The Stata Journal (2001) **1**, Number 1, pp. 1–20

Parameters behind "non-parametric" statistics: Kendall's τ_a, Somers' D and median differences

Roger Newson
King's College, London, UK
roger.newson@kcl.ac.uk

Abstract.

So-called "non-parametric" statistical methods are often in fact based on population parameters, which can be estimated (with confidence limits) using the corresponding sample statistics. This article reviews the uses of three such parameters, namely Kendall's τ_a, Somers' D and the Hodges-Lehmann median difference. Confidence intervals for these are demonstrated using the `somersd` package. It is argued that confidence limits for these parameters, and their differences, are more informative than the traditional practice of reporting only P-values. These three parameters are also important in defining other tests and parameters, such as the Wilcoxon test, the area under the receiver operating characteristic (ROC) curve, Harrell's C, and the Theil median slope.

Keywords: notag1, confidence intervals, Gehan test, Harrell's C, Hodges-Lehmann median difference, Kendall's tau, non-parametric methods, rank correlation, rank-sum test, ROC area, Somers' D, Theil median slope, Wilcoxon test.

1 Introduction

Rank-based statistical methods are sometimes called "non-parametric" statistical methods. However, they are usually in fact based on population parameters, which can be estimated using confidence intervals around the corresponding sample statistics. Traditionally, these sample statistics are used for significance tests of the hypothesis that the population parameter is zero. However, statisticians increasingly recommend confidence intervals in preference to P-values alone, for rank-based parameters as well as for regression parameters such as mean differences and relative risks.

Three important rank-based parameters are Kendall's τ_a, Somers' D (which is defined in terms of Kendall's τ_a), and the Hodges-Lehmann median difference (which is defined in terms of Somers' D). This review aims to summarize the use and estimation of these parameters, and their links to methods possibly more familiar.

1.1 The `somersd` package

The methods will be demonstrated using the `somersd` package. In its present form, the package contains two programs, `somersd` (which calculates confidence intervals for Kendall's τ_a and Somers' D) and `cendif` (which calculates confidence limits for median and other percentile differences). The original version of `somersd` was presented (with

methods and formulae) in Newson (2000a) and updated by Newson (2000b, 2000c). The original version of `cendif` (with methods and formulae) was presented in Newson (2000d). The most up-to-date version of the `somersd` package at any time is downloadable from SSC. `somersd` offers a choice of normalizing and/or variance-stabilizing transformations, notably the arcsine and the hyperbolic arctangent. It also offers a `cluster` option.

2 Kendall's τ_a and Somers' D

Given two variables X and Y, sampled jointly from a bivariate distribution, the population value of Kendall's τ_a (Kendall, 1938; Kendall and Gibbons, 1990) is defined as

$$\tau_{XY} = E\left[\text{sign}(X_1 - X_2)\,\text{sign}(Y_1 - Y_2)\right], \tag{1}$$

where (X_1, Y_1) and (X_2, Y_2) are bivariate random variables sampled independently from the same population, and $E[\cdot]$ denotes expectation. The population value of Somers' D (Somers, 1962) is defined as

$$D_{YX} = \frac{\tau_{XY}}{\tau_{XX}}. \tag{2}$$

Therefore, τ_{XY} is the difference between two probabilities, namely the probabilities of concordance and discordance between the X-values and the Y-values. The X-values and Y-values are said to be concordant if the larger of the two X-values is associated with the larger of the two Y-values, and they are said to be discordant if the larger X-value is associated with the smaller Y-value. D_{YX} is the difference between the two corresponding *conditional* probabilities, given that the two X-values are not equal.

Kendall's τ_a is the covariance between $\text{sign}(X_1 - X_2)$ and $\text{sign}(Y_1 - Y_2)$, whereas Somers' D is the regression coefficient of $\text{sign}(Y_1 - Y_2)$ with respect to $\text{sign}(X_1 - X_2)$. The corresponding correlation coefficient between $\text{sign}(X_1 - X_2)$ and $\text{sign}(Y_1 - Y_2)$ is known as Kendall's τ_b, and is defined as

$$\tau_{XY}^{(b)} = \text{sign}(\tau_{XY}) \times \sqrt{D_{XY} D_{YX}}\,, \tag{3}$$

the geometric mean of the two regression coefficients D_{YX} and D_{XY} multiplied by their common sign. Kendall's τ_a and τ_b are both calculated by `ktau`, but τ_b is more commonly quoted than either Kendall's τ_a or Somers' D. However, τ_a is more easily interpreted in words to non-statisticians. For instance, if two medical statistics lecturers (Lecturer A and Lecturer B) are double-marking exam scripts, and Kendall's τ_a between their two marks is 0.7, then this means that, given two exam scripts and asked which of the two is better, the two statisticians are 70% more likely to agree than to disagree. (Agreement and disagreement are defined in the strictest sense of concordance and discordance, respectively, excluding cases where tied marks are awarded by either lecturer.)

Differences between concordance and discordance probabilities (such as Somers' D and Kendall's τ_a) have the attractive property that they lie on a scale from -1 to 1, where values of 1, -1 and 0 signify a perfect positive relationship, a perfect neg-

ative relationship, and no overall ordinal relationship at all, respectively. Concordance/discordance ratios, on the other hand, are on a scale from 0 to ∞, with values of 1 in the case of statistical independence. If both X and Y are binary, then their concordance/discordance ratio is their odds ratio.

An alternative parameter used in defining rank methods is Spearman's r_S, defined as the product-moment correlation coefficient between the respective cumulative distribution functions (CDFs) of the X_i and the Y_i, and estimated by the correlation coefficient of the corresponding ranks. r_S is on a scale from -1 to 1, but is not interpretable as a difference between probabilities. As Kendall and Gibbons (1990) argue, confidence intervals for Spearman's r_S are less reliable and less interpretable than confidence intervals for Kendall's τ-parameters, but the sample Spearman's r_S is much more easily calculated without a computer. This was an important consideration when Spearman's r_S was originally advocated (Spearman, 1904). Kendall's τ-parameters were introduced under their present name by Kendall (1938), but parameters based on concordance and discordance probabilities were discussed even earlier (e.g. Fechner (1897)). Kruskal (1958) gives a good account of Kendall's τ_a, Spearman's r_S and other ordinal correlation measures, including historical references.

2.1 Confidence intervals vs. significance tests

The population parameters described above can be estimated by the corresponding sample statistics, such as the sample Kendall's τ_a ($\hat{\tau}_{XY}$) or the sample Somers' D ($\hat{D}_{YX}$). Traditionally, however, these sample statistics are used only to test the null hypothesis that the corresponding population parameter is zero. In Stata, `ktau` tests the hypothesis that Kendall's τ_a is zero, using the sample τ_a.

A confidence interval for Kendall's τ_a (or Somers' D) is more informative, for two main reasons.

- If the null hypothesis is not compatible with the data, then we might ask which hypotheses *are* compatible with the data. For instance, in the case of the two lecturers double-marking exam scripts, it is not very helpful just to be told that the Kendall's τ_a between their marks is "significantly positive", because this only shows that, given two exam scripts and asked which is best, they are more likely to agree than to disagree, and that the excess of agreement over disagreement is too large to be explained by chance. It is more informative to be told that their Kendall's τ_a is 0.70 (95% CI, 0.67 to 0.72), because this shows, with 95% confidence, that they are at least 67% more likely to agree than to disagree, and possibly as much as 72% more likely to agree than to disagree.
- If the null hypothesis *is* compatible with the data, then we might ask what other hypotheses are *also* compatible with the data. As a statistical referee, I find that the most common single mistake made by naive medics is to carry out a "non-parametric" test on a small sample, and to find a large P-value, and then to argue that the high P-value proves the null hypothesis. This is definitely *not* the case

if the two lecturers have double-marked a sample of 17 exam scripts, and their Kendall's τ_a is "non-significant" at 0.35 (95% CI, −0.11 to 0.69; $P = 0.17$).

2.2 Differences between τ_a or Somers' D values

Given an outcome variable Y and two positive predictors W and X, we may want to ask whether W or X is a better predictor of Y. This might be done by defining a confidence interval for the difference $\tau_{WY} - \tau_{XY}$, or for half of that difference. For instance, suppose three statisticians are treble-marking exam scripts, and W, X and Y are the marks awarded by Lecturers A and B and Professor C respectively, and $\tau_{WY} = 0.73$, and $\tau_{XY} = 0.67$. Then the difference between the τ_a values is 0.06, and half that difference is 0.03. This means that, given two exam scripts to place in order, Professor C is (approximately) 3% more likely to agree with Lecturer A and to disagree with Lecturer B than she is to agree with Lecturer B and to disagree with Lecturer A. This might be thought important if Professor C represents a "gold standard".

To understand this point, suppose that trivariate data points (W_i, X_i, Y_i) are sampled independently from a common population, and define $\mathrm{Con}(X, Y)$, $\mathrm{Dis}(X, Y)$ and $\mathrm{Tie}(X, Y)$ as the events that (X_1, Y_1) and (X_2, Y_2) are concordant, discordant or neither, respectively, and similarly for $\mathrm{Con}(W, Y)$, $\mathrm{Dis}(W, Y)$ and $\mathrm{Tie}(W, Y)$. Then the difference between the two τ_a values is

$$\begin{aligned}\tau_{WY} - \tau_{XY} = \quad & 2\,\{\,\Pr[\mathrm{Con}(W,Y) \text{ and } \mathrm{Dis}(X,Y)] - \Pr[\mathrm{Con}(X,Y) \text{ and } \mathrm{Dis}(W,Y)]\,\} \\ & + \Pr[\mathrm{Tie}(X,Y) \text{ and } \mathrm{Con}(W,Y)] - \Pr[\mathrm{Tie}(X,Y) \text{ and } \mathrm{Dis}(W,Y)] \\ & - \Pr[\mathrm{Tie}(W,Y) \text{ and } \mathrm{Con}(X,Y)] + \Pr[\mathrm{Tie}(W,Y) \text{ and } \mathrm{Dis}(X,Y)]\,. \quad (4)\end{aligned}$$

In particular, if the marginal distributions of W and X are both continuous, then only the first term (in the curly braces) is non-zero, and then we have

$$(\tau_{WY} - \tau_{XY})/2 = \Pr[\mathrm{Con}(W,Y) \text{ and } \mathrm{Dis}(X,Y)] - \Pr[\mathrm{Con}(X,Y) \text{ and } \mathrm{Dis}(W,Y)]\,. \quad (5)$$

Whether or not W and X are continuous, Kendall's τ_a has the advantageous property that a larger τ_a cannot be secondary to a smaller τ_a. That is to say, if a positive τ_{XY} is caused entirely by a monotonic positive relationship of both variables with W, then τ_{WX} and τ_{WY} must both be greater than τ_{XY}. If we can show that $\tau_{XY} - \tau_{WY} > 0$ (or, equivalently, that $D_{XY} - D_{WY} > 0$), then this implies that the correlation between X and Y is not caused entirely by the influence of W. This feature is a good reason for preferring Somers' D and Kendall's τ_a to other measures of ordinal trend. To understand this point, suppose that the (W_i, X_i, Y_i) have a discrete probability mass function $f_{W,X,Y}(\cdot, \cdot, \cdot)$ and a marginal probability mass function $f_{W,X}(\cdot, \cdot)$. Define the conditional expectation

$$Z(w_1, x_1, w_2, x_2) = E[\mathrm{sign}(Y_2 - Y_1) | W_1 = w_1, X_1 = x_1, W_2 = w_2, X_2 = x_2] \quad (6)$$

for any w_1 and w_2 in the range of W-values and any x_1 and x_2 in the range of X-values. If we state that the positive relationship between X_i and Y_i is caused entirely by a

monotonic positive relationship between both variables and W_i, then that is equivalent to stating that

$$Z(w_1, x_1, w_2, x_2) \geq 0 \tag{7}$$

whenever $w_1 \leq w_2$ and $x_2 \leq x_1$. However, (4) can then be rewritten

$$\begin{aligned} \tau_{WY} - \tau_{XY} = \quad & 4 \sum_{w_1<w_2} \sum_{x_2<x_1} f_{W,X}(w_1, x_1)\, f_{W,X}(w_2, x_2)\, Z(w_1, x_1, w_2, x_2) \\ & +2 \sum_{x} \sum_{w_1<w_2} f_{W,X}(w_1, x)\, f_{W,X}(w_2, x)\, Z(w_1, x, w_2, x) \\ & +2 \sum_{w} \sum_{x_2<x_1} f_{W,X}(w, x_1)\, f_{W,X}(w, x_2)\, Z(w, x_1, w, x_2). \end{aligned} \tag{8}$$

This difference must be non-negative whenever the inequality (7) applies, and depends on the ordering of Y-values in pairs of data points where the W-values are non-concordant with the X-values.

The program `somersd` calculates Somers' D or Kendall's τ_a between one variable X and a list of others $Y^{(1)} \ldots Y^{(p)}$, and saves the estimation results as for a model fit. Confidence intervals for differences can then be calculated using `lincom`. For instance, in the `auto` data set distributed with official Stata, we might `generate` a new variable `gpm=1/mpg` to represent fuel consumption in gallons/mile, and use Kendall's τ_a estimates and their differences to find out if fuel consumption is predicted better by the weight of the car (in pounds) or by its displacement (in cubic inches):

```
. somersd gpm weight displacement,taua
Kendall's tau-a with variable: gpm
Transformation: Untransformed
Valid observations: 74
Symmetric 95% CI
```

gpm	Coef.	Jackknife Std. Err.	z	P>|z|	[95% Conf.	Interval]
gpm	.9470566	.0077145	122.76	0.000	.9319366	.9621767
weight	.685672	.0445194	15.40	0.000	.5984156	.7729283
displacement	.5942244	.0601971	9.87	0.000	.4762403	.7122085

```
. lincom (weight-displacement)/2
 ( 1)  .5 weight - .5 displacement = 0.0
```

gpm	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
(1)	.0457238	.0229597	1.99	0.046	.0007236	.090724

We note that `somersd`, with the `taua` option, calculates the Kendall τ_a estimates between `gpm` and three other variables, namely `gpm` itself, `weight` and `displacement`. (The τ_a of `gpm` with itself is simply the probability that two independently-sampled `gpm` values are not equal.) We find that it is 60% to 77% more likely that a heavier car consumes more fuel per mile than less fuel per mile, and that it is 48% to 71% more likely that a higher-volume car consumes more fuel per mile than less fuel per mile. Finally, we use `lincom` to compute a confidence interval for the half-difference. As `weight` and `displacement` are nearly continuous, we conclude that, if we sample two cars at random,

then fuel consumption is (approximately) 0% to 9% more likely to be concordant with weight (but not with displacement) than with displacement (but not with weight). It therefore seems that heavier but less voluminous cars typically consume more fuel than lighter but more voluminous cars. It follows that more massive cars consume more fuel, and that this is not just because of their typically higher volume.

3 Kendall's τ_a and product-moment correlations

Compared with the standard Pearson product-moment correlation ρ_{XY}, Kendall's τ_a is slightly easier to interpret in words, and is certainly a lot more robust to extreme observations and to non-linearity. In particular, if X predicts Y by a perfectly monotonic non-linear relationship, then τ_{XY} will be equal to ± 1, whereas ρ_{XY} may have a lower magnitude than ρ_{WY} if W is an imperfect linear predictor that is less useful in practice. However, ρ_{XY} is much easier than τ_{XY} to calculate without a computer, and may be more impressively large than τ_{XY} if the true relationship between X and Y is fairly linear. In the case where X and Y are sampled from a bivariate normal distribution, the two correlation measures are associated by Greiner's relation

$$\rho_{XY} = \sin\left(\frac{\pi}{2}\tau_{XY}\right). \tag{9}$$

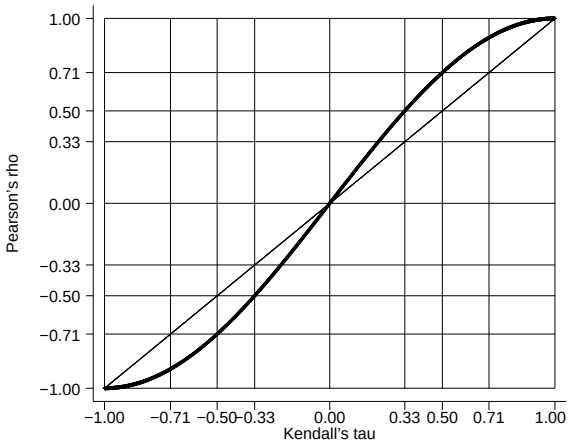

Figure 1: Greiner's relation between Pearson's ρ and Kendall's τ_a.

This relation is discussed in Kendall (1949) and depicted in Figure 1. Note that Kendall's τ_a-values of 0, $\pm\frac{1}{3}$, $\pm\frac{1}{2}$ and ± 1 correspond to Pearson's correlations of 0, $\pm\frac{1}{2}$, $\pm\frac{1}{\sqrt{2}}$ and ± 1, respectively. The Pearson ρ is therefore of greater magnitude than the Kendall τ.

Greiner's relation (or something similar) is expected to hold under a wide range of continuous bivariate distributions, as well as under the bivariate normal. Kendall

Table 1: τ_{XY} and ρ_{XY} for X and Y defined as sums and differences of "hidden variables" U, V and W, sampled independently from any common continuous distribution.

X	Y	τ_{XY}	ρ_{XY}
U	$\pm V$	0	0
$V+U$	$W \pm U$	$\pm\frac{1}{3}$	$\pm\frac{1}{2}$
U	$V \pm U$	$\pm\frac{1}{2}$	$\pm\frac{1}{\sqrt{2}}$
U	$\pm U$	± 1	± 1

(1949) showed that Greiner's relation is not affected by odd-numbered moments (such as skewness). Newson (1987), using a simpler line of argument, examined the case where the observed variables X and Y are defined as sums or differences of three hidden variables U, V and W, sampled independently from the same arbitrary continuous univariate distribution. It was shown that different definitions of X and Y implied values of Kendall's τ_{XY} and Pearson's ρ_{XY} on various points on the Greiner curve. These are listed in Table 1.

If X and Y are continuous and we expect Greiner's relation to hold, we can then calculate a confidence interval for τ_{XY}, and then define an "outlier-resistant" confidence interval for ρ_{XY} by transforming the confidence interval for τ_{XY} using Greiner's relation. This is especially helpful if we expect X and Y to be transformed to a bivariate normal form by a pair of monotonic transformations $g(X)$ and $h(Y)$. We then no longer have to hunt for such a pair of transformations, because, *if* such transformations exist, *then* it follows that $\tau_{g(X),h(Y)} = \tau_{XY}$, and therefore the correlation $\rho_{g(X),h(Y)}$ will be as implied by Greiner's relation (9).

In the case of the two lecturers double-marking exam scripts, their Kendall τ_a of 0.70 (95% CI, 0.67 to 0.72) could be transformed, using Greiner's relation, to an "equivalent" Pearson correlation of 0.89 (95% CI, 0.87 to 0.90). The latter form would be less explicable in terms of probabilities of agreement and disagreement, but more impressive when presented to an audience accustomed to Pearson correlations. Such an audience might include the two lecturers' superiors, or an external examiner.

4 Somers' D for binary X-variables

The Somers' D parameter D_{YX} is defined whether X and/or Y are discrete or continuous. However, in practice, it is most often used when X is discrete, and used most often of all if X is a binary variable with values 0 ("negative") and 1 ("positive"). D_{YX} is then equal to the difference between two probabilities. Given two individual Y-values Y_1 and Y_0, randomly sampled from the populations with "positive" and "negative" X-values respectively, Somers' D is defined as

$$D_{YX} = \Pr(Y_1 > Y_0) - \Pr(Y_0 > Y_1), \tag{10}$$

and is the parameter tested by a Wilcoxon test. If both X and Y are binary, then Somers' D is simply the difference between proportions

$$D_{YX} = \Pr(Y_1 = 1) - \Pr(Y_0 = 1). \tag{11}$$

4.1 Somers' D and Wilcoxon tests

Traditionally, Somers' D is usually used to define significance tests, using the sample Somers' D ($\hat{D}_{YX}$) to test the hypothesis that the population Somers' D (D_{YX}) is zero. In Stata (as in much other software), this is usually done using Wilcoxon tests. If X is a binary variable and Y is a quantitative variable, then `ranksum` (implicitly) uses a two-sample Wilcoxon test to test the hypothesis that D_{YX} is zero, using the sample Somers' D. If there are two paired variables U and V, and we define $X = \text{sign}(U - V)$ and $Y = |U - V|$, then (implicitly) the Wilcoxon matched pairs signed rank test carried out by `signrank` tests the hypothesis that $D_{YX} = 0$.

It would be more informative to have confidence limits for the population Somers' D values themselves, and their differences. For instance, in the `auto` data, we might define the binary X-variable `us=!foreign`, and compare weight, fuel consumption and price in American and non-American cars, using `ranksum`:

```
. ranksum weight,by(us) porder
Two-sample Wilcoxon rank-sum (Mann-Whitney) test
          us |       obs    rank sum    expected
-------------+---------------------------------
           0 |        22       395.5         825
           1 |        52      2379.5        1950
-------------+---------------------------------
    combined |        74        2775        2775
unadjusted variance     7150.00
adjustment for ties       -1.06
                        ---------
adjusted variance       7148.94
Ho: weight(us==0) = weight(us==1)
             z =  -5.080
    Prob > |z| =   0.0000
P{weight(us==0) > weight(us==1)} = 0.125
```

(*continued on the next page*)

```
. ranksum gpm,by(us) porder
Two-sample Wilcoxon rank-sum (Mann-Whitney) test
          us |       obs    rank sum    expected
-------------+---------------------------------
           0 |        22       563.5         825
           1 |        52      2211.5        1950
-------------+---------------------------------
    combined |        74        2775        2775
unadjusted variance     7150.00
adjustment for ties      -36.95
                        ---------
adjusted variance       7113.05
Ho: gpm(us==0) = gpm(us==1)
             z =  -3.101
    Prob > |z| =   0.0019
P{gpm(us==0) > gpm(us==1)} = 0.271
. ranksum price,by(us) porder
Two-sample Wilcoxon rank-sum (Mann-Whitney) test
          us |       obs    rank sum    expected
-------------+---------------------------------
           0 |        22         913         825
           1 |        52        1862        1950
-------------+---------------------------------
    combined |        74        2775        2775
unadjusted variance     7150.00
adjustment for ties        0.00
                        ---------
adjusted variance       7150.00
Ho: price(us==0) = price(us==1)
             z =   1.041
    Prob > |z| =   0.2980
P{price(us==0) > price(us==1)} = 0.577
```

We note that American cars are typically heavier, and consume more miles per gallon, than cars from elsewhere, but we cannot conclude that, in the population of car types at large, they are typically more or less expensive. Note also that we have used the `porder` option, introduced into Stata 7 on 13 April 2001. The `porder` option causes `ranksum` to output the sample value of $\Pr(Y_0 > Y_1)$, where Y_0 is the Y-value of a randomly-sampled non-US car and Y_1 is the Y-value of a randomly-sampled US-made car. This quantity appears in the formula for Somers' D (10), and, for a continuous Y-variable, is equal to $(D_{YX}+1)/2$, where X is an indicator of non-US origin. However, there are no confidence intervals of any kind.

`somersd` is more informative, allowing us to define confidence intervals for the population Somers' D values, and for their differences (using `lincom`):

(*continued on the next page*)

```
. somersd us weight gpm price
Somers' D with variable: us
Transformation: Untransformed
Valid observations: 74
Symmetric 95% CI
```

us	Coef.	Jackknife Std. Err.	z	P>|z|	[95% Conf.	Interval]
weight	.7508741	.0832485	9.02	0.000	.58771	.9140383
gpm	.4571678	.135146	3.38	0.001	.1922866	.7220491
price	-.1538462	.1496016	-1.03	0.304	-.4470598	.1393675

```
. lincom (weight-gpm)/2
 ( 1)  .5 weight - .5 gpm = 0.0
```

us	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
(1)	.1468531	.0442198	3.32	0.001	.0601838	.2335224

We note that, given a randomly-chosen American car and a randomly-chosen non-American car, the American car is 59% to 91% more likely to be heavier than the other car than to be lighter, 19% to 72% more likely to consume more gallons per mile than to consume fewer, and 45% less likely to 14% more likely to be more expensive than to be less expensive. Using `lincom`, we compare the association with weight with the association with fuel consumption. As weight and fuel consumption are nearly continuous, we can conclude that the American car is approximately 6% to 23% more likely to move more mass with less gas than to move less mass with more gas. Therefore, most of the time, American cars tend to be more efficient for their weight than cars from elsewhere. This has been shown in stronger terms than would be possible using a regression model, because the method does not use possibly contentious assumptions such as linearity or additivity.

4.2 ROC curves and dominance diagrams

Sometimes, we may want to use a quantitative variable Y to predict a binary variable X, rather than *vice versa*. For instance, in the medical world, we may want to use a quantitative clinical diagnostic test result to give a binary answer to the effect that the patient has tested positive or negative for a disease. Once again, D_{YX} can be used as a general measure of predictive power.

Typically, given a quantitative test result and asked for a binary prediction of disease, a medical statistician defines a threshold and says that the test result is "positive" if the quantitative result exceeds the threshold, and "negative" otherwise. The sensitivity of the test is defined as the probability that a patient tests positive, assuming that the said patient has the disease. The specificity of the test is defined as the probability that the patient tests negative, assuming that the said patient does not have the disease. Typically, the lower the threshold chosen, the higher the sensitivity and the lower the

specificity. There is therefore a trade-off.

Medical statisticians visualize this trade-off using the sensitivity-specificity curve, otherwise known as the receiver operating characteristic (ROC) curve (Hanley and McNeil, 1982). An example of such a curve is given in Figure 2, where the "patients" are cars in the `auto` data, and they are being tested, using fuel consumption (`gpm`) as a quantitative diagnostic test, for the "disease" of being made in the USA. By convention, the vertical axis is sensitivity (true positive rate), and the horizontal axis is the quantity (1−specificity), otherwise known as the false positive rate. The data points correspond to candidate thresholds, equal to the values of `gpm` occurring in the data, and connected in descending order from the highest to the lowest. The curve gives the true positive rate that can be purchased at the price of each possible false positive rate. The lower the threshold that must be exceeded for a car to be diagnosed as American, the greater will be the false positive rate, but, on the other hand, the true positive rate will also increase. The choice of a threshold depends on the perceived costs of mis-diagnosis in each direction, and also on the perceived prior probability that a car suffers from the "disease" of being American. For each candidate threshold y_{crit}, the corresponding point on the *population* ROC curve has horizontal co-ordinate $1 - F_0(y_{\text{crit}})$ and vertical co-ordinate $1 - F_1(y_{\text{crit}})$, where $F_0(\cdot)$ and $F_1(\cdot)$ are the cumulative distribution functions of the diagnostic measure for the populations of non-diseased and diseased individuals, respectively. For the *sample* ROC curve, the co-ordinates of the point corresponding to y_{crit} are $1 - \hat{F}_0(y_{\text{crit}})$ and $1 - \hat{F}_1(y_{\text{crit}})$, where $\hat{F}_0(\cdot)$ and $\hat{F}_1(\cdot)$ are the sample cumulative distribution functions.

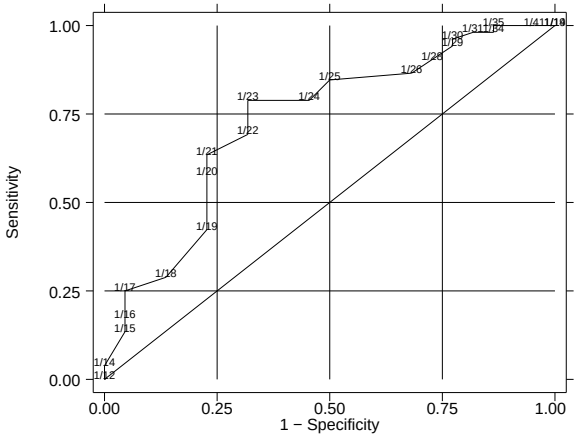

Figure 2: Receiver-operator characteristic (ROC) curve for `gpm` as a predictor of US origin.

The area under the ROC curve is frequently viewed as a good robust "performance indicator" for a quantitative diagnostic measure. If there are two quantitative diagnostic measures to choose from, and one yields a higher sensitivity than the other for every

possible false positive rate, then it is obviously to be preferred to the other, and obviously will have a higher ROC curve and therefore a greater ROC area. Figure 3 shows the ROC curves for `gpm` and `weight` as predictors of US origin. The ROC curve for weight is higher than that for fuel consumption for most (but not all) false positive rates, and the ROC area for weight is greater than that for fuel consumption.

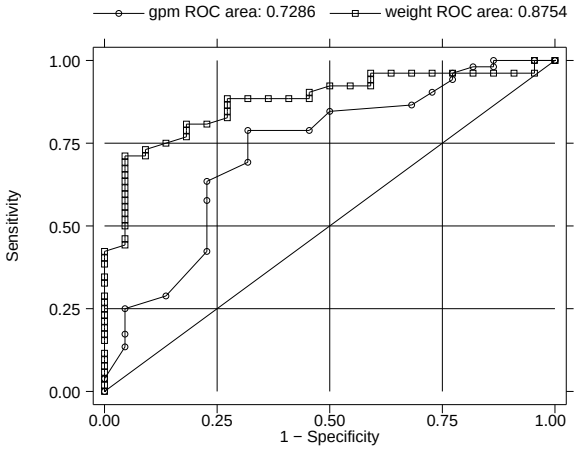

Figure 3: ROC curves for `gpm` and `weight` as predictors of US origin.

The area under the ROC curve for a quantitative clinical measure Y to predict a binary disease indicator X can be defined as

$$A_{YX} = \Pr(Y_0 < Y_1) + \frac{1}{2}\Pr(Y_0 = Y_1), \tag{12}$$

and the area over the ROC curve is equal to

$$1 - A_{YX} = \Pr(Y_0 > Y_1) + \frac{1}{2}\Pr(Y_0 = Y_1), \tag{13}$$

where Y_0 and Y_1 are values of the diagnostic measure sampled at random from the populations of negatives and positives, respectively. The corresponding Somers' D is

$$D_{YX} = \Pr(Y_0 < Y_1) - \Pr(Y_0 > Y_1) = 2A_{YX} - 1. \tag{14}$$

Therefore, the ROC area is a performance indicator equivalent to Somers' D, and the difference between two ROC areas is half the difference between the corresponding Somers' D values, which we measured for `weight` and `gpm` in the previous sub-section. Somers' D has the advantage that a perfect positive predictor, a perfect negative predictor and a completely useless predictor have Somers' D values of 1, -1 and 0, respectively, whereas their ROC areas are 1, 0 and 0.5. (A completely useless predictor is defined as a predictor whose ROC curve is the diagonal line from (0,0) to (1,1).)

The derivation of (12) and (13) can be made clearer by looking at Figure 4, which is a dominance diagram of the relation between US origin and fuel consumption. The

dominance diagram is essentially a re-invention of the ROC curve for the behavioral sciences, discussed in Fisher (1983), Cliff (1993) and Cliff (1996). The vertical axis is the `gpm` rank (highest values first) of an American car within the set of 52 American cars, whereas the horizontal axis is the `gpm` rank of a non-American car within the set of 22 non-American cars. The graphical area is therefore divided into a matrix of $52 \times 22 = 1144$ cells, and the cell (i, j) is assigned a plus-sign, a minus-sign or a zero, depending on whether the jth American car consumes more, less or the same amount of fuel, respectively, compared with the ith non-American car. The experiment of sampling a car at random from each group and measuring their fuel consumption is equivalent to sampling a point at random from the area of Figure 4. If we superimpose Figure 2 on Figure 4, then we will find that the area covered by plus-signs is below the ROC curve, the area covered by minus-signs is above the ROC curve, and the areas covered by zeros are bisected diagonally by the ROC curve. This implies that the areas below and above the ROC curve are given by (12) and (13).

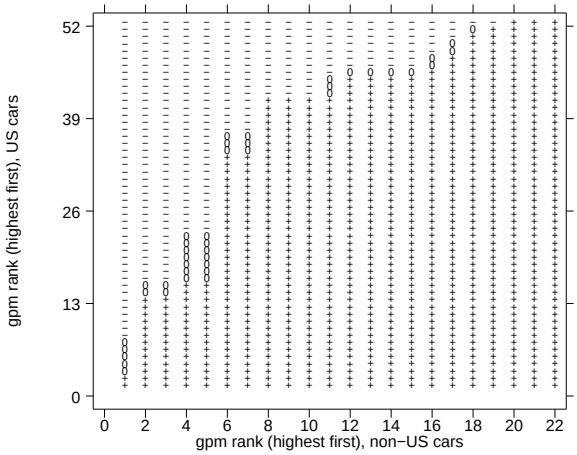

Figure 4: Dominance diagram for the relationship between fuel consumption and US origin.

Figure 2 was generated by `roctab`, whereas Figure 3 was generated by `roccomp`. Both of these programs belong to the `roc` package of official Stata, documented in [R] **roc**. `roctab` and `roccomp` calculate confidence intervals for ROC areas by a method similar to that used in default by `somersd` to calculate confidence intervals for Somers' D, due to DeLong, DeLong and Clarke-Pearson (1982). `roccomp` also gives chi-squared tests (but not confidence intervals) for the differences between ROC areas. The `roc` package is complementary to the `somersd` package, just as, in the regression statistics field, specialist programs such as `logit` are complementary to `glm`. The `roc` package is a specialist package for a special case, whereas `somersd` is a "grand unified solution", which offers the user extra options. These include a choice of normalizing and variance-stabilizing transformations for more accurate confidence intervals, such as the hyperbolic arctangent or z-transformation recommended by Edwardes (1995), and a `cluster` op-

tion for the case where there are multiple measurements per primary sampling unit, as discussed in Obuchowski (1997) and Beam (1998). Figure 4 was generated by the program `domdiag`, written by Nicholas J. Cox (who very kindly sent me a copy) and soon to be downloadable from SSC (at the time of writing). `domdiag` is complementary to the other two packages, and is especially useful for teaching purposes.

5 Extensions to survival data

Kendall's τ_a and Somers' D can be generalized to the case where the X-variable, the Y-variable, or both are possibly-censored lifetimes, rather than known values. The most general case is discussed extensively in Newson (1987). In general, given possibly-censored survival times X and Y, and censorship indicator variables R and S set to 1 if the lifetime terminates from the cause of interest and 0 if the lifetime is censored, we proceed as follows. For r_i and s_i equal to 0 or 1, and numbers x_i and y_i, we define

$$t(x_1,r_1,y_1,s_1,x_2,r_2,y_2,s_2)=\begin{cases}1 & \text{if } x_1<x_2,\ r_1=1,\ y_1<y_2 \text{ and } s_1=1,\\ 1 & \text{if } x_2<x_1,\ r_2=1,\ y_2<y_1 \text{ and } s_2=1,\\ -1 & \text{if } x_1<x_2,\ r_1=1,\ y_2<y_1 \text{ and } s_2=1,\\ -1 & \text{if } x_2<x_1,\ r_2=1,\ y_1<y_2 \text{ and } s_1=1,\\ 0 & \text{otherwise.}\end{cases} \tag{15}$$

We can then define Kendall's τ_a as

$$\tau_{X,R,Y,S}=E\left[t(X_1,R_1,Y_1,S_1,X_2,R_2,Y_2,S_2)\right], \tag{16}$$

where (X_1,R_1,Y_1,S_1) and (X_2,R_2,Y_2,S_2) are sampled independently from the same (X,R,Y,S) vector population distribution, the R_i and S_i must have values 0 or 1, and $E[\cdot]$ denotes expectation. We can define Somers' D as

$$D_{Y,S,X,R}=\tau_{X,R,Y,S}/\tau_{X,R,X,R}. \tag{17}$$

In principle, X, Y or both of them may be censored, and the latter might be the case if they are lifetimes of related organisms, as with the data analysed in Newson (1987). However, more attention has usually been paid to the case where only Y is a lifetime, whereas X is an uncensored predictor. Two common applications of Somers' D, available in Stata, are the Gehan test and Harrell's C. The Gehan test (Gehan, 1965), available as output from `sts test`, is similar to the Wilcoxon test, and tests the hypothesis that $D_{Y,S,X,1}=0$ in the case where X is a binary variable. William Gould's program `stcstat`, downloadable from SSC, calculates Harrell's C (Harrell *et al.*, 1982; Harrell *et al.*, 1996). If X is a continuous predictor variable, then Harrell's C is related to Somers' D by

$$D_{X,1,Y,S}=2C-1 \text{ or } C=(D_{X,1,Y,S}+1)/2. \tag{18}$$

Comparing this formula with (14), we see that Harrell's C is a reparameterization of Somers' D similar to the ROC area, but measures the ability of a continuous X to predict survival, rather than the ability of a continuous Y to predict disease. Note that the Gehan test is based on the Somers' D of Y with respect to X, whereas Harrell's C is based on Somers' D of X with respect to Y. The `somersd` package has not yet been extended to the case of possibly censored variables.

6 Median differences and slopes

Kendall's τ_a and Somers' D may be useful purely for scientific inference, in order to show that an association exists and that some associations are stronger than others. However, to be able to make economic or other practical decisions, we usually need to estimate a difference in units of the outcome variable. For instance, if we wish to know whether the difference in blood pressure between patients on Treatment A and Treatment B is large enough to justify the increased cost of Treatment B, then we need to have a difference in blood pressure units (e.g. millimetres of mercury) and a cost difference in dollars, rather than a Somers' D between treatment groups.

Fortunately, Somers' D (and the `somersd` package) can help us here as well. Somers' D is used in the definition of median differences and slopes, and can be used to define confidence limits for these.

6.1 The Hodges-Lehmann median difference

The Hodges-Lehmann median difference was introduced by Hodges and Lehmann (1963), and popularized by Conover (1980), Campbell and Gardner (1988) and Gardner and Altman (1989). Given two sub-populations A and B, the Hodges-Lehmann median difference is the median value of $Y_1 - Y_2$, where Y_1 is a value of an outcome variable Y sampled at random from Population A and Y_2 is a value of Y sampled at random from Population B. As Newson (2000d) pointed out, it can be defined in terms of Somers' D. In general, for $0 < q < 1$, a $100q$th percentile difference in Y can be defined as a value θ satisfying

$$D_{Y^*(\theta),X} = 1 - 2q, \tag{19}$$

where X is a binary variable equal to 1 for Population A and 0 for Population B, and $Y^*(\theta)$ is defined as Y if $X = 1$ and as $Y + \theta$ if $X = 0$. In particular, if $q = 0.5$, then the $100q$th percentile difference is known as a Hodges-Lehmann median difference, and satisfies

$$D_{Y^*(\theta),X} = 0. \tag{20}$$

Confidence intervals for the general $100q$th percentile difference (including the median difference) can be calculated using the program `cendif`, which is part of the `somersd` package. The statistical methods used, and the program `cendif` itself, are summarized in detail by Newson (2000d).

In the special case where the distributions of Y in Populations A and B differ only in location, the median difference is also the mean difference, which is the difference between the two population means, and also the difference between the two population medians. Traditionally, confidence intervals for the Hodges-Lehmann median difference have been calculated assuming that the two distributions differ only in location, so that the confidence interval is also a confidence interval for the difference between medians. In Stata, this is done using the STB program `npshift` (Wang, 1999) or by Patrick Royston's program `cid`, downloadable from SSC. The method used by `cendif` does not make this assumption, as the confidence interval is intended to be robust to the possi-

bility that the two populations differ in ways other than location. For instance, Y might be unequally variable between the two populations. Therefore, the difference between the method used by `cendif` and the method used by `npshift` is very similar to the difference between the unequal-variance t-test and the equal-variance t-test. `npshift`, like the equal-variance t-test, assumes that you can use data from the larger of two samples to estimate the population variability of the smaller of two samples.

I have carried out a few simulations of sampling from two normal populations, with a view to finding coverage probabilities of the confidence intervals generated by `cendif` and `npshift`. I have found that, even with small sample sizes, `cendif` gives coverage probabilities closer to the nominal ones when variances are unequal, in which case the traditional method gives confidence intervals either too wide or too narrow, depending on whether the larger or the smaller sample has the greater population variance, respectively. Usually, the difference between coverage probabilities has been small (2% or less), so the traditional method does not perform badly, in spite of its false assumption. However, if a sample of 20 is compared with a sample of 10, and the population standard deviation of the smaller sample is three times that of the larger sample, then the nominal 95% confidence interval has a true coverage probability of 90% using the traditional method and 94% using the `cendif` method. (Such a case is similar to sampling from two lognormal income distributions from two different countries, and taking a sample of 10 from a country whose 75th percentile is 8 times its 25th percentile, and a sample of 20 from a country whose 75th percentile is only twice its 25th percentile.) On the other hand, the two methods perform similarly when population variances are equal. From the results so far, I would therefore recommend the `cendif` method.

In the `auto` data, we might compare weight between American and non-American cars, using `npshift` and `cendif` to calculate a Hodges-Lehmann median difference:

```
. npshift weight,by(foreign)
Hodges-Lehmann Estimates of Shift Parameters
------------------------------------------------------------------
Point Estimate of Shift : Theta = Pop_2 - Pop_1 = -1095
95% Confidence Interval for Theta:          [-1350    ,     -720]
------------------------------------------------------------------

. cendif weight,by(foreign) tdist
Y-variable: weight (Weight (lbs.))
Grouped by: foreign (Car type)
Group numbers:
   Car type |      Freq.     Percent        Cum.
------------+-----------------------------------
   Domestic |          52       70.27       70.27
    Foreign |          22       29.73      100.00
------------+-----------------------------------
      Total |          74      100.00
Transformation: Fisher's z
Degrees of freedom: 73
95% confidence interval(s) for percentile difference(s)
between values of weight in first and second groups:
     Percent  Pctl_Dif   Minimum   Maximum
r1        50      1095       750      1330
```

We note that `npshift` and `cendif` estimate the same median difference, although

`npshift` gives the negative difference (−1,095 lb) between non-American and American cars, whereas `cendif` gives the positive difference (1,095 lb) between American and non-American cars. However, `cendif` gives slightly narrower confidence limits, because the larger group (52 American cars) is more variable in weight than the smaller group (22 non-American cars). A similar difference in confidence interval width is seen if we use `ttest` to calculate equal-variance and unequal-variance confidence limits for the mean difference (not shown).

As well as median differences, `cendif` can calculate median ratios, using logged data and the `eform` option:

```
. gene logwt=log(weight)
. cendif logwt,by(foreign) tdist eform
Y-variable: logwt
Grouped by: foreign (Car type)
Group numbers:
   Car type |      Freq.     Percent        Cum.
------------+-----------------------------------
   Domestic |         52       70.27       70.27
    Foreign |         22       29.73      100.00
------------+-----------------------------------
      Total |         74      100.00
Transformation: Fisher's z
Degrees of freedom: 73
95% confidence interval(s) for percentile ratio(s)
between values of exp(logwt) in first and second groups:
      Percent   Pctl_Rat    Minimum    Maximum
r1         50  1.4806389  1.3090908  1.6323524
```

We note that an American car typically has 131% to 163% of the weight of a non-American car.

6.2 The Theil median slope

The Theil median slope is a generalization of the Hodges-Lehmann median difference to the case of a non-binary X-variable. It was first defined by Theil (1950), and a good account of it appears in Sprent and Smeeton (2001). Supposing that (X_1, Y_1) and (X_2, Y_2) are sampled independently from a common bivariate distribution, the Theil median slope is usually defined as the median value of the slope $(Y_1 - Y_2)/(X_1 - X_2)$, or at least as its conditional median, assuming that $X_1 \neq X_2$. Sen (1968) argued that the Theil slope could be defined in terms of Kendall's τ, so the use of the Theil slope is often referred to as the Theil-Kendall method. The population Theil median slope is usually estimated using the sample Theil median slope, which is less affected by outliers than the ordinary least squares linear regression slope.

The Theil slope can also be defined in terms of Somers' D. In the general case, a $100q$th percentile slope can be defined as a value β such that

$$D_{Y-\beta X,X} = 1 - 2q. \tag{21}$$

In the case of $q = 0.5$, β is a median slope, such that

$$D_{Y-\beta X,X} = 0. \tag{22}$$

If (X_1, Y_1) and (X_2, Y_2) are sampled from the same bivariate (X, Y)-distribution, then (22) is equivalent to

$$\Pr\left[(Y_1 - Y_2)/(X_1 - X_2) > \beta \,|\, X_1 > X_2\right] = \Pr\left[(Y_1 - Y_2)/(X_1 - X_2) < \beta \,|\, X_1 > X_2\right], \tag{23}$$

where $\Pr[\cdot \,|\, \cdot]$ denotes conditional probability. This is a property we would expect of a median slope.

It is possible to generalize the method of `cendif` to calculate a sample Theil median slope, with confidence limits for the population Theil median slope, but I have not yet implemented this method in Stata. Traditionally, confidence intervals for the Theil median slope have been calculated assuming that the "residual" $Y - \beta X$ is not only "Kendall-uncorrelated" with X, but also independent of X. This, of course, implies that the "residual" $Y - \beta X$ has the same conditional variance regardless of X. A confidence interval for the Theil slope based on a modified `cendif` method would not use this assumption. It would therefore be robust to heteroskedasticity, like the Huber confidence interval for the least-squares regression slope.

Given that rank methods can be used to define confidence intervals for between-group differences and for linear quasi-regression slopes, it is natural to ask whether they could be used to define confidence intervals for anything similar to multivariate regression coefficients. Hussain and Sprent (1983) explored this question. They concluded that, if there were k different X-variables, then, instead of calculating the median of the slopes for all pairs of data points with different X-values, we would have to calculate median adjusted slopes for all sets of $k + 1$ data points. This would use an amount of computer time of the order of n^{k+1}, where n is the sample number. This suggests that regression-based methods, such as generalized linear models, will remain in business, at least for the important work of multivariate modelling.

7 Acknowledgment

I would like to thank Nicholas J. Cox for drawing my attention to Norman Cliff's work on dominance diagrams and for sending me a copy of his own program `domdiag`.

8 References

Beam, C. A. 1998. Analysis of clustered data in receiver operating characteristic studies. *Statistical Methods in Medical Research* 7: 324–336.

Campbell, M. J. and M. J. Gardner. 1988. Calculating confidence intervals for some non-parametric analyses. *British Medical Journal* 296: 1454–1456.

Cliff, N. 1993. Dominance statistics: ordinal analyses to answer ordinal questions. *Psychological Bulletin* 114, 494–509.

Cliff, N. 1996. *Ordinal Methods for Behavioral Data Analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

Conover, W. J. 1980. *Practical Nonparametric Statistics*. 2nd ed. New York: Wiley.

DeLong, E. R., D. M. Delong and D. L. Clarke-Pearson. 1982. Comparing the areas under two or more receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44: 837–845.

Edwardes, M. D. deB. 1995. A confidence interval for $\Pr(X < Y) - \Pr(X > Y)$ estimated from simple cluster samples. *Biometrics* 51: 571–578.

Fechner, G. T. 1897. *Kollectivmasslehre*. Leipzig: Wilhelm Engelmann. (Published posthumously, completed and edited by G. F. Lipps.)

Fisher, N. I. 1983. Graphical methods in nonparametric statistics: a review and annotated bibliography. *International Statistical Review* 51, 25–38.

Gardner, M. J. and D. G. Altman. 1989. *Statistics with confidence – confidence intervals and statistical guidelines*. London: British Medical Journal.

Gehan, E. A. 1965. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* 52: 203–223.

Hanley, J. A. and B. J. McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143: 29–36.

Harrell, F. E., R. M. Califf, D. B. Pryor, K. L. Lee and R. A. Rosati. 1982. Evaluating the yield of medical tests. *Journal of the American Medical Association* 247: 2543–2546.

Harrell, F. E., K. L. Lee and D. B. Mark. 1996. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 15: 361–387.

Hodges, J. L. and E. L. Lehmann. 1963. Estimates of location based on rank tests. *Annals of Mathematical Statistics* 34: 598–611.

Hussain, S. S. and P. Sprent. 1983. Non-parametric regression. *Journal of the Royal Statistical Society, Series A (General)* 146: 182–191.

Kendall, M. G. 1938. A new measure of rank correlation. *Biometrika* 30: 81–93.

Kendall, M. G. 1949. Rank and product-moment correlation. *Biometrika* 36: 177–193.

Kendall, M. G. and J. D. Gibbons. 1990. *Rank Correlation Methods*. 5th ed. London: Griffin.

Kruskal, W. H. 1958. Ordinal measures of association. *Journal of the American Statistical Association* 53: 814–861.

Newson, R. B. 1987. An analysis of cinematographic cell division data using U-statistics [D.Phil. dissertation]. Brighton, UK: Sussex University.

Newson, R. 2000a. snp15: `somersd` – Confidence intervals for nonparametric statistics and their differences. *Stata Technical Bulletin* 55: 47–55. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, pp. 312–322.

Newson, R. 2000b. snp15.1: Update to `somersd`. *Stata Technical Bulletin* 57: 35. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, pp. 322–323.

Newson, R. 2000c. snp15.2: Update to `somersd`. *Stata Technical Bulletin* 58: 30. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, p. 323.

Newson, R. 2000d. snp16: Robust confidence intervals for median and other percentile differences between groups. *Stata Technical Bulletin* 58: 30–35. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, pp. 324–331.

Obuchowski, N. A. 1997. Nonparametric analysis of clustered ROC data. *Biometrics* 53: 567–578.

Sen, P. K. 1968. Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association* 63: 1379–1389.

Somers, R. H. 1962. A new asymmetric measure of association for ordinal variables. *American Sociological Review* 27: 799–811.

Spearman, C. 1904. The proof and measurement of association between two things. *American Journal of Psychology* 15: 72–101.

Sprent, P. and N. C. Smeeton. 2001. *Applied nonparametric statistical methods.* Third ed. London: Chapman and Hall/CRC.

Theil, H. 1950. A rank invariant method of linear and polynomial regression analysis, I, II, III. *Proceedings of the Koninklijke Nederlandse Akademie Wetenschappen, Series A – Mathematical Sciences* 53: 386–392, 521–525, 1397–1412.

Wang, D. 1999. sg123: Hodges-Lehmann estimation of a shift in location between two populations. *Stata Technical Bulletin* 52: 52–53. Reprinted in *Stata Technical Bulletin Reprints*, vol. 9, pp. 255–257.

About the Author

Roger Newson is a medical statistician working at King's College, London, UK, principally in asthma research. He wrote the `somersd` package.