Chapter 5
Probability Theory

In the last chapter, we discussed the first goal of statistics: summarization. Summarization of data is an incredibly important aspect of statistics, but our goal in summarizing sample data is not usually to simply to report the characteristics of a sample. Instead, we are usually interested in using our sample to draw some conclusions about the population from which the sample was drawn, *and to place limits on the conclusions we can reach*. This is the role of statistical inference. For example, suppose I had a sample of 50 persons drawn at random from the population and had measured the heights and weights of all of the sample members. Suppose that the mean height was 70 in., with a standard deviation of 5 in. As we discussed in the previous chapter, if the sample is random, then our sample mean ($\bar{x}$) and standard deviation (s) may be a good guess about the population mean (μ) and standard deviation (σ). However, it is unreasonable to expect that this sample mean would be a perfect reflection of the average height in the population, because we may have a few people in our sample who were unusually tall (or short). In other words, every sample we draw from a population is likely to have slightly different means and standard deviations. The goal of statistical inference in this example would be to attempt to quantify our uncertainty about the true mean and standard deviation in the population, given the sample data that we have. Thus, we might end-up with a statement like: we are 95 % confident that the mean height in the population is 70 in., give or take an inch. In common language, this "give or take an inch" is called the "margin of error," as we will discuss in more detail in the next chapter.

Statistical inference relies on probability; in fact, inference simply inverts probabilistic reasoning. As we will discuss, probabilistic reasoning involves knowing something about a population and using that information to *deduce* characteristics of samples. Inference involves knowing something about a sample and using that information to *induce* (i.e., infer) something about the population. For example, if we know a population mean, we can use probability theory to determine the most likely values for the means of samples drawn from that population. But what if we

S.M. Lynch, *Using Statistics in Social Research: A Concise Approach*,
DOI 10.1007/978-1-4614-8573-5_5, © Springer Science+Business Media New York 2013

have a sample mean? We can "reverse" the probabilistic approach and infer what the population mean is most likely to be using statistical theory. This process forms the basis for testing hypotheses in statistics.

5.1 Probability Rules

When we refer to probability, we are generally talking about the chance of some event occurring in a trial or experiment. This definition of probability is insufficient, because it simply begs the question: isn't "chance" defined in terms of probability? One way of conceptualizing what we mean by probability is to imagine an "experiment" like flipping a (fair) coin. When we flip a coin, there are two possible, and equally likely, outcomes of the experiment (heads and tails), and the chance of obtaining a head is the ratio of the number of possible ways to obtain a successful outcome (a head) to the total number of possible outcomes. It is common knowledge that, with a fair coin, the probability of obtaining a head on a single flip is $1/2$. This conclusion is reached because (1) there are two possible (equally likely) outcomes (heads and tails), and (2) heads constitutes one of them. So, there is a one-out-of-two chance of obtaining heads on a given flip. Similarly, with a single roll of a six-sided die, there is a one-out-of-six chance of obtaining a three. It is not always the case that all outcomes are equally likely, as in these examples, but we will discuss this momentarily.

Given this basic view of probability, it is clear that probability involves (1) counting the number of ways a "success" can be obtained, (2) counting the total number of possible events that can occur in a given "trial," and (3) forming a ratio of the two. From this basic conceptualization, several terms can be defined and rules derived:

1. The collection of all of the possible events $E_1 \ldots E_n$ that can occur in a given trial is called the "sample space," which is denoted as S. We denote the probability of event i occurring as: $p(E_i)$.

   Regarding coins, for example, the sample space is $S = \{\text{Heads}, \text{Tails}\}$. Some measure of the size of the sample space forms the denominator of our ratio indicating probability. Here, the size of the sample space is 2 equally likely events.
2. Probabilities are bounded between 0 and 1. An event that will definitely occur has a probability of 1, and an event that will definitely not occur has a probability of 0.

   Obviously, if a probability is a ratio of the count of ways a success can be obtained out of a given number of possible outcomes, the ratio cannot be larger than 1: there cannot be more events labeled as "successes" than there are total

events possible. Also, if an event cannot occur in a trial, then the implication is that the event does not exist in the numerator. Thus the probability of such an event occurring is 0.

3. The sum of the probabilities of all the events in a sample space must be 1 ($\sum_{E_i \in S} p(E_i) = 1$).

   This rule should be intuitive: If the sample space consists of all possible events that can occur in a trial, one of the events *must* occur if we go through with the trial. So, the probability that one of the events occurs is 1.

4. The probability that two events A and B both occur—a "joint probability"—is represented as $p(A, B)$ and equals $p(A) \times p(B)$ if A and B are *independent,* that is, if the occurrence of A has no bearing on the occurrence of B. For example, if I flip two coins, whether I get heads on one coin has nothing to do with obtaining a heads on the other. Thus, the probability of obtaining two heads is the product of the probability of obtaining heads on each:

$$p(H, H) = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{1}{4}. \tag{5.1}$$

5. When two events are not independent, $p(A, B) = p(A \mid B)p(B)$, where "$p(A \mid B)$" is the *conditional probability* of A, given that we know B has occurred.

   This rule is often rearranged algebraically to appear as:

$$p(A \mid B) = \frac{p(A, B)}{p(B)}. \tag{5.2}$$

   In this arrangement, the equation says that the probability that an event A will occur, given that we know B has occurred, is the probability that both events will occur, divided by the total probability that event B would occur. This division essentially reduces the sample space for A to the sample space that only includes B, which we already know to have occurred.

6. The probability that, of two events A and B, at least one will occur (the "union" of two events), is $p(A \cup B) = p(A) + p(B) - p(A, B)$.

The Venn diagram in Fig. 5.1 helps us understand some of these rules. For example, why do we subtract $p(A, B)$ in the last rule? We do so because it is added twice when we sum the probability of A and B. So, based on the diagram, the probability of being male is .5 (the entire lefthand circle), and the probability of being obese is .3 (the entire righthand circle). The probability of being in either circle (male *or* obese) is not .8, because that would double count the overlap region of the two circles. Instead: $p(M \ or \ O) = .5 + .3 - .1 = .7$.

In terms of the joint probability rule, we can rearrange the rule for non-independent events ($p(A, B) = p(A \mid B)p(B)$) as a conditional probability rule:

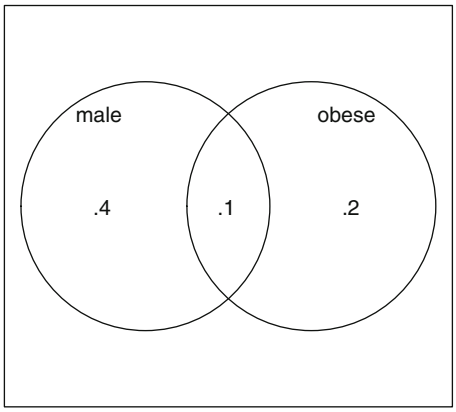

Fig. 5.1. Sample Venn diagram

$p(A \mid B) = p(A, B)/p(B)$. The figure shows why this calculation works: if we already know B has occurred/is true (say, we have been told that the person is male), then the total sample space has been reduced from the overall rectangle in the figure to the circle for being male.

As an example, consider the Venn diagram's representation of the probabilities for being male and being obese. The probability of being male is $.4 + .1 = .5$. The probability of being obese is $.2 + .1 = .3$. The probability of being an obese male ($p(\text{male, obese})$) is .1 (the overlap region). The probability of being obese if a person is male (conditional on being male) is:

$$p(O|M) = \frac{p(M, O)}{p(M)} \tag{5.3}$$

$$= \frac{.1}{.5} \tag{5.4}$$

$$= .2. \tag{5.5}$$

In words, under the conditional formulation, we already know that the person is male, and so the probability that he is also obese is the ratio of the probability for being obese that is also within the male circle.

What is the probability of being a nonobese female? Given the rules above, we know that the total sample space must sum to 1. Thus, the portion of the diagram that lies outside the circles is .3 What does this area represent? It is the proportion that is neither male nor obese.

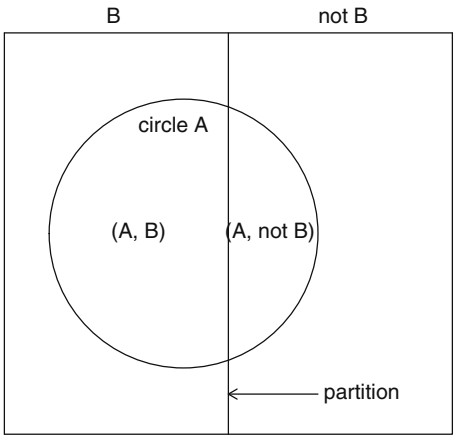

Fig. 5.2. Heuristic for the law of total probability.

5.1.1 Total Probability and Bayes' Theorem

The rules above can be used to derive more complex, important rules, like the law of total probability and Bayes' Theorem, both of which are commonly used in probability calculations. Consider Fig. 5.2, a Venn diagram with a "partition" breaking the sample space into two components—the part of the world that is B and the part that is not B ($\neg B$). If we are interested in the total probability of event A, we can use the rule for unions of events ("or"), coupled with the rule for joint probabilities of nonindependent events to compute this total probability:

$$p(A) = p(A, B) + p(A, \neg B) \tag{5.6}$$

$$= p(A|B)p(B) + p(A|\neg B)p(\neg B) \tag{5.7}$$

The first equation shows that the total probability of A is the sum of the joint probabilities of A with B and A with $\neg B$. Under the usual rule for unions, we should subtract out the joint probability of the events on both sides of the plus sign; however, here, these two events are "disjoint." That is, there is no joint probability of being in state B and $\neg B$. In words, the total probability of A is the probability of A to the left of the partition plus the probability of A on the right side of the partition.

The second equation expands the joint probability using the conditional probability rule. This equation says that the total probability of A is the probability of A occurring, given that the world is in state B, times (or weighted by) the probability the world is in state B, plus the probability of A occurring, given that the world is in state $\neg B$, times the probability the world is in state $\neg B$.

As an example of the law of total probability, suppose the weather forecast says there is a 30 % chance for rain today. I decide that, if it rains, there is an 80 % chance

I will skip class, because I don't like to walk to class in the rain. On the other hand, if it does not rain, there is a 40 % chance I will skip class, because I hate to sit in class when the weather is nice.

All things considered, am I more likely to attend class or skip it? The probability that it will rain is .3. Under that state of the world, there is a probability of .8 that I will not go to class. The probability that it will not rain is .7. Under that state of the world, there is a probability of .4 that I will not go to class. Thus, under the law of total probability:

$$p(skip) = p(skip|rain)p(rain) + p(skip|sun)p(sun) \tag{5.8}$$

$$= (.8)(.3) + (.4)(.7) \tag{5.9}$$

$$= .52 \tag{5.10}$$

All things considered, I am slightly more likely to skip class than to attend it.

The law of total probability can be extended to more than two states of the world (say n states) as follows:

$$p(A) = \sum_{i=1}^{n} p(A|B_i)p(B_i). \tag{5.11}$$

The only requirement is that $\sum B_i = 1$; that is, the probability of being in *some* state must be 1 (all the states of the world must be covered).

Bayes' Theorem uses the law of total probability and allows us to reverse conditional probabilities. What does it mean to "reverse a conditional probability?" Suppose, for example, we go to the doctor's office for a blood test for some rare disease X (occuring in 1 out of 10,000 people), and we test positive. Virtually all tests have false positive rates and false negative rates; in other words, medical tests are not infallible. So, we would like to know, in this circumstance, what our probability is for having X conditional on the positive test result. We may look online and find that the test has a 10 % false positive rate and a 10 % false negative rate. The false positive rate means that, among those who do not have the disease, 10 % will receive positive test results. The false negative rate means that, among those with the disease, 10 % will test negative. The converse of the false negative rate is that 90 % of those with the disease will test positive.

So, we know p(test +|have disease). We are interested in p(have disease|test +). We therefore need a way to reverse the known conditional in order to obtain the probability of interest. Bayes' Theorem supplies the recipe:

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)}. \tag{5.12}$$

Term	Meaning	Value
$p(A\|B)$	prob. of testing + if have disease	.9
$p(B)$	prob. of having disease	.0001
$p(A\|\neg B)$	false + (test + but no disease)	.1
$p(\neg B)$	prob. of not having disease	.9999

Table 5.1. Elements of Bayes' Theorem in disease example.

In this equation, $p(A)$ is the total probability of A, which can be found using the law of total probability:

$$p(A) = p(A|B)p(B) + p(A|\neg B)p(\neg B) \tag{5.13}$$

(or its extended version).

The theorem is easily proven by multiplying both sides of the equation by $p(A)$ and recognizing (1) the lefthand side is the joint probability $p(B, A)$, (2) the righthand side is the joint probability $p(A, B)$, and (3) these two joint probabilities are the same, only written in reverse.

Returning to the disease example, if B is "has disease", and A is "tests positive," then $\neg B$ is "does not have disease". We have all the information we need to compute $p(B|A)$ as shown in Table 5.1. Thus:

$$p(B|A) = \frac{(.90)(.0001)}{(.90)(.0001) + (.10)(.9999)} \tag{5.14}$$

$$= .0009. \tag{5.15}$$

Based on these results, even after the positive test result, the probability of having the disease (the "posterior probability") is miniscule. Why? The posterior probability is heavily influenced by two factors: (1) the marginal, or "prior" probability of having the disease, not knowing anything else (.0001), and (2) the high false positive rate. It is most likely the case that an individual does not have the disease in general, and there's a reasonably large probability (.1) under that probability that one will obtain a positive test result anyway. Put another way, we know we obtained a positive result. It is much more likely to have occured because of the false positive rate than because of the disease, given how rare the disease is.

5.2 How to Count

As noted earlier, an important part of computing probabilities is being able to count successes and the size of sample spaces. Counting seems a very basic task, but there are special ways to count in probability that often require careful thought in

order to get the numerator (successes) and denominator (sample spaces) correct for computing probabilities. When trying to obtain the total number of events that may occur in a sample space, we often rely on permutation and combination calculations. We discussed combinations briefly before when discussing how to obtain a given number of colored marbles out of a box.

The key distinction between permutations and combinations is whether the order in which events occurs matters. For combinations, order does not matter; for permutations, order does matter. For example, if we have 10 persons' names in a hat and are putting together a committee of three persons drawn at random, the committee with Scott, Chris, and Brian as members would be the same committee as the one with Chris, Brian, and Scott as members. In that case, the order of selection does not matter. In contrast, suppose we were arranging these 10 people on one side of a long table. Determining how many different ways we could arrange the persons obviously involves considering the order in which individuals are placed. In that context, that's the whole point!

As we discussed earlier, the combination formula yields the number of unique sets of x items that can be drawn from a collection of n items. The calculation, again, is:

$$C(n, x) = \frac{n!}{x!(n-x)!}. \tag{5.16}$$

Let's discuss this calculation in some detail in order to differentiate how to compute combinations and permutations. The numerator of the calculation tells us how many ways n items can be arranged, ultimately into n positions. In the seating arrangement example, where order matters, determining the total number of arrangements of n people requires us to decide who will take the first seat, followed by who will take the next seat, and so on. At first, there are n persons who could take the first seat. After that person is chosen, there are $n - 1$ persons who could take the next seat. After that person is chosen, there are $n - 2$ persons who could take the next seat, and so on. Thus, there are $n!$ arrangements of the n people into the n positions at the table. This is what the numerator of the combination formula "does," and it is a very basic permutation calculation.

The denominator of the combination calculation factors out the order of (1) the persons selected out of the n persons available, and (2) the persons not selected out of the n persons.

To make this idea concrete, we must expand the example slightly. Suppose now that our table has only $x = 4$ seats. We still have $n = 10$ persons, but we now have to choose who will be selected to sit at the table and we still care about the arrangement, but only of those who are selected. In that case, we have n persons we can choose for the first seat, $n - 1$ for the second, $n - 2$ for the third, and $n - 3$ for the fourth. Thus, the number of possible ways to arrange 4 people out of 10 total is:

$$n \times (n-1) \times (n-2) \times (n-3). \tag{5.17}$$

An alternative, generic way to write this is:

$$n \times (n-1) \times (n-2) \times \ldots \times \ldots (n-x+1) = \frac{n!}{(n-x)!}. \tag{5.18}$$

In this equation, $(n-x)!$ cancels the remainder of the n persons we are not selecting. Put another way, it is factoring out of the numerator the order of the persons who are not selected to be seated at the table. Thus, another permutation calculation—one in which we are ordering x persons selected out of n persons (and discarding the rest)—denoted as $P(n, x)$ is:

$$P(n, x) = \frac{n!}{(n-x)!} \tag{5.19}$$

Let's extend this scenario one more step. Suppose now that the order in which we place the selected persons at the table does not matter; only selection vs. nonselection matters. In that case, we need to factor out the arrangement of the x persons who end up seated at the table. This involves simply reducing the problem to what remains: how many ways can we arrange the x people? Obviously, if there are $n!$ ways to arrange n people, then there are $x!$ ways to arrange x people! Thus, if we are selecting x people out of n people, and arrangement of them does not matter, we are left with the original combination formula.

All of these calculation formulas assume *sampling without replacement*. That is, once a person/object is chosen, it is no longer possible to select it again. In extremely large—or infinite—populations, combination and permutation formulas are often useless, because sampling from an infinite population is akin to sampling with replacement. However, in situations involving finite populations, these formulas are sufficient for most, if not all, problems.

To differentiate these two types of populations—finite vs. infinite—consider computing the probability of obtaining a four digit "pick 4" lottery number. In such a lottery, one pays a specified fee in order to select a four digit number, with the digits each ranging between 0 and 9. The winning lottery number is then selected at random. In a typical lottery hopper, there are four chambers, each filled with 10 balls labeled with the digits 0–9. A fan circulates the balls in each chamber, and the digit painted on the first ball that rises to the top is that position's digit in the winning number. Under that approach, there are 10 possibilities for the first digit, 10 for the second, 10 for the third, and 10 for the fourth. Given that these are independent selections, there are $10^4 = 10,000$ possible outcomes, ranging from 0000 to 9999. Thus, the probability of any given four digit number is $1/10,000$. The numbers are considered to be selected with replacement, because, even if a 1 is selected as the first digit, this selection does not affect the probability of obtaining a 1 as the second digit. It is as if there are an infinite number of 1's in the population (or as if the 1 had been replaced before the second number was picked), so the selection of a 1 on the first draw has no impact on the probability of the selection of a 1 on subsequent draws.

An alternative approach to this lottery would be that the digits are selected without replacement: that the chamber with the digits contains the digits 0–9,

and that once a digit is selected, it cannot be used again. Under that scenario, there would be $P(10, 4)$ possible outcomes. There are 10 possibilities for the first number (0–9), 9 for the second (whatever is left), and so on. In total, there are $10 \times 9 \times 8 \times 7 = 5{,}040$ possible outcomes instead of 10,000. Thus, if you pick a number that has nonrepeating digits, there is a 1/5,040 probability of winning in this type of lottery.

It is difficult for many people to understand that the chance, under the first scenario (sampling with replacement), of obtaining 0000 is the same as the chance of obtaining 1,234 or 5,972. This may be the result of the fact that there is a larger probability of obtaining a set of digits that do not seem to follow a pattern than the probability of obtaining a set of digits that do. In particular, clearly, the sampling-without-replacement options constitute more than half of the total possible outcomes in the sample space of sampling with replacement. In other words, of the total 10,000 possible outcomes under sampling with replacement—the largest possible set of outcomes—more than half of these outcomes (5,040) involves non-redundant digits. In contrast, there are only 10 outcomes that involve sequences of four identical digits. Thus, *as a set*, it is less likely to draw a patterned number than a non-patterned (or less obviously patterned) number, but *each* sequence itself has the exact same probability of being selected in a single lottery.

When we attempt to determine the probability for an event (or sets of events), we must decide whether counting the events in the sample space, as well as the ways successes can be obtained, requires us to consider the order in which events occur, and we must keep consistent in both numerator and denominator calculations.

Often, counting can be done either way, but the difference involves recognizing that, if order is not taken into account but should have been, that all events in the sample space may not be equally likely. For example, in a classic mistake, a famous mathematician computed the probability that one would obtain two heads on two coin flips as 1/3. His reasoning was that there are three options when flipping two coins: two heads, a head and a tail, or two tails. Yet this conclusion is based on a miscount in the denominator. When one flips two coins, there are four *equally likely* possible outcomes if order is taken into consideration: head-head, head-tail, tail-head, and tail-tail. In other words, the order matters in this situation. If one wishes to ignore order, then one has to recognize that the probability of obtaining a head and a tail is actually 2/4.

Let's consider an example involving rolling a pair of dice. Each die has from one to six dots ("pips") on its face. If we roll the pair of dice, what is the probability of obtaining a sum of seven pips between the two dice? There are two ways to answer this question. Under one approach, we would determine the complete sample space of rolls, where order matters. Following that approach, given that there are 6 possible outcomes on the first die, and 6 possible outcomes on the second die, there are 36 possible rolls. The sample space looks like:

$$S = \{(1,1), (1,2), \ldots, (1,6), (2,1), \ldots (2,6), \ldots, (6,1), \ldots, (6,6)\}. \quad (5.20)$$

The number of pairs in this set constitutes the denominator of our probability calculation. What remains is to count the number of ways a success—that is, a sum of 7—can occur from these 36 outcomes. That set looks like:

$$S_7 = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}. \tag{5.21}$$

There are six pairs in this set. Thus, the probability of rolling a sum of 7 on a pair of dice is $6/36 = 1/6$.

The second method of solving this problem involves recognizing that the order of the dice does not matter; for all intents and purposes, they are interchangeable. Thus, we could consider the sample space to be the possible sums that can arise. The sample space represented in this fashion would be:

$$S = \{2,3,4,5,6,7,8,9,10,11,12\}. \tag{5.22}$$

Under this approach, in finding the probability that we would roll a sum of 7, we must take into account the fact that these sums *are not equally likely to occur*. For example, there is only one way to roll a sum of 2: both dice must come up ones. Three are two ways to roll a three: a 1 followed by a 2 or a 2 followed by a 1. There are three ways to roll a four: a 1 followed by a 3, two twos, or a 3 followed by a one. And so on. Overall, computing the probability of rolling a sum of 7, then, involves mapping the possible events in the sample space onto a set of probabilities associated with those events and then summing up the relevant probabilities. The set of probabilities associated with the events in this sample space is:

$$p(E) = \left\{\frac{1}{36}, \frac{1}{18}, \frac{1}{12}, \frac{1}{9}, \frac{5}{36}, \frac{1}{6}, \frac{5}{36}, \frac{1}{9}, \frac{1}{12}, \frac{1}{18}, \frac{1}{36}\right\} \tag{5.23}$$

To some extent, even under this approach, in order to determine the probabilities of particular sums, we had to return to considering the order of rolls, but sometimes we can develop calculations (or others already have) that simplify the process. We now discuss such cases.

5.3 Probability Density/Mass Functions

When sample spaces are large, we can use algebraic functions to assign probabilities to events within the sample space. These functions must obey the rules of probability outlined above, including that all events have probabilities between 0 and 1 and that the sum of the probabilities of all events must be 1. However, aside from those key rules, the form of such functions is quite flexible. Two common such functions—called probability density functions (pdfs) if the sample space is continuous and probability mass functions if the sample space is discrete (like counting numbers;

integers)—are the binomial mass function and the normal density function. These two probability distributions are the most common ones used in statistics, and so some in-depth discussion of them is warranted.

5.3.1 Binomial Mass Function

The sample space for a single coin flip is pretty simple, consisting of only two outcomes, and so computing the probability of the two outcomes is easy. But, what if we were interested in knowing the probability of obtaining, say, 5 heads in 10 coin flips? In that case, the sample space is considerably larger. The possible outcomes (counts of heads) in 10 flips are: $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. Furthermore, this is a case in which the probabilities of the outcomes are not equal. For example, there is only one possible way to obtain 0 heads in 10 tosses: Every toss must be a tail. But, consider the possible ways one can obtain 1 head. The head could occur on the first toss, the second toss, the third toss, and so on. The binomial mass function allows for straightforward computation of such probabilities. Its mass function is:

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}. \tag{5.24}$$

The random variable (the count of successes; the quantity that is random) in this function is x, while n and p are the "parameters" of the distribution. n is the number of trials, and p is the probability of success on any given trial. In mathematical shorthand, we say: $x \sim Binomial(n, p)$ ("x is distributed binomially with parameters n and p"). The sample space for x is determined by n. x must be between 0 and n, and x is restricted to nonnegative integers. Given a value for n and p, the probabilities for x can be computed. For example, if we wanted to know the probability of obtaining 3 heads on 10 flips of a fair coin, we would solve $p(x = 3) = \binom{10}{3}.5^3(1-.5)^{10-3} \approx .12$. If we want to know the probability of obtaining a range of successes (e.g., three or fewer successes, six or more successes, etc.), we simply compute the probabilities for all relevant events and sum them.

The binomial mass function may appear somewhat complicated at first, but it involves the basic ideas of counting outlined in the previous section and the probability rule for independent events. Suppose we want to know the probability of obtaining 3 heads in a row. This computation would simply involve multiplying the probability of obtaining a head on each flip: $(.5)(.5)(.5) = .125$. Suppose instead, we want to know the probability of obtaining 2 heads and 1 tail. The probability of getting two heads on two flips is .25. The probability of getting a tail on a single flip is .5. It may seem as if we could simply multiply .25 by .5 and obtain the final answer. However, doing so ignores that the tail may come in any position in the three flips. It could be first, second, or third. In other words, there are three ways

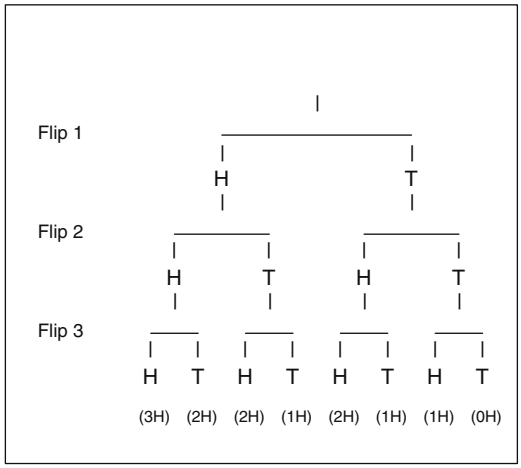

Fig. 5.3. Tree diagram showing outcomes of three coin flips

the heads and tail could be ordered. The combinatorial at the beginning of the mass function accounts for these different arrangements.

Figure 5.3 shows a tree diagram that illustrates this point. The figure shows the possible sequences of outcomes of three successive flips. The total number of heads on the three flips for each possible branch of the tree are shown in parentheses at the bottom. As the figure indicates, there is only one branch that produces three heads (similarly, there is only one branch that produces three tails—0 heads). There are three branches that produce 2 heads and three branches that produce 1 head. In this scenario, each outcome on a given trial is equally likely. Thus, each branch is equally likely, and so the probability that we will obtain two heads and a tail is $3/8$. In the event that the outcomes on a given trial are not equally likely (i.e., p is not .5), we can insert the success/failure probabilities in the branches as appropriate and multiply them along a branch to obtain the probability of a particular branch. However, this is what the righthand side of the binomial mass function does for us.

Figure 5.4 shows three examples of binomial distributions. The upper left distribution represents the probability of different counts of successes out of 10 trials when the success probability on any given trial is .5. The upper right distribution shows the probability when $p = .8$, and the bottom left distribution shows the probability when $p = .2$.

5.3.1.1 Pascal's Triangle

The combinatorial expression at the beginning of the binomial mass function is sometimes called the "binomial coefficient." For small values of n, it may be easier to use Pascal's Triangle to determine how many different ways a given count of successes can be obtained than to use the combinatorial formula. Pascal's triangle

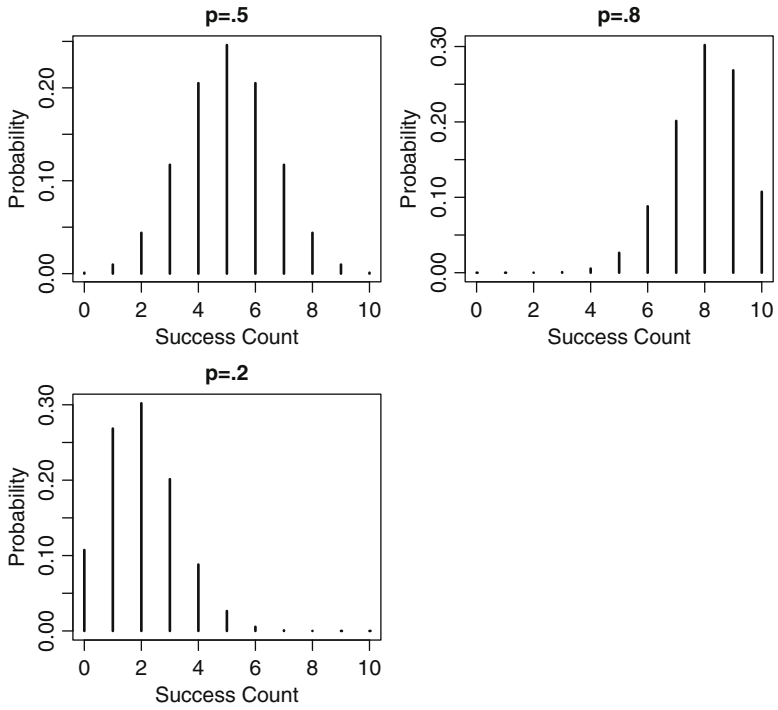

Fig. 5.4. Some binomial distributions ($n = 10$)

Trials	Ways to Obtain Sucess Counts										
0						1					
1					1		1				
2				1		2		1			
3			1		3		3		1		
4		1		4		6		4		1	
5	1		5		10		10		5		1

Fig. 5.5. Pascal's triangle. Each row shows the number of ways to obtain x successes out of n trials for different values of x, where x increases from left to right from 0 to n

is a collection of numbers with ones along the left and right edges of the triangle and sums filling in the rest of the triangle. The number at each location in a row in the triangle is simply the sum of the two numbers above it in the previous row. Figure 5.5 shows the first six rows of the triangle. To interpret the triangle, consider the last row in the figure. This row tells us how many ways there are to obtain different counts of successes (x) in a binomial distribution with an n parameter of 5 trials. Thus, there is 1 way to obtain 0 successes, 5 ways to obtain 1 success, 10 ways to obtain 2 successes, 10 ways to obtain 3 successes, 5 ways to obtain 4 successes, and 1 way to obtain 5 successes.

Combination	Equivalent	Equals
$\binom{n}{x}$	$\binom{n}{(n-x)}$	$\frac{n!}{x!(n-x)!}$
$\binom{n}{n}$	$\binom{n}{0}$	1
$\binom{n}{(n-1)}$	$\binom{n}{1}$	n

Table 5.2. Some helpful shortcuts for calculating combinations.

Pascal's triangle can be a useful shortcut to determining counts of combinations for small n. If n is large, it may take longer to draw the triangle than to simply compute the combinatorial. For large n, Table 5.2 shows some helpful shortcuts for calculating combinations.

5.3.2 Normal Density Function

Many phenomena in the natural (and social) world are distributed so that the majority cluster around some "middle" value, with more extreme cases occurring less frequently, with frequency declining rapidly with distance from the center. Consider, for example, the sample histogram for education in the previous chapter: most values of education were clustered around 12 years of schooling, with fewer persons having many more or many fewer years. Despite the slight multimodality and the skew, the distribution of schooling followed this general pattern. Many phenomena match the pattern more closely, like the distributions for height and weight in the population.

The normal distribution (also called the bell curve) represents this pattern. The normal density function is:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}. \tag{5.25}$$

The random variable in this function is x, while μ and σ^2 are the parameters (the mean and variance, respectively), and we say $x \sim N(\mu, \sigma^2)$ ("x is distributed normally with a mean of μ and variance of σ^2"). $\exp\{c\}$ is simply an alternative way to write e^c, with e being the base of the natural logarithms—the exponential function.

Although this density function looks complicated, it simply defines a bell-shaped curve with tails that asymptote to 0 (i.e., they diminish toward 0 the further one moves from the center, but they never reach 0). Recall from algebra that a general formula for a parabola is:

$$y = a(x-h)^2 + k. \tag{5.26}$$

The point (h, k) is the vertex of the parabola, and a determines whether the parabola is narrow or wide. The interior of the "kernel" of the normal distribution (the part inside the exponential function) is simply a parabola. The x coordinate of the vertex (technically, the "abscissa") is μ, and $-1/2\sigma^2$ is a. The negative sign flips the parabola so that it opens downward. The exponential function wrapped around this downward-facing parabola bends the tails of the parabola outward (toward $+\infty$ and $-\infty$), because the exponential function always produces a non-negative result. Consequently, the bell curve sits entirely above the x-axis. The expression in front of the exponential function determines the height of the inverted parabola. If you are familiar with the exponential function, you may recall that $e^a e^b = e^{a+b}$. Thus, the leading term $1/\sqrt{2\pi\sigma^2}$ can be logged and placed within the exponential function. From there, it is easy to see that term becomes k—the y coordinate of the vertex—in the general formula for a parabola.

Notice that the left side of this function is written as $f(x)$ rather than $p(x)$ as in the binomial distribution; the reason for this is that the normal distribution is a continuous distribution, and thus the probability of any particular value for x is 0. In a continuous distribution, the denominator of the usual probability ratio (successes over sample space) is infinitely large: there are an infinite number of real numbers between any two values in the sample space. As a consequence, we cannot determine probabilities by simply computing one application of the function. Instead, we must determine probabilities for ranges of x using integral calculus. For given values of μ and σ^2, we can compute probabilities for x falling in some desired range. The domain of x is unrestricted; x can take any real value.

The normal distribution is the most important distribution in statistics, and much of our discussion regarding statistics will focus on this distribution. Although you need not memorize the density function, it is important to know a few things about this distribution. First, the mean, median and mode of the distribution are equal (μ). Second, the distribution is perfectly symmetric around the mean, so that $p(-x < -z) = p(x > z)$. In English: the probability that x falls above some value z is the same as the probability that $-x$ falls below $-z$. Third, the width of the distribution is governed by the variance parameter σ^2. Larger values of σ^2 imply a wider and shorter distribution than smaller values of σ^2. Figure 5.6 shows three normal distributions with different means and variances.

If a variable is normally distributed in the population—that is, its histogram follows a curve like those shown in Fig. 5.6—and we know μ and σ^2, then we can determine the probability of obtaining x values in any range. In order to find probabilities, we standardize our desired range for x and look up the appropriate probabilities in a "z table" like the one in Appendix A. To standardize x we simply subtract off the mean and divide by the standard deviation:

$$z = \frac{x - \mu}{\sigma}. \tag{5.27}$$

This process gives us a value z, the distribution for which has a mean of 0 and standard deviation of 1 (called the standard normal distribution). If you consider

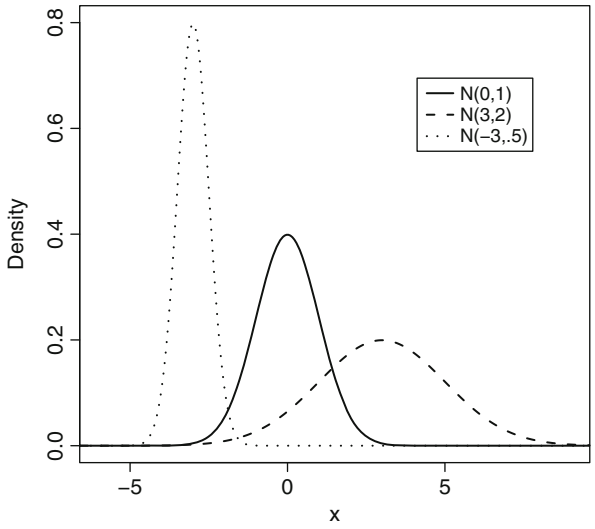

Fig. 5.6. Some normal distributions

this for a moment, z is really just a count of the number of standard deviations away from the mean that a given value of x is. Some standard probabilities from the z distribution are well-known. For example, we know (and you should memorize this) that approximately 68 % of the mass of a normal distribution falls within 1 standard deviation of the mean, 90 % falls within 1.645 standard deviations, 95 % falls within 1.96 standard deviations, and 99 % falls within 2.58 standard deviations (see Fig. 5.7). For simplicity (but not exactness), we often round the standard deviations and probabilities and use a "1, 2, 3" rule of thumb: 68 % of the mass is within 1 standard deviation, 95 % is within 2, and almost 100 % is within 3 (99.7 %). We will follow this convention throughout the remainder of the book as a matter of convenience.

For example, if I claimed that IQs were normally distributed in the population with a mean of 100 and a standard deviation of 15, then I could conclude that only about 2.5 % of the population have IQs above 130. How? If 95 % of the mass of the distribution falls within two (1.96) standard deviations of the mean, and I know that the distribution is symmetric, then there is only 2.5 % of the mass of the distribution beyond two standard deviations on either end of the distribution.

Often we are given a value of x that, when transformed to z scale, is not one of these values for which the probability is immediately known. In those cases, we simply look the z up in a z table and find the associated probability. Doing so may be a tedious process, especially given that most z tables—in order to save space—only provide probabilities for one half of the distribution. In this book, the z table provides only the probability an observation falls below a value of $-Z$.

As an example of using the z table, suppose I select a person at random from the population and want to know the probability that person has an IQ greater than 125.

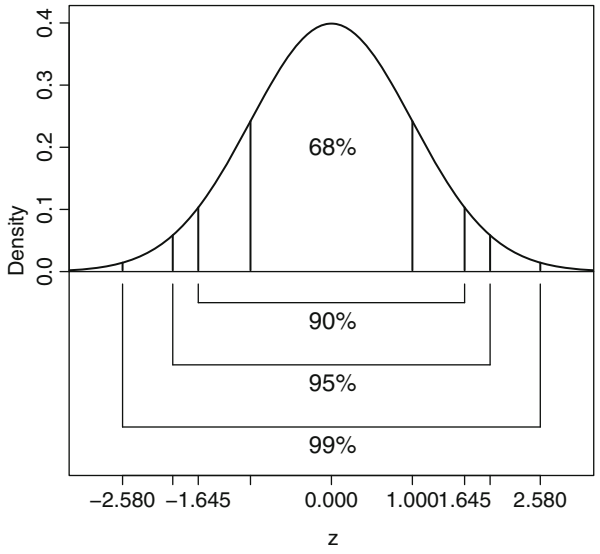

Fig. 5.7. Some common areas under the normal distribution at given numbers of standard deviations from the mean

In that case, I need to standardize the value of 125 so that I can find the probability in the z table:

$$z = \frac{x - \mu}{\sigma} \tag{5.28}$$

$$= \frac{125 - 100}{15} \tag{5.29}$$

$$= 1.67. \tag{5.30}$$

If I want to know $p(IQ > 125)$, this is equivalent to $p(z > 1.67)$. The z table in the appendix only shows negative values for z. Under the symmetry of the distribution, I know that $p(z > Z) = p(-z < -Z)$; here, $p(z > 1.67) = p(-z < -1.67)$. So, I can find $z = -1.67$ in the table. I find that $p(z < -1.67) = .047$. Again, by symmetry, if there is a probability of .047 of obtaining a person who is at least 1.67 standard deviations (z units) below the mean, then there is the same probability of obtaining a person who is 1.67 standard deviations above the mean.

Given the tedium of determining probabilities using the z distribution, I strongly recommend, whenever faced with this type of probability problem, that you take a moment to sketch a normal distribution and mark the area that you are attempting to find. Doing so is especially important when the probability an event falls in some central region is of interest. For example, suppose we would like to determine the probability that a person selected at random has an IQ between 95 and 110. In that case, we need to compute two z scores:

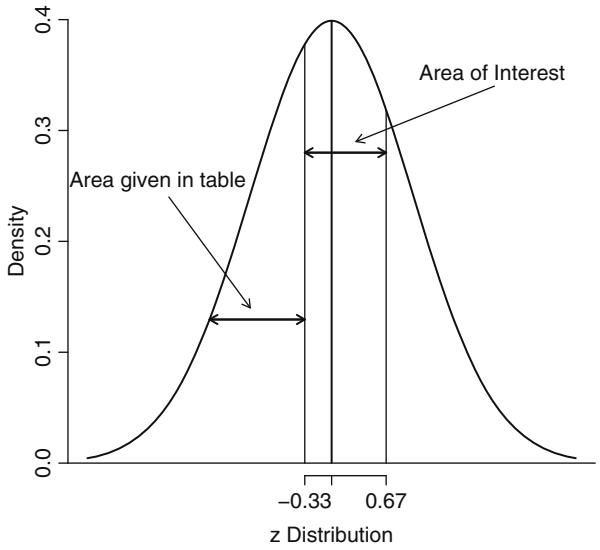

Fig. 5.8. Normal distribution area shown in z table vs. the area of interest in IQ example

$$z_1 = \frac{95 - 100}{15} \tag{5.31}$$

$$= -.33 \tag{5.32}$$

$$z_2 = \frac{110 - 100}{15} \tag{5.33}$$

$$= .67 \tag{5.34}$$

Figure 5.8 shows the region of interest and the region that the z table in the appendix provides. It is clear, from the figure, that we can find the area to the left of $-.33$ in the table, but the remaining areas must be found via using the symmetry property of the normal distribution and subtraction. From the table, the area to the left of $-.33$ is .371. Furthermore, given that $p(z > .67) = p(-z < -.67)$, the area to the right of .67 is .251. Finally, given that the total area under the normal distribution is 1, the area between $-.33$ and .67 is: $1 - (.371 + .251) = .378$. This is the probability of interest.

5.3.3 Normal Approximation to the Binomial

As we have seen in the previous two sections, computing cumulative probabilities may involve multiple computations when using the binomial distribution, but is fairly easy with the normal distribution. For example, determining the probability

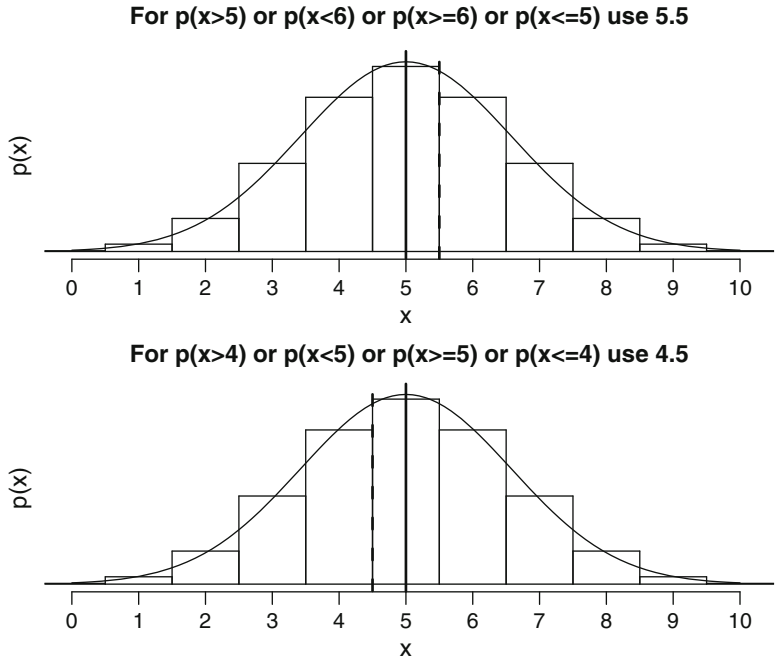

Fig. 5.9. Illustration of continuity correction factor in binomial distribution with $n = 10$ and $p = .5$

of obtaining 100 heads or fewer on 210 coin flips would involve summing the probabilities obtained from 101 applications of the binomial density function (from 0 through 100). In contrast, determining the proportion of the population that has IQs above, say, 50 would only require the calculation and evaluation of one standardized score.

In cases where the binomial distribution is unwieldy, it is possible to use a normal distribution approximation to it. If n is sufficiently "large," say 10 or so, and p is close to .5—or more, generally, if $np > 5$—then the following approximation works pretty well:

$$z = \frac{(x \pm .5) - np}{\sqrt{np(1-p)}}. \tag{5.35}$$

This equation looks similar to our previous standardized score calculation, but we have substituted np for μ and $\sqrt{np(1-p)}$ for σ. np is the expected count of successes, given a particular sample size n and success probability p. $p(1-p)$ is the variance of a proportion, while n times that is the variance of a count.

We have also added a "continuity correction" (the "$\pm .5$") to compensate for the fact that the binomial distribution is discrete, while the normal distribution is continuous. The rationale for the correction factor is displayed in Fig. 5.9. The

figure shows two plots. In both plots, a normal distribution (represented via lines) is superimposed over a binomial distribution (represented via bars) with parameters $n = 10$ and $p = .5$. The normal distribution appears to match the binomial distribution quite well. However, notice that, while $p(x < 5) = 1 - p(x > 5)$ in the normal distribution because there is no probability associated with the exact value of 5, we must make a choice whether to include the probability associated with exactly 5 successes in the binomial distribution computation. The upper plot shows that, if we are interested in the probability that x is greater than 5, less than 6, greater than or equal to 6 or less than or equal to 5, then we should *add* .5 as a correction. The lower plot shows the cases in which we should subtract .5 as a correction. The logic of choosing whether to add or subtract the .5 correction factor is fairly easy to visualize, and so I recommend drawing a figure like the one shown when trying to determine the appropriate computation.

Applying the normal approximation formula is straightforward. For example, determining the probability of obtaining at least 100 heads on 210 coin flips would involve the following steps. First, realize that the probability of obtaining 100+ heads is equivalent to 1− the probability of obtaining 99 heads or fewer. Then, compute the z-score associated with $x = 99$ using the above approximation:

$$z = \frac{99.5 - (210)(.5)}{\sqrt{(210)(.5)(.5)}} = -.76 \tag{5.36}$$

This area is approximately .224. After subtracting from 1, we obtain .776. In fact, using the binomial distribution, we would obtain .7761.

It is important to note that, if n is small, or p varies much from .5, the approximation may not work well. In particular, when p varies considerably from .5, the binomial distribution is not symmetric, *unless n is large enough to offset the asymmetry*. When n is small, even if p is very close (or equal) to .5, the normal approximation may perform poorly because of the distinction between continuous and discrete calculations, even with the continuity correction.

5.4 Conclusions

In this chapter, we discussed probability theory in considerable depth, leading up to the discussion of two important probability distributions used in statistics—the binomial and the normal. These distributions are two of the most commonly used distributions in social science research, and so we spent considerable time developing and discussing them. Importantly, we established that, if a variable follows a normal distribution, we can standardize it and evaluate probabilities of obtaining particular ranges of scores using the z table. This computation will become extremely important in subsequent chapters as we extend it for use in problems of inference.

5.5 Items for Review

- Probability and its rules
- Sample space
- Event
- Joint probability of independent events
- Joint probability of non-independent events
- Conditional probability
- Law of total probability
- Bayes' Theorem
- Probability of one *or* another event
- Venn diagram
- Combinations and permutations
- Sampling with and without replacement
- Probability density function
- Binomial distribution
- Tree diagram
- Pascal's triangle
- Normal distribution
- Standardized score
- Normal approximation to the binomial

5.6 Homework

1. What is the probability that a family with two children has two girls?
2. Now, suppose I introduce you to one of the children, and it is a girl. What is the probability that both children are girls?
3. Suppose license plates in a particular state consist of three letters followed by three numbers, and that both the letters and numbers can repeat. How many license plates can the state produce before repeating?
4. Now suppose license plates in a particular state consist of three letters followed by three numbers, but letters and numbers cannot be repeated. How many plates can the state produce without repeating?
5. Suppose a state produces licence plates that consist of four letters that can be repeated, and your name is "John." What is the probability, assuming you're the first person issued a plate, that the plate will have your name on it?
6. I have a set of three dice. One is a typical 6-sided die, one is a 4-sided die, and one is 10-sided die. First, I roll the 4-sided die. If that result is a 1, 2, or a 3, then I roll the 6-sided die. If the result of the roll of the 4-sided die is a 4, then I roll the 10-sided die instead of the 6-sided die. What is the probability, overall, that in this process I will roll a 5?
7. I rolled a 5. What is the probability that I rolled the 10-sided die?

8. If IQ is normally distributed in the population with a mean of 100 and standard deviation of 15, what is the probability of randomly selecting a person with an IQ greater than 130?
9. What is the probability of obtaining two such people in a row?
10. There are six marbles in a box; four are red and two are green. Draw a histogram showing the frequency of red marbles in samples of size n = 3.
11. What is the probability of obtaining 4 heads in a row, followed by a tail, on 5 flips of a fair coin?
12. What is the probability of obtaining 4 heads and 1 tail on 5 flips of a fair coin, regardless of the order of heads and tails?
13. What if the coin were weighted so that the probability of obtaining a head on any given flip were .8?
14. In the game Yahtzee, a "yahtzee" happens when you obtain the same number on each of five dice. What is the probability of getting a yahtzee on a single throw (i.e., getting 5 of the same number on one roll)?
15. Male body weight is approximately normally distributed in the population with a mean of 190 pounds and a standard deviation of 59 pounds. What proportion of males weighs between 175 and 200 pounds?
16. What is the probability of obtaining a sample of five men, all of whom are in that weight range?
17. Approximately 50 % of the population is male, and 30 % is obese. Twenty percent of males are obese. Draw a Venn diagram illustrating these proportions, as well as the proportion of the population that are non-obese females.
18. Based on the above problem, if I randomly select an individual from the population, what is the probability the person would be either a male or obese?
19. Based on the above problem, if I randomly select an individual from the population, what is the probability the person would be an obese male?
20. Based on the above problem, if I randomly select an individual from the population, and I know the person is obese, what is the probability that the person is male?
21. Height in the male population is normally distributed with a mean height of 5–11 (71 in.) and a standard deviation of 4 in. I am 5–8 (68 in.). In what percentile of the height distribution do I fall?
22. I like to flip coins, so I flip a quarter 100 times and obtain 55 heads. What is the probability, assuming the coin is fair, that I would obtain 55 or more heads in 100 flips of a fair coin?
23. A deck of cards consists of four "suits" (clubs, diamonds, hearts, and spades, with clubs and spades being black, and diamonds and hearts being red), each of which has 13 cards, including numbered cards from 2 through 10, three "face" cards (Jack, Queen, and King), and an ace, which can be considered a one or a high card above the King. In five card stud poker, players are each dealt five cards, and they must make the best hand possible out of the cards dealt. How many unique five card stud poker hands are possible?
24. If I deal you five cards, what is the probability of obtaining a royal flush? (royal flush is defined by having the 10, Jack, Queen, King, and Ace of a single suit).

25. What is the probability of obtaining a flush? (all cards are of the same suit. Don't exclude royal and straight flushes).
26. What is the probability that you are dealt all red OR all black cards?
27. What is the probability that you are dealt a four-of-a-kind (all four of a given number or face card)?
28. If I roll two dice, what is the probability that at least one of them will show a number greater than 4?
29. The breakfast buffet I went to last week had an omelet station with ham, green peppers, onions, cheese, bacon, mushrooms, and spinach as possible toppings. Assuming I can have as many of these toppings as I want (no repeats), how many different omelets are possible?
30. What is the probability that two people in a row will order the same omelet (assume they do not know each other and are unaware of the other's order)?
31. I flipped a fair coin 10 times in a row and got heads on every flip. What is the probability the next toss will land on heads?
32. What is the probability of obtaining 5 heads in a row on 5 flips of a fair coin?
33. There is a probability of .6 that I will give a tough final exam. If the final is easy, there is a probability of .8 that you will make an A. If the final is tough, there is a probability of .3 that you will make an A. What is the probability that you will make an A?
34. You find out you received an A. What is the probability that the exam was tough?
35. What is the probability of obtaining exactly 5 sums that exceed 8 on 10 rolls of a pair of dice?
36. If I draw two cards from a deck, what is the probability that I will obtain a four on the first draw followed by an ace on the second draw?
37. My local hardware store has a supply of 100 light bulbs, of which 5 are defective. If I buy two bulbs, what is the probability that both of them will be defective?
38. A certain machine has a triple redundancy system to minimize its probability of failure. If a particular component fails, a second one takes over the task. If the second one fails, the third takes over the task. If that component fails, then the machine fails. There is a probability of .8 that the first component will fail. If the second component is called, there is a probability of .5 that it will fail. If the second component fails, there is a probability of .3 that the third will fail. What is the probability that the machine will actually work?
39. I want to arrange 5 people out of a class of 10 on one side of a table. How many ways can I do this?
40. The probability that a Republican will win the presidential election is 0 if he does not win the state of Ohio. The probability that he will win if he does win Ohio is .35. If I tell you that the Republican won the election, what is the probability that he won Ohio?
41. The probability that a person will become rich if s/he does not finish college is (1/10,000). The probability that a person will become rich if s/he does finish

college is (1/10). The probability of college completion is .3. I introduce you to a rich person. What is the probability that she finished college?

42. In a Major League Baseball season, each team plays 162 games. Suppose a team's probability of winning each game is .5. What is the probability that a team will win more than 90 games?
43. What is the probability of rolling doubles (matching numbers) on a pair of dice?
44. I have three large koi in my outdoor pond. I'm hoping that they might produce baby fish, but they cannot do this if they're all the same sex (exclude the possibility of parthenogensis, which is known to happen in some fish). What is the probability that they're all the same sex? (assume the probability of .5 for each sex).
45. I'm trying to determine the proportion of persons in a community who tested HIV-positive in a recent community-wide blood test. This is a highly sensitive question, and many may not answer it directly. So, I ask the survey respondents to roll a die (and not show me the result). If the roll comes up 1 or 2, they are to answer "yes" to the question, regardless of whether they have HIV. If the die roll comes up 3 through 6, they are to answer truthfully. After collecting my data, I find that 50 % of respondents answers "yes" to the question. What proportion of the community in fact has HIV? (assume honesty, given the indirect method used, and assume the HIV test is 100 % accurate)
46. What is the probability that a person who said "yes" actually is HIV positive?
47. Assume that final exam scores in a given course are approximately normally distributed with a mean of 75 and a standard deviation of 10 points. In a class of 200 students, how many would you expect to fail the exam (i.e., score less than 60) (note: scores can only be approximately normally distributed, because the minimum score is 0, but this should not alter the answer, given how large the mean is relative to the standard deviation).
48. How many students can be expected to make an A; i.e., score 90 or higher?
49. The length of life of a certain brand of light bulb is normally distributed with a mean of 1,000 h and a standard deviation of 200 h (note: this is a very unreliable set of light bulbs!). If I buy new light bulbs for two lamps before I leave for a month-long vacation (30 days), and I leave both lamps on when I leave, what is the probability that at least one will be working when I return?
50. As described in a previous problem, height in the male population is normally distributed with a mean height of 5–11 (71 in.) and a standard deviation of 4 in. If I randomly selected 5 men from the population, what is the probability that I would have a collection of 5 men over 6 ft tall?