THE USE OF MAXIMUM LIKELIHOOD ESTIMATES IN χ^2 TESTS FOR GOODNESS OF FIT[1]

By Herman Chernoff and E. L. Lehmann

Stanford University and University of California

Summary. The usual test that a sample comes from a distribution of given form is performed by counting the number of observations falling into specified cells and applying the χ^2 test to these frequencies. In estimating the parameters for this test, one may use the maximum likelihood (or equivalent) estimate based (1) on the cell frequencies, or (2) on the original observations. This paper shows that in (2), unlike the well known result for (1), the test statistic does not have a limiting χ^2-distribution, but that it is stochastically larger than would be expected under the χ^2 theory. The limiting distribution is obtained and some examples are computed. These indicate that the error is not serious in the case of fitting a Poisson distribution, but may be so for the fitting of a normal.

1. Introduction. When using χ^2 for testing that a sample comes from a distribution of specified functional form such as a Poisson or normal distribution, the problem arises as to what estimates of the population parameters to use. If only the numbers m_i of observations falling into the ith of the k cells are available, there is no difficulty. Let p_i $(i = 1, \cdots, k)$ denote the probability of an observation falling into the ith cell, and let $\tilde{p}_i$ be any best asymptotically normal (b.a.n.) estimate of p_i such as the minimum χ^2 or maximum likelihood estimate. Then it is known [1], [2] that under suitable regularity conditions the asymptotic distribution of

$$\tilde{R} = \sum (m_i - m\tilde{p}_i)^2/n\tilde{p}_i \tag{1}$$

is that of χ^2 with $k - s - 1$ degrees of freedom, where s is the number of (independent) population parameters being estimated.

If, however, the original observations $x_1, \cdots, x_n$ are available, one is tempted to use more efficient estimates, such as the maximum likelihood estimates $\hat{p}_i$ based on all the data. One may reasonably expect this procedure to provide more powerful tests than those based only on the m_i; at the same time the estimates usually are simpler and easier to obtain. This is in fact the procedure recommended in many textbooks, particularly for the fitting of Poisson distributions, either as an approximation to the one with known theory described above, or more often without comment.

It is the purpose of the present paper to obtain the distribution of

$$\hat{R} = \sum (m_i - n\hat{p}_i)^2/n\hat{p}_i\,, \tag{2}$$

Received 7/24/53.

[1] This work was prepared with the partial support of the Office of Naval Research.

which differs from that of $\tilde{R}$. If we let

$$R = \sum (m_i - np_i)^2/np_i, \tag{3}$$

which has a limiting χ^2-distribution with $k - 1$ degrees of freedom, we shall show that the limiting distribution of $\hat{R}$ lies between those of $\tilde{R}$ and of R. More specifically, we shall show in Section 3 that under suitable regularity conditions we have

THEOREM 1. *The asymptotic distribution of $\hat{R}$ is that of*

$$\sum_{i=1}^{k-s-1} y_i^2 + \sum_{i=k-s}^{k-1} \lambda_i y_i^2 \tag{4}$$

where the y_i are independently normally distributed with mean zero and unit variance, and the λ_i are between 0 and 1 and may depend on the s parameters $\theta_1, \cdots, \theta_s$.

This result indicates that the recommended procedure of rejecting the hypothesis of goodness of fit when $\hat{R} > C$, where C is obtained from the χ^2-distribution with $k - s - 1$ degrees of freedom, will lead to a probability of rejection which, when the hypotheses is true, is greater than the desired level of significance. However, a numerical investigation of a few special cases indicates that, at least in the Poisson problem, this excess of probability of type I error will be so small as not to be serious. The situation appears to be not quite so favorable in the normal case.

Throughout this paper, the notation and background material given in Section 2 of the preceding paper [3] will be used.

2. Example. Before proceeding to the main result, let us treat the special example where the observations are independently and normally distributed with unknown mean and variance 1, and where the cells are $(-\infty, 0)$ and $(0, \infty)$. In this case it is obvious that $\tilde{R} = 0$. However,

$$\hat{R} = \sum_{i=1}^{2} \frac{(m_i - n\hat{p}_i)^2}{n\hat{p}_i} = \frac{(m_1 - n\hat{p}_1)^2}{n\hat{p}_1\hat{p}_2}$$

where

$$p_1(\mu) = \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} e^{-(y-\mu)^2/2}\, dy = \int_{-\infty}^{-\mu} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy$$

$$p_2(\mu) = 1 - p_1(\mu), \qquad \hat{p}_i = p_i(\bar{x}).$$

We have

$$\hat{R} = \frac{1}{p_1p_2 + o_p(1)}\left(\frac{m_1 - np_1}{\sqrt{n}} - \frac{n(\hat{p}_1 - p_1)}{\sqrt{n}}\right)^2 = \frac{(\epsilon - \nu)^2}{p_1p_2}[1 + o_p(1)]$$

where $\epsilon = (m_1 - np_1)/\sqrt{n}$ and $\nu = \sqrt{n}\,(\hat{p}_1 - p_1)$. Using the first order Taylor expansion of $p_1(\bar{x})$ about $p_1(\mu)$, we have

$$\nu = -\sqrt{n}\,(\bar{x} - \mu)\, e^{-\mu^2/2}/\sqrt{2\pi} + o_p(1) = \nu' + o_p(1)$$

where $\nu' = -\sqrt{n}(\bar{x} - \mu)\, e^{-\mu^2/2}/\sqrt{2\pi}$. Let $g(x) = 1$ for $x < 0$, and $= 0$ otherwise. The central limit theorem tells us that, since

$$(m_1, n\bar{x}) = \sum_{\alpha=1}^{n} [g(x_\alpha), x_\alpha], \qquad d\infty(\epsilon, \nu') = N(0, \Sigma)$$

where $N(0, \Sigma)$ denotes the normal distribution with mean 0 and covariance matrix Σ given by

$$\Sigma = \begin{pmatrix} p_1 p_2 & e^{-\mu^2}/2\pi \\ e^{-\mu^2}/2\pi & e^{-\mu^2}/2\pi \end{pmatrix}.$$

Hence

$$d\infty(\epsilon, \nu) = N(0, \Sigma), \qquad d\infty(\epsilon - \nu) = N(0, p_1 p_2 - e^{-\mu^2}/2\pi)$$

and in particular $\epsilon - \nu = O_p(1)$. It follows that $d\infty(\hat{R}) = d(\lambda y^2)$, where $d(y) = N(0, 1)$ and $\lambda = 1 - e^{-\mu^2}/2\pi p_1 p_2 < 1$. The fact that $\lambda \geqq 0$ follows from the fact that Σ is nonnegative definite.

NOTE. A general proof of Theorem 1 cannot be based solely on the fact that $\hat{p}_1$ is a better estimate of p_1 than $\tilde{p}_1$ is. Suppose, in fact, that we use $p_1^* = p_1(2 - \bar{x})$ as our estimate of p_1. In the event that $\mu = 1$, p_1^* has the same distribution as $\hat{p}_1$. The above argument repeated for R^* shows that ν would be replaced by $-\nu$ and λ by $\lambda^* = 1 + 3e^{-\mu^2}/2\pi p_1 p_2$.

3. The general case. We shall now prove Theorem 1 under the following regularity conditions:

(i) The $p_i(\theta)$ satisfy the condition on pages 426–427 of Cramér's *Mathematical Methods of Statistics.*

(ii) Let $z = (z_1, \cdots, z_k)$ where $z_i = 1$ if the observation falls in the ith cell and 0 otherwise. Let $f(z, \theta) = \prod p_i^{z_i}$, and let us assume that the value w of our chance variable x determines z, and that the density of x is given by

$$f^*(w, \theta) = \prod p_i^{z_i}\, g(w|z, \theta)$$

where g is the conditional density of x given z. Then we assume that f^* satisfies the condition $\mathcal{R}$ of the preceding paper [3].

Let

$$m_i - np_i = \sqrt{np_i}\,\epsilon_i, \tag{5}$$

$$n(\tilde{p}_i - p_i) = \sqrt{np_i}\,\tilde{\nu}_i, \tag{6}$$

$$n(\hat{p}_i - p_i) = \sqrt{np_i}\,\hat{\nu}_i. \tag{7}$$

Then

$$R = \sum \epsilon_i^2 = \epsilon'\epsilon, \tag{8}$$

$$\tilde{R} = \sum (\epsilon_i - \tilde{\nu}_i)^2\,[1 + o_p(1)], \tag{9}$$

$$\hat{R} = \sum (\epsilon_i - \hat{\nu}_i)^2\,[1 + o_p(1)]. \tag{10}$$

We shall first compute $\tilde{\nu}_i$ to show that $\tilde{R}$ is asymptotically a sum of squares of the components of a normally distributed chance variable, and then do the same for $\hat{R}$. We have

$$\frac{\partial \log f(z, \theta)}{\partial \theta_j} = \sum_{i=1}^{k} \frac{z_i}{p_i} \frac{\partial p_i}{\partial \theta_j}.$$

The information matrix referred to in Section 2 of [3] is given by

$$\tilde{J} = \left\| \sum_{r=1}^{k} \frac{1}{p_r} \frac{\partial p_r}{\partial \theta_i} \frac{\partial p_r}{\partial \theta_j} \right\| = D'D \tag{11}$$

where

$$D = \left\| \frac{1}{\sqrt{p_i}} \frac{\partial p_i}{\partial \theta_j} \right\|. \tag{12}$$

The corresponding A vector $\tilde{A}$ has elements

$$\tilde{A}_i = \frac{1}{n} \sum_{r=1}^{k} \frac{m_r}{p_r} \frac{\partial p_r}{\partial \theta_i} = \sum_{r=1}^{k} \frac{m_r - np_r}{np_r} \frac{\partial p_r}{\partial \theta_i}.$$

Therefore

$$\tilde{A} = (1/\sqrt{n})D'\epsilon \tag{13}$$

and

$$\tilde{\nu}_i = \frac{\sqrt{n}(\tilde{p}_i - p_i)}{\sqrt{p_i}} = \sum_{j=1}^{s} \sqrt{n}(\tilde{\theta}_j - \theta_j) \frac{1}{\sqrt{p_i}} \frac{\partial p_i}{\partial \theta_j} + o_p(1),$$

$$\tilde{\nu} = D\sqrt{n}(\tilde{\theta} - \theta) + o_p(1) = D\tilde{J}^{-1}D'\epsilon - o_p(1).$$

Finally

$$\tilde{R} = (\tilde{F}\epsilon)'(\tilde{F}\epsilon) + o_p(1) \tag{14}$$

where

$$\tilde{F} = I - D\tilde{J}^{-1}D'. \tag{15}$$

Now

$$\frac{\partial \log f^*(w, \theta)}{\partial \theta_j} = \sum_{i=1}^{k} \frac{z_i}{p_i} \frac{\partial p_i}{\partial \theta_j} + \frac{\partial \log g(w \mid z, \theta)}{\partial \theta_j}. \tag{16}$$

Since the conditional expectation, given z, of

$$\left[\sum_{i=1}^{k} \frac{z_i}{p_i} \frac{\partial p_i}{\partial \theta_j} \right] \cdot \frac{\partial \log g(w \mid z, \theta)}{\partial \theta_l}$$

is zero, we have

$$\hat{J} = \tilde{J} + J^* \tag{17}$$

$$\hat{A} = \tilde{A} + A^* \tag{18}$$

where

$$J^* = \left\| E\left[\frac{\partial \log g(x \mid z, \theta)}{\partial \theta_i} \cdot \frac{\partial \log g(x \mid z, \theta)}{\partial \theta_j}\right] \right\|, \tag{19}$$

$$A_i^* = \frac{1}{n} \sum_{\alpha=1}^{n} \frac{\partial \log g(x \mid z^{(\alpha)}, \theta)}{\partial \theta_i} \tag{20}$$

and $z^{(\alpha)}$ is the αth observation on z. Now

$$\begin{aligned} \hat{\nu}_i &= \sqrt{n}\, \frac{\hat{p}_i - p_i}{\sqrt{p_i}} = \sum \sqrt{n}(\hat{\theta}_j - \theta_j)\, \frac{1}{\sqrt{p_i}} \frac{\partial p_i}{\partial \theta_j} + o_p(1), \\ \hat{\nu} &= D\sqrt{n}(\hat{\theta} - \theta) + o_p(1). \\ \hat{\nu} &= D(\tilde{J} + J^*)^{-1}(D'\epsilon + \sqrt{n}A^*) + o_p(1). \end{aligned} \tag{21}$$

Hence

$$\hat{R} = (\hat{F}\epsilon + \hat{G}\eta)'(\hat{F}\epsilon + \hat{G}\eta) + o_p(1) \tag{22}$$

where $\eta = \sqrt{n}A^*$, while $\hat{F} = I - D(\tilde{J} + J^*)^{-1} D'$ and $\hat{G} = D(\tilde{J} + J^*)^{-1}$. The asymptotic distributions of R, $\tilde{R}$, and $\hat{R}$ are those of $\epsilon'\epsilon$, $(\tilde{F}\epsilon)'(\tilde{F}\epsilon)$, and $(\hat{F}\epsilon + \hat{G}\eta)'(\hat{F}\epsilon + \hat{G}\eta)$, respectively. To find these distributions we must know the asymptotic distribution of (ϵ, η). Applying the central limit theorem to

$$\left[z_1, z_2, \cdots, z_k, \frac{\partial \log g(w \mid z, \theta)}{\partial \theta_1}, \cdots, \frac{\partial \log g(w \mid z, \theta)}{\partial \theta_s}\right]$$

we see that

$$d\infty(\epsilon, \eta) = N\left[0, \begin{pmatrix} I - qq' & 0 \\ 0 & J^* \end{pmatrix}\right] \tag{23}$$

where q is the vector whose ith component is $\sqrt{p_i}$. (Note that $D'q = 0$.)

From one of the Mann-Wald results it follows that the asymptotic distributions we desire are those obtained by assuming that (ϵ, η) actually have the above joint normal distribution. That is we assume that

$$d(\epsilon) = N(0, \Sigma), \qquad d(\tilde{F}\epsilon) = N(0, \tilde{\Sigma}), \qquad d(\hat{F}\epsilon + \hat{G}\eta) = N(0, \hat{\Sigma})$$

where

$$\Sigma = I - qq' \tag{24}$$

$$\tilde{\Sigma} = I - qq' - D\tilde{J}^{-1} D' \tag{25}$$

$$\hat{\Sigma} = I - qq' - D(\tilde{J} + J^*)^{-1} D'. \tag{26}$$

If for symmetric matrices we write $K \geqq L$ whenever $K - L$ is nonnegative definite, then

$$\Sigma \geqq \hat{\Sigma} \geqq \tilde{\Sigma}. \tag{27}$$

We digress to present

LEMMA 1. *If* $d(y) = N(0, U)$ *where the characteristic roots of* U *are* $\lambda_1, \lambda_2, \cdots, \lambda_k$, *then*

$$d(y'y) = d(\Sigma \lambda_i z_i^2)$$

where $d(z) = N(0, I)$.

PROOF. Expressing U in canonical form, we have $U = P\Lambda P'$, where P is orthogonal and Λ is the diagonal matrix whose diagonal elements are the λ_i. Since U is nonnegative definite, the λ_i are nonnegative and we may define $\Lambda^{\frac{1}{2}}$ in the obvious way. Let $d(z) = N(0, I)$ and $y^* = P\Lambda^{\frac{1}{2}}z$. Then $d(y^*) = N(0, U)$ and $d(y^{*\prime}y^*) = d(y'y)$. But

$$y^{*\prime}y^* = z'\Lambda^{\frac{1}{2}}P'P\Lambda^{\frac{1}{2}}z = z'\Lambda z$$

and the lemma follows.

As a consequence of this lemma, it follows that the distributions of R $\tilde{R}$, and $\hat{R}$ are those of $z'\Lambda z$, $z'\tilde{\Lambda}z$, and $z'\hat{\Lambda}z$ where Λ, $\tilde{\Lambda}$, $\hat{\Lambda}$ are the diagonal matrices of characteristic values corresponding to R, $\tilde{R}$, and $\hat{R}$, respectively. From the known results on R and $\tilde{R}$ it follows that Σ has for characteristic roots $k - 1$ ones and 1 zero, while $\tilde{\Sigma}$ has for characteristic roots $k - s - 1$ ones and $s + 1$ zeros. Since $\Sigma \geqq \hat{\Sigma} \geqq \tilde{\Sigma}$, it follows that $\hat{\Sigma}$ has for characteristic roots: $k - s - 1$ ones, 1 zero, and s roots $\lambda_1, \lambda_2, \cdots, \lambda_s$ between zero and one. Our Theorem 1 follows.

REMARK. A direct proof of the above-mentioned properties of the characteristic roots of Σ and $\tilde{\Sigma}$ may be given by showing that qq' and $D\tilde{J}^{-1}D'$ are projection operators on orthogonal manifolds of dimensions 1 and s respectively, that is,

$$(qq')(qq') = q(\textstyle\sum p_i)q' = qq'$$

$$(D\tilde{J}^{-1}D')(D\tilde{J}^{-1}D') = D\tilde{J}^{-1}\tilde{J}\tilde{J}^{-1}D' = D\tilde{J}^{-1}D'$$

$$(D\tilde{J}^{-1}D')(qq') = 0.$$

The roots $\lambda_1, \cdots, \lambda_s$ which determine the distribution of the test criterion $\hat{R}$ can be obtained from

THEOREM 2. *If* $\mu_i = 1 - \lambda_i$, *then the* μ_i *are the characteristic roots of the determinantal equation* $|\tilde{J} - \mu\hat{J}| = 0$.

PROOF. We shall use the fact that if the vectors $t_1, \cdots, t_k$ form an orthonormal basis of k-dimensional space, then the matrix $\sum \tau_i t_i t_i'$ has the characteristic roots $\tau_1, \cdots, \tau_k$. This implies, in particular, that $\sum t_i t_i'$ is the identity matrix.

Given $\hat{J}$ and $\tilde{J}$, there exists a nonsingular $(s \times s)$ matrix S and a diagonal matrix M such that

$$\tilde{J}^{-1} = SS', \qquad \hat{J}^{-1} = SMS'$$

where the diagonal elements of M are the roots of $| \hat{J}^{-1} - \mu \tilde{J}^{-1} | = 0$ and hence of $| \tilde{J} - \mu \hat{J} | = 0$, and are all between 0 and 1, since $\hat{J} \leqq \tilde{J}$.

If $u_1, \cdots, u_s$ are the columns of DS,

$$D\tilde{J}^{-1}D' = (DS)(DS)' = \sum u_i u_i'.$$

Since $(DS)'(DS) = S'\tilde{J}S = I$ and $D'q = 0$, it follows that $q, u_1, \cdots, u_s$ are mutually orthogonal unit vectors. If we let $v_1, \cdots, v_{k-s-1}$ be a complementary set of orthogonal unit vectors we have

$$\begin{aligned}\hat{\Sigma} &= I - qq' - DSMS'D' = I - qq' - \sum_{i=1}^{s} \mu_i u_i u_i' \\ &= \sum_{j=1}^{k-s-1} v_j v_j' + \sum_{i=1}^{s} (1 - \mu_i) u_i u_i'.\end{aligned}$$

It follows that the characteristic roots of $\hat{\Sigma}$ consist of $k - s - 1$ ones, one zero, and $\lambda_i = 1 - \mu_i$ for $i = 1, \cdots, s$.

4. Some Numerical Examples. By using the maximum likelihood estimates based on the full sample, one is operating at a higher significance level than the one stated. One can, however, on the basis of the above results, make an adjustment which asymptotically provides the correct value.

Given θ, let $C(\theta)$ be such that

$$P\left\{\sum_{i=1}^{k-s-1} y_i^2 + \sum_{i=k-s}^{k-1} \lambda_i(\theta) y_i^2 \neq C(\theta)\right\} = \alpha.$$

Clearly $C(\theta)$ is a continuous function of θ and hence $C(\hat{\theta}) \to C(\theta)$ in probability as $n \to \infty$. It follows that the probability of

$$\sum_{i=1}^{k-s-1} y_i^2 + \sum_{i=k-s}^{k-1} \lambda_i(\theta) y_i^2 \geqq C(\hat{\theta})$$

tends to α as $n \to \infty$. Here $C(\hat{\theta})$ can be computed, at least in theory, to an arbitrary degree of accuracy using the results of Pitman and Robbins [4].

Theoretically, the error committed by using the maximum likelihood estimates based on the full sample without an adjustment can be quite serious in the case of a small number of cells. For example, if $s = 1$ and $\lambda(\theta)$ is close to 1, we have essentially one extra degree of freedom, and when the number k of cells is small so that $k - 2 = 1$, 2 or 3, the actual probability α^* of type I error would vary from 15 per cent to 10 per cent when the level of significance is supposed to be $\alpha = 5$ per cent.

In practice, however, at least for fitting a Poisson distribution, the error does not appear to be so serious. Some values of $\lambda(\theta)$ and the true probability of rejection $\alpha^*(\theta)$ in the Poisson case are given below for groupings $x = 0, 1, \geqq 2$ and $x = 0, 1, 2, \geqq 3$, and level of significance supposed to be $\alpha = .05$.

x	$0,1, \geqq 2$		$0,1,2, \geqq 3$	
θ	1	2	2	3
$\lambda(\theta)$ $\alpha^*(\theta)$	.12 .054	.35 .067	.14 .055	.32 .065

As a second example, consider the fitting of a normal distribution with mean ζ and variance σ^2, both unknown. For the case of the four cells $(-\infty, -1)$, $(-1, 0)$, $(0, 1)$, $(1, \infty)$ and two combinations of ζ and σ we obtain the following values for the two roots λ_1 and λ_2 :

$$\zeta = 0, \quad \sigma = 2.5; \quad \lambda_1 = .80, \quad \lambda_2 = .20$$

$$\zeta = .5, \quad \sigma = 2.0; \quad \lambda_1 = .74, \quad \lambda_2 = .15.$$

The probability α^* is then given by

$$\alpha^* = P\{U + \lambda_1 V + \lambda_2 W \geqq C_\alpha\}$$

where U, V, W are χ^2 variables with 1 degree of freedom and C_α is such that $P\{U \geqq C_\alpha\} = \alpha$. As a lower bound of α^* in the first case we have computed $P\{U + .8V \geqq C_{.05}\} = .12$. This indicates that in the normal case the use of maximum likelihood estimates in χ^2 may lead to a more serious underestimate of the probability of type I error.

REFERENCES

[1] 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.

[2] H. Cramér, *Mathematical Methods of Statistics*, Princeton University Press, 1946.

[3] H. Chernoff, "On the distribution of the likelihood ratio," *Ann. Math. Stat.*, Vol. 25 (1954), pp. 573-578.

[4] E. J. G. Pitman and H. Robbins, "Application of the method of mixtures to quadratic forms in normally correlated variables," *Ann. Math. Stat.*, Vol. 20 (1949), pp. 552-560.