The Annals of Statistics
1975, Vol. 3, No. 4, 965–968

A NOTE ON CHI-SQUARE STATISTICS WITH RANDOM CELL BOUNDARIES[1]

BY F. H. RUYMGAART

Mathematisch Centrum, Amsterdam

Moore (1971) derives the limiting distribution of chi-square statistics for testing goodness of fit to k-variate parametric families, where the cell boundaries are allowed to be functions of the estimated parameter values. The only point at which random cells, multivariate observations etc. require a deviation from methods of proof given by Cramér (1946) is in the proof of the asymptotic negligibility of the remainder terms. Attention is drawn to an alternative proof of this asymptotic negligibility, which turns out to be an immediate consequence of a modification of Lemma 1 by Bahadur (1966) in more dimensions.

1. Introduction and summary. Suppose that $X_1, X_2, \cdots$ is a sequence of mutually independent and identically distributed k-dimensional random vectors. All random vectors are supposed to be defined on a single probability space and their common distribution function (df) F_θ depends on an m-dimensional parameter θ which is restricted to an open set $\mathscr{T} \subset \mathbb{R}^m$. (If p is an arbitrary positive integer, p-dimensional number space is denoted by $\mathbb{R}^p$.) Given any positive integer n, the empirical df F_n based on the first n random vectors in the sequence is defined in the usual way.

In the context of testing goodness of fit, as described in a paper by Moore [4], $\mathbb{R}^k$ is partitioned into a fixed finite number of cells, where the cell boundaries are allowed to be functions of the estimated parameter values. Proceeding along the lines of Moore's paper, for $i = 1, 2, \cdots, k$ a nonrandom partition of the x_i-axis is defined by functions of $\theta \in \mathscr{T}$, satisfying

$$-\infty = \xi_{i,0}(\theta) < \xi_{i,1}(\theta) < \cdots < \xi_{i,\nu_i-1}(\theta) < \xi_{i,\nu_i}(\theta) = \infty .$$

These partitions of the axes induce $\nu = \prod_{i=1}^k \nu_i$ cells in $\mathbb{R}^k$, formed by the Cartesian products

$$\text{(1.1)} \qquad \mathsf{X}_{i=1}^k \, (\xi_{i,j_i}(\theta), \xi_{i,j_i+1}(\theta)) ,$$

where j_i is an arbitrary number in $\{0, 1, \cdots, \nu_i - 1\}$ for $i = 1, \cdots, k$. According to a specific enumeration the cells in (1.1) will be denoted by $I_\sigma(\theta)$ for $\sigma = 1, 2, \cdots, \nu$. Suppose that $\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, \cdots, X_n)$ is an estimator of θ for each n. To each $I_\sigma(\theta)$ there corresponds the random cell $I_\sigma(\hat{\theta}_n)$ when θ is replaced by $\hat{\theta}_n$ in (1.1).

Received January 1974; revised August 1974.

[1] This paper is a revised version of MCR SW 21/73.

AMS 1970 *subject classifications.* Primary 62E20; Secondary 62F05.

Key words and phrases. Chi-square statistic, goodness of fit, cell boundaries depending on estimated parameter values, asymptotic negligibility, empirical distribution function.

The mass assigned to any Borel set $B \subset \mathbb{R}^k$ by the df F_θ will be denoted by $F_\theta\{B\}$ and, similarly, the mass assigned to B by the empirical df F_n will be denoted by $F_n\{B\}$. The latter, of course, equals the number of elements in the set $B \cap \{X_1, X_2, \cdots, X_n\}$, divided by n. For any two subsets $A, B \subset \mathbb{R}^k$ the difference $A - B$ is defined as $A \cap B^c$ and for any $y \in \mathbb{R}^m$ the norm is denoted by $||y||$.

In the search for the asymptotic distribution of chi-square statistics

$$T_n = \sum_{\sigma=1}^{\nu} n[F_n\{I_\sigma(\hat{\theta}_n)\} - F_{\hat{\theta}_n}\{I_\sigma(\hat{\theta}_n)\}]^2[F_{\hat{\theta}_n}\{I_\sigma(\hat{\theta}_n)\}]^{-1}, \tag{1.2}$$

one can write $n^{\frac{1}{2}}[F_n\{I_\sigma(\hat{\theta}_n)\} - F_{\hat{\theta}_n}\{I_\sigma(\hat{\theta}_n)\}] = \sum_{i=1}^{2}(A_{i\sigma n} + B_{i\sigma n})$, where

$$\begin{aligned}
A_{1\sigma n} &= n^{\frac{1}{2}}[F_n\{I_\sigma(\theta_0)\} - F_{\theta_0}\{I_\sigma(\theta_0)\}], \\
A_{2\sigma n} &= n^{\frac{1}{2}}[F_{\theta_0}\{I_\sigma(\hat{\theta}_n)\} - F_{\hat{\theta}_n}\{I_\sigma(\hat{\theta}_n)\}], \\
B_{1\sigma n} &= n^{\frac{1}{2}}[F_n\{I_\sigma(\hat{\theta}_n) - I_\sigma(\theta_0)\} - F_{\theta_0}\{I_\sigma(\hat{\theta}_n) - I_\sigma(\theta_0)\}], \\
B_{2\sigma n} &= n^{\frac{1}{2}}[F_{\theta_0}\{I_\sigma(\theta_0) - I_\sigma(\hat{\theta}_n)\} - F_n\{I_\sigma(\theta_0) - I_\sigma(\hat{\theta}_n)\}].
\end{aligned}$$

Here, and throughout the sequel, $\theta_0 \in \mathscr{T}$ is the true parameter value. The expression on the left of Moore's formula (2.2) equals $B_{1\sigma n} + B_{2\sigma n}$, but here the terms are arranged somewhat differently for reasons that will become clear in Section 2.

Attention will be restricted to the proof of

$$B_{1\sigma n} + B_{2\sigma n} = o_P(1), \qquad \text{as } n \to \infty. \tag{1.3}$$

This is the only point at which random cells, multivariate observations etc. require a deviation from methods of proof given by Cramér [2] and this also constitutes one of the main mathematical difficulties.

To prove (1.3), Moore essentially uses the assumptions

(1) $\xi_{i,j}$ is continuous at θ_0 for $i = 1, 2, \cdots, k$ and $j = 1, 2, \cdots, \nu_i - 1$;
(2) $||\hat{\theta}_n - \theta_0|| = o_P(1)$, as $n \to \infty$;
(3) F_{θ_0} is continuous on $\mathbb{R}^k$.

Actually, in Moore's paper stronger conditions are imposed, implying

(1′) $\partial\xi_{i,j}(\theta)/\partial\theta_s$ exists and is continuous in a neighborhood of θ_0 for $i = 1, 2, \cdots, k$, $j = 1, 2, \cdots, \nu_i - 1$ and $s = 1, 2, \cdots, m$;
(2′) $||\hat{\theta}_n - \theta_0|| = O_P(n^{-\frac{1}{2}})$, as $n \to \infty$;
(3′) F_{θ_0} has a continuous density with respect to Lebesgue measure on $\mathbb{R}^k$.

These conditions, however, are only used to deal with the A-terms.

Moore obtains (1.3) by appealing to rather advanced papers by Dudley [3] and Neuhaus [5]. It is the purpose of this note to draw attention to an alternative proof of (1.3), by showing that it is an almost immediate consequence of a modification of Lemma 1 by Bahadur [1] in more dimensions. In this form Bahadur's lemma has been given by W. R. van Zwet. For completeness the lemma is formulated. A proof may be found in [6], [7] for $k = 2$. (The proof for $k > 2$ is similar.)

For each $n = 1, 2, \cdots$ let be given a random sample of size n from an arbitrary k-variate df F. The corresponding k-variate empirical df will be denoted by F_n. By an interval in $\mathbb{R}^k$ we understand the Cartesian product of k intervals on the real line.

LEMMA (van Zwet). *Let $I_1, I_2, \cdots$ be a sequence of intervals in $\mathbb{R}^k$ and let $\mathscr{I}_n = \{I_n^* : I_n^*$ is an interval contained in $I_n\}$ for $n = 1, 2, \cdots$. Then*

$$\sup_{I_n^* \in \mathscr{I}_n} |F_n\{I_n^*\} - F\{I_n^*\}| = O_P(n^{-\frac{1}{2}}[F\{I_n\}]^{\frac{1}{2}}) , \qquad \text{as } n \to \infty , \tag{1.4}$$

uniformly in all sequences of intervals $I_1, I_2, \cdots$ and all k-variate df's *F.*

In Section 2 it will be shown that under assumptions (1)—(3) the lemma yields that $B_{1\sigma n} + B_{2\sigma n} = o_P(1)$, as $n \to \infty$, and even $B_{1\sigma n} + B_{2\sigma n} = O_P(n^{-\frac{1}{4}})$, as $n \to \infty$, under assumptions (1′)—(3′).

This illustrates once more the usefulness of (this modification of) Bahadur's lemma, which has also proved an essential tool for handling some of the second order terms occurring in the proofs of asymptotic normality, under fixed alternatives, of certain nonparametric test statistics (see e.g. [8] and [6], [7]).

2. Proof of the asymptotic negligibility. Let σ be fixed. By symmetry it suffices to consider $B_{1\sigma n}$ (see (1.3)). For an arbitrary sequence of positive numbers $b_1, b_2, \cdots$ define the sets

$$\Omega_{1n} = \textstyle\bigcap_{i=1}^{k} \bigcap_{j=1}^{\nu_i - 1} \{|\xi_{i,j}(\hat{\theta}_n) - \xi_{i,j}(\theta_0)| \leqq b_n\} , \tag{2.1}$$

and (for all $i = 1, 2, \cdots, k$ and $j = 1, 2, \cdots, \nu_i - 1$) the intervals

$$I_{n,i,j} = \mathbb{R}^{i-1} \times [\xi_{i,j}(\theta_0) - b_n, \xi_{i,j}(\theta_0) + b_n] \times \mathbb{R}^{k-i} . \tag{2.2}$$

Note that for all $\omega \in \Omega_{1n}$ the relation $\{I_\sigma(\hat{\theta}_n) - I_\sigma(\theta_0)\} = \bigcup_{\rho=1}^{K} I_{n,\sigma,\rho}^*$ holds. Here K is fixed and the $I_{n,\sigma,\rho}^*$ are disjoint possibly empty intervals contained in the $I_{n,i,j}$. For brevity denote

$$c_n = c_n(b_n) = \max_{i,j} \{F_{\theta_0}\{I_{n,i,j}\}\} . \tag{2.3}$$

From now on let $\varepsilon > 0$ be arbitrary but fixed. The lemma of Section 1 ensures the existence of a number $M = M_\varepsilon$ such that the set

$$\Omega_{2n} = \textstyle\bigcap_{\rho=1}^{K} \{|F_n\{I_{n,\sigma,\rho}^*\} - F_{\theta_0}\{I_{n,\sigma,\rho}^*\}| \leqq Mn^{-\frac{1}{2}}c_n^{\frac{1}{2}}\} \tag{2.4}$$

has probability $P(\Omega_{2n}) \geqq 1 - \frac{1}{2}\varepsilon$ for all $n = 1, 2, \cdots$. Denoting the indicator function of the set $\Omega_{1n} \cap \Omega_{2n}$ by $\chi(\Omega_{1n} \cap \Omega_{2n})$ it follows that

$$\chi(\Omega_{1n} \cap \Omega_{2n})|B_{1\sigma n}| \leqq KMc_n^{\frac{1}{2}} . \tag{2.5}$$

If assumptions (1)—(3) are satisfied, (1) and (2) entail the existence of a sequence $b_{1n} = b_{1n\varepsilon} = o(1)$ such that $P(\Omega_{1n}) \geqq 1 - \frac{1}{2}\varepsilon$ for all n, choosing $b_n = b_{1n}$ in (2.1). Substitution of b_{1n} in (2.2) and (2.3) yields that $c_{1n} = c_n(b_{1n}) = o(1)$ because of (3). Consequently the quantity on the right in (2.5) is $o(1)$. Since, moreover, $P(\Omega_{1n} \cap \Omega_{2n}) \geqq 1 - \varepsilon$ for arbitrary $\varepsilon > 0$ and all n, it follows from (2.5) that $B_{1\sigma n} = o_P(1)$, as $n \to \infty$.

Under assumptions (1′)—(3′) there exists a sequence $b_{2n} = b_{2n\varepsilon} = O(n^{-\frac{1}{2}})$ such that $P(\Omega_{1n}) \geqq 1 - \frac{1}{2}\varepsilon$ for all n, by choosing $b_n = b_{2n}$ in (2.1). This follows from (1′) and (2′). Substitution of b_{2n} in (2.2) and (2.3) entails $c_{2n} = c_n(b_{2n}) = O(n^{-\frac{1}{2}})$ by (3′). In a similar way as before it follows that in this case even $B_{1\sigma n} = O_P(n^{-\frac{1}{4}})$, as $n \to \infty$.

REFERENCES

[1] Bahadur, R. R. (1966). A note on quantiles in large samples. *Ann. Math. Statist.* **37** 577–580.

[2] Cramér, H. (1946). *Mathematical Methods of Statistics.* Princeton Univ. Press.

[3] Dudley, R. M. (1966). Weak convergence of probabilities on nonseparable metric spaces and empirical measures on Euclidean spaces. *Illinois J. Math.* **10** 109–126.

[4] Moore, D. S. (1971). A chi-square statistic with random cell boundaries. *Ann. Math. Statist.* **42** 147–156.

[5] Neuhaus, G. (1971). On weak convergence of stochastic processes with multidimensional time parameter. *Ann. Math. Statist.* **42** 1285–1295.

[6] Ruymgaart, F. H. (1973). *Asymptotic Theory of Rank Tests for Independence.* Mathematical Centre Tracts 43, Amsterdam.

[7] Ruymgaart, F. H. (1974). Asymptotic normality of nonparametric tests for independence. *Ann. Statist.* **2** 892–910.

[9] Sen, P. K. (1970). Asymptotic distribution of a class of multivariate rank order statistics. *Calcutta Statist. Assoc. Bull.* **19** 23–31.

AFD. Statistiek
Mathematisch Centrum
2e Boerhaavestraat 49
Amsterdam-0
Holland