stichting
mathematisch
centrum

MC

AFDELING MATHEMATISCHE STATISTIEK SW 70/80 NOVEMBER
(DEPARTMENT OF MATHEMATICAL STATISTICS)

W.R. VAN ZWET

ON THE EDGEWORTH EXPANSION FOR THE SIMPLE LINEAR RANK STATISTIC

Preprint

kruislaan 413 1098 SJ amsterdam

Printed at the Mathematical Centre, 413 Kruislaan, 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.).

1980 Mathematics subject classification: 62E20, 62F05

On the Edgeworth expansion for the simple linear rank statistic [*)]

by

W.R. van Zwet [**)]

ABSTRACT

In this paper we consider the behavior of the characteristic function of a properly standardized simple linear rank statistic for large values of the argument. Under mild assumptions on the statistic and on the underlying probability distributions we obtain an upper bound for this characteristic function. This result makes it possible to obtain an Edgeworth expansion for the simple linear rank statistic.

KEY WORDS & PHRASES: *simple linear rank statistic, Edgeworth expansion, characteristic function*

[*)] This report will be submitted for publication elsewhere.

[**)] University of Leiden

1. INTRODUCTION

Let $X_1, X_2, \ldots, X_N$ be independent random variables with probability density functions $f_1, f_2, \ldots, f_N$ respectively. If $Z_1 < Z_2 < \ldots < Z_N$ denotes the sequence $X_1, \ldots, X_N$ arranged in increasing order, then the rank R_j of X_j is defined by $X_j = Z_{R_j}$, $j = 1, \ldots, N$. For sequences of real numbers $c_1, \ldots, c_N$ (regression constants) and $a_1, \ldots, a_N$ (scores),

$$(1.1) \qquad T_N = \sum_{j=1}^{N} c_j \, a_{R_j}$$

is called a simple linear rank statistic. It may be used for testing the hypothesis $H : f_1 = f_2 = \ldots = f_N$ against certain classes of alternatives indicated by the choice of scores and regression constants. This test is obviously distributionfree and under H the random vector $(R_1, \ldots, R_N)$ equals each permutation of $1, \ldots, N$ with probability $1/N!$. Well-known special cases are the two-sample statistics which have $c_1 = \ldots = c_m = 0$, $c_{m+1} = \ldots = c_N = 1$ and are used to test H against alternatives of the form $f_1 = \ldots = f_m$, $f_{m+1} = \ldots = f_N$.

Define

$$(1.2) \qquad \mu_N = E\, T_N , \qquad \sigma_N^2 = \sigma^2(T_N) ,$$

$$(1.3) \qquad T_N^* = \frac{T_N - \mu_N}{\sigma_N} ,$$

$$(1.4) \qquad F_N(x) = P(T_N^* \leq x) .$$

Under certain conditions (cf. Hājek and Šidāk (1967), chapter VI) T_N^* is asymptotically normal as $N \to \infty$, i.e.

$$\lim_{N\to\infty} \sup_x |F_N(x) - \Phi(x)| = 0 ,$$

where Φ is the standard normal distribution function. This result justifies the use of the normal approximation to compute the critical value and the power of a simple linear rank test for large N . On a more theoretical plane it enables us to find the limiting power of the test against contiguous alternatives and make comparisons with other tests on that basis.

Quite often, however, one needs more precise information than asymptotic normality can provide. On the one hand one may need more accurate numerical approximations and on the other one may wish to compare the performance of a simple linear rank test with that of other tests with the same limiting power. To achieve this one needs an asymptotic expansion for F_N with a uniform remainder $o(N^{-1})$ (cf. Hodges and Lehmann (1970)). Such expansions are usually of a type called Edgeworth expansions, i.e. they are of the form

$$(1.5) \qquad \tilde{F}_N(x) = \Phi(x) + \phi(x)\{N^{-\frac{1}{2}}Q_1(x) + N^{-1}Q_2(x)\}$$

where ϕ is the standard normal density and Q_1 and Q_2 are polynomials. To establish such an expansion for F_N one has to compute $\tilde{F}_N$ and show that as $N \to \infty$,

$$(1.6) \qquad \sup_x |F_N(x) - \tilde{F}_N(x)| = o(N^{-1}) .$$

A standard approach to this problem which has been successful for many statistics other than T_N^* , is as follows. Let

$$(1.7) \qquad \psi_N(t) = E\, e^{itT_N^*} = \int e^{itx}\, dF_N(x)$$

be the characteristic function of T_N^* . First of all one has to obtain an expansion for ψ_N of the form

$$(1.8) \qquad \tilde{\psi}_N(t) = e^{-\frac{1}{2}t^2}\{1 + N^{-\frac{1}{2}}R_1(t) + N^{-1}R_2(t)\} ,$$

where R_1 and R_2 are polynomials, and then show that for some $\varepsilon \in (0,\frac{1}{2}]$ this expansion satisfies

$$(1.9) \qquad \int_{-N^\varepsilon}^{N^\varepsilon} \left|\frac{\psi_N(t)-\tilde{\psi}_N(t)}{t}\right| dt = o(N^{-1}) .$$

This is generally a difficult and highly technical part of the proof, the difficulty lying not so much in finding $\tilde{\psi}_N$ and proving (1.9) but in doing so under reasonably mild assumptions.

The next step is to show that for some sequence $\eta_N \to \infty$,

$$(1.10) \qquad \int_{N^\varepsilon \le |t| \le \eta_N N} \left|\frac{\psi_N(t)}{t}\right| dt = o(N^{-1}) .$$

This is a problem of an entirely different nature because (1.10) is essentially a smoothness property of the distribution function F_N and generally applicable methods for establishing it are not available. The proper method of attack seems to depend very much on the structure of the particular statistic one is considering and - except in the case of sums of i.i.d. random variables - a satisfactory sufficient condition for (1.10) is usually hard to obtain.

A trivial consequence of (1.8) is that

$$\int_{|t| \ge N^\varepsilon} \left|\frac{\tilde{\psi}_N(t)}{t}\right| dt = o(N^{-1})$$

and hence (1.9) and (1.10) imply that

$$\int_{-\eta_N N}^{\eta_N N} \left|\frac{\psi_N(t)-\tilde{\psi}_N(t)}{t}\right| dt = o(N^{-1}) .$$

But then Esseen's smoothing lemma (cf.Feller (1971), p.538) yields (1.6) with $\tilde{F}_N$ as obtained by Fourier inversion of

$$\tilde{\psi}_N(t) = \int e^{itx} \, d\, \tilde{F}_N(x) \ .$$

The aim of the present paper is to find a satisfactory sufficient condition for (1.10) in the case of the simple linear rank statistic. As we have explained, this is a crucial step in obtaining an Edgeworth expansion for this statistic; it shows that, in principle, such an expansion can be obtained but, of course, much work remains to be done in connection with (1.9). The latter problem and the resulting Edgeworth expansion will be discussed in the forthcoming Ph.D. thesis of R.J.M.M. Does (1981). We note that for the special case of the two-sample statistics, Edgeworth expansions were obtained in Bickel and Van Zwet (1978). In Van Zwet (1977) an analysis similar to the one in the present paper was carried out for linear combinations of order statistics.

2. A BOUND ON THE CHARACTERISTIC FUNCTION

Define

$$(2.1) \qquad \bar{c} = \frac{1}{N} \sum_{j=1}^{N} c_j \ , \qquad \bar{a} = \frac{1}{N} \sum_{j=1}^{N} a_j \ ,$$

and for $\zeta > 0$, let $\gamma(a_1,\ldots,a_N;\zeta)$ denote the Lebesgue measure λ of the ζ - neighborhood of the set $\{a_1,\ldots,a_N\}$, thus

$$(2.2) \qquad \gamma(a_1,\ldots,a_N;\zeta) = \lambda\{x : \exists_j |x-a_j| < \zeta\} \ .$$

Furthermore, define

$$(2.3) \qquad \phi_N(t) = E \exp\{it\, N^{-\frac{1}{2}}(T_N-\mu_N)\}$$

with T_N and μ_N as in (1.1) and (1.2). We shall prove the following result.

THEOREM 2.1

Suppose that there exist positive numbers c, C, a, A *and* δ , *a density* f *and a sequence* $\varepsilon_N \downarrow 0$ *such that*

$$(2.4)\qquad \sum_{j=1}^{N} |c_j-\bar{c}| \geq cN , \qquad \sum_{j=1}^{N} (c_j-\bar{c})^2 \leq CN ,$$

$$(2.5)\qquad \sum_{j=1}^{N} |a_j-\bar{a}| \geq aN , \qquad \sum_{j=1}^{N} (a_j-\bar{a})^2 \leq AN ,$$

$$(2.6)\qquad \gamma(a_1,\ldots,a_N;\zeta) \geq \delta N\zeta \quad \textit{for some} \quad \zeta \geq N^{-3/2}\log N ,$$

$$(2.7)\qquad \sum_{j=1}^{N} \int_{-\infty}^{\infty} \frac{(f_j(x)-f(x))^2}{f(x)}\, dx \leq \varepsilon_N N .$$

Then there exist positive numbers b, B *and* β *depending only on* c, C, a, A, δ *and the sequence* ε_N *and such that*

$$(2.8)\qquad |\phi_N(t)| \leq BN^{-\beta\log N} \quad \textit{for} \quad \log N \leq |t| \leq bN^{3/2} .$$

The proof of this theorem is a technically complicated affair and we shall split it up in a series of lemmas. To avoid the laborious formulation of the theorem in each of these lemmas we adopt the following conventions. Whenever we assume that one or more of the conditions (2.4) - (2.7) are satisfied, it will be tacitly understood that the numbers c, C, a, A and δ occurring in these conditions are indeed positive and that $\varepsilon_N \downarrow 0$. In each lemma where they appear, B_ν and β_ν are positive numbers which may depend on c, C, a, A, δ, $\{\varepsilon_N\}$ and other quantities specified in that lemma. Furthermore we define

$$\delta_1 = \frac{c^2}{64C} , \qquad \delta_2 = \frac{\delta}{16} \min\left\{\frac{\delta c}{3c+8}, 4\delta_1\right\}$$

$$(2.9)\qquad \delta_3 = \frac{a^2}{64A} , \qquad \delta_4 = \frac{\delta_3^2}{8} \min\{4\delta_1, \delta_3\}$$

$$n = [\tfrac{1}{2}N] ,$$

where $[x]$ denotes the integer part of x. Similarly, $[x]^*$ will denote the smallest integer $\geq x$.

We begin by noting that the assumptions as well as the conclusion of the theorem are invariant under simultaneous permutation of the c_j, X_j and f_j, $j = 1,\ldots,N$. By choosing a convenient permutation we arrive at

LEMMA 2.1

To prove theorem 2.1 *it suffices to prove it under the additional assumption that*

$$(2.10) \qquad c_{2j} - c_{2j-1} \geq \frac{c}{4} \qquad \textit{for} \quad j = 1,\ldots,[\delta_1 N]^* .$$

Proof

Let $c_{(1)} \leq c_{(2)} \leq \ldots \leq c_{(N)}$ be the nondecreasing rearrangement of $c_1,\ldots,c_N$ and define $b_j = c_{(N-j+1)} - c_{(j)} \geq 0$ for $j = 1,\ldots,n = [N/2]$. Since we are allowed to permute $c_1,\ldots,c_N$ provided that we permute $X_1,\ldots,X_N$ and $f_1,\ldots,f_N$ in the same way, the lemma is proved if we show that (2.4) implies that

$$(2.11) \qquad b_j \geq \frac{c}{4} \qquad \text{for at least } [\delta_1 N]^* \text{ indices } j .$$

Let $\tilde{c}$ be a median of $c_1,\ldots,c_N$. Then

$$\sum_{j=1}^{n} b_j = \sum_{j=1}^{n} (c_{(N-j+1)} - \tilde{c}) + \sum_{j=1}^{n} (\tilde{c} - c_{(j)}) = \sum_{j=1}^{N} |c_j - \tilde{c}| ,$$

and because

$$\sum_{j=1}^{N} |c_j - \bar{c}| \leq \sum_{j=1}^{N} |c_j - \tilde{c}| + N|\bar{c} - \tilde{c}| \leq 2 \sum_{j=1}^{N} |c_j - \tilde{c}| ,$$

we see that (2.4) implies that

$$(2.12) \qquad \sum_{j=1}^{n} b_j \geq \tfrac{1}{2} cN .$$

Also,

$$\sum_{j=1}^{n} b_j^2 \le 2 \sum_{j=1}^{n} (c_{(N-j+1)}-\bar{c})^2 + 2 \sum_{j=1}^{n} (c_{(j)}-\bar{c})^2 \le 2 \sum_{j=1}^{N} (c_j-\bar{c})^2 ,$$

and (2.4) yields

$$(2.13) \qquad \sum_{j=1}^{n} b_j^2 \le 2CN .$$

A straightforward calculation shows that for nonnegative $b_1,\ldots,b_n$ satisfying (2.12) and (2.13), (2.11) must hold with δ_1 as in (2.9). The lemma is proved. □

In what follows we may therefore restrict attention to the case where (2.10) is satisfied. With this in mind and recalling that $f_1,\ldots,f_N$, $R_1,\ldots,R_N$ and $Z_1,\ldots,Z_N$ denote the densities, the ranks and the order statistics of $X_1,\ldots,X_N$, we define, for $j = 1,\ldots,n = [N/2]$, the random variables

$$(2.14) \qquad P_j = \frac{f_{2j-1}(X_{2j-1})f_{2j}(X_{2j})}{f_{2j-1}(X_{2j-1})f_{2j}(X_{2j}) + f_{2j-1}(X_{2j})f_{2j}(X_{2j-1})} ,$$

$$(2.15) \qquad D_j = (c_{2j}-c_{2j-1})(a_{R_{2j}}-a_{R_{2j-1}}) .$$

LEMMA 2.2

For every N *and* t , *and* n = [N/2] ,

$$(2.16) \qquad |\phi_N(t)| \le E \prod_{j=1}^{n} [1 - 2P_j(1-P_j)\{1 - \cos(N^{-\frac{1}{2}}tD_j)\}]^{\frac{1}{2}} .$$

Proof

For $j = 1,\ldots,n$, let

$$U_j = \min(R_{2j-1},R_{2j}) , \quad V_j = \max(R_{2j-1},R_{2j})$$

and consider the conditional distribution of T_N given the ordered sample $Z_1 <\ldots< Z_N$ as well as the pairs (U_j,V_j) for $j = 1,\ldots,n$. Thus, the sets $\{X_{2j-1},X_{2j}\}$ are given to us (as well as X_N if N is odd) but in each set we donot know which of the elements corresponds to X_{2j-1} or to X_{2j} respectively. According to this conditional distribution the n pairs (R_{2j-1},R_{2j}) are independent and take the values (U_j,V_j) and (V_j,U_j) with probabilities

$$\tilde{P}_j = \frac{f_{2j-1}(Z_{U_j})f_{2j}(Z_{V_j})}{f_{2j-1}(Z_{U_j})f_{2j}(Z_{V_j}) + f_{2j-1}(Z_{V_j})f_{2j}(Z_{U_j})}$$

and $(1-\tilde{P}_j)$ respectively. Since

$$T_N = \tfrac{1}{2} \sum_{j=1}^{n} (c_{2j}+c_{2j-1})(a_{R_{2j}}+a_{R_{2j-1}})$$

$$+ \tfrac{1}{2} \sum_{j=1}^{n} (c_{2j}-c_{2j-1})(a_{R_{2j}}-a_{R_{2j-1}})$$

$$(+\ c_N\, a_{R_N} \quad \text{if } N \text{ is odd})$$

and $(a_{R_{2j}}+a_{R_{2j-1}})$ and a_{R_N} are given, we have

$$|E(\exp\{it\, N^{-\frac{1}{2}}(T_N-\mu_N)\}|Z_1,\ldots,Z_N,U_1,V_1,\ldots,U_n,V_n)|$$

$$= \prod_{j=1}^{n} |E\ \exp\{\tfrac{1}{2}it\ N^{-\frac{1}{2}}(c_{2j}-c_{2j-1})(a_{V_j}-a_{U_j})W_j\}| \ ,$$

where $P(W_j=1) = 1-P(W_j = -1) = \tilde{P}_j$. This, in turn, equals

$$\prod_{j=1}^{n} [1 - 2\tilde{P}_j(1-\tilde{P}_j)\{1 - \cos(N^{-\frac{1}{2}}t(c_{2j}-c_{2j-1})(a_{V_j}-a_{U_j}))\}]^{\frac{1}{2}}$$

$$= \prod_{j=1}^{n} [1 - 2P_j(1-P_j)\{1 - \cos(N^{-\frac{1}{2}}tD_j)\}]^{\frac{1}{2}}$$

because $\tilde{P}_j(1-\tilde{P}_j) = P_j(1-P_j)$ and $|(c_{2j}-c_{2j-1})(a_{V_j}-a_{U_j})| = |D_j|$. The lemma follows upon taking expectations. □

We note that the idea of bounding $|\phi_N|$ by this conditioning argument occurs in some unpublished notes of H. Kesten concerning a local limit theorem for T_N .

Consider real numbers $d_1,\dots,d_m$ and $p_1,\dots,p_m$ with $0 \le p_j \le 1$ for $j = 1,\dots,m$. For $\zeta > 0$ and $0 < \varepsilon < \frac{1}{2}$, let $\gamma(d_1,\dots,d_m;p_1,\dots,p_m;\zeta,\varepsilon)$ denote the Lebesgue measure λ of the ζ - neighborhood of the set of those d_j for which the corresponding p_j satisfy $\varepsilon \le p_j \le 1-\varepsilon$, thus

$$\gamma(d_1,\dots,d_m;p_1,\dots,p_m;\zeta,\varepsilon) = \lambda\{x : \exists_j |x-d_j| < \zeta,\ \varepsilon \le p_j \le 1-\varepsilon\} .$$

LEMMA 2.3

Suppose that positive numbers d, D, δ' *and* ε *exist such that*

$$(2.17)\qquad \sum_{j=1}^{m} p_j(1-p_j)d_j^2 \ge dm ,\quad \sum_{j=1}^{m} d_j^4 \le Dm ,$$

$$(2.18)\qquad \gamma(d_1,\dots,d_m;p_1,\dots,p_m;\zeta',\varepsilon) \ge \delta' m\zeta'$$

for some $\zeta' \ge m^{-3/2}\log m$. *Then, for every positive* b_1 , *there exist positive numbers* b_2 , B *and* β *depending only on* d, D, δ', ε *and* b_1 *and such that*

$$(2.19)\qquad \prod_{j=1}^{m} [1 - 2p_j(1-p_j)\{1 - \cos(m^{-\frac{1}{2}}td_j)\}]^{\frac{1}{2}} \le Bm^{-\beta\log m}$$

for $b_1\log m \le |t| \le b_2 m^{3/2}$.

Proof

The present lemma is a trivial modification of lemma 2.2 in Albers, Bickel and Van Zwet (1976). In the proof of that lemma it is shown that under the present conditions

$$(2.20)\qquad \prod_{j=1}^{m} [1 - 2p_j(1-p_j)\{1 - \cos(\tau_m^{-1}td_j)\}]^{\frac{1}{2}} \le Bm^{-\beta\log m}$$

for $\tau_m^2 = \Sigma p_j(1-p_j)d_j^2$ and $\log(m+1) \le |t| \le bm^{3/2}$. Also, inspection of this proof reveals at once that for any $\tilde{b} > 0$, (2.20) will continue to hold for $\tilde{b}\log m \le |t| \le bm^{3/2}$ with possibly different B and β . In view of (2.17) we have $dm \le \tau_m^2 \le D^{\frac{1}{2}}m$ and hence (2.20) for $\tilde{b}\log m \le |t| \le bm^{3/2}$ implies (2.19) for $b_1\log m \le |t| \le b_2 m^{3/2}$ with $b_1 = \tilde{b}d^{-\frac{1}{2}}$ and $b_2 = bD^{-\frac{1}{4}}$. This proves the lemma. □

We should perhaps point out that lemma 2.2 in Albers, Bickel and Van Zwet (1976) served to prove the analogue of (1.10) for the one-sample rank statistic. The same lemma was invoked in Bickel and Van Zwet (1978) to deal with this problem for the two-sample rank statistic. As this is a special case of the simple linear rank statistic discussed in the present paper, we should not be surprised to see the same result turn up again as a major tool in the form of lemma 2.3.

A comparison of (2.16) and (2.19) clearly reveals our plan of attack. We shall show that with large probability there is a large set of indices $J \subset \{1,\dots,n\}$ such that the sequences P_j and D_j for $j \in J$ satisfy conditions (2.17) and (2.18). We may then use (2.19) to bound the product on the right in (2.16) on a set of large probability and this will yield the desired bound for $|\phi_N(t)|$. As a first step we prove

LEMMA 2.4

If (2.7) *holds then for every* $\varepsilon \in (0,\frac{1}{2})$ *and* $\eta \in (0,\frac{1}{2})$,

$$P(\varepsilon \le P_j \le 1-\varepsilon \text{ for at least } [\eta N]^* \text{ indices } j\ ) \ge 1-B_1 e^{-\beta_1 N} .$$

Proof

For $j = 1,\dots,n = [N/2]$,

$$E|2P_j-1| = E\left[\frac{|f_{2j-1}(X_{2j-1})f_{2j}(X_{2j}) - f_{2j-1}(X_{2j})f_{2j}(X_{2j-1})|}{f_{2j-1}(X_{2j-1})f_{2j}(X_{2j}) + f_{2j-1}(X_{2j})f_{2j}(X_{2j-1})}\right]$$

and since the expression in square brackets is symmetric in X_{2j-1} and X_{2j} , we find

$$E|2P_j-1| = \tfrac{1}{2}\iint |f_{2j-1}(x)f_{2j}(y) - f_{2j-1}(y)f_{2j}(x)|\ dxdy$$

$$\le \int |f_{2j}(x) - f_{2j-1}(x)|\ dx .$$

By Markov's inequality

$$P(P_j \notin [\varepsilon,1-\varepsilon]) = P(|2P_j-1| \geq 1-2\varepsilon) \leq \frac{\int |f_{2j}-f_{2j-1}|}{1-2\varepsilon}$$

and because (2.7) yields

$$\frac{1}{N}\sum_{j=1}^{n}\int |f_{2j}-f_{2j-1}| \leq \frac{1}{N}\sum_{j=1}^{N}\int |f_j-f| \leq \left[\frac{1}{N}\sum_{j=1}^{N}\int \frac{(f_j-f)^2}{f}\right]^{\frac{1}{2}} \leq \varepsilon_N^{\frac{1}{2}} ,$$

we have

$$\sum_{j=1}^{n} P(P_j \notin [\varepsilon,1-\varepsilon]) \leq \frac{\varepsilon_N^{\frac{1}{2}} N}{1-2\varepsilon} .$$

The lemma now follows from Bernstein's inequality (cf. Hoeffding (1963)). □

Under the model we are discussing, $X_1,\ldots,X_N$ are independent with densities $f_1,\ldots,f_N$ and probabilities and expectations under this model are indicated by P and E. We now introduce an auxiliary model under which $X_1,\ldots,X_N$ are independent and identically distributed with a common density f and we shall write P_0 and E_0 for probabilities and expectations under this model. Note that the model P_0 satisfies the hypothesis $H : f_1=\ldots=f_N$ discussed in section 1. Of course, most probabilistic calculations are much easier under P_0 than under P. Lemma 2.5 will allow us to do the remainder of our computations under the easier model.

LEMMA 2.5

If assumption (2.7) *is satisfied, then for every event* A *in the* σ *- algebra generated by* $X_1,\ldots,X_N$,

$$(2.21) \qquad P(A) \leq 2\left\{e^{\varepsilon_n N} P_0(A)\right\}^{\frac{1}{2}} .$$

Proof

Define the likelihood ratio

$$L = \prod_{j=1}^{N} \frac{f_j(X_j)}{f(X_j)} .$$

Clearly (2.7) implies that $P(L=\infty) = 0$ and

$$EL = \prod_{j=1}^{N} \int \frac{f_j^2(x)}{f(x)} dx = \prod_{j=1}^{N} \left\{1 + \int \frac{(f_j(x)-f(x))^2}{f(x)} \mathbf{dx} \right\}$$

$$\le \exp\left\{\sum_{j=1}^{N} \int \frac{(f_j-f)^2}{f} \right\} \le \exp\{\varepsilon_N N\} .$$

Hence for $\xi > 0$, Markov's inequality yields

$$P(A) \le P(A \cap \{L \le \xi\}) + P(L > \xi)$$

$$\le \xi\, P_0(A) + \xi^{-1}\exp\{\varepsilon_N N\} ,$$

and (2.21) follows by taking $\xi^2 = \exp\{\varepsilon_N N\}/P_0(A)$. □

Now we may continue our task of checking conditions (2.17) and (2.18) for (subsequences of) P_j and D_j .

LEMMA 2.6

Suppose that assumptions (2.6), (2.7) *and* (2.10) *are satisfied and let* δ_2 *be defined by* (2.9). *Then for some* $\zeta \ge N^{-3/2}\log N$,

$$(2.22) \qquad P(\gamma(D_1,\ldots,D_n;\zeta) \ge \delta_2 N\zeta) \ge 1 - B_2 e^{-\beta_2 N} .$$

Proof

Let us first consider the situation under the model P_0 , so that $(R_1,\ldots,R_N)$ equals each permutation of $(1,\ldots,N)$ with probability $1/N!$. Take ζ as in (2.6) and

$$(2.23) \qquad r = \left[\min\left(\frac{\delta c N}{4(3c+8)} , \delta_1 N\right)\right] .$$

Given $R_1,R_3,\ldots,R_{2r-1}$, we build up $\gamma(D_1,\ldots,D_r;\zeta)$ in r steps by successively choosing $R_2,R_4,\ldots,R_{2r}$ at random without replacement from

$\{1,\dots,N\} \setminus \{R_1,R_3,\dots,R_{2r-1}\}$ and running through the sequence $\gamma(D_1;\zeta)$, $\gamma(D_1,D_2;\zeta)$, ..., $\gamma(D_1,\dots,D_r;\zeta)$. If we choose R_{2k} in such a way that D_k is not contained in the 2ζ - neighborhood of $\{D_1,\dots,D_{k-1}\}$, then $\gamma(D_1,\dots,D_k;\zeta) = \gamma(D_1,\dots,D_{k-1};\zeta) + 2\zeta$. This is the case unless $|D_k - D_j| < 2\zeta$ for some $j = 1,\dots,k-1$, i.e. unless

$$(2.24) \qquad (a_{R_{2k}} - a_{R_{2k-1}}) \in \bigcup_{j=1}^{k-1} \left(\frac{D_j-2\zeta}{c_{2k}-c_{2k-1}}, \frac{D_j+2\zeta}{c_{2k}-c_{2k-1}}\right).$$

Since $k \le r \le [\delta_1 N]$, (2.10) ensures that $c_{2k} - c_{2k-1} \ge c/4 > 0$ and hence (2.24) restricts $a_{R_{2k}}$ to a set A_k which is the union of $(k-1)$ intervals of length $\le 16\zeta/c$. The set of a_j in A_k has a ζ - neighborhood of Lebesgue measure at most $(k-1)\{(16\zeta/c) + 2\zeta\}$, so (2.6) implies that the number of j for which $a_j \notin A_k$ equals at least

$$\frac{1}{2\zeta}\{\delta N\zeta - 2(k-1)\zeta(1 + \frac{8}{c})\} = \tfrac{1}{2}\delta N - (k-1)(1 + \frac{8}{c}) .$$

Subtracting the $(r+k-1)$ indices $R_1,R_3,\dots,R_{2r-1},R_2,R_4,\dots,R_{2k-2}$ chosen before R_{2k} and for which the corresponding a_j may well be outside A_k, we find that the conditional probability that $a_{R_{2k}} \notin A_k$ given $R_1,R_3,\dots,R_{2r-1},R_2,R_4,\dots,R_{2k-2}$ equals at least

$$\frac{\tfrac{1}{2}\delta N - (k-1)(1 + \frac{8}{c}) - (r+k-1)}{N - (r+k-1)} \ge \tfrac{1}{2}\delta - \frac{r}{N}(3 + \frac{8}{c}) \ge \frac{\delta}{4}$$

in view of (2.23). As $a_{R_{2k}} \notin A_k$ implies that 2ζ is added to γ at the k-th step, we see that $\gamma(D_1,\dots,D_r;\zeta)/2\zeta$ is stochastically larger than a binomial random variable with parameters r and $\delta/4$. Since $\gamma(D_1,\dots,D_n;\zeta) \ge \gamma(D_1,\dots,D_r;\zeta)$ and $2\zeta r\delta/4 = 2\delta_2 N\zeta$, Bernstein's inequality ensures that for positive B and β

$$P_0(\gamma(D_1,\dots,D_n;\zeta) \le \delta_2 N\zeta) \le B\, e^{-\beta N} .$$

Application of lemma 2.5 yields

$$P(\gamma(D_1,\dots,D_n;\zeta) \le \delta_2 N\zeta) \le 2B^{\frac{1}{2}}\, e^{-\frac{1}{2}(\beta-\varepsilon_N)N}$$

and the proof is complete. □

Finally we need

LEMMA 2.7

If conditions (2.5), (2.7) *and* (2.10) *are satisfied and* δ_4 *is given by* (2.9), *then*

(2.25) $$P(|D_j| \ge \frac{ac}{16} \text{ for at least } [\delta_4 N]^* \text{ indices } j) \ge 1 - B_3\, e^{-\beta_3 N} .$$

Proof

Let $a_{(1)} \le \dots \le a_{(N)}$ denote the ordered $a_1,\dots,a_N$ and let $\delta_3 = a^2/(64A)$ as in (2.9). If we replace $c_1,\dots,c_N$ by $a_1,\dots,a_N$ in the proof of lemma 2.1, we find that $a_{(N-j+1)} - a_{(j)} \ge a/4$ for $j \le [\delta_3 N]^*$, i.e. that there are two sets of at least $[\delta_3 N]^*$ points a_j each, with a distance of at least $a/4$ in between. Take $r = [\min(\delta_1 N, \delta_3 N/4)]^*$ and let $j = 1,\dots,r$. Under P_0 and given $R_1,\dots,R_{2j-2}$, the conditional probability that $|a_{R_{2j}} - a_{R_{2j-1}}| \ge a/4$ is easily seen to be at least

$$\frac{2[\delta_3 N]^*[\delta_3 N-2j+2]^*}{N^2} \ge \delta_3^2$$

because $j-1 \le r-1 \le [\delta_3 N/4]$. It follows that the number of indices $j \le r$ for which $|a_{R_{2j}} - a_{R_{2j-1}}| \ge a/4$ is stochastically larger under P_0 than a binomial random variable with parameters r and δ_3^2 . Since $r\delta_3^2 \ge 2\delta_4 N$, Bernstein's inequality yields positive B and β such that

(2.26) $$P_0(|a_{R_{2j}} - a_{R_{2j-1}}| \ge \frac{a}{4} \text{ for at least } [\delta_4 N]^* \text{ indices } j \le r) \ge 1 - B\, e^{-\beta N} .$$

But if $j \leq r$, then $j \leq [\delta_1 N]^*$ and $c_{2j} - c_{2j-1} \geq c/4$ by (2.10). Hence (2.26) implies

$$P_0(|D_j| \geq \frac{ac}{16} \text{ for at least } [\delta_4 N]^* \text{ indices } j\ ) \geq 1 - B\, e^{-\beta N} .$$

The transition from P_0 to P is again achieved by applying lemma 2.5. □

We have now assembled the necessary machinery for establishing the theorem.

Proof of theorem 2.1

In view of lemma 2.1 we may assume that (2.4) - (2.7) and (2.10) are satisfied for positive c, C, a, A and δ, f and $\varepsilon_N \downarrow 0$. Take δ_2 and δ_4 as in (2.9), choose $\varepsilon \in (0,\frac{1}{2})$ and define

$$\text{(2.27)} \quad \begin{aligned} \delta_5 &= \frac{1}{8}\min(\delta_2, 2\delta_4) , & D &= \left\{\frac{2(A+C)}{\delta_5}\right\}^4 , \\ d &= 2^{-8}\, \varepsilon(1-\varepsilon)a^2c^2\delta_4 , & \delta' &= \frac{1}{8}\,\delta_2 . \end{aligned}$$

Let $J \subset \{1,\ldots,n\}$ be the random set of indices j for which $|D_j| \leq D^{\frac{1}{4}}$ and let M be the cardinality of J , thus

$$J = \{j : |D_j| \leq D^{\frac{1}{4}}\} , \qquad M = |J| .$$

Because of (2.4) and (2.5),

$$\sum_{j=1}^{n} c_{2j} - c_{2j-1})^2 \leq 2 \sum_{j=1}^{N} (c_j - \bar{c})^2 \leq 2CN ,$$

$$\sum_{j=1}^{n} (a_{R_{2j}} - a_{R_{2j-1}})^2 \leq 2 \sum_{j=1}^{N} (a_j - \bar{a})^2 \leq 2AN$$

and hence the sets $\{j : |c_{2j} - c_{2j-1}| \geq D^{1/8}\}$ and $\{j : |a_{R_{2j}} - a_{R_{2j-1}}| \geq D^{1/8}\}$ have cardinalities at most $2CND^{-\frac{1}{4}}$ and $2AND^{-\frac{1}{4}}$ respectively. It follows that $M \geq n - 2(A+C)ND^{-\frac{1}{4}}$ and because of (2.27)

$$\text{(2.28)} \quad n - \delta_5 N \leq M \leq n$$

with probability 1. Since the assumptions of the theorem trivially imply that $N \geq 2$ so that $n \geq N/3$ and since certainly $\delta_5 \leq 1/12$, we also have the following crude but useful bounds in terms of N

$$(2.29) \qquad \frac{1}{4} N \leq M \leq \frac{1}{2} N .$$

Take $\zeta = N^{-3/2} \log N$ and define the event F by

$$F = \{\varepsilon \leq P_j \leq 1-\varepsilon \text{ for at least } [(\tfrac{1}{2}-\delta_5)N]^* \text{ indices } j\} \cap$$
$$\cap \{\gamma(D_1,\ldots,D_n;\zeta) \geq \delta_2 N\zeta\} \cap$$
$$\cap \{|D_j| \geq \frac{ac}{16} \text{ for at least } [\delta_4 N]^* \text{ indices } j\} .$$

Application of lemma 2.4 for $\eta = \frac{1}{2} - \delta_5$ and of lemmas 2.6 and 2.7 yields

$$(2.30) \qquad P(F) \geq 1 - B_4 e^{-\beta_4 N}$$

where $B_4 = B_1 + B_2 + B_3$ and $\beta_4 = \min(\beta_1,\beta_2,\beta_3)$ are positive numbers depending only on c, C, a, A, δ and the sequence ε_N.

On the set F in our sample space, the number of indices $j \in \{1,\ldots,n\}$ for which $\varepsilon \leq P_j \leq 1-\varepsilon$ as well as $|D_j| \geq ac/16$, equals at least $(\delta_4-\delta_5)N$. Because of (2.28), $(\delta_4-2\delta_5)N$ of these indices must also belong to J. Combining this with (2.29) and (2.27) we find that

$$(2.31) \qquad \sum_{j\in J} P_j(1-P_j)D_j^2 \geq \varepsilon(1-\varepsilon)\left(\frac{ac}{16}\right)^2 (\delta_4-2\delta_5)N \geq dM$$

for every sample point in F. Similarly, we see that on F the number of indices j for which either $j \notin J$ or $P_j \notin [\varepsilon,1-\varepsilon]$, equals at most $2\delta_5 N$ and hence

$$\gamma(D_j,j \in J;\ P_j,j \in J;\ \zeta,\varepsilon) \geq (\delta_2-4\delta_5)N\zeta \geq \delta_2 M\zeta .$$

Take $\zeta' = M^{-3/2}\log M$. If $M \geq 2$, then (2.29) ensures that $1/8 \leq \zeta/\zeta' < 1$ and as γ is obviously nondecreasing in ζ

$$(2.32)\qquad \gamma(D_j, j \in J;\ P_j, j \in J;\ \zeta', \varepsilon) \geq \frac{1}{8}\,\delta_2 M\zeta' = \delta' M\zeta' \ .$$

Since this is trivially true for $M = 1$ also, (2.32) holds for every sample point in F . Finally, the definition of J implies that

$$(2.33)\qquad \sum_{j\in J} D_j^4 \leq DM \ .$$

We have shown that on the set F the sequences D_j and P_j, $j \in J$, satisfy the assumptions of lemma 2.3 for values of d, D, δ' and ε which depend only on c, C, a, A, δ and the sequence ε_N . Application of lemma 2.3 with $b_1 = \frac{1}{2}$ yields the existence of positive numbers b_2, B_5 and β_5 depending only on c, C, a, A, δ and $\{\varepsilon_N\}$ and such that for every sample point in F ,

$$\prod_{j=1}^{n} \left[1 - 2\,P_j(1-P_j)\{1 - \cos(N^{-\frac{1}{2}}tD_j)\}\right]^{\frac{1}{2}}$$

$$\leq \prod_{j\in J} \left[1 - 2\,P_j(1-P_j)\{1 - \cos(M^{-\frac{1}{2}}\left(\frac{M}{N}\right)^{\frac{1}{2}}tD_j)\}\right]^{\frac{1}{2}}$$

$$\leq B_5\, M^{-\beta_5 \log M}$$

for $\frac{1}{2}\log M \leq (M/N)^{\frac{1}{2}}|t| \leq b_2 M^{3/2}$. An easy calculation based on (2.29) shows that this implies that positive B_6 and β_6 exist depending only on B_5 and β_5 , such that on F ,

$$(2.34)\qquad \prod_{j=1}^{n} \left[1 - 2\,P_j(1-P_j)\{1 - \cos(N^{-\frac{1}{2}}tD_j)\}\right]^{\frac{1}{2}} \leq B_6\, N^{-\beta_6 \log N}$$

for $\log N \leq |t| \leq bN^{3/2}$, where $b = b_2/4$. Combining (2.30), (2.34) and lemma 2.2 we find that for $\log N \leq |t| \leq bN^{3/2}$,

$$|\phi_N(t)| \leq B_6\, N^{-\beta_6 \log N} + B_4\, e^{-\beta_4 N} \leq B\, N^{-\beta \log N}$$

with $B = B_4 + B_6$ and $\beta = \min(\beta_4, \beta_6)$. The theorem is proved. □

3. COMMENTS

In this section we provide a discussion of theorem 2.1. First of all we should point out that the standardization of T_N in the theorem is different from the one in section 1. If $\sigma^2(T_N)$ is of exact order N, then the difference between $N^{-\frac{1}{2}}(T_N-\mu_N)$ and T_N^* is of no importance and (1.10) follows immediately from (2.8). Under the hypothesis $H : f_1=\ldots=f_N$ we know that the variance $\sigma_0^2(T_N)$ of T_N is given by

$$\sigma_0^2(T_N) = \frac{1}{N-1} \sum_{j=1}^{N} (c_j-\bar{c})^2 \sum_{j=1}^{N} (a_j-\bar{a})^2$$

and assumptions (2.4) and (2.5) imply that for $N \geq 2$

$$a^2c^2N \leq \sigma_0^2(T_N) \leq 2ACN ,$$

so that under H no problems can arise. In general, the conditions of theorem 2.1 imply that $\sigma^2(T_N) \geq \alpha N$ for some $\alpha > 0$; to see this use (2.31) to obtain a lower bound for the variance of the conditional distribution of T_N discussed in lemma 2.2. However, the conditions of the theorem would seem to be too weak to guarantee that $\sigma^2(T_N)$ is not of larger order than N. We shall not pursue this matter further because one also has to prove (1.9) to establish an Edgeworth expansion and the much stronger conditions needed to do this will typically imply that $\sigma^2(T_N)$ is indeed of exact order N.

A second general remark is that, even though theorem 2.1 is formulated in terms of bounds for an arbitrary but fixed value of N, it is strictly an asymptotic result since b, B and β are not specified. The fact that these numbers depend on the sequences of regression constants, scores and densities only through c, C, a, A, δ and $\{\varepsilon_N\}$, allows us to apply the theorem to triangular arrays $c_{1,N},\ldots,c_{N,N}$, $a_{1,N},\ldots,a_{N,N}$ and $f_{1,N},\ldots,f_{N,N}$, $N = 1,2,\ldots$, provided they satisfy (2.4) - (2.7) for $N = 1,2,\ldots$, for fixed values of c, C, a, A, δ and $\varepsilon_N \downarrow 0$.

Next, let us comment on each of the conditions of theorem 2.1 separately. Assumption (2.4) may be replaced by

$$(3.1) \qquad \sum|c_j-\bar{c}|^r \geq c'N \ , \qquad \sum|c_j-\bar{c}|^s \leq C'N$$

for positive c' and C' and for some $s \geq 2$ and $0 < r < s$. For $s = 2$ this is equivalent to (2.4) and for $s > 2$ it is stronger. When proving (1.9) to establish the Edgeworth expansion, one will typically require (3.1) for $r = 2$ and $s = 4$ and (2.4) will then automatically be satisfied. The same remark applies to assumption (2.5).

Assumption (2.6) is well-known from previous work on one- and two-sample rank statistics (cf. Albers, Bickel and Van Zwet (1976) and Bickel and Van Zwet (1978)). Its role is to ensure that the scores $a_1,\ldots,a_N$ donot cluster too much around too few points, thus preventing a too pronounced lattice character of the distribution of T_N even in the case where the c_j equal 0 or 1 as they do in the two-sample problem. An equivalent formulation of (2.6) is that there exists a positive fraction of the scores $a_1,\ldots,a_N$, which are at a distance of at least $N^{-3/2}\log N$ apart from each other.

The density f that minimizes the left-hand side of (2.7) is given by $f^2 = (KN)^{-1}\sum f_j^2$ and for this choice of f , (2.7) reduces to

$$(3.2) \qquad \int\left\{\frac{1}{N}\sum_{j=1}^{N} f_j^2(x)\right\}^{\frac{1}{2}} dx \leq (1+\varepsilon_N)^{\frac{1}{2}} \ ,$$

which is therefore equivalent to the final assumption of the theorem. Taking $\bar{f} = N^{-1}\sum f_j$ and $\varepsilon_N' = (1+\varepsilon_N)^{\frac{1}{2}} - 1 \downarrow 0$, one easily verifies that a sufficient condition for (3.2) is

$$(3.3) \qquad \int\left\{\frac{1}{N}\sum_{j=1}^{N} (f_j(x) - \bar{f}(x))^2\right\}^{\frac{1}{2}} dx \leq \varepsilon_N' \downarrow 0 \ .$$

To see how restrictive this assumption is, one should realize that for power computations, Edgeworth expansions are of interest mainly for sequences of alternatives $(f_{1,N},\ldots,f_{N,N})$, $N = 1,2,\ldots$, which are contiguous to the hypothesis, i.e. for which the sequence of joint densities $\Pi_{j=1}^{N} f_{j,N}(x_j)$ is contiguous to a sequence $\Pi_{j=1}^{N} f_N(x_j)$ for some choice of f_N , $N = 1,2,\ldots$. A simple computation based on theorem 1 in Oosterhoff and Van Zwet (1979) shows that this contiguity assumption implies that for some positive C'' and all N ,

$$(3.4) \qquad \int\left\{\frac{1}{N}\sum_{j=1}^{N}(f_{j,N}(x) - f_N(x))^2\right\}^{\frac{1}{2}} dx \le C'' N^{-\frac{1}{2}}$$

for some f_N and therefore certainly for $f_N = \bar{f}_N = \frac{1}{N}\sum f_{j,N}$. It follows that contiguity is a much stronger assumption than (3.3), (3.2) or (2.7) and we conclude that the latter condition doesn't really restrict the scope of the theorem at all.

Let us finally consider theorem 2.1 for the special case of the two-sample rank statistic where $c_1=\ldots=c_m = 0$, $c_{m+1}=\ldots=c_N = 1$, $f_1=\ldots=f_m = g_N$ and $f_{m+1}=\ldots=f_N = \tilde{g}_N$. In this case (2.4) reduces to the requirement that m/N is bounded away from 0 and 1 and if this holds (3.3) is satisfied if $\int|g_N-\tilde{g}_N| \to 0$. Combined with (2.5) and (2.6) these conditions appear to be comparable to those needed in Bickel and Van Zwet (1978) for establishing (1.10) for the two-sample rank statistic.

REFERENCES

[1] ALBERS, W., BICKEL, P.J. and VAN ZWET, W.R. (1976). Asymptotic expansions for the power of distributionfree tests in the one-sample problem. *Ann. Statist.* 4, 108-156.

[2] BICKEL, P.J. and VAN ZWET, W.R. (1978). Asymptotic expansions for the power of distributionfree tests in the two-sample problem. *Ann. Statist.* 6, 937-1004.

[3] FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications*, Vol.2, 2^{nd} Edition, Wiley, New York.

[4] HÁJEK, J. and ŠIDÁK (1967). *Theory of Rank Tests*. Academic Press, New York.

[5] HODGES, J.L. Jr. and LEHMANN, E.L. (1970). Deficiency. *Ann. Math. Statist.* 41, 783-801.

[6] HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* 58, 13-30.

[7] OOSTERHOFF, J. and VAN ZWET, W.R. (1979). A note on contiguity and Hellinger distance. *Contributions to Statistics, J. Hájek Memorial Volume*, J. Jurečková editor. Academia, Prague, 157-166.

[8] VAN ZWET, W.R. (1977). Asymptotic expansions for the distribution functions of linear combinations of order statistics. *Statistical Decision Theory and Related Topics* II, S.S. Gupta and D.S. Moore editors. Academic Press, New York, 421-437.