

A SAS® Macro for Improved Correlation Coefficient Inference

Stephen W. Looney, Georgia Regents University

ABSTRACT

We present a SAS macro for improved statistical inference for measures of association, including Pearson's correlation, Spearman's coefficient, and Kendall's coefficient. While the CORR procedure is a powerful tool for calculating and testing these coefficients, some analyses are lacking. For example, the CORR procedure does not incorporate recent theoretical improvements in confidence interval estimation for Spearman's rho, nor does it provide any confidence interval at all for Kendall's tau-b. We have written a SAS macro that incorporates these new developments; it produces confidence intervals, as well as p -values for testing any null value of these coefficients.

Key words: confidence interval, Fisher z-transformation, Kendall coefficient, p -value, Spearman coefficient

INTRODUCTION

Correlation is one of the most commonly used statistical methods, and the Pearson correlation coefficient (PCC) is the most commonly used measure of the association between two variables X and Y . If both variables X and Y appear to be normally distributed, the PCC is the appropriate measure to use. If the data for either variable appear to be non-normally distributed, then a distribution-free method for measuring association should be considered. (For our purposes, "distribution free" refers to a test statistic whose null distribution does not depend on any underlying distributional assumption.) In this presentation, we consider the two most commonly used distribution-free measures of association: Spearman's correlation coefficient (SCC) and Kendall's coefficient of concordance (KCC).

STATISTICAL INFERENCE FOR CORRELATION COEFFICIENTS

Let ρ denote the true value of the PCC. Then there is an exact test of $H_0 : \rho = 0$ based on the t distribution. The test statistic is given by:

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, \quad (1)$$

which is compared with the Student's t -distribution with $n - 2$ degrees of freedom to obtain the exact p -value. The problem with using this approach is that the test statistic in (1) can be used to test only the null value $\rho_0 = 0$. This means that there is no corresponding confidence interval procedure that will yield a conclusion identical to that obtained when (1) is used as the test statistic. An alternative approach is to use the Fisher z-transformation to transform the sample correlation r to a new random variable that is approximately normally distributed. This transformed random variable can then be used to derive a test statistic for testing $H_0 : \rho = \rho_0$, where ρ_0 is any hypothesized value of the PCC (zero, negative, or positive). This transformed random variable can also be used to derive an approximate confidence interval for the true value of the PCC.

The Fisher z-transformed PCC is given by:

$$z(r) = \frac{1}{2} \log \frac{1+r}{1-r}, \quad (2)$$

where r is the sample value of the PCC. This new random variable has an approximate normal distribution with mean 0 and variance $1/(n - 3)$. This result can be used to derive a test statistic for testing $H_0 : \rho = \rho_0$:

$$z_0 = \frac{z(r) - z(\rho_0)}{\sqrt{n-3}}, \quad (3)$$

where n is the sample size, $z(r)$ is the Fisher z-transformation applied to the sample value of the PCC, and $z(\rho_0)$ is the Fisher z-transformation applied to the hypothesized value (usually 0, but any value between -1 and 1 is possible). The value of z_0 in (3) is then compared with the standard normal distribution to obtain an approximate p -value. Unless n is extremely small, there will be very little difference in the p -values obtained when using the test statistics in

(1) and (3).

The Fisher z-transformation can also be used to derive an approximate $100(1-\alpha)\%$ confidence interval (C.I.) for ρ that will always yield a conclusion consistent with that obtained using the hypothesis test based on (3). To derive the formula for the C.I. for ρ , one must first derive a Wald-type approximate C.I. for the Fisher-z transformed value of the population correlation $z(\rho) = \frac{1}{2} \log \frac{1+\rho}{1-\rho}$. This C.I. is given by:

$$z(r) \pm z_{\alpha/2} \sqrt{1/(n-3)}, \quad (4)$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ -percentage point of the standard normal.

Next, the endpoints of the C.I. in (4) are "back transformed" using the inverse of the Fisher-z transformation. (In other words, the equation in (2) is solved for r in terms of z .) This yields the following approximate $100(1-\alpha)\%$ for ρ :

$$\left(\frac{\exp[2(LCL)] - 1}{\exp[2(LCL)] + 1}, \frac{\exp[2(UCL)] - 1}{\exp[2(UCL)] + 1} \right), \quad (5)$$

where

$$\begin{aligned} LCL &= z(r) - z_{\alpha/2} \sqrt{1/(n-3)} \quad \text{and} \\ UCL &= z(r) + z_{\alpha/2} \sqrt{1/(n-3)}. \end{aligned} \quad (6)$$

The Fisher z-transformation can also be applied to the Spearman and Kendall coefficients (Fieller et al. 1957, Bonett and Wright 2000). Let r_s denote the sample value of the SCC, let ρ_s denote the true value, and let ρ_{s0} denote the null value to be tested. Similarly, let t_s denote the sample value of the KCC, let τ_s denote the true value, and let τ_{s0} denote the null value to be tested. The test statistic for testing $H_0 : \rho_s = \rho_{s0}$ is given by:

$$z_{s0} = \frac{z(r_s) - z(\rho_{s0})}{\sqrt{(1 + \rho_{s0}^2 / 2) / (n-3)}}, \quad (7)$$

and the test statistic for testing $H_0 : \tau_b = \tau_{b0}$ is given by:

$$z_{t0} = \frac{z(t_b) - z(\tau_{b0})}{\sqrt{0.437 / (n-4)}}. \quad (8)$$

Note that the more accurate approximation to the distribution of $z(r_s)$ proposed by Bonett and Wright (2000) has been incorporated into the test statistic for the SCC in Equation (7). In the more traditional approach, $1 + \rho_{s0}^2 / 2$ in (7) would be replaced by 1.06 (Fieller et al. 1957). (The CORR procedure in SAS uses the test statistic in (1) with the value of r replaced with r_s to calculate the p-value for their "exact" test of $H_0 : \rho_s = 0$. It also uses a null distribution for the sample KCC different from the one used here when testing $H_0 : \tau_b = 0$.)

Approximate confidence limits for the true values of ρ_s and τ_b can be derived using the same approach used in deriving the confidence limits in (5) for the PCC. Thus, an approximate $100(1-\alpha)\%$ confidence interval for $z(\rho_s)$ is given by:

$$z(r_s) \pm z_{\alpha/2} \sqrt{(1 + r_s^2 / 2) / (n-3)} \quad (9)$$

and an approximate $100(1-\alpha)\%$ confidence interval for $z(\tau_b)$ is given by:

$$z(t_b) \pm z_{\alpha/2} \sqrt{0.437 / (n-4)}. \quad (10)$$

Note that the more accurate approximation to the distribution of $z(r_s)$ proposed by Bonett and Wright (2000) has been incorporated into the confidence interval formula in Equation (9). In the more traditional approach, $1 + r_s^2 / 2$ in (9) would be replaced by 1.06 (Fieller et al. 1957). (Note that the CORR procedure in SAS uses a value of 1 instead of 1.06 when calculating confidence limits for ρ_s .) Approximate confidence limits for ρ_s and τ_b are obtained by applying the "back-transform" in (5) to the limits obtained in (9) and (10). For additional details on these approaches, see Bonett and Wright (2000).

SAS MACRO FOR SPEARMAN INFERENCE

```

%macro spearman(data=, coeff=, var1=, var2=, rho=, c_sq=);
*   Generate standard results for Pearson correlation   ;
proc corr data = &data pearson fisher(biasadj=no RH00=&rho) nosimple;
  var &var1;
  with &var2 ;
  title 'PROC CORR Results for Pearson Correlation';
*   Generate default results for Spearman correlation   ;
proc corr data = &data spearman fisher(biasadj=no RH00=&rho) nosimple outs = spearman;
  var &var1;
  with &var2 ;
  title 'PROC CORR Results for Spearman Correlation';
  title2 'This Analysis is Not Recommended';
*   Perform improved analysis for Spearman using Bonett & Wright transformation;
data corr;
set spearman;
if _TYPE_ = 'CORR';
rs = &var1;
i+1;
data samp;
set spearman;
if _TYPE_ = 'N';
n = &var1;
i+1;
data combined;
merge corr samp;
by i;
run;
data _null_; set combined;
call symput("n",n); call symput("rs",rs);
run;
data fisher_rs;
n_=&n;
rs_=&rs;
rho_=&rho;
coeff_ = &coeff;
z_crit=probit((1+(coeff_/100))/2);
c_sq_ = &c_sq;
if &c_sq = . then c_sq_ = 1+(rs_**2)/2;
rs_fisher=0.5*(log((1+rs_)/(1-rs_)));
rho_fisher=0.5*(log((1+rho_)/(1-rho_)));
low_rsz=rs_fisher-z_crit*sqrt(c_sq_)/(sqrt(n_-3));
up_rsz=rs_fisher+z_crit*sqrt(c_sq_)/(sqrt(n_-3));
low_rs=(exp(2*low_rsz)-1)/(exp(2*low_rsz)+1);
up_rs=(exp(2*up_rsz)-1)/(exp(2*up_rsz)+1);
se_rs = sqrt((c_sq_)/(n_-3));
z_rs = (rs_fisher - rho_fisher)/se_rs;
p_val_rs = 2*(1-probnorm(abs(z_rs)));
label rs_="Spearman@Coefficient"
      n_="Sample@Size"
      rs_fisher="Fisher-z"
      c_sq_="c squared"
      se_rs="Standard@Error"
      coeff_="Confidence@Coefficient"
      low_rs="Lower@Confidence@Limit"
      up_rs="Upper@Confidence@Limit"
      rho_ = "Null@Value"
      p_val_rs="p-value";
run;
title 'Spearman Correlation Analysis Using Bonett and Wright Transformation';
proc print label split='@' noobs;
  var rs_ n_ rs_fisher c_sq_ se_rs coeff_ low_rs up_rs rho_ p_val_rs;
run;
quit;
%mend spearman;

```

SAS MACRO FOR KENDALL INFERENCE

```

%macro kendall(data=, coeff=, var1=, var2=, kappa= );
*   Generate standard results for Pearson correlation   ;
proc corr data = &data pearson fisher(biasadj=no RHO0=&kappa) nosimple;
  var &var1;
  with &var2 ;
  title 'PROC CORR Results for Pearson Correlation';
*   Generate default results for Kendall coefficient   ;
proc corr data = &data kendall nosimple outk = kendall;
  var &var1;
  with &var2 ;
  title 'PROC CORR Results for Kendall Correlation';
  title2 'Note that SAS 9.4 Does Not Perform Inference Other Than Test Kappa = 0';
*   Perform analysis for Kendall using Fieller transformation   ;
data corr;
set kendall;
if _TYPE_ = 'CORR';
rk = &var1;
i+1;
data samp;
set kendall;
if _TYPE_ = 'N';
n = &var1;
i+1;
data combined;
  merge corr samp;
  by i;
  run;
data _null_; set combined;
call symput("n",n); call symput("rk",rk);
run;
data fisher_rk;
n_=&n;
rk_=&rk;
kappa_=&kappa;
coeff_ = &coeff;
z_crit=probit((1+(coeff_/100))/2);
rk_fisher=0.5*(log((1+rk)/(1-rk)));
kappa_fisher=0.5*(log((1+kappa_)/(1-kappa_)));
low_rkz=rk_fisher-z_crit*sqrt(0.437)/(sqrt(n_-4));
up_rkz=rk_fisher+z_crit*sqrt(0.437)/(sqrt(n_-4));
low_rk=(exp(2*low_rkz)-1)/(exp(2*low_rkz)+1);
up_rk=(exp(2*up_rkz)-1)/(exp(2*up_rkz)+1);
se_rk = sqrt((0.437)/(n_-4));
z_rk = (rk_fisher - kappa_fisher)/se_rk;
p_val_rk = 2*(1-probnorm(abs(z_rk)));
label rk_="Kendall@Coefficient"
      n_="Sample@Size"
      rk_fisher="Fisher-z"
      se_rk="Standard@Error"
      coeff_="Confidence@Coefficient"
      low_rk="Lower@Confidence@Limit"
      up_rk="Upper@Confidence@Limit"
      kappa_ = "Null@Value"
      p_val_rk="p-value";
  run;
title 'Kendall Coefficient Analysis Using Bonett and Wright Transformation';
proc print label split='@' noobs;
var rk_ n_ rk_fisher se_rk coeff_ low_rk up_rk kappa_ p_val_rk;
run;
quit;
%mend kendall;

```

MACRO EXECUTION

In order to use either the spearman or kendall macros, the user first creates a SAS data set containing their data.

The spearman macro is then called using the following arguments:

data: original data set where the variable to be correlated can be found;

coeff: confidence coefficient, expressed as a percentage between 0 and 100 (exclusive)

var1: name of the X variable

var2: name of the Y variable

rho: null value for the hypothesis test of the SCC; any value between -1 and 1, exclusive.

c_sq: desired value of the c^2 parameter in the Fisher z-transformation of the SCC (Bonett and Wright 2000). If the spearman macro is called with c_sq set equal to missing (.), the value of c^2 recommended by Bonett and Wright is automatically calculated by the macro. Otherwise, the user may specify another desired value. The traditional value is 1.06.

The kendall macro is called using the **data**, **coeff**, **var1**, and **var2** arguments as described above, plus the following additional argument:

kappa: null value of the hypothesis test for the KCC; any value between -1 and 1, exclusive.

EXAMPLE

In an unpublished Masters thesis, Stuart (2013) evaluated several potential biomarkers for measuring severity of symptoms of dry mouth (xerostomia). As part of her assessment of the validity of the biomarkers, she examined the association of one of the potential biomarkers, p21, with a measure of severity of xerostomia symptoms. Severity was assessed using 5 items measured on a visual analog scale (VAS). A sample item is "How difficult is it for you to speak without drinking liquids?"; the possible response ranged from 1 (not difficult) to 100 (very difficult), depending on where the respondent placed a mark on a 100 mm line. The scores of the 5 items on the VAS were totaled to yield an overall score, which had a theoretical range from 5 to 500. (Patients who were not currently experiencing any xerostomia symptoms were assigned a score of zero on the VAS.) Table 1 contains the values for the total VAS score and the p21 values for 20 patients currently suffering from xerostomia.

Patient	VAS	p21
1	47	0.4831
2	59	0.6234
3	24	0.5885
4	83	0.5886
5	315	0.5193
6	85	0.7022
7	93	0.5665
8	9	0.3120
9	211	0.6340
10	81	0.5609
11	226	0.7317
12	91	0.5763
13	126	0.5134
14	31	0.4056
15	17	0.3472
16	0	0.4860
17	113	0.5226
18	257	0.5051
19	69	0.3523
20	68	0.5140

Data courtesy of Dr. Mary Stuart.

Table 1. Unpublished Data on p21 Biomarker and Severity of Symptoms for 20 Xerostomia Patients

SAS CODE FOR EXAMPLE

The following SAS code was used to create the SAS data set containing the input data provided in Table 1, as well as perform statistical inference for the Pearson, Spearman and Kendall coefficients. Note that the calculations for the Spearman coefficient were performed using both the improved Bonett and Wright approach (specified by the argument `c_sq=.`), as well as for the traditional approach (specified by the argument `c_sq=1.06`).

```
data biomarkers;
input Patient VAS p21;
datalines;
1      47      0.4831
2      59      0.6234
3      24      0.5885
4      83      0.5886
5      315     0.5193
6      85      0.7022
7      93      0.5665
8      9       0.3120
9      211     0.6340
10     81      0.5609
11     226     0.7317
12     91      0.5763
13     126     0.5134
14     31      0.4056
15     17      0.3472
16     0       0.4860
17     113     0.5226
18     257     0.5051
19     69      0.3523
20     68      0.5140
;
run;

%spearman(data=biomarkers, coeff=95, var1=VAS, var2=p21, rho=0, c_sq=.);
%spearman(data=biomarkers, coeff=95, var1=VAS, var2=p21, rho=0, c_sq=1.06);
%kendall(data=biomarkers, coeff=95, var1=VAS, var2=p21, kappa=0);
```

RESULTS

A scatterplot of p21 vs. VAS is given in Figure 1. Table 2 contains a summary of the analyses described above for the Pearson, Spearman, and Kendall coefficients, considered as measures of the association of p21 with VAS. Contrary to popular belief, the sample value of the SCC for the correlation of p21 with VAS in Table 2 is actually larger than the sample value of the PCC ($r_s = 0.477$ vs. $r = 0.402$). Furthermore, the SCC is significantly different from zero ($p = 0.0427$), whereas the PCC is not ($p = 0.0790$). (Note that if the traditional value of 1.06 is used in the calculations for the SCC, the results are even stronger ($p = 0.0325$). Even though the KCC is smaller than the PCC ($t_b = 0.316$ vs. $r = 0.402$), it does achieve statistical significance at the 0.05 level ($p = 0.0479$), unlike the PCC. As far as determining the correct measure of association to use with these data, given that there is strong evidence of a violation of normality for the VAS data (Shapiro-Wilk test, $p = 0.010$), either the SCC or the KCC would be preferable to the PCC for measuring the association between p21 and VAS. The results in Table 2 provide only modest support for the validity of p21 as a biomarker for severity of xerostomia symptoms.

Measure of Association	Estimated Coefficient	95% C.I. for ξ	p -value for Test of $H_0: \xi = 0$
Pearson	0.402	(-0.049, 0.717)	0.0790
Spearman	0.477	(0.017, 0.770)	0.0427
Kendall	0.316	(0.003, 0.572)	0.0479

ξ : Population value of measure of association ($\xi = \rho$ for Pearson's, ρ_s for Spearman's, and t_b for Kendall's)

Table 2. Results for Fisher z-Transformed Correlations of p21 with Severity of Symptoms in 20 Xerostomia Patients

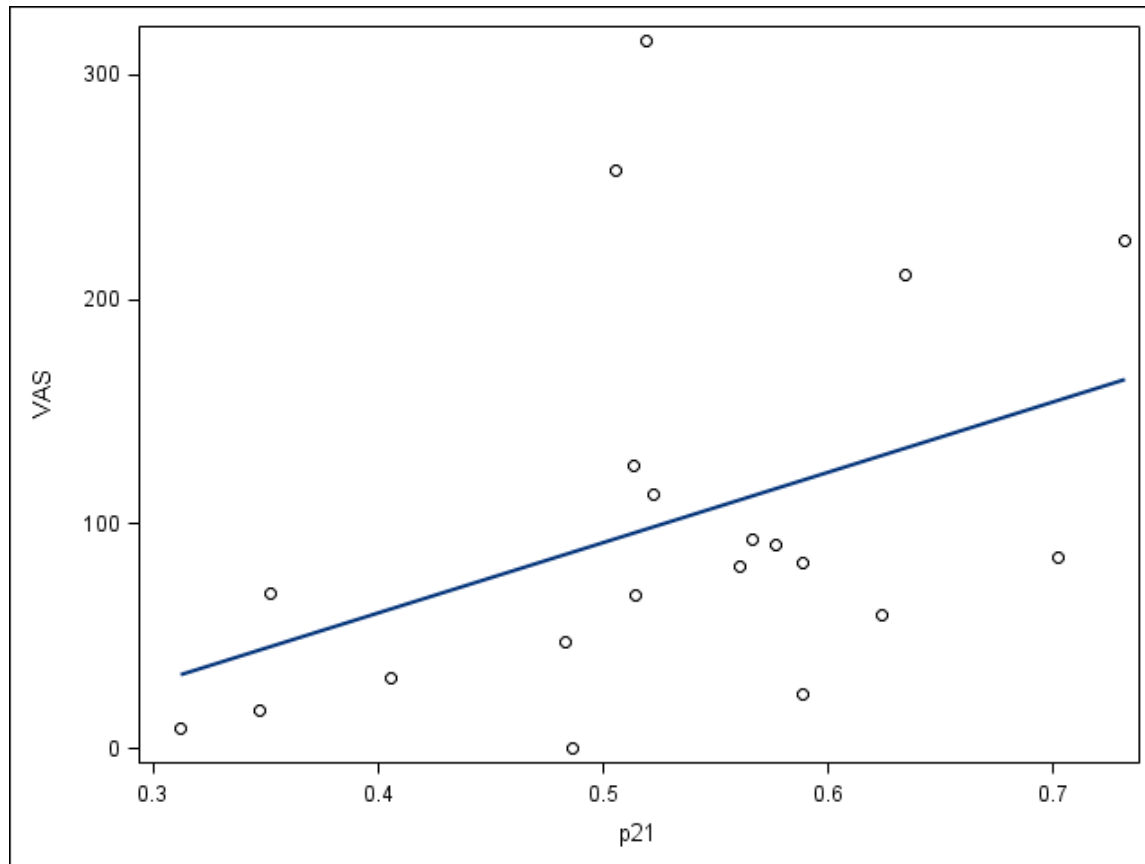


Figure 1. Scatterplot of VAS vs. p21 for 20 Xerostomia Patients

REFERENCES

- Bonett D.G., and Wright T.A. (2000). "Sample Size Requirements for Estimating Pearson, Kendall, and Spearman Correlations." *Psychometrika*, 65, 23-28.
- Fieller, E.C., Hartley, H.O., and Pearson, E.S. (1957). "Tests for Rank Correlation Coefficients." *Biometrika*, 44, 470-481.
- Stuart, M. (2013). "Identification of Novel Molecular Biomarkers for Diagnosis of Salivary Dysfunction." Unpublished Masters thesis, Georgia Regents University, Dept. of Oral Biology.

ACKNOWLEDGMENT

We gratefully acknowledge Jennifer Waller, who graciously provided basic SAS code for calculating improved confidence intervals for correlation coefficients, which we expanded into the SAS macros we present here.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Stephen W. Looney, Ph.D.
 Dept. of Biostatistics and Epidemiology
 Georgia Regents University
 1120 15th Street, AE-1014
 Augusta, GA 30912-4900
 (706) 721-4846
slooney@gru.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.