Thus, there are 720 different ways of arranging six slides. If we want to present all possible arrangements to each participant, we are going to need 720 trials, or some multiple of that. That is a lot of trials. For the second problem, where we have six slides but show only three to any one subject, we have

$$P_3^6 = \frac{6!}{(6-3)!} = \frac{6!}{3!} = \frac{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{6} = 120$$

If we want to present all possible arrangements to each subject, we need 120 trials, a result that may still be sufficiently large to lead us to modify our design. This is one reason we often use random orderings rather than try to present all possible orderings.

Combinations

To return to the ice-cream lottery, suppose we now decide that we will award only single-dip cones to the two winners. We will still draw the names of two winners out of a hat, *but we will no longer care which of the two names was drawn first*—the result AB is for all practical purposes the same as the result BA because in each case Aaron and Barbara win a cone. When the order in which names are drawn is no longer important, we are no longer interested in permutations. Instead, we are now interested in what are called **combinations.** We want to know the number of possible combinations of winning names, but not the order in which they were drawn.

combinations

We can enumerate these combinations as

A B	B C
A C	B D
A D	C D

There are six of them. In other words, out of four people, we could compile six different sets of winners. (If you look back to the previous enumeration of permutations of winners, you will see that we have just combined outcomes containing the same names.)

Normally, we do not want to enumerate all possible combinations just to find out how many of them there are. To calculate the number of *combinations* of N things taken r at a time C_r^N, we will define

$$C_r^N = \frac{N!}{r!(N-r)!}$$

For our example,

$$C_2^4 = \frac{4!}{2!(4-2)!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 2 \cdot 1} = 6$$

Let's return to the example involving slides to be presented to subjects. When we were dealing with permutations, we worried about the way in which each set of slides was arranged; that is, we worried about all possible orderings. Suppose we no longer care about the order of the slides within sets, but we need to know how many different sets of slides we could form if we had six slides but took only three at a time. This is a question of combinations.

For six slides taken three at a time, we have

$$C_3^6 = \frac{6!}{3!(6-3)!} = \frac{\overset{2}{\cancel{6}} \cdot 5 \cdot \overset{2}{\cancel{4}} \cdot \cancel{3} \cdot \cancel{2} \cdot \cancel{1}}{\cancel{3} \cdot \cancel{2} \cdot \cancel{1} \cdot \cancel{3} \cdot \cancel{2} \cdot 1} = 20$$

If we wanted every subject to get a different set of three slides but did not care about the order within a set, we would need 20 subjects.

Later in the book we will discuss procedures, called *permutation tests,* in which we imagine that the data we have are all the data we could collect, but we want to imagine what the sample means would likely be if the N scores fell into our two different experimental groups (of n_1 and n_2 scores) *purely at random*. To solve that problem, we could calculate the number of different ways the observations could be assigned to groups, which is just the number of combinations of N things taken n_1 and n_2 at a time. (Please don't ask why it's called a permutation test if we are dealing with combinations—I haven't figured that out yet.) Knowing the number of different ways that data could have occurred at random, we will calculate the percentage of those outcomes that would have produced differences in means at least as extreme as the difference we found. That would be the probability of the data given H_0: true, often written $p(D|H_0)$. I mention this here only to give you an illustration of when we would want to know how to calculate permutations and combinations.

5.7 Bayes' Theorem

Bayes' theorem

We have one more basic element of probability theory to cover before we go on to use those basics in particular applications. This section is new to this edition, not because **Bayes' theorem** is new (it was developed by Thomas Bayes and first read before the Royal Society in London in 1764—3 years after Bayes' death), but because it is becoming important that people in the behavioral sciences know what the theorem is about, even if they forget the details of how to use it. (You can always look up the details.)

Bayes' theorem tells us how to accumulate information to revise estimates of probabilities. By "accumulate information" I mean a process in which you continually revise a probability estimate as more information comes in. Suppose that I tell you that Fred was murdered and ask you for your personal (subjective) probability that Willard committed the crime. You think he is certainly capable of it and not a very nice person, so you say $p = .15$. Then I say that Willard was seen near the crime that night, and you raise your probability to $p = .20$. Then I say that Willard owns the right type of gun, and you might raise your probability to $p = .25$. Then I say that a fairly reliable witness says Willard was at a baseball game with him at the time, and you drop your probability to $p = .10$. And so on. This is a process of accumulating information to come up with a probability that some event occurred. For those interested in Bayesian statistics, probabilities are usually subjective or personal probabilities, meaning that they are a statement of personal belief, rather than having a frequentist or analytic basis as defined at the beginning of the chapter. Bayes' theorem will work perfectly well with any kind of probability, but it is most often seen with subjective probabilities.

Let's take a simple example. Suppose that I asked you to give me your estimate of the probability that I am writing this section in April. You have no idea what month it is, so you would probably say that the probability is $1/12 = .083$, if we ignore the fact that some months have more days than others. (The probability that it is not April is $11/12 = .917 = 1 - .083$.) Now I tell you that I can look out my office window and see that it snowed in the mountains in the last few days. Although that is not definitive information, it certainly should be helpful. (You probably would be inclined to doubt that it is July.) What you want to do is to revise your earlier estimate ($p = .083$) on the basis of this new information. That is what Bayes' theorem allows you to do.

prior probability

First, define $p(A)$ as the **prior probability** that it is April. We call it a prior probability because it is the probability you estimate *before* I tell you anything about snow. Define $p(NA)$ as the probability that it is not April, which we have specified as .917. We will define

posterior probability

$p(A|S)$ as the **posterior probability** that it is April given that you know that it snowed in the last few days. This is the probability that we ultimately want to estimate. We call it a posterior probability because it is your revised probability *after* receiving information about snow. Last, we need to consider the probability that it would snow last night if the month really is April ($p(S|A)$ and the probability it would snow last night if this is not April (NA), which is represented by $p(S|NA)$. You could either look these probabilities up in meteorological tables, or you could just take a reasonable guess—making them subjective probabilities. I will choose to guess based on what I know about the annual snow patterns where I live. I will guess that the probability of having snow in the last few days *given that it is April* is .20, and the probability that it snowed *given that it is not April* is .10 (remember all those summer months when it doesn't snow).[2]

So now I have the following probabilities

$$p(A) = 1/12 = .083 \qquad p(NA) = 11/12 = .917$$
$$p(S|A) = .20 \qquad p(S|NA) = .10$$

What we now want is the probability of it being April, *given the data about snow*. Bayes' theorem tells us that this probability is given by

$$p(A|S) = \frac{p(S|A)p(A)}{p(S|A)p(A) + p(S|NA)p(NA)}$$

Substituting what we already know, we have

$$p(A|S) = \frac{p(S|A)p(A)}{p(S|A)p(A) + p(S|NA)p(NA)}$$
$$= \frac{(.20)(.083)}{(.20)(.083) + (.10)(.917)}$$
$$= \frac{.0166}{.0166 + .0917} = \frac{.0166}{.1083} = .1533$$

When you didn't know anything about snow, your best estimate of the probability was .083. Once you know that it snowed, you were able to revise that probability to .153, nearly doubling it. This it the kind of task that Bayes' theorem was designed to solve. It allows you to accumulate information and update your estimates.

A lot of work in human decision making has been based on applications of Bayes' theorem. Much of it focuses on comparing what people *should* say in a situation, with what they *actually* say, for the purpose of characterizing how people really make decisions. A famous problem was posed to decision makers by Tversky and Kahneman (1980). This problem involved deciding which cab company was involved in an accident. We are told that there was an accident involving one of the two cab companies (Green Cab and Blue Cab) in the city, but we are not told which one it was. We know that 85% of the cabs in a given city are Green, and 15% are Blue. The prior probabilities then, based on the percentage of Green and Blue cabs, are .85 and .15. If that were all you knew and were then told that someone was just run over by a cab, your best estimate would be that the probability that it was a Green cab is .85. Then a witness comes along who thinks that it was a

[2] The probability of snow in April may look high to you, but I live in the Colorado Rockies, and those mountains are at 10,000 to 12,000 feet. (So why do I give my affiliation as the University of Vermont? Because I retired from there and am now "Professor Emeritus." They give me the title, and I give them the credit. Fair trade.)

Blue cab. You might think that was conclusive, but identifying colors at night is not a foolproof task, and the insurance company tested our informant and found that he was able to identify colors at night with only 80% accuracy. Thus, if you show him a Blue cab, the probability that he will correctly say Blue is .80, and the probability that he will incorrectly say Green is .20. (Similarly, if the cab is Green.) So our conditional probability that the cab was a Blue cab, given that he said it was Blue is .80, and the conditional probability that it was Green, given that he said it was Blue is .20. This information is sufficient to allow you to calculate the posterior probability that the cab was a Blue cab given that the witness said it was blue.

In the following formula, let B stand for the event that it was a Blue cab, and let b stand for the event that the witness called it blue. Similarly for G and g.

$$p(B|b) = \frac{p(b|B)p(B)}{p(b|B)p(B) + p(g|B)p(G)}$$
$$= \frac{(.80)(.15)}{(.80)(.15) + (.20)(.85)}$$
$$= \frac{.12}{.12 + .17} = \frac{.12}{.29} = .414$$

Most of the participants in Tversky and Kahneman's experiment guessed that the probability that it was the Blue Cab was around .80, when in fact the correct answer is approximately .41. Thus, Kahneman and Tversky concluded that judges place too much weight on the witness' testimony, and not enough weight on the prior probabilities. Here is a situation where the discrepancy between what judges say and what they should say gives us clues about the strategies that judges use and where they go wrong.

A Generic Formula

The formulae given previously were framed in terms of the specific example under discussion. It may be helpful to have a more generic formula that you can adapt to your own purposes. Suppose that we are asking about the probability that some hypothesis (H) is true, given certain data (D). For our examples, H represented "the month is April" or "it was the Blue Cab company." The D represent "it snowed" or "the witness reported that the cab was blue" The symbol $\overline{H}$ is read "not H" and stands for the case where the hypothesis is false. Then

$$p(H|D) = \frac{p(D|H)p(H)}{p(D|H)p(H) + p(D|\overline{H})p(\overline{H})}$$

Back to the Hypothesis Testing

In Chapter 4, we discussed hypothesis testing and different approaches to it. Bayes' theorem has an important contribution to make to that discussion, although I am only going to touch on the issue here. (I want you to understand the nature of the argument, but it is not reasonable to expect you to go much beyond that.) Recall that I said that in some ways a hypothesis test is not really designed to answer the question we would ideally like to answer. We want to collect some data and then ask about the probability that the null hypothesis is true given the data. But instead, our statistical procedures tell us the probability that we would obtain those data given that the null hypothesis (H_0) is true. In other words, we want

$p(H_0|D)$ when what we really have is $p(D|H_0)$. Many people have pointed out that we could have the answer we seek if we simply apply Bayes' theorem,

$$p(H_0|D) = \frac{p(D|H_0)p(H_0)}{p(D|H_0)p(H_0) + p(D|H_1)p(H_1)}$$

where H_0 stands for the null hypothesis, H_1 stands for the alternative hypothesis, and D stands for the data.

The problem here is that we don't know most of the necessary probabilities. We could *estimate* those probabilities, but those would only be estimates. It is one thing to be able to calculate that the probability of April is .083 because April is one of 12 months in the year. But it is quite a different thing to be able to estimate the probability that the null hypothesis is true. Using the example of waiting times in parking lots, you and I might have quite different prior probability estimates that people leave a parking space at the same speed whether or not there is someone waiting. In addition, our statistical test is designed to give us $p(D|H_0)$, which is helpful. But where do we obtain $p(D|H_1)$ from if we don't have a specific alternative hypothesis is mind (other than the negation of the null)? It was one thing to estimate it when we had something concrete (like all months but April), but considerably more difficult when the alternative is that people leave *more slowly* when someone is waiting if we don't know *how much* more slowly. The probabilities would differ dramatically if we think of "5 seconds more slowly" compared with "25 seconds more slowly." That these probabilities we need are hard, or impossible, to determine has stood in the way of developing this as a general approach to hypothesis testing—though many have tried. (One approach is to chose a variety of reasonable estimates, and note how the results hold up under those different estimates. If most believable estimates lead to the same conclusion, that tells us something useful.)

Bayesian statistics

I don't mean to suggest that the application of Bayes' theorem (known as **Bayesian statistics**) is hopeless—it certainly is not. Many people are very interested in that approach, though its use is mostly restricted to situations where the null and alternative hypotheses are sharply defined, such as H_0: $\mu = 0$ and H_1: $\mu = 3$. But I have never seen clearly specified alternative hypotheses in the behavioral sciences.

5.8 The Binomial Distribution

binomial distribution

We now have all the information on probabilities and combinations that we need for understanding one of the most common probability distributions—the **binomial distribution.** This distribution will be discussed briefly, and you will see how it can be used to test simple hypotheses. I don't think that I can write a chapter on probability without discussing the binomial distribution, but many students and instructors would be more than happy if I skipped this topic. There certainly are many applications for it (the sign test to be discussed shortly is one example), but I would easily forgive you for not wanting to memorize the necessary formulae—you can always look them up.

Bernoulli trial

The binomial distribution deals with situations in which each of a number of independent trials results in one of two mutually exclusive outcomes. Such a trial is called a **Bernoulli trial** (after a famous mathematician of the same name). The most common example of a Bernoulli trial is flipping a coin, and the binomial distribution could be used to give us the probability of, for example, 3 heads out of 5 tosses of a coin. Most people don't get excited by the prospect of flipping coins, so think of calculating the probability that 20 out of your 30 cancer patients will survive a diagnosis of lung cancer if the probability of survival for any one of them is .70.

The binomial distribution is an example of a discrete, rather than a continuous, distribution because one can flip coins and obtain 3 heads or 4 heads, but not, for example, 3.897 heads. Similarly one can have 21 survivors or 22 survivors, but not anything in between.

Mathematically, the binomial distribution is defined as

$$p(X) = C_X^N p^X q^{(N-X)} = \frac{N!}{X!(N-X)!} p^X q^{(N-X)}$$

where

$p(X)$ = The probability of X successes
N = The number of trials
p = The probability of a success on any one trial
$q = (1 - p)$ = The probability of a failure on any one trial
C_X^N = The number of combinations of N things taken X at a time

The notation for combinations has been changed from r to X because the symbol X is used to refer to data. Whether we call something r or X is arbitrary; the choice is made for convenience or intelligibility.

success

failure

The words **success** and **failure** are used as arbitrary labels for the two alternative outcomes. If we are talking about cancer, the meaning is obvious. If we are talking about whether a driver will turn left or right at a fork, the designation is arbitrary. We will require that the trials be independent of one another, meaning that the result of trial$_i$ has no influence on trial$_j$.

To illustrate the application of this formula, suppose we are interested in studying the art of wine tasting and the relationship between quality and price. As part of our study, we ask a judge to taste two glasses of wine and pick the one that she thinks is the more expensive one. This task is repeated 10 times, each time with a different pair of wines. Assume for the moment that our wine taster really does not know the first thing about wines (or that quality and price are completely unrelated—which is sometimes the case). Assuming that there are no extraneous factors to bias the judge's decision (such as a tendency to choose the darker-colored wine), then on each trial the probability of her being correct (i.e., correctly identifying the more expensive wine) is .50 because there are only two wines to choose from. Now suppose we want to know the probability that our judge will somehow manage to make 9 (X) correct choices out of 10 (N) trials when the null hypothesis ($p = .50$) is true. The probability of being correct on any one trial is denoted p and equals .50, whereas the probability of being incorrect on any one trial is denoted q and also equals .50. Then we have

$$p(X) = \frac{N!}{X!(N-X)!} p^X q^{(N-X)}$$

$$p(9) = \frac{10!}{9!1!}(.50^9)(.50^1)$$

But $10! = 10 \cdot 9 \cdot 8 \cdot \cdots \cdot 2 \cdot 1 = 10 \cdot 9!$ so

$$p(9) = \frac{10 \cdot \not{9}!}{\not{9}!1!}(.50^9)(.50^1)$$

$$= 10(.001953)(.50) = .0098$$

Thus, the probability of making 9 correct choices out of 10 trials with $p = .50$ is remote, occurring approximately 1 time out of every 100 tasting sessions.