

Interpretation of Somers' D under four simple models

Roger B. Newson

04 October, 2020

1 Introduction

Somers' D is an ordinal measure of association introduced by Somers (1962)[9]. It can be defined in terms of Kendall's τ_a (Kendall and Gibbons, 1990)[4]. Given a sequence of bivariate random variables $(X, Y) = \{(X_i, Y_i)\}$, sampled using a sampling scheme for sampling pairs of bivariate pairs from a population of pairs of bivariate pairs, Kendall's τ_a is defined as

$$\tau(X, Y) = E [\text{sign}(X_i - X_j)\text{sign}(Y_i - Y_j)] \quad (1)$$

(where $E[\cdot]$ denotes expectation), or, equivalently, as the difference between the probability that the two X, Y -pairs are concordant and the probability that the two X, Y -pairs are discordant. A pair of X, Y -pairs is said to be concordant if the larger X -value is paired with the larger Y -value, and is said to be discordant if the larger X -value is paired with the smaller Y -value. Somers' D of Y with respect to X is defined as

$$D(Y|X) = \tau(X, Y)/\tau(X, X) \quad (2)$$

or, equivalently, as the difference between the two *conditional* probabilities of concordance and discordance, assuming that the 2 X -values are unequal. Note that Kendall's τ_a is symmetric in X and Y , whereas Somers' D is asymmetric in X and Y .

Somers' D plays a central role in rank statistics, and is the parameter behind most of these "nonparametric" methods, and can be estimated with confidence limits like other parameters. It can also be generalized to sampling-probability weighted and/or clustered and/or possibly censored data. (See Newson (2002)[7] and Newson (2006)[6] for details.) However, many non-statisticians appear to have a problem interpreting Somers' D , even though a difference between proportions is arguably a simpler concept than an odds ratio, which many of them claim to understand better. Parameters are often easier to understand if they play a specific role in a specific model. Fortunately, in a number of simple standard models, Somers' D can be derived from another parameter by a transformation. A confidence interval for Somers' D can therefore be transformed, using inverse end-point transformation, to give a robust, outlier-resistant confidence interval for the other parameter, assuming that the model is true.

We will discuss 4 simple models for bivariate X, Y -pairs:

- **Binary X , binary Y .** Somers' D is then the difference between proportions.
- **Binary X , continuous Y , constant hazard ratio.** Somers' D is then a transformation of the hazard ratio.
- **Binary X , Normal Y , equal variances.** Somers' D is then a transformation of the mean difference divided by the common standard deviation (SD). (Or, equivalently, a transformation of an interpercentile odds ratio of X with respect to Y .)
- **Bivariate Normal X and Y .** Somers' D is then a transformation of the Pearson correlation coefficient.

Each of these cases has its own Section, and a Figure (or Figures) illustrating the transformation. In each case, the alternative parameter (or its log) is nearly a linear function of Somers' D , for values of Somers' D between -0.5 and 0.5.

2 Binary X , binary Y

We assume that there are two subpopulations, Subpopulation A and Subpopulation B , and that X is a binary indicator variable, equal to 1 for observations in Subpopulation A and 0 for observations in Subpopulation B , and that Y is also a binary variable, equal to 1 for “successful” observations and 0 for “failed” observations. Define

$$p_A = \Pr(Y = 1|X = 1), \quad p_B = \Pr(Y = 1|X = 0) \quad (3)$$

to be the probabilities of “success” in Subpopulations A and B , respectively. Then Somers' D is simply

$$D(Y|X) = p_A - p_B, \quad (4)$$

or the difference between the two probabilities of “success”. Figure 1 gives Somers' D as the trivial identity function of the difference between proportions. Note that Somers' D is expressed on a linear scale of multiples of $1/12$, which (as we will see) is arguably a natural scale of reference points for Somers' D .

3 Binary X , continuous Y , constant hazard ratio

We assume, again, that X indicates membership of Subpopulation A instead of Subpopulation B , and assume, this time, that Y has a continuous distribution in each of the two subpopulations, with cumulative distribution functions $F_A(\cdot)$ and $F_B(\cdot)$, and probability density functions $f_A(\cdot)$ and $f_B(\cdot)$. We imagine Y to be a survival time variable, although we will not consider the possibility of censorship. In the two subpopulations, the survival functions and the hazard functions are given, respectively, by

$$S_k(y) = 1 - F_k(y), \quad h_k(y) = f_k(y)/S_k(y), \quad (5)$$

where y is in the range of Y and $k \in \{A, B\}$. Suppose that the hazard ratio $h_A(y)/h_B(y)$ is constant in y , and denote its value as R . (This is trivially the case if both subpopulations have an exponential distribution, with $h_k(y) = 1/\mu_k$, where μ_k is the subpopulation mean. However, it can also be the case if we assume some other distributional families, such as the Gompertz or Weibull families, or even if we do not assume any specific distributional family, but still assume the proportional hazards model of Cox (1972)[1].) Somers' D is then derived as

$$D(Y|X) = \frac{\int_y h_B(y)S_A(y)S_B(y)dy - \int_y R h_B(y)S_A(y)S_B(y)dy}{\int_y h_B(y)S_A(y)S_B(y)dy + \int_y R h_B(y)S_A(y)S_B(y)dy} = (1 - R)/(1 + R). \quad (6)$$

Note that Somers' D is then the parameter that is zero under the null hypothesis tested using the method of Gehan (1965)[2]. For finite R , this formula can be inverted to give the hazard ratio R as a function of Somers' D by

$$R = [1 - D(Y|X)] / [1 + D(Y|X)] = [1 - c(Y|X)] / c(Y|X), \quad (7)$$

where $c(Y|X) = [D(Y|X) + 1]/2$ is Harrell's c -index, which reparameterizes Somers' D to a probability scale from 0 to 1 (Harrell *et al.*, 1982)[3]. Note that, for continuous Y and binary X , Harrell's c is the probability of concordance, and that a constant hazard ratio R is the corresponding odds *against* concordance.

Figure 2 gives Somers' D as a function of R . Note that R is expressed on a log scale, similarly to the standard practice with logistic regression. Somers' D of lifetime with respect to membership of Population A is seen to be a decreasing logistic sigmoid function of the Population A /Population B log hazard ratio, equal to 0 when the log ratio is 0 and the ratio is therefore 1. A hazard ratio of 2 (as typically found when comparing the lifetimes of cigarette smokers as Population A to lifetimes of nonsmokers as Population B) corresponds to a Somers' D of $-1/3$, or a Harrell's c of $1/3$. Similarly, a hazard ratio of $1/2$ (as typically found when comparing lifetimes of nonsmokers as Population A to lifetimes of cigarette smokers as Population B) corresponds to a Somers' D of $1/3$, or a Harrell's c of $2/3$. Therefore, although a smoker may possibly survive a non-smoker of the same age, the odds are 2 to 1 against this happening. A hazard ratio of 3 corresponds to a Somers' D of -0.5 , and a hazard ratio of $1/3$ corresponds to a Somers' D of 0.5 . For even more spectacular hazard ratios in either direction, the linearity breaks down, even though the hazard ratio is logged.

4 Binary X , Normal Y , equal variances

We assume, again, that X indicates membership of Subpopulation A instead of Subpopulation B , and assume, this time, that Y has a Normal distribution in each of the two subpopulations, with respective means μ_A and μ_B and standard deviations (SDs) σ_A and σ_B . Then, the probability of concordance (Harrell's c) is

the probability that a random member of Population A has a higher Y -value than a random member of Population B , or (equivalently) the probability that the difference between these two Y -values is positive. This difference has a Normal distribution, with mean $\mu_A - \mu_B$ and variance $\sigma_A^2 + \sigma_B^2$. Somers' D is therefore given by the formula

$$D(Y|X) = 2\Phi\left(\frac{\mu_A - \mu_B}{\sqrt{\sigma_A^2 + \sigma_B^2}}\right) - 1, \quad (8)$$

where $\Phi(\cdot)$ is the cumulative standard Normal distribution function. If the two SDs are both equal (to $\sigma = \sigma_A = \sigma_B$), then this formula simplifies to

$$D(Y|X) = 2\Phi\left(\frac{\mu_A - \mu_B}{\sigma\sqrt{2}}\right) - 1 = 2\Phi\left(\frac{\delta}{\sqrt{2}}\right) - 1, \quad (9)$$

where $\delta = (\mu_A - \mu_B)/\sigma$ is the difference between the two means, expressed in units of the common SD. The parameter δ can therefore be defined as a function of Somers' D by the inverse formula

$$\delta = \sqrt{2}\Phi^{-1}\left(\frac{D(Y|X) + 1}{2}\right), \quad (10)$$

where $\Phi^{-1}(\cdot)$ is the inverse standard Normal cumulative distribution function.

Figure 3 gives Somers' D as a function of the mean difference, expressed in SD units. Again, we see a sigmoid curve, but this time Somers' D is increasing with the alternative parameter. Note that a mean difference of 1 SD corresponds to a Somers' D just above 1/2 (approximately .52049988), corresponding to a concordance probability (or Harrell's c) just above 3/4, whereas a mean difference of -1 SDs corresponds to a Somers' D of just below -1/2 (approximately -.52049988), or a Harrell's c just below 1/4. For mean differences between -1 SD and 1 SD, the corresponding Somers' D can be interpolated (approximately) in a linear fashion, with Somers' D approximately equal to half the mean difference in SDs. A mean difference of 2 SDs corresponds to a Somers' D of approximately .84270079 (slightly more than 5/6), or a Harrell's c of approximately .9213504 (slightly over 11/12). A mean difference of 3 SDs corresponds to a Somers' D of approximately .96610515 (slightly more than 19/20), or a Harrell's c of .98305257 (over 49/50). For mean differences over 3 SDs, the probability of discordance falls to fractions of a percent, and Somers' D becomes progressively less useful, as there is very little overlap between subpopulations for Somers' D to measure. Within the range of approximate linearity (± 1 SDs), a Somers' D of 0, $\pm 1/4$, $\pm 1/3$ or $\pm 1/2$ corresponds to a difference of 0, $\pm .45062411$, $\pm .60914039$ or $\pm .95387255$ SDs, respectively.

Note that the above argument applies equally if Y is calculated using a Normalizing and variance-stabilizing monotonic transformation on an untransformed variable whose distribution is not Normal. As Somers' D is a rank parameter, it is preserved by monotonically-increasing transformations. Therefore, *if* a Normalizing and variance-stabilizing increasing transformation exists, *then* we can estimate δ from the Somers' D of the untransformed variable with respect to the binary X by end-point transformation of Somers' D and its confidence limits, using (10). This can be done without even knowing the Normalizing and variance-stabilizing transformation. Therefore, a lot of work can be saved this way, if it was the mean difference in SDs that we wanted to know.

4.1 Likelihood and odds ratios for diagnostic and case-control data

The equal-variance Normal model is also often used as a "toy model" for the problem of defining diagnostic tests, based on a continuous marker variable Y , for membership of Subpopulation A instead of Subpopulation B . In this case, Subpopulation A might be diseased individuals, Subpopulation B might be non-diseased individuals, and Y might be a quantitative test result. If the true distribution of Y in each subpopulation is Normal, with a common subpopulation variance and different subpopulation means, then the log of the likelihood ratio between Subpopulations A and B is a linear function of the Y -value, given by the formula

$$\log \text{LR} = \frac{\mu_B^2 - \mu_A^2}{2\sigma^2} + \left(\frac{\mu_A - \mu_B}{\sigma^2}\right)Y. \quad (11)$$

Note that the intercept term is the value of the log likelihood ratio if $Y = 0$, whereas the slope term is equal to δ/σ in our notation. Therefore, δ , the mean difference expressed in standard deviations, is the slope of $\log \text{LR}$ with respect to Y/σ , the Y -value expressed in standard deviations. In Bayesian inference, this log likelihood ratio is added to the log of the prior (pre-test) odds of membership of Subpopulation A instead of Subpopulation B , to derive the log of the posterior (post-test) odds of membership of Subpopulation A instead of Subpopulation B . The role of Somers' D in diagnostic tests is discussed in Newson (2002)[7]. Briefly, if we graph true-positive rate (sensitivity) against false-positive rate (1-specificity), and choose

points on the graph corresponding to the various possible test thresholds, and join these points in ascending order of candidate threshold value, then the resulting curve is known as the sensitivity–specificity curve, or (alternatively) as the receiver–operating characteristic (ROC) curve. Harrell's c is the area below the ROC curve, whereas Somers' D is the difference between the areas below and above the ROC curve. The likelihood ratio is the slope of the ROC curve (defined as the derivative of true–positive rate with respect to false–positive rate), and, in the equal–variance Normal model, is computed by exponentiating (11). The equal–variance Normal model, predicting the test result from the disease status, is therefore combined with Bayes' theorem to imply a logistic regression model for estimating the disease status from the test result. Figure 3 therefore implies (unsurprisingly) that, other things being equal, the ROC curve becomes higher as the mean difference (in SDs) becomes higher.

The equal–variance Normal “toy model” is also useful for interpreting case–control studies. In these studies, Subpopulation A is the cases, Subpopulation B is the controls, and Y is a continuous disease predictor. The rare–disease assumption implies that the distribution of Y in the controls is a good approximation to the distribution of Y in the population, and that a population relative risk can be estimated using the corresponding odds ratio. The 100 q th percentile of Y in the control population is then $\xi_q = \mu_B + \sigma\Phi^{-1}(q)$ for q in the interior of the unit interval. It follows from (11) that, for $q < 0.5$, the log of the interpercentile odds ratio between ξ_{1-q} and ξ_q is defined as the difference between the log likelihood ratios at ξ_{1-q} and ξ_q . If we express Y in SD units, then, by (11), this difference is given by

$$\log \text{OR}_q = \delta [\Phi^{-1}(1-q) - \Phi^{-1}(q)] = 2\delta\Phi^{-1}(1-q). \quad (12)$$

For instance, if $q = 1/4 = 0.25$, then OR_q is the interquartile odds ratio, equal to

$$\text{OR}_{0.25} = \exp \{ \delta [\Phi^{-1}(0.75) - \Phi^{-1}(0.25)] \} = \exp [2\delta\Phi^{-1}(0.75)]. \quad (13)$$

The plot of Somers' D against the interquartile odds ratio $\text{OR}_{0.25}$ (on a log scale) is given as Figure 4. A Somers' D of $1/12$, $1/6$, $1/4$, $1/3$, $5/12$ or $1/2$ is equivalent to an interquartile odds ratio of 1.2209313 , 1.4939801 , 1.8365388 , 2.2744037 , 2.847498 or 3.6210155 , respectively. Again, this argument is still valid if Y is derived from an original non–Normal variable, using a Normalizing and variance–stabilizing monotonically–increasing transformation. And, again, Somers' D can be estimated from the original untransformed variable, without knowing the transformation.

In the range of Somers' D values from -0.5 to 0.5 , the relationship between Somers' D and the log of the interpercentile odds ratio is approximately linear (see Figure 4 and observe the linearity from interquartile odds ratios $1/4 = 0.25$ to 4), corresponding approximately to Somers' D from -0.5 to 0.5). From (12) and (10), we have for $q \in (0, 1)$,

$$\log \text{OR}_q = 2\sqrt{2}\Phi^{-1}(1-q)\Phi^{-1} \left(\frac{D(Y|X) + 1}{2} \right) = 2\sqrt{2}\Phi^{-1}(1-q)\Phi^{-1}[c(Y|X)], \quad (14)$$

where $c(Y|X) = [D(Y|X) + 1]/2$ is Harrell's c of Y with respect to X (Harrell *et al.*, 1982)[3]. For $c(Y|X) \in [0.25, 0.75]$ (corresponding to $D(Y|X) \in [-0.5, 0.5]$), (14) can be approximated linearly by the interpolation formula

$$\log \text{OR}_q \approx 4\sqrt{2}\Phi^{-1}(1-q) \{ [c(Y|X) - 0.25] \Phi^{-1}(0.75) + [0.75 - c(Y|X)] \Phi^{-1}(0.25) \}. \quad (15)$$

Figure 5 shows the exact and approximate interquartile range (implied by exponentiating (15)) as a function of Harrell's c under the two–sample equal–variance Normal model. Note that the horizontal axis is labelled at non–negative multiples of $k/24$ for Harrell's c , corresponding to the possibly–negative multiples of $k/12$ for Somers' D . Note that the dotted line for the approximate formula is a good approximation to the exact formula for Harrell's c from 0.25 to 0.75 , but not outside that range.

This formula can be useful in interpreting differences between Harrell's c indices for different predictors. If we have 2 predictors Y and Z , then confidence intervals for differences between $D(Y|X)$ and $D(Z|X)$, or for differences between $c(Y|X)$ and $c(Z|X)$, may be used to eliminate the possibility that a positive association between Y and X might be secondary to a positive association of both X and Y with Z . (See Newson (2006)[6] and Newson (2002)[7].) Unfortunately, a lot of statistically–minded scientists think that they understand odds ratios (and their ratios), but do not think that they understand Somers' D and Harrell's c (and their differences). To prevent culture shock in such scientists, we can re–interpret the difference between Harrell's c indices in terms of ratios of odds ratios, at least if the Harrell's c indices being compared are between 0.25 and 0.75 , and if we also assume that Y and Z can both be transformed to standard Normal variables. For $q \in (0, 1)$, define $\text{OR}_q(Y|X)$ as the odds ratio for X between percentiles $100(1-q)$ and $100q$ of Y , and define $\text{OR}_q(Z|X)$ as the odds ratio for X between percentiles $100(1-q)$ and $100q$ of Z . From (15), we have the linear approximation

$$\log [\text{OR}_q(Y|X)/\text{OR}_q(Z|X)] \approx 8\sqrt{2}\Phi^{-1}(1-q)\Phi^{-1}(0.75) [c(Y|X) - c(Z|X)]. \quad (16)$$

This parameter is a ratio between odds ratios, and implies, in the case of a ratio between interquartile ratios, that

$$\text{OR}_{0.25}(Y|X)/\text{OR}_{0.25}(Z|X) = \exp \left\{ 8\sqrt{2} [\Phi^{-1}(0.75)]^2 [c(Y|X) - c(Z|X)] \right\}. \quad (17)$$

This means that differences between c -indices of 0.05, 0.01, 0.005, 0.001, 0.0005 and 0.0001 correspond to ratios between interpercentile odds ratios of 1.2934989, 1.0528178, 1.0260691, 1.0051603, 1.0025768 and 1.0005148, respectively. So, although differences between c -indices may be very small and come with narrow confidence intervals, and may be good at proving that an association cannot be secondary to another association, they are not very good at proving non-inferiority of a predictive score to another predictive score, unless we have an enormous dataset, because a difference in c -index of a fraction of a percent does not rule out an increase in predictive power that may be important, at least for some of the people, some of the time.

The above arguments about interpercentile odds ratios might possibly be extended to the case where X is a possibly-censored survival-time variable instead of an uncensored binary variable. In this case, the interpercentile odds ratio would be replaced by an interpercentile hazard ratio, the ratio between interpercentile odds ratios would be replaced by the ratio between interpercentile hazard ratios, and the equally-variable Normal distributions would be assumed to belong to non-survivors and survivors, instead of to cases and controls. Note that, unlike the case with a constant hazard ratio for a continuous Y between binary X -groups, it would then be X that is the lifetime variable, and the continuous Y that is the predictor variable.

5 Bivariate Normal X and Y

We assume, this time, that X and Y have a joint bivariate Normal distribution, with means μ_X and μ_Y , SDs σ_X and σ_Y , and a Pearson correlation coefficient ρ . As both X and Y are continuous, Somers' D is equal to Kendall's τ_a , and is therefore given by the formula

$$D(Y|X) = \frac{2}{\pi} \arcsin(\rho), \quad (18)$$

which can be inverted to give

$$\rho = \sin \left[\frac{\pi}{2} D(Y|X) \right]. \quad (19)$$

This relation is known as Greiner's relation. The curve of (18) is illustrated in Figure 6. Note that Pearson correlations of $-\sqrt{1/2}$, $-1/2$, 0 , $1/2$ and $\sqrt{1/2}$ correspond to Kendall correlations (and therefore Somers' D values) of $-1/2$, $-1/3$, 0 , $1/3$ and $1/2$, respectively. This implies that audiences accustomed to Pearson correlations may be less impressed when presented with the same correlations on the Kendall-Somers scale. A possible remedy for this problem is to use the end-point transformation (19) on confidence intervals for Somers' D or Kendall's τ_a to define outlier-resistant confidence intervals for the Pearson correlation.

This practice of end-point transformation is also useful if we expect variables X and Y not to have a bivariate Normal distribution themselves, but to be transformed to a pair of bivariate Normal variables by a pair of monotonically increasing transformations $g_X(X)$ and $g_Y(Y)$. As Somers' D and Kendall's τ_a are rank parameters, they will not be affected by substituting X for $g_X(X)$ and/or substituting Y for $g_Y(Y)$. Therefore, the end-point transformation method can be used to estimate the Pearson correlation between $g_X(X)$ and $g_Y(Y)$, without even knowing the form of the functions $g_X(\cdot)$ and $g_Y(\cdot)$. This can save a lot of work, if the Pearson correlation between the transformed variables was what we wanted to know.

Greiner's relation, or something very similar, is expected to hold for a lot of other bivariate continuous distributions, apart from the bivariate Normal. Kendall (1949)[5] showed that Greiner's relation is not affected by odd-numbered moments, such as skewness. Newson (1987)[8], using a much simpler argument, discussed the case where two variables X and Y are defined as sums or differences of up to 3 latent variables (hidden variables) U , V and W , which were assumed to be sampled independently from an arbitrary common continuous distribution. It was shown that different definitions of X and Y implied the values of Kendall's τ_a and Pearson's correlation displayed in Table 1. These pairs of values all lie along the line of Greiner's relation, as displayed in Figure 6.

6 Acknowledgement

I would like to thank Raymond Boston of Pennsylvania University, PA, USA for raising the issue of interpretations of Somers' D , and for prompting me to summarize these multiple interpretations in a single document.

Table 1: Kendall and Pearson correlations for X and Y defined in terms of independent continuous latent variables U , V and W .

X	Y	Kendall's τ_a	Pearson's ρ
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

References

- [1] Cox DR. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B* 1972; **34(2)**: 187–220.
- [2] Gehan EA. A generalized Wilcoxon test for comparing arbitrarily single-censored samples. *Biometrika* 1965; **52(1/2)**: 203–223.
- [3] Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *Journal of the American Medical Association* 1982; **247(18)**: 2543–2546.
- [4] Kendall MG, Gibbons JD. *Rank Correlation Methods. 5th Edition*. New York, NY: Oxford University Press; 1990.
- [5] Kendall MG. Rank and product-moment correlation. *Biometrika* 1949; **36(1/2)**: 177–193.
- [6] Newson R. Confidence intervals for rank statistics: Somers' D and extensions. *The Stata Journal* 2006; **6(3)**: 309–334.
- [7] Newson R. Parameters behind “nonparametric” statistics: Kendall's tau, Somers' D and median differences. *The Stata Journal* 2002; **2(1)**: 45–64.
- [8] Newson RB. *An analysis of cinematographic cell division data using U -statistics [DPhil dissertation]*. Brighton, UK: Sussex University; 1987.
- [9] Somers RH. A new asymmetric measure of association for ordinal variables. *American Sociological Review* 1962; **27(6)**: 799–811.

Figure 1: Somers' D and difference between proportions in the two-sample binary model.

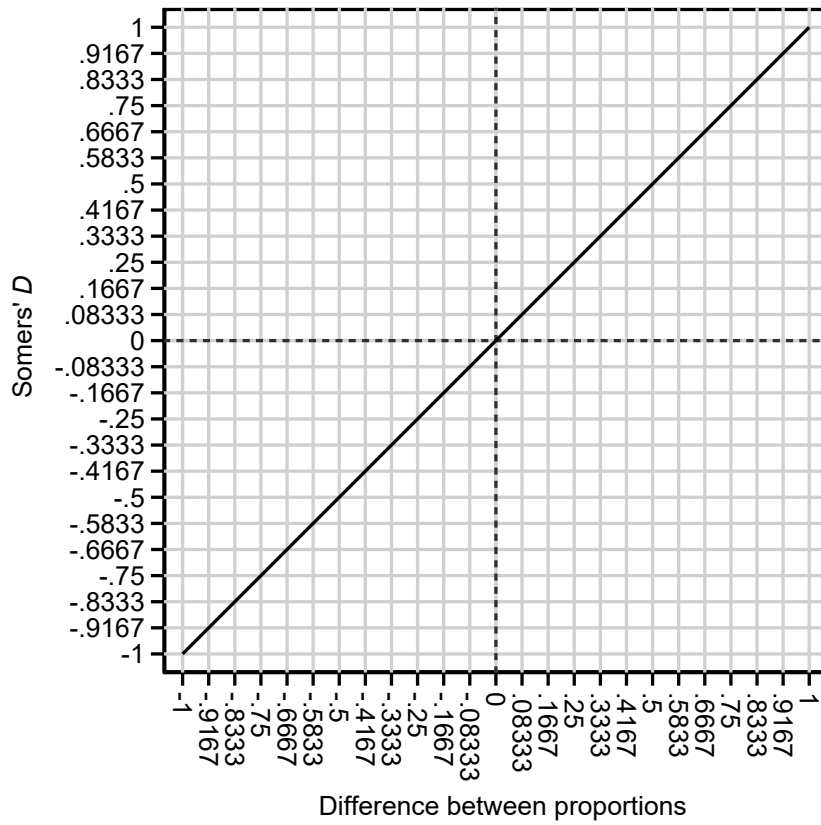


Figure 2: Somers' D and hazard ratio in the two-sample constant hazard ratio model.

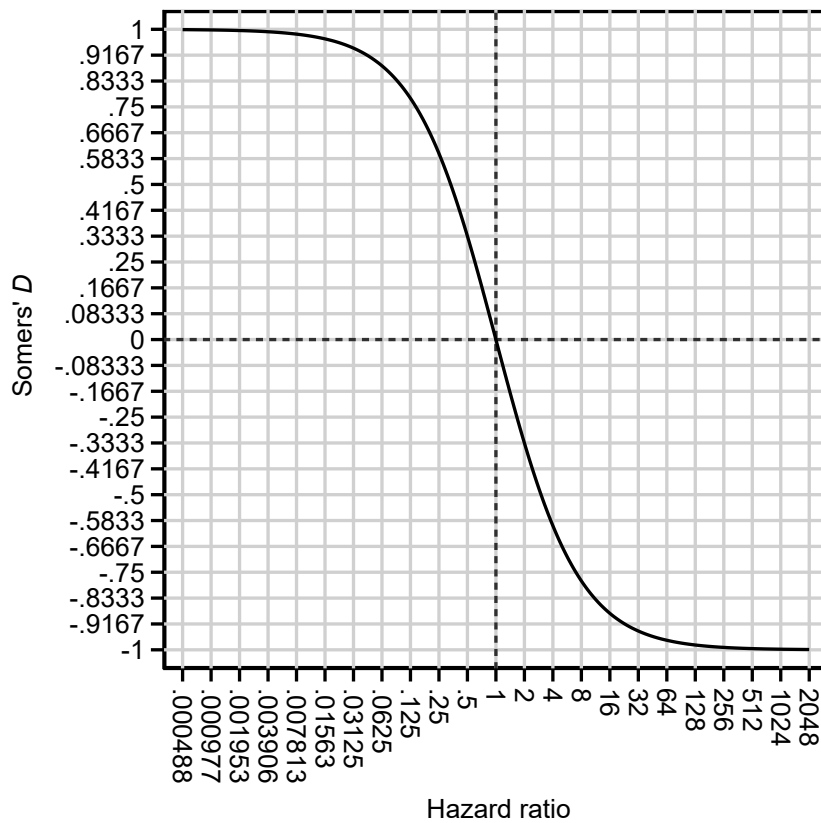


Figure 3: Somers' D and mean difference in SDs in the two-sample equal-variance Normal model.

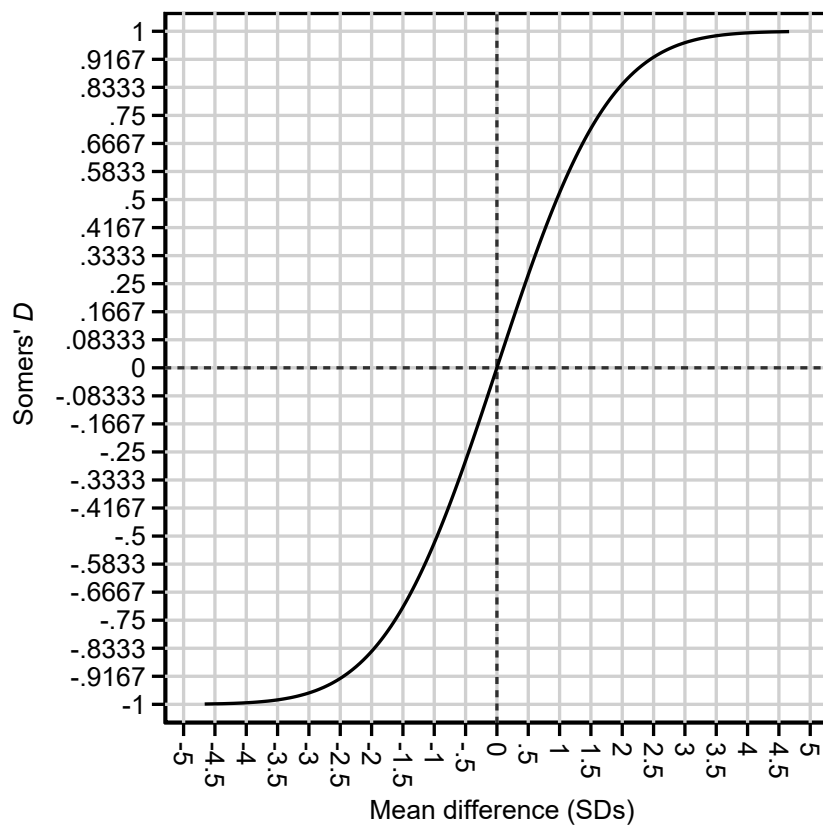


Figure 4: Somers' D and interquartile odds ratio in the two-sample equal-variance Normal model.

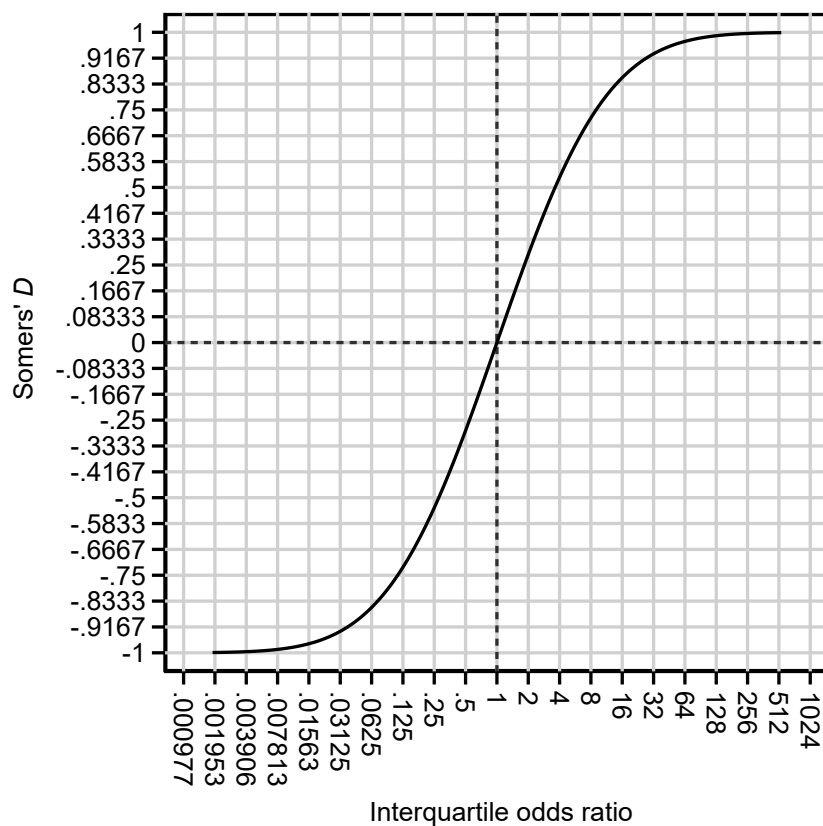


Figure 5: Interquartile odds ratio and Harrell's c in the two-sample equal-variance Normal model.

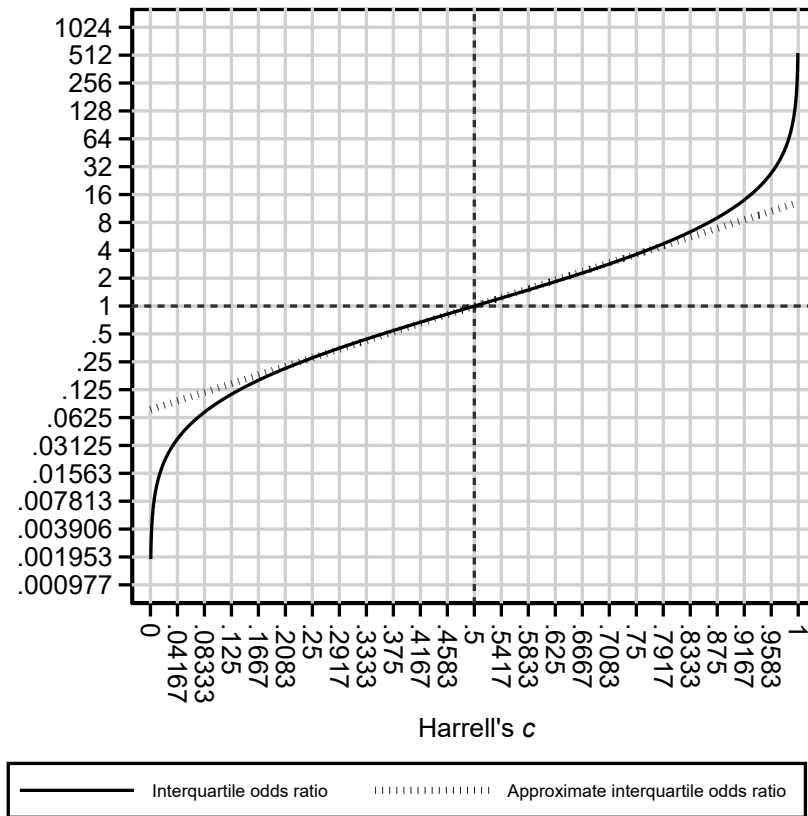


Figure 6: Somers' D and Pearson correlation in the bivariate Normal model.

