The Annals of Statistics
1990, Vol. 18, No. 3, 1483–1490

BERRY–ESSEEN-TYPE BOUNDS FOR SIGNED LINEAR RANK STATISTICS WITH A BROAD RANGE OF SCORES[1]

BY MUNSUP SEOH

Wright State University

The Berry–Esseen-type bounds of order $N^{-1/2}$ for the rate of convergence to normality are derived for the signed linear rank statistics under the hypothesis of symmetry. The results are obtained with a broad range of regression constants and scores (allowed to be generated by discontinuous score generating functions, but not necessarily) restricted by only mild conditions, while almost all previous results are obtained with continuously differentiable score generating functions. Furthermore, the proof is very short and elementary, based on the conditioning argument.

1. Introduction. Let X_{Nj}, $1 \le j \le N$, be independent and identically distributed (iid) random variables (rv's) with a continuous cumulative distribution function (cdf) $F_N(x)$, symmetric about zero. We consider the signed linear rank statistic

$$T_N^+ = \sum_{j=1}^{N} c_{Nj} a_N(R_{Nj}^+) \operatorname{sgn} X_{Nj}, \tag{1.1}$$

where R_{Nj}^+ is the rank of $|X_{Nj}|$ among $\{|X_{Nk}|: 1 \le k \le N\}$, c_{Nj}, $1 \le j \le N$, are known regression constants, $a_N(j)$ (interchangeably, a_{Nj}), $1 \le j \le N$, are known constants called scores and $\operatorname{sgn} x = 1$ or -1 according as $x \ge 0$ or $x < 0$. Note that T_N^+ reduces to the well-known one-sample Wilcoxon signed rank statistic when $c_{Nj} = N^{-1/2}$ and $a_{Nj} = j$, for $j = 1, 2, \dots, N$.

Let E and $U_{N:j}$ denote, respectively, the expectation and the jth order statistic among a random sample of size N from the uniform distribution over the unit interval $(0, 1)$. Then, scores are usually generated by some known function (called the score generating function) $J(t)$, $0 < t < 1$, in one of the following three ways:

$$a_{Nj} = E(J(U_{N:j})), \qquad 1 \le j \le N \text{ (exact scores)}, \tag{1.2}$$

$$a_{Nj} = J(E(U_{N:j})), \qquad 1 \le j \le N \text{ (approximate scores)}, \tag{1.3}$$

$$a_{Nj} = \int_{(j-1)/N}^{j/N} J(t)\, dt, \qquad 1 \le j \le N \text{ (approximate scores)}. \tag{1.4}$$

The asymptotic normality of T_N^+ is well known by previous works [see, e.g., Hájek (1962), Hájek and Šidák (1967) and Hušková (1970)]. In fact, under

Received January 1989; revised September 1989.

[1]Research partially completed when the author was an associate professor at POSTECH, Pohang, Korea.

AMS 1980 *subject classifications*. Primary 62E20; secondary 60F05, 62G10.

Key words and phrases. Berry–Esseen bound, rate of convergence, signed linear rank statistic, discontinuous score generating function.

suitable assumptions, $\Lambda_N \equiv \sup_x |G_N(x) - \Phi(x)| \to 0$ as $N \to \infty$, where $G_N(x)$ is the cdf of T_N^+ (suitably normalized) and $\Phi(x)$ is the standard normal cdf. However, one often needs more precise information than the asymptotic normality can provide and may try to find a suitable rate for Λ_N. To this end, Puri and Wu (1986) have obtained the order of $N^{-1/2+\delta}$, $\delta > 0$, with bounded score generating functions and Puri and Seoh (1984) obtained the order of $N^{-1/2}$ with unbounded score generating functions [which can be the Chi-quantile function $J(t) = \Phi^{-1}((t+1)/2)$], adapting the ideas of van Zwet (1982) and Does (1982).

Consider now the unsigned linear rank statistic (the counterpart of the statistic T_N^+) $T_N = \sum_{j=1}^N c_{Nj} a_N(R_{Nj})$, where R_{Nj}, $1 \le j \le N$, is the rank of Y_{Nj} among an independent sample $Y_{N1}, Y_{N2}, \dots, Y_{NN}$ with continuous cdf's. Under suitable assumptions, Jurečková and Puri (1975), Bergström and Puri (1977) and Hušková (1977, 1979) derived bounds of

$$O(N^{-1/2+\delta}) \quad \text{for } 0 < \delta < \tfrac{1}{2}, \qquad O(N^{-1/2} \log N) \quad \text{and} \quad O(N^{-1/2}),$$

respectively. In all these papers the score generating function J was assumed to be bounded and continuous. Later, the restrictive boundedness was dropped by Does (1982) for the case of iid rv's.

We also note that Puri and Seoh (1985), dealing with the so-called generalized rank statistic (which includes the statistics T_N^+, T_N, and some others as special cases), derived the order of $N^{-1/2}$ with bounded score generating functions assuming the underlying distributions are only independent.

However, all results mentioned are obtained by taking scores generated by a score generating function J which has bounded first derivatives or continuous second derivatives.

In this paper, we now derive the Berry–Esseen bound of order $N^{-1/2}$ for the statistic T_N^+ given by (1.1) with a broad range of scores with mild regularity conditions [see Assumption B and (2.7)]. When scores generated by a known function J are taken, the results obtained extend most of the previous results for the statistic (1.1), covering some discontinuous score generating function tending to infinity in the neighborhoods of 0 and 1 at the rate of $\{t(1-t)\}^{-1/4+\varepsilon}$, $\varepsilon > 0$. We also note that our method is based on an elementary conditioning argument, while the method of all previous works claiming the optimal bound for rank statistics is to approximate the characteristic function of the statistics (suitably normalized) by that of the standard normal rv, and then to invoke Esseen's smoothing lemma. Naturally this requires laborious computations.

To conclude this section, we note that von Bahr (1976) has also derived the order of $N^{-1/2}$ dealing with the so-called rank combinatorial statistic which includes T_N^+ and T_N as special cases. His result, when applied to linear rank statistics T_N^+ and T_N, ensures the optimal bound when scores are uniformly bounded, while ours allows them to be unbounded.

2. Assumptions and main result. Let $R_N^+ = (R_{N1}^+, R_{N2}^+, \dots, R_{NN}^+)$ be the vector of ranks and $D_N = (D_{N1}, D_{N2}, \dots, D_{NN})$ be the vector of antiranks defined by the inverse permutation of R_N^+, i.e., $D_N = (R_N^+)^{-1}$. We also denote $\operatorname{sgn} X_{D_N} = (\operatorname{sgn} X_{D_{N1}}, \operatorname{sgn} X_{D_{N2}}, \dots, \operatorname{sgn} X_{D_{NN}})$. Then the statistic T_N^+ is equivalently expressible [in its dual form to (1.1)] as

$$T_N^+ = \sum_{j=1}^{N} c_{D_{Nj}} a_{Nj} \operatorname{sgn} X_{D_{Nj}}. \tag{2.1}$$

Under the hypothesis of symmetry, it is well known that R_N^+ as well as D_N is uniformly distributed over the set of all permutations of $(1, 2, \dots, N)$. Also note that $\operatorname{sgn} X_{D_{Nj}}$, $1 \le j \le N$, are iid rv's with the common symmetric Bernoulli distribution and that the vector D_N is stochastically independent of the vector $\operatorname{sgn} X_{D_N}$ [see Theorems 19A and 19C in Hájek (1969)]. Hence, we have

$$ET_N^+ = 0 \quad \text{and} \quad \operatorname{Var} T_N^+ \equiv \tau_N^2 = \frac{1}{N} \sum_{j=1}^{N} c_{Nj}^2 \sum_{j=1}^{N} a_{Nj}^2. \tag{2.2}$$

Throughout this paper, we make the following three assumptions, with an absolute constant $\alpha > 1$:

Assumption A. The regression constants satisfy

$$\sum_{j=1}^{N} c_{Nj}^2 = 1, \qquad \max_{1 \le j \le N} c_{Nj}^2 \le \alpha N^{-1}. \tag{2.3}$$

Assumption B. The scores satisfy

$$\frac{1}{N} \sum_{j=1}^{N} a_{Nj}^2 > \alpha^{-1}, \quad \text{for all sufficiently large } N, \tag{2.4}$$

and

$$\max_{1 \le j \le N} a_{Nj}^2 = O(N^{1-2\delta}), \qquad 0 < \delta \le \tfrac{1}{2}. \tag{2.5}$$

Assumption C. The scores satisfy

$$\sum_{j=1}^{N} |a_{Nj}|^3 = O(N). \tag{2.6}$$

We now state our main theorems.

Theorem 2.1. *Under Assumptions* A, B *and* C, *we have*

$$\left\| P(T_N^+ \tau_N^{-1} \le \cdot) - \Phi(\cdot) \right\| = O(N^{-(1/2+\delta)/2}),$$

where $\|\cdot\|$ *denotes the usual supremum norm*, Φ *is the standard normal cdf and* δ *is given in* (2.5).

THEOREM 2.2. *In addition to Assumptions* A *and* B, *suppose*

$$\sum_{j=1}^{N} a_{Nj}^4 = O(N). \tag{2.7}$$

Then we have that $\| P(T_N^+ \tau_N^{-1} \le \cdot) - \Phi(\cdot) \| = O(N^{-1/2})$.

To consider the statistic T_N^+ with scores generated by a known function J, we consider the following assumption.

ASSUMPTION D. The set $\{t\colon J(t) \ne 0\}$ has a positive Lebesgue measure and J can be expressed as a finite linear combination of monotone functions $J_1, J_2, \dots, J_m$, with

$$|J_j(t)| = O\big(\{t(1-t)\}^{-1/4+\varepsilon}\big), \qquad 0 < \varepsilon \le \tfrac{1}{4},\ j = 1, 2, \dots, m. \tag{2.8}$$

Let the scores a_{Nj}, $1 \le j \le N$, be generated in either an exact or an approximate way [of (1.2)–(1.4)] with a known function J satisfying Assumption D. Then, it follows by Theorem V.1.4a and Lemma V.1.6a in Hájek and Šidák (1967) that $\max_{1 \le j \le N} a_{Nj}^2 = O(N^{1/2-2\varepsilon})$, and that

$$\begin{aligned} &\lim_{N\to\infty} N^{-1} \sum_{j=1}^{N} a_{Nj}^2 = \int_0^1 J^2(t)\,dt < \infty, \\ &\lim_{N\to\infty} N^{-1} \sum_{j=1}^{N} a_{Nj}^4 = \int_0^1 J^4(t)\,dt < \infty. \end{aligned} \tag{2.9}$$

Hence, the assumptions of Theorem 2.2 are satisfied, proving the following corollary.

COROLLARY 2.3. *In addition to Assumption* A, *assume that the scores are generated by a known function J satisfying Assumption* D *in either the exact way of* (1.2) *or the approximate ways of* (1.3) *and* (1.4). *Then we have*

$$\big\| P\big(T_N^+ \tau_N^{-1} \le \cdot\big) - \Phi(\cdot) \big\| = O(N^{-1/2}).$$

Proving Theorems 2.1 and 2.2 in the next section, we conclude this section with the following remarks.

REMARK 2.1. Condition (2.4) prohibits scores from taking too many values near zero. It is trivially satisfied by (2.9), if the scores are generated by (1.2)–(1.4) by a nonzero monotonic function J.

REMARK 2.2. Note that the Chi-quantile function $J(t) = \Phi^{-1}((1+t)/2)$ satisfies condition (2.8) [see Does (1982) and Puri and Seoh (1984)]. Hence, our theorems apply to the exact normal-scores statistic as well as van der Waerden's approximate normal-scores statistic for testing the hypothesis of symmetry.

REMARK 2.3. Theorem 2.1 ensures the optimal Berry–Esseen bound under the milder condition (2.6) than (2.7), when the scores are uniformly bounded ($\delta = \frac{1}{2}$). Condition (2.6) is considered the optimal one to ensure the bound, in the sense that the same bound is obtained under a similar condition (finite third moments of the summands) in the classical case of sums of independent rv's. It is hoped that the optimal bound is obtained under condition (2.6) instead of (2.7). However, with unbounded scores, the latter is the mildest condition yet used under which the optimal bound is obtained even with scores generated by a continuously differentiable function [see Does (1982) and Puri and Seoh (1984)].

REMARK 2.4. The Berry–Esseen bound can be obtained under the milder conditions on the regression constants, i.e.,

$$\sum_{j=1}^{N} c_{Nj}^2 = 1, \qquad \sum_{j=1}^{N} |c_{Nj}|^3 = O(N^{-1/2}),$$

used by several authors: see, for example, Does (1982), Hušková (1977, 1979) and Puri and Seoh (1984). However, scores used by these authors are generated by continuously differentiable functions, while ours are somewhat arbitrary, restricted by only mild conditions of (2.4)–(2.7). But our assumptions on the regression constants are stronger, as might be expected because the restriction on them is counterbalanced by those on scores.

REMARK 2.5. Corollary 2.3 is Theorem 2.2 in Puri and Seoh (1984). They obtained it, under the existence of continuous second derivative of the score generating function, by approximating laboriously the characteristic function of the statistic (1.1) to that of a normal one, and then invoking Esseen's smoothing lemma.

3. Proofs of the main theorems. From now on, we drop the subscript N in c_{Nj}, a_{Nj}, etc., whenever this causes no confusion.

LEMMA 3.1. *Under Assumption* A, *we have*

$$E\left(\sum_{j=1}^{N} c_{D_j}^2 a_j^2 - \frac{1}{N}\sum_{j=1}^{N} a_j^2\right)^2 = \frac{1}{N-1}\sum_{j=1}^{N}\left(c_j^2 - \frac{1}{N}\right)^2 \sum_{j=1}^{N}\left(a_j^2 - \overline{a^2}\right)^2,$$

where $\overline{a^2} = N^{-1}\sum_{j=1}^{N} a_j^2$.

PROOF. This is easily verified [or see Theorem II.3.1.c in Hájek and Šidák (1967), page 61]. □

Note that the vector of antiranks $D_N = (D_1, D_2, \dots, D_N)$ is uniformly distributed over Δ_N, the set of all permutations of $(1, 2, \dots, N)$. We now define a subset Ω_N of Δ_N by

$$\Omega_N = \left\{ d \in \Delta_N \colon \sum_{j=1}^{N} c_{d_j}^2 a_j^2 \geq (2\alpha)^{-1} \right\}, \tag{3.1}$$

where α is the absolute constant given in (2.4) and $d = (d_1, d_2, \dots, d_N)$. Then we have the following lemma.

LEMMA 3.2. *Under Assumptions* A, B *and* C, *we have*

$$P(D_N \in \Omega_N) \geq 1 - KN^{-1/2},$$

where Ω_N *is given by* (3.1) *and* K *is an absolute constant.*

PROOF. Let $\tilde{\Omega}_N = \{d \in \Delta_N \colon |\sum_{j=1}^N c_{d_j}^2 a_j^2 - (1/N)\sum_{j=1}^N a_j^2| \leq (2\alpha)^{-1}\}$. Then, by Chebyshev's inequality and Lemma 3.1, we have

$$\begin{aligned} 1 - P\big(D_N \in \tilde{\Omega}_N\big) &= P\left(\left| \sum_{j=1}^{N} c_{D_j}^2 a_j^2 - \frac{1}{N} \sum_{j=1}^{N} a_j^2 \right| \geq (2\alpha)^{-1} \right) \\ &\leq O\left(E\left(\sum_{j=1}^{N} c_{D_j}^2 a_j^2 - \frac{1}{N} \sum_{j=1}^{N} a_j^2 \right)^2 \right) = O\left(N^{-2} \sum_{j=1}^{N} a_j^4 \right) \\ &= O\left(N^{-2} \max_{1 \leq j \leq N} |a_{N_j}| \sum_{j=1}^{N} |a_j|^3 \right) = O(N^{-1/2-\delta}), \end{aligned}$$

which ensures that $P(D_N \in \tilde{\Omega}_N) \geq 1 - KN^{-1/2}$. Since $\tilde{\Omega}_N \subset \Omega_N$, the proof of Lemma 3.2 is complete. $\square$

We now prove the theorems.

PROOF OF THEOREM 2.1. Using conditional probability, we derive

$$\begin{aligned} &|P(T_N^+ \leq x\tau_N) - \Phi(x)| \\ &\quad = |E\{P(T_N^+ \leq x\tau_N | D_N) - \Phi(x)\}| \\ &\quad \leq \sum_{d \in \Omega_N} |P(T_N^+ \leq x\tau_N | D_N = d) - \Phi(x)| P(D = d) + O(N^{-1/2}), \end{aligned} \tag{3.2}$$

where Ω_N is given by (3.1) and the last inequality follows by Lemma 3.2.

For any permutation $d = (d_1, d_2, \dots, d_N)$ in Δ_N, we denote

$$S_N(d) = \sum_{j=1}^{N} c_{d_j} a_j \operatorname{sgn} X_{d_j}. \tag{3.3}$$

Then, since two vectors sgn X_D and D_N are independent under the hypothesis of symmetry, the distribution of T_N^+ conditionally given $D_N = d \in \Delta_N$, is that of $S_N(d)$, i.e., $P(T_N^+ \leq x\tau_N | D_N = d) = P(S_N(d) \leq x\tau_N)$.

Note that $S_N(d)$ is a sum of independent rv's with

$$ES_N(d) = 0, \qquad \operatorname{Var} S_N(d) \equiv \sigma_N^2 = \sum_{j=1}^{N} c_{d_j}^2 a_j^2. \tag{3.4}$$

By (3.1), $\sigma_N^2 > (2\alpha)^{-1} > 0$ for any $d \in \Omega_N$. Hence, we may apply Theorem V.2.3 in Petrov (1975) to obtain that, for any $d \in \Omega_N$,

$$\begin{aligned}
&|P(T_N^+ \le x\tau_N | D_N = d) - \Phi(x)| \\
&\quad = |P(S_N(d) \le x\tau_N) - \Phi(x)| \\
&\quad \le |P(S_N(d)\sigma_N^{-1} \le x\tau_N\sigma_N^{-1}) - \Phi(x\tau_N\sigma_N^{-1})| \\
&\qquad + |\Phi(x\tau_N\sigma_N^{-1}) - \Phi(x)| \\
&\quad = O\left(\sigma_N^{-3/2} \sum_{j=1}^{N} E|c_{d_j} a_j \operatorname{sgn} X_{d_j}|^3\right) \\
&\qquad + |\Phi(x\tau_N\sigma_N^{-1}) - \Phi(x)| \\
&\quad = O\left(N^{-3/2} \sum_{j=1}^{N} |a_j|^3\right) + |\Phi(x\tau_N\sigma_N^{-1}) - \Phi(x)| \\
&\quad = O(N^{-1/2}) + |\Phi(x\tau_N\sigma_N^{-1}) - \Phi(x)|,
\end{aligned} \tag{3.5}$$

in view of Assumptions A, B and C.

We now estimate the last term in (3.5). Since both x and $x\tau_N\sigma_N^{-1}$ are either nonpositive or nonnegative, it follows by Taylor's expansion that, uniformly on Ω_N,

$$\begin{aligned}
|\Phi(x) - \Phi(x\tau_N\sigma_N^{-1})| &\le |(1 - \tau_N\sigma_N^{-1})x\{\psi(x) + \psi(x\tau_N\sigma_N^{-1})\}| \\
&= O(|1 - \tau_N\sigma_N^{-1}| + |1 - \tau_N^{-1}\sigma_N|),
\end{aligned} \tag{3.6}$$

where $\psi(\cdot)$ denotes the standard normal density function. Because $|1 - x| \le |1 - x^2|$ for any $x \ge 0$, we have, uniformly on Ω_N,

$$\left|1 - \frac{\tau_N}{\sigma_N}\right| \le \left|1 - \frac{1}{\sigma_N^2}\left(\frac{1}{N}\sum_{j=1}^{N} a_j^2\right)\right| = O\left(\left|\sum_{j=1}^{N} c_{d_j}^2 a_j^2 - \frac{1}{N}\sum_{j=1}^{N} a_j^2\right|\right),$$

$$\left|1 - \frac{\sigma_N}{\tau_N}\right| \le \left|1 - \sigma_N^2\left(\frac{1}{N}\sum_{j=1}^{N} a_j^2\right)^{-1}\right| = O\left(\left|\sum_{j=1}^{N} c_{d_j}^2 a_j^2 - \frac{1}{N}\sum_{j=1}^{N} a_j^2\right|\right),$$

which, together with (3.5) and (3.6), ensure

$$\begin{aligned}
&|P(T_N^+ \le x\tau_N | D_N = d) - \Phi(x)| \\
&\quad = O(N^{-1/2}) + O\left(\left|\sum_{j=1}^{N} c_{d_j}^2 a_j^2 - \frac{1}{N}\sum_{j=1}^{N} a_j^2\right|\right),
\end{aligned} \tag{3.7}$$

uniformly on Ω_N.

It now follows by (3.2), (3.7) and Lemma 3.1 that

$$
\begin{aligned}
&|P(T_N^+ \le x\tau_N) - \Phi(x)| \\
&\quad = O(N^{-1/2}) + O\left(\left(E\left\{\sum_{j=1}^{N} c_{D_j}^2 a_j^2 - \frac{1}{N}\sum_{j=1}^{N} a_j^2\right\}^2\right)^{1/2}\right) \qquad (3.8)\\
&\quad = O(N^{-1/2}) + O\left(N^{-1}\left\{\sum_{j=1}^{N}\left(a_j^2 - \overline{a^2}\right)^2\right\}^{1/2}\right)
\end{aligned}
$$

Finally, the last term is of order $N^{-(1/2+\delta)/2}$, in view of (2.5) and (2.6). The proof is complete. □

Proof of Theorem 2.2. The proof is exactly same as the Proof of Theorem 2.1 except for the estimation of the last term in (3.8): it is of order $N^{-1/2}$ under condition (2.7). The proof is complete. □

REFERENCES

Bergström, H. and Puri, M. L. (1977). Convergence and remainder terms in liner rank statistic. *Ann. Statist.* **5** 671–680.

Does, R. J. M. M. (1982). Berry–Esseen theorems for simple linear rank statistics under the null-hypothesis. *Ann. Probab.* **10** 982–991.

Hájek, J. (1962). Asymptotically most powerful rank order tests. *Ann. Math. Statist.* **33** 1124–1147.

Hájek, J. (1969). *Nonparametric Statistics*. Holden-Day, San Francisco.

Hájek, J. and Šidák, Z. (1967). *Theory of Rank Tests*. Academic, New York.

Hušková, M. (1970). Asymptotic distribution of simple linear rank statistics for testing symmetry. *Z. Wahrsch. Verw. Gebiete* **14** 308–322.

Hušková, M. (1977). The rate of convergence of simple liner rank statistics under hypothesis and alternatives. *Ann. Statist.* **5** 658–670.

Hušková, M. (1979). The Berry–Esseen theorem for rank statistics. *Comment. Math. Univ. Carolin.* **20** 399–415.

Jurečková, J. and Puri, M. L. (1975). Order of normal approximation for rank test statistics distribution. *Ann. Probab.* **3** 526–533.

Petrov, V. V. (1975). *Sums of Independent Random Variables*. Springer, Berlin.

Puri, M. L. and Seoh, M. (1984). Berry–Esseen theorems for signed linear rank statistics with regression constants. In *Colloquium on Limit Theorems in Probability and Statistics, Veszprém, Hungary* (P. Révész, ed.) 875–906. North-Holland, Amsterdam.

Puri, M. L. and Seoh, M. (1985). On the rate of convergence to normality for generalized linear rank statistics. *Ann. Inst. Statist. Math.* **37** (Part A) 51–69.

Puri, M. L. and Wu, T. J. (1986). The order of normal approximation for signed linear rank statistics. *Teor. Verojatnost. i Primen.* **31** 145–151.

von Bahr, B. (1976). Remainder term estimate in a combinatorial limit theorem. *Z. Wahrsch. Verw. Gebiete* **35** 131–139.

van Zwet, W. R. (1982). On the Edgeworth expansion for the simple linear rank statistic. In *Nonparametric Statistical Inference* (B. V. Gnedenko, M. L. Puri and I. Vincze, eds.) **2** 889–909. North-Holland, Amsterdam.

Department of Mathematics and Statistics
Wright State University
Dayton, Ohio 45435