BIOMETRY

THE PRINCIPLES AND PRACTICE OF STATISTICS IN BIOLOGICAL RESEARCH

THIRD EDITION

Robert R. SOKAL and F. James ROHLF
State University of New York at Stony Brook

W. H. FREEMAN AND COMPANY
New York

CONTENTS

To our parents of blessed memory
Klara and Siegfried Sokal
Harriet and Gilbert Rohlf

Library of Congress Cataloging-in-Publication Data

Sokal, Robert R.
Biometry : the principles and practice of statistics in biological research / Robert R. Sokal and F. James Rohlf.—3d ed.
p. cm.
Includes bibliographical references (p. 850) and index.
ISBN 0-7167-2411-1
1. Biometry. I. Rohlf, F. James, 1936– . II. Title.
QH323.5.S63 1995
574'.01'5195—dc20 94-11120
CIP

Printed in the United States of America

Eighth printing, 2003

2 DATA IN BIOLOGY

In Section 2.1 we explain the statistical meaning of "sample" and "population," terms used throughout this book. Then we come to the types of observations obtained from biological research material, with which we shall perform the computations in the rest of this book (Section 2.2). In obtaining data we shall run into the problem of the degree of accuracy necessary for recording the data. This problem and the procedure for rounding off figures are discussed in Section 2.3, after which we will be ready to consider in Section 2.4 certain kinds of derived data, such as ratios and indices, frequently used in biological science, which present peculiar problems with respect to their accuracy and distribution. Knowing how to arrange data as frequency distributions is important, because such arrangements permit us to get an overall impression of the shape of the variation present in a sample. Frequency distributions, as well as the presentation of numerical data, are discussed in the last section (2.5) of this chapter.

2.1 SAMPLES AND POPULATIONS

We shall now define a number of important terms necessary for an understanding of biological data. The data in a biometric study are generally based on **individual observations,** which are *observations or measurements taken on the smallest sampling unit.* These smallest sampling units frequently, but not necessarily, are also individuals in the ordinary biological sense. If we measure weight in 100 rats, then the weight of each rat is an individual observation; the hundred rat weights together represent the **sample of observations,** defined as *a collection of individual observations selected by a specified procedure.* In this instance, one individual observation is based on one individual in a biological sense—that is, one rat. However, if we had studied weight in a single rat over a period of time, the sample of individual observations would be all the weights recorded on one rat at successive times. In a study of temperature in ant colonies, where each

DATA IN BIOLOGY

tion 2.1 we explain the statistical meaning of "sample" and used throughout this book. Then we come to the types of ed from biological research material, with which we shall tions in the rest of this book (Section 2.2). In obtaining data problem of the degree of accuracy necessary for recording em and the procedure for rounding off figures are discussed which we will be ready to consider in Section 2.4 certain such as ratios and indices, frequently used in biological peculiar problems with respect to their accuracy and ow to arrange data as frequency distributions is impor- angements permit us to get an overall impression of the sent in a sample. Frequency distributions, as well as the data, are discussed in the last section (2.5) of this

ULATIONS

of important terms necessary for an understanding a biometric study are generally based on **individ-** *observations or measurements taken on the small-* st sampling units frequently, but not necessarily, ary biological sense. If we measure weight in 100 at is an individual observation; the hundred rat **mple of observations,** defined as *a collection of by a specified procedure.* In this instance, one one individual in a biological sense—that is, ed weight in a single rat over a period of time, ions would be all the weights recorded on one of temperature in ant colonies, where each

colony is a basic sampling unit, each temperature reading for one colony is an individual observation, and the sample of observations is the temperatures for all the colonies considered. An estimate of the DNA content of a single mammalian sperm cell is an individual observation, and the corresponding sample of observations is the estimates of DNA content of all other sperm cells studied in one individual mammal. A synonym for individual observation is "item."

Up to now we have carefully avoided specifying the particular variable being studied because "individual observation" and "sample of observations" as we just used them define only the structure but not the nature of the data in a study. *The actual property measured by the individual observations* is the **variable,** or *character.* The more common term employed in general statistics is *variable.* In evolutionary and systematic biology however, *character* is frequently used synonymously. More than one variable can be measured on each smallest sampling unit. Thus, in a group of 25 mice we might measure the blood pH and the erythrocyte count. The mouse (a biological individual) would be the smallest sampling unit; blood pH and cell count would be the two variables studied. In this example the pH readings and cell counts are individual observations, and two samples of 25 observations on pH and erythrocyte count would result. Alternatively, we may call this example a **bivariate sample** of 25 observations, each referring to a pH reading paired with an erythrocyte count.

Next we define **population.** The biological definition of this term is well known: It refers to all the individuals of a given species (perhaps of a given life history stage or sex) found in a circumscribed area at a given time. In statistics, population always means the *totality of individual observations about which inferences are to be made, existing anywhere in the world or at least within a definitely specified sampling area limited in space and time.* If you take five humans and study the number of leucocytes in their peripheral blood and you are prepared to draw conclusions about all humans from this sample of five, then the population from which the sample has been drawn represents the leucocyte counts of all humankind—that is, all extant members of the species *Homo sapiens.* If, on the other hand, you restrict yourself to a more narrowly specified sample, such as five male Chinese, aged 20, and you are restricting your conclusions to this particular group, then the population from which you are sampling will be leucocyte numbers of all Chinese males of age 20. The population in this statistical sense is sometimes referred to as the *universe.* A population may refer to variables of a concrete collection of objects or creatures—such as the tail lengths of all the white mice in the world, the leucocyte counts of all the Chinese men in the world of age 20, or the DNA contents of all the hamster sperm cells in existence—or it may refer to the outcomes of experiments—such as all the heartbeat frequencies produced in guinea pigs by injections of adrenalin. In the first three cases the population is finite. Although in practice it would be impossible to collect, count, and examine all white mice, all Chinese men of age 20, or all hamster sperm cells in the world, these populations are finite. Certain smaller populations, such as all the whooping cranes in North America or all the pocket

ch is the sum of the probabilities of having a positive test among those who e cancer and among those who do not have cancer—each weighted by the luencies of the two populations. Substituting these two results into Expres- 1 (5.7) yields

$$P[\mathrm{C}|\mathrm{T}] = \frac{P[\mathrm{T}|\mathrm{C}]P[\mathrm{C}]}{P[\mathrm{T}|\mathrm{C}]P[\mathrm{C}] + P[\mathrm{T}|\mathrm{C}^{\mathrm{C}}]P[\mathrm{C}^{\mathrm{C}}]} \tag{5.8}$$

This expression is known as **Bayes' theorem** and can be generalized to allow an event C having more than just two states (the denominator is summed over events C_i rather than just C and its complement). This famous formula, lished posthumously by the eighteenth-century English clergyman Thomas es, has led to much controversy over the interpretation of the quantity C|T].

Earlier we defined "probability" as the proportion that an event occurs out of rge number of trials. In the current example we have only a single patient, either does or does not have cancer. The patient does not have cancer some portion of the time. Thus the meaning of $P[\mathrm{C}|\mathrm{T}]$ in this case is the degree of 's belief, or the likelihood that this patient has cancer. It is this alternative rpretation of probability and the question of how it should be applied to stics that is controversial. Kotz and Stroup (1983) give a good introduction he idea that probability refers to uncertainty of knowledge rather than of its.

Consider the following example, in which Bayes' theorem was applied to a gnostic test. The figures are based on Watson and Tang (1980). The sensitiv- of the radioimmunoassay for prostatic acid phosphatase (RIA-PAP) as a test prostatic cancer is 0.70. Its specificity is 0.94. The prevalence of prostatic cer in the white male population is 35 per 100,000, or 0.00035. Applying e values to Expression (5.8), we find

$$P[\mathrm{C}|\mathrm{T}] = \frac{P[\mathrm{T}|\mathrm{C}]P[\mathrm{C}]}{P[\mathrm{T}|\mathrm{C}]P[\mathrm{C}] + P[\mathrm{T}|\mathrm{C}^{\mathrm{C}}]P[\mathrm{C}^{\mathrm{C}}]}$$

$$= \frac{0.70 \times 0.00035}{(0.70 \times 0.00035) + [(1 - 0.94)(1 - 0.00035)]} = 0.0041$$

rather surprising result is that the likelihood that a white male who tests tive for the RIA-PAP test actually has prostate cancer is only 0.41%. This ability is known in epidemiology as the **positive predictive value.** Even if test had been much more sensitive, say, 0.95 rather than 0.70, the positive lictive value would have been low—0.55 percent. Only for a perfect test , sensitivity and specificity both = 1) would a positive test imply with cer- y that the patient had prostate cancer.

The paradoxically low positive predictive value is a consequence of its de- lence on the prevalence of the disease. Only if the prevalence of prostatic

cancer were 7895 per 100,000 would there be a 50 : 50 chance that a patient with a positive test result has cancer. This is more than 127 times the highest prevalence ever reported from a population in the United States. Watson and Tang (1980) use these findings (erroneously reported as 1440 per 100,000) and further analyses to make the point that using the RIA-PAP test as a routine screening procedure for prostate cancer is not worthwhile.

Readers interested in extending their knowledge of probability should refer to general texts such as Galambo (1984) or Kotz and Stroup (1983) for a simple introduction.

5.2 THE BINOMIAL DISTRIBUTION

For the discussion to follow, we will simplify our sample space to consist of only two elements, foreign and American students, represented by {F, A}, and ignore whether they are undergraduates or graduates. Let us symbolize the probability space by $\{p, q\}$, where $p = P[\mathrm{F}]$, the probability of being a foreign student, and $q = P[\mathrm{A}]$, the probability of being an American student. As before, we can compute the probability space of samples of two students as follows:

$$\{\mathrm{FF},\ \mathrm{FA},\ \mathrm{AA}\}$$

$$\{p^2,\ 2pq,\ q^2\}$$

If we were to sample three students independently, the probability space of the sample would be

$$\{\mathrm{FFF},\ \mathrm{FFA},\ \mathrm{FAA},\ \mathrm{AAA}\}$$

$$\{p^3,\ 3p^2q,\ 3pq^2,\ q^3\}$$

Samples of three foreign or three American students can be obtained in only one way, and their probabilities are p^3 and q^3, respectively. In samples of three, however, there are three ways of obtaining two students of one kind and one student of the other. As before, if A stands for American and F stands for foreign, then the sampling sequence could be AFF, FAF, and FFA for two foreign students and one American. Thus the probability of this outcome will be $3p^2q$. Similarly, the probability for two Americans and one foreign student is $3pq^2$.

A convenient way to summarize these results is by the binomial expansion, which is applicable to samples of any size from populations in which objects occur independently in only two classes—students who may be foreign or American, individuals who may be dead or alive, male or female, black or white, rough or smooth, and so forth. This summary is accomplished by expanding the binomial term $(p + q)^k$, where k equals sample size, p equals the probability of occurrence of the first class, and q equals the probability of occurrence of the

ɔd class. By definition, $p + q = 1$; hence q is a function of p: $q = 1 - p$. vill expand the expression for samples of k from 1 to 3:

For samples of 1 $(p + q)^1 = p + q$

For samples of 2 $(p + q)^2 = p^2 + 2pq + q^2$

For samples of 3 $(p + q)^3 = p^3 + 3p^2q + 3pq^2 + q^3$

e expressions yield the same outcomes discussed previously. The coeffi- s (the numbers before the powers of p and q) express the number of ways a ɔular outcome is obtained.

general formula that gives both the powers of p and q, as well as the nial coefficients, is

$$\binom{k}{Y} p^Y q^{k-Y} = \frac{k!}{Y!(k-Y)!} p^Y (1-p)^{k-Y} \tag{5.9}$$

is formula k, p, and q retain their earlier meaning, while Y stands for the ɔer or count of "successes," the items that interest us and whose probability ɔcurrence is symbolized by p. In our example, Y designates the number of gn students. The expression $\binom{k}{Y}$ stands for the number of combinations can be formed from k items taken Y at a time. This expression can be ıated as $k!/[Y!(k-Y)!]$, where ! means factorial. In mathematics, k facto- s the product of all the integers from 1 up to and including k. Thus: $5! = 2 \times 3 \times 4 \times 5 = 120$. By convention, $0! = 1$. In working out fractions ıining factorials, note that a factorial always cancels against a higher facto- Thus $5!/3! = (5 \times 4 \times 3!)/3! = 5 \times 4$. For example, the binomial coeffi- for the expected frequency of samples of 5 students containing 2 foreign ɔnts is $\binom{5}{2} = 5!/2!3! = (5 \times 4)/2 = 10$.

ɔw let us turn to a biological example. Suppose we have a population of ts, exactly 40% of which are infected with a given virus X. If we take ıles of $k = 5$ insects each and examine each insect separately for the pres- of virus, what distribution of samples could we expect if the probability of tion of each insect in a sample were independent from that of other insects e sample? In this case $p = 0.4$, the proportion infected, and $q = 0.6$, the ɔrtion not infected. The population is assumed to be so large that the ques- ɔf whether sampling is with or without replacement is irrelevant for practical ɔses. The expected frequencies would be the expansion of the binomial:

$$(p + q)^k = (0.4 + 0.6)^5$$

the aid of Expression (5.9) this expansion is

$$p^5 + 5p^4q + 10p^3q^2 + 10p^2q^3 + 5pq^4 + q^5$$

or

$$(0.4)^5 + 5(0.4)^4(0.6) + 10(0.4)^3(0.6)^2 + 10(0.4)^2(0.6)^3 + 5(0.4)(0.6)^4 + (0.6)^5$$

representing the expected proportions of samples of five infected insects, four infected and one noninfected insects, three infected and two noninfected insects, and so on.

By now you have probably realized that the terms of the binomial expansion yield a type of frequency distribution for these different outcomes. Associated with each outcome, such as "five infected insects," is a probability of occurrence—in this case $(0.4)^5 = 0.01024$. This is a theoretical frequency distribution, or **probability distribution,** of events that can occur in two classes. It describes the expected distribution of outcomes in random samples of five insects, 40% of which are infected. The probability distribution described here is known as the **binomial distribution;** the binomial expansion yields the expected frequencies of the classes of the binomial distribution.

A convenient layout for presentation and computation of a binomial distribution is shown in Table 5.1, based on Expression (5.9). In the first column, which lists the number of infected insects per sample, note that we have revised the order of the terms to indicate a progression from $Y = 0$ successes (infected insects) to $Y = k$ successes. The second column features the binomial coefficient as given by the combinatorial portion of Expression (5.9). Column 3 shows

Table 5.1 EXPECTED FREQUENCIES OF INFECTED INSECTS IN SAMPLES OF 5 INSECTS SAMPLED FROM AN INFINITELY LARGE POPULATION WITH AN ASSUMED INFECTION RATE OF 40%.

(1) Number of infected insects per sample Y	(2) Binomial coefficients $\binom{k}{Y}$	(3) Powers of $p = 0.4$	(4) Powers of $q = 0.6$	(5) Relative expected frequencies $\hat{f}_{rel}$	(6) Absolute expected frequencies $\hat{f}$	(7) Observed frequencies f
0	1	1.00000	0.07776	0.07776	188.4	202
1	5	0.40000	0.12960	0.25920	628.0	643
2	10	0.16000	0.21600	0.34560	837.4	817
3	10	0.06400	0.36000	0.23040	558.3	535
4	5	0.02560	0.60000	0.07680	186.1	197
5	1	0.01024	1.00000	0.01024	24.8	29
		$\sum \hat{f}$ or $\sum f(=n)$		1.00000	2423.0	2423
		Mean		2.00000	2.00004	1.98721
		Standard deviation		1.09545	1.09543	1.11934

reasing powers of p from p^0 to p^5, and column (4) shows decreasing powers of rom q^5 to q^0. The **relative expected frequencies,** which are the probabilities the various outcomes, are shown in column (5). We label such expected quencies $\hat{f}_{rel}$. They are the product of columns (2), (3), and (4), and their sum is ıal to 1.0, since the events in column (1) exhaust the possible outcomes. We from column (5) that only about 1% of the samples are expected to consist of nfected insects, and 25.9% are expected to contain 1 infected and 4 noninted insects. We will now test whether these predictions hold in an actual ıeriment.

ERIMENT 5.1. Simulate the case of the infected insects by using a table of dom numbers such as Statistical Table **FF**. These are randomly chosen one-digit ıbers in which each digit 0 through 9 has an equal probability of appearing. The ıbers are grouped in blocks of 25 for convenience. Such numbers can also be obtained n random number keys on some pocket calculators and by pseudorandom numbererating algorithms in computer programs. Since there is an equal probability for any digit to appear, you can let any four digits (say 0, 1, 2, 3) stand for the infected insects the remaining digits (4, 5, 6, 7, 8, 9) stand for the noninfected insects. The probability any one digit selected from the table represents an infected insect (that is, a 0, 1, 2, or . therefore 40% or 0.4, since these are four of the ten possible digits. Also, successive ts are assumed to be independent of the values of previous digits. Thus the assumps of the binomial distribution should be met in this experiment. Enter the table of om numbers at an arbitrary point (not always at the beginning!) and look at succesgroups of five digits, noting in each group how many of the digits are 0, 1, 2, or 3, e as many groups of five as you can find time to do, but no fewer than 100 groups. sons with computer experience can easily generate the data required by this exercise out using Table **FF**. There are also some programs that specialize in simulating)ling experiments.)

'olumn (7) in Table 5.1 shows the results of such an experiment by a bioy class. A total of 2423 samples of five numbers were obtained from Statis- Table **FF**, and the distribution of the four digits simulating the percentage of ction is shown in this column. The observed frequencies are labeled f. To ulate the expected frequencies for this example, we multiplied the relative cted frequencies, $\hat{f}_{rel}$, of column (5) by $n = 2423$, the number of samples n. These calculations resulted in **absolute expected frequencies,** $\hat{f}$, shown)lumn (6). When we compare the observed frequencies in column (7) with xpected frequencies in column (6), we note general agreement between the columns of figures. The two distributions are illustrated in Figure 5.2. If the rved frequencies did not fit expected frequencies, we might believe that the of fit was due to chance alone. Or we might be led to reject one or more of ollowing hypotheses: (1) that the true proportion of digits 0, 1, 2, and 3 is 0.4 ction of this hypothesis would normally not be reasonable, for we may rely ıe fact that the proportion of digits 0, 1, 2, and 3 in a table of random

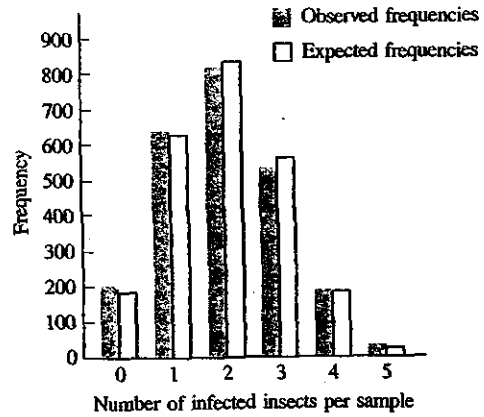

FIGURE 5.2 *Bar diagram of observed and expected frequencies given in Table 5.1.*

numbers *is* 0.4 or very close to it); (2) that sampling was random; and (3) that events are independent.

These statements can be reinterpreted in terms of the original infection model with which we started this discussion. If, instead of a sampling experiment of digits by a biometry class, this had been a real sampling experiment of insects, we would conclude that the insects had indeed been randomly sampled and that we had no evidence to reject the hypothesis that the proportion of infected insects was 40%. If the observed frequencies had not fit the expected frequencies, the lack of fit might be attributed to chance or to the conclusion that the true proportion of infection is not 0.4, or we would have to reject one or both the following assumptions: (1) that sampling was at random, and (2) that the occurrence of infected insects in these samples was independent.

Experiment 5.1 was designed to yield random samples and independent events. How could we simulate a sampling procedure in which the occurrences of the digits 0, 1, 2, and 3 were not independent? We could, for example, instruct the sampler to sample as indicated previously, but every time he found a 3, to search through the succeeding digits until he found another one of the four digits standing for infected individuals and to incorporate this in the sample. Thus, once a 3 was found, the probability would be 1.0 that another one of the indicated digits would be included in the sample. After repeated samples, this procedure would result in higher frequencies of classes of two or more indicated digits and in lower frequencies than expected (on the basis of the binomial distribution) of classes of one event. Many such sampling schemes could be devised. It should be clear that the probability of the second event occurring would be different from and dependent on that of the first.

How would we interpret a large departure of the observed frequencies from pected frequencies in another example? We have not yet learned techniques r testing whether observed frequencies differ from those expected by more an can be attributed to chance alone. This topic will be taken up in Chapter 17. ssume that such a test has been carried out and that it has shown that our served frequencies are significantly different from the expected frequencies. vo main types of departure from expectation are likely: (1) **clumping** and (2) **pulsion,** shown in fictitious examples in Table 5.2. In real examples we would ve no a priori notions about the magnitude of p, the probability of one of the o possible outcomes. In such cases it is customary to obtain p from the ob- ved sample and to calculate the expected frequencies using the sample p. This uld mean that the hypothesis that p is a given value cannot be tested, since by sign the expected frequencies will have the same p value as the observed quencies. Therefore, the hypotheses tested are whether the samples are ran- m and the events independent.

The clumped frequencies in Table 5.2 have an excess of observations at the ls of the frequency distribution and consequently a shortage of observations at : center. Such a distribution is also called **contagious.** Remember that the total mber of items must be the same in both observed and expected frequencies in ler to make them comparable. In the repulsed frequency distribution there are re observations than expected at the center of the distribution and fewer at the s. These discrepancies are most obvious in columns (4) and (6) of Table 5.2,

able 5.2 ARTIFICIAL DISTRIBUTIONS TO ILLUSTRATE CLUMPING AND REPULSION.

ected frequencies from Table 5.1.

(1) Number of fected insects per sample Y	(2) Absolute expected frequencies $\hat{f}$	(3) Clumped (contagious) frequencies f	(4) Deviation from expectation	(5) Repulsed frequencies f	(6) Deviation from expectation
0	188.4	225	+	143	−
1	628.0	703	+	618	−
2	837.4	663	−	943	+
3	558.3	558	0	548	−
4	186.1	227	+	157	−
5	24.8	47	+	14	−
$\sum \hat{f}$ or n	2423.0	2423		2423	
Mean	2.00004	2.00000		2.00000	
ıdard deviation	1.09543	1.20074		1.01435	

where the deviations of observed from expected frequencies are shown as plus or minus signs. (These two types of distributions are also called overdispersed and underdispersed, but there has been some confusion in the literature about the meaning of these terms, so we will not use them here.)

What do these phenomena imply? In the clumped frequencies more samples were entirely infected (or largely infected) and similarly more samples were entirely noninfected (or largely noninfected) than you would expect if probabilities of infection were independent. This result could be due to poor sampling design. If, for example, the investigator, in collecting samples of five insects, always tended to pick out like ones—that is, infected ones or noninfected ones—then such a result would likely appear. If the sampling design is sound, however, the results become more interesting. Clumping would then mean that the samples of five are in some way related—that is, if one insect is infected, others in the same sample are more likely to be infected. This relation could be true if the insects came from adjacent locations in a situation in which neighbors are easily infected. Or the insects could be siblings exposed simultaneously to a source of infection. Or the infection could spread among members of a sample between the time the insects are sampled and the time they are examined.

The opposite phenomenon, repulsion, is more difficult to interpret biologically. There are fewer homogeneous groups and more mixed groups in such a distribution, which implies a compensatory phenomenon: If some of the insects in a sample are infected, the others in the sample are less likely to be. If the infected insects in the sample could transmit immunity to their associates in the sample, such a situation could arise logically, but it is biologically improbable. A more reasonable interpretation of such a finding is that for each sampling unit there are a limited number of pathogens available and that once several insects have become infected, the others go free of infection simply because there is no more infectious agent. This situation is unlikely in microbial infections, but in situations in which a limited number of parasites enter the body of the host, repulsion is more reasonable.

From the expected and observed frequencies in Table 5.1, we may calculate the mean and standard deviation of the number of infected insects per sample. These values are given at the bottom of columns (5), (6), and (7) in Table 5.1. We note that the means and standard deviations in columns (5) and (6) are almost identical and differ only trivially because of rounding errors. Column (7), however, being a sample from a population whose parameters are the same as those of the expected frequency distribution in columns (5) or (6), differs. The mean is slightly smaller and the standard deviation is slightly greater than in the expected frequencies. If we wish to know the mean and standard deviation of expected binomial frequency distributions, we need not go through the computations shown in Table 5.1. The mean and standard deviation of a binomial frequency distribution are, respectively,

$$\mu = kp \qquad \sigma = \sqrt{kpq}$$

ostituting the values $k = 5, p = 0.4$, and $q = 0.6$ from the example above, we ain $\mu = 2.0$ and $\sigma = 1.09545$, which are identical to the values computed m column (5) in Table 5.1. Note that we use the Greek parametric notation e because μ and σ are parameters of an expected frequency distribution, not nple statistics, as are the mean and standard deviation in column (7). The portions p and q are parametric values also and strictly speaking should be tinguished from sample proportions. In fact, in later chapters we resort to $\hat{p}$ $\hat{q}$ for parametric proportions (rather than π, which conventionally is used as ratio of the circumference to the diameter of a circle). Here, however, we fer to keep our notation simple.

It is interesting to look at the standard deviations of the clumped and repulsed quency distributions of Table 5.2. We note that the clumped distribution has a ndard deviation greater than expected, and that of the repulsed one is less than pected. Comparison of sample standard deviations with their expected values a useful measure of dispersion in such instances. If we wish to express our iable as a proportion rather than as a count—that is, to indicate mean inci- ice of infection in the insects as 0.4, rather than as 2 per sample of 5—we can other formulas for the mean and standard deviation in a binomial distribu- n:

$$\mu = p \qquad \sigma = \sqrt{pq/k}$$

We will now use the binomial distribution to solve a biological problem. On basis of our knowledge of the cytology and biology of species A, we expect sex ratio among its offspring to be 1 : 1. The study of a litter in nature reveals t of 17 offspring, 3 were males and 14 were females. What conclusions can draw from this evidence? Assuming that p_{δ} (the probability of being a male spring) = 0.5 and that this probability is independent among the members of sample, the pertinent probability distribution is the binomial for sample size = 17. Expanding the binomial to the power 17 is a nontrivial task, which, as shall see, fortunately need not be done in its entirety.

The setup of this example is shown in Table 5.3. For the purposes of our oblem, we need not proceed beyond the term for 4 males and 13 females. lculating the relative expected frequencies in column (3), we note that the obability of 3 males and 14 females is 0.005,188,40, a very small value. If we d to this value all "worse" outcomes—that is, all outcomes that are even ore unlikely than 14 females and 3 males on the assumption of a 1 : 1 hypothe- —we obtain a probability of 0.006,363,42, still a very small value. In statis- s one often needs to calculate the probability of observing a deviation as large larger than a given value.

On the basis of these findings one or more of the following assumptions is likely: (1) that the true sex ratio in species A is 1 : 1; (2) that we have sampled random in the sense of obtaining an unbiased sample; or (3) that the sexes of offspring are independent of one another. Lack of independence of events ay mean that although the average sex ratio is 1 : 1, the individual sibships, or

Table 5.3 SOME EXPECTED FREQUENCIES OF MALES AND FEMALES FOR SAMPLES OF 17 OFFSPRING ON THE ASSUMPTION THAT THE SEX RATIO IS 1 : 1 [$p_{\delta} = 0.5$, $q_{\varphi} = 0.5$; $(p_{\delta} + q_{\varphi})^k = (0.5 + 0.5)^{17}$].

(1) Y ♂♂	(2) $k - Y$ ♀♀	(3) Relative expected frequencies $\hat{f}_{rel}$	
0	17	0.000,007,63	0.006,363,42
1	16	0.000,129,71	
2	15	0.001,037,68	
3	14	0.005,188,40	
4	13	0.018,157,91	

litters, are largely unisexual—that is, the offspring from a given mating tend to be all (or largely) males or all (or largely) females. To confirm this hypothesis we would need to have more samples and then examine the distribution of samples for clumping, which would indicate a tendency for unisexual sibships.

We must be very precise about the questions we ask of our data. There are really two questions we can ask about the sex ratio: (1) Are the sexes unequal in frequency so that females appear more often than males? and (2) Are the sexes unequal in frequency? We may be concerned with only the first of these questions, since we know from past experience that in this particular group of organisms the males are never more frequent than females. In such a case the reasoning followed above is appropriate. However, if we know very little about this group of organisms and if our question is simply whether the sexes among the offspring are unequal in frequency, then we have to consider both tails of the binomial frequency distribution; departures from the 1 : 1 ratio could occur in either direction. We should then consider not only the probabilities of samples with 3 males and 14 females (and all worse cases) but also the probability of samples of 14 males and 3 females (and all worse cases in that direction). Since this probability distribution is symmetrical (because $p_{\delta} = q_{\varphi} = 0.5$), we simply double the cumulative probability of 0.006,363,42 obtained previously, which results in 0.012,726,84. This new value is still very small, making it quite unlikely that the true sex ratio is 1 : 1.

This is your first experience with one of the most important applications of statistics—hypothesis testing. A formal introduction to this field will be deferred until Section 7.8. We simply point out here that the two approaches just described are known as **one-tailed tests** and **two-tailed tests,** respectively. Students sometimes have difficulty knowing which of the two tests to apply. In

uture examples, we will try to explain why a one-tailed or a two-tailed test is eing used.

We have said that a tendency for unisexual sibships would result in a clumped istribution of observed frequencies. An actual case of this phenomenon in na-ıre is a classic in the literature, the sex ratio data obtained by Geissler (1889) :om hospital records in Saxony. Table 5.4 shows the sex ratios of 6115 sibships f 12 children, each from the more extensive study by Geissler. All columns of ıe table should by now be familiar. To keep you on your toes, and to conform to ıe layout of the original publication, the meaning of p and q have been reversed :om that in the earlier sex ratio example. Now $p_{♀}$ is the proportion of females nd $q_{♂}$ that of males. In a binomial, which of the two outcomes is p and which is is simply a matter of convenience.

The expected frequencies in this example were not calculated on the basis of 1 : 1 hypothesis, since it is known that in human populations the sex ratio at irth is not 1 : 1. Because the sex ratio varies in different human populations, ıe best estimate of it for the population in Saxony was obtained simply by sing the mean proportion of males in these data. This value can be obtained by

Table 5.4 SEX RATIOS IN 6115 SIBSHIPS OF 12 IN SAXONY.

(1) Y ♀♀	(2) $k-Y$ ♂♂	(3) Relative expected frequencies $\hat{f}_{rel}$	(4) Absolute expected frequencies $\hat{f}$	(5) Observed frequencies f	(6) Deviation from expectation $f-\hat{f}$
0	12	0.000384	2.3	7	+
1	11	0.004264	26.1	45	+
2	10	0.021725	132.8	181	+
3	9	0.067041	410.0	478	+
4	8	0.139703	854.3	829	−
5	7	0.206973	1265.6	1112	−
6	6	0.223590	1367.3	1343	−
7	5	0.177459	1085.2	1033	−
8	4	0.102708	628.1	670	+
9	3	0.042280	258.5	286	+
10	2	0.011743	71.8	104	+
11	1	0.001975	12.1	24	+
12	0	0.000153	0.9	3	+
Total		0.999998	6115.0	6115	

$\bar{Y} = 5.76942 \qquad s^2 = 3.48985$

URCE: Data from Geissler (1889).

calculating the average number of females per sibship ($\bar{Y} = 5.76942$) for the 6115 sibships and converting this into a proportion. This value is 0.480,785. Consequently, the proportion of males = 0.519,215. In the deviations of the observed frequencies from the absolute expected frequencies shown in column (6) of Table 5.4, we notice considerable clumping. There are many more instances of families with all female or all male children (or nearly so) than independent probabilities would indicate. The genetic basis for this is not clear, but it is evident that there are some families that "run to girls" and similarly others that "run to boys." Other evidence of clumping is the fact that s^2 is much larger than we would expect on the basis of the binomial distribution [$\sigma^2 = kpq = 12(0.480785)0.519215 = 2.99557$].

There is a distinct contrast between the data in Table 5.1 and those in Table 5.4. In the insect infection data of Table 5.1, we had a hypothetical proportion of infection based on outside knowledge. In the sex ratio data of Table 5.4 we had no such knowledge; we used an *empirical value of p obtained from the data,* rather than a *hypothetical value external to the data.* The importance of this distinction will become apparent later. In the sex ratio data of Table 5.3, as in much work in Mendelian genetics, a hypothetical value of p is used.

An alternative, efficient method for calculating expected binomial frequencies is given in Box 5.1. The instructions are self-explanatory. Program BIOM-pc includes an option for computing expected binomial frequencies.

5.3 THE POISSON DISTRIBUTION

In the typical application of the binomial, we had relatively small samples (2 students, 5 insects, 17 offspring, 12 siblings), in which two alternative states occurred at varying frequencies (American and foreign, infected and noninfected, male and female). Quite frequently, however, we study cases in which sample size k is very large, and one of the events (represented by probability q) is much more frequent than the other (represented by p). We have seen that the expansion of the binomial $(p+q)^k$ is quite tiresome when k is large. Suppose you had to expand the expression $(0.001 + 0.999)^{1000}$. In such cases we are generally interested in one tail of the distribution only. This is the tail represented by the terms

$$p^0q^k, \binom{k}{1}p^1q^{k-1}, \binom{k}{2}p^2q^{k-2}, \binom{k}{3}p^3q^{k-3}, \ldots$$

The first term represents no rare events and k frequent events in a sample of k events, the second term represents 1 rare event and $k-1$ frequent events, the third term 2 rare events and $k-2$ frequent events, and so forth. The expressions of the form $\binom{k}{i}$ are the binomial coefficients, discussed in the previous section.

concern yourself with these, but it might be well for you to be familiar by name at least with those we will discuss here briefly.

The **hypergeometric distribution** is the distribution equivalent to the binomial case but sampled from a *finite population without replacement.* In Section 5.1 we estimated the probability of finding a second foreign student in Matchless University. We pointed out that if there are 10,000 students at Matchless University, 400 of whom are foreign, then the probability of sampling one foreigner is indeed 0.04. Once a foreign student has been sampled, however, the probability of sampling another foreign student is no longer 0.04 (even when independent) but is 399/9999, or 0.0399. This probability would be 0.04 only *with replacement*—that is, if we returned the first foreign student sampled to the university population before we sampled again.

The binomial distribution is entirely correct only in cases of sampling with replacement or with infinite population size (which amounts to the same thing). For practical purposes, when small samples are taken from large populations, as in the case of Matchless University, these populations can be considered infinite. But if you have a population of 100 animals, 4% of which carry a mutation, a sample of one mutant reduces the population to 3 out of 99, or from 4% to 3.03%. Thus repeated samples of 5 from this population would follow not the binomial distribution but a different distribution, the hypergeometric distribution. The individual terms of the hypergeometric distribution are given by the expression

$$\frac{\binom{pN}{r}\binom{qN}{k-r}}{\binom{N}{k}}$$

which gives the probability of sampling r items of the type represented by probability p out of a sample of k items from a population of size N. The mean and variance of a hypergeometric distribution are kp and $kpq(N-k)/(N-1)$.

Note that the mean is the same as that of the binomial distribution, and the variance is that of the binomial multiplied by $(N-k)/(N-1)$. When N is very large as compared with k, this term is approximately 1, expected since the hypergeometric distribution approximates the binomial. The expected frequencies for any sizable distribution are tedious to compute. We suggest using a digital computer with facilities for double-precision arithmetic to evaluate expected hypergeometric frequencies for a sizable distribution.

In biology, sampling of small samples from a finite distribution occurs in certain problems of evolutionary genetics. Another application is in mark–recapture studies, in which a certain proportion of a population is caught, marked, released, and subsequently recaptured, leading to estimates of the number of the entire population.

A number of probability distributions have been employed as underlying mathematical models for cases of contagious distribution. The difficulty of

fitting these varies with the distribution. Greig-Smith (1964) gives a nontechnical account of the application of these distributions to ecology and provides references that may lead the interested reader deeper into the subject. More quantitative accounts are given by Pielou (1977) and Krebs (1989). Many ecological examples of the application of contagious distributions are given in Williams (1964).

We have used the binomial and Poisson distributions to test whether given data are random or whether they show marked departure from random expectations—either clumping or repulsion. In some cases it may seem unreasonable to assume random occurrences. In social organisms, for example, aggregation is a given. Statisticians and biologists have developed models for such cases which lead to the so-called contagious distributions. Although we cannot discuss these in detail in this volume, we mention two here briefly.

The **negative binomial distribution** has been used probably more frequently than any other contagious distribution. The theoretical conditions that would give rise to a negative binomial distribution are discussed by Bliss and Calhoun (1954), who also give methods of calculating expected frequencies. The account in Bliss and Fisher (1953) is somewhat more rigorous. Krebs (1989) describes methods for estimation and significance testing and provides computer software for the calculations.

The **logarithmic distribution** (or logarithmic series) has been used extensively in studying the distribution of taxonomic units in faunal samples. This distribution has been employed frequently by C. B. Williams (see Williams, 1964). Johnson and Kotz (1969) give general information on many types of discrete distributions.

EXERCISES 5

5.1 In humans the sex ratio of newborn infants is about 100♀♀ : 105♂♂. If we were to take 10,000 random samples of 6 newborn infants each from the total population of such infants for one year, what would be the expected frequency of groups of 6 males, 5 males, 4 males, and so on? *Answer:* For 4 males $\hat{f} = 2456.5$.

5.2 Show algebraically why the computational method of Box 5.1 works.

5.3 The two columns below give fertility of eggs of the CP strain of *Drosophila melanogaster* raised in 100 vials of 10 eggs each (data from R. R. Sokal). Find the expected frequencies, assuming that the mortality for each egg in a vial is independent. Use the observed mean. Calculate the expected variance and compare it with the observed variance. Interpret the results, knowing that the eggs of each vial are siblings and that the different vials contain descendants from different parent pairs. *Answer:* $\sigma^2 = 2.417$, $s^2 = 6.628$.

intended anova, must be employed. These are the nonparametric or distribution-free techniques, which are sometimes used by preference even when the parametric method (anova in this case) can be legitimately employed. Ease of computation and a preference for the generally simple assumptions of the nonparametric analyses cause many research workers to turn to them. When the assumptions of the anova are met, however, these methods are less powerful than analysis of variance. Section 13.11 examines several nonparametric methods in lieu of single-classification anova and Section 13.12 features nonparametric methods in lieu of two-way anova.

13.1 A FUNDAMENTAL ASSUMPTION

All anovas require that sampling of individuals be random. Thus, in a study of the effects of three doses of a drug (plus a control) on five rats each, the five rats allocated to each treatment must be selected at random. If the five rats employed as controls are either the youngest or the smallest or the heaviest rats, while those allocated to some other treatment are selected in some other way, the results are not apt to yield an unbiased estimate of the true treatment effects. Nonrandomness of sample selection may well be reflected in lack of independence of the items (see Section 13.2), in heterogeneity of variances (Section 13.3), or in nonnormal distribution (Section 13.4). Adequate safeguards to ensure random sampling during the design of an experiment or when sampling from natural populations are essential.

13.2 INDEPENDENCE

An assumption stated in each explicit expression for the expected value of a variate [for example, Expression (8.2) was $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$] is that the error term ϵ_{ij} is a random normal variable. In addition, for completeness we should add that it is assumed that the ϵ's are independently and identically (see Section 13.3) distributed.

Thus, if the variates within any one group are arranged in a logical order independent of their magnitude (such as the order in which the measurements were obtained), we expect the ϵ_{ij}'s to succeed each other in a random sequence. Consequently, we assume a long sequence of large positive values followed by an equally long sequence of negative values to be quite unlikely. We would also not expect positive and negative values to alternate with regularity.

How could departures from independence arise? An obvious example is an experiment in which the experimental units are plots of ground laid out in a field. In such a case adjacent plots of ground often give similar yields. It would thus be important not to group all the plots containing the same treatment into an adjacent series of plots, but rather to randomize the allocation of treatments among

the experimental plots. The physical process of randomly allocating the treatments to the experimental plots ensures that the ϵ's will be independent.

Lack of independence of the ϵ's can result from correlation in time rather than in space. In an experiment we might measure the effect of a treatment by recording weights of ten individuals. The balance we use may suffer from a maladjustment that results in giving successive underestimates, compensated for by several overestimates. Conversely, compensation by the operator of the balance may result in regularly alternating over- and underestimates of the true weight. Here again randomization may overcome the problem of nonindependence of errors. For example, we may determine the sequence in which individuals of the various groups are weighed according to some random procedure.

Both of these examples—the spatial and the temporal—are instances of positive **autocorrelation,** the self-similarity of variates adjacent in space or time. Regular alternation of positive and negative errors is a manifestation of negative autocorrelation.

Independence of errors in a sequence of continuous variates may be tested as first proposed by the well-known mathematician John von Neumann (von Neumann et al., 1941), with critical values tabulated by Young (1941). The test is based on successive differences between normal variates, $d_i = Y_{i+1} - Y_i$, which are squared. In Section 15.3 you will learn why the expected sum of such squared differences is twice the sum of squares of variable Y if the variates are independent. Thus in the case of independent errors the ratio $\eta = \Sigma d^2/\Sigma y^2$ should approximate 2. If there are sequences of similar variates, their differences will be less than what they would have been if the variates were randomly ordered, and the ratio η will be less than 2. Conversely, if there is a nonrandom alternation of the magnitudes of the variates, the variance of the differences will be greater than expected and η will be greater than 2. In Statistical Table **HH** we expanded a shorter table of critical values of $|1 - \eta/2|$ by Young (1941) up to a sample size of $n = 50$. When $n > 50$ we can use the normal approximation

$$t_s = \frac{|1 - \eta/2|}{\sqrt{(n - 2)/(n^2 - 1)}}$$

The computations are summarized in Box 13.1, where we examine the sequence of 25 numbers representing the aphid stem mother femur lengths from Box 2.1. We compute first differences to match all but the last observation; then we square and sum these d's. The result is an estimate of $\Sigma d^2 = 9.3700$. When we divide this value by the sum of squares of the femur lengths ($\Sigma y^2 = 0.1337$), we obtain $\eta = 2.9194$. Since, $\eta > 2$, a nonrandom alternation of variates is indicated. Computing $|1 - \eta/2|$, we obtain 0.459683, which in Table **HH** yields a two-tailed $P < 0.02$ for $n = 25$. Had this example been based on more than 50 observations, we could have tested it by using the normal approximation. If we do so in any case, we obtain $t_s = 2.394$, which yields $0.01 < P < 0.02$. We conclude that the observations occur in a sequence that appears to be nonrandom,

Box 13.1 TEST FOR SERIAL INDEPENDENCE OF A CONTINUOUS VARIABLE.

Twenty-five aphid stem mother femur lengths. Data from Box 2.1.

Y_i	$d_i^2 = (Y_{i+1} - Y_i)^2$
3.8	0.04
3.6	0.49
4.3	0.64
3.5	0.64
4.3	1.00
3.3	1.00
4.3	0.16
3.9	0.16
4.3	0.25
3.8	0.01
3.9	0.25
4.4	0.36
3.8	0.81
4.7	1.21
3.6	0.25
4.1	0.09
4.4	0.01
4.5	0.81
3.6	0.04
3.8	0.36
4.4	0.09
4.1	0.25
3.6	0.36
4.2	0.09
3.9	
$\Sigma y^2 = 3.2096$	$\Sigma d^2 = 9.3700$

Computation

1. Make a column of the observations. Construct a second column of first differences between the observations and square them as shown.
2. Compute the sum of squares of the observations and the sum of the squared differences (shown at the bottom of the columns).
3. Compute $\eta = \Sigma d^2/\Sigma y^2 = 9.3700/3.2096 = 2.9194$.

BOX 13.1 CONTINUED

4. Evaluate $|1 - \eta/2| = 0.459683$. If $n \leq 50$, consult Statistical Table **HH** for significance. In our case the two-tailed probability is $0.01 < P < 0.02$. For illustrative purposes we also evaluate

$$\frac{|1 - \eta/2|}{\sqrt{(n-2)/(n^2-1)}}$$

and compare with $t_{\alpha[\infty]}$. This is the approximation we would use with $n > 50$. Since n for this example is 25, the approximation should be close. We find that $t_s = 0.459683/\sqrt{23/(25^2-1)} = 2.394348$, which in Statistical Table **B** is significant at $0.01 < P < 0.02$. The observations are not serially independent. The fact that η is greater than 2 suggests a nonrandom alternation of the observations. Values of $\eta < 2$ indicate serial correlation (= autocorrelation) between adjacent variates.

We used a two-tailed significance test here, since we had no a-priori notion of the nature of the departure from serial independence. In some instances our alternative hypothesis would be one-tailed, in which case the probabilities at the head of the columns of Table **HH**, or of Table **B** when using the normal approximation, should be halved.

so we question the assumption of independence in these data. Recent work has indicated that the femur lengths may have come from a dimorphic sample. Possibly the technician mounting the aphids on slides for measurement alternated between the two types of galls in a conscious (but misguided) attempt to strike a balance. A ratio of η significantly less than 2 would have indicated some serial correlation (= autocorrelation)—succeeding variates would be more similar to each other because of technician or instrument bias.

For a nonparametric serial correlation test of continuous variates, or when the variates are nominal, employ a runs test (see Section 18.2).

There is no simple adjustment or transformation to overcome the lack of independence or errors. The basic design of the experiment or the way in which it was performed must be changed. We have seen how a randomized-blocks design often overcomes lack of independence of error by randomizing the effects of differences in soils or cages. Similarly, in the experiment with the biased balance we could obtain independence of errors by redesigning the experiment, using different times of weighing as blocks. Of course, if a source of error is suspected or known, attempts can be made to remove it; if we know, for example, that the balance is biased, we may have it fixed. If the ϵ's are not independent, the validity of the usual F-test of significance can be seriously impaired.

13.3 HOMOGENEITY OF VARIANCES

In Section 9.4 and Box 9.6, in which we described the t-test for the difference between two means, we said that the statistical test was valid only if we could