ON SOME PROPERTIES OF A CLASS OF SPEARMAN RANK STATISTICS WITH APPLICATIONS

MALAY GHOSH

(Received Oct. 29, 1971)

Summary

The object of the present investigation is to study some properties of a class of Spearman rank statistics and to apply these results in studying the properties of a sequential procedure proposed in Section 3. The problem is one of bounded length confidence intervals for simple regression coefficients in linear models where both variables are subject to error. It is shown that the proposed procedure is asymptotically 'consistent' and 'efficient' in the sense of Chow and Robbins [3].

1. Introduction

Let $\{X_i,\ i=1, 2, \cdots\}$ and $\{Y_i,\ i=1, 2, \cdots\}$ be two independent sequences of independent and identically distributed random variables (iidrv). Define the sequences $\{Z_i(b)=Y_i-bX_i,\ i=1, 2, \cdots\}$ of iidrv for all real b. Let R_{ni} be the rank of X_i among $X_1, \cdots, X_n$ $(1\leqq i\leqq n;\ n\geqq 1)$, i.e. $R_{ni}=\frac{1}{2}+\sum_{j=1}^{n} u(X_i-X_j)$, $1\leqq i\leqq n$, where $u(t)=1$, $1/2$ or 0 according as $t>$, $=$, or <0. Similarly let $R'_{ni}(b)=\frac{1}{2}+\sum_{j=1}^{n} u(Z_i(b)-Z_j(b))$, $1\leqq i\leqq n$, $n\geqq 1$. Then the Spearman rank correlation coefficients between the first n X's and $Z(b)$'s $(n\geqq 1)$ are defined by

$$S_n(b)=\frac{12}{n(n^2-1)}\sum_{i=1}^{n}\left(R_{ni}-\frac{n+1}{2}\right)\left(R'_{ni}(b)-\frac{n+1}{2}\right), \tag{1.1}$$

$$n\geqq 1,\ b \text{ real}.$$

In Section 2, we shall study some properties of the class of statistics $S_n(b)$. In Section 3, we start with a linear regression model where both variables are subject to error. The problem is to find a bounded length confidence interval for the regression coefficient with prescribed confidence coefficient. The results of Section 2 are utilized there in studying the asymptotic properties of the proposed procedure.

2. The properties of $S_n(b)$

Let $\mathcal{F}_n$ denote the σ-field generated by $R_{n1}, \cdots, R_{nn}$ and $\mathcal{G}_n(b)$ the one generated by $R'_{n1}(b), \cdots, R'_{nn}(b)$. Let $\mathcal{H}_n(b)=\mathcal{F}_n \cap \mathcal{G}_n(b)$. Clearly $\{\mathcal{F}_n, n\geqq 1\}$, $\{\mathcal{G}_n(b), n\geqq 1\}$ and consequently $\{\mathcal{H}_n(b), n\geqq 1\}$ are non-decreasing sequences. Further, from our set up, $\mathcal{F}_n$ is independent of $\mathcal{G}_n(0)$ for all $n\geqq 1$. Define now

$$(2.1)\qquad T_n(b)=\frac{n(n^2-1)}{12(n+1)^2}S_n(b)=\sum_{i=1}^{n}\left(\frac{R_{ni}}{n+1}-\frac{1}{2}\right)\left(\frac{R'_{ni}(b)}{n+1}-\frac{1}{2}\right), \qquad n\geqq 1,\ b \text{ real}.$$

We first prove the following theorem.

THEOREM 2.1. *$\{(T_n(0), \mathcal{H}_n(0)), n\geqq 1\}$ forms a martingale sequence.*

PROOF. It is sufficient to show that $\mathrm{E}[T_{n+1}|\mathcal{H}_n(0)]=T_n$ (since $\mathcal{H}_n(0)$ is $\uparrow$ in n). But

$$\begin{aligned}
&\mathrm{E}[T_{n+1}(0)|\mathcal{H}_n(0)]\\
&=\sum_{i=1}^{n+1}\mathrm{E}\left[\frac{R_{n+1,i}R'_{n+1,i}(0)}{(n+2)^2}\,\middle|\,\mathcal{H}_n(0)\right]-\frac{n+1}{4}\\
&=\mathrm{E}\left[\frac{R_{n+1,n+1}R'_{n+1,n+1}(0)}{(n+2)^2}\right]+\sum_{i=1}^{n}\mathrm{E}\left[\frac{R_{n+1,i}R'_{n+1,i}(0)}{(n+2)^2}\,\middle|\,\mathcal{H}_n(0)\right]-\frac{n+1}{4}\\
&=\mathrm{E}\left[\frac{R_{n+1,n+1}}{n+2}\right]\mathrm{E}\left[\frac{R'_{n+1,n+1}(0)}{n+2}\right]+\sum_{i=1}^{n}\mathrm{E}\left(\frac{R_{n+1,i}}{n+2}\,\middle|\,\mathcal{F}_n\right)\\
&\quad\cdot\mathrm{E}\left(\frac{R'_{n+1,i}(0)}{n+2}\,\middle|\,\mathcal{G}_n(0)\right)-\frac{n+1}{4}\\
&=\frac{1}{2}\cdot\frac{1}{2}+\sum_{i=1}^{n}\frac{R_{ni}(1-R_{ni}/(n+1))+(R_{ni}+1)\cdot R_{ni}/(n+1)}{n+2}\\
&\quad\cdot\frac{R'_{ni}(0)(1-R'_{ni}(0)/(n+1))+(R'_{ni}(0)+1)\cdot R'_{ni}(0)/(n+1)}{n+2}-\frac{n+1}{4}\\
&=\sum_{i=1}^{n}\frac{R_{ni}R'_{ni}(0)}{(n+1)^2}-\frac{n}{4}\\
&=T_n(0)\,.
\end{aligned}$$

The above result is now utilized in finding the asymptotic distribution of $T_n(0)$ where the sample size N is random. More precisely, the following theorem is proved.

THEOREM 2.2. *Let $\{N_r\}$ be a sequence of random variables (can be viewed as stopping variables) assuming values $1, 2, \cdots$ such that $N_r\to\infty$ a.s. as $r\to\infty$. Also let $\{n_r\}$ be a sequence of positive integers $\to\infty$ as $r\to\infty$. If*

(2.2) $\frac{N_r}{n_r} \to 1$ *in probability as* $r \to \infty$,

then,

(2.3) $N_r^{1/2} S_{N_r}(0)$ *is asymptotically normal* $(0, 1)$.

PROOF. It is well-known (see e.g. Kendall [8] or Hoeffding [5]) that

(2.4) $\sqrt{n}\, S_n(0)$ is asymptotically (as $n \to \infty$) normal $(0, 1)$.

In view of (2.2), (2.4) and Theorem 1 of Anscombe [1], it is now sufficient to prove that the sequence $\{S_n(0)\}$ is uniformly continuous in probability with respect to $\{n^{-1/2}\}$ i.e. for every positive ε and η, there exist a δ (>0) and a positive integer n_0 such that for all $n \geqq n_0$, and n' assuming integer values,

$$\text{(2.5)} \qquad \mathrm{P}\{\sup_{|n'-n|<\delta n} |n^{1/2}(S_{n'}(0)-S_n(0))| > \eta\} < \varepsilon .$$

To prove this, first note that

$$\text{(2.6)} \qquad \sqrt{n}\, S_n(b) = n^{-1/2} T_n(b) + O(n^{-1/2}) , \qquad \text{for all real } b .$$

Hence, for all n' satisfying $|n'-n| < \delta n$,

$$\text{(2.7)} \qquad \sqrt{n}\,(S_{n'}(0)-S_n(0)) = n^{-1/2}(T_{n'}(0)-T_n(0)) + O(n^{-1/2}) .$$

Thus, all we need prove is that for every positive ε and η, there exist a δ (>0) and an n_0 such that for all $n \geqq n_0$, and n' assuming integer values

$$\text{(2.8)} \qquad \mathrm{P}\{\sup_{|n'-n|<\delta n} |n^{-1/2}(T_{n'}(0)-T_n(0))| > \eta\} < \varepsilon .$$

We prove the situation when $0 < n'-n < \delta n$ as the case when $n-\delta n < n' < n$ can be tackled similarly.

From Theorem 2.1 and the Kolmogorov inequality for martingales (cf. Loève [7], p. 386), it follows

$$\begin{aligned} \text{(2.9)} \quad & \mathrm{P}\{\sup_{n<n'<n+\delta n} |T_{n'}(0)-T_n(0)| > \eta n^{1/2}\} \\ & \leqq \mathrm{P}\{\max_{1 \leqq k \leqq [\delta n]} |T_{n+k}(0)-T_n(0)| > \eta n^{1/2}\} \\ & \qquad ([\delta n] \text{ being the largest integer contained in } \delta n) \\ & \leqq (\eta n)^{-1} \mathrm{E}\,[T^2_{n+[\delta n]}(0)-T_n^2(0)] . \end{aligned}$$

But

$$\mathrm{E}\, T_n^2(0) = \frac{n^2(n-1)^2}{(n+1)^2}\, \mathrm{E}\, S_n^2(0) = \frac{n^2(n-1)}{(n+1)^2} = n + O(1) .$$

Hence, $\mathrm{E}\, T^2_{n+[\delta n]}(0)=n+[\delta n]+O(1)$, and the right-hand side of (2.9) is bounded above by $(\delta/\eta)+O(n^{-1})$. Thus for any positive δ and η, $\delta\ (>0)$ and n_0 can be so chosen that (2.8) is satisfied. This proves the theorem.

The above result is comparable to an analogous result of Koul [9] concerning the asymptotic normality of random rank statistics. Koul's statistics are different from ours as is his method of proof. The present author, however, feels that even in his (Koul's) case, an analogous (to Theorem 2.2) and simpler proof can be constructed using the martingale inequalities and uniform continuity in probability.

The next theorem proves the asymptotic almost sure (a.s.) linearity of the class of statistics $S_n(b)$. Roughly it means that for a suitable range of b, $\sqrt{n}(S_n(b)-S_n(0))$ is asymptotically a linear function of b a.s. The result will be utilized in the following section for constructing a bounded length confidence interval for the regression coefficient in the problem mentioned earlier. The result, however, has some independent interest in as much as it proves an important asymptotic property of the Spearman rank statistics. Similar findings for other rank statistics are available in Ghosh and Sen ([4], [10]).

Before proving the theorem, the following representation of $S_n(b)$ is needed. We can write (see e.g. Hoeffding [5])

$$S_n(b)=\frac{3}{n+1}W_n(b)+\frac{n-2}{n+1}U_n(b)\,, \qquad \text{for all real } b\,, \tag{2.10}$$

where

$$W_n(b)=\frac{1}{n(n-1)}\sum\sum_{1\le i\ne j\le n}\operatorname{sgn}(X_i-X_j)\operatorname{sgn}(Z_i(b)-Z_j(b))\,,$$

$$U_n(b)=\frac{1}{n(n-1)(n-2)}\sum\sum\sum_{1\le i\ne j\ne l\le n}3\operatorname{sgn}(X_i-X_j)\operatorname{sgn}(Z_i(b)-Z_l(b))\,.$$

Then, for all real b,

$$\sqrt{n}\,S_n(b)=\sqrt{n}\,U_n(b)+O(n^{-1/2})\,, \qquad (\text{since } |U_n(b)|\le 1,\ |W_n(b)|\le 1)\,. \tag{2.11}$$

Also, $\mathrm{E}\, U_n(b)=3[4G(b)-1]$, where $G(b)=\mathrm{P}\{X_1>X_2,\ Z_1(b)>Z_3(b)\}$. We assume

(A) $G'(b)$, $G''(b)$ exist and are bounded for all real b. Then, from the Taylor expansion, $G(b)=G(0)+bG'(0)+(b^2/2)G''(\theta b)$ $0<\theta<1$. Now, $G(0)=1/4$. Hence, for, $|b|\le Cn^{-1/2}\log n$,

$$\sqrt{n}\ \mathrm{E}\, U_n(b)=12\sqrt{n}\,bG'(0)+O(n^{-1/2}(\log n)^2)\,. \tag{2.12}$$

The following theorem is now proved.

THEOREM 2.3. *Under assumption* (A), *for every* $\delta>0$, *there exist* K_1, K_2, δ_1 $(<1/4)$ *and* n_0 (*all depending on* δ) *such that for* $n\geqq n_0$,

$$(2.13)\qquad P\{\sup_{|b|\leqq Cn^{-1/2}\log n}\sqrt{n}\,|U_n(b)-U_n(0)-E\,U_n(b)|\geqq K_1 n^{-\delta_1}\log n\}\leqq K_2 n^{-1-\delta}.$$

Before proving the theorem, we study some of its implications. An immediate use of Borel-Cantelli lemma gives from (2.13),

$$\sup_{|b|\leqq Cn^{-1/2}\log n}\sqrt{n}\,|U_n(b)-U_n(0)-E\,U_n(b)|=o(1)\ \text{a.s.}\qquad \text{as } n\to\infty.$$

Since $E\,S_n(0)=0$, one gets from (2.11),

$$(2.14)\qquad \sup_{|b|\leqq Cn^{-1/2}\log n}\sqrt{n}\,[S_n(b)-S_n(0)-E\,\{S_n(b)-S_n(0)\}]=o(1)\ \text{a.s.}\quad \text{as } n\to\infty.$$

This is the so-called asymptotic linearity property of $S_n(b)$. By virtue of (2.12), one may also write

$$(2.15)\qquad \sup_{|b|\leqq Cn^{-1/2}\log n}\sqrt{n}\,[S_n(b)-S_n(0)-12bG'(0)]=o(1)\ \text{a.s.}\qquad \text{as } n\to\infty.$$

Since, we know that $\sqrt{n}S_n(0)$ is asymptotically normal (0, 1), for any real b such that $|b|\leqq Cn^{-1/2}\log n$, $\sqrt{n}\,[S_n(b)-12bG'(0)]$ is asymptotically normal (0, 1). This gives the non-null asymptotic distribution of the Spearman rank statistics.

PROOF OF THEOREM 2.3. One can rewrite $U_n(b)$ as

$$(2.16)\qquad U_n(b)=\binom{n}{3}^{-1}\sum\sum\sum_{1\leqq i\leqq j<l\leqq n}\phi_b(X_i, X_j, X_l, Y_i, Y_j, Y_l)$$

where $\phi_b(X_i, X_j, X_l, Y_i, Y_j, Y_l)$ reduces on simplification to $6\,\mathrm{sgn}\,(Z_r(b)-Z_s(b))=6\,\mathrm{sgn}\,(Y_r-Y_s-b(X_r-X_s))$, $(X_r=\max(X_i, X_j, X_l)$, $X_s=\min(X_i, X_j, X_l))$. Thus $U_n(b)$ is $\downarrow$ in b. Put now $m_n=[n^{\delta_1}]$, the largest integer contained in n^{δ_1}, $\eta_{r,n}=(r/m_n)(Cn^{-1/2}\log n)$, $r=-m_n,\cdots,-1, 0, 1,\cdots, m_n$. Then, for $\eta_{r-1,n}\leqq b\leqq\eta_{r,n}$, $(r=-m_n+1,\cdots, m_n)$, $U_n(\eta_{r,n})-U_n(0)-E\,U_n(\eta_{r-1,n})\leqq U_n(b)-U_n(0)-E\,U_n(b)\leqq U_n(\eta_{r-1,n})-U_n(0)-E\,U_n(\eta_{r,n})$. Hence, $|U_n(b)-U_n(0)-E\,U_n(b)|\leqq\max_{j=r-1,r}|U_n(\eta_{j,n})-U_n(0)-E\,U_n(\eta_{j,n})|+E\,[U_n(\eta_{r-1,n})-U_n(\eta_{r,n})]$. But from (2.12) it follows that $E\,[U_n(\eta_{r-1,n})-U_n(\eta_{r,n})]=-(12Cn^{-1/2}\log n/m_n)G'(0)+O(n^{-1}(\log n)^2)$. Hence,

$$(2.17)\qquad \sup_{|b|\leqq Cn^{-1/2}\log n}|U_n(b)-U_n(0)-E\,U_n(b)|\leqq\max_{j=-m_n,\cdots,-1,0,1,\cdots,m_n}|U_n(\eta_{j,n})-U_n(0)-E\,U_n(\eta_{j,n})|+O(n^{-1/2-\delta_1}\log n).$$

But

$$(2.18)\qquad P\{\max_{j=-m_n,\cdots,-1,0,1,\cdots,m_n}|U_n(\eta_{j,n})-U_n(0)-E\,U_n(\eta_{j,n})|>t_n\}$$
$$\leqq\sum_{j=-m_n}^{m_n}P\{|U_n(\eta_{j,n})-U_n(0)-E\,U_n(\eta_{j,n})|>t_n\}\ .$$

Now for each $j=-m_n,\cdots,-1,1,\cdots,m_n$, $U_n(\eta_{j,n})-U_n(0)-E\,U_n(\eta_{j,n})$ is a U-statistic (with a kernel of degree 3) minus its expectation, while for $j=0$, the above expression vanishes. Since $(X_i, Z_i(\eta_{j,n}))$, $i=1,2,\cdots$, n are n iid pairs of rv's, using Hoeffding's [6] result on exponential deviations of U-statistics, one gets,

$$(2.19)\qquad P\{U_n(\eta_{j,n})-U_n(0)-E\,U_n(\eta_{j,n})>t_n\}$$
$$\leqq\left(\left[\frac{\mu_{j,n}}{\mu_{j,n}+t_n}\right]^{\mu_{j,n}+t_n}\left[\frac{1-\mu_{j,n}}{1-\mu_{j,n}-t_n}\right]^{1-\mu_{j,n}-t_n}\right)^{[n/3]}$$

for $0<t_n<1-\mu_{j,n}$, $\mu_{j,n}=E\,U_n(\eta_{j,n})$, $j=-m_n,\cdots,-1,1,\cdots,m_n$. The logarithm of the right-hand side of (2.19) can be expressed as

$$(2.20)\qquad -\left[\frac{n}{3}\right]\left[(\mu_{j,n}+t_n)\log\left(1+\frac{t_n}{\mu_{j,n}}\right)+(1-\mu_{j,n}-t_n)\log\left(1-\frac{t_n}{1-\mu_{j,n}}\right)\right]$$

for $0<t_n<1-\mu_{j,n}$. But from (2.12), $\mu_{j,n}=12\eta_{j,n}G'(0)+O(n^{-1}(\log n)^2)$. Thus, $\mu_{j,n}=O(n^{-1/2}\log n)$ uniformly in j. Hence, putting $t_n=K_1n^{-1/2-\delta_1}\log n$, we see that (2.19) is valid for large n uniformly in j. Also if C is so chosen that $|K_1/CG'(0)|<1$, then $t_n<|\mu_{j,n}|$ for large n uniformly in j, and the logarithmic expansion in (2.20) is valid. (2.20) reduces now to

$$-\left[\frac{n}{3}\right]\left\{\frac{t_n^2}{2}\left(\frac{1}{1-\mu_{j,n}}+\frac{1}{\mu_{j,n}}\right)+\frac{t_n^3}{6}\left(\frac{1}{(1-\mu_{j,n})^2}-\frac{1}{\mu_{j,n}^2}\right)+\frac{t_n^4}{12}\left(\frac{1}{(1-\mu_{j,n})^3}+\frac{1}{\mu_{j,n}^3}\right)\right.$$
$$\left.+\cdots\right\}=-K_1'n^{1/2-2\delta_1}\log n+O(n^{1/2-3\delta_1}\log n),$$

where K_1' is some constant >0. It is easily seen that a similar inequality holds for $P\{U_n(b)-U_n(0)-E\,U_n(b)<-t_n\}$. It follows now from (2.18)-(2.20) that

$$(2.21)\qquad P\{\max_{j=-m_n,\cdots,-1,1,\cdots,m_n}|U_n(\eta_{j,n})-U_n(0)-E\,U_n(\eta_{j,n})|>K_1n^{-1/2-\delta_1}\log n\}$$
$$\leqq 2m_n 2\exp(-K_1'n^{1/2-2\delta_1})\log n+O(n^{1/2-3\delta_1}\log n)$$
$$\leqq 4n^{\delta_1}\exp(-K_1'n^{1/2-2\delta_1})\log n(1+O(n^{-\delta_1}\log n))\ .$$

This shows that for every $\delta_2>0$, K_1, K_2, δ_1 and n_0 can be so chosen that (2.13) is satisfied. Hence, the theorem.

3. Construction of bounded length confidence intervals

Applications will be made of the above results in the following problem:

Suppose we have a sequence $\{(Y_i, X_i),\ i=1, 2, \cdots\}$ of pairs of observations on a random vector (Y, X) satisfying the linear regression model $Y=\alpha+\beta X+e$, X being distributed independently of e. Our problem is to find a confidence interval $I_n=(\hat{\beta}_{L,n}, \hat{\beta}_{U,n})$ for β (based on a sample of size n) such that $P\{\beta \in I_n\}=1-\varepsilon$ and $0<\hat{\beta}_{U,n}-\hat{\beta}_{L,n}\leqq 2d$. If the joint distribution of (Y, X) say $F(y, x)$ were known, the classical least square approach for such a problem would be to consider the least square estimator

$$\beta_n^*=\sum_{i=1}^{n}(X_i-\bar{X}_n)(Y_i-\bar{Y}_n)\Big/\sum_{i=1}^{n}(X_i-\bar{X}_n)^2$$

$$\left(\bar{X}_n=n^{-1}\sum_{i=1}^{n}X_i,\ \ \bar{Y}_n=n^{-1}\sum_{i=1}^{n}Y_i\right)$$

for β based on a sample of size n and consider the interval $(\beta_n^*-d, \beta_n^*+d)$. The sample size n is so determined that $P\{\beta_n^*-d\leqq\beta\leqq\beta_n^*+d\}=1-\varepsilon$.

Now β_n^* can be interpreted as the solution for b on equating the product moment correlation coefficient between X_i and $Z_i(b)=Y_i-bX_i$ $(i=1, 2, \cdots, n)$ to zero. A natural question would be whether similar estimator can be obtained by equating the Spearman correlation coefficient $S_n(b)$ (defined in (1.1)) to zero. Fortunately for us, the answer is in the affirmative. To see this, first observe that $S_n(b)$ is $\downarrow$ in b. This follows from the representation (2.10) if each $W_n(b)$ and $U_n(b)$ is $\downarrow$ in b. We have already shown that $U_n(b)$ is $\downarrow$ in b, while for $W_n(b)$ the same is obviously true. Then defining $\hat{\beta}_{n1}=\sup\{b: S_n(b)>0\}$, $\hat{\beta}_{n2}=\inf\{b: S_n(b)<0\}$, $\hat{\beta}_n=(\hat{\beta}_{n1}+\hat{\beta}_{n2})/2$ can be defined as the point estimator of β. Roughly this is the solution for b by equating the Spearman statistic to zero. So a procedure would be to take $(\hat{\beta}_n-d, \hat{\beta}_n+d)$ as the proposed confidence interval and determine n eventually. A more direct approach in finding the confidence interval would be as follows:

$S_n(b)$ is $\downarrow$ in b and $S_n(\beta)$ has a known distribution symmetric about zero. Hence, there exist ε_n and $S_{n,\varepsilon}$ such that

$$(3.1)\qquad P_\beta\{-S_{n,\varepsilon}<S_n(\beta)<S_{n,\varepsilon}\}=1-\varepsilon_n \qquad (\to 1-\varepsilon \text{ as } n\to\infty)\,.$$

Define

$$(3.2)\qquad \hat{\beta}_{U,n}=\inf\{b: S_n(b)\leqq -S_{n,\varepsilon}\} \qquad \hat{\beta}_{L,n}=\sup\{b: S_n(b)\geqq S_{n,\varepsilon}\}\,.$$

Then $P_\beta\{\hat{\beta}_{L,n}<\beta<\hat{\beta}_{U,n}\}=P_\beta\{-S_{n,\varepsilon}<S_n(\beta)<S_n(\varepsilon)\}=1-\varepsilon_n$. (The distribution of $S_n(\beta)$ being discrete, it is not possible to achieve $P_\beta\{\hat{\beta}_{L,n}<\beta<\hat{\beta}_{U,n}\}=1-\varepsilon$ for all finite n.) The above procedure is not the same as other procedure of taking the confidence interval as $(\hat{\beta}_n-d, \hat{\beta}_n+d)$ for

finite samples. But both agree for large samples (see the comments after (3.9)).

If F is unknown, no fixed sample size procedure for our problem seems valid for all F. A sequential procedure for this problem is proposed in the same fashion as of Chow and Robbins [3]. We define the stopping variable $N=N(d)$ to be the first positive integer $m\geqq 3$ for which $0<\hat{\beta}_{U,m}-\hat{\beta}_{L,m}\leqq 2d$. The proposed confidence interval for β is

$$I_{N(d)}=(\hat{\beta}_{L,N(d)},\ \hat{\beta}_{U,N(d)})\ . \tag{3.3}$$

Some properties (mostly asymptotic i.e. as $d\to 0$) of the above procedure are provided in the following theorem.

THEOREM 3.1. *Under the assumption* (A)

(I) $N\ (=N(d))$ *is a non-increasing function of* $d\ (>0)$; $N(d)<\infty$ *a.s.*, $\mathrm{E}\,N(d)<\infty$ *for all* $d>0$; $\lim_{d\to 0}\mathrm{E}\,N(d)=\infty$;

(II) $\lim_{d\to 0} N(d)/\nu(d)=1$ *a.s.*;

(III) $\lim_{d\to 0}\mathrm{P}_\beta\{\beta\in I_{N(d)}\}=1-\varepsilon$ *for all* F;

(IV) $\lim_{d\to 0}\mathrm{E}\,N(d)/\nu(d)=1$,

where $\nu(d)=\tau^2_{\varepsilon/2}/144d^2[G'(0)]^2$.

One more lemma is needed for proving the theorem. This we prove below.

LEMMA 3.1. *For every* $\delta>0$, *there exist* C *and* n_0 (*all depending on* δ) *such that for* $n\geqq n_0$,

$$\mathrm{P}\{\hat{\beta}_{U,n}>\beta+Cn^{-1/2}\log n\}\leqq n^{-1-\delta}\ ; \tag{3.4}$$

$$\mathrm{P}\{\hat{\beta}_{L,n}<\beta-Cn^{-1/2}\log n\}\leqq n^{-1-\delta}\ . \tag{3.5}$$

PROOF. We prove only (3.4) as (3.5) follows analogously. We have

$$\begin{aligned}\mathrm{P}\{\hat{\beta}_{U,n}>\beta+Cn^{-1/2}\log n\}&=\mathrm{P}\{S_n(\beta+Cn^{-1/2}\log n)>-S_{n,\varepsilon}\}\\&\leqq\mathrm{P}\left\{\frac{3W_n(\beta+Cn^{-1/2}\log n)}{n+1}>-\frac{1}{2}S_{n,\varepsilon}\right\}\\&\quad+\mathrm{P}\left\{U_n(\beta+Cn^{-1/2}\log n)>-\frac{1}{2}S_{n,\varepsilon}\right\}\ .\end{aligned} \tag{3.6}$$

But $(3/(n+1))W_n(\beta+Cn^{-1/2}\log n)=O(n^{-1})$, whereas $\sqrt{n}\,S_n(\beta)$ being asymptotically normal (0, 1), for large n, $\sqrt{n}\,S_{n,\varepsilon}\approx\tau_{\varepsilon/2}$, the upper $100(\varepsilon/2)\%$ point of a normal (0, 1) distribution. So it is sufficient to consider only

$$(3.7)\quad \mathrm{P}\left\{U_n(\beta+Cn^{-1/2}\log n)>-\frac{1}{2}S_{n,\epsilon}\right\}$$
$$=\mathrm{P}\left\{U_n(\beta+Cn^{-1/2}\log n)-\mathrm{E}\,U_n(\beta+Cn^{-1/2}\log n)>-\frac{1}{2}S_{n,\epsilon}\right.$$
$$\left.-\mathrm{E}\,U_n(\beta+Cn^{-1/2}\log n)\right\}.$$

Now,

$$\mathrm{E}\,U_n(\beta+Cn^{-1/2}\log n)=12Cn^{-1/2}\log nG'(0)(1+O(n^{-1/2}\log n)).$$

Also, for large n, $S_{n,\epsilon}=O(n^{-1/2})$. Put

$$t_n=-\frac{1}{2}S_{n,\epsilon}-\mathrm{E}\,U_n(\beta+Cn^{-1/2}\log n)$$
$$=12Cn^{-1/2}\log n(-G'(0))(1+O(\log n)^{-1})$$

for large n, and so $t_n>0$ for large n. Hence, for large n, Hoeffding's bounds (cf. (5.7) in [6]) majorize (3.7) by

$$(3.8)\quad \exp\left[-\frac{2[n/3]t_n^2}{(3+3)^2}\right]\leqq\exp\left[-\frac{(n-2)t_n^2}{54}\right]$$
$$=\exp\left[-\frac{(n-2)K^2n^{-1}(\log n)^2}{54}(1+O(\log n)^{-2})\right]$$

where K is a constant depending on C. Hence, for any given $\delta>0$, C can be so chosen as the expression on the right-hand side of (3.8) $\leqq n^{-1-\delta}$. Hence the lemma.

In view of Theorem 2.3, an immediate implication of Lemma 3.1 is that as $n\to\infty$,

$$\sqrt{n}S_n(\hat{\beta}_{U,n})-\sqrt{n}S_n(\beta)-12\sqrt{n}(\hat{\beta}_{U,n}-\beta)G'(0)\to 0 \text{ a.s.}$$
$$\sqrt{n}S_n(\hat{\beta}_{L,n})-\sqrt{n}S_n(\beta)-12\sqrt{n}(\hat{\beta}_{L,n}-\beta)G'(0)\to 0 \text{ a.s.}$$

Using (3.2) it follows now on subtraction that

$$-2\sqrt{n}S_{n,\epsilon}-12\sqrt{n}(\hat{\beta}_{U,n}-\hat{\beta}_{L,n})G'(0)\to 0 \text{ a.s.}$$

From the fact $\sqrt{n}S_{n,\epsilon}\approx\tau_{\epsilon/2}$ as $n\to\infty$, one gets,

$$(3.9)\quad \sqrt{n}(\hat{\beta}_{U,n}-\hat{\beta}_{L,n})\to-\frac{\tau_{\epsilon/2}}{6G'(0)} \text{ a.s.}$$
$$(G'(0)<0 \text{ since } G(b) \text{ is } \downarrow \text{ in } b).$$

It can also be similarly shown that $\sqrt{n}(\hat{\beta}_{U,n}-\hat{\beta}_n)\to-\tau_{\epsilon/2}/12G'(0)$ a.s., $\sqrt{n}(\hat{\beta}_{L,n}-\hat{\beta}_n)\to+\tau_{\epsilon/2}/12G'(0)$ a.s. so that the comments made after (3.2) are true as roughly for large samples, the point estimator is taken to

be the midpoint of the confidence interval.

We are now in a position to prove the main theorem.

(I) It is true from definition that N is a non-increasing function of d (>0). Also,

$$\begin{aligned}\mathrm{P}(N=\infty)&=\lim_{n\to\infty}\mathrm{P}(N>n)\leqq\lim_{n\to\infty}\mathrm{P}(\hat{\beta}_{U,n}-\hat{\beta}_{L,n}>2d)\\&=\lim_{n\to\infty}\mathrm{P}(\sqrt{n}(\hat{\beta}_{U,n}-\hat{\beta}_{L,n})>2d\sqrt{n})=0 \qquad \text{from (3.9)}.\end{aligned}$$

Hence $N<\infty$ a.s. and also $\mathrm{E}\,N<\infty$. Again from definition $\lim_{d\to 0} N(d)=\infty$ a.s. and the Monotone Convergence theorem implies that $\lim_{d\to 0}\mathrm{E}\,N(d)=\infty$.

(II) We have $\hat{\beta}_{U,N(d)}-\hat{\beta}_{L,N(d)}\leqq 2d<\hat{\beta}_{U,N(d)-1}-\hat{\beta}_{L,N(d)-1}$. Hence,

$$\limsup_{d\to 0} 2d\sqrt{N(d)}\leqq\limsup_{d\to 0}\sqrt{N(d)}(\hat{\beta}_{U,N(d)-1}-\hat{\beta}_{L,N(d)-1})=-\frac{\tau_{\varepsilon/2}}{6G'(0)} \text{ a.s.}$$

$$\liminf_{d\to 0} 2d\sqrt{N(d)}\geqq\liminf_{d\to 0}\sqrt{N(d)}(\hat{\beta}_{U,N(d)}-\hat{\beta}_{L,N(d)})=-\frac{\tau_{\varepsilon/2}}{6G'(0)} \text{ a.s.}$$

Hence (II) is proved.

(III) It follows from Theorem 2.3, Lemma 3.1 and the result that $\sqrt{n}S_n(\beta)$ is asymptotically normal $(0, 1)$ that $\sqrt{n}(\hat{\beta}_{U,n}-\beta)$ is asymptotically normal $(-\tau_{\varepsilon/2}/4G'(0), 1/16[G'(0)]^2)$. Also, from Theorem 2.3, and the uniform continuity in probability of $\{S_n(\beta)\}$ wrt $\{n^{-1/2}\}$ it can be shown easily that $\{\hat{\beta}_{U,n}\}$ is uniformly continuous in probability wrt $\{n^{-1/2}\}$. Hence, $\sqrt{N}(\hat{\beta}_{U,N}-\beta)$ is asymptotically normal $(-\tau_{\varepsilon/2}/12G'(0), 1/144(G'(0))^2)$. Similarly, $\sqrt{N}(\hat{\beta}_{L,N}-\beta)$ is asymptotically normal $(\tau_{\varepsilon/2}/12G'(0), 1/144(G'(0))^2)$. Hence,

$$\begin{aligned}\lim_{d\to 0}\mathrm{P}\{\beta\in I_N\}&=\lim_{d\to 0}[\mathrm{P}\{\beta\leqq\hat{\beta}_{U,N}\}-\mathrm{P}\{\beta\leqq\hat{\beta}_{L,N}\}]\\&=\lim_{d\to 0}\mathrm{P}\left\{-12G'(0)\left(\sqrt{N}(\hat{\beta}_{U,N}-\beta)+\frac{\tau_{\varepsilon/2}}{12G'(0)}\right)\geqq-\tau_{\varepsilon/2}\right\}\\&\quad+\lim_{d\to 0}\mathrm{P}\left\{-12G'(0)\left(\sqrt{N}(\hat{\beta}_{L,N}-\beta)-\frac{\tau_{\varepsilon/2}}{12G'(0)}\right)\geqq\tau_{\varepsilon/2}\right\}\\&=1-\frac{\varepsilon}{2}-\frac{\varepsilon}{2}=1-\varepsilon\,.\end{aligned}$$

(IV) In view of (II) all we need prove is that $\{Nd^2\}_{d>0}$ is uniformly integrable. From Lemma 3.2 of Bickel and Yahav [2], it is sufficient to prove that $\sum_{m=1}^{\infty}\sup_{0<d<d_0}\mathrm{P}\{N(d)d^2>m\}<\infty$ for some $d_0>0$. Writing $m(d)=[m/d^2]$,

$$\mathrm{P}\{N(d)d^2>m\}=\mathrm{P}\{N(d)>m(d)\}\leqq\mathrm{P}\{\hat{\beta}_{U,m(d)}-\hat{\beta}_{L,m(d)}>2d\}$$

$$\leqq P\{\hat{\beta}_{U,m(d)}>d\}+P\{\hat{\beta}_{L,m(d)}<-d\}$$
$$=P\{S_{m(d)}(d)>-S_{m(d),\epsilon}\}+P\{S_{m(d)}(-d)<S_{m(d),\epsilon}\}\ .$$

We consider only $P\{S_{m(d)}(d)>-S_{m(d),\epsilon}\}$ as the other probability can be tackled similarly.

$$(3.10)\qquad P\{S_{m(d)}(d)>-S_{m(d),\epsilon}\}\leqq P\left\{\frac{3}{m(d)+1}W_{m(d)}(d)>-\frac{1}{2}S_{m(d),\epsilon}\right\}$$
$$+P\left\{U_{m(d)}(d)>-\frac{1}{2}S_{m(d),\epsilon}\right\}\ .$$

Since for large m, $S_{m(d),\epsilon}=O(m^{-1/2})$, while $(3/(m(d)+1))W_{m(d)}(d)=O(m^{-1})$, the first term on the right-hand side of (3.10) can be neglected, while,

$$(3.11)\qquad P\left(U_{m(d)}(d)>-\frac{1}{2}S_{m(d),\epsilon}\right)$$
$$=P\left(U_{m(d)}(d)-E\,U_{m(d)}(d)>-\frac{1}{2}S_{m(d),\epsilon}-E\,U_{m(d)}(d)\right)\ .$$

Using Hoeffding's bounds once again, one gets

$$(3.12)\qquad P\{U_{m(d)}(d)-E\,U_{m(d)}(d)>t\}\leqq e^{-2m(d)t^2}\ ,$$

where

$$t\equiv t(m,d)=-\frac{1}{2}S_{m(d),\epsilon}-E\,U_{m(d)}(d)\ ,$$

valid for large m. But for large m, $S_{m(d),\epsilon}=dO(m^{-1/2})$ while $E\,U_{m(d)}(d)=12d(-G'(0)(1+O(d)))$, for $0<d<d_0$; we can find a $\delta\ (>0)$ such that $t>d\delta$. Hence, the right-hand side of (3.12) is bounded above for $m\geqq M$ (say) by $e^{-2m\delta^2}$. Since $\sum_{m=M}^{\infty}e^{-2m\delta^2}$ converges, the uniform integrability of $\{Nd^2\}_{d>0}$ follows.

We have remarked at the summary that the procedure is asymptotically 'consistent' and 'efficient' in the sense of Chow and Robbins [3]. By asymptotic consistency we mean that as $d\to 0$, the coverage probability of the confidence interval $\to 1-\alpha$ (the prescribed confidence coefficient). This is guaranteed by (III) of the theorem. Asymptotic 'efficiency' roughly means that the average sample number (ASN) is asymptotically the same as the sample size one would have asymptotically obtained had F' been known and a fixed sample size procedure were used. To see this we need recall (3.9) which says

$$\sqrt{n}(\hat{\beta}_{U,n}-\hat{\beta}_{L,n})\to-\frac{\tau_{\epsilon/2}}{6G'(0)}\ \text{a.s.}\qquad\text{as } n\to\infty\ .$$

Since the width of the confidence interval is needed to be $\leqq 2d$, n is taken

to be the smallest positive integer $\geqq[-\tau_{\epsilon/2}/12dG'(0)]^2=\tau^2_{\epsilon/2}/144d^2[G'(0)]^2$. Thus (IV) of Theorem 3.1 guarantees asymptotic efficiency.

The expression for $\nu(d)$ depends on the form of the parent distribution. When (X, Y) has a bivariate normal distribution with means μ_1, μ_2, standard deviations σ_1^2, σ_2^2 and correlation coefficient ρ, one gets,

$$G(b)=\frac{1}{4}+\frac{1}{2\pi}\sin^{-1}\left(\frac{1}{2}\tau(b)\right),$$

where

$$\tau(b)=\frac{\rho\sigma_2-b\sigma_1}{(\sigma_2^2+b^2\sigma_1^2-2\rho b\sigma_1\sigma_2)^{1/2}}.$$

Then,

$$G'(0)=\frac{1}{2\pi}\frac{\tau'(0)}{\sqrt{4-\tau^2(0)}}=\frac{1}{2\pi}\left(-\frac{\sigma_1(1-\rho^2)}{\sigma_2}\right)\frac{1}{\sqrt{4-\rho^2}}=-\frac{\sigma_1(1-\rho^2)}{2\pi\sigma_2\sqrt{4-\rho^2}}.$$

Hence,

$$\nu(d)=\frac{\tau^2_{\epsilon/2}4\pi^2\sigma_2^2(4-\rho^2)}{144d^2\sigma_1^2(1-\rho^2)^2}=\frac{\pi^2\tau^2_{\epsilon/2}\sigma_2^2(4-\rho^2)}{36d^2\sigma_1^2(1-\rho^2)^2}.$$

INDIAN STATISTICAL INSTITUTE

REFERENCES

[1] Anscombe, F. J. (1952). Large sample theory of sequential estimation, *Proc. Camb. Phil. Soc.*, **48**, 600-607.

[2] Bickel, P. J. and Yahav, J. A. (1968). Asymptotically optimal Bayes and minimax procedures in sequential estimation, *Ann. Math. Statist.*, **39**, 442-456.

[3] Chow, Y. S. and Robbins, H. (1965). On the asymptotic theory of fixed width sequential confidence intervals for the mean, *Ann. Math. Statist.*, **36**, 457-462.

[4] Ghosh, M. and Sen, P. K. (1972). On bounded length confidence interval for the regression coefficient based on a class of rank statistics, *Sankhyā* A, **34**, 33-52.

[5] Hoeffding, W. (1948). A class of statistics with asymptotically normal distributions, *Ann. Math. Statist.*, **19**, 293-325.

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

[7] Loève, M. (1962). *Probability Theory* (3rd ed.), Van Nostrand, Princeton.

[8] Kendall, M. G. (1955). *Rank Correlation Methods*, Charles Griffin and Company, London, 2nd ed.

[9] Koul, H. L. (1970). Asymptotic normality of random rank statistics, *Ann. Math. Statist.*, **41**, 2144-2149.

[10] Sen, P. K. and Ghosh, M. (1971). On bounded length sequential confidence intervals based on one-sample rank order statistics, *Ann. Math. Statist.*, **42**, 189-203.