Somers' d

ST Corner
October 2011

Clay Ford, M.S.

Transforming Clinical Trials

Background

- Working on a Phase II study
- QC programming of efficacy tables
  - One table requested Somers' d
  - Had no idea what it was
  - An ST Corner was born

What is Somers' d?

- Measure of association between two ordinal variables
- Makes distinction between independent and dependent variables (asymmetric)
- Ranges in value from -1 to 1
- Named for Robert H. Somers
  - *A New Asymmetric Measure of Association for Ordinal Variables.* American Sociological Review, Vol. 27, No. 6 (Dec., 1962), pp. 799-811

The Setting

- An I × J contingency table with both classifications ordinal
- When both variables are ordinal, a monotone trend in association is common. As the level of X increases…
  - responses on Y tend to increase toward higher levels, or
  - responses on Y tend to decrease toward lower levels

The Goals

- Measure the association between the ordinal variables (i.e., describe the degree to which the relationship is monotone)
- We want to be able to accommodate a response variable

Why not use chi-square test?

- Does not measure association
  - using sample to test the inference of statistical independence
  - not a useful descriptive statistic
- It is insensitive to ordering
  - less powerful against population hypotheses of monotone correlation
  - when monotone correlation exists, it is less likely to reject the null hypothesis

Getting Started

- Classify each pair of subjects as concordant, discordant, or tied
  - A pair is concordant if the subject ranked higher on X ranks higher on Y.
  - A pair is discordant if the subject ranked higher on X ranks lower on Y.
  - A pair is tied if subjects have the same classification on X and/or Y.

Example

	Condition (Y)		
Dose (X)	Worse	Same	Better
Placebo	4	15	3
5 mg	2	11	12
10 mg	1	9	15

Concordant Pairs

Discordant Pairs

Tied Pairs

- A pair is concordant if the subject ranked higher on X ranks higher on Y.
- A pair is discordant if the subject ranked higher on X ranks lower on Y.
- A pair is tied if subjects have the same classification on X and/or Y.

Somers' d

- C = total number of concordant pairs
- D = total number of discordant pairs
- n = number of subjects
- Somers' d:

$$d = \frac{C - D}{\binom{n}{2} - \sum_{i=1}^{k} \binom{n_i}{2}}$$

All possible pairs

All pairs tied on X (i.e., the independent variable)

$$\boxed{-1 \le d \le 1}$$

Counting Concordant Pairs

	Condition		
Dose	Worse	Same	Better
Placebo	4	15	3
5 mg	2	11	12
10 mg	1	9	15

- Concordant pairs:
  C = 4*(11 + 12 + 9 + 15) + 2*(9 + 15) + 15*(12 + 15) + 11(15) = 806

Counting Discordant Pairs

	Condition		
Dose	Worse	Same	Better
Placebo	4	15	3
5 mg	2	11	12
10 mg	1	9	15

- Discordant pairs:
  C = 3*(2+ 11 + 1 + 9) + 15*(2 + 1) + 12*(1 + 19) + 11(1)= 245

Counting All Pairs tied on X

	Condition			
Dose	Worse	Same	Better	sum of subjects on dose
Placebo	4	15	3	22
5 mg	2	11	12	25
10 mg	1	9	15	25

$$\left[\binom{22}{2} = 231\right] + \left[\binom{25}{2} = 300\right] + \left[\binom{25}{2} = 300\right] = 831$$

Calculating Somers' d

	Condition		
Dose	Worse	Same	Better
Placebo	4	15	3
5 mg	2	11	12
10 mg	1	9	15

Interpretation:
surplus of concordant pairs as a percentage of concordant, discordant, and relevant tied pairs of observations

$$d = \frac{C - D}{\binom{n}{2} - \sum_{i=1}^{k} \binom{n_i}{2}} = \frac{806 - 245}{\binom{72}{2} - 831} = \frac{561}{1725} = 0.325$$

Calculating Somers' d (2)

	Condition		
Dose	Worse	Same	Better
Placebo	4	15	3
5 mg	2	11	12
10 mg	1	9	15

Concordant pairs: 806

Discordant pairs: 245

Relevant pairs: 1725

$$d = \left(\frac{806}{1725} - \frac{245}{1725} \right) = 0.467 - 0.142 = 0.325$$

About 33% of relevant pairs are surplus concordant pairs

Calculating Somers' d – SAS style

- SAS formula for Somers' d:
- Same as first formula:

$$\boxed{d = \frac{2C - 2D}{n^2 - \sum_{i=1}^{k} n_i^2}}$$

  - lose the factorials in denominator and simplify:

$$\binom{n}{2} - \sum_{i=1}^{k} \binom{n_i}{2} = \frac{n!}{2!(n-2)!} - \sum \frac{n_i}{2!(n_i - 2)!}$$

$$= \frac{n(n-1)}{2} - \sum \frac{n_i(n_i - 1)}{2} = \frac{n^2 - n}{2} - \sum \frac{n_i^2}{2} + \frac{n}{2} = \frac{n^2}{2} - \sum \frac{n_i^2}{2}$$

  - multiply by 2/2 = 1

$$\frac{C - D}{\frac{n^2}{2} - \frac{\sum n_i^2}{2}} \times \frac{2}{2} = \frac{2C - 2D}{n^2 - \sum n_i^2}$$

Somers' d in SAS

- Use PROC FREQ

```
proc freq data=study;
  weight cnt;
  table dose*cond;
  test smdcr;
run;
```

If column variable is response, use `smdcr`

If row variable is response, use `smdrc`

- Output

```
              Somers' D C|R
----------------------------------------
Somers' D C|R                    0.3252
ASE                              0.0872
95% Lower Conf Limit             0.1543
95% Upper Conf Limit             0.4961

   Test of H0: Somers' D C|R = 0

ASE under H0                     0.0877
Z                                3.7098
One-sided Pr >  Z                0.0001
Two-sided Pr > |Z|               0.0002

          Sample Size = 72
```

Somers' d Inference

- In addition to Somers' d, SAS provides
  - standard error
  - 95% CI (d ± 1.96 * SE)
  - Hypothesis test: H_0: d = 0
    - test statistic: Z = d/SE
    - has N(0,1) distribution

Somers' D C|R

Somers' D C|R	0.3252
ASE	0.0872
95% Lower Conf Limit	0.1543
95% Upper Conf Limit	0.4961

Test of H0: Somers' D C|R = 0

ASE under H0	0.0877
Z	3.7098
One-sided Pr > Z	0.0001
Two-sided Pr > |Z|	0.0002

Sample Size = 72

- The standard error formulas
  - can be found in SAS documentation
  - look awful

Somers’ d simulation

- To find evidence for using d ± 1.96 * SE to construct a 95% CI:
  - Simulate 10,000 classification tables
  - Calculate Somers’ d each time
  - Take the standard deviation of the 10,000 Somers’ d values
  - find the 2.5% and 97.5% quantiles to determine a 95% interval
  - compare to SAS output

Somers' d simulation (2)

- R code:

```
d <- c()
sxi <- choose(22,2) + choose(25,2) + choose(25,2)
N <- choose(72,2)
for (i in 1:10000){
          a <- rmultinom(1, 22, c(4/22,15/22,3/22))
          b <- rmultinom(1, 25, c(2/25,11/25,12/25))
          c <- rmultinom(1, 25, c(1/25,9/25,15/25))
          xy <- t(cbind(a,b,c))
          C <- xy[1,1]*sum(xy[2:3,2:3])+ xy[2,1]*sum(xy[3,2:3])+
                    xy[1,2]*sum(xy[2:3,3])+ xy[2,2]*xy[3,3]
          D <- xy[1,3]*sum(xy[2:3,1:2])+ xy[1,2]*sum(xy[2:3,1])+
                    xy[2,3]*sum(xy[3,1:2])+ xy[2,2]*xy[3,1]
          d[i] <- (C-D)/(N - sxi)
          }
sd(d) # compare to SAS output
quantile(d,c(0.025,0.975)) # 95% CI
hist(d) # looks normal
```

- Output:

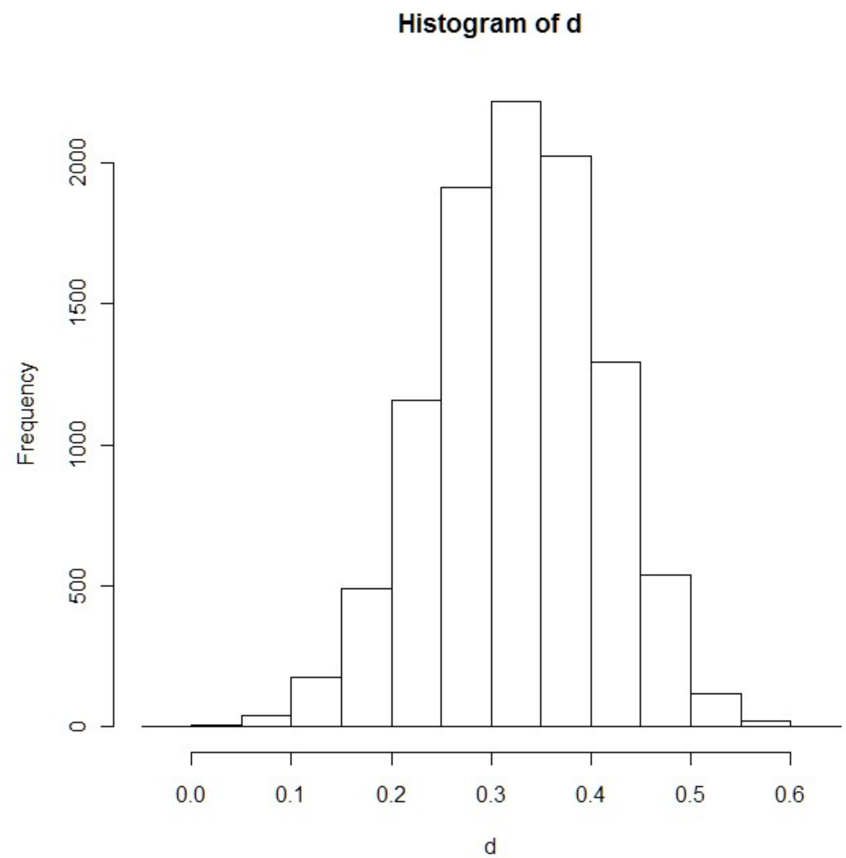
```
> sd(d) # compare to SAS output
[1] 0.08485863
> quantile(d,c(0.025,0.975)) # 95% CI
     2.5%     97.5%
0.1553623 0.4857971
```

Somers' D C|R

Somers' D C|R	0.3252
ASE	0.0872
95% Lower Conf Limit	0.1543
95% Upper Conf Limit	0.4961

Somers' d simulation (3)

- Frequency histogram of 10,000 Somers' d



Relation to Gamma

$$\gamma = \frac{C - D}{C + D}$$

- Does not account for ties, therefore it tends to exaggerate the actual strength of association.
- Treats variables symmetrically; do not need to identify one variable as a response.
- Somers' d has same numerator, larger denominator (penalizes for pairs tied on response)

Relation to Gamma (2)

$$\gamma = \frac{C - D}{C + D}$$

- Somers' d has same numerator, larger denominator (penalizes for pairs tied on response)

$$d = \frac{C - D}{\binom{n}{2} - \sum_{i=1}^{k} \binom{n_i}{2}} = \frac{C - D}{C + D + Y_t}$$

where Y_t = pairs tied on response

References

Agresti, A. (1990), *Categorical Data Analysis, 2nd Edition*, New York: John Wiley & Sons, Inc. Pages 57 – 59, 68.

SAS STAT User's Guide – The FREQ Procedure – Details – Statistical Computations – Measures of Association

Somers, R.H. (1962), "A New Asymmetric Measure of Association for Ordinal Variables," *American Sociological Review*, 27, 799 - 811.

Thank you!

Questions or comments?