On Multivariate Chi-Square Statistics with Random Cell Boundaries

by

D.S. Moore

Department of Statistics
Division of Mathematical Sciences

Mimeo Series No. 236

August 1970

On Multivariate Chi-Square

Statistics with Random Cell Boundaries

by

D.S. Moore

Purdue University

SUMMARY

The limiting distribution of the chi-square goodness of fit statistic with parameters estimated is not effected by the use of cells which are functions of the observations, whether the parameters are estimated from grouped or ungrouped data. More precise statements of these results are found in Watson (1959). Unfortunately, rigorous proofs are tedious and have been given only for univariate observations. We show here that modern weak convergence methods greatly simplify the proofs and extend the results to rectangular cells in any number of dimensions.

1. INTRODUCTION

The theory of asymptotic distributions of chi-square goodness-of-fit statistics based on a fixed number of cells for univariate observations is quite well known. If one uses M cells and estimates m parameters by means of maximum likelihood estimators based on the grouped data, the asymptotic distribution under the null hypothesis is χ^2_{M-m-1} (Cramer (1946), section 30.3). If the parameters are estimated by maximum likelihood or other asymptotically efficient means from the ungrouped data, Chernoff and Lehmann (1954) showed that the asymptotic distribution is not chi-square. It is rather the law of a random variable of the form

(1) $$\sum_{i=1}^{M-m-1} Z_i^2 + \sum_{i=M-m}^{M-1} \lambda_i Z_i^2$$

where the Z_i are independent standard normal and the constants λ_i lie between 0 and 1.

These results are derived from the multinomial distribution and are therefore independent of the dimension of the observations. It is, however, often desirable not to use fixed cell boundaries, which are assumed by the results above. For example, in testing goodness of fit to the normal family $N(\mu,\sigma^2)$ with both parameters unknown a statistician may wish to use cell boundaries of the form $\overline{X} + a_i s$. Here $\overline{X}$ and s^2 are the sample mean and variance and the constants a_i would typically be chosen as the boundaries of equiprobable cells in the N(0,1) case. Work of A.R. Roy (1956) and G.S. Watson (1958), (1959) shows that in the case of univariate observations we can say generally that use of cell boundaries which are functions of asymptotically efficient estimators of unknown parameters does not effect the asymptotic distribution of the chi-square statistic under the null hypothesis.

To be more precise, suppose that $\hat{\theta}_n$ is an asymptotically efficient sequence of estimators for θ (both $\hat{\theta}_n$ and θ may be vectors) based on a random sample $X_1,\ldots, X_n$. We choose cell boundaries to be smooth functions of $\hat{\theta}_n$. Let N_i, $i=1,\ldots,M$ be the number of $X_1,\ldots,X_n$ falling in the ith cell. Since the cell boundaries are functions of $\hat{\theta}_n$, the cell probabilities under the null hypothesis are functions $p_i(\theta_o,\hat{\theta}_n)$, where θ_o is the true parameter value. Now suppose we ignore the random nature of the cells and proceed as usual to compute a chi-square statistic. There are two cases. If, following Chernoff and Lehmann, we estimate θ_o at this stage by $\hat{\theta}_n$, the chi-square statistic is

$$T_n = \sum_{i=1}^{M} \frac{[N_i - np_i(\hat{\theta}_n, \hat{\theta}_n)]^2}{np_i(\hat{\theta}_n, \hat{\theta}_n)} .$$

Under suitable regularity conditions, the law of T_n again approaches the law of (1). However, if we choose to estimate θ_o from the grouped data as in the standard case, we obtain estimators $\bar{\theta}_n$ by solving the equations

$$(2) \qquad \sum_{i=1}^{M} \frac{N_i}{p_i(\theta, \hat{\theta}_n)} \frac{\partial}{\partial \theta_j} P_i(\theta, \hat{\theta}_n) = 0 \qquad j=1,\ldots,m$$

for θ. Here θ_j is the jth component of the m-dimensional parameter θ. Note that (2) are just the usual equations for the maximum likelihood estimators from grouped data (Cramer (1946), equation 30,3,3a), ignoring the presence of the random $\hat{\theta}_n$. The chi-square statistic is now

$$S_n = \sum_{i=1}^{M} \frac{[N_i - np_i(\bar{\theta}_n, \hat{\theta}_n)]^2}{np_i(\bar{\theta}_n, \hat{\theta}_n)} .$$

Watson (1959), section 4, observed that $\bar{\theta}_n$ have the same asymptotic behavior as the corresponding estimators from fixed cells and hence that the law of S_n under the null hypothesis approaches χ^2_{M-m-1} just as in the fixed cell case.

The proof of these results depends at a crucial point on the assumption that $X_1,\ldots,X_n$ are univariate. The proof is also quite long, being given in full only in Roy's unpublished thesis. It is now possible, by using modern tools for weak convergence of processes, to greatly simplify the univariate proof and to extend it to the general case of k-variate observations. This was done for T_n by the author in Moore (1970). The present paper seeks to simplify the method and observe that it applies equally to S_n. It may be read as a supplement to section 4 of the excellent survey paper of Watson (1959).

2. BASIC RESULTS

Let $X_1,\ldots,X_n$ be a random sample from a k-variate distribution function $F(x|\theta)$ having density $f(x|\theta)$ which is continuous in both variables and continuously differentiable in θ. Here $\theta'=(\theta_1,\ldots,\theta_m)$. (Vectors are assumed to be column vectors, with prime denoting transpose.) We partition k-space by rectangular cells formed by partitioning the x_i-axis by

$$-\infty = \xi_{i0}(\hat{\theta}_n) < \xi_{i1}(\hat{\theta}_n) < \ldots < \xi_{i\nu_i}(\hat{\theta}_n) = \infty$$

where each $\xi_{ij}(\theta)$ is continuously differentiable in θ. There are $M = \Pi_{i=1}^{k} \nu_i$ cells in the resulting partition, which we index 1 to M in an arbitrary manner. The probability $p_i(\theta,\hat{\theta}_n)$ that an observation on $F(x|\theta)$ falls in the ith cell can be expressed by a familiar difference operator applied to $F(x|\theta)$ (see Cramer (1946), 8.3.3, for example.) We define this operator by writing

$$p_i(\theta,\hat{\theta}) = \Delta_i^{\hat{\theta}} F(x|\theta)$$

so that the superscript denotes the value $\hat{\theta}$ at which the partitioning functions $\xi_{ij}(\hat{\theta})$ are evaluated.

Let N_i be the number of observations falling in the ith cell. N_i is <u>not</u> multinomial, since the cells are themselves functions of the observations. But if we let n_i be the number of observations falling in the cells formed by $\xi_{ij}(\theta_o)$, where θ_o is the true parameter value, the n_i are multinomial with cell probabilities $p_i(\theta_o,\theta_o)$. The primary difficulty in these proofs is the asymptotic evaluation of N_i-n_i.

We will meet this difficulty with a simple result on weak convergence of processes. Let S be a separable metric space of real valued functions on a subset E of euclidean k-space R_k. We assume that S contains the space C(E) of functions continuous on E and that the topology of S is such that convergence in S to a member of C(E) is uniform.

Lemma 1. Suppose $X_n(\cdot)$ are random elements of S which converge weakly to a random element $X(\cdot)$ such that $P[X(\cdot) \varepsilon C(E)]=1$. Then if ξ_n are random variables taking values in E and $\xi_n \to c$ in probability, $X_n(\xi_n)-X_n(c) \to 0$ in probability.

Proof. The terminology is that of Billingsley (1968). Consider the random element (X_n,ξ_n) of the separable metric space SxR_k. Since both marginal distributions converge weakly and that of ξ_n converges in probability, Theorem 4.4 of Billingsley (1968) implies that (X_n,ξ_n) converges weakly on SxR_k to (X,c). Since $X(\cdot)$ is continuous w.p. 1 and convergence in S to a continuous limit is uniform, the function ϕ: $SxR_k \to R_1$ defined by $\phi(f,a)=f(a)-f(c)$ is continuous w.p. 1 with respect to the distribution of (X,c). Thus by Billingsley's Theorem 5.1, $\phi(X_n,\xi_n)$ converges in law to $\phi(X,c)$. But $\phi(X,c) \equiv 0$, so we have the result of the lemma.

Let now $u_i(\theta,\hat{\theta})$ be the m-vector with components

$$u_{ij}(\theta,\hat{\theta}) = \frac{\partial}{\partial\theta_j} p_i(\theta,\hat{\theta}) \qquad j=1,\ldots,m$$

and let $v_i(\theta,\hat{\theta})$ be the m-vector with components

$$v_{ij}(\theta,\hat{\theta}) = \frac{\partial}{\partial\hat{\theta}_j} p_i(\theta,\hat{\theta}).$$

Our basic result, due to Roy in the univariate case with a lengthy proof, is the following.

Lemma 2. If F and the ξ_{ij} satisfy the conditions stated above, and if the $v_{ij}(\theta,\hat{\theta})$ are continuous in $\hat{\theta}$, then when θ_o is true

$$n^{-1/2}(N_i-n_i) = \sqrt{n}\, v_i'(\theta_o,\theta_o)(\hat{\theta}_n-\theta_o)+o_p(1). \tag{3}$$

Proof. Let $F_n(x)$ be the empiric df of $X_1,\ldots,X_n$ and $W_n(x)=\sqrt{n}\,(F_n(x)-F(x|\theta_o))$ the empiric df process. Then

$$\begin{aligned} n^{-1/2}(N_i-n_i) &= \sqrt{n}\,[\Delta_i^{\hat{\theta}} F_n(x)-\Delta_i^{\theta_o} F_n(x)] \\ &= \{\Delta_i^{\hat{\theta}} W_n(x)-\Delta_i^{\theta_o} W_n(x)\}+\sqrt{n}\{\Delta_i^{\hat{\theta}} F(x|\theta_o)-\Delta_i^{\theta_o}F(x|\theta_o)\} \end{aligned} \tag{4}$$

By Taylor's theorem the second term on the right in (4) is

$$\sqrt{n}\,\{p_i(\theta_o,\hat{\theta}_n)-p_i(\theta_o,\theta_o)\} = \sqrt{n}\sum_{j=1}^{m} v_{ij}(\theta_o,\theta^*)(\hat{\theta}_n-\theta_o)_j$$

where θ^* lies between $\hat{\theta}_n$ and θ_o. Since $\sqrt{n}\,(\hat{\theta}_n-\theta_o)$ is bounded in probability (this is actually the only property of $\hat{\theta}_n$ used here) and v_{ij} is continuous, this expression is just the right side of (3). It remains to show that the first term on the right in (4) is $o_p(1)$. Since ξ_{ij} are continuous and $\hat{\theta}_n\to\theta_o$ in probability, this will follow if $\xi_n\to c$ in probability implies $W_n(\xi_n)-W_n(c)=o_p(1)$. This is the result of Lemma 1, but S must be properly chosen. Define a continuous function H mapping R_k onto the unit cube in such a way that the random variables $Y_i=H(X_i)$ have uniform marginal distributions in each direction. If U(u) is the df of any Y_i and

$U_n(u)$ the empiric df of $Y_1,\ldots,Y_n$ then the empiric process $W_n^*(u)=\sqrt{n}\,(U_n(u)-U(u))$ satisfies $W_n(x)=W_n^*(H(x))$. The path functions of W_n^* fall in the space D_k of functions on the unit cube having only jump discontinuities. For k=1, D_k is the familiar space D[0,1] of Billingsley (1968), Chapter 3. For any k it is possible to define on D_k a metric satisfying the conditions of Lemma 1 under which W_n^* converges weakly to a Gaussian process W* which has continuous paths w.p. 1. The topological results are due to Neuhaus (1970) while even stronger convergence results are contained in Dudley (1966). Since $H(\xi_n)\to H(c)$ in probability, we have by Lemma 1 that

$$W_n(\xi_n)-W_n(c) = W_n^*(H(\xi_n))-W_n^*(H(c)) = o_p(1).$$

3. LIMITING DISTRIBUTION OF T_n and S_n

Lemma 2 can be combined with elementary arguments to yield the multivariate random cells version of the Chernoff-Lehmann theorem. Suppose we have the following regularity conditions in addition to those already stated.

(A1) For i=1,...,m

$$\frac{\partial}{\partial\theta_i}\int f(x|\theta)\,dx = \int\frac{\partial}{\partial\theta_i} f(x|\theta)\,dx.$$

(A2) For all θ the information integrals

$$J_{ij} = \frac{\partial \log f}{\partial\theta_i}\frac{\partial \log f}{\partial\theta_j} f(x|\theta)dx$$

are finite and the mxm matrix $J=||J_{ij}||$ is positive definite.

(A3) $M > m$ and for any θ_o the $m \times M$ matrix with rows $u_i(\theta_o,\theta_o)'$ has rank m.

(A4) The estimators θ_n satisfy

$$\sqrt{n}\,(\hat{\theta}_n-\theta_o) = n^{-1/2} \sum_{i=1}^{n} J^{-1} A(X_i)+o_p(1)$$

where

$$A(x)' = \left(\frac{\partial \log f}{\partial \theta_1}, \ldots, \frac{\partial \log f}{\partial \theta_m}\right)\Big|_{\theta=\theta_o} .$$

(A5) All u_{ij} and v_{ij} are continuous in both variables.

Condition (A4) is of course satisfied by the maximum likelihood estimators from ungrouped data in the regular case.

Theorem 1. Under the stated regularity conditions T_n converges in distribution to the law of a random variable of form (1).

Proof. Using Taylor's theorem, convergence in law of $\sqrt{n}\,(\hat{\theta}_n-\theta_o)$ and continuity of u_i and v_i as well as Lemma 2 we obtain

$$\begin{aligned} n^{-1/2}(N_i-np_i(\hat{\theta}_n,\hat{\theta}_n)) &= n^{-1/2}(n_i-np_i(\theta_o,\theta_o))+n^{-1/2}(N_i-n_i) \\ &\quad -\sqrt{n}\,\{p_i(\hat{\theta}_n,\hat{\theta}_n)-p_i(\theta_o,\theta_o)\} \\ &= n^{-1/2}(n_i-np_i)+v_i'\,\sqrt{n}\,(\hat{\theta}_n-\theta_o) \\ &\quad -\{u_i'\,\sqrt{n}\,(\hat{\theta}_n-\theta_o)+v_i'\,\sqrt{n}\,(\hat{\theta}_n-\theta_o)\}+o_p(1) \\ &= n^{-1/2}(n_i-np_i)-u_i'\,\sqrt{n}\,(\hat{\theta}_n-\theta_o)+o_p(1) \end{aligned}$$

where the argument θ_o is assumed wherever suppressed. Since p_i is continuous, $p_i(\hat{\theta}_n,\hat{\theta}_n)/p_i(\theta_o,\theta_o)\to 1$ in probability. Applying this fact and (A4) gives

$$\frac{N_i - np_i(\hat{\theta}_n,\hat{\theta}_n)}{[np_i(\hat{\theta}_n,\hat{\theta}_n)]^{1/2}} = \frac{n_i - np_i}{[np_i]^{1/2}} - \frac{u_i'}{\sqrt{p_i}} n^{-1/2} \sum_{s=1}^{n} J^{-1}A(X_s) + o_p(1)$$

which if we define

$$\begin{aligned} C_i(x) &= 1 - p_i(\theta_o,\theta_o) && \text{if } x \text{ falls in the ith cell} \\ &= -p_i(\theta_o,\theta_o) && \text{otherwise} \end{aligned}$$

becomes

$$n^{-1/2} \sum_{s=1}^{n} \frac{1}{\sqrt{p_i}} \{C_i(X_s) - u_i' J^{-1}A(X_s)\} + o_p(1). \tag{5}$$

The M quantities (5) for $i=1,\ldots,M$ are clearly asymptotically multivariate normal by the multivariate central limit theorem. Since T_n is the sum of squares of these quantities, it only remains to compute the covariance matrix and investigate its eigenvalues. Details of this are given by Moore (1970) using methods outlined in section 4 of Watson (1959).

Turning to S_n, we need only add to the method of proof used above knowledge of the behavior of the estimators $\bar{\theta}_n$ obtained by solving the equations (2). These estimators were first studied in the univariate case by Watson. The proof, which depends on Lemma 2, is much the same in the k-variate case. Since Watson did not publish a proof, we sketch it here. Recall the condition that θ_o is understood where arguments are suppressed and define B as the Mxm matrix with entries

$$B_{ij} = p_i^{-1/2} u_{ij}.$$

Lemma 3 (Watson). Under the stated regularity conditions

(6) $$\sqrt{n}\,(\overline{\theta}_n-\theta_o) = (B'B)^{-1} B' q_n+o_p(1)$$

where q_n has components $(n_i-np_i)/(np_i)^{1/2}$, i=1,...,M.

Proof. Note that B'B is nonsingular by (A3). Note also that (6) is exactly the asymptotic behavior of maximum likelihood estimators from data grouped in fixed cells (Cramer (1946), equation 30.3.17). To prove (6), rewrite (2) as

(7) $$\sum_{i=1}^{M} \frac{(N_i-np_i(\hat{\theta}_n,\hat{\theta}_n))\; u_{ij}(\overline{\theta}_n,\hat{\theta}_n)}{\sqrt{n}\; p_i(\overline{\theta}_n,\hat{\theta}_n)} = \sum_{i=1}^{M} \frac{\sqrt{n}\;(p_i(\overline{\theta}_n,\hat{\theta}_n)-p_i(\hat{\theta}_n,\hat{\theta}_n))}{p_i(\overline{\theta}_n,\hat{\theta}_n)} u_{ij}(\overline{\theta}_n,\hat{\theta}_n).$$

The left side of (7) can be expressed as

(8) $$\sum_{i=1}^{M} \frac{(n_i-np_i)u_{ij}(\overline{\theta}_n,\hat{\theta}_n)}{\sqrt{n}\; p_i(\overline{\theta}_n,\hat{\theta}_n)} + \sum_{i=1}^{M} \frac{(N_i-n_i)\; u_{ij}(\overline{\theta}_n,\hat{\theta}_n)}{\sqrt{n}\; p_i(\overline{\theta}_n,\hat{\theta}_n)}$$

$$- \sum_{i=1}^{M} \sqrt{n}\;(p_i(\hat{\theta}_n,\hat{\theta}_n)-p_i)\; \frac{u_{ij}(\overline{\theta}_n,\hat{\theta}_n)}{p_i(\overline{\theta}_n,\hat{\theta}_n)}.$$

Consistency of $\overline{\theta}_n$ can be proved as is done for the fixed cell case by Rao (1965), section 5e.2, using consistency of $\hat{\theta}_n$. Routine computations making use of Lemma 2 in the second term of (8) and Taylor's theorem in the third term show that (8) is

$$\sum_{i=1}^{M} \frac{n_i - np_i}{\sqrt{np_i}} B_{ij} - \sum_{i=1}^{M} [u_i' \sqrt{n}\ (\hat{\theta}_n - \theta_o)] \frac{B_{ij}}{\sqrt{p_i}} + o_p(1).$$

This is just the jth entry of the vector

(9) $$B'q_n - B'B \sqrt{n}\ (\hat{\theta}_n - \theta_o) + o_p(1).$$

Another use of Taylor's theorem shows that the right side of (7) is

(10) $$\sum_{s=1}^{m} \sqrt{n}\ (\bar{\theta}_n - \hat{\theta}_n)_s \sum_{i=1}^{M} \frac{u_{is}(\theta_n^*, \hat{\theta}_n)\ u_{ij}(\bar{\theta}_n, \hat{\theta}_n)}{p_i(\bar{\theta}_n, \hat{\theta}_n)}$$

where θ_n^* lies between $\bar{\theta}_n$ and $\hat{\theta}_n$. Write

$$\bar{\theta}_n - \hat{\theta}_n = (\bar{\theta}_n - \theta_o) - (\hat{\theta}_n - \theta_o)$$

and observe that $\sqrt{n}\ (\bar{\theta}_n - \theta_o)$ is bounded in probability since all other terms in the expanded version of (7) are. (10) therefore becomes the jth entry of the vector

(11) $$B'B\sqrt{n}\,(\bar{\theta}_n-\theta_o)-B'B\sqrt{n}\,(\hat{\theta}_n,\theta_o) + o_p(1)$$

and equating (9) and (11) proves the lemma.

Theorem 2. Under the stated regularity conditions the law of S_n converges to χ^2_{M-m-1}.

Proof. Proceeding exactly as in the proof of Theorem 1 with $\bar{\theta}_n$ replacing $\hat{\theta}_n$ where appropriate gives

$$n^{-1/2}(N_i-np_i(\bar{\theta}_n,\hat{\theta}_n)) = n^{-1/2}(n_i-np_i)-u_i'\sqrt{n}\,(\bar{\theta}_n-\theta_o) + o_p(1).$$

Substituting the result of Lemma 3 shows that the vector with components $(N_i-np_i(\bar{\theta}_n,\hat{\theta}_n))/[np_i(\bar{\theta}_n,\hat{\theta}_n)]^{1/2}$ is equal to

$$(I-B(B'B)^{-1}B')\ q_n + o_p(1)$$

where I is the identity matrix. This is the same result as in the fixed cell case ([3], equation 30.3.18) and therefore S_n has the limiting distribution χ^2_{M-m-1} obtained in that case.

REFERENCES

Billingsley, Patrick (1968). Convergence of Probability Measures. Wiley, New York.

Chernoff, H. and Lehmann, E.L. (1954). The use of maximum likelihood estimates in χ^2 test for goodness of fit. Ann. Math. Statist. **25**, 579-586.

Cramer, H. (1946). Mathematical Methods of Statistics. Princeton University Press, Princeton, New Jersey.

Dudley, R.M. (1966). Weak convergence of probabilities on nonseparable metric spaces and empirical measures on euclidean spaces. Illinois Jour. Math. **10**, 109-126.

Moore, D.S. (1970). A chi-square statistic with random cell boundaries. To appear in Ann. Math. Statist.

Neuhaus, Georg (1969). On weak convergence of stochastic processes with multi-dimensional time parameter. Technical report, Institut für Mathematische Statistik, University of Münster.

Rao, C.R. (1965). Linear Statistical Inference and its Applications. Wiley, New York.

Roy, A.R. (1956). On χ^2 statistics with variable intervals. Technical report, Department of Statistics, Stanford University.

Watson, G.S. (1958). On chi-square goodness-of-fit tests for continuous distributions. J. Roy. Statist. Soc., Series B, **20**, 44-61.

Watson, G.S. (1959). Some recent results in χ^2 goodness-of-fit tests. Biometrics **15**, 440-468.