THE SPEARMAN FOOTRULE AND A MARKOV CHAIN PROPERTY

by

Pranab Kumar Sen and Ibrahim A. Salama

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1431

February 1983

THE SPEARMAN FOOTRULE AND A MARKOV CHAIN PROPERTY*

by

Pranab Kumar Sen and Ibrahim A. Salama
Department of Biostatistics, University of North Carolina,
Chapel Hill, NC 27514 USA

SUMMARY

An equivalent representation of the Spearman footrule is considered and a characterization in terms of a Markov chain is established. A martingale approach is thereby incorporated in the study of the asymptotic normality of the statistics.

AMS Subject Classifications: 62E10, 62E20, 60J99

Key words: Asymptotic normality; Markov chain; martingale; Spearman footrule; uniform (permutational) distribution.

*Work supported partially by the National Heart, Lung and Blood Institute, Contract NIH-NHLBI-71-2243-L and partially by the Grant 5-732-ES07018-01 from the National Institute of Environmental Health Sciences.

1. INTRODUCTION

Let S_n be the set of all (n!) permutations of the first n integers $\{1,\ldots,n\}$. As in Diaconis and Graham (1977), we may define a metric (D_n) on S_n by

$$D_n(\underset{\sim}{\pi}_n,\underset{\sim}{\sigma}_n) = \Sigma_{i=1}^{n}|\sigma_n(i)-\pi_n(i)| \quad , \tag{1.1}$$

where $\underset{\sim}{\sigma}_n$ $(= (\sigma_n(1),\ldots,\sigma_n(n))$ and $\underset{\sim}{\pi}_n$ $(= (\pi_n(1),\ldots,\pi_n(n))$ are elements of S_n. D_n is known as the *Spearman* (1904) *Footrule*. Its relationships with other commonly used nonparametric measures of association (such as the Kendall tau and Spearman rho coefficients) and its asymptotic normality (under the assumption that $\underset{\sim}{\sigma}_n$ and $\underset{\sim}{\pi}_n$ are chosen independently and distributed uniformly in S_n) have been studied by Diaconis and Graham (1977). The object of the present investigation is to consider an equivalent representation of D_n, to characterize a *Markovian structure* for the same, and to incorporate a *martingale approach* in the proof of the asymptotic normality of the statistics. The representation is considered in Section 2. Along with the Markovian structure, some distributional results are presented in Section 3. The concluding section deals with the asymptotic normality result.

2. A REPRESENTATION FOR D_n

Note that for every $\underset{\sim}{\sigma}_n \in S_n$, we have

$$D_n(\underset{\sim}{\sigma}_n) = D_n(\underset{\sim}{1}_n,\underset{\sim}{\sigma}_n) = \Sigma_{i=1}^{n}|i-\sigma_n(i)|$$

$$= \Sigma_{i=1}^{n}\Sigma_{j=1}^{n}|i-j|w_{ij} \tag{2.1}$$

where $\underset{\sim}{1}_n = (1,\ldots,n)$ and

$$w_{ij} = \begin{cases} 1, & \sigma_n(j) = i, \\ 0, & \text{otherwise;} \end{cases} \quad \text{for} \quad i,j = 1,\ldots,n \quad . \tag{2.2}$$

Define then

$$T_{n,i} = T_{n,i}(\underset{\sim}{\sigma}_n) = T_{n,i}(\underset{\sim}{1}_n, \underset{\sim}{\sigma}_n) = \Sigma_{j=1}^{i} I(\sigma_n(j) \le i) \quad , \tag{2.3}$$

for $i = 1,2,\ldots,n$, and let

$$T_n = T_n(\underset{\sim}{\sigma}_n) = T_n(\underset{\sim}{1}_n, \underset{\sim}{\sigma}_n) = \Sigma_{i=1}^{n} T_{n,i} \quad . \tag{2.4}$$

Then, we have the following.

Theorem 1. For every $\underset{\sim}{\sigma}_n \in S_n$ *and* $n(\ge 1)$,

$$T_n(\underset{\sim}{\sigma}_n) + \tfrac{1}{2} D_n(\underset{\sim}{\sigma}_n) = \binom{n+1}{2} \quad . \tag{2.5}$$

Proof. Note that $\Sigma_{i=1}^{n}\Sigma_{j=1}^{n} w_{ij} = n$, $\Sigma_{j=1}^{n} w_{ij} = 1$ $(\forall i)$, $\Sigma_{i=1}^{n} w_{ij} = 1$ $(\forall j)$, so that on letting $u(t)$ be equal to 1 or 0, according as t is $\ge$ or < 0, we have

$$\begin{aligned} T_n(\underset{\sim}{\sigma}_n) &= \Sigma_{i=1}^{n}\Sigma_{j=1}^{i} u(i-\sigma_n(j)) \\ &= \Sigma_{i=1}^{n}\Sigma_{j=1}^{i}\Sigma_{k=1}^{n} u(i-k) w_{jk} \\ &= \Sigma_{j=1}^{n}\Sigma_{k=1}^{n} w_{jk}\{\Sigma_{i=j}^{n} u(i-k)\} \\ &= \Sigma_{j=1}^{n}\Sigma_{k=1}^{n} w_{jk}\{(n+1) - k\vee j)\} \end{aligned} \tag{2.6}$$

where $a\vee b = \max(a,b)$ and $a\wedge b = \min(a,b)$. Therefore,

$$\begin{aligned} T_n(\underset{\sim}{\sigma}_n) + \tfrac{1}{2} D_n(\underset{\sim}{\sigma}_n) &= \Sigma_{j=1}^{n}\Sigma_{k=1}^{n} w_{jk}\{n+1-k\vee j+\tfrac{1}{2}|j-k|\} \\ &= \Sigma_{j=1}^{n}\Sigma_{k=1}^{n} w_{jk}\{n+1-k\vee j+\tfrac{1}{2}(k\vee j - k\wedge j)\} \\ &= \Sigma_{j=1}^{n}\Sigma_{k=1}^{n} w_{jk}\{n+1-\tfrac{1}{2}(k\vee j + k\wedge j)\} \end{aligned}$$

$$= \Sigma_{j=1}^{n}\Sigma_{k=1}^{n} w_{jk}\{n+1-\tfrac{1}{2}(k+j)\}$$

$$= n(n+1) - \tfrac{1}{2}\Sigma_{j=1}^{n}\Sigma_{k=1}^{n} w_{jk}(k+j)$$

$$= n(n+1) - \tfrac{1}{2}\{\Sigma_{j=1}^{n} j \Sigma_{k=1}^{n} w_{jk} + \Sigma_{k=1}^{n} k \Sigma_{j=1}^{n} w_{jk}\}$$

$$= n(n+1) - \tfrac{1}{2}\{\Sigma_{j=1}^{n} j + \Sigma_{k=1}^{n} k\} = n(n+1) - \tfrac{1}{2}n(n+1) = \binom{n+1}{2} \qquad (2.7)$$

Q.E.D.

By virtue of Theorem 1, T_n may be regarded as a complementary part of D_n. Also, note that the distribution of $D_n(\underset{\sim}{\sigma}_n, \underset{\sim}{\pi}_n)$ when $\underset{\sim}{\sigma}_n$ and $\underset{\sim}{\pi}_n$ are independent and each distributed uniformly over S_n is the same as the distribution of $D_n(\underset{\sim}{\sigma}_n)$ $(= D_n(1, \underset{\sim}{\sigma}_n))$ when $\underset{\sim}{\sigma}_n$ has a uniform distribution on S_n. The same is true for $T_n(\underset{\sim}{\sigma}_n)$ (and $T_n(\underset{\sim}{\sigma}_n, \underset{\sim}{\pi}_n)$). Hence, we shall consider only the case of $D_n(\underset{\sim}{\sigma}_n)$ and $T_n(\underset{\sim}{\sigma}_n)$.

3. A MARKOVIAN PROPERTY OF THE $T_{n,i}$

The main result of this section is the following

Theorem 2. For every $n(\geq 1)$, *whenever* $\underset{\sim}{\sigma}_n$ *is distributed uniformly on* S_n, $\{T_{n,i}; i \leq n\}$ *is a Markov chain, i.e., for every* $k(\leq n-1)$ *and* $0 \leq r_1 \leq \cdots \leq r_k \leq r_{k+1}$ $(\leq n)$,

$$P\{T_{n,k+1} = r_{k+1} \mid T_{n,j} = r_j, j \leq k\}$$

$$= P\{T_{n,k+1} = r_{k+1} \mid T_{n,k} = r_k\} \quad . \qquad (3.1)$$

Proof. Let P be the set of all permutations (of $\{1,\ldots,n\}$) satisfying the condition $\{T_{n,1}=r_1,\ldots,T_{n,k}=r_k\}$. It is easy to see that for any $p \in P$, $T_{n,k+1}$ can only assume the values r_k, r_k+1 and r_k+2. If $(\alpha_1,\ldots,\alpha_n) \in P$, then among the set $\{\alpha_1,\ldots,\alpha_k\}$, we have $k-r_k$ elements of the set

$\{k+1,\ldots,n\}$. If we denote this set by A, then, we may have either of the following:

(i) $k+1 \in A$. This happens with the (conditional) probability

$$\binom{n-k-1}{k-r_k-1}/\binom{n-k}{k-r_k} = (k-r_k)/(n-k) \quad , \tag{3.2}$$

and (ii) $k+1 \notin A$. This happens with the (conditional) probability

$$1 - (k-r_k)/(n-k) = (n-2k+r_k)/(n-k) \quad . \tag{3.3}$$

In case (i), $T_{n,k+1}$ can assume only the values r_k+1 or r_k+2 with probabilities $\{(n-k) - (k-r_k)\}/(n-k)$ and $(k-r_k)/(n-k)$, respectively, while in case (ii), $T_{n,k+1}$ can assume only the values r_k or r_k+1 with respective probabilities $\{(n-k) - (k-r_k) - 1\}/(n-k)$ and $\{(k-r_k) + 1\}/(n-k)$. Thus, the assumable values of $T_{n,k+1}$ (viz., r_k, r_k+1 and r_k+2) and their respective (conditional) probabilities (given the $T_{n,i}$, $i \leq k$) depend only on the value r_k assumed by $T_{n,k}$. Q.E.D.

Actually, by very similar (elementary) arguments, we have the following:

Lemma 3. If $\underset{\sim n}{\sigma}$ *has a uniform distribution on* S_n, *then for every* k $(1 \leq k \leq n)$ *and* r,

$$P\{T_{n,k}=r\} = \binom{n}{k}^{-1}\binom{k}{r}\binom{n-k}{k-r}, \quad \text{for} \quad r = 0 \vee (2k-n),\ldots,k \ , \tag{3.4}$$

and, for every $k<q$, $r \leq s$,

$$P\{T_{n,k}=r, T_{n,q}=s\} = \frac{\binom{n-q}{q-s}\sum_{u \geq r}\binom{k}{u}\binom{q-k}{s-u}\binom{u}{r}\binom{q-u}{k-r}}{n!\{k!(q-k)!(n-q)!\}^{-1}} \quad . \tag{3.5}$$

For our subsequent analysis, we may note that by (3.4) and (3.5),

$$P\{T_{n,k+1}=s \mid T_{n,k}=r\} = \frac{\binom{n-k-1}{k+1-s}}{(n-k)\binom{n-k}{k-r}} \sum_{u \geq r}\binom{1}{s-u}\binom{k-r}{u-r}\binom{k+1-u}{k-r} \quad , \tag{3.6}$$

for $s \geq r$ (and 0, for $s<r$), so that

$$P\{T_{n,k+1}=s \mid T_{n,k}=r\} = \begin{cases} (n-2k+r)(n-2k+r-1)/(n-k)^2, & s=r \\ (n-2k+r)(2k-2r+1)/(n-k)^2, & s=r+1 \\ (k-r)^2/(n-k)^2, & s=r+2 \\ 0, \ s \geq r+3 \ \text{ or } \ s<r \quad . \end{cases} \tag{3.7}$$

Hence, from (3.7) we have

$$E\{T_{n,k+1} \mid T_{n,k}\} = \frac{(n-k-1)^2}{(n-k)^2} T_{n,k} + \frac{2k+1}{n-k} - \frac{k}{(n-k)^2} \tag{3.8}$$

for $k = 0,1,\ldots,n-1$, where, conventionally, we let $T_{n,0} = 0$. Finally, by (3.4) and (3.5), we have

$$\mu_{n,k} = ET_{n,k} = n^{-1}k^2 \ , \quad \gamma^2_{n,k} = \mathrm{Var}(T_{n,k}) = k^2(n-k)^2/n^2(n-1) \ , \tag{3.9}$$

$$\gamma_{n,kq} = \mathrm{Cov}(T_{n,k},T_{n,q}) = (n(k \wedge q)-kq)^2/n^2(n-1) \ , \tag{3.10}$$

for every $k,q = 1,\ldots,n$, so that

$$\mu_n = ET_n = n(2n+1)/6 \quad \text{and} \quad \gamma^2_n = \mathrm{Var}(T_n) = n(2n^2+7)/180 \quad . \tag{3.11}$$

We also write

$$\nu_{n,k} = ET^2_{n,k} = \gamma^2_{n,k} + \mu^2_{n,k} \ , \quad 0 \leq k \leq n \quad . \tag{3.12}$$

4. ASYMPTOTIC PERMUTATIONAL NORMALITY OF T_n

Motivated by the Markovian property in (3.1), we would like to prove the following result via a martingale central limit theorem.

Theorem 4. If for every n, $\underset{\sim}{\sigma}_n$ *has the uniform distribution over* S_n, *then as* $n \to \infty$,

$$T_n^* = (T_n - \mu_n)/\gamma_n \xrightarrow{\mathcal{D}} N(0,1) \quad . \tag{4.1}$$

Proof. For every k $(1 \leq k \leq n-1)$, we let

$$d_{nk} = \{(n-k+1)(2k-1) - (k-1)\}/\{(n-k)(n-k+1)\}^2; \quad d_{nk}^* = \Sigma_{j=1}^{k} d_{nk} \quad . \tag{4.2}$$

Then, by (3.8), we have on letting

$$Y_{nk} = (n-k)^{-2} T_{n,k} - d_{nk}^* \quad , \quad \mathcal{B}_{nk} = \mathcal{B}(T_{n,j};\ j \leq k) \quad , \tag{4.3}$$

for $k = 0,1\ldots,n$, that

$$E(Y_{nk} | \mathcal{B}_{nk-1}) = Y_{nk-1} \quad , \quad \forall\ 1 \leq k \leq n-1 \quad . \tag{4.4}$$

Therefore, if we let

$$Z_{nk} = Y_{nk} - Y_{nk-1} \quad , \quad 1 \leq k \leq n-1 \quad , \tag{4.5}$$

then the Z_{nk} are martingale differences. By (2.4), (4.2), (4.3) and (4.5), we have

$$\begin{aligned} T_n - \mu_n &= \Sigma_{i=1}^{n-1} (n-i)^2 Y_{ni} \\ &= \Sigma_{i=1}^{n-1} (n-i)^2 (\Sigma_{j=1}^{i} Z_{ni}) \\ &= \Sigma_{k=1}^{n-1} Z_{nk} (\Sigma_{j=k}^{n-1} (n-j)^2) \\ &= \Sigma_{k=1}^{n-1} \{(n-k)(n-k+1)(2n-2k+1)/6\} Z_{nk} \quad . \end{aligned} \tag{4.6}$$

Thus, if we let

$$c_{ni} = \frac{(n-i)(n-i+1)(2n-2i+1)}{6\{(n+1)(2n^2+7)/180\}^{\frac{1}{2}}} \quad , \quad 1 \leq i \leq n-1 \quad , \tag{4.7}$$

$$U_{ni} = c_{ni} Z_{ni} \quad , \quad i = 1,\ldots,n-1; \quad U_{n0} = 0 \quad , \tag{4.8}$$

then, by (3.11), (4.6), (4.7) and (4.8), we have

$$T_n^* = \Sigma_{k=1}^{n-1} U_{nk} \quad , \tag{4.9}$$

where by (4.4), (4.5) and (4.8),

$$E\{U_{nk} | \mathcal{B}_{nk-1}\} = 0 \quad , \quad \forall\ 1 \leq k \leq n-1 \tag{4.10}$$

and

$$\Sigma_{i=1}^{n} E\, U_{ni}^2 = 1 \quad , \quad \forall\ n \geq 1 \quad . \tag{4.12}$$

Thus, T_n^* relates to a martingale array, normalized by (4.12), and to establish (4.1), we need only to verify the following:

$$\tilde{U}_n = \Sigma_{i=1}^{n-1} \tilde{U}_{ni} = \Sigma_{i=1}^{n-1} E(U_{ni}^2 | \mathcal{B}_{ni-1}) \xrightarrow{p} 1 \quad , \tag{4.13}$$

and, for every $\varepsilon > 0$,

$$\Sigma_{i=1}^{n} E\{U_{ni}^2\, I(|U_{ni}| > \varepsilon) | \mathcal{B}_{ni-1}\} \xrightarrow{p} 0 \quad . \tag{4.14}$$

Since the $T_{n,k}$ are nonnegative, $T_{n,k} \leq k$, $\forall\ k \leq n$, by using (3.7), (4.2), (4.3), (4.5), (4.7) and (4.8), it can be easily seen that

$$|U_{ni}| \leq C n^{-\frac{1}{2}} \quad , \quad \text{with probability 1} \quad , \tag{4.15}$$

for every i: $1 \leq i \leq n-1$, where C does not depend on n $(0 < C < \infty)$. Hence, (4.15) ensures that (4.14) holds for n adequately large. Further, by virtue of (4.12), to prove (4.13), it suffices to show that

$$E(\tilde{U}_n - 1)^2 \to 0 \quad , \quad \text{as} \quad n \to \infty \quad . \tag{4.16}$$

Towards this, note that for every i: $1 \leq i \leq n-1$,

$$\tilde{U}_{ni} = E(U_{ni}^2 | \mathcal{B}_{ni-1})$$

$$= a_{ni} T_{ni-1}^2 + B_{ni} T_{ni-1} + g_{ni} \ , \quad \text{say,} \tag{4.17}$$

where, by (4.7),

$$a_{ni} = O(n^{-3}) \ , \quad b_{ni} = O(n^{-2}) \quad \text{and} \quad g_{ni} = O(n^{-1}) \tag{4.18}$$

for every i $(= 1,\ldots,n-1)$. Further,

$$\tilde{U}_n - 1 = \tilde{U}_n - E\,\tilde{U}_n$$

$$= \sum_{i=1}^{n-1} \{a_{ni}(T_{n,i-1}^2 - \nu_{n,i-1}) + b_{ni}(T_{n,i-1} - \mu_{n,i-1}) \ , \tag{4.19}$$

so that on noting that (by virtue of (3.4)-(3.5))

$$E(T_{n,i-1}^2 - \nu_{n,i-1})^2 = O(n^3) \ , \tag{4.20}$$

$$E(T_{n,i-1} - \mu_{n,i-1})^2 = O(n) \ , \tag{4.21}$$

$$E(T_{n,i-1}^2 - \nu_{n,i-1})(T_{n,i-1} - \mu_{n,i-1}) = O(n^2) \ , \tag{4.22}$$

$$E(T_{n,i-1} - \mu_{n,i-1})(T_{n,j-1} - \mu_{n,j-1}) = O(n) \ , \tag{4.23}$$

for every i $(= 1,\ldots,n-1)$, we obtain from (4.19)-(4.23) that

$$E(\tilde{U}_n - 1)^2 = O(n^{-1}) \ , \tag{4.24}$$

and hence, (4.16) holds. This completes the proof of the theorem.

We conclude this section with the remark that the martingale approach also enables us to prove an invariance principle relating to the partial sequence $\{\Sigma_{i=1}^{k}(T_{n,i} - \mu_{n,i})/\gamma_n, \ k \leq n\}$, where the earlier method of Diaconis

and Graham (1977) may not lead to a simple solution. This invariance principle will particularly be useful to extend (4.1) to the case where the sample size is itself a (positive integer valued) random variable.

REFERENCES

Persi Diaconis and R.L. Graham (1977). Spearman footrule as a measure of discovery. *J. Roy. Statist. Soc. B.* 39, 262-268.

Charles Spearman (1904). The proof and measurement of association between two things. *Amer. J. Psychology* 15, 72-101.