ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
Hugo Steinhaus Jubilee Volume
X (1969)

V. DUPAČ and J. HÁJEK (Praha)

ASYMPTOTIC NORMALITY OF THE WILCOXON STATISTIC UNDER DIVERGENT ALTERNATIVES

1. Introduction. Throughout this paper, $X_{\nu 1}, \ldots, X_{\nu m_\nu}$ and $X_{\nu, m_\nu+1}, \ldots, X_{\nu, m_\nu+n_\nu}$ denote two samples from populations with continuous distribution functions F_ν and G_ν, respectively. We write N_ν for $m_\nu + n_\nu$, $R_{\nu i}$ for the rank of $X_{\nu i}$ in the joint sample $X_{\nu 1}, \ldots, X_{\nu N_\nu}$, and S_ν for the Wilcoxon statistic $\sum_{i=1}^{m_\nu} R_{\nu i}$; all symbols being defined for $\nu = 1, 2, \ldots$

We shall be interested in the problem of asymptotic normality of the Wilcoxon statistic: Under which conditions concerning m_ν, n_ν, F_ν, G_ν is the distribution of S_ν asymptotically normal? A sufficient condition given by the Chernoff-Savage theorem [1] reads: N_ν tends to $+\infty$ and each of the sequences

$$\frac{m_\nu}{N_\nu}, \quad \frac{n_\nu}{N_\nu}, \quad \operatorname{var} G_\nu(X_{\nu 1}), \quad \operatorname{var} F_\nu(X_{\nu N_\nu}) \tag{1}$$

is bounded away from zero.

In a special case (location alternatives), when $F_\nu \equiv F$, $G_\nu(x) = F(x - \Delta_\nu)$, F' exists and its carrier is an interval of length $\Delta \leqslant +\infty$, the condition on variances reduces to the boundedness of $|\Delta_\nu|$ away from Δ.

In the present paper we try to get closer to necessary conditions. It will be shown that

(i) all the boundedness-from-zero conditions imposed on sequences (1) can be relaxed simultaneously (so that, especially, for the location alternatives, the asymptotic normality holds even for $|\Delta_\nu|$ tending — sufficiently slowly — to Δ);

(ii) more detailed statements concerning location alternatives are available for normal or rectangular F.

As to the methodology, the main tools are an inequality on projections of linear rank order statistics (Hájek [2]) and a lemma presenting sufficient conditions for asymptotic normality in terms of conditional distributions.

2. General alternatives.

THEOREM 1. *If*

$$\lim_{\nu\to+\infty}\frac{m_\nu n_\nu}{N_\nu}\left(\frac{n_\nu}{N_\nu}\operatorname{var}\,G_\nu(X_{\nu 1})+\frac{m_\nu}{N_\nu}\operatorname{var}\,F_\nu(X_{\nu N_\nu})\right)=+\infty, \tag{2}$$

then S_ν *is asymptotically normal* $(ES_\nu, \operatorname{var} S_\nu)$ *and*

$$\lim_{\nu\to+\infty}\operatorname{var}S_\nu/\left(m_\nu n_\nu^2\operatorname{var}G_\nu(X_{\nu 1})+m_\nu^2 n_\nu\operatorname{var}F_\nu(X_{\nu N_\nu})\right)=1. \tag{3}$$

Proof. The essential points of the proof have been already prepared in [2]. We first approximate S (we shall drop the index ν if there is no danger of confusion) by the statistic $\hat{S}=ES+\sum_{i=1}^{N}Y_i$, where $Y_i=E(S\,|\,X_i)-ES$. The Y_i's are independent, have zero expectations, and can be expressed as

$$Y_i=\sum_{j=1}^{m}\left(E(R_j\,|\,X_i)-ER_j\right)=\begin{cases}-nG(X_i)+\varkappa_i, & i=1,\dots,m,\\ mF(X_i)+\varkappa_i', & i=m+1,\dots,N\end{cases} \tag{4}$$

with some constants $\varkappa_i, \varkappa_i'$. It easily follows that var $\hat{S}$ is given by the denominator in (3) and that $|Y_i|\leqslant N$, $i=1,\dots,N$, which together with (2) entails that ess. sup $|Y_i|/(\operatorname{var}\hat{S})^{\frac{1}{2}}$ tends to zero. Thus the Lindeberg condition for the asymptotic normality of $\hat{S}$ with parameters $(E\hat{S}, \operatorname{var}\hat{S})$ is trivially satisfied. The asymptotic normality of S with parameters $(ES, \operatorname{var} S)$ and the relation $\operatorname{var} S/\operatorname{var}\hat{S}\to 1$ will be entailed by $E(S-\hat{S})^2/\operatorname{var}\hat{S}\to 0$. This however, follows from an inequality (Th. 4.1. in [2]) which — specialized to our case — gives $E(S-\hat{S})^2\leqslant\frac{1}{2}mn$, and from the relation (2).

THEOREM 2. *If for some* A_ν, $\nu\geqslant 1$, *the sequence*

$$P(X_{\nu i}<A_\nu<X_{\nu j},\ 1\leqslant i\leqslant m_\nu,\ m_\nu+1\leqslant j\leqslant N_\nu) \tag{5}$$

is bounded away from zero, then S_ν *is not asymptotically normal* (μ_ν, σ_ν^2) *for any* μ_ν, $\sigma_\nu>0$.

Proof is evident, since (5) implies that $P\left(S_\nu=\frac{1}{2}m_\nu(m_\nu+1)\right)$, $\nu\geqslant 1$, are bounded away from zero, which makes the asymptotic normality of S_ν impossible.

3. Location alternatives.

From now on, we shall confine ourselves to the case

$$F_\nu(x)=F(x),\qquad G_\nu(x)=F(x-\Delta_\nu)\quad\text{for all } x \text{ and } \nu. \tag{6}$$

THEOREM 3. *Let the distribution* F *be symmetric (about the median). Let*

$$\lim_{\nu\to\infty}\min(m_\nu, n_\nu)\operatorname{var}F(X_{\nu 1}-\Delta_\nu)=+\infty. \tag{7}$$

Then S_ν is asymptotically normal (ES_ν, var S_ν) and

$$\lim_{\nu\to+\infty} \operatorname{var} S_\nu/\left(m_\nu n_\nu N_\nu \operatorname{var} F(X_{\nu 1}-\Delta_\nu)\right) = 1.$$

Proof. For all x we have $F(x) = 1-F(2c-x)$, where c is the median. Hence

$$\int_{G_\nu(u)<t} dF(u) = \int_{1-F(2c-u+\Delta_\nu)<t} dF(u) = \int_{1-F(v)<t} dG_\nu(t),$$

i.e., $G(X_{\nu 1})$ and $1-F(X_{\nu N_\nu})$ have the same distributions, which implies that

$$\operatorname{var} F(X_{\nu N_\nu}) = \operatorname{var} G_\nu(X_{\nu 1}) = \operatorname{var} F(X_{\nu 1}-\Delta_\nu).$$

Inserting this into (2) and (3), we get the statement of the theorem.

4. Normal location alternatives. In addition to (6), assume that

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt.$$

Without loss of generality suppose that $m_\nu \leqslant n_\nu$, $\nu \geqslant 1$; further, let $m_\nu \to +\infty$.

THEOREM 4.

(i) *Let*

$$\limsup_{\nu\to+\infty} |\Delta_\nu|/(3\log m_\nu)^{\frac{1}{2}} < 1;$$

then S_ν is asymptotically normal (ES_ν, var S_ν). If, moreover, $|\Delta_\nu| \to +\infty$, then

$$\lim_{\nu\to+\infty} \operatorname{var} S_\nu/(3^{3/2}(2\pi)^{-1}\Delta^{-2}e^{-\Delta^2/3} m_\nu n_\nu N_\nu) = 1.$$

(ii) *Let*

$$\liminf_{\nu\to+\infty} |\Delta_\nu|/(8\log m_\nu)^{\frac{1}{2}} > 1,$$

let m_ν/N_ν be bounded away from zero; then S_ν is not asymptotically normal (μ_ν, σ_ν^2) for any μ_ν, $\sigma_\nu > 0$.

Proof. We can restrict ourselves to $\Delta_\nu \to +\infty$, since other cases are either known (Theorem 3) or can be obtained by combining known results with results for $\Delta_\nu \to +\infty$. We write $F' = f$ and drop indices ν.

(i) The main part of the proof consists in finding bounds for var $G(X_1)$, $G(x) = F(x-\Delta)$. We have

$$\operatorname{var} G(X_1) =$$

$$= \int_{-\infty}^{+\infty} G^2(x)f(x)\,dx - \Big(\int_{-\infty}^{+\infty} G(x)f(x)\,dx\Big)^2 = \int_{-\infty}^{q_1\Delta} + \int_{q_1\Delta}^{q_2\Delta} + \int_{q_2\Delta}^{+\infty} - \Big(\int_{-\infty}^{+\infty}\Big)^2 =$$

$$= I_1 + I_2 + I_3 - (I_4)^2,$$

where we shall assume that

$$0 < q_1 < \tfrac{2}{3} < q_2 < 1.$$

Using the formula

$$1-F(x) = \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{x}\left(1-\frac{\delta}{x^2}\right) e^{-\frac{x^2}{2}},$$

holding true for $x > 0$ with some $0 < \delta = \delta_x < 1$, we get

$$G^2(x)f(x)|_{x=q\Delta} = \left[1-F\left((1-q)\Delta\right)\right]^2 f(q\Delta) =$$

$$= (2\pi)^{-\frac{3}{2}} (1-q)^{-2}\Delta^{-2}\left(1-\frac{\delta}{(1-q)^2\Delta^2}\right)^2 \exp\left[-\Delta^2\{(1-q)^2+\tfrac{1}{2}q^2\}\right].$$

Since $(1-q)^2+\frac{1}{2}q^2 = \frac{3}{2}(q-\frac{2}{3})^2+\frac{1}{3}$,

$$I_1 < G^2(q_1\Delta) \int_{-\infty}^{+\infty} f(x)\,dx = o(e^{-\Delta^2(1-q_1)^2}),$$

$$I_3 < \int_{q_2\Delta}^{+\infty} f(x)\,dx = o(e^{-\Delta^2 \cdot q_2^2/2}),$$

$$I_2 = \Delta \int_{q_1}^{q_2} G^2(q\Delta)f(q\Delta)\,dq = (2\pi)^{-\frac{3}{2}}\Delta^{-1} \int_{q_1}^{q_2} (1-q)^{-2} \times$$

$$\times \left(1-\frac{\delta}{(1-q)^2\Delta^2}\right)^2 e^{-\Delta^2\left\{\frac{3}{2}\left(q-\frac{2}{3}\right)^2+\frac{1}{3}\right\}} dq$$

$$= 3^{-1/2}(2\pi)^{-1}\Delta^{-2} e^{-\frac{\Delta^2}{3}} \int_{q_1-2/3}^{q_2-2/3} \left(\frac{1}{3}-u\right)^{-2} \times$$

$$\times \left(1-\frac{\delta}{(\frac{1}{3}-u)^2\Delta^2}\right)^2 \frac{1}{(2\pi)^{1/2}3^{-1/2}\Delta^{-1}}\, e^{-u^2/2.3^{-1}\Delta^{-2}} du$$

$$= 3^{3/2}(2\pi)^{-1}\Delta^{-2} e^{-\frac{\Delta^2}{3}}\{1+o(1)\}.$$

By similar calculations we get $(I_4)^2 = o(e^{-4\Delta^2/9})$.

Choosing $q_1 < 1-3^{-1/2}$, $q_2 > 2^{1/2}3^{-1/2}$, we have ($\varepsilon > 0$) $I_1+I_3-(I_4)^2 = o(\exp[-\Delta^2(1+\varepsilon)/3])$, hence

$$\operatorname{var} G(X_1)/(3^{3/2}(2\pi)^{-1}\Delta^{-2}e^{-\Delta^2/3}) \to 1.$$

Now, if $\Delta \leqslant a\sqrt{3\log m}$ ($a < 1$), then

$$m\Delta^{-2}e^{-\Delta^2/3} \geqslant \text{const}\,\frac{m^{1-a^2}}{\log m} \to +\infty;$$

hence, the condition (7) is satisfied and the assertion (i) is thus proved.

(ii) As m/N is bounded away from zero and as $n/N \geqslant \frac{1}{2}$, we have $a'\sqrt{2\log m} > \sqrt{2\log n}$ for each $a' > 1$ and for m, n sufficiently large, and, further, $\Delta \geqslant a'\sqrt{8\log m}$ implies $\Delta > \sqrt{2\log m}+\sqrt{2\log n}$. Put $A = \sqrt{2\log m}$ and apply Theorem 2. We get

$$P(X_i < A < X_j,\ 1 \leqslant i \leqslant m,\ m+1 \leqslant j \leqslant N) = F^m(A)(1-G(A))^n$$
$$= F^m(A)F^n(\Delta - A) \geqslant F^m(\sqrt{2\log m})F^n(\sqrt{2\log n});$$

but

$$F(\sqrt{2\log k}) = 1 - \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{2\log k}}\left(1 - \frac{\delta}{2\log k}\right)e^{-\log k} = 1 - \frac{o(1)}{k},$$

hence $F^k(\sqrt{2\log k}) \to 1$ for $k \to +\infty$, which completes the proof.

5. Rectangular location alternatives. In this section we shall assume, in addition to (6), that

$$F(x) = \begin{cases} 0 & \text{for} \quad x \leqslant 0, \\ x & \text{for} \quad 0 < x \leqslant 1, \\ 1 & \text{for} \quad 1 < x. \end{cases}$$

Again, we suppose $m_\nu \leqslant n_\nu$, $m_\nu \to +\infty$.

THEOREM 5.

(i) *Let* $m_\nu(1-|\Delta_\nu|)$ *tend to* $+\infty$. *Then* S_ν *is asymptotically normal* (ES_ν, var S_ν) *with* $E\,S_\nu = \frac{1}{2}\left(m_\nu(m_\nu+1)+m_\nu n_\nu(1-|\Delta_\nu|)^2\right)$ *and*

$$\lim_{\nu\to+\infty} \operatorname{var} S_\nu \Big/ \left(\frac{1}{12}\, m_\nu n_\nu N_\nu (1+3|\Delta_\nu|)(1-|\Delta_\nu|)^3\right) = 1.$$

(*If* $|\Delta_\nu| \to 1$, *the denominator may be reduced to* $\frac{1}{3}m_\nu n_\nu N_\nu(1-|\Delta_\nu|)^3$.)

(ii) *Let* $m_\nu(1-|\Delta_\nu|)$ *be bounded* (*from above*). *Then* S_ν *is not asymptotically normal* (μ_ν, σ_ν^2) *for any* μ_ν, $\sigma_\nu > 0$.

The proof makes use of the following lemma.

LEMMA 1. *Let* T_ν *be random variables, let* Z_ν *be* k*-dimensional random vectors, let* a_ν, $s_\nu > 0$ *be functions of* k *variables,* $\nu \geqslant 1$. *Suppose that*

(A) *for each* $\varepsilon > 0$, *and each real* $a < b$,

$$\lim_{\nu\to+\infty} P\left(\max_{a\leqslant t\leqslant b}\left|E\left(\exp\left(it\big(T_\nu - a_\nu(Z_\nu)\big)/s_\nu(Z_\nu)\right)\Big|Z_\nu\right) - e^{-t^2/2}\right| < \varepsilon\right) = 1;$$

(B) $a_\nu(Z_\nu)$ *is asymptotically normal* (μ, b^2);

(C) $s_\nu^2(Z_\nu) \xrightarrow{P} \sigma^2$, $0 < \sigma^2 < +\infty$.

Then T_ν *is asymptotically normal* (μ, σ^2+b^2).

Proof of Lemma 1. Define $R_\nu(t, Z_\nu)$ by the relation

$$E\left(\exp\left(it(T_\nu - a_\nu)/s_\nu\right)|Z_\nu\right) = e^{-t^2/2} + R_\nu(t, Z_\nu).$$

Writing here ts_ν instead of t, then multiplying both sides by e^{ita_ν} and taking expectations, we get

$$E(e^{itT_\nu}) = E(e^{ita_\nu - \frac{1}{2}t^2 s_\nu^2}) + E\left(e^{ita_\nu} R_\nu(ts_\nu, Z_\nu)\right).$$

The first term on the right-hand side tends to $e^{it\mu - \frac{1}{2}t^2(b^2+\sigma^2)}$, which is easily seen from the inequality

$$|E(e^{ita_\nu - \frac{1}{2}t^2 s_\nu^2}) - e^{it\mu - \frac{1}{2}t^2(\sigma^2+b^2)}| \leqslant e^{-\frac{1}{2}t^2\sigma^2}|E(e^{ita_\nu}) - e^{it\mu + \frac{1}{2}t^2 b^2}| + \\ + E(|e^{-\frac{1}{2}t^2 s_\nu^2} - e^{-\frac{1}{2}t^2\sigma^2}|)$$

and from (B) and (C). The second term is less or equal to $E(|R_\nu(ts_\nu, Z_\nu)|)$ in absolute value. To prove its convergence to zero, it suffices — because of $|R_\nu| \leqslant 2$ — to show that $R_\nu(ts_\nu, Z_\nu) \xrightarrow{P} 0$. But this follows from the inequality $P(|R_\nu(ts_\nu, Z_\nu)| > \varepsilon) \leqslant P(\max_{t\sigma-\delta \leqslant u \leqslant t\sigma+\delta} |R(u, Z_\nu)| > \varepsilon, |ts_\nu - t\sigma| \leqslant \delta) + P(|ts_\nu - t\sigma| > \delta)$ and from (A) and (C).

Proof of Theorem 5. We confine ourselves to the case $\Delta \to 1-$, for analogous reasons as in the proof of Theorem 4.

(i) Suppose that in the interval $\langle \Delta, 1\rangle$ there lie exactly m' values of the first sample and n' values of the second one. Denote them $X'_1, \ldots, X'_{m'}, X'_{m'+1}, \ldots, X'_{m'+n'}$; further denote $R'_i = \sum_{j=1}^{m'+n'} u(X'_i - X'_j)$, where

$$u(x) = \begin{cases} 1 & \text{for} \quad x \geqslant 0, \\ 0 & \text{for} \quad x < 0, \end{cases}$$

and put $S' = \sum_{i=1}^{m'} R'_i$. Now,

$$S = S' + \tfrac{1}{2}(m-m')(m-m'+1) + m'(m-m') = \\ = S' + \tfrac{1}{2}(m-m')(m+m'+1). \tag{8}$$

Here m' and n' are independent binomial variables with parameters $(m, 1-\Delta)$ and $(n, 1-\Delta)$, respectively. Conditioned by $m' = k$, $n' = l$, the distribution of S' is that of the Wilcoxon statistic from samples of sizes k, l, under the null hypothesis; it is asymptotically normal ($\frac{1}{2}k(k+l+1)$, $\frac{1}{12}kl(k+l+1)$ for $k, l \to +\infty$. This fact can be written as

$$\lim_{k,l\to+\infty} \max_{a\leqslant t\leqslant b} \left| E\left(\exp\left(it\,\frac{S' - \frac{1}{2}m'(m'+n'+1)}{\left(\frac{1}{12}m'n'(m'+n'+1)\right)^{1/2}}\right) \middle| m' = k, n' = l\right) - e^{-t^2/2}\right| = 0, \tag{9}$$

for each real $a < b$.

Let us define

$$d^2 = \tfrac{1}{3}mnN(1-\Delta)^3,$$
$$T = (S - \tfrac{1}{2}\{m(m+1)+mn(1-\Delta)^2\})/d,$$
$$a(k,l) = \tfrac{1}{2}(kl - mn(1-\Delta)^2)/d,$$
$$s^2(k,l) = \tfrac{1}{12}kl(k+l+1)/d^2,$$

(all these quantities being dependent on ν). We shall verify that T, a, s satisfy the conditions of Lemma 1.

(A) According to (8),

$$\frac{T-a(m',n')}{s(m',n')} = \frac{S'-\frac{1}{2}m'(m'+n'+1)}{\left(\frac{1}{12}m'n'(m'+n'+1)\right)^{1/2}},$$

so that we can rewrite the assertion (9) as follows: To each $\varepsilon > 0$ and $a < b$, there exist K, L (independent of ν) such that

$$\max_{a\leqslant t\leqslant b}\left|E\left(\exp\left(it\frac{T-a(m',n')}{s(m',n')}\right)\Big| m'=k, n'=l\right) - e^{-t^2/2}\right| < \varepsilon$$

holds for all $k \geqslant K, l \geqslant L$.

Now, from $m(1-\Delta) \to +\infty$ (and from $m < n$) it easily follows that $m' \xrightarrow{P} +\infty$, $n' \xrightarrow{P} +\infty$, which means that $P(m' \geqslant K, n' \geqslant L) \to 1$. Consequently,

$$\lim_{\nu\to+\infty} P\left(\max_{a\leqslant t\leqslant b}\left|E\left(\exp\left(it\frac{T-a(m',n')}{s(m',n')}\right)\Big| m',n'\right) - e^{-t^2/2}\right| < \varepsilon\right) = 1.$$

(B) We have

$$m'n' - mn(1-\Delta)^2 = \big(m'-m(1-\Delta)\big)n(1-\Delta) + \big(n'-n(1-\Delta)\big)m(1-\Delta) + \\ + \big(m'-m(1-\Delta)\big)\big(n'-n(1-\Delta)\big) = U+V+W,$$

say. The assumptions $\Delta \to 1$ and $m(1-\Delta) \to +\infty$, together with the central limit theorem, entail that $U+V$ is asymptotically normal $\big(0, mnN(1-\Delta)^3\big)$; further,

$$EW^2 = mn\Delta^2(1-\Delta)^2.$$

Hence,

$$a(m',n') = (U+V+W)/2d$$

is asymptotically normal $(0, \frac{3}{4})$, considering that $E(W^2/4d^2) \to 0$.

(C) We have

$$s^2(m',n') = \frac{m'n'(m'+n'+1)}{4mnN(1-\Delta)^3} \xrightarrow{P} \frac{1}{4},$$

since the coefficients of variation of m' and n' tend to 0 and thus each of the sequences

$$\frac{m'}{m(1-\Delta)}, \quad \frac{n'}{n(1-\Delta)}, \quad \frac{m'+n'+1}{N(1-\Delta)}$$

tends to 1 in probability.

The random variables T, a, s satisfy thus the conditions of Lemma 1 with constants $\mu = 0$, $b^2 = \frac{3}{4}$, $\sigma^2 = \frac{1}{4}$. Consequently, T is asymptotically normal $(0, 1)$, which is equivalent to the assertion (i) [for $\Delta \to 1$].

(ii) Put $\lambda = m(1-\Delta)$ and denote by C its upper bound (which exists by assumption). Then apply Theorem 2 with $A = \Delta$; we get

$$P(X_i < A < X_j, 1 \leqslant i \leqslant m, m+1 \leqslant j \leqslant N) =$$

$$= \Delta^m = \left(1-\frac{\lambda}{m}\right)^m > (1-\varepsilon)e^{-C}$$

for each $\varepsilon > 0$ and large m. This completes the proof.

REMARK ADDED IN PROOF. The proof of Theorem 1 is based on the original version of [2]. From Theorem 2.1 of the published revised version, the theorem follows immediately.

References

[1] H. Chernoff and I. R. Savage, *Asymptotic normality and efficiency of certain nonparametric test statistics*, Ann. Math. Statist. 29 (1958), pp. 972-994.

[2] J. Hájek, *Asymptotic normality of simple linear rank statistics under alternatives*, Ann. Math. Statist. 39 (1968), pp. 325-346.

CHARLES UNIVERSITY, PRAGUE

Received on 29. 5. 1967