RASHI 2 (1) : 06-17 (2017)

Estimating Rank Correlation in Sample Surveys

Arijit Chaudhuri and Tathagata Dutta

Indian Statistical Institute, Kolkata
and Indian Institute of Technology, Kanpur.

ABSTRACT

If a sample is taken with unequal selection probabilities from a finite population and sample ranks are obtained, it is illustrated how rank correlation coefficients may suitably be estimated, thereby providing appropriate tools for inference making.

AMS Subject Classication : 62 D05

Keywords : Rank correlation, unequal probability sampling

1. Introduction

Dubey and Gangopadhyay(1998) in the context of an Indian exercise in ‘‘Counting the poor” dealt with ‘rank correlation coefficients’. This led Chaudhuri and Shaw (2016) to attend to a possible efficacy in employing rank correlation coefficients in inference making when sampling may be from finite populations permitting ‘unequal probability selection’. The relevant variance estimation is a crucial problem here. Chaudhuri and Shaw (2016) have provided a limited solution. We present here a follow-up we consider worthy of attention in the next two sections.

2. Estimating Spearman’s rank correlation coefficient and Kendall’s τ in finite populations from sample survey data.

Given a finite population U = (1, ..., i,...., N) on which two real variables (x, y), are defined having values (x_i, y_i), i = 1, ..., N the product-moment correlation coefficient R_N between x and y is then given

$$R_N = \frac{\sum_1^N (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_1^N (x_i - \bar{X})^2}\sqrt{\sum_1^N (y_i - \bar{Y})^2}}, \text{ on writing}$$

$$\bar{X} = \frac{1}{N}\sum_1^N x_i \text{ and } \bar{Y} = \frac{1}{N}\sum_1^N y_i.$$

This simplifies to

$$R_N = \frac{N\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{N\sum x_i^2 - (\sum x_i)^2}\sqrt{N\sum y_i^2 - (\sum y_i)^2}}$$

If a sample s of units of U is selected with a probability $p(s)$ for which 1st and 2nd order inclusion probabilities $\pi_i = \sum_{s\ni i} p(s)$ and $\pi_{ij} = \sum_{s\ni i,j} p(s)$ are positive, then for a population total $L = \sum_{i=1}^N l_i$ an unbiased estimator

$$t = \sum_{i\in s} l_i b_{si}$$

is usually employed taking b_{si}’s as constants free of $\underline{L} = (l_1..., l_i, ..., l_N)$ subject to $\sum_{s\ni i} p(s) b_{si} = 1 \forall i \in U$.

The variance of t may then be unbiasedly estimated if $\pi_{ij} > 0 \ \forall i \neq j$. For the six totals

$$\text{N} = \sum 1, \sum_1^N x_i, \sum_1^N y_i, \sum_1^N x_i^2, \sum_1^N y_i^2, \sum_1^N x_i y_i$$

involved in R_N the same estimator of the form t above may be employed to take a nonlinear function of six unbiased estimator t_j (j = 1, ..., 6) to propose an estimator, as $r = f(t_1, t_2, t_3, t_4, t_5, t_6)$ to estimate R_N, the same function f(..., ..., ..., ..., ..., ...,) of the six corresponding population totals.

E-mail : manoj_iasri@yahoo.com

If the individuals i in U are ranked according to their (x, y) values with u_i, v_i as their ranks respectively, then taking $d_i = u_i - v_i$, Spearman's rank correlation coefficient, R_N with $x_i = u_i$ and $y_i = v_i$ is given by the formula

$$R_{sp} = 1 - \frac{6\sum_{1}^{N} d_i^2}{N(N^2-1)}$$

But as noted by Chaudhuri and Shaw(2016) this cannot be estimated as r provides an estimator for R_N.

But as a saving grace using the ranks u_i, v_i for $i = 1, \ldots, N$ Kendall's rank correlation τ defined as

$$\tau = \left(\sum_{\substack{i=1 \\ i<j}}^{N} \sum_{j=1}^{N} a_{ij} b_{ij} \right) \Big/ \left[\sqrt{\sum_{i<j}^{N} \sum a_{ij}^2} \sqrt{\sum \sum_{i<j} b_{ij}^2} \right]$$

on writing

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

and

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

For simplicity, let $\underline{\theta} = (\theta_1, \theta_2, \theta_3)$,

$$\theta_1 = \sum_{i<j}^{N} \sum_{i<j}^{N} a_{ij} b_{ij}, \ \theta_2 = \sum_{i<j}^{N} \sum_{i<j}^{N} a_{ij}^2, \ \theta_3 = \sum \sum_{i<j} b_{ij}^2$$

and

$$\tau = f(\underline{\theta}), \text{ say,}$$

$$= \frac{\theta_1}{\sqrt{\theta_2 \theta_3}}$$

Let a sample s be taken with n units $(2 \le n < N)$, each distinct, with probability $p(s)$ and $(u'_1, \ldots, u'_n)$, $(v'_1, \ldots, v'_n)$ be the ranks of those individuals according to x and y.

Then, using these u'_i, v'_i for $i = 1, \ldots, n$, θ_1, θ_2 and θ_3 may be unbiasedly estimated using t_j in terms of them for $j = 1; 2; 3$. Consequently τ may be estimated by $\hat{\tau} = f(t_1, t_2, t_3)$.

3. Chaudhuri and Shaw's (2016) estimation of Kendall's τ and its further study.

Chaudhuri and Shaw (2016) considered only Horvitz and Thompson's (1952) form of t as, say, $t_{HT} = \sum_{i \in s} \frac{l_i}{\pi_i}$ for $L = \sum_{1}^{N} l_i$ and studied the corresponding special case of $\hat{\tau}$ and examined the corresponding variance estimation problem. This necessitated the use of 3rd and 4th order inclusion-probabilities $\pi_{ijk} = \sum_{s \ni i,j,k} p(s)$ and $\pi_{ijkl} = \sum_{s \ni i,j,k,l} p(s)$ for i, j, k, l in U with $i \ne j \ne k \ne l$. Though they presented extensive numerical data including results for simulation based confidence intervals for τ, average coverage percentages, average estimated coefficients of variation, average lengths of confidence intervals *etc.*, it was evident that the results from the study cannot be used efficiently especially because employing a sampling scheme admitting positive-valued inclusion probabilities of the first four orders is cumbersome.

We are therefore interested to examine alternative procedures to estimate $\tau = \frac{\theta_1}{\sqrt{\theta_2 \theta_3}}$ in case, for example, a sample of n distinct units of U be chosen following Rao, Hartley and Cochran's (1962) (RHC) scheme and employing (1) RHC estimator $t_{RHC} = \sum_{n} \frac{Q_i}{P_i} l_i$ for L and (2) also t for L based on an RHC sample.

To cover (1), one may proceed in a straight-forward way but to cover (2), we find it convenient to utilize Chaudhuri, Bose and Dihidar's (2009) results concerning uses of Horvitz & Thompson's estimators from RHC-samples. We present the relevant calculations in short in the Appendix Section.For now, we may place on record the following gist of our numerical exercise.

We consider $N = 79$, $n = 23$, draw z_i-values from a distribution with density $g(z) = e^{-(z-2)}$, $z \ge 2$, $i = 1 \ldots, N$, draw (x, y)-values from the bivariate normal distribution $N_2(0, 0, 1, 1, 0.67)$. Then take an RHC sample of 23 units using z_i's as size-measures and then work out u'_i, v'_i, for $i = 1, 2, \ldots, 23$ and also u_i, v_i for $i = 1, \ldots, 79$ to observe specimen values of τ using both the estimators based on -

1) RHC estimator 2) HT estimator

to cover both the cases.

Appendix

Estimation of Kendall's τau in RHC Scheme employing:

1. An estimator *t* based on t_{RHC} :

As we assumed RHC scheme, where

$$\theta_1 = \sum_{i<i'}^{n} \sum_{j=1}^{N_i} \sum_{k=1}^{N_{i'}} a_{ij i'_k} b_{ij i'_k}$$

$$\theta_2 = \sum_{i<i'}^{n} \sum_{j=1}^{N_i} \sum_{k=1}^{N_{i'}} a^2_{ij i'_k}$$

$$\theta_3 = \sum_{i<i'}^{n} \sum_{j=1}^{N_i} \sum_{k=1}^{N_{i'}} b^2_{ij i'_k}$$

Let us use the following estimators to estimate θ_1, θ_2, & θ_3, respectively

$$t_1 = \sum_{\substack{i_j \\ i<i'}} \sum_{i'_k \in S} \frac{a_{ij i'_k} b_{ij i'_k}}{\frac{p_{ij}}{Q_i} \cdot \frac{p_{i'k}}{Q_{i'}}} \cdot \frac{N(N-N_i)}{N_i N_{i'}}$$

$$t_2 = \sum_{\substack{i_j \\ i<i'}} \sum_{i'_k \in S} \frac{a^2_{ij i'_k}}{\frac{p_{ij}}{Q_i} \cdot \frac{p_{i'k}}{Q_{i'}}} \cdot \frac{N(N-N_i)}{N_i N_{i'}}$$

$$t_3 = \sum_{\substack{i_j \\ i<i'}} \sum_{i'_k \in S} \frac{b^2_{ij i'_k}}{\frac{p_{ij}}{Q_i} \cdot \frac{p_{i'k}}{Q_{i'}}} \cdot \frac{N(N-N_i)}{N_i N_{i'}}$$

Now

$$E_P(t_1) = E_G[E_C(t_1)] = E_G\left[E_C\left(\sum_{\substack{i_j \\ i<i'}} \sum_{i'_k \in S} \frac{a_{ij i'_k} b_{ij i'_k}}{\frac{p_{ij}}{Q_i} \cdot \frac{p_{i'k}}{Q_{i'}}} \cdot \frac{N(N-N_i)}{N_i N_{i'}}\right)\right]$$

$$= E_G\left[\sum_{i<i'}^{n} \sum_{j=1}^{N_i} \sum_{k=1}^{N_{i'}} \frac{a_{ij i'_k} b_{ij i'_k}}{\frac{p_{ij}}{Q_i} \cdot \frac{p_{i'k}}{Q_{i'}}} \cdot \frac{p_{ij}}{Q_i} \cdot \frac{p_{i'k}}{Q_{i'}} \cdot \frac{N(N-N_i)}{N_i N_{i'}}\right]$$

$$= E_G\left[\sum_{i<i'}^{n} \sum_{j=1}^{N_i} \sum_{k=1}^{N_{i'}} \frac{a_{ij i'_k} b_{ij i'_k} (N-N_i).N}{N_i N_{i'}}\right]$$

$$= \sum_{i<i'}^{n} \sum_{j=1}^{N_i} \sum_{k=1}^{N_{i'}} a_{ij i'_k} b_{ij i'_k} \cdot \frac{N(N-N_i)}{N_i N_{i'}} \cdot \frac{N_i N_{i'}}{N(N-N_i)}$$

$$= \theta_1$$

Similarly,

$E_P(t_2) = \theta_2$, $E_P(t_3) = \theta_3$

Now, by Taylor Series expansion, we know,

$$f(\underset{\sim}{t}) \approx f(\underset{\sim}{\theta}) + \sum_{j=1}^{3} \lambda_j (t_j - \theta_j) \quad \text{[Ignoring the higher order terms]}$$

where, $\lambda_j = \dfrac{\partial f(\underset{\sim}{t})}{\partial t_j}\Big|_{\underset{\sim}{t}=\underset{\sim}{\theta}}$

Hence,

$$\begin{aligned} E_P(f(\underset{\sim}{t})) &\approx E_P(f(\underset{\sim}{\theta})) \quad [\because E(t_j - \theta_j) = 0 \ \forall \ j = 1,2,3] \\ &= f(\underset{\sim}{\theta}) \end{aligned}$$

Hence, we find,

$$\widehat{\tau} = \widehat{f(\underset{\sim}{\theta})} = f(\underset{\sim}{t}) = \frac{t_1}{\sqrt{t_2 t_3}}$$

$$= \frac{\sum\limits_{\substack{i_j \\ i<i'}} \sum\limits_{i'_k \in S} \dfrac{a_{i_j i'_k} b_{i_j i'_k}}{\left(\frac{p_{ij}}{Q_i} \cdot \frac{p_{i'k}}{Q_{i'}}\right)} \cdot \dfrac{N(N-N_i)}{N_i N_{i'}}}{\sqrt{\sum\limits_{\substack{i_j \\ i<i'}} \sum\limits_{i'_k \in S} \dfrac{a^2_{i_j i'_k}}{\frac{p_{ij}}{Q_i} \cdot \frac{p_{i'k}}{Q_{i'}}} \cdot \dfrac{N(N-N_i)}{N_i N_{i'}}} \sqrt{\sum\limits_{\substack{i_j \\ i<i'}} \sum\limits_{i'_k \in S} \dfrac{b^2_{i_j i'_k}}{\frac{p_{ij}}{Q_i} \cdot \frac{p_{i'k}}{Q_{i'}}} \cdot \dfrac{N(N-N_i)}{N_i N_{i'}}}}$$

as an estimator for τ & it is approximately unbiased for τ.

Calculation of $V_P(\hat{\tau})$:

Using Taylor Series expansion & neglecting higher order terms, we get approximately,

$$f(\underset{\sim}{t}) = f(\underset{\sim}{\theta}) + \sum_{j=1}^{3} \frac{\partial f(\underset{\sim}{t})}{\partial t_j}\Big|_{\underset{\sim}{t}=\underset{\sim}{\theta}} (t_j - \theta_j)$$

Hence, $$V_P\{f(\underset{\sim}{t})\} = V_P\left\{ \frac{\partial f(\underset{\sim}{t})}{\partial t_1}\Big|_{\underset{\sim}{t}=\underset{\sim}{\theta}} t_1 + \frac{\partial f(\underset{\sim}{t})}{\partial t_2}\Big|_{\underset{\sim}{t}=\underset{\sim}{\theta}} t_2 + \frac{\partial f(\underset{\sim}{t})}{\partial t_3}\Big|_{\underset{\sim}{t}=\underset{\sim}{\theta}} t_3 \right\}$$

Now,

$$\begin{aligned} \frac{\partial f(\underset{\sim}{t})}{\partial t_1}\Big|_{\underset{\sim}{t}=\underset{\sim}{\theta}} &= \frac{1}{\sqrt{\theta_2 \theta_3}} \\ \frac{\partial f(\underset{\sim}{t})}{\partial t_2}\Big|_{\underset{\sim}{t}=\underset{\sim}{\theta}} &= \frac{-\theta_1}{2\theta_2 \sqrt{\theta_2 \theta_3}} \\ \frac{\partial f(\underset{\sim}{t})}{\partial t_3}\Big|_{\underset{\sim}{t}=\underset{\sim}{\theta}} &= \frac{-\theta_1}{2\theta_3 \sqrt{\theta_2 \theta_3}} \end{aligned}$$

Now,
$$\begin{aligned}V_P\{f(\underline{t})\} &= E_G\{V_C(f(\underline{t}))\} + V_G\{E_C(f(\underline{t}))\}\\ &= E_G[V_C\{f(\underline{t})\}] \quad [\because V_G[E_C\{f(\underline{t})\}] \approx V_G[\text{constant}] = 0]\\ &= E_G\left[V_C\left\{\frac{1}{\sqrt{\theta_2\theta_3}}t_1 + \frac{-\theta_1}{2\theta_2\sqrt{\theta_2\theta_3}}t_2 + \frac{-\theta_1}{2\theta_3\sqrt{\theta_2\theta_3}}t_3\right\}\right]\\ &= E_G\left[V_C\left\{\frac{1}{\sqrt{\theta_2\theta_3}}t_1\right\} + V_C\left\{\frac{-\theta_1}{2\theta_2\sqrt{\theta_2\theta_3}}t_2\right\} + V_C\left\{\frac{-\theta_1}{2\theta_3\sqrt{\theta_2\theta_3}}t_3\right\}\right.\\ &\quad +2\left\{\mathrm{Cov}_C\left(\frac{1}{\sqrt{\theta_2\theta_3}}t_1, \frac{-\theta_1}{2\theta_2\sqrt{\theta_2\theta_3}}t_2\right) + \mathrm{Cov}_C\left(\frac{1}{\sqrt{\theta_2\theta_3}}t_1, \frac{-\theta_1}{2\theta_3\sqrt{\theta_2\theta_3}}t_3\right)\right.\\ &\quad \left.\left.+\mathrm{Cov}_C\left(\frac{-\theta_1}{2\theta_2\sqrt{\theta_2\theta_3}}t_2, \frac{-\theta_1}{2\theta_3\sqrt{\theta_2\theta_3}}t_3\right)\right\}\right]\end{aligned}$$

Now,

$$E_G\left[V_C\left\{\frac{1}{\sqrt{\theta_2\theta_3}}t_1\right\}\right]$$
$$= \frac{1}{\theta_2\theta_3}\sum_{i<i'}^{n}\frac{(N_i-1)(N_{i'}-1).N(N-N_i)}{(N-1)(N-N_i-1)N_iN_{i'}}\sum_{j<j'}^{N_i}\sum_{k<k'}^{N_{i'}}\frac{p_{ij}p_{i'k}}{Q_iQ_{i'}}.\frac{p_{ij'}p_{i'k'}}{Q_iQ_{i'}}\left(\frac{a_{i_ji'_k}b_{i_ji'_k}}{\frac{p_{ij}}{Q_i}.\frac{p_{i'k}}{Q_{i'}}} - \frac{a_{i_{j'}i'_{k'}}b_{i_{j'}i'_{k'}}}{\frac{p_{ij'}}{Q_i}.\frac{p_{i'k'}}{Q_{i'}}}\right)^2$$

$$E_G\left[V_C\left\{\frac{-\theta_1}{2\theta_2\sqrt{\theta_2\theta_3}}t_2\right\}\right]$$
$$= \frac{{\theta_1}^2}{4{\theta_2}^2\theta_3}\sum_{i<i'}^{n}\frac{(N_i-1)(N_{i'}-1).N(N-N_i)}{(N-1)(N-N_i-1)N_iN_{i'}}\sum_{j<j'}\sum_{k<k'}\frac{p_{ij}}{Q_i}.\frac{p_{i'k}}{Q_{i'}}.\frac{p_{ij'}}{Q_i}.\frac{p_{i'k'}}{Q_{i'}}\left(\frac{a_{i_ji'_k}{}^2}{\frac{p_{ij}}{Q_i}.\frac{p_{i'k}}{Q_{i'}}} - \frac{a_{i_{j'}i'_{k'}}{}^2}{\frac{p_{ij}}{Q_i}.\frac{p_{i'k'}}{Q_{i'}}}\right)^2$$

$$E_G\left[V_C\left\{\frac{-\theta_1}{2\theta_3\sqrt{\theta_2\theta_3}}t_3\right\}\right]$$
$$= \frac{{\theta_1}^2}{4\theta_2{\theta_3}^2}\sum_{i<i'}^{n}\frac{(N_i-1)(N_{i'}-1).N(N-N_i)}{(N-1)(N-N_i-1)N_iN_{i'}}\sum_{j<j'}\sum_{k<k'}\frac{p_{ij}}{Q_i}.\frac{p_{i'k}}{Q_{i'}}.\frac{p_{ij'}}{Q_i}.\frac{p_{i'k'}}{Q_{i'}}\left(\frac{b_{i_ji'_k}{}^2}{\frac{p_{ij}}{Q_i}.\frac{p_{i'k}}{Q_{i'}}} - \frac{b_{i_ji'_{k'}}{}^2}{\frac{p_{ij}}{Q_i}.\frac{p_{i'k'}}{Q_{i'}}}\right)^2$$

$$E_G\left[\mathrm{Cov}_C\left(\frac{1}{\sqrt{\theta_2\theta_3}}t_1, \frac{-\theta_1}{2\theta_2\sqrt{\theta_2\theta_3}}t_2\right)\right]$$
$$= \frac{-\theta_1}{2{\theta_2}^2\theta_3}\left[\sum_{i<i'}^{n}\sum_{j=1}^{N_i}\sum_{k=1}^{N'_i}t_1t_2.\frac{p_{ij}}{Q_i}.\frac{p_{i'k}}{Q_{i'}}.\frac{N_iN_{i'}}{N(N-N_i)} - \sum_{i<i'}^{n}\frac{N(N-N_i)}{N_iN_{i'}}.\theta_1\theta_2\right]$$

$$E_G\left[\mathrm{Cov}_C\left(\frac{1}{\sqrt{\theta_2\theta_3}}t_1, \frac{-\theta_1}{2\theta_3\sqrt{\theta_2\theta_3}}t_3\right)\right]$$
$$= \frac{-\theta_1}{2\theta_2{\theta_3}^2}\left[\sum_{i<i'}^{n}\sum_{j=1}^{N_i}\sum_{k=1}^{N'_i}t_1t_3.\frac{p_{ij}}{Q_i}.\frac{p_{i'k}}{Q_{i'}}.\frac{N_iN_{i'}}{N(N-N_i)} - \sum_{i<i'}^{n}\frac{N(N-N_i)}{N_iN_{i'}}.\theta_1\theta_3\right]$$

$$E_G\left[\mathrm{Cov}_C\left(\frac{-\theta_1}{2\theta_2\sqrt{\theta_2\theta_3}}t_2, \frac{-\theta_1}{2\theta_3\sqrt{\theta_2\theta_3}}t_3\right)\right]$$
$$= \frac{{\theta_1}^2}{2{\theta_2}^2{\theta_3}^2}\left[\sum_{i<i'}^{n}\sum_{j=1}^{N_i}\sum_{k=1}^{N'_i}t_2t_3.\frac{p_{ij}}{Q_i}.\frac{p_{i'k}}{Q_{i'}}.\frac{N_iN_{i'}}{N(N-N_i)} - \sum_{i<i'}^{n}\frac{N(N-N_i)}{N_iN_{i'}}.\theta_2\theta_3\right]$$

Estimation of $V_P(\hat{\tau})$:

We separately estimate the three covariance terms & the three variance terms present in $V_P(\hat{\tau})$. We use the estimators $\hat{V}_1, \hat{V}_2, \hat{V}_3$ to estimate $E_G[\text{Var}_C(t_1)]$, $E_G[\text{Var}_C(t_2)]$, $E_G[\text{Var}_C(t_3)]$,

where,

$$\widehat{V_1} = \frac{1}{n-1} \times \frac{\sum_{i<i'} f(N, N_i, N_{i'})}{\sum_{i<i'} f(N, N_i, N_{i'}) - 1}(t_1^2 - e_1)$$

$$\widehat{V_2} = \frac{1}{n-1} \times \frac{\sum_{i<i'} f(N, N_i, N_{i'})}{\sum_{i<i'} f(N, N_i, N_{i'}) - 1}(t_2^2 - e_2)$$

$$\widehat{V_3} = \frac{1}{n-1} \times \frac{\sum_{i<i'} f(N, N_i, N_{i'})}{\sum_{i<i'} f(N, N_i, N_{i'}) - 1}(t_3^2 - e_3)$$

$$e_1 = \sum_{\substack{i_j \\ i<i'}} \sum_{i'_k \in S} \frac{a_{ij i'_k}^2 b_{ij i'_k}^2}{p_{ij}^2 . p_{i'k}^2} . Q_i Q_{i'} . \frac{N(N-N_i)}{N_i N_{i'}}$$

$$e_2 = \sum_{\substack{i_j \\ i<i'}} \sum_{i'_k \in S} \frac{a_{ij i'_k}^4}{p_{ij}^2 . p_{i'k}^2} . Q_i Q_{i'} . \frac{N(N-N_i)}{N_i N_{i'}}$$

$$e_3 = \sum_{\substack{i_j \\ i<i'}} \sum_{i'_k \in S} \frac{b_{ij i'_k}^4}{p_{ij}^2 . p_{i'k}^2} . Q_i Q_{i'} . \frac{N(N-N_i)}{N_i N_{i'}}$$

$$f(N, N_i, N_{i'}) = \frac{(N_i - 1)(N_{i'} - 1)N(N - N_i)}{(N-1)(N - N_i - 1)N_i N_{i'}}$$

Clearly,

$$E_P(e_1) = \sum_{i<i'}^{n} \sum_{j=1}^{N_i} \sum_{k=1}^{N_{i'}} \frac{a_{ij i'_k}^2 b_{ij i'_k}^2}{p_{ij} p_{i'k}}$$

& $$E_P(t_1^2 - v_1) = \left(\sum_{i<i'}^{n} \sum_{j=1}^{N_i} \sum_{k=1}^{N_{i'}} a_{ij i'_k} b_{ij i'_k} \right)^2$$ [v_1 is such that $E_P(v_1) = V_1$]

which leads us to find an unbiased estimator for

$$\sum_{i<i'}^{n} \sum_{j=1}^{N_i} \sum_{k=1}^{N_{i'}} \frac{\left(a_{ij i'_k} b_{ij i'_k} \right)^2}{p_{ij}\, p_{i'k}} - \left(\sum_{i<i'}^{n} \sum_{j=1}^{N_i} \sum_{k=1}^{N_{i'}} a_{ij i'_k} b_{ij i'_k} \right)^2$$

& hence V_1.

Now, we estimate the covariance terms.

We know,

$$\begin{aligned} \mathrm{Cov}(X,Y) &= E(XY) - E(X)\,E(Y) \\ \Rightarrow \widehat{\mathrm{Cov}(X,Y)} &= XY - \widehat{E(X)\,E(Y)} \end{aligned}$$

In this case,

$$E_G\left[\widehat{\mathrm{Cov}_C(t_1,t_2)}\right] = (t_1t_2 - \widehat{\theta_1\theta_2}) \sum_{i<i'} g(N,N_i,N_{i'})$$

We use that technique to find an approximately unbiased estimator of $E_G[\mathrm{Cov}_C(t_1, t_2)]$.

We find an approximately unbiased estimator of $\widehat{\theta_1\theta_2}$ by Taylor Series expansion.

We know,

$$\begin{aligned} f(t_1,t_2) \approx f(\theta_1,\theta_2) &+ \sum_i \frac{\partial f(\underset{\sim}{t})}{\partial t_i}\Big|_{\underset{\sim}{t}=\underset{\sim}{\theta}} (t_i-\theta_i) \\ &+ \frac{1}{2}\left[\frac{\partial^2 f(\underset{\sim}{t})}{\partial t_1^2}\Big|_{\underset{\sim}{t}=\underset{\sim}{\theta}} (t_1-\theta_1)^2 + 2\frac{\partial^2 f(\underset{\sim}{t})}{\partial t_1\,\partial t_2}\Big|_{\underset{\sim}{t}=\underset{\sim}{\theta}} (t_1-\theta_1)(t_2-\theta_2)\right. \\ &\left. + \frac{\partial^2 f(\underset{\sim}{t})}{\partial t_2^2}\Big|_{\underset{\sim}{t}=\underset{\sim}{\theta}} (t_2-\theta_2)^2\ \right] \end{aligned}$$

which leads to,

$$\widehat{\theta_1\theta_2} = t_1t_2 - \frac{1}{t_2\sqrt{t_2t_3}}\,E_G\left[\widehat{\mathrm{Cov}_C(t_1,t_2)}\right] + \frac{3t_1}{8t_2^2\sqrt{t_2t_3}}\,E_G\left[\widehat{\mathrm{Var}_C}(t_2)\right]$$

which further gives,

$$E_G\left[\widehat{\mathrm{Cov}_C(t_1,t_2)}\right] = \frac{3t_1/8t_2^2\sqrt{t_2t_3}}{1+\frac{\sum_{i<i'} g(N,N_i,N_{i'})}{t_2\sqrt{t_2t_3}}}\,\widehat{V_2}\sum_{i<i'} g(N,N_i,N_{i'})$$

Similarly,

$$E_G\left[\widehat{\mathrm{Cov}_C(t_1,t_3)}\right] = \frac{3t_1/8t_3^2\sqrt{t_2t_3}}{1+\frac{\sum_{i<i'} g(N,N_i,N_{i'})}{t_3\sqrt{t_2t_3}}}\,\widehat{V_3}\sum_{i<i'} g(N,N_i,N_{i'})$$

$$E_G\left[\widehat{\mathrm{Cov}_C(t_2,t_3)}\right] = \frac{3t_1t_2t_3}{8t_2t_3\sqrt{t_2t_3} - t_1\sum_{i<i'} g(N,N_i,N_{i'})} \cdot \left(\frac{\widehat{V_2}}{t_2^2} + \frac{\widehat{V_3}}{t_3^2}\right) \sum_{i<i'} g(N,N_i,N_{i'})$$

where, $$g(N,N_i,N_{i'}) = \frac{N(N-N_i)}{N_iN_{i'}}$$

2. An estimator t′ based on t_{HT} :

Here, in this case, for estimating Kendall's τ, we choose a sample of size *n* by RHC scheme. Then the sample units are ranked w.r.t *X*-data as $u'_1, u'_2, \ldots u'_n$ & w.r.t *Y* -data as $v'_1, v'_2, \ldots v'_n$

$$\tau = \frac{\sum_{i}^{N}\sum_{<j}^{N} a_{ij}b_{ij}}{\sqrt{\sum_{i}^{N}\sum_{<j}^{N} a_{ij}^2}\sqrt{\sum_{i}^{N}\sum_{<j}^{N} b_{ij}^2}} = \frac{\theta_1}{\sqrt{\theta_2\theta_3}} = f(\underset{\sim}{\theta}), \quad \underset{\sim}{\theta} = (\theta_1, \theta_2, \theta_3).$$

Clearly, as we assume RHC scheme, we can write,

$$\theta_1 = \sum_{i<i'}^{n}\sum_{j=1}^{N_i}\sum_{k=1}^{N_{i'}} a_{ij i'_k} b_{ij i'_k}$$

$$\theta_2 = \sum_{i<i'}^{n}\sum_{j=1}^{N_i}\sum_{k=1}^{N_{i'}} a^2_{ij i'_k}$$

$$\theta_3 = \sum_{i<i'}^{n}\sum_{j=1}^{N_i}\sum_{k=1}^{N_{i'}} b^2_{ij i'_k}$$

Let us use the following estimators to estimate θ_1, θ_2, & θ_3 respectively.

$$t'_1 = \sum_{\substack{i_j \\ i<i'}}\sum_{i'_k \in S} \frac{a_{ij i'_k} b_{ij i'_k}}{\pi_{ij i'_k}} \cdot \frac{N(N-N_i)}{N_i N_{i'}}$$

$$t'_2 = \sum_{\substack{i_j \\ i<i'}}\sum_{i'_k \in S} \frac{a^2_{ij i'_k}}{\pi_{ij i'_k}} \cdot \frac{N(N-N_i)}{N_i N_{i'}}$$

$$t'_3 = \sum_{\substack{i_j \\ i<i'}}\sum_{i'_k \in S} \frac{b^2_{ij i'_k}}{\pi_{ij i'_k}} \cdot \frac{N(N-N_i)}{N_i N_{i'}}$$

Case-1 : [*N*/*n*] = *m*

$$\pi_{ij} = \frac{G_2}{G}\left[S^{N-2}_{m-1}\frac{p_i}{p_i + \sum_{l=1}^{m-1} p_{i_l}}\left(S^{N-m-1}_{m-1}\frac{p_j}{p_j + \sum_{t=1}^{m-1} p_{j_t}}\right)\right]$$

$$G = \frac{N!}{(m!)^n\, n!},\ G_2 = \frac{(N-2m)!}{(m!)^{n-2}\,(n-2)!}$$

In this case, *n* disjoint groups, each with *m* units, can be formed from the *N* units in $\frac{N!}{(m!)^n\, n!}$ ways & a group of *m* units containing unit *i* can be formed from the *N* units in $\alpha\,(N-1, m-1)$ ways. Let S_{m-1}^{N-1} denote summation over these $\alpha\,(N-1, m-1)$ groups. Let P_{il} be the normed size measure of the l^{th} unit i_l in the i^{th} group consisting of *m* distinct units ($i_l \neq i$).

Again, the no. of *m*-tuples of distinct units of *U* with unit *i* included but unit *j* not included is equal to $\alpha\,(N-2, m-1)$ while the no. of *m*-tuples ($j = j_1, \ldots, j_{m-1}$) of distinct units of *U*, such that $j \neq i$, $j_t \neq i \neq j \neq i_l$, $l = 1, \ldots, m-1$; $t = 1, \ldots, m-1$ is $\alpha\,(N-m-1, m-1)$. Let S_{m-1}^{N-2} & S_{m-1}^{N-m-1} respectively denote summations over these $\alpha\,(N-2, m-1)$ & $\alpha\,(N-m-1, m-1)$ *m*-tuples. Hence this result.

Case 2 : *N/n* ≠ not an integer; [*N/n*] = *m*.

In this case *K* groups contain *m* units each & $(n-K)$ groups have $(m+1)$ units each. Then,

$$\pi_{ij} = A_1 + A_2 + A_3 + A_4.$$

where,

$$A_1 = \frac{G_2^{(1)}}{G'}\left[S_{m-1}^{N-2}\frac{p_i}{p_i+\sum_{l=1}^{m-1}p_{i_l}}\left(S_{m-1}^{N-m-1}\frac{p_j}{p_j+\sum_{t=1}^{m-1}p_{j_t}}\right)\right]$$

$$A_2 = \frac{G_2^{(2)}}{G'}\left[S_{m-1}^{N-2}\frac{p_i}{p_i+\sum_{l=1}^{m-1}p_{i_l}}\left(S_{m}^{N-m-1}\frac{p_j}{p_j+\sum_{t=1}^{m}p_{j_t}}\right)\right]$$

$$A_3 = \frac{G_2^{(2)}}{G'}\left[S_{m}^{N-2}\frac{p_i}{p_i+\sum_{l=1}^{m}p_{i_l}}\left(S_{m-1}^{N-m-2}\frac{p_j}{p_j+\sum_{t=1}^{m-1}p_{j_t}}\right)\right]$$

$$A_4 = \frac{G_2^{(3)}}{G'}\left[S_{m}^{N-2}\frac{p_i}{p_i+\sum_{l=1}^{m}p_{i_l}}\left(S_{m}^{N-m-2}\frac{p_j}{p_j+\sum_{t=1}^{m}p_{j_t}}\right)\right]$$

$$G' = \frac{N!}{(m!)^k\ K!\ (m+1)!^{n-K}\ (n-K)!}$$

$$G_2^{(1)} = \frac{(N-2m)!}{(m!)^{K-2}\ (K-2)!\ (m+1)!^{n-K}\ (n-K)!}$$

$$G_2^{(2)} = \frac{(N-2m-1)!}{(m!)^{K-1}\ (K-1)!\ (m+1)!^{n-K-1}\ (n-K-1)!}$$

$$G_2^{(3)} = \frac{(N-2m-2)!}{(m!)^{K}\ K!\ (m+1)!^{n-K-2}\ (n-K-2)!}$$

Now,

$$E_p(t_1') = E_G\left[E_C(t_1')\right] = E_G\left[E_C\left(\sum_{i_j}\sum_{\substack{i'_{K\in s}\\ i<i'}}\frac{a_{i_j i'_K}\, b_{i_j i'_K}}{\pi_{i_j i'_K}}\,\frac{(N-N_i)N}{N_i N_{i'}}\right)\right]$$

$$= E_G\left[\sum_{i<i'}^{n}\sum_{j=1}^{N_i}\sum_{K=1}^{N_{i'}}\frac{a_{i_j i'_K}\, b_{i_j i'_K}}{\cancel{\pi_{i_j i'_K}}}\cdot\cancel{\pi_{i_j i'_K}}\cdot\frac{N(N-N_i)}{N_i N_{i'}}\right]$$

$$= \sum_{i<i'}^{n} \sum_{j=1}^{N_i} \sum_{K=1}^{N_{i'}} a_{i_j i'_K} b_{i_j i'_K} \cdot \frac{N(N-N_i)}{N_i N_{i'}} \cdot \frac{N_i N_{i'}}{N(N-N_i)}$$
$$= \sum_{i<i'}^{n} \sum_{j=1}^{N_i} \sum_{K=1}^{N_{i'}} a_{i_j i'_K} b_{i_j i'_K} = \theta_1$$

Similarly, $E_P(t'_2) = \theta_2$, $E_P(t'_3) = \theta_3$

Now, by Taylor Series expansion, we know,

$$f(\underset{\sim}{t}') \approx f(\underset{\sim}{\theta}) + \sum_{j=1}^{3} \lambda_j (t'_j - \theta_j) \text{ [Ignoring the higher order terms]}$$

where, $\lambda_j = \left.\frac{\partial f(\underset{\sim}{t})}{\partial t'_j}\right|_{\underset{\sim}{t}=\underset{\sim}{\theta}}$

Hence, $E_P(f(\underset{\sim}{t})) \approx E_P(f(\underset{\sim}{\theta})) \quad [\because E(t'_j - \theta_j) = 0 \ \forall \ j = 1, 2, 3]$

Hence we find

$$\widehat{\tau} = \widehat{f(\underset{\sim}{\theta})} = f(\underset{\sim}{t}') = \frac{\sum_{i_j} \sum_{\substack{i'_K \in s \\ i<i'}} \frac{a_{i_j i'_K} b_{i_j i'_K}}{\pi_{i_j i'_K}} \cdot \frac{N(N-N_i)}{N_i N_{i'}}}{\sqrt{\sum_{i_j} \sum_{\substack{i'_K \in s \\ i<i'}} a^2_{i_j i'_K} / \pi_{i_j i'_K} \cdot \frac{(N-N_i)N}{N_i N_{i'}}} \sqrt{\sum_{i_j} \sum_{i<i'} \sum_{i'_K} \frac{b^2_{i_j i'_K}}{\pi_{i_j i'_K}} \cdot \frac{N(N-N_i)}{N_i N_{i'}}}}$$

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

Calculation of $V_P(\hat{\tau})$:

In this case, also, we proceed in a similar way to find $V_P(\hat{\tau})$.

We use Taylor Series expansion to find $V_P(\hat{\tau})$.

Clearly,

$$\begin{aligned} V_P\{f(\underset{\sim}{t}')\} &= E_G[V_C\{f(\underset{\sim}{t}')\}] \\ &= E_G\left[V_C\left\{\frac{1}{\sqrt{\theta_2\theta_3}} t'_1\right\} + V_C\left\{\frac{-\theta_1}{2\theta_2\sqrt{\theta_2\theta_3}} t'_2\right\} + V_C\left\{\frac{-\theta_1}{2\theta_3\sqrt{\theta_2\theta_3}} t'_3\right\}\right. \\ &\quad + 2\left\{\mathrm{Cov}_C\left(\frac{1}{\sqrt{\theta_2\theta_3}} t'_1, \frac{-\theta_1}{2\theta_2\sqrt{\theta_2\theta_3}} t'_2\right) + \mathrm{Cov}_C\left(\frac{1}{\sqrt{\theta_2\theta_3}} t'_1, \frac{-\theta_1}{2\theta_3\sqrt{\theta_2\theta_3}} t'_3\right)\right. \\ &\quad \left.\left. + \mathrm{Cov}_C\left(\frac{-\theta_1}{2\theta_2\sqrt{\theta_2\theta_3}} t'_2, \frac{-\theta_1}{2\theta_3\sqrt{\theta_2\theta_3}} t'_3\right)\right\}\right] \end{aligned}$$

Now

$$E_G\left[V_C\left\{\frac{1}{\sqrt{\theta_2\theta_3}}t'_1\right\}\right]$$
$$=\frac{1}{\theta_2\theta_3}\sum_{i<i'}^{n}\frac{(N_i-1)(N_{i'}-1)N(N-N_i)}{(N-1)(N-N_i-1)N_iN_{i'}}\sum_{j<j'}^{N_i}\sum_{k<k'}^{N_{i'}}\pi_{i_ji'_k}\pi_{i_{j'}i'_{K'}}\left(\frac{a_{i_ji'_k}b_{i_ji'_k}}{\pi_{i_ji'_k}}-\frac{a_{i_{j'}i'_{k'}}b_{i_{j'}i'_{k'}}}{\pi_{i_{j'}i'_{K'}}}\right)^2$$

$$E_G\left[V_C\left\{\frac{-\theta_1}{2\theta_2\sqrt{\theta_2\theta_3}}t'_2\right\}\right]$$
$$=\frac{{\theta_1}^2}{4{\theta_2}^3\theta_3}\sum_{i<i'}^{n}\frac{(N_i-1)(N_{i'}-1)N(N-N_i)}{(N-1)(N-N_i-1)N_iN_{i'}}\sum_{j<j'}^{N_i}\sum_{k<k'}^{N_{i'}}\pi_{i_ji'_k}\pi_{i_{j'}i'_{K'}}\left(\frac{a^2_{i_ji'_k}}{\pi_{i_ji'_k}}-\frac{a^2_{i_{j'}i'_{k'}}}{\pi_{i_{j'}i'_{K'}}}\right)^2$$

$$E_G\left[V_C\left\{\frac{-\theta_1}{2\theta_3\sqrt{\theta_2\theta_3}}t'_3\right\}\right]$$
$$=\frac{{\theta_1}^2}{4\theta_2{\theta_3}^3}\sum_{i<i'}^{n}\frac{(N_i-1)(N_{i'}-1)N(N-N_i)}{(N-1)(N-N_i-1)N_iN_{i'}}\sum_{j<j'}^{N_i}\sum_{k<k'}^{N_{i'}}\pi_{i_ji'_k}\pi_{i_{j'}i'_{K'}}\left(\frac{b^2_{i_ji'_k}}{\pi_{i_ji'_k}}-\frac{b^2_{i_{j'}i'_{k'}}}{\pi_{i_{j'}i'_{K'}}}\right)^2$$

$$E_G\left[\mathrm{Cov}_C\left(\frac{1}{\sqrt{\theta_2\theta_3}}t'_1,\frac{-\theta_1}{2\theta_2\sqrt{\theta_2\theta_3}}t'_2\right)\right]$$
$$=\frac{-\theta_1}{2{\theta_2}^2\theta_3}\left[\sum_{i<i'}^{n}\sum_{j=1}^{N_i}\sum_{k=1}^{N'_i}t'_1t'_2.\pi_{i_ji'_k}.\frac{N_iN_{i'}}{N(N-N_i)}-\sum_{i<i'}^{n}\frac{N(N-N_i)}{N_iN_{i'}}\theta_1\theta_2\right]$$

$$E_G\left[\mathrm{Cov}_C\left(\frac{1}{\sqrt{\theta_2\theta_3}}t'_1,\frac{-\theta_1}{2\theta_3\sqrt{\theta_2\theta_3}}t'_3\right)\right]$$
$$=\frac{-\theta_1}{2\theta_2{\theta_3}^2}\left[\sum_{i<i'}^{n}\sum_{j=1}^{N_i}\sum_{k=1}^{N'_i}t'_1t'_3.\pi_{i_ji'_k}.\frac{N_iN_{i'}}{N(N-N_i)}-\sum_{i<i'}^{n}\frac{N(N-N_i)}{N_iN_{i'}}\theta_1\theta_3\right]$$

$$E_G\left[\mathrm{Cov}_C\left(\frac{-\theta_1}{2\theta_2\sqrt{\theta_2\theta_3}}t'_2,\frac{-\theta_1}{2\theta_3\sqrt{\theta_2\theta_3}}t'_3\right)\right]$$
$$=\frac{{\theta_1}^2}{2{\theta_2}^2{\theta_3}^2}\left[\sum_{i<i'}^{n}\sum_{j=1}^{N_i}\sum_{k=1}^{N'_i}t'_2t'_3.\pi_{i_ji'_k}.\frac{N_iN_{i'}}{N(N-N_i)}-\sum_{i<i'}^{n}\frac{N(N-N_i)}{N_iN_{i'}}\theta_2\theta_3\right]$$

Estimation of $V_p(\hat{\tau})$:

Again, in a similar manner, we find the estimates of three variance terms & three covariance terms, which leads us to an approximately unbiased estimator of $V_p(\hat{\tau})$:.

We use $\widehat{V_1'}, \widehat{V_2'}, \widehat{V_3'}, \widehat{C_{12}'}, \widehat{C_{13}'}, \widehat{C_{23}'}$ to estimate $E_G[\mathrm{Var}_C(t_1')]$, $E_G[\mathrm{Var}_C(t_2')]$, $E_G[\mathrm{Var}_C(t_3')]$, $E_G[\mathrm{Cov}_C(t_1', t_2')]$, $E_G[\mathrm{Cov}_C(t_1', t_3')]$, $E_G[\mathrm{Cov}_C(t_2', t_3')]$ respectively.

$$\widehat{V_1'} = \frac{1}{n-1} \times \frac{\sum f(N, N_i, N_{i'})}{\sum f(N, N_i, N_{i'}) - 1}(t_1'^2 - e_1')$$

$$\widehat{V_2'} = \frac{1}{n-1} \times \frac{\sum f(N, N_i, N_{i'})}{\sum f(N, N_i, N_{i'}) - 1}(t_2'^2 - e_2')$$

$$\widehat{V_3'} = \frac{1}{n-1} \times \frac{\sum f(N, N_i, N_{i'})}{\sum f(N, N_i, N_{i'}) - 1}(t_3'^2 - e_3')$$

$$\widehat{C_{12}'} = \frac{3t_1'/8t_2'^2\sqrt{t_2't_3'}}{\left(1 + \frac{\sum_{i<i'} g(N,N_i,N_{i'})}{t_2'\sqrt{t_2't_3'}}\right)} \widehat{V_2'} \sum_{i<i'} g(N, N_i, N_{i'})$$

$$\widehat{C_{13}'} = \frac{3t_1'/8t_3'^2\sqrt{t_2't_3'}}{\left(1 + \frac{\sum_{i<i'} g(N,N_i,N_{i'})}{t_3'\sqrt{t_2't_3'}}\right)} \widehat{V_3'} \sum_{i<i'} g(N, N_i, N_{i'})$$

$$\widehat{C_{23}'} = \frac{3t_1't_2't_3'}{8t_2't_3'\sqrt{t_2't_3'} - t_1' \sum_{i<i'} g(N, N_i, N_{i'})} \cdot \left(\frac{\widehat{V_2'}}{t_2'^2} + \frac{\widehat{V_3'}}{t_3'^2}\right) \sum_{i<i'} g(N, N_i, N_{i'}).$$

Population Tau	RHC estimator	S.e(RHC est)	length of CI (RHC)	HT estimator	S.e (HT est)	length of CI (HT)
0.5507952	0.4632655	0.12048	0.4722816	0.5006925	0.13125	0.5145
0.5806556	0.4255564	0.110525	0.433258	0.468665	0.122854	0.48158768
0.4203181	0.2775423	0.071242	0.27926864	0.2145921	0.056247	0.22048824
0.4644596	0.4195417	0.108935	0.4270252	0.4289857	0.112444	0.44078048
0.3742291	0.4598341	0.119575	0.468734	0.4970397	0.130341	0.51093672
0.4625122	0.3156524	0.081403	0.31909976	0.3261124	0.085485	0.3351012
0.4579682	0.5243403	0.13658	0.5353936	0.4642051	0.121684	0.47700128
0.5475495	0.4229314	0.109833	0.43054536	0.4456906	0.116831	0.45797752
0.4066861	0.454988	0.118295	0.4637164	0.4851398	0.127172	0.49851424
0.396949	0.5469992	0.142546	0.55878032	0.5373155	0.14085	0.552132

REFERENCES

Chaudhuri, A. and Shaw, P. 2016. Accuracy in estimating Kendall's tau in sampling nite populations. *J. Ind. Soc. Agri. Stat*. **70 :** 1-9 .

Chaudhuri, A., Bose, M. and Dihidar, K. 2009. Rao-Hartley and Cochran's sampling with competitive estimators. *Cal. Stat. Assoc. Bull*, 227-42.

Dubey, A. and Gangopadhyay 1998. Counting the poor. Sarvekshana Analytical Report. No.1, Dept. Stat. Govt. of India.

Horvitz, D.G. and Thompson, D.J. 1952. A generalization of sampling without replacement from a nite universe. *J. Amer. Stat. Asso*.**47 :** 663-85.

Rao, JNK, Hartley, H.O. and Cochran, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *J. Roy. Stat.* **24:** 482-91.