C. Tolmie

REFIGURING ANTHROPOLOGY

First Principles Of Probability & Statistics

David Hurst Thomas
American Museum of Natural History

Waveland Press, Inc.
Prospect Heights, Illinois

For information about this book, write or call:

Waveland Press, Inc.
P.O. Box 400
Prospect Heights, Illinois 60070
(312) 634-0081

For permission to use copyrighted material, the author is indebted to the following:

FIG. 2.1. By permission of the Trustees of the British Museum (Natural History).
TABLE 2.3. (p. 24) *From Physical Anthropology: An Introduction* by A. J. Kelso. Reprinted by permission of the publisher, J. B. Lippincott Company. Copyright © 1974. (p. 25) Reproduced by permission of the Society for American Archaeology from *Memoirs of the Society for American Archaeology*, Vol. 11, 1956.
FIG. 3.1. From Hulse, Frederick S. *The Human Species: An Introduction to Physical Anthropology.* Copyright © 1963 by Random House, Inc.
FIG. 3.2. From Dozier, Edward P., *The Pueblo Indians of North America.* Copyright © 1970 by Holt, Rinehart and Winston, Inc. Reproduced by permission of Holt, Rinehart and Winston. [This book reissued 1983 by Waveland Press, Inc.]
FIG. 3.4. Reproduced by permission of the Society for American Archaeology from *American Antiquity,* Vol. 35 (4), 1970.
FIG. 3.5. Reproduced by permission of the American Anthropological Association from the *American Anthropologist,* Vol. 73 (3), 1971.
FIG. 13.14. From *Biometry* by Robert R. Sokal and F. James Rohlf. W. H. Freeman and Company. Copyright © 1969.

Second Printing
The 1976 version of this book was entitled *Figuring Anthropology.*

ISBN 0-88133-223-2

Printed in the United States of America.

11 Nonparametric Statistics: Nominal Scales

● *THE LAW OF NATURAL PERVERSITY: You cannot successfully determine beforehand which side of bread to butter.*—L. Peter

11.1 INTRODUCTION TO NONPARAMETRIC STATISTICS

The theoretical models underlying the *t*- and *z*-statistics are grounded in a few explicit and rather important assumptions. By way of review, the following is assumed by the simple test for a difference between two means (see Section 10.9):

1. The variable is defined on an interval or ratio scale.
2. The samples exhibit independent errors.
3. The sample variates have been randomly selected from a normally distributed population.

These conditions are rarely tested outright. They are usually just *assumed* to hold for the case at hand. As long as these requirements are reasonably satisfied, the parametric model remains a powerful tool, enabling us to test hypotheses and to establish confidence limits.

But must we set aside our elaborate parametric machinery when these conditions cannot be assumed? Section 10.9 discussed one aspect of this problem, noting that some degree of violation is permissible, *as long as the sample sizes are sufficiently large and the hypotheses are nondirectional.* That is, parametric methods are valid as long as the assumptions are at least *approximately* true. The *t*-test, for example, requires only that (1) the population is *approximately* normal, (2) the variables exhibit *largely* independent errors, and (3) the scale is *close enough* to an interval scale. Normality need not be assumed for the *t*-test as long as the sample size is sufficiently large that the

Central Limit Theorem comes into play. Fudge factors such as this allow analysts to proceed under the parametric model, even though the specifics are something less than ideal.

But real data often place unacceptable constraints upon parametric assumptions, constraints so severe that the model simply does not apply, regardless of how inclined one might be to fudge the assumptions. It is precisely in samples of smaller size that the normal distribution is most likely to be violated, and about which one is forced to make an assumption of normality. When the sample is small, the Central Limit Theorem is of no assistance. Once the basic conditions underlying the parametric model prove untenable, the statistical inferences based upon these false assumptions become likewise suspect. When one's assumptions do not hold, the computed levels of probability no longer bear a credible relationship to true probability values. Although a *t*-test can physically be computed on a nonnormal population or upon ordinal variables, the resulting levels of significance are worse than incorrect. They are downright misleading and confounding. The parametric model has a built-in gray zone which permits a certain flexibility regarding assumptions. But there is a point beyond which assumptions should not be stretched, a point at which parametric methods must be scrapped in favor of a more realistic model.

This chapter introduces a sorely needed alternative to normal theory statistics, since both the *t*-test and the standardized normal deviate assume (1) interval or better measurement and (2) a normal distribution. The *nonparametric* family of statistical methods assumes neither condition. Nonparametric statistics comprise a large battery of techniques derived to free us from unrealistic and restrictive assumptions. There was surprisingly little interest in nonparametric methods until the mid-1940s when Frank Wilcoxon proposed a test distinguished by its simplicity. Wilcoxon's test assumed neither interval measurement nor normal distribution of population variates, yet produced excellent results when compared to the common *t*-test (Wilcoxon's test is presented in Chapter 12). Over the past three decades, literally dozens of nonparametric devices have been derived to cope with the social science problems. Unfortunately, most nonparametric methods lack the efficiency of Wilcoxon's test. In fact, some tests extract a dear price indeed in terms of information lost, but at least they offer a viable alternative to the parametric assumptions.

A statistic is *nonparametric* if any *one* of the following conditions applies (after Conover 1971: 94):

1. The statistic can be used on *nominal* scale data; *or*
2. The statistic can be used on *ordinal* scale data; *or*
3. The statistic can be used on a random variable of unspecified distribution.

The first two conditions allow the valid analyses of nominal and ordinal variables. This is especially important for anthropologists, who are often forced to deal with rather crude scales of observation. The third condition, that data can arise from a distribution of unspecified shape, has led some statisticians to call these tests *distribution-free*.

Nonparametric statistics have several advantages beyond mere freedom from unrealistic assumption. For one thing, nonparametrics usually require fewer computations than their parametric counterparts. Some nonparametric tests require only one operation to count plus and minus signs. Thus, the theory underlying nonparametric tests is usually easier for the beginner to follow. Do not be misled by this simplicity. Parametric methods can produce elegant results in the hands of the skilled statisticians, but to the uninitiated these more advanced methods can prove disastrous. The normal theory of statistics has been compared to an expensive camera, equipped with dozens of complex options. Trained photographers use such costly equipment to produce results worthy of the lofty price tag. But to the beginner, just learning the fundamentals of photography, a new Honeywell Pentax ESII with a 50 mm f/1.4 Super-Multi-Coated Takumar lens, self-timer, FP and X sync, battery checker, PC terminal, hot shoe, and shutter-release lock produces more confusion than well-exposed negatives. There are times when a small Brownie box camera is preferable to a more expensive model costing 20 times the price. Nonparametric statistics have much in common with the modest, yet dependable, Brownie camera. Both are cheap and easy to understand, difficult to abuse, and rather easy to explain to one's friends. Small wonder that the term "quick-and-dirty" is lovingly bestowed upon the nonparametric statistics.

Nonparametric analysis can also facilitate a more efficient collection of data. If one strongly suspects that a given population is asymmetrical or otherwise nonnormally distributed, then ordinal or even presence–absence methods of recording data might be just as useful as measurements accurate to 0.01 mm. Normal theory should not be applied to extremely nonnormal populations, regardless of how precise are one's measurements. The nonparametric methods also allow one to use smaller samples, sometimes saving additional costly fieldwork. And the resulting probabilities from nonparametric computations are often *exact*, avoiding the arbitrary cutoff points (critical regions) necessary with the z- and t-statistics.

But all these obvious virtues of the nonparametric approach must not detract from its role as a second-best substitute for normal theory. When information exists on the population distribution, and when level of measurement is satisfactory, the normal theory should be used forthwith. To apply nonparametric methods to such situations is an ill-advised waste of information. Furthermore, the comforting phrases "nonparametric" and "distribution-free" must not be misread to imply "assumption-free." Nonparametric methods make a couple of rather critical assumptions which cannot be ignored.

Although you might not have realized it, a nonparametric statistical test has already been introduced. The *binomial test* (Chapter 6) assumed neither a normal distribution nor an interval level of measurement. Hence, the binomial test qualifies as nonparametric on two counts. Binomials such as heads–tails, male–female, or blood type are only nominal level variates, and a moment's reflection reveals that a Bernoulli variable could not be distributed normally because only two possibilities exist for each variable.[1]

Several additional nominal-level nonparametric tests are presented in this chapter, including the ubiquitous chi-square test. Chapter 12 considers further

[1] Be careful here not to confuse the binomial statistic with the variables themselves. The binomial *statistic* becomes distributed in normal fashion as the sample size increases (in fact, this is a characteristic common to many nonparametric statistics), but this is a very different matter indeed from assuming that the variates themselves distribute normally.

nonparametric methods which are suitable for ordinal level variates. More advanced nonparametric methods of correlation are presented in conjunction with their parametric counterparts.

> ● *General Grant only knew two songs—one was* Yankee Doodle *and the other wasn't.*—A. Gingrich

11.2 THE CHI-SQUARE TEST

Chi-square gets my vote as anthropology's most used (and abused) statistic. The technique is flexible, and the computations are elementary and easily carried out without computational machinery. As long as certain limitations and assumptions are satisfied, the chi-square techniques can play a pivotal role in quantitative anthropology.

Recall how useful the binomial distribution was when a given trial had but two possible outcomes—success or failure. Several examples from Mendelian genetics were discussed earlier. One of Mendel's experiments considered round and wrinkled peas (see Example 11.1), which were expected to occur in the ratio of 3:1. Outcomes of this sort were characterized as simply R (success) or W (failure). Mendel's breeding experiments involved a simple null hypothesis: H_o: $p = 0.75$, where p is the probability of a round seed on a given trial. The associated probabilities were computed and compared with the theoretical binomial probability.

Viewed another way, binomial experiments compare empirically derived *observed (O)* values with theoretically *expected (E)* figures. The normal approximation to the binomial can also be used to test H_o, provided n is sufficiently large (n being the total number of seeds observed).

But suppose there are more than two possible outcomes. Many genetic situations involve several significant phenotypes, too many outcomes to be succinctly characterized as success or failure. There are, for instance, *four* equally likely blood types for the offspring of a heterozygous A and a heterozygous B.

	Offspring	
Parents	Genotype	Phenotype
ao	*ab*	*AB*
	ao	*A*
bo	*bo*	*B*
	oo	*O*

Mendelian theory tells us that, in the long run, unions of this sort should produce offspring with blood types AB:A:B:O in approximately the ratios of 1:1:1:1. The *expected frequencies* for a sample of $n = 100$ such offspring would be ($E = np$).

Type AB: $E_1 = 100(0.25) = 25$
Type A: $E_2 = 100(0.25) = 25$
Type B: $E_3 = 100(0.25) = 25$
Type O: $E_4 = 100(0.25) = 25$

These expected figures can then be tested upon an actual sample of 100 such offspring. Suppose the empirical data consist of the following observed (O_i) values:

Type AB: $O_1 = 32$
Type A: $O_2 = 13$
Type B: $O_3 = 24$
Type O: $O_4 = 31$

The observed values are not equal to exactly 25 for each blood type, but random sampling theory predicts that some degree of deviation is likely. We must decide whether these observations conform to the expected Mendelian frequencies or whether the deviation is too great for the theory to hold.

Had this situation been expressed in terms of success and failure—such as the probability of obtaining Type AB blood as opposed to all other types—then p and q could have been defined as before and the binomial theorem used to compare the expected with the observed values. But introducing more than two possible alternatives (E_k with $k > 2$) vitiates the binomial theorem as we know it.

Fortunately, the χ^2 (to be read "chi-square") test was designed for just such situations:

$$\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{11.1}$$

where O_i are the experimentally observed values and the E_i are the theoretically expected frequencies for the kth class. There is no limit to the magnitude of k in the χ^2 distribution,[2] as there was in the case of the binomial (where $k = 2$).

The chi-square statistic sums the deviations for each class in the frequency distribution. The $(O_i - E_i)$ differences are squared to produce a nonzero sum. The squared deviations are then divided by the expected number of cases in each measurement class. This *standardizes* the chi-square statistic, just as the variates in a normal distribution were standardized into z-scores. Dividing by E_i weights the contribution of each class so that the biggest proportion of the chi-square sum does not always come from the most numerous class.

The value of the χ^2 statistic is best computed from the following conventional tabular format (χ^2 in this case is 9.20).

[2] Some introductory textbooks label the chi-square statistic X^2 rather than χ^2 and, in a strict sense, this procedure is more accurate. The values listed in chi-square tables are really *statistical estimations* of true chi-square parameters. The computed values of the chi-square estimator can vary somewhat because it is sometimes necessary to "correct for continuity" (Section 11.4). While these considerations are germane to a truly exhaustive consideration of this technique, such rigor is beyond the current scope. The symbol χ^2 is used here to indicate both the estimates obtained from computation and the tabled values of the chi-square statistic. By so doing, we can avoid any confusion between χ^2 and the symbol for the common variable X.

Blood Typo	O_i	E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
AB	32	25	7	49	1.96
A	13	25	−12	144	5.76
B	24	25	− 1	1	0.04
O	31	25	6	36	1.44
	100	100	0		$\chi^2 = 9.20$

But computing a chi-square statistic is only half the battle. So far, no decision can be made about the *probability* of any observed value of chi-square. That is, a sampling distribution for the chi-square statistic is necessary so that we can judge the acceptability of a null hypothesis. Just as with the normal distribution, the probability of obtaining *exactly* the expected outcome is zero for a continuous random variable. A certain amount of variability is expected in the chi-square statistic, just as variability was expected in the ABO blood type experiment itself.

But how much variability should we expect? To answer this question, statisticians have repeated randomized experiments literally hundreds of times and then constructed histograms of the chi-square sampling distribution. Two variables are involved in the sample chi-square experiments: the number of experimental cases (n) and the number of observed-expected comparisons (k). As long as n is kept above critical minimum values, the frequency distribution of χ^2 stablilizes within each level of k. But instead of dealing directly with k, we must follow a procedure similar to that of the t-test (Chapter 10), and instead consider the number of *degrees of freedom*, where df $= (k - 1)$. Degrees of freedom in this case refers to the number of classes within a chi-square table, which may be filled arbitrarily without altering the expectations.

Note that degrees of freedom for the χ^2 distribution is determined by k, the number of independent observed-expected comparisons, rather than by sample size (n). For the ABO blood-type experiment,

$$n = 100, \qquad k = 4, \qquad \text{df} = (4 - 1) = 3$$

The chi-square distribution for 3 degrees of freedom is known to follow the distribution given in Fig. 11.1.[3] The x-axis represents the range of possible χ^2 values; chi-square cannot drop below zero and the right-hand tail asymptotes toward positive infinity. The ordinate, scaled in probabilities, ranges between zero and unity. Although chi-square distributions are generally quite asymmetrical, there exists a close parallel between normal and chi-square distributions. Both curves represent probabilities. The higher the curve, the more probable is the interval represented. As χ^2 becomes larger and larger, the probability of observing this or larger values diminishes. Figure 11:1 indicates that 50 percent of all observed χ^2 (with df $= 3$) are expected to exceed 2.4. Only 5 percent of the χ^2 variates should exceed about 7.8 and only 1 percent of the χ^2 values should be greater than about 9.21. But because there is a different graph for each

[3] The actual derivation of this curve, and its formula, are beyond the scope of the present text (see Hays 1973: 432–436).

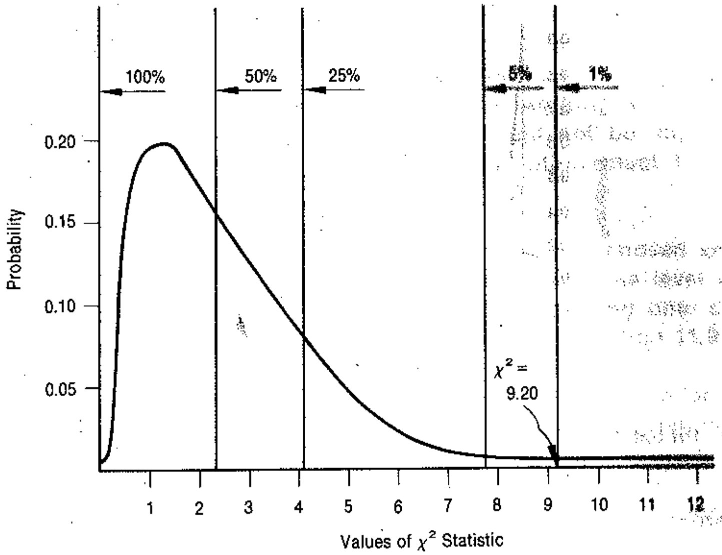

Fig. 11.1 Probability distribution function of χ^2 values with 3 degrees of freedom.

change in the number of degrees of freedom, these figures have been recorded on Table A.5 (Appendix).

Figure 11.1 enables us to evaluate the results obtained in the ABO blood group experiment. Chi-square was computed to be $\chi^2 = 9.20$, but until now we had no way of relating this figure to a probability statement. We could not tell whether this value represented a significant departure from expectation or whether the observed deviations were likely by chance alone. Figure 11.1 shows that an observed $\chi^2 = 9.20$ or greater can be expected to occur less than 5 percent of the time. Actually, Table A.5 shows this figure more accurately to be $\chi^2_{0.05} = 7.81473$. The probability of any particular χ^2 value is zero, as with any continuous random variable. It is the probability of *exceeding* the given value that is of interest in the chi-square test.

A different chi-square probability distribution curve results for each change in degrees of freedom. Several of these curves appear in Fig. 11.2. Each curve is asymmetrical, commencing at zero and tailing off toward positive infinity. Table A.5 presents the probability functions for the common significance levels and for degrees of freedom up to 100.

Now we are in a position to use the chi-square statistic as a hypothesis-testing device. Consider χ^2 in terms of the six steps of hypothesis testing presented in Section 9.2.

Step I. *Statistical hypotheses:* Mendelian theory predicted that blood types AB, A, B, and O should occur in the ratio of 1 : 1 : 1 : 1. The null hypothesis for $n = 100$ trials is therefore

$$\begin{aligned} H_o: \quad & E_1 = 100(0.25) = 25 \\ & E_2 = 100(0.25) = 25 \\ & E_3 = 100(0.25) = 25 \\ & E_4 = 100(0.25) = 25 \end{aligned}$$

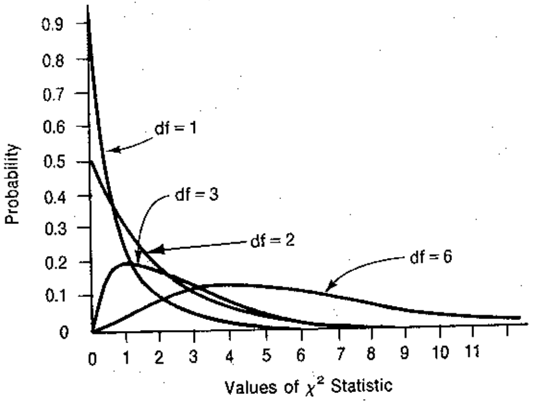

Fig. 11.2 Probability distribution functions for values of χ^2 with several degrees of freedom (after Sokal and Rohlf 1969: fig. 7.12).

The alternative hypothesis states that H_o is false:

$$H_1: \quad E_1 \neq 25; \qquad E_2 \neq 25; \qquad E_3 \neq 25; \qquad E_4 \neq 25$$

Chi-square deals only with generalized deviation and the alternative hypothesis in the chi-square test does not specify just which class (or classes) will deviate from expectation. Any observed value with a large deviation from expectation is sufficient to reject H_o.

Another phrasing of the statistical hypotheses expresses *probabilities* rather than *expectations*. There are four classes in this example, $k = 4$. The probability of class 1 occurring on a given trial is $p_1 = 0.25$; this is the probability of a given offspring having Type AB blood. The probability of Type A blood—class 2—is $p_2 = 0.25$, and so forth. The statistical hypotheses can be expressed in terms of these theoretical relative frequencies (for any sample of size n).

$$H_o: \quad p_1 = p_2 = p_3 = p_4 \qquad H_1: \quad p_1 \neq p_2 \neq p_3 \neq p_4$$

This second version is usually easier to frame when the various probabilities are equal, but in many cases (such as Example 11.2), the expected frequency null hypothesis is easier to visualize.

Once again we should mention the relationship between the chi-square and the binomial distributions. So long as $k = 2$, then the binomial distribution is identical to the chi-square distribution (see Example 11.1).

Step II. *The statistical model:* The chi-square probability distributions (such as those of Fig. 11.2) provide us with a statistical model. This model changes with every level of degree of freedom, so a number of different curves are necessary; Table A.5 summarizes several of the appropriate curves. Thus, our statistical model consists of the χ^2 probability distribution when all assumptions (including H_o) are met. A region of rejection for observed values of the chi-square statistic can be defined, just as with the t-statistic and the standardized normal deviate z. If the observed χ^2 does not deviate from expectation, then we have no reason to question any of our assumptions, and H_o survives. But when an observed chi-square falls into the critical region under the probability distribution, we must search for an invalid assumption. The chi-square statistic is a *nonparametric* statistic, as defined earlier in Section 11.1. The assumptions of the chi-square test are discussed in Section 11.5. As long as these simple assumptions are intact—and a statistical test should not be applied if the assumptions are not valid—then our faulty assumption must be the null hypothesis. All statistical tests operate in this manner.

Step III. *Level of statistical significance:* The alpha level is chosen as before. Although the same general principles for selecting the alpha level apply to chi-square testing, some confusion seems to arise regarding one- and two-tailed alternatives. These difficulties will be discussed in Section 11.9.

Step IV. *Region of rejection:* The critical region is that area under the chi-square sampling distribution which contains unacceptable deviations, given alpha. In the example at hand, with df = 4 − 1 = 3, the 0.05 critical region is given by Table A.5 to be $\chi^2_{0.05}=7.815$.This means that any observed chi-square *greater than or equal to* 7.815 is unacceptably large, given a significance level of 0.05. This is the statistical model against which the actual data are juxtaposed.

Step V. *Calculations and statistical decision:* Formula (11.1) is used to compute the actual observed sample value of the chi-square statistic. In this case, $\chi^2 = 9.20$, a value falling into the region of rejection. Thus, the sample tends to favor H_1 over H_o, given α.

Step VI. *Nonstatistical decision:* As before, these quantitative findings must be rephrased in terms of the research situation. The hypothetical random sample of $n = 100$ offspring has contradicted Mendelian theory. Because such a large deviation will occur by chance fewer than 5 in 100 times of such experiments, we reject the Mendelian theory in this case and search for alternative genetic explanations for our deviant results.

Example 11.1

In 1859, Gregor Mendel conducted a genetic experiment with pea plants (*Pisum*) which were all known to be heterozygous for wrinkled seeds. Mendel found that upon plant maturation, 5474 seeds from his experimental plants were round, while only 1850 seeds were wrinkled. Do these results support Mendel's theory that round seeds should outnumber wrinkled seeds in a 3:1 ratio?

Step I. *Statistical hypotheses:*

$$H_o\colon\ p = 0.75 \qquad H_1\colon\ p \neq 0.75$$

where p is the relative frequency of round seeds.

Step II. *Statistical model:* The chi-square method is appropriate for comparing these two discrete classes (round versus wrinkled seeds). The assumptions of nominal nonparametric tests apply (discussed at the end of Chapter 12).

Step III. *Significance level:* Let $\alpha = 0.05$ for a two-tailed (nondirectional) test.

Step IV. *Region of rejection:* Table A.5 provides the sampling distribution of the chi-square statistic. The degrees of freedom in this case are $\text{df} = k - 1 = 2 - 1 = 1$. The critical region thus contains all values of the chi-square statistic greater than or equal to $\chi^2_{0.05} = 3.841$.

Step V. *Calculations and statistical decision:* The standard chi-square format is as follows:

Outcome	Observed value O_i	Expected value E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
Round	5474	$7324(0.75) = 5493$	−19	361	0.066
Wrinkled	1850	$7324(0.25) = 1831$	19	361	0.197
	7324	7324			$\chi^2 = 0.263$

The observed chi-square statistic does not fall into the critical region. The sample results hence favor H_o at $\alpha = 0.05$.

Step VI. *Nonstatistical decision:* This experiment does not represent a significant departure from the predicted 3:1 Mendelian ratios at $\alpha = 0.05$.

For the simple case of $k = 2$, the chi-square and binomial methods produce identical results. For illustration, the same Mendelian sample can be tested using the normal approximation to the binomial distribution.

Step I. *Statistical hypotheses:*

$$H_o: \quad \mu = np = 7324(0.75) = 5493 \qquad H_1: \quad \mu \neq np \neq 5493$$

where p is the relative frequency of round seeds, and n is the total number of seeds.

Step II. *Statistical model:* The normal approximation to the binomial distribution. Assumptions of nominal level nonparametric tests apply.

Step III. *Significance level:* Let $\alpha = 0.05$ for a two-tailed (nondirectional) test.

Step IV. *Region of rejection:* Any value of $z \geq 1.96$ will fall into the critical region for $\alpha = 0.05$.

Step V. *Calculation and statistical decision:* The experimentally observed results must first be standardized:

$$z = \frac{X_i - \mu}{\sigma} = \frac{5475 - 5493}{37} = -0.49$$

where $\mu = np = 7324(0.75) = 5493$ and $\sigma = \sqrt{npq} = \sqrt{7324(0.75)(0.25)} = 37$.

The observed value of z does not fall within the region of rejection and H_o is retained.

Step VI. *Nonstatistical decision:* This experiment does not represent a significant departure from the expected 3:1 Mendelian ratio, at $\alpha = 0.05$.

Example 11.2

Suppose that a particular theory predicts that, in the long run, hunter-gatherer marriages tend to occur in the following percentage proportions:

Spouse from own village,	25
Spouse's village within 50 miles,	25
Spouse's village more than 50 miles,	50

Julian Steward (1938: 67) collected the following data for the Northern Paiute of the Fish Lake Valley of eastern California:

Spouse from own village,	4
Spouse within valley,	15
Spouse from another valley,	13

Assuming that the radius of the Fish Lake Valley is about 50 miles, are these data consistent with the above theory?

Step I. *Statistical hypotheses:* The expectations arise from preexisting theory: Marriages are predicted to occur in a 1:1:2 ratio for spouse from own village, nearby village, and distant village. In other words, there are three different groups ($k = 3$), each with a distinct probability: $p_1 = 0.25$, $p_2 = 0.25$, $p_3 = 0.50$.

$$\begin{aligned} H_o:\quad & E_1 = np_1 = 32(0.25) = 8 \\ & E_2 = np_2 = 32(0.25) = 8 \\ & E_3 = np_3 = 32(0.50) = 16 \\ H_1:\quad & E_1 \neq 8;\ E_2 \neq 8;\ E_3 \neq 16 \end{aligned}$$

Note here how the alternative hypothesis is composite. H_i simply states that one or more propositions of H_o are false.

Step II. *Statistical model:* The binomial model is no longer applicable because more than two discrete classes are involved ($k > 2$). This is why H_o is expressed as p_1, p_2, and p_3; the p versus q notation of the binomial applies only when $k = 2$. The chi-square sampling distribution is relevant here and nominal level nonparametric assumptions apply.

Step III. *Significance level:* Let $\alpha = 0.05$ for a nondirectional test. Note that we do not specify which of the E_i classes is deviant. Any significant deviation will reject H_o.

Step IV. *Region of rejection:* This example has df $= k - 1 = 3 - 1 = 2$. The critical region thus contains all chi-square statistics $\geq \chi^2_{0.05} = 5.99147$.

Step V. *Calculations and statistical decision:*

Outcome	O_i	E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
Own village	4	8	−4	16	2.000
Within valley	15	8	7	49	6.125
Another valley	13	16	−3	9	0.563
	32	32			$\chi^2 = 8.688$

This χ^2 exceeds the critical value of $\chi^2_{0.05} = 5.99147$ and falls into the region of rejection. The sample data favor H_1, so we reject H_o.

Step VI. *Nonstatistical decision:* The Fish Lake Paiute data depart significantly from the marriage theory at $\alpha = 0.05$. Be sure to note here that chi-square tells us only about the *overall* agreement with theory. By examining the actual data, we see that the Fish Lake Paiute have a much higher rate of spouses from within the valley than the theory predicted.

11.3 TWO-BY-TWO CONTINGENCY TABLES

Section 11.2 introduced the logic for the chi-square statistic, but we have considered only the *univariate* case. As the name implies, the univariate chi-square test treats a single dimension, such as blood type, marriage practices, or seed shape in pea plants. Although univariate chi-squares can ultimately handle an infinity of variables, each dimension must be considered *one at a time*. We will now examine the *bivariate* form of the chi-square test, beginning with the simplest application, the *2 × 2 contingency table*.

In their study of urbanization and its impact upon family structure, Stanley Freed and Ruth Freed collected data in Shanti Nagar, a small village in northern India (Freed and Freed 1969). The Freeds were particularly concerned with the response of traditional family organization to increasing industrialization. They interviewed a random sample of 107 families to determine precisely how the introduction of wage labor influenced traditional family structure.

Family head, 39 years and younger		
Traditional job	26	
Nontraditional cash income	15	41
Family head, 40 years and older		
Traditional job	59	
Nontraditional cash income	7	66
		107

TABLE 11.2 **Chi-square test for goodness of fit for normality. Data are 99 breadth measurements on the first lower molar of pygmy chimpanzee (see Table 4.3).**

Class Interval, mm	O_i	E_i	(O_i-E_i)	$(O_i-E_i)^2$	$(O_i-E_i)^2/E_i$
<7.9	3	4.14	1.14	1.30	0.314
8.0–8.1	7	5.28	1.72	2.96	0.056
8.2–8.3	8	9.07	1.07	1.14	0.126
8.4–8.5	12	12.76	0.76	0.58	0.045
8.6–8.7	12	15.89	3.89	15.13	0.952
8.8–8.9	16	15.91	0.11	0.01	0.001
9.0–9.1	9	14.12	5.12	26.21	1.857
9.2–9.3	18	10.26	7.74	59.91	5.839
9.4–9.5	9	6.27	1.26	1.59	0.253
>9.6	5	5.32	0.32	0.10	0.019
	99	99.02			$\chi^2 = 9.462$

df = 7; $\chi^2_{0.05} = 14.0671$.

area corresponds to the probability of a randomly selected variate being smaller than 8.0 mm. The expected frequency for $n = 99$ is

$$E_1 = np_1 = 99(0.0418) = 4.1$$

The expected frequency of class "8.0–8.1" is determined in a similar manner:

$$z_2 = \frac{X_2 - \bar{X}}{S} = \frac{8.2 - 8.83}{0.479} = -1.31$$

$$A_2 = 0.4582 - 0.4049 = 0.0533$$

The second expected frequency is

$$E_2 = np_2 = 99(0.0533) = 5.3$$

Expected frequencies can be similarly computed for the remaining classes in Table 11.2. As a final check upon these calculations, the summation of the expected values must be equal to n within a small rounding error.

Both expected and observed frequencies are now available. Chi-square can be computed to determine whether the observed deviations are of significant magnitude to reject the null hypothesis.

As mentioned earlier, testing for normality always involves a loss of 3 degrees of freedom because both μ and σ must be estimated from sample statistics. Assuming a significance level of 0.05, Table A.5 indicates that since $\chi^2 = 9.461 < \chi^2_{0.05} = 14.067$. The results are not significant, which means that this sample of 99 variates could easily have been from a normally distributed population of variates.

Example 11.7

In Exercise 3.5, the sample mean and standard deviation were computed for a series of 256 fluted projectile points from Virginia. Before performing a detailed attribute analysis on these data, archaeologist James Fitting tested the variates for normality (Fitting 1965: table 1). Do these length measurements sufficiently follow a normal distribution at the 0.05 level?

In order to keep all cell frequencies above 5, the first two categories (less than 2.9 cm) and the last two divisions (longer than 10.0 cm) were pooled, thereby leaving nine categories ($k = 9$). The mean and standard deviation of the sample are known to be $\bar{X} = 5.8$ cm and $S = 2.07$ cm. The probability of any randomly selected variate from the population with $\mu = \bar{X} = 5.8$ and $\sigma = S = 2.07$ falling into the first category ("less than 2.9 cm") is the area to the left of $X_1 = 3.0$ cm in the z-distribution:

$$z_1 = \frac{3.0 - 5.8}{2.07} = -1.35 \quad A_1 = 0.0885 .$$

The expected frequency of this class is

$$E_1 = 256(0.0885) = 22.66$$

Other expectations are found in a similar manner.

Length, cm	O_i	E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
<2.9	8	22.66	−14.66	214.92	9.484
3.0–3.9	30	26.68	3.32	11.02	0.413
4.0–4.9	56	39.99	16.01	256.32	6.410
5.0–5.9	60	49.02	10.98	120.56	2.459
6.0–6.9	35	45.88	−10.88	118.37	2.580
7.0–7.9	33	34.92	− 1.92	3.69	0.106
8.0–8.9	12	21.50	− 9.50	90.25	4.198
9.0–9.9	13	10.09	2.91	8.47	0.839
>10.0	9	5.43	3.57	12.74	2.347
	256	256.17			$\chi^2 = 28.836$

The degrees of freedom in this case are given by df $= (k - 3) = 6$, and the critical value of chi-square is $\chi^2_{0.05} = 12.5916$. Since the observed value is over twice this expected value, we can conclude that the fluted projectile point lengths were probably not drawn from a normally distributed population. Specifically, this significant departure is caused by the absence of very large points (that is, larger than 10.0 cm) and also by the lack of points shorter than 2.9 cm. There also seems to be an overabundance of points between about 4.0 and 7.0 cm.

11.9 SMALL VALUES OF χ^2: THE STRANGE CASE OF MENDEL'S PEAS

The chi-square variants discussed so far involved only the right-hand tail of the chi-square distribution (Fig. 11.1). The reason for this is that the χ^2 statistic is computed by summing the squares of the deviations. All these deviations from the expected values are positive; therefore, the larger the summed deviations the greater the chi-square statistic. Only the right-hand tail of the distribution is thus involved.

We have also stressed the importance of directionality in the alternative hypothesis for 2×2 tables. Directionality in this case is not identical to one- or two-tailed hypothesis testing of the normal curve. The significance testing of the chi-square distribution involves only the right-hand tail (regardless of directionality). As long as a priori directions were specified in 2×2 tables (either $ad > bc$ or $ad < bc$), the alpha level must be halved. A directional result significant at the tabled 0.10 value, for example, is in fact known to be significant at $\alpha = 0.05$.

But do not conclude from all this that χ^2 methods must *always* be concerned with only the right-hand tail of the distribution. There are some unusual instances when one might wish to determine whether the sum of the squared deviations (as reflected in the χ^2 statistic) could be *too small* to be attributed to chance alone. Suppose that an experiment with 10 degrees of freedom produced a chi-square of, say, 3.00. Table A.5 indicates a probability of greater than 0.975, corresponding to $\chi^2 = 3.00$. That is, there is more than a 97.5 percent chance of a randomly generated chi-square statistic occurring to the *right* of 3.00. However, this situation can also be reversed: There is less than 2.5 percent chance of a randomly generated chi-square value falling to the left of $\chi^2 = 3.00$. Very small chi-squares are themselves rare events, with a known probability of occurrence. In a strictly probabilistic sense, an extremely small chi-square disproves the null hypothesis just as surely as would a very large chi-square value. But within the conventional hypothesis-testing format discussed in Chapter 9, chi-squares with associated probabilities of $p = 0.975$ or larger (while still a rare occurrence) do not allow the rejection of H_o. In fact, the null hypothesis appears to be strongly supported by such results. To resolve this problem, it is necessary to stress again that statistical hypothesis testing must be problem-oriented. Procedures must be designed to answer specific questions at hand rather than merely to blindly follow rigid formats of inquiry.

The problem of the small χ^2 is well illustrated by the classic genetic experiments of Gregor Mendel. Every beginning anthropology student has been subjected to the traditional tale of how Gregor Mendel, an obscure monk, discovered the laws of inheritance. His work was relegated to obscurity until 16 years after his death, when it was miraculously rediscovered by no fewer than three independent scientists. This parable is taken as a tribute to the self-correcting nature of science (the truth will out), and the tale also seems to support the doctrine of independent invention. Independent cultural trajectories can be parallel and often repetitive.

At any rate, statistician Sir Ronald Fisher has posed a highly heretical question. Should we take Mendel literally? Fisher, himself an accomplished geneticist, reconstructed Mendel's famous experiments from various contemporary notes and reports. According to Fisher's reconstruction, Mendel took eight years to complete his experiments. Mendel apparently discovered the critical 3:1 phenotypic ratio rather early in the experimentation, and Fisher wondered aloud whether the actual published experiments represented a true *discovery* or a staged *demonstration* to illustrate previous findings.

Fisher analyzed Mendel's later results, using the chi-square test to scrutinize the role of chance in the genetic experiments. One of Mendel's experiments, for instance, was designed to illustrate the independent segregation of genetic factors—in this case, seed shape and color. In Mendel's notation, the following conditions were involved:

Seed Shape	Seed Color
AA Round (homozygous)	*BB* Yellow (homozygous)
Aa Round (heterozygous)	*Bb* Yellow (heterozygous)
aA Round (heterozygous)	*bB* Yellow (heterozygous)
aa Wrinkled (homozygous)	*bb* Green (homozygous)

Mendel's theory predicts that if the two traits truly segregate randomly in the same plants, then the progeny should appear in a fixed ratio 9:3:3:1, as follows:

(round, yellow):	(round, green):	(wrinkled, yellow):	(wrinkled, green)
9 :	3 :	3 :	1

In 1862, Mendel harvested the seeds of 15 plants known to be heterozygous on both seed shape and color. The offspring seeds were harvested the following year with the following results:

	Seed Shape			
Seed Color	*AA*	*Aa*	*aa*	Total
BB	38	60	28	126
Bb	65	138	68	271
bb	35	67	30	132
Total	138	265	126	529

The predicted ratios (9:3:3:1) were found in the 1863 experiment to be 9.1:3.1:2.9:0.9. When Mendel published these and other findings, he did not analyze the element of chance in experimentation. (The chi-square distribution was unknown at the time, but enough was known about the binomial distribution to estimate the probability of obtaining such satisfactory results.) Mendel ignored random effects and declared that the experimentally devised ratios overwhelmingly confirmed his predictions. These experiments eventually established the independent segregation of genetic traits.

As science progressed throughout the early twentieth century, the role of chance became an important criterion in experimental design. Writing in 1936, Fisher wondered if, given the normal exigencies of genetic experimentation, Mendel's results could be *too good* ? In effect, Fisher tested the *left-hand* side of the chi-square distribution to see if there was *less* deviation (that is, too low a χ^2 value) than one should expect under chance conditions. Fisher's actual computations (Fisher 1936) were quite similar to the bivariate R × C tables considered in Section 11.3. But in this case, the expected probabilities were computed from Mendel's theoretically predicted ratios rather than from marginal totals.

Genotype	Phenotype	O_i	$E_i = np$	$(O_i - E_i)$	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
BB–AA	Round, yellow	38	529 (1/16) = 33.06	4.94	24.40	0.7380
BB–Aa	Round, yellow	60	529 (2/16) = 66.12	6.12	37.45	0.5565
Bb–AA	Round, yellow	65	529 (2/16) = 66.12	1.12	1.25	0.0190
Bb–Aa	Round, yellow	138	529 (4/16) = 132.24	5.76	33.18	0.2510
bb–AA	Round, green	67	529 (2/16) = 66.12	0.88	0.77	0.0120
bb–Aa	Round, green	35	529 (1/16) = 33.06	1.94	3.76	0.1140
BB–aa	Wrinkled, yellow	28	529 (1/16) = 33.06	5.06	25.60	0.7740
Bb–aa	Wrinkled, yellow	68	529 (2/16) = 66.12	1.88	3.53	0.0530
bb–aa	Wrinkled, green	30	529 (1/16) = 33.06	3.06	9.36	0.2830
		529	528.96			$\chi^2 = 2.8005$

These calculations produce a chi-square value of $\chi^2 = 2.8005$, with df $= (k-1) = 8$. Table A.5 indicates that almost 95 percent of all chi-square values (with 8 degrees of freedom) are expected to fall to the right of $\chi^2 = 2.73264$. In other words, we expect such a low chi-square only about one time in every twenty independent experiments.

So, in this particular bifactorial genetic experiment, Mendel obtained results which were uncommonly close to expectation. Fisher went on to investigate the remainder of Mendel's experiments conducted during this eight-year interval. When all these experiments are combined into a single chi-square figure, the computed value of $\chi^2 = 41.606$ with 84 degrees of freedom, a figure corresponding to a probability of $p = 0.9993$. Remembering that chi-square tables refer only to the right-hand tail of the chi-square distribution, the actual probability associated with Mendel's complete results is only $p = 1.000 - 0.9993 = 0.0007$. That is, Fisher demonstrated that there are fewer than 7 in 1,000 chances of obtaining Mendel's results by chance alone.

How do we account for Mendel's near-perfect findings? The early experiments (probably in 1858) may have come as such a revelation to Mendel that he knew enough at that time essentially to frame his entire theory of genetic factor and gametic segregation. His confidence in his early discovery can be seen in several ways: He conducted no further experiments directly to test the 3:1 ratio, since he had already established this to his satisfaction; he ignored the then-current body of statistical inference, through which he could have "tested" his results; he conducted no tests to establish the equivalence of contribution from each parent, preferring simply to assume the 3:1 ratio once again. While it is unfair to accuse Mendel of directly doctoring his results, Fisher contended that chi-square analysis clearly indicates that most (if not all) of Mendel's experiments must have been falsified to agree with expectation. Perhaps Mendel's figures were intended only to illustrate his general principles. Perhaps the results were not to be taken seriously. Mendel could have purposely modified the counts, in order to support a principle he knew to be correct. It is even possible that Mendel was deceived by overly loyal assistants who knew all too well what results the good monk Mendel expected.

Whatever the ultimate explanation, the point is not to deride Mendel's contribution to genetic theory. His insights alone qualify him for plaudits, regardless of the experimental evidence. This intriguing case merely illustrates another application of the chi-square statistic, the situation in which χ^2 is too small to allow for a normal amount of randomness. Mendel's experiments seem to illustrate Fisher's generalization that "fictitious data can seldom survive a careful scrutiny, as, since most men underestimate the frequency of large deviation arising by chance, such data may be expected to agree more closely with expectation than genuine data would" (Fisher 1936).

11.10 FISHER'S EXACT TEST

It was mentioned earlier that although the chi-square distribution is estimated by a continuous curve, observed frequencies are always compared with expected frequencies along discrete intervals. All chi-square statistics therefore only approximate the chi-square continuous curve. These approximations are suitable for most purposes, as long as n is kept suitably large, but in many anthropological cases the frequencies of interest are simply too small to be tested for significance by chi-square methods. Ronald Fisher, the same Fisher who investigated Mendel's genetic experiments, derived a technique for computing the *exact probability* of contingency tables. The approximations are thereby avoided altogether, and this procedure is known as ***Fisher's Exact Test.***

Consider again the generalized 2×2 contingency table:

	First Variable		
Second Variable	+	–	Total
+	a	b	$(a+b)$
–	c	d	$(c+d)$
Total	$(a+c)$	$(b+d)$	n

To determine whether a particular set of results is too rare to have arisen by chance alone, it is necessary to find the probability of obtaining these frequencies in a random experiment. One proceeds in such cases as though this sample were really a population, so the statistical situation can be rephrased: Given the observed marginal totals (which are regarded as fixed), what is the probability of getting random observations within a contingency table as extreme as the observed a, b, c, and d, or results even more extreme?

The number of "successes" for a 2×2 contingency table is defined as the number of possible ways in which the observed cell frequencies could have been randomly selected. Although the derivation of the formula is beyond the scope of this text,[4] it is known that

$$\text{number of successes} = \frac{n!}{a!\,b!\,c!\,d!}$$

[4] This formula has been derived from the coefficients of the *multinomial distribution*, which is analogous to the binomial situation except that the possible outcomes are not limited strictly to success and failure. In this case, there are four possible outcomes for each trial, namely, a, b, c, or d.

The probability fraction associated with this event, the number of successes, becomes the numerator. The denominator in this case is the total number of ways in which a 2 × 2 table could be constructed with the same marginal totals. Beginning with the row totals, $(a + b)$ and $(c + d)$, how many ways can n items be randomly selected such that $(a + b)$ is a success?

Because the order of selection is irrelevant to the probability fraction, the number of randomly selected successes is given by $C_{n,(a+b)}$. Similarly for the column totals, there are exactly $C_{n,(a+c)}$ possible successes. You should convince yourself that identical results would be obtained had either $(c + d)$ been selected for the rows, or $(b + c)$ for the columns. To obtain the total possible combinations between rows and columns, it is necessary to multiply the individual outcomes. Hence, the total numerator of the probability is given by

$$C_{n(a+b)} \cdot C_{n(a+c)} = \frac{n!}{(a+b)!\,(c+d)!} \cdot \frac{n!}{(a+c)!\,(b+d)!}$$

From the coefficients of the multinomial distribution it can be shown that there are

$$\frac{n!}{a!b!c!d!}$$

ways of obtaining the observed cell frequencies. Thus, the probability of obtaining a contingency table with cell frequencies *a,b,c,* and *d* can be computed as the ratio of the two quantities given above. This can be simplified to

$$= \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!\,a!\,b!\,c!\,d!} \tag{11.7}$$

This equation facilitates computation of the exact probability of obtaining the frequencies observed in the 2 × 2 table. But we need to test a null hypothesis which considers not only the frequencies observed but also results to be potentially *more extreme* than those actually obtained. It becomes necessary to compute each individual probability associated with the more extreme results. The summation of all these exact probabilities comprises Fisher's Exact Test.

Consider the following example, which illustrates the computational procedures involved. Suppose that 14 mummies (6 males and 8 females) were discovered in a prehistoric habitation cave in western Nevada. Of these burials, 9 were found to be lacking heads (a frequent custom in this area). From the data at hand, is it justifiable to conclude that males were more frequently decapitated than females (at the 0.05 level)?

Burials	With Skulls	Without Skulls	Total
Male	0	6	6
Female	5	3	8
Total	5	9	14 burials

Specifically, we are interested in whether the two sets of dichotomous categories (sex and burial condition) sort independently of one another. Had this sample been larger, the question could have readily been resolved by the chi-square test, but because $n = 14$ and there seems to be no chance of increasing the sample size, chi-square must be ruled out (for precise statements on the minimum sizes recommended for the chi-square test, see Section 11.11). It is precisely this sort of situation in which Fisher's Exact Test proves to be valuable.

Fisher's test is designed to answer one question: What are the chances of obtaining observed results as extreme (or more extreme) as those obtained in the experiment? The null hypothesis here is that males are just as likely to be decapitated as females. The alternative hypothesis considered the probability that males are *more frequently* decapitated. Specifically, under the alternative hypothesis, we expect:

Few males with skulls (cell *a*)
Several males without skulls (cell *b*)
Several females with skulls (cell *c*)
Few females without skulls (cell *d*)

So, in this case, the alternative hypothesis is directional.[5]

$$H_1: \quad ad < bc$$

We can see by inspection that indeed $ad < bc$, as expected, but we need further to determine whether this could occur by chance alone. (Had the reverse situation occurred in the observed data ($ad > bc$), then H_1 is obviously wrong and no test of statistical significance is required.)

The level of significance must be halved because the alternative hypothesis is directional (are more males decapitated?). In order to reach significance at $\alpha = 0.05$, the exact probability must be less than or equal to $p = 0.025$. This is the region of rejection for Fisher's test.

The exact probability is found by substituting the observed archaeological frequencies into Formula (11.7):

$$p = \frac{6!\,8!\,5!\,9!}{14!\,0!\,6!\,5!\,3!}$$

$$= \frac{4}{143} = 0.028$$

Remember that the final probability in Fisher's Exact Test refers to the observed arrangement, or *more extreme arrangements*. Since H_1 predicts that $ad < bc$, there can be no more extreme arrangement than having a zero in cell *a*. In fact, whenever a zero occurs in any cell of a 2×2 table, the single probability computation covers the most extreme case because no frequency can be more extreme than zero.

[5] When cells *a* and *d* are predicted to be common, a "positive" association is said to exist. Conversely, when the variables are inversely proportional, with cells *b* and *c* more common, then a "negative" association exists. The terms "negative" and "positive" are arbitrarily assigned to distinguish the two sorts of directional alternative hypotheses which might exist in 2 × 2 contingency tables (Coult 1965).

The computed value of $p = 0.028 > p$, $\alpha = 0.025$, an observed figure falling outside the region of rejection, and the results are judged not to be a significant departure from randomness. The excavator cannot justifiably conclude that men within this particular cave site tended to be decapitated more frequently than women.

Suppose that the data occurred in the following configuration:

Burials	With Skulls	Without Skulls	Total
Male	8	2	10
Female	3	5	8
Total	11	7	18

The alternative hypothesis in this case is that $bc < ad$. The probability of obtaining exactly such an arrangement is

$$p_2 = \frac{10!\,8!\,11!\,7!}{18!\,8!\,2!\,3!\,5!} = 0.079$$

Fisher's test is also concerned with more extreme probabilities, so we must also consider all cases in which the product of bc is lower than the observed case; that is, when $bc < 3(2) = 6$. The overall probability of occurrence is given by

$$p = p_0 + p_1 + \cdots + p_r$$

where r = frequency of the rarest cell + 1.

In other words, the overall probability is the sum of the individual observed frequencies plus all other less likely probabilities, given constant marginal totals.

These additional probabilities can be determined by subtracting 1 from both b and c (and, of course, adding 1 to a and d in order to keep the row and column totals constant). The second most extreme arrangement is

$$bc = (2-1)(3-1) = 2$$

and the third most extreme arrangement is

$$bc = (1-1)(2-1) = 0$$

Cell b is empty in the third case, so there can be no more extreme arrangements along the diagonal bc.

For the present example, p_2 has already been computed to be 0.079.

$$p_1 = \frac{10!\,8!\,11!\,7!}{18!\,9!\,1!\,2!\,6!} = 0.0088$$

$$p_0 = \frac{10!8!11!7!}{18!10!0!1!7!} = 0.00025$$

The total probability of the observed frequency or those more extreme is

$$p = p_2 + p_1 + p_0 = 0.079 + 0.0088 + 0.00025 = 0.088$$

This final probability figure of $p = 0.088$ is sufficiently greater than $p_\alpha = 0.025$ assigned so that the null hypothesis cannot be rejected. The observed frequency of this second example could well have arisen simply by chance.

So far, we have used Fisher's Exact Test only to examine the directional alternative hypothesis; in the strict sense, this test should be reserved for directional cases. But there are times when an alternative hypothesis can be stated only about the *existence* of an association rather than its direction. The alternative hypothesis in such cases is simply

$$H_1: \quad ad \neq bc$$

In addition to the probability of obtaining the observed frequency, one must also compute the more extreme positions of positive ($ad > bc$) and negative ($ad < bc$) associations.

Suppose that the last problem had been expressed differently? Does decapitation appear to have any association with sex? No direction is expressed in this statement, so both positive and negative associations must be considered. The probability of more positive associations was computed above, so all that remains to solve the nondirectional hypothesis is to determine the probability of the more extreme negative associations.

The most extreme negative association would be when cell d is empty, rendering $ad = 0$. The probability for this case is

$$p_{0\,(\text{negative})} = \frac{10!\,8!\,11!\,7!}{18!\,3!\,7!\,8!\,0!} \cong 0.0038$$

The next smallest extreme negative association (with a 1 in cell d) has the probability of

$$p_{1\,(\text{negative})} = \frac{10!\,8!\,11!\,7!}{18!\,4!\,7!\,6!\,1!} \cong 0.0528$$

Both cases are "more extreme" than the observed arrangement of frequencies because their probabilities are smaller than the observed frequencies:

$$p_{0\,(\text{negative})} = 0.0038 \qquad p_2 \cong 0.079$$
$$p_{1\,(\text{negative})} = 0.0528 \qquad p_2 \cong 0.079$$

The next largest negative association (with a 2 in cell d) has a probability of

$$p_{2\,(\text{negative})} = \frac{10!\,8!\,11!\,7!}{18!\,5!\,5!\,6!\,2!} \cong 0.2217$$

This probability is not "more extreme" than that of the observed arrangement ($p_2 = 0.079$), and the value of $p_{2\,(\text{negative})} = 0.2217$ should not be included in the summary probability statement.

The total probability of the two-tailed alternative is given as the sum of (1) the probability of the observed case, (2) the probabilities of more extreme positive associations, and (3) the probabilities of more extreme negative associations:

$$\begin{aligned} p &= p_2 + p_1 + p_0 + p_{0\,(\text{negative})} + p_{1\,(\text{negative})} \\ &= 0.079 + 0.0088 + 0.00025 + 0.0038 + 0.0528 \\ &= 0.145 \end{aligned}$$

Because this final probability $p = 0.145 > p_\alpha = 0.05$, the null hypothesis cannot be rejected for the nondirectional case.

Fisher's Exact Test involves a prodigious amount of calculation when the smallest cell frequency on the relevant diagonal is much greater than about 3. In such cases, one is well advised to process these data upon a computer, and adequate programs are readily available (for example, Sokal and Rohlf 1969: 702–703). There are also tables to cover some of the values for Fisher's Exact Test (Siegel 1956: tables; row and column totals smaller than 15). Unfortunately, these tables partially vitiate the precision of the "exact" test, since one reads only significance levels rather than the exact probabilities. But for most hypothesis-testing purposes, these levels of significance are adequate.

Example 11.8

Noncommercial societies often practice a unilocal residential pattern in which the newly married couple moves to a prescribed setting: patrilocal, matrilocal, or avunculocal. Both societies may also simultaneously practice two or more patterns of consanguinal residence, in a situation termed *multilocal.* Carol Ember and Melvin Ember (1972) conducted a cross-cultural study to test several explanations of the multilocal residential pattern. One theory holds that multilocal societies tend to have undergone recent depopulation so that choice of spouses is limited to a survival population. For purposes of this test, the Embers operationally defined a population as depopulated if the population had dropped more than 25 percent in the 30-year period prior to fieldwork.

Do the following 27 cases, randomly selected from the Human Relations Area Files, support the depopulation hypothesis at the 0.05 level?

Data of this sort are commonly presented in a special form of the 2 × 2 contingency table, in which the actual society name rather than simply the cell frequency is entered into the cell. Presentation in this fashion allows investigators to examine the sample societies for other variables of interest, such as geographic area, subsistence base, or linguistic stock.

	Multilocal Residence		
Depopulation	Present	Absent	Total
Present	Chukchee Comanche Ila Lau Mandan Nambicuara Yaruro	Crow Kaska Tapirape Tehuelche Tlingit	12

	Multilocal		
Depopulation	Present	Absent	Total
Absent	Burmese	Annamese Aymara Bikinians Burusho Cuna Kikuyu Koi Lepcha Pukapukans Seri Somali Tikopia Toda Wogeo	15
Total	8	19	27

The directional alternative hypothesis (do multilocal societies tend to be depopulated?) involves a "positive" association:

$$H_1: \quad ad > bc$$

For these results to be significant, the observed probability must exceed $\alpha/2 = 0.05/2 = 0.025$.

Fisher's Exact Test is appropriate in this case because of the low expected frequency in cell d: ($E_d = 15(8)/27 \approx 4.44$). Two separate probabilities are computed: the observed case in which cell c contains one case, and the more extreme instance in which cell c would have been empty (but with the row and column totals fixed).

The probability associated with the observed frequencies is

$$p_1 = \frac{12!\,15!\,19!\,8!}{27!\,5!\,7!\,14!\,1!} = 0.00535$$

and the probability of the more extreme case on the ad diagonal is

$$p_0 = \frac{12!\,15!\,19!\,8!}{27!\,4!\,8!\,15!\,1!} = 0.00022$$

So the total probability of observing data this extreme or more extreme by chance alone is

$$\begin{aligned} p = p_1 + p_0 &= 0.00535 + 0.00022 \\ &= 0.00557 \end{aligned}$$

Since $p = 0.006 < p = 0.025$, the null hypothesis is rejected, and the conclusion is that the Embers' cross-cultural test supports an association between residence and recent depopulation.

11.11 GENERAL SIZE CONSIDERATIONS

The following size recommendations can serve as guidelines for applying the chi-square and Fisher's Exact tests (Cochran 1954; Grizzle 1967).

Guideline I. *Two-by-two contingency tables.*

A. Use *Fisher's Exact Test* if
1. n is less than 20; or
2. n is between 20 and 40, and the smallest E_i is less than 5.

B. Use the chi-square test if
1. n is greater than 40; *or*
2. n is between 20 and 40, and the smallest E_i is greater than 5.
3. Yates' Correction for Continuity is necessary only when the smallest E_i is less than 10.

Guideline II. *$R \times C$ contingency tables* (*where* $R > 2$ *or* $C > 2$). Chi-square is permissible if

A. All E_i are greater than 5; *or*

B. No more than about 20 percent of the cells have E_i less than 5 *and* no E_i is less than 1; *or*

C. More than about 20 percent of the cells have E_i less than 5 *and* no E_i is less than 2.

11.12 THE McNEMAR TEST FOR CORRELATED PROPORTIONS

The chi-square and Fisher's Exact tests are the most common methods for examining relationships between two variables in the 2×2 format. Both tests assume two conditions: (1) the sample has been randomly selected from its population, and (2) the two samples are *mutually independent*. All previous 2×2 tables have implicitly conformed to these assumptions. The null hypothesis has been that all cell frequencies should be in relative proportion to their corresponding row and column totals. Any disproportionate cell will inflate the sample statistic and hence cast doubt upon H_o.

But suppose that the second assumption has been violated and that the samples lack mutual independence. Are the chi-square and Fisher's Exact tests then invalid? Quite simply: yes, they are. Because these standard contingency tests cannot be used when the variables are mutually dependent, an alternative is recommended.

What specifically is meant by *mutual independence* within a contingency table? Two variables are dependent when the frequencies of one variable logically influence the values of the second variable. A prime example of this relationship is the familiar "before–after" research design of psychological or educational experiments. Suppose that 150 college sophomores are questioned whether or not they think marijuana should be legalized. Each subject is then shown a recorded television segment in which a number of heroin addicts testify that marijuana led them directly into abuse of hard drugs. Graphic examples of suicides "under the influence of marijuana" are presented along with clinical discussions of the links between marijuana usage, lung cancer, and chromosome damage. These same 150 subjects are then questioned again: "Do you now believe that marijuana should be legalized?"

Some students who formerly favored legalization will probably change their attitudes because of the potential hazards. But others who were originally opposed to legalization will undoubtedly resent the biased television presentation as "brainwashing" and will favor legalization largely as a means of protest. The results from the experiment can be arrayed in the familiar 2 × 2 format.

	After Television Segment	
Before TV Segment	Favor Legalization	Oppose Legalization
Favor legalization	*a*	*b*
Oppose legalization	*c*	*d*

Is there a significant change in attitude due to the television program?

Meaningless chi-square or Fisher's Exact statistics could easily be computed from these data. This experiment violates the assumption of mutual independence between samples, and hence both tests are invalid. The *same* subjects have been asked the *same* questions, so the "before" variable influences the "after" variable.

The *McNemar Test for Correlated Proportions* is specifically designed to assess the significance of change between dependent variables. The chi-square test is sensitive to changes in *all four cells*; any major deviation from expectation inflates chi-square. But in the above cases, interest is only in those cells denoting change, that is, the number of students who have changed their attitudes toward legalization of marijuana. Cells *a* and *d* represent continuity, those individuals who declined to alter their opinions regarding legalization. Only cells *b* and *c* represent changes in attitude, and the McNemar test provides a statistical method for assessing the relative significance of change. The null hypothesis of *no change* states simply that the frequencies of cells *b* and *c* should be roughly similar. The larger the discrepancy between cells *b* and *c*, the less tenable is the null hypothesis. As long as the sample remains relatively large, any particular probability can be approximated by the chi-square distribution where

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c} \tag{11.8}$$

with a single degree of freedom.[6]

Consider an archaeological application of the McNemar statistic. Three archaeologists—Tom, Dick, and Harriet— have convened to discuss the prehistoric cultural sequence of the Yahoo Basin. A total of 75 archaeological sites are known from this area, and the session begins with Dick and Harriet comparing their analyses of these sites. Harriet classifies 50 of the 75 sites into the Early

[6] Note that although the *chi-square distribution* is used to find the probability of the McNemar statistic, the assumptions and methods of the McNemar test are quite distinct from those of the 2 × 2 χ^2 test.

Period, and the remaining 25 sites into the Late Period; Dick classifies these same sites as 47 Early and 28 Late. Because of this surprising amount of disagreement, they tabulate their findings for site-by-site comparisons. Of the 50 sites that Harriet has called Early, Dick has agreed with only 40, calling the remaining 10 sites Late. Harriet has classified 25 sites Late, while Dick has listed 28 Late sites. In other words, Harriet and Dick have agreed on only 40 Early sites and 18 Late sites, and they have disagreed on the temporal affinity of the remaining 17 sites.

	Harriet		
Dick	Early	Late	Total
Early	40	7	47
Late	10	18	28
Total	50	25	75

Most archaeologists realize how much subjectivity is involved in such cases, and error will never be eliminated. Random errors are of less concern, since they tend to cancel one another in the long run, but systematic errors of classification are more serious and can disrupt entire cultural sequences. Harriet and Dick disagree on 17 sites. How much disagreement is due to random errors of classification and how much to a systematic bias resulting from differing conceptions of Early and Late phases?

The McNemar test is useful here because only cells *b* and *c* of the contingency table are involved; these are the cells of disagreement. Clearly, the typology lacks precision, but do systematic errors appear? The McNemar statistic is computed from Formula (11.8),

$$\chi^2 = \frac{(|7-10|-1)^2}{10+7} = 0.235$$

This small value of chi-square (with a single degree of freedom) does not approach the significant values in Table A.5, so the null hypothesis is not in danger. Harriet and Dick do not appear to be classifying the sites in significantly different ways. Their differences are due simply to random errors, and future research will surely reduce this random component.

The situation is somewhat different when Harriet's typology is compared with Tom's list. Tom and Harriet likewise disagree on the temporal placement of 17 sites. The percentage disagreement is exactly the same as that between Dick and Harriet (23%). But take a closer look at the nature of the disagreement.

	Harriet		
Tom	Early	Late	Total
Early	36	3	39
Late	14	22	36
Total	50	25	75

Harriet has classified only three of Tom's Late sites as Early, and Tom has classified 14 of Harriet's Early sites as Late. The McNemar comparison is as follows:

$$\chi^2 = \frac{(|14-3|-1)^2}{17} = 5.88$$

This value of chi-square is significant past the 0.05 level.

The peculiar situation illustrated above needs to be considered in a bit more detail because an important concept is at stake. Two comparisons were made based upon a single sample of 75 sites. Harriet's classifications agreed with Dick's by 58/75 = 77 percent, and Harriet's also agreed with Tom's by 77 percent. The two situations were identical in overall agreement. The character of disagreement is quite different. Harriet and Dick essentially split their differences in half. They disagreed randomly. The McNemar test evaluated the cell frequencies for *b* and *c* (7 and 10, respectively) and concluded that these proportions could readily be due to random error. Harriet and Tom likewise disagreed on 17 of the sites, but the proportion of cases between the critical cells seems to be out of line, with $b = 3$ and $c = 14$. The McNemar test concluded that this disproportionate outcome will occur by chance in fewer than 5 in 100 random samples. Some systematic source of error is probably at work here: Either Tom consistently calls Harriet's Early sites as Late, or Harriet consistently classifies Tom's Late sites Early. The difference cannot be distinguished on these data alone. Although both comparisons involved an error of 23 percent, the errors between Harriet's and Tom's typology seem more grievous because a systematic bias results.

● *Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.*—H. G. Wells

Example 11.9

A matrilocal residence system is one in which a newly married couple takes up residence in the village of the bride's mother. Anthropologists have attempted to explain the origin of specific matrilocal systems for well over a decade, but few have devised a set of specific causes which can explain *all* matrilocal systems. A recent study by Divale (1974) attempts just such a universal explanation.

Divale's argument goes as follows: When a population migrates into an already inhabited region, warfare usually results, and the society best equipped to fight such battles will have an adaptive advantage. Matrilocal residence selects for more efficient warfare because the agnatically related males are scattered over several communities; patrilocal systems do not fare so well because the females rather than the male warriors are scattered. Thus, matrilocality is caused by migration and is an adaptation to the resulting disequilibrium. Does the following cross-cultural sample support Divale's contention that recently migrated societies tend to change to matrilineality?

	After Migration		
Before Migration	Matrilocal	Patrilocal	Total
Matrilocal	39	12	51
Patrilocal	35	32	67
Total	74	44	118

The sample consists of 118 societies which are known to have recently migrated. The post-marital residential pattern remains unchanged in 71 of these societies, while 47 societies have changed their patterns (12 from matrilocal to patrilocal and 35 in the other direction).

This contingency table cannot be tested using the chi-square statistic because the same variable (residence) has been measured twice on each society. McNemar's test is the appropriate statistic to test the significance of the residential change:

$$\chi^2 = \frac{(|35-12|-1)^2}{35+12} = 10.298$$

with a single degree of freedom. The result is significant beyond the 0.01 level. Chance phenomena do not seem sufficient to explain this change, and the data from these 118 societies do not conflict with Divale's hypothesis of matrilocal residence patterns.

SUGGESTIONS FOR FURTHER READING

Conover (1971: chapter 4)
Hays (1973: chapter 11)
Morrison and Henkel (1970). A collection documenting abuses of chi-square and other significance tests in the social sciences.
Siegel (1956: chapters 4, 6, 8)

EXERCISES

11.1 The following figures reveal the cross-cultural prevalence of riddles in 137 societies (data from Roberts and Forman 1971):

	Level of Political Integration			
	Absent	Autonomous Local	Minimal State	State
Riddles absent	9	40	16	31
Riddles present	0	8	10	23

(a) Is there a significant difference in riddling behavior between societies with autonomous local political organization and societies with the state? (Use a chi-square statistic.)

(b) Recompute part (a) using the binomial distribution.

11.2 Investigators from the Government Hospital Tel-Hashomer, Israel, conducted a long-range study on colorblindness among Jews and Arabs living in Israel. The following cases of red–green blindness were noted in a large sample of informants living in central Israel (data from Adam, Doron, and Modan 1967):

	Normal Vision	Red/Green Blindness
Arabs	638	75
Jews	1085	43

Does this study indicate a significant difference in the frequency of colorblindness?

11.3 In a study on the cultural patterning of sexual beliefs and behavior, Minturn *et al.* (1969) generated a cross-cultural sample of 135 societies using the Human Relations Area File.

(a) Do the following data, extracted from their survey, support the hypothesis that divorce is more difficult in societies in which the nuclear family is the primary social unit?

	Ease of Divorce	
Family Organization	Difficult	Easy
Extended	12	25
Nuclear	7	3

(b) Why did you select the coefficient you did?

11.4 The Graduate Division of the University of California, Berkeley, processed a total of 12,763 applications for graduate study for the fall of 1973 (data from Bickel, Hammel, and O'Connell 1975: table 1).

	Outcome	
Applicants	Admit	Deny
Men	3738	4704
Women	1494	2827

(a) Do these data indicate that sexual bias is operative in the admission process?

(b) What is the danger of applying the chi-square statistic in this case?

*11.5 Some social anthropologists have hypothesized a functional relationship between Hawaiian kinship terminology and prohibition of cross cousin marriage. Do these data from the *Ethnographic Atlas* support such a hypothesis (data from Goody 1970: table 6)?

	Prohibition on Cross-Cousin Marriage	
Hawaiian Kin Terms	Present	Absent
Present	200	39
Absent	219	206

11.6 The following data characterize phenotypic frequencies of the *ABO* blood system among three Macro-Maya speaking societies in southern Mexico (data from Cordova, Lisker, and Loria 1967: 59)?

	ABO System Phenotype			
	A	*B*	*O*	*AB*
Chol	16	1	135	0
Chontol	10	3	88	0
Totonac	9	0	70	0

(a) Is there a significant difference in *ABO* phenotypes among the three groups?

(b) In the above calculation, which phenotypes (if any) must be excluded from the chi-square calculation? Why?

11.7 The acculturational study of rural Buganda discussed earlier (Chapter 2) generated the following data (Robbins and Pollnac 1969: table 5).

	Beverage Choice		
Age	Traditional	Mixed	Modern
17–40	13	16	11
40+	29	7	2

Do these data support the notion that the younger members of Buganda society prefer nontraditional alcoholic beverages?

11.8 Based upon a sample from the *Ethnographic Atlas*, Ember and Ember (1971) determined the following relationship between warfare and residence:

	Pattern of Residence	
Warfare	Matrilocal	Patrilocal
External	Callinago Cherokee Creek Kaska Navaho Miskito	
Internal	Mataco Yao	Azanda Ganda Jivaro Kapauku Murngin Nama Nootka Nuer Tallensi Tiv

Is there a significant relationship between warfare and residence?

11.9 As part of a study on blood-group frequencies in the higher primates, a team of scientists tested the chimpanzees at the Edinburgh Zoo for the ability to taste PTC. A total of 27 chimps were tested (data from Fisher, Ford, and Huxley 1939).

	Males	Females
Taster	11	9
Nontaster	3	4

(a) Do the male chimps seem to have a greater ability to taste PTC than the females?

(b) Why is the chi-square statistic an invalid measure in this case?

*11.10 The following mortality figures come from three North American archaeological sites. The archaic population is from Indian Knoll, Kentucky; the Hopewellian series is from the Pete Klunk Mounds in southwestern Illinois; and the Middle Mississippian sample is from the Dickson Mounds, also in Illinois (data from Blakely 1971: table 3).

	Age at Death		
	0–19	20–39	40+
Archaic	60	18	23
Hopewell	106	80	108
Middle Mississippian	215	150	114

(a) Is the mortality rate significantly different between the archaic and Hopewellian samples?

(b) Is the age at death different between Hopewell and Middle Mississippian samples?

(c) Are the three samples significantly different from one another?

(d) What levels of measurement are involved?