Available online at www.isas.org.in/jisas

JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 70(1) 2016 1-6

Accuracy in Estimating Kendall's Tau in Sampling Finite Populations*

Arijit Chaudhuri[1] and Purnima Shaw[2]
[1]*Indian Statistical Institute, Kolkata*
[2]*Reserve Bank of India, Bandra-Kurla Complex, Bandra (East), Mumbai*

Received 13 July 2015; Revised 18 January 2016; Accepted 8 February 2016

SUMMARY

From a general unequal probability sample a standard estimator for Karl Pearson's product-moment correlation coefficient between two variables in a finite population is taken as a non-linear function of unbiased estimators respectively for six specific population totals. By Taylor series expansion an approximate variance estimator for it is also available. The corresponding Spearman's rank correlation coefficient has no such facility because sample ranks bear no discernible relations to individual-wise population ranks. But Kendall's rank correlation coefficient "Tau" has no such shortcoming. Rather, it is still simpler involving only 'totals of three variables, instead of six' and the corresponding estimators. Applying Taylor series expansion its accuracy level is examined. Simulation-based numerical results are also presented that look encouraging.

Keywords: Linearization, Product-moment correlation coefficient, Rank correlation, Unequal probability sampling.

1. INTRODUCTION

Let x and y be two real variables with values x_i, y_i for individuals labelled i in a finite survey population $U = (1,2,3,\ldots,i,\ldots,N)$. The product-moment correlation coefficient between them is

$$R_N = \frac{N\sum_{i=1}^{N} x_i y_{i-} \left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right)}{\sqrt{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} y_i\right)^2}\sqrt{\sum_{i=1}^{N} y_i^2 - \left(\sum_{i=1}^{N} y_i\right)^2}} \quad (1.1)$$

This is a non-linear function of six population totals, namely $\theta_1 = N$, $\theta_2 = \sum_{i=1}^{N} x_i y_i$, $\theta_3 = \sum_{i=1}^{N} x_i$, $\theta_4 = \sum_{i=1}^{N} y_i$, $\theta_5 = \sum_{i=1}^{N} x_i^2$, $\theta_6 = \sum_{i=1}^{N} y_i^2$. If a sample s is taken from U with a probability $p(s)$ according to a design admitting positive first order and second order inclusion-probabilities $\pi_i = \sum_{s \ni i} p(s)$ and $\pi_{ij} = \sum_{s \ni i,j} p(s)$, then a standard estimator r for R_N is taken as

$$r = \frac{\left(\sum_{i\in s}\frac{1}{\pi_i}\right)\left(\sum_{i\in s}\frac{x_i y_i}{\pi_i}\right) - \left(\sum_{i\in s}\frac{x_i}{\pi_i}\right)\left(\sum_{i\in s}\frac{y_i}{\pi_i}\right)}{\sqrt{\left(\sum_{i\in s}\frac{1}{\pi_i}\right)\left(\sum_{i\in s}\frac{x_i^2}{\pi_i}\right) - \left(\sum_{i\in s}\frac{x_i}{\pi_i}\right)^2}\sqrt{\left(\sum_{i\in s}\frac{1}{\pi_i}\right)\left(\sum_{i\in s}\frac{y_i^2}{\pi_i}\right) - \left(\sum_{i\in s}\frac{y_i}{\pi_i}\right)^2}} \quad (1.2)$$

Like R_N this r also takes values in the closed interval [-1, +1]. Writing $r = f(t_1, t_2, t_3, t_4, t_5, t_6) = f(\mathbf{t})$ as a function of the respective unbiased estimators t_j for θ_j, j=1, 2, 3, 4, 5, 6, assuming large sample-size, writing $\theta = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6)$, one may expand $f(\mathbf{t})$ about $f(\boldsymbol{\theta}) = R_N$ and by Taylor series expansion neglecting higher order terms write

---

*The views expressed in this paper are of the author alone and not of the institution, author is working in.

Corresponding author: Purnima Shaw
E-mail address: purnimashaw2011@gmail.com

$$f(\mathbf{t}) \simeq f(\boldsymbol{\theta}) + \sum_{j=1}^{6} \frac{\partial f(\mathbf{t})}{\partial t_j}\bigg|_{\mathbf{t}=\boldsymbol{\theta}} (t_j - \theta_j)$$

$$= f(\boldsymbol{\theta}) + \sum_{j=1}^{6} \lambda_j (t_j - \theta_j), \text{ writing } \lambda_j = \frac{\partial f(\mathbf{t})}{\partial t_j}\bigg|_{\mathbf{t}=\boldsymbol{\theta}}$$

This well-known result yields a convenient approximate formula for V(r) = Vf($\mathbf{t}$) leading to a simple formula for an estimator for it which is approximately unbiased for V(r). If, we have x_i, y_i as the values of ranks of the units of U according to two qualitative characteristics A and B, say, then R_N is given by the Spearman's rank correlation coefficient

$$R_S = 1 - \frac{6\sum_{i=1}^{N} {d_i}^2}{N(N^2-1)},\ d_i = y_i - x_i, i \epsilon U. \tag{1.3}$$

This is useful; for example, Dubey and Gangopadhyay (1998) used this in an important Indian context. Unfortunately, even though R_S and R_N are same, a corresponding simple estimator like r cannot be employed for R_S by the above Taylor series approach. The reason is "sample ranks bear no natural relations to the population ranks" and t_j's corresponding to the θ_j's cannot be obtained. Dubey *et al*. (1998) did not say anything about the accuracy of Spearman's rank correlation coefficients they referred to.

In this new work we intend to make it a point that if circumstances demand (i) obtaining rank correlation coefficients and (ii) assessing their accuracy levels from a sample chosen according to a general sampling design, one may safely resort to using Kendall's (1938) rank correlation coefficient called "Tau", denoted by τ.

From Kendall (1955), τ may be regarded as a product-moment coefficient. Writing u_i = rank according to A and v_i = rank according to B, for the pair (i,j), with $i < j$, let a_{ij} and b_{ij} be such that

$$a_{ij} = \begin{cases} +1 \text{ if } u_i < u_j \\ 0 \text{ if } u_i = u_j \\ -1 \text{ if } u_i > u_j \end{cases}$$

$$b_{ij} = \begin{cases} +1 \text{ if } v_i < v_j \\ 0 \text{ if } v_i = v_j \\ -1 \text{ if } v_i > v_j \end{cases}$$

Then,

$$\tau = \frac{\sum_i^N \sum_{<j}^N a_{ij} b_{ij}}{\sqrt{\sum_i^N \sum_{<j}^N a_{ij}^2}\sqrt{\sum_i^N \sum_{<j}^N b_{ij}^2}} \tag{1.4}$$

is Kendall's Tau.

In Section 2 we propose an estimator for τ, along with variance estimator. Confidence Intervals for τ are also derived. Numerical calculations regarding the estimation of Kendall's τ using hypothetical data are presented.

2. ESTIMATION FROM SAMPLES

2.1 Estimation of Kendall's τ

A sample s is chosen from U with a pre-assigned probability p(s) with first order inclusion probability $\pi_i = \sum_{s \ni i} \mathrm{p}(s) > 0,\ \ i$, second order inclusion probability $\pi_{ij} = \sum_{s \ni i,j} \mathrm{p}(s) > 0,\ \ i \neq j$, third order inclusion probability $\pi_{ijk} = \sum_{s \ni i,j,k} \mathrm{p}(s) > 0,\ \ i \neq j \neq k$ and fourth order inclusion probability $\pi_{ijkl} = \sum_{s \ni i,j,k,l} \mathrm{p}(s) > 0,\ \ i \neq j \neq k \neq l$. Assume each s contains n units each distinct. The sampled units are ranked with respect to A as $u'_1, u'_2, \ldots, u'_n$ and with respect to B as $v'_1, v'_2, \ldots, v'_n$. Let us write

$$\tau = \frac{\sum_i^N \sum_{<j}^N a_{ij} b_{ij}}{\sqrt{\sum_i^N \sum_{<j}^N a_{ij}^2}\sqrt{\sum_i^N \sum_{<j}^N b_{ij}^2}} = \frac{\theta_1}{\sqrt{\theta_2 \theta_3}} = \mathrm{f}(\theta) \tag{2.1.1}$$

where $\theta_1 = \sum_i^N \sum_{<j}^N a_{ij} b_{ij}$,

$\theta_2 = \sum_i^N \sum_{<j}^N a_{ij}^2$

and $\theta_3 = \sum_i^N \sum_{<j}^N b_{ij}^2$.

for $i<j$, let a'_{ij} and b'_{ij} be such that

$$a'_{ij} = \begin{cases} +1 \text{ if } u'_i < u'_j \\ 0 \text{ if } u'_i = u'_j \\ -1 \text{ if } u'_i > u'_j \end{cases}$$

$$b'_{ij} = \begin{cases} +1 \text{ if } v'_i < v'_j \\ 0 \text{ if } v'_i = v'_j \\ -1 \text{ if } v'_i > v'_j \end{cases}$$

Clearly, $a'_{ij} = a_{ij}$ and $b'_{ij} = b_{ij}$ irrespective of the change in ranks of the units in the population and the sample.

Let us use the following Horvitz Thompson (1952) unbiased estimators for θ_1, θ_2 and θ_3.

Consider $t_1 = \sum_i \sum_{<j\in s} \frac{a'_{ij}\, b'_{ij}}{\pi_{ij}}$

$$E_p(t_1) = \sum_s p(s) \sum_i \sum_{<j\in s} \frac{a'_{ij}\, b'_{ij}}{\pi_{ij}}$$

$$= \sum_i^N \sum_{<j}^N \frac{a'_{ij} b'_{ij}}{\pi_{ij}} \sum_{s\ni i,j} p(s)$$

$$= \sum_i^N \sum_{<j}^N \frac{a'_{ij}\, b'_{ij}}{\pi_{ij}} \pi_{ij}$$

$$= \sum_i^N \sum_{<j}^N a'_{ij}\, b'_{ij}$$

$$= \sum_i^N \sum_{<j}^N a_{ij} b_{ij} = \theta_1;$$

$$t_2 = \sum_i \sum_{<j\in s} \frac{a'^2_{ij}}{\pi_{ij}}$$

and $t_3 = \sum_i \sum_{<j\in s} \frac{b'^2_{ij}}{\pi_{ij}}$, which are unbiased estimators of θ_2 and θ_3 respectively.

We take

$$\hat{\tau} = \hat{f}(\theta) = f(\mathbf{t}) = \frac{t_1}{\sqrt{t_2\, t_3}}$$

$$= \frac{\sum_i \sum_{<j\in s} \frac{a'_{ij}\, b'_{ij}}{\pi_{ij}}}{\sqrt{\sum_i \sum_{<j\in s} \frac{a'^{\,2}_{ij}}{\pi_{ij}}}\sqrt{\sum_i \sum_{<j\in s} \frac{b'^{\,2}_{ij}}{\pi_{ij}}}} \qquad (2.1.2)$$

as an estimator for τ and it is approximately unbiased for τ for large sample size n.

By Cauchy-Schwartz Inequality,

$$\left[\sum_i \sum_{j\in s} \left(\frac{a'_{ij}}{\sqrt{\pi_{ij}}}\right)\left(\frac{b'_{ij}}{\sqrt{\pi_{ij}}}\right)\right]^2 \le \left[\sum_i \sum_{j\in s} \frac{a'^2_{ij}}{\pi_{ij}}\right]\left[\sum_i \sum_{j\in s} \frac{b'^2_{ij}}{\pi_{ij}}\right]$$

$$\Rightarrow \left[\sum_i \sum_{j\in s} \left(\frac{a'_{ij}\, b'_{ij}}{\pi_{ij}}\right)\right]^2 \le \left[\sum_i \sum_{j\in s} \frac{a'^2_{ij}}{\pi_{ij}}\right]\left[\sum_i \sum_{j\in s} \frac{b'^2_{ij}}{\pi_{ij}}\right]$$

$$\Rightarrow \frac{\left[\sum_i \sum_{<j\in s} \left(\frac{a'_{ij}\, b'_{ij}}{\pi_{ij}}\right)\right]^2}{\left[\sum_i \sum_{<j\in s} \frac{a'^2_{ij}}{\pi_{ij}}\right]\left[\sum_i \sum_{<j\in s} \frac{b'^2_{ij}}{\pi_{ij}}\right]} \le 1$$

$$\Rightarrow \hat{\tau}^2 \le 1$$

$$\Rightarrow -1 \le \hat{\tau} \le 1$$

2.2 Calculation of $V_p(\hat{\tau})$ and its Estimate using Linearization Technique

$$\hat{\tau} = f(\mathbf{t}) = \frac{t_1}{\sqrt{t_2 t_3}}$$

Using Taylor series expansion and neglecting higher order terms, we get, approximately

$$f(\mathbf{t}) = f(\boldsymbol{\theta}) + \left.\frac{\partial f(\mathbf{t})}{\partial t_1}\right|_{\mathbf{t}=\boldsymbol{\theta}} (t_1-\theta_1) + \left.\frac{\partial f(\mathbf{t})}{\partial t_2}\right|_{\mathbf{t}=\boldsymbol{\theta}} (t_2-\theta_2) + \left.\frac{\partial f(\mathbf{t})}{\partial t_3}\right|_{\mathbf{t}=\boldsymbol{\theta}} (t_3-\theta_3)$$

$$V_p\{f(\mathbf{t})\} = V_p\left\{\left.\frac{\partial f(\mathbf{t})}{\partial t_1}\right|_{\mathbf{t}=\boldsymbol{\theta}} t_1 + \left.\frac{\partial f(\mathbf{t})}{\partial t_2}\right|_{\mathbf{t}=\boldsymbol{\theta}} t_2 + \left.\frac{\partial f(\mathbf{t})}{\partial t_3}\right|_{\mathbf{t}=\boldsymbol{\theta}} t_3\right\};$$

$$\text{Now, } \left.\frac{\partial f(\mathbf{t})}{\partial t_1}\right|_{\mathbf{t}=\boldsymbol{\theta}} = \frac{1}{\sqrt{\theta_2\theta_3}},$$

$$\left.\frac{\partial f(\mathbf{t})}{\partial t_2}\right|_{\mathbf{t}=\boldsymbol{\theta}} = \frac{-\theta_1}{2\theta_2\sqrt{\theta_2\theta_3}}$$

$$\left.\frac{\partial f(\mathbf{t})}{\partial t_3}\right|_{t=\theta} = \frac{-\theta_1}{2\theta_3\sqrt{\theta_2\theta_3}};$$

$$V_p\{f(\mathbf{t})\} = V_p\left\{\begin{array}{l}\frac{1}{\sqrt{\theta_2\theta_3}}\sum_i \sum_{<j\in s} \frac{a'_{ij}\, b'_{ij}}{\pi_{ij}} + \frac{-\theta_1}{2\theta_2\sqrt{\theta_2\theta_3}} \\ \sum_i \sum_{<j\in s} \frac{a'^2_{ij}}{\pi_{ij}} + \frac{-\theta_1}{2\theta_3\sqrt{\theta_2\theta_3}} \sum_i \sum_{<j\in s} \frac{b'^2_{ij}}{\pi_{ij}}\end{array}\right\}$$

$$= V_p\left(\sum_i \sum_{j\in s} \frac{\Psi_{ij}}{\pi_{ij}}\right)$$

$$\left[\text{taking } \Psi_{ij} = \frac{1}{\sqrt{\theta_2\theta_3}} a'_{ij} b'_{ij} + \frac{-\theta_1}{2\theta_2\sqrt{\theta_2\theta_3}} a'^2_{ij} + \frac{-\theta_1}{2\theta_3\sqrt{\theta_2\theta_3}} b'^2_{ij}\right]$$

$$= E_p\left(\sum_i \sum_{<j\in s} \frac{\Psi_{ij}}{\pi_{ij}}\right)^2 - \left(\sum_i^N \sum_{<j}^N \Psi_{ij}\right)^2$$

$$= \sum_s p(s) \left(\sum_i \sum_{<j\in s} \frac{\Psi_{ij}}{\pi_{ij}}\right)^2 - \left(\sum_i^N \sum_{<j}^N \Psi_{ij}\right)^2$$

$$= \sum_s p(s) \sum_i \sum_{<j\in s} \frac{\Psi_{ij}{}^2}{\pi_{ij}{}^2}$$
$$+2\sum_s p(s) \sum_i \sum_{<j} \sum_{<l\in s} \frac{\Psi_{ij}\Psi_{il}}{\pi_{ij}\pi_{il}}$$
$$+2\sum_s p(s) \sum_i \sum_{<k} \sum_{<j\in s} \frac{\Psi_{ij}\Psi_{kj}}{\pi_{ij}\pi_{kj}}$$
$$+2\sum_s p(s) \sum_i \sum_{<j} \sum_{<l\in s} \frac{\Psi_{ij}\Psi_{jl}}{\pi_{ij}\pi_{jl}}$$
$$+\sum_s p(s) \sum_i \sum_{<j} \sum_{\neq k} \sum_{<l\in s} \frac{\Psi_{ij}\Psi_{kl}}{\pi_{ij}\pi_{kl}}$$
$$-\sum_i^N \sum_{<j}^N \Psi_{ij}{}^2$$
$$-2\sum_i^N \sum_{<j}^N \sum_{<i}^N \Psi_{ij}\Psi_{il}$$

$$-2\sum_{i}^{N}\sum_{<k}^{N}\sum_{<j}^{N}\Psi_{ij}\Psi_{kj}$$

$$-2\sum_{i}^{N}\sum_{<j}^{N}\sum_{<i}^{N}\Psi_{ij}\Psi_{jl}$$

$$-2\sum_{i}^{N}\sum_{<j}^{N}\sum_{<\neq k}^{N}\sum_{i}^{N}\Psi_{ij}\Psi_{kj}$$

$$= \sum_{i}^{N}\sum_{<j}^{N}\frac{\Psi_{ij}^{2}}{\pi_{ij}} + 2\sum_{i}^{N}\sum_{<j}^{N}\sum_{<l}^{N}\frac{\Psi_{ij}\Psi_{il}\,\pi_{ijl}}{\pi_{ij}\,\pi_{il}}$$

$$+ 2\sum_{i}^{N}\sum_{<k}^{N}\sum_{<j}^{N}\frac{\Psi_{ij}\Psi_{kj}\pi_{ijk}}{\pi_{ij}\pi_{kj}}$$

$$+ 2\sum_{i}^{N}\sum_{<j}^{N}\sum_{<l}^{N}\frac{\Psi_{ij}\Psi_{jl}\pi_{ijl}}{\pi_{ij}\pi_{jl}}$$

$$+\sum_{i}^{N}\sum_{<j}^{N}\sum_{\neq k}^{N}\sum_{<l}^{N}\frac{\Psi_{ij}\Psi_{kl}\pi_{ijkl}}{\pi_{ij}\pi_{kl}} - \sum_{i}^{N}\sum_{<j}^{N}\Psi_{ij}^{2}$$

$$-2\sum_{i}^{N}\sum_{<j}^{N}\sum_{<l}^{N}\Psi_{ij}\Psi_{il} - 2\sum_{i}^{N}\sum_{<k}^{N}\sum_{<j}^{N}\Psi_{ij}\Psi_{kj}$$

$$-2\sum_{i}^{N}\sum_{<j}^{N}\sum_{<l}^{N}\Psi_{ij}\Psi_{jl}$$

$$-\sum_{i}^{N}\sum_{<j}^{N}\sum_{\neq k}^{N}\sum_{<l}^{N}\Psi_{ij}\Psi_{kl}$$

$$=\sum_{i}^{N}\sum_{<j}^{N}\Psi_{ij}^{2}\frac{(1-\pi_{ij})}{\pi_{ij}}$$

$$+2\sum_{i}^{N}\sum_{<j}^{N}\sum_{<l}^{N}\Psi_{ij}\Psi_{il}\frac{(\pi_{ijl}-\pi_{ij}\pi_{il})}{\pi_{ij}\pi_{il}}$$

$$+2\sum_{i}^{N}\sum_{<k}^{N}\sum_{<j}^{N}\Psi_{ij}\Psi_{kj}\frac{(\pi_{ijk}-\pi_{ij}\pi_{kj})}{\pi_{ij}\pi_{kj}}$$

$$+2\sum_{i}^{N}\sum_{<j}^{N}\sum_{<l}^{N}\Psi_{ij}\Psi_{jl}\frac{(\pi_{ijl}-\pi_{ij}\pi_{jl})}{\pi_{ij}\pi_{jl}}+$$

$$\sum_{i}^{N}\sum_{<j}^{N}\sum_{\neq k}^{N}\sum_{<l}^{N}\Psi_{ij}\Psi_{kl}\frac{(\pi_{ijkl}-\pi_{ij}\pi_{kl})}{\pi_{ij}\pi_{kl}} \quad (2.2.1)$$

$$\widehat{V}_p(\hat{\tau})=\widehat{V}_p\{f(t)\}=\sum_{i}\sum_{<j\in s}\widehat{\Psi}_{ij}^{2}\frac{(1-\pi_{ij})}{\pi_{ij}^{2}} +$$

$$2\sum_{i}\sum_{<j}\sum_{<l\in s}\widehat{\Psi}_{ij}\widehat{\Psi}_{il}\frac{(\pi_{ijl}-\pi_{ij}\pi_{il})}{\pi_{ijl}\pi_{ij}\pi_{il}}$$

$$+2\sum_{i}\sum_{<k}\sum_{<j\in s}\widehat{\Psi}_{ij}\widehat{\Psi}_{kj}\frac{(\pi_{ijk}-\pi_{ij}\pi_{kj})}{\pi_{ijk}\pi_{ij}\pi_{kj}}+$$

$$2\sum_{i}\sum_{<j}\sum_{<l\in s}\widehat{\Psi}_{ij}\widehat{\Psi}_{jl}\frac{(\pi_{ijl}-\pi_{ij}\pi_{jl})}{\pi_{ijl}\pi_{ij}\pi_{jl}}$$

$$+\sum_{i}\sum_{<j}\sum_{\neq k}\sum_{<l\in s}\widehat{\Psi}_{ij}\widehat{\Psi}_{kl}\frac{(\pi_{ijkl}-\pi_{ij}\pi_{kl})}{\pi_{ijkl}\pi_{ij}\pi_{kl}}, (2.2.2)$$

where $\widehat{\Psi}_{ij}=\frac{1}{\sqrt{t_2t_3}}a'_{ij}b'_{ij} + \frac{-t_1}{2t_2\sqrt{t_2t_3}}a'^2_{ij} + \frac{-t_1}{2t_3\sqrt{t_2t_3}}b'^2_{ij}$

and similarly $\widehat{\Psi}_{il}$, $\widehat{\Psi}_{kj}$, $\widehat{\Psi}_{jl}$ and $\widehat{\Psi}_{kl}$ are defined.

$$\Rightarrow E_p[\widehat{V}_p(\hat{\tau})] \simeq V_p(\hat{\tau}).$$

2.3 Confidence Interval (CI) for τ

A 100 (1-α)% Confidence Interval (CI) for τ can be obtained by two methods:

(a) Method 1

Using chebychev's in inequality we can approximately write, negelechng the bas term

$$P\left[|\hat{\tau} - \tau| \geq t\sqrt{Vp(\hat{\tau})}\right] \leq \frac{1}{t^2} \text{ for } t > 0$$

$$\Rightarrow P\left[|\hat{\tau} - \tau| \leq t\sqrt{Vp(\hat{\tau})}\right] \geq 1 - \frac{1}{t^2}$$

Taking $\frac{1}{t^2} = \alpha$

$$\Rightarrow t = +\sqrt{\frac{1}{\alpha}}\ .$$

Then,

$$P\left[\hat{\tau} - \sqrt{\frac{Vp(\hat{\tau})}{\alpha}} \leq \tau \leq \hat{\tau} + \sqrt{\frac{Vp(\hat{\tau})}{\alpha}}\right] \geq 1 - \alpha \quad (2.3.1)$$

An approximate 100 (1-α)% Confidence Interval for τ is given by

$$\left(\hat{\tau} - \sqrt{\frac{\widehat{V}_p(\hat{\tau})}{\alpha}}, \hat{\tau} + \sqrt{\frac{\widehat{V}_p(\hat{\tau})}{\alpha}}\right).$$

(b) Method 2

Assuming $\hat{\tau}\sim$Normal $(\tau, Vp(\hat{\tau}))$, it is implied that

$$\frac{\hat{\tau}-\tau}{\sqrt{\widehat{V}_p(\hat{\tau})}} \sim t_{n-1}$$

where t_{n-1} is the Student's t-distribution with n-1 degrees of freedom.

An approximate 100 (1-α)% Confidence Interval for τ is derived from:

$$P\left[\frac{|\hat{\tau}-\tau|}{\sqrt{\widehat{V}_p(\hat{\tau})}} \leq t_{\frac{\alpha}{2},n-1}\right] \geq 1-\alpha$$

where $t_{\frac{\alpha}{2},n-1}$ is the upper 100 $(\frac{\alpha}{2})$% point of the Student's t-distribution with n-1 degrees of freedom.

$$\text{or, } P\left[\hat{\tau} - \sqrt{\widehat{V}_p(\hat{\tau})}\quad t_{\frac{\alpha}{2},n-1} \leq \tau \leq \hat{\tau} + \sqrt{\widehat{V}_p(\hat{\tau})}\quad t_{\frac{\alpha}{2},n-1}\right] \geq 1-\alpha. \quad (2.3.2)$$

An approximate 100 (1-α)% Confidence Interval for τ is given by

$$\left(\hat{\tau} - \sqrt{\hat{V}_p(\hat{\tau})}\ \ t_{\frac{\alpha}{2},n-1}, \hat{\tau} + \sqrt{\hat{V}_p(\hat{\tau})}\ t_{\frac{\alpha}{2},n-1}\right).$$

Average length of the Confidence Interval in Method 2 comes out to be $t_{\frac{\alpha}{2},n-1}$ which is smaller than that obtained from Method 1 which is $\frac{1}{\sqrt{\alpha}}$ for all n (considering $n \geq 3$).

3. NUMERICAL PRESENTATION

Consider the following hypothetical population consisting of *N*=37 households. The values corresponding to A and B are '*y*' and '*x*' respectively where A is the 'monthly expenditure on household' and B is the 'necessary medical expenses of the household'. Let '*w*', the 'number of household members' taken as the size measure for sample selection.

1000 samples each of size *n*=11 are chosen by employing a sampling scheme by Seth (1966) as described by Chaudhuri and Pal (2002). In this sampling scheme, the first two units are chosen according to Brewer (1963) and the next 9 units following Seth (1966). The first unit *i* is chosen with a probability proportional to

$$q_i = \frac{p_i(1-p_i)}{1-2p_i} \text{ where } p_i = \frac{w_i}{\sum_{i=1}^{N} w_i}.$$

From the remaining units, a second unit $j(\neq i)$ is chosen with a probability $\frac{p_j}{1-p_i}$.

For this scheme, π_i and π_{ij}'s based on the first two draws are

$$\pi_i(2)=2p_i \qquad (3.1)$$

$$\text{and } \pi_{ij}(2)=\frac{2p_ip_j}{1+D}\left(\frac{1}{1-2p_i}+\frac{1}{1-2p_j}\right)$$

$$\text{where } D = \sum_{i=1}^{N} \frac{p_i}{1-2p_i} \qquad (3.2)$$

The next *n*-2 = 9 units are chosen from the remaining *N*-2 = 35 units by SRSWOR as done by Seth (1966). For the above sampling scheme of choosing *n* units out of *N*, the following were derived (cf Chaudhuri and Pal 2002) for the π_i and π_{ij}'s based on n draws:

$$\pi_i(n)=\frac{1}{N-2}\left[(n-2)+(N-n)\pi_i(2)\right] \qquad (3.3)$$

$$\text{and } \pi_{ij}(n) = \pi_{ij}(2) + \left(\frac{n-2}{N-2}\right)[\pi_i(2) + \pi_j(2) - 2\pi_{ij}(2)]$$

$$+\left(\frac{n-2}{N-2}\right)\left(\frac{n-3}{N-3}\right)[1-\pi_i(2) - \pi_j(2) + \pi_{ij}(2)]. \qquad (3.4)$$

Clearly, third and fourth order inclusion probabilities are also required for our calculations. We have further derived:

$$\pi_{ijk}(n)=\left(\frac{n-2}{N-2}\right)[\pi_{ij}(2)+\pi_{ik}(2)+\pi_{jk}(2)]$$

$$+\left(\frac{n-2}{N-2}\right)\left(\frac{n-3}{N-3}\right)\begin{matrix}[\pi_i(2)+\pi_j(2)+\pi_k(2)\\ -2\pi_{ij}(2)-2\pi_{ik}(2)-2\pi_{jk}(2)]\end{matrix}$$

$$+\left(\frac{n-2}{N-2}\right)\left(\frac{n-3}{N-3}\right)\left(\frac{n-4}{N-4}\right)\begin{matrix}[1-\pi_i(2)-\pi_j(2)-\pi_k(2)\\ +\pi_{ij}(2)+\pi_{ik}+2\pi_{jk}(2)]\end{matrix} \qquad (3.5)$$

$$\text{and}\pi_{ijkl}(n)= \left(\frac{n-2}{N-2}\right)\left(\frac{n-3}{N-3}\right)\begin{matrix}[\pi ij(2)+\pi ik(2)+\pi il(2)\\ +\pi jk(2)+\pi jl(2)+\pi kl(2)]\end{matrix}$$

$$+\left(\frac{n-2}{N-2}\right)\left(\frac{n-3}{N-3}\right)\left(\frac{n-4}{N-4}\right)\begin{matrix}[\pi i(2)+\pi j(2)+\pi k(2)+\pi l(2)\\ -2\pi ij(2)-2\pi ik(2)-2\pi il(2)\\ -2\pi jk(2)-2\pi jl(2)-2\pi kl(2)]\end{matrix}$$

$$+\left(\frac{n-2}{N-2}\right)\left(\frac{n-3}{N-3}\right)\left(\frac{n-4}{N-4}\right)\left(\frac{n-5}{N-5}\right)\begin{matrix}[1-\pi i(2)-\pi j(2)-\pi k(2)\\ -\pi l(2)+\pi ij(2)+\pi ik+\pi il(2)\\ +\pi jk(2)+\pi jl(2)+\pi kl(2)]\end{matrix} \qquad (3.6)$$

such that $\sum_{k(\neq i,j)}^{N} \pi_{ijk}(n)=(n-2)\ \pi_{ij}(n)$

and $\sum_{l(\neq i,j,k)}^{N} \pi_{ijkl}(n)=(n-3)\ \pi_{ijk}(n)$.

We calculate $\hat{\tau}$, $\hat{V}_p(\hat{\tau})$, Coefficient of Variation (CV) = $100\frac{\sqrt{\hat{V}_p(\hat{\tau})}}{\hat{\tau}}$ and 95% approximate Confidence Intervals (CI) for τ by both the methods 1 and 2 for all the 1000 samples. Then based on the 1000 samples we calculate:

ACV (Average Coefficient of Variation) = the average of the coefficient of variation over the 1000 replicates,

Table 1

Unit	w	y(Rs.)	x(Rs.)	Unit	w	y(Rs.)	x(Rs.)	Unit	w	y(Rs.)	x(Rs.)	Unit	w	y(Rs.)	x(Rs.)
1	5	5700	2700	11	3	3030	1650	21	3	3885	1573	31	2	3615	1402
2	6	4020	1875	12	4	2835	1239	22	2	2175	1277	32	9	11062	4500
3	2	2145	1200	13	3	2775	1425	23	5	2436	1110	33	4	4200	1589
4	3	2190	1320	14	5	3510	1680	24	1	1695	900	34	7	9200	3999
5	4	4500	2100	15	2	2730	1360	25	5	2115	947	35	8	8125	3000
6	5	3210	1350	16	2	4080	1500	26	5	3105	1260	36	3	3135	1125
7	3	3600	1877	17	5	4600	1429	27	9	6037	2748	37	2	2910	1307
8	2	2199	975	18	6	10375	2751	28	5	3255	1426				
9	5	2790	1275	19	2	4230	1453	29	9	13500	7998				
10	2	2400	1353	20	2	2625	1155	30	5	3120	1479				

ARB (Absolute Relative Bias) = $\left|\frac{\bar{e}-\tau}{\tau}\right|$, where $\bar{e} = \frac{1}{1000}\sum_{i=1}^{1000}\hat{\tau}_i$, $\hat{\tau}_i$ being the ithsample estimate of τ,

ACP (Actual Coverage Proportion) = percentage of replicates out of 1000 for which the CI covers τ and

AL (Average Length) = average length of the CI over 1000 replicates.

AVE (Avergae Variance estimate) $= \frac{1}{1000}\sum_{i=1}^{1000}\hat{V}_p(\hat{\tau})_i$.

The results are tabulated below:

4. SUMMARY TABLE FINDINGS FOR ACCURACY IN ESTIMATION

Table 2

τ = 0.718, V_p $(\hat{\tau})$ = 0.0145	
ACV	16.550
ARB	0.00261
ACP(method 1)	95.10%
AL(method 1)	0.993
ACP(method 2)	84.70%
AL(method 2)	0.435
AVE	0.0142

Table 3. A few out of the 1000 sample estimates of τ(= 0.718)

0.742	0.740	0.728	0.714	0.690	0.737	0.662	0.695	0.731	0.732
0.662	0.673	0.694	0.665	0.724	0.741	0.743	0.691	0.747	0.704
0.691	0.734	0.694	0.732	0.744	0.729	0.700	0.701	0.661	0.740
0.724	0.733	0.737	0.677	0.665	0.725	0.700	0.737	0.724	0.702
0.661	0.698	0.748	0.688	0.693	0.694	0.709	0.733	0.669	0.733
0.665	0.719	0.702	0.732	0.737	0.684	0.701	0.746	0.710	0.679
0.665	0.661	0.727	0.724	0.662	0.685	0.699	0.741	0.734	0.708
0.729	0.703	0.660	0.740	0.700	0.661	0.690	0.661	0.693	0.695
0.748	0.722	0.735	0.699	0.742	0.690	0.700	0.677	0.660	0.674
0.728	0.745	0.661	0.671	0.698	0.727	0.690	0.698	0.710	0.726

5. CONCLUSION

From the above tables it can be concluded that the proposed estimator is not only good but also provides a very accurate estimate of its variance as well as coefficient of variation. The relative bias of the estimate is extremely low which is desirable. Estimation of Kendall's Rank Correlation Coefficient for a finite population is worth applying because accuracy of the estimator is now easy to calculate. Although ACP calculated using Confidence Interval for Tau by using Chebychev's Inequality is closer to 95% than that calculated by assuming Normality, AL is always much smaller while using method 2 than the case when method 1 is used.

REFERENCES

Brewer, K.R.W. (1963). A model of systematic sampling with unequal probabilities. *Austr. J. Statist.*, **5,** 5-13.

Chaudhuri, A. (2010). *Essentials of Survey Sampling.* Prentice Hall Publications Private Limited.

Chaudhuri, A. and Pal, S. (2002). On certain alternative mean square error estimators in complex survey sampling. *J. Statist. Plan. Inf.*, **104,** 363-375.

Dubey, A. and Gangopadhyay, S. (1998). *Counting the Poor, Sarvekshana Analytical Report Number* 1. Department of Statistics, Govt. of India.

Kendall, M.G. (1955). *Rank Correlation Methods.* Charles Griffin and Company Limited.

Seth, G.R. (1966). On estimators of variance of estimate of population total in varying probabilities. *J. Ind. Soc. Agric. Statist.*, **18(2),** 52-56.