Chapter 1
Introduction

In this chapter we introduce basic ideas of probability and some related mathematical concepts that are used in statistics. Values to be analyzed statistically are generally thought of as random variables; these are numbers that result from random events. The mean (average) value of a population is defined in terms of the expected value of a random variable. The variance is introduced as a measure of the variability in a random variable (population). We also introduce some special distributions (populations) that are useful in modeling statistical data. The purpose of this chapter is to introduce these ideas, so they can be used in analyzing data and in discussing statistical models.

In writing statistical models, we often use symbols from the Greek alphabet. A table of these symbols is provided in Appendix B.6.

Rumor has it that there are some students studying statistics who have an aversion to mathematics. Such people might be wise to focus on the concepts of this chapter and not let themselves get bogged down in the details. The details are given to provide a more complete introduction for those students who are not math averse.

1.1 Probability

Probabilities are numbers between zero and one that are used to explain random phenomena. We are all familiar with simple probability models. Flip a standard coin; the probability of heads is 1/2. Roll a die; the probability of getting a three is 1/6. Select a card from a well-shuffled deck; the probability of getting the queen of spades is 1/52 (assuming there are no jokers). One way to view probability models that many people find intuitive is in terms of random sampling from a fixed population. For example, the 52 cards form a fixed population and picking a card from a well-shuffled deck is a means of randomly selecting one element of the population. While we will exploit this idea of sampling from fixed populations, we should also note its limitations. For example, blood pressure is a very useful medical indicator, but even with a fixed population of people it would be very difficult to define a useful population of blood pressures. Blood pressure depends on the time of day, recent diet, current emotional state, the technique of the person taking the reading, and

many other factors. Thinking about populations is very useful, but the concept can be very limiting both practically and mathematically. For measurements such as blood pressures and heights, there are difficulties in even specifying populations mathematically.

For mathematical reasons, probabilities are defined not on particular outcomes but on sets of outcomes (events). This is done so that continuous measurements can be dealt with. It seems much more natural to define probabilities on outcomes as we did in the previous paragraph, but consider some of the problems with doing that. For example, consider the problem of measuring the height of a corpse being kept in a morgue under controlled conditions. The only reason for getting morbid here is to have some hope of defining what the height is. Living people, to some extent, stretch and contract, so a height is a nebulous thing. But even given that someone has a fixed height, we can never know what it is. When someone's height is measured as 177.8 centimeters (5 feet 10 inches), their height is not really 177.8 centimeters, but (hopefully) somewhere between 177.75 and 177.85 centimeters. There is really no chance that anyone's height is *exactly* 177.8 cm, or exactly 177.8001 cm, or exactly 177.800000001 cm, or exactly 56.5955π cm, or exactly $(76\sqrt{5} + 4.5\sqrt{3})$ cm. In any neighborhood of 177.8, there are more numerical values than one could even imagine counting. The height should be somewhere in the neighborhood, but it won't be the particular value 177.8. The point is simply that trying to specify all the possible heights and their probabilities is a hopeless exercise. It simply cannot be done.

Even though individual heights cannot be measured exactly, when looking at a population of heights they follow certain patterns. There are not too many people over 8 feet (244 cm) tall. There are lots of males between 175.3 cm and 177.8 cm ($5'9''$ and $5'10''$). With continuous values, each possible outcome has no chance of occurring, but outcomes do occur and occur with regularity. If probabilities are defined for sets instead of outcomes, these regularities can be reproduced mathematically. Nonetheless, initially the best way to learn about probabilities is to think about outcomes and their probabilities.

There are five key facts about probabilities:

1. Probabilities are between 0 and 1.

2. Something that happens with probability 1 is a sure thing.

3. If something has no chance of occurring, it has probability 0.

4. If something occurs with probability, say, .25, the probability that it will not occur is $1 - .25 = .75$.

5. If two events are mutually exclusive, i.e., if they cannot possibly happen at the same time, then the probability that either of them occurs is just the sum of their individual probabilities.

Individual outcomes are always mutually exclusive, e.g., you cannot flip a coin and get both heads and tails, so probabilities for outcomes can always be added together. Just to be totally correct, I should mention one other point. It may sound silly, but we need to assume that *something* occurring is always a sure thing. If we flip a coin, we must get either heads or tails with probability 1. We could even allow for the coin landing on its edge as long as the probabilities for all the outcomes add up to 1.

Example 1.1.1. Consider the nine outcomes that are all combinations of three heights, tall (T), medium (M), short (S) and three eye colors, blue (Bl), brown (Br) and green (G). The combinations are displayed below.

Height–eye color combinations

		Eye color		
		Blue	Brown	Green
Height	Tall	T, Bl	T, Br	T, G
	Medium	M, Bl	M, Br	M, G
	Short	S, Bl	S, Br	S, G

The set of all outcomes is

$$\{(\mathrm{T},\mathrm{Bl}),(\mathrm{T},\mathrm{Br}),(\mathrm{T},\mathrm{G}),(\mathrm{M},\mathrm{Bl}),(\mathrm{M},\mathrm{Br}),(\mathrm{M},\mathrm{G}),(\mathrm{S},\mathrm{Bl}),(\mathrm{S},\mathrm{Br}),(\mathrm{S},\mathrm{G})\}\,.$$

The event that someone is tall consists of the three pairs in the first row of the table, i.e.,

$$\{\mathrm{T}\} = \{(\mathrm{T},\mathrm{Bl}),(\mathrm{T},\mathrm{Br}),(\mathrm{T},\mathrm{G})\}\,.$$

This is the union of the three outcomes (T, Bl), (T, Br), and (T, G). Similarly, the set of people with blue eyes is obtained from the first column of the table; it is the union of (T, Bl), (M, Bl), and (S, Bl) and can be written

$$\{\mathrm{Bl}\} = \{(\mathrm{T},\mathrm{Bl}),(\mathrm{M},\mathrm{Bl}),(\mathrm{S},\mathrm{Bl})\}\,.$$

If we know that {T} *and* {Bl} both occur, there is only one possible outcome, (T, Bl).

The event that {T} *or* {Bl} occurs consists of all outcomes in either the first row or the first column of the table, i.e.,

$$\{(\mathrm{T},\mathrm{Bl}),(\mathrm{T},\mathrm{Br}),(\mathrm{T},\mathrm{G}),(\mathrm{M},\mathrm{Bl}),(\mathrm{S},\mathrm{Bl})\}\,. \qquad \square$$

EXAMPLE 1.1.2. Table 1.1 contains probabilities for the nine outcomes that are combinations of height and eye color from Example 1.1.1.

TABLE 1.1. Height–eye color probabilities

		Eye color		
		Blue	Brown	Green
Height	Tall	.12	.15	.03
	Medium	.22	.34	.04
	Short	.06	.01	.03

Note that each of the nine numbers is between 0 and 1 and that the sum of all nine equals 1. The probability of blue eyes is

$$\begin{aligned}\Pr(\mathrm{Bl}) &= \Pr[(\mathrm{T},\mathrm{Bl}),(\mathrm{M},\mathrm{Bl}),(\mathrm{S},\mathrm{Bl})]\\ &= \Pr(\mathrm{T},\mathrm{Bl})+\Pr(\mathrm{M},\mathrm{Bl})+\Pr(\mathrm{S},\mathrm{Bl})\\ &= .12+.22+.06\\ &= .4\,.\end{aligned}$$

Similarly, Pr(Br) = .5 and Pr(G) = .1. The probability of not having blue eyes is

$$\begin{aligned}\Pr(\text{not Bl}) &= 1-\Pr(\mathrm{Bl})\\ &= 1-.4\\ &= .6\,.\end{aligned}$$

Note also that $\Pr(\text{not Bl}) = \Pr(\text{Br}) + \Pr(\text{G})$.

The (*marginal*) probabilities for the various heights are:

$$\Pr(\text{T}) = .3, \quad \Pr(\text{M}) = .6, \quad \Pr(\text{S}) = .1\,. \qquad \square$$

Even if there are a countable (but infinite) number of possible outcomes, one can still define a probability by defining the probabilities for each outcome. It is only for measurement data that one really needs to define probabilities on sets.

Two random events are said to be independent if knowing that one of them occurs provides no information about the probability that the other event will occur. Formally, two events A and B are *independent* if

$$\Pr(A \text{ and } B) = \Pr(A)\Pr(B).$$

Thus the probability that *both* events A and B occur is just the product of the individual probabilities that A occurs and that B occurs. As we will begin to see in the next section, independence plays an important role is statistics.

Example 1.1.3. Using the probabilities of Table 1.1 and the computations of Example 1.1.2, the events tall and brown eyes are independent because

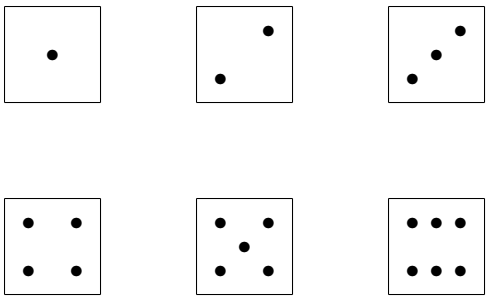
$$\Pr(\text{tall and brown}) = \Pr(\text{T}, \text{Br}) = .15 = (.3)(.5) = \Pr(\text{T}) \times \Pr(\text{Br}).$$

On the other hand, medium height and blue eyes are *not* independent because

$$\Pr(\text{medium and blue}) = \Pr(\text{M}, \text{Bl}) = .22 \neq (.6)(.4) = \Pr(\text{M}) \times \Pr(\text{Bl}). \qquad \square$$

1.2 Random variables and expectations

A *random variable* is simply a function that relates outcomes with numbers. The key point is that any probability associated with the outcomes induces a probability on the numbers. The numbers and their associated probabilities can then be manipulated mathematically. Perhaps the most common and intuitive example of a random variable is rolling a die. The outcome is that a face of the die with a certain number of spots ends up on top. These can be pictured as



Without even thinking about it, we define a random variable that transforms these six faces into the numbers 1, 2, 3, 4, 5, 6.

In statistics we think of observations as random variables. These are often some number associated with a randomly selected member of a population. For example, one random variable is the height of a person who is to be randomly selected from among University of New Mexico students. (A random selection gives the same probability to every individual in the population. This random variable presumes that we have well-defined methods for measuring height and defining UNM students.) Rather than measuring height, we could define a different random variable by giving the person a score of 1 if that person is female and 0 if the person is male. We can also perform mathematical operations on random variables to yield new random variables. Suppose we plan to select a random sample of 10 students, then we would have 10 random variables with female and male scores. The sum of these random variables is another random variable that tells us the (random) number of females in the sample. Similarly, we would have 10 random variables for heights and we can define a new random variable consisting of the average of the 10 individual height random variables. Some random variables are related in obvious ways. In our example we measure both a height and a sex score on each person. If the sex score variable is a 1 (telling us that the person is female), it suggests that the height may be smaller than we would otherwise suspect. Obviously some female students are taller than some male students, but knowing a person's sex definitely changes our knowledge about their probable height.

We do similar things in tossing a coin.

EXAMPLE 1.2.1. Consider tossing a coin twice. The four outcomes are ordered pairs of heads (H) and tails (T). The outcomes can be denoted as

$$(H,H) \qquad (H,T) \qquad (T,H) \qquad (T,T)$$

where the outcome of the first toss is the first element of the ordered pair.

The standard probability model has the four outcomes equally probable, i.e., $1/4 = \Pr(H,H) = \Pr(H,T) = \Pr(T,H) = \Pr(T,T)$. Equivalently

		Second toss		
		Heads	Tails	Total
First toss	Heads	1/4	1/4	1/2
	Tails	1/4	1/4	1/2
	Total	1/2	1/2	1

The probability of heads on each toss is 1/2. The probability of tails is 1/2. We will define two random variables:

$$y_1(r,s) = \begin{cases} 1 & \text{if } r = H \\ 0 & \text{if } r = T \end{cases}$$

$$y_2(r,s) = \begin{cases} 1 & \text{if } s = H \\ 0 & \text{if } s = T \end{cases} .$$

Thus, y_1 is 1 if the first toss is heads and 0 otherwise. Similarly, y_2 is 1 if the second toss is heads and 0 otherwise.

The event $y_1 = 1$ occurs if and only if we get heads on the first toss. We get heads on the first toss by getting either of the outcome pairs (H,H) or (H,T). In other words, the

event $y_1 = 1$ is equivalent to the event $\{(H,H),(H,T)\}$. The probability of $y_1 = 1$ is just the sum of the probabilities of the outcomes in $\{(H,H),(H,T)\}$.

$$\begin{aligned} \Pr(y_1 = 1) &= \Pr(H,H) + \Pr(H,T) \\ &= 1/4 + 1/4 = 1/2. \end{aligned}$$

Similarly,

$$\begin{aligned} \Pr(y_1 = 0) &= \Pr(T,H) + \Pr(T,T) \\ &= 1/2 \\ \Pr(y_2 = 1) &= 1/2 \\ \Pr(y_2 = 0) &= 1/2\,. \end{aligned}$$

Now define another random variable,

$$W(r,s) = y_1(r,s) + y_2(r,s)\,.$$

The random variable W is the total number of heads in two tosses:

$$\begin{aligned} W(H,H) &= 2 \\ W(H,T) &= W(T,H) \;=\; 1 \\ W(T,T) &= 0\,. \end{aligned}$$

Moreover,

$$\begin{aligned} \Pr(W = 2) &= \Pr(H,H) \;=\; 1/4 \\ \Pr(W = 1) &= \Pr(H,T) + \Pr(T,H) \;=\; 1/2 \\ \Pr(W = 0) &= \Pr(T,T) \;=\; 1/4\,. \end{aligned}$$

These three equalities define a probability on the outcomes 0, 1, 2. In working with W, we can ignore the original outcomes of head-tail pairs and work only with the new outcomes 0, 1, 2 and their associated probabilities. We can do the same thing for y_1 and y_2. The probability table given earlier can be rewritten in terms of y_1 and y_2.

		y_2		
		1	0	y_1 totals
y_1	1	1/4	1/4	1/2
	0	1/4	1/4	1/2
	y_2 totals	1/2	1/2	1

Note that, for example, $\Pr[(y_1,y_2) = (1,0)] = 1/4$ and $\Pr(y_1 = 1) = 1/2$. This table shows the *distribution* of the probabilities for y_1 and y_2 both separately (marginally) and jointly. □

For any random variable, *a statement of the possible outcomes and their associated probabilities is referred to as the (marginal)* probability distribution *of the random variable. For two or more random variables, a table or other statement of the possible joint outcomes and their associated probabilities is referred to as the* joint probability distribution *of the random variables.*

All of the entries in the center of the distribution table given above for y_1 and y_2 are independent. For example,

$$\Pr[(y_1, y_2) = (1, 0)] \equiv \Pr(y_1 = 1 \text{ and } y_2 = 0) = \Pr(y_1 = 1)\Pr(y_2 = 0).$$

We therefore say that y_1 and y_2 are independent. In general, *two random variables y_1 and y_2 are independent if any event involving only y_1 is independent of any event involving only y_2.*

Independence is an extremely important concept in statistics. Observations to be analyzed are commonly assumed to be independent. This means that *the random aspect of one observation contains no information about the random aspect of any other observation.* (However, every observation tells us about fixed aspects of the underlying population such as the population center.) *For most purposes in applied statistics, just this intuitive understanding of independence is sufficient.*

1.2.1 Expected values and variances

The *expected value (population mean)* of a random variable is a number characterizing the middle of the distribution. For a random variable y with a discrete distribution (i.e., one having a finite or countable number of outcomes), the expected value is

$$\mathrm{E}(y) \equiv \sum_{\text{all } r} r \Pr(y = r).$$

Example 1.2.2. Let y be the result of picking one of the numbers 2, 4, 6, 8 at random. Because the numbers are chosen at random,

$$1/4 = \Pr(y = 2) = \Pr(y = 4) = \Pr(y = 6) = \Pr(y = 8).$$

The expected value in this simple example is just the mean (average) of the four possible outcomes.

$$\begin{aligned} \mathrm{E}(y) &= 2\left(\frac{1}{4}\right) + 4\left(\frac{1}{4}\right) + 6\left(\frac{1}{4}\right) + 8\left(\frac{1}{4}\right) \\ &= (2 + 4 + 6 + 8)/4 \\ &= 5. \end{aligned}$$

□

Example 1.2.3. Five pieces of paper are placed in a hat. The papers have the numbers 2, 4, 6, 6, and 8 written on them. A piece of paper is picked at random. The expected value of the number drawn is the mean of the numbers on the five pieces of paper. Let y be the random variable that relates a piece of paper to the number on that paper. Each piece of paper has the same probability of being chosen, so, because the number 6 appears twice, the distribution of the random variable y is

$$\frac{1}{5} = \Pr(y = 2) = \Pr(y = 4) = \Pr(y = 8)$$

$$\frac{2}{5} = \Pr(y = 6).$$

The expected value is

$$\begin{aligned} \mathrm{E}(y) &= 2\left(\frac{1}{5}\right) + 4\left(\frac{1}{5}\right) + 6\left(\frac{2}{5}\right) + 8\left(\frac{1}{5}\right) \\ &= (2+4+6+6+8)/5 \\ &= 5.2\,. \end{aligned}$$

□

EXAMPLE 1.2.4. Consider the coin tossing random variables y_1, y_2, and W from Example 1.2.1. Recalling that y_1 and y_2 have the same distribution,

$$\begin{aligned} \mathrm{E}(y_1) &= 1\left(\frac{1}{2}\right) + 0\left(\frac{1}{2}\right) = \frac{1}{2} \\ \mathrm{E}(y_2) &= \frac{1}{2} \\ \mathrm{E}(W) &= 2\left(\frac{1}{4}\right) + 1\left(\frac{1}{2}\right) + 0\left(\frac{1}{4}\right) = 1\,. \end{aligned}$$

The variable y_1 is the number of heads in the first toss of the coin. The two possible values 0 and 1 are equally probable, so the middle of the distribution is $1/2$. W is the number of heads in two tosses; the expected number of heads in two tosses is 1. □

The expected value indicates the middle of a distribution, but does not indicate how spread out (dispersed) a distribution is.

EXAMPLE 1.2.5. Consider three gambles that I will allow you to take. In game z_1 you have equal chances of winning 12, 14, 16, or 18 dollars. In game z_2 you can again win 12, 14, 16, or 18 dollars, but now the probabilities are .1 that you will win either \$14 or \$16 and .4 that you will win \$12 or \$18. The third game I call z_3 and you can win 5, 10, 20, or 25 dollars with equal chances. Being no fool, I require you to pay me \$16 for the privilege of playing any of these games. We can write each game as a random variable.

z_1	outcome	12	14	16	18
	probability	.25	.25	.25	.25
z_2	outcome	12	14	16	18
	probability	.4	.1	.1	.4
z_3	outcome	5	10	20	25
	probability	.25	.25	.25	.25

I try to be a good casino operator, so none of these games is fair. You have to pay \$16 to play, but you only expect to win \$15. It is easy to see that

$$\mathrm{E}(z_1) = \mathrm{E}(z_2) = \mathrm{E}(z_3) = 15\,.$$

But don't forget that I'm taking a loss on the ice-water I serve to players and I also have to pay for the pictures of my extended family that I've decorated my office with.

Although the games z_1, z_2, and z_3 have the same expected value, the games (random variables) are very different. Game z_2 has the same outcomes as z_1, but much more of its probability is placed farther from the middle value 15. The extreme observations 12 and 18 are much more probable under z_2 than z_1. If you currently have \$16, need \$18 for your grandmother's bunion removal, and anything less than \$18 has no value to you, then z_2 is obviously a better game for you than z_1.

Both z_1 and z_2 are much more tightly packed around 15 than is z_3. If you needed \$25 for the bunion removal, z_3 is the game to play because you can win it all in one play with probability .25. In either of the other games you would have to win at least five times to get \$25, a much less likely occurrence. Of course you should realize that the most probable result is that Grandma will have to live with her bunion. You are unlikely to win either \$18 or \$25. While the ethical moral of this example is that a fool and his money are soon parted, the statistical point is that there is more to a random variable than its mean. The variability of random variables is also important. □

The *(population) variance* is a measure of how spread out a distribution is from its expected value. Let y be a random variable having a discrete distribution with $\mathrm{E}(y) = \mu$, then the variance of y is

$$\mathrm{Var}(y) \equiv \sum_{\text{all } r} (r - \mu)^2 \Pr(y = r) .$$

This is the average squared distance of the outcomes from the center of the population. More technically, it is the expected squared distance between the outcomes and the mean of the distribution.

EXAMPLE 1.2.6. Using the random variables of Example 1.2.5,

$$\begin{aligned}
\mathrm{Var}(z_1) &= (12-15)^2(.25) + (14-15)^2(.25) \\
&\quad + (16-15)^2(.25) + (18-15)^2(.25) \\
&= 5 \\
\mathrm{Var}(z_2) &= (12-15)^2(.4) + (14-15)^2(.1) \\
&\quad + (16-15)^2(.1) + (18-15)^2(.4) \\
&= 7.4 \\
\mathrm{Var}(z_3) &= (5-15)^2(.25) + (10-15)^2(.25) \\
&\quad + (20-15)^2(.25) + (25-15)^2(.25) \\
&= 62.5
\end{aligned}$$

The increasing variances from z_1 through z_3 indicate that the random variables are increasingly spread out. However, the value $\mathrm{Var}(z_3) = 62.5$ seems too large to measure the relative variabilities of the three random variables. More on this later. □

EXAMPLE 1.2.7. Consider the coin tossing random variables of Examples 1.2.1 and 1.2.4.

$$\begin{aligned}
\mathrm{Var}(y_1) &= \left(1 - \frac{1}{2}\right)^2 \frac{1}{2} + \left(0 - \frac{1}{2}\right)^2 \frac{1}{2} = \frac{1}{4} \\
\mathrm{Var}(y_2) &= \frac{1}{4} \\
\mathrm{Var}(W) &= (2-1)^2 \left(\frac{1}{4}\right) + (1-1)^2 \left(\frac{1}{2}\right) + (0-1)^2 \left(\frac{1}{4}\right) = \frac{1}{2} .
\end{aligned}$$

□

A problem with the variance is that it is measured on the wrong scale. If y is measured in meters, $\mathrm{Var}(y)$ involves the terms $(r-\mu)^2$; hence it is measured in meters squared. To get things back on the original scale, we consider the *standard deviation* of y

$$\text{Std. dev. } (y) \equiv \sqrt{\mathrm{Var}(y)}\,.$$

EXAMPLE 1.2.8. Consider the random variables of Examples 1.2.5 and 1.2.6.

$$\begin{array}{lcll} \text{Std. dev. } (z_1) & = & \sqrt{5} & \doteq 2.236 \\ \text{Std. dev. } (z_2) & = & \sqrt{7.4} & \doteq 2.720 \\ \text{Std. dev. } (z_3) & \equiv & \sqrt{62.5} & \doteq 7.906 \end{array}$$

The standard deviation of z_3 is 3 to 4 times larger than the others. From examining the distributions, the standard deviations seem to be more intuitive measures of relative variability than the variances. The variance of z_3 is 8.5 to 12.5 times larger than the other variances; these values seem unreasonably inflated. □

Standard deviations and variances are useful as measures of the relative dispersions of different random variables. The actual numbers themselves do not mean much. Moreover, there are other equally good measures of dispersion that can give results that are somewhat inconsistent with these. One reason standard deviations and variances are so widely used is because they are convenient mathematically. In addition, normal (Gaussian) distributions are widely used in applied statistics and are completely characterized by their expected values (means) and variances (or standard deviations). Knowing these two numbers, the mean and variance, one knows everything about a normal distribution.

1.2.2 Chebyshev's inequality

Another place in which the numerical values of standard deviations are useful is in applications of Chebyshev's inequality. Chebyshev's inequality gives a lower bound on the probability that a random variable is within an interval. Chebyshev's inequality is important in quality control work (control charts) and in evaluating prediction intervals.

Let y be a random variable with $\mathrm{E}(y) = \mu$ and $\mathrm{Var}(y) = \sigma^2$. Chebyshev's inequality states that for any number $k > 1$,

$$\Pr[\mu - k\sigma < y < \mu + k\sigma] \geq 1 - \frac{1}{k^2}.$$

Thus the probability that y will fall within k standard deviations of μ is at least $1-(1/k^2)$.

The beauty of Chebyshev's inequality is that it holds for absolutely any random variable y. Thus we can always make some statement about the probability that y is in a symmetric interval about μ. In many cases, for particular choices of y, the probability of being in the interval can be much greater than $1-k^{-2}$. For example, if $k=3$ and y has a normal distribution as discussed in the next section, the probability of being in the interval is actually .997, whereas Chebyshev's inequality only assures us that the probability is no less than $1-3^{-2} = .889$. However, we know the lower bound of .889 applies regardless of whether y has a normal distribution.

1.2.3 Covariances and correlations

Often we take two (or more) observations on the same member of a population. We might observe the height and weight of a person. We might observe the IQs of a wife and husband.

(Here the population consists of married couples.) In such cases we may want a numerical measure of the relationship between the pairs of observations. Data analysis related to these concepts is known as regression analysis and is discussed in Chapters 7, 13, 14, and 15. These ideas are also briefly used for testing normality in Section 2.4.

The *covariance* is a measure of the linear relationship between two random variables. Suppose y_1 and y_2 are discrete random variables. Let $\mathrm{E}(y_1) = \mu_1$ and $\mathrm{E}(y_2) = \mu_2$. The covariance between y_1 and y_2 is

$$\mathrm{Cov}(y_1, y_2) \equiv \sum_{\text{all } (r,s)} (r - \mu_1)(s - \mu_2) \Pr(y_1 = r, y_2 = s) .$$

Positive covariances arise when relatively large values of y_1 tend to occur with relatively large values y_2 and small values of y_1 tend to occur with small values of y_2. On the other hand, negative covariances arise when relatively large values of y_1 tend to occur with relatively small values y_2 and small values of y_1 tend to occur with large values of y_2. It is simple to see from the definition that, for example,

$$\mathrm{Var}(y_1) = \mathrm{Cov}(y_1, y_1) .$$

In an attempt to get a handle on what the numerical value of the covariance means, it is often rescaled into a *correlation coefficient*.

$$\mathrm{Corr}(y_1, y_2) \equiv \mathrm{Cov}(y_1, y_2) / \sqrt{\mathrm{Var}(y_1)\mathrm{Var}(y_2)} .$$

Positive values of the correlation have the same qualitative meaning as positive values of the covariance, but now a *perfect* increasing linear relationship is indicated by a correlation of 1. Similarly, negative correlations and covariances mean similar things, but a perfect decreasing linear relationship gives a correlation of -1. The absence of any linear relationship is indicated by a value of 0.

A perfect linear relationship between y_1 and y_2 means that an increase of one unit in, say, y_1 dictates an exactly proportional change in y_2. For example, if we make a series of very accurate temperature measurements on something and simultaneously use one device calibrated in Fahrenheit and one calibrated in Celsius, the pairs of numbers should have an essentially perfect linear relationship.

EXAMPLE 1.2.9. Let z_1 and z_2 be two random variables defined by the following probability table:

		z_2			
		0	1	2	z_1 totals
	6	0	1/3	0	1/3
z_1	4	1/3	0	0	1/3
	2	0	0	1/3	1/3
	z_2 totals	1/3	1/3	1/3	1

Then

$$\begin{aligned} E(z_1) &= 6\left(\frac{1}{3}\right) + 4\left(\frac{1}{3}\right) + 2\left(\frac{1}{3}\right) = 4, \\ E(z_2) &= 0\left(\frac{1}{3}\right) + 1\left(\frac{1}{3}\right) + 2\left(\frac{1}{3}\right) = 1, \end{aligned}$$

$$\begin{aligned}\mathrm{Var}(z_1) &= (2-4)^2\left(\frac{1}{3}\right)+(4-4)^2\left(\frac{1}{3}\right)+(6-4)^2\left(\frac{1}{3}\right)\\ &= 8/3,\\ \mathrm{Var}(z_2) &= (0-1)^2\left(\frac{1}{3}\right)+(1-1)^2\left(\frac{1}{3}\right)+(2-1)^2\left(\frac{1}{3}\right)\\ &= 2/3,\end{aligned}$$

$$\begin{aligned}\mathrm{Cov}(z_1,z_2) &= (2-4)(0-1)(0)+(2-4)(1-1)(0)+(2-4)(2-1)\left(\frac{1}{3}\right)\\ &\quad +(4-4)(0-1)\left(\frac{1}{3}\right)+(4-4)(1-1)(0)+(4-4)(2-1)(0)\\ &\quad +(6-4)(0-1)(0)+(6-4)(1-1)\left(\frac{1}{3}\right)+(6-4)(2-1)(0)\\ &= -2/3,\end{aligned}$$

$$\begin{aligned}\mathrm{Corr}(z_1,z_2) &= (-2/3)\Big/\sqrt{(8/3)(2/3)}\\ &= -1/2.\end{aligned}$$

This correlation indicates that relatively large z_1 values tend to occur with relatively small z_2 values. However, the correlation is considerably greater than -1, so the linear relationship is less than perfect. Moreover, the correlation measures the linear relationship and *fails to identify the perfect nonlinear relationship* between z_1 and z_2. If $z_1 = 2$, then $z_2 = 2$. If $z_1 = 4$, then $z_2 = 0$. If $z_1 = 6$, then $z_2 = 1$. If you know one random variable, you know the other, but because the relationship is nonlinear, the correlation is not ± 1. □

Example 1.2.10. Consider the coin toss random variables y_1 and y_2 from Example 1.2.1. We earlier observed that these two random variables are independent. If so, there should be no relationship between them (linear or otherwise). We now show that their covariance is 0.

$$\begin{aligned}\mathrm{Cov}(y_1,y_2) &= \left(0-\frac{1}{2}\right)\left(0-\frac{1}{2}\right)\frac{1}{4}+\left(0-\frac{1}{2}\right)\left(1-\frac{1}{2}\right)\frac{1}{4}\\ &\quad +\left(1-\frac{1}{2}\right)\left(0-\frac{1}{2}\right)\frac{1}{4}+\left(1-\frac{1}{2}\right)\left(1-\frac{1}{2}\right)\frac{1}{4}\\ &= \frac{1}{16}-\frac{1}{16}-\frac{1}{16}+\frac{1}{16}=0.\end{aligned}$$

□

In general, whenever two random variables are independent, their covariance (and thus their correlation) is 0. However, just because two random variables have 0 covariance does not *imply that they are independent. Independence has to do with not having any kind of relationship; covariance examines only linear relationships. Random variables with nonlinear relationships can have zero covariance but not be independent.*

1.2.4 Rules for expected values and variances

We now present some extremely useful results that allow us to show that statistical estimates are reasonable and to establish the variability associated with statistical estimates. These

results relate to the expected values, variances, and covariances of linear combinations of random variables. A linear combination of random variables is something that only involves multiplying random variables by fixed constants, adding such terms together, and adding a constant.

Proposition 1.2.11. Let y_1, y_2, y_3, and y_4 be random variables and let a_1, a_2, a_3, and a_4 be real numbers.

1. $\mathrm{E}(a_1y_1 + a_2y_2 + a_3) = a_1\mathrm{E}(y_1) + a_2\mathrm{E}(y_2) + a_3$.
2. If y_1 and y_2 are independent, $\mathrm{Var}(a_1y_1 + a_2y_2 + a_3) = a_1^2\mathrm{Var}(y_1) + a_2^2\mathrm{Var}(y_2)$.
3. $\mathrm{Var}(a_1y_1 + a_2y_2 + a_3) = a_1^2\mathrm{Var}(y_1) + 2a_1a_2\mathrm{Cov}(y_1, y_2) + a_2^2\mathrm{Var}(y_2)$.
4. $\mathrm{Cov}(a_1y_1 + a_2y_2, a_3y_3 + a_4y_4) = a_1a_3\mathrm{Cov}(y_1, y_3) + a_1a_4\mathrm{Cov}(y_1, y_4) + a_2a_3\mathrm{Cov}(y_2, y_3) + a_2a_4\mathrm{Cov}(y_2, y_4)$.

All of these results generalize to linear combinations involving more than two random variables.

Example 1.2.12. Recall that when independently tossing a coin twice, the total number of heads, W, is the sum of y_1 and y_2, the number of heads on the first and second tosses respectively. We have already seen that $\mathrm{E}(y_1) = \mathrm{E}(y_2) = .5$ and that $\mathrm{E}(W) = 1$. We now illustrate item 1 of the proposition by finding $\mathrm{E}(W)$ again. Since $W = y_1 + y_2$,

$$\mathrm{E}(W) = \mathrm{E}(y_1 + y_2) = \mathrm{E}(y_1) + \mathrm{E}(y_2) = .5 + .5 = 1.$$

We have also seen that $\mathrm{Var}(y_1) = \mathrm{Var}(y_2) = .25$ and that $\mathrm{Var}(W) = .5$. Since the coin tosses are independent, item 2 above gives

$$\mathrm{Var}(W) = \mathrm{Var}(y_1 + y_2) = \mathrm{Var}(y_1) + \mathrm{Var}(y_2) = .25 + .25 = .5\,.$$

The key point is that this is an easier way of finding the expected value and variance of W than using the original definitions. □

We now illustrate the generalizations referred to in Proposition 1.2.11. We begin by looking at the problem of estimating the mean of a population.

Example 1.2.13. Let y_1, y_2, y_3, and y_4 be four random variables each with the same (population) mean μ, i.e., $\mathrm{E}(y_i) = \mu$ for $i = 1, 2, 3, 4$. We can compute the *sample mean* (average) of these, defining

$$\begin{aligned} \bar{y}_{\cdot} &\equiv \frac{y_1 + y_2 + y_3 + y_4}{4} \\ &= \frac{1}{4}y_1 + \frac{1}{4}y_2 + \frac{1}{4}y_3 + \frac{1}{4}y_4. \end{aligned}$$

The $\cdot$ in the subscript of $\bar{y}_{\cdot}$ indicates that the sample mean is obtained by summing over the subscripts of the y_is. The $\cdot$ notation is not necessary for this problem but becomes useful in dealing with the analysis of variance problems treated later in the book.

Using item 1 of Proposition 1.2.11 we find that

$$\begin{aligned}\mathrm{E}(\bar{y}_\cdot) &= \mathrm{E}\left(\frac{1}{4}y_1 + \frac{1}{4}y_2 + \frac{1}{4}y_3 + \frac{1}{4}y_4\right)\\ &= \frac{1}{4}\mathrm{E}(y_1) + \frac{1}{4}\mathrm{E}(y_2) + \frac{1}{4}\mathrm{E}(y_3) + \frac{1}{4}\mathrm{E}(y_4)\\ &= \frac{1}{4}\mu + \frac{1}{4}\mu + \frac{1}{4}\mu + \frac{1}{4}\mu\\ &= \mu.\end{aligned}$$

Thus one observation on $\bar{y}_\cdot$ would make a reasonable estimate of μ.

If we also assume that the y_is are independent with the same variance, say, σ^2, then from item 2 of Proposition 1.2.11

$$\begin{aligned}\mathrm{Var}(\bar{y}_\cdot) &= \mathrm{Var}\left(\frac{1}{4}y_1 + \frac{1}{4}y_2 + \frac{1}{4}y_3 + \frac{1}{4}y_4\right)\\ &= \left(\frac{1}{4}\right)^2 \mathrm{Var}(y_1) + \left(\frac{1}{4}\right)^2 \mathrm{Var}(y_2)\\ &\quad + \left(\frac{1}{4}\right)^2 \mathrm{Var}(y_3) + \left(\frac{1}{4}\right)^2 \mathrm{Var}(y_4)\\ &= \left(\frac{1}{4}\right)^2 \sigma^2 + \left(\frac{1}{4}\right)^2 \sigma^2 + \left(\frac{1}{4}\right)^2 \sigma^2 + \left(\frac{1}{4}\right)^2 \sigma^2\\ &= \frac{\sigma^2}{4}.\end{aligned}$$

The variance of $\bar{y}_\cdot$ is only one fourth of the variance of an individual observation. Thus the $\bar{y}_\cdot$ observations are more tightly packed around their mean μ than the y_is are. This indicates that one observation on $\bar{y}_\cdot$ is more likely to be close to μ than an individual y_i. □

These results for $\bar{y}_\cdot$ hold quite generally; they are not restricted to the average of four random variables. *If* $\bar{y}_\cdot = (1/n)(y_1 + \cdots + y_n) = \sum_{i=1}^n y_i/n$ *is the sample mean of* n *independent random variables all with the same population mean* μ *and population variance* σ^2,

$$\mathrm{E}(\bar{y}_\cdot) = \mu$$

and

$$\mathrm{Var}(\bar{y}_\cdot) = \frac{\sigma^2}{n}.$$

In fact, proving these general results uses exactly the same ideas as the proofs for a sample of size 4.

As with a sample of size 4, the general results on $\bar{y}_\cdot$ are very important in statistical inference. If we are interested in determining the population mean μ from future data, the obvious estimate is the average of the individual observations, $\bar{y}_\cdot$. The observations are random, so the estimate $\bar{y}_\cdot$ is also a random variable and the middle of its distribution is $\mathrm{E}(\bar{y}_\cdot) = \mu$, the original population mean. Thus $\bar{y}_\cdot$ is a reasonable estimate of μ. Moreover, $\bar{y}_\cdot$ is a better estimate than any particular observation y_i because $\bar{y}_\cdot$ has a smaller variance, σ^2/n as opposed to σ^2 for y_i. With less variability in the estimate, any one observation of $\bar{y}_\cdot$ is more likely to be near its mean μ than a single observation y_i. In practice, we obtain data and compute a sample mean. This constitutes one observation on the random variable $\bar{y}_\cdot$. If our sample mean is to be a good estimate of μ, our one look at $\bar{y}_\cdot$ had better have

a good chance of being close to μ. This occurs when the variance of $\bar{y}_\cdot$ is small. Note that the larger the sample size n, the smaller is σ^2/n, the variance of $\bar{y}_\cdot$. We will return to these ideas later.

Generally, we will use item 1 of Proposition 1.2.11 to show that estimates are *unbiased*. In other words, we will show that the expected value of an estimate is what we are trying to estimate. In estimating μ, we have $\mathrm{E}(\bar{y}_\cdot) = \mu$, so $\bar{y}_\cdot$ is an unbiased estimate of μ. All this really does is show that $\bar{y}_\cdot$ is a reasonable estimate of μ. More important than showing unbiasedness is using item 2 to find variances of estimates. Statistical inference depends crucially on having some idea of the variability of an estimate. Item 2 is the primary tool in finding the appropriate variance for different estimates.

1.3 Continuous distributions

As discussed in Section 1.1, many things that we would like to measure are, in the strictest sense, not measurable. We cannot find a building's exact height even though we can approximate it *extremely* accurately. This theoretical inability to measure things exactly has little impact on our practical world, but it has a substantial impact on the theory of statistics.

The data in most statistical applications can be viewed as counts of how often some event has occurred or as measurements. Probabilities associated with count data are easy to describe. We discuss some probability models for count data in Sections 1.4 and 1.5. With measurement data, we can never obtain an exact value, so we don't even try. With measurement data, we assign probabilities to intervals. Thus we do not discuss the probability that a person has the height 177.8 cm or 177.8001 cm or 56.5955π cm, but we do discuss the probability that someone has a height *between* 177.75 cm and 177.85 cm. Typically, we think of doing this in terms of pictures. We associate probabilities with areas under curves. (Mathematically, this involves integral calculus and is discussed in a brief appendix at the end of the chapter.) Figure 1.1 contains a picture of a continuous probability distribution (*a density*). Probabilities must be between 0 and 1, so the curve must always be nonnegative (to make all areas nonnegative) and the area under the entire curve must be 1.

Figure 1.1 also shows a point $K(1-\alpha)$. This point divides the area under the curve into two parts. The probability of obtaining a number less than $K(1-\alpha)$ is $1-\alpha$, i.e., the area under the curve to the left of $K(1-\alpha)$ is $1-\alpha$. The probability of obtaining a number greater than $K(1-\alpha)$ is α, i.e., the area under the curve to the right of $K(1-\alpha)$. $K(1-\alpha)$ is a particular number, so the probability is 0 that $K(1-\alpha)$ will actually occur. There is no area under a curve associated with any particular point.

Pictures such as Figure 1.1 are often used as models for populations of measurements. With a fixed population of measurements, it is natural to form a histogram, i.e., a bar chart that plots intervals for the measurement against the proportion of individuals that fall into a particular interval. Pictures such as Figure 1.1 can be viewed as approximations to such histograms. The probabilities described by pictures such as Figure 1.1 are those associated with randomly picking an individual from the population. Thus, randomly picking an individual from the population modeled by Figure 1.1 yields a measurement less than $K(1-\alpha)$ with probability $1-\alpha$.

Ideas similar to those discussed in Section 1.2 can be used to define expected values, variances, and covariances for continuous distributions. These extensions involve integral calculus and are discussed in the appendix. In any case, Proposition 1.2.11 continues to apply.

The most commonly used distributional model for measurement data is the *normal* distribution (also called the *Gaussian* distribution). The bell shaped curve in Figure 1.1 is

FIGURE 1.1. A continuous probability density.

referred to as the standard normal curve. The formula for writing the curve is not too ugly, it is

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

Here e is the base of natural logarithms. Unfortunately, even with calculus it is very difficult to compute areas under this curve. Finding standard normal probabilities requires a table.

By itself, the standard normal curve has little value in modeling measurements. For one thing, the curve is centered about 0. I don't take many measurements where I think the central value should be 0. To make the normal distribution a useful model, we need to expand the standard normal into a family of distributions with different centers (expected values) μ and different spreads (standard deviations) σ. By appropriate recentering and rescaling of the plot, all of these curves will have the same shape as Figure 1.1.

The standard normal distribution is the special case of a normal with $\mu = 0$ and $\sigma = 1$. The standard normal plays an important role because it is the only normal distribution that we need tabled. (Obviously, we could not table normal distributions for every possible value of μ and σ.) Suppose a measurement y has a normal distribution with mean μ, standard deviation σ, and variance σ^2. We write this as

$$y \sim N(\mu, \sigma^2).$$

Normal distributions have the property that

$$\frac{y - \mu}{\sigma} \sim N(0, 1),$$

cf. Exercise 1.6.2. This standardization process allows us to get by with only the standard normal table for finding probabilities for all normal distributions.

The standard normal distribution is sometimes used in constructing statistical inferences but more often a similar distribution is used. When data are normally distributed, statistical inferences often require something called Student's t distribution. (Student was the pen name of W. S. Gosset.) The t distribution is a family of distributions all of which look

roughly like Figure 1.1. They are all symmetric about 0, but they have slightly different amounts of dispersion (spread). The amount of variability in each distribution is determined by a positive integer parameter called the *degrees of freedom*. With only 1 degree of freedom, the mathematical properties of a t distribution are fairly bizarre. (This special case is called a Cauchy distribution.) As the number of degrees of freedom get larger, the t distributions get better behaved and have less variability. As the degrees of freedom gets arbitrarily large, the t distribution approximates the standard normal distribution.

Two other distributions that come up later are the chi-squared distribution (χ^2) and the F distribution. These arise naturally when drawing conclusions about the population variance from data that are normally distributed. Both distributions differ from those just discussed in that both are asymmetric and both are restricted to positive numbers. However, the basic idea of probabilities being areas under curves remains unchanged.

In Section 1.2, we introduced Chebyshev's inequality. Shewhart (1931, p. 177) discusses work by Camp and Meidell that allows us to improve on Chebyshev's inequality for continuous distributions. Once again let $\mathrm{E}(y) = \mu$ and $\mathrm{Var}(y) = \sigma^2$. If the density, i.e., the function that defines the curve, is symmetric, unimodal (has only one peak), and always decreases as one moves farther away from the mode, then the inequality can be sharpened to

$$\Pr[\mu - k\sigma < y < \mu + k\sigma] \geq 1 - \frac{1}{(2.25)k^2}.$$

As discussed in the previous section, with y normal and $k = 3$, the true probability is .997, Chebyshev's inequality gives a lower bound of .889, and the new improved Chebyshev inequality gives a lower bound of .951. By making some relatively innocuous assumptions, we get a substantial improvement in the lower bound.

1.4 The binomial distribution

There are a few distributions that are used in the vast majority of statistical applications. The reason for this is that they tend to occur naturally. The normal distribution is one. As discussed in the next chapter, the normal distribution occurs in practice because a result called The Central Limit Theorem dictates that many distributions can be approximated by the normal. Two other distributions, the binomial and the multinomial, occur in practice because they are very simple. In this section we discuss the binomial. The next section introduces the multinomial distribution. The results of this section are only used in Chapter 8 and in discussions of transformations.

If you have independent identical random trials and count how often something (anything) occurs, the appropriate distribution is the binomial. What could be simpler?

EXAMPLE 1.4.1. Being somewhat lonely in my misspent youth, I decided to go to a dating service. The service was to provide me with five dates. Being a very open-minded soul, I convinced myself that the results of one date would not influence my opinion about other dates. From my limited experience with the opposite sex, I have found that I enjoy about 40% of such brief encounters. I decided that my money would be well spent if I enjoyed two or more of the five dates. Unfortunately, my loan shark repossessed my 1954 Studebaker before I could indulge in this taste of nirvana. Back in those days, we chauvinists believed: no wheels – no women. Nevertheless, let us compute the probability that I would have been satisfied with the dating service. Let W be the number of dates I would have enjoyed. The simplest way to find the probability of satisfaction is

$$\begin{aligned}\Pr(W \geq 2) &= 1 - \Pr(W < 2) \\ &= 1 - \Pr(W = 0) - \Pr(W = 1)\,,\end{aligned}$$

but that is much too easy. Let's compute

$$\Pr(W \geq 2) = \Pr(W = 2) + \Pr(W = 3) + \Pr(W = 4) + \Pr(W = 5)\,.$$

In particular, we compute each term on the right-hand side.

Write the outcome of the five dates as an ordered collection of Ls and Ds. For example, (L, D, L, D, D) indicates that I like the first and third dates, but dislike the second, fourth, and fifth.

To like five dates, I must like everyone of them.

$$\Pr(W = 5) = \Pr(L, L, L, L, L)\,.$$

Remember, I assumed that the dates were independent and that the probability of my liking any one is .4. Thus,

$$\begin{aligned}\Pr(W = 5) &= \Pr(L)\Pr(L)\Pr(L)\Pr(L)\Pr(L) \\ &= (.4)^5\,.\end{aligned}$$

The probability of liking four dates is a bit more complicated. I could only dislike one date, but there are five different choices for the date that I could dislike. It could be the fifth, the fourth, the third, the second, or the first. Any pattern of 4 Ls and a D excludes the other patterns from occurring, e.g., if the only date I dislike is the fourth, then the only date I dislike cannot be the second. Since the patterns are mutually exclusive (disjoint), the probability of disliking one date is the sum of the probabilities of the individual patterns.

$$\begin{aligned}\Pr(W = 4) &= \Pr(L, L, L, L, D) \\ &\quad + \Pr(L, L, L, D, L) \\ &\quad + \Pr(L, L, D, L, L) \\ &\quad + \Pr(L, D, L, L, L) \\ &\quad + \Pr(D, L, L, L, L)\,.\end{aligned} \tag{1.4.1}$$

By assumption $\Pr(L) = .4$, so $\Pr(D) = 1 - \Pr(L) = 1 - .4 = .6$. The dates are independent, so

$$\begin{aligned}\Pr(L, L, L, L, D) &= \Pr(L)\Pr(L)\Pr(L)\Pr(L)\Pr(D) \\ &= (.4)^4.6\,.\end{aligned}$$

Similarly,

$$\begin{aligned}\Pr(L, L, L, D, L) &= \Pr(L, L, D, L, L) \\ &= \Pr(L, D, L, L, L) \\ &= \Pr(D, L, L, L, L) \\ &= (.4)^4.6\,.\end{aligned}$$

Summing up the values in equation (1.4.1),

$$\Pr(W = 4) = 5(.4)^4(.6)\,.$$

Computing the probability of liking three dates is even worse.

$$\begin{aligned}\Pr(W=3) &= \Pr(L,L,L,D,D)\\ &\quad+\Pr(L,L,D,L,D)\\ &\quad+\Pr(L,D,L,L,D)\\ &\quad+\Pr(D,L,L,L,D)\\ &\quad+\Pr(L,L,D,D,L)\\ &\quad+\Pr(L,D,L,D,L)\\ &\quad+\Pr(D,L,L,D,L)\\ &\quad+\Pr(L,D,D,L,L)\\ &\quad+\Pr(D,L,D,L,L)\\ &\quad+\Pr(D,D,L,L,L)\end{aligned}$$

Again all of these patterns have exactly the same probability. For example, using independence

$$\Pr(D,L,D,L,L) = (.4)^3(.6)^2\,.$$

Adding up all of the patterns

$$\Pr(W=3) = 10(.4)^3(.6)^2\,.$$

By now it should be clear that

$$\Pr(W=2) = (\text{no. of patterns with 2 Ls and 3 Ds})(.4)^2(.6)^3\,.$$

The number of patterns can be computed as

$$\binom{5}{2} \equiv \frac{5!}{2!(5-2)!} \equiv \frac{5\cdot 4\cdot 3\cdot 2\cdot 1}{(2\cdot 1)(3\cdot 2\cdot 1)} = 10\,.$$

The probability that I would be satisfied with the dating service is

$$\begin{aligned}\Pr(W\geq 2) &= 10(.4)^2(.6)^3 + 10(.4)^3(.6)^2 + 5(.4)^4.6 + (.4)^5\\ &= .663\,.\end{aligned}$$

□

Binomial random variables can also be generated by sampling from a fixed population. If we were going to make 20 random selections from the UNM student body, the number of females would have a binomial distribution. Given a set of procedures for defining and sampling the student body, there would be some fixed number of students of which a given number would be females. Under random sampling, the probability of selecting a female on any of the 20 trials would be simply the proportion of females in the population. Although it is very unlikely to occur in this example, the sampling scheme must allow the possibility of students being selected more than once in the sample. If people were not allowed to be chosen more than once, each successive selection would change the proportion of females available for the subsequent selection. Of course, when making 20 selections out of a population of over 20,000 UNM students, even if you did not allow people to be reselected, the changes in the proportions of females are insubstantial and the binomial distribution makes a good approximation to the true distribution. On the other hand, if the entire student population was 40 rather than 20,000+, it might not be wise to use the binomial approximation when people are not allowed to be reselected.

Typically, the outcome of interest in a binomial is referred to as a success. If the probability of a success is p for each of N independent identical trials, then the number of successes y has a binomial distribution with parameters N and p. Write

$$y \sim \text{Bin}(N, p)\,.$$

The distribution of y is

$$\Pr(y = r) = \binom{N}{r} p^r (1-p)^{N-r}$$

for $r = 0, 1, \ldots, N$. Here

$$\binom{N}{r} \equiv \frac{N!}{r!(N-r)!}$$

where for any positive integer m, $m! \equiv m(m-1)(m-2)\cdots(2)(1)$ and $0! \equiv 1$. The notation $\binom{N}{r}$ is read "N choose r" because it is the number of distinct ways of choosing r individuals out of a collection containing N individuals.

EXAMPLE 1.4.2. The random variables in Example 1.2.1 were y_1, the number of heads on the first toss of a coin, y_2, the number of heads on the second toss of a coin, and W, the combined number of heads from the two tosses. These have the following distributions:

$$\begin{aligned} y_1 &\sim \text{Bin}\left(1, \frac{1}{2}\right) \\ y_2 &\sim \text{Bin}\left(1, \frac{1}{2}\right) \\ W &\sim \text{Bin}\left(2, \frac{1}{2}\right)\,. \end{aligned}$$

Note that W, the $\text{Bin}\left(2, \frac{1}{2}\right)$, was obtained by adding together the two independent $\text{Bin}\left(1, \frac{1}{2}\right)$ random variables y_1 and y_2. This result is quite general. Any $\text{Bin}\,(N, p)$ random variable can be written as the sum of N independent $\text{Bin}\,(1, p)$ random variables. □

Given the probability distribution of a binomial, we can find the mean (expected value) and variance. By definition, if $y \sim \text{Bin}(N, p)$, the mean is

$$\text{E}(y) = \sum_{r=0}^{N} r \binom{N}{r} p^r (1-p)^{N-r}\,.$$

This is difficult to evaluate directly, but by writing y as the sum of N independent $\text{Bin}\,(1, p)$ random variables and using Exercise 1.6.1 and Proposition 1.2.11, it is easily seen that

$$\text{E}(y) = Np\,.$$

Similarly, the variance of y is

$$\text{Var}(y) = \sum_{r=0}^{N} (r - Np)^2 \binom{N}{r} p^r (1-p)^{N-r}$$

but by again writing y as the sum of N independent $\text{Bin}\,(1, p)$ random variables and using Exercise 1.6.1 and Proposition 1.2.11, it is easily seen that

$$\text{Var}(y) = Np(1-p)\,.$$

Exercise 1.6.8 consists of proving these mean and variance formulae.

On occasion we will need to look at both the number of successes from a group of N trials and the number of failures at the same time. If the number of successes is y_1 and the number of failures is y_2, then

$$\begin{aligned} y_2 &= N - y_1 \\ y_1 &\sim \text{Bin}(N, p) \end{aligned}$$

and

$$y_2 \sim \text{Bin}(N, 1-p)\,.$$

The last result holds because, with independent identical trials, the number of outcomes that we call failures must also have a binomial distribution. If p is the probability of success, the probability of failure is $1-p$. Of course,

$$\begin{aligned} E(y_2) &= N(1-p) \\ \text{Var}(y_2) &= N(1-p)p\,. \end{aligned}$$

Note that $\text{Var}(y_1) = \text{Var}(y_2)$ regardless of the value of p. Finally,

$$\text{Cov}(y_1, y_2) = -Np(1-p)$$

and

$$\text{Corr}(y_1, y_2) = -1\,.$$

There is a perfect linear relationship between y_1 and y_2. If y_1 goes up one count, y_2 goes down one count. When we look at both successes and failures write

$$(y_1, y_2) \sim \text{Bin}\big(N, p, (1-p)\big)\,.$$

This is the simplest case of the multinomial distribution discussed in the next section.

1.5 The multinomial distribution

The multinomial distribution is a generalization of the binomial allowing more than two categories. The results in this section are only used in Chapter 8.

EXAMPLE 1.5.1. Consider the probabilities for the nine height and eye color categories given in Example 1.1.2. The probabilities are repeated below.

Height–eye color probabilities

		Eye color		
		Blue	Brown	Green
Height	Tall	.12	.15	.03
	Medium	.22	.34	.04
	Short	.06	.01	.03

Suppose a random sample of 50 individuals was obtained with these probabilities. For example, one might have a population of 100 people in which 12 were tall with blue eyes, 15 were tall with brown eyes, 3 were short with green eyes, etc. We could randomly select one of the 100 people as the first individual in the sample. Then, returning that individual

to the population, take another random selection from the 100 to be the second individual. We are to proceed in this way until 50 people are selected. Note that with a population of 100 and a sample of 50 there is a substantial chance that some people would be selected more than once. The numbers of selections falling into each of the nine categories has a multinomial distribution with $N = 50$ and these probabilities.

It is unlikely that one would actually perform sampling from a population of 100 people as described above. Typically, one would not allow the same person to be chosen more than once. However, if we had a population of 10,000 people where 1200 were tall with blue eyes, 1500 were tall with brown eyes, 300 were short with green eyes, etc., with a sample size of 50 we might be willing to allow the possibility of selecting the same person more than once simply because it is extremely unlikely to happen. Technically, to obtain the multinomial distribution with $N = 50$ and these probabilities, when sampling from a fixed population we need to allow individuals to appear more than once. However, when taking a small sample from a large population, it does not matter much whether or not you allow people to be chosen more than once, so the multinomial often provides a good approximation even when individuals are excluded from reappearing in the sample. □

Consider a group of N independent identical trials in which each trial results in the occurrence of one of q events. Let y_i, $i = 1, \ldots, q$ be the number of times that the ith event occurs and let p_i be the probability that the ith event occurs on any trial. The p_is must satisfy $p_1 + p_2 + \cdots + p_q = 1$. We say that $(y_1, \ldots, y_q)$ has a multinomial distribution with parameters $N, p_1, \ldots, p_q$. Write

$$(y_1, \ldots, y_q) \sim \text{Mult}(N, p_1, \ldots, p_q)\,.$$

The distribution is given by the probabilities

$$\begin{aligned} \Pr(y_1 = r_1, \ldots, y_q = r_q) &= \frac{N!}{r_1! \cdots r_q!} p_1^{r_1} \cdots p_q^{r_q} \\ &= \left(N! \Big/ \prod_{i=1}^{q} r_i! \right) \prod_{i=1}^{q} p_i^{r_i}. \end{aligned}$$

Here the r_is are allowed to be any whole numbers with each $r_i \geq 0$ and $r_1 + \cdots + r_q = N$. Note that if $q = 2$, this is just a binomial distribution. In general, each individual component y_i of a multinomial consists of N trials in which category i either occurs or does not occur, so individual components have the marginal distributions

$$y_i \sim \text{Bin}(N, p_i).$$

It follows that

$$\text{E}(y_i) = Np_i$$

and

$$\text{Var}(y_i) = Np_i(1 - p_i)\,.$$

It can also be shown that

$$\text{Cov}(y_i, y_j) = -Np_ip_j \qquad \text{for} \qquad i \neq j\,.$$

Example 1.5.2. Suppose that the 50 individuals from Example 1.5.1 fall into the categories as listed below.

Height–eye color observations

		Eye color		
		Blue	Brown	Green
Height	Tall	5	8	2
	Medium	10	18	2
	Short	3	1	1

The probability of getting this particular table is

$$\frac{50!}{5!8!2!10!18!2!3!1!1!}(.12)^5(.15)^8(.03)^2(.22)^{10}(.34)^{18}(.04)^2(.06)^3(.01)^1(.03)^1.$$

This number is zero to over 5 decimal places. The fact that this is a very small number is not surprising. There are a lot of possible tables, so the probability of getting any particular table is very small. In fact, many of the possible tables are *much* less likely to occur than this table.

Let's return to thinking about the observations as random. The expected number of observations for each category is given by Np_i. It is easily seen that the expected counts for the cells are as given below.

Height–eye color expected values

		Eye color		
		Blue	Brown	Green
Height	Tall	6.0	7.5	1.5
	Medium	11.0	17.0	2.0
	Short	3.0	0.5	1.5

Note that the expected counts need not be integers.

The variance for, say, the number of tall blue-eyed people in this population is $50(.12)(1-.12) = 5.28$. The variance of the number of short green-eyed people is $50(.03)(1-.03) = 1.455$. The covariance between the number of tall blue-eyed people and the number of short green-eyed people is $-50(.12)(.03) = -.18$. The correlation between the numbers of tall blue-eyed people and short green-eyed people is $-.18/\sqrt{(5.28)(1.455)} = -0.065$. □

Appendix: probability for continuous distributions

As stated in Section 1.3, probabilities are sometimes defined as areas under a curve. The curve, called a probability density function or just a density, must be defined by some nonnegative function $f(\cdot)$. (Nonnegative to ensure that probabilities are always positive.) Thus the probability that a random observation y is between two numbers, say a and b, is the area under the curve measured between a and b. Using calculus, this is

$$\Pr[a < y < b] = \int_a^b f(y)\,\mathrm{d}y.$$

Because we are measuring areas under curves, there is no area associated with any one point, so $\Pr[a < y < b] = \Pr[a \le y < b] = \Pr[a < y \le b] = \Pr[a \le y \le b]$. The area under the entire curve must be 1, i.e.,

$$1 = \Pr[-\infty < y < \infty] = \int_{-\infty}^{\infty} f(y)\,\mathrm{d}y.$$

Figure 1.1 indicates that the probability below $K(1-\alpha)$ is $1-\alpha$, i.e.,

$$1-\alpha = \Pr[y < K(1-\alpha)] = \int_{-\infty}^{K(1-\alpha)} f(y)\,\mathrm{d}y$$

and that the probability above $K(1-\alpha)$ is α, i.e.,

$$\alpha = \Pr[y > K(1-\alpha)] = \int_{K(1-\alpha)}^{\infty} f(y)\,\mathrm{d}y.$$

The expected value of y is defined as

$$\mathrm{E}(y) = \int_{-\infty}^{\infty} yf(y)\,\mathrm{d}y.$$

For any function $g(y)$, the expected value is

$$\mathrm{E}[g(y)] = \int_{-\infty}^{\infty} g(y)f(y)\,\mathrm{d}y.$$

In particular, if we let $\mathrm{E}(y) = \mu$ and $g(y) = (y-\mu)^2$, we define the variance as

$$\mathrm{Var}(y) = E[(y-\mu)^2] = \int_{-\infty}^{\infty} (y-\mu)^2 f(y)\,\mathrm{d}y.$$

To define the covariance between two random variables, say y_1 and y_2, we need a joint density $f(y_1, y_2)$. We can find the density for y_1 alone as

$$f_1(y_1) = \int_{-\infty}^{\infty} f(y_1, y_2)\,\mathrm{d}y_2$$

and we can write $\mathrm{E}(y_1)$ in two equivalent ways

$$\mathrm{E}(y_1) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y_1 f(y_1, y_2)\,\mathrm{d}y_1\,\mathrm{d}y_2 = \int_{-\infty}^{\infty} y_1 f_1(y_1)\,\mathrm{d}y_1.$$

Writing $\mathrm{E}(y_1) = \mu_1$ and $\mathrm{E}(y_2) = \mu_2$ we can now define the covariance between y_1 and y_2 as

$$\mathrm{Cov}(y_1, y_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (y_1-\mu_1)(y_2-\mu_2) f(y_1, y_2)\,\mathrm{d}y_1\,\mathrm{d}y_2.$$

1.6 Exercises

Exercise 1.6.1. Use the definitions to find the expected value and variance of a $\mathrm{Bin}(1, p)$ distribution.

Exercise 1.6.2. Let y be a random variable with $\mathrm{E}(y) = \mu$ and $\mathrm{Var}(y) = \sigma^2$. Show that

$$\mathrm{E}\left(\frac{y-\mu}{\sigma}\right) = 0$$

and

$$\mathrm{Var}\left(\frac{y-\mu}{\sigma}\right) = 1.$$

Let $\bar{y}_\cdot$ be the sample mean of n independent observations y_i with $\mathrm{E}(y_i) = \mu$ and $\mathrm{Var}(y_i) = \sigma^2$. What is the expected value and variance of

$$\frac{\bar{y}_\cdot - \mu}{\sigma/\sqrt{n}}?$$

Hint: For the first part, write

$$\frac{y-\mu}{\sigma} \quad \text{as} \quad \frac{1}{\sigma}y - \frac{\mu}{\sigma}$$

and use Proposition 1.2.11.

EXERCISE 1.6.3. Let y be the random variable consisting of the number of spots that face up upon rolling a die. Give the distribution of y. Find the expected value, variance, and standard deviation of y.

EXERCISE 1.6.4. Consider your letter grade for this course. Obviously, it is a random phenomenon. Define the 'grade point' random variable: $y(\mathrm{A}) = 4$, $y(\mathrm{B}) = 3$, $y(\mathrm{C}) = 2$, $y(\mathrm{D}) = 1$, $y(\mathrm{F}) = 0$. If you were lucky enough to be taking the course from me, you would find that I am an easy grader. I give 5% As, 10% Bs, 35% Cs, 30% Ds, and 20% Fs. I also assign grades at random, that is to say, my tests generate random scores. Give the distribution of y. Find the expected value, variance, and standard deviation of the grade points a student would earn in my class. (Just in case you hadn't noticed, I'm being sarcastic.)

EXERCISE 1.6.5. Referring to Exercise 1.6.4, suppose I have a class of 40 students, what is the joint distribution for the numbers of students who get each of the five grades? Note that we are no longer looking at how many grade points an individual student might get, we are now counting how many occurrences we observe of various events. What is the distribution for the number of students who get Bs? What is the expected value of the number of students who get Cs? What is the variance and standard deviation of the number of students who get Cs? What is the probability that in a class of 5 students, 1 gets an A, 2 get Cs, 1 gets a D, and 1 fails?

EXERCISE 1.6.6. Graph the function $f(x) = 1$ if $0 < x < 1$ and $f(x) = 0$ otherwise. This is known as the uniform density on $(0, 1)$. If we use this curve to define a probability function, what is the probability of getting an observation larger than $1/4$? Smaller than $2/3$? Between $1/3$ and $7/9$?

EXERCISE 1.6.7. Arthritic ex-football players prefer their laudanum made with Old Pain-Killer Scotch by two to one. If we take a random sample of 5 arthritic ex-football players, what is the distribution of the number who will prefer Old Pain-Killer? What is the probability that only 2 of the ex-players will prefer Old Pain-Killer? What is the expected number who will prefer Old Pain-Killer? What are the variance and standard deviation of the number who will prefer Old Pain-Killer?

EXERCISE 1.6.8. Let $W \sim \mathrm{Bin}(N, p)$ and for $i = 1, \ldots, N$ take independent y_is that are $\mathrm{Bin}(1, p)$. Argue that W has the same distribution as $y_1 + \cdots + y_N$. Use this fact, along with Exercise 1.6.1 and Proposition 1.2.11, to find $\mathrm{E}(W)$ and $\mathrm{Var}(W)$.

EXERCISE 1.6.9. Appendix B.1 gives probabilities for a family of distributions that all look roughly like Figure 1.1. All members of the family are symmetric about zero and the

members are distinguished by having different numbers of degrees of freedom (df). They are called t distributions. For $0 \leq \alpha \leq 1$, the α percentile of a t distribution with df degrees of freedom is the point x such that $\Pr[t(df) \leq x] = \alpha$. For example, from Table B.1 the row corresponding to $df = 10$ and the column for the .90 percentile tells us that $\Pr[t(10) \leq 1.372] = .90$.

(a) Find the .99 percentile of a $t(7)$ distribution.

(b) Find the .975 percentile of a $t(50)$ distribution.

(c) Find the probability that a $t(25)$ is less than or equal to 3.450.

(d) Find the probability that a $t(100)$ is less than or equal to 2.626.

(e) Find the probability that a $t(16)$ is greater than 2.92.

(f) Find the probability that a $t(40)$ is greater than 1.684.

(g) Recalling that t distributions are symmetric about zero, what is the probability that a $t(40)$ distribution is less than -1.684?

(h) What is the probability that a $t(40)$ distribution is between -1.684 and 1.684?

(i) What is the probability that a $t(25)$ distribution is less than -3.450?

(j) What is the probability that a $t(25)$ distribution is between -3.450 and 3.450?

Exercise 1.6.10. Consider a random variable that takes on the values 25, 30, 45, and 50 with probabilities .15, .25, .35, and .25, respectively. Find the expected value, variance, and standard deviation of this random variable.

Exercise 1.6.11. Consider three independent random variables X, Y, and Z. Suppose $\mathrm{E}(X) = 25$, $\mathrm{E}(Y) = 40$, and $\mathrm{E}(Z) = 55$ with $\mathrm{Var}(X) = 4$, $\mathrm{Var}(Y) = 9$, and $\mathrm{Var}(Z) = 25$.

(a) Find $\mathrm{E}(2X + 3Y + 10)$ and $\mathrm{Var}(2X + 3Y + 10)$.

(b) Find $\mathrm{E}(2X + 3Y + Z + 10)$ and $\mathrm{Var}(2X + 3Y + Z + 10)$.

Exercise 1.6.12. As of 1994, Duke University had been in the final four of the NCAA's national basketball championship tournament seven times in nine years. Suppose their appearances were independent and that they had a probability of .25 for winning the tournament in each of those years.

(a) What is the probability that Duke would win two national championships in those seven appearances?

(b) What is the probability that Duke would win three national championships in those seven appearances?

(c) What is the expected number of Duke championships in those seven appearances?

(d) What is the variance of the number of Duke championships in those seven appearances?

Exercise 1.6.13. Graph the function $f(x) = 2x$ if $0 < x < 1$ and $f(x) = 0$ otherwise. If we use this curve to define a probability function, what is the probability of getting an observation larger than 1/4? Smaller than 2/3? Between 1/3 and 7/9?

Exercise 1.6.14. A pizza parlor makes small, medium, and large pizzas. Over the years they make 20% small pizzas, 35% medium pizzas, and 45% large pizzas. On a given Tuesday night they were asked to make only 10 pizzas. If the orders were independent and representative of the long-term percentages, what is the probability that the orders would be for four small, three medium, and three large pizzas. On such a night, what is the expected number of large pizzas to be ordered and what is the expected number of small pizzas to be ordered? What is the variance of the number of large pizzas to be ordered and what is the variance of the number of medium pizzas to be ordered?

Exercise 1.6.15. When I order a limo, 65% of the time the driver is male. Assuming independence, what is the probability that 6 of my next 8 drivers are male? What is the expected number of male drivers among my next eight? What is the variance of the number of male drivers among my next eight?

Exercise 1.6.16. When I order a limo, 65% of the time the driver is clearly male, 30% of the time the driver is clearly female, and 5% of the time the gender of the driver is indeterminant. Assuming independence, what is the probability that among my next 8 drivers 5 are clearly male and 3 are clearly female? What is the expected number of indeterminant drivers among my next eight? What is the variance of the number of clearly female drivers among my next eight?