





1 INTRODUCTION

In the study of probability theory, we assign a probability measure to the possible outcomes of an experiment. We then make probability predictions relating to the experiment. For example, a coin is tossed ten times. We assign an equal weight to all possible sequences of heads and tails. We then compute the probability that exactly six heads turn up. We find that this probability is $\binom{10}{6} \cdot \binom{1}{2}^{10} = .205$. Statistics deals with the inverse problem. We do not know the basic probability measure, but we are able to carry out certain chance experiments, from which we obtain information about the underlying measure.

As an example, assume that in a large population each person holds an opinion on the question of legalizing marijuana. They either favor this or are opposed. We choose at random 20 people and ask them their opinions. Choosing "at random" means that we have an equal chance of obtaining any group of 20 people from the entire population. If the size of the population is large, the effect of knowing certain of the opinions will not significantly change the chance that the next person sampled will say "yes." Thus it is reasonable to assume that the underlying chance model is an independent trials model with probability p for success (answer "yes") on each trial, where p is the proportion in the entire population that favor legalizing marijuana.

On the basis of the sample we would like to estimate p. The intuitive estimate for the parameter p would be simply the fraction \overline{p} of persons in the sample that say "yes." In Figure 1 we show the result of drawing ten samples of 20 each in a case where p = .4. While our estimates are in general near the true value .4, our worst estimate is .2, only half the true value.

Experiment number	Number of "yes" answers	Fraction \overline{p}
1	6	.3
2	. 8	.4
3	9	.4 .45
4	4	.2
5	7	.35
6	6	.3
7	6	.3
8	9	.45
9	5	.25
10	7	.35

Figure 1

From the binomial measure (see Chapter 3, Section 8) we can calculate the exact probability that the observed fraction \overline{p} will lie in a given range. For example, the values of f(20, x; .4) for x between 6 and 10 add up to .747. Thus with probability .747 our estimate will be between .3 and .5.

We recall from our study of independent trials that the expected number of successes in a sample of size *n* is *np* and the standard deviation for the number of successes is \sqrt{npq} (see Chapter 3, Section 9). Further, the probability of a deviation of more than 3 standard deviations from the expected number is very unlikely (.001). Thus if we increase the sample size to 2400 the expected number of "yes" responses would be 960 and the standard deviation $\sqrt{2400 \times .4 \times .6} = 24$. Thus our estimate would with high probability lie between $\frac{960 - 72}{2400} = .37$ and $\frac{960 + 72}{2400} = .43$, in the interval [37, 43]

interval [.37, .43].

This suggests that when p is unknown we should try to estimate from the sample an interval within which we believe the true p lies. We shall show in Section 4 that this can indeed be done.

In some situations we need to make a choice between two estimates for p. For example, the incidence of colds may be known and we wish to test the claim that this can be decreased if people take large doses of vitamin C. Thus we have to determine whether the incidence of colds among those taking vitamin C is the same as for the whole population or a smaller value. Or a manufacturer may assume that his production process is operating correctly if it produces no more than 1 percent defective items but is not operating correctly if it produces as many as 5 percent defective items. He is interested in devising a test to see if the system is operating correctly.

Perhaps the largest statistical test ever conducted was the test designed in the early '50s to see if the vaccine developed by Jonas Salk would effectively cut down the incidence of polio. The average incidence of polio at that time was about 50 per 100,000 persons. It was not expected that the vaccine would be 100 percent effective, but it was hoped that it would cut down the incidence of polio by at least 50 percent. Thus we can view the experiment as a test of the hypothesis that a person vaccinated will have a significantly lower probability of being afflicted with polio than a person not vaccinated. This type of hypothesis testing will be studied in Section 3.

In applying probability models, predictions from the model are only reliable if the assumptions made in describing the model are reasonably met. Similarly, our statistical inferences are based upon certain assumptions. We have already mentioned the assumption of randomness in a sample. There are many pitfalls that one can fall into if care is not taken. Perhaps the most famous example of this is the celebrated prediction of the *Literary Digest* that Alfred Landon would defeat Franklin Roosevelt in the 1936 presidential election. In this poll the sample was chosen from names obtained from telephone books and car registrations. In 1936 this was not at all a "random sample" and the prediction was badly in error. Opinion polls are still trying to recover from this blunder. We shall discuss this and other pitfalls in more detail in Section 5.

Before we continue our discussion of statistics we shall need one important result from probability theory called the central limit theorem. This will be studied in the next section.

For use in the exercises and in later sections we show the probabilities for ten independent trials and various values of p in Figure 2.

x	0.1	0.25	0.4	0.5	0.6	0.75	0.9
0	0.349	0.056	0.006	0.001	0	0	0
1	0.387	0.188	0.040	0.010	0.002	0	0
2	0.194	0.282	0.121	0.044	0.011	0	0
3	0.057	0.250	0.215	0.117	0.042	0.003	0
4	0.011	0.146	0.251	0.205	0.111	0.016	0
5	0.001	0.058	0.201	0.246	0.201	0.058	0.001
6	0	0.016	0.111	0.205	0.251	0.146	0.011
7	0	0.003	0.042	0.117	0.215	0.250	0.057
8	0	0	0.011	0.044	0.121	0.282	0.194
9	0	0	0.002	0.010	0.040	0.188	0.387
10	0	0	0	0.001	0.006	0.056	0.349

Table of values of f(10, x; p)

Figure 2

EXERCISES

- A random sample of ten persons is chosen in New York City at a time when 60 percent are in favor of Kelly for mayor and 40 percent are in favor of McGrath. What is the probability that the sample will show less than 50 percent in favor of Kelly? [Ans. .166.]
- 2. An independent trials experiment is repeated ten times with six suc-

cesses. Which value of p in Figure 2 gives the highest probability of obtaining the outcome of six successes—i.e., the observed outcome?

- 3. In Exercise 1 assume that the sample size is increased to 9600. Find the expected number and the standard deviation for the number of those in favor of Kelly. What could we say about the range of our estimates if the number of "yes" responses does not deviate by more than three standard deviations from the expected number?
- 4. In a city there are 100,000 persons who are going to vote on the question of legalizing marijuana. Of these, 90,000 are under 50 years of age and 10,000 are 50 or over. Assume that 75 percent of those under 50 favor legalizing marijuana and of those 50 or older only 20 percent are in favor. What is the probability that a person chosen at random will favor legalizing marijuana? In a sample of 100 chosen at random what is the expected number that will answer "yes"? What is the expected number if a random sample of 100 is chosen, 50 from each of the two groups?
- 5. In an experiment where the probability distribution depends on a single number, or parameter, the following is a standard method of estimating this parameter. Choose the value of the parameter which gives the highest probability of obtaining the observed result. This method is called the method of *maximum likelihood*. On the basis of the result of Exercise 2, what would you guess to be the maximum likelihood estimator for an independent trials experiment for the probability p of success when x successes are observed in n trials?
- 6. A box has ten items, eight good and two defective. A sample of five is chosen with replacement—that is, after each item is chosen and inspected it is replaced (i.e., put back) before the next item is drawn. Find the probability that the sample has exactly one defective item. [Ans. .410.]
- 7. Answer the same question as in Exercise 6 if the sampling is done without replacement. That is, a set of five is chosen at random from all possible subsets of five items of the box. Find the probability that the sample has exactly one defective item and compare your answer to that obtained in Exercise 6.

$$[Ans. \frac{\binom{2}{1}\binom{8}{4}}{\binom{10}{5}} = .556.]$$

- 8. A sample of three items is chosen from a box of 1000 of which 80 percent are defective. Show that the probability of obtaining exactly one defective item is essentially the same whether we sample with or without replacement.
- 9. Assume that the incidence of lung cancer among smokers is estimated to be 20 per 100,000 and among heavy smokers to be 200 per 100,000.

Estimate the probability that a person who smokes will *not* get lung cancer and compare this with the estimate for a heavy smoker.

[Ans. .9998, .998.]

- 10. A hardware store receives boxes of 50 bolts. Experience has shown that they occasionally get a bad lot. When they get a box they choose two bolts at random and if either is defective they return the box. Assume that a box has five defective bolts. What is the probability that the box will be sent back?
- 11. Referring to Exercise 10, assume that the store receives shipments of ten boxes each with 50 bolts. It combines the 500 bolts and then chooses two bolts at random; if either is defective it sends back the entire lot. If the shipment contains 50 defectives in all, is the probability of the lot being returned larger than, equal to, or smaller than the probability in Exercise 10 of a single box with five defectives being returned?
- 12. Toss a coin 100 times. In each group of ten tosses count the number of heads. Compare the results with Figure 2.
- 13. Toss a pair of coins 100 times. In each group of ten tosses count the number of times two heads turn up. Compare the results with Figure 2.

2 THE CENTRAL LIMIT THEOREM

As we have indicated, to go further in our discussion of statistics we shall need an important theorem of probability theory called the *central limit theorem*. While this is a very general theorem, we shall discuss it in this section only as it applies to independent trials processes.

As usual, let p be the probability of success on a trial and f(n, p; x) the probability of exactly x successes in n trials.

In Figure 3 we have plotted bar graphs which represent f(n, .3; x) for n = 10, 50, 100, and 200. We note first of all that the graphs are drifting off to the right. This is not surprising, since their peaks occur at np, which is steadily increasing. We also note that while the total area is always 1, this area becomes more and more spread out.

We want to redraw these graphs in a manner that prevents the drifting and the spreading out. First of all, we replace x by x - np, assuring that our peak always occurs at 0. Next we introduce a new unit for measuring the deviation, which depends on n, and which gives comparable scales. As we saw in Chapter 3, Section 9, the standard deviation \sqrt{npq} is such a unit.

We must still insure that probabilities are represented by areas in the graph. In Figure 3 this is achieved by having a unit base for each rectangle, and having the probability f(n, p; x) as height. Since we are now representing a standard deviation as a single unit on the horizontal axis, we must take $f(n, p; x)\sqrt{npq}$ as the heights of our rectangles. The resulting curves for n = 50 and n = 200 are shown in Figures 4 and 5, respectively.





We note that the two figures look very much alike. We have also shown in Figure 5 that it can be approximated by a bell-shaped curve. This curve represents the function*

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

and is known as the *normal curve*. It is a fundamental theorem of probability theory that as n increases, the appropriately rescaled bargraphs more and more closely approach the normal curve. The theorem is known as the *central limit theorem*, and we have illustrated it graphically.

More precisely, the theorem states that for any two numbers a and b, with a < b,

*The number e is the base of natural logarithms and its numerical value is 2.71828182.... Its derivation and most important properties are discussed in most calculus books.





$$\Pr\left[a < \frac{x - np}{\sqrt{npq}} < b\right]$$

approaches the area under the normal curve between a and b, as n increases. This theorem is particularly interesting in that the normal curve is symmetric about 0, while f(n, p; x) is symmetric about the expected value np only for the case $p = \frac{1}{2}$. It should also be noted that we always arrive at the same normal curve, no matter what the value of p is.

In Figure 6 we give a table for the area under the normal curve between 0 and d. Since the total area is 1, and since it is symmetric about the origin, we can compute arbitrary areas from this table. For example, suppose that





d	A(d)	d	A(d)	d	A(d)	d	A(d)
.0	.000	1.1	.364	2.1	.482	3.1	.4990
.1	.040	1.2	.385	2.2	.486	3.2	.4993
.2	.079	1.3	.403	2.3	.489	3.3	.4995
.3	.118	1.4	.419	2.4	.492	3.4	.4997
.4	.155	1.5	.433	2.5	.494	3.5	.4998
.5	.191	1.6	.445	2.6	.495	3.6	.4998
.6	.226	1.7	.455	2.7	.497	3.7	.4999
.7	.258	1.8	.464	2.8	.497	3.8	.49993
.8	.288	1.9	.471	2.9	.498	3.9	.49995
.9	.316	2.0	.477	3.0	.4987	4.0	.49997
1.0	.341					5.0	.49999997

Figure 6

we wish the area between -1 and +2. The area between 0 and 2 is given in the table as .477. The area between -1 and 0 is the same as between 0 and 1, and hence is given as .341. Thus the total area is .818. The area outside the interval (-1, 2) is then 1 - .818 = .182.

EXAMPLE 1 Let us find the probability that s differs from the expected value np by as much as d standard deviations.

$$\Pr\left[|x - np| \ge d\sqrt{npq}\right] = \Pr\left[\left|\frac{x - np}{\sqrt{npq}}\right| \ge d\right],$$

and hence the approximate answer should be the area outside the interval (-d, d) under the normal curve. For d = 1, 2, 3 we obtain

 $1 - (2 \times .341) = .318, \quad 1 - (2 \times .477) = .046$

and

$$1 - (2 \times .4987) = .0026$$

respectively. These agree with the values given in Chapter 3, Section 9, to within rounding errors. In fact, the central limit theorem is the basis of those estimates.

EXAMPLE 2 In Chapter 3, Section 9, we considered the example of tossing a coin 10,000 times. The expected number of heads that turn up is 5000, and the standard deviation is $\sqrt{10,000 \cdot \frac{1}{2} \cdot \frac{1}{2}} = 50$. We observed that the probability of a deviation of more than two standard deviations (or 100) is very unlikely.

[Ans. .006.]

On the other hand, consider the probability of a deviation of less than .1 standard deviation—that is, of a deviation of less than 5. The area from 0 to .1 under the normal curve is .040, and hence the probability of a deviation from 5000 of less than 5 is approximately .08. Thus, while a deviation of 100 is very unlikely, it is also very unlikely that a deviation of less than 5 will occur.

EXAMPLE 3 The normal approximation can be used to estimate the individual probability f(n, x; p) for large n. For example, let us estimate f(200, 65; .3). The graph of the probabilities f(200, x; .3) was given in Figure 5 together with the normal approximation. The desired probability is the area of the bar corresponding to x = 65. An inspection of the graph suggests that we should take the area under the normal curve between 64.5 and 65.5 as an estimate for this probability. In normalized units this is the area between

$$\frac{4.5}{\sqrt{200(.3)(.7)}}$$
 and $\frac{5.5}{\sqrt{200(.3)(.7)}}$,

or between .6944 and .8487. Our table is not fine enough to find this area, but from more complete tables, or by machine computation, this area may be found to be .046 to three decimal places. The exact value to three decimal places is .045. This procedure gives us a good estimate.

If we check all of the values of f(200, x; .3) we find in each case that we would make an error of at most .001 by using the normal approximation. There is unfortunately no simple way to estimate the error caused by the use of the central limit theorem. The error will clearly depend upon how large n is, but it also depends upon how near p is to 0 or 1. The greatest accuracy occurs when p is near $\frac{1}{2}$.

EXERCISES

1. Let x be the number of successes in n trials of an independent trials

process with probability p for success. Let $x^* = \frac{x - np}{\sqrt{npq}}$. For large

n estimate the following probabilities:

- (a) $\Pr[x^* < -2.5]$.
- (b) $\Pr[x^* < 2.5].$ (c) $\Pr[x^* \ge -.5].$
- (d) $\Pr[1 [x \ge -...]]$

(d) $\Pr[-1.5 < x^* < 1]$. [Ans. .774.]

- 2. A coin is biased in such a way that a head comes up with probability .8 on a single toss. Use the normal approximation to estimate the probability that in a million tosses there are more than 800,400 heads.
- 3. Plot a graph of the probabilities f(10, x; .5). Plot a graph also of the normalized probabilities as in Figures 4 and 5.
- 4. An ordinary coin is tossed 1 million times. Let x be the number of heads which turn up. Estimate the following probabilities:

- (a) $\Pr[499,500 \le x \le 500,500].$
- **(b)** $\Pr[499,000 \le x \le 501,000].$
- (c) $\Pr[498,500 \le x \le 501,500].$

[Ans. .682; .954; .997 (approximate answers).]

- 5. Assume that a baseball player has probability .37 of getting a hit each time he comes to bat. Find the probability of getting an average of .388 or better if he comes to bat 300 times during the season. (In 1957 Ted Williams had a batting average of .388 and Mickey Mantle had an average of .353. If we assume this difference is due to chance, we may estimate the probability of a hit as the combined average, which is about .37.) [Ans. .242.]
- 6. A true-false examination has 48 questions. Assume that the probability that a given student knows the answer to any one question is $\frac{3}{4}$. A passing score is 30 or better. Estimate the probability that the student will fail the exam.
- 7. In Example 3 of Section 9 in Chapter 3, assume that the school decides to admit 1296 students. Estimate the probability that they will have to have additional dormitory space. [Ans. Approximately .115.]
- 8. Peter and Paul each have 20 pennies. They each toss a coin and Peter wins a penny if his coin matches Paul's, otherwise he loses a penny; they do this 400 times, keeping score but not paying until the 400 matches are over. What is the probability that one of the players will not be able to pay? Answer the same question for the case in which Peter has 10 pennies and Paul has 30.
- 9. In tossing a coin 100 times, the probability of getting 50 heads is, to three decimal places, .080. Estimate this same probability using the central limit theorem. [Ans. .080.]
- 10. A standard medicine has been found to be effective in 80 percent of the cases where it is used. A new medicine for the same purpose is found to be effective in 90 of the first 100 patients on which the medicine is used. Could this be taken as good evidence that the new medication is better than the old?
- 11. Two railroads are competing for the passenger traffic of 1000 passengers by operating similar trains at the same hour. If a given passenger is equally likely to choose one train as the other, how many seats should the railroad provide if it wants to be sure that its seating capacity is sufficient in 99 out of 100 cases? [Ans. 537.]

3 TEST OF HYPOTHESES

We turn now to our first typical statistical problem. As we indicated in the introductory section, our problem is often to decide between two or more competing probability measures. We shall illustrate this in terms of an example.

EXAMPLE Smith claims that he has the ability to distinguish ale from beer and has bet Jones a dollar to that effect. Now Smith does not mean that he can distinguish beer from ale with 100 percent accuracy, but rather that he believes that he can distinguish them a proportion of the time which is significantly greater than $\frac{1}{2}$.

Assume that it is possible to assign a number p which represents the probability that Smith can pick out the ale from a pair of glasses, one containing ale and one beer. We identify $p = \frac{1}{2}$ with his having no ability, $p > \frac{1}{2}$ with his having some ability, and $p < \frac{1}{2}$ with his being able to distinguish, but having the wrong idea which is the ale. If we knew the value of p, we would award the dollar to Jones if p were $\leq \frac{1}{2}$, and to Smith if p were $>\frac{1}{2}$. As it stands, we have no knowledge of p and thus cannot make a decision. We perform an experiment and make a decision as follows.

Smith is given a pair of glasses, one containing ale and the other beer, and is asked to identify which is the ale. This procedure is repeated ten times, and the number of correct identifications is noted. If the number correct is at least eight, we award the dollar to Smith, and if it is less than eight, we award the dollar to Jones.

We now have a definite procedure and shall examine this procedure from both Jones's and Smith's points of view. We can make two kinds of errors. We may award the dollar to Smith when in fact the appropriate value of p is $\leq \frac{1}{2}$, or we may award the dollar to Jones when the appropriate value for p is $>\frac{1}{2}$. There is no way that these errors can be completely avoided. We hope that our procedure is such that each of the bettors will be convinced that, if he is right, he will very likely win the bet.

Jones believes that the true value of p is $\frac{1}{2}$. We shall calculate the probability of Jones winning the bet if this is indeed true. We assume that the individual tests are independent of each other and all have the same probability $\frac{1}{2}$ for success. (This assumption will be unreasonable if the glasses are too large.) We have then an independent trials process with $p = \frac{1}{2}$ to describe the entire experiment. The probability that Jones will win the bet is the probability that Smith gets fewer than eight correct. From the table in Figure 2 we compute that this probability is approximately .945. Thus Jones sees that, if he is right, it is very likely that he will win the bet.

Smith, on the other hand, believes that p is significantly greater than $\frac{1}{2}$. If he believes that p is as high as .9, we see from Figure 2 that the probability of his getting eight or more correct is .930. Then both men will be satisfied by the bet.

Suppose, however, that Smith thinks the value of p is only about .75. Then the probability that he will get eight or more correct and thus win the bet is .526. There is then only an approximately even chance that the experiment will discover his abilities, and he probably will not be satisfied with this. If Smith really thinks his ability is represented by a p value of about $\frac{3}{4}$, we would have to devise a different method of awarding the dollar. We might, for example, propose that Smith win the bet if he gets seven

or more correct. Then, if he has probability $\frac{3}{4}$ of being correct on a single trial, the probability that he will win the bet is approximately .776. If $p = \frac{1}{2}$, the probability that Jones will win the bet is about .828 under this new arrangement. Jones's chances of winning are thus decreased, but Smith may be able to convince him that it is a fairer arrangement than the first procedure.

In the theory of hypothesis testing it is common to refer to one hypothesis, say $p = \frac{1}{2}$, as the null hypothesis H_0 , and an alternate hypothesis as H_1 .

In the above example, it was possible to make two kinds of errors. The probability of making these errors depended on the way we designed the experiment and the method we used for the required decision. In some cases we are not too worried about the errors and can make a relatively simple experiment. In other cases, errors are very important, and the experiment must be designed with that fact in mind. For example, the possibility of error is certainly important in the case that a vaccine for a given disease is proposed and the statistician is asked to help in deciding whether or not it should be used. In this case it might be assumed that there is a certain probability p that a person will get the disease if not vaccinated and a probability r that he will get it if he is vaccinated. If we have some knowledge of the approximate value of p, we are then led to construct an experiment to decide whether r is greater than p, equal to p, or less than p. The first case would be interpreted to mean that the vaccine actually tends to produce the disease, the second that it has no effect, and the third that it prevents the disease; so that we can make three kinds of errors. We could recommend acceptance when it is actually harmful, we could recommend acceptance when it has no effect, or finally we could reject it when it actually is effective. The first and third might result in the loss of lives, the second in the loss of time and money of those administering the test. Here it would certainly be important that the probability of the first and third kinds of errors be made small. To see how it is possible to make the probability of both errors small, we return to the case of Smith and Jones.

EXAMPLE Suppose that, instead of demanding that Smith make at least eight correct identifications out of ten trials, we insist that he make at least 60 correct identifications out of 100 trials. (The glasses must now be very small.) Then, if $p = \frac{1}{2}$, the probability that Jones wins the bet is about .98; so that we are extremely unlikely to give the dollar to Smith when in fact it should go to Jones. (If $p < \frac{1}{2}$, it is even more likely that Jones will win.) If $p > \frac{1}{2}$, we can also calculate the probability that Smith will win the bet. These probabilities are shown in the graph in Figure 7. The dashed curve gives for comparison the corresponding probabilities for the test requiring eight out of ten correct. Note that with 100 trials, if p is $\frac{3}{4}$, the probability that Smith wins the bet is nearly 1, while in the case of eight out of ten, it was only about $\frac{1}{2}$. Thus in the case of 100 trials, it would be easy to convince

Statistics 285



both Smith and Jones that whichever one is correct is very likely to win the bet.

Thus we see that the probability of both types of errors can be made small at the expense of having a large number of experiments.

In applications it is important to have some estimate of the number of experiments that are necessary to reduce the probabilities of errors to acceptable levels. Assume, for example, that we are trying to decide for an independent trials process whether the true probability is p_0 or p_1 . Assume that $p_0 < p_1$. We want to design a test so that the probability of error under either hypothesis is at most a. We choose a number s so that the area under the normal curve beyond s is a. We perform n experiments. If p_0 is correct, the probability of the number of successes x exceeding the expected number np_0 by s standard deviations is a. That is, if p_0 is correct, then

$$\Pr\left[x > np_0 + s\sqrt{np_0q_0}\right] = a.$$

On the other hand, if p_1 is correct, the probability that the number of successes will be more than s standard deviations below the expected value of np_1 is also a. That is, if p_1 is correct, then

$$\Pr\left[x < np_1 - s\sqrt{np_1q_1}\right] = a.$$

Assume, then, that we can choose n so that

$$np_0 + s\sqrt{np_0q_0} < np_1 - s\sqrt{np_1q_1}.$$

Then we can choose a value t greater than the first number but such that t-1 is less than the second number. We accept p_0 if $x \le t - 1$ and p_1 if $x \ge t$. The test will have a probability of error of at most a under either hypothesis. We can achieve this inequality if

 $\sqrt{n}p_0 + s\sqrt{p_0q_0} < \sqrt{n}p_1 - s\sqrt{p_1q_1}$

or

$$s\left[\frac{\sqrt{p_0q_0} + \sqrt{p_1q_1}}{p_1 - p_0}\right] < \sqrt{n}$$

or

$$n > s^2 \left[\frac{\sqrt{p_0 q_0} + \sqrt{p_1 q_1}}{p_1 - p_0} \right]^2.$$

For example, in our beer and ale example, assume that $p_0 = .5$ and $p_1 = .75$. We would like to be 90 percent certain of being correct. Then from Figure 6 (Section 2) we see that s = 1.3. Thus we must have

$$n > (1.3)^2 \left[\frac{\sqrt{.5 \times .5} + \sqrt{.75 \times .25}}{.75 - .5} \right]^2 = 23.5.$$

We would need only a moderate number of experiments, namely 24. Then $np_0 + s\sqrt{np_0q_0} = 15.18$ and $np_1 - s\sqrt{np_1q_1} = 15.24$. Jones is 90 percent sure that Smith will have fewer than 16 correct guesses, while Smith is 90 percent sure that he will have more than 15 correct guesses. Thus we award the bet to Smith if he guesses correctly at least 16 times out of 24 experiments.

Consider, however, the Salk vaccine experiment. In this experiment we want to test $p_0 = .00025$ against $p_1 = .00050$ —that is, whether the vaccine will reduce the incidence of polio from 50 to 25 per 100,000. We would want a great deal of reliability for such a test. Let us choose s so that the probability of error is less than .001. We can have this by choosing s = 3.1. Then we must have

$$n \ge (3.1)^2 \times \left[\frac{\sqrt{.00025 \times .99975} + \sqrt{.0005 \times .9995}}{.00025}\right] = 223,956.$$

In one of the major parts of the Salk vaccine experiment the vaccine was given to 200,000 students. Of these vaccinated students 57 contracted polio. In Exercise 10 you are asked to design an experiment to test the hypothesis $p_1 = .00050$ against the hypothesis $p_0 = .00025$.

EXERCISES

- 1. Assume that in the beer and ale experiment Jones agrees to pay Smith if Smith gets at least nine out of ten correct.
 - (a) What is the probability of Jones paying Smith even though Smith cannot distinguish beer and ale, and guesses? [Ans. .011.]
 - (b) Suppose that Smith can distinguish with probability .9. What is the probability of his not collecting from Jones? [Ans. .264.]
- 2. Suppose that in the beer and ale experiment Jones wishes the probability to be less than .1 that Smith will be paid if, in fact, he guesses. How many of ten trials must he insist that Smith get correct to achieve this?
- 3. In the analysis of the beer and ale experiment, we assume that the various trials were independent. Discuss several ways that error can enter, because of the nonindependence of the trials, and how this error

can be eliminated. (For example, the glasses in which the beer and ale were served might be distinguishable.)

- 4. Consider the following two procedures for testing Smith's ability to distinguish beer from ale.
 - (a) Four glasses are given at each trial, three containing beer and one ale, and he is asked to pick out the one containing ale. This procedure is repeated ten times. He must guess correctly seven or more times.
 - (b) Ten glasses are given him, and he is told that five contain beer and five ale, and he is asked to name the five which he believes contain ale. He must choose all five correctly.

In each case, find the probability that Smith establishes his claim by guessing. Is there any reason to prefer one test over the other?

[Ans. (a) .003; (b) .004.]

- 5. A testing service claims to have a method for predicting the order in which a group of freshmen will finish in their scholastic record at the end of college. The college agrees to try the method on a group of five students, and says that it will adopt the method if, for these five students, the prediction is either exactly correct or can be changed into the correct order by interchanging one pair of *adjacent* men in the predicted order. If the method is equivalent to simply guessing, what is the probability that it will be accepted? [Ans. $\frac{1}{24}$.]
- 6. The standard treatment for a certain disease leads to a cure in $\frac{1}{4}$ of the cases. It is claimed that a new treatment will result in a cure in $\frac{3}{4}$ of the cases. The new treatment is to be tested on ten people having the disease. If seven or more are cured, the new treatment will be adopted. If three or fewer people are cured, the treatment will not be considered further. If the number cured is four, five, or six, the results will be called inconclusive, and a further study will be made. Find the probabilities for each of these three alternatives first, under the assumption that the new treatment has the same effectiveness as the old, and second, under the assumption that the claim made for the treatment is correct.
- 7. Three upperclassmen debate the intelligence of the freshmen class. One claims that most freshmen (say 90 percent of them) are intelligent. A second claims that very few (say 10 percent) of them are intelligent, while a third one claims that a freshman is just as likely to be intelligent as not. They administer an intelligence test to ten freshmen, classifying them as intelligent or not. They agree that the first man wins the bet if eight or more are intelligent, the second if two or fewer, the third in all other cases. For each man, calculate the probability that he wins the bet, if he is right.
- 8. Ten men take a test with ten problems. Each man on each question has probability $\frac{1}{2}$ of being right, if he does not cheat. The instructor determines the number of students who get each problem correct. If he finds on four or more problems there are fewer than three or more

than seven correct, he considers this convincing evidence of communication between the students. Give a justification for the procedure. [*Hint:* The table in Figure 2 must be used twice, once for the probability of fewer than three or more than seven correct answers on a given problem, and the second time to find the probability of this happening on four or more problems.]

- 9. An instructor claims that a certain student knows only 70 percent of the material. The student claims that he knows 85 percent. Design a test that will settle the argument with probability .9.
- [Ans. 50 questions, student must get 40 correct answers.] 10. Assume that the Salk vaccine is to be given to 225,000 students. It is claimed that the probability of getting polio is $\leq .00025$ if vaccinated and .00050 if not vaccinated. Design a test to decide between these two alternatives. In the actual experiment there were 28 cases per 100,000 of polio among the 200,000 vaccinated. This would suggest 63 cases in 225,000 students. Would your test establish the claim that the Salk vaccine was effective, if this few cases of polio occurred in the experiment?

4 CONFIDENCE INTERVALS

Consider *n* independent trials with probability *p* for success on each trial. We assume that we do not know *p* but want to make, on the basis of our observations, some estimate of *p*. Let *a* be any number between 0 and 1. Then from Figure 6 we can find a number *s* such that the area under the normal curve beyond *s* is a/2. For example, if a = .05 then we can choose s = 2. By the central limit theorem, if *x* is the number of successes, then

$$\Pr\left[\left|\frac{x-np}{\sqrt{npq}}\right| \le s\right] \approx 1-a.$$

This is the same as saying that

$$\Pr\left[\left|\frac{x/n-p}{\sqrt{pq/n}}\right| \le s\right] \approx 1-a.$$

Putting $\overline{p} = x/n$, we have

$$\Pr\left[|\overline{p}-p| \le s\sqrt{pq/n}\right] \approx 1-a.$$

Using the fact that $pq \leq \frac{1}{4}$ for all p (see Exercise 9), we have

$$\Pr\left[|\overline{p} - p| \le s/2\sqrt{n}\right] \ge 1 - a.$$

Thus, no matter what p is, with probability at least 1 - a, the true value will not deviate from \overline{p} by more than $s/2\sqrt{n}$. We say then that

$$\overline{p} - \frac{s}{2\sqrt{n}} \le p \le \overline{p} + \frac{s}{2\sqrt{n}}$$

.63

with confidence 1 - a. We call the interval

$$\left[\overline{p} - \frac{s}{2\sqrt{n}}, \ \overline{p} + \frac{s}{2\sqrt{n}}\right]$$

a 100(1 - a) percent confidence interval. For example, the 95 percent confidence interval requires s = 2, and hence is $\left[p - \frac{1}{\sqrt{n}}, p + \frac{1}{\sqrt{n}}\right]$.

For example, if in 400 trials a drug is found effective 124 times or .31 of the time, the 95 percent confidence interval for p is

$$\left[.31 - \frac{1}{20}, \ .31 + \frac{1}{20}\right],$$

or [.26, .36]. The 99 percent confidence interval would be found by using s = 2.6. This gives

$$\left[.31 - \frac{2.6}{40}, \ .31 + \frac{2.6}{40}\right],$$

or [.245, .375]. Of course, as we demand more confidence our prediction is more conservative, i.e., the interval is larger.

It is important to realize that the interval obtained depends upon the value .63 .83 of \overline{p} , which in turn depends upon the value of x. Thus \overline{p} is a chance quantity. .59 .79 We are assuming that the true value p, though unknown, is not a chance .67 .87 quantity. Thus our confidence interval itself is a chance quantity which may .65 .85 or may not cover the true value p. When we choose a 95 percent confidence .59 .79 interval we mean that the probability is .95 that the interval will cover the .65 true value p. Thus by the law of large numbers we expect this to be the .85 .61 .81 case about 95 percent of the time.

.51 .71 In Figure 8 we give the results of computing the 95 percent confidence .81 .61 intervals based upon several experiments with n = 100 trials for a true value .57 .77 of p = .7. We carried out this experiment 20 times. It will be noted that .58 in each case the interval does include the true value, though sometimes just .78 .59 .79 barely. We should not have been surprised if in one or two cases it did .61 .81 not.

.60 .80 The use of the inequality $pq \leq \frac{1}{4}$ was for convenience and simplicity of our computations. It results in slightly larger confidence intervals than are .68 .88 .68 .88 necessary for a given confidence level. Without making this approximation it is possible to transform the first inequality into an inequality about \overline{p} to .66 .86 .56 .76 obtain a more exact confidence interval (see Exercise 14). .60

.80 As a second example of confidence intervals consider the following prob-.83 lem. In a small town lottery tickets numbered from 1 to N are being sold Figure 8 weekly and a prize is given to the person who holds the ticket having the lucky number drawn at random from the numbers from 1 to N. The value of N is not publicly announced, but is the same every week. A man buys lottery tickets for ten weeks, receiving numbers 27, 46, 77, 85, 34, 24, 34, 46, 34, and 89. Before buying a ticket the following week, he wants to

 n	$\frac{1}{(.05)^{1/n}}$
5	1.821
6	1.648
7	1.534
8	1.454
9	1.395
10	1.349

Figure	9
--------	---

M	<i>M</i> /(.05 ^{.1})
100) 134
96	129
89	120
88	118
99	133
93	125
97	130
85	114
74	99
99	133
82	110
98	132
97	130
91	122
93	125
96	129
97	130
90	121
91	122
98	132
Figur	e 10

estimate his chance of winning; i.e., he wants to estimate N. Of course he knows that N is at least 89, the highest number that he has drawn.

Let us see how we would obtain confidence intervals for the unknown "parameter" N. The man has in effect drawn a number from the N possible numbers n times. Let M be the maximum of the numbers drawn. Then for fixed n and N, M may be considered a chance quantity. For any A,

$$\Pr\left[M \le A\right] = \left(\frac{A}{N}\right)^n.$$

As before, let a be any number between 0 and 1. We can choose A so that $A = a^{1/n}N$. Then

$$\Pr[M \le a^{1/n}N] = \frac{(a^{1/n}N)^n}{N^n} = a$$

 $\Pr\left[\frac{M}{a^{1/n}} \le N\right] = a.$

That is,

or

$$\Pr\left[N < \frac{M}{a^{1/n}}\right] = 1 - a.$$

Since $M \leq N$, we can write this as

$$\Pr\left[M \le N < \frac{M}{a^{1/n}}\right] = 1 - a.$$

Thus the interval $[M, M/a^{1-n}]$ has probability 1 - a of covering N and hence is a 100(1 - a) percent confidence interval for N. In any given example, for a 95 percent confidence interval, we choose a = .05 and hence $M/a^{1/n} = 89/(.5)^{1/10} = 120.1$. Hence the man can be 95 percent sure that there are at most 120 lottery tickets.

For such calculations the table in Figure 9 is useful.

In Figure 10 we have indicated the result of twenty experiments with N equal in each case to 100 and n = 10. We have computed the 95 percent confidence intervals. In this case we see that one interval does not include the true value of N. Thus the intervals include the true value of N precisely 95 percent of the time.

EXAMPLE The Fish and Game Department is interested in estimating the number of trout in a pond (which contains only trout). They take out a sample of 1000 fish and mark them. Later they take another sample of 1600 and find that 120 of them are marked. What is a reasonable estimate for the total number of trout?

Let *n* be the unknown total. Since 1000 of them were marked, there is probability p = 1000/n that a fish in the second sample will be marked.

The observed fraction is $\overline{p} = 120/1600$, and the 95 percent confidence interval yields

$$\frac{120}{1600} - \frac{1}{\sqrt{1600}} \le p \le \frac{120}{1600} + \frac{1}{\sqrt{1600}},$$

or

$$\frac{3}{40} - \frac{1}{40} \le p \le \frac{3}{40} + \frac{1}{40}$$

or

$$\frac{1}{20} \le p \le \frac{1}{10}.$$

Hence

$$\frac{1}{20} \le \frac{1000}{n} \le \frac{1}{10},$$

and we obtain the estimate $10,000 \le n \le 20,000$.

EXERCISES

- 1. A prospective college student visits a college and sits in on a class of 50 students. She notes that there are 39 men and 10 women in the class. She decides to compute the 95 percent confidence interval for the proportion of women in the school. She will reject the school if this interval excludes the possibility that $\frac{1}{3}$ of the students are women. Does she reject the school for this reason?
- 2. A young ballplayer in his first season is at bat 400 times and gets 100 hits for a batting average of .250. Find 90 percent confidence limits for his batting average based upon his first season. Is it reasonable to believe that he may in fact be a .300 batter?

[Ans. [.209, .291]; maybe, next year.]

- 3. A large company has as many as a million accounts. It wishes to estimate the number that are at least three months delinquent in their payments. A thousand accounts are randomly selected and of these it is observed that 30 are at least three months delinquent. Find the 95 percent confidence limits for the proportion of customers that are at least three months behind in their payments.
- 4. Opinion pollsters in election years usually poll about 3000 voters. Suppose that in an election year 51 percent favor candidate A and 49 percent favor candidate B in a poll. Construct 95 percent confidence limits on the true percentage of the population in favor of A. [Ans. .492, .528.]
- 5. An experimenter has an independent trials process and she has a hypothesis that the true value of p is p_0 . She decides to carry out a number of trials, and from the observed \overline{p} calculate the 95 percent

confidence interval of p. She will reject p_0 if it does not fall within these limits. What is the probability that she will reject p_0 when in fact it is correct? Should she accept p_0 if it does fall within the confidence interval?

- 6. A coin is tossed 100 times and turns up heads 61 times. Using the method of Exercise 5, test the hypothesis that the coin is a fair coin. [Ans. Reject.]
- 7. In an experiment with independent trials we are going to estimate p by the fraction \overline{p} of successes. We wish our estimate to be within .02 of the correct value with probability .95. Show that 2500 observations will always suffice. Show that if it is known that p is approximately .1, then 900 observations would be sufficient.
- 8. In the Weldon dice experiment, 12 dice were thrown 26,306 times and the appearance of a 5 or a 6 was considered to be a success. The mean number of successes observed was, to four decimal places, 4.0524. Is this result significantly different from the expected average number of 4?
 [Ans. Yes.]
- 9. Prove that $pq \leq \frac{1}{4}$. [*Hint*: write $p = \frac{1}{2} + x$.]
- 10. Suppose that out of 1000 persons interviewed 650 said that they would vote for Mr. Big for mayor. Construct the 99 percent confidence interval for p, the proportion in the city that would vote for Mr. Big.
- 11. In a pond 400 fish are marked. If in a subsequent sample of 225 there are 45 marked fish, find the 90 percent confidence interval for the total number of fish.
- 12. In a large city each taxi is assigned a number. A man observes the numbers 125, 135, 356, 344, 25, 299, and 320 on seven occasions that he takes a cab. On the basis of this, compute the 95 percent confidence limits for the number of cabs in the city. If he knows that the number of cabs is a multiple of 100, can be determine the total?
- 13. Suppose that the man in Exercise 12 takes three more cabs numbered 76, 421, and 211. Can he be 95 percent sure of the total?

[Ans. Yes.]

14. In this section we have approximated confidence limits on p such that $\Pr\left[|\bar{p} - p| \le s \sqrt{\frac{pq}{n}}\right] = 1 - a$. The expression inside the brackets

is equivalent to $(\overline{p} - p)^2 \le s^2 \left(\frac{p(1-p)}{n} \right)$. Substituting equality for

inequality we obtain a quadratic equation which can be solved for p in terms of \overline{p} , s, and n. There will be two roots r_1 and r_2 , where $r_1 \leq r_2$ and the above inequality will be satisfied for all p such that $r_1 \leq p \leq r_2$. Use this information to obtain more exact confidence intervals than that obtained by setting $pq = \frac{1}{4}$.

15. A hundred names are picked at random out of a large telephone book. It is found that 70 of these names have eight letters or less. Place 95 .586

.591

.594

.59

.594

.598

percent confidence limits on the fraction of names in that telephone book containing eight letters or less:

- (a) Using the estimate developed in the text.
- (b) Using the limits developed in exercise 14. [Ans. .602, .783.]
- (c) Suppose we were using the method of Exercise 5 to test the hypothesis that 79 percent of the names have eight or less letters. Which of the above intervals would be better?

5 SOME PITFALLS

Statistics properly used is a very powerful tool. If it is not properly used it can lead to incorrect predictions and thereby cause considerable distrust in its methods. We have already mentioned the example of the poll of the .618 Literary Digest in the 1936 presidential election between Roosevelt and .623 Landon. In this poll about 10 million postcards were sent to persons whose .626 names were obtained from telephone directories and car registrations. .622 Several million cards were returned, with 40.9 percent in favor of Franklin .626 Roosevelt. A few weeks later in the actual election, Roosevelt obtained 60.7 .63 percent of the vote. .623

.591 There are two obvious flaws in the above procedure. The first, and the .591 .623 one which is normally blamed for the error, is that people who had tele-.592 .625 phones or cars at that time were not truly representative of the voting .595 .627 population as a whole. The second is the possibility that people change .59 .622 their minds between the time a poll is taken and the election takes place. .583 .616 They may even deliberately tell the poll taker one thing and vote another. .588 .62 Of course, with many millions of people it is difficult to choose a truly .591 .623 random sample. However, let us assume that there were in fact 60.7 percent .603 .635 of the people in favor of Roosevelt at the time of the poll and that we could .59 .622 choose a random sample of only 10,000 voters. In Figure 11 we indicate .592 .624 the result of simulating 30 such samples and determining the 99 percent .597 .629 confidence intervals. We see that in every case we would have picked .595 .627 Roosevelt to win. This is on the basis of only 10,000 samples rather than .588 .62 the millions which led to a wrong answer. Thus if statistics can be properly Figure 11 used it is a very powerful tool.

Because of the difficulties indicated above there is still, with some justice, skepticism of polls. However, there is also some danger in refusing to use statistical methods. For example, assume that an all-male college wishes to know the opinion of its alumni on the question of becoming a coeducational institution. Assume that there are 30,000 alumni and in fact 60 percent are in favor of the college admitting women. This is a situation in which any person not asked could conceivably challenge the poll. Assume that it is decided to poll by mail *all* the alumni. Also assume that a proportion p of those who favor coeducation will respond to the query and a proportion 2p of those who oppose will respond because they feel more strongly about the matter. Then the expected number of yes answers would

be 18,000p and the no answers would be 24,000p. Thus, neglecting sampling errors, the vote would be 18,000p/42,000p = .43 in favor, and coeducation would be defeated. On the other hand, as we have seen, a relatively small random sample in which the response of each person sampled was recorded would give a much more reliable indication of the true feelings of the alumni. For example, a sample of 1000 was taken in such a poll and a 95 percent confidence interval of (.559, .621) was obtained. Here is a situation where one could also use the method of hypothesis testing discussed in Section 3.

As we have indicated previously, a number of precautions had to be taken in the experiment to test the effectiveness of Salk vaccine. First, although initially the incidence of polio was only about 50 per 100,000, there was considerable variability from year to year and from region to region. So a reduction from 50 to 25 in a sample of 100,000 could easily be caused by reasons having nothing to do with the effectiveness of the vaccine. Thus it was decided to have control groups. In one part of the experiment a population of students was divided into two groups of about 200,000 each. All were inoculated at the same time. The first group received the Salk vaccine and the second a harmless and useless salt solution (a "placebo"). The decision as to which students received the real vaccine was made randomly, and the knowledge of whether a student was given the vaccine or the placebo was not made known to the student or to the physician observing the student. The reason for this is that in such experiments knowledge of whether the subject has been treated or not has been found to introduce a bias in the diagnosis and in the behavior of the subject. As indicated earlier, the test did show a significantly lower rate among those vaccinated. The test led to further development of vaccines and the virtual elimination of polio in the United States.

There have been a large number of statistical studies to determine if smoking is injurious to one's health. It is now widely believed that this is the case. However, the problem of establishing this has been exceedingly difficult, and there are still statisticians who feel that more testing must be done. In the case of the polio vaccine it was possible to select two groups and randomly give one half the vaccine and the other half a placebo. Random selection eliminates the effect of biases which can creep in, such as differences in age, place of residence, economic status, etc. To do the corresponding experiment for smoking would require one randomly selected group to become heavy smokers and the rest to abstain. This is clearly not possible, and many of the studies have had to rely on choosing groups in a less random way and studying their smoking and health patterns. In the exercises some of these methods are briefly mentioned and you are asked to consider possible pitfalls. While statisticians who criticize these tests or refuse to accept their conclusions are often accused of being overly cautious, their criticisms have led to the development of more careful methods of statistical tests in these very difficult areas. It should be emphasized that demonstrating that more heavy smokers than nonsmokers get lung cancer

does not demonstrate that smoking is a cause of lung cancer. It seems likely that there will still be controversy about this question until more knowledge is obtained as to what is the essential cause of cancer.

In the *Literary Digest* poll the particular people that responded to the postcard inquiry was a chance quantity. In effect, the *size* of the sample was random. While pollsters do not intentionally take advantage of this, the results under such circumstances can be distorted. We illustrate this in an extreme case where the experimenter deliberately tries to take advantage of the randomness of the size of the experiment.

Assume that Mr. Esp claims that he has extrasensory perception. An experiment is arranged in which he is to tell, when a card is placed face down, whether it has a circle or a square on it. Of course we would want to run a large number of experiments, but for the point we are trying to make we can take a small number, say four. If Mr. Esp is just guessing, we can find his expected score (percentage correct) in the usual manner. The tree and tree measure are shown in Figure 12, and his expected score is $1 \times \frac{1}{16} + \frac{3}{4} \times \frac{4}{16} + \frac{1}{2} \times \frac{6}{16} + \frac{1}{4} \times \frac{4}{16} + 0 \times \frac{1}{16} = \frac{1}{2}$.



Figure 12

Assume now that the experimenter, eager to find a good subject, stops the experiment the first time (if any) that Mr. Esp's score is greater than $\frac{1}{2}$. Then the new tree measure, still assuming guessing, is shown in Figure 13.

We see now that his expected score is

$$\frac{1}{2} \times \frac{1}{2} + \frac{2}{3} \times \frac{1}{8} + \frac{1}{2} \times \frac{1}{8} + \frac{1}{4} \times \frac{3}{16} + 0 \times \frac{1}{16} = \frac{133}{192} = .69,$$

which is considerably better than before.

It is extremely important in designing a statistical test to decide upon the criteria for acceptance or rejection before the test is carried out. Of course, we should not be surprised if we find some unlikely feature of an experiment by looking after the fact for something of small probability. A local expert on probability theory would occasionally be roused from bed at one in the



Figure 13

morning to have an excited colleague ask, "What is the probability of being dealt a hand of all hearts in bridge." He would answer, "The same as any other hand," and then go back to sleep.

EXERCISES

- 1. In the tests of the Salk vaccine 400,000 students volunteered to be vaccinated. Half were vaccinated and half given placebo. Among these 400,000 people 199 got polio. Assuming that a person who had polio was equally likely to be in either of the groups, place 99 percent confidence limits on the number of people in the vaccinated group that got polio. Does the fact that of the 199 reported cases 142 were in the placebo group suggest that the vaccine was effective?
- 2. Referring to Exercise 1, data was taken also on 340,000 students who did not volunteer to be inoculated. Assuming that these people had the same probability of getting polio as those who received the placebo, place 99 percent confidence limits on the number of people among this group to get polio. What does the fact that among this group there were 157 polio cases suggest? How might we explain this result?
- 3. In many of the major studies of smoking and health the samples are obtained by interviewing whomever happens to be at home when the interviewer calls. This person answers questions relating to everyone in the family over 21. Comment on some possible defects in this method of sampling.
- 4. In one major study of smoking and health two groups were compared, one that had lung cancer and another that was chosen by virtue of having similar backgrounds to the group that had cancer. Comment on this technique of sampling.
- 5. This exercise is designed to show that optional stopping in sampling can significantly change the results. Consider the following game. A box contains five balls, three of which are red and two blue. If a red ball is drawn we lose a dollar, if a blue ball we win a dollar.
 - (a) Find the expected value of the game if one ball is drawn.

[Ans. -20 cents.]

- (b) Show that the game becomes increasingly unfavorable if two, three, four, or five balls are drawn.
- (c) Show that the game is favorable if you are allowed to stop at any

time. Use the following strategy: If the first draw is blue, stop. Otherwise, play until you are even or until all five balls are drawn. [Partial ans. Value is +20 cents.]

- 6. In a certain college 25 out of 324 faculty members with Ph.D.'s are women. Nationally approximately 20 percent of all Ph.D.'s are awarded to women. Test the hypothesis that the faculty members were picked from the national pool without regard to sex.
- 7. In regarding the significance of Exercise 6 as evidence of discrimination, what other factors would have to be taken into account? For example, is it sufficient to know the present percentage of women among Ph.D.'s? And should one know something about the distribution among disciplines?
- 8. The President of the United States announces a major policy decision. His mail the following week contains 25,000 irate letters and 10,000 favoring his decision. Would it be reasonable to conclude that a majority of people oppose his decision?

6 COMPUTER APPLICATIONS

In many problems in statistics the theory is straightforward but the computations are very difficult. This makes statistics an important area for computer applications. We shall first illustrate this by the computation of confidence intervals considered earlier in this chapter, then we shall introduce the important technique of *simulation*.

In Section 4, for the example of the lottery ticket, the only difficulty in computing the confidence interval for the total number of tickets is the necessity of raising a decimal fraction to the 1/n power. This is simply done in BASIC. In the program LOTTERY we supply N (the number of tickets bought), M (the largest number observed), and C (the percentage confidence

LOTTERY

```
10 READ N,M,C

20 LET A = 1-C

30 PRINT M, M/A+(1/N)

90 DATA 10,89,.95

99 END

READY

RUN

LOTTERY

89 120.086

0.060 SEC.

READY
```

desired). The entire computation is carried out in two instructions. The RUN shows the 95 percent confidence interval if among ten tickets bought the highest number was 89.

The program CONFIDE computes three confidence intervals for an independent trials experiment in which we observed x successes in n trials. The

CONFIDE

```
10 READ X,N

20 LET P = X/N

30 LET D = 2*SQR(N)

40 FOR K = 1 TO 3

50 READ S

60 PRINT P-S/D,P+S/D

70 NEXT K

80 DATA 61,100

90 DATA 1.65,2,2.6

99 END

READY
```

RUN

CONFIDE

ؕ52 75	0.6925
ؕ51	0.71
0•48	0.74

```
Ø•Ø7Ø SEC•
READY
```

DATA in Line 90 supplies the number of standard deviations for 90, 95, and 99 percent confidence. The rest of this simple program is a direct translation of the formula obtained in Section 4. The RUN shows that if we observe 61 successes in 100 trials we can be 90 percent sure that the true value of p is below .7, but we cannot be 95 percent sure. It also shows that we can be 95 percent sure that p > .5, but not 99 percent sure.

Other formulas in this chapter may similarly be translated into simple computer programs. (See the Exercises.)

Probabilistic models prevail in the social sciences. While many of them can, in principle, be treated by the methods studied in this book, in practice they frequently are much too complicated to obtain precise theoretical results. In such cases, simulation by a high-speed computer may be a powerful tool. Simulation is a process during which the computer acts out a situation from real life. Typically, the relevant facts about an experiment are supplied to the computer, and it is instructed to run through a large series of experiments, perhaps under varying conditions. This enables the scientist to carry out in an hour a series of experiments that would otherwise take years, and at the same time all the important information is automatically tabulated by the computer.

Of course, the computer cannot duplicate the exact circumstances of an experiment. The facts fed to it are based on a model (or theory) formed by the scientist, and the value of the simulation depends on the accuracy of the model. Thus the main significance of simulation is that it enables a scientist to study the kind of behavior predicted by his model. For very complicated models this may be the only procedure open to him.

In addition to the use of simulation for theoretical studies, there are two very important types of pragmatic uses of simulation: (1) It can be used as a planning device. If there are various alternative courses of action open, the computer is asked to try out the various alternatives under different conditions, and report the advantages and disadvantages of each course. (2) Simulation may be used as a training device. For example, business schools are making increasing use of "business games" in which fledgling executives may try their skill at decision making under realistic circumstances.

We shall first discuss how computers simulate stochastic processes, and then illustrate simulation in terms of examples previously considered in this book. Simulation depends on the generation of so-called *random numbers*. In BASIC this is achieved by an instruction using "RND," such as

LET X = RND.

Every time this instruction is executed, BASIC generates a real number between 0 and 1 by a process that is reasonably random.

Actually, the computer is forced to cheat, in that it has only a finite capacity for expressing numbers. Thus it may in reality divide the unit interval into a million (or more) numbers, and give them in a quite random order. When its supply is exhausted, it will start again giving the same numbers in the same order. However, if one needs only 100,000 numbers, or even a million numbers, the results are highly satisfactory.

We illustrate this by means of the program RANDOM, which generates 30 random numbers. In looking at the output the reader should recall that E-2 indicates multiplication by 10^{-2} ; thus 8.5 E-2 = .085. The distribution is reasonably random. For example, 6 out of 30 numbers lie between .3 and .5, which is what we would expect. However, there are "too many" numbers between .2 and .3. Whether this is statistically significant needs to be checked (see Exercise 5).

Very often instead of random numbers we need random integers. For example, to simulate the roll of a die we need random integers from 1 to 6. We show the process for generating these in the other two RUNSs of

RUN				
RANDOM				
0.406533	0.272549	Ø.85Ø262	0.595677	9.421672
0.697993	0.251284	0.365032	61	.69229
0.757194	8.53103 E-2	0.244524	.2595	.61212
0.240634	0.540498	96	1594	.27508
0.351612	0.929495	ഗ	.592732	.32189
0.840201	0.111732	0.557491	3524	.94036
0.097 SEC.				

0.097 SEC READY

RANDOM

10 FOR I = 1 TO 30 20 LET X = RND 30 PRINT X, 40 NEXT I 99 END READY

20 LET X = 6*RND RUN

RANDOM

• 5 74 0	4.67751	.1209	.55621	.33211				すちこー ユー	
5.10157	• 4671	.1837	.8721	• 3449				ଦ୍ୟ ସ ଅନ୍ୟ	
1.63529	.51186	.2429	•57697	•6703		INT(6*RND)+1		こ こ ー 4 ら ー	
2.4392 4.18799	4.54316	I • 4438	2.12967	5.04121	Ø.Ø95 SEC. Ready	20 LET X = RUN	RANDOM	ოიაია	Ø.Ø90 SEC. Ready

005000

2.53883 4.15378 3.67273 1.6585 1.9314 5.64839

301

RANDOM. First we print out 6 times the random numbers. They are now spread evenly on the interval (0, 6). Thus if we take their integer parts, the numbers 0, 1, 2, 3, 4, 5 will turn up with equal probabilities, at random. By taking integer parts and adding one, we obtain the "roll of a die."

One common use of random numbers is to simulate an independent trials process. Such a process with p = .3 may be simulated as follows:

100 IF RND < .3 THEN 200 100 PRINT ''FAILURE'' ... 200 PRINT ''SUCCESS''

It is in the nature of the process that generates random numbers that the probability of RND < .3 is precisely .3. Of course any other probability may be used in place of .3.

EXAMPLE 1 Craps. Let us simulate the game of shooting craps. This is carried out by the program CRAPS, which closely follows the flow diagram in Figure 14, on page 304.

CRAPS

```
5
   FOR N = 1 TO 12
10 LET D1 = INT(6*RND)+1
20 LET D2 = INT(6*RND)+1
30 \text{ LET } D = D1 + D2
35 PRINT D;
40 IF D < 4 THEN 300
50 IF D = 12 THEN 300
60 \text{ IF } D = 7 \text{ THEN } 200
70
100 REM
          TRY TO MAKE POINT
110 \text{ LET } X = D
120 LET D1 = INT(6*RND)+1
130 \text{ LET } D2 = INT(6*RND)+1
140 LET D = D1 + D2
150 PRINT D;
160 IF D = X THEN 200
170 IF D = 7 THEN 300
180 GOTO 120
19Ø
200 REM
          PLAYER WINS
210 PRINT "YOU WIN".
220 GOTO 400
230
```

```
300 REM
          PLAYER LOSES
310 PRINT "YOU LOSE"
320
400 REM
          START OVER
420 NEXT N
430
999 END
READY
RUN
CRAPS
 5
    1Ø
         8
            5 YOU WIN
 10
     6
         4
            6
                7 YOU LOSE
 3
   YOU LOSE
 9
    5
        8
           5
               7 YOU LOSE
 4
    2
        11
            7 YOU LOSE
 8
        8 YOU WIN
    4
 7 YOU WIN
 2
  YOU LOSE
 5
    5 YOU WIN
 7 YOU WIN
0.117 SEC.
READY
```

One may consider running a program like CRAPS a large number of times, keeping count of the amount won or lost, and use it to estimate the expected value of the game. (In Chapter 3, Section 11, this was found to be -.0141.) Let us suppose that we try to simulate 10,000 games. How good an estimate can we expect? We know that the 95 percent confidence interval for a probability near .5 is $1/\sqrt{n}$. But if the fraction of successes is high by that amount, the fraction of losses will be low by the same amount, and vice versa. Thus we should expect errors up to $2/\sqrt{n}$ on either side of the expected value. For n = 10,000 this is an error of .02. In five such simulations the values obtained were: -.0238, -.0298, -.0090, +.0016, and -.0084. All are within the 95 percent confidence interval, but one simulation shows a loss twice the expected size and one actually shows a profit.

Thus, while a simulation provides an easy rough approximation to the answer, a good approximation requires a substantial computer effort. A simulation of 250,000 games requires about three minutes of computing time, much longer than the other examples we have shown in this book. Two such RUNs produced values of -.0115 and -.0154, which are much closer to the real value (see Exercise 4).



- Figure 14
- **EXAMPLE 2** *Poker.* In the exercises of Section 3 in Chapter 3 we computed the probabilities for various poker hands. Let us obtain estimates for the same by simulation.

Our problem here amounts to selecting five cards at random from a deck of 52 cards. We first number the cards from 1 to 52, in any convenient manner. Then we select one card by generating a random integer from the set 1 through 52. (This can be achieved by computing INT(52*RND + 1).) Next we select one of the 51 remaining cards at random, etc. When we have five cards, we determine how good a hand we drew.

This simulation was carried out for 10,000 poker hands on the Dartmouth Computer. The results were as in Figure 15.

You will be asked, in the exercises, to compare these figures with the expected values.

EXAMPLE 3 Land of Oz. Models in the social sciences often depend on Markov chain processes. While there are powerful theoretical tools for treating Markov chains, sufficiently complex models may have to be simulated. We shall

Type of hand	Number of times
Bust	5046
One pair	4169
Two pairs	508
Three of a kind	191
Straight	43
Flush	11
Full house	25
Four of a kind	6
Straight flush	1

Figure 15

illustrate this for a simple Markov chain, which we have already treated theoretically.

Consider the Land of Oz (Chapter 4, Section 7, Exercise 12). Suppose that we wished to find the fraction of times that the weather is "nice," "rain," or "snow," by simulation. We would first pick a starting state, say "rain." We then know that the probability of "rain" is $\frac{1}{2}$, of "nice" $\frac{1}{4}$, and of "snow" $\frac{1}{4}$. We can achieve this by generating an RND; if it is less than $\frac{1}{2}$ we decide on "rain," if it is between $\frac{1}{2}$ and $\frac{3}{4}$ then "nice" is next, while if RND $> \frac{3}{4}$ then "snow" is next.

The program RANDOMOZ carries out 1000 simulations for each starting state. After reading the transition probabilities, it starts a loop on S, the starting state. S1 is the current state. The list N is used for counting—e.g., N(1) is the total number of nice days. The only other comment needed is the explanation of line 90. Suppose that the probabilities of stepping into the three states is currently .25, .5, and .25. Then we should compare the random number successively with .25, .25 + .5 = .75, and .25 + .5 + .25 = 1. The same result may more simply be achieved by successive subtraction of .25, .5, and .25 until the number turns negative.

The program prints the number of times in each state for each starting state. While the values are reasonably close to the expected values of 200, 400, and 400, they are not close enough to be convincing. We show a second RUN with 10,000 simulations for each starting state and this time the fractions are much closer to the limiting probabilities .2, .4, and .4.

EXAMPLE 4 Central Limit Theorem. By simulating an independent trials process a large number of times we can hope to obtain an approximation of the central limit theorem. The program CLTH uses this method to approximate four values in Figure 6. Since the same distribution is obtained for any value of P, its choice is not crucial. The program uses P = .3. It carries out 100 experiments and counts the number of successes, noting how many standard deviations we are off the expected value. It repeats this 1000 times to get a frequency distribution.

RANDOMOZ

```
10 MAT READ P(3,3)
20 DATA 0,.5,.5
21 DATA .25,.5,.25
22 DATA .25..25..5
30 \text{ FOR S} = 1 \text{ TO } 3
40 \text{ LET S1} = \text{S}
50 LET N(1) = N(2) = N(3) = 0
60 \text{ FOR N} = 1 \text{ TO } 1000
70 LET X = RND
80 \text{ FOR I} = 1 \text{ TO } 3
90 LET X = X - P(S_{1}, I)
100 IF X<0 THEN 120
110 NEXT I
120 \text{ LET N(I)} = \text{N(I)}+1
130 \text{ LET } S1 = 1
140 NEXT N
150 PRINT N(1);N(2);N(3)
160 NEXT S
999 END
READY
RUN
RANDOMOZ
 211
       391 398
 186 420 394
 199
       378 423
1.019 SEC.
READY
60 \text{ FOR } N = 1 \text{ TO } 10000
RUN
RANDOMOZ
 2006 3997 3997
 1976 3982 4042
 2068 3973
                3959
9.286 SEC.
READY
```

CLTH starts by setting up P, N, the expected value E, and the standard deviation S. Then the loop of 1000 repetitions is started. For each repetition, X, the number of successes, is initially set to 0. Lines 50-80 count the number of successes in 100 trials. On line 60, if RND > .3 we have a failure, and hence the next line is skipped. For a success, X is increased by 1. Line 90 computes the number of standard deviations. We shall keep track only whether it is between 0 and 1, between 1 and 2, etc. This is accomplished on lines 100 and 110. When all 1000 repetitions are completed, we wish

CLTH

10 LET P = 20 LET N = 1 25 LET E = 1 30 LET S = 5 40 FOR I = 1 45 LET X = 0 50 FOR J = 1 60 IF RND >	100 N*P SQR(N*P*(1-P)) 1 TO 1000 3 1 TO N
70 LET X = 2 80 NEXT J 90 LET Y = 4	K+1 Abs((X-e)/s)
100 LET K = 110 LET N(K) 120 NEXT I	INT(Y)+1 = N(K)+1
140 FOR D =	A+N(D)/2000
199 END READY RUN	
CLTH	
STD'S 1 2 3 4	AREA ؕ334 ؕ4795 ؕ4975 ؕ5
5.466 SEC. READY	

to print the approximate areas. A comment concerning line 150 is in order. We wish to compute the cumulative areas, as in Figure 5; hence we keep adding the new area to the previous value of A. The reason for dividing the total number of occurrences N(D) by 2000 rather than 1000 is that we want the area on one side of the expected value, while our counting method lumped the two sides together.

We note that the computed values agree quite well with the true values. The true values are .341, .477, .4987, .49997.

EXAMPLE 5 Baseball. The game of baseball is a good example of a game having a model for which a complete theoretical treatment is not practical, and hence much can be gained from simulation.

How would we build a simulation model for a given team, in order to study the way they produce runs? Fortunately, some very detailed statistics are kept, over long periods, which are ideal for building such a model. Let us suppose that a given batter comes to bat. We know from past experience what the probabilities are for his making an out, getting a walk, or getting a hit of various kinds. We simply generate an RND, and use it to decide what the batter did.

For example, if he has probabilities .1 for a walk, .64 for an out, .2 for a single, .03 for a double, .01 for a triple, and .02 for a home run, we can generate a random integer from 1 through 100, and interpret it as in Figure 16.

Range	Result	Probability
1-10	Walk	.1
11-74	Out	.64
75-94	Single	.2
95-97	Double	.03
98	Triple	.01
99-100	Home run	.02

Figure 16

We can then bring the next batter to bat, and arrive at a result based on *his* past performance. The running on the bases may be simulated similarly. For example, we can feed into the machine the probability that a man on first reaches third on a single. Just how realistic we wish to make the model depends entirely on how much work we are willing to do.

It should be noted that we are simulating only the batting of *one* team. We do not here consider the batting of the other team, or questions of defensive play.

Such a model would be most useful in training young managers. The computer could make all decisions (many of them stochastic) having to do with the performance of the players, while the manager could make all

decisions normally open to managers. For example, he could call for a hit-and-run play, and the machine would simulate the results. He could call for a steal, or send in a pinch hitter, or tell a batter to try to hit a long fly ball.

By the use of a computer a new manager could gain an entire season's experience in a few days—and he would not be learning at the expense of his team.

The model is also useful for planning purposes, as we shall illustrate here. One important task of the manager is to decide on his batting order. He could feed a variety of batting orders to the computer, have it try each for a season's games (or more), and report back the results.

This was actually done on the Dartmouth Computer.

The team used in the simulation was the starting line-up of the 1963 world champion Los Angeles Dodgers. The line-up of Figure 17 was used throughout.

Line-up	Batting average	Slugging average	
1. Wills	.302	.349	
2. Gilliam	.282	.383	
3. W. Davis	.245	.365	
4. T. Davis	.326	.457	
5. Howard	.273	.518	
6. Fairly	.271	.388	
7. McMullen	.236	.339	
8. Roseboro	.236	.351	
9. Pitcher (average)	.117	.152	

Figure 17

An entire season of 162 games was simulated, keeping detailed records for each player. Of course, this simulation differed from the normal year in a few respects. For instance, the first eight players played every inning of every game. Since only the batting was simulated, no allowance was made for defensive play, nor did the game stop after eight innings if the home team was ahead. Games were not called on account of rain, and there were no extra-inning games. But many important features concerning batting were recreated quite realistically. We shall cite a few of the more interesting results.

Seven of the batters ended up with batting averages close to their actual ones, but two did not. Tommy Davis, the league's leading hitter, had an even more spectacular year during simulation: he batted an even .350 (compared with .326 in 1963). On the other hand, Fairly, who had batted .271 in actuality, had a bad simulated year, batting only .250. This shows how much a batting average can change due to purely random factors.

Howard was far ahead in home runs, with 54. This is much higher than the 28 he had in actuality, but he was only used part-time in 1963, while in the simulated year he played all the time. Two of the home runs were hit by pitchers—just as in real life. In one game Howard hit three home runs. But mostly it was the balance of the Dodger team that showed up; there were ten games in which three different players hit home runs.

There were no really spectacular slumps, though Gilliam once went 15 consecutive at-bats without getting a hit. The total number of runs scored was 652, in excellent agreement with the actual 640. On the other hand, the 1352 men left on base compared very poorly with the Dodgers' league-leading performance of leaving only 1034 men on base. Two factors in this were the absence of double-plays and pinch hitters in the simulation model. But there is probably some other relevant attribute of the team that was missed in the model.

Perhaps the most interesting result is the number of shutouts (of the Dodgers, of course). There were 11 in the simulation, as compared to the league-leading performance of only 8 shutouts. In the simulation, two of the shutouts occurred in the final two games. Thus, if the season ended in 160 games, the simulation would have been off by only one shutout. This shows how hard it is to get an accurate estimate for a small probability through simulation! And there was a four-game stretch late in the season in which three of the games ended in shutouts. If this had happened in real life, all the Los Angeles papers would have carried headlines about a Dodger batting slump.

To compare various possible batting orders, several line-ups were simulated for ten entire seasons. The seven line-ups are shown in the first column of Figure 18, and the results in the second column. The standard deviation of the average number of runs per game was about .07. Since the difference between the best and the worst line-ups is over three standard deviations, one is tempted to conclude that the batting order really makes a difference —though not very much of a difference.

However, this simulation—though time-consuming—is not conclusive. We may still entertain the hypothesis that any line-up averages about 3.95 runs per game, and all seven outcomes are within two standard deviations of this. We are forced into an even more substantial simulation run.

The simulation was repeated; this time every line-up had seven sets of ten entire seasons simulated. The newly computed averages are shown in the third column of Figure 18, while the maximum and minimum values obtained for a set of ten seasons are shown in the last column. Since we have simulated seven times as many games for each line-up, the standard deviation is reduced by a factor of $\sqrt{7}$, to less than .03. The differences in the averages now look more significant. Also, we note that the ranges obtained for the first five line-ups don't overlap (or hardly overlap) the ranges for the last two line-ups. We may therefore conclude that we have five "good" and two "poor" line-ups. And this hypothesis stands up under more sophisticated tests.

What characterizes the poor line-ups? Most noticeably, the pitcher is first, rather than being last. But also we note that the Dodgers had three weak

	Average number of runs per game		
Line-up	10 seasons	7×10 seasons	Range
1, 2, 3, 4, 5, 6, 7, 8, 9	4.06	4.00	3.91-4.06
1, 4, 2, 5, 6, 3, 8, 7, 9	4.07	4.02	3.92-4.07
4, 5, 6, 1, 2, 3, 7, 8, 9	4.00	3.98	3.90-4.04
2, 1, 3, 5, 4, 6, 8, 7, 9	3.98	4.01	3.95-4.08
1, 4, 7, 2, 5, 8, 3, 6, 9	3.90	3.98	3.90-4.05
9, 8, 7, 6, 5, 4, 3, 2, 1	3.89	3.82	3.72-3.89
9, 6, 3, 8, 5, 2, 7, 4, 1	3.83	3.83	3.76-3.92

Figure 18

hitters (numbers 3, 7, and 8), and two of these are near the top of the bad line-ups. We therefore conclude that poor hitters should be near the end of the line-up. But little else can be concluded.

We should also note that the difference between best and worst is surprisingly little, and drastic changes in the "best" have practically no effect. Thus we conclude that the importance of the batting order has been greatly exaggerated.

One additional remark may be of interest: The first line-up in Figure 18 is, of course, the one actually chosen by the manager. The last five are simply permutations chosen according to simple patterns. However, the second line-up was chosen by one of the authors, a Dodger fan, as his attempt to "manage" the team. He was most pleased that it turned out best! Of course, .02 is only $\frac{2}{3}$ of a standard deviation, which represents about three runs per year, and is not significant.

EXERCISES

- 1. Use the RUN of RANDOM to simulate an independent trials process with probability .4 of success, for 30 trials. How many successes do you obtain? [Ans. 15.]
- 2. Simulate three games of craps as follows. To imitate the roll of a pair of dice choose pairs of outcomes of the last RUN of RANDOM reading from left to right in successive rows and then proceed according to the rules of craps.
- [Partial Ans. On first game player rolls a 5 and wins.] 3. From Chapter 3, Section 3, Exercises 16, 17, and 18, compute the expected number of bust, straight, flush, and full house hands in 10,000 poker hands. Also compute the standard deviation for each. Do the figures given in Example 2 for the simulation look reasonable? [Partial Ans. Bust: expect 5012; off by less than one standard deviation.]
- 4. What would be reasonable 95 percent confidence limits for the deviation from the expected number of wins in 250,000 games of craps?

Do the simulated results of -.0115 and -.0154 mentioned in the text fall within these limits?

- 5. Using the data of the 30 random numbers between 0 and 1 generated by RANDOM, test the hypothesis that the probability that a random number generated this way has probability .1 of falling between .2 and .3.
- 6. Use the random numbers produced by the program RANDOM to simulate 30 days' weather in the Land of Oz, following a rainy day; see Example 3.
- 7. Change the random numbers generated by RANDOM between 0 and 1 to random numbers between 1 and 100.
- 8. Suppose that we have a baseball team whose batters each performs according to the simulation scheme in Figure 16. Use the random integers obtained in Exercise 7 to simulate the performance of the first 30 batters on one team. How does the team stand after 30 men have come to bat? [Ans. End of six innings; four runs scored.]
- 9. In 1951, Gil Hodges of the Brooklyn Dodgers was officially at bat 582 times and hit 40 home runs. Estimate his probability of hitting a home run each time he was at bat. How large a fluctuation in his annual home-run output is attributable to pure chance?
- 10. From 1949 through 1959, Gil Hodges had the following number of home runs: 23, 32, 40, 32, 31, 42, 27, 32, 27, 22, 25. Is there a case for his having had "good" and "bad" years, or may we assign the differences entirely to chance fluctuations: [*Hint*: Estimate the expected value from the data and use Exercise 10.]

[Ans. Explainable as chance fluctuations.]

11. In Exercise 14 of Section 4, you were asked to derive the more exact confidence intervals

$$\frac{1}{1+s^2/n} \left[\overline{p} + \frac{s^2}{2n} \pm s \left(\frac{\overline{p}(1-\overline{p})}{n} + \frac{s^2}{4n^2} \right)^{1/2} \right].$$

Write a program to compute these more exact intervals given n, \overline{p} , and s.

- 12. Use the program of Exercise 11 to rework Exercises 2 and 6 of Section 4. Also rework each of these exercises using the program CLTH given in this section. For each exercise give one possible value for p which is ruled out by the more exact confidence interval but is not ruled out by the approximation used in the program in this section.
- 13. Write a program to test the hypothesis p_0 against p_1 , given p_0 , p_1 , and s. Have the program print both the number of experiments needed and the number of those experiments that must be successful in order to accept hypothesis p_1 .
- 14. Use the program of Exercise 13 to rework Exercises 9 and 11 of Section 3.
- 15. Write a program which, given p, simulates 100 tosses of a coin which comes up heads with probability p. Combine this with the program for confidence intervals given in the text and compute 95 percent

confidence limits for p, given the simulated data, and see whether p is within the confidence interval. Do the same using the more exact confidence limits which the program of Exercise 11 computes.

16. Run the program of Exercise 15 a total of 500 times and find, for each method, what fraction of the time p lies within the confidence interval.

SUGGESTED READING

Blackwell, David. Basic Statistics. New York: McGraw-Hill, 1969.

Hodges, J. L., Jr., and Lehmann, E. L. Basic Concepts of Probability and Statistics. 2nd ed. San Francisco: Holden-Day, 1971.

Kurtz T. E. Basic Statistics. Englewood Cliffs, N.J.: Prentice-Hall, 1963.

Mosteller, F.: Rourke, Robert E. K.; and Thomas, G. B. Probability with Statistical Applications. Reading, Mass.: Addison-Wesley, 1970.

Tanur, Judith, et al., eds. Statistics: A Guide to the Unknown. San Francisco: Holden-Day, 1972.