Asymptotic Normality of χ^2 in $m \times n$ tables with n large
and small cell expectations*

by Jack Cuzick

Department of Mathematical Statistics

Columbia University, New York

NOTICE
This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Department of Energy, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

Summary: Asymptotic normality for χ^2 used as a test for homogeneity is established under nonstandard conditions. The case of an $m \times n$ table with m fixed and the total number of observations proportional to n is studied for n large. Results are obtained under very mild assumptions on the marginal totals.

Some key words: contingency tables; chi-squared statistic; asymptotic normality.

MASTER

There is no objection from the patent point of view to the publication or dissemination of the document(s) listed in this report.

EX-100 PATENT GROUP

11/12/78 By *exl*

*Support for this work was provided by SIMS (SIAM Institute for Mathematics and Society).

Prepared For
U.S. ENERGY RESEARCH AND DEVELOPMENT ADMINISTRATION
UNDER CONTRACT NO: EY-76-S-02-2874

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED *EB*

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

1. Introduction

The classical analysis of Pearson's χ^2 statistic provides its asymptotic distribution under the assumption of a fixed number of cells and a large number of observations in each cell. The object here is to obtain asymptotic results in cases where the number of cells and the number of observations both get large while cell expectations remain small. Under these circumstances the χ^2 statistic does not have the χ^2 distribution. Certain results in this direction have been obtained by Morris (1975) when χ^2 is used as a goodness-of-fit test. We shall study the case in which χ^2 is used to test homogeneity. More specifically this note limits discussion to an $m \times n$ table with m fixed ($m \geq 2$) and n tending to infinity at the same rate as the number of observations. Such a case occurs when repeated data classifications occur. An example which considers daily visits in different New York City hospitals for a particular disease will be presented elsewhere.

Let N_{ij} denote the entries in an $m \times n$ table with $m \geq 2$. Define $N_j = \sum_{i=1}^m N_{ij}$, $N_i = \sum_{j=1}^n N_{ij}$, $N = \sum_{i=1}^m \sum_{j=1}^n N_{ij}$.

and $p_j = N_j/N$, where the last three variables have an implicit dependence on n . Haldane (1939) has given the exact mean and variance for χ^2 when used as a test of homogeneity. For our asymptotic use they take the form

$$E(\chi^2) = (m-1)(n-1) \frac{N}{N-1}$$

$$\text{Var}(\chi^2) = 2(m-1)(n-1-N^*) + (p^*-m^2)(N^* - \frac{n^2}{N}) +$$

$$2(\frac{n}{N})(1-p^*) - (\frac{n}{N})^2 (2m^2 - 2m + 1 - p^*) + O(\frac{1}{n}), \quad (1)$$

$$\text{where } p^* = \sum_{i=1}^m p_i^{-1} \text{ and } N^* = \sum_{j=1}^n N_j^{-1}.$$

Note that the mean is not much changed from its usual value but the variance can be strikingly different. The following theorem asserts that under mild conditions on the marginal totals, χ^2 is asymptotically normal as $n \rightarrow \infty$.

Theorem Assume that as $n \rightarrow \infty$

- i) $p_i \rightarrow \bar{p}_i > 0$
- ii) $\limsup n^{-1} N^* < 1$
- iii) $\limsup n^{-1} N < \infty$.

Then, under the hypothesis of homogeneity, χ^2 is asymptotically standard normal after centering and rescaling.

Proof The proof uses methods of Steck (1957) and Morris (1975) to compare the given χ^2 to that which would be obtained if the p_i were known and not estimated from the marginal data. To do this, let $X_j = (X_{1j}, \dots, X_{mj})$, $j=1, \dots, n$ denote independent multinomial vectors with N_j trials and cell probabilities

p_i , $i=1, \dots, m$. Let

$$\chi_j^2(X_{ij}) = \sum_{i=1}^m \frac{(X_{ij} - p_i N_j)^2}{p_i N_j} \quad (2)$$

Then the $\chi_j^2(X_{ij})$ are all independent and from Haldane (1937) we have

$$\begin{aligned} E(\chi_j^2(X_{ij})) &= m-1 \\ \text{Var}(\chi_j^2(X_{ij})) &= 2(m-1)(1-N_j^{-1}) + (p^*-m^2)N_j^{-1}. \end{aligned} \quad (3)$$

For fixed p_i this variance is bounded, and since $p^* \geq m^2$, it is bounded away from zero if $N_j \geq 2$. Thus under assumptions i) and ii), the uniformly asymptotically negligible condition holds, so that after centering and rescaling,

$$\chi^2(X_{ij}) = \sum_{j=1}^n \chi_j^2(X_{ij}) \quad (4)$$

is asymptotically $\mathcal{N}(0,1)$.

Limiting results are more difficult when the X_{ij} are replaced by N_{ij} . The difficulties arise because the $\chi_j^2(N_{ij})$ are no longer independent. However the N_{ij} have the same joint distribution as the X_{ij} conditioned on the sums $X_{i.} = \sum_{j=1}^n X_{ij} = N_{i.}$, $i=1, \dots, m$. Following Morris (1975) define $V_n = (V_{1n}, \dots, V_{mn})$ by

$$V_{in} = N^{-1/2}(X_{i.} - p_i N) \quad i=1, \dots, m$$

and let

$$U_n^* = \chi^2(X_{ij}) - E_w(\chi^2(X_{ij}) | V_n)$$

where E_w denotes wide sense conditional expectation, i.e.

$$U_n^* = \chi^2(X_{ij}) - E(\chi^2(X_{ij})) - \sum_{i=1}^{m-1} \gamma_i N^{1/2} V_{in}$$

with γ_i chosen so that U_n^* is orthogonal to the vector V_n .

(The sum has upper limit $m-1$ since the V_{in} are constrained to the subspace $\sum_{i=1}^m V_{in} = 0$.) It is easy to verify that the γ_i remain bounded as $n \rightarrow \infty$. If we define

$$g_j(X_{ij}) = \chi_j^2(X_{ij}) - E(\chi_j^2(X_{ij})) - \sum_{i=1}^{m-1} \gamma_i (X_{ij} - p_i N_j) \quad (5)$$

then the $g_j(X_{ij})$ are independent and $U_n^* = \sum_{j=1}^n g_j(X_{ij})$. Clearly $\text{Var}(g_j(X_{ij}))$ is bounded so that $s_n^2 = \text{Var}(U_n^*)$ grows no faster than linearly. Thus we will have that

$$U_n = s_n^{-1} U_n^* \xrightarrow{n \rightarrow \infty} \mathcal{N}(0, 1)$$

if we can show that s_n^2 grows at least linearly. However s_n^2 is the wide sense conditional variance and is larger than the expected conditional variance given by

$$E \left\{ \text{var}(\chi^2(X_{ij}) | X_{i.}, i=1, \dots, m) \right\}.$$

Since $\sum_{i=1}^m X_{i.} = N$, from (1) we see that this is proportional to n

for n large. This follows because $p^* \geq m^2$ and $N^* \geq n^2 N^{-1}$ by

Jensen's inequality and the first term in (1) is of the order of n by

assumption ii). Since U_n is orthogonal to V_n , by a simple modif-

ication of lemma 2.1 of Morris (1975), we have that $(U_n, V_n) \rightarrow$

(U, V) which is $\mathcal{N}(0, \Sigma)$ with

$$\Sigma = \left(\begin{array}{c|c} 1 & 0 \dots 0 \\ \hline 0 & \\ \vdots & p_i \delta_{ik} - p_i p_k \\ 0 & \end{array} \right)$$

Thus from theorem 2.1 of Steck (1957), the normalized $\chi^2(N_{ij})$, which is equal

in law to U_n given $V_n = 0$, will converge to a standard normal,

if it can be shown that for each λ

$$f_n(v) = E \{ \exp (i \lambda Z_n(v)) \}$$

is an equicontinuous family for v in compact sets in the domain

of V_n . Here $Z_n(v)$ is U_n conditioned by $V_n = v$. Since V_n is lattice,

the f_n are trivially continuous, and we need only concern ourselves with large n in checking equicontinuity. This is established by constructing the random variables $Z_n(v)$ and $Z_n(v+h)$ on the same probability space as follows. First observe that

$$V_n = v \Leftrightarrow X_{i.} = p_i N + N^{1/2} v_i = \tilde{p}_i N$$

$$V_n = v + h \Leftrightarrow X_{i.} = p_i N + h_i N^{1/2}$$

and that $\sum_{i=1}^m h_i = 0$. Take n so large that

$$N^{1/2} \sum_{i=1}^m |h_i| \leq N \min_i p_i.$$

Now construct the variables N_{ij} under the homogeneity assumption with marginals $N_{.j}$ and $N_{i.} = p_i N$ from which we obtain $Z_n(v)$.

Then $Z_n(v+h)$ is obtained by shuffling some of the entries. If h_i is positive, randomly chose $h_i N^{1/2}$ elements $\{e_k\}_{k=1}^{h_i N^{1/2}}$ from the m^{th} row, and if e_k is in the j^{th} column place it in cell ij . If h_i is negative, take samples from the i^{th} row and place them in the m^{th} row analogously. Do this for $i=1, \dots, m-1$, all operations being independent. The resulting cell counts, denoted $N_{ij} + L_{ij}$, will be homogeneous with marginals $N_{.j}$ and $N_{i.} + h_i N^{1/2}$. Lastly $Z_n(v+h)$ is computed from this table. By means of this construction we see that

$$\begin{aligned} |f_n(v) - f_n(v+h)| &= |E\{\exp(i\lambda Z_n(v)) - \exp(i\lambda Z_n(v+h))\}| \\ &\leq \frac{|\lambda|}{s_n} E \left| \sum_{j=1}^n g_j(N_{ij}) - g_j(N_{ij} + L_{ij}) \right|. \end{aligned}$$

Expanding g_j by (4) and (5), this equals

$$\frac{|\lambda|}{s_n} E \left| \sum_{j=1}^n \sum_{i=1}^m \frac{2N_{ij}L_{ij} + L_{ij}^2 - E(2N_{ij}L_{ij} + L_{ij}^2)}{N_j p_i} - \sum_{i=1}^{m-1} \gamma_i h_i N_i \right| \quad (6)$$

Since s_n is proportional to $n^{\frac{1}{2}}$ and the γ_i are bounded we see that the last term in (6) is negligible and the remaining expression is bounded by a constant times

$$n^{-\frac{1}{2}} \sum_{j=1}^n \sum_{i=1}^m E \left| \frac{N_{ij}L_{ij}}{N_j p_i} + \frac{EL_{ij}^2}{N_j p_i} \right| \quad (7)$$

Since L_{ij} is hypergeometric

$$EL_{ij}^2 \leq E(L_{ij}) + E^2(L_{ij}) = N_j h_i N^{-\frac{1}{2}} + (N_j h_i N^{-\frac{1}{2}})^2$$

so that the second term in (7) goes to zero as $\max |h_i| \rightarrow 0$ uniformly in n . To estimate the first term in (7) first consider the case $i \leq m$ and $h_i > 0$. Then N_{ij} and L_{ij} can be considered as separate cell counts in an $(m+1) \times n$ table, where the extra row is obtained by segregating the elements of row m transferred into row i . Thus

$$\begin{aligned} E|N_{ij}L_{ij}| &= E|N_{ij}E(L_{ij}|N_{ij})| \\ &= E|N_{ij}(N_j - N_{ij})(h_i N^{-\frac{1}{2}})(1 - \tilde{p}_i)^{-1}| \\ &\leq h_i N^{-\frac{1}{2}} N_j (1 - \tilde{p}_i)^{-1} E|N_{ij}| = O(h_i N_j^2 N^{-\frac{1}{2}}). \end{aligned} \quad (8)$$

When $i < m$, $h_i < 0$, L_{ij} is negative, and $(N_{ij} - |L_{ij}|)$ and the

$|L_{ij}|$ can be considered as entries in an $(m+1) \times n$ table in the same manner. Writing

$N_{ij} = (N_{ij} + L_{ij}) + (-L_{ij})$ the same estimate (8) can be obtained. An analogous argument can be constructed when $i=m$, allowing one to conclude the first term in (7) is asymptotically small. This completes the proof.

3. Remarks

i). Assumption iii) is an artifact of the present method of proof and possibly can be completely done away with. However it is satisfied in most applications.

ii). Comparing (1) and (3) we see that

$n^{-1} \text{Var}(\chi^2)$ converges to different values depending on whether the p_j are known in advance or are estimated from the data. Thus the effects of not knowing the p_j are retained asymptotically even though we can estimate them arbitrarily well.

iii). Analogous results can be obtained for the likelihood ratio statistic but their usefulness appears to be hampered by the difficulties in computing centering and scaling constants.

iv) Restricting attention to the $2 \times n$ case, it is often of interest to estimate some difference in the p -vectors obtained by considering the two rows as separate multinomials. If a trend is suspected, the methods of Cox (1970 chap. 4) would be useful.

In the general case an estimate of the difference of the two p -vectors is some suitable metric is called for.

This work was supported by the SIAM Institute for Mathematics and Society.

References

Cox, D.R. (1970). The Analysis of Binary Data, London: Methuen.

Haldane, J.B.S. (1937). 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 29, 133-143.

Haldane, J.B.S. (1939). The mean and variance of χ^2 , when used as a test of homogeneity, when expectations are small. Biometrika 31, 346-355.

Morris, C. (1975). Central limit theorems for multinomial sums. Ann. Statistics 3, 165-188.

Steck, G. P. (1957). Limit theorems for conditional distributions. Univ. California Publ. Statist. 2, 237-284.