SOME RECENT RESULTS IN CHI-SQUARE GOODNESS-OF-FIT TESTS

by

G. S. Watson

Institute of Statistics

Mimeo. Series No. 209

August, 1958

Some Recent Results in Chi-Square Goodness-of-Fit Tests

by

G. S. Watson
Australian National University

Abstract

The aim of this paper is to relate and extend some recent work on chi-square goodness-of-fit tests. There is no discussion of any problems which are specifically associated with more than one categorical variable. The main topics are the effect of estimation on chi-square and its partitions and their relation to Neyman's smooth goodness-of-fit test, and the effect of grouping a univariate distribution according to the disposition of the sample on the distribution of the chi-square statistic and on the smooth test statistic.

Some Recent Results on Chi-Square Goodness-of-Fit Tests

G. S. Watson

Australian National University*

*(This paper was written while the author was visiting the Department of Experimental Statistics, North Carolina State College.)

1. Introduction

The aim of this paper is to review, and make some additions to, some recent work on chi-square goodness-of-fit tests. It is restricted to a discussion of the testing of univariate distributions and the choice and emphasis of topics is a personal one. The main subject is the asymptotic effect of estimation on chi-square and its partitions. This summarizes a sequence of papers by Chernoff and Lehmann (1954), Barton (1956), Roy (1956), Watson (1957b, 1958). The main new result deals with the effect of forming the classes by examination of the data -- it will be shown that this has no effect on the asymptotic distribution of the chi-square statistic. Mention is also made of efforts to make conditional chi-square tests easier to apply and of several papers on improvements to the asymptotic distribution of chi-square. Finally some remarks are made on the metrical aspects of chi-square.

It was found convenient to divide the discussion of the effects of estimation into two parts. In Section 2, the true multinomial distribution is examined and the effect of estimates on the asymptotic distribution of the chi-square statistic -- written X^2 here -- and of linear forms in the class frequencies is studied. This theory has, of course, many more uses than merely for testing univariate distributions. However to have considered bivariate categorical variables would have greatly increased the length of the review since it would have been necessary to discuss the canonical analysis and other aspects of contingency tables, a subject of perennial interest. In Section 4, the special problems which arise in testing the distributions of continuous variates, which depend on unknown parameters, are

discussed. In particular the effect of allowing the boundaries of the class intervals to vary from sample to sample is discussed.

In Section 3, χ^2 tests, conditional upon certain sufficient statistics being held constant, are discussed for both small and large samples and a reference made to asymptotic expansions of the distribution of χ^2 . Finally in Section 5, several papers by Barton and Lancaster are related to our general treatment. Barton's work on Neyman's smooth goodness of fit test (1937), in the case where parameters must be estimated and where it is convenient to group the observations, contains several very general theorems which are referred to in Sections 2 and 4. It is however convenient to delay any discussion of smooth goodness-of-fit tests until Section 5.

2. The Multinomial Distribution - Asymptotic Theory

Let N observations be drawn from a multinomial distribution of k classes with class probabilities $p_1, p_2, \dots, p_k$ such that the class frequencies are $n_1, n_2, \dots, n_k$, $\sum_{\ell=1}^k n_{\ell} = N$. It will be supposed that the class probabilities are functions of s parameters, $\theta_1, \theta_2, \dots, \theta_s$, sufficiently smooth to satisfy the requirements of Cramér (1954, para. 30.3). Cramér then shows that the maximum likelihood estimators, $\hat{\theta}_1, \dots, \hat{\theta}_s$, or any estimators asymptotically equivalent to them, satisfy

$$\hat{\theta} - \theta = \frac{1}{\sqrt{N}} (B'B)^{-1} B'x + o_p\left(\frac{1}{\sqrt{N}}\right) \quad (2.1)$$

where $\hat{\theta}, \theta, x$ are column vectors of components $\hat{\theta}_i, \theta_i$, and $\frac{n_{\ell} - Np_{\ell}(\theta)}{\sqrt{Np_{\ell}(\theta)}}$ respectively, where

$$B = \begin{bmatrix} \frac{1}{\sqrt{p_1(\theta)}} \left(\frac{\partial p_1(\theta)}{\partial \theta_1} \right) & \dots & \frac{1}{\sqrt{p_1(\theta)}} \left(\frac{\partial p_1(\theta)}{\partial \theta_s} \right) \\ \dots & \dots & \dots \\ \frac{1}{\sqrt{p_k(\theta)}} \left(\frac{\partial p_k(\theta)}{\partial \theta_1} \right) & \dots & \frac{1}{\sqrt{p_k(\theta)}} \left(\frac{\partial p_k(\theta)}{\partial \theta_s} \right) \end{bmatrix} \quad (2.2)$$

and where $o_p(\frac{1}{\sqrt{N}})$ is any quantity which, when multiplied by $\sqrt{N}$, tends to zero in probability as $N \rightarrow \infty$. It will be noted that the expected value of $(B'B)^{-1}B'x$ is zero so that $\sqrt{N} E(\hat{\theta} - \theta) \rightarrow 0$ as $N \rightarrow \infty$. The same conditions ensure that the covariance matrix of $\hat{\theta} - \theta$ is of order N^{-1} . We will discuss later the effect of having estimators whose bias is $O(N^{-1/2})$ and also the effect of using estimators whose covariance matrix is $o(\frac{1}{N})$.

The next stage of Cramér's analysis, referred to above, is to show that the terms in

$$X^2 = \sum_{\ell=1}^k \frac{[n_{\ell} - Np_{\ell}(\hat{\theta})]^2}{Np_{\ell}(\hat{\theta})},$$

i.e.

$$= \sum_{\ell=1}^k y_{\ell}^2$$

may be written

$$y_{\ell} = \frac{n_{\ell} - Np_{\ell}(\theta) - N \sum_{r=1}^s (\hat{\theta}_r - \theta) \frac{\partial p_{\ell}(\theta)}{\partial \theta_r}}{\sqrt{Np_{\ell}(\theta)}} + o_p(1)$$

i.e.

$$y = x - \sqrt{N} B(\hat{\theta} - \theta) + o_p(1) \quad (2.3)$$

Now using (2.1), it is seen that

$$y = (I - B(B'B)^{-1}B')x + o_p(1)$$

Since $X^2 = y'y$, the fact that the asymptotic joint distribution of the x_{ℓ} is multivariate normal with mean zero and a covariance matrix $I - (\sqrt{p})(\sqrt{p})'$ where $(\sqrt{p})' = [\sqrt{p_1}, \dots, \sqrt{p_k}]$ and $B'\sqrt{p} = 0$, leads to X^2 being distributed asymptotically as a sum of squares of variables which are normal with zero means and the idempotent covariance matrix

$$I - \sqrt{p} \sqrt{p}' = B(B'B)^{-1}B' \quad (2.4)$$

Since the matrix (2.4) has a trace of $k-s-1$, the well known result, that X^2 is asymptotically distributed as a chi-square variate with $k-s-1$ degrees of freedom, is proven.

The simplicity of this result arises from being able to neglect certain terms and from the occurrence of the same matrix B in (2.1) and (2.3). Taking up the latter point first suppose that different estimators, $\theta_1^*, \dots, \theta_s^*$, say, had been used but that we can still have

$$\theta^* \rightarrow \theta = \sqrt{\frac{1}{N}} Ax + o_p\left(\frac{1}{\sqrt{N}}\right) \quad (2.5)$$

with $E(Ax) = 0$; the covariance matrix of the θ^* will still be $O(N^{-1})$. In this case, (2.3) is replaced by

$$y = x - \sqrt{N} B(\theta^* - \theta) + o_p(1) \quad (2.6)$$

so that

$$y = (I - BA)x + o_p(1)$$

and $X^2 = y'y$ is distributed as the sum of squares of variables which are asymptotically normal with zero means but with a covariance matrix given by

$$(I - BA) (I - (\sqrt{p})(\sqrt{p})') (I - A'B')$$

which is equal to

$$I - \sqrt{p}\sqrt{p}' - BA(I - \sqrt{p}\sqrt{p}') - (I - \sqrt{p}\sqrt{p}') A'B' + BA(I - \sqrt{p}\sqrt{p}')A'B' \quad (2.7)$$

The latent roots λ_i of (2.7) are no longer obvious. They may however be studied using a device due to Barton (1956). Defining $\mu_i = 1 - \lambda_i$, the μ_i are the latent roots of a matrix which may be written as

$$\sqrt{p}\sqrt{p}' + B \left\{ A(I - \sqrt{p}\sqrt{p}') - \frac{1}{2} J^{*-1} B' \right\} + \left\{ (I - \sqrt{p}\sqrt{p}') A' - \frac{1}{2} B J^{*-1} \right\} B',$$

where we have written the asymptotic covariance matrix of the θ^*

$$J^{*-1} = A(I - \sqrt{p}\sqrt{p}')A' \quad (2.8)$$

in terms of the information matrix for these estimators. Hence the problem of finding the latent roots of (2.7) is reduced to finding those of

$$\begin{bmatrix} \sqrt{p'} \\ A(I - \sqrt{p/p'}) - \frac{1}{2} J^{*-1} B' \\ B' \end{bmatrix} [\sqrt{p}, B, (I - \sqrt{p/p'})A' - \frac{1}{2} B J^{*-1}]$$

i.e. of the $(2s+1) \times (2s+1)$ matrix

$$\begin{bmatrix} 1, & 0, & 0 \\ 0, & AB - \frac{1}{2} J^{*-1} \hat{J}, & J^{*-1} - \frac{1}{2} AB J^{*-1} - \frac{1}{2} J^{*-1} B' A' + \frac{1}{4} J^{*-1} \hat{J} J^{*-1} \\ 0, & \hat{J}, & B' A' - \frac{1}{2} \hat{J} J^{*-1} \end{bmatrix}$$

using again that $B'\sqrt{p} = 0$ and writing $B'B = \hat{J}$, the information matrix for the multinomial maximum likelihood estimators, $\hat{\theta}$. Thus $k-2s-1$ of the λ_i are unity, one is zero, and the remaining $2s$ are the latent roots of

$$\begin{bmatrix} AB - \frac{1}{2} J^{*-1} \hat{J}, & J^{*-1} - \frac{1}{2} AB J^{*-1} - \frac{1}{2} J^{*-1} B' A' + \frac{1}{4} J^{*-1} \hat{J} J^{*-1} \\ \hat{J}, & B' A' - \frac{1}{2} \hat{J} J^{*-1} \end{bmatrix} \quad (2.9)$$

Hence asymptotically

$$X^2 = X_{k-2s-1}^2 + \sum_{i=1}^{2s} \lambda_i y_i^2$$

where the y_i are standard normal deviates independent of X_{k-2s-1}^2 and each other. This result is intimately related to Theorem 1 of Barton (1956) on the distribution of Neyman's smooth goodness-of-fit statistic when the parameters are estimated.

If $A = (B'B)^{-1}B'$ so that $\theta^* = \hat{\theta}$, $BA = B J^{*-1} B'$, and $J^* = \hat{J}$, (2.9) becomes

$$\begin{bmatrix} \frac{I_s}{2}, & \frac{\hat{J}^{-1}}{4} \\ \hat{J}, & \frac{I_s}{2} \end{bmatrix}$$

which may easily be shown (since it is idempotent and has a trace of s) to have s zero and s unit latent roots. We thus return to the classical result. In the most usual case where the correct (i.e. frequency maximum likelihood) estimators are not used, they are replaced by the sample maximum likelihood estimators. In this case the results of Chernoff and Lehmann (1954) imply that s of the roots of (2.9) are unity and the other s roots are all between zero and one. For example, in the case of a continuous distribution with density function $f(x, \theta_1, \dots, \theta_s)$ and fixed class boundaries $-\infty < Z_1 < \dots < Z_{k-1} < \infty$ it is easily seen that

$$A = J^{*-1} \begin{bmatrix} \int_{-\infty}^{Z_1} \frac{1}{\sqrt{p_1(\theta)}} \frac{\partial f}{\partial \theta_1} dx, & \dots, & \int_{Z_{k-1}}^{\infty} \frac{1}{\sqrt{p_k(\theta)}} \frac{\partial f}{\partial \theta_1} dx \\ \int_{-\infty}^{Z_1} \frac{1}{\sqrt{p_1(\theta)}} \frac{\partial f}{\partial \theta_s} dx, & \dots, & \int_{Z_{k-1}}^{\infty} \frac{1}{\sqrt{p_k(\theta)}} \frac{\partial f}{\partial \theta_s} dx \end{bmatrix}$$

$$= J^{*-1} B',$$

and

$$J^* = \begin{bmatrix} \int_{-\infty}^{\infty} \left(\frac{\partial f}{\partial \theta_1} \right)^2 \frac{dx}{f}, & \dots, & \int_{-\infty}^{\infty} \left(\frac{\partial f}{\partial \theta_1} \right) \left(\frac{\partial f}{\partial \theta_s} \right) \frac{dx}{f} \\ \dots & & \dots \\ \int_{-\infty}^{\infty} \left(\frac{\partial f}{\partial \theta_s} \right) \left(\frac{\partial f}{\partial \theta_1} \right) \frac{dx}{f}, & \dots, & \int_{-\infty}^{\infty} \left(\frac{\partial f}{\partial \theta_s} \right)^2 \frac{dx}{f} \end{bmatrix}$$

Thus $AB = J^{*-1} \hat{J}$ and to find the latent roots it is better to return from (2.9) to (2.7) which becomes simply

$$I - \sqrt{p/p'} - B J^{*-1} B'.$$

The latent roots of this matrix are $k-s-1$ unities, one zero and the s roots of the determinantal equation

$$\left| \hat{J} - (1-\lambda) J^* \right| = 0,$$

which lie between zero and one because the θ^* contain more information than the $\hat{\theta}$.
Hence

$$X^2 = \chi_{k-s-1}^2 + \sum_{i=1}^s \lambda_i y_i^2,$$

the result of Chernoff and Lehmann and the analogue of Theorem II of Barton. A similar result is obtained when a discrete distribution, such as a Poisson distribution, is grouped.

Thus the distribution of X^2 when the correct estimators are not used is in general not a χ^2 and contains the unknown parameters θ . Nothing is known of the magnitudes of the $2s$ latent roots λ_i in the most general case where they must be found by (2.9). When the multinomial has been obtained by grouping a discrete or continuous distribution and the sample maximum likelihood estimators are used, s of these roots are unity and the other s roots will decrease as the grouping is made finer and finer since then $\hat{J} \rightarrow J$, so that finally $X^2 \sim \chi_{k-s-1}^2$. In the general case one would presume that the roots of (2.9) fall usually into two groups -- s roots near unity and s roots near zero. It follows from (2.15) below that the sum of the $2s$ roots is always greater than s .

We see that (2.3) would still be true for estimators which converge more rapidly to θ . But if $\sqrt{N}(\theta^* - \theta) = o_p(1)$, y is asymptotically equivalent to x , so that $X^2 \sim \chi_{k-1}^2$ i.e. estimation of parameters has no effect at all on the distribution of X^2 . Proceeding in the other direction suppose that the estimators θ^* of type (2.5) with the additional term $bN^{-1/2}$ -- i.e. the $E(\theta^* - \theta)$ is not $o(N^{-1/2})$ as before but $O(N^{-1/2})$ -- and a covariance matrix of $O(N^{-1})$ as before. Then it follows from (2.6), that the y still have an asymptotically normal distribution with covariance matrix (2.7) but with a non-zero mean vector -- $(\sqrt{p})Bb$, where $\sqrt{p}$ is a diagonal matrix of elements $\sqrt{p_1}, \dots, \sqrt{p_k}$. Thus X^2 tends to non-central quadratic

form. This is the full analogue of Barton's Theorem 1, except that he deals with the non-null case.

We may now discuss the partitioning of X^2 i.e. we wish to consider linear forms in the frequencies

$$L_i = \sum_{\ell=1}^k L_{i\ell} \frac{n_{\ell} - Np_{\ell}(\theta^*)}{\sqrt{Np_{\ell}(\theta^*)}} \quad , \quad (2.10)$$

the sum of whose squares adds up exactly to X^2 . Choosing

$$L_{k\ell} = \sqrt{p_{\ell}(\theta^*)} \quad (\ell=1, \dots, k)$$

it is seen that $L_k = 0$ identically. If the other forms are chosen to be normalized and orthogonal to L_k i.e. so that, if $L_i^* = [L_{i1}, \dots, L_{ik}]$

$$\left. \begin{aligned} L_i^* (\sqrt{p}) &= 0 \quad (i=1, \dots, k-1) \\ L_i^* L_j &= \delta_{ij} \quad (i, j=1, \dots, k-1) \end{aligned} \right\} \quad (2.11)$$

then

$$X^2 = L_1^2 + \dots + L_{k-1}^2 \quad . \quad (2.12)$$

The conditions (2.10) mean that the coefficients $L_{i\ell}$ are functions of θ^* . However, for the same reason as we may take the denominators of elements of X^2 to be their true values, we may suppose, for asymptotic results, that the $L_{i\ell}$ take their true values i.e. the first equation of (2.11) may be taken as $\sum L_{i\ell} \sqrt{p_{\ell}}(\theta) = 0$. Thus the joint distribution of any subset of the L_i is asymptotically multivariate normal with zero means and, [assuming that θ^* is the estimator first envisaged in (2.5)], with covariances given by

$$\text{cov}(L_i, L_j) = L_i^* V L_j$$

i.e.

$$\text{cov}(L_i, L_j) = \delta_{ij} - L_i' B A L_j - L_i' A' B' L_j + L_i' B J^{*-1} B' L_j \quad (2.13)$$

after using (2.7) and (2.11). Since $(\sqrt{p_1}, \dots, \sqrt{p_k})$ is a latent vector of V , the coefficient vectors L_i can be taken, if suitable, as the latent vectors of V . If such a choice is made $L_i' V L_j = 0$ ($i \neq j$) and $L_i' V L_i = \lambda_i$, the latent root of V associated with L_i . Hence, in the general case, $k - 2s - 1$ forms will have unit variance and $2s$ forms will have variances $\lambda_1, \dots, \lambda_{2s}$ which are the roots of (2.9). In fact the last $2s$ forms can be used "to partition out" the disturbance due to estimation because, if the sum of their squares is subtracted from X^2 , the result is a statistic which is distributed as χ_{k-2s-1}^2 . However the loss of degrees of freedom is s more than it need be.

Suppose now that the last s forms are defined by the linear forms used in (2.1), namely by

$$\begin{aligned} L &= [L_{k-s}, \dots, L_{k-1}], \\ &= \hat{B} J^{-1/2}, \end{aligned} \quad (2.14)$$

so that, as required by (2.11), $L' L = I_s$ and $L' \sqrt{p} = 0$. Then

$$\begin{aligned} L' V L &= \hat{J}^{-1/2} B' V B \hat{J}^{-1/2} \\ &= I_s - \hat{J}^{1/2} A B \hat{J}^{-1/2} - \hat{J}^{-1/2} B' A' \hat{J}^{1/2} + \hat{J}^{1/2} J^{*-1} \hat{J}^{1/2}, \end{aligned}$$

whose trace is given by

$$s - 2 \text{ trace } (AB) + \text{trace } (\hat{J} J^{*-1})$$

But

$$\begin{aligned} \text{trace } V &= k - \left\{ 1 + 2 \text{ trace } (AB) - \text{trace } (\hat{J} J^{*-1}) \right\} \\ &= (k-s-1) + \text{trace } (L' V L). \end{aligned} \quad (2.15)$$

If then forms $L, \dots, L_{k-s-1}$ are constructed so that with the above definition of forms $L_{k-s}, \dots, L_{k-1}$, (2.11) is satisfied, it is seen that

$$\sum_{i=1}^{k-s-1} \text{var} (L_i) = k-s-1 ,$$

so that it would appear that $X^2 - \sum_{i=k-s}^{k-1} L_i^2 = L_1^2 + \dots + L_{k-s-1}^2$ is now distributed as χ_{k-s-1}^2 i.e. that the disturbing effect of incorrect estimation has been partitioned out. For this to be true we must show that a typical pair of vectors L_i, L_j ($i, j=1, \dots, k-s-1$) satisfying (2.11) and $\hat{J}^{-1/2} B' L_i = 0$

$$\text{i.e.} \quad B' L_i = 0 \quad (2.16)$$

$$\text{also give} \quad L_i' V L_j = \delta_{ij} \quad (2.17)$$

But this is seen to be true from the form (2.7) for V .

Since only $k-2s-1$ latent roots of V are necessarily unity, it seems strange that the disturbing effect of incorrect estimation on X^2 can be partitioned out with only s linear forms. However this choice of the last s degrees of freedom served as the basis for Fisher's (1928) most general account of the effect of estimation on the distribution of X^2 . It can be shown that the subtraction of these terms is equivalent to calculating X^2 with the estimators which result from one cycle of iteration to the maximum likelihood equations, using the estimators θ^* as the starting values. This works because in large samples one iteration should be enough.

This completes the asymptotic investigation of the effect of estimation on the partitioning of X^2 . If the linear forms of interest satisfy (2.11) and (2.16), they may be taken, on the null hypothesis, as independent standard normal deviates. If this is not so, (2.13) must be used to compute their variances and they may not be independent. When the correct estimators are used so that $\theta^* = \hat{\theta}$, (2.13) reduces to

$$\text{Cov} (L_i, L_j) = \delta_{ij} - L_i' B \hat{J}^{-1} B' L_j , \quad (2.18)$$

so that it is still necessary to insist on (2.16) being satisfied if the linear forms are to have unit variance.

The theory of this section is adequate to deal with true multinomial situations, cases when a discrete distribution has been grouped (for example, the Poisson distribution), and the case of a continuous distribution which has been grouped by using intervals that are fixed for all samples. Chernoff and Lehmann (1954) give two worked examples showing the effect of using sample rather than frequency maximum likelihood estimators when testing normal and Poisson distributions. In the first case the effect may be large and in the second case it will rarely be appreciable. But, in fact, one should never use such estimators because it is very little trouble to find the correct ones, using the sample maximum likelihood estimators as starting values in the iterative solution of the frequency maximum likelihood equations.

Cochran (1954, 1955) has considered the variances of linear forms in the class frequencies. His formulae are particular cases of those given here. It should be noted that his "variance chi-square" is not of the type considered here and so the above remarks on estimation do not apply to it. However in his ordinary goodness-of-fit X^2 , it does appear as if he has used the sample estimators though, in his cases, the perturbation would be small.

3. Multinomial Distributions - More Exact Theory

The above results are all asymptotic. For finite N , the distribution of X^2 or any part of it is discrete and depends on the unknown parameters even if maximum likelihood estimation is used. The latter difficulty may be overcome if $\theta_1, \dots, \theta_s$ admit sufficient statistics. Generalizing a result of Gani (1955), we may suppose that, for $s \leq r < k-1$, the class probabilities have the form,

$$p_{\ell}(\theta) = \frac{\alpha_{\ell} \prod_{j=1}^k [\lambda_j(\theta)]^{K_{j\ell}}}{\sum_{\ell=1}^k \alpha_{\ell} [\lambda_j(\theta)]^{K_{j\ell}}} \quad (3.1)$$

where it may, without loss of generality, be assumed $\sum_{\ell=1}^k \alpha_{\ell} = 1$.^{*} Then the joint distribution of $n_1, \dots, n_k$ given $K_1 = \sum_{\ell} K_{1\ell} n_{\ell}, \dots, K_r = \sum_{\ell} K_{r\ell} n_{\ell}$ is the same as if the $n_1, \dots, n_k$ were drawn, under the same restrictions, from the multinomial population with class probabilities $\alpha_1, \dots, \alpha_k$. Since the $K_{1\ell}, \dots, K_{r\ell}, \alpha_{\ell}$ ($\ell=1, \dots, k$) are independent of $\theta_1, \dots, \theta_s$, the conditional distribution of any function (e.g. X^2) of $n_1, \dots, n_k$ for given values of $K_1, \dots, K_r$ is independent of $\theta_1, \dots, \theta_s$. The restrictions may so drastically limit the number of possible samples that it may be easily possible to compute, for each such sample, the value of the statistic and the probability in the conditional distribution and so to perform an "exact" test in the same way as is done in the well known case of 2 x 2 tables. Such tests are "similar" in the Neyman-Pearson sense since the Type I error rate will not depend on the unknown parameters. Unless some modification, such as randomization, is used, discreteness will prevent the attainment of a preassigned Type I error rate.

If the number of possible values of the statistic is too large for an enumeration even of the "more significant samples", it seems worthwhile trying to approximate the conditional distribution of the statistic for N large. It seems to be assumed in the literature that the conditional distribution of

* A more elaborate discussion could be given using recent results of Halmos and Savage (1949) (for an elementary exposition, see Watson (1957a) but this would not be in keeping with our present point of view.

$$X^2 = \sum \frac{[n_\ell - N p_\ell(\hat{\theta})]^2}{N p_\ell(\hat{\theta})}, \quad (3.2)$$

$$= X_{M.L.}^2, \text{ say,}$$

given $K_1, \dots, K_r$, will be asymptotically χ_{k-r-1}^2 but no previous proof of this is known to the writer. Writing $K' = [K_1, \dots, K_r]$, the same applies to an alternative statistic

$$X^2 = \sum \frac{[n_\ell - E(n_\ell | K)]^2}{E(n_\ell | K)}, \quad (3.3)$$

$$= X_R^2,$$

mentioned by Rao and Chakravathi (1956). The form of the conditional distribution of the $n_\ell, \dots, n_k$ given $K_1, \dots, K_r$, suggests the much simpler statistic

$$X^2 = \sum \frac{[n_\ell - N \alpha_\ell]^2}{N \alpha_\ell}. \quad (3.4)$$

The statistic (3.4) appears superior because it does not require the solution of the maximum likelihood equations, as does (3.2) and does not require the evaluation of the $E(n_\ell | K)$, as does (3.3).

The evaluation of the $E(n_\ell | K)$, or more generally the expectation of any function ϕ of the n_ℓ 's may be carried out by a generalization of the method of Rao and Chakravathi. From (3.1), the probability of observing K , $P_\lambda(K)$ say, is given by

$$P_\lambda(K) = \frac{\lambda_1^{K_1} \dots \lambda_r^{K_r}}{\sum_{\ell} \lambda_1^{K_{1\ell}} \dots \lambda_r^{K_{r\ell}}} P_1(K) \quad (3.5)$$

where

$$P_1(K) = \sum \frac{N!}{n_1! \dots n_k!} \gamma_1^{n_1} \dots \gamma_k^{n_k}, \quad (3.6)$$

the summation being over all samples which given simultaneously

$\sum K_{1\ell} n_\ell = K_1, \dots, \sum K_{r\ell} n_\ell = K_r$. Writing $E(\phi|K)$ for the conditional mean of ϕ given K ,

$$\sum E(\phi|K) P_\lambda(K) = E(\phi), \quad (3.7)$$

where the summation is over all possible values of K . Combining (3.7) with (3.5) and rearranging, we find

$$\sum E(\phi|K) P_1(K) \lambda_1^{K_1} \dots \lambda_r^{K_r} = E(\phi) \left(\sum \gamma_\ell \lambda_1^{K_{1\ell}} \dots \lambda_r^{K_{r\ell}} \right)^N. \quad (3.8)$$

Since the $K_{1\ell}, \dots, K_{rk}$ may be taken as integers, it follows that

$$E(\phi|K) = \text{coeff. of } \lambda_1^{K_1} \dots \lambda_r^{K_r} \text{ in } E(\phi) \left(\sum \gamma_\ell \lambda_1^{K_{1\ell}} \dots \lambda_r^{K_{r\ell}} \right)^N. \quad (3.9)$$

It will, however, often be difficult to evaluate (3.9).

The asymptotic distribution of (3.4) may be determined by methods similar to those used in section 2. It may be assumed that for large N , $n_1, \dots, n_k$, $\sum K_{1\ell} n_\ell, \dots, \sum K_{r\ell} n_\ell$ have a multivariate normal distribution with means and covariances determined from the unconditional multinomial distribution $(\gamma_1 + \dots + \gamma_k)^N$. It is also necessary to consider a sequence of values for K which converge to $E(K)$ at the proper rate. Since

$$\begin{aligned} E(K_i) &= N \sum K_{i\ell} \gamma_\ell, \\ \text{cov}(K_i, K_j) &= N \left[\sum K_{i\ell} K_{j\ell} \gamma_\ell - \left(\sum K_{i\ell} \gamma_\ell \right) \left(\sum K_{j\ell} \gamma_\ell \right) \right] \\ &= N V_{ij}, \text{ say,} \end{aligned} \quad (3.10)$$

we will assume that the sequence of values is such that

$$[K-E(K)]' V^{-1} [K-E(K)] \quad (3.11)$$

converges to a non-zero finite limit. Unconditionally, the quantity (3.11) is asymptotically a χ^2_r variable. On this basis, it is easy to show that, asymptotically,

$$X^2 \sim \chi^2_{k-r-1} + N^{-1} [K-E(K)]' V^{-1} [K-E(K)] \quad (3.12)$$

This suggests that in practice the statistic

$$X^2_c = X^2 - N^{-1} [K-E(K)]' V^{-1} [K-E(K)] \quad (3.13)$$

should be referred to the χ^2 tables with $k-r-1$ degrees of freedom. This result is of some general interest. Roy and Mitra (1955) mention that Mitra has found a case where (3.4) is not distributed as a χ^2_{k-r-1} . Our result is a possible answer to their problem.

The result (3.12) may be rewritten, by analogy with the identity in regression analysis,

$$\text{Total S.S.} = \text{Error S.S.} + \text{Regression S.S.} .$$

To find a vector β_ℓ so that $n_\ell - N\alpha_\ell - \beta_\ell' [K-E(K)]$ is uncorrelated with $K-E(K)$, we must solve the equations

$$E \left\{ [n_\ell - N\alpha_\ell] [K-E(K)]' \right\} = \beta_\ell' V N \quad (3.14)$$

But

$$\begin{aligned} E \left\{ [n_\ell - N\alpha_\ell] \left[\sum K_{r\ell} n_\ell - N \sum K_{r\ell} \alpha_\ell \right] \right\} \\ = N\alpha_\ell \left\{ K_{r\ell} - \sum K_{r\ell} \alpha_\ell \right\} , \end{aligned}$$

so that

$$\beta_\ell' = \alpha_\ell [K_{1\ell} - \sum K_{1\ell} \alpha_\ell, \dots, K_{r\ell} - \sum K_{r\ell} \alpha_\ell] V^{-1} . \quad (3.15)$$

Now

$$\begin{aligned}
 & \sum_{\ell=1}^k \frac{1}{N\alpha_{\ell}} \left\{ n_{\ell} - N\alpha_{\ell} - \beta_{\ell}' [K-E(K)] \right\}^2, \\
 &= \sum_{\ell=1}^k \frac{(n_{\ell} - N\alpha_{\ell})^2}{N\alpha_{\ell}} - 2 \sum_{\ell=1}^k \frac{(n_{\ell} - N\alpha_{\ell}) \beta_{\ell}' [K-E(K)]}{N\alpha_{\ell}}, \\
 &+ \sum_{\ell=1}^k \frac{\{ \beta_{\ell}' [K-E(K)] \}^2}{N\alpha_{\ell}}, \\
 &= \sum_{\ell=1}^k \frac{(n_{\ell} - N\alpha_{\ell})^2}{N\alpha_{\ell}} - N^{-1} [K-E(K)]' V^{-1} [K-E(K)], \\
 &= x_C^2.
 \end{aligned}$$

Thus we have shown that

$$E(n_{\ell} | K) \sim N\alpha_{\ell} + \beta_{\ell}' [K-E(K)], \quad (3.16)$$

and that

$$x_C^2 \sim \sum \frac{[n_{\ell} - E(n_{\ell} | K)]^2}{N\alpha_{\ell}}. \quad (3.17)$$

However the last term of (3.16) is $O(\sqrt{N})$ so that if (3.16) is used in (3.3), the $E(n_{\ell} | K)$ in the denominator may be replaced asymptotically by $N\alpha_{\ell}$ so that (3.3), i.e. x_R^2 , becomes equivalent to (3.17) i.e. x_C^2 and so is also asymptotically x_{k-r-1}^2 . It is worth noting that this result implies the usual χ^2 statistic for two way tables is asymptotically χ^2 when the marginal totals are kept fixed.

It remains to verify that $x_{M,L}^2$ is asymptotically x_{k-r-1}^2 in the conditional distribution with K fixed. However a stronger result is true. Returning to the beginning of Section 2, it is seen that $x_{M,L}^2 = y'y$ where the elements of the vector y given by

$$y = (I - B(B'B)^{-1}B')x + o_p(1)$$

are asymptotically normal and have an idempotent covariance matrix and zero means. Moreover the asymptotic joint distribution of y and $\sqrt{N}(\hat{\theta} - \theta)$ is multivariate normal since, by (2.1)

$$\sqrt{N}(\hat{\theta} - \theta) = (B'B)^{-1}B'y + o_p(1).$$

The asymptotic distribution of $X_{M.L.}^2$ given $\hat{\theta}$ must be computed using the distribution of y for fixed $\hat{\theta}$. But y and $\hat{\theta}$ are asymptotically independent since

$$\begin{aligned} E \{ y' N(\hat{\theta} - \theta) \} &= E \{ (I - B(B'B)^{-1}B')' y' B(B'B)^{-1} \}, \\ &= (I - B(B'B)^{-1}B') (I - \sqrt{p} \sqrt{p'}) B(B'B)^{-1}, \\ &= 0, \end{aligned}$$

since $\sqrt{p'}B = 0$. Hence $X_{M.L.}^2$ is unaffected by fixing $\hat{\theta}$, and so is still χ_{k-s-1}^2 . Here it is not assumed that the $p_\theta(\theta)$ are defined as in (3.1) - the result follows from the fact that maximum likelihood estimators have the sufficiency property in large samples.

Thus if $r=s$ and $N \rightarrow \infty$, $X_{M.L.}^2$, X_R^2 and X_C^2 are equivalent and are distributed as χ_{k-s-1}^2 . They may however differ in small samples and one may be better approximated by χ_{k-s-1}^2 than the others. To investigate this, we consider one of Gani's examples:

$$p_1 = \theta, \quad p_2 = 2\theta, \quad p_3 = 1 - 3\theta, \quad (0 < \theta < 1/3).$$

This has the form (3.1) with $s=r=1$, $\alpha_1 = \frac{1}{4}$, $\alpha_2 = \frac{1}{2}$, $\alpha_3 = \frac{1}{4}$, $K_{11} = 1$, $K_{12} = 1$, $K_{13} = 0$. Thus $K_1 = K = n_1 + n_2$, $(\theta) = \theta(1-3\theta)^{-1}$. Suppose that we wish to make an χ^2 goodness of fit test of the sample $n_1 = 2$, $n_2 = 9$, $n_3 = 9$ so that $N = 20$, $K = 11$. In the conditional distribution the possible samples are $(0, 11, 9)$, $(1, 10, 9)$, ..., $(11, 0, 9)$ and the associated values of X_C^2 , $X_{M.L.}^2$ and P , the probability of obtaining the sample from the conditional distribution, are given in Table 1. It is easily verified that here

$$E(n_1|K) = Np_1(\hat{\theta}) = \frac{K}{3}, E(n_2|K) = Np_r(\hat{\theta}) = \frac{2K}{3}$$

so that $X_{R.}^2 = X_{M.L.}^2$.

Table 1

n_1	n_2	n_3	n	X_c^2	X_{ML}^2	X_{LR}^2
0	11	9	.011 556	4.033	5.500	8.920
1	10	9	.063 636	2.133	2.909	3.605
2	9	9	.158 943	0.833	1.136	1.262
3	8	9	.238 396	0.133	.182	0.188
4	7	9	.238 433	0.033	.046	0.046
5	6	9	.166 888	0.533	.727	0.693
6	5	9	.083 444	1.633	2.227	2.079
7	4	9	.029 797	2.333	4.546	4.204
8	3	9	.007 450	5.633	7.682	7.119
9	2	9	.001 242	8.533	11.637	10.966
10	1	9	.000 259	12.033	16.409	16.081
11	0	9	.000 006	16.133	22.000	24.169

Thus the chance of X_c^2 exceeding the tabular 5% (1%) point of X_1^2 , 3.841 (6.635) is 0.0205 (.0015) whereas for X_{ML}^2 the corresponding figures are 0.0504 and 0.00895. So the X_c^2 statistic is inferior to the X_{ML}^2 statistic. For comparison the likelihood ratio statistic

$$X_{LR}^2 = 2 \sum n_\ell \log_\ell \left(\frac{n_\ell}{Np_\ell(\hat{\theta})} \right)$$

have been computed and is seen to be very similar to X_{ML}^2 .

Finally we mention approximations to exact unconditional distributions. Hoel (1938) first found an asymptotic expansion for the distribution of X^2 . Darwin (1958) has reconsidered the problem and tried to find an asymptotic expansion with discontinuous factors. Good (1957) has given a general method

and a number of results in this difficult field. A much simpler task is to find an asymptotic expansion for the distribution of a linear function of multinomial frequencies. Cochran (1956) has given a fully worked numerical trinomial example. From his results it is clear that the normal approximation is surprisingly good for the two-tailed test and that its inadequacy comes more from the discontinuities of the true distribution than from its failure to follow the overall pattern of the true distribution. Corrections could easily be made for skewness and kurtosis by using a series in the first four Hermite polynomials. A proper treatment requires a study of the relation

$$\text{Prov} \left(\sum_{\ell=1}^k a_{\ell} n_{\ell} \geq A \right) = \frac{1}{2\pi i} \int_C \frac{\left(\sum_{\ell} p_{\ell} \lambda^{a_{\ell}} \right)^N}{\lambda^A (\lambda-1)} d\lambda,$$

where the a_{ℓ} and A are all integers and c is a closed curve about the origin and on which $|\lambda| > 1$.

4. Continuous Distributions

Given a sample $x_1, \dots, x_N$ from a continuous distribution with density $f(x, \theta_1, \dots, \theta_s) = f(x, \theta)$, the parameters $\theta_1, \dots, \theta_s$ being unknown, an χ^2 goodness-of-fit test may be made in several ways. The range of x may be split up into k segments by $g_0 = -\infty < g_1 < \dots < g_{k-1} < g_k = +\infty$ and the number of x_i in each class found to be $n_1, n_2, \dots, n_k$ say. Estimators $\theta_1^*, \dots, \theta_s^*$ for $\theta_1, \dots, \theta_s$ may be found in some way and

$$\chi^2 = \sum \frac{[n_{\ell} - Np_{\ell}(\theta^*)]^2}{Np_{\ell}(\theta^*)} \quad (4.1)$$

computed. If the class boundaries $g_1, \dots, g_{k-1}$ are fixed without reference to the sample, the class frequencies are multinomial with class probabilities

$$p_{\ell}(\theta) = \int_{g_{\ell-1}}^{g_{\ell}} f(x, \theta) dx \quad (4.2)$$

and the theory of section 2 may be applied. In practice the class intervals are to some extent the result of an examination of the data, though it is perhaps not often that they are strictly defined by it, a fact that is usually ignored. It may even be possible to define variable intervals so that the distribution of X^2 , using the sample maximum likelihood estimators, is asymptotically free of the unknown θ -- it is intuitively evident that this is so when the parameters are those of scale and location. This question has been examined by A. R. Roy (1956), G. S. Watson (1957, 1958). As Roy's treatment is both more general and more rigorous, it will be sketched below.

We will use the notation that, for a function $h(\theta)$,

$$\begin{bmatrix} \frac{\partial h(\theta)}{\partial \theta_1} \\ \vdots \\ \frac{\partial h(\theta)}{\partial \theta_s} \end{bmatrix} = \frac{\partial h(\theta)}{\partial \theta}, \left[\frac{\partial h(\theta)}{\partial \theta_1}, \dots, \frac{\partial h(\theta)}{\partial \theta_s} \right] = \frac{\partial h(\theta)}{\partial \theta^t}.$$

The result of Cramér (1954, ch. 32) for regular maximum likelihood estimators may then be written, analogously with (2.1),

$$\hat{\theta} - \theta = \frac{1}{N} \sum_{i=1}^N J^{-1} \frac{\partial \log f(x_i, \theta)}{\partial \theta} + o_p\left(\frac{1}{\sqrt{N}}\right) \quad (4.3)$$

and where J is the information matrix with (i, j) element

$$J_{ij} = E \left\{ - \frac{\partial^2 \log f(x, \theta)}{\partial \theta_i \partial \theta_j} \right\} \quad (4.4)$$

It is well known that

$$E \left\{ \frac{\partial \log f(x, \theta)}{\partial \theta} \right\} = 0 \quad (4.5)$$

where 0 is a zero vector. Roy considers, analogously with (2.5), the class of estimators defined by

$$\theta^* - \theta = \frac{1}{N} \sum_{i=1}^N h(x_i) + o_p\left(\frac{1}{\sqrt{N}}\right) \quad (4.6)$$

where $h(x_i)$ is a vector valued function of x with s components such that

$$E \{ h(x) \} = 0 \quad (4.6)$$

and such that

$$E \{ h(x) h'(x) \} = \Sigma, \quad (4.7)$$

the $s \times s$ covariance matrix of the components of $h(x)$. Clearly (4.6) reduces to (4.3) by putting

$$h(x) = J^{-1} \frac{\partial}{\partial \theta} \log f(x, \theta) \quad (4.8)$$

and then

$$\Sigma = J^{-1}$$

As in section 2.2, the investigation consists of showing that the

$$y_\ell = \frac{n_\ell - N p_\ell(\theta^*)}{\sqrt{N p_\ell(\theta^*)}} \quad (4.10)$$

have a joint multivariate normal distribution with zero means and finding their covariance matrix. As before the $p_\ell(\theta^*)$ in the denominator of (4.10) may be replaced by $p_\ell(\theta)$ and the $p_\ell(\theta^*)$ in the numerator taken to be

$$p_\ell(\theta) + (\theta^* - \theta)' \frac{\partial}{\partial \theta} p_\ell(\theta), \quad (4.11)$$

where, by (4.2), with g_ℓ replaced by $g_\ell(\theta)$

$$\frac{\partial p_\ell(\theta)}{\partial \theta} = \int_{g_{\ell-1}(\theta)}^{g_\ell(\theta)} \frac{\partial f(x, \theta)}{\partial \theta} dx + \left\{ f(g_\ell(\theta), \theta) \frac{\partial g_\ell(\theta)}{\partial \theta} - f(g_{\ell-1}(\theta), \theta) \frac{\partial g_{\ell-1}(\theta)}{\partial \theta} \right\},$$

i.e.

$$\frac{\partial p_\ell(\theta)}{\partial \theta} = u_\ell(\theta) + v_\ell(\theta), \quad \text{say.} \quad (4.12)$$

The n_ℓ are the number of x_ℓ 's in $\{g_{\ell-1}(\theta^*), g_\ell(\theta^*)\}$ and are no longer multinomial though their dominant part, the number of x_ℓ 's in $\{g_{\ell-1}(\theta), g_\ell(\theta)\}$, m_ℓ say, clearly is. The crux of Roy's method is his proof that

$$n_\ell - m_\ell = N v_\ell'(\theta^* - \theta) + o_p(\sqrt{N}). \quad (4.13)$$

Considering the case of θ scalar ($s=1$) to see the plausibility of this result, it would be reasonable to suppose that the number of x_ℓ 's between $g_\ell(\theta)$ and $g_\ell(\theta^*)$ to be approximately binomial with mean

$$N \left| \int_{g_\ell(\theta)}^{g_\ell(\theta^*)} f(x, \theta) dx \right| \approx N \left| (\theta^* - \theta) g_\ell'(\theta) \right| f(g_\ell(\theta), \theta)$$

The result would then follow from considering the normal approximation to binomial distribution wherein the probability of success is $O(N^{-1/2})$. This is the only point in the argument where it seems really necessary to assume that the random variable x is a scalar.

The rest of the discussion is similar to that of section 2. The y_ℓ of (4.10) is simplified by using (4.13), (4.11), (4.10) and then $\theta^* - \theta$ is eliminated by (4.6). The only additional task is the proof of asymptotic normality. This is achieved by expressing the m_ℓ as the sums of N indicator variables so that the column of y_ℓ 's is finally expressed as the sum of N independent vectors and hence a multivariate form of the central limit theorem gives the result. Writing

$$P = \begin{bmatrix} p_1(\theta) & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & p_k(\theta) \end{bmatrix}, \quad P^{1/2} = \begin{bmatrix} \sqrt{p_1(\theta)} & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & \sqrt{p_k(\theta)} \end{bmatrix},$$

$$p = \begin{bmatrix} p_1(\theta) \\ \vdots \\ p_k(\theta) \end{bmatrix}, \quad w_\ell(\theta) = \int_{g_{\ell-1}(\theta)}^{g_\ell(\theta)} h(x) f(x, \theta) dx,$$

$$W = [w_1(\theta), \dots, w_k(\theta)], \quad U = [u_1(\theta), \dots, u_k(\theta)]$$

$$\sum^* = P - pp^s - V^s W - W^s U + W^s \sum U,$$

the final result is that X^2 is asymptotically distributed as $\sum_{i=1}^k \lambda_i z_i^2$ where the λ_i are the latent roots of $P^{-1/2} \sum^* P^{-1/2}$. The discussion of the roots of $P^{-1/2} \sum^* P^{-1/2}$ is the same as that of (2.7). If the sample maximum likelihood estimators are used, so that (4.8) and (4.9) are satisfied, the s non-negative roots are between 0 and 1 and are usually functions of θ . If the parameters are those of scale and location, it is easily shown that this is not so when the functions $g_\ell(\theta)$ are chosen such that the $p_\ell(\theta)$ are constant.

This however does not represent the solution to the practical problem which may be stated better as follows: what is the effect on the distribution of X^2 , calculated by using maximum likelihood estimation with the class frequencies, when the class boundaries are to some extent, conditioned by the sample? If we can show that X^2 , so calculated, is asymptotically χ_{k-s-1}^2 then the intuitive procedure (i.e. ignoring the variability of the class boundaries) is proved to be valid.

For this purpose, we may take the formulation of Roy and use the class boundaries $g_\ell(\theta^*)$. However with the class frequencies so obtained, $n_1, \dots, n_k$, we would, in the intuitive procedure, carry on as follows. Define

$$p_\ell(\theta, \theta^*) = \int_{g_{\ell-1}(\theta^*)}^{g_\ell(\theta^*)} f(x, \theta) dx, \quad (4.14)$$

and then solve the equations

$$\sum_{\ell=1}^k \frac{n_\ell}{p_\ell(\theta, \theta^*)} \frac{\partial p_\ell(\theta, \theta^*)}{\partial \theta} = 0 \quad (4.15)$$

for θ , obtaining $\theta = \theta_f^*$, say. Finally

$$\chi^2 = \sum_{\ell=1}^k \frac{[n_\ell - N p_\ell(\theta_f^*, \theta^*)]^2}{N p_\ell(\theta_f^*, \theta^*)}, \quad (4.16)$$

could be computed and treated as χ_{k-s-1}^2 . The asymptotic distribution of (4.16) can be easily found by a trivial variation of Roy's discussion. In the present case, (4.13) remains true since only θ^* is relevant here. However (4.11) must be replaced by

$$\begin{aligned} p_\ell(\theta_f^*, \theta^*) &= p_\ell(\theta, \theta) + (\theta_f^* - \theta)' \left(\frac{\partial p_\ell(\theta_f^*, \theta^*)}{\partial \theta_f^*} \right)_{\theta_f^* = \theta, \theta^* = \theta} + \\ &+ (\theta^* - \theta)' \left(\frac{\partial p_\ell(\theta_f^*, \theta^*)}{\partial \theta^*} \right)_{\theta_f^* = \theta, \theta^* = \theta} \end{aligned} \quad (4.17)$$

By consideration of (3.14) and by comparison with (4.12), (4.17) becomes

$$p_\ell(\theta_f^*, \theta^*) = p_\ell(\theta, \theta) + (\theta_f^* - \theta)' u_\ell + (\theta^* - \theta)' v_\ell \quad (4.18)$$

so that finally

$$\frac{n_{\ell} - Np_{\ell}(\theta_f^*, \theta^*)}{\sqrt{Np_{\ell}(\theta_f^*, \theta^*)}} \sim \frac{m_{\ell} - Np_{\ell}(\theta, \theta) - N(\theta_f^* - \theta)' u_{\ell}}{\sqrt{Np_{\ell}(\theta, \theta)}}$$

But this is precisely the expression which would arise if one began with the class boundaries $g_{\ell}(\theta)$, found the m_{ℓ} , then forgot the true value of θ and proceeded to estimate and calculate the X^2 statistic in the correct way for fixed class intervals. So the intuitive method leads to the result that

$$X^2 \sim \chi_{k-s-1}^2$$

This last stage of the proof can, of course, be carried out formally by the methods given above.

The generality of the above argument seems to mean that in practice we may form our class intervals with full knowledge of the data and then proceed in the classical manner as though these intervals were known a priori. This seems intuitively plausible since any fixed set of class intervals leads to the same asymptotic distribution in the null case, though in the non-null case the powers will be different, and in the case of small samples the approximation to the χ^2 -distribution will be very different.

There is therefore no need to go beyond the theory given in section 2 -- in particular, the partitioning of X^2 requires no further discussion.

5. Some Metrical Aspects of X^2 Tests

In the preceding sections, only distributions on the null hypothesis have been considered. X^2 is not designed for specific alternatives, although partitions of X^2 - or non-independent parts of X^2 - are used often to test particular deviations from the null hypothesis. The asymptotic distribution of X^2 may be considered (see Cochran, (1952)) for general alternatives whose class probabilities differ

from the null probabilities by terms of order $N^{-1/2}$. This result could be sharpened to take account of more specific alternatives and of estimation difficulties, but it would then resemble very closely the discussion of Neyman's smooth goodness-of-fit test given, in a fine series of papers, by Barton (1953a, 1953b, 1955, 1956). In view of the intimate relationship of Barton's results with those above, his work will be described in some detail. It will be possible, using the final result of section 4, to complete Barton's Theorem III.

The family of distributions envisaged by Neyman and Barton is defined by

$$\left\{ 1 + \sum_{r=1}^K \frac{\lambda_r}{\sqrt{N}} \pi_r(z) \right\} f(x, \theta_1, \dots, \theta_s) \quad (5.1)$$

where the λ_r are constants, not depending on $\theta_1, \dots, \theta_s$, where the $\pi_r(z)$ are Legendre polynomials standardized on the interval $(-\frac{1}{2}, \frac{1}{2})$, and where

$$z = \int_{-\infty}^x f(t, \theta_1, \dots, \theta_s) dt - \frac{1}{2}. \quad (5.2)$$

The null hypothesis is defined by $\lambda_1 = \dots = \lambda_K = 0$; if it is true the probability integral transformation (5.2), when applied to a sample $x_1, \dots, x_N$ will lead to a set, $z_1, \dots, z_N$, uniformly distributed on $(-\frac{1}{2}, \frac{1}{2})$ - provided θ is known. Since the $\pi_r(z)$ are smooth functions of z (and so of x), a small value of k and small λ_r will mean that (5.1) differs smoothly from the null hypothesis. For θ known, it is easily seen that $u_r = N^{-1/2} \sum_{i=1}^N \pi_r(z_i)$ ($r=1, \dots, k$) are asymptotically uncorrelated normal variates with unit variances and means λ_r . Hence $\psi_k^2 = \sum_{r=1}^K u_r^2$ is a good statistic for testing $\lambda_1 = \dots = \lambda_K = 0$, since it has a non-central χ^2 distribution with parameter $\sum_{r=1}^K \lambda_r^2$.

When θ is unknown, and estimators $\theta_1^*, \dots, \theta_s^*$ are used in (5.2) to find $z_1^*, \dots, z_N^*$ and hence ψ_K^{*2} , Barton shows that, even on the null hypothesis, the distribution of ψ_K^{*2} is not χ_{k-s-1}^2 , as might be supposed, at least for maximum

likelihood estimators, but of a form close to that of Section 2. The latent roots involved will be functions of the unknown parameters except when they are location and scale parameters. Because the distribution of ψ_K^{*2} is never simple, it is not likely to enjoy much practical application.

Although one of Neyman's aims in suggesting this approach was to avoid the arbitrariness of class intervals, they are undoubtedly convenient with the large samples that are likely to be used in these tests. Moreover, it will appear from our extension of Barton's Theorem III, that they enable us to get a test statistic with a workable distribution. As with the discussion in Section 4, two types of grouping may be used. Barton uses the terms pre- and post-grouping.

In pregrouping, fixed class boundaries are used -- say, $-\infty < g_1 < \dots < g_{k-1} < \infty$ -- to divide the range of x into k intervals. If ξ_ℓ is the median of the ℓ th class, all the observations in that class will be considered as having this common value ξ_ℓ . Writing p_ℓ for the probability constant of the ℓ th class, ξ_ℓ , which replaces z in the ungrouped case, is defined by

$$\left. \begin{aligned} \sum_{\ell=1}^{l-1} p_\ell + \frac{p_l}{2} &= \int_{-\infty}^{\xi_r} f(x, \theta) dx, \\ &= \xi_r + \frac{1}{2} \end{aligned} \right\} \quad (5.3)$$

The generalization of u_r is

$$u_r = \frac{1}{\sqrt{N}} \sum_{\ell=1}^k n_\ell P_r(\xi_\ell) \quad (5.4)$$

where n_ℓ is the frequency in the ℓ th class and $P_r(\xi_\ell)$ are orthogonal polynomials replacing the $\sqrt{N_r}(z)$. If the $P_r(\xi_\ell)$ are defined to satisfy

$$\left. \begin{aligned} P_0(\xi_\ell) &= 1, \quad (\ell=1, \dots, k), \\ \sum_{\ell=1}^k p_\ell P_r(\xi_\ell) P_s(\xi_\ell) &= \delta_{rs}, \quad (r, s=0, \dots, k) \end{aligned} \right\} \quad (5.4)$$

then it is easily seen from Section 2 that, when θ is known, and the null hypothesis is true,

$$\psi_k^2 = \sum_{r=1}^K u_r^2, \quad \sim \chi_K^2 \quad (5.5)$$

as $N \rightarrow \infty$ and where $K < k-1$. Barton however finds the non-null distribution of ψ_K^2 and discusses the effect of estimation on this procedure. His main result, for our purposes, is that if θ is replaced by its maximum likelihood estimator θ_f^* based on the class frequencies, the null distribution of ψ_K^{*2} is χ_{K-s}^2 . This corresponds to the wellknown result for X^2 .

By post-grouping, Barton means that the group boundaries are adjusted to give constant estimated class probabilities. His theorem for this case may be extended and completed by assuming, as in Section 4, that the new class intervals are defined by functions $g_\ell(\theta^*)$ where θ^* is some estimator of θ . Suppose now that we proceed as in the pregrouped case. The class probabilities will be taken as (4.14) and the estimator θ_f^* found from (4.15). Instead of (5.3), we will write

$$\left. \begin{aligned} \sum_{\ell=1}^{l-1} p_\ell^* + \frac{p_\ell^*}{2} &= \int_{-\infty}^{S_\ell^*} f(x, \theta_f^*) dx \\ &= S_r^* + \frac{1}{2} \end{aligned} \right\} \quad (5.6)$$

where

$$p_\ell^* = p_\ell(\theta_f^*, \theta^*).$$

The orthogonal system $P_r(S_\ell^*)$ will be found as before so that (5.4) is satisfied but, in doing so, θ^* must be taken as an adjustable parameter so that (5.4) is true for all θ^* . Finally we may define

$$v_r^* = \frac{1}{\sqrt{N}} \sum_{\ell=1}^k n_{\ell} P_r(S_{\ell}^*) \quad (5.7)$$

and

$$\psi_K^{*2} = \sum_{r=1}^K v_r^{*2} \quad (5.8)$$

To find the asymptotic distribution of ψ_K^{*2} , it is merely necessary to examine (5.7). Only a sketch will be given since the methods required are now familiar. The n_{ℓ} will again have the representation (4.13) while, to the order required,

$$P_r(S_{\ell}^*) = P_r(S_{\ell}) + \frac{\partial P_r(S_{\ell})}{\partial S_{\ell}} \left\{ (\theta^* - \theta)' \frac{\partial S_{\ell}^*}{\partial \theta^*} \Big|_{\theta^* = \theta, \theta_f^* = \theta} + (\theta_f^* - \theta)' \frac{\partial S_{\ell}^*}{\partial \theta_f^*} \Big|_{\theta^* = \theta, \theta_f^* = \theta} \right\} \quad (5.9)$$

where unstarred symbols are those which result from putting $\theta^* = \theta$, $\theta_f^* = \theta$. In the notation of (4.12),

$$\left. \begin{aligned} \left(\frac{\partial S_{\ell}^*}{\partial \theta^*} \right)_{\theta^* = \theta, \theta_f^* = \theta} &= \sum_{\ell=1}^{\ell-1} v_{\ell}' + \frac{v_{\ell}}{2} \\ \left(\frac{\partial S_{\ell}^*}{\partial \theta_f^*} \right)_{\theta^* = \theta, \theta_f^* = \theta} &= \sum_{\ell=1}^{\ell-1} u_{\ell}' + \frac{u_{\ell}}{2} \end{aligned} \right\} \quad (5.10)$$

Using (4.13), (5.9), (5.10) in (5.7) and retaining only terms $O_p(1)$, we find that

$$\begin{aligned} v_r^* &\sim \frac{1}{\sqrt{N}} \sum_{\ell=1}^k m_{\ell} \left[P_r(S_{\ell}) + \frac{\partial P_r}{\partial S_{\ell}} (\theta_f^* - \theta)' \left(\sum_{\ell=1}^{\ell-1} u_{\ell}' + \frac{u_{\ell}}{2} \right) \right] \\ &= \sum_{\ell=1}^k \sqrt{N} (\theta^* - \theta)' \left[v_{\ell} P_r(S_{\ell}) + \frac{m_{\ell}}{N} \frac{\partial P_r}{\partial S_{\ell}} \left(\sum_{\ell=1}^{\ell-1} v_{\ell}' + \frac{v_{\ell}}{2} \right) \right] \end{aligned} \quad (5.11)$$

The first term in (5.11) is that which appears in the pregrouping case using intervals $g_{\ell}(\theta)$. The second term tends to zero in probability since $\sqrt{N} (\theta^* - \theta) = O_p(1)$

and the other factor has a variance which is $O(N^{-1})$ and a zero mean. To see the latter point, the mean in question is

$$\sum_{\ell=1}^k v_{\ell} P_r(S_{\ell}) + p_{\ell} \frac{\partial P_r}{\partial S_{\ell}} \left(\sum_{\ell=1}^{\ell-1} v_{\ell} + \frac{v_{\ell}}{2} \right). \quad (5.12)$$

But the relations (5.4) now hold for all θ^* and when the identity $\sum_{\ell=1}^k p_{\ell} P_r(S_{\ell}) = 0$ is differentiated with respect to θ^* , the left-hand side, with $\theta^* = \theta$ is simply (5.12). Thus only the first term of (5.11) contributes to V_r^* which then has the same distribution as in the pregrouped case. It follows then that ψ_K^{*2} is again χ_{K-S}^2 on the null hypothesis. This result corresponds to the last result of Section 4, and provides the solution to a slightly more general problem than that posed by Barton since the estimated class probabilities are here no longer required to be constant. Again we interpret this result as meaning that grouping the sample according to rules based on the sample estimators of the parameters has no influence on the asymptotic distribution of Barton's grouped form of Neyman's statistic. Finally we remark that the U_r^{*2} are partitions of the X^2 of Section 4 (in particular $\psi_{k-1}^{*2} = X^2$) and that this result could have been obtained from the theory given there. But to check that this was so would have taken as much space as, and been less informative than, the direct argument given above.

The Barton-Neyman test has much to recommend it -- greater power than X^2 if K is small -- with almost equal ease of application. In the hope of encouraging its further investigation for practical use, we quote Barton's formulae for $P_1(S)$, $P_2(S)$.

$$\left. \begin{aligned} P_1(S) &= c_1 \sqrt{3} \{2S\}, \\ P_2(S) &= c_2 \sqrt{5} \left\{ 6S^2 - \frac{1}{2}(1-S_{03}) + 18c_{113}^2 S \right\}, \end{aligned} \right\} \quad (5.13)$$

where

$$\left. \begin{aligned} s_{rs} &= \sum_{\ell=1}^k s_{\ell}^r p_{\ell}^s, \\ c_1 &= (1-s_{03})^{-1/2} \\ c_2 &= 2 \left\{ 4 - 36s_{23} - 9s_{05} + 10s_{03} - 5s_{03}^2 - 54s_{13}^2 c_1^2 \right\}^{-1/2} \end{aligned} \right\} (5.14)$$

As Barton has given no worked examples, his test will be applied to the data in Cramér (1954), Table 30.4.3, on the breadths of 12000 beans (Phaseolus vulgaris).

It may be shown here that, with a null hypothesis of normality,

$$P_1(\xi_{\ell}) = 3.49032 \xi_{\ell},$$

$$P_2(\xi_{\ell}) = 17.59329 (6 \xi_{\ell}^2 - 0.492543 + .000006 \xi_{\ell}),$$

$$U_1 = -2.5503, \quad U_2 = -10.7979,$$

so that

$$U_1^2 = 6.5, \quad U_2^2 = 116.6$$

In this example there are 16 classes so that the effects of differing estimators and the method of determining the class boundaries may be neglected (Watson, (1957)). Further, since N is 12000, we may assume that Cramér's $X^2 = 196.5$ is accurately a Chi-square with 13 degrees of freedom. U_1^2 and U_2^2 are components of this Chi-square, each with one degree of freedom, which are both significant, U_2^2 very much so, but the remainder, with eleven degrees of freedom,

$$X^2 - U_1^2 - U_2^2 = 73.4$$

is still significant. Cramer has shown by fitting an Edgeworth series to this observed distribution that the introduction of the term in g_1 alone reduces X^2 to 34.3 while the inclusion of the further term in g_2 (and g_1) makes X^2 only 14.9, a non-significant value. There are several lessons to be learnt from this example.

Cramer's analysis shows that the deviation from normality is "smooth"; in fact, that all but 34.3 of X^2 is due to skewness. Barton's components $U_1^2, U_2^2, \dots$ have no such immediately helpful interpretations and they may not always be in decreasing order of importance. Thus we have the situation, common in statistics, of having a test which would be ideal if we knew how many components should be used (i.e. the value of K) falling down in practice where we do not know this. Nor is the total reduction 123.1 as large as that obtained by Cramer, 181.6, for two degrees of freedom. However it is unfair to expect that a partitioning of X^2 of general utility should be as efficient - in the sense of reducing of X^2 - as a partitioning specifically designed for a given null hypothesis distribution. For the Edgeworth series analysis by Cramer arises naturally out of the orthogonal polynomials for the normal density function.

This leads us to the suggestion made by Lancaster (1953, 1957, 1958) that partitioning of X^2 should often be based on the series of orthogonal polynomials whose weight function is the null hypothesis density. Lancaster has not shown how to deal with the effects of grouping and estimation. Any theory taking these into account, as Barton has done for his partitioning, runs into difficulties unless the unknown parameters are those of scale and location. For suppose that the density on the alternative hypothesis may be written as

$$h(x) = f(x) \left\{ \sum_{r=0}^{\infty} \lambda_r f_r(x) \right\} \quad (5.15)$$

where $\lambda_0 = 1$ and the $f_r(x)$ are polynomials such that

$$\left. \begin{array}{l} f_r(x) \text{ has degree } r, \\ \int_{-\infty}^{\infty} f_r(x) f_s(x) f(x) dx = \begin{cases} 0, & r \neq s \\ 1, & r = s \end{cases} \end{array} \right\} \quad (5.16)$$

Then if $f(x)$ contains parameters, the restrictions (5.16) will mean that the $f_r(x)$ must also contain them. If the parameters are only introduced into (5.15) by replacing x by $(x-\theta_1)\theta_2^{-1}$, this is not so and standard families of polynomials may often be used. However the theory given in Section 2 to which, we have shown, all cases may be reduced - may be used to construct suitable linear functions for the partitioning of X^2 . With class boundaries and probabilities g_ℓ and p_ℓ as before, quantities

$$F_{r\ell} = \frac{1}{p_\ell} \int_{g_{\ell-1}}^{g_\ell} f_r(x) f(x) dx$$

may be defined, leading to linear forms $N^{-1/2} \sum_{\ell=1}^k n_\ell F_{r\ell}$ which are the natural analogues of the linear forms which are appropriate when the sample is ungrouped.

They are not satisfactory here because they are not uncorrelated with unit variance.

By using the earlier theory this could be overcome, along with the troubles due to the fact that the $F_{r\ell}$ will contain estimated parameters, but no further discussion can be given here.

References

- Barton, D. E., (1953a) Skand. Akt. Tid., 36, 24-63.
- Barton, D. E., (1953b) Trabajos d'Estadistica, 4, 199-207.
- Barton, D. E., (1955) Skand. Akt. Tid., 39, 1-17.
- Barton, D. E., (1956) Skand. Akt. Tid., 40, 216-245.
- Chernoff, H. and Lehmann, E. L., (1954) Ann. Math. Stat., 25, 579-586.
- Cochran, W. G., (1952) Ann. Math. Stat., 23, 315-345.
- Cochran, W. G., (1954) Biometrics, 10, 417-451.
- Cochran, W. G., (1955) J. Amer. Stat. Soc., 50, 377-397.
- Cramer, H. (1954) Mathematical Methods of Statistics. Princeton, Princeton University Press.
- Darwin, J., (1958) J. R. Stat. Soc. (to be published).
- Fisher, R. A. (1928) Atti de Congress Internazionale dei Mathematici, Bologna, 6, 94-100.
- Gani, J. (1955) Biometrika, 42, 342-359.
- Good, I. J., (1957) Ann. Math. Stat., 28, 861-881.
- Halmos, P. and Savage, L. T. (1949) Ann. Math. Stat., 20, 225-241.
- Hoel, P. G., (1938) Ann. Math. Stat., 2, 158-165.
- Lancaster, H. O., (1953) Sankhya, 13, 1-10.
- Lancaster, H. O., (1957) Biometrika, 44, 289-292.
- Lancaster, H. O., (1958) Ann. Math. Stat. (to be published).
- Neyman, J., (1937) Skand. Akt. Tid., 20, 150-199.
- Rao, C. R., and Chakravathi, I. M., (1956) Biometrics, 12, 264-282.
- Roy, A. R., (1956) Department of Statistics, Stanford University, Technical Report.
- Roy, S. N. and Mitra, (1955) North Carolina Institute of Statistics Mimeograph Series 139.
- Watson, G. S., (1957a) J. R. Stat. Soc., Series B, 19, 262-267.
- Watson, G. S., (1957b) Biometrika, 44, 336-348.
- Watson, G. S., (1958) J. R. Stat. Soc., Series B, 20, (to be published).