Bivariate ridits and distribution theory for Kendall's tau–a

Roger B. Newson
Imperial College London
r.newson@imperial.ac.uk
http://www.rogernewsonresources.org.uk/

February 22, 2019

1 Introduction

This document is motivated by the need to do power calculations for estimation of Kendall's tau–a (or τ_a), using bivariate data. To do power calculations, we need an expression for the sampling variance of the sample Kendall's tau–a in its distribution around the population Kendall's tau–a, which does not have to be zero, even under the null hypothesis used in the power calculations. The sampling distribution of Kendall's tau–a is defined in terms of the bivariate ridit function, which is in turn defined by analogy with the univariate ridit function. We therefore need first to define univariate and bivariate ridits, and then to define Kendall's tau–a in terms of bivariate ridits, and then to define the sampling distribution of the sample Kendall's tau–a, and then to explain how this theory can be used in power calculations.

2 Univariate and bivariate ridits

Given a random variable X with a cumulative distribution function $F_X(\cdot)$, its distribution can be defined alternatively using the Brockett–Levene ridit function[1] $R_X(\cdot)$, defined in turn by

$$R_X(x) \quad = \quad \Pr(X < x) - \Pr(X > x) \quad = \quad E\left[\text{sign}(x - X)\right], \tag{1}$$

where $E[\cdot]$ denotes expectation. This function is defined on a probability–difference scale from -1 to 1. (Note that a third alternative is the Bross ridit function[2], defined, on a scale from 0 to 1, by averaging the Brockett–Levene ridit with 1. Bross stated, in this reference, that the word ridit stood for "with respect to an identified distribution", but later stated that it was named after his wife Rida.)

Given a bivariate random variable (X, Y), with distribution function $F_{X,Y}(\cdot, \cdot)$, we can define the bivariate ridit function $B_{X,Y}(\cdot, \cdot)$ (on a scale from -1 to 1) by the formula

$$\begin{array}{rcl} B_{X,Y}(x,y) & = & E\left[\text{sign}(x - X)\,\text{sign}(y - Y)\right] \\ & = & \Pr\left[\left[\text{sign}(x - X)\,\text{sign}(y - Y) = 1\right] \quad - \quad \Pr\left[\text{sign}(x - X)\,\text{sign}(y - Y) = -1\right]\right., \end{array} \tag{2}$$

or (in other words) as the difference between the probability that the bivariate (X, Y) is concordant with (x, y) and the probability that (X, Y) is discordant with (x, y). Note that the bivariate ridit function does not define the bivariate dstribution, as a univariate ridit function defines its univariate distribution.

The mean bivariate ridit of (X, Y) with respect to its own distribution is known as Kendall's tau–a, defined as

$$\tau_{X,Y} \quad = \quad E\left[B_{X,Y}(X, Y)\right], \tag{3}$$

or (in other words) as the difference between the probabilities of concordance and discordance between 2 bivariate values independently sampled from the joint distribution of X and Y. Kendall's tau–a is available in multiple versions for multiple sampling schemes, discussed in Newson (2002)[3] and Newson (2006)[5]. However, we will concentrate on the case where the bivariate (X, Y)–pairs are sampled independently from a common distribution. Kendall's tau–a is a member of a class of distributional parameters known as regular Hoeffding functionals, whose estimation (using corresponding sample statistics with asymptotically Normal distributions) is discussed in Section 3.2 of Puri and Sen (1971)[6].

3 Power calculations for Kendall's tau–a

We will use the unified power calculations methods of Newson (2004)[4]. These methods use a quantity known as the standard deviation of the influence function, which can be divided by the square root of the sample number to compute the sampling standard error of the sample statistic. In our case, the sample statistic is the sample Kendall's tau–a, which (in the terminology of Section 3.2 of Puri and Sen (1971)[6] s a U–statistic. The standard deviation of the influence function is asymptotically equal to

$$SD_{\inf}(\tau_{XY}) \quad = \quad \sqrt{4Var\left[B_{X,Y}(X,Y)\right]} \quad = \quad 2SD\left[B_{X,Y}(X,Y)\right], \tag{4}$$

where $Var[\cdot]$ and $SD[\cdot]$ denote the population variance and standard deviation, respectively. The key to power calculations is therefore the estimation of the variance and standard deviation of $B_{X,Y}(X,Y)$. This can be done using numerical integration, if we specify a distribution for $B_{X,Y}(X,Y)$.

We will assume that there exist monotonic transformations $g(\cdot)$ and $h(\cdot)$, with the feature that $U = g(X)$ and $V = h(Y)$ are variables with a bivariate standard Normal joint distribution with correlation coefficient ρ. As Kendall's tau–a is invariant under monotonic transformations, it follows that

$$\tau_{X,Y} \quad = \quad \tau_{U,V} \quad = \quad \left(\frac{2}{\pi}\right)\arcsin\left(\rho\right). \tag{5}$$

(This is a consequence of Greiner's relation between Kendall's tau–a and Pearson's rho under a bivariate Normal distribution, discussed in Section 3 of Newson (2002)[3].) This assumption allows the possibility that X and Y are non–Normal, and skewed in one direction or other, but it also implies that there are no 2–way relationships, and that the conditional variance of V given that $U = u$ is the same for all u, and that the conditional variance of U given that $V = v$ is the same for all v.

If we can assume this, then the bivariate ridit of X and Y (with respect to their own bivariate distribution) is equal to the bivariate ridit of U and V (with respect to their own bivariate distribution). Standard Normal bivariate ridits are given by the formula

$$\begin{array}{rcl} B_{U,V}(u,v) & = & \Pr\left[\left[\mathrm{sign}(u-U)\,\mathrm{sign}(v-V) \;=\; 1\right]\right] \quad - \quad \Pr\left[\mathrm{sign}(u-U)\,\mathrm{sign}(v-V) \;=\; -1\right] \\ & = & 2\Pr\left[\left[\mathrm{sign}(u-U)\,\mathrm{sign}(v-V) = 1\right]\right] \quad - \quad 1 \\ & = & 2\left[\Pr(U<u \quad \& \quad V<v) \;+\; \Pr(U>u \quad \& \quad V>v)\right] \quad - \quad 1 \\ & = & 2\left[\Phi(u,v|\rho) \;+\; \Phi(-u,-v|\rho)\right] \quad - \quad 1, \end{array} \tag{6}$$

where $\Phi(\cdot,\cdot|\rho)$ is the cumulative bivariate standard Normal distribution function with Pearson correlation coefficient ρ. (The first equality follows from (2), the second equality follows from the fact that the bivariate standard Normal distribution is continuous, the third equality follows from the definition of the sign function, and the fourth equality follows from the fact that the bivariate standard Normal distribution is symmetrical around 0 in both arguments.)

It follows that Kendall's tau–a between X and Y is given by

$$\tau_{X,Y} \quad = \quad \tau_{U,V} \quad = \quad E\left\{2\left[\Phi(U,V|\rho) \;+\; \Phi(-U,-V|\rho)\right] \quad - \quad 1\right\} \tag{7}$$

and the standard deviation of its influence function is given by

$$SD_{\inf}(\tau_{XY}) \quad = \quad 2SD\left\{2\left[\Phi(U,V|\rho) \;+\; \Phi(-U,-V|\rho)\right] \;-\; 1\right\}. \tag{8}$$

The value of $\tau_{X,Y}$ is alternatively given by the Greiner relation (5). The standard deviation of the influence function is equal to 2/3 if $\rho = 0$ (and therefore $\tau_{X,Y} = 0$), and has lower values for other values of ρ, and is zero in the limit as ρ tends to 1 or -1.

It is possible to define confidence intervals for Kendall's tau–a using Normalizing and variance–stabilizing transformations, such as Daniels' arcsine or Fisher's z–transform (also known as the hyperbolic arctangent). The confidence interval for the transformed Kendall's tau–a will usually have a coverage probability nearer to the advertized level than a confidence interval for the untransformed Kendall's tau–a, and can be back–transformed using the inverse function of the transformation function to give an asymmetric confidence interval for the untransformed Kendall's tau–a. A list of such transformations is given in Table 1 of Newson (2006)[5]. If we do this, then the power calculations have to be for the transformed Kendall's tau–a, and the standard deviation of the influence function (given by twice the standard deviation of the bivariate ridits) must be multiplied by $d\zeta(\tau_{X,Y})/d\tau_{XY}$, where $\zeta(\cdot)$ is the transformation function. Note that, if the null hypothesis being tested specifies a non–zero Kendall's tau–a, then the null tau–a must also be transformed, in order for the detectable difference to be specified correctly as a difference between the alternative and null tau–a values.

3.1 Implementation in the Stata statistical software

If we are using the Stata statistical software[7] to do the power and sample size calculations, then we can estimate $SD_{\text{inf}}(\tau_{XY})$ using numerical integration, and input the result into the `powercal` package of Newson (2004)[4]. The numerical integration is done using the SSC add–on package `expgen` to expand each observation (representing a power–calculation scenario) to a large number of new observations (representing combinations of power–calculation scenarios and values of the bivariate variables U and V). The variables U and V can be generated either by random sampling (using the `runiform()` function) or by using `expgen` to expand the datset to have 1 observation per power calculation scenario per (U, X)-pair, and sampling probability weights proportional to the bivariate standard Normal probability density. We can then use the add–on SSC command `normalbvr` to compute the Normal bivariate ridits. To collapse the dataset to have 1 observation per power–calculation scenario, and data on the mean and standard deviation of the Normal bivariate ridits under that scenario, we use the `collapse` command. We then double the standard deviaton of the Normal bivariate ridits, under each scenario, to compute the standard deviation of the influence function for Kendall's tau–a under that scenario. Some examples are given in the on–line help for `normalbvr`.

References

[1] Brockett PL, Levene A. On a characterization of ridits. *The Annals of Statistics* 1977; **5(6)**: 1245–1248.

[2] Bross IDJ. How to use ridit analysis. *Biometrics* 1958; **14(1)**: 18–38.

[3] Newson R. Parameters behind "nonparametric" statistics: Kendall's tau, Somers' D and median differences. *The Stata Journal* 2002; **2(1)**: 45–64.

[4] Newson R. Generalized power calculations for generalized linear models and more. *The Stata Journal* 2004; **4(4)**: 379–401.

[5] Newson R. Confidence intervals for rank statistics: Somers' D and extensions. *The Stata Journal* 2006; **6(3)**: 309–334.

[6] Puri ML, Sen PK. Nonparametric Methods in Multivariate Statistics. New York: John Wiley & Sons Inc.; 1971.

[7] StataCorp. *Stata: Release 15. Statistical Software.* College Station, TX: StataCorp LLC; 2017.