A Law of the Iterated Logarithm and Invariance Principle for Regression Rank Statistics

by

David M. Mason
Purdue University

Department of Statistics
Division of Mathematical Sciences
Mimeograph Series #522

January, 1978

A Law of the Iterated Logarithm and Invariance Principle for Regression Rank Statistics

by

David M. Mason
Purdue University

A law of the iterated logarithm and invariance principle for regression rank statistics is given. These results are an extension of analogous results of Sen and Ghosh (1972) with simplified proofs. An inequality of Hušková (1977) is also extended.

Key words: Rank statistics, law of the iterated logarithm, invariance principle.

1. Summary and Notation.

We will use more or less the notation of Sen and Ghosh (1972). Let $\{X_i;\ i \geq 1\}$ be a sequence of independent random variables defined on the same probability space $(\Omega,\mathcal{F},P)$ with common continuous distribution function F. Let $I(u) = 1$ or 0 according to whether $u \geq 0$ or $u < 0$. Define $R_{in} = \sum_{k=1}^{n} I(X_i-X_k)$ to be the rank of X_i among $X_1,\ldots,X_n$ for $i=1,\ldots,n$. Let J be a nondecreasing, absolutely continuous function inside (0,1) such that $\int_0^1 J^2(u)du < \infty$. Set $J_n(i/(n+1)) = EJ_n(U_n^{(i)})$ where $U_n^{(1)} \leq \ldots \leq U_n^{(n)}$ are the ordered values of n independent uniform (0,1) random variables. Without loss of generality we will assume $\int_0^1 J(u)du = 0$.

By a regression rank statistic we will mean a statistic of the form:

$$T_n = \sum_{i=1}^{n} (c_i-\bar{c}_n)J_n(i/(n+1)), \text{ where } c_1,\ldots,c_n \text{ are}$$

constants not all equal and $\bar{c}_n = \sum_{i=1}^{n} c_i/n$. Set $\sigma_n^2 = \text{Var } T_n$ and $C_n^2 = \sum_{i=1}^{n}(c_i-\bar{c}_n)^2$.

Denote by $y_i = (c_i-\bar{c}_{i-1})J_i(R_{ii}/(i+1))$ for $i \geq 1$, $M_n = \sum_{i=1}^{n} y_i$, $\phi_n = T_n - T_{n-1} - M_n + M_{n-1}$ and $s_n^2 = \text{Var } M_n$. We will set $T_0 = M_0 = \bar{c}_0 = 0$.

A law of the iterated logarithm and invariance principle will be proven for T_n. The method of proof will be to show that to obtain our results, M_n is a sufficient approximation to T_n. M_n is a sum of independent random variables (See Lemma A.3 Appendix). So law of the logarithm results for M_n should apply to T_n, if M_n is sufficiently close to T_n. The main tool will be an inequality (Proposition A.1 Appendix) that will allow us to obtain a sufficient approximation of closeness of M_n to T_n. The inequality is an extension of Lemma 2.5 Hušková (1977).

2. Main Theorems.

Law of the Iterated Logarithms

Theorem 1. (Unbounded Score Function)

If

1.i. $J'(u) \leq K(u(1-u))^{-3/2+\delta}$ for some $K > 0$ and $0 < \delta < 1/2$

1.ii. $\underline{\lim}_{n\to\infty} \; n^{-1}\sigma_n^2 > 0$

1.iii. $C_n^2 \leq nC$ for some constant $C > 0$ and all $n \geq 1$

1.iv. $|c_n - \bar{c}_{n-1}| n^{-1/2} = O([(\ell n\, n)^{-1}(\ell n \ell n\, n)^{-1-r}]^{1/\varepsilon})$ for some $r > 0$ and $\varepsilon > 0$ such that $(2+\varepsilon)(1/2-\delta) < 1$

then

1.v. $\overline{\lim}_{n\to\infty} \; T_n/\sqrt{2\sigma_n^2 \ell n \ell n\, \sigma_n^2} \overset{a.s.}{=} 1.$

Remark 1.1.

Theorem 1 is an improvement on Theorem 1.2 Sen and Ghosh (1972). Among other additional conditions, they require that $\max_{1\leq i\leq n} |c_i - \bar{c}_n| C_n^{-1} = O(n^{-1/2})$. 1.iv. is a relaxation of this condition. Sen and Ghosh's proof is essentially a verification of the conditions of a martingale law of the iterated logarithm of Strassen (1967). The proof of Theorem 1 utilizes an entirely different technique.

Remark 1.2. (Rate of Convergence to Normality)

Let $F_n(x) = P(T_n \leq \sigma_n x)$. The machinery developed in this paper to prove Theorem 1 can be used to obtain a rate of convergence to normality for T_n; that is, under the conditions of Theorem 1 with 1.iv. replaced by $|c_n - \bar{c}_{n-1}| = O(1)$, $\sup_x |F_n(x) - \Phi(x)| = O(n^{-2\delta/3})$. The proof of this is along the lines of Bergström and Puri (1977).

Theorem 1'. (Bounded Score Function)

If

1'.i. $\quad J'(u) \leq M$ for some $M > 0$.

1.ii, 1.iii.

and

1'.iv $\quad (c_n - \bar{c}_{n-1})C_n^{-1} = o((\ln\ln(C_n^2))^{-1/2})$

then 1.v. holds.

Remark 1'.1.

Theorem 1' can be proved by the verification of theconditions of the law of the iterated logarithm for martingales of Stout (1970). The proof given here though will be basically the same as the proof of Theorem 1 with a few modifications.

The proofs of Theorems 1 and 1' will be delayed until Section 3.

Invariance Principle

For each $n \geq 1$ let V_n and W_n be random functions on [0,1] defined as follows:

$$V_n(t) = s_n^{-1}[M_k + (M_{k+1}-M_k)(ts_n^2-s_k^2)/(s_{k+1}^2-s_k^2)]$$

and

$$W_n(t) = s_n^{-1}[T_k + (T_{k+1}-T_k)(ts_n^2-s_k^2)/(s_{k+1}^2-s_k^2)]$$

whenever $s_k^2 \leq ts_n^2 \leq s_{k+1}^2$ for $k = 0,\ldots,n-1$.

Theorem 2.

If 1.i or 1'.i, 1.ii, 1.iii, and $\max_{1\leq i\leq n} |c_i-\bar{c}_n|C_n^{-1} = o(1)$, then

$W_n \Rightarrow W$ where W is a standard Wiener process on $[0,1]$.

Remark 2.1.

See Theorem 1.2 Sen and Ghosh (1972) for an analogous invariance principle proven under the conditions described in Remark 1.1.

Proof of Theorem 2.

It is easy to show that under the conditions of Theorem 2 that the V_n process satisfies the conditions of Theorem 2.1 Prokhorov (1956) to give $V_n \Rightarrow W$.

Note
$$\sup_{0\leq t\leq 1} |V_n(t) - W_n(t)| \tag{2.1}$$

$$= \sup_{0\leq t\leq 1} s_n^{-1} \, |M_k-T_k + (M_{k+1}-M_k- T_{k+1}+T_k)\, (ts_n^2-s_k^2)/(s_{k+1}^2-s_k^2)\, |$$

$$\leq s_n^{-1} \max_{1\leq k\leq n} |M_k-T_k|.$$

Now since M_k-T_k, $k=0,\ldots,n$ is a martingale (See Lemmas A.2 and A.3 Appendix), for all $\varepsilon > 0$

$$P(\sup_{0\leq t\leq 1} |V_n(t)-W_n(t)| > \varepsilon) \leq \varepsilon^{-2}s_n^{-2}E(T_n-M_n)^2. \tag{2.2}$$

Under 1.i or 1.i', 1.ii and 1.iii, Corollary A.2 Appendix gives (2.2) = o(1). Hence (2.1) = $o_p(1)$, which implies that $W_n \Rightarrow W$.

3 Proofs of Theorems 1 and 1'.

Proof of Theorem 1.

Let $G_n(x) = P(M_n \leq s_n x)$ and $R_n = \sup_x |G_n(x) - \Phi(x)|$.

Theorem 6 page 115 Petrov (1975) coupled with the fact that M_n is a sum of independent random variables (see Lemma A.3 Appendix) gives

$$R_n = As_n^{-(2+\varepsilon)} \sum_{j=1}^{n} (c_j-\bar{c}_{j-1})^{2+\varepsilon} \sum_{i=1}^{j} J_j^{2+\varepsilon}(i/(j+1))/j \text{ for some}$$

constant A > 0, which is

$$\leq As_n^{-2} \sum_{j=1}^{n} (c_j-\bar{c}_{j-1})^2 \sum_{i=1}^{j} J_j^{2+\varepsilon}(i/(j+1))/j \max_{1\leq j\leq n} |c_j-\bar{c}_{j-1}|^{\varepsilon} s_n^{-\varepsilon} .$$

A.2.v. of Corollary A.2 Appendix , 1.ii., 1.iii., and 1.iv. imply that

$$\max_{1\leq j\leq n} |c_j-\bar{c}_{j-1}|^{\varepsilon} s_n^{-\varepsilon} = O((\ln n)^{-1}(\ln\ln n)^{-1-r}).$$

A.2.v. along with 1.i, 1.ii, 1.iii, and the assumption that $0 < (2+\varepsilon)(1/2-\delta) < 1$ imply that

$$s_n^{-2} \sum_{j=1}^{n} (c_j-\bar{c}_{j-1})^2 \sum_{i=1}^{j} J_j^{2+\varepsilon}(i/(j+1))/j = O(1).$$

Hence $R_n = O((\ln n)^{-1}(\ln\ln n)^{-1-r})$.

Thus by the remarks on page 305 Petrov (1975),

$\overline{\lim_{n\to\infty}} M_n/\sqrt{2s_n^2 \ln\ln s_n^2} \overset{a.s.}{=} 1$. A.2.v. also implies that

$$\overline{\lim_{n\to\infty}} M_n/\sqrt{2\sigma_n^2 \ln\ln \sigma_n^2} \overset{a.s.}{=} 1.$$

The following lemma will now give l.v.

Lemma 1.1.

$$\lim_{n\to\infty}(T_n-M_n)/\sqrt{2\sigma_n^2 \ln \ln \sigma_n^2} \overset{a.s.}{=} 0$$

Proof.

It is sufficient to show that

$$\lim_{n\to\infty} (T_n-M_n)/\sqrt{2s_n^2 \ln \ln s_n^2} \overset{a.s.}{=} 0.$$

Note that $\{(T_n-M_n)/\sqrt{2s_n^2 \ln \ln s_n^2}$ does not converge to zero$\} \subset \{(T_n-M_n)^2 > s_n^2$ i.o.$\}$, which since s_n^2 is a nondecreasing function of n is

$$\subset \bigcup_{j=1}^{\infty} \{ \max_{n_j<k\leq n_{j+1}} (T_k-M_k)^2 > s_{n_j}^2 \}. \quad \text{Where } n_j = (j+1)^{1/\delta}.$$

Now since $(T_n-M_n)^2$ is a submartingale (See Lemmas A.2 and A.3 Appendix), the maximal inequality gives

$$P(\max_{n_j<k\leq n_{j+1}} (T_k-M_k)^2 > s_{n_j}^2) < E(T_{n_{j+1}}-M_{n_{j+1}})^2/s_{n_j}^2.$$

By Corollary A.2 Appendix,

$$E(T_{n_{j+1}}-M_{n_{j+1}})^2/s_{n_{j+1}}^2 = O(n_{j+1}^{-2\delta}). \quad \text{Also note that}$$

$$s_{n_{j+1}}^2/s_{n_j}^2 = O(((j+1)/j)^{1/\delta}) = O(1).$$

Hence $P(\max_{n_j<k\leq n_{j+1}} (T_k-M_k)^2 > s_{n_j}^2) = O((j+1)^{-2})$.

Application of the Borel-Cantelli lemma completes the proof.□

Proof of Theorem 1'.

Theorem 1' is proven almost exactly as Theorem 1, except that Kolmogorov's Theorem (See Theorem 1 page 292 Petrov (1975)) is used to obtain the law of the iterated logarithm for M_n. This is where condition 1'.iv. comes into play. □

4 Appendix.

For fixed integers $k \geq 1$ and $1 \leq m \leq 2k$, let S be any set of indices $\{\ell_1,\ldots,\ell_m\}$ such that $1 \leq \ell_1 \leq \ldots \leq \ell_m$ are integers and $\sum_{i=1}^{m} \ell_i = 2k$.

Let $\mathscr{S}$ = the class of all such S for $k \geq 1$ fixed and $1 \leq m \leq 2k$. For $n \geq 1$, let $\mathscr{S}_n$ be the subclass of $\mathscr{S}$ where $1 \leq m \leq 2k \wedge n$.

Suppose $W_1,\ldots,W_n$ are random variables such that for each $S \in \mathscr{S}_n$ there exists an η_S such that

$$E(W_{i_1}^{\ell_1} \ldots W_{i_m}^{\ell_m}) = \eta_S \text{ for all permutations } i_1,\ldots,i_m \text{ of } 1,\ldots,n$$

taken m at a time.

With the above notation and assumptions we will now prove the following inequality.

Propositon A.1. (An Inequality)

For each $k \geq 1$, there exists a constant C(k) independent of $c_1,\ldots,c_n$ such that $E\phi_n^{2k} \leq C_n^{2k}\, EW_1^{2k}$, where $\phi_n = \sum_{i=1}^{n} (c_i-\bar{c}_n)W_i$.

Remark A.1.

Proposition A.1 is analogous to Lemma 2.5 Hušková (1977), but with less specified assumptions and a simplified proof.

Proof of Proposition A.1.

Observe that $E\phi_n^{2k} =$

$$\sum_{j_1+\ldots+j_n=2k} \binom{2k}{j_1,\ldots,j_n} (c_1-\bar{c}_n)^{j_1}\ldots(c_n-\bar{c}_n)^{j_n} E(W_1^{j_1}\ldots W_n^{j_n})$$

$$= \sum_{S\in\mathscr{S}_n} \binom{2k}{\ell_1,\ldots,\ell_m} \sum_{\substack{i_1,\ldots,i_m \\ \text{distinct}}} (c_{i_1}-\bar{c}_n)^{\ell_1}\ldots(c_{i_m}-\bar{c}_n)^{\ell_m} \eta_S .$$

The proof will now follow directly from:

Lemma A.1.

For all $S \in \mathscr{S}_n$,

$$\Big|\sum_{\substack{i_1,\ldots,i_m \\ \text{distinct}}} (c_{i_1}-\bar{c}_n)^{\ell_1}\ldots(c_{i_m}-\bar{c}_n)^{\ell_m}\Big| \le (2k)!C_n^{2k} \tag{A.1.1}$$

Proof.

Case 1. Suppose all the $\ell_1,\ldots,\ell_m$ are even integers. Then

$$\sum_{\substack{i_1,\ldots,i_m \\ \text{distinct}}} (c_{i_1}-\bar{c}_n)^{\ell_1}\ldots(c_{i_m}-\bar{c}_n)^{\ell_m} =$$

$$\sum_{\substack{i_1,\ldots,i_m \\ \text{distinct}}} (c_{i_1}-\bar{c}_n)^{2\ell_1^*}\ldots(c_{i_m}-\bar{c}_n)^{2\ell_m^*}, \tag{A.1.2}$$

where $\ell_i^* = \ell_i/2$ for $i=1,\ldots,m$, and the ℓ_i^* are integers.

Since $\sum_{i=1}^{m} \ell_i^* = k$, (A.1.2) is obviously less than C_n^{2k}.

Case 2. Suppose $\ell_1,\ldots,\ell_m$ consist of $2r$ odd integers ≥ 3 and $m-2r$ even integers where $r \ge 1$.

Let us relabel $\ell_1,\ldots,\ell_m$, so that $e_1,\ldots,e_{m-2r}$ are the even integers and $d_1,\ldots,d_{2r}$ are the odd integers.

We can now rewrite our sum as

$$\Big|\sum_{\substack{i_1,\ldots,i_m \\ \text{distinct}}} (c_{i_1}-\bar{c}_n)^{e_1} \ldots (c_{i_{m-2r}}-\bar{c}_n)^{e_{m-2r}}(c_{i_{m-2r+1}}-\bar{c}_n)^{d_1}\ldots c_{i_m}-\bar{c}_n)^{d_{2r}}\Big|.$$

But since each $|c_i-\bar{c}_n| \leq C_n$, the above is

$$\leq \Big|\sum_{\substack{i_1,\ldots,i_m \\ \text{distinct}}} (c_{i_1}-\bar{c}_n)^{e_1}\ldots(c_{i_{m-2r}}-\bar{c}_n)^{e_{m-2r}}(c_{i_{m-2r+1}}-\bar{c}_n)^{d_1-1}\ldots$$

$$(c_{i_m}-\bar{c}_n)^{d_{2r}-1}\Big|C_n^{2r}.$$

Observe that $e_1,\ldots,e_{m-2r},d_1-1,\ldots,d_{2r}-1$ are now all positive even integers which add up to $2k-2r$. Application of Case 1 gives us that the above is $\leq C_n^{2k}$.

<u>Case 3.</u> Suppose $\ell_1,\ldots,\ell_m$ consist of $\ell_1 = \ldots = \ell_p = 1$ and $\ell_{p+1} > 1,\ldots, \ell_m > 1$ where $1 \leq p \leq m$.

Assume first that $p = 1$.

Then $$\Big|\sum_{\substack{i_1,\ldots,i_m \\ \text{distinct}}} (c_{i_1}-\bar{c}_n)^{\ell_1}\ldots(c_{i_m}-\bar{c}_n)^{\ell_m}\Big| =$$

$$\Big|-\sum_{\substack{i_2,\ldots,i_m \\ \text{distinct}}} ((c_{i_2}-\bar{c}_n) + \ldots + (c_{i_m}-\bar{c}_n))(c_{i_2}-\bar{c}_n)^{\ell_2}\ldots(c_{i_m}-\bar{c}_n)^{\ell_m}\Big| \leq$$

$$\Big|\sum_{\substack{i_2,\ldots,i_m \\ \text{distinct}}} (c_{i_2}-\bar{c}_n)^{\ell_1+1}\ldots(c_{i_m}-\bar{c}_n)^{\ell_m}\Big| +\ldots+ \Big|\sum_{\substack{i_2,\ldots,i_m \\ \text{distinct}}} (c_{i_2}-\bar{c}_n)^{\ell_2}\ldots(c_{i_m}-\bar{c}_n)^{\ell_m+1}\Big|.$$

Which by Cases 1 and 2 is $\leq (m-1)C_n^{2k}$.

Proceeding inductively in the same manner as above, we get for all $1 \leq p \leq m$

$$\Big|\sum_{\substack{i_1,\ldots,i_m\\ \text{distinct}}} (c_{i_1}-\bar{c}_n)^{\ell_1}\cdots(c_{i_m}-\bar{c}_n)^{\ell_m}\Big| \leq (m-1)\ldots(m-p)C_n^{2k}.$$

Now by noting that $m \leq 2k$, we get (A.2.1). □

To complete the proof of Proposition A.1, application of Lemma A.2 gives:

$$E\phi_n^{2k} \leq \sum_{S \in \mathscr{S}_n} \binom{2k}{\ell_1,\ldots,\ell_m}(2k)!C_n^{2k}|\eta_S|$$

$$\leq ((2k)!)^2 \text{ Card}\,\mathscr{S}\,C_n^{2k} \max_{S \in \mathscr{S}} |\eta_S|.$$

Note that each $|\eta_S| \leq EW_1^{2k}$, also the card $\mathscr{S}$ depends only on k. Let $C(k) = ((2k)!)^2 \text{ card}\,\mathscr{S}$. □

Let $\mathscr{F}_n$ be the σ-field generated by $\mathscr{R}_n = (R_{1n},\ldots,R_{nn})$ for $n \geq 1$ and $\mathscr{F}_0 = \{\phi,\Omega\}$.

Lemma A.2.

$\{T_n,\mathscr{F}_n, n\geq 0\}$ is a martingale ($T_0 \equiv 0$)

Proof. See Lemma 2.1 Sen and Ghosh (1972).

Lemma A.3.

For each $n \geq 1$, $y_1,\ldots,y_n$ are independent random variables where $y_i = (c_i-\bar{c}_{i-1})J_i(R_{ii}/(i+1))$ for $i = 1,\ldots,n$.

Proof.

Suppose the lemma is true for some $n \geq 1$. Note that it is true for $n = 1$. We will show that it is true for $n + 1$. Pick any set of reals $\{t_1,\ldots,t_{n+1}\}$. Then

$$\text{(A.3.1)} \qquad E \exp(i \sum_{j=1}^{n+1} y_j t_j) = E(E(\exp(i \sum_{j=1}^{n+1} y_j t_j)|\mathscr{F}_n)) =$$

$$E(\exp(i \sum_{j=1}^{n} y_j t_j)E(\exp(i\, y_{n+1}t_{n+1})|\mathscr{F}_n)).$$

But $\qquad E(\exp(i\, y_{n+1}t_{n+1})|\mathscr{F}_n) \overset{a.s.}{=}$

$$(n+1)^{-1} \sum_{i=1}^{n+1} \exp(i\, t_{n+1}(c_{n+1}-\bar{c}_n)J_{n+1}(j/(n+2))) \qquad = E \exp(i\, y_{n+1}t_{n+1})$$

Hence (A.3.1) $= E \exp(i \sum_{j=1}^{n} y_j t_j) E \exp(i\, t_{n+1}y_{n+1})$,

which by the inductive hypothesis equals $\sum_{j=1}^{n+1} E \exp(i\, t_j y_j)$. □

Lemma A.4.

Suppose for some constants $K > 0$ and $0 < \delta < 1/2$

$$J'(u) \leq K(u(1-u))^{-3/2+\delta} \text{for all } u \in (0,1).$$

Then there exists a constant $K' > 0$ such that for all $n \geq 2$ and $1 \leq j \leq n-1$

$$J_n((j+1)/(n+1))-J_n(j/(n+1)) \leq K'(n+1)^{-1}[(j+1)(n+1-j)/(n+1)^2]^{-3/2+\delta}.$$

Proof.

Pick any $1 \leq j \leq n-1$, $n \geq 2$. Note that there exists a $K_1 > 0$ such that

$$K(u(1-u))^{-3/2+\delta} \leq K_1(u^{-3/2+\delta} + (1-u)^{-3/2+\delta}) \text{ for all } u \in (0,1).$$

Hence $J_n((j+1)/(n+1)) - J_n(j/(n+1)) = E \int_{U_n^{(j)}}^{U_n^{(j+1)}} J'(u)du.$

$$< K_1 E \int_{U_n^{(j)}}^{U_n^{(j+1)}} (u^{-3/2+\delta} + (1-u)^{-3/2+\delta})du =$$

$$K_1(1/2-\delta)^{-1} E[(1-U_n^{(j+1)})^{-1/2+\delta} - (1-U_n^{(j)})^{-1/2+\delta} - (U_n^{(j+1)})^{-1/2+\delta} + (U_n^{(j)})^{-1/2+\delta}] =$$

$$K_1(1/2-\delta)^{-1}[\prod_{i=n-j}^{n} i/(i+\delta-1/2) - \prod_{i=n+1-j}^{n} i/(i+\delta-1/2)$$

$$- \prod_{i=j+1}^{n} i/(i+\delta-1/2) + \prod_{i=j}^{n} i/(i+\delta-1/2)] =$$

$$K_1(1/2-\delta)^{-1}[((n-j)/(n-j+\delta-1/2)-1) \prod_{i=n+1-j}^{n} i/(i+\delta-1/2)$$

$$+ (j/(j+\delta-1/2)-1) \prod_{i=j+1}^{n} i/(i+\delta-1/2)]. \tag{A.4.1}$$

Now it is easy to show that there exists a $K_\delta > 0$ such that for all $1 \leq k \leq n-1$ and $n \geq 2$

$$\prod_{i=n+1-k}^{n} i/(i+\delta-1/2) \leq K_\delta((n+1-k)/(n+1))^{-1/2+\delta},$$ which implies that

(A.4.1) $\leq$

$$K_1 K_\delta[(1/(n-j+\delta-1/2))((n+1-j)/(n+1))^{-1/2+\delta} +$$

$$(1/(j+\delta-1/2))((j+1)/(n+1))^{-1/2+\delta}]$$

$$\le K_1K_\delta[(1/(n-j+\delta-1/2))((n+1-j)/(n+1))^{-1/2+\delta} +$$

$$(1/(j+\delta-1/2))(j/(n+1))^{-1/2+\delta}]$$

$$= \frac{K_1K_\delta}{n+1}[((n+1-j)/(n-j+\delta-1/2))((n+1-j)/(n+1))^{-3/2+\delta} +$$
$$(j/(j+\delta-1/2))(j/(n+1))^{-3/2+\delta}]. \tag{A.4.2}$$

Note that $k/(k+\delta-1/2) = 1/(1+(\delta-1/2)/k) \le (1/2+\delta)^{-1}$ for $1 \le k \le n$. Hence

$$\text{(A.4.2)} \quad \le \quad K_1K_\delta(1/2+\delta)^{-1}(n+1)^{-1}[(1-j/(n+1))^{-3/2+\delta} + (j/(n+1))^{-3/2+\delta}]$$

It is simple to verify that there exists a constant $K_2 > 0$ such that $(1-u)^{-3/2+\delta} + u^{-3/2+\delta} \le K_2(u(1-u))^{-3/2+\delta}$ for all $u \in (0,1)$. Now let $K' = K_1K_2K_\delta(1/2+\delta)^{-1}$. □

Lemma A.5.

Suppose for some constants $K > 0$ and $0 < \delta < 1/2$

$$J'(u) \le K(u(1-u))^{-3/2+\delta} \text{ for all } u \in (0,1).$$

Set $W_{1n} = J_n(R_{1n}/(n+1)) - J_{n-1}(R_{1n-1}/n)$.
Then for every integer $k \ge 1$, there exists a $D(k) > 0$ dependent only on J and k such that for all $n \ge 2$

$$EW_{1n}^{2k} \le D(k)(n+1)^{k-2\delta k-2} .$$

Proof.

Pick $k \ge 1$.

Note that $E(W_{1n}^{2k}|\mathscr{F}_{n-1}) =$

$$n^{-1}(n-R_{1n-1})[J_n(R_{1n-1}/(n+1))-J_{n-1}(R_{1n-1}/n)]^{2k} +$$

$$n^{-1}R_{1n-1}[J_n((R_{1n-1}+1)/(n+1))-J_{n-1}(R_{1n-1}/n)]^{2k}. \qquad \text{(A.5.1)}$$

Now by application of the identity: for $1 \leq i \leq n-1$

$$J_{n-1}(i/n) = n^{-1}(n-i)J_n(i/(n+1)) + n^{-1}iJ_n((i+1)/(n+1)),$$

we get (A.5.1) =

$$(1-R_{1n-1}/n)(R_{1n-1}/n)[(R_{1n-1}/n)^{2k-1} + (1-R_{1n-1}/n)^{2k-1}] \cdot$$
$$[J_n((R_{1n-1}+1)/(n+1))-J_n(R_{1n-1}/(n+1))]^{2k}$$

$$\leq (1-R_{1n-1}/n)(R_{1n-1}/n)[J_n((R_{1n-1}+1)/(n+1))-J_n(R_{1n-1}/(n+1))]^{2k}.$$

Thus $EW_{1n}^{2k} \leq$

$$(n-1)^{-1} \sum_{i=1}^{n-1} (1-i/n)(i/n)[J_n((i+1)/(n+1))-J_n(i/(n+1))]^{2k}. \qquad \text{(A.5.2)}$$

Now by Lemma A.4 there exists a $K' > 0$ such that for all $n \geq 2$,
(A.5.2) $\leq$

$$(K')^{2k}(n-1)^{-1} \sum_{i=1}^{n-1} (1-i/n)(i/n)(n+1)^{-2k}[(1-i/(n+1))(i/(n+1))]^{-3k+2\delta k}$$

which is $\leq$

$$K''(n+1)^{-2k} \sum_{i=1}^{n} [(1-i/(n+1))(i/(n+1))]^{-3k+2\delta k+1}/(n-1) \qquad \text{(A.5.3)}$$

for some $K'' > 0$.

But (A.5.3) is in turn $\leq$

$D(k)(n+1)^{k-2\delta k-2}$ for some $D(k) > 0$ for all $n \geq 2$. $\square$

Proposition A.2.

Suppose for some constants $K > 0$ and $0 < \delta < 1/2$

$$J'(u) \leq K(u(1-u))^{-3/2+\delta} \text{ for all } u \in (0,1).$$

Then for every integer $k > 0$, there exists a constant $A(k) > 0$ dependent only on J and k such that

$$E(M_n-T_n)^{2k} \leq n^{k-1}A(k) \sum_{j=2}^{n} C_{j-1}^{2k} (j+1)^{k-2\delta k-2}$$

Proof.

Set $\phi_j = T_j-T_{j-1}-(M_j-M_{j-1})$ for $j = 1,\ldots,n$.

Note that $\sum_{j=1}^{n} \phi_j = T_n-M_n$, $E\phi_j = 0$ for $j = 1,\ldots,n$ and by Lemmas A.2 and A.3 $\{T_j-M_j, \mathscr{F}_j, 1 \leq j \leq n\}$ is a martingale.

Direct application of the moment inequality for martingales of Dharmadikari, Fabian, and Jogdeo (1968) gives

$$E(T_n-M_n)^{2k} \leq n^{k-1}B(k) \sum_{j=1}^{n} E\phi_j^{2k}, \text{ where } B(k) > 0 \text{ is a constant}$$

dependent only on k.

Observe that for each $2 \leq j \leq n$, $\phi_j =$

$$\sum_{i=1}^{j} (c_i-\bar{c}_j)J_j(R_{ij}/(j+1)) - \sum_{i=1}^{j-1} (c_i-\bar{c}_{j-1})J_{j-1}(R_{ij-1}/j)$$

$$- (c_j-\bar{c}_{j-1})J_j(R_{jj}/(j+1)) =$$

$$\sum_{i=1}^{j-1} (c_i-\bar{c}_{j-1})W_{ij}, \text{ where } W_{ij} = J_j(R_{ij}/(j+1)) - J_{j-1}(R_{ij-1}/j).$$

Also observe that $\phi_1 = 0$.

Note that for each choice of integers $\ell_1,\ldots,\ell_m \geq 0, 1 \leq m \leq j-1$, $E(W_{i_1 j}^{\ell_1} \cdots W_{i_m j}^{\ell_m})$ is independent of all permutations $i_1,\ldots,i_m$ of $1,\ldots,j-1$ taken m at a time. This is enough to apply Proposition A.1.

Therefore

$E\phi_j^{2k} \leq C(k)C_{j-1}^{2k}EW_{1j}^{2k}$, where $C(k) > 0$ is a constant dependent only on k; and by Lemma A.5

$EW_{1j}^{2k} \leq D(k)(j+1)^{k-2\delta k-2}$ for some constant $D(k) > 0$ dependent only on J and k.

Now let $A(k) = B(k)C(k)D(k)$. □

Corollary A.2.

Suppose J satisfies the condition in Proposition A.2.

If

A.2.i. $\varliminf_{n\to\infty} n^{-1}\sigma_n^2 > 0$

and

A.2.ii. $C_n^2 \leq nC$ for some constant $C > 0$ and all $n \geq 1$

then

A.2.iii. $E(T_n-M_n)^2/\sigma_n^2 = O(n^{-2\delta})$

A.2.iv. $E(T_n-M_n)^2/s_n^2 = O(n^{-2\delta})$

A.2.v. $(s_n/\sigma_n-1)^2 = O(n^{-2\delta})$

and

A.2.vi. $(\sigma_n/s_n-1)^2 = O(n^{-2\delta})$

Proof.

Set k = 1 in Proposition A.2, then for $n \geq 2$

$$E(T_n-M_n)^2 \leq nA(1) \sum_{j=2}^{n} C_{j-1}^2 (j+1)^{-2\delta-1}/n, \text{ which by A.2.ii is}$$

$$\leq nA(1)C \sum_{j=2}^{n} (j+1)^{-2\delta}/n.$$

But $\sum_{j=2}^{n} (j+1)^{-2\delta}/n = O(n^{-2\delta})$ and by A.2.i

$\sigma_n^{-2} = O(n^{-1})$. Hence we have A.2.iii.

Note that $E(T_n-M_n)^2 \geq (s_n-\sigma_n)^2$. Thus by A.2.iii. we have A.2.v., from which we immediately get A.2.iv. and A.2.vi. □

References

Bergstrom, H. and Puri, M. L. (1977). Convergence and remainder terms in linear rank statistics. Ann. Statist. 5 671-680.

Breiman, L. (1968). Probability. Addison-Wesley, Reading, Massachusetts.

Dharmadhikari, S. W., Fabian, V. and Jogdeo, K. (1968). Bounds on the moments of martingales. Ann. Math. Statist. 39 1719-1723.

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

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

Petrov, V. V. (1975). Sums of Independent Random Variables. Springer-Verlag, New York.

Prokhorov, Y. V., (1956). Convergence of random processes and limit theorems in probability thoery. Theor. Probability Appl. Vol. 1 177-238.

Sen, P. K. and Ghosh, M. (1972). On strong convergence of regression rank statistics. Sankhya A, 34 335-348.

Stout, W. (1970). A martingale analogue of Kolmogorov's law of the iterated logarithm. Z. Wahr. and Verw. Gebiete 15 276-290.

Strassen, V. (1967). Almost sure behavior of sums of independent random variables and martingales. Proc. Fifth Berkeley Symp. Math. Statist. Prob. 2 315-343 University of California Press.

SMALL SAMPLE RESULTS FOR THE 27% RULE

George P. McCabe, Jr., Purdue University

I. MOTIVATING PROBLEM

In social science research the following type of problem is occasionally encountered. A fairly large collection of individuals, e.g. all students in an introductory psychology class, are measured on a variable, denoted by X. The observations are ordered and two groups are formed: one corresponding to low-X scores and the other corresponding to high-X. Sometimes the split is performed at the median; in other instances, the upper and lower thirds are used. In the latter case, no further observations are taken on the individuals in the middle third of the collection.

The designations low-X and high-X are used subsequently as a dichotomous variable, with the actual X-score being ignored. This dichotomous variable is then treated as a two-level factor in an experimental design. The simplest case of such a design, which will be the only one studied here in detail, is when a single additional variable, denoted by Y, is measured. The two sample t-test is then used to compare the performance of the low-X and high-X groups on Y.

The purpose of this investigation is not to defend the appropriateness of the above procedure. Rather, the properties of the procedure are examined and some useful information for determining efficient low-X and high-X groups is given.

II. MODEL ASSUMPTIONS AND NOTATION

Let (X_i, Y_i), $i = 1,\ldots,N$ be bivariate normal random variables. In what follows, it can be assumed without loss of generality, that the means are zero and the variances are one. Let $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(N)}$ denote the order statistics of the X variable. For any $n \leq N/2$, let

$$X_{1j} = X_{(j)}, \quad \text{for } j = 1,\ldots,n$$

and

$$X_{2j} = X_{(n-j+1)}, \text{ for } j = 1,\ldots,n$$

The Y observation paired with X_{ij} will be denoted by Y_{ij} for $i = 1,2$ and $j = 1,\ldots,n$.

Let α be the fraction of observations in each tail to be designated low-X and high-X. Thus, we let $\alpha \in (0.5]$ and define n by $n = [\alpha N]$, where $[\cdot]$ is the greatest integer function. The sample means and variances for the Y variable are

$$\bar{Y}_i(\alpha) = n^{-1} \textstyle\sum_{i=1}^{n} Y_{ij}, \quad i = 1,2$$

and

$$s_i^2(\alpha) = (n-1)^{-1} \textstyle\sum_{i=1}^{n} (Y_{ij} - \bar{Y}_i(\alpha))^2, \; i=1,2.$$

The t-statistic for comparing the Y means of the low-X and high-X groups is

$$t(\alpha) = \frac{\bar{Y}_1(\alpha) - \bar{Y}_2(\alpha)}{\sqrt{\dfrac{s_1^2(\alpha) + s_2^2(\alpha)}{n-1}}}.$$

III. LARGE SAMPLE RESULTS

In [1] it is shown that

$$\lim_{N\to\infty} N^{-1} t^2(\alpha) = \frac{2\rho^2(1-\Phi(a))}{M^2(a) + \rho^2(aM(a)-1)},$$

where ρ is the correlation between X and Y, $\Phi(\cdot)$ is the standard normal cumulative distribution function, a is defined by $\alpha = 1-\Phi(a)$ and $M(\cdot)$ is the Mills ratio, $M(x) = (1-\Phi(x))/\varphi(x)$, with $\varphi(\cdot)$ being the standard normal density function.

For a given value of ρ, the value of α which maximizes the above expression can be found. Details are given in [1]. As $|\rho^2|$ approaches zero, the optimal α approaches 27% from below. For $|\rho^2| = .5$ the optimal value is about 24% and falls off to about 20% for $|\rho^2| = .8$. For $|\rho^2| = .95$, $\alpha = 16\%$.

In summary, the large sample results indicate that a choice of $\alpha = 25\%$ is effective for a reasonable range of values of ρ that one might expect to encounter in practice.

IV. SMALL SAMPLE RESULTS

Of course, choosing α which maximizes the limiting value of $N^{-1}t^2(\alpha)$ is not equivalent to finding the α which maximizes the power of the t-test for detecting nonzero values of ρ. In addition, the relevance of the asymptotic calculations for reasonable size samples must be examined.

To address these questions, several simulations were run. Values of N chosen were 10, 20, 30, 40, 50, 60 and 100. For the first four values of N, 10,000 simulations were run; for the next two, 5,000 and for the last 4,000. The following values of $|\rho^2|$ were used: 0, .05, .1, .2, .3, .4, .5, .6, .7, .8, .9, .99. The five percent and one percent powers were estimated for all possible values of α.

In the simulations, the same generated random variables were used for all values of $|\rho^2|$ by considering the appropriate conditional distributions. The normal random numbers were generated using the routine described in [2].

Inspection of the results of the simulations reveals that the large sample results are applicable for practical values of N, i.e. it is reasonable to choose $\alpha = 25\%$ for cases where $|\rho^2|$ is expected to be small or moderate and slightly lower values of α when $|\rho^2|$ is expected to be large.

A very interesting fact revealed by the simulations is that the power as a function of α is very flat. The difference in performance between the optimal α and close values is often negligible from a practical point of view. This observation led to the construction of Tables 1 and 2. In Table 1 values of α^* are given for all values of $|\rho^2|$ and N considered. The quantity α^* is defined to be the smallest α giving power that is not less than .01 less than the power of the optimal α, where power is the power of the t-test using a type I error of 5%. Table 2 gives the powers for these α^*. Results for a Type I error of 1% are qualitatively similar.

Reprinted from the 1978 Social Statistics Section Proceedings of the American Statistical Association.

Since observations on the variable Y are taken on $2n = 2[\alpha N]$ cases, substantial savings can result by choosing α as small as possible while still retaining good power. As can be seen from Table 2, it is often possible to have excellent power with very few observations. For example, with $N = 50$ and $|\rho^2| = .5$, one only needs $\alpha = 8\%$, i.e. $n = 4$ to get 99% power (rounded to 2 places.)

REFERENCES

[1] McCabe, George P. Jr. (1977) Use of the 27% rule in experimental design. Purdue University Department of Statistics Mimeo Series No. 499.

[2] Rubin, Herman (1974) RVP-Random variable package. Purdue University Computing Center.

Table 1

Values of α^*

		N						
$\alpha^*(\%)$		10	20	30	40	50	60	100
	.05	30	25	23.3	22.5	24	23.3	27
	.10	30	25	26.7	25	24	25	22
	.20	30	30	23.3	25	22	21.7	13
	.30	30	30	23.3	22.5	20	15	7
	.40	40	30	23.3	15	12	10	4
$\lvert\rho^2\rvert$	.50	40	25	20	12.5	8	6.7	3
	.60	40	25	20	10	8	5	3
	.70	40	20	10	7.5	6	5	3
	.80	40	15	10	7.5	6	3.3	2
	.90	30	15	6.7	5	4	3.3	2
	.99	30	10	6.7	5	4	3.3	2

Table 2

Power of the 5% t-test for $\alpha = \alpha^*$

		N						
Power(%)		10	20	30	40	50	60	100
	.05	13	20	27	33	40	45	66
	.10	19	32	44	55	64	71	89
	.20	28	54	70	82	90	94	99
	.30	39	72	87	95	98	99	99
$\lvert\rho^2\rvert$	.40	51	85	95	98	99	100	99
	.50	62	93	99	99	99	100	99
	.60	73	98	100	100	100	100	100
	.70	83	99	100	100	100	100	100
	.80	92	99	100	100	100	100	100
	.90	98	100	100	100	100	100	100
	.99	100	100	100	100	100	100	100