The Annals of Statistics
1989, Vol. 17, No. 1, 443–447

ON EDGEWORTH EXPANSIONS IN THE MIXTURE CASES

BY G. J. BABU AND K. SINGH

Pennsylvania State University and Rutgers University

Let X be a random vector with at least one marginal having a lattice distribution. For a wide class of statistics which can be written as a function of means of independent copies of X, it is established in this article that the one-term Edgeworth expansion is typically the same as the usual one-term expansion in the pure nonlattice case.

Introduction. Consider a bivariate mean $\bar{Z}_n' = (\bar{X}_n, \bar{Y}_n)$ of i.i.d. bivariate r.v.'s. If the underlying population is strongly nonlattice, i.e., $E(e^{it'\bar{Z}_n}) \neq 1$ for all $t' = (t_1, t_2) \neq 0$, then the Edgeworth expansions are known for $P_n(A) = P(\sqrt{n}\,(\bar{Z}_n - E(\bar{Z}_n)) \in A)$ for a fairly rich class of Borel measurable sets A in $\mathbb{R}^2$, where $'$ denotes transpose.

In the lattice case, where each univariate population assigns its entire mass to countably many equidistant points, the expansion is different and an explicit form is known only for a restricted class of sets A. Consider an univariate statistic of the form $H(\bar{Z}_n)$, for a smooth function H. Bhattacharya and Ghosh (1978) used the expansion for $\bar{Z}_n$ to establish the validity of the formal Edgeworth expansion for $\sqrt{n}\,[H(\bar{Z}_n) - H(E(\bar{Z}_n))]$ in the nonlattice case. In the lattice case, an explicit form of the expansion is typically not available for nonlinear univariate statistics. See Yarnold (1972) for instance, where an explicit expansion for $P_n(A)$ is given, when A is an ellipsoid. In this paper we consider Edgeworth expansions for $H(\bar{Z}_n)$, when X has a lattice distribution and Y has a continuous distribution. Situations like this do arise in statistical experiments. As a simple example, X could be the age of a plant or animal in years and Y could be its weight; $H(\bar{Z}_n) = \bar{Y}_n/\bar{X}_n$. In this article we establish that no correction factor is needed for the lattice character of X in the one-term Edgeworth expansion. Beyond one term, the answer is not known to us at present.

Our approach for establishing the above mentioned fact is as follows: Let η be a symmetric random variable having finite third moment and whose characteristic function vanishes outside a compact interval. It is first shown that in the lattice case $\{\sqrt{n}\,(\bar{X}_n - E(\bar{X}_n)) + \eta n^{-3/8}, \sqrt{n}\,(\bar{Y}_n - E(\bar{Y}_n))\}$ has the usual (nonlattice) one-term bivariate Edgeworth expansion. Now, under a smoothness condition on H, one can write $\sqrt{n}\,[H(\bar{Z}_n) - H(E(\bar{Z}_n))]$ as $\sqrt{n}\, l(\bar{Z}_n - E(\bar{Z}_n)) + \sqrt{n}\,(\bar{Z}_n - E(\bar{Z}_n))' L(\bar{Z}_n - E(\bar{Z}_n)) + o_p(n^{-1/2})$, where l is a 2×1 vector and L is a 2×2 matrix. The remainder $o_p(n^{-1/2})$, typically, does not affect the one-term Edgeworth expansion. (Here and in what follows T' denotes the transpose of T.) One can write the above linear term as $\sqrt{n}\,(\bar{W}_n - E(\bar{W}_n))$, where $W_i = l(X_i, Y_i)'$. Suppose $l = (l_1, l_2)$ and $l_2 \neq 0$. Since $Y_i = (W_i - l_i X_i)/l_2$, by simple algebra we

Received October 1986; revised April 1988.

AMS 1980 *subject classification.* 62E20.

Key words and phrases. Edgeworth expansions, characteristic functions, ratio estimators, lattice distributions.

can replace the linear and the quadratic terms above, respectively, by $\sqrt{n}(\overline{W}_n - E(\overline{W}_n))$ and $\overline{V}_n \Sigma \overline{V}_n'$, where

$$\overline{V}_n = \left(\overline{X}_n - E(\overline{X}_n), \overline{W}_n - E(\overline{W}_n)\right), \qquad \Sigma = NLN' \text{ and } N = \begin{pmatrix} 1 & -l_1/l_2 \\ 0 & 1/l_2 \end{pmatrix}.$$

One can now look at W as the continuous part and X as the lattice part in the bivariate expansion. Thus the smoothing through η appears only in the quadratic term. As a result the smoothing is of the order $O_p(n^{-7/8})$. Hence, it does not affect the one-term expansion.

The above ideas extend naturally to the general cases of k-dimensional means. Let $k = r + q$, where r of the marginal populations are continuous and the remaining q are each separately lattice. On the r continuous marginal populations, one needs to assume, further, that any nonzero linear combination has a continuous component. Then, using arguments similar to the ones used in the bivariate case, one can obtain Edgeworth expansions for statistics of the form $H(\overline{Z}_n)$.

We discuss two specific examples in the article: (i) The ratio estimator $\overline{Y}_n/\overline{X}_n$, where the Y population is continuous and the X population is lattice and (ii) the coefficient of correlation between a continuous random variable and a lattice one. The correlation coefficient example is reduced to the case of a 4-dimensional mean after writing $\sqrt{n}(\gamma - \rho)$ as a linear term + $(1/\sqrt{n})$ a quadratic term + $o_p(n^{-1/2})$.

Main results. Consider a bivariate random vector (X, Y), of which X is a nondegenerate lattice variable and Y is a continuous random variable. Let $\{(X_i, Y_i)\}$ be a sequence of independent copies of (X, Y). Let η be a real valued symmetric random variable with finite absolute third moment and whose characteristic function vanishes outside a compact interval. [See Theorem 10.1 of Bhattacharya and Ranga Rao (1986), for the existence of such random variables.] It is also assumed that η is independent of X_i and Y_i for all i.

We require some notation before stating the main theorem. Let for x in $\mathbb{R}^2$, $\|x\|$ denote the Euclidean norm of x. Let Σ denote the dispersion matrix of (X, Y) and let φ_Σ denote the density of normal distribution with mean zero and variance covariance matrix Σ. Note that, in view of the conditions on (X, Y), Σ is a positive definite matrix. Let $(\overline{X}_n, \overline{Y}_n)$ denote the sample mean of (X_i, Y_i), $i = 1, \dots, n$. For $\varepsilon > 0$ and any measurable function f, let

$$\omega(f, x, \varepsilon) = \sup\{|f(x) - f(y)| : \|x - y\| < \varepsilon\}$$

and let

$$\overline{\omega}(f, \Sigma, \varepsilon) = \int \omega(f, x, \varepsilon) \varphi_\Sigma(x)\, dx.$$

We have the following

THEOREM. *Let* $E\|(X, Y)\|^3 < \infty$, $E(X) = E(Y) = 0$ *and let* f *be a real-valued measurable function on* $\mathbb{R}^2$ *bounded by* 1. *Suppose* P_n *denotes the* *distribution of* $(\sqrt{n}\,\overline{X}_n + \eta n^{-3/8}, \sqrt{n}\,\overline{Y}_n)$ *and* Q_n *denotes the distribution with density* $(1 + pn^{-1/2})\varphi_{\Sigma}$, *where p is a third degree polynomial, whose coefficients are continuous functions of the moments of* (X, Y) *of the orders not more than* 3. [*Here p is the same polynomial which appears in the Edgeworth expansions of* $(\sqrt{n}\,\overline{X}_n, \sqrt{n}\,\overline{Y}_n)$, *when* (X, Y) *has strongly nonlattice distribution.*] *Then*

$$\left|\int f\,d(P_n - Q_n)\right| \le c\bar{\omega}(f, \Sigma, \delta_n) + o(n^{-1/2}),$$

where c is an absolute constant and $\delta_n = o(n^{-1/2})$, *not depending upon* f.

The proof of the theorem essentially involves the expansion of the derivatives $D^{\alpha}\psi_n(t, s)$ of the characteristic function ψ_n of the vector $(\sqrt{n}\,\overline{X}_n + \eta n^{-3/8}, \sqrt{n}\,\overline{Y}_n)$, where $\alpha = (\alpha_1, \alpha_2)$, $\alpha_1 + \alpha_2 \le 3$ and α_1 and α_2 are nonnegative integers. See the proofs of Theorems 20.8 and 24.2 of Bhattacharya and Ranga Rao (1986). The derivatives of the characteristic function φ_n of $(\sqrt{n}\,\overline{X}_n, \sqrt{n}\,\overline{Y}_n)$ can be obtained in the range $|t| \le \sqrt{n}/\log n$ and $|s| < \varepsilon\sqrt{n}$, for some fixed $\varepsilon > 0$, as in the classical case of strongly nonlattice distributions. As the characteristic function of η is 0 outside a compact interval, the derivatives of $\psi_n(t, s)$, of orders $\alpha = (\alpha_1, \alpha_2)$ with $\alpha_1 + \alpha_2 \le 3$, are 0 for $|t| > \sqrt{n}/\log n$. Further, they are exponentially decaying in the range $|t| \le \sqrt{n}/\log n$ and $\varepsilon < |s/\sqrt{n}\,| < M$ for all large n and for any $M > \varepsilon > 0$. The rest of the proof is similar to that of the classical case of strongly nonlattice distributions.

COROLLARY. *Let* $q' = (a, b)$ *be a nonzero vector and* L *be a* 2×2 *matrix. Then under the conditions of the theorem we have*

$$\begin{aligned}(1)\quad &\sup_{\hat{x}}\Big|P\big((a\overline{X}_n + b\overline{Y}_n)\sqrt{n} + a\eta n^{-3/8} + \sqrt{n}\,(\overline{X}_n, \overline{Y}_n)L(\overline{X}_n, \overline{Y}_n)' < xq'\Sigma q\big)\\ &\qquad\qquad - \int_{-\infty}^{x}(1 + n^{-1/2}r(y))\varphi(y)\,dy\Big| = o(n^{-1/2}),\end{aligned}$$

where r is a polynomial of degree less than or equal to 3 *and* φ *is the density of the standard normal variable.*

The corollary follows from lemma 3 of Babu and Singh (1984) which is a modification of a result of Bhattacharya and Ghosh (1978). Note that $P(|\eta| > n^{3/8}(\log n)^{-2}) = o(n^{-1/2})$ and by a moderate deviation result of Michel (1976), $P(\sqrt{n}\,|\overline{Y}_n| > \log n) < n^{-1}$. As a result $P(|\eta n^{-3/8}\sqrt{n}\,\overline{Y}_n| > (\log n)^{-1}) = o(n^{-1/2})$. So the effect of smoothing by $\eta n^{-3/8}$ from the quadratic term is of the order $o(n^{-1/2})$ and so is negligible.

REMARK. In our applications $a = 0$, so the term containing η in (1) does not appear. See the examples below. The theorem can easily be extended to higher dimensions.

Examples.

1. *The ratio estimator.* The ratio estimator $T_n = \sqrt{n}\,[(\bar{Z}_n/\bar{X}_n)E(X) - E(Z)]$ is often encountered in sample surveys. Quite frequently, the auxiliary variable X is lattice and the variable Z under study is continuous. The statistic T_n can be written as

$$\sqrt{n}\left(\bar{Z}_n - \bar{X}_n(E(Z)/E(X))\right)\left(1 - \left((\bar{X}_n - E(X))/E(X)\right)\right) + \text{higher order terms.}$$

These higher order terms contribute an error of the order $o(n^{-1/2})$ in (1), and hence can be ignored.

Note that if W has a continuous distribution and V has a discrete distribution then $W + V$ has a continuous distribution. This follows from noting that

$$P(W + V = x) = \sum_{a \in B} P(W = x - a, V = a) \leq \sum_{a \in B} P(W = x - a) = 0,$$

where $B = \{a\colon P(V = a) > 0\}$, which is countable.

An Edgeworth expansion for T_n can now be obtained by applying the corollary to $\{(X_i, Y_i)\}$ with $Y_i = Z_i - X_i(E(Z)/E(X))$ and using the fact that Y_i has a continuous distribution.

2. *Sample correlation coefficient.* Let (W, V) be a random vector, where W is a lattice random variable. Further assume the existence of the sixth moment for $\|(W, V)\|$. Let ρ denote the correlation coefficient of (W, V). Without loss of generality we shall assume that $E(W) = E(V) = 0$ and $E(W^2) = E(V^2) = 1$. Consider the sample correlation coefficient R_n based on n observations (W_i, V_i), $i = 1, \ldots, n$ on (W, V). Let $\bar{G}$ denote the sample mean of the random variables $G_1, \ldots, G_n$. Now

$$R_n - \rho = \left(\overline{WV} - \overline{W}\overline{V}\right)\left[\left(\overline{W^2} - (\overline{W})^2\right)\left(\overline{V^2} - (\overline{V})^2\right)\right]^{-1/2} - \rho = T_n + B_n,$$

where

$$T_n = (1 - \rho)\left(\overline{WV - \rho}\right) - (\rho/2)\left(\overline{(W - V)^2 - 2(1 - \rho)}\right) - \overline{W}\overline{V} - \left[\overline{WV - \rho}\right]^2$$
$$-\left(\left(\overline{WV - \rho}\right)/2\right)\left(\overline{(W - V)^2 - 2(1 - \rho)}\right)$$

and $P(|B_n| > \theta n^{-1/2}) = o(n^{-1/2})$ for any $\theta > 0$. Note that $E(W - V)^2 = 2(1 - \rho)$.

Clearly T_n is a second degree polynomial in the sample mean of n observations from

$$S = \left(W, V, WV - \rho, (V - W)^2 - 2(1 - \rho)\right).$$

Since W is assumed to be lattice, it is enough to check that $Z = tV + sWV + r(V - W)^2$ has a continuous component, whenever (t, s, r) is a nonzero vector. This can be seen from the following argument. If $r \neq 0$, then for some

$b, a_i, a_2, \ldots, d_1, d_2, \ldots,$ we have

$$\begin{aligned} P((Z/r) = a) &= \sum_i P\big((V - a_i)^2 = d_i, W = b + hi\big) \\ &\leq \sum_i P\big((V - a_i)^2 = d_i\big) = 0. \end{aligned}$$

So Z is continuous if $r \neq 0$. If $r = 0$ and $|t| + |s| \neq 0$, then $Z = (t + sW)V$ has a continuous component, as $(t + sW)$ is lattice or degenerate at t and V has a continuous distribution. Using the theorem we get an Edgeworth expansion for $T_n - \eta \bar{V} n^{-7/8}$, where η is as in the Introduction. From this an Edgeworth expansion for R_n can easily be obtained.

REFERENCES

Babu, G. J. and Singh, K. (1984). On one term Edgeworth correction by Efron's bootstrap. *Sankhyā Ser. A* **46** 219–232.

Bhattacharya, R. N. and Ghosh, J. K. (1978). On the validity of the formal Edgeworth expansion. *Ann. Statist.* **6** 434–451.

Bhattacharya, R. N. and Ranga Rao, R. (1986). *Normal Approximation and Asymptotic Expansions*, 2nd ed. Wiley, New York.

Michel, R. (1976). Nonuniform central limit bounds with applications to probabilities of deviations. *Ann. Probab.* 4 102–106.

Yarnold, J. K. (1972). Asymptotic approximations for the probability that a sum of lattice random vectors lies in a convex set. *Ann. Math. Statist.* **43** 1566–1580.

Department of Statistics
219 Pond Laboratory
Pennsylvania State University
University Park, Pennsylvania 16802

Department of Statistics
Hill Center, Busch Campus
Rutgers University
New Brunswick, New Jersey 08903