PROBABILITY

STAT 220/230 COURSE NOTES FALL 2021 Edition

MARKET SNAPSHOT NIKKEI ▼ FTSE 100 ▼ DOW ▼ MORE DATA

% ↓ Nasdaq 2,458.14 -2.93% ↓ Dow 11,198.10 -2.15% ↓ S&P 500 1,167.92 -2.62% ↓ FTSE 100 5,157.81 -1.70% ↓ STOXX 50 2,345.38 -1.25%

STAT 220/230 COURSE NOTES

By Chris Springer Revised by Jerry Lawless, Don McLeish and Cyntha Struthers

Fall 2021 Edition

Contents

[1]

[1] Sections 5.3, 5.5, 5.8, 5.9, Chapter 6, Sections 8.4, 9.3, 10.2 and 10.3 are usually optional for STAT 220. Chapter 6, Sections 8.4, 9.3 and 10.3 are usually optional for STAT 230.

1. INTRODUCTION TO PROBABILITY

1.1 Definitions of Probability

You are the product of a random universe. From the Big Bang to your own conception and birth, random events have determined who we are as a species, who you are as a person, and much of your experience to date. Ironic therefore that we are not well-tuned to understanding the randomness around us, perhaps because millions of years of evolution have cultivated our ability to see regularity, certainty and deterministic cause-and-effect in the events and environment about us. We are good at finding patterns in numbers and symbols, or relating the eating of certain plants with illness and others with a healthy meal. In many areas, such as mathematics or logic, we assume we know the results of certain processes with certainty (e.g., $2 + 3 = 5$), though even these are often subject to assumed axioms. Most of the real world, however, from the biological sciences to quantum physics[2], involves variability and uncertainty. 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; the number of requests to a web server is uncertain. Uncertainty or "randomness" (that is, 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, hair colour, blood type, etc.), and (2) *variability in processes* or phenomena (e.g., the random selection of six numbers from forty-nine numbers in a lottery draw can lead to a very large number of different outcomes). Which of these would you use to describe the fluctuations in stock prices or currency exchange rates?

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 simulation

[2] "As far as the laws of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality" Albert Einstein, 1921.

of queues and networks or the spread of epidemics. Of course we do not have the time in this course to develop these applications in detail, but some of the problems at the end of the chapter 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 two pips or dots will be on the upturned face 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, 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 two pips on the upturned face is $\frac{1}{6}$ because only one of the six faces has two pips.

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 two pips on the upturned face when a die is rolled 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.

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 can't 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: 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 this course, 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 and use in the remaining chapters 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 (numbers between 0 and 1) to events is specified.

Of course in a given run of the random experiment, a particular event may or may not occur.

In order to understand the material in these notes, you may need to review your understanding of basic counting arguments, elementary set theory as well as some of the important series that you have encountered in Calculus that provide a basis for some of the distributions discussed in these notes. In Chapter 2, we begin a more mathematical description of probability theory.

1.2 Chapter 1 Problems

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

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

   (a) A person's birthday is in April

   (b) A driver makes a claim on their car insurance in the next year

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

   (d) The disk in a personal computer crashes

3. Give examples of how probability applies to each of the following areas.

   (a) Lottery draws

   (b) Public opinion polls

   (c) Sending data over a network

   (d) Auditing of expense items in a financial statement

   (e) Disease transmission (e.g. measles, tuberculosis, STD's)

4. Which of the following can be accurately described by a "deterministic" model, that is, 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 lifetime of a heavy smoker

   (d) The value of a stock which was purchased for $20 one month ago

   (e) The number of servers at a large data center which crash on a given day

2. MATHEMATICAL PROBABILITY MODELS

2.1 Sample Spaces and Probability

Consider some phenomenon or process which is repeatable, at least in theory, and suppose that certain events or outcomes $A_1, A_2, A_3, \ldots$ are defined. We will often term the phenomenon or process an **"experiment"** and refer to a single repetition of the experiment as a **"trial"**. The probability of an event A, denoted $P(A)$, is a number between 0 and 1. For probability to be a useful mathematical concept, it should possess some other properties. For example, 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 $P(A_1) = 0.6$ and $P(A_2) = 0.6$, so that $P(A_1) + P(A_2) > 1$. (Why is this so? Is there a fundamental reason or have we simply adopted 1 as a convenient scale?) To avoid this sort of thing we begin with the following definition.

Definition 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, as the following example shows.

Example: Roll a six-sided die, and define the events

$$a_i = \text{there are } i \text{ pips on the top face, for } i = 1, 2, \ldots, 6$$

Then we could take the sample space as $S = \{a_1, a_2, \ldots, a_6\}$. (Note we use the curly brackets "$\{\ldots\}$" to indicate the elements of a set). Instead of using this definition of the sample space we could instead define the events

E : the event that there are an even number of pips on the top face

O : the event that there are an odd number of pips on the top face and take $S = \{E, O\}$. Both sample spaces satisfy the definition. 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 "there are less than three pips on the top face" 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 preferred 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 simple events. Recall that a countably infinite sequence is one that can be put into a one-to-one correspondence with 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 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. For the next few chapters we consider only discrete sample spaces. For discrete sample spaces it is much easier to specify the class of events to which we may wish to assign probabilities; we will allow all possible subsets of the sample space. For example if $S = \{a_1, a_2, a_3, a_4, a_5, a_6\}$ is the sample space then $A = \{a_1, a_2, a_3, a_4\}$ and $B = \{a_6\}$ and S itself are all examples of events.

Definition 2 *An event in a discrete sample space is a subset $A \subset S$. 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 such as $A = \{a_1, a_2\}$ is called a **compound event.***

Note that the notation $A \subset B$ means $a \in A$ implies $a \in B$.

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. When we mean the probability of the event $A_1 = \{a_1\}$, we should write $P(A_1)$ or $P(\{a_1\})$ but the latter is often shortened to $P(a_i)$. In the case of a discrete sample space it is easy to specify probabilities of events since they are determined by the probabilities of simple events.

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

(1) $0 \leq P(a_i) \leq 1$

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

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

Note that $P(*)$ is a function whose domain is the sample space S. 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

$\{a_i\}$ 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 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 $P(A) = \sum\limits_{a \in A} P(a)$.*

For example, the probability of the compound event $A = \{a_1, a_2, a_3\}$ is $P(a_1) + P(a_2) + P(a_3)$. Probability theory does not say what numbers 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.

Example: Suppose a six-sided die is rolled, and let the sample space be $S = \{1, 2, \ldots, 6\}$, where i represents the simple event that there are i pips on the top face, $i = 1, 2, \ldots, 6$. If the die is an ordinary one, (a *fair* die) we would likely define probabilities as

$$P(i) = \frac{1}{6} \text{ for } i = 1, 2, \ldots, 6 \tag{2.1}$$

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 outcome 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 (2.1), if we wish to consider some compound event, the probability is easily obtained. For example, if $A =$ "there are an even number of pips on the top face" then because $A = \{2, 4, 6\}$ we get $P(A) = P(2) + P(4) + P(6) = \frac{1}{2}$.

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, find $P(A)$ by adding the probabilities of all the simple events that make up A.

Later we will discover that having a detailed specification or list of the elements of the sample space may be difficult. Indeed in many cases the sample space is so large that at best we can describe it in words. For the present 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 $P(A)$.

When S has only a few points, one of the easiest methods for finding the probability of an event is to list all outcomes. 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: Draw one card from a standard well-shuffled deck (13 cards of each of 4 suits - spades, hearts, diamonds, clubs). Find the probability that the card is a club.

Solution 1: Let $S =$\{spade, heart, diamond, club\}. Then S has 4 points, with 1 of them being "club", so P(club) $= \frac{1}{4}$.

Solution 2: Let $S = \{$ 2♦,3♦,4♦$, \ldots, A$♦$, 2\heartsuit, \ldots, A$♣$\}$. Then each of the 52 cards in S has probability $\frac{1}{52}$. The event A of interest is

$$A = \{2♣, 3♣, \ldots, A♣\}$$

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

$$P(A) = \frac{1}{52} + \frac{1}{52} + \cdots + \frac{1}{52} = \frac{13}{52} = \frac{1}{4}$$

Note 1: A sample space is not necessarily unique, as mentioned earlier. The two solutions illustrate this. Note that in the first solution the event $A =$ "the card is a club" is a simple event because of the way the sample space was defined, but in the second it is a compound event.

Note 2: In solving the problem we have assumed that each simple event in S is equally probable. For example in Solution 1 each simple event has probability $1/4$. 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 equiprobable.

The term "odds" is sometimes used in describing probabilities. In this card example the odds in favour of clubs are $1 : 3$; we could also say the odds against clubs are $3 : 1$. In general,

Definition 5 *The odds in favour of an event A is the probability the event occurs divided by the probability it does not occur or $\frac{P(A)}{1-P(A)}$. The odds against the event is the reciprocal of this, $\frac{1-P(A)}{P(A)}$.*

If the odds against a given horse winning a race are 20 to 1 (or $20 : 1$), what is the corresponding probability that the horse will win the race? According to the definition above $\frac{1-P(A)}{P(A)} = 20$, which gives $P(A) = \frac{1}{21}$. Note that these odds are derived from bettor's collective opinion and therefore subjective.

Example: Toss a coin twice. Find the probability of getting one head. (In this course, "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 1^{st} toss and tails on the 2^{nd}.) 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 $P\,(1 \text{ head}) = \frac{1}{3}$.

Which solution 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 event $\{1 \text{ head}\}$ occurs more often than either $\{0 \text{ heads}\}$ or $\{2 \text{ heads}\}$ in actual repeated trials.

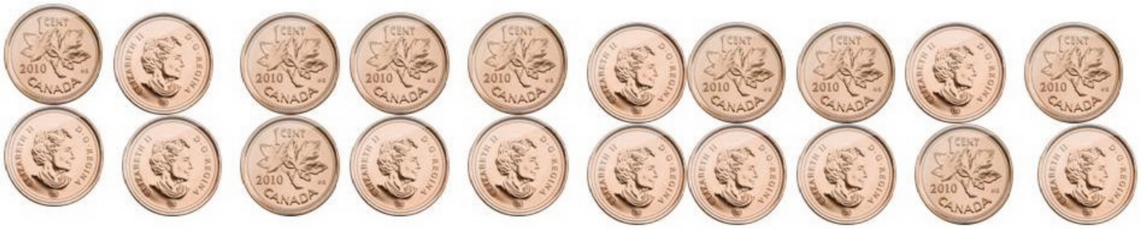


Figure 2.1: Ten tosses of two coins.

You can experiment to verify this (for example of the 10 replications of the experiment in Figure 2.1, 2 heads occurred 2 of the 10 times, 1 head occurred 7 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 $\frac{1}{4}$. 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: Roll a red die and a green die. Find the probability of the event $A =$ "the total number of pips showing on the top 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 is

$$\begin{array}{lccccc} S=\{ & (1,1) & (1,2) & (1,3) & \cdots & (1,6) \\ & (2,1) & (2,2) & (2,3) & \cdots & (2,6) \\ & (3,1) & (3,2) & (3,3) & \cdots & (3,6) \\ & & -- & -- & -- & \\ & (6,1) & (6,2) & (6,3) & \cdots & (6,6)\} \end{array}$$

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

Example: Suppose the 2 dice were identical in colour. Find the probability of the event A.

Solution 1: Since we can no longer distinguish between (x, y) and (y, x), the only distinguishable points in S are:

$$\begin{array}{lccccc} S=\{ & (1,1) & (1,2) & (1,3) & \cdots & (1,6) \\ & & (2,2) & (2,3) & \cdots & (2,6) \\ & & & (3,3) & \cdots & (3,6) \\ & & & & .. & .. \\ & & & & & (6,6)\} \end{array}$$

Using this sample space, we have $A = \{(1, 4), (2, 3)\}$. If we assign equal probability $\frac{1}{21}$ to each point (simple event) then we get $P(A) = \frac{2}{21}$.

At this point you should be suspicious since $\frac{2}{21} \neq \frac{4}{36}$. 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 $(1, 1)$. So the only sensible way to use this sample space so it is consistent with the real world is to assign probabilities $\frac{1}{36}$ 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. On the website *http://www.math.duke.edu/education/postcalc/probability/dice/index.html* you may throw virtual dice up to $10,000$ times and record the results. For example on 1000 throws of two dice (see Figure 2.2), there were 121 occasions when the total on the dice was 5, indicating the probability of the event A is close to $\frac{121}{1000}$ or 0.121. This compares with the probability $P(A) = \frac{4}{36} = 0.111$.

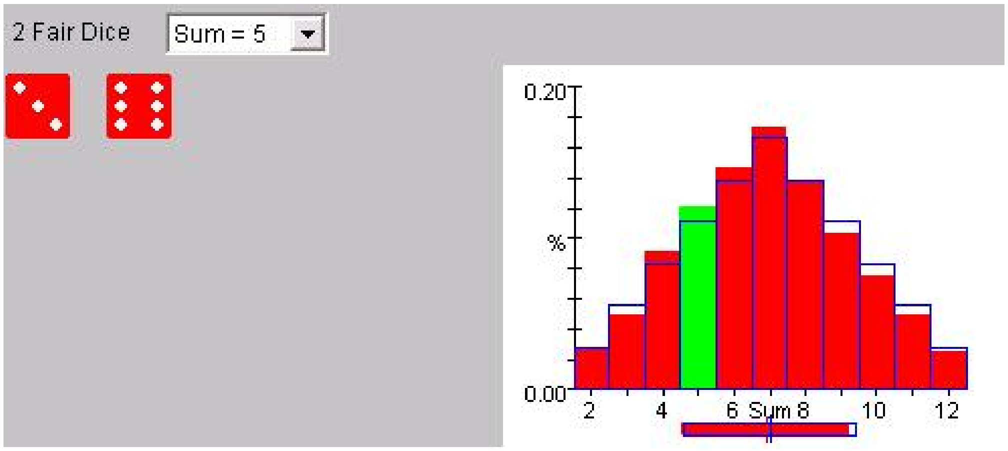


Figure 2.2: Results of 1000 throws of 2 dice

Solution 2: For a more straightforward solution to the above problem, pretend the dice can be distinguished. (Imagine, for example, that we put tiny mark on one die, or label one of them differently.) We then get the same 36 sample points as in the example with the red die and the green die. The fact that one die has a tiny mark cannot change the probabilities so that $P(A) = \frac{4}{36}$. 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 can't be, in order to get equally likely sample points.

While the method of finding 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 three dice were tossed there would be 216 points in S). We need to have more efficient ways of determining the number of outcomes in S or in a compound event without having to list them all. Chapter 3 considers ways to do this, and then Chapter 4 develops other ways to manipulate and calculate probabilities.

Although we often consider simple problems involving things such as coins, dice and simple games, probability is used to deal with a huge variety of practical problems from finance to clinical trials. In some settings such as in Problems 6 and 7 below, we need to rely on previous repetitions of an experiment, or on related scientific data, to assign numerical probabilities to events.

2.2 Chapter 2 Problems

1. Students in a particular program have the same four math professors. Two students in the program each independently ask one of their math professors[3] for a letter of reference. Assume each is equally likely to ask any of the math professors.

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

   (b) Use this sample space to find the probability both students ask the same professor.

2. A fair coin is tossed three times.

   (a) List a sample space for this experiment.

   (b) Find the probability of two heads.

   (c) Find the probability of exactly two consecutive tails.

3. Two numbers are chosen at random without replacement from the set $\{1, 2, 3, 4, 5\}$.

   (a) List a sample space for this experiment.

   (b) Find the probability both numbers are odd.

   (c) Find the probability the numbers chosen differ by one, that is, the two numbers are consecutive.

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 the 24 equally probable outcomes 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";
   D: "exactly three letters go into the correct envelopes".

   (c) Find the probability of each event in (b).

[3] "America believes in education: the average professor earns more money in a year than a professional athlete earns in a whole week." Evan Esar (1899 - 1995)

5. Three balls are placed at random in three boxes, with no restriction on the number of balls per box.

   (a) List the 27 equally probable outcomes of this experiment. Be sure to explain your notation.

   (b) Find the probability of each of the following events:
   A: "the first box is empty";
   B: "the first two boxes are empty";
   C: "no box contains more than one ball".

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

   (d) Find the probabilities of events A, B and C when k balls are placed in n boxes $(n \geq k)$.

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 (the person does not have the disease but the test says they do) or a "false negative" result (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:

   920 persons without the disease tested negative

   60 persons without the disease tested positive

   18 persons with the disease tested positive

   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.

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:

   $$\begin{aligned} &p(0,6) = p(6,0) = 0.004; \;\; p(0,0) = p(6,6) = 0.096 \\ &p(5,9) = p(9,5) = 0.005; \;\; p(5,5) = p(9,9) = 0.095 \\ &p(4,7) = p(7,4) = 0.002; \;\; p(4,4) = p(7,7) = 0.098 \\ &p(y,y) = 0.100 \;\text{ for } y = 1, 2, 3, 8 \end{aligned}$$

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

8. In Problems 4-7, what can you say about how appropriate you think the probability model is for the experiment being modelled?

9. **Challenge Problem:** Professor X has an integer $(1 \leq m \leq 9)$ in mind and asks two students, Allan and Beth to pick numbers between 1 and 9. Whichever is closer to m gets 90% and the other 80% in STAT 230. If they are equally close, they both get 85%. If the professor's number and that of Allen are chosen purely at random and Allen announces his number out loud, describe a sample space and a strategy which leads Beth to the highest possible mark.

3. PROBABILITY AND COUNTING TECHNIQUES

Some probability problems can be solved by specifying a sample space $S = \{a_1, a_2, \ldots, a_n\}$ in which each simple event has probability $\frac{1}{n}$, that is, each event is "equally likely". This is referred to as a uniform distribution over the set $\{a_1, a_2, \ldots, a_n\}$. In a ***uniform probability model,*** we can calculate the probability of any event A by counting the number of outcomes in the event A,

$$P(A) = \frac{\text{number of outcomes in } A}{\text{number of outcomes in } S}$$

In other words, we need to be able to count the number of events in S which are in A. We now look at techniques for counting outcomes from "experiments".

3.1 Addition and Multiplication Rules

There are two helpful rules for counting, phrased in terms of "jobs" which are to be done.

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 class has 30 men and 25 women. There are $30 + 25 = 55$ ways the instructor can pick one student to answer a question. If there are 5 vowels and 20 consonants on a list and I must pick one letter, this can be done in 5+20 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.*

For example, if there are 5 vowels and 20 consonants and I must choose one consonant followed by one vowel for a two-letter word, this can be done in 20×5 ways (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, which gives $3 \times 7 = 21$ 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.

Example: Suppose we pick 2 numbers from digits 1, 2, 3, 4, 5 with replacement. (Note: "with replacement" means that after the first number is picked it is "replaced" in the set of numbers, so it could be picked again as the second number.) Assume a uniform distribution on the sample space, that is, assume that every pair of numbers has the same probability. Let us find the probability that one number is even. This can be reworded as: "The first number is even AND the second is odd (this can be done in 2×3 ways) OR the first 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 number can be chosen in 5 ways AND the second in 5 ways, S contains $5 \times 5 = 25$ points and since each point has the same probability, they all have probability $\frac{1}{25}$. Therefore

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

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** being replaced, some special rules may simplify the solution.

Note: The phrases *at random*, or *uniformly* are often used to mean that all of the points in the sample space are equally likely so that in the above problem, every possible pair of numbers chosen from this set has the same probability $\frac{1}{25}$.

Problems

3.1.1 (a) A course has 4 sections with no limit on how many can enrol in each section. Three students each pick a section at random.

(i) Specify the sample space S.

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

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

(iv) Find the probability nobody picks section 1.

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

3.1.2 Canadian postal codes consist of 3 letters (of 26 possible letters) alternated with 3 digits (of the 10 possible), starting with a letter (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:

(a) all 3 letters are the same?

(b) the digits are all even or all odd? Treat 0 as being neither even nor odd.

3.1.3 Suppose a password has to contain between six and eight digits, with each digit either a letter or a number from 1 to 9. The password must contain at least one number.

(a) What is the total number of possible passwords?

(b) If you started to try passwords in random order, what is the probability you would find the correct password for a given situation within the first 1,000 passwords you tried?

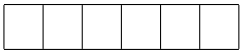
3.2 Counting Arrangements or Permutations

In many problems, the sample space is a set of arrangements or sequences. These are classically called *permutations*. A key step in the argument is to be sure to understand what it is you are counting. It is helpful to invent a notation for the outcomes in the sample space and the events of interest (these are the objects you are counting).

Example: Suppose the letters a,b,c,d,e,f are arranged at random to form a six-letter word (an arrangement) – we must use each letter once only. The sample space

$$S = \{\text{abcdef, abcdfe, } \dots \text{, fedcba}\}$$

has a large number of outcomes and, because we formed the word "at random", we assign the same probability to each. To count the number of words in S, count the number of ways that we can construct such a word – each way corresponds to a unique word. Consider filling the boxes



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. Thus 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 there are

$6 \times 5 \times 4 = 120$ ways to fill the first three boxes. Applying the same reasoning, we 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 is e or f so

$$A = \{\text{afbcde, aebcdf}, \ldots, \text{efdcba}\}$$

We can count the number of outcomes in A using a similar argument if we start with the second box. We can fill the second box in 2 ways, that is, with an e or f. For each of these choices, we can then fill the first box in 5 ways, so now we can 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 so the number of outcomes in A is $10 \times 24 = 240$. Since we have a uniform probability model

$$P(A) = \frac{\text{number of outcomes in } A}{\text{number of outcomes in } S} = \frac{240}{720} = \frac{1}{3}$$

In determining the number of outcomes in A, it is important that we start with the second box. Suppose, instead, we start 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. If we used e or f in the first box, there is only one way to fill the second. If we used a, b, c or d for the first box, there are 2 ways of filling the second. We avoid this complication by starting with the second box.

We can generalize the above problem in several ways. In each case we count the number of arrangements by counting the number of ways we can fill the positions in the arrangement. Suppose we start with n symbols. Then we can make:

- $n \times (n-1) \times \cdots \times 1$ arrangements of length n using each symbol once and only once. This product is denoted by $n!$ ("n factorial"). Note that $n! = n \times (n-1)!$.
- $n \times (n-1) \times \cdots \times (n-\ k+1)$ arrangements of length k using each symbol at most once. This product is denoted by $n^{(k)}$ ("n to k factors"). Note that $n^{(k)} = \frac{n!}{(n-k)!}$.
- $n \times n \times \cdots \times n = n^k$ arrangements of length k using each symbol as often as we wish.

In Table 3.1 we see how quickly $n!$ increases as n increases.

n	1	2	3	4	5	6	7	8	9	10
$n!$	1	2	6	24	120	720	5040	40320	362880	3628800
$(n/e)^n\sqrt{2\pi n}$	0.9	1.9	5.8	23.5	118.0	710.1	4980.4	39902.4	359536.9	3598695.6

Table 3.1

Stirling's Approximation: For large n there is an approximation to $n!$ called Stirling's approximation. Note that the sequence $\{a_n\}$ is *asymptotically equivalent* to the sequence $\{b_n\}$ if $\lim\limits_{n\to\infty} \frac{a_n}{b_n} = 1$.

Stirling's approximation says that $n!$ is asymptotically equivalent to $(n/e)^n \sqrt{2\pi n}$. The relative error in the approximation $n! \approx (n/e)^n \sqrt{2\pi n}$ decreases quickly as n increases as can be seen in Table 3.1 and is less than 0.01 if $n \geq 8$.

For many problems involving sampling from a deck of cards or a reasonably large population, counting the number of cases by simple conventional means is virtually impossible, and we need the counting arguments dealt with here. The extraordinarily large size of populations, in part due to the large size of quantities like n^n and $n!$, is part of the reason that statistics, sampling, counting methods and probability calculations play such an important part in modern science and business.

Example: A pin number of length four is formed by randomly selecting four digits from the set $\{0, 1, 2, \ldots, 9\}$ **with replacement**. Find the probability of the events:

A: the pin number is even
B: the pin number has only even digits
C: all of the digits are unique
D: the pin number contains at least one 1.

Solution: Since we pick the digits with replacement, the outcomes in the sample space can have repeated digits. The sample space is

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

with 10^4 equally probable outcomes.

For the event $A = \{0000, 0002, \ldots, 9998\}$, we can select the last digit to be any one of $0, 2, 6, 4, 8$ in 5 ways. Then for each of these choices, we can select the first digit in 10 ways and so on. There are 5×10^3 outcomes in A and

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

For the event $B = \{0000, 0002, \ldots, 8888\}$, we can select the first digit in 5 ways, and for each of these choices, the second in 5 ways, and so on. There are 5^4 outcomes in B and

$$P(B) = \frac{5^4}{10^4} = \frac{1}{16}$$

For the event $C = \{0123, 0124, \ldots, 9876\}$, we can select the first digit in 10 ways and for each of these choices, the second in 9 ways and so on. There are $10 \times 9 \times 8 \times 7$ outcomes in C and so

$$P(C) = \frac{10 \times 9 \times 8 \times 7}{10^4} = \frac{10^{(4)}}{10^4} = \frac{63}{125}$$

For the event $D = \{0001, 0011, 0111, 1111, \ldots\}$, it is easier to count the number of outcomes in the complement of D, that is, the set of all outcomes in S but not in D. We denote this event

$\bar{D} = \{0000, 0002, \ldots, 9999\}$. There are 9^4 outcomes in $\bar{D}$ and so there are $10^4 - 9^4$ outcomes in D and

$$P(D) = \frac{10^4 - 9^4}{10^4} = \frac{3439}{10000}$$

For a general event A, the *complement* of A, denoted $\bar{A}$, is the set of all outcomes in S which are not in A. It is sometimes easier to count outcomes in the complement rather than in the event itself.

Example: A pin number of length four is formed by randomly selecting four digits from the set $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ **without replacement**. Find the probability of the events:

A: the pin number is even.

B: the pin number has only even digits

C: the pin number begins or ends with a 1

D: the pin number contains 1.

Solution: The sample space is

$$S = \{0123, 0132, \ldots, 6789\}$$

with $10^{(4)}$ equally probable outcomes. For the event $A = \{1230, 0134, \ldots, 9876\}$, we can select the last digit to be any one of $0, 2, 6, 4, 8$ in 5 ways. Then for each of these choices, we can select the first digit in 9 ways, the third in 8 ways and so on. There are $5 \times 9 \times 8 \times 7$ outcomes in A and

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

The event $B = \{0246, 0248, \ldots, 8642\}$. The pin numbers in B are all $5^{(4)}$ arrangements of length 4 using only the even digits $\{0, 2, 4, 6, 8\}$ and so

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

The event $C = \{1023, 0231, \ldots, 9871\}$. There are 2 positions for the 1. For each of these choices, we can fill the remaining three positions in $9^{(3)}$ ways and so

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

The event $D = \{1234, 2134, \ldots, 9871\}$. We can use the complement and count the number of pin numbers that do not contain a 1. There are $9^{(4)}$ pin numbers that do not contain 1 and so there are $10^{(4)} - 9^{(4)}$ that do contain a 1. Therefore

$$P(D) = 1 - P\left(\bar{D}\right) = 1 - \frac{9^{(4)}}{10^{(4)}} = \frac{2}{5}$$

3.3 Counting Subsets or Combinations

In some problems, the outcomes in the sample space are subsets of a fixed size. Here we look at counting such subsets. Again, it is useful to write a short list of the subsets you are counting.

Example: Suppose we randomly select a subset of three digits from the set $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ so that the sample space is

$$S = \{\{1, 2, 3\}, \{0, 1, 3\}, \{0, 1, 4\}, \ldots, \{7, 8, 9\}\}$$

All the digits in each outcome are unique, that is, we do not consider $\{1, 1, 2\}$ to be a subset of S. Also, the order of the elements in a subset is not relevant. This is true in general for sets; the subsets $\{1, 2, 3\}$ and $\{3, 1, 2\}$ are the same. To count the number of outcomes in S, we use what we have learned about counting arrangements. Suppose there are m such subsets. Using the elements of any subset of size 3, we can form 3! arrangements of length 3. For example, the subset $\{1, 2, 3\}$ generates the $3! = 6$ arrangements $123, 132, 213, 231, 312, 321$ and any other subset generates a different 3! arrangements so that the total number of arrangements of 3 digits taken without replacement from the set $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ is $3! \times m$. But we know the total number of arrangements is $10^{(3)}$ so $3! \times m = 10^{(3)}$. Solving we get

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

Number of subsets of size k: We use the combinatorial symbol $\binom{n}{k}$ ("n *choose* k") to denote the number of subsets of size k that can be selected from a set of n objects. By an argument similar to that above, if m denotes the number of subsets of size k that can be selected from n things, then $m \times k! = n^{(k)}$ and so we have

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

In the example above we selected the subset at random so each of the $\binom{10}{3} = 120$ subsets has the same probability $\frac{1}{120}$. We now find the probability of the following events:

A: the digit 1 is included in the selected subset
B: all the digits in the selected subset are even
C: at least one of the digits in the selected subset is less than or equal to 5

To count the outcomes in event A, we must have 1 in the subset and we can select the other two elements from the remaining 9 digits in $\binom{9}{2}$ ways. And so

$$P(A) = \frac{\binom{9}{2}}{\binom{10}{3}} = \frac{9^{(2)}/2!}{10^{(3)}/3!} = \frac{3}{10}$$

The event $B = \{\{0,2,4\},\{0,2,6\},\ldots\}$. We can form the outcomes in B by selecting 3 digits from the 5 even digits $\{0,2,4,6,8\}$ in $\binom{5}{3}$ ways. And so

$$P(B) = \frac{\binom{5}{3}}{\binom{10}{3}}$$

The event $C = \{\{0,1,2\},\{0,1,6\},\{0,6,7\},\ldots\}$. Here it is convenient to consider the complement $\overline{C}$ in which the outcomes are $\{\{6,7,8\},\{6,7,9\},\ldots\}$, that is, the subsets with all elements greater than 5. We can form the subsets in $\overline{C}$ by selecting a subset of size 3 from the set $\{6,7,8,9\}$ in $\binom{4}{3}$ ways. Therefore

$$\begin{aligned} P(C) &= 1 - P(\overline{C}) \\ &= 1 - \frac{\binom{4}{3}}{\binom{10}{3}} \end{aligned}$$

Example: Suppose a box contains 10 balls of which 3 are red, 4 are white and 3 are green. A sample of 4 balls is selected at random without replacement. Find the probability of the events:

E: the sample contains 2 red balls
F: the sample contains 2red, 1 white and 1 green ball
G: the sample contains 2 or more red balls

Solution: Imagine that we label the balls from 1 to 10 with labels $1,2,3$ being red, $4,5,6,7$ being white and $8,9,10$ being green. Construct a uniform probability model in which all subsets of size 4 are equally probable. The sample space is

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

and each outcome has probability $\frac{1}{\binom{10}{4}}$.

The event E: To count the number of outcomes in E, we can construct a subset with two red balls by first choosing the two red balls from the three in $\binom{3}{2}$ ways. For each of these choices we can select the other two balls from the seven non-red balls in $\binom{7}{2}$ ways so there $\binom{3}{2} \times \binom{7}{2}$ are outcomes in E and

$$P(E) = \frac{\binom{3}{2}\binom{7}{2}}{\binom{10}{4}} = \frac{3}{10}$$

The event $F = \{\{1,2,4,8\},\{1,2,4,9\},\ldots\}$. To count the number of outcomes in F, we can select the two red balls in $\binom{3}{2}$ ways, then the white ball in $\binom{4}{1}$ ways and the green ball in $\binom{3}{1}$ ways. So we have

$$P(F) = \frac{\binom{3}{2}\binom{4}{1}\binom{3}{1}}{\binom{10}{4}} = \frac{6}{35}$$

The event $G = \{\{1,2,3,4\},\{1,2,4,5\},\ldots\}$ has outcomes with both 2 and 3 red balls. We need to count these separately (see below). There are $\binom{3}{2}\binom{7}{2}$ outcomes with exactly two red balls and $\binom{3}{3}\binom{7}{1}$ outcomes with three red balls. Hence we have

$$P(G) = \frac{\binom{3}{2}\binom{7}{2} + \binom{3}{3}\binom{7}{1}}{\binom{10}{4}} = \frac{1}{3}$$

A **common mistake** is to count the outcomes in G as follows. There are $\binom{3}{2}$ ways to select two red balls and then for each of these choices we can select the remaining two balls from the remaining eight in $\binom{8}{2}$ ways. So the number of outcomes in G is $\binom{3}{2} \times \binom{8}{2}$. You can easily check that this 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 G 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 $\{1,2,3,4\}$. Or 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 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. two red balls, three red balls.

Properties of $\binom{n}{k}$: You should be able to prove the following for n and k non-negative integers:

1. $n^{(k)} = \frac{n!}{(n-k)!} = n(n-1)^{(k-1)}$ for $k \geq 1$
2. $\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n^{(k)}}{k!}$
3. $\binom{n}{k} = \binom{n}{n-k}$ for all $k = 0, 1, \ldots, n$
4. If we define $0! = 1$, then the formulas hold with $\binom{n}{0} = \binom{n}{n} = 1$.
5. $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$
6. **Binomial Theorem:** $(1+x)^n = \binom{n}{0} + \binom{n}{1}x + \binom{n}{2}x^2 + \ldots + \binom{n}{n}x^n$

In many problems, we can combine counting arguments for arrangements and subsets as in the next example.

Example: A binary sequence is an 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?

Solution: The sample space is

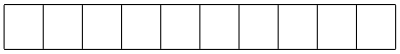
$$S = \{0000000000, 0000000001, \ldots, 1111111111\}$$

We can fill each of the 10 positions in the sequence in 2 ways and hence S has 2^{10} equally possible outcomes.

The event E with exactly 5 zeros and 5 ones is

$$E = \{0000011111, 1000001111, \ldots, 1111100000\}$$

To count the outcomes in E, think of constructing the sequence by filling the ten boxes



We can choose the 5 boxes for the zeros in $\binom{10}{5}$ ways and then the ones go in the remaining boxes in 1 way.

Hence we have

$$P(E) = \frac{\binom{10}{5}}{2^{10}}$$

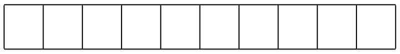
3.4 Number of Arrangements When Symbols Are Repeated

Example: Suppose the letters of the word STATISTICS are arranged at random. Find the probability of the event G that the arrangement begins and ends with S.

Solution: The sample space is

$$S = \{SSSTTTIIAC, SSSTTTIICA, \ldots\}$$

Here we need to count arrangements when some of the elements are the same. We construct the arrangements by filling ten boxes corresponding to the positions in the arrangement.



We can choose the three positions for the three S's in $\binom{10}{3}$ ways. For each of these choices, we can choose the positions for the three T's in $\binom{7}{3}$ ways. Then we can place the two Is in $\binom{4}{2}$ ways, then the

C in $\binom{2}{1}$ ways and finally the A in $\binom{1}{1}$ ways. The number of equally probable outcomes in the sample space S is

$$\binom{10}{3}\binom{7}{3}\binom{4}{2}\binom{2}{1}\binom{1}{1} = \frac{10!}{3!7!}\frac{7!}{3!4!}\frac{4!}{2!2!}\frac{2!}{1!1!}\frac{1!}{1!0!} = \frac{10!}{3!3!2!1!1!}$$

The event $G = \{$SSTTTIIACS,SSTTTIICAS,$\ldots\}$. To count the outcomes in G we must have S in the first and last position

S									S

Now we can use the same technique to arrange the remaining eight letters. Having placed two of the S's, there remain eight free boxes, in which we are to place three T's in $\binom{8}{3}$ ways, two I's in $\binom{5}{2}$ ways, one C in $\binom{3}{1}$ ways, one A in $\binom{2}{1}$ ways and finally the remaining S in the last empty box in $\binom{1}{1}$ way. There are

$$\binom{8}{3}\binom{5}{2}\binom{3}{1}\binom{2}{1}\binom{1}{1} = \frac{8!}{3!2!1!1!1!} = 3360$$

elements in G and

$$\begin{aligned} P(G) &= \frac{\frac{8!}{3!2!1!1!1!}}{\frac{10!}{3!3!2!1!1!}} = \frac{3360}{50400} \\ &= \frac{1}{15} \end{aligned}$$

Number of arrangements when symbols are repeated:

If we have n_i symbols of type i, $i = 1, 2, \ldots, k$ with $n_1 + n_2 + \cdots + n_k = n$, then the number of arrangements using all of the symbols is

$$\begin{aligned} &\binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdots\binom{n_k}{n_k} \\ &= \frac{n!}{n_1!n_2!\cdots n_k!} \end{aligned}$$

Example: Suppose we make a random arrangement of length 3 using letters from the set $\{a, b, c, d, e, f, g, h, i, j\}$. What is the probability of the event $B =$ "letters are in alphabetic order" if

(a) letters are selected without replacement?

(b) letters are selected with replacement?

Solution: For (a), the sample space is

$$\{abc, acb, bac, bca, cab, cba, \ldots, hij\}$$

with $10^{(3)}$ equally probable outcomes. The event $B = \{abc, abd, \ldots, hij\}$. To count the outcomes in B, we first select the three (different) letters to form the arrangement in $\binom{10}{3}$ ways. There is then one way to make an arrangement with the selected letters in alphabetic order. So we have

$$P(B) = \frac{\binom{10}{3}}{10^{(3)}} = \frac{1}{6}$$

For (b), the sample space is

$$\{aaa, aab, baa, aba, abc, acb, bac, bca, cab, cba, \ldots, hij\}$$

with 10^3 equally probable outcomes. To count the elements in B, consider the following cases:

Case 1: all three letters are the same. There are ten such arrangements $\{aaa, bbb, ccc, \ldots\}$ all in alphabetic order.

Case 2: there are two different letters e.g. $\{aab, aba, baa, abb, bab, bba\}$. We can choose the two letters in $\binom{10}{2}$ ways. For each of these choices, we can then make two arrangements with the letters in alphabetic order e.g. $\{aab, abb\}$ There are $2\binom{10}{2}$ arrangements in this case.

Case 3: all three letters are different. We can select the three letters in $\binom{10}{3}$ ways and then make one arrangement that is in alphabetic order (as in part (a)).

Combining the three cases, we have

$$P(B) = \frac{10 + 2\binom{10}{2} + \binom{10}{3}}{10^3} = \frac{11}{50}$$

Example: Suppose we make a four digit number by randomly selecting and arranging four digits from $1, 2, \ldots, 7$ without replacement. Find the probability that the number formed is

(a) even

(b) over 3000

(c) an even number over 3000.

Solution: Since we are forming a four digit number, the order in which the numbers are selected is important. We choose the sample space S to be the set of all possible arrangements of four numbers selected without replacement from the numbers $1, 2, \ldots, 7$. The sample space is

$$S = \{1234, 1243, 1324, 1342, \ldots, 4567, 4576, \ldots, 7654\}$$

with $7^{(4)}$ equally probable outcomes.

(a) For a number to be even, the last digit must be even. We can fill the last position in three ways with a 2, 4, or 6. The first three positions can be filled by choosing and arranging three of the six digits not used in the final position in $6^{(3)}$ ways. Then there are $3 \times 6^{(3)}$ ways to fill the final position AND the first three positions to produce an even number. Therefore the probability the number is even is $\frac{3\times 6^{(3)}}{7^{(4)}} = \frac{3}{7}$. Alternatively, the four digit number is even if and only if the last digit is even. The last digit is equally likely to be any one of the numbers $1, 2, \ldots, 7$ so the probability it is even is the probability it is either $2, 4$, or 6 or $\frac{3}{7}$.

(b) To get a number over 3000, we require the first digit to be 3, 4, 5, 6, or 7, that is, the first position can be filled in five ways. The remaining three positions can be filled in $6^{(3)}$ ways. Therefore the probability the number is greater than 3000 is $\frac{5\times 6^{(3)}}{7^{(4)}} = \frac{5}{7}$. Alternatively, note that the four digit number is over 3000 if and only if the first digit is one of 3, 4, 5, 6 or 7. Since each of $1, 2, \ldots, 7$ is equally likely to be the first digit, we get the probability the number is greater than 3000 is $\frac{5}{7}$.

In both (a) and (b) we dealt with positions which had restrictions first, before considering positions with no restrictions. This is generally the best approach to follow in applying counting techniques.

(c) This part has restrictions on both the first and last positions. To illustrate the complication this introduces, suppose we decide to fill positions in the order 1 then 4 then the middle two. We can fill position 1 in 5 ways. How many ways can we then fill position 4? The answer is either 2 or 3 ways, depending on whether the first position was filled with an even or odd digit. Whenever we encounter a situation such as this, we have to break the solution into separate cases. One case is where the first digit is even. The positions can be filled in 2 ways for the first (that is, with a 4 or 6), 2 ways for the last, and then $5^{(2)}$ ways to arrange 2 of the remaining 5 digits in the middle positions. This first case then occurs in $2 \times 2 \times 5^{(2)}$ ways. The second case has an odd digit in position one. There are 3 ways to fill position one (3, 5, or 7), 3 ways to fill position four (2, 4, or 6), and $5^{(2)}$ ways to fill the remaining positions. Case 2 then occurs in $3 \times 3 \times 5^{(2)}$ ways. We need case 1 OR case 2. Therefore the probability we obtain an even number greater than 3000 is

$$\frac{2 \times 2 \times 5^{(2)} + 3 \times 3 \times 5^{(2)}}{7^{(4)}} = \frac{13 \times 5^{(2)}}{7 \times 6 \times 5^{(2)}} = \frac{13}{42}$$

Another way to do this is to realize that we need only to consider the first and last digit, and to find P(first digit is ≥ 3 **and** last digit is even). There are $7 \times 6 = 42$ different choices for (first digit, last digit) and it is easy to see there are 13 choices for which first digit ≥ 3, last digit is even (5×3 minus the impossible outcomes $(4, 4)$ and $(6, 6)$). Thus the desired probability is $\frac{13}{42}$.

Exercise: Try to solve (c) by filling positions in the order 4, 1, middle. You should get the same answer.

Exercise: Can you spot the flaw in the following argument? There are $3 \times 6^{(3)}$ ways to get an even number (part (a)). There are $5 \times 6^{(3)}$ ways to get a number ≥ 3000 (part (b)). Therefore by the multiplication rule there are $\left(3 \times 6^{(3)}\right) \times \left(5 \times 6^{(3)}\right)$ ways to get a number which is even and > 3000.

Example: Five red balls and three white balls are arranged at random in a row. Find the probability that:

(a) the same colour is at each end

(b) the three white balls are together.

Solution: There are eight objects, five of one type and three of another, that is, five R's and three W's, so our sample space has $\frac{8!}{5!3!} = 56$ equally possible outcomes.

(a) To get the same colour at each end we need either

R							R	OR	W							W

The number of distinct arrangements with R at each end is $\frac{6!}{3!3!} = 20$, since we are arranging three R's and three W's in the middle six positions. The number with W at each end is $\frac{6!}{5!1!} = 6$. Thus

$$P(\text{same colour at each end}) = \frac{20+6}{56} = \frac{13}{28}$$

(b) Treating WWW as a single unit, we are arranging six objects, five R's and one object we might call "WWW". There are $\frac{6!}{5!1!} = 6$ arrangements. Thus,

$$P(\text{three white balls are together}) = \frac{6}{56} = \frac{3}{28}$$

Problems

3.4.1 Digits $1, 2, 3, \ldots, 7$ are arranged at random to form a 7 digit number. Find the probability that

(a) the even digits occur together, in any order

(b) the digits at the 2 ends are both even or both odd.

3.4.2 The letters of the word EXCELLENT are arranged in a random order. Find the probability that

(a) the word begins and ends with the same letter.

(b) X,C, and N occur together, in any order.

(c) the letters occur in alphabetical order.

3.5 Examples

Example: In the Lotto 6/49 lottery, six numbers are drawn at random, without replacement, from the numbers 1 to 49. Find the probability that

(a) the numbers $\{1, 2, 3, 4, 5, 6\}$ are drawn in any order.
(b) no even number is drawn.

Solution:

(a) Let the sample space S consist of all subsets of six numbers from $1, 2, \ldots, 49$; there are $\binom{49}{6}$ of them. Since $\{1, 2, 3, 4, 5, 6\}$ is one of these subsets, the probability of this particular set is

$$\frac{1}{\binom{49}{6}}$$

which is about 1 in 13.9 million.

(b) There are 25 odd and 24 even numbers, so there are $\binom{25}{6}$ choices in which all the numbers are odd. Therefore the probability no even number is drawn is the probability they are all odd, or

$$\frac{\binom{25}{6}}{\binom{49}{6}} = 0.0127$$

Example: Find the probability a bridge hand (13 cards picked at random from a standard deck[4] without replacement) has:

(a) 3 aces
(b) at least 1 ace
(c) 6 spades, 4 hearts, 2 diamonds, 1 club
(d) a 6-4-2-1 split between the 4 suits
(e) a 5-4-2-2 split.

Solution: Since order of selection does not matter, we take S to have $\binom{52}{13}$ outcomes, each with the same probability.

(a) We can choose 3 aces in $\binom{4}{3}$ ways. We also have to choose 10 other cards from the 48 non-aces. This can be done in $\binom{48}{10}$ ways. Hence the probability of exactly three aces is

$$\frac{\binom{4}{3}\binom{48}{10}}{\binom{52}{13}}$$

[4] A standard deck has 13 cards in each of four suits, hearts, diamonds, clubs and spades for a total of 52 cards. There are four aces in the deck (one of each suit).

(b) **Solution 1:** At least 1 ace means 1 ace or 2 aces or 3 aces or 4 aces. Calculate each part as in (a) and use the addition rule to get that the probability of at least one ace is

$$\frac{\binom{4}{1}\binom{48}{12}+\binom{4}{2}\binom{48}{11}+\binom{4}{3}\binom{48}{10}+\binom{4}{4}\binom{48}{9}}{\binom{52}{13}}$$

Solution 2: If we subtract all cases with 0 aces from the $\binom{52}{13}$ points in S we are left with all points having at least 1 ace. There are $\binom{4}{0}\binom{48}{13}=\binom{48}{13}$ possible hands with 0 aces since all cards must be drawn from the non-aces. (The term $\binom{4}{0}$ can be omitted since $\binom{4}{0}=1$, but was included here to show that we were choosing 0 of the 4 aces.) This gives that the probability of at least one ace is

$$\frac{\binom{52}{13}-\binom{48}{13}}{\binom{52}{13}}=1-\frac{\binom{48}{13}}{\binom{52}{13}}$$

Incorrect Solution: Choose 1 of the 4 aces then any 12 of the remaining 51 cards. This guarantees we have at least 1 ace, so the probability of at least one ace is $\frac{\binom{4}{1}\binom{51}{12}}{\binom{52}{13}}$. This is a common *incorrect* solution. The flaw in this solution is that it counts some points more than once by partially keeping track of order. For example, we could get the ace of spades on the first choice and happen to get the ace of clubs in the last 12 draws. We also could get the ace of clubs on the first draw and then get the ace of spades in the last 12 draws. Though in both cases we have the same outcome, they would be counted as 2 different outcomes. The strategies in solution 1 and 2 above are safer. We often need to inspect a solution carefully to avoid double or multiple counting.

(c) Choose the 6 spades in $\binom{13}{6}$ ways and the hearts in $\binom{13}{4}$ ways and the diamonds in $\binom{13}{2}$ ways and the clubs in $\binom{13}{1}$ ways. Therefore the probability of 6 spades, 4 hearts, 2 diamonds and one clubs is

$$\frac{\binom{13}{6}\binom{13}{4}\binom{13}{2}\binom{13}{1}}{\binom{52}{13}}=0.00196$$

(d) The split in (c) is only 1 of several possible 6-4-2-1 splits. In fact, filling in the numbers 6, 4, 2 and 1 in the spaces below

Spades	Hearts	Diamonds	Clubs

defines a 6-4-2-1 split. There are $4!$ ways to do this, and having done this, there are $\binom{13}{6}\binom{13}{4}\binom{13}{2}\binom{13}{1}$ ways to pick the cards from these suits. Therefore the probability of a a 6-4-2-1 split between the 4 suits is

$$\frac{4!\binom{13}{6}\binom{13}{4}\binom{13}{2}\binom{13}{1}}{\binom{52}{13}}=0.047$$

(e) This is the same question as (d) except the numbers 5-4-2-2 are not all different. There are $\frac{4!}{2!}$ different arrangements of 5-4-2-2 in the spaces below.

Spades	Hearts	Diamonds	Clubs

Therefore, the probability of a a 5-4-2-2 split is

$$\frac{\frac{4!}{2!}\binom{13}{5}\binom{13}{4}\binom{13}{2}\binom{13}{2}}{\binom{52}{13}} = 0.1058$$

Notes:

1. While $n^{(k)}$ only has a physical interpretation when n and k are non-negative integers with $n \geq k$, $n^{(k)}$ can still be defined when n is a real number and k is a non-negative integer. In general we can define

$$n^{(k)} = n(n-1)\cdots(n-k+1)$$

For example,

$$(-2)^{(3)} = (-2)(-2-1)(-2-2) = (-2)(-3)(-4) = -24$$

and

$$(1.3)^{(2)} = (1.3)(1.3-1) = 0.39$$

2. In order for $\binom{n}{0} = \binom{n}{n} = 1$ we must define

$$n^{(0)} = \frac{n!}{(n-0)!} = 1$$

and

$$0! = 1$$

3. When n is not a non-negative integer $\geq k$, $\binom{n}{k}$ loses its physical meaning. If n is a real number and k is a non-negative integer then we use

$$\binom{n}{k} = \frac{n^{(k)}}{k!} = \frac{n(n-1)\cdots(n-k+1)}{k!}$$

For example,

$$\binom{\frac{1}{2}}{3} = \frac{(\frac{1}{2})^{(3)}}{3!} = \frac{(\frac{1}{2})(-\frac{1}{2})(-\frac{3}{2})}{3!} = \frac{1}{16}$$

and

$$\binom{-7}{5} = \frac{(-7)^{(5)}}{5!} = \frac{(-7)\,(-8)(-9)(-10)\,(-11)}{(5)\,(4)\,(3)\,(2)\,(1)} = -462$$

4. When n and k are non-negative integers and $k > n$ then

$$\binom{n}{k} = \frac{n^{(k)}}{k!} = \frac{n(n-1)\cdots(1)\,(0)\cdots(n-k+1)}{k!} = 0$$

   For example,

$$\binom{3}{5} = 0$$

Problems

3.5.1 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 cars on the lot.

(a) Give an expression for the probability that at least 3 of these 8 cars will have faulty emission controls.

(b) What assumption does your answer to (a) require? How likely is it that this assumption holds if the inspector hopes to catch as many cars with faulty controls as possible?

3.5.2 In a race, the 15 runners are randomly assigned the numbers $1, 2, \ldots, 15$. Find the probability that

(a) 4 of the first 6 finishers have single digit numbers.

(b) the fifth runner to finish is the 3rd finisher with a single digit number.

(c) number 13 is the highest number among the first 7 finishers.

3.6 Useful Series and Sums

In remaining chapters the following series and sums will be useful. You will have seen some of these results in previous courses.

1. **Geometric Series:**

$$\sum_{i=0}^{n-1} t^i = 1 + t + t^2 + \cdots + t^{n-1} = \frac{1-t^n}{1-t} \quad \text{for} \;\; t \neq 1$$

If $|t| < 1$, then

$$\sum_{x=0}^{\infty} t^x = 1 + t + t^2 + \cdots = \frac{1}{1-t}$$

Other identities can be obtained from this one by differentiation. For example

$$\frac{d}{dt} \sum_{x=0}^{\infty} t^x = \frac{d}{dt} \left(\frac{1}{1-t} \right)$$

or

$$\sum_{x=1}^{\infty} x t^{x-1} = \frac{1}{(1-t)^2} \quad \text{for} \;\; |t| < 1$$

You should be able to determine other identities by taking second and higher derivatives.

2. **Binomial Theorem:** There are various forms of this theorem. We will use the form

$$\begin{aligned}(1+t)^n &= 1 + \binom{n}{1} t^1 + \binom{n}{2} t^2 + \cdots + \binom{n}{n} t^n \\ &= \sum_{x=0}^{n} \binom{n}{x} t^x\end{aligned}$$

where n is a positive integer and t is any real number.

Justification: One way of verifying this formula uses the counting arguments of this chapter. Imagine a product of the individual terms:

$$(1+t)(1+t)(1+t) \cdots (1+t)$$

To evaluate this product we must add together all of the possibilities obtained by taking one of the two possible terms from the first bracketed expression, that is, one of $\{1, t\}$, multiplying by one of $\{1, t\}$ taken from the second bracketed expression, etc. In how many ways do we obtain the term t^x where $x = 0, 1, 2, \ldots, n$? We might choose t from each of the first x terms above and then 1 from the remaining $(n-x)$ terms, or indeed we could choose t from any x of the n terms in $\binom{n}{x}$ ways and then 1 from the remaining $(n-x)$ terms.

3. **Binomial Theorem:** There is a more general version of the Binomial Theorem that results in an infinite series and that holds when n is not a positive integer:

$$(1+t)^n = \sum_{x=0}^{\infty} \binom{n}{x} t^x \quad \text{if } |t| < 1$$

Proof: Recall from Calculus the Maclaurin's series which says that a sufficiently smooth function $f(t)$ can be written as an infinite series using an expansion around $t = 0$,

$$f(t) = f(0) + \frac{f'(0)}{1}t + \frac{f''(0)}{2!}t^2 + \cdots$$

provided that this series is convergent. If $f(t) = (1+t)^n$, then $f\ (0) = 1$ and $f^{(k)}(0) = n^{(k)}$ for $k = 1, 2, \ldots$ and we obtain

$$(1+t)^n = 1 + \frac{n}{1}t + \frac{n(n-1)}{2!}t^2 + \cdots + \frac{n^{(k)}}{k!}t^k + \cdots = \sum_{x=0}^{\infty} \binom{n}{x} t^x$$

It is not difficult to show that this converges whenever $|t| < 1$ using the Ratio Test.

4. **Multinomial Theorem:** A generalization of the Binomial Theorem is

$$(t_1 + t_2 + \cdots + t_k)^n = \sum \frac{n!}{x_1!x_2!\cdots x_k!} t_1^{x_1} t_2^{x_2} \cdots t_k^{x_k}$$

where the summation is over all non-negative integers $x_1, x_2, \ldots, x_k$ such that $\sum\limits_{i=1}^{k} x_i = n$ where n is a positive integer.

Justification: Again we could verify this formula using a counting argument. Consider the product:

$$(t_1 + t_2 + \cdots + t_k)\,(t_1 + t_2 + \cdots + t_k) \cdots (t_1 + t_2 + \cdots + t_k)$$

To evaluate this product we must add together all of the possibilities obtained by taking one of the terms from the first bracketed expression, that is, one of $\{t_1, t_2, \ldots, t_k\}$, multiplying by one $\{t_1, t_2, \ldots, t_k\}$ taken from the second bracketed expression. etc. In how many ways do we obtain the term $t_1^{x_1} t_2^{x_2} \cdots t_k^{x_k}$ where $\sum\limits_{i=1}^{k} x_i = n$? We can choose t_1 a total of $x_1$times from any of the n terms in $\binom{n}{x_1}$ ways, and then t_2 from any of the remaining $n - x_1$ terms in $\binom{n-x_1}{x_2}$ ways, and so on so there are

$$\binom{n}{x_1}\binom{n-x_1}{x_2}\binom{n-x_1-x_2}{x_3}\cdots\binom{x_k}{x_k} = \frac{n!}{x_1!x_2!\cdots x_k!}$$

ways or obtaining this term in the product. The case $k = 2$ gives the Binomial Theorem in the form

$$(t_1 + t_2)^n = \sum_{x_1=0}^{n} \binom{n}{x_1} t_1^{x_1} t_2^{n-x_1}$$

5. **Hypergeometric Identity:**

$$\sum_{x=0}^{\infty}\binom{a}{x}\binom{b}{n-x}=\binom{a+b}{n}$$

There will not be an infinite number of terms if a and b are positive integers since the terms become 0 eventually. For example

$$\binom{4}{5}=\frac{4^{(5)}}{5!}=\frac{(4)(3)(2)(1)(0)}{5!}=0$$

Proof: We prove this in the case that a and b are non-negative integers. Obviously

$$(1+y)^{a+b}=(1+y)^a(1+y)^b$$

If we expand each term using the Binomial Theorem we obtain

$$\sum_{k=0}^{a+b}\binom{a+b}{k}y^k=\sum_{i=0}^{a}\binom{a}{i}y^i\times\sum_{j=0}^{b}\binom{b}{j}y^j$$

Note that the coefficient of y^k on the right side is $\sum\limits_{i=0}^{a}\binom{a}{i}\binom{b}{k-i}$ and so this must equal $\binom{a+b}{k}$, the coefficient of y^k on the left side.

6. **Exponential Series:** This is another example of a Maclaurin series expansion, if we let $f(x)=e^x$, then $f^{(k)}(0)=1$ for $k=1,2,\ldots$ and so

$$e^t=\frac{t^0}{0!}+\frac{t^1}{1!}+\frac{t^2}{2!}+\frac{t^3}{3!}+\cdots=\sum_{n=0}^{\infty}\frac{t^n}{n!}\quad\text{for all } t\in\Re$$

We will also use the limit definition of the exponential function:

$$e^t=\lim_{n\to\infty}\left(1+\frac{t}{n}\right)^n\quad\text{for all } t\in\Re$$

7. **Special series involving integers:**

$$\begin{aligned}1+2+3+\cdots+n&=\frac{n(n+1)}{2}\\1^2+2^2+3^2+\cdots+n^2&=\frac{n(n+1)(2n+1)}{6}\\1^3+2^3+3^3+\cdots+n^3&=\left[\frac{n(n+1)}{2}\right]^2\end{aligned}$$

Example: Find

$$\sum_{x=0}^{\infty} x\,(x-1)\binom{a}{x}\binom{b}{n-x}$$

Solution: For $x=0$ or 1 the term becomes 0, so we can start summing at $x=2$. For $x \geq 2$, we can expand $x!$ as $x(x-1)(x-2)!$

$$\sum_{x=0}^{\infty} x(x-1)\binom{a}{x}\binom{b}{n-x} = \sum_{x=2}^{\infty} x(x-1)\frac{a!}{x(x-1)(x-2)!(a-x)!}\binom{b}{n-x}$$

Cancel the $x(x-1)$ terms and try to re-group the factorial terms as “something choose something”.

$$\frac{a!}{(x-2)!(a-x)!} = \frac{a(a-1)(a-2)!}{(x-2)!\left[(a-2)-(x-2)\right]!} = a(a-1)\binom{a-2}{x-2}$$

Then

$$\sum_{x=0}^{\infty} x(x-1)\binom{a}{x}\binom{b}{n-x} = \sum_{x=2}^{\infty} a(a-1)\binom{a-2}{x-2}\binom{b}{n-x}$$

Factor out $a(a-1)$ and let $y = x-2$ to get

$$a(a-1)\sum_{y=0}^{\infty}\binom{a-2}{y}\binom{b}{n-(y+2)} = a(a-1)\binom{a+b-2}{n-2}$$

by the Hypergeometric Identity.

Problems

3.6.1 Use the Binomial Theorem to show that

$$\sum_{x=0}^{n} x\binom{n}{x}p^x(1-p)^{n-x} = np$$

3.6.2 Show that

$$\sum_{x=2}^{\infty} x\,(x-1)\,t^{x-2} = \frac{2}{(1-t)^3} \quad \text{for } |t|<1$$

3.6.3 For k a non-negative real number and $0<p<1$ show that

$$\sum_{x=0}^{\infty}\binom{-k}{x}p^k(p-1)^x = 1$$

3.7 Chapter 3 Problems

1. Six digits from $2, 3, 4, \ldots, 8$ are chosen and arranged in a row without replacement. Find the probability that

   (a) the number is divisible by 2

   (b) the digits 2 and 3 appear consecutively in the proper order (that is, in the order 23)

   (c) digits 2 and 3 appear in the proper order but not consecutively.

2. There are 6 stops on a subway line and 4 passengers on a subway car. Assume the passengers are each equally likely to get off at any stop. Find the probability that

   (a) the passengers all get off at different stops

   (b) 2 passengers get off at stop two and 2 passengers get off at stop five

   (c) 2 passengers get off at one stop and the other 2 passengers get off at another same stop

3. Five tourists plan to attend Octoberfest. Each tourist attends a location selected at random from the choices: Alpine Club, Bingemans, Concordia Club, Kitchener Memorial Auditorium, Queensmount, Schwaben Club, Transylvania Club (7 locations). Find the probability that

   (a) the tourists all attend different locations

   (b) the tourists all attend the same location

   (c) two tourists attend one location and three tourists attend another same location

   (d) at least one of the tourists attends Queensmount

   (e) both Queensmount and Concordia are unattended by the tourists

4. Suppose k people get on an elevator at the basement floor. There are n floors above the basement floor which are numbered $1, 2, 3, \ldots, n$ where people may get off.

   (a) Find the probability

      (i) nobody gets off at floor 1

      (ii) the people all get off at different floors $(n \geq k)$.

   (b) What assumption(s) underlies your answer to (a)? Comment briefly on how likely it is that the assumption(s) is(are) valid.

5. Give an expression for the probability a bridge hand of 13 cards contains 2 aces, 4 face cards (Jack, Queen or King) and 7 others.

6. The letters of the word STATISTICS are arranged in a random order to form a "word". Find the probability that

   (a) the word is STATISTICS

   (b) the word begins and ends with an I

   (c) the word begins and ends with the same letter

   (d) the T's occur together in the word

   (e) the word begins with an S and the T's occur together

7. Three digits are chosen in order from $0, 1, 2, \ldots, 9$. Find the probability the digits are drawn in increasing order (first digit $<$ second digit $<$ third digit) if

   (a) draws are made without replacement

   (b) draws are made with replacement.

8. **The Birthday Problem:**[5] Suppose there are n people in a room. Ignore birthdays that fall on February 29 and assume that every person is equally likely to have been born on any of the other 365 days in a year. (Is this a reasonable assumption? There are several images available on the internet illustrating the frequency of birthdays throughout the year.)

   (a) Find the probability that every person in the room has a different birthday.

   (b) Let $p(n)$ be the probability that at least two people in a room containing n people have the same birthday. Plot $p(n)$ for $n = 1, 2, \ldots, 80$.

   (c) For what value of n does $p(n)$ exceed 0.5? This surprising result is called the *Birthday Paradox*.

9. You have n identical looking keys on a chain, and one opens your office door. Suppose you try the keys in random order.

   (a) What is the probability the k'th key opens the door?

   (b) What is the probability one of the first two keys opens the door (assume $n \geq 3$)?

   (c) Determine numerical values for the answer in part (b) for $n = 3, 5, 7$.

[5] "My birthday was a natural disaster, a shower of paper full of flattery under which one almost drowned" Albert Einstein, 1954 on his seventy-fifth birthday.

10.

(a) Suppose a set of nine tickets are numbered $1, 2, \ldots, 9$. Three tickets are selected at random without replacement. Find the probability that the numbers of the tickets form an arithmetic progression. The *order* in which the tickets are selected does *not* matter.

(b) Suppose a set of $2n+1$ tickets are numbered $1, 2, \ldots, 2n+1$. Three tickets are selected at random without replacement. Find the probability that the numbers of the tickets form an arithmetic progression.

11. The 10,000 tickets for a lottery are numbered 0000 to 9999. A four-digit winning number is drawn and a prize is paid on each ticket whose four-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. A ticket costs $1 and each prize is $500.

(a) What is the probability of winning a prize

(i) with ticket number 7337?

(ii) with ticket number 7235?

(b) Based on your calculations in (a), what advice would you give to someone buying a ticket for this lottery?

(c) Assuming that all tickets are sold, what is the probability that the operator will lose money on the lottery?

12. **Capture/Recapture:**

(a) There are 25 deer in a certain forested area, and 6 have been caught temporarily and tagged. Some time later, 5 deer are caught. Find the probability that 2 of them are tagged. (What assumption did you make to do this?)

(b) Suppose that the total number of deer in the area was unknown to you. Describe how you could estimate the number of deer based on the information that 6 deer were tagged earlier, and later when 5 deer are caught, 2 are found to be tagged. What estimate do you get?

13. **Lotto 6/49:** In Lotto 6/49 you purchase a lottery ticket with 6 different numbers, selected from the set $\{1, 2, \ldots, 49\}$. In the draw, six (different) numbers are randomly selected. Find the probability that

(a) your ticket matches exactly 3 of the 6 numbers drawn.

(b) your ticket matches none of the numbers drawn.

(c) your ticket matches exactly x of the 6 numbers drawn, $x = 0, 1, \ldots, 6$.

14. The Enigma machine was used in World War II to send encrypted messages. There were five rotors to choose from. Three rotors were chosen and placed in order in the machine, and each was set to one of 26 letters (A-Z) to create the starting position. (There were also other complications such as a plugboard which swapped ten pairs of letters, but we will ignore this here.) Suppose you are a cryptanalyst trying to break Enigma.

    (a) How many possible starting positions are there?

    (b) Now suppose you know which three rotors are being used (and the order), but you don't know the letters on each one that form the setting. Find the probability that:

    (i) the setting is DKS

    (ii) the setting contains 2 vowels

    (iii) the setting begins and ends with a consonant

    (iv) the setting contains at least one A

    (c) Repeat the calculations in (b) if no two rotors can have the same letter.

15. **Hash Tables:** In computer science, a dictionary is a collection of key-value pairs (k, v), such that each pair consists of a unique key k and some data v. For example, for a collection of student records the key k would be the student ID number and v would be the data associated with that student. Suppose a data structure is to be designed to maintain such a collection. Let U = the set of possible keys, N = the number of elements in U, and n = number of keys used in the dictionary. If $n << N$ then a data structure of size N is very wasteful in terms of space. In such a situation, a hash table of size $M < N$ can be used together with a hash function $h : U \to \{0, 1, \ldots, M-1\}$. The pair (k, v) is stored in slot $h(k)$ of the table. Ideally, a hash function should have the property that each key is equally likely to be mapped to any of the M slots in the table independently of any other key.

    (a) Consider the hash function which chooses the slot for key k by randomly selecting a number with replacement from the set $\{0, 1, \ldots, M-1\}$. When a key-value pair is mapped to a slot which has already been assigned to another key-value pair a collision is said to have occurred. Show that, for n keys and a hash table of size M, the probability of at least one collision is equal to

$$1 - \left(1 - \frac{1}{M}\right)\left(1 - \frac{2}{M}\right)\cdots\left(1 - \frac{n-1}{M}\right)$$

    Hint: See Problem 8.

(b) Show that the probability in (a) is approximately equal to

$$1 - e^{-\frac{1}{2}n(n-1)/M}$$

Hint: $e^{-x} \approx 1 - x$ for x close to zero.

(c) Use the approximation in (b) to show that if you want the probability of a least one collision to be at most 0.5 then n should be less than $\sqrt{2M \log 2}$.

(d) Use the result in (d) to show that if there are $n = 2^{L/2}$ distinct keys, a hash table of size at least $M = 2^L$ should be used to ensure that the probability of a collision is less than 0.5.

16. **Texas Hold-em:** Texas Hold-em is a poker game in which players are each dealt two cards face down (called your hole or pocket cards), from a standard deck of 52 cards, followed by a round of betting, and then five cards are dealt face up on the table with various breaks to permit players to bet the farm. These are communal cards that anyone can use in combination with their two pocket cards to form a poker hand. Players can use any five of the face-up cards and their two cards to form a five card poker hand. Probability calculations for this game are not only required at the end, but also at intermediate steps and are quite complicated so that usually simulation is used to determine the odds that you will win given your current information, so consider a simple example. Suppose I was dealt two Jacks in the first round.

    (a) What is the probability that the next three cards (face up) include at least one Jack?

    (b) Given that there was no Jack among these next three cards, what is the probability that there is at least one among the last two cards dealt face-up?

    (c) What is the probability that the five face-up cards show two Jacks, given that I have two in my pocket cards?

17. I have a quarter which turns up heads with probability 0.6, and a fair dime. The quarter is flipped until a head occurs. Independently the dime is flipped until a head occurs. Find the probability that the number of flips is the same for both coins.

18. Players A and B decide to play chess until one of them wins. Assume games are independent with $P(A \text{ wins}) = 0.3$, $P(B \text{ wins}) = 0.25$ and $P(\text{draw}) = 0.45$ on each game. If the game ends in a draw another game will be played. Find the probability A wins before B.

4. PROBABILITY RULES AND CONDITIONAL PROBABILITY

4.1 General Methods

Recall that a probability model consists of a sample space S, a set of events or subsets of the sample space to which we can assign probabilities and a mechanism for assigning these probabilities. The probability of an arbitrary event A can be determined by summing the probabilities of simple events in A and so we have the following rules:

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

Proof: $P(S) = \sum\limits_{a \in S} P(a) = \sum\limits_{\text{all } a} P(a) = 1$.

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

Proof: $P(A) = \sum\limits_{a \in A} P(a) \leq \sum\limits_{a \in S} P(a) = 1$ and since each $P(a) \geq 0$, we have $0 \leq P(A) \leq 1$.

Rule 3: If A and B are two events with $A \subseteq B$ (that is, all of the points in A are also in B), then $P(A) \leq P(B)$.

Proof: $P(A) = \sum\limits_{a \in A} P(a) \leq \sum\limits_{a \in B} P(a) = P(B)$ so $P(A) \leq P(B)$.

Before continuing with the set-theoretic description of a probability model, let us review some of the basic ideas in set theory. First what do sets have to do with the occurrence of events? Suppose a random experiment having sample space S is conducted (for example a die is thrown with $S = \{1, 2, 3, 4, 5, 6\}$). When would we say an event $A \subset S$, or in the case of the die, the event $A = \{2, 4, 6\}$ occurs? In the latter case, the event A means that the number showing is even, that is, in general that *one of the simple outcomes in A occurred.*

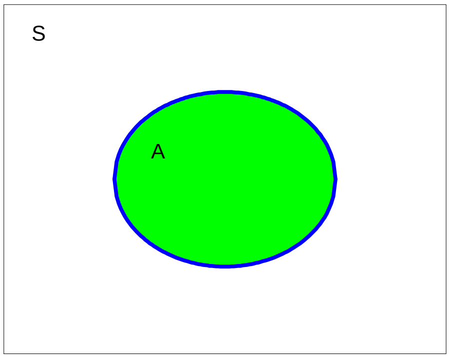


Figure 4.1: Event A in sample space S

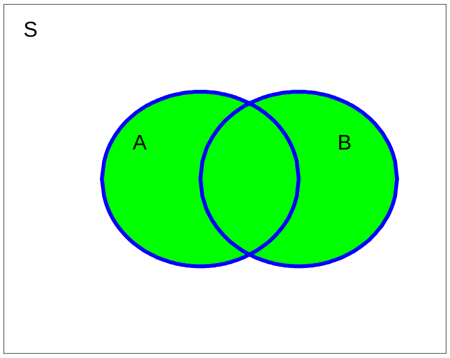


Figure 4.2: The union of two events $A \cup B$

We often illustrate the relationship among sets using *Venn diagrams*. In the drawings below, think of S consisting of all of the points in a rectangle of area one[6]. To illustrate the event A we can draw a region within the rectangle with area roughly proportional to the probability of the event A. We might think of the random experiment as throwing a dart at the rectangle in Figure 4.1, and we say the event A occurs if the dart lands within the region A.

What if we combine two events A, B by including all of the points in either A or B or both. This is the union of the two events or $A \cup B$ illustrated in Figure 4.2. The union of the events occurs if one of the outcomes in either A or B or both occurs. In language we refer to this as the event "A or B" with

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

the understanding that in this course we will use the word “or” inclusively to also permit both. Another way of expressing a union is $A \cup B$ occurs if at least one of A, B occurs. Similarly if we have three events A, B, C, the event $A \cup B \cup C$ means “at least one of A, B, C”.

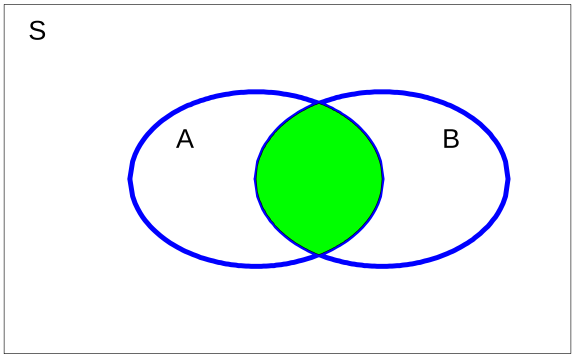


Figure 4.3: The intersection of two events $A \cap B$

What about the intersection of two events $(A \cap B)$ or the set of all points in S that are in both A **and** B? This is illustrated in Figure 4.3. The event $A \cap B$ occurs if and only if a point in the intersection occurs which means **both A and B occur**.

Note: The sets $A \cap B$ and $A \cap B \cap C$ are often written more simply as AB and ABC respectively.

Finally the complement of the event A, denoted by $\bar{A}$, is the set of all points which are in S but **not in** A as in Figure 4.4.

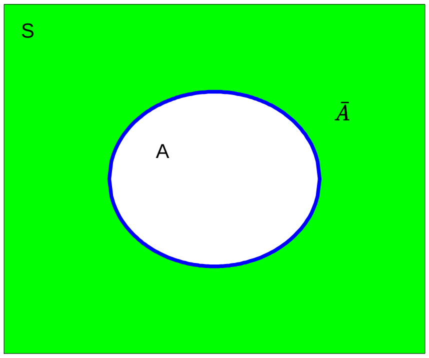


Figure 4.4: $\bar{A}$ = the complement of the event A

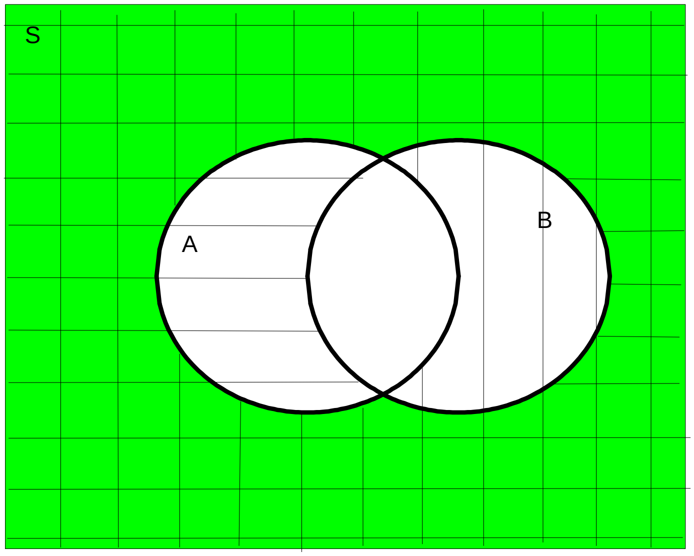


Figure 4.5: Illustration of De Morgan's Law using a Venn Diagram

There are two special events in a probability model that we will use. One is the whole sample space S. Because $P(S) = 1$, this event is certain to occur. Another is the *empty event*, or the *null set* $\emptyset$. This is a set with no elements at all and so it must have probability 0. Notice that $\emptyset = \overline{S}$.

The illustrations above showing the relationship among sets are examples of Venn diagrams. Since probability theory is built from the relationships among sets, it is often helpful to use Venn diagrams in solving problems. For example there are rules governing taking the complements of unions and intersections that can easily be verified using Venn diagrams.

De Morgan's Laws:

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

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

Proof of (a): One can argue such set theoretic rules using the definitions of the sets. For example when is a point a in the set $\overline{A \cup B}$. This means $a \in S$ but a is not in $A \cup B$, which in turn implies a is not in A **and** it is not in B, or $a \in \overline{A}$ **and** $a \in \overline{B}$, equivalently $a \in \overline{A} \cap \overline{B}$. As and alternative demonstration, we can use a Venn diagram (Figure 4.5) in which $\overline{A}$ is indicated with vertical lines, $\overline{B}$ with horizontal lines and so $\overline{A} \cap \overline{B}$ is the region with both vertical and horizontal lines. This agrees with the shaded region $\overline{A \cup B}$.

The following example demonstrates solving a problem using a Venn diagram.

Example: Suppose for students finishing second year Math 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 greater than 80%, 13% have both an overall average and STAT 230 greater than 80%, 10% have all 3 of these averages greater than 80%, and 67% have none of these 3 averages greater than 80%. Find the probability a randomly chosen math student finishing 2A has math and overall averages both greater than 80% and STAT 230 less than or equal to 80%.

Solution: When using rules of probability it is generally helpful to begin by labeling the events of interest. Imagine a student is chosen at random from all students finishing second year Math. For this student, let

A	be the event	"math average greater than 80%"
B	be the event	"overall average greater than 80%"
C	be the event	"STAT 230 mark greater than 80%"

In terms of these symbols, we are given:

$$\begin{array}{ll} P(A) = 0.22 & P(B) = 0.20 \\ P(A \cap C) = 0.14 & P(B \cap C) = 0.13 \\ P(A \cap B \cap C) = 0.1 & \\ P(\overline{A} \cap \overline{B} \cap \overline{C}) = 0.67 & \end{array}$$

Let us interpret some of these expressions; for example $\overline{A} \cap \overline{B} \cap \overline{C}$ or (not A) **and** (not B) **and** (not C), means that none of the marks or averages are greater than 80% for the randomly chosen student. We are asked to find $P(A \cap B \cap \overline{C})$, the region labelled with '(5) x' in Figure 4.6. We have filled in the following information on the Venn diagram, in the order indicated by (1), (2), (3), etc.

(1) $P(A \cap B \cap C) = 0.1$ is given
(2) $P(A \cap C) - P(A \cap B \cap C) = 0.14 - 0.1 = 0.04$
(3) $P(B \cap C) - P(A \cap B \cap C) = 0.13 - 0.1 = 0.03$
(4) $P(C) - P(A \cap C) - 0.03 = 0.24 - 0.14 - 0.03 = 0.07$
(5) $P(A \cap B \cap \overline{C})$ is unknown, so let $P(A \cap B \cap \overline{C}) = x$
(6) $P(A) - P(A \cap C) - P(A \cap B \cap \overline{C}) = 0.22 - 0.14 - x = 0.08 - x$
(7) $P(B) - P(B \cap C) - P(A \cap B \cap \overline{C}) = 0.20 - 0.13 - x = 0.07 - x$
(8) $P(\overline{A \cup B \cup C}) = 0.67$ is given.

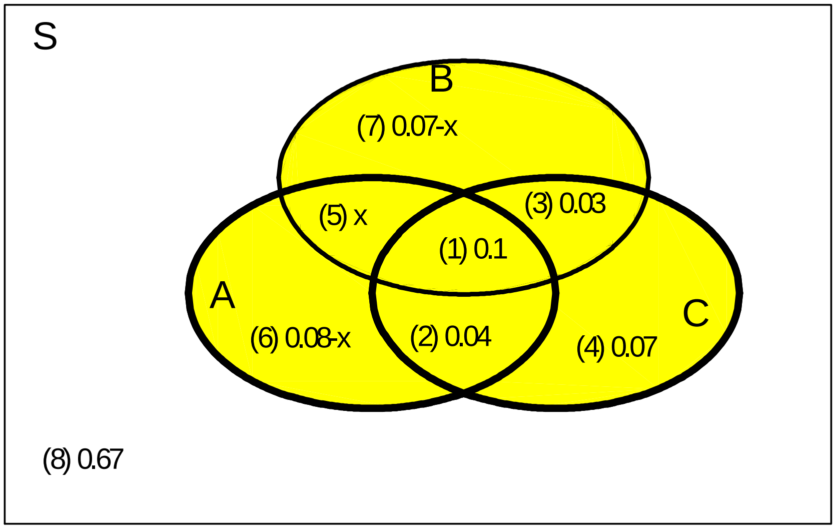


Figure 4.6: Venn Diagram for Math Averages Example

Adding all probabilities from (1) to (8) we obtain, since $P(S) = 1$,

$$0.1 + 0.04 + 0.03 + 0.07 + x + 0.08 - x + 0.07 - x + 0.67 = 1$$

giving $1.06 - x = 1$ and solving for x, $P(A \cap B \cap \overline{C}) = x = 0.06$.

Problems

4.1.1 At a large company 40% of the employees do not have black hair, 75% of the employees have brown eyes, and 55% of the employees have black hair and brown eyes. If an employee is chosen at random what is the probability the employee does not have black hair and brown eyes?

4.1.2 In a class of 500 students who speak English, 50% also speak French, 25% also speak Spanish and 40% also speak Mandarin or Cantonese. 10% speak both French and Spanish, 12% speak both French and Mandarin or Cantonese and 2% speak French, Spanish and Mandarin or Cantonese. 8% of the students only speak English. Find the probability that a randomly chosen student from this class speaks Spanish and Mandarin or Cantonese but not French.

4.2 Rules for Unions of Events

In addition to the two rules which govern probabilities listed in Section 4.1, we have the following

Rule 4 a: Addition Law of Probability or the Sum Rule

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

Proof: Suppose we denote set differences by $A/B = A \cap \overline{B} =$ the set of points which are in A but not in B. Then

$$\begin{aligned} P(A) + P(B) &= \sum_{a \in A} P(a) + \sum_{a \in B} P(a) \\ &= \left(\sum_{a \in A/B} P(a) + \sum_{a \in AB} P(a) \right) + \left(\sum_{a \in B/A} P(a) + \sum_{a \in AB} P(a) \right) \\ &= \left(\sum_{a \in A/B} P(a) + \sum_{a \in AB} P(a) + \sum_{a \in B/A} P(a) \right) + \sum_{a \in AB} P(a) \\ &= \sum_{a \in A \cup B} P(a) + \sum_{a \in AB} P(a) \\ &= P(A \cup B) + P(A \cap B) \end{aligned}$$

Rearranging $P(A)+P(B) = P(A \cup B)+P(A \cap B)$ we obtain $P(A \cup B) = P(A)+P(B)-P(A \cap B)$ as required. This can also be justified by using a Venn diagram. Each point in $A \cup B$ must be counted once. In the expression $P(A) + P(B)$, however, points in $A \cap B$ have their probability counted twice, once in $P(A)$ and once in $P(B)$, so they need to be subtracted once.

Rule 4 b: Probability of the Union of Three Events

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC) \qquad (4.1)$$

Proof: See Figure 4.7. In the sum $P(A) + P(B) + P(C)$ those points in the regions labelled D, H, J in Figure 4.7 lie in only one of the events and their probabilities are added only once. However points in the regions labelled G, E, I, for example, lie in two of the events. We can compensate for this double counting by subtracting these probabilities once, e.g. using $P(A) + P(B) + P(C) - [P(AB) + P(AC) + P(BC)]$. However, now those points in all three sets, that is, those points in $F = ABC$ have their probabilities added in three times and then subtracted three times so they are not included at all. We must correct the formula to give (4.1).

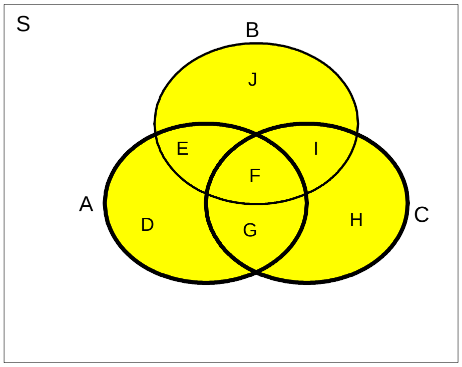


Figure 4.7: The union of three events $A \cup B \cup C$

Rule 4 c: Probability of the Union of n Events:

There is an obvious generalization of the above formula to n events $A_1, A_2, \ldots, A_n$. This is often referred to as the *inclusion-exclusion principle* because of the process discussed above for constructing it:

$$P(A_1 \cup A_2 \cup A_3 \cup \cdots \cup A_n) = \sum_i P(A_i) - \sum_{i<j} P(A_iA_j) + \sum_{i<j<k} P(A_iA_jA_k) \qquad (4.2)$$
$$- \sum_{i<j<k<l} P(A_iA_jA_kA_l) + \cdots$$

(where the subscripts are all distinct, for example $i < j < k < l$).

Proof: This can be proved using rule 4a and induction. Let $B_n = A_1 \cup A_2 \cup A_3 \cup \cdots \cup A_n$ for $n = 1, 2, \ldots$. Then 4a shows that (4.2) holds for $n = 2$. Suppose the rule is true for n. Then

$$\begin{aligned}
&P(A_1 \cup A_2 \cup A_3 \cup \cdots \cup A_n \cup A_{n+1}) \\
&= P(B_n \cup A_{n+1}) \\
&= P(B_n) + P(A_{n+1}) - P(B_nA_{n+1}) \\
&= \sum_{i \le n} P(A_i) - \sum_{i<j\le n} P(A_iA_j) + \sum_{i<j<k\le n} P(A_iA_jA_k) + \ldots + P(A_{n+1}) \\
&- \sum_{i \le n} P(A_iA_{n+1}) + \sum_{i<j\le n} P(A_iA_jA_{n+1}) - \sum_{i<j<k\le n} P(A_iA_jA_kA_{n+1}) + \ldots
\end{aligned}$$

We will use (4.2) rarely in this course[7].

[7] i.e. do not memorize

Definition 6 *Events A and B are **mutually exclusive** if $A \cap B = \emptyset$ (the empty event).*

Since mutually exclusive events A and B have no common points, $P(A \cap B) = P(\emptyset) = 0$.

In general, events $A_1, A_2, \ldots, A_n$ are mutually exclusive if $A_i \cap A_j = \emptyset$ for all $i \neq j$. This means that there is no chance of two or more of these events occurring together, we either have exactly one of the events occur, or none. For example, if a die is rolled twice, the events $A =$ "2 occurs on the 1st roll" and $B =$ "total is 10" are mutually exclusive events. Similarly the events $A_2, A_3, \ldots, A_{12}$ where A_j is the event that the total on the two dice is j are all mutually exclusive events. In the case of mutually exclusive events, Rule 4 above simplifies to Rule 5 below.

Rule 5 a: Probability of the Union of Two Mutually Exclusive Events

Let A and B be mutually exclusive events. Then

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

This is a consequence of Rule 4a and the fact that $P(A \cap B) = P(\emptyset) = 0$.

Rule 5 b: Probability of the Union of n Mutually Exclusive Events

In general, let $A_1, A_2, \ldots, A_n$ be mutually exclusive events. Then

$$P\left(A_1 \cup A_2 \cup \cdots \cup A_n\right) = \sum_{i=1}^{n} P(A_i)$$

This is easily proven from Rule 5a above using induction or as an immediate consequence of Rule 4c.

Rule 6: Probability of the Complement of an Event

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

Proof: A and $\bar{A}$ are mutually exclusive and $A \cup \bar{A} = S$, so by Rule 5a,

$$P(A \cup \overline{A}) = P(A) + P(\overline{A})$$

But since $P(A \cup \bar{A}) = P(S) = 1$,

$$1 = P(A) + P(\overline{A}) \text{ or}$$
$$P(A) = 1 - P(\overline{A})$$

This result is useful whenever $P(\overline{A})$ is easier to obtain than $P(A)$.

Example: Two ordinary dice are rolled. Find the probability that at least one of them turns up a six.

Solution 1: The sample space is $S = \{(1,1),(1,2),(1,3),\ldots,(6,6)\}$. Let A be the event that we obtain 6 on the first die, B be the event that we obtain 6 on the second die and note (by rule $4a$) that

$$\begin{aligned} P(\text{at least one die shows } 6) &= P(A \cup B) \\ &= P(A) + P(B) - P(A \cap B) \\ &= \frac{1}{6} + \frac{1}{6} - \frac{1}{36} = \frac{11}{36} \end{aligned}$$

Solution 2: This is an example where it is perhaps somewhat easier to obtain the complement of the event $A \cup B$ since the complement is the event that there is no six showing on either die, and there are exactly 25 such points, $\{(1,1),\ldots,(1,5),(2,1),\ldots,(2,5),\ldots,(5,5)\}$. Therefore

$$\begin{aligned} P(\text{at least one die shows } 6) &= 1 - P(\text{no 6 on either die}) \\ &= 1 - \frac{25}{36} = \frac{11}{36} \end{aligned}$$

Example: Roll a die 3 times. Find the probability of getting at least one 6.

Solution 1: Let A be the event "least one die shows 6". Then $\bar{A}$ is the event that no 6 on any die shows. Using counting arguments, there are 6 outcomes on each roll, so $S = \{(1,1,1),(1,1,2),\ldots,(6,6,6)\}$ has $6 \times 6 \times 6 = 216$ points. For $\bar{A}$ to occur we cannot have a 6 on any roll. Then $\bar{A}$ can occur in $5 \times 5 \times 5 = 125$ ways. Therefore

$$P(\bar{A}) = \frac{125}{216}$$

and

$$P(A) = 1 - \frac{125}{216} = \frac{91}{216}$$

Solution 2: Can you spot the flaw in the following argument? Let

A be the event that 6 occurs on the first roll

B be the event that 6 occurs on the second roll

C be the event that 6 occurs on the third roll

Then

$$\begin{aligned} P(\text{one or more sixes}) &= P(A \cup B \cup C) \\ &= P(A) + P(B) + P(C) \\ &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2} \end{aligned}$$

You should have noticed that A, B, and C are **not** mutually exclusive events, so we should have used

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC)$$

Each of AB, AC, and BC occurs 6 times in the sample space of 216 points and so

$$P(AB) = \frac{1}{36} = P(BC) = P(AC)$$

Also

$$P(ABC) = \frac{1}{216}$$

Therefore

$$\begin{aligned} P(A \cup B \cup C) &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} - \frac{1}{36} - \frac{1}{36} - \frac{1}{36} + \frac{1}{216} \\ &= \frac{91}{216} \end{aligned}$$

Note: These rules link the concepts of addition of probabilities with unions of events, and complements. The next segment will consider intersection, multiplication of probabilities, and a concept known as independence. Making these linkages will make problem solving and the construction of probability models easier.

Problems

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

(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?

4.2.2 Prove that $P(A \cup B) = 1 - P(\overline{A} \cap \overline{B})$ for arbitrary events A and B in S.

4.3 Intersections of Events and Independence

Dependent and Independent Events:

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 Canada Goose is sucked into one jet engine, then this event would affect the probability that the airplane safely reaches its destination, that is, it affects the probability that should be assigned to the event B.

Suppose we toss a fair coin twice. Consider the events A : head on first toss and B : head on both tosses. Again there appears to be some dependence. On the other hand, if we define the event B as B : head on second toss, we do not think that the occurrence of A affects the chances that B will occur.

If we need to reassess the probability of an event B if we are given that the event A 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 7 *Events A and B are* ***independent events*** *if and only if $P(A \cap B) = P(A)P(B)$. If the events are not independent, we call the events* ***dependent****.*

When we use Venn diagrams, we 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 the same properties. Let us continue this tradition, so that if two events are independent, then the “size” of their intersection, as represented by the area in a Venn diagram, must equal the product of the 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 4.8, the area of region A is equal to $(0.4)\,(0.5) = 0.2$, the area of region B is equal to $(0.6)\,(0.5) = 0.3$ and the area of region $A \cap B$ is equal to $(0.3)\,(0.2) = 0.06$. Since $P(A)P(B) = (0.2)\,(0.3) = 0.06 = P\,(A \cap B)$, the events A and B are independent events. If you were to hold the rectangle A in place and move the rectangle B down and to the right, the probability of the intersection as represented by the area would decrease and the events would become dependent.

Exercise: Suppose the events A and B are mutually exclusive events with $P\,(A) > 0$ and $P\,(B) > 0$. Can A and B be independent events?

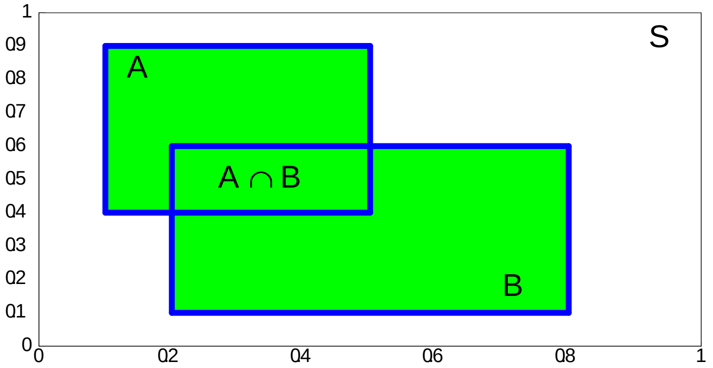


Figure 4.8: A Venn diagram illustrating independent events

Example: Suppose we toss a fair coin twice. Let $A =$ "head on 1st toss" and $B =$ "head on 2nd toss". Clearly A and B are independent events since the outcome on each toss is unrelated to other tosses, so $P(A) = \frac{1}{2}$, $P(B) = \frac{1}{2}$, and $P(A \cap B) = \frac{1}{4} = P(A)P(B)$.

Example: Suppose we roll a die once and let $A =$ "the number is even" and $B =$ "number > 3". The events will be dependent since

$$P(A) = \frac{1}{2}, \quad P(B) = \frac{1}{2} \text{ and}$$
$$P(A \cap B) = P(4 \text{ or } 6 \text{ occurs}) = \frac{2}{6} \neq P(A)P(B)$$

(Rationale: B only happens half the time. If A occurs we know the number is 2, 4, or 6. So B occurs $\frac{2}{3}$ of the time when A occurs. The occurrence of A does affect the chances of B occurring so A and B are **not** independent.)

When there are more than two events, the above definition generalizes to[8]:

Definition 8 *The events $A_1, A_2, \ldots, A_n$ are mutually independent if and only if*

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

for all sets $(i_1, i_2, \ldots, i_k)$ of distinct subscripts chosen from $(1, 2, \ldots, n)$.

[8] We need all subsets so that events are independent of combinations of other events. For example if A_1 is independent of A_2 and A_4 is to be independent of A_1A_2 then, $P(A_1A_2A_4) = P(A_1A_2)P(A_4) = P(A_1)P(A_2)P(A_4)$

For example, for $n = 3$, we need

$$\begin{aligned} P(A_1 \cap A_2) &= P(A_1)P(A_2) \\ P(A_1 \cap A_3) &= P(A_1)P(A_3) \\ P(A_2 \cap A_3) &= P(A_2)P(A_3) \end{aligned}$$

and

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3)$$

We will shorten “mutually independent” to “independent” to reduce confusion with “mutually exclusive.”

Remark: The definition of independence works two ways. If we can find $P(A)$, $P(B)$, and $P(A \cap B)$ then we can determine whether A and B are independent. Conversely, if we know (or assume) that A and B are independent, then we can use the definition as a rule of probability to calculate $P(A \cap B)$. Examples of each follow.

Example: Toss a die twice. Let A be the event that the first toss is a 3 and B the event that the total is 7. Are A and B independent? (What do you think?) We use the definition to check. Now $P(A) = \frac{1}{6}$, $P(B) = \frac{6}{36}$ since $B = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$ and $P(A \cap B) = \frac{1}{36}$ since $A \cap B = \{(3,4)\}$. Therefore,

$$P(A \cap B) = \frac{1}{36} = P(A)P(B) = \left(\frac{1}{6}\right)\left(\frac{6}{36}\right)$$

and so A and B are independent events.

Now suppose we define C to be the event that the total is 8. This is a minor change from the definition of B. Then

$$P(A) = \frac{1}{6}, \;\; P(C) = \frac{5}{36} \;\text{ and }\; P(A \cap C) = \frac{1}{36}$$

Therefore $P(A \cap C) \neq P(A)P(C)$ and A and C are dependent events.

This example often puzzles students. Why are they independent if B is a total of 7 but dependent for C is a total of 8? The key is that regardless of the first toss, there is always one number on the 2nd toss which makes the total 7. Since the probability of getting a total of 7 started off being $\frac{6}{36} = \frac{1}{6}$, the outcome of the 1st toss doesn't affect the chances. However, for any total other than 7, the outcome of the 1st toss does affect the chances of getting that total (e.g., a first toss of 1 guarantees the total cannot be 8)[9].

[9] This argument is in terms of “conditional probability” closely related to independence and to be treated in the next section.

Example: Show if A and B are independent events, then $\bar{A}$ and B are independent events.

Solution: Since $A \cap B$ and $\bar{A} \cap B$ are mutually exclusive events $B = (A \cap B) \cup \left(\bar{A} \cap B\right)$ so

$$P(B) = P(A \cap B) + P\left(\bar{A} \cap B\right)$$

Therefore

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

Example: A pseudo random number generator on a computer can give a sequence of independent random digits chosen from $S = \{0, 1, \ldots, 9\}$. This means that (i) each digit has probability of $\frac{1}{10}$ of being any of $0, 1, \ldots, 9$, and (ii) events determined by the different trials are independent of one another. We call this 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) Define the events A_i**:** digit from trial i is odd, $i = 1, 2, \ldots, 5$. Then

$$P(\text{all digits are odd}) = P(A_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5) = \prod_{i=1}^{5} P(A_i)$$

since the A_i's are mutually independent. Since $P(A_i) = \frac{1}{2}$, we get $P(\text{all digits are odd}) = \frac{1}{2^5}$.

(b) Define events A_i: 9 occurs on trial i, for $i = 1, 2, \ldots$. Then we want

$$\begin{aligned} P(\overline{A}_1 \cap \overline{A}_2 \cap \cdots \cap \overline{A}_9 \cap A_{10}) &= P(\overline{A}_1)P(\overline{A}_2)\cdots P(\overline{A}_9)P(A_{10}) \\ &= (0.9)^9(0.1) \end{aligned}$$

because the A_i's are independent, and $P(A_i) = 1 - P(\overline{A}_i) = 0.1$.

Note: We implicitly assumed independence of events in some of our earlier probability calculations. For example, suppose a coin is tossed 3 times, and we consider the sample space

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

Assuming that the outcomes on the three tosses are independent, and $P(H) = P(T) = \frac{1}{2}$ on each toss, we obtain

$$P(HHH) = P(H)P(H)P(H) = \left(\frac{1}{2}\right)^3 = \frac{1}{8}$$

Similarly, all the other simple events have probability $\frac{1}{8}$. In earlier calculations we implicitly assumed this was true by assigning the same probability $\frac{1}{8}$ to all possible outcomes without thinking directly about independence. However, it is clear that if somehow the 3 tosses were not independent then it might be a bad idea to assume each outcome had 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; now consider 3 consecutive days. Would you want to assign a probability of $\frac{1}{8}$ to each of the 8 simple events even if this were in a season when the probability of rain on a day was $\frac{1}{2}$?)

Note: The definition of independent events can be used either to check for independence or, if events are known to be independent, to calculate $P(A \cap B)$. Many problems are not obvious, and 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.

Problems

4.3.1 A weighted die is such that $P(1) = P(2) = P(3) = 0.1$, $P(4) = P(5) = 0.2$, and $P(6) = 0.3$. Assume that events determined by different throws of the die are independent.

(a) If the die is thrown twice what is the probability the total is 9?

(b) If a die is thrown twice, and this process repeated 4 times, what is the probability the total will be 9 on exactly 1 of the 4 repetitions?

4.3.2 Sixty percent of undergraduate UWaterloo students are coop students and 20% are Math students. Twenty-five percent of coop students are Math students. A committee of 10 students is formed by randomly selecting from UWaterloo students. What is the probability there will be at least one coop students and at least one Math student on the committee[10]?

4.3.3 Prove that $\bar{A}$ and $\bar{B}$ are independent events if and only if $\bar{A}$ and B are independent.

[10] Although the sampling is conducted without replacement, because the population is very large, whether we replace or not will make little difference. Therefore assume in your calculations that sampling is *with replacement* so the 10 draws are independent.

4.4 Conditional Probability

In many situations we may want to determine the probability of some event A, while knowing that some other event B has already occurred. For example, what is the probability a randomly selected person is over 6 feet tall, given that she is female? Let the symbol $P(A|B)$ represent the probability that event A occurs, when we know that B occurs. We call this the conditional probability of A given B. While we will give a definition of $P(A|B)$, let's first consider an example we looked at earlier, to get some sense of why $P(A|B)$ is defined as it is.

Example: Suppose we roll a die once so that sample space is $S = \{1, 2, 3, 4, 5, 6\}$. Let A be the event that the number is even and B the event that the number is greater than 3. If we know that B occurs, this tells us that we have a 4, 5, or 6. Of the times when B occurs, we have an even number $\frac{2}{3}$ of the time. So $P(A|B) = \frac{2}{3}$. More formally, we could obtain this result by calculating $\frac{P(A\cap B)}{P(B)}$, since $P(A \cap B) = P(4 \text{ or } 6) = \frac{2}{6}$ and $P(B) = \frac{3}{6}$.

Definition 9 *The conditional probability of event A, given event B, is*

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \;\; \textit{provided} \;\; P(B) > 0$$

Note: If A and B are independent then

$$\begin{aligned} P(A \cap B) &= P(A)P(B) \\ \text{so } \; P(A|B) &= \frac{P(A)P(B)}{P(B)} = P(A) \;\; \text{provided} \;\; P(B) > 0 \end{aligned}$$

This result leads us to the following theorem:

Theorem 10 *Suppose A and B are two events defined on a sample space S such that $P(A) > 0$ and $P(B) > 0$. Then A and B are independent events if and only if either of the following statements is true*

$$P(A|B) = P(A) \quad \textit{or} \quad P(B|A) = P(B)$$

Note: We could have taken the definition of independent events to be: A and B are independent events if $P(A|B) = P(A)$. In some sense, this definition is more intuitive than the original definition. However this definition does not hold in the case that $P(B) = 0$ whereas the original definition does hold.

Example: If 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 is $S = \{HHH, HHT, HTH, \ldots, TTT\}$. Define the events $A =$ "one Head" and $B =$ "at least one Head". We need to find

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

Now

$$\begin{aligned} P(B) &= 1 - P\left(\bar{B}\right) = 1 - P(\text{no heads}) \\ &= \frac{7}{8} \end{aligned}$$

and

$$\begin{aligned} P(A \cap B) &= P(\text{we obtain one head AND we obtain at least one head}) \\ &= P(\text{we obtain one head}) \\ &= P(\{HTT, THT, TTH\}) \\ &= \frac{3}{8} \end{aligned}$$

using either the sample space with equally probably points, or the fact that the 3 tosses are independent. Thus,

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{\frac{3}{8}}{\frac{7}{8}} = \frac{3}{7} \end{aligned}$$

Example: 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 is the fraction that 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 $F = \overline{M}$ the event that the person selected is female. We are asked to find $P(C)$. We are given that $P(C|M) = 0.05$, $P(C|F) = 0.0025$, and $P(M) = 0.5 = P(F)$. From the definition of conditional probability

$$\begin{aligned} P(C|M)P(M) &= \frac{P(C \cap M)}{P(M)} P(M) \\ &= P(C \cap M) \end{aligned}$$

and similarly $P(C|F)P(F) = P(C \cap F)$. To obtain $P(C)$ we can therefore use the fact that $C = (C \cap M) \cup \left(C \cap \overline{M}\right)$ and the events $C \cap M$ and $C \cap \overline{M}$ are mutually exclusive so

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

4.5 Product Rules, Law of Total Probability and Bayes' Theorem

The preceding example suggests two more useful probability rules. They are based on the idea of breaking down the event of interest into mutually exclusive pieces.

Rule 7: Product Rules *Let $A, B, C, D, \ldots$ be arbitrary events in a sample space. Assume that* $P(A) > 0$, $P(A \cap B) > 0$, and $P(A \cap B \cap C) > 0$. *Then*

$$\begin{aligned} P(AB) &= P(A)P(B|A) \\ P(ABC) &= P(A)P(B|A)P(C|AB) \\ P(ABCD) &= P(A)P(B|A)P(C|AB)P(D|ABC) \end{aligned}$$

and so on.

Proof: The first rule comes directly from the definition $P(B|A)$ since

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

assuming $P(A) > 0$. The right hand side of the second rule equals (assuming $P(AB) > 0$ and $P(A) > 0$)

$$\begin{aligned} P(A)P(B|A)P(C|AB) &= P(A)\frac{P(AB)}{P(A)}P(C|AB) \\ &= P(AB)P(C|AB) \\ &= P(AB)\frac{P(CAB)}{P(AB)} \\ &= P(ABC) \end{aligned}$$

and so on.

In order to remember these rules you can imagine that the events unfold in some chronological order, even if they do not. For example,

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

could be interpreted as the probability that “A occurs” (first) and then “given A occurs, that B occurs” (next), etc.

Rule 8: Law of Total Probability *Let $A_1, A_2, \ldots, A_k$ be a partition of the sample space S into disjoint (mutually exclusive) events, that is*

$$A_1 \cup A_2 \cup \cdots \cup A_k = S \text{ and } A_i \cap A_j = \emptyset \text{ if } i \neq j$$

Let B be an arbitrary event in S. Then

$$\begin{aligned} P(B) &= P(BA_1) + P(BA_2) + \cdots + P(BA_k) \\ &= \sum_{i=1}^{k} P(B|A_i)P(A_i) \end{aligned}$$

Proof: Note that the events $BA_1, BA_2, \ldots, BA_k$ are all mutually exclusive and their union is B, that is $B = (BA_1) \cup \cdots \cup (BA_k)$. Therefore by Rule 5b

$$P(B) = P(BA_1) + P(BA_2) + \cdots + P(BA_k).$$

By the product rule, $P(BA_i) = P(B|A_i)P(A_i)$ so this becomes

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_k)P(A_k)$$

Example: In an insurance portfolio 10% of the policy holders are in Class A_1 (high risk), 40% are in Class A_2 (medium risk), and 50% are in Class A_3 (low risk). The probability there is a claim on a Class A_1 policy in a given year is 0.10; similar probabilities for Classes A_2 and A_3 are 0.05 and 0.02. Find the probability that if a claim is made, it is made on a Class A_1 policy.

Solution: For a randomly selected policy, let $B =$ “policy has a claim” and $A_i =$ “policy is of Class A_i”, $i = 1, 2, 3$. We are asked to find $P(A_1|B)$. Note that

$$P(A_1|B) = \frac{P(A_1 \cap B)}{P(B)}$$

and that

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B)$$

We are given that

$$P(A_1) = 0.10, \ P(A_2) = 0.40, \ P(A_3) = 0.50$$

and

$$P(B|A_1) = 0.10,\ P(B|A_2) = 0.05,\ P(B|A_3) = 0.02$$

Therefore

$$P(A_1 \cap B) = P(A_1)P(B|A_1) = 0.01$$
$$P(A_2 \cap B) = P(A_2)P(B|A_2) = 0.02$$
$$P(A_3 \cap B) = P(A_3)P(B|A_3) = 0.01$$

This gives $P(B) = 0.01 + 0.02 + 0.01 = 0.04$ and

$$P(A_1|B) = \frac{P(A_1 \cap B)}{P(B)} = \frac{0.01}{0.04} = 0.25$$

Tree Diagrams

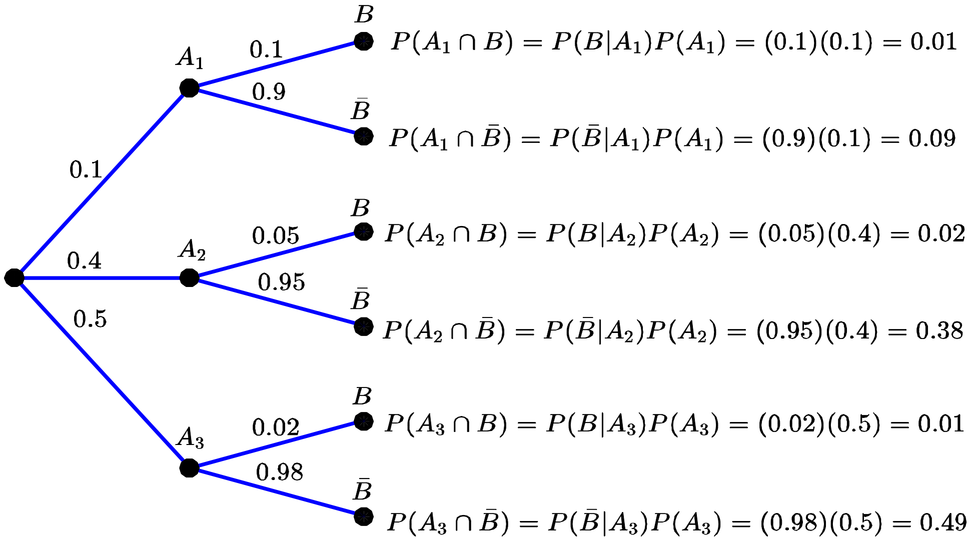

B
A_1
0.1
$P(A_1 \cap B) = P(B|A_1)P(A_1) = (0.1)(0.1) = 0.01$
0.9
$\bar{B}$
$P(A_1 \cap \bar{B}) = P(\bar{B}|A_1)P(A_1) = (0.9)(0.1) = 0.09$
0.1
B
$P(A_2 \cap B) = P(B|A_2)P(A_2) = (0.05)(0.4) = 0.02$
0.05
0.4
A_2
0.95
$\bar{B}$
$P(A_2 \cap \bar{B}) = P(\bar{B}|A_2)P(A_2) = (0.95)(0.4) = 0.38$
0.5
B
$P(A_3 \cap B) = P(B|A_3)P(A_3) = (0.02)(0.5) = 0.01$
A_3
0.02
0.98
$\bar{B}$
$P(A_3 \cap \bar{B}) = P(\bar{B}|A_3)P(A_3) = (0.98)(0.5) = 0.49$

Figure 4.9: Tree diagram for insurance example

Tree diagrams can be a useful device for keeping track of conditional probabilities when using multiplication and partition rules. 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 branches leading to it. The probability at any node of the tree is obtained by multiplying the probabilities on the branches leading to the node, and equals the probability of the intersection of the events leading to it. For example, the immediately preceding example could be represented by the tree in Figure 4.9. Note that the probabilities on the terminal nodes must add up to 1.

Here is another example involving diagnostic tests for disease. See if you can represent the problem using a tree diagram.

Example: Testing for HIV

Tests used to diagnose medical conditions are often imperfect, and give false positive or false negative results, as described in Problem 2.6 of Chapter 2. A fairly cheap blood test for the Human Immunodeficiency Virus (HIV) 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 around 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: Suppose a male is randomly selected from the population. Define the events $A =$ "selected male has HIV" and $B =$ "blood test is positive". We are asked to find $P(A|B)$.

From the information given we know that

$$\begin{aligned} P(A) = 0.0004, &\quad P(\bar{A}) = 0.9996 \\ P(B|A) = 0.98, &\quad P(B|\bar{A}) = 0.005 \end{aligned}$$

Therefore we can find

$$\begin{aligned} P(AB) &= P(A)P(B|A) = 0.000392 \\ P(\bar{A}B) &= P(\bar{A})P(B|\bar{A}) = 0.004998 \end{aligned}$$

Thus

$$P(B) = P(AB) + P(\bar{A}B) = 0.00539$$

and

$$P(A|B) = \frac{P(AB)}{P(B)} = 0.0727$$

Thus, if a randomly selected male tests positive, there is still only a small probability (0.0727) that they actually have HIV!

Exercise: Try to explain in ordinary words why this is the case.

Bayes' Theorem: Suppose A and B are events defined on a sample space S. Suppose also that $P(B) > 0$. Then

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

Proof:

$$\begin{aligned}\frac{P(B|A)P(A)}{P(B|\overline{A})P(\overline{A}) + P(B|A)P(A)} &= \frac{P(AB)}{P(\overline{A}B) + P(AB)} \quad \text{by the Product Rule} \\ &= \frac{P(AB)}{P(B)} \quad \text{by the Law of Total Probability} \\ &= P(A|B)\end{aligned}$$

Remark: Bayes' Theorem allows us to write conditional probabilities in terms of similar conditional probabilities but with the order of conditioning reversed. It is a simple theorem, but it has inspired approaches to problems in statistics and other areas such as machine learning, classification and pattern recognition. In these areas the term "Bayesian methods" is often used. The result is named after a mathematician[11] who proved it in the 1700's.

Problems

4.5.1 If you take a bus to work in the morning there is a 20% chance you'll arrive late. When you go by bicycle there is a 10% chance you'll be late. 70% of the time you go by bike, and 30% by bus. Given that you arrive late, what is the probability you took the bus?

4.5.2 A box contains 4 coins – 3 fair coins and 1 biased coin for which $P(\text{heads}) = 0.8$. A coin is picked at random and tossed 6 times. It shows 5 heads. Find the probability this coin is fair.

4.5.3 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?

[11] (Rev) Thomas Bayes (1702-1761) was an English Nonconformist minister, turned Presbyterian. He may have been tutored by De Moivre. His famous paper introducing this rule was published after his death. "Bayesians" are statisticians who opt for a purely probabilistic view of inference. All unknowns obtain from some distribution and ultimately, the distribution says it all.

4.6 Chapter 4 Problems

1. Suppose A and B are mutually exclusive events with $P(A) = 0.25$ and $P(B) = 0.4$. Determine the probabilities of the following events:

$$\overline{A}, \quad \overline{B}, \quad A \cup B, \quad A \cap B, \quad \overline{A} \cup \overline{B}, \quad \overline{A} \cap \overline{B}, \quad \overline{A \cap B}$$

2. Three digits are chosen at random with replacement from $0, 1, \ldots, 9$. Define the following events:

A: "all three digits are the same"	D: "the digits all exceed 4"
B: "all three digits are different"	E: "digits all have the same parity (all odd or all even)"
C: "the digits are all nonzero"	

Determine probabilities of the events A, B, C, D, E and the events

$$B \cap E, \quad B \cup D, \quad B \cup D \cup E, \quad (A \cup B) \cap D, \quad A \cup (B \cap D)$$

Show the events $(A \cup B) \cap D$ and $A \cup (B \cap D)$ in a Venn diagram.

3. Let A and B be events defined on the same 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?

4. A die is loaded to give the probabilities:

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

The die is rolled 8 times. Rolls of the die are assumed to be independent. Find the probability that

(a) the number 1 does not occur in the 8 rolls

(b) the number 2 does not occur in the 8 rolls

(c) the number 1 and the number 2 do not occur in the 8 rolls

(d) the numbers 1 and 2 both occur at least once in the 8 rolls.

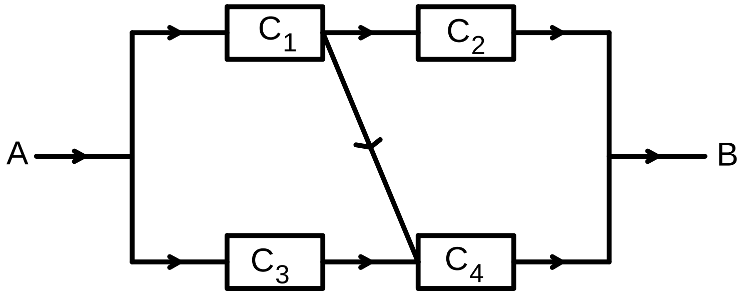
5. Events A and B are independent with $P(A) = 0.3$ and $P(B) = 0.2$. Find $P(A \cup B)$.

6. Let E and F be independent events with $E = A \cup B$ and $F = A \cap B$. Prove that either $P(A \cap B) = 0$ or $P\left(\overline{A} \cap \overline{B}\right) = 0$.

7. A population consists of F females and M males; the 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 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.

8. Suppose A, B, C, and D are events defined on a sample space such that A and C are mutually exclusive events, A and $\overline{B}$ are independent events, B and D are not independent events and $A \subset D$. Suppose also that $P(A) = 0.15$, $P\left(\overline{B}\right) = 0.3$, $P(C) = 0.1$, and $P\left(D|\overline{B}\right) = 0.8$. Determine the probabilities of the following events:
$$A \cup B, \quad B \cap D|A, \quad B \cup \overline{D}, \quad C|\overline{A} \cup B$$

9. Consider a system of independent components shown in the figure below. The system functions properly if all components along at least one path from point A to point B are working. The probabilities that the components C_1, C_2, C_3, C_4 are working are $0.9, 0.8, 0.7, 0.6$ respectively. What is the probability that the system is functioning properly?



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

    (a) 3 out of 5 customers pay by debit card

    (b) the 5th customer is the first one to pay by cash

    (c) the 5th customer is the 3rd one to pay by debit card.

11. Students A, B and C each independently answer a question on a test. The probability of getting the correct answer is 0.9 for A, 0.7 for B and 0.4 for C.

    (a) What is the probability that all three students get the correct answer?

    (b) Find the probability that exactly two students get the correct answer.

    (c) If exactly two students get the correct answer, what is the probability student C got the wrong answer?

12. Suppose you are playing a game where you flip a coin to determine who plays first. You know that when you play first, you win the game 60% of the time and that when you play second, you lose 52% of the time.

    (a) What is the probability that you win the game given that you played second?

    (b) What is the probability that you win the game?

    (c) If you won the game, what is the probability that you played first?

13. In a large population, people are one of three genetic types A, B and C: 30% are type A, 60% type B and 10% 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.

    (a) What is the probability a randomly selected person is susceptible for the disease?

    (b) If ten unrelated persons are selected at random, what is the probability at least one is susceptible for the disease?

14. 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 A's field, the next three games on B's field, and the last two on A's field. The probability that A wins a game is 0.7 at home and 0.5 away. Assume that the results of the games are independent. Find the probability that:

    (a) A wins the series in 4 games; in 5 games;

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

15. An experiment has three possible outcomes A, B and C with respective probabilities p, q and r, where $p + q + r = 1$. The experiment is repeated until either outcome A or outcome B occurs. Show that A occurs before B with probability $p/(p + q)$. A tree diagram is useful for solving this problem.

16. In the game of craps, a player rolls two dice. The player wins at once if the total is 7 or 11, and loses at once if the total is 2, 3, or 12. Otherwise, the player continues rolling the dice until they either win by throwing their initial total again, or lose by rolling 7. Show that the probability the player wins is 0.493. **Hint**: Use the result of Problem 15.

17. **Slot machines:** Standard slot machines have three wheels, each marked with some number of symbols at equally spaced positions around the wheel. For this problem suppose there are ten 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, house) is used in a total of ten positions across the three wheels. A payout occurs whenever all three symbols showing are the same.

    (a) If wheels 1, 2, 3 have 2, 6, 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.

18. The following table of probabilities are based on data from the 2011 Canadian census data. The probabilities are for Canadians aged $25 - 34$.

Highest level of education attained	Employed	Unemployed
No certificate, diploma or degree	0.066	0.010
High school diploma or equivalent	0.185	0.016
Postsecondary certificate, diploma or degree	0.683	0.040

    (a) What proportion of Canadians aged $25 - 34$ are unemployed?

    (b) What proportion of Canadians aged $25 - 34$ have no certificate, diploma or degree?

    (c) What proportion of Canadians aged $25 - 34$ have at least a high school diploma or equivalent?

    (d) What proportion of Canadians aged $25 - 34$ who are employed have at least a high school diploma or equivalent?

    (e) Are the events, “unemployed” and “no certificate, diploma or degree”, independent events? Why?

19. 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 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. 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[12]. 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?

20. **Diagnostic tests:** See Chapter 2, Problem 6. For a randomly selected person let $D =$ “person has the disease” and $R =$ “the test result is positive”. Give estimates of the following probabilities: $P(R|D)$, $P(R|\overline{D})$, $P(R)$.

21. **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 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 (“regular”)”
A	=	“Message contains the word Viagra”

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

    (a) Find $P\left(A\right) = P$(email message contains the word Viagra).

    (b) Find P(Spam$|A$) and P(Not Spam$|A$).

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

[12] “A foolish faith in authority is the worst enemy of truth.” Albert Einsten, 1901.

22. **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 three binary features, and we define events
$A_i =$ feature i appears in a message, $i = 1, 2, 3$.

    Assume that A_1, A_2, A_3 are independent events, given that a message is spam, and that they are also independent events, given that a message is not spam.

    From a study of email messages coming into a certain system it is estimated that P(Spam) $= 0.5$ and

    $P(A_1|$ Spam$) = 0.2$ $P(A_1|$ Not Spam$) = 0.005$
    $P(A_2|$ Spam$) = 0.1$ $P(A_2|$ Not Spam$) = 0.004$
    $P(A_3|$ Spam$) = 0.1$ $P(A_3|$ Not Spam$) = 0.005$

    (a) Find $P\,(\text{message has all three features}) = P\,(A_1A_2A_3)$.

    (b) Find P(Spam $|A_1A_2A_3)$.

    (c) Suppose a message has features 1 and 2 present, but feature 3 is not present. Determine P(Spam $|A_1A_2\overline{A}_3$).

    (d) If you declare any message with one or more of features 1, 2 or 3 as spam, what fraction of spam emails would you detect? Compare this with Problem 21(c).

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

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

23. **Online fraud detection:** Methods like those in Problems 21 and 22 are also used in monitoring events such as credit card transactions for potential fraud. Unlike the case of spam email, however, the fraction of transactions that are fraudulent is usually very small. What we hope to do in this case is to "flag" certain transactions so that they can be checked for potential fraud, and perhaps to block (deny) certain transactions. This is done by identifying features of a transaction so that if $F =$ "transaction is fraudulent", then

    $$r = \frac{P(\text{feature present}|F)}{P(\text{feature present}|\bar{F})}$$

    is large.

    (a) Suppose $P(F) = 0.0005$ and that $P(\text{feature present}|\overline{F}) = 0.02$. Determine $P(F|$ feature present) as a function of r, and give the values when $r =$ 10, 30 and 50.

(b) Suppose $r = 50$ and you decide to flag transactions if the feature is present. What percentage of transactions would be flagged? Does this seem like a good idea?

24. **Challenge problem:** n music lovers have reserved seats in a theatre containing a total of $n + k$ seats (k seats are unassigned). The first person who enters the theatre, however, lost his seat assignment and chooses a seat at random. Subsequently, people enter the theatre one at a time and sit in their assigned seat unless it is already occupied. If it is, they choose a seat at random from the remaining empty seats. What is the probability that person n, the last person to enter the theatre, finds their seat already occupied?

25. **Challenge problem (Monty Hall):** You have been chosen as finalist on a television show. For your prize, the host shows you three doors. Behind one door is a sports car, and behind the other two are goats. After you choose one door, the host, who knows what is behind each of the three doors, opens one (never the one you chose or the one with the car) and then says:“You are allowed to switch the door you chose if you find that advantageous”. Should you switch?

5. DISCRETE RANDOM VARIABLES

5.1 Random Variables and Probability Functions

Probability models are used to describe outcomes associated with random processes. So far we have used sets $A, B, C, \ldots$ in sample spaces to describe such outcomes. In this chapter we introduce numerical-valued variables $X, Y, \ldots$ to describe outcomes. This allows probability models to be manipulated easily using ideas from algebra, calculus, or geometry.

A random variable (r.v.) is a numerical-valued variable that represents outcomes in an experiment or random process. For example, suppose an experiment consists of tossing a coin 3 times. Then

$$X = \text{number of heads that occur}$$

would be a random variable. Associated with any random variable is a **range** A, which is the set of possible values for the variable. For example, the random variable $X =$ number of heads that occur, has range $A = \{0, 1, 2, 3\}$.

Random variables are denoted by capital letters like $X, Y, \ldots$ and their possible values are denoted by $x, y, \ldots$. This gives a nice short-hand notation for outcomes. For example, "$X = 2$" in the experiment above stands for "2 heads occurred".

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

Definition 11 *A **random variable** is a function that assigns a real number to each point in a sample space S.*

To understand this definition, consider the experiment in which a coin is tossed 3 times, and suppose that we use the sample space

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

and define a random variable as $X =$ number of heads that occur. The range of the random variable X is $A = \{0, 1, 2, 3\}$. For points in the sample space, for example $a = THH$, the value of the function $X(a)$ is obtained by counting the number of heads, $X(a) = 2$ in this case. Each of the outcomes "$X = x$" represents an event (either simple or compound). For example they are as follows:

Table 4.1

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

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. For example the events in the second column of Table 4.1 are mutually exclusive (for example $\{TTT\} \cap \{HTT, THT, TTH\} = \emptyset$) and their union is the whole 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 given in words rather than by a formula. As mentioned above, we generally denote random variables, in the abstract, by capital letters $(X, Y,$ etc.$)$ and denote the actual numbers taken by random variables by small letters $(x, y,$ etc.). You should know that there is a difference between a **function** ($f(x)$ or $X(a)$) and the **value of a function** (for example $f(2)$ or $X(a) = 2$).

Since "$X = x$" represents an event of some kind, we will be interested in its probability, which we write as $P(X = x)$. In the above example in which a fair coin is tossed three times, we might wish the probability that X is equal to 2, or $P(X = 2)$. This is $P(\{HHT, HTH, THH\}) = \frac{3}{8}$ in the example. We classify random variables into two types, according to how big their range of values is:

Discrete random variables take integer values or, more generally, values in a countable set. Recall that a set is countable if its elements can be placed in a one-one correspondence with a subset of the positive integers.

Continuous random variables take values in some interval of real numbers like $(0, 1)$ or $(0, \infty)$ or $(-\infty, \infty)$. You should be aware that the cardinality of the real numbers in an interval is NOT countable.

Examples of each might be (where we assume the values in the second column are the actual values, not rounded in any way),

Discrete	**Continuous**
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.

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 ignore this possibility here and concentrate first on discrete random variables. Continuous random variables are considered in Chapter 8.

Our aim is to set up general models which describe how the probability is distributed among the possible values in the range of a random variable X. To do this we define for any discrete random variable X the probability function.

Definition 12 *Let X be a discrete random variable with range$(X) = A$. The **probability function** (p.f.) of X is the function*

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

The set of pairs $\{(x, f(x)) : x \in A\}$ is called the **probability distribution** of X.

All probability functions must have two properties:

1. $f(x) \geq 0$ for all $x \in A$
2. $\sum\limits_{\text{all } x \in A} f(x) = 1$

By implication, these properties ensure that $f(x) \leq 1$ for all x.

We consider a few simple examples before dealing with more complicated problems.

Example: Let X be the number obtained when a die is thrown. We would normally use the probability function $f(x) = 1/6$ for $x = 1, 2, \ldots, 6$. In fact there is probably no absolutely perfect die in existence. For most dice, however, the 6 sides will be close enough to being equally likely that the model $f\,(x) = 1/6$ is a satisfactory one for the distribution of probability among the possible outcomes.

Example: Suppose a “fair” coin is tossed 3 times, with the results on the three tosses independent, and let X be the total number of heads occurring. Refer to Table 4.1 and compute the probabilities of the four events listed there; you obtain

Table 4.2

Events	Definition of this event	$P(X = x)$
$X = 0$	$\{TTT\}$	$\frac{1}{8}$
$X = 1$	$\{HTT, THT, TTH\}$	$\frac{3}{8}$
$X = 2$	$\{HHT, HTH, THH\}$	$\frac{3}{8}$
$X = 3$	$\{HHH\}$	$\frac{1}{8}$

Thus the probability function has values $f(0) = \frac{1}{8}$, $f(1) = \frac{3}{8}$, $f(2) = \frac{3}{8}$, $f(3) = \frac{1}{8}$. In this case it is easy to see that the number of points in each of the four events of the form "$X = x$" is $\binom{3}{x}$ using the counting arguments of Chapter 3, so we can give a simple algebraic expression for the probability function,

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

Example: Find the value of k which makes $f(x)$ below a probability function.

x	0	1	2	3
$f(x)$	k	$2k$	0.3	$4k$

Since the probability of all possible outcomes must add to one, $\sum\limits_{x=0}^{3} f(x) = 1$ giving $7k + 0.3 = 1$. Hence $k = 0.1$.

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

Definition 13 *The cumulative distribution function (c.d.f.) of X is the function usually denoted by $F(x)$*

$$F(x) = P(X \leq x) \quad \textit{defined for all } x \in \Re$$

In the last example, with $k = 0.1$, the range of values for the random variable is $A = \{0, 1, 2, 3\}$. For $x \in A$ we have

x	$f(x) = P(X = x)$	$F(x) = P(X \leq x)$
0	0.1	0.1
1	0.2	0.3
2	0.3	0.6
3	0.4	1

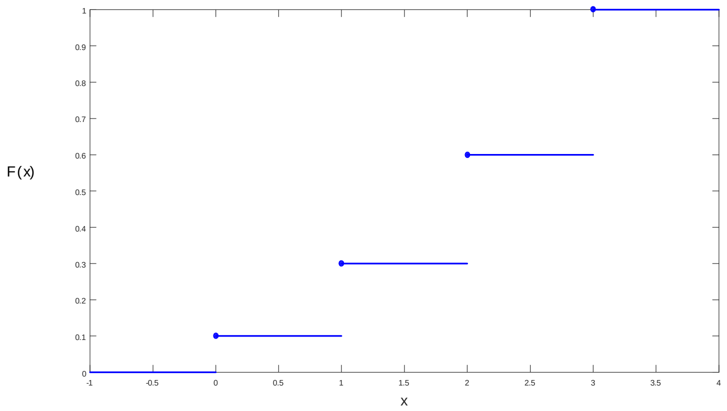


Figure 5.1: A simple cumulative distribution function

Note that the values in the third column are partial sums of the values of the probability function in the second column. For example,

$$F(1) = P(X \leq 1) = P(X = 0) + P(X = 1) = f(0) + f(1) = 0.3$$
$$F(2) = P(X \leq 2) = f(0) + f(1) + f(2) = 0.6.$$

$F(x)$ is defined for real numbers $x \notin A$ not in the range of the random variable, for example

$$F(2.5) = F(2) = 0.6 \text{ and } F(3.8) = 1$$

The cumulative distribution function for this example is plotted in Figure 5.1.

In general, $F(x)$ can be obtained from $f(x)$ using

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

A cumulative distribution function $F(x)$ has certain properties, just as a probability function $f(x)$ does. Obviously, since it represents a probability, $F(x)$ must be between 0 and 1. In addition it must be a non-decreasing function (e.g. $P(X \leq 8)$ cannot be less than $P(X \leq 7)$). Thus we note the following properties of a cumulative distribution function $F(x)$**:**

1. $F(x)$ is a non-decreasing function of x for all $x \in \Re$.

2. $0 \leq F(x) \leq 1$ for all $x \in \Re$.

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

We have noted above that $F(x)$ can be obtained from $f(x)$. The opposite is also true; for example the following result holds:

If X takes on integer values then for values x such that $x \in A$ and $x-1 \in A$,

$$f(x) = F(x) - F(x-1)$$

This says that $f(x)$ is the size of the jump in $F(x)$ at the point x. To prove this, just note that

$$F(x) - F(x-1) = P(X \le x) - P(X \le x-1) = P(X = x)$$

When a random variable has been defined it is sometimes simpler to find its probability function $f(x)$ first, and sometimes it is simpler to find $F(x)$ first. The following example gives two approaches for the same problem.

Example: 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. Find the probability function for the random variable $X =$ largest number selected.

Solution 1: If $X = x$ then we must select the number x plus $n-1$ numbers from the set $\{1, 2, \ldots, x-1\}$. (Note that this means we need $x \ge n$.) This gives

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

Solution 2: First find $F(x) = P(X \le x)$. Noting that $X \le x$ if and only if all n balls selected are from the set $\{1, 2, \ldots, x\}$, we get

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

We can now find

$$\begin{aligned} f(x) &= F(x) - F(x-1) \\ &= \frac{\binom{x}{n} - \binom{x-1}{n}}{\binom{N}{n}} \\ &= \frac{\binom{x-1}{n-1}}{\binom{N}{n}} \end{aligned}$$

as before.

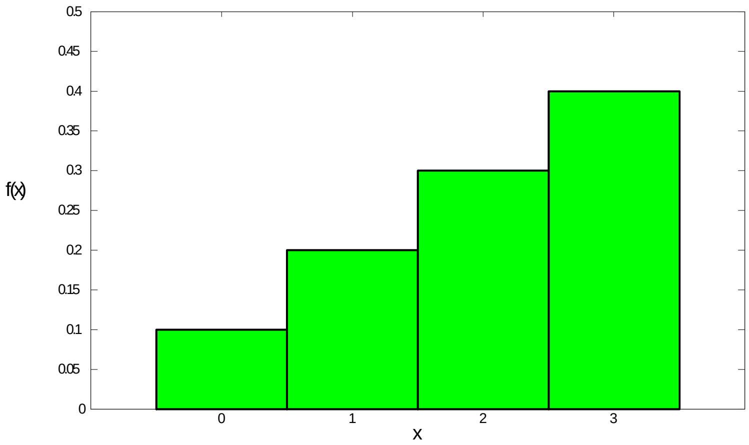


Figure 5.2: Probability histogram for $f(x) = \frac{x+1}{10}$, $x = 0, 1, 2, 3$

Remark: When you write down a probability function, remember to state its domain (that is, the possible values of the random variable, or the values x for which $f(x)$ is defined). This is an essential part of the function's definition.

We frequently graph the probability function $f(x)$ using a probability **histogram**. For now, we'll define this only for random variables whose range is some set of consecutive integers $\{0, 1, \ldots\}$. A histogram of $f(x)$ is then a graph consisting of adjacent bars or rectangles. At each x we place a rectangle with base on $(x - 0.5, x + 0.5)$ and with height $f(x)$. In the above example, a histogram of $f(x)$ looks like that in Figure 5.2.

Notice that the areas of these rectangles correspond to the probabilities, so for example $P(X = 1)$ is the area of the bar above and centered around the value 1 and $P(1 \leq X \leq 3)$ is the sum of the area of the three rectangles above the points $1, 2$, and 3 (actually the area of the region above between the points $x = 0.5$ and $x = 3.5$). In general in a probability histogram, probabilities are depicted by areas.

Model Distributions:

Many processes or problems have the same structure. In the remainder of this course we will identify common types of problems and develop probability distributions that represent them. In doing this it is important to be able to strip away the particular wording of a problem and look for its essential features. For example, the following three problems are all essentially the same.

(a) A fair coin is tossed 10 times and the "number of heads obtained" (X) is recorded.

(b) Twenty seeds are planted in separate pots and the "number of seeds germinating" (X) is recorded.

(c) Twelve items are picked at random from a factory's production line and examined for defects. The number of items having no defects (X) is recorded.

What are the common features? In each case the process consists of "trials" which are repeated a stated number of times: 10, 20, and 12. In each repetition there are two types of outcomes: heads/tails, germinate/don't germinate, and no defects/defects. These repetitions are independent (as far as we can determine), with the probability of each type of outcome remaining constant for each repetition. The random variable we record is the number of times one of these two types of outcome occurred.

Six model distributions for discrete random variables will be developed in the rest of this chapter. Students often have trouble deciding which one (if any) to use in a given setting, so be sure you understand the physical setup which leads to each one. Also, as illustrated above you will need to learn to focus on the essential features of the situation as well as the particular content of the problem.

Statistical Computing

A number of major software systems have been developed for Probability and Statistics. We will use a system called R, which has a wide variety of features and which has Unix and Windows versions. Chapter 6 gives a brief introduction to R, and how to access it. For this course, R can compute probabilities for all the distributions we consider, can graph functions or data, and can simulate random processes. In the sections below we will indicate how R can be used for some of these tasks.

Problems

5.1.1 Find c if X is a random variable with probability function

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

5.1.2 Suppose that 5 people, including you and a friend, line up at random. Let X be the number of people standing between you and your friend. Tabulate the probability function and the cumulative distribution function for X.

5.2 Discrete Uniform Distribution

We define each model in terms of an abstract "physical setup", or setting, and then consider specific examples of the setup.

Physical Setup: Suppose the range of X is $\{a, a+1, \ldots, b\}$ where a and b are integers and suppose all values are equally probable. Then X has a Discrete Uniform distribution on the set $\{a, a+1, \ldots, b\}$. The variables a and b are called the parameters of the distribution.

Illustrations:

1. If X is the number obtained when a die is rolled, then X has a discrete Uniform distribution with $a = 1$ and $b = 6$.

2. Computer random number generators generate observations from a Discrete Uniform distribution on the set $\{1, 2, \ldots, N\}$, for a specified positive integer N. These are used for many purposes, e.g. generating lottery numbers or providing automated random sampling from a set of N items.

Probability Function: There are $b - a + 1$ values in the set $\{a, a + 1, \ldots, b\}$ so the probability of each value must be $\frac{1}{b-a+1}$ in order that $\sum\limits_{x=a}^{b} f(x) = 1$. Therefore

$$f(x) = P\left(X = x\right) = \begin{cases} \frac{1}{b-a+1} & \text{for } x = a, a + 1, \ldots, b \\ 0 & \text{otherwise} \end{cases}$$

Example: Suppose a fair die is thrown once and let X be the number on the face. Find the cumulative distribution function of X.

Solution: This is an example of a Discrete Uniform distribution on the set $\{1, 2, 3, 4, 5, 6\}$ having $a = 1$, $b = 6$ and probability function

$$f(x) = P\left(X = x\right) = \begin{cases} \frac{1}{6} & \text{for } x = 1, 2, \ldots, 6 \\ 0 & \text{otherwise} \end{cases}$$

The cumulative distribution function is $F(x) = P(X \leq x)$,

$$F(x) = P\left(X \leq x\right) = \begin{cases} 0 & \text{for } x < 1 \\ \frac{[x]}{6} & \text{for } 1 \leq x < 6 \\ 1 & \text{for } x \geq 6 \end{cases}$$

where $[x]$ is the largest integer less than or equal to x.

Many distributions are constructed using Discrete Uniform random variables. For example we might throw two dice and sum the values on their faces.

Example: Suppose two fair dice (suppose for simplicity one is red and the other is green) are thrown. Let X be the sum of the values on their faces. Find the cumulative distribution function of X.

Solution: In this case we can consider the sample space to be

$$S = \{(1,1), (1,2), (1,3), \ldots, (5,6), (6,6)\}$$

where for example the outcome (i, j) means we obtained i on the red die and j on the green. There are 36 outcomes in this sample space, all with the same probability $\frac{1}{36}$. The probability function of X is easily found. For example $f(5)$ is the probability of the event $X = 5$ or the probability of $\{(1,4), (2,3), (3,2), (4,1)\}$ so $f(5) = \frac{4}{36}$. The probability function and the cumulative distribution function are given below:

x	2	3	4	5	6	7	8	9	10	11	12
$f(x) = P(X = x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$
$F(x) = P(X \le x)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{15}{36}$	$\frac{21}{36}$	$\frac{26}{36}$	$\frac{30}{36}$	$\frac{33}{36}$	$\frac{35}{36}$	1

Although it is a bit more difficult to give a formula for the cumulative distribution function for general argument x in this case, it is clear for example that $F(x) = F([x])$ and $F(x) = 0$ for $x < 2$, $F(x) = 1$ for $x \ge 12$.

Example: Let X be the largest number when a die is rolled 3 times. First find the cumulative distribution function, and then find the probability function of X.

Solution: This is another example of a distribution constructed from the Discrete Uniform. In this case the sample space

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

consists of all 6^3 possible outcomes of the three dice, with each outcome having probability $\frac{1}{216}$. Suppose that x is an integer between 1 and 6. What is the probability that the largest of these three numbers is less than or equal to x? This requires that all three of the dice show numbers less than or equal to x, and there are exactly x^3 points in S which satisfy this requirement. Therefore the probability that the largest number is less than or equal to x is

$$F(x) = \frac{x^3}{6^3}$$

for $x = 1, 2, 3, 4, 5, 6$. Here is the cumulative distribution function for all real values of x:

$$F(x) = P(X \le x) = \begin{cases} \frac{[x]^3}{216} & \text{for } 1 \le x < 6 \\ 0 & \text{for } x < 1 \\ 1 & \text{for } x \ge 6 \end{cases}$$

To find the probability function we may use the fact that for $x \in \{1, 2, 3, 4, 5, 6\}$ we have $P(X = x) = P(X \le x) - P(X < x)$ so that

$$\begin{aligned} f(x) &= F(x) - F(x-1) \\ &= \frac{x^3 - (x-1)^3}{216} \\ &= \frac{[x - (x-1)][x^2 + x(x-1) + (x-1)^2]}{216} \\ &= \frac{3x^2 - 3x + 1}{216} \quad \text{for } x = 1, 2, 3, 4, 5, 6 \end{aligned}$$

5.3 Hypergeometric Distribution

[13]**Physical Setup:** We have a collection of N objects which can be classified into two distinct types. Call one type "success"[14] (S) and the other type "failure" (F). There are r successes and $N - r$ failures. Pick n objects at random **without replacement**. Let X be the number of successes obtained. Then X has a Hypergeometric distribution. The parameters of the distribution are N, r, and n.

Illustrations:

1. The number of aces X in a bridge hand has a Hypergeometric distribution with $N = 52$, $r = 4$, and $n = 13$.

2. In a fleet of 200 trucks there are 12 which have defective brakes. In a safety check 10 trucks are picked at random for inspection. The number of trucks X with defective brakes chosen for inspection has a Hypergeometric distribution with $N = 200$, $r = 12$, and $n = 10$.

Probability Function: Using counting techniques we note there are $\binom{N}{n}$ points in the sample space S if we don't consider order of selection. There are $\binom{r}{x}$ ways to choose the x success objects from the r available and $\binom{N-r}{n-x}$ ways to choose the remaining $(n - x)$ objects from the $(N - r)$ failures. Hence

$$f(x) = P(X = x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$$

The range of values for x is somewhat complicated. Of course, $x \ge 0$. However if the number, n, picked exceeds the number, $N - r$, of failures, the difference, $n - (N - r)$ must be successes. So

[13] This section optional for STAT 220.

[14] "If A is a success in life, then A equals x plus y plus z. Work is x; y is play; and z is keeping your mouth shut." Albert Einstein, 1950

$x \geq \max(0, n-N+r)$. Also, $x \leq r$ since we can't get more successes than the number available. But $x \leq n$, since we can't get more successes than the number of objects chosen. Therefore $x \leq \min(r, n)$.

Example: In Lotto 6/49 a player selects a set of six numbers (with no repeats) from the set $\{1, 2, \dots, 49\}$. In the lottery draw six numbers are selected at random. Find the probability function for X, the number from your set which are drawn.

Solution: Think of your numbers as the S objects and the remainder as the F objects. Then X has a Hypergeometric distribution with $N = 49, r = 6$ and $n = 6$, so

$$f(x) = P(X = x) = \frac{\binom{6}{x}\binom{43}{6-x}}{\binom{49}{6}} \quad \text{for } x = 0, 1, \dots, 6$$

For example, you win the jackpot prize if $X = 6$; the probability of this is $\binom{6}{6}/\binom{49}{6}$, or about 1 in 13.9 million.

Remark: When parameter values are large, Hypergeometric probabilities may be tedious to compute using a basic calculator. The `R` functions `dhyper` and `phyper` can be used to evaluate $f(x)$ and $F(x)$. In particular, `dhyper(x,r,N-r,n)` gives $f(x)$ and `phyper(x,r,N-r,n)` gives $F(x)$. Using this we find for the Lotto 6/49 problem here, for example, that $f(6)$ is calculated by typing `dhyper(6,6,43,6)` in `R`, which returns the answer 7.151124×10^{-8} or $1/13,983,186$.

For all of our model distributions we can also confirm that $\sum\limits_{\text{all } x} f(x) = 1$. To do this here we use a summation result from Chapter 4 called the Hypergeometric Identity. Letting $a = r$ and $b = N - r$ we obtain

$$\sum_{\text{all } x} f(x) = \sum_{\text{all } x} \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} = \frac{1}{\binom{N}{n}} \sum_{\text{all } x} \binom{r}{x}\binom{N-r}{n-x} = \frac{\binom{r+N-r}{n}}{\binom{N}{n}} = 1$$

Problems

5.3.1 A box of 12 tins of tuna contains d which are tainted. Suppose 7 tins are opened for inspection and none of these 7 is tainted.

(a) Calculate the probability that none of the 7 is tainted for $d = 0, 1, 2, 3$.

(b) Do you think it is likely that the box contains as many as 3 tainted tins?

5.3.2 Suppose our sample space distinguishes points with different orders of selection. For example suppose that $S = \{SSSSFFF \dots .,\}$ consists of all words of length n where letters are drawn without replacement from a total of r S's and $N - r$ F's. Derive a formula for the probability that the word contains exactly X S's. In other words, determine the Hypergeometric probability function using a sample space in which order of selection is considered.

5.4 Binomial Distribution

Physical Setup:
Suppose an "experiment" has two types of distinct outcomes. Call these types "success" (S) and "failure" (F). Let $P(S) = p$ and $P(F) = 1 - p$. Repeat the experiment n **independent** times. Let X be the number of successes obtained. Then X has what is called a **Binomial distribution**. The parameters of the distribution are n and p. We write $X \sim Binomial(n, p)$ as a shorthand for "X is distributed according to a Binomial distribution with n repetitions and probability p of success" or "X has a Binomial distribution with parameters n and p". The n individual experiments in the process just described are often called "trials" or "Bernoulli trials" and the process is called a Bernoulli[15] process or a Binomial process.

Illustrations:

1. Toss a fair die 10 times and let X be the number of sixes that occur. Then $X \sim Binomial(10, 1/6)$.
2. In a microcircuit manufacturing process, 90% of the chips produced work (10% are defective). Suppose we select 25 chips, independently[16] and let X be the number that work. Then $X \sim Binomial(25, 0.9)$.

Comment: We must think carefully whether the physical process we are considering is closely approximated by a Binomial 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 we would need to think about the manufacturing process. Microcircuit chips are 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 25 chips from the same wafer, or from only 2 or 3 wafers, that the "trials" (chips) might not be independent, or perhaps that the probability of defectives changes.

[15] After James (Jakob) Bernoulli (1654 – 1705), a Swiss member of a family of eight mathematicians. Nicolaus Bernoulli was an important citizen of Basel, being a member of the town council and a magistrate. Jacob Bernoulli's mother also came from an important Basel family of bankers and local councillors. Jacob Bernoulli was the brother of Johann Bernoulli and the uncle of Daniel Bernoulli. He was compelled to study philosophy and theology by his parents, graduated from the University of Basel with a master's degree in philosophy and a licentiate in theology but against his parents wishes, studied mathematics and astronomy . He was offered an appointment in the Church he turned it down instead taught mechanics at the University in Basel from 1683, giving lectures on the mechanics of solids and liquids. Jakob Bernoulli is responsible for many of the combinatorial results dealing with independent random variables which take values 0 or 1 in these notes. He was also a fierce rival of his younger brother Johann Bernoulli, also a mathematician, who would have liked the chair of mathematics at Basel which Jakob held.

[16] for example we select at random with replacement or without replacement from a *very large number* of chips.

Probability Function: There are $\frac{n!}{x!(n-x)!} = \binom{n}{x}$ different arrangements of x S's and $(n-x)$ F's over the n trials. The probability for each of these arrangements has p multiplied together x times and $(1-p)$ multiplied $(n-x)$ times, in some order, since the trials are independent. So each arrangement has probability $p^x(1-p)^{n-x}$. Therefore

$$f(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, \ldots, n \text{ and } 0 < p < 1$$

Checking that $\sum\limits_{\text{all } x} f(x) = 1$ for $0 < p < 1$:

$$\begin{aligned}\sum_{x=0}^{n} f(x) &= \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} \\ &= (1-p)^n \sum_{x=0}^{n} \binom{n}{x} \left(\frac{p}{1-p}\right)^x \\ &= (1-p)^n \left(1 + \frac{p}{1-p}\right)^n \quad \text{by the Binomial Theorem if } 0 < p < 1 \\ &= (1-p)^n \left(\frac{1-p+p}{1-p}\right)^n \\ &= 1^n = 1\end{aligned}$$

Figure 5.3 shows the probability histogram for the Binomial distribution with parameters $n = 20$ and $p = 0.3$. Although the formula for $f(x)$ may seem complicated the shape of the histogram is simple since it increases to a maximum value near np and then decreases thereafter.

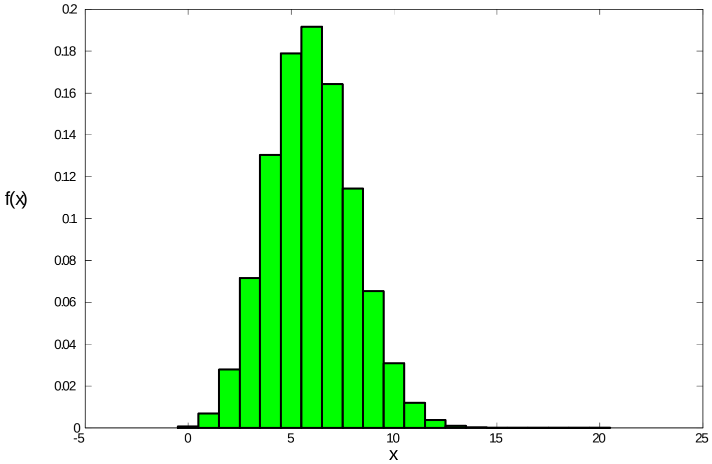


Figure 5.3: Probability histogram for a $Binomial(20, 0.3)$ random variable

Computation: Many software packages and some calculators give Binomial probabilities. In R we use the function `dbinom(x,n,p)` to compute $f(x)$ and `pbinom(x,n,p)` to compute the cumulative distribution function $F(x) = P(X \le x)$.

Example: Suppose that in a weekly lottery you have probability 0.02 of winning a prize with a single ticket. If you buy 1 ticket per week for 52 weeks, what is the probability that (a) you win no prizes, and (b) that you win 3 or more prizes?

Solution: Let X be the number of weeks that you win; then $X \sim Binomial(52, 0.02)$. We find

(a)

$$\begin{aligned} P(X = 0) &= \binom{52}{0}(0.02)^0(0.98)^{52} \\ &= 0.350 \end{aligned}$$

(b)

$$\begin{aligned} P(X \ge 3) &= 1 - P(X \le 2) \\ &= 1 - [P(X = 0) + P(X = 1) + P(X = 2)] \\ &= 0.0859 \end{aligned}$$

Note that $P(X \le 2)$ is given by the R command `pbinom(2,52,0.02)`.

Comparison of Binomial and Hypergeometric Distributions: These distributions are similar in that an experiment with two types of outcome (S and F) is repeated n times and X is the number of successes. The key difference is that the Binomial requires independent repetitions with the same probability of S, whereas the draws in the Hypergeometric are made from a fixed collection of objects **without** replacement. The trials (draws) are therefore not independent. For example, if there are $r = 10$ S objects and $N - r = 10$ F objects, then the probability of getting an S on draw two depends on what was obtained in draw one. If these draws had been made **with** replacement, however, they would be independent and we would use the Binomial rather than the Hypergeometric model.

If N is large and the number, n, 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. So 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 the Hypergeometric models should produce similar probabilities. As the Binomial is easier to calculate, it is often used as an approximation to the Hypergeometric in such cases.

Example: Suppose we have 15 cans of soup with no labels, but 6 are tomato and 9 are pea soup. We randomly pick 8 cans and open them. Find the probability 3 of the cans picked are tomato.

Solution: Let $X =$ number of tomato soup cans picked. The correct solution uses the Hypergeometric distribution and is

$$P(X = 3) = \frac{\binom{6}{3}\binom{9}{5}}{\binom{15}{8}} = 0.3916$$

If we incorrectly used the Binomial distribution, we would obtain

$$\binom{8}{3}\left(\frac{6}{15}\right)^3\left(\frac{9}{15}\right)^5 = 0.2787$$

As expected, this is a poor approximation since we are picking over half of a fairly small collection of cans.

However, if we had 1500 cans: 600 tomato and 900 pea, we are not likely to get the same can again even if we did replace each of the 8 cans after opening it. (Put another way, the probability we get a tomato soup on each pick is very close to 0.4, regardless of what the other picks give.) The Hypergeometric probability is

$$\frac{\binom{600}{3}\binom{900}{5}}{\binom{1500}{8}} = 0.2794$$

The Binomial probability is

$$\binom{8}{3}\left(\frac{600}{1500}\right)^3\left(\frac{900}{1500}\right)^5 = 0.2787$$

which is a very good approximation.

Problems

5.4.1 Megan audits 130 clients during a year and finds irregularities for 26 of them.

(a) Give an expression for the probability that 2 clients will have irregularities when 6 of her clients are picked at random,

(b) Evaluate your answer to (a) using a suitable approximation.

5.4.2 The flash mechanism on camera A fails on 10% of shots, while that of camera B fails on 5% of shots. The two cameras being identical in appearance, a photographer selects one at random and takes 10 indoor shots using the flash.

(a) Give the probability that the flash mechanism fails exactly twice. What assumption(s) are you making?

(b) Given that the flash mechanism failed exactly twice, what is the probability camera A was selected?

5.5 Negative Binomial Distribution

[17]**Physical Setup:**
The setup for this distribution is almost the same as for Binomial; that is, an experiment (trial) has two distinct types of outcome, S and F, and is repeated independently with $P(S) = p$ on each trial. Continue doing the experiment until a specified number, k, of successes have been obtained. Let X be the number of failures obtained before the kth success. Then X has a Negative Binomial distribution. We write $X \sim Negative\ Binomial(k, p)$ to denote this. The parameters of the distribution are k and p.

Illustrations:

(1) If a fair coin is tossed until we get the fifth head, the number of tails we obtain has a Negative Binomial distribution with $k = 5$ and $p = \frac{1}{2}$.

(2) As a rough approximation, the number of half credit failures a student collects before successfully completing 40 half credits for an honours degree has a Negative Binomial distribution. (Assume all course attempts are independent, with the same probability of being successful, and ignore the fact that getting more than 6 half credit failures prevents a student from continuing toward an honours degree.)

Probability Function: In all there will be $x + k$ trials (x F's and k S's) and the last trial must be a success. In the first $x + k - 1$ trials we therefore need x failures and $(k - 1)$ successes, in any order. There are $\frac{(x+k-1)!}{x!(k-1)!} = \binom{x+k-1}{x}$ different orders. Each order will have probability $p^k(1-p)^x$ since there must be x trials which are failures and k which are success. Hence

$$f(x) = P(X = x) = \binom{x+k-1}{x} p^k (1-p)^x \quad \text{for } x = 0, 1, \ldots \text{ and } 0 < p < 1$$

Note: An alternate version of the Negative Binomial distribution defines X to be the total number of trials needed to get the kth success. This is equivalent to our version. For example, asking for the probability of getting 3 tails before the fifth head is exactly the same as asking for a total of 8 tosses in order to get the fifth head. You need to be careful to read how X is defined in a problem rather than mechanically "plugging in" numbers in the above formula for $f(x)$.

Checking that $\sum\limits_{\text{all } x} f(x) = 1$ for $0 < p < 1$:
This requires somewhat more work as compared to the Binomial distribution.

[17] This section optional for STAT 220

We first re-arrange the $\binom{x+k-1}{x}$ term,

$$\binom{x+k-1}{x} = \frac{(x+k-1)^{(x)}}{x!} = \frac{(x+k-1)(x+k-2)\cdots(k+1)(k)}{x!}$$

Factor a (-1) out of each of the x terms in the numerator, and re-write these terms in reverse order,

$$\begin{aligned}\binom{x+k-1}{x} &= (-1)^x \frac{(-k)(-k-1)\cdots(-k-x+2)(-k-x+1)}{x!} \\ &= (-1)^x \frac{(-k)^{(x)}}{x!} \\ &= (-1)^x \binom{-k}{x}\end{aligned}$$

If $0 < p < 1$ then by the Binomial Theorem

$$\begin{aligned}\sum_{x=0}^{\infty} f(x) &= \sum_{x=0}^{\infty} \binom{-k}{x} (-1)^x p^k (1-p)^x \\ &= p^k \sum_{x=0}^{\infty} \binom{-k}{x} \left[(-1)(1-p)\right]^x \\ &= p^k \left[1 + (-1)(1-p)\right]^{-k} \quad \text{if } 0 < p < 1 \\ &= p^k p^{-k} \\ &= 1\end{aligned}$$

Comparison of Binomial and Negative Binomial Distributions

These should be easily distinguished because they reverse what is specified or known in advance and what is variable.

- **Binomial:** We know the number n of trials in advance but we do not know the number of successes we will obtain until after the experiment.
- **Negative Binomial:** 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.

Example: The fraction of a large population that has a specific blood type T is 0.08 (8%). For blood donation purposes it is necessary to find five people with type T blood. If randomly selected individuals from the population are tested one after another, then (a) What is the probability y persons have to be tested to get five type T persons, and (b) What is the probability that over 80 people have to be tested?

Solution: Think of a type T person as a success (S) and a non-type T as an F. Let $Y =$ number of persons who have to be tested and let $X =$ number of non-type T persons in order to get five S's. Then X has a Negative Binomial distribution with $k = 5$ and $p = 0.08$ and

$$P(X = x) = f(x) = \binom{x+4}{x}(0.08)^5(0.92)^x \text{ for } x = 0, 1, \ldots$$

We are actually asked here about $Y = X + 5$. Thus

$$\begin{aligned} P(Y = y) &= P(X = y - 5) \\ &= f(y - 5) \\ &= \binom{y-1}{y-5}(0.08)^5(0.92)^{y-5} \quad \text{for } y = 5, 6, 7, \ldots \end{aligned}$$

Thus we have the answer to (a) as given above, and for (b)

$$\begin{aligned} P(Y > 80) &= P(X > 75) \\ &= 1 - P(X \leq 75) \\ &= 1 - \sum_{x=0}^{75} f(x) \\ &= 0.2235 \end{aligned}$$

Note: Calculating such probabilities is easy with R. To get $f(x)$ we use `dnbinom(x,k,p)` and to get $F(x) = P(X \leq x)$ we use `pnbinom(x,k,p)`.

Problems

5.5.1 You can get a group rate on tickets to a play if you can find 25 people to go. Assume each person you ask responds independently and has a 20% chance of agreeing to buy a ticket. Let X be the total number of people you have to ask in order to find 25 who agree to buy a ticket. Find the probability function of X.

5.5.2 A shipment of 2500 car headlights contains 200 which are defective. You choose from this shipment without replacement until you have 18 which are not defective. Let X be the number of defective headlights you obtain.

(a) Give the probability function, $f(x)$.

(b) Using a suitable approximation, find $f(2)$.

5.6 Geometric Distribution

Physical Setup: Consider the Negative Binomial distribution with $k = 1$. In this case we repeat independent Bernoulli trials with two types of outcome, S and F, and $P(S) = p$ each time until we obtain the first success. Let X be the number of failures obtained before the first success. We write $X \sim Geometric\,(p)$. The parameter of the distribution is p.

Illustrations:

(1) The probability you win a lottery prize in any given week is a constant p. The number of weeks **before** you win a prize for the first time has a Geometric distribution.

(2) If you take STAT 230 until you pass it and attempts are independent with the same probability of a pass each time[18], then the number of failures would have a Geometric distribution. (Thankfully these assumptions are unlikely to be true for most persons! Why is this?)

Probability Function: There is only the one arrangement with x failures followed by 1 success. This arrangement has probability

$$f(x) = P(X = x) = (1-p)^x p \;\text{ for }\; x = 0, 1, \ldots \text{ and } 0 < p < 1$$

Alternatively if we substitute $k = 1$ in the probability function for the Negative Binomial, we obtain

$$\begin{aligned} f(x) &= \binom{x+1-1}{x} p^1 (1-p)^x \\ &= p(1-p)^x \;\text{ for }\; x = 0, 1, 2, \ldots \text{ and } 0 < p < 1 \end{aligned}$$

which is the same. To check that $\sum f(x) = 1$ for $0 < p < 1$, we need to evaluate a Geometric series,

$$\begin{aligned} \sum_{x=0}^{\infty} f(x) &= \sum_{x=0}^{\infty} (1-p)^x p \\ &= p \sum_{x=0}^{\infty} (1-p)^x \\ &= \frac{p}{1-(1-p)} \quad \text{if } 0 < p < 1 \;\text{ by the Geometric series} \\ &= 1 \end{aligned}$$

Note: The names of the models so far derive from the summation results which show $f(x)$ sums to one. The Geometric distribution involved a Geometric series; the Hypergeometric distribution used the

[18] you burn all notes and purge your memory of the course after each failure

Hypergeometric Identity; both the Binomial and Negative Binomial distributions used the Binomial Theorem.

Bernoulli Trials: Once again remember that the Binomial, Negative Binomial and Geometric models all involve trials (experiments) which:

(1) are independent

(2) have 2 distinct types of outcome (S and F)

(3) have the same probability p of "success" (S) each time.

Such trials are known as Bernoulli trials.

Problem

5.6.1 Suppose there is a 30% chance of a car from a certain production line having a leaky windshield. The probability an inspector will have to check at least n cars to find the first one with a leaky windshield is 0.05. Find n.

5.7 Poisson Distribution from Binomial

The **Poisson**[19] distribution has probability function of the form

$$f(x) = P\left(X = x\right) = e^{-\mu}\frac{\mu^x}{x!} \quad \text{for } x = 0, 1, \ldots$$

where $\mu > 0$ is a parameter of the distribution whose value depends on the setting for the model. Mathematically, we can see that $f(x)$ has the properties of a probability function, since $f(x) \geq 0$ for $x = 0, 1, \ldots$ and

$$\sum_{x=0}^{\infty} f(x) = e^{-\mu}\sum_{x=0}^{\infty}\frac{\mu^x}{x!} = e^{-\mu}(e^{\mu}) = 1$$

by the Exponential series.

The Poisson distribution arises in physical settings where the random variable X represents the number of events of some type. In this section we show how it arises from a Binomial process, and in the following section we consider another derivation of the model. We write $X \sim Poisson(\mu)$ to denote that X has the probability function above.

[19] After Siméon Denis Poisson (1781-1840), a French mathematician who was supposed to become a surgeon but, fortunately for his patients, failed medical school for lack of coordination. He was forced to do theoretical research, being too clumsy for anything in the lab. He wrote a major work on probability and the law, *Recherchés sur la probabilité des jugements en matière criminelle et matière civile (1837),* discovered the Poisson distribution (called law of large numbers) and to him is ascribed one of the more depressing quotes in our discipline "Life is good for only two things: to study mathematics and to teach it."

Physical Setup: One way the Poisson distribution arises is as a limiting case of the Binomial distribution as $n \to \infty$ and $p \to 0$. In particular, we keep the product np fixed at some constant value, μ, while letting $n \to \infty$. This automatically makes $p \to 0$. Let us see what the limit of the Binomial probability function $f(x)$ is in this case.

Probability Function: Since $np = \mu$, $p = \frac{\mu}{n}$ and for x fixed,

$$\begin{aligned}
f(x) &= \binom{n}{x} p^x (1-p)^{n-x} = \frac{n^{(x)}}{x!}\left(\frac{\mu}{n}\right)^x \left(1-\frac{\mu}{n}\right)^{n-x} \\
&= \frac{\mu^x}{x!} \frac{\overbrace{n(n-1)(n-2)\cdots(n-x+1)}^{x \text{ terms}}}{(n)(n)\quad(n)\quad\cdots\quad(n)} \left(1-\frac{\mu}{n}\right)^{n-x} \\
&= \frac{\mu^x}{x!}\left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\left(\frac{n-2}{n}\right)\cdots\left(\frac{n-x+1}{n}\right)\left(1-\frac{\mu}{n}\right)^n\left(1-\frac{\mu}{n}\right)^{-x} \\
&= \frac{\mu^x}{x!}(1)\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{x-1}{n}\right)\left(1-\frac{\mu}{n}\right)^n\left(1-\frac{\mu}{n}\right)^{-x} \\
\lim_{n\to\infty} f(x) &= \frac{\mu^x}{x!}\underbrace{(1)(1)(1)\cdots(1)}_{x \text{ terms}} e^{-\mu}(1)^{-x} \quad \left(\text{since } e^k = \lim_{n\to\infty}\left(1+\frac{k}{n}\right)^n\right) \\
&= \frac{\mu^x e^{-\mu}}{x!}; \quad \text{for } x = 0, 1, 2, \ldots
\end{aligned}$$

(For the Binomial the upper limit on x is n, but we are letting $n \to \infty$.)
This result allows us to use the Poisson distribution with $\mu = np$ as a close approximation to the Binomial distribution in processes for which n is large and p is small.

Example: There are 200 people at a party. What is the probability that 2 of them were born on January 1?

Solution: Assuming all days of the year are equally likely for a birthday (and ignoring February 29) and that the birthdays are independent (e.g. no twins!) we can use the Binomial distribution with $n = 200$ and $p = 1/365$ for X = number born on January 1, giving

$$f(2) = \binom{200}{2}\left(\frac{1}{365}\right)^2\left(1-\frac{1}{365}\right)^{198} = 0.086767$$

Since n is large and p is close to 0, we can use the Poisson distribution to approximate this Binomial probability, with $\mu = np = \frac{200}{365}$, giving

$$f(2) = \frac{\left(\frac{200}{365}\right)^2 e^{-\left(\frac{200}{365}\right)}}{2!} = 0.086791$$

As might be expected, this is a very good approximation.

Notes:

(1) If p is close to 1 we can also use the Poisson distribution to approximate the Binomial. By interchanging the labels "success" and "failure", we can get the probability of "success" (formerly labelled "failure") close to 0.

(2) The Poisson distribution used to be very useful for approximating Binomial probabilities with n large and p near 0 since the calculations are easier. (This assumes values of e^x to be available.) With the advent of computers, it is just as easy to calculate the exact Binomial probabilities as the Poisson probabilities. However, the Poisson approximation is useful when employing a calculator without a built in Binomial function.

(3) The R functions `dpois(x,`μ`)` and `ppois(x,`μ`)` give $f(x)$ and $F(x)$ respectively.

Problem

5.7.1 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.

(a) 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 answer given by the Poisson approximation.

(b) What assumptions does your answer depend on? How well would you expect these assumptions to be met?

5.8 Poisson Distribution from Poisson Process

[20]We now derive the Poisson distribution as a model for the number of a certain kind of event or occurrence (e.g. births, insurance claims, web site hits) that occur at points in time or in space. To this end, we use the "order" notation $g(\Delta t) = o(\Delta t)$ as $\Delta t \to 0$ to mean that the function g approaches 0 faster than Δt as Δt approaches zero, or that

$$\frac{g(\Delta t)}{\Delta t} \to 0 \text{ as } \Delta t \to 0$$

For example $g(\Delta t) = (\Delta t)^2 = o(\Delta t)$ but $(\Delta t)^{1/2}$ is not $o(\Delta t)$.

Physical Setup: Consider a situation in which a certain type of event occurs at random points in time (or space) according to the following conditions:

1. **Independence:** the number of occurrences in non-overlapping intervals are independent.

[20] This section optional for STAT 220.

2. **Individuality:** for sufficiently short time periods of length Δt, the probability of 2 or more events occurring in the interval is close to zero, that is, events occur singly not in clusters. More precisely, as $\Delta t \to 0$, the probability of two or more events in the interval of length Δt must go to zero faster than $\Delta t \to 0$ or

$$P\left(\text{2 or more events in } (t, t+\Delta t)\right) = o(\Delta t) \text{ as } \Delta t \to 0$$

3. **Homogeneity or Uniformity:** events occur at a uniform or homogeneous rate λ over time so that the probability of one occurrence in an interval $(t, t+\Delta t)$ is approximately $\lambda\Delta t$ for small Δt for any value of t. More precisely,

$$P\left(\text{one event in } (t, t+\Delta t)\right) = \lambda\Delta t + o(\Delta t)$$

These three conditions together define a **Poisson Process.**

Let X be the number of event occurrences in a time period of length t. Then it can be shown (see below) that X has a Poisson distribution with parameter $\mu = \lambda t$.

Illustrations:

(1) The emission of radioactive particles from a substance follows a Poisson process. (This is used in medical imaging and other areas.)

(2) Hits on a web site during a given time period often follow a Poisson process.

(3) Occurrences of certain non-communicable diseases sometimes follow a Poisson process.

Probability Function: We can derive the probability function for X from the conditions above. We are interested in time intervals of arbitrary length t, so as a temporary notation, let $f_t(x)$ be the probability of x occurrences in a time interval of length t. We now relate $f_t(x)$ and $f_{t+\Delta t}(x)$. From that we can determine what $f_t(x)$ is. To find $f_{t+\Delta t}(x)$ we note that for Δt small there are only two ways to get a total of x event occurrences by time $t + \Delta t$. Either there are x events by time t and no more from t to $t + \Delta t$ or there are $x - 1$ by time t and 1 more from t to $t + \Delta t$. (since P(2 or more events in $(t, t + \Delta t)) = o(\Delta t)$, other possibilities are negligible if Δt is small). This and condition 1 above (independence) imply that

$$f_{t+\Delta t}(x) \approx f_t(x)(1 - \lambda\Delta t) + f_t(x-1)(\lambda\Delta t) + o(\Delta t)^2$$

Re-arranging gives

$$\frac{f_{t+\Delta t}(x) - f_t(x)}{\Delta t} \approx \lambda \left[f_t(x-1) - f_t(x) \right] + o(\Delta t)$$

Taking the limit as $\Delta t \to 0$ we get

$$\frac{d}{dt} f_t(x) = \lambda \left[f_t(x-1) - f_t(x) \right] \tag{5.1}$$

This provides a "differential-difference" equation that needs to be solved for the functions $f_t(x)$ as functions of t for each fixed integer value of x. We know that in interval of length 0, zero events will occur, so that $f_0(0) = 1$ and $f_0(x) = 0$ for $x = 1, 2, \ldots$. At the moment we may not know how to solve such a system but let's approach the problem using the Binomial approximation of the last section. Suppose that the interval $(0, t)$ is divided into $n = \frac{t}{\Delta t}$ small subintervals of length Δt. The probability that an event falls in any subinterval (record this as a success) is approximately $p = \lambda \Delta t$ provided the interval length is small. The probability of two or more events falling in any one subinterval is less than $nP\left[2 \text{ or more events in} (t, t + \Delta t)\right] = n \times o(\Delta t)$ which goes to 0 as $\Delta t \to 0$ so we can ignore the possibility that one of the subintervals has 2 or more events in it. Also the "successes" are independent on the n different subintervals or "trials", and so the total number of successes recorded, X, is approximately Binomial(n, p). Therefore

$$P(X = x) \approx \binom{n}{x} p^x (1-p)^{n-x} = \frac{n^{(x)} p^x}{x!} (1-p)^n \left(\frac{1}{1-p} \right)^x$$

Notice that for fixed t, x, as $\Delta t \to 0$, $p = \lambda \Delta t \to 0$ and $n = \frac{t}{\Delta t} \to \infty$, and $(1-p)^n \to e^{-\lambda t}$. Also, for fixed x, $n^{(x)} p^x \to (\lambda t)^x$. This yields the approximation

$$P(X = x) \approx \frac{(\lambda t)^x e^{-\lambda t}}{x!}$$

You can confirm that

$$f_t(x) = f(x) = \frac{(\lambda t)^x e^{-\lambda t}}{x!} \text{ for } x = 0, 1, \ldots$$

provides a solution to the system (5.1) with the required initial conditions. If we let $\mu = \lambda t$, we can re-write $f(x)$ as $f(x) = \frac{\mu^x e^{-\mu}}{x!}$, which is the Poisson distribution from Section 5.7. That is:

In a Poisson process with rate of occurrence λ, the number of event occurrences X in a time interval of length t has a Poisson distribution with $\mu = \lambda t$.

Interpretation of μ and λ: λ is referred to as the **intensity** or **rate of occurrence** parameter for the events. It represents the average rate of occurrence of events per unit of time (or area or volume, as discussed below). Then $\lambda t = \mu$ represents the average number of occurrences in t units of time. It is important to note that the value of λ depends on the units used to measure time. For example, if phone calls arrive at a store at an average rate of 20 per hour, then $\lambda = 20$ when time is in hours and the average in 3 hours will be 3×20 or 60. However, if time is measured in minutes then $\lambda = 20/60 = 1/3$; the average in 180 minutes (3 hours) is still $(1/3)(180) = 60$.

Example: Suppose earthquakes recorded in Ontario each year follow a Poisson process with an average of 6 per year. What is the probability that 7 will be recorded in a two year period?

Solution: In this case $t = 2$ (years) and the intensity of earthquakes is $\lambda = 6$. Therefore X, the number of earthquakes in the two year period follows a Poisson distribution with parameter $\mu = \lambda t = 12$. The probability that 7 earthquakes will be recorded in a two year period is $f(7) = \frac{12^7 e^{-12}}{7!} = 0.0437$.

Example: At a nuclear power station an average of 8 leaks of heavy water are reported per year. Find the probability of 2 or more leaks in one month, if leaks follow a Poisson process.

Solution: Assume leaks satisfy the conditions for a Poisson process and that a month is $1/12$ of a year. Let X be the number of leaks in one month. Then X has the the Poisson distribution with $\lambda = 8$ and $t = 1/12$, so $\mu = \lambda t = 8/12$. Thus

$$\begin{aligned} P(X \geq 2) &= 1 - P(X < 2) \\ &= 1 - [f(0) + f(1)] \\ &= 1 - \left[\frac{(8/12)^0 \, e^{-8/12}}{0!} + \frac{\left(\frac{8}{12}\right)^1 e^{-8/12}}{1!} \right] \\ &= 0.1443 \end{aligned}$$

Random Occurrence of Events in Space: The Poisson process also applies when "events" occur randomly in space (either 2 or 3 dimensions). For example, the "events" might be bacteria in a volume of water or blemishes in the finish of a paint job on a metal surface. If X is the number of events in a volume or area in space of size v and if λ is the average number of events per unit volume (or area), then X has a Poisson distribution with $\mu = \lambda v$. For this model to be valid, it is assumed that the Poisson process conditions given previously apply here, with "time" replaced by "volume" or "area". Once again, note that the value of λ depends on the units used to measure volume or area.

Example: Coliform bacteria occur in river water with an average intensity of 1 bacteria per 10 cubic centimeters of water. Find (a) the probability there are no bacteria in a 20 cubic centimeter sample of water which is tested, and (b) the probability there are 5 or more bacteria in a 50 cubic centimeter sample. (To do this assume that a Poisson process describes the location of bacteria in the water at any given time.)

Solution: Let X = number of bacteria in a v cubic centimeter sample of water. Since $\lambda = 0.1$ bacteria per 1 cubic centimeter (1 per 10 cubic centimeters) the probability function of X is Poisson with $\mu = 0.1v$,

$$f(x) = e^{-0.1v}\frac{(0.1v)^x}{x!} \text{ for } x = 0, 1, 2, \ldots$$

Thus we find

(a) With $v = 20$, $\mu = 2$ so $P(X = 0) = f(0) = e^{-2} = 0.135$

(b) With $v = 50$, $\mu = 5$ so $f(x) = e^{-5}5^x/x!$ and $P(X \geq 5) = 1 - P(X \leq 4) = 1 - 0.440 = 0.560$

(Note: we can use the R command `ppois(4,5)` to get $P(X \leq 4)$.)

Exercise: In each of the above examples, how well are each of the conditions for a Poisson process likely to be satisfied?

Distinguishing Poisson from Binomial and Other Distributions

Students often have trouble knowing when to use the Poisson distribution and when not to use it. To be certain, the three conditions for a Poisson process need to be checked. However, a quick decision can often be made by asking yourself the following questions:

1. *Can we specify in advance the maximum value which X can take?*
   If we can, then the distribution is <u>not</u> Poisson. If there is no fixed upper limit, the distribution might be Poisson, but is certainly not Binomial or Hypergeometric, e.g. the number of seeds which germinate out of a package of 25 does not have a Poisson distribution since we know in advance that $X \leq 25$. The number of cardinals sighted at a bird feeding station in a week might be Poisson since we can't specify a fixed upper limit on X. At any rate, this number would not have a Binomial or Hypergeometric distribution. Of course if it is Binomial with a very large value of n and a small value of p we may still use the Poisson distribution, but in this case it is being used to approximate a Binomial.

2. *Does it make sense to ask how often the event did <u>not</u> occur?*
   If it does make sense, the distribution is not Poisson. If it does not make sense, the distribution might be Poisson. For example, it does not make sense to ask how often a person did not hiccup during an hour. So the number of hiccups in an hour might have a Poisson distribution. It would certainly not be Binomial, Negative Binomial, or Hypergeometric. If a coin were tossed until the 3rd head occurs it does make sense to ask how often heads did not come up. So the distribution would not be Poisson. (In fact, we'd use Negative Binomial for the number of non-heads or tails.)

Problems

5.8.1 Suppose that emergency calls to 911 follow a Poisson process with an average of 3 calls per minute. Find the probability there will be

(a) 6 calls in a period of 2.5 minutes.

(b) 2 calls in the first minute of a 2.5 minute period, given that 6 calls occur in the entire period.

5.8.2 Misprints are distributed randomly and uniformly in a book, at a rate of 2 per 100 lines.

(a) What is the probability a line is free of misprints?

(b) Two pages are selected at random. One page has 80 lines and the other 90 lines. What is the probability that there are exactly 2 misprints on each of the two pages?

5.9 Combining Other Models with the Poisson Process

[21]While we've considered the model distributions in this chapter one at a time, we will sometimes need to use two or more distributions to answer a question. To handle this type of problem you'll need to be very clear about the characteristics of each model. Here is a somewhat artificial illustration. Lots of other examples are given in the problems at the end of the chapter.

Example: A very large (essentially infinite) number of ladybugs is released in a large orchard. They scatter randomly so that on average a tree has 6 ladybugs on it. Trees are all the same size.

(a) Find the probability a tree has > 3 ladybugs on it.

(b) When 10 trees are picked at random, what is the probability 8 of these trees have > 3 ladybugs on them?

(c) Trees are checked until 5 with > 3 ladybugs are found. Let X be the total number of trees checked. Find the probability function, $f(x)$.

(d) Find the probability a tree with > 3 ladybugs on it has exactly 6.

(e) On 2 trees there are a total of t ladybugs. Find the probability that x of these are on the first of these 2 trees.

[21] This section optional for STAT 220.

Solution:

(a) If the ladybugs are randomly scattered the most suitable model is the Poisson distribution with $\lambda = 6$ and $v = 1$ (that is, any tree has a "volume" of one unit), so $\mu = 6$ and

$$\begin{aligned} P(X > 3) &= 1 - P(X \leq 3) = 1 - [f(0) + f(1) + f(2) + f(3)] \\ &= 1 - \left[\frac{6^0 e^{-6}}{0!} + \frac{6^1 e^{-6}}{1!} + \frac{6^2 e^{-6}}{2!} + \frac{6^3 e^{-6}}{3!}\right] = 0.8488 \end{aligned}$$

(b) Using the Binomial distribution where "success" means > 3 ladybugs on a tree, we have $n = 10$, $p = 0.8488$ and

$$f(8) = \binom{10}{8}(0.8488)^8(1 - 0.8488)^2 = 0.2772$$

(c) Using the Negative Binomial distribution, we need the number of successes, k, to be 5, and the number of failures to be $(x - 5)$. Then

$$\begin{aligned} f(x) &= \binom{x - 5 + 5 - 1}{x - 5}(0.8488)^5(1 - 0.8488)^{x-5} \\ &= \binom{x-1}{x-5}(0.8488)^5(1 - 0.8488)^{x-5} \\ &= \binom{x-1}{4}(0.8488)^5(0.1512)^{x-5} \quad x = 5, 6, 7, \ldots \end{aligned}$$

(d) This is conditional probability. Let $A = \{6 \text{ ladybus}\}$ and $B = \{\text{more than 3 ladybugs}\}$. Then

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(\text{6 lady bugs})}{P(\text{more than 3 ladybugs})} = \frac{\frac{6^6 e^{-6}}{6!}}{0.8488} = 0.1892$$

(e) Again we need to use conditional probability.

$$\begin{aligned} P(x \text{ on } 1^{\text{st}} \text{ tree}|\text{total of } t) &= \frac{P(x \text{ on } 1^{\text{st}} \text{ tree and total of } t)}{P\left(\text{total of } t\right)} \\ &= \frac{P(x \text{ on } 1^{\text{st}} \text{ tree and } t - x \text{ on } 2^{\text{nd}} \text{ tree})}{P\left(\text{total of } t\right)} \\ &= \frac{P(x \text{ on } 1^{\text{st}} \text{ tree})P(t - x \text{ on } 2^{\text{nd}} \text{ tree})}{P\left(\text{total of } t\right)} \end{aligned}$$

Use the Poisson distribution to calculate each, with $\mu = 6 \times 2 = 12$ in the denominator since there are 2 trees.

$$\begin{aligned} P(x \text{ on } 1^{\text{st}} \text{ tree} \mid \text{total of } t) &= \frac{\left(\frac{6^x e^{-6}}{x!}\right)\left(\frac{6^{t-x} e^{-6}}{(t-x)!}\right)}{\frac{12^t e^{-12}}{t!}} \\ &= \frac{t!}{x!(t-x)!}\left(\frac{6}{12}\right)^x \left(\frac{6}{12}\right)^{t-x} \\ &= \binom{t}{x}\left(\frac{1}{2}\right)^x \left(1-\frac{1}{2}\right)^{t-x} \quad \text{for } x = 0, 1, \ldots, t \end{aligned}$$

Caution: Don't forget to state the range of X. If the total is t, there couldn't be more than t ladybugs on the 1^{st} tree.

Exercise: The answer to (e) is a Binomial probability function. Can you reach this answer by general reasoning rather than using conditional probability to derive it?

Problems

5.9.1 In a Poisson process the average number of occurrences is λ per minute. Independent 1 minute intervals are observed until the first minute with no occurrences is found. Let X be the number of 1 minute intervals required, including the last one. Find the probability function, $f(x)$.

5.9.2 Calls arrive at a telephone distress centre during the evening according to the conditions for a Poisson process. On average there are 1.25 calls per hour.

(a) Find the probability there are no calls during a 3 hour shift.

(b) Give an expression for the probability a person who starts working at this centre will have the first shift with no calls on the fifteenth shift.

(c) A person works one hundred 3 hour evening shifts during the year. Give an expression for the probability there are no calls on at least 4 of these 100 shifts. Calculate a numerical answer using a Poisson approximation.

5.10 Summary of Probability Functions for Discrete Random Variables

Name	Probability Function
Discrete Uniform	$f(x) = \frac{1}{b-a+1}$ for $x = a, a+1, a+2, \ldots, b;\ b > a$
Hypergeometric	$f(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$ for $x = \max(0, n-(N-r)), \ldots, \min(n, r)$
Binomial	$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$ for $x = 0, 1, 2, \ldots, n;\ 0 < p < 1$
Negative Binomial	$f(x) = \binom{x+k-1}{x} p^k (1-p)^x$ for $x = 0, 1, 2, \ldots;\ 0 < p < 1$
Geometric	$f(x) = p(1-p)^x$ for $x = 0, 1, 2, \ldots;\ 0 < p < 1$
Poisson	$f(x) = \frac{e^{-\mu}\mu^x}{x!}$ for $x = 0, 1, 2, \ldots;\ \mu > 0$

Notes:

1. If $p = 0$ for the Binomial distribution then we obtain the probability function $f(x) = 1$ for $x = 0$ and $f(x) = 0$ otherwise. Since all the probability is assigned to one value $x = 0$ the random variable X is said to be a degenerate random variable. If $p = 1$ then $f(x) = 1$ for $x = n$ and $f(x) = 0$ otherwise which is also a degenerate distribution. The values $p = 0$ and $p = 1$ are included as possible parameter values for the Binomial distribution in the Table of Distributions in Chapter 15.

2. If $p = 1$ for the Negative Binomial distribution then we obtain the probability function $f(x) = 1$ for $x = k$ and $f(x) = 0$ otherwise which is also a degenerate distribution. The value $p = 1$ is included as a possible parameter value for the Negative Binomial and Geometric distributions in the Table of Distributions in Chapter 15.

3. If $\mu = 0$ for the Poisson Distribution then we obtain the probability function $f(x) = 1$ for $x = 0$ and $f(x) = 0$ otherwise which is also a degenerate distribution. The value $\mu = 0$ is included as a possible parameter value for the Poisson distribution in the Table of Distributions in Chapter 15.

5.11 Chapter 5 Problems

1. The random variable X has probability function given by

x	0	1	2	3	4
$f(x)$	$0.1c$	$0.2c$	$0.5c$	c	$0.2c$

   (a) Find c and $P(X > 2)$.

   (b) Find $F(x) = P(X \leq x)$, the cumulative distribution function for X.

2. The range of the random variable X is $A = \{1, 2, 3, 4, 5\}$. For $x \in A$ the cumulative distribution function for X is given by

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

   (a) Find k and $P(2 < X \leq 4)$.

   (b) Find $f(x) = P(X = x)$, the probability function for X. Draw a probability histogram for $f(x)$.

3. The range of the random variable X is $A = \{0, 1, 2, \ldots\}$. For $x = 0, 1, 2, \ldots$, the cumulative distribution function of X is given by

$$F(x) = P(X \leq x) = 1 - 2^{-x}$$

   (a) Find $P(X = 5)$ and $P(X \geq 5)$.

   (b) Find $f(x) = P(X = x)$, the probability function of X.

4. Two balls are drawn at random from a box containing ten balls numbered $0, 1, \ldots, 9$. Let the random variable X be the maximum of the two numbers drawn and let the random variable Y be the total of the two numbers drawn.

   (a) If sampling is done **without** replacement, determine

      (i) the probability function of X

      (ii) the probability function of Y.

   (b) If sampling is done **with** replacement, determine

      (i) the probability function of X

      (ii) the probability function of Y.

5. Suppose X is a discrete random variable with probability function

$$f(x) = P(X = x) = p(1-p)^x, \ \ 0 < p < 1, \ \text{for} \ \ x = 0, 1, 2, \ldots.$$

   (a) Verify that $\sum\limits_{x=0}^{\infty} f(x) = 1$.

   (b) Find $P(X < x)$ for $x = 0, 1, \ldots$.

   (c) Find the probability that X is an odd number.

   (d) Find the probability that X is divisible by 3.

   (e) Find the probability function of the random variable R, where R is the remainder when X is divided by 4.

6. In a box of 1000 computer chips, 5% are defective. Twenty computer chips are drawn at random without replacement and tested for defects. Let the random variable X be the number of defective computer chips found.

   (a) Give the probability function for X.

   (b) Give an expression for the probability that at least two chips are defective.

   (c) Approximate the probability in (b) using a suitable approximation. Justify the approximation.

7. **Jury selection:** 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 there were 74 men and 76 women present for jury selection. Twelve people are chosen at random without replacement for a jury of 12 people. Let Y be the number of men chosen. Give an expression for $P(Y = y)$.

   (c) For the trial in part (b), the number of men selected turned out to be two. Find $P(Y \leq 2)$. What might you conclude from this?

8. A string of zeros and ones of length 10^4 is sent over a network. Suppose each bit has probability 10^{-5} of being corrupted, independently for each bit.

   (a) Give an expression for the probability that no bits are corrupted.

   (b) Give an expression for the probability that at most one bit is corrupted.

   (c) Approximate the probabilities in (a) and (b) using a suitable approximation. Justify the approximation.

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 10 fill-ups during the contest, give an expression for the probability that he or she wins at least one prize. Approximate this probability using a suitable approximation. Justify the approximation.

   (b) If a particular gas bar distributes 2000 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. Justify the approximation in each case.

10. Let $f(x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$. We want to determine $\lim\limits_{N\to\infty} f(x)$ such that $p = \frac{r}{N}$ is held fixed.

    (a) Use $\binom{a}{x} = \frac{a^{(x)}}{x!}$ and $r = pN$ to show that

    $$\frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} = \binom{n}{x}\frac{(pN)^{(x)}\left[(1-p)N\right]^{(n-x)}}{N^{(x)}(N-x)^{(n-x)}}$$

    (b) Show that

    $$\lim_{N\to\infty}\binom{n}{x}\frac{(pN)^{(x)}\left[(1-p)N\right]^{(n-x)}}{N^{(x)}(N-x)^{(n-x)}} = \binom{n}{x}p^x(1-p)^{n-x}$$

    (c) What is the importance of the result in (b)?

11. A bin at a hardware store contains 35 forty watt lightbulbs and 70 sixty watt bulbs. A customer wants to buy 8 sixty watt bulbs, and withdraws bulbs without replacement until these 8 bulbs have been found. Let the random variable X be the number of 40 watt bulbs drawn from the bin. Find the probability function of X.

12. Todd buys a lottery ticket every week. Suppose 1% of the tickets win some prize.

    (a) Todd buys one ticket every week for the next 20 weeks. What is the probability he bought no winning tickets? at least one winning ticket?

    (b) What is the probability he does not get a winning ticket in the first 30 weeks?

    (c) Todd realizes he is spending too much money on lottery tickets so he decides he will continue buying tickets only until he has bought 4 winning tickets. Find an expression for the probability he will have to buy a ticket every week for at least the next 100 weeks in order to achieve his goal.

13. A coffee chain claims that you have a 1 in 9 chance of winning a prize on their "roll up the edge" promotion, where you roll up the edge of your paper cup to see if you win.

    (a) What is the probability that you have no winners in a one week period in which you bought fifteen cups of coffee?

    (b) What is the probability that you get your first win when you buy your twenty-fifth cup of coffee?

    (c) Over the last week of a month long promotion you and your friends bought 60 cups of coffee, but there was only one winner. Find the probability that there would be one or fewer winners. What would you conclude?

14. Suppose $X \sim Geometric\,(p)$.

    (a) Find an expression for $P(X \geq x)$, and show that $P(X \geq s+t|X \geq s) = P(X \geq t)$ for all non-negative integers s, t. Explain why this property of the Geometric distribution is called the "memoryless" property.

    (b) What is the most probable value of X?

15. Requests to a web server are assumed to follow a Poisson process. On average there are two requests per second.

    (a) Discuss briefly whether or not you think the three assumptions for a Poisson process would hold reasonably well in this situation.

    (b) Find the probability of three or more requests in a one second interval.

    (c) Given an expression for the probability of more than 125 requests in a one minute interval.

16. A waste disposal company averages 6.5 spills of toxic waste per month. Assume spills occur randomly at a uniform rate, and independently of each other, with a negligible chance of two or more occurring at the same time. Find the probability there are four or more spills in a two month period.

17. Coliform bacteria are distributed randomly and uniformly throughout river water at the average concentration of one per 20 cubic centimeters of water.

    (a) What is the probability of finding exactly 2 coliform bacteria in a 10 cubic centimeters sample of the river water?

    (b) What is the probability of finding at least 1 coliform bacterium in a 1 cubic centimeter sample of the river water?

    (c) In testing for the concentration (average number per unit volume) of bacteria it is possible to determine cheaply whether a sample has **any** bacteria present or not. Suppose the average concentration of bacteria in a body of water is λ per cubic centimeter. If 10 independent water samples of 10 cubic centimeters each are tested, let the random variable Y be the number of samples with **no** bacteria. Find $P(Y = y)$.

    (d) Suppose that in 10 independent samples, there were exactly 3 samples with no bacteria. Give an estimate for the value of λ.

18. In a group of policy holders for house insurance, the average number of claims per 100 policies per year is $\lambda = 8.0$. The number of claims for an individual policy holder is assumed to follow a Poisson distribution.

    (a) In a given year, what is the probability an individual policy holder has at least 1 claim?

    (b) In a group of 20 policy holders, what is the probability there are no claims in a given year? What is the probability there are 2 or more claims?

19. Assume power failures occur independently of each other at a uniform rate through the months of the year, with little chance of 2 or more occurring simultaneously. Suppose that 80% of months have no power failures.

    (a) Seven months are picked at random. What is the probability that 5 of these months have no power failures?

    (b) Months are picked at random until 5 months without power failures have been found. What is the probability that 7 months will have to be picked?

    (c) What is the probability a month has more than one power failure?

20. Spruce budworms are distributed through a forest according to a Poisson process so that the average is λ per hectare.

    (a) Give an expression for the probability that at least 1 of n one hectare plots contains at least k spruce budworms.

    (b) Discuss briefly which assumption(s) for a Poisson process may not be well satisfied in this situation.

21. A person working in telephone sales has a 20% chance of making a sale on each call, with calls being independent. Assume calls are made at a uniform rate, with the numbers made in non-overlapping periods being independent. On average there are 20 calls made per hour.

    (a) Find the probability there are 2 sales in 5 calls.

    (b) Find the probability exactly 8 calls are needed to make 2 sales.

    (c) If 8 calls were needed to make 2 sales, what is the probability there was 1 sale in the first 3 of these calls?

    (d) Find the probability of 3 calls being made in a 15 minute period.

22. During rush hour the number of cars passing through a particular intersection[22] is assumed to follow a Poisson process. On average there are 540 cars per hour.

    (a) Discuss briefly whether or not you think the three assumptions for a Poisson process would hold reasonably well in this situation.

    (b) Find the probability that 11 cars passed through the intersection in a thirty second interval.

    (c) Find the probability that 11 or more cars passed through the intersection in a thirty second interval.

    (d) Find the probability that when 20 disjoint thirty second intervals are studied, exactly 2 of them had 11 cars pass through the intersection.

    (e) We want to find 12 disjoint thirty second intervals in which 11 cars passed through the intersection.

        (i) Give an exact expression for the probability that 1000 disjoint 30 second intervals have to be observed to find the 12 having the desired traffic flow.

        (ii) Use an appropriate approximation to evaluate this probability and justify why this approximation is suitable.

[22] "Traffic signals in New York are just rough guidelines." David Letterman (1947 -)

23. Bubbles are distributed in sheets of glass, as a Poisson process, at an intensity of $\lambda = 1.2$ bubbles per square metre. Sheets of glass, each of area $0.8m^2$, are manufactured.

    (a) What is the probability a sheet of glass has no bubbles?

    (b) What is the probability a sheet of glass has more than one bubble?

    (c) Let X be the number of sheets of glass, in a shipment of n sheets, which have no bubbles. What is the probability function of the random variable X?

    (d) In a shipment of 100 sheets, what is the probability more than ten sheets have more than one bubble?

    (e) If the glass manufacturer wants to have at least 50% of the sheets of glass with no bubbles, how small should the intensity λ be to achieve this?

    (f) If the glass manufacturer wants to ensure that 95% of all sheets manufactured have fewer than two bubbles, how small should the intensity λ be to achieve this?

24. **Polls and Surveys**: Polls or surveys in which people are selected and their opinions or other characteristics are determined are very widely used. For example, in a survey on cigarette use among teenage girls, we might select a random sample of n girls from the population in question, and determine the number X who are regular smokers. If p is the fraction of girls who smoke, then $X \sim Binomial(n, p)$. Since p is unknown (that is why we do the survey) we then estimate it as $\hat{p} = X/n$. (In statistics a “hat” is used to denote an estimate of a model parameter based on data.) The $Binomial$ distribution can be used to study how “good” such estimates are, as follows:

    (a) Suppose $p = 0.3$ and $n = 100$. Find the probability $P\left(0.27 \le \frac{X}{n} \le 0.33\right)$. Many surveys try to get an estimate X/n which is within 3% (0.03) of p with high probability. What would you conclude here?

    (b) Repeat the calculation in (a) if $n = 400$ and $n = 1000$. What do you conclude?

    (c) If $p = 0.5$ instead of 0.3, find $P\left(0.47 \le \frac{X}{n} \le 0.53\right)$ when $n = 400$ and 1000.

    (d) Your employer asks you to design a survey to estimate the fraction p of persons age 25-34 who download music via the internet. The objective is to get an estimate accurate to within 3%, with probability close to 0.95. What size of sample n would you recommend?

25. **Telephone surveys**: In some "random digit dialing" surveys, a computer phones randomly selected telephone numbers. However, not all numbers are "active" (belong to a telephone account) and some numbers belong to businesses.

    Suppose that for a given large set of telephone numbers, 57% are active residential or individual numbers. We will call these "personal" numbers.

    Suppose that we wish to interview (over the phone) 1000 persons in a survey.

    (a) Suppose that the probability a call to a personal number is answered is 0.8, and that the probability the person answering agrees to be interviewed is 0.7. Give the probability distribution for X, the number of calls needed to obtain 1000 interviews.

    (b) Use the statistical software R (see Chapter 6) to find $P(X \leq x)$ for $x = 2900, 3000, 3100, 3200$.

    (c) Suppose instead that 3200 randomly selected numbers were dialed. Give the probability function for Y, the number of interviews obtained, and find $P(Y \geq 1000)$.

26. **Hash tables continued:** See Chapter 3, Problem 15. When a hash function is used to create a data structure for a dictionary a collision can occur when a key-value pair is mapped to a slot which has already been assigned to another key-value pair. One strategy for handling collisions is to create a *linked list* of all the key-value pairs which are mapped to the same slot. This is called *collision resolution by separate chaining*. Suppose separate chaining is used with a hash table of size M and that the slot for key k is chosen by randomly selecting a number with replacement from the set $\{0, 1, \ldots, M-1\}$.

    (a) For n keys show that the probability that a given list contains exactly x keys is equal to

    $$\binom{n}{x}\left(\frac{1}{M}\right)^x\left(1-\frac{1}{M}\right)^{n-x} \quad \text{for } x = 0, 1, \ldots, n$$

    (b) Let $\alpha = n/M$ (called the load factor). Under what conditions can the probability in (a) be approximated by the Poisson probability

    $$\frac{\alpha^x e^{-\alpha}}{x!}$$

    (c) If $X \sim Poisson\,(\alpha)$ then the Chernoff bound for tail probabilities gives

    $$P\,(X \geq x) \leq e^{-\alpha}\left(\frac{\alpha e}{x}\right)^x$$

    if $x \geq \alpha$. If $\alpha = 10$, then use this inequality to bound the probability that a list has

    (i) 15 or more keys (ii) 20 or more keys

    What are the implications of these results?

27. The ALOHA protocol for sending messages over wireless connections works as follows:

    Messages of length t are sent by multiple users, without checking if the frequency is busy.
    If two messages are sent during overlapping time intervals, both messages fail.
    If a message fails, the user waits a random amount of time and then tries again.

    Suppose messages are sent according to a Poisson process with a constant rate of λ messages per t units of time.

    (a) Discuss briefly whether or not you think the three assumptions for a Poisson process would hold reasonably well in this situation.

    (b) Find the probability that 3 or more users send a message within t units of time, if $\lambda = 0.75$.

    (c) Find the probability a message sent at time x succeeds if $\lambda = 0.75$. **Hint**: For this event to happen, there must be *no* other messages sent between $x - t$ and $x + t$.

    (d) If $\lambda = 0.75$ and a message has just been sent, find the probability of waiting at least $3t$ units of time until the next message is sent.

    (e) Slotted ALOHA is an updated protocol for sending messages where discrete timeslots of length t are set up, and messages can only be sent at the beginning. If $\lambda = 0.75$, find the probability a timeslot has a message successfully sent in it. Find the probability that of 10 slots, exactly 7 have successful messages sent.

28. **Error Correcting Codes 1:** A message consisting of a string of zeros and ones is sent over a network. Due to interference in the network a bit can randomly "flip" from 0 to 1 or vice versa during transmission. An error correcting code (ECC) is a process of adding redundant data to the message by the transmitter which allows error detection or correction by the receiver. The Triple Repetition Code (TRC) is an ECC for which each bit is sent three times. For example the message 0110 is sent as 000111111000. The receiver looks at the received message in groups of three and decodes each group to the bit that occurs most often in the group (called the majority rule). For example, 000, 001, 010, 100 are decoded as 0 while 111, 110, 101, 011 are decoded as 1. TRC allows the correction of at most one error in each group of three bits. Suppose each bit has probability p of flipping, independently for each bit.

    (a) What is the probability a group of three repeated bits will be decoded correctly?

    (b) What is the probability an original message of length four (e.g. 0110) is decoded correctly if no ECC is used?

    (c) What is the probability an original message of length four (e.g. 0110) is decoded correctly if TRC is used?

(d) Compare the probabilities in (b) and (c) for $p = 0.2$, 0.1, 0.05 and 0.01.

(e) Suppose each bit is repeated five times and the majority rule is used. What is the probability a group of five repeated bits will be decoded correctly?

(f) Suppose each bit is repeated k times, where k is an odd number, and the majority rule is used. Let $P(k,p)$ be the probability a group of k repeated bits will be decoded correctly. Find an expression for $P(k,p)$. Calculate $P(k,p)$ for $p = 0.2$, 0.1, and 0.05 and $k = 3, 5, 7, 9$. What do you notice?

29. **Error Correcting Codes 2:** Hamming(7,4) is another type of ECC in which three parity bits are added to a four bit string so a total of seven bits are transmitted. The three parity bits can be used to correct at most one error in the string of seven bits received. Suppose each bit has probability p of flipping, independently for each bit.

    (a) What is the probability a correctable message is received using Hamming(7,4) (at most one bit is flipped)?

    (b) Calculate the probability in (a) for $p = 0.2$, 0.1, 0.05 and 0.01.

    (c) Compare the probabilities in (b) with the probabilities obtained in Problem 28(d). What do you notice?

30. **Challenge problem:** Suppose that n independent tosses of a coin with $P(\text{Head}) = p$ are made. Show that the probability of an even number of heads is given by $\frac{1}{2}[1+(q-p)^n]$ where $q = 1-p$.

6. COMPUTATIONAL METHODS AND THE STATISTICAL SOFTWARE R

One of the giant steps towards democracy in the last century was the increased democratization of knowledge[23], facilitated by the personal computer, *Wikipedia* and the advent of free open-source (GNU) software such as *Linux*. The statistical software package R implements a dialect of the S language that was developed at AT&T Bell Laboratories by Rick Becker, John Chambers and Allan Wilks. Versions of R are available, at no cost, for 32-bit versions of Microsoft Windows for Linux, for Unix and for Macintosh systems. It is available through the Comprehensive R Archive Network (CRAN) (downloadable for unix, windows or MAC platforms at *http://cran.r-project.org/*). This means that a community of interested statisticians voluntarily maintain and updates the software. Like the licensed software *Matlab* and *Splus*, R permits easy matrix and numerical calculations, as well as a programming environment for high-level computations. The R software also provides a powerful tool for handling probability distributions, generating random variables, and graphical display. Because it is freely available and used by statisticians world-wide, high level programs in R are often available on the web. These notes provide a glimpse of a few of the features of R. Web resources have much more information and more links can be found on the Stat 230 web page. We will provide a brief description of commands on a windows machine here, but the MAC and UNIX commands will generally be similar once R is started.

6.1 Preliminaries

Begin by installing R on your personal computer and then invoke it on Math Unix machines by typing R or on a windows machine by clicking on the R icon. For these notes, we will simply describe typing commands into the R command window following the R prompt ">" in interactive mode.

Objects include variables, functions, vectors, arrays, lists and other items. To see online documenta-

[23] "Knowledge is the most deomocratic source of power." Alvin Toffler

tion about something, we use the "help" function. For example, to see documentation on the function mean(), type

```
help(mean).
```

In some cases help.search() is helpful. For example

```
help.search("matrix")
```

lists all functions whose help pages have a title or alias in which the text string "matrix" appears.

The <- is a left diamond bracket (<) followed by a minus sign (-). It means "is assigned to", for example,

```
x <- 15
```

assigns the value 15 to variable x. To quit an R session, type

```
q()
```

You need the brackets () because you wish to run the function "q". Typing q on its own, without the parentheses, displays the text of the function on the screen. Try it! Alternatively to quit R, you can click on the "File" menu and then on Exit or on the x in the top right corner of the R window. You are asked whether you want to save the workspace image. Clicking "Yes" (safer) will save all the objects that remain in the workspace both those at the start of the session and those added.

6.2 Vectors

Vectors can consist of either numbers or other symbols like characters; we will consider only numbers here. Vectors are defined using c(): for example,

```
x<-c(1,3,5,7,9)
```

defines a vector of length 5 with the elements given. Vectors and other classes of objects possess certain attributes. For example, typing

```
length(x)
```

will give the length of the vector x. Vectors are a convenient way to store values of a function (e.g. a probability function or a c.d.f) or values of a random variable that have been recorded in some experiment or process. We can also read a table of values from a text file that we created earlier called say "mydata.txt" on a disk in drive c:

```
> mydata <- read.table("c:/mydata.txt", header=TRUE)
```

Use of “header=TRUE” causes R to use the first line of the text file to get header information for the columns. If column headings are not included in the file, the argument can be omitted and we obtain a table with just the data. The R object “mydata” is a special form known as a “data frame”. Data frames that consist entirely of numeric data have a structure that is similar to that of numeric matrices. The names of the columns can be displayed with the command

```
> names(mydata)
```

6.3 Arithmetic Operations

The following R commands and responses should explain the most basic arithmetic operations.

```
> 7+3
[1] 10
> 7*3
[1] 21
> 7/3
[1] 2.333333
> 2^3
[1] 8
```

In the last example the result is 8. The [1] says basically “first requested element follows” but here there is just one element. The “>” indicates that R is ready for another command.

6.4 Some Basic Functions

Functions of many types exist in R. Many operate on vectors in a transparent way, as do arithmetic operations. For example, if x and y are vectors then x+y adds the vectors element-wise; thus x and y must be the same length. Some examples, with comments, follow. Note that anything that follows a # on the command line is taken as comment and ignored by R.

```
> x<- c(1,3,5,7,9)   # Defines a vector x
> x                  # displays x
[1] 1 3 5 7 9
> y<- seq(1,2,.25) # defines vector whose elements are an
                   # arithmetic progression
> y
```

```
[1] 1.00 1.25 1.50 1.75 2.00
> y[2]             #displays the second element of vector y
[1] 1.25
> y[c(2,3)]        #displays 2nd and 3rd elements of vector y
[1] 1.25 1.50
> mean(x)          #computes mean of the elements of vector x
[1] 5
> summary(x)    #function which summarizes features of a vector x
 Min.    1st Qu. Median  Mean 3rd Qu. Max.
  1       3        5      5     7     9
> var(x)    # Computes the (sample) variance of the elements of x
[1] 10
> exp(1)    # The exponential function
[1] 2.718282
> exp(y)
[1] 2.718282 3.490343 4.481689 5.754603 7.389056
# round(y,n) rounds elements of vector y to n decimals
> round(exp(y),2)
[1] 2.72 3.49 4.48 5.75 7.39
> x+2*y
[1]  3.0  5.5  8.0 10.5 13.0
```

6.5 R Objects

Type “ls()” to see a list of names of all objects, including functions and data structures, in your workspace.

If you type the name of an object, vector, matrix or function, you are returned its contents. (Try typing “q” or “mean”).

Before you quit, you may remove objects that you no longer require with “rm()” and then save the workspace image. The workspace image is automatically loaded when you restart R in that directory.

6.6 Graphs

To open a graphics window in Unix, type x11(). Note that in R, a graphics window opens automatically when a graphical function is used.

There are various plotting and graphical functions. Two useful ones are

```
plot(x,y)  # Gives a scatterplot of x versus y; thus x and y must be
           #vectors of the same length.

hist(x)    # Creates a frequency histogram based on the values in the
           #vector x. To get a relative frequency histogram (areas of
           #rectangles sum to one) use
hist(x,prob=T).
```

Graphs can be tailored with respect to axis labels, titles, numbers of plots to a page etc. Type help(plot), help(hist) or help(par) for some information. Try

```
x<-(0:20)*pi/10
plot(x, sin(x))
```

Is it obvious that these points lie on a sine curve? One can make it more obvious by changing the shape of the graph. Place the cursor over the lower border of the graph sheet, until it becomes a double-sided and then drag the border in towards the top border, to make the graph sheet short and wide.
To save/print a graph in R using UNIX, you generate the graph you would like to save/print in R using a graphing function like plot() and type:

```
dev.print(device,file="filename")
```

where device is the device you would like to save the graph to and filename is the name of the file that you would like the graph saved to. To look at a list of the different graphics devices you can save to,

```
type help(Devices).\newpage
```

To save/print a graph in R using Windows, you can do one of two things.

a) You can go to the File menu when the graph window is active and save the graph using one of several formats (that is, postscript, jpeg, etc.) or print it. You may also copy the graph to the clipboard using one of the formats and then paste to an editor, such as MS Word.

b) You can right click on the graph. This gives you a choice of copying the graph and then pasting to an editor, such as MS Word, or saving the graph as a metafile or bitmap or print directly to a printer.

6.7 Distributions

There are functions which compute values of probability or probability density functions, cumulative distribution functions, and quantiles for various distributions. It is also possible to generate (pseudo) random samples from these distributions. Some examples follow for Binomial and Poisson distributions. For other distribution information, type

```
help(rhyper),
help(rnbinom)
```

and so on. Note that R does not have any function specifically designed to generate random samples from a discrete Uniform distribution (although there is one for a continuous Uniform distribution). To generate n random samples from a discrete $Uniform(a,b)$, use

```
sample(a:b,n,replace=T).\smallskip
> y<- rbinom(10,100,0.25)    # Generate 10 random values from the
                             # Binomial(100,0.25) distribution
                             # The values are  stored in the vector y.
> y                          # Display the values
 [1] 24 24 26 18 29 29 33 28 28 28
> pbinom(3,10,0.5)       # Compute P(Y<=3) for a Binomial(10,0.5) r.v.
[1] 0.171875
> qbinom(.95,10,0.5)        # Find the 0.95 quantile (95th percentile)
                            # for Binomial(10,0.5) distribution
[1] 8
> z<- rpois(10,10)           # Generate 10 random values from the
                             # Poisson(10) distribution.
                             # The values are stored in the vector z.
> z                          # Display the values
 [1]  6  5 12 10  9  7  9 12  5  9
> ppois(3,10)                # Compute P(Y<=3) for a Poisson(10) r.v.
[1] 0.01033605
> qpois(.95,10)              # Find the 0.95 quantile (95th percentile)
                             # for Poisson(10) distribution
[1] 15\smallskip
```

To illustrate how to plot the probability function for a random variable, a $Binomial(10, 0.5)$ random variable is used.

```
# Assign all possible values of the Binomial(10,0.5) r.v.
x <- seq(0,10,by=1)
# Determine the value of the p.f. for possible values of X
x.pf <- dbinom(x,10,0.5)
# Plot the probability function
barplot(x.pf,xlab="X",ylab="Probability Function",
names.arg=c("0","1","2","3","4","5","6","7","8","9","10"))\smallskip
```

Loops in R are easy to construct but long loops can be slow and should be avoided where possible. For example

```
x=0
for (i in 1:10) x<- c(x,i)
```

can be replaced by

```
x=c(0:10)
```

Commonly used functions.

```
print()      # Prints a single R object
cat()        # Prints multiple objects, one after the other
length()     # Number of elements in a vector or of a list
mean()       # mean of a vector of data
median()     # median of a vector of data
range()      # Range of values of a vector of data
unique()     # Gives the vector of distinct values
diff()       # the vector of first differences so diff(x) has
             # one less element than x
sort()       # Sort elements into order, omitting NAs
order()      # x[order(x)] orders elements of x, with NAs last
cumsum()     # vector of partial or cumuulative sums
cumprod()    # vector of partial or cumuulative products
rev()        # reverse the order of vector element
```

6.8 Chapter 6 Problems

1. Input the following data, on damage that had occurred in space shuttle launches prior to the Challenger space shuttle launch of January 28, 1986.

Date	**Temperature (Frahenheit)**	**Number of Damage Incidents**
4/12/81	66	0
11/12/81	70	1
3/22/82	69	0
6/27/82	80	NA
1/11/82	68	0
4/4/83	67	0
6/18/83	72	0
8/30/83	73	0
11/28/83	70	0
2/3/84	57	1
4/6/84	63	1
8/30/84	70	1
10/5/84	78	0
11/8/84	67	0
1/24/85	53	3
4/12/85	67	0
4/29/85	75	0
6/17/85	70	0
7/29/85	81	0
8/27/85	76	0
10/3/85	79	0
10/30/85	75	2
11/26/85	76	0
1/12/86	58	1

This was then followed by the disasterous Challenger incident on 1/28/86.

(a) Enter the temperature data into a data frame, with (for example) column names: date, temperature, damage.

(b) Plot total incidents against temperature. Do you see any relationship? On the date of the challenger incident the temperature at launch was 31 degrees F. What would you expect for the number of damage incidents?

2. The following ten observations, taken during the years $1970 - 79$, are on October snow cover for Eurasia. (Snow cover is in millions of square kilometers).

Year	Snow.cover
1970	6.5
1971	12
1972	14.9
1973	10
1974	10.7
1975	7.9
1976	21.9
1977	12.5
1978	14.5
1979	9.2

(a) Enter the data into R. To save keystrokes, enter the successive years as 1970:1979

(b) Plot snow.cover versus year.

(c) Use "hist()" to plot a histogram of the snow cover values.

(d) Repeat (b) and (c) after taking logarithms of snow cover.

3. **Benford's Law** Simon Newcomb, born in Wallace, Nova Scotia, discovered "Benford's" law in 1881. He observed that the first digit "1" in naturally occuring datasets is more common than the others and occurs about 30% of the time and about 17.5% of the time, it is the digit 2. Of course if all leading numbers (1 through 9) had equal probability, each would occur $\frac{1}{9}$ of the time. Benford's law states that the probability $f(x)$ that the first digit in a set of numbers is x, $x = 1, 2, \ldots, 9$ is proportional to $\log_{10}(1 + \frac{1}{x})$, or, in terms of the natural logarithm,

$$f(x) = \frac{\ln(1 + \frac{1}{x})}{\ln(10)} \quad \text{for } x = 1, 2, ...9$$

Consider, for example, the first $n = 20$ Fibonacci numbers defined by

$$F_n = F_{n-1} + F_{n-2}, \quad F_1 = F_2 = 1$$

These are

$$1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377, 610, 987, 1597, 2584, 4181, 6765$$

Notice that 5 or 25% of these numbers have their first digit equal to 1. Use R to verify numerically that as n increases the proportion of Fibonacci numbers which have their first digit equal to 1 approaches $\frac{\ln(2)}{\ln(10)}$. Use R to determine the proportions for $x = 2, 3, \ldots, 9$. Repeat this experiment for the sequence 2^x, $x = 1, 2, \ldots, n$. (**Note:** Because it is a common property of naturally occuring data, Benford's law has been used in many areas, including fraud detection, where artificially changing the data or a digital image will usually result in a failure of Benford's law.)

7. EXPECTED VALUE AND VARIANCE

7.1 Summarizing Data on Random Variables

When we return midterm tests, someone almost always asks what the average was. While we could list out all marks to give a picture of how students performed, this would be tedious. It would also give more detail than could be immediately digested. 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 giving full details of every outcome.

To illustrate some of the ideas involved, suppose we were to observe cars crossing a toll bridge, and record the number, X, of people in each car. Suppose in a small study[24] data on 25 cars were collected. We could list out all 25 numbers observed, but a more helpful way of presenting the data would be in terms of the **frequency distribution** below, which gives the number of times (the "frequency") each value of X occurred.

X	Frequency Count	Frequency
1	𝍸𝍷	6
2	𝍸𝍷𝍷𝍷	8
3	𝍸	5
4	𝍷𝍷𝍷	3
5	𝍷𝍷	2
6	𝍷	1

We could also draw a *frequency* histogram of these frequencies (see Figure 7.1).

Frequency distributions or histograms are good summaries of data because they show the variability in the observed outcomes very clearly. Sometimes, however, we might prefer a single-number summary. The most common such summary is the average, or arithmetic mean of the outcomes. The mean

[24] "Study without desire spoils the memory, and it retains nothing that it takes in." Leonardo da Vinci

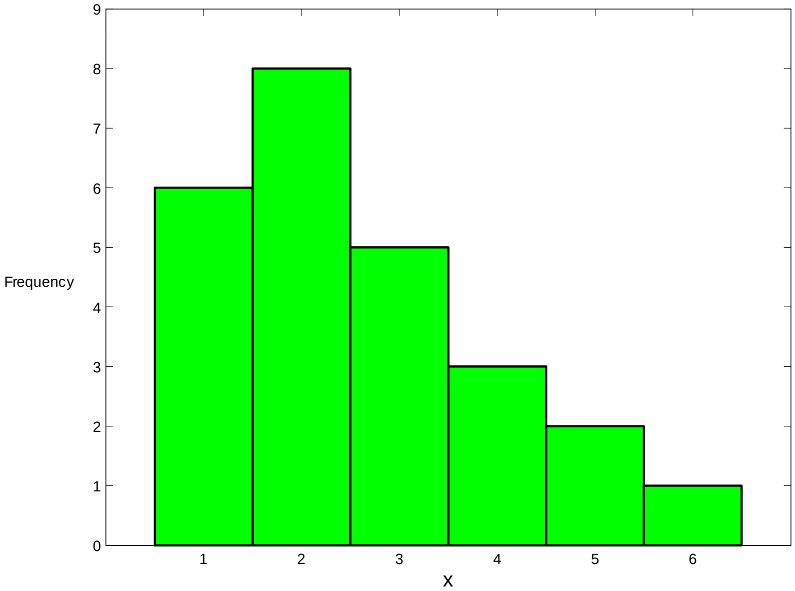


Figure 7.1: Frequency histogram for number of people in a car at a toll bridge

of n outcomes $x_1, \ldots, x_n$ for a random variable X is $\sum_{i=1}^{n} x_i/n$, and is denoted by $\bar{x}$. The arithmetic mean for the example above can be calculated as

$$\frac{(6 \times 1) + (8 \times 2) + (5 \times 3) + (3 \times 4) + (2 \times 5) + (1 \times 6)}{25} = \frac{65}{25} = 2.60$$

That is, there was an average of 2.6 persons per car. A set of observed outcomes $x_1, \ldots, x_n$ for a random variable X is termed a **sample** in probability and statistics. To reflect the fact that this is the average for a particular sample, we refer to it as the **sample mean**. Unless somebody deliberately "cooked" the study, we would not expect to get precisely the same sample mean if we repeated it another time. Note also that $\bar{x}$ is not in general an integer, even though X is.

Two other common summary statistics are the median and mode.

Definition 14 *The* ***median*** *of a sample is a value such that half the results are below it and half above it, when the results are arranged in numerical order.*

If these 25 results were written in order, the thirteenth outcome would be a 2. So the median is 2. By convention, we go half way between the middle two values if there are an even number of observations.

Definition 15 *The* ***mode*** *of the sample is the value which occurs most often. In this case the mode is 2. There is no guarantee there will be only a single mode.*

Exercise: Give a data set with a total of 11 values for which the median$<$mode$<$mean.

7.2 Expectation of a Random Variable

The statistics in the preceding section summarize features of a sample of observed X-values. The same idea can be used to summarize the probability distribution of a random variable X. To illustrate, consider the previous example, where X is the number of persons in a randomly selected car crossing a toll bridge. Note that we can re-arrange the expression used to calculate $\overline{x}$ for the sample, as

$$\begin{aligned}
&\frac{(6\times 1)+(8\times 2)+(5\times 3)+(3\times 4)+(2\times 5)+(1\times 6)}{25}\\
&=(1)\left(\frac{6}{25}\right)+(2)\left(\frac{8}{25}\right)+(3)\left(\frac{5}{25}\right)+(4)\left(\frac{3}{25}\right)+(5)\left(\frac{2}{25}\right)+(6)\left(\frac{1}{25}\right)\\
&=\sum_{x=1}^{6} x\times \text{fraction of times } x \text{ occurs}
\end{aligned}$$

Now suppose we know that the probability function of X is given by

x	1	2	3	4	5	6
$f(x)$	0.30	0.25	0.20	0.15	0.09	0.01

Using the relative frequency "definition" of probability, if we observed a very large number of cars, the fraction (or relative frequency) of times $X = 1$ would be 0.30, for $X = 2$, this proportion would be 0.25, etc. So, *in theory*, (according to the probability model) we would expect the mean to be

$$(1)(0.30)+(2)(0.25)+(3)(0.20)+(4)(0.15)+(5)(0.09)+(6)(0.01)=2.51$$

if we observed an infinite number of cars. This "theoretical" mean is usually denoted by μ or $E(X)$, and requires us to know the distribution of X. With this background we make the following mathematical definition.

Definition 16 *Let X be a discrete random variable with range$(X) = A$ and probability function $f(x)$. The **expected value** (also called the mean or the expectation) of X is given by*

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

The expected value of X is also often denoted by the Greek letter μ. The expected value [25] of X can be thought of physically as the average of the X-values that would occur in an infinite series of repetitions of the process where X is defined. This value not only describes one aspect of a probability distribution, but is also very important in certain types of applications. For example, if you are playing

[25] Oft expectation fails, and most oft where most it promises; and oft it hits where hope is coldest; and despair most sits. William Shakespeare (1564 - 1616)

a casino game in which X represents the amount you win in a single play, then $E(X)$ represents your average winnings (or losses!) per play.

Sometimes we may not be interested in the average value of X itself, but in some function of X. Consider the toll bridge example once again, and suppose there is a toll which depends on the number of car occupants. For example, a toll of $1 per car plus 25 cents per occupant would produce an average toll for the 25 cars in the study of Section 7.1 equal to

$$(1.25)\left(\frac{6}{25}\right)+(1.50)\left(\frac{8}{25}\right)+(1.75)\left(\frac{5}{25}\right)+(2.00)\left(\frac{3}{25}\right)+(2.25)\left(\frac{2}{25}\right)+(2.50)\left(\frac{1}{25}\right)=\$1.65$$

If X has the theoretical probability function $f(x)$ given above, then the average value of this $\$(0.25X+1)$ toll would be defined in the same way, as,

$$(1.25)(0.30)+(1.50)(0.25)+(1.75)(0.20)+(2.00)(0.15)+(2.25)(0.09)+(2.50)(0.01)=\$1.6275$$

We call this the expected value of $(0.25X+1)$ and write $E\left(0.25X+1\right)=1.6275$.

As a further illustration, suppose a toll designed to encourage car pooling charged $\$12/x^2$ if there were x people in the car. This scheme would yield an average toll, in theory, of

$$\left(\frac{12}{1}\right)(0.30)+\left(\frac{12}{4}\right)(0.25)+\left(\frac{12}{9}\right)(0.20)+\left(\frac{12}{16}\right)(0.15)+\left(\frac{12}{25}\right)(0.09)+\left(\frac{12}{36}\right)(0.01)$$
$$=\$4.7757$$

In other words

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

is the "expected value" of $\left(\frac{12}{X^2}\right)$.

With this as background, we can now make a formal definition.

Theorem 17 *Let X be a discrete random variable with range$(X)=A$ and probability function $f(x)$. The* ***expected value*** *of some function $g(X)$ of X is given by*

$$E\left[g(X)\right]=\sum_{x\in A}g(x)f(x)$$

Proof: To use Definition 16, we need to determine the expected value of the random variable $Y=g(X)$ by first finding the probability function of Y, say $f_Y(y)=P(Y=y)$ and then computing

$$E[g(X)]=E(Y)=\sum_{y\in B}yf_Y(y) \tag{7.1}$$

where range$(Y) = B$. Let $D_y = \{x; g(x) = y\}$ be the set of x values with a given value y for $g(x)$, then

$$f_Y(y) = P\left[g(X) = y\right] = \sum_{x \in D_y} f(x)$$

Substituting this in (7.1) we obtain

$$\begin{aligned} E[g(X)] &= \sum_{y \in B} y f_Y(y) \\ &= \sum_{y \in B} y \sum_{x \in D_y} f(x) \\ &= \sum_{y \in B} \sum_{x \in D_y} g(x) f(x) \\ &= \sum_{x \in A} g(x) f(x) \end{aligned}$$

Notes:

(1) You can interpret $E[g(X)]$ as the average value of $g(X)$ in an infinite series of repetitions of the process where X is defined.

(2) $E\left[g(X)\right]$ is also known as the "expected value" of $g(X)$. This name is somewhat misleading since the average value of $g(X)$ may be a value which $g(X)$ never takes - hence unexpected!

(3) The case where $g(x) = x$ reduces to our earlier definition of $E(X)$.

(4) Confusion sometimes arises because we have two notations for the mean of a probability distribution: μ and $E(X)$ mean the same thing. There is a small advantage to using the (lower case) letter μ. It makes it visually clearer that the expected value is NOT a random variable like X but a non-random constant.

(5) When calculating expectations, look at your answer to be sure it makes sense. Suppose for example that X takes values from 1 to 10. Then since

$$1 = \sum_{x=1}^{10} (1)\, P\left(X = x\right) \le \sum_{x=1}^{10} x P\left(X = x\right) = E\left(X\right) \le \sum_{x=1}^{10} (10)\, P\left(X = x\right) = 10\,(1) = 10$$

you should know you've made an error if you get $E(X) > 10$ or $E(X) < 1$. In physical terms, $E(X)$ is the balance point for the probability histogram of $f(x)$.

Let us note a couple of mathematical properties of expected value that can help to simplify calculations.

Linearity Properties of Expectation: If your linear algebra is good, it may help if you think of E as being a linear operator, and this may save memorizing these properties.

1. For constants a and b,
$$E\left[ag(X)+b\right] = aE\left[g(X)\right] + b$$

   Proof:
$$\begin{aligned} E\left[ag(X)+b\right] &= \sum_{\text{all } x} \left[ag(x)+b\right] f(x) \\ &= \sum_{\text{all } x} \left[ag(x)f(x)+bf(x)\right] \\ &= a\sum_{\text{all } x} g(x)f(x) + b\sum_{\text{all } x} f(x) \\ &= aE\left[g(X)\right] + b \quad \text{ since } \sum_{\text{all } x} f(x) = 1 \end{aligned}$$

2. Similarly for constants a and b and two functions g_1 and g_2, it is also easy to show
$$E\left[ag_1(X)+bg_2(X)\right] = aE\left[g_1(X)\right] + bE\left[g_2(X)\right]$$

Don't let expected value intimidate you. Much of it is common sense. For example, using property 1, with we let $a = 0$ and $b = 13$ we obtain $E(13) = 13$. The expected value of a constant b is, of course, equal to b. The property also implies $E\left(2X\right) = 2E(X)$ if we use $a = 2$, $b = 0$, and $g(X) = X$. This is obvious also. Note, however, that for $g(x)$ a *nonlinear* function, it is NOT generally true that $E[g(X)] = g\left[E(X)\right]$; this is a common mistake. (Check this for the example above when $g(X) = 12/X^2$.)

7.3 Some Applications of Expectation

Because expected value is an average value, it is frequently used in problems where costs or profits are connected with the outcomes of a random variable X. It is also used as a summary statistic; for example, one often hears about the expected life (expectation of lifetime) for a person or the expected return on an investment. Be cautious however. The expected value does NOT tell the whole story about a distribution. One investment could have a higher expected value than another but much much larger probability of large losses.

The following are examples.

Example: Expected Winnings in a Lottery A small lottery[26] sells 1000 tickets numbered $000, 001, \ldots, 999$; the tickets cost $\$10$ each. When all the tickets have been sold the draw takes place: this consists of a single ticket from 000 to 999 being chosen at random. For ticket holders the prize structure is as follows:

- Your ticket is drawn - win $\$5000$.
- Your ticket has the same first two number as the winning ticket, but the third is different - win $\$100$.
- Your ticket has the same first number as the winning ticket, but the second number is different - win $\$10$.
- All other cases - win nothing.

Let the random variable X represent the winnings from a given ticket. Find $E(X)$.

Solution: The possible values for X are 0, 10, 100, 5000 (dollars). First, we need to find the probability function for X. We find (make sure you can do this) that $f(x) = P(X = x)$ has values

$$f(0) = 0.9, \quad f(10) = 0.09, \quad f(100) = 0.009, \quad f(5000) = 0.001$$

The expected winnings are thus the expected value of X, or

$$E(X) = \sum_{\text{all x}} xf(x) = \$6.80$$

Thus, the gross expected winnings per ticket are $\$6.80$. However, since a ticket costs $\$10$ your expected net winnings are negative, $-\$3.20$ (that is, an expected loss of $\$3.20$).

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

Remark: The random variable associated with a given problem may be defined in different ways but the expected winnings will remain the same. For example, instead of defining X as the amount won we could have defined $X = 0, 1, 2, 3$ as follows:

[26] "Here's something to think about: How come you never see a headline like 'Psychic Wins Lottery'?" Jay Leno (1950 -)

$X = 3$	all 3 digits of number match winning ticket
$X = 2$	1st 2 digits (only) match
$X = 1$	1st digit (but not the 2nd) match
$X = 0$	1st digit does not match

Now, we would define the function $g(x)$ as the winnings when the outcome $X = x$ occurs. Thus,

$$g(0) = 0, \qquad g(1) = 10, \qquad g(2) = 100, \qquad g(3) = 5000$$

The expected winnings are then

$$E\left[g(X)\right] = \sum_{x=0}^{3} g(x)f(x) = \$6.80$$

the same as before.

Example: Diagnostic medical Tests Often there are cheaper, less accurate tests for diagnosing the presence of some conditions in a person, along with more expensive, accurate tests. Suppose we have two cheap tests and one expensive test, with the following characteristics. All three tests are positive if a person has the condition (there are no "false negatives"), but the cheap tests give "false positives". Let a person be chosen at random, and let $D = \{\text{person has the condition}\}$. For the three tests the probability of a false positive and cost are:

Test	$P\left(\text{positive test } \vert \overline{D}\right)$	Cost (in dollars)
1	0.05	5
2	0.03	8
3	0	40

We want to check a large number of people for the condition, and have to choose among three testing strategies:

(i) Use Test 1, followed by Test 3 if Test 1 is positive[27].

(ii) Use Test 2, followed by Test 3 if Test 2 is positive.

(iii) Use Test 3.

Determine the expected cost per person under each of strategies (i), (ii) and (iii). We will then choose the strategy with the lowest expected cost. It is known that about 0.001 of the population have the condition ($P(D) = 0.001$, $P(\overline{D}) = 0.999$).

[27] Assume that given D or $\overline{D}$, tests are independent of one another.

Solution: For a person tested chosen at random and tested, define the random variable X as follows:

$$\begin{aligned} X &= 1 \qquad \text{if the initial test is negative} \\ X &= 2 \qquad \text{if the initial test is positive} \end{aligned}$$

Let $g(x)$ be the total cost of testing the person. The expected cost per person is then

$$E[g(X)] = \sum_{x=1}^{2} g(x) f(x)$$

The probability function $f(x)$ for X and function $g(x)$ differ for strategies (i), (ii) and (iii). Consider for example strategy (i). Then

$$\begin{aligned} P(X = 2) &= P\,(\text{initial test positive}) \\ &= P(D) + P(\text{positive}|\overline{D})P(\overline{D}) \\ &= 0.001 + (0.05)(0.999) \\ &= 0.0510 \end{aligned}$$

The rest of the probabilities, associated values of $g(X)$ and $E[g(X)]$ are obtained below.

$$\begin{aligned} \text{(i)} \quad & f(2) = 0.0510 \text{ (obtained above)} \\ & f(1) = P(X = 1) = 1 - f(2) = 1 - 0.0510 = 0.949 \\ & g(1) = 5 \qquad\qquad g(2) = 45 \\ & E[g(X)] = 5(0.949) + 45(0.0510) = \$7.04 \end{aligned}$$

$$\begin{aligned} \text{(ii)} \quad & f(2) = 0.001 + (0.03)(0.999) = 0.03097 \\ & f(1) = 1 - f(2) = 0.96903 \\ & g(1) = 8 \qquad\qquad g(2) = 48 \\ & E[g(X)] = 8(0.96903) + 48(0.03097) = \$9.2388 \end{aligned}$$

$$\begin{aligned} \text{(iii)} \quad & f(2) = 0.001, \; f(1) = 0.999 \\ & g(2) = g(1) = 40 \\ & E[g(X)] = \$40.00 \end{aligned}$$

Therefore the cheapest strategy is strategy (i).

Problem

7.3.1 A lottery[28] has tickets numbered 000 to 999 which are sold for \$1 each. One ticket is selected at random and a prize of \$200 is given to any person whose ticket number is a permutation of the selected ticket number. All 1000 tickets are sold. What is the expected profit or loss to the organization running the lottery?

7.4 Means and Variances of Distributions

It is useful to know the means, $\mu = E(X)$ of the probability models derived in Chapter 5.

Example: Expected value of a Binomial random variable Let $X \sim Binomial(n, p)$. Find $E(X)$.

Solution:

$$\begin{aligned} E(X) &= \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_{x=0}^{n} x \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \end{aligned}$$

When $x = 0$ the value of the expression is 0. We can therefore begin our sum at $x = 1$. Provided $x \neq 0$, we can expand $x!$ as $x(x-1)!$ (so it is important to eliminate the term when $x = 0$). Therefore

$$\begin{aligned} E(X) &= \sum_{x=1}^{n} \frac{n(n-1)!}{(x-1)!\,[(n-1)-(x-1)]!} p p^{x-1} (1-p)^{(n-1)-(x-1)} \\ &= np(1-p)^{n-1} \sum_{x=1}^{n} \binom{n-1}{x-1} \left(\frac{p}{1-p}\right)^{x-1} \end{aligned}$$

Let $y = x - 1$ in the sum, to get

$$\begin{aligned} E(X) &= np(1-p)^{n-1} \sum_{y=0}^{n-1} \binom{n-1}{y} \left(\frac{p}{1-p}\right)^{y} \\ &= np\,(1-p)^{n-1} \left(1 + \frac{p}{1-p}\right)^{n-1} \quad \text{by the Binomial Theorem if } 0 < p < 1 \\ &= np\,(1-p)^{n-1} \frac{(1-p+p)^{n-1}}{(1-p)^{n-1}} \\ &= np \end{aligned}$$

[28] "I've done the calculation and your chances of winning the lottery are identical whether you play or not." Fran Lebowitz (1950 -)

Exercise: Does this result make sense? If you try something 100 times and there is a 20% chance of success each time, how many successes do you expect to get, on average?

Example: Expected value of the Poisson random variable Let X have a Poisson distribution where λ is the average rate of occurrence and the time interval is of length t. Find $\mu = E(X)$.

Solution: Since the probability function of X is

$$f(x) = \frac{(\lambda t)^x \, e^{-\lambda t}}{x!} \quad \text{for } \; x = 0, 1, \ldots$$

then

$$E(X) = \sum_{x=0}^{\infty} x \frac{(\lambda t)^x \, e^{-\lambda t}}{x!}$$

As in the Binomial example, we can eliminate the term when $x = 0$ and expand $x!$ as $x(x-1)!$ for $x = 1, 2, \ldots$ to obtain

$$\begin{aligned}
E(X) &= \sum_{x=1}^{\infty} x \frac{(\lambda t)^x \, e^{-\lambda t}}{x!} = \sum_{x=1}^{\infty} x \frac{(\lambda t)^x \, e^{-\lambda t}}{x(x-1)!} \\
&= \sum_{x=1}^{\infty} (\lambda t) e^{-\lambda t} \frac{(\lambda t)^{x-1}}{(x-1)!} \\
&= (\lambda t) \, e^{-\lambda t} \sum_{x=1}^{\infty} \frac{(\lambda t)^{x-1}}{(x-1)!} \\
&= (\lambda t) e^{-\lambda t} \sum_{y=0}^{\infty} \frac{(\lambda t)^y}{y!} \text{ letting } y = x - 1 \text{ in the sum} \\
&= (\lambda t) e^{-\lambda t} e^{\lambda t} \quad \text{by the Exponential series} \\
&= \lambda t
\end{aligned}$$

Note that we used the symbol $\mu = \lambda t$ earlier in connection with the Poisson model; this was because we knew (but couldn't show until now) that $E(X) = \mu$.

Exercise: These techniques can also be used to work out the mean for the Hypergeometric or Negative Binomial distributions. Looking back at how we proved that $\sum f(x) = 1$ shows the same method of summation used to find μ. However, in Chapter 9 we will give a simpler method of finding the means of these distributions, which are $E(X) = nr/N$ (Hypergeometric) and $E(X) = k(1-p)/p$ (Negative Binomial).

Variability: 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. We give a way of measuring the amount of variability next. 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,

$$\begin{aligned} E\left(X-\mu\right) &= E(X)-\mu \quad \text{since } \mu \text{ is a constant} \\ &= 0 \end{aligned}$$

We soon realize that for a measure of variability, we can use the expected value of a function that has the same sign for $X > \mu$ and for $X < \mu$. One might try the expected value of the distance between X and its mean, e.g. $E(|X-\mu|)$. An alternative, more mathematically tractable version squares the distance (much as Euclidean distance in $\Re^n$ involves a sum of squared distances) is the variance.

Definition 18 *The **variance** of a random variable X, denoted by $Var(X)$ or by σ^2, is*

$$\sigma^2 = Var(X) = E\left[(X-\mu)^2\right]$$

In words, the variance is the average square of the distance from the mean. This turns out to be a very useful measure of the variability of X.

The basic definition of variance is often awkward to use for mathematical calculation of $Var(X)$, whereas the following two results are often useful:

$$\begin{array}{ll} (1) & Var(X) = E\left(X^2\right) - \left[E\left(X\right)\right]^2 = E\left(X^2\right) - \mu^2 \\ (2) & Var(X) = E\left[X(X-1)\right] + E\left(X\right) - \left[E\left(X\right)\right]^2 = E\left[X(X-1)\right] + \mu - \mu^2 \end{array}$$

Proof:

(1) Using properties of expected value,

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

(2) Since $X^2 = X(X-1) + X$,

$$\begin{aligned} Var(X) = E\left(X^2\right) - \mu^2 &= E\left[X\left(X-1\right) + X\right] - \mu^2 \\ &= E\left[X(X-1)\right] + E(X) - \mu^2 \\ &= E\left[X(X-1)\right] + \mu - \mu^2 \end{aligned}$$

Formula (2) is most often used when there is an $x!$ term in the denominator of $f(x)$. Otherwise, formula (1) is generally easier to use.

Suppose the random variable X is the number of dollars that a person wins if they play a certain game. We notice that the units of measurement for $E(X)$ will also be dollars but the units of measurement for $Var(X)$ will be $(\text{dollars})^2$. We can regain the original units by taking the square root of $Var(X)$. This is called the standard deviation of X, and is denoted by σ, or by $sd(X)$.

Definition 19 *The standard deviation of a random variable X is*

$$\sigma = sd(X) = \sqrt{Var(X)} = \sqrt{E\left[(X-\mu)^2\right]}$$

Both variance and standard deviation are commonly used to measure variability.

Example: Suppose X is a random variable with probability function given by

x	1	2	3	4	5	6	7	8	9	Total
$f(x)$	0.07	0.10	0.12	0.13	0.16	0.13	0.12	0.10	0.07	1

The probability histogram for X is given in Figure 7.2.

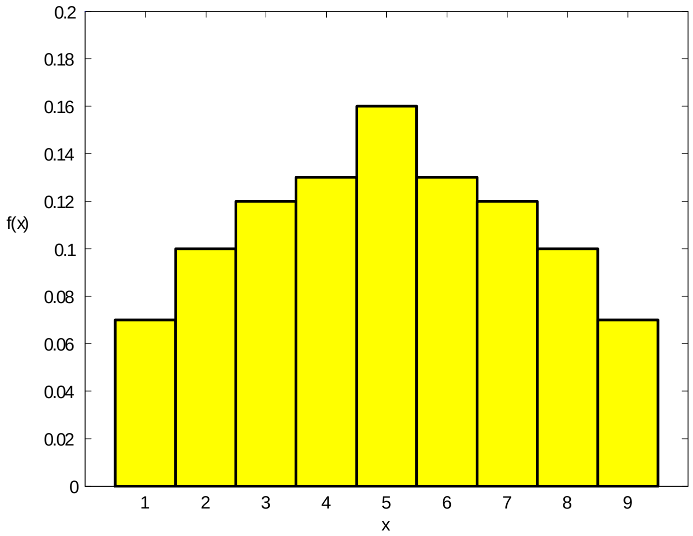


Figure 7.2: Probability Histogram for X

Find $E(X)$ and $Var(X)$.

Solution:

$$\begin{aligned}
\mu &= E\left(X\right) \\
&= 1\left(0.07\right) + 2\left(0.1\right) + 3\left(0.12\right) + 4\left(0.13\right) + 5\left(0.16\right) \\
&+ 6\left(0.13\right) + 7\left(0.12\right) + 8\left(0.1\right) + 9\left(0.07\right) \\
&= 5
\end{aligned}$$

$E\left(X\right) = 5$ should be obvious by looking at the histogram. If a probability histogram is symmetric about the line $x = \mu$ then $E\left(X\right) = \mu$ without any calculation.

Without doing any calculations we also know that $Var(X) = \sigma^2 \leq 16$. This is because the possible values of X are $\{1, 2, \ldots, 9\}$ and so the maximum possible value for $\left(X - \mu\right)^2$ is $\left(9 - 5\right)^2$ or $\left(1 - 5\right)^2 = 16$. Therefore

$$\begin{aligned}
Var\left(X\right) = E\left[\left(X - 5\right)^2\right] &= \sum_{x=1}^{9} \left(x - 5\right)^2 P\left(X = x\right) \\
&\leq \sum_{x=1}^{9} \left(9 - 5\right)^2 P\left(X = x\right) = 16 \sum_{x=1}^{9} P\left(X = x\right) = 16\left(1\right) = 16
\end{aligned}$$

An expected value of a function, say $E\left[g(X)\right]$ is always somewhere between the minimum and the maximum value of the function $g(x)$ so in this case $0 \leq Var(X) \leq 16$. Since

$$\begin{aligned}
E\left(X^2\right) &= \left(1\right)^2\left(0.07\right) + \left(2\right)^2\left(0.1\right) + \left(3\right)^2\left(0.12\right) + \left(4\right)^2\left(0.13\right) + \left(5\right)^2\left(0.16\right) \\
&+ \left(6\right)^2\left(0.13\right) + \left(7\right)^2\left(0.12\right) + \left(8\right)^2\left(0.1\right) + \left(9\right)^2\left(0.07\right) \\
&= 30.26
\end{aligned}$$

Therefore

$$\begin{aligned}
\sigma^2 = Var\left(X\right) &= E\left(X^2\right) - \mu^2 \\
&= 30.26 - \left(5\right)^2 \\
&= 5.26
\end{aligned}$$

and

$$\begin{aligned}
\sigma &= \sqrt{Var\left(X\right)} \\
&= \sqrt{5.26} \\
&= 2.2935
\end{aligned}$$

To see how $\sigma^2 = Var\left(X\right)$ or $\sigma = sd\left(X\right)$ reflects the shape of a probability histogram see Figure 7.3. In each case the range of the random variable X is $\{1, 2, \ldots, 9\}$ and the mean is $\mu = E\left(X\right) = 5$.

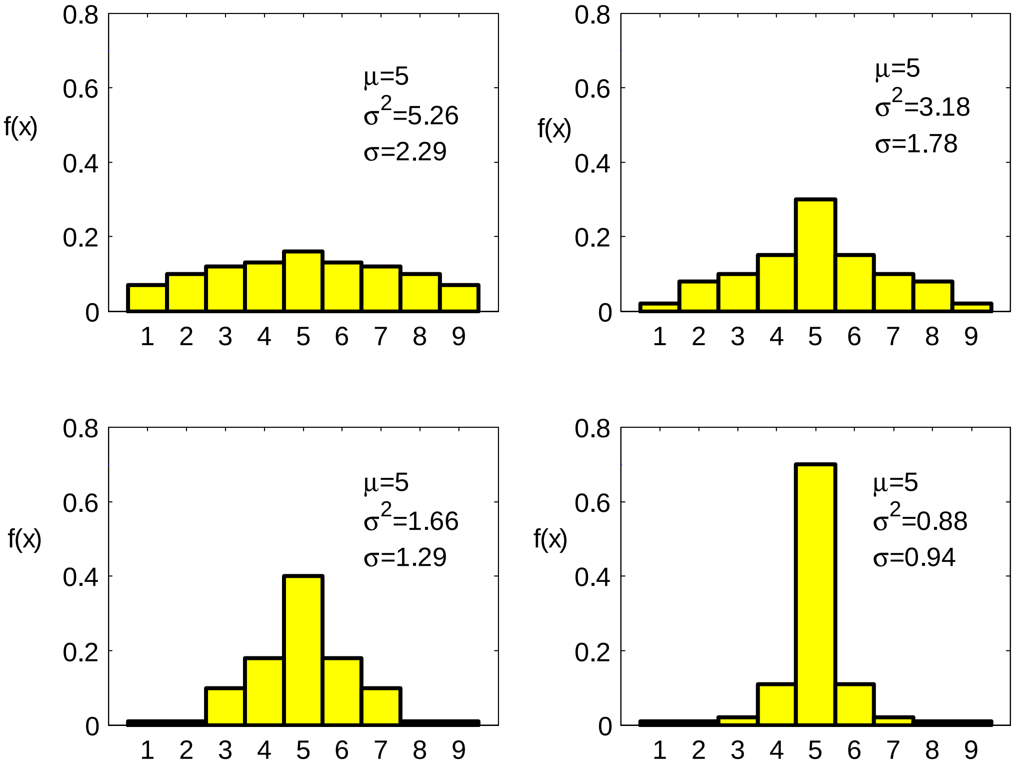


Figure 7.3: How $Var\,(X)$ or $sd\,(X)$ reflects the spread of a probability histogram

Each histogram is labeled with its corresponding values for $\sigma^2 = Var\,(X)$ and $\sigma = sd\,(X)$. We can see from the histograms that a small value of $\sigma^2\,(\sigma)$ suggests there is a greater probability of getting an observed value of the random variable near the mean $\mu = E\,(X)$. A large value of $\sigma^2\,(\sigma)$ suggests there is a greater probability of getting an observed value of the random variable that is not close to the mean $\mu = E\,(X)$.

Example: Variance of Binomial random variable Let $X \sim Binomial(n, p)$. Find $Var(X)$.

Solution: The probability function for X is

$$f(x) = \frac{n!}{x!(n-x)!}p^x\,(1-p)^{n-x} \quad \text{for } x = 0, 1, \ldots, n$$

so we'll use formula (2) above,

$$E\,[X(X-1)] = \sum_{x=0}^{n} x(x-1)\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}$$

If $x = 0$ or $x = 1$ the value of the term is 0, so we can begin summing at $x = 2$. For $x \neq 0$ or 1, we can expand the $x!$ as $x(x-1)(x-2)!$

$$\text{Therefore } E\left[X\left(X-1\right)\right] = \sum_{x=2}^{n} \frac{n!}{(x-2)!(n-x)!} p^x (1-p)^{n-x}$$

Now re-group to fit the Binomial Theorem, since that was the summation technique used to show $\sum f(x) = 1$ and to derive $\mu = np$.

$$\begin{aligned} E\left[X(X-1)\right] &= \sum_{x=2}^{n} \frac{n(n-1)(n-2)!}{(x-2)!\left[(n-2)-(x-2)\right]!} p^2 p^{x-2} (1-p)^{(n-2)-(x-2)} \\ &= n(n-1)p^2(1-p)^{n-2} \sum_{x=2}^{n} \binom{n-2}{x-2} \left(\frac{p}{1-p}\right)^{x-2} \end{aligned}$$

Let $y = x - 2$ in the sum, giving

$$\begin{aligned} E\left[X(X-1)\right] &= n(n-1)p^2(1-p)^{n-2} \sum_{y=0}^{n-2} \binom{n-2}{y} \left(\frac{p}{1-p}\right)^{y} \\ &= n(n-1)p^2(1-p)^{n-2} \left(1 + \frac{p}{1-p}\right)^{n-2} \quad \text{by the Binomial Theorem} \\ &= n(n-1)p^2(1-p)^{n-2} \frac{(1-p+p)^{n-2}}{(1-p)^{n-2}} \\ &= n(n-1)p^2 \end{aligned}$$

Then

$$\begin{aligned} \sigma^2 &= E\left[X\left(X-1\right)\right] + \mu - \mu^2 \\ &= n(n-1)p^2 + np - (np)^2 \\ &= n^2p^2 - np^2 + np - n^2p^2 \\ &= np(1-p) \end{aligned}$$

Remember that the variance of a Binomial distribution is $np(1-p)$, since we'll be using it later in the course.

In Figure 7.4 there are four Binomial probability histograms for various values of n and p along with their means, variances and standard deviations. From the top two panels we can see that a $Binomial\left(10, 0.1\right)$ random variable and a $Binomial\left(10, 0.9\right)$ random variable have the same variance and standard deviation and that the probability histograms are mirror images of each other. This is what you might expect given that the Binomial distribution arises as the number of successes in n Bernoulli

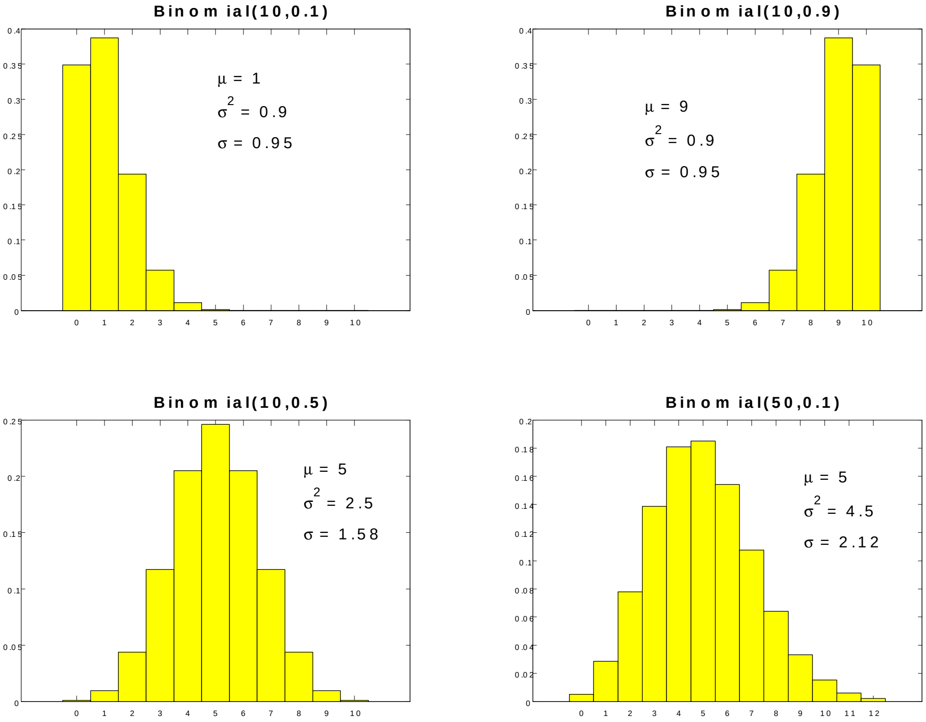


Figure 7.4: Probability histograms, means and variances for various Binomial(n, p) random variables

trials and the only difference between a $Binomial\,(10, 0.1)$ random variable and a $Binomial\,(10, 0.9)$ random variable is which outcome is labeled a Success (S) and which is labeled a Failure (F).

The lower left hand panel contains the probability histogram for a $Binomial\,(10, 0.5)$ random variable which we notice is symmetric about its mean $\mu = 10\,(0.5) = 5$. In fact the probability histogram for a $Binomial\,(n, 0.5)$ random variable will always be symmetric about its mean $\mu = np$. Note also that for fixed n, a $Binomial\,(n, 0.5)$ random variable has the largest variance since $Var\,(X) = np(1-p)$ is maximized for $p = 0.5$.

The lower right hand panel contains the probability histogram for a $Binomial\,(50, 0.1)$ random variable which we notice is fairly symmetric about its mean $\mu = 50\,(0.1) = 5$ even though $p = 0.1$ is not close to 0.5. This observation leads to an approximation to the Binomial distribution which is discussed in Section 10.1.

Example: Variance of Poisson random variable Suppose X has a Poisson(μ) distribution. Find $Var(X)$.

Solution: The probability function for X is

$$f(x) = \frac{\mu^x e^{-\mu}}{x!} \text{ for } x = 0, 1, 2, \ldots$$

from which we obtain

$$\begin{aligned} E\left[X(X-1)\right] &= \sum_{x=0}^{\infty} x(x-1)\frac{\mu^x e^{-\mu}}{x!} \\ &= \sum_{x=2}^{\infty} x(x-1)\frac{\mu^x e^{-\mu}}{x(x-1)(x-2)!}, \quad \text{set the lower limit to 2 and expand } x! \\ &= \mu^2 e^{-\mu} \sum_{x=2}^{\infty} \frac{\mu^{x-2}}{(x-2)!} \end{aligned}$$

Let $y = x - 2$ in the sum, giving

$$\begin{aligned} E\left[X(X-1)\right] &= \mu^2 e^{-\mu} \sum_{y=0}^{\infty} \frac{\mu^y}{y!} = \mu^2 e^{-\mu} e^{\mu} \quad \text{by the Exponential series} \\ &= \mu^2 \\ \text{so } \sigma^2 &= E\left[X(X-1)\right] + \mu - \mu^2 \\ &= \mu^2 + \mu - \mu^2 = \mu \end{aligned}$$

(*For the Poisson distribution, the variance equals the mean.*)

In Figure 7.5 there are four Poisson probability histograms for various values of μ along with their means, variances and standard deviations. In the top left histogram $\mu = 0.5$ and we see that much of the probability is concentrated on the values $x = 0$ and $x = 1$. As the value of μ approaches 0, more and more of the probability will be concentrated on the value $x = 0$ since

$$\lim_{\mu \to 0} P\left(X = 0\right) = \lim_{\mu \to 0} \frac{\mu^0 e^{-\mu}}{0!} = \lim_{\mu \to 0} e^{-\mu} = 1.$$

This is consistent with $Var\left(X\right) = \mu$ approaching 0 as well.

As μ increases $Var\left(X\right) = \mu$ increases and the spread of the probability histogram increases as illustrated in the top right histogram where $\mu = 1.5$ and the bottom two histograms where $\mu = 5$ and $\mu = 9.5$. Note that for $\mu = 5$ the histogram is quite symmetric about the mean $E\left(X\right) = \mu = 5$ and even more symmetric for $\mu = 9.5$. This observation leads to an approximation to the Poisson distribution which is discussed in Section 10.1.

Properties of Mean and Variance:

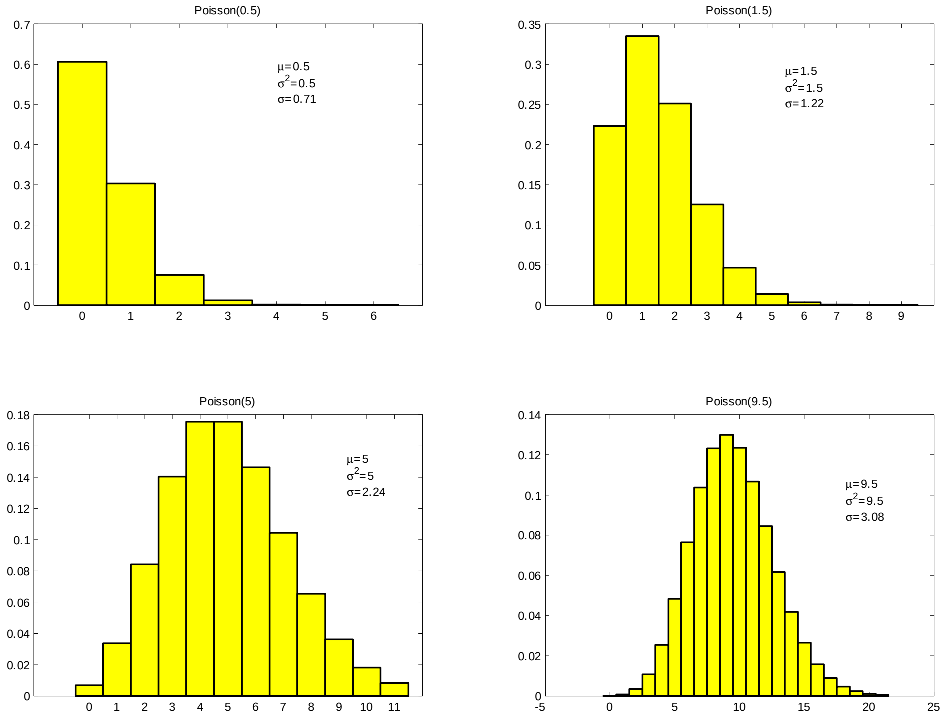


Figure 7.5: Probability histograms, means and variances for various Poisson(μ) random variables

If a and b are constants and $Y = aX + b$, then

$$\mu_Y = E(Y) = aE(X) + b = a\mu_X + b$$

and

$$\sigma_Y^2 = Var(Y) = a^2 Var(X) = a^2 \sigma_X^2$$

where $\mu_X = E(X)$, $\sigma_X^2 = Var(X)$, $E(Y) = \mu_Y$, and $Var(Y) = \sigma_Y^2$.

Proof:
We already showed that $E\left(Y\right) = E(aX+b) = a\mu_X + b = \mu_Y$. Then

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

This result is to be expected. Adding a constant, b, to all values of X has no effect on the amount of variability. So it makes sense that $Var(aX+b)$ doesn't depend on the value of b. Also since variance is in squared units, multiplication by a constant results in multiplying the variance by the constant squared. A simple way to relate to this result is to consider a random variable X which represents a temperature in degrees Celsius (even though this is a continuous random variable which we don't study until Chapter 9). Now let Y be the corresponding temperature in degrees Fahrenheit. We know that

$$Y = \frac{9}{5}X + 32$$

and it is clear that $\mu_Y = (\frac{9}{5})\mu_X + 32$ and that $\sigma_Y^2 = (\frac{9}{5})^2\sigma_X^2$.

Problems

7.4.1 An airline knows that there is a 97% chance a passenger for a certain flight will show up, and assumes passengers arrive independently of each other. Tickets cost $\$100$, but if a passenger shows up and can't be carried on the flight the airline has to refund the $\$100$ and pay a penalty of $\$400$ to each such passenger. How many tickets should they sell for a plane with 120 seats to maximize their expected ticket revenues after paying any penalty charges? Assume ticket holders who don't show up get a full refund for their unused ticket.

7.4.2 A typist typing at a constant speed of 60 words per minute makes a mistake in any particular word with probability 0.04, independently from word to word. Each incorrect word must be corrected; a task which takes 15 seconds per word.

(a) Find the mean and variance of the time (in seconds) taken to finish a 450 word passage.

(b) Would it be less time consuming, on average, to type at 45 words per minute if this reduces the probability of an error to 0.02?

7.5 Chapter 7 Problems

1. For Chapter 5, Problems 1 and 2 find $E\left(X\right)$, $E\left(X^2\right)$, and $Var\left(X\right)$.

2. Let X have probability function

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

   Find $E\left(X\right)$, $E\left(X^2\right)$, and $Var\left(X\right)$.

3. A person plays a game in which a fair coin is tossed until the first tail occurs. The person wins $\$2^x$ if x tosses are needed for $x = 1, 2, 3, 4, 5$ but loses $\$256$ if $x > 5$.

   (a) Determine the expected winnings.

   (b) Determine the variance of the winnings.

4. Yasmin and Zack are undergraduate mathematics students currently taking the same five courses. Let X be the number of assignments they have in one week. The probability function of X is:

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 Yasmin and Zack drink in one week both depend on the number of assignments they have. Yasmin drinks about $2X^2$ cups per week and Zack drinks about $|2X - 1|$ cups per week.

   (a) Find the expected number of cups of coffee that Yasmin will drink in a week. Find the expected number of cups of coffee that Zack will drink in a week.

   (b) Find the variance of the number of cups of coffee that Yasmin will drink in a week. Find the variance of the number of cups of coffee that Zack will drink in a week.

5. Suppose $X \sim Geometric\left(p\right)$.

   (a) Find the mean and variance of X. **Hint**: See Problem 3.6.2.

   (b) Use your result in (a) to show that if p is the probability of "success" (S) in a sequence of Bernoulli trials, then the expected number of trials until the first S occurs is $1/p$. Explain why this is "obvious".

6. **Diagnostic medical tests I:** Consider diagnostic tests like those discussed in the example of Section 7.3. Assume that for a randomly selected person, $P(D) = 0.02$, $P(R|D) = 1$, $P(R|\overline{D}) = 0.05$, so that the inexpensive test only gives false positive, and not false negative, results. Suppose that this inexpensive test costs $\$10$. If a person tests positive then they are also given a more expensive test, costing $\$100$, which correctly identifies all persons with the disease. What is the expected cost per person if a population is tested for the disease using the inexpensive test followed, if necessary, by the expensive test?

7. **Diagnostic medical tests II:** Two percent of the population has a certain condition for which there are two diagnostic tests. Test A, which costs $\$1$ per person, gives positive results for 80% of persons with the condition and for 5% of persons without the condition. Test B, which costs $\$100$ per person, gives positive results for all persons with the condition and negative results for all persons without it.

   (a) Suppose that test B is given to 150 persons, at a cost of $\$15,000$. How many cases of the condition would one expect to detect?

   (b) Suppose that 2000 persons are given test A, and then only those who test positive are given test B. Show that the expected cost is $\$15,000$ but that the expected number of cases detected is much larger than in part (a).

8. **Diagnostic medical tests III:** Suppose that n people take a blood test for a disease, where each person has probability p of having the disease, independent of other persons. To save time and money, blood samples from k people are pooled and analyzed together. If none of the k persons has the disease then the test will be negative, but otherwise it will be positive. If the pooled test is positive then each of the k persons is tested separately (so $k+1$ tests are done in that case).

   (a) Let X be the number of tests required for a group of k people. Show that

   $$E(X) = k + 1 - k(1-p)^k$$

   (b) What is the expected number of tests required for n/k groups of k people each? If $p = 0.01$, evaluate this for the cases $k = 1, 5, 10$.

   (c) Show that if p is small, the expected number of tests in part (b) is approximately $n(kp + k^{-1})$, and is minimized for $k \approx p^{-1/2}$. **Hint**: Use the linear approximation $(1-p)^k \approx 1 - kp$ for p close to 0.

9. The probability that a roulette wheel stops on a red number is $18/37$. Suppose you bet x dollars on "red". If the wheel stops on a red number then you are paid $2x$ dollars and your net winnings are x dollars. If the wheel does not stop on a red number then you lost your bet and your net winnings are $-x$ dollars.

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

   (b) For each of the two cases in part (a), calculate the probability that you made a profit (that is, your winnings are positive, not negative).

10. Consider the slot machine discussed in Chapter 4, Problem 17. 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
Symbols	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 $\$10$, and if all three wheels stop on a house, you win $\$5$. Otherwise you win nothing. Find the expected value of your winnings **per dollar spent**.

11. Suppose a slot machine has $n+1$ possible outcomes $A_1, A_2, \ldots, A_{n+1}$ for a single play. A single play costs $\$1$. If outcome A_i occurs, a player wins $\$a_i$, for $i = 1, 2, \ldots, n$. If outcome A_{n+1} occurs, the player wins nothing. In other words, if outcome A_i, $i = 1, 2, \ldots, n$ occurs the player's net winnings are $\$\,(a_i - 1)$ and if A_{n+1} occurs the player's net winnings are $\$\,(-1)$.

   (a) Give a formula for the expected value of your net winnings from a single play, if the probabilities of the $n+1$ outcomes are $p_i = P(A_i)$, $i = 1, 2, \ldots, n+1$.

   (b) The slot machine owner wants the expected value of the player's net winnings to be negative. Suppose $n = 4$, 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 $\$3$ when outcome A_1 occurs, and $\$5$ when outcomes A_2, A_3 or A_4 occur, determine the expected value of the player's net winnings from a single play.

   (c) The slot machine owner wants the player's net winnings (equivalently the owner's net payout) to be $\$db_i$ when outcome A_i occurs, where $b_i = 1/p_i$ and d is a number between 0 and 1. The owner also wants the expected value of the player's net winnings to be $\$\,(-0.05)$ per play. Find d as a function of n and p_{n+1}. What is the value of d if $n = 10$ and $p_{n+1} = 0.7$?

12. 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 she receives $\$100$ and if she answers B correctly 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 winnings?

    (b) Suppose that she must now pay a $\$50$ penalty if she gets the the first question wrong. What question should she attempt first?

13. A manufacturer of car radios ships them to retailers in cartons of n radios. The profit per radio is $\$59.50$, less shipping cost of $\$25$ per carton, so the profit is $\$\,(59.5n-25)$ per carton. To promote sales by assuring high quality, the manufacturer promises to pay the retailer $\$200X^2$ if X radios in the carton are defective. (The retailer is then responsible for repairing any defective radios.) Suppose radios are produced independently and that 5% of radios are defective. How many radios should be packed per carton to maximize expected net profit per carton?

14. On Halloween trick-or-treaters arrive at a house from 5:30pm until 9pm according to a Poisson Process with an average of 12 trick-or-treaters per hour.

    (a) What is the probability that between 5 and 7 trick-or-treaters (inclusive) arrive in the first half hour?

    (b) How many trick-or-treaters would be expected to arrive over the whole evening?

    (c) What number of trick-or-treaters is most likely to arrive?

    (d) What is the variance of the number of trick-or-treaters that arrive over the whole evening?

15. Assume that each week a stock either increases in value by $\$1$ with probability $\frac{1}{2}$ or decreases by $\$1$, these moves independent of the past. The current price of the stock is $\$50$. I wish to purchase a call option which allows me (if I wish to do so) the option of buying the stock 13 weeks from now at a "strike price" of $\$55$. Of course if the stock price at that time is $\$55$ or less there is no benefit to the option and it is not exercised. Assume that the return from the option is

    $$R = \max(S-55, 0)$$

    where S is the price of the stock after 13 weeks. What is the fair price of the option today assuming no transaction costs and 0% interest, that is, what is $E\,(R)$?
    Hint: Let $X =$ the number of times the stock increases in value by $\$1$ during the 13 weeks and determine S in terms of X.

16. **Web cache:** Web browsers often use a cache (locally-stored information which can be accessed very quickly) to improve performance. When a user requests information, the cache is searched first. If the information is found in the cache (known as a "cache hit"), it takes 10 ms (milliseconds) for the information to be displayed to the user. If the information is not found in the cache (a "cache miss"), a request is made to a web server (50 ms), a data base is searched (70 ms), and the information is returned and displayed to the user (50 ms).

    (a) If no cache is used, what is the expected time for the information to be displayed to the user?

    (b) If a cache is used and there is a 20% chance of a cache hit, what is the expected time for the information to be displayed to the user?

    (c) How small would the probability of a cache hit need to be to have these expected times be equal?

17. **Analysis of algorithms - Quicksort:** Suppose we have a set S of distinct numbers and we wish to sort them from smallest to largest. The Quicksort algorithm works as follows: When $n = 2$ it just compares the numbers and puts the smallest one first. For $n > 2$ it starts by choosing a random "pivot" number from the n numbers. It then compares each of the other $n - 1$ numbers with the pivot and divides them into groups S_1 (numbers smaller than the pivot) and $\bar{S}_1$ (numbers bigger than the pivot). It then does the same thing with S_1 and $\bar{S}_1$ as it did with S, and repeats this recursively until the numbers are all sorted. (Try this out with, say $n = 10$ numbers to see how it works.) In computer science it is common to analyze such algorithms by finding the expected number of comparisons (or other operations) needed to sort a list. Thus, let

$$C_n = \text{expected number of comparisons for lists of length } n$$

    (a) Show that if X is the number of comparisons needed,

$$C_n = \sum_{i=1}^{n} E(X \mid \text{initial pivot is } i\text{th smallest number}) \left(\frac{1}{n}\right)$$

    (b) Show that

$$E(X \mid \text{initial pivot is } i\text{th smallest number}) = n - 1 + C_{i-1} + C_{n-i}$$

    and thus that C_n satisfies the recursion (note $C_0 = C_1 = 0$)

$$C_n = n - 1 + \frac{2}{n}\sum_{k=1}^{n-1} C_k \quad \text{for } n = 2, 3, \ldots$$

(c) Show that

$$(n+1)C_{n+1} = 2n + (n+2)C_n \quad \text{for } n = 1, 2, \ldots$$

(d) (Harder) Use the result of part (c) to show that for large n,

$$\frac{C_{n+1}}{n+1} \sim 2\log(n+1)$$

(Note: $a_n \sim b_n$ means $a_n/b_n \to 1$ as $n \to \infty$.) This proves a result from computer science which says that for Quicksort, $C_n \sim O(n \log n)$.

18. **Challenge problem:** Let X_n be the number of ascents in a random permutation of the integers $\{1, 2, \ldots n\}$. For example, the number of ascents in the permutation 213546 is three, since $2, 135, 46$ form ascending sequences.

(a) Show the following recursion for the probabilities $p_n(k) = P\left(X_n = k\right)$

$$p_n(k+1) = \frac{k+1}{n} p_{n-1}(k) + \frac{n-k}{n} p_{n-1}(k-1)$$

(b) Cards numbered $1, 2, \ldots, n$ are shuffled, drawn and put into a pile as long as the card drawn has a number lower than its predecessor. A new pile is started whenever a higher card is drawn. Show that the distribution of the number of piles that we end with is that of X_n and that the expected number of piles is $\frac{n+1}{2}$.

8. CONTINUOUS RANDOM VARIABLES

8.1 General Terminology and Notation

For **continuous** random variables the range (set of possible values) is an interval (or a collection of intervals) on the real number line. Continuous random variables must be treated a little differently than discrete random variables because $P(X = x)$ is zero for each x. To illustrate a random variable with a continuous distribution, consider the simple spinning pointer in Figure 8.1 operating in a frictionless environment. Suppose we assume that the pointer is equally likely to stop at any point in the interval $(0, 4]$. If we assume this probability is $p > 0$, then for $A = \{x : 0 < x \leq 4\}$, $P(A) = \sum\limits_{x \epsilon (0,4]} p = \infty$ since the set A is uncountably infinite. This implies that probability that the pointer stops precisely at any given number x must be zero. Note however that it seems reasonable to assign a probability of $\frac{1}{16}$ to the event that the spinner stops at some value x in the interval $(0, \frac{1}{4}]$ or $(1\frac{3}{4}, 2]$. For continuous random variables we specify the probability of intervals, rather than individual points.

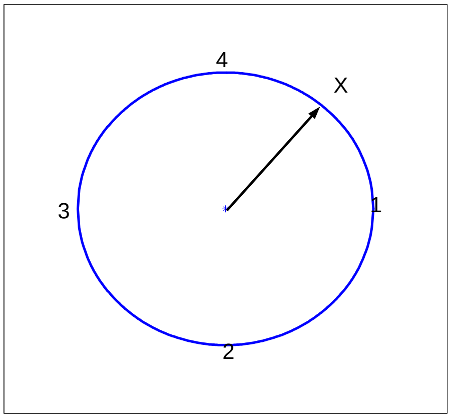


Figure 8.1: Spinner: a device for generating a continuous random variable

Consider another example produced by choosing a “random point” in a region. Suppose we plot a graph of a function $f(x)$ as in Figure 8.2 (assume the function is positive and has finite integral) and then generate a point at random by closing our eyes and firing a dart from a distance until at least one lands in the shaded region under the graph. We assume such a point, here denoted “*” is “uniformly” distributed under the graph. This means that the point is equally likely to fall in any one of many possible regions of a given area located in the shaded region so we only need to know the area of a region to determine the probability that a point falls in it. Consider the x-coordinate X of the point “*” as our random variable. (In Figure 8.2 it appears to be around 6.) Notice that the probability that X falls in a particular interval (a, b) is measured by the area of the region above this interval, that is, $\int\limits_a^b f(x)dx$ and so the probability of any particular point $P(X = a)$ is the area of the region immediately above this single point $\int\limits_a^a f(x)dx = 0$. This is another example of a random variable X which has a continuous distribution.

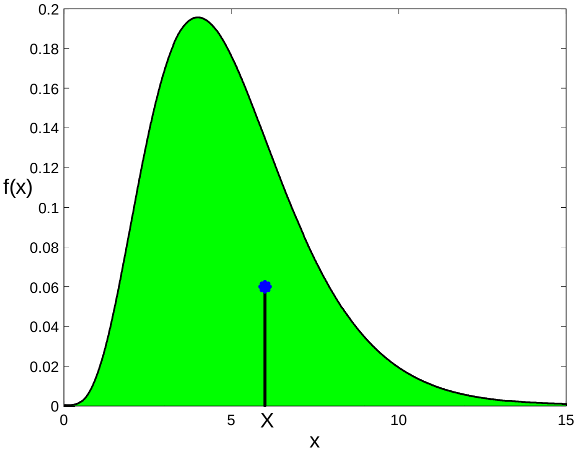


Figure 8.2: Graph of $f\ (x)$

For a continuous random variable X, there are two commonly used functions which describe its distribution. The first is the cumulative distribution function, used before for discrete distributions, and the second is the probability density function, the derivative of the cumulative distribution function.

Cumulative Distribution Function:

For discrete random variables we defined the cumulative distribution function, $F(x) = P\,(X \leq x)$. For continuous random variables we can also define the cumulative distribution function. For the spinner example, the probability the pointer stops between 0 and 1 is $1/4$ if all values x are equally “likely”;

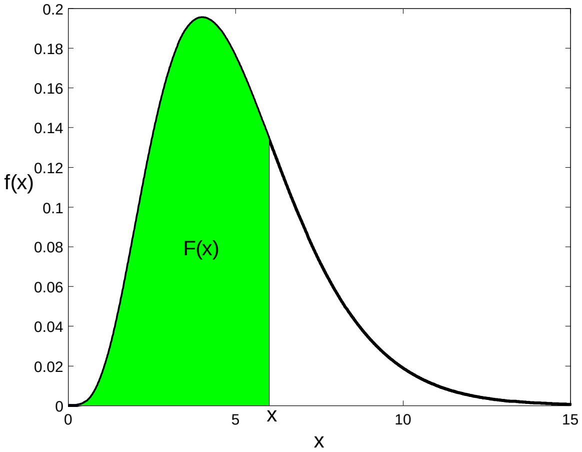


Figure 8.3: Area of shaded region equals $F(x) = P(X \leq x)$

between 0 and 2 the probability is $1/2$, between 0 and 3 it is $3/4$; and so on. In general, $F(x) = x/4$ for $0 < x \leq 4$. Also, $F(x) = 0$ for $x \leq 0$ since there is no chance of the pointer stopping at a number ≤ 0, and $F(x) = 1$ for $x > 4$ since the pointer is certain to stop at number below x if $x > 4$.

Suppose in the second example in which we generated a point at random under the graph of a function $f(x)$, we assume that the total area under the graph is one, then the cumulative distribution function $F(x)$ is the area under the graph but to the left of the point x as in Figure 8.3.

Most properties of a cumulative distribution function are the same for continuous variables as for discrete variables. These are:

1. $F(x)$ is defined for all real x
2. $F(x)$ is a non-decreasing function of x for all real x
3. $\lim_{x \to -\infty} F(x) = 0$; and $\lim_{x \to \infty} F(x) = 1$
4. $P(a < X \leq b) = F(b) - F(a)$.

Note that, as indicated before, for a continuous random variable, we have

$$0 = P(X = a) = \lim_{\varepsilon \to 0} P(a - \varepsilon < X \leq a) = \lim_{\varepsilon \to 0} F(a) - F(a - \varepsilon)$$

This means that $\lim_{\varepsilon \to 0} F(a - \varepsilon) = F(a)$ or that the distribution function F is a continuous function (in the sense of continuity in calculus). Also, since the probability is 0 at each point:

$$P(a < X < b) = P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b) = F(b) - F(a)$$

(For a discrete random variable, each of these 4 probabilities could be different.). For the continuous distributions in this chapter, we do not worry about whether intervals are open, closed, or half-open since the probability of these intervals is the same.

Probability Density Function: While the cumulative distribution function 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 X-values, $[x, x + \Delta x]$. The probability X lies in the interval is

$$P(x \leq X \leq x + \Delta x) = F(x + \Delta x) - F(x)$$

To compare the probabilities for two intervals, each of length Δx, is easy. Now suppose we consider what happens as Δx becomes small, and we divide the probability by Δx. This leads to the following definition.

Definition 20 *The* ***probability density function*** *(p.d.f.)* $f(x)$ *for a continuous random variable* X *is the derivative*

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

where $F(x)$ *is the cumulative distribution function for* X.

If the derivative of F does not exist at $x = a$ we usually define $f(a) = 0$ for convenience. Note that if the function $f(x)$ graphed in Figure 8.3 has total integral one, the cumulative distribution function or the area to the left of a point x is given by $F(x) = \int\limits_{-\infty}^{x} f(u)du$ and so the derivative of the cumulative distribution function is $F'(x) = f(x)$. It is clear from the way in which X was generated that $f(x)$ represents the relative likelihood of (small intervals around) different x-values. To do this we first note some properties of a probability density function. It is assumed that $f(x)$ is a continuous function of x at all points for which $0 < F(x) < 1$.

Properties of a probability density function:

1. $P(a \leq X \leq b) = F(b) - F(a) = \int\limits_a^b f(x)dx$. (This follows from the definition of $f(x)$)

2. $f(x) \geq 0$. (Since $F(x)$ is non-decreasing, its derivative is non-negative)

3. $\int\limits_{-\infty}^{\infty} f(x)dx = \int\limits_{\text{all} x} f(x)dx = 1$. (This is because $P(-\infty \leq X \leq \infty) = 1$)

4. $F(x) = \int\limits_{-\infty}^{x} f(u)du$. (This is just property 1 with $a = -\infty$)

To see that $f(x)$ represents the relative likelihood of different outcomes, we note that for Δx small,

$$P\left(x - \frac{\Delta x}{2} \leq X \leq x + \frac{\Delta x}{2}\right) = F\left(x + \frac{\Delta x}{2}\right) - F\left(x - \frac{\Delta x}{2}\right) \approx f(x)\Delta x$$

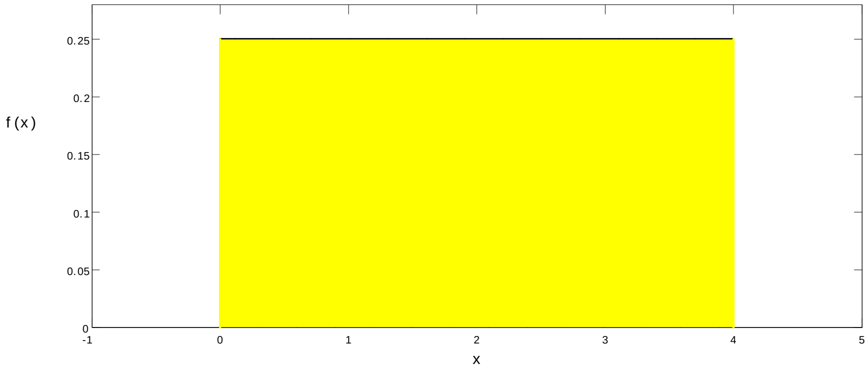


Figure 8.4: Probability density function for spinner example

Thus, $f(x) \neq P(X = x)$ **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 the function $f(x)$ shows such values clearly and for this reason it is very common to plot the probability density functions of continuous random variables.

Example: Consider the spinner example, where

$$F(x) = \begin{cases} 0 & x \leq 0 \\ \frac{x}{4} & 0 < x \leq 4 \\ 1 & x > 4 \end{cases}$$

Thus, the probability density function is $f(x) = F'(x)$, or

$$f(x) = \frac{1}{4} \text{ if } 0 < x < 4$$

and outside this interval the probability density function is defined to be 0. Figure 8.4 shows the probability density function $f(x)$; for obvious reasons this is called a “Uniform” distribution.

Remark: Continuous probability distributions are, like discrete distributions, mathematical[29] **models**. Thus, the Uniform distribution assumed for the spinner above is a model, and it seems likely it would be a good model for many real spinners.

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 “experiments” with continuous variables. The catch is that all

[29] “How can it be that mathematics, being after all a product of human thought which is independent of experience, is so admirably appropriate to the objects of reality? Is human reason, then, without experience, merely by taking thought, able to fathom the properties of real things?” Albert Einstein.

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.

To summarize, in measurements we are actually observing something like

$$P(x - 0.5\Delta \leq X \leq x + 0.5\Delta)$$

where Δ may be very small, but not zero. The probability of this outcome is **not** zero: it is (approximately) $f(x)\Delta$.

We now consider a more complicated mathematical example of a continuous random variable. Remember that it is always a good idea to sketch or plot the probability density function $f(x)$ for a random variable.

Example:
Let X be a continuous random variable with probability density function

$$f(x) = \begin{cases} kx^2 & 0 < x \leq 1 \\ k(2-x) & 1 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

Find**:** (a) the constant k

(b) the cumulative distribution function $F(x) = P\left(X \leq x\right)$

(c) $P\left(0.5 < X < 1.5\right)$

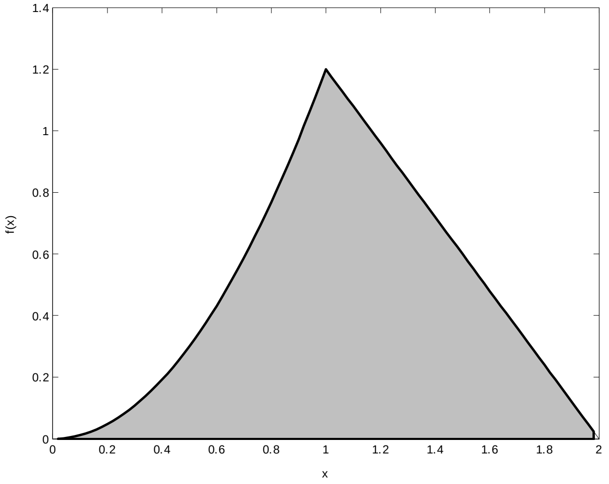


Figure 8.5: Area of shaded region equals one

Solution:

(a) When finding the area of a region bounded by different functions we split the integral into pieces.

$$\begin{aligned}1 &= \int\limits_{-\infty}^{\infty} f(x)dx \\ &= \int\limits_{-\infty}^{0} 0dx + \int\limits_{0}^{1} kx^2dx + \int\limits_{1}^{2} k(2-x)dx + \int\limits_{2}^{\infty} 0dx \\ &= 0 + k\int\limits_{0}^{1} x^2dx + k\int\limits_{1}^{2}(2-x)dx + 0 \\ &= k\frac{x^3}{3}|_0^1 + k\left(2x - \frac{x^2}{2}\ |_1^2\right) \\ &= \frac{5k}{6} \text{ and therefore } k = \frac{6}{5}\end{aligned}$$

(b) Let us start with the easy pieces (which are unfortunately often left out) first:

$F(x) = P(X \leq x) = 0$ if $x \leq 0$

$F(x) = P(X \leq x) = 1$ if $x \geq 2$ since the probability density function equals 0 for all $x \geq 2$

By looking at Figure 8.6 we have

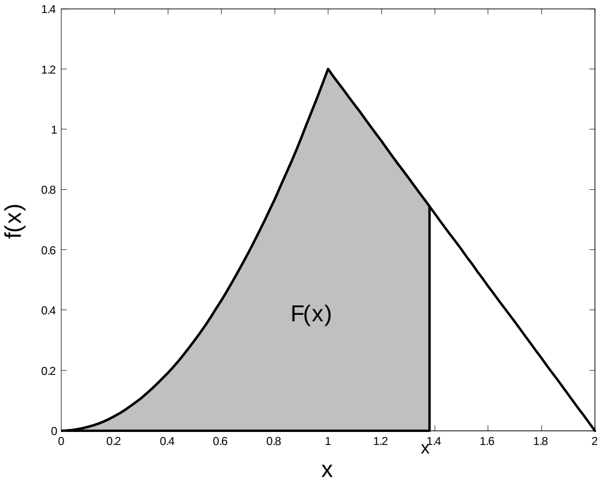


Figure 8.6: Area of shaded region equals $F(x) = P(X \leq x)$

$$F(x) = P(X \leq x) = \int\limits_{-\infty}^{x} f(z)\,dz = 0 + \int\limits_{0}^{x} \frac{6}{5}z^2dz = \frac{6}{5}\frac{x^3}{3}\ |_0^x = \frac{2x^3}{5} \quad \text{if } 0 < x < 1$$

$$F(x) = P\left(X \le x\right) = 0 + \int_0^1 \frac{6}{5} z^2 dz + \int_1^x \frac{6}{5}\left(2 - z\right) dz = \frac{6}{5}\frac{x^3}{3}|_0^1 + \frac{6}{5}(2x - \frac{x^2}{2})|_1^x$$
$$= \frac{12x - 3x^2 - 7}{5} \quad \text{if } 1 < x < 2$$

Therefore

$$F(x) = P\left(X \le x\right) = \begin{cases} 0 & x \le 0 \\ \frac{2x^3}{5} & 0 < x \le 1 \\ \frac{12x-3x^2-7}{5} & 1 < x < 2 \\ 1 & x \ge 2 \end{cases}$$

As a rough check, since for a continuous distribution there is no probability at any point, $F(x)$ should have the same value as we approach each boundary point from above and from below. For example,

$$\begin{array}{ll} \text{as } x \to 0^+, & \frac{2x^3}{5} \to 0 \\ \text{as } x \to 1^-, & \frac{2x^3}{5} \to \frac{2}{5} \\ \text{as } x \to 1^+, & \frac{12x-3x^2-7}{5} \to \frac{2}{5} \\ \text{as } x \to 2^-, & \frac{12x-3x^2-7}{5} \to 1 \end{array}$$

This quick check won't prove your answer is right, but will detect many careless errors.

(c)

$$P\left(0.5 < X < 1.5\right) = \int_{0.5}^{1.5} f(x)dx = F\left(1.5\right) - F\left(0.5\right)$$
$$= \frac{12\left(1.5\right) - 3\left(1.5\right)^2 - 7}{5} - \frac{2\left(0.5\right)^3}{5} = 0.8$$

Definition 21 ***Quantiles and Percentiles:*** *Suppose X is a continuous random variable with cumulative distribution function $F\left(x\right)$. The pth quantile of X (or the pth quantile of the distribution) is the value $q\left(p\right)$, such that $P\left[X \le q\left(p\right)\right] = p$. The value $q\left(p\right)$ is also called the $100p$th percentile of the distribution. If $p = 0.5$ then $m = q\left(0.5\right)$ is called the median of X or the median of the distribution.*

Example: For the example above find:

(a) the 0.4 quantile (40th percentile) of the distribution

(b) the median of the distribution

Solution:

(a) Since $F(1) = 0.4$, the 0.4 quantile is equal to 1.

(b) The median is the solution to $F(x) = \frac{12x-3x^2-7}{5} = 0.5$ or $24x - 6x^2 - 19 = 0$ which has two solutions. Since $F(1) = 0.4$ we know that the median must lie between 1 and 2 and we choose the solution $x \approx 1.087$. The median is approximately equal to 1.087.

Defined Variables or Change of Variable:
When we know the probability density function or cumulative distribution function for a continuous random variable X we sometimes want to find the probability density function or cumulative distribution function for some other random variable Y which is a function of X. The procedure for doing this is summarized below. It is based on the fact that the cumulative distribution function $F_Y(y)$ for Y equals $P(Y \leq y)$, and this can be rewritten in terms of X since Y is a function of X. Thus:

1) Write the cumulative distribution function of Y as a function of X.

2) Use $F_X(x)$ to find $F_Y(y)$. Then if you want the probability density function $f_Y(y)$, you can differentiate the expression for $F_Y(y)$.

3) Find the range of values of y.

Example: In the earlier spinner example,

$$f(x) = \begin{cases} \frac{1}{4} & 0 < x \leq 4 \\ 0 & \text{otherwise} \end{cases}$$

and

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

Find the probability density function of $Y = X^{-1}$.

Solution:
Step 1 above becomes:

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P\left(X^{-1} \leq y\right) \\ &= P\left(X \geq y^{-1}\right) = 1 - P\left(X < y^{-1}\right) \\ &= 1 - F_X\left(y^{-1}\right) \end{aligned}$$

For step (2), we can substitute $\frac{1}{y}$ in place of x in $F_X(x)$ giving:

$$F_Y(y) = 1 - \frac{y^{-1}}{4} = 1 - \frac{1}{4y}$$

and then differentiate to obtain the probability density function

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{1}{4y^2} \text{ for } y \geq \frac{1}{4}$$

(Note that as x goes from 0 to 4, $y = \frac{1}{x}$ goes between ∞ and $\frac{1}{4}$.)

Alternatively, and a little more generally, we can use the chain rule:

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) = \frac{d}{dy} \left[1 - F_X\left(y^{-1}\right)\right] \\ &= -f_X\left(y^{-1}\right) \frac{d}{dy}\left(y^{-1}\right) \text{ since } \frac{d}{dx} F_X(x) = f_X(x) \\ &= -f_X\left(y^{-1}\right)\left(-y^{-2}\right) = -\frac{1}{4}\left(-y^{-2}\right) \\ &= \frac{1}{4y^2} \text{ for } y \geq \frac{1}{4} \end{aligned}$$

Generally if $F_X(x)$ is known in some easy form, it is easier to substitute first, then differentiate. If $F_X(x)$ is more complicated, for example an integral that can't be easily found, it is usually easier to differentiate first, then substitute $f_X(x)$.

Expectation, Mean, and Variance for Continuous Random Variables

Definition 22 *When X is a continuous random variable we define*

$$E\left[g(X)\right] = \int\limits_{-\infty}^{\infty} g(x) f(x) dx$$

Note that this is analogous to the definition for discrete random variables $E\left[g(X)\right] = \sum\limits_{\text{all } x} g(x) f(x)$. With this definition, all of the earlier properties of expected value and variance still hold. For example with $\mu = E(X)$,

$$\sigma^2 = Var(X) = E[(X-\mu)^2] = E(X^2) - \mu^2 = E\left(X^2\right) - \left[E\left(X\right)\right]^2$$

(This definition can be justified by writing $\int\limits_{-\infty}^{\infty} g(x) f(x) dx$ as a limit of a Riemann sum and recognizing the Riemann sum as being in the form of an expected value for discrete random variables.)

Example: For the earlier spinner example,

$$f(x) = \begin{cases} \frac{1}{4} & 0 < x \leq 4 \\ 0 & \text{otherwise} \end{cases} .$$

Find $E(X)$ and $Var(X)$.

Solution:

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx = 0 + \int_{0}^{4} x \frac{1}{4} dx + 0 = \frac{1}{4} \left(\frac{x^2}{2} \right) |_0^4 = 2$$

Since

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x) dx = 0 + \int_{0}^{4} x^2 \frac{1}{4} dx + 0 = \frac{1}{4} \left(\frac{x^3}{3} \right) |_0^4 = \frac{16}{3}$$

then

$$Var(X) = E(X^2) - [E(X)]^2 = \frac{16}{3} - (2)^2 = \frac{4}{3}$$

Example: Let X have probability density function

$$f(x) = \begin{cases} \frac{6x^2}{5} & 0 < x \leq 1 \\ \frac{6}{5}(2 - x) & 1 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

Find $E(X)$ and $Var(X)$.

Solution:

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f(x) dx = 0 + \int_{0}^{1} x \frac{6}{5} x^2 dx + \int_{1}^{2} x \frac{6}{5}(2 - x) dx + 0 \\ &= \frac{6}{5} \left[\frac{x^4}{4} |_0^1 + \left(x^2 - \frac{x^3}{3} \right) |_1^2 \right] \\ &= \frac{11}{10} = 1.1 \end{aligned}$$

Since

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} x^2 f(x) dx = 0 + \int_{0}^{1} x^2 \frac{6}{5} x^2 dx + \int_{1}^{2} x^2 \frac{6}{5}(2 - x) dx + 0 \\ &= \frac{6}{5} \left(\frac{x^5}{5} |_0^1 + 2 \left(\frac{x^3}{3} \right) |_1^2 - \frac{x^4}{4} |_1^2 \right) \\ &= \frac{67}{50} \end{aligned}$$

therefore

$$\begin{aligned} Var\left(X\right) &= E\left(X^2\right) - \left[E\left(X\right)\right]^2 \\ &= \frac{67}{50} - \left(\frac{11}{10}\right)^2 \\ &= \frac{13}{100} \\ &= 0.13 \end{aligned}$$

Problems

8.1.1 Let X be a continuous random variable with probability density function

$$f(x) = \begin{cases} kx^2 & -1 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Find

(a) k

(b) the cumulative distribution function, $F(x)$

(c) $P\left(-0.1 < X \le 0.2\right)$

(d) the mean and variance of X.

(e) the probability density function of $Y = X^2$.

8.1.2 Let X be a continuous random variable with cumulative distribution function

$$F(x) = \begin{cases} \frac{kx^n}{1+x^n} & x > 0, \;\; n > 0 \\ 0 & x \le 0 \end{cases}$$

(a) Find k.

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

(c) Find the median of the distribution.

8.2 Continuous Uniform Distribution

Just as we did for discrete random variables, we now consider some special types of continuous probability distributions. These distributions arise in certain settings, described below. This section considers what we call Uniform distributions.

Physical Setup:
Suppose X takes values in some interval $[a, b]$ (it doesn't actually matter whether the interval is open or closed) with all subintervals of a fixed length being equally likely. Then X has a **continuous Uniform distribution**. We write $X \sim Uniform\,(a, b)$ where $b > a$ and $a, b \in \Re$. The parameters of the distribution are a and b.

Illustrations:
(1) In the spinner example $X \sim Uniform\,(0, 4)$.
(2) Computers can generate a random number X which appears as though it is drawn from the distribution $Uniform\,(0, 1)$. This is the starting point for many computer simulations of random processes; an example is given below.

The probability density function and the cumulative distribution function:
Since all points are equally likely (more precisely, intervals contained in $[a, b]$ of a given length, say 0.01, all have the same probability), the probability density function must be a constant function $f(x) = k$ for all $a \leq x \leq b$ for some constant k. Now

$$\begin{aligned} 1 &= \int\limits_a^b f(x)dx = \int\limits_a^b kdx = k\int\limits_a^b dx = k\left(x|_a^b\right) \\ &= k\,(b-a) \end{aligned}$$

and therefore $k = \frac{1}{b-a}$.

The probability density function is

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

which is pictured in Figure 8.7.

It easy to verify that $\int\limits_{-\infty}^{\infty} f(x)dx = 1$ since the area of the shaded region is a rectangle of area $(b-a)\left(\frac{1}{b-a}\right) = 1$.

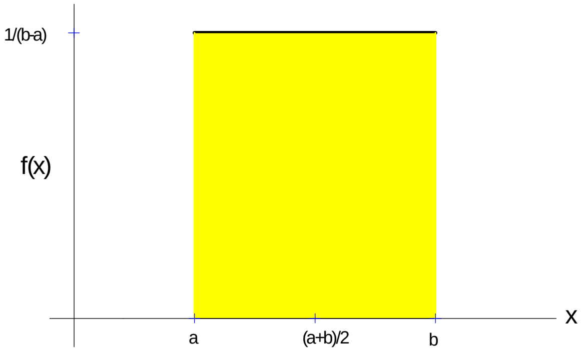


Figure 8.7: Probability density function for a $Uniform\,(a,b)$ random variable

The cumulative distribution function pictured in Figure 8.8 is

$$F(x) = \begin{cases} 0 & x < a \\ \int\limits_a^x \frac{1}{b-a}dx = \frac{x-a}{b-a} & a \le x \le b \\ 1 & x > b \end{cases}$$

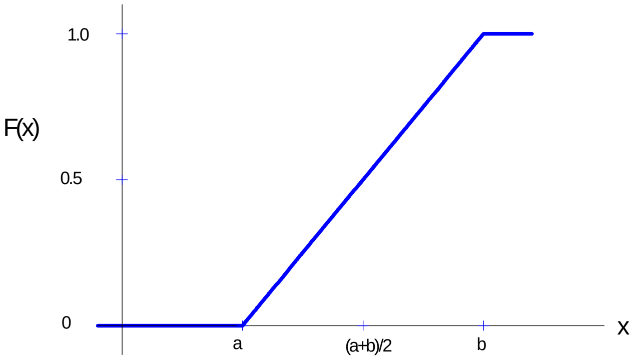


Figure 8.8: The cumulative distribution function for a $Uniform\,(a.b)$ random variable

Mean and Variance:

The mean of a $Uniform\,(a,b)$ random variable can easily be determined by noting that the graph of the probability density function is symmetric about the line $x=(a+b)/2$. Since the integral $\int\limits_a^b xdx$ exists (why?) therefore $E\,(X)$ exists and by symmetry $E\,(X)=(a+b)/2$.

To determine $Var\,(X)$ we note that

$$\begin{aligned} E(X^2) &= \int\limits_{-\infty}^{\infty} x^2 f(x)dx = \int\limits_a^b x^2 \frac{1}{b-a}dx = \frac{1}{(b-a)}\left(\frac{x^3}{3}\,|_a^b\right) = \frac{b^3-a^3}{3(b-a)} \\ &= \frac{(b-a)\left(b^2+ab+a^2\right)}{3(b-a)} = \frac{b^2+ab+a^2}{3} \end{aligned}$$

and therefore

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

In summary

$$\text{If } X \sim Uniform\,(a,b) \text{ then } E\,(X) = \frac{a+b}{2} \text{ and } Var\,(X) = \frac{(b-a)^2}{12}$$

Example: Suppose X has the continuous probability density function

$$f(x) = 0.1e^{-0.1x} \quad \text{for } x > 0$$

and zero otherwise. (This is called an Exponential distribution and is discussed in the next section. It is used in areas such as queueing theory and reliability.) We will show that the new random variable

$$Y = e^{-0.1X}$$

has a $Uniform\,(0,1)$ distribution. To see this, we follow the steps in Section 8.1:

$$\begin{aligned} F_Y(y) &= P(Y \le y) \\ &= P(e^{-0.1X} \le y) \\ &= P(X \ge -10\ln y) \\ &= 1 - P(X < -10\ln y) \\ &= 1 - F_X(-10\ln y) \end{aligned}$$

Since for $x > 0$

$$F_X(x) = \int_0^x 0.1e^{-0.1u}du = 1 - e^{-0.1x}$$

we have

$$F_Y(y) = 1 - \left[1 - e^{-0.1(-10\ln y)}\right] = y \quad \text{for } 0 < y < 1$$

The range of Y is $(0, 1)$ since the range of X is $(0, \infty)$. Thus

$$f_Y(y) = \begin{cases} \frac{d}{dy}F_Y(y) = 1 & 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

which implies $Y \sim Uniform(0, 1)$.

Many computer software systems have "random number generator" functions that will simulate observations Y from a $Uniform(0, 1)$ distribution. These are more properly called **pseudo-random number generators** because they are based on deterministic algorithms. In addition they give observations Y that have finite precision so they cannot be **exactly** like continuous $Uniform(0, 1)$ random variables. However, good generators give Y's that appear indistinguishable in most ways from $Uniform(0, 1)$ random variables. Given such a generator, we can also simulate random variables X with the Exponential distribution above by the following algorithm:

1. Generate $Y \sim Uniform(0, 1)$ using the computer random number generator.

2. Compute $X = -10\ln Y$.

Then X has the desired distribution. This is a particular case of a method described in Section 8.4 for generating random variables from a general distribution. In R software the command `runif(n)` produces a vector consisting of n independent $Uniform(0, 1)$ values.

Problem

8.2.1 If X has cumulative distribution function $F(x)$, then show $Y = F(X)$ has a $Uniform(0, 1)$ distribution. Suppose you want to simulate observations from a distribution with probability density function $f(x) = 1.5x^2$ for $-1 < x < 1$ and zero otherwise, by using the random number generator on a computer to generate $Uniform(0, 1)$ numbers. What value would X take when you generate the random number $y = 0.27125$?

8.3 Exponential Distribution

The continuous random variable X is said to have an **Exponential distribution** if its probability density function is of the form

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\lambda \in \Re^+$. This distribution arises in various problems involving the time until some event occurs. The following gives one such setting.

Physical Setup:
In a Poisson process for events in time let X be the length of time we wait for the first event occurrence. We'll show that X has an Exponential distribution. (Recall that the number of occurrences in a fixed time has a Poisson distribution. The difference between the Poisson and Exponential distributions lies in what is being measured.)

Illustrations:
(1) The length of time X we wait with a Geiger counter until the emission of a radioactive particle is recorded follows an Exponential distribution.
(2) The length of time between phone calls to a fire station (assuming calls follow a Poisson process) follows an Exponential distribution.

Probability density function and the cumulative distribution function:
For $x > 0$

$$\begin{aligned} F(x) = P(X \leq x) &= P(\text{time to } 1^{\text{st}} \text{ occurrence } \leq x) \\ &= 1 - P(\text{time to } 1^{\text{st}} \text{ occurrence } > x) \\ &= 1 - P(\text{no occurrence in the interval } (0, x)) \end{aligned}$$

Check that you understand this last step. If the time to the first occurrence is greater than x, then there must be no occurrences in $(0, x)$, and vice versa.

We have now expressed $F(x)$ in terms of the number of occurrences in a Poisson process by time x. But the number of occurrences has a Poisson distribution with mean $\mu = \lambda x$, where λ is the average rate of occurrence. Therefore

$$F(x) = \begin{cases} 1 - \frac{(\lambda x)^0 e^{-\lambda x}}{0!} = 1 - e^{-\lambda x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Also since $\frac{d}{dx}\left(1 - e^{-\lambda x}\right) = \lambda e^{-\lambda x}$ we have

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

which is the formula we gave above.

Alternate Form: It is common to use the parameter $\theta = 1/\lambda$ in the Exponential distribution. (We'll see below that $\theta = E(X)$.) This gives

$$F(x) = \begin{cases} 1 - e^{-x/\theta} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

and

$$f(x) = \begin{cases} \frac{1}{\theta} e^{-x/\theta} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

where $\theta > 0$. We write $X \sim Exponential\,(\theta)$. The parameter of this distribution is θ.

A graph of the probability density function $f\,(x)$ is given in Figure 8.9. It obvious from the graph why this distribution is called the Exponential distribution. The distribution is said to be positively skewed (skewed to the right) or have a long right tail.

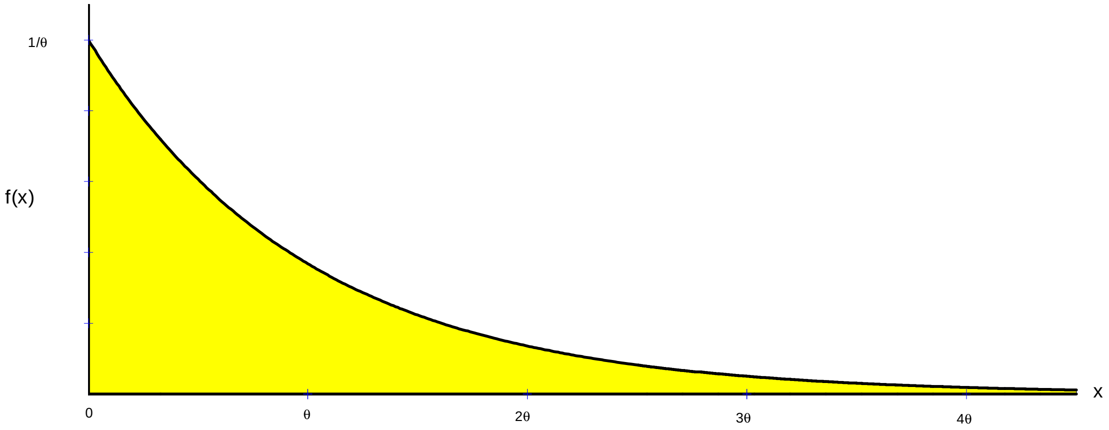


Figure 8.9: Graph of the probability density function of a $Exponential\,(\theta)$ random variable

A graph of the cumulative distribution function is given in Figure.8.10.

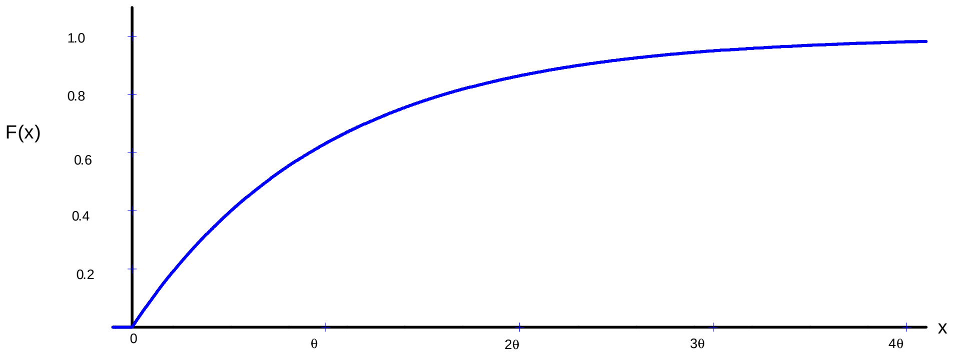


Figure 8.10: Cumulative distribution function for a $Exponential\left(\theta\right)$ random variable

Exercise:

Suppose trees in a forest are distributed according to a Poisson process. Let X be the distance from an arbitrary starting point to the nearest tree. The average number of trees per square metre is λ. Derive $f(x)$ the same way we derived the Exponential probability density function. You are now using the Poisson distribution in two dimensions (area) rather than one dimension (time).

Mean and Variance:

Finding $E\left(X\right)$ and $Var\left(X\right)$ directly involves integration by parts. An easier solution uses properties of the **gamma function**, which extends the notion of factorials beyond the integers to the positive real numbers.

Definition 23 ***The Gamma Function:***

$$\Gamma(\alpha) = \int\limits_0^\infty y^{\alpha-1}e^{-y}dy$$

is called the gamma function of α, where $\alpha > 0$.

Note that α is 1 more than the power of y in the integrand. For example,

$$\Gamma(5) = \int\limits_0^\infty y^4e^{-y}dy$$

Properties of the gamma function:

1. $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$ for $\alpha > 1$

   Proof: Using integration by parts,

   $$\int_0^\infty y^{\alpha-1} e^{-y} dy = -\lim_{y\to\infty} y^{\alpha-1} e^{-y} + (\alpha - 1)\int_0^\infty y^{\alpha-2} e^{-y} dy$$

   and provided that $\alpha > 1$, $\lim\limits_{y\to\infty} y^{\alpha-1} e^{-y} = 0$. Therefore

   $$\int_0^\infty y^{\alpha-1} e^{-y} dy = (\alpha - 1)\int_0^\infty y^{\alpha-2} e^{-y} dy = (\alpha - 1)\Gamma(\alpha - 1)$$

2. $\Gamma(\alpha) = (\alpha - 1)!$ if α is a positive integer.

   Proof: It is easy to show that $\Gamma(1) = 1$. Using property 1 repeatedly, we obtain

   $$\begin{aligned} \Gamma(2) &= 1\Gamma(1) = 1 \\ \Gamma(3) &= 2\Gamma(2) = 2! \\ \Gamma(4) &= 3\Gamma(3) = 3! \quad \text{etc.} \end{aligned}$$

   In general, $\Gamma(n + 1) = n!$ for $n = 0, 1, \ldots$

3. $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$

   (This can be proved using double integration.)

Returning to the Exponential distribution we have:

$$\begin{aligned} E(X) &= \int_{-\infty}^\infty xf(x)dx = \int_0^\infty x\frac{1}{\theta}e^{-x/\theta}dx \quad \text{let } y = \frac{x}{\theta} \text{ with } dx = \theta dy \\ &= \int_0^\infty ye^{-y}\theta dy = \theta\int_0^\infty y^1 e^{-y} dy = \theta\Gamma(2) = \theta\,(1!) \\ &= \theta \end{aligned}$$

Note: Read questions carefully. If you are given the average **rate** of occurrence in a Poisson process, then this is the parameter λ. If you are given the average waiting **time** for an occurrence, then this is the parameter θ.

To obtain $Var(X)$, we first find

$$E\left(X^2\right) = \int_{-\infty}^{\infty} x^2 f(x)dx = \int_{0}^{\infty} x^2 \frac{1}{\theta} e^{-x/\theta} dx \text{ let } y = x/\theta$$

$$= \int_{0}^{\infty} \theta^2 y^2 \frac{1}{\theta} e^{-y} \theta dy = \theta^2 \int_{0}^{\infty} y^2 e^{-y} dy = \theta^2 \Gamma(3) = \theta^2 (2!) = 2\theta^2$$

Then

$$Var(X) = E\left(X^2\right) - \left[E\left(X\right)\right]^2 = 2\theta^2 - \theta^2 = \theta^2$$

In summary

$$\text{If } X \sim Exponential\,(\theta) \text{ then } E\left(X\right) = \theta \text{ and } Var\left(X\right) = \theta^2$$

In Figure 8.11, the probability density functions for different values of θ are pictured to see the effect of changing θ.

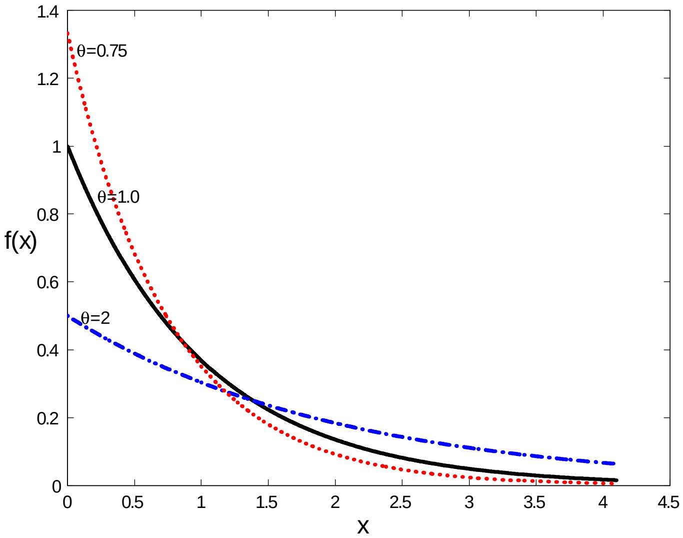


Figure 8.11: Exponential probability density functions for different values of θ

Example:

Suppose buses arrive at a bus stop according to a Poisson process with an average of 5 buses per hour. ($\lambda = 5$/hour so $\theta = 1/5$ hour or 12 minutes).

Find the probability:

(a) you have to wait longer than 15 minutes for a bus

(b) you have to wait more than 15 minutes longer, having already waited for 6 minutes.

Solution:

(a)

$$P(X > 15) = 1 - P(X \leq 15) = 1 - F(15) = 1 - \left(1 - e^{-15/12}\right) = e^{-1.25} = 0.2865$$

(b) If X is the total waiting time, the question asks for the probability

$$\begin{aligned} P(X > 21 | X > 6) &= \frac{P(X > 21 \text{ and } X > 6)}{P(X > 6)} \\ &= \frac{P(X > 21)}{P(X > 6)} \end{aligned}$$

$$\begin{aligned} &= \frac{1 - \left(1 - e^{-21/12}\right)}{1 - \left(1 - e^{-6/12}\right)} \\ &= \frac{e^{-21/12}}{e^{-6/12}} \\ &= e^{-1.25} = 0.2865 \end{aligned}$$

Does this surprise you? The fact that you're already waited 6 minutes doesn't seem to matter.

Memoryless Property of the Exponential Distribution

The example above illustrates the "memoryless property" of the Exponential distribution:

$$P(X > c + b | X > b) = P(X > c)$$

In other words for a Poisson process , given that you have waited b units of time for the next event, the probability you wait an additional c units of time does not depend on b but only depends on c.

Fortunately, buses don't follow a Poisson process so the example need not cause you to stop using the bus.

Problems

8.3.1 In a bank with on-line terminals, the time the system runs between disruptions has an Exponential distribution with mean θ hours. One quarter of the time the system shuts down within 8 hours of the previous disruption. Find θ.

8.3.2 Flaws in painted sheets of metal occur over the surface according to the conditions for a Poisson process, at an intensity of λ per m^2. Let X be the distance from an arbitrary starting point to the second closest flaw. (Assume sheets are of infinite size!)

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

(b) What is the average distance to the second closest flaw?

8.4 A Method for Computer Generation of Random Variables

[30]Most computer software has a built-in "pseudo-random number[31] generator" that will simulate observations U from a $Uniform(0,1)$ distribution, or at least a reasonable approximation to this Uniform distribution. If we wish a random variable with a non-Uniform distribution, the standard approach is to take a suitable function of U. By far the simplest and most common method for generating non-Uniform variates is based on the inverse cumulative distribution function. For arbitrary cumulative distribution function $F(x)$, define $F^{-1}(y) = min\{x; F(x) \geq y\}$. This is a real inverse (that is, $F(F^{-1}(y)) = F^{-1}(F(y)) = y$) in the case that the cumulative distribution function is continuous and strictly increasing. However, in the more general case of a possibly discontinuous non-decreasing cumulative distribution function (such as the cumulative distribution function of a discrete distribution) the function continues to enjoy at least some of the properties of an inverse. F^{-1} is useful for generating a random variables having cumulative distribution function $F(x)$ from U, a Uniform random variable on the interval $[0,1]$.

Theorem 24 *If F is an arbitrary cumulative distribution function and $U \sim Uniform\,(0,1)$, then the random variable defined by $X = F^{-1}(U)$ has cumulative distribution function $F(x)$.*

Proof:
The proof is a consequence of the fact that

$$[U < F(x)] \subset [X \leq x] \subset [U \leq F(x)] \text{ for all } x$$

[30] This section optional for STAT 220.

[31] "The generation of random numbers is too important to be left to chance." Robert R. Coveyou, Oak Ridge National Laboratory

You can check this graphically be checking, for example, that if $[U < F(x)]$ then $[F^{-1}(U) \leq x]$ (this confirms the left hand "$\subset$"). Taking probabilities on all sides of this, and using the fact that $P[U \leq F(x)] = P[U < F(x)] = F(x)$, we discover that $P\left(X \leq x\right) = F(x)$.

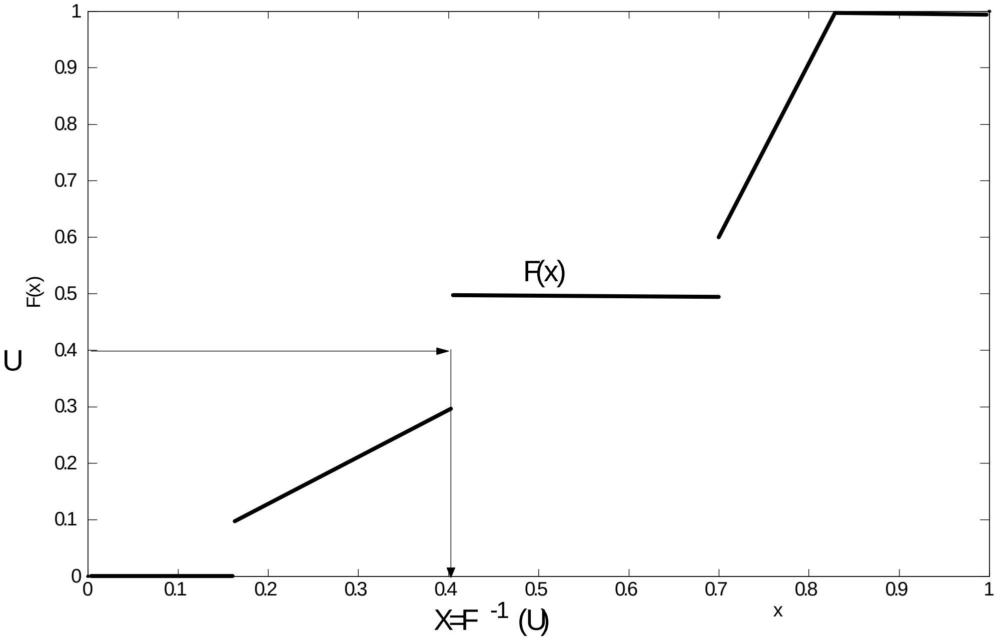


Figure 8.12: Inverting a Cumulative Distribution Function

The relation $X = F^{-1}(U)$ implies that $F(X) \geq U$ and for any point $z < X$, $F(z) < U$. For example, for the rather unusual looking piecewise linear cumulative distribution function in Figure 8.12, we find the solution $X = F^{-1}(U)$ by drawing a horizontal line at U until it strikes the graph of the cumulative distribution function (or where the graph would have been if we had joined the ends at the jumps) and then X is the $x - coordinate$ of this point. This is true in general, X is the coordinate of the point where a horizontal line first strikes the graph of the cumulative distribution function We provide one simple example of generating random variables by this method, for the Geometric distribution.

Example: A Geometric random number generator

For the Geometric distribution, the cumulative distribution function is given by $F(x) = 1-(1-p)^{x+1}$ for $x = 0, 1, 2, \ldots$. Then if U is a Uniform random number in the interval $[0, 1]$, we seek an integer X such that $F(X-1) < U \leq F(X)$. (You should confirm that this is the value of X at which the above horizontal line strikes the graph of the cumulative distribution function) and solving these inequalities gives

$$\begin{aligned} 1-(1-p)^X < U &\leq 1-(1-p)^{X+1} \\ (1-p)^X > 1-U &\geq (1-p)^{X+1} \\ X\ln(1-p) > \ln(1-U) &\geq (X+1)\ln(1-p) \\ X < \frac{\ln(1-U)}{\ln(1-p)} &\leq X+1 \end{aligned}$$

so we compute the value of $\ln(1-U)/\ln(1-p)$ and round down to the next lower integer.

Exercise: An Exponential random number generator.

For the Exponential(θ) distribution show that the inverse transform method above results in

$$X = -\theta\ln(1-U)$$

8.5 Normal Distribution

Physical Setup:
A random variable X has a Normal[32] distribution if it has probability density function of the form

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{for } x \in \Re$$

where $\mu \in \Re$ and $\sigma \in \Re^+$ are parameters of the distribution. It turns out (and is shown below) that $E(X) = \mu$ and $Var(X) = \sigma^2$ for this distribution; that is why its probability density function is written using the symbols μ and σ^2.

We write $X \sim N(\mu, \sigma^2)$ to denote that X has a Normal distribution with mean μ and variance σ^2 (standard deviation σ). The parameters of the distribution are μ and σ^2.

The Normal distribution is the most widely used distribution in probability and statistics. Physical processes leading to the Normal distribution exist but are a little complicated to describe. (For example, it arises in physics via statistical mechanics and maximum entropy arguments.) It is used for many processes where X represents a physical dimension of some kind, but also in many other settings.

Illustrations:
(1) Heights or weights of males (or of females) in large populations tend to follow a Normal distribution.
(2) The logarithms of stock prices are often assumed to have a Normal distribution.

[32] "The only normal people are the ones you don't know very well." Joe Ancis

The graph of the probability density function $f(x)$, shown in Figure 8.13, is symmetric about the line $x = \mu$. The shape of the probability density function is often termed a "bell shape" or "bell curve". You should be able to verify the shape of the function using the first and second derivatives of $f\,(x)$.

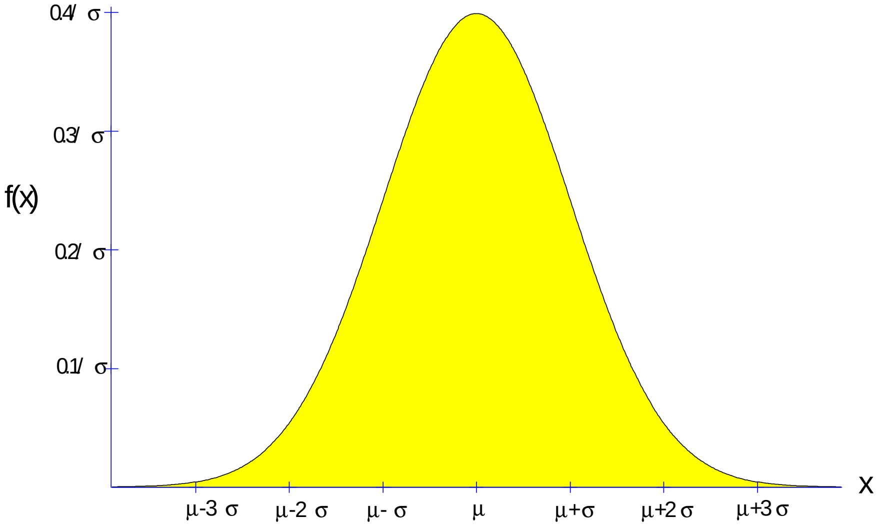


Figure 8.13: The $N(\mu, \sigma^2)$ probability density function

We can show that $f(x)$ integrates to 1:

$$
\begin{aligned}
&\int\limits_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx \quad \text{let } z = (x-\mu)/\sigma \\
&= \int\limits_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 2\int\limits_{0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \quad \text{let } y = \frac{1}{2}z^2 \text{ and } dz = \frac{dy}{\sqrt{2}y^{\frac{1}{2}}} \\
&= 2\int\limits_{0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y} \frac{dy}{\sqrt{2}y^{\frac{1}{2}}} \\
&= \frac{1}{\sqrt{\pi}} \int\limits_{0}^{\infty} y^{-\frac{1}{2}} e^{-y} dy \\
&= \frac{1}{\sqrt{\pi}} \Gamma(\frac{1}{2}) \quad \text{where } \Gamma \text{ is the gamma function} \\
&= 1 \quad \text{since } \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}
\end{aligned}
$$

The cumulative distribution function:

The cumulative distribution function of the Normal distribution $N(\mu, \sigma^2)$ is

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy \quad \text{for } x \in \Re$$

as shown in Figure 8.14. This integral cannot be given a simple mathematical expression so numerical methods are used to compute its value for given values of x, μ and σ. This function is included in many software packages and some calculators.

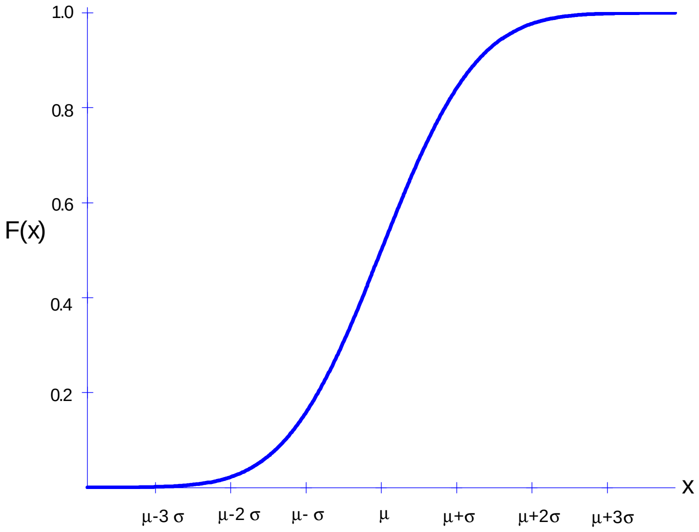


Figure 8.14: The Normal cumulative distribution function

In the statistical packages R we get $F(x)$ above using the function pnorm(x,μ,σ). Before computers, people needed to produce tables of probabilities $F(x)$ by numerical integration, using mechanical calculators. Fortunately it is necessary to do this only for a single Normal distribution: the one with $\mu = 0$ and $\sigma = 1$. This is called the **“standard” Normal distribution** and is denoted $N(0, 1)$.

It is easy to see that if $X \sim N(\mu, \sigma^2)$ then the “new” random variable $Z = (X - \mu)/\sigma$ is distributed as $Z \sim N(0, 1)$. (Just use the change of variables methods in Section 8.1.) We’ll use this result to compute probabilities for X, and to show that $E(X) = \mu$ and $Var(X) = \sigma^2$.

Mean:

Recall that an odd function, $f(x)$, has the property that $f(-x) = -f(x)$. If $f(x)$ is an odd function then $\int\limits_{-\infty}^{\infty} f(x)dx = 0$, provided the integral exists.

Consider

$$E\left(X-\mu\right)=\int\limits_{-\infty}^{\infty}\left(x-\mu\right)\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^{2}}dx$$

Let $y = x - \mu$. Then

$$E\left(X-\mu\right)=\int\limits_{-\infty}^{\infty} y\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^{2}}dy$$

where the integrand is an odd function so that $E\left(X-\mu\right) = 0$. But since $E\left(X-\mu\right) = E(X) - \mu$, this implies $E(X) = \mu$ and so μ is the mean of the $N(\mu, \sigma^2)$ distribution.

Variance:

To obtain the variance we have

$$\begin{aligned}
Var(X) &= E\left[\left(X-\mu\right)^{2}\right] \\
&= \int\limits_{-\infty}^{\infty}\left(x-\mu\right)^{2}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^{2}}dx \\
&= 2\int\limits_{\mu}^{\infty}(x-\mu)^{2}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^{2}}dx \quad \text{(since the function is symmetric about } \mu\text{)}.
\end{aligned}$$

We can obtain a gamma function by letting $y = \frac{(x-\mu)^2}{2\sigma^2}$ and noting that

$$\begin{aligned}
(x-\mu)^{2} &= 2\sigma^{2}y \;\text{ or }\; (x-\mu) = \sigma\sqrt{2y} \quad \text{since } x > \mu \\
dx &= \frac{\sigma\sqrt{2}dy}{2\sqrt{y}} = \frac{\sigma}{\sqrt{2y}}dy
\end{aligned}$$

Then

$$\begin{aligned}
Var(X) &= 2\int_0^\infty \left(2\sigma^2 y\right)\frac{1}{\sigma\sqrt{2\pi}}e^{-y}\left(\frac{\sigma}{\sqrt{2y}}dy\right) \\
&= \frac{2\sigma^2}{\sqrt{\pi}}\int_0^\infty y^{1/2}e^{-y}dy \\
&= \frac{2\sigma^2}{\sqrt{\pi}}\Gamma\left(\frac{3}{2}\right) = \frac{2\sigma^2}{\sqrt{\pi}}\left(\frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right) \\
&= \frac{\sigma^2\sqrt{\pi}}{\sqrt{\pi}} \quad \text{since } \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi} \\
&= \sigma^2
\end{aligned}$$

and so σ^2 is the variance of the $N(\mu, \sigma^2)$ distribution.

Finding Normal Probabilities Using $N(0,1)$ Tables: As noted above, $F(x)$ does not have an explicit closed form so numerical computation is needed. The following result shows that if we can compute the cumulative distribution function for the standard Normal distribution $N(0,1)$, then we can compute it for any other Normal distribution $N(\mu, \sigma^2)$ as well.

Theorem 25 *Let $X \sim N(\mu, \sigma^2)$ and define $Z = (X - \mu)/\sigma$. Then $Z \sim N(0,1)$ and*

$$P(X \leq x) = P\left(Z \leq \frac{x-\mu}{\sigma}\right)$$

Proof: The fact that $Z \sim N(0,1)$ has probability density function

$$\phi(z) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2} \quad \text{for } z \in \Re$$

follows immediately by change of variables.

Alternatively, we can just note that

$$\begin{aligned}
P\left(X \leq x\right) &= \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}dy \quad \text{let } \ z = \frac{y-\mu}{\sigma} \\
&= \int_{-\infty}^{(x-\mu)/\sigma} \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}dz \\
&= P\left(Z \leq \frac{x-\mu}{\sigma}\right)
\end{aligned}$$

A table of probabilities $P(Z \leq z)$ is given at the end of these Course Notes. A space-saving feature is that only the values for $z \geq 0$ are shown; for negative values we use the fact that $N(0, 1)$ probability density function is symmetric about 0.

The following examples illustrate how to get probabilities for Z using the tables.

Examples: Find the following probabilities, where $Z \sim N(0, 1)$.

(a) $P(Z \leq 2.11)$

(b) $P(Z < 3.40)$

(c) $P(Z > 1.06)$

(d) $P(Z \leq -1.06)$

(e) $P(-1.06 < Z \leq 2.11)$

Solution:

(a) Look up 2.11 in the table by going down the left column to 2.1 then across to the heading 0.01. We find the number 0.98257. Then $P(Z \leq 2.11) = 0.98257$. See Figure 8.15.

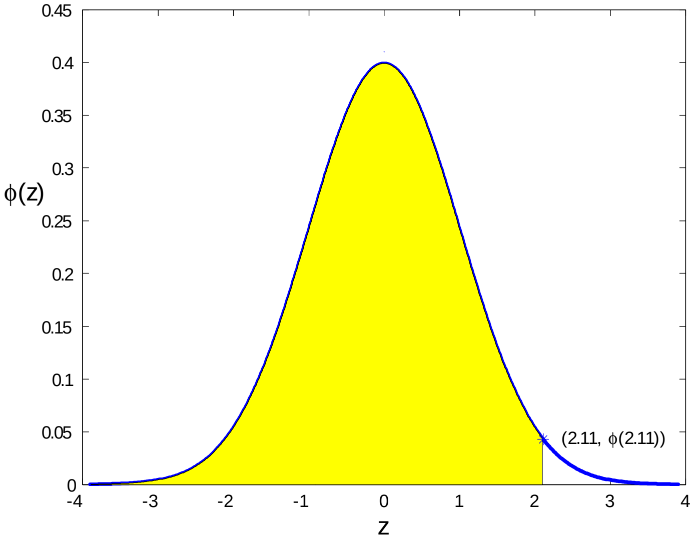


Figure 8.15: Area of shaded region equals $P(Z \leq 2.11) = 0.9826$

(b) $P\left(Z<3.40\right)=P\left(Z\leq 3.40\right)=0.99966$

(c) $P\left(Z>1.06\right)=1-P\left(Z\leq 1.06\right)=1-0.85543=0.14457$

(d) Now we have to use symmetry:

$$P\left(Z\leq -1.06\right)=P\left(Z>1.06\right)=1-P\left(Z\leq 1.06\right)=1-0.85543=0.14457$$

See Figure 8.16.

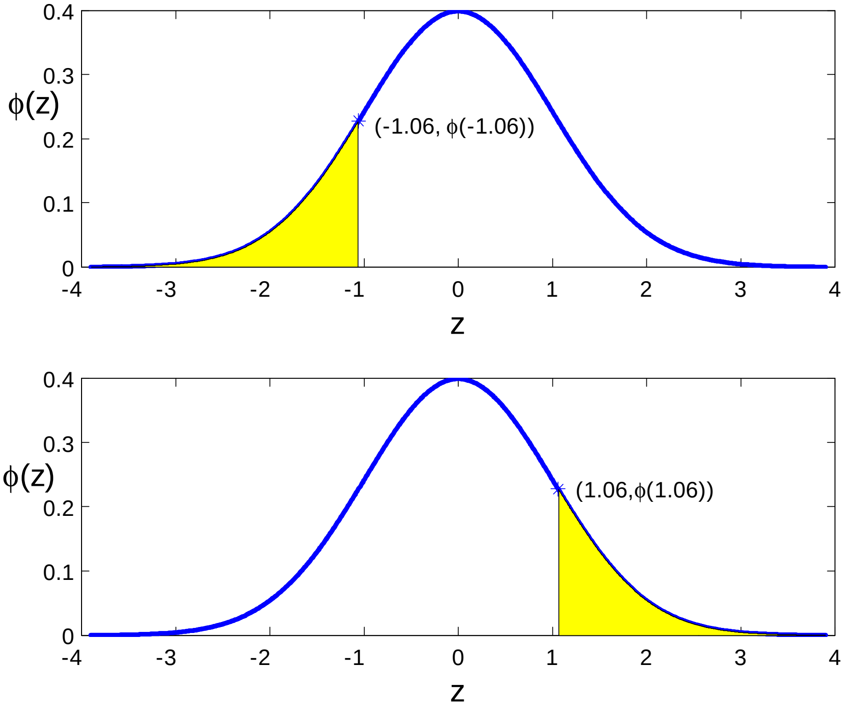


Figure 8.16: Calculation of $P\left(Z<-1.06\right)$ and $P\left(Z>1.06\right)$

(e)

$$\begin{aligned}P\left(-1.06<Z\leq 2.11\right)&=P(Z\leq 2.11)-P(Z\leq -1.06)\\&=P(Z\leq 2.11)-P\left(Z>1.06\right)\\&=P(Z\leq 2.11)-\left[1-P\left(Z\leq 1.06\right)\right]\\&=0.98257-(1-0.85543)=0.83800\end{aligned}$$

In addition to using the tables to find the probabilities for given numbers, we sometimes are given the probabilities and asked to find the number. With R, the function qnorm(p, μ, σ) gives the 100 pth percentile (where $0 < p < 1$). We can also use tables to find desired values.

Examples:

(a) Find a number c such that $P(Z \leq c) = 0.85$.

(b) Find a number d such that $P(Z > d) = 0.90$.

(c) Find a number b such that $P(|Z| \leq b) = 0.95$.

Solutions:

(a) We can look in the body of the table to get an entry close to 0.85. This occurs for z between 1.03 and 1.04; $z = 1.04$ gives the closest value to 0.85. For greater accuracy, the table at the bottom of the last page is designed for finding numbers, given the probability. Looking beside the entry 0.85 we find $z = 1.0364$.

(b) See Figure 8.17. Since $P(Z > d) = 0.90$ we have $P(Z \leq d) = 1 - P(Z > d) = 0.10$. There is no entry for which $P(Z \leq z) = 0.10$ so we again have to use symmetry, since d will be negative. From the table we have $P(Z \leq 1.2816) = 0.90$. By symmetry $P(Z > -1.2816) = 0.90$ and

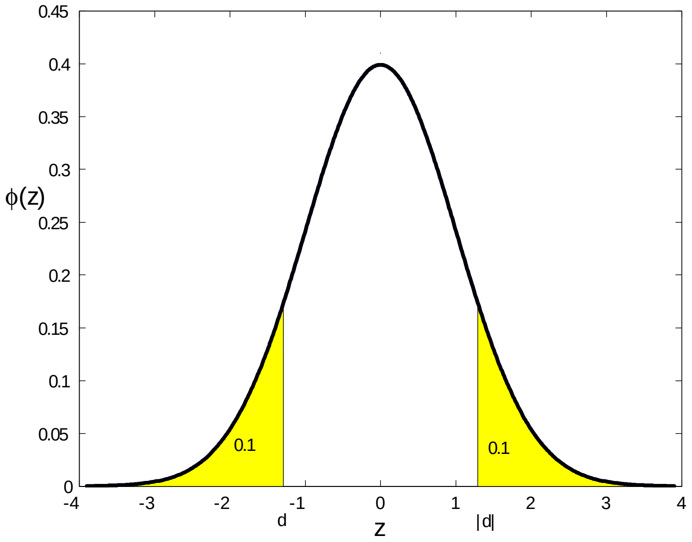


Figure 8.17: Picture to determine $P(Z > d) = 0.90$

therefore $d = -1.2816$.

The key to this solution lies in recognizing that d will be negative. If you can picture the situation it will probably be easier to handle the question than if you rely on algebraic manipulations.

Exercise: Will a be positive or negative if $P(Z > a) = 0.05$? What if $P(Z \leq a) = 0.99$?

(c) We first note that $P(|Z| \leq b) = P(-b < Z < b) = 0.95$. By symmetry, the probability outside the interval $(-b, b)$ must be 0.05, and this is evenly split between the area above b and the area below $-b$. Therefore (see Figure 8.18)

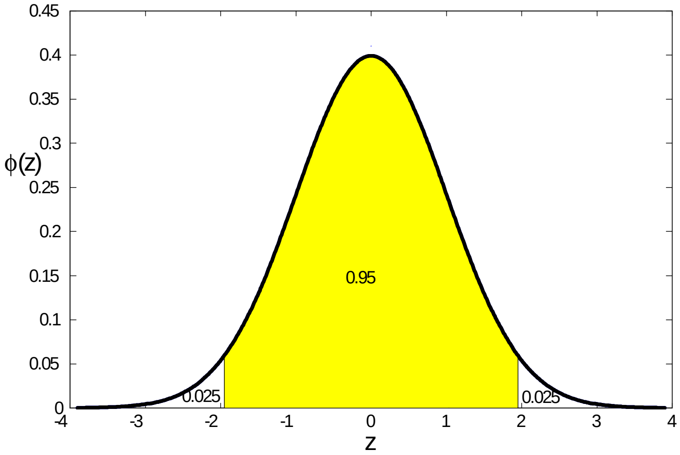


Figure 8.18: Picture to determine $P(-b < Z < b) = 0.95$

$$P(Z \leq -b) = P(Z > b) = 0.025$$

and

$$P(Z \leq b) = 0.975$$

Looking within the body of the top table, we can see $P(Z \leq 1.96) = 0.975$ so $b = 1.96$.

Exercise: Find b such that $P(|Z| \leq b) = 0.9$ and $P(|Z| \leq b) = 0.99$.

Finding $N(\mu, \sigma^2)$ probabilities: To find $N(\mu, \sigma^2)$ probabilities in general, we use the theorem given earlier, which implies that if $X \sim N(\mu, \sigma^2)$ then

$$\begin{aligned} P(a \leq X \leq b) &= P\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right) \\ &= P\left(Z \leq \frac{b-\mu}{\sigma}\right) - P\left(Z \leq \frac{a-\mu}{\sigma}\right) \end{aligned}$$

where $Z \sim N(0,1)$.

Example: Let $X \sim N(3, 25)$.

(a) Find $P\left(X < 2\right)$

(b) Find a number c such that $P\left(X > c\right) = 0.95$.

Solution:

(a)

$$\begin{aligned} P\left(X < 2\right) &= P\left(\frac{X-\mu}{\sigma} < \frac{2-3}{5}\right) \\ &= P\left(Z < -0.20\right) \text{ where } Z \sim N(0,1) \\ &= 1 - P\left(Z < 0.20\right) = 1 - 0.57926 = 0.42074 \end{aligned}$$

(b)

$$\begin{aligned} P\left(X > c\right) &= P\left(\frac{X-\mu}{\sigma} > \frac{c-3}{5}\right) \\ &= P\left(Z > \frac{c-3}{5}\right) \text{ where } Z \sim N(0,1) \\ &= 0.95 \end{aligned}$$

Therefore (see Figure (8.19)), $\left(c-3\right)/5 = -1.6449$ or $c = -5.2245$.

Gaussian Distribution: The Normal distribution is also known as the Gaussian[33] distribution. The notation $X \sim G(\mu, \sigma)$ means that X has Gaussian (Normal) distribution with mean μ and standard deviation σ. So, for example, if $X \sim N(1, 4)$ then we could also write $X \sim G(1, 2)$.

[33] After Johann Carl Friedrich Gauss (1777-1855), a German mathematician, physicist and astronomer, discoverer of Bode's Law, the Binomial Theorem and a regular 17-gon. He discovered the prime number theorem while an 18 year-old student and used least-squares (what is called statistical regression in most statistics courses) to predict the position of Ceres.

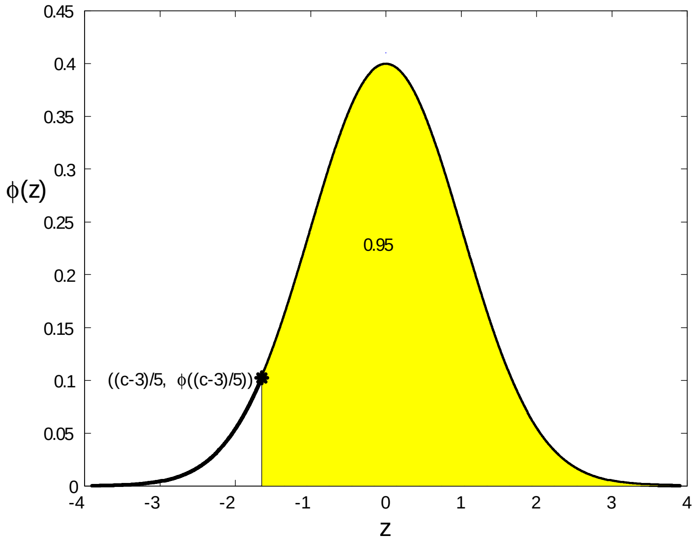


Figure 8.19: Picture for determining $P\left(Z > \frac{c-3}{5}\right) = 0.95$

Example: The distribution of heights of adult males in Canada is well approximate by a Gaussian distribution with mean $\mu = 69.0$ inches and standard deviation $\sigma = 2.4$ inches. Find the 10th and 90th percentiles of the height distribution.

Solution: We are being told that if X is the height of a randomly selected Canadian adult male, then $X \sim G(69.0, 2.4)$, or equivalently $X \sim N(69.0, 5.76)$.

To find the 90th percentile c, we use

$$\begin{aligned} P(X \le c) &= P\left(\frac{X-69.0}{2.4} \le \frac{c-69.0}{2.4}\right) \\ &= P\left(Z \le \frac{c-69.0}{2.4}\right) \quad \text{where } Z \sim G(0,1) \\ &= 0.90 \end{aligned}$$

From the table we see $P(Z \le 1.2816) = 0.90$ so we need

$$\frac{c-69.0}{2.4} = 1.2816$$

which gives $c = 72.08$ inches as the 90th percentile.

Similarly, to find c such that $P(X \le c) = 0.10$ we find that $P(Z \le -1.2816) = 0.10$, so we need

$$\frac{c-69.0}{2.4} = -1.2816$$

or $c = 65.92$ inches, as the 10th percentile.

Problem

8.5.1 Let X have a Normal distribution. What percent of the time does X lie within one standard deviation of the mean? Two standard deviations? Three standard deviations?

8.6 Chapter 8 Problems

1. A continuous random variable X has probability density function

$$f(x) = \begin{cases} k(1-x^2) & -1 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

   (a) Find k and the cumulative distribution function of X. Graph $f(x)$ and the cumulative distribution function.

   (b) Find the value of c such that $P(-c \leq X \leq c) = 0.95$.

   (c) Find $\mu = E\left(X\right)$ and $\sigma = sd\left(X\right)$.

   (d) Find the probability density function of $Y = X^2$.

2. When people are asked to make up a random number between 0 and 1, it has been found that the distribution of the numbers, X, has probability density function close to

$$f(x) = \begin{cases} 4x & 0 < x \leq 1/2 \\ 4\left(1-x\right) & \frac{1}{2} < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

   (rather than the $Uniform\left(0,1\right)$ distribution which would be expected).

   (a) Graph the probability density function and show that $\int\limits_{-\infty}^{\infty} f\left(x\right)dx = 1$ without evaluating an integral.

   (b) Find $P\left(0.25 \leq X \leq 0.8\right)$.

   (c) Find the median and 10th percentile of the distribution.

   (d) Find the mean and variance of X.

   (e) Find the probability density function of $Y = 2\left(X - 1/2\right)$.

   (f) Find the probability density function of $Z = X^3$.

3. Let X have probability density function

$$f(x) = \begin{cases} \frac{1}{20} & -10 \leq x \leq 10 \\ 0 & \text{otherwise} \end{cases}$$

   Show that $Y = \left(X+10\right)/20 \sim Uniform\left(0,1\right)$.

4. Suppose X is a continuous random variable with finite mean μ and standard deviation σ. Suppose also this its probability density function is symmetric about the line $x = \mu$. Show that $P\left(|X-\mu| \leq 2\sigma\right) = 2P\left(X \leq \mu + 2\sigma\right) - 1$.

5. Suppose that X is the lifetime in years of a Canadian born in 1995. When dealing with lifetimes, a function of interest is the survivor function defined as

$$S(x) = P(X > x) = 1 - P(X \leq x) = 1 - F(x)$$

Values of $S(x)$, based on data collected by the Canadian government, are given in the table below for $x = 30, 40, \ldots, 100$.

x	30	40	50	60	70	80	90	100
Females: $S(x)$	0.996	0.987	0.971	0.939	0.863	0.704	0.396	0.075
Males: $S(x)$	0.989	0.975	0.949	0.903	0.801	0.603	0.273	0.034

(a) If a female born in 1995 lives to over age 30, what is the probability she lives to over age 80? Over age 90? What are the corresponding probabilities for males?

(b) If 51% of persons born in 1995 were male, find the fraction of the total population (males and females) that will live to over age 90.

6. A continuous random variable X has probability density function

$$f(x) = \begin{cases} (\theta + 1)\, x^{\theta} & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

where θ is a real-valued parameter of the distribution.

(a) For what values of θ is this a probability density function? Explain.

(b) Find $P(X \leq 0.5)$.

(c) Find $E\left(X^k\right)$ for $k = 0, 1, \ldots$ and use this to find $E(X)$ and $Var(X)$.

(d) Find the probability density function of $Y = 1/X$.

7. A continuous random variable X has probability density function

$$f(x) = \begin{cases} kxe^{-x^2/\theta} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\theta > 0$ is a real-valued parameter of the distribution.

(a) Find k and the cumulative distribution function of X.

(b) Find the mean and variance of X. **Hint**: Use the method of substitution and the Gamma function.

(c) Show that $Y = X^2/\theta \sim Exponential(1)$.

8. The diameters in centimeters of spherical particles produced by a machine are randomly distributed according to a $U\left(0.6, 1.0\right)$ distribution. Find the probability density function for the volume of a particle.

9. The magnitudes of earthquakes in a region of North America can be modelled by an Exponential distribution with mean 2.5 measured on the Richter scale.

   (a) What is the probability an earthquake has a magnitude greater than 5 on the Richter scale?

   (b) Suppose 3 earthquakes occur in a given month. What is the probability that none of the earthquakes have a magnitude greater than 5 on the Richter scale?

   (c) If the magnitude of an earthquake exceeds 4, what is the probability it also exceeds 5?

10. The lifetime of a certain type of light bulb follows an Exponential distribution with mean 1000 hours.

    (a) What are the mean and standard deviation of the lifetime of this type of light bulb?

    (b) What are the mean and standard deviation of the lifetime of this type of light bulb in days?

    (c) Find the median lifetime in hours.

11. Traffic accidents at the intersection of University Avenue and Westmount Road occur according to a Poisson process with an average rate of 0.5 accidents per day. Suppose an accident has just occurred.

    (a) What is the expected waiting time until the next accident?

    (b) What is the probability that the waiting time until the next accident is less than 12 hours?

    (c) If there have been no accidents before noon on a particular day what is the probability that there are no accidents before midnight on the same day?

12. Jamie figures that the total number of thousands of kilometers that an auto can be driven before it needs to be junked is an Exponential random variable with mean 20 thousand kilometers. Smith has a used car that he claims has been driven 10 thousand kilometers.

    (a) If Jamie purchases the car, what is the probability that Jamie would get at least 20 thousand additional kilometers out of the car?

    (b) Repeat the calculation in (a) under the assumption that the lifetime kilometer-age of the car (in thousands of kilometers) is a $Uniform(0, 40)$ random variable.

13. Server crashes at a giant data center are assumed to follow a Poisson process. On average there are three server crashes per day (24 hours).

    (a) What is the probability that the waiting time between two consecutive crashes is greater than 8 hours?

    (b) Suppose there have been no server crashes for the last 8 hours. What is the probability that the time until the next crash exceeds one hour?

    (c) What is the probability that there are fewer than three crashes in a day?

14. **Gamma distribution**: A continuous random variable X is said to have the Gamma distribution with parameters $\alpha > 0$ and $\beta > 0$ if

$$f(x) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

    (a) Show that f is a legitimate probability density function.

    (b) Use the properties of the Gamma function to obtain $E(X)$ and $Var(X)$.

    (c) Verify that setting $\alpha = 1$ results in the Exponential distribution with parameter $\lambda = 1/\beta$.

15. 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 proportion of students who obtain each of the following letter grades:

$$A(\geq 80\%),\ B(70 - 80\%),\ C(60 - 70\%),\ D(50 - 60\%),\ F(< 50\%)$$

16. Suppose $X \sim N(10, 16)$. Find the 20th, 40th, 60th, and 80th percentiles of the distribution.

17. The number of liters X that a filling machine in a water bottling plant deposits in a nominal two liter bottle follows a Normal distribution $N(\mu, \sigma^2)$, where $\sigma = 0.01$ liters and μ is the setting (in liters) on the machine.

    (a) If $\mu = 2$, what is the probability a bottle has less than 2 liters of water in it?

    (b) Find 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 liters be less than 0.01?

18. A manufacturer produces bolts that are specified to be between 1.19 and 1.21 centimeters in diameter. If the production process results in a bolt's diameter being Normally distributed with mean 1.20 centimeters and standard deviation 0.005 centimeters, what percentage of bolts will not meet specifications?

19. Suppose that the diameters in millimeters of the eggs laid by a large flock of hens can be modelled by a Normal distribution with a mean of 40 millimeters and a variance of 4 (millimeters)^2. The wholesale selling price is 5 cents for an egg less than 37 millimeters in diameter, 6 cents for eggs between 37 and 42 millimeters, and 7 cents for eggs over 42 millimeters. What is the average wholesale price per egg?

20. The manufacturer of computer chips advertises that the lifetimes of the computer chips that it produces are Normally distributed with mean $\mu = 5 \times 10^6$ hours and standard deviation $\sigma = 5 \times 10^5$ hours.

    (a) What proportion of computer chips last less than 6×10^6 hours?

    (b) What proportion of computer chips last longer than 4×10^6 hours?

    (c) The manufacturer is working on improvements to the computer chips to increase the average lifetime. The manufacturer wishes to ensure that at least 95 percent of the computer chips last longer than 4.5×10^6 hours. What should the value of μ be to achieve this? (Assume the value of σ is unchanged by the improvements.)

21. The temperature of a CPU (central processing unit) while operating, X, has a Normal distribution with mean $\mu = 60$ degrees Celsius and standard deviation $\sigma = 5$ degrees Celsuis. If the temperature reaches over 75 degrees, it will throttle (slow down) to avoid damage to the CPU. If it reaches 95 degrees, it will shut down.

    (a) What is the probability the CPU will slow down?

    (b) Find c such that $P(X \leq c) = 0.9$.

    (c) You can overclock your CPU (make it run faster than it's supposed to) but that increases its average operating temperature μ. What should μ be set at to make the probability the CPU shuts down be 0.01? (Assume σ does not change and that there is no slow down at 75 degrees.)

22. **Binary classification**: Many situations require that we "classify" a unit of some type as being one of two types, which for convenience we will call positive and negative. For example, a diagnostic test for a disease might be positive or negative; an email message may be spam or not spam; a credit card transaction may be fraudulent or not. The problem is that in many cases we cannot tell for certain whether a unit is positive or negative, so when we have to classify the unit, we may make errors. The following framework helps us to deal with these problems.

    For a randomly selected unit from the population being considered, define the random variable Y such that $Y = 1$ if the unit is positive and $Y = 0$ if the unit is negative. Suppose that we are unable to determine if $Y = 0$ or $Y = 1$ for a given unit, but that we can obtain a measurement X with the property that

$$\begin{aligned}&\text{if } Y = 1 \text{ then } X \sim N(\mu_1, \sigma_1^2)\\&\text{if } Y = 0 \text{ then } X \sim N(\mu_0, \sigma_0^2)\end{aligned}$$

where $\mu_1 > \mu_0$. We then classify the unit based on the measurement X according to the following rule:

$$\begin{aligned}&\text{if } X \geq d \text{ then classify the unit as positive}\\&\text{if } X < d \text{ then classify the unit as negative}\end{aligned}$$

where d is a value chosen to between μ_0 and μ_1. Such a rule can obviously result in errors. If a unit is actually positive and we wrongly classified it as negative then we call this a "false negative". If a unit is actually negative and we wrongly classified it as positive then we call this a "false positive".

(a) Find the probability of a false negative and the probability of a false positive if $\mu_0 = 0$, $\mu_1 = 10$, $\sigma_0 = 4$, $\sigma_1 = 6$ and $d = 5$.

(b) Find the probability of a false negative and the probability of a false positive if $\mu_0 = 0$, $\mu_1 = 10$, $\sigma_0 = 3$, $\sigma_1 = 3$. Explain in words why the false negative and false positive missclassification probabilities are smaller than in (a).

23. **Binary classification and spam detection**: Chapter 4, Problems 21 and 22 discussed methods of spam detection using binary features. Suppose that for a given email message we compute a measure X, designed so that X tends to be high for spam messages and low for regular (non-spam) messages. (For example X can be a composite measure based on the presence or absence of certain words in a message, as well as other features.) In this problem we assume that X is a continuous random variable.

Suppose that for spam messages, the distribution of X is approximately $N(\mu_1, \sigma_1^2)$, and that for regular messages, it is approximately $N(\mu_0, \sigma_0^2)$, where $\mu_1 > \mu_0$. This is the setup described in Problem 22. We will filter spam by picking a value d, and then filtering any message for which $X \geq d$. The trick here is to decide what value of d to use.

(a) Suppose that $\mu_0 = 0$, $\mu_1 = 10$, $\sigma_1 = 3$, $\sigma_2 = 3$. What is the probability of a false positive (filtering a message that is not spam) and a false negative (not filtering a message that is spam) under each of the three choices (i) $d = 5$ (ii) $d = 4$ (iii) $d = 6$?

(b) What factors would determine which of the three choices of d would be best to use?

24. Let $T_1, T_2, \ldots, T_n$ denote the first n interarrival times of a Poisson process X_t (recall that X_t is the number of hits in $[0, t)$) with intensity λ.

    (a) What is the interpretation of $S_n = \sum\limits_{i=1}^{n} T_i$?

    (b) Argue that the two events $\{S_n \leq t\}$ and $\{X_t \geq n\}$ are identical.

    (c) Use (b) to show that
    $$P(S_n \leq t) = 1 - \sum_{j=0}^{n-1} e^{-\lambda t} \frac{(\lambda t)^j}{j!}$$

    (d) By differentiating the cumulative distribution function of S_n given in (c), show that S_n is a Gamma random variable (see Problem 8.12) with parameters $\alpha = n$ and $\beta = \lambda$.

25. **Cauchy distribution**: A random variable X is said to have a Cauchy distribution with parameter $\alpha > 0$ if
    $$f(x) = \frac{\alpha}{\pi(\alpha^2 + x^2)} \quad \text{for } x \in \Re$$

    (a) Show that $E(X)$ does not exist. Explain why this also implies $Var(X)$ does not exist

    (b) Let $Y = 1/X$. Show that Y has a Cauchy distribution with parameter α^{-1}.

    (c) Find the cumulative distribution function F and the inverse cumulative distribution function F^{-1} for the random variable X.

    (d) Suppose $U \sim Uniform\,(-1, 0)$. Find a function g such that $g(U)$ is a Cauchy random variable with parameter α.

26. **Challenge Problem**: Given a circle, find the probability that a chord chosen at random be longer than the side of an inscribed equilateral triangle. For example in Figure 8.20, the line joining A and B satisfies the condition, the other lines do not. This is called Bertrand's paradox (see

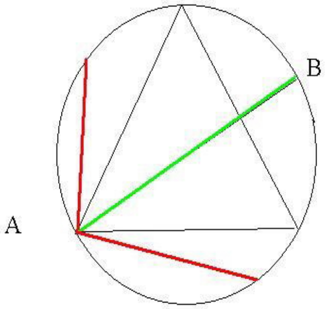


Figure 8.20: Bertrand's Paradox

http://www.cut-the-knot.org/bertrand.shtml) and there various possible solutions, depending on exactly how you interpret the phrase “a chord chosen at random”. For example, since the only important thing is the position of the second point relative to the first one, we can fix the point A and consider only the chords that emanate from this point. Then it becomes clear that 1/3 of the outcomes (those with angle with the tangent at that point between 60 and 120 degrees) will result in a chord longer than the side of an equilateral triangle. But a chord is fully determined by its midpoint. Chords whose length exceeds the side of an equilateral triangle have their midpoints inside a smaller circle with radius equal to 1/2 that of the given one. If we choose the midpoint of the chord at random and uniformly from the points within the circle, what is the probability that corresponding chord has length greater than the side of the triangle? Can you think of any other interpretations which lead to different answers?

9. MULTIVARIATE DISTRIBUTIONS

9.1 Basic Terminology and Techniques

Many problems involve more than a single random variable. When there are multiple random variables associated with an experiment or process we usually denote them as $X, Y, \ldots$ or as $X_1, X_2, \ldots$. For example, your final mark in a course might involve $X_1 =$ your assignment mark, $X_2 =$ your midterm test mark, and $X_3 =$ your exam mark. We need to extend the ideas introduced for single variables to deal with multivariate problems. In Sections $9.1 - 9.5$ we consider discrete multivariate problems. Continuous multivariate variables are also common in daily life (e.g. consider a person's height X and weight Y, or $X_1 =$ the return from Stock 1, $X_2 =$ return from stock 2). In Section 9.6 we will only consider the special case of a linear combination of independent Normal random variables.

To introduce the ideas in a simple setting, we will first consider an example in which there are only a few possible values of the variables. Later we will apply these concepts to more complex examples. In particular we will look at the Multinomial distribution which is a generalization of the Binomial distribution that we saw in Chapter 5. The ideas themselves are simple even though some applications can involve fairly messy algebra.

Joint Probability Functions:

First, suppose there are two discrete random variables X and Y, and define the function

$$\begin{aligned} f(x,y) &= P(X = x \text{ and } Y = y) \\ &= P(X = x, Y = y) \end{aligned}$$

We call $f(x,y)$ the joint probability function of (X,Y). The properties of a joint probability function are similar to those for a single variable; for two random variables we have $f(x,y) \geq 0$ for all (x,y) and

$$\sum_{\text{all(x,y)}} f(x,y) = 1$$

In general,

$$f(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

if there are n random variables $X_1, \ldots, X_n$.

Example: Consider the following numerical example, where we show $f(x, y)$ in a table.

			x		
$f(x,y)$		0	1	2	
y	1	0.1	0.2	0.3	
	2	0.2	0.1	0.1	
					1

For example $f(0, 2) = P(X = 0,Y = 2) = 0.2$. We can check that $f(x, y)$ is a proper joint probability function since $f(x, y) \geq 0$ for all 6 combinations of (x, y) and the sum of these 6 probabilities is 1. When there are only a few values for X and Y it is often easier to tabulate $f(x, y)$ than to find a formula for it. We'll use this example below to illustrate other definitions for multivariate distributions, but first we give a short example where we need to find $f(x, y)$.

Example: Suppose a fair coin is tossed 3 times. Define the random variables X = number of Heads and $Y = 1$ (0) if Heads (Tails) occurs on the first toss. Find the joint probability function for (X, Y).

Solution: First we should note the range for (X, Y), which is the set of possible values (x, y) which can occur. Clearly X can be 0, 1, 2, or 3 and Y can be 0 or 1, but we'll see that not all 8 combinations (x, y) are possible. We can find $f(x, y) = P(X = x, Y = y)$ by just writing down the sample space

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

that we have used before for this process. Then simple counting gives $f(x, y)$ as shown in the following table:

				x		
$f(x,y)$		0	1	2	3	
y	0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	0	
	1	0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	
						1

For example, $(X, Y) = (0, 0)$ if and only if the outcome is TTT; $(X, Y) = (1, 0)$ if and only if the outcome is either THT or TTH.

Note that the joint probability function for (X, Y) is a little awkward to write down in a formula, so we just use a table.

Marginal Distributions:

We may be given a joint probability function involving more variables than we're interested in using. How can we eliminate any which are not of interest? Look at the first example above. If we're only interested in X, and don't care what value Y takes, we can see that

$$\begin{aligned} P(X=0) &= P(X=0, Y=1) + P(X=0, Y=2) \\ &= f(0,1) + f(0,2) \\ &= 0.3 \end{aligned}$$

Similarly

$$\begin{aligned} & P(X=1) = f(1,1) + f(1,2) = 0.3 \\ \text{and} \quad & P(X=2) = f(2,1) + f(2,2) = 0.4 \end{aligned}$$

The distribution of X obtained in this way from the joint probability function is called the marginal probability function of X:

x	0	1	2	Total
$f_1(x) = P(X = x)$	0.3	0.3	0.4	1

In the same way, if we were only interested in Y, we obtain

$$P(Y=1) = f(0,1) + f(1,1) + f(2,1) = 0.6$$

since X can be 0, 1, or 2 when $Y = 1$. The marginal probability function of Y would be:

y	1	2	Total
$f_2(y) = P(Y = y)$	0.6	0.4	1

Note that we use the notation $f_1(x) = P(X = x)$ and $f_2(y) = P(Y = y)$ to avoid confusion with $f(x, y) = P(X = x, Y = y)$. An alternative notation that you may see is $f_X(x)$ and $f_Y(y)$.

In general, to find $f_1(x)$ we add over all values of y where $X = x$, and to find $f_2(y)$ we add over all values of x with $Y = y$. Then

$$\begin{aligned} & f_1(x) = \sum_{\text{all } y} f(x,y) \\ \text{and} \quad & f_2(y) = \sum_{\text{all } x} f(x,y) \end{aligned}$$

This reasoning can be extended beyond two variables. For example, with three variables (X_1, X_2, X_3),

$$f_1(x_1) = \sum_{\text{all } (x_2, x_3)} f(x_1, x_2, x_3)$$

$$\text{and} \quad f_{1,3}(x_1, x_3) = \sum_{\text{all } x_2} f(x_1, x_2, x_3) = P(X_1 = x_1, X_3 = x_3)$$

where $f_{1,3}(x_1, x_3)$ is the marginal joint probability function of (X_1, X_3).

Note that if the joint probability function is given in a table then the marginal probability functions are obtained by simply summing over the rows and columns as shown in the table below for the coin example above:

		x				
$f(x,y)$		0	1	2	3	$f_2(y)$
y	0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	0	$\frac{4}{8}$
	1	0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{4}{8}$
$f_1(x)$		$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	1

Independent Random Variables:

For events A and B, we have defined A and B to be independent if and only if $P(AB) = P(A)P(B)$. This definition can be extended to random variables (X, Y). Two random variables are independent if their joint probability function is the product of the marginal probability functions.

Definition 26 *X and Y are **independent** random variables if $f(x, y) = f_1(x)f_2(y)$ for all values (x, y).*

Definition 27 *In general, $X_1, X_2, \ldots, X_n$ are independent random variables if and only if*

$$f(x_1, x_2, \ldots, x_n) = f_1(x_1)f_2(x_2) \cdots f_n(x_n) \quad \textit{for all } x_1, x_2, \ldots x_n$$

In the first example X and Y are not independent since $f_1(x)f_2(y) \neq f(x, y)$ for any of the 6 combinations of (x, y) values; e.g., $f(1, 1) = 0.2 \neq f_1(1)f_2(1) = (0.3)\,(0.6)$. Be careful applying this definition. You can only conclude that X and Y are independent after checking all (x, y) combinations. Even a single case where $f_1(x)f_2(y) \neq f(x, y)$ makes X and Y dependent random variables.

Conditional Probability Functions:

Again we can extend a definition from events to random variables. For events A and B, recall that

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

Since

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} \text{ provided } P(Y = y) > 0$$

we make the following definition.

Definition 28 *The conditional probability function of X given $Y = y$ is*

$$f_1(x|y) = \frac{f(x,y)}{f_2(y)} \textit{ provided } f_2(y) > 0$$

Similarly, the conditional probability function of Y given $X = x$ is

$$f_2(y|x) = \frac{f(x,y)}{f_1(x)} \textit{ provided } f_1(x) > 0$$

Example: Suppose X and Y have joint probability function:

			x		
$f(x,y)$		0	1	2	$f_2(y)$
y	1	0.1	0.2	0.3	0.6
	2	0.2	0.1	0.1	0.4
$f_1(x)$		0.3	0.3	0.4	1

Find $f_1(x|Y = 1) = P(X = x|Y = 1)$.

Solution: Since

$$f_1(x|Y = 1) = \frac{f(x,1)}{f_2(1)}$$

we obtain

x	0	1	2	Total
$f_1(x\|Y = 1)$	$\frac{0.1}{0.6} = \frac{1}{6}$	$\frac{0.2}{0.6} = \frac{1}{3}$	$\frac{0.3}{0.6} = \frac{1}{2}$	1

As you would expect, marginal and conditional probability functions are probability functions in that they are always ≥ 0 and their sum is 1.

Functions of Random Variables:

Your final mark in a course might be a function of the three variables X_1, X_2, X_3 - assignment, midterm, and exam marks[34]. We often encounter problems where we need to find the probability distribution of a function of two or more random variables. The most general method for finding the probability function for some function of random variables X and Y involves looking at every combination (x, y) to see what value the function takes.

Example: Suppose X and Y have joint probability function

		x			
	$f(x,y)$	0	1	2	
y	1	0.1	0.2	0.3	
	2	0.2	0.1	0.1	
					1

and we want to find the probability function of $U = 2(Y - X)$. The possible values of U are seen by looking at the value of $u = 2(y - x)$ for each (x, y) in the range of (X, Y).

		x		
	u	0	1	2
y	1	2	0	-2
	2	4	2	0

Since

$$
\begin{aligned}
&P(U=-2) = P(X=2, Y=1) = f(2,1) = 0.3\\
&P(U=0) = P(X=1, Y=1) + P(X=2, Y=2) = f(1,1) + f(2,2) = 0.3\\
&P(U=2) = f(0,1) + f(1,2) = 0.2\\
&P(U=4) = f(0,2) = 0.2
\end{aligned}
$$

the probability function of U is

u	-2	0	2	4	Total
$P(U=u)$	0.3	0.3	0.2	0.2	1

[34] "Don't worry about your marks. Just make sure that you keep up with the work and that you don't have to repeat a year. It's not necessary to have good marks in everything." Albert Einstein in letter to his son, 1916.

For some functions it is possible to approach the problem more systematically. One of the most common functions of this type is the total. Let $T = X + Y$. This gives:

			x	
t		0	1	2
y	1	1	2	3
	2	2	3	4

Then $P(T = 3) = f(1, 2) + f(2, 1) = 0.4$, for example. Continuing in this way, we obtain

t	1	2	3	4	Total
$P(T = t)$	0.1	0.4	0.4	0.1	1

In fact, to find $P(T = t)$ we are simply adding the probabilities for all (x, y) combinations with $x + y = t$. This could be written as:

$$P(T = t) = \sum_{\substack{\text{all } (x,y) \\ \text{with } x+y=t}} f(x, y)$$

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 over values of x and then substitute $t - x$ for y. Then,

$$P(T = t) = \sum_{\text{all } x} f(x, t - x) = \sum_{\text{all } x} P(X = x, Y = t - x)$$

So $P(T = 3)$ would be

$$P(T = 3) = \sum_{\text{all } x} f(x, 3 - x) = f(0, 3) + f(1, 2) + f(2, 1) = 0.4$$

(note $f(0, 3) = 0$ since Y can't be 3.)

We can summarize the method of finding the probability function for a function $U = g(X, Y)$ of two random variables X and Y as follows:
Let $f(x, y) = P(X = x, Y = y)$ be the probability function for (X, Y). Then the probability function for U is

$$P(U = u) = \sum_{\substack{\text{all}(x,y): \\ g(x,y)=u}} f(x, y)$$

This can also be extended to functions of three or more random variables $U = g(X_1, X_2, \ldots, X_n)$:

$$P(U = u) = \sum_{\substack{(x_1,\ldots,x_n): \\ g(x_1,\ldots,x_n)=u}} f(x_1, \ldots, x_n)$$

Note: Do not get confused between the functions f and g in the above: $f(x, y)$ is the joint probability function of the random variables X, Y whereas $U = g(X, Y)$ defines the "new" random variable that is a function of X and Y, and whose distribution we want to find.

Theorem 29 *If $X \sim Poisson\,(\mu_1)$ and $Y \sim Poisson\,(\mu_2)$ independently then $T = X + Y \sim Poisson\,(\mu_1 + \mu_2)$.*

Proof. Since $X \sim Poisson\,(\mu_1)$ and independently $Y \sim Poisson\,(\mu_2)$ their joint probability function is given by

$$f(x, y) = \frac{\mu_1^x e^{-\mu_1}}{x!} \cdot \frac{\mu_2^y e^{-\mu_2}}{y!} \text{ for } x = 0, 1, \ldots \text{ and } y = 0, 1, \ldots.$$

The probability function of T is

$$\begin{aligned}
P\,(T = t) &= P(T = t) = P(X + Y = t) \\
&= \sum_{\text{all } x} P(X = x, Y = t - x) \\
&= \sum_{x=0}^{t} \frac{\mu_1^x e^{-\mu_1}}{x!} \frac{\mu_2^{t-x} e^{-\mu_2}}{(t - x)!} \\
&= \mu_2^t e^{-(\mu_1 + \mu_2)} \sum_{x=0}^{t} \frac{1}{x!(t - x)!} \left(\frac{\mu_1}{\mu_2}\right)^x \\
&= \frac{\mu_2^t e^{-(\mu_1 + \mu_2)}}{t!} \sum_{x=0}^{t} \binom{t}{x} \left(\frac{\mu_1}{\mu_2}\right)^x \\
&= \frac{\mu_2^t e^{-(\mu_1 + \mu_2)}}{t!} \left(1 + \frac{\mu_1}{\mu_2}\right)^t \quad \text{by the Binomial Theorem.} \\
&= \frac{\mu_2^t e^{-(\mu_1 + \mu_2)}}{t!} \frac{(\mu_1 + \mu_2)^t}{\mu_2^t} \\
&= \frac{(\mu_1 + \mu_2)^t}{t!} e^{-(\mu_1 + \mu_2)} \text{ for } t = 0, 1, 2, \ldots
\end{aligned}$$

which we recognize as the probability function of a $Poisson\,(\mu_1 + \mu_2)$ and we have proved the desired result. ■

Exercise: Prove the following theorem.

Theorem 30 *If $X \sim Binomial\,(n, p)$ and $Y \sim Binomial\,(m, p)$ independently then $T = X + Y \sim Binomial\,(n + m, p)$.*

Problems

9.1.1 The joint probability function of (X, Y) is:

			x	
	$f(x,y)$	0	1	2
	0	0.09	0.06	0.15
y	1	0.15	0.05	0.20
	2	0.06	0.09	0.15

(a) Are X and Y independent random variables? Why?

(b) Tabulate the conditional probability function of Y given $X = 0$.

(c) Tabulate the probability function of $D = X - Y$.

9.1.2 Suppose X and Y are independent random variables with $f_1(x) = \binom{x+k-1}{x} p^k (1-p)^x$ and $f_2(y) = \binom{y+\ell-1}{y} p^\ell (1-p)^y$. Let $T = X + Y$. Find the probability function of T. **Hint**: use the result $\binom{a+b-1}{a} = (-1)^a \binom{-b}{a}$.

9.2 Multinomial Distribution

There is only one multivariate model distribution introduced in this course, though other multivariate distributions do exist. The Multinomial distribution defined below is very important. It is a generalization of the Binomial model to the case where each trial has k possible outcomes. Before defining the Multinomial distribution we consider the following example:

Example: Three sprinters, A, B and C, compete against each other in 10 independent 100 m. races. The probabilities of winning any single race are 0.5 for A, 0.4 for B, and 0.1 for C. Let X_1, X_2 and X_3 be the number of races A, B and C win respectively.

(a) Find the joint probability function, $f(x_1, x_2, x_3)$

(b) Find the marginal probability function, $f_1(x_1)$

(c) Find the conditional probability function, $f(x_2|x_1)$

(d) Are X_1 and X_2 independent? Why?

(e) Find the probability function of $T = X_1 + X_2$.

Solution: Before starting, note that $x_1 + x_2 + x_3 = 10$ since there are 10 races in all. We really only have two variables since $x_3 = 10 - x_1 - x_2$. However it is convenient to use x_3 to save writing and preserve symmetry.

(a) The reasoning will be similar to the way we found the Binomial distribution in Chapter 5 except that there are now 3 types of outcome. There are $\frac{10!}{x_1!x_2!x_3!}$ different outcomes (that is, results for races 1 to 10) in which there are x_1 wins by A, x_2 by B, and x_3 by C. Each of these arrangements has a probability of (0.5) multiplied x_1 times, (0.4) multiplied x_2 times, and (0.1) multiplied x_3 times in some order, that is, $(0.5)^{x_1}(0.4)^{x_2}(0.1)^{x_3}$.

Therefore

$$f(x_1, x_2, x_3) = \frac{10!}{x_1!x_2!x_3!}(0.5)^{x_1}(0.4)^{x_2}(0.1)^{x_3}$$

The domain of f is the set $\{(x_1, x_2, x_3);\ x_i = 0, 1, \ldots, 10,\ i = 1, 2, 3 \text{ and } x_1 + x_2 + x_3 = 10\}$.

(b) It would also be acceptable to drop x_3 as a variable and write down the probability function for X_1, X_2 only; this is

$$f(x_1, x_2) = \frac{10!}{x_1!x_2!(10 - x_1 - x_2)!}(0.5)^{x_1}(0.4)^{x_2}(0.1)^{10-x_1-x_2}$$

because of the fact that X_3 must equal $10 - X_1 - X_2$. For this probability function $x_1 = 0, 1, \cdots, 10$; $x_2 = 0, 1, \cdots, 10$ and $x_1 + x_2 \leq 10$. This simplifies finding $f_1(x_1)$ a little. We now have $f_1(x_1) = \sum\limits_{x_2} f(x_1, x_2)$. The limits of summation need care: x_2 could be as small as 0, but since $x_1 + x_2 \leq 10$, we also require $x_2 \leq 10 - x_1$. (For example if $x_1 = 7$ then B can win $0, 1, 2$, or 3 races.) Thus,

$$\begin{aligned} f_1(x_1) &= \sum_{x_2=0}^{10-x_1} \frac{10!}{x_1!x_2!(10 - x_1 - x_2)!}(0.5)^{x_1}(0.4)^{x_2}(0.1)^{10-x_1-x_2} \\ &= \frac{10!}{x_1!}(0.5)^{x_1}(0.1)^{10-x_1} \sum_{x_2=0}^{10-x_1} \frac{1}{x_2!(10 - x_1 - x_2)!}\left(\frac{0.4}{0.1}\right)^{x_2} \end{aligned}$$

(Hint: In $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ the two terms in the denominator add to the term in the numerator, if we ignore the ! sign.) Multiply top and bottom by $[x_2 + (10 - x_1 - x_2)]! = (10 - x_1)!$ This gives

$$
\begin{aligned}
f_1(x_1) &= \frac{10!}{x_1!(10-x_1)!}(0.5)^{x_1}(0.1)^{10-x_1}\sum_{x_2=0}^{10-x_1}\binom{10-x_1}{x_2}\left(\frac{0.4}{0.1}\right)^{x_2} \\
&= \binom{10}{x_1}(0.5)^{x_1}(0.1)^{10-x_1}\left(1+\frac{0.4}{0.1}\right)^{10-x_1} \quad \text{by the Binomial Theorem} \\
&= \binom{10}{x_1}(0.5)^{x_1}(0.1)^{10-x_1}\frac{(0.1+0.4)^{10-x_1}}{(0.1)^{10-x_1}} \\
&= \binom{10}{x_1}(0.5)^{x_1}(0.5)^{10-x_1} \quad \text{for } x_1 = 0,1,2,\ldots,10
\end{aligned}
$$

Note: While this derivation is included as an example of how to find marginal distributions by summing a joint probability function, there is a much simpler method for this problem. Note that each race is either won by A (“success”) or it is not won by A (“failure”). Since the races are independent and X_1 is now just the number of “success” outcomes, X_1 must have a Binomial distribution, with $n = 10$ and $p = 0.5$. Hence

$$f_1(x_1) = \binom{10}{x_1}(0.5)^{x_1}(0.5)^{10-x_1} \quad \text{for } x_1 = 0,1,\ldots,10 \text{ as above.}$$

(c) Remember that $f(x_2|x_1) = P(X_2 = x_2|X_1 = x_1)$, so that

$$
\begin{aligned}
f(x_2|x_1) = \frac{f(x_1,x_2)}{f_1(x_1)} &= \frac{\frac{10!}{x_1!x_2!(10-x_1-x_2)!}(0.5)^{x_1}(0.4)^{x_2}(0.1)^{10-x_1-x_2}}{\frac{10!}{x_1!(10-x_1)!}(0.5)^{x_1}(0.5)^{10-x_1}} \\
&= \frac{(10-x_1)!}{x_2!\,(10-x_1-x_2)!}\frac{(0.4)^{x_2}\,(0.1)^{10-x_1-x_2}}{(0.5)^{x_2}\,(0.5)^{10-x_1-x_2}} \\
&= \binom{10-x_1}{x_2}\left(\frac{4}{5}\right)^{x_2}\left(\frac{1}{5}\right)^{10-x_1-x_2} \quad \text{for } x_2 = 0,1,\ldots,(10-x_1)
\end{aligned}
$$

The range of X_2 depends on the value x_1, which makes sense: if B wins x_1 races then the most A can win is $10 - x_1$.

Note: As in (b), this result can be obtained more simply by general reasoning. Once we are given that A wins x_1 races, the remaining $(10 - x_1)$ races (“trials”) are all won by either B or C. Since $P(B \text{ wins}) = 0.4$ and $P(C \text{ wins}) = 0.1$ then, for the races won by either B or C, the probability that B wins (“Success”) is

$$P(B \text{ wins } |B \text{ or } C \text{ wins}) = \frac{P\,(B \text{ wins})}{P\,(B \text{ or } C \text{ wins})} = \frac{0.4}{0.4+0.1} = 0.8$$

So the probability function of the number of wins (“Successes”) in $(10 - x_1)$ races (“trials”) is

$$f(x_2|x_1) = \binom{10-x_1}{x_2}(0.8)^{x_2}\,(0.2)^{10-x_1-x_2} \quad \text{for } x_2 = 0,1,\ldots,10-x_1$$

(d) X_1 and X_2 are clearly not independent random variables since the more races A wins, the fewer races there are for B to win. More formally,

$$f_1(x_1)f_2(x_2) = \binom{10}{x_1}(0.5)^{x_1}(0.5)^{10-x_1}\binom{10}{x_2}(0.4)^{x_2}(0.6)^{10-x_2} \neq f(x_1,x_2)$$

(In general, if the range for X_1 depends on the value of X_2, then X_1 and X_2 cannot be independent random variables.)

(e) If $T = X_1 + X_2$ then

$$\begin{aligned} f_T(t) = P(T=t) &= \sum_{x_1}^{t} f(x_1, t-x_1) \\ &= \sum_{x_1=0}^{t} \frac{10!}{x_1!(t-x_1)!\underbrace{(10-x_1-(t-x_1))!}_{(10-t)!}}(0.5)^{x_1}(0.4)^{t-x_1}(0.1)^{10-t} \end{aligned}$$

The upper limit on x_1 is t because, for example, if $t = 7$ then A could not have won more than 7 races. Then

$$f_T(t) = P(T=t) = \frac{10!}{(10-t)!}(0.4)^t(0.1)^{10-t}\sum_{x_1=0}^{t}\frac{1}{x_1!(t-x_1)!}\left(\frac{0.5}{0.4}\right)^{x_1}$$

What do we need to multiply by on the top and bottom? Can you spot it before looking below?

$$\begin{aligned} f_T(t) = P(T=t) &= \frac{10!}{t!(10-t)!}(0.4)^t(0.1)^{10-t}\sum_{x_1=0}^{t}\frac{t!}{x_1!(t-x_1)!}\left(\frac{0.5}{0.4}\right)^{x_1} \\ &= \binom{10}{t}(0.4)^t(0.1)^{10-t}\left(1+\frac{0.5}{0.4}\right)^t \\ &= \binom{10}{t}(0.4)^t(0.1)^{10-t}\frac{(0.4+0.5)^t}{(0.4)^t} \\ &= \binom{10}{t}(0.9)^t(0.1)^{10-t} \quad \text{for } t = 0,1,\ldots,10 \end{aligned}$$

Exercise: Explain to yourself how the answer in (e) can be obtained from the Binomial distribution, as we did for parts (b) and (c)

We now generalize this example to the case in which there are k types of outcome rather than three.

Physical Setup for the Multinomial distribution: Suppose an experiment is repeated independently n times with k distinct types of outcome each time. Let the probabilities of these k types be $p_1, p_2, \ldots, p_k$ each time. Let X_1 be the number of times the 1st type occurs, X_2 the number of times the 2nd occurs, ..., X_k the number of times the k'th type occurs. Then $(X_1, X_2, \ldots, X_k)$ has a Multinomial distribution.

Notes:

(1) $p_1 + p_2 + \cdots + p_k = 1$

(2) $X_1 + X_2 + \cdots + X_k = n$,

If we wish we can drop one of the variables (say the last), and just note that
$X_k = n - X_1 - X_2 - \cdots - X_{k-1}$.

Illustrations:

(1) If $k = 2$ and there are two possible outcomes (Success and Failure) then we simply have a Binomial distribution.

(2) In the example above with sprinters A, B, and C running 10 races we had a Multinomial distribution with $n = 10$ and $k = 3$. Since there were $k = 3$ possible outcomes this distribution is also called the Trinomial distribution.

(3) Suppose student marks are given in letter grades as A, B, C, D, or F. In a class of 80 students the number getting A, B, ..., F might have a Multinomial distribution with $n = 80$ and $k = 5$.

Joint Probability Function: The joint probability function of $X_1, X_2, \ldots, X_k$ is given by extending the argument in the sprinters example from $k = 3$ to general k. There are $\frac{n!}{x_1!x_2!\cdots x_k!}$ different outcomes of the n trials in which x_1 are of the 1st type, x_2 are of the 2nd type, etc. 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, etc. Therefore

$$f(x_1, x_2, \ldots, x_k) = \frac{n!}{x_1!x_2!\cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

The restriction on the x_i's are $x_i = 0, 1, \ldots, n$ and $\sum\limits_{i=1}^{k} x_i = n$.
As a check that $\sum f(x_1, x_2, \ldots, x_k) = 1$ we use the Multinomial Theorem to get

$$\sum \frac{n!}{x_1!x_2!\cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} = (p_1 + p_2 + \cdots + p_k)^n = 1$$

Example: Every person is one of four blood types: A, B, AB and O. (This is important in determining, for example, who may give a blood transfusion to a person.) In a large population let the fraction that has type A, B, AB and O, respectively, be p_1, p_2, p_3, p_4. Then, if n persons are randomly selected from the population, the numbers X_1, X_2, X_3, X_4 of types A, B, AB, O have a Multinomial distribution with $k = 4$. (In Caucasian people the values of the p_i's are approximately $p_1 = 0.45$, $p_2 = 0.08$, $p_3 = 0.03$, $p_4 = 0.44$.)

Remark: We sometimes use the notation $(X_1, \ldots, X_k) \sim Multinomial(n; p_1, p_2, \ldots, p_k)$ to indicate that $(X_1, X_2, \ldots, X_k)$ have a Multinomial distribution.

Remark: For some types of problems its helpful to write formulas in terms of $x_1, x_2, \ldots, x_{k-1}$ and $p_1, p_2, \ldots, p_{k-1}$ using the fact that

$$x_k = n - x_1 - x_2 - \cdots - x_{k-1} \text{ and } p_k = 1 - p_1 - p_2 \cdots - -p_{k-1}$$

In this case we can write the joint probability function as $f(x_1, x_2, \ldots, x_{k-1})$ but we must remember then that $x_1, x_2, \ldots, x_{k-1}$ satisfy the condition $0 \leq x_1 + x_2 + \cdots + x_{k-1} \leq n$.

The Multinomial distribution can also arise in combination with other models, and students often have trouble recognizing it then.

Example: A potter is producing teapots one at a time. Assume that they are produced independently of each other and with probability p the pot produced will be "satisfactory"; the rest are sold at a lower price. The number, X, of rejects before producing a satisfactory teapot is recorded. When 12 satisfactory teapots are produced, what is the probability the 12 values of X will consist of six 0's, three 1's, two 2's and one value which is ≥ 3?

Solution: Each time a "satisfactory" pot is produced the value of X falls in one of the four categories $X = 0$, $X = 1$, $X = 2$, $X \geq 3$. Under the assumptions given in the question, X has a Geometric distribution with

$$P\left(X = x\right) = f\left(x\right) = p(1-p)^x \quad \text{for } x = 0, 1, 2, \ldots$$

so we can find the probability for each of these categories.

We have

$$\begin{aligned}
P(X = 0) &= f(0) = p \\
P(X = 1) &= f(1) = p\left(1 - p\right) \\
P(X = 2) &= f(2) = p\left(1 - p\right)^2
\end{aligned}$$

and

$$\begin{aligned}P\left(X\geq 3\right) &= f(3)+f(4)+f(5)+\cdots\\ &= p(1-p)^3+p(1-p)^4+p(1-p)^5+\cdots\\ &= \frac{p(1-p)^3}{1-(1-p)}\\ &= (1-p)^3 \quad \text{by the Geometric series.}\end{aligned}$$

Therefore

$$\begin{aligned}&P\left(\text{six 0's, three 1's, two 2's and one value } \geq 3\right)\\ &= \frac{12!}{6!3!2!1!}\left[p\right]^6\left[p\left(1-p\right)\right]^3\left[p\left(1-p\right)^2\right]^2\left[(1-p)^3\right]^1\\ &= \frac{12!}{6!3!2!1!}p^{6+3+2}(1-p)^{3+4+3}\\ &= \frac{12!}{6!3!2!1!}p^{11}(1-p)^{10}\end{aligned}$$

Problems

9.2.1 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. Give expressions for the probability that 25 randomly chosen policy holders will include

(a) 3A's, 11B's, 7C's, and 4D's.

(b) 3A's and 11B's.

(c) 3A's and 11B's, given that there are 4D's.

9.2.2 Chocolate chip cookies are made from batter containing an average of 0.6 chips per c.c. Chips are distributed according to the conditions for a Poisson process. Each cookie uses 12 c.c. of batter. Give expressions for the probabilities that in a dozen cookies:

(a) 3 have fewer than 5 chips.

(b) 3 have fewer than 5 chips and 7 have more than 9.

(c) 3 have fewer than 5 chips, given that 7 have more than 9.

9.3 Markov Chains

[35]Consider a sequence of (discrete) random variables $X_1, X_2, \ldots$ each of which takes integer values $1, 2, \ldots, N$ (called *states*). We assume that for a certain matrix P (called the *transition probability matrix*), the conditional probabilities are given by corresponding elements of the matrix; that is,

$$P\left(X_{n+1} = j | X_n = i\right) = P_{ij}, i = 1, \ldots N, j = 1, \ldots N$$

and furthermore that the chain only uses the last state occupied in determining its future; that is,

$$P\left(X_{n+1} = j | X_n = i, X_{n-1} = i_1, X_{n-2} = i_2 \ldots X_{n-l} = i_l\right) = P\left(X_{n+1} = j | X_n = i\right) = P_{ij}$$

for all $j, i, i_1, i_2, \ldots, i_l$, and $l = 2, 3, \ldots$. Then the sequence of random variables X_n is called a *Markov*[36] *Chain*. Markov Chain models are the most common simple models for dependent variables, and are used to predict weather as well as movements of security prices. They allow the future of the process to depend on the present state of the process, but the past behaviour can influence the future only through the present state.

Example. Rain-No rain

Suppose that the probability that tomorrow is rainy given that today is not raining is α (and it does not otherwise depend on whether it rained in the past) and the probability that tomorrow is dry given that today is rainy is β. If tomorrow's weather depends on the past only through whether today is wet or dry, we can define random variables

$$X_n = \begin{cases} 1 & \text{if} \quad \text{Day } n \text{ is wet} \\ 0 & \text{if} \quad \text{Day } n \text{ is dry} \end{cases}$$

(beginning at some arbitrary time origin, day $n = 0$). Then the random variables $X_n, n = 0, 1, 2, \ldots$ form a Markov chain with $N = 2$ possible states and having probability transition matrix

$$P = \left[\begin{array}{cc} 1-\alpha & \alpha \\ \beta & 1-\beta \end{array} \right]$$

[35] This section optional for STAT 220 and STAT 230.

[36] After Andrei Andreyevich Markov (1856-1922), a Russian mathematician, Professor at Saint Petersburg University. Markov studied sequences of mutually dependent variables, hoping to establish the limiting laws of probability in their most general form and discovered Markov chains, launched the theory of stochastic processes. As well, Markov applied the method of continued fractions, pioneered by his teacher Pafnuty Chebyshev, to probability theory, completed Chebyschev's proof of the central limit theorem (see Chapter 10) for independent non-identically distributed random variables. For entertainment, Markov was also interested in poetry and studied poetic style.

Properties of the Transition Matrix P

Note that $P_{ij} \geq 0$ for all i, j and $\sum_{j=1}^{N} P_{ij} = 1$ for all i. This last property holds because given that $X_n = i$, X_{n+1} must occupy one of the states $j = 1, 2, \ldots, N$.

The distribution of X_n

Suppose that the chain is started by randomly choosing a state for X_0 with distribution $P[X_0 = i] = q_i$, $i = 1, 2, \ldots, N$. Then the distribution of X_1 is given by

$$\begin{aligned} P(X_1 = j) &= \sum_{i=1}^{N} P(X_1 = j, X_0 = i) \\ &= \sum_{i=1}^{N} P(X_1 = j | X_0 = i) P(X_0 = i) \\ &= \sum_{i=1}^{N} P_{ij} q_i \end{aligned}$$

and this is the j'th element of the vector $\underline{q}^T P$ where $\underline{q}$ is the column vector of values q_i. To obtain the distribution at time $n = 1$, premultiply the transition matrix P by a vector representing the distribution at time $n = 0$. Similarly the distribution of X_2 is the vector $\underline{q}^T P^2$ where P^2 is the product of the matrix P with itself and the distribution of X_n is $\underline{q}^T P^n$. Under very general conditions, it can be shown that these probabilities converge because the matrix P^n converges pointwise to a limiting matrix as $n \to \infty$. In fact, in many such cases, the limit does not depend on the initial distribution $\underline{q}$ because the limiting matrix has all of its rows identical and equal to some vector of probabilities $\underline{\pi}$. Identifying this vector $\underline{\pi}$ when convergence holds is reasonably easy.

Definition 31 *A* limiting distribution *of a Markov chain is a vector ($\underline{\pi}$ say) of long run probabilities of the individual states such that*

$$\pi_i = \lim_{t \to \infty} P[X_t = i]$$

Now let us suppose that convergence to this distribution holds for a particular initial distribution $\underline{q}$ so we assume that

$$\underline{q}^T P^n \to \underline{\pi}^T \text{ as } n \to \infty$$

Then notice that

$$(\underline{q}^T P^n) P \to \underline{\pi}^T P$$

but also

$$(\underline{q}^T P^n) P = \underline{q}^T P^{n+1} \to \underline{\pi}^T \text{ as } n \to \infty$$

so $\underline{\pi}^T$ must have the property that

$$\underline{\pi}^T P = \underline{\pi}^T$$

Any limiting distribution must have this property and this makes it easy in many examples to identify the limiting behaviour of the chain.

Definition 32 *A* stationary distribution *of a Markov chain is the column vector ($\underline{\pi}$ say) of probabilities of the individual states such that $\underline{\pi}^T P = \underline{\pi}^T$.*

Example: (weather continued)

Let us return to the weather example in which the transition probabilities are given by the matrix

$$P = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}$$

What is the long-run proportion of rainy days? To determine this we need to solve the equations

$$\underline{\pi}^T P = \underline{\pi}^T$$

$$\begin{bmatrix} \pi_0 & \pi_1 \end{bmatrix} \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix} = \begin{bmatrix} \pi_0 & \pi_1 \end{bmatrix}$$

subject to the conditions that the values π_0, π_1 are both probabilities (non-negative) and add to one. It is easy to see that the solution is

$$\pi_0 = \frac{\beta}{\alpha+\beta}$$
$$\pi_1 = \frac{\alpha}{\alpha+\beta}$$

which is intuitively reasonable in that it says that the long-run probability of the two states is proportional to the probability of a switch to that state from the other. So the long-run probability of a dry day is the limit

$$\pi_0 = \lim_{n\to\infty} P(X_n = 0) = \frac{\beta}{\alpha+\beta}$$

You might try verifying this by computing the powers of the matrix P^n for $n = 1, 2, \ldots.$ and show that P^n approaches the matrix

$$\begin{bmatrix} \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \\ \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \end{bmatrix}$$

as $n \to \infty$. There are various mathematical conditions under which the limiting distribution of a Markov chain is unique and independent of the initial state of the chain but roughly they assert that the chain is such that it forgets the more and more distant past.

Independent Random Variables

Consider a Markov chain with transition probability matrix

$$P = \begin{bmatrix} 1-\alpha & \alpha \\ 1-\alpha & \alpha \end{bmatrix}$$

Notice that both rows of this matrix are identical so $P(X_{n+1} = 1|X_n = 0) = \alpha = P(X_{n+1} = 1|X_n = 1)$. For this chain the conditional distribution of X_{n+1} given $X_n = i$ evidently does not depend on the value of i. This demonstrates independence. Indeed if X and Y are two discrete random variables and if the conditional probability function $f_{y|x}(y|x)$ of Y given X is identical for all possible values of x then it must be equal to the unconditional (marginal) probability function $f_y(y)$. If $f_{y|x}(y|x) = f_y(y)$ for all values of x and y then X and Y are independent random variables. Therefore if a Markov Chain has transition probability matrix with all rows identical, it corresponds to **independent random variables** $X_1, X_2, \ldots$.. This is the most forgetful of all Markov chains. It pays no attention whatever to the current state in determining the next state.

Is the stationary distribution unique? One might wonder whether it is possible for a Markov chain to have more than one stationary distribution and consequently possibly more than one limiting distribution. We have seen that the 2×2 Markov chain with transition probability matrix

$$P = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}$$

has a solution of $\underline{\pi}^T P = \underline{\pi}^T$ and $\pi_0 + \pi_1 = 1$ given by $\pi_0 = \frac{\beta}{\alpha+\beta}$, $\pi_1 = \frac{\alpha}{\alpha+\beta}$. Is there is any other solution possible? Rewriting the equation $v^T P = v^T$ in the form $v^T(P - I) = 0$, note that the dimension of the subspace of solutions v^T is one provided that the rank of the matrix $P - I$ is one (that is, the solutions v^T are all scalar multiples of the vector $\underline{\pi}^T$), and the dimension is 2 provided that the rank of the matrix $P - I$ is 0. Only if rank$(P - I) = 0$ will there be two linear independent solutions and hence two possible candidates for equilibrium distributions. But if $P - I$ has rank 0, then $P = I$, the transition probability matrix of a very stubborn Markov chain which **always stays in the state currently occupied**. For two-dimensional Markov Chains, only in the case $P = I$ is there more than one stationary distribution and any probability vector $\underline{\pi}^T$ satisfies $\underline{\pi}^T P = \underline{\pi}^T$ and is a stationary distribution. This is at the opposite end of the spectrum from the independent case above which pays no attention to the current state in determining the next state. The chain with $P = I$ never leaves the current state.

Example (Gene Model) A simple form of inheritance of traits occurs when a trait is governed by a pair of genes A and a. An individual may have an AA of an Aa combination (in which case they are indistinguishable in appearance, or "A dominates a”$)$. Let us call an AA individual *dominant*, aa, *recessive* and Aa *hybrid*. When two individuals mate, the offspring inherits one gene of the pair from each parent, and we assume that these genes are selected at random. Now let us suppose that two individuals of opposite sex selected at random mate, and then two of their offspring mate, etc. Here the state is determined by a pair of individuals, so the states of our process can be considered to be objects like (AA, Aa) indicating that one of the pair is AA and the other is Aa (we do not distinguish the order of the pair, or male and female-assuming these genes do not depend on the sex of the individual)

Number	State
1	(AA, AA)
2	(AA, Aa)
3	(AA, aa)
4	(Aa, Aa)
5	$Aa, aa)$
6	(aa, aa)

For example, consider the calculation of $P(X_{t+1} = j|X_t = 2)$. In this case each offspring has probability $1/2$ of being a dominant AA, and probability of $1/2$ of being a hybrid (Aa). If two offspring are selected independently from this distribution the possible pairs are $(AA, AA), (AA, Aa), (Aa, Aa)$ with probabilities $1/4, 1/2, 1/4$ respectively. So the transitions have probabilities below:

	(AA, AA)	(AA, Aa)	(AA, aa)	(Aa, Aa)	(Aa, aa)	(aa, aa)
(AA, AA)	1	0	0	0	0	0
(AA, Aa)	0.25	0.5	0	0.25	0	0
(AA, aa)	0	0	0	1	0	0
(Aa, Aa)	0.0625	0.25	0.125	0.25	0.25	0.0625
(Aa, aa)	0	0	0	0.25	0.5	0.25
(aa, aa)	0	0	0	0	0	1

and transition probability matrix

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0.25 & .5 & 0 & 0.25 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0.0625 & 0.25 & 0.125 & 0.25 & 0.25 & 0.0625 \\ 0 & 0 & 0 & 0.25 & 0.5 & 0.25 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

What is the long-run behaviour in such a system? For example, the two-generation transition probabilities are given by

$$P^2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0.3906 & 0.3125 & 0.0313 & 0.1875 & 0.0625 & 0.01156 \\ 0.0625 & 0.25 & 0.125 & 0.25 & 0.25 & 0.0625 \\ 0.1406 & 0.1875 & 0.0312 & 0.3125 & 0.1875 & 0.14063 \\ 0.01562 & 0.0625 & 0.0313 & 0.1875 & 0.3125 & 0.3906 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

which seems to indicate a drift to one or other of the extreme states 1 or 6. To confirm the long-run behaviour calculate:

$$P^{100} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0.75 & 0 & 0 & 0 & 0 & 0.25 \\ 0.5 & 0 & 0 & 0 & 0 & 0.5 \\ 0.5 & 0 & 0 & 0 & 0 & 0.5 \\ 0.25 & 0 & 0 & 0 & 0 & 0.75 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

which shows that eventually the chain is absorbed in either of state 1 or state 6, with the probability of absorption depending on the initial state. This chain, unlike the ones studied before, has more than one possible stationary distribution, for example, $\pi^T = (1, 0, 0, 0, 0, 0)$ and $\pi^T = (0, 0, 0, 0, 0, 1)$, and in these circumstances the chain does not have the same limiting distribution for all initial states.

9.4 Expectation for Multivariate Distributions: Covariance and Correlation

Recall that for a discrete random variable X with probability function $f(x) = P(X = x)$ we defined

$$E\left[g\left(X\right)\right] = \sum_{\text{all } x} g(x)f(x)$$

It is easy to extend the definition of expected value to multiple discrete random variables.

Definition 33

$$E\left[g\left(X,Y\right)\right] = \sum_{\text{all } (x,y)} g(x,y)f(x,y)$$

and

$$E\left[g\left(X_1, X_2, \cdots, X_n\right)\right] = \sum_{\text{all } (x_1,x_2,\cdots,x_n)} g\left(x_1, x_2, \cdots x_n\right) f\left(x_1, \cdots, x_n\right)$$

As before, these represent the average value of $g(X,Y)$ and $g(X_1, X_2, \ldots, X_n)$. $E\left[g\left(X,Y\right)\right]$ could also be determined by finding the probability function $f_Z(z)$ of $Z = g(X,Y)$ and then using the definition of expected value $E(Z) = \sum_{\text{all } z} z f_Z(z)$.

Example: Let the joint probability function, $f(x,y)$, be given by

			x		
$f(x,y)$		0	1	2	$f_2(y)$
y	1	0.1	0.2	0.3	0.6
	2	0.2	0.1	0.1	0.4
	$f_1(x)$	0.3	0.3	0.4	1

Find $E(XY)$ and $E(X)$.

Solution:

$$\begin{aligned}
E\left(XY\right) &= \sum_{\text{all } (x,y)} xyf(x,y) \\
&= \left(0 \times 1\right)\left(0.1\right) + \left(1 \times 1\right)\left(0.2\right) + \left(2 \times 1\right)\left(0.3\right) + \left(0 \times 2\right)\left(0.2\right) + \left(1 \times 2\right)\left(0.1\right) + \left(2 \times 2\right)\left(0.1\right) \\
&= 1.4
\end{aligned}$$

To find $E(X)$ we have a choice of methods. First, taking $g(x,y)=x$ we get

$$\begin{aligned}E(X) &= \sum_{\text{all }(x,y)} xf(x,y)\\ &= (0\times 0.1)+(1\times 0.2)+(2\times 0.3)+(0\times 0.2)+(1\times 0.1)+(2\times 0.1)\\ &= 1.1\end{aligned}$$

Alternatively, since $E(X)$ only involves X, we could find $f_1(x)$ and use

$$E(X)=\sum_{x=0}^{2} xf_1(x)=(0\times 0.3)+(1\times 0.3)+(2\times 0.4)=1.1$$

Example: In the example of Section 9.2 with sprinters A, B, and C we had (using only X_1 and X_2 in our formulas)

$$f\left(x_1,x_2\right)=\frac{10!}{x_1!x_2!(10-x_1-x_2)!}(0.5)^{x_1}(0.4)^{x_2}(0.1)^{10-x_1-x_2}$$

where A wins x_1 times and B wins x_2 times in 10 races. Find $E\left(X_1X_2\right)$.

Solution: This will be similar to the way we derived the mean of the Binomial distribution but, since this is a Multinomial distribution, we'll be using the Multinomial Theorem to evaluate the sum.

$$E\left(X_1X_2\right)=\sum x_1x_2 f(x_1,x_2)$$

$$=\sum_{\substack{x_1\neq 0\\ x_2\neq 0}} x_1x_2\frac{10!}{x_1(x_1-1)!x_2(x_2-1)!(10-x_1-x_2)!}(0.5)^{x_1}(0.4)^{x_2}(0.1)^{10-x_1-x_2}$$

$$=\sum_{\substack{x_1\neq 0\\ x_2\neq 0}}\frac{(10)(9)(8!)}{(x_1-1)!(x_2-1)!\left[(10-2)-(x_1-1)-(x_2-1)\right]!}$$

$$\times(0.5)(0.5)^{x_1-1}(0.4)(0.4)^{x_2-1}(0.1)^{(10-2)-(x_1-1)-(x_2-1)}$$

$$=90(0.5)(0.4)\sum_{\substack{x_1\neq 0\\ x_2\neq 0}}\frac{8!}{(x_1-1)!(x_2-1)!\left[8-(x_1-1)-(x_2-1)\right]!}(0.5)^{x_1-1}(0.4)^{x_2-1}(0.1)^{8-(x_1-1)-(x_2-1)}$$

Let $y_1=x_1-1$ and $y_2=x_2-1$ in the sum and we obtain

$$\begin{aligned}E\left(X_1X_2\right) &= 18\sum_{(y_1,y_2)}\frac{8!}{y_1!y_2!(8-y_1-y_2)!}(0.5)^{y_1}(0.4)^{y_2}(0.1)^{8-y_1-y_2}\\ &= 18(0.5+0.4+0.1)^8=18\quad\text{by the Multinomial Theorem}\end{aligned}$$

Property of Multivariate Expectation: It is easily proved (make sure you can do this) that

$$E\left[ag_1(X,Y)+bg_2(X,Y)\right]=aE\left[g_1(X,Y)\right]+bE\left[g_2(X,Y)\right]$$

This can be extended beyond 2 functions g_1 and g_2, and beyond 2 variables X and Y.

Relationships between Variables

Independence is a "yes/no" way of defining a relationship between variables. We all know that there can be different types of relationships between variables which are dependent. For example, if X is your height in inches and Y your height in centimeters the relationship 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'll look at two ways of measuring the strength of the relationship between two random variables. The first is called covariance.

Definition 34 *The* ***covariance*** *of X and Y, denoted $Cov(X,Y)$ or σ_{XY}, is*

$$Cov(X,Y)=E\left[(X-\mu_X)(Y-\mu_Y)\right]$$

Note that

$$\begin{aligned}Cov(X,Y)&=E\left[\left(X-\mu_X\right)\left(Y-\mu_Y\right)\right]\\&=E\left(XY-\mu_XY-X\mu_Y+\mu_X\mu_Y\right)\\&=E(XY)-\mu_XE(Y)-\mu_YE(X)+\mu_X\mu_Y\\&=E(XY)-E(X)E(Y)-E(Y)E(X)+E(X)E(Y)\\&=E(XY)-E(X)E(Y)\end{aligned}$$

and $Cov(X,Y)=E(XY)-E(X)E(Y)$ is the formula we usually use for calculation purposes.

Example: Find $Cov(X,Y)$ if X and Y have joint probability function:

			x		
$f(x,y)$		0	1	2	$f_2(y)$
y	1	0.1	0.2	0.3	0.6
	2	0.2	0.1	0.1	0.4
	$f_1(x)$	0.3	0.3	0.4	1

Solution: We previously calculated $E(XY) = 1.4$ and $E(X) = 1.1$. Similarly, $E(Y) = (1 \times 0.6) + (2 \times 0.4) = 1.4$. Therefore

$$Cov(X, Y) = 1.4 - (1.1)(1.4) = -0.14$$

Exercise: Calculate the covariance of X_1 and X_2 for the sprinter example. We have already found that $E(X_1 X_2) = 18$. The marginal distributions of X_1 and of X_2 are models for which we've already derived the mean. If your solution takes more than a few lines you're missing an easier solution.

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 $(X - \mu_X)$ and $(Y - \mu_Y)$ will tend to be of the same sign, whether positive or negative. Thus $(X - \mu_X)(Y - \mu_Y)$ will be positive. Hence $Cov(X, Y) > 0$. For example in Figure 9.1 we see several hundred points plotted. Notice that the majority of the points are in the two quadrants (lower left and upper right) labelled with "+" so that for these $(X - \mu_X)(Y - \mu_Y) > 0$. A minority of points are in the other two quadrants labelled "-" and for these $(X - \mu_X)(Y - \mu_Y) < 0$. 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 $average((X - \mu_X)(Y - \mu_Y)) > 0$. Presumably this implies that over the joint distribution of (X, Y), $E[(X - \mu_X)(Y - \mu_Y)] > 0$ or $Cov(X, Y) > 0$.

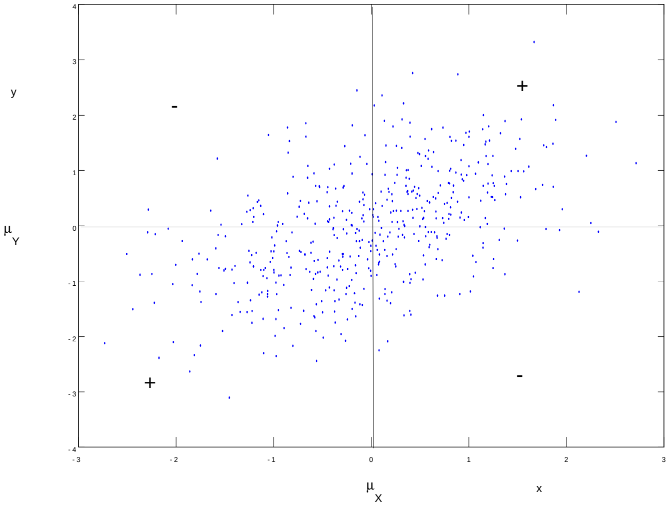


Figure 9.1: Random points (X, Y) with covariance 0.5, variances 1.

For example if X = person's height and Y = person's weight, then these two random variables will have a positive covariance.

(2) Suppose large values of X tend to occur with small values of Y and small values of X with large values of Y. Then $(X - \mu_X)$ and $(Y - \mu_Y)$ will tend to be of opposite signs. Thus $(X - \mu_X)(Y - \mu_Y)$ tends to be negative. Hence $Cov(X, Y) < 0$. For example see Figure 9.2.

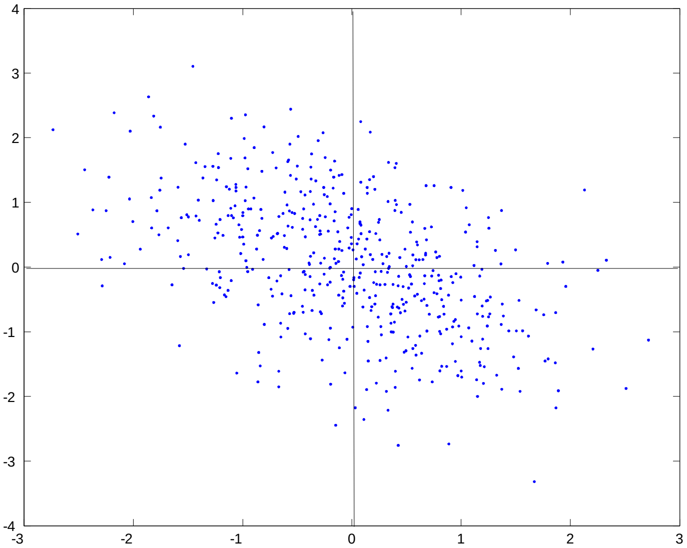


Figure 9.2: Covariance $= -0.5$, variances $= 1$

For example if X = thickness of attic insulation in a house and Y = heating cost for the house, then $Cov(X, Y) < 0$.

Theorem 35 *If X and Y are independent then $Cov(X, Y) = 0$.*

Proof: Recall $E(X - \mu_X) = E(X) - \mu_X = 0$. Let X and Y be independent.
Then $f(x, y) = f_1(x)f_2(y)$.

$$Cov(X,Y) = E[(X - \mu_X)(Y - \mu_Y)] = \sum_{\text{all } y}\left[\sum_{\text{all } x}(x - \mu_X)(y - \mu_Y) f_1(x)f_2(y)\right]$$

$$= \sum_{\text{all } y}\left[(y - \mu_Y) f_2(y) \sum_{\text{all } x}(x - \mu_X) f_1(x)\right]$$

$$= \sum_{\text{all } y}\left[(y - \mu_Y) f_2(y) E(X - \mu_X)\right]$$

$$= \sum_{\text{all } y} 0 = 0$$

The following theorem gives another way to proof the above theorem, and is useful in many other situations.

Theorem 36 *Suppose random variables X and Y are independent random variables. Then, if $g_1(X)$ and $g_2(Y)$ are any two functions,*

$$E[g_1(X)g_2(Y)] = E[g_1(X)]E[g_2(Y)]$$

Proof: Since X and Y are independent, $f(x,y) = f_1(x)f_2(y)$. Thus

$$\begin{aligned} E[g_1(X)g_2(Y)] &= \sum_{\text{all}(x,y)} g_1(x)g_2(y)f(x,y) \\ &= \sum_{\text{all } x}\sum_{\text{all } y} g_1(x)f_1(x)g_2(y)f_2(y) \\ &= [\sum_{\text{all } x} g_1(x)f_1(x)][\sum_{\text{all } y} g_2(y)f_2(y)] \\ &= E[g_1(X)]E[g_2(Y)] \end{aligned}$$

To prove result Theorem 35, we just note that if X and Y are independent then by Theorem 36

$$\begin{aligned} Cov(X,Y) &= E[(X-\mu_X)(Y-\mu_Y)] \\ &= E(X-\mu_X)E(Y-\mu_Y) \\ &= 0 \times 0 = 0 \end{aligned}$$

Caution: This result is not reversible. If $Cov(X,Y) = 0$ we can not conclude that X and Y are independent random variables. For example suppose that the random variable Z has a discrete Uniform distribution on the values $\{-1, -0.9, \ldots, 0.9, 1\}$ and define $X = \sin(2\pi Z)$ and $Y = \cos(2\pi Z)$. It is easy to see that $Cov(X,Y) = 0$ but the two random variables X, Y are clearly related because the points (X,Y) are always on a circle.

Example: Let (X,Y) have the joint probability function $f(0,0) = 0.2$, $f(1,1) = 0.6$, $f(2,0) = 0.2$; that is, (X,Y) only takes three values.

Then

x	0	1	2	Total
$f_1(x)$	0.2	0.6	0.2	1

and

y	0	1	Total
$f_2(y)$	0.4	0.6	1

are the marginal probability functions.

Since $f_1(x)f_2(y) \neq f(x,y)$, therefore, X and Y are not independent. However,

$$E\left(XY\right) = (0 \times 0 \times 0.2) + (1 \times 1 \times 0.6) + (2 \times 0 \times 0.2) = 0.6$$
$$E(X) = (0 \times 0.2) + (1 \times 0.6) + (2 \times 0.2) = 1$$

and

$$E(Y) = (0 \times 0.4) + (1 \times 0.6) = 0.6$$

Therefore $Cov(X,Y) = E(XY) - E(X)E(Y) = 0.6 - (1)(0.6) = 0$. So X and Y have covariance 0 but are not independent. If $Cov(X,Y) = 0$ we say that X and Y are uncorrelated, because of the definition of correlation [37] given below.

Exercise:

(a) Look back at the example in which $f(x,y)$ was tabulated and $Cov(X,Y) = -0.14$. Considering how covariance is interpreted, does it make sense that $Cov(X,Y)$ would be negative?

(b) Without looking at the actual covariance for the sprinter exercise, would you expect $Cov\left(X_1, X_2\right)$ to be positive or negative? (If A wins more of the 10 races, will B win more races or fewer races?)

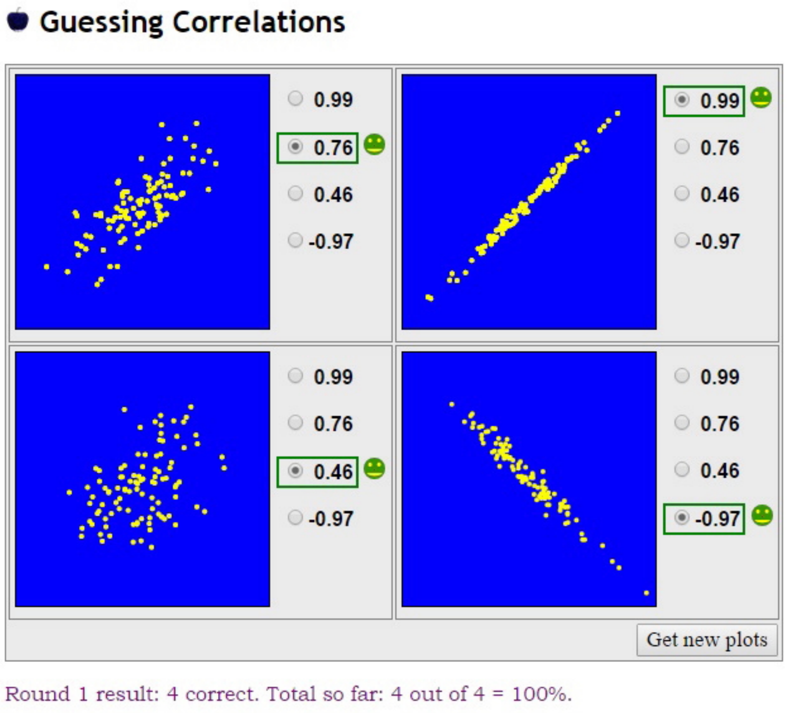
The actual numerical value of $Cov(X,Y)$ has no interpretation, so covariance is of limited use in measuring relationships. We now consider a second, related way to measure the strength of relationship between X and Y.

Definition 37 *The **correlation coefficient** of X and Y is*

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

The correlation coefficient measures the strength of the linear relationship between X and Y and is simply a rescaled version of the covariance, scaled to lie in the interval $[-1,1]$. You can attempt to guess the correlation between two variables based on a scatter diagram of values of these variables at *http://www.istics.net/Correlations/*. For example in Figure 9.3 you can see four correct guesses.

[37] "The finest things in life include having a clear grasp of correlations." Albert Einstein, 1919.



Round 1 result: 4 correct. Total so far: 4 out of 4 = 100%.

Figure 9.3: Guessing the correlation based on a scatter diagram of points

Properties of ρ:

(1) Since σ_X and σ_Y, the standard deviations of X and Y, are both positive, ρ will have the same sign as $Cov(X, Y)$. Hence the interpretation of the sign of ρ is the same as for $Cov(X, Y)$, and $\rho = 0$ if X and Y are independent. When $\rho = 0$ we say that X and Y are uncorrelated.

(2) $-1 \leq \rho \leq 1$ and as $\rho \to \pm 1$ the relation between X and Y becomes one-to-one and linear.

Proof of (2): Define a new random variable $S = X + tY$, where t is some real number. We'll show that the fact that $Var(S) \geq 0$ gives us the desired result. We have

$$
\begin{aligned}
Var(S) &= E\left[(S - \mu_S)^2\right] \\
&= E\{[(X + tY) - (\mu_X + t\mu_Y)]^2\} \\
&= E\{[(X - \mu_X) + t(Y - \mu_Y)]^2\} \\
&= E\left[(X - \mu_X)^2 + 2t(X - \mu_X)(Y - \mu_Y) + t^2(Y - \mu_Y)^2\right] \\
&= \sigma_X^2 + 2tCov(X, Y) + t^2\sigma_Y^2
\end{aligned}
$$

Since $Var(S) \geq 0$ for any real number t, this quadratic equation must have at most one real root (value of t for which it is zero). Therefore

$$[2Cov(X,Y)]^2 - 4\sigma_X^2\sigma_Y^2 \leq 0$$

leading to the inequality

$$\left|\frac{Cov(X,Y)}{\sigma_X\sigma_Y}\right| \leq 1$$

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 quadratic equation. This means that there exists one real number t^* for which

$$Var(S) = Var(X + t^*Y) = 0$$

But for $Var(X + t^*Y)$ to be zero, $X + t^*Y$ must equal a constant c. Thus X and Y satisfy a linear relationship.

Exercise: Calculate ρ for the sprinter example. Does your answer make sense? (You should already have found $Cov\,(X_1, X_2)$ in a previous exercise, so little additional work is needed.)

Problems

9.4.1 The joint probability function of (X,Y) is:

		x			
$f(x,y)$		0	1	2	
y	0	0.06	0.15	0.09	
	1	0.14	0.35	0.21	
					1

Calculate the correlation coefficient, ρ. What does it indicate about the relationship between X and Y?

9.4.2 Suppose that X and Y are random variables with joint probability function:

		x			
$f(x,y)$		2	4	6	
y	-1	$1/8$	$1/4$	p	
	1	$1/4$	$1/8$	$\frac{1}{4} - p$	
					1

(a) For what value of p are X and Y uncorrelated?

(b) Show that there is no value of p for which X and Y are independent.

9.5 Mean and Variance of a Linear Combination of Random Variables

Many problems require us to consider linear combinations of random variables; examples will be given below and in Chapter 10. Although writing down the formulas is somewhat tedious, we give here some important results about their means and variances.

Results for Means:

1. $E\left(aX+bY\right)=aE(X)+bE(Y)=a\mu_X+b\mu_Y$, when a and b are constants. (This follows from the definition of expected value.) In particular, $E\left(X+Y\right)=\mu_X+\mu_Y$ and $E\left(X-Y\right)=\mu_X-\mu_Y$.

2. Let a_i be constants (real numbers) and $E\left(X_i\right)=\mu_i$, $i=1,2,\ldots,n$. Then $E\left(\sum\limits_{i=1}^{n}a_iX_i\right)=\sum\limits_{i=1}^{n}a_i\mu_i$. In particular, $E\left(\sum\limits_{i=1}^{n}X_i\right)=\sum\limits_{i=1}^{n}E\left(X_i\right)$.

3. Let $X_1,X_2,\ldots,X_n$ be random variables which have mean μ. (You can imagine these being some sample results from an experiment such as recording the number of occupants in cars travelling over a toll bridge.) The sample mean is $\overline{X}=\frac{1}{n}\sum\limits_{i=1}^{n}X_i$. Then $E\left(\overline{X}\right)=\mu$.

Proof of (3): From (2), $E\left(\sum\limits_{i=1}^{n}X_i\right)=\sum\limits_{i=1}^{n}E\left(X_i\right)=\sum\limits_{i=1}^{n}\mu=n\mu$. Thus

$$E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n}E\left(\sum_{i=1}^{n}X_i\right)=\frac{1}{n}n\mu=\mu$$

Results for Covariance:

1. $Cov\left(X,X\right)=E\left[\left(X-\mu_X\right)\left(X-\mu_X\right)\right]=E\left[\left(X-\mu\right)^2\right]=Var(X)$

2. $Cov\left(aX+bY,cU+dV\right)=acCov\left(X,U\right)+adCov\left(X,V\right)+bcCov\left(Y,U\right)+bdCov\left(Y,V\right)$ where a,b,c, and d are constants.

Proof:

$$\begin{aligned}Cov\left(aX+bY,cU+dV\right)&=E\left[\left(aX+bY-a\mu_X-b\mu_Y\right)\left(cU+dV-c\mu_U-d\mu_V\right)\right]\\&=E\left\{\left[a\left(X-\mu_X\right)+b\left(Y-\mu_Y\right)\right]\left[c\left(U-\mu_U\right)+d\left(V-\mu_V\right)\right]\right\}\\&=acE\left[\left(X-\mu_X\right)\left(U-\mu_U\right)\right]+adE\left[\left(X-\mu_X\right)\left(V-\mu_V\right)\right]\\&+bcE\left[\left(Y-\mu_Y\right)\left(U-\mu_U\right)\right]+bdE\left[\left(Y-\mu_Y\right)\left(V-\mu_V\right)\right]\\&=acCov\left(X,U\right)+adCov\left(X,V\right)+bcCov\left(Y,U\right)+bdCov\left(Y,V\right)\end{aligned}$$

This type of result can be generalized, but the results become messy to write out.

Results for Variance:

1. **Variance of a linear combination:**

$$Var\left(aX+bY\right)=a^2Var(X)+b^2Var(Y)+2abCov\left(X,Y\right)$$

Proof:

$$\begin{aligned}
Var\left(aX+bY\right)&=E\left[\left(aX+bY-a\mu_X-b\mu_Y\right)^2\right]\\
&=E\left\{\left[a\left(X-\mu_X\right)+b\left(Y-\mu_Y\right)\right]^2\right\}\\
&=E\left[a^2\left(X-\mu_X\right)^2+b^2\left(Y-\mu_Y\right)^2+2ab\left(X-\mu_X\right)\left(Y-\mu_Y\right)\right]\\
&=a^2E\left[\left(X-\mu_X\right)^2\right]+b^2E\left[\left(Y-\mu_Y\right)^2\right]+2abE\left[\left(X-\mu_X\right)\left(Y-\mu_Y\right)\right]\\
&=a^2\sigma_X^2+b^2\sigma_Y^2+2abCov\left(X,Y\right)
\end{aligned}$$

Exercise: Try to prove this result by writing $Var\left(aX+bY\right)$ as $Cov\left(aX+bY,aX+bY\right)$ and using properties of covariance.

2. **Variance of a sum of independent random variables:** Let X and Y be independent. Since $Cov\left(X,Y\right)=0$, result 1. gives

$$Var\left(X+Y\right)=\sigma_X^2+\sigma_Y^2$$

that is, for independent variables, the *variance of a sum is the sum of the variances*. Also note

$$Var\left(X-Y\right)=\sigma_X^2+(-1)^2\sigma_Y^2=\sigma_X^2+\sigma_Y^2$$

that is, for independent variables, the variance of a difference is the sum of the variances.

3. **Variance of a general linear combination of random variables:** Let a_i be constants and $Var\left(X_i\right)=\sigma_i^2$. Then

$$Var\left(\sum_{i=1}^n a_iX_i\right)=\sum_{i=1}^n a_i^2\sigma_i^2+2\sum_{i=1}^n\sum_{j=i+1}^n a_ia_jCov\left(X_i,X_j\right)$$

This is a generalization of result 1. and can be proved using either of the methods used for 1.

4. **Variance of a linear combination of independent random variables:** Special cases of result 3. are:

   a) If $X_1, X_2, \cdots, X_n$ are independent then $Cov\,(X_i, X_j) = 0$, so that

   $$Var\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \sigma_i^2$$

   b) If $X_1, X_2, \cdots, X_n$ are independent and all have the same variance σ^2, then

   $$Var\left(\overline{X}\right) = \frac{\sigma^2}{n}$$

Proof of 4 (b): $\overline{X} = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$. From 4(a), $Var\left(\sum\limits_{i=1}^{n} X_i\right) = \sum\limits_{i=1}^{n} Var\,(X_i) = n\sigma^2$. Using $Var\,(aX+b) = a^2 Var(X)$, we get:

$$Var\left(\overline{X}\right) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Remark: This result is a very important one in probability and statistics. To recap, it says that if $X_1, \ldots, X_n$ are independent random variables with the same mean μ and same variance σ^2, then the sample mean $\bar{X} = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$ has

$$E(\bar{X}) = \mu \text{ and } Var(\bar{X}) = \frac{\sigma^2}{n}$$

This shows that the average $\bar{X}$ of n 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 just take the height of one randomly selected person. A sample of $n = 20$ persons would be better still. There is an applet at *http://users.ece.gatech.edu/users/gtz/java/samplemean/notes.html* which allows one to sample and explore the rate at which the sample mean approaches the expected value. In Section 9.7 we will see how to decide how large a sample we should take for a certain degree of precision. Also note that as $n \to \infty$, $Var(\bar{X}) \to 0$, which means that $\bar{X}$ becomes arbitrarily close to μ. This is sometimes called the "law of averages[38]". There is a formal theorem which supports the claim that for large sample sizes, sample means approach the expected value, called the "law of large numbers".

[38] "I feel like a fugitive from the law of averages."
William H. Mauldin (1921 - 2003)

Problems

9.5.1 The joint probability function of (X, Y) is given by:

			x		
$f(x,y)$		0	1	2	
y	0	0.15	0.1	0.05	
	1	0.35	0.2	0.15	
					1

Calculate $E(X)$, $Var(X)$, $Cov(X,Y)$ and $Var(3X - 2Y)$.
You may use the fact that $E(Y) = 0.7$ and $Var(Y) = 0.21$ without verifying these figures.

9.5.2 Suppose $Var(X) = 1.69$, $Var(Y) = 4$, $\rho = 0.5$. Find the standard deviation of $U = 2X - Y$.

9.5.3 Let $Y_0, Y_1, \ldots, Y_n$ be uncorrelated random variables with $E\left(Y_i\right) = 0$ and $Var\left(Y_i\right) = \sigma^2$, $i = 0, 1, \ldots, n$. Let $X_1 = Y_0 + Y_1$, $X_2 = Y_1 + Y_2, \ldots, X_n = Y_{n-1} + Y_n$.
Find $Cov\left(X_{i-1}, X_i\right)$ for $i = 2, 3, \ldots, n$ and $Var\left(\sum\limits_{i=1}^{n} X_i\right)$.

9.6 Linear Combinations of Independent Normal Random Variables

For continuous multivariate distributions we focus on linear combinations of Normal random variables which have many important applications. The following theorem gives us the results that we need for these applications.

Theorem 38 ***Linear Combinations of Independent Normal Random Variables***

(1) Let $X \sim N(\mu, \sigma^2)$ and $Y = aX + b$, where a and b are constant real numbers. Then $Y \sim N(a\mu + b, a^2\sigma^2)$

(2) Let $X \sim N\left(\mu_1, \sigma_1^2\right)$ and $Y \sim N\left(\mu_2, \sigma_2^2\right)$ independently, and let a and b be constants. Then $aX + bY \sim N\left(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2\right)$. In general if $X_i \sim N\left(\mu_i, \sigma_i^2\right)$, $i = 1, 2, \ldots, n$ independently and $a_1, a_2, \ldots, a_n$ are constants, then $\sum\limits_{i=1}^{n} a_i X_i \sim N\left(\sum\limits_{i=1}^{n} a_i \mu_i, \sum\limits_{i=1}^{n} a_i^2 \sigma_i^2\right)$.

(3) Let $X_1, X_2, \ldots, X_n$ be independent $N\left(\mu, \sigma^2\right)$ random variables. Then $\sum\limits_{i=1}^{n} X_i \sim N\left(n\mu, n\sigma^2\right)$ and $\overline{X} \sim N\left(\mu, \sigma^2/n\right)$.

Result (1) follows easily from the change of variable method discussed in Section 8.1. Result (2) is proved in Section 10.2 using moment generating functions. Result (3) is a special case of

(2). Note that the means and variances of these linear combinations of random variables can be obtained using the results of Section 9.5.

Example: Suppose $X \sim N(3,5)$ and $Y \sim N(6,14)$ independently. Find $P\left(X > Y\right)$.

Solution: Whenever we have variables on both sides of the inequality we should collect them on one side, leaving us with a linear combination. For example $P\left(X > Y\right) = P\left(X - Y > 0\right)$. Since $X - Y \sim N\left(3-6, 5+14\right) = N\left(-3, 19\right)$

$$\begin{aligned} P\left(X - Y > 0\right) &= P\left(Z > \frac{0-(-3)}{\sqrt{19}}\right) \text{ where } Z \sim N\left(0,1\right) \\ &= P\left(Z > 0.69\right) \\ &= 1 - P\left(Z \leq 0.69\right) \\ &= 1 - 0.75490 \\ &= 0.2451 \end{aligned}$$

Example: Three cylindrical parts are joined end to end to make up a shaft in a machine; 2 type A parts and 1 type B. The lengths of the parts vary a little, and have the distributions: $A \sim N\left(6, 0.4\right)$ and $B \sim N\left(35.2, 0.6\right)$. The overall length of the assembled shaft must lie between 46.8 and 47.5 or else the shaft has to be scrapped. Assume the lengths of different parts are independent. What percent of assembled shafts have to be scrapped?

Exercise: Why would it be wrong to represent the length of the shaft as $2A + B$? How would this length differ from the solution given below?

Solution: Let L, the length of the shaft, be $L = A_1 + A_2 + B$. Then

$$L \sim N\left(6+6+35.2,\ 0.4+0.4+0.6\right) = N(47.2, 1.4)$$

and so

$$\begin{aligned} &P\left(46.8 < L < 47.5\right) \\ &= P\left(\frac{46.8-47.2}{\sqrt{1.4}} < \frac{L-47.2}{\sqrt{1.4}} < \frac{47.5-47.2}{\sqrt{1.4}}\right) \\ &= P\left(-0.34 < Z < 0.25\right) \text{ where } Z \sim N\left(0,1\right) \\ &= P\left(Z < 0.25\right) - \left[1 - P\left(Z < 0.34\right)\right] \\ &= 0.59871 + 0.63307 - 1 \\ &= 0.23178 \end{aligned}$$

that is, 23.18% are acceptable and 76.82% must be scrapped. Obviously we have to find a way to reduce the variability in the lengths of the parts. This is a common problem in manufacturing.

Exercise: How could we reduce the percent of shafts being scrapped? (What if we reduced the variance of A and B parts each by 50%?)

Example: The heights of adult females in a large population is well represented by a Normal distribution with mean 64 inches and variance 6.2 (inches)^2.

(a) Find the proportion of females whose height is between 63 and 65 inches.

(b) Suppose 10 women are randomly selected, and let $\bar{X}$ be their average height, that is, $\bar{X} = \frac{1}{10}\sum\limits_{i=1}^{10} X_i$, where $X_1, X_2, \ldots, X_{10}$ are the heights of the 10 women. Find $P(63 \leq \bar{X} \leq 65)$.

(c) Suppose $\bar{X} = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$, is the average height of n women selected at random. Find the smallest value of n such that $P\left(\left|\bar{X} - 64\right| \leq 1\right) \geq 0.95$.

Solution:

(a) Let $X \sim N(64, 6.2)$ be the height X of a randomly chosen female. Then

$$\begin{aligned}
&P(63 \leq X \leq 65)\\
&= P\left(\frac{63-64}{\sqrt{6.2}} \leq \frac{X-64}{\sqrt{6.2}} \leq \frac{65-64}{\sqrt{6.2}}\right)\\
&= P(-0.40 \leq Z \leq 0.40) \quad \text{where } Z \sim N(0,1)\\
&= 2P(Z \leq 0.40) - 1\\
&= 2(0.65542) - 1\\
&= 0.31084
\end{aligned}$$

and therefore 31% of females have a height between 63 and 65 inches.

(b) $\bar{X} \sim N\left(64, \frac{6.2}{10}\right)$ so

$$\begin{aligned}
&P\left(63 \leq \bar{X} \leq 65\right)\\
&= P\left(\frac{63-64}{\sqrt{0.62}} \leq \frac{\bar{X}-64}{\sqrt{0.62}} \leq \frac{65-64}{\sqrt{0.62}}\right)\\
&= P(-1.27 \leq Z \leq 1.27) \quad \text{where } Z \sim N(0,1)\\
&= 2P(Z \leq 1.27) - 1\\
&= 2(0.89796) - 1 = 0.79592
\end{aligned}$$

(c) Since $\bar{X} \sim N\left(64, \frac{6.2}{n}\right)$ we want

$$\begin{aligned}
&P\left(|\bar{X} - 64| \leq 1\right)\\
&= P\left(\frac{|\bar{X} - 64|}{\sqrt{6.2/n}} \leq \frac{1}{\sqrt{6.2/n}}\right)\\
&= P\left(|Z| \leq \sqrt{\frac{n}{6.2}}\right) \quad \text{where } Z \sim N\left(0, 1\right)\\
&\geq 0.95
\end{aligned}$$

But $P\left(|Z| \leq 1.96\right) = 0.95$, so we need $\sqrt{\frac{n}{6.2}} \geq 1.96$ or $n \geq \left(1.96\right)^2\left(6.2\right) = 23.82$. Since n must be an integer, the smallest value of n is 24.

Remark: This shows that if we were to select a random sample of $n = 24$ persons, then their average height $\bar{X}$ would be within 1 inch of the average height $\mu = 64$ of the whole population of women. So if we did not know μ then we could estimate it to within ± 1 inch (with probability 0.95) by taking a sample of only $n = 24$ persons which is not very large.

Exercise: Find the smallest value of n such that $P(|\bar{X} - 64| \leq 0.5) \geq 0.95$.

These ideas form the basis of statistical sampling and estimation of unknown parameter values in populations and processes (STAT 231). If $X \sim N(\mu, \sigma^2)$ and we know roughly what σ is, but don't know μ, then we can use the fact that $\bar{X} \sim N(\mu, \sigma^2/n)$ to find the probability that the mean $\bar{X}$ from a sample of size n will be within a given distance of the unknown mean μ.

Problems

9.6.1 Let $X \sim N(10, 4)$ and $Y \sim N(3, 100)$ be independent. Find:

(a) $P\left(8.4 < X < 12.2\right)$

(b) $P\left(2Y > X\right)$

(c) $P\left(\overline{Y} < 0\right)$ where $\overline{Y}$ is the sample mean of 25 independent observations on Y.

9.6.2 Suppose $X \sim N(5, 4)$ and independently $Y \sim G\left(7, 3\right)$. Find:

(a) The probability $2X$ differs from Y by more than 4.

(b) Suppose $\bar{X} = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$ where $X_i \sim N(5, 4)$, $i = 1, 2, \ldots, n$ independently. Find the smallest value of n such that
$P\left(|\overline{X} - 5| < 0.1\right) \geq 0.98$.

9.7 Indicator Random Variables

The results for linear combinations of random variables provide a way of breaking up more complicated problems, involving mean and variance, into simpler pieces using indicator variables; an indicator variable is just a binary variable (0 or 1) that indicates whether or not some event occurs. We'll illustrate this important method with 3 examples.

Example: Mean and Variance of a Binomial Random Variable
Let $X \sim Binomial(n, p)$. Define new random variables X_i by

$$\begin{aligned} X_i &= 0 \text{ if the } i\text{'th trial was a failure} \\ X_i &= 1 \text{ if the } i\text{'th trial was a success.} \end{aligned}$$

The random variable X_i indicates whether the outcome "success" occurred on the i'th trial. The trick we use is that the total number of successes, X, is the sum of the X_i's:

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

We can find the mean and variance of X_i and then use our results for the mean and variance of a sum to get the mean and variance of X. First,

$$E\left(X_i\right) = \sum_{x_i=0}^{1} x_i f\left(x_i\right) = 0f(0) + 1f(1) = f(1)$$

But $f(1) = p$ since the probability of success is p on each trial. Therefore $E\left(X_i\right) = p$. Since $X_i = 0$ or 1, $X_i = X_i^2$, and therefore

$$E\left(X_i^2\right) = E\left(X_i\right) = p.$$

Thus

$$Var\left(X_i\right) = E\left(X_i^2\right) - \left[E\left(X_i\right)\right]^2 = p - p^2 = p(1-p).$$

In the Binomial distribution the trials are independent so the X_i's are also independent. Thus

$$\begin{aligned} E(X) &= E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E\left(X_i\right) = \sum_{i=1}^{n} p = np \\ Var(X) &= Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var\left(X_i\right) = \sum_{i=1}^{n} p(1-p) = np(1-p) \end{aligned}$$

These, of course, are the same as we derived previously for the mean and variance of the Binomial distribution. Note how simple the derivation here is!

Remark: If X_i is a binary random variable with $P(X_i = 1) = p = 1 - P(X_i = 0)$ then $E(X_i) = p$ and $Var(X_i) = p(1-p)$, as shown above. (Note that $X_i \sim Binomial(1,p)$ is actually a Binomial random variable.) In some problems the X_i's are not independent, and then we also need covariances.

Example: Let X have a Hypergeometric distribution. Find the mean and variance of X.

Solution: As above, let us think of the setting, which involves drawing n items at random from a total of N, of which r are "S" and $N - r$ are "F" items. Define

$$X_i = \begin{cases} 0 & \text{if } i\text{'th draw is a failure } (F) \text{ item} \\ 1 & \text{if } i\text{'th draw is a success } (S) \text{ item.} \end{cases}$$

Then $X = \sum_{i=1}^{n} X_i$ as for the Binomial example, but now the X_i's are dependent. (For example, what we get on the first draw affects the probabilities of S and F for the second draw, and so on.) Therefore we need to find $Cov(X_i, X_j)$ for $i \neq j$ as well as $E(X_i)$ and $Var(X_i)$ in order to use our formula for the variance of a sum.

We see first that $P(X_i = 1) = r/N$ for each of $i = 1, 2 \ldots, n$. (If the draws are random then the probability an S occurs in draw i is just equal to the probability position i is an S when we arrange r S's and $N - r$ F's in a row.) This immediately gives

$$\begin{aligned} E(X_i) &= \frac{r}{N} \\ Var(X_i) &= \frac{r}{N}\left(1 - \frac{r}{N}\right) \end{aligned}$$

since

$$Var(X_i) = E(X_i^2) - [E(X_i)]^2 = E(X_i) - [E(X_i)]^2$$

The covariance of X_i and X_j $(i \neq j)$ is equal to $E(X_i X_j) - E(X_i)E(X_j)$, so we need

$$\begin{aligned} E(X_i X_j) &= \sum_{x_i=0}^{1} \sum_{x_j=0}^{1} x_i x_j f(x_i, x_j) \\ &= f(1,1) \\ &= P(X_i = 1, X_j = 1) \end{aligned}$$

The probability of an S on both draws i and j is just

$$\left(\frac{r}{N}\right)\left(\frac{r-1}{N-1}\right) = P(X_i = 1)P(X_j = 1|X_i = 1)$$

Thus,

$$\begin{aligned}
Cov\left(X_i, X_j\right) &= E\left(X_i X_j\right) - E\left(X_i\right) E\left(X_j\right) \\
&= \frac{r(r-1)}{N(N-1)} - \left(\frac{r}{N}\right)\left(\frac{r}{N}\right) = \left(\frac{r}{N}\right)\left(\frac{r-1}{N-1} - \frac{r}{N}\right) \\
&= -\frac{r(N-r)}{N^2(N-1)}
\end{aligned}$$

(Does it make sense that $Cov\left(X_i, X_j\right)$ is negative? If you draw a success in draw i, are you more or less likely to have a success on draw j?)

Now we find $E(X)$ and $Var(X)$. First,

$$E(X) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E\left(X_i\right) = \sum_{i=1}^{n} \frac{r}{N} = n\left(\frac{r}{N}\right)$$

Before finding $Var(X)$, how many combinations X_i, X_j are there for which $i < j$? Each i and j takes values from $1, 2, \ldots, n$ so there are $\binom{n}{2}$ different combinations of (i, j) values. Each of these can only be written in one way to make $i < j$. There are $\binom{n}{2}$ combinations with $i < j$ (e.g. if $i = 1, 2, 3$ and $j = 1, 2, 3$, the combinations with $i < j$ are $(1, 2)$, $(1, 3)$ and $(2, 3)$. So there are $\binom{3}{2} = 3$ different combinations.) Therefore

$$\begin{aligned}
Var(X) = Var\left(\sum_{i=1}^{n} X_i\right) &= \sum_{i=1}^{n} Var\left(X_i\right) + 2 \sum_{i<j} Cov\left(X_i, X_j\right) \\
&= n \frac{r(N-r)}{N^2} + 2\binom{n}{2}\left[-\frac{r(N-r)}{N^2(N-1)}\right] \\
&= n\left(\frac{r}{N}\right)\left(\frac{N-r}{N}\right)\left[1 - \frac{(n-1)}{(N-1)}\right] \quad \left(\text{since } 2\binom{n}{2} = \frac{2n(n-1)}{2} = n(n-1)\right) \\
&= n\left(\frac{r}{N}\right)\left(1 - \frac{r}{N}\right)\left(\frac{N-n}{N-1}\right)
\end{aligned}$$

In the last two examples, we know $f(x)$, and could have found $E(X)$ and $Var(X)$ without using indicator variables. In the next example $f(x)$ is not known and is difficult to find, but we can still use indicator variables for obtaining μ and σ^2. The following example is a famous problem in probability.

Example: We have N letters to N different people, and N envelopes addressed to those N people. One letter is put in each envelope at random. Find the mean and variance of the number of letters placed in the right envelope.

Solution:

$$\text{Let } X_i = \begin{cases} 0 & \text{if letter } i \text{ is not in envelope } i \\ 1 & \text{if letter } i \text{ is in envelope } i. \end{cases}$$

Then $\sum\limits_{i=1}^{N} X_i$ is the number of correctly placed letters. Once again, the X_i's are dependent (Why?).

First

$$E\left(X_i\right) = \sum_{x_i=0}^{1} x_i f(x_i) = f(1) = \frac{1}{N} = E\left(X_i^2\right)$$

since there is 1 chance in N that letter i will be put in envelope i and then,

$$Var\left(X_i\right) = E\left(X_i\right) - \left[E\left(X_i\right)\right]^2 = \frac{1}{N} - \frac{1}{N^2} = \frac{1}{N}\left(1 - \frac{1}{N}\right)$$

Exercise: Before calculating $Cov\left(X_i, X_j\right)$, what sign do you expect it to have? If letter i is correctly placed does that make it more or less likely that letter j will be placed correctly?

Next, $E\left(X_i X_j\right) = f(1,1)$. (As in the last example, this is the only non-zero term in the sum.) Now, $f(1,1) = \frac{1}{N}\frac{1}{N-1}$ since once letter i is correctly placed there is 1 chance in $N-1$ of letter j going in envelope j. Therefore

$$E\left(X_i X_j\right) = \frac{1}{N(N-1)}$$

For the covariance we have

$$\begin{aligned} Cov\left(X_i, X_j\right) &= E\left(X_i X_j\right) - E\left(X_i\right) E\left(X_j\right) \\ &= \frac{1}{N(N-1)} - \left(\frac{1}{N}\right)\left(\frac{1}{N}\right) = \frac{1}{N}\left(\frac{1}{N-1} - \frac{1}{N}\right) \\ &= \frac{1}{N^2(N-1)} \end{aligned}$$

Therefore

$$\begin{aligned} E\left(\sum_{i=1}^{N} X_i\right) &= \sum_{i=1}^{N} E\left(X_i\right) \\ &= \sum_{i=1}^{N} \frac{1}{N} = \left(\frac{1}{N}\right) N \\ &= 1 \end{aligned}$$

and

$$
\begin{aligned}
Var\left(\sum_{i=1}^{N} X_i\right) &= \sum_{i=1}^{N} Var\left(X_i\right) + 2\sum_{i<j} Cov\left(X_i, X_j\right) \\
&= \sum_{i=1}^{N} \frac{1}{N}\left(1-\frac{1}{N}\right) + 2\binom{N}{2}\frac{1}{N^2(N-1)} \\
&= N\frac{1}{N}\left(1-\frac{1}{N}\right) + 2\binom{N}{2}\frac{1}{N^2(N-1)} \\
&= 1-\frac{1}{N} + 2\frac{N(N-1)}{2}\frac{1}{N^2(N-1)} \\
&= 1
\end{aligned}
$$

Common sense often helps in this course, but for this example there is no way of being able to say this result is obvious. On average one letter will be correctly placed and the variance will be one, regardless of how many letters there are.

Problems

9.7.1 In a row of 25 switches, each is considered to be “on” or “off”. The probability of being on is 0.6 for each switch, independently of other switch. Find the mean and variance of the number of unlike pairs among the 24 pairs of adjacent switches.

9.7.2 A plastic fabricating company produces items in strips of 24, with the items connected by a thin piece of plastic:

$$\boxed{\text{Item 1}}-\boxed{\text{Item 2}}-\boxed{\text{Item 3}}-\cdots-\boxed{\text{Item 23}}-\boxed{\text{Item 24}}$$

A cutting machine then cuts the connecting pieces to separate the items, with the 23 cuts made independently. There is a 10% chance the machine will fail to cut a connecting piece. Find the mean and standard deviation of the number of the 24 items which are completely separate after the cuts have been made. (**Hint**: Let $X_i = 0$ if item i is not completely separate, and $X_i = 1$ if item i is completely separate.)

9.8 Chapter 9 Problems

1. The joint probability function of (X, Y) is given by:

			x		
$f(x,y)$		0	1	2	
y	0	0.15	0.1	0.05	
	1	0.35	0.2	0.15	
					1

   (a) Find the marginal probability function of X and the marginal probability function of Y.

   (b) Are X and Y independent random variables? Why?

   (c) Find $P(X > Y)$.

   (d) Find the conditional probability function of X given $Y = 0$.

   (e) Find the probability function of $T = X + Y$.

2. Consider Chapter 2, Problem 7, which concerned machine recognition of handwritten digits. Recall that $p(x, y)$ was the probability that the number actually written was x, and the number identified by the machine was y.

   (a) Are X and Y independent random variables? Why?

   (b) Find $P(X = Y) =$ probability that a random number is correctly identified.

   (c) If the number written is a 5, what is the probability that it is incorrectly identified?

3. 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 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 probability function of X and Y.

   (b) Find the marginal probability function of X and the marginal probability function of Y.

   (c) Find $P(X = Y)$.

   (d) Find $P(X = 1|Y = 0)$.

4. A box contains 5 yellow and 3 red balls, from which 4 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 Y the number of yellow balls on all 4 draws.

   (a) Find the joint probability function of X and Y.

   (b) Find the marginal probability function of X and the marginal probability function of Y.

   (c) Are X and Y independent random variables? Why?

5. For a person whose car insurance and house insurance are with the same company, let X and Y represent the number of claims on the car and house policies, respectively, in a given year. Suppose that for a certain group of individuals, $X \sim Poisson\,(0.1)$ and $Y \sim Poisson\,(0.05)$.

   (a) If X and Y are independent random variables, find $P(X + Y > 1)$.

   (b) If X and Y are independent random variables, find the mean and variance of $X + Y$,

   (c) Suppose it was learned that $P(X = 0, Y = 0)$ was very close to 0.94. Show why X and Y cannot be independent random variables in this case. What might explain the non-independence?

6. Let X and Y be discrete random variables with joint probability function

$$f(x, y) = \frac{2^{x+y}e^{-4}}{x!y!} \text{ for } x = 0, 1, 2, \ldots \text{ and } y = 0, 1, 2, \ldots$$

   (a) Find the marginal probability function of X and the marginal probability function of Y without evaluating any sums.

   (b) Find the probability function of the random variable $T = X + Y$.

7. In an auto parts company an average of μ defective parts are produced per shift. The number, X, of defective parts produced has a Poisson distribution. An inspector checks all parts prior to shipping them, but there is a 10% chance that a defective part will slip by undetected. Let Y be the number of defective parts the inspector finds on a shift. Find the conditional probability function of X given $Y = y$. (The company wants to know how many defective parts are produced, but can only know the number which were actually detected.)

8. Guitars are strung with six different strings, numbered $1, 2, \ldots, 6$. Strings are sold in a package, each containing all six strings. Suppose each of the six strings are equally likely to break.

   (a) Suppose at a concert, a guitarist had five guitar strings break (for example, string 1, then string 3, then string 1 again, etc.). What is the probability that there were three or more strings of the same number?

   (b) If the number of strings that break during a concert has a $Poisson\,(1)$ distribution, how many packages should the guitarist take to the concert to ensure that with probability at least 0.99 they have enough strings?

9. (**Infection Control**) The basic reproductive rate for a disease, R_0 is the expected number of cases directly generated by one case in a population where all individuals are susceptible to infection. For a highly contagious disease spread by respiratory droplets such as COVID-19, you might think of this as the expected number of "close" contacts an infected individual has over the period in which this individual sheds virus (in the early stages of the COVID-19 pandemic it appeared to be around 2.5). Of course behaviour such as social distancing and the wearing of masks, as well as immunity achieved through vaccination or previous exposure to the disease will reduce the effective value of R_0. (This reduced value is often denoted R_e). Suppose that the actual number of an individual's close contacts, N, is a random variable with the Poisson(R_0) distribution. Now suppose that the fraction of the population that is NOT susceptible to infection either through previous contact with the virus or through immunization, is h and the fraction that is susceptible is $s = 1 - h$. If the population is randomly mixed, then an infected individual passes the infection to a total of X others, where the conditional distribution of X, given the number of close contacts N, has a $Binomial\,(N, s)$ distribution.

   (a) Find the marginal distribution of X.

   (b) Determine how large h needs to be so that $E(X) < 1$. ($E(X) < 1$ means that in the next period, the expected number of infected individuals has decreased.) Find the variance of X.

   (c) Now suppose that $m = 10000$, that there are initially $x_0 = 10$ infected individuals, and the initial fraction of susceptibles is $s_0 = 1$. What is the expected total number of cases after 4 periods (for COVID-19, 4 periods is less than a month).

10. In a breeding experiment involving horses the offspring are of four genetic types with probabilities:

Type	1	2	3	4	Total
Probability	$\frac{3}{16}$	$\frac{5}{16}$	$\frac{5}{16}$	$\frac{3}{16}$	1

A group of 40 independent offspring are observed.

(a) Find the probability that there are 10 offspring of each type.

(b) Find the probability that the total number of types 1 and 2 is 16.

(c) Find the probability that there are exactly 10 offspring of type 1, given that the total number of types 1 and 2 is 16.

11. Bacteria are distributed through river water according to a Poisson process with an average of 5 per 100 cc of water. Five 50 cc samples of water are collected. Find the probability that exactly one sample has no bacteria and exactly two samples have one bacterium.

12. A certain type of light bulb has lifetimes that can be modelled by an Exponential distribution with mean 1000 hours.

(a) What proportion of light bulbs last less than 500 hours? between 500 and 1000 hours? between 1000 and 1500 hours? longer than 1500 hours?

(b) For a carton of 50 light bulbs, find the probability that 15 light bulbs last less than 500 hours, 15 light bulbs last between 500 and 1000 hours, and 10 light bulbs last between 1000 and 1500 hours.

(c) For a carton of 50 light bulbs find the probability that 10 or more light bulbs last longer than 1500 hours.

13. For Chapter 8, Problem 15 suppose you have a class of 50 students.

(a) What is the probability that exactly 5 students receive A's, 15 students receive B's, 10 students receive C's and 15 students receive D's.

(b) What is the probability that at least 45 students receive marks above an F?

(c) What is the joint probability function of the number of students who receive A's and the number of students who receive B's?

14. In a particular city, the probability a call to a fire department concerns various situations is as given below:

Type	**Probability**
1. fire in a detached home	$p_1 = 0.10$
2. fire in a semi detached home	$p_2 = 0.05$
3. fire in an apartment or multiple unit residence	$p_3 = 0.05$
4. fire in a non-residential building	$p_4 = 0.15$
5. non-fire-related emergency	$p_5 = 0.15$
6. false alarm	$p_6 = 0.50$

Let X_i represent the numbers of calls of type i, $i = 1, 2, \ldots, 6$ in a set of 10 calls.

(a) Give the joint probability function for $X_1, X_2, \ldots, X_6$.

(b) What is the probability there is at least one apartment fire, given that there are 4 fire-related calls?

(c) If the average costs of calls of types $1, 2, \ldots, 6$ are (in \$100 units) 5, 5, 7, 20, 4, 2 respectively, what is the expected total cost of the 10 calls?

15. Blood donors arrive at a clinic and are classified as type A, type O, or other types. The blood types of donors are independent with P(type A) = p, P(type O) = q, and P(other type) = $1-p-q$. Let $X =$ number of of type A donors and $Y =$ number of type O donors arriving before the tenth other type.

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

(b) Find the marginal probability function of X.

(c) Find the conditional probability function of Y given $X = x$.

16. Accidents occur on Wednesday's at a particular intersection at random at the average rate of θ accidents per Wednesday according to a Poisson process. Define the random variable
$X_i =$ number of accidents on Wednesday at this intersection in week i, $i = 1, 2, \ldots, n$.

(a) Suppose $n = 6$ and the number of accidents observed on 6 consecutive Wednesday's was 0, 2, 0, 1, 3, 1. What is the probability of observing these data if $\theta = 1$? (Remember the Poisson process assumption that the number of events in non-overlapping time intervals are independent.)

(b) Suppose θ is unknown. What is the probability of observing these data as a function of θ?

17. For the joint probability function in Problem 1, find the correlation coefficient of X and Y.

18. If X and Y are random variables with $Var(X) = 13$, $Var(Y) = 34$ and $\rho = -0.7$ then find $Var\,(X - 2Y)$.

19. Let X and Y be independent random variables with $E(X) = E(Y) = 0$, $Var(X) = 1$ and $Var(Y) = 2$. Find $Cov\,(X + Y, X - Y)$.

20. Jane and Jack each toss a fair coin twice. Let X be the number of heads Jane obtains and Y the number of heads Jack obtains. Define $U = X + Y$ and $V = X - Y$.

    (a) Find $E\,(U)$ and $E\,(V)$.

    (b) Find $Var\,(U)$ and $Var\,(V)$.

    (c) Find $Cov(U, V)$. Are U and V independent random variables? Why?

21. Let X and Y be random variables with joint probability function

$$f(x,y) = \frac{n!}{x!y!(n-x-y)!} p^x q^y (1-p-q)^{n-x-y} \quad \text{for} \quad \begin{array}{l} x = 0, 1, \ldots, n \\ y = 0, 1, \ldots, n \\ \text{and } x + y \leq n \end{array}$$

    (a) What is the distribution of $T = X + Y$? Either explain why or derive this result.

    (b) Find $E(T)$ and $Var(T)$?

    (c) Using (b) find $Cov(X, Y)$, and explain why you expect it to have the sign it does.

22. In a particular city, let the random variable X represent the number of children in a randomly selected household, and let Y represent the number of female children. Assume that the probability a child is female is 0.5, regardless of the size of the household they live in, and that the marginal distribution of X is as follows:

x	0	1	2	3	4	5	6	7	8	Total
$P\,(X = x)$	0.20	0.25	0.35	0.10	0.05	0.02	0.01	0.01	0.01	1

    (a) Find $E(X)$.

    (b) Find the marginal probability function for the random variable Y = number of girls in a randomly chosen family. Find $E(Y)$.

23. Suppose X and Y are discrete random variables with joint probability function $f(x, y)$. If $g(x, y)$ is a function such that $a \leq g(x, y) \leq b$ for all (x, y) in the range of (X, Y)

    then show that $a \leq E[g(X, Y)] \leq b$.

24. Let $X_i =$ the return on stock i, $i = 1, 2, 3$. Suppose $E(X_i) = 0.08$, $i = 1, 2, 3$ and $Var(X_1) = (0.2)^2$, $Var(X_2) = (0.3)^2$, $Var(X_3) = (0.4)^2$. Assuming X_1, X_2, X_3 are independent random variables, find portfolio weights w_1, w_2, w_3 so that the linear combination $w_1X_1 + w_2X_2 + w_3X_3$ has the smallest variance among all such linear combinations subject to the constraint $w_1 + w_2 + w_3 = 1$.

25. Suppose $X_i \sim Poisson\,(\theta)$, $i = 1, 2, \ldots, n$ independently. Let

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Find $E(\bar{X})$ and $Var(\bar{X})$. What happens to $Var(\bar{X})$ as $n \to \infty$?

26. Suppose $X_i \sim Geometric\,(\theta)$, $i = 1, 2, \ldots, n$ independently. Find $E(\bar{X})$ and $Var(\bar{X})$. What happens to $Var(\bar{X})$ as $n \to \infty$?

27. Suppose $X_i \sim Exponential\,(\theta)$, $i = 1, 2, \ldots, n$ independently. Find $E(\bar{X})$ and $Var(\bar{X})$. What happens to $Var(\bar{X})$ as $n \to \infty$?

28. Suppose $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables with $E\left(X_i\right) = \mu$ and $Var\left(X_i\right) = \sigma^2$, $i = 1, 2, \ldots, n$.

    (a) Find $E\left(X_i^2\right)$.

    (b) Find $E(\bar{X})$, $Var(\bar{X})$ and $E\left[(\bar{X})^2\right]$.

    (c) Use (a) and (b) to show that $E\left(S^2\right) = \sigma^2$ where

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\left(\bar{X}\right)^2\right]$$

29. Let $X \sim G(-1.4, 1.5)$ (recall G stands for the Gaussian distribution introduced in Section 8.5) and $Y \sim N(-2.1, 4)$ independently. Find:

    (a) $P(X + Y > -6)$

    (b) $P\left(-2X + Y < 3\right)$

    (c) $P(Y < X)$

30. An automobile driveshaft is assembled by placing parts three independent pieces A, B and C end to end in a straight line. The standard deviation in the lengths of parts A, B and C are 0.6, 0.8, and 0.7 respectively.

    (a) Find the standard deviation of the length of the assembled driveshaft.

    (b) What percent reduction would there be in the standard deviation of the assembled driveshaft if the standard deviation of the length of part B were cut in half?

31. The amount of wine in a bottle has a Normal distribution with mean 1.05 liters and variance 0.0004 (liters)^2.

    (a) A bottle is labelled as containing 1 liter. What is the probability the bottle contains less than 1 liter?

    (b) The volume of a cask has a Normal distribution with mean 22 liters and variance 0.16 (liters)^2. What is the probability the contents of 20 randomly chosen wine bottles will fit inside a randomly chosen cask?

32. A turbine shaft is made up of four sections. The lengths of the sections are independent and have Normal distributions with different 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 shaft meets the specifications 28 ± 0.26?

33. 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%.

    (a) Find the probability that the average score in a random group of 25 students exceeds 70%.

    (b) Find the probability that the average scores of two distinct random groups of 25 students differ by more than 5%.

34. Suppose $X_i \sim G(\mu, \sigma)$, $i = 1, 2, \dots, n$ independently.

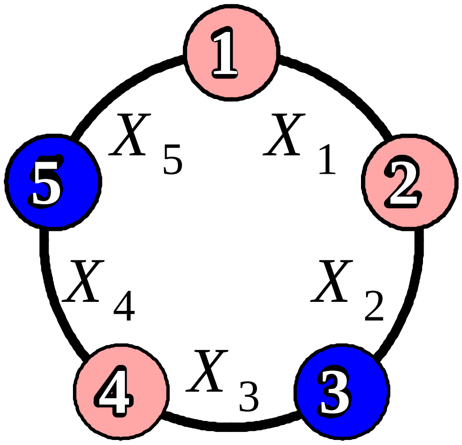
    (a) What is the distribution of
    $$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

    (b) Find $E(\bar{X})$, and $Var(\bar{X})$. What happens to $Var(\bar{X})$ as $n \to \infty$?

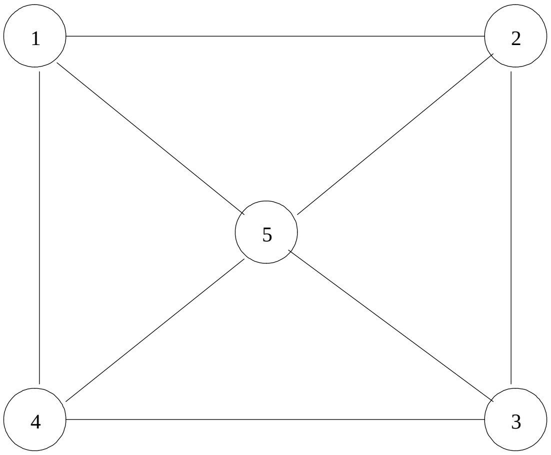
    (c) Calculate $P\left(\left|\bar{X} - \mu\right| \leq 1.96\sigma/\sqrt{n}\right)$.

    (d) If $\sigma = 12$, how large should n be to ensure that $P\left(\left|\bar{X} - \mu\right| \leq 1.0\right)$ is greater than 0.95?

35. A necklace consists of 5 beads on a string. The beads for making the necklace are drawn at random from a box containing a very large number of beads. Two-thirds of the beads are pink and one-third of the beads are blue. Let $X_1 = 1(0)$ if beads 1 and 2 are of different (same) colour, $X_2 = 1(0)$ if beads 2 and 3 are of different (same) colour, $\ldots$, and $X_5 = 1(0)$ if beads 5 and 1 are of different (same) colour. In the figure is an example of a possible necklace. For this necklace $X_1 = 0$, $X_2 = 1$, $X_3 = 1$, $X_4 = 1$, and $X_5 = 1$. Find the mean and variance of the number of unlike pairs of adjacent beads in the necklace.



36. 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.



On any day, a bridge can be closed, with probability p, for restoration work. Assuming that the eight bridges are closed independently, find the mean and variance of the number of islands which are completely cut off because of restoration work.

37. A multiple choice exam has 100 questions, each with 5 possible answers. One mark is awarded for a correct answer and 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 other questions.

    (a) Suppose that when the student does not know the answer to a question, s/he guesses randomly. Let T be the student's final mark on the exam. Show that

    $$E(T) = \sum_{i=1}^{100} p_i \text{ and } Var(T) = \sum_{i=1}^{100} p_i (1 - p_i) + \frac{1}{4}\left(100 - \sum_{i=1}^{100} p_i\right)$$

    (b) Let S be the student's final mark on the exam if s/he do not guess. Show that

    $$E(S) = \sum_{i=1}^{100} p_i \text{ and } Var(S) = \sum_{i=1}^{100} p_i (1 - p_i)$$

    (c) Compare the variances in (a) and (b) when

    (i) $p_i = 0.9$ for $i = 1, 2, \ldots, n$

    (ii) $p_i = 0.5$ for $i = 1, 2, \ldots, n$

38. **Hash Tables Continued:** See Chapter 5, Problem 26. For a hash table of size M and n keys determine the following:

    (a) the expected number of keys in a given list

    (b) the expected number of empty slots

    (c) the expected number of collisions

    (d) the expected number of keys in the table when the event "every slot has at least one key" occurs for the first time **Hint**: Let X_i = number of keys in the table when a total of i slots are assigned for the first time, $i = 1, 2, \ldots, M$ and use the approximation $\sum_{j=1}^{M} \frac{1}{j} \approx \ln M$.

39. A Markov chain has a *doubly stochastic* transition matrix if both the row sums and the column sums of the transition matrix P are all 1. Show that for such a Markov chain, the Uniform distribution on $\{1, 2, \ldots, N\}$ is a stationary distribution.

40. A salesperson named Chen sells in three cities A, B, and C. Chen never sells in the same city on successive weeks. If Chen sells in city A, then the next week Chen always sells in B. However if Chen sells in either B or C, then the next week Chen is twice as likely to sell in city A as in the other city. What is the long-run proportion of time Chen spends in each of the three cities?

41. Find $\lim\limits_{n\to\theta} P^n$ where

$$P = \begin{bmatrix} 0 & 1 & 0 \\ \frac{1}{6} & \frac{1}{2} & \frac{1}{3} \\ 0 & \frac{2}{3} & \frac{1}{3} \end{bmatrix}$$

42. Waterloo in January is blessed by many things, but not by good weather. There are never two nice days in a row. If there is a nice day, we are just as likely to have snow as rain the next day. If we have snow or rain, there is an even chance of having the same the next day. If there is change from snow or rain, only half of the time is this a change to a nice day. Taking as states the kinds of weather R, N, and S. the transition probabilities P are as follows

$$P = \left[\begin{array}{cccc} & \text{R} & \text{N} & \text{S} \\ \text{R} & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \text{N} & \frac{1}{2} & 0 & \frac{1}{2} \\ \text{S} & \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{array}\right]$$

If today is raining, find the probability of Rain, Nice, Snow three days from now. Find the probabilities of the three states in five days, given (i) today is raining (ii) today is nice (iii) today is snowing.

43. **One-card Poker**: A card game, which we will call Metzler Poker, is played as follows. Each of two players bets an initial $\$1$ and is dealt a card from a deck of 13 cards numbered $1, 2, \ldots, 13$. Upon looking at their card, each player then decides (unaware of the other's decision) whether or not to increase their bet by $\$5$ (to a total stake of $\$6$). If both increase the stake ("raise"), then the player with the higher card wins both stakes, that is, they get their money back as well as the other player's $\$6$. If one person increases and the other does not, then the player who increases automatically wins the pot. If neither person increases the stake, then it is considered a draw - each player receives their own $\$1$ back. Suppose Player A and Player B play Metzler Poker according to the following strategy. Player A choses an integer number a between 1 and 13 and similarly Player B also choses an integer number b between 1 and 13. Player A always raises whenever their dealt card is greater than or equal to their chosen value for a. Player B always raises whenever their dealt card is greater than or equal to their chosen value for b.

   (a) Let R be the event that both players raise. Determine $P\,(\text{player } B \text{ wins}|R)$ and $P\,(\text{player } A \text{ wins}|R)$. (Hint: Consider the cases $b \leq a$ and $b > a$.) Determine player A's expected winnings if both players raise. (Note: Winnings can be negative.)

   (b) Suppose player B always raises, that is, suppose that $b = 1$. Determine player A's expected winnings for the different possible values of $a = 1, 2, \ldots, 13$.

(c) If $b = 11$, determine player A's optimal expected winnings.

(d) What is the optimal (minimax) strategy for both players if both players know the values of a and b?

44. **Searching a database**: Suppose that we are given 3 records, R_1, R_2, R_3 *initially* stored in that order. The cost of accessing the j'th record in the list is j so we would like the more frequently accessed records to be near the front of the list. Whenever a request for record j is processed, the "move-to-front" heuristic stores R_j at the front of the list and the others in the original order. For example if the first request is for record 2, then the records will be re-stored in the order R_2, R_1, R_3. Assume that on each request, record j is requested with probability p_j, for $j = 1, 2, 3$.

    (a) Show that if X_j is the permutation that obtains after j requests for records (e.g. $X_2 = (2, 1, 3)$), then X_j, $j = 1, 2, \ldots$ is a Markov chain.

    (b) Find the stationary distribution of this Markov chain. **Hint:** What is the probability that X_j takes the form $(2, *, *)$ for large j?

    (c) Find the expected long-run cost per record accessed in the case $(p_1, p_2, p_3) = (0.1, 0.3, 0.6)$.

    (d) How does this expected long-run cost compare with keeping the records in random order, and with keeping them in order of decreasing values of p_j(only possible if we know p_j).

45. **Secretary Problem:** Suppose you are to interview N candidates for a job, one at a time. You must decide immediately after each interview whether to hire the current candidate or not and you wish to maximize your chances of choosing the best person for the job (there is no benefit from choosing the second or third best). For simplicity, assume candidate i has numerical value X_i chosen without replacement from $\{1, 2, \ldots, N\}$ where $1 =$ worst, $N =$ best. Our strategy is to interview k candidates first, and then pick the first of the remaining $N - k$ that has value greater than $\max(X_1, X_2, \ldots, X_k)$.

    (a) What is the best choice of k? **Hint**: use the approximation $\sum\limits_{j=1}^{n-1} \frac{1}{j} \approx \ln(n)$.

    (b) For the value of k found in (a), what is the approximate probability that you do choose the maximum?

46. **Challenge problem:** A drunken probabilist stands n steps from a cliffs edge. He takes random steps, either towards or away from the cliff, each step independent of the previous step. On each step the probability he takes a step away from the cliff is $\frac{2}{3}$ and the probability he takes a step towards the cliff is $\frac{1}{3}$. What is the probability he does not fall off the cliff?

47. **Challenge problem:** Let X be a continuous random variable with probability density function $f_1(x)$ and let Y be a discrete random variable. We define the conditional probability density function of X given $Y = y$ as

$$f_1(x|y) = \frac{d}{dx} P(X \leq x|Y = y)$$

and the conditional probability function of Y given $X = x$ as

$$f_2\left(y|x\right) = P\left(Y = y|X = x\right) = \frac{f_1(x|y)P\left(Y = y\right)}{f_1(x)}$$

(a) Show that

$$f_2\left(y\right) = P\left(Y = y\right) = \int\limits_{-\infty}^{\infty} f_2\left(y|x\right) f_1(x) dx$$

(b) Assume we have a coin which is not fair and we do not know the probability of a head for the coin. As such, we model the probability of heads by a random variable $X \sim Uniform(0, 1)$. This is an appropriate model as the probability of heads can be any number in the interval $[0, 1]$. We want to find the probability function of $Y =$ the number of heads in n tosses of the coin. Clearly,

$$P(Y = y|X = x) = \binom{n}{y} x^y (1 - x)^{n-y}$$

Use the result in part (a) to show that

$$P(Y = y) = \frac{1}{n+1}$$

Hint: Use the identity

$$\int\limits_0^1 \theta^a (1 - \theta)^b d\theta = \frac{a!b!}{(a + b + 1)!}$$

(c) Let Y be an indicator variable with $Y = 1$ if $|X| > 1$ and $Y = 0$ if $|X| \leq 1$. Find the conditional probability density function of X given $Y = 0$.

48. **Challenge problem:** Suppose $U_i \sim U\left(0, 1\right)$, $i = 1, 2, \ldots, n$ independently. Define the random variable

$$N = \min_n \left(\textstyle\sum_{i=1}^n U_i \geq k\right)$$

where k is a positive real number. What is the expected value of N? How would you approximate this expected value if k were large?

10. CENTRAL LIMIT THEOREM and MOMENT GENERATING FUNCTIONS

10.1 Central Limit Theorem

The Normal distribution can, under certain conditions, be used to approximate probabilities for linear combinations of variables having a non-Normal distribution. This remarkable property follows from an amazing result called the Central Limit Theorem (C.L.T.). There are actually several versions of the Central Limit Theorem. The version given below is one of the simplest.

Example: The major reason that the Normal distribution is so commonly used is that it tends to approximate the distribution of sums of random variables. For example, if we throw n fair dice and S_n is the sum of the outcomes, what is the distribution of S_n? The tables below provide the number of ways in which a given value can be obtained. The corresponding probability is obtained by dividing by 6^n. For example on the throw of $n = 1$ dice the probable outcomes are $1, 2, \ldots, 6$ with probabilities all $1/6$ as indicated in the first panel of the histogram in Figure 10.1

For $S_2 =$ the sum of 2 fair dice, the possible values are $\{2, 3, \ldots, 12\}$. The probability function for S_2 is:

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 is shown in the second panel of Figure 10.1.

For $S_3 =$ the sum of 3 fair dice, the possible values are $\{3, 4, \ldots, 18\}$. The probability function for S_3 is:

s	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$P(S_3 = s)$	$\frac{1}{216}$	$\frac{3}{216}$	$\frac{6}{216}$	$\frac{10}{216}$	$\frac{15}{216}$	$\frac{21}{216}$	$\frac{25}{216}$	$\frac{27}{216}$	$\frac{27}{216}$	$\frac{25}{216}$	$\frac{21}{216}$	$\frac{15}{216}$	$\frac{10}{216}$	$\frac{6}{216}$	$\frac{3}{216}$	$\frac{1}{216}$

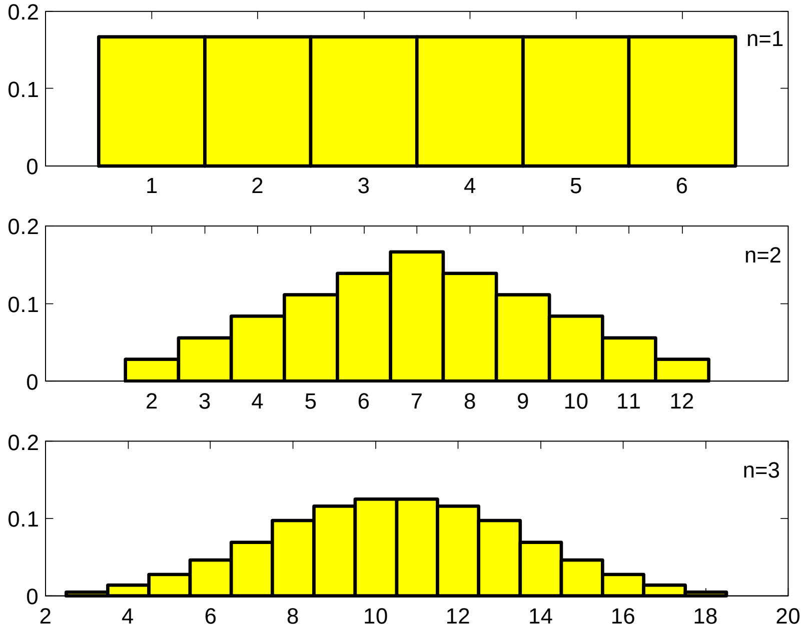


Figure 10.1: Probability histograms for the sum of n rolls of a dice for $n = 1, 2, 3$

The probability histogram is shown in the third panel of Figure 10.1. The probability histogram for S_3 already resembles a Normal probability density function. In general, these distributions show a simple pattern. For $n = 1$, the probability function is a constant (polynomial degree 0). For $n = 2$, the probability histogram can be constructed from two linear functions spliced together (polynomials of degree 1). For $n = 3$, the probability histogram can be constructed from three quadratic functions (polynomials of degree 2). The shapes of these probability histograms rapidly approach the shape of the Normal probability density function as n increases.

You can simulate the throws of any number of dice and illustrate the behaviour of the sums at *http://www.math.csusb.edu/faculty/stanton/probstat/clt.html.*

This example illustrate what happens in general with the distribution of the sum of independent random variables from any distribution. If $X_1, X_2, \ldots, X_n$ are independent discrete random variables all having the same distribution, with mean μ and variance σ^2, then as $n \to \infty$, the shape of the probability histogram for the random variable $S_n = \sum_{i=1}^{n} X_i$ approaches the shape of a $N\left(n\mu, n\sigma^2\right)$ probability density function or equivalently the shape of the probability histogram for $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ approaches the shape of a $N\left(\mu, \frac{\sigma^2}{n}\right)$ probability density function. If $X_1, X_2, \ldots, X_n$ are independent continuous random variables all having the same distribution, with mean μ and variance σ^2, then as $n \to \infty$, the shape of the probability density function of the random variable $S_n = \sum\limits_{i=1}^{n} X_i$ approaches the shape of a $N\left(n\mu, n\sigma^2\right)$ probability density function and the shape of the probability density function of $\overline{X}$ approaches the shape of a $N\left(\mu, \frac{\sigma^2}{n}\right)$ probability density function.

The following theorem is the mathematical statement of these results.

Theorem 39 [39] ***Central Limit Theorem*** *If $X_1, X_2, \ldots, X_n$ are independent random variables all having the same distribution, with mean μ and variance σ^2, then as $n \to \infty$, the cumulative distribution function of the random variable*

$$\frac{\sum\limits_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

approaches the $N(0,1)$ cumulative distribution function. Similarly, the cumulative distribution function of

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

approaches the $N(0,1)$ cumulative distribution function.

This is a theorem about limits. We will use it when n is large, but finite, to approximate the distribution of S_n or $\overline{X}$ by a Normal distribution. That is, we will use

$$S_n = \sum_{i=1}^{n} X_i \text{ has approximately a } N\left(n\mu, n\sigma^2\right) \text{ distribution for large } n$$

and

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ has approximately a } N\left(\mu, \frac{\sigma^2}{n}\right) \text{ distribution for large } n$$

Note that as $n \to \infty$, both distributions $N\left(n\mu, n\sigma^2\right)$ and $N\left(\mu, \sigma^2/n\right)$ fail to exist. (The former because both $n\mu$ and $n\sigma^2 \to \infty$, the latter because $\sigma^2/n \to 0$.)

Notes:

(1) The Central Limit Theorem does not hold if the common mean μ and variance σ^2 do not exist. The Cauchy distribution introduced in Problem 8.20 is an example of such a distribution.

(2) We use the Central Limit Theorem to approximate the distribution of the sum $S_n = \sum\limits_{i=1}^{n} X_i$ or average $\overline{X} = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$. The accuracy of the approximation depends on n (bigger is better) and

[39] A proof is given in Section 10.3.

also on the actual distribution of the X_i's. The approximation works better for small n when the shape of the probability function/probability density function of X_i is symmetric (for example, the $U(a,b)$ probability density function) or nearly symmetric (for example, the $Poisson\,(5)$ probability function).

(3) In Section 9.6, the distributions of linear combinations of independent Normal random variables were given. In particular if $X_1, X_2, \ldots, X_n$ are independent $N\left(\mu, \sigma^2\right)$ random variables then

$$S_n = \sum_{i=1}^{n} X_i \sim N\left(n\mu, n\sigma^2\right) \text{ and } \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \sigma^2/n\right)$$

Thus, if the X_i's themselves have a Normal distribution, then S_n and $\overline{X}$ have exactly Normal distributions for all values of n. If the X_i's do not have a Normal distribution themselves, then S_n and $\overline{X}$ have approximately Normal distributions when n is large. From this distinction you should be able to guess that if the shape of the probability (density) function of X_i is somewhat Normal shaped then the approximation will be good for smaller values of n. If the shape of the probability (density) function of X_i is very non-Normal shaped (for example, an $Exponential\,(\theta)$ probability density function) then the approximation will be poor for small n. (This is related to the second remark in (2)).

Example: Hamburger patties are packed eight to a box, and each box is supposed to have 1 kilogram of meat in it. The weights of the patties vary a little because they are mass produced, and the weight X of a single patty is actually a random variable with mean $\mu = 0.128$ kilogram and standard deviation $\sigma = 0.005$ kilogram. Find the probability a box has at least 1 kilogram of meat, assuming that the weights of the eight patties in any given box are independent.

Solution: Let $X_1, X_2, \ldots, X_8$ be the weights of the eight patties in a box, and $S_8 = X_1 + X_2 + \cdots + X_8$ be their total weight. By the Central Limit Theorem, S_8 has approximately a $N(8\mu, 8\sigma^2)$ distribution. We will assume this approximation is reasonable even though $n = 8$ is small. (This is likely okay because the distribution of X is likely fairly close to Normal.)

Thus $S_8 \sim N(1.024, 0.0002)$ approximately and

$$\begin{aligned} P(S_8 > 1) &\approx P\left(Z > \frac{1 - 1.024}{\sqrt{0.0002}}\right) \quad \text{where } Z \sim N(0,1) \\ &= P(Z > -1.70) = P(Z \leq 1.70) \\ &= 0.95543 \end{aligned}$$

Note: We see that only about 96% of the boxes actually have 1 kilogram or more of hamburger. What would you recommend be done to increase this probability to 99%?

Example: Suppose fires reported to a fire station satisfy the conditions for a Poisson process, with a mean of 1 fire every 4 hours. Find the probability the 500'th fire of the year is reported on the 84^{th} day of the year.

Solution: Let X_i be the time between the $(i-1)$'st and i'th fires (X_1 is the time to the first fire). Then X_i has an Exponential distribution with $\theta = 1/\lambda = 4$ hours, or $\theta = 1/6$ day. Since $S_{500} = \sum\limits_{i=1}^{500} X_i$ is the time until the 500th fire, we want to find $P\left(83 < S_{500} \leq 84\right)$. While the Exponential probability density function is very skewed and not very Normal shaped, we are summing a large number of independent Exponential variables. Hence, by the Central Limit Theorem, S_{500} 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 Exponential distribution, $\mu = \theta = 1/6$ and $\sigma^2 = \theta^2 = 1/36$ so

$$\begin{aligned}
P\left(83 < S_{500} \leq 84\right) &\approx P\left(\frac{83 - \frac{500}{6}}{\sqrt{\frac{500}{36}}} < Z \leq \frac{84 - \frac{500}{6}}{\sqrt{\frac{500}{36}}}\right) \quad \text{where } Z \sim N\left(0,1\right) \\
&= P\left(-0.09 < Z \leq 0.18\right) \\
&= P\left(Z \leq 0.18\right) - P\left(Z \leq -0.09\right) \\
&= P\left(Z \leq 0.18\right) - \left[1 - P\left(Z \leq 0.09\right)\right] \\
&= 0.57142 + 0.53586 - 1 \\
&= 0.10728
\end{aligned}$$

Example: This example is frivolous but shows how the Normal distribution can approximate even sums of discrete random variables. In an apple orchard, suppose the number X of worms in an apple has probability function:

x	0	1	2	3	Total
$f(x)$	0.4	0.3	0.2	0.1	1

Find the probability a basket with 250 apples in it has between 225 and 260 (inclusive) worms in it.

Solution:

$$\mu = E\left(X\right) = \sum_{x=0}^{3} x f(x) = 0\left(0.4\right) + 1\left(0.3\right) + 2\left(0.2\right) + 3\left(0.1\right) = 1$$

$$E\left(X^2\right) = \sum_{x=0}^{3} x^2 f(x) = \left(0\right)^2\left(0.4\right) + \left(1\right)^2\left(0.3\right) + \left(2\right)^2\left(0.2\right) + \left(3\right)^2\left(0.1\right) = 2$$

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

By the Central Limit Theorem, $S_{250} = \sum\limits_{i=1}^{250} X_i$ has approximately a $N\left(250\mu, 250\sigma^2\right)$ distribution, where X_i is the number of worms in the i'th apple. Therefore S_{250} has approximately a $N\left(250, 250\right)$

distribution and

$$\begin{aligned} P(225 \le S_{250} \le 260) &\approx P\left(\frac{225-250}{\sqrt{250}} \le Z \le \frac{260-250}{\sqrt{250}}\right) \quad \text{where } Z \sim N(0,1) \\ &= P(-1.58 \le Z \le 0.63) \\ &= P(Z \le 0.63) - [1 - P(Z \le 1.58)] \\ &= 0.73565 + 0.94295 - 1 \\ &= 0.67860 \end{aligned}$$

While this approximation is adequate, we can improve its accuracy, as follows. When X_i has a discrete distribution, as it does here, $S_n = \sum_{i=1}^{n} X_i$ will always remain discrete no matter how large n gets. So the shape of the probability histogram of S_n, while Normal shaped, will never be exactly Normal shaped. Consider a probability histogram for the random variable S_{250}, as shown in Figure 10.2. (Only part of the histogram is shown.)

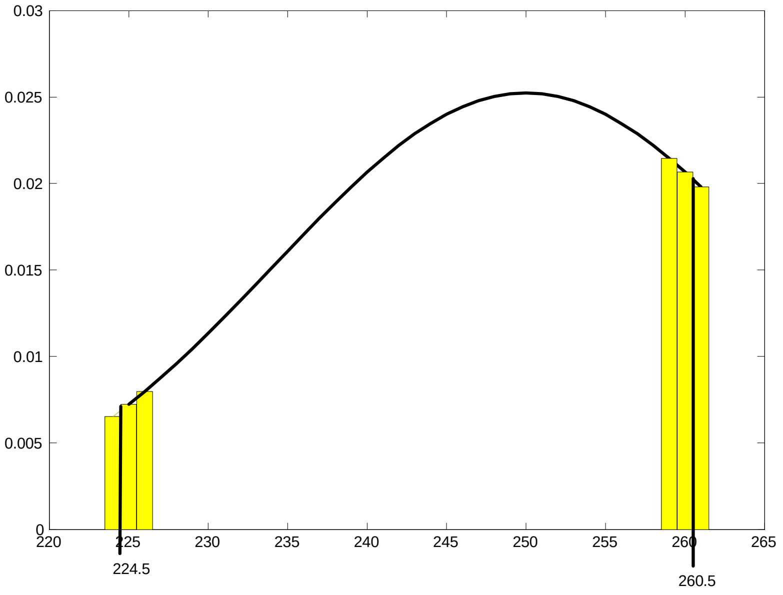


Figure 10.2: Probability histogram for S_{250}

The area of the bar on the interval $[s-0.5, s+0.5]$ is equal to $P(S_{250} = s)$. The smooth curve is the probability density function for the approximating Normal distribution. Then $\sum_{s=225}^{260} P(S_{250} = s)$ is the total area of all bars of the histogram for $s = 225, 226, \ldots, 260$. These bars actually span the interval of values $[224.5, 260.5]$. The left and right end bars are more easily seen in Figure 10.3.

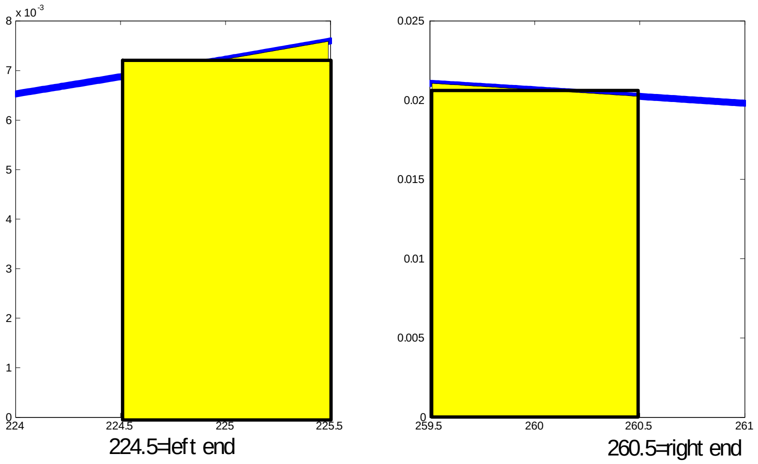


Figure 10.3: A magnification of the bars at the left and right hand end of the interval

We can obtain a more accurate approximation by finding the area under the Normal probability density function from 224.5 to 260.5, that is,

$$\begin{aligned} P\left(225 \leq S_{250} \leq 260\right) &= P\left(224.5 < S_{250} < 260.5\right) \\ &\approx P\left(\frac{224.5-250}{\sqrt{250}} < Z < \frac{260.5-250}{\sqrt{250}}\right) \quad \text{where } Z \sim N\left(0,1\right) \\ &= P\left(-1.61 < Z < 0.66\right) = 0.74537 + 0.94630 - 1 \\ &= 0.69167 \end{aligned}$$

Unless making this adjustment greatly complicates the solution, it is preferable to make this "**continuity correction**".

Suppose you required the probability of a single value such as

$$P\left(S_{250} = 225\right)$$

Then without the continuity correction we obtain the silly approximation

$$P\left(S_{250} = 225\right) \approx P\left(Z = \frac{225-250}{\sqrt{250}}\right) = 0$$

In such a case we need to use the continuity correction. We obtain

$$\begin{aligned} P\left(S_{250} = 225\right) &= P\left(224.5 < S_{250} < 225.5\right) \\ &\approx P\left(\frac{224.5-250}{\sqrt{250}} < Z < \frac{225.5-250}{\sqrt{250}}\right) \quad \text{where } Z \sim N\left(0,1\right) \\ &= P\left(-1.61 < Z < -1.55\right) = 0.9463 - 0.93953 = 0.0068 \end{aligned}$$

and although this is small, it is certainly not zero.

Notes:

(1) A continuity correction should <u>not</u> 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 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 distribution, such as $P(X = 50)$ where X is $Binomial\,(100, 0.5)$ you need to use the continuity correction. Otherwise, for approximating the Binomial with large n, it is not necessary to use the correction.

Normal Approximation to the Poisson Distribution

Let X be a random variable with a $Poisson(\mu)$ distribution and suppose μ is large. For the moment suppose that μ is an integer and recall that if we add μ independent Poisson random variables, each with parameter 1, then the sum has the Poisson distribution with parameter μ. In general, a Poisson random variable with large expected value can be written as the sum of a large number of independent random variables, and so the Central Limit Theorem implies that it must be close to Normally distributed.

Theorem 40 ***Normal Approximation to Poisson:*** *Suppose $X \sim Poisson(\mu)$. Then the cumulative distribution function of the standardized random variable*

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

approaches that of a standard Normal random variable as $\mu \to \infty$.

We prove this theorem in Section 10.2.

Example: Suppose $X \sim Poisson(\mu)$. Use the Normal approximation to approximate

$$P(X > \mu)$$

Compare this approximation with the true value when $\mu = 9$.

Solution: Theorem 40 implies that the cumulative distribution function of the standardized random variable

$$Z_\mu = \frac{X - \mu}{\sqrt{\mu}}$$

(note: identify $E(X)$ and $Var(X)$ in the above standardization) approaches the cumulative distribution function of a **standard Normal** random variable Z. In particular, without a continuity correction,

$$P(X \leq \mu) = P(Z_\mu \leq 0) \rightarrow P(Z \leq 0) = 0.5 \text{ as } \mu \rightarrow \infty$$

Computing the true value when $\mu = 9$ we obtain

$$P(X > 9) = 1 - P(X \leq 9) = 1 - \left(e^{-9} + 9e^{-9} + \frac{9^2}{2!}e^{-9} + \cdots + \frac{9^9}{9!}e^{-9}\right) = 1 - 0.5874 = 0.4126$$

The Normal approximation without a continuity correction gives

$$P(X > 9) \approx P\left(Z_\mu > \frac{9-9}{3}\right) \approx P(Z > 0) = 0.5$$

which is not close to the true value 0.4126. The Normal approximation with a continuity correction gives

$$P(X > 9) \approx P\left(Z_\mu > \frac{9.5-9}{3}\right) \approx P(Z > 0.17) = 0.4324$$

which is much closer to the true value.

Normal Approximation to the Binomial Distribution

It is well-known that the probability histogram for a Binomial distribution, at least for large values of $n,$ resembles a bell-shaped or Normal curve. The most common demonstration of this is with a mechanical device common in science museums called a "Galton board" or "Quincunx"[40] which drop balls through a mesh of equally spaced pins (see Figure 10.4). Notice that if balls either go to the right or left at each of the eight levels of pins, independently of the movement of the other balls, then
$X =$ number of moves to right has a $Binomial(8, 0.5)$ distribution. If the balls are dropped from location 0 (on the $x-$axis) then the ball eventually rests at location $2X - 8$ which is approximately Normally distributed since X is approximately Normal.

The following result is proved using the Central Limit Theorem.

Theorem 41 ***Normal Approximation to Binomial:*** *Suppose $X \sim Binomial(n, p)$. Then for n large, the random variable*

$$W = \frac{X - np}{\sqrt{np\,(1-p)}} \text{ has approximately a } N(0,1) \text{ distribution}$$

[40] The word comes from Latin quinque (five) unicia (twelve) and means five twelfths.

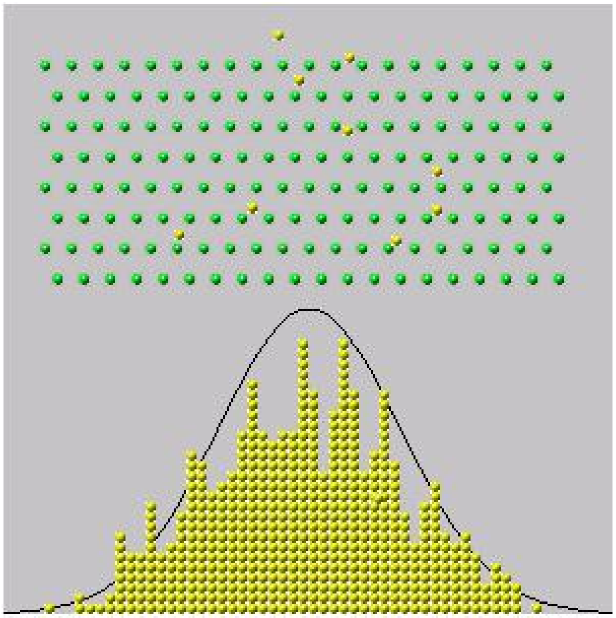


Figure 10.4: A "Galton Board" or "Quincunx"

Proof: We use indicator variables X_i, $i = 1, 2, \ldots, n$ where $X_i = 1$ if the ith trial in the Binomial process is an "S" outcome and $X_i = 0$ if it is an "F" outcome. Then $X = \sum\limits_{i=1}^{n} X_i$ and we can use the Central Limit Theorem. Since

$$\mu = E(X_i) = p, \text{ and } \sigma^2 = Var(X_i) = p(1-p)$$

we have that as $n \to \infty$ the cumulative distribution function of the random variable

$$W = \frac{\sum\limits_{i=1}^{n} X_i - np}{\sqrt{np(1-p)}} = \frac{X - np}{\sqrt{np(1-p)}}$$

approaches the cumulative distribution function of a $N(0,1)$ random variable, as required.

Remark: We can write the Normal approximation either as $\frac{X-np}{\sqrt{np(1-p)}} \sim N(0,1)$ approximately or as $X \sim N(np, np(1-p))$ approximately.

Remark: The continuity correction method can be used here. The following numerical example illustrates the procedure.

Example: Suppose $X \sim Binomial(n, p)$.
(a) If $n = 20$ and $p = 0.4$, approximate the probability $P(4 \le X \le 12)$. Compare the answer with the exact value.

(b) If $n = 100$ and $p = 0.4$, approximate the probability $P(34 \leq X \leq 48)$. Compare the answer with the exact value.

Solution (a) By the Normal approximation to the Binomial we have $X \sim N(8, 4.8)$ approximately. Without the continuity correction we have

$$\begin{aligned} P(4 \leq X \leq 12) &= P\left(\frac{4-8}{\sqrt{4.8}} \leq \frac{X-8}{\sqrt{4.8}} \leq \frac{12-8}{\sqrt{4.8}}\right) \\ &\approx P(-1.826 \leq Z \leq 1.826) \quad \text{where } Z \sim N(0,1) \\ &= 0.932 \end{aligned}$$

Using the continuity correction method, we get

$$\begin{aligned} P(4 \leq X \leq 12) &= P\left(\frac{3.5-8}{\sqrt{4.8}} \leq Z \leq \frac{12.5-8}{\sqrt{4.8}}\right) \quad \text{where } Z \sim N(0,1) \\ &\approx P(-2.054 \leq Z \leq 2.054) \\ &= 0.960 \end{aligned}$$

The exact probability is

$$\sum_{x=4}^{12} \binom{20}{x} (0.4)^x (0.6)^{20-x} = 0.963$$

which was calculated using the R function `pbinom()`. As expected the continuity correction method gives a more accurate approximation.

(b) By the Normal approximation to the Binomial we have $X \sim N(40, 24)$ approximately so without the continuity correction we have

$$\begin{aligned} P(34 \leq X \leq 48) &\approx P\left(\frac{34-40}{\sqrt{24}} \leq Z \leq \frac{48-40}{\sqrt{24}}\right) \quad \text{where } Z \sim N(0,1) \\ &= P(-1.225 \leq Z \leq 1.633) \\ &= 0.9488 - (1 - 0.8897) = 0.8385 \end{aligned}$$

With the continuity correction

$$\begin{aligned} P(34 \leq X \leq 48) &\approx P\left(\frac{33.5-40}{\sqrt{24}} \leq Z \leq \frac{48.5-40}{\sqrt{24}}\right) \quad \text{where } Z \sim N(0,1) \\ &= P(-1.327 \leq Z \leq 1.735) \\ &= 0.9586 - (1 - 0.9076) = 0.866 \end{aligned}$$

The exact value to three decimal places is

$$\sum_{x=34}^{48} \binom{100}{x} (0.4)^x (0.6)^{40-x} = 0.866$$

so the approximation and exact answer agree to three decimal places.

Note: The error of the Normal approximation decreases as n increases, but it is a good idea to use the continuity correction when it is convenient. 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 is less than 10, then the continuity correction makes a difference of $0.5/10 = 0.05$ to the number we look up in the table. 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 use the continuity correction whenever the standard deviation of the integer-valued random variable being approximated is less than 10.

Example 10.1: Binomial sample size calculation Let p be the proportion of Canadians who think Canada should adopt the US dollar.
(a) Suppose 400 Canadians are randomly chosen and asked their opinion. Let X be the number who say yes. Find the probability that the proportion, $\frac{X}{400}$, of people who say yes is within 0.02 of p, if $p = 0.20$.
(b) Suppose for a future opinion poll we want to determine the number, n, to survey to ensure that there is a 95% chance that $\frac{X}{n}$ lies within 0.02 of p. Suppose $p = 0.20$ is known.
(c) Repeat (b) when the value of p is unknown. (Note that this would be the more realistic situation in the case of conducting an opinion poll.)

Solution:

a) Since $X =$ the number of Canadians who say yes it is reasonable to assume $X \sim Binomial\,(400, 0.2)$. (How well do you think the assumptions for a Binomial distribution hold in this case?) By the Normal approximation to the Binomial we have that X has approximately a Normal distribution with mean $\mu = np = (400)(0.2) = 80$ and variance $\sigma^2 = np(1-p) = (400)(0.2)(0.8) = 64$. Therefore

$$\begin{aligned}
&P\left(\left|\frac{X}{400} - 0.2\right| \leq 0.02\right)\\
&= P\left(|X - 80| \leq 8\right) = P\left(|X - 80| \leq 8.5\right)\\
&\approx P\left(|Z| \leq \frac{8.5}{\sqrt{64}}\right) = P\left(|Z| \leq 1.06\right)\\
&= 2P\left(Z \leq 1.06\right) - 1\\
&= 2\left(0.85543\right) - 1\\
&= 0.71086
\end{aligned}$$

b) Since n is unknown, it is difficult to apply a continuity correction. If n is large the continuity correction changes the answer by very little so we do not apply a continuity correction in this case. By the Normal approximation to the Binomial we have that X has approximately a Normal distribution with mean $\mu = np = 0.2n$ and variance $\sigma^2 = np(1-p) = 0.16n$. We want to find n such that

$$P\left(\left|\frac{X}{n} - 0.2\right| \leq 0.02\right) \geq 0.95$$

Now

$$\begin{aligned} & P\left(\left|\frac{X}{n} - 0.2\right| \leq 0.02\right) \\ & = P\left(|X - 0.2n| \leq 0.02n\right) \\ & = P\left(\frac{|X - 0.2n|}{\sqrt{0.16n}} \leq \frac{0.02n}{\sqrt{0.16n}}\right) \\ & \approx P\left(|Z| \leq 0.05\sqrt{n}\right) \quad \text{where } Z \sim N\left(0,1\right) \end{aligned}$$

Since $P\left(|Z| \leq 1.96\right) = 0.95$ we want n such that

$$0.05\sqrt{n} \geq 1.96 \text{ or } n \geq \left(\frac{1.96}{0.05}\right)^2 = 1536.64$$

In other words, we need to survey 1537 people to be at least 95% sure that $\frac{X}{n}$ lies within 0.02 of $p = 0.2$. Note that $n = 1537$ is large so using a continuity correction would not affect the final answer.

c) By the Normal approximation to the Binomial, $X \sim N\left(np, np(1-p)\right)$ approximately. We want to find n such that

$$P\left(\left|\frac{X}{n} - p\right| \leq 0.02\right) \geq 0.95$$

Now

$$\begin{aligned} & P\left(\left|\frac{X}{n} - p\right| \leq 0.02\right) \\ & = P\left(|X - np| \leq 0.02n\right) \\ & = P\left(\frac{|X - np|}{\sqrt{np\left(1-p\right)}} \leq \frac{0.02n}{\sqrt{np\left(1-p\right)}}\right) \\ & \approx P\left(|Z| \leq \frac{0.02\sqrt{n}}{\sqrt{p\left(1-p\right)}}\right) \quad \text{where } Z \sim N\left(0,1\right) \end{aligned}$$

Since $P\left(|Z| \leq 1.96\right) = 0.95$ we want n such 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 don't know p. The way out of this dilemma is to find the maximum value for

$$\left(\frac{1.96}{0.02}\right)^2 p(1-p)$$

If we choose n this large, then we can be sure of having the required precision in our estimate, $\frac{X}{n}$, for any p. It's easy to see that $p(1-p)$ is a maximum when $p = 0.5$. Therefore we take

$$n \geq \left(\frac{1.96}{0.02}\right)^2 \left(\frac{1}{2}\right)\left(1-\frac{1}{2}\right) \geq 2401$$

that is, if we survey $n = 2401$ people we can be 95% sure that $\frac{X}{n}$ lies within 0.02 of p, regardless of the value of p.

Remark: This method is used when poll results are reported in the media: you often see or hear that "this poll is accurate to with 3 percent 19 times out of 20". This is saying that n was big enough so that $P(p - 0.03 \leq X/n \leq p + 0.03)$ was 95%. (This requires n of about 1067.)

Problems

10.1.1 Tomato seeds germinate (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 soil germinate. Evaluate this using a suitable approximation.

10.1.2 A metal parts manufacturer inspects each part produced. 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$.

10.2 Moment Generating Functions

Univariate Discrete Distributions

We have seen two functions which characterize a distribution of a random variable, the probability function/probability density function and the cumulative distribution function. If we are given the probability function/probability density function of a random variable X or the cumulative distribution function of the random variable X then we can determine everything there is to know about the distribution of X. There is a third type of function, the *moment generating function*, which also uniquely determines a distribution. The moment generating function is closely related to other transforms used in mathematics, the Laplace and Fourier transforms.

We first consider moment generating functions for discrete random variables.

Definition 42 *Consider a discrete random variable X with probability function $f(x)$. The moment generating function (m.g.f.) of X is defined as*

$$M(t) = E(e^{tX}) = \sum_{\text{all } x} e^{tx} f(x)$$

We will assume that the moment generating function is defined and finite for values of t in an interval around 0 (that is, for some $a > 0$, $\sum\limits_{x} e^{tx} f(x) < \infty$ for all $t \in [-a, a]$).

The *moments* of a random variable X are the expectations of the functions X^k for $k = 1, 2, \ldots$. The expected value $E(X^k)$ is called the k'th moment of X. The mean $\mu = E(X)$ is therefore the first moment, $E(X^2)$ is the second moment and so on. It is often easy to find the moments of a probability distribution mathematically by using the moment generating function. This often gives easier derivations of means and variances than the direct summation methods in Chapter 7. The following theorem gives a useful property of moment generating functions.

Theorem 43 *Suppose the random variable X has moment generating function $M(t)$ defined for all $t \in [-a, a]$ for some $a > 0$. Then*

$$E(X^k) = M^{(k)}(0) \quad \text{for } k = 1, 2, \ldots$$

where

$$M^{(k)}(0) = \frac{d^k}{dt^k} M\left(t\right)|_{t=0} \quad \text{for } k = 1, 2, \ldots$$

Proof:
$M(t) = \sum\limits_{\text{all } x} e^{tx} f(x)$ and if the sum converges for all $t \in [-a, a]$, then

$$\begin{aligned} M^{(k)}(t) &= \frac{d^k}{dt^k} \sum_{\text{all } x} e^{tx} f(x) \\ &= \sum_{\text{all } x} \frac{d}{dt^k} (e^{tx}) f(x) \\ &= \sum_{\text{all } x} x^k e^{tx} f(x) \end{aligned}$$

Therefore $M^{(k)}(0) = \sum\limits_{\text{all } x} x^k f(x) = E(X^k)$, as stated.

Theorem 43 gives us another way to find the moments for a distribution.

Example: Suppose X has a $Binomial(n, p)$ distribution. Then its moment generating function is

$$\begin{aligned} M(t) &= \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x (1-p)^{n-x} \\ &= (pe^t + 1 - p)^n \quad \text{by the Binomial Theorem for all } t \in \Re \end{aligned}$$

Therefore

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

and so

$$\begin{aligned} E(X) &= M'(0) = np, \\ E(X^2) &= M\text{"}(0) = np + n(n-1)p^2 \\ Var(X) &= E(X^2) - E(X)^2 = np(1-p) \end{aligned}$$

Exercise: Moment generating function for a Poisson random variable
Use the Exponential series to show that if $X \sim Poisson\,(\mu)$ then the moment generating function is

$$M(t) = \exp\left(-\mu + \mu e^t\right) \quad \text{for all } t \in \Re$$

Use this to show that $E(X) = \mu$ and $Var(X) = \mu$.

The moment generating function uniquely identifies a distribution in the sense that if two random variables have the same moment generating function, they have the same distribution (so the same probability function, cumulative distribution function, moments, etc.). Of course the moment generating functions must match for all values of t, in other words they agree as *functions*, not just at a few points. For example if we can show somehow that the moment generating function of a random variable X is

$$M(t) = e^{2(e^t-1)} \quad \text{for all } t \in \Re$$

then we know from the previous example, the random variable must have a $Poisson(2)$ distribution. This means that if we are able to determine the moment generating function for a given random variable then the moment generating function can be used to identify its distribution. This gives us another technique for finding the distribution of a random variable.

Theorem 44 ***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 then X and Y have the same distribution.*

Moment generating functions can also be used to determine that a sequence of distributions gets closer and closer to some limiting distribution. To show this (albeit a bit loosely), suppose that a sequence of probability functions $f_n(x)$ have corresponding moment generating functions

$$M_n(t) = \sum_{\text{all } x} e^{tx} f_n(x)$$

Suppose moreover that the probability functions $f_n(x)$ converge to another probability function $f(x)$ pointwise in x as $n \to \infty$. This is what we mean by convergence of discrete distributions. Then since

$$f_n(x) \to f(x) \text{ as } n \to \infty \text{ for each } x, \tag{10.1}$$

$$\sum_{\text{all } x} e^{tx} f_n(x) \to \sum_{\text{all } x} e^{tx} f(x) \text{ as } n \to \infty \text{ for each } t \tag{10.2}$$

which says that $M_n(t)$ converges to $M(t)$ the moment generating function of the limiting distribution. It shouldn't be too surprising that a very useful converse to this result also holds. (This is strictly an aside and may be of interest only to those with a thing for infinite series, but is it always true that because the individual terms in a series converge as in (10.1) does this guarantee that the sum of the series also converges (10.2)?)

Suppose conversely that X_n has moment generating function $M_n(t)$ and $M_n(t) \to M(t)$ for each t such that $M(t) < \infty$. For example we saw in Chapter 5 that a $Binomial(n,p)$ distribution with very large n and very small p is close to a Poisson distribution with parameter $\mu = np$. Consider the moment generating function of such a Binomial random variable

$$\begin{aligned} M(t) &= \left(pe^t + 1 - p\right)^n \\ &= \left[1 + p\left(e^t - 1\right)\right]^n \\ &= \left[1 + \frac{\mu}{n}\left(e^t - 1\right)\right]^n \end{aligned}$$

Now take the limit of this expression as $n \to \infty$. Since in general

$$\lim_{n\to\infty}\left(1 + \frac{c}{n}\right)^n \to e^c$$

we have

$$\lim_{n\to\infty}\left[1 + \frac{\mu}{n}\left(e^t - 1\right)\right]^n = e^{\mu(e^t - 1)}$$

and this is the moment generating function of a Poisson distribution with parameter μ. This shows a little more formally than we did earlier that the Binomial distribution with small p approaches the Poisson distribution with mean $\mu = np$ as $n \to \infty$.

Moment Generating Function of a Continuous Random Variable

For continuous random variables the moment generating function is defined in a manner analogous to discrete random variables.

Definition 45 *Consider a continuous random variable X with probability density function $f(x)$. The moment generating function (m.g.f.) of X is defined as*

$$M(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f(x)dx$$

We will assume that the moment generating function is defined and finite for values of t in an interval around 0 (that is, for some $a > 0$, $\int_{-\infty}^{\infty} e^{tx} f(x)dx < \infty$ for all $t \in [-a, a]$).

Example: Moment generating function of a $N(\mu, \sigma^2)$ random variable If X has the $N(\mu, \sigma^2)$ distribution, then

$$\begin{aligned}
M(t) = E(e^{tX}) &= \int_{-\infty}^{\infty} e^{tx} f(x) dx \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2\sigma^2}\left(x^2 - 2\mu x - 2xt\sigma^2 + \mu^2\right)\right] dx \\
&= e^{\mu t + \sigma^2 t^2/2} \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}\left[x^2 - 2\left(\mu + t\sigma^2\right)x + \left(\mu + t\sigma^2\right)^2\right]\right\} dx \\
&= e^{\mu t + \sigma^2 t^2/2} \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2\sigma^2}\left[x - \left(\mu + t\sigma^2\right)\right]^2\right] dx \\
&= e^{\mu t + \sigma^2 t^2/2} \quad \text{for all } t \in \Re
\end{aligned}$$

where the last step follows since

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2\sigma^2}\left[x - \left(\mu + t\sigma^2\right)\right]^2\right] dx$$

is just the integral of a $N(\mu + t\sigma^2, \sigma^2)$ probability density function and is therefore equal to one. This confirms the values we obtained before for the mean and the variance of the Normal distribution

$$\begin{aligned}
M'(0) &= e^{\mu t + \sigma^2 t^2/2}\left(\mu + t\sigma^2\right)|_{t=0} = \mu \\
M''(0) &= e^{\mu t + \sigma^2 t^2/2}\left[\sigma^2 + \left(\mu + t\sigma^2\right)^2\right]|_{t=0} = \sigma^2 + \mu^2 = E(X^2)
\end{aligned}$$

from which we obtain

$$Var(X) = \sigma^2$$

Exercise: Suppose $X \sim Exponential\,(\theta)$. Use the moment generating function of X given by

$$M\,(t) = \frac{1}{1 - \theta t} \quad \text{for } t < \frac{1}{\theta}$$

(see Problem 15) to show $E\,(X) = \theta$ and $Var\,(X) = \theta^2$.

Note: The moment generating function uniquely identifies continuous distributions as well, so that Theorem 44 also holds for continuous random variables. We use this property in the proof below.

Proof of Theorem: 40 The moment generating function of a Poisson(μ) random variable X is given by

$$M_X(t) = e^{-\mu+\mu e^t}$$

Then the standardized random variable is

$$Z_\mu = \frac{X-\mu}{\sqrt{\mu}}$$

and this has moment generating function

$$\begin{aligned} M_Z(t) = E(e^{tZ}) &= E\left[e^{t(X-\mu)/\sqrt{\mu}}\right] \\ &= e^{-t\sqrt{\mu}}E(e^{tX/\sqrt{\mu}}) \\ &= e^{-t\sqrt{\mu}}M_X(t/\sqrt{\mu}) \end{aligned}$$

This is easier to work with if we take logarithms,

$$\begin{aligned} \ln\left[M_Z(t)\right] &= -t\sqrt{\mu} - \mu + \mu e^{t/\sqrt{\mu}} \\ &= \mu\left(e^{t/\sqrt{\mu}} - 1 - \frac{t}{\sqrt{\mu}}\right) \end{aligned}$$

Now as $\mu \to \infty$,

$$\frac{t}{\sqrt{\mu}} \to 0$$

and

$$e^{t/\sqrt{\mu}} = 1 + \frac{t}{\sqrt{\mu}} + \frac{1}{2}\frac{t^2}{\mu} + o\left(\mu^{-1}\right)$$

so

$$\begin{aligned} \ln\left[M_Z(t)\right] &= \mu\left(e^{t/\sqrt{\mu}} - 1 - \frac{t}{\sqrt{\mu}}\right) \\ &= \mu\left[\frac{t^2}{2\mu} + o\left(\mu^{-1}\right)\right] \to \frac{t^2}{2} \quad \text{as } \mu \to \infty \end{aligned}$$

where $o\left(\mu^{-1}\right)$ represents terms that go to zero faster than μ^{-1} as $\mu \to \infty$. Therefore the moment generating function of the standardized Poisson random variable Z_μ approaches $e^{t^2/2}$, which is the moment generating function of the standard Normal and this implies that the Poisson distribution approaches the Normal as $\mu \to \infty$.

We can similarly use moment generating functions to prove convergence of the distribution of a Binomial random variable to the Normal distribution (see Problem 10.13).

10.3 Multivariate Moment Generating Functions

Suppose we have two possibly dependent random variables (X, Y) and we wish to characterize their joint distribution using a moment generating function. Just as the probability function and the cumulative distribution function are, in tis case, functions of two arguments, so is the moment generating function.

Definition 46 *The joint moment generating function of* (X, Y) *is*

$$M(s,t) = E\left(e^{sX+tY}\right)$$

Recall that if X, Y are independent random variables and $g_1(X)$ and $g_2(Y)$ are any two functions, then

$$E[g_1(X)g_2(Y)] = E[g_1(X)]E[g_2(Y)] \tag{10.3}$$

and so with $g_1(X) = e^{sX}$ and $g_2(Y) = e^{tY}$ we obtain, for independent random variables X, Y

$$M(s,t) = M_X(s)M_Y(t)$$

the product of the moment generating functions of X and Y respectively.

There is another labour-saving property of moment generating functions for independent random variables. Suppose X, Y are independent discrete random variables with moment generating functions $M_X(t)$ and $M_Y(t)$ respectively. Suppose you wish the moment generating function of the sum $Z = X + Y$. One could attack this problem by first determining the probability function of Z,

$$\begin{aligned} f_Z(z) = P(Z = z) &= \sum_{\text{all } x} P(X = x, Y = z - x) \\ &= \sum_{\text{all } x} P(X = x)P(Y = z - x) \\ &= \sum_{\text{all } x} f_X(x) f_Y(z - x) \end{aligned}$$

and then calculating

$$E(e^{tZ}) = \sum_{\text{all } z} e^{tZ} f_Z(z)$$

Evidently lots of work! On the other hand, using (10.3) with

$$g_1(X) = e^{tX} \text{ and } g_2(Y) = e^{tY}$$

gives

$$\begin{aligned} M_Z(t) &= E\left[e^{t(X+Y)}\right] \\ &= E\left(e^{tX}\right)E\left(e^{tY}\right) \\ &= M_X(t)M_Y(t) \end{aligned}$$

Theorem 47 *The moment generating function of the sum of independent random variables is the product of the individual moment generating functions.*

Example: If X and Y are independent Bernoulli random variables with probability function

x	0	1
$f(x)$	$1-p$	p

then both have moment generating function

$$M_X(t) = M_Y(t) = \left(1 - p + pe^t\right)$$

and so the moment generating function of the sum Z is $M_X(t)M_Y(t) = (1-p+pe^t)^2$. Similarly if we add another independent Bernoulli the moment generating function is $\left(1-p+pe^t\right)^3$ and in general the sum of n independent Bernoulli random variables is $\left(1-p+pe^t\right)^n$ which is the moment generating function of a Binomial(n, p) distribution. This confirms that the sum of independent Bernoulli random variables has a $Binomial(n, p)$ distribution.

We now prove that a linear combination of independent Normal random variables has a Normal distribution.

Theorem 48 *If $X_i \sim N\left(\mu_i, \sigma_i^2\right)$, $i = 1, 2, \ldots, n$ independently and $a_1, a_2, \ldots, a_n$ are constants, then $\sum\limits_{i=1}^{n} a_i X_i \sim N\left(\sum\limits_{i=1}^{n} a_i\mu_i, \sum\limits_{i=1}^{n} a_i^2\sigma_i^2\right)$.*

Proof: Recall that the moment generating function for a $N\left(\mu, \sigma^2\right)$ random variable is $M(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$ for all $t \in \Re$. Since $X_i \sim N\left(\mu_i, \sigma_i^2\right)$ the moment generating function of $a_i X_i$ is

$$E\left(e^{ta_iX_i}\right) = E\left(e^{(ta_i)X_i}\right) = \exp\left[\mu_i\left(a_i t\right) + \frac{1}{2}\sigma_i^2\left(a_i t\right)^2\right], \text{ for all } t \in \Re$$

Since $X_i, i = 1, 2, \ldots, n$ are independent random variables then by Theorem 47 the moment generating of $Y = \sum\limits_{i=1}^{n} a_i X_i$ is the product of the individual moment generating functions so

$$E\left(e^{tY}\right) = \prod_{i=1}^{n} \exp\left[\mu_i\left(a_i t\right) + \frac{1}{2}\sigma_i^2\left(a_i t\right)^2\right] = \exp\left[\left(\sum_{i=1}^{n} a_i\mu_i\right)t + \frac{1}{2}\left(\sum_{i=1}^{n} a_i^2\sigma_i^2\right)t^2\right]$$

which we recognize as the moment generating function of a $N\left(\sum\limits_{i=1}^{n} a_i\mu_i, \sum\limits_{i=1}^{n} a_i^2\sigma_i^2\right)$ random variable. Therefore by the Uniqueness Theorem $\sum\limits_{i=1}^{n} a_i X_i \sim N\left(\sum\limits_{i=1}^{n} a_i\mu_i, \sum\limits_{i=1}^{n} a_i^2\sigma_i^2\right)$.

Proof of the Central Limit Theorem

We wish to now use our knowledge of moment generating functions to prove the Central Limit Theorem. We prove that if X_i are independent identically distributed random variables with $E(X_i) = \mu$, $var(X_i) = \sigma^2$, then

$$S_n^* = \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu)$$

converges in distribution to $N(0,1)$ by showing that the corresponding moment generating functions converge (assuming that they are finite).

To do this we note that the random variable $Y_i = (X_i - \mu)/\sigma$ has $E(Y_i) = 0$ and $Var(Y_i) = 1$. By a Taylor series expansion of e^{Yt} with remainder term $r(x)$, we have

$$\begin{aligned} E\left(e^{tY}\right) &= E\left[1 + tY + \frac{t^2}{2}Y^2 + r(tY)\right] \quad \text{where } \frac{r(x)}{x^2} \to 0 \text{ as } x \to 0 \\ &= 1 + tE(Y) + \frac{t^2}{2}E(Y^2) + o(t^2) \ \text{ as } t \to 0 \\ &= 1 + \frac{t^2}{2} + o(t^2) \ \text{ as } t \to 0 \end{aligned} \tag{10.4}$$

where $o(t^2)$ means terms which go to zero faster than t^2 or $\frac{o(t^2)}{t^2} \to 0$ as $t \to 0$.[41]

Proof of Theorem 39 - Central Limit Theorem:

The $X_i's$ are independent and identically distributed random. We assume that there common moment generating function $M_X(t)$ exists. This implies that the moment generating function of the random variable $Y_i = (X_i - \mu)/\sigma$ also exists and is given by

$$M(t) = E\left(e^{t(X_i - \mu)/\sigma}\right) = e^{-\mu t/\sigma} M_X\left(\frac{t}{\sigma}\right).$$

By (10.4),

$$M(t) = 1 + \frac{1}{2}t^2 + o(t^2) \ \text{ where } \ \frac{o(t^2)}{t^2} \to 0 \ \text{ as } \ t \to 0$$

Replacing t by $t/\sqrt{n}$ we have

$$M\left(\frac{t}{\sqrt{n}}\right) = 1 + \frac{1}{2}\left(\frac{t}{\sqrt{n}}\right)^2 + o\left(\frac{1}{n}\right) \ \text{ where } \ \frac{o\left(n^{-1}\right)}{n^{-1}} \to 0 \ \text{ as } \ n \to \infty \tag{10.5}$$

Since the moment generating function of the sum of independent random variables is the product of the individual moment generating functions, then the moment generating function of $\sum_{i=1}^{n} Y_i$ is equal to $[M(t)]^n$.

[41] If you were REALLY paying attention here, you might wonder about the logic of these steps. If $r(Yt)/t^2 \to 0$ for random variable Y, how do we know $E[r(Yt)]/t^2 \to 0$ as $t \to 0$? The proof of this is unfortunately beyond the scope of this course.

Let $M_n(t)$ be the moment generating function of $S_n^* = \frac{1}{\sqrt{n}} \sum\limits_{i=1}^{n} Y_i$. Then

$$M_n(t) = E\left[\exp\left(t \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i\right)\right] = \left[M\left(\frac{t}{\sqrt{n}}\right)\right]^n$$

and by (10.5)

$$M_n(t) = \left[1 + \frac{1}{2}\left(\frac{t}{\sqrt{n}}\right)^2 + o\left(\frac{1}{n}\right)\right]^n \quad \text{as } n \to \infty$$

Using $\ln(1+x) = x + o\left(x^2\right)$ as $x \to 0$ we have

$$\begin{aligned}
\ln\left[M_n(t)\right] &= \ln\left\{\left[1 + \frac{1}{2}\left(\frac{t}{\sqrt{n}}\right)^2 + o\left(\frac{1}{n}\right)\right]^n\right\} \\
&= n \ln\left[1 + \frac{1}{2}\left(\frac{t}{\sqrt{n}}\right)^2 + o\left(\frac{1}{n}\right)\right] \\
&= n\left[\frac{1}{2}\left(\frac{t}{\sqrt{n}}\right)^2 + o\left(n^{-1}\right) + o\left(n^{-2}\right)\right] \\
&= n\left[\frac{1}{2}\left(\frac{t}{\sqrt{n}}\right)^2 + o\left(n^{-1}\right)\right] \\
&= \frac{1}{2}t^2 + \frac{o\left(n^{-1}\right)}{n^{-1}} \quad \text{as } n \to \infty
\end{aligned}$$

This implies that

$$M_n(t) = \left[1 + \frac{1}{2}\left(\frac{t}{\sqrt{n}}\right)^2 + o\left(\frac{1}{n}\right)\right]^n \to e^{t^2/2} \quad \text{as } n \to \infty$$

Since $e^{t^2/2}$ is the moment generating function of a $N(0,1)$ random variable, then by the uniqueness of moment generating functions we have that S_n^* converges in distribution to the $N(0,1)$ distribution.

10.4 Chapter 10 Problems

1. When people are asked to make up a random number between 0 and 1, it has been found that the distribution of the numbers, X, has probability density function close to

$$f(x) = \begin{cases} 4x; & 0 < x \leq 1/2 \\ 4\left(1-x\right) & \frac{1}{2} < x < 1 \end{cases}$$

(rather than the $U\left(0,1\right)$ distribution which would be expected). See Chapter 8, Problem 2 for $E\left(X\right)$ and $Var\left(X\right)$.

   (a) Let X_i be the i'th "random number", $i = 1, 2, \ldots, 100$ and $S = \sum\limits_{i=1}^{100} X_i$. Approximate $P\left(49.0 \leq S \leq 50.5\right)$.

   (b) Approximate $P\left(49.0 \leq S \leq 50.5\right)$ if $X_i \sim U\left(0,1\right), i = 1, 2, \ldots, 100$.

2. For Chapter 9, Problem 37, approximate the probability of passing the exam, both with and without guessing if (a) each $p_i = 0.45$; (b) each $p_i = 0.55$. What is the best strategy for passing the course if (a) $p_i = 0.45$ (b) $p_i = 0.55$?

3. In a survey of n voters from a given riding in Canada, the proportion x/n who say they would vote Conservative is used to estimate p, the probability a voter votes Conservative (x is the number of Conservative supporters in the survey.) 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? **Hint**: See Example 10.1.

4. When blood samples are tested for the presence of a disease, samples from 20 people are pooled and analysed together. If the analysis is negative, none of the 20 people is infected. If the pooled sample is positive, at least one of the 20 people is infected so they must each be tested separately; that is, a total of 21 tests is required. The probability a person has the disease is 0.02.

   (a) Find the mean and variance of the number of tests required for each group of 20.

   (b) For 2000 people, tested in groups of 20, find the mean and variance of the total number of tests. What assumption(s) has been made about the pooled samples?

   (c) Find the approximate probability that more than 800 tests are required for the 2000 people.

5. Suppose 80% of people who buy a new car say they are satisfied with the car when surveyed one year after purchase. Let X be the number of people in a group of 60 randomly chosen new car buyers who report satisfaction with their car. Let Y be the number of satisfied owners in a second (independent) survey of 62 randomly chosen new car buyers. Using a suitable approximation, find $P\left(\left|X - Y\right| \geq 3\right)$. A continuity correction is expected.

6. 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 needed.

   (b) How large a random sample would be required so that, with probability 0.95, the proportion of unemployed persons in the sample is between 6.9% and 7.1%? **Hint:** See Example 10.1.

7. Requests to a web server are assumed to follow a Poisson process. On average there are two requests per second.

   (a) Give an expression for the probability that between 110 and 135 (inclusive) requests are received in a one minute interval. Approximate this probability using a suitable approximation.

   (b) Suppose the web server crashes if more than 150 requests are received in a one minute interval. Give an expression for the probability that the web server crashes. Approximate this probability using a suitable approximation.

   (c) Suppose requests are observed beginning at midnight. Approximate the probability that the waiting time until the 600'th request is less than four and a half minutes.

8. The following calculations will be useful for STAT 221/231/241.

   (a) Suppose $X \sim Binomial\,(n,p)$ where n is large. Approximate

   $$P\left(\frac{X}{n} - 1.645\sqrt{\frac{p\,(1-p)}{n}} \leq p \leq \frac{X}{n} + 1.645\sqrt{\frac{p\,(1-p)}{n}}\right)$$

   You may ignore the continuity correction.

   (b) Suppose $X_i \sim Poisson\,(\mu)$, $i = 1, 2, \ldots, n$ where n is large. Let $\bar{X} = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$. Approximate

   $$P\left(\bar{X} - 1.96\sqrt{\frac{\mu}{n}} \leq \mu \leq \bar{X} + 1.96\sqrt{\frac{\mu}{n}}\right)$$

   You may ignore the continuity correction.

   (c) Suppose $X_i \sim Exponential\,(\theta)$, $i = 1, 2, \ldots, n$ where n is large. Let $\bar{X} = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$. Approximate

   $$P\left(\bar{X} - 2.576\sqrt{\frac{\theta^2}{n}} \leq \theta \leq \bar{X} + 2.576\sqrt{\frac{\theta^2}{n}}\right)$$

9. **Gambling:** Your chances of winning or losing money can be calculated in many games of chance as described here.

   Suppose each time you play a game (or place a bet) of $\$1$ that the probability you win (thus ending up with a profit of $\$1$) is 0.49 and the probability you lose (meaning your "profit" is $-\$1$) is 0.51.

   (a) Let S represent your profit after n independent plays or bets. Give a Normal approximation for the distribution of S.

   (b) If $n = 20$, determine $P(S \geq 0)$. (This is the probability you are "ahead" after 20 plays.) Also find $P(S \geq 0)$ if $n = 50$ and $n = 100$. What do you conclude?

   Note: For many casino games (roulette, blackjack) there are bets for which your probability of winning is only a little less than 0.5. However, as you play more and more times, the probability you lose (end up "behind") approaches 1.

   (c) Suppose now you are the casino owner. If all players combined place $n = 100,000$ one dollar bets in an evening, let Y be your profit. Find the value c with the property that $P(Y > c) = 0.99$. Explain in words what this means.

10. **Gambling: 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 three dice, in which each die has its six sides labelled with a crown, an anchor, and the four card suits club, diamond, heart and spade, respectively. You bet an amount (let's say $\$1$) on one of the six symbols: let's suppose you bet on "heart". The three dice are then rolled simultaneously and you win $\$t$ if t hearts turn up ($t = 0, 1, 2, 3$).

    (a) Let S represent your profits from playing the game n times. Give a Normal approximation for the distribution of S.

    (b) Find (approximately) the probability that $S > 0$ if

    (i) $n = 10$

    (ii) $n = 50$

11. Suppose $X \sim Geometric\,(p)$.

    (a) Find the moment generating function of X.

    (b) Use the moment generating function to determine $E\,(X)$ and $Var\,(X)$.

12. Suppose X has a discrete Uniform distribution on $\{a, a+1, \ldots, b\}$ with probability function

$$P(X = x) = \frac{1}{b - a + 1} \quad \text{for } x = a, a+1, \ldots, b$$

   (a) Find the moment generating function of X.

   (b) Use the moment generating function to determine $E(X)$ and $E\left(X^2\right)$.

13. Let X be a discrete random variable taking values in the set $\{0, 1, 2\}$ with $E(X) = 1$, and $E(X^2) = 1.5$.

   (a) Find $P(X = x)$, $x = 0, 1, 2$ and thus determine the moment generating function of X.

   (b) Determine $E(X^3)$ and $E(X^4)$.

   (c) Show that any probability distribution on $\{0, 1, 2\}$ is completely determined by its first two moments.

14. Find the distributions that correspond to the following moment-generating functions:

   (a)

$$M(t) = \frac{1}{3e^{-t} - 2} \quad \text{for } t < \ln(3/2)$$

   (b)

$$M(t) = e^{2(e^t - 1)} \quad \text{for } t \in \Re$$

15. Suppose $X \sim Exponential\,(\theta)$.

   (a) Find the moment generating function of X.

   (b) Use the moment generating function to determine $E(X)$ and $Var(X)$.

16. Let $X_1, X_2, \ldots, X_n$ be independent $N(1, 2)$ random variables. For each of the following random variables, find the moment generating function and use the Uniqueness Theorem to determine its distribution.

   (a) $Y = -3X_1 + 4$

   (b) $T = X_1 + X_2$

   (c) $S_n = X_1 + X_2 + \ldots + X_n$

   (d) $Z = n^{-1/2}(S_n - n)$

17. Suppose $X \sim Poisson(\lambda_1)$ and $Y \sim Poisson(\lambda_2)$ independently. Use moment generating functions to prove that $X + Y \sim Poisson\,(\lambda_1 + \lambda_2)$.

18. Suppose X is a continuous random variable with the probability density function

$$f(x) = \frac{1}{\theta^2} x e^{-x/\theta} \quad \text{for } x > 0 \text{ and } \theta > 0$$

   (a) Find the moment generating function of X.

   (b) Suppose $X \sim Exponential\,(\theta)$ and independently $Y \sim Exponential\,(\theta)$. Use moment generating functions to find the distribution of $S = X + Y$.

19. Recall that the moment generating function of the Binomial random variable X is

$$M(t) = (1 - p + pe^t)^n \quad \text{for } t \in \Re$$

Find the moment generating function of the standardized random variable

$$Z_n = \frac{X - np}{\sqrt{np(1-p)}}$$

Assume p is fixed and $n \to \infty$. Show that the moment generating function of Z_n is

$$E(e^{Z_n t}) \to e^{t^2/2} \text{ as } n \to \infty$$

This result implies that the standardized Binomial random variable Z_n approaches the standard Normal distribution.

20. **A model for stock returns:** A common model for stock returns is as follows: the number of trades N of stock XXX in a given day has a Poisson distribution with parameter λ. At each trade, say the i'th trade, the change in the price of the stock is X_i and has a Normal distribution with mean 0 and variance σ^2, say and these changes are independent of one another and independent of N. Find the moment generating function of the total change in stock price over the day. Is this a distribution that you recognise? What is its mean and variance?

11. SOLUTIONS TO SECTION PROBLEMS

3.1.1 (a) Each student can choose in 4 ways and they each get to choose.

(i) Suppose we list the points in S in a specific order, for example (choice of student A, choice of student B, choice of student C) so that the point $(1, 2, 3)$ indicates A chose section 1, B chose section 2, and C chose section 3. Then

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

Since each student can choose in 4 ways regardless of the choice of the other two students, by the multiplication rule S has $4 \times 4 \times 4 = 64$ points.

(ii) To satisfy the condition, the first student can choose in 4 ways and the others then only have 1 section they can go in. Therefore

$$P\left(\text{all in the same section}\right) = \frac{(4)\,(1)\,(1)}{64} = \frac{1}{16}$$

(iii) To satisfy the condition, the first to pick has 4 ways to choose, the next has 3 sections left, and the last has 2 sections left. Therefore

$$P\left(\text{all in different sections}\right) = \frac{(4)\,(3)\,(2)}{64} = \frac{3}{8}$$

(iv) To satisfy the condition, each has 3 ways to choose a section. Therefore

$$P\left(\text{nobody in section 1}\right) = \frac{(3)\,(3)\,(3)}{64} = \frac{27}{64}$$

(b) (i) The sample space S has n^s points, each a sequence like $(1, 2, 3, 2, \ldots)$ of length s.

(ii)

$$P(\text{all in the same section}) = \frac{(n)\,(1)\,(1)\cdots(1)}{n^s} = \frac{1}{n^{s-1}}$$

(iii)

$$P\,(\text{all in different sections}) = \frac{(n)\,(n-1)\,(n-2)\cdots(n-s+1)}{n^s} = \frac{n^{(s)}}{n^s}$$

(iv)

$$P\,(\text{nobody in section }1) = \frac{(n-1)\,(n-1)\,(n-1)\cdots(n-1)}{n^s} = \frac{(n-1)^s}{n^s}$$

3.1.2 (a) There are 26 ways to choose each of the 3 letters, so in all the letters can be chosen in $(26)\,(26)\,(26) = 26^3$ ways. If all letters are the same, there are 26 ways to choose the first letter, and only 1 way to choose the remaining 2 letters. Therefore

$$P(\text{all letters the same}) = \frac{(26)\,(1)\,(1)}{26^3} = \frac{1}{26^2}$$

(b) There are $10 \times 10 \times 10$ ways to choose the 3 digits. The number of ways to choose all even digits is $(4)\,(4)\,(4) = 4^3$. The number of ways to choose all odd digits is $(5)\,(5)\,(5) = 5^3$. Therefore

$$P(\text{all even or all odd}) = \frac{4^3 + 5^3}{10^3} = 0.189$$

3.1.3 (a) There are 35 symbols in all (26 letters + 9 numbers). The number of different 6-symbol passwords is $35^6 - 26^6$ (we need to subtract off the 26^6 arrangements in which only letters are used, since there must be at least one number). Similarly, we get the number of 7-symbol and 8-symbol passwords as $35^7 - 26^7$ and $35^8 - 26^8$. The total number of possible passwords is then

$$(35^6 - 26^6) + (35^7 - 26^7) + (35^8 - 26^8)$$

(b) Let N be the answer to part (a) (the total number of possible passwords). Assuming you never try the same password twice, the probability you find the correct password within the first $1,000$ tries is

$$P\,(\text{first password works}) + P\,(\text{second password works}) + \cdots + P\,(\text{1000'th password works})$$
$$= \frac{1}{N} + \frac{N-1}{N}\frac{1}{N-1} + \cdots + \frac{N-1}{N}\frac{N-2}{N-1}\cdots\frac{1}{N-999} = \frac{1000}{N}$$
$$= \frac{1000}{(35^6 - 26^6) + (35^7 - 26^7) + (35^8 - 26^8)}$$

3.4.1 There are $7!$ different arrangements of the 7 digits.

(a) We can arrange the 3 even digits in $3!$ ways. This block of even digits plus the 4 odd digits can be arranged in $5!$ orders. Therefore

$$P(\text{even together}) = \frac{3!5!}{7!} = \frac{1}{7}$$

(b) For even at ends, there are 3 ways to fill the first place, and 2 ways to fill the last place and $5!$ ways to arrange the middle 5 digits. For odd at ends there are 4 ways to fill the first place and 3 ways to fill the last place and $5!$ ways to arrange the middle 5 digits.

$$P(\text{even or odd at ends}) = \frac{(3)(2)(5!) + (4)(3)(5!)}{7!} = \frac{3}{7}$$

3.4.2 The total number of arrangements is $\frac{9!}{3!2!}$.

(a) E at each end gives $\frac{7!}{2!}$ arrangements of the middle 7 letters. L at each end gives $\frac{7!}{3!}$ arrangements of the middle 7 letters. Therefore

$$P(\text{word begins and ends with the same letter}) = \frac{\frac{7!}{2!} + \frac{7!}{3!}}{\frac{9!}{3!2!}} = \frac{1}{9}$$

(b) The X, C and N can be "stuck" together in $3!$ ways to form a single unit. We can then arrange the $3E$'s, $2L$'s, T, and (XCN) in $\frac{7!}{3!2!}$ways. Therefore

$$P(XCN \text{ together}) = \frac{\frac{7!}{3!2!} \times 3!}{\frac{9!}{3!2!}} = \frac{1}{12}$$

(c) There is only 1 way to arrange the letters in the order CEEELLNTX. Therefore

$$P(\text{alphabetical order}) = \frac{1}{\frac{9!}{3!2!}} = \frac{12}{9!}$$

3.5.1 (a) The 8 cars can be chosen in $\binom{160}{8}$ ways. We can choose x with faulty emission controls and $(8 - x)$ with good ones in $\binom{35}{x}\binom{125}{8-x}$ ways. Therefore

$$P(\text{at least 3 faulty emission controls}) = \frac{\sum_{x=3}^{8} \binom{35}{x}\binom{125}{8-x}}{\binom{160}{8}}$$

(b) This assumes all $\binom{160}{8}$ combinations are equally likely. This assumption probably doesn't hold since the inspector would tend to select older cars or those in bad shape.

3.5.2 (a) The first 6 finishes can be chosen in $\binom{15}{6}$ ways. Choose 4 from numbers $1, 2, \ldots, 9$ in $\binom{9}{4}$ ways and 2 from numbers $10, 11, \ldots, 15$ in $\binom{6}{2}$ ways. Therefore

$$P(4 \text{ single digits in top } 6) = \frac{\binom{9}{4}\binom{6}{2}}{\binom{15}{6}} = \frac{54}{143}$$

(b) We need 2 single digits and 2 double digit numbers in the first four digits and then a single digit. This occurs in $\binom{9}{2}\binom{6}{2}(7)$ ways. Therefore

$$P(\text{fifth digit is the third single digit}) = \frac{\binom{9}{2}\binom{6}{2}(7)}{\binom{15}{4}(11)} = \frac{36}{143}$$

Alternate Solution: There are $15^{(5)}$ ways to choose the first 5 in order. We can choose in order, 2 double digit and 3 single digit finishers in $6^{(2)}9^{(3)}$ ways, and then choose which 2 of the first 4 places have double digit numbers in $\binom{4}{2}$ ways. Therefore

$$P(\text{fifth digit is the third single digit}) = \frac{6^{(2)}9^{(3)}\binom{4}{2}}{15^{(5)}} = \frac{36}{143}$$

(c) Choose the number 13 in 1 way and the other 6 numbers in $\binom{12}{6}$ ways (from $1, 2, \ldots, 12$). Therefore

$$P(13 \text{ is highest}) = \frac{\binom{12}{6}}{\binom{15}{7}} = \frac{28}{195}$$

Alternate Solution: From the $\binom{13}{7}$ ways to choose 7 numbers from $1, 2, \ldots, 13$ subtract the $\binom{12}{7}$ which don't include 13 (that is, all 7 numbers chosen from $1, 2, \ldots, 12$). Therefore

$$P(13 \text{ is highest}) = \frac{\binom{13}{7} - \binom{12}{7}}{\binom{15}{7}} = \frac{28}{195}$$

3.6.1 By the Binomial Theorem

$$\sum_{x=0}^{n} \binom{n}{x} a^x = (1+a)^n \text{ for all } n \in \mathbb{Z}^+ \text{ and } a \in \Re$$

Differentiate with respect to a on both sides:

$$\sum_{x=0}^{n} x \binom{n}{x} a^{x-1} = n(1+a)^{n-1}$$

Multiply by a to get

$$\sum_{x=0}^{n} x \binom{n}{x} a^{x} = na(1+a)^{n-1}$$

Let $a = \left(\frac{p}{1-p}\right)$. Then

$$\begin{aligned} \sum_{x=0}^{n} x \binom{n}{x} \left(\frac{p}{1-p}\right)^x &= n\left(\frac{p}{p-1}\right)\left(1+\frac{p}{1-p}\right)^{n-1} \\ &= \frac{np}{(1-p)^n}(1)^{n-1} \end{aligned}$$

Multiply by $(1-p)^n$:

$$\begin{aligned} \sum_{x=0}^{n} x \binom{n}{x} \left(\frac{p}{1-p}\right)^x (1-p)^n = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} \\ = \frac{np}{(1-p)^n}(1-p)^n = np \end{aligned}$$

3.6.3

$$\begin{aligned} \sum_{x=0}^{\infty} \binom{-k}{x} p^k (p-1)^x &= p^k \sum_{x=0}^{\infty} \binom{-k}{x} (p-1)^x \quad \text{converges since } |p-1| < 1 \\ &= p^k\,(1+p-1)^{-k} \quad \text{by the Binomial Theorem} \\ &= 1 \end{aligned}$$

4.1.1 Let A be the event “black hair” and let B be the event “brown eyes”. We are given that $P(A \cap B) = 0.55$, $P(A) = 1 - 0.4 = 0.6$, and $P(B) = 0.75$. See Figure 11.1. We wish to determine $P\left(\overline{A} \cap \overline{B}\right)$. Since $0.05 + 0.55 + 0.20 + P\left(\overline{A} \cap \overline{B}\right) = 1$, therefore $P\left(\overline{A} \cap \overline{B}\right) = 0.2$.

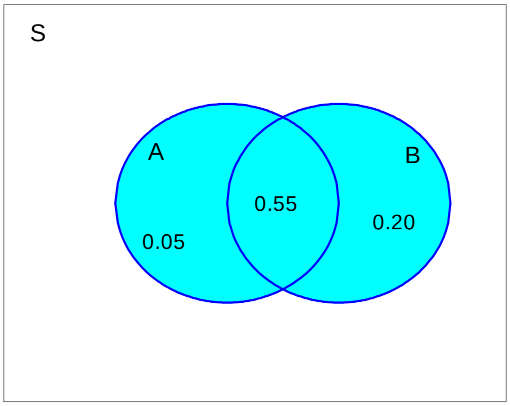


Figure 11.1: Venn diagram for Problem 4.1.1

4.1.2 Let A be the event “Mandarin or Cantonese speaking”, let B be the event “Spanish speaking”, and let C be the event “French speaking”. We are given

$$P(A) = 0.4, \quad P(B) = 0.25, \quad P(C) = 0.5$$
$$P(B \cap C) = 0.1, \quad P(A \cap C) = 0.12$$
$$P(A \cap B \cap C) = 0.02 \text{ and } P\left(\overline{A} \cap \overline{B} \cap \overline{C}\right) = 0.08$$

See Figure 11.2. We wish to determine $x = P\left(A \cap B \cap \overline{C}\right)$.
Since

$$\begin{aligned} 1 &= 0.08 + 0.3 + 0.08 + 0.1 + 0.02 + 0.28 - x + x + 0.15 - x \\ &= 1.01 - x \end{aligned}$$

therefore $x = 0.01$ and $P\left(A \cap B \cap \overline{C}\right) = 0.01$.

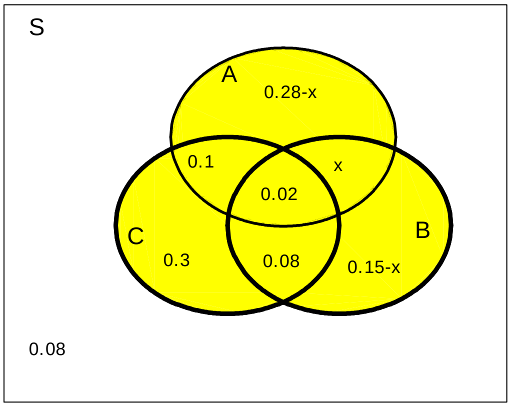


Figure 11.2: Venn diagram for Problem 4.1.2

4.2.1 (a)

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC) \\ &= 1 - 0.1 - [P(AC) + P(BC) - P(ABC)] \\ &= 0.9 - P(AC \cup BC) \end{aligned}$$

Therefore $P(A \cup B \cup C) = 0.9$ is the largest value, and this occurs when $P(AC \cup BC) = 0$.

(b) If each point in the sample space has strictly positive probability then if $P(AC \cup BC) = 0$, then $AC = \emptyset$ and $BC = \emptyset$ so that A and C are mutually exclusive and B and C are mutually exclusive. Otherwise we cannot make this determination. While A and C could be mutually exclusive, it can't be determined for sure.

4.2.2

$$\begin{aligned} P\left(A \cup B\right) &= P\left(A \text{ or } B \text{ occurs}\right) \\ &= 1 - P\left(A \text{ does not occur and } B \text{ does not occur}\right) \\ &= 1 - P\left(\overline{A} \cap \overline{B}\right) \end{aligned}$$

Alternatively, sinc $S = (A \cup B) \cup (\overline{A} \cap \overline{B})$ is a partition, and $P(S) = 1$, it follows that $P(A \cup B) + P(\overline{A} \cap \overline{B}) = 1$ and therefore $P\left(A \cup B\right) = 1 - P\left(\overline{A} \cap \overline{B}\right)$.

4.3.1 (a) Points giving a total of 9 are: $(3,6)$, $(4,5)$, $(5,4)$ and $(6,3)$. The probabilities are $(0.1)(0.3) = 0.03$ for $(3,6)$ and $(6,3)$, and $(0.2)(0.2) = 0.04$ for $(4,5)$ and $(5,4)$. Therefore

$$P\left[(3,6) \text{ or } (4,5) \text{ or } (5,4) \text{ or } (6,3)\right] = 0.03 + 0.04 + 0.04 + 0.03 = 0.14$$

(b) There are $\binom{4}{1}$ arrangements with 1 nine and 3 non-nines. Each arrangement has probability $(0.14)(0.86)^3$. Therefore

$$P(\text{total of nine on 1 of 4 repetitions}) = \binom{4}{1}(0.14)(0.86)^3 = 0.3562$$

4.3.2 Let C be the event a randomly chosen student is a coop student and let M be the event a randomly chosen student is a Math student. We are given $P(C) = 0.6$, $P(M) = 0.2$, $P(C \cap M) = (0.25)(0.6) = 0.15$. See Figure 11.3. From this information we can also determine that $P\left(\overline{C} \cap \overline{M}\right) = 1 - (0.45 + 0.15 + 0.05) = 0.35$.

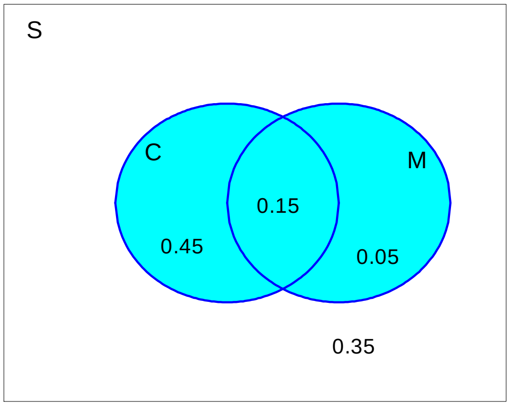


Figure 11.3: Venn diagram for Problem 4.3.2

Let A be the event “at least one coop student on the committee” and let B be the event “at least one Math student on the committee”. We wish to determine $P(A \cap B)$. This can be done most easily by noting that

$$P(A \cap B) = 1 - P\left(\overline{A \cap B}\right) = 1 - P\left(\overline{A} \cup \overline{B}\right) = 1 - \left[P\left(\overline{A}\right) + P\left(\overline{B}\right) - P\left(\overline{A} \cap \overline{B}\right)\right]$$

Now $\overline{A}$ is the event that there are no coop students on the committee so

$$P\left(\overline{A}\right) = (1 - 0.6)^{10} = (0.4)^{10}$$

Similarly $\overline{B}$ is the event that there are no Math students on the committee so

$$P\left(\overline{B}\right) = (1-0.2)^{10} = (0.8)^{10}$$

Finally $\overline{A} \cap \overline{B}$ is the event that there are no coop students and no Math students on the committee so $P\left(\overline{A} \cap \overline{B}\right) = (0.35)^{10}$. Therefore

$$P\left(A \cap B\right) = 1 - \left[(0.4)^{10} + (0.8)^{10} - (0.35)^{10}\right] = 0.8925$$

4.3.3 Since $B = (A \cap B) \cup \left(\bar{A} \cap B\right)$ and $P\left(B\right) = P\left(A \cap B\right) + P\left(\bar{A} \cap B\right)$ then

$$P(\overline{A} \cap B) = P(B) - P(A \cap B) \quad (1)$$

By De Morgan's Laws

$$P(\overline{A} \cap \overline{B}) = P(\overline{A \cup B}) \quad (2)$$

$$\begin{aligned} P(\overline{A} \cap \overline{B}) = P(\overline{A})P(\overline{B}) & \quad \text{since } \overline{A} \text{ and } \overline{B} \text{ are independent events} \\ & \Leftrightarrow P(\overline{A \cup B}) = P(\overline{A} \cap \overline{B}) = P(\overline{A})P(\overline{B}) \text{ by } (2) \\ & \Leftrightarrow 1 - P(A \cup B) = P(\overline{A})P(\overline{B}) \\ & \Leftrightarrow 1 - [P(A) + P(B) - P(A \cap B)] = P(\overline{A})\left[1 - P(B)\right] \\ & \Leftrightarrow [1 - P(A)] - [P(B) - P(A \cap B)] = P(\overline{A}) - P(\overline{A})P(B) \\ & \Leftrightarrow P(\overline{A}) - P(\overline{A} \cap B) = P(\overline{A}) - P(\overline{A})P(B) \text{ by } (1) \\ & \Leftrightarrow P(\overline{A})P(B) = P(\overline{A} \cap B) \end{aligned}$$

Therefore $\overline{A}$ and $\overline{B}$ are independent events if and only if $\overline{A}$ and B are independent events.

4.5.1 Let B be the event you go bus and let L be the event you are late.

$$P(B|L) = \frac{P(B \cap L)}{P(L)} = \frac{P(L|B)P(B)}{P(L|B)P(B) + P(L|\overline{B})P(\overline{B})} = \frac{(0.3)(0.2)}{(0.3)(0.2) + (0.7)(0.1)} = \frac{6}{13}$$

4.5.2 Let F be the event the coin is fair and let H be the event you get 5 heads in 6 tosses of the coin.

$$\begin{aligned} P(F|H) &= \frac{P(F \cap H)}{P(H)} = \frac{P(H|F)P(F)}{P(H|F)P(F) + P(H|\bar{F})P(\bar{F})} \\ &= \frac{(\frac{3}{4})\binom{6}{5}(\frac{1}{2})^6}{(\frac{3}{4})\binom{6}{5}(\frac{1}{2})^6 + (\frac{1}{4})\binom{6}{5}(0.8)^5(0.2)^1} = 0.4170 \end{aligned}$$

4.5.3 Let H be the event the car has defective headlights and let M be the event the car has a defective muffler

$$P(M|H) = \frac{P(M \cap H)}{P(H)} = \frac{P(M \cap H)}{P((M \cap H) \cup (\overline{M} \cap H))} = \frac{0.1}{0.1 + 0.15} = 0.4$$

5.1.1 We need $f(x) \geq 0$ and $\sum\limits_{x=0}^{2} f(x) = 1$

$$\begin{aligned} 9c^2 + 9c + c^2 &= 1 \\ 10c^2 + 9c - 1 &= 0 \\ (10c - 1)(c + 1) &= 0 \end{aligned}$$

Therefore $c = 1/10$ or -1. But if $c = -1$ we have $f(1) < 0$ which is impossible. Therefore $c = 1/10$.

5.1.2 We are arranging the letters YFOOO where Y = “You”, F = “Friend”, O = “Other”. There are $\frac{5!}{3!} = 20$ distinct arrangements.

$X = 0$: YFOOO,OYFOO,OOYFO,OOOYF, FYOOO,OFYOO,OOFYO,OOOFY
$X = 1$: YOFOO,OYOFO,OOYOF,FOYOO,OFOYO,OOFOY
$X = 2$: YOOFO,OYOOF,FOOYO,OFOOY
$X = 3$: YOOOF,FOOOY

x	0	1	2	3
$f(x)$	0.4	0.3	0.2	0.1
$F(x)$	0.4	0.7	0.9	1

5.3.1 (a) Using the Hypergeometric distribution,

$$g(d) = P(\text{none of 7 cans are tainted if there are } d \text{ tainted cans}) = \frac{\binom{d}{0}\binom{12-d}{7}}{\binom{12}{7}}$$

Therefore

d	0	1	2	3
$g(d)$	1	$\frac{5}{12}$	$\frac{5}{33}$	$\frac{5}{110}$

(b) While we could find no tainted tins if d is as big as 3, it is not likely to happen. This implies the box is not likely to have as many as 3 tainted tins.

5.3.2 We are sampling **without replacement**. Suppose that **we can distinguish** the r S's (for example $S_1, S_2, ..., S_r$) and the $(N-r)$ F's (for example $F_1, F_2, ..., F_{N-r}$). (Whether or not the S's and F's can be distinguished does not affect the probabilities.) There are a total of $N^{(n)}$ different arrangements of N symbols. How many of these possible arrangements have exactly x S's? First we may choose x places to place the S's which can be done in $\binom{n}{x}$ ways. Of course the F's will be placed in the remaining $(n-x)$ positions. Then arrange the x S's in these positions in $r^{(x)}$ ways. Finally arrange the $(n-x)$ F's in the remaining positions which can be done in $(N-r)^{(n-x)}$ ways. Therefore there are $\binom{n}{x}r^{(x)}(N-r)^{(n-x)}$ different arrangements of exactly x S's and $(n-x)$ F's. The probability of exactly x S's is therefore

$$P(X=x) = f(x) = \frac{\binom{n}{x}r^{(x)}(N-r)^{(n-x)}}{N^{(n)}}$$

with x ranging from $max(0, n-(N-r))$ to $min(n, r)$.

5.4.1 (a) Using Hypergeometric, with $N = 130$, $r = 26$, $n = 6$,

$$f(2) = \frac{\binom{26}{2}\binom{104}{4}}{\binom{130}{6}} = 0.2506$$

(b) Using the Binomial approximation to the Hypergeometric

$$f(2) \approx \binom{6}{2}\left(\frac{26}{130}\right)^2\left(\frac{104}{130}\right)^4 = 0.2458$$

5.4.2 (a) Let A be the event camera A is picked and let B be the event camera B is picked. Assume shots are independent with a constant failure probability.

$$\begin{aligned} P(\text{fail twice}) &= P(A)P(\text{fail twice}|A) + P(B)P(\text{fail twice}|B) \\ &= \left(\frac{1}{2}\right)\binom{10}{2}(0.1)^2(0.9)^8 + \left(\frac{1}{2}\right)\binom{10}{2}(0.05)^2(0.95)^8 = 0.1342 \end{aligned}$$

(b)

$$P(A|\text{fail twice}) = \frac{P(A \text{ and fail twice})}{P(\text{fail twice})} = \frac{\left(\frac{1}{2}\right)\binom{10}{2}(0.1)^2(0.9)^8}{0.1342} = 0.7219$$

5.5.1 We need $(x-25)$ "failures" before our 25th "success".

$$\begin{aligned} f(x) &= \binom{x-1}{x-25}(0.2)^{25}(0.8)^{x-25} \\ &= \binom{x-1}{24}(0.2)^{25}(0.8)^{x-25} \quad \text{for } x = 25, 26, 27, \ldots \end{aligned}$$

5.5.2 (a) In the first $(x+17)$ selections we need to get $x$defective (use Hypergeometric distribution) and then we need a good one on the $(x+18)$'th draw. Therefore

$$f(x) = \frac{\binom{200}{x}\binom{2300}{17}}{\binom{2500}{x+17}} \times \frac{2283}{2500-(x+17)} \quad \text{for } x = 0, 1, \ldots, 200$$

(b) Since 2500 is large and we're only choosing a few of them, we can approximate the Hypergeometric portion of $f(x)$ using Binomial

$$f(2) \approx \binom{19}{2}\left(\frac{200}{2500}\right)^2\left(1-\frac{200}{2500}\right)^{17}\left(\frac{2283}{2481}\right) = 0.2440$$

5.6.1 Using the Geometric distribution we have

$$P(x \text{ not leaky found before first leaky}) = (0.7)^x(0.3) = f(x)$$

$$\begin{aligned} P(X \geq n-1) = &\ f(n-1) + f(n) + f(n+1) + \cdots \\ &= (0.7)^{n-1}(0.3) + (0.7)^n(0.3) + (0.7)^{n+1}(0.3) + \cdots \\ &= \frac{(0.7)^{n-1}(0.3)}{1-0.7} = (0.7)^{n-1} = 0.05 \\ &(n-1)log\,(0.7) = log\,(0.05) \quad \text{so } n = 9.4 \end{aligned}$$

At least 9.4 cars means 10 or more cars must be checked. Therefore $n = 10$.

5.7.1 (a) Let X be the number who don't show. Then $X \sim Binomial(122, 0.03)$

$$\begin{aligned} P(\text{not enough seats}) &= P(X=0 \text{ or } X=1) \\ &= \binom{122}{0}(0.03)^0(0.97)^{122} + \binom{122}{1}(0.03)^1(0.97)^{121} \\ &= 0.1161 \end{aligned}$$

(To use a Poisson approximation we need p near 0. That is why we defined "success" as not showing up).

For Poisson, $\mu = np = (122)(0.03) = 3.66$

$$f(0) + f(1) = e^{-3.66} + 3.66e^{-3.66} = 0.1199$$

(b) Binomial requires all passengers to be independent as to showing up for the flight, and that each passenger has the same probability of showing up. Passengers are not likely independent since people from the same family or company are likely to all show up or all not show. Even strangers arriving on an earlier incoming flight would not miss their flight independently if the flight was delayed. Passengers may all have roughly the same probability of showing up, but even this is suspect. People travelling in different fare categories or in different classes (e.g. charter fares versus first class) may have different probabilities of showing up.

5.8.1 (a) Since $\lambda = 3$ and $t = 2.5$, $\mu = \lambda t = 7.5$ and

$$f(6) = \frac{7.5^6 e^{-7.5}}{6!} = 0.1367$$

(b)

$$\begin{aligned} &P(2 \text{ in 1st minute} \mid 6 \text{ in } 2.5 \text{ minutes}) \\ &= \frac{P(2 \text{ in 1st minute and } 6 \text{ in } 2.5 \text{ minutes})}{P(6 \text{ in } 2.5 \text{ minute})} \\ &= \frac{P(2 \text{ in 1st minute and } 4 \text{ in last } 1.5 \text{ minutes})}{P(6 \text{ in } 2.5 \text{ minute})} \\ &= \frac{\left(\frac{3^2 e^{-3}}{2!}\right)\left(\frac{4.5^4 e^{-4.5}}{4!}\right)}{\left(\frac{7.5^6 e^{-7.5}}{6!}\right)} \\ &= \binom{6}{2}\left(\frac{3}{7.5}\right)^2\left(\frac{4.5}{7.5}\right)^4 = 0.3110 \end{aligned}$$

Note this is a Binomial probability.

5.8.2 Assume that the conditions for a Poisson process are met, with lines as units of “time”:

(a) $\lambda = .02$ per line; $t = 1$ line; $\mu = \lambda t = 0.02$

$$f(0) = \frac{\mu^0 e^{-\mu}}{0!} = e^{-0.02} = 0.9802$$

(b) $\mu_1 = 80 \times 0.02 = 1.6$; $\mu_2 = 90 \times 0.02 = 1.8$

$$\left(\frac{\mu_1^2 e^{-\mu_1}}{2!}\right)\left(\frac{\mu_2^2 e^{-\mu_2}}{2!}\right) = 0.0692$$

5.9.1 Consider a 1 minute period with no occurrences as a “success”. Then X has a Geometric distribution. The probability of “success” is

$$P\,(\text{no occurrencs}) = \frac{\lambda^0 e^{-\lambda}}{0!} = e^{-\lambda}$$

Therefore

$$f(x) = (e^{-\lambda})(1 - e^{-\lambda})^{x-1} \quad \text{for } x = 1, 2, 3, \ldots$$

(There must be $(x - 1)$ failures before the first success.)

5.9.2 (a) $\mu = (3)(1.25) = 3.75$

$$f(0) = \frac{3.75^0 e^{-3.75}}{0!} = 0.0235$$

(b) $\left(1 - e^{-3.75}\right)^{14} e^{-3.75}$, using a Geometric distribution

(c) Use a Binomial distribution

$$f(x) = \binom{100}{x} \left(e^{-3.75}\right)^x \left(1 - e^{-3.75}\right)^{100-x}$$

Approximate this by Poisson (since n large, p small) with $\mu = np = 100e^{-3.75} \approx 2.35$.

$$f(x) \approx e^{-2.35} \frac{2.35^x}{x!} \quad \text{for } x = 0, 1, \ldots$$

Therefore

$$P(X \geq 4) = 1 - P(X \leq 3) = 1 - 0.789 = 0.211.$$

7.3.1 Let X = the organization's profit. The profit depends on the ticket number picked. Since a 3 digit number can have either all digits equal, two different digits or three different digits, there are 3 possible cases to consider.

Case 1: All digits are the same. There are 10 such tickets, 000,111, 222, $\ldots$, 999 so the probability of drawing such a ticket is $10/1000$. If one of these tickets is drawn the profit is $X = 1000 - 200 = 800$ since the organization takes in \$1000 and pays out \$200 for the one winning ticket.

Case 2: There are 2 digits the same, for example, 211, 121, 112, or 662, 626, 266, etc. There are $10\,(9) = 90$ ways to chose the 2 numbers and 3 ways to arrange them so there are a total of $90\,(3) = 270$ such tickets and so the probability of drawing such a ticker is $270/1000$. If one of these tickets is drawn the profit is $X = 1000 - 3\,(200) = 400$ since there are 3 winners.

Case 3: Since there are a total of 1000 tickets and only 3 types of tickets, we can find the total number of tickets with 3 different numbers by subtraction. There are $1000 - 10 - 270 = 720$ tickets with 3 different digits. Therefore the probability of drawing a ticker with 3 different digits is $720/1000$. Note that if ticket 123 is drawn then the tickets 123,132,213,231,312,321 all win. If one of these tickets is drawn the profit is $X = 1000 - 6\,(200) = -200$.

Therefore the expected profit is

$$E\,(X) = 800\left(\frac{10}{1000}\right) + 400\left(\frac{270}{1000}\right) + (-200)\left(\frac{720}{1000}\right) = -28$$

that is, on average the organization loses \$28.

7.4.1 Suppose n tickets are sold. Let the random variable X_n be the number of people who show up. Then X_n has a Binomial(n, p) distribution with $p = 0.97$. For the Binomial distribution, $E(X_n) = np$. The expected revenue as a function of $n > 120$ is

$$\begin{aligned} h(n) &= 100E(X_n) - 500E\left(X_n - 120\right)^+ \\ &= 100np - 500\sum_{x=121}^{n}(x-120)\binom{n}{x}p^x(1-p)^{n-x} \end{aligned}$$

where $(X_n - 120)^+ = \max(0, X_n - 120)$. If $n \leq 120$, then $h(n) = 100np$ and since this is an increasing function of n, for $n \leq 120$, we need only consider the case $n > 120$ in attempting to maximize the function $h(n)$. Consider the values of $h\left(n\right)$ for $n = 121, 122, 123, 124$ since the number of tickets sold must be a positive integer.

$$\begin{aligned} &h(121) = 100(121)\,(0.97) - 500(0.97)^{121} = 11,724.46 \\ &h(122) = 100(122)\,(0.97) - 1000(0.97)^{122} - (500)\,122(0.97)^{121}\left[1 - (0.97)\right] = 11,763.77 \\ &h(123) = 100\,(123)\,(0.97) - 500\sum_{x=121}^{123}(x-120)\binom{123}{x}(0.97)^x(0.03)^{123-x} = 11,721.13 \\ &h(124) = 100\,(124)\,(0.97) - 500\sum_{x=121}^{124}(x-120)\binom{124}{x}(0.97)^x(0.03)^{124-x} = 11579 \end{aligned}$$

It would appear that the function $h(n)$ for $n = 121, 122, \ldots$ has a maximum at $n = 122$ which would indicate that the optimal number of tickets to be sold is $n = 122$.

Can we prove that $n = 122$ does indeed correspond to a maximum? Note that $X_{n+1} = X_n + Y$ where X_n, Y are independent random variables and Y has a Bernoulli(p) distribution with mean $E\left(Y\right) = p$. Now

$$h\left(n+1\right) = 100E\left(X_n + Y\right) - 500E\left(X_n + Y - 120\right)^+$$

and

$$\begin{aligned} h(n+1) - h(n) &= 100E\left(X_n + Y\right) - 500E\left(X_n + Y - 120\right)^+ - 100E\left(X_n\right) + 500E\left(X_n - 120\right)^+ \\ &= 100E\left(Y\right) - 500\left[E\left(X_n + Y - 120\right)^+ - E\left(X_n - 120\right)^+\right] \\ &= 100p - 500\left[E\left(X_n + Y - 120\right)^+ - E\left(X_n - 120\right)^+\right] \end{aligned}$$

Let

$$g\left(x, y\right) = (x + y - 120)^+ - (x - 120)^+ = \begin{cases} 1 & x \geq 120, y = 1 \\ 0 & \text{otherwise} \end{cases}$$

then

$$\begin{aligned} h(n+1) - h(n) &= 100p - 500\left[E\left(X_n + Y - 120\right)^+ - E\left(X_n - 120\right)^+\right] \\ &= 100p - 500E\left[g\left(X_n, Y\right)\right] \end{aligned}$$

Since

$$E\left[g\left(X_n, Y\right)\right] = P\left(X_n \geq 120, Y = 1\right) = pP\left(X_n \geq 120\right)$$

we have

$$h(n+1) - h(n) = 100p - 500pP\left(X_n \geq 120\right)$$

Since $P(X_n \geq 120)$ is an increasing function of n, $h(n+1) - h(n)$ is a decreasing function of n. That is if $h(n_0+1) - h(n_0) < 0$ for some value of $n_0 > 120$ (in our case above $n_0 + 1 = 123$ or $n_0 = 122$) then $h(n+1) - h(n)$ is negative for all $n > n_0$ and $h(n) \leq h(n_0)$ for all $n > n_0$ which proves that the maximum value occurs at n_0.

7.4.2 (a) Let X be the number of words needing correction and let T be the time to type the passage. Then $X \sim Binomial(450, 0.04)$ and $T = 450 + 15X$.
X has mean $np = 18$ and variance $np(1-p) = 17.28$.
$E(T) = E(450 + 15X) = 450 + 15E(X) = 450 + (15)(18) = 720$
$Var(T) = Var(450 + 15X) = 15^2 Var(X) = 3888.$

(b) At 45 words per minute, each word takes $\frac{4}{3}$ seconds. $X \sim Binomial(450, 0.02)$ and $T = (450)\left(\frac{4}{3}\right) + 15X = 600 + 15X$
$E(X) = 450\,(0.02) = 9;\;\; E(T) = 600 + (15)(9) = 735$, so it takes longer on average.

8.1.1 (a) Since

$$\int\limits_{-1}^{1} kx^2 dx = k\frac{x^3}{3}|_{-1}^{1} = \frac{2k}{3} = 1$$

therefore $k = \frac{3}{2}$

(b)

$$F(x) = \begin{cases} 0 \;\text{ for } x \leq -1 \\ \int\limits_{-1}^{x} \frac{3}{2}x^2 dx = \frac{x^3}{2}|_{-1}^{x} = \frac{x^3}{2} + \frac{1}{2} \;\text{ for} -1 < x < 1 \\ 1 \;\text{ for } x \geq 1 \end{cases}$$

(c)

$$P(-0.1 < X < 0.2) = F(0.2) - F(-0.1) = 0.504 - 0.4995 = 0.0045$$

(d)

$$E(X) = \int_{-1}^{1} x\frac{3}{2}x^2 dx = \frac{3}{2}\int_{-1}^{1} x^3 dx = \frac{3}{8}x^4|_{-1}^{1} = 0$$
$$E(X^2) = \int_{-1}^{1} x^2\frac{3}{2}x^2 dx = \frac{3}{10}x^5|_{-1}^{1} = \frac{3}{5}$$
$$Var(X) = E(X^2) - [E(X)]^2 = \frac{3}{5}$$

(e) For $0 \le y < 1$

$$\begin{aligned} F_Y(y) &= P(Y \le y) = P(X^2 \le y) = P(-\sqrt{y} \le X \le \sqrt{y}) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}) \\ &= \left[\frac{(\sqrt{y})^3}{2} + \frac{1}{2}\right] - \left[\frac{(-\sqrt{y})^3}{2} + \frac{1}{2}\right] \\ &= y^{3/2} \end{aligned}$$

Since $\frac{d}{dy}y^{3/2} = \frac{3}{2}\sqrt{y}$, the probability density function of Y is

$$f(y) = \begin{cases} \frac{3}{2}\sqrt{y} & \text{for } 0 \le y < 1 \\ 0 & \text{otherwise} \end{cases}$$

8.1.2 (a) Since

$$\lim_{x\to\infty} F(x) = 1 = \lim_{x\to\infty}\frac{kx^n}{1+x^n} = \lim_{x\to\infty}\frac{k}{\frac{1}{x^n}+1} = k$$

therefore $k = 1$.

(b)

$$f(x) = \frac{d}{dx}F(x) = \frac{nx^{n-1}}{(1+x^n)^2} \quad \text{for } x > 0$$

and 0 otherwise.

(c) Let m be the median. Then

$$F(m) = 0.5 = \frac{m^n}{1+m^n}$$

Therefore $m^n = 1$ and so the median equals 1.

8.2.1

$$F(x) = \int_{-1}^{x}\frac{3}{2}u^2 du = \frac{x^3+1}{2} \quad \text{for } x > -1$$

If $y = F(x) = \frac{x^3+1}{2}$ is a random number between 0 and 1, then $x = (2y-1)^{1/3}$. For $y = 0.27125$ we get $x = (-0.4574)^{1/3} = -0.77054$.

8.3.1 Let the time to disruption be X. Then

$$P(X \leq 8) = F(8) = 1 - e^{-8/\theta} = 0.25$$

Therefore $e^{-8/\theta} = 0.75$ or $\theta = -8/\ln(0.75) = 27.81$ hours.

8.3.2 (a) The number of flaws in a circle of radius x has a Poisson$\left(\lambda\pi x^2\right)$ distribution. If X is the distance from an arbitrary starting point to the second closest flaw then

$$\begin{aligned} F(x) &= P(X \leq x) = 1 - P(X > x) \\ &= 1 - P(0 \text{ or } 1 \text{ flaws within a circle of radius } x) \\ &= 1 - \frac{\left(\lambda\pi x^2\right)^0 e^{-\lambda\pi x^2}}{0!} - \frac{\left(\lambda\pi x^2\right)^1 e^{-\lambda\pi x^2}}{1!} \\ &= 1 - e^{-\lambda\pi x^2}\left(1 + \lambda\pi x^2\right) \end{aligned}$$

Therefore

$$f(x) = \frac{d}{dx}F(x) = 2\lambda^2\pi^2 x^3 e^{-\lambda\pi x^2} \quad \text{for } x > 0$$

and 0 otherwise

(b)

$$\begin{aligned} \mu = E(X) &= \int_{-\infty}^{\infty} xf(x)\,dx = \int_0^{\infty} x\left(2\lambda^2\pi^2 x^3 e^{-\lambda\pi x^2}\right)dx \\ &= \int_0^{\infty} 2\lambda^2\pi^2 x^4 e^{-\lambda\pi x^2}dx \quad \text{let } y = \lambda\pi x^2 \text{ with } dx = \frac{dy}{2\sqrt{\lambda\pi y}} \\ &= \int_0^{\infty} 2y^2 e^{-y}\frac{dy}{2\sqrt{\lambda\pi y}} = \frac{1}{\sqrt{\lambda\pi}}\int_0^{\infty} y^{3/2}e^{-y}dy \\ &= \frac{1}{\sqrt{\lambda\pi}}\Gamma\left(\frac{5}{2}\right) = \frac{1}{\sqrt{\lambda\pi}}\left(\frac{3}{2}\right)\Gamma\left(\frac{3}{2}\right) \\ &= \frac{1}{\sqrt{\lambda\pi}}\left(\frac{3}{2}\right)\left(\frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right) = \frac{\left(\frac{3}{2}\right)\left(\frac{1}{2}\right)\sqrt{\pi}}{\sqrt{\lambda\pi}} = \frac{3}{4\sqrt{\lambda}} \end{aligned}$$

8.5.1

$$\begin{aligned}
P\left(|X-\mu|<\sigma\right) &= P\left(-\sigma<X-\mu<\sigma\right)=P\left(-1<Z<1\right)\\
&= F(1)-[1-F(1)]=0.8413-(1-0.8413)\\
&= 68.26\% \quad \text{(about 2/3)}\\
P\left(|X-\mu|<2\sigma\right) &= P\left(-2\sigma<X-\mu<2\sigma\right)=P\left(-2<Z<2\right)\\
&= F(2)-[1-F(2)]=0.9772-(1-0.9772)\\
&= 95.44\% \quad \text{(about 95\%)}
\end{aligned}$$

Similarly

$$P\left(|X-\mu|<3\sigma\right)=P\left(-3<Z<3\right)=99.73\% \quad \text{(over 99\%)}$$

9.1.1 (a) The marginal probability functions are:

x	0	1	2
$f_1(x)$	0.3	0.2	0.5

and

y	0	1	2
$f_2(y)$	0.3	0.4	0.3

Note that, for example, $f_1(1)f_2(1)=0.08\neq 0.05$. Since $f_1(x)f_2(y)\neq f(x,y)$ for all (x,y), therefore X and Y are not independent random variables.

(b)

$$f(y|X=0)=\frac{f(0,y)}{f_1(0)}=\frac{f(0,y)}{0.3}$$

y	0	1	2
$f(y\|X=0)$	0.3	0.5	0.2

(c)

d	-2	-1	0	1	2
$f(d)$	0.06	0.24	0.29	0.26	0.15

(e.g. $P(D=0)=f(0,0)+f(1,1)+f(2,2)$)

9.1.2

$$\begin{aligned}
f(x,y) &= f(x)f(y) = \binom{x+k-1}{x}\binom{y+\ell-1}{y}p^{k+\ell}(1-p)^{x+y} \\
f(t) &= \sum_{x=0}^{t} f(x, y=t-x) \\
&= \sum_{x=0}^{t}\binom{x+k-1}{x}\binom{t-x+\ell-1}{t-x}p^{k+\ell}(1-p)^{t} \\
&= \sum_{x=0}^{t}(-1)^x\binom{-k}{x}(-1)^{t-x}\binom{-\ell}{t-x}p^{k+\ell}(1-p)^{t} \\
&= (-1)^t p^{k+\ell}(1-p)^t\sum_{x=0}^{t}\binom{-k}{x}\binom{-\ell}{t-x} \\
&= (-1)^t p^{k+\ell}(1-p)^t\binom{-k-\ell}{t} \quad \text{using the Hypergeometric Identity} \\
&= \binom{t+k+\ell-1}{t}p^{k+\ell}(1-p)^t \quad \text{for } t=0,1,2,\ldots
\end{aligned}$$

using the given identity on $(-1)^t\binom{-k-\ell}{t}$. T has a Negative Binomial distribution.

9.2.1 (a) Use a Multinomial distribution.

$$P\left(3\ A\text{'}s,\ 11\ B\text{'}s, 7C\text{'}s \text{ and } 4D\text{'}s\right) = \frac{25!}{3!11!7!4!}(0.1)^3(0.4)^{11}(0.3)^7(0.2)^4$$

(b) Group C's and D's into a single category.

$$P\left(3\ A\text{'}s \text{ and } 11\ B\text{'}s\right) = \frac{25!}{3!11!11!}(0.1)^3(0.4)^{11}(0.5)^{11}$$

(c) Of the 21 non D's we need 3 A's, 11 B's and 7 C's. The (conditional) probabilities for the non-D's are: $1/8$ for A, $4/8$ for B, and $3/8$ for C.
(e.g. $P(A|\overline{D}) = P(A)/P(\overline{D}) = 0.1/0.8 = 1/8$)
Therefore

$$f(3\ A\text{'}s,\ 11\ B\text{'}s, 7C\text{'}s|4\ D\text{'}s) = \frac{21!}{3!11!7!}\left(\frac{1}{8}\right)^3\left(\frac{4}{8}\right)^{11}\left(\frac{3}{8}\right)^7$$

9.2.2 $\mu = (0.6)\,(12) = 7.2$

$$p_1 = P(\text{fewer than } 5 \text{ chips}) = \sum_{x=0}^{4}\frac{7.2^x e^{-7.2}}{x!}$$

$$p_2 = P(\text{more than } 9 \text{ chips}) = 1 - \sum_{x=0}^{9}\frac{7.2^x e^{-7.2}}{x!}$$

(a)

$$\binom{12}{3} p_1^3 (1-p_1)^9$$

(b)

$$\frac{12!}{3!7!2} p_1^3 p_2^7 (1-p_1-p_2)^2$$

(c) Given that 7 have > 9 chips, the remaining 5 are of 2 types - under 5 chips, or 5 to 9 chips

$$P(< 5 | \le 9 \text{ chips}) = \frac{P(< 5 \text{ and } \le 9)}{P(\le 9)} = \frac{p_1}{1-p_2}$$

Using a Binomial distribution

$$P(3 \text{ under } 5 | 7 \text{ over } 9) = \binom{5}{3} \left(\frac{p_1}{1-p_2}\right)^3 \left(1 - \frac{p_1}{1-p_2}\right)^2$$

9.4.1

x	0	1	2
$f_1(x)$	0.2	0.5	0.3

y	0	1
$f_2(y)$	0.3	0.7

$$\begin{aligned}
E(X) &= (0)(0.2) + (1)(0.5) + (2)(0.3) = 1.1 \\
E(Y) &= (0)(0.3) + (1)(0.7) = 0.7 \\
E(X^2) &= \left(0^2\right)(0.2) + \left(1^2\right)(0.5) + \left(2^2\right)(0.3) = 1.7 \\
E(Y^2) &= 0.7 \\
Var(X) &= 1.7 - 1.1^2 = 0.49 \\
Var(Y) &= 0.7 - (0.7)^2 = 0.21 \\
E(XY) &= (1)(1)(0.35) + (2)(1)(0.21) = 0.77 \\
Cov(X,Y) &= 0.77 - (1.1)(0.7) = 0 \\
\text{Therefore } \rho &= \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = 0
\end{aligned}$$

While $\rho = 0$ indicates X and Y may be independent (and indeed are in this case), it does not prove that they are independent. It only indicates that there is no linear relationship between X and Y.

9.4.2

(a)

x	2	4	6
$f_1(x)$	$3/8$	$3/8$	$1/4$

y	-1	1
$f_2(y)$	$\frac{3}{8}+p$	$\frac{5}{8}-p$

$$\begin{aligned}
E(X) &= (2)\left(\frac{3}{8}\right)+(4)\left(\frac{3}{8}\right)+(6)\left(\frac{1}{4}\right)=15/4\\
E(Y) &= -\frac{3}{8}-p+\frac{5}{8}-p=\frac{1}{4}-2p\\
E(XY) &= (-2)\left(\frac{1}{8}\right)+(-4)\left(\frac{1}{4}\right)+\cdots+(6)\left(\frac{1}{4}-p\right)=\frac{5}{4}-12p
\end{aligned}$$

Since $Cov(X,Y)=0=E(XY)-E(X)E(Y)$ therefore

$$\frac{5}{4}-12p=\frac{15}{16}-\frac{15}{2}p$$

or $p=5/72$.

(b) If X and Y are independent then $Cov(X,Y)=0$, and so p must be $5/72$. But if $p=5/72$ then

$$f_1(2)f_2(-1)=\left(\frac{3}{8}\right)\left(\frac{4}{9}\right)=\frac{1}{6}\neq f(2,-1)$$

Therefore X and Y cannot be independent for any value of p.

9.5.1

x	0	1	2
$f_1(x)$	0.5	0.3	0.2

$$\begin{aligned}
E(X) &= (0)(0.5)+(1)(0.3)+(2)(0.2)=0.7\\
E(X^2) &= \left(0^2\right)(0.5)+\left(1^2\right)(0.3)+\left(2^2\right)(0.2)=1.1\\
Var(X) &= E(X^2)-[E(X)]^2=0.61
\end{aligned}$$

$$\begin{aligned}
E(XY) &= \textstyle\sum_{\text{all } x,y} xyf(x,y) \quad \text{and this has only two non-zero terms}\\
&= (1)(1)(0.2)+(2)(1)(0.15)=0.5
\end{aligned}$$

$$\begin{aligned}
Cov(X,Y) &= E(XY)-E(X)E(Y)=0.01\\
Var(3X-2Y) &= 9Var(X)+(-2)^2Var(Y)+2(3)(-2)Cov(X,Y)\\
&= 9(0.61)+4(0.21)-12(0.01)=6.21
\end{aligned}$$

9.5.2

$$\begin{aligned}
\rho &= \frac{Cov(X,Y)}{\sigma_x\sigma_y}=0.5\\
Cov(X,Y) &= 0.5\sqrt{1.69\times 4}=1.3\\
Var(U) &= Var(2X-Y)=4\sigma_X^2+\sigma_Y^2-4Cov(X,Y)=5.56
\end{aligned}$$

Therefore the standard deviation of $U = 2.36$

9.5.3

$$\begin{aligned} Cov\left(X_{i-1}, X_i\right) &= Cov\left(Y_{i-2} + Y_{i-1}, Y_{i-1} + Y_i\right) \\ &= Cov\left(Y_{i-2}, Y_{i-1}\right) + Cov\left(Y_{i-2}, Y_i\right) + Cov\left(Y_{i-1}, Y_{i-1}\right) + Cov\left(Y_{i-1}, Y_i\right) \\ &= 0 + 0 + Var\left(Y_{i-1}\right) + 0 = \sigma^2 \\ Cov\left(X_i, X_j\right) &= 0 \text{ for } j \neq i \pm 1 \end{aligned}$$

and

$$Var(X_i) = Var(Y_{i-1}) + Var(Y_i) = 2\sigma^2$$

so

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) + 2\sum_{i=2}^{n} Cov\left(X_{i-1}, X_i\right) = n(2\sigma^2) + 2(n-1)\sigma^2 = (4n-2)\sigma^2$$

9.6.1 (a)

$$\begin{aligned} P\left(8.4 < X < 12.2\right) &= P\left(\frac{8.4 - 10}{2} < Z < \frac{12.2 - 10}{2}\right) \text{ where } Z \sim N\left(0, 1\right) \\ &= P\left(-0.8 < Z < 1.1\right) = P\left(Z < 1.1\right) - P\left(Z < -0.8\right) \\ &= P\left(Z < 1.1\right) - \left[1 - P\left(Z < 0.8\right)\right] \\ &= 0.8643 + 0.7881 - 1 = 0.6524 \end{aligned}$$

(b) Since $2Y - X$ is Normally distributed with mean $2(3) - 10 = -4$, and variance $2^2(100) + (-1)^2(4) = 404$ then

$$\begin{aligned} P(2Y > X) = P(2Y - X > 0) &= P\left(Z > \frac{0 - (-4)}{\sqrt{404}} = 0.20\right) \\ &= P(Z > 0.20) = 1 - P(Z < 0.20) = 1 - 0.5793 = 0.4207 \end{aligned}$$

(c) $\overline{Y}$ is Normally distributed with mean 3, and variance $100/25 = 4$. Therefore

$$\begin{aligned} P(\overline{Y} < 0) &= P\left(Z < \frac{0 - 3}{2} = -1.5\right) \text{ where } Z \sim N\left(0, 1\right) \\ &= P(Z > 1.5) = 1 - P(Z \leq 1.5) = 1 - 0.9332 = 0.0668 \end{aligned}$$

9.6.2 (a) Since $2X-Y$ is Normally distributed with mean $2(5)-7=3$ and variance $2^2(4)+9=25$ then

$$\begin{aligned}P\left(|2X-Y|>4\right) &= P(2X-Y>4)+P(2X-Y<-4)\\ &= P\left(Z>\frac{4-3}{5}=0.20\right)+P\left(Z<\frac{-4-3}{5}=-1.40\right)\\ &= 0.42074+0.08076=0.5015\end{aligned}$$

(b) Since $\overline{X}\sim N(5,4/n)$

$$P\left(\left|\overline{X}-5\right|<0.1\right)=P\left(|Z|<\frac{0.1}{2/\sqrt{n}}\right)=P\left(|Z|<0.05\sqrt{n}\right)=0.98$$

Since $P\left(|Z|<2.3263\right)=0.98$ we solve $0.05\sqrt{n}=2.3263$ to obtain $n=2164.7$ so $n=2165$.

9.7.1 Let

$$X_i=\begin{cases}0 & \text{if the } i\text{'th pair is alike}\\ 1 & \text{if the } i\text{'th pair is unalike}\end{cases}\qquad \text{for } i=1,2,\ldots,24$$

$$E(X_i)=\sum_{x_i=0}^{1}x_if(x_i)=1f(1)=P(\text{ON OFF }\cup\text{OFF ON})=(0.6)(0.4)+(0.4)(0.6)=0.48$$

$$E(X_i^2)=E(X_i)=0.48$$

$$Var(X_i)=0.48-(0.48)^2=0.2496$$

Consider a pair which has no common switch such as X_1,X_3. Since X_1 depends on switches 1 and 2 and X_3 on switches 3 and 4 and since the switches are set independently, X_1 and X_3 are independent and so $Cov(X_1,X_3)=0$. In fact all pairs are independent if they have no common switch, but may not be independent if the pairs are adjacent. In this case, for example, since X_iX_{i+1} is also an indicator random variable,

$$\begin{aligned}E(X_iX_{i+1}) &= P(X_iX_{i+1}=1)\\ &= P(\text{ON OFF ON}\cup\text{OFF ON OFF})\\ &= (0.6)(0.4)(0.6)+(0.4)(0.6)(0.4)=0.24\end{aligned}$$

Therefore

$$\begin{aligned}Cov(X_i,X_{i+1}) &= E(X_iX_{i+1})-E(X_i)E(X_{i+1})\\ &= 0.24-(0.48)^2=0.0096\end{aligned}$$

$$E\left(\sum_{i=1}^{24}X_i\right)=\sum_{i=1}^{24}E(X_i)=24\times 0.48=11.52$$

$$\begin{aligned} Var\left(\sum_{i=1}^{24} X_i\right) &= \sum_{i=1}^{24} Var(X_i) + 2\sum_{i=1}^{23} Cov(X_i, X_{i+1}) = (24 \times 0.2496) + (2 \times 23 \times 0.0096) \\ &= 6.432 \end{aligned}$$

9.7.2 Using X_i as defined, $E(X_i) = \sum_{x_i=0}^{1} x_i f(x_i) = f(1) = E\left(X_i^2\right)$ since $X_i = X_i^2$ we have

$$\begin{aligned} E\left(X_1\right) &= E\left(X_{24}\right) = 0.9 \text{ since only one cut is needed} \\ E\left(X_2\right) &= E\left(X_3\right) = \cdots = E\left(X_{23}\right) = (0.9)^2 = 0.81 \text{ since two cuts are needed} \\ Var\left(X_1\right) &= Var\left(X_{24}\right) = 0.9 - (0.9)^2 = 0.09 \\ Var\left(X_2\right) &= Var\left(X_3\right) = \cdots = Var\left(X_{23}\right) = 0.81 - (0.81)^2 = 0.1539 \end{aligned}$$

Also

$$Cov\left(X_i, X_j\right) = 0 \text{ if } j \neq i \pm 1 \text{ since there are no common pieces and cuts are independent.}$$

Since

$$\begin{aligned} E\left(X_i X_{i+1}\right) &= \sum x_i x_{i+1} f\left(x_i, x_{i+1}\right) = f(1,1) \\ &= \begin{cases} (0.9)^2 & \text{for } i = 1 \text{ or } i = 23 \text{ two cuts are needed} \\ (0.9)^3 & \text{for } i = 2, 3, \ldots, 22 \text{ three cuts are needed} \end{cases} \end{aligned}$$

we have

$$\begin{aligned} Cov\left(X_i, X_{i+1}\right) &= E\left(X_i X_{i+1}\right) - E(X_i) E\left(X_{i+1}\right) \\ &= \begin{cases} (0.9)^2 - (0.9)(0.9)^2 = 0.081 & \text{for } i = 1 \text{ or } i = 23 \\ (0.9)^3 - (0.9)^2(0.9)^2 = 0.0729 & \text{for } i = 2, 3, \ldots, 22 \end{cases} \end{aligned}$$

Therefore

$$\begin{aligned} E\left(\sum_{i=1}^{24} X_i\right) &= \sum_{i=1}^{24} E(X_i) = (2 \times 0.9) + (22 \times 0.81) = 19.62 \\ Var\left(\sum_{i=1}^{24} X_i\right) &= \sum_{i=1}^{24} Var(X_i) + 2\sum_{i<j} Cov\left(X_i, X_j\right) \\ &= (2 \times 0.09) + (22 \times 0.1539) + 2\left[(2 \times 0.081) + (21 \times 0.0729)\right] = 6.9516 \end{aligned}$$

and the standard deviation of $\sum_{i=1}^{24} X_i$ equals $\sqrt{6.9516} = 2.64.$

10.1.1 Let X be the number germinating. Then $X \sim Binomial(100, 0.8)$. Approximate using a Normal distribution with $\mu = np = 80$ and $\sigma^2 = np(1-p) = 16$.

$$\begin{aligned} P(X \geq 75) &= \sum_{x=75}^{100} \binom{100}{x} (0.8)^x (0.2)^{100-x} \\ &\approx P\left(Z > \frac{74.5 - 80}{4}\right) \text{ where } Z \sim N(0,1) \\ &= P(Z > -1.38) \\ &= P(Z \leq 1.38) = 0.9162 \end{aligned}$$

10.1.2 Let X_i be the cost associated with inspecting part i

$$\begin{aligned} E(X_i) &= (0)(0.6) + (10)(0.3) + (100)(0.1) = 13 \\ E\left(X_i^2\right) &= \left(0^2\right)(0.6) + \left(10^2\right)(0.3) + \left(100^2\right)(0.1) = 1030 \\ Var(X_i) &= 1030 - 13^2 = 861 \end{aligned}$$

By the Central Limit Theorem $S_{80} = \sum\limits_{i=1}^{80} X_i$ is Normal with mean $80(13) = 1040$ and variance $80(861) = 68880$ approximately. Since S_{80} increases in \$10 increments,

$$\begin{aligned} P(S_{80} > 1200) &\approx P\left(Z > \frac{1205 - 1040}{\sqrt{68880}}\right) \text{ where } Z \sim N(0,1) \\ &= P(Z > 0.63) \\ &= 1 - P(Z \leq 0.63) = 0.2643 \end{aligned}$$

12. DISTRIBUTIONS AND $N(0,1)$ TABLES

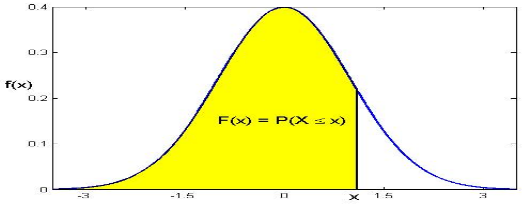
Summary of Discrete Distributions

Notation and Parameters	**Probability Function** $f(x)$	**Mean** $E(X)$	**Variance** $Var(X)$	**Moment Generating Function** $M(t)$
Discrete Uniform(a,b) $b \geq a$ a,b integers	$\frac{1}{b-a+1}$ $x = a, a+1, \ldots, b$	$\frac{a+b}{2}$	$\frac{(b-a+1)^2-1}{12}$	$\frac{1}{b-a+1}\sum_{x=a}^{b} e^{tx}$ $t \in \Re$
Hypergeometric(N,r,n) $N = 1,2,\ldots$ $n = 0,1,\ldots,N$ $r = 0,1,\ldots,N$	$\frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$ $x = \max(0, n-N+r),$ $\ldots, \min(r,n)$	$\frac{nr}{N}$	$\frac{nr}{N}(1-\frac{r}{N})\frac{N-n}{N-1}$	Not tractable
Binomial(n,p) $0 \leq p \leq 1,\ q = 1-p$ $n = 1,2,\ldots$	$\binom{n}{x}p^x q^{n-x}$ $x = 0,1,\ldots,n$	np	npq	$(pe^t+q)^n$ $t \in \Re$
Bernoulli(p) $0 \leq p \leq 1,\ q = 1-p$	$p^x q^{1-x}$ $x = 0,1$	p	pq	$pe^t + q$ $t \in \Re$
Negative Binomial(k,p) $0 < p \leq 1,\ q = 1-p$ $k = 1,2,\ldots$	$\binom{x+k-1}{x}p^k q^x$ $= \binom{-k}{x}p^k(-q)^x$ $x = 0,1,\ldots$	$\frac{kq}{p}$	$\frac{kq}{p^2}$	$\left(\frac{p}{1-qe^t}\right)^k$ $t < -\ln q$
Geometric(p) $0 < p \leq 1,\ q = 1-p$	pq^x $x = 0,1,\ldots$	$\frac{q}{p}$	$\frac{q}{p^2}$	$\frac{p}{1-qe^t}$ $t < -\ln q$
Poisson(λ) $\lambda \geq 0$	$\frac{e^{-\lambda}\lambda^x}{x!}$ $x = 0,1,\ldots$	λ	λ	$e^{\lambda(e^t-1)}$ $t \in \Re$
Multinomial$(n; p_1, p_2, \ldots, p_k)$ $0 \leq p_i \leq 1$ $i = 1,2,\ldots,k$ and $\sum_{i=1}^{k} p_i = 1$	$f(x_1, x_2, \ldots, x_k) =$ $\frac{n!}{x_1!x_2!\cdots x_k!}p_1^{x_1}p_2^{x_2}\cdots p_k^{x_k}$ $x_i = 0,1,\ldots,n$ $i = 1,2,\ldots,k$ and $\sum_{i=1}^{k} x_i = n$	$E(X_i) = np_i$ $i = 1,2,\ldots,k$	$Var(X_i)$ $= np_i(1-p_i)$ $i = 1,2,\ldots,k$	$M(t_1, t_2, \ldots, t_k)$ $=(p_1e^{t_1}+p_2e^{t_2}+\cdots$ $+p_{k-1}e^{t_{k-1}}+p_k)^n$ $t_i \in \Re$ $i = 1,2,\ldots,k-1$

Summary of Continuous Distributions

Notation and Parameters	**Probability Density Function** $f(x)$	**Mean** $E(X)$	**Variance** $Var(X)$	**Moment Generating Function** $M(t)$
Uniform(a,b) $b > a$	$\frac{1}{b-a}$ $a \le x \le b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{bt}-e^{at}}{(b-a)t} \quad t \neq 0$ $1 \quad t = 0$
Exponential(θ) $\theta > 0$	$\frac{1}{\theta}e^{-x/\theta}$ $x \ge 0$	θ	θ^2	$\frac{1}{1-\theta t}$ $t < \frac{1}{\theta}$
$N(\mu,\sigma^2) = G(\mu,\sigma)$ $\mu \in \Re,\ \sigma^2 > 0$	$\frac{1}{\sqrt{2\pi}\,\sigma}e^{-(x-\mu)^2/(2\sigma^2)}$ $x \in \Re$	μ	σ^2	$e^{\mu t+\sigma^2 t^2/2}$ $t \in \Re$

N(0,1) Cumulative Distribution Function



This table gives values of $F(x) = P(X \leq x)$ for $X \sim \mathrm{N}(0,1)$ and $x \geq 0$

x	**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	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	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

N(0,1) Quantiles: This table gives values of $F^{-1}(p)$ for $p \geq 0.5$

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