Kendall Tau, Goodman Gamma

τ_K, γ_G

Simple Definitions:

Let π^c be the probability of a concordant pair, and π^d be the probability of a discordant pair.
For ties:
* π^t_A be the probability of a tie on the A variable (dependent)
* π^t_B be the probability of a tie on the B variable (independent), and
* π^t_{AB} be the probability of a tie on both variables (dependent)

Then

- Kendals τ_a is π^c - π^d
- Goodman's γ is $\frac{\tau_a}{\pi^c+\pi^d}$
- Somers d_{AB} is $\frac{\tau_a}{\pi^c+\pi^d+\pi^t_A}$
- Wilson's e is $\frac{\tau_a}{\pi^c+\pi^d+\pi^t_A+\pi^t_B}$
- Kendall's τ_b is $\frac{\tau_a}{\sqrt{(\pi^c+\pi^d+\pi^t_A)(\pi^c+\pi^d+\pi^t_B)}}$
- Kendall's τ_c is $\frac{\tau_a}{1-\frac{1}{\min\ (n\ rows, n\ cols)}}$
  These are all standard definitions. One can also "Rao-Kupperize" the ties by using a Dale Global Cross Ratio estimating a $2x2$ underlying table.

WRT ties:

- γ conditions on the untied (as if the data were absolutely continuous)
- τ_a doesn't include them in the numerator but does in the denominator
- τ_b and τ_c are intermediate, including some. (Note that, for square tables, $\tau_b \approx \tau_c$)
  In the (realistic) case of discrete or binned data $\gamma_G \geq \tau_b \geq \tau_a$ (for positive associations)
  As the number of bins increases (for a bivariate normal) the measures converge towards ρ.

Note: The variances will depend on the sample size (e.g. γ ignores all ties), so a correlation matrix will have multiple sample sizes, which would screw-up calculating partials. ?Use a harmonic mean?

Converting to ρ.

If one wishes to compare to ρ under the assumption of bivariate normality, use the following relationships:

- $\tau = \frac{2}{\pi} \arcsin \rho$
- $\rho_\tau = \sin \frac{\pi}{2} \tau$ This is also known as Greiner's relation.[1]

Partial τ_a

from NIST Dataplot page:

$$\tau_{12.3} = \frac{\tau_{12} - \tau_{13}\tau_{23}}{\sqrt{(1 - \tau_{13}^2)(1 - \tau_{23}^2)}}$$

from [p. 327, 2]. See also [3].

> *Quade [p 7, 4] notes that*
>
> > *"Kendall and Stuart use this relationship to **define** partial correlation for **any** parent distribution" [p. 317ff, 5]*

test for partial correlation

Same old, same old:

$$t = \tau_{ij|S} \sqrt{\frac{n - 2 - |S|}{1 - \tau_{ij|S}^2}}$$

$$\dot{\sim} t_{n-2-k}$$

However, Kim [6, 6] uses

```
if(p.method == "Kendall"){
statistic <- pcor/sqrt(2*(2*(n-gn)+5)/(9*(n-gn)*(n-1-gn)))
p.value <- 2*pnorm(-abs(statistic))
```

citing [p. 7, 7]. Note that there is no adjustment for the number of ties. *Also note:* The Abdi reference given below is not the source of this. I have the same reference she does and it doesn't match the Encyclopedia entry. [8] gives the conditional based on counts and no test statistics. They do note:

> *However, the association measures are usually not invariant with the breakdown of categories.*

demonstrating that the association measure changes when you merge categories. It can go either way.

> *The point is that the associations are conditional to the retained association level.*

This is relevant to consumer statistics, as the straight use of k-box questions and plain old correlations assumes that this is exact scale of interest.

They suggest a Kendall tau_b (pg 49):

> *By analogy for ordinal variables, we suggest then to use Kendall's rb, which is symmetric, for the raw association and Somers's d, which is asymmetric, for the path coefficients.*

They also note that, for Somer's d, $d_{AB}d_{BA} = tau_b^2$, analogously to $beta_x$ for the OLS regression of y on x and vice-versa. See also: [9] which does give asymptotic standard errors for cell counts, versus Quade's [10] paper, which uses raw data.

$$beta_y = r_{xy}^2$$

Paired Comparisons

Need to note that the pairs are pre-made in paired comparisons (and partially in ranked). It becomes a simple matter to count for tau_b. The diagonal counts are the concordant pairs, the anti-diagonal counts are the discordant pairs and marginals supply the denominator terms.

Agresti [11] Category Choice

Abstract

> *Several ordinal measures of association for cross-classification tables are compared with respect to their stability when various grlds are placed on a bivariate normal distribution. Kendall's tau b usually fares better than Kendall's tau c, Goodman and Kruskal's gamma, and three extensions of Spearman's rho for cross-classfication tables, in terms of approximating an associated measure for ungrouped data. The loss of efficiency of tau b due to grouping in testing the hypothesis of no association is considered and observed to be strongly related to the proportion of tied pairs of observations.*

For a square table (r=c), define the probability of a tie as

$$P_t = \sum_i p_{i.}^2 - \sum_j p_{.j}^2 - \sum_i \sum_j p_{ij}^2$$

also, crudely,

$$(1 - RE)$$
$$dotsim 1 -$$
$$sqrt(1 - P_t)$$

he also suggests that, for n_0 ungrouped observations, the equivalent grouped n is defined by

$$n_0(n_0 - 1)/2$$
$$approx(1 - P_t)n(n - 1)/2$$

> *For all values of pearsons rho, as r and c increase in such a way that P_t decreases toward zero, RE increases toward one. For example, when rho = .5 with $p_{i.}$ = I / r and p_{.j} = I/c, the test for the 4 X 4 table is about twice as efficient as the test for the 2 X 2 table (.778 vs. .380), and RE = .926 for the 10 X 10 table.A more thorough inspection of the 226 grids for each value of rho reveals that the RE (of tau_b) values are linearly related to the 1 - P_t values to a good approximation.*

Earlier he notes that when $rho = .2$ the RE is .434 and RE = .106 when $rho = .8$. As rho increases away from 0, RE decreases. Also decreases marginal asymmetry.

Conclusion

In summary, Kendall's tau_b seems more stable overall than the others in terms of approximating the corre- sponding measure for ungrouped data. Also, whenever possible, care should be exercised in choosing a grid for a table. When the number of rows or columns is increased or the marginal proportions are selected to minimize the proportion of tied pairs of observations, the efficiencyof a hypothesis test of no association tends to improve and the value of the measure tends to be closer to the value for the underlying continuous distribution.

Quade Nonparametric Partial Correlation [4] [10] [12]

Concepts of Control

1. *C(X,Y) Holding Z constant* — conditional correlations. Partial correlation as average correlation. See Davis' partial Gamma $\gamma(X, Y|Z)$

$$sum_i P_{C_i} -$$
$$sum_i P_{Di}$$
$$oversum_i P_{Ci} +$$
$$sum_i P_{Di}$$

$$P_{C_i} - P_{Di}$$
$$overP_{Ci} + P_{Di}$$

$fracab$

where $P_{C_i}(P_{Di})$ is the probability of a concordant (discordant) pair in the i-th stratum. Then γ is a weighted average of the γ_i.
2. *C(X,Y) adjusting for* Z - correlation between residuals from "regression" on Z. The familiar partial correlation formula is what you get when you assume a OLS with homscedastic variances. There is an implicit assumption that the conditional correlations are $approx$ constant.
3. *Kendall partial* phi: two randomly chosen triple (X,Y,Z) crossclassified as (X,Z) concordant or discordant and (Y,Z) similarly, to form a partial phi coefficient.
4. * Partial Proportional Reduction in Error* what do you gain from knowing X in addition to Z when predicting Y?

> *WJR: The difference between "Holding Z Constant" and "adjusting for* Z*" is that the former doesn't require or assume a model while the later does assume things like additivity and linearity and normality. When all those hold, they are the same. Otherwise they are not (e.g. discrete multivariate matching captures interactions among the matching variables, the usual partial doesn't*

Concepts of Correlation

- **Conditional Correlation** index of conditional correlation

$$tau_{\underset{vertZ}{XY}}(z) = P_{C|Z=z} - P_{D|Z=z},$$

  where $P_{C|Z=z}$ is the probability that a randomly chosen pair of observations (X_1, Y_1, Z_1) and (X_2, Y_2, Z_2) will be concordant, conditional upon the event that Z_1 and Z_2 are both fixed at the same point z. Z is unspecified and need not be random.
- **Partial Correlation** $tau_{\underset{vertZ}{XY}}$ is a weighted average of the conditional correlations $tau_{\underset{vertZ}{XY}}(z)$ over the values z of Z.

PARTIAL CORRELATION BASED ON MATCHING

Let X and Y be at least ordinal and Z be unrestricted.

Then an intuitively reasonable way of measuring the partial correlation … is to find out much more probable it is to get like than unlike orders with respect to X and Z when pairs of observations are chosen at random from the population.

> *WJR: Note for discrete choice paired comparisons: The "matching" is given and so two products are matched with respect to a person and Z when $Z = 0$. All comparisons are within person, and are averaged across people who are strata.*

He introduces [4] the W function:

$$W((x_1, y_1), (x_2, y_2)) = 1 \text{ if } x_1 < x_2, y_1 < y_2 \text{ or } x_1 > x_2, y_1 > y_2,$$

and

$$W((x_1, y_1), (x_2, y_2)) = 0 \text{ if } \ x_1 = x_2 \text{ or } \ y_1 = y_2 \text{ or both},$$

and

$$W((x_1, y_1), (x_2, y_2)) = -1 \text{ if } x_1 < x_2, y_1 > y_2 \text{ or } x_1 > x_2, y_1 < y_2,$$

so that

$$\tau_{XY} = \int \int W((x_1, y_1), (x_2, y_2)) dF(x_1, y_1) dF(x_2, y_2)$$

and

$$\tau_{XY|Z}(z) = \int \int W((x_1, y_1), (x_2, y_2)) dF(x_1, y_1|z) dF(x_2, y_2|z)$$

Historically, the confusion between conditional and partial arises because of the multivariate normal where they are equal, because the conditional is constant. Quade quotes Yule [13] to the effect that the partial correlation should be interpreted as a weighted average.

Quade's index of matched correlation

$$\theta(X, Y|Z) = P(C|\text{MATCH}) - P(D|\text{MATCH}),$$

where $textrmMATCH$ is the event that a (randomly) chosen pair is matched on Z.

The estimator is then

$$T(X, Y|Z) = N_{CM} - N_{DM},$$
$$overN_M$$

where N_M is the number that are matched and $ {N*{CM}} ({N*{DM}})$ are the number of concordant and discordant pairs among the matched observations. Note that Quade is pooling here.

In [4] he provides some example distances for matching on a continuous variable:

- maximum component: $$ D(\underline{z}_i, \underline{z} *j) = \max_k \vert z*{i}^{k} - z_{j}^{k} \vert $$
- city-block: $ D(\underline{z}_i, \underline{z}_j) = \sum_k \vert z_i^{k-z_j}k \vert $
- Weighted Euclidean for a weight matrix Q
- Mahalanobis

In discrete case this could be a k of n definition, which is a sum of indicators for each z^k.

Note: In the published paper there is a slight notation shift and $sum_i M_i$ becomes $sum_i R_i$ and M_M becomes R, where R stands for "relevant". This emphasizes the generality of this formula. Some examples are
- "all pairs are relevant" leads to tau_a
- "pairs are relevant unless they are tied" leads to $gamma$
- "pairs are relevant unless they are tied on X," leads to Sommers' d_{yx}
- "pairs are relevant unless they are tied on both X and Y simultaneously," leads to Wilson's e, which is

$$e = \frac{C - D}{overC + D + T_Y + T_X},$$

where T_X is the number of pairs tied on X but not on Y
- "pairs are relevant if they are tied on the control variable Z but not tied on X or on Y" produces Davis' partial correlation coefficient ($gamma$).

sampling distribution

Let M_i be the number of observations $(X_j, Y_j, Z_j), j$ that are matched with (X_i, Y_i, Z_i), and let W_i be the number concordant less the number discordant. Then $sum_i M_i = 2N_M$ and $\sum_i W_i = 2(N*{CM} - N*{DM}) $. The factor 2 appears because each matched pair is counted twice; hence

$$T(X, Y|Z) = \frac{sum_i W_i}{oversum_i M_i},$$

> *WJR Note: In the cases of paired comparisons (or blocks) the index i indexes products NOT persons, so each person contributes 2 product evaluations!*
> *Also, this discussion prefigured the poset ideas present in the 2018 version of the pim package in R, where you can specify the pairs that are relevant*

The sampling distribution of each index in this family is asymptotically normal with standard error

$$S(X, Y vertZ) = 2 over(sum_i M_i)^2 sqrtsum_i W_i^2 (sumM_i)^2 - 2sumW_i(sumW_iM_i) sumM_i) + (sumM_i^2 sum_i W_i)^2$$

The test statistic is $\frac{T-\tau}{S(X,Y|Z)}$ [p. 15, 4].

T is always a consistent estimator of $theta$.

> *The index of matched correlation may be regarded as a somewhat generalized version of partial correlation in the sense of average conditional correlation.*

If there are n relevant sets, he proposes[p. 23, 10] an alternate statistic for the null hypothesis that $theta = 0$

{{\\overline{W} \\sqrt{n}}\\over{2\\sqrt{\\sum (W_i - \\overline{W})^2}}} \\ge Z_\\alpha,

where $overlineW = sum_i W_i$. Note the 2! For the (–1,0,1) coding in discrete choice pairs, $overline(W)={{\sum_i W_i}} overn$

Recommends n pairs gt 200. [p. 382, 12].

He also gives an upper bound on the asymptotic variance of T:
$$ \sigma_\theta = sqrt{ {{2(1-\theta_0^{2)}\over{n p_R}} },$$ where $theta_0$ is the population value of $theta$, maximized when $theta = 0$, n is the number of items, and p_R is the proportion of pairs that are relevant.

On [p. 28, 10] for tau_a he notes the standard error simplifies to

$$2 overn(n-1) sqrtsumW_i^2 - (sumW_i)^2 overn$$

while $gamma$ and Davis' Partial $gamma$ can be simplified to a special form of the generalized variance estimator.

He also discusses allocating ties (Rao-Kupper is not mentioned)
1. 50/50 split leads to tau_a
2. Proportional allocation leads to $gamma$
3.
General comment:

> *in general, those measures which include ties may be regarded as conservative or pessimistic, since they tend to underestimate the stgrength of any underlying correlation; whereas, those which discard ties are optimistic, tending to overestimate its strength.*

He also discusses tau_b as being a compromise. It is not a special case of the index of matched correlation. He considers a modified form of the index

$$T(X, Y|Z) = N_{CM} - N_{DM} \qquad ,$$
$$over N_{CM} + N_{DM} + (N_{XM} + N_{YM})/2$$

where N_{XM} are the number tied on X but not Y, similarly for N_{YM}. Compare this with Wilson's e, given above. Wilson explicitly excludes the pairs tied on both X and Y, while Quade's modified tau_b includes them once. Quade notes in passing that this index is quite close to tau_b.

For the standard error results, replace M_i by the number of observations which are matched with the observation (X_i, Y_i, Z_i) and either concordant or discordant with it, plus half the number of matched observations which are tied with it on X but not Y, $oron\$Y$ but not X for $i = 1, 2, \ldots, n$;, leaving W_i unchanged. The asymptotic results then hold without further modification.

He notes that while the partial correlation formula is numerically unstable in the presence of highly correlated variables and that this is not obvious, while problems with the matched correlation show up in the small sample sizes.

Appendix

He derives the variance via U-statistics.

Note on Wilson's e [14]

Wilson derives his measure under an hypothesis of strict monotone relations:
- Positive: if x increases then y increases and if x does not vary, then neither does y
- Negative: if x increases then y decreases and if x does not vary, then neither does y

and proposes his e as a measure of that. Count the positive forms (concordant) and take the difference from the negative form (discordant), and normalize by the total number of pairs Minus the ones that don't change at all.

References

1. Newson, R. (2002). Parameters behind“ nonparametric” statistics: Kendall’s tau, Somers’ D and median differences. *The Stata Journal* 1(1) 1–20.
2. Conover (1999), “Practical Nonparametric Statistics,” Third Edition, Wiley, p. 327.
3. Davis, J.A. (1967). A partial coefficient for Goodman and Kruskal’s gamma. Journal of the American Statistical Association, 62, 189–193 (§4.4 in Conover).
4. Quade, D. (1967). “Nonparametric Partial Correlation.” Institute of Statistics mimeo series 526
5. Kendall, M. G., and Stuart, Alan, (1961) *The Advanced Theory of Statistics, Vol 2* Hefner Publishing Company, New York.
6. Kim, S. (2015). ppcor: an R package for a fast calculation to semi-partial correlation coefficients. Communications for statistical applications and methods, 22(6), 665.
7. Abdi H. Kendall rank correlation. In: Salkind NJ, editor. Encyclopedia of Measurement and Statistics. Thousand Oaks (CA): Sage; 2007. pp. 508–510.
8. Ritschard, G., Kellerhals, J., Olszak, M., & Sardi, M. (1996). Path analysis with partial association measures. Quality and Quantity, 30(1), 37–60.
9. Olszak, M., & Ritschard, G. (1995). The Behaviour of Nominal and Ordinal Partial Association Measures. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 44(2), 195–212. doi:10.2307/2348444
0. Quade, D. (1971). “Nonparametric Partial Correlation.” Report SW 13/71. Mathematisch Centrum, Amsterdam
1. Agresti, A. (1976). The effect of category choice on some ordinal measures of association. Journal of the American Statistical Association, 71(353), 49–55.
2. Quade, D. (1974). Nonparametric partial correlation, in H. M. Blalock Jr. (ed.), Measurement in the Social Sciences, Aldine-Atherton, Chicago, pp. 369–397.
3. Yule, G. U. (1911) *An Introduction to the Theory of Statistics*, Charles Griffin and Company, London.
4. Wilson, T. P. (1974). Measures of association for bivariate ordinal hypotheses. Measurement in the social sciences, 327–342.