compound) from S. In this particular example, we would have the following:

Outcomes	Definition of the event
$X = 0$	$\{TTT\}$
$X = 1$	$\{HTT, THT, TTH\}$
$X = 2$	$\{HHT, HTH, THH\}$
$X = 3$	$\{HHH\}$

Table 3.1.1: Mapping of outcomes in S to each numerical value of X.

Since some value of X in the range A must occur, the events of the form "$X = x$" for $x \in A$ form a partition of the sample space S. In particular, note that the events in the second column of Table 3.1.1 are mutually exclusive and their union is the entire sample space:

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

As you may recall, a function is a mapping of each point in a domain into a unique point. For example, the function $f(x) = x^3$ maps the point $x = 2$ in the domain into the point $f(2) = 8$ in the range. We are familiar with this rule for mapping being defined by a mathematical formula. However, the rule for mapping a point in the sample space (domain) into the real number in the range of a random variable is often expressed in words rather than by a formula. As a convention, we will generally denote random variables by upper-case letters (such as X or Y) and denote the actual numbers taken by random variables (i.e., their realized values) by lower-case letters (such as x or y).

We generally classify random variables into two types, according to how large their range of values is:

(1) **Discrete random variables** take on values in a countable set. Recall that a set is countable if its elements can be placed in a one-to-one correspondence with a subset of the positive integers.

(2) **Continuous random variables** take on values in some interval of real numbers, such as $(0, 1)$ or $(0, \infty)$ or $(-\infty, \infty)$. Keep in mind that the cardinality of the real numbers in an interval is not countable.

Some examples of each type of random variable are given below:

Discrete random variable	Continuous random variable
number of people in a car	total weight of people in a car
number of cars in a parking lot	distance between cars in a parking lot
number of phone calls to 911	time between calls to 911

Remarks:

(1) We assume that the quantities described in the second column of the above table are measured "exactly" and are not rounded (i.e., discretized) in any way.

(2) In theory, there could also be **mixed** random variables which are discrete-valued over part of their range and continuous-valued over some other portion of their range. We will not explore this possibility here. Instead, we will focus our efforts in this chapter on the treatment of discrete random variables. Continuous random variables will be considered later on in Chapter 5.

Since "$X = x$" represents an event of some kind, we will be interested in its probability, which we will write as $P(X = x)$. Our aim is to set up general models which describe how the probability is distributed among the possible values in the range A of a random variable X. To do this, we define for any discrete random variable X its so-called *probability mass function.*

Definition 3.1.2. *The **probability mass function** (pmf) of a random variable X is the function*

$$f(x) = P(X = x), \text{ defined for all } x \in A.$$

The set of pairs $\{(x, f(x)) : x \in A\}$ is called the **probability distribution** of X. In the earlier example, assuming that the coin being tossed is a fair one, we can determine the probability distribution of X. Referring to Table 3.1.1, the pmf of X would be computed as follows:

$$\begin{aligned}
f(0) &= P(X = 0) = P(\{TTT\}) = \frac{1}{8}, \\
f(1) &= P(X = 1) = P(\{HTT, THT, TTH\}) = \frac{3}{8}, \\
f(2) &= P(X = 2) = P(\{HHT, HTH, THH\}) = \frac{3}{8}, \\
f(3) &= P(X = 3) = P(\{HHH\}) = \frac{1}{8}.
\end{aligned}$$

Alternatively, it is straightforward to see that the number of outcomes in S corresponding to each of the four events of the form "$X = x$" is given by $\binom{3}{x}$ using the counting arguments of Section 1.3, meaning that we can specify a simple algebraic expression for the pmf, namely

$$f(x) = \frac{\binom{3}{x}}{8} \text{ for } x = 0, 1, 2, 3.$$

Sometimes, we may choose to include the name of the random variable in the subscript of its pmf, meaning that we may write $f_X(x)$ to be clear that $f_X(x)$ represents the pmf of X. This will become more important as we move towards considering more than one random variable at a time. Furthermore, in what follows, we adopt the convention that if $x \notin A$, then $f(x) = 0$. Therefore, all probability mass functions must satisfy the following two properties:

(1) $f(x) \geq 0$ for all $x \in \mathbb{R}$,

(2) $\sum\limits_{all\ x} f(x) = \sum\limits_{x \in A} f(x) = 1.$

By implication, these properties ensure that $f(x) \leq 1$ for all $x \in \mathbb{R}$. We now consider some further examples of discrete random variables and their associated probability distributions.

Example 3.1.1. Suppose that two fair six-sided dice are thrown. Let X be the sum of the values on their upturned faces. Determine the probability distribution of X.

Solution: We consider the sample space of this experiment to be

$$S = \{(1, 1), (1, 2), \ldots, (6, 6)\},$$

where the outcome (i, j) refers to getting upturned face i on the first die and upturned face j on the second die. There are $6^2 = 36$ equally likely outcomes in S. The mapping of outcomes in S to each numerical value of X is summarized below:

Outcomes	Definition of the event
$X = 2$	$\{(1, 1)\}$
$X = 3$	$\{(1, 2), (2, 1)\}$
$X = 4$	$\{(1, 3), (2, 2), (3, 1)\}$
$X = 5$	$\{(1, 4), (2, 3), (3, 2), (4, 1)\}$
$X = 6$	$\{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\}$
$X = 7$	$\{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$
$X = 8$	$\{(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)\}$
$X = 9$	$\{(3, 6), (4, 5), (5, 4), (6, 3)\}$
$X = 10$	$\{(4, 6), (5, 5), (6, 4)\}$
$X = 11$	$\{(5, 6), (6, 5)\}$
$X = 12$	$\{(6, 6)\}$

As a result, it immediately follows that the pmf of X is given by

$$\begin{aligned}
f(2) &= P(X=2) = P(\{(1,1)\}) = \frac{1}{36},\\
f(3) &= P(X=3) = P(\{(1,2),(2,1)\}) = \frac{2}{36} = \frac{1}{18},\\
f(4) &= P(X=4) = P(\{(1,3),(2,2),(3,1)\}) = \frac{3}{36} = \frac{1}{12},\\
f(5) &= P(X=5) = P(\{(1,4),(2,3),(3,2),(4,1)\}) = \frac{4}{36} = \frac{1}{9},\\
f(6) &= P(X=6) = P(\{1,5),(2,4),(3,3),(4,2),(5,1)\}) = \frac{5}{36},\\
f(7) &= P(X=7) = P(\{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}) = \frac{6}{36} = \frac{1}{6},\\
f(8) &= P(X=8) = P(\{(2,6),(3,5),(4,4),(5,3),(6,2)\}) = \frac{5}{36},\\
f(9) &= P(X=9) = P(\{(3,6),(4,5),(5,4),(6,3)\}) = \frac{4}{36} = \frac{1}{9},\\
f(10) &= P(X=10) = P(\{(4,6),(5,5),(6,4)\}) = \frac{3}{36} = \frac{1}{12},\\
f(11) &= P(X=11) = P(\{(5,6),(6,5)\}) = \frac{2}{36} = \frac{1}{18}, \text{ and}\\
f(12) &= P(X=12) = P(\{(6,6)\}) = \frac{1}{36}.
\end{aligned}$$

The probability distribution of X can be neatly summarized using the following table:

x	2	3	4	5	6	7	8	9	10	11	12
$f(x)$	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{9}$	$\frac{5}{36}$	$\frac{1}{6}$	$\frac{5}{36}$	$\frac{1}{9}$	$\frac{1}{12}$	$\frac{1}{18}$	$\frac{1}{36}$

We remark that $\sum_{x=2}^{12} f(x) = 1$, as expected. ■

Example 3.1.2. Consider an experiment in which two digits are randomly chosen without replacement from the set of digits $\{1,2,3,4,5\}$. Determine the probability distribution of X, where X represents the *absolute difference* of the two digits chosen.

Solution: A possible sample space to use for this experiment is

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

consisting of the possible pairs of size 2 chosen from 5 distinct objects. In other words, S contains $\binom{5}{2} = 10$ equally likely outcomes. The following table shows how each numerical value of X is matched to the outcomes in S:

Outcomes	Definition of the event
$X = 1$	$\{(1,2),(2,3),(3,4),(4,5)\}$
$X = 2$	$\{(1,3),(2,4),(3,5)\}$
$X = 3$	$\{(1,4),(2,5)\}$
$X = 4$	$\{(1,5)\}$

With the use of the above table, the pmf of X is easily calculated:

$$\begin{aligned}
f(1) &= P(X=1) = P(\{(1,2),(2,3),(3,4),(4,5)\}) = \frac{4}{10} = \frac{2}{5},\\
f(2) &= P(X=2) = P(\{(1,3),(2,4),(3,5)\}) = \frac{3}{10},\\
f(3) &= P(X=3) = P(\{(1,4),(2,5)\}) = \frac{2}{10} = \frac{1}{5},\\
f(4) &= P(X=4) = P(\{(1,5)\}) = \frac{1}{10}.
\end{aligned}$$

Clearly, $\sum_{x=1}^{4} f(x) = 1$. Moreover, note that we can succinctly represent the pmf of X as follows:

$$f(x) = \frac{5-x}{10} \text{ for } x = 1,2,3,4.$$

■

Remarks:

(1) When specifying the formula for a pmf, it is important that one does not forget to provide its domain (i.e., the values x for which $f(x)$ is non-zero, or equivalently, the set A of possible values that the random variable can take on). This is an essential part of the definition of the pmf.

(2) We frequently plot the pmf of a discrete random variable X using a *probability histogram*. For convenience, we will only describe how it is done for random variables whose range is some subset of the integers. With this in mind, a probability histogram of $f(x)$ is a graph consisting of adjacent bars or rectangles. At each value $x \in A$, we place a rectangle with base on $(x - 0.5, x+0.5)$ and having a height of $f(x)$. Figure 3.1.1 illustrates the probability histogram of $f(x)$ corresponding to Example 3.1.1. In general, probabilities are depicted by areas in a probability histogram. Note that the areas of the rectangles in Figure 3.1.1 correspond to probabilities, so for example, $P(5 \leq X \leq 7)$ is the sum of the areas of the three rectangles which are centered above the points 5, 6, and 7.

While the pmf is the most common way of describing a probability model, there are other possibilities. One of them is by using the so-called *cumulative distribution function*.

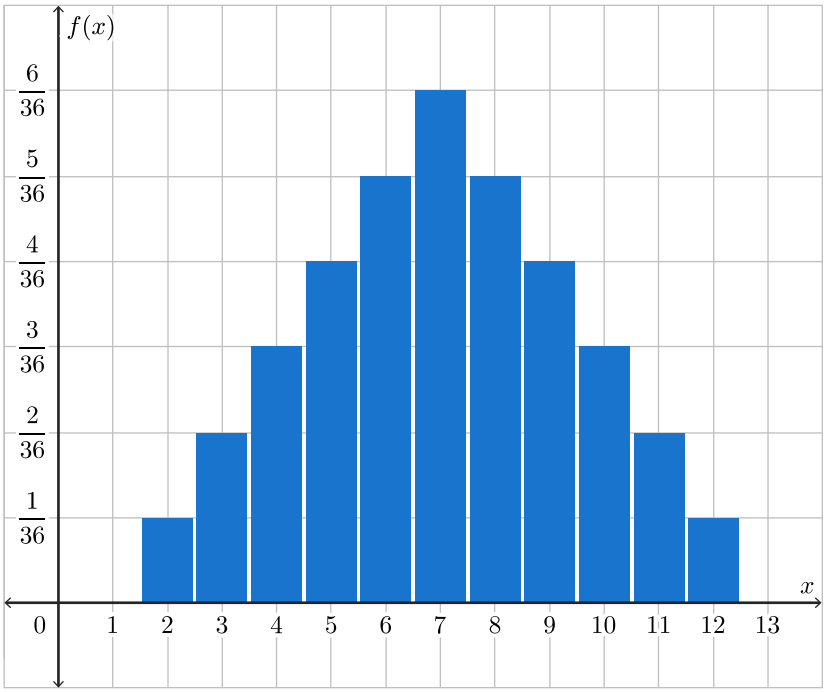


Figure 3.1.1: Probability histogram for Example 3.1.1

Definition 3.1.3. *The **cumulative distribution function** (cdf) of a random variable X is the function*

$$F(x) = P(X \leq x), \text{ defined for all } x \in \mathbb{R}.$$

In Example 3.1.2, the range of values for the random variable X is $A = \{1, 2, 3, 4\}$. For $x \in A$, consider the following table:

x	$f(x)$	$F(x) = P(X \leq x)$
1	$\frac{2}{5}$	$\frac{2}{5}$
2	$\frac{3}{10}$	$\frac{7}{10}$
3	$\frac{1}{5}$	$\frac{9}{10}$
4	$\frac{1}{10}$	1

Note that the values in the third column of the above table are partial sums of the values of the pmf in the second column. For instance, we have:

$$F(2) = P(X \leq 2) = P(X = 1) + P(X = 2) = f(1) + f(2) = \frac{7}{10},$$
$$F(3) = P(X \leq 3) = P(X = 1) + P(X = 2) + P(X = 3) = f(1) + f(2) + f(3) = \frac{9}{10}.$$

In addition, we remark that $F(x)$ is even defined for real numbers x such that $x \notin A$. In reference again to Example 3.1.2, note that

$$F(2.5) = P(X \leq 2.5) = F(2) = \frac{7}{10} \text{ and } F(4.8) = P(X \leq 4.8) = F(4) = 1.$$

The cdf of X corresponding to Example 3.1.2 is plotted in Figure 3.1.2. This is an example of a *step function*, a piecewise-defined function that consists of constant pieces, so that its plot resembles a set of steps.

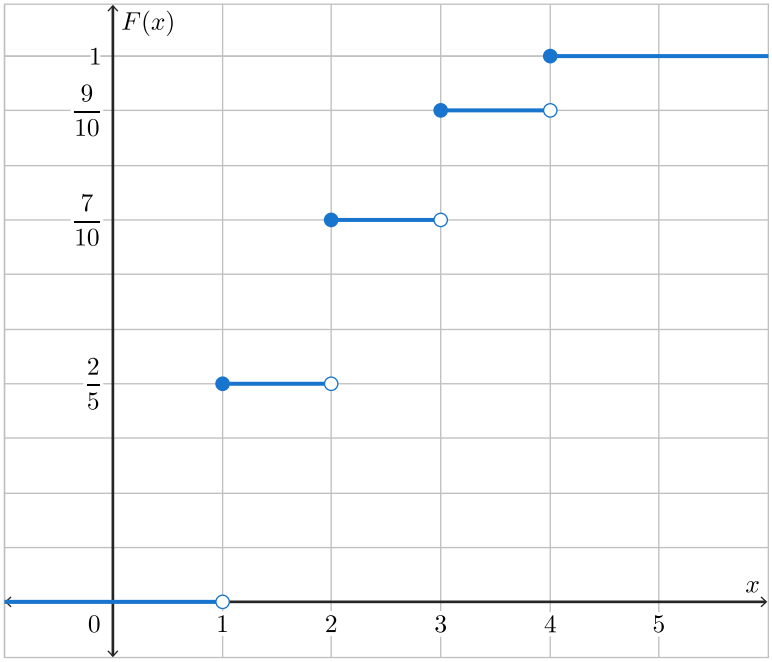


Figure 3.1.2: Plot of the cdf for Example 3.1.2

In general, the cdf $F(x)$ can be obtained from the pmf $f(x)$ via the formula

$$F(x) = P(X \leq x) = \sum_{u \leq x} f(u). \tag{3.1.1}$$

A cdf $F(x)$ has certain properties, just as a pmf $f(x)$ does. Obviously, since it represents a probability, $F(x) \in [0, 1]$ for all $x \in \mathbb{R}$. In addition, $F(x)$ must be a non-decreasing function in x (e.g., $P(X \leq 8)$ cannot possibly be less than $P(X \leq 7)$). In summary, we take note of the following important properties that a cdf $F(x)$ possesses:

(1) $0 \leq F(x) \leq 1$ for all $x \in \mathbb{R}$,

(2) $F(x)$ is a non-decreasing function of x for all $x \in \mathbb{R}$,

(3) $\lim_{h\to 0^+} F(x+h) = F(x)$,

(4) $\lim_{x\to -\infty} F(x) = 0$ and $\lim_{x\to \infty} F(x) = 1$.

Remarks:

(1) The third property above asserts that the function $F(x)$ is *continuous from the right*. For example, note that in Figure 3.1.2 the only discontinuities occur at the values $1, 2, 3$, and 4, and the limit as x approaches these values from the right is the value of $F(x)$ at these values. On the other hand, as x approaches these values from the left, the limit of $F(x)$ is the value of $F(x)$ on the lower step, so in general $F(x)$ is not continuous from the left.

(2) We have noted that $F(x)$ can be obtained from $f(x)$ via (3.1.1). The converse is also true, meaning that $f(x)$ can be obtained from $F(x)$. In particular, suppose that X takes on values in the range A of the form $A = \{a_1, a_2, a_3, \ldots\}$ where $a_1 < a_2 < a_3 < \cdots$. We can recover the pmf from knowledge of the cdf through the following relation:

$$f(a_1) = P(X = a_1) = P(X \le a_1) = F(a_1),$$

$$f(a_i) = P(X = a_i) = P(X \le a_i) - P(X \le a_{i-1}) = F(a_i) - F(a_{i-1}) \text{ for } i = 2, 3, 4, \ldots.$$

This asserts that $f(a_i)$ is the size of the "jump" in $F(x)$ at the point $x = a_i$. This is nicely visualized when looking at the plot of the cdf given in Figure 3.1.2.

(3) When a random variable has been defined, it is sometimes simpler to find its pmf directly, whereas other times it might be easier to find its cdf first and then its pmf. The following example demonstrates two approaches to the same problem.

Example 3.1.3. Suppose that N balls labelled $1, 2, \ldots, N$ are placed in a box, and n balls ($n \le N$) are randomly selected without replacement. Define the random variable X to be the *largest* number of the balls selected. Find the pmf of X.

Solution 1: If the event "$X = x$" is to occur, then of the n balls selected, the ball numbered x must be included as well as $n-1$ other balls with numbers chosen from the set $\{1, 2, \ldots, x-1\}$. Note that this implies that we require $x \ge n$. Therefore, this leads to

$$f(x) = P(X = x) = \frac{\binom{1}{1}\binom{x-1}{n-1}}{\binom{N}{n}} = \frac{\binom{x-1}{n-1}}{\binom{N}{n}} \text{ for } x = n, n+1, \ldots, N.$$

■

Solution 2: We proceed by finding the cdf $F(x) = P(X \leq x)$ first. Noting that the event "$X \leq x$" occurs if and only if all n balls selected are from the set of numbers $\{1, 2, \ldots, x\}$, we get

$$F(x) = \frac{\binom{x}{n}}{\binom{N}{n}} \text{ for } x = n, n+1, \ldots N.$$

With this cdf and $A = \{n, n+1, \ldots, N\}$, we can now determine the pmf using the relation in Remark (2) above:

$$f(n) = F(n) = \frac{\binom{n}{n}}{\binom{N}{n}} = \frac{1}{\binom{N}{n}}$$

and

$$f(x) = F(x) - F(x-1) = \frac{\binom{x}{n} - \binom{x-1}{n}}{\binom{N}{n}} \text{ for } x = n+1, n+2, \ldots, N.$$

However, note that the above numerator can be expressed as

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

which ultimately leads to

$$f(x) = \frac{\binom{x-1}{n-1}}{\binom{N}{n}} \text{ for } x = n, n+1, \ldots N,$$

just as before. ■

Section 3.1 Problems

3.1.1 Consider a random variable X having the following pmf:

x	0	1	2
$f(x)$	$9c^2$	$9c$	c^2

(a) Determine the value of c that makes the above pmf valid.

(b) Plot the cdf of X.

3.1.2 Suppose that a random variable X takes on values in the set $\{1, 2, 3, 4, 5\}$ and has cdf given by the following table:

x	1	2	3	4	5
$F(x)$	$0.1k$	0.2	$0.5k$	k	$4k^2$

(a) Determine the value of k that makes the above cdf valid.

(b) Calculate $P(2 < X \leq 4)$.

(c) Determine the pmf of X and plot its histogram.

3.1.3 Let X be a positive, integer-valued random variable with cdf of the form

$$F(x) = 1 - 2^{-x} \text{ for } x = 1, 2, 3, \ldots.$$

(a) Calculate $P(X \geq 5)$.

(b) Determine the pmf $f(x)$ and plot its histogram.

(c) Verify that $\sum_{x=1}^{\infty} f(x) = 1$.

3.1.4 Suppose that eight people, including you and a friend, randomly line up in single file. Let X be the number of people standing between you and your friend. Determine the pmf and cdf of X.

3.1.5 Two balls are drawn at random from a box containing seven balls numbered $1, 2, \ldots, 7$. Let X be the random variable representing the *smaller* of the numbers on the two drawn balls. Let Y be the random variable representing the *sum* of the numbers on the two drawn balls.

(a) Determine the pmf of X and of Y if the draws are made *without replacement.*

(b) Repeat part (a) if the draws are made *with replacement.*

3.1.6 A bin at a hardware store contains 35 forty-watt light bulbs and 70 sixty-watt light bulbs. A customer wants to buy 8 sixty-watt light bulbs, and randomly selects light bulbs from the bin without replacement until these 8 light bulbs have been found. Let X be the number of forty-watt light bulbs drawn from the bin. Determine the pmf of X.

3.2 Functions of Random Variables

The probability distribution of a random variable X describes how probabilities are assigned to the possible values that X can take on. In various applications, however, we may not be interested in the random variable X itself, but instead a different random variable Y which is related to X. For instance, consider a weather application in which a random variable X measures the daily high temperature in degrees Celsius. Instead, we may be interested in the transformation $Y = 1.8X + 32$, which alternatively specifies the temperature in degrees Fahrenheit. As a second example, X could represent the weight in kilograms of a randomly chosen person from a given population, and $Y = 2.2X$ would be the function which transforms kilograms to pounds. In both of these examples, note that Y happens to be a linear function of X, having the form $Y = aX + b$ where a and b are real constants. However, it is also possible to consider non-linear functions of the general form $Y = g(X)$. In the weather application, for example, if we wanted to display temperatures on a logarithmic scale, we might choose to use the function $g(X) = \ln X$.

Functions of random variables play an important role in probability and statistics, and we will begin to learn how to work with them in this section. The key observation to realize is that if $Y = g(X)$ is a function of a random variable X, then Y itself is also a random variable, since it provides a numerical value for each possible outcome. This is because every outcome in the sample space defines a numerical value x for X, and hence the corresponding numerical value $y = g(x)$ for Y. Note, however, that the probability distribution of Y will differ from that of X.

If X is a discrete random variable (with range A), then $Y = g(X)$ is also a discrete random variable, and its pmf can be determined using the pmf of X. To obtain $f_Y(y) = P(Y = y)$ for any value of y, we simply add the probabilities of all values of $x \in A$ such that $g(x) = y$ (denoted by the set $\{x \in A : g(x) = y\}$). In other words, if we let $f_X(x) = P(X = x)$ denote the pmf of X, we have that

$$f_Y(y) = \sum_{\{x \in A : g(x) = y\}} f_X(x). \tag{3.2.1}$$

Let us now revisit two examples from Section 3.1 to demonstrate how (3.2.1) can be used.

Example 3.2.1. Consider Example 3.1.1 in which X represents the sum of the upturned faces on the two thrown fair six-sided dice. Determine the probability distribution of $Y = X \pmod 5$.

Solution: Recall that for two numbers a and b, the number $a \pmod b$ is the remainder when a is divided by b. As a result, it immediately follows that the possible values of Y lie in the set $\{0, 1, 2, 3, 4\}$. To determine $f_Y(y)$ for a given value of y from this set, we must add $f_X(x)$ over all values x such that $x \pmod 5 = y$. The following table provides a helpful summary to summarize the situation at hand:

x	$y = x \pmod 5$	$f_X(x)$
2	2	$\frac{1}{36}$
3	3	$\frac{1}{18}$
4	4	$\frac{1}{12}$
5	0	$\frac{1}{9}$
6	1	$\frac{5}{36}$
7	2	$\frac{1}{6}$
8	3	$\frac{5}{36}$
9	4	$\frac{1}{9}$
10	0	$\frac{1}{12}$
11	1	$\frac{1}{18}$
12	2	$\frac{1}{36}$

With the use of the above table, the pmf of Y is easily calculated via (3.2.1):

$$\begin{aligned}
f_Y(0) &= f_X(5) + f_X(10) = \frac{1}{9} + \frac{1}{12} = \frac{7}{36},\\
f_Y(1) &= f_X(6) + f_X(11) = \frac{5}{36} + \frac{1}{18} = \frac{7}{36},\\
f_Y(2) &= f_X(2) + f_X(7) + f_X(12) = \frac{1}{36} + \frac{1}{6} + \frac{1}{36} = \frac{2}{9},\\
f_Y(3) &= f_X(3) + f_X(8) = \frac{1}{18} + \frac{5}{36} = \frac{7}{36},\\
f_Y(4) &= f_X(4) + f_X(9) = \frac{1}{12} + \frac{1}{9} = \frac{7}{36}.
\end{aligned}$$

Noting that $f_Y(y) = 0$ for $y \notin \{0, 1, 2, 3, 4\}$, we clearly have that $\sum\limits_{all\ y} f_Y(y) = \sum_{y=0}^{4} f_Y(y) = 1$. ■

Example 3.2.2. Consider Example 3.1.2 in which X represents the absolute difference of the two digits randomly chosen without replacement from the set of digits $\{1, 2, 3, 4, 5\}$. Determine the probability distribution of $Y = 2X - 5$.

Solution: Recall that $A = \{1, 2, 3, 4\}$ represents the range of the random variable X. Note that since Y is a linear function of X, each value $x \in A$ will map to a unique point of the form $y = 2x - 5$, and this immediately implies that $\{-3, -1, 1, 3\}$ is the corresponding range of the random variable Y. Due to this one-to-one mapping of points, the pmf of Y given by (3.2.1) simply gives rise to

y	-3	-1	1	3
$f_Y(y)$	$\frac{2}{5}$	$\frac{3}{10}$	$\frac{1}{5}$	$\frac{1}{10}$

■

Remark: Recall that in Example 3.1.2, an explicit formula was found for $f_X(x)$, namely

$$f_X(x) = \frac{5-x}{10} \text{ for } x = 1, 2, 3, 4.$$

For each value of $y \in \{-3, -1, 1, 3\}$, note that

$$\begin{aligned} f_Y(y) &= P(Y = y) \\ &= P(2X - 5 = y) \\ &= P\left(X = \frac{y+5}{2}\right) \\ &= f_X\left(\frac{y+5}{2}\right) \\ &= \frac{5 - \frac{y+5}{2}}{10} \\ &= \frac{5-y}{20}. \end{aligned}$$

In other words, we are able to obtain an explicit formula for $f_Y(y)$ via the known formula for the pmf of X. More generally, if $Y = g(X)$ where g is an *injective* (i.e., one-to-one) function that maps distinct elements of its domain to distinct elements of its range, then its inverse function, g^{-1}, exists and this leads to the following important connection:

$$f_Y(y) = P(Y = y) = P(g(X) = y) = P(X = g^{-1}(y)) = f_X(g^{-1}(y)). \tag{3.2.2}$$

In the case of a linear function of the form $Y = g(X) = aX + b$ where $a \neq 0$, (3.2.2) would yield

$$f_Y(y) = f_X\left(\frac{y-b}{a}\right).$$

Section 3.2 Problems

3.2.1 Suppose that X is a random variable with pmf of the form

$$f_X(x) = cx^2 \text{ for } x = 1, 2, \ldots, 9.$$

(a) Determine the value of c that makes the above pmf valid.

(b) Determine the pmf of $Y = 6 \pmod X$.

3.2.2 Consider a random variable X with pmf of the form

$$f_X(x) = \left(\frac{1}{2}\right)^x \text{ for } x = 1, 2, 3, \ldots.$$

Determine the cdf of $Y = 1/X$.

3.2.3 Consider an experiment in which a fair coin is independently tossed four times.

(a) Find the pmf of X representing the number of tails obtained.

(b) Use the pmf in part (a) to determine the probability distribution of Y representing the number of tails minus the number of heads obtained.

3.2.4 Three cards are randomly chosen without replacement from a deck of 14 cards consisting of 8 face cards and 6 non-face cards. Let X be the number of face cards chosen.

(a) Determine the pmf of X.

(b) A game is constructed such that \$3 is won for each face card selected and \$1 is lost for each non-face card selected. Use the pmf in part (a) to determine the probability distribution of Y representing the winnings of the game.

3.2.5 Consider the following game: A box contains seven coloured balls, two blue and five yellow. A player pays a \$9 entry fee and draws balls successively without replacement from the box until the last blue ball is obtained. For every draw made, the player receives \$2. Determine the probability distribution of the player's net winnings (i.e., winnings minus the \$9 entry fee) from this game. What is the probability that the player makes a profit?

3.2.6 Suppose that an integer x is randomly chosen from the set of integers $D = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$. With this chosen value of x, two subsets of D can be formed:

$$D_1 = \{0, 1, \ldots, x\} \text{ and } D_2 = \{x + 1, x + 2, \ldots, 9\}.$$

Determine the probability distribution of the random variable R representing the *ratio* of the number of elements in the smaller subset to the larger subset.

3.3 Expectation of a Random Variable

In both daily language and scientific reporting, quantitative properties are often described in terms of averages. For example, when midterm marks are released to a class of students, someone almost invariably asks what the average was. While we could list out all the marks to give a complete picture of how students performed, this would be somewhat tedious. It would also give more detail than could be immediately digested. On the other hand, if we summarize the results by telling a class the average mark, students immediately get a sense of how well the class performed. For this reason, "summary statistics" are often more helpful than providing full details of every outcome, or in the case of a random variable, its entire probability distribution.

To illustrate some of the ideas involved, suppose we were to observe vehicles crossing a toll bridge, and our interest lied in the random variable X representing the number of people in a vehicle. Suppose in a small study (or sample) that data on 25 vehicles were collected (i.e., the number of people in each of the 25 vehicles was recorded). We could list out all 25 numbers observed, but a more helpful way of presenting the sample data would be in terms of the *frequency distribution* (see Table 3.3.1), which specifies the number of times (i.e., frequency) each value of X occurred. Along these same lines, we could also provide a *frequency histogram* for this count data, as demonstrated in Figure 3.3.1.

X	Frequency count	Frequency
1	𝍸 |	6
2	𝍸 |||	8
3	𝍸	5
4	|||	3
5	||	2
6	|	1

Table 3.3.1: Frequency distribution for the toll bridge example

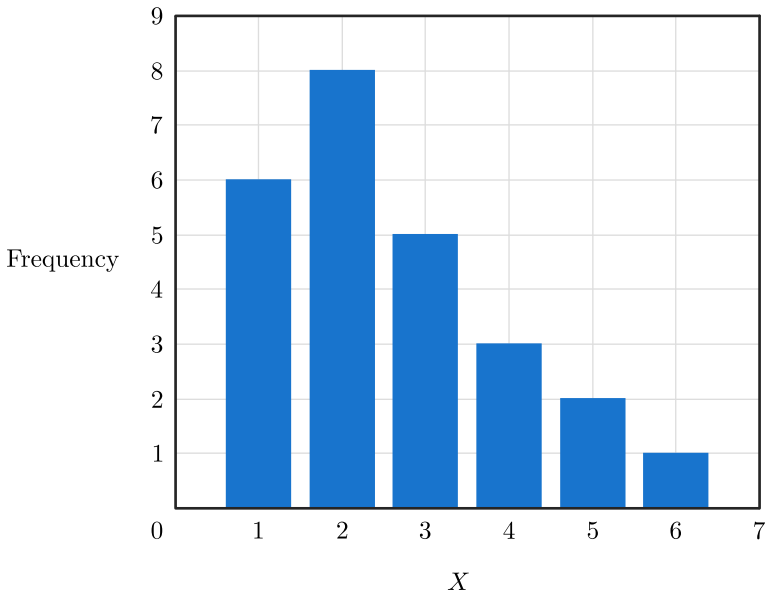


Figure 3.3.1: Frequency histogram for the toll bridge example

Frequency distributions or histograms are good summaries of data because they show the *variability* in the observed outcomes very clearly. The same can be said when looking at the probability

histogram of a random variable X. However, as we mentioned at the outset of this section, we might sometimes prefer a "single-number summary" of the data. The most common such summary is the average (or arithmetic mean) of the observed outcomes. More precisely, the arithmetic mean of n observed outcomes $x_1, x_2, \ldots, x_n$ from a random variable X is given by

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

and is generally referred to as the *sample mean*. The sample mean in our toll bridge example can be calculated as

$$\bar{x} = \frac{(6 \times 1) + (8 \times 2) + (5 \times 3) + (3 \times 4) + (2 \times 5) + (1 \times 6)}{25} = \frac{65}{25} = 2.6.$$

Alternatively, we can rearrange the above calculation as follows:

$$\begin{aligned} \bar{x} &= \frac{(6 \times 1) + (8 \times 2) + (5 \times 3) + (3 \times 4) + (2 \times 5) + (1 \times 6)}{25} \\ &= (1)\left(\frac{6}{25}\right) + (2)\left(\frac{8}{25}\right) + (3)\left(\frac{5}{25}\right) + (4)\left(\frac{3}{25}\right) + (5)\left(\frac{2}{25}\right) + (6)\left(\frac{1}{25}\right) \\ &= \sum_{x=1}^{6} x \cdot p_x, \end{aligned}$$

where p_x denotes the fraction of times the value of x occurred in the sample. In this particular example, we see that $p_1 = \frac{6}{25}$, $p_2 = \frac{8}{25}$, $p_3 = \frac{5}{25}$, $p_4 = \frac{3}{25}$, $p_5 = \frac{2}{25}$, and $p_6 = \frac{1}{25}$. Note that $\sum_{x=1}^{6} p_x = 1$.

Now, let us suppose that we actually know that the pmf of X is given by

x	1	2	3	4	5	6
$f(x)$	0.3	0.25	0.2	0.15	0.09	0.01

Thinking about the relative frequency definition of probability introduced in Section 1.1, if we observed a very large number of vehicles crossing the toll bridge, the fraction (or relative frequency) of times $X = 1$ would be 0.3, for $X = 2$ this proportion would be 0.25, and so on. In other words, *in theory* (i.e., according to the probability model), we would expect the mean to be equal to

$$(1)(0.3) + (2)(0.25) + (3)(0.2) + (4)(0.15) + (5)(0.09) + (6)(0.01) = 2.51,$$

if we observed an infinite number of vehicles crossing the toll bridge. Note that the determination of this "theoretical" mean requires us to know the probability distribution of X. With this as a backdrop, we now introduce the following mathematical definition.

Definition 3.3.1. *The **expected value** (also called the **mean** or the **expectation**) of a discrete random variable X with range A and pmf $f(x)$ is given by*

$$E(X) = \sum_{all\ x} xf(x) = \sum_{x \in A} xf(x).$$

Remarks:

(1) We point out that $E(X)$ being known as the "expected value" of X is somewhat misleading, since the formula for $E(X)$ in Definition 3.3.1 may yield a value which the random variable X never actually takes – hence unexpected! With that said, what we do clearly see in the above formula is that $E(X)$ can be viewed as a weighted average of the possible values of X, with the corresponding probabilities serving as weights. It is also useful to think of $E(X)$ as a "representative" value of X, which typically lies somewhere near the middle of its range. More precisely though, $E(X)$ represents the "balancing point" of the distribution. If a fulcrum was to be placed on the x-axis of the plot of the pmf $f(x)$, then $E(X)$ would represent the position of the fulcrum in order that the distribution be balanced.

(2) The expected value of X is often denoted by the Greek letter μ. In line with our earlier notational convention concerning the pmf, we may sometimes write μ_X to represent the mean of X.

(3) Confusion sometimes arises because we have two notations for the mean of a probability distribution, namely μ and $E(X)$. Perhaps one slight advantage to using the (lower-case) letter μ is that it makes it visually clearer that the expected value is <u>not</u> a random quantity like X, but instead is a non-random number (i.e., constant).

(4) When dealing with discrete random variables that take on a countably infinite number of values (i.e., the cardinality of A is countably infinite), one has to deal with the possibility that the infinite sum $\sum_{x \in A} xf(x)$ is not well-defined. More precisely, we will say that the expectation is well-defined if $\sum_{x \in A} |x| f(x) < \infty$. In this case, it is known that the infinite sum $\sum_{x \in A} xf(x)$ converges to a finite value that is independent of the order in which the various terms are summed. Throughout these course notes, in the absence of an indication to the contrary, we implicitly assume that the expected value of a random variable is well-defined.

Example 3.3.1. A small lottery sells 1000 tickets numbered $000, 001, \ldots, 999$. Tickets cost \$10 each. When all the tickets have been sold, the draw takes place. This draw consists of a single ticket from 000 to 999 being chosen at random. For ticket holders, the prize structure is as follows:

(i) If your exact ticket is drawn, then \$5000 is won.

(ii) If your ticket has the same first two numbers as the winning ticket (but the third number is different), then \$100 is won.

(iii) If your ticket has the same first number as the winning ticket (but the second number is different and the third number is irrelevant), then \$10 is won.

(iv) In all other cases, nothing is won.

What is the expected *net winnings* from a given ticket?

Solution: Let X be the random variable representing the net winnings from a given ticket. We observe that the range of the random variable X is given by $A = \{-10, 0, 90, 4990\}$. The pmf of X is determined as follows:

$$f(4990) = P(\text{exact ticket is drawn}) = \frac{1}{1000},$$

$$f(90) = P(1^{\text{st}} \text{ two numbers match and } 3^{\text{rd}} \text{ number is different}) = \frac{1 \times 1 \times 9}{1000} = \frac{9}{1000},$$

and

$$f(0) = P(1^{\text{st}} \text{ number matches and } 2^{\text{nd}} \text{ number is different}) = \frac{1 \times 9 \times 10}{1000} = \frac{9}{100},$$

so that $f(-10) = 1 - f(0) - f(90) - f(4990) = \frac{9}{10}$. Therefore, the expected net winnings from a given ticket is equal to the expected value of X, which we compute as

$$\begin{aligned} E(X) &= \sum_{x \in A} x f(x) \\ &= (-10)\left(\frac{9}{10}\right) + (0)\left(\frac{9}{100}\right) + (90)\left(\frac{9}{1000}\right) + (4990)\left(\frac{1}{1000}\right) \\ &= -\frac{16}{5}. \end{aligned}$$

Since the expected net winnings are negative, this means that a ticket holder is expected to incur a loss of \$3.20 when playing this particular lottery. ■

Remark: For any lottery or game of chance, the expected net winnings per play is a key value of interest. A fair game is one for which this value is equal to 0. Needless to say, casino games and lotteries are never fair, as the expected net winnings for a player are always negative.

Sometimes we may not be interested in the expected value of X itself, but in some function of X. Consider the toll bridge example once again, and suppose that there is a toll which depends on the number of vehicle occupants. For instance, a toll of \$3 per vehicle plus \$2 per occupant would produce an average toll for the 25 cars in the study equal to

$$(5)\left(\frac{6}{25}\right) + (7)\left(\frac{8}{25}\right) + (9)\left(\frac{5}{25}\right) + (11)\left(\frac{3}{25}\right) + (13)\left(\frac{2}{25}\right) + (15)\left(\frac{1}{25}\right) = \$8.20.$$

If X has the theoretical pmf $f(x)$ provided earlier, then the average value of this toll would be defined in the same manner, namely

$$(5)(0.3) + (7)(0.25) + (9)(0.2) + (11)(0.15) + (13)(0.09) + (15)(0.01) = \$8.02.$$

We would call this the expected value of $2X + 3$ and write $E(2X + 3) = 8.02$. As a further illustration, suppose that a toll designed to encourage "car pooling" charged $\$12/x^2$ if there were x people in the vehicle. This scheme would yield an average toll, in theory, of

$$\left(\frac{12}{1}\right)(0.3) + \left(\frac{12}{4}\right)(0.25) + \left(\frac{12}{9}\right)(0.2) + \left(\frac{12}{16}\right)(0.15) + \left(\frac{12}{25}\right)(0.09) + \left(\frac{12}{36}\right)(0.01) = \$4.7757.$$

In other words,

$$E\left(\frac{12}{X^2}\right) = 4.7757$$

is the expected value of the random variable $\frac{12}{X^2}$. Using this as motivation, we now introduce a formal procedure.

Theorem 3.3.1. *Suppose that X is a discrete random variable with range A and pmf $f(x)$. Then, the expected value of some real-valued function $g(X)$ of X is given by*

$$E\left(g(X)\right) = \sum_{x \in A} g(x) f(x). \tag{3.3.1}$$

Proof: To determine the expected value of the random variable $Y = g(X)$, we use Definition 3.3.1 and the pmf of Y given by (3.2.1) to get

$$E\left(g(X)\right) = E(Y) = \sum_{all\ y} y f_Y(y), \tag{3.3.2}$$

where

$$f_Y(y) = \sum_{\{x \in A : g(x) = y\}} f(x).$$

Let $D_y = \{x \in A : g(x) = y\}$ be the set of x-values from A with a given value y for $g(x)$, so that

$$f_Y(y) = \sum_{x \in D_y} f(x).$$

Substituting the above expression into (3.3.2), we obtain

$$\begin{aligned} E\left(g(X)\right) &= \sum_{all\ y} y \sum_{x \in D_y} f(x) \\ &= \sum_{all\ y} \sum_{x \in D_y} y f(x) \\ &= \sum_{all\ y} \sum_{x \in D_y} g(x) f(x) \\ &= \sum_{x \in A} g(x) f(x). \end{aligned}$$

■

With the above general result in place, we take note of two additional mathematical properties of expected value that can help to simplify calculations.

Theorem 3.3.2. *For real constants $c_1, c_2, \ldots, c_n$ and real-valued functions $g_1, g_2, \ldots, g_n$,*

$$E\left(\sum_{i=1}^{n} c_i g_i(X)\right) = \sum_{i=1}^{n} c_i E\left(g_i(X)\right).$$

Proof: Define the function $h(X) = \sum_{i=1}^{n} c_i g_i(X)$. Applying Theorem 3.3.1 and the properties of summation, we obtain

$$\begin{aligned} E\left(h(X)\right) &= \sum_{x\in A} h(x)f(x) \\ &= \sum_{x\in A}\left(\sum_{i=1}^{n} c_i g_i(x)\right) f(x) \\ &= \sum_{i=1}^{n} c_i \sum_{x\in A} g_i(x)f(x) \\ &= \sum_{i=1}^{n} c_i E\left(g_i(X)\right). \end{aligned}$$

■

Corollary 3.3.1. *For real constants a and b,*

$$E(aX + b) = aE(X) + b.$$

Proof: If we let $c_1 = a$, $c_2 = b$, $g_1(X) = X$, and $g_2(X) = 1$, then applying the result of Theorem 3.3.2 immediately yields

$$E(aX + b) = aE(X) + bE(1) = aE(X) + b\sum_{x\in A} 1 \cdot f(x) = aE(X) + b.$$

Note that the last equality follows since $\sum_{x\in A} f(x) = 1$. ■

Remarks:

(1) From Corollary 5.2.1, we see that the expected value of a constant b is, of course, equal to b.

(2) The above two results highlight the linearity properties that the expected value operator "E" possesses. In particular, Theorem 3.3.2 asserts that the expected value of a linear combination (of functions of a random variable) is equal to the linear combination of individual expected values. Likewise, Corollary 5.2.1 states that the expected value of a linear function of X is equal to the linear function of the expected value of X.

(3) If $g(x)$ is a *non-linear* function of x, then it is not generally true that $E\left(g(X)\right) = g\left(E(X)\right)$. It is a common pitfall to think that they should be equal. In the earlier example when the non-linear function $g(X) = \frac{12}{X^2}$ represented a particular toll bridge charge to use, we found that $E\left(\frac{12}{X^2}\right) = 4.7757$. However, note that

$$\frac{12}{E(X)^2} = \frac{12}{(2.51)^2} = \frac{12}{6.3001} \approx 1.90473,$$

and we see in this case that $E\left(g(X)\right) \neq g\left(E(X)\right)$.

While an average or expected value is a useful summary of a set of observations (or a probability distribution), it omits another important piece of information, namely the amount of variability. For example, it would be possible for car doors to be the right width, on average, and still have no doors fit properly. In the case of fitting car doors, we would also want the door widths to all be close to this correct average.

In what follows, we seek to provide a way of measuring the amount of variability (or spread) associated with a probability distribution. At first glance, you might think we could use the average difference between X and μ to indicate the amount of variation. In terms of expectation, this would be $E\left(X - \mu\right)$. However, since μ is a constant, we have that

$$E\left(X - \mu\right) = E(X) - \mu = \mu - \mu = 0.$$

We soon realize that for a measure of variability, we would like to use the expected value of a function that has the same sign for $X > \mu$ and for $X < \mu$. One might try the expectation of the absolute value of the distance between X and its mean, namely $E(|X-\mu|)$. An alternative, more mathematically tractable version squares the distance (in much the same way as Euclidean distance in $\mathbb{R}^n$ involves a sum of squared distances). This leads to the following definition.

Definition 3.3.2. *The **variance** of a random variable X is given by*

$$Var(X) = E\left((X - \mu)^2\right).$$

Remarks:

(1) Looking at Definition 3.3.2, the variance represents the average squared distance of the random variable X away from its mean, and as we will see, it turns out to be a very convenient measure of the variability of X.

(2) In keeping with the Greek letter notation we introduced for the mean, the variance of X is often denoted by σ^2 (or σ_X^2).

(3) Since $(X-\mu)^2$ can never be negative, $Var(X)$ is guaranteed to always be non-negative. However, could it possibly be equal to 0? Since every term in the formula $\sum_{x\in A}(x-\mu)^2 f(x)$ for the variance is non-negative, the sum is zero if and only if $(x-\mu)^2 f(x) = 0$ for every $x \in A$. This condition implies that $x = \mu$ and the random variable X is not really "random" – its only value is the mean μ, with probability 1.

Since we have squared the values in Definition 3.3.2, we note that variance is not measured on the same scale as X and μ. For example, if X and μ are weights in kilograms, the unit of measure of $Var(X)$ is square kilograms which does not have a clear meaning. We can regain the original units by taking the square root of variance. By doing so, we obtain another commonly-used measure of variability.

Definition 3.3.3. *The **standard deviation** of a random variable X is given by*

$$\sigma = SD(X) = \sqrt{Var(X)} = \sqrt{E\left((X-\mu)^2\right)}.$$

Example 3.3.2. Suppose that X is a discrete random variable with pmf given by

x	1	2	3	4	5	6
$f(x)$	0.3	0.25	0.2	0.15	0.09	0.01

Calculate $Var(X)$ and $SD(X)$.

Solution: We remark that the above pmf is the same one as the theoretical pmf provided earlier in our toll bridge example, and as such, we had previously computed $\mu = 2.51$. As a result, we can calculate the variance of X in the following manner:

$$\begin{aligned}
Var(X) &= E\left((X-\mu)^2\right) \\
&= \sum_{x=1}^{6}(x-2.51)^2 f(x) \\
&= (1-2.51)^2(0.3) + (2-2.51)^2(0.25) + (3-2.51)^2(0.2) + (4-2.51)^2(0.15) + (5-2.51)^2(0.09) + (6-2.51)^2(0.01) \\
&= (2.2801)(0.3) + (0.2601)(0.25) + (0.2401)(0.2) + (2.2201)(0.15) + (6.2201)(0.09) + (12.1801)(0.01) \\
&= 1.8099.
\end{aligned}$$

Hence, $SD(X) = \sqrt{1.8099} \approx 1.3453$. ■

As Example 3.3.2 demonstrates, the basic definition of variance can be awkward (i.e., cumbersome) to use in the calculation of $Var(X)$, whereas the following two formulas are often much easier to use.

Theorem 3.3.3.

$$Var(X) = E(X^2) - \mu^2 \tag{3.3.3}$$

and

$$Var(X) = E\left(X(X-1)\right) + \mu - \mu^2. \tag{3.3.4}$$

Proof: Making use of the linearity properties of expected value, we first observe that

$$\begin{aligned} Var(X) &= E\left((X-\mu)^2\right) \\ &= E(X^2 - 2\mu X + \mu^2) \\ &= E(X^2) - 2\mu E(X) + \mu^2 \text{ since } \mu \text{ is a constant} \\ &= E(X^2) - 2\mu^2 + \mu^2 \text{ since } E(X) = \mu \\ &= E(X^2) - \mu^2. \end{aligned}$$

Using this result and noting that $X^2 = X(X-1) + X$, we immediately obtain

$$Var(X) = E\left(X(X-1) + X\right) - \mu^2 = E\left(X(X-1)\right) + E(X) - \mu^2 = E\left(X(X-1)\right) + \mu - \mu^2.$$

■

Remark: The formula given by (3.3.4) is most often used when there is an $x!$ term present in the denominator of the pmf $f(x)$. Otherwise, the formula given by (3.3.3) is generally the more preferable one to use in practice.

Example 3.3.3. Suppose that X is a discrete random variable with pmf given by

x	1	2	3	4	5	6	7	8	9
$f(x)$	0.07	0.1	0.12	0.13	0.16	0.13	0.12	0.1	0.07

Calculate μ_X, σ_X^2, and σ_X.

Solution: We begin by calculating μ_X as follows:

$$\begin{aligned} \mu_X &= \sum_{x=1}^{9} x f(x) \\ &= (1)(0.07) + (2)(0.1) + (3)(0.12) + (4)(0.13) + (5)(0.16) + (6)(0.13) + (7)(0.12) + (8)(0.1) + (9)(0.07) \\ &= 5. \end{aligned}$$

Next, we calculate $E(X^2)$:

$$\begin{aligned} E(X^2) &= \sum_{x=1}^{9} x^2 f(x) \\ &= (1)(0.07) + (4)(0.1) + (9)(0.12) + (16)(0.13) + (25)(0.16) + (36)(0.13) + (49)(0.12) + (64)(0.1) + (81)(0.07) \\ &= 30.26. \end{aligned}$$

Using (3.3.3), we immediately obtain

$$\sigma_X^2 = E(X^2) - \mu^2 = 30.26 - 5^2 = 5.26.$$

Therefore, $\sigma_X = \sqrt{5.26} \approx 2.29347$. ■

Remarks:

(1) The probability histogram for Example 3.3.3 is given in Figure 3.3.2. In looking at the histogram, it should be obvious that $\mu_X = 5$. If a probability histogram is *symmetric* about the line $x = a$, then the mean of X is equal to a without any calculation required.

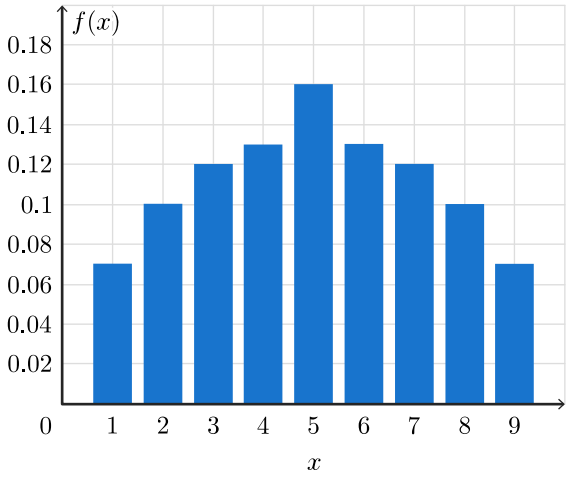


Figure 3.3.2: Probability histogram for Example 3.3.3

(2) To see how $Var(X)$ or $SD(X)$ influences the shape of a probability histogram, Figure 3.3.3 displays the probability histograms associated with four discrete random variables X_1, X_2, X_3, and X_4. For each of the four probability distributions, the range of the random variable is $\{1, 2, \ldots, 9\}$ and the mean is equal to 5, but the variance (or standard deviation) values do vary. Each probability histogram is labelled with its corresponding values for variance and standard deviation. We see that a small standard deviation value suggests that there is a greater probability of getting an observed value of the random variable near its mean. On the other hand, a large standard

deviation value translates to a greater probability of getting an observed value of the random variable that is further away from its mean.

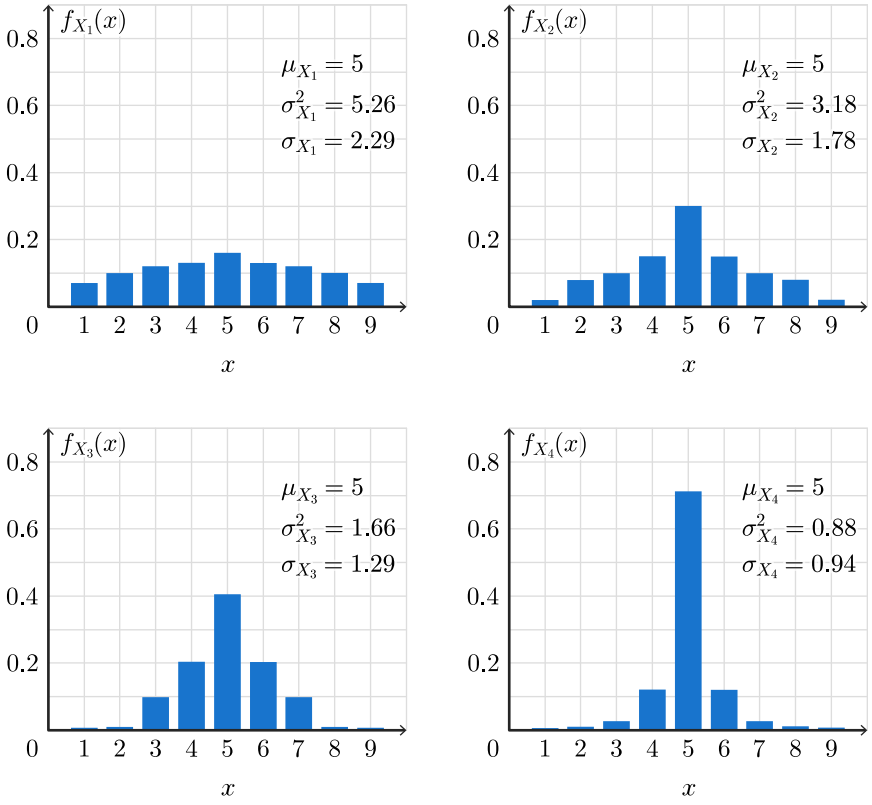


Figure 3.3.3: Comparison of variability over four discrete probability distributions

Recall earlier from Corollary 5.2.1 that if $Y = aX + b$ where a and b are real constants, then the mean of Y is equal to $aE(X) + b$. The next theorem specifies what happens when one considers the variance and standard deviation of a linear function of a random variable.

Theorem 3.3.4. *For real constants a and b,*

$$Var(aX + b) = a^2 Var(X)$$

and

$$SD(aX + b) = |a|\, SD(X).$$

Proof: Letting $Y = aX + b$, we established previously that $\mu_Y = a\mu_X + b$. Therefore, it follows that

$$\begin{aligned} Var(Y) &= E\left((Y - \mu_Y)^2\right) \\ &= E\left([aX + b - (a\mu_X + b)]^2\right) \\ &= E\left((aX - a\mu_X)^2\right) \\ &= E\left(a^2(X - \mu_X)^2\right) \\ &= a^2 E\left((X - \mu_X)^2\right) \\ &= a^2 Var(X). \end{aligned}$$

Taking square roots of both sides of the above equation (and remembering that standard deviation is not to be negative) yields $SD(Y) = |a|\, SD(X)$. ■

Remark: The result of Theorem 3.3.4 is to be expected. Adding a constant b to all values of a random variable has no effect on the amount of variability, so it makes sense that the formula for $Var(aX + b)$ would not depend on the value of b. Moreover, since variance is measured in squared units, multiplication of a random variable by a constant simply results in multiplying the variance by the constant squared.

Section 3.3 Problems

3.3.1 Let X be a discrete random variable with pmf of the form

$$f(x) = \begin{cases} \frac{11}{40} & \text{for } x = 1, \\ \frac{1}{2x} & \text{for } x = 2, 3, 4, 5, 6. \end{cases}$$

Calculate the mean, variance, and standard deviation of X.

3.3.2 A person plays a game in which a fair coin is tossed until the first instance of heads occurs. The person wins $\$2^x$ if x tosses are needed for $x = 1, 2, 3, 4, 5$, but loses \$256 if $x > 5$. Determine the mean and variance of the person's winnings.

3.3.3 The probability that a roulette wheel at a casino stops on a red number is 18/37. Suppose that you bet $\$b$ on "red". If the wheel stops on a red number, then you are paid $\$2b$ (so that your net winnings are $\$b$). If the wheel does not stop on a red number, then you have lost your bet.

(a) If you bet \$1 on each of 10 consecutive (independent) plays, what is your expected net winnings? What is your expected net winnings if you bet \$10 on a single play?

(b) For each of the two scenarios in part (a), calculate the probability that you make a profit (i.e., your net winnings are positive).

3.3.4 Consider the following game: Three cards are labelled \$1, \$5, and \$8, respectively. A player pays a \$5 entry fee, selects two cards at random with replacement, and then receives the sum of the winnings indicated on the two cards.

(a) Determine the mean and standard deviation of a player's net winnings.

(b) Suppose that the \$8 card is replaced by a different card, labelled \$$k$, and the game is replayed from the beginning. Determine the value of k which gives an expected net winnings of \$7.

3.3.5 Consider the slot machine discussed in Problem 2.4.10. Suppose that the number of each type of symbol on wheels 1, 2, and 3 is as given below:

	Wheel		
	1	2	3
Flower	2	6	2
Dog	4	3	3
House	4	1	5

If all three wheels stop on a flower, you win \$20 for a \$1 bet. If all three wheels stop on a dog, you win \$10. If all three wheels stop on a house, you win \$5. Otherwise, you win nothing. Determine the expected value and standard deviation of your winnings per dollar spent.

3.3.6 Consider the experiment described in Problem 3.2.3. Determine the mean and variance of Y representing the number of tails minus the number of heads obtained.

3.3.7 Consider the game described in Problem 3.2.5. Determine the mean and standard deviation of a player's net winnings.

3.3.8 Abbott and Costello are undergraduate UW Math students currently taking the same five courses. Let X be the number of assignments they have in one week. The pmf of X is given by the following table:

x	0	1	2	3	4	5
$f(x)$	0.09	0.10	0.25	0.40	0.15	0.01

The number of cups of coffee that Abbott and Costello drink in one week both depend on the number of assignments they have. Abbott drinks about $2X^2$ cups per week and Costello drinks

about $|2X-1|$ cups per week. Determine the mean and variance of the number of cups of coffee that each student will drink in a week.

3.3.9 A contestant on a game show has two questions, one from category A and one from category B. She may choose which category to attempt first, but she must answer the first question correctly to be able to attempt the remaining question. If she answers A correctly, then she receives \$100. If she answers B correctly, then she receives \$200. She knows the answer to A with probability 0.8 and the answer to B with probability 0.6 (assume independence in knowing the answers to the two questions).

(a) Which question should she attempt first to maximize her expected winnings?

(b) Suppose that she must now pay a \$50 penalty if she answers the first question incorrectly. What question should she attempt first?

3.3.10 If X is a non-negative, integer-valued random variable, show that

$$E(X) = \sum_{x=0}^{\infty} P(X > x)$$

and

$$E\left(X(X-1)\right) = 2\sum_{x=1}^{\infty} xP(X > x).$$

3.4 Moment Generating Functions

We have encountered two functions which characterize the probability distribution of a discrete random variable, namely the pmf and the cdf. In other words, if we are given the pmf/cdf of a discrete random variable X, then we can determine everything there is to know about the probability distribution of X. There is a third type of function, the *moment generating function*, which also uniquely characterizes a probability distribution. In fact, the moment generating function is closely related to other transforms used in mathematics such as the Laplace and Fourier transforms. Moreover, as the following definition specifies, the moment generating function is nothing more than a particular expected value.

Definition 3.4.1. *Consider a discrete random variable X with range A and pmf $f(x)$. The **moment generating function** (mgf) of X is defined as*

$$M(t) = E(e^{tX}) = \sum_{x\in A} e^{tx} f(x). \tag{3.4.1}$$

We will assume that the mgf is defined (i.e., finite) for values of t in an open neighbourhood of 0 (i.e., there exists $a > 0$ such that $\sum_{x\in A} e^{tx} f(x) < \infty$ for all $t \in (-a, a)$).

Remark: The formula given by (3.4.1) is simply an application of (3.3.1) with $g(X) = e^{tX}$. Note, however, that the function g contains a parameter t which can vary, and this is why we view $M(t)$ as a function of t. For an mgf to exist, it must exist in an open interval about 0, implying that we always have $M(0) = E(e^{0X}) = E(1) = 1$.

Example 3.4.1. Suppose that X is a discrete random variable with pmf given by

$$f(x) = \frac{12}{25(x+2)} \text{ for } x = -1, 0, 1, 2.$$

Find the mgf of X.

Solution: Applying (3.4.1), the mgf of X is given by

$$\begin{aligned} M(t) &= \sum_{x=-1}^{2} e^{tx} \frac{12}{25(x+2)} \\ &= \frac{12}{25}e^{-t} + \frac{6}{25}e^{0t} + \frac{4}{25}e^{t} + \frac{3}{25}e^{2t} \\ &= \frac{6}{25} + \frac{12}{25}e^{-t} + \frac{4}{25}e^{t} + \frac{3}{25}e^{2t}. \end{aligned}$$

Note that $M(t)$ exists for all $t \in \mathbb{R}$. ■

Example 3.4.2. Consider tossing a fair coin repeatedly and let X be the number of tosses until the first occurrence of heads appears. Find the mgf of X.

Solution: Assuming that coin tosses are independent of one another, it follows that the pmf of X is given by

$$\begin{aligned} f(x) &= P(1^{\text{st}}\ x-1 \text{ tosses are tails and } x^{\text{th}} \text{ toss is heads}) \\ &= \left(\frac{1}{2}\right)^{x-1}\left(\frac{1}{2}\right) \\ &= \left(\frac{1}{2}\right)^{x} \text{ for } x = 1, 2, 3, \ldots. \end{aligned}$$

Applying (3.4.1), the mgf of X is given by

$$M(t) = \sum_{x=1}^{\infty} e^{tx}\left(\frac{1}{2}\right)^{x} = \underbrace{\sum_{x=1}^{\infty}\left(\frac{e^t}{2}\right)^{x}}_{\text{infinite geometric series}} = \frac{e^t}{2}\left(1 - \frac{e^t}{2}\right)^{-1},$$

provided that $\frac{e^t}{2} < 1$, or equivalently, $t < \ln 2$. Since this last interval is an open interval of 0, the mgf of X exists and is given by the final expression in the above derivation. ■

As we alluded to at the outset of this section, one of the main uses of the mgf lies in its ability to characterize a probability distribution. It turns out that the mgf uniquely identifies a probability distribution in the sense that if two random variables have the same mgf, then they must also have the same probability distribution (i.e., the same pmf/cdf). It bears mentioning that moment generating functions must match for all values of t (in other words, they agree as *functions*) and not just for a few values. For example, if we were to somehow show that the mgf of a random variable Y was given by

$$M_Y(t) = \frac{e^t}{2}\left(1 - \frac{e^t}{2}\right)^{-1} \text{ for all } t < \ln 2,$$

then we would know that the random variable Y must have the same formula for its pmf that X has from Example 3.4.2, namely

$$f_Y(y) = \left(\frac{1}{2}\right)^y \text{ for } y = 1, 2, 3, \ldots.$$

This ultimately means that if we are able to determine the mgf for a given random variable, then (in theory) that mgf can be used to identify its probability distribution. What this result does is provide us with another tool for finding the probability distribution of a random variable. We state this result formally in the form of the following theorem (its proof, however, is omitted since it is beyond the scope of this course).

Theorem 3.4.1. ***(Uniqueness Theorem for Moment Generating Functions):*** *Suppose that random variables X and Y have moment generating functions $M_X(t)$ and $M_Y(t)$, respectively. If $M_X(t) = M_Y(t)$ for all $t \in (-a, a)$ for some $a > 0$, then X and Y have the same probability distribution.*

Example 3.4.3. Suppose that the random variable X has mgf of the form

$$M(t) = \frac{1}{4} + \frac{11}{20}e^{-7t} + \frac{1}{5}e^{11t} \text{ for } t \in \mathbb{R}.$$

Determine the pmf of X.

Solution: Based on the form of $M(t)$, we conjecture a possible pmf of X as follows:

x	-7	0	11
$f(x)$	$\frac{11}{20}$	$\frac{1}{4}$	$\frac{1}{5}$

Clearly, $A = \{-7, 0, 11\}$ with $f(-7) + f(0) + f(11) = \frac{11}{20} + \frac{1}{4} + \frac{1}{5} = 1$. Moreover, note that

$$\sum_{x \in A} e^{tx} f(x) = \frac{11}{20}e^{-7t} + \frac{1}{4}e^{0t} + \frac{1}{5}e^{11t} = M(t) \text{ for } t \in \mathbb{R}.$$

By Theorem 3.4.1, X indeed has the proposed pmf given in the above table. ■

The other main use of moment generating functions stems from its very name – these functions generate the *moments* of a probability distribution. How is this done? Perhaps more importantly, what are moments exactly? In short, moments are an important class of expectations, serving as quantitative metrics that help describe particular aspects of the shape of a probability distribution.

Definition 3.4.2. *For a random variable X, the quantity $E(X^n)$, $n = 1, 2, 3, \ldots$, is called the* $\boldsymbol{n^{th}}$ ***moment of the probability distribution****, or simply, the* $\boldsymbol{n^{th}}$ ***moment of*** *X.*

As we see, moments of a random variable X are simply expected values of functions of the form X^n for $n = 1, 2, 3, \ldots$. The mean $\mu = E(X)$ is therefore the first moment, $E(X^2)$, which is used in the calculation of variance, is the second moment, and so on. Generally speaking, moments quantify three characteristics of a probability distribution: location, scale, and shape. By convention, we plot probability histograms with their support values (i.e., values that do not have a probability mass of 0) on the horizontal axis and each supported value's probability on the vertical axis. A distribution's location refers to where its "center of mass" is along the horizontal axis. The scale refers to how spread out a distribution is. Scale stretches or compresses a distribution along the horizontal axis. Finally, the shape of a distribution refers to its overall geometry: is the distribution *bimodal*, *asymmetric*, *heavy-tailed*? In effect, the first moment describes a distribution's location, the second moment describes its scale, and all higher moments describe its shape. In most practical applications, third and fourth moments are sometimes of interest, but there is usually little statistical reason for examining higher moments than these.

Armed now with an understanding of what moments are and the kind of information they supply, the following theorem specifies precisely how the mgf generates moments, thereby providing us with another way to find the moments of a probability distribution.

Theorem 3.4.2. *If X is a discrete random variable with mgf $M(t)$, then*

$$E(X^n) = M^{(n)}(0) \quad \text{for } n = 1, 2, 3, \ldots,$$

where

$$M^{(n)}(0) = \frac{d^n}{dt^n} M(t)\Big|_{t=0}.$$

In other words, the n^{th} moment of X is equal to the n^{th} derivative of $M(t)$ evaluated at $t = 0$.

Proof: First of all, the assumed existence of $M(t) = \sum_{all\ x} e^{tx} f(x)$ for $t \in (-a, a)$ ensures that (i) we may interchange the order of summation and differentiation (with respect to t), and (ii) derivatives of

$M(t)$ of all orders exist at $t = 0$. Therefore, it follows that

$$\begin{aligned} M^{(n)}(t) &= \frac{d^n}{dt^n} \sum_{all\ x} e^{tx} f(x) \\ &= \sum_{all\ x} \frac{d^n}{dt^n}(e^{tx}) f(x) \\ &= \sum_{all\ x} x^n e^{tx} f(x). \end{aligned}$$

Evaluating this last expression at $t = 0$, we obtain $M^{(n)}(0) = \sum_{all\ x} x^n f(x) = E(X^n)$, as required. ■

Remark: For some probability distributions, it may be easier mathematically to find the moments through repeated differentiation of the mgf instead of direct summation involving the pmf. The following example illustrates a situation where this is the case.

Example 3.4.4. Suppose that X has the pmf derived in Example 3.4.2, namely

$$f(x) = \left(\frac{1}{2}\right)^x \text{ for } x = 1, 2, 3, \ldots.$$

Calculate the mean and variance of X.

Solution: By definition, the n^{th} moment of X is given by

$$E(X^n) = \sum_{x=1}^{\infty} x^n \left(\frac{1}{2}\right)^x.$$

However, even with $n = 1$, the above series is tricky to evaluate. It certainly becomes more challenging when $n = 2$. On the other hand, we found that the mgf of X in Example 3.4.2 is given by

$$M(t) = \frac{e^t}{2}\left(1 - \frac{e^t}{2}\right)^{-1} \text{ for } t < \ln 2.$$

Using both the product rule and chain rule for differentiation, let us take the first two derivatives of $M(t)$:

$$\begin{aligned} M^{(1)}(t) &= \frac{e^t}{2}(-1)\left(1 - \frac{e^t}{2}\right)^{-2}\left(-\frac{e^t}{2}\right) + \frac{e^t}{2}\left(1 - \frac{e^t}{2}\right)^{-1} \\ &= \frac{e^{2t}}{4}\left(1 - \frac{e^t}{2}\right)^{-2} + \frac{e^t}{2}\left(1 - \frac{e^t}{2}\right)^{-1}, \end{aligned}$$

$$\begin{aligned} M^{(2)}(t) &= \frac{e^{2t}}{4}(-2)\left(1 - \frac{e^t}{2}\right)^{-3}\left(-\frac{e^t}{2}\right) + \frac{2e^{2t}}{4}\left(1 - \frac{e^t}{2}\right)^{-2} + \frac{e^{2t}}{4}\left(1 - \frac{e^t}{2}\right)^{-2} + \frac{e^t}{2}\left(1 - \frac{e^t}{2}\right)^{-1} \\ &= \frac{e^{3t}}{4}\left(1 - \frac{e^t}{2}\right)^{-3} + \frac{3e^{2t}}{4}\left(1 - \frac{e^t}{2}\right)^{-2} + \frac{e^t}{2}\left(1 - \frac{e^t}{2}\right)^{-1}. \end{aligned}$$

By Theorem 3.4.2, we immediately obtain

$$\mu = E(X) = M^{(1)}(0) = \frac{e^{2(0)}}{4}\left(1-\frac{e^0}{2}\right)^{-2} + \frac{e^0}{2}\left(1-\frac{e^0}{2}\right)^{-1} = \frac{1}{4}\left(\frac{1}{2}\right)^{-2} + \frac{1}{2}\left(\frac{1}{2}\right)^{-1} = 1+1 = 2,$$

$$\begin{aligned} E(X^2) &= M^{(2)}(0) \\ &= \frac{e^{3(0)}}{4}\left(1-\frac{e^0}{2}\right)^{-3} + \frac{3e^{2(0)}}{4}\left(1-\frac{e^0}{2}\right)^{-2} + \frac{e^0}{2}\left(1-\frac{e^0}{2}\right)^{-1} \\ &= \frac{1}{4}\left(\frac{1}{2}\right)^{-3} + \frac{3}{4}\left(\frac{1}{2}\right)^{-2} + \frac{1}{2}\left(\frac{1}{2}\right)^{-1} \\ &= 2+3+1 \\ &= 6, \end{aligned}$$

and

$$Var(X) = E(X^2) - \mu^2 = 6 - 2^2 = 2.$$

■

We conclude this section by stating a useful property of the mgf pertaining to the case of a linear transformation.

Theorem 3.4.3. *Let X be a random variable with mgf $M_X(t)$. If a and b are real constants such that $Y = aX + b$, then Y has mgf given by $M_Y(t) = e^{bt}M_X(at)$.*

Proof: Since $Y = aX + b$, we have that

$$\begin{aligned} M_Y(t) &= E(e^{tY}) \\ &= E(e^{t(aX+b)}) \\ &= E(e^{atX+bt}) \\ &= E(e^{bt}e^{atX}) \\ &= e^{bt}E(e^{(at)X}) \text{ since } e^{bt} \text{ is a constant} \\ &= e^{bt}M_X(at). \end{aligned}$$

■

Remark: As we will see later, the result of Theorem 3.4.3 is of special importance when $a = \frac{1}{\sigma}$ and $b = -\frac{\mu}{\sigma}$, where $\mu \in \mathbb{R}$ and $\sigma > 0$ represent the mean and standard deviation of X, respectively. In this case, Theorem 3.4.3 specifies that the random variable $\frac{X-\mu}{\sigma}$ has mgf of the form

$$M_{\frac{X-\mu}{\sigma}}(t) = e^{-\frac{\mu t}{\sigma}}M_X\left(\frac{t}{\sigma}\right).$$

Section 3.4 Problems

3.4.1 Explain why there can be no random variable X for which $M_X(t) = \frac{t}{1-t}$ for $|t| < 1$.

3.4.2 Let X be a discrete random variable with mgf $M(t)$ for $t \in (-a, a)$ where $a > 0$.

(a) Prove that $P(X \geq a) \leq e^{-at}M(t)$ for $0 < t < a$.

(b) Prove that $P(X \leq a) \leq e^{-at}M(t)$ for $-a < t < 0$.

3.4.3 Let X be a discrete random variable having mgf of the form

$$M_X(t) = \frac{1}{2} + \frac{e^{-3t}}{3} + \frac{e^{2t}}{6} \text{ for } t \in \mathbb{R}.$$

Determine the cdf of X.

3.4.4 Let Y be a discrete random variable having mgf of the form

$$M_Y(t) = \frac{2}{3 - e^t} \text{ for } t < \ln 3.$$

Determine the pmf of Y.

3.4.5 Let $M(t)$ be the mgf of a random variable X. Define the function $C(t) = \ln M(t)$.

(a) Show that

$$C^{(1)}(0) = \frac{d}{dt}C(t)\Big|_{t=0} = E(X) \text{ and } C^{(2)}(0) = \frac{d^2}{dt^2}C(t)\Big|_{t=0} = Var(X).$$

(b) Use the results in part (a) to calculate the mean and variance of a random variable X with mgf $M(t) = e^{3(e^{4t}-1)}$ for $t \in \mathbb{R}$.

3.4.6 Suppose that X is a random variable with mgf of the form

$$M(t) = \frac{1}{125}\left(4 + e^t\right)^3 \text{ for } t \in \mathbb{R}.$$

Find the mgf of $Y = \frac{1}{2}(X - 4)$ and use it to calculate the mean and variance of Y.

3.4.7 Find the mgf of the discrete random variable X which has pmf given by

$$f(x) = 2\left(\frac{1}{3}\right)^{x+1} \text{ for } x = 0, 1, 2, \ldots.$$

Use the mgf to calculate the mean and variance of X.

3.4.8 Let Y be a discrete random variable having mgf of the form

$$M_Y(t) = \frac{e^{2t}}{4}\left(1 - \frac{e^t}{2}\right)^{-2} \text{ for } t < \ln 2.$$

Determine the pmf of Y. (*Hint:* Try and rewrite $M_Y(t)$ in terms of the result derived in Example 3.4.2.)

3.5 Special Discrete Probability Distributions

As we have seen throughout these course notes, many processes or problems share similar attributes or even have the exact same structure to them. For this reason, it is important to identify such common types of problems and develop general probability distributions that represent them. In this section, we present six *special* (i.e., commonly-used) model distributions for discrete random variables and study their key distributional properties. We introduce each model in terms of an abstract "physical setup" (or setting), and then consider specific illustrations of the setup.

3.5.1 Discrete Uniform Distribution

Physical Setup: Consider a random variable X which takes on values in the set $A = \{a, a+1, a+2, \ldots, b\}$, where a and b are integers such that $a \leq b$. Suppose that all values in A are equally likely. In this situation, X is said to have a discrete uniform distribution on the set $\{a, a+1, a+2, \ldots, b\}$. We write $X \sim DU(a, b)$ as a shorthand for "X is distributed according to a discrete uniform distribution with parameters a and b".

Illustrations:

(1) If X is the number obtained when a fair six-sided die is rolled once, then $X \sim DU(1, 6)$.

(2) Computer random number generators emit $DU(1, N)$ values for a specified positive integer N. These are used for many purposes, such as generating lottery numbers or providing automated random sampling from a set of N items.

Probability Distribution: There are $b-a+1$ values X can take on with equal probability. Therefore, the probability at each of these values must be $\frac{1}{b-a+1}$, and this immediately yields the following form for the pmf of X:

$$f(x) = \frac{1}{b-a+1} \text{ for } x = a, a+1, \ldots, b.$$

With the above pmf, we can also determine other pertinent distributional quantities. For example, if $x \in A$, note that

$$P(X \leq x) = \sum_{u \leq x} f(u) = \sum_{u=a}^{x} \frac{1}{b-a+1} = \frac{1}{b-a+1} \sum_{u=a}^{x} 1 = \frac{x-a+1}{b-a+1}.$$

As a result, the "full" cdf of X can be expressed as

$$F(x) = \begin{cases} 0 & \text{for } x < a, \\ \frac{\lfloor x \rfloor - a + 1}{b-a+1} & \text{for } a \leq x < b, \\ 1 & \text{for } x \geq b, \end{cases}$$

where $\lfloor x \rfloor$, commonly known as the *floor function* of x, denotes the greatest integer less than or equal to x. In addition, we can also obtain the mean and variance of X as follows:

$$\begin{aligned} E(X) &= \sum_{x=a}^{b} x \, \frac{1}{b-a+1} \\ &= \frac{1}{b-a+1} \sum_{y=1}^{b-a+1} (y + a - 1) \text{ if we let } y = x - a + 1 \\ &= \frac{1}{b-a+1} \left(\sum_{y=1}^{b-a+1} y + (a-1)(b-a+1) \right) \\ &= \frac{1}{b-a+1} \cdot \frac{(b-a+1)(b-a+2)}{2} + a - 1 \\ &= \frac{b-a+2+2(a-1)}{2} \\ &= \frac{a+b}{2} \end{aligned}$$

and

$$\begin{aligned}
Var(X) &= E(X^2) - (E(X))^2 \\
&= \sum_{x=a}^{b} x^2 \frac{1}{b-a+1} - \left(\frac{a+b}{2}\right)^2 \\
&= \frac{1}{b-a+1} \sum_{y=1}^{b-a+1} (y+a-1)^2 - \left(\frac{a+b}{2}\right)^2 \text{ if we let } y = x-a+1 \\
&= \frac{1}{b-a+1} \left(\sum_{y=1}^{b-a+1} y^2 + 2(a-1) \sum_{y=1}^{b-a+1} y + (a-1)^2(b-a+1) \right) - \left(\frac{a+b}{2}\right)^2 \\
&= \frac{1}{b-a+1} \left(\frac{(b-a+1)(b-a+2)[2(b-a+1)+1]}{6} + 2(a-1)\frac{(b-a+1)(b-a+2)}{2} \right) + (a-1)^2 - \left(\frac{a+b}{2}\right)^2 \\
&= \frac{(b-a+2)(2b-2a+3)}{6} + (a-1)(b-a+2) + \left(a-1-\frac{a+b}{2}\right)\left(a-1+\frac{a+b}{2}\right) \\
&= \frac{(b-a+2)(4a+2b-3)}{6} + \left(\frac{a-b-2}{2}\right)\left(\frac{3a+b-2}{2}\right) \\
&= \frac{(b-a+2)[8a+4b-6-(9a+3b-6)]}{12} \\
&= \frac{(b-a)(b-a+2)}{12}.
\end{aligned}$$

3.5.2 Binomial Distribution

Physical Setup: Suppose that an experiment has two types of distinct outcomes. Let us label these two outcomes as "s" (for success) and "f" (for failure), and let their respective probabilities be p (for s) and $1-p$ (for f). Consider a process in which we repeat the experiment n independent times, where n is positive integer. Let X be the number of successes obtained over this process. In this situation, X is said to have a binomial distribution. We write $X \sim Bin(n, p)$ as a shorthand for "X is distributed according to a binomial distribution with n repetitions and probability p of success". The n individual experiments (or repetitions) in the process just described are called *Bernoulli trials* (or simply "trials" for short), and the process is often referred to as a Bernoulli process.

Illustrations:

(1) If a fair six-sided die is rolled 10 times and X is the number of ones that occur, then $X \sim Bin(10, \frac{1}{6})$.

(2) In a microcircuit manufacturing process, suppose that 90% of the chips produced work correctly and 10% are defective and do not work correctly. If we independently select 25 chips from a

very large allotment of chips and let X be the number of selected chips that work, then $X \sim Bin(25, 0.9)$.

Remark: We must think carefully whether the physical process we are considering is closely approximated by a Bernoulli process, for which the key assumptions are that (i) the probability p of success is constant over the n trials, and (ii) the outcome (s or f) on any trial is independent of the outcome on the other trials. For Illustration (1), these assumptions seem appropriate. For Illustration (2), however, we would need to think about the manufacturing process involved. Microcircuit chips tend to be produced on "wafers" containing a large number of chips, and it is common for defective chips to cluster on wafers. This could mean that if we selected the 25 chips from the same wafer, or from only two or three wafers, that the "trials" (i.e., chips) might not be independent, or perhaps that the probability of a chip being defective changes.

Probability Distribution: First of all, the random variable X is modelling the number of observed successes over n trials, and so it is clear that $A = \{0, 1, \ldots, n\}$. Next, the number of different arrangements of x s's and $(n - x)$ f's over the n trials is given by

$$\binom{n}{x} \times \binom{n-x}{n-x} = \binom{n}{x}.$$

The probability for any one of these arrangements would have p multiplied together x times and $1-p$ multiplied together $n-x$ times, in some order, since the trials are independent of each other. In other words, each arrangement has probability $p^x(1-p)^{n-x}$. Therefore, we immediately obtain the following form for the pmf of X:

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \text{ for } x = 0, 1, \ldots, n. \tag{3.5.1}$$

Remarks:

(1) With the use of the well-known *binomial series formula*

$$\sum_{x=0}^{n} \binom{n}{x} a^x b^{n-x} = (a+b)^n \text{ for } a, b \in \mathbb{R},$$

it is an elementary exercise to verify that $\sum_{x=0}^{n} f(x) = 1$.

(2) Figure 3.5.1 displays the probability histogram for the binomial distribution with parameters $n = 20$ and $p = 0.3$. While the formula for $f(x)$ given by (3.5.1) may seem somewhat complex at first glance, the shape of the histogram is rather well-behaved since it increases to a maximum value near $np = 6$ and then decreases thereafter.

(3) Unfortunately, a closed-form expression for the cdf of a binomial distribution does not exist. Instead, one has to sum probabilities term by term in order to calculate cumulative probabilities.

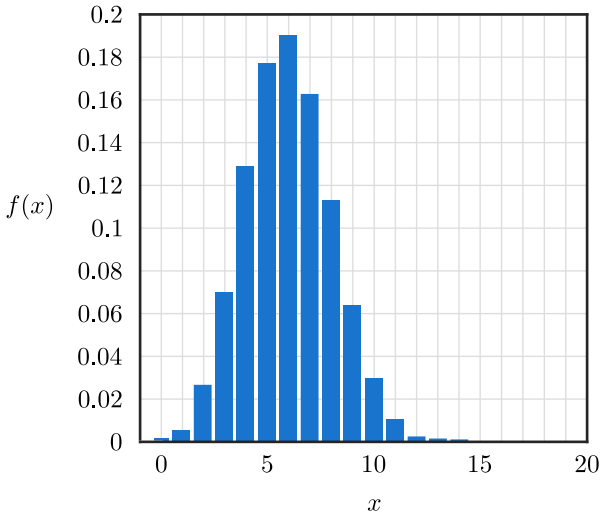


Figure 3.5.1: Probability histogram for a $Bin(20, 0.3)$ random variable

With the pmf of X given by (3.5.1), we can apply the binomial series formula to obtain the mgf of X:

$$\begin{aligned} M(t) &= \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x (1-p)^{n-x} \\ &= (pe^t + 1 - p)^n, \ t \in \mathbb{R}. \end{aligned}$$

It immediately follows that

$$M^{(1)}(t) = n(pe^t + 1 - p)^{n-1} pe^t$$

and

$$M^{(2)}(t) = n(pe^t + 1 - p)^{n-1} pe^t + npe^t(n-1)(pe^t + 1 - p)^{n-2} pe^t.$$

Thus, the mean and variance of X are given by

$$E(X) = M^{(1)}(0) = n(pe^0 + 1 - p)^{n-1} pe^0 = np$$

and

$$\begin{aligned} Var(X) &= E(X^2) - (E(X))^2 \\ &= M^{(2)}(0) - (np)^2 \\ &= n(pe^0 + 1 - p)^{n-1} pe^0 + npe^0(n-1)(pe^0 + 1 - p)^{n-2} pe^0 - n^2p^2 \\ &= np + np(n-1)p - n^2p^2 \\ &= np + n^2p^2 - np^2 - n^2p^2 \\ &= np(1-p). \end{aligned}$$

3.5.3 Hypergeometric Distribution

Physical Setup: Consider a population of N objects which can be classified into one of two distinct types. Let us refer to one type of object as type "success" and the other as type "failure". Among this collection of N objects, suppose that r of them are success-type objects and $N - r$ of them are failure-type objects. If a sample of $n \leq N$ objects are selected at random *without replacement* from this population and X denotes the number of success-type objects obtained, then X is said to have a hypergeometric distribution. We write $X \sim HG(N, r, n)$ as a shorthand for "X is distributed according to a hypergeometric distribution with parameters N, r, and n".

Illustrations:

(1) If five cards are randomly dealt from a standard deck of 52 playing cards and X is the number of diamond cards drawn, then $X \sim HG(52, 13, 5)$.

(2) In a fleet of 60 trucks, suppose that 8 of them have defective brakes. As part of a safety check, 10 trucks from the fleet are randomly selected for inspection. If X represents the number of selected trucks with defective brakes, then $X \sim HG(60, 8, 10)$.

Probability Distribution: We begin by noticing that the range of values for X is somewhat complicated. Of course, it is obvious that $X \geq 0$. However, if the size of the sample (i.e., the value of n) happens to exceed the number of failure-type objects in the population (i.e., the value of $N-r$), then their difference, namely $n-(N-r)$, must necessarily be of success type. In other words, $X \geq \max(0, n-N+r)$. On the other hand, we have that $X \leq r$ as it is impossible to select more success-type objects than the total number available. But $X \leq n$, since we cannot get more successes than the number of objects sampled. As a result, $X \leq \min(r, n)$. Therefore, we ultimately arrive at $A = \{\max(0, n-N+r), \ldots, \min(r, n)\}$.

If we choose not to consider order among the selected objects, we note that there are $\binom{N}{n}$ outcomes in the sample space corresponding to this experiment. For $x \in A$, there are $\binom{r}{x}$ ways to choose the x

success-type objects from the r available and $\binom{N-r}{n-x}$ ways to choose the remaining $n-x$ failure-type objects from the $N-r$ available. Therefore, this leads to the following form for the pmf of X:

$$f(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} \text{ for } x = \max(0, n-N+r), \ldots, \min(r, n).$$

Remarks:

(1) If we adopt the standard mathematical convention that $\binom{n}{k} = 0$ for non-negative integers n and k such that $n < k$, then it is a straightforward exercise to verify that $\sum\limits_{x \in A} f(x) = 1$ via the use of the well-known *hypergeometric identity*

$$\sum_{x=0}^{m} \binom{a}{x}\binom{b}{m-x} = \binom{a+b}{m} \text{ for non-negative integers } a, b, \text{ and } m. \tag{3.5.2}$$

(2) Just as with the binomial distribution, a closed-form expression for the cdf of a hypergeometric distribution does not exist. Therefore, one must once again sum probabilities term by term in order to calculate cumulative probabilities.

Turning our attention to the mean and variance of X, we have that

$$\begin{aligned}
E(X) &= \sum_{x \in A} x \, \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} \\
&= \frac{1}{\binom{N}{n}} \sum_{x=1}^{n} x \, \frac{r^{(x)}}{x!} \cdot \frac{(N-r)^{(n-x)}}{(n-x)!} \\
&= \frac{1}{\binom{N}{n}} \sum_{x=1}^{n} \frac{r(r-1)^{(x-1)}}{(x-1)!} \cdot \frac{(N-r)^{(n-x)}}{(n-x)!} \\
&= \frac{r}{\binom{N}{n}} \sum_{y=0}^{n-1} \frac{(r-1)^{(y)}}{y!} \cdot \frac{(N-r)^{(n-1-y)}}{(n-1-y)!} \text{ if we let } y = x-1 \\
&= \frac{r}{\binom{N}{n}} \sum_{y=0}^{n-1} \binom{r-1}{y}\binom{N-r}{n-1-y} \\
&= \frac{r}{\binom{N}{n}} \binom{N-1}{n-1} \text{ using the result of (3.5.2)} \\
&= \frac{rn!(N-n)!}{N!} \cdot \frac{(N-1)!}{(n-1)!(N-n)!} \\
&= \frac{nr}{N}
\end{aligned}$$

and

$$\begin{aligned}
Var(X) &= E(X(X-1)) + E(X) - (E(X))^2 \\
&= \sum_{x\in A} x(x-1)\,\frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} + \frac{nr}{N} - \left(\frac{nr}{N}\right)^2 \\
&= \frac{1}{\binom{N}{n}} \sum_{x=2}^{n} x(x-1)\,\frac{r^{(x)}}{x!}\cdot\frac{(N-r)^{(n-x)}}{(n-x)!} + \frac{nr}{N}\left(1-\frac{nr}{N}\right) \\
&= \frac{1}{\binom{N}{n}} \sum_{x=2}^{n} \frac{r(r-1)(r-2)^{(x-2)}}{(x-2)!}\cdot\frac{(N-r)^{(n-x)}}{(n-x)!} + \frac{nr(N-nr)}{N^2} \\
&= \frac{r(r-1)}{\binom{N}{n}} \sum_{y=0}^{n-2} \frac{(r-2)^{(y)}}{y!}\cdot\frac{(N-r)^{(n-2-y)}}{(n-2-y)!} + \frac{nr(N-nr)}{N^2} \text{ if we let } y = x-2 \\
&= \frac{r(r-1)}{\binom{N}{n}} \sum_{y=0}^{n-2} \binom{r-2}{y}\binom{N-r}{n-2-y} + \frac{nr(N-nr)}{N^2} \\
&= \frac{r(r-1)}{\binom{N}{n}} \binom{N-2}{n-2} + \frac{nr(N-nr)}{N^2} \text{ using the result of (3.5.2)} \\
&= \frac{r(r-1)n!(N-n)!}{N!}\cdot\frac{(N-2)!}{(n-2)!(N-n)!} + \frac{nr(N-nr)}{N^2} \\
&= \frac{r(r-1)n(n-1)}{N(N-1)} + \frac{nr(N-nr)}{N^2} \\
&= \frac{nr}{N^2(N-1)}\Big(N(r-1)(n-1) + (N-1)(N-nr)\Big) \\
&= \frac{nr}{N^2(N-1)}\Big(Nrn - Nr - Nn + N + N^2 - Nrn - N + nr\Big) \\
&= \frac{nr(N-r)(N-n)}{N^2(N-1)}.
\end{aligned}$$

Comparison of the Binomial and Hypergeometric Distributions: These distributions are similar in that an experiment with 2 types of outcomes (i.e., success and failure) is repeated n times, with X representing the number of successes obtained. The key difference is that the binomial distribution requires independent repetitions with the same probability of success on each trial, whereas the draws associated with the hypergeometric distribution are made from a fixed collection of objects **without replacement**. These trials (i.e., draws) are therefore not independent. For example, if there are 10 success-type objects and 10 failure-type objects in a population, then the probability of getting a success-type object on draw 2 depends on what was obtained in draw 1. If these draws had been made **with replacement**, however, they would be independent and we would use the binomial distribution rather than the hypergeometric model.

If N is large and n, the number of objects being drawn, is relatively small in the hypergeometric setup, then we are unlikely to get the same object more than once even if we do replace it. In other words, it makes little practical difference whether we draw with or without replacement. This suggests that when we are drawing a fairly small proportion of a large collection of objects, the binomial and hypergeometric probability models should produce similar probabilities. Since the pmf of the binomial distribution is generally easier to calculate, it is often used as an approximation to the hypergeometric distribution in such cases. The following example demonstrates this point.

Example 3.5.1. Consider 15 cans of soup having no labels on them. Suppose that 6 of the cans are tomato and the other 9 are mushroom. We randomly select 8 cans and open them. Find the probability that 3 of the opened cans are tomato. How would this probability change if instead we had 1500 total cans of soup of which 600 were tomato and 900 were mushroom?

Solution: Let X be the number of opened cans which are tomato. The correct model to use here is the hypergeometric distribution. In the first scenario, since $X \sim HG(15, 6, 8)$, it follows that

$$P(X = 3) = \frac{\binom{6}{3}\binom{9}{5}}{\binom{15}{8}} = 0.391608.$$

Note that if we incorrectly assume that $X \sim Bin(8, \frac{6}{15})$, we would get

$$P(X = 3) = \binom{8}{3}\left(\frac{6}{15}\right)^3\left(\frac{9}{15}\right)^5 = 0.278692.$$

As expected, this is a poor approximation since we are picking over half of a fairly small collection of cans. However, in the much larger second scenario, we note that we are not likely to select the same can again, even if we did replace each of the 8 cans after opening it. In other words, the probability that we get a can of tomato soup on each selection is very close to $\frac{600}{1500} = 0.4$, regardless of what the other picks give. Note that the exact hypergeometric probability now becomes

$$P(X = 3) = \frac{\binom{600}{3}\binom{900}{5}}{\binom{1500}{8}} = 0.279407,$$

whereas the binomial probability remains exactly the same, namely

$$\binom{8}{3}\left(\frac{600}{1500}\right)^3\left(\frac{900}{1500}\right)^5 = 0.278692.$$

Comparing values now, we see that the binomial distribution provides a very good approximation in this case. ■

3.5.4 Geometric Distribution

Physical Setup: Similar to the setup for the binomial distribution, consider an experiment having two distinct types of outcomes labelled s (for success) and f (for failure). Suppose that the experiment is repeated independently with the same (positive) probability, p, of success on each trial. Continue to perform the experiment until the very first success occurs. Let X be the number of failures observed before the first success occurs. In this situation, X is said to have a geometric distribution. We write $X \sim Geo(p)$ as a shorthand for "X is distributed according to a geometric distribution with probability p of success".

Illustrations:

(1) If the probability you win a lottery prize in any given week is 0.05 and X counts the number of weeks before you win a prize for the first time, then $X \sim Geo(0.05)$.

(2) If you take a driving test until you pass it and attempts are independent of each other with the same probability of 0.75 of passing each time, then the number of failed attempts would have a $Geo(0.75)$ distribution.

Probability Distribution: First of all, we note that the range of possible values X can take on is given by the countably infinite set $A = \{0, 1, 2, \ldots\}$, since we do not know in advance how many trials will be needed in order to observe the first success. Moreover, there is only the one arrangement to consider here, namely the one with x f's followed by the single s. Since trials are independent of each other, we immediately obtain the following pmf of X:

$$f(x) = (1-p)^x p \text{ for } x = 0, 1, 2, \ldots. \tag{3.5.3}$$

Note that if $p = 1$, then $f(0) = 1$ and $f(x) = 0$ for $x = 1, 2, 3, \ldots$ (i.e., the random variable X is simply the constant 0). If $p \in (0, 1)$, then we obtain the expected result that

$$\sum_{x=0}^{\infty} f(x) = p \underbrace{\sum_{x=0}^{\infty} (1-p)^x}_{\text{infinite geometric series}} = p \, \frac{1}{1-(1-p)} = \frac{p}{p} = 1.$$

Given the rather simple form of the above pmf, it is not surprising that we can also determine the cdf of X. In particular, if $x \in A$ and $p \in (0, 1)$, note that

$$P(X \leq x) = \sum_{u \leq x} f(u) = p \underbrace{\sum_{u=0}^{x} (1-p)^u}_{\text{finite geometric series}} = p \, \frac{1-(1-p)^{x+1}}{1-(1-p)} = 1 - (1-p)^{x+1}.$$

Therefore, the "full" cdf of X can be expressed as

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

To obtain the mean and variance of X when $p \in (0,1)$, we first find the mgf of X as follows:

$$M(t) = \sum_{x=0}^{\infty} e^{tx}\,(1-p)^x p = p \underbrace{\sum_{x=0}^{\infty} \left((1-p)e^t\right)^x}_{\text{infinite geometric series}} = \frac{p}{1-(1-p)e^t},$$

provided that $(1-p)e^t < 1$, or equivalently, $t < \ln(1-p)^{-1}$. Taking the first two derivatives of $M(t)$ yields

$$M^{(1)}(t) = p(1-p)e^t\left(1-(1-p)e^t\right)^{-2}$$

and

$$M^{(2)}(t) = 2p(1-p)^2e^{2t}\left(1-(1-p)e^t\right)^{-3} + p(1-p)e^t\left(1-(1-p)e^t\right)^{-2}.$$

Therefore, we readily obtain

$$E(X) = M^{(1)}(0) = p(1-p)e^0\left(1-(1-p)e^0\right)^{-2} = p(1-p)p^{-2} = \frac{1-p}{p}$$

and

$$\begin{aligned} Var(X) &= E(X^2) - (E(X))^2 \\ &= M^{(2)}(0) - \left(\frac{1-p}{p}\right)^2 \\ &= 2p(1-p)^2e^0\left(1-(1-p)e^0\right)^{-3} + p(1-p)e^0\left(1-(1-p)e^0\right)^{-2} - \frac{(1-p)^2}{p^2} \\ &= 2p(1-p)^2p^{-3} + p(1-p)p^{-2} - \frac{(1-p)^2}{p^2} \\ &= \frac{2(1-p)^2 + p(1-p) - (1-p)^2}{p^2} \\ &= \frac{1-p}{p^2}. \end{aligned}$$

3.5.5 Negative Binomial Distribution

Physical Setup: The setup for this distribution is the exact same as that described for the geometric distribution. The only difference is that now we will continue to conduct Bernoulli trials until k successes have been obtained, where k is a pre-specified positive integer. Let X be the number of failures observed before the k^{th} success occurs. In this situation, X is said to have a negative binomial

distribution. We write $X \sim NB(k, p)$ as a shorthand for "X is distributed according to a negative binomial distribution with parameters k and p". Keep in mind that k is the parameter representing the number of successes that occur before the experiment is stopped and p is the parameter denoting the probability of success on a single trial.

Illustrations:

(1) If a fair coin is tossed until we get the sixth head and X is the number of tails obtained, then $X \sim NB(6, \frac{1}{2})$.

(2) Suppose that a telemarketer has a 15% chance of making a sale on any given phone call. If phone calls are independent of each other and the telemarketer is required to make 10 successful sales before leaving for the day, then the number of unsuccessful calls made has a $NB(10, 0.15)$ distribution.

Probability Distribution: Just as in the case of the geometric distribution, the range of possible values for the number of observed failures before the k^{th} success occurs is given by the countably infinite set $A = \{0, 1, 2, \ldots\}$. For $x \in A$, there will be, in all, $x + k$ conducted trials (i.e., x f's and k s's) and the very last trial must be a success. Therefore, among the first $x + k - 1$ trials, we require x f's and $(k - 1)$ s's, occurring in any order. The number of different arrangements of x f's and $(k - 1)$ s's over the first $x + k - 1$ trials is given by

$$\binom{x+k-1}{x} \times \binom{k-1}{k-1} = \binom{x+k-1}{x}.$$

Each arrangement will have probability $p^k(1-p)^x$ since the trials are independent and there must be x trials which are failures and k which are successes (including the very last trial). Thus, the pmf of X is given by

$$f(x) = \binom{x+k-1}{x} p^k(1-p)^x \text{ for } x = 0, 1, 2, \ldots. \tag{3.5.4}$$

Remarks:

(1) As a special case, if we substitute $k = 1$ into (3.5.4), note that we immediately retrieve the pmf of the geometric distribution given by (3.5.3). Unlike the geometric distribution, however, a closed-form expression for the cdf of a negative binomial distribution does not exist. As such, we need to sum probabilities term by term in order to calculate cumulative negative binomial probabilities.

(2) With the aid of the *extended binomial series formula* given by

$$\sum_{x=0}^{\infty} \binom{\alpha}{x} a^x = (1+a)^\alpha \text{ for } \alpha \in \mathbb{R} \text{ and } |a| < 1, \tag{3.5.5}$$

in which the term $\binom{\alpha}{x}$ is calculated as

$$\binom{\alpha}{x} = \frac{\alpha^{(x)}}{x!}$$

for non-integer values of α, it is possible to show that $\sum_{x=0}^{\infty} f(x) = 1$.

(3) An alternate version of the negative binomial distribution defines the random variable of interest to be the *total number of trials* (not the number of failures) needed to get the k^{th} success. However, this is equivalent to our version of the negative binomial distribution. For example, asking for the probability of getting 3 tails before the 5^{th} head is exactly the same as asking for a total of 8 tosses in order to get the 5^{th} head.

(4) When distinguishing between the binomial and negative binomial distributions, it is useful to keep in mind that they essentially reverse what is specified or known in advance and what is variable. In the case of the binomial distribution, we know the number n of trials in advance but do not know the number of successes we will obtain until after the experiment. Looking at the negative binomial distribution, we know the number k of successes in advance but do not know the number of trials that will be needed to obtain this number of successes until after the experiment.

With the pmf of X given by (3.5.4), we can derive the mgf of X as follows:

$$\begin{aligned}
M(t) &= \sum_{x=0}^{\infty} e^{tx} \binom{x+k-1}{x} p^k (1-p)^x \\
&= p^k \sum_{x=0}^{\infty} \binom{x+k-1}{x} \left((1-p)e^t\right)^x \\
&= p^k \sum_{x=0}^{\infty} \binom{x+k-1}{x} \left(1 - \left(1 - (1-p)e^t\right)\right)^x \\
&= \frac{p^k}{\left(1-(1-p)e^t\right)^k} \sum_{x=0}^{\infty} \binom{x+k-1}{x} \left(1-(1-p)e^t\right)^k \left(1 - \left(1-(1-p)e^t\right)\right)^x \\
&= \left(\frac{p}{p^*}\right)^k \sum_{x=0}^{\infty} \binom{x+k-1}{x} p^{*k} (1-p^*)^x \text{ where we define } p^* = 1-(1-p)e^t. \qquad (3.5.6)
\end{aligned}$$

Note that in (3.5.6) we are summing a $NB(k, p^*)$ pmf provided that the condition $0 < p^* < 1$ is satisfied. However, it is straightforward to show that this condition is equivalent to requiring $t < \ln(1-p)^{-1}$, in which case the sum in (3.5.6) is equal to 1 and we simply obtain

$$M(t) = \left(\frac{p}{1-(1-p)e^t}\right)^k \text{ for } t < \ln(1-p)^{-1}.$$

Taking the first two derivatives of $M(t)$ yields

$$M^{(1)}(t) = kp^k(1-p)e^t\left(1-(1-p)e^t\right)^{-(k+1)}$$

and

$$M^{(2)}(t) = k(k+1)p^k(1-p)^2e^{2t}\left(1-(1-p)e^t\right)^{-(k+2)} + kp^k(1-p)e^t\left(1-(1-p)e^t\right)^{-(k+1)}.$$

Thus, the mean and variance of X are given by

$$E(X) = M^{(1)}(0) = kp^k(1-p)e^0\left(1-(1-p)e^0\right)^{-(k+1)} = kp^k(1-p)p^{-(k+1)} = \frac{k(1-p)}{p}$$

and

$$\begin{aligned}
Var(X) &= E(X^2) - (E(X))^2 \\
&= M^{(2)}(0) - \left(\frac{k(1-p)}{p}\right)^2 \\
&= k(k+1)p^k(1-p)^2e^0\left(1-(1-p)e^0\right)^{-(k+2)} + kp^k(1-p)e^0\left(1-(1-p)e^0\right)^{-(k+1)} - \frac{k^2(1-p)^2}{p^2} \\
&= k(k+1)p^k(1-p)^2p^{-(k+2)} + kp^k(1-p)p^{-(k+1)} - \frac{k^2(1-p)^2}{p^2} \\
&= \frac{k(k+1)(1-p)^2}{p^2} + \frac{k(1-p)}{p} - \frac{k^2(1-p)^2}{p^2} \\
&= \frac{k(1-p)^2}{p^2} + \frac{k(1-p)}{p} \\
&= \frac{k(1-p)}{p^2}.
\end{aligned}$$

3.5.6 Poisson Distribution

The Poisson distribution arises as a limiting case of the binomial distribution as $n \to \infty$ and $p \to 0$. In particular, we keep the product np (which happens to be the mean of the binomial distribution) fixed at some constant positive value (call it μ) while letting $n \to \infty$. This automatically makes $p = \frac{\mu}{n} \to 0$.

To see what the limit of the binomial pmf becomes in this case, suppose that $x \in \{0, 1, \ldots, n\}$ is

fixed and consider

$$\begin{aligned}
\binom{n}{x}p^x(1-p)^{n-x} &= \frac{n^{(x)}}{x!}\left(\frac{\mu}{n}\right)^x\left(1-\frac{\mu}{n}\right)^{n-x} \\
&= \frac{\mu^x}{x!}\frac{\overbrace{n(n-1)(n-2)\cdots(n-x+1)}^{x \text{ terms}}}{\underbrace{(n)(n)(n)\cdots(n)}_{x \text{ terms}}}\left(1-\frac{\mu}{n}\right)^{n-x} \\
&= \frac{\mu^x}{x!}\left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\left(\frac{n-2}{n}\right)\cdots\left(\frac{n-x+1}{n}\right)\left(1-\frac{\mu}{n}\right)^{n}\left(1-\frac{\mu}{n}\right)^{-x} \\
&= \frac{\mu^x}{x!}(1)\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{x-1}{n}\right)\left(1-\frac{\mu}{n}\right)^{n}\left(1-\frac{\mu}{n}\right)^{-x}. \qquad (3.5.7)
\end{aligned}$$

Taking the limit of (3.5.7) as $n \to \infty$ yields

$$\begin{aligned}
\lim_{n\to\infty}\binom{n}{x}p^x(1-p)^{n-x} &= \frac{\mu^x}{x!}\underbrace{(1)(1)(1)\cdots(1)}_{x \text{ terms}}e^{-\mu}(1)^{-x} \text{ since } e^z = \lim_{n\to\infty}\left(1+\frac{z}{n}\right)^n \text{ for } z \in \mathbb{R} \\
&= \frac{\mu^x e^{-\mu}}{x!}.
\end{aligned}$$

Since the upper limit on x for the binomial pmf is n and we are letting $n \to \infty$, the above limit is defined for $x \in \{0, 1, 2, \ldots\}$.

Therefore, using the above result as motivation, let us now formally introduce a discrete random variable X having pmf of the form

$$f(x) = \frac{\mu^x e^{-\mu}}{x!} \text{ for } x = 0, 1, 2, \ldots, \qquad (3.5.8)$$

where $\mu > 0$. Mathematically, we can see that $f(x)$ has the properties of a pmf, since $f(x) \geq 0$ for $x = 0, 1, 2, \ldots$ and

$$\sum_{x=0}^{\infty} f(x) = e^{-\mu}\underbrace{\sum_{x=0}^{\infty}\frac{\mu^x}{x!}}_{\text{exponential power series}} = e^{-\mu}\cdot e^{\mu} = 1.$$

We say that a random variable X has a Poisson distribution if its pmf is given by (3.5.8). We write $X \sim Poi(\mu)$ as a shorthand for "X is distributed according to a Poisson distribution with parameter μ". In particular, we can use the Poisson distribution with $\mu = np$ as a close approximation to the binomial distribution in situations when n is large and p is small.

Example 3.5.2. Suppose that 200 guests are in attendance at a wedding ceremony. What is the probability that exactly two of the guests were born on January 1, assuming all days of the year (excluding February 29) are equally likely for a birthday?

Solution: Let X be the random variable representing the number of guests born on January 1. Assuming all 365 days of the year are equally and that the birthdays of guests are independent of each other, we can use the binomial distribution with $n = 200$ and $p = \frac{1}{365}$ to get

$$P(X = 2) = \binom{200}{2}\left(\frac{1}{365}\right)^2\left(1 - \frac{1}{365}\right)^{198} = 0.086767.$$

Since n is large and p is close to 0, we can use the Poisson distribution to approximate this binomial probability. With $\mu = np = \frac{200}{365}$, we would obtain

$$P(X = 2) = \frac{\left(\frac{200}{365}\right)^2 e^{-\left(\frac{200}{365}\right)}}{2!} = 0.086791.$$

As we might expect, this is a very good approximation. ■

Remark: If instead, the value of p is close to 1, then we can also use the Poisson distribution to approximate the binomial distribution. By interchanging the labels "success" and "failure", we can get the probability of success (formerly labelled failure) close to 0.

Like the binomial distribution, a closed-form expression for the cdf of a $Poi(\mu)$ random variable does not exist. However, its mgf does exist and is given by

$$\begin{aligned} M(t) &= \sum_{x=0}^{\infty} e^{tx}\,\frac{\mu^x e^{-\mu}}{x!} \\ &= e^{-\mu}\underbrace{\sum_{x=0}^{\infty}\frac{(\mu e^t)^x}{x!}}_{\text{exponential power series}} \\ &= e^{-\mu}\cdot e^{\mu e^t} \\ &= e^{\mu(e^t-1)} \text{ for } t \in \mathbb{R}. \end{aligned}$$

Since the first two derivatives of $M(t)$ are given by

$$M^{(1)}(t) = \mu e^t e^{\mu(e^t-1)}$$

and

$$M^{(2)}(t) = \mu^2 e^{2t} e^{\mu(e^t-1)} + \mu e^t e^{\mu(e^t-1)},$$

the mean and variance of X can be easily obtained as

$$E(X) = M^{(1)}(0) = \mu e^0 e^{\mu(e^0-1)} = \mu$$

and

$$\begin{aligned} Var(X) &= E(X^2) - (E(X))^2 \\ &= M^{(2)}(0) - \mu^2 \\ &= \mu^2 e^0 e^{\mu(e^0-1)} + \mu e^0 e^{\mu(e^0-1)} - \mu^2 \\ &= \mu^2 + \mu - \mu^2 \\ &= \mu. \end{aligned}$$

Section 3.5 Problems

3.5.1 A computer random number generator is used to create a four-digit number using the set of digits $\{1, 2, \ldots, 9\}$. Digit selections are independent of one another, and each digit selected from this set has an equally likely chance of being chosen. Determine the pmf of the random variable X, representing the *smallest* digit in the four-digit number that is created.

3.5.2 A box of twelve tins of tuna contains d tins which are tainted. Suppose that seven tins are opened for inspection and none of these seven is tainted.

(a) Calculate the probability that none of the seven tins are tainted for $d = 0, 1, 2$, and 3.

(b) Do you think it is likely that the box contains as many as three tainted tins?

3.5.3 Megan audits 130 clients during a year and finds irregularities for 26 of them.

(a) What is the probability that two clients will have irregularities when six of her clients are picked at random?

(b) Evaluate your answer to part (a) using a suitable approximation.

3.5.4 The fraction of a large population that has specific blood type O^+ is 0.38. For blood donation purposes, it is necessary to find five people with type O^+ blood. If randomly selected individuals from the population are tested one after another, what is the probability that over ten people have to be tested?

3.5.5 Suppose that there is a 30% chance of a car from a certain production line having a leaky windshield. The probability that an inspector will have to check at least n cars to find the first one with a leaky windshield is 0.05. Determine the value of n.

3.5.6 An airline knows that 97% of the passengers who buy tickets for a certain flight will show up on time. The plane has 120 seats. Suppose that the airline sells 122 tickets. Find the probability

that more people will show up than can be carried on the flight. Compare this answer with the probability given by the Poisson approximation.

3.5.7 Suppose that $X \sim Geo(p)$.

(a) What is the probability that X is an odd number?

(b) What is the probability that X is divisible by 3?

(c) Find the pmf of the random variable $R = X \pmod 4$.

3.5.8 During jury selection, a large number of people are asked to be present, then persons are selected one by one in a random order until the required number of jurors has been chosen. Because the prosecution and defense teams can each reject a certain number of persons, and because some individuals may be exempted by the judge, the total number of persons selected before a full jury is found can be quite large.

(a) Suppose that you are one of 150 persons asked to be present for the selection of a jury. If it is necessary to select 40 persons in order to form the jury, what is the probability you are chosen?

(b) In a recent trial, the numbers of men and women present for jury selection were 74 and 76, respectively. If Y is the number of men picked for a jury of 12 persons, determine the pmf of Y, assuming that men and women are equally likely to be picked.

(c) For the trial in part (b), the number of men selected turned out to be two. Calculate $P(Y \leq 2)$. What might you conclude from this?

3.5.9 An oil company runs a contest in which there are 500,000 tickets. A motorist receives one ticket with each fill-up of gasoline, and 500 of the tickets are winners.

(a) If a motorist has ten fill-ups during the contest, what is the probability that he or she wins at least one prize?

(b) If a particular gas bar distributes 2,000 tickets during the contest, give an expression for the probability that there is at least one winner among the gas bar's customers. Use two different approximations to approximate this probability.

3.5.10 Let $X \sim NB(k, p)$, representing the number of *failures* obtained before achieving the k^{th} success in a sequence of independent Bernoulli trials. If Y represents the total number of *trials* needed to achieve k successes, determine its pmf, mgf, mean, and variance.

Chapter 4

Multivariate Discrete Probability Distributions

4.1 Basic Terminology and Techniques

Many real-life applications often involve more than a single random variable. For example, your final mark in a course might involve your assignment mark X_1, your midterm test mark X_2, and your final exam mark X_3. In a medical diagnosis context, the results of various tests $Y_1, Y_2, \ldots, Y_m$ may be significant. In a networking context, the workloads of several routers $Z_1, Z_2, \ldots, Z_n$ may be of interest. All of these random variables are associated with the same experiment and sample space, and as such, their values may relate in interesting ways. This motivates us to consider probabilities of events involving simultaneously several random variables. Therefore, we need to extend the ideas introduced in the previous chapter for univariate discrete probability distributions to deal with multivariate problems.

Joint Probability Mass Functions: Let us consider two discrete random variables X and Y associated with the same experiment, and define the function

$$f(x, y) \quad = \quad P(\{X = x\} \cap \{Y = y\}),\ x \in A_X, y \in A_Y,$$

where A_X and A_Y represent the ranges of X and Y, respectively. We call $f(x, y)$ the **joint probability mass function (pmf) of** (X, Y) and write

$$f(x, y) = P(X = x, Y = y)$$

as a shorthand for this function. The properties of a joint pmf are similar to those for a single discrete

random variable. In particular, we have that $0 \leq f(x, y) \leq 1$ for all (x, y) and

$$\sum_{all\ (x,y)} f(x, y) = 1.$$

More generally, for a collection of n discrete random variables $X_1, X_2, \ldots, X_n$, the joint pmf of $(X_1, X_2, \ldots, X_n)$ is defined as

$$f(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n).$$

As a simple demonstration, consider the following table summarizing the values of $f(x, y)$ for two discrete random variables X and Y:

		x		
$f(x,y)$		0	1	2
	1	0.06	0.18	0.35
y				
	2	0.15	0.14	0.12

From the table, for example, we can easily read off that $f(1, 2) = P(X = 1, Y = 2) = 0.14$. Moreover, it is not difficult to see that $f(x, y)$ is a proper joint pmf since $0 \leq f(x, y) \leq 1$ for all six combinations of (x, y) and the sum of these six probabilities equals 1. When there are only a few values for X and Y (as is the case here), it is often easier to tabulate $f(x, y)$ than to find a formula for it. The next example also demonstrates this point as well.

Example 4.1.1. Suppose a fair coin is tossed 3 times. If we define X to be the number of heads obtained and Y to *indicate* whether heads or tails occurs on the first toss (i.e., $Y = 1$ if heads occurs and $Y = 0$ if tails occurs), determine the joint pmf of (X, Y).

Solution: First of all, we should note the range for (X, Y), which is the set of possible values (x, y) which can occur. Although X can clearly be 0, 1, 2, or 3 and Y can be 0 or 1, not all eight combinations of (x, y) have a positive probability. For instance, the ordered pair $(0, 1)$ is impossible to obtain (i.e., we cannot have a situation in which 3 tails were obtained simultaneous to getting heads on the first toss). We can find $f(x, y)$ by simply writing out the sample space for this experiment, namely

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

which we have used previously for this experiment. Simple counting then leads to $f(x, y)$ as given in the following table:

		x			
$f(x,y)$		0	1	2	3
	0	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	0
y					
	1	0	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$

Note that $(X, Y) = (0, 0)$ if and only if the outcome is TTT, $(X, Y) = (1, 0)$ if and only if the outcome is either THT or TTH, and so on. ■

The joint pmf can be used to determine the probability of any event that can be specified in terms of the random variables X and Y. For example, suppose that B is a collection of ordered pairs taken from the range of (X, Y) that have a certain property. It immediately follows that

$$P((X, Y) \in B) = \sum_{(x,y) \in B} f(x, y).$$

Example 4.1.2. Consider the earlier table summarizing the values of $f(x, y)$ for two discrete random variables X and Y:

		x		
$f(x,y)$		0	1	2
	1	0.06	0.18	0.35
y				
	2	0.15	0.14	0.12

Calculate the following probabilities: (i) $P(X + Y \leq 3)$, (ii) $P(XY > 0)$, and (iii) $P(Y > X)$.

Solution: Note that each of the 3 events can be represented in terms of the points (x, y) taken from the range of (X, Y):

$$\begin{aligned} \{X + Y \leq 3\} &= \{(0, 1), (1, 1), (2, 1), (0, 2), (1, 2)\}, \\ \{XY > 0\} &= \{(1, 1), (2, 1), (1, 2), (2, 2)\}, \\ \{Y > X\} &= \{(0, 1), (0, 2), (1, 2)\}. \end{aligned}$$

As a result, we immediately obtain

$$\begin{aligned} P(X + Y \leq 3) &= 0.06 + 0.18 + 0.35 + 0.15 + 0.14 = 0.88, \\ P(XY > 0) &= 0.18 + 0.35 + 0.14 + 0.12 = 0.79, \\ P(Y > X) &= 0.06 + 0.15 + 0.14 = 0.35. \end{aligned}$$

■

Marginal Distributions: We may be given a joint pmf involving more random variables than we are actually interested in using. The natural question becomes: *How can we eliminate from the joint distribution any random variables which are not of interest to us?* If we consider Example 4.1.2 and

suppose that we are only interested in the random variable X (i.e., we do not care what value Y takes on), then we see that

$$\begin{aligned} P(X = 0) &= P(X = 0, Y = 1) + P(X = 0, Y = 2) \\ &= f(0, 1) + f(0, 2) \\ &= 0.21. \end{aligned}$$

Similarly,

$$\begin{aligned} P(X = 1) &= f(1, 1) + f(1, 2) = 0.32 \\ \text{and} \quad P(X = 2) &= f(2, 1) + f(2, 2) = 0.47. \end{aligned}$$

The probability distribution of X obtained in this fashion from the joint distribution of (X, Y) is referred to as the **marginal pmf of** X, and is given by

x	0	1	2
$f_X(x)$	0.21	0.32	0.47

In the same way, if we are only interested in the random variable Y, we would obtain

$$P(Y = 1) = f(0, 1) + f(1, 1) + f(2, 1) = 0.59,$$

since X can be 0, 1, or 2 when $Y = 1$. A similar calculation would hold for $P(Y = 2)$, and so the **marginal pmf of** Y would be given by

y	1	2
$f_Y(y)$	0.59	0.41

In general, to find $f_X(x)$, we add over all values of y where $X = x$, and to find $f_Y(y)$ we add over all values of x with $Y = y$. This leads to

$$\begin{aligned} f_X(x) &= \sum_{all\ y} f(x, y) \\ \text{and} \quad f_Y(y) &= \sum_{all\ x} f(x, y). \end{aligned}$$

Example 4.1.3. Suppose that X and Y have joint pmf given by

$$f(x, y) = \frac{1}{6}\left(\frac{1}{2}\right)^x \left(\frac{2}{3}\right)^y \quad \text{for } x = 0, 1, 2, \ldots \text{ and } y = 0, 1, 2 \ldots.$$

Find the marginal pmf of X and the marginal pmf of Y.

Solution: The marginal pmf of X is given by

$$\begin{aligned} f_X(x) &= \sum_{all\ y} f(x,y) \\ &= \sum_{y=0}^{\infty} \frac{1}{6}\left(\frac{1}{2}\right)^x \left(\frac{2}{3}\right)^y \\ &= \frac{1}{6}\left(\frac{1}{2}\right)^x \sum_{y=0}^{\infty}\left(\frac{2}{3}\right)^y \\ &= \frac{1}{6}\left(\frac{1}{2}\right)^x \frac{1}{1-\frac{2}{3}} \\ &= \left(\frac{1}{2}\right)^{x+1} \quad \text{for } x = 0,1,2,\ldots. \end{aligned}$$

Similarly, the marginal pmf of Y is given by

$$\begin{aligned} f_Y(y) &= \sum_{all\ x} f(x,y) \\ &= \sum_{x=0}^{\infty} \frac{1}{6}\left(\frac{1}{2}\right)^x \left(\frac{2}{3}\right)^y \\ &= \frac{1}{6}\left(\frac{2}{3}\right)^y \sum_{x=0}^{\infty}\left(\frac{1}{2}\right)^x \\ &= \frac{1}{6}\left(\frac{2}{3}\right)^y \frac{1}{1-\frac{1}{2}} \\ &= \left(\frac{1}{3}\right)\left(\frac{2}{3}\right)^y \quad \text{for } y = 0,1,2,\ldots. \end{aligned}$$

In looking at the results we derived, note that $X \sim Geo(\frac{1}{2})$ and $Y \sim Geo(\frac{1}{3})$. ■

Remark: The above reasoning can be extended beyond the case of two random variables. For example, in the case of three discrete random variables X_1, X_2, and X_3, we would have

$$f_1(x_1) = P(X_1 = x_1) = \sum_{all\ (x_2,x_3)} f(x_1,x_2,x_3)$$

$$\text{and} \quad f_{1,3}(x_1,x_3) = P(X_1 = x_1, X_3 = x_3) = \sum_{all\ x_2} f(x_1,x_2,x_3),$$

where $f_1(x_1)$ is the marginal pmf of X_1 and $f_{1,3}(x_1,x_3)$ denotes the (marginal) joint pmf of (X_1, X_3).

Independent Random Variables: Recall that two events A and B are said to be independent if and only if $P(A \cap B) = P(A)P(B)$. This definition can be extended to random variables X and Y in the following way.

Definition 4.1.1. *X and Y are **independent random variables** if and only if*

$$f(x, y) = f_X(x)f_Y(y) \text{ for all values } (x, y).$$

When n random variables $X_1, X_2, \ldots, X_n$ are under consideration, the above definition generalizes in a very natural way.

Definition 4.1.2. *$X_1, X_2, \ldots, X_n$ are **independent random variables** if and only if*

$$f(x_1, x_2, \ldots, x_n) = f_1(x_1)f_2(x_2)\cdots f_n(x_n) \quad \text{for all } (x_1, x_2, \ldots x_n),$$

where $f_i(x_i) = P(X_i = x_i)$ denotes the marginal pmf of X_i for $i = 1, 2, \ldots, n$.

In Example 4.1.2, we observe that X and Y are not independent since $f_X(x)f_Y(y) \neq f(x, y)$ for any of the six combinations of (x, y) values (e.g., $0.18 = f(1, 1) \neq f_X(1)f_Y(1) = (0.32)(0.59)$). On the other hand, in Example 4.1.3, note that

$$f_X(x)f_Y(y) = \left(\frac{1}{2}\right)^{x+1}\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)^{y} = \frac{1}{6}\left(\frac{1}{2}\right)^{x}\left(\frac{2}{3}\right)^{y} = f(x, y),$$

for each $x, y \in \{0, 1, 2, \ldots\}$. In this case, X and Y are independent random variables. In general, care should be taken applying the definition of independence. You can only conclude that X and Y are independent after checking <u>all</u> (x, y) combinations. Even a single case where $f_X(x)f_Y(y) \neq f(x, y)$ results in X and Y being *dependent* random variables.

Functions of Random Variables: We often encounter situations where we need to find the probability distribution of a function of two or more random variables. The most general method for finding the pmf for some function of random variables X and Y involves looking at every outcome (x, y) to see what value the function produces. For example, if we let $U = 2(Y - X)$ in Example 4.1.2, the possible values of u are seen by looking at the value of $U = 2(y - x)$ for each outcome (x, y) in the range of (X, Y). The following table summarizes the situation:

		x		
	u	0	1	2
	1	2	0	−2
y				
	2	4	2	0

Based on the values of u in this table, we readily obtain the following probabilities:

$$\begin{aligned} P(U=-2) &= f(2,1) = 0.35, \\ P(U=0) &= f(1,1) + f(2,2) = 0.18 + 0.12 = 0.3, \\ P(U=2) &= f(0,1) + f(1,2) = 0.06 + 0.14 = 0.2, \\ P(U=4) &= f(0,2) = 0.15. \end{aligned}$$

Therefore, the pmf of U readily follows:

u	-2	0	2	4
$f_U(u)$	0.35	0.3	0.2	0.15

For some functions, it is possible to approach the problem more systematically. One of the most common functions of this type is the total of the random variables. In particular, let $T = X + Y$. For Example 4.1.2, the values that the total T could take on is summarized below:

		x		
	t	0	1	2
	1	1	2	3
y				
	2	2	3	4

Then, for instance, $P(T=3) = f(1,2) + f(2,1) = 0.14 + 0.35 = 0.49$. Continuing in this fashion, we would ultimately obtain the pmf of T:

t	1	2	3	4
$f_T(t)$	0.06	0.33	0.49	0.12

In fact, to find $f_T(t) = P(T=t)$, we simply add the probabilities for all (x,y) combinations with $x+y=t$, which could be expressed as

$$P(T=t) = \sum_{\substack{\text{all } (x,y): \\ x+y=t}} f(x,y).$$

However, if $x+y=t$, then $y=t-x$. To systematically pick out the right combinations of (x,y), all we really need to do is sum the joint pmf of (X,Y) over all values of x with $t-x$ substituted for y. This leads to the following formula:

$$P(T=t) = \sum_{all\ x} f(x, t-x).$$

Therefore, using this formula, $P(T = 3)$ would be computed as

$$P(T = 3) = \sum_{x=0}^{2} f(x, 3 - x) = \underbrace{f(0, 3)}_{=0} + f(1, 2) + f(2, 1) = 0.49.$$

We can summarize the method of finding the pmf of a function $U = g(X, Y)$ of two random variables X and Y. If $f(x, y)$ represents the joint pmf of (X, Y), then the pmf of U is given by

$$f_U(u) = P(U = u) = \sum_{\substack{\text{all } (x,y): \\ g(x,y)=u}} f(x, y).$$

This can also be extended to functions of three or more random variables, such as $U = g(X_1, X_2, \ldots, X_n)$:

$$f_U(u) = P(U = u) = \sum_{\substack{\text{all } (x_1,x_2,\ldots,x_n): \\ g(x_1,x_2,\ldots,x_n)=u}} f(x_1, x_2, \ldots, x_n).$$

Example 4.1.4. Let X_1 and X_2 be independent random variables having Poisson distributions with means of μ_1 and μ_2, respectively. Determine the pmf of $T = X_1 + X_2$.

Solution: We first need to find $f(x_1, x_2)$. Since X_1 and X_2 are independent, we know that

$$f(x_1, x_2) = f_1(x_1) f_2(x_2) \text{ for all } (x_1, x_2).$$

Using the form of the Poisson pmf, we have

$$f(x_1, x_2) = \frac{\mu_1^{x_1} e^{-\mu_1}}{x_1!} \cdot \frac{\mu_2^{x_2} e^{-\mu_2}}{x_2!} \text{ for } x_1 = 0, 1, 2, \ldots \text{ and } x_2 = 0, 1, 2, \ldots.$$

Now, the pmf of T is given by

$$\begin{aligned}
f_T(t) &= P(X_1 + X_2 = t) \\
&= \sum_{\text{all } x_1} f(x_1, t - x_1) \\
&= \sum_{x_1=0}^{t} \frac{\mu_1^{x_1} e^{-\mu_1}}{x_1!} \cdot \frac{\mu_2^{t-x_1} e^{-\mu_2}}{(t-x_1)!} \\
&= \mu_2^t e^{-(\mu_1+\mu_2)} \sum_{x_1=0}^{t} \frac{1}{x_1!(t-x_1)!} \left(\frac{\mu_1}{\mu_2}\right)^{x_1} \\
&= \frac{\mu_2^t e^{-(\mu_1+\mu_2)}}{t!} \sum_{x_1=0}^{t} \frac{t!}{x_1!(t-x_1)!} \left(\frac{\mu_1}{\mu_2}\right)^{x_1} \\
&= \frac{\mu_2^t e^{-(\mu_1+\mu_2)}}{t!} \sum_{x_1=0}^{t} \binom{t}{x_1} \left(\frac{\mu_1}{\mu_2}\right)^{x_1} 1^{t-x_1} \\
&= \frac{\mu_2^t e^{-(\mu_1+\mu_2)}}{t!} \left(\frac{\mu_1}{\mu_2} + 1\right)^t \quad \text{by the binomial series formula} \\
&= \frac{(\mu_1+\mu_2)^t e^{-(\mu_1+\mu_2)}}{t!} \quad \text{for } t = 0, 1, 2, \ldots.
\end{aligned}$$

Looking at the form of above pmf, note that we have just shown that the sum of two independent Poisson random variables also has a Poisson distribution. ■

Section 4.1 Problems

4.1.1 Suppose that the joint pmf of (X, Y) is given by

		x	
$f(x,y)$	0	1	2
0	0.15	0.1	0.05
y			
1	0.35	0.2	0.15

(a) Find the marginal pmf of X and the marginal pmf of Y.

(b) Are X and Y independent random variables? Justify your response.

(c) Calculate $P(X > Y)$.

(d) Determine the pmf of $T = X + Y$.

4.1.2 The joint pmf of (X, Y) is given by

		x		
$f(x,y)$		1	2	3
	2	0.09	0.06	0.15
y	4	0.15	0.05	0.20
	6	0.06	0.09	0.15

(a) Calculate $P\left(X^2 < 2Y\right)$.

(b) Are X and Y independent random variables? Justify your response.

(c) Determine the pmf of $D = |X - Y|$.

4.1.3 Britney owns an ice cream company that makes a cherry ice cream containing chocolate chips and cherries. Suppose that X is the number of cherries in a random scoop of ice cream, and Y is the number of chocolate chips in a random scoop. Suppose that X and Y have joint pmf given by

		y		
$f(x,y)$		2	3	4
	1	0.2	0.15	0.1
x				
	2	0.15	0.15	0.25

(a) Calculate $P(X + Y \geq 5)$.

(b) Calculate the mean number of chocolate chips in a random scoop.

(c) Are X and Y independent random variables? Justify your response.

(d) If the production cost of a scoop of ice cream is 3 cents plus 5 cents for every cherry and 1 cent for every chocolate chip, calculate the expected production cost of a scoop.

4.1.4 Suppose that random variables X and Y have joint pmf of the form

$$f(x,y) = kxy^2 \text{ for } x = 1,2,3 \text{ and } y = 1,2,$$

where k is a positive constant that makes the above joint pmf valid.

(a) Find the value of k.

(b) Find marginal pmf of X and the marginal pmf of Y.

(c) Determine, and justify, whether or not X and Y are independent random variables.

4.1.5 In a quality control inspection, items are classified as having a minor defect, a major defect, or as being acceptable. A carton of 10 items contains 2 with a minor defect, 1 with a major defect, and 7 that are acceptable. Three items are chosen at random without replacement. Let X be the number selected with a minor defect and let Y be the number selected with a major defect.

(a) Find the joint pmf of X and Y.

(b) Determine the marginal pmf of X and the marginal pmf of Y.

(c) Calculate $P(X = Y)$.

4.1.6 Let X and Y be discrete random variables with joint pmf

$$f(x, y) = \frac{e^{-2}}{x!(y-x)!} \text{ for } y = 0, 1, 2, \ldots \text{ and } x = 0, 1, \ldots, y.$$

(a) Find the marginal pmf of X and the marginal pmf of Y.

(b) Are X and Y independent random variables? Justify your response.

4.1.7 If X and Y are independent random variables such that $X \sim Geo(p)$ and $Y \sim Geo(p)$, determine the probability distribution of $T = X + Y$.

4.1.8 Consider an experiment in which a fair coin is independently flipped m times, j times by player A and $m - j$ times by player B. Show that the probability that players A and B flip the same number of heads is equal to the probability that there are a total of j heads flipped.

4.2 Multinomial Distribution

Unlike our treatment of various special univariate discrete probability distributions of interest, we only consider one special multivariate probability distribution in this course, although other multivariate distributions of interest do exist. The multinomial distribution defined below is a very important one. It is a generalization of the binomial distribution introduced in Section 3.5.2 to the situation where each trial now has k possible outcomes instead of just two.

In terms of its physical setup, the multinomial distribution has the same essential setup as the binomial distribution. An experiment is repeated independently n times. On each repetition, the experiment results in only one of k possible mutually exclusive and exhaustive outcomes. Let p_i be the probability that the outcome is of type i, $i = 1, 2, \ldots, k$, and we assume that each p_i remains constant throughout the n independent repetitions. By construction, we also have $p_1 + p_2 + \cdots + p_k = 1$. For

$i = 1, 2, \ldots, k$, let X_i be the number of times the i^{th} type of outcome occurs. In this case, we say that $(X_1, X_2, \ldots, X_k)$ has a **multinomial distribution**.

To determine the joint pmf of $(X_1, X_2, \ldots, X_k)$ given by

$$f(x_1, x_2, \ldots, x_k) = P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k),$$

we first take note of the existence of the condition

$$x_1 + x_2 + \cdots + x_k = n,$$

or equivalently,

$$x_k = n - x_1 - x_2 - \cdots - x_{k-1}.$$

In other words, the number of times each type of outcome is observed must necessarily be equal to the number of repetitions run. With this condition in mind, the number of ways of arranging x_1 type-1 outcomes, x_2 type-2 outcomes, $\ldots$, x_k type-k outcomes over a sequence of n outcomes in total is given by

$$\begin{aligned}
& \binom{n}{x_1} \times \binom{n-x_1}{x_2} \times \cdots \times \binom{n-x_1-x_2-\cdots-x_{k-2}}{x_{k-1}} \times \binom{n-x_1-x_2-\cdots-x_{k-1}}{x_k} \\
= \; & \frac{n!}{x_1!(n-x_1)!} \cdot \frac{(n-x_1)!}{x_2!(n-x_1-x_2)!} \cdots \frac{(n-x_1-x_2-\cdots-x_{k-2})!}{x_{k-1}!(n-x_1-x_2-\cdots-x_{k-1})!} \cdot \frac{(n-x_1-x_2-\cdots-x_{k-1})!}{x_k!(n-x_1-x_2-\cdots-x_k)!} \\
= \; & \frac{n!}{x_1!x_2!\cdots x_k!}.
\end{aligned}$$

Due to the independence of trials, each of these arrangements has probability $p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$, since p_1 is multiplied x_1 times in some order, p_2 is multiplied x_2 times in some order, $\ldots$, p_k is multiplied x_k times in some order. Therefore, we ultimately obtain

$$f(x_1, x_2, \ldots, x_k) = \frac{n!}{x_1!x_2!\cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}.$$

The restriction on the variables $x_1, x_2, \ldots, x_n$ is such that $x_i = 0, 1, \ldots, n$ for each $i = 1, 2, \ldots, k$ subject to the condition that $\sum_{i=1}^{k} x_i = n$.

Remarks:

(1) As a check that the joint pmf properly sums to 1, one could apply the well-known *multinomial series formula* to show that

$$\sum_{\substack{0 \le x_1, x_2, \ldots, x_k \le n \\ x_1 + x_2 + \cdots + x_k = n}} \frac{n!}{x_1!x_2!\cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} = (p_1 + p_2 + \cdots + p_k)^n = 1.$$

(2) We sometimes use the shorthand notation $(X_1, X_2, \ldots, X_k) \sim MN(n; p_1, p_2, \ldots, p_k)$ to indicate that $(X_1, X_2, \ldots, X_k)$ has a multinomial distribution with number of trials parameter n and outcome type probabilities $p_1, p_2, \ldots, p_k$.

(3) One could argue that it is better to represent our derived expressions solely in terms of $x_1, x_2, \ldots, x_{k-1}$ and $p_1, p_2, \ldots, p_{k-1}$ using the fact that

$$x_k = n - x_1 - \cdots - x_{k-1} \quad \text{and} \quad p_k = 1 - p_1 - \cdots - p_{k-1}.$$

In this case, we would express our joint pmf as a function of $x_1, x_2, \ldots, x_{k-1}$, keeping in mind now that $x_1, x_2, \ldots, x_{k-1}$ must satisfy the condition that $0 \leq x_1 + x_2 + \cdots + x_{k-1} \leq n$. When considering the multinomial distribution from this perspective, it is immediately evident that in the case of $k = 2$ the joint pmf simplifies to become the $Bin(n, p_1)$ pmf.

(4) If one is only interested in the random variable X_i, the number of type-i outcomes that have occurred, then we can view the repetitions of the experiment as a simple Bernoulli process, one in which an outcome is either of type i (with probability p_i) or it is not (with probability $1 - p_i$). In other words, the marginal probability distribution of X_i is simply that of a $Bin(n, p_i)$ distribution.

Example 4.2.1. Consider an experiment in which three fair dice are tossed. In ten independent tosses, let X be the number of times all three faces are alike and let Y be the number of times only two faces are alike. Determine the joint pmf of (X, Y).

Solution: In a single toss of three fair dice, we can easily compute

$$P(\text{all 3 faces are alike}) = \frac{6}{6^3} = \frac{1}{36}$$

and

$$P(\text{only 2 faces are alike}) = \frac{\binom{6}{1} \times 5 \times 3}{6^3} = \frac{15}{36} = \frac{5}{12}.$$

Therefore, it immediately follows that

$$P(\text{all 3 faces are distinct}) = 1 - \frac{1}{36} - \frac{5}{12} = \frac{5}{9}.$$

In ten independent tosses, let Z be the number of times that all three faces are distinct. It follows that $(X, Y, Z) \sim MN(10; \frac{1}{36}, \frac{5}{12}, \frac{5}{9})$. Since $Z = 10 - X - Y$, the joint pmf of (X, Y) is given by

$$f(x, y) = \frac{10!}{x!y!(10 - x - y)!} \left(\frac{1}{36}\right)^x \left(\frac{5}{12}\right)^y \left(\frac{5}{9}\right)^{10-x-y},$$

where $x, y = 0, 1, \ldots, 10$ such that $0 \leq x + y \leq 10$. ■

Remark: In Example 4.2.1, we know that $X \sim Bin(10, \frac{1}{36})$ and $Y \sim Bin(10, \frac{5}{12})$. As a result, it is straightforward to show that $f(x,y) \neq f_X(x)f_Y(y)$, implying that X and Y are dependent random variables. Moreover, if one were interested in the random variable $T = X + Y$, we could obtain the pmf of T as follows:

$$\begin{aligned}
f_T(t) &= \sum_{all\ x} f(x, t-x) \\
&= \sum_{x=0}^{t} \frac{10!}{x!(t-x)!(10-x-(t-x))!} \left(\frac{1}{36}\right)^x \left(\frac{5}{12}\right)^{t-x} \left(\frac{5}{9}\right)^{10-x-(t-x)} \\
&= \frac{10!}{t!(10-t)!} \left(\frac{5}{9}\right)^{10-t} \sum_{x=0}^{t} \underbrace{\frac{t!}{x!(t-x)!}}_{=\binom{t}{x}} \left(\frac{1}{36}\right)^x \left(\frac{5}{12}\right)^{t-x} \\
&= \binom{10}{t} \left(\frac{5}{9}\right)^{10-t} \left(\frac{1}{36} + \frac{5}{12}\right)^t \text{ by the binomial series formula} \\
&= \binom{10}{t} \left(\frac{4}{9}\right)^t \left(\frac{5}{9}\right)^{10-t} \quad \text{for } t = 0, 1, \ldots, 10.
\end{aligned}$$

In other words, $T = X + Y \sim Bin(10, \frac{4}{9})$. This result makes sense intuitively, since we can view each toss of three fair dice as producing an outcome which has either at least 2 faces being alike (with probability $p_1 + p_2 = \frac{1}{36} + \frac{5}{12} = \frac{4}{9}$) or none alike (with complementary probability $\frac{5}{9}$).

Section 4.2 Problems

4.2.1 An insurance company classifies policy holders as class A, B, C, or D. The probabilities of a randomly selected policy holder being in these categories are 0.1, 0.4, 0.3, and 0.2, respectively. Provide expressions for the probability that 25 randomly chosen policy holders will include:

(a) 3 A's, 11 B's, 7 C's, and 4 D's.

(b) 3 A's and 11 B's.

4.2.2 Three sprinters, Alvin, Barry, and Charlie, compete against each other in 9 independent 100-meter races. The probabilities of winning any single race are 0.5 for Alvin, 0.4 for Barry, and 0.1 for Charlie. Let X_1, X_2, and X_3 be the number of races Alvin, Barry, and Charlie win, respectively.

(a) What is the probability that all three sprinters win the same number of races?

(b) What is the probability that at most 4 races are won by Alvin or Charlie?

(c) What is the probability that exactly two of the three sprinters win the same number of races?

4.2.3 In a breeding experiment involving horses, the offspring can be one of four genetic types with the following probabilities:

Type	1	2	3	4
Probability	3/16	5/16	5/16	3/16

A group of 40 independent offspring are observed.

(a) Provide an expression for the probability that there are 10 offspring of each type.

(b) Provide an expression for the probability that the total number of type 1 or type 2 offspring is 16.

4.2.4 A certain type of battery has lifetimes that can be modelled by a geometric distribution with a mean of 100 days.

(a) Calculate the proportion of batteries which last:

(i) less than or equal to 50 days.

(ii) more than 50 days but less than or equal to 100 days.

(iii) more than 100 days but less than or equal to 150 days.

(iv) more than 150 days.

(b) In a small shipment of 20 batteries, provide an expression for the probability that 5 batteries last less than or equal to 50 days, 8 batteries last more than 50 days but less than or equal to 100 days, and 2 batteries last more than 150 days.

(c) In a large shipment of 50 batteries, what is the probability that 4 or more batteries last longer than 150 days?

4.2.5 In a particular city, the probability a call to a fire department concerns various situations is provided in the following table. Let X_i represent the number of calls of type i, $i = 1, 2, \ldots, 6$, in a set of 10 calls.

(a) Specify the joint pmf of $(X_1, X_2, \ldots, X_6)$.

(b) What is the probability that over half the calls are related to the first four situational types?

(c) What is the probability there is at least two non-fire-related emergencies and at least two false alarms?

Type of situation	Probability
fire in a detached home	$p_1 = 0.1$
fire in a semi-detached home	$p_2 = 0.05$
fire in an apartment or multiple unit residence	$p_3 = 0.05$
fire in a non-residential building	$p_4 = 0.15$
non-fire-related emergency	$p_5 = 0.15$
false alarm	$p_6 = 0.5$

4.3 Expectation, Covariance, and Correlation

When there are multiple random variables of interest, it is possible to generate new random variables by considering functions involving several of these random variables. As we witnessed in Section 4.1, a function $U = g(X, Y)$ of the random variables X and Y defines another random variable, and its pmf can be calculated from the joint pmf of (X, Y) according to the formula

$$f_U(u) = P(U = u) = \sum_{\substack{all\ (x,y):\\ g(x,y)=u}} f(x, y).$$

Therefore, the expected value rule for functions naturally extends and takes the form

$$E(U) = E\left(g(X, Y)\right) = \sum_{all\ (x,y)} g(x, y) f(x, y). \tag{4.3.1}$$

The verification of this result is very similar to the earlier case involving a function of a single random variable, as demonstrated in the proof of Theorem 3.3.1. Moreover, this result can be extended to functions of n random variables, such as $U = g(X_1, X_2, \ldots, X_n)$, in the following way:

$$E(U) = E\left(g(X_1, X_2, \ldots, X_n)\right) = \sum_{all\ (x_1,x_2,\ldots,x_n)} g(x_1, x_2, \ldots, x_n) f(x_1, x_2, \ldots, x_n).$$

Example 4.3.1. Suppose that the joint pmf of (X, Y) is given by the following table:

		x		
$f(x,y)$		1	2	3
y	-1	0.3	0.1	0.1
	1	0.2	0.2	0.1

Calculate the mean of $U = \frac{X^2 Y}{X+Y^2}$.

Solution: We wish to calculate

$$\begin{aligned}
E(U) &= E\left(\frac{X^2Y}{X+Y^2}\right) \\
&= \sum_{all\ (x,y)} \frac{x^2y}{x+y^2} f(x,y) \\
&= \frac{(1)^2(-1)}{1+(-1)^2}(0.3) + \frac{(2)^2(-1)}{2+(-1)^2}(0.1) + \frac{(3)^2(-1)}{3+(-1)^2}(0.1) + \frac{(1)^2(1)}{1+(1)^2}(0.2) + \frac{(2)^2(1)}{2+(1)^2}(0.2) + \frac{(3)^2(1)}{3+(1)^2}(0.1) \\
&= \left(-\frac{1}{2}\right)(0.3) + \left(-\frac{4}{3}\right)(0.1) + \left(-\frac{9}{4}\right)(0.1) + \left(\frac{1}{2}\right)(0.2) + \left(\frac{4}{3}\right)(0.2) + \left(\frac{9}{4}\right)(0.1) \\
&= -\frac{1}{20} + \frac{4}{30} \\
&= \frac{1}{12}.
\end{aligned}$$

■

Example 4.3.2. Suppose that $(X_1, X_2, \ldots, X_k) \sim MN(n; p_1, p_2, \ldots, p_k)$. For $i, j \in \{1, 2, \ldots, k\}$ and $i < j$, determine $E(X_iX_j)$.

Solution: It follows that the joint pmf of (X_i, X_j) is given by

$$f_{ij}(x_i, x_j) = \frac{n!}{x_i!x_j!(n-x_i-x_j)!} p_i^{x_i} p_j^{x_j} (1-p_i-p_j)^{n-x_i-x_j},$$

where $x_i, x_j = 0, 1, \ldots, n$ such that $x_i + x_j \leq n$. We wish to determine

$$\begin{aligned}
E(X_iX_j) &= \sum_{all\ (x_i,x_j)} x_ix_jf_{ij}(x_i,x_j) \\
&= \sum_{\substack{x_i,x_j=0,1,\ldots,n \\ x_i+x_j\leq n}} x_ix_j \cdot \frac{n!}{x_i!x_j!(n-x_i-x_j)!} p_i^{x_i} p_j^{x_j} (1-p_i-p_j)^{n-x_i-x_j} \\
&= \sum_{\substack{x_i,x_j=1,2,\ldots,n \\ x_i+x_j\leq n}} \frac{n(n-1)(n-2)!}{(x_i-1)!(x_j-1)!((n-2)-(x_i-1)-(x_j-1))!} p_i^{(x_i-1)+1} p_j^{(x_j-1)+1} (1-p_i-p_j)^{(n-2)-(x_i-1)-(x_j-1)} \\
&= n(n-1)p_ip_j \sum_{\substack{y_i,y_j=0,1,\ldots,n-2 \\ y_i+y_j\leq n-2}} \frac{(n-2)!}{y_i!y_j!((n-2)-y_i-y_j)!} p_i^{y_i} p_j^{y_j} (1-p_i-p_j)^{(n-2)-y_i-y_j} \text{ if we let } y_i = x_i - 1 \text{ and } y_j = x_j - 1 \\
&= n(n-1)p_ip_j(p_i+p_j+1-p_i-p_j)^{n-2} \text{ by the multinomial series formula} \\
&= n(n-1)p_ip_j.
\end{aligned}$$

■

Let us next consider a scenario in which $c_1, c_2, \ldots, c_n$ are real constants and $g_1, g_2, \ldots, g_n$ are arbitrary real-valued functions. Note that

$$\begin{aligned} E\left(\sum_{i=1}^{n} c_i g_i(X_1, X_2, \ldots, X_n)\right) &= \sum_{all\ (x_1,x_2,\ldots,x_n)} \left(\sum_{i=1}^{n} c_i g_i(x_1, x_2, \ldots, x_n)\right) f(x_1, x_2, \ldots, x_n) \\ &= \sum_{i=1}^{n} c_i \sum_{all\ (x_1,x_2,\ldots,x_n)} g_i(x_1, x_2, \ldots, x_n) f(x_1, x_2, \ldots, x_n) \\ &= \sum_{i=1}^{n} c_i E\left(g_i(X_1, X_2, \ldots, X_n)\right). \end{aligned}$$

In the special case when $g_i(X_1, X_2, \ldots, X_n) = X_i$ for each $i = 1, 2, \ldots, n$, the above result simplifies to give

$$E\left(\sum_{i=1}^{n} c_i X_i\right) = \sum_{i=1}^{n} c_i E(X_i). \tag{4.3.2}$$

In other words, the expected value of a linear combination of random variables is equal to the linear combination of individual expected values. As a further property of multivariate expectation, the following theorem also plays an important role in a variety of situations.

Theorem 4.3.1. *Suppose that X and Y are independent random variables. If g_1 and g_2 are two real-valued functions, then*

$$E(g_1(X)g_2(Y)) = E(g_1(X))E(g_2(Y)).$$

Proof: Since X and Y are independent random variables, we have that $f(x, y) = f_X(x)f_Y(y)$ for all values (x, y). Therefore, it follows that

$$\begin{aligned} E(g_1(X)g_2(Y)) &= \sum_{all\,(x,y)} g_1(x)g_2(y)f(x, y) \\ &= \sum_{all\,x} \sum_{all\,y} g_1(x)g_2(y)f_X(x)f_Y(y) \\ &= \left(\sum_{all\,x} g_1(x)f_X(x)\right)\left(\sum_{all\,y} g_2(y)f_Y(y)\right) \\ &= E(g_1(X))E(g_2(Y)). \end{aligned}$$

■

Remark: Theorem 4.3.1 can be extended to the case of n independent random variables. In particular, if $X_1, X_2, \ldots, X_n$ are independent random variables and $g_1, g_2, \ldots, g_n$ are arbitrary real-valued

functions, then

$$E\left(\prod_{i=1}^{n} g_i(X_i)\right) = \prod_{i=1}^{n} E(g_i(X_i)). \tag{4.3.3}$$

The following example demonstrates the usefulness of this result.

Example 4.3.3. Let $X_1, X_2, \ldots, X_k$ be independent random variables where $X_i \sim Geo(p)$ for $i = 1, 2, \ldots, k$. Determine the probability distribution of $T = \sum_{i=1}^{k} X_i$.

Solution: Recall that if each $X_i \sim Geo(p)$, then its mgf is given by

$$M_{X_i}(t) = \frac{p}{1-(1-p)e^t} \text{ for } t < \ln(1-p)^{-1}.$$

To determine the probability distribution of $T = \sum_{i=1}^{k} X_i$, let us consider the mgf of T given by

$$\begin{aligned}
M_T(t) &= E(e^{tT}) \\
&= E(e^{t(X_1+X_2+\cdots+X_k)}) \\
&= E(e^{tX_1}e^{tX_2}\cdots e^{tX_k}) \\
&= E(e^{tX_1})E(e^{tX_2})\cdots E(e^{tX_k}) \text{ using (4.3.3) since } X_1, X_2, \ldots, X_k \text{ are independent} \\
&= M_{X_1}(t)M_{X_2}(t)\cdots M_{X_k}(t) \\
&= \left(\frac{p}{1-(1-p)e^t}\right)^k \text{ for } t < \ln(1-p)^{-1}.
\end{aligned}$$

However, we recognize the above mgf as that of a $NB(k, p)$ distributed random variable. Therefore, by Theorem 3.4.1, we conclude that $T = \sum_{i=1}^{k} X_i \sim NB(k, p)$. ■

Remark: The same approach used in the solution of Example 4.3.3 can be applied to establish the following important distributional results:

(1) If $X_1, X_2, \ldots, X_k$ are independent random variables where $X_i \sim Bin(n_i, p)$ for $i = 1, 2, \ldots, k$, then $T = \sum_{i=1}^{k} X_i \sim Bin(\sum_{i=1}^{k} n_i, p)$.

(2) If $X_1, X_2, \ldots, X_k$ are independent random variables where $X_i \sim Poi(\mu_i)$ for $i = 1, 2, \ldots, k$, then $T = \sum_{i=1}^{k} X_i \sim Poi(\sum_{i=1}^{k} \mu_i)$.

Independence is a "yes/no" way of defining a relationship between random variables. We all know that there can be different types of relationships between random variables which are dependent. For example, if X is your height in inches and Y your height in centimeters, the relationship between X and Y is one-to-one and linear. More generally, two random variables may be related (non-independent) in a probabilistic sense. For example, a person's weight Y is not an exact linear function of their height X, but Y and X are nevertheless related. We will look at two ways of measuring the strength of the relationship between two random variables. The first measure we introduce is called the covariance.

Definition 4.3.1. *The **covariance** of random variables X and Y is given by*

$$Cov(X, Y) = E\left((X - \mu_X)(Y - \mu_Y)\right).$$

In keeping with the Greek letter notation we introduced for the mean and variance in Section 3.3, the covariance of X and Y is often denoted by σ_{XY}. Clearly, $Cov(X, Y) = Cov(Y, X)$ and $Cov(X, X) = Var(X)$. In addition, note that

$$\begin{aligned} Cov(X, Y) &= E\left[(X - \mu_X)(Y - \mu_Y)\right] \\ &= E\left(XY - \mu_X Y - X\mu_Y + \mu_X\mu_Y\right) \\ &= E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X\mu_Y \\ &= E(XY) - \mu_X\mu_Y - \mu_Y\mu_X + \mu_X\mu_Y \\ &= E(XY) - \mu_X\mu_Y. \end{aligned} \tag{4.3.4}$$

The formula given by (4.3.4) is the one we generally use for calculation purposes.

Example 4.3.4. Suppose that the joint pmf of (X, Y) is given by the following table:

		x		
$f(x,y)$		0	1	2
	1	0.3	0.2	0.1
y	2	0.1	0.1	0.2

Calculate the covariance of X and Y.

Solution: First of all, we have that

$$\begin{aligned} E(XY) &= \sum_{all\ (x,y)} xyf(x, y) \\ &= (0)(1)(0.3) + (1)(1)(0.2) + (2)(1)(0.1) + (0)(2)(0.1) + (1)(2)(0.1) + (2)(2)(0.2) \\ &= 1.4. \end{aligned}$$

To calculate $E(X)$, we have a choice of approaches. On the one hand, we could choose $g(X, Y) = X$ and use (4.3.1) to obtain

$$\begin{aligned} E(X) &= \sum_{all\ (x,y)} xf(x, y) \\ &= (0)(0.3) + (1)(0.2) + (2)(0.1) + (0)(0.1) + (1)(0.1) + (2)(0.2) \\ &= 0.9. \end{aligned}$$

Alternatively, we could sum down the columns of the above joint pmf table to find the marginal pmf $f_X(x)$ and then use

$$E(X) = \sum_{all\ x} x f_X(x) = (0)(0.4) + (1)(0.3) + (2)(0.3) = 0.9.$$

Either of these two approaches could be adopted to calculate $E(Y)$. Opting for the first approach, we get

$$E(Y) = \sum_{all\ (x,y)} y f(x,y) = (1)(0.3) + (1)(0.2) + (1)(0.1) + (2)(0.1) + (2)(0.1) + (2)(0.2) = 1.4.$$

Using (4.3.4), we therefore obtain $Cov(X,Y) = 1.4 - (0.9)(1.4) = 0.14$. ■

Interpretation of Covariance:

(1) Suppose large values of X tend to occur with large values of Y and small values of X with small values of Y. Then, the quantities $(X - \mu_X)$ and $(Y - \mu_Y)$ will tend to be of the same sign, whether positive or negative. Therefore, their product $(X - \mu_X)(Y - \mu_Y)$ will be positive, and so $Cov(X,Y) > 0$. Looking at Figure 4.3.1, for instance, we see several hundred pairs of (X,Y) points plotted. Notice that the majority of these points lie in the two quadrants (lower left and upper right) labelled with "+", meaning that for these points $(X - \mu_X)(Y - \mu_Y) > 0$. A minority of points lie in the other two quadrants labelled "−", and for these points we have $(X - \mu_X)(Y - \mu_Y) < 0$. Moreover, the points in the latter two quadrants appear closer to the mean (μ_X, μ_Y), indicating that on average, over all points generated, the average value of the product $(X - \mu_X)(Y - \mu_Y)$ is positive. Presumably, this implies that over the joint distribution of (X,Y), $E((X - \mu_X)(Y - \mu_Y)) > 0$, or equivalently, $Cov(X,Y) > 0$.

(2) Conversely, suppose large values of X tend to occur with small values of Y and small values of X with large values of Y. In this case, the quantities $(X - \mu_X)$ and $(Y - \mu_Y)$ will tend to be of opposite signs. As a result, their product $(X - \mu_X)(Y - \mu_Y)$ will tend to be negative, and hence $Cov(X,Y) < 0$. As a visual demonstration, Figure 4.3.2 depicts a plot of (X,Y) points indicating behaviour in this negative direction.

Example 4.3.5. Suppose that $(X_1, X_2, \ldots, X_k) \sim MN(n; p_1, p_2, \ldots, p_k)$. For $i, j \in \{1, 2, \ldots, k\}$ and $i < j$, determine $Cov(X_i, X_j)$. Do you expect $Cov(X_i, X_j)$ to be positive or negative?

Solution: In Example 4.3.2, we determined that $E(X_i X_j) = n(n-1)p_i p_j$. Since $(X_1, X_2, \ldots, X_k)$ has the aforementioned multinomial distribution, it immediately follows that $X_i \sim Bin(n, p_i)$ and

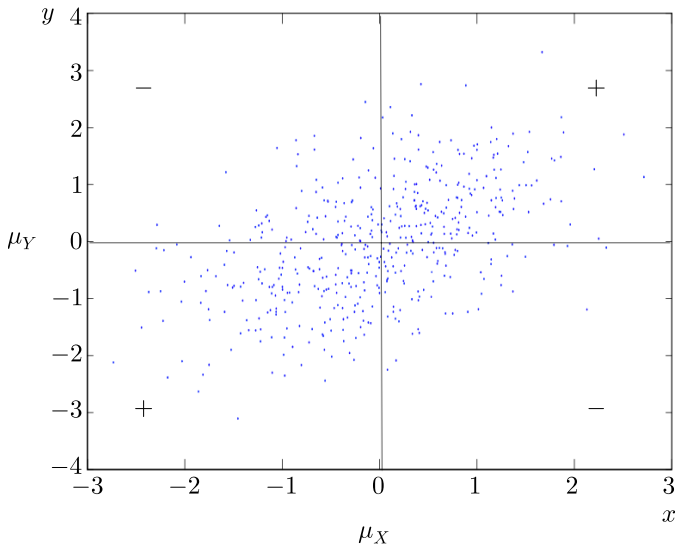


Figure 4.3.1: Random points (X, Y) indicating a positive covariance

$X_j \sim Bin(n, p_j)$, thereby implying that $E(X_i) = np_i$ and $E(X_j) = np_j$. As a result, we obtain

$$\begin{aligned} Cov(X_i, X_j) &= E(X_iX_j) - E(X_i)E(X_j) \\ &= n(n-1)p_ip_j - (np_i)(np_j) \\ &= n^2p_ip_j - np_ip_j - n^2p_ip_j \\ &= -np_ip_j, \end{aligned}$$

which is a negative value. However, this negative value is expected, since if X_i increases (meaning the number of type-i outcomes observed becomes greater), then there are less opportunities for type-j outcomes to be observed (meaning that there would be a tendency for X_j to decrease). ■

As we know, independent random variables are not related. In this situation, the next theorem confirms that this would lead to a zero covariance value.

Theorem 4.3.2. *If X and Y are independent random variables, then $Cov(X, Y) = 0$.*

Proof: Using (4.3.4), note that

$$\begin{aligned} Cov(X, Y) &= E(XY) - \mu_X\mu_Y \\ &= E(X)E(Y) - \mu_X\mu_Y \text{ by Theorem 4.3.1} \\ &= \mu_X\mu_Y - \mu_X\mu_Y \\ &= 0. \end{aligned}$$

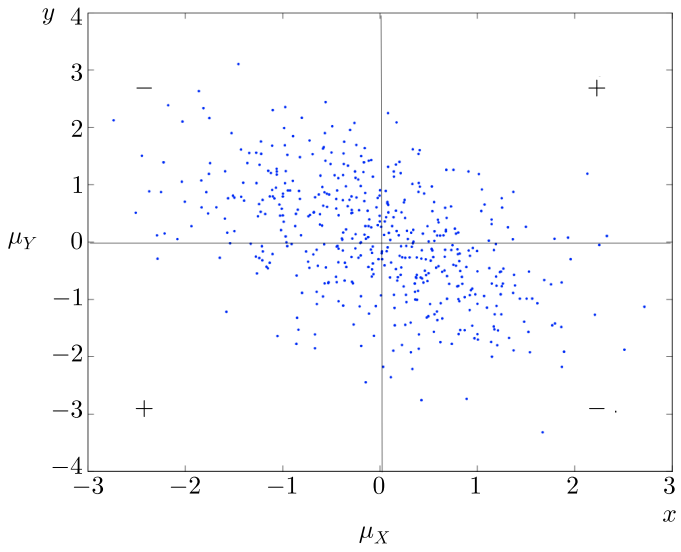


Figure 4.3.2: Random points (X, Y) indicating a negative covariance

■

It turns out that Theorem 4.3.2 is not reversible. In other words, if $Cov(X, Y) = 0$, then we can not conclude that X and Y are independent random variables. For example, suppose that Z is a random variable such that $Z \sim DU(-10, 10)$. If we define $X = \sin(0.2\pi Z)$ and $Y = \cos(0.2\pi Z)$, it is possible to show that $Cov(X, Y) = 0$, but the two random variables X and Y are clearly related (i.e., dependent) because the points (X, Y) are always on a circle. An even simpler example demonstrating this point follows below.

Example 4.3.6. Suppose that the joint pmf of (X, Y) is given by the following table:

		x		
$f(x,y)$		0	1	2
	0	0.2	0	0.2
y	1	0	0.6	0

Show that $Cov(X, Y) = 0$, but X and Y are not independent random variables.

Solution: Summing down the columns and across the rows of the above joint pmf table, the marginal probability distributions of X and Y are easily obtained:

x	0	1	2
$f_X(x)$	0.2	0.6	0.2

and

y	0	1
$f_Y(y)$	0.4	0.6

Note that

$$0.2 = f(0,0) \neq f_X(0)f_Y(0) = (0.2)(0.4) = 0.08,$$

and so X and Y are not independent random variables. However, we have that

$$\begin{aligned} E(XY) &= (0)(0)(0.2) + (1)(1)(0.6) + (2)(0)(0.2) = 0.6, \\ E(X) &= (0)(0.2) + (1)(0.6) + (2)(0.2) = 1, \text{ and} \\ E(Y) &= (0)(0.4) + (1)(0.6) = 0.6. \end{aligned}$$

Therefore, $Cov(X,Y) = E(XY) - E(X)E(Y) = 0.6 - (1)(0.6) = 0$. In other words, X and Y have zero covariance, but are not independent. ■

As a measure of strength of dependence, covariance does have a major drawback. In particular, for real constants a and b, note that

$$\begin{aligned} Cov(aX, bY) &= E((aX)(bY)) - E(aX)E(bY) \\ &= E(abXY) - aE(X) \cdot bE(Y) \\ &= abE(XY) - abE(x)E(Y) \\ &= abCov(X,Y). \end{aligned}$$

In other words, a large value of covariance can be simply the result of the size of the random variables themselves, and not signify an especially strong dependence. In fact, the actual numerical value of $Cov(X,Y)$ has no real interpretation, and this limits the use of covariance as a way of measuring the strength of relationship between X and Y. We now consider a second, related way which normalizes the covariance in the right way.

Definition 4.3.2. *The **correlation coefficient** of random variables X and Y having positive standard deviations σ_X and σ_Y is given by*

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}.$$

Most of the time, the correlation coefficient of X and Y is simply denoted by the Greek letter ρ, or ρ_{XY} if we wish to explicitly state the random variables X and Y under consideration. Note that, by construction, the correlation coefficient is a unitless quantity. In addition, we see that the correlation coefficient is simply a rescaled version of the covariance. Since σ_X and σ_Y are both positive, ρ will have the same sign as $Cov(X,Y)$. Therefore, the interpretation of the sign of ρ is the same as for $Cov(X,Y)$. Moreover, $\rho = 0$ if X and Y are independent. However, if $\rho = 0$, we cannot say that X and Y are independent. Instead, we simply say that X and Y are *uncorrelated* random variables.

While the value of covariance is unbounded, the same is not true for the correlation coefficient. In fact, we can show that $-1 \leq \rho \leq 1$. To do so, let us define a new random variable $Z = Y - tX$, where t is some real number. Note that

$$\begin{aligned} Var(Z) &= E\left((Z-\mu_Z)^2\right) \\ &= E\left([(Y-tX)-(\mu_Y-t\mu_X)]^2\right) \\ &= E\left([(Y-\mu_Y)-t(X-\mu_X)]^2\right) \\ &= E\left((Y-\mu_Y)^2-2t(X-\mu_X)(Y-\mu_Y)+t^2(X-\mu_X)^2\right) \\ &= \sigma_X^2 t^2 - 2Cov(X,Y)t + \sigma_Y^2. \end{aligned}$$

Since $Var(Z) \geq 0$ for all $t \in \mathbb{R}$, this quadratic equation in t must have at most one real root. This implies that the discriminant of this quadratic equation must be less than or equal to 0. In other words, we must have that

$$(-2Cov(X,Y))^2 - 4\sigma_X^2\sigma_Y^2 \leq 0,$$

which immediately leads to the inequality

$$\left|\frac{Cov(X,Y)}{\sigma_X\sigma_Y}\right| \leq 1,$$

or equivalently,

$$-1 \leq \rho \leq 1.$$

We observe that as $\rho \to \pm 1$, the relation between X and Y becomes one-to-one and linear. To see that $\rho = \pm 1$ corresponds to a one-to-one linear relationship between X and Y, note that $\rho = \pm 1$ corresponds to a zero discriminant in the aforementioned quadratic equation in t. This implies that there exists one real number t^* for which

$$Var(Z) = Var(Y - t^*X) = 0.$$

But in order for $Var(Y - t^*X)$ to be zero, this means that $Y - t^*X$ must be equal to a constant c. In order words, $Y = t^*X + c$ and X and Y satisfy a linear relationship. As a result, this is why the correlation coefficient serves as a measure of the strength of the *linear relationship* between X and Y. The closer the value of ρ is to ± 1, the stronger the linear dependence. The closer the value of ρ is to 0, the more evidence there is to indicate a weak linear relationship between the two random variables.

Example 4.3.7. Suppose that $X \sim DU(-2, 2)$ and let $Y = a + bX + cX^2$ for real constants a, b, and c. Find the correlation coefficient of X and Y.

Solution: First of all, the pmf of X is given by

$$f_X(x) = \frac{1}{5} \text{ for } x = -2, -1, 0, 1, 2.$$

As a result, the first four moments of X are given by

$$E(X) = \sum_{x=-2}^{2} x f_X(x) = \frac{1}{5}(-2 - 1 + 0 + 1 + 2) = 0,$$

$$E(X^2) = \sum_{x=-2}^{2} x^2 f_X(x) = \frac{1}{5}(4 + 1 + 0 + 1 + 4) = 2,$$

$$E(X^3) = \sum_{x=-2}^{2} x^3 f_X(x) = \frac{1}{5}(-8 - 1 + 0 + 1 + 8) = 0,$$

and

$$E(X^4) = \sum_{x=-2}^{2} x^4 f_X(x) = \frac{1}{5}(16 + 1 + 0 + 1 + 16) = \frac{34}{5}.$$

Therefore, $\sigma_X^2 = E(X^2) - E(X)^2 = 2 - 0^2 = 2$. Next, we determine the mean and variance of Y as follows:

$$E(Y) = E(a + bX + cX^2) = a + bE(X) + cE(X^2) = a + b(0) + c(2) = a + 2c$$

and

$$\begin{aligned}
\sigma_Y^2 &= E\left((a + bX + cX^2)^2\right) - (a + 2c)^2 \\
&= E(a^2 + abX + acX^2 + abX + b^2X^2 + bcX^3 + acX^2 + bcX^3 + c^2X^4) - (a^2 + 4ac + 4c^2) \\
&= a^2 + ab(0) + ac(2) + ab(0) + b^2(2) + bc(0) + ac(2) + bc(0) + c^2\left(\frac{34}{5}\right) - (a^2 + 4ac + 4c^2) \\
&= 2b^2 + \frac{14}{5}c^2.
\end{aligned}$$

Finally, we have that

$$Cov(X, Y) = E(XY) - E(X)E(Y) = E\left(X(a + bX + cX^2)\right) = E(aX + bX^2 + cX^3) = 2b.$$

Therefore, the correlation coefficient of X and Y is given by

$$\rho = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} = \frac{2b}{\sqrt{2}\sqrt{2b^2 + \frac{14}{5}c^2}} = \frac{b}{\sqrt{b^2 + \frac{7}{5}c^2}}.$$

Interestingly, $\rho = 0$ if $b = 0$, implying that X and Y are uncorrelated (although a non-linear relationship of the form $Y = a + cX^2$ is in place). On the other hand, $\rho = \pm 1$ (depending on the sign of b) if $c = 0$, which is to be expected since the relationship between X and Y is linear in this case. ■

Section 4.3 Problems

4.3.1 Let X and Y be discrete random variables with joint pmf of the form

$$f(x,y) = \frac{4}{5xy} \text{ for } x = 1,2 \text{ and } y = 2,3.$$

(a) Calculate the mean and variance of the random variable $Z = \frac{Y}{X}$.

(b) Calculate the correlation coefficient of X and Y. What does it indicate about the relationship between X and Y?

4.3.2 For real constants a and b, show that $Cov(X + a, Y + b) = Cov(X, Y)$.

4.3.3 Suppose that the joint pmf of (X, Y) is given by

		x	
$f(x,y)$	0	1	2
0	0.06	0.15	0.09
y			
1	0.14	0.35	0.21

Calculate the correlation coefficient of X and Y. What does it indicate about the relationship between X and Y?

4.3.4 Suppose that X and Y are random variables with joint pmf of the following form:

		x	
$f(x,y)$	2	4	6
-1	1/8	1/4	p
y			
1	1/4	1/8	$\frac{1}{4} - p$

(a) What are the possible values for p?

(b) Determine the value(s) of p which result in X and Y being uncorrelated.

(c) Show that there is no value of p for which X and Y are independent.

4.3.5 Let X be the number of ones and Y the number of twos that occur in n rolls of a fair six-sided die. Calculate the correlation coefficient of X and Y.

4.3.6 Consider the joint probability distribution given in Problem 4.1.5. Calculate the correlation coefficient of X and Y.

4.3.7 Consider the joint probability distribution given in Problem 4.1.6. Calculate the correlation coefficient of X and Y.

4.3.8 In a slot machine at the local casino, suppose that there are $n+1$ possible outcomes $A_1, A_2, \ldots, A_{n+1}$ for a single play. A single play costs $\$d$. If outcome A_i occurs, you win $\$a_i$ for $i = 1, 2, \ldots, n$. If outcome A_{n+1} occurs, you win nothing. In other words, if outcome A_i, $i = 1, 2, \ldots, n$, occurs, your net profit is $a_i - d$; if A_{n+1} occurs, your net profit is $-d$.

(a) Give a formula for your expected net profit from m independent plays of the slot machine, assuming that the probabilities of the $n + 1$ outcomes are given by $p_i = P(A_i)$, $i = 1, 2, \ldots, n + 1$.

(b) The owner of the slot machine would like the player's expected net profit to be negative. Suppose that $n = 4$ with $p_1 = 0.1$, $p_2 = p_3 = p_4 = 0.04$, and $p_5 = 0.78$. If the slot machine is set to pay \$10 when outcome A_1 occurs and \$25 when either of outcomes A_2, A_3, or A_4 occur, what amount per play should the owner charge to ensure that the player's expected net profit will be negative?

4.4 Linear Combinations of Random Variables

Many problems of a practical nature require us to consider linear combinations of random variables. Specifically, if $X_1, X_2, \ldots, X_n$ are n random variables such that $\mu_i = E(X_i)$ and $\sigma_i^2 = Var(X_i)$ for $i = 1, 2, \ldots, n$, then we are often interested in the mean and variance of the random variable

$$V = \sum_{i=1}^{n} c_i X_i,$$

where $c_1, c_2, \ldots, c_n$ are real constants. In fact, we have already derived a formula for $E(V)$ in the previous section, namely (see (4.3.2))

$$E(V) = E\left(\sum_{i=1}^{n} c_i X_i\right) = \sum_{i=1}^{n} c_i \mu_i. \tag{4.4.1}$$

As an immediate consequence, the following special cases are obtained:

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mu_i, \tag{4.4.2}$$

$$E(X_1 + X_2) = \mu_1 + \mu_2,$$

$$E(X_1 - X_2) = \mu_1 - \mu_2.$$

To determine a formula for $Var(V)$, let us first examine a different problem. In particular, consider another sequence of random variables, $Y_1, Y_2, \ldots, Y_r$, and define

$$W = \sum_{j=1}^{r} d_j Y_j,$$

where $d_1, d_2, \ldots, d_r$ are real constants and $\alpha_j = E(Y_j)$ for $j = 1, 2, \ldots, r$. We wish to determine an expression for $Cov(V, W)$. Using Definition 4.3.1, we have that

$$\begin{aligned}
Cov(V,W) &= E\left((V - E(V))\,(W - E(W))\right) \\
&= E\left(\left(\sum_{i=1}^{n} c_i X_i - \sum_{i=1}^{n} c_i \mu_i\right)\left(\sum_{j=1}^{r} d_j Y_j - \sum_{j=1}^{r} d_j \alpha_j\right)\right) \\
&= E\left(\left(\sum_{i=1}^{n} c_i (X_i - \mu_i)\right)\left(\sum_{j=1}^{r} d_j (Y_j - \alpha_j)\right)\right) \\
&= E\left(\sum_{i=1}^{n}\sum_{j=1}^{r} c_i d_j (X_i - \mu_i)(Y_j - \alpha_j)\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{r} c_i d_j E\left((X_i - \mu_i)(Y_j - \alpha_j)\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{r} c_i d_j Cov(X_i, Y_j). \qquad (4.4.3)
\end{aligned}$$

We highlight two special cases of interest. First of all, in the case when $n = r = 2$, (4.4.3) yields the result

$$Cov(c_1X_1+c_2X_2, d_1Y_1+d_2Y_2) = c_1d_1Cov(X_1, Y_1)+c_1d_2Cov(X_1, Y_2)+c_2d_1Cov(X_2, Y_1)+c_2d_2Cov(X_2, Y_2).$$

Secondly, suppose that $n = r$, and for $i = 1, 2, \ldots, n$, we set $X_i = Y_i$ and $c_i = d_i$. In other words, we set

$W = V$. Since $Cov(V, V) = Var(V)$, (4.4.3) simplifies to give

$$\begin{aligned}
Var(V) &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j Cov(X_i, X_j) \\
&= \sum_{i=1}^{n} c_i^2 Cov(X_i, X_i) + \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} c_i c_j Cov(X_i, X_j) \\
&= \sum_{i=1}^{n} c_i^2 Var(X_i) + \sum_{i=1}^{n}\sum_{j=1}^{i-1} c_i c_j Cov(X_i, X_j) + \sum_{i=1}^{n}\sum_{j=i+1}^{n} c_i c_j Cov(X_i, X_j) \\
&= \sum_{i=1}^{n} c_i^2 \sigma_i^2 + \sum_{i=2}^{n}\sum_{j=1}^{i-1} c_i c_j Cov(X_i, X_j) + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_i c_j Cov(X_i, X_j) \\
&= \sum_{i=1}^{n} c_i^2 \sigma_i^2 + \sum_{j=1}^{n-1}\sum_{i=j+1}^{n} c_i c_j Cov(X_i, X_j) + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_i c_j Cov(X_i, X_j) \text{ by interchanging the order of summation} \\
&= \sum_{i=1}^{n} c_i^2 \sigma_i^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_i c_j Cov(X_i, X_j). \qquad (4.4.4)
\end{aligned}$$

In the case when $c_i = 1$ for $i = 1, 2, \ldots, n$, (4.4.4) gives rise to

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \sigma_i^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} Cov(X_i, X_j), \qquad (4.4.5)$$

which, when $n = 2$, becomes

$$Var(X_1 + X_2) = \sigma_1^2 + \sigma_2^2 + 2Cov(X_1, X_2).$$

Another special case of fundamental importance is when $X_1, X_2, \ldots, X_n$ are independent random variables. From Theorem 4.3.2, we have that $Cov(X_i, X_j) = 0$ for all $i \neq j$, and this fact results in (4.4.5) simplifying to become

$$Var(V) = \sum_{i=1}^{n} c_i^2 \sigma_i^2. \qquad (4.4.6)$$

This gives rise to the following widely-used results in the case when $X_1, X_2, \ldots, X_n$ are independent:

$$\begin{aligned}
Var\left(\sum_{i=1}^{n} X_i\right) &= \sum_{i=1}^{n} \sigma_i^2, \qquad (4.4.7) \\
Var(X_1 + X_2) &= \sigma_1^2 + \sigma_2^2, \\
Var(X_1 - X_2) &= \sigma_1^2 + (-1)^2\sigma_2^2 = \sigma_1^2 + \sigma_2^2.
\end{aligned}$$

In other words, for independent random variables, the *variance of a sum is the sum of the variances* and the variance of a difference is the sum of the variances.

Example 4.4.1. Suppose that $X_1, X_2, \ldots, X_n$ are independent random variables which have the same mean μ and same variance σ^2. The sample mean, which we introduced briefly in Section 3.3, of these n random variables is defined as

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

Determine the mean and variance of $\bar{X}$.

Solution: First of all, we note that we can represent $\bar{X}$ as

$$\bar{X} = \sum_{i=1}^{n}\left(\frac{1}{n}\right)X_i,$$

which is a particular linear combination of $X_1, X_2, \ldots, X_n$. To determine $E(\bar{X})$, we apply (4.4.1) to get

$$\begin{aligned} E(\bar{X}) &= \sum_{i=1}^{n}\left(\frac{1}{n}\right)\mu_i \\ &= \sum_{i=1}^{n}\left(\frac{1}{n}\right)\mu \ \text{ since } \mu_i = \mu \text{ for each } i = 1, 2, \ldots, n \\ &= \frac{\mu}{n}\cdot\underbrace{\sum_{i=1}^{n} 1}_{=n} \\ &= \mu. \end{aligned}$$

To obtain $Var(\bar{X})$ when $X_1, X_2, \ldots, X_n$ are independent, we apply (4.4.6) to obtain

$$\begin{aligned} Var(\bar{X}) &= \sum_{i=1}^{n}\left(\frac{1}{n}\right)^2\sigma_i^2 \\ &= \sum_{i=1}^{n}\left(\frac{1}{n}\right)^2\sigma^2 \ \text{ since } \sigma_i^2 = \sigma^2 \text{ for each } i = 1, 2, \ldots, n \\ &= \frac{\sigma^2}{n^2}\underbrace{\sum_{i=1}^{n} 1}_{=n} \\ &= \frac{\sigma^2}{n}. \end{aligned}$$

■

Remarks:

(1) The results of Example 4.4.1 are very important ones in probability and statistics. It establishes that if $X_1, X_2, \ldots, X_n$ are independent random variables with the same mean μ and same variance σ^2, then the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ has mean μ and variance $\frac{\sigma^2}{n}$.

(2) Since $Var(\bar{X}) = \frac{\sigma^2}{n}$, this implies that the average $\bar{X}$ of n independent random variables with the same distribution is less variable than any single observation X_i, and that the larger n is, the less variability there is. This explains mathematically why, for example, that if we want to estimate the unknown mean height μ in a population of people, we are better to take the average height for a random sample of $n = 10$ persons than to simply take the height of one randomly selected person. A sample of $n = 20$ persons would be better still. In Chapter 6, we will see how to decide how large a sample we should take in order to attain a certain degree of precision.

(3) Note that as $n \to \infty$, $Var(\bar{X}) \to 0$, which means that $\bar{X}$ becomes arbitrarily close to μ. This is sometimes referred to as the "law of averages". In actual fact, there is a formal theorem, called the *law of large numbers*, which supports the claim that for large sample sizes, sample means approach the expected value.

Indicator Random Variables:

The above results we have accumulated for linear combinations of random variables find some of their greatest use in providing a way of breaking up more complicated problems, involving mean and variance, into simpler, more manageable pieces. This is particularly the case with the use of *indicator random variables*. Simply put, an indicator random variable is a $Bin(1, p)$ random variable – that is, a binary random variable (0 or 1) used to indicate whether some event of interest occurs (with probability p) or does not occur (with probability $1 - p$). We will illustrate the usefulness of such random variables by looking at several examples.

Example 4.4.2. Let $Y \sim Bin(n, p)$. Use indicator random variables to derive the mean and variance of Y.

Solution: If $Y \sim Bin(n, p)$, then recall that Y represents the number of successes obtained in n independent Bernoulli trials. For $i = 1, 2, \ldots, n$, let us define the indicator random variable X_i as follows:

$$X_i = 0 \text{ if the } i^{\text{th}} \text{ trial is a failure, and}$$
$$X_i = 1 \text{ if the } i^{\text{th}} \text{ trial is a success.}$$

In other words, the random variable X_i indicates whether the outcome "success" occurs on the i^{th} trial. Based on the meaning of each X_i and the above definition of Y, it immediately follows that

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

Therefore, we can derive the mean and variance of Y by using our earlier results concerning the mean and variance of a sum of random variables. Since $X_i \sim Bin(1, p)$ for each $i = 1, 2, \ldots, n$, we have that

$\mu_i = E(X_i) = p$ and $\sigma_i^2 = Var(X_i) = p(1-p)$. Moreover, since the trials are independent of each other, it means that $X_1, X_2, \ldots, X_n$ are independent. As a result, (4.4.2) and (4.4.7) subsequently give rise to

$$E(Y) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mu_i = \sum_{i=1}^{n} p = np$$

and

$$Var(Y) = Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \sigma_i^2 = \sum_{i=1}^{n} p(1-p) = np(1-p).$$

These formulas are, of course, identical to the ones we derived previously for the mean and variance of a $Bin(n, p)$ distribution in Section 3.5. However, note how simple the derivation here is! ■

In Example 4.4.2, we were aided by the fact that the indicator random variables $X_1, X_2, \ldots, X_n$ were independent. In some problems, such as the next one, this is not the case. If $X_1, X_2, \ldots, X_n$ are dependent, we will also require their covariances in order to use the variance formula given by (4.4.4).

Example 4.4.3. Let $Y \sim HG(N, r, n)$. Use indicator random variables to derive the mean and variance of Y.

Solution: As in Example 4.4.2, let us, at the outset, think of the setting of this problem, which involves drawing n objects at random without replacement from a population of N total objects, of which r are type "s" (for success) and $N - r$ are type "f" (for failure). For $i = 1, 2, \ldots, n$, let us define

$$X_i = 0 \text{ if the } i^{\text{th}} \text{ drawn object is of type } f, \text{ and}$$
$$X_i = 1 \text{ if the } i^{\text{th}} \text{ drawn object is of type } s.$$

Since Y represents the number of success-type objects obtained, it follows that $Y = \sum_{i=1}^{n} X_i$, just as in Example 4.4.2. However, since the draws are made without replacement, the random variables $X_1, X_2, \ldots, X_n$ are now dependent (for example, what we get on the first draw affects the probability of getting a success-type object on the second draw, and so on). Thus, we will need to find $Cov(X_i, X_j)$ for $i \neq j$ in order to use (4.4.5) to calculate the variance of a sum of dependent random variables.

We see first that $X_i \sim Bin(1, p_i)$ for $i = 1, 2, \ldots, n$, where

$$p_i = P(X_i = 1) = P(i^{\text{th}} \text{ drawn object is of type } s) = \frac{r \times (N-1)^{(n-1)}}{N^{(n)}} = \frac{r \times (N-1)^{(n-1)}}{N(N-1)^{(n-1)}} = \frac{r}{N}.$$

The above result makes sense, since if the draws are random, the probability a success-type object occurs in draw i is simply equal to the probability that position i is occupied by an s if we were to

arrange r s's and $(N-r)$ f's in a row. In a similar fashion, for $i \neq j$, we would have

$$\begin{aligned} P(X_i = 1, X_j = 1) &= P(i^{\text{th}} \text{ drawn object is of type } s \text{ and } j^{\text{th}} \text{ drawn object is of type } s) \\ &= \frac{r \times (r-1) \times (N-2)^{(n-2)}}{N^{(n)}} \\ &= \frac{r \times (r-1) \times (N-2)^{(n-2)}}{N(N-1)(N-2)^{(n-2)}} \\ &= \frac{r(r-1)}{N(N-1)}. \end{aligned}$$

With these results, we immediately obtain

$$\mu_i = E(X_i) = p_i = \frac{r}{N},\ i = 1, 2, \ldots, n,$$

$$\sigma_i^2 = Var(X_i) = p_i(1-p_i) = \frac{r}{N}\left(1 - \frac{r}{N}\right) = \frac{r(N-r)}{N^2},\ i = 1, 2, \ldots, n,$$

and

$$\begin{aligned} Cov(X_i, X_j) &= E(X_i X_j) - \mu_i \mu_j \\ &= \sum_{x_i=0}^{1} \sum_{x_j=0}^{1} x_i x_j P(X_i = x_i, X_j = x_j) - \left(\frac{r}{N}\right)^2 \\ &= (1)(1) P(X_i = 1, X_j = 1) - \left(\frac{r}{N}\right)^2 \\ &= \frac{r(r-1)}{N(N-1)} - \left(\frac{r}{N}\right)^2 \\ &= \frac{r}{N}\left(\frac{r-1}{N-1} - \frac{r}{N}\right) \\ &= \frac{r}{N} \cdot \frac{Nr - N - Nr + r}{N(N-1)} \\ &= -\frac{r(N-r)}{N^2(N-1)},\ i \neq j. \end{aligned}$$

Finally, to find $E(Y)$ and $Var(Y)$, we use (4.4.2) and (4.4.5), respectively, to get

$$E(Y) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mu_i = \sum_{i=1}^{n} \frac{r}{N} = \frac{nr}{N}$$

and

$$\begin{aligned}
Var(Y) &= Var\left(\sum_{i=1}^{n} X_i\right) \\
&= \sum_{i=1}^{n} \sigma_i^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} Cov(X_i, X_j) \\
&= \sum_{i=1}^{n} \frac{r(N-r)}{N^2} - 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \frac{r(N-r)}{N^2(N-1)} \\
&= \frac{nr(N-r)}{N^2} - \frac{2r(N-r)}{N^2(N-1)} \underbrace{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} 1}_{\text{number of subsets of size 2 chosen from } \{1,2,\ldots,n\}} \\
&= \frac{nr(N-r)}{N^2} - \frac{2r(N-r)}{N^2(N-1)} \binom{n}{2} \\
&= \frac{nr(N-r)}{N^2} - \frac{2r(N-r)}{N^2(N-1)} \cdot \frac{n(n-1)}{2} \quad \text{since } \binom{n}{2} = \frac{n^{(2)}}{2!} = \frac{n(n-1)}{2} \\
&= \frac{nr(N-r)}{N^2}\left(1 - \frac{n-1}{N-1}\right) \\
&= \frac{nr(N-r)(N-n)}{N^2(N-1)}.
\end{aligned}$$

Just as with Example 4.4.2, these formulas agree with the results we obtained for the hypergeometric distribution in Section 3.5. ■

In the last two examples, it bears mentioning that the pmf of Y happens to be known, which means we could have determined $E(Y)$ and $Var(Y)$ without the use of indicator random variables. In the next example, however, the pmf of the random variable of interest is not known and is difficult to find. Despite this, indicator random variables can still be used to obtain the mean and variance. This particular example is a famous problem in probability.

Example 4.4.4. Consider a situation in which we have N letters to N different people, and N envelopes addressed to those N people. One letter is placed into each envelope at random. Determine the mean and variance of the number of letters placed in their correct envelope.

Solution: For $i = 1, 2, \ldots, N$, let us define the indicator random variable X_i as follows:

X_i = 0 if letter i is not correctly placed in its envelope, and
X_i = 1 if letter i is correctly placed in its envelope.

Therefore, $Y = \sum_{i=1}^{N} X_i$ represents the number of letters which are correctly placed. Just as in Example 4.4.3, we note that the random variables $X_1, X_2, \ldots, X_N$ are dependent (since incorrectly placing a let-

ter in an envelope negatively affects the chance of matching another letter with its envelope).

We begin again by noting that $X_i \sim Bin(1, p_i)$ for $i = 1, 2, \ldots, n$, where

$$p_i = P(\text{letter } i \text{ is correctly placed in its envelope}) = \frac{1 \times (N-1)!}{N!} = \frac{1}{N}.$$

Similarly, for $i \neq j$, we have that

$$P(X_i = 1, X_j = 1) = P(\text{letters } i \text{ and } j \text{ are correctly placed in their envelopes}) = \frac{1 \times 1 \times (N-2)!}{N!} = \frac{1}{N(N-1)}.$$

Therefore, for $i = 1, 2, \ldots, n$, we have that

$$\mu_i = E(X_i) = p_i = \frac{1}{N} \text{ and } \sigma_i^2 = Var(X_i) = p_i(1-p_i) = \frac{1}{N}\left(1 - \frac{1}{N}\right).$$

Furthermore, for $i \neq j$, we see that

$$\begin{aligned} Cov(X_i, X_j) &= E(X_i X_j) - \mu_i \mu_j \\ &= \sum_{x_i=0}^{1} \sum_{x_j=0}^{1} x_i x_j P(X_i = x_i, X_j = x_j) - \left(\frac{1}{N}\right)^2 \\ &= (1)(1)P(X_i = 1, X_j = 1) - \frac{1}{N^2} \\ &= \frac{1}{N(N-1)} - \frac{1}{N^2} \\ &= \frac{N - (N-1)}{N^2(N-1)} \\ &= \frac{1}{N^2(N-1)}. \end{aligned}$$

At last, we once again use (4.4.2) and (4.4.5), respectively, to obtain

$$E(Y) = E\left(\sum_{i=1}^{N} X_i\right) = \sum_{i=1}^{N} \mu_i = \sum_{i=1}^{N} \frac{1}{N} = \left(\frac{1}{N}\right) N = 1$$

and

$$\begin{aligned}
Var(Y) &= Var\left(\sum_{i=1}^{n} X_i\right) \\
&= \sum_{i=1}^{N} \sigma_i^2 + 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} Cov(X_i, X_j) \\
&= \sum_{i=1}^{N} \frac{1}{N}\left(1 - \frac{1}{N}\right) + 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \frac{1}{N^2(N-1)} \\
&= \frac{1}{N}\left(1 - \frac{1}{N}\right)N + \frac{2}{N^2(N-1)}\binom{N}{2} \\
&= 1 - \frac{1}{N} + \frac{2}{N^2(N-1)} \cdot \frac{N(N-1)}{2} \\
&= 1 - \frac{1}{N} + \frac{1}{N} \\
&= 1.
\end{aligned}$$

■

Remark: Common sense often helps in this course, but we have found no way of being able to say that the results of Example 4.4.4 are obvious. What is obvious though is the important role that indicator random variables play in tackling a problem of this nature. With its use, we are able to show that, rather remarkably, 1 letter on average will be correctly placed and the variance will be 1 as well, regardless of how many letters there are.

Section 4.4 Problems

4.4.1 Suppose that the joint pmf of (X, Y) is given by

		x		
$f(x,y)$		0	1	2
	0	0.15	0.1	0.05
y				
	1	0.35	0.2	0.15

Calculate the variance of the random variable $U = 3X - 2Y$.

4.4.2 Suppose that X and Y are random variables such that $Var(X) = 1.69$, $Var(Y) = 4$, and $\rho_{XY} = 0.5$. Calculate the standard deviation of the random variable $Z = 2X + Y$.

4.4.3 If X and Y are random variables with $Var(X) = 13$, $Var(Y) = 34$, and $\rho_{XY} = -0.7$, then calculate $Var(X - 2Y)$.

4.4.4 Let X and Y be independent random variables with $Var(X) = 1$ and $Var(Y) = 2$. Calculate the correlation coefficient of $2X + Y$ and $X - 2Y$.

4.4.5 Jack and Jill each toss a fair coin three times. Let X be the number of heads Jack obtains and Y the number of heads Jill obtains. Define $U = X + Y$ and $V = X - Y$.

(a) Calculate the means and variances of U and V.

(b) Calculate $Cov(U, V)$.

(c) Are U and V independent random variables? Justify your response.

4.4.6 A multiple choice exam has 100 questions, each with 5 possible answers. One mark is awarded for a correct answer and $\frac{1}{4}$ mark is deducted for an incorrect answer. A particular student has probability p_i of knowing the correct answer to the i^{th} question, independently of the other questions.

(a) Suppose that on a question where the student does not know the answer, he or she guesses randomly. Show that the student's total mark has mean

$$\sum_{i=1}^{100} p_i$$

and variance

$$\sum_{i=1}^{100} p_i (1 - p_i) + \frac{1}{4}\left(100 - \sum_{i=1}^{100} p_i\right).$$

(b) Show that the total mark for a student who refrains from guessing also has mean

$$\sum_{i=1}^{100} p_i,$$

but with variance

$$\sum_{i=1}^{100} p_i (1 - p_i).$$

Compare the variances when (i) $p_i = 0.9$ for $i = 1, 2, \ldots, 100$, and (ii) $p_i = 0.5$ for $i = 1, 2, \ldots, 100$.

4.4.7 Suppose that $X_1, X_2, \ldots, X_n$ are independent random variables which have the same mean μ and same variance σ^2. The *sample variance* of these n random variables is defined as

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2,$$

where $\bar{X}$ denotes the sample mean.

(a) Show that

$$S^2 = \frac{1}{n-1} \left(\sum_{i=1}^{n} X_i^2 - n\left(\bar{X}\right)^2 \right).$$

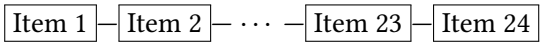
(b) Use the representation in part (a) to determine the mean of S^2.

4.4.8 Let $Y_0, Y_1, \ldots, Y_n$ be $n+1$ uncorrelated random variables with $E(Y_i) = 0$ and $Var(Y_i) = \sigma^2$ for $i = 0, 1, \ldots, n$. Define random variables $X_1, X_2, \ldots, X_n$ such that $X_i = Y_{i-1} + Y_i$ for $i = 1, 2, \ldots, n$.

(a) Calculate $Cov(X_{i-1}, X_i)$ for $i = 2, 3, \ldots, n$.

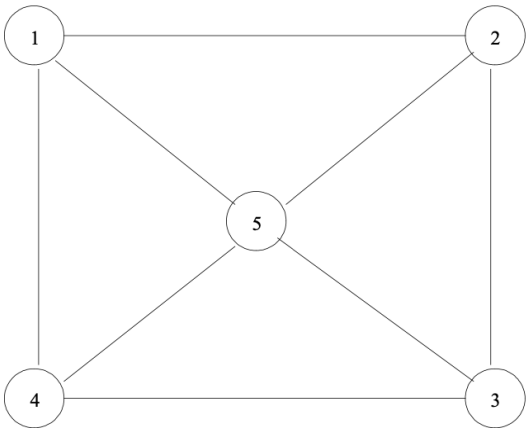
(b) Calculate $Var\left(\sum_{i=1}^{n} X_i\right)$.

4.4.9 A plastic fabricating company produces items in strips of 24, with the items connected by a thin piece of plastic as follows:



A cutting machine then cuts the connecting pieces to separate the items, with the 23 cuts made independently. There is a 10% chance that the machine will fail to cut a connecting piece. Calculate the mean and standard deviation of the number of the 24 items which are completely separate after the cuts have been made. (*Hint:* Define $X_i = 0$ if item i is not completely separate, and $X_i = 1$ if item i is completely separate.)

4.4.10 The inhabitants of the beautiful and ancient canal city of Pentapolis live on 5 islands separated from each other by water. Bridges cross from one island to another as shown below:



On any given day, a bridge can be closed, with probability p, for restoration work. Assuming that the 8 bridges are closed independently of each other, determine the mean and variance of the number of islands which are completely cut off because of restoration work.

4.5 Conditional Probability Distributions

One of the most useful concepts in probability theory is that of conditional probability and conditional expectation. There are two main reasons. First of all, in practice, we are often interested in calculating probabilities and expectations when some partial information is available, resulting in conditional probabilities and expectations. Secondly, in calculating a desired probability or expectation, it is often very helpful to first condition on some appropriate random variable.

As a starting point, we begin by extending a definition from events to random variables. For events A and B, recall that $P(A|B) = \frac{P(A\cap B)}{P(B)}$ provided that $P(B) > 0$. In the context of discrete random variables X and Y, we can simply apply this definition of conditional probability by choosing $A = \{X = x\}$ and $B = \{Y = y\}$. This gives rise to the following definition.

Definition 4.5.1. *Suppose that X and Y are discrete random variables with joint pmf $f(x, y)$. Let $f_X(x)$ be the marginal pmf of X and let $f_Y(y)$ be the marginal pmf of Y. The* ***conditional probability mass function (pmf) of X given $Y = y$*** *is*

$$f_{X|Y}(x|y) = P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{f(x, y)}{f_Y(y)},$$

provided that $f_Y(y) > 0$. Similarly, the ***conditional pmf of Y given $X = x$*** *is*

$$f_{Y|X}(y|x) = P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{f(x, y)}{f_X(x)},$$

provided that $f_X(x) > 0$.

Remarks:

(1) Clearly, $f_{X|Y}(x|y) \geq 0$ and $f_{Y|X}(y|x) \geq 0$. Moreover, it is straightforward to verify that

$$\sum_{all\ x} f_{X|Y}(x|y) = 1 \text{ and } \sum_{all\ y} f_{Y|X}(y|x) = 1.$$

(2) If X and Y are *independent* random variables, then $f(x,y) = f_X(x)f_Y(y)$ for all (x,y), and so $f_{X|Y}(x|y) = f_X(x)$ and $f_{Y|X}(y|x) = f_Y(y)$.

(3) These ideas extend beyond the simple bivariate case in a natural way. For example, suppose that X_1, X_2, and X_3 are discrete random variables with joint pmf $f(x_1,x_2,x_3)$. We can define the conditional joint pmf of (X_1,X_2) given $X_3 = x_3$ as follows:

$$f_{12|3}(x_1,x_2|x_3) = \frac{P(X_1 = x_1, X_2 = x_2, X_3 = x_3)}{P(X_3 = x_3)} = \frac{f(x_1,x_2,x_3)}{f_3(x_3)},$$

provided that $f_3(x_3) > 0$. Alternatively, we can define the conditional pmf of X_2 given $(X_1 = x_1, X_3 = x_3)$ by

$$f_{2|13}(x_2|x_1,x_3) = \frac{f(x_1,x_2,x_3)}{f_{13}(x_1,x_3)},$$

where $f_{13}(x_1,x_3)$ denotes the joint pmf of (X_1,X_3) with $f_{13}(x_1,x_3) > 0$.

With the notion of a conditional probability distribution defined as above, we can naturally introduce the concept of a *conditional mean.*

Definition 4.5.2. *The **conditional mean of** X **given** $Y = y$ is*

$$E(X|Y=y) = \sum_{all\ x} x f_{X|Y}(x|y).$$

*Similarly, the **conditional mean of** Y **given** $X = x$ is*

$$E(Y|X=x) = \sum_{all\ y} y f_{Y|X}(y|x).$$

Remarks:

(1) We can take the definition of conditional mean and extend it more broadly. For example, if $g(X,Y)$ is a function of the random variables X and Y, then we can define

$$E(g(X,Y)|Y=y) = \sum_{all\ x} g(x,y) f_{X|Y}(x|y) = E(g(X,y)|Y=y). \tag{4.5.1}$$

If we now take $g(X,Y) = h_1(X)h_2(Y)$ where h_1 and h_2 are arbitrary real-valued functions, then (4.5.1) would yield

$$E(h_1(X)h_2(Y)|Y=y) = E(h_1(X)\underbrace{h_2(y)}_{\text{a constant}}|Y=y) = h_2(y)E(h_1(X)|Y=y).$$

(2) Suppose once again that X_1, X_2, and X_3 are discrete random variables with joint pmf $f(x_1, x_2, x_3)$. In Section 4.3, we saw that

$$E(X_1 + X_2) = \sum_{all\ x_1} \sum_{all\ x_2} (x_1 + x_2) f_{12}(x_1, x_2),$$

where $f_{ij}(x_i, x_j)$ denotes the joint pmf of (X_i, X_j) for $i \neq j$. If we let $f_{12|3}(x_1, x_2|x_3)$ denote the conditional joint pmf of (X_1, X_2) given $X_3 = x_3$, then it correspondingly follows that

$$\begin{aligned}
E(X_1 + X_2|X_3 = x_3) &= \sum_{all\ x_1} \sum_{all\ x_2} (x_1 + x_2) f_{12|3}(x_1, x_2|x_3) \\
&= \sum_{all\ x_1} \sum_{all\ x_2} (x_1 + x_2) \frac{f(x_1, x_2, x_3)}{f_3(x_3)} \\
&= \sum_{all\ x_1} \sum_{all\ x_2} x_1 \cdot \frac{f(x_1, x_2, x_3)}{f_3(x_3)} + \sum_{all\ x_1} \sum_{all\ x_2} x_2 \cdot \frac{f(x_1, x_2, x_3)}{f_3(x_3)} \\
&= \sum_{all\ x_1} \frac{x_1}{f_3(x_3)} \sum_{all\ x_2} f(x_1, x_2, x_3) + \sum_{all\ x_2} \frac{x_2}{f_3(x_3)} \sum_{all\ x_1} f(x_1, x_2, x_3) \\
&= \sum_{all\ x_1} \frac{x_1}{f_3(x_3)} f_{13}(x_1, x_3) + \sum_{all\ x_2} \frac{x_2}{f_3(x_3)} f_{23}(x_2, x_3) \\
&= \sum_{all\ x_1} x_1 f_{1|3}(x_1|x_3) + \sum_{all\ x_2} x_2 f_{2|3}(x_2|x_3) \\
&= E(X_1|X_3 = x_3) + E(X_2|X_3 = x_3).
\end{aligned}$$

In other words, conditional expectation possesses the same linearity properties that the "standard" (i.e., unconditional) expected value operator possesses. In fact, more generally, if $c_1, c_2, \ldots, c_n$ are real constants and $g_1, g_2, \ldots, g_n$ are arbitrary real-valued functions, then the same essential approach above can be used to show that

$$E\left(\sum_{i=1}^{n} c_i g_i(X_i) \middle| Y = y \right) = \sum_{i=1}^{n} c_i E(g_i(X_i)|Y = y).$$

Example 4.5.1. Suppose that the joint pmf of (X, Y) is given by the following table:

		x		
$f(x,y)$		2	4	6
	0	$\frac{1}{15}$	$\frac{1}{5}$	$\frac{1}{15}$
y	1	$\frac{1}{15}$	$\frac{1}{5}$	$\frac{2}{15}$
	2	$\frac{1}{5}$	$\frac{1}{15}$	0

Determine the conditional pmf of X given $Y = 1$ and calculate its conditional mean. In addition, calculate $E((Y + 1)^2|T = 6)$ where $T = X + 2Y$.

Solution: First of all, we wish to determine the conditional pmf of X given $Y = 1$, which is given by

$$f_{X|Y}(x|1) = \frac{f(x,1)}{f_Y(1)}.$$

Using the above joint pmf table, note that the $f_Y(1) = \frac{1}{15} + \frac{1}{5} + \frac{2}{15} = \frac{2}{5}$. Furthermore, we have:

$$\begin{aligned}
f_{X|Y}(2|1) &= P(X = 2|Y = 1) = \frac{f(2,1)}{f_Y(1)} = \frac{\frac{1}{15}}{\frac{2}{5}} = \frac{1}{6}, \\
f_{X|Y}(4|1) &= P(X = 4|Y = 1) = \frac{f(4,1)}{f_Y(1)} = \frac{\frac{1}{5}}{\frac{2}{5}} = \frac{1}{2}, \\
f_{X|Y}(6|1) &= P(X = 6|Y = 1) = \frac{f(6,1)}{f_Y(1)} = \frac{\frac{2}{15}}{\frac{2}{5}} = \frac{1}{3}.
\end{aligned}$$

Therefore, the conditional pmf of X given $Y = 1$ can be represented as follows:

x	2	4	6
$f_{X\|Y}(x\|1)$	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{1}{3}$

Moreover, its conditional mean is given by

$$E(X|Y = 1) = \sum_{all\ x} x f_{X|Y}(x|1) = 2\left(\frac{1}{6}\right) + 4\left(\frac{1}{2}\right) + 6\left(\frac{1}{3}\right) = \frac{13}{3}.$$

Next, we turn our attention to first finding the conditional pmf of Y given $T = 6$. We obtain this conditional pmf as follows:

$$\begin{aligned}
f_{Y|T}(y|6) &= P(Y = y|T = 6) \\
&= \frac{P(Y = y, T = 6)}{P(T = 6)} \\
&= \frac{P(X + 2Y = 6, Y = y)}{P(X = 6, Y = 0) + P(X = 4, Y = 1) + P(X = 2, Y = 2)} \\
&= \frac{P(X + 2y = 6, Y = y)}{f(6,0) + f(4,1) + f(2,2)} \\
&= \frac{P(X = 6 - 2y, Y = y)}{\frac{1}{15} + \frac{1}{5} + \frac{1}{5}} \text{ using the joint pmf table} \\
&= \frac{15}{7} f(6 - 2y, y). \qquad (4.5.2)
\end{aligned}$$

Substituting the values of $y = 0, 1, 2$ into (4.5.2) yields the following representation for the conditional pmf of Y given $T = 6$:

y	0	1	2
$f_{Y\|T}(y\|6)$	$\frac{1}{7}$	$\frac{3}{7}$	$\frac{3}{7}$

With the above conditional pmf, we can calculate the desired conditional expectation:

$$E((Y+1)^2|T=6) = \sum_{y=0}^{2}(y+1)^2 f_{Y|T}(y|6) = (1)^2\left(\frac{1}{7}\right) + (2)^2\left(\frac{3}{7}\right) + (3)^2\left(\frac{3}{7}\right) = \frac{40}{7}.$$

■

Example 4.5.2. For $i = 1, 2$, suppose that $X_i \sim Bin(n_i, p)$ where X_1 and X_2 are independent. Find the conditional pmf of X_1 given $X_1 + X_2 = m$ as well as its conditional mean.

Solution: We want to determine the conditional pmf of X_1 given $Y = m$, where we define $Y = X_1 + X_2$. Let this conditional pmf be denoted by $f_{X_1|Y}(x_1|m) = P(X_1 = x_1|Y = m)$. Recalling from Section 4.3 the distributional result that $X_1 + X_2 \sim Bin(n_1 + n_2, p)$, we obtain

$$\begin{aligned}
f_{X_1|Y}(x_1|m) &= \frac{P(X_1 = x_1, X_1 + X_2 = m)}{P(X_1 + X_2 = m)} \\
&= \frac{P(X_1 = x_1, X_2 = m - x_1)}{\binom{n_1+n_2}{m}p^m(1-p)^{n_1+n_2-m}} \\
&= \frac{f(x_1, m - x_1)}{\binom{n_1+n_2}{m}p^m(1-p)^{n_1+n_2-m}} \text{ where } f(x_1, x_2) \text{ denotes the joint pmf of } (X_1, X_2) \\
&= \frac{f_1(x_1)f_2(m - x_1)}{\binom{n_1+n_2}{m}p^m(1-p)^{n_1+n_2-m}} \text{ since } X_1 \text{ and } X_2 \text{ are independent} \\
&= \frac{\binom{n_1}{x_1}p^{x_1}(1-p)^{n_1-x_1} \cdot \binom{n_2}{m-x_1}p^{m-x_1}(1-p)^{n_2-(m-x_1)}}{\binom{n_1+n_2}{m}p^m(1-p)^{n_1+n_2-m}} \text{ provided that } 0 \le x_1 \le n_1 \text{ and } 0 \le m - x_1 \le n_2 \\
&= \frac{\binom{n_1}{x_1}\binom{n_2}{m-x_1}}{\binom{n_1+n_2}{m}} \text{ for } x_1 = \max\{0, m - n_2\}, \ldots, \min\{n_1, m\}. \qquad (4.5.3)
\end{aligned}$$

Based on the form of (4.5.3), we recognize that X_1 given $X_1 + X_2 = m$ is distributed according to the $HG(n_1 + n_2, n_1, m)$ distribution. As a result, it immediately follows that the conditional mean is given by

$$E(X_1|X_1 + X_2 = m) = \frac{mn_1}{n_1 + n_2}.$$

■

Remark: The result that X_1 given $X_1 + X_2 = m$ has a hypergeometric distribution should not be all that surprising. To see this, consider the sequence of $n_1 + n_2$ Bernoulli trials represented visually by Figure 4.5.1. Of these $n_1 + n_2$ trials in which m of them are known to be successes, we want x_1 of those successes to have occurred among the first n_1 trials (thereby implying that $m - x_1$ successes are obtained during the final n_2 trials). Any of these trials are equally likely to be a success (since the success probability p does not change from trial to trial), and so the desired result ends up being the obtained hypergeometric probability.

$$\boxed{1}\ \boxed{2}\ \cdots\ \boxed{n_1}\ \boxed{1}\ \boxed{2}\ \cdots\ \boxed{n_2}$$

$$n_1 + n_2 \text{ trials}$$

Figure 4.5.1: A sequence of Bernoulli trials

Example 4.5.3. Let X and Y be random variables such that $X \sim Poi(\mu)$ and Y given $X = x$ follows a $Bin(x, p)$ distribution. Find the conditional pmf of X given $Y = y$.

Solution: We want to determine the conditional pmf of X given $Y = y$, to be denoted by

$$f_{X|Y}(x|y) = P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}. \tag{4.5.4}$$

First of all, note that

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)},$$

which implies that

$$\begin{aligned} P(X = x, Y = y) &= P(Y = y|X = x)P(X = x) \\ &= \frac{e^{-\mu}\mu^x}{x!} \cdot \binom{x}{y} p^y (1-p)^{x-y} \text{ for } x = 0, 1, 2, \ldots \text{ and } y = 0, 1, \ldots, x. \end{aligned} \tag{4.5.5}$$

Note that the range of y clearly depends on the values of x. A graphical display of this region is depicted in Figure 4.5.2.

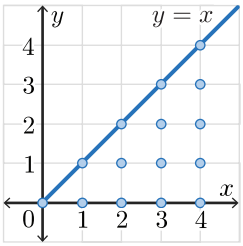


Figure 4.5.2: Set of (x, y) points where the joint pmf of (X, Y) is defined for Example 4.5.3

We may rewrite this region with the range of x depending on the values of y. Specifically, note that $x = 0, 1, 2, \ldots$ and $y = 0, 1, \ldots, x$ is equivalent to $y = 0, 1, 2, \ldots$ and $x = y, y + 1, y + 2, \ldots$. We use this

alternative representation to find the marginal pmf of Y:

$$\begin{aligned}
P(Y=y) &= \sum_{all\ x} P(X=x, Y=y) \\
&= \sum_{x=y}^{\infty} e^{-\mu}\frac{\mu^x}{x!}\binom{x}{y}p^y(1-p)^{x-y} \\
&= \frac{e^{-\mu}(\mu p)^y}{y!}\sum_{x=y}^{\infty}\frac{(\mu(1-p))^{x-y}}{(x-y)!} \\
&= \frac{e^{-\mu}(\mu p)^y}{y!}\underbrace{\sum_{z=0}^{\infty}\frac{(\mu(1-p))^z}{z!}}_{\text{exponential power series}} \quad \text{if we let } z = x-y \\
&= \frac{e^{-\mu}(\mu p)^y}{y!}e^{\mu(1-p)} \\
&= \frac{e^{-\mu p}(\mu p)^y}{y!} \text{ for } y = 0, 1, 2, \ldots. \qquad (4.5.6)
\end{aligned}$$

Interestingly, we see that $Y \sim Poi(\mu p)$. Therefore, if we now substitute (4.5.5) and (4.5.6) into (4.5.4), we ultimately obtain

$$f_{X|Y}(x|y) = \frac{\frac{e^{-\mu}\mu^x}{x!}\cdot\frac{x!}{y!(x-y)!}p^y(1-p)^{x-y}}{\frac{e^{-\mu p}(\mu p)^y}{y!}} = \frac{e^{-\mu(1-p)}(\mu(1-p))^{x-y}}{(x-y)!} \text{ for } x = y, y+1, y+2, \ldots.$$

The above conditional pmf is recognized as that of a *shifted* (y units to the right) Poisson distribution. Specifically, the conditional random variable X given $Y = y$ has the same probability distribution as $W+y$, where $W \sim Poi(\mu(1-p))$. Making use of this keen observation, we easily obtain the conditional mean as

$$E(X|Y=y) = E(W+y) = E(W) + y = \mu(1-p) + y.$$

■

Example 4.5.4. Suppose that $(X_1, X_2, \ldots, X_k) \sim MN(n; p_1, p_2, \ldots, p_k)$. Determine the conditional joint probability distribution of $(X_1, X_2, \ldots, X_{k-1})$ given $X_k = x_k$.

Solution: For the sake of notational convenience, let $g(x_1, x_2, \ldots, x_{k-1}|x_k)$ denote the conditional joint pmf of $(X_1, X_2, \ldots, X_{k-1})$ given $X_k = x_k$. If $f(x_1, x_2, \ldots, x_k)$ denotes the joint pmf of $(X_1, X_2, \ldots, X_k)$,

then it follows that

$$\begin{aligned}
g(x_1, x_2, \ldots, x_{k-1}|x_k) &= \frac{f(x_1, x_2, \ldots, x_{k-1}, x_k)}{f_k(x_k)} \\
&= \frac{f(x_1, x_2, \ldots, x_{k-1}, x_k)}{\binom{n}{x_k} p_k^{x_k}(1-p_k)^{n-x_k}} \text{ since } X_k \sim Bin(n, p_k) \\
&= \frac{n!}{x_1!x_2!\cdots x_{k-1}!x_k!} p_1^{x_1} p_2^{x_2} \cdots p_{k-1}^{x_{k-1}} p_k^{x_k} \cdot \frac{x_k!(n-x_k)!}{n! p_k^{x_k}(1-p_k)^{n-x_k}} \\
&= \frac{(n-x_k)!}{x_1!x_2!\cdots x_{k-1}!} \cdot \frac{p_1^{x_1} p_2^{x_2} \cdots p_{k-1}^{x_{k-1}}}{(1-p_k)^{x_1+x_2+\cdots+x_{k-1}+x_k-x_k}} \text{ since } x_1 + x_2 + \cdots + x_{k-1} + x_k = n \\
&= \frac{(n-x_k)!}{x_1!x_2!\cdots x_{k-1}!} \left(\frac{p_1}{1-p_k}\right)^{x_1} \left(\frac{p_2}{1-p_k}\right)^{x_2} \cdots \left(\frac{p_{k-1}}{1-p_k}\right)^{x_{k-1}},
\end{aligned}$$

where $x_i = 0, 1, \ldots, n - x_k$ for each $i = 1, 2, \ldots, k-1$ and $\sum_{i=1}^{k-1} x_i = n - x_k$. Based on the above form of $g(x_1, x_2, \ldots, x_{k-1}|x_k)$, we conclude that the conditional joint probability distribution of $(X_1, X_2, \ldots, X_{k-1})$ given $X_k = x_k$ is $MN(n^*, p_1^*, p_2^*, \ldots, p_{k-1}^*)$ where $n^* = n - x_k$ and

$$p_i^* = \frac{p_i}{1-p_k} \text{ for } i = 1, 2, \ldots, k-1.$$

As required, note that

$$\sum_{i=1}^{k-1} p_i^* = \frac{1}{1-p_k} \sum_{i=1}^{k-1} p_i = \frac{1}{1-p_k}\left(\underbrace{\sum_{i=1}^{k} p_i}_{=1} - p_k\right) = 1.$$

■

Remarks:

(1) The result of Example 4.5.4 is quite intuitive, in the sense that we obtain another multinomial distribution, but one that is rescaled due to the fact that $X_k = x_k$. Considering the physical setup of the multinomial distribution along with the fact that we now know $X_k = x_k$, what we are seeing happen is that our original n independent trials changes to $n - x_k$ independent trials where one of the outcome types (i.e., the k^{th} type) can no longer occur. As a result, the remaining outcome probabilities are weights which are normalized to account for the fact that type-k outcomes are no longer possible.

(2) In Example 4.5.4, we could replace the condition "$X_k = x_k$" with any $X_i = x_i$ for $i = 1, 2, \ldots, k$, and the same approach would be used to establish an analogous multinomial result.

Section 4.5 Problems

4.5.1 Consider the joint probability distribution given in Problem 4.1.2.

(a) Determine the conditional pmf of X given $Y = 6$ as well as its conditional mean.

(b) Determine the conditional pmf of Y given $X = 2$ as well as its conditional mean.

(c) Determine the conditional pmf of Y given $D = |X - Y| = 1$ as well as its conditional mean.

4.5.2 Consider the joint probability distribution given in Problem 4.1.6.

(a) Determine the conditional pmf of X given $Y = y$ where y is a non-negative integer.

(b) Determine the conditional pmf of Y given $X = x$ where x is a non-negative integer.

(c) Using the results of parts (a) and (b), determine $E(X|Y = y)$ and $E(Y|X = x)$.

4.5.3 Consider the policy holder classification system described in Problem 4.2.1. Among 25 randomly chosen policy holders, calculate the following probabilities:

(a) there are 3 A's and 11 C's given that there are 4 D's.

(b) there are 11 C's given that there is a combined total of 9 B's and D's.

4.5.4 A box contains five yellow and three red balls, from which four balls are drawn one at a time, at random, without replacement. Let X be the number of yellow balls on the first two draws and let Y be the number of yellow balls on all four draws.

(a) Find the joint pmf of (X, Y).

(b) Determine the conditional pmf of X given $Y = y$ where $y \in \{1, 2, 3, 4\}$.

(c) Use the result of part (b) to determine $E(X|Y = y)$.

4.5.5 A box contains three white, six red, and five black balls. Six of these balls are randomly selected without replacement from the box. Let X and Y denote the number of white and black balls selected, respectively.

(a) Determine the conditional pmf of X given $Y = 1$.

(b) Calculate $E(X|Y = 1)$.

(c) Redo parts (a) and (b) but under the assumption that when a ball is selected, its colour is noted and it is then replaced in the box before the next selection is made.

4.5.6 For $i = 1, 2$, suppose that $X_i \sim Poi(\mu_i)$ where X_1 and X_2 are independent. Find the conditional pmf of X_1 given $X_1 + X_2 = m$ as well as its conditional mean.

4.5.7 For $i = 1, 2, 3$, suppose that X_i has pmf given by

$$f_i(x_i) = p(1-p)^{x_i - 1} \text{ for } x_i = 1, 2, 3, \ldots.$$

Assume that X_1, X_2, and X_3 are independent random variables.

(a) Show that X_1 given $X_1 + X_2 = n$ is distributed according to the $DU(1, n-1)$ distribution.

(b) For $i = 1, 2, 3$, determine the conditional pmf of X_i given $X_1 + X_2 + X_3 = n$.

(c) Use the result of part (b) to determine $E(X_i | X_1 + X_2 + X_3 = n)$ for $i = 1, 2, 3$.

4.6 Law of Total Expectation

The law of total expectation is a fundamental result in probability theory which highlights how the overall expected value of a random variable can be influenced by different possible outcomes and their probabilities. This result is particularly useful when dealing with complex problems that can be simplified by conditioning on another random variable. As we have seen previously with the use of indicator random variables in the previous section, oftentimes taking a problem and breaking it down into smaller pieces turns out to be a better strategy than trying to solve the original problem.

We begin by making a simple but important observation regarding the conditional expected value $E(g(X)|Y = y)$, where g is an arbitrary real-valued function. The key observation is this: *The quantity $E(g(X)|Y = y)$, generally speaking, depends on the conditioned value y, and as such, $E(g(X)|Y = y)$ would be some function of y.* We have seen instances of this. In Example 4.5.2 from the previous section, we found that

$$E(X_1 | X_1 + X_2 = m) = \frac{mn_1}{n_1 + n_2},$$

which depends on the conditioned value of m. In Example 4.5.3, we found that

$$E(X|Y = y) = \mu(1-p) + y,$$

which depends on the conditioned value of y. Therefore, in general, we recognize that $E(g(X)|Y = y) = v(y)$, where $v(y)$ is some function of y.

With this in mind, let us make the following definition:

$$E(g(X)|Y) = E(g(X)|Y = y)\Big|_{y=Y} = v(y)\Big|_{y=Y} = v(Y).$$

In other words, we simply replace any instance of a lower-case y in the expression for $E(g(X)|Y = y)$ with an upper-case Y. However, capital letters are reserved for random variables, and functions of random variables are, once again, random variables themselves. Therefore, $E(g(X)|Y)$ is a random variable, and being a random variable, it would make sense to consider its expected value. On this point, we would obtain

$$E(E(g(X)|Y)) = E(v(Y)) = \sum_{all\ y} v(y) f_Y(y) = \sum_{all\ y} E(g(X)|Y = y) f_Y(y). \tag{4.6.1}$$

This brings us to the following important result, which is regarded as the law of total expectation.

Theorem 4.6.1. ***(Law of Total Expectation):*** *Let g be an arbitrary real-valued function. For random variables X and Y, $E(g(X)) = E(E(g(X)|Y))$.*

Proof: Using (4.6.1) as a starting point, note that

$$\begin{aligned}
E(E(g(X)|Y)) &= \sum_{all\ y} E(g(X)|Y = y) f_Y(y) \\
&= \sum_{all\ y} \left\{ \sum_{all\ x} g(x) f_{X|Y}(x|y) \right\} f_Y(y) \\
&= \sum_{all\ y} \sum_{all\ x} g(x) \frac{f(x,y)}{f_Y(y)} f_Y(y) \\
&= \sum_{all\ x} \sum_{all\ y} g(x) f(x,y) \\
&= \sum_{all\ x} g(x) \left\{ \sum_{all\ y} f(x,y) \right\} \\
&= \sum_{all\ x} g(x) f_X(x) \\
&= E(g(X)).
\end{aligned}$$

■

Remark: Simply put, the law of total expectation states that the expected value of a random variable can be calculated by considering all possible outcomes of another random variable that may influence it. This principle breaks down the overall expectation into manageable parts based on the conditional expectations given various scenarios, highlighting the relationship between different random variables and their influence on outcomes. Its usefulness is well-demonstrated in the next two examples.

Example 4.6.1. You need to pick up some groceries and there are 3 supermarkets to choose from on your way home. If you choose supermarket 1, then the number of items you will purchase is a Poisson random variable with mean 9.4. If you choose supermarket 2, then the number of items you will purchase is a Poisson random variable with mean 11.9. If you choose supermarket 3, then the number of items you will purchase is a Poisson random variable with mean 13.5. Assuming that you are twice as likely to choose supermarket 1 over each of the other two supermarkets, calculate the mean and variance of the number of items purchased.

Solution: Let X be the number of items purchased. Now, define a random variable Y such that

$$Y = \begin{cases} 1 & \text{if supermarket 1 is chosen,} \\ 2 & \text{if supermarket 2 is chosen, and} \\ 3 & \text{if supermarket 3 is chosen.} \end{cases}$$

Since supermarket 1 is twice as likely to be chosen over each of supermarkets 2 and 3, it follows that the pmf of Y is given by

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

Based on the above formulation, it follows that X given $Y = 1$ has a $Poi(9.4)$ distribution, X given $Y = 2$ has a $Poi(11.9)$ distribution, and X given $Y = 3$ has a $Poi(13.5)$ distribution. For convenience, let W_i be a $Poi(\mu_i)$ random variable with $\mu_1 = 9.4$, $\mu_2 = 11.9$, and $\mu_3 = 13.5$. Therefore, by the law of total expectation, we obtain

$$\begin{aligned} E(X) &= E(X|Y=1)f_Y(1) + E(X|Y=2)f_Y(2) + E(X|Y=3)f_Y(3) \\ &= E(W_1)\left(\frac{1}{2}\right) + E(W_2)\left(\frac{1}{4}\right) + E(W_3)\left(\frac{1}{4}\right) \\ &= (9.4)\left(\frac{1}{2}\right) + (11.9)\left(\frac{1}{4}\right) + (13.5)\left(\frac{1}{4}\right) \\ &= 11.05 \end{aligned}$$

and

$$\begin{aligned} E(X^2) &= E(X^2|Y=1)f_Y(1) + E(X^2|Y=2)f_Y(2) + E(X^2|Y=3)f_Y(3) \\ &= E(W_1^2)\left(\frac{1}{2}\right) + E(W_2^2)\left(\frac{1}{4}\right) + E(W_3^2)\left(\frac{1}{4}\right) \\ &= \Big(Var(W_1) + (E(W_1))^2\Big)\left(\frac{1}{2}\right) + \Big(Var(W_2) + (E(W_2))^2\Big)\left(\frac{1}{4}\right) + \Big(Var(W_3) + (E(W_3))^2\Big)\left(\frac{1}{4}\right) \\ &= (97.76)\left(\frac{1}{2}\right) + (153.51)\left(\frac{1}{4}\right) + (195.75)\left(\frac{1}{4}\right) \\ &= 136.195. \end{aligned}$$

Hence,

$$Var(X) = E(X^2) - (E(X))^2 = 136.195 - (11.05)^2 = 14.0925.$$

■

Example 4.6.2. Suppose that $X \sim Geo(p)$. Calculate $E(X)$ and $Var(X)$ using the law of total expectation.

Solution: Recall that if $X \sim Geo(p)$, then X models the number of failures observed before the occurrence of the first success. Let us define the indicator random variable Y as follows:

$$\begin{aligned} Y &= 0 \text{ if the } 1^{\text{st}} \text{ trial is a failure, and} \\ Y &= 1 \text{ if the } 1^{\text{st}} \text{ trial is a success.} \end{aligned}$$

Clearly, $Y \sim Bin(1, p)$, so that $f_Y(0) = 1 - p$ and $f_Y(1) = p$. By the law of total expectation, we obtain

$$E(X) = E(E(X|Y)) = \sum_{y=0}^{1} E(X|Y = y) f_Y(y) = (1 - p)E(X|Y = 0) + pE(X|Y = 1). \tag{4.6.2}$$

Let us consider the conditional random variable X given $Y = 1$. The event $\{Y = 1\}$ means that the first trial is a success, which in turn implies that the number of failures observed before this first success would surely be equal to 0. In other words, X given $Y = 1$ is equal to 0 with probability 1.

On the other hand, looking at the conditional random variable X given $Y = 0$, the first trial is now known to be a failure. Therefore, the number of failures observed is guaranteed to be equal to 1 plus however many failures are observed, starting from the second trial onwards, until the first success occurs. But trials are independent of each other, and waiting for the first success to occur starting from the second trial looks exactly like the original problem in which we wish to count the total number of failures observed until the first success occurs. Based on this reasoning, we conclude that X given $Y = 0$ has the same probability distribution as the random variable $1 + X$.

As a result of the above observations, (4.6.2) now becomes

$$\begin{aligned} E(X) &= (1 - p)E(1 + X) + p \cdot 0 \\ &= (1 - p)(1 + E(X)) \\ &= (1 - p)E(X) + 1 - p. \end{aligned}$$

Rearranging the above equation for $E(X)$ leads to

$$(1 - (1 - p))E(X) = 1 - p,$$

or simply

$$E(X) = \frac{1-p}{p}.$$

In a similar fashion, the law of total expectation also yields

$$\begin{aligned} E(X^2) &= E(E(X^2|Y)) \\ &= (1-p)E(X^2|Y=0) + pE(X^2|Y=1) \\ &= (1-p)E((1+X)^2) + p \cdot 0^2 \\ &= (1-p)E(X^2+2X+1) \\ &= (1-p)E(X^2) + 2(1-p)E(X) + 1 - p \\ &= (1-p)E(X^2) + \frac{2(1-p)^2}{p} + 1 - p, \end{aligned}$$

which implies that

$$(1-(1-p))E(X^2) = \frac{2(1-p)^2 + p(1-p)}{p} = \frac{(1-p)(2-p)}{p},$$

or simply

$$E(X^2) = \frac{(1-p)(2-p)}{p^2}.$$

Finally,

$$Var(X) = \frac{(1-p)(2-p)}{p^2} - \left(\frac{1-p}{p}\right)^2 = \frac{(1-p)(2-p-1+p)}{p^2} = \frac{1-p}{p^2}.$$

■

Remarks:

(1) Note that the obtained mean and variance in Example 4.6.2 agree with known results from Section 3.5 concerning the geometric distribution. Moreover, the above procedure relied only on basic manipulations and did not involve any complicated sums or the differentiation of an mgf.

(2) As part of the solution of Example 4.6.2, we argued that X given $Y = 0$ was identically distributed to the random variable $Z = 1 + X$, and this in turn implied that $E(X^2|Y = 0) = E((1+X)^2)$. To see why this holds true formally, consider first that

$$f_{X|Y}(x|0) = P(X = x|Y = 0) = \frac{P(X = x, Y = 0)}{P(Y = 0)} = \frac{P(X = x, Y = 0)}{1-p}.$$

For the numerator, note that

$$\begin{aligned} & P(X = x, Y = 0) \\ = \; & P(1^{\text{st}} \text{ trial is a failure and } x \text{ total failures observed before } 1^{\text{st}} \text{ success occurs}) \\ = \; & P(1^{\text{st}} \text{ trial is a failure, next } x-1 \text{ trials are failures, and } (x+1)^{\text{th}} \text{ trial is a success}) \\ = \; & (1-p)(1-p)^{x-1}p \text{ due to independence of trials.} \end{aligned}$$

Thus,

$$f_{X|Y}(x|0) = \frac{(1-p)(1-p)^{x-1}p}{1-p} = (1-p)^{x-1}p \text{ for } x = 1, 2, 3, \ldots.$$

On the other hand, since $X \sim Geo(p)$, the pmf of Z is simply given by

$$\begin{aligned} f_Z(z) &= P(Z = z) \\ &= P(1 + X = z) \\ &= P(X = z - 1) \\ &= (1-p)^{z-1}p \text{ for } z = 1, 2, 3, \ldots. \end{aligned}$$

Since these two probability mass functions are identical, it follows that X given $Y = 0$ has the same probability distribution as the random variable Z. As a further consequence, for an arbitrary real-valued function g, we must have that

$$\begin{aligned} E(g(X)|Y = 0) &= \sum_{all\ x} g(x) f_{X|Y}(x|0) \\ &= \sum_{all\ x} g(x) f_Z(x) \\ &= E(g(Z)) \\ &= E(g(1 + X)). \end{aligned}$$

Section 4.6 Problems

4.6.1 Suppose that a and b are positive real constants. Let $X \sim Poi(a)$ and define $Y = X + b$. Suppose that Z is a random variable such that Z given $Y = y$ follows a $Poi(y^2 + ab)$ distribution. Use the law of total expectation to determine an expression for $E(Z)$ in terms of a and b.

4.6.2 Suppose that weather on any February day is classified as being one of three possible types: clear, rainy, or snowy. On a clear February day, the number of traffic accidents in Waterloo is a Poisson random variable with a mean of 2.5. The number of traffic accidents in Waterloo on a rainy February day has a $Bin(8, 1/2)$ distribution, whereas on a snowy February day it has a

$Geo(1/7)$ distribution. If it is equally likely that the weather tomorrow will be clear, rainy, or snowy, what is the mean and standard deviation of the number of traffic accidents in Waterloo tomorrow?

4.6.3 A rat is placed at the start of a maze having four doors. If the rat selects door i, $i = 1, 2, 3$, it will travel around the maze and return to where it started after a number of minutes determined by a $Poi(0.5i)$ random variable. If the rat selects door 4, it will find the hidden cheese after travelling exactly 10 seconds. Assume that the rat is always equally likely to select one of the four doors, and that the random time to return to the start of the maze after selecting any of doors $1, 2$, or 3 is independent of all past door selections and travel times. Calculate the mean and variance of the total time in minutes that the rat will travel before finding the cheese.

4.6.4 Consider a deck of k shuffled cards (labelled $1, 2, \ldots, k$) in which one card is randomly selected. If the selected card is not the number 1, the card is placed back among the other cards, the deck is reshuffled, and another card is randomly chosen. This process continues until the card numbered 1 is selected. Let Y be the number of card selections made. Following that, a coin (with probability p of coming up heads) is repeatedly flipped until the number of heads obtained equals the number of cards selected. On average, how many coin flips will be made?

4.6.5 Suppose that the number of people who get on an elevator on the ground floor of a building is a Poisson random variable with mean μ. If there are m floors above the ground floor and if each person is equally likely to get off at any one of these m floors, independently of where the others get off, determine an expression for the expected number of stops that the elevator will make before discharging all of its passengers.

Chapter 5

Univariate Continuous Probability Distributions

5.1 Continuous Random Variables

When introducing the concept of a random variable in Chapter 3, we stated that random variables are generally classified into one of two types, according to the size of their range of possible values. In this chapter, we focus our attention on continuous random variables, in which the range of possible values is an interval (or a collection of intervals) on the real number line. As we will see shortly, continuous random variables must be treated a bit differently than discrete random variables, and the primary reason for this is due to the fact that $P(X = x) = 0$ for each value of x in the range of a continuous random variable X.

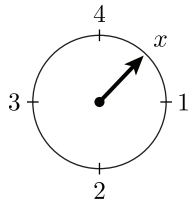


Figure 5.1.1: Spinning pointer used for generating a continuous random variable

To see why this must be the case, consider the simple spinning pointer in Figure 5.1.1. For convenience, let us assume that the spinning pointer operates in a frictionless environment, so that it is equally likely to stop at any point x in the interval $(0, 4]$. The probability of the pointer stopping precisely at point x must be zero, because if each number were to have the same probability $p > 0$,

then the probability of the event $A = \{x : 0 < x \leq 4\}$ would be equal to the sum

$$\sum_{x \in (0,4]} p = \infty,$$

since A is comprised of an uncountably infinite number of values (it can be shown that, in general, the sum of uncountably many positive values diverges). For a continuous random variable, the probability of each individual point is 0 and probability mass functions cannot be used to describe a continuous probability distribution. On the other hand, it seems reasonable that one would assign a probability of $\frac{1}{16}$ to the event that the pointer stops at some value x in the interval $(0, \frac{1}{4}]$ or $[\frac{7}{4}, 2)$. For continuous random variables, this highlights a very important feature: *We specify the probability of intervals, rather than individual points.*

Let us consider another example produced by choosing a "random point" in a region. Specifically, suppose that we plot the graph of a function $f(x)$ as depicted in Figure 5.1.2. Let us further assume that the function is positive and has a finite integral. We generate a point at random by closing our eyes and firing a dart from a distance until one lands in the shaded region under the graph. We assume that such a point, which we denote by "$*$", is "uniformly distributed" under the graph. This means that the point is equally likely to fall in any one of the many possible regions of a given area located in the shaded region, implying that we only need to know the area of a region to determine the probability that a point falls in it. Consider the x-coordinate of the point "$*$" as our random variable X. In Figure 5.1.2, this value appears to be around 6.

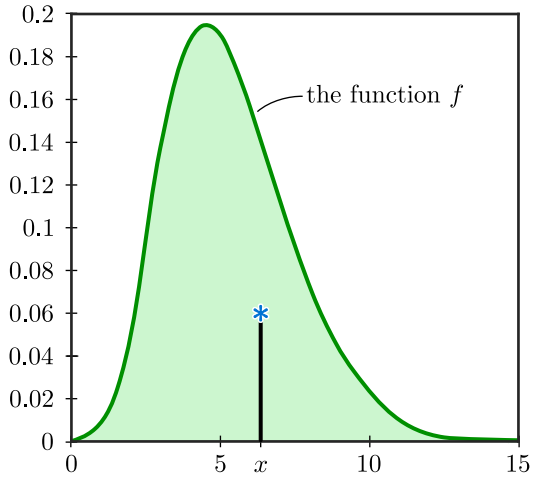


Figure 5.1.2: Graph of a function $f(x)$

Note that the probability that X falls in a particular interval (a, b) is measured by the area of the

region above this interval, which is given by

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

Therefore, the probability of any particular point, $P(X = a)$, is the area of the region immediately above this single point, namely $\int_a^a f(x)dx = 0$. As you can see, this demonstrates another example of a random variable X which has a continuous probability distribution.

For a continuous random variable X, there are two commonly-used functions which describe its probability distribution. The first one we will discuss is the *cumulative distribution function (cdf),* which we have used previously for discrete probability distributions. Recall that for a discrete random variable X, we defined its cdf as $F(x) = P(X \leq x)$. For continuous random variables, we can also define the cdf in the exact same fashion. In the spinning pointer example, if all values $x \in (0, 4]$ are "equally likely", then the probability that the pointer stops between 0 and $\frac{1}{2}$ is $\frac{1}{8}$, between 0 and 1 the probability is $\frac{1}{4}$, between 0 and 2 it is $\frac{1}{2}$, and so on. In general, $F(x) = P(X \leq x) = \frac{x}{4}$ for $0 < x \leq 4$. Moreover, $F(x) = 0$ for $x \leq 0$ since there is no chance of the pointer stopping at a non-positive number, and $F(x) = 1$ for $x > 4$ since the pointer is certain to stop at a number below x if $x > 4$. In the second example where we generated a point at random under the graph of a function $f(x)$, suppose now that the total area under the graph is equal to 1. In this situation, the cdf $F(x)$ would represent the area under the graph but to the left of the point x as shown in Figure 5.1.3.

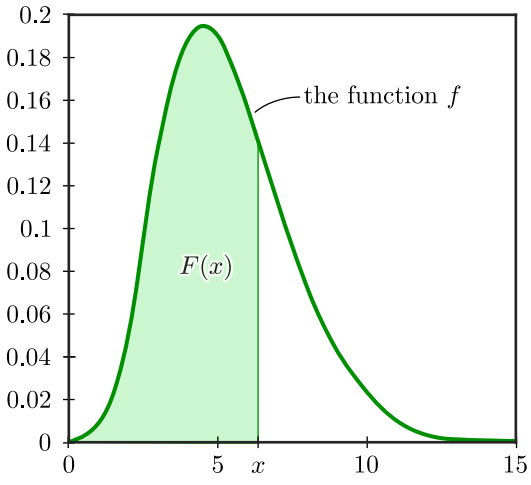


Figure 5.1.3: Area of shaded region equals $F(x) = P(X \leq x)$

Most properties of a cdf are the same for continuous random variables as for discrete random

variables. These properties include:

(1) $F(x)$ is defined for all $x \in \mathbb{R}$,

(2) $F(x)$ is a non-decreasing function of x for all $x \in \mathbb{R}$,

(3) $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$,

(4) $P(a < X \leq b) = F(b) - F(a)$.

For a continuous random variable, as indicated earlier, we have that $P(X = a) = 0$ for any $a \in \mathbb{R}$. Using property (4), note that

$$0 = P(X = a) = \lim_{\epsilon \to 0} P(a - \epsilon < X \leq a) = \lim_{\epsilon \to 0} \left(F(a) - F(a - \epsilon)\right) = F(a) - \lim_{\epsilon \to 0} F(a - \epsilon).$$

This gives us

$$\lim_{\epsilon \to 0} F(a - \epsilon) = F(a).$$

In a similar fashion, we can also show that

$$\lim_{\epsilon \to 0} F(a + \epsilon) = F(a).$$

Therefore, we have that $\lim_{x \to a} F(x) = F(a)$, which implies that the cdf $F(x)$ is *continuous* at the point $x = a$. Since the value of a is arbitrary, the cdf $F(x)$ is a continuous function on the entire real number line. Note that this is in contrast with the behaviour of a cdf for a discrete random variable, which is generally a step function with discontinuities at points which have positive probability.

In addition, since the probability is 0 at each individual point for a continuous random variable, it follows that

$$P(a < X < b) = P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b) = F(b) - F(a).$$

Note that for a discrete random variable, each of these four probabilities could potentially produce a different value. For continuous probability distributions, however, it does not matter whether intervals are open, closed, or half-open since the probability of these intervals is the same.

While the cdf can be used to find probabilities, it does not give an intuitive picture of which values of x are more likely, and which are less likely. To develop such a picture, suppose that we take a short interval of values, say $[x, x + \Delta x]$, from the range of a continuous random variable X. The probability that X lies in this interval is given by

$$P(x \leq X \leq x + \Delta x) = F(x + \Delta x) - F(x).$$

Therefore, the comparison of probabilities for two intervals, each of length Δx, is direct using the above formula. Now suppose we consider what happens as Δx becomes small, and we divide the probability by Δx. This leads to the introduction of the second important function used in the study of continuous random variables.

Definition 5.1.1. *The **probability density function** (pdf) of a continuous random variable X is the derivative of the cdf of X, which is given by*

$$f(x) = F'(x) = \frac{dF(x)}{dx} = \lim_{\Delta x \to 0} \frac{F(x + \Delta x) - F(x)}{\Delta x}.$$

Remarks:

(1) The notation "f" is used for the pdf of a continuous random variable, just as f was used to denote the pmf for a discrete random variable. Similar to the convention we adopted for a pmf, we may sometimes choose to include the name of the random variable in the subscript of its pdf.

(2) If the derivative of $F(x)$ does not exist at $x = a$, then we adopt the convention that $f(a) = 0$.

(3) If the function $f(x)$ graphed in Figure 5.1.3 is actually a pdf and has a total integral (i.e., $\int_{-\infty}^{\infty} f(x)dx$) equal to one, then the cdf (or the area to the left of a point x) is given by

$$F(x) = \int_{-\infty}^{x} f(z)dz.$$

If we take the derivative of this cdf, then the *Fundamental Theorem of Calculus* ensures that

$$F'(x) = \frac{d}{dx}\left(\int_{-\infty}^{x} f(z)dz \right) = f(x).$$

(4) Note that for a small value of Δx, we have

$$P\left(x - \frac{\Delta x}{2} \leq X \leq x + \frac{\Delta x}{2}\right) = F\left(x + \frac{\Delta x}{2}\right) - F\left(x - \frac{\Delta x}{2}\right) \approx f(x)\Delta x.$$

Therefore, $f(x) \neq P(X = x)$ **but** $f(x)\Delta x$ is the *approximate probability* that X is inside an interval of length Δx centered about the value x when Δx is small. A plot of a pdf $f(x)$, such as the one given by Figure 5.1.4, shows this quite clearly.

We now take note of a few of the important properties that a pdf possesses (along with a brief justification as to why they hold true). They include:

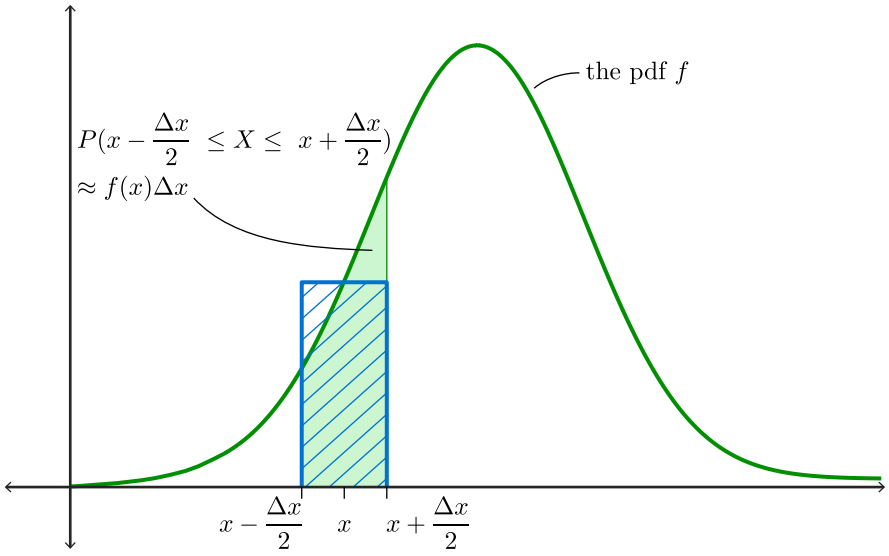


Figure 5.1.4: Visualization that $P\left(x - \frac{\Delta x}{2} \leq X \leq x + \frac{\Delta x}{2}\right) \approx f(x)\Delta x$

(1) $\int_{-\infty}^{\infty} f(x)dx = 1$. *(This is due to the fact that $P(-\infty \leq X \leq \infty) = 1$.)*

(2) $f(x) \geq 0$ for all $x \in \mathbb{R}$. *(Since $F(x)$ is a non-decreasing function, its derivative is non-negative.)*

(3) $P(a \leq X \leq b) = F(b) - F(a) = \int_{a}^{b} f(z)dz$. *(This follows from the definition of $f(x)$.)*

(4) $F(x) = \int_{-\infty}^{x} f(z)dz$. *(This is simply property (3) with $a = -\infty$ and $b = x$.)*

Remark: It is important to realize that even though a pdf $f(x)$ is used to calculate event probabilities associated with a continuous random variable, $f(x)$ is not the probability of any specific event. In particular, $f(x)$ is not even restricted to be less than or equal to 1. As long as properties (1) and (2) above are satisfied, a pdf exceeding the value of 1 is not infeasible as long as it is on a sufficiently small interval length.

Returning to the spinning pointer example, we had reasoned that the cdf of X is given by

$$F(x) = \begin{cases} 0 & \text{for } x \leq 0, \\ \frac{x}{4} & \text{for } 0 < x \leq 4, \\ 1 & \text{for } x > 4. \end{cases}$$

Therefore, the pdf of X is given by

$$f(x) = F'(x) = \frac{1}{4} \text{ for } 0 < x < 4,$$

and outside this interval, the pdf is defined to be 0. Figure 5.1.5 illustrates what the pdf $f(x)$ looks like in this case. Note that the area under $f(x)$, represented by the shaded region in Figure 5.1.5, is equal to 1 (as it is simply the area of a rectangle with base 4 and height 0.25). For obvious reasons, this is an example of a "uniform distribution", which we will study in greater detail later in Section 5.4.1.

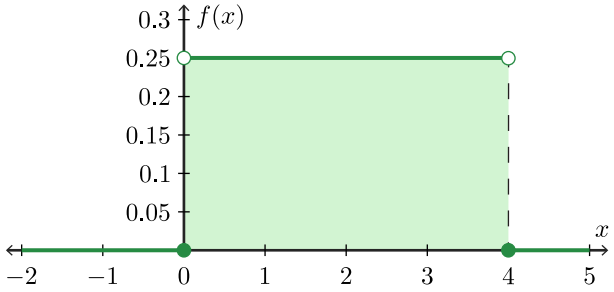


Figure 5.1.5: Plot of the pdf $f(x)$ for the spinning pointer example

Remark: It may seem paradoxical that $P(X = x) = 0$ for a continuous random variable and yet we record the outcomes $X = x$ in real-life "experiments" with continuous variables. The catch is that all measurements have finite precision; they are in effect discrete. For example, the height $60 + \pi$ inches is within the range of the height X of people in a population, but we could never observe the outcome $X = 60 + \pi$ if we selected a person at random and measured their height. In measurements, we are actually observing something like

$$P\left(x - \frac{\epsilon}{2} \leq X \leq x + \frac{\epsilon}{2}\right),$$

where ϵ may be very small, but not zero. The probability of this outcome is **not** zero: it is (approximately) equal to $f(x) \cdot \epsilon$.

At some point in your life, you have most likely been told that you fall within some range or percentage of values with respect to a particular measure. For example, if you are tall, you might have been told that you are in the 95^{th} percentile in height, meaning that you are taller than 95% of the population. Perhaps if you took a college entrance exam, you might have been informed that you placed in the 80^{th} percentile in math ability, meaning that you scored better than 80% of the population on the math portion of the exam. We will now formally define what a *percentile* is within the framework of probability theory.

Definition 5.1.2. *Suppose that X is a continuous random variable with cdf $F(x)$. For $p \in (0, 1)$, the $100p^{\text{th}}$ **percentile** of X (or sometimes called the $100p^{\text{th}}$ percentile of the distribution) is the value $q(p)$ such that $F(q(p)) = p$. If $p = 0.5$, then $m = q(0.5)$ is called the **median** of X (or the median of the distribution).*

Remark: Another name which is frequently used for the $100p^{\text{th}}$ percentile of X is the **p-quantile** of X. Simply put, the p-quantile of X is the point at which $100p\%$ of the distribution of X falls below this point, as shown in Figure 5.1.6.

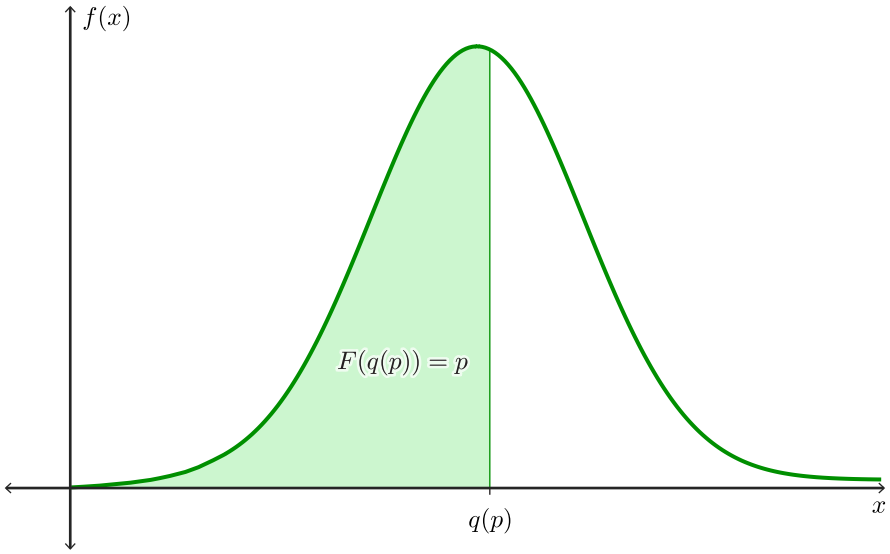


Figure 5.1.6: Visualization of the p-quantile of a continuous random variable

Let us now consider another example demonstrating some of the key concepts we have introduced thus far for continuous random variables.

Example 5.1.1. Let X be a continuous random variable with pdf of the form

$$f(x) = \begin{cases} kx^2 & \text{for } 0 < x < 1, \\ k(2-x) & \text{for } 1 < x < 2, \\ 0 & \text{otherwise.} \end{cases}$$

Determine the following quantities of interest: (i) the constant k, (ii) the cdf of X, (iii) $P(0.5 < X < 1.5)$, (iv) the 0.4-quantile of X, and (v) the median of X.

Solution: Let us begin by finding the value of k. To do so, we make use of the property that $\int_{-\infty}^{\infty} f(x)dx = 1$ for a continuous random variable. However, when finding the area of a region

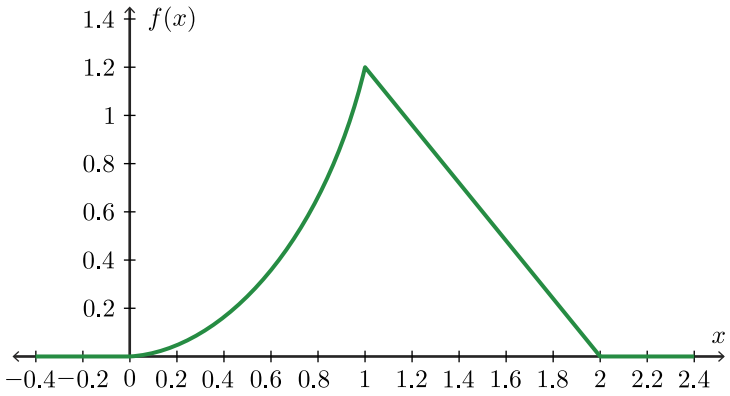


Figure 5.1.7: Plot of the pdf for Example 5.1.1

bounded by different functions, we must split the integral into pieces as follows:

$$\begin{aligned}
1 &= \int_{-\infty}^{\infty} f(x)dx \\
&= \int_{-\infty}^{0} 0dx + \int_{0}^{1} kx^2dx + \int_{1}^{2} k(2-x)dx + \int_{2}^{\infty} 0dx \\
&= 0 + k\left(\frac{x^3}{3}\right)\Big|_0^1 + k\left(2x - \frac{x^2}{2}\right)\Big|_1^2 + 0 \\
&= \frac{k}{3} + \frac{k}{2} \\
&= \frac{5k}{6}.
\end{aligned}$$

Solving for k, we immediately obtain $k = \frac{6}{5}$. The plot of $f(x)$ is given in Figure 5.1.7. Note that over the short interval $(\sqrt{\frac{5}{6}}, \frac{7}{6})$, the value of the pdf is greater than 1.

With $k = \frac{6}{5}$, we next proceed to determine the cdf of X given by $F(x) = P(X \leq x)$. First of all, for $x \leq 0$, note that

$$F(x) = \int_{-\infty}^{x} f(z)dz = \int_{-\infty}^{x} 0dz = 0.$$

Next, for $0 < x \leq 1$, we have

$$F(x) = \int_{-\infty}^{x} f(z)dz = \int_{-\infty}^{0} 0dz + \int_{0}^{x} \frac{6}{5}z^2dz = 0 + \frac{6}{5}\left(\frac{z^3}{3}\right)\Big|_0^x = \frac{2x^3}{5}.$$

For $1 < x \leq 2$, we have

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} f(z)dz \\ &= \int_{-\infty}^{0} 0\,dz + \int_{0}^{1} \frac{6}{5}z^2 dz + \int_{1}^{x} \frac{6}{5}(2-z)\,dz \\ &= 0 + \frac{6}{5}\left(\frac{z^3}{3}\right)\Big|_0^1 + \frac{6}{5}\left(2z - \frac{z^2}{2}\right)\Big|_1^x \\ &= \frac{2}{5} + \frac{6}{5}\left(2x - \frac{x^2}{2} - 2 + \frac{1}{2}\right) \\ &= \frac{12x - 3x^2 - 7}{5}. \end{aligned}$$

Finally, we have that $F(x) = F(2) = 1$ for $x > 2$ since the pdf equals 0 for all $x > 2$. Putting everything together, we ultimately obtain

$$F(x) = \begin{cases} 0 & \text{for } x \leq 0, \\ \frac{2x^3}{5} & \text{for } 0 < x \leq 1, \\ \frac{12x-3x^2-7}{5} & \text{for } 1 < x \leq 2, \\ 1 & \text{for } x > 2. \end{cases}$$

Figure 5.1.8 displays a plot of the cdf of X. Note that it is clearly continuous everywhere.

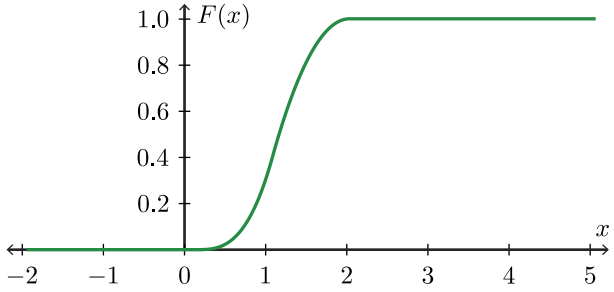


Figure 5.1.8: Plot of the cdf for Example 5.1.1

With $F(x)$ now known, it immediately follows that

$$P(0.5 < X < 1.5) = F(1.5) - F(0.5) = \frac{12(1.5) - 3(1.5)^2 - 7}{5} - \frac{2(0.5)^3}{5} = 0.85 - 0.05 = 0.8.$$

Looking at the formula we derived for $F(x)$, note that $F(1) = \frac{2(1)^3}{5} = 0.4$. Hence, we immediately conclude that $q(0.4)$, the 0.4-quantile of X, is equal to 1. In order to find the median $m = q(0.5)$, we

set $F(m) = 0.5$ and solve for m. We must use the form of $F(x)$ for $1 < x \leq 2$ (since $F(1) = 0.4 < 0.5$), and this leads to

$$\begin{aligned} \frac{12m - 3m^2 - 7}{5} &= 0.5 \\ 24m - 6m^2 - 14 &= 5 \\ 6m^2 - 24m + 19 &= 0. \end{aligned}$$

Applying the quadratic formula to find the roots of the above equation, we obtain

$$\begin{aligned} m &= \frac{24 \pm \sqrt{24^2 - 4(6)(19)}}{2(6)} \\ m &= \frac{24 \pm \sqrt{120}}{12} \\ m &= 2 \pm \frac{\sqrt{4 \cdot 30}}{12} \\ m &= \underbrace{2 - \frac{\sqrt{30}}{6}}_{\approx 1.087} \text{ or } \underbrace{2 + \frac{\sqrt{30}}{6}}_{\approx 2.913}. \end{aligned}$$

Since the median must lie between 1 and 2, we choose the solution $m = 2 - \frac{\sqrt{30}}{6} \approx 1.087$. ■

Section 5.1 Problems

5.1.1 Suppose that X is a continuous random variable with pdf given by

$$f(x) = \begin{cases} k(1-x)^2 & \text{for } -1 < x < 3, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Determine the value of k that makes the above pdf valid.

(b) Determine the cdf of X and plot its graph.

(c) Calculate $P(-0.1 < X < 0.2)$.

(d) Find the 95^{th} percentile of the distribution of X.

5.1.2 Let X be a continuous random variable with cdf of the form

$$F(x) = \begin{cases} 0 & \text{for } x \leq 0, \\ \frac{kx^\alpha}{1+x^\alpha} & \text{for } x > 0, \end{cases}$$

where $\alpha > 0$ is a real-valued parameter of the probability distribution.

(a) Determine the value of k that makes the above cdf valid.

(b) Determine the pdf of X and plot its graph when $\alpha = 1$ and $\alpha = 2$.

(c) Find the median of X.

5.1.3 Suppose that X is a continuous random variable with pdf given by

$$f(x) = \begin{cases} 4x & \text{for } 0 < x < \frac{1}{2}, \\ 4(1-x) & \text{for } \frac{1}{2} < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Graph the pdf of X and justify why $\int_{-\infty}^{\infty} f(x)dx = 1$ without evaluating an integral.

(b) Determine the cdf of X and plot its graph.

(c) Calculate $P(0.25 \leq X < 0.8)$.

(d) Find the median and 10^{th} percentile of the distribution of X.

5.1.4 A continuous random variable X has pdf of the form

$$f(x) = \begin{cases} (\theta+1)\, x^{\theta} & \text{for } 0 < x < 1, \\ 0 & \text{otherwise,} \end{cases}$$

where θ is a real-valued parameter of the probability distribution.

(a) For what values of θ is this a valid pdf? Explain your answer.

(b) Determine $P(X > 0.5)$.

(c) Determine the p-quantile of X where $p \in (0, 1)$.

5.1.5 A continuous random variable X has pdf of the form

$$f(x) = \begin{cases} kxe^{-x^2/\theta} & \text{for } x > 0, \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is a real-valued parameter of the probability distribution.

(a) Determine the value of k that makes the above pdf valid.

(b) Determine the cdf of X.

(c) Calculate $P\left(\sqrt{\theta} \leq X < 2\sqrt{\theta}\right)$.

(d) Find the median of X.

5.2 Functions of Random Variables

When we know the pdf or cdf of a continuous random variable X, we oftentimes want to find the pdf or cdf of some other random variable of interest Y which happens to be a function of X, namely $Y = g(X)$. The approach for doing this is summarized below. It is based on the idea that the cdf of Y can be expressed in terms of the cdf of X since Y is a function of X. More specifically, the general procedure can be described as follows:

(1) Express the cdf of Y, denoted by $F_Y(y)$, in terms of the cdf of X.

(2) Use the known form of the cdf of X to find an explicit formula for $F_Y(y)$. If the pdf of Y is desired, then differentiate the obtained expression for $F_Y(y)$.

(3) Determine the range of possible values for y.

We demonstrate the above procedure with a couple of examples.

Example 5.2.1. In the spinning pointer example from the previous section, we assumed that the random variable X had pdf of the form

$$f_X(x) = \begin{cases} \frac{1}{4} & \text{for } 0 < x < 4, \\ 0 & \text{otherwise,} \end{cases}$$

and cdf of the form

$$F_X(x) = \begin{cases} 0 & \text{for } x \leq 0, \\ \frac{x}{4} & \text{for } 0 < x \leq 4, \\ 1 & \text{for } x > 4. \end{cases}$$

Find the pdf of $Y = \frac{1}{X}$.

Solution: Starting with the cdf of Y, note that

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P\left(X^{-1} \leq y\right) \\ &= P\left(X \geq y^{-1}\right) \\ &= 1 - P\left(X < y^{-1}\right) \\ &= 1 - F_X\left(y^{-1}\right). \end{aligned} \tag{5.2.1}$$

We can now substitute y^{-1} in place of x in $F_X(x)$ to give

$$F_Y(y) = 1 - \frac{y^{-1}}{4} = 1 - \frac{1}{4y},$$

and then differentiate this function to obtain the desired pdf:

$$f_Y(y) = F'_Y(y) = \frac{d}{dy}\left(1 - \frac{1}{4y}\right) = \frac{1}{4y^2} \text{ for } y > \frac{1}{4}.$$

The above range for y is determined by noting that as x goes from 0 to 4, $y = \frac{1}{x}$ goes from ∞ to $\frac{1}{4}$. ■

Remark: Alternatively, and a little more generally, we could have applied the chain rule and differentiated (5.2.1) to get

$$\begin{aligned} f_Y(y) &= \frac{d}{dy}F_Y(y) \\ &= \frac{d}{dy}\left(1 - F_X\left(y^{-1}\right)\right) \\ &= -F'_X(y^{-1}) \cdot \frac{d}{dy}\left(y^{-1}\right) \\ &= -f_X(y^{-1})\left(-y^{-2}\right) \\ &= \frac{1}{4y^2} \text{ for } y > \frac{1}{4}. \end{aligned}$$

As a general rule of thumb, if the cdf $F_X(x)$ is known in some straightforward form, then it is easier to substitute in $F_X(x)$ first and then differentiate. However, if $F_X(x)$ is not readily known or is more complicated to obtain (e.g., an integral that cannot be easily solved), then it is usually easier to differentiate first and then substitute in the form of the pdf $f_X(x)$. This latter approach is illustrated in the next example.

Example 5.2.2. Let X be a continuous random variable with pdf $f_X(x)$ and cdf $F_X(x)$. Determine a general formula for the pdf of $Y = X^2$.

Solution: Since $Y = X^2$, we immediately note that Y must be a non-negative valued continuous random variable. Therefore, for $y \geq 0$, we have

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

If we now apply the chain rule and differentiate the above expression, then we obtain

$$\begin{aligned} f_Y(y) &= F'_Y(y) \\ &= \frac{d}{dy}\Big(F_X(\sqrt{y}) - F_X(-\sqrt{y})\Big) \\ &= F'_X(\sqrt{y}) \cdot \frac{d}{dy}\left(y^{\frac{1}{2}}\right) - F'_X(-\sqrt{y}) \cdot \frac{d}{dy}\left(-y^{\frac{1}{2}}\right) \\ &= f_X(\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} - f_X(-\sqrt{y}) \cdot \frac{-1}{2\sqrt{y}} \\ &= \frac{1}{2\sqrt{y}} f_X(\sqrt{y}) + \frac{1}{2\sqrt{y}} f_X(-\sqrt{y}) \ \text{ for } \ y > 0. \end{aligned} \tag{5.2.2}$$

■

Remark: We emphasize that (5.2.2) is a general formula which applies to any continuous random variable X. For example, if X happens to have pdf of the form

$$f_X(x) = \begin{cases} \frac{1}{4} & \text{for } 0 < x < 4, \\ 0 & \text{otherwise,} \end{cases}$$

then the pdf of $Y = X^2$ given by (5.2.2) simplifies to become

$$f_Y(y) = \frac{1}{2\sqrt{y}} \cdot \frac{1}{4} + \frac{1}{2\sqrt{y}} \cdot 0 = \frac{1}{8\sqrt{y}} \ \text{ for } \ 0 < y < 16.$$

On the other hand, if X is a continuous random variable with pdf

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

then the pdf of $Y = X^2$ is now given by

$$f_Y(y) = \frac{1}{2\sqrt{y}} \cdot \frac{2}{3}(\sqrt{y})^2 + \frac{1}{2\sqrt{y}} \cdot \frac{2}{3}(-\sqrt{y})^2 = \frac{2}{3}\sqrt{y} \ \text{ for } \ 0 < y < 1.$$

For strictly monotonic functions g, it is possible to develop a convenient analytical formula for the pdf of the function $Y = g(X)$.

Theorem 5.2.1. *Let X be a continuous random variable with pdf $f_X(x)$. Suppose that g is a strictly monotonic differentiable function on the range of X. Then, the pdf of $Y = g(X)$ in the region where $f_Y(y) > 0$ is given by*

$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{dg^{-1}(y)}{dy}\right|, \tag{5.2.3}$$

where g^{-1} denotes the inverse function of g.

Proof: We assume first that g is monotonically increasing. In this case, we have

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

If we now apply the chain rule and differentiate the above relation, we obtain

$$f_Y(y) = F'_Y(y) = \frac{d}{dy}\Big(F_X\left(g^{-1}(y)\right)\Big) = F'_X\left(g^{-1}(y)\right)\frac{dg^{-1}(y)}{dy} = f_X\left(g^{-1}(y)\right)\frac{dg^{-1}(y)}{dy}.$$

Since g is monotonically increasing, g^{-1} is also monotonically increasing, so that its derivative is positive and we have

$$\left|\frac{dg^{-1}(y)}{dy}\right| = \frac{dg^{-1}(y)}{dy}.$$

This justifies the pdf formula in the monotonically increasing case. In the case of a monotonically decreasing function, we would first obtain the relation

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

Differentiation subsequently leads to

$$f_Y(y) = F'_Y(y) = \frac{d}{dy}\Big(1 - F_X\left(g^{-1}(y)\right)\Big) = -F'_X\left(g^{-1}(y)\right)\frac{dg^{-1}(y)}{dy} = f_X\left(g^{-1}(y)\right)\left(-\frac{dg^{-1}(y)}{dy}\right).$$

However, g^{-1} is now monotonically decreasing, implying that its derivative is negative and so

$$\left|\frac{dg^{-1}(y)}{dy}\right| = -\frac{dg^{-1}(y)}{dy}.$$

Therefore, in either case, we obtain the formula given by (5.2.3). ■

Example 5.2.3. Let X be a continuous random variable with pdf of the form

$$f_X(x) = \begin{cases} \frac{9-x^2}{18} & \text{for } 0 < x < 3, \\ 0 & \text{otherwise.} \end{cases}$$

Find the pdf of $Y = X^2$.

Solution: We begin by noting that within the interval $(0, 3)$, $g(x) = x^2$ is strictly monotonic, and its inverse is given by $g^{-1}(y) = \sqrt{y}$. Therefore, for $y \in (0, 9)$, we have

$$f_X\left(g^{-1}(y)\right) = f_X(\sqrt{y}) = \frac{9 - (\sqrt{y})^2}{18} = \frac{9-y}{18}$$

and

$$\left|\frac{dg^{-1}(y)}{dy}\right| = \left|\frac{d}{dy}y^{\frac{1}{2}}\right| = \frac{1}{2\sqrt{y}}.$$

Therefore, by Theorem 5.2.1, we obtain

$$f_Y(y) = \frac{9-y}{18} \cdot \frac{1}{2\sqrt{y}} = \frac{9-y}{36\sqrt{y}} \text{ for } 0 < y < 9.$$

■

Remarks:

(1) Note that the pdf of Y derived in Example 5.2.3 is a valid pdf. Clearly, $f_Y(y) \geq 0$ for all $y \in (0, 9)$ and

$$\begin{aligned}\int_0^9 \frac{9-y}{36\sqrt{y}} dy &= \int_0^9 \left(\frac{1}{4}y^{-\frac{1}{2}} - \frac{1}{36}y^{\frac{1}{2}}\right) dy \\ &= \left(\frac{y^{\frac{1}{2}}}{2} - \frac{y^{\frac{3}{2}}}{54}\right)\Bigg|_0^9 \\ &= \frac{3}{2} - \frac{27}{54} \\ &= 1.\end{aligned}$$

(2) Alternatively, if the random variable X had pdf of the form (which we saw earlier)

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

then we would not be able to directly use Theorem 5.2.1, as the function $g(x) = x^2$ is **not** strictly monotonic for $x \in (-1, 1)$. Instead, we would have to rely on the general formula for $f_Y(y)$ we derived in Example 5.2.2.

We conclude this section by stating an important result pertaining to the pdf of a linear function of a continuous random variable.

Corollary 5.2.1. *Let X be a continuous random variable with pdf $f_X(x)$. If a and b are real constants such that $a \neq 0$, then $Y = aX + b$ has pdf given by*

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right).$$

Proof: If we define $y = g(x) = ax + b$ (so that $x = g^{-1}(y) = \frac{y-b}{a}$), then applying the result of Theorem 5.2.1 immediately yields

$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{dg^{-1}(y)}{dy}\right| = f_X\left(\frac{y-b}{a}\right)\left|\frac{d}{dy}\left(\frac{y-b}{a}\right)\right| = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right),$$

as required. ■

Section 5.2 Problems

5.2.1 Consider the pdf of X given in Problem 5.1.1.

(a) Determine the pdf of $Y = 2X + 3$.

(b) Determine the pdf of $Z = X^2$.

5.2.2 Consider the pdf of X given in Problem 5.1.3.

(a) Determine the pdf of $Y = 2(X - 1)$.

(b) Determine the pdf of $Z = X^3$.

5.2.3 Consider the pdf of X given in Problem 5.1.4.

(a) Determine the pdf of $Y = 1 - X$.

(b) Determine the pdf of $Z = \frac{1}{X}$.

5.2.4 Suppose that X is a continuous random variable with pdf given by

$$f_X(x) = \begin{cases} 4x^3 & \text{for } 0 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Determine the pdf and cdf of $Y = -\ln X^4$.

5.2.5 Suppose that X is a continuous random variable with pdf given by

$$f_X(x) = \begin{cases} \frac{|x|}{4} & \text{for } -2 < x < 2, \\ 0 & \text{otherwise.} \end{cases}$$

Determine the pdf and cdf of $Y = X^3$.

5.3 Expectation of a Random Variable

The **expected value** or **mean** of a continuous random variable X with pdf $f_X(x)$ is defined by

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

Note that this is analogous to the formula given in Definition 3.3.1 regarding the expected value of a discrete random variable Y with pmf $f_Y(y)$, namely

$$\mu_Y = E(Y) = \sum_{all\, y} y f_Y(y).$$

In other words, the concept of expectation for a continuous random variable is similar to the discrete case except that the pmf is replaced by the pdf, and summation is replaced by integration. As was discussed in Chapter 3, we can continue to interpret μ_X as the "balancing point" or "center of mass" of the probability distribution of X.

It is important to mention that one must consider the possibility that the integral $\int_{-\infty}^{\infty} x f_X(x)dx$ is not well-defined. In particular, we will say that the expected value of X is well-defined if

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

In this case, it is known that the integral $\int_{-\infty}^{\infty} x f_X(x)dx$ takes on a finite and unambiguous value. Just as in the discrete case, we implicitly assume (unless otherwise stated) that the expected value of a continuous random variable is well-defined.

If X is a continuous random variable with pdf $f_X(x)$, then any real-valued function $Y = g(X)$ of X is also a random variable. Note that Y can be a continuous random variable, as we witnessed in all of the examples considered in Section 5.2. However, Y can also turn out to be discrete. For example, suppose that $g(x) = 1$ for $x > 0$, and $g(x) = 0$ otherwise. In this case, $Y = g(X)$ is a discrete random variable, taking on two possible values (0 or 1). In either scenario, the mean of $Y = g(X)$ satisfies the *expected value rule*

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

which is in complete analogy with the discrete case.

The mathematical properties of expectation for a continuous random variable X are practically identical to those we established in the discrete case. After all, an integral is just a limiting form of a sum, meaning that such properties can be formally justified by expressing a quantity like $\int_{-\infty}^{\infty} g(x) f_X(x)dx$ as a limit of a Riemann sum and recognizing the Riemann sum as being in the form of an expected value for a discrete random variable. As an immediate consequence, we can simply list a few of the more important facts which are essentially identical to their discrete counterparts:

(1) The mgf of X is defined as

$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f_X(x)dx,$$

provided that it exists for all values of t within some open interval around 0. Theorem 3.4.1 concerning the uniqueness of moment generating functions continues to hold true for continuous probability distributions.

(2) The n^{th} moment of X, $n = 1, 2, 3, \ldots$, is defined as

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

Theorem 3.4.2 describing how the mgf can be used to obtain the moments of a discrete random variable is even true for continuous random variables, so that

$$E(X^n) = M_X^{(n)}(0) = \frac{d^n}{dt^n} M_X(t)\Big|_{t=0} \quad \text{for } n = 1, 2, 3, \ldots.$$

(3) The variance of X is defined as

$$\sigma_X^2 = Var(X) = E\left((X - \mu_X)^2\right) = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx = E(X^2) - \mu_X^2,$$

along with its standard deviation $\sigma_X = SD(X) = \sqrt{Var(X)}$.

(4) For real constants a and b,

$$\begin{aligned} E(aX + b) &= a\mu_X + b, \\ Var(aX + b) &= a^2\sigma_X^2, \\ SD(aX + b) &= |a|\sigma_X, \\ M_{aX+b}(t) &= e^{bt} M_X(at). \end{aligned}$$

Example 5.3.1. Suppose that X is a continuous random variable with pdf given by

$$f_X(x) = \begin{cases} \frac{6x^2}{5} & \text{for } 0 < x < 1, \\ \frac{6}{5}(2 - x) & \text{for } 1 < x < 2, \\ 0 & \text{otherwise.} \end{cases}$$

Determine the mean, variance, and standard deviation of $Y = 8 - 2X$.

Solution: We first calculate the mean and variance of X as follows:

$$\begin{aligned}
\mu_X &= \int_{-\infty}^{\infty} x f_X(x) dx \\
&= \int_0^1 x \, \frac{6}{5} x^2 dx + \int_1^2 x \, \frac{6}{5}(2-x) dx \\
&= \frac{6}{5}\left(\frac{x^4}{4}\right)\Big|_0^1 + \frac{6}{5}\left(x^2 - \frac{x^3}{3}\right)\Big|_1^2 \\
&= \frac{3}{10} + \frac{6}{5}\left(4 - \frac{8}{3} - 1 + \frac{1}{3}\right) \\
&= \frac{3}{10} + \frac{6}{5} \cdot \frac{2}{3} \\
&= \frac{11}{10},
\end{aligned}$$

$$\begin{aligned}
E(X^2) &= \int_{-\infty}^{\infty} x^2 f_X(x) dx \\
&= \int_0^1 x^2 \, \frac{6}{5} x^2 dx + \int_1^2 x^2 \, \frac{6}{5}(2-x) dx \\
&= \frac{6}{5}\left(\frac{x^5}{5}\right)\Big|_0^1 + \frac{6}{5}\left(\frac{2x^3}{3} - \frac{x^4}{4}\right)\Big|_1^2 \\
&= \frac{6}{25} + \frac{6}{5}\left(\frac{16}{3} - 4 - \frac{2}{3} + \frac{1}{4}\right) \\
&= \frac{6}{25} + \frac{6}{5} \cdot \frac{11}{12} \\
&= \frac{67}{50},
\end{aligned}$$

$$\sigma_X^2 = E\left(X^2\right) - \mu_X^2 = \frac{67}{50} - \left(\frac{11}{10}\right)^2 = \frac{13}{100}.$$

Therefore, we can now obtain

$$\mu_Y = 8 - 2\mu_X = 8 - \frac{22}{10} = \frac{58}{10} = \frac{29}{5},$$

$$\sigma_Y^2 = (-2)^2 \sigma_X^2 = \frac{52}{100} = \frac{13}{25},$$

and

$$\sigma_Y = \sqrt{Var(Y)} = \frac{\sqrt{13}}{5}.$$

■

Example 5.3.2. Let X be a continuous random variable with pdf of the form

$$f(x) = \frac{1}{2b} e^{-|x|/b} \text{ for } -\infty < x < \infty,$$

where $b > 0$ is a real-valued parameter of the probability distribution. Determine the mgf of X and use it to calculate the mean and variance of X.

Solution: First of all, we remark that the random variable X takes on values on the entire real number line. For $x \in \mathbb{R}$, note that $f(x) > 0$ and

$$\begin{aligned}
\int_{-\infty}^{\infty} f(x)dx &= \underbrace{\int_{-\infty}^{0} \frac{1}{2b} e^{x/b} dx}_{\text{Let } y = -x \text{ so that } dy = -dx} + \int_{0}^{\infty} \frac{1}{2b} e^{-x/b} dx \\
&= \int_{0}^{\infty} \frac{1}{2b} e^{-y/b} dy + \int_{0}^{\infty} \frac{1}{2b} e^{-x/b} dx \\
&= \frac{1}{b} \int_{0}^{\infty} e^{-x/b} dx \\
&= \frac{1}{b} \left(\frac{e^{-x/b}}{-1/b} \right) \Big|_{0}^{\infty} \\
&= 1 - \underbrace{\lim_{x \to \infty} e^{-x/b}}_{=0} \\
&= 1.
\end{aligned}$$

Figure 5.3.1 shows plots of $f(x)$ for a few values of b.

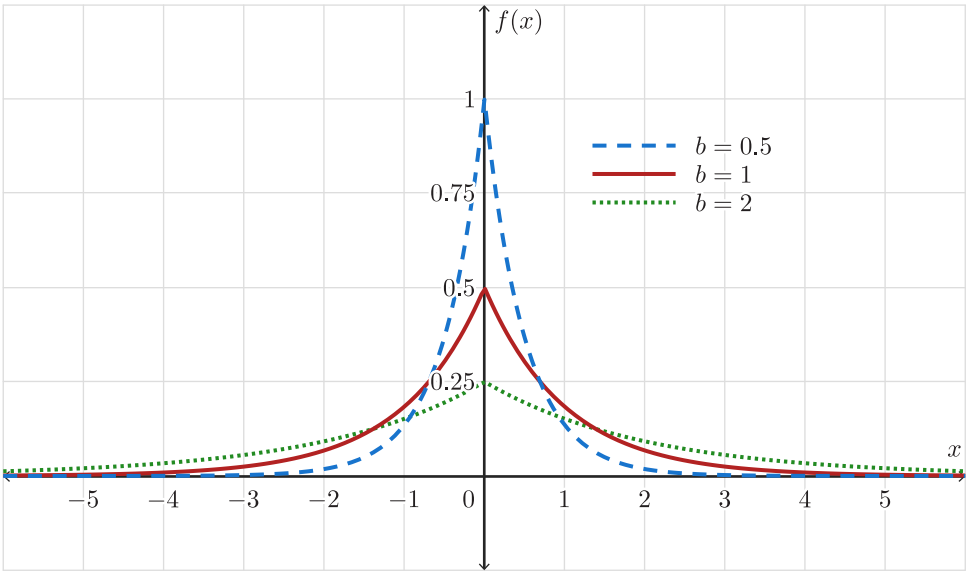


Figure 5.3.1: Plots of the pdf $f(x)$ in Example 5.3.2 for different values of b

We now proceed to find the mgf of X, which is given by

$$\begin{aligned}
M(t) &= E(e^{tX}) \\
&= \int_{-\infty}^{\infty} e^{tx} \frac{1}{2b} e^{-|x|/b} dx \\
&= \int_{-\infty}^{0} e^{tx} \frac{1}{2b} e^{x/b} dx + \int_{0}^{\infty} e^{tx} \frac{1}{2b} e^{-x/b} dx \\
&= \frac{1}{2b} \int_{-\infty}^{0} e^{(t+1/b)x} dx + \frac{1}{2b} \int_{0}^{\infty} e^{(t-1/b)x} dx \\
&= \frac{1}{2b} \left(\frac{e^{(t+1/b)x}}{t+1/b} \right) \Big|_{-\infty}^{0} + \frac{1}{2b} \left(\frac{e^{(t-1/b)x}}{t-1/b} \right) \Big|_{0}^{\infty} \\
&= \frac{1}{2b(t+1/b)} \Big(1 - \underbrace{\lim_{x \to -\infty} e^{(t+1/b)x}}_{= 0 \text{ if } t > -1/b} \Big) + \frac{1}{2b(t-1/b)} \Big(\underbrace{\lim_{x \to \infty} e^{(t-1/b)x}}_{= 0 \text{ if } t < 1/b} - 1 \Big) \\
&= \frac{(t-1/b) - (t+1/b)}{2b(t+1/b)(t-1/b)} \text{ provided that } |t| < 1/b \\
&= \frac{1}{b^2(t+1/b)(-t+1/b)} \\
&= \frac{1}{1-b^2t^2}.
\end{aligned}$$

Taking the first two derivatives of $M(t)$, we get

$$M^{(1)}(t) = -(1-b^2t^2)^{-2}(-2b^2t) = 2b^2t(1-b^2t^2)^{-2}$$

and

$$M^{(2)}(t) = -4b^2t(1-b^2t^2)^{-3}(-2b^2t) + 2b^2(1-b^2t^2)^{-2} = 8b^4t^2(1-b^2t^2)^{-3} + 2b^2(1-b^2t^2)^{-2}.$$

Therefore, we immediately obtain

$$E(X) = M^{(1)}(0) = 2b^2(0)(1-b^2(0)^2)^{-2} = 0$$

and

$$Var(X) = E(X^2) - 0^2 = M^{(2)}(0) = 8b^4(0)^2(1-b^2(0)^2)^{-3} + 2b^2(1-b^2(0)^2)^{-2} = 2b^2.$$

■

Section 5.3 Problems

5.3.1 Consider the pdf of X given in Problem 5.1.1. Calculate the mean and variance of X.

5.3.2 Consider the pdf of X given in Problem 5.1.3. Calculate the mean and standard deviation of X.

5.3.3 Consider the pdf of X given in Problem 5.1.4. Determine a formula for the n^{th} moment of X where $n = 1, 2, 3, \ldots$.

5.3.4 Suppose that X is a continuous random variable with pdf given by

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

Consider a random rectangle whose sides are X and $1 - X$. Determine the mean and variance of the area of the rectangle.

5.3.5 Let X be a continuous random variable having cdf of the form

$$F_X(x) = \begin{cases} 0 & \text{for } x \leq 0, \\ \frac{x(8-x)}{16} & \text{for } 0 < x \leq 4, \\ 1 & \text{for } x > 4. \end{cases}$$

Calculate the variance of the random variable $Y = 2(9 - X^2)$.

5.3.6 Suppose that X is a continuous random variable with pdf given by

$$f(x) = \begin{cases} a + bx + cx^2 & \text{for } 0 < x < 3, \\ 0 & \text{otherwise.} \end{cases}$$

(a) If $E(X) = 2$ and $Var(X) = 1$, determine the values of a, b, and c.

(b) Plot the graph of $f(x)$.

5.3.7 Let X be a random variable with pdf $f(x)$ and mgf $M(t)$. Suppose that f is symmetric about 0 (i.e., $f(-x) = f(x)$). Show that $M(-t) = M(t)$.

5.3.8 Let X be a continuous random variable with pdf of the form

$$f(x) = \frac{3}{2}e^{-x} - \frac{3}{2}e^{-cx} \text{ for } x > 0.$$

(a) Determine the value of c that makes the above pdf valid.

(b) Determine the mgf of X.

(c) Use the mgf of X to calculate $E(X)$ and $Var(X)$.

5.4 Special Continuous Probability Distributions

Similar to our treatment of the special discrete probability distributions in Section 3.5, we now turn our attention to some of the more popular continuous probability distributions used in practice. Such distributions arise in certain settings and find wide use in a variety of real-life applications. In this section, we present three *special* model distributions for continuous random variables and study their key distributional properties.

5.4.1 Continuous Uniform Distribution

To describe the setup associated with this probability distribution, consider a random variable X which takes on values in the interval (a, b), where a and b are real numbers such that $a < b$. If the probability distribution of X satisfies the property that all subintervals of (a, b) with the same (fixed) length are *equally likely*, then X is said to have a **continuous uniform distribution**. We write $X \sim U(a, b)$ as a shorthand for "X is distributed according to a continuous uniform distribution on the interval (a, b)". In terms of an immediate illustration, the spinning pointer example in Section 5.1 featured a random variable X such that $X \sim U(0, 4)$.

Since all values in (a, b) are, in some sense, equally likely (more precisely, all subintervals contained within (a, b) of a given length have the same probability), then it follows that the pdf of X must be a constant. In other words, we must have $f(x) = k$ for all $a < x < b$ where k is an appropriate constant. In order to ensure that $\int_a^b f(x)dx = 1$, it is obvious that we require $k = \frac{1}{b-a}$. Therefore, the pdf of X is given by

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

A plot of the $U(a, b)$ pdf is depicted in Figure 5.4.1. As expected, the area under $f(x)$ corresponds to the area of a rectangle with base $b - a$ and height $\frac{1}{b-a}$, producing a value of 1.

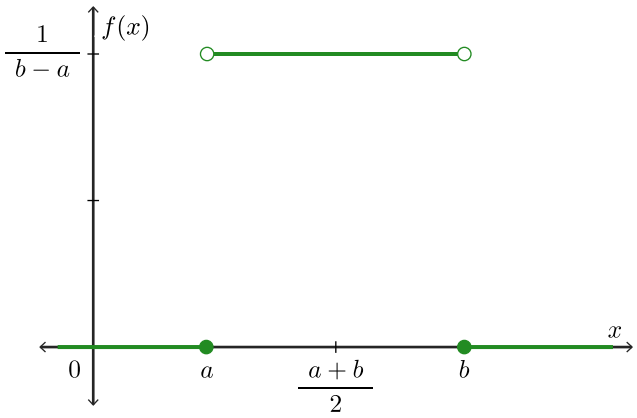


Figure 5.4.1: A plot of the $U(a, b)$ pdf

Turning our attention to the cdf of X, we have for $a < x \leq b$:

$$F(x) = \int_a^x \frac{1}{b-a} dz = \frac{1}{b-a}(z)\Big|_a^x = \frac{x-a}{b-a}. \tag{5.4.1}$$

As a result, the "full" cdf of X can be expressed as

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

A plot of the $U(a, b)$ cdf is depicted in Figure 5.4.2. Since the cdf is a *linear* function on the interval (a, b), the determination of the p-quantile of X is straightforward. In particular, from the equation $F(q(p)) = p$, we simply obtain

$$\frac{q(p) - a}{b-a} = p \Longrightarrow q(p) = a(1-p) + bp.$$

In addition, we can also obtain the mean and variance of X as follows:

$$E(X) = \int_a^b x \frac{1}{b-a} dx = \frac{1}{b-a}\left(\frac{x^2}{2}\right)\Big|_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{(b-a)(b+a)}{2(b-a)} = \frac{a+b}{2}$$

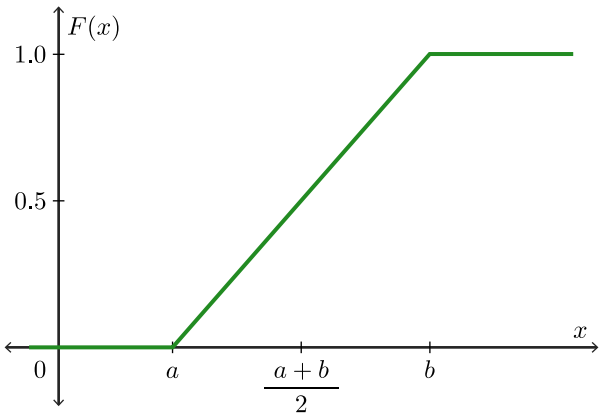


Figure 5.4.2: A plot of the $U(a, b)$ cdf

and

$$\begin{aligned}
Var(X) &= E(X^2) - (E(X))^2 \\
&= \int_a^b x^2 \frac{1}{b-a} dx - \left(\frac{a+b}{2}\right)^2 \\
&= \frac{1}{b-a}\left(\frac{x^3}{3}\right)\Big|_a^b - \frac{a^2+2ab+b^2}{4} \\
&= \frac{b^3-a^3}{3(b-a)} - \frac{a^2+2ab+b^2}{4} \\
&= \frac{(b-a)(b^2+ab+a^2)}{3(b-a)} - \frac{a^2+2ab+b^2}{4} \\
&= \frac{b^2+ab+a^2}{3} - \frac{a^2+2ab+b^2}{4} \\
&= \frac{4b^2+4ab+4a^2-3a^2-6ab-3b^2}{12} \\
&= \frac{b^2-2ab+a^2}{12} \\
&= \frac{(b-a)^2}{12}.
\end{aligned}$$

One of the most interesting and key results involving the uniform distribution is stated in the next theorem.

Theorem 5.4.1. *Suppose that X is a continuous random variable with cdf $F_X(x)$. Suppose that $F_X(x)$ is strictly increasing in the following sense: If $x_1 < x_2$ such that $0 < F_X(x_1)$ and $F_X(x_2) < 1$, then $F_X(x_1) < F_X(x_2)$. If $Y = F_X(X)$, then $Y \sim U(0, 1)$.*

Proof: If $F_X(x)$ is strictly increasing in the sense described, then for $y \in (0,1)$ the equation $F_X(x) = y$ has a unique solution, call it $x = F_X^{-1}(y)$. Moreover, the function F_X^{-1} is strictly increasing on the interval $(0,1)$ where it is defined. Therefore, for $y \in (0,1)$, the cdf of Y is given by

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(F_X(X) \leq y) \\ &= P(F_X^{-1}(F_X(X)) \leq F_X^{-1}(y)) \\ &= P(X \leq F_X^{-1}(y)) \\ &= F_X(F_X^{-1}(y)) \\ &= y \\ &= \frac{y-0}{1-0}, \end{aligned}$$

which, from (5.4.1), is the cdf of a $U(0,1)$ random variable. Thus, $Y = F_X(X) \sim U(0,1)$. ■

Remark: Many computer software programs have "random number generator" functions that will simulate observations from a $U(0,1)$ distribution. These are more properly called **pseudo-random number generators**, since they are based on deterministic algorithms. As a result, they yield observations that have finite precision so they cannot be **exactly** like $U(0,1)$ random variables. However, good generators produce values that appear indistinguishable in most ways from $U(0,1)$ random variables. Therefore, given such a generator, the result of Theorem 5.4.1 provides us with a way to simulate random variables having a general continuous probability distribution (denoted by the random variable X) via the following algorithm:

(1) Generate a value y from the $U(0,1)$ distribution using the computer random number generator.

(2) Calculate the corresponding value from the distribution of X via the relation $x = F_X^{-1}(y)$.

5.4.2 Exponential Distribution

One of the most commonly-used continuous distributions, both in terms of its application as well as its mathematical tractability, is the **exponential distribution**. It serves as a building block for other continuous distributions of interest that are studied in more advanced probability courses. It is also a very important probability distribution typically used to model times between events or arrivals. The exponential distribution is used in a wide range of applications such as survival analysis, queueing theory, and actuarial risk theory, to name a few.

The pdf of an exponential distribution is given by

$$f(x) = \frac{1}{\theta} e^{-x/\theta} \text{ for } x > 0,$$

where $\theta > 0$ is a real-valued parameter of the probability distribution. In fact, $\frac{1}{\theta}$ is often referred to as the *rate* parameter of the exponential distribution. We write $X \sim Exp(\theta)$ as a shorthand for "X is distributed according to an exponential distribution with parameter θ". Clearly, we see that

$$\int_0^\infty \frac{1}{\theta} e^{-x/\theta} dx = \frac{1}{\theta} \left(\frac{e^{-x/\theta}}{-\frac{1}{\theta}} \right) \Bigg|_0^\infty = -\Big(\underbrace{\lim_{x \to \infty} e^{-x/\theta}}_{=0} - 1 \Big) = 1.$$

Several plots of the $Exp(\theta)$ pdf are depicted in Figure 5.4.3 for different values of θ. We can also obtain a formula for the cdf of X. Note that when $x > 0$,

$$F(x) = \int_0^x \frac{1}{\theta} e^{-z/\theta} dz = \frac{1}{\theta} \left(\frac{e^{-z/\theta}}{-\frac{1}{\theta}} \right) \Bigg|_0^z = 1 - e^{-x/\theta}.$$

As a result, the "full" cdf of X is given by

$$F(x) = \begin{cases} 0 & \text{for } x \leq 0, \\ 1 - e^{-x/\theta} & \text{for } x > 0. \end{cases}$$

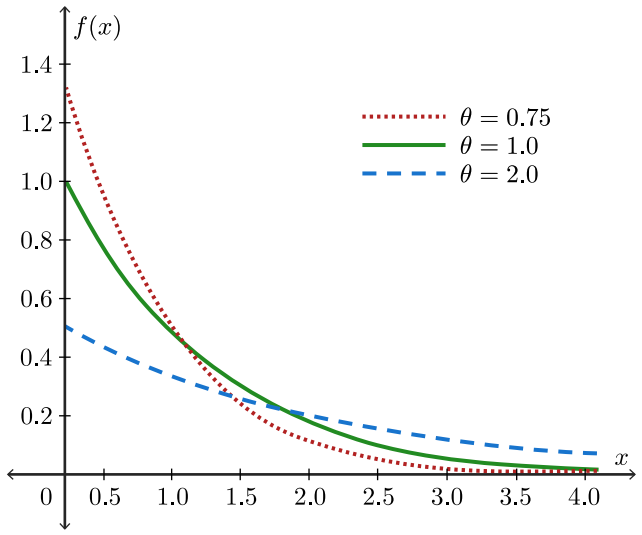


Figure 5.4.3: Plots of the $Exp(\theta)$ pdf for different values of θ

A plot of the $Exp(\theta)$ cdf is depicted in Figure 5.4.4. With such a convenient formula for the cdf, the determination of the p-quantile of X is also accessible. In particular, from the equation $F(q(p)) = p$,

we have that

$$1 - e^{-q(p)/\theta} = p \Longrightarrow -\frac{q(p)}{\theta} = \ln(1-p) \Longrightarrow q(p) = \theta \ln(1-p)^{-1}.$$

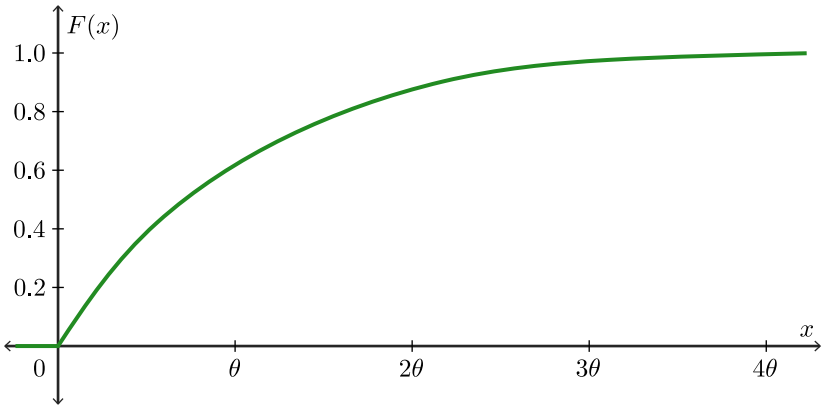


Figure 5.4.4: A plot of the $Exp(\theta)$ cdf

To obtain the mean and variance of X, we first find the mgf of X as follows:

$$\begin{aligned}
M(t) &= \int_0^\infty e^{tx}\,\frac{1}{\theta}e^{-x/\theta}dx \\
&= \frac{1}{\theta}\int_0^\infty e^{(t-1/\theta)x}dx \\
&= \frac{1}{\theta}\left(\frac{e^{(t-1/\theta)x}}{t-1/\theta}\right)\Big|_0^\infty \\
&= \frac{1}{\theta t-1}\Big(\lim_{x\to\infty} e^{(t-1/\theta)x} - 1\Big) \\
&= \frac{1}{\theta t-1}\big(0-1\big) \text{ provided that } t-\frac{1}{\theta} < 0 \\
&= \frac{1}{1-\theta t} \text{ for } t < \frac{1}{\theta}.
\end{aligned}$$

Taking the first two derivatives of $M(t)$ yields

$$M^{(1)}(t) = -(1-\theta t)^{-2}(-\theta) = \theta(1-\theta t)^{-2}$$

and

$$M^{(2)}(t) = -2\theta(1-\theta t)^{-3}(-\theta) = 2\theta^2(1-\theta t)^{-3}.$$

Therefore, we readily obtain

$$E(X) = M^{(1)}(0) = \theta(1-0)^{-2} = \theta$$

and

$$Var(X) = E(X^2) - (E(X))^2 = 2\theta^2(1-0)^{-3} - \theta^2 = 2\theta^2 - \theta^2 = \theta^2.$$

Example 5.4.1. Upon arrival to a bus stop every morning, past history indicates that the time you have to wait for a bus to arrive is exponentially distributed with a mean of 12 minutes. What is the probability that you have to wait longer than 15 minutes for a bus to arrive? In addition, if you have already been waiting for 10 minutes, what is the probability that you have to wait at least an additional 15 minutes for a bus to arrive?

Solution: Let X represent the waiting time (in minutes) for a bus to arrive. Since it takes, on average, 12 minutes for a bus to arrive, it follows that $X \sim Exp(12)$. We first wish to calculate

$$\begin{aligned} P(X > 15) &= 1 - P(X \leq 15) \\ &= 1 - F(15) \\ &= 1 - \left(1 - e^{-15/12}\right) \\ &= e^{-1.25} \\ &\approx 0.2865. \end{aligned}$$

In addition, we wish to calculate a particular conditional probability. Specifically, we consider

$$\begin{aligned} P(X > 25|X > 10) &= \frac{P(\{X > 25\} \cap \{X > 10\})}{P(X > 10)} \\ &= \frac{P(X > 25)}{P(X > 10)} \\ &= \frac{1 - \left(1 - e^{-25/12}\right)}{1 - (1 - e^{-10/12})} \\ &= \frac{e^{-25/12}}{e^{-10/12}} \\ &= e^{-1.25} \\ &\approx 0.2865. \end{aligned}$$

■

Remark: Do the calculated probabilities in Example 5.4.1 surprise you? In comparing them, does it not seem strange to discover that already waiting 10 minutes for a bus has no effect on your remaining waiting time, as these two probabilities are identical? This illustrates an important mathematical property that the exponential distribution possesses, namely the so-called "memoryless property". It is stated formally as follows: for non-negative real numbers a and b, if $X \sim Exp(\theta)$, then

$$P(X > a + b|X > a) = P(X > b).$$

If we think of X as the waiting time for some event of interest to occur, then what the memoryless property is telling us is given that you have waited a units of time for the next event to occur, the

probability that you have to wait an additional b units of time does not depend on a whatsoever and only depends on the value of b. This is a rather rare quality for a random variable to have. In fact, the exponential distribution is the *unique* continuous probability distribution possessing the memoryless property.

5.4.3 Normal Distribution

The **normal distribution** is generally considered to be the most important and most widely used probability distribution in statistics. It is sometimes referred to as the "Gaussian distribution", named after the German mathematician Carl Friedrich Gauss who is credited with its discovery in 1809. The normal distribution finds extensive use in many processes where X represents a physical dimension of some kind (e.g., the heights or weights of individuals in large populations tend to follow normal distributions), but also in other diverse applications (e.g., logarithms of stock prices are often assumed to be normally distributed).

A random variable X defined on $\mathbb{R}$ is said to have a normal distribution if its pdf is of the form

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \text{ for } -\infty < x < \infty,$$

where $\mu \in \mathbb{R}$ and $\sigma > 0$ are parameters of the probability distribution. The choice to use μ and σ, which we typically reserve for mean and standard deviation of a random variable, respectively, is not coincidental. It turns out (as we will verify shortly) that $E(X) = \mu$ and $Var(X) = \sigma^2$ for this particular distribution. Therefore, we write $X \sim N(\mu, \sigma^2)$ as a shorthand to denote that "X has a normal distribution with mean μ and variance σ^2".

A plot of the $N(\mu, \sigma^2)$ pdf is shown in Figure 5.4.5. The shape of $f(x)$ is what is often termed a "bell shape" or "bell curve", symmetric about the point $x = \mu$. In other words, the normal distribution is one in which most points cluster around μ, while the rest taper off symmetrically toward either extreme.

To verify that the area under the curve of $f(x)$ is equal to 1, we consider

$$\begin{aligned}
\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy \text{ if we let } y = \frac{x-\mu}{\sigma} \text{ so that } dy = \frac{dx}{\sigma} \\
&= \frac{1}{\sqrt{2\pi}} \cdot 2 \int_{0}^{\infty} e^{-\frac{y^2}{2}} dy \text{ since } e^{-\frac{y^2}{2}} \text{ is an even function} \\
&= \frac{\sqrt{2}}{\sqrt{\pi}} \int_{0}^{\infty} e^{-\frac{y^2}{2}} dy.
\end{aligned}$$

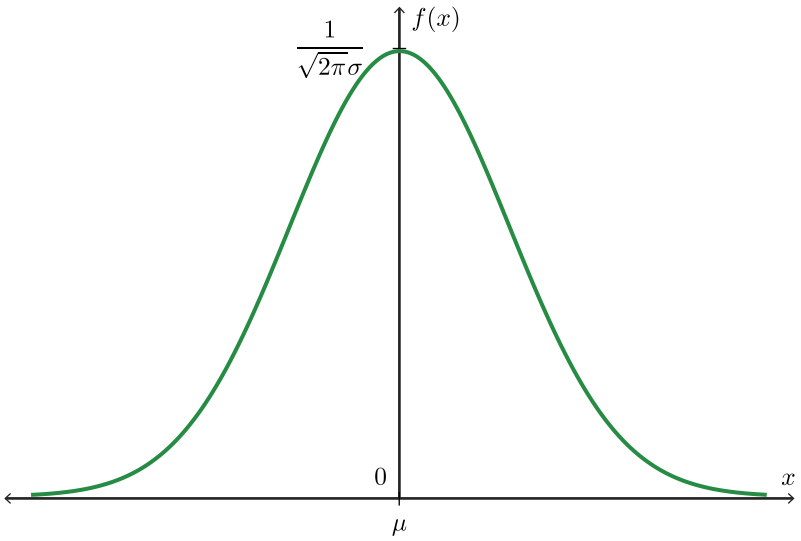


Figure 5.4.5: A plot of the $N(\mu, \sigma^2)$ pdf

However, note that

$$\begin{aligned}
\left(\frac{\sqrt{2}}{\sqrt{\pi}}\int_0^\infty e^{-\frac{y^2}{2}}dy\right)^2 &= \frac{2}{\pi}\left(\int_0^\infty e^{-\frac{y^2}{2}}dy\right)\left(\int_0^\infty e^{-\frac{z^2}{2}}dz\right) \\
&= \frac{2}{\pi}\int_0^\infty\int_0^\infty e^{-\frac{y^2+z^2}{2}}dydz \\
&= \frac{2}{\pi}\int_0^\infty\int_0^{\pi/2} e^{-\frac{r^2}{2}}rd\theta dr \text{ if we let } y = r\cos\theta \text{ and } z = r\sin\theta \text{ so that } dydz = rd\theta dr \\
&= \int_0^\infty re^{-\frac{r^2}{2}}dr \\
&= \int_0^\infty e^{-s}ds \text{ if we let } s = \frac{r^2}{2} \text{ so that } ds = rdr \\
&= (-e^{-s})\Big|_0^\infty \\
&= 1.
\end{aligned}$$

Therefore, it immediately follows that $\int_{-\infty}^{\infty} f(x)dx = 1$, as required.

We also stated at the outset of this subsection that the parameters μ and σ represented the mean

and standard deviation, respectively, of X. Let us now justify this earlier assertion. In order to do so, we begin by introducing a special case of the normal distribution, one having parameters $\mu = 0$ and $\sigma = 1$. Specifically, suppose that $Z \sim N(0,1)$ with pdf

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \text{ for } -\infty < z < \infty. \tag{5.4.2}$$

Such a normal distribution with parameters 0 and 1 is referred to as the **standard normal distribution**. Its mgf is given by

$$\begin{aligned}
M_Z(t) &= \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z^2 - tz)} dz \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z^2 - tz + t^2)} \cdot e^{\frac{t^2}{2}} dz \\
&= e^{\frac{t^2}{2}} \int_{-\infty}^{\infty} \underbrace{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-t)^2}}_{N(t,1) \text{ pdf}} dz \\
&= e^{\frac{t^2}{2}} \cdot 1 \\
&= e^{\frac{t^2}{2}} \text{ for } t \in \mathbb{R}.
\end{aligned}$$

Thus, the mean and variance of Z can be readily obtained as follows:

$$E(Z) = M_Z^{(1)}(0) = \frac{d}{dt}\left(e^{\frac{t^2}{2}}\right)\Big|_{t=0} = \left(te^{\frac{t^2}{2}}\right)\Big|_{t=0} = 0,$$

$$Var(Z) = E(Z^2) - 0^2 = M_Z^{(2)}(0) = \frac{d}{dt}\left(te^{\frac{t^2}{2}}\right)\Big|_{t=0} = \left(t^2 e^{\frac{t^2}{2}} + e^{\frac{t^2}{2}}\right)\Big|_{t=0} = 0 + 1 = 1.$$

To connect the above expressions pertaining to the $N(0,1)$ distribution to the more general $N(\mu, \sigma^2)$ distribution, the following important result provides the necessary framework to establish this connection.

Theorem 5.4.2. *If $X \sim N(\mu, \sigma^2)$ and $Z = \frac{X-\mu}{\sigma}$, then $Z \sim N(0,1)$.*

Proof: Clearly, we may express Z in the form $Z = aX + b$ where $a = \frac{1}{\sigma}$ and $b = -\frac{\mu}{\sigma}$. By Corollary 5.2.1, Z has pdf given by

$$\begin{aligned}
f_Z(z) &= \frac{1}{\left|1/\sigma\right|} f_X\Big(\frac{z + \mu/\sigma}{1/\sigma}\Big) \\
&= \sigma f_X(\sigma z + \mu) \\
&= \sigma \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{\sigma z + \mu - \mu}{\sigma}\right)^2} \text{ since } X \sim N(\mu, \sigma^2) \\
&= \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \text{ for } -\infty < z < \infty.
\end{aligned} \tag{5.4.3}$$

Comparing the forms of (5.4.2) and (5.4.3), we conclude that $Z \sim N(0,1)$. ■

As a result of Theorem 5.4.2, $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$ if $X \sim N(\mu, \sigma^2)$, so that

$$M_X(t) = M_{\sigma Z+\mu}(t) = e^{\mu t} M_Z(\sigma t) = e^{\mu t} \cdot e^{\frac{(\sigma t)^2}{2}} = e^{\mu t + \frac{\sigma^2 t^2}{2}} \text{ for } t \in \mathbb{R},$$

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

and

$$Var(X) = Var(\sigma Z + \mu) = \sigma^2 Var(Z) = \sigma^2 \cdot 1 = \sigma^2.$$

Therefore, we have indeed verified that $\mu = E(X)$ and $\sigma^2 = Var(X)$.

While we are able to evaluate the integral expression $\int_{-\infty}^{\infty} f(x)dx$ and show that it is equal to 1, the same cannot be said when it comes to looking at the cdf associated with a $N(\mu, \sigma^2)$ random variable X, which is given by

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2} dz \text{ for } -\infty < x < \infty.$$

A closed-form expression for this integral does not exist, and as a result, numerical methods have to be used to calculate its value for given values of x, μ, and σ. Fortunately, this function is included in many computer software packages and even some advanced calculators. Prior to such technological advances, however, one had to formulate tables of probabilities for $F_X(x)$ through the use of numerical integration procedures. Thankfully, due to the result of Theorem 5.4.2, it is only necessary to do this for a single normal distribution – namely, the standard normal distribution. The essential reason is due to the following basic relationship we have: for any $x \in \mathbb{R}$, note that

$$F_X(x) = P(X \le x) = P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) = P\left(Z \le \frac{x-\mu}{\sigma}\right) = F_Z\left(\frac{x-\mu}{\sigma}\right).$$

In other words, we can always "convert" a probability statement for a $N(\mu, \sigma^2)$ random variable into an equivalent one which is expressed in terms of the standard normal distribution. This conversion process involves "standardizing" the random variable X by subtracting off its mean and dividing by its standard deviation, so as to create the $N(0,1)$ random variable Z.

To this end, a table of probabilities for $F_Z(z) = P(Z \le z)$ is displayed in Table 5.4.1. A space-saving feature of this table is that only the values for $z \ge 0$ are displayed. For a negative value of z, we would simply use the fact that the $N(0,1)$ pdf is symmetric about 0. In other words, for any $z \in \mathbb{R}$, we always have $P(Z \le z) = P(Z \ge -z)$ by symmetry. Before tackling problems involving a general $N(\mu, \sigma^2)$ distribution, let us first consider an example illustrating how to obtain standard normal probabilities with the aid of Table 5.4.1.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56750	0.57142	0.57534
0.2	0.57926	0.58317	0.58706	0.59095	0.59484	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983

Table 5.4.1: Table of cdf probabilities pertaining to the $N(0, 1)$ distribution, corresponding to the shaded area in Figure 5.4.6

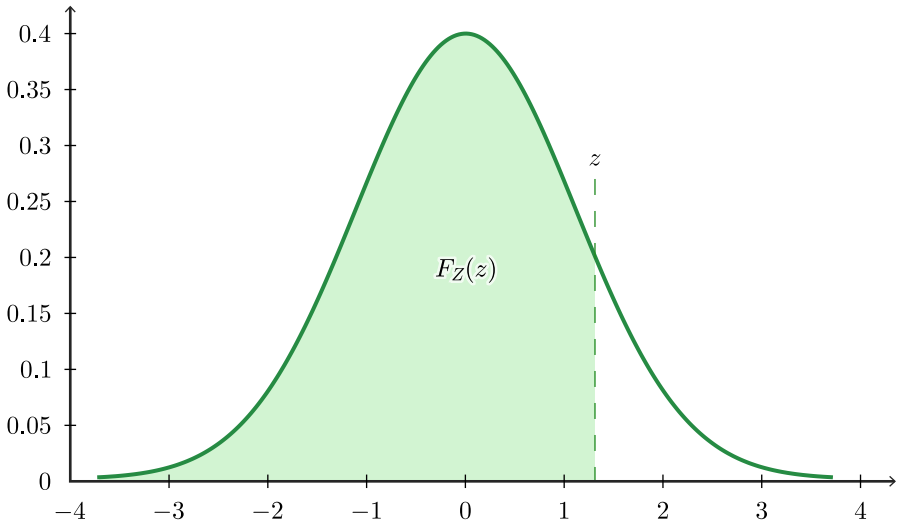


Figure 5.4.6: A plot of the $N(0, 1)$ pdf with the cdf probability $F_Z(z) = P(Z \leq z)$ shaded

Example 5.4.2. For $Z \sim N(0, 1)$, determine the following probabilities:

(i) $P(Z < 2.18)$,

(ii) $P(Z \geq 1.06)$,

(iii) $P(Z > -1.49)$,

(iv) $P(Z \leq -0.63)$, and

(v) $P(-1.32 < Z < 1.95)$.

Solution: First of all, in order to determine $P(Z < 2.18) = F_Z(2.18)$, we look up the value "2.18" in Table 5.4.1 by going down the left column to 2.1 and then across to the heading 0.08. We find the number 0.98537 in that location. Hence, we simply obtain $P(Z < 2.18) = 0.98537$.

Using the same approach to look up probabilities in Table 5.4.1, we have

$$P(Z \geq 1.06) = 1 - P(Z < 1.06) = 1 - F_Z(1.06) = 1 - 0.85543 = 0.14457.$$

Next, we would first apply a symmetry argument to get

$$P(Z > -1.49) = P(Z < 1.49) = F_Z(1.49) = 0.93189.$$

Again, a symmetry argument comes in handy to calculate

$$P(Z \leq -0.63) = P(Z \geq 0.63) = 1 - P(Z < 0.63) = 1 - F_Z(0.63) = 1 - 0.73565 = 0.26435.$$

p	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.075	0.08	0.09	0.095
0.5	0.0000	0.0251	0.0502	0.0753	0.1004	0.1257	0.1510	0.1764	0.1891	0.2019	0.2275	0.2404
0.6	0.2533	0.2793	0.3055	0.3319	0.3585	0.3853	0.4125	0.4399	0.4538	0.4677	0.4959	0.5101
0.7	0.5244	0.5534	0.5828	0.6128	0.6433	0.6745	0.7063	0.7388	0.7554	0.7722	0.8064	0.8239
0.8	0.8416	0.8779	0.9154	0.9542	0.9945	1.0364	1.0803	1.1264	1.1503	1.1750	1.2265	1.2536
0.9	1.2816	1.3408	1.4051	1.4758	1.5548	1.6449	1.7507	1.8808	1.9600	2.0537	2.3263	2.5758

Table 5.4.2: Table of p-quantiles pertaining to the $N(0, 1)$ distribution

Finally, note that

$$\begin{aligned} P(-1.32 < Z < 1.95) &= P(Z < 1.95) - P(Z \leq -1.32) \\ &= F_Z(1.95) - P(Z \geq 1.32) \\ &= F_Z(1.95) - (1 - P(Z < 1.32)) \\ &= F_Z(1.95) + F_Z(1.32) - 1 \\ &= 0.97441 + 0.90658 - 1 \\ &= 0.88099. \end{aligned}$$

■

In addition to using Table 5.4.1 to look up standard normal probabilities for given z-values, we sometimes are given the probabilities and asked to find the associated z-values which produce those probabilities. To help in this regard, Table 5.4.2 provides the $N(0, 1)$ p-quantiles for values of $p \geq 0.5$. The following example demonstrates how this table can be used.

Example 5.4.3. For $Z \sim N(0, 1)$, find:

(i) the number a such that $P(Z < a) = 0.85$,

(ii) the number b such that $P(Z > b) = 0.9$, and

(iii) the number c such that $P(-c < Z < c) = 0.95$.

Solution: In order to find the value of a, one approach would be to look in the body of Table 5.4.1 to find an entry close to 0.85. Note that this occurs for z between 1.03 and 1.04. We would be inclined to take $a = 1.04$ since it gives the closest value to 0.85. However, for greater accuracy, we should be using Table 5.4.2 which is designed for finding such numbers, given the probability. Looking up the entry "0.85" in this table, we find $a = 1.0364$.

Turning our attention next to finding b, note that

$$0.9 = P(Z > b) = 1 - P(Z \leq b) \Longrightarrow P(Z \leq b) = 0.1.$$

Unfortunately, there is no entry in Table 5.4.2 for which $P(Z \leq z) = 0.1$. As a result, we again have to rely on a symmetry argument, since we know that b will be a negative number. Specifically, from Table 5.4.2, we have that $P(Z \leq 1.2816) = 0.9$. By symmetry, $P(Z > -1.2816) = 0.9$, which implies that $P(Z \leq -1.2816) = 0.1$. Therefore, we have that $b = -1.2816$. The key to this solution lies in recognizing that b will be negative. If you are able to picture the situation at hand (as in Figure 5.4.7), then it will likely be easier to handle questions of this nature than if you strictly rely on algebraic manipulations.

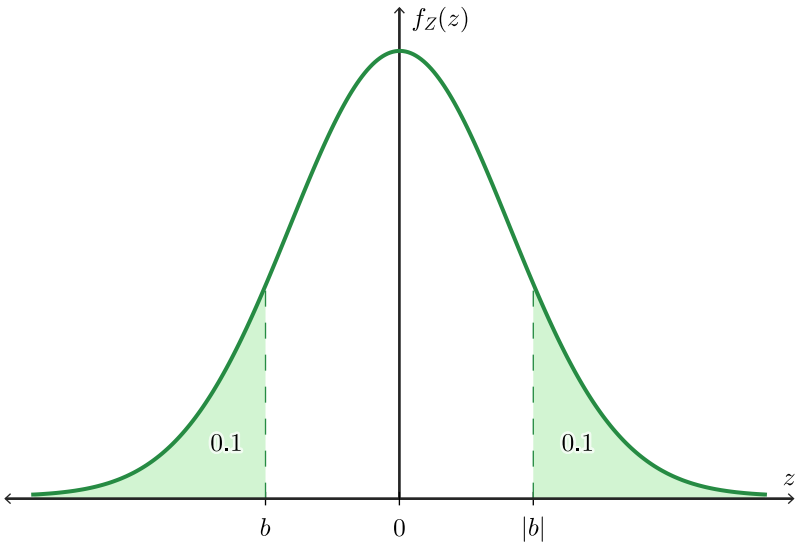


Figure 5.4.7: Useful picture in helping to find b satisfying $P(Z > b) = 0.9$ in Example 5.4.3

Finally, if we wish to have $P(-c < Z < c) = 0.95$, this means that the probability of Z being outside the interval $(-c, c)$ must be equal to 0.05, and by symmetry, this is evenly split between the area above c and the area below $-c$, as demonstrated by Figure 5.4.8.
Therefore,

$$P(Z < -c) = P(Z > c) = 0.025,$$

and so

$$P(Z \leq c) = 0.975.$$

Looking up the value for $p = 0.975$ using Table 5.4.2, we see that $P(Z \leq 1.96) = 0.975$. Hence, $c = 1.96$. ■

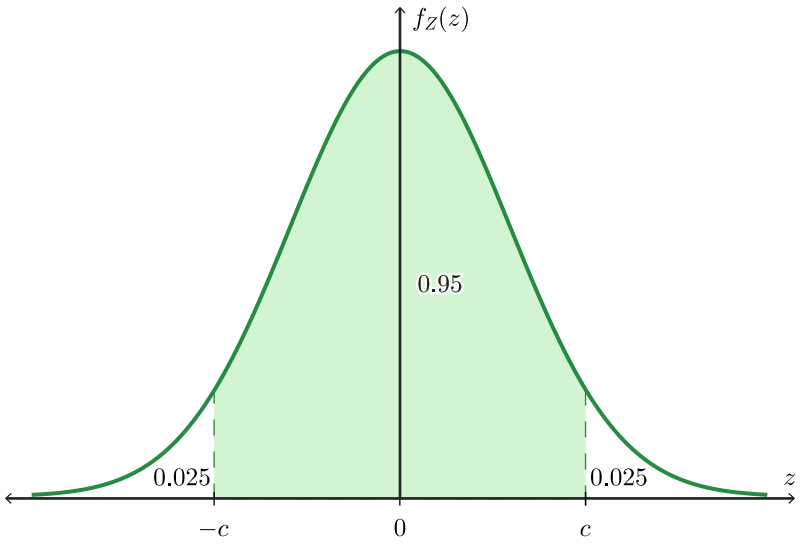


Figure 5.4.8: Useful picture in helping to find c satisfying $P(-c < Z < c) = 0.95$ in Example 5.4.3

Now that we are comfortable working with both standard normal tables, let us conclude this subsection with two examples involving non-standard normal distributions.

Example 5.4.4. Suppose that $X \sim N(3, 25)$. Calculate $P(X < 2)$ and find the number c such that $P(X > c) = 0.95$.

Solution: Since $X \sim N(3, 25)$, we have that $\mu = 3$ and $\sigma = 5$. If we apply the standardizing procedure mentioned earlier, then we obtain

$$\begin{aligned} P(X < 2) &= P\left(\frac{X-\mu}{\sigma} < \frac{2-3}{5}\right) \\ &= P(Z < -0.20) \quad \text{where } Z \sim N(0,1) \\ &= P(Z > 0.20) \\ &= 1 - F_Z(0.20) \\ &= 1 - 0.57926 \quad \text{using Table 5.4.1} \\ &= 0.42074. \end{aligned}$$

Next, the number c must satisfy

$$0.95 = P(X > c) = P\left(\frac{X-\mu}{\sigma} > \frac{c-3}{5}\right) = P\left(Z > \frac{c-3}{5}\right) \quad \text{where } Z \sim N(0,1).$$

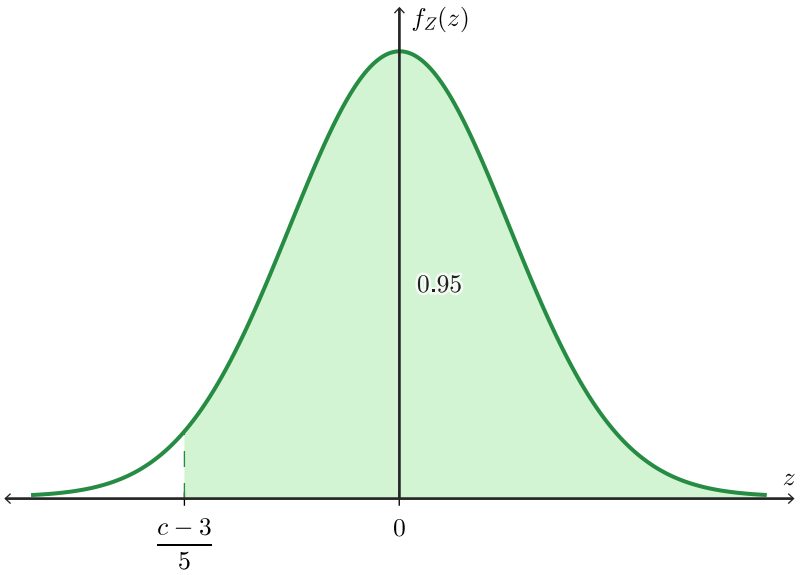


Figure 5.4.9: Useful picture in helping to find c satisfying $P\left(Z > \frac{c-3}{5}\right) = 0.95$ in Example 5.4.4

Figure 5.4.9 depicts the above probability statement. From Table 5.4.2, we also have that

$$0.95 = P(Z < 1.6449) = P(Z > -1.6449) \text{ by symmetry.}$$

Therefore, it follows that

$$\frac{c-3}{5} = -1.6449 \Longrightarrow c = 3 - 5(1.6449) = -5.2245.$$

■

Example 5.4.5. The distribution of heights of adult males in Canada is well-approximated by a normal distribution with a mean of 69 inches and a standard deviation of 2.4 inches. Find the 20^{th} and 80^{th} percentiles of the height distribution of adult males in Canada.

Solution: We are told that if X is the height of a randomly selected Canadian adult male, then $X \sim N(69, 2.4^2)$. To find $q(0.8)$, the 80^{th} percentile of X, we require $q(0.8)$ to satisfy

$$0.8 = P(X \leq q(p)) = P\left(\frac{X-69}{2.4} \leq \frac{q(p)-69}{2.4}\right) = P\left(Z \leq \frac{q(p)-69}{2.4}\right) \text{ where } Z \sim N(0,1).$$

From Table 5.4.2, we see that $P(Z \leq 0.8416) = 0.8$, and this immediately leads to

$$\frac{q(0.8)-69}{2.4} = 0.8416 \Longrightarrow q(0.8) = 69 + 2.4(0.8416) = 71.0198.$$

In other words, 71.0198 inches is the 80^{th} percentile of the height distribution.

Similarly, to find $q(0.2)$ such that $P(X \leq q(0.2)) = P\left(Z \leq \frac{q(0.2)-69}{2.4}\right) = 0.2$, we now use the fact that

$$0.8 = P(Z < 0.8416) = P(Z > -0.8416) = 1 - P(Z \leq -0.8416) \Longrightarrow P(Z \leq -0.8416) = 0.2.$$

Therefore, we now set

$$\frac{q(0.2) - 69}{2.4} = -0.8416$$

and solve to get $q(0.2) = 69 - 2.4(0.8416) = 66.9802$ inches as the 20^{th} percentile of the height distribution. ■

Section 5.4 Problems

5.4.1 Suppose that $X \sim U(a, b)$. Find the probability distribution of the random variable $Y = cX + d$ where c and d are real constants.

5.4.2 The diameters (in centimeters) of spherical particles produced by a machine are randomly distributed according to a $U(0.6, 1.0)$ distribution. Determine the probability distribution of the volume of a particle.

5.4.3 Suppose that you want to simulate observations from a probability distribution having pdf

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

by using the random number generator on a computer to generate $U(0, 1)$ numbers. What value from the above distribution would you take if you generated the random number $y = 0.27125$?

5.4.4 The magnitudes of earthquakes in a region of North America can be modelled by an exponential distribution with a mean of 2.5 (measured on the Richter scale).

(a) If 3 earthquakes occur independently in a given month, what is the probability that none of them exceed 5 on the Richter scale?

(b) If an earthquake exceeds 4 on the Richter scale, what is the probability it also exceeds 5?

5.4.5 Jamie believes that the total number of thousands of kilometers that a used car can be driven before it needs to be scrapped is an exponentially distributed random variable with mean 20 thousand kilometres. Sam has a used car that he claims has been driven 10 thousand kilometers.

(a) If Jamie purchases the car from Sam, what is the probability that Jamie would get at least 20 thousand additional kilometers out of it?

(b) Repeat the calculation in part (a) but under the assumption that the lifetime age of the car (measured in thousands of kilometres) is uniformly distributed over the interval $(0, 40)$.

5.4.6 In a bank with online terminals, the time the system runs between disruptions has an exponential distribution with a mean of θ hours. One quarter of the time the system shuts down within 8 hours of the previous disruption. Determine the value of θ.

5.4.7 A certain type of light bulb has lifetimes that follow an exponential distribution with a mean of 1000 hours. Determine the 25^{th}, 50^{th}, and 75^{th} percentiles of the lifetime distribution.

5.4.8 Suppose that $X \sim Exp(\theta)$. Use the pdf of X and integration by parts to directly find $E(X)$ and $E(X^2)$.

5.4.9 Consider the pdf of X given in Problem 5.1.5. Suppose that $Y = \frac{X^2}{\theta}$. Show that $Y \sim Exp(1)$.

5.4.10 A police station is to be located along a road of infinite length – stretching from point 0 outward to ∞. If the distance of an emergency call from point 0 is exponentially distributed with mean θ, where should the police station be located so as to minimize the expected distance from the emergency call? In other words, what is the value of a which minimizes $E(|X - a|)$ when $X \sim Exp(\theta)$?

5.4.11 Suppose that $X \sim N(\mu, \sigma^2)$. What percent of the time does X lie within one standard deviation of the mean? Two standard deviations? Three standard deviations?

5.4.12 Suppose that $X \sim N(10, 16)$. Find the 20^{th}, 40^{th}, 60^{th}, and 80^{th} percentiles of X.

5.4.13 The examination scores obtained by a large group of students can be modelled by a normal distribution with a mean of 65% and a standard deviation of 10%. Find the percentage of students who obtain each of the following letter grades:

$$A\ (\geq 80\%),\ B\ (70 - 80\%),\ C\ (60 - 70\%),\ D\ (50 - 60\%),\ \text{and}\ F\ (< 50\%).$$

5.4.14 The number of litres X that a filling machine in a water bottling plant deposits in a nominal two litre bottle follows a $N(\mu, \sigma^2)$ distribution, where $\sigma = 0.01$ and μ is the setting on the machine.

(a) If $\mu = 2$, what is the probability that a bottle has less than 2 litres of water in it?

(b) Determine c such that $P(|X - \mu| \leq c) = 0.9$.

(c) What should μ be set at to make the probability a bottle has less than 2 litres be less than 0.01?

5.4.15 A manufacturer produces bolts that are specified to be between 1.19 and 1.21 inches in diameter. If the production process results in a bolt's diameter being normally distributed with mean 1.20 inches and standard deviation 0.005 inches, what percentage of bolts will not meet specifications?

5.4.16 Suppose that the diameters in millimeters (mm) of the eggs laid by a large flock of hens can be modelled by a normal distribution with a mean of 40 mm and a variance of 4 mm^2. The wholesale selling price is 5 cents for an egg less than 37 mm in diameter, 6 cents for eggs between 37 and 42 mm, and 7 cents for eggs over 42 mm. What is the average wholesale price per egg?

5.4.17 If $Z \sim N(0,1)$, find the pdf of $Y = Z^2$. (*Note*: The probability distribution you obtain is known in the literature as the *chi-squared distribution with 1 degree of freedom.*)

5.5 Use of the Normal Distribution in Approximations

One of the main reasons why the normal distribution is so commonly used in practice is that it tends to accurately approximate the probability distribution of sums of random variables. This remarkable property follows from a very important result known as the **Central Limit Theorem**.

To motivate this idea, let us consider an experiment in which we throw n fair six-sided dice. Suppose that we are interested in the random variable S_n, representing the sum of the upturned faces on the n thrown dice. For the case when $n = 1$, it is clear that $S_1 \sim DU(1,6)$. The shape of the pmf of S_1 is indicated by the histogram in the first panel of Figure 5.5.1. In the case when $n = 2$, the possible values of S_2 lie in the set $\{2, 3, \ldots, 12\}$ as shown in the following table, along with their corresponding probabilities:

s	2	3	4	5	6	7	8	9	10	11	12
$P(S_2 = s)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

The probability histogram for S_2 is displayed in the second panel of Figure 5.5.1. Finally, for the sum of the upturned faces on three thrown dice, the possible values of S_3 lie in the set $\{3, 4, \ldots, 18\}$ as shown in the following table, along with their corresponding probabilities:

s	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$P(S_3 = s)$	$\frac{1}{216}$	$\frac{3}{216}$	$\frac{6}{216}$	$\frac{10}{216}$	$\frac{15}{216}$	$\frac{21}{216}$	$\frac{25}{216}$	$\frac{27}{216}$	$\frac{27}{216}$	$\frac{25}{216}$	$\frac{21}{216}$	$\frac{15}{216}$	$\frac{10}{216}$	$\frac{6}{216}$	$\frac{3}{216}$	$\frac{1}{216}$

The probability histogram for S_3 is shown in the third panel of Figure 5.5.1. Note that this histogram is already starting to resemble the bell-shaped features of a normal distribution.

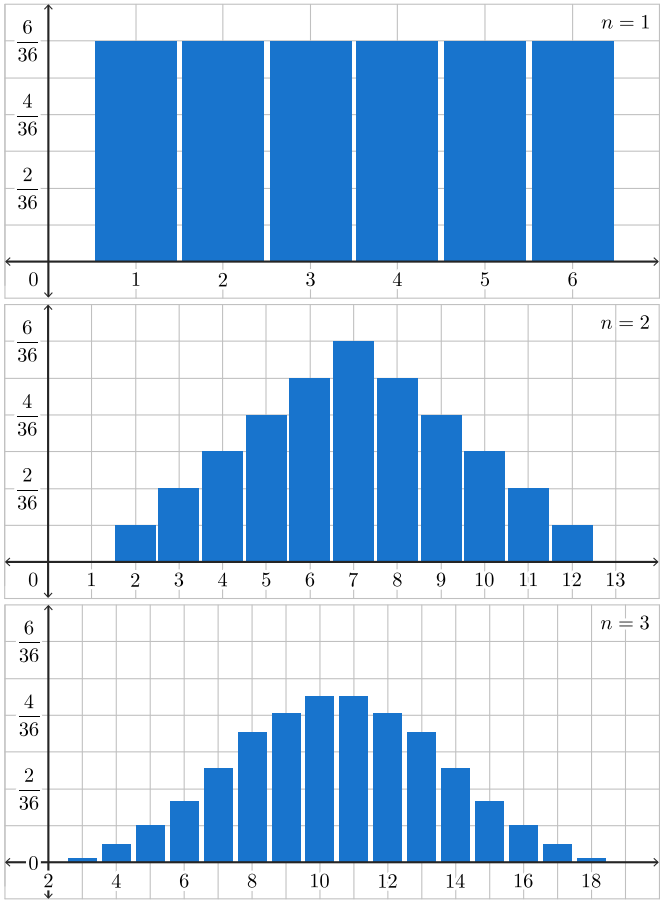


Figure 5.5.1: Probability histograms for the sum of the upturned faces on n thrown dice for $n = 1, 2, 3$

This experiment illustrates what happens in general with the probability distribution of a sum of independent random variables chosen from any underlying distribution. The following theorem is the mathematical statement of this phenomenon (its proof, however, is omitted since it is beyond the scope of this course).

Theorem 5.5.1. ***(Central Limit Theorem (CLT)):*** *Let $X_1, X_2, \ldots, X_n$ be independent random variables that all have the same probability distribution, with common mean μ and common variance σ^2. Let $S_n = \sum_{i=1}^{n} X_i$. Then, as $n \to \infty$, the cdf of the random variable*

$$\frac{S_n - n\mu}{\sigma\sqrt{n}}$$

approaches the $N(0,1)$ cdf.

Remarks:

(1) An equivalent way to express the result of the CLT is through the *sample mean* instead of the sum. Specifically, if the conditions of the CLT are satisfied and we let

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{S_n}{n},$$

then the cdf of the random variable

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

approaches the $N(0,1)$ cdf as $n \to \infty$.

(2) The CLT works for essentially all distributions (discrete or continuous) which $X_1, X_2, \ldots, X_n$ could be chosen from. The only exception occurs when X_i has a probability distribution whose mean or variance is not finite. Such probability distributions do exist, but they are rare.

(3) From a practical standpoint, the CLT can be used to approximate the probability distribution of sums or averages. As such, we will apply it when n is large, but finite, to approximate the probability distribution of $\sum_{i=1}^{n} X_i$ or $\bar{X}_n$ by a normal distribution. In other words, we will use

$$\sum_{i=1}^{n} X_i \sim N\left(n\mu, n\sigma^2\right) \text{ approximately for large } n$$

and

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right) \text{ approximately for large } n.$$

(4) The accuracy of the approximation depends on n (i.e., bigger is better) and also on the actual probability distribution of the random variables $X_1, X_2, \ldots, X_n$. The approximation works better for small values of n when the shape of the pmf/pdf of X_i is symmetric (e.g., the $U(a,b)$ distribution) or nearly symmetric (e.g., the $Poi(5)$ distribution). A general rule of thumb is that $n \geq 30$ is considered "sufficiently large" for the CLT to apply.

(5) It is worth noting that in the special case when $X_1, X_2, \ldots, X_n$ are independent $N(\mu, \sigma^2)$ random variables, the CLT actually becomes an exact result (holding true for all n) and not simply an approximate one for large n. Stated even more generally, it can be shown (although we will not prove this formally) that if $X_1, X_2, \ldots, X_n$ are independent random variables such that $X_i \sim N(\mu_i, \sigma_i^2)$ for $i = 1, 2, \ldots, n$, then

$$\sum_{i=1}^{n} c_i X_i \sim N\left(\sum_{i=1}^{n} c_i\mu_i, \sum_{i=1}^{n} c_i^2\sigma_i^2\right) \text{ for all } n = 1, 2, 3, \ldots.$$

As a result, if $X_1, X_2, \ldots, X_n$ themselves have a normal distribution, then $\sum_{i=1}^{n} X_i$ and $\bar{X}_n$ have exact normal distributions regardless of the value of n. On the other hand, if the $X_1, X_2, \ldots, X_n$ are not normally distributed, then $\sum_{i=1}^{n} X_i$ and $\bar{X}_n$ are approximately normally distributed when n is large due to the CLT. From this distinction, and echoing the comment made in remark (4), you should be able to surmise that if the distribution of $X_1, X_2, \ldots, X_n$ is somewhat "normally-shaped", then the approximation will be good for smaller values of n than if the distribution of $X_1, X_2, \ldots, X_n$ is very "non-normal" in shape.

Example 5.5.1. Hamburger patties are packed 40 to a box, and each box is supposed to have 5 kilograms of meat inside it. The weights of the patties vary a little because they are mass produced, and the weight of a single patty is actually a continuous random variable with mean $\mu = 0.128$ kg and standard deviation $\sigma = 0.03$ kg. Find the probability a box has at least 5 kg of meat, assuming that the weights of the 40 hamburger patties in any given box are independent.

Solution: Let $X_1, X_2, \ldots, X_{40}$ be the weights of 40 independent hamburger patties. Therefore, the random variable $Y = X_1 + X_2 + \cdots + X_{40}$ represents their total weight in a box. Note that $n = 40 > 30$, implying that it is reasonable to assume that Y has approximately a $N(\underbrace{40\mu}_{=5.12}, \underbrace{40\sigma^2}_{=0.036})$ distribution by the CLT. We wish to calculate

$$\begin{aligned} P(Y \geq 5) &\approx P\left(Z \geq \frac{5 - 5.12}{\sqrt{0.036}}\right) \text{ where } Z \sim N(0,1) \\ &= P(Z \geq -0.63) \\ &= P(Z \leq 0.63) \\ &= 0.73565. \end{aligned}$$

■

Example 5.5.2. Starting on the first day of the new year, suppose that the times between any two consecutively-reported fires to a fire station are independent exponentially distributed random variables with a mean of 4 hours. Find the probability the 500^{th} fire of the year is reported on the 84^{th} day of the year.

Solution: First of all, let X_1 be the time to the first reported fire. For $i = 2, 3, \ldots, 500$, let X_i be the time between the $(i-1)^{\text{th}}$ and i^{th} reported fires. Based on the given information (and noting that 4 hours equals $\frac{1}{6}$ of a day), it follows that $X_1, X_2, \ldots, X_{500}$ are independent random variables where $X_i \sim Exp\left(\frac{1}{6}\right)$ for each $i = 1, 2, \ldots, 500$. Since $\sum_{i=1}^{500} X_i$ represents the time of the 500^{th} fire, we wish to calculate

$$P\left(83 < \sum_{i=1}^{500} X_i \leq 84\right).$$

Although the exponential distribution is quite skewed (i.e., non-symmetric) and not at all bell-shaped, we are summing a very large number of independent random variables. Hence, by the CLT, $\sum_{i=1}^{500} X_i$ has approximately a $N\left(500\mu, 500\sigma^2\right)$ distribution, where $\mu = E(X_i)$ and $\sigma^2 = Var(X_i)$. For the $Exp\left(\frac{1}{6}\right)$ distribution, we immediately have that $\mu = \frac{1}{6}$ and $\sigma^2 = \left(\frac{1}{6}\right)^2 = \frac{1}{36}$. Therefore, this leads to

$$\begin{aligned}
P\left(83 < \sum_{i=1}^{500} X_i \leq 84\right) &\approx P\left(\frac{83 - \frac{500}{6}}{\sqrt{500/36}} < Z \leq \frac{84 - \frac{500}{6}}{\sqrt{500/36}}\right) \text{ where } Z \sim N(0,1) \\
&= P(-0.09 < Z \leq 0.18) \\
&= P(Z \leq 0.18) - P(Z \leq -0.09) \\
&= P(Z \leq 0.18) - P(Z \geq 0.09) \\
&= P(Z \leq 0.18) - (1 - P(Z < 0.09)) \\
&= 0.57142 + 0.53586 - 1 \\
&= 0.10728.
\end{aligned}$$

■

The previous two examples showed how the CLT can be used approximate the distribution of a sum of continuous random variables. This next example, albeit a bit frivolous, demonstrates how the normal distribution can even approximate sums of discrete random variables.

Example 5.5.3. In an orchard, suppose that the number X of worms in an apple has pmf given by

x	0	1	2	3
$f(x)$	0.4	0.3	0.2	0.1

Find the probability that a basket with 250 apples has between 225 and 260 (inclusive) worms in it.

Solution: Let X_i count the number of worms present in the i^{th} apple, $i = 1, 2, \ldots, 250$. For each value of i, we have that

$$\begin{aligned}
\mu = E(X_i) &= \sum_{x=0}^{3} x f(x) = 0(0.4) + 1(0.3) + 2(0.2) + 3(0.2) = 1, \\
E(X_i^2) &= \sum_{x=0}^{3} x^2 f(x) = (0)^2(0.4) + (1)^2(0.3) + (2)^2(0.2) + (3)^2(0.2) = 2, \text{ and} \\
\sigma^2 = Var(X_i) &= E(X_i^2) - \mu^2 = 2 - (1)^2 = 1.
\end{aligned}$$

With these computed values for μ and σ^2, we know that $Y = \sum_{i=1}^{250} X_i$ has approximately a $N(250, 250)$

distribution by the CLT. Therefore, it follows that

$$\begin{aligned} P(225 \le Y \le 260) &\approx P\left(\frac{225-250}{\sqrt{250}} \le Z \le \frac{260-250}{\sqrt{250}}\right) \text{ where } Z \sim N(0,1) \\ &= P(-1.58 \le Z \le 0.63) \\ &= P(Z \le 0.63) - P(Z < -1.58) \\ &= P(Z \le 0.63) - P(Z > 1.58) \\ &= P(Z \le 0.63) - (1 - P(Z \le 1.58)) \\ &= 0.73565 + 0.94295 - 1 \\ &= 0.6786. \end{aligned}$$

Although this approximation is adequate, we can improve its accuracy in the following manner. When X_i has a discrete probability distribution, as it does here, $\sum_{i=1}^{n} X_i$ will always remain discrete no matter how large n gets. Therefore, the distribution of $\sum_{i=1}^{n} X_i$, while bell-shaped and approximately normal, will never be precisely normal. In this particular example, consider the probability histogram for the random variable $Y = \sum_{i=1}^{250} X_i$, as shown in Figure 5.5.2. Note that only part of the histogram is displayed.

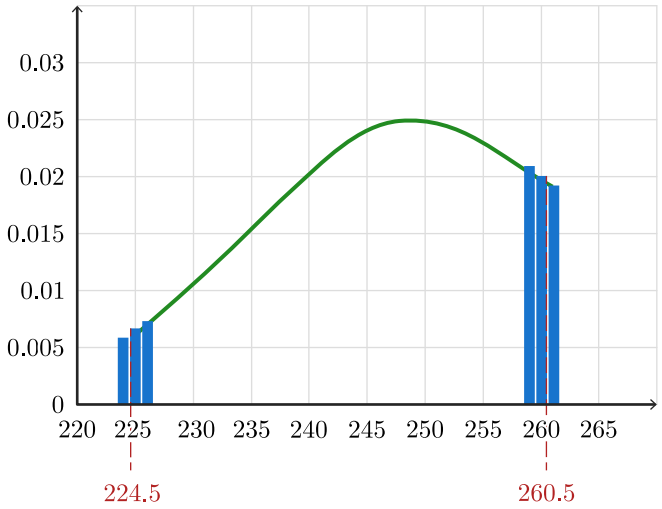


Figure 5.5.2: A portion of the probability histogram for $Y = \sum_{i=1}^{250} X_i$ in Example 5.5.3

The area of each bar of this histogram represents the probability at the Y-value in the centre of the interval (which has length 1). The smooth curve is the pdf of the approximating normal distribution. Therefore, the required probability, given by $\sum_{y=225}^{260} P(Y = y)$, would be the total area of all bars of the

histogram corresponding to Y ranging from 225 to 260. Upon closer inspection, these bars actually span continuous y-values from 224.5 to 260.5. The left and right end bars are more easily seen in Figure 5.5.3.

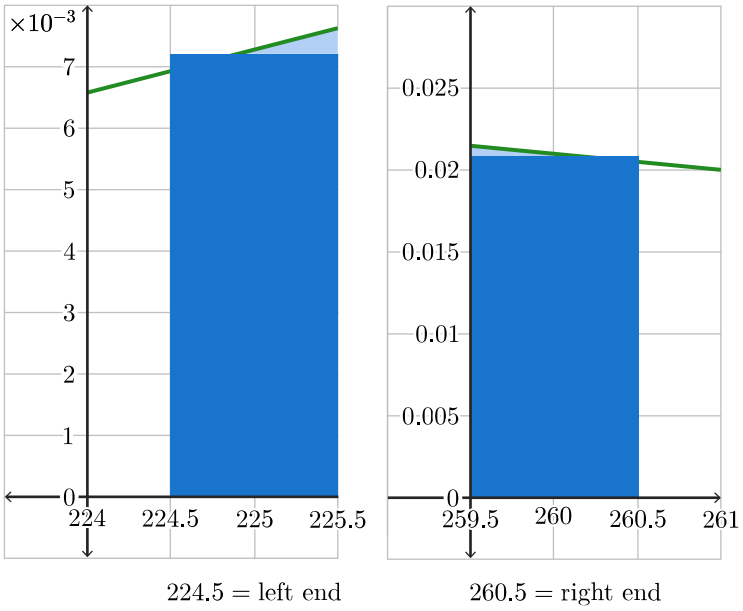


Figure 5.5.3: A magnification of the bars at the extreme ends of the desired interval in Example 5.5.3

Therefore, we could then obtain a more accurate approximation by calculating the area under the normal pdf curve from 224.5 to 260.5, namely

$$\begin{aligned}
P(225 \le Y \le 260) &= P(224.5 \le Y \le 260.5) \\
&\approx P\left(\frac{224.5-250}{\sqrt{250}} \le Z \le \frac{260.5-250}{\sqrt{250}}\right) \text{ where } Z \sim N(0,1) \\
&= P(-1.61 \le Z \le 0.66) \\
&= P(Z \le 0.66) - P(Z < -1.61) \\
&= P(Z \le 0.66) - P(Z > 1.61) \\
&= P(Z \le 0.66) - (1 - P(Z \le 1.61)) \\
&= 0.74537 + 0.94630 - 1 \\
&= 0.69167.
\end{aligned}$$

Unless making this adjustment greatly complicates the solution, it is generally preferable to make this so-called "**continuity correction**". ■

Remarks:

(1) A continuity correction should not be applied when approximating a **continuous** distribution by the normal distribution. Since the correction involves going halfway to the next possible value, there would be no adjustment to make if the random variable takes on a continuum of real values.

(2) Rather than trying to guess or remember when to add 0.5 and when to subtract 0.5, it is often helpful to sketch a histogram and shade the bars you wish to include. It should then be obvious which value to use.

(3) Whenever approximating the probability of a single value for a discrete random variable, such as the probability that Y is equal to 225 from Example 5.5.3, we do need to apply the continuity correction. Without it, we obtain the uninformative approximation

$$P(Y = 225) \approx P\left(Z = \frac{225 - 250}{\sqrt{250}}\right) = P(Z = -1.58) = 0 \text{ since } Z \sim N(0, 1).$$

In such a situation, we should be using the continuity correction to obtain

$$\begin{aligned} P\left(\sum_{i=1}^{500} X_i = 225\right) &= P\left(224.5 \leq \sum_{i=1}^{500} X_i \leq 225.5\right) \\ &\approx P\left(\frac{224.5 - 250}{\sqrt{250}} \leq Z \leq \frac{225.5 - 250}{\sqrt{250}}\right) \text{ where } Z \sim N(0, 1) \\ &= P(-1.61 \leq Z \leq -1.55) \\ &= P(1.55 \leq Z \leq 1.61) \text{ by the symmertry of the } N(0, 1) \text{ distribution} \\ &= P(Z \leq 1.61) - P(Z < 1.55) \\ &= 0.94630 - 0.93953 \\ &= 0.00677, \end{aligned}$$

and although this is small, it is certainly not zero.

Recall from Example 4.4.2 that if $Y \sim Bin(n, p)$, then it is possible to express Y as

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

where $X_1, X_2, \ldots, X_n$ are independent $Bin(1, p)$ random variables with $\mu = E(X_i) = p$ and $\sigma^2 = Var(X_i) = p(1 - p)$. Therefore, by the CLT, the cdf of the random variable

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} = \frac{Y - np}{\sqrt{np(1 - p)}}$$

approaches the $N(0,1)$ cdf as $n \to \infty$. In other words, for large n, Y is approximately normally distributed with mean np and variance $np(1-p)$. In a similar fashion (in particular, see Example 4.3.3 and the remark that follows), one could readily verify that the following results also hold true:

(1) If $Y \sim Poi(n)$, then the cdf of the random variable

$$\frac{Y-n}{\sqrt{n}}$$

approaches the $N(0,1)$ cdf as $n \to \infty$.

(2) If $Y \sim NB(n,p)$, then the cdf of the random variable

$$\frac{Y-\frac{n(1-p)}{p}}{\sqrt{n(1-p)/p^2}} = \frac{pY-n(1-p)}{\sqrt{n(1-p)}}$$

approaches the $N(0,1)$ cdf as $n \to \infty$.

Example 5.5.4. Suppose that $X \sim Poi(9)$. Apply the normal approximation to calculate $P(X > 9)$ and then compare the approximation with the exact probability.

Solution: If we use the normal approximation without applying a continuity correction, then we simply obtain

$$P(X > 9) \approx P\left(Z > \frac{9-9}{3}\right) = P(Z > 0) = 0.5 \text{ since } Z \sim N(0,1).$$

Calculating the exact value to 6 decimal places of accuracy, we get

$$P(X > 9) = 1 - P(X \leq 9) = 1 - \sum_{i=0}^{9} \frac{e^{-9}9^i}{i!} = 1 - 0.587408 = 0.412592.$$

Note that there is a considerable difference here between the exact value 0.412592 and the normal approximation of 0.5 since the value of $n = 9$ is still quite small. However, if we use the continuity correction when we apply the normal approximation, then we obtain

$$\begin{aligned}
P(X > 9) &= P(X \geq 10) \\
&= P(X > 9.5) \\
&\approx P\left(Z > \frac{9.5-9}{3}\right) \text{ where } Z \sim N(0,1) \\
&= P(Z > 0.17) \\
&= 1 - P(Z \leq 0.17) \\
&= 1 - 0.56749 \\
&= 0.43251,
\end{aligned}$$

which is much closer to the true value of 0.412592. ■

Example 5.5.5. Suppose that $X \sim Bin(100, 0.4)$. Apply the normal approximation to calculate $P(34 \leq X \leq 48)$ and then compare the approximation with the exact probability.

Solution: We assume that X is approximately normal with mean $100(0.4) = 40$ and variance $100(0.4)(0.6) = 24$. Without applying the continuity correction, we obtain

$$\begin{aligned} P(34 \leq X \leq 48) &\approx P\left(\frac{34-40}{\sqrt{24}} \leq Z \leq \frac{48-40}{\sqrt{24}}\right) \text{ where } Z \sim N(0,1) \\ &= P(-1.23 \leq Z \leq 1.63) \\ &= P(Z \leq 1.63) - P(Z < -1.23) \\ &= P(Z \leq 1.63) - P(Z > 1.23) \\ &= P(Z \leq 1.63) - (1 - P(Z \leq 1.23)) \\ &= 0.94845 + 0.89065 - 1 \\ &= 0.8391. \end{aligned}$$

With the continuity correction applied, we get

$$\begin{aligned} P(34 \leq X \leq 48) &= P(33.5 \leq X \leq 48.5) \\ &\approx P\left(\frac{33.5-40}{\sqrt{24}} \leq Z \leq \frac{48.5-40}{\sqrt{24}}\right) \text{ where } Z \backsim N(0,1) \\ &= P(-1.33 \leq Z \leq 1.74) \\ &= P(Z \leq 1.74) - P(Z < -1.33) \\ &= P(Z \leq 1.74) - P(Z > 1.33) \\ &= P(Z \leq 1.74) - (1 - P(Z \leq 1.33)) \\ &= 0.95907 + 0.90824 - 1 \\ &= 0.86731. \end{aligned}$$

The exact value to 6 decimal places of accuracy is

$$P(34 \leq X \leq 48) = \sum_{x=34}^{48} \binom{100}{x} (0.4)^x (0.6)^{100-x} = 0.866445.$$

Once again, the approximation with the continuity correction is more accurate. ■

Remark: The error of the normal approximation decreases as n increases, but it is generally a good idea to use the continuity correction when it is convenient to do so. For example, if we are using a normal approximation to a discrete distribution like the binomial which takes integer values and the standard deviation of the binomial distribution is less than 10, then the continuity correction makes

a difference of $0.5/10 = 0.05$ to the number we look up in Table 5.4.1. This can result in a difference in the probability of up to around 0.02. If you are willing to tolerate errors in probabilities of that magnitude, your rule of thumb might be to apply the continuity correction whenever the standard deviation of the integer-valued random variable being approximated is less than 10.

Example 5.5.6. Suppose that p represents the true (but unknown) proportion of Canadians who think Canada should adopt the US dollar. If a sample of n Canadians are randomly selected and asked whether they think Canada should adopt the US dollar, let X be the number who say yes so that $\frac{X}{n}$ is the sample proportion of people who say yes. Determine the number n who must be surveyed so that there is at least a 95% chance that $\frac{X}{n}$ lies within 0.02 of p.

Solution: We begin by noting that although the exact distribution of X is hypergeometric, it is reasonable to view X as approximately binomial, since the total population of Canada from which we are sampling is large. As a result, we simply assume that $X \sim Bin(n, p)$. If we now apply the normal approximation to the binomial distribution, then X is approximately normal with mean np and variance $np(1-p)$. We wish to find the value of n such that

$$P\left(\left|\frac{X}{n} - p\right| \leq 0.02\right) \geq 0.95,$$

or equivalently,

$$\begin{aligned}
P\left(|X - np| \leq 0.02n\right) &\geq 0.95 \\
P\left(\frac{|X - np|}{\sqrt{np\left(1-p\right)}} \leq \frac{0.02n}{\sqrt{np\left(1-p\right)}}\right) &\geq 0.95 \\
P\left(|Z| \leq \frac{0.02\sqrt{n}}{\sqrt{p\left(1-p\right)}}\right) &\geq 0.95 \text{ where } Z \sim N\left(0,1\right) \\
2P\left(Z \leq \frac{0.02\sqrt{n}}{\sqrt{p\left(1-p\right)}}\right) - 1 &\geq 0.95 \\
P\left(Z \leq \frac{0.02\sqrt{n}}{\sqrt{p\left(1-p\right)}}\right) &\geq 0.975.
\end{aligned}$$

From Table 5.4.2, we have that $P(Z \leq 1.96) = 0.975$, which immediately implies that

$$\frac{0.02\sqrt{n}}{\sqrt{p\left(1-p\right)}} \geq 1.96 \text{ or } n \geq \left(\frac{1.96}{0.02}\right)^2 p\left(1-p\right).$$

Unfortunately, this does not give us an explicit expression for n because we do not know the value of p. However, the way out of this dilemma is to use the maximum value for

$$\left(\frac{1.96}{0.02}\right)^2 p(1-p).$$

If we choose n to be this large, then we can be sure of having the required precision in our estimate, $\frac{X}{n}$, for any p. It is straightforward to see that $p(1-p)$ is maximized when $p = 0.5$. Taking $p = 0.5$, we therefore obtain

$$n \geq \left(\frac{1.96}{0.02}\right)^2 (0.5)(0.5) \geq 2401.$$

In other words, if we survey $n = 2401$ Canadians, then we can be at least 95% sure that $\frac{X}{n}$ lies within 0.02 of p, regardless of the value of p. ■

Remark: This method of determining n is used when poll results are reported in the media. For instance, you often see or hear that "this poll is accurate to within 2 percent 19 times out of 20". This is saying that n was big enough so that $P(p - 0.02 \leq \frac{X}{n} \leq p + 0.02)$ was 95%. As we just discovered, this requires a value of n of 2401.

Section 5.5 Problems

5.5.1 When people are asked to make up a "random" number between 0 and 1, rather than the $U(0,1)$ distribution which would be expected, it has been found that the distribution of the made-up number X has pdf given by

$$f(x) = \begin{cases} 4x & \text{for } 0 < x < \frac{1}{2}, \\ 4(1-x) & \text{for } \frac{1}{2} < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

(a) For 100 "random" numbers independently selected from the above distribution, approximate the probability that their sum lies between 49.0 and 50.5.

(b) What would the answer to part (b) be if instead the 100 numbers selected were truly $U(0,1)$?

5.5.2 Tomato seeds germinate (i.e., sprout to produce a plant) independently of each other, with probability 0.8 of each seed germinating. Give an expression for the probability that at least 75 seeds out of 100 which are planted in the soil germinate. Evaluate this probability using a suitable approximation.

5.5.3 A metal parts manufacturer inspects each part produced. Suppose that 60% are acceptable as produced, 30% have to be repaired, and 10% are beyond repair and must be scrapped. It costs the manufacturer \$10 to repair a part, and \$100 (in lost labour and materials) to scrap a part. Find the approximate probability that the total cost associated with inspecting 80 parts will exceed \$1200.

5.5.4 Student examination scores can be modelled by a normal distribution with a mean of 65% and a standard deviation of 10%. Find the probability that the average score in a random group of n students exceeds 70% when $n = 10$ and $n = 25$.

5.5.5 A turbine shaft is made up of four different sections. The lengths of those sections are independent and have normal distributions with different values of μ and σ: (8.10, 0.22), (7.25, 0.20), (9.75, 0.24), and (3.10, 0.20). What is the probability an assembled turbine shaft meets the desired specifications of 28 ± 0.26?

5.5.6 In a survey of n voters from a given riding in Canada, the sample proportion who say they would vote Conservative is used to estimate p, the probability a voter would vote Conservative. If Conservative support is actually 16%, how large should n be so that with probability 0.95, the estimate will be in error at most 0.03?

5.5.7 Suppose that the unemployment rate in Canada is 7%.

(a) Find the approximate probability that in a random sample of 10,000 persons in the labour force, the number of unemployed will be between 675 and 725 inclusive. Since $n = 10,000$ is large, a continuity correction is not required.

(b) How large a random sample would it be necessary to choose so that, with probability 0.95, the proportion of unemployed persons in the sample is between 6.9% and 7.1%?

5.5.8 **Crown and Anchor.** *Crown and Anchor* is a game that is sometimes played at charity casinos or just for fun. It can be played with a "wheel of fortune" or with 3 dice, in which each die has its 6 sides labelled with a crown, an anchor, and the four card suits club, diamond, heart and spade, respectively. You bet an amount (say \$1) on one of the 6 symbols: let us suppose you bet on "heart". The 3 dice are then rolled simultaneously and you win \$$x$ if x hearts turn up, $x = 0, 1, 2, 3$. Let X represent your profits from playing the game n times. Use a normal approximation to calculate the probability that X is positive when $n = 10, 25$, and 50.