J. Austral. Math. Soc. (Series A) 60 (1996), 355–362

RATES OF CONVERGENCE TO NORMALITY FOR SAMPLES FROM A FINITE SET OF RANDOM VARIABLES

R. D. JOHN and J. ROBINSON

(Received 1 February 1991; revised 24 November 1994)

Communicated by R. A. Maller

Abstract

Rates of convergence to normality of $O(N^{-1/2})$ are obtained for a standardized sum of m random variables selected at random from a finite set of N random variables in two cases. In the first case, the sum is randomly normed and the variables are not restricted to being independent. The second case is an alternative proof of a result due to von Bahr, which deals with independent variables. Both results derive from a rate obtained by Höglund in the case of sampling from a finite population.

1991 *Mathematics subject classification* (*Amer. Math. Soc.*): 60F05.
Keywords and phrases: Berry-Esseen lemma, finite sampling.

1. Introduction

The two results of this paper make use of a theorem due to Höglund [6] relating to the rate of approach to normality of a sum of a set of elements randomly selected from a finite population. Specifically, if $x_1, \dots, x_N$ are real numbers, with $\sum_{i=1}^{N}(x_i - \bar{x})^2 > 0$, and $m < N$, then

$$(1)\quad \sup_v \left| P\left(\frac{\sum_{i=1}^{m}(x_{R_i} - \bar{x})}{\left(pq\sum_{i=1}^{N}(x_i - \bar{x})^2\right)^{1/2}} \le v \right) - \Phi(v) \right| \le \frac{B}{(pq)^{1/2}} \frac{\sum_{i=1}^{N}|x_i - \bar{x}|^3}{\left(\sum_{i=1}^{N}(x_i - \bar{x})^2\right)^{3/2}}$$

where $R_1, \dots, R_N$ is a uniform random permutation of $1, \dots, N$, $\bar{x} = N^{-1}\sum_{i=1}^{N} x_i$, $p = 1 - q = m/N$, Φ is the standard normal distribution function and B is an absolute constant. Both results investigate the consequences of replacing the constants $x_1, \dots, x_N$ by random variables $X_1, \dots, X_N$.

https://doi.org/10.1017/S1446788700037861 Published online by Cambridge University Press

The first deals with the statistic

$$T = \sum_{i=1}^{m} (Y_i - \bar{Y}) \left(pq \sum_{i=1}^{N} (Y_i - \bar{Y})^2 \right)^{-1/2} \tag{2}$$

in which

$$Y_i = X_{R_i}, \quad i = 1, \ldots, N, \tag{3}$$

where $R_1, \ldots, R_N$ are as above and independent of $X_1, \ldots, X_N$ and $\bar{Y} = N^{-1} \sum_{i=1}^{N} Y_i$. That T converges in distribution to a standard normal variable when $Y_1, \ldots, Y_N$ are finitely exchangeable was proved by Chernoff and Teicher [2]. The first result provides a rate for this convergence for a special class of exchangeable random variables, namely those obtained from a set of fairly general random variables via (3). In fact, the X_i we consider will be of the form

$$X_i = V_i + W_i \tag{4}$$

in which $V_1, \ldots, V_N$ are assumed to be independent random variables, independent of $W_1, \ldots, W_N$, which are not assumed independent.

The second result is an alternative proof of a theorem due to von Bahr [1], which deals with a rate of convergence to normality of

$$T^* = \left(\sum_{i=1}^{m} Y_i - a \right) / b \tag{5}$$

where Y_i is as in (3), a and b are now some norming constants, as opposed to the random norming of the previous result, and the X_i of (3) are independent but not necessarily identically distributed random variables. The proof of von Bahr involves a combinatoric argument of some complexity, whereas that given here proceeds via a conditioning argument using (1). An essential part of Höglund's proof of (1) is the use of the Erdös-Rényi form of the characteristic function of $\sum_{i=1}^{m} Y_i$ conditional on $X_1, \ldots, X_N$ (Erdös and Rényi [4]) and it is of some interest to note that there is a proof of the von Bahr result which uses this form. Von Bahr, himself, mentions that he could not see how it could be used. Actually, the result given here is not quite as general as von Bahr's, requiring that the X_i have finite fourth moments, compared with his requirement of finite third moments. Also, our bound is in terms of $(Npq)^{-1/2}$ whereas von Bahr's involves $(Np)^{-1/2}$. Our result would seem to imply that p should be bounded away from 1 as well as 0 for a rate of $O(N^{-1/2})$ to apply, but clearly, since the X's are assumed to be independent here, the Berry-Esseen result (Feller [5, p. 544]) ensures this rate when p is close to 1.

The first result is used in situations such as permutation tests for two sample problems under randomization where an assumption of independence and equal variance for the plots may be unacceptable, but where condition (4) may be assumed. The second arises in two stage sampling as stated by von Bahr [1].

2. A rate for the statistic T

Let Y_i be given by (3), where $X = (X_1, \ldots, X_N)$ is a random vector with $E(X_i) = \mu_i$, $\operatorname{var}(X_i) = \sigma_i^2$ and $E|X_i - \mu_i|^3 = \mu_{3,i}$. Put

$$S^2 = N^{-1} \sum_{i=1}^{N} (X_i - \bar{X})^2 \tag{6}$$

where $\bar{X} = N^{-1} \sum_{i=1}^{N} X_i$ and define $\bar{\mu} = N^{-1} \sum_{i=1}^{N} \mu_i$. Let T be given by (2).

THEOREM 1. *For arbitrary $\tau > 0$, there exists a constant B, depending on τ, such that*

$$\sup_v |P(T \le v) - \Phi(v)| \le B(pq)^{-1/2} N^{-3/2} \left(\sum_{i=1}^{N} \mu_{3,i} + \sum_{i=1}^{N} |\mu_i - \bar{\mu}|^3 \right) + P(S^2 < \tau).$$

PROOF. We denote by I_H and H^c the indicator function and complement, respectively, of an arbitrary set H. The constant B, here and in the sequel, is not necessarily the same at each occurrence. Let E_τ be the set where $S^2 \ge \tau$. Then using (1),

$$\begin{aligned} |P(T \le v) - \Phi(v)| &\le E(I_{E_\tau} |P(T \le v | X) - \Phi(v)|) + E(I_{E_\tau^c} |P(T \le v | X) - \Phi(v)|) \\ &\le B(pq)^{-1/2} E \left(I_{E_\tau} \sum_{i=1}^{N} \left| X_i - \bar{X} \right|^3 N^{-3/2} S^{-3} \right) + P\left(S^2 < \tau\right) \\ &\le B(pq)^{-1/2} N^{-3/2} \sum_{i=1}^{N} E \left| X_i - \bar{X} \right|^3 + P\left(S^2 < \tau\right) \end{aligned}$$

and the result follows applications of the C_r and Hölder inequalities.

We note that the first term in the bound of Theorem 1 is $O(N^{-1/2})$ subject to some condition on p which ensures it is bounded away from 0 and 1. In this case, since the X's are not independent, the p close to 1 exclusion is necessary. The term $P(S^2 < \tau)$ may well be of $O(N^{-1/2})$ for a large class of variables. As yet we have imposed no conditions on the joint distribution of $X_1, \ldots, X_N$. However the particular model we propose for the X_i, namely (4), incorporates some degree of dependency and

https://doi.org/10.1017/S1446788700037861 Published online by Cambridge University Press

non-stationarity. This model is motivated by the context of randomised agricultural experiments where the plot error is considered as the sum of an independent random error and a 'soil' error (see for example Neymann, Iwaskiewicz and Kolodziejczyk [7]).

COROLLARY 1. *Suppose* $X_i = V_i + W_i, i = 1, \dots, N$ *where* $V_1, \dots, V_N$ *are independent random variables with* $E(V_i) = 0$, *and independent of the random variables* $W_1, \dots, W_N$. *Put* $E(W_i) = E(X_i) = \mu_i$, *and for* $j = 2, 3$, $E|X_i - \mu_i|^j = \mu_{j,i}$, $E|W_i - \mu_i|^j = \omega_{j,i}$ *and* $E|V_i|^j = \nu_{j,i}$. *Suppose there exists a positive constant* δ *such that*

$$\sum_{i=1}^{N} \nu_{2,i} \geq \delta N. \tag{7}$$

Then there exists a constant B, *depending on* δ, *such that*

$$\sup_v |P(T \leq v) - \Phi(v)| \leq B(pq)^{-1/2} N^{-3/2} \left(\sum_{i=1}^{N} \mu_{3,i} + \sum_{i=1}^{N} |\mu_i - \bar{\mu}|^3 \right).$$

PROOF. With $W = (W_1, \dots, W_N)$, we have by (7), $E(S^2|W) \geq \frac{1}{2}\delta$ and so, by Chebychev's inequality and Lemma A with $k = 3/2$ (see Appendix),

$$\begin{aligned} P\left(S^2 < \tfrac{1}{4}\delta\right) &\leq P\left(\left|S^2 - E\left(S^2|W\right)\right| > \tfrac{1}{4}\delta\right) \\ &\leq BE\left(E(|S^2 - E(S^2|W)|^{3/2}|W)\right) \\ &\leq BN^{-3/2}\left(\sum_{i=1}^{N} \nu_{3,i} + \sum_{i=1}^{N} E\left|W_i - \bar{W}\right|^3\right). \end{aligned}$$

The result follows from Theorem 1 by noting that since $X_i - \mu_i = W_i - \mu_i + V_i$ and W_i and V_i are independent, we have $\nu_{3,i} \leq \mu_{3,i}$ and $\omega_{3,i} \leq \mu_{3,i}$.

3. A result due to von Bahr

Let $X_1, \dots, X_N$ be independent random variables with $E(X_i) = \mu_i$, and for $1 < j \leq 4$, $E|X_i|^j = \mu'_{j,i}$. We adopt the same scale as von Bahr by insisting that

$$\sum_{i=1}^{N} \mu_i = 0$$

and

$$N^{-1} \sum_{i=1}^{N} \mu'_{2,i} = 1.$$

https://doi.org/10.1017/S1446788700037861 Published online by Cambridge University Press

Put

$$\mu^2 = N^{-1} \sum_{i=1}^{N} \mu_i^2.$$

Let S^2 be given by (6) and T^* by (5) with $a = 0$ and $b = (m(1 - p\mu^2))^{1/2}$, where $p = m/N$.

THEOREM 2. *There exists an absolute constant B such that*

$$\sup_v |P(T^* \leq v) - \Phi(v)| \leq B(Npq)^{-1/2} \max_i(\mu'_{4,i})(1 - \mu^2)^{-2}.$$

PROOF. Let Λ be the set where $S^2 > 1/4$. Then

$$\begin{aligned} |P(T^* \leq v) - \Phi(v)| &= |E(P(T^* \leq v|X)) - \Phi(v)| \\ &\leq |E_1| + |E_2| + |E_3| + |E_4| \end{aligned}$$

where

$$\begin{aligned} E_1 &= E\left(I_\Lambda\left\{P\left(\frac{\sum_{i=1}^m Y_i - m\bar{X}}{(m(1-p)S^2)^{1/2}} \leq \frac{v(1-p\mu^2)^{1/2} - m^{1/2}\bar{X}}{((1-p)S^2)^{1/2}}|X\right)\right.\right. \\ &\qquad \left.\left. -\Phi\left(\frac{v(1-p\mu^2)^{1/2} - m^{1/2}\bar{X}}{((1-p)S^2)^{1/2}}\right)\right\}\right), \\ E_2 &= E\left(I_\Lambda\left\{\Phi\left(\frac{v(1-p\mu^2)^{1/2} - m^{1/2}\bar{X}}{((1-p)S^2)^{1/2}}\right) - \Phi\left(\frac{v(1-p\mu^2)^{1/2} - m^{1/2}\bar{X}}{(1-p)^{1/2}}\right)\right\}\right), \\ E_3 &= E\left(I_\Lambda\left\{\Phi\left(\frac{v(1-p\mu^2)^{1/2} - m^{1/2}\bar{X}}{(1-p)^{1/2}}\right) - \Phi(v)\right\}\right), \end{aligned}$$

and

$$E_4 = E(I_{\Lambda^c}\{P(T^* \leq v|X) - \Phi(v)\}).$$

From (1), we have

$$\begin{aligned} |E_1| &\leq B(pq)^{-1/2}N^{-3/2}E\left(I_\Lambda \sum_{i=1}^{N} \left|X_i - \bar{X}\right|^3 S^{-3}\right) \\ &\leq B(pq)^{-1/2}N^{-3/2} \sum_{i=1}^{N} E\left|X_i - \bar{X}\right|^3 \\ &\leq B(pq)^{-1/2}N^{-3/2} \sum_{i=1}^{N} \mu'_{3,i}. \end{aligned} \tag{8}$$

Putting F as the distribution function of $m^{1/2}\bar{X}((1-\mu^2)p)^{-1/2}$, we have

$$\begin{aligned}
&\left|E\Phi\left(\frac{v(1-p\mu^2)^{1/2}-m^{1/2}\bar{X}}{(1-p)^{1/2}}\right)-\Phi(v)\right| \\
&\quad=\left|\int_{-\infty}^{\infty}\Phi\left(\frac{v(1-p\mu^2)^{1/2}-u}{(1-p)^{1/2}}\right)dF\left(\frac{u}{((1-\mu^2)p)^{1/2}}\right)-\Phi(v)\right| \\
&\quad=\left|\int_{-\infty}^{\infty}F\left(\frac{v(1-p\mu^2)^{1/2}-u}{((1-\mu^2)p)^{1/2}}\right)d\Phi\left(\frac{u}{(1-p)^{\frac{1}{2}}}\right)\right. \\
&\qquad\left.-\int_{-\infty}^{\infty}\Phi\left(\frac{(v-u)(1-p\mu^2)^{\frac{1}{2}}}{((1-\mu^2)p)^{\frac{1}{2}}}\right)d\Phi\left(\frac{u(1-p\mu^2)^{1/2}}{(1-p)^{\frac{1}{2}}}\right)\right| \\
&\quad\le\int_{-\infty}^{\infty}\left|F\left(\frac{v(1-p\mu^2)^{1/2}-u}{((1-\mu^2)p)^{1/2}}\right)-\Phi\left(\frac{v(1-p\mu^2)^{1/2}-u}{((1-\mu^2)p)^{1/2}}\right)\right|d\Phi\left(\frac{u}{(1-p)^{1/2}}\right) \\
&\quad\le B(1-\mu^2)^{-3/2}N^{-3/2}\sum_{i=1}^{N}\mu'_{3,i},
\end{aligned}$$

where the last inequality follows from Berry-Esseen rate results for independent, non-identically distributed random variables (Feller [5, p. 544]). Thus

$$(9)\qquad |E_3|\le B(1-\mu^2)^{-3/2}N^{-3/2}\sum_{i=1}^{N}\mu'_{3,i}+P(\Lambda^c).$$

Now, by Chebychev's inequality and Lemma A, with $k=3/2$, since $ES^2=(N-1+\mu^2)/N>1/2$, we have

$$(10)\qquad P(\Lambda^c)\le P(|S^2-ES^2|\ge 1/4)\le BN^{-3/2}\sum_{i=1}^{N}\mu'_{3,i}.$$

As $|E_4|$ is bounded by $P(\Lambda^c)$ we have only $|E_2|$ left to consider. So far, we have only needed to assume the existence of finite third moments for the X_i . In dealing with the term $|E_2|$ it appears that we need finite fourth moments to obtain a rate of $O(N^{-1/2})$. We have, by Taylor's theorem

$$\begin{aligned}
|E_2|&\le BE|S^2-1| \\
&\le B\left\{E\left|S^2-ES^2\right|+\frac{1-\mu^2}{N}\right\} \\
&\le B\left\{\left(E\left|S^2-ES^2\right|^2\right)^{1/2}+\frac{1-\mu^2}{N}\right\}.
\end{aligned}$$

Now, Lemma A with $k=2$ and Holder's inequality ensure

$$(11)\qquad |E_2|\le BN^{-3/2}\sum_{i=1}^{N}\mu'_{4,i}.$$

https://doi.org/10.1017/S1446788700037861 Published online by Cambridge University Press

The inequalities (8), (9), (10) and (11) essentially establish the result.

Appendix

LEMMA (A). *Let $X_1, \ldots, X_N$ be independent random variables $E(X_i) = \mu_i$, and for $1 < j \leq 2k$, $E|X_i - \mu_i|^j = \mu_{j,i} < \infty$. Put $\bar{\mu} = N^{-1}\sum_{i=1}^N \mu_i$. Then if $k \geq 1$ there exists an absolute constant B, depending on k, such that*

$$E\left|N^{-1}\sum_{i=1}^N (X_i - \bar{X})^2 - E\left(N^{-1}\sum_{i=1}^N (X_i - \bar{X})^2\right)\right|^k \leq BN^{-\lambda}\left\{\sum_{i=1}^N \mu_{2k,i} + \sum_{i=1}^N |\mu_i - \bar{\mu}|^{2k}\right\}$$

where $\lambda = \min(k, \frac{1}{2}k + 1)$.

PROOF. First we see that

(12)

$$\begin{aligned}
&N^{-k}E\left|\sum_{i=1}^N (X_i - \bar{X})^2 - E\sum_{i=1}^N (X_i - \bar{X})^2\right|^k \\
&\leq 2^{k-1}\left(N^{-k}E\left|\sum_{i=1}^N (X_i - \bar{\mu})^2 - E\sum_{i=1}^N (X_i - \bar{\mu})^2\right|^k + E\left|(\bar{X} - \bar{\mu})^2 - E(\bar{X} - \bar{\mu})^2\right|^k\right)
\end{aligned}$$

and

$$\begin{aligned}
&E\left|\sum_{i=1}^N \{(X_i - \bar{\mu})^2 - E(X_i - \bar{\mu})^2\}\right|^k \\
&\leq 2^{k-1}E\left|\sum_{i=1}^N \{(X_i - \mu_i)^2 - E(X_i - \mu_i)^2\}\right|^k + 2^{2k-1}E\left|\sum_{i=1}^N (\mu_i - \bar{\mu})(X_i - \mu_i)\right|^k.
\end{aligned}$$

Now, from the Marcinkiewicz-Zygmund-Chung inequality (Chung [3]) and the Hölder inequality, putting $\nu = \max(0, \frac{1}{2}k - 1)$,

$$\begin{aligned}
E\left|\sum_{i=1}^N (\mu_i - \bar{\mu})(X_i - \mu_i)\right|^k &\leq BN^{\nu}\left(\sum_{i=1}^N |\mu_i - \bar{\mu}|^{2k}\right)^{1/2}\left(\sum_{i=1}^N \mu_{2k,i}\right)^{1/2} \\
&\leq BN^{\nu}\left(\sum_{i=1}^N \mu_{2k,i} + \sum_{i=1}^N (\mu_i - \bar{\mu})^{2k}\right)
\end{aligned}$$

and

$$E\left|\sum_{i=1}^N \{(X_i - \mu_i)^2 - E(X_i - \mu_i)^2\}\right|^k < BN^{\nu}\sum_{i=1}^N \mu_{2k,i}.$$

A similar result holds for the last term of (12) and this establishes the lemma.

References

[1] B. von Bahr, 'On sampling from a finite set of independent random variables', *Z. Wahrsch. Verw. Gebiete.* **24** (1972), 279–286.

[2] H. Chernoff and H. Teicher, 'A central limit theorem for sums of interchangeable random variables', *Ann. Math. Statist.* **29** (1958), 118–130.

[3] K. L. Chung, 'The strong law of large numbers', in: *Proc. 2nd Berkeley Symp. on Math. Statist. and Prob.* (ed. J. Neyman) (University of California, Berkeley, 1951), 341–352.

[4] P. Erdös and A. Rényi, 'On the central limit theorem for samples from a finite population', *Magyar Tud. Akad. Mat. Kutató Intéz. Közl.* **4** (1959), 49–61.

[5] W. Feller, *An introduction to probability theory and its applications*, vol. II, 2nd edition (Wiley, New York, 1971).

[6] T. Höglund, 'Sampling from a finite population. A remainder term estimate', *Scand. J. Statist.* **5** (1978), 69–71.

[7] J. Neymann, K. Iwaszkiewicz and St. Kolodziejczyk, 'Statistical problems in agricultural experimentation', *Roy. Stat. Soc. Jour. Suppl.* **2** (1935), 107–180.

Statistical Consulting Group
Department of Mathematics
University of Western Australia
Nedlands, WA 6009
Australia

School of Mathematics and Statistics
University of Sydney
NSW 2006
Australia
e-mail: robinson_j@maths.su.oz.au