28. [*40*] (A. G. Waterman.) Experiment with linear congruential sequences with m the square or cube of the computer word size, while a and c are single-precision numbers.

▸ **29.** [*40*] Find a good way to compute the function $f(x_1, \ldots, x_k)$ defined by Martin's sequence in exercise 17, given only the k-tuple $(x_1, \ldots, x_k)$.

30. [*M37*] (R. P. Brent.) Let $f(x) = x^k - a_1x^{k-1} - \cdots - a_k$ be a primitive polynomial modulo 2, and suppose that $X_0, \ldots, X_{k-1}$ are integers not all even.

a) Prove that the period of the recurrence $X_n = (a_1X_{n-1} + \cdots + a_kX_{n-k}) \bmod 2^e$ is $2^{e-1}(2^k - 1)$ for all $e \geq 1$ if and only if $f(x)^2 + f(-x)^2 \not\equiv 2f(x^2)$ and $f(x)^2 + f(-x)^2 \not\equiv 2(-1)^kf(-x^2)$ (modulo 8). [*Hint:* We have $x^{2^k} \equiv -x$ (modulo 4 and $f(x)$) if and only if $f(x)^2 + f(-x)^2 \equiv 2f(x^2)$ (modulo 8).]

b) Prove that this condition always holds when the polynomial $f(x) = x^k \pm x^l \pm 1$ is primitive modulo 2 and $k > 2$.

31. [*M30*] (G. Marsaglia.) What is the period length of the sequence (7′) when $m = 2^e \geq 8$? Assume that $X_0, \ldots, X_{54}$ are not all $\equiv \pm 1$ (modulo 8).

32. [*M21*] What recurrences are satisfied by the elements of the subsequences $\langle X_{2n} \rangle$ and $\langle X_{3n} \rangle$, when $X_n = (X_{n-24} + X_{n-55}) \bmod m$?

▸ **33.** [*M23*] (a) Let $g_n(z) = X_{n+30} + X_{n+29}z + \cdots + X_nz^{30} + X_{n+54}z^{31} + \cdots + X_{n+31}z^{54}$, where the X's satisfy the lagged Fibonacci recurrence (7). Find a simple relation between $g_n(z)$ and $g_{n+t}(z)$. (b) Express X_{500} in terms of $X_0, \ldots, X_{54}$.

34. [*M25*] Prove that the inversive congruential sequence (12) has period $p + 1$ if and only if the polynomial $f(x) = x^2 - cx - a$ has the following two properties: (i) $x^{p+1} \bmod f(x)$ is a nonzero constant, when computed with polynomial arithmetic modulo p; (ii) $x^{(p+1)/q} \bmod f(x)$ has degree 1 for every prime q that divides $p+1$. [*Hint:* Consider powers of the matrix $\left(\begin{smallmatrix} 0 & 1 \\ a & c \end{smallmatrix}\right)$.]

35. [*HM35*] How many pairs (a, c) satisfy the conditions of exercise 34?

36. [*M25*] Prove that the inversive congruential sequence $X_{n+1} = (aX_n^{-1} + c) \bmod 2^e$, $X_0 = 1$, $e \geq 3$, has period length 2^{e-1} whenever $a \bmod 4 = 1$ and $c \bmod 4 = 2$.

▸ **37.** [*HM32*] Let p be prime and assume that $X_{n+1} = (aX_n^{-1} + c) \bmod p$ defines an inversive congruential sequence of period $p + 1$. Also let $0 \leq b_1 < \cdots < b_d \leq p$, and consider the set

$$V = \{(X_{n+b_1}, X_{n+b_2}, \ldots, X_{n+b_d}) \mid 0 \leq n \leq p \text{ and } X_{n+b_j} \neq \infty \text{ for } 1 \leq j \leq d\}.$$

This set contains $p + 1 - d$ vectors, any d of which lie in some $(d-1)$-dimensional hyperplane $H = \{(v_1, \ldots, v_d) \mid r_1v_1 + \cdots + r_dv_d \equiv r_0 \text{ (modulo } p)\}$, where $(r_1, \ldots, r_d) \neq (0, \ldots, 0)$. Prove that no $d + 1$ vectors of V lie in the same hyperplane.

3.3. STATISTICAL TESTS

OUR MAIN PURPOSE is to obtain sequences that behave as if they are random. So far we have seen how to make the period of a sequence so long that for practical purposes it never will repeat; this is an important criterion, but it by no means guarantees that the sequence will be useful in applications. How then are we to decide whether a sequence is sufficiently random?

If we were to give some randomly chosen man a pencil and paper and ask him to write down 100 random decimal digits, chances are very slim that he would produce a satisfactory result. People tend to avoid things that seem nonrandom, such as pairs of equal adjacent digits (although about one out of every 10 digits should equal its predecessor). And if we would show that same man a table of truly random digits, he would quite probably tell us they are not random at all; his eye would spot certain apparent regularities.

According to Dr. I. J. Matrix (as quoted by Martin Gardner in *Scientific American*, January, 1965), "Mathematicians consider the decimal expansion of π a random series, but to a modern numerologist it is rich with remarkable patterns." Dr. Matrix has pointed out, for example, that the first repeated two-digit number in π's expansion is 26, and its second appearance comes in the middle of a curious repetition pattern:

$$3.14159\overset{\frown}{26}53589\underset{\smile}{79}\underset{\smile}{32}\underset{\smile}{38}46\overset{\frown}{26}43\underset{\smile}{38}\underset{\smile}{32}\underset{\smile}{79}50 \qquad (1)$$

After listing a dozen or so further properties of these digits, he observed that π, when correctly interpreted, conveys the entire history of the human race!

We all notice patterns in our telephone numbers, license numbers, etc., as aids to memory. The point of these remarks is that we cannot be trusted to judge by ourselves whether a sequence of numbers is random or not. Some unbiased mechanical tests must be applied.

The theory of statistics provides us with some quantitative measures for randomness. There is literally no end to the number of tests that can be conceived; we will discuss the tests that have proved to be most useful, most instructive, and most readily adapted to computer calculation.

If a sequence behaves randomly with respect to tests $T_1, T_2, \ldots, T_n$, we cannot be *sure* in general that it will not be a miserable failure when it is subjected to a further test T_{n+1}. Yet each test gives us more and more confidence in the randomness of the sequence. In practice, we apply about half a dozen different kinds of statistical tests to a sequence, and if it passes them satisfactorily we consider it to be random — it is then presumed innocent until proven guilty.

Every sequence that is to be used extensively should be tested carefully, so the following sections explain how to administer the tests in an appropriate way. Two kinds of tests are distinguished: *empirical tests*, for which the computer manipulates groups of numbers of the sequence and evaluates certain statistics; and *theoretical tests*, for which we establish characteristics of the sequence by using number-theoretic methods based on the recurrence rule used to form the sequence.

If the evidence doesn't come out as desired, the reader may wish to try the techniques in *How to Lie With Statistics* by Darrell Huff (Norton, 1954).

3.3.1. General Test Procedures for Studying Random Data

A. "Chi-square" tests. The chi-square test (χ^2 test) is perhaps the best known of all statistical tests, and it is a basic method that is used in connection with many other tests. Before considering the idea in general, let us consider a particular example of the chi-square test as it might be applied to dice throwing. Using two "true" dice (each of which, independently, is assumed to yield the values 1, 2, 3, 4, 5, or 6 with equal probability), the following table gives the probability of obtaining a given total, s, on a single throw:

value of $s =$	2	3	4	5	6	7	8	9	10	11	12
probability, $p_s =$	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{9}$	$\frac{5}{36}$	$\frac{1}{6}$	$\frac{5}{36}$	$\frac{1}{9}$	$\frac{1}{12}$	$\frac{1}{18}$	$\frac{1}{36}$

(1)

For example, a value of 4 can be thrown in three ways: $1+3$, $2+2$, $3+1$; this constitutes $\frac{3}{36} = \frac{1}{12} = p_4$ of the 36 possible outcomes.

If we throw the dice n times, we should obtain the value s approximately np_s times on the average. For example, in 144 throws we should get the value 4 about 12 times. The following table shows what results were *actually* obtained in a particular sequence of 144 throws of the dice:

value of $s =$	2	3	4	5	6	7	8	9	10	11	12
observed number, $Y_s =$	2	4	10	12	22	29	21	15	14	9	6
expected number, $np_s =$	4	8	12	16	20	24	20	16	12	8	4

(2)

Notice that the observed number was different from the expected number in all cases; in fact, random throws of the dice will hardly ever come out with *exactly* the right frequencies. There are 36^{144} possible sequences of 144 throws, all of which are equally likely. One of these sequences consists of all 2s ("snake eyes"), and anyone throwing 144 snake eyes in a row would be convinced that the dice were loaded. Yet the sequence of all 2s is just as probable as any other particular sequence if we specify the outcome of each throw of each die.

In view of this, how can we test whether or not a given pair of dice is loaded? The answer is that we can't make a definite yes-no statement, but we can give a *probabilistic* answer. We can say how probable or improbable certain types of events are.

A fairly natural way to proceed in the example above is to consider the squares of the differences between the observed numbers Y_s and the expected numbers np_s. We can add these together, obtaining

$$V = (Y_2 - np_2)^2 + (Y_3 - np_3)^2 + \cdots + (Y_{12} - np_{12})^2. \tag{3}$$

A bad set of dice should result in a relatively high value of V; and for any given value of V we can ask, "What is the probability that V is this high, using true dice?" If this probability is very small, say $\frac{1}{100}$, we would know that only about one time in 100 would true dice give results so far away from the expected numbers, and we would have definite grounds for suspicion. (Remember, however, that even *good* dice would give such a high value of V about one time in a hundred, so a cautious person would repeat the experiment to see if the high value of V is repeated.)

The statistic V in (3) gives equal weight to $(Y_7 - np_7)^2$ and $(Y_2 - np_2)^2$, although $(Y_7 - np_7)^2$ is likely to be a good deal higher than $(Y_2 - np_2)^2$ since 7s occur about six times as often as 2s. It turns out that the "right" statistic, at least one that has proved to be most important, will give $(Y_7 - np_7)^2$ only $\frac{1}{6}$ as much weight as $(Y_2 - np_2)^2$, and we should change (3) to the following formula:

$$V = \frac{(Y_2 - np_2)^2}{np_2} + \frac{(Y_3 - np_3)^2}{np_3} + \cdots + \frac{(Y_{12} - np_{12})^2}{np_{12}}. \tag{4}$$

This is called the "chi-square" statistic of the observed quantities $Y_2, \ldots, Y_{12}$ in the dice-throwing experiment. For the data in (2), we find that

$$V = \frac{(2-4)^2}{4} + \frac{(4-8)^2}{8} + \cdots + \frac{(9-8)^2}{8} + \frac{(6-4)^2}{4} = 7\frac{7}{48}. \tag{5}$$

The important question now is, of course, "Does $7\frac{7}{48}$ constitute an improbably high value for V to assume?" Before answering this question, let us consider the general application of the chi-square method.

In general, suppose that every observation can fall into one of k categories. We take *n independent observations*; this means that the outcome of one observation has absolutely no effect on the outcome of any of the others. Let p_s be the probability that each observation falls into category s, and let Y_s be the number of observations that actually *do* fall into category s. We form the statistic

$$V = \sum_{s=1}^{k} \frac{(Y_s - np_s)^2}{np_s}. \tag{6}$$

In our example above, there are eleven possible outcomes of each throw of the dice, so $k = 11$. (Eq. (6) is a slight change of notation from Eq. (4), since we are numbering the possibilities from 1 to k instead of from 2 to 12.)

By expanding $(Y_s - np_s)^2 = Y_s^2 - 2np_sY_s + n^2p_s^2$ in (6), and using the facts that

$$\begin{aligned} Y_1 + Y_2 + \cdots + Y_k &= n, \\ p_1 + p_2 + \cdots + p_k &= 1, \end{aligned} \tag{7}$$

we arrive at the formula

$$V = \frac{1}{n} \sum_{s=1}^{k} \left(\frac{Y_s^2}{p_s} \right) - n, \tag{8}$$

which often makes the computation of V somewhat easier.

Table 1

SELECTED PERCENTAGE POINTS OF THE CHI-SQUARE DISTRIBUTION

	$p = 1\%$	$p = 5\%$	$p = 25\%$	$p = 50\%$	$p = 75\%$	$p = 95\%$	$p = 99\%$
$\nu = 1$	0.00016	0.00393	0.1015	0.4549	1.323	3.841	6.635
$\nu = 2$	0.02010	0.1026	0.5754	1.386	2.773	5.991	9.210
$\nu = 3$	0.1148	0.3518	1.213	2.366	4.108	7.815	11.34
$\nu = 4$	0.2971	0.7107	1.923	3.357	5.385	9.488	13.28
$\nu = 5$	0.5543	1.1455	2.675	4.351	6.626	11.07	15.09
$\nu = 6$	0.8721	1.635	3.455	5.348	7.841	12.59	16.81
$\nu = 7$	1.239	2.167	4.255	6.346	9.037	14.07	18.48
$\nu = 8$	1.646	2.733	5.071	7.344	10.22	15.51	20.09
$\nu = 9$	2.088	3.325	5.899	8.343	11.39	16.92	21.67
$\nu = 10$	2.558	3.940	6.737	9.342	12.55	18.31	23.21
$\nu = 11$	3.053	4.575	7.584	10.34	13.70	19.68	24.72
$\nu = 12$	3.571	5.226	8.438	11.34	14.85	21.03	26.22
$\nu = 15$	5.229	7.261	11.04	14.34	18.25	25.00	30.58
$\nu = 20$	8.260	10.85	15.45	19.34	23.83	31.41	37.57
$\nu = 30$	14.95	18.49	24.48	29.34	34.80	43.77	50.89
$\nu = 50$	29.71	34.76	42.94	49.33	56.33	67.50	76.15
$\nu > 30$	$\nu + \sqrt{2\nu}x_p + \frac{2}{3}x_p^2 - \frac{2}{3} + O(1/\sqrt{\nu})$						
$x_p =$	-2.33	-1.64	$-.674$	0.00	0.674	1.64	2.33

(For further values, see *Handbook of Mathematical Functions*, edited by M. Abramowitz and I. A. Stegun (Washington, D.C.: U.S. Government Printing Office, 1964), Table 26.8. See also Eq. (22) and exercise 16.)

Now we turn to the important question, "What constitutes a reasonable value of V?" This is found by referring to a table such as Table 1, which gives values of "the chi-square distribution with ν degrees of freedom" for various values of ν. The line of the table with $\nu = k-1$ is to be used; *the number of "degrees of freedom" is $k-1$, one less than the number of categories.* (Intuitively, this means that $Y_1, Y_2, \ldots, Y_k$ are not completely independent, since Eq. (7) shows that Y_k can be computed if $Y_1, \ldots, Y_{k-1}$ are known; hence, $k-1$ degrees of freedom are present. This argument is not rigorous, but the theory below justifies it.)

If the table entry in row ν under column p is x, it means, "The quantity V in Eq. (8) will be less than or equal to x with approximate probability p, if n is large enough." For example, the 95 percent entry in row 10 is 18.31; we will have $V > 18.31$ only about 5 percent of the time.

Let us assume that our dice-throwing experiment has been simulated on a computer using some sequence of supposedly random numbers, with the following results:

$$\begin{array}{lcccccccccccc} & \text{value of } s = & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ \text{Experiment 1,} & Y_s = & 4 & 10 & 10 & 13 & 20 & 18 & 18 & 11 & 13 & 14 & 13 \\ \text{Experiment 2,} & Y_s = & 3 & 7 & 11 & 15 & 19 & 24 & 21 & 17 & 13 & 9 & 5 \end{array} \qquad (9)$$

We can compute the chi-square statistic in the first case, getting the value $V_1 = 29\frac{59}{120}$, and in the second case we get $V_2 = 1\frac{17}{120}$. Referring to the table entries for 10 degrees of freedom, we see that V_1 is *much too high*; V will be greater than 23.21 only about one percent of the time! (By using more extensive tables, we find in fact that V will be as high as V_1 only 0.1 percent of the time.) Therefore Experiment 1 represents a significant departure from random behavior.

On the other hand, V_2 is quite low, since the observed values Y_s in Experiment 2 are quite close to the expected values np_s in (2). The chi-square table tells us, in fact, that V_2 is *much too low*: The observed values are so close to the expected values, we cannot consider the result to be random! (Indeed, reference to other tables shows that such a low value of V occurs only 0.03 percent of the time when there are 10 degrees of freedom.) Finally, the value $V = 7\frac{7}{48}$ computed in (5) can also be checked with Table 1. It falls between the entries for 25 percent and 50 percent, so we cannot consider it to be significantly high or significantly low; thus the observations in (2) are satisfactorily random with respect to this test.

It is somewhat remarkable that the same table entries are used no matter what the value of n is, and no matter what the probabilities p_s are. Only the number $\nu = k-1$ affects the results. In actual fact, however, the table entries are not exactly correct: *The chi-square distribution is an approximation that is valid only for large enough values of n.* How large should n be? A common rule of thumb is to take n large enough so that each of the expected values np_s is five or more; preferably, however, take n much larger than this, to get a more powerful test. In our examples above we took $n = 144$, so np_2 was only 4, violating the stated rule of thumb. This was done only because the author tired of throwing the dice; it makes the entries in Table 1 less accurate for our application. Experiments run on a computer, with $n = 1000$, or 10000, or even 100000, would be much better than this. We could also combine the data for $s = 2$ and $s = 12$; then the test would have only nine degrees of freedom but the chi-square approximation would be more accurate.

We can get an idea of how crude an approximation is involved by considering the case when there are only two categories, having probabilities p_1 and p_2. Suppose $p_1 = \frac{1}{4}$ and $p_2 = \frac{3}{4}$. According to the stated rule of thumb, we should have $n \geq 20$ to have a satisfactory approximation, so let's check that out. When $n = 20$, the possible values of V are $(Y_1 - 5)^2/5 + (5 - Y_1)^2/15 = \frac{4}{15}r^2$ for $-5 \leq r \leq 15$; we wish to know how well the row $\nu = 1$ of Table 1 describes the distribution of V. The chi-square distribution varies continuously, while the actual distribution of V has rather big jumps, so we need some convention for representing the exact distribution. If the distinct possible outcomes of the experiment lead to the values $V_0 \le V_1 \le \cdots \le V_n$ with respective probabilities $\pi_0, \pi_1, \ldots, \pi_n$, suppose that a given percentage p falls in the range $\pi_0 + \cdots + \pi_{j-1} < p < \pi_0 + \cdots + \pi_{j-1} + \pi_j$. We would like to represent p by a "percentage point" x such that V is less than x with probability $\le p$ and V is greater than x with probability $\le 1-p$. It is not difficult to see that the only such number is $x = V_j$. In our example for $n = 20$ and $\nu = 1$, it turns out that the percentage points of the exact distribution, corresponding to the approximations in Table 1 for $p = 1\%, 5\%, 25\%, 50\%, 75\%, 95\%$, and 99%, respectively, are

$$0, \quad 0, \quad .27, \quad .27, \quad 1.07, \quad 4.27, \quad 6.67$$

(to two decimal places). For example, the percentage point for $p = 95\%$ is 4.27, while Table 1 gives the estimate 3.841. The latter value is too low; it tells us (incorrectly) to reject the value $V = 4.27$ at the 95% level, while in fact the probability that $V \ge 4.27$ is more than 6.5%. When $n = 21$, the situation changes slightly because the expected values $np_1 = 5.25$ and $np_2 = 15.75$ can never be obtained exactly; the percentage points for $n = 21$ are

$$.02, \quad .02, \quad .14, \quad .40, \quad 1.29, \quad 3.57, \quad 5.73.$$

We would expect Table 1 to be a better approximation when $n = 50$, but the corresponding tableau actually turns out to be further from Table 1 in some respects than it was for $n = 20$:

$$.03, \quad .03, \quad .03, \quad .67, \quad 1.31, \quad 3.23, \quad 6.$$

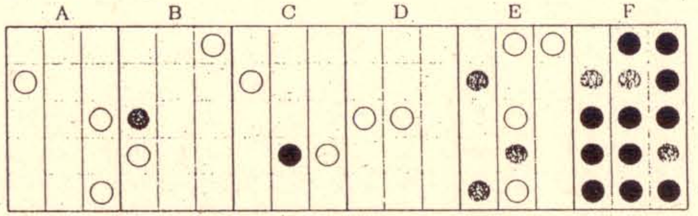
Here are the values when $n = 300$:

$$0, \quad 0, \quad .07, \quad .44, \quad 1.44, \quad 4, \quad 6.42.$$

Even in this case, when np_s is ≥ 75 in each category, the entries in Table 1 are good to only about one significant digit.

The proper choice of n is somewhat obscure. If the dice are actually biased, the fact will be detected as n gets larger and larger. (See exercise 12.) But large values of n will tend to smooth out *locally* nonrandom behavior, when blocks of numbers with a strong bias are followed by blocks of numbers with the opposite bias. Locally nonrandom behavior is not an issue when actual dice are rolled, since the same dice are used throughout the test, but a sequence of numbers generated by computer might very well display such anomalies. Perhaps a chi-square test should be made for several different values of n. At any rate, n should always be rather large.

We can summarize the chi-square test as follows. A fairly large number, n, of independent observations is made. (It is important to avoid using the chi-square method unless the observations are independent. See, for example, exercise 10, which considers the case when half of the observations depend on the other half.) We count the number of observations falling into each of k categories and compute the quantity V given in Eqs. (6) and (8). Then V is compared with the numbers in Table 1, with $\nu = k - 1$. If V is less than the 1% entry or greater than the 99% entry, we reject the numbers as not sufficiently random. If V lies



Range of V	Indication	Code
0–1 percent, 99–100 percent	Reject	●
1–5 percent, 95–99 percent	Suspect	▨
5–10 percent, 90–95 percent	Almost suspect	○

Fig. 2. Indications of "significant" deviations in 90 chi-square tests (see also Fig. 5).

between the 1% and 5% entries or between the 95% and 99% entries, the numbers are "suspect"; if (by interpolation in the table) V lies between the 5% and 10% entries, or the 90% and 95% entries, the numbers might be "almost suspect." The chi-square test is often done at least three times on different sets of data, and if at least two of the three results are suspect the numbers are regarded as not sufficiently random.

For example, see Fig. 2, which shows schematically the results of applying five different types of chi-square tests on each of six sequences of random numbers. Each test in this illustration was applied to three different blocks of numbers of the sequence. Generator A is the MacLaren–Marsaglia method (Algorithm 3.2.2M applied to the sequences in 3.2.2–(13)); Generator E is the Fibonacci method, 3.2.2–(5); and the other generators are linear congruential sequences with the following parameters:

Generator B: $X_0 = 0, \quad a = 3141592653, \quad c = 2718281829, \quad m = 2^{35}$.
Generator C: $X_0 = 0, \quad a = 2^7 + 1, \quad c = 1, \quad m = 2^{35}$.
Generator D: $X_0 = 47594118, \quad a = 23, \quad c = 0, \quad m = 10^8 + 1$.
Generator F: $X_0 = 314159265, \quad a = 2^{18} + 1, \quad c = 1, \quad m = 2^{35}$.

From Fig. 2 we conclude that (so far as these tests are concerned) Generators A, B, D are satisfactory, Generator C is on the borderline and should probably be rejected, Generators E and F are definitely unsatisfactory. Generator F has, of course, low potency; Generators C and D have been discussed in the literature, but their multipliers are too small. (Generator D is the original multiplicative generator proposed by Lehmer in 1948; Generator C is the original linear congruential generator with $c \ne 0$ proposed by Rotenberg in 1960.)

Instead of using the "suspect," "almost suspect," etc., criteria for judging the results of chi-square tests, one can employ a less *ad hoc* procedure discussed later in this section.