# SOME METHODS FOR STRENGTHENING THE COMMON $\chi^2$ TESTS*

WILLIAM G. COCHRAN
*The Johns Hopkins University*

## 1. INTRODUCTION

Since the $\chi^2$ tests of goodness of fit and of association in contingency tables are presented in many courses on statistical methods for beginners in the subject, it is not surprising that $\chi^2$ has become one of the most commonly-used techniques, even by scientists who profess only a smattering of knowledge of statistics. It is also not surprising that the technique is sometimes misused, e.g. by calculating $\chi^2$ from data that are not frequencies or by errors in counting the number of degrees of freedom. A good catalogue of mistakes of this kind has been given by Lewis and Burke (1).

In this paper I want to discuss two kinds of failure to make the best use of $\chi^2$ tests which I have observed from time to time in reading reports of biological research. The first arises because $\chi^2$ tests, as has often been pointed out, are not directed against any specific alternative to the null hypothesis. In the computation of $\chi^2$, the deviations $(f_i - m_i)$ between observed and expected frequencies are squared, divided by $m_i$ in order to equalize the variances (approximately), and added. No attempt is made to detect any particular pattern of deviations $(f_i - m_i)$ that may hold if the null hypothesis is false. One consequence is that the usual $\chi^2$ tests are often insensitive, and do not indicate significant results when the null hypothesis is actually false. Some forethought about the kind of alternative hypothesis that is likely to hold may lead to alternative tests that are more powerful and appropriate. Further, when the ordinary $\chi^2$ test does give a significant result, it does not direct attention to the way in which the null hypothesis disagrees with the data, although the pattern of deviations may be informative and suggestive for future research. The remedy here is to supplement the ordinary test by additional tests that help to reveal the significant type of deviation.

In this paper a number of methods for strengthening or supplementing the most common uses of the ordinary $\chi^2$ test will be presented and illustrated by numerical examples. The principal devices are as follows:

*Work assisted by a contract with the Office of Naval Research, U. S. Navy Department. Dept. of Biostatistics paper no. 278.

1. Use of small expectations in computing $\chi^2$.
2. Use of a single degree of freedom, or a group of degrees of freedom, from the total $\chi^2$.
3. Use of alternative tests.

Most of the techniques have been available in the literature for some time: indeed, most of them stem from early editions of Fisher's "Statistical Methods for Research Workers." Research which has clarified the problem of subdividing $\chi^2$ in contingency tables is more recent, and still continues.

In the hope of avoiding some confusion, the symbol $X^2$ will be used for the quantity that we calculate from the sample in a chi-square test. The symbol $\chi^2$ itself will refer to a random variate that follows the distribution in the $\chi^2$ tables, and will sometimes be used in the phrase "the $\chi^2$ test."

## 2. USE OF SMALL EXPECTATIONS IN COMPUTING $X^2$

In order to prove that the quantity $X^2 = \sum(f_i - m_i)^2/m_i$ is distributed as $\chi^2$ when the null hypothesis is true, it is necessary to postulate that the expectations $m_i$ are large: in fact, the proof is strictly valid only as a limiting result when the $m_i$ tend to infinity. For this reason many writers recommend that the $m_i$ be not less than 5 when applying the test in practice, and that neighboring classes be combined if this requirement is not met in the original data. Some writers recommend a lower limit of 10 for the $m_i$ .

It is my opinion that these recommendations are too conservative, and that their application may on occasion result in a substantial loss of power in the test. I give this as an opinion, because not enough research has been done to make the situation quite clear. However, the exact distribution of $\sum(f_i - m_i)^2/m_i$ , when the expectations are small, has been worked out in a number of particular cases by Sukhatme (2), Neyman and Pearson (3) and Cochran (4), (5). These results indicate that the $\chi^2$ tables give an adequate approximation to the exact distribution even when some $m_i$ are much lower than 5.

Loss of power from following a rule that $m_i \geq 5$ occurs because this rule tends to require grouping of classes at the tails or extremes of the distribution. These are often the places where the difference between the alternative hypothesis and the null hypothesis stands out most clearly, so that the grouping may cover up the most marked difference between the two hypotheses. Information about the extent of the loss of power is unfortunately very scanty, because the power function of $X^2$ is known only as a limiting result when the $m_i$ are large. The following illustration suggests that the loss can be large.

Suppose that we have a sample of $N = 100$. The null hypothesis is that the data follow a Poisson distribution with mean $m$ known to be 1, when actually the data follow the negative binomial distribution $(q - p)^{-n}$, where $n = 2$, $q = 1.5$, $p = 0.5$. What is the chance of rejecting the null hypothesis at the 5% level of significance?

The expected frequencies of 0, 1, 2, 3, and 4 or more occurrences on the two hypotheses are shown in table 1.

TABLE 1.
EXPECTED FREQUENCIES ON THE NULL AND ALTERNATIVE HYPOTHESIS.

| | Expected frequencies | | $\frac{(m'_i - m_i)^2}{m_i}$ | | |
|---|---|---|---|---|---|
| | | | | Grouped | |
| $i$ | Poisson $m_i$ | Neg. Bin. $m'_i$ | Ungrouped | $m_i \geq 5$ | $m_i \geq 10$ |
| 0 | 36.79 | 44.44 | 1.590 | 1.590 | 1.590 |
| 1 | 36.79 | 29.63 | 1.393 | 1.393 | 1.393 |
| 2 | 18.39 | 14.82 | 0.693 | 0.693 | 0.009 |
| 3 | 6.13 | 6.58 | 0.033 | 1.181 | |
| 4+ | 1.90 | 4.53 | 3.640 | | |
| Totals $N$ = | 100.00 | 100.00 | $\lambda = 7.349$ | $\lambda = 4.857$ | $\lambda = 2.992$ |

If an $m_i$ as low as 1.90 is allowed, we can use 5 classes in the $\chi^2$ test, with 4 degrees of freedom since the $m_i$ are known. To make all $m_i \geq 5$, we must pool the last two classes and have 3 degrees of freedom; and to make all $m_i \geq 10$, we must pool the last three classes and have 2 degrees of freedom. In order to obtain an approximation to the powers of these three ways of applying the $\chi^2$ test, we shall use the asymptotic result for the power function. This is a non-central $\chi^2$ distribution, with parameter $\lambda$ of non-centrality, where

$$\lambda = \sum \frac{(m'_i - m_i)^2}{m_i}$$

The larger the value of $\lambda$, the higher is the power. The contributions to $\lambda$ from each class are shown in the right-hand columns of table 1. For the ungrouped case, note that the extreme class 4+ is much the largest contributor to $\lambda$. Grouping the last two classes considerably reduces this contribution, while grouping the last three classes diminishes it almost to zero. The approximate probabilities of rejecting the null hypothesis may be read from Fix's tables (6) of the non-central $\chi^2$ distribution. These probabilities are 0.56 for the ungrouped case, 0.43 for $m_i \geq 5$ and 0.32 for $m_i \geq 10$. The loss in sensitivity from grouping is evident. Perhaps a more revealing comparison is to compute from Fix's tables the sizes of sample $N$ that would be needed in the two grouped cases in order to bring the probabilities of rejection up to 0.56, the value for the ungrouped case. The results are $N = 136$ when the last two classes are grouped and $N = 191$ when the last three classes are grouped, as against $N = 100$ in the ungrouped case.

This example is only suggestive, and probably favors the ungrouped case slightly, because the computations are based on large-sample results. However, the losses in power from grouping, measured in terms of equivalent sample sizes, are impressive.

### 2.1 *Recommendations about minimum expectations*

Elsewhere, (7), I have given recommendations about the minimum expectation to be used in $\chi^2$ tests. These working rules may be summarized, in slightly revised form, as follows:

(*a*) *Goodness of fit tests of unimodal distributions* (*such as the normal or Poisson*). Here the expectations will be small only at one or both tails. Group so that the minimum expectation at each tail is at least 1.

(*b*) *The* $2 \times 2$ *table.* Use Fisher's exact test (i) if the total $N$ of the table $< 20$, (ii) if $20 < N < 40$ and the smallest expectation is less than 5. Mainland (8) has given useful tables of the exact test for these cases. If $N > 40$ use $X^2$, corrected for continuity.

(*c*) *Contingency tables with more than* 1 *d.f.* If relatively few expectations are less than 5 (say in 1 cell out of 5 or more, or 2 cells out of 10 or more), a minimum expectation of 1 is allowable in computing $X^2$.

Contingency tables with most or all expectations below 5 are harder to prescribe for. With very small expectations, the exact distribution of $X^2$ can be calculated without too much labor. Computing methods have been given by Freeman and Halton (9). If $X^2$ has less than 30 degrees of freedom and the minimum expectation is 2 or more, use of the ordinary $\chi^2$ tables is usually adequate. If $X^2$ has more than 30 degrees of freedom, it tends to become normally distributed, but when the expectations are low, the mean and variance are different from those of the tabular $\chi^2$. Expressions for the exact mean and variance have been given by Haldane (10). Compute the exact mean and variance, and treat $X^2$ as normally distributed with that mean and variance.*

*In a previous paper (7) I recommended this procedure only when the degrees of freedom in $X^2$ exceed 60. Some unpublished research suggests that 30 is a better division point.

Further research will presumably change these recommendations, but I do not believe that the recommendations will lead users far astray.

Succeeding sections will deal with some of the common applications of the $\chi^2$ test to goodness of fit problems and to contingency tables. The alternative or supplementary tests to be presented are those that seem most often useful, but they by no means exhaust the possibilities. The important guiding rule is to think about the type of alternative that is likely to hold if the null hypothesis is false, and to select a test that will be sensitive to this kind of alternative.

## 3. THE GOODNESS OF FIT TEST OF THE POISSON DISTRIBUTION

This is the test already referred to in table 1, except that in practice the parameter $m$ must usually be estimated from the data. We have

$$X^2 = \sum \frac{(f_i - m_i)^2}{m_i}$$

where $f_i$ is the observed and $m_i$ the expected frequency of an observation equal to $i(i = 0, 1, 2 \cdots)$.

If the data do not follow the Poisson distribution, two common alternatives are as follows:

(1) The data follow some other single distribution, such as the negative binomial or one of the "contagious" distributions. Another way of discribing this case is to postulate that the individual observations, say $x_i$ , follow Poisson distributions, but with means that vary from observation to observation so as to follow some fixed frequency distribution. For instance, as has been shown, the negative binomial distribution can be produced by assuming that these means follow a Pearson type III distribution.

(2) The means of the observations $x_i$ follow some systematic pattern. With data gathered over several days, the means might be constant within a day, but vary from day to day, or they might exhibit a slow declining trend.

### 3.1 *Test of the variance*

In both cases (1) and (2), a comparison between the observed variance of the observations $x_i$ and the variance predicted from Poisson theory will frequently be more sensitive than the goodness of fit test. The variance test is made by calculating

$$X_v^2 = \sum_{i=1}^{N} \frac{(x_i - \bar{x})^2}{\bar{x}} ,$$

or if the calculation is made from the frequency distribution of the $x_i$ ,

$$X_v^2 = \sum_i \frac{f_i(i - m)^2}{m},$$

where $m = \bar{x}$ is the sample mean. The quantity $X_v^2$ is referred to the $\chi^2$ tables with $(N - 1)$ degrees of freedom. The variance test is an old one: it was introduced by Fisher in the first edition of "Statistical Methods for Research Workers" under the heading "Small samples of the Poisson series", because it can be used when the sample is too small to permit use of the goodness of fit test.

The increased power of the variance test over the goodness of fit test was strikingly shown in some sampling experiments conducted by Berkson (20), in a situation in which the data followed a binomial distribution, which has a *smaller* variance than the Poisson. A rough calculation for the example in table 1 gives 0.76 as the probability of rejecting the Poisson hypothesis by the variance test, as compared with 0.56 for the best of the goodness of fit tests. In practice, I have often found the variance test significant when the goodness of fit test was not. Berkson (21) presents some data that illustrate this point.

### 3.2 *Subdivision of degrees of freedom in the variance test*

If it is suspected that the means of the observations $x_i$ may change in some *systematic* manner, as in case (2) above, more specific tests of significance can be obtained by selecting certain degrees of freedom from $X_v^2$ . The ordinary rules of the analysis of variance are followed in subdividing $\sum(x_i - \bar{x})^2$, and the denominator $\bar{x}$ is used to convert the partial sum of squares approximately to $\chi^2$. A few examples will be given. (The formulas presented are intended to make clear the structure of the $\chi^2$ components, but they are not always the speediest formulas to use in computing the components.)

### 3.3 *Test for a change in level*

To test for an abrupt change in the mean of the distribution, occurring after the first $N_1$ observations, we take

$$X^2 = \frac{N_1 N_2}{(N_1 + N_2)} \frac{(\bar{x}_1 - \bar{x}_2)^2}{\bar{x}} \qquad \text{(1 d.f.)}$$

where $N_2 = (N - N_1)$, $\bar{x}_1$ , and $\bar{x}_2$ are the sample means in the two parts of the series and $\bar{x}$ is the overall mean. This test may be extended to compare a group of means.

### 3.4 *Test for a linear trend*

We may anticipate that the means will follow a linear regression on some variate $z_i$ (frequently a time-variable). In this case

$$X^2 = \frac{[\sum (x_i - \bar{x})(z_i - \bar{z})]^2}{\bar{x} \sum (z_i - \bar{z})^2} \qquad \text{(1 d.f.)}$$

### 3.5 *Detecting the point at which a change in level occurs*

This problem has been illustrated by Lancaster (11) in an experiment in which increasing concentrations of disinfectant are poured on a series of suspensions of a bacterial culture, each suspension having a constant amount of bacteria. The observations $x_i$ represent numbers of colonies per plate. The problem is to find the value of $j$ for which the disinfectant is strong enough to begin reducing the number of colonies. To this end we compare each observation with the mean of all previous observations, looking for the first value of $j$ at which $X^2$ becomes large owing to a drop in the number of colonies. We thus obtain a set of independent single degrees of freedom:

$$X_1^2 = \frac{(x_1 - x_2)^2}{2\bar{x}}; \qquad X_2^2 = \frac{(x_1 + x_2 - 2x_3)^2}{6\bar{x}};$$

and in general

$$X_r^2 = \frac{(x_1 + x_2 + \cdots + x_r - rx_{r+1})^2}{r(r+1)\bar{x}}$$

*Note.* The above set of $X^2$ values will add up to the total variance $X_v^2$, but as Lancaster has pointed out, there are other natural ways of subdividing $X_v^2$ in which the separate $X^2$ do not add up to $X_v^2$. When comparing $x_1$ with $x_2$, we might decide to disregard the remainder of the observations and compute $X_1^2$ as

$$X_1'^2 = \frac{(x_1 - x_2)^2}{2\bar{x}_{12}}$$

where $\bar{x}_{12}$ is the mean of $x_1$ and $x_2$. A set of successive $X^2$ values computed in this way will not add up to $X_v^2$, because the denominator changes from term to term, whereas $X_v^2$ carries the denominator $\bar{x}$.

The practice of computing $X^2$ components by using only those parts of the data that are immediately involved has something in its favor (despite the non-additive feature), at least if the total $X_v^2$ has already been shown to be significant. For in that event we have already concluded that the data as a whole do not follow a single Poisson distribution, and the overall mean $\bar{x}$ is of dubious validity as an estimate of the Poisson variance for a part of the data. On the other hand, if the total $X_v^2$ is not significant, but we suspect that some component is, the additive partition is convenient and should be satisfactory in a preliminary examination.

### 3.6 *Single degrees of freedom in the goodness of fit test*

In the goodness of fit comparison of the observed frequency $f_i$ with the expected frequency $m_i$ of $i$ occurrences, we can test any linear function of the deviations

$$L = \sum g_i(f_i - m_i)$$

where the $g_i$ are numbers chosen in advance.

In the case in which the mean of the Poisson, and hence the $m_i$, are given in advance, the variance of $L$ is

$$V(L) = \sum g_i^2 m_i - \frac{(\sum g_i m_i)^2}{N} \tag{1}$$

In the more common situation in which the Poisson mean $m$ is estimated from the data,

$$V(L) = \sum g_i^2 m_i - \frac{(\sum g_i m_i)^2}{N} - \frac{[\sum g_i m_i(i - m)]^2}{Nm} \tag{2}$$

where the sums are over the values 0, 1, 2, $\cdots$ of $i$. In either case

$$X_1^2 = \frac{L^2}{V(L)} \tag{3}$$

is approximately distributed as $\chi^2$ with 1 d.f. I plan to publish justification for formula (2), which appears to be new. By appropriate choice of the $g_i$, a test specific for a given pattern of deviations is obtained.

In particular, to test any single deviation $(f_i - m_i)$ when $m$ is estimated from the data, we take

$$L = (f_i - m_i) : V(L) = m - \frac{m_i^2}{N}\left\{1 + \frac{(i - m)^2}{m}\right\} \tag{4}$$

As an illustration, the data in table 2 are for a sample which gave a satisfactory fit to a Poisson distribution. However, in copying down the frequencies before fitting the Poisson, the frequency of 3 occurrences was erroneously written as 52 instead of 32.

TABLE 2.
GOODNESS OF FIT TEST FOR A SAMPLE WITH A GROSS ERROR.

| $i$ | $f_i$ | $m_i$ | Contribution to $X^2$ |
|---|---|---|---|
| 0 | 52 | 47.65 | 0.40 |
| 1 | 67 | 77.04 | 1.31 |
| 2 | 58 | 62.28 | 0.29 |
| 3 | 52 | 33.56 | 10.13 |
| 4 | 7 | 13.56 | 3.17 |
| 5 | 3 | 4.39 | 0.44 |
| 6+ | 1 | 1.52 | 0.18 |
| Total | 240 | 240.00 | 15.92 |

The value of $m$ is 1.6167. The total $X^2$, 15.92 with 5 d.f., is significant at the 1 per cent level. The large contribution to $X^2$ from $i = 3$ excites notice. In order to test this deviation, we take

$$L = f_3 - m_3 = 52 - 33.56 = 18.44$$

$$V(L) = 33.56 - \frac{(33.56)^2}{240}\left\{1 + \frac{(3 - 1.6167)^2}{1.6167}\right\} = 23.31$$

$$X_1^2 = \frac{(18.44)^2}{23.31} = 14.59$$

This comparison accounts for the major part of the total $X^2$. It must be pointed out, however, that the $X_1^2$ test applies only to a deviation picked out before seeing the data. Thus the test can be applied validly, for $i = 0$, say, if we suspect beforehand that the data follow the Poisson distribution for $i \geq 1$, but that the frequency for $i = 0$ may be anomalous. If the test is applied, as here, to a deviation selected because it looks abnormally large, the significance $P$ obtained is too low. I do not have an expression for the correct significance probability when we select the largest deviation. It appears intuitively that the correct probability lies between $P$ and $kP$, where $k$ is the number of classes in the goodness of fit test. Since $P$ is about 0.00013 and $k$ is 7, the upper limit is .00091, which is still highly significant statistically.

## 4. THE GOODNESS OF FIT TEST OF THE BINOMIAL DISTRIBUTION

For the binomial distribution, there is a series of tests analogous to those given for the Poisson distribution. A typical observation consists of the number of successes $x_i$ out of $n$ independent trials. We have a sample of $N$ such observations. The ordinary goodness of fit test is made by recording the frequency $f_i$ with which $i$ successes occur in the sample, fitting a binomial to these frequencies, and calculating $X^2$ as

$$X^2 = \sum \frac{(f_i - m_i)^2}{m_i},$$

where $m_i$ is the corresponding expected frequency.

As in the Poisson case, departures from the binomial frequently occur either because

(1) the data follow a different frequency distribution, usually with a larger variance (or the probabilities of success $p_j$ show some kind of random variation from observation to observation).

(2) the probabilities $p_j$ are affected by a systematic source of variation.

In both cases, a comparison of the observed and expected variances is likely to be more sensitive than the goodness of fit test.

### 4.1 *Test of the variance*

The test criteria, all distributed approximately as $\chi^2$, are as follows.

*n constant, p given in advance*

$$X_v^2 = \frac{\sum_j (x_j - np)^2}{npq} = \frac{\sum_i f_i(i - np)^2}{npq}, \qquad N \text{ d.f.}$$

*n constant, p estimated*

$$X_v^2 = \frac{\sum_j (x_j - \bar{x})^2}{n\hat{p}\hat{q}} = \frac{\sum_i f_i(i - n\hat{p})^2}{n\hat{p}\hat{q}}, \qquad (N - 1) \text{ d.f.}$$

where $n\hat{p} = \bar{x}$, and $\hat{q} = 1 - \hat{p}$.

*n varying, p given in advance*

In this case we cannot make a simple goodness of fit test (unless the sample is large enough to be divided into batches, each with $n$ constant, so that the test can be made separately for each batch). If $p_j = x_j/n_j$ is the observed proportion of successes in the $j$th member of the sample,

$$X_v^2 = \frac{\sum n_j(p_j - p)^2}{pq} \qquad N \text{ d.f.}$$

*n varying, p estimated*

$$X_v^2 = \frac{\sum n_i(p_i - \hat{p})^2}{\hat{p}\hat{q}} = \frac{\sum \frac{x_i^2}{n_i} - \frac{(\sum x_i)^2}{(\sum n_i)}}{\hat{p}\hat{q}}, \qquad (N - 1) \text{ d.f.}$$

where $\hat{p} = (\sum x_i)/(\sum n_i)$ is the estimate of $p$ from the total sample.

There is another way of deriving the variance test. Arrange the data in a 2 × $N$ contingency table, as follows.

| | | | | |
|---|---|---|---|---|
| Successes<br>Failures | $x_1$<br>$n_1 - x_1$ | $x_2$<br>$n_2 - x_2$ | | $x_N$<br>$n_N - x_N$ |
| Total | $n_1$ | $n_2$ | | $n_N$ |

Then the $X^2$ that is used to test for association in this 2 × $N$ table may be shown to be identical with $X_v^2$ .

If $N$ exceeds 30 and the expectations are small, it was pointed out in section 2 that $X^2$ in contingency tables tends to a normal distribution with a mean and variance somewhat different from those of $\chi^2$. Use of Haldane's correct expressions for the mean and variance of $X^2$ was recommended in this case. The same procedure is recommended in the variance test if $N$ exceeds 30 and the average $n_i$ is less than 10. Haldane's expressions are rather complicated when the $n_i$ vary. In the fairly common situation in which $n$ is constant and $p$ is estimated from the data, the following results suffice in almost all cases.

$$E(X_v^2) \doteq (N - 1)\left(1 + \frac{1}{Nn}\right) \tag{5}$$

$$V(X_v^2) \doteq 2(N - 1)\left(\frac{n-1}{n}\right)\left(1 - \frac{1 - 7\hat{p}\hat{q}}{Nn\hat{p}\hat{q}}\right) \tag{6}$$

The important correction term is that in $(n - 1)/n$ in the variance: the terms in $1/Nn$ are usually small. These results will be used in the numerical example which follows.

The data in table 3, taken from a previous paper, (12), illustrate the application of the goodness of fit test and the variance test to the same observations. The original data, due to Dr. J. G. Bald, consisted of 1440 tomato plants in a field having 24 rows with 60 plants in a row. For each plant it was recorded whether the plant was healthy or attached by spotted wilt as of a given date. As one method of examining whether the distribution of diseased plants was random over the field, the plants were divided into 160 groups of 9, each group consisting of 3 plants × 3 rows. Thus $N = 160$, $n = 9$. The choice of $n = 9$ was of course arbitrary, and I do not know what would have been the best choice. The obvious alternative to a binomial distribution is that the values of $p_i$ vary from one group of 9 to another, indicating a patchiness in distribution.

TABLE 3.
NUMBERS OF DISEASED PLANTS IN GROUPS OF 9 PLANTS.

| $i$ | $f_i$ | $m_i$ | Contr. to $X^2$ |
|---|---|---|---|
| 0 | 36 | 26.45 | 3.45 |
| 1 | 48 | 52.70 | 0.42 |
| 2 | 38 | 46.67 | 1.61 |
| 3 | 23 | 24.11 | 0.05 |
| 4 | 10 | 8.00 | 0.50 |
| 5 | 3 | 2.05 | 4.25 |
| 6 | 1 | | |
| 7 | 1 | | |
| 8 | 0 | | |
| | N = 160 | 159.98 | 10.28 |

Allowing a minimum expectation of 1, we must pool the last 4 classes in table 3. The value of $X^2$ is 10.28, with 4 d.f., since $p$ is estimated; the significance $P$ is 0.036. (If we pooled the last 5 classes in order to have a minimum expectation of 5, we would obtain a significance $P$ of 0.046.)

For the variance test, we compute $X_v^2$ from the observed frequency distribution in table 3. We have

$$\sum if_i = 261 = Nn\hat{p}, \qquad \text{so that } \hat{p} = 0.18125.$$

$$X_v^2 = \frac{\sum i^2 f_i - (\sum if_i)^2/N}{n\hat{p}\hat{q}} = \frac{727 - (261)^2/160}{9(0.18125)(0.81875)}$$

$$= 225.55, \text{ with 159 d.f.}$$

Since $N = 160$ and $n = 9$, we use the normal approximation to the distribution of $X_v^2$ , based on the correct mean and variance. In expressions (5) and (6), the terms in $1/Nn$ are negligible ($Nn = 1440$). Hence we take

$$E(X_v^2) = 159 : V(X_v^2) = 2(159)(8)/9 = 282.66$$

The approximate normal deviate is

$$\frac{225.55 - 159}{\sqrt{282.66}} = 3.96$$

This has a significance $P$ less than 0.0001, much lower than that obtained from the goodness of fit tests.

Subdivision of the sum of squares for $X^2_v$ , which may be useful in testing for systematic variation of the $p_i$ , will be discussed in section 6, which deals with 2 × $N$ contingency tables.

### 4.2 *Single degrees of freedom in the goodness of fit test*

Let

$$L = \sum g_i(f_i - m_i)$$

be a specified linear function of the deviations of observed from expected frequencies. If $p$ is given, formula (1) in section 3.6 holds for the variance of $L$. If $p$ is estimated from the data,

$$V(L) = \sum g_i^2 m_i - \frac{(\sum g_i m_i)^2}{N} - \frac{[\sum g_i m_i(i - n\hat{p})]^2}{Nn\hat{p}\hat{q}} \qquad (7)$$

For a single deviation, $(f_i - m_i)$, selected in advance,

$$V(L) = m_i - \frac{m_i^2}{N}\left\{1 + \frac{(i - n\hat{p})^2}{n\hat{p}}\right\} \qquad (8)$$

Then $L^2/V(L)$ is approximately distributed as $\chi^2$ with 1 d.f.

## 5. THE GOODNESS OF FIT TEST OF THE NORMAL DISTRIBUTION

When the normal distribution is fitted to a body of data, both the mean and the variance are estimated from the sample: consequently, no variance test is possible. However, the variance test is just an application of the general procedure in which we compare the lowest moments (or cumulants) in which the sample can differ from the theoretical distribution that is being fitted. In this sense, the analogue of the variance test is the test for skewness (as given e.g. in Fisher's "Statistical Methods for Research Workers," §14), which will often detect a departure from normality that escapes the goodness of fit test. The test for kurtosis is also useful in this connection.

As with the Poisson and binomial distributions, a specified linear function $L$ of the deviations in the goodness of fit test can be scrutinized. The formula for the variance of $L$ is somewhat more complicated, because the observed frequencies are subject to 3 constraints. If

$$L = \sum g_i(f_i - m_i)$$

then

$$V(L) = \sum g_i^2 m_i - \frac{(\sum g_i m_i)^2}{N} - \frac{(\sum g_i d_i m_i)^2}{Ns^2} - \frac{[\sum g_i m_i(d_i^2 - s^2)]^2}{2Ns^4}$$

where

$d_i$ = (midpoint of $i$th class) − (sample mean)

$s^2$ = sample estimate of variance

In order to apply this test, construct 3 additional columns containing the quantities $g_i m_i$, $d_i$ and $(d_i^2 - s^2)$, respectively. $V(L)$ is then easily computed.

For testing a *single* deviation, $V(L)$ simplifies to

$$V(L) = m_i - \frac{3m_i^2}{2N} - \frac{m_i^2 d_i^4}{2Ns^4}$$

The series of supplementary tests described above for the Poisson, binomial and normal distributions can be extended to other distributions. In particular, variance and skewness tests for the negative binomial distribution have been developed by Anscombe (18) and further illustrated by Bliss (19).

## 6. SUBDIVISION OF DEGREES OF FREEDOM IN THE 2 x $N$ CONTINGENCY TABLE

This section describes some useful ways in which the total $X^2$ for a 2 × $N$ contingency table may be subdivided. The notation, which continues that already used for the binomial distribution, is as follows.

| | Number of | | | Proportion of Successes |
|---|---|---|---|---|
| | Successes | Failures | Total | |
| | $x_1$ | $n_1 - x_1$ | $n_1$ | $p_1 = x_1/n_1$ |
| | $x_2$ | $n_2 - x_2$ | $n_2$ | $p_2 = x_2/n_2$ |
| | - - - - | - - - - - - - | - - - - - - - | - - - - - - - - |
| | $x_N$ | $n_N - x_N$ | $n_N$ | $p_N = x_N/n_N$ |
| Totals | $T_x$ | $T - T_x$ | $T$ | $\bar{p} = T_x/T$ |

For many purposes, a formula due to Brandt and Snedecor is useful in interpreting the total $X^2$. The formula is:

$$X^2 = \frac{\sum_{i=1}^{N} n_i(p_i - \hat{p})^2}{\hat{p}\hat{q}} \tag{9}$$

Thus the total $X^2$ is seen to be a weighted sum of squares of the deviations of the individual proportions of success $p_i$ from their mean, with weights $n_i/\hat{p}\hat{q}$. Consequently, if we subdivide this weighted sum of squares into a set of independent components by the rules of the analysis of variance, we obtain a corresponding subdivision of $X^2$ into independent components.

### 6.1 *Test for a change in the level of* p

In order to test whether the value of $p$ is different in the first $N_1$ rows from that in the subsequent $N_2$ rows ($N = N_1 + N_2$), we may subdivide $X^2$ into the following 3 components.

| | d.f. |
|---|---|
| Difference between $p$'s in first $N_1$ and last $N_2$ rows | 1 |
| Variation among $p_i$ within the first $N_1$ rows | $(N_1 - 1)$ |
| Variation among $p_i$ within the last $N_2$ rows | $(N_2 - 1)$ |

For the following example I am indebted to Dr. Douglas P. Murphy. A group of women known to have cancer of the uterus and a corresponding 'control' group (primarily from a dental clinic and several women's clubs) were selected. From each of a defined set of relatives of the selected person (the proband), data were secured about the presence of cancer. A higher proportion of cancer cases among relatives of the cancer proband would indicate some kind of familial aggregation of the disease, perhaps of genetic origin (13).

In one table, data were presented separately for those relatives who were of the same generation as the proband (e.g. sister) and for those relatives who were one generation earlier (e.g. mother). Some breakdown of this kind is advisable, because cancer attacks mainly in middle or old age, and the 'cancer' and 'control' groups might be found to differ in the proportions of young relatives which they contained. The data, which are a small part of a much more intensive investigation, appear in table 4.

There are several ways of computing the total $X^2$ in a 2 × $N$ table. One form of the Brandt-Snedecor formula is

$$X^2 = \frac{\sum x_i p_i - \hat{p}T_x}{\hat{p}\hat{q}} \tag{10}$$

TABLE 4.
CANCER AMONG RELATIVES OF 'CANCER' AND 'CONTROL' PROBANDS

| | | No. of relatives | | | Proportion with cancer |
|---|---|---|---|---|---|
| | Proband | With | Without | Total | |
| | | $x_i$ | | $n_i$ | $p_i$ |
| Earlier Generation | Cancer | 86 | 814 | 900 | 0.095556 |
| | Control | 117 | 1038 | 1155 | 0.101299 |
| Same Generation | Cancer | 49 | 1475 | 1524 | 0.032152 |
| | Control | 61 | 1580 | 1641 | 0.037172 |
| | Total | 313 | 4907 | 5220 | $\hat{p} = 0.059962$<br>$\hat{q} = 0.940038$<br>$\hat{p}\hat{q} = 0.056367$ |

This expression is useful when there is some question of a systematic variation in the $p_i$, because we will want to compute the $p_i$ in order to have a look at them. When the $p_i$ have been computed, the numerator of $X^2$ can be obtained from formula (10) directly on the computing machine, without any intermediate writing down. The only disadvantage is that a substantial number of decimal places must be retained in the $p_i$.

If this method is to be used, first compute $\hat{p}$, $\hat{q}$ and the product $\hat{p}\hat{q}$ (bottom right of table 4). Since $\hat{p}\hat{q}$ is about 1/20, the numerator of $X^2$ must be correct to at least 3 decimal places if we want $X^2$ correct to 2 decimal places. Further, since $T_x$ is 313, we should have 6 decimal places in the $p_i$. (The symbol $x_i$ should be assigned to the column with the *smaller* numbers: this makes the computations lighter and necessitates fewer decimals in the $p_i$).

From inspection of the $p_i$ in table 4, a large difference in cancer rates between the two generations is evident. Within each generation, the differences in rates between the cancer and control groups appear tiny. To illustrate the methods, the total $X^2$ will be partitioned into the 3 relevant components. All that is necessary is to subdivide the sum of squares in the numerator, and then divide each component by $\hat{p}\hat{q}$. For the numerator of the total $X^2$, we have from formula (10),

$$\begin{aligned}\text{Total } S.S. &= (86)(0.095556) + \cdots + (61)(0.037172) - (313)(0.059962)\\ &= 5.1447\end{aligned}$$

For the comparison of the two generations, we form the auxiliary 2 × 2 table.

| | With<br>$x_j$ | Without | Total<br>$n_j$ | Proportion<br>$p_j$ |
|---|---|---|---|---|
| Earlier generation | 203 | 1852 | 2055 | 0.098783 |
| Same generation | 110 | 3055 | 3165 | 0.034755 |
| Total | 313 | 4907 | 5220 | 0.055962 |

The same formula can be used for this table.

$$\begin{aligned}\text{Generations } SS &= (203)(0.098783) + (110)(0.034755) - \\ &\quad (313)(0.055962) \\ &= 5.1079.\end{aligned}$$

For the comparisons between cancer and control groups within generations, we have

$$\begin{aligned}\text{First generation } SS &= (86)(0.095556) + (117)(0.101299) - \\ &\quad (203)(0.098783) = 0.0169.\end{aligned}$$

The second generation $SS$, obtained similarly, gives 0.0199. In table 5 the results are summarized and converted to $X^2$ values on division by $\hat{p}\hat{q}$.

TABLE 5.
SUBDIVISION OF $X^2$ INTO COMPONENTS (2 x 4 TABLE)

| Component | d.f. | S.S. | $X^2$ | From 2 x 2 tables |
|---|---|---|---|---|
| First vs. later generation | 1 | 5.1079 | 90.62 | 90.62 |
| Cancer vs. Control: first gen. | 1 | 0.0169 | 0.30 | 0.19 |
| Cancer vs. Control: later gen. | 1 | 0.0199 | 0.35 | 0.59 |
| Total | 3 | 5.1447 | 91.27 | 91.40 |

There is no indication of any difference in cancer rates between the cancer and control groups within either generation. In order to complete this analysis, we should make a combined test of Cancer vs. Control from the two generations. Methods for making tests of this kind are discussed in section 8.

As mentioned previously in connection with the corresponding subdivision for the Poisson distribution, separation of $X^2$ into *additive* components is convenient for a preliminary examination of the data. But if differences in $p$ are found between groups of rows, the $X^2$ values for comparisons made *within* groups need to be recomputed. This is clear in the present example. The additive method requires the assumption that the estimated variance of any $p_i$ is $\hat{p}\hat{q}/n_i$ . However, the huge $X^2$ value of 90.62 between generations shows that the combined $\hat{p}$ cannot be regarded as a valid estimate of the proportion of cancer cases within either of the individual generations.

The procedure is to recompute the two 'within-generation' $X^2$, each from its own 2 × 2 table. These values, which no longer add to the original total $X^2$, are given in the right hand column of table 5. The interpretation of the data is not altered. The $X^2$ values computed from individual parts of the table seldom differ greatly from the additive $X^2$, but they can do so in certain circumstances and are worth looking at, as a precaution, in analyses of this type.

In this example the rows were divided into 2 groups. The same methods may be applied to test the variation in $p$ among any number of groups, and also within each group. To obtain an additive separation, we subdivide the numerator of $X^2$ as indicated in the example. Alternatively, for a non-additive separation, we can form an auxiliary table in which each row is a group total and obtain the $X^2$ for this table, and a further $X^2$ for each group considered by itself.

### 6.2 *Test for a linear regression of p.*

In some contingency tables we may expect that the $p_i$ will bear a linear relation to a variate $z_i$ that is defined for each of the $N$ rows of the table. In others, where the rows fall into a natural order, it is not unreasonable to assign scores $z_i$ to the rows, in an attempt to convert the ordering into a continuous scale, with which the $p_i$ may show a linear relation. Since $p_i$ is assigned a weight $n_i/\hat{p}\hat{q}$, the regression coefficient $b$ of $p_i$ on $z_i$ is obtained by the standard formula for weighted regressions:

$$b = \frac{\sum n_i(p_i - \hat{p})(z_i - \bar{z}_w)}{\sum n_i(z_i - \bar{z}_w)^2} \tag{11}$$

where $\bar{z}_w$ is the weighted mean of the $z_i$ .

For computing purposes, the numerator and denominator of $b$ are conveniently expressed as follows (note that $n_i p_i = x_i$):

$$\text{Num.} = \sum x_i z_i - \frac{T_x(\sum n_i z_i)}{T} \tag{12}$$

$$\text{Den.} = \sum n_i z_i^2 - \frac{(\sum n_i z_i)^2}{T} \tag{13}$$

The $X^2$ for regression, with 1 d.f., is

$$X^2 = \frac{(\text{Num.})^2}{\hat{p}\hat{q}(\text{Den.})} \tag{14}$$

As an illustration, the data in table 6, for which I am indebted to the Leonard Wood Memorial, (American Leprosy Foundation) are taken from an experiment on the use of drugs (sulfones and streptomycin) in the treatment of leprosy. The rows denote the *change* in the overall clinical condition of the patient during 48 weeks of treatment: the columns indicate the degree of infiltration (a measure of a certain type of skin damage) present at the beginning of the experiment. The question of interest is whether patients with much initial infiltration progressed differently from those with little infiltration. Patients did not all receive the same drugs, but since no difference in the effects of drugs could be detected, it was thought that the data for different drugs could be combined for this analysis.

TABLE 6.
196 PATIENTS CLASSIFIED ACCORDING TO CHANGE IN CONDITION AND DEGREE OF INFILTRATION

| Clinical change | | Score $z_j$ | Degree of infiltration | | Total $n_j$ | $p_j = x_j/n_j$ (in %) | $n_j z_j$ |
|---|---|---|---|---|---|---|---|
| | | | $0 - 7$ | $8 - 15$ $x_j$ | | | |
| Im-provement | Marked | 3 | 11 | 7 | 18 | 39 | 54 |
| | Moderate | 2 | 27 | 15 | 42 | 36 | 84 |
| | Slight | 1 | 42 | 16 | 58 | 28 | 58 |
| | Stationary | 0 | 53 | 13 | 66 | 20 | 0 |
| | Worse | −1 | 11 | 1 | 12 | 8 | −12 |
| Total | | | 144 | 52 $T_x$ | 196 $T$ | 0.26531 $\hat{p}$ | 184 |

The total $X^2$ is 6.88, with 4 d.f. ($P$ about 0.16). However, the $p_j$ (the proportions of patients with severe infiltration) decline steadily from 39% in the "markedly improved" class to 8% in the "worse" class. This suggests that a regression of the $p_j$ on the clinical change might furnish a more sensitive test.

The data are typical of many tables in that the rows (clinical changes) are ordered. In order to compute a regression, this ordering must be replaced by a numerical scale. I have supposed that scores 3, 2, 1, 0, −1, as shown in table 6, have been assigned to the five classes of clinical change. Such scores are to some extent subjective and arbitrary, and some scientists may feel that the assignment of scores is slightly unscrupulous, or at least they are uncomfortable about it. Actually, any set of scores gives a *valid* test, provided that they are constructed without consulting the results of the experiment. If the set of scores is poor, in that it badly distorts a numerical scale that really does underlie the ordered classification, the test will not be sensitive. The scores should therefore embody the best insight available about the way in which the classification was constructed and used. In the present example, I considered an alternative set 4, 2, 1, 0, −2, on the grounds that the doctor seemed to deliberate very carefully before assigning a patient to the "markedly improved" or to the "worse" class, but I decided that the presumption in favor of this scale was not strong enough.

To compute the regression $X^2$, the only supplementary column needed is that of the products $n_iz_i$ , shown on the right of table 6. From equations (12) and (13) for $b$,

$$\begin{aligned}\text{Num.} &= \sum x_iz_i - \frac{T_x(\sum n_iz_i)}{T}\\ &= (7)(3) + (15)(2) + \cdots + (1)(-1) - \frac{(52)(184)}{196}\\ &= 17.1837 \qquad (15)\end{aligned}$$

$$\begin{aligned}\text{Den.} &= \sum n_iz_i^2 - \frac{(\sum n_iz_i)^2}{T}\\ &= (54)(3) + (84)(2) + \cdots + (-12)(-1) - \frac{(184)^2}{196}\\ &= 227.2653 \qquad (16)\end{aligned}$$

$$X^2 \text{ for regression} = \frac{(17.1837)^2}{(227.2653)\hat{p}\hat{q}} = 6.666 \qquad (1 \text{ d.f.})$$

This is significant at the 1% level. The total $X^2$ has now been subdivided as follows.

| | d.f. | $X^2$ |
|---|---|---|
| Regression of $p_i$ on $z_i$ | 1 | 6.67 |
| Deviations from regression | 3 | 0.21 |
| Total | 4 | 6.88 |

6.3 *Comparison of mean scores.*

There is another way of looking at the relation between degree of infiltration and clinical progress. We might ask whether the degree of improvement is, on the average, better for patients with severe infiltration than for patients with less infiltration. In many applications, including the present one, this is a more natural way of posing the scientific question than by asking whether the proportion of patients with severe infiltration changes from class to class.

Yates (14) has shown how to compute $X^2$, with 1 d.f., for comparing the mean scores, and has proved, in a more general case, that this $X^2$ is identical with the regression $X^2$ which we have just considered. In the 'mean score' approach, it is best to think of the two groups of patients as representing two independent samples. Each patient has been assigned a variate or score $z_i$ , and the $x_i$ now represent the frequencies with which these scores occur in one of the samples. The mean score for a sample is

$$\bar{z} = \frac{\sum x_i z_i}{\sum x_i} = \frac{\sum x_i z_i}{T_x}$$

In table 6, this has the value 0.8194 for patients with 0–7 degrees of infiltration and 1.2692 for patients with 8–15 degrees. The formula for the variance of a mean score is, from Yates (14),

$$V(\bar{z}) = \frac{1}{T_x T}\left[\sum z_i^2 n_i - \frac{(\sum z_i n_i)^2}{T}\right] \tag{17}$$

Note that, except for the term in $T_x$ , this formula is exactly the same for each of the two means.

The value of the term in square brackets has already been obtained in equation (16) as 227.2653. Hence, from equation (17) and the data in table 6,

$$V(\bar{z}_1) = \frac{227.2653}{(144)(196)} : V(\bar{z}_2) = \frac{227.2653}{(52)(196)}$$

$$V(\bar{z}_1 - \bar{z}_2) = \frac{227.2653}{196}\left\{\frac{1}{144} + \frac{1}{52}\right\} = 0.03035$$

The $X^2$ value for the difference in mean scores is

$$X^2 = \frac{(\bar{z}_1 - \bar{z}_2)^2}{V(\bar{z}_1 - \bar{z}_2)} = \frac{(1.2692 - 0.8194)^2}{0.03035} = 6.666$$

in agreement with the regression $X^2$.

### 6.4 *Step by step comparisons of the $p_i$ .*

Occasionally it is useful to compare $p_1$ with $p_2$ , the weighted mean of $p_1$ and $p_2$ with $p_3$ , and so on. To obtain an additive subdivision of $X^2$, we partition the sum of squares in the numerator of $X^2$ as follows:

$$\frac{n_2\{n_1p_1 - n_1p_2\}^2}{n_1(n_1 + n_2)},$$

$$\frac{n_3\{n_1p_1 + n_2p_2 - (n_1 + n_2)p_3\}^2}{(n_1 + n_2)(n_1 + n_2 + n_3)},$$

and for the general term,

$$\frac{n_{r+1}\{n_1p_1 + \cdots + n_rp_r - (n_1 + \cdots + n_r)p_{r+1}\}^2}{(n_1 + \cdots + n_r)(n_1 + \cdots + n_{r+1})}$$

Each term is then divided by $\hat{p}\hat{q}$ to convert it to $X^2$ with 1 d.f.

For the corresponding non-additive subdivision, we calculate $X^2$, by the standard methods, for each of the following set of 2 × 2 tables.

| | | |
|---|---|---|
| $x_1$ | $y_1$ | $n_1$ |
| $x_2$ | $y_2$ | $n_2$ |
| $x_1 + x_2$ | $y_1 + y_2$ | |

| | | |
|---|---|---|
| $x_1 + x_2$ | $y_1 + y_2$ | $n_1 + n_2$ |
| $x_3$ | $y_3$ | $n_3$ |
| $x_1 + x_2 + x_3$ | $y_1 + y_2 + y_3$ | |

and so on, where we have written $y_i$ for $(n_i - x_i)$.

## 7. THE GENERAL TWO-WAY CONTINGENCY TABLE

For considerations of space, this case will be discussed very briefly. The following notation will be used for the frequencies in the $r \times c$ cells of a table with $c$ columns and $r$ rows.

| | | | | | Total |
|---|---|---|---|---|---|
| | $x_{11}$ | $x_{21}$ | | $x_{c1}$ | $n_1$ |
| | $x_{12}$ | $x_{22}$ | | $x_{c2}$ | $n_2$ |
| | $x_{1r}$ | $x_{2r}$ | | $x_{cr}$ | $n_r$ |
| Total | $T_1$ | $T_2$ | | $T_c$ | $T$ |

One fairly common situation is that the rows can be divided into $g$ groups. We expect that within any group there is no association between rows and columns, but there is reason to examine for association among the group totals. To put it another way, the ratios of the $p_{ij}$ from column to column may be constant within a group but vary from group to group. For this situation, a non-additive separation of the total $X^2$ is obtained by: (1), dividing the contingency table into the appropriate $g$ contingency tables, and calculating the $X^2$ in the ordinary way for each separate table; (2), forming a new $g \times c$ table in which the entries are column totals within each group, and calculating the ordinary $X^2$ for this table.

Lancaster (11) and Irwin (15) have shown how to make a complete (non-additive) separation into single d.f. by subdividing the table into a set of $2 \times 2$ contingency tables, $(r - 1)(c - 1)$ in number, and computing $X^2$ in the ordinary way for each $2 \times 2$ table. The method of forming the $2 \times 2$ tables is easily grasped. The same authors have given general methods for the production of additive subdivisions of the total $X^2$. In the case of the partition into single d.f. just mentioned, Kimball (16) gives an expeditious method of computing the individual $X^2$ values. As in the $2 \times N$ case, the basic difference between additive and non-additive partitions is that in the former the estimated variances used to convert the sums of squares into approximate $\chi^2$ values are obtained from the margins of the whole table, while in the non-additive partition they are obtained separately from the margins of each part.

### 7.1 *Score methods.*

If the rows, say, represent an ordered classification, and scores $z_j$ can be assigned to them, we can compute mean scores

$$\bar{z}_i = \frac{\sum_j z_j x_{ij}}{T_i} = \frac{U_i}{T_i} \qquad (i = 1, 2, \cdots, c)$$

for each column, where $U_i$ may be called the total score for the $i$th column. From equation (17), section 6.3, noting that $T_x$ is now replaced by $T_i$ in our notation, we have

$$V(\bar{z}_i) = \frac{1}{T_i T}\left[\sum z_j^2 n_j - \frac{(\sum z_j n_j)^2}{T}\right] = \frac{D}{T_i}, \qquad (17)'$$

where the factor $D$ is the same for all columns.

These results provide a test of the hypothesis that there is no difference in mean score from column to column to column. We have

$$X^2 = \frac{\sum T_i(\bar{z}_i - \bar{\bar{z}})^2}{D} = \frac{\sum \frac{U_i^2}{T_i} - \frac{(\sum U_i)^2}{T}}{D}. \qquad (18)$$

The second form is better for computing. The d.f. are $(c - 1)$. These d.f. may in turn be subdivided for more specific comparisons among the columns. In particular, if scores have also been assigned to each column, we may test the regressions of the column means on these scores. General methods have been given by Yates (14).

### 7.2 *Use of the analysis of variance.*

For the reader who is familiar with the analysis of variance, but less so with $\chi^2$ tests in contingency tables, it is worth noting that the analyses based on scores may be performed quite satisfactorily by ordinary analysis of variance methods. The approach can be illustrated from the comparison of the mean scores of the two groups of patients which was made from the data in table 6.

Let us regard the data in table 6 as representing two independent samples of data from frequency distributions in which the variates take only the values 3, 2, 1, 0, −1. The orthodox analysis of variance, using a pooled estimate of error, is given in table 7.

TABLE 7.
COMPARISON OF MEAN SCORES BY ANALYSIS OF VARIANCE

| Source of variation | d.f. | S.S. | M.S. |
|---|---|---|---|
| Between sample means | 1 | 7.7289 | 7.7289 |
| Within samples | 194 | 219.5364 | 1.1316 |

The $F$-value is 6.830 with 1 and 194 d.f., as compared with the $X^2$ value of 6.666, which we may regard as an $F$-value of 6.666 with 1 and $\infty$ d.f. Clearly, the significance probabilities differ by a negligible amount.

For an $r \times c$ contingency table, the general relation between the $F$ and $X^2$ values for comparing the mean scores for the $c$ columns is

$$F = \frac{X^2}{df} \cdot \frac{(T - c)}{(T - X^2)} \qquad (19)$$

where $df$ is the number of degrees of freedom in $X^2$.

This relation may be verified numerically in the present example. Since $X^2 = 6.666$, $df = 1$, $T = 196$, $c = 2$, we obtain 6.830 for the right side of equation (19), in agreement with the actual $F$-value. In most tables, $T$ will be much larger than either $c$ or $X^2$, so that $F$ and $X^2$ will be practically equal. Further, since $F$ usually has a substantial number of d.f. in the denominator, the significance probabilities obtained from $F$ and $X^2$ will usually agree closely.

The $F$-test has one aesthetic objection. As Yates showed, the $X^2$ test gives *exactly* the same result, as it should, for the regression of the $p_i$ on the row scores as for the comparison of the mean scores in the two columns. The $F$-test can also be performed both ways. We can assign a score 0 to one column and a score 1 to the other, and make an $F$-test of the regression of the *row* means on the row scores. This $F$-value, however, is not identical with the $F$-value obtained in table 7. However, the significance levels differ only by trivial amounts except when $T$ is small.

## 8. THE COMBINATION OF 2 x 2 CONTINGENCY TABLES

Suppose that we are comparing the frequencies of some occurrence in two independent samples, and that the whole procedure is repeated a number of times under somewhat differing environmental conditions. The data then consist of a series of $2 \times 2$ tables, and the problem is to make a combined test of significance of the difference in occurrence rates in the two samples. The data obtained in comparing the effectiveness of two agents in dosage-mortality experiments are a typical example, in which the repetitions of the experiment are made under a series of different dosage levels. My concern here, however, is with cases where there is no variate corresponding to dosage level, and no well-established theory of how to combine the data.

One method that is sometimes usǫd is to combine all the data into a single $2 \times 2$ table, for which $X^2$ is computed in the usual way. This procedure is legitimate only if the probability $p$ of an occurrence (on the null hypothesis) can be assumed to be the same in all the individual $2 \times 2$ tables. Consequently, if $p$ obviously varies from table to table, or we suspect that it may vary, this procedure should not be used.

Another favorite technique is to compute the usual $X^2$ separately for each table, and add them, using the fact that the sum of $g$ values of $\chi^2$, each with 1 d.f., is distributed as $\chi^2$ with $g$ d.f. This is a poor method. It takes no account of the signs of the differences $(p_1 - p_2)$ in the two samples, and consequently lacks power in detecting a difference that shows up consistently in the same direction in all or most of the individual tables.

An alternative is to compute the $X$ values, and add them, taking account of the signs of the differences. Since $X$ is approximately normally distributed with mean 0 and unit S.D., the sum of $g$ independent $X$ values is approximately normally distributed with mean 0 and S.D. $\sqrt{g}$. Hence the test criterion,

$$\frac{\sum X}{\sqrt{g}}$$

is referred to the standard normal tables.

This method has much to commend it if the total $N$'s of the individual tables do not differ greatly (say by more than a ratio of 2 to 1) and if the $p$'s are all in the range 20-80%. For the following illustrative data I am indebted to Dr. Martha Rogers. The comparison is between mothers of children in the Baltimore schools who had been referred by their teachers as presenting behavior problems, and mothers of a comparable group of control children who had not been so referred. For each mother, it was recorded whether she had suffered any infant losses (e.g. stillbirths) previous to the birth of the child in the study. The comparison is part of a study of possible associations between behavior problems in children and complications of pregnancy of the mother. Since these loss rates increase with later birth orders, and since the samples might not be comparable in birth orders, the data were examined separately, as a precaution, for 3 birth-order classes (see table 8). The two groups of children are referred to as 'Problems' and 'Controls'.

TABLE 8.
DATA ON NUMBER OF MOTHERS WITH PREVIOUS INFANT LOSSES

| Birth Order | | No. of mothers with | | Total | % Loss |
|---|---|---|---|---|---|
| | | Losses | None | | |
| 2 | Problems<br>Controls | 20<br>10 | 82<br>54 | 102<br>64 | 19.6<br>15.6 |
| | | 30 | 136 | 166 = $N_1$ | 18.1 |
| 3-4 | Problems<br>Controls | 26<br>16 | 41<br>30 | 67<br>46 | 38.8<br>34.8 |
| | | 42 | 71 | 113 = $N_2$ | 37.2 |
| 5+ | Problems<br>Controls | 27<br>14 | 22<br>23 | 49<br>37 | 55.1<br>37.8 |
| | | 41 | 45 | 86 = $N_3$ | 47.7 |

Note that the loss rate is higher in the 'Problems' sample in all 3 tables. Since the $N$'s in the separate tables lie within a 2:1 ratio, and the $p$'s are between 18% and 48%, addition of the $X$ values is indicated. The individual $X$ values are, respectively, 0.650, 0.436, 1.587, all being given the same sign since the difference is in the same direction. For this test, the $X$ values are computed without the correction for continuity. Hence the test criterion is the approximate normal deviate

$$\frac{0.650 + 0.436 + 1.587}{\sqrt{3}} = 1.54$$

The $P$ value is just above 0.10. Addition of the $X^2$ values gives 3.131, with 3 d.f., corresponding to a $P$ of about 0.38.

If the $N$'s and $p$'s do not satisfy the conditions mentioned, addition of the $X$'s tends to lose power. Tables that have very small $N$'s cannot be expected to be of much use in detecting a difference, yet they receive the same weight as tables with large $N$'s. Where differences in the $N$'s are extreme, we need some method of weighting the results from the individual tables. Further, if the $p$'s vary from say 0 to 50%, the difference that we are trying to detect, if present, is unlikely to be constant at all levels of $p$. A large amount of experience suggests that the difference is more likely to be constant on the probit or logit scale. As a further complication, the term $pq$ in the variance of a difference will change from one 2 × 2 table to another.

Perhaps the best method for a combined analysis is to transform the data to a probit or logit scale. Examples of this type of analysis are given by Winsor (17) and Dyke and Patterson (22): it is recommended if the data are extensive enough to warrant a searching examination. As an alternative, the following test of significance in the original scale will, I believe, be satisfactory under a wide range of variations in the N's and $p$'s from table to table.

For the $i$th 2 × 2 table, let

$$\begin{aligned}
n_{i1}\,,\ n_{i2} &= \text{sample sizes}\\
p_{i1}\,,\ p_{i2} &= \text{observed proportions in the two samples}\\
\hat{p}_i &= \text{combined proportion from the margins}\\
d_i &= p_{i1} - p_{i2} = \text{observed difference in proportions}\\
w_i &= \frac{n_{i1}n_{i2}}{n_{i1} + n_{i2}}: \qquad w = \sum w_i
\end{aligned}$$

Then we compute the weighted mean difference

$$\bar{d} = \frac{\sum w_i d_i}{w}$$

TABLE 9.
MORTALITY BY SEX OF DONOR AND SEVERITY OF DISEASE

| Degree of disease | Sex of donor | Number of | | Total | % deaths |
|---|---|---|---|---|---|
| | | Deaths | Surv. | | |
| None | M | 2 | 21 | 23 | 8.7 |
| | F | 0 | 10 | 10 | 0.0 |
| | Total | 2 | 23 | $33 = N_1$ | $6.1 = \hat{p}_1$ |
| Mild | M | 2 | 40 | 42 | 4.8 |
| | F | 0 | 18 | 18 | 0.0 |
| | Total | 2 | 58 | $60 = N_2$ | $3.3 = \hat{p}_2$ |
| Moderate | M | 6 | 33 | 39 | 15.4 |
| | F | 0 | 10 | 10 | 0.0 |
| | Total | 6 | 43 | $49 = N_3$ | $12.2 = \hat{p}_3$ |
| Severe | M | 17 | 16 | 33 | 51.5 |
| | F | 0 | 4 | 4 | 0.0 |
| | Total | 17 | 20 | $37 = N_4$ | $45.9 = \hat{p}_4$ |

This has a standard error

$$\text{S.E.} = \frac{\sqrt{\sum w_i \hat{p}_i \hat{q}_i}}{w}$$

The test criterion is

$$\frac{\bar{d}}{\text{S.E.}} = \frac{\sum w_i d_i}{\sqrt{\sum w_i \hat{p}_i \hat{q}_i}}$$

This is referred to the tables of the normal distribution. As explained in the Appendix, which gives supporting reasons for this criterion, the criterion was constructed so that it would be powerful if the alternative hypothesis implies a constant difference on either the probit or the logit scale. The form of the criterion is not one that I would have selected intuitively, and the reader who feels the same way should consult the Appendix.

The test will be illustrated from data published by Diamond et al. (23). Erythroblastosis foetalis is a disease of newborn infants, sometimes fatal, caused by the presence in the blood of an *Rh*+ baby, of anti-*Rh* antibody transmitted by his *Rh*− mother. One form of treatment is an "exchange transfusion," in which as much as possible of the infant's blood is replaced by a donor's blood that is free of anti-*Rh* antibody. In 179 cases in which this treatment was used in a Boston hospital, the rather startling finding was made that there were no infant deaths out of 42 cases in which a female donor was used, but 27 infant deaths out of 137 cases in which a male donor was used. Since there seemed no *a priori* reason why there should be less hazard with female donors, a statistical investigation was made and reported in the reference.

One possibility was that male donors had been used primarily in the more severe cases. Consequently, the data were classified according to the stage of disease at birth, giving the four 2 × 2 tables shown in table 9.

The $N$'s do not vary greatly, but the $p$'s range from 3 to 46%. The combined test of significance is made from the supplementary data in table 10.

TABLE 10.
COMPUTATIONS FOR THE COMBINED TEST

| Stage | $d_i$ | $\hat{p}_i$ | $\hat{p}_i\hat{q}_i$ | $w_i = \dfrac{n_{i1}n_{i2}}{n_{i1} + n_{i2}}$ |
|---|---|---|---|---|
| None | + 8.7 | 6.1 | 573 | 7.0 |
| Mild | + 4.8 | 3.3 | 319 | 12.6 |
| Moderate | +15.4 | 12.2 | 1071 | 8.0 |
| Severe | +51.5 | 45.9 | 2483 | 3.6 |

$$\frac{\bar{d}}{\text{S.E.}} = \frac{\sum w_i d_i}{\sqrt{\sum w_i \hat{p}_i \hat{q}_i}} = \frac{429.88}{\sqrt{25{,}537}} = 2.69$$

The significance probability is 0.0072. In data of this kind, a nonsignificant result would indicate that the surprising phenomenon can be explained by differences in the selection of cases. A significant result must be interpreted with caution, since conditions were not necessarily comparable for male and female donors even within cases of a given degree of severity. However, a significant result does encourage further study, e.g. by examining results in other hospitals.

The expectations in table 9 are so low that one might well doubt the validity of a normal approximation. At least, I did. However, the exact distribution of $\sum w_i d_i$ can be worked out, with some labor, by writing down the probabilities of all possible configurations for each $2 \times 2$ table. The total numbers of configurations are 3, 3, 7 and 5, respectively, for the four tables, so that the total number of possible samples is 315, of which some have negligible probabilities. The value of the test criterion and the probability was worked out for each sample. The exact significance level was found to be 0.0095, as against the normal approximation of 0.0072. The degree of agreement is reassuring, considering the extreme smallness of the expectations. I am indebted to Mrs. Leah Barron for performing the computations.

## 9. THE EFFECT OF EXTRANEOUS VARIATION

Sometimes count data are subject to extraneous variation of a non-Poisson or non-binomial type, especially when the data are samples from an extensive population about which inferences are to be made. In these circumstances, an answer to the question that is really of interest frequently requires a test of significance which takes account of the extraneous variation. For this purpose, $\chi^2$ tests are inappropriate, because they allow only for Poisson or binomial deviations from the null hypothesis. A few simple examples will be considered.

Suppose that we have a sample of data of the Poisson type and that the sample can be divided on some rational basis into $g$ groups. If we wish to examine whether the mean varies from group to group, it was pointed out (section 3.3) that the variance $X^2$ can be divided into components as follows.

| | d.f. |
|---|---|
| Between groups | $(g - 1)$ |
| Within groups | $(N - g)$ |

(The latter component may be divided into a contribution from each group.) If the "Within groups" $X^2$ is statistically significant, this is a warning that the "Between groups" $\chi^2$ test may be invalid. Further thought about the situation usually leads us to conclude that we do not want to declare that there are real differences between the true group means unless the observed group means differ by more than can be accounted for from the variation within groups. A more appropriate test is either an $F$-test of the "Between-groups" mean square against the "Within groups" mean square, or; if the observations are small and highly variable, an $F$-test in the square root scale. In fact, if there is reason to expect that the within-group variation will contain a non-Poisson component, use of an $F$-test without troubling to compute $X^2$ is a safer procedure. The same issue may arise in testing whether $p$ varies from group to group within a $2 \times N$ contingency table, where the $\chi^2$ test may have to be replaced by an $F$-test based on the original $p_i$ or on the equivalent angles. If the $n_i$ vary substantially from one proportion to another, the problem of obtaining the most efficient type of $F$-test is discussed in (24).

Another example occurs in the combination of $2 \times 2$ contingency tables discussed in section 8. Suppose that the percentages of success under two procedures are compared repeatedly over a variety of conditions that are planned as a sample of some population of conditions about which we wish to draw conclusions. In this case the mean difference $\bar{d} = \bar{p}_1 - \bar{p}_2$ should be tested against the interaction with replications, instead of by the tests that were presented in section 8.

A thorough discussion of when not to use $\chi^2$ tests in problems of this kind, and of the best alternatives, would be lengthy. It is hoped that the few words of warning given in this section will encourage critical thinking about the suitability of a $\chi^2$ test.

## 10. CONCLUDING REMARKS

In presenting the examples in this paper, I have not attempted to give hard and fast rules as to when the supplementary tests should be used. The most useful principle is to think in advance about the way in which the data seem likely to depart from the null hypothesis. This often leads to the selection of a single test (e.g. the variance test) as the only one that appears appropriate. In other situations we may apply both the variance test (or some breakdown of it) and the goodness of fit test, on the grounds that the latter may reveal some types of departure that would not be discovered by the variance test.

When several tests are applied simultaneously to the same data, the chance that at least one of them will be significant is greater, and sometimes much greater, than the presumed 5% probability. This danger of misleading ourselves about the significance level is now widely recognized, and methods for avoiding it have been produced in some of the simpler problems. Although such methods need further development for the applications discussed in this paper, I believe that an awareness of the problem helps to prevent at least the worst distortions of the significance level.

In conclusion, I wish to thank Dr. Paul Meier for some useful suggestions.

## APPENDIX

The purpose of this Appendix is to give some justification for the method proposed in section 8 for making a combined test of significance of the difference between two proportions in a group of independent $2 \times 2$ tables. In the $i$th table, the observed proportions $p_{i1}$, $p_{i2}$ are based on samples of size $n_{i1}$, $n_{i2}$, respectively, and the observed difference is $d_i = p_{i1} - p_{i2}$. The test criterion proposed is

$$\frac{\sum w_i d_i}{\sqrt{\sum w_i \hat{p}_i \hat{q}_i}}, \qquad \text{where } w_i = \frac{n_{i1} n_{i2}}{n_{i1} + n_{i2}}$$

and

$$\hat{p}_i = \frac{n_{i1} p_{i1} + n_{i2} p_{i2}}{n_{i1} + n_{i2}}$$

is the overall proportion from the margins of the $i$th table.

On the null hypothesis,

$$E(d_i) = 0; \qquad V(d_i) \doteq \hat{p}_i \hat{q}_i \left\{ \frac{1}{n_{i1}} + \frac{1}{n_{i2}} \right\} = \frac{\hat{p}_i \hat{q}_i}{w_i},$$

so that the test criterion is approximately normally distributed with mean 0 and s.d. 1. The problem is to show that the test criterion is sensitive in detecting departures from the null hypothesis.

The principal fact to be taken into account is that if the true $p_i$ vary over a wide range, the true difference $\delta_i$ is unlikely to be constant on the observed (proportions) scale, but more likely to be constant on a probit or a logit scale. Suppose, for the moment, that we know the values of the $\delta_i$ on the alternative hypothesis, and consider the more general test criterion

$$Y = \frac{\sum W_i d_i}{\sqrt{\sum W_i^2 \hat{p}_i \hat{q}_i / w_i}}$$

where the $W_i$ are any set of weights. This criterion reduces to the recommended criterion if $W_i = w_i$. It reduces to the sum of the individual $X_i$, divided by the square root of their number, if we put $W_i = 1/\sqrt{\hat{p}_i \hat{q}_i / w_i}$, and can ignore the contribution to the variance of $Y$ arising from variation in the $W_i$. (This restriction is needed because we wish to treat the $W_i$ as fixed weights).

On the null hypothesis, $Y$ is approximately normally distributed with mean 0 and s.d. 1. On the alternative hypothesis,

$$E(\sum W_i d_i) = \sum W_i \delta_i$$

$$V(\sum W_i d_i) = \sum W_i^2 \left\{ \frac{p'_{i1} q'_{i1}}{n_{i1}} + \frac{p'_{i2} p'_{i2}}{n_{i2}} \right\}$$

where $'$ denotes a true proportion. Hence, if we can ignore the contribution to the variance of $Y$ arising from its denominator (as we do in using $\chi^2$ in a 2 × 2 table) we have

$$\frac{E(Y)}{\sigma(Y)} = \frac{\sum W_i \delta_i}{\sqrt{\sum W_i^2 \left\{\frac{p'_{i1}q'_{i1}}{n_{i1}} + \frac{p'_{i2}q'_{i2}}{n_{i2}}\right\}}}.$$

We will maximize the power of $Y$, on the alternative hypothesis, if we choose the $W_i$ so as to maximize the ratio $E(Y)/\sigma(Y)$. By ordinary calculus methods, we find that we must have

$$W_i \propto \frac{\delta_i}{\frac{p'_{i1}q'_{i1}}{n_{i1}} + \frac{p'_{i2}q'_{i2}}{n_{i2}}}$$

Now if the $\delta_i$ are fairly small, so that $p'_{i1}$ and $p'_{i2}$ are not too far apart, the products $p'_{i1}q'_{i1}$ and $p'_{i2}q'_{i2}$ will both be approximately equal to $p'_iq'_i$ , where $p'_i$ is the mean of $p'_{i1}$ and $p'_{i2}$ . This gives, as an approximation to the best weights,

$$W_i \propto \frac{\delta_i}{p'_iq'_i\left(\frac{1}{n_{i1}} + \frac{1}{n_{i2}}\right)} = \frac{\delta_i w_i}{p'_iq'_i}$$

This result is not practically useful, since the $\delta_i$ are unknown. It happens, however, that if the true difference is constant on either the probit or the logit scale, the quantity $\delta_i/p'_iq'_i$ is close to constant over practically the whole range of $p'_i$ from 0 to 100%.

In table 11, which illustrates this result, $p'_{i1}$ has been given a range of values from 0 to 99%. To represent a constant effect on the probit scale, it is assumed that the true probit for the second sample is always 0.5 probit units higher than the probit of $p'_{i1}$ . For the logit scale, the true logit for the second sample is taken as 0.8 logit units higher than the logit of $p'_{i1}$ .

The sizes of the effects are roughly equal on the two scales for $p'_{i1} = 50\%$, where the alternative hypothesis gives about 69% successes, so that $\delta_i$ is about 19. For both the probit and logit cases, $\delta_i$ varies widely from this value as $p'_{i1}$ varies, but the variation in $\delta_i$ is closely matched by that in $p'_iq'_i$ so that the ratio is practically constant for the logit case and varies only slightly for the probit case. If the $\delta_i$ are made smaller, the constancy of the ratio is improved; if larger, the ratio varies more.

The conclusion from this argument is that if effects are constant, for varying levels of $p'_{i1}$ , on either the probit or logit scale, the choice of $W_i = w_i$ gives a test criterion that should be close to the optimum in power.

TABLE 11.
CHECK ON CONSTANCY OF $\delta_i/p'_iq'_i$

| $p'_{i1}(\%)$ | Constant probit effect | | | Constant logit effect | | |
|---|---|---|---|---|---|---|
| | $\delta_i$ | $p'_iq'_i$ | $10^4\delta_i/p'_iq'_i$ | $\delta_i$ | $p'_iq'_i$ | $10^4\delta_i/p'_iq'_i$ |
| 1 | 2.38 | 214 | 111 | 1.20 | 157 | 76 |
| 5 | 7.61 | 803 | 95 | 5.49 | 714 | 77 |
| 10 | 11.71 | 1334 | 88 | 9.83 | 1269 | 77 |
| 30 | 19.04 | 2390 | 80 | 18.82 | 2388 | 79 |
| 50 | 19.15 | 2408 | 80 | 19.00 | 2410 | 79 |
| 70 | 14.71 | 1751 | 84 | 13.85 | 1775 | 78 |
| 90 | 6.26 | 640 | 98 | 5.24 | 683 | 77 |
| 95 | 3.40 | 319 | 107 | 2.69 | 352 | 76 |
| 99 | 0.76 | 62 | 122 | 0.55 | 72 | 76 |

## REFERENCES

(1) D. Lewis and C. J. Burke, "The use and misuse of chi-square test," *Psych. Bull.*, *46* (1949), 433–498.

(2) P. V. Sukhatme, "On the distribution of $\chi^2$ in small samples of the Poisson series," *Jour. Roy. Stat. Soc. Suppl.*, *5* (1938), 75–79.

(3) J. Neyman and E. S. Pearson, "Further notes on the $\chi^2$ distribution," *Biometrika*, *22* (1931), 298–305.

(4) W. G. Cochran, "The $\chi^2$ correction for continuity," *Iowa State Coll. Jour. Sci.*, *16* (1942), 421–436.

(5) W. G. Cochran, "The $\chi^2$ distribution for the binomial and Poisson series, with small expectations," *Annals of Eugenics*, *7* (1936), 207–217.

(6) E. Fix, "Tables of the Noncentral $\chi^2$," *Univ. California Publ. Stat.*, *1* (1949), 15–19.

(7) W. G. Cochran, "The $\chi^2$ test of goodness of fit," *Annals of Math. Stat.*, *23* (1952), 315–345.

(8) D. Mainland, "Statistical methods in medical research," *Canadian Jour. Res.*, *E*, *26* (1948), 1–166.

(9) G. H. Freeman and J. H. Halton, "Note on an exact treatment of contingency, goodness of fit and other problems of significance," *Biometrika*, *38* (1951), 141–149.

(10) J. B. S. Haldane, The mean and variance of $\chi^2$ when used as a test of homogeneity, when expectations are small," *Biometrika*, *31* (1939), 346–355.

(11) H. O. Lancaster, "The derivation and partition of $\chi^2$ in certain discrete distributions," *Biometrika*, *36* (1949), 117–129.

(12) W. G. Cochran, "Statistical analysis of field counts of diseased plants," *Jour. Roy. Stat. Soc. Suppl.*, *3* (1936), 49–67.

(13) D. P. Murphy, *Heredity in uterine cancer*, Harvard University Press, 1952.

(14) F. Yates, "The analysis of contingency tables with groupings based on quantitative characters," *Biometrika*, *35* (1948), 176–181.

(15) J. O. Irwin, "A note on the subdivision of $\chi^2$ into components," *Biometrika*, *36* (1949), 130–134.

(16) A. W. Kimball, "Short-cut formulas for the exact partition of $\chi^2$ in contingency tables," *Biometrics*, *10* (1954), 452–458.

(17) C. P. Winsor, "Factorial analysis of a multiple dichotomy," *Human Biology*, *20* (1948), 195–204.

(18) F. J. Anscombe, "Sampling theory of the negative binomial and logarithmic distributions," *Biometrika*, *37* (1950), 358–382.

(19) C. I. Bliss, "Fitting the negative binomial distribution to biological data," *Biometrics*, *9* (1953), 176–196.

(20) J. Berkson, "A note on the chi-square test, the Poisson and the binomial," *Jour. Amer. Stat. Ass.*, *35* (1940), 362–367.

(21) J. Berkson, "Some difficulties of interpretation encountered in the application of the chi-square test," *Jour. Amer. Stat. Ass.*, *33* (1938), 526–536.

(22) G. V. Dyke and H. D. Patterson, "Analysis of factorial arrangements when the data are proportions," *Biometrics*, *8* (1952), 1–12.

(23) Fred H. Allen, Jr., Louis K. Diamond and Joseph B. Watrous, Jr., "Erythroblastosis fetalis. V. The value of blood from female donors for exchange transfusion," *The New England Jour. of Med.*, *241* (1949), 799–806.

(24) W. G. Cochran, "Analysis of variance for percentages based on unequal numbers," *Jour. Amer. Stat. Assoc.*, *38* (1943), 287–301.