An Introduction to Probability and Statistics

(statistics 325)

J. Calvin Berry
Mathematics Department
University of Louisiana at Lafayette

http://www.ucs.louisiana.edu/~jcb0773/

May 2022 revision

20220627

Table of Contents

1 Introduction toc

1.1 Basic ideas toc

The entity of primary interest is a group

Statistical methods deal with properties of groups or aggregates. In many applications the entity of primary interest is an actual, physical group (population) of objects. These objects may be animate (*e.g.*, people or animals) or inanimate (*e.g.*, farm field plots, trees, or days). We will refer to the individual objects that comprise the group of interest as **units**. In certain contexts we may refer to the unit as a *population unit, a sampling unit, an experimental unit, or a treatment unit.*

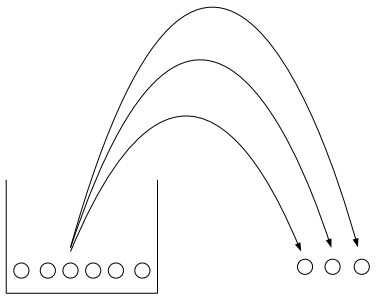
Information about a unit – variables

In order to obtain information about a group of units we first need to obtain information about each of the units in the group. A **variable** is a measurable characteristic of an individual unit. Since our goal is to learn something about the group, we are most interested in the **distribution of the variable**, *i.e.*, the way in which the possible values of the variable are distributed among the units in the group.

The population and the sample

The population is the collection of all of the units that we are interested in. The sample is the subset of the population that we will examine. (We will define a sample more precisely when we discuss random sampling.)

Figure 1.1 Population (box of balls) and sample. A ball represents a population unit. The balls removed from the box represent the sample.



When the units are actual, physical objects we define the **population** as the collection of all of the units that we are interested in. In most applications it is unnecessary or undesirable to examine the entire population. Thus we define a **sample** as a subset or part of the population for which we have or will obtain data. The collection of observed values of one or more variables corresponding to the individual units in the sample constitute the **data**. Once the data are obtained we can use the distributions of the variables among the units in the sample to characterize the sample itself and to make inferences or generalizations about the entire population, *i.e.*, inferences about the distributions of these variables among the units in the population.

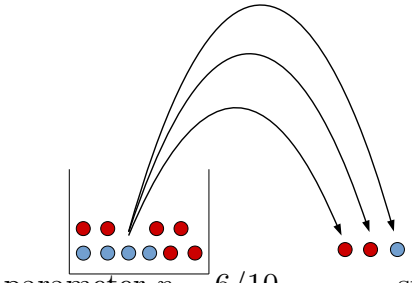
In some applications, such as experimental studies, the population is best viewed as a hypothetical population of values of one or more variables. For example, suppose that we are interested in the effects of an alternative diet on weight gain in some population of experimental animals. We might conduct an experiment by randomly assigning animals to two groups; feeding one group a standard diet and the other group the alternative diet; and then recording the weight gained by each animal over some fixed period of time. In this example we can envision two hypothetical populations of weight gains: The population of weight gains we would have observed if all of the animals were given the standard diet; and, the population of weight gains we would have observed if all of the animals were given the alternative diet.

Information about a group – parameters and statistics

Recall that a **variable** is a measurable characteristic of an individual unit. One way to characterize a group of units is to examine the values of the variable corresponding to all of the units in the group and determine one or more suitable summary values. For example, given a group of adults, we might compute the average age of the group or the proportion who have full–time jobs. A **parameter** is a numerical characteristic of the population. A **statistic** is a numerical characteristic of the sample. That is, a parameter is a number which characterizes a population and a statistic is a number which characterizes a sample.

An illustration is provided in Figure 1.2 for a population of 10 balls and sample of 3 balls. In this figure the characteristic of interest is the color of the ball and the color red (darker shade) is of particular interest. The parameter is the proportion of red balls in the population, 6/10, and the statistic is the proportion of red balls in the sample, 2/3.

Figure 1.2 Parameter and statistic for a population of 10 balls and sample of 3 balls.



parameter $p = 6/10$ statistic $\hat{p} = 2/3$
(population proportion red) (sample proportion red)

Example 1.1 NHANES The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States. The survey is unique in that it combines interviews and physical examinations. We will use some data from the 2013–2014 NHANES to illustrate the basic ideas we are discussing. For now we will concentrate our attention on some body size measurements for the 5588 adults (age 20 and over) in the 2013–2014 NHANES.

For present purposes we will view this group of $N = 5588$ adults as the population. In the original context of the survey this is a sample. Thus, for our purposes:

1. A unit is an individual adult.

2. The population is the collection of $N = 5588$ adults about whom we have information.

3. The sample is a collection of $n = 50$ individuals (units) which I selected at random from the population of $N = 5588$ adults.

4. We will consider six variables:
The sex of the person (male or female);
The age of the person (years);
The weight of the person (pounds);
The height of the person (inches);
The BMI (body mass index) of the person (kg/m^2); and,
The waist circumference of the person (inches).

Table 1.1 contains the values of the six variables for the $n = 50$ people in the sample. The values in a particular row correspond to an individual (one unit).

Table 1.1 NHANES 2013-2014 simple random sample of $n = 50$.

line	sex	age	weight	height	bmi	waist
1	male	48	285.34	70.1181	40.9	48.8976
2	female	80	172.48	63.5433	30.1	41.6535
3	male	48	186.78	69.8819	26.9	37.5197
4	male	80	167.2	67.5591	25.8	37.0079
5	female	20	156.2	67.7953	23.9	32.2835
6	female	43	196.02	66.1811	31.5	46.8504
7	male	54	200.86	64.8425	33.7	42.5197
8	female	24	153.78	63.4252	26.9	40.7874
9	female	25	122.32	64.1732	20.9	29.0945
10	male	58	142.34	69.0157	21.1	34.252
11	female	74	154.22	63.1496	27.2	37.8346
12	female	74	173.8	61.378	32.5	45.3937
13	male	78	154.22	67.4016	23.9	33.4646
14	female	72	89.1	56.9291	19.4	29.1339
15	male	48	178.2	72.2835	24	36.811
16	male	64	231.66	62.874	41.3	51.7323
17	male	41	164.34	68.0315	25	36.7717
18	female	39	143.88	60.7874	27.4	33.7402
19	male	49	305.14	72.1654	41.3	missing
20	male	73	141.24	70.3937	20.1	37.4409
21	female	67	203.94	62.0079	37.4	44.7244
22	female	26	101.86	59.8031	20.1	29.0157
23	female	73	150.7	65.7087	24.6	38.189
24	male	60	199.32	72.1654	27	40.0394
25	male	40	206.58	67.4409	32	43.2677
26	male	27	181.94	67.3622	28.2	38.622
27	male	62	199.98	67.3622	31.1	43.3858
28	female	71	176.66	64.6063	29.8	42.9134
29	female	80	166.1	62.3228	30.1	39.3701
30	male	39	280.72	71.9291	38.2	52.0472
31	female	48	127.6	61.4173	23.8	31.2598
32	male	46	156.64	67.7953	24	missing
33	female	80	147.4	61.3386	27.6	40.7087
34	female	35	183.04	65.2362	30.3	41.2205
35	female	57	140.36	62.2047	25.6	34.9213

continued below

Table 1.1 continuation of NHANES 2013-2014 simple random sample of $n = 50$.

line	sex	age	weight	height	bmi	waist
36	female	40	170.06	65.9055	27.6	35.748
37	female	56	182.6	59.685	36.1	39.4882
38	male	34	192.5	70.3937	27.4	37.9921
39	male	75	163.9	64.7244	27.6	48.8189
40	female	76	147.84	60.4724	28.5	36.6535
41	male	57	254.32	70.7087	35.8	43.8189
42	male	77	187.22	69.2126	27.5	40.1181
43	male	45	179.3	70.7087	25.3	36.5354
44	male	48	249.48	67.126	39	51.063
45	female	47	214.06	64.6457	36.1	42.5984
46	male	72	205.26	70.3543	29.2	42.8346
47	male	80	127.38	66.4567	20.3	missing
48	male	24	199.32	71.2205	27.7	40.6299
49	male	80	162.8	66.7323	25.8	39.4094
50	male	21	119.24	66.6929	18.9	26.3386

Some parameters and statistics (population and sample means) for the numerical variables in this example are given in Table 1.2. With respect to the categorical variable sex of the person; there are 2919 females and 2669 males in the population and there are 22 females and 28 males in the sample. This gives the (parameter) population percentage female as 52.24% and the (statistic) sample percentage female as 44%.

Table 1.2 Population and sample means for the HANES example.

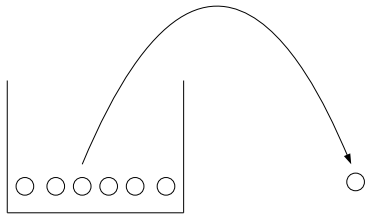
variable	**parameter population mean**	**statistic sample mean**
age	49.15	54.70
weight	179.22	177.94
height	65.77	66.11
bmi	29.10	28.53
waist	39.05	39.47

Which direction? Probability versus statistics

Probability theory is used to model the outcomes of random processes. The basic probability situation is illustrated in Figure 1.3. Here we know all there is to know about the

characteristics of the balls in the box and we want to make a statement about what will happen when we select a ball at random from the box and examine it. For example, for the box of Figure 1.2, where 60% of the balls in the box are red, if we select one ball at random, there is a 60% chance (probability) that it will be red.

Figure 1.3 Probability: What will happen when we select a ball at random?

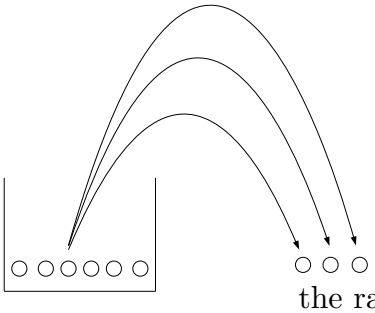


the population

If we know all about the balls in the box, then we can assign probabilities to the outcomes we may observe when we select a ball at random.

Statistical theory is used to make inferences from a random sample to a population. The basic statistics situation is illustrated in Figure 1.4. Here we know all there is to know about the characteristics of the balls in the random sample and we want to make a statement about what we would find if we examined all of the balls in the box (the entire population).

Figure 1.4 Statistics: What can we say about the balls in the box?



the random sample

the population

If we know all about the balls in the random sample, then we can use statistics to make inferences about the balls in the box.

1.2 Sampling toc

A sampling study is conducted by selecting a random sample of units from a population, observing the values of a variable for the units in the sample, and then making inferences or generalizations about the population. More specifically, the distribution of the values

of the variable among the units in a random sample is used to make inferences about the distribution of the variable among the units in the population. The first consideration in planning or interpreting the results of a sampling study is the determination of exactly which units could be in the sample. The second consideration concerns the proper selection of the units which constitute the sample. We will discuss these considerations in more depth in the rest of this section.

Sampling is the process of obtaining a sample from a population. Our ultimate goal is to use the sample (which we will examine) to make inferences about the population (which we will not examine in its entirety). If the sample is selected from the population in an *appropriate fashion*, then we can use the information in the sample to make reliable and quantifiable inferences about the population. When the sample is obtained we will use the distribution of the variable among the units in the sample to make inferences about the distribution of the variable among the units in the population. If the distribution of the variable in the sample was exactly the same as the distribution of the variable in the population, then it would be easy to make inferences about the population; but, this is clearly too much to ask. Therefore we need to determine how to select a sample so that the sample is representative of the population.

The first step in deciding whether a method of choosing a sample will yield a representative sample requires a distinction between two populations. Before we obtain a sample we need to decide exactly which population we are interested in. The **target population** is the collection of all of the units that we want to make inferences about. We then need to determine which population our sample actually comes from. The **sampled population** is the collection of all of the units that *could* be in the sample. Notice that the sampled population is determined by the method used to select the sample.

Ideally the sampling method is chosen so that the sampled population is exactly the same as the target population and we can refer to this collection as the population. In practice, there may be some differences between the target population and the sampled population. When the sampled population is not identical to the target population we cannot be confident that the sample (which comes from the sampled population) will be representative of the target population. Furthermore, we cannot be confident that the statistic (which is based on a sample from the sampled population) will be suitable for inference about the parameter (which corresponds to the target population).

If there is a difference between the sampled population and the target population, in the sense that the distribution of the variable in the sampled population is different from the distribution of the variable in the target population, then a sample (obtained from the sampled population) is said to be **biased** for making inferences about the target population. If we use a biased sample to make inferences about the target population, the resulting inferences will not be appropriate. For example, a statistic based on a biased sample, may provide a suitable estimate of the corresponding parameter in the sampled population; but, it may not provide a suitable estimate of the corresponding parameter in the target population. Therefore, if the sampled population is different from the target population, then we must modify our goals by redefining the target population or we must change the sampled population by modifying our sampling method, since we want these two populations to be the same so that our inferences will be valid for our target population. It may be possible to change the method of obtaining our sample so that all of the units in the target population could be in our sample and these two populations are the same. If it is not possible to change the sampling method, then we must change our goals by restricting our inferences to the sampled population. In any case, once a sampling method has been chosen, the sampled population is determined and we should restrict our inferences to this sampled population. In conclusion, when making inferences from a sample we must carefully consider the restrictions imposed by the sampling method, since statistical theory can only justify inferences about the sampled population.

Assuming that we have defined a method of selecting a sample so that the sampled population is the same as the target population, we next need to consider exactly how we should select the units that constitute the sample. Since we are assuming that the sampled and target populations are the same, we do not need to worry about the type of bias described above. However, we might introduce bias if we do not select the units for the sample in an appropriate fashion. The approach to sampling that we will adopt is called random sampling. The idea behind random sampling is to eliminate potential bias (intentional or unintentional) from the selection process by using *impersonal random chance* to select the sample. In addition to eliminating bias random sampling also provides the basis for theoretical justification and quantification of inferences based on the sample.

All of the sampling situations we consider can be viewed as being abstractly the same as the simple situation of selecting a sample of balls from a box of balls. This scenario was illustrated in Figure 1.1.

The simplest type of random sample is called a simple random sample. A **simple random sample of size n** is a sample of n units selected from the population in such a way that every possible sample of n units has the same chance of being selected. This definition of a simple random sample can be refined to distinguish between two versions of simple random samples. If we require that the possible samples of n units are such that a particular unit can occur at most once in a sample, then we refer to the sample as being a **simple random sample of size n, selected without replacement.** On the other hand, if we allow a particular unit to occur more than once in the sample, then we refer to the sample as a **simple random sample of size n, selected with replacement.**

To obtain a **simple random sample of size n** from the balls in our box, we first mix the balls in the box and select one ball at random (so that each ball in the box has the same chance of being selected). We then determine the value of the variable for the selected ball giving us the value of the variable for one of the balls in our random sample. If we are sampling with replacement we return the ball to the box before the next draw. If we are sampling without replacement we do not return the ball to the box. We then mix the balls in the box and continue this process of selecting a ball from the box at random until n balls have been selected. These n balls (or the values of the variable for these balls) form the simple random sample of size n.

If the population from which we wish to select a random sample is not too large, then it is possible to envision actually labeling each unit in the population, placing these labels on a collection of balls, placing these labeled balls in a box, and selecting a simple random sample of these balls as described above. In fact, state lotteries, where a simple random sample of numbers is selected from a collection of allowable numbers (the units), are conducted in this way. If you have ever observed the complicated mechanisms used to select winning lottery numbers, you know that it is difficult to convince people that a method of "drawing balls from a box" yields a proper simple random sample. For most purposes it is best to use a computer or calculator to select a simple random sample. The computer will simulate the process of drawing balls at random from a box.

When we take a simple random sample, all of the possible samples have the same chance of being selected. There are situations where it is not appropriate for all of the possible samples to have the same chance of being selected. Suppose that there are two or more identifiable subsets of the population (subpopulations). If we obtain a simple random

sample from the whole population, then it is possible for the resulting sample to come entirely from one of the subpopulations so that the sample does not contain any units from one or more of the subpopulations. If we know or suspect that the distribution of the variable of interest varies among the subpopulations, then a sample which does not contain any units from some of the subpopulations will not be representative of the whole population. Therefore, in a situation like this we should not use a simple random sample to make inferences about the whole population. Instead we should use a more complex kind of random sample. One possibility is to use a sampling method known as **stratified random sampling** which is described below in the context of a simple example.

Suppose we wish to estimate the proportion of all registered voters in the United States who favor a particular candidate in an upcoming presidential election. We might expect there to be differences in the proportion of registered voters who favor this candidate among the various states. For example, we might expect support for this candidate to be particularly strong in his or her home state. Because we are interested in the proportion of all registered voters in the United States who favor this candidate, we want to be sure that all of the states are represented fairly in our sample.

We can use the states to define **strata** (subpopulations), take a random sample from each state (stratum), and then combine these samples to get a sample that is representative of the entire country. This is an example of a stratified random sample. The simplest type of **stratified random sample** is obtained as described in the following three steps.

1. Divide the population into appropriate strata (subpopulations).

2. Obtain a simple random sample within each stratum.

3. Combine these simple random samples to get the stratified random sample from the whole population.

To obtain a representative sample in the opinion poll example, we would need to determine the number of registered voters in each state and select simple random samples of sizes that are proportional to the numbers of registered voters in the states.

1.3 Experimentation toc

An experimental study differs from a sampling study in that the units used in the experimental study are manipulated and the responses of the units to this experimental manipulation are recorded. For an experimental study the relevant population or populations are hypothetical populations of values of the variable defined by the experimental treatment(s) and corresponding to all of the units available for use in the experiment. That is, the relevant population(s) is the population(s) of values of the variable which would be observed if all of the available units were subjected to the experimental treatment(s). In the context of a comparative experiment we cannot properly quantify inferences unless the units are assigned to the treatments being compared using an appropriate method of random assignment. This random assignment of units to treatments is analogous to the random sampling of a sampling study.

In an **experimental study** we manipulate the units and observe their response to this manipulation. In the experimental context, a particular combination of experimental conditions is known as a **treatment**. The purpose of an experiment is to obtain information about how the units in the population would respond to a treatment; or, to compare the responses of the units to two or more treatments. The response of a unit to a particular treatment is determined by measuring the value of a suitable **response variable**.

The steps involved in conducting a simple experimental study based on a random sample are summarized below.

1. Obtain a random sample of units from the population of interest.

2. Subject the units in the sample to the experimental treatment of interest.

3. Obtain the data. That is, determine the values of the response variable for the units in the sample.

4. Use the data to make inferences about the how the units in the population would respond to the treatment. More specifically, use the distribution of the response variable in the sample to make inferences about the distribution of the response variable in the population from which the sample was taken. In this context it may be easiest to think of the population as the hypothetical population of values of the response variable which would result if all of the units in the population were subjected to the treatment.

We will now discuss the basic ideas of experimentation in more detail in the context of a simple hypothetical experiment. Suppose that a new drug has been developed to reduce the blood pressure of hypertensive patients. The treatment of interest is the administration of the new drug to a hypertensive patient. The change in a patient's blood pressure will be used as the response variable. For this example the plan of the simple experiment described above is summarized in the steps below.

1. Obtain a random sample of n hypertensive patients.

2. Measure the blood pressure of each patient before the new drug is administered.

3. Administer the new drug to each of these patients.

4. After a suitable period of time, measure the blood pressure of each patient.

5. For each patient determine the change in his or her blood pressure by computing the difference between the patient's blood pressure before the drug was administered and the patient's blood pressure after the new drug was administered. This change or difference will serve as the response variable for assessing the effects of the new drug. In this example the relevant population is the hypothetical population of changes in blood pressure that we would observe if all of the hypertensive patients in the population from which the sample was selected had been subjected to this experiment.

Suppose that we actually conducted this experiment. Furthermore, suppose that the data indicate that the hypertensive patients' blood pressures tend to decrease after they are given the new drug, *i.e.*, suppose that the data indicate that most of the patients experienced a meaningful reduction in blood pressure. We can conclude that there is an association between the new drug and a reduction in blood pressure. This association is clear, since the patients (as a group) tended to experience a decrease in blood pressure after they received the new drug. Can we conclude that the new drug caused this decrease in blood pressure? The support for the contention that the new drug caused the decrease in blood pressure is not so clear. In addition to the new drug there may be other factors associated with the observed decrease in blood pressure. For example, the decrease in blood pressure might be explained, in whole or in part, as the physical manifestation of the psychological effect of receiving medication. In other words, we might observe a similar decrease in blood pressure if we administered a placebo to the patients instead of the new drug. It is also possible that some other aspects of the experimental protocol are

affecting the patients' blood pressures. The way that this experiment is being conducted does not allow us to separate out the effects of the many possible causes of the decrease in blood pressure. If we hope to establish a cause and effect relationship between taking the new drug and observing a decrease in blood pressure, then we need to use a comparative experiment.

In a **randomized comparative experiment** baseline data is obtained at the same time as the data concerning the treatment of interest. This is done by randomly dividing the available units (patients) into two or more groups and comparing the responses for these groups. In the drug example there is one treatment of interest, administer the new drug. Therefore, in this situation we only need two groups, a control group and a treatment group. The units (patients) in the **control group** do not receive the treatment (do not receive the new drug). The units (patients) in the **treatment group** do receive the treatment (do receive the drug). During the course of the experiment we need to keep all aspects of the experiment, other than the treatment itself, as similar as possible for all of the units in the study. The idea is that, if the only difference between the units in the control group and the units in the treatment group is that the units in the treatment group received the treatment, then any observed differences between the responses of the two groups must be caused by the treatment. In the drug example it would be a good idea to administer a placebo to the patients in the control group, so that they do not know that they did not receive the new drug. It would also be a good idea to "blind" the patients and the people administering the drug or placebo by not telling them which patients are receiving the placebo and which patients are receiving the new drug. The purpose of such blinding is to eliminate intentional or unintentional effects due to patient or administrative actions which might affect a patient's response. The steps for conducting such a **randomized comparative experiment** are given below.

1. Randomly divide the group of available patients into two groups: a group of n_1 patients to serve as the control group and a group of n_2 patients to serve as the treatment group. These two groups are random samples of sizes n_1 and n_2 from the group of available patients. The samples sizes n_1 and n_2 may be different.

2. Administer the placebo to the patients in the control group and administer the new drug to the patients in the treatment group.

3. Obtain the data. That is, measure the response variable for each of the $n_1 + n_2$ patients in the experiment. For example, we could determine the change (difference) in each patient's blood pressure as measured before and after administration of the placebo or new drug.

4. Compare the responses of the patients in the treatment group to the responses of the patients in the control group and make inferences about the effects of the new drug versus the placebo.

In this example there are two hypothetical populations of changes in blood pressure. The hypothetical population of changes in blood pressure that we would observe if all of the available hypertensive patients were subjected to this experiment and given the placebo and the hypothetical population of changes in blood pressure that we would observe if all of the available hypertensive patients were subjected to this experiment and given the new drug. Notice that, strictly speaking, our inferences in this example only apply to the hypertensive patients who were available for assignment to the groups used in the experiment. If we want to make inferences about a larger population of hypertensive patients, then the group of available patients used in the study should form a random sample from this larger population.

The experiment described above is designed to compare the effects of the new drug to the effects of a placebo. Suppose that we wanted to compare the effects of the new drug to the effects of a standard drug. To make this comparison we could design the experiment with three groups: a control group, a treatment group for the new drug, and a treatment group for the standard drug. If our only goal is to compare the two drugs (treatments), then we could eliminate the placebo control group and run the experiment with the two treatment groups alone.

2 Descriptive Statistics toc

2.1 Tabular and graphical summary toc

Consider the problem of using data to learn something about the characteristics of the group of units which comprise the sample. Recall that the distribution of a variable is the way in which the possible values of the variable are distributed among the units in the group. A variable is chosen to measure some characteristic of the units in the group; therefore, the distribution of a variable contains all of the available information about the characteristic (as measured by that variable) for the group. Other variables, either alone or in conjunction with the primary variable, may also contain information about the characteristic of interest. A meaningful summary of the distribution of a variable provides an indication of the overall pattern of the distribution and serves to highlight possible unusual or particularly interesting aspects of the distribution.

Generally speaking, it is hard to tell much about the distribution of a variable by examining the data in raw form. Therefore, the first step in summarizing the distribution of a variable is to tabulate the frequencies with which the possible values of the variable appear in the sample. A **frequency distribution** is a table listing the possible values of the variable and their frequencies (counts of the number of times each value occurs). A frequency distribution provides a decomposition of the total number of observations (the sample size) into frequencies for each possible value. In general, especially when comparing two distributions based on different sample sizes, it is preferable to provide a decomposition in terms of relative frequencies. A **relative frequency distribution** is a table listing the possible values of the variable along with their relative frequencies (proportions). A relative frequency distribution provides a decomposition of the total relative frequency of one (100%) into proportions or relative frequencies (percentages) for each possible value.

Many aspects of the distribution of a variable are most easily communicated by a graphical representation of the distribution. The basic idea of a graphical representation of a distribution is to use area to represent relative frequency. The total area of the graphical representation is taken to be one (100%) and sections with area equal to the relative frequency (percentage) of occurrence of a value are used to represent each possible value of the variable.

2.2 Types of variables toc

When discussing the distribution of a variable we need to consider the structure possessed by the possible values of the variable. This leads to the following classification of variables into four basic types.

A **qualitative** variable (categorical variable) classifies a unit into one of several possible categories. The possible values of a qualitative variable are names for these categories. We can distinguish between two types of qualitative variables. A qualitative variable is said to be **nominal** if there is no inherent ordering among its possible values. If there is an inherent ordering of the possible values of a qualitative variable, then it is said to be **ordinal**. For example the sex (female or male) of a college student is nominal while the classification (freshman, sophomore, junior, senior) is ordinal.

A **quantitative** variable (numerical variable) assigns a meaningful numerical value to a unit. Because the possible values of a quantitative variable are meaningful numerical quantities, they can be viewed as points on a number line. If the possible values of a quantitative variable correspond to isolated points on the number line, then there is a discrete jump between adjacent possible values and the variable is said to be a **discrete** quantitative variable. The most common example of a discrete quantitative variable is a count such as the number of babies in a litter of animals or the number of plants in a field plot. If there is a continuous transition from one value of the variable to the next, then the variable is said to be a **continuous** quantitative variable. For a continuous quantitative variable there is always another possible value between any two possible values, no matter how close together the values are. In practice all quantitative variables are discrete in the sense that the observed values are rounded to a reasonable number of decimal places. Thus the distinction between a continuous quantitative variable and a discrete quantitative variable is often more conceptual than real. If a value of the variable represents a measurement of the size of a unit, such as height, weight, or length, or the amount of some quantity, then it is reasonable to think of the possible values of the variable as forming a continuum of values on the number line and to view the variable as continuous.

We can also classify variables with respect to the roles they play in a statistical analysis. That is, we can distinguish between response variables and explanatory variables. A **response variable** is a variable that measures the response of a unit to natural or experimental stimuli. A response variable provides us with a measurement or observation that

characterizes a unit with respect to a characteristic of primary interest. An **explanatory variable** is a variable that can be used to explain, in whole or in part, how a unit responds to natural or experimental stimuli. This terminology is clearest in the context of an experimental study. Consider an experiment where a unit is subjected to a treatment (some combination of conditions) and the response of the unit to the treatment is recorded. A variable that describes the treatment conditions is called an explanatory variable, since it may be used to explain the outcome of the experiment. A variable that measures the outcome of the experiment is called a response variable, since it measures the response of the unit to the treatment. An explanatory variable may also be used to subdivide a group so that the distributions of a response variable can be compared among subgroups.

2.3 Qualitative data

toc

Bar graphs are used to summarize the distribution of a qualitative variable. A **bar graph** consists of a collection of bars (rectangles) such that the combined area of all the bars is one (100%) and the area of a particular bar is the relative frequency of the corresponding value of the variable. Two other common forms for such a graphical representation are segmented bar graphs and pie graphs. A **segmented bar graph** consists of a single bar of area one (100%) that is divided into segments with a segment of the appropriate area for each observed value of the variable. A segmented bar graph can be obtained by joining the separate bars of a bar graph. If the bar of the segmented bar graph is replaced by a disk, the result is a pie graph or pie chart. In a **pie graph** or pie chart the interior of a disk (the pie) is used to represent the total area of one (100%); and the pie is divided into slices of the appropriate area or relative frequency, with one slice for each observed value of the variable.

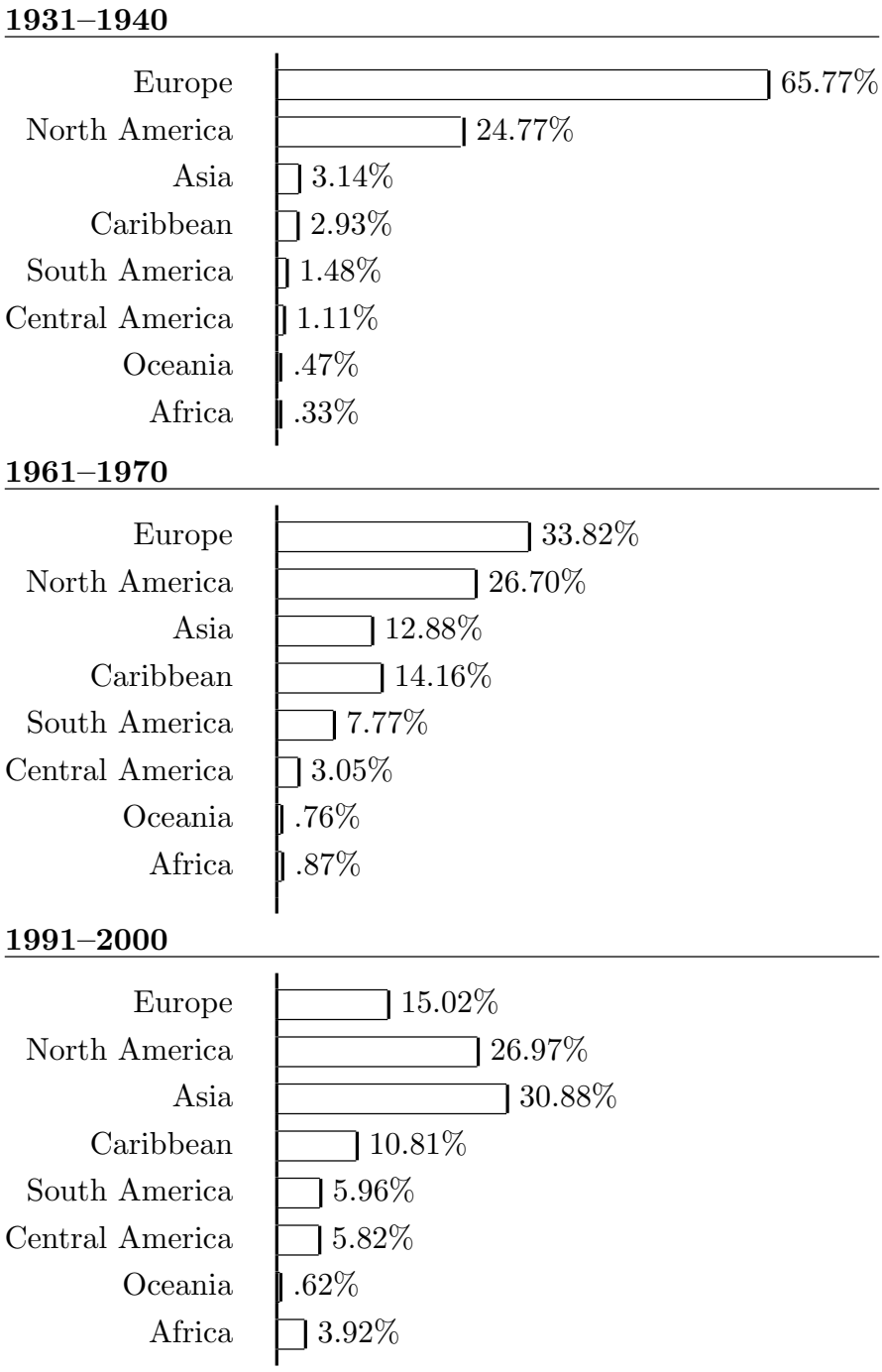
Example 2.1 Immigrants to the United States. The data concerning immigrants admitted to the United States summarized by decade as raw frequency distributions in Table 2.1 were taken from the *2002 Yearbook of Immigration Statistics*, USCIS, (`www.uscis.gov`). Immigrants for whom the country of last residence was unknown are omitted. For this example a unit is an individual immigrant and these data correspond to a census of the entire population of immigrants, for whom the country of last residence was known, for these decades. Because the region of last residence of an immigrant is a nominal variable and its values do not have an inherent ordering, the values in the bar graphs (and relative frequency distributions) in Figure 2.1 have been arranged so that the percentages for the 1931–1940 decade are in decreasing order.

Table 2.1 Region of last residence for immigrants to USA.

	period		
region	**1931–1940**	**1961–1970**	**1991–2000**
Europe	347,566	1,123,492	1,359,737
Asia	16,595	427,692	2,795,672
North America	130,871	886,891	2,441,448
Caribbean	15,502	470,213	978,787
Central America	5,861	101,330	526,915
South America	7,803	257,940	539,656
Africa	1,750	28,954	354,939
Oceania	2,483	25,122	55,845
total	528,431	3,321,634	9,052,999

Two aspects of the distributions of region of origin of immigrants which are apparent in these bar graphs are: The decrease in the proportion of immigrants from Europe; and, the increase in the proportion of immigrants from Asia. In 1931–1940 a large majority (65.77%) of the immigrants were from Europe but for the later decades this proportion steadily decreases. On the other hand, the proportion of Asians (only 3.14% in 1931–1940) steadily increases to 30.88% in 1991–2000. Also note that the proportion of immigrants from North America is reasonably constant for these three decades. The patterns we observe in these distributions may be attributable to several causes. Political, social, and economic pressures in the region of origin of these people will clearly have an impact on their desire to immigrate to the US. Furthermore, political pressures within the US have effects on immigration quotas and the availability of visas.

Figure 2.1 Region of last residence for immigrants to USA, by decade.



2.4 Quantitative data toc

The tabular representations used to summarize the distribution of a discrete quantitative variable, *i.e.,* the frequency and relative frequency distributions, are defined the same as they were for qualitative data. Since the values of a quantitative variable can be viewed as points on the number line, we need to indicate this structure in a tabular representation. In the frequency or relative frequency distribution the values of the variable are listed in order and all possible values within the range of the data are listed even if they do not appear in the data.

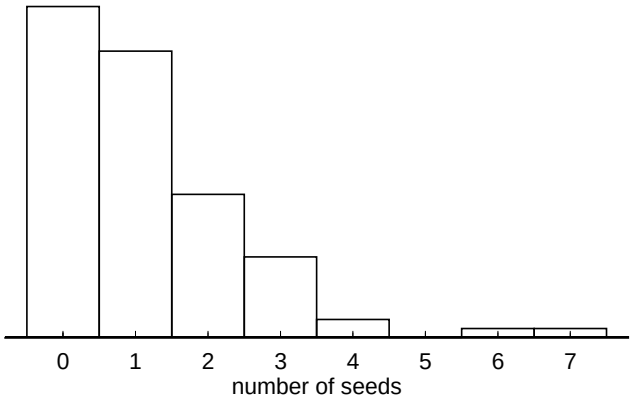
We will use a graphical representation called a histogram to summarize the distribution of a discrete quantitative variable. Like the bar graph we used to represent the distribution of a qualitative variable, the histogram provides a representation of the distribution of a quantitative variable using area to represent relative frequency. A **histogram** is basically a bar graph modified to indicate the location of the observed values of the variable on the number line. For ease of discussion we will describe histograms for situations where the possible values of the discrete quantitative variable are equally spaced (the distance between any two adjacent possible values is always the same). We will use the following weed seed example to illustrate the methodology.

Example 2.2 Weed seeds. C. W. Leggatt counted the number of seeds of the weed *potentilla* found in 98 quarter–ounce batches of the grass *Phleum praetense*. This example is taken from Snedecor and Cochran, *Statistical Methods*, Iowa State, (1980), 198; the original source is C. W. Leggatt, *Comptes rendus de l'association international d'essais de semences*, **5** (1935), 27. The 98 observed numbers of weed seeds, which varied from 0 to 7, are summarized in the relative frequency distribution of Table 2.2 and the histogram of Figure 2.2. In this example a unit is a batch of grass and the number of seeds in a batch is a discrete quantitative variable with possible values of $0, 1, 2, \ldots$.

Table 2.2 Weed seeds relative frequency distribution.

number of seeds	frequency	relative frequency
0	37	.3776
1	32	.3265
2	16	.1633
3	9	.0918
4	2	.0204
5	0	.0000
6	1	.0102
7	1	.0102
total	98	1.0000

Figure 2.2 Histogram for number of weed seeds.



Consider the histogram for the number of weed seeds in a batch of grass of Figure 2.2. This histogram is made up of rectangles of equal width, centered at the observed values of the variable. The heights of these rectangles are chosen so that the area of a rectangle is the relative frequency of the corresponding value of the variable. There is not a gap between two adjacent rectangles in the histogram unless there is an unobserved possible value of the variable between the corresponding adjacent observed values. For this example there is a gap at 5 since none of the batches had exactly 5 weed seeds.

In this histogram we are using an interval of values on the number line to indicate a single value of the variable. For example, the rectangle centered over 1 in the histogram of Figure 2.2 represents the relative frequency that a batch of grass contains exactly 1 weed seed;

but, its base extends from .5 to 1.5 on the number line. Because it is impossible for the number of weed seeds to be strictly between 0 and 1 or strictly between 1 and 2, we are identifying the entire interval from .5 to 1.5 on the number line with the actual value of 1. This identification of an interval of values with the possible value at the center of the interval eliminates gaps in the histogram that would incorrectly suggest the presence of unobserved, possible values.

The histogram for the distribution of the number of weed seeds in Figure 2.2 has a mound shaped appearance with a single peak at zero, indicating that the most common number of weed seeds is zero. In fact, 37.76% of the batches of grass contain no weed seeds. Among the batches that do contain weed seeds we see that 32.65% contain one weed seed and 16.33% contain two. Thus, 86.74% of the 98 batches of grass contain two or fewer weed seeds and 95.92% contain three or fewer weed seeds. In summary, the majority of these batches of grass have a small number of weed seeds; but, there are a few batches with relatively high numbers of weed seeds.

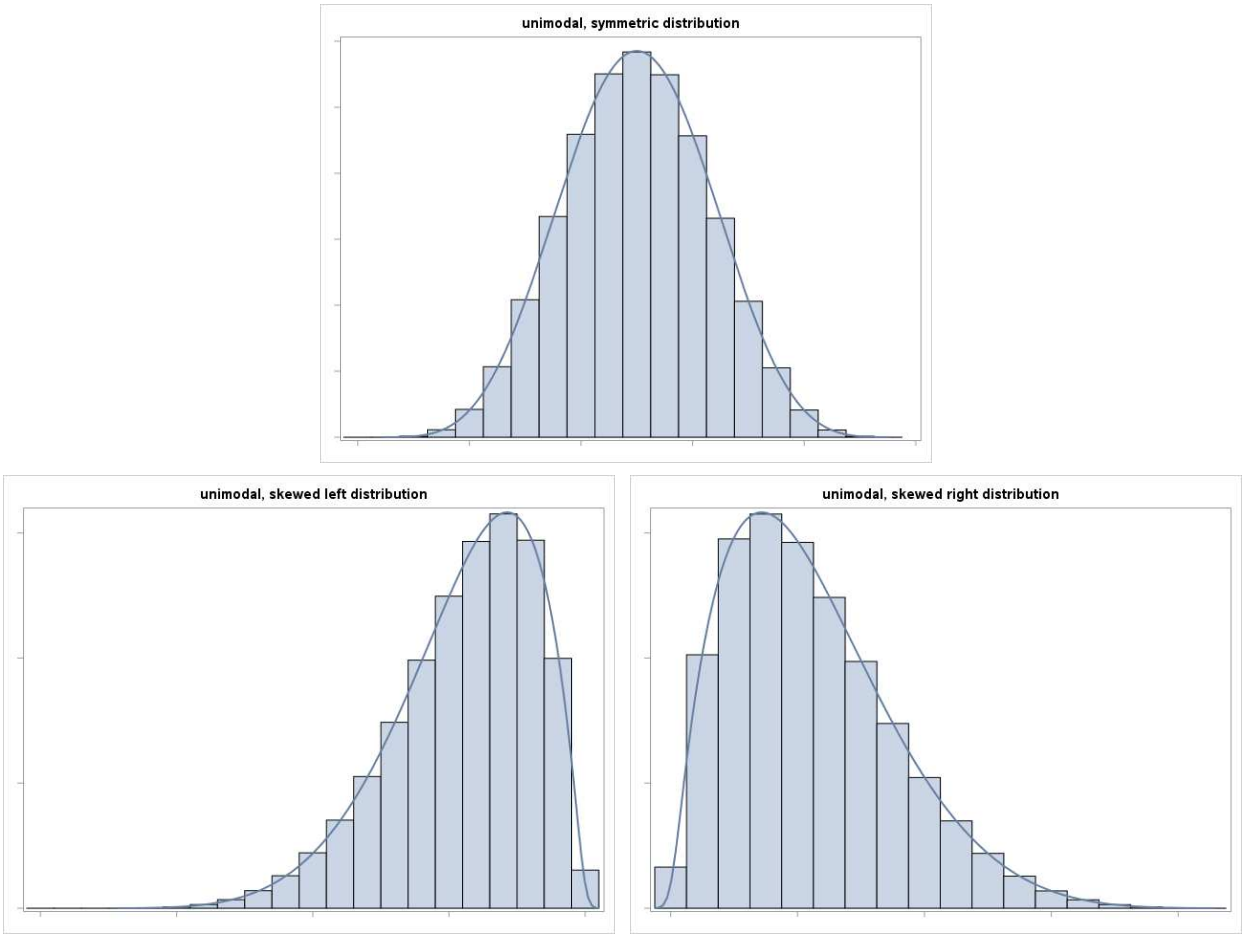
The histogram of Figure 2.2, or the associated distribution, is not symmetric. That is, the histogram (distribution) is not the same on the left side (smaller values) as it is on the right side (larger values). This histogram or distribution is said to be skewed to the right. The concept of a distribution being skewed to the right is often explained by saying that the right "tail" of the distribution is "longer" than the left "tail". That is, the area in the histogram is more spread out along the number line on the right than it is on the left. For this example, the smallest 25% of the observed values are zeros and ones while the largest 25% of the observed values include values ranging from two to seven. In the present example we might say that there is essentially no left tail in the distribution.

The number of weed seeds histogram provides an example of a very common type of histogram (distribution) which is mound shaped and has a single peak. (A distribution with a single peak is said to be unimodal.) This type of distribution arises when there is a single value (or a few adjacent values) which occurs with highest relative frequency, causing the histogram to have a single peak at this location, and when the relative frequencies of the other values taper off (decrease) as we move away from the location of the peak.

Three examples of mound shaped distributions with a single peak are provided in Figure 2.3. For these illustrations a smooth curve is used to indicate the shape of the histogram. The **symmetric** distribution is such that the histogram has two mirror image halves. The

skewed distributions are more spread out along the number line on one side (the direction of the skewness) than they are on the other side.

Figure 2.3 Distribution shapes – symmetry and skewness



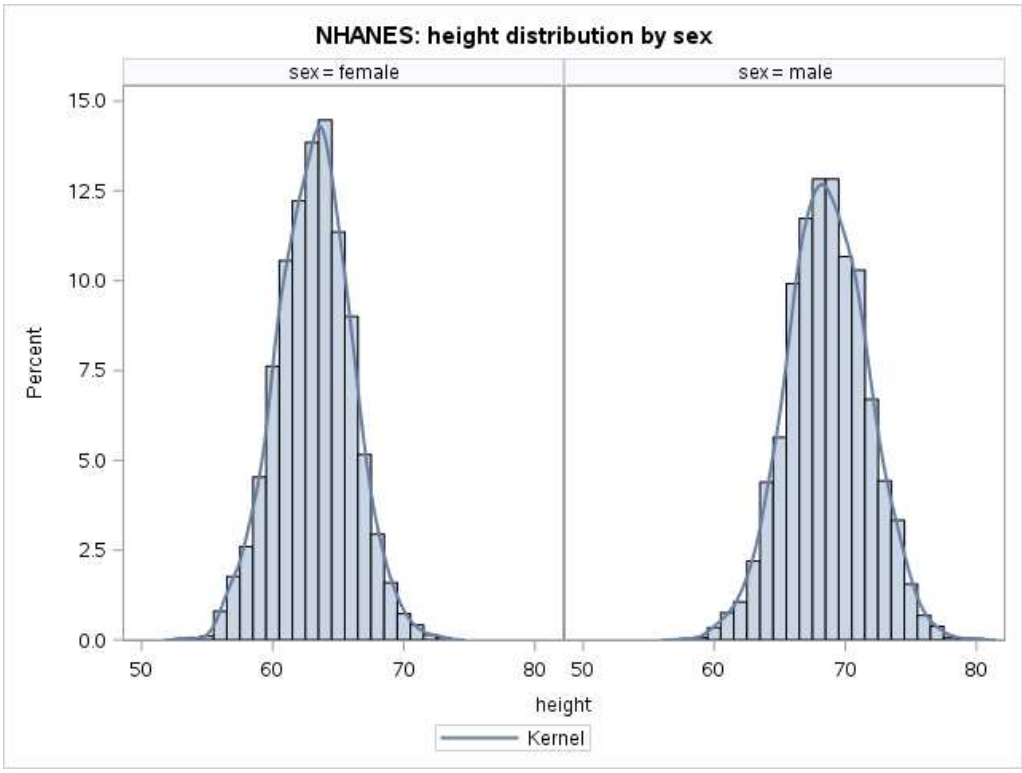
skewed left skewed right

There is a fundamental difference between summarizing and describing the distribution of a discrete quantitative variable and summarizing and describing the distribution of a continuous quantitative variable. Since a continuous quantitative variable has an infinite number of possible values, it is not possible to list all of these values. Therefore, some changes to the tabular and graphical summaries used for discrete variables are required.

In practice, the observed values of a continuous quantitative variable are discretized, *i.e.,* the values are rounded so that they can be written down. Therefore, there is really no difference between summarizing the distribution of a continuous variable and summarizing

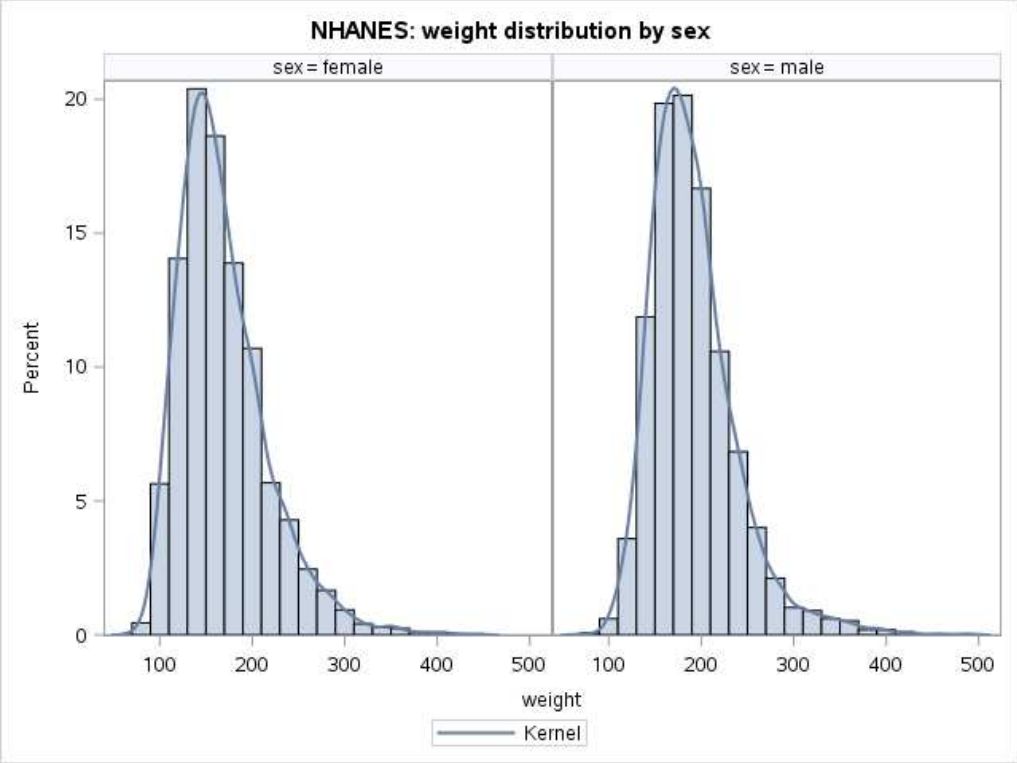
the distribution of a discrete variable with a large number of possible values. In either case, it may be impossible or undesirable to actually list all of the possible values of the variable within the range of the observed data. Thus, when summarizing the distribution of a continuous variable, we will group the possible values into intervals.

Figure 2.4 NHANES 2013–2014 adult height distribution histograms.



Example 1.1 NHANES (revisited). We will use some data from the 2013–2014 NHANES to illustrate the basic ideas we are discussing. For this application we will use all of the 5588 adults (age 20 and over) in the 2013–2014 NHANES for whom a height or weight measurement is available. This group forms a random sample from the population of all adults (age 20 and over) in the US at the time of the survey. Histograms for the heights and weights of the adult participants in the 2013–2014 NHANES, grouped by sex, are given in Figures 2.4 and 2.5. A smooth version of each histogram (smooth curve) is also provided. The height distributions are both unimodal and reasonably symmetric. The weight distributions are both unimodal and skewed to the right. We will discuss these distributions in more depth shortly.

Figure 2.5 NHANES 2013–2014 adult weight distribution histograms.



Notice that the data have been grouped into intervals in order to construct these histograms. For the height distributions the intervals are of length one inch. For the weight distributions the intervals are of length 20 pounds. (In the context of histograms these intervals are also know as bins.) For example, in the weight histograms the area of the rectangle centered over 100 is the proportion of the individuals in the group who had a weight between 90 and 110 pounds.

2.5 Numerical summary

toc

For many purposes a few well–chosen numerical summary values (statistics) will suffice as a description of the distribution of a quantitative variable. A **statistic** is a numerical characteristic of a sample. More formally, a statistic is a numerical quantity computed from the values of a variable, or variables, corresponding to the units in a sample. Thus a statistic serves to quantify some interesting aspect of the distribution of a variable in a sample. Summary statistics are particularly useful for comparing and contrasting the distribution of a variable for two different samples.

If we plan to use a small number of summary statistics to characterize a distribution or to compare two distributions, then we first need to decide which aspects of the distribution are of primary interest. If the distributions of interest are essentially mound shaped with a single peak (unimodal), then there are three aspects of the distribution which are often of primary interest. The first aspect of the distribution is its location on the number line. Generally, when speaking of the location of a distribution we are referring to the location of the "center" of the distribution. The location of the center of a symmetric, mound shaped distribution is clearly the point of symmetry. There is some ambiguity in specifying the location of the center of an asymmetric, mound shaped distribution and we shall see that there are at least two standard ways to quantify location in this context. The second aspect of the distribution is the amount of variability or dispersion in the distribution. Roughly speaking, we would say that a distribution exhibits low variability if the observed values tend to be close together on the number line and exhibits high variability if the observed values tend to be more spread out in some sense. For example, the female height distribution histogram of Figure 2.4 is more peaked than the male height distribution histogram, which is somewhat flatter or more spread out. This indicates that, for this NHANES data, there is less variability among the heights of the females than there is among the heights of the males. The weight distribution histograms of Figure 2.5 suggest that the variability among the weights of the females is similar to the variability among the weights of the males. The third aspect is the shape of the distribution and in particular the degree of skewness in the distribution.

As a starting point consider the **minimum** (smallest observed value) and **maximum** (largest observed value) as statistics. We know that all of the data values lie between the minimum and the maximum, therefore, the minimum and the maximum provide a crude quantification of location and variability. In particular, we know that all of the values of the variable are restricted to the interval from the minimum to the maximum; however, the minimum and the maximum alone tell us nothing about how the data values are distributed within this interval. If the distribution is reasonably symmetric and mound shaped, then the **midrange**, defined as the average of the minimum and the maximum, may provide a suitable quantification of the location of the center of the distribution. The median and mean, which are defined below, are generally better measures of the center of a distribution.

The **range**, defined as the distance from the minimum to the maximum can be used to quantify the amount of variability in the distribution. Note that the range is the positive number obtained by subtracting the minimum from the maximum. When comparing two distributions the distribution with the larger range will generally have more variability than the distribution with the smaller range; however, the range is very sensitive to extreme observations so that one or a few unusually large or small values can lead to a very large range.

We will now consider an approach to the quantification of the shape, location, and variability of a distribution based on the division of the histogram of the distribution into sections of equal area. This is equivalent to dividing the data into groups, each containing the same number of values. We will first use a division of the histogram into halves. We will then use a division of the histogram into fourths.

The median is used to quantify the location of the center of the distribution. In terms of area, the **median** is the number (point on the number line) with the property that the area in the histogram to the left of the median is equal to the area to the right of the median. Here and in the sequel we will use a lower case n to denote the sample size, *i.e.*, n will denote the number of units in the sample. In terms of the n observations, the **median** is the number with the property that at least $n/2$ of the observed values are less than or equal to the median and at least $n/2$ of the observed values are greater than or equal to the median.

A simple procedure for finding the median, which is easily generalized to fractions other that 1/2, is outlined below.

Median computation procedure.
step 1. Arrange the data (observations) in increasing order from the smallest (obs. no. 1) to the largest (obs. no. n). Be sure to include all n values in this list, including repeats if there are any.
step 2. Compute the quantity $n/2$.
step 3a. If $n/2$ is not a whole number, round it up to the next largest integer. The observation at the location indicated by the rounded–up value in the ordered listing of the data is the median.
step 3b. If $n/2$ is a whole number, then we need to average two values to get the median. The two observations to be averaged are obs. no. $n/2$ and the next observation (obs. no.

$n/2+1$) in the ordered listing of the data. Find these two observations and average them to get the median.

We can use the distance between the minimum and the median and the distance between the median and the maximum to quantify the amount of skewness in the distribution. The distance between the minimum and the median is the range of the lower (left) half of the distribution, and the distance between the median and the maximum is the range of the upper (right) half of the distribution. If the distribution is symmetric, then these two distances (median – minimum) and (maximum – median) will be equal. If the distribution is skewed, then we would expect to observe a larger range (indicating more variability) for the half of the distribution in the direction of the skewness. Thus if the distribution is skewed to the left, then we would expect (median – minimum) to be greater than (maximum – median). On the other hand, if the distribution is skewed to the right, then we would expect (maximum – median) to be greater than (median – minimum).

Example 2.2 Weed seeds (revisited). Recall that this example is concerned with the number of weed seeds found in $n = 98$ quarter–ounce batches of grass. Since $98/2 = 49$, the median for this example is the average of observations 49 and 50. Referring to Table 2.2 we find that the minimum number of weed seeds is 0, the maximum is 7, and the median is 1, since observations 49 and 50 are each 1. The range for this distribution is $7 - 0 = 7$. Notice that the range of the right half of this distribution (maximum – median) $= 7 - 1 = 6$ is much larger than the range of the left half (median – minimum) $= 1 - 0 = 1$ confirming our observation that this distribution is strongly skewed to the right.

A more detailed quantification of the shape and variability of a distribution can be obtained from a division of the distribution into fourths. In order to divide a distribution into fourths, we need to specify three numbers or points on the number line. These statistics are called **quartiles**, since they divide the distribution into quarters. In terms of area, the **first quartile**, denoted by Q_1 (read this as Q sub one), is the number (point on the number line) with the property that the area in the histogram to the left of Q_1 is equal to one fourth and the area to the right of Q_1 is equal to three fourths. The **second quartile**, denoted by Q_2, is the median. The **third quartile**, denoted by Q_3, is the number (point on the number line) with the property that the area in the histogram to the left of Q_3 is equal to three fourths and the area to the right of Q_3 is equal to one fourth. In terms of the n observations, Q_1 is the number with the property that at least $n/4$ of the observed

values are less than or equal to Q_1 and at least $3n/4$ of the observed values are greater than or equal to Q_1. Similarly, Q_3 is the number with the property that at least $3n/4$ of the observed values are less than or equal to Q_3 and at least $n/4$ of the observed values are greater than or equal to Q_3.

The method for finding the median given above is readily modified for finding the first and third quartiles. For Q_1, we simply replace $n/2$ by $n/4$ and replace the words 'the median' by Q_1. To find Q_3, use exactly the same method but count down from the largest value instead of counting up from the smallest value. Some calculators and computer programs use variations of the methods given above for finding Q_1 and Q_3. These variations may give slightly different values for Q_1 and Q_3.

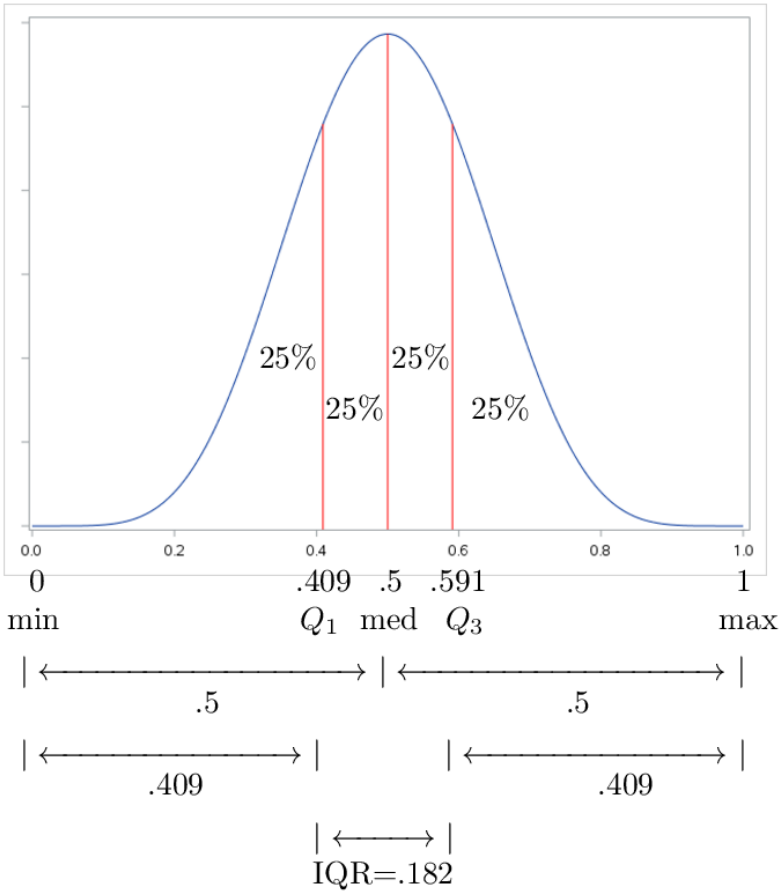
Example 2.1 Weed seeds (revisited). Since $98/4 = 24.5$, the quartiles Q_1 and Q_3 for this example are the observations located at position 25 counting up for Q_1 and counting down for Q_3. Referring to Table 2.2 we find that $Q_1 = 0$ and $Q_3 = 2$. Notice that the range of the lower three–fourths of this distribution, Q_3 – minimum, is 2 while the range of the upper fourth, maximum – Q_3 is 5. This indicates that 75% (a large proportion) of the batches of grass have relatively few weed seeds, and the skewness in this distribution is due to the high amount of variability among the numbers of weed seeds in the 25% of the batches with between 2 and 7 weed seeds.

Previously we introduced the range as a measure of variability. An alternative measure of variability is provided by the interquartile range. The **interquartile range** (IQR) is the distance between the first quartile Q_1 and the third quartile Q_3, *i.e.*, the interquartile range is the positive number obtained by subtracting Q_1 from Q_3. Notice that the **interquartile range** is the range of the middle half of the distribution. The interquartile range is less sensitive to the presence of a few extreme observations in the data than is the range. For example, if there are one or two unusually large or unusually small values, then these values may have the effect of making the range much larger than it would be if these unusual values were not present. In such a situation, we might argue that the range is too large to be deemed an appropriate overall measure of the variability of the distribution. The interquartile range is not affected by a few unusual values, since it only depends on the middle half of the data. We could use the range of a larger part of the middle of the distribution, say the middle 75% or 90%, as a compromise between the range and the interquartile range.

The five summary statistics: the minimum (min), the first quartile (Q_1), the median (med), the third quartile (Q_3), and the maximum (max), constitute the **five number summary** of the distribution. Each of these five statistics provides a quantification of a particular aspect of the distribution. They quantify where the distribution begins, where the first quarter of the distribution ends, and so on. Furthermore, the distances between these five statistics can be used to quantify the shape (skewness) of the distribution.

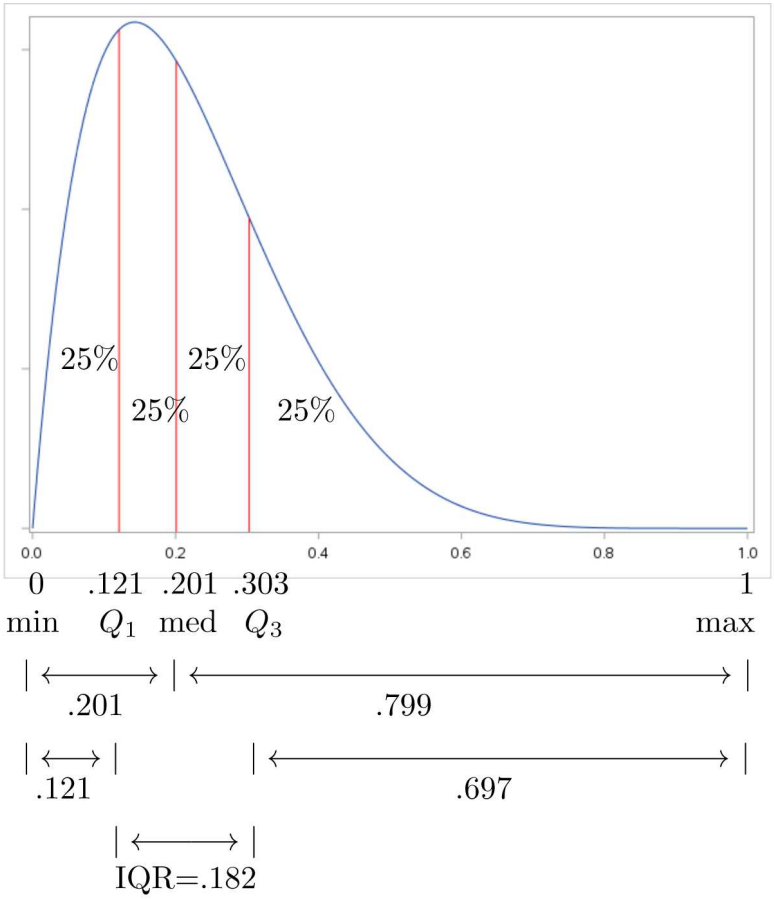
The four distances: (Q_1 – min), (med – Q_1), (Q_3 – med), and (max – Q_3), are the ranges of the first, second, third, and fourth quarters of the distribution, respectively. These distances can be used to quantify the amount of variability in the corresponding parts of the distribution. Comparisons of appropriate pairs of these distances provide indications of certain aspects of the shape of the distribution. The relationship between (med – Q_1) and (Q_3 – med) can be used to quantify the shape (skewness) of the middle half of the distribution. Since (Q_1 – min) and (max – Q_3) are the lengths of the tails (lower and upper fourths) of the distribution, the relationship between these numbers can be used to quantify skewness in the tails of the distribution.

Figure 2.6. Mound shaped, single peak, symmetric distribution



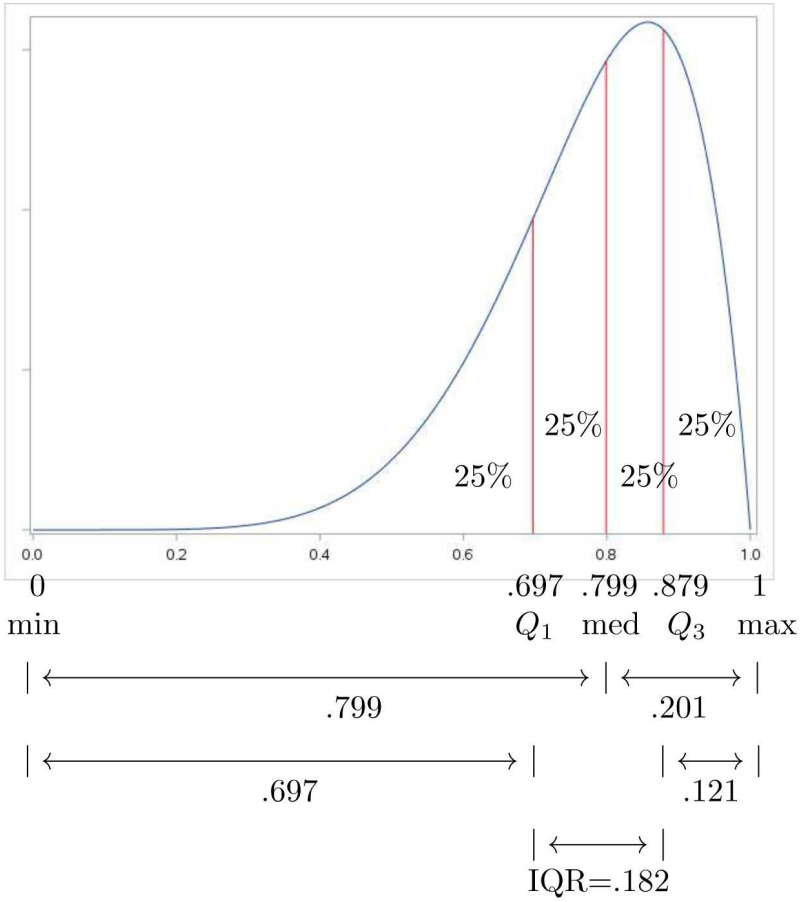
The distribution of Figure 2.6 is mound shaped with a single peak (mode) at .5. This distribution is symmetric. Since this distribution is symmetric we see that the range of the left half of this distribution .5 is equal to the range of the right half; the range of the left tail .409 is equal to the range of the right tail; and, the median .5 is exactly half way between $Q_1 = .409$ and $Q_3 = .591$.

Figure 2.7 Mound shaped, single peak, skewed right distribution



The distribution of Figure 2.7 is mound shaped with a single peak (mode) around .15. This distribution is clearly skewed right. The fact that the range of the right half of this distribution .799 is about 4 times .201 the range of the left half shows extreme skewness to the right. This skewness is mostly due to the fact that the range of the right tail .697 is almost 6 times .121 the range of the left tail. Notice that the middle half of the distribution is reasonably symmetric since the median .201 is about half way between $Q_1 = .121$ and $Q_3 = .303$.

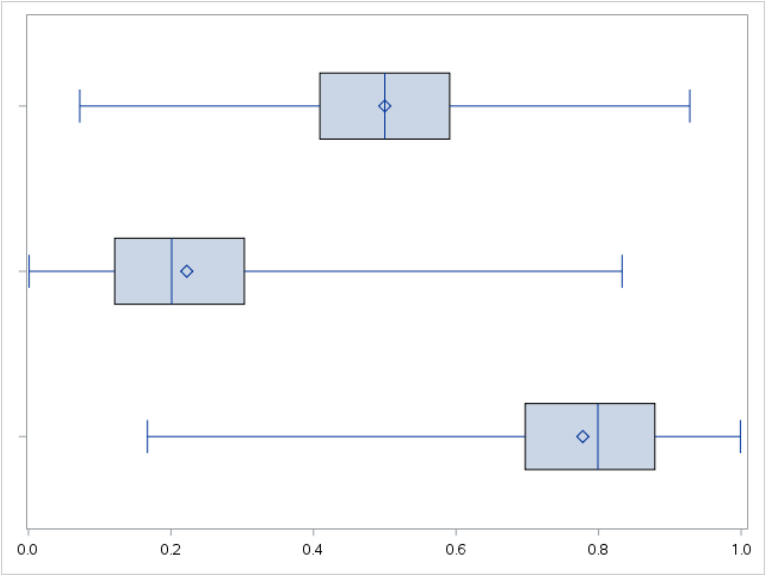
Figure 2.8 Mound shaped, single peak, skewed left distribution



The distribution of Figure 2.8 is mound shaped with a single peak (mode) around .85. This distribution is clearly skewed left. The fact that the range of the left half of this distribution .799 is about 4 times .201 the range of the right half shows extreme skewness to the left. This skewness is mostly due to the fact that the range of the left tail .697 is almost 6 times .121 the range of the right tail. Notice that the middle half of the distribution is reasonably symmetric since the median .799 is about half way between $Q_1 = .697$ and $Q_3 = .879$.

We can use the five number summary values to form a simple graphical representation of a distribution known as a **boxplot** or a box and whiskers plot. A boxplot provides a useful graphical impression of the shape of the distribution as well as its location and variability. Simple boxplots for unimodal mound shaped distributions similar to the distributions of Figures 2.6, 2.7, and 2.8 are provided in Figure 2.9.

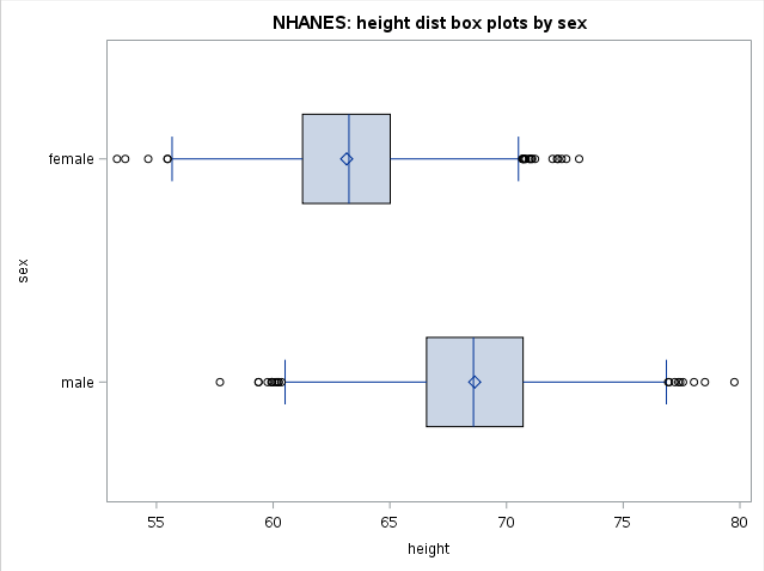
Figure 2.9 Box plots for unimodal mound shaped distributions.



Each boxplot has five vertical marks indicating the locations of the five number summary values. The box which extends from the first quartile to the third quartile and is divided into two parts by the median gives an impression of the distribution of the values in the middle half of the distribution. In particular, a glance at this box indicates whether the middle half of the distribution is skewed or symmetric and indicates the magnitude of the interquartile range (the length of the box). The line segments (whiskers) which extend from the ends of the box to the extreme values (the minimum and the maximum) give an impression of the distribution of the values in the tails of the distribution. The relative lengths of the whiskers indicate the contribution of the tails of the distribution to the symmetry or skewness of the distribution.

Example 1.1 NHANES (revisited). Boxplots for the NHANES height and weight distributions are given in Figures 2.10 and 2.11. In these boxplots the whiskers are modified to show extreme observations (indicated by o's in these plots) in the tails of the distribution. Summary information for the NHANES height and weight distributions is given in Tables 2.3 and 2.4.

Figure 2.10 NHANES height distribution boxplots.



reasonably symmetric

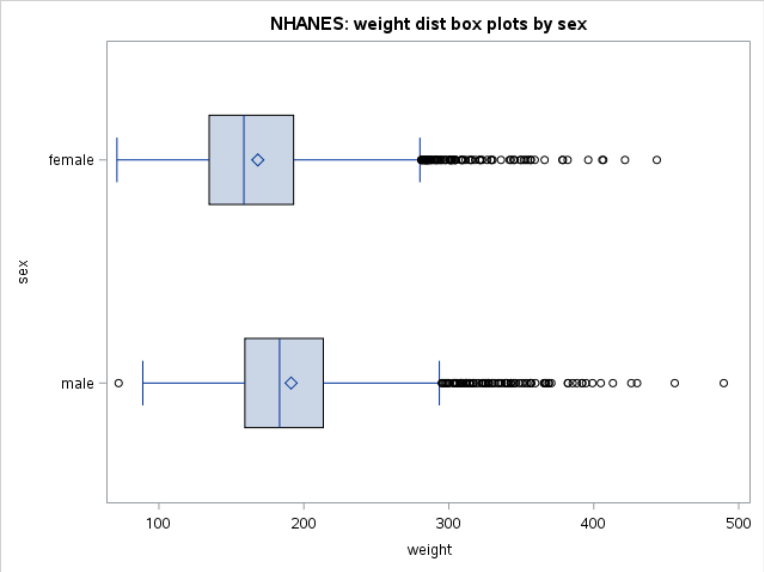
reasonably symmetric

Table 2.3 NHANES height distribution summary information. (height in inches)

	group	
statistic	**females**	**males**
n	2888	2642
location		
mean	63.15	68.63
median	63.25	68.58
variability		
std deviation	2.82	3.06
variance	7.97	9.39
range	19.80	22.05
IQR	3.76	4.13
5 number summary		
min	53.31	57.72
Q_1	61.26	66.57
median	63.25	68.58
Q_3	65.02	70.71
max	73.11	79.76
distances		
Q_1-min	7.95	8.85
med-Q_1	1.99	2.01
Q_3-med	1.77	2.13
max-Q_3	8.09	9.05

As noted earlier, the height distribution histograms for the adult males and the adult females of Figure 2.4 are both unimodal and reasonably symmetric. The boxplots of Figure 2.10 also indicate reasonable symmetry. We will now use the distances from Table 2.3 to quantify this claim of reasonable symmetry. For the adult males we see that the range of the left tail 7.95 is almost the same as 8.09 the range of the right tail. Looking at the middle half of the male height distribution we find that the range of the left side $\text{med} - Q_1 = 1.99$ is only slightly larger than the range of the right side $Q_3 - \text{med} = 1.77$. Similarly, for the adult females the range of the left tail 8.85 is almost the same as 9.05 the range of the right tail; and, $\text{med} - Q_1 = 2.01$ is only slightly smaller than $Q_3 - \text{med} = 2.13$.

Figure 2.11 NHANES weight distribution boxplots.



strongly skewed right

strongly skewed right

Turning to the weight distributions, recall that the weight distribution histograms for the adult males and the adult females of Figure 2.5 are both unimodal and strongly skewed to the right. The boxplots of Figure 2.11 also indicate strong skewness to the right. From each of these boxplots it is clear that in both cases the skewness is due to the extreme variability in the right tail of the distribution, that is, due to high variability among the weights of the heaviest 25% of the group. Note also that in each case the middle half of the distribution is reasonably symmetric. The distances in Table 2.4 readily support these observations. For the male weights $\max - Q_3 = 250.58$ is much larger than $Q_1 - \min = 63.58$, while, relatively speaking, $\text{med} - Q_1 = 23.98$ is only slightly smaller than $Q_3 - \text{med} = 34.32$. Similarly,

for the female weights $\max - Q_3 = 276.32$ is much larger than $Q_1 - \min = 87.12$, while, relatively speaking, $\text{med} - Q_1 = 23.98$ is only slightly smaller than $Q_3 - \text{med} = 30.14$.

Table 2.4 NHANES weight distribution summary information. (weight in pounds)

	group	
statistic	**females**	**males**
n	2888	2645
location		
mean	168.23	191.21
median	158.62	183.26
variability		
std deviation	47.70	46.80
variance	2275	2190
range	372.46	417.56
IQR	58.30	54.12
5 number summary		
min	71.06	72.16
Q_1	134.64	159.28
median	158.62	183.26
Q_3	192.94	213.40
max	443.52	489.72
distances		
Q_1-min	63.58	87.12
med-Q_1	23.98	23.98
Q_3-med	34.32	30.14
max-Q_3	250.58	276.32

2.6 Quantiles and percentiles in general

toc

We will now provide an extension of the method we used to compute the median and quartiles to allow an arbitrary fraction. Given a proportion p (a fraction between zero and one), the pth **quantile** ($p \times 100th$ **percentile**) of the distribution of X is the value Q_p with the property that if we choose a value of X at random, then X will be less than Q_p with probability p and X will be greater than Q_p with probability $1 - p$, *i.e.*, $p \times 100\%$ of the time X will be less than Q_p and $(1 - p) \times 100\%$ of the time X will be greater than Q_p. In terms of the histogram of the distribution of X this means that the area in the histogram to the left of Q_p is p and the area to the right of Q_p is $1 - p$. Note that the first quartile is the $25th$ percentile, the median is the $50th$ percentile, and the third quartile is the $75th$ percentile.

In terms of a sample of n values, the pth quantile is the number Q_p with the property that at least pn of the sample values are less than or equal to Q_p and at least $(1-p)n$ of the sample values are greater than or equal to Q_p. In order to compute a quantile we first need to sort (order) the data. Let $x_1, x_2, \ldots, x_n$ denote the (unordered) data. Let $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ denote the ordered values with $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. These ordered values are know as **order statistics**.

Quantile computation procedure. For $0 < p < 1$, the pth quantile Q_p of the n sample values $x_1, \ldots, x_n$ of X is computed as follows. Let $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ denote the sample values ordered from smallest to largest. There are two cases
case 1: If there is an integer k such that $pn = k$, then $Q_p = \frac{x_{(k)}+x_{(k+1)}}{2}$
case 2: If there is an integer k such that $k < pn < k+1$, then $Q_p = x_{(k+1)}$.

2.7 The mean and standard deviation toc

The approach that we have been using to form summary statistics is to select a single representative value from the observed values of the variable (or the average of two adjacent observed values) to quantify a particular aspect of the distribution. We have also considered statistics that are distances between two such representative values.

An alternative approach to forming a summary statistic is to combine all of the observed values to get a suitable statistic. The first statistic of this type that we consider is the mean. The **mean**, which is the simple arithmetic average of the n data values, is used to quantify the location of the center of the distribution. You could compute the mean by adding all n data values together and dividing this sum by n; however, it is better to use a calculator or a computer. The sample mean is often denoted by the symbol $\overline{X}$ (read this as X bar).

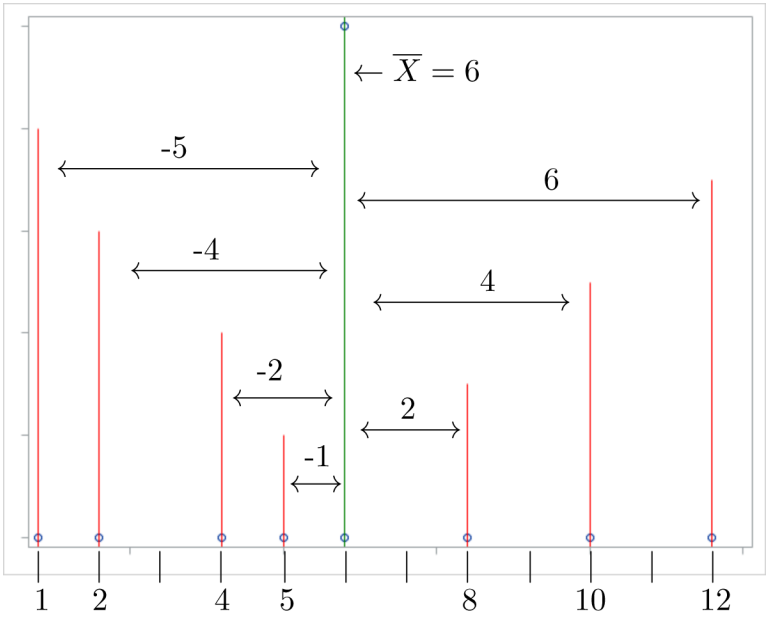
Recall that the median is the number (point on the number line) with the property that the area in the histogram to the left of the median is equal to the area to the right of the median. The mean is the number (point on the number line) where the histogram would balance. To understand what we mean by the balance point, imagine the histogram as being cut out of a piece of cardboard. The mean is located at the point along the number line side of this cutout where the histogram cutout would balance. These geometric characterizations of the mean and the median imply that when the distribution is symmetric the mean will be equal to the median. Furthermore, if the distribution is skewed to the right, then the mean (the balance point) will be larger than the median (to the right of the median).

Similarly, if the distribution is skewed to the left, then the mean (the balance point) will be smaller than the median (to the left of the median).

The primary use of the mean, like the median, is to quantify the location of the center of a distribution and to compare the locations (centers) of two distributions. Since both the mean and the median can be used to quantify the location of the center of a distribution, it seems reasonable to ask which is more appropriate. If the distribution is approximately symmetric, then the mean and the median will be approximately equal. On the other hand, if the distribution is not symmetric, then the median is likely to provide a better indication of the center of the distribution. For example, if the distribution is strongly skewed to the right, then the mean may be much larger than the median and the mean may not be a good indication of the center of the distribution. For a specific application it is a good idea to mark the locations of the mean and the median on a histogram of the distribution and consider which seems more reasonable as an indicator of the center of the distribution.

The two measures of variability we discussed earlier, the range and the interquartile range, are distances between two representative values, the minimum and maximum for the range and the first and third quartiles for the interquartile range. We will now discuss a more complex measure of variability which is based on the distances between each of the observations and a single representative value. If the mean $\overline{X}$ is deemed suitable as a measure of the center of the distribution of X, then the deviations $(X-\overline{X})$ of the observed values of X from their mean $\overline{X}$ contain information about the amount of variability in the distribution. If there is little variability (the observed values of X are close together and they are close to the mean $\overline{X}$), then the deviations $(X-\overline{X})$ will tend to be small in magnitude (absolute value). On the other hand, if there is a lot of variability (at least some of the observed values of X are far apart and they are not all close to the mean $\overline{X}$), then the deviations $(X-\overline{X})$ will tend to be large in magnitude. It is this observation which suggests that a summary statistic based on the distances between each of the observed values of the variable and their mean can be used to measure the variability in the distribution. These deviations from the mean are illustrated, for a small sample of values, in Figure 2.12.

Figure 2.12 An example showing deviations from the mean.
X values: 1, 2, 4, 5, 8, 10, 12; $n = 7$; mean: $\overline{X} = 6$



The most commonly used measure of variability based on these deviation from the mean is the **standard deviation**. The **standard deviation** is the square root of the "average" of the squared deviations of the observed values of the variable from their mean. The formula for the standard deviation given below is not intended for computation purposes; you should use a calculator or a computer to compute the standard deviation. The standard notation for the **sample standard deviation** of the distribution of the variable X is S_X (read this as S sub X). The defining formula is

$$S_X = \sqrt{\frac{\Sigma(X - \overline{X})^2}{n - 1}}$$

In this formula the capital Greek letter sigma, Σ, represents the statement "the sum of", and $(X - \overline{X})^2$ denotes the square of the distance from the observed value X to the mean $\overline{X}$. Therefore, as mentioned above, the expression under the square root sign in the formula is the "average" of the squared deviations of the observed values of the variable from their mean. The reason for the square root is so that the standard deviation of X and the variable X are in the same units of measurement. The quantity

$$S_X^2 = \frac{\Sigma(X - \overline{X})^2}{n - 1}$$

is the **sample variance** of the distribution of the variable X. The calculations corresponding to these formulae are demonstrated below for the sample used in Figure 2.12.

Example 2.3 Computation of the standard deviation. Consider the sample of seven ($n = 7$) X values: 1, 2, 4, 5, 8, 10, 12. The sample mean is

$$\overline{X} = \frac{1+2+4+5+8+10+12}{7} = \frac{42}{7} = 6.$$

The sample variance is

$$\begin{aligned} S_X^2 &= \frac{(1-6)^2+(2-6)^2+(4-6)^2+(5-6)^2+(8-6)^2+(10-6)^2+(12-6)^2}{7-1} \\ &= \frac{(-5)^2+(-4)^2+(-2)^2+(-1)^2+(2)^2+(4)^2+(6)^2}{6} \\ &= \frac{25+16+4+1+4+16+36}{6} = \frac{102}{6} = 17. \end{aligned}$$

The sample standard deviation is

$$S_X = \sqrt{17}.$$

The standard deviation is positive, unless there is no variability at all in the data. That is, unless all of the observations are exactly the same, the standard deviation is a positive number. The standard deviation is a very widely used measure of variability. Unfortunately, the standard deviation does not have a simple, direct interpretation. The important thing to remember is that larger values of the standard deviation indicate that there is more variability in the data.

There are quotation marks around the word average in the definition of the sample standard deviation because we divided by $n-1$ even though there are n squared deviations in the average. When the standard deviation (variance) is computed for the population this divisor is changed to n and a lower case Greek sigma is used instead of an S. That is, the **population standard deviation** is defined as

$$\sigma_X = \sqrt{\frac{\Sigma(X-\overline{X})^2}{n}}$$

and the **population variance** is σ_X^2.

Example 1.1 NHANES (revisited). We noted earlier that the shapes of the female and male height distributions are very similar and that the shapes of the female and male weight

distributions are also very similar. We will now use summary statistics from Tables 2.3 and 2.4 to compare and contrast the locations and the variability in these distributions.

We will first look at the height distributions. As you would expect the males tend to be taller than the females. On average the males are about 5.5 inches taller than the females (male mean height of 68.63 inches versus female mean height of 63.15 inches). Since the height distributions are reasonably symmetric the median heights are very similar to the mean heights. In terms of variability, there is slightly more variability among the heights of the males: male height standard deviation of 3.06 inches verses female height standard deviation of 2.82 inches; male height range of 22.05 inches verses female height range of 19.80 inches; and, male height interquartile range of 4.13 inches verses female height interquartile range of 3.76 inches.

Now we will look at the weight distributions. Again, as you would expect, the males tend to be heavier than the females. On average the males are about 23 pounds heavier than the females (male mean weight of 191.21 pounds versus female mean weight of 168.23 pounds). Since the both weight distributions are strongly skewed right each median weight is smaller than the corresponding mean weight; but, as it turns out, the difference between the median weights is similar to the difference between the mean weights (the median weight difference is about 25 pounds). The weight distributions are quite similar in terms of variability.

2.8 A measure of relative position

toc

Percentiles can be used to quantify the location of a particular value of X relative to a group. Another widely used measure of the relative position of a value within a group is its Z–score. The Z–score of X quantifies the location of X relative to the mean $\overline{X}$ of the sample in terms of the standard deviation S_X of the sample. Since the Z–score is based on $\overline{X}$ and S_X, the Z–score is only appropriate when $\overline{X}$ and S_X are appropriate measures of the center and variability in the sample, respectively. We will develop the Z–score in two stages.

First, we need a measure of the location of X relative to the center of the distribution as determined by the mean $\overline{X}$. The deviation, $X - \overline{X}$, of X from the mean $\overline{X}$ is such a measure. The deviation $X - \overline{X}$ is the signed distance from the particular value X to the mean $\overline{X}$. If $X - \overline{X}$ is negative, then X is below (smaller than) the mean. If $X - \overline{X}$ is positive,

then X is above (larger than) the mean. In summary, the sign of the deviation $X - \overline{X}$ indicates the location of X relative to the mean $\overline{X}$; and the magnitude of the deviation $|X - \overline{X}|$ is the distance from X to the mean $\overline{X}$, measured in the units of measurement used for the observation X.

Second, we want a measure of the location of X relative to the mean $\overline{X}$ which takes the amount of variability in the data into account. We will obtain such a measure by using the standard deviation S_X of the sample to standardize the deviation $X - \overline{X}$. Given a particular value X, the sample mean $\overline{X}$, and the sample standard deviation S_X, the **Z–score** corresponding to X is

$$Z = \frac{X - \overline{X}}{S_X}.$$

The sign of the Z–score indicates the location of X relative to the mean $\overline{X}$ and the magnitude of the Z–score is the distance from X to the mean $\overline{X}$ in terms of standard deviation units. For example, if $Z = 2$, then X is two standard deviation units above the mean ($X = \overline{X} + 2S_X$), and, if $Z = -2$, then X is two standard deviation units below the mean ($X = \overline{X} - 2S_X$).

Interpretation of a Z–score requires some knowledge of the connection between Z–scores and percentiles. The $68\% - 95\% - 99.7\%$ rule given below allows us to associate a percentage with a Z–score. This rule works best for distributions that are unimodal (single peaked), mound shaped, and symmetric. A formal statement of the rule follows.

The 68%-95%-99.7% rule. *For a distribution that is unimodal (has a single peak), mound shaped, and reasonably symmetric:*
i) Approximately 68% of the observed values will be within one standard deviation unit of the mean. That is, approximately 68% of the observed values will have a Z–score that is between -1 and 1.
ii) Approximately 95% of the observed values will be within two standard deviation units of the mean. That is, approximately 95% of the observed values will have a Z–score that is between -2 and 2.
iii) Approximately 99.7% of the observed values will be within three standard deviation units of the mean. That is, approximately 99.7% of the observed values will have a Z–score that is between -3 and 3. Notice that this indicates that almost all of the observed values will be within three standard deviations of the mean.

When it is applicable, the $68\% - 95\% - 99.7\%$ rule, can be used to determine the relative position of a particular value of a variable based on the corresponding Z–score. Notice that this rule indicates that a fairly large proportion (68%) of the sample will lie within one standard deviation of the mean; a very large proportion (95%) of the sample will lie within two standard deviations of the mean; and, almost all (99.7%) of the sample will lie within three standard deviations of the mean.

An aside – Chebyshev's rule

Another connection between Z–scores and percentages is provided by Chebyshev's rule. Chebyshev's rule is a mathematical fact that is true for any distribution. Unfortunately, the universal applicability of Chebyshev's rule forces its conclusions to be of more theoretical than practical interest. That is, the conclusions of Chebyshev's rule are valid for any distribution; but, they are often so imprecise that they are of limited practical use.

Chebyshev's rule. *For any distribution:*
i) At least 75% of the observed values will be within two standard deviation units of the mean. That is, at least 75% of the observed values will have a Z–score that is between -2 and 2.
ii) At least 89% of the observed values will be within three standard deviation units of the mean. That is, at least 89% of the observed values will have a Z–score that is between -3 and 3.
iii) In general, given a number $k > 1$, at least $[1 - (1/k^2)]100\%$ of the observed values will be within k standard deviation units of the mean, i.e., at least this percentage of the observed values will have a Z–score that is between -k and k.

---

3 Probability toc

3.1 The setting toc

Probability theory is used to model the behavior of random experiments. In this context a **random experiment** is any process of observation or experimentation for which the particular outcome is not known with certainty in advance of performance of the experiment. For example we might consider the experiment of tossing a coin or a die, tossing a pair of dice, dealing a hand of cards from a deck of cards, drawing balls from a box of balls, selecting a person from a population and measuring the person's height, age, or weight, or testing a machine to determine if it works properly.

Our goal is to formulate a theory which can be used to specify a formal model for the randomness in the outcomes of the experiment by assigning probabilities to events associated with the experiment. An **event** is a description of the outcome of the experiment. It is useful to distinguish between simple events and compound events. A **simple event** is an event which cannot be decomposed into simpler events. We will refer to simple events as **elementary outcomes**. A **compound event** is an event which can be decomposed into two or more events. For example, if we toss a die once, then the elementary outcomes, "observe a 1", "observe a 2", *etc.*, can be represented by the integers 1, 2, 3, 4, 5, and 6. The compound event "observe an even number" is the collection (set) $\{2, 4, 6\}$ of the three elementary outcomes corresponding to even numbers.

The first step in forming a probability model for a particular experiment is the specification of a **sample space** containing all the possible elementary outcomes of the experiment. Relevant events can then be viewed as subsets of the sample space. The elementary outcomes (simple events) are the singleton sets (sets containing a single elementary outcome) and compound events are sets containing two or more elementary outcomes.

3.2 Some illustrative examples toc

Example 3.1 Tossing a die. As noted above, we can represent the 6 elementary outcomes of a single toss of a die by the integers 1, 2, 3, 4, 5, and 6. Now suppose that a die is tossed twice. The 36 elementary outcomes for this experiment can be represented by ordered pairs of the integers $1, \ldots, 6$. For example the ordered pair $(1, 3)$ indicates that the first toss yielded a 1 and the second yielded a 3. The sample space is the set containing the 36 elementary outcomes (ordered pairs) shown in Table 3.1.

Table 3.1 Elementary outcomes (ordered pairs) when a die is tossed twice.

	second toss					
first toss	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Two events of potential interest are:

1. The event that the sum of the two numbers observed is 7. The ordered pairs on the upper right to lower left diagonal of the table are favorable for this event. Thus the event that the sum is 7 is represented by the set

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

Read this as: the set containing the ordered pairs $(1,6),(2,5),(3,4),(4,3),(5,2)$*, and* $(6,1)$*.*

2. The event that the number on the first toss is even. The ordered pairs in the second, fourth, and sixth rows of the table are favorable to this event. Thus the event that the number on the first toss is even is represented by the set

$$\left\{\begin{array}{l}(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),\\(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),\\(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\end{array}\right\}.$$

Example 3.2 Tossing a coin. If we toss a coin once, then, letting H denote "heads" and T denote "tails", the sample space is the set containing the 2 elementary outcomes $\{H,T\}$. Letting H and T denote the values of a dichotomy, such as the sex of a person (male or female), the quality of an item (acceptable or unacceptable), or the outcome of a medical procedure (successful or not), we can use the H and T notation of this and the three toss example below to represent a wide variety of experiments.

If we toss a coin three times, then, letting H denote "heads" and T denote "tails", we can use an ordered triple (with the order indicating the outcomes of the first, second, and

third tosses) to represent an elementary outcome. The sample space is the set containing the 8 elementary outcomes (ordered triples):

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

Two events of potential interest are:
1. The event that there are exactly two heads. The ordered triples with two $H's$ and one T are favorable to this event. Thus the event is represented by the set

$$\{HHT, HTH, THH\}.$$

2. The event that there are at least two heads. The ordered triples with two $H's$ and one T or three $H's$ are favorable to this event. Thus the event is represented by the set

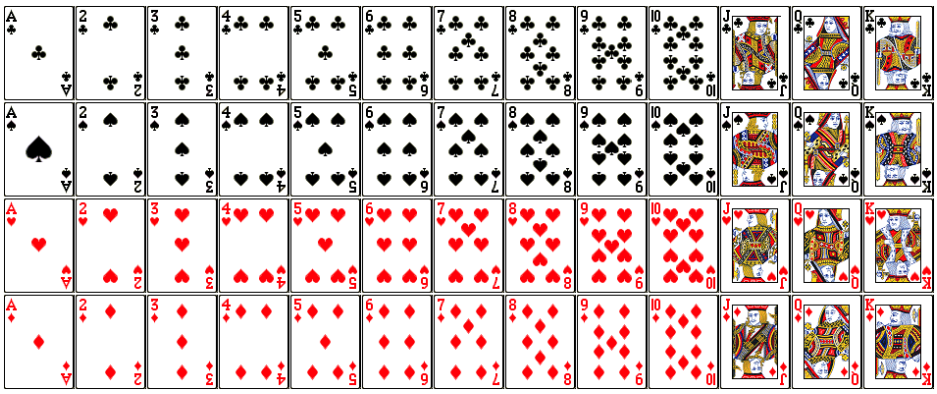
$$\{HHH, HHT, HTH, THH\}.$$

An aside – playing cards, bridge, and poker

A standard deck of playing cards, as shown in Figure 3.1, contains 52 cards. The cards appear in four suits, hearts ♡, diamonds ♢, clubs ♣, and spades ♠, of these the heart and diamond cards are colored red and the club and spade cards are colored black. The 13 cards of each suit have face values (ranks) of $2, 3, \ldots, 10$, jack, queen, king, and ace. An ace is often treated as having a face value (rank) of one. Cards with the same face value are said to be of the same kind. The jacks, queens, and kings, which typically (as with the deck in Figure 3.1) have "faces" on them are called face cards.

Bridge is a four–player partnership game. The four players in bridge are known as North, East, South, and West. North and South form one team and East and West form the second team. All 52 cards are dealt with each player receiving 13 cards. Thus, for our purposes, bridge means dividing the 52 cards into four hands (sets) of 13 cards and a bridge hand is a collection of 13 cards. There are many variations of poker. For our purposes, a poker hand is a collection of five cards.

Figure 3.1 A standard deck of 52 playing cards



Example 3.3 Bridge or poker. If we assign the numbers from 1 to 52 as labels for the 52 cards in the deck (for example we might label the cards according to their positions in Figure 3.1), then we can use an ordered arrangement of the numbers from 1 to 52 to represent a particular deal of cards for a bridge game with the first 13 numbers corresponding to North, the next 13 to East, and so on. If we are only concerned with characteristics of one hand we could use a collection of 13 different numbers selected from the labels 1 to 52. Similarly, we could use a collection (set) of 5 different numbers to represent a poker hand.

Consider the event that a poker hand contains exactly two aces. The sets of five cards containing two aces and three non–aces are favorable for this event. There are too many such hands to list here. In case you wondered, using notation defined in Section 5, there are $\binom{4}{2}\binom{48}{3} = 103,776$ hands containing exactly two aces.

Example 3.4 Tossing a coin or die repeatedly. Suppose that a coin is tossed repeatedly until a head occurs. In this application we can represent an elementary outcome by a sequence terminating with an H and with enough $T's$ to indicate the outcome. For example; H indicates that we got heads on the first toss, TH indicates that we first got heads on the second toss, TTH indicates that we first got heads on the third toss, and so on. Thus the sample space is the countably infinite set

$$\{H, TH, TTH, TTTH, \ldots\}.$$

As noted earlier, we can use the H and T notation of this example to represent a wide variety of experiments. Three events of potential interest are:

1. The event that exactly three tosses are required to get a head. Since TTH is the only outcome favorable for this event, it is represented by the set

$$\{TTH\}.$$

2. The event that at most three tosses are required to get a head. The three outcomes favorable for this event, heads first, second, or third, are the elements of the set

$$\{H, TH, TTH\}.$$

3. The event that more than three tosses are required to get a head. The representation of this event is provided by the infinite set

$$\{TTTH, TTTTH, TTTTTH, \ldots\}$$

containing all sequences which consist of a sequence of three or more $T's$ followed by a single H.

If a die is tossed repeatedly until a 1 (an ace) occurs, then this same representation can be used with H indicating that a 1 occurs and T indicating that a 2, 3, 4, 5, or 6 occurs.

Similarly, if parts selected from a production line are tested sequentially until a defective part is found, then this same representation can be used with H indicating that the part is defective and T indicating that the part is acceptable.

Example 3.5 Sampling for a numerical value. Suppose that we are interested in the distribution of weights among individuals in a particular population. Further suppose that we choose one individual from this population and determine the weight of this individual. We can use a positive real number x to represent an elementary outcome (the weight of the individual). If we know the weights of all of the individuals in the population we can use these weights as the sample space. Since it is unlikely that we would know the weights of all the individuals, it is more convenient to use an interval of values on the number line as our sample space. If we know that no one in the population weighs more than 300 pounds, then we can use the interval from zero to 300 ($(0, 300]$ in interval notation) as our sample space. If we do not want to specify a maximal weight, we can use the positive part of the number line ($(0, \infty)$ in interval notation) to represent the sample space. Events can be represented by appropriate intervals. For example, the event "the individual weighs

between 100 and 150 pounds" corresponds to the interval $[100, 150]$ and the event "the individual weighs at least 100 pounds" corresponds to the interval $[100, \infty)$ (or $[100, 300)$ with the sample space $(0, 300)$).

3.3 Sample spaces and events toc

The examples given above illustrate the process of defining an elementary outcome, a sample space, and events for several specific experiments. Recall that for our purposes an experiment is a process of observation or experimentation which results in one of several possible elementary outcomes. We will now begin a more formal treatment of these aspects of an experiment. We will assume that all of the possible elementary outcomes of the experiment are known in advance but the actual outcome of the experiment will not be known with certainty until after the experiment is performed.

The **sample space** Ω of a particular experiment is the collection of all possible elementary outcomes for the experiment. Recall that an elementary outcome or simple event is an event which cannot be decomposed into simpler events. We will assume that these elementary outcomes are mutually exclusive so that two distinct elementary outcomes cannot occur at the same time. A generic elementary outcome will be denoted ω. In set terminology the objects which form a set are called elements, thus, we may refer to an elementary outcome as an element of the sample space or of an event. (Ω and ω are the upper and lower case versions of the Greek letter omega.)

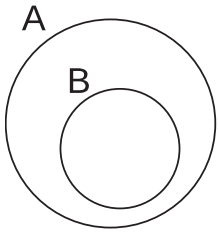
An **event** A is a collection of elementary outcomes, *i.e.*, a subset of Ω. If the experiment is conducted and the **elementary outcome** ω occurs, then if ω is an element of A (in symbols, $\omega \in A$) we say that event A has occurred. On the other hand, if ω is not an element of A (in symbols, $\omega \notin A$) we say that A has not occurred . For example, in the die tossing example, if we observe $\omega = 2$, then $2 \in \{2, 4, 6\}$ and the event *observe an even number* occurred. But, $2 \notin \{1, 3, 5\}$ so the event *observe an odd number* did not occur.

For convenience we define the **null event** (empty set), $\emptyset$, as the event with no elements. Viewed as an event, the sample space Ω is the **sure event**, since it contains all possible outcomes and therefore must occur.

Given two events A and B we write $B \subset A$ (B is contained in A or B is a **subset** of A) when every element of B is also an element of A. More formally, $B \subset A$ means that if $\omega \in B$, then $\omega \in A$. This is illustrated in Figure 3.2. Notice that if $B \subset A$, then the

occurrence of B implies the occurrence of A. In particular, if B is a proper subset of A, that is, if A contains some elements that are not in B, then the event B is a "special case" of the event A. Also note that for every event A, $A \subset \Omega$ (A is a subset of the sample space Ω), $A \subset A$ (A is a subset of itself), and $\emptyset \subset A$ (the empty set is a subset of A).

Figure 3.2 An event and a subevent.
In this Venn diagram event B is a subevent (subset) of event A.

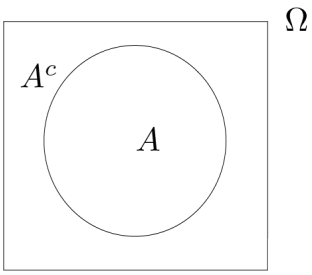


Example 3.6a Tossing a die once. For one toss of a die we have $\Omega = \{1, 2, 3, 4, 5, 6\}$. Consider the following events:
A: "the number observed is 4 or less" $A = \{1, 2, 3, 4\}$
B: "the number observed is 1, 3, or 4" $B = \{1, 3, 4\}$
C: "the number observed is 5 or 6" $C = \{5, 6\}$
D: "the number observed is 3 or more" $D = \{3, 4, 5, 6\}$.
For these events, $B \subset A$, since observing a 1, 3, or 4 implies that we observed a number which is 4 or less, and $C \subset D$ since observing 5 or 6 implies that we we observed a number which is 3 or larger.

Example 3.6b Tossing a coin three times. For three tosses of a coin we have
$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$. Consider the following events:
A: "the first toss yields a head" $A = \{HHH, HHT, HTH, HTT\}$
B: "the first two tosses yield heads" $B = \{HHH, HHT\}$
C: "there are exactly two heads" $C = \{HHT, HTH, THH\}$
D: "there are at least two heads" $D = \{HHH, HHT, HTH, THH\}$.
For these events, $B \subset A$, since observing heads on the first two tosses is a special case of observing heads on the first toss, $B \subset D$, since observing heads on the first two tosses is a special case observing at least two heads, and $C \subset D$ since observing exactly two heads is a special case of observing at least two heads.

Given an event A the complementary event A^c (the **complement** of A) is the event containing all elements of Ω that do not belong to A. This is illustrated in Figure 3.3. In symbols, $A^c = \{\omega \in \Omega : \omega \notin A\}$ *Read this as the set of* ω *in* Ω *such that* ω *is not in* A. In words, A^c (A complement) is the event that A does not occur.

Figure 3.3 An event A and its complement A^c.
In this Venn diagram, the rectangle represents the sample space Ω, the interior of the circle the event A, and the region outside the circle A^c, the complement of A.



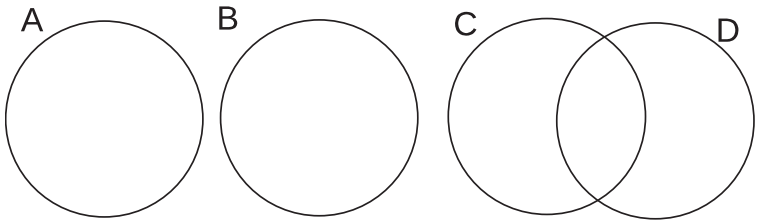
Example 3.6a Tossing a die once (continued). The complements of the events $A = \{1, 2, 3, 4\}$, $B = \{1, 3, 4\}$, $C = \{5, 6\}$, and $D = \{3, 4, 5, 6\}$ are: $A^c = C = \{5, 6\}$, $B^c = \{2, 5, 6\}$, $C^c = A$, and $D^c = \{1, 2\}$

Example 3.6b Tossing a coin three times (continued). The complements of the events $A = \{HHH, HHT, HTH, HTT\}$, $B = \{HHH, HHT\}$, $C = \{HHT, HTH, THH\}$, and $D = \{HHH, HHT, HTH, THH\}$ are:
$A^c = \{THH, THT, TTH, TTT\}$ "the first toss yields tails"
$B^c = \{HTH, HTT, THH, THT, TTH, TTT\}$ "at least one of the first two tosses yields tails"
$C^c = \{HHH, HTT, THT, TTH, TTT\}$ "there are zero, one, or three heads"
$D^c = \{HTT, THT, TTH, TTT\}$ "there is at most one head".

Note that, for any event A, events A and A^c are mutually exclusive in the sense that they cannot both occur at the same time. Furthermore, since they are complementary, exactly one of A and A^c must occur. Note also that $\Omega^c = \emptyset$ (the complement of everything is nothing) and $\emptyset^c = \Omega$ (the complement of nothing is everything).

In general, two events are said to be **mutually exclusive** or **disjoint** if they do not have any elements in common. In other words, two events are mutually exclusive (disjoint) if they cannot occur at the same time. This is illustrated in Figure 3.4.

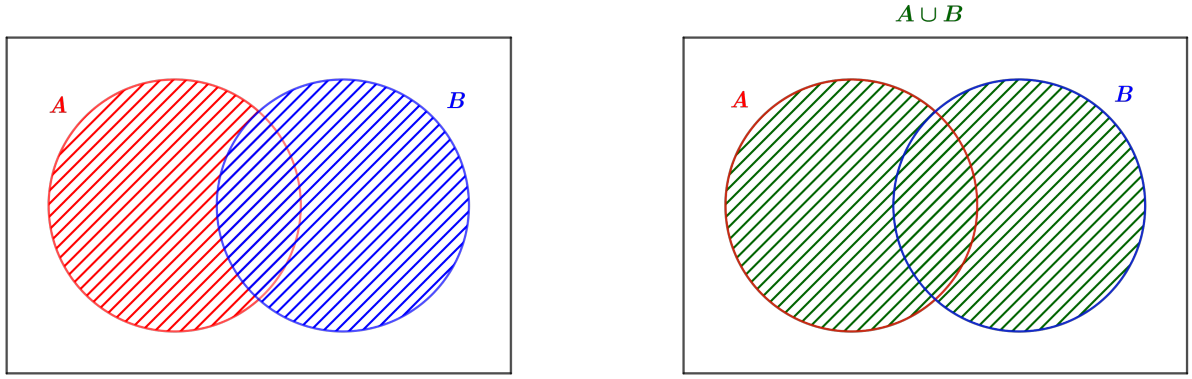
Figure 3.4 Mutually exclusive (disjoint) events.
In this Venn diagram events A and B are mutually exclusive or disjoint while events C and D are not mutually exclusive.



Example 3.6a Tossing a die once (continued). There are two pairing of the events $A = \{1, 2, 3, 4\}$, $B = \{1, 3, 4\}$, $C = \{5, 6\}$, and $D = \{3, 4, 5, 6\}$ which give mutually exclusive events. A and C are mutually exclusive and B and C are mutually exclusive.

Example 3.6b Tossing a coin three times (continued). Each of the events $A = \{HHH, HHT, HTH, HTT\}$, $B = \{HHH, HHT\}$, $C = \{HHT, HTH, THH\}$, and $D = \{HHH, HHT, HTH, THH\}$ contains HHT; thus, no pairing of these events gives mutually exclusive events. On the other hand, the event $E = \{TTT, TTH, THT, HTT\}$ (observe at least two tails) shares no elements with B, C, or D; thus, E and B are mutually exclusive, E and C are mutually exclusive, and E and D are mutually exclusive.

Figure 3.5 The union of A and B.



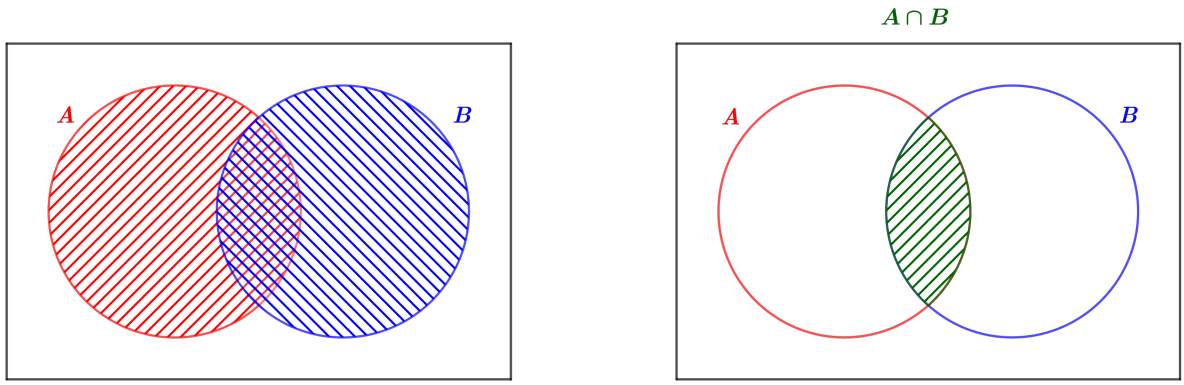
The **union** of events A and B, denoted $A \cup B$, is the collection of elementary outcomes which belong to A or B. The *or* in this statement is the *logical or* meaning one or the other or both. That is, when we say A or B we mean A alone or B alone or both A and B. This is illustrated in Figure 3.5.
In symbols, $A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}$
In words, $A \cup B$ is the event "A or B" in the sense that A or B or both A and B occur.

Example 3.6a Tossing a die once (continued). Recall that $A = \{1, 2, 3, 4\}$,
$B = \{1, 3, 4\}$, $C = \{5, 6\}$, and $D = \{3, 4, 5, 6\}$ The unions of these events are:
$A \cup B = \{1, 2, 3, 4\} \cup \{1, 3, 4\} = \{1, 2, 3, 4\} = A$ (B is a subset of A)
$A \cup C = \{1, 2, 3, 4\} \cup \{5, 6\} = \{1, 2, 3, 4, 5, 6\}$
$A \cup D = \{1, 2, 3, 4\} \cup \{3, 4, 5, 6\} = \{1, 2, 3, 4, 5, 6\}$
$B \cup C = \{1, 3, 4\} \cup \{5, 6\} = \{1, 3, 4, 5, 6\}$
$B \cup D = \{1, 3, 4\} \cup \{3, 4, 5, 6\} = \{1, 3, 4, 5, 6\}$
$C \cup D = \{5, 6\} \cup \{3, 4, 5, 6\} = \{3, 4, 5, 6\} = D$ (C is a subset of D)

Example 3.6b Tossing a coin three times (continued). Recall that
$A = \{HHH, HHT, HTH, HTT\}$, $B = \{HHH, HHT\}$, $C = \{HHT, HTH, THH\}$,
$D = \{HHH, HHT, HTH, THH\}$ The unions of these events are:
$A \cup B = \{HHH, HHT, HTH, HTT\} \cup \{HHH, HHT\} = A$ (B is a subset of A)
$A \cup C = \{HHH, HHT, HTH, HTT\} \cup \{HHT, HTH, THH\} = A$ (C is a subset of A)
$A \cup D = \{HHH, HHT, HTH, HTT\} \cup \{HHH, HHT, HTH, THH\}$
$= \{HHH, HHT, HTH, HTT, THH\}$
$B \cup C = \{HHH, HHT\} \cup \{HHT, HTH, THH\} = \{HHH, HHT, HTH, THH\}$
$B \cup D = \{HHH, HHT\} \cup \{HHH, HHT, HTH, THH\} = D$ (B is a subset of D)
$C \cup D = \{HHT, HTH, THH\} \cup \{HHH, HHT, HTH, THH\} = D$ (C is a subset of D)

The **intersection** of events A and B, denoted $A \cap B$, is the collection of elementary outcomes which belong to both A and B. This is illustrated in Figure 3.6.
In symbols, $A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$
In words, $A \cap B$ is the event "A and B" in the sense that both A and B occur.

Figure 3.6 The intersection of A and B.



Example 3.6a Tossing a die once (continued). Recall that $A = \{1, 2, 3, 4\}$, $B = \{1, 3, 4\}$, $C = \{5, 6\}$, and $D = \{3, 4, 5, 6\}$. The intersections of these events are:
$A \cap B = \{1, 2, 3, 4\} \cap \{1, 3, 4\} = \{1, 3, 4\} = B$ (B is a subset of A)
$A \cap C = \{1, 2, 3, 4\} \cap \{5, 6\} = \emptyset$ (A and C are mutually exclusive (disjoint)).
$A \cap D = \{1, 2, 3, 4\} \cap \{3, 4, 5, 6\} = \{3, 4\}$
$B \cap C = \{1, 3, 4\} \cap \{5, 6\} = \emptyset$ (B and C are mutually exclusive (disjoint)).
$B \cap D = \{1, 3, 4\} \cap \{3, 4, 5, 6\} = \{3, 4\}$
$C \cap D = \{5, 6\} \cap \{3, 4, 5, 6\} = \{5, 6\} = C$ (C is a subset of D)

Example 3.6b Tossing a coin three times (continued). Recall that
$A = \{HHH, HHT, HTH, HTT\}$, $B = \{HHH, HHT\}$, $C = \{HHT, HTH, THH\}$,
$D = \{HHH, HHT, HTH, THH\}$ The intersections of these events are:
$A \cap B = \{HHH, HHT, HTH, HTT\} \cap \{HHH, HHT\} = B$ (B is a subset of A)
$A \cap C = \{HHH, HHT, HTH, HTT\} \cap \{HHT, HTH, THH\} = C$ (C is a subset of A)
$A \cap D = \{HHH, HHT, HTH, HTT\} \cap \{HHH, HHT, HTH, THH\} = \{HHH, HHT, HTH\}$
$B \cap C = \{HHH, HHT\} \cap \{HHT, HTH, THH\} = \{HHT\}$
$B \cap D = \{HHH, HHT\} \cap \{HHH, HHT, HTH, THH\} = B$ (B is a subset of D)
$C \cap D = \{HHT, HTH, THH\} \cap \{HHH, HHT, HTH, THH\} = C$ (C is a subset of D)

Substantial simplifications of probability calculations are often possible using the representations of the complement of a union and the complement of an intersection contained in DeMorgan's laws.

DeMorgan's laws. *For any pair of events A and B:*
(1) The complement of the union of the events is the intersection of their complements. This is illustrated in Figure 3.7. In symbols, $(A \cup B)^c = A^c \cap B^c$
(2) The complement of the intersection of the events is the union of their complements. In symbols, $(A \cap B)^c = A^c \cup B^c$.

Example 3.6 Tossing a die once (continued). Recall that $\Omega = \{1,2,3,4,5,6\}$, $A = \{1,2,3,4\}$, $B = \{1,3,4\}$, $C = \{5,6\}$, and $D = \{3,4,5,6\}$. We will use A and B to demonstrate DeMorgan's laws. The complements are $A^c = \{5,6\}$, $B^c = \{2,5,6\}$.
First consider $A \cup B = \{1,2,3,4\}$. For these events the complement of the union is $(A \cup B)^c = \{1,2,3,4\}^c = \{5,6\}$ and the intersection of the complements is $A^c \cap B^c = \{5,6\} \cap \{2,5,6\} = \{5,6\}$.
Next consider $A \cap B = \{1,3,4\}$. For these events the complement of the intersection is $(A \cap B)^c = \{1,3,4\}^c = \{2,5,6\}$ and the union of the complements is $A^c \cup B^c = \{5,6\} \cup \{2,5,6\} = \{2,5,6\}$.

Figure 3.7a Illustration of DeMorgan's law $(A \cup B)^c = A^c \cap B^c$.
From the shading it is clear that $(A \cup B)^c$ is the intersection of A^c and B^c.

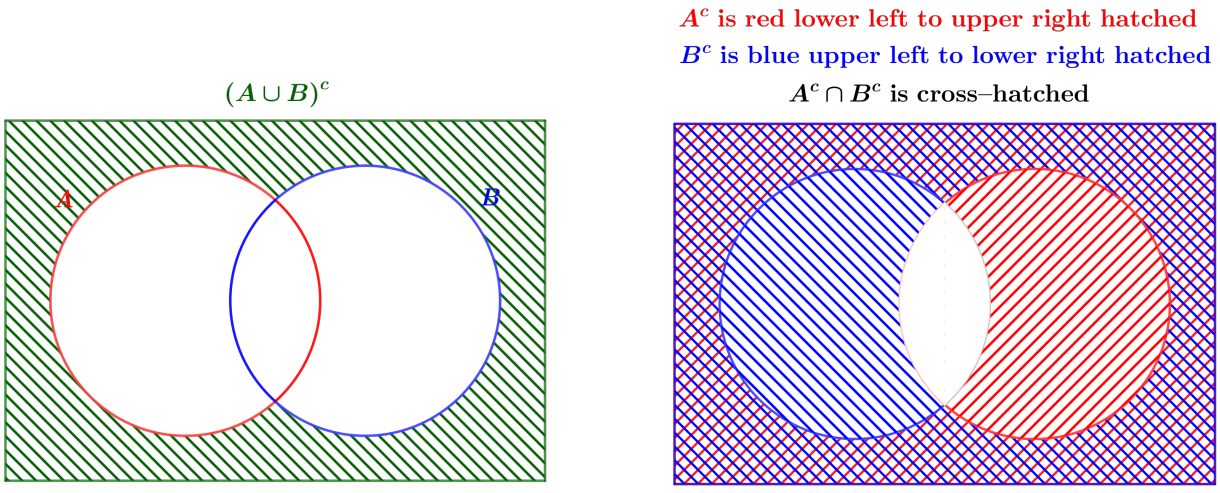
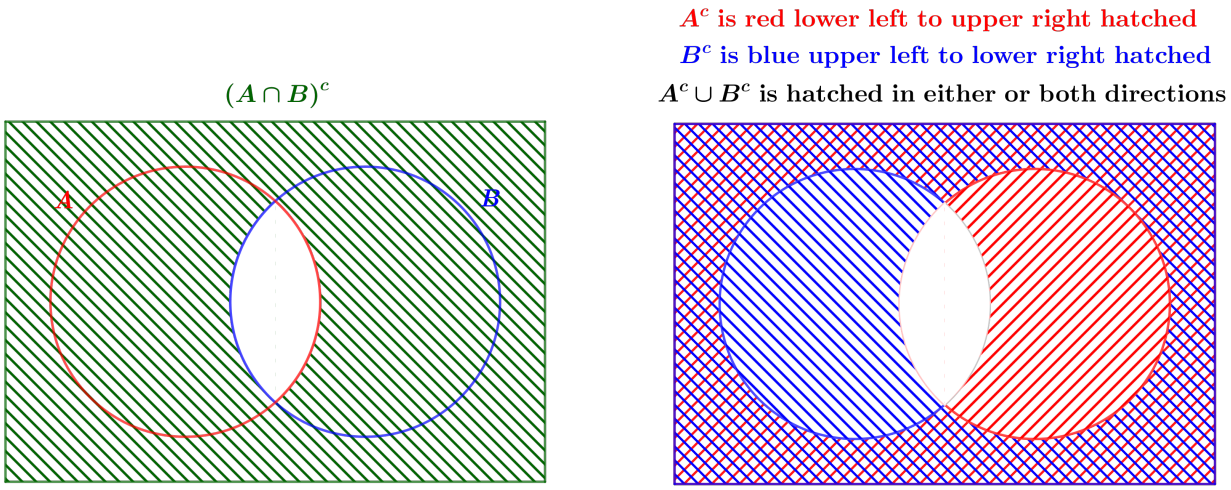


Figure 3.7b Illustration of DeMorgan's law $(A \cap B)^c = A^c \cup B^c$.
From the shading it is clear that $(A \cap B)^c$ is the union of A^c and B^c.



Example 3.6b Tossing a coin three times (continued). We will use A and D to demonstrate DeMorgan's laws. Recall that $A = \{HHH, HHT, HTH, HTT\}$
and $D = \{HHH, HHT, HTH, THH\}$.
First consider $A \cup D = \{HHH, HHT, HTH, HTT, THH\}$ "the first toss yields a head OR there are at least two heads". The complement of this union is $(A \cup D)^c = \{THT, TTH, TTT\}$, that is, "the first toss does not yield a head AND there are less than two heads". Notice that this is the intersection of the complements $A^c = \{THH, THT, TTH, TTT\}$ "the first toss yields tails (not heads)" and $D^c = \{HTT, THT, TTH, TTT\}$: "there are less than two heads" Thus, $(A \cup D)^c = A^c \cap D^c$.
Next consider $A \cap D = \{HHH, HHT, HTH\}$ "the first toss yields a head AND there are at least two heads". The complement of this intersection is
$(A \cap D)^c = \{HTT, THH, THT, TTH, TTT\}$ "the first toss does not yield a head OR there are less than two heads". Notice that this is the union of the complements $A^c = \{THH, THT, TTH, TTT\}$ "the first toss yields tails (not heads)"
and $D^c = \{HTT, THT, TTH, TTT\}$: "there are less than two heads" Thus, $(A \cap D)^c = A^c \cup D^c$.

Some properties of set operations

Some basic properties of set operations are summarized here for completeness and ease of reference.

Commutative properties. *For any events A and B*
(a) $A \cup B = B \cup A$
(b) $A \cap B = B \cap A$

Associative properties. *For any events A, B, and C*
(a) $(A \cup B) \cup C = A \cup (B \cup C)$ *(thus the notation $A \cup B \cup C$ is unambiguous)*
(b) $(A \cap B) \cap C = A \cap (B \cap C)$ *(thus the notation $A \cap B \cap C$ is unambiguous)*

Distributive properties. *For any events A, B, and C*
(a) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
(b) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

Additional properties.
For any event A: $A \cap A = A,\ A \cup A = A,\ A \cap \Omega = A,\ A \cup \Omega = \Omega,\ A \cap \emptyset = \emptyset,\ A \cup \emptyset = A,$ $(A^c)^c = A,\ A \cap A^c = \emptyset,$ *and* $A \cup A^c = \Omega$.
If $A \subset B$, then $A \cap B = A$ and $A \cup B = B$.

3.4 Partitioning an event toc

Given two events, we will often find it useful to use one of the events to decompose the other event into disjoint subevents. A decomposition of an event into disjoint subevents yields a **partition** of the event. Given events A and B, we can partition (decompose) A into the subevents $A \cap B$ and $A \cap B^c$. These two subevents form a partition of A, since they are disjoint (they cannot share any outcomes since $A \cap B$ is a subset of B while $A \cap B^c$ is a subset of B^c) and their union is A (each outcome in A is either an element of B or an element of its complement B^c).

In symbols, $A = (A \cap B) \cup (A \cap B^c)$ and $(A \cap B) \cap (A \cap B^c) = \emptyset$.

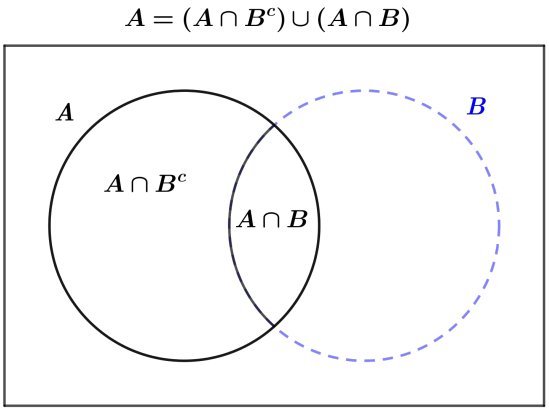
In words, there are two mutually exclusive ways in which the event A can occur:

(1) Event A and event B both occur ($A \cap B$ occurs); or,

(2) Event A occurs but event B does not occur ($A \cap B^c$ occurs).

This basic decomposition of A is illustrated in Figure 3.8. Figures 3.9 and 3.10 illustrate how this relationship can be used to partition a union of two or three events.

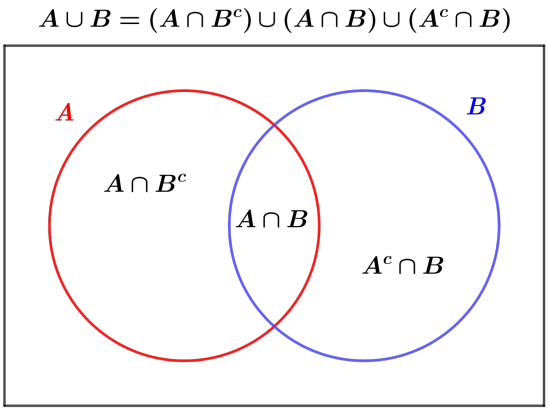
Figure 3.8 Use of event B to decompose event A into two disjoint parts.



Example 3.7 Characteristics of a person. Consider the experiment of selecting a person from a population of adults. Let A denote the event that the person selected is a college student and let B denote the event that the person is 20 years old. The decomposition $A = (A \cap B) \cup (A \cap B^c)$ indicates that a college student is either 20 years old, event $A \cap B$, or some other age, event $A \cap B^c$. The fact that this is a partition of A simply indicates that there is no other option, *i.e.* the college student is either 20 or not.

The Venn diagram of Figure 3.9 shows a partition (decomposition) of $A \cup B$ into three disjoint parts. For this partition we have $A \cup B = (A \cap B^c) \cup (A \cap B) \cup (A^c \cap B)$. This shows the three mutually exclusive ways in which the event $A \cup B$ can occur:
(1) (event $A \cap B^c$) A occurs and B does not occur;
(2) (event $A \cap B$) both A and B occur; or,
(3) (event $A^c \cap B$) B occurs and A does not occur.

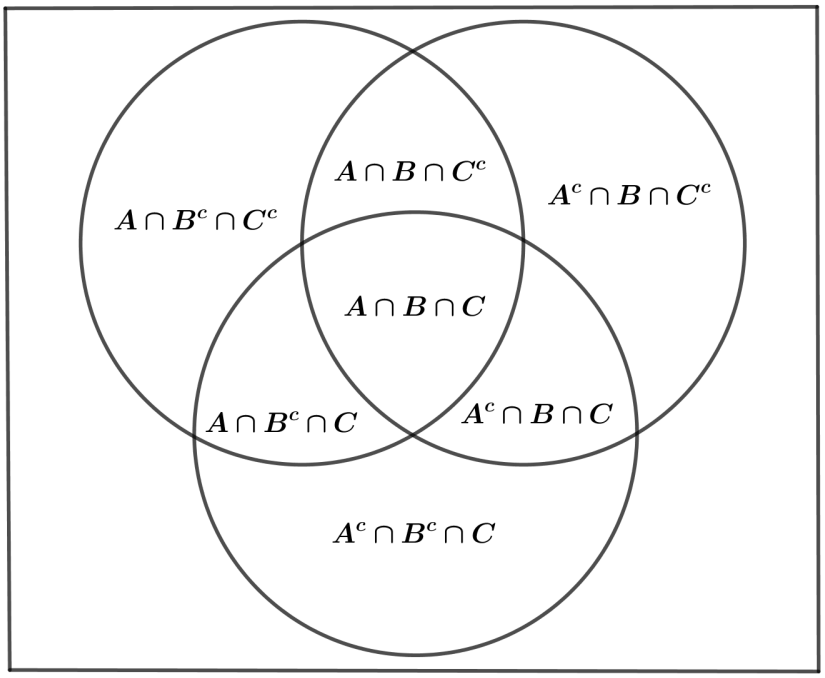
Figure 3.9 Decomposition of $A \cup B$ into three disjoint parts.



Example 3.7 Characteristics of a person, revisited. As before, consider the experiment of selecting a person from a population of adults. Let A denote the event that the person selected is a college student and let B denote the event that the person is 20 years old. The union $A \cup B$ is the event that the person selected is a college student or age 20 or both. This union is partitioned into the three disjoint events: event $A \cap B^c$ the person is a college student but is not 20 years old, event $A \cap B$ the person is a college student and is also 20 years old, and event $A^c \cap B$ the person is 20 years old but is not a college student.

Figure 3.10 Decomposition of $A \cup B \cup C$ into seven disjoint parts.

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



The Venn diagram of Figure 3.10 shows a partition (decomposition) of $A \cup B \cup C$ into seven disjoint parts. The seven mutually exclusive ways in which the event $A \cup B \cup C$ can occur are:

(1) (event $A \cap B^c \cap C^c$) A occurs but B and C do not occur;
(2) (event $A \cap B \cap C^c$) A and B occur but C does not occur;
(3) (event $A^c \cap B \cap C^c$) B occurs but A and C do not occur;
(4) (event $A \cap B \cap C$) A and B and C all occur;
(5) (event $A \cap B^c \cap C$) A and C occur but B does not occur;
(6) (event $A^c \cap B \cap C$) B and C occur but A does not occur; or,
(7) (event $A^c \cap B^c \cap C$) C occurs but A and B do not occur.

Example 3.7 Characteristics of a person, revisited. As before, consider the experiment of selecting a person from a population of adults. Let A denote the event that the person selected is a college student, let B denote the event that the person is 20 years old, and let C denote the event that the person owns a car. The union $A \cup B \cup C$ is the event that the person selected is a college student or age 20 or owns a car, *i.e.*, the person possesses

at least one of these three characteristics. This union is partitioned into the seven disjoint events:

(1) (event $A \cap B^c \cap C^c$) is a college student, is not age 20, and does not own a car;

(2) (event $A \cap B \cap C^c$) is a college student, is age 20, and does not own a car;

(3) (event $A^c \cap B \cap C^c$) is not a college student, is age 20, and does not own a car;

(4) (event $A \cap B \cap C$) is a college student, is age 20, and owns a car;

(5) (event $A \cap B^c \cap C$) is a college student, is not age 20, and owns a car;

(6) (event $A^c \cap B \cap C$) is not a college student, is age 20, and owns a car;

(7) (event $A^c \cap B^c \cap C$) is not a college student, is not age 20, and owns a car;

An aside – notation for more complicated unions and intersections

On occasion it is useful to have a compact notation for the union or intersection of a collection of events.

Given a set of n events $\{A_1, \ldots, A_n\}$:

$$\bigcup_{i=1}^{n} A_i = A_1 \cup A_2 \cup \cdots \cup A_n \text{ denotes the union of these events}$$

$$\bigcap_{i=1}^{n} A_i = A_1 \cap A_2 \cap \cdots \cap A_n \text{ denotes the intersection of these events}$$

Given an infinite sequence of events $\{A_1, A_2, \ldots\}$:

$$\bigcup_{i=1}^{\infty} A_i = A_1 \cup A_2 \cup \cdots \text{ denotes the union of this sequence of events}$$

$$\bigcap_{i=1}^{\infty} A_i = A_1 \cap A_2 \cap \cdots \text{ denotes the intersection of this sequence of events}$$

---

4 Probability measure toc

4.1 Definition of a probability measure toc

Given an experiment and an event A we need to associate a probability $P(A)$ with the event. The formal (axiomatic) definition of a **probability measure** below indicates the restrictions we will impose on any such assignment of probabilities to events.

Definition. *A probability measure P is a function which assigns probabilities to events (subsets of Ω) and satisfies the following axioms.*
Axiom 1: For every event A, $0 \leq P(A) \leq 1$.
Axiom 2: $P(\Omega) = 1$.
Axiom 3: For every finite collection of mutually exclusive events $A_1, A_2, \ldots, A_n$,
$P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n)$.
Furthermore, for every infinite sequence of mutually exclusive events $\{A_1, A_2, \ldots\}$,
$P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$, *more formally* $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

The meaning and desirability of these axioms is straightforward. The first two axioms place some basic restrictions on the possible values of probabilities. Axiom 1 simply requires that the probability of an event must be a number between zero and one. (If we think of probabilities as percentages, this says that a probability must be between zero and 100 percent.) By assigning the value 1 (100%) to the sample space, Axiom 2 simply requires that something must happen when the experiment is conducted. The third axiom requires a certain type of consistency in the assignment of probabilities. Recall that when events are mutually exclusive (disjoint), they cannot occur at the same time. Axiom 3 states that if an event can be partitioned (decomposed) into a collection of mutually exclusive subevents (the A_i), then the probability of the event must be equal to the sum of the probabilities of the mutually exclusive subevents of the partition.

4.2 Properties of probability measures toc

In this section we will provide some basic properties of probability measures.

The null event has probability zero

As noted above, Axiom 2, $P(\Omega) = 1$, indicates that something must occur when the experiment is conducted. In other words, the sure event Ω must occur. As you would

expect, this implies that the null event $\emptyset$, will not occur. Formally this is indicated by saying that the null event has probability zero. That is,

$$P(\emptyset) = 0.$$

(Technical remark: This result follows from the infinite sequence part of axiom 3.)

The probability of the complement of an event

For any event A, the sample space can be partitioned as $\Omega = A \cup A^c$. Thus, by axioms 2 and 3 we see that for any event A, $P(\Omega) = P(A) + P(A^c) = 1$. It follows that the probability of the complement of an event is one minus the probability of the event. In symbols,

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

An important decomposition of the probability of an event

In Figure 3.8 we illustrated the use of one event (event B) to partition another event (event A), *viz*, $A = (A \cap B) \cup (A \cap B^c)$. This partition and axiom 3 imply that for any events A and B,

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

Probabilities of nested events

Recall that if $A \subset B$, then the occurrence of A implies the occurrence of B. Therefore, if A is a subset of B, then the probability of B cannot be less than the probability of A. More formally, if $A \subset B$, then $A \cap B = A$ and the decomposition we just discussed becomes $P(B) = P(A \cap B) + P(A^c \cap B) = P(A) + P(A^c \cap B)$. Since $P(A^c \cap B) \geq 0$, it follows that

$$\text{if } A \subset B, \text{ then } P(A) \leq P(B).$$

We will now provide some important expressions for **the probability of a union** in terms of the probabilities of certain subevents.

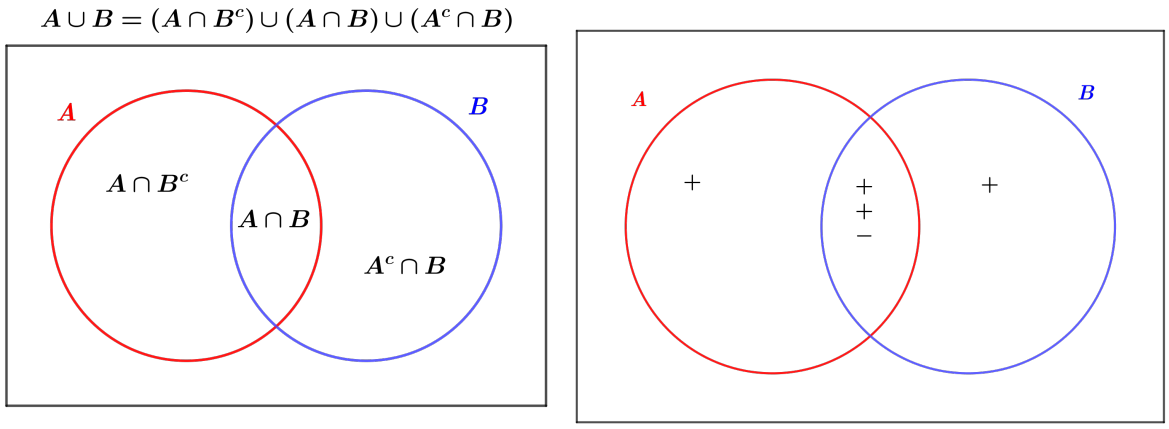
The probability of the union of 2 events

In Figure 3.9 we illustrated the partition of $A \cup B$ into three disjoint parts, *viz*, $A \cup B = (A \cap B^c) \cup (A \cap B) \cup (A^c \cap B)$. This partition implies that $P(A \cup B) = P(A \cap B^c) + P(A \cap B) + P(A^c \cap B)$. The partitions $A = (A \cap B) \cup (A \cap B^c)$ and $B = (A \cap B) \cup (A^c \cap B)$ imply that $P(A) = P(A \cap B^c) + P(A \cap B)$ and $P(B) = P(A \cap B) + P(A^c \cap B)$. These observations lead to the expression

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

for the union of any two events A and B. This decomposition is illustrated in Figure 4.1.

Figure 4.1 Illustration – the union of 2 events The pluses and minuses indicate how the probabilities for each section enter into and are removed from the sum.



$$\begin{aligned} P(A \cup B) &= P(A \cap B^c) + P(A \cap B) + P(A^c \cap B) \\ &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

Example 4.1 Dogs and cats. Among households in the United States (in 2006), 44% of the households have a dog, 29% have a cat, 17% have both, and 44% do not have a dog or a cat. Let D denote the event that a household has a dog and let C denote the event that a household has a cat. If we select a household at random, so that each household has the same chance of being selected, then $P(D) = .44$, $P(C) = .29$, and $P(D \cap C) = .17$. We will now find $P(D \cup C)$, the probability that a household has a dog or a cat. Application of the formula for the probability of a union yields $P(D \cup C) = P(D) + P(C) - P(D \cap C) = .44 + .29 - .17 = .56$, that is, 56% have a dog or a cat or both.

Some readers may find the following detailed derivation of this probability instructive. Since $P(D) = P(D \cap C) + P(D \cap C^c)$, we find that $P(D \cap C^c) = P(D) - P(D \cap C) = .44 - .17 = .27$, that is, 27% have a dog but not a cat.
Similarly $P(C) = P(D \cap C) + P(D^c \cap C)$, so that $P(D^c \cap C) = P(C) - P(D \cap C) = .29 - .17 = .12$, that is, 12% have a cat but not a dog.
Combining these results we have
$P(D \cup C) = P(D \cap C^c) + P(D \cap C) + P(D^c \cap C) = .27 + .17 + .12 = .56.$

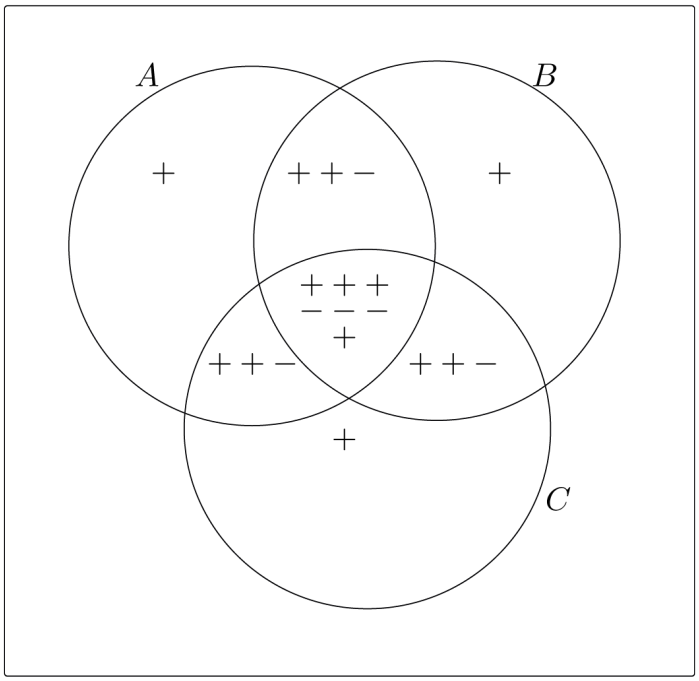
The probability of the union of 3 events

In Figure 3.10 we illustrated the partition of $A \cup B \cup C$ into seven disjoint parts, *viz*, $A \cup B \cup C = (A \cap B^c \cap C^c) \cup (A^c \cap B \cap C^c) \cup (A \cap B \cap C^c) \cup (A \cap B \cap C) \cup (A \cap B^c \cap C) \cup (A^c \cap B \cap C) \cup (A^c \cap B^c \cap C)$. This partition implies that $P(A \cup B \cup C) = P(A \cap B^c \cap C^c) + P(A^c \cap B \cap C^c) + P(A \cap B \cap C^c) + P(A \cap B \cap C) + P(A \cap B^c \cap C) + P(A^c \cap B \cap C) + P(A^c \cap B^c \cap C)$. Similar decompositions for the probabilities of the events A, B, C, $A \cap B$, $A \cap C$, $B \cap C$, and $A \cap B \cap C$, and a bit of algebra yields the decomposition

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

This decomposition is illustrated in Figure 4.2.

Figure 4.2 Illustration – the union of 3 events The pluses and minuses indicate how the probabilities for each section enter into and are removed from the sum.



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

An aside – The probability of the union of many events

The decomposition of the probability of 3 events above can be extended to 4 or more events. This extension is straightforward (but tedious) with alternating inclusions (pluses) and exclusions (minuses). Here is the expression for n events. For any events $A_1, \ldots, A_n$,

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_i P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) + \cdots + (-1)^{n+1} P(A_1 \cap \cdots \cap A_n).$$

This expression indicates that the probability of the union is obtained by first adding the probabilities of each event, then subtracting the probabilities of each intersection of 2 events, then adding the probabilities of each intersection of 3 events, with this process continuing (alternating adding and subtracting) until the probability of the intersection of all n events is entered.

4.3 Probabilities on discrete sample spaces toc

A sample space Ω is said to be **discrete** if it contains a finite or countably infinite number of elementary outcomes. That is, either $\Omega = \{\omega_1, \ldots, \omega_N\}$ for some positive integer N or the elements of Ω can be arranged in a sequence $\Omega = \{\omega_1, \omega_2, \ldots\}$.

A **probability distribution** on a finite sample space $\Omega = \{\omega_1, \ldots, \omega_N\}$ is an assignment of probabilities to the elementary outcomes (elements) of Ω. More formally, given a finite sample space $\Omega = \{\omega_1, \ldots, \omega_N\}$, a collection of probabilities $p_1, \ldots, p_N$, with $0 \leq p_i \leq 1$ and $p_1 + \cdots + p_N = 1$, determines a probability distribution on Ω with $P(\omega_i) = p_i$ for $i = 1, \ldots, N$. Note that in most situations we can remove any elements with zero probability and there are at least two elements with positive probability, thus, we can assume that $0 < p_i < 1$.

Similarly, if $\Omega = \{\omega_1, \omega_2, \ldots\}$ is countably infinite, then a **probability distribution** on Ω is a sequence $p_1, p_2, \ldots$ of probabilities ($P(\omega_i) = p_i$) with $0 \leq p_i \leq 1$ for all i and $\sum_{i=1}^{\infty} p_i = p_1 + p_2 + \cdots = 1$.

Given an event A, *i.e.*, given $A \subset \Omega$, the probability of the event A is the sum of the probabilities of the elementary outcomes which belong to A, *i.e.*, if the elements of Ω are labeled so that $A = \{\omega_1, \omega_2, \ldots, \omega_m\}$, then

$$P(A) = P(\omega_1) + P(\omega_2) + \cdots + P(\omega_m) = p_1 + p_2 + \cdots + p_m.$$

The simplest way to assign probabilities to the elements of a finite sample space $\Omega = \{\omega_1, \ldots, \omega_N\}$ is to assume that the N elementary outcomes are **equally probable** (**equally likely**) so that $P(\omega_i) = 1/N$ for $i = 1, \ldots, N$. When the N elementary outcomes are assumed equally probable, the probability of an event A is $P(A) = N(A)/N$, where $N(A)$ is the number of elementary outcomes which belong to A. In other words, with equally probable outcomes, the probability of event A is the ratio of the number of outcomes "favorable" for A to the number of "possible" outcomes. This simple situation is convenient for demonstrating concepts; but, the usefulness of the assumption of a finite sample space with equally probable outcomes as a model for an idealized version of reality is limited.

Example 4.2 Fiber breaks (from Schervish and DeGroot). Consider an experiment in which five fibers having different lengths are subjected to a testing procedure to see which will

break first. Suppose that the lengths of these five fibers are 1, 2, 3, 4, and 5 inches, respectively. Suppose also that the probability that any given fiber will be the first to break is proportional to its length. We will find the probability that the length of the first fiber to break is no more than 3 inches.

For $i = 1, \ldots, 5$, let ω_i be the outcome that the fiber of length i inches breaks first. Then $\Omega = \{\omega_1, \ldots, \omega_5\}$ and $p_i = ki$ for $i = 1, \ldots, 5$, where k is a proportionality factor. Since we need $p_1 + \ldots + p_5 = 1$ and $k1 + k2 + k3 + k4 + k5 = 15k$, we know that $k = 1/15$. Let A denote the event that the length of the first fiber to break is no more than 3 inches. Then $A = \{\omega_1, \omega_2, \omega_3\}$, and

$$P(A) = p_1 + p_2 + p_3 = \frac{1}{15} + \frac{2}{15} + \frac{3}{15} = \frac{2}{5}.$$

5 Combinatorics – counting toc

5.1 Counting basics toc

We will start with a basic (illustrative) counting problem. Consider a box containing 5 balls labeled with the integers $1, 2, 3, 4, 5$. If we select one ball from the box, then, obviously, there are 5 possible outcomes. Let's consider a slightly more interesting setup. Suppose we select 2 balls from the box in the following way, first we select a ball and note its number, then we return the ball to the box and make a second selection. How many possible outcomes are there now? We can think of this selection process as an experiment with two stages and we can represent the possible outcomes as ordered pairs of the form (a, b) where a denotes the number on the first ball selected and b the number on the second. The answer is that there are 5 times 5 equals 25 possible outcomes. These 25 possible outcomes are listed in Table 5.1a. In this table there are 5 rows, one for each of the 5 possibilities at the first stage, and 5 columns, one for each of the 5 possibilities at the second stage.

Table 5.1a Ordered pairs (a, b) with $a, b \in \{1, 2, 3, 4, 5\}$ (repeats allowed).

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)

How does the answer change if we do not return the first ball to the box before selecting the second? In this case, regardless of which ball was selected first, there are only 4 possibilities at the second stage. As shown in Table 5.1b, in terms of the listing in Table 5.1a, the ordered pairs on the diagonal, (1,1), (2,2), *etc.*, are not possible. Thus, in this case, there are 5 times 4 equals 20 possible outcomes. We will now formalize this fundamental rule of counting.

Table 5.1b Ordered pairs (a, b) with $a, b \in \{1, 2, 3, 4, 5\}$ and $a \neq b$ (repeats not allowed).

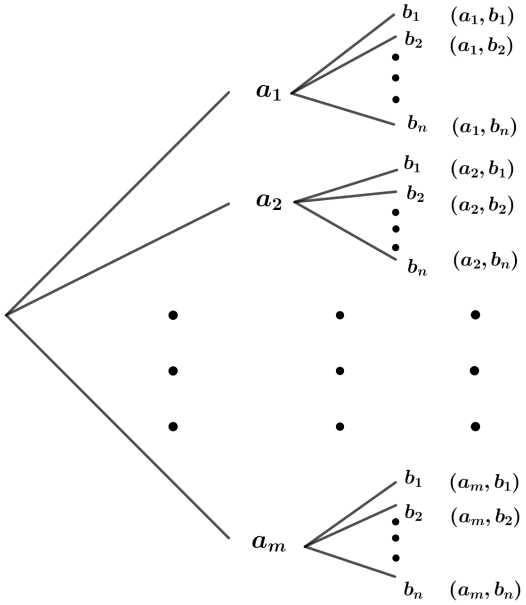
	(1,2)	(1,3)	(1,4)	(1,5)
(2,1)		(2,3)	(2,4)	(2,5)
(3,1)	(3,2)		(3,4)	(3,5)
(4,1)	(4,2)	(4,3)		(4,5)
(5,1)	(5,2)	(5,3)	(5,4)	

Fundamental rule of counting – multiplication rule for counting. *If an experiment consists of two stages (parts) a first stage which can be performed in m ways, and, regardless of the particular outcome of the first stage, a second stage which can be performed in n ways, then the experiment itself can be performed in mn ways.*

If we use ordered pairs to represent the possible outcomes of the experiment, as we did in Tables 5.1a and 5.1b, we have the following alternate statement.

Alternate statement of the fundamental rule of counting. *Given m labels (outcomes) $a_1, \ldots, a_m$ and n labels (outcomes) $b_1, \ldots, b_n$, there are mn ordered pairs of the form (a_i, b_j) with $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$.*

Figure 5.1 Multiplication rule: a tree diagram with mn paths.



We can also envision the choices in the two stages of the experiment as branches in the tree diagram of Figure 5.1. Here a path (choice of an a branch, say a_i, and choice of a b branch, say b_j) corresponds to the outcome associated with the label (a_i, b_j). The fundamental rule of counting says that there are mn paths through this tree diagram. That is, there are mn paths through a tree diagram with m branches at the first level (stage) and n branches at the second level (stage).

Example 5.1 Some basic counting problems.
1. Each year starts on one of the seven days (Sunday through Saturday). Each year is either a leap year (*i.e.*, it includes February 29) or not. How many different calendars are possible for a year?
2. John has 4 clean shirts and 2 clean pairs of jeans. How many clean shirt / clean jeans combinations does John have to choose from?

solutions
1. There are two steps in forming a calendar. First we need to select a day of the week for January 1 (7 choices) and then we need to decide whether to include February 29 (2 choices). Thus there are $7 \times 2 = 14$ possible calendars.
2. John has 4 choices for his shirt and 2 choices for his jeans. Therefore, John has $4 \times 2 = 8$ clean shirt / clean jeans combinations to choose from.

Extension of the fundamental rule of counting. *If an experiment consists of k stages, so that, for $i = 1, \ldots, k$, regardless of the exact outcomes of the other stages, the i^{th} stage can be performed in n_i ways, then the experiment itself can be performed in $n_1 n_2 \cdots n_k$ ways. In other words, thinking of the possible outcomes as ordered k–tuples, there are $n_1 n_2 \cdots n_k$ ways in which an ordered k–tuple can be formed when there are n_1 choices for the first element, n_2 choices for the second element, and so on ending with n_k choices for the k^{th} element.*

Example 5.2 Some more basic counting problems. Suppose that a license plate "number" consists of a combination of three letters followed by four digits, such as ABC–1234.
1. How many such license plate "numbers" are possible if there is no restriction on the choices for the letters or digits?
2. How many such license plate "numbers" are possible if the three letters must be distinct

(three different letters) but there is no restriction on the choices the digits?
3. How many such license plate "numbers" are possible if the three letters must be distinct and the four digits must be distinct?

solutions
There are 26 letters (A, B, ..., Z) and ten digits (0,1, ..., 9).
1. With no restrictions, there are 26 choices for each letter and 10 for each digit resulting in $26 \cdot 26 \cdot 26 \cdot 10 \cdot 10 \cdot 10 \cdot 10 = 175,760,000$ possible license plate "numbers".
2. If the letters must be distinct, then at each letter selection stage there is one fewer choices. Hence, there are $26 \cdot 25 \cdot 24 \cdot 10 \cdot 10 \cdot 10 \cdot 10 = 156,000,000$ possible license plate "numbers".
3. Similarly, if the letters must be distinct and the digits must be distinct, then there are $26 \cdot 25 \cdot 24 \cdot 10 \cdot 9 \cdot 8 \cdot 7 = 78,624,000$ possible license plate "numbers".

5.2 Ordered samples

toc

Consider a population (collection) consisting of N distinct objects.

An **ordered sample** of size n is an ordered collection of n objects selected from the N objects in the population. An ordered sample of this type can be represented by an ordered n–tuple. The multiplication rule can be used to determine the number of possible ordered samples. We can think of the selection of an ordered sample of size n as a sequence of n steps, where the first element of the sample is selected at step one, the second element is selected at step two, and so on until all n elements have been selected. Since the selection of the sample will entail n such steps the multiplication rule will yield the number of possible samples as a product of n values (one for each of the n steps).

Before we start counting, notice that there are two ways in which a sample can be selected. The elements of the sample can be **selected with replacement**, so that at each step the element is selected from the entire population, or the elements of the sample can be **selected without replacement**, so that at each step an object is removed from the population and there is one less object in the population for the next selection. When sampling with replacement a particular object can appear more than once in the sample and there is no restriction on the sample size n. When sampling without replacement no particular object can appear more than once in the sample and the sample size n clearly cannot exceed the population size N.

When sampling with replacement there are N choices at each of the n steps. Thus there are N^n possible ordered samples of size n when the sample is selected with replacement from a population of size N. This result is illustrated for a population of size $N = 4$ and a sample of size $n = 3$ in Table 5.2.

Table 5.2 The $4^3 = 64$ ordered samples of size $n = 3$, selected with replacement from $\{a, b, c, d\}$

(a,a,a)	(a,a,b)	(a,a,c)	(a,a,d)	(a,b,a)	(a,b,b)	(a,b,c)	(a,b,d)
(a,c,a)	(a,c,b)	(a,c,c)	(a,c,d)	(a,d,a)	(a,d,b)	(a,d,c)	(a,d,d)
(b,a,a)	(b,a,b)	(b,a,c)	(b,a,d)	(b,b,a)	(b,b,b)	(b,b,c)	(b,b,d)
(b,c,a)	(b,c,b)	(b,c,c)	(b,c,d)	(b,d,a)	(b,d,b)	(b,d,c)	(b,d,d)
(c,a,a)	(c,a,b)	(c,a,c)	(c,a,d)	(c,b,a)	(c,b,b)	(c,b,c)	(c,b,d)
(c,c,a)	(c,c,b)	(c,c,c)	(c,c,d)	(c,d,a)	(c,d,b)	(c,d,c)	(c,d,d)
(d,a,a)	(d,a,b)	(d,a,c)	(d,a,d)	(d,b,a)	(d,b,b)	(d,b,c)	(d,b,d)
(d,c,a)	(d,c,b)	(d,c,c)	(d,c,d)	(d,d,a)	(d,d,b)	(d,d,c)	(d,d,d)

When sampling without replacement there are N choices at the first step, $N-1$ choices at the second step, and so on, with $N-n+1$ choices at the n^{th} step (assuming that $n \leq N$). Thus, assuming that $n \leq N$, there are $N(N-1)\cdots(N-n+1)$ possible ordered samples of size n when the sample is selected without replacement from a population of size N. Notice that, as before, there are n terms in this product. This result is illustrated for a population of size $N = 4$ and a sample of size $n = 3$ in Table 5.3.

Table 5.3 The $4 \cdot 3 \cdot 2 = 24$ ordered samples of size $n = 3$, selected without replacement from $\{a, b, c, d\}$

(a,b,c)	(a,b,d)	(a,c,b)	(a,c,d)	(a,d,b)	(a,d,c)
(b,a,c)	(b,a,d)	(b,c,a)	(b,c,d)	(b,d,a)	(b,d,c)
(c,a,b)	(c,a,d)	(c,b,a)	(c,b,d)	(c,d,a)	(c,d,b)
(d,a,b)	(d,a,c)	(d,b,a)	(d,b,c)	(d,c,a)	(d,c,b)

For ease of reference these counting results are summarized below.

The number of ordered samples. *Refer to the ordered samples of Tables 5.2 and 5.3 for an illustration of this result. For a population of size N:*
(1) There are N^n ordered samples of size n when the sample is selected with replacement.
(2) For $n \leq N$, there are $N(N-1)\cdots(N-n+1)$ ordered samples of size n when the sample is selected without replacement.

Example 5.3 Some basic ordered sample counting problems. Consider a club consisting of 35 members.
1. In how many ways can this club select three officers (a president, a secretary, and a treasurer), if the three officers are required to be three different members?
2. How does the answer to part 1 change if we allow one person to hold two or even three offices?

solutions
We can use an ordered triple (a, b, c) to represent the choice of the officers (a is president, b is secretary, c is treasurer). Since there are 35 members and we need 3 officers, $N = 35$ and $n = 3$.
1. If a member can hold at most one office, then there are $N(N-1)(N-2) = 35 \cdot 34 \cdot 33 = 39,270$ ways to select the three officers.
2. If a member is allowed to hold two or more offices, then there are $N^n = 35^3 = 42,875$ ways to select the three officers.

An ordered sample selected without replacement can also be viewed as an ordered arrangement of a set of objects. A **permutation of a set** is an ordered arrangement of the elements of the set. Note that if an ordered sample of size $n = N$ is selected without replacement from a population of size N, then the ordered sample is a permutation (ordered arrangement) of the N objects which comprise the population. There are $N! = N(N-1)\cdots 1$ (read this as N **factorial**) permutations of N objects. By convention $0! = 1$.

The $24 = 4 \cdot 3 \cdot 2$ ordered samples of size $n = 3$ selected without replacement from $\{a, b, c, d\}$ listed in Table 5.3 are the permutations of the four letters $\{a, b, c, d\}$ taken three at a time. In general, the **permutations of N objects taken n at a time** are the ordered samples of size n selected without replacement from a population of N objects. Thus there are

$N(N-1)\cdots(N-n+1)$ permutations of N objects taken n at a time. This result is restated below for ease of reference.

The number of permutations – ordered samples. *The number of permutations of N objects taken n at a time, which is also the number of ordered samples of size n selected without replacement from a population of size N, is equal to*

$$ {}_N P_n = \frac{N!}{(N-n)!} = N(N-1)\cdots(N-n+1). $$

Notice that there are n terms in the product determining the value of ${}_N P_n$.

Example 5.4 Arranging people.

1. In how many ways can 6 people be seated in a row of 6 theater seats?

We can represent a seating arrangement by a 6–tuple $(x_1, x_2, x_3, x_4, x_5, x_6)$, where x_i is the label of the seat assigned to person i. There are 6 choices for the first person's seat, 5 choices for the second person's seat, and so on, down to 1 choice for the sixth person's seat. Thus, the number of ways to seat 6 people in a row of 6 theater seats is $6! = 6\cdot5\cdot4\cdot3\cdot2\cdot1 = 720$. Notice that this is the number of permutations of 6 objects and can also be written as ${}_6P_6$.

2. Now suppose that there are 12 seats in the row. In how many ways can 6 people be seated in a row of 12 theater seats?

Arguing as before, in this situation there are 12 choices for the first person's seat, 11 choices for the second person's seat, and so on, down to 7 choices for the sixth person's seat. Thus, the number of ways to seat 6 people in a row of 12 theater seats is ${}_{12}P_6 = 12\cdot11\cdot10\cdot9\cdot8\cdot7 = 665,280$.

Calculator aside: Many calculators use the notation nPr (the number of permutations of n objects taken r at a time). For these calculators $6! = (6\ nPr\ 6)$ *and* ${}_{12}P_6 = 12\cdot11\cdot10\cdot9\cdot8\cdot7 = (12\ nPr\ 6)$.

Example 5.5 Choosing different numbers.
Consider a box containing N balls, numbered $1, 2, \ldots, N$. Suppose a sample of n balls is selected at random with replacement from this box. Assuming that the N^n possible ordered samples of size n are equally likely, we will find the probability P that the numbers on the n balls are all different? If $n > N$, then $P = 0$, since the sample size is greater than the

number of balls in the box and at least one number must occur more than once. If $n \leq N$, then the ${}_N P_n$ permutations of N objects taken n at a time are the samples of size n for which the the numbers on the n balls are all different. Hence, the probability of interest is

$$P = \frac{{}_N P_n}{N^n} = \frac{N!}{(N-n)!\, N^n} = \frac{N(N-1)\cdots(N-n+1)}{N^n}.$$

We will consider an interesting application of this result in the next example.

Example 5.6 The birthday problem. In this example we will find the probability P_n that at least two people in a group of n people have the same birthday (month and day). We need to make a few simplifying assumptions. First assume that there are no twins in the group and assume that no one in the group was born on February 29. Next assume that the 365 birthdays are equally likely. We will use the argument from the preceding example to solve the birthday problem. The complement of the event that at least two people in the group of n have the same birthday is the event that no two people have the same birthday. Since there are $N = 365$ possible birthdays, we have

$$P_n = 1 - \frac{{}_{365} P_n}{365^n} = 1 - \frac{365!}{(365-n)!\, 365^n}$$

Numerical values of this probability for several choices of n are given in Table 5.4. If you have never encountered this problem before, you may find these numbers surprisingly large. If fact, you might reasonably guess that we would need a large number of people, say $n > 100$, in order for P_n to be greater than $1/2$. However, you can see that we only need $n = 23$ people in the group to get $P_n > 1/2$. Notice that for $n = 50$ people there is a 97% chance that at least two people have the same birthday!

Table 5.4 The probability P_n that at least two people in a group of n people will have the same birthday.

n	P_n	n	P_n
5	0.027	23	0.507
10	0.117	25	0.569
15	0.253	30	0.706
20	0.411	40	0.891
22	0.476	50	0.970

5.3 Unordered samples toc

An **unordered sample** of size n is a collection (set) of n objects selected without replacement from the N objects in the population. In other words, an unordered sample is a subset or subpopulation of the original population. We can use set notation to represent an unordered sample. We will now determine, for fixed $n < N$, the number of unordered samples of size n that can be formed from the objects in a population of size N. Note that the selection of a unordered sample of size n can also be viewed as a partitioning (division) of the population into two complementary subpopulations — one of size n (the objects in the sample) and the other of size $N - n$ (the objects not in the sample).

A **combination of N objects taken n at a time** is an unordered sample of size n selected without replacement from a population of N objects, *i.e.*, a subpopulation (subset) of size n. We will now find the number of such combinations.

We know that there are ${}_NP_n = N(N-1)\cdots(N-n+1)$ ordered samples of size n selected without replacement from a population of size N, *i.e.*, there are ${}_NP_n$ permutations of N objects taken n at a time. Each of these ordered samples of size n can be ordered (permuted) in $n!$ ways, *i.e.*, there are $n!$ permutations of a particular set of n distinct objects. This connection is illustrated in the table of Example 5.7 where the ${}_4P_3 = 24$ permutations of $\{a, b, c, d\}$ taken 3 at a time are arranged in rows of $3! = 6$ permutations, one row for each of the $24/6 = 4$ combinations.

Example 5.7 Permutations and combinations. The connection between the permutations of the four letters $\{a, b, c, d\}$ taken three at a time and the combinations of the four letters $\{a, b, c, d\}$ taken three at a time is indicated in this table. Within each row, the first 6 columns contain the 6 permutations of the combination in column 7.

The permutations and combinations of the four letters $\{a, b, c, d\}$ taken three at a time.

permutations (ordered samples)						**combinations (unordered samples)**
(a, b, c)	(a, c, b)	(b, a, c)	(b, c, a)	(c, a, b)	(c, b, a)	$\{a, b, c\}$
(a, b, d)	(a, d, b)	(b, a, d)	(b, d, a)	(d, a, b)	(d, b, a)	$\{a, b, d\}$
(a, c, d)	(a, d, c)	(c, a, d)	(c, d, a)	(d, a, c)	(d, c, a)	$\{a, c, d\}$
(b, c, d)	(b, d, c)	(c, b, d)	(c, d, b)	(d, b, c)	(d, c, b)	$\{b, c, d\}$

If we divide the number of ordered samples of size n by the number of ways that each set of n objects can be ordered, then we obtain the number of unordered samples. For example, there are four combinations in the table of Example 5.7 since ${}_4P_3/3! = 24/6 = 4$. A general version of this expression is given below.

The number of combinations – unordered samples. *The number of combinations of N objects taken n at a time, which is also the number of unordered samples of size n selected without replacement (number of subpopulations of size n) from a population of size N, is given by the binomial coefficient (read $\binom{N}{n}$ as N choose n)*

$$\binom{N}{n} = \frac{N(N-1)\cdots(N-n+1)}{n(n-1)\cdots 1} = \frac{N!}{n!(N-n)!}.$$

The terminology N choose n indicates that $\binom{N}{n}$ is the number of ways to choose n objects without replacement from N objects. The binomial coefficient $\binom{N}{n}$ is sometimes denoted by ${}_NC_n$ with C indicating combinations. Note that in the ratio of products expression for $\binom{N}{n}$ the numerator product $N(N-1)\cdots(N-n+1)$ and the denominator product $n!$ both contain n terms.

Calculator aside: Many calculators use the notation nCr (the number of combinations of n objects taken r at a time). For these calculators $\binom{N}{n} = {}_NC_n = (N\ nCr\ n)$, e.g., $\binom{7}{3} = \frac{7\cdot 6\cdot 5}{3\cdot 2\cdot 1} = 35 = (7\ nCr\ 3)$.

As noted above, a combination of N objects taken n at a time can also be viewed as a subpopulation of size n selected from a population of size N. Thus the **binomial coefficient** $\binom{N}{n}$ is the number of subpopulations of size n from a population of size N. Note also that selecting a subpopulation of size n from a population of size N is equivalent to selecting a collection of $N-n$ objects which are excluded from the subpopulation. Thus, recalling the convention $0! = 1$, which leads to the convention $\binom{N}{0} = \binom{N}{N} = 1$, for $0 \le n \le N$, we have the identity

$$\binom{N}{n} = \binom{N}{N-n}.$$

Continuing along this line of thought we see that there are $\binom{N}{n}$ ways to partition a population of size N into a primary subpopulation of size n and a complementary subpopulation

of size $N - n$. Note that this method of counting partitions takes the order of the subpopulations into account. For example, with $N = 4$ and $n = 2$ the six ordered partitions, expressed as ordered pairs, are:
$(\{a, b\}, \{c, d\}), (\{a, c\}, \{b, d\}), (\{a, d\}, \{b, c\}), (\{b, c\}, \{a, d\}), (\{b, d\}, \{a, c\}), (\{c, d\}, \{a, b\})$.

We will now extend the binomial coefficient to provide a general expression for the number of partitions of a population. For $m \geq 2$ and $n_1, \ldots, n_m$ such that $n_1 + \cdots + n_m = N$, we might ask: In how many ways can a population of size N be partitioned into m subpopulations of respective sizes $n_1, \ldots, n_m$? (Note that, as with the case when $m = 2$ and $n_1 = n_2$, there is an implied ordering of the subpopulations in this statement.) To answer this question consider the formation of such a partition via a sequence of m steps. At the first step there are N objects to choose from and $\binom{N}{n_1}$ ways to select the first subpopulation. At the second step there are $N - n_1$ objects to choose from and $\binom{N-n_1}{n_2}$ ways to select the second subpopulation. Continuing with this reasoning when we reach the last (m^{th}) step we find that there are $N - n_1 - \cdots - n_{m-1} = n_m$ objects to choose from and $\binom{n_m}{n_m} = 1$ ways to select the final subpopulation. The answer to our question is the **multinomial coefficient** given in the following result.

The number of partitions. *Consider a population of size N. For $n_1, \ldots, n_m$ such that each $n_i \geq 1$ and $n_1 + \cdots + n_m = N$, the number of partitions of the population into m subpopulations of respective sizes $n_1, \ldots, n_m$ is given by the multinomial coefficient*

$$\binom{N}{n_1, n_2, \ldots, n_m} = \binom{N}{n_1}\binom{N - n_1}{n_2}\binom{N - n_1 - n_2}{n_3} \cdots \binom{n_m}{n_m} = \frac{N!}{n_1! n_2! \cdots n_m!}.$$

Note that if we allow $n_i = 0$ in this theorem, then the expression is still valid but the number of nontrivial subpopulations in the partition is reduced by the number of i for which $n_i = 0$. Note also that a binomial coefficient is a special case of a multinomial coefficient. In particular,

$$\binom{N}{n, N - n} = \binom{N}{n}.$$

Example 5.7a An application. A child has 20 colored beads, of which 9 are red, 5 are green, 4 are blue, and 2 are black. If the child puts the beads on a string to form a necklace, how many arrangements are possible?

We will assume that the beads of the same color are indistinguishable (all look the same). The elementary outcomes can be represented by ordered sequences of the form

($RRRRRRRRRGGGGGBBBBKK$), where R, G, B, and K denote the respective colors red, green, blue, and black. As noted above, switching the positions of two beads of the same color does not result in a new outcome since such a change would be undetectable once the beads were on the string. We will count by selecting positions in the sequence (on the string) sequentially for the four colors. Initially there are 20 positions to choose from. Thus, there are $\binom{20}{9}$ choices for the 9 red beads. Once this is done, there are 11 positions to choose from, so we have $\binom{11}{5}$ choices for the green beads. Continuing in this way, there are $\binom{6}{4}$ position choices for the blue beads and there is $\binom{2}{2} = 1$ choice for the black beads. Hence, there are

$$\begin{aligned}\binom{20}{9,5,4,2} &= \binom{20}{9}\binom{11}{5}\binom{6}{4}\binom{2}{2} \\ &= 167,960 \cdot 462 \cdot 15 \cdot 1 = 1,163,962,800\end{aligned}$$

possible arrangements of the 20 beads.

The next few sections are devoted to some probabilistic applications of the combinatorial techniques we have been discussing.

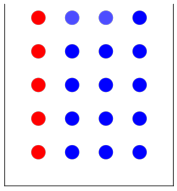
5.4 Sampling with replacement – The binomial distribution toc

We will first consider probabilities associated with the number of objects of a specified type in a random sample selected with replacement.

We will use the selection of balls from a box of suitably labeled balls to represent a random experiment. Consider a box (the population) containing $N = N_1 + N_2$ balls (individual objects) of which N_1 are red (of one type) and $N_2 = N - N_1$ are blue (of a second type). Suppose that a simple random sample of n balls is selected at random with replacement from this population. For an integer x, between 0 and n, we will find the probability $P(x)$ that the sample contains exactly x red balls (and consequently exactly $n - x$ blue balls).

At first glance the use of a box of balls as a model for a population and the drawing of balls from the box as sampling from the population may not appear to lend itself to great generality in application. As we shall see, this is not the case. The simple device of using the selection of balls from a box — a so–called box model — can be adapted to a very wide spectrum of applications.

Figure 5.2 A box containing 5 red and 15 blue balls.



$N_1 = 5$ balls of the first type (red)
$N_2 = 15$ balls of the second type (blue)

Let's start with a simple example. Consider the box (population), illustrated in figure 5.2, containing $N = 20$ balls (objects) of which $N_1 = 5$ are red (of one type) and $N_2 = 15$ are blue (of a second type). Suppose that a simple random sample of $n = 3$ balls is selected at random with replacement from this population. The characteristic of interest here is the number of red balls among the $n = 3$ balls in the sample. The possible numbers of red balls in a sample of $n = 3$ are $0, 1, 2, 3$.

Before we start counting, consider a representation of the possible outcomes of the three draws by ordered triples of the letters R for red and B for blue. *Note well that these 8 elementary outcomes are not equally likely!*

$$\Omega = \{BBB, RBB, BRB, BBR, RRB, RBR, BRR, RRR\}$$

Since we are drawing with replacement, there are always 5 choices for each R and 15 choices for each B in these outcomes. Furthermore, there are always a total of 20 choices for each draw. Thus, the probabilities for these outcomes are as shown below.

$$
\begin{aligned}
P(BBB) &= \frac{15 \cdot 15 \cdot 15}{20 \cdot 20 \cdot 20} = \left(\frac{15}{20}\right)^3 \\
P(RBB) &= \frac{5 \cdot 15 \cdot 15}{20 \cdot 20 \cdot 20} = \left(\frac{5}{20}\right)^1 \left(\frac{15}{20}\right)^2 \\
P(BRB) &= \frac{15 \cdot 5 \cdot 15}{20 \cdot 20 \cdot 20} = \left(\frac{5}{20}\right)^1 \left(\frac{15}{20}\right)^2 \\
P(BBR) &= \frac{15 \cdot 15 \cdot 5}{20 \cdot 20 \cdot 20} = \left(\frac{5}{20}\right)^1 \left(\frac{15}{20}\right)^2 \\
P(RRB) &= \frac{5 \cdot 5 \cdot 15}{20 \cdot 20 \cdot 20} = \left(\frac{5}{20}\right)^2 \left(\frac{15}{20}\right)^1 \\
P(RBR) &= \frac{5 \cdot 15 \cdot 5}{20 \cdot 20 \cdot 20} = \left(\frac{5}{20}\right)^2 \left(\frac{15}{20}\right)^1 \\
P(BRR) &= \frac{15 \cdot 5 \cdot 5}{20 \cdot 20 \cdot 20} = \left(\frac{5}{20}\right)^2 \left(\frac{15}{20}\right)^1 \\
P(RRR) &= \frac{5 \cdot 5 \cdot 5}{20 \cdot 20 \cdot 20} = \left(\frac{5}{20}\right)^3
\end{aligned}
$$

Notice that there is $\binom{3}{0} = 1$ way to get $x = 0$ red (BBB); there are $\binom{3}{1} = 3$ ways to get $x = 1$ red (RBB, BRB, BBR); there are $\binom{3}{2} = 3$ ways to get $x = 2$ red (RRB, RBR, BRR); and, there is $\binom{3}{3} = 1$ way to get $x = 3$ red (RRR). Also notice that for $x = 1$ red and $x = 2$ red, the three probabilities corresponding to the three ways to order the $R's$ and $B's$ are the same. These observations lead to the probabilities $P(x)$ as summarized below.

These probabilities depend on $N_1 = 5$ and $N_2 = 15$ only through the proportions $p_1 = \frac{5}{20} = \frac{1}{4}$ (proportion red) and $p_2 = 1 - p_1 = \frac{15}{20} = \frac{3}{4}$ (proportion blue). This implies that if we had used a box with any values of N_1 and N_2 that yielded these proportions, then the probabilities would be the same!

$$
\begin{aligned}
P(0) = P(0 \text{ red}) &= 1 \cdot \frac{15 \cdot 15 \cdot 15}{20 \cdot 20 \cdot 20} = \binom{3}{0}\left(\frac{5}{20}\right)^0\left(\frac{15}{20}\right)^3 = \binom{3}{0}\left(\frac{1}{4}\right)^0\left(\frac{3}{4}\right)^3 = \frac{27}{64} \\
P(1) = P(1 \text{ red}) &= 3 \cdot \frac{5 \cdot 15 \cdot 15}{20 \cdot 20 \cdot 20} = \binom{3}{1}\left(\frac{5}{20}\right)^1\left(\frac{15}{20}\right)^2 = \binom{3}{1}\left(\frac{1}{4}\right)^1\left(\frac{3}{4}\right)^2 = \frac{27}{64} \\
P(2) = P(2 \text{ red}) &= 3 \cdot \frac{5 \cdot 5 \cdot 15}{20 \cdot 20 \cdot 20} = \binom{3}{2}\left(\frac{5}{20}\right)^2\left(\frac{15}{20}\right)^1 = \binom{3}{2}\left(\frac{1}{4}\right)^2\left(\frac{3}{4}\right)^1 = \frac{9}{64} \\
P(3) = P(3 \text{ red}) &= 1 \cdot \frac{5 \cdot 5 \cdot 5}{20 \cdot 20 \cdot 20} = \binom{3}{3}\left(\frac{5}{20}\right)^3\left(\frac{15}{20}\right)^0 = \binom{3}{3}\left(\frac{1}{4}\right)^3\left(\frac{3}{4}\right)^0 = \frac{1}{64}
\end{aligned}
$$

We are now ready to generalize the expression for $P(x)$ from our specific example with $N_1 = 5$ red, $N_2 = 15$ blue, and a sample of size $n = 3$ to the general situation.

In general, we have a population containing $N = N_1 + N_2$ individual objects (the balls in the box) of which N_1 are of one type (red) and $N_2 = N - N_1$ are of a second type (blue). The proportion of individual objects of the first type (red) is $p_1 = N_1/N$ and the proportion of individual objects of the second type (blue) is $p_2 = N_2/N = 1 - p_1$. We are assuming that a simple random sample of n individual objects is selected at random with replacement from this population.

For $x = 0, 1, \ldots, n$, the probability that the sample of n objects contains exactly x objects of the first type (red) is

$$\begin{aligned} P(x) &= \binom{n}{x} \left(\frac{N_1}{N}\right)^x \left(\frac{N_2}{N}\right)^{n-x} \\ &= \binom{n}{x} p_1^x p_2^{n-x} = \binom{n}{x} p_1^x (1-p_1)^{n-x}. \end{aligned}$$

Notice that this expression depends on N, N_1, and N_2 only through $p_1 = \frac{N_1}{N}$, the probability that a single object (ball) selected at random is of the first type (red), and $1 - p_1 = \frac{N_2}{N} = \frac{N-N_1}{N}$, the probability that a single object (ball) selected at random is of the second type (blue). Hence, for fixed values of n and x, the probability $P(x)$ does not depend on the size of the population; it only depends on the proportion of red balls in the population.

Let's review the terms in this expression for the binomial probability $P(x)$. It is helpful to split this expression into two terms.

$$P(x) = \binom{n}{x} \left(p_1^x p_2^{n-x}\right) = \binom{n}{x} \left(\frac{N_1^x \cdot N_2^{n-x}}{N^x}\right)$$

(1) $\binom{n}{x}$ is counting the number of ways that x $R's$ and $n - x$ $B's$ can be arranged in a sequence of length n. This is analogous to counting the number of ways to string x red beads and $n - x$ blue beads like we did in Example 5.7a.

(2) $p_1^x p_2^{n-x} = \frac{N_1^x \cdot N_2^{n-x}}{N^x}$ is the probability that we observed x $R's$ followed by $n - x$ $B's$, expressed as the ratio $N(A)/N$ of the number of favorable outcomes to the number of possible outcomes. As noted above, this probability is the same for each of the $\binom{n}{x}$ ways to order x $R's$ and $n - x$ $B's$.

Expressing the probability $P(x)$ as the product of terms (1) and (2) is equivalent to summing the probabilities (as expressed in term (2)) for each of the $\binom{n}{x}$ ways that x $R's$ and $n-x$ $B's$ can be arranged in a sequence of length n.

We will reconsider this binomial distribution and discuss it in more detail in a later section.

Example 5.8 Tossing a die Tossing a fair (balanced) die and noting the number on the upturned face is abstractly the same as selecting a number (ball) at random from the set $\{1,2,3,4,5,6\}$. Tossing the die n times is equivalent to selecting a random sample of size n with replacement from the set $\{1,2,3,4,5,6\}$.

If a fair die is tossed 10 times, then, for $x = 0,1,\ldots,10$, the probability of observing exactly x aces (ones) is

$$P(x) = \binom{10}{x}\left(\frac{1}{6}\right)^x\left(\frac{5}{6}\right)^{10-x}.$$

In particular:
The probability of observing exactly 0 aces is $P(0) = \binom{10}{0}(\frac{1}{6})^0(\frac{5}{6})^{10} = \frac{1\cdot 5^{10}}{6^{10}} \approx .1615$.
The probability of observing exactly 1 ace is $P(1) = \binom{10}{1}(\frac{1}{6})^1(\frac{5}{6})^{9} = \frac{10\cdot 5^{9}}{6^{10}} \approx .3230$.
The probability of observing exactly 2 aces is $P(2) = \binom{10}{2}(\frac{1}{6})^2(\frac{5}{6})^{8} = \frac{45\cdot 5^{8}}{6^{10}} \approx .2907$.
The probability of observing exactly 3 aces is $P(3) = \binom{10}{3}(\frac{1}{6})^3(\frac{5}{6})^{7} = \frac{120\cdot 5^{7}}{6^{10}} \approx .1550$.

Example 5.9 Sampling from a box of balls Consider a box containing 600 balls of which 100 are red and 500 are blue. Since $p = 1/6$ of the balls in the box are red and the probability of getting an ace when a fair die is tossed once is $p = 1/6$, selecting 10 balls at random with replacement from these 600 balls and counting the number of red balls is equivalent to tossing a fair die 10 times and counting the number of aces tossed. Thus the probabilities of Example 5.8 apply here as well. That is, if 10 balls are selected at random with replacement from these 600 balls, then, for $x = 0,1,\ldots,10$, the probability of observing exactly x red balls is

$$P(x) = \binom{10}{x}\left(\frac{1}{6}\right)^x\left(\frac{5}{6}\right)^{10-x}.$$

5.5 Sampling without replacement – The hypergeometric distribution toc

We will now consider probabilities associated with the number of objects of a specified type in a random sample selected without replacement.

As before, we will use the selection of balls from a box of suitably labeled balls to represent a random experiment and we will consider a box (the population) containing $N = N_1 + N_2$ balls (individual objects) of which N_1 are red (of one type) and $N_2 = N - N_1$ are blue (of a second type).

This time, however, suppose that a simple random sample of n balls is selected at random without replacement from this population. For an integer x, between 0 and n, we will find the probability $P(x)$ that the sample contains exactly x red balls (and consequently exactly $n - x$ blue balls). Note that, since the population contains N_1 red balls and $N_2 = N - N_1$ blue balls and we are sampling without replacement certain values of x may not be attainable. In particular, we must have $P(x) = 0$ when $x > N_1$ and when $n - x > N - N_1$, since these values of x are not possible.

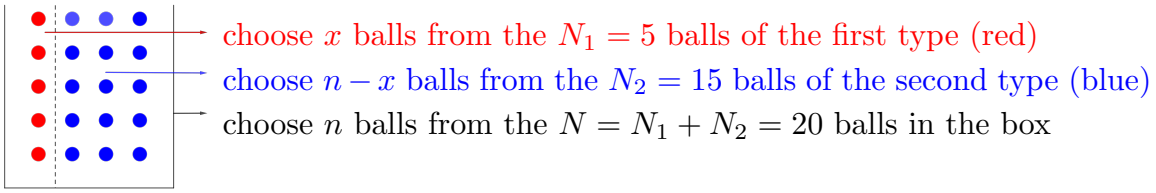
Let's start with the same simple example as we used when introducing the binomial distribution. Consider the box (population), illustrated in figure 5.2, containing $N = 20$ balls (objects) of which $N_1 = 5$ are red (of one type) and $N_2 = 15$ are blue (of a second type). Suppose that a simple random sample of $n = 3$ balls is selected at random without replacement from this population. The characteristic of interest here is the number of red balls among the $n = 3$ balls in the sample. The possible numbers of red balls in a sample of $n = 3$ are $0, 1, 2, 3$. *In this application, since the box contains more than 3 balls of both colors, all values of x are attainable.*

For the probabilities we want to compute, when sampling without replacement we can think of selecting all n balls at once and we can use unordered outcomes in our representation of the sample space. There are $\binom{20}{3} = 1,140$ ways to choose 3 balls (without keeping track of order) from a box containing 20 balls. In order to formally list these elementary outcomes we would need to add labels to the 20 balls so that they would be distinguishable. This is tedious and not very helpful so let's just consider the counting argument. Before we start counting, note that there are three basic types of elementary outcomes here, using R for red and B for blue, the possibilities are: a set of the form $\{B, B, B\}$ (three distinct blue balls); a set of the form $\{R, B, B\}$ (one red ball and two distinct blue balls); a set of the form $\{R, R, B\}$ (two distinct red balls and one blue ball); and, a set of the form $\{R, R, R\}$

(three distinct red balls). *Note well that, as when sampling with replacement, these 4 types of elementary outcomes are not equally likely!*

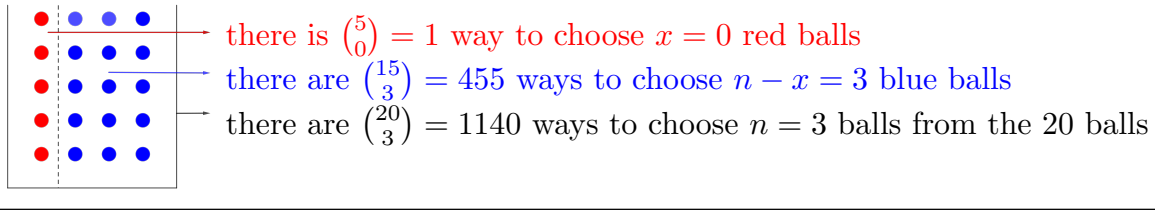
It is helpful to think of the process of choosing the balls as series of three stages (select the desired number, x, of red balls, select the desired number, $n-x$, of blue balls, and select the desired number, n, of balls). These three selections are illustrated in Figure 5.3.

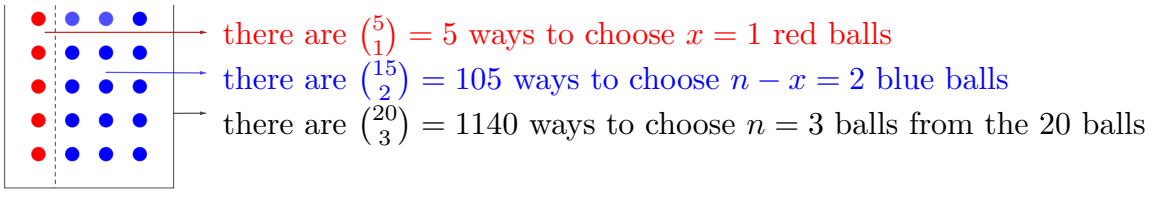
Figure 5.3 Choosing balls from a box containing 5 red and 15 blue balls.

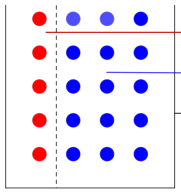


There is $\binom{5}{0} = 1$ way to choose $x = 0$ red from 5 red balls; there are $\binom{5}{1} = 5$ ways to choose $x = 1$ red from 5 red balls; there are $\binom{5}{2} = 10$ ways to choose $x = 2$ red from 5 red balls; and, there are $\binom{5}{3} = 10$ ways to choose $x = 3$ red from 5 red balls. Similarly, there are $\binom{15}{3} = 455$ ways to choose $n - x = 3 - 0 = 3$ blue from 15 blue balls; there are $\binom{15}{2} = 105$ ways to choose $n - x = 3 - 1 = 2$ blue from 15 blue balls; there are $\binom{15}{1} = 15$ ways to choose $n - x = 3 - 2 = 1$ blue from 15 blue balls; and, there is $\binom{15}{0} = 1$ way to choose $n - x = 3 - 3 = 0$ blue from 15 blue balls. Finally, as noted earlier, there are $\binom{20}{3} = 1,140$ ways to choose $n = 3$ balls from the $N = 20$ balls in the box. These observations are represented graphically in Figure 5.4.

Figure 5.4 Choosing balls from a box containing 5 red and 15 blue balls.

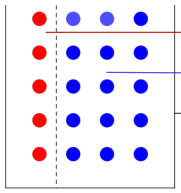






there are $\binom{5}{2} = 10$ ways to choose $x = 2$ red balls
there are $\binom{15}{1} = 15$ ways to choose $n - x = 1$ blue balls
there are $\binom{20}{3} = 1140$ ways to choose $n = 3$ balls from the 20 balls

---



there are $\binom{5}{3} = 10$ ways to choose $x = 3$ red balls
there is $\binom{15}{0} = 1$ way to choose $n - x = 0$ blue balls
there are $\binom{20}{3} = 1140$ ways to choose $n = 3$ balls from the 20 balls

---

Combining these observations and computations yields the probabilities $P(x)$ as summarized below.

$$P(0) = P(0 \text{ red}) = \frac{\binom{5}{0}\binom{15}{3}}{\binom{20}{5}} = \frac{1 \cdot 455}{1140} = \frac{91}{228}$$

$$P(1) = P(1 \text{ red}) = \frac{\binom{5}{1}\binom{15}{2}}{\binom{20}{5}} = \frac{5 \cdot 105}{1140} = \frac{105}{228}$$

$$P(2) = P(2 \text{ red}) = \frac{\binom{5}{2}\binom{15}{1}}{\binom{20}{5}} = \frac{10 \cdot 15}{1140} = \frac{30}{228}$$

$$P(3) = P(3 \text{ red}) = \frac{\binom{5}{3}\binom{15}{0}}{\binom{20}{5}} = \frac{10 \cdot 1}{1140} = \frac{2}{228}$$

We are now ready to generalize the expression for $P(x)$ from our specific example with $N_1 = 5$ red, $N_2 = 15$ blue, and a sample of size $n = 3$ to the general situation.

In general, we have a population containing $N = N_1 + N_2$ individual objects (the balls in the box) of which N_1 are of one type (red) and $N_2 = N - N_1$ are of a second type (blue). We are assuming that a simple random sample of n individual objects is selected at random without replacement from this population.

For $x = 0, 1, \ldots, n$ (subject to the restrictions described below) the probability that the sample of n objects contains exactly x objects of the first type (red) is

$$P(x) = \frac{\binom{N_1}{x}\binom{N-N_1}{n-x}}{\binom{N}{n}} = \frac{\binom{N_1}{x}\binom{N_2}{n-x}}{\binom{N_1+N_2}{n}}.$$

The restriction mentioned above says that this hypergeometric probability formula only works when $x \leq N_1$ and $n - x \leq N_2$. This restriction reflects the fact that we are sampling without replacement and that it may be possible to run out of objects (balls) of one type (color) before we obtain a sample of the desired size n.

Notice the pattern in the binomial coefficients in the expression for the hypergeometric probability $P(x)$. The "top" numbers, N_1 and N_2, in the numerator binomial coefficients sum to give the "top" number, N, in the denominator binomial coefficient. Similarly, the "bottom" numbers, x and $n - x$, in the numerator binomial coefficients sum to give the "bottom" number, n, in the denominator binomial coefficient. This pattern reflects the facts that the N objects (balls) are comprised of N_1 objects of the first type (red balls) and N_2 objects of the second type (blue balls). And, the n objects (balls) in the sample are comprised of x objects of the first type (red balls) and $n - x$ objects of the second type (blue balls).

Also notice that, unlike the binomial probability formula, this hypergeometric probability formula does depend on the values of N_1 and N_2. Therefore, these hypergeometric probabilities cannot be computed unless we know the sizes of the two subpopulations.

We will reconsider this hypergeometric distribution and discuss it in more detail in a later section.

Example 5.10 Poker Consider the number of aces in a 5–card poker hand. Dealing a 5–card hand is abstractly the same as selecting five cards at random without replacement from a collection of 52 balls (cards). For this problem, the relevant partition of the 52 cards in the deck is into the 4 aces and the 48 non–aces. Since there are only 4 aces the x values of interest are $x = 0, 1, 2, 3, 4$. For $x = 0, 1, 2, 3, 4$, the probability that the 5–card hand contains exactly x aces is

$$P(x) = \frac{\binom{4}{x}\binom{48}{5-x}}{\binom{52}{5}}.$$

In particular:
The probability that the hand contains exactly 0 aces is $P(0) = [\binom{4}{0}\binom{48}{5}]/\binom{52}{5} \approx .6588$
The probability that the hand contains exactly 1 ace is $P(1) = [\binom{4}{1}\binom{48}{4}]/\binom{52}{5} \approx .2995$
The probability that the hand contains exactly 2 aces is $P(2) = [\binom{4}{2}\binom{48}{3}]/\binom{52}{5} \approx .0399$
The probability that the hand contains exactly 3 aces is $P(3) = [\binom{4}{3}\binom{48}{2}]/\binom{52}{5} \approx .0017$
The probability that the hand contains exactly 4 aces is $P(4) = [\binom{4}{4}\binom{48}{1}]/\binom{52}{5} \approx .00002$

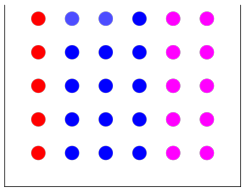
5.6 Sampling with replacement – The multinomial distribution toc

We will now extend the binomial distribution by allowing three types of objects. As before, suppose that a simple random sample of n balls (objects) is selected at random with replacement from a population of N balls (objects). However, now suppose that N_1 balls are red (of one type), N_2 are blue (of a second type), and N_3 are magenta (of a third type), with $N = N_1 + N_2 + N_3$.

For $n = x_1 + x_2 + x_3$, we will find the probability $P(x_1, x_2, x_3)$ that the sample contains x_1 red balls, x_2 blue balls, and x_3 magenta balls. As in the binomial case, we will think of obtaining the balls in the sample one at a time and use ordered outcomes in our representation of the sample space.

As we did with the binomial distribution, let's start with a simple example. Consider the box (population), illustrated in Figure 5.5, containing $N = 30$ balls (objects) of which $N_1 = 5$ are red (of one type), $N_2 = 15$ are blue (of a second type), and $N_3 = 10$ are magenta (of a third type). Suppose that a simple random sample of $n = 6$ balls is selected at random with replacement from this population. The characteristics of interest here are the number of balls of each of the three colors (objects of each of the three types) among the $n = 6$ balls in the sample. There are 28 ways in which a sample of $n = 6$ balls can be distributed among the three colors. For example, we might observe ($x_1 = 0$ red, $x_2 = 0$ blue, and $x_3 = 6$ magenta) or ($x_1 = 1$ red, $x_2 = 2$ blue, and $x_3 = 3$ magenta) or ($x_1 = 2$ red, $x_2 = 2$ blue, and $x_3 = 2$ magenta).

Figure 5.5 A box containing 5 red, 15 blue, and 10 magenta balls.



$N_1 = 5$ balls of the first type (red)
$N_2 = 15$ balls of the second type (blue)
$N_3 = 10$ balls of the third type (magenta)

Before we start counting, consider a representation of the possible outcomes of the six draws by ordered 6–tuples of the letters R for red, B for blue, and M for magenta. *Note*

well that these 28 elementary outcomes are not equally likely!

$$\begin{aligned}\Omega = \{&MMMMMM, BMMMMM, BBMMMM, BBBMMM, BBBBMM, BBBBBM,\\ &BBBBBB, RMMMMM, RBMMMM, RBBMMM, RBBBMM, RBBBBM,\\ &RBBBBB, RRMMMM, RRBMMM, RRBBMM, RRBBBM, RRBBBB,\\ &RRRMMM, RRRBMM, RRRBBM, RRRBBB, RRRRMM, RRRRBM,\\ &RRRRBB, RRRRRM, RRRRRB, RRRRRR\}\end{aligned}$$

Since we are drawing with replacement, there are always 5 choices for each R, 15 choices for each B, and 10 choices for each M in these outcomes. Furthermore, there are always a total of 30 choices for each draw. The probabilities for a few of these outcomes are

$$P(RRBBMM) = \frac{5\cdot 5\cdot 15\cdot 15\cdot 10\cdot 10}{30\cdot 30\cdot 30\cdot 30\cdot 30\cdot 30} = \left(\frac{5}{30}\right)^2\left(\frac{15}{30}\right)^2\left(\frac{10}{30}\right)^2$$

$$P(RBBMMM) = \frac{5\cdot 15\cdot 15\cdot 10\cdot 10\cdot 10}{30\cdot 30\cdot 30\cdot 30\cdot 30\cdot 30} = \left(\frac{5}{30}\right)^1\left(\frac{15}{30}\right)^2\left(\frac{10}{30}\right)^3$$

$$P(RRRBBB) = \frac{5\cdot 5\cdot 5\cdot 15\cdot 15\cdot 15}{30\cdot 30\cdot 30\cdot 30\cdot 30\cdot 30} = \left(\frac{5}{30}\right)^3\left(\frac{15}{30}\right)^3\left(\frac{10}{30}\right)^0$$

$$P(RRRRRR) = \frac{5\cdot 5\cdot 5\cdot 5\cdot 5\cdot 5}{30\cdot 30\cdot 30\cdot 30\cdot 30\cdot 30} = \left(\frac{5}{30}\right)^6\left(\frac{15}{30}\right)^0\left(\frac{10}{30}\right)^0$$

For each collection of values x_1, x_2, x_3 with $x_1 + x_2 + x_3 = 6$, we now need to determine how many ways we can order a collection of x_1 $R's$, x_2 $B's$, and x_3 $M's$. As we noted in the binomial section, this is analogous to asking in how many ways can we string a collection of x_1 red beads, x_2 blue beads, and x_3 magenta beads (when $x_1 + x_2 + x_3 = 6$). In other words, we need count the number of ways to partition the $n = 6$ draws (elements of the 6–tuple) into a group of x_1 when a red ball is drawn, a group of x_2 when a blue ball is drawn, and a group of x_3 when a magenta ball is drawn. The solution is provided by the multinomial coefficient

$$\binom{6}{x_1, x_2, x_3} = \binom{6}{x_1}\binom{6-x_1}{x_2}\binom{6-x_1-x_2}{x_3}.$$

As in the binomial case, we can now combine these observations and provide the following expression for these multinomial probabilities in the application when $n = 6$, $p_1 = N_1/N = 5/30$, $p_2 = N_2/N = 15/30$, and $p_3 = N_3/N = 10/30$. For values of x_1, x_2, x_3 between 0

and 6 with $x_1 + x_2 + x_3 = 6$, the probability of observing x_1 red balls, x_2 blue balls, and x_3 magenta balls in a sample of size $n = 6$ is

$$\begin{aligned} P(x_1, x_2, x_3) &= \binom{6}{x_1, x_2, x_3} \left(\frac{5}{30}\right)^{x_1} \left(\frac{15}{30}\right)^{x_2} \left(\frac{10}{30}\right)^{x_3} \\ &= \binom{6}{x_1}\binom{6-x_1}{x_2}\binom{6-x_1-x_2}{x_3} \left(\frac{5}{30}\right)^{x_1} \left(\frac{15}{30}\right)^{x_2} \left(\frac{10}{30}\right)^{x_3} \end{aligned}$$

Generalizing this expression for $P(x_1, x_2, x_3)$ from our specific example with $N_1 = 5$ red, $N_2 = 15$ blue, $N_3 = 10$ magenta, and a sample of size $n = 6$ to the general situation is straightforward.

In general, we have a population containing $N = N_1 + N_2 + N_3$ individual objects (the balls in the box) of which N_1 are of one type (red) and N_2 are of a second type (blue), and N_3 are of the third type (magenta). The proportion of individual objects of the first type (red) is $p_1 = N_1/N$, the proportion of individual objects of the second type (blue) is $p_2 = N_2/N$, and the proportion of individual objects of the third type (magenta) is $p_3 = N_3/N$. Note well that $p_1 + p_2 + p_3 = 1$, since $N_1 + N_2 + N_3 = N$. We are assuming that a simple random sample of n individual objects is selected at random with replacement from this population.

For values of x_1, x_2, x_3 between 0 and n with $x_1 + x_2 + x_3 = n$, the probability of observing x_1 objects of the first type (red balls), x_2 objects of the second type (blue balls), and x_3 objects of the third type (magenta balls) in a sample of size n is

$$\begin{aligned} P(x_1, x_2, x_3) &= \binom{n}{x_1, x_2, x_3} \left(\frac{N_1}{N}\right)^{x_1} \left(\frac{N_2}{N}\right)^{x_2} \left(\frac{N_3}{N}\right)^{x_3} \\ &= \binom{n}{x_1}\binom{n-x_1}{x_2}\binom{n-x_1-x_2}{x_3} \left(\frac{N_1}{N}\right)^{x_1} \left(\frac{N_2}{N}\right)^{x_2} \left(\frac{N_3}{N}\right)^{x_3} \\ &= \binom{n}{x_1}\binom{n-x_1}{x_2}\binom{n-x_1-x_2}{x_3} p_1^{x_1} p_2^{x_2} p_3^{x_3} \end{aligned}$$

Notice that this expression depends on N, N_1, N_2 and N_3 only through $p_1 = \frac{N_1}{N}$, the probability that a single object (ball) selected at random is of the first type (red), $p_2 = \frac{N_2}{N}$, the probability that a single object (ball) selected at random is of the second type (blue), and $p_3 = \frac{N_3}{N}$, the probability that a single object (ball) selected at random is of the third type (magenta). Hence, for fixed values of n, x_1, x_2, and x_3, the probability $P(x_1, x_2, x_3)$ does not depend on the size of the population; it only depends on the proportions of

objects (balls) in the three subpopulations. *Computational aside:* $n - x_1 - x_2 = x_3$ *and thus* $\binom{n-x_1-x_2}{x_3} = \binom{x_3}{x_3} = 1$.

As you might expect, this is a trinomial distribution, since we have three types of objects (colors). The extension of this expression to give multinomial probabilities when there are four or more types of objects should be reasonably clear. We will reconsider this multinomial distribution and discuss it in more detail in a later section.

Example 5.11 Tossing a die Suppose a fair die is tossed 10 times. Since the population (possible outcomes for one toss) $\{1, 2, 3, 4, 5, 6\}$ contains six values, in order to apply the multinomial probability formula, we need to partition these six values into a fixed number of values ("colors").

First consider the probability of observing exactly 3 ones and exactly 4 twos. Since $3+4 = 7$ we also need 3 numbers other than one or two. Thus we need the partition $\{1, 2, 3, 4, 5, 6\} = \{1\} \cup \{2\} \cup \{3, 4, 5, 6\}$. Here we have $N_1 = 1$, $N_2 = 1$ and $N_3 = 4$ and $x_1 = 3$, $x_2 = 4$, and $x_3 = 3$. The probability of observing exactly 3 ones and exactly 4 twos is

$$\begin{aligned} P(\text{3 ones and 4 twos}) &= P(\text{3 ones, 4 twos, and 3 others}) \\ &= \binom{10}{3}\binom{7}{4}\binom{3}{3}\left(\frac{1}{6}\right)^3\left(\frac{1}{6}\right)^4\left(\frac{4}{6}\right)^3 = \frac{120 \cdot 35 \cdot 1 \cdot 4^3}{6^{10}} \approx .0044. \end{aligned}$$

Now let's consider an event which involves four values. In particular, let's find the probability of observing exactly 2 ones, exactly 3 twos, and exactly 4 threes. Since $2+3+4 = 9$ we also need 1 number other than one, two, or three. Here we have $N_1 = 1$, $N_2 = 1$, $N_3 = 1$, and $N_4 = 3$ and $x_1 = 2$, $x_2 = 3$, $x_3 = 4$, and $x_4 = 1$. The probability of observing exactly 2 ones, exactly 3 twos, and exactly 4 threes is

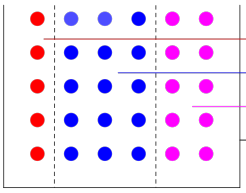
$$\begin{aligned} P(\text{2 ones, 3 twos, and 4 threes}) &= P(\text{2 ones, 3 twos, 4 threes, and 1 other}) \\ &= \binom{10}{2}\binom{8}{3}\binom{5}{4}\binom{1}{1}\left(\frac{1}{6}\right)^2\left(\frac{1}{6}\right)^3\left(\frac{1}{6}\right)^4\left(\frac{3}{6}\right)^1 \\ &= \frac{45 \cdot 56 \cdot 5 \cdot 1 \cdot 3^1}{6^{10}} \approx .00062. \end{aligned}$$

5.7 Sampling without replacement – The multiple hypergeometric dist.toc

The hypergeometric distribution is also readily generalized to allow for balls of three or more colors. As before, suppose that a simple random sample of n balls (objects) is selected at random without replacement from a box (population) of N balls (objects). However, now suppose that N_1 balls are red (of one type), N_2 are blue (of a second type), and N_3 are magenta (of a third type), with $N = N_1 + N_2 + N_3$.

As we did with the hypergeometric distribution, let's start with a simple example. Consider the box (population), illustrated in Figure 5.6, containing $N = 30$ balls (objects) of which $N_1 = 5$ are red (of one type), $N_2 = 15$ are blue (of a second type), and $N_3 = 10$ are magenta (of a third type). Suppose that a simple random sample of $n = 6$ balls is selected at random without replacement from this population.

Figure 5.6 Choosing balls from a box containing 5 red and 15 blue balls.



choose x_1 balls from the $N_1 = 5$ balls of the first type (red)
choose x_2 balls from the $N_2 = 15$ balls of the second type (blue)
choose x_3 balls from the $N_3 = 10$ balls of the third type (magenta)
choose $n = 6$ balls from the $N = N_1 + N_2 + N_3 = 30$ balls in the box

In this application there are $\binom{30}{6} = 593,775$ equally likely possible outcomes (subsets of size $n = 6$). Our task is to determine how many of these possible outcomes are favorable for each event of the form "the sample contains x_1 red balls, x_2 blue balls, and x_3 magenta balls" so that we can compute the corresponding probability $P(x_1, x_2, x_3)$.

The ingredients we need are:

1. There are $\binom{5}{x_1}$ ways to choose x_1 red balls from the $N_1 = 5$ red balls in the box.

2. There are $\binom{15}{x_2}$ ways to choose x_2 blue balls from the $N_2 = 15$ blue balls in the box.

3. There are $\binom{10}{x_3}$ ways to choose x_3 magenta balls from the $N_3 = 10$ magenta balls in the box.

The obvious extension of the hypergeometric probability expression yields

$$P(x_1, x_2, x_3) = \frac{\binom{5}{x_1}\binom{15}{x_2}\binom{10}{x_3}}{\binom{30}{6}}$$

Note well that the sizes of the subpopulations (numbers of balls of each color in the box) place some restrictions on which values of x_1, x_2, and x_3, will work in this expression. Since there are only 5 red balls we must have $x_1 \in \{0, 1, 2, 3, 4, 5\}$. Since there are more than six blue balls and more than six magenta balls, there is no restriction on the values of x_2 and x_3 other than the requirement that $x_1 + x_2 + x_3 = 6$.

Generalizing this expression for $P(x_1, x_2, x_3)$ from our specific example with $N_1 = 5$ red, $N_2 = 15$ blue, $N_3 = 10$ magenta, and a sample of size $n = 6$ to the general situation is straightforward.

In general, we have a population containing $N = N_1 + N_2 + N_3$ individual objects (the balls in the box) of which N_1 are of one type (red) and N_2 are of a second type (blue), and N_3 are of the third type (magenta). We are assuming that a simple random sample of n individual objects is selected at random without replacement from this population.

For values of x_1, x_2, x_3 between 0 and n with $x_1 + x_2 + x_3 = n$ (subject to the restrictions that $x_1 \leq N_1, x_2 \leq N_2$, and $x_3 \leq N_3$), the probability of observing x_1 objects of the first type (red balls), x_2 objects of the second type (blue balls), and x_3 objects of the third type (magenta balls) in a sample of size n is

$$P(x_1, x_2, x_3) = \frac{\binom{N_1}{x_1}\binom{N_2}{x_2}\binom{N_3}{x_3}}{\binom{N_1+N_2+N_3}{x_1+x_2+x_3}} = \frac{\binom{N_1}{x_1}\binom{N_2}{x_2}\binom{N_3}{x_3}}{\binom{N}{n}}$$

As with the hypergeometric distribution, notice the pattern in the binomial coefficients in the expression for the multiple hypergeometric probability $P(x_1, x_2, x_3)$. The "top" numbers, N_1, N_2, and N_3, in the numerator binomial coefficients sum to give the "top" number, N, in the denominator binomial coefficient. Similarly, the "bottom" numbers, x_1, x_2, and x_3, in the numerator binomial coefficients sum to give the "bottom" number, n, in the denominator binomial coefficient. This pattern reflects the facts that the N objects (balls) are comprised of N_1 objects of the first type (red balls), N_2 objects of the second type (blue balls), and N_3 objects or the third type (magenta balls). And, the n objects (balls) in the sample are comprised of x_1 objects of the first type (red balls), x_2 objects of the second type (blue balls) and x_3 objects of the third type (magenta balls). Also notice that, unlike the multinomial probability formula, this multiple hypergeometric probability formula does depend on the values of N_1, N_2, and N_3. Therefore, as with hypergeometric

probabilities, these multiple hypergeometric probabilities cannot be computed unless we know the sizes of the three subpopulations.

The extension to a population with four or more types of objects (balls) should be clear.

Example 5.12 Poker Now we can find the probabilities of some more interesting poker hands. The 52 card deck contains 4 aces, 4 kings, and 44 other cards. Using this partition yields the following probabilities.
The probability that a poker hand contains exactly 2 aces and exactly 1 king is

$$\frac{\binom{4}{2}\binom{4}{1}\binom{44}{2}}{\binom{52}{5}} \approx .0087.$$

The probability that a poker hand contains exactly 2 aces and exactly 2 kings is

$$\frac{\binom{4}{2}\binom{4}{2}\binom{44}{1}}{\binom{52}{5}} \approx .0006.$$

The probability that a poker hand contains exactly 2 aces and exactly 3 kings is

$$\frac{\binom{4}{2}\binom{4}{3}\binom{44}{0}}{\binom{52}{5}} \approx .000009.$$

6 Conditional probability and independence toc

6.1 Conditional probability toc

In many situations we have partial information about the outcome of an experiment and we need to update the probability measure to reflect this additional information. More formally, consider two events A and B which can occur at the same time, if we know that event A has occurred, then we need to determine how to update the probability of event B to take this information about A into account. That is, we need to determine the conditional probability of event B conditioning on the fact that event A has occurred. A couple of simple examples will help motivate the formal definition of conditional probability given below.

Example 6.1 Selecting one card Suppose we select 1 card at random from a 52 card deck. The deck of 52 cards contains 4 kings. Thus, the probability that we will select a king is $P(\text{king}) = 4/52 = 1/13$.

Now suppose we know that the card selected is a face card. How does this partial information about the card affect the probability that it is a king? That is, what is the conditional probability that the card is a king given the information that it is a face card? Since there are 12 face cards (the 4 kings, 4 queens, and 4 jacks) and we know that one of these 12 cards has been selected at random, there is a 4 out of 12 probability that the card is a king. Thus, when we know that the card is a face card the probability that it is a king (conditioning on the fact that it is a face card) is $P(\text{king given the card is a face card}) = 4/12 = 1/3$.

Next suppose that we know that the card selected is a spade. How does this partial information about the card affect the probability that it is a king? That is, what is the conditional probability that the card is a king given the information that it is a spade? Since there are 13 spades and we know that one of these 13 spades has been selected at random, and there is only one king of spades, there is a 1 out of 13 probability that the card is a king. Thus, when we know that the card is a spade the probability that it is a king (conditioning on the fact that it is a spade) is $P(\text{king given the card is a spade}) = 1/13$.

It is helpful to consider the three probabilities we just computed in terms of the representation of the 52 cards in the image of Figure 3.1. The first probability we computed, the

unconditional probability that the card is a king, is 1/13 because we are selecting at random from all 52 cards and 4 of these are kings. For the second probability, we conditioned on the event that the card was a face card. That is, we restricted our selection to the 12 cards in the last three columns of Figure 3.1. Since 1/3 of the cards in these columns are kings, we argued that the conditional probability of selecting a king given that the card is a face card is 1/3. For the third probability, we conditioned on the event that the card was a spade. That is, we restricted our selection to the 13 spades in the third row of Figure 3.1. Since one of the cards in the spade row is a king, we argued that the conditional probability of selecting a king given that the card is a spade is 1/13.

Notice that the unconditional probability of selecting a king changes when we condition on the event that the card is a face card; but, it does not change when we condition on the event that the card is a spade. When two events, say A and B, have the property that the unconditional probability of A is the same as the conditional probability of A given that B has occurred, the events are said to be independent. We will discuss independence in Section 6.2

Example 6.2 Bachelor's degree majors In 2004–2005, degree–granting institutions in the US conferred $N = 1,439,264$ bachelor's degrees (see Table 6.1). Of these degrees, $N(S) = 174,980$ had a Science or Math major. Thus, if we select a bachelor's degree recipient at random, then the probability that that person is a Science or Math major is $P(\text{Science or Math major}) = N(S)/N = .1216$. As you might expect this probability will change if we know that the person selected is a woman. For these N degrees, $N(F) = 826,264$ were awarded to women and out of these $N(F)$ degrees $N(S \cap F) = 77,380$ had a Science or Math major. Thus, if we know that a woman was selected, then the conditional probability that she has a Science or Math major is
$P(\text{Science or Math major given the person is a woman}) = N(S \cap F)/N(F) = .0936$. We will return to this example after we define some relevant notation. (source: National Center for Education Statistics (NCES))

Consider a population of N objects and two subpopulations (events) A and B. Let $N(A)$ denote the number of objects in subpopulation A and let $N(B)$ denote the number of objects in subpopulation B. If an object is selected at random from this population, then $P(A) = N(A)/N$ and $P(B) = N(B)/N$. Now suppose that it is known that event A has

occurred. That is suppose that we know that the object belongs to subpopulation (event) A. Using subpopulation A as our reference space (new sample space) we note that the event B occurs if and only if the object is the intersection $A \cap B$. Thus letting $N(A \cap B)$ denote the number of objects in the population which belong to subpopulation A and subpopulation B, we see that the conditional probability of the selected object belonging to subpopulation B given that the selected object belongs to subpopulation A is

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

Definition – Conditional probability. *Given events A and B with $P(A) > 0$, the conditional probability of B given A is*

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

Example 6.2 Bachelor's degree majors revisited The $N = 1,439,264$ bachelor's degrees conferred by degree–granting institutions in the US in 2004–2005 are classified by field of major and sex in Table 6.1.

Table 6.1. Bachelor's degrees conferred in 2004-05, by sex and major. (counts)

	sex		
Major	male	female	**total**
Business	162,669	176,358	339,027
Social Sciences	91,533	220,410	311,943
Humanities	80,163	126,431	206,594
Science/Math	97,600	77,380	174,980
Art	61,901	101,006	162,907
Education	34,215	94,124	128,339
Engineering	84,919	30,555	115,474
total	613,000	826,264	1,439,264

Suppose that a person is selected at random from this group of $N = 1,439,264$ bachelor's degree recipients. The (unconditional) probabilities for each major–sex combination,

obtained by dividing the entries in Table 6.1 by N, are given in Table 6.2. The margins of this table contain the (unconditional) probabilities for each major (the column of row probabilities) and for each sex (the row of column probabilities).

Table 6.2. Bachelor's degrees conferred in 2004-05, by sex and major. (proportions)

	sex		
Major	male	female	**total**
Business	0.1130	0.1225	0.2356
Social Sciences	0.0636	0.1531	0.2167
Humanities	0.0557	0.0878	0.1435
Science/Math	0.0678	0.0538	0.1216
Art	0.0430	0.0702	0.1132
Education	0.0238	0.0654	0.0892
Engineering	0.0590	0.0212	0.0802
total	0.4259	0.5741	1

Of these N bachelor's degrees, $N(M) = 613,000$ were awarded to men and $N(F) = 826,264$ were awarded to women. We will use conditional probabilities to explore some differences and similarities in the distributions of the tabulated majors among these groups of men and women. Note that we can use the counts in Table 6.1 or the probabilities (proportions) in Table 6.2 to compute conditional probabilities.

First consider education majors. Let E denote the event that the person selected had an education major. There are $N(E) = 128,339$ education majors and $P(E) = \frac{128,339}{1,439,264} = 0.0892$. Thus, there is a 8.92% chance of selecting an education major. Let F denote the event that the person selected is a woman, and let M denote the event that the person selected is a man. From Table 6.2 we see that $P(F) = 0.5741$ and $P(M) = 0.4259$. Thus, there is a 57.41% chance of selecting a woman and a 42.59% chance of selecting a man. Now consider how the probability of selecting an education major changes when we know the sex of the selected person. In order to compute the desired conditional probabilities we need the counts or probabilities of the intersections $E \cap F$ and $E \cap M$.

First consider education majors among women. In terms of the counts in the "Female" column of Table 6.1 we have $N(E \cap F) = 94,124$ and $N(F) = 826,264$. Thus

$$P(E|F) = \frac{N(E \cap F)}{N(F)} = \frac{94,124}{826,264} = 0.1139.$$

In terms of the probabilities in the "Female" column of Table 6.2 we have $P(E \cap F) = 0.0654$ and $P(F) = 0.5741$. Thus

$$P(E|F) = \frac{0.0654}{0.5741} = 0.1139.$$

Now consider education majors among men. In terms of the counts in the "Male" column of Table 6.1 we have $N(E \cap M) = 34,215$ and $N(M) = 613,000$. Thus

$$P(E|M) = \frac{N(E \cap M)}{N(M)} = \frac{34,215}{613,000} = 0.0558.$$

In terms of the probabilities in the "Male" column of Table 6.2 we have $P(E \cap M) = 00238.$ and $P(M) = 0.4259$. Thus

$$P(E|F) = \frac{0.0238}{0.4259} = 0.0558.$$

As noted earlier, there is a 8.92% chance of selecting an education major. However, if we know that the person selected is a woman, then there is an 11.39% chance that we have selected an education major, and, if we know that the person selected is a man, then there is a 5.58% chance that we have selected an education major. As you probably expected, we see that the conditional probability of selecting an education major given that we have selected a woman, $P(E|F) = 0.1139$, is substantially larger than the conditional probability of selecting an education major given that we have selected a man, $P(E|M) = 0.0558$.

If we performed analogous computations for the other majors we could find the entire conditional distribution of majors among the $826,264$ women and the entire conditional distribution of majors among the $613,000$ men. All we need to do is extract the numbers (counts or proportions) in the "Female" column (respectively "Male" column) of Table 6.1 or 6.2 and normalize them by dividing each by their sum. These conditional distributions are given in Table 6.3.

Note in particular the probabilities for engineering majors. Let G denote the event that the person selected had an engineering major. The unconditional probability of selecting an

engineering major is $P(G) = 0.0802$ and the conditional probabilities are $P(G|F) = 0.0370$ and $P(G|M) = 0.1385$.

Table 6.3 The conditional distributions of majors for each sex.

men		**women**	
major	**probability**	**major**	**probability**
Business	0.2654	Business	0.2134
SocialSciences	0.1493	SocialSciences	0.2668
Humanities	0.1308	Humanities	0.1530
Science/Math	0.1592	Science/Math	0.0936
Art	0.1010	Art	0.1222
Education	0.0558	Education	0.1139
Engineering	0.1385	Engineering	0.0370
total	1.0000	total	0.9999

Given events A and B, with $P(A) > 0$, the conditional probability of B given A is $P(B|A) = P(A \cap B)/P(A)$. Multiplying the conditional probability $P(B|A)$ by $P(A)$ yields the following **multiplication rule for probabilities**.

The multiplication rule. *Given events A and B with $P(A) > 0$,*

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

Note that if $P(A \cap B) > 0$, then $P(A) > 0$ and $P(B) > 0$ and we have
$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$.

The multiplication rule–three events. *Given events A, B, C with $P(A \cap B \cap C) > 0$,*

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

The multiplication rule–several events. *Given events $A_1, \ldots, A_n$ with $P(A_1 \cap A_2 \cap \cdots \cap A_n) > 0$,*

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

The multiplication rule is especially useful for computing the probability of an intersection when the events involved occur sequentially in time.

6.2 Independence

toc

Intuitively we say that the events A and B are **independent** (**stochastically independent**) when knowing that B has occurred has no effect on the probability of occurrence of A, *i.e.* when $P(A) = P(A|B)$. For mathematical convenience the formal definition of independence is in terms of a product so that it does not depend on the existence of conditional probabilities.

Definition – Independence. *The events A and B are said to be independent (stochastically independent) when $P(A \cap B) = P(A)P(B)$.*

Example 6.1 Selecting one card, revisited Recall that we found that the probability of selecting a king and the conditional probability of selecting a king given that the card selected is a spade are both 1/13. Thus, the events K – the card is a king and S – the card is a spade are independent. Note also that $P(K \cap S) = 1/52 = (1/4)(1/13) = P(S)P(K)$.

Example 6.3 Tossing a die If a fair die is tossed once and we let $A = \{2, 4, 6\}$ denote the event that an even value occurs and $B = \{1, 2, 3, 4\}$ the event that the value is four or less, then $P(A) = \frac{1}{2}$, $P(B) = \frac{2}{3}$, and $P(A \cap B) = \frac{1}{3}$. Thus $P(A \cap B) = P(A)P(B)$ and A and B are independent. We can also verify the independence of A and B by noting that $P(A) = 3/6 = 1/2$ and $P(A|B) = 2/4 = 1/2$ so that $P(A|B) = P(A)$.

If two events are disjoint (mutually exclusive), then they cannot occur at the same time; thus if A and B are mutually exclusive, then they cannot be independent unless at least one of them is the null event. That is, if A and B are mutually exclusive, with $P(A) > 0$ and $P(B) > 0$, then $P(A \cap B) = 0$ cannot be equal to $P(A)P(B)$.

In the next result we note some ways in which the independence of two events implies the independence of some associated pairs of events.

Some implications of independence. *If the events A and B are independent, then the events A and B^c are independent, the events A^c and B are independent, and the events A^c and B^c are independent.*

An aside – independence in general and conditional independence

When more than two events are involved the definition of independence is somewhat more complicated. The reason for this complexity is that we need to be sure that all possible subsets of the events exhibit appropriate independence properties. We first define **independence** (**mutual independence**) for several events.

Definition – Independence of several events. *The events $A_1, \ldots, A_n$ are said to be independent (mutually independent) when*
$P(A_i \cap A_j) = P(A_i)P(A_j)$ *for all pairs* (i, j) *with distinct elements*
$P(A_i \cap A_j \cap A_k) = P(A_i)P(A_j)P(A_k)$ *for all triples* (i, j, k) *with distinct elements*
and so on for sets of four, five, ..., up to
$P(A_i \cap \cdots \cap A_n) = P(A_1) \cdots P(A_n)$.

This definition of mutual independence simply says that in order for the events $A_1, \ldots, A_n$ to be independent, every possible subset of the A_i must satisfy the condition that the probability of the intersection of the events in the subset is equal to the product of the probabilities of the individual events which form the subset. In particular, this factorization of the probability of an intersection must hold for every possible pair of events, triple of events, and so on up to the collection of all n events.

There is a weaker type of independence, **pairwise independence**, which arises in some applications. As the name suggests pairwise independence only requires that the factorization property holds for pairs.

Definition – Pairwise independence. *The events $A_1, \ldots, A_n$ are said to be pairwise independent when*
$P(A_i \cap A_j) = P(A_i)P(A_j)$ *for all pairs* (i, j) *with* $i \neq j$.

Example 6.4 Independent events Let $P(\omega) = 1/8$ for $\omega \in \Omega = \{1, 2, 3, 4, 5, 6, 7, 8\}$, let $A = \{1, 2, 3, 4\}$, $B = \{1, 2, 5, 6\}$, and $C = \{1, 3, 5, 7\}$. Then $P(A) = P(B) = P(C) = \frac{1}{2}$, $P(A \cap B) = P(A \cap C) = P(B \cap C) = \frac{1}{4} = (\frac{1}{2})^2$, and $P(A \cap B \cap C) = \frac{1}{8} = (\frac{1}{2})^3$. Thus, in this example, the events A, B, and C are independent (mutually independent).

Example 6.5 Pairwise independent events that are not independent Let $P(\omega) = 1/8$ for $\omega \in \Omega = \{1,2,3,4,5,6,7,8\}$, let $A = \{1,2,3,4\}$, $B = \{1,2,5,6\}$, and $C = \{1,2,7,8\}$. Then $P(A) = P(B) = P(C) = \frac{1}{2}$ and $P(A \cap B) = P(A \cap C) = P(B \cap C) = \frac{1}{4} = (\frac{1}{2})^2$. But $P(A \cap B \cap C) = \frac{1}{4} \neq (\frac{1}{2})^3$. Thus, in this example, the events A, B, and C are pairwise independent but not mutually independent.

Definition – Conditional independence. *Given events A, B, and C with $P(A \cap B \cap C) > 0$, the events A and B are said to be conditionally independent given the event C when $P(A \cap B|C) = P(A|C)P(B|C)$.*

6.3 The law of total probability – Bayes' theorem

toc

In some situations we may find it convenient to compute the probability of an event by first decomposing the event into disjoint subevents and then adding the probabilities of these subevents. For example, given events A and B, we can partition the sample space as $\Omega = B \cup B^c$ and the event A as $A = (A \cap B) \cup (A \cap B^c)$. Since B and B^c are disjoint, $A \cap B$ and $A \cap B^c$ are also disjoint. This gives the decomposition of the probability of A as $P(A) = P(A \cap B) + P(A \cap B^c)$. This result is the simplest case of the following law of total probability.

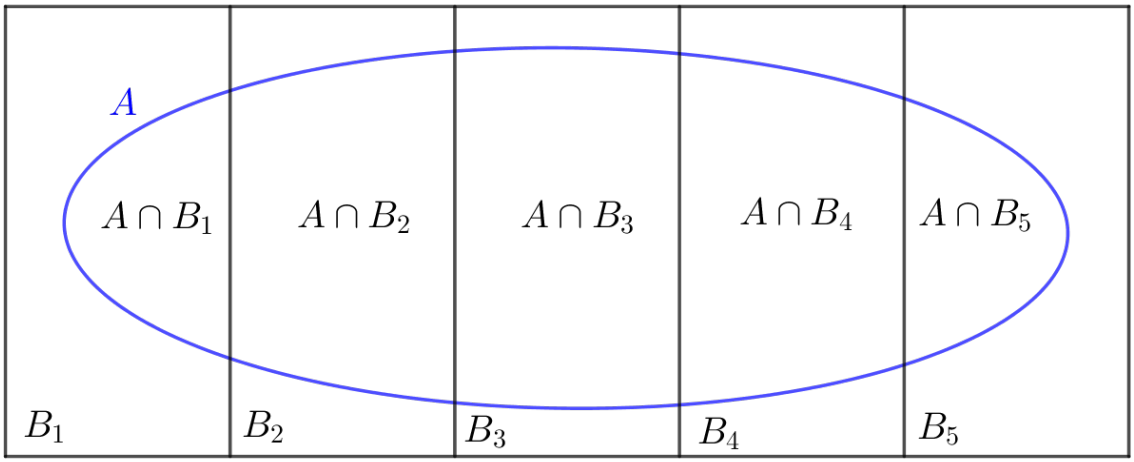
The law of total probability. *Refer to Figure 6.1 for an illustration of this result. If the events $B_1, \ldots, B_n$ form a partition of Ω, i.e. if $B_1, \ldots, B_n$ are disjoint and $\Omega = B_1 \cup \cdots \cup B_n$, then, for any event A,*

$$\begin{aligned} P(A) &= P(A \cap B_1) + P(A \cap B_2) + \cdots + P(A \cap B_n) \\ &= \sum_{i=1}^{n} P(A \cap B_i). \end{aligned}$$

Alternate statement of the law of total probability. *If the events $B_1, \ldots, B_n$ form a partition of Ω, and $P(B_i) > 0$ for $i = 1, \ldots, n$, then*

$$\begin{aligned} P(A) &= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_n)P(B_n) \\ &= \sum_{i=1}^{n} P(A|B_i)P(B_i). \end{aligned}$$

Figure 6.1 The law of total probability. The ellipse represents event A and the rectangular sections represent the events $B_1, \ldots, B_5$ of the partition.



Here we see that Ω is partitioned into 5 disjoint parts, A is similarly partitioned, and the probabilities of these part are added.

$$P(A) = P(A \cap B_1) + \cdots + P(A \cap B_5)$$

$$P(A) = P(A|B_1)P(B_1) + \cdots + P(A|B_5)P(B_5)$$

We will now provide a variation of the law of total probability, called Bayes' theorem, which allows use to compute conditional probabilities of events at the first stage of an experiment given the outcome at the second stage.

Bayes' theorem. *If the events $B_1, \ldots, B_n$ form a partition of Ω and $P(B_i) > 0$ for $i = 1, \ldots, n$, then for any event A with $P(A) > 0$, we have*

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_n)P(B_n)}.$$

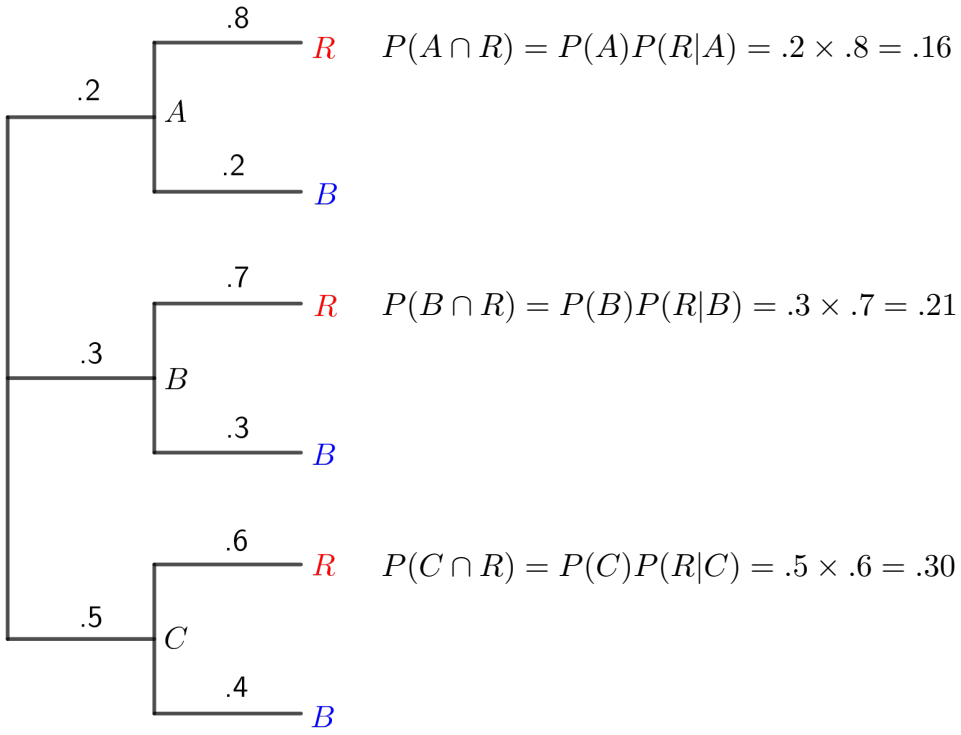
Bayes' theorem is particularly useful for a situation where the occurrence of event A follows the occurrence of one of the events B_i in time and we are interested in the conditional probability that a particular B_i, say B_1, has occurred given that event A has occurred.

Note that if A and B are events and $0 < P(B) < 1$, then the events B and B^c form a partition of Ω. Thus $P(A) = P(A \cap B) + P(A \cap B^c)$ and Bayes' theorem reduces to

$$P(B|A) = \frac{P(A \cap B)}{P(A \cap B) + P(A \cap B^c)} = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}.$$

Example 6.6 Balls from a box Consider a box containing 100 balls of which 20 are labeled A, 30 are labeled B, and 50 are labeled C, and three other boxes labeled A, B, and C such that: box A contains 8 red and 2 blue balls; box B contains 7 red and 3 blue balls; and, box C contains 6 red and 4 blue balls. Now suppose that a ball is chosen at random from the box containing 100 balls, the letter (A, B, or C) on the ball is noted, and then a ball is chosen at random from the 10 balls in the box with the appropriate letter label. Clearly, the conditional probabilities of choosing a red ball given the letter label are: $P(R|A) = .8$, $P(R|B) = .7$, and $P(R|C) = .6$. It is also clear that the probabilities of selecting the label (A, B, or C) are: $P(A) = .2$, $P(B) = .3$, and $P(C) = .5$. Thus, $P(A \cap R) = P(R|A)P(A) = .16$, $P(B \cap R) = P(R|B)P(B) = .21$, and $P(C \cap R) = P(R|C)P(C) = .30$ are obtained by multiplication of the probabilities along the appropriate path in the tree diagram as indicated in Figure 6.2. The law of total probability says that the probability of choosing a red ball $P(R)$ is obtained, as indicated at the bottom of the figure, by summing these three probabilities.

Figure 6.2 The law of total probability example.



$$P(R) = P(R|A)P(A) + P(R|B)P(B) + P(R|C)P(C) = .16 + .21 + .30 = .67$$

The values of conditional probabilities of the form $P(A|R)$, the conditional probability that the ball was selected from box A given that it was red, are less obvious than the conditional probabilities of selecting red from a specified box. However, these conditional probabilities are readily computed using Bayes' Theorem. Here

$$P(R) = P(R|A)P(A) + P(R|B)P(B) + P(R|C)P(C) = .16 + .21 + .30 = .67$$

and we have

$$P(A|R) = \frac{P(R \cap A)}{P(R)} = \frac{P(R|A)P(A)}{P(R)} = \frac{16}{67} \approx .24$$

$$P(B|R) = \frac{P(R \cap B)}{P(R)} = \frac{P(R|B)P(B)}{P(R)} = \frac{21}{67} \approx .31$$

$$P(C|R) = \frac{P(R \cap C)}{P(R)} = \frac{P(R|C)P(C)}{P(R)} = \frac{30}{67} \approx .45$$

Notice that each of these conditional probabilities is of the form $\frac{a}{a+b+c}$, where a, b, and c are the three probabilities summed to get $P(R)$ in the tree diagram of Figure 6.2.

It is interesting to compare the unconditional probabilities of drawing from boxes A, B, and C, $P(A) = .2$, $P(B) = .3$, and $P(C) = .5$, to the corresponding conditional probabilities given that the ball drawn is known to be red, $P(A|R) = \frac{16}{67} \approx .24$, $P(B|R) = \frac{21}{67} \approx .31$, and $P(C|R) = \frac{30}{67} \approx .45$. The initial probabilities (before we obtain the additional information that the ball drawn was red) are known as **prior probabilities** and the updated probabilities (conditional on the added information) are known as **posterior probabilities**. In this example, the conditional probability of drawing a red ball is highest when drawing from box A and lowest when drawing from box C. Thus, the added information that the selected ball was red increases the likelihood that the ball came from box A and decreases the likelihood that the ball came from box C.

7 Discrete random variables toc

7.1 Random variables toc

In some of the examples we have considered the sample space is a set of integers, *e.g.*, $\Omega = \{1, 2, 3, 4, 5, 6\}$ corresponding to one toss of a die. In other examples we restricted our attention to events described in terms of a numerical value, *e.g.*, the number of heads in n tosses of a coin. We will now consider a more formal treatment of such assignments of numerical values to the outcomes of an experiment.

A function which assigns numerical values (real numbers) to the elements of a sample space is known as a **random variable** (denoted **r.v.**). Given an experiment with sample space Ω, a random variable associates a numerical value with each elementary outcome (element) ω of the sample space Ω. In the term random variable: the word variable indicates that the values of the function are numbers which are assigned to elementary outcomes; and, the adjective random is used as it is used in random experiment to indicate that the value of the random variable or outcome of the experiment is not known with certainty before the experiment is conducted at which time the value of the random variable or outcome of the experiment is determined.

Given a random experiment with sample space Ω and a random variable X defined on Ω, the r.v. X defines a new sample space Ω_X comprised of all of the possible values of X. An event defined in terms of X (a subset of Ω_X) can be identified with the equivalent event from the underlying experiment (a subset of Ω). For example, the event $X = 2$ corresponds to the event consisting of all elements ω of the underlying sample space Ω which are assigned the value $X = 2$. More formally, the event $[X = x]$ (a subset of Ω_X) corresponds to the event $\{\omega \in \Omega : X(\omega) = x\}$ (a subset of Ω). In this last expression we are using the function notation $X(\omega)$ to indicate the value of X assigned to the elementary outcome ω.

Recall that a set is said to be discrete if it contains a finite or countably infinite number of elements. That is, the elements of a discrete set can be arranged in a list of the form $x_1, x_2, \ldots, x_N$, if the set is finite, or $x_1, x_2, \ldots$ if the set is infinite. If Ω_X is discrete, then X is said to be a **discrete random variable**. In this chapter we will restrict our attention to discrete random variables.

Probabilities of events defined in terms of the random variable X are determined by finding the probability of the equivalent event in terms of the underlying experiment. That is, given a random variable X and an event A (a subset of Ω_X), the probability that X takes on a value in A, $P(X \in A)$, is found by adding the probabilities of each of the elementary outcomes in Ω for which the corresponding X values are in A.

We will use a simple example to clarify the relationship between Ω and Ω_X, the relationship between events defined in terms of X and events of the original experiment, and the computation of probabilities of events defined in terms of X.

Example 7.1 Tossing a coin. Suppose we toss a coin three times. Letting H denote "heads", T denote "tails", and using 3–tuples of the letters H and T to represent the elementary outcomes, the sample space of the experiment is

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

Now let X denote the random variable "the number of heads in the three tosses". The sample space for this X is

$$\Omega_X = \{0, 1, 2, 3\}.$$

As noted above, an event defined in terms of X is a subset of Ω_X and each such event can be identified with an equivalent event from the underlying experiment. For example:
The event $[X = 0]$ is equivalent to the event $\{TTT\}$;
The event $[X = 1]$ is equivalent to the event $\{HTT, THT, TTH\}$;
The event $[X = 2]$ is equivalent to the event $\{HHT, HTH, THH\}$;
The event $[X = 3]$ is equivalent to the event $\{HHH\}$; and,
The event $[X \leq 2]$ is equivalent to the event $\{TTT, HTT, THT, TTH, HHT, HTH, THH\}$.

If we assume that the coin we are tossing is fair so that each of the eight elementary outcomes in Ω has probability 1/8, then
$P(X = 0) = P(\{TTT\}) = 1/8$;
$P(X = 1) = P(\{HTT, THT, TTH\}) = 3/8$;
$P(X = 2) = P(\{HHT, HTH, THH\}) = 3/8$;
$P(X = 3) = P(\{HHH\}) = 1/8$; and,
$P(X \leq 2) = P(\{TTT, HTT, THT, TTH, HHT, HTH, THH\}) = 7/8$.

We can characterize the distribution of the discrete r.v. X by specifying a **probability mass function** (denoted p.m.f.) p_X. The probability mass function p_X assigns a probability to each potential value of the random variable X, *i.e.*,

$$p_X(x) = P(X = x).$$

The p.m.f. assigns positive probabilities to the elements of Ω_X and zero probabilities to values not in Ω_X. Furthermore, since all of the values of X which have positive probability of occurrence belong to Ω_X, the probabilities of the elements of Ω_X must sum to one. The requisite properties of a p.m.f. are provided in the following definition.

Definition: probability mass function. *Given a discrete sample space* $\Omega_X = \{x_1, \ldots, x_N\}$ *any function* p_X *with the properties that:*
(1) $p_X(x) > 0$ *for all* $x \in \{x_1, \ldots, x_N\}$*;*
(2) $p_X(x) = 0$ *for all* $x \notin \{x_1, \ldots, x_N\}$*; and,*
(3) $p_X(x_1) + \cdots + p_X(x_N) = 1$
can be viewed as the probability mass function (p.m.f.) of a random variable X *with sample space* Ω_X*. The obvious modifications apply when* $\Omega_X = \{x_1, x_2, \ldots\}$ *is countably infinite.*

For any event A defined in terms of the r.v. X, *i.e.*, for any $A \subset \Omega_X$, the probability of event A is equal to the sum of the probabilities of each of the distinct values of X which form the event (the sum of the $p_X(x)$ corresponding to each element $x \in A$). In symbols, the probability of A is

$$P(A) = \sum_{x \in A} p_X(x).$$

We will now return to the coin tossing example to clarify this definition of a probability mass function.

Example 7.1 Tossing a coin, revisited. As noted above, assuming that the coin we are tossing is fair so that each of the eight elementary outcomes in Ω has probability 1/8, $P(X = 0) = 1/8$, $P(X = 1) = 3/8$, $P(X = 2) = 3/8$, and $P(X = 3) = 1/8$. Hence, in function notation, X has probability mass function

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

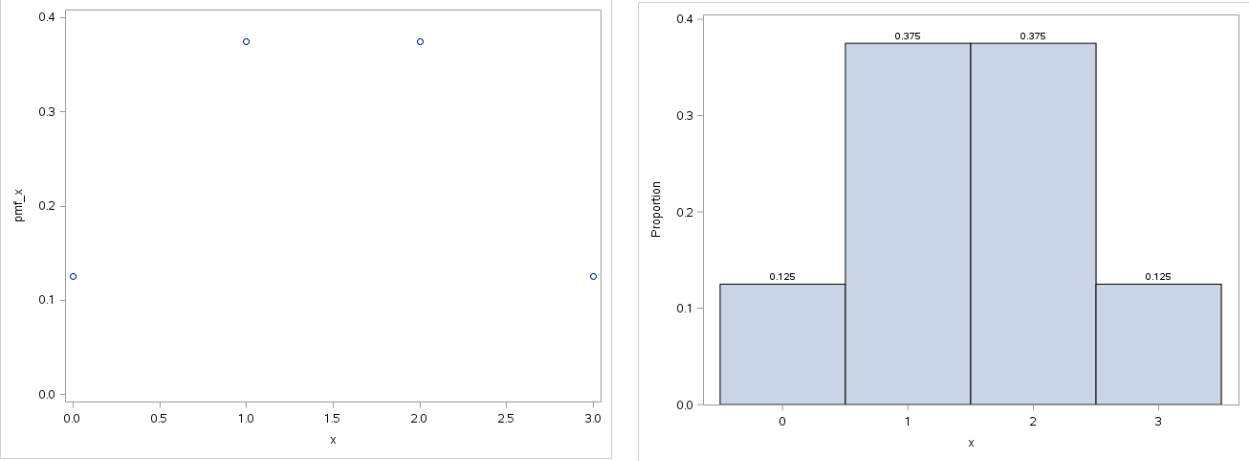
We can also express this p.m.f. in tabular form as shown in Table 7.1.

Table 7.1 The p.m.f. of X = the number of heads in 3 tosses of a fair coin.

x	$p_X(x)$
0	1/8
1	3/8
2	3/8
3	1/8

Note that observing the value of X is equivalent to selecting a ball at random from a box containing eight balls of which 1 is numbered zero, 3 are numbered one, 3 are numbered two, and 1 is numbered three. This is an example of a binomial distribution. This binomial p.m.f. is shown in the graph on the left in Figure 7.1; the graph on the right represents this distribution as a probability histogram. The probability histogram representation allows us to easily visualize the probability of an event as the sum of the areas of the rectangles over the values of X which comprise the event.

Figure 7.1 Binomial distribution $n = 3$ $p = .5$ – p.m.f. and probability histogram.



The distribution of X can also be characterized in terms of its **cumulative distribution function** (denoted c.d.f.) F_X, where

$$F_X(x) = P(X \leq x).$$

For discrete random variables the p.m.f. usually gives the more convenient characterization of the distribution. The c.d.f. is useful for computing probabilities.

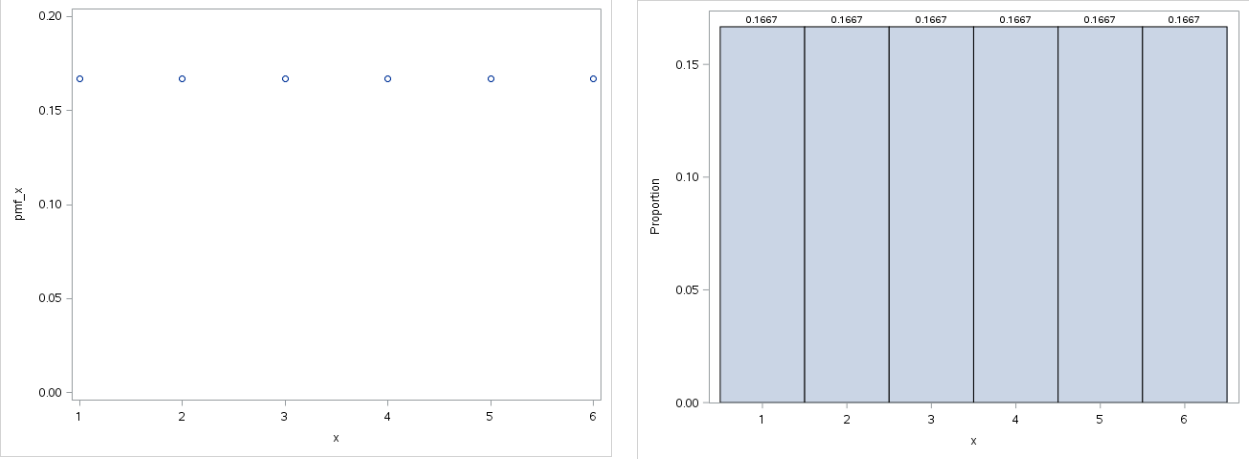
7.2 Some examples toc

Example 7.2 Tossing a die – uniform distribution. Suppose that a fair (balanced) die is tossed and let X denote the number on the upturned face. The sample space for X is $\Omega_X = \{1, 2, 3, 4, 5, 6\}$. Since the die is assumed to be fair we will assume that these 6 outcomes are equally probable. (Note that tossing a fair die once is equivalent to selecting a ball at random from a box containing six balls numbered from one to six.) Thus, the p.m.f. of X is given by

$$p_X(x) = \begin{cases} 1/6 & \text{if } x \in \{1, 2, 3, 4, 5, 6\} \\ 0 & \text{otherwise} \end{cases} .$$

This is an example of a uniform distribution on a finite set of integers. This uniform p.m.f. is shown in the graph on the left in Figure 7.2; the graph on the right represents this distribution as a probability histogram.

Figure 7.2 Uniform distribution on $\{1, 2, 3, 4, 5, 6\}$ – p.m.f. and probability histogram.



Example 7.3 The sum when a die is tossed twice – triangular distribution. Suppose that a fair (balanced) die is tossed twice and let X denote the sum of the numbers observed. The elementary outcomes for this experiment can be represented by the 36 ordered pairs of the form (a, b) where $a, b \in \{1, 2, 3, 4, 5, 6\}$. The sums and their relationship with these elementary outcomes are shown in Table 7.2. Here the row label represents the number on the first toss, the column label represents the number on the second toss, and the value in the body of the table is the corresponding sum. Because the sums range from 2 to 12 the sample space for X is $\Omega_X = \{2, 3, \ldots, 12\}$.

Table 7.2 Sums when a die is tossed twice.

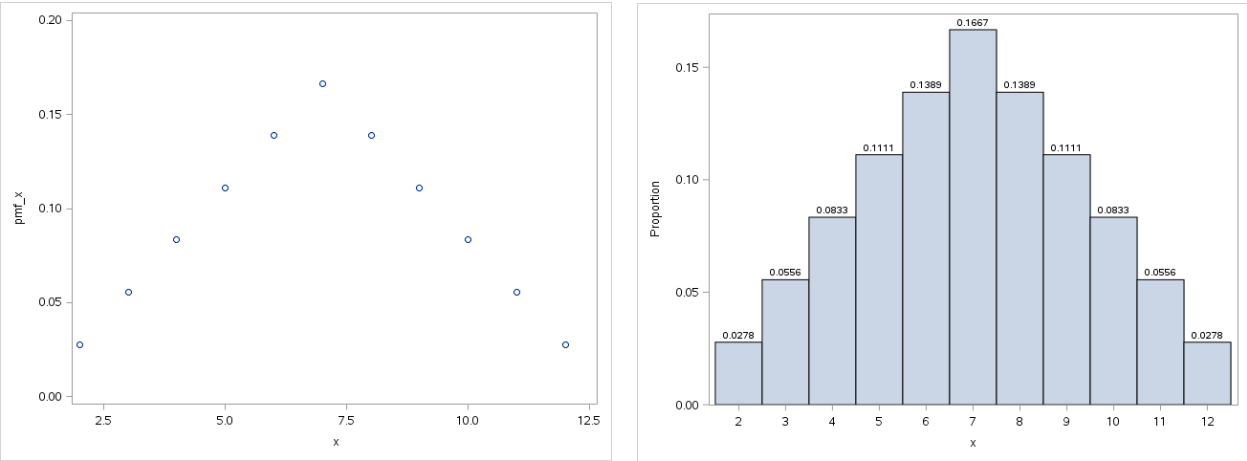
	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Since the die is assumed to be fair we will assume that these 36 elementary outcomes are equally probable. Counting the number of favorable outcomes for each case we find that the p.m.f. of X is given by

$$p_X(x) = \begin{cases} 1/36 & \text{if } x = 2 \text{ or } x = 12 \\ 2/36 & \text{if } x = 3 \text{ or } x = 11 \\ 3/36 & \text{if } x = 4 \text{ or } x = 10 \\ 4/36 & \text{if } x = 5 \text{ or } x = 9 \\ 5/36 & \text{if } x = 6 \text{ or } x = 8 \\ 6/36 & \text{if } x = 7 \\ 0 & \text{otherwise} \end{cases} .$$

In this example, we can think of selecting a ball at random from a box containing 36 balls of which 1 is numbered two, 2 are numbered three, 3 are numbered four, and so on. This p.m.f. is shown in the graph on the left in Figure 7.3; the graph on the right represents this distribution as a probability histogram. This is an example of a triangular distribution on a finite set of integers.

Figure 7.3 Distribution of sum of two fair dice – p.m.f. and probability histogram. (Triangular distribution on $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$)



7.3 The binomial distribution

toc

In Section 5.4 we introduced the binomial distribution as the distribution of the number of objects of one type in a sample selected at random with replacement. We also discussed two typical applications of the binomial distribution as the distribution of a random variable X denoting the number of objects of a specified type in a sample. In Example 5.8 X denoted the number of aces observed when a fair die was tossed 10 times. In Example 5.9 X denoted the number of red balls in a sample selected at random with replacement from a population containing 100 red balls and 500 blue balls.

We will now provide a slightly different description of the binomial distribution. Consider an experiment consisting of a sequence of n independent, dichotomous trials, where a trial is a process of observation or experimentation which results in the occurrence of one of two possible outcomes. The two possible outcomes are generically known as "success" and "failure". There is no connotation of "goodness" associated with the term "success", this term is simply used to indicate a success in the sense that the outcome of interest has occurred. In Example 5.8 a trial is a toss of the die and tossing an ace (one) is a success. In Example 5.9 a trial is the selection of a ball from the population and obtaining a red ball is a success. The success probability p is the probability of observing a success on a single trial and $1-p$ is the probability of observing a failure on a single trial. The random variable of interest X is the number of successes observed in the sequence of n independent trials.

A sequence of independent Bernoulli trials

Given a positive integer n and a probability p $(0 < p < 1)$, a sequence of independent Bernoulli trials with success probability p is a sequence of n dichotomous (success or failure) trials with the properties that:

1. On each trial the probability of success is p; and,
2. The outcomes of the trials are independent.

A binomial r.v. X denotes the number of successes in a sequence of n independent Bernoulli trials with success probability p.

The binomial probability mass function

Given a positive integer n and a probability p $(0 < p < 1)$, the binomial random variable X denotes the number of successes in a sequence of n independent Bernoulli trials with success probability p. For $x = 0, 1, \ldots, n$, the **binomial probability mass function** is of the form

$$p_X(x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

Example 7.4 Number of red balls in a sample. Suppose we select a random sample of $n = 4$ balls with replacement from a box containing 60 balls of which 10 are red. We can think of the selection of a single ball from the box as a trial. Since the balls are being selected with replacement the outcomes of the four trials are independent and the probability of obtaining a red ball is 1/6 for each selection, we can view these four selections of a ball as forming a sequence of $n = 4$ Bernoulli trials with success probability $p = 1/6$. Hence, letting X denote the number of red balls in the four selections, X is the binomial random variable with p.m.f.

$$p_X(x) = \binom{4}{x} \left(\frac{1}{6}\right)^x \left(\frac{5}{6}\right)^{4-x} \text{ for } x = 0, 1, 2, 3, 4.$$

Example 7.5 Number of aces when tossing a fair die. Suppose a fair (balanced) die is tossed $n = 4$ times and the number of aces (ones) is determined. Since the outcomes of these four tosses are independent and the probability of obtaining an ace is 1/6 on each toss, we can view these four tosses as forming a sequence of $n = 4$ Bernoulli trials with success probability $p = 1/6$. Hence, letting X denote the number of aces in the four tosses, X is the binomial random variable with p.m.f.

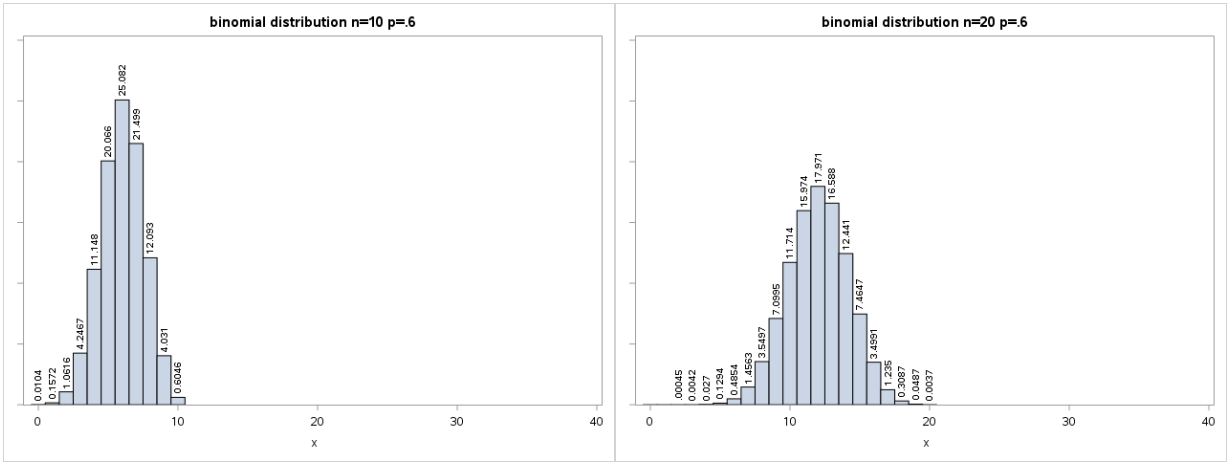
$$p_X(x) = \binom{4}{x} \left(\frac{1}{6}\right)^x \left(\frac{5}{6}\right)^{4-x} \text{ for } x = 0, 1, 2, 3, 4.$$

Some properties of binomial distributions are illustrated in Figures 7.4 and 7.5. Figure 7.4 shows that the binomial distribution is centered at np and the variability in the distribution increases as the sample size n increases. In this figure the success probability is fixed at $p = .6$ and histograms are provided for several sample sizes ($n = 10, 20, 30,$ and 40). Figure 7.5 shows how the shape of the binomial distribution and the variability in the distribution depend on the value of the success probability p. In this figure the sample size

is fixed at $n = 20$ and histograms are provided for several values of the success probability ($p = .5, .6, .7, .8$, and $.9$). When $p = .5$, the distribution is exactly symmetric about $np = 10$. When $p \neq .5$ the binomial distribution is skewed and as the value of p moves away from .5 the distribution becomes more skewed. In this figure the values $p = .6, .7, .8$, and .9 are greater than .5 and the distributions are skewed left. if we had used values less than .5 the distributions would be skewed right. Notice that in addition to the increase in skewness there is a decrease in variability as p moves away from .5.

Figure 7.4 Binomial distributions for $p = .6$. Histograms are provided for $n = 10, 20, 30$, and 40. These distributions are centered at np for $p = .6$ and the appropriate value of n. Notice how the variability in the distribution increases as n increases.

$n = 10$ $n = 20$



$n = 30$ $n = 40$

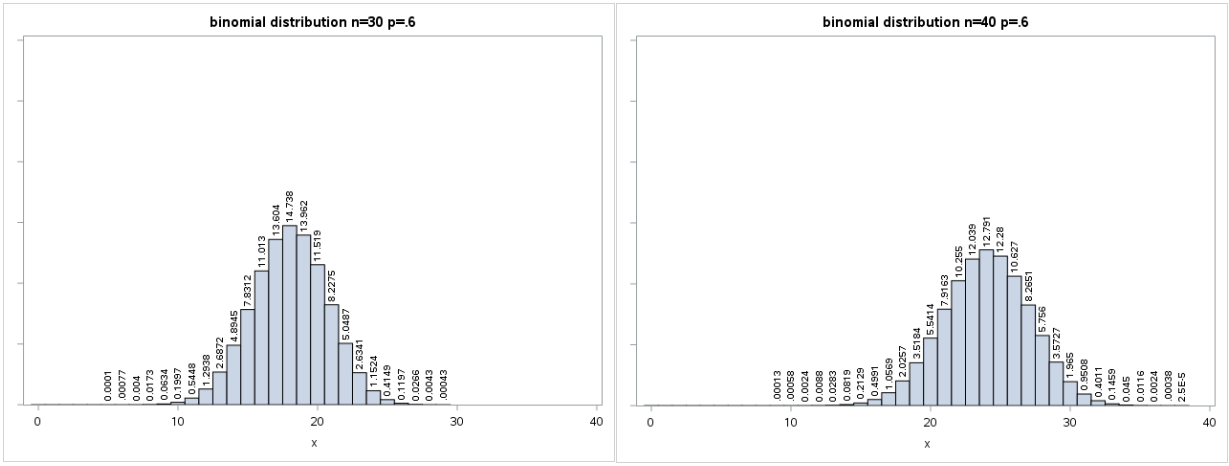
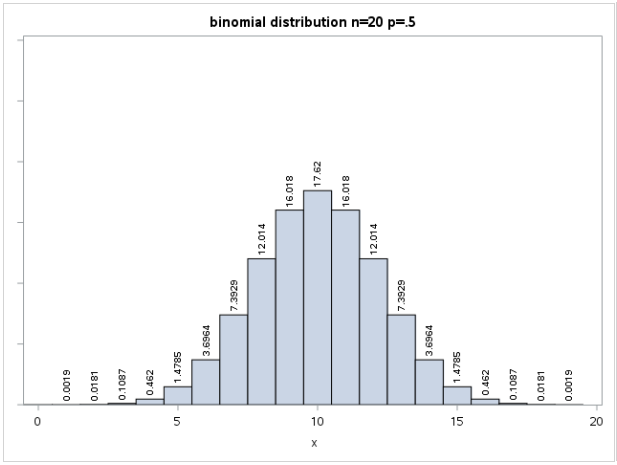
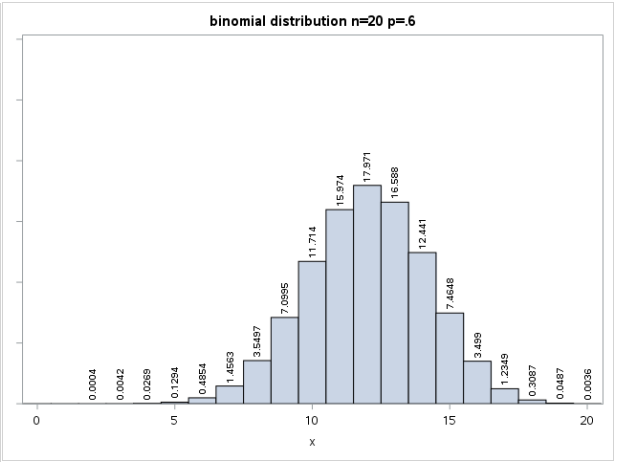


Figure 7.5 Binomial distributions for $n = 20$. Histograms for $p = .5, .6, .7, .8$, and .9. Each distribution is centered at $20p$ for the appropriate p. Notice how the variability in the distribution decreases as p moves away from .5. Notice also how the distribution becomes more skewed left as p moves away from .5.

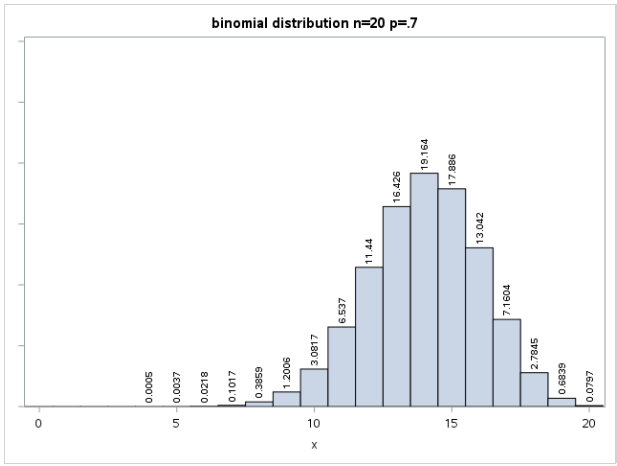
$p = .5$



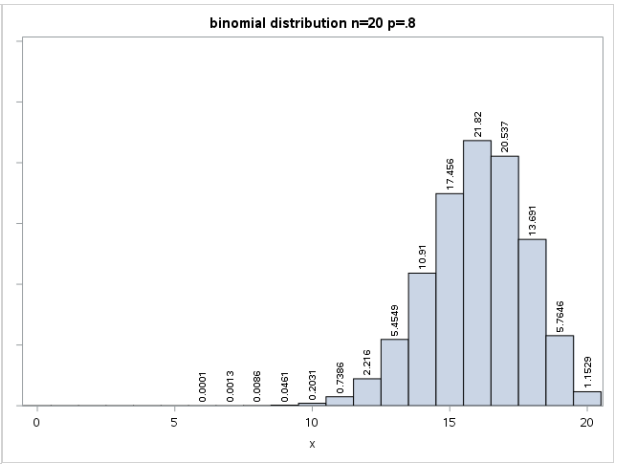
$p = .6$



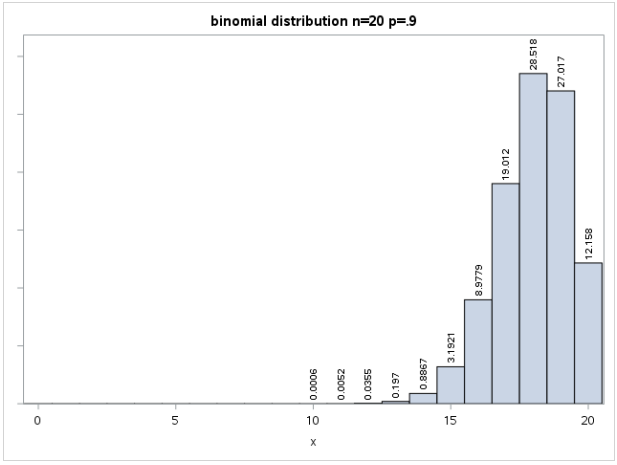
$p = .7$



$p = .8$



$p = .9$



When $n = 1$ our sequence of trials reduces to a single Bernoulli trial with $X = 1$ indicating that a success has occurred and $X = 0$ indicating a failure. The binomial distribution for the case when $n = 1$ is also know as a **Bernoulli distribution**. We will now consider an interesting application.

Example 7.6 Mendelian inheritance models. In his investigations, during the years 1856 to 1868, of the chromosomal theory of inheritance Gregor Mendel performed a series of experiments on ordinary garden peas. One characteristic of garden peas that Mendel studied was the color of the flowers (red or white). When Mendel crossed a plant with red flowers with a plant with white flowers, the resulting offspring all had red flowers. But when he crossed two of these first generation plants, he observed plants with white as well as red flowers.

The gene which determines the color of the flower occurs in two forms (alleles). Let R denote the allele for red flowers (which is dominant) and r denote the allele for white flowers (which is recessive). When two plants are crossed the offspring receives one allele from each parent, thus there are four possible genotypes (combinations) RR, Rr, rR, and rr. The three genotypes RR, Rr, and rR, which include the dominant R allele, will yield red flowers while the fourth genotype rr will yield white flowers. If a red flowered RR genotype parent is crossed with a white flowered rr genotype parent, then all of the offspring will have genotype Rr and will produce red flowers. The basic Mendelian inheritance model assumes that a pair of alleles is formed by randomly choosing one allele from each parent. Under this model, if two of these first generation Rr genotype plants are crossed, each of the four possible genotypes RR, Rr, rR, and rr is equally likely and plants with white as well as red flowers will occur. Under this simple model, with each of the four genotypes having the same probability of occurring, the probability that a plant will have red flowers is $P(\text{red}) = 3/4$ and the probability that a plant will have white flowers is $P(\text{white}) = 1/4$.

The distribution of the r.v. X denoting flower color, with $X = 1$ indicating the plant has red flowers and $X = 0$ indicating the plant has white flowers, follows the Bernoulli distribution with $p = 3/4$.

We could test the validity of this model, as Mendel did, by crossing n pairs of peas plants and determining the color of the flowers for these n crosses. If we let X denote the number of these n crosses which result in a red flowered offspring, then X will follow the binomial distribution with parameters n and p, where, according to Mendel's model $p = 3/4$.

7.4 The hypergeometric distribution

toc

In Section 5.5 we introduced the hypergeometric distribution distribution as the distribution of the number of objects of one type in a sample selected at random without replacement. We discussed a typical application of the hypergeometric distribution as the distribution of a random variable X denoting the number of objects of a specified type in a sample. In Example 5.10 X denoted the number of aces in a five card poker hand. For ease of reference the hypergeometric probability mass function is given below. Here we have simplified the notation compared to that of Section 5.5.

The hypergeometric probability mass function

Given positive integers n, A, and B with $n \leq A+B$, the hypergeometric random variable X denotes the number of successes in random sample of size n selected without replacement from a population containing $A + B$ objects of which A are classified as successes and B are classified as failures. Subject to some restrictions noted below, for $x = 0, 1, \ldots, n$, the **hypergeometric probability mass function** is of the form

$$p_X(x) = \frac{\binom{A}{x}\binom{B}{n-x}}{\binom{A+B}{n}}.$$

Notice that the values of A and B may impose restrictions on the values of x for which this expression works. Specifically, we must have $x \leq A$ and $n - x \leq B$.

Example 7.7 The number of aces in a poker hand. Suppose that a five card poker hand is dealt from a well–shuffled deck. Dealing a five card hand in this way is equivalent to selecting five balls (cards) at random from a box containing 52 balls (cards) labeled so that they represent the cards in the deck. Let X denote the number of aces in the hand. In this application the 52 cards are partitioned into the set of $A = 4$ aces and the set of $B = 48$ non–aces. For $x = 0, 1, 2, 3, 4$, the p.m.f. for this X is

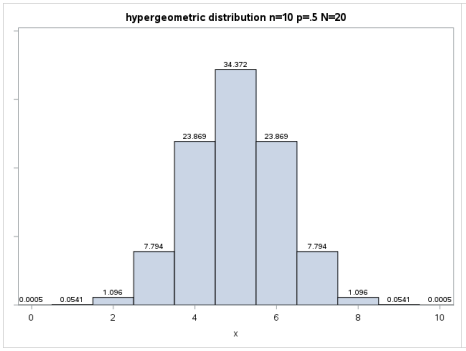
$$p_X(x) = \frac{\binom{4}{x}\binom{48}{5-x}}{\binom{52}{5}}.$$

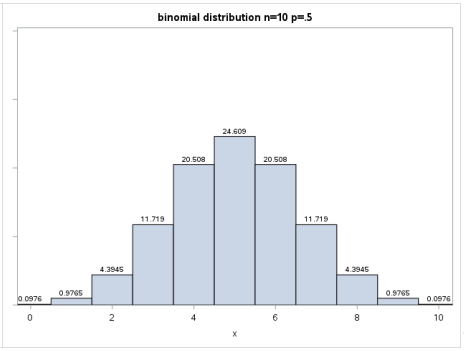
For example, the probability of getting exactly 4 aces is

$$p_X(4) = \frac{\binom{4}{4}\binom{48}{1}}{\binom{52}{5}} = \frac{1 \cdot 48}{2598960} \approx .00002.$$

Figure 7.6 Hypergeometric and binomial distributions compared, with $n = 10$, $p = .5$. Left column: hypergeometric histograms for $N = A + B = 20, 50, 200$, and 500. Middle column: binomial histogram for comparison. Right column: probabilities.

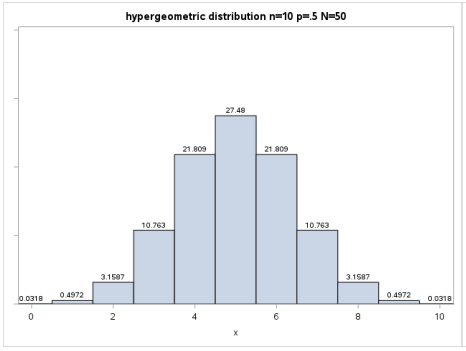
$N = 20$ binomial probabilities

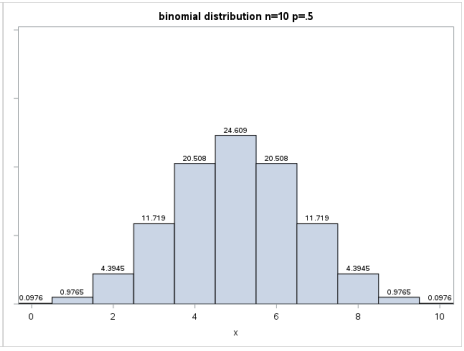




x	hyper	bino
0	0.0000	0.0010
1	0.0005	0.0098
2	0.0110	0.0439
3	0.0779	0.1172
4	0.2387	0.2051
5	0.3437	0.2461
6	0.2387	0.2051
7	0.0779	0.1172
8	0.0110	0.0439
9	0.0005	0.0098
10	0.0000	0.0010

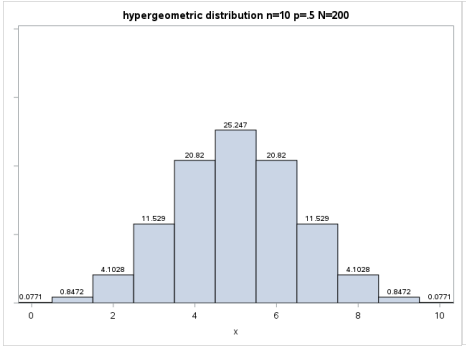
$N = 50$ binomial probabilities

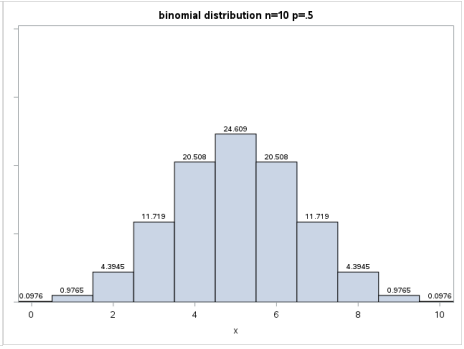




x	hyper	bino
0	0.0003	0.0010
1	0.0050	0.0098
2	0.0316	0.0439
3	0.1076	0.1172
4	0.2181	0.2051
5	0.2748	0.2461
6	0.2181	0.2051
7	0.1076	0.1172
8	0.0316	0.0439
9	0.0050	0.0098
10	0.0003	0.0010

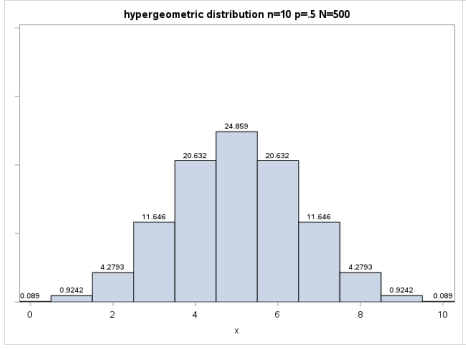
$N = 200$ binomial probabilities

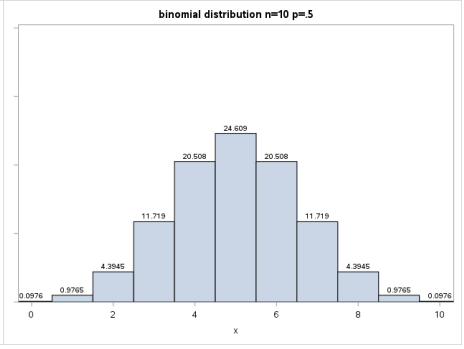




x	hyper	bino
0	0.0008	0.0010
1	0.0085	0.0098
2	0.0410	0.0439
3	0.1153	0.1172
4	0.2082	0.2051
5	0.2525	0.2461
6	0.2082	0.2051
7	0.1153	0.1172
8	0.0410	0.0435
9	0.0085	0.0098
10	0.0008	0.0010

$N = 500$ binomial probabilities





x	hyper	bino
0	0.0009	0.0010
1	0.0092	0.0098
2	0.0428	0.0439
3	0.1165	0.1172
4	0.2063	0.2051
5	0.2486	0.2461
6	0.2063	0.2051
7	0.1165	0.1172
8	0.0428	0.0439
9	0.0092	0.0098
10	0.0009	0.0010

For given values of the sample size n and the success proportion p, there is more variability in a hypergeometric distribution than a binomial distribution. The pattern of dependence of the basic shape of hypergeometric distributions on the values on n and p are similar to those for binomial distributions. In particular, when $p = 1/2$ the hypergeometric distribution is symmetric, when $p < 1/2$ it is skewed right, and when $p > 1/2$ it is skewed left. When the population size $N = A + B$ is large, the hypergeometric and binomial distribution, for given values of n and p, are very similar. This relationship between the hypergeometric and binomial distribution is illustrated, graphically and numerically, for $n = 10$, $p = 1/2$ in Figure 7.6.

7.5 The geometric distribution

toc

The binomial and hypergeometric distributions can be used to model the distribution of the number of red balls (successes) in a sample of fixed size. We will now explore a variation on this sampling procedure where balls are sampled, one at a time with replacement, until a ball of the specified color (a success) is obtained.

Consider a box containing 20 balls of which 5 are red and 15 are blue. First a single ball is selected at random from the 20 balls in the box and its color is determined. If the ball is red we stop sampling. On the other hand, if the ball is blue, it is returned to the box and this procedure is repeated until a red ball is obtained. Let X denote the number of the trial (draw) on which the red ball is obtained. For example, $X = 1$ means we got a red ball on the first draw and $X = 10$ means we got a blue ball on each of the first 9 draws and then a red ball on the tenth draw. Notice that there is no upper limit on how many draws might be needed to obtain a red ball so that, in this situation, the sample space for X is the positive integers (the counting numbers) $1, 2, 3, \ldots$

We can view the results of this sampling process as forming a, potentially infinite, sequence of trials, where a trial is the selection of a ball from the box. In the present context, since we are sampling with replacement, the outcomes of the trials (draws) are independent and on every trial the probability of selecting a red ball is $p = \frac{5}{20} = \frac{1}{4}$. In other words, we are considering a potentially infinite sequence of independent Bernoulli trials with success probability $p = \frac{1}{4}$ which ends when the first success is obtained. In this context, using R to denote a red ball and G to denote a blue ball, the results of the sequence of trials (the elementary outcomes) can be represented by sequences of the form R, BR, BBR, ... Notice that, letting x denote the number of the trial on which the red ball was obtained,

these elementary outcome sequences contain $x-1$ $B's$ followed by a single R. Notice also that, since the outcomes of the trials are independent, the probability of observing $x-1$ $B's$ ($x-1$ blue balls) followed by 1 R (one red ball) is

$$p_X(x) = \left(\frac{15}{20}\right)^{x-1}\left(\frac{5}{20}\right).$$

The geometric probability mass function

Given a probability p, with $0 < p < 1$, consider a possibly infinite sequence of independent Bernoulli trials with success probability p. The geometric random variable X denotes the number of the trial on which the first success occurs. For $x = 1, 2, 3\ldots$, the **geometric probability mass function** is of the form

$$p_X(x) = (1-p)^{x-1}p.$$

Example 7.8 Tossing a fair die until we get an ace. If a fair die is tossed repeatedly until an ace (a one) appears, then the probability of observing this initial ace on the toss x is $(\frac{5}{6})^{x-1}(\frac{1}{6})$. For example:
The probability that the first ace appears on the first toss is $(\frac{5}{6})^0(\frac{1}{6}) \approx .1667$;
The probability that the first ace appears on the second toss is $(\frac{5}{6})^1(\frac{1}{6}) \approx .1389$;
The probability that the first ace appears on the third toss is $(\frac{5}{6})^2(\frac{1}{6}) \approx .1157$;
while, the probability that the first ace appears on the tenth toss is $(\frac{5}{6})^9(\frac{1}{6}) \approx .0323$;
and, the probability that the first ace appears on the twentieth toss is $(\frac{5}{6})^{19}(\frac{1}{6}) \approx .0052$.

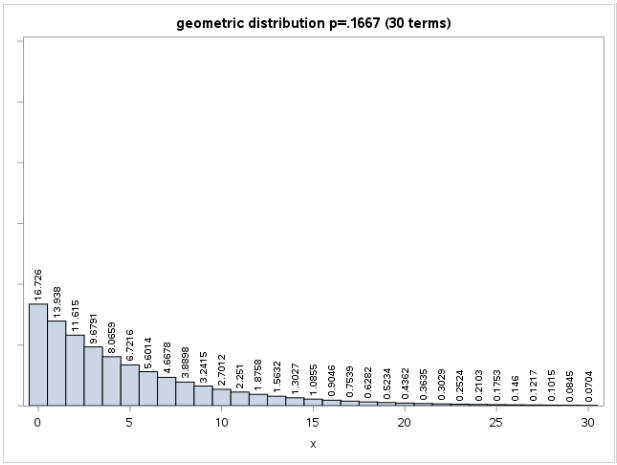
Example 7.9 Tossing a fair coin until we get a head. If a fair coin is tossed repeatedly until a head appears, then the probability of observing this initial head on the toss x is $(\frac{1}{2})^{x-1}(\frac{1}{2}) = (\frac{1}{2})^x$. For example:
The probability that the first ace appears on the first toss is $(\frac{1}{2})^1 = .5$;
The probability that the first ace appears on the second toss is $(\frac{1}{2})^2 = .25$;
The probability that the first ace appears on the third toss is $(\frac{1}{2})^3 = .125$;
while, the probability that the first ace appears on the tenth toss is $(\frac{1}{2})^{10} \approx .0010$;
and, the probability that the first ace appears on the twentieth toss is $(\frac{1}{2})^{20} \approx .000001$.

Graphical representations (with percentages) of the geometric distributions with $p = 1/6$ (the die tossing Example 7.8), $p = 1/4$ (the 5 red and 15 blue balls example), and $p = 1/2$ (the coin tossing Example 7.9) are provided in Figure 7.7. Note that in these plots the x–axis is "$X - 1$" the number of failures before first success.

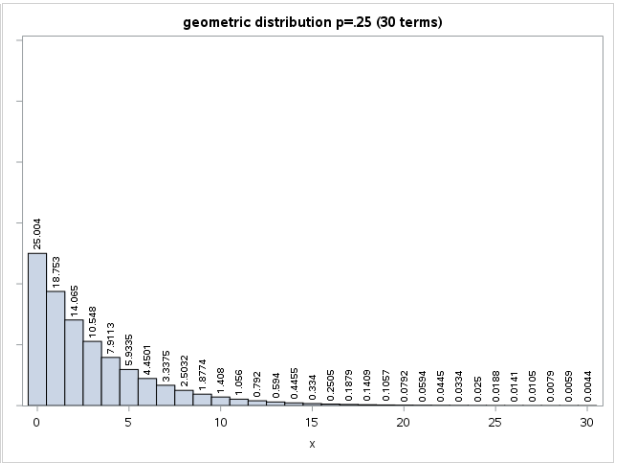
Terminology note: Some authors define the geometric random variable as $Y = X - 1$, the number of failures before the first success.

Figure 7.7 Geometric distributions for $p = 1/6, p = 1/4, p = 1/2$ (first 30 terms)
In these plots the x–axis is "$X - 1$" the number of failures before first success.

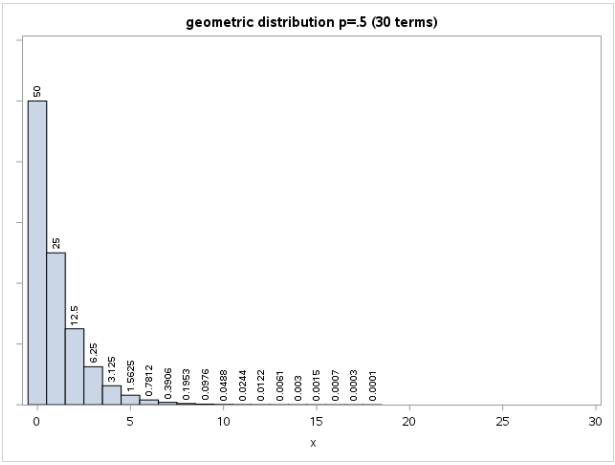
$p = 1/6$



$p = 1/4$



$p = 1/2$



7.6 The Poisson distribution

toc

In many settings, the Poisson distribution provides a realistic model for a random variable representing the number of occurrences of a "rare event". Consider a sequence of events occurring randomly in time or space and a count such as the number of radioactive particle emissions per unit time, the number of meteorites that collide with a satellite during a single orbit, the number of defects per unit length of some material, or the number of weed seeds per unit volume in a large batch of wheat seeds. We can picture the time (or location) of each occurrence as a point on the positive part of the number line. Consider the following assumptions about the times (locations) of these occurrences:

1. The probability of exactly one occurrence in a small interval of length t is approximately νt, where $\nu > 0$ is the mean rate at which events occur per unit time (the mean rate of occurrence).

2. The probability of more than one occurrence in a small interval of length t is negligible compared to the probability of exactly one occurrence in a small interval of length t.

3. The numbers of occurrences in non–overlapping intervals are independent in the sense that information concerning the number of events in one interval reveals nothing about the number of events in any other interval.

If we let X denote the number of occurrences in a period of length t, then these three assumptions imply that X follows the Poisson distribution with parameter $\lambda = \nu t$. The possible values of X are $0, 1, \ldots$, with no theoretical upper bound on the value.

The Poisson probability mass function

Given a constant $\lambda > 0$. For $x = 1, 2, 3 \ldots$, the **Poisson probability mass function** is of the form

$$P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda},$$

where $e \approx 2.718$ is the base of the natural logarithm.

Two applications of the Poisson distribution are provided in the following examples.

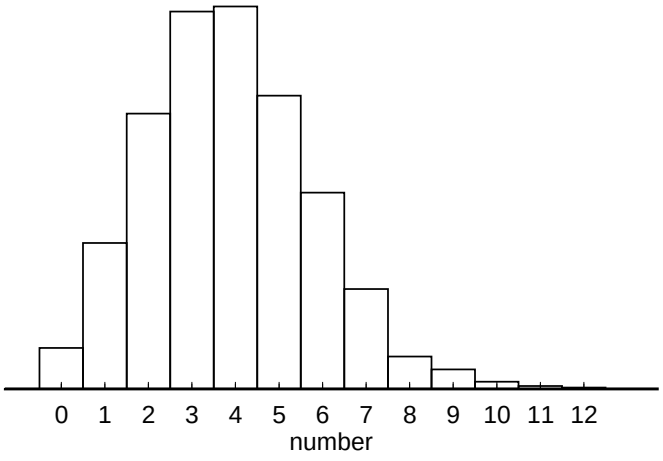
Example 7.10 Example. Radioactive disintegrations. This example is taken from Feller (1957), p. 149 and Cramér (1946) p. 436. In a famous experiment by Rutherford, Chadwick, and Ellis (*Radiations from Radioactive Substances*, Cambridge, 1920) a radioactive

substance was observed during 2608 consecutive time intervals of length $t = 7.5$ seconds each. The number of particles reaching a counter was recorded for each period. The results are summarized in Table 7.3 and Figure 7.8. (In this table the observations greater than or equal to 10 are grouped together. The data actually contain 10 tens, 4 elevens, and 2 twelves.) The last column of Table 7.3 contains expected relative frequencies (probabilities) computed using a Poisson model with λ estimated from these data. These Poisson probabilities appear to match the observed relative frequencies fairly well. A formal test of the goodness of fit of this Poisson model to these data indicates that the model does fit well ($\chi^2 = 12.885$, 9 d.f., P–value .17).

Table 7.3 Relative frequency distribution for radioactive disintegrations.

number	observed frequency	observed relative frequency	expected relative frequency
0	57	.0219	.0209
1	203	.0778	.0807
2	383	.1469	.1562
3	525	.2013	.2015
4	532	.2040	.1949
5	408	.1564	.1509
6	273	.1047	.0973
7	139	.0533	.0538
8	45	.0173	.0260
9	27	.0104	.0112
≥ 10	16	.0051	.0065
total	2608	.9991	.9999

Figure 7.8 Histogram for radioactive disintegrations (with ≥ 10 expanded).

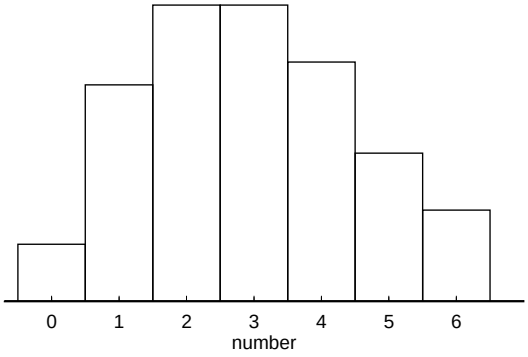


Example 7.11 Bacteria counts. This example is taken from Feller (1957), p.153. The original source is T. Matuszewsky, J. Supinska, and J. Neyman (1936), *Zentralblatt für Bakteriologie, Parasitenkunde und Infektionskrankrankheiten*, II Abt., **95**. A Petri dish with bacteria colonies was examined under a microscope. The dish was divided into small squares and the number of bacteria colonies, visible as dark spots, was recorded for each square. In this example t is the area the square within which the count is determined and we will take this area to be one. If the bacteria colonies were randomly distributed over the Petri dish, without being clustered together, then the Poisson model should hold. The results for one of several experiments are summarized in Table 7.4 and Figure 7.9. The last column of Table 7.4 contains expected relative frequencies (probabilities) computed using a Poisson model, with λ estimated from these data. In this example the observed relative frequency in the "≥ 6" line is for "exactly 6", but, the expected relative frequency is for all values greater than or equal to 6. These Poisson probabilities appear to match the observed relative frequencies fairly well. Therefore, the evidence supports the contention that the bacteria colonies are randomly distributed over the Petri dish. A formal test of the goodness of fit of this Poisson model to these data indicates that the model does fit well ($\chi^2 = .8386$, 5 d.f., P–value .9745).

Table 7.4 Relative frequency distribution for bacteria counts.

number	observed frequency	observed relative frequency	expected relative frequency
0	5	.0424	.0533
1	19	.1610	.1562
2	26	.2203	.2290
3	26	.2203	.2239
4	21	.1780	.1641
5	13	.1102	.0962
≥ 6	8	.0678	.0772
total	118	1.0000	.9999

Figure 7.9 Histogram for bacteria counts.



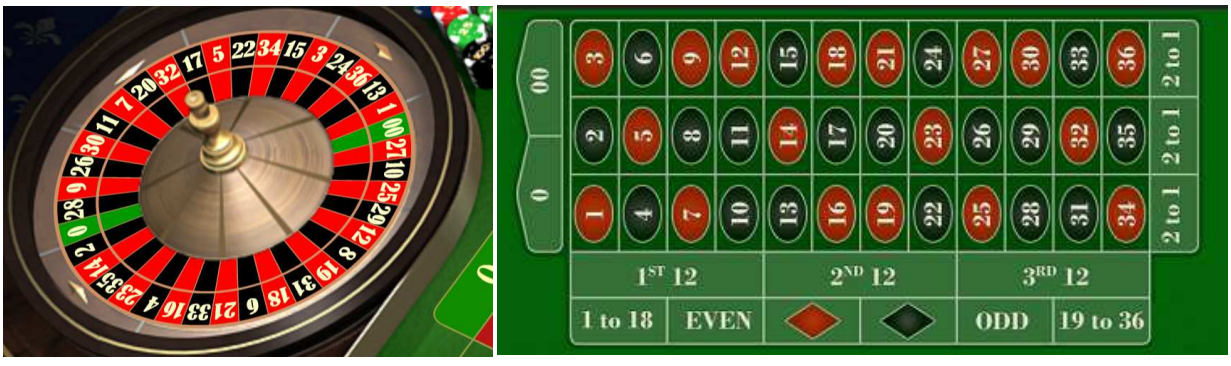
7.7 Expected value

toc

There are many situations where we might wonder what value to expect when we perform an experiment and observe the value of a particular random variable. Before we can address this topic more formally we need to explain exactly what we mean when we say "what value to expect". The expected value of a random variable can be viewed as the long run average value of the random variable. That is, the average of the values of the random variable we would obtain if we conducted the experiment and observed the value of the random variable a large number of times. Before we introduce a formal definition it is helpful to consider a simple example of a real world application of an expected value.

Example 7.11 Expected winnings in roulette An American style roulette wheel, as shown in Figure 7.10, has 38 pockets of which 18 are red, 18 are black, and 2 are green. We will consider the expected winnings for a one dollar bet on red (a bet that the ball will land in one of the 18 red pockets).

Figure 7.10 An American roulette wheel and betting table



Let the random variable W denote our winnings. This random variable assumes two values: $W = 1$ indicates that the ball lands in a red pocket and we win one dollar; while $W = -1$ indicates that the ball lands in a black or green pocket and we lose one dollar. Since 18 of the 38 pockets are red, the probability mass function for W is given by

$$p_W(w) = \begin{cases} 18/38 & \text{if } w = 1 \\ 20/38 & \text{if } w = -1 \\ 0 & \text{otherwise} \end{cases}.$$

If we imagine placing this bet a large number of times, then we see that in about 18 out of 38 tries (about 47.37% of the time) we will win a dollar (we will observe $W = 1$) and in about 20 out of 38 tries (about 52.63% of the time) we will lose a dollar (we will observe $W = -1$). Thus, on average in the long run, we expect to see an average winnings of $(18 - 20)/38 = -2/38$ dollars (about -.0526 dollars or -5.26 cents). That is, if we made a large number of bets, on average, we would lose 5.26 cents per bet. The long run average winnings we computed, $(18 - 20)/38 = -2/38$ dollars, is called the expected values of W. The notation for the expected value of W is $E(W)$.

We will now provide a formal definition of the expected value of a discrete random variable.

If X is a discrete r.v. with finite sample space $\Omega_X = \{x_1, \ldots, x_N\}$ and p.m.f. p_X, then the **expected value** of X, denoted by $E(X)$ or μ_X, is defined as the weighted average of the possible values of X (the elements of Ω_X) obtained using the associated probabilities (the corresponding values of $p_X(x)$) as weights. In symbols we have

$$E(X) = x_1 p_X(x_1) + x_2 p_X(x_2) + \cdots + x_N p_X(x_N) = \sum_{i=1}^{N} x_i p_X(x_i).$$

Example 7.12 Two distributions to illustrate computations Let X and Y denote r.v.'s with the following p.m.f.'s

$$p_X(x) = \begin{cases} .1 & \text{if } x = 1 \\ .2 & \text{if } x = 2 \\ .4 & \text{if } x = 3 \\ .2 & \text{if } x = 4 \\ .1 & \text{if } x = 5 \\ 0 & \text{otherwise} \end{cases} \quad p_Y(y) = \begin{cases} .1 & \text{if } y = 1 \\ .1 & \text{if } y = 2 \\ .2 & \text{if } y = 3 \\ .2 & \text{if } y = 4 \\ .4 & \text{if } y = 5 \\ 0 & \text{otherwise} \end{cases}.$$

The expected values of X and Y are:

$$E(X) = .1(1) + .2(2) + .4(3) + .2(4) + .1(5) = .1 + .4 + 1.2 + .8 + .5 = 3$$

$$E(Y) = .1(1) + .1(2) + .2(3) + .2(4) + .4(5) = .1 + .2 + .6 + .8 + 2.0 = 3.7.$$

A tabular representation of these computations may be helpful. In these tabular representations the expected value is the sum of the $xp_X(x)$ column.

x	$p_X(x)$	$xp_X(x)$
1	.1	$1 \times .1 = .1$
2	.2	$2 \times .2 = .4$
3	.4	$3 \times .4 = 1.2$
4	.2	$4 \times .2 = .8$
5	.1	$5 \times .1 = .5$
		$E(X) = 3$

y	$p_Y(y)$	$yp_Y(y)$
1	.1	$1 \times .1 = .1$
2	.1	$2 \times .1 = .2$
3	.2	$3 \times .2 = .6$
4	.2	$4 \times .2 = .8$
5	.4	$5 \times .4 = 2.0$
		$E(Y) = 3.7$

If the sample space is countably infinite (as with the geometric or Poisson distribution), then $\Omega_X = \{x_1, x_2, \ldots\}$ is countably infinite and the sum in the definition of $E(X)$ has an infinite number of terms (does not stop at a finite N).

An aside – existence of expected values

Some technicalities may arise when Ω_X is countably infinite. The problem is that for some choices of the p.m.f. p_X the series (infinite sum) in the definition of the expected value may not exist. (More formally, this series may not converge.) We do not need to worry about this technicality, since the expected value does exist for all of the discrete random variables with countably infinite sample spaces we will encounter.

Expected value as center of mass

If X is a discrete r.v. with finite sample space $\Omega_X = \{x_1, \ldots, x_N\}$ and p.m.f. p_X, then, as the name suggests, we can think of the probabilities $p_X(x_1), \ldots, p_X(x_N)$ as masses located at points $x_1, \ldots, x_N$ on a segment of the number line. With this interpretation the center of mass of the distribution of X is located at $E(X) = x_1 p_X(x_1) + \cdots + x_N p_X(x_N)$. Similarly, if $\Omega_X = \{x_1, x_2, \ldots\}$, then the center of mass of the distribution of X is located at $E(X) = x_1 p_X(x_1) + x_2 p_X(x_2) + \cdots$.

Expected value as population mean

If an individual is selected at random from a population of N people and X represents the individual's height in inches so that Ω_X represents the collection of distinct heights for this population and $p_X(x_i)$ is the proportion of individuals in the population with height x_i, then $E(X)$ is the population mean height (the average height of all N people in the population).

The expected value of a function of X

We will now show how the definition of the expected value of X can be extended to the expected value of a function of X. The expected value of a function of X may be of interest in some applications. Also, as we will see shortly, the variance of X is defined as the expected value of a specific function of X. Let g denote the function of interest and let Ω_Y denote the range of the function, that is, Ω_Y is the collection of values obtained when the function g is applied to the elements of Ω_X. If Ω_X and Ω_Y are discrete, then $Y = g(X)$ is a discrete r.v. and there are two ways to compute the expected value of Y. Letting $\Omega_X = \{x_1, x_2, \ldots x_N\}$ and $\Omega_Y = \{y_1, y_2, \ldots y_M\}$:
(1) We can find the p.m.f. p_Y of Y and compute the expected value of Y as

$$E(Y) = y_1 p_Y(y_1) + y_2 p_Y(y_2) + \cdots + y_M p_Y(y_M) = \sum_{i=1}^{M} y_i p_Y(y_i).$$

(2) We can use the p.m.f. p_X of X to compute the expected value of Y as

$$E(Y) = E(g(X)) = g(x_1)p_X(x_1) + g(x_2)p_X(x_2) + \cdots + g(x_N)p_X(x_N) = \sum_{i=1}^{N} g(x_i)p_X(x_i).$$

The analogous expressions with infinite sums apply in the countably infinite case. We will demonstrate these computations in the context of finding the expected winnings of a bet on red in the roulette example.

Example 7.11 Expected winnings in roulette (revisited) The 36 pockets on the roulette wheel shown in Figure 7.10, are numbered as follows: pockets 0 and 00 are green;
pockets 1,3,5,7,9,12,14,16,18,19,21,23,25,27,30,32,34, and 36 are red; and,
pockets 2,4,6,8,10,11,13,15,17,20,22,24,26,28,29,31,33, and 35 are black. If we let X denote the number on the pocket the ball lands in, then the r.v. X follows the uniform distribution with $\Omega_X = \{0, 00, 1, 2, \ldots, 36\}$. In terms of this X the function g which defines the

winnings for a bet on red assigns the value 1 to the numbers associated with the red pockets and assigns the value -1 to the numbers associated with the green and black pockets so that the r.v. $W = g(X)$ is the winnings for a bet on red. As noted earlier, $E(W) = (1)\frac{18}{38} + (-1)\frac{20}{38} = -\frac{2}{38}$. Using the relationship $W = g(X)$ and the uniform distribution of X, with some regrouping, we have

$$\begin{aligned} E(W) &= (1)\left[p_X(1) + p_X(3) + \cdots + p_X(34) + p_X(36)\right] \\ &\quad + (-1)\left[p_X(2) + p_X(4) + \cdots + p_X(33) + p_X(35) + p_X(0) + p_X(00)\right] \\ &= (1)\frac{18}{38} + (-1)\frac{20}{38} = -\frac{2}{38}, \end{aligned}$$

since $p_X(x) = \frac{1}{38}$ for the 38 values in Ω_X and there are 18 values with $g(x) = 1$ and 20 values with $g(x) = -1$.

Example 7.12 Two distributions to illustrate computations, revisited Recall that X and Y denote r.v.'s with the following p.m.f.'s

$$p_X(x) = \begin{cases} .1 & \text{if } x = 1 \\ .2 & \text{if } x = 2 \\ .4 & \text{if } x = 3 \\ .2 & \text{if } x = 4 \\ .1 & \text{if } x = 5 \\ 0 & \text{otherwise} \end{cases} \quad p_Y(y) = \begin{cases} .1 & \text{if } y = 1 \\ .1 & \text{if } y = 2 \\ .2 & \text{if } y = 3 \\ .2 & \text{if } y = 4 \\ .4 & \text{if } y = 5 \\ 0 & \text{otherwise} \end{cases}.$$

We will now find the expected values of X^2 and Y^2. The reason that these expected values are of interest will become clear shortly.

$$\begin{aligned} E(X^2) &= .1(1^2) + .2(2^2) + .4(3^2) + .2(4^2) + .1(5^2) \\ &= .1(1) + .2(4) + .4(9) + .2(16) + .1(25) \\ &= .1 + .8 + 3.6 + 3.2 + 2.5 = 10.2 \end{aligned}$$

$$\begin{aligned} E(Y^2) &= .1(1^2) + .1(2^2) + .2(3^2) + .2(4^2) + .4(5^2) \\ &= .1(1) + .1(4) + .2(9) + .2(16) + .4(25) \\ &= .1 + .4 + 1.8 + 3.2 + 10.0 = 15.5. \end{aligned}$$

As before, a tabular representation of these computations may be helpful. In these tabular representations the expected value of X^2 is the sum of the $x^2 p_X(x)$ column.

x	$p_X(x)$	$x^2 p_X(x)$
1	.1	$1 \times .1 = .1$
2	.2	$4 \times .2 = .8$
3	.4	$9 \times .4 = 3.6$
4	.2	$16 \times .2 = 3.2$
5	.1	$25 \times .1 = 2.5$
		$E(X^2) = 10.2$

y	$p_Y(y)$	$y^2 p_Y(y)$
1	.1	$1 \times .1 = .1$
2	.1	$4 \times .1 = .4$
3	.2	$9 \times .2 = 1.8$
4	.2	$16 \times .2 = 3.2$
5	.4	$25 \times .4 = 10.0$
		$E(Y^2) = 15.5$

7.8 Variance

toc

The expected value or mean of the distribution of X, $E(X) = \mu_X$, provides an indication of where the "center" of the distribution is located on the number line. It would also be useful to have a measure of the amount of variability in the distribution of X. The variance is one such measure of variability. The variance of the distribution of X is the long run average value of the square of the distance between the points in the sample space Ω_X and the mean μ_X of the distribution. More formally, letting $E(X) = \mu_X$ denote the mean of the distribution of X, and assuming that this mean exists, the function $g(x) = (x - \mu_X)^2$ (the squared deviation of x from the mean of the distribution of X) can be used to define a measure of the variability in the distribution of X. The expected value of the r.v. $g(X) = (X - \mu_X)^2$ is the variance of the distribution of X.

If X is a discrete r.v. with sample space $\Omega_X = \{x_1, \ldots, x_N\}$, p.m.f. p_X, and mean $\mu_X = E(X)$, then the **variance** of X, denoted by $\text{var}(X)$ or σ_X^2, is defined by

$$\text{var}(X) = (x_1 - \mu_X)^2 p_X(x_1) + \cdots + (x_N - \mu_X)^2 p_X(x_N) = \sum_{i=1}^{N} (x_i - \mu_X)^2 p_X(x_i).$$

Similarly, if $\Omega_X = \{x_1, x_2, \ldots\}$, then the variance is

$$\text{var}(X) = (x_1 - \mu_X)^2 p_X(x_1) + (x_2 - \mu_X)^2 p_X(x_2) + \cdots = \sum_{i=1}^{\infty} (x_i - \mu_X)^2 p_X(x_i).$$

Since squared distances are used in its definition the scale of measurement for the variance of X is the square of the scale of measurement for X itself. For example, if X denotes

the height of a person in inches, then the variance of X, which provides a measure of variability among the heights of different individuals, is measured in inches squared. This change of scale is undesirable in practical applications where we really need a measure of variability on the original scale of measurement. To obtain a measure of variability on the original scale of measurement we simply take the square root of the variance to obtain the standard deviation. The principal (positive) square root of the variance of X, $\text{SD}(X) = \sigma_X = \sqrt{\text{var}(X)}$, is know as the **standard deviation** of X.

Example 7.12 Two distributions to illustrate computations, revisited Recall that X and Y denote r.v.'s with the following p.m.f.'s

$$p_X(x) = \begin{cases} .1 & \text{if } x = 1 \\ .2 & \text{if } x = 2 \\ .4 & \text{if } x = 3 \\ .2 & \text{if } x = 4 \\ .1 & \text{if } x = 5 \\ 0 & \text{otherwise} \end{cases} \quad p_Y(y) = \begin{cases} .1 & \text{if } y = 1 \\ .1 & \text{if } y = 2 \\ .2 & \text{if } y = 3 \\ .2 & \text{if } y = 4 \\ .4 & \text{if } y = 5 \\ 0 & \text{otherwise} \end{cases}$$

and that $E(X) = 3$ and $E(Y) = 3.7$. We will now find the variances and standard deviations of X and Y.

$$\begin{aligned} \text{var}(X) &= .1(1-3)^2 + .2(2-3)^2 + .4(3-3)^2 + .2(4-3)^2 + .1(5-3)^2 \\ &= .1(-2)^2 + .2(-1)^2 + .4(0)^2 + .2(1)^2 + .1(2)^2 \\ &= .4 + .2 + 0 + .2 + .4 = 1.2 \\ \text{var}(Y) &= .1(1-3.7)^2) + .1(2-3.7)^2 + .2(3-3.7)^2 + .2(4-3.7)^2 + .4(5-3.7)^2 \\ &= .1(-2.7)^2 + .1(-1.7)^2 + .2(-0.7)^2 + .2(0.3)^2 + .4(1.3)^2 \\ &= .729 + .289 + .098 + .018.676 = 1.81. \end{aligned}$$

The standard deviations are $\text{SD}(X) = \sigma_X = \sqrt{1.2} \approx 1.0954$ and $\text{SD}(Y) = \sigma_Y = \sqrt{1.81} \approx 1.3454$.

Again, a tabular representation of these computations may be helpful. In these tabular representations the variance of X is the sum of the $(x - E(X))^2 p_X(x)$ column.

x	$p_X(x)$	$(x - E(X))^2 p_X(x)$
1	.1	$(-2)^2 \times .1 = .4$
2	.2	$(-1)^2 \times .2 = .2$
3	.4	$(0)^2 \times .4 = 0$
4	.2	$(1)^2 \times .2 = .2$
5	.1	$(2)^2 \times .1 = .4$
		$\text{var}(X) = 1.2$

y	$p_Y(y)$	$(y - E(Y))^2 p_Y(y)$
1	.1	$(-2.7)^2 \times .1 = .729$
2	.1	$(-1.7)^2 \times .1 = .289$
3	.2	$(-.7)^2 \times .2 = .098$
4	.2	$(.3)^2 \times .2 = .018$
5	.4	$(1.3)^2 \times .4 = .676$
		$\text{var}(Y) = 1.81$

A computational formula for the variance

The variance of X can be computed by subtracting the square of the expected value of X from the expected value of the square of X. In symbols,

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

Example 7.12 Two distributions to illustrate computations, revisited For the random variables X and Y of this example. We have $E(X) = 3$, $E(Y) = 3.7$, $E(X^2) = 10.2$, and $E(Y^2) = 10.5$. Thus $\text{var}(X) = 10.2 - (3)^2 = 1.2$ and $\text{var}(Y) = 15.5 - (3.7)^2 = 15.5 - 13.69 = 1.81$.

Some useful properties of expected values and variances

For these properties a and b denote constants, X and Y are discrete r.v.'s, and $E(X)$, $E(Y)$, and $\text{var}(X)$ are assumed to exist.

(1) The expected value of a constant is that constant. In symbols,

$$E(a) = a.$$

(2) If X is bounded below by a (X is always greater than or equal to a or more formally X is greater than or equal to a with probability one), then the expected value of X is greater than or equal to a. In symbols,

$$\text{if } P(X \geq a) = 1, \text{ then } E(X) \geq a.$$

(3) Similarly, if X is bounded above by b (X is always less than or equal to b or more formally X is less than or equal to b with probability one), then the expected value of X is less than or equal to b. In symbols, if

$$\text{if } P(X \leq b) = 1, \text{ then } E(X) \leq b.$$

(4) Combining (2) and (3) we see that if X is always between a and b or more formally X takes on a value between a and b, inclusive, with probability one, then the expected value of X is between a and b, inclusive. In symbols,

$$\text{if } P(a \leq X \leq b) = 1, \text{ then } a \leq E(X) \leq b.$$

(5) If we transform the value of X by first multiplying X by a constant b and then adding a constant a, then the expected value of X is transformed in the same way. In symbols,

$$E(a + bX) = a + bE(X).$$

(6) The expected value of the sum of two (or more) random variables is equal to the sum of their expected values. In symbols,

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

(7) We will now note what happens to the variance of X when we transform the value of X by first multiplying X by a constant b and then adding a constant a. Since variance is a measure of variability adding the constant a has no effect on the variance. Multiplying X by the constant b changes the scaling of the values of X, if $|b| > 1$ the values get more spread out on the number line, if $|b| < 1$ the values get closer together (less spread out) on the number line, and if $b < 0$ then the ordering of the values is reversed. Therefore, the transformation of X to $a+bX$ has a multiplicative effect on the variance of X corresponding to a multiplication by the square of the constant b. In symbols,

$$\text{var}(a + bX) = b^2\text{var}(X).$$

(8) The variance of a constant is zero,

$$\text{var}(a) = 0.$$

7.9 Means, variances, and pmf's for several families of distributions toc

A family of distributions is a collection of distributions of a specified form which is known up to the values of one or more parameters. For the families summarized below, the p.m.f. is expressed as a function of one or more parameters and the family is obtained by varying the parameter or parameters over all suitable values.

Binomial distribution

The binomial distribution with parameters n (number of trials) and p (success probability).

Restrictions on the parameters: n is a positive integer $(1, 2, 3, \ldots)$ and $0 < p < 1$.
Probability mass function:

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

Mean and variance:

$$E(X) = np \text{ and } \operatorname{var}(X) = np(1-p)$$

Hypergeometric distribution

The hypergeometric distribution with parameters A (number of units of the first type in the population), B (number of units of the second type in the population), and n (sample size).

Restrictions on the parameters: A, B, and n are positive integers $(1, 2, 3, \ldots)$ with $n \leq A + B$.
Probability mass function:

$$p_X(x) = \frac{\binom{A}{x}\binom{B}{n-x}}{\binom{A+B}{n}} \text{ for } x = 0, 1, \ldots, n \text{ subject to the restrictions below.}$$

Notice that the values of A and B may impose restrictions on the values of x for which this expression works. Specifically, we must have $x \leq A$ and $n - x \leq B$.
Mean and variance:

$$E(X) = n\left(\frac{A}{A+B}\right) \text{ and } \operatorname{var}(X) = n\left(\frac{A}{A+B}\right)\left(\frac{B}{A+B}\right)\left(\frac{A+B-n}{A+B-1}\right)$$

Geometric distribution

The geometric distribution with parameter p (success probability).

Restrictions on the parameter: $0 < p < 1$.
Probability mass function:

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

Mean and variance:

$$E(X) = \frac{1}{p} \text{ and } \text{var}(X) = \frac{1-p}{p^2}$$

Uniform distribution

The discrete uniform distribution on the integers $1, 2, \ldots, N$.

Restrictions on the parameter: The parameter N is a positive integer $(1, 2, 3, \ldots)$.
Probability mass function:

$$p_X(x) = \frac{1}{N} \text{ for } x = 1, 2, \ldots, N$$

Mean and variance:

$$E(X) = \frac{N+1}{2} \text{ and } \text{var}(X) = \frac{N^2-1}{12}$$

Poisson distribution

The Poisson distribution with parameter λ.

Restrictions on the parameter: $\lambda > 0$.
Probability mass function:

$$p_X(x) = \frac{\lambda^x}{x!}e^{-\lambda} \text{ for } x = 0, 1, 2, 3, \ldots$$

Mean and variance:

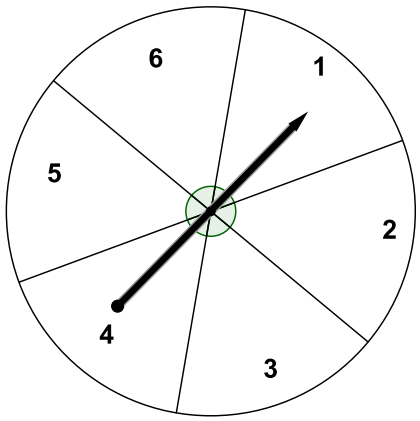
$$E(X) = \lambda \text{ and } \text{var}(X) = \lambda$$

8 Continuous random variables toc

8.1 Moving from a discrete to a continuous random variable toc

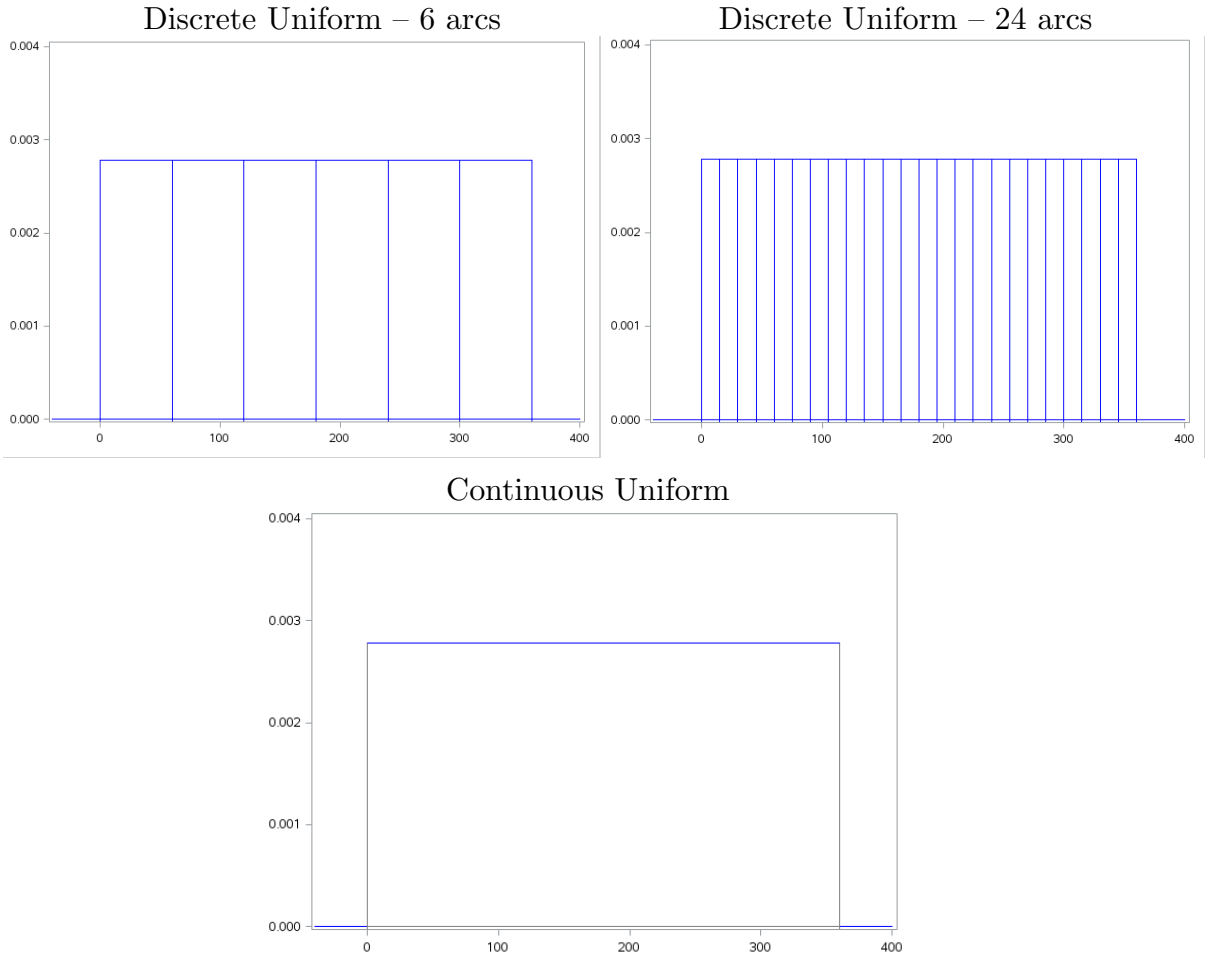
A continuous random variable is a random variable X for which the sample space Ω_X is a (bounded or unbounded) interval of values on the number line. The term continuous here refers to the fact that when Ω_X is an interval the possible values of the random variable form a continuum. That is, there is a continuous transition from one possible value to the next in contrast to the jumps between values with a discrete transition. We will begin with a simple example showing how we can move from a discrete uniform distribution on six integers to a continuous uniform distribution on an interval.

Figure 8.1 A spinner on a disk with six equiangular regions



Consider a spinner atop a disk and an experiment consisting of spinning the pointer and noting its location on the circumference of the disk. In particular consider the spinner in Figure 8.1 where the circumference of the disk is divided into six equal length arcs. If we assume that the spinner is very well made and balanced, then it is reasonable to claim that the six values on the disk are equally probable. (If the pointer lands exactly on one of the boundaries, then we can simply spin again until the pointer lands inside a region.) Hence, under these assumptions, the discrete random variable X, denoting the number on the region where the pointer lands, follows the uniform distribution on the integers in the sample space $\Omega_X = \{1, 2, 3, 4, 5, 6\}$.

Figure 8.2 Uniform distributions on the interval from 0 to 360.
The first two graphs show probability histograms for discrete uniform distributions on the midpoints of 6 and 24 equal length subintervals (arcs).
The third graph shows the continuous uniform density curve (p.d.f.).



On the other hand, if we think of the circumference of the disk as a continuum, then we can represent an elementary outcome as a real number representing the angle of the radius to the point where the pointer stops. Let the points on the circumference of the disk be labeled in degrees starting with zero on the radius between 6 and 1 and moving clockwise and let Y denote the angle corresponding to the point where the pointer lands. With this convention the sample space is the interval from 0 to 360 degrees, *i.e.*, $\Omega_Y = [0, 360)$, with $0 < Y < 60$ corresponding to $X = 30$, $60 < Y < 120$ corresponding to $X = 90$, $120 < Y < 180$ corresponding to $X = 150$, $180 < Y < 240$ corresponding to $X = 210$, $240 < Y < 300$ corresponding to $X = 270$, and $300 < Y < 360$ corresponding to $X = 330$.

If we think of these six arcs as six discrete outcomes corresponding to these six values of X, then we can label and scale the probability histogram of X so that it lies on the interval from 0 to 360 on the number line. As noted above, with this scaling the midpoints of the arcs 30, 90, 150, 210, 270, and 330 are the values of X. This probability histogram is shown in the first graph of Figure 8.2. Each of the rectangles in the probability histogram has height 1/360.

It seems natural to argue that the continuous random variable Y should be uniformly distributed on the sample space $\Omega_Y = [0, 360)$. More formally, it seems natural to require that the probability that Y takes on a value within any arc of length θ ($0 < \theta < 360$) is equal to the length of the arc divided by 360, *i.e.*, for $0 < \theta < 360$ and any arc of length θ, the probability that Y belongs to this arc (that the pointer stops within this arc) is $\frac{\theta}{360}$. As shown in Figure 8.2, the probability histogram for this continuous uniform distribution on $\Omega_Y = [0, 360)$ is a single rectangle of height 1/360 located over the interval from 0 to 360.

The probability histogram for the discrete uniform distribution on 24 arcs in Figure 8.2 is provided to indicate how we can think of the continuous uniform probability histogram as the limiting version of the discrete uniform probability histogram (on this same interval) which we would obtain if we made the rectangles narrower and narrower. That is, to indicate that by increasing the number of arcs that the circle (interval) is partitioned into the histogram of the discrete uniform distribution would approach the histogram of the continuous uniform distribution.

Notice that with this continuous uniform distribution on the circumference of the disk the probability that the pointer lands at a specified point is zero but the probability that it lands in a specified arc, of length θ, which contains the point is $\frac{\theta}{360} > 0$. Notice also that that if the length of the arc containing the point is made shorter and shorter, then the arc degenerates to the point and the probability $\frac{\theta}{360}$ approaches zero; thus, the notion of a point having probability zero is consistent with this assignment of probability to an arc.

8.2 Continuous random variables

toc

We will now consider probability models for the distribution of a continuous random variable. Recall that a continuous random variable is a random variable X for which the sample space Ω_X is a (bounded or unbounded) interval of values on the number line. In

Section 8.1 we showed how a continuous uniform distribution on an interval arises as a sort of limiting version of a discrete uniform distribution on the same interval obtained by letting the number of possible values increase without bound. If we modified our division of the disk by allowing the lengths of the arcs to vary, our limiting argument would lead to a non–uniform continuous distribution. We will now consider continuous random variables more generally.

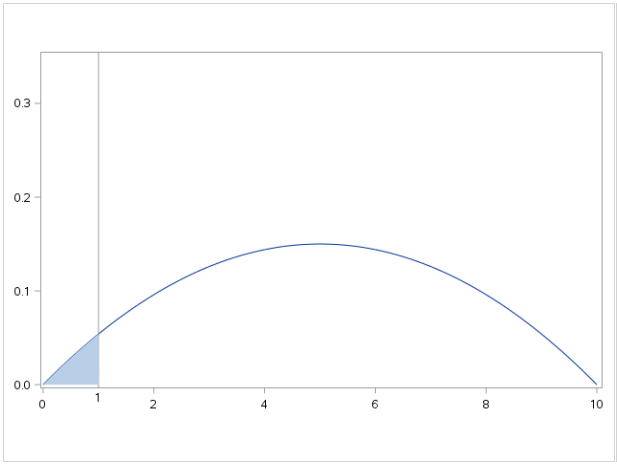
As noted above, the sample space of a continuous random variable is a (bounded or unbounded) interval of values on the number line. For simplicity, the sample space is often the entire number line or the nonnegative part of the number line. We can characterize the distribution of the continuous r.v. X by specifying a **probability density function** (denoted p.d.f.) f_X. We will emphasize the graph of the p.d.f. and we will also refer to the p.d.f. as a density curve. A probability density function f_X is a nonnegative valued function with the property that the area under the graph of the function (the density curve) is one. That is, the area between the x–axis and the density curve is one. We can think of the density curve as a smooth version of a probability histogram with the rectangles of the histogram replaced by a smooth curve indicating where the tops of the rectangles would be.

As noted above, the area under a density curve (p.d.f.) is one. The probability of an event is equal to the area under the density curve over the interval corresponding to the event. Thus, for any event A defined in terms of the continuous r.v. X with p.d.f f_X, the probability of event A is equal to the area under the graph of f_X over the region A. For example, if the event A is an interval, say $A = (a, b)$, then the probability that X belongs to A, $P(a < X < b)$ is equal to the area under the p.d.f. f_X (under the density curve) over the interval (a, b). We can use a computer or a calculator to find such probabilities (areas).

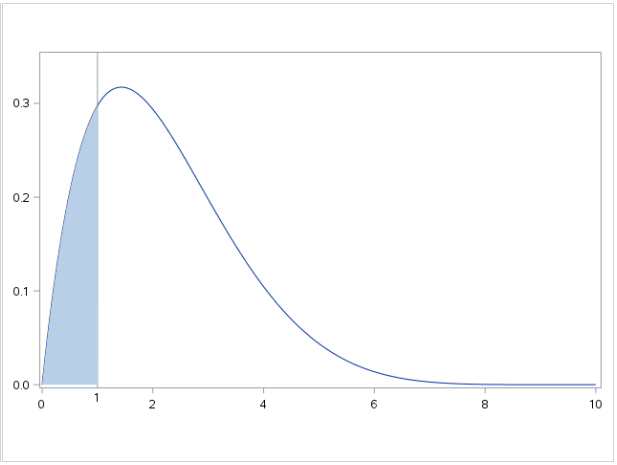
This identification of the probability of an event with an area under a density curve is illustrated in Figures 8.3 and 8.4. In these figures, X_1 is a continuous random variable with a symmetric distribution and X_2 is a continuous random variable with a skewed right distribution. In Figure 8.3 the probabilities are of the form $P(X \leq a)$ and in Figure 8.4 the probabilities are of the form $P(a \leq X \leq b)$.

Figure 8.3 Continuous distributions – probability as area
symmetric distribution: X_1, skewed right distribution: X_2

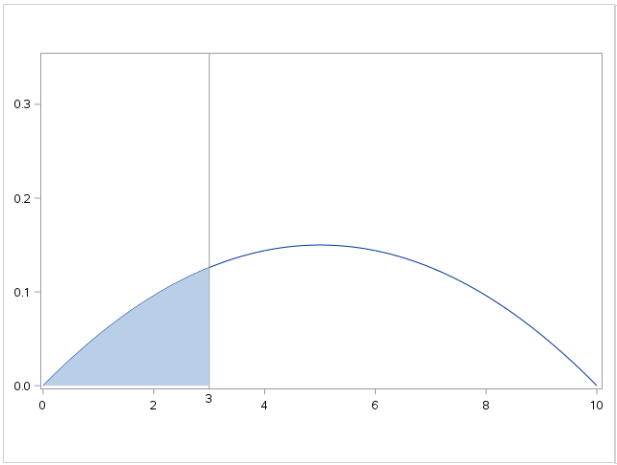
$P(X_1 \leq 1) = .028$



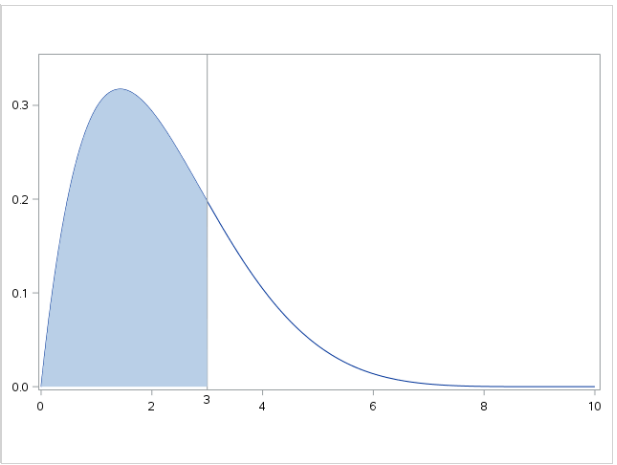
$P(X_2 \leq 1) = .1869$



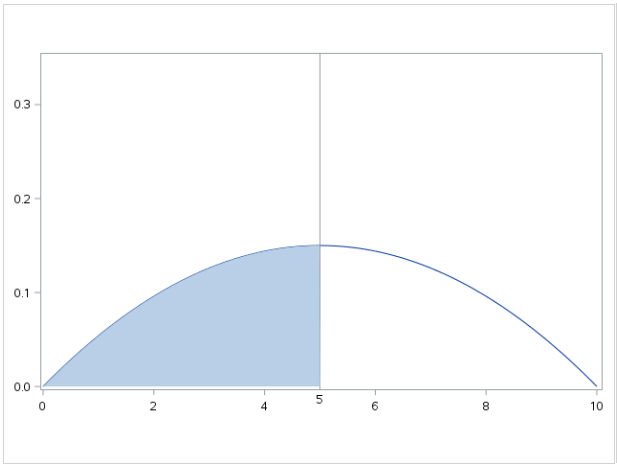
$P(X_1 \leq 3) = .216$



$P(X_2 \leq 3) = .7447$



$P(X_1 \leq 5) = .5$



$P(X_2 \leq 5) = .9648$

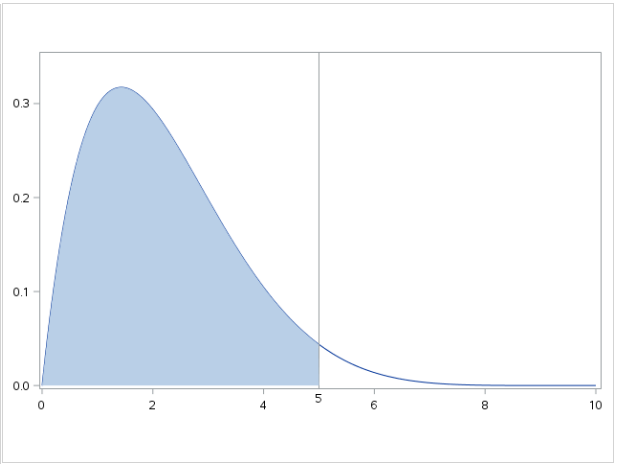
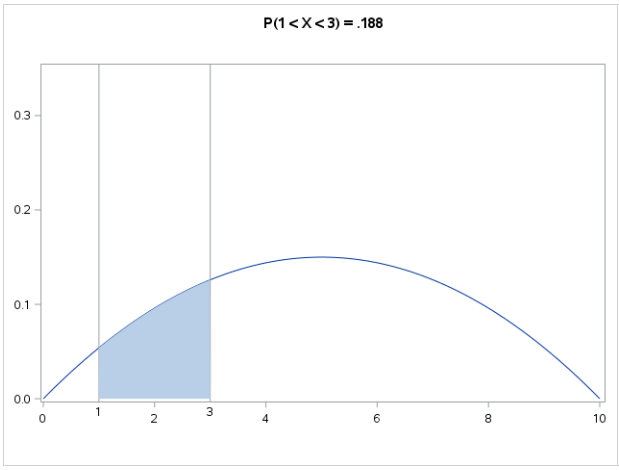
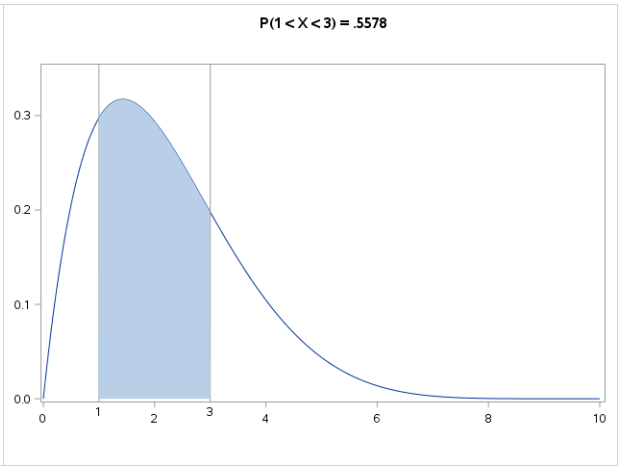


Figure 8.4 Continuous distributions – probability as area

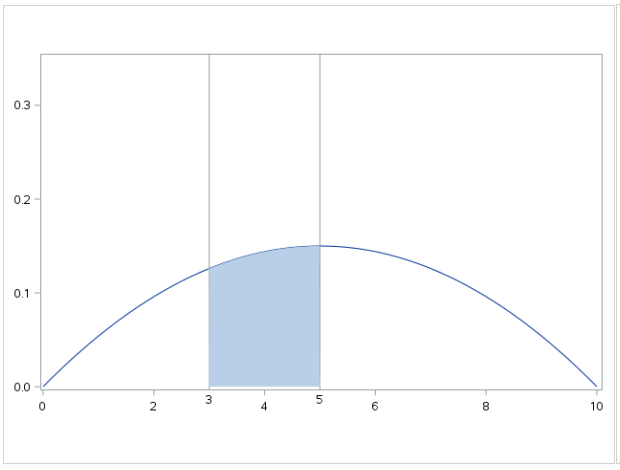
$P(1 \leq X_1 \leq 3) = .188$



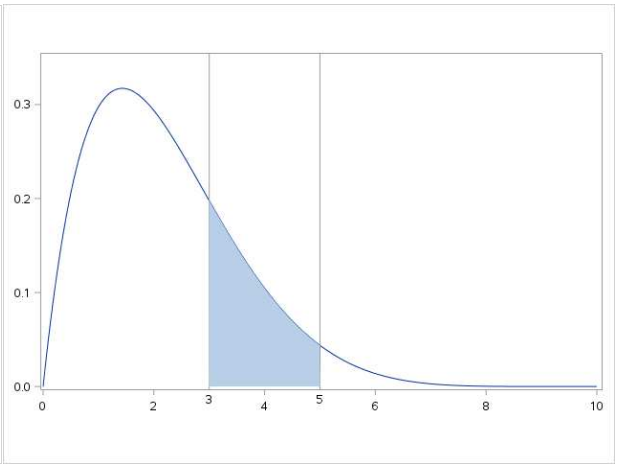
$P(1 \leq X_2 \leq 3) = .5578$



$P(3 \leq X_1 \leq 5) = .284$



$P(3 \leq X_2 \leq 5) = .2201$



Notice that $f_X(x)$ is not a probability! With a continuous random variable only events which can be represented as non–degenerate intervals (or unions of non–degenerate intervals) can have positive probability, since probability is defined as the area under the density curve and we need at least one non–degenerate interval to obtain a positive area.

Given a probability model for the distribution of a continuous random variable X, *i.e.*, given a density curve for the distribution of X, we can define population parameters which characterize relevant aspects of the distribution. For example, we can define the population mean μ as the balance point of the unit mass bounded by the density curve and the number line. We can also think of the population mean as the weighted average of the infinite collection of possible values of X with weights determined by the density

curve. We can similarly define the population median M as the point on the number line where a vertical line would divide the area under the density curve into two equal areas (each of size one–half).

The distribution of X can also be characterized in terms of its **cumulative distribution function** (denoted c.d.f.) F_X, the c.d.f. gives the probability of events of the form $[X \leq x]$, thus

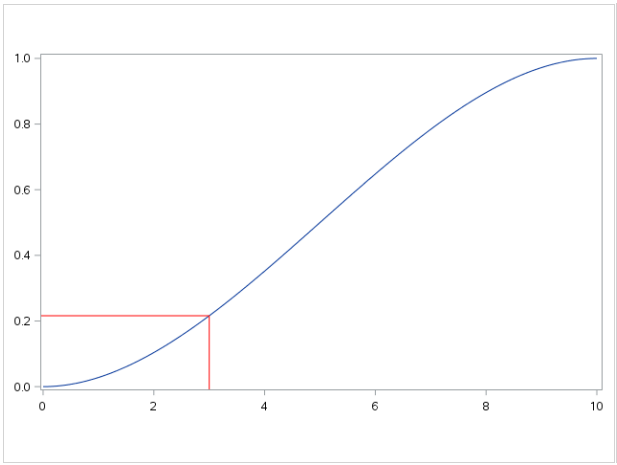
$$F_X(x) = P(X \leq x).$$

For a continuous random variable the value of the c.d.f. $F_X(x)$ is the area under the density curve over the interval which ranges from minus infinity to x. As noted earlier, we can use a computer or a calculator to determine such areas (integrals).

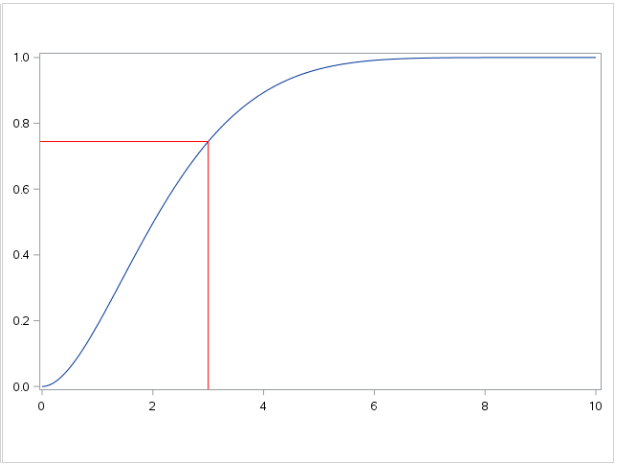
The cumulative distribution functions corresponding to the probability density functions of Figure 8.3 are provided in Figure 8.5. In this figure lines at $x = 3$ and $F_X(3)$ are included to show how the c.d.f yields the value of the corresponding area under the p.d.f over the interval $(0, 3)$ as shown in Figure 8.3.

Figure 8.5 Continuous distributions – cumulative distribution functions

cdf of the symmetric distribution of X_1 of Figure 8.3 lines show $x = 3$ and $F_{X_1}(3) = .216$



cdf of the skewed right distribution of X_2 of Figure 8.3 lines show $x = 3$ and $F_{X_2}(3) = .7447$



8.3 The normal distribution toc

The most widely used continuous probability model is the normal probability model or normal distribution. The normal distribution with mean μ and standard deviation σ can be characterized by its density curve. The density curve for the normal distribution with mean μ and standard deviation σ is the familiar bell shaped curve. The standard normal density curve, which has mean $\mu = 0$ and standard deviation $\sigma = 1$, is shown in Figure 8.6. The standard normal cumulative distribution function is shown in Figure 8.7.

Figure 8.6 The standard normal p.d.f. (density curve).

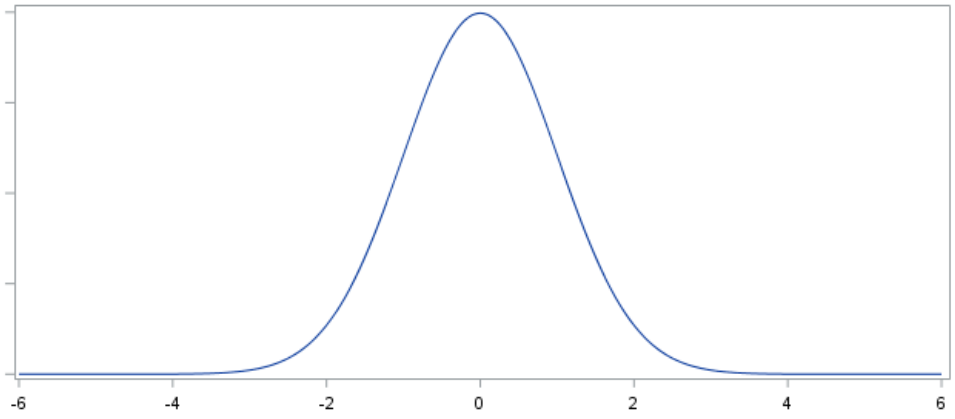
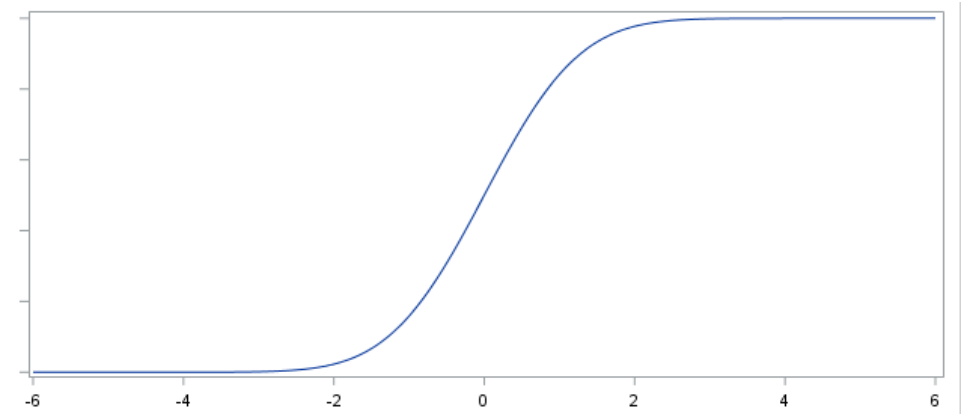


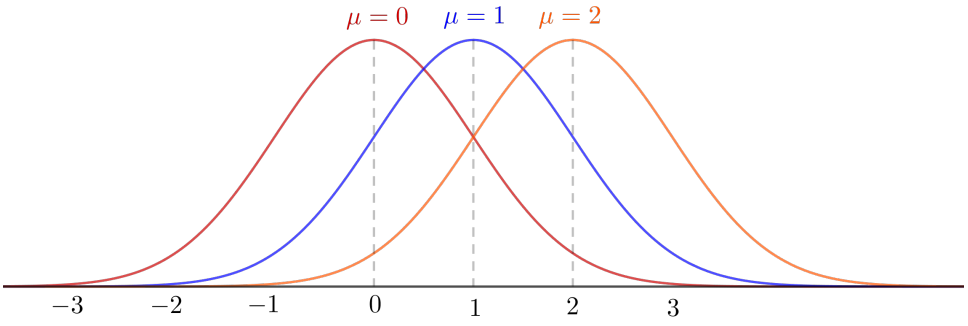
Figure 8.7 The standard normal c.d.f.



The normal distribution with mean μ and its density curve are symmetric around μ, *i.e.*, if we draw a vertical line through μ, then the two sides of the density curve are mirror images of each other. Therefore the mean of a normal distribution μ is also the median of the normal distribution. The mean μ locates the normal distribution on the number line so that if we hold σ constant and change the mean μ, the normal distribution is simply shifted

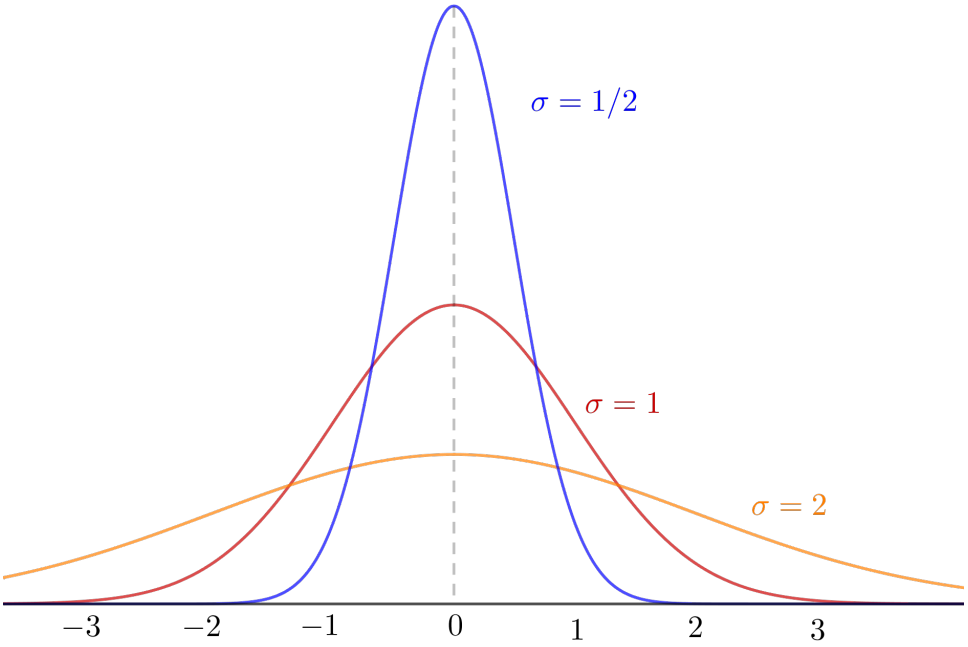
along the number line until it is centered at the new mean. In other words, holding σ fixed and changing μ simply relocates the density curve on the number line; it has no effect on the shape of the curve. Figure 8.8 provides the density curves for normal distributions with respective means $\mu = 0$, $\mu = 1$, and $\mu = 2$ and common standard deviation $\sigma = 1$.

Figure 8.8 Normal distributions with common standard deviation one and means of zero, one, and two.



The standard deviation σ indicates the amount of variability in the normal distribution. If we hold μ fixed and increase the value of σ, then the normal density curve becomes flatter, while retaining its bell–shape, indicating that there is more variability in the distribution. Similarly, if we hold μ fixed and decrease the value of σ, then the normal density curve becomes more peaked around the mean μ, while retaining its bell–shape, indicating that there is less variability in the distribution. Normal distributions with mean $\mu = 0$ and respective standard deviations $\sigma = 1/2$, $\sigma = 1$, and $\sigma = 2$ are plotted in Figure 8.9.

Figure 8.9 Normal distributions with common mean zero and standard deviations one–half, one, and two.



Computer programs and calculators can be used to compute normal probabilities or equivalently to compute areas under the normal density curve. These probabilities can also be calculated using tables of standard normal distribution probabilities such as Table 8.12. To use this table or a calculator which only handles the standard normal distribution, you need re–express probability statements in terms of a standard normal random variable. The relationship between the standard normal random variable Z and the general normal random variable X, when X has mean $E(X) = \mu$ and standard deviation σ, is

$$Z = \frac{X - \mu}{\sigma} \quad \text{or equivalently} \quad X = \mu + Z\sigma.$$

This relationship implies that a probability statement about the normal variable X can be re–expressed as a probability statement about the standard normal variable Z by re–expressing the statement in terms of standard deviation units from the mean. Given two constants $a < b$, observing a value of X between a and b (observing $a \leq X \leq b$) is equivalent to observing a value of $Z = (X - \mu)/\sigma$ between $(a - \mu)/\sigma$ and $(b - \mu)/\sigma$ (observing $(a - \mu)/\sigma \leq (X - \mu)/\sigma \leq (b - \mu)/\sigma$). Furthermore, $Z = (X - \mu)/\sigma$ behaves in accordance with the standard normal distribution so that the probability of observing

a value of X between a and b, denoted by $P(a \leq X \leq b)$, is equal to the probability that the standard normal variable Z takes on a value between $(a-\mu)/\sigma$ and $(b-\mu)/\sigma$, *i.e.*,

$$P(a \leq X \leq b) = P\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right).$$

In terms of areas this probability equality says that the area under the normal density curve with mean μ and standard deviation σ over the interval from a to b is equal to the area under the standard normal density curve over the interval from $(a-\mu)/\sigma$ to $(b-\mu)/\sigma$. Similarly, given constants $c < d$, we have the analogous result that

$$P(c \leq Z \leq d) = P(\mu + c\sigma \leq X \leq \mu + d\sigma).$$

Table 8.12 provides cumulative standard normal probabilities of the form $P(Z \leq a)$ for values of a (Z in the table) between 0 and 3.69. Computer programs usually produce cumulative probabilities like these. To use these cumulative probabilities to compute a probability of the form $P(a \leq Z \leq b)$ note that

$$P(a \leq Z \leq b) = P(Z \leq b) - P(Z \leq a)$$

and note that the symmetry of the normal distribution implies that

$$P(Z \leq -a) = P(Z \geq a) = 1 - P(Z \leq a).$$

Calculators often provide probabilities of the form $P(a \leq Z \leq b)$ directly.

Example 8.1 NHANES The National Health and Nutrition Examination Survey (NHANES) We will use some data from the 2013–2014 NHANES (see Example 1.1) to illustrate some cases where the normal distribution does and does not provide a suitable model for the distribution of a variable. We will consider the distributions of three variables: age, height, and weight. All available data for the 5588 adults (age 20 and over) in the 2013–2014 NHANES are used. For females the samples sizes are: $n = 2919$ for age and $n = 2888$ for height and weight. For males we have: $n = 2669$ for age, $n = 2642$ for height, and $n = 2645$ for weight. The data for the females is summarized graphically in Figure 8.10 and that for the males in Figure 8.11. Two density curves are superimposed on the histograms in the left column of these figures. There is a normal density curve, with μ and σ

chosen to match the data, and there is a "kernel" density curve which is basically a smooth version of the histogram. It is somewhat tricky to properly compare distributions based on histograms or density curves. Cumulative distribution functions, and associated graphs and statistics, form a better basis for such comparisons. The graphs in the right column contain an empirical (sample) cumulative distribution function and the cumulative distribution function for the fitted normal distribution. The empirical c.d.f is a "step function" based on the observed data values. The "steps" in this function are most pronounced in the graphs for the age distributions. The fact that the empirical (sample) and theoretical (normal) c.d.f.'s for the height distributions (females and males) are essentially indistinguishable indicates that the normal distribution provides a reasonable model for the height distribution. The weight distributions (females and males) are not modeled very well by a normal distribution. For both sexes the weight distributions are too strongly skewed right for the normal model to fit well. The age distributions are clearly non-normal. Actually, the reason that I included these age distributions is so that you could more easily see the appearance (steps) of an empirical c.d.f.

Figure 8.10 NHANES 2013–2014 height, weight, and age distributions for females

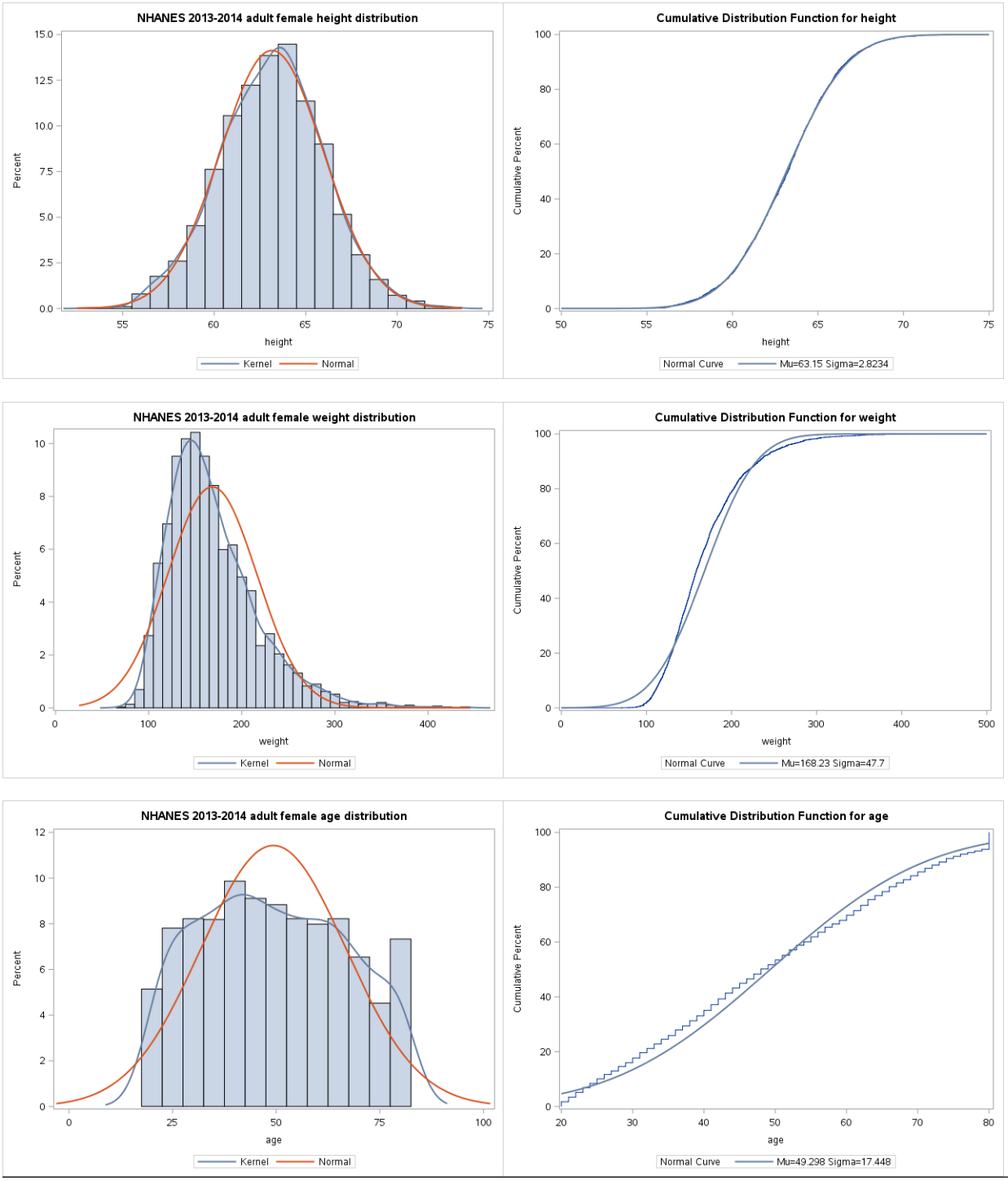
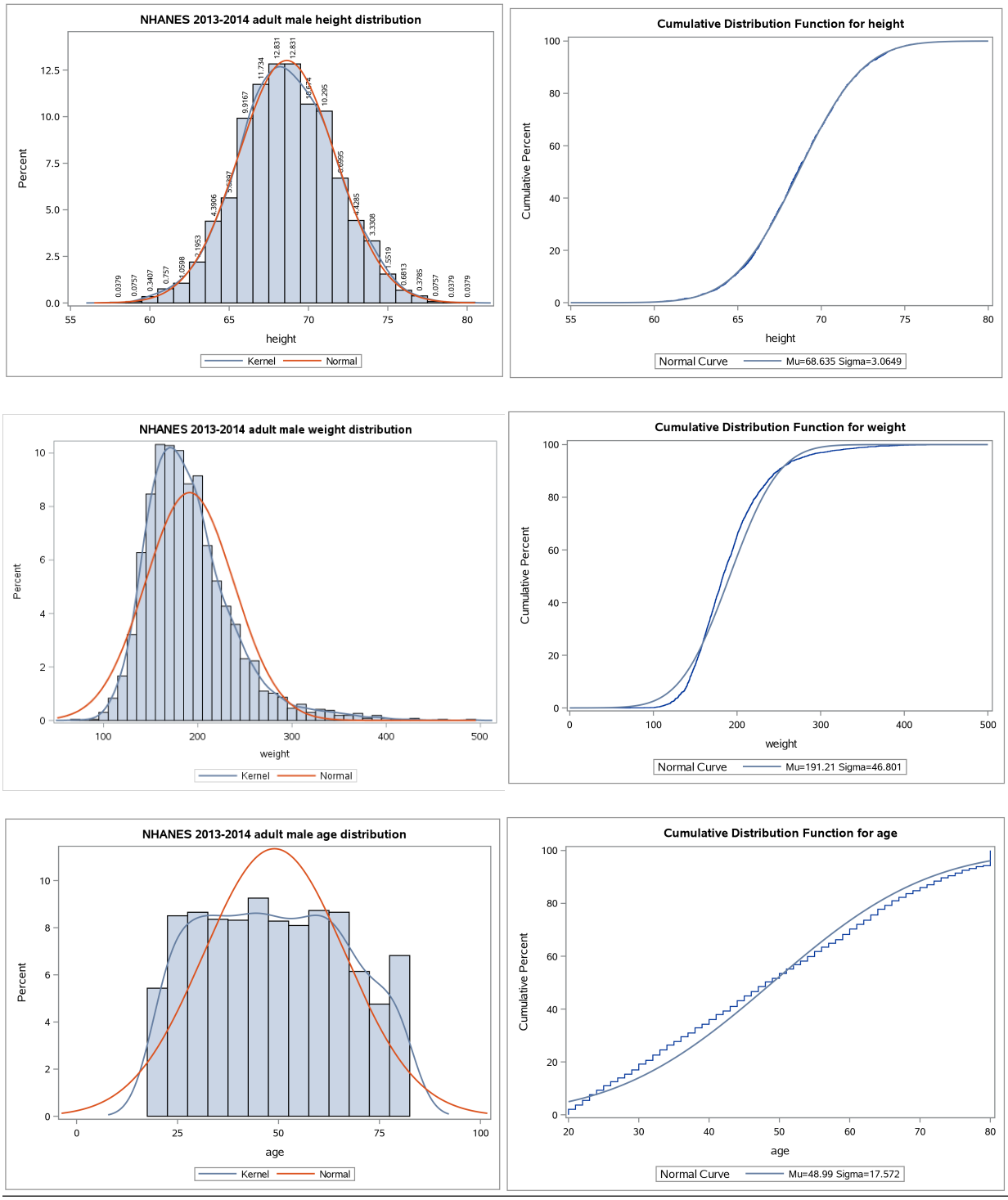


Figure 8.11 NHANES 2013–2014 height, weight, and age distributions for males



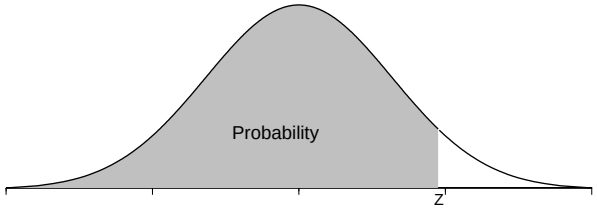


Table 8.12 Cumulative normal probabilities.
(Areas under the standard normal curve to the left of Z.)

Z	Second decimal place in Z									
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767

continued on next page

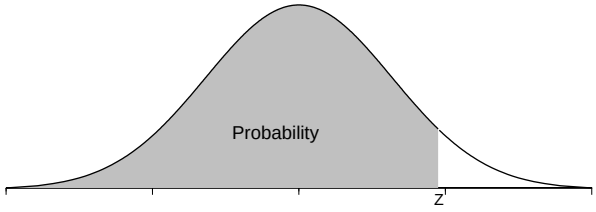


Table 8.12 (continuation) Cumulative normal probabilities.
(Areas under the standard normal curve to the left of Z.)

Z	Second decimal place in Z									
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998
3.6	.9998	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999

8.4 The exponential distribution toc

Exponential distributions are often used to model the time until the occurrence of an event. For example, we might use an exponential distribution to model the amount of time (starting from now) until an earthquake occurs or until a machine malfunctions.

The exponential distribution is defined for nonnegative values, that is, the sample space for an exponential random variable is the nonnegative part of the x–axis. An exponential distribution is determined by the value of a single parameter $\lambda > 0$. The p.d.f. of the exponential distribution with parameter $\lambda > 0$ is given by

$$f_X(x) = \lambda e^{-\lambda x}$$

for $x \geq 0$. If X denotes a random variable which follows this exponential distribution with parameter λ, then $E(X) = 1/\lambda$ and $\text{var}(X) = 1/\lambda^2$. The p.d.f's of exponential distributions with respective means 1/2, 1, and 2 (respective parameters $\lambda = 2$, $\lambda = 1$, and $\lambda = 1/2$) are graphed in Figure 8.13. Figure 8.14 shows the p.d.f and c.d.f of the exponential distribution with mean one.

Figure 8.13 Exponential distributions with means one–half, one, and two.

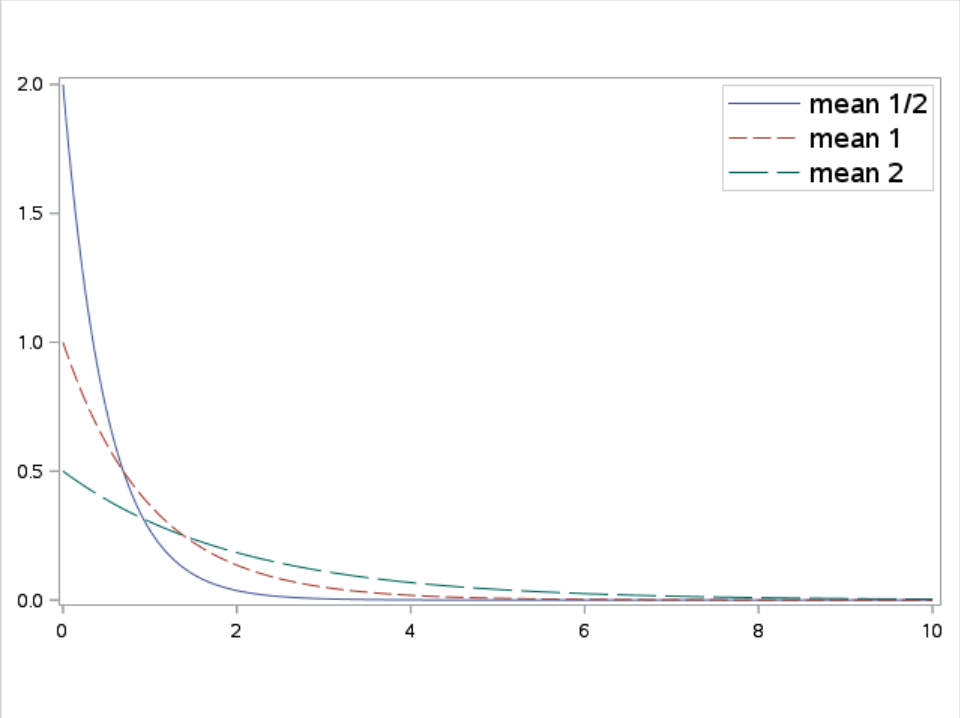
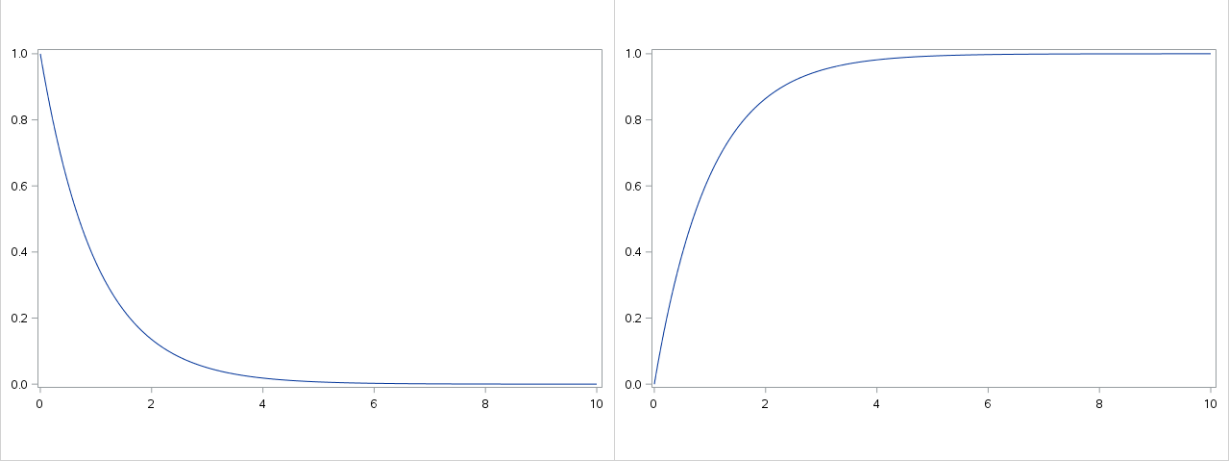


Figure 8.14 p.d.f. and c.d.f. of exponential distribution with mean one



There is an important connect between the Poisson distribution and the exponential distribution. As noted earlier, exponential distributions are often used to model the time until the occurrence of an event. In Section 7.6 we listed three assumptions about the times of occurrence of events which justify the use of the Poisson distribution to model a count of the number of occurrences of an event. If the events of interest behave in accordance with these assumptions, then the time until the occurrence of an event and the time between any two successive events will follow an exponential distribution.

9 Inference for a Proportion toc

9.1 Introduction toc

A **dichotomous population** is a collection of units which can be divided into two nonoverlapping subcollections corresponding to the two possible values of a dichotomous variable, *e.g.* male or female, dead or alive, pass or fail. It is conventional to refer to one of the two possible values which dichotomize the population as "success" and the other as "failure." These generic labels are not meant to imply that success is good. Rather, we can think of choosing one of the two possible classifications and asking "does the unit belong to the subcollection of units with this classification?" with the two possibilities being yes (a success) and no (a failure). When a unit is selected from the population and the unit is found to belong to the success subgroup we say that a **success** has occurred. Similarly, when a member of the failure subgroup is selected we say that a **failure** has occurred. The proportion of units in the population that belong to the success subgroup (the units classified as successes) is the **population success proportion**. This population success proportion is denoted by the lower case letter p. The population success proportion p is a parameter, since it is a numerical characteristic of the population. The **sample success proportion** or observed proportion of successes in a sample from a dichotomous population is denoted by $\hat{p}$ (read this as p hat). The observed success proportion $\hat{p}$ is a statistic, since it is a numerical characteristic of the sample.

We can use the selection of balls from a box of balls as a model for sampling from a dichotomous population. Consider a box containing balls of which some are red (successes) and the rest are green (failures). The population success proportion, p, is the proportion of red balls in the box. If we select a random sample of n balls from this box, then the sample success proportion, $\hat{p}$, is the proportion of red balls in the sample. Thus, in this model, a ball is a unit, the box of balls is the population, selecting a red ball is a success, the proportion of red balls in the box p is the parameter (the population success proportion), and the proportion of red balls in the sample $\hat{p}$ is the statistic (the sample success proportion).

9.2 The sampling distribution and the normal approximation toc

We will now discuss how the sample proportion $\hat{p}$ can be used to make inferences about the population proportion p. Assuming that the sample size n is reasonably large, it seems reasonable to view the sample proportion $\hat{p}$ as an estimate of the population proportion p. Clearly there will be some variability from sample to sample in the computed values of the statistic $\hat{p}$. That is, if we took several random samples from the population, we would not expect the observed sample success proportions, the $\hat{p}$'s, to be exactly the same. In the box of balls example, if we took several samples from the box we would expect the proportion of red balls in the sample to vary from sample to sample.

Two questions we might ask about the sample proportion $\hat{p}$ as an estimator of the population proportion p are:
(1) Can we expect the sample proportion $\hat{p}$ to be close to the population proportion p?
(2) Can we quantify how close $\hat{p}$ will be to p?
The sampling distribution of $\hat{p}$, which describes the sample to sample variability in $\hat{p}$, can be used to address these questions.

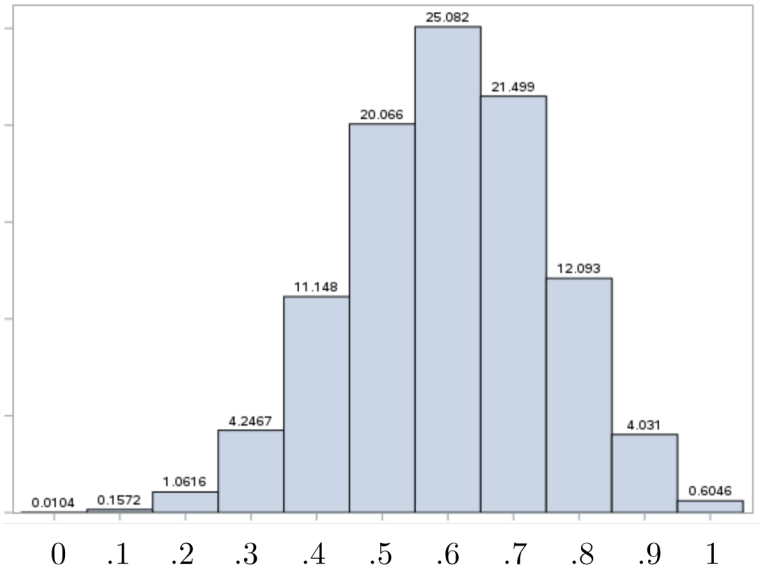
In general, the **sampling distribution of a statistic** is the distribution of the possible values of the statistic that could be obtained from random samples. We can think of the sampling distribution of a statistic as a theoretical relative frequency distribution for the possible values of the statistic which describes the sample to sample variability in the statistic. The form of the sampling distribution of a statistic depends on the nature of the population the sample is taken from, the size of the sample, and the method used to select the sample.

For example, suppose that we select a simple random sample of $n = 10$ balls, with replacement, from a box containing six red balls and four green balls. If we identify red as a success, then, in this example, the population success proportion is $p = .6$, since 60% of the balls in the box are red. The possible values of the sample success proportion $\hat{p}$ when $n = 10$ are: $0, .1, .2, \ldots, 1$. The sampling distribution of $\hat{p}$ summarized in Figure 9.1 gives the probabilities corresponding to these possible values. As you would expect, the probabilities are highest for values near $p = .6$ and very small for values far away from $p = .6$. In particular, when we select a simple random sample of size $n = 10$ from this box, 25.08% of the time we would observe $\hat{p} = .6$, 20.07% of the time we would observe $\hat{p} = .5$, and 21.50% of the time we would observe $\hat{p} = .7$. On the other hand, it would be possible

but very unlikely to observe $\hat{p} = .2$ (only 1.06% of the time) or $\hat{p} = .9$ (only 4.03% of the time).

Figure 9.1 Sampling distribution of $\hat{p}$ when $n = 10$ and $p = .6$

$\hat{p}$	probability
0	.0001
.1	.0016
.2	.0106
.3	.0425
.4	.1115
.5	.2007
.6	.2508
.7	.2150
.8	.1209
.9	.0403
1	.0060



The mean and the standard deviation of the sampling distribution are of particular interest. The mean of the sampling distribution indicates whether the statistic is biased as an estimator of the parameter of interest. If the mean of the sampling distribution is equal to the parameter of interest, then the statistic is said to be **unbiased** as an estimator of the parameter. Otherwise, the statistic is said to be **biased** as an estimator of the parameter. To say that a statistic is **unbiased** means that, even though the statistic will overestimate the parameter for some samples and will underestimate the parameter for other samples, it will do so in such a way that, in the long run, the values of the statistic will average to give the correct value of the parameter. When the statistic is **biased** the statistic will tend to consistently overestimate or consistently underestimate the parameter; therefore, in the long run, the values of a biased statistic will not average to give the correct value of the parameter. The standard deviation of the sampling distribution is known as the **standard error** of the statistic. The standard error of the statistic provides a measure of the sample to sample variability in the values of the statistic. The standard error of the statistic can be used to quantify how close we can expect the value of the statistic to be to the value of the parameter.

Returning to our discussion of the sampling distribution of $\hat{p}$ we first present two important properties of this sampling distribution. The observed proportion $\hat{p}$ in a simple random sample selected with replacement from a population with population proportion p has a sampling distribution with the following properties.

1. The mean of the sampling distribution of $\hat{p}$ is the population probability p. Therefore, $\hat{p}$ is unbiased as an estimator of p.

2. The population standard error of $\hat{p}$, denoted by $\text{SE}(\hat{p})$, is

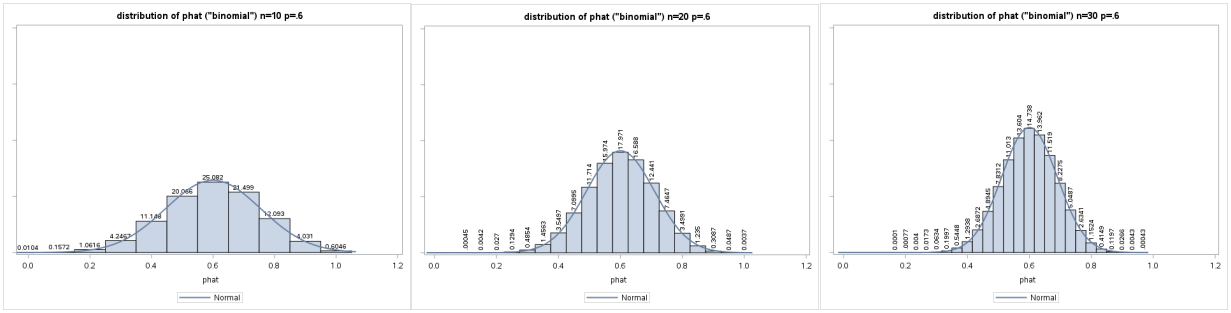
$$\text{SE}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}.$$

The inferential methods we will consider are based on a large sample size normal approximation to the sampling distribution of $\hat{p}$. A detailed discussion of the exact form of the sampling distribution of $\hat{p}$, the normal distribution, and the normal approximation to the sampling distribution of $\hat{p}$ can be found in a separate document. Here we will provide an indication of some basic properties of the sampling distribution of $\hat{p}$ in terms of some graphical representations of the probability histogram of the distribution of $\hat{p}$ with superimposed fitted normal density curves.

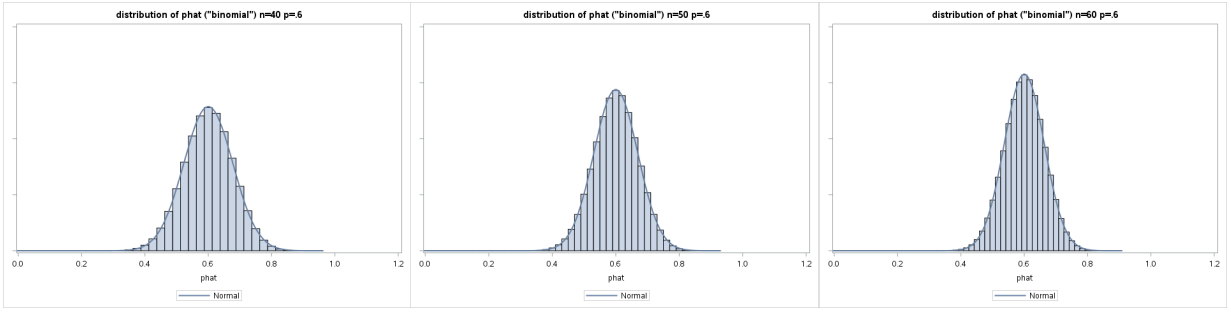
First consider how the sampling distribution of $\hat{p}$ depends on the sample size n. From the expression given above for the population standard error of $\hat{p}$ we can see that, as you would expect, the variability in $\hat{p}$ as an estimator of p decreases as the sample size increases. This behavior of the sampling distribution is illustrated in Figure 9.2.

Figure 9.2 The sampling distribution of $\hat{p}$ with normal approximation for $p = .6$. Histograms are provided for $n = 10, 20, 30, 40, 50$, and 60. All of these distributions are centered at $p = .6$. Notice how the variability in the distribution decreases as the sample size n increases and how much better the normal density curve matches the histogram.

$n = 10$ $n = 20$ $n = 30$



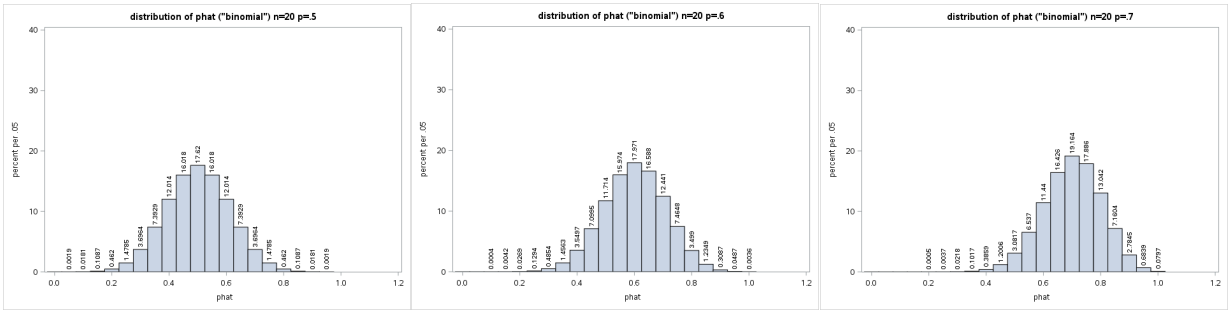
$n = 40$ $n = 50$ $n = 60$



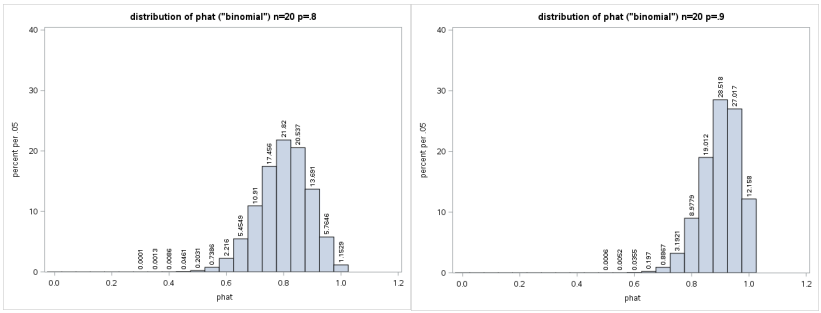
Next consider how the sampling distribution of $\hat{p}$ depends on the value of the population success proportion p. The sampling distribution of $\hat{p}$ is unimodal (single peaked) with its peak at p. If $p = .5$, then the sampling distribution of $\hat{p}$ is symmetric. If $p < .5$, then sampling distribution of $\hat{p}$ is skewed right. If $p > .5$, then sampling distribution of $\hat{p}$ is skewed left. Since $\hat{p}$ is unbiased as an estimator of p, the mean of the distribution of $\hat{p}$ is p. Thus, the probability histogram of the distribution of $\hat{p}$ has its balance point at p. From the expression given above for the population standard error of $\hat{p}$ we can see that, for fixed n, the variability in $\hat{p}$ as an estimator of p is highest when $p = .5$ and decreases as p moves away from .5. These properties of the sampling distribution are illustrated in Figure 9.3.

Figure 9.3 The sampling distribution of $\hat{p}$ with normal approximation for $n = 20$. Histograms are provided for $p = .5, .6, .7, .8$, and $.9$. Each distribution is centered at its p. Notice how the variability in the distribution decreases as p moves away from .5. Notice also how the distribution becomes more skewed left as p moves away from .5.

$p = .5$ $p = .6$ $p = .7$

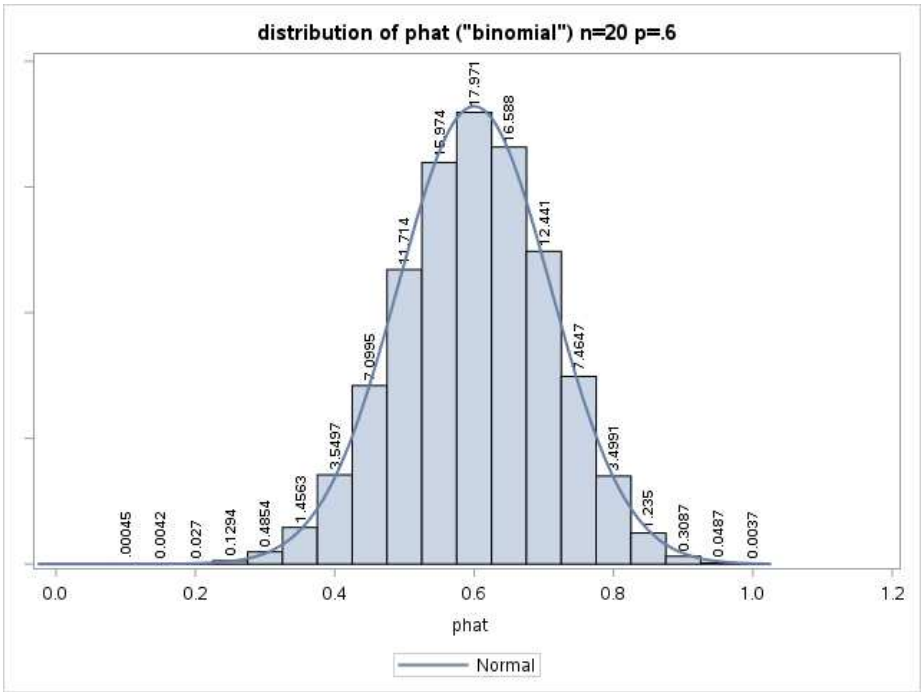


$p = .8$ $p = .9$



The normal approximation to the sampling distribution of $\hat{p}$ is illustrated graphically in Figure 9.4. In this figure, the probability histogram of the sampling distribution of $\hat{p}$ for the situation when a simple random sample of size $n = 20$ is selected with replacement from a dichotomous population with $p = .6$ is given along with the approximating normal density curve. The exact probabilities which correspond to the areas of the rectangles of the histogram are given as percentages, *e.g.*, the probability of observing $\hat{p} = .06$ is .17971 (17.97%). The normal approximation, explained more fully below, basically says that, over a specified range of $\hat{p}$ values, the area under the normal density curve and the area in the rectangles of the histogram are similar. You can see that the curve matches the histogram reasonably well in Figure 9.4.

Figure 9.4 The sampling distribution of $\hat{p}$ with normal approximation for $n = 20$ and $p = .6$. This assumes a simple random sample selected with replacement. If the $\hat{p}$ values were converted to counts, $X = n\hat{p}$, then this would be a binomial distribution.



The normal approximation to the sampling distribution of $\hat{p}$ says that, for large values of n, the standardized value of $\hat{p}$ obtained by subtracting the population proportion p from $\hat{p}$ and dividing this difference by the population standard error of $\hat{p}$, behaves in approximate accordance with the standard normal distribution. That is, for large values of n the quantity $Z = (\hat{p} - p)/\text{SE}(\hat{p})$ behaves like a standard normal variable. This means, as will be shown below, that we can use an area under the standard normal density curve (a probability in terms of Z) to approximate an area in the probability histogram of the sampling distribution of $\hat{p}$ (a probability in terms of $\hat{p}$). The relationship between $\hat{p}$ and p indicated by this expression for Z and the normal distribution itself allow us to use $\hat{p}$ to make formal, quantifiable inferences about p.

9.3 Estimation of p

toc

As noted above, the sample proportion $\hat{p}$ is an unbiased estimator of the population proportion p. We can think of $\hat{p}$ as our "best guess" of the value of p. To allow for sampling variability it would be more useful to report a range or interval of plausible values for p.

In particular, given the data we would like to be able to say, with a reasonable level of confidence, that the true value of p is between two particular limiting values. We will now use the normal approximation to the sampling distribution of $\hat{p}$ to develop such an interval estimate of p.

The probability that a standard normal variable Z takes on a value between -1.96 and 1.96 is equal to .95, *i.e.*, $P(-1.96 \leq Z \leq 1.96) = .95$. Thus, when we observe the value of a standard normal variable Z, 95% of the time we will find that $-1.96 \leq Z \leq 1.96$. Graphically this means that the area under the standard normal density curve over the interval from -1.96 to 1.96 is .95. Thus, for sufficiently large values of n we have the approximation,

$$P\left[-1.96 \leq \frac{\hat{p} - p}{\text{SE}(\hat{p})} \leq 1.96\right] = .95$$

or equivalently

$$P\left[p - 1.96 \cdot \text{SE}(\hat{p}) \leq \hat{p} \leq p + 1.96 \cdot \text{SE}(\hat{p})\right] = .95.$$

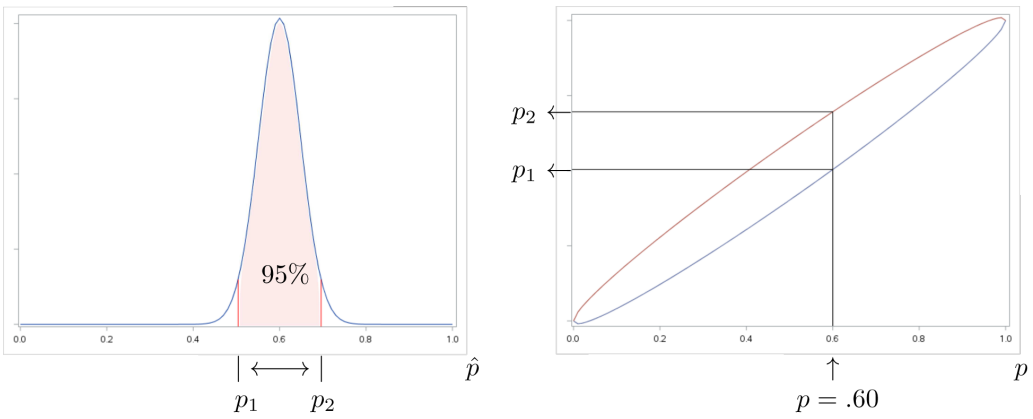
Note that this indicates that 95% of the time when a simple random sample is selected and $\hat{p}$ is computed the observed value of $\hat{p}$ will be between $p - 1.96 \cdot \text{SE}(\hat{p})$ and $p + 1.96 \cdot \text{SE}(\hat{p})$, *i.e.*, $\hat{p}$ will be within 1.96 population standard error units of p. We will refer to the interval from $p - 1.96 \cdot \text{SE}(\hat{p})$ to $p + 1.96 \cdot \text{SE}(\hat{p})$ as the central 95% interval of the distribution of $\hat{p}$, since it is centered at p and it will contain the observed value of $\hat{p}$ 95% of the time.

The relationship between $\hat{p}$ and p is illustrated, for $n = 100$ and $p = .6$, in Figure 9.5. In this case, $p = .6$, $\text{SE}(\hat{p}) = \sqrt{.6(.4)/100} = .0490$, and $p_1 = .6 - 1.96 \cdot \text{SE}(\hat{p}) = .5040$ and $p_2 = .6 + 1.96 \cdot \text{SE}(\hat{p}) = .6960$ are the limits of the central 95% interval of the distribution of $\hat{p}$. The shaded region with area .95 in Figure 9.5 indicates that 95% of all samples will yield a sample proportion $\hat{p}$ which is between p_1 and p_2. That is, when $p = .6$ and $n = 100$, 95% of all samples will yield a sample proportion $\hat{p}$ which is between $p_1 = .5040$ and $p_2 = .6960$. In terms of a box of balls this means that, if exactly 60% of the balls in the box were red, then 95% of the time when we selected $n = 100$ balls from the box, at random and with replacement, we would find between 51 and 69 red balls (between 50.40 and 69.60) among the 100 balls in the sample.

The plot on the right in Figure 9.5 shows how the endpoints of the central 95% interval of the distribution of $\hat{p}$ depend on p. In this plot the sample size is $n = 100$. (The pattern is similar for other values of n.) The upper (red) curve gives values of $p + 1.96 \cdot \text{SE}(\hat{p}) = p + 1.96\sqrt{p(1-p)/n}$ as a function of p and the lower (blue) curve gives values of

$p - 1.96 \cdot \text{SE}(\hat{p}) = p - 1.96\sqrt{p(1-p)/n}$ as a function of p. The intersections of a vertical line draw at a particular value of p with these curves are the endpoints of the central 95% interval of the distribution of $\hat{p}$ for that value of p. The lines drawn in the figure demonstrate this for the case $p = .6$. Note that, as mentioned above, $p_1 = .5040$ and $p_2 = .6960$, and these are the endpoints of the interval in the plot on the left in Figure 9.5.

Figure 9.5 The plot on the left shows the central 95% interval of the distribution of $\hat{p}$ for $n = 100$ and $p = .6$.
The curves in the plot on the right show the endpoints, $p \pm 1.96\sqrt{p(1-p)/n}$, of the central 95% interval of the distribution of $\hat{p}$ as a function of p for $n = 100$.
The endpoints for the case $p = .6$ are indicated by the lines marking the intersections at $p_1 = .5040$ and $p_2 = .6960$.



A confidence interval for p

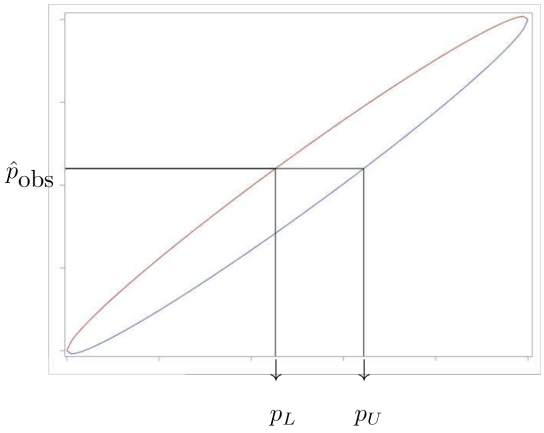
As noted above, for a particular value of p, 95% of all samples will yield a value of $\hat{p}$ within the corresponding central 95% interval from $p - 1.96\text{SE}(\hat{p})$ to $p + 1.96\text{SE}(\hat{p})$. Thus for each possible value of p we can find an interval of likely values for $\hat{p}$. But we need an interval of plausible values for p not for $\hat{p}$! We will now show how the central 95% intervals of the distribution of $\hat{p}$ can be used to form a 95% confidence interval estimate of p. We will provide a formal definition of a 95% confidence interval estimate later.

Suppose that a simple random sample has been selected and let $\hat{p}_{\text{obs}}$ denote the observed value of $\hat{p}$. Given this $\hat{p}_{\text{obs}}$, we want to know which values of p are plausible. More precisely, we want to know which values of p determine sampling distributions for $\hat{p}$ under which seeing $\hat{p} = \hat{p}_{\text{obs}}$ would not be surprising. We can formalize this goal by saying that

we want to know which values of p determine a sampling distributions for $\hat{p}$ for which the central 95% interval of the distribution of $\hat{p}$ contains the observed value $\hat{p}_{\text{obs}}$.

The plot on the right in Figure 9.5 shows how the central 95% interval of the distribution of $\hat{p}$ depends on the value of p. We want to determine which values of p yield central 95% intervals which contain $\hat{p}_{\text{obs}}$. This requires using the graph of Figure 9.5 in the other direction. In Figure 9.6 a horizontal line is drawn at $\hat{p}_{\text{obs}}$ and its intersections, p_L and p_U, with the two curves are indicated. Notice that p_L is the smallest value of p for which the central 95% interval of the distribution of $\hat{p}$ contains $\hat{p}_{\text{obs}}$ and p_U is the largest value of p for which the central 95% interval of the distribution of $\hat{p}$ contains $\hat{p}_{\text{obs}}$. (Figure 9.6 is drawn for $n = 100$ and $\hat{p}_{\text{obs}} = .55$. In this case, $p_L = .4524$ and $p_U = .6439$.) Thus, if we draw a vertical line, as in Figure 9.5, at any value of p between p_L and p_U, then the corresponding central 95% interval of the distribution of $\hat{p}$ contains $\hat{p}_{\text{obs}}$.

Figure 9.6 The smallest and largest values of p, p_L and p_U, for which the central 95% interval of the distribution of $\hat{p}$ contains $\hat{p}_{\text{obs}}$.
This example is drawn for $n = 100$ and $\hat{p}_{\text{obs}} = .55$.



In order to determine the value of p_L we simply set $p + 1.96\text{SE}(\hat{p})$ equal to $\hat{p}_{\text{obs}}$ and solve for p (draw the horizontal line as in Figure 6 and project down at the intersection with the upper (red) curve). To determine the value of p_U we set $p - 1.96\text{SE}(\hat{p})$ equal to $\hat{p}_{\text{obs}}$ and solve for p (draw the horizontal line as in Figure 9.6 and project down at the intersection with the lower (blue) curve).

Table 9.1 Central 95% intervals $[p_1, p_2]$ of the distribution of $\hat{p}$ for selected values of p when $n = 100$. The intervals for $p = .46$ through $p = .64$ contain $\hat{p}_{\text{obs}} = .55$.

	central 95% interval		
p	p_1	p_2	
.10	.041	.159	
.20	.122	.278	
.30	.210	.390	
.40	.304	.496	
.45	.352	.548	***
.46	.362	.558	
.47	.372	.568	
.48	.382	.578	
.49	.392	.588	
.50	.402	.598	
.51	.412	.608	
.52	.422	.618	
.53	.432	.628	
.54	.442	.638	
.55	.452	.648	
.56	.463	.657	
.57	.473	.667	
.58	.483	.677	
.59	.494	.686	
.60	.504	.696	
.61	.514	.706	
.62	.525	.715	
.63	.535	.725	
.64	.546	.734	
.65	.556	.743	***
.70	.610	.790	
.80	.722	.878	
.90	.841	.959	

A simple example will help to clarify this discussion. Consider a box containing a large number of balls. Suppose that some of the balls in the box are red. Let p denote the proportion of red balls in the box. Now suppose that a simple random sample of $n = 100$ balls has been selected and 55 of the 100 balls found to be red. In this example 55% of the balls in the sample are red, *i.e.*, $\hat{p}_{\text{obs}} = .55$. Table 9.1 contains central 95% intervals of the distribution of $\hat{p}$ for $n = 100$ and selected values of p. Notice that when p is between .46 and .64 the intervals contain $\hat{p}_{\text{obs}} = .55$.

For example, when $p = .50$, 95% of all samples will yield a $\hat{p}$ value between .402 and .598. In a case like this, where the central interval contains .55, it would not be surprising to see $\hat{p}_{\text{obs}} = .55$ and the correspond value of p would be deemed plausible. That is, if exactly 50% of the balls in the box were red, then 95% of the time when we selected a random sample of $n = 100$ balls we would find that between 40.2% and 59.8% of the 100 balls in the sample were red. Since 55% belongs to this range it is plausible that exactly 50% of all the balls in the box are red ($p = .50$).

On the other hand, if $p = .30$, 95% of all samples will yield a $\hat{p}$ value between .210 and .390. Thus, if exactly 30% of the balls in the box were red, then 95% of the time when we selected a random sample of $n = 100$ balls we would find that between 21.0% and 39.0% of the 100 balls in the sample were red. Since this entire interval is less than 55%, it would be surprising to see $\hat{p}$ as large as .55 when p was .30. Therefore, when we see $\hat{p}_{\text{obs}} = .55$ it is reasonable to conclude that the percentage of red balls in the box is not 30% ($p \neq .30$).

Similarly, if $p = .70$, 95% of all samples will yield a $\hat{p}$ value between .610 and .790. Thus, if exactly 70% of the balls in the box were red, then 95% of the time when we selected a random sample of $n = 100$ balls we would find that between 61.0% and 79.0% of the 100 balls in the sample were red. Since this entire interval is greater than 55%, it would be surprising to see $\hat{p}$ as small as .55 when p was .70. Therefore, when we see $\hat{p}_{\text{obs}} = .55$ it is reasonable to conclude that the percentage of red balls in the box is not 70% ($p \neq .70$).

In cases like the last two, where the entire interval is less than .55 or the entire interval is greater than .55, it would be surprising to see $\hat{p}_{\text{obs}} = .55$ and the correspond value of p would not be deemed plausible. From the table we can see that $p_L \approx .45$, since the upper endpoint of the central 95% interval p_2 is approximately .55 when $p = .45$, and $p_U \approx .64$, since the lower endpoint p_1 of the central 95% interval is approximately .55 when $p = .64$. More precise computation indicates that $p_L = .4524$ and $p_U = .6439$. Using this reasoning, we see that when $n = 100$ and we observe $\hat{p}_{\text{obs}} = .55$ we can argue that values of p between $p_L = .4524$ and $p_U = .6439$ are plausible. In other words, if we selected a simple random sample of $n = 100$ balls from the box and found that 55% of the balls in the sample were red, then we would conclude that somewhere between 45.24% and 64.39% of all the balls in the box are red.

The interval of plausible values for p with endpoints p_L and p_U discussed above is a 95% confidence interval estimate of p. In the preceding example, the more formal way of stating

the conclusion is as follows. If we selected a simple random sample of $n = 100$ balls from the box and found that 55% of the balls in the sample were red, then we would conclude that we were 95% confident that the percentage of red balls in the box was somewhere between 45.24% and 64.39% (in terms of p, $.4524 \leq p \leq .6439$).

A confidence interval for p – more formally

We will now give a more formal definition of a confidence interval estimate of p. The purpose of a confidence interval estimate is to provide a range or interval of plausible values for p. In particular, given the data we would like to be able to say, with a reasonable level of confidence (95% in the example above), that the true value of p is between two particular values (p_L and p_U in the example above). A confidence interval estimate of p consists of two parts. There is an interval of plausible values for p and a corresponding level of confidence. We will adopt the usual convention of using a confidence level of 95%. The confidence level indicates our confidence that the unknown p actually belongs to the corresponding interval.

There is some chance for confusion about what it means to say we are 95% confident that p is between p_L and p_U. The important thing to remember is that it is the endpoints of the interval p_L and p_U that vary from sample to sample. The population proportion p is a fixed, unknown parameter which does not vary. Therefore, the 95% confidence level applies to the method used to generate the confidence interval estimate. That is, the method (obtain a simple random sample and compute the numbers p_L and p_U forming the confidence interval is such that 95% of the time it will yield a pair of confidence interval limits which bracket the population success proportion p. Therefore, when we obtain a sample, compute the confidence interval, and say that we are 95% confident that this interval contains p what we mean is that we feel "pretty good" about claiming that p is in this interval, since the method used to construct the interval works for 95% of all possible samples and so it probably worked for our sample.

A 95% confidence interval for p – summary and computation.

As noted above, a 95% confidence interval estimate of p is an interval of plausible values for p constructed using a method of generating such intervals with the property that this method will work, in the sense of generating an interval that contains p, for 95% of all possible samples.

The Wilson (score) 95% confidence interval for p.

The 95% confidence interval estimate of the population proportion p discussed above is usually called the Wilson interval or the score interval. The graphical derivation above is readily modified for a confidence level other that 95%. The requisite modification is to change the 95% confidence level multiplier 1.96 to the value appropriate for the desired confidence level, *e.g.*, the multiplier 1.645 leads to a 90% confidence interval.

Computation of the Wilson (score) 95% confidence interval for p.

Computations of the Wilson 95% confidence interval estimate of p for the box of balls example with $n = 100$ and 55 red balls are illustrated in Figures 9.7–9.10. Recall that in this example, we selected a simple random sample of $n = 100$ balls from the box and found that 55% of the balls in the sample were red, then we concluded that we were 95% confident that the percentage of red balls in the box was somewhere between 45.24% and 64.39% (in terms of p, $.4524 \leq p \leq .6439$).

Figure 9.7 SAS output giving the Wilson 95% confidence interval for the box of balls example with $n = 100$ and $\hat{p}_{\text{obs}} = .55$ with SAS command file.

box of balls example

The FREQ Procedure

outcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
red	55	55.00	55	55.00
green	45	45.00	100	100.00

Binomial Proportion	
outcome = red	
Proportion	0.5500
ASE	0.0497

Confidence Limits for the Binomial Proportion		
Proportion = 0.5500		
Type	95% Confidence Limits	
Wilson	0.4524	0.6439

SAS command file: boxofballs_onepropexample.sas

```
/* box of balls example */
data boxofballs;
   input outcome $ count;
/* the coding outcome $ indicates that outcome is character valued */
cards;
red 55
green 45
;
proc freq data=boxofballs order=data;
   tables outcome / alpha=.05 binomial(cl=wilson level='red');
   weight count;
title 'box of balls example';
/* tables option
   -- alpha=.05 requests 95% confidence level
        this is the default Use alpha=.10 to
        get a 90% confidence interval and use
        the appropriate endpoint of the 90% CI
        as a 95% lower/upper confidence bound */
/* binomial options
   -- cl=wilson Wilson confidence interval
   -- level='red' define "success" default is
         the first level which is "red" in this example */
run;
/* ---------------------------------------------------------*/
```

Figure 9.8 R commands and output giving the Wilson 95% confidence interval for the box of balls example with $n = 100$ and $\hat{p}_{\text{obs}} = .55$.

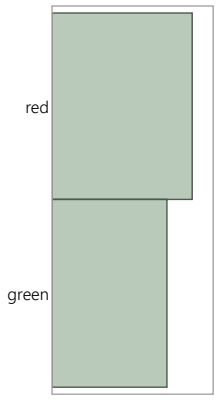
R commands and output for the box of balls example

```
R Commands:

# box of balls example
prop.test(55,100,correct=0)

R Output:

1-sample proportions test without continuity correction

data:  55 out of 100, null probability 0.5
X-squared = 1, df = 1, p-value = 0.3173
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4524460 0.6438546
sample estimates:
   p
0.55
```

Figure 9.9 JMP commands and output giving the Wilson 95% confidence interval for the box of balls example with $n = 100$ and $\hat{p}_{\text{obs}} = .55$.

Distributions

color



Frequencies		
Level	Count	Prob
green	45	0.45000
red	55	0.55000
Total	100	1.00000

N Missing 0

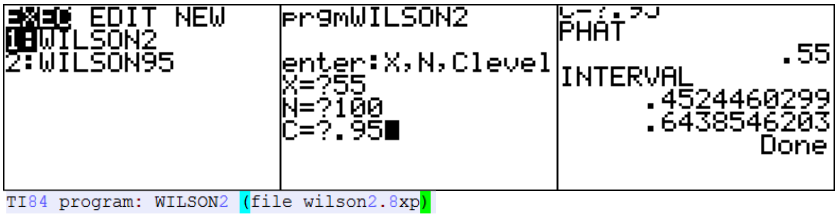
2 Levels

Confidence Intervals					
Level	Count	Prob	Lower CI	Upper CI	1-Alpha
green	45	0.45000	0.356145	0.547554	0.950
red	55	0.55000	0.452446	0.643855	0.950
Total	100				

Note: Computed using score confidence intervals.

```
JMP directions for the box of balls example

step 1: create a data table: file --> new --> data table
input in columns 1 and 2 as:

color count
red    55
green  45

step 2: analyze --> distribution
assign variables as:

y, columns : color
freq : count

under the "color tab" select options
display options -->  horizontal layout
histogram options -->  vertical
confidence interval -->  .95
```

Figure 9.10 TI84 program "WILSON2" screen captures followed by command listing. This output shows how to get the Wilson 95% confidence interval for the box of balls example with $n = 100$ and $\hat{p}_{\text{obs}} = .55$.



TI84 program: WILSON2 (file wilson2.8xp)

```
Disp "enter:X,N,Clevel"
Prompt X,N,C
X/N→R
C+(1-C)/2→P
invNorm(P)→K
(X+K^2/2)/(N+K^2)→Y
(Y^2-X^2/(N*(N+K^2)))^.5→W
Y-W→L
Y+W→U
Disp "PHAT",R
Disp "INTERVAL",L,U
```

Example 9.1 Insects in an apple orchard. The manager of a large apple orchard is concerned with the presence of a particular insect pest in the apple trees in the orchard. An insecticide

that controls this particular insect pest is available. However, application of this insecticide is rather expensive. It has been determined that the cost of applying the insecticide is not economically justifiable unless more than 20% of the apple trees in the orchard are infested. The manager has decided to assess the extent of infestation in the orchard by examining a simple random sample of 200 apple trees. In this example a unit is an apple tree and the target population is all of the apple trees in this orchard. We will assume that the simple random sample is selected from all of the apple trees in the orchard so that the sampled population is the same as the target population. We will also assume that the 200 trees in the sample form a small proportion of all of the trees in the entire orchard so that we do not need to worry about whether the sample is chosen with or without replacement. An appropriate dichotomous variable is whether an apple tree is infested with possible values of yes (the tree is infested) and no (the tree is not infested). Since we are interested in the extent of the infestation we will view a tree that is infested as a success. Thus, the population success proportion p is the proportion of all of the apple trees in the entire orchard that are infested.

Two (related) questions of interest in this situation are:

(1) What proportion of all of the trees in this orchard are infested? (What is p?)

(2) Is there sufficient evidence to justify the application of the insecticide? (Is $p > .20$?)

We will consider four hypothetical outcomes for this scenario to demonstrate how a 95% confidence interval estimate can be used to address these questions. The SAS output for these examples is provided in Figures 9.11a and 9.11b.

Figure 9.11a SAS output for the apple orchard example – confidence intervals

apple orchard example

The FREQ Procedure

case=1

outcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
infested	35	17.50	35	17.50
notinfested	165	82.50	200	100.00

Binomial Proportion	
outcome = infested	
Proportion	0.1750
ASE	0.0269

Confidence Limits for the Binomial Proportion		
Proportion = 0.1750		
Type	95% Confidence Limits	
Wilson	0.1286	0.2336

apple orchard example

The FREQ Procedure

case=2

outcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
infested	26	13.00	26	13.00
notinfested	174	87.00	200	100.00

Binomial Proportion	
outcome = infested	
Proportion	0.1300
ASE	0.0238

Confidence Limits for the Binomial Proportion		
Proportion = 0.1300		
Type	95% Confidence Limits	
Wilson	0.0903	0.1837

Case 1. Suppose that 35 of the 200 apple trees in the sample are infested so that $\hat{p} = .175$. In this case we know that 17.5% of the 200 trees in the sample are infested and we can conjecture that a similar proportion of all of the trees in the entire orchard are infested. However, we need a confidence interval estimate to get a handle on which values of the population success proportion p are plausible when we observe 17.5% infested trees in a sample of size 200. Using the Wilson method we get a 95% confidence interval ranging from .1286 to .2336 (see Figure 9.11a). Thus we can conclude that we are 95% confident that between 12.86% and 23.36% of all of the trees in this orchard are infested. Notice that this confidence interval does not exclude the possibility that more than 20% of the trees in the entire orchard are infested, since the upper limit of the confidence interval 23.36% is greater than 20%. In other words, even though less than 20% of the trees in the sample were infested, when we take sampling variability into account we find that it is possible that more than 20% (as high as 23.36%) of the trees in the entire orchard are infested. Of course the interval also indicates that it is possible that less than 20% (as low as 12.86%) of the trees in the entire orchard are infested.

Case 2. Suppose that 26 of the 200 apple trees in the sample are infested so that $\hat{p} = .13$. In this case we know that 13% of the 200 trees in the sample are infested. Using the Wilson method we get a 95% confidence interval ranging from .0903 to .1837 (see Figure 9.11a). Thus we can conclude that we are 95% confident that between 9.03% and 18.37% of all of the trees in this orchard are infested. In this case the entire confidence interval is below 20% excluding the possibility that more than 20% of the trees in the entire orchard are infested. Therefore, in this case we have sufficient evidence to conclude that less than 20% of the trees in the entire orchard are infested, *i.e.*, that $p < .20$.

Figure 9.11b SAS output for the apple orchard example – confidence intervals

apple orchard example

The FREQ Procedure

case=3

outcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
infested	45	22.50	45	22.50
notinfested	155	77.50	200	100.00

Binomial Proportion	
outcome = infested	
Proportion	0.2250
ASE	0.0295

Confidence Limits for the Binomial Proportion		
Proportion = 0.2250		
Type	95% Confidence Limits	
Wilson	0.1726	0.2877

apple orchard example

The FREQ Procedure

case=4

outcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
infested	54	27.00	54	27.00
notinfested	146	73.00	200	100.00

Binomial Proportion	
outcome = infested	
Proportion	0.2700
ASE	0.0314

Confidence Limits for the Binomial Proportion		
Proportion = 0.2700		
Type	95% Confidence Limits	
Wilson	0.2132	0.3354

Case 3. Suppose that 45 of the 200 apple trees in the sample are infested so that $\hat{p} = .225$. In this case we know that 22.5% of the 200 trees in the sample are infested. Using the Wilson method we get a 95% confidence interval ranging from .1726 to .2877 (see Figure 9.11b). Thus we can conclude that we are 95% confident that between 17.26% and 28.77% of all of the trees in this orchard are infested. Notice that this confidence interval does not exclude the possibility that less than 20% of the trees in the entire orchard are infested, since the lower limit of the confidence interval 17.26% is less than 20%. In other words, even though more than 20% of the trees in the sample were infested, when we take sampling variability into account we find that it is possible that less than 20% (as low as 17.26%)

of the trees in the entire orchard are infested. Of course the interval also indicates that it is possible that more than 20% (as high as 28.77%) of the trees in the entire orchard are infested.

Case 4. Finally, suppose that 54 of the 200 apple trees in the sample are infested so that $\hat{p} = .27$. In this case we know that 27% of the 200 trees in the sample are infested. Using the Wilson method we get a 95% confidence interval ranging from .2132 to .3354 (see Figure 9.11b). Thus we can conclude that we are 95% confident that between 21.32% and 33.54% of all of the trees in this orchard are infested. In this case the entire confidence interval is above 20% excluding the possibility that less than 20% of the trees in the entire orchard are infested. Therefore, in this case we have sufficient evidence to conclude that more than 20% of the trees in the entire orchard are infested, *i.e.*, that $p > .20$.

A formula for the Wilson (score) 95% confidence interval for p.

Some readers may find an algebraic expression for the Wilson confidence interval useful. Refer to the computer code and output of Figures 9.7–9.10 for an illustration of the computations of the Wilson 95% confidence interval for a simple example. For greater generality we will provide these expressions in terms of k, where k is the multiplier for the desired confidence level. For a 95% confidence level $k = 1.96$ and for a 90% confidence level $k = 1.645$. The Wilson confidence interval estimate of p is given by

$$\tilde{p}_k - \text{ME}(\tilde{p}_k) \leq p \leq \tilde{p}_k + \text{ME}(\tilde{p}_k),$$

(read $\tilde{p}_k$ as p tilde sub k) where

$$\tilde{p}_k = \frac{n\hat{p} + \frac{k^2}{2}}{n + k^2}$$

determines the center of the interval, and the **margin of error of** $\tilde{p}_k$

$$\text{ME}(\tilde{p}_k) = \sqrt{\tilde{p}_k^2 - \frac{n\hat{p}^2}{n + k^2}}$$

determines the length of the interval.

If we use $k = 1.96$ in these expressions, then we can claim that we are 95% confident that the population success proportion p is between $\tilde{p}_k - \text{ME}(\tilde{p}_k)$ and $\tilde{p}_k + \text{ME}(\tilde{p}_k)$. As noted above, there is some chance for confusion about what this statement actually means. The important thing to remember is that it is the statistic $\tilde{p}_k$ and the margin of error

$\text{ME}(\tilde{p}_k)$ that vary from sample to sample. The population proportion p is a fixed, unknown parameter which does not vary. Therefore, the 95% confidence level applies to the method used to generate the confidence interval estimate. That is, the method (obtain a simple random sample and compute the numbers $\tilde{p} - \text{ME}(\tilde{p})$ and $\tilde{p} + \text{ME}(\tilde{p})$) used to generate the limits of the confidence interval is such that 95% of the time it will yield a pair of confidence interval limits which bracket the population success proportion p. Therefore, when we obtain a sample, compute the confidence interval, and say that we are 95% confident that this interval contains p what we mean is that we feel "pretty good" about claiming that p is in this interval, since the method works for 95% of all possible samples and so it probably worked for our sample.

Aside – derivation of the Wilson interval formula.

This aside contains an algebraic derivation of the Wilson confidence interval for p. The starting point for this derivation is the interval $p - k\text{SE}(\hat{p}) \leq \hat{p} \leq p + k\text{SE}(\hat{p})$ for $\hat{p}$. Notice that we can re–express this relationship as $|\hat{p} - p| \leq k\text{SE}(\hat{p})$. Since $|\hat{p} - p|$, k, and $\text{SE}(\hat{p}) = \sqrt{p(1-p)/n}$ are positive, we can square each side of the inequality

$$|\hat{p} - p| \leq k\text{SE}(\hat{p})$$

to get the equivalent inequality

$$(\hat{p} - p)^2 \leq \frac{k^2}{n}(p - p^2).$$

Straightforward algebra allows us to re–express this inequality as the following quadratic inequality in p

$$(n + k^2)p^2 - 2(n\hat{p} + \frac{k^2}{2})p + n\hat{p}^2 \leq 0.$$

Treating this inequality as an equality and solving for p gives the two values

$$\tilde{p}_k \pm \text{ME}(\tilde{p}_k),$$

$$\text{where} \quad \tilde{p}_k = \frac{n\hat{p} + \frac{k^2}{2}}{n + k^2} \quad \text{and} \quad \text{ME}(\tilde{p}_k) = \sqrt{\tilde{p}_k^2 - \frac{n\hat{p}^2}{n + k^2}}.$$

Thus, letting C denote the desired confidence level, the probability statement

$$P\big[|\hat{p} - p| \leq k\text{SE}(\hat{p})\big] = C.$$

is equivalent to the probability statement

$$P\big[\tilde{p}_k - \text{ME}(\tilde{p}_k) \le p \le \tilde{p}_k + \text{ME}(\tilde{p}_k)\big] = C.$$

The endpoints of this interval, which are functions of $n, \hat{p}$, and k, are computable. Therefore, the Wilson confidence interval is given by

$$\tilde{p}_k - \text{ME}(\tilde{p}_k) \le p \le \tilde{p}_k + \text{ME}(\tilde{p}_k).$$

9.4 Testing a hypothesis about p

toc

In the apple orchard example we used a confidence interval estimate of p to decide whether the data supported the contention that more than 20% of the apple trees in the entire orchard were infested. We will now develop some formal methodology for assessing the evidence in favor of such a contention. We will start with some terminology.

A **hypothesis** (statistical hypothesis) is a conjecture about the nature of the population. When the population is dichotomous, a hypothesis is a conjecture about the value of the population success proportion p.

A **hypothesis test** (test of significance) is a formal procedure for deciding between two complementary hypotheses. These hypotheses are known as the null hypothesis (H_0 for short) and the research (or alternative) hypothesis (H_1 for short). The research hypothesis is the hypothesis of primary interest, since the testing procedure is designed to address the question: "Do the data support the research hypothesis?" The null hypothesis is defined as the negation of the research hypothesis. The test begins by tentatively assuming that the null hypothesis is true (the research hypothesis is false). The data are then examined to determine whether the null hypothesis can be rejected in favor of the research hypothesis. The probability of observing data as unusual (surprising) or more unusual as that actually observed under the tentative assumption that the null hypothesis is true is computed. This probability is known as the P–value of the test. (The P in P–value indicates that it is a probability it does not refer to the population success proportion p.) A small P–value indicates that the observed data would be unusual (surprising) if the null hypothesis was actually true. Thus if the P–value is small enough, then the null hypothesis is judged untenable and the test rejects the null hypothesis in favor of the research (alternative) hypothesis. On the other hand, a large (not small) P–value indicates that the observed data would not be unusual (not surprising) if the null hypothesis was actually true. Thus

if the P–value is large (not small enough), then the null hypothesis is judged tenable and the test fails to reject the null hypothesis.

There is a strong similarity between the reasoning used for a hypothesis test and the reasoning used in the trial of a defendant in a court of law. In a trial the defendant is presumed innocent (tentatively assumed to be innocent) and this tentative assumption is not rejected unless sufficient evidence is provided to make this tentative assumption untenable. In this situation the research hypothesis states that the defendant is guilty and the null hypothesis states that the defendant is not guilty (is innocent). The P–value of a hypothesis test is analogous to a quantification of the weight of the evidence that the defendant is guilty with small values indicating that the evidence is unlikely under the assumption that the defendant is innocent.

We will introduce hypothesis testing in the context of the apple orchard example. Details and formalization will follow the example.

Example 9.1 Insects in an apple orchard (revisited). Recall that the manager of a large apple orchard examined a simple random sample of 200 apple trees to gauge the extent of insect infestation in the orchard. The manager has determined that applying the insecticide is not economically justifiable unless more than 20% of the apple trees in the orchard are infested. Since the manager does not want to apply the insecticide unless there is evidence that it is needed, the question of interest here is: "Is there sufficient evidence to justify application of the insecticide?" In terms of the population success proportion p (the proportion of all of the apple trees in this orchard that are infested) **the research hypothesis** is $H_1 : p > .20$ (more than 20% of all the trees in the orchard are infested); and **the null hypothesis** is $H_0 : p \leq .20$ (no more than 20% of all the trees in the orchard are infested).

A test of the null hypothesis $H_0 : p \leq .20$ versus the research hypothesis $H_1 : p > .20$ begins by tentatively assuming that no more than 20% of all the trees in the orchard are infested. Under this tentative assumption it would be surprising to observe a proportion of infested trees in the sample, $\hat{p}$, that was much larger than .20. For example, when $n = 200$ and $p = .2$, the central 95% interval for $\hat{p}$ ranges from .1446 to.2554 and the central 99% interval for $\hat{p}$ ranges from .1272 to.2728. Therefore, if exactly 20% of all the trees in the orchard were infested, then it would be surprising to see much more than about

25% infested trees in the sample (above the central 95% interval) and it would be very surprising to see more than about 27% infested trees in the sample (above the central 99% interval). On the other hand, if exactly 30% of all the trees in the orchard were infested, then the central 95% interval for the sample percentage ranges from 23.65% to 36.35% and the central 99% interval for the sample percentage ranges from 21.65% to 38.35%, so that sample percentages around 25% to 28% would not be surprising. Thus the test should reject $H_0 : p \leq .20$ in favor of $H_1 : p > .20$ if the observed value of $\hat{p}$ is sufficiently large relative to .20. That is, if $\hat{p}$ is large enough to make us doubt our tentative assumption that p itself is not larger than .20.

Figure 9.12a SAS output for the apple orchard example – tests and confidence intervals. The commands to produce this output are provided in Figure 9.12b.

apple orchard example

The FREQ Procedure

case=5

outcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
infested	52	26.00	52	26.00
notinfested	148	74.00	200	100.00

Binomial Proportion	
outcome = infested	
Proportion	0.2600
ASE	0.0310

Confidence Limits for the Binomial Proportion		
Proportion = 0.2600		
Type	95% Confidence Limits	
Wilson	0.2041	0.3249

Test of H0: Proportion = 0.2	
ASE under H0	0.0283
Z	2.1213
One-sided Pr > Z	0.0169
Two-sided Pr > |Z|	0.0339

Sample Size = 200

apple orchard example

The FREQ Procedure

case=6

outcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
infested	45	22.50	45	22.50
notinfested	155	77.50	200	100.00

Binomial Proportion	
outcome = infested	
Proportion	0.2250
ASE	0.0295

Confidence Limits for the Binomial Proportion		
Proportion = 0.2250		
Type	95% Confidence Limits	
Wilson	0.1726	0.2877

Test of H0: Proportion = 0.2	
ASE under H0	0.0283
Z	0.8839
One-sided Pr > Z	0.1884
Two-sided Pr > |Z|	0.3768

Sample Size = 200

Case 1. Suppose that 52 of the 200 apple trees in the sample are infested so that $\hat{p} = .26$. In this case we know that 26% of the 200 trees in the sample are infested and we need to decide whether this suggests that the proportion of all the trees in the orchard that are infested, p, exceeds .20. More specifically, we need to determine whether observing 52 or more infested trees in a simple random sample of 200 trees would be surprising if in fact no more than 20% of all the trees in the orchard were infested. Assuming that

exactly 20% of all the trees in the orchard are infested, we find that the probability of observing 52 or more infested trees in a sample of 200 trees (seeing $\hat{p} \geq .26$), is .0169 (this is the P–value of the test). In the SAS output of Figure 9.12a (case 5) this P–value is labeled "one–sided Pr $> Z$". In other words, if no more than 20% of all the trees in the orchard were infested, then a simple random sample of 200 trees would give $\hat{p} \geq .26$ about 1.69% of the time. Therefore, observing 52 infested trees in a sample of 200 would be very surprising if no more than 20% of all the trees in the orchard were infested and we have sufficient evidence to reject the null hypothesis $H_0 : p \leq .20$ in favor of the research hypothesis $H_1 : p > .20$. In the case we would conclude that there is sufficient evidence to contend that more than 20% of all the trees in the orchard are infested and, in this sense, application of the insecticide is justifiable. Referring to the SAS output, we see that we can also conclude with 95% confidence that between 20.41% and 32.49% of all the trees in the entire orchard are infested. As expected, we find that the entire 95% confidence interval is above 20%. Note, however, that values as low as 20.41%, which is not much above 20%, are deemed plausible here.

The SAS commands which produced the output in Figure 9.12a are provided in Figure 9.12b.

Case 2. Next suppose that 45 of the 200 apple trees in the sample are infested so that $\hat{p} = .225$. Assuming that exactly 20% of all the trees in the orchard are infested, we find that the probability of observing 45 or more infested trees in a sample of 200 trees (seeing $\hat{p} \geq .225$), is .1884 (this is the P–value of the test). In the SAS output of Figure 9.12a (case 6) this P–value is labeled "one–sided Pr $> Z$". In other words, if no more than 20% of all the trees in the orchard were infested, then a simple random sample of 200 trees would give $\hat{p} \geq .225$ about 18.84% of the time. Therefore, observing 45 infested trees in a sample of 200 would not be very surprising if no more than 20% of all the trees in the orchard were infested and we do not have sufficient evidence to reject the null hypothesis $H_0 : p \leq .20$ in favor of the research hypothesis $H_1 : p > .20$. In the case we would conclude that there is not sufficient evidence to contend that more than 20% of all the trees in the orchard are infested and, in this sense, application of the insecticide not is justifiable. Referring to the SAS output, we see that we can also conclude with 95% confidence that between 17.26% and 28.77% of all the trees in the entire orchard are infested. As expected, we see that the 95% confidence interval contains values which are both above and below 20%.

Figure 9.12b SAS commands for the apple orchard example – tests and confidence intervals

SAS command file: Apple orchard examples cases 5 and 6

```
/* apple orchard example */
data orchard;
   input case outcome : $ 11. count;
/* the coding outcome : $ 11. indicates that we want
                      to allocate 11 characters for outcome */
cards;
5 infested 52
5 notinfested 148
6 infested 45
6 notinfested 155
;
proc freq data=orchard order=data;
   tables outcome / alpha=.05 binomial(cl=wilson p=.2 level='infested');
   weight count;
   by case;
title 'apple orchard example';
/* binomial options
   -- cl=wilson Wilson confidence interval
   -- p=.2 use p_0=.2 in the hypothesis
   -- level='infested' define "success" default is
         the first level which is infested in this example */
run;
```

The research hypothesis in the apple orchard example is a **directional hypothesis** of the form $H_1 : p > p_0$, where $p_0 = .20$. We will now discuss the details of a hypothesis test for a directional research hypothesis of this form. For the test procedure to be valid the specified value p_0 and the direction of the research hypothesis must be motivated from subject matter knowledge before looking at the data that are to be used to perform the test.

Testing a directional hypothesis of the form $p > p_0$

Research question. Is there sufficient evidence to conclude that the population proportion p is greater than the hypothesized value p_0?

Research hypothesis. $H_1 : p > p_0$, The population proportion p is greater than the hypothesized value p_0.

Tentative assumption – null hypothesis. $H_0 : p \leq p_0$, We tentatively assume that the population proportion p is not greater than the hypothesized value p_0.

Evidence in favor of the research hypothesis. The relationship between the observed proportion of successes in the sample $\hat{p}$ and the hypothesized value p_0 will be used to assess the strength of the evidence in favor of the research hypothesis. Generally, we would expect to observe larger values of $\hat{p}$ more often when the research hypothesis $H_1 : p > p_0$ is true than when the null hypothesis $H_0 : p \leq p_0$ is true. In particular, we can view the observation of a value of $\hat{p}$ that is sufficiently large relative to p_0 as constituting evidence against the null hypothesis $H_0 : p \leq p_0$ and in favor of the research hypothesis $H_1 : p > p_0$.

Assessment of the strength of the evidence – the P–value of the test. Deciding whether the observed value of $\hat{p}$ is "sufficiently large relative to p_0" is based on the P–value of the test. The P–value for testing the null hypothesis $H_0 : p \leq p_0$ versus the research hypothesis $H_1 : p > p_0$ is the probability of observing a value of $\hat{p}$ as large or larger than the value of $\hat{p}$ that we actually do observe. The P–value quantifies the consistency of the observed data with the null hypothesis and may be interpreted as a, somewhat indirect, measure of the strength of the evidence in the data in favor of the research hypothesis and against the null hypothesis. Because the P–value is computed under the assumption that the null hypothesis is true (and the research hypothesis is false), the smaller the P–value is, the less consistent the observed data are with the null hypothesis. Therefore, since one of the hypotheses must be true, when we observe a small P–value we can conclude that the research hypothesis is more consistent with the observed data than is the null hypothesis.

Computation of the P–value. The P–value of the test is computed under the assumption that the research hypothesis $H_1 : p > p_0$ is false and the null hypothesis $H_0 : p \leq p_0$ is true. Because the null hypothesis only specifies that $p \leq p_0$, we need to choose a particular value of p (that is no larger than p_0) in order to compute the P–value. It is most appropriate to use $p = p_0$ for this computation. (Recall that in the apple orchard example we used $p_0 = .20$ to compute the P–value.) Using $p = p_0$, which defines the boundary between $p \leq p_0$, where the null hypothesis is true, and $p > p_0$, where the research hypothesis is true, provides some protection against incorrectly rejecting $H_0 : p \leq p_0$.

The derivation below is meant to clarify the procedure. We can use a suitable calculator or computer program to perform these computations. We will use the normal approximation to the sampling distribution of $\hat{p}$ to compute the P–value. As noted above we will use the

hypothesized value p_0 in our computation of the P–value. Thus we will use the population standard deviation of $\hat{p}$

$$\text{SE}(\hat{p}) = \sqrt{\frac{p_0(1-p_0)}{n}}$$

in our computation of the Z–score. The calculated Z statistic or Z score corresponding to the observed value of $\hat{p}$, denoted by Z_{calc}, is

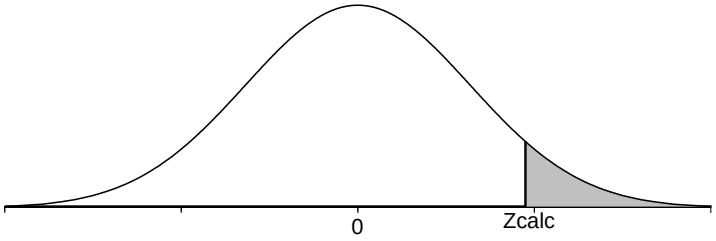
$$Z_{calc} = \frac{\hat{p} - p_0}{\text{SE}(\hat{p})} = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}.$$

Recall that the P–value for testing the null hypothesis $H_0 : p \leq p_0$ versus the research hypothesis $H_1 : p > p_0$ is the probability of observing a value of $\hat{p}$ as large or larger than the value of $\hat{p}$ that we actually do observe, computed assuming that $p = p_0$. Using the normal approximation, this P–value is equal to the probability that a standard normal variable takes on a value at least as large as Z_{calc}. This P–value is

$$P\text{–value} = P(Z \geq Z_{calc}),$$

where Z denotes a standard normal variable, *i.e.*, this P–value is the area under the standard normal density curve to the right of Z_{calc}, as shown in Figure 9.13. Notice that the P value (the area to the right of Z_{calc}) is small when Z_{calc} is far to the right of zero which is equivalent to $\hat{p}$ being far to the right of p_0.

Figure 9.13 P–value for $H_0 : p \leq p_0$ versus $H_1 : p > p_0$.



Once the P–value has been computed we need to decide whether the P–value is small enough to justify rejecting the null hypothesis in favor of the research hypothesis. In the apple orchard example we argued that observing 52 infested trees in a sample of 200 would be very surprising if no more than 20% of all the trees in the orchard were infested, since the corresponding P–value of .0169 was very small. We also argued that observing 45

infested trees in a sample of 200 would not be very surprising if no more than 20% of all the trees in the orchard were infested, since the corresponding P–value of .1884 is fairly large. Deciding whether a P–value is small enough to reject a null hypothesis requires a subjective judgment by the investigator in the context of the problem at hand. Some guidelines for interpreting a P–value are provided below.

Returning to our discussion for the directional research hypothesis $H_1 : p > p_0$. The final steps for performing a hypothesis test for

$$H_0 : p \leq p_0 \quad \text{versus} \quad H_1 : p > p_0$$

are summarized below.

1. Use a suitable calculator or computer program to find the P–value $= P(Z \geq Z_{calc})$, where Z denotes a standard normal variable, $Z_{calc} = (\hat{p} - p_0)/\text{SE}(\hat{p})$, and $\text{SE}(\hat{p}) = \sqrt{p_0(1-p_0)/n}$. This P–value is the area under the standard normal density curve to the right of Z_{calc}, as shown in Figure 9.13. This P–value is the "one–sided Pr > Z" of the SAS output.

2a. If the P–value is small enough (less than .05 for a test at the 5% level of significance), conclude that the data favor $H_1 : p > p_0$ over $H_0 : p \leq p_0$. That is, if the P–value is small enough, then there is sufficient evidence to conclude that the population success proportion p is greater than p_0.

2b. If the P–value is not small enough (is not less than .05 for a test at the 5% level of significance), conclude that the data do not favor $H_1 : p > p_0$ over $H_0 : p \leq p_0$. That is, if the P–value is not small enough, then there is not sufficient evidence to conclude that the population success proportion p is greater than p_0.

Remarks and guidelines about P–values.

The following general remarks regarding the use of P–values to assess the evidence against a null hypothesis and in favor of a research hypothesis apply to hypothesis tests in general, not just hypothesis tests for a proportion.

One approach to hypothesis testing is to use a fixed cutoff value to decide whether the P–value is "large" or "small". The most common application of this approach is to conclude that there is sufficient evidence to reject the null hypothesis in favor of the research

hypothesis only when the P–value is less than .05. When a fixed cutoff value like .05 (5%) is used to decide whether to reject the null hypothesis in favor of the research hypothesis this cutoff value is known as the **significance level** of the test. Hence, if we adopt the rule of rejecting the null hypothesis in favor of the research hypothesis only when the P–value is less than .05, then we are performing a hypothesis test at the 5% level of significance. In accordance with this terminology, the P–value is also known as the **observed significance level** of the test and if the P–value is less than the prescribed significance level, then the results are said to be **statistically significant**.

To perform a hypothesis test at the 5% level of significance we compute the appropriate P–value and compare it to the fixed significance level .05. If the P–value is less than .05, then we conclude that there is sufficient evidence, at the 5% level of significance, to reject the null hypothesis H_0 in favor of the research hypothesis H_1, *i.e.*, if the P–value **is less than** .05, then the data **do** support H_1. If the P–value is not less than .05, then we conclude that there is not sufficient evidence, at the 5% level of significance, to reject the null hypothesis H_0 in favor of the research hypothesis H_1, *i.e.*, if the P–value **is not less than** .05, then the data **do not** support H_1.

Instead of, or in addition to, using a fixed significance level like 5% we can use the P–value as a measure of the evidence (in the data) against the null hypothesis H_0 and in favor of the research hypothesis H_1. Some guidelines for deciding how strong the evidence is in favor of the research hypothesis H_1 are given below.

Guidelines for interpreting a P–value:

1. If the P–value is greater than .10, there is no evidence in favor of H_1.

2. If the P–value is between .05 and .10, there is suggestive but very weak evidence in favor of H_1.

3. If the P–value is between .04 and .05, there is weak evidence in favor of H_1.

4. If the P–value is between .02 and .04, there is moderately strong evidence in favor of H_1.

5. If the P–value is between .01 and .02, there is strong evidence in favor of H_1.

6. If the P–value is less than .01, there is very strong evidence in favor of H_1.

Whether you choose to use a fixed significance level or the preceding guidelines based on the P–value you should always report the P–value since this allows someone else to interpret the evidence in favor of H_1 using their personal preferences regarding the size of a P–value.

In the U.S. legal system there is a similar set of guidelines for assessing the level of proof or weight of the evidence against the null hypothesis of innocence and in favor of the research hypothesis of guilt. The weakest level of proof is "the preponderance of the evidence" (this is similar to a reasonably small P–value), the next level of proof is "clear and convincing evidence" (this is similar to a small P–value), and the highest level of proof is "beyond a reasonable doubt" (this is similar to a very small P–value).

Example 9.2 Acceptance sampling for electronic devices. A large retailer receives a shipment of 10,000 electronic devices from a supplier. The supplier guarantees that no more than 6% of these devices are defective. In fact, if more than 6% of the devices in the shipment are defective, then the supplier will allow the retailer to return the entire shipment, provided this is done with 10 days of receiving the shipment. Therefore, the retailer needs to decide between accepting the shipment and returning the shipment to the supplier. This decision will be based on the information provided by examining a simple random sample of electronic devices selected from the shipment.

In this example one of these electronic devices is a unit and the collection of 10,000 units constituting the shipment is the population. Notice that, in this example, the target population and the sampled population are the same (each is the shipment of 10,000 devices). A suitable variable for the indicated objective is whether an electronic device is defective with the two possible values: yes (it is defective) and no (it is not defective). A relevant parameter is the proportion p of defective devices in the shipment of 10,000 devices. The corresponding statistic $\hat{p}$ is the proportion of defective devices in the sample of devices that is examined.

The boundary between the null and research hypotheses is clearly $p_0 = .06$, since we need to decide whether the population proportion of defective devices p exceeds .06. Assuming that the supplier is trustworthy, it would seem to be a reasonable business practice to accept the shipment of electronic devices unless we find sufficient evidence, by examining the sample of devices, to conclude that more than 6% of the devices in the shipment

are defective. Hence, we will use a hypothesis test to determine whether there is sufficient evidence to conclude that the population defective proportion p exceeds .06. More formally, our research hypothesis is $H_1 : p > .06$ and our null hypothesis is $H_0 : p \leq .06$.

To continue with this example we need to know the sample size n and the results of the examination of the sample of electronic devices. Suppose that the simple random sample contains $n = 200$ electronic devices. For a sample of size $n = 200$ the standard error of $\hat{p}$ for testing a hypothesis with $p_0 = .06$ is

$$\text{SE}(\hat{p}) = \sqrt{\frac{(.06)(.94)}{200}} = .0168.$$

Figure 9.14 SAS output for the acceptance sampling example

acceptance sampling example

The FREQ Procedure

case=1

outcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
defective	16	8.00	16	8.00
notdefective	184	92.00	200	100.00

Binomial Proportion	
outcome = defective	
Proportion	0.0800
ASE	0.0192

Confidence Limits for the Binomial Proportion		
Proportion = 0.0800		
Type	95% Confidence Limits	
Wilson	0.0498	0.1260

Test of H0: Proportion = 0.06	
ASE under H0	0.0168
Z	1.1910
One-sided Pr > Z	0.1168
Two-sided Pr > |Z|	0.2337

Sample Size = 200

acceptance sampling example

The FREQ Procedure

case=2

outcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
defective	20	10.00	20	10.00
notdefective	180	90.00	200	100.00

Binomial Proportion	
outcome = defective	
Proportion	0.1000
ASE	0.0212

Confidence Limits for the Binomial Proportion		
Proportion = 0.1000		
Type	95% Confidence Limits	
Wilson	0.0657	0.1494

Test of H0: Proportion = 0.06	
ASE under H0	0.0168
Z	2.3820
One-sided Pr > Z	0.0086
Two-sided Pr > |Z|	0.0172

Sample Size = 200

Case 1. Suppose that 16 of the 200 devices in the sample are defective so that $\hat{p} = .08$. In this case we know that 8% of the 200 devices in the sample are defective and we need to decide whether this suggests that more than 6% of all the devices in the shipment are defective. The calculated Z statistic is

$$Z_{calc} = \frac{\hat{p} - p_0}{\text{SE}(\hat{p})} = \frac{.08 - .06}{.0168} = 1.1910$$

and the P–value is

$$P\text{–value} = P(Z \geq Z_{calc}) = P(Z \geq 1.1910) = .1168.$$

In the SAS output of Figure 9.14 this P–value is labeled "one–sided Pr $> Z$. Since this P–value is large there is not sufficient evidence to reject the null hypothesis $p \leq .06$ in favor of the research hypothesis $p > .06$. Therefore, if we observe 16 defective devices in a random sample of $n = 200$ devices, then we should accept the shipment of devices, since there is not sufficient evidence to conclude that more than 6% of the shipment of 10,000 devices is defective.

Case 2. Now suppose that 20 of the 200 devices in the sample are defective so that $\hat{p} = .10$. In this case

$$Z_{calc} = \frac{\hat{p} - p_0}{\text{SE}(\hat{p})} = \frac{.10 - .06}{.0168} = 2.3820$$

and the P–value is

$$P\text{–value} = P(Z \geq Z_{calc}) = P(Z \geq 2.3820) = .0086.$$

In the SAS output of Figure 9.14 this P–value is labeled "one–sided Pr $> Z$. This P–value is very small indicating that we have strong evidence against the null hypothesis $p \leq .06$ and in favor of the research hypothesis $p > .06$. Therefore, if we observe 20 defective devices in a random sample of $n = 200$ devices, then we are justified in returning the shipment of devices, since there is strong evidence that more than 6% of the shipment of 10,000 devices is defective.

In both of the cases described above, in addition to the conclusion of the hypothesis test the retailer might also wonder exactly what proportion of devices in the shipment of 10,000 devices are defective. We can use a 95% confidence interval estimate of p to answer this question.

In the first case there are 16 defective devices in the sample of $n = 200$ giving an observed proportion of defective devices of $\hat{p} = .08$. From the SAS output in Figure 9.14, we are 95% confident that the actual proportion of defective devices in the shipment of 10,000 is between .0498 and .1260. As expected, since we did not reject the tentative assumption that $p \leq .06$, we see that this confidence interval includes proportions that are both less than .06 and greater than .06.

In the second case there are 20 defective devices in the sample of $n = 200$ giving $\hat{p} = .10$. From the SAS output in Figure 9.14, we are 95% confident that the actual proportion of defective devices in the shipment of 10,000 is between .0656 and .1494. As expected, since we did reject the tentative assumption that $p \leq .06$, we see that all of the values in this confidence interval are greater than .06. Notice that in this case the P–value .0086 is quite small indicating that there is very strong evidence that the proportion of defective devices in the shipment is larger than .06. However, from the 95% confidence interval estimate of p we find that this proportion of defective devices might actually be as small as .0656, which is not much larger than .06. Thus, the small P–value indicates strong evidence that p is greater than .06 but it does not necessarily indicate that p is a lot larger than .06. Of course the 95% confidence interval estimate also indicates that p may be as large as .1494 which is a good bit larger than .06.

The scenario in the acceptance sampling example where there is strong evidence that $p > .06$ (P–value .0086) but the lower limit of the 95% confidence interval .0656 is not much larger than .06 highlights the need for a confidence interval to estimate the value of p in addition to a hypothesis test to clarify the practical importance of the result of the test. Bear in mind that a hypothesis test addresses a very formal distinction between two complementary hypotheses and that in some situations the results may be statistically significant (in the sense that the P–value is small) but of little practical significance (in the sense that p is not very different from p_0).

Testing a directional hypothesis of the form p < p_0

The procedure for testing the null hypothesis $H_0 : p \leq p_0$ versus the research hypothesis $H_1 : p > p_0$ given above is readily modified for testing the null hypothesis $H_0 : p \geq p_0$ versus the research hypothesis $H_1 : p < p_0$. The essential modification is to change the direction of the inequality in the definition of the P–value. Consider a situation where the research hypothesis specifies that the population success proportion p is less than the particular, hypothesized value p_0, *i.e.*, consider a situation where the research hypothesis is $H_1 : p < p_0$ and the null hypothesis is $H_0 : p \geq p_0$. For these hypotheses values of the observed success proportion $\hat{p}$ that are sufficiently small relative to p_0 provide evidence in favor of the research hypothesis $H_1 : p < p_0$ and against the null hypothesis $H_0 : p \geq p_0$. Therefore, the P–value for testing $H_0 : p \geq p_0$ versus $H_1 : p < p_0$ is the probability of observing a value of $\hat{p}$ as small or smaller than the value actually observed. As before, the

P–value is computed under the assumption that $p = p_0$. The calculated Z statistic Z_{calc} is defined as before; however, in this situation the P–value is the area under the standard normal density curve to the left of Z_{calc}, since values of $\hat{p}$ that are small relative to p_0 constitute evidence in favor of the research hypothesis.

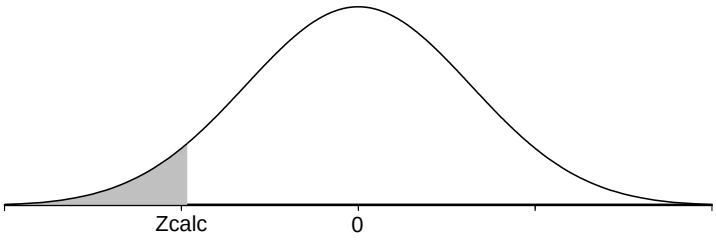
The steps for performing a hypothesis test for

$$H_0 : p \geq p_0 \quad \text{versus} \quad H_1 : p < p_0$$

are summarized below.

1. Use a suitable calculator or computer program to find the P–value $= P(Z \leq Z_{calc})$, where Z denotes a standard normal variable, $Z_{calc} = (\hat{p} - p_0)/\text{SE}(\hat{p})$, and $\text{SE}(\hat{p}) = \sqrt{p_0(1-p_0)/n}$. This P–value is the area under the standard normal density curve to the left of Z_{calc} as shown in Figure 9.15. This P–value is the "one–sided Pr < Z" of the SAS output.

Figure 9.15 P–value for $H_0 : p \geq p_0$ versus $H_1 : p < p_0$.



2a. If the P–value is small enough (less than .05 for a test at the 5% level of significance), conclude that the data favor $H_1 : p < p_0$ over $H_0 : p \geq p_0$. That is, if the P–value is small enough, then there is sufficient evidence to conclude that the population success proportion p is less than p_0.

2b. If the P–value is not small enough (is not less than .05 for a test at the 5% level of significance), conclude that the data do not favor $H_1 : p < p_0$ over $H_0 : p \geq p_0$. That is, if the P–value is not small enough, then there is not sufficient evidence to conclude that the population success proportion p is less than p_0.

Bernoulli trials

Some applications of these inferential methods for a proportion, such as the following machine parts example, correspond to a sequence of n trials. In this context a dichotomous trial is a process of observation or experimentation which results in one of two distinct outcomes (success or failure).

A sequence of n trials is said to constitute a sequence of **n Bernoulli trials with success probability p** if the following conditions are satisfied.

1. There is a common probability of success p for every trial. That is, on every trial the probability that the outcome of the trial will be a success is p.

2. The outcomes of the trials are independent of each other. That is, if we knew the outcome of a particular trial or trials this would provide no additional information about the probability of observing a success (or failure) on any other trial. For example, if we knew that a success (or failure) occurred in the first trial, this would not change the probability of success in any other trial.

The simple example below will help to clarify the definition of a sequence of n Bernoulli trials and the connection between sampling from a dichotomous population and Bernoulli trials.

Example 9.3 Tossing a fair die. Let a trial consist of tossing a fair (balanced) die and observing the number of dots on the upturned face. Define a success to be the occurrence of a 1, 2, 3, or 4. Since the die is fair, the probability of a success on a single trial is $p = 4/6 = 2/3$. Furthermore, if the die is always tossed in the same fashion, then the outcomes of the trials are independent. Therefore, with success defined as above, tossing the fair die n times yields a sequence of n Bernoulli trials with success probability $p = 2/3$.

Note that this process of tossing a die is abstractly the same as the process of selecting a ball at random from a box containing six balls with the balls numbered from 1 to 6. Thus tossing the die n times is equivalent to selecting a simple random sample of size n with replacement from this box containing six balls.

Example 9.4 Machine parts. The current production process used to manufacture a particular machine part is known (from past experience) to produce parts which are unacceptable, in the sense that they require further machining, 35% of the time. A new production process has been developed with the hope that it will reduce the chance of producing unacceptable parts. Suppose that 200 parts are produced using the new production process and that 54 of these parts are found to be unacceptable.

In this example we have a sequence of 200 dichotomous trials, where a trial consists of producing a part with the new production process and determining whether it is unacceptable. In this example p denotes the probability that a part produced using the new production process will be unacceptable. We will model these 200 trials as a sequence of $n = 200$ Bernoulli trials with population success probability p. This assumption is reasonable provided: (1) the probability that a part is unacceptable is essentially constant from part to part; and, (2) whether a specific part is unacceptable or not has no effect on the probability that any other part is unacceptable.

In this example the boundary between the null and research hypotheses is clearly $p_0 = .35$. Since these data were collected to determine if the new production process is better than the old process, we want to know whether there is sufficient evidence to conclude that less than 35% of the parts produced using the new production process would be unacceptable. Thus our research hypothesis is $H_1 : p < .35$ and our null hypothesis is $H_0 : p \geq .35$. Since 54 of the 200 parts in our sample are unacceptable we know that $\hat{p} = .27$ and we need to determine whether this is small enough to suggest that the corresponding population probability p is also less than .35. For a sample of size $n = 200$ the standard error of $\hat{p}$ for testing a hypothesis with $p_0 = .35$ is

$$\text{SE}(\hat{p}) = \sqrt{\frac{(.35)(.65)}{200}} = .0337.$$

The calculated Z statistic is

$$Z_{calc} = \frac{\hat{p} - p_0}{\text{SE}(\hat{p})} = \frac{.27 - .35}{.0337} = -2.3739$$

and the P–value is

$$P\text{–value} = P(Z \leq Z_{calc}) = P(Z \leq -2.3739) = .0088.$$

You can find these values in the SAS output of Figure 9.16. Since this P–value is very small, there is sufficient evidence to reject the null hypothesis $p \geq .35$ in favor of the research hypothesis $p < .35$. Hence, based on this sample of 200 parts there is very strong evidence that the new production process is superior in the sense that the probability of producing an unacceptable part is less than .35.

Clearly this conclusion should be accompanied by an estimate of how much smaller this probability is likely to be. Observing 54 unacceptable parts in the sample of $n = 200$ gives $\hat{p} = .27$ and a 95% confidence interval ranging from .2132 to .3354. Therefore, we are 95% confident that the probability of a part produced using the new production process being unacceptable is between .2132 and .3354. As expected, since we did reject the tentative assumption that $p \geq .35$, we see that all of the values in this confidence interval are less than .35. The P–value .0088 is quite small indicating that there is very strong evidence that the probability of producing an unacceptable part is less than .35. However, from the 95% confidence interval estimate of p we find that this probability might actually be as large as .3354 which is not much smaller than .35. Of course the 95% confidence interval estimate also indicates that p may be as small as .2132 which is a good bit smaller than .35.

Figure 9.16 SAS output for the machine parts example

machine parts example

The FREQ Procedure

case=1

outcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
unacceptable	54	27.00	54	27.00
acceptable	146	73.00	200	100.00

Binomial Proportion	
outcome = unacceptable	
Proportion	0.2700
ASE	0.0314

Confidence Limits for the Binomial Proportion		
Proportion = 0.2700		
Type	95% Confidence Limits	
Wilson	0.2132	0.3354

Test of H0: Proportion = 0.35	
ASE under H0	0.0337
Z	-2.3720
One-sided Pr < Z	0.0088
Two-sided Pr > |Z|	0.0177

Sample Size = 200

Testing a nondirectional research hypothesis

The hypothesis tests we have discussed thus far are only appropriate when we have enough *a priori* information, *i.e.*, information that does not depend on the data to be used for the hypothesis test, to postulate that the population success proportion p is on one side of a particular value p_0. That is, we have only considered situations where the research hypothesis is directional in the sense of specifying either that $p > p_0$ or that $p < p_0$. In some situations we will not have enough *a priori* information to allow us to choose the appropriate directional research hypothesis. Instead, we might only conjecture that the population success proportion p is different from some particular value p_0. In a situation like this our research hypothesis specifies that the population success proportion p is different from p_0, *i.e.*, $H_1 : p \neq p_0$ and the corresponding null hypothesis specifies that p is exactly equal to p_0, *i.e.*, $H_0 : p = p_0$. As we will see in the inheritance model considered below, when testing to see whether p is equal to a specified value p_0 the null hypothesis $H_0 : p = p_0$ often corresponds to the validity of a particular theory or model and the research hypothesis or alternative hypothesis specifies that the theory is invalid.

Testing a nondirectional research (alternative) hypothesis of the form $p \neq p_0$.

Research question. Is there sufficient evidence to conclude that the population proportion p is different from the hypothesized value p_0?

Research hypothesis. $H_1 : p \neq p_0$, The population proportion p is not equal to the hypothesized value p_0.

Tentative assumption – null hypothesis. $H_0 : p = p_0$, We tentatively assume that the population proportion p is exactly equal to the hypothesized value p_0.

Evidence in favor of the research hypothesis. As with the directional hypothesis cases, the relationship between the observed proportion of successes in the sample $\hat{p}$ and the hypothesized value p_0 will be used to assess the strength of the evidence in favor of the research hypothesis. Generally, we would expect to observe values of $\hat{p}$ farther away from p_0 more often when the research hypothesis $H_1 : p \neq p_0$ is true than when the null hypothesis $H_0 : p = p_0$ is true. In particular, we can view the observation of a value of $\hat{p}$ that is sufficiently far away from p_0, in either direction, as constituting evidence against the null hypothesis $H_0 : p = p_0$ and in favor of the research hypothesis $H_1 : p \neq p_0$.

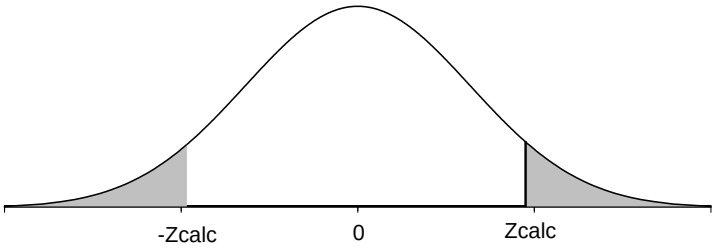
Assessment of the strength of the evidence – the P–value of the test. Deciding whether the observed value of $\hat{p}$ is "sufficiently far away from p_0 in either direction" is based on the P–value of the test. The P–value for testing the null hypothesis $H_0 : p = p_0$ versus the research hypothesis $H_1 : p \neq p_0$ is the probability of observing a value of $\hat{p}$ for which the distance $|\hat{p} - p_0|$ (the absolute value of the difference between $\hat{p}$ and p_0) is as large or larger than the value of this distance that we actually do observe.

Computation of the P–value. The P–value of the test is computed under the assumption that the research hypothesis $H_1 : p \neq p_0$ is false and the null hypothesis $H_0 : p = p_0$ is true. In this situation the calculated Z statistic Z_{calc} is the absolute value of the Z statistic that would be used for testing a directional hypothesis. That is, the calculated Z statistic is

$$Z_{calc} = \left| \frac{\hat{p} - p_0}{\text{SE}(\hat{p})} \right| = \left| \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} \right|.$$

In terms of this Z statistic the P–value is the probability that the absolute value of a standard normal variable Z would take on a value as large or larger than Z_{calc} assuming that $p = p_0$. This probability is the sum of the area under the standard normal density curve to the left of $-Z_{calc}$ and the area under the standard normal density curve to the right of Z_{calc}, as shown in Figure 9.17. We need to add these two areas (probabilities) since we are finding the probability that the observed success proportion $\hat{p}$ would be as far or farther away from p_0 in either direction as is the value that we actually observe, when $p = p_0$. Notice that this P value (the area to the left of $-Z_{calc}$ plus the area to the right of Z_calc) is small when Z_{calc} is far away from zero in one direction or the other which is equivalent to $\hat{p}$ being far away from p_0 in one direction or the other.

Figure 9.17 P–value for $H_0 : p = p_0$ versus $H_1 : p \neq p_0$.



The steps for performing a hypothesis test for

$$H_0 : p = p_0 \quad \text{versus} \quad H_1 : p \neq p_0$$

are summarized below.

1. Use a suitable calculator or computer program to find the P–value $= P(|Z| \geq Z_{calc}) = P(Z \leq -Z_{calc}) + P(Z \geq Z_{calc})$, where Z denotes a standard normal variable, $Z_{calc} = |(\hat{p} - p_0)/\text{SE}(\hat{p})|$, and $\text{SE}(\hat{p}) = \sqrt{p_0(1-p_0)/n}$. This P–value is the sum of the area under the standard normal density curve to the left of $-Z_{calc}$ and the area under the standard normal density curve to the right of Z_{calc} as shown in Figure 9.17. This P–value is the "two–sided Pr $> |Z|$" of the SAS output.

2a. If the P–value is small enough (less than .05 for a test at the 5% level of significance), conclude that the data favor $H_1 : p \neq p_0$ over $H_0 : p = p_0$. That is, if the P–value is small enough, then there is sufficient evidence to conclude that the population success proportion p is different from p_0.

2b. If the P–value is not small enough (is not less than .05 for a test at the 5% level of significance), conclude that the data do not favor $H_1 : p \neq p_0$ over $H_0 : p = p_0$. That is, if the P–value is not small enough, then there is not sufficient evidence to conclude that the population success proportion p is different from p_0.

Example 9.5 Inheritance in peas (flower color). In his investigations, during the years 1856 to 1868, of the chromosomal theory of inheritance Gregor Mendel performed a series of experiments on ordinary garden peas. One characteristic of garden peas that Mendel studied was the color of the flowers (red or white). When Mendel crossed a plant with red flowers with a plant with white flowers, the resulting offspring all had red flowers. But when he crossed two of these first generation plants, he observed plants with white as well as red flowers. We will use the results of one of Mendel's experiments to test a simple model for inheritance of flower color. Mendel observed 929 pea plants arising from a cross of two of these first generation plants. Of these 929 plants he found 705 plants with red flowers and 224 plants with white flowers.

The gene which determines the color of the flower occurs in two forms (alleles). Let R denote the allele for red flowers (which is dominant) and r denote the allele for white flowers

(which is recessive). When two plants are crossed the offspring receives one allele from each parent, thus there are four possible genotypes (ordered combinations) RR, Rr, rR, and rr. The three genotypes RR, Rr, and rR, which include the dominant R allele, will yield red flowers while the fourth genotype rr will yield white flowers. If a red flowered RR genotype parent is crossed with a white flowered rr genotype parent, then all of the offspring will have genotype Rr and will produce red flowers. If two of these first generation Rr genotype plants are crossed, each of the four possible genotypes RR, Rr, rR, and rr is equally likely and plants with white as well as red flowers will occur. Under this simple model for inheritance, with each of the four genotypes having the same probability of occurring (and with each plant possessing only one genotype), the probability that a plant will have red flowers is $p = 3/4$ and the probability that a plant will have white flowers is $1 - p = 1/4$. In other words, this model for inheritance of flower color says that we would expect to see red flowers 3/4 of the time and white flowers 1/4 of the time.

We can test the validity of this model by testing the null hypothesis $H_0 : p = 3/4$ versus the alternative hypothesis $H_1 : p \neq 3/4$. Notice that the model is valid under the null hypothesis and the model is not valid under the alternative hypothesis. Mendel observed 705 plants with red flowers out of the $n = 929$ plants giving an observed proportion of plants with red flowers of $\hat{p} = 705/929 = .7589$. The standard error of $\hat{p}$, computed under the assumption that $p = p_0 = 3/4$, is

$$\text{SE}(\hat{p}) = \sqrt{\frac{(.75)(.25)}{929}} = .0142$$

and the calculated Z statistic is $Z_{calc} = .6251$ giving a P–value of

$$P\text{–value} = P(|Z| \geq Z_{calc}) = P(|Z| \geq .6251) = .5319.$$

You can find these values in the SAS output of Figure 9.18. This P–value is quite large and we are not able to reject the null hypothesis; therefore, we conclude that the observed data are consistent with Mendel's model. Technically, we should say that the data are not inconsistent with the model in the sense that we cannot reject the hypothesis that $p = 3/4$. In this example, the 95% confidence interval estimate of p ranges from .7303 to .7853.

Figure 9.18 SAS output for the Mendel pea flower color example

Mendel pea flower color example

The FREQ Procedure

case=1

outcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
red	705	75.89	705	75.89
white	224	24.11	929	100.00

Binomial Proportion	
outcome = red	
Proportion	0.7589
ASE	0.0140

Confidence Limits for the Binomial Proportion		
Proportion = 0.7589		
Type	95% Confidence Limits	
Wilson	0.7303	0.7853

Test of H0: Proportion = 0.75	
ASE under H0	0.0142
Z	0.6251
One-sided Pr > Z	0.2660
Two-sided Pr > |Z|	0.5319

Sample Size = 929

9.5 Directional confidence bounds toc

In our discussion of hypothesis testing we considered directional research hypotheses of the form $p > p_0$ and $p < p_0$ as well as nondirectional research hypotheses of the form $p \neq p_0$. However, in our discussion of 95% confidence intervals for p we only considered "nondirectional" confidence intervals of the form $p_L \leq p \leq p_U$. A 95% confidence interval of this form, consisting of a lower bound p_L for p and an upper bound p_U for p, gives a range of plausible values for p. In a situation where we have enough *a priori* information to justify a directional research hypothesis we might argue that it would be more appropriate to determine a 95% confidence bound (a lower bound or an upper bound) for p instead of a range of values.

A lower confidence bound on p allows us to estimate the smallest value of p which is plausible in light of the observed value of $\hat{p}$. Similarly, an upper confidence bound on p allows us to estimate the largest value of p which is plausible in light of the observed value of $\hat{p}$. For example, in the acceptance sampling example we might argue that we are less concerned with a limit on how large p might be than with a limit on how small it might be.

Therefore, we might be satisfied with an estimate of the smallest value of p which would be consistent with the data, *i.e.*, we might only need a 95% confidence lower bound for p.

The reasoning which led to the "nondirectional" Wilson interval (a range of values for p) can be adapted to yield a directional interval (a lower or upper confidence bound for p). Recall that our development of the Wilson 95% confidence interval estimate of p was based on the observation that, for each possible value of p, we can view the central 95% interval from $p - 1.96\text{SE}(\hat{p})$ to $p + 1.96\text{SE}(\hat{p})$ as an interval which is likely to contain $\hat{p}$. This starting point led us to a confidence interval estimate of p which provided a range of values between a lower limit and an upper limit. If we use an upper 95% interval for $\hat{p}$ or a lower 95% interval for $\hat{p}$ instead, we will end up with a directional confidence interval, *i.e.*, we will end up with a 95% confidence bound (lower or upper) for p. In practice, 95% confidence bounds are usually computed by selecting the appropriate endpoint of a 90% confidence interval. That is, if $p_L \leq p \leq p_U$ is a 90% confidence interval estimate of p, then P_L is a 95% lower confidence bound for p and p_U is a 95% upper confidence bound for p.

We will illustrate some applications of confidence bounds in the context of some of the examples we discussed earlier. Detailed derivations of 95% lower and upper confidence bounds are provided after the examples.

Example 9.2 Acceptance sampling for electronic devices (revisited). Recall that in this example the retailer had received a shipment of 10,000 electronic devices from a supplier with a guarantee that no more than 6% of these devices were defective. The retailer is interested in p, the value of the proportion of defective devices in the shipment of 10,000 devices. In particular, the retailer is concerned that this proportion might be too large (greater than .06). Thus we proposed a test of the null hypothesis $H_0 : p \leq .06$ versus the research hypothesis $H_1 : p > .06$. We considered two cases to illustrate the corresponding hypothesis testing procedure. We will now show how a directional confidence bound can supplement the formal hypothesis test.

As noted above, in this example, we are less concerned with a limit on how large p might be than with a limit on how small it might be. Therefore, we might be satisfied with an estimate of the smallest value of p which would be consistent with the data, *i.e.*, we might only need a 95% confidence lower bound for p. Note that a lower confidence bound will

always include values greater than .06 (6%). The important consideration is whether it also includes values less than .06.

Figure 9.19 SAS output for the acceptance sampling example

acceptance sampling example

The FREQ Procedure

case=1

outcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
defective	16	8.00	16	8.00
notdefective	184	92.00	200	100.00

Binomial Proportion	
outcome = defective	
Proportion	0.0800
ASE	0.0192

Confidence Limits for the Binomial Proportion		
Proportion = 0.0800		
Type	90% Confidence Limits	
Wilson	0.0538	0.1174

acceptance sampling example

The FREQ Procedure

case=2

outcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
defective	20	10.00	20	10.00
notdefective	180	90.00	200	100.00

Binomial Proportion	
outcome = defective	
Proportion	0.1000
ASE	0.0212

Confidence Limits for the Binomial Proportion		
Proportion = 0.1000		
Type	90% Confidence Limits	
Wilson	0.0703	0.1404

Case 1 If there are 16 defective devices in a sample of $n = 200$, then the observed proportion of defective devices is $\hat{p} = .08$. The lower endpoint .0538 of the 90% confidence interval in the SAS output for case 1 in Figure 9.19 is the 95% confidence lower bound for p. Hence, we can conclude that we are 95% confident that the actual percentage of defective devices in the shipment of 10,000 is at least 5.38%. More importantly, with 95% confidence we can say that the percentage could be as low as 5.38%. Since this lower bound is less than 6% we do not have sufficient evidence to claim that more than 6% of the 10,00 devices are defective.

Recall that, in this case, the P–value for testing $H_0 : p \leq .06$ versus $H_1 : p > .06$ was .1168 and we concluded that there is not sufficient evidence to claim that more than 6% of the 10,000 devices are defective. Thus the lower confidence bound and the formal hypothesis test lead to the same conclusion.

Case 2 If there are 20 defective devices in a sample of $n = 200$, then the observed proportion of defective devices is $\hat{p} = .10$. The lower endpoint .0703 of the 90% confidence

interval in the SAS output for case 2 in Figure 9.19 is the 95% confidence lower bound for p. In this case we can conclude that we are 95% confident that the actual percentage of defective devices in the shipment of 10,000 is at least 7.03%. That is, with 95% confidence we can say that the percentage is not less than 7.03%. Since this lower bound is greater than 6% we have sufficient evidence to claim that more than 6% of the 10,000 devices are defective. Note also that the lower bound 7.03% indicates that the percentage of defective devices among the 10,000 is at least 1.03 percentage points higher than the 6% cutoff value.

In this case, the P–value for testing $H_0 : p \leq .06$ versus $H_1 : p > .06$ was .0086 and we concluded that there is strong evidence that more than 6% of the 10,000 devices are defective. Again, the lower confidence bound and the test lead to the same conclusion.

Example 9.4 Machine parts (revisited). Recall that, in this example, the current production process used to manufacture a particular machine part is known (from past experience) to produce parts which are unacceptable, in the sense that they require further machining, 35% of the time. A new production process has been developed with the hope that it will reduce the chance of producing unacceptable parts. In this example p denotes the probability that a part produced using the new production process will be unacceptable and our goal is to decide whether this probability is less than .35. A sample of 200 parts was produced using the new production process and 54 of these parts were found to be unacceptable. The SAS output for this example is provided in Figure 9.20. In this example, the P–value for testing $H_0 : p \geq .35$ versus $H_1 : p < .35$ is .0088 and we concluded that there is very strong evidence that the new production process is superior in the sense that the probability of producing an unacceptable part is less than .35. Using the upper endpoint of the 90% confidence interval in the SAS output of Figure 9.20, we can conclude that we are 95% confident that the probability that a part produced using the new production process will be unacceptable is no larger than .3245. Since this value is less than .35, the 95% confidence upper bound leads to the conclusion that the new production process is superior to the old process and indicates the extent of the improvement.

Figure 9.20 SAS output for the machine parts example

machine parts example

The FREQ Procedure

case=1

outcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
unacceptable	54	27.00	54	27.00
acceptable	146	73.00	200	100.00

Binomial Proportion	
outcome = unacceptable	
Proportion	0.2700
ASE	0.0314

Confidence Limits for the Binomial Proportion		
Proportion = 0.2700		
Type	90% Confidence Limits	
Wilson	0.2217	0.3245

Test of H0: Proportion = 0.35	
ASE under H0	0.0337
Z	-2.3720
One-sided Pr < Z	0.0088
Two-sided Pr > |Z|	0.0177

For completeness, we will now provide detailed derivations of 95% lower and upper confidence bounds. The probability that a standard normal variable Z takes on a value less than 1.645 is equal to .95, $P(Z \leq 1.645) = .95$. That is, when we observe the value of a standard normal variable Z, 95% of the time we will find that $Z \leq 1.645$. Graphically this means that the area under the standard normal density curve over the interval from $-\infty$ to 1.645 is .95. Thus, for sufficiently large values of n we have the approximation,

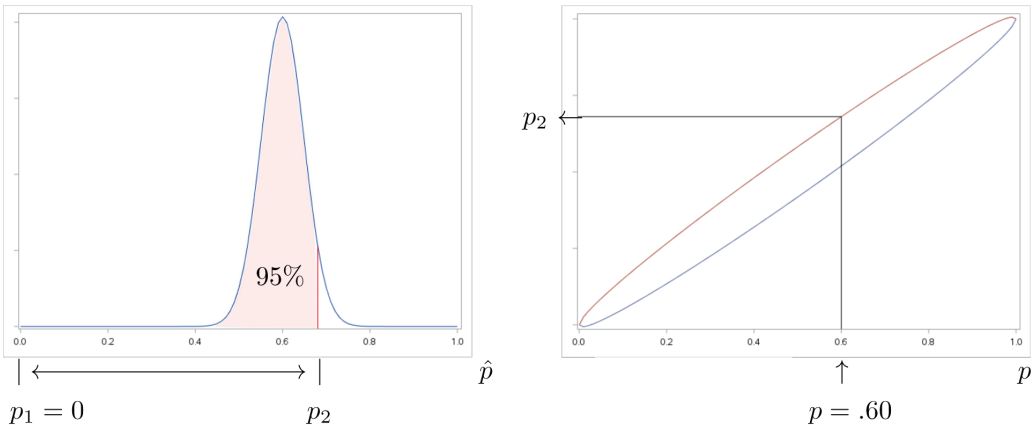
$$P\left[\frac{\hat{p} - p}{\mathrm{SE}(\hat{p})} \leq 1.645\right] = .95.$$

Note that this indicates that 95% of the time when a simple random sample is selected and $\hat{p}$ is computed the observed value of $\hat{p}$ will be between zero and $p + 1.645\mathrm{SE}(\hat{p})$, *i.e.*, $\hat{p}$ will be no more than 1.645 population standard error units above p. We will refer to the interval from zero to $p + 1.645\mathrm{SE}(\hat{p})$ as the lower 95% interval of the distribution of $\hat{p}$, since it will contain the observed value of $\hat{p}$ 95% of the time.

The plots in Figure 9.21 show how the lower 95% interval of the distribution of $\hat{p}$ depends on the value of p. On the left we have a representation of the lower 95% interval, the interval from $p_1 = 0$ to p_2. The plot on the right shows how p_2 depends on p. We want to determine the values of p which yield lower 95% intervals which contain $\hat{p}_{\mathrm{obs}}$. To do this

we need to use the graph on the right of Figure 9.21 in the other direction, as shown in Figure 9.22. In Figure 9.22 a horizontal line is drawn at $\hat{p}_{\text{obs}}$ and its intersection, p_L, with the upper curve is indicated. Notice that p_L is the smallest value of p for which the lower 95% interval of the distribution of $\hat{p}$ contains $\hat{p}_{\text{obs}}$. (Figure 9.22 is drawn for $n = 100$ and $\hat{p}_{\text{obs}} = .55$. In this case, $p_L = .4679$.) Thus, if we draw a vertical line, as in Figure 9.21, at any value of p between p_L and $p_U = 1$, then the corresponding lower 95% interval of the distribution of $\hat{p}$ contains $\hat{p}_{\text{obs}}$. Note that p_L is the 95% confidence lower bound for p.

Figure 9.21 The plot on the left shows the lower 95% interval of the distribution of $\hat{p}$ for $n = 100$ and $p = .6$.
The upper (red) curve in the plot on the right shows the upper endpoint, $p + 1.645\sqrt{p(1-p)/n}$, of the lower 95% interval of the distribution of $\hat{p}$ as a function of p for $n = 100$. The upper endpoint for the case $p = .6$ is indicated by the line marking the intersections at $p_2 = .6806$. (The lower endpoint is zero.)



In order to determine the value of p_L we simply set $p + 1.645\text{SE}(\hat{p})$ equal to $\hat{p}_{\text{obs}}$ and solve for p (draw the horizontal line as in Figure 9.22 and project down at the intersection with the upper (red) curve).

An analogous argument starting with the upper 95% interval of the distribution of $\hat{p}$, *i.e.*, the interval from $p - 1.645\text{SE}(\hat{p})$ to one, leads to a 95% confidence upper bound for p. In this case, as shown in Figure 9.23, we start with the upper 95% interval of the distribution of $\hat{p}$, the interval from $p_1 = p - 1.645\text{SE}(\hat{p})$ to one. The graph on the right in Figure 9.23 shows how the upper 95% interval of the distribution of $\hat{p}$ depends on value of p.

Figure 9.22 The smallest value of p, p_L, for which the lower 95% interval of the distribution of $\hat{p}$ contains $\hat{p}_{\text{obs}}$. (The interval goes from p_L to $p_U = 1$.)
This example is drawn for $n = 100$ and $\hat{p}_{\text{obs}} = .55$. Here $p_L = .4679$.

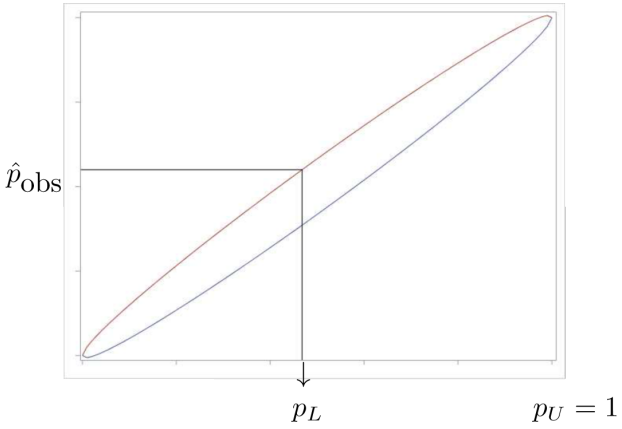
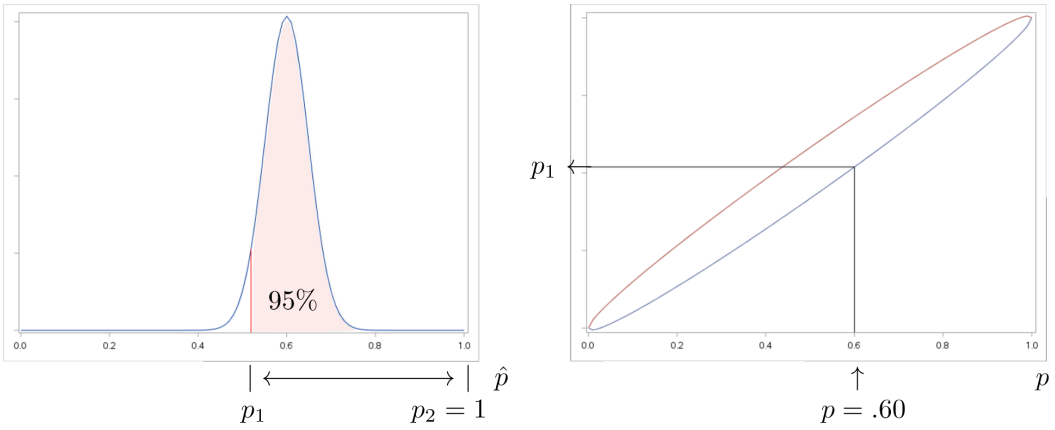


Figure 9.23 The plot on the left shows the upper 95% interval of the distribution of $\hat{p}$ for $n = 100$ and $p = .6$.
The lower (blue) curve in the plot on the right shows the lower endpoint, $p - 1.645\sqrt{p(1-p)/n}$, of the upper 95% interval of the distribution of $\hat{p}$ as a function of p for $n = 100$. The lower endpoint for the case $p = .6$ is indicated by the line marking the intersections at $p_1 = .6806$. (The upper endpoint is one.)

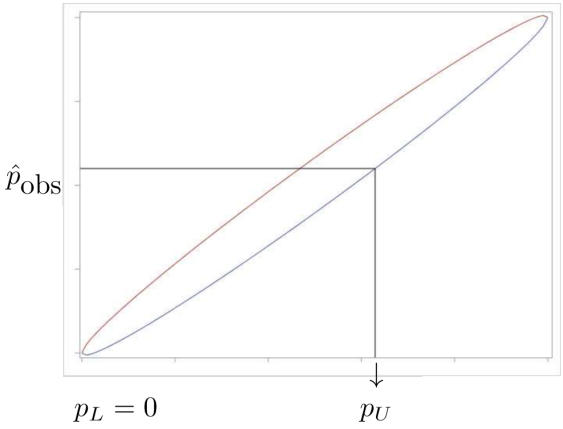


In Figure 9.24 a horizontal line is drawn at $\hat{p}_{\text{obs}}$ and its intersection, p_U, with the lower curve is indicated. Notice that p_U is the largest value of p for which the upper central 95% interval of the distribution of $\hat{p}$ contains $\hat{p}_{\text{obs}}$. (Figure 9.24 is drawn for $n = 100$ and $\hat{p}_{\text{obs}} = .55$. In this case, $p_U = .6806$.) Thus, if we draw a vertical line, as in Figure 9.23, at any value of p between $p_L = 0$ and p_U, then the corresponding upper 95% interval of

the distribution of $\hat{p}$ contains $\hat{p}_{\text{obs}}$. Note that p_U is the 95% confidence upper bound for p.

In order to determine the value of p_U we simply set $p - 1.645\text{SE}(\hat{p})$ equal to $\hat{p}_{\text{obs}}$ and solve for p (draw the horizontal line as in Figure 9.24 and project down at the intersection with the lower (blue) curve).

Figure 9.24 The largest value of p, p_U, for which the upper 95% interval of the distribution of $\hat{p}$ contains $\hat{p}_{\text{obs}}$. (The interval goes from $p_L = 0$ to p_U.)
This example is drawn for $n = 100$ and $\hat{p}_{\text{obs}} = .55$. Here $p_U = .6294$.



Index

toc