CHAPTER 9. USING RANDOM DIGITS.

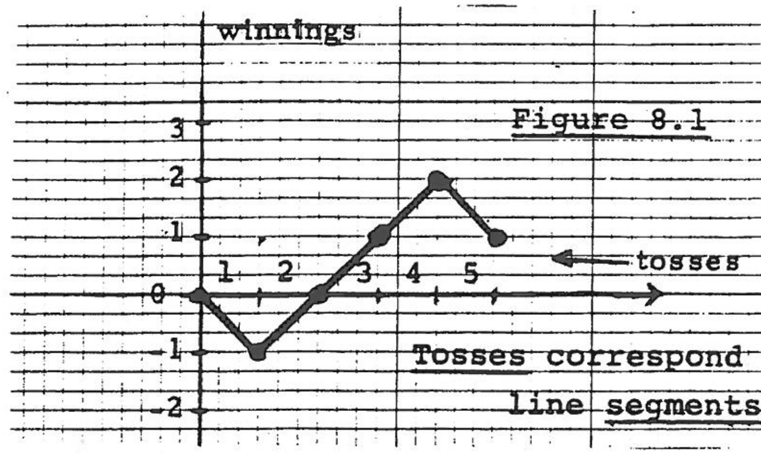
The basic concepts of probability (and of statistics as well) have to do with the carrying out of experiments. Theory and experiment are therefore closely related, and it is often useful, when working in this part of mathematics, *actually to carry out probabilistic experiments*. We can use such experiments for several different purposes. We can do experiments to confirm a theory or calculation; or to get an approximate value for a quantity that would otherwise require a long and difficult computation; or to explore the properties of a given probability space; or to obtain data which may help to suggest a new theory or concept; or to help make random choices as part of a larger observational or decision-making procedure. In the present section, we look at these uses in more detail; we give some examples; and we present an important special method for getting experimental data - the method of *random digit tables*.

Let us look at some of these purposes in more detail.

1. *To confirm a theoretical result or to illustrate a previously observed experimental fact*. In Chapter 1 we saw examples where experiments were carried out to illustrate the stability of relative frequencies. (These experiments were described as tossing a thumbtack or tossing a coin.) Results of such experiments were given in Chapter 1.

2. *To get, quickly, an approximate value for a quantity when the computation of that quantity from theory appears difficult*. In particular, experiments can be especially useful when we have no idea at all ahead of time of the size of a desired

probability. Consider the following situation. Two individuals, A and B, play a game where a coin is tossed repeatedly. Each time heads appears, B pays A one dollar. Each time tails appears, A pays B one dollar. At any moment, A has a certain total of net winnings equal to the difference between the total amount received from B so far and the total amount paid to B so far. (The net winnings may be negative.) We say that A is ahead for a particular toss if, either before or after that toss, A's net winnings are positive. We say that B is ahead for a particular toss if, either before or after that toss, A's net winnings are negative. It is easy to see that for every toss in a given sequence of tosses, either A or B is ahead but not both. For example, in the sequence of five tosses THHHT, B is ahead for the first two tosses and A is ahead for the last three. In the sequence HTHHT, on the other hand, A is ahead for all five tosses. Results of a sequence of tosses can be graphed by drawing a connected sequence of line segments to show the changes in A's net winnings with each play. For the sequence THHHT we would have:



Each segment corresponds to a toss. Hence, to say that A is ahead for a certain toss is the same as to say that the line segment for that toss lies above the horizontal axis in the graph.

We now ask the following question: in a sequence of 20 tosses, what is the probability that either A is ahead for all 20 tosses or else B is ahead for all 20 tosses? While the sample space is clear (there are 2^{20} equally likely possible sequences), there is no evident easy way to calculate the probability of this event. On the basis of our experience in previous chapters with binomial distributions and normal approximations, we might intuitively expect this probability to be extremely small.

We can get an estimate for this probability by carrying out a series of experiments. In each experiment, A and B do a sequence of 20 tosses. We then note for each sequence whether or not the event that either A is ahead for the full 20 tosses or B is ahead for the full 20 tosses occurs.

We now carry out a series of 25 such experiments, and we get the following data. (Details of the data are given in Example 3 below.) We observe that in 5 of these experiments A is ahead for all 20 tosses while in 4 of these experiments B is ahead for all 20 tosses. Thus the desired event has occurred 9 times in 25 experiments for an observed relative frequency of $9/25 = 0.36$. This relative frequency cannot be taken as a good estimate of the probability we seek, since, by weak stability, we can only expect the relative frequency to be within $1/\sqrt{n} = 0.2$ of the correct probability. The observed relative frequency does tell us, however, (to our possible surprise) that the value of the probability is not small. If we wished to, we could get a more accurate estimate by using a larger number of experiments (that is to say, a larger value of n). We shall calculate the correct (theoretical) value later in this chapter.

3. _To study, experimentally, the properties of a given probability space_. For example, in the game between A and B described above, we note that in a sequence of 20 tosses A can be ahead on either 0, 2, 4, ... 18, or 20 tosses. We can ask how the probability varies over this range of possible results. Our previous experience might lead us to expect that the most likely results are in the middle of the range, where A is ahead about half the time. In the actual experiment of 25 trials described above (where each trial is a sequence of 20 tosses) we observe the following:

x = no. of tosses A ahead	0	2	4	6	8	10	12	14	16	18	20
$y = \frac{x}{20}$ = fraction of time A ahead	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
observed frequency in 25 trials	4	4	1	1	1	2	1	3	1	2	5

These observations suggest, contrary to what we may have expected, that the probability is greater at the ends of this range and smaller in the middle.

4. To suggest new theoretical results. When the coin tossing problem described above is proposed, experienced mathematicians often expect that the probability distribution will (after an appropriate change of scale) have the shape of the standard normal curve. Actual observations, similar to those obtained above, show that this cannot be the case. Indeed if we make more and more observations, we see that the probabilities follow a symmetrical U-shaped curve. A theoretical investigation to find the true shape of this curve

is complex but can be carried out. (See the end of this chapter.) The result is as follows. Consider a bar graph for the theoretical probability distribution drawn with y on the horizontal axis (where y is the fraction of time that A is ahead on a given trial of 20 tosses) and with bar height taken so that the probability of each possible y value is given by the area of the bar above it. Then, as the number of tosses in each sequence is increased, the bar graph approaches the silhouette given by the curve.

$$f(y) = \frac{1}{\pi\sqrt{y(1-y)}} .$$

Because the area under this curve between $y = a$ and $y = b$ is given by

$$\int_a^b \frac{1}{\pi\sqrt{y(1-y)}}\, dy = \frac{1}{\pi} \sin^{-1}(2y - 1)\Big]_a^b ,$$

this theoretical result is called the arc sine law.

5. To make random choices as part of a larger observational or decision-making procedure. We give two examples. First, a group of 10 students is to be randomly selected from a body of 500 students. We could proceed as follows. Take 500 identical slips of paper, write each student's name on a slip, mix the slips in a box, and make a blindfold drawing of 10 slips. (This method, known as drawing lots, has been a traditional one for selecting conscripts for military service.) The result of such a procedure is called a random sample. Second, 20 rats are grouped into 10 pairs, with each pair of similar weight and age. One member of each pair is to be randomly chosen for a special diet. The remaining members are to form a control group that gets a normal diet. Here a single coin toss for each pair will make the choice for us.

The method of random digit tables. If experimental results are needed, it is often easier to make use of data from physical experiments that have already been performed than to conduct new experiments. A useful source of such data can be found in tables of random digits. Such a table is initially created in the following way. Repeated independent trials are made of an experiment with 10 equally likely basic outcomes which we designate as $0, 1, \ldots, 9$. Results of these trials are recorded as a sequence of digits. The record of a large number of successive trials is called a table of random digits, or, sometimes, a table of random numbers. Such tables can be obtained in published form. As many as a million digits may be included in a single published volume. A typical line in such a table appears as follows:

91562 18812 81267 67384 77024 35639 05559 47634 40763 34680

The digits are grouped in blocks of 5 for ease in reading and using the tables. We emphasize that such a table is the record of repeated trials of an actual physical experiment that has been carried out in the past. A brief table of random digits is given at the end of this chapter.

The special importance of random digit tables arises from the fact that they can be easily used to simulate a wide variety of probabilistic experiments. We now illustrate this by showing how various of the experiments described above can be carried out by means of random digits.

Note. In working with random digits, it is convenient to have random digit tables stored in a computer. The computer can then be programmed to carry out various desired simulations quickly and easily. In current computer-science practice, however, a table of experimentally obtained random digits is usually not stored in the computer. Instead, the computer is given a special additional program for generating a deterministic sequence of digits which appears to have a random pattern quite similar to the patterns found in experimentally obtained sequences of digits. This deterministic sequence of so-called pseudo-random digits is then used, in exactly the same way as tables of random digits, for simulating probabilistic experiments. Later below, we shall further consider pseudo-random digits and some of the controversy that surrounds their use.

Example 1. We wish to make 10 repeated trials of the binomial experiment of tossing a coin 5 times and seeing how many heads occur. We use the above line of random digits as follows. In each block of 5, we take odd digits to mean heads and even digits to mean tails. Then the number of successes for the ten trials of the binomial experiment are:

3, 2, 2, 2, 2, 4, 4, 2, 2, 1.

Example 2. A thumbtack is known to have probability 0.58 of falling on its side and 0.42 of falling on its back. We wish to carry out 25 independent trials of tossing the tack. From a table of random digits we take the following line:

03070 19190 90627 03821 57788 00640 61392 67042 40441 33293

We simulate the tack tossing experiment as follows. We divide the line into successive pairs of digits

03 07 01 91 90 ...

Each pair of the form 00, 01, ..., 57 counts as success (side) and each pair of the form 58, 59, ..., 99 counts as failure (back). By the multiplication law for independent events, each possible pair has probability 0.01;

hence we are simulating a binomial experiment with $n = 25$ and $p = 0.58$. We obtain the result

SSSFFFFFSSSFFSSFSSFSSSSSF.

The observed number of 15 successes gives a relative frequency of $15/25 = 0.60$ which agrees well with our probability value of 0.58. (In fact, the formula $2/\sqrt{n}$ from Chapter 1, for the weak stability of relative frequencies, would lead us to expect only that the observed value should fall between 0.38 and 0.78.)

Example 3. Consider the coin tossing game described above where we asked for the probability that either A is ahead for all 20 tosses or B is ahead for all 20 tosses. We can use random digit tables to conduct our experiment as follows. We take the following lines from a table.

83859	26028	09875	91192	79804	37849	25470	31328	36448	35040
50544	24432	31980	54495	67467	98564	73568	46852	06740	70182
31985	14308	52244	72821	06894	63611	34813	72992	41403	90753
38192	40488	51390	23732	96267	37126	77043	03465	49410	60711
46186	85771	26633	46162	98815	97171	31414	70074	95699	75577
04636	03212	69251	80531	38720	95551	77997	73905	34319	33735
87894	96447	03480	67122	18493	09631	96310	34117	03988	42617
34597	71020	66965	10069	08508	68820	46419	93443	05775	27300
72756	00456	35886	91555	39362	80164	41610	96236	39254	30930
19483	86555	10938	68930	81873	69305	38285	50131	34190	56609

For convenience in reading, we use the first two lines in the first two columns for the first sequence of 20 tosses, then the next 2 lines in the first two columns, etc. We take <u>even</u> as heads and <u>odd</u> as tails. In these 25 sequences of 20 tosses each, we find 9 cases where either A is always ahead or B is always ahead. (Third case in first pair of columns, third and fourth cases in second pair, second case in third pair, last three cases in fourth pair, and third and fifth cases in last pair.) It is this simulated experiment that was reported earlier in this section when the game was first described.

<u>Remark</u>. A theoretical analysis of this game yields the following formula for the probability desired (see the end of this chapter):

$$P(\underline{\text{A ahead on all 20 tosses or B ahead on all 20 tosses}}) =$$

$$\binom{20}{10}\frac{1}{2^{19}} = 0.35$$

Thus the observed result (0.36) agrees well with the theoretical value. In fact, this is better agreement that we would usually expect to find.

<u>Example 4</u>. To get information as to how the probability varies over the range of possible results, we can record the result of each of the 25 sequences (of 20 tosses each) obtained from the above random digits. This yields the table of experimental results given earlier in this chapter. We reproduce that table here and add a line, based on a deeper mathematical analysis of the probability space, which gives the theoretical average number of occurrences of each result that we would expect to observe in the long run in

every 25 sequences of 20 tosses each. We do not give details of the mathematical analysis here.

x = no. of tosses A ahead	0	2	4	6	8	10	12	14	16	18	20
$y = \frac{x}{20}$ = fraction of time A ahead	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
observed frequency in 25 trials	4	4	1	1	1	2	1	3	1	2	5
theoretical frequency in 25 trials	4.4	2.3	1.8	1.6	1.5	1.5	1.5	1.6	1.8	2.3	4.4

The agreement between theory and observation is close.

Example 5. One of the most common uses of tables of random digits is in the selection of random samples. For example, if we wish to select 10 students from among 500, we associate numbers from 1 to 500 with the individual students, and we then turn to a table of random digits to select 10 of these numbers. In particular, we look at successive triples of numbers and consider only those which fall between 001 and 500 in numerical order. The first ten distinct such triples that we find form our sample. Thus using the random digits

70533 44415 27129 29130 11215 33658 61558 82697 96584 97213
94610 34001 56631 57082 80810 93288 94441 87871 61274 01288

we would get

334, 441, 129, 291, 301, 121, 269, 139, 461, 034.

It easy to show that all samples of 10 (without replacement) are equally likely under this procedure.

As we shall later see, random choices can be an important part of data gathering procedures in statistical studies. For example, in a physical comparison test of consumer preferences among three different soft drinks, we might use random digit tables to determine the order in which the 3 drinks are presented to each experimental subject. We discuss this further in Chapter 13.

Remark. Random digits are not always used when a single sample of people is to be drawn from a larger group. In drawing such a sample, it is a common practice, for example, to list the individuals by name in alphabetical order, and then to select the members of the sample by choosing every m^{th} individual where m is the largest integer not greater than $\frac{n}{r}$, n is the size of the larger group, and r is the size of the sample. Thus for a sample of 10 from 500, we would take numbers 50, 100, 150, ... in alphabetical order. There are some theoretical difficulties with this procedure, but it is convenient and, on the whole, works well. It is possible, if one is not careful, to devise procedures that fail to be random. For example, in selecting people from a phone book, we might use random numbers to select an entry in the book and then, if this entry is for an organization or office rather than for a person, go on to select the first following entry which is for a person. This is not a random selection because an individual whose name happens to occur first after a lengthy list of organization names will have a much higher chance of being chosen than most other persons.

Crossing out. It is a common practice among researchers using random digit tables to cross out digits as they are used so that they will never be used again. This assures that all simulated experiments are new for that investigator and that there is no way in which results can be biased by the unintended repetition of some long pattern of digits.

Monte Carlo methods. Random digits are also useful in parts of mathematics other than probability and statistics. They can be used, for example, to estimate the area of a set D of points in the plane defined by certain algebraic inequalities. A rectangle is taken which includes the set D, and then a random digit table is used to generate coordinates for a sequence of points in the rectangle, where points in the sequence may fall anywhere in the rectangle. The proportion of these points which also satisfy the given inequalities gives an estimate of the proportion of the total area of the rectangle which is taken up by D. Estimation methods of this kind are called Monte Carlo methods. Computers are almost always used in order to obtain data for a sequence of sufficiently many points so that the resulting estimate has accuracy to several decimal places.

The number of points necessary to yield a desired accuracy can be estimated by normal approximation (or, more briefly, by the weak stability law.)

Origin of tables. A variety of experimental setups can be used to generate random digit tables. For example, a die can be rolled and a coin can be tossed at the same time. If outcomes with the die showing six are ignored, the remaining ten outcomes can be taken as the digits 0,1,...,9, and we will expect these outcomes to occur about equally often. A record of outcomes of successive trials then gives us a table of random digits. (We can take the outcomes 1H,2H,...,5H as the digits 1,2,...,5 and the outcomes 1T,2T,...,5T as the digits 6,7,...,0.) More elaborate, faster, and more standardized experiments have been used to produce published tables. Usually such experiments take one of two forms. An SP-generator involves (i) a physical apparatus which cycles through the ten digits 0,1,...,9 at a rapid and constant rate (this could be a rotating cylinder with the ten digits marked on its surface and a fixed pointer past which the digits move, or it could be an appropriate electronic circuit); and (ii) a method for getting irregular points in time at which digits are read from the apparatus (this could be the accidental closing of circuits during irregular motions of a human hand, or it could be an appropriately instrumented monitoring of the occasional escape of electrons from a heated filament (shot noise)). A PS-generator involves

(i) a physical apparatus which cycles through the ten digits 0,1,...,9 at a rapid but irregular rate (this rate is governed by some method for getting irregular points in time as described in (ii) for the SP-generator); and (ii) a method for reading digits from the apparatus at regular intervals. SP-generators were used for early tables of random digits. The PS-generator has theoretical advantages, however, and has now come into general use.

(Note. "SP" stands for Steady/Poisson and refers to the fact that (i) occurs at a steady rate while (ii) is obtained from a physical Poisson process (a Poisson experiment in which events occur randomly in time). Similarly "PS" means that (i) is governed by a Poisson process while (ii) occurs at a steady rate. The case "PP" turns out to be theoretically inferior to both PS and SP.)

Pseudo-random digits. Sequences of digits which appear to be very similar to tables of random digits can also be generated by computers operating under certain special deterministic programs. As mentioned above, such a sequence is called a sequence of pseudo-random digits. Pseudo-random digits have been found to be satisfactory for many of the same purposes as tables of random digits. The method of middle digits is one example of a method for generating such a pseudo-random sequence of digits. We take some initial number to 9 decimal places (say the number π). We then carry out a mixture of transcendental and algebraic operations on it and display the result of each step to 9 decimal places in scientific notation. At each step we take the 3rd through 7th decimal digits (the middle digits) as the next five digits for our table. For example, the operation could be successively to square until a number greater than 10 is reached, successively to take natural logarithms until a number less than 3 is reached, successively to square until a number greater than 10 is reached, etc. Starting with π, this particular program yields

15926	96044	09091	89195	14630
48498	85298	13842	73793	11607
69102	18256	03537	40175	52665
01345	47013	42754	83432	21854

as its first 100 digits. Note that the digits 1, 2, ..., 0 have occurred with frequencies

13, 9, 10, 13, 11, 8, 7, 9, 10, 10.

This is the sort of variation one would expect to find in a genuine table of random digits. For example, the following two lines from a genuine random digit table

30220	72938	62117	83140	84679	94416	93338	29255	25554	28666
30838	04547	77014	06276	58503	91848	45112	95180	04065	48652

give the frequencies

9, 11, 9, 12, 12, 10, 7, 12, 7, 11.

Pseudo-random digits have the evident advantages (a) that much less space is required in computer memory, (b) that the same sequence can be regenerated whenever needed, and (c) that no probabilistic physical experiment is required. Pseudo-random digits have the disadvantage that there may be (in a given sequence) hidden regularities and periodicities which we have failed to notice.

Testing for randomness. The usefulness of pseudorandom digits raises interesting philosophical questions about the nature of randomness in a sequence of digits. Can we test for randomness? After all, any table of random digits is a given finite sequence, and any given single finite sequence has the same probability of being produced in an experiment as any other single sequence of the same length (including, for example, the sequence which consists entirely of 0's). What makes one finite sequence seem more random than another? Philosophers and scientists do not yet have satisfactory and complete answers to questions of this kind. Some philosophers would relate these matters to the system of language and concepts that we use to describe and predict the physical world around us. More specifically, one widely-used approach is to define the degree of randomness of a given finite sequence in terms of the minimum length of program needed to generate the sequence on some standard computer.

Mathematicians have developed a number of partial tests for randomness. We expect, for example, that each digit will occur about 1/10 of the time in a long sequence. We can go on and test the frequency of occurrence of pairs of digits and of triples of digits. More sensitive tests have also been studied which count the lengths and frequencies of runs. (A run is a portion of the sequence consisting of repetitions of a single digit). None of these tests is conclusive, however, and long sequences passing all of these tests but having deeper repetitive patterns can be devised.

Theoretical note. The exact formula for Example 3 can be derived as follows. We first calculate the probability that A is ahead for all 20 tosses. It will follow by symmetry that the probability that either A is ahead for all 20 tosses or B is ahead for all 20 tosses will be twice this quantity.

We begin by considering sequences of 21 tosses (rather than 20 tosses). Each of these can be graphed as in Figure 8.1. Since 21 is an odd number, either A or B must be ahead at the end of 21 tosses. Hence all of the 2^{21} graphical paths must end either above or below the horizontal axis. It is evident that half of them, 2^{20}, will end above the axis. Now consider the points S, P, Q, and R in the graphical figure shown in Figure 8.2.

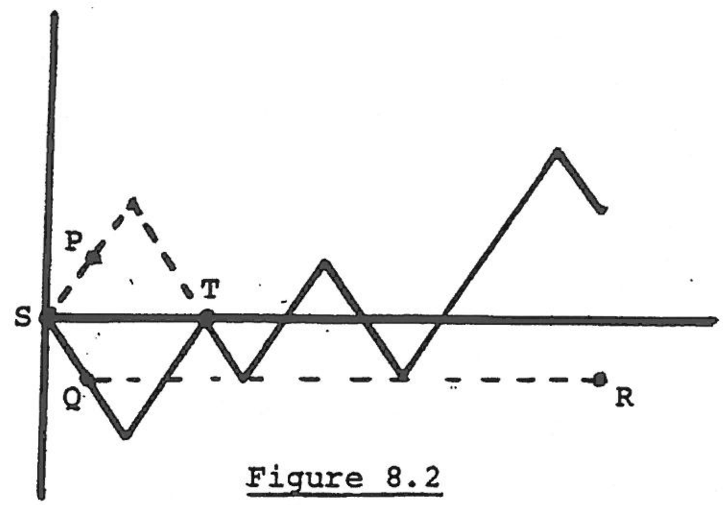


Figure 8.2

S is the starting point. P is the point reached after one toss if A wins that toss. Q is the point reached after one toss if A loses that toss. R is the final point reached if, after 21 tosses, A ends up one dollar behind. We note the following facts.

(a) The number of possible sequences of 20 tosses in which A is always ahead (this is what we seek) is the same as the number of sequences of 21 tosses in which A wins the first toss and then, on the remaining 20 tosses, is always ahead by at least one dollar. Hence it is the same as the number of paths in Figure 8.2 that go through P and never again touch the horizontal axis.

(b) The number of all paths through Q is 2^{20}, and the number of paths through Q which end at R must be $\binom{20}{10}$. Hence the number of paths through Q which do not end at R must be $2^{20} - \binom{20}{10}$.

(c) Hence, by symmetry about the line $\overline{QR}$, the number of paths through Q which end above the axis must be $\frac{1}{2}[2^{20} - \binom{20}{10}]$. (Remember that no path can end on the axis, since the total number of tosses is odd.)

(d) Any path through Q which ends above the axis must cross the axis. Let T be the first point after S where such a path touches the axis. If we make a mirror reflection (in the horizontal axis) of the portion of the path between S and T but keep the remainder fixed, we get a new path through P which touches the horizontal axis for the first time at T and then goes on to end above the axis. This gives a one-one correspondence between (i) the set of paths through P which touch the axis at least once after S and end above it and (ii) the set of paths through Q which end above the axis. Hence from (c), the number of paths through P which touch the axis after S and end above must also be $\frac{1}{2}[2^{20} - \binom{20}{10}]$.

(e) The number of all paths from S which end above the axis must be $\frac{1}{2}(2^{21}) = 2^{20}$. Hence the number of paths through P which do not touch the axis after S and end above the axis [(a)] must equal the total number ending above [(e)] less the number through Q which end above [(c)] less the number through P which touch after S and end above [(d)]. Thus we have the number of paths through P which do not again touch the axis after S:

$$(a) = (e) - (c) - (d) = 2^{20} - \tfrac{1}{2}[2^{20} - \binom{20}{10}] - \tfrac{1}{2}[2^{20} - \binom{20}{10}]$$
$$= \binom{20}{10}.$$

Since the probability of any particular sequence of 20 tosses is $1/2^{20}$, it immediately follows that the probability that A is always ahead in a sequence of 20 tosses is $\binom{20}{10}\frac{1}{2^{20}}$. By symmetry, the probability that B is always ahead is also $\binom{20}{10}\frac{1}{2^{20}}$. Hence the probability that either A is always ahead or B is always ahead is

$$\binom{20}{10}\frac{1}{2^{19}}.$$

The above argument obviously gives, for any sequence of length 2n, the result that the number of paths in which A is always ahead is $\binom{2n}{n}$.

A similar but somewhat more complex argument, which we do not present here, shows that the number of sequences of length 2n for which A is ahead on a total of 2x tosses and B is hence ahead on a total of 2n - 2x tosses is

$$\binom{2x}{x}\binom{2(n-x)}{n-x}.$$

(This formula was used to calculate the theoretical frequencies for the table given in Example 4 above.) Substitution of Stirling's formula in this expression immediately yields the arc sine law.

Ruin problems. Assume that A and B play the game of 20 tosses described at the beginning of this chapter, where each player bets one dollar on each toss. Assume further that A begins with total resources of a dollars and that B begins with b dollars. If A's losses ever reach size a or if B's losses ever reach size b, then the game must end, since one player or the other will be out of money at that point and unable to continue. We say that the player who is out of money has been ruined. Various interesting questions can now be asked. For example, in the game of 20 tosses (now, more precisely, a game of at most 20 tosses), we can ask for: P(A is ruined on or before the 20th toss), P(B is ruined on or before the 20th toss), and P(neither player is ruined on or before the 20th toss). We can also, if we wish, allow the game to go on beyond 20 tosses (if neither player is ruined at that point) and to end only when one player or the other is ruined. We can then ask for P(A is ruined) and P(B is ruined) for this indefinitely long game. This last problem is known as the classical symmetrical ruin problem (symmetrical because the Bernoulli trials have $p = q = 1/2$, and classical because the problem was first posed very early in the history of probability theory). We consider ruin problems further in an exercise below.

Random walks. An individual W stands at the origin of a number line and begins a sequence of coin tosses. Each time heads appears, W moves one unit (of distance) to the right (the positive direction) and each time tails appears, W moves one unit to the left (the negative direction). Each move is called a step. The entire process is called a random walk or, more specifically, a symmetrical one-dimensional random walk (symmetrical because the probabilities for the two directions are equal). A random walk can be limited to some fixed finite number of steps, or it can be infinite. The game of 20 tosses described early in this chapter can be thought of as a random walk of 20 steps, with the size of A's winnings at any time corresponding to the position of W in the random walk. A graphical picture such as Figure 8.1 can then be thought of as a graph of position (vertical axis) against time (horizontal axis) for a particular random walk path.

To apply probability theory to a random walk (finite or infinite), we take the set of all possible paths for that random walk as our sample space. In the finite case, we have finitely many possible paths, and we get a probability function in the obvious way: we give each path probability value $1/2^n$, since the probability of having that particular sequence of outcomes for the coin tosses must be $1/2^n$, where n is the number of steps in the walk. We can now find the probability of any event. For example, in the walk of 20 steps, the event W never moves to left of origin in 20 steps is the same as the event A is never behind in 20 plays in the game described earlier, and from the theoretical note above, we have

$$P(\text{A is never behind in 20 plays}) = \binom{20}{10}\frac{1}{2^{20}} .$$

In the infinite case, there are infinitely many points in the sample space, and each sample point is itself a particular infinite path. We are thus, for the first time, concerned with a stochastic process of the kind described in Chapter 2 on pages 57 and 59. We shall not go into technical detail about the full definition of a probability measure for this space. Instead, we shall merely note that not all subsets of the space can be assigned a probability value and that the probability for certain simple events in the infinite space can be obtained by taking the limit as $n \longrightarrow \infty$ of the probabilities for certain corresponding events for random walks of n steps. For example, we can find P(W never moves to left of origin) (in an infinite walk) by taking the limit as $n \longrightarrow \infty$ of P(W does not move to left of origin in n steps). See Exercise 8-15 below. We can go on in this way, proceeding to more and more complex events, and consistently define a unique probability measure for a broad class of events in the infinite sample space. This process is described in an area of mathematics known as measure theory. We do not consider the theory further here.

A variety of interesting questions about random walks can be asked. For example, we can ask for P(W does not return to origin in n steps), and, for an infinite walk, we can ask for P(W never returns to origin). We can also ask questions about how long, on the average, we should expect to wait for a return to the origin. Ruin problems can also be viewed as special kinds of random walk problems. For example, in the game of 20 tosses where A starts with resources a and B starts with resources b, P(A is ruined on or before the 20th toss) is the same as P(W reaches position + a on or before the 20th step and before reaching position -b). Ruin problems are sometimes described as problems about random walks with absorbing barriers. We return to random walk problems in the exercises below.

Markov chains. Consider a random walk of at least five steps. Let A be the event: W at position +3 immediately after the fifth step. Then P(A) is given by the binomial formula as $\binom{5}{1}\frac{1}{2^5} = 5/16$. Now let B_1 be the event: W at position +2 immediately after the fourth step, and let B_2 be the event: W at origin immediately after the fourth step. Evidently, $P(A|B_1) = 1/2$, but $P(A|B_2) = 0$. Let C be the event: W at position +1 immediately after the third step. Evidently, $P(A|C) = 1/4$. On the other hand, $P(A|B_1 \cap C) = P(A|B_1) = 1/2$. Thus the probability that we give to event A is affected by the knowledge that we have of W's previous positions. However, if we know W's immediately preceding position, then no further knowledge of W's other previous positions (before that) will change the probability that we give to A. Such a stochastic process, in which a physical system (in this case W) is pictured as moving from state to state in successive steps in time (in this case the possible states are the possible positions for W) and in which knowledge of the present state provides probabilities (for transitions to other states on the next step) that are unaffected by any additional knowledge of states previous to the present state, is called a Markov chain. A random walk is an example of a Markov chain. Observation of sunny or rainy weather from day to day (see Example 2 in Chapter 4) is a time series experiment for which a Markov chain would not normally provide a good model. P(sunny tomorrow | sunny today) will in general be somewhat different from P(sunny tomorrow | sunny today and sunny yesterday).

It is possible to alter and generalize the concept of Markov chain by replacing the picture of changes at successive given steps in time by a picture in which change occurs at arbitrary moments in (continuously flowing) time. It is also possible, if desired, to have the changes themselves be of a continuously varying nature (in time) rather than of a sudden and discontinuous nature. Such a generalized mathematical picture is called a Markov process. The crucial Markovian assumption is preserved: that the changes are probabilistic, and that the probabilities for further change, given knowledge of the present state, are not altered by having additional knowledge of past states. The mathematical picture of Brownian motion referred to in Chapter 2 (pages 57 and 59) is a Markov process. A notable feature of modern theoretical physics is that the basic mathematical models given by quantum theory for physical systems at the molecular, atomic, and nuclear levels take the form of Markov processes.

<u>Philosophical note</u>. The fact that most current mathematical models for atomic physics are essentially probabilistic has been a major source of discussion and controversy among physicists and philosophers since the introduction of those models in the early part of this century. On one side, many physicists who work with quantum mechanics argue that ultimate physical reality must be, in some sense, probabilistic. It can only be described by pictures in which probability plays a central role. On the other side, some have argued that probabilistic pictures must be incomplete. There must be further pictures, at a level of finer detail, which will describe physical phenomena in deterministic terms. These finer details are sometimes referred to as <u>hidden variables</u>. A chief proponent of the second view was Albert Einstein who asserted his convictions on the matter in the words "God does not play at dice". (Is it, indeed, possible for a mathematical picture (of the physical world) that is deterministic in its fine detail to become probabilistic at a coarser level of detail and then to exhibit such features as the stability laws? Questions of this kind are considered in the area of mathematics known as <u>ergodic theory</u>.) Even if one accepts the view that ultimate microscopic physical reality is probabilistic, it may still be the case that certain macroscopic phenomena which appear probabilistic may, at a finer

level of detail, prove to be deterministic. Current physical and mathematical research in connection with turbulent fluid flow is concerned with this possibility.

The issue of whether or not the ultimate ground of physical reality at the microscopic level can only be pictured in probabilistic terms has been seen by some philosophers as having important implications for such concerns as free will, religion, and the supernatural. It clearly has significant implications for probability theory itself. More specifically, if reality is ultimately probabilistic, then the models of probability theory at all levels of size and detail take on a deeper significance, and it may be that empirical facts like the stability laws express some of the most fundamental aspects of the physical universe.

TABLE OF RANDOM DIGITS

91562	18812	81267	67384	77024	35639	05559	47634	40763	34680
03070	19190	90627	03821	57788	00640	61392	67042	40441	33293
75772	29219	18540	59310	80266	01740	16761	89766	33144	89512
80825	28175	30992	25100	89614	90287	23461	35780	96468	27642
09232	43057	18972	24243	71022	38403	48556	66628	10651	64539
83859	26028	09875	91192	79804	37849	25470	31328	36448	35040
50544	24432	31980	54495	67467	98564	73568	46852	06740	70182
31985	14308	52244	72821	06894	63611	34813	72992	41403	90753
38192	40488	51390	23732	96267	37126	77043	03465	49410	60711
46186	85771	26633	46162	98815	97171	31414	70074	95699	75577
04636	03212	69251	80531	38720	95551	77997	73905	34319	33735
87894	96447	03480	67122	18493	09631	96310	34117	03988	42617
34597	71020	66965	10069	08508	68820	46419	93443	05775	27300
72756	00456	35886	91555	39362	80164	41610	96236	39254	30930
19483	86555	10938	68930	81873	69305	38285	50131	34190	56609
70533	44415	27129	29130	11215	33658	61558	82697	96584	97213
94610	34001	56631	57082	80810	93288	94441	87871	61274	01288
03566	40771	02429	03606	76943	24529	80448	42264	37636	15013
44406	06840	26754	55879	98780	84144	69602	71037	19115	35779
96376	05071	95681	18087	18569	12507	53596	90818	49468	53984
30220	72938	62117	83140	84679	94416	93338	29255	25554	28666
30838	04547	77014	06276	58503	91848	45112	95180	04065	48652
89084	80870	15301	03620	06725	07093	41462	40868	04347	01601
28548	10932	18599	85789	37953	20417	69775	28598	87766	03009
18716	80946	46668	69949	89219	22083	99327	11475	71702	29507
95045	41179	83811	74367	07759	06618	32577	35080	26521	79039
56966	34670	19259	28155	24994	38443	39638	20655	23036	03096
56584	35913	69517	11649	50292	48618	66427	26956	67203	04876
03901	99839	23757	67426	61494	72493	92383	71087	89028	08575
62409	45728	13678	59996	40600	27642	58805	48678	32418	61539
00886	69183	25748	97794	90894	17612	49914	93317	57860	16732
11389	89262	31885	05959	45085	42831	48157	07432	38921	70387
12106	06929	89715	14563	46688	31021	45910	13787	56973	58009
64031	91935	07527	97173	92426	65452	71239	22317	87583	10126
45393	74329	27206	43811	67120	01173	93908	25039	63407	73160
18773	34029	42225	90736	01546	11701	13962	74820	44195	78818
81464	20782	23586	41398	22608	51637	82298	08243	62721	97484
44704	39341	19925	67590	99102	94542	65993	34553	24943	91241
59656	97108	13663	62634	51426	07648	58443	93131	66903	04937
68678	40166	90278	11790	95966	47287	74905	87915	36473	52559
46453	15535	19132	08319	53664	95528	18461	97898	78105	87833
01548	11764	75520	25220	23572	90765	39578	95647	52820	38313
93822	76984	67982	94605	68710	24661	79177	53759	35555	50917
69224	03058	47961	92336	17007	32400	87638	08970	56141	67545
52341	94468	94636	39911	16340	24841	86829	18240	91895	72043
41109	59306	37191	53544	05968	35487	13288	75726	50063	67575
50782	50396	02795	86305	31080	05963	39705	94830	29017	09065
42524	15025	23273	66839	24253	73137	20998	33919	16371	39238
05492	49426	59915	87246	86062	38104	55218	82668	18441	82851
68762	92129	59374	89719	82529	50388	56832	04977	94355	62835

57323	99793	10404	96963	94149	09436	81289	21954	61399	86562
34720	06356	13309	04958	72741	75575	56567	21262	77996	05409
07578	65410	00035	32693	98319	68032	28659	80408	27007	63030
47371	37336	22937	51738	37503	89855	45893	74636	29767	52720
75165	97575	83615	07281	58542	36009	59936	18517	73157	11300
04885	81136	90574	88200	34719	33343	91636	24256	34494	31587
83239	54826	75030	75527	97006	47758	42133	19846	64341	70126
66474	06298	38096	44376	26037	63524	62251	88179	23447	34741
87329	30156	01551	50210	65238	34063	01982	09946	32295	47462
79155	84829	48006	33146	60458	69990	05107	28048	82959	22878
78986	67245	27473	25972	56124	65964	78422	24307	21716	11185
48528	89326	57091	61161	30877	68643	23416	46300	38171	39607
81006	92708	24803	09953	06949	38763	85506	82871	16774	23654
46570	62179	45157	39433	70012	25561	44522	58486	16495	91575
82073	60459	92791	53955	11120	36410	14834	22564	11477	07221
33835	66284	64182	70403	76105	01726	62048	19155	27851	22094
12640	75442	76862	68499	33486	57958	09867	95452	61593	64290
21169	93933	69194	76324	89108	02521	47389	88952	84196	61334
96268	83080	76688	17712	61708	43603	48754	11164	25733	26622
23625	85536	27462	62992	88022	75867	83873	65408	17795	82751
94668	37558	66240	14759	16472	23333	50819	06367	95422	45043
31377	07132	84367	03961	31082	64493	72267	82722	61919	77949
27290	95250	76361	89702	08367	88061	08887	26817	77449	64561
96474	56573	07318	94244	11211	39106	85678	68921	37696	15046
19982	08416	70690	39842	54764	43373	62609	51968	73501	33840
93680	47532	69086	33632	42387	56096	93777	48942	13020	41687
15302	75898	21162	37786	66961	86425	54004	75370	47379	26955
25188	10433	98394	63340	13059	51567	99759	15205	97136	06467
97521	53578	46447	18822	78202	50881	27649	56352	10471	11864
50659	09388	66058	59280	32170	32237	61958	90519	92639	23562
07479	15283	61125	14652	17993	19537	93660	64835	22592	79821
64529	34694	22619	38324	48400	75378	09240	36861	98773	72209
48766	41753	28432	77542	01540	51697	35739	36545	27139	41492
62371	43589	54464	76751	76286	99431	53709	81291	36388	18239
39778	15508	71011	89910	57399	58438	85383	52950	08688	55506
03572	40870	80928	43010	44354	46546	62850	55804	69579	56033
48323	14620	62144	87071	34851	35772	58560	87164	60469	44358
73561	94254	15586	79022	80518	35385	02413	16608	97986	86378
40933	90142	33973	05994	41921	00538	51677	61707	31505	83978
01687	16267	82704	10086	38959	31971	63458	48903	91708	78642
63237	85898	17680	04691	05891	99990	17838	32989	18486	48168
32710	07942	46986	40972	93529	95560	69704	76599	82420	77828
01698	69360	77089	88169	67613	10902	86726	33163	67216	79745
03257	24670	57515	58366	72192	84483	85997	95869	29868	84265
53952	37074	63163	49843	38295	57415	01467	44343	99594	69159
29491	05725	55666	61541	93590	52618	77207	58627	09852	45554
20765	57055	76350	16220	13411	94537	88567	51673	18179	81825
51533	72854	25221	27548	48080	83387	14392	75371	55010	31938
36917	25339	57165	74158	75581	71166	59216	87606	04612	89915
57055	76561	84511	36531	35237	93892	93545	69497	78415	94475

96124	73355	01925	17210	81719	74603	30305	29383	69753	61156
31283	54371	20985	00299	71681	22496	71241	35347	37285	02028
49988	48558	20397	60384	24574	14852	26414	10767	60334	36911
82790	45529	48792	31384	55649	08779	94194	62843	11182	49766
51473	13821	75776	24401	00445	61570	80687	39454	07628	94806
07785	02854	91971	63537	84671	03517	28914	48762	76952	96837
16624	68335	46052	07442	41667	62897	40326	75187	36639	21396
28718	92405	07123	22008	83082	28526	49117	96627	38470	78905
33373	90330	67545	74667	20398	58239	22772	34500	34392	92989
36535	48606	11139	82646	18600	53898	70267	74970	35100	01291
47408	62155	47467	14813	56684	56681	31779	30441	19883	17044
56129	36513	41292	82142	13717	49966	35367	43255	06993	17418
35459	10460	33925	75946	26708	63004	89286	24880	38838	76022
61955	55992	36520	08005	48783	08773	45424	44359	25248	75881
85374	69791	18857	92948	90933	90290	97232	61348	22204	43440
15556	39555	09325	16717	74724	79343	26313	39585	56285	22525
75454	90681	73339	08810	89716	99234	36613	43440	60269	90899
27582	90856	04254	23715	00086	12164	16943	62099	32132	93031
89658	47708	01691	22284	50446	05451	68947	34932	81628	22716
57194	77203	26072	92538	85097	58178	46391	58980	12207	94901
64219	53416	03811	11439	80876	38314	77078	85171	06316	29523
53166	78592	80640	58248	68818	78915	57288	85310	43287	89223
58112	88451	22892	29765	20908	49267	18968	39165	03332	94932
14548	36314	05831	01921	97159	55540	00867	84294	54653	81281
21251	15618	40764	99303	38995	97879	98178	03701	70069	80463
30953	63369	05445	20240	35362	82072	29280	72468	94845	97004
12764	79194	36992	74905	85867	18672	28716	17995	63510	67901
72393	71563	42596	87316	80039	75647	66121	17083	07327	39209
11031	40757	10904	22385	39813	63111	33237	95008	09057	50820
91948	69586	45045	67557	86629	67943	23405	86552	17393	24221
18537	07384	13059	47389	97265	11379	24426	09528	36035	02501
66885	11985	38553	97029	88433	78988	88864	03876	48791	72613
96177	71237	08744	38483	16602	94343	18593	84747	57469	08334
37321	96867	64979	89159	33269	06367	09234	77201	92195	89547
77905	69703	77702	90176	04883	84487	88688	09360	42803	88379
53814	14560	43698	86631	87561	90731	59632	52672	24519	10966
16963	37320	40740	79330	04318	56078	23196	49668	80418	73842
87558	58885	65475	25295	59946	47877	81764	85986	61687	04373
84269	55068	10532	43324	39407	65004	35041	20714	20880	19385
94907	08019	05159	64613	26962	30688	51677	05111	51215	53285
45735	14319	78439	18033	72250	87674	67405	94163	16622	54994
11755	40589	83489	95820	70913	87328	04636	42466	68427	79135
51242	05075	80028	35144	70599	92270	62912	08859	87405	08266
00281	25893	94848	74342	45848	10404	28635	92136	42852	40812
12233	65661	10625	93343	21834	95563	15070	99901	09382	01498
88817	57827	02940	66788	76246	85094	44885	72542	31695	83843
75548	53699	90888	94921	04949	80725	72120	80838	38409	72270
42860	40656	33282	45677	05003	46597	67666	70858	41314	71100
71208	72822	17662	50330	32576	95030	87874	25965	05261	95727
44319	22313	89649	47415	21065	42846	78055	64776	64993	48051

EXERCISES FOR CHAPTER 9.

Remark. Normally, in problems of simulation or estimation by random digits, a computer should be used to get answers that are highly reliable to several decimal places, since the use of a computer will permit the carrying out of sufficiently many trials in a short time. We know, from our study of weak stability and normal approximation, that this reliability can be made extremely high. For example, accuracy to two decimal places (in a probability estimate) will fail at most 5 percent of the time if 10,000 trials are used, and will fail at most 0.3 percent of the time if 22,500 trials are used. It can be shown, using areas under the normal curve, that the chance of failure is less than 1 in 10^4 if 40,000 trials are used, and less than 1 in 10^{22} if 250,000 trials are used. Thus, if a computer with an adequate source of random digits is available, these numerical estimates become, for practical purposes, as precise and reliable as theoretical calculations. A theoretical calculation of a probability (such as the probability in Example 3 above) may provide rewarding insights, but it will provide no practical information (about the specific example) that is not already given by a random digit estimate done on a computer. (With sufficient ingenuity in programming pseudorandom digits, a high level of accuracy for simple problems can also be obtained on a programmable calculator.)

In the exercises which follow, the estimates by random digits are based on a small number of trials and do not require a computer. The accuracy of the estimates is correspondingly low. It is hoped, however, that the exercises will illustrate the general nature and flavor of random digit estimates.

Note. There are often a variety of ways in which a table of random digits can be used for any single problem, and considerable ingenuity can be shown in choosing one of them. In some of the exercises below, we suggest, for convenience in classroom use, specific ways to use the table given in this chapter. The ways suggested may not be as good as some other ways that the reader may think of. They also violate the crossing out principle (see page 215), which asserts that digits should be used once and then discarded. (This is a principle that the reader should normally follow.) The ways suggested in the exercises have the advantage, however, that student solutions to the exercises can be easily compared with a single set of correct solutions.

9-1. (a) Use random digits to simulate 100 throws of two dice. What observed values do you obtain for relative frequencies of the event 7 or 11 occurs and of the event 2, 3, or 12 occurs?

(b) Use random digits to conduct 20 trials of the experiment: roll two dice until either 10 or 7 occurs. What value do you observe for the relative frequency of the event 10 occurs before 7?

(c) For each of the above, give the correct (theoretical) probability value. (See Exercise 4-9.)

(Suggestion. For (a), begin at the top line of the table (page 231), and on each line go all the way across the page, removing digits other than 1, 2, 3, 4, 5, 6. Group the remaining digits in pairs. Use this same sequence of pairs, as far as necessary, for (b).)

9-2. A certain town has an average of 7 fire alarms per week. Use random digits to simulate the Poisson experiment of observing the number x of alarms in each of 50 successive weeks.

(a) What values do you observe for the relative frequencies of the events: $\underline{x = 7}$, $\underline{x = 9}$, and $\underline{x \leq 3}$?

(b) Give the correct probability value for each of these events.

(Suggestion: Consider a binomial experiment with $n = 70$ and $p = 0.1$. From Poisson approximation, we know that the true probabilities for this binomial experiment agree, to two decimal places, with the true probabilities for the Poisson experiment. Therefore, beginning on line 11 of the table, take 50 successive blocks of 70 digits and count the number of 0's occurring in each block.)

9-3. (a) Suppose that in a human population, the four blood types A, B, C, and D occur in 40, 30, 20, and 10 percent of the people respectively. It is proposed that each hospital have, at all times, 4 donors, randomly selected from the population, available to give blood. Assume that applicants to receive blood are selected randomly from the population. Use random digits to simulate 50 hospitals, and estimate the fraction of the applicants, on the average, who would not find correct blood available at the first randomly selected hospital.

(Suggestion. Use lines 8 through 12 in the table. Let digits 1, 2, 3, 4 be type A; 5, 6, 7 be type B; 8, 9 be type C; and 0 be type D. Let each block of five digits represent a hospital, with the first four digits representing the randomly chosen donors and the fifth digit representing a random applicant for blood at that hospital.)

(b) Calculate the true value for P(applicant does not find correct type at first hospital tried). (Hint. For each applicant, failure to find correct blood is the same as failure on all of four corresponding Bernoulli trials. The calculation, surprisingly, is quite elementary.)

9-4. Use random digits to simulate 5 sequences of 100 coin tosses each. In each sequence, check, for $n = 25$, 50, and 100, whether the observed relative frequencies for heads lie within the interval given by the weak stability law. (Suggestion. Use lines 13 through 22 on the first page of the table. Let even digits represent heads and odd digits represent tails.)

9-5. Consider the following experiment. In a block of 20 random digits, observe the number x of distinct digits that appear. (x will be an integer from 1 to 10.)

(a) Conduct 25 trials of this experiment. What relative frequency do you observe for the event $x = 10$?

(Suggestion. Use the top 10 lines on the second page of the table. For each block of 20, use two lines in two columns.)

(b) What are the relative frequencies with which the other values of x have occurred?

(c) Give the correct probability value for the event $x = 10$. (Hint. This is a simple application of the inclusion/exclusion principle; see Exercise 3-15.)

9-6. Describe a convenient and fully correct procedure, using random digits, for randomly selecting the names of 3 individuals listed in the white pages of a telephone book. (Assume that you can always tell the difference between the name of an individual and the name of a business, office, or other institution. Note that the total number of entries on a page may vary from page to page.)

9-7. Use random digits to simulate 50 random draws, without replacement, of 2 cards from the deck of 8 cards described in Example 1 of Chapter 4.

(a) What relative frequencies do you observe corresponding to the events: _two queens_, _at least one queen_, and _at least one red queen_?

(b) What relative frequency values correspond to the conditional probabilities P(_two queens_ | _at least one queen_) and P(_two queens_ | _at least one red queen_)?

(_Suggestion_. Draws without replacement can be done very simply as follows (as the reader may easily justify). Let 1, 2, 3, 4 be the queens, and let 1 and 2 be the red queens. Let 5, 6, 7, 8 be the kings. Discard the digits 9 and 0. Begin with line 11 on the second page. For each draw, take a digit and then the next following digit distinct from it.)

9-8. (a) Carry out a Monte Carlo calculation for π as follows. In the coordinate plane, find coordinates (x,y) to two decimal places for 100 randomly chosen points in the square: $0 \leq x < 1$, $0 \leq y < 1$. Take, as your estimate of π, 4 times the observed frequency of points whose distance from the origin is less than 1. (If a point, such as (0.60, 0.80) lies at distance exactly 1 from the origin, count it (for the relative frequency) as half a point.)

(Suggestion. Use lines 23 through 30 on the first page of the table. Interpret each successive block of 4 digits as a pair of two-digit decimal expressions for the x and y coordinates of a point.)

(b) Estimate the number of points that you would need in order to have probability 0.95 of getting three decimal place accuracy for π. (You may assume that $P(x^2 + y^2 \leq 1) = \pi/4$ to 4 decimal places.)

9-9. Consider the following experiment: a fair coin is tossed until heads appears.

(a) Use random digits to simulate 50 trials of this experiment. What observed value do you get for the average number of tosses up to and including the first heads?

(Suggestion. Begin with line 31 on page 1 of the random digit table. Take even digits as heads and odd digits as tails.)

(b) Let X be the number of tosses in a single trial. Find E_X. (Hint. Apply $\frac{d}{dx}$ to both sides of the identity $\frac{1}{1-x} = 1 + x + x^2 + \ldots$.)

9-10. Consider the following game, where, on each play of the game, a bettor pays 4 dollars for the privilege of playing. The bettor then tosses a coin until heads appears, and receives 2^x dollars, where x is the number of tosses.

Use random digits to simulate 50 plays of this game. What observed value do you get for the bettor's average winnings for these 50 plays. (Suggestion. Use the same observed values for x as obtained for Exercise 9-9.)

(Recall from page 216 of Chapter 8 that if Y is the bettor's winnings for a single play, then E_Y does not exist. In fact, it is possible to show that if the game is played repeatedly, the observed average winnings per play will eventually exceed any given bound. This result is known, for historical reasons, as the St. Petersburg paradox, and the game is called the St. Petersburg game.)

9-11. Use random digits to simulate 25 trials of each of the following experiments, and use the data obtained, in each case, to estimate the desired probability.

(a) Six cards are drawn, with replacement, from a shuffled bridge deck. Estimate P(all four suits appear).

(b) Six cards are drawn, <u>without</u> replacement, from a shuffled bridge deck. Estimate P(<u>all four suits appear</u>).

(c) Five cards are drawn, <u>without</u> replacement, from a shuffled bridge deck. Estimate P(<u>exactly one suit is missing</u>).

(d) For each of (a), (b), and (c), give the correct theoretical value. (See Exercises 3-15c, 3-15d, and 3-18a.)

(<u>Suggestion</u>. For each case, begin with line 34 of the first page. For (a), let 1, 2, 3, 4 be the suits, and discard the other digits. For (b) and (c), let the pairs of digits 01 through 13, 14 through 26, 27 through 39, and 40 through 52 represent the four suits. Discard the other pairs and ignore repeated pairs on any trial.)

9-12. In the classical symmetrical ruin problem, consider the case where A begins with resources $\underline{a} = 2$ dollars, and B begins with resources $\underline{b} = 3$ dollars.

(a) Use random digits to simulate 36 games in which either A or B is eventually ruined. What relative frequency do you observe for the event <u>A is ruined</u>?

(<u>Suggestion</u>. Use lines 46 through 50 on page 1 of the table. Use even digits for <u>heads</u> and odd digits for <u>tails</u>. A is betting on heads.)

(b) Find the correct (theoretical) value for P(A is ruined). (Hint. For all integers z, $-2 \leq z \leq 3$, let $A(z)$ = P(A is eventually ruined | A now has net winnings z). Then (i) $A(-2) = 1$ and $A(3) = 0$. Show that (ii): if $-2 < z < 3$, $A(z) = \frac{1}{2}A(z+1) + \frac{1}{2}A(z-1)$. Use (i) and (ii) to find $A(0)$ = P(A is ruined).)

(c) Show that the general solution to the classical symmetrical ruin problem, for arbitrary a and b, is P(A is ruined) $= \frac{b}{b+a}$. (Hint. For all z, $-a \leq z \leq b$, define $A(z)$ as in the hint for (b). $A(z)$ must satisfy appropriate versions of conditions (i) and (ii). It will be enough to show that for all z, $-a \leq z \leq b$, $A(z) = \frac{b-z}{b+a}$. To show this, verify that the function $\frac{b-z}{b+a}$ satisfies (i) and (ii) and that any other function that does so must differ from it by a function which is zero for all z, such that $-a \leq z \leq b$.)

9-13. Consider a non-symmetrical ruin problem which is identical with the classical symmetrical problem except that the coin being tossed has probability 4/5 of coming up heads.

(a) Assume $a = 1$ dollar and $b = 2$ dollars. Use random digits to simulate games in which either A or B is eventually ruined. What relative frequency do you observe for the event A is ruined? (Suggestion. Use lines 14 through 18 of page 2 of the table. Use digits 0,1,...,7 for heads and 8, 9 for tails.)

(b) Find the correct (theoretical) value for P(A is ruined). (Hint. As in the hints for Exercise 8-12, define A(z) = P(A is eventually ruined | A now has net winnings z). Then (i) $A(-1) = 1$ and $A(2) = 0$. Show that (ii): If $-1 < z < 2$, $A(z) = 4/15\,A(z+1) + 1/5\,A(z-1)$. Use (i) and (ii) to find A(0) = P(A is ruined).)

(c) Assume that the coin being tossed has probability p of heads and q of tails, where $p \neq q$. Show that the general solution to this classical non-symmetrical ruin problem, for arbitrary a and b, is $P(\text{A is ruined}) = \dfrac{(\frac{p}{q})^{b} - 1}{(\frac{p}{q})^{b+a} - 1}$. (Hint. Proceed as in the hint for Exercise 8-12c, using

$$A(z) = \frac{(\frac{p}{q})^{b-z} - 1}{(\frac{p}{q})^{b+a} - 1}.)$$

9-14. Ten votes are cast in an election for president of a club. Candidate A receives 7 votes and candidate B receives 3 votes. The voters have voted by writing their choices on identical pieces of paper called ballots. Thus there are ten ballots, 7 for A and 3 for B. The ballots are counted in random order and, while the ballots are being counted, cumulative totals are kept of the votes that each candidate has received so far. We seek to find the probability that, beginning with the counting of the first ballot, the total for A is always greater than the total for B. We shall call this P(A always ahead).

(a) Use random digits to simulate 20 trials of the experiment of counting the ballots. What relative frequency do you observe for the event A always ahead? (Suggestion. Use lines 19 through 28 on page 2. Let 0,1,2,...,6 be ballots for A and let 7,8,9 be ballots for B. In each experiment, discard repetitions of ballots already counted.)

(b) Find the correct (theoretical) value for P(A always ahead). (Hint. Consider a rectangular grid 7 spaces wide and 3 spaces high. Assume that as the votes are counted, a marker-point is moved along the lines of the grid from the lower left corner P to the upper right corner Q of the grid. Each vote counted for A causes a horizontal move of one space to the right; each vote counted for B causes a vertical move of one space upwards. Thus any particular counting of the votes generates a path from P to Q.

(i) Show that there are $\binom{10}{7}$ distinct paths of this kind and that they have equal probability of occurring. (ii) Show that the number of these paths in which the first vote counted goes to B is $\binom{9}{7}$. (iii) Show by a reflection argument similar to that used with Figure 8.2 that the number of paths in which A gets the first vote but is tied at some later point by B must also be $\binom{9}{7}$. (iv) Hence conclude that the number of paths for which A is always ahead is $\binom{10}{7} - 2\binom{9}{7}$. (v) Find P(A always ahead) from (iv) and (i).)

(c) Generalize the result of (b) to the case where, in an election in which m + n votes are cast, A receives m votes and B receives n votes, with m > n. (This classical problem is known as the ballot-box problem).

9-15. (a) Use random digits to simulate 20 trials of a symmetrical one-dimensional random walk of 10 steps. What relative frequency do you observe for the event W returns to origin on or before the 10th step? (Suggestion. Use lines 29-33 on page 2. Take even digits for moves in the positive direction and odd digits for moves in the negative direction.)

(b) Use random digits to simulate 20 trials of a random walk of 20 steps. What relative frequency do you observe for the event W returns to origin on or before the 20th step? (Suggestion. Use lines 33-40 on page 2 of the table.)

(c) Find the correct (theoretical) values for P(W returns to origin on or before 10th step) and P(W returns to origin on or before 20th step).

(d) Show that P(W returns on or before $2n^{th}$ step) = P(W returns on or before $(2n+1)^{st}$ step). Use Stirling's formula to show P(W returns on or before $2n^{th}$ step) $\approx 1 - \frac{1}{\sqrt{n\pi}}$. Let $n \longrightarrow \infty$ and find, for an infinite symmetrical one-dimensional random walk, P(W returns to origin).

(e) Conclude from (d) that the probability of returning to the origin from any point P must be 1 and hence, given any P, that the probability of starting at the origin and passing through P is also 1.

(f) Use (e) to conclude that if two random walkers, W_1 and W_2, start simultaneously at two points P_1 and P_2, the probability is 1 that they eventually meet. (Hint. Let x_n be the position of W_1 after the n^{th} step and let y_n be the position of W_2 after the n^{th} step. Consider successive changes in the quantity $x_n - y_n$.)

Comment. (i) The arc sine law can be modified in a minor way to show that in the fair coin-tossing game between A and B, it is more probable that A is ahead by at least one dollar for the whole time than that A is ahead by at least one dollar for some given smaller positive fraction of the time. (ii) The results of 8-15d and 8-15e above now show that no matter how far ahead A gets, the probability is 1 that eventually, at some later time, A's winnings will be reduced to zero. (i) and (ii) appear to be contrary in intuitive content, yet both are true. Together they give added insight into the fluctuations that must occur in tossing a fair coin. (ii) tells us that a graph of the Figure 8.1 kind must return to the horizontal axis again and again. At the same time, (i) suggests that these returns must occur rarely, and usually after long intervals. (The average length of these intervals can, in fact, be shown to increase without bound as an infinite random walk continues.)

9-16. A symmetrical two-dimensional random walk occurs when W begins at the origin of the coordinate plane and makes successive moves of unit length in a direction parallel to a coordinate axis. Each of the four directions occurs with probability 1/4, and the successive moves are independent. Similarly, a symmetrical three-dimensional random walk occurs when W begins at the origin of three-dimensional cartesian coordinates and, on each step, moves 1 unit in one of the six possible directions parallel to a coordinate axis. The moves are independent, and each direction occurs with probability 1/6.

(a) Use random digits to simulate 20 trials of a two-dimensional random walk of 20 steps. What relative frequency do you observe for the event W returns to origin on or before 20th step? (*Suggestion*. Use the digits 1,2,3,4 for the directions *north*, *south*, *east*, *west*, respectively. Discard the other digits. Use lines 1-22 on page 3 of the table.)

(b) Use random digits to simulate 20 trials of a three-dimensional random walk of 20 steps. What relative frequency do you observe for the event W returns to origin on or before 20th step. (*Suggestion*. Use the digits 1,2,3,4,5,6 for the directions *north*, *south*, *east*, *west*, *up*, *down*, respectively. Discard 0,7,8,9. Use lines 23-36 on page 3 of the table.)

Remark. It can be shown theoretically that for the infinite, symmetric, two-dimensional random walk, P(W returns to origin) = 1, but that for the three-dimensional walk, P(W returns to origin) ≈ 0.35. It can be shown to follow, as in (e) and (f) of Exercise 8-15, that the probability that two simultaneous random walkers eventually meet is always equal to 1 in two dimensions, but that this probability can be much less than 1 in three dimensions.

9-17. Assume that in a certain country each day is either sunny or rainy and that the sequence of sunny and rainy days can be correctly pictured as a Markov chain. Assume further that in this Markov chain P(sunny tomorrow | sunny today) = 0.7 and P(sunny tomorrow | rainy today) = 0.2.

(a) Use random digits to simulate the experiment of observing 200 successive days. What relative frequency do you observe for the event sunny? (*Suggestion*. Use lines 41-44 on page 2. Given that today is sunny, let the digits 0,1,...,6 indicate that tomorrow is sunny. Given that today is rainy, let the digits 0,1 indicate that tomorrow is sunny. Assume that you begin on a sunny day; but do not count this initial day in your observed relative frequency.)

(b) Find a theoretical value for the probability that a randomly chosen day in this country is sunny. (*Hint*. Begin by assuming that P(sunny) has a single value which is the same for each individual day. Call this probability s. Verify from the laws of probability that

P(tomorrow sunny) =

P(today sunny)P(tomorrow sunny | today sunny)

+ P(today rainy)P(tomorrow sunny | today rainy).

Substitute s in this equation in an appropriate way and solve for s.)

(c) A run is a block of successive days of the same kind which is immediately preceded and immediately followed by a day of the opposite kind. In the data obtained in (a), what is the average length of the runs of sunny days observed and the average length of the runs of rainy days?

(d) Use random digits to simulate a sequence of 200 Bernoulli trials with $p = 0.4$. Define runs of successes and runs of failures as in (c). What is the average length of the runs of successes observed and the average length of the runs of failures? (Suggestion. Use lines 35-38 on page 2 with digits 0,1,2,3 for success.)

(e) Find the expected value for the length of a run of successes, and the expected value for the length of a run of failures.

(Hint. The calculations are similar to that for Exercise 9-9b.)

(f) Consider Bernoulli trials with individual success probability p. Let $\bar{s}$ be the expected length of a run of successes and let $\bar{f}$ be the expected length of a run of failures.
Show that $\frac{1}{\bar{s}} + \frac{1}{\bar{f}} = 1.$

Comment. The fact that the observed averages in (c) do not come close to satisfying the algebraic relation in (f) would be a strong indication, in the absence of other information, that the observed sequence of sunny and rainy days could not be correctly pictured as a sequence of independent Bernoulli trials. Questions of this kind, about whether models do or do not give reasonable pictures on the basis of observed data, are the subject matter of mathematical statistics, where they are treated quantitatively. We begin a study of statistics in the following chapters.