stichting
mathematisch
centrum

MC

AFDELING MATHEMATISCHE STATISTIEK SW 21/73 NOVEMBER

F.H. RUYMGAART
A NOTE ON A PAPER BY D.S. MOORE
ON CHI-SQUARE STATISTICS

Prepublication

2e boerhaavestraat 49 amsterdam

Printed at the Mathematical Centre, 49, 2e Boerhaavestraat, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O), by the Municipality of Amsterdam, by the University of Amsterdam, by the Free University at Amsterdam, and by industries.

AMS (MOS) subject classification scheme (1970): 62E20, 62F05

A note on a paper by D.S. Moore on chi-square statistics *)

by

F.H. Ruymgaart

Summary

In this note we draw attention to an elementary proof of the asymptotic negligibility of the remainder terms in a paper by D.S. Moore (1971) on the limiting distribution of chi-square statistics. The asymptotic negligibility turns out to be an immediate consequence of a modification of Lemma 1 by Bahadur (1966) in more dimensions.

*) This paper is not for review; it is meant for publication in a journal.

1. INTRODUCTION

Suppose that we are given a sequence $X_1, X_2, \ldots$ of mutually independent and identically distributed k-dimensional random vectors. All random vectors are supposed to be defined on a single probability space (Ω, A, P) and their common distribution function (df) F_θ depends on an m-dimensional parameter θ which is restricted to an open subset T of m-dimensional number space $\mathbb{R}^m$. Given any positive integer n, we define the empirical df F_n based on the first n random vectors of the sequence in the usual way.

In the context of testing goodness of fit, as described in a paper by Moore (1971), $\mathbb{R}^k$ is partitioned into a fixed finite number of cells, where the cell boundaries are allowed to be functions of the estimated parameter values. Let us proceed along the lines of Moore's paper and define for $i = 1,2,\ldots, k$ a non-random partition of the x_i - axis by functions of $\theta \in T$, satisfying

$$(1.1) \qquad -\infty = \xi_{i,0}(\theta) < \xi_{i,1}(\theta) < \ldots < \xi_{i,\nu_i-1}(\theta) < \xi_{i,\nu_i}(\theta) = \infty .$$

The partitions of the axes induce a partition of $\mathbb{R}^k$ into $\nu = \prod_{i=1}^{k} \nu_i$ cells. According to a specific enumeration these cells will be denoted by $I_\sigma(\theta)$, $\sigma = 1,2,\ldots, \nu$. Suppose that for each positive integer n we have an estimator $\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, \ldots, X_n)$ of θ. To $I_\sigma(\theta)$ there corresponds the random cell $I_\sigma(\hat{\theta}_n)$ when θ is replaced by $\hat{\theta}_n$ in (1.1). 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 $\{X_1, X_2, \ldots, X_n\}$ contained in B, divided by n.

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

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

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

$$A_{1n} = n^{\frac{1}{2}}\,[F_n\{I_\sigma(\theta_0)\} - F_{\theta_0}\{I_\sigma(\theta_0)\}],$$

$$A_{2n} = n^{\frac{1}{2}}\,[F_{\theta_0}\{I_\sigma(\hat{\theta}_n)\} - F_{\hat{\theta}_n}\{I_\sigma(\hat{\theta}_n)\}],$$

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

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

and where θ_0 is the true parameter value. The expression on the left of Moore's formula (2.2) equals $B_{1n} + B_{2n}$, but we have arranged the terms somewhat differently for purposes that will become clear below.

An essential part of the proof of Theorem 1 by Moore (1971) consists of showing that

$$(1.3) \qquad B_{1n} + B_{2n} \to_P 0\,,$$

as $n \to \infty$. Moore derives this result by appealing to rather advanced papers by Dudley (1966) and Neuhaus (1969). It is the purpose of this note to draw attention to a more elementary proof of (1.3), by showing that it is an immediate consequence of a modification of Lemma 1 by Bahadur (1966) in more dimensions. In this form Bahadur's lemma has been given by W.R. van Zwet. For completeness we shall formulate the lemma, a proof of which may

be found in Ruymgaart (1972, 1973) for k = 2. (The proof for k > 2 is completely similar.) Suppose that for each n = 1,2,... we are given a random sample of size n from an arbitrary fixed k-variate df F (continuous or not). The corresponding k-variate empirical df will be denoted by F_n. By an interval I in $\mathbb{R}^k$ we understand the product set of k intervals on the real line.

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

$$\sup_{I_n^* \in \mathcal{I}_n} |F_n\{I_n^*\} - F\{I_n^*\}| = O_P([n^{-1} F\{I_n\}]^{\frac{1}{2}}),$$

uniformly in all sequences of intervals I_1, I_2,... and all k-variate dfs F (continuous or not).

Let us for the moment restrict attention to regularity conditions on F_{θ_0}, although some other conditions will also be needed (see Section 2). Using only continuity of F_{θ_0} it follows almost immediately from the lemma that $B_{1n} + B_{2n} = o_P(1)$, as $n \to \infty$, which is Moore's result. In Moore's paper it is assumed, for other purposes, that F_{θ_0} has a continuous density. Under the latter stronger assumption we deduce from the lemma in quite the same way that $B_{1n} + B_{2n} = O_P(n^{-1/4})$, as $n \to \infty$. The above illustrates once more the usefulness of (this modification of) Bahadur's lemma, which has also proved essential for handling some of the second order terms occurring in the proofs of asymptotic normality, under fixed alternatives, of certain nonparametric test statistics (Sen (1970), Ruymgaart (1972, 1973)).

2. PROOF OF THE ASYMPTOTIC NEGLIGIBILITY

The first assumption needed for the proof of (1.3) is that the function

$$(2.1) \qquad \partial \xi_{i,j}(\theta) \,/\, \partial \theta_1$$

exists and is continuous for $\theta \in T$ and $i = 1,2,\ldots, k$, $j = 1,2,\ldots, \nu_i - 1$, $1 = 1,2,\ldots, m$.

The second assumption is that the sequence of estimators $\hat{\theta}_1, \hat{\theta}_2, \ldots$ satisfies

$$(2.2) \qquad |\hat{\theta}_n - \theta_0| = O_P(n^{-\frac{1}{2}}),$$

as $n \to \infty$, where θ_0 is the true parameter value.

These assumptions guarantee for each $\varepsilon > 0$ the existence of a constant $M_1 = M_{1\varepsilon}$ such that the set

$$(2.3) \qquad \Omega_{1n} = \cap_{i=1}^{k} \cap_{j=1}^{\nu_i - 1} \{|\xi_{i,j}(\hat{\theta}_n) - \xi_{i,j}(\theta_0)| \le M_1 n^{-\frac{1}{2}}\}$$

has probability $P(\Omega_{1n}) \ge 1 - \varepsilon/2$ for all $n = 1,2,\ldots$

By symmetry we need only consider B_{1n}. Let us fix σ and introduce for all $i = 1,2,\ldots, k$ and $j = 1,\ldots, \nu_i - 1$ the intervals

$$I_{n,i,j} = \mathbb{R}^{i-1} \times [\xi_{i,j}(\theta_0) - M_1 n^{-\frac{1}{2}}, \xi_{i,j}(\theta_0) + M_1 n^{-\frac{1}{2}}] \times \mathbb{R}^{k-i},$$

$$I^*_{n,i,j} = I_{n,i,j} \cap \{I_\sigma(\hat{\theta}_n) - I_\sigma(\theta_0)\}.$$

Note that for all $\omega \in \Omega_{1n}$ we have $\{I_\sigma(\hat{\theta}_n) - I_\sigma(\theta_0)\} = \cup_{i=1}^{k} \cup_{j=1}^{\nu_i - 1} I^*_{n,i,j}$.

If F_{θ_0} is given to be only continuous it follows that

$\max_{i,j} F_{\theta_0}\{I_{n,i,j}\} = c_n$, where $c_n \to 0$ as $n \to \infty$. The lemma of Section 1 ensures the existence of a number $M_2 = M_{2\varepsilon}$, such that the set

$$\Omega_{2n} = \cap_{i=1}^{k} \cap_{j=1}^{\nu_i - 1} \{|F_n\{I^*_{n,i,j}\} - F_{\theta_0}\{I^*_{n,i,j}\}| \le M_2 n^{-\frac{1}{2}} c_n^{\frac{1}{2}}\} \tag{2.4}$$

has probability $P(\Omega_{2n}) \ge 1 - \varepsilon/2$ for all $n = 1,2,\ldots$ Denoting the characteristic function of the set $\Omega_{1n} \cap \Omega_{2n}$ by $\chi(\Omega_{1n} \cap \Omega_{2n})$ it follows that

$$\chi(\Omega_{1n} \cap \Omega_{2n}) \mid B_{1n} \mid \le [\sum_{i=1}^{k} (\nu_i - 1)] M_2 c_n^{\frac{1}{2}} \to 0, \tag{2.5}$$

as $n \to \infty$. Because $P(\Omega_{1n} \cap \Omega_{2n}) \ge 1 - \varepsilon$ for all $n = 1,2,\ldots$ and $\varepsilon > 0$ is arbitrary we may conclude from (2.5) that $B_{1n} = o_P(1)$.

In the case where F_{θ_0} has a continuous density with respect to Lebesgue measure we find that $\max_{i,j} F_{\theta_0}\{I_{n,i,j}\} \le M_3 n^{-\frac{1}{2}}$ for some constant M_3 and all $n = 1,2,\ldots$ The lemma applies in the same way so that for some constant $M_2' = M_{2\varepsilon}'$ we may use (2.4) and (2.5) with M_2 replaced by M_2' and c_n by $M_3 n^{-\frac{1}{2}}$. Consequently we now have that $B_{1n} = o_P(n^{-1/4})$.

REFERENCES

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

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

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

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

[5] Ruymgaart, F.H. (1972). Asymptotic normality of nonparametric tests for independence. To appear in *Ann. Statist.*.

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

[7] Sen, P.K. (1970). Asymptotic distribution of a class of multivariate rank order statistics. *Calcutta Statist. Ass. 19*, 23-31.