The χ^2 Test of Goodness of Fit
Author(s): William G. Cochran
Source: *The Annals of Mathematical Statistics*, Vol. 23, No. 3 (Sep., 1952), pp. 315-345
Published by: Institute of Mathematical Statistics
Stable URL: http://www.jstor.org/stable/2236678
Accessed: 10/03/2014 13:46

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at
http://www.jstor.org/page/info/about/policies/terms.jsp

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve and extend access to *The Annals of Mathematical Statistics*.

http://www.jstor.org

This content downloaded from 128.97.55.209 on Mon, 10 Mar 2014 13:46:32 PM
All use subject to JSTOR Terms and Conditions

THE χ^2 TEST OF GOODNESS OF FIT[1]

By William G. Cochran

Johns Hopkins University

1. Summary. This paper contains an expository discussion of the chi square test of goodness of fit, intended for the student and user of statistical theory rather than for the expert. Part I describes the historical development of the distribution theory on which the test rests. Research bearing on the practical application of the test—in particular on the minimum expected number per class and the construction of classes—is discussed in Part II. Some varied opinions about the extent to which the test actually is useful to the scientist are presented in Part III. Part IV outlines a number of tests that have been proposed as substitutes for the chi square test (the ω^2 test, the smooth test, the likelihood ratio test) and Part V a number of supplementary tests (the run test, tests based on low moments, subdivision of chi square into components).

2. Introduction. In the standard applications of the test, the n observations in a random sample from a population are classified into k mutually exclusive classes. There is some theory or null hypothesis which gives the probability p_i that an observation falls into the ith class ($i = 1, 2, \cdots, k$). Sometimes the p_i are completely specified by the theory as known numbers, and sometimes they are less completely specified as known functions of one or more parameters $\alpha_1, \alpha_2, \cdots$ whose actual values are unknown. The quantities $m_i = np_i$ are called the *expected* numbers, where

$$\sum_{i=1}^{k} p_i = 1, \qquad \sum_{i=1}^{k} m_i = n.$$

The starting point in the theory is the joint frequency distribution of the *observed* numbers x_i falling in the respective classes. If the theory is correct, these observed numbers follow a multinomial distribution with the p_i as probabilities. The joint distribution of the x_i is therefore specified by the probabilities

$$\frac{n!}{x_1!\, x_2! \cdots x_k!}\, p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}. \tag{1}$$

As a test criterion for the null hypothesis that the theory is correct, Karl Pearson [1] proposed the quantity

$$X^2 = \sum_{i=1}^{k} \frac{(x_i - m_i)^2}{m_i} = \sum_{i=1}^{k} \frac{x_i^2}{m_i} - n. \tag{2}$$

---

[1] Department of Biostatistics Paper No. 282.

Editor's Note: This paper was presented to the Boston meeting of the Institute of Mathematical Statistics, December 28, 1951, and is published in the *Annals* by invitation of the Institute Committee on Special Invited Papers.

This content downloaded from 128.97.55.209 on Mon, 10 Mar 2014 13:46:32 PM
All use subject to JSTOR Terms and Conditions

Pearson did not mention any particular alternative hypothesis. The test has usually been regarded as applicable to situations in which the alternative hypothesis is described only in rather vague and general terms.

As with any test of a hypothesis, certain properties of the test must be worked out before it is ready for practical application. We need to know the frequency distribution of the test criterion when the null hypothesis is correct, in order that tables of significant values can be constructed. As much as possible should also be known about the performance of the test when the null hypothesis does not hold. Practically all the results in the literature deal only with the limiting distribution of X^2 as $n \to \infty$, the p_i remaining fixed. When the null hypothesis holds, this limiting distribution is the χ^2 distribution,

$$\frac{1}{2^{\nu/2}\left(\frac{\nu}{2}-1\right)!}(\chi^2)^{(\nu/2)-1}e^{-\frac{1}{2}\chi^2}\,d\chi^2, \tag{3}$$

where ν is the number of degrees of freedom in χ^2. This distribution is also known to be that followed by the quantity

$$y_1^2 + y_2^2 + \cdots + y_\nu^2,$$

where the y_i are normally and independently distributed with zero means and unit variances.

To avoid confusion, the symbol X^2 will be used for the quantity in equation (2) which is calculated from the data when a chi square test is performed. The symbol χ^2 will refer to any random variate which follows the tabular chi square distribution given in (3).

Part I. Historical Development of the Test

3. Karl Pearson's 1900 paper. This remarkable paper is one of the foundations of modern statistics. Its style has always impressed me as unusual for a pioneering paper. Pearson writes with the air of a man who knows exactly what he is doing. The exposition, although clear, is slightly hurried and brusque, as if the reader will not wish to be troubled by elaborate details of a problem that is routine and straightforward. One misses any discussion of how Pearson came to choose the X^2 test criterion, and of when he first came to realize that this criterion would, under certain circumstances, follow the χ^2 distribution.

The paper opens by proving that if a set of ν correlated variates z_i, with zero means, follow a multivariate normal distribution

$$Ce^{-\frac{1}{2}Q}dz_1dz_2\cdots dz_\nu,$$

then the quadratic form Q is distributed as χ^2 with ν degrees of freedom. This proof is accomplished, in about half a page, by a now familiar geometrical argument. Pearson points out that the ellipsoid Q can be "squeezed" into a sphere. A transformation to polar coordinates is made, where χ is the radius of the sphere and $Q = \chi^2$. He then remarks that all the angles introduced in

This content downloaded from 128.97.55.209 on Mon, 10 Mar 2014 13:46:32 PM
All use subject to JSTOR Terms and Conditions

the transformation will integrate out to a constant factor, so that the probability that Q exceeds χ_0^2, say, reduces to

$$P = \frac{\int_{\chi_0}^{\infty} e^{-\frac{1}{2}\chi^2}\chi^{\nu-1}\,d\chi}{\int_0^{\infty} e^{-\frac{1}{2}\chi^2}\chi^{\nu-1}\,d\chi}.$$

This, of course, is the tabular χ^2 distribution, expressed as an integral of χ rather than χ^2. The result is a generalization of the result reached by Helmert in 1876, and also of the result which Student later developed in 1908 as groundwork for the t-distribution.

The next step is to express the probability integral in power series form, this being necessary to construct a table of the probability integral. The paper contains a table giving P to 6 decimal places, for degrees of freedom from 2 to 19, and for various integral values of χ^2.

Pearson now turns to the problem of testing goodness of fit. He deals first with the case in which the expectations m_i are known numbers. The data have been classified as described in Section 2, so that the observations x_i follow a multinomial distribution. Pearson assumes without more ado that the x_i may be taken as normally distributed. It is at this point, therefore, that he is committed to the assumption that the expectations m_i are large in all cells. He assigns to the x_i their correct variances and covariances from the multinomial distribution, that is,

$$\sigma_{ii} = np_i(1 - p_i), \qquad \sigma_{ij} = -np_ip_j \qquad (i \neq j) \tag{4}$$

The remainder of the proof consists in writing down the presumed multivariate normal distribution of the quantities $(x_i - m_i)$. From this comes the pleasing result that the quadratic form Q in the exponent is simply

$$Q = \sum \frac{(x_i - m_i)^2}{m_i} = X^2.$$

This may be shown as follows. Since the x_i are constrained to add to n, we must omit one of them, say x_k, in considering their joint distribution. If the joint frequency function of the first $(k - 1)$ of the x's is $Ce^{-\frac{1}{2}Q}$, it is well known that

$$Q = \sum_{i=1}^{k-1}\sum_{j=1}^{k-i} \sigma^{ij}(x_i - m_i)(x_j - m_j),$$

where σ^{ij} is the inverse of the matrix σ_{ij} given in (4).

Now consider X^2, with $(x_k - m_k)$ replaced by $-\sum_{i=1}^{m-1}(x_i - m_i)$.

$$X^2 = \frac{(x_1 - m_1)^2}{m_1} + \cdots + \frac{(x_{k-1} - m_{k-1})^2}{m_{k-1}} + \frac{\{(x_1 - m_1) + \cdots + (x_{k-1} - m_{k-1})\}^2}{m_k}.$$

This content downloaded from 128.97.55.209 on Mon, 10 Mar 2014 13:46:32 PM
All use subject to JSTOR Terms and Conditions

Hence, if we write

$$X^2 = \sum_{i=1}^{k-1} \sum_{j=1}^{k-1} a_{ij} (x_i - m_i)(x_j - m_j),$$

the matrix a_{ij} is

$$a_{ii} = \frac{1}{m_i} + \frac{1}{m_k}, \qquad a_{ij} = \frac{1}{m_k} \qquad (i \neq j). \tag{5}$$

The remainder of the proof consists in showing that (5) is the inverse of (4). Pearson does this by a rather complicated polar transformation, but the student who has some familiarity with the evaluation of determinants will find it a fairly easy exercise, as Hotelling [2] has pointed out. It may be helpful to write (4) as

$$\sigma_{ii} = m_i \left(1 - \frac{m_i}{n}\right), \qquad \sigma_{ij} = -\frac{m_i m_j}{n} \qquad (i \neq j), \tag{4'}$$

and to invert (5) rather than (4), or to prove that the product of the matrices a_{ij} , σ_{jk} is the unit matrix.

Hence, by the first part of Pearson's paper, we reach the result that in the limit as n becomes large, X^2 follows the χ^2 distribution with $(k - 1)$ degrees of freedom.

An approach which avoids most of the mathematical complexities in Pearson's argument has been pointed out by Fisher [3]. If the observations x_i are regarded as following independent Poisson distributions, their joint frequency function is

$$\prod_{i=1}^{k} \frac{e^{-m_i} m_i^{x_i}}{x_i\,!} = e^{-n} \prod_{i=1}^{k} \frac{m_i^{x_i}}{x_i\,!}, \tag{6}$$

since $\sum m_i = n$.

Under this assumption, their total $T = \sum x_i$ also follows a Poisson distribution, with mean $\sum m_i = n$. The frequency function of T is therefore

$$\frac{e^{-n} n^T}{T!}. \tag{6'}$$

Hence, on dividing (6) by (6') the *conditional* frequency function of the x_i , given that their total T has the value n, is

$$\frac{n!}{x_1!\, x_2! \cdots x_k!} \left(\frac{m_1}{n}\right)^{x_1} \left(\frac{m_2}{n}\right)^{x_2} \cdots \left(\frac{m_k}{n}\right)^{x_k}.$$

This is the same as the basic multinomial (1).

This argument implies that in an investigation of the distribution of X^2 we may start by regarding the x_i as following independent Poisson distributions, subject to the restriction that $\sum x_i = n$.

This content downloaded from 128.97.55.209 on Mon, 10 Mar 2014 13:46:32 PM
All use subject to JSTOR Terms and Conditions

In the limit, as the m_i become large, the quantities

$$y_i = \frac{x_i - m_i}{\sqrt{m_i}}$$

become normally distributed with means zero and unit standard deviations, since the Poisson distribution of x_i has mean m_i and standard deviation $\sqrt{m_i}$. Hence the limiting distribution of X^2 is that of the quantity

$$y_1^2 + y_2^2 + \cdots + y_k^2,$$

where the y_i are independently distributed but are subject to the single linear restriction

$$\sum_{i=1}^{k} y_i \sqrt{m_i} = \sum_{i=1}^{k} (x_i - m_i) \doteq 0.$$

The fact that in the limit X^2 follows the χ^2 distribution with $(k - 1)$ degrees of freedom can now be established by integration or by quoting well known theorems on the analysis of variance. This approach also makes it clear that if further homogeneous linear restrictions are imposed on the variates $(x_i - m_i)$, either by the structure of the data or in the process of fitting, the effect will merely be to reduce the degrees of freedom in χ^2.

Pearson next considers the situation in which the m_i depend on parameters that have to be estimated from the sample. Denoting by m_i' the expectations derived from sample estimates of these parameters, and by m_i the true expectations, he discusses the difference

$$X^2 - X'^2 = \sum_{i=1}^{k} \frac{x_i^2}{m_i} - \sum_{i=1}^{k} \frac{x_i^2}{m_i'}.$$

He suggests that this difference will usually be positive, because we ought to be able to do a better job of fitting when we can adjust the estimates of the parameters to suit the vagaries of the sample. He argues, however, that the difference will be small enough so that if we regard X'^2 as also distributed as χ^2 with $(k - 1)$ degrees of freedom, the error in this approximation will not affect practical decisions.

In this conclusion, which is reached with some sign of hesitation, he may well have been justified for many applications. We now know that the number of degrees of freedom must be reduced in order to give the correct limiting distribution. Perhaps the most common of all uses of the X^2 test is for the 2×2 contingency table. Unfortunately, Pearson's suggestion works rather poorly in this case, since he attributed 3 degrees of freedom to X^2, whereas it should receive only 1. This point caused some confusion and controversy in practical applications, and was not settled for over 20 years.

Finally, the paper contains eight numerical applications of the new technique. In two of these he pokes fun at Sir George Airy and Professor Merriman. They had both published series of observations which they claimed to be good illus-

This content downloaded from 128.97.55.209 on Mon, 10 Mar 2014 13:46:32 PM
All use subject to JSTOR Terms and Conditions

trations of variates that follow the normal distribution. In the absence of any test of goodness of fit, Airy and Merriman could judge this question only by eye inspection. Pearson showed that the significance probability for Airy's data was 0.014, although the data from which Pearson calculated X^2 had already been smoothed by Airy. Merriman fared worse, his probability being $1\frac{1}{2}$ parts in a million. These examples show the weak position in which the scientist was placed when he had to judge goodness or badness of fit in the absence of an objective test of significance.

To summarize, Pearson established the necessary distribution theory for finding significance levels when the null hypothesis provides the exact values of the m_i, except that he did not show that the exact distribution of X^2, which is discontinuous, actually approaches χ^2 as a limiting distribution. A fully rigorous proof may be given by the use of moment-generating functions [8].

4. The distribution of X^2 when the expectations are estimated from the sample. This problem is much more difficult and was not elucidated until the appearance of Fisher's 1924 paper. In the intervening period, a paper by Greenwood and Yule [4] in 1915 illustrates the perplexity which existed among critical users of the test and which led to the "degrees of freedom" battle. The authors were attempting to examine the effects of inoculation against typhoid and cholera. They present a substantial number of 2 × 2 tables containing subjects classified as inoculated or not, and also as to whether they contracted the disease following exposure to it. The following is an example.

Kalain (Cholera)

	Not Attacked	Attacked	Total
Inoculated	1625	5	1630
Not	1022	11	1033
Total	2647	16	2663

$X^2 = 6.08$.

Following Pearson's rule, they assign 3 degrees of freedom to X^2. This gives a P of 0.108, whereas with 1 degree of freedom, P is 0.015. They realised, however, that the hypothesis that inoculation is without effect could be tested alternatively, by calculating the difference $(p_1 - p_2)$ between the percent ill among the inoculated and the non-inoculated. On the null hypothesis, the ratio

$$R = \frac{p_1 - p_2}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}}$$

is approximately a normal deviate with mean zero and unit standard deviation. This test, as they found, gave more statistically significant results than Pear-

This content downloaded from 128.97.55.209 on Mon, 10 Mar 2014 13:46:32 PM
All use subject to JSTOR Terms and Conditions

son's test. The quantity R is exactly equal to X if we use the pooled percent ill, $\bar{p}$, in estimating the two variances in the denominator, so that the "normal deviate" test and the X^2 test should be identical. It is not clear that Greenwood and Yule recognised this in 1915.

Although giving the impression of being somewhat in a quandary as to which test to employ, they content themselves with the decision to adopt Pearson's test, pointing out that it is the more conservative of the two, and adding that the issue deserves further theoretical investigation.

After some controversy, the matter was cleared up in theoretical papers [3], [5] by Fisher in 1922 and 1924, supported by sampling experiments which were published by Yule [6] and Brownlee [7]. Fisher's 1922 paper included a discussion of 2×2 contingency tables, and showed that for this case X^2 is the square of a single quantity which had a limiting normal distribution, and that the X^2 test and the test of $(p_1 - p_2)$ by the normal distribution are identical.

Fisher's 1924 paper is much more general. He points out that the limiting distribution of X^2 depends on the method of estimation. With a poor method of estimation, X^2 may frequently have a large value even if the theory is correct. It is therefore necessary, in a general proof of the distribution of X^2, to state what is to be the method of estimation. At first sight, the natural method would seem to be to choose the unknown parameters so that X^2 is as small as possible. Fisher shows that in the limit in large samples, this method becomes equivalent to the method of maximum likelihood. For his main proof, this result serves as an ingenious lemma, since at one point in the main proof he assumes that estimation is by maximum likelihood, while at another he assumes that it is by minimum X^2.

Although Fisher's main proof is not fully rigorous, it is worthwhile to outline the principal steps, because the proof does reveal the core of the problem, and a rigorous proof requires advanced methods. Fisher starts in the same way as Pearson, by considering

$$X^2 - X'^2 = \sum_{i=1}^{k} \frac{x_i^2}{m_i} - \sum_{i=1}^{k} \frac{x_i^2}{m_i'} = \sum_{i=1}^{k} x_i^2 \left(\frac{1}{m_i} - \frac{1}{m_i'} \right),$$

where m_i is a specified function of a single unknown parameter α, with $m_i = m_i(\alpha)$, $m_i' = m_i(\alpha')$. He expands in a Taylor series about the point α'. The first two terms give

$$X^2 - X'^2 = -\sum \frac{x_i^2}{m_i'^2} \left(\frac{\partial m_i'}{\partial \alpha'} \right) \delta\alpha + \sum x_i^2 \left\{ \frac{2}{m_i'^3} \left(\frac{\partial m_i'}{\partial \alpha'} \right)^2 - \frac{1}{m_i'^2} \frac{\partial^2 m_i'}{\partial \alpha'^2} \right\} \frac{(\delta\alpha)^2}{2!},$$

where $\delta\alpha = (\alpha - \alpha')$. Since the method of estimation consists in choosing α' so that X'^2 is a minimum, we have

$$\sum \frac{x_i^2}{m_i'} \left(\frac{\partial m_i'}{\partial \alpha'} \right) = 0,$$

so that the first term on the right vanishes.

In the second term on the right, Fisher replaces x_i by m_i'. The error intro-

This content downloaded from 128.97.55.209 on Mon, 10 Mar 2014 13:46:32 PM
All use subject to JSTOR Terms and Conditions

duced by this step may be shown to be of the same order as the third term in the Taylor series, which has already been ignored. Hence,

$$X^2 - X'^2 = \sum \left\{ \frac{2}{m_i'} \left(\frac{\partial m_i'}{\partial \alpha'} \right)^2 - \frac{\partial^2 m_i'}{\partial \alpha'^2} \right\} \frac{(\delta\alpha)^2}{2!}.$$

But if the identity $\sum m_i' = n$ is differentiated twice, we find

$$\sum \left(\frac{\partial^2 m_i'}{\partial \alpha'^2} \right) = 0.$$

Hence,

$$X^2 - X'^2 = \sum \left\{ \frac{1}{m_i'} \left(\frac{\partial m_i'}{\partial \alpha'} \right)^2 \right\} (\delta\alpha)^2.$$

If α' is regarded as a maximum likelihood estimate of α, we may use the standard result that the error of estimate $(\alpha' - \alpha)$ has a limiting normal distribution, with mean zero, and variance given by

$$\frac{1}{\sigma_{\alpha'}^2} = \sum \left\{ \frac{1}{m_i} \left(\frac{\partial m_i}{\partial \alpha} \right)^2 \right\} \doteq \sum \left\{ \frac{1}{m_i'} \left(\frac{\partial m_i'}{\partial \alpha'} \right)^2 \right\}.$$

This gives the neat result

$$X^2 - X'^2 = \frac{(\alpha' - \alpha)^2}{\sigma_{\alpha'}^2}.$$

Our object is to find the limiting distribution of X'^2. At this point the facts in our possession are: (i) X^2 is distributed as χ^2 with $(k - 1)$ degrees of freedom (this follows from Pearson's results, since X^2 is calculated from the correct m's); and (ii) $X^2 - X'^2$ is distributed as χ^2 with 1 degree of freedom (from Fisher's argument). These facts are not sufficient to determine the distribution of X'^2. However, Fisher points out that the limiting distributions of X'^2 and $(\alpha' - \alpha)^2$ must be independent, since X'^2 was obtained by minimizing X^2 with respect to α'. Given this additional result, it is easily shown that X'^2 must be distributed as χ^2 with $(k - 2)$ degrees of freedom.

The argument leads to two further results. Any method of estimation that is efficient gives estimates which become, in the limit, identical with the maximum likelihood estimate. Thus the χ^2 distribution, with the appropriate reduction in degrees of freedom, is valid for any efficient method of estimation. The argument also provides the limiting mean value of X'^2 when the estimation is inefficient. An interesting corollary is that the mean value of X'^2 exceeds that of X^2 when the efficiency is less than 50 percent.

Rigorous proofs of the general limiting distribution, when several parameters are being estimated, are scarce in the literature. For the student, one of the best is that given by Cramér [8]. The restrictions under which he proves his result are stated below. He assumes maximum likelihood estimation.

This content downloaded from 128.97.55.209 on Mon, 10 Mar 2014 13:46:32 PM
All use subject to JSTOR Terms and Conditions

THEOREM. *Suppose that the k probabilities $p_i(\alpha_1, \alpha_2, \cdots, \alpha_s)$ are known functions of $s < k$ parameters $\alpha_1, \alpha_2, \cdots, \alpha_s$. For all points of a nondegenerate interval A in the s-dimensional space of the α_j, the p_i satisfy the following conditions:*

(a) $$\sum_{i=1}^{k} p_i(\alpha_1, \cdots, \alpha_s) = 1;$$

(b) $$p_i(\alpha_1, \cdots, \alpha_s) > C^2 > 0 \qquad \text{for all } i;$$

(c) *every p_i has continuous derivatives $\dfrac{\partial p_i}{\partial \alpha_j}$ and $\dfrac{\partial^2 p_i}{\partial \alpha_j \partial \alpha_h}$;*

(d) *the matrix $D = \left(\dfrac{\partial p_i}{\partial \alpha_j}\right)$ is of rank s. Then X^2 is distributed in the limit, as $n \to \infty$, in a χ^2 distribution with $(k - s - 1)$ degrees of freedom.*

5. The limiting power function of the test. The literature does not contain much discussion of the power function of the X^2 test. There has been little demand for this from applications, because the test is most commonly used when we do not have a clear-cut alternative in mind, and are not in a position to make computations of the power.

Suppose that we test the null hypothesis that the expectations are m_i when in fact they are m'_i. If the values of m_i, m'_i and the significance level are kept fixed, then as n increases, it turns out, as would be expected, that the power of the test tends to 1. This has been shown by Neyman [9]. In order to examine the situation in which the power is not close to 1 in large samples, we must somehow make the task continually harder for the test as n becomes larger. This can be accomplished either by making the significance probability decrease steadily as n increases, thus reducing the chance of an error of type I, or by moving the alternative hypothesis steadily closer to the null hypothesis. The second method will be discussed here. Let

$$m'_i - m_i = c_i\sqrt{n}; \qquad \text{that is, } p'_i - p_i = c_i/\sqrt{n},$$

where the quantities c_i remain fixed as n increases.

The nature of the limiting power distribution of X^2 is indicated by the following argument, for which I am indebted to J. W. Tukey. We may write

$$\frac{x_i - m_i}{\sqrt{m_i}} = \frac{(x_i - m'_i)}{\sqrt{m'_i}}\sqrt{\frac{m'_i}{m_i}} + \frac{m'_i - m_i}{\sqrt{m_i}}. \tag{7}$$

Now

$$\sqrt{\frac{m'_i}{m_i}} = \sqrt{1 + \frac{m'_i - m_i}{m_i}} = \sqrt{1 + \frac{c_i}{p_i\sqrt{n}}}.$$

This content downloaded from 128.97.55.209 on Mon, 10 Mar 2014 13:46:32 PM
All use subject to JSTOR Terms and Conditions

This tends to 1 as n becomes large, since c_i and p_i are presumed to remain fixed. If we adopt Fisher's approach to the distribution theory (Section 3), the quantities

$$\frac{x_i - m'_i}{\sqrt{m'_i}}$$

tend to become normally and independently distributed with means zero and unit standard deviations, as n becomes large. Consequently, so do the quantities

$$y_i = \frac{(x_i - m'_i)}{\sqrt{m'_i}} \sqrt{\frac{m'_i}{m_i}}.$$

Finally, by equation (7),

$$X^2 = \sum_{i=1}^{k} \frac{(x_i - m_i)^2}{m_i} = \sum_{i=1}^{k} \left\{ y_i + \frac{m'_i - m_i}{\sqrt{m_i}} \right\}^2 = \sum_{i=1}^{k} (y_i + a_i)^2,$$

where the y_i are subject to the linear restriction

$$\sum_{i=1}^{k} y_i \sqrt{m_i} = \sum_{i=1}^{k} (x_i - m'_i) = 0.$$

Thus, in the limit, X^2 is distributed as a sum of squares of variates $(y_i + a_i)$ independently and normally distributed with unit variances, but where the means a_i are not all zero. The variates are subject to one linear restriction when the m_i are completely specified.

This type of distribution is known as a noncentral χ^2. It depends on two parameters—the degrees of freedom, in this case $(k - 1)$, and a parameter of non-centrality $(a_1^2 + a_2^2 + \cdots + a_k^2)$, which has the value

$$\sum_{i=1}^{k} \frac{(m'_i - m_i)^2}{m_i} = \sum_{i=1}^{k} \frac{c_i^2 n}{p_i n} = \sum_{i=1}^{k} \frac{c_i^2}{p_i}.$$

Tables of the noncentral distribution have been provided by Fix [10] and approximations studied by Patnaik [11].

When the m_i have to be estimated from the data, the limiting noncentral χ^2 distribution still holds, with a reduced number of degrees of freedom. A rigorous proof is obtained from Wald's derivation of the limiting distribution of the likelihood ratio test criterion [12], which becomes equivalent to X^2 in large samples.

6. Conditional X^2 tests. As has been mentioned, additional homogeneous linear restrictions imposed on the deviations $(x_i - m_i)$ have the effect of reducing the number of degrees of freedom attributed to χ^2 in the limiting distribution of X^2. These restrictions may arise in the process of fitting, or by the nature of the data. They may also be deliberately imposed by the statistician in the device known as a conditional test. This device is illustrated by the 2×2 contingency table, in which it has created some stimulating discussion [13].

This content downloaded from 128.97.55.209 on Mon, 10 Mar 2014 13:46:32 PM
All use subject to JSTOR Terms and Conditions

The data are classified according to two different axes, A and B.

	B_1	B_2	Totals
A_1	x_{11}	x_{12}	r_1
A_2	x_{21}	x_{22}	r_2
Totals.....	c_1	c_2	n

Data of this kind occur in at least three distinct experimental situations.

(*i*) We select a random sample of n from some population and classify every observation into one of the four cells. The symbol x_{ij} denotes the observed number falling in class A_iB_j, while p_{ij} denotes the corresponding probability of falling in this class, where the sum of the four p's is unity. The null hypothesis that the two classifications are *independent* amounts to the relation

$$p_{11}/p_{12} = p_{21}/p_{22}. \tag{8}$$

The joint probability of this group of observations is the usual multinomial

$$\frac{n!}{x_{11}!\,x_{12}!\,x_{21}!\,x_{22}!}\,p_{11}^{x_{11}}\,p_{12}^{x_{12}}\,p_{21}^{x_{21}}\,p_{22}^{x_{22}}. \tag{9}$$

Only two of the p_{ij} need to be estimated from the data, because of equation (8) and the fact that the p_{ij} add to 1. Thus X^2 has $(4 - 2 - 1)$ or 1 degree of freedom.

(*ii*) We take a random sample of size r_1 from a population denoted by A_1, and an *independent* random sample of size r_2 from another population denoted by A_2. The null hypothesis states that the probability p of an observation falling in B_1 is the same in both populations A_1 and A_2. Given the null hypothesis, the probability of the sample is the product of the two binomials

$$\left\{\frac{r_1!}{x_{11}!\,x_{12}!}\,p^{x_{11}}q^{x_{12}}\right\}\left\{\frac{r_2!}{x_{21}!\,x_{22}!}\,p^{x_{21}}q^{x_{22}}\right\}. \tag{10}$$

This is not the same as the multinomial (9). Given data of type (*i*), however, let us arbitrarily impose the restriction that in repeated sampling we will consider only those tables which have the same marginal totals r_1, r_2 as our data. Then (9) must be replaced by the conditional distribution of the x_{ij}, given r_1 and r_2. This conditional distribution is easily seen to be the same as (10). For, starting with (9), the distribution of r_1 (and hence r_2) is the binomial

$$\frac{n!}{r_1!\,r_2!}\,(p_{11} + p_{12})^{r_1}(p_{21} + p_{22})^{r_2}. \tag{11}$$

This content downloaded from 128.97.55.209 on Mon, 10 Mar 2014 13:46:32 PM
All use subject to JSTOR Terms and Conditions

To obtain the conditional distribution, we divide (9) by (11). The quotient reduces to (10) if we note that from (8),

$$\frac{p_{11}}{p_{11} + p_{12}} = \frac{p_{21}}{p_{21} + p_{22}} = p \quad \text{(say)}.$$

(*iii*) A third case is obtained if *both* sets of marginal totals are regarded as fixed in repeated sampling. Fisher's tea-tasting experiment [14] is an example. The A classification tells whether the milk or the tea was added first, and the B classification tells whether the lady guessed that the milk or the tea was added first. In Fisher's original experiment, he recommended that the lady be informed how many cups were of each kind, and pointed out that she would presumably match her guesses to those two numbers. Thus in repeated trials it is natural to regard both sets of margins as fixed.

The appropriate basic distribution of the x_{ij} is the conditional distribution which develops from (10) if we keep c_1 and c_2 fixed. This is found to be

$$\frac{r_1!\, r_2!\, c_1!\, c_2!}{n!\, x_{11}!\, x_{12}!\, x_{21}!\, x_{22}!}. \tag{12}$$

Case (*i*) has 2 unknown parameters and 1 linear restriction on the x_{ij} ; case (*ii*) has 1 unknown parameter and 2 restrictions, while case (*iii*) has no unknown parameters and 3 restrictions.

Is the same X^2 test to be used for all cases? In large samples there is no conflict, because X^2 has the same limiting distribution however the linear restrictions arise. This is not so in small samples, where the distribution of X^2 differs in the three cases. Fisher [15] recommends that the distribution of X^2 obtained in case (*iii*) be taken as the exact small-sample distribution for all three types of data. Questions have been raised about this recommendation.

Originally, part of the objection came perhaps from a feeling that there is something improper in keeping the marginal totals fixed in cases (*i*) and (*ii*), because if we actually drew repeated samples of the same size by the same methods, the margins would not all remain fixed. For a rational appraisal, however, the only relevant factors are the effects of the marginal restrictions on the significance probabilities (or type I errors) and on the power (or type II errors). As regards type I errors, Fisher's recommendation has the great advantage that in case (*iii*) the significance probabilities can be computed exactly, whereas in cases (*i*) and (*ii*) the distribution of X^2 involves nuisance parameters, so that "exact" probabilities are not available.

The issue thus reduces to the question whether any loss of power occurs if the case (*iii*) test is employed with the first two types of data. For case (*ii*) data, Barnard [13] proposed a different test which in some circumstances appeared to give a small increase in power. More recently K. D. Tocher [16], has proved the remarkable result that a modification of Fisher's test is the most powerful, in the sense of Neyman and Pearson, for one-tailed tests with any

This content downloaded from 128.97.55.209 on Mon, 10 Mar 2014 13:46:32 PM
All use subject to JSTOR Terms and Conditions

of the three types of data. The modification is necessary to make the problem amenable to Neyman and Pearson's techniques.

The modification may be illustrated by the example which Tocher presents.

TABLE 1

Tocher's illustration

Original table

2	5	7
3	2	5
5	7	12

More extreme cases

1	6	7
4	1	5
5	7	12

0	7	7
5	0	5
5	7	12

Given the data on the left, we wish to make a one-tailed test at the 5% level. The two possible sets of data which deviate more from the null hypothesis are shown on the right. In Fisher's exact test, we add the probabilities of the three tables as computed by formula (12). This gives

$$0.26515 + 0.04399 + 0.00126 = 0.31040.$$

This value is regarded as the significance probability.

In Tocher's modification, we also compute the total probability of all more extreme cases, that is,

$$0.04399 + 0.00126 = 0.04525.$$

If these numbers, 0.31040 and 0.04525, are both *below* the stated significance level, 0.05, we reject the hypothesis. If they are both *above* 0.05, we accept. If one is above and one is below, as in the present example, we calculate the ratio

$$\frac{0.05 - 0.04525}{0.26515} = 0.01791.$$

Now draw a random number between 0 and 1. If this number is less than 0.01791, we reject; if greater, we accept.

Although this procedure may appear somewhat startling at first sight, the idea is basically simple. Consider how we can obtain a one-tailed test at the 0.05 level from the 2×2 table in this example. If the null hypothesis is rejected only when the two most extreme cases on the right of Table 1 occur, the significance level is actually 0.04525. The third most extreme case, represented by the data on the left of Table 1, occurs with probability 0.26515. Consequently, by the computation above, we obtain a test at the 0.05 level if we also declare as significant a fraction 0.01791 of the cases in which the data on the left are encountered. Tocher selects this fraction by a table of random numbers. There seems no other logical basis for deciding which particular fraction to select.

This content downloaded from 128.97.55.209 on Mon, 10 Mar 2014 13:46:32 PM
All use subject to JSTOR Terms and Conditions

Part II. Some Aspects of the Practical Use of the Test

7. The minimum expectation. Since χ^2 has been established as the limiting distribution of X^2 in large samples, it is customary to recommend, in applications of the test, that the smallest expected number in any class should be 10 or (with some writers) 5. If this requirement is not met in the original classification, combination of neighboring classes until the rule is satisfied is recommended. This topic has recently been subject to vigorous discussion among the psychologists [17], [18]. The numbers 10 and 5 appear to have been arbitrarily chosen. A few investigations throw some light on the appropriateness of the rule. The approach has been to examine the exact distribution of X^2, when some or all expectations are small, either by mathematical methods or from sampling experiments.

The investigations are scanty and narrow in scope, as is to be expected since work of this type is time-consuming. Thus the recommendations given below may require modification when new evidence becomes available.

To digress for a moment, the problem of investigating the behavior of X^2 when expectations are small is an example of a whole class of problems that are relevant to applied statistics. In applications it is an everyday occurrence to use the results of a body of theory in situations where we know, or strongly suspect, that some of the assumptions in the theory are invalid. Thus the literature contains investigations of the t-distribution when the parent population is nonnormal, and of the performance of linear regression estimates when the regression in the population is actually nonlinear. Fortunately for applications, the results of theory sometimes remain substantially true even when some assumptions fail to hold. This fact tends to make statistics a more confusing subject than pure mathematics, in which a result is usually either right or wrong.

In any problem of this kind, it is important to define what is meant by saying the results remain "substantially true." I stress this point because a reader who becomes interested in a specific problem and tries to summarize the available knowledge may encounter considerable difficulty. Definitions vary from writer to writer and are sometimes entirely subjective, so that the researches may be presented in a form which baffles any attempt to apply a uniform definition. This remark is not intended as a criticism of work on the X^2 problem, where the task of summarizing is comparatively easy. However, I believe that the usefulness of this kind of research would be enhanced by careful attention to the questions: (i) how are we going to measure the disturbance caused by a failure in assumptions, and (ii) when are we going to call this disturbance "serious."

In the present instance, my criterion is to compare the exact P and the P from the χ^2 table, when the null hypothesis is true, in the region in which the tabular P lies between 0.05 and 0.01. This criterion is not ideal, but it does appraise the performance of the tabular approximation in the borderline region between statistical significance and nonsignificance. A disturbance is regarded as unimportant if when the P is 0.05 in the χ^2 table, the exact P lies between 0.04 and 0.06, and if when the tabular P is 0.01, the exact P lies between 0.007 and

This content downloaded from 128.97.55.209 on Mon, 10 Mar 2014 13:46:32 PM
All use subject to JSTOR Terms and Conditions

0.015. These limits are, of course, arbitrary; some would be content with less conservative limits.

The results suggest that four cases need to be considered separately.

(i) *Goodness of fit tests of bell-shaped curves such as the normal distribution.* The distinguishing feature of this case is that usually only one or two expectations at the tails are small, the others being above the conventional limits of 5 or 10. Cochran [19] has shown that there is little disturbance to the 5% level when a *single* expectation is as low as $\frac{1}{2}$.

This is also true for the 1% level if the number of degrees of freedom in X^2 exceeds 6. Two expectations as low as 1 may be allowed with negligible disturbance to the 5% level. Since the discrepancy between an observed and a postulated distribution is often most apparent at the tails, the sensitivity of the X^2 test is likely to be decreased by an overdose of pooling at the tails. Thus considerations of the power of the test urge us to use cells with as small expectations as we dare from distributional considerations. The inflexible use of minimum expectations of 5 or 10 may be harmful.

(ii) 2×2 *contingency tables.* This case is the most thoroughly worked out and can be regarded as solved for practical purposes. Fisher [15] has given the method of obtaining an exact solution, which is not too laborious in samples up to size 30. Tables such as Mainland's [20] give the probability levels of the exact distribution for two samples each of size up to $n = 40$, and Yates' table [21] gives almost exact tests based on X^2 after correction for continuity.

(iii) *Tests in which all expectations may be small.* This case occurs from time to time, for example, in genetical studies in which a Mendelian ratio is being compared over a number of small families. Results by Neyman and Pearson [22], Cochran [23] and Sukhatme [24] suggest tentatively that the tabular χ^2 is tolerably accurate provided that all expectations are at least 2.

With very scanty data, there is one danger—that only a few different values of X^2 are possible, so that the effects of discontinuity become noticeable. For example, consider the 2×4 contingency table with marginal totals shown below. All expectations are exactly 2.

					8
					8
	4	4	4	4	16

If we construct all tables which satisfy these marginal totals, only seven different values of X^2 are found. The exact distribution of X^2 and the χ^2 approximation (with 3 degrees of freedom) are shown in Table 2. The agreement is not good, the tabular P's being fairly consistently too low.

With such a small number of values of X^2, a correction for continuity comes to mind. To apply this for $X^2 = 2$, we read the χ^2 table at $\chi^2 = 1$, this being

This content downloaded from 128.97.55.209 on Mon, 10 Mar 2014 13:46:32 PM
All use subject to JSTOR Terms and Conditions

half way between 2 and 0 (the next largest value of X^2). The corrected P's show a considerable improvement in fit.

In practice, a small table of this kind can be handled by computing the exact distribution of X^2 in cases of doubt about the adequacy of the χ^2 approximation. For more complex contingency tables, a systematic method of computation has been given by Freeman and Halton [25].

TABLE 2

Exact distribution of X^2 for a 2 × 4 contingency table. $P\{X^2 \geqq X_0^2\}$

X_0^2	Exact	χ^2 Table	Corrected
0	1.000	1.000	1.000
2	.899	.572	.801
4	.362	.261	.391
6	.243	.112	.172
8	.064	.046	.072
10	.030	.019	.029
16	.0005	.0011	.004

(iv) *Tests in which all expectations are small and X^2 has many degrees of freedom (say >60).* Examples occur in genetical research. The data are presented in, say, a 2 × 200 contingency table, with all 400 expectations small.

In this case, the exact distributions of X^2 and χ^2 are both approximately normal, since the degrees of freedom are large. However, the two distributions have different variances, and the normal approximation to the exact distribution is sometimes quite different from the normal approximation to χ^2.

This problem has been studied by Haldane [26], [27] who has worked out the exact mean and variance of X^2 for several types of data. His results are given below for the two cases that are perhaps most common.

(a) We have g groups, each containing s individuals, classified into one of two classes. The null hypothesis specifies a *known* constant probability p that any individual falls into the first class. If x_i individuals fall into this class in the ith group, and if we wish to test against the alternative that p varies from class to class, a familiar extension of the X^2 test is to calculate

$$X^2 = \sum_{i=1}^{g} \frac{(x_i - sp)^2}{spq}$$

with g degrees of freedom. Haldane shows that

$$E(X^2) = g,$$

$$V(X^2) = 2g\left(1 + \frac{1 - 6pq}{2spq}\right).$$

This content downloaded from 128.97.55.209 on Mon, 10 Mar 2014 13:46:32 PM
All use subject to JSTOR Terms and Conditions

(b) Same data, but X^2 computed as for a $2 \times g$ contingency table, since p is unknown.

		n_1
		n_2
s	$s \quad \cdots \quad s$	n

Then

$$E(X^2) = \frac{(g-1)n}{n-1},$$

$$V(X^2) = \frac{2(g-1)n^3(n-g)}{(n-1)^2(n-2)(n-3)}\left(1 - \frac{(n-1)}{n_1 n_2}\right).$$

To take an extreme case, suppose $s = 2$, $g = 160$, $n = 320$, $n_1 = 64$, $n_2 = 256$. The mean and variance of X^2 are 159.5 and 159.4 respectively, whereas χ^2 has mean 159 and variance 318, twice as large. The normal approximation to X^2 is satisfactory, but χ^2 gives very poor agreement [23]. In practice, such data may be dealt with by use of the normal approximation to the exact distribution, using Haldane's expressions for the mean and variance.

The question remains: where does case (iii) shade into case (iv)? The available data suggest that case (iii) may apply when X^2 has less than 15 degrees of freedon, while case (iv) may hold if X^2 has more than 60 degrees of freedom. The intervening gap needs investigation.

8. The correction for continuity. The exact distribution of X^2 is always discontinuous. When all expectations are small, the number of distinct values of X^2 may be very limited. In such cases the χ^2 table may give a poor approximation to the exact $P\{X^2 \geqq X_0^2\}$, mainly because the area of a continuous curve is used to approximate the sum of a small number of discrete probabilities. The correction for continuity, introduced by Yates [28], is an attempt to remove this source of error. It amounts to reading the χ^2 table, not at the point X_0^2, but at a point halfway between X_0^2 and the value of X^2 immediately below X_0^2 in the discrete series of values.

In practice, the correction is seldom needed except when X^2 has 1 degree of freedom, as when testing a single binomial ratio or a 2×2 contingency table. In the 2×2 table there are various ways of computing the correction, depending on how one likes to compute X^2. My own preference is to find the difference d between x_i and m_i, which is the same, apart from sign, in all four cells. The absolute value of d is reduced by $\frac{1}{2}$, and X^2 is computed as

$$X^2 = (|d| - \tfrac{1}{2})^2 \sum \frac{1}{m_i}.$$

This content downloaded from 128.97.55.209 on Mon, 10 Mar 2014 13:46:32 PM
All use subject to JSTOR Terms and Conditions

Note that in the 2×2 table it is X that is corrected for continuity, not X^2, since the successive values of d differ by unity.

When X^2 has 1 degree of freedom, a good rule is to apply the correction whenever it produces any appreciable difference in the significance probability. The correction has a tendency to over-correct, changing the tabular P from too small to too large, but is usually an improvement.

If a number of X^2 values, each with 1 degree of freedom, are added to form a total X^2, the individual X^2 values should *not* be corrected for continuity, because the over-correction mounts up in a disconcerting manner [19]. After it has been obtained, the total X^2 is corrected, if this is necessary, by the method given in the following paragraph.

Compute the next largest value of X^2 which the structure of the data permits. Read the χ^2 table at a point halfway between this value and the observed X^2. This procedure was illustrated for a 2×4 table in the preceding section. Sometimes the next largest value of X^2 is not immediately obvious and trial and error is required to find it.

9. The construction of classes. When X^2 is used to test the hypothesis that the observations follow a *continuous* frequency distribution, the first step is to group the observations into classes. Both the number of classes and the division points between classes are at the disposal of the investigator, and the choices that he makes will affect the sensitivity of the test. I believe that the common practice is to have a moderate number of classes, say between 10 and 25, and to make the class intervals equal. Although information about the best rule for constructing the classes is still meager, the recommendations of those who have looked into this problem are contrary to current practice.

With regard to class intervals, Mann and Wald [29] and Gumbel [30] suggest that these be chosen so that the *expected* number is the same ($= n/k$) in all classes. These authors do not claim that this will increase the power of the test, but merely suggest that it is likely to be a good procedure. Gumbel points out that if this method is used in conjunction with a rule for choosing the value of k, much of the arbitrariness that accompanies the construction of class intervals is removed. Under this method, the computational steps are first to estimate the constants (mean, s.d., etc.) which determine the fitted curve, then find the class boundaries which give equal expectations in each class, and finally count the *observed* numbers x_i which fall in the respective classes. The value of X^2 is given as

$$X^2 = \frac{k}{n} \sum x_i^2 - n.$$

The paper by Mann and Wald deals with the choice of the number of classes, k. The null hypothesis is assumed to specify the distribution completely, and n is assumed large enough so that the limiting χ^2 distribution is applicable.

Some criterion is required to define what is meant by a "best" value of k.

This content downloaded from 128.97.55.209 on Mon, 10 Mar 2014 13:46:32 PM
All use subject to JSTOR Terms and Conditions

It seems natural to try to maximize some property of the power function of the test. The criterion set up by Mann and Wald is a little complex to describe, but this stems from the complexity of the problem.

They define the distance Δ between the null distribution and any alternative distribution as the maximum difference between the heights of the two cumulative distribution functions. It becomes evident, after some examination of the problem, that there is no hope of choosing k so as to maximize the power function of X^2 at all points along its course. They decide to concentrate on maximizing the power at about the point where the power is $\frac{1}{2}$. This is an arbitrary but reasonable choice. The two principal properties possessed by their "best" k are as follows.

(i) For a value of Δ which they determine, the power of the X^2 test is at least $\frac{1}{2}$ for all alternative distributions whose distance from the null distribution is at least Δ. This value Δ is a simple function of sample size and, as would be expected, it decreases steadily with increasing sample size.

(ii) If any k other than the "best" is chosen, the power of X^2 is less than $\frac{1}{2}$ for at least one alternative whose distance from the null distribution exceeds Δ.

The best k is given by the formula

$$k = 4\left[\frac{2(n-1)^2}{c^2}\right]^{\frac{1}{5}},$$

where

$$\frac{1}{\sqrt{2\pi}}\int_c^\infty e^{-(x^2/2)}\,dx = \alpha,$$

where α is the significance level. Thus $c = 1.64$ for a test at the 5% level.

The optimum values of k are substantially higher than those customary in practice. For a test at the 5% level, k rises slowly from 31 at $n = 200$ to 78 at $n = 2{,}000$.

TABLE 3

Optimum number of classes (Mann and Wald)

n	200	400	600	800	1,000	1,500	2,000
k	31	41	48	54	59	70	78

A good exposition and critique of the Mann-Wald paper has been given by Williams [31]. The Mann-Wald method is more tedious to compute than the usual procedure, partly because of the increased number of classes and partly because of the fitting with equal *expected* numbers. Williams shows, however, that the optimum is a broad one, and that the value of k in Table 3 can probably be halved with little loss in sensitivity.

The Mann-Wald paper, although an able performance in a difficult field, is far from a complete investigation of the optimum number of classes. Such an

This content downloaded from 128.97.55.209 on Mon, 10 Mar 2014 13:46:32 PM
All use subject to JSTOR Terms and Conditions

investigation is unlikely to be forthcoming soon. What is the user of the X^2 test to conclude from their results? My own reaction has been to put more computational work into X^2 tests of continuous distributions, by increasing the number of intervals and using unequal lengths of interval where this is necessary in order to avoid classes with high expected numbers. For sample sizes between 200 and 1000, their recommended expected numbers per class in Table 3 range from 6 to 16. With Williams' modification, the range is from 12 to 32. This does suggest that there is an appreciable loss of power if classes with expectations of more than 50 are commonly used.

10. Summary recommendations. The following is an attempt to set down in brief form the recommendations about the computation of X^2 which flow from the discussion in this part and from practical experience. The recommendations are not as explicit as I should like. They can, I believe, be made more explicit, but this requires detailed study that goes beyond the scope of the present paper. The total number of observations is n.

I. Attribute data. The data come to us in grouped form. Pooling of classes is considered undesirable because of loss of power.

(a) *The* 2×2 *table.* Use Fisher's exact test (i) if $n < 20$, (ii) if $20 < n < 40$ and the smallest expectation is less than 5. Mainland's tables [20] are helpful in all such cases. If $n > 40$, use X^2, corrected for continuity if the smallest expectation is less than 500.

(b) *Tables with degrees of freedom between* 2 *and* 60 *and all expectations less than* 5. If n is so small that Fisher's exact test can be computed without excessive labor, use this. Otherwise use X^2, considering whether this needs correction for continuity by finding the next largest value of X^2.

(c) *Tables with degrees of freedom greater than* 60 *and all expectations less than* 5. Try to obtain the exact mean and variance of X^2 and use the normal approximation to the exact distribution.

(d) *Tables with more than* 1 *degree of freedom and some expectations greater than* 5. Use X^2 without correction for continuity.

II. Continuous data. The data must first be grouped. Use enough cells to keep the expectations down to the levels recommended by Williams (12 per cell for $n = 200$, 20 per cell for $n = 400$, 30 per cell for $n = 1{,}000$). At the tails, pool (if necessary) so that the minimum expectation is 1.

Part III. Utility of the Test

11. Criticisms and limitations of the test. A competent appraisal of the utility of the X^2 test would require a sampling survey of scientists and others who try to draw conclusions from data. In such a survey the object would be to discover how frequently these workers have occasion to use a X^2 test, and to what extent the application of the test really seems to help them. In fact, such a survey, directed at the use of statistical techniques in general and not merely at the X^2 test, might be very illuminating to statisticians if it could be carried out despite the obvious difficulties. Statisticians are, I think, rather quick to jump to con-

This content downloaded from 128.97.55.209 on Mon, 10 Mar 2014 13:46:32 PM
All use subject to JSTOR Terms and Conditions

clusions about the kinds of problems which scientists in other fields are supposed to face, and about their presumed uses and misuses of statistical methods and ideas.

In the absence of survey data of this kind, the statistician can give only a personal opinion, based on such contacts as he has had with the users of the X^2 test. I will content myself with the cautious statement that since the construction of hypotheses and their continued modification or rejection in the light of new data is one of the standard tools of science, some kind of test of the agreement between theory and data must often be useful. The experiences of Airy and Merriman illustrate the uncomfortable position in which the scientist is placed when he has to state, without the benefit of such a test, whether his observations are in accordance with the predictions of some theory.

On the other hand, a reading of the literature reveals the opinion, expressed by several writers, that the X^2 test is of restricted usefulness. The reasons for this critical verdict seem to be diverse. Some of the criticism is directed at the X^2 test itself, but some seems to apply to composite, or "general purpose" tests of significance as a whole, and some to *all* tests of significance.

Considering first the criticisms of X^2 itself, the name "goodness of fit" is misleading, because the power of the test to detect an underlying disagreement between theory and data is controlled largely by the size of the sample. With a small sample, an alternative hypothesis which departs violently from the null hypothesis may still have a small probability of yielding a significant value of X^2. In a very large sample, small and unimportant departures from the null hypothesis are almost certain to be detected. Consequently, when X^2 is nonsignificant, the amount by which the null hypothesis has been strengthened depends mainly on the size of sample. This is one of the principal reasons for such misuse of the test as exists. Authors sometimes write as if the validity of their null hypothesis has been greatly strengthened, if not definitely established, by a goodness of fit test made on very scanty data.

Secondly, as Gumbel has pointed out, the X^2 test for a continuous frequency distribution is not unique, because of the freedom of choice of number of intervals and end-points of the intervals. Although this is an argument for more standardization in the application of the test, the objection perhaps is minor rather than major. At least, statisticians have not seen any overwhelming advantage in having just one test of a given null hypothesis against a given alternative. In recent years there has been active research in the development of quick, though inefficient, tests for problems in which satisfactory, but less speedy, tests already exist. The tests will give different values of P from the same data, but no serious objections to this situation seem to have been noticed.

There are two available substitute tests which resemble the X^2 test in that they are not directed against any specific alternative. One is the ω^2 test (Section 13). This was constructed in order to avoid the grouping of continuous data that is necessary with X^2. The other, for data that are in grouped form, is the likelihood ratio test against a completely general alternative hypothesis (Section 14).

One limitation of X^2, or of any nonspecific test, is that when the alternative

This content downloaded from 128.97.55.209 on Mon, 10 Mar 2014 13:46:32 PM
All use subject to JSTOR Terms and Conditions

hypotheses can be fairly clearly defined, we may hope to obtain another test, directed against these alternatives, that will be more powerful than X^2. An example is Neyman's "smooth" tests (Section 15). These were constructed to detect alternative hypotheses that depart from the null hypothesis in some continuous or smooth fashion. Like X^2, the smooth tests are still general, since they do not demand detailed knowledge of the nature of the alternatives. Further down the scale there is a variety of supplementary tests to X^2.

Finally, the X^2 test is sometimes used when what is needed is not a test of significance of the usual type. There are numerous occasions when the null hypothesis is not expected to be exactly true, but at best approximately true. The argument against X^2 in this situation has been developed amusingly by Berkson [32]. He writes "I make the following dogmatic statement, referring for illustration to the normal curve: 'If the normal curve is fitted to a body of data representing any real observations whatever of quantities in the physical world, then if the number of observations is extremely large—for instance, on the order of 200,000—the chi-square P will be small beyond any usual limit of significance.'"

If this statement is granted—and counter-evidence, to put it mildly, is not abundant—then Berkson proceeds to the Socratic conclusion. What is the point of applying a X^2 test to a moderate or small sample if we already know that a large sample would show P highly significant?

In his original paper, Karl Pearson was aware of this issue, and did not seem to feel uncomfortable about it. He writes, "Nor again does it appear to follow that if the number be largely increased the same curve will still be a good fit. Roughly the χ^2's of two samples appear to vary for the same grouping as their total contents. Hence if a curve be a good fit for a large sample, it will be good for a small one, but the converse is not true, and a larger sample may show that our theoretical frequency gives only an approximate law for samples of a certain size. In practice we must attempt to obtain a good fitting frequency for such groupings as are customary or utile. To ascertain the ultimate law of distribution of a population for any groupings, however small, seems a counsel of perfection." Although it is hazardous to try to read another man's mind, his attitude was apparently the defensible one that any theory is at best approximately true, but nevertheless, if we are going to reject a theory, we do so because it does not fit the data that we have, not because it would not fit a much larger sample of data that we do not have.

Nevertheless, I would agree with Berkson that in this situation an ordinary test of significance is not very useful. It is more difficult to say just what we do want. One attack would be to reformulate the null hypothesis so that, instead of testing whether a binomial p equals p_0, we try to construct a test of the null hypothesis that p lies in the specified range p_0, p_1.

As an alternative approach, fiducial or confidence limits seem to be helpful. Suppose that these limits are set up for the difference between two percentages in a 2×2 contingency table, the ordinary null hypothesis being that the true difference is zero. If the two limits are far from zero, then even when X^2 is non-

This content downloaded from 128.97.55.209 on Mon, 10 Mar 2014 13:46:32 PM
All use subject to JSTOR Terms and Conditions

significant we are warned that the data do not establish the null hypothesis as approximately true. If the limits are both near zero, on the other hand, we may be able to conclude that the null hypothesis, although presumably not exactly true, is close enough to the correct hypothesis for all practical purposes.

In testing goodness of fit of a frequency distribution, the extension of this approach is Kolmogorov's method [33] for constructing confidence bounds for the cumulative frequency distribution, given a sample.

To summarize, the X^2 test is helpful primarily in the exploratory stages of an investigation, when there is no very clear knowledge of the alternative hypotheses. It is well to remember that the size of sample determines whether the test really is a severe test of the null hypothesis.

12. Interpretation of high *P*'s. The question of the interpretation to be placed on very high P's, say those greater than 0.99, has been raised from time to time. In the few instances of this kind that have come my way, my practice has been to give the data further scrutiny before regarding the result as evidence in favor of the null hypothesis.

Events have justified this practice. In nearly every instance, something wrong was discovered, most frequently a numerical mistake or an error in the formula used to compute X^2. In one set of data assembled by geneticists, a whole group of X^2 showed P's of the order of 0.999. The reason was that these X^2 values had been obtained by adding a large number of X^2 values, each with 1 degree of freedom, and all the original (1 d.f.) X^2 had been corrected for continuity. The over-correction which is a feature of this device had piled up to such an extent that their total X^2's were much smaller than those following the χ^2 distribution. In another case, after discussion with the assistants of the scientist in charge, I surmised that the observations had been influenced by the anticipations of the scientist. Fisher [34] has raised a similar speculation with respect to some of Mendel's results, without any suggestion of improper scientific conduct on the part of Mendel.

Part IV. Tests Which Are Competitive to X^2

13. The ω^2 test. Alternatives that have been proposed to the X^2 test are of two kinds. Several of the tests, like X^2, are "general" tests. Then there is a battery of supplementary tests that are intended for situations where the alternative hypothesis is more definitely known.

The *general* substitute tests that have been proposed have not given X^2 very serious competition. This is understandable because of the long history of X^2 and of its inclusion as standard doctrine in most elementary courses, and because some of the substitute tests are limited in the type of hypothesis with which they can cope. Moreover, despite the weaknesses of the X^2 test discussed in Part III, the advantages of the alternative tests have not yet been clearly enough demonstrated to win many converts. Consequently, Part IV contains only a brief and rather noncommital introduction to these tests.

This content downloaded from 128.97.55.209 on Mon, 10 Mar 2014 13:46:32 PM
All use subject to JSTOR Terms and Conditions

The first general test, developed by Cramér [35], von Mises [36] and Smirnov [37], was constructed mainly for use with small samples. The null hypothesis completely specifies the frequency distribution followed by the observations. Unlike X^2, the ω^2 test requires no grouping of the observations, an obvious advantage with small samples. The test is based on a comparison of the cumulative frequency function $F(x)$ specified by the null hypothesis with an estimate of the cumulative frequency made from the sample. This estimate, $F^*(x)$, is simply r/n, where r is the number of observations in the sample which are $\leqq x$. The test criterion proposed is the Stieltjes integral

$$\omega^2 = \int_{-\infty}^{\infty} [F(x) - F^*(x)]^2 \, dF(x).$$

If $F(x)$ is continuous, this may be shown to satisfy

$$\omega^2 = \frac{1}{12n^2} + \frac{1}{n} \sum_{r=1}^{n} \left[F(x_r) - \frac{2r - 1}{2n} \right]^2,$$

where the values $x_1, x_2, \cdots, x_n$ are now arranged in increasing order.

The mean and variance of ω^2 are known, and also its limiting distribution (which is nonnormal) as $n \to \infty$. A table of this distribution by Darling and Anderson [38] has appeared recently. Practical use of this test is restricted by the condition that $F(x)$ must be known and by lack of information about the small-sample distribution of ω^2.

14. The likelihood ratio test. If the data are presented in grouped form, and if the alternative hypothesis is completely general, it is known that in large samples the X^2 test and the likelihood ratio test become equivalent [9]. We start from the usual multinomial

$$Pr = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}.$$

The likelihood on the null hypothesis is found from $p_i = m_i/n$, where m_i are the expectations estimated by maximum likelihood (unless they are explicitly given). The likelihood on the unrestricted alternative is found from $p_i = x_i/n$. Hence the likelihood ratio becomes

$$\left(\frac{x_1}{m_1}\right)^{x_1} \left(\frac{x_2}{m_2}\right)^{x_2} \cdots \left(\frac{x_k}{m_k}\right)^{x_k}.$$

Its logarithm is

$$L = \sum_{i=1}^{k} x_i \log \left(\frac{x_i}{m_i}\right) = \sum_{i=1}^{k} x_i \log \left\{1 + \frac{x_i - m_i}{m_i}\right\}.$$

When this is expanded in a power series in the $(x_i - m_i)$, the leading term is X^2 for the maximum likelihood estimates of the parameters.

In view of the equivalence of the two criteria in large samples, there seems no advantage, except one of taste or convenience, in one test over the other.

This content downloaded from 128.97.55.209 on Mon, 10 Mar 2014 13:46:32 PM
All use subject to JSTOR Terms and Conditions

For small samples, the suggestion has been made from time to time that the likelihood ratio is to be preferred. Examples worked by both tests have been presented and discussed by Neyman and Pearson [22] and Fisher [39]. Since users of statistical methods naturally do not wish to learn more tests than are necessary, a movement to replace X^2 by the likelihood ratio seems unlikely to gather momentum unless some definite advantages can be shown to follow. The advantage in computing time is at most small, but there may be an increase in power. The striking way in which many different configurations of the data turn out to give exactly the same value of X^2 in small samples suggests an element of coarseness in the X^2 test. This coalescence happens to a much reduced extent with the likelihood ratio. However, not enough data about relative power has accumulated to permit a verdict on this issue.

15. Neyman's smooth tests. As in the ω^2 test, Neyman [40] postulates that the cumulative frequency $F(x)$ (assumed continuous) is known exactly from the null hypothesis. The first step is to replace the observations x_i by the familiar "probability integral" transforms y_i, where

$$y_i = F(x_i).$$

If the null hypothesis is correct, the variates y_i follow a rectangular distribution in the interval (0, 1). The problem, therefore, reduces to that of finding a test for this transformed hypothesis.

Neyman points out that the conceivable alternatives to the null hypothesis fall into two broad classes. The first class, of "smooth" alternatives, contains frequency functions which are continuous and which depart in some gradual and regular manner from the null hypothesis. The second class contains all other alternatives, whose deviation from the null hypothesis is in some respects erratic or discontinuous. The X^2 test is not directed specifically at either class, and is to some extent effective against both types of departure from the null hypothesis. Neyman's object is to develop tests sensitive to the first class of alternatives.

If a "smooth" alternative holds, the transforms y_i will no longer follow a rectangular distribution, but will presumably follow a continuous distribution with a limited number of maxima and minima. The proposal is, therefore, to test the y_i for polynomial trends, on the assumption that a polynomial of fairly small degree will satisfactorily represent the smooth alternative. The computations involved and a discussion of the appropriate degree of the polynomial are presented in [40].

Part V. Tests Which Are Supplementary to X^2

16. A supplementary test based on runs. X^2 takes no account of the succession of + and − signs in the deviations $(x_i - m_i)$ between observations and expectations. When a smooth alternative holds, it seems likely that the succession of signs will exhibit some systematic features, such as a long run of +'s followed

This content downloaded from 128.97.55.209 on Mon, 10 Mar 2014 13:46:32 PM
All use subject to JSTOR Terms and Conditions

by a run of $-$'s, and this has often been observed in applications of X^2. David [41] has adapted the now familiar "run" test as a supplementary test to X^2. In the run test, we count the number of runs and refer to a table which shows the significance levels of this quantity, given the total numbers of $+$'s and $-$'s in the series. David has shown that the limiting distribution of the number of runs is independent of that of X^2.

The run tests will, of course, be most effective for alternatives which produce few runs, such as a shift in the mean of the distribution. In the reference cited, the test is restricted to the case where the null hypothesis completely specifies the distribution; David states that a test has been developed for the case in which some parameters must be estimated.

17. Tests based on low moments. When the null hypothesis postulates that the observations follow a normal, binomial or Poisson distribution, an alternative to X^2 that is in fairly common use is to compare the lower moments of the theoretical distribution with estimated moments from the sample. With the normal distribution, the actual values of the mean and variance are rarely given by the null hypothesis, so that a comparison of these moments is not usually possible. Tests of skewness, derived from the third moment, and of kurtosis, derived from the fourth moment, can be made [15].

In the binomial distribution, if p is specified we can compare the sample mean and variance with the theoretical mean and variance. If p is not specified, we can compare variances. Suppose that we have g series of trials, and that each series contains s trials. The number of successes (out of s) in the ith trial is x_i. For a test of the mean proportion of successes, we regard

$$\frac{\left(\frac{\bar{x}}{s} - p\right)}{\sqrt{\frac{pq}{gs}}}$$

as a normal deviate.

18. Dispersion tests. Turning to the variance of x_i, if p is specified the estimated variance is $\sum (x_i - sp)^2/g$, while the theoretical variance is spq. An appropriate test criterion for the variance is, therefore,

$$\frac{\sum (x_i - sp)^2}{gspq}.$$

If p must be estimated from the data, either because it is unspecified or because the sample estimate disagrees with the postulated p, the test criterion becomes

$$\frac{\sum (x_i - \bar{x})^2}{gs\left(\frac{\bar{x}}{s}\right)\left(1 - \frac{\bar{x}}{s}\right)} = \frac{s\sum (x_i - \bar{x})^2}{g\bar{x}(s - \bar{x})}.$$

This content downloaded from 128.97.55.209 on Mon, 10 Mar 2014 13:46:32 PM
All use subject to JSTOR Terms and Conditions

As is well known, the related quantities

$$\frac{\sum (x_i - sp)^2}{spq} \quad \text{and} \quad \frac{s \sum (x_i - \bar{x})^2}{\bar{x}(s - \bar{x})} \tag{13}$$

are distributed approximately as χ^2 with g and $(g - 1)$ degrees of freedom, respectively, when the null hypothesis is true.

The variance test can also be made when the number of trials s_i varies from series to series. The test criterion becomes

$$\frac{\sum s_i \left(\frac{x_i}{s_i} - \frac{\bar{x}}{\bar{s}}\right)^2}{\frac{\bar{x}}{\bar{s}}\left(1 - \frac{\bar{x}}{\bar{s}}\right)} = \frac{\bar{s}^2 \left\{\sum \frac{x_i^2}{s_i} - \frac{(\sum x_i)^2}{\sum s_i}\right\}}{\bar{x}(\bar{s} - \bar{x})}.$$

This test criterion can be shown to be identical with X^2 as calculated for the $2 \times g$ contingency table.

	x_1	x_2	$\cdots$ x_g	$\sum x_i$
	$s_1 - x_1$	$s_2 - x_2$	$\cdots$ $s_g - x_g$	$\sum s_i - \sum x_i$
	s_1	s_2	$\cdots$ s_g	$\sum s_i$

It is important to distinguish clearly between this variance X^2 test, which is sometimes called a *dispersion* test, and the ordinary goodness of fit X^2 test. Suppose that we have 200 families each of size 4, and that every individual belongs to one of two classes A and a. The null hypothesis states that the probability p of an A is constant for all families. We may tally the numbers of families that have 0, 1, 2, 3, 4 A's, respectively, and test this frequency distribution against the binomial $(q + p)^4$. This is the ordinary goodness of fit test, which has 3 degrees of freedom if p is unknown. The dispersion test, computed by formula (13), compares the observed variance of this frequency distribution with the theoretical binomial variance.

The dispersion tests frequently prove more sensitive than X^2 when the binomial null hypothesis fails because the probability of an A varies from one family to another. The notion of a measure of dispersion of this kind is due to Lexis and antedates the goodness of fit test.

19. Subdivision of X^2 into components. In the analysis of variance, the subdivision of a sum of squares into single components, or "single degrees of freedom," is a familiar device. If variates y_i are normally and independently distributed with mean 0 and variance σ^2 on some null hypothesis, these components are obtained by any linear transformation of the form

$$z_i = \sum_{j=1}^{k} l_{ij} y_j \qquad (i = 1, 2, \cdots k)$$

This content downloaded from 128.97.55.209 on Mon, 10 Mar 2014 13:46:32 PM
All use subject to JSTOR Terms and Conditions

where

$$\sum_j l_{ij} l_{hj} = \begin{cases} 0, & i \neq h, \\ 1, & i = h. \end{cases} \tag{14}$$

All z_i are normally and independently distributed with mean 0 and variance σ^2. This transformation enables us to select those z_i that are likely to be sensitive to a particular alternative hypothesis. Often only one or two of the z_i are examined, because it is hard to imagine any feasible alternative that would make the other z's large. Thus the device replaces a "sum of squares" test by a few more specialized tests.

The corresponding subdivision of X^2 is easily obtained from Fisher's device of regarding the observed numbers x_i in the cells as following independent Poisson distributions, subject to a single linear restriction. Thus when all expectations are large, we may take

$$y_i = \frac{x_i - m_i}{\sqrt{m_i}}$$

as the set of unit normal deviates. Since these are subject to the linear restriction

$$\sum_{j=1}^{k} (x_j - m_j) = \sum_{j=1}^{k} \sqrt{m_j}\, y_j = 0,$$

we must take

$$Z_1 = \sum \sqrt{m_j}\, y_j / n.$$

Let the remaining Z_i $(i = 2, 3, \cdots k)$ be

$$Z_i = \sum_j l'_{ij} x_j = \sum l'_{ij} \sqrt{m_j}\, y_j + \sum l'_{ij} m_j.$$

If these are to have means zero, we must have

$$\sum_j l'_{ij} m_j = 0. \tag{15}$$

Note that this relation makes all the remaining Z_i orthogonal with Z_1. Since $l_{ij} = l'_{ij} \sqrt{m_j}$, equations (14) become

$$\sum_j l'_{ij} l'_{hj} m_j = \begin{cases} 0 & i \neq h, \\ 1 & i = h. \end{cases} \tag{16}$$

Any set of Z_i whose coefficients satisfy equations (15) and (16) provide a breakdown of X^2 into $(k - 1)$ single components. Then as an algebraic identity,

$$X^2 = \sum_{i=1}^{k} \frac{(x_i - np_i)^2}{np_i} = \sum_{i=1}^{k-1} Z_i^2$$

As n increases, the individual terms on the right become in the limit independently distributed as χ^2 with 1 degree of freedom. In genetic analysis,

This content downloaded from 128.97.55.209 on Mon, 10 Mar 2014 13:46:32 PM
All use subject to JSTOR Terms and Conditions

where simple interpretations can be attached to the Z_i, this tool has proved useful [15].

The application of this breakdown to contingency tables, which requires care, has been elucidated by Lancaster [42] and Irwin [43]. In an $r \times c$ contingency table, X^2 can be partitioned into $(r-1)(c-1)$ single components. Each of these represents the usual X^2 for a 2×2 table which is formed by amalgamation of cells in the original table. This breakdown is illustrated below for a 3×3 table.

Original table

x_{11}	x_{12}	x_{13}	r_1
x_{21}	x_{22}	x_{23}	r_2
x_{31}	x_{32}	x_{33}	r_3
c_1	c_2	c_3	n

Components

x_{11}	x_{12}	r_{12}	r_{12}	x_{13}	r_1
x_{21}	x_{22}	r_{22}	r_{22}	x_{23}	r_2
c_{21}	c_{22}	n_{22}	n_{22}	c_{23}	n_{23}
c_{21}	c_{22}	n_{22}	n_{22}	c_{23}	n_{23}
x_{31}	x_{32}	r_{32}	r_{32}	x_{33}	r_3
c_1	c_2	n_{32}	n_{32}	c_3	n

If the X^2 are calculated in the usual way for each 2×2 table, the partition is only approximate, in that in finite samples the individual X^2 do not add up to the total X^2 for the 3×3 table. The authors show how to obtain a partition which adds up exactly, that is, which satisfies the sets of equations (15) and (16). It appears that the approximate partition is adequate for most tests of significance; in fact, it has not been shown that the additive partition is really preferable to the approximate partition in small samples.

Another application of the breakdown of X^2 is to contingency tables in which numerical scores can be attached to one or both of the classifications. Yates [44] shows how to isolate and compare the regressions of the observations on these scores.

In conclusion, the testing of individual components of X^2 is analogous to the use of a set of independent t-tests instead of, or in addition to, an F-test in the analysis of variance.

I wish to thank T. W. Anderson, E. L. Lehmann and J. W. Tukey for many helpful suggestions.

This content downloaded from 128.97.55.209 on Mon, 10 Mar 2014 13:46:32 PM
All use subject to JSTOR Terms and Conditions

REFERENCES

[1] KARL PEARSON, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *Philos. Mag. Series 5*, Vol. 50 (1900), pp. 157–172.

[2] H. HOTELLING, "The consistency and ultimate distribution of optimum statistics," *Trans. Am. Math. Soc.*, Vol. 32 (1930), pp. 847–859.

[3] R. A. FISHER, "On the interpretation of chi square from contingency tables, and the calculation of P," *Jour. Roy. Stat. Soc.*, Vol. 85 (1922), pp. 87–94.

[4] M. GREENWOOD AND G. U. YULE, "The statistics of anti-typhoid and anti-cholera inoculations and the interpretation of such statistics in general," *Proc. Roy. Soc. Med.*, Vol. 8 (1915), pp. 113–190.

[5] R. A. FISHER, "The conditions under which chi square measures the discrepancy between observation and hypothesis," *Jour. Roy. Stat. Soc.*, Vol. 87 (1924), pp. 442–450.

[6] G. U. YULE, "On the application of the χ^2 method to association and contingency tables, with experimental illustrations," *Jour. Roy. Stat. Soc.*, Vol. 87 (1922), pp. 76–82.

[7] J. BROWNLEE, "Some experiments to test the theory of goodness of fit," *Jour. Roy. Stat. Soc.*, Vol. 87 (1924), pp. 76–82.

[8] H. CRAMÉR, *Mathematical Methods of Statistics*, Princeton University Press, 1946, p. 424.

[9] J. NEYMAN, "Contribution to the theory of the χ^2 test," *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 1949, pp. 239–273.

[10] E. FIX, "Tables of the Noncentral χ^2," *Univ. California Publ. Stat.*, Vol. 1 (1949), pp. 15–19.

[11] P. B. PATNAIK, "The non-central χ^2- and F-distributions and their applications," *Biometrika*, Vol. 36 (1949), pp. 202–232.

[12] A. WALD, "Tests of statistical hypotheses concerning several parameters when the number of observations is large," *Trans. Am. Math. Soc.*, Vol. 54 (1943), pp. 426–482.

[13] G. A. BARNARD, "Significance tests for 2 × 2 tables," *Biometrika*, Vol. 34 (1947), pp. 123–138.

[14] R. A. FISHER, *The Design of Experiments*, Oliver and Boyd Ltd., Edinburgh, 1935.

[15] R. A. FISHER, *Statistical Methods for Research Workers*, Edinburgh, 5th and subsequent editions, Oliver and Boyd Ltd., Edinburgh, 1934, Section 21.02.

[16] K. D. TOCHER, "Extension of the Neyman-Pearson theory of tests to discontinuous variates," *Biometrika*, Vol. 37 (1950), pp. 130–144.

[17] D. LEWIS AND C. J. BURKE, "The use and misuse of the chi-square test," *Psych. Bull.*, Vol. 46 (1949), pp. 433–489.

[18] A. L. EDWARDS, "On the use and misuse of the chi-square test—the case of the 2 × 2 contingency table," *Psych. Bull.*, Vol. 47 (1950), pp. 341–346.

[19] W. G. COCHRAN, "The χ^2 correction for continuity," *Iowa State Coll. Jour. Sci.*, Vol. 16 (1942), pp. 421–436.

[20] D. MAINLAND, "Statistical methods in medical research," *Canadian Jour. Res., E*, Vol. 26 (1948), pp. 1–166.

[21] R. A. FISHER AND F. YATES, *Statistical Tables for Biological, Agricultural and Medical Research*, Oliver and Boyd, Ltd., Edinburgh, 1938, Table VIII.

[22] J. NEYMAN AND E. S. PEARSON, "Further notes on the χ^2 distribution," *Biometrika*, Vol. 22 (1931), pp. 298–305.

[23] W. G. COCHRAN, "The χ^2 distribution for the binomial and Poisson series, with small expectations," *Annals of Eugenics*, Vol. 7 (1936), pp. 207–217.

This content downloaded from 128.97.55.209 on Mon, 10 Mar 2014 13:46:32 PM
All use subject to JSTOR Terms and Conditions

[24] P. V. SUKHATME, "On the distribution of χ^2 in small samples of the Poisson series," *Jour. Roy. Stat. Soc. Suppl.*, Vol. 5 (1938), pp. 75–79.
[25] G. H. FREEMAN AND J. H. HALTON, "Note on an exact treatment of contingency, goodness of fit and other problems of significance," *Biometrika*, Vol. 38 (1951), pp. 141–149.
[26] J. B. S. HALDANE, "The exact value of the moments of the distribution of χ^2, used as a test of goodness of fit, when expectations are small," *Biometrika*, Vol. 29 (1937), pp. 133–143.
[27] J. B. S. HALDANE, "The mean and variance of χ^2, when used as a test of homogeneity, when expectations are small," *Biometrika*, Vol. 31 (1939), pp. 346–355.
[28] F. YATES, "Contingency tables involving small numbers and the χ^2 test," *Jour. Roy. Stat. Soc. Suppl.*, Vol. 1 (1934), pp. 217–235.
[29] H. B. MANN AND A. WALD, "On the choice of the number of class intervals in the application of the chi square test," *Annals of Math. Stat.*, Vol. 13 (1942), pp. 306–317.
[30] E. J. GUMBEL, "On the reliability of the classical χ^2 test," *Annals of Math. Stat.*, Vol. 14 (1943), pp. 253–263.
[31] C. ARTHUR WILLIAMS, "On the choice of the number and width of classes for the chi-square test of goodness of fit," *Jour. Am. Stat. Assn.*, Vol. 45 (1950), pp. 77–86.
[32] J. BERKSON, "Some difficulties of interpretation encountered in the application of the chi-square test," *Jour. Am. Stat. Assn.*, Vol. 33 (1938), pp. 526–536.
[33] A. KOLMOGOROV, "Confidence limits for an unknown distribution function," *Annals of Math. Stat.*, Vol. 12 (1941), pp. 461–465.
[34] R. A. FISHER, "Has Mendel's work been re-discovered?" *Annals of Science*, Vol. 1 (1936), pp. 115–137.
[35] H. CRAMÉR, "On the composition of elementary errors," *Skandinavisk Aktuarietidskrift*, Vol. 11 (1928), pp. 13–74, 141–180.
[36] R. VON MISES, *Wahrscheinlichkeitsrechnung und ihre Anwendung in der Statistik und theoretischen Physik*, Deuticke, Leipzig und Wien, 1931, pp. 316–335.
[37] N. SMIRNOV, "Sur la distribution de ω^2," *C. R. Acad. Sci.* Paris, Vol. 202 (1936), p. 449.
[38] D. A. DARLING AND T. W. ANDERSON, "Asymptotic theory of certain goodness of fit criteria based on stochastic processes," *Annals of Math. Stat.*, Vol. 23 (1952), pp. 193–212.
[39] R. A. FISHER, "The significance of deviations from expectation in a Poisson series," *Biometrics*, Vol. 6 (1950), pp. 17–24.
[40] J. NEYMAN, "Smooth test for goodness of fit," *Skandinavisk Aktuarietidskrift*, Vol. 20 (1937), pp. 150–199.
[41] F. N. DAVID, "A χ^2 'smooth' test for goodness of fit," *Biometrika*, Vol. 34 (1947), pp. 299–310.
[42] H. O. LANCASTER, "The derivation and partition of χ^2 in certain discrete distributions," *Biometrika*, Vol. 36 (1949), pp. 117–129.
[43] J. O. IRWIN, "A note on the subdivision of χ^2 into components," *Biometrika*, Vol. 36 (1949), pp. 130–134.
[44] F. YATES, "The analysis of contingency tables with groupings based on quantitative characters," *Biometrika*, Vol. 35 (1948), pp. 176–181.

This content downloaded from 128.97.55.209 on Mon, 10 Mar 2014 13:46:32 PM
All use subject to JSTOR Terms and Conditions