Paper 170-31

Computation of Correlation Coefficient and Its Confidence Interval in SAS®

David Shen, WCI, Inc.
Zaizai Lu, AstraZeneca Pharmaceuticals

ABSTRACT Correlation measures the association between variables. However, the correlation coefficient is not normally distributed and its variance is not constant. This paper presents the principle of Fisher transformation to normalize the distribution and stabilize the variance, and its application in computation of confidence interval. The calculation is implemented with SAS codes in both 8.2 and 9.1.3 versions.

Keywords: correlation coefficient, Fisher Transformation, Confidence Interval

1. CORRELATION

Correlation is a measure of the strength of relationship between random variables. The population correlation between two variables X and Y is defined as:

ρ (X, Y) = Covariance (X, Y) / {Variance (X) * Variance (Y)} ^{½}

ρ is called the Product Moment Correlation Coefficient or simply the Correlation Coefficient. It is a number that summarizes the direction and closeness of linear relations between two variables. The sample value is called *r*, and the population value is called ρ (rho). The correlation coefficient can take values between -1 through 0 to +1. The sign (+ or -) of the correlation defines the direction of the relationship. When the correlation is positive ($r > 0$), it means that as the value of one variable increases, so does the other. For example, as the dose amount of an oncology medicine increases, so does the survival time, in a certain range. If a correlation is negative ($r < 0$),, it indicates that when one variable **increases**, the other variable **decreases**. This means there is an inverse relationship between the two variables. For example, as the dose amount of an anti-hypertensive medicine increases, the diastolic blood pressure decreases.

The formula for the population Pearson product-moment correlation, denoted by ρ_{xy}, is

$$\rho_{xy} = \frac{\operatorname{Cov}(x,y)}{\sqrt{V(x)V(y)}} = \frac{\mathrm{E}(\,(x-\mathrm{E}(x))(y-\mathrm{E}(y))\,)}{\sqrt{\mathrm{E}(x-\mathrm{E}(x))^2\,\mathrm{E}(y-\mathrm{E}(y))^2}}$$

The formula for the sample Pearson product-moment correlation is

$$r_{xy} = \frac{\sum_i(\,(x_i-\bar{x})(y_i-\bar{y})\,)}{\sqrt{\sum_i(x_i-\bar{x})^2\,\sum_i(y_i-\bar{y})^2}}$$

The correlation coefficient can be computed with PROC CORR procedure in SAS.
Probability values for the Pearson correlation are computed by treating

$$t = (n-2)^{1/2} \left([(r^2)/(1-r^2)]\right)^{1/2}$$

as coming from a t distribution with (n-2) degrees of freedom, where r is the sample correlation.

2. DISTRIBUTION OF r

The population correlation (ρ) is usually not known. Therefore, the sample statistic r is used to estimate ρ and to carry out tests of hypotheses. The tests of hypotheses rest on statements of probability. For example, we might say that the observed r of 0.75 is very unlikely if ρ =0. A sampling distribution is when you take repeated samples from a population and compute a statistic each time you take a sample. The long run average of the sampling distribution of means is the population means, that is, the expected value of the average of the distribution of sample means is the parameter. In other words, the sample mean is an unbiased estimator of the population mean. Unbiased estimators have the property that the expected value (mean) of the sampling distribution is the parameter. If the expected value does not equal the parameter, the estimator is biased. As the sample size increases, the sampling distribution of means becomes normal in shape according to the Central Limit Theorem. This is really important since we can use the normal distribution to calculate probabilities and confidence interval of correlation coefficient. However, most of the time the mean of the sampling distribution of r does not equal, nor is the sampling distribution of r ever normal. Let's look at the sampling distribution of r. Recall that r is bounded by +1 and -1, that is, it can take no larger or smaller values. When ρ=0, r is distributed around 0 symmetrically, and the mean of the sampling distribution does equal the parameter. As ρ increases from zero (becomes more positive), the sampling distribution becomes negatively skewed. As ρ becomes negative, the sampling distribution becomes positively skewed. As ρ approaches +1 or -1, the sampling variance decreases, so that when ρ is either at +1 or -1, all sample values equal the parameter and the sampling variance is zero. Note that as |ρ| approaches 1, the sampling variance approaches zero. The shape of the sampling distribution depends on N. The shape becomes increasingly normal with large values of N, and becomes increasingly skewed with increasing |ρ|. So you can see, the sampling variance and the significance test depend upon (1) the size of the population correlation and (2) the sample size.

3. FISHER TRANSFORMATION

Fisher developed a transformation of r that tends to become normal quickly as N increases. It is called the r to z transformation. We use it to conduct tests of the correlation coefficient and calculate the confidence interval.

$$z = .5 \log_e \left(\frac{1+r}{1-r} \right)$$

For the transformed z, the approximate variance V(z) = 1/(n-3) is independent of the correlation. Furthermore, even the distribution of z is not strictly normal; it tends to be normal rapidly as the sample size increases for any values of ρ.

4. CONFIDENCDE INTERVALS

A confidence interval gives an estimated range of r values which is likely to include an unknown population ρ, the estimated range being calculated from a given set of sample data. Confidence intervals are calculated at a confidence level (a certain percentage), usually 95% (α =0.05), but we can also produce 90%, 99%, 99.9% and other confidence intervals for the unknown parameter ρ. Confidence intervals are more informative than the simple results of hypothesis tests (where we decide 'reject H0' or 'don't reject H0') since they provide a range of plausible values for the unknown parameter. If the confidence interval includes 0 we can say that the population ρ is not significantly different from zero, at a given level of confidence. The confidence intervals of correlation coefficient are computed based on the sample mean r and sample standard deviation. As mentioned above, the variance is dependent on both sample size and ρ size, so the CI can not be computed directly. Fisher transformation makes it possible to calculate the CI indirectly.

Step 1: Fisher transformation.
Step 2: The two-sided confidence limits for ζ are computed as

$$\zeta_l = z_r - z_{(1-\alpha/2)}\sqrt{\frac{1}{n-3}}$$

$$\zeta_u = z_r + z_{(1-\alpha/2)}\sqrt{\frac{1}{n-3}}$$

where $z_{(1-\alpha/2)}$ is the $100(1-\alpha/2)$ percentage point of the standard normal distribution.

Step 3: These computed confidence limits of ζ_l and ζ_u are then transformed back to derive the confidence limits for the correlation ρ :

$$r_l = \tanh(\zeta_l) = \frac{\exp(2\zeta_l)-1}{\exp(2\zeta_l)+1}$$

$$r_u = \tanh(\zeta_u) = \frac{\exp(2\zeta_u)-1}{\exp(2\zeta_u)+1}$$

5. SAS CODES

SAS codes (v8.2) for correlation coefficient, hypothesis test and confidence interval are provided in the following section.

```
proc corr data=a outp=corr; *outp with pearson correlation coefficient;
  var x y;
run;

data corr_ci;
```

```
   set corr (rename=(x=corr) drop=y _name_);
   retain n;
   if _type_='N' then n=corr;
   if _type_='CORR'and corr ^= 1;

   fishersz=0.5*(log(1+corr)-log(1-corr)); *Fisher Z transformation;
   sigmaz=1/sqrt(n-3);                     *variance;
   l95=fishersz-1.96*sigmaz;               *α=0.05, i.e. at 95% level;
   u95=fishersz+1.96*sigmaz;

   l95=(exp(2*l95)-1)/(exp(2*l95)+1);      *inverse of Fisher Z ;
   u95=(exp(2*u95)-1)/(exp(2*u95)+1);      *transformation to get CI;
run;

proc print data=corr_ci;
run;
```

The SAS output would look like

Obs	_TYPE_	corr	n	fishersz	sigmaz	l95	u95
1	CORR	0.87982	30	1.37496	0.19245	0.76065	0.94162

The CORR procedure in SAS v9.1.3 provides the following new features:

(1) The FISHER option in the PROC CORR statement offers confidence limits and p-values for Pearson correlation coefficients based on Fisher's z transformation. Using the FISHER option, you can specify an alpha value and a null hypothesis value. You can also specify the type of confidence limit (upper, lower, or two-sided) and whether the bias adjustment should be used for the confidence limits.

```
title 'Calculation and Test of Correlations, 95% CI';
ods output FisherPearsonCorr=corr;

proc corr data=a  fisher ( biasadj=no );
  var  x y;
run;
```

SAS output result:

Calculation and Test of Correlations, 95% CI

Obs	Var	With Var	NObs	Corr	ZVal	Lcl	Ucl	pValue
1	x	y	30	0.87982	1.37496	0.760653	0.941620	<.0001

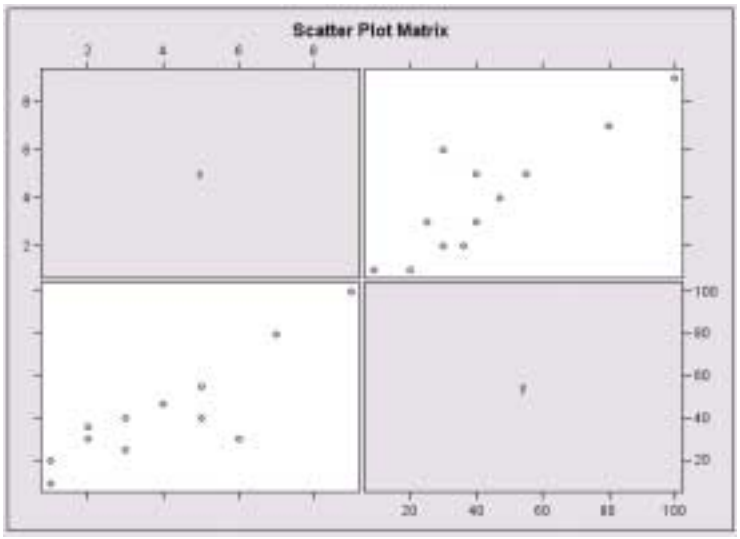
(2) Create scatter plots: The PLOTS=MATRIX option in the PROC CORR statement uses ODS graphics to produce the symmetric matrix plot.

```
ods html;
ods graphics on;

proc corr data=a noprint plots=matrix;
 var x y;
run;

ods graphics off;
```

```
ods html close;
```

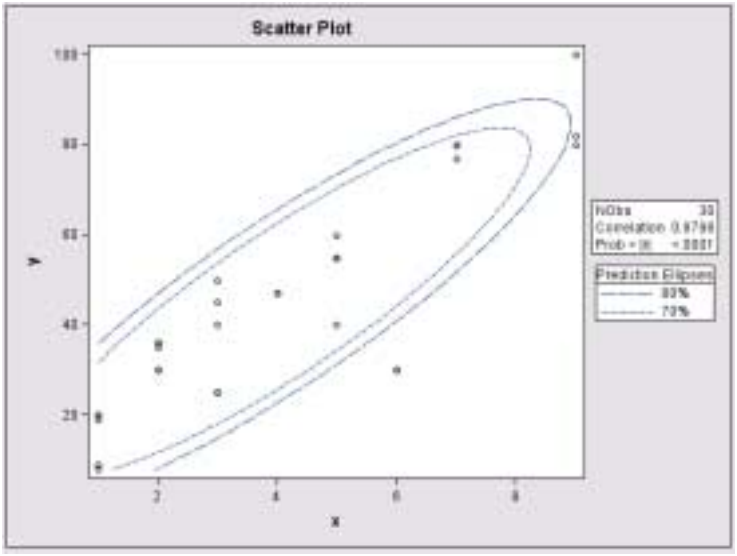


The PLOTS=SCATTER option in the PROC CORR statement uses ODS graphics to produce scatter plots for variables. By default, the scatter plot also includes a 95% prediction ellipse. You can use the ELLIPSE= option with the PLOTS=SCATTER option to include prediction ellipses for new observations, confidence ellipses for the mean, or no ellipses.

```
ods html;
ods graphics on;

proc corr data=a noprint plots=scatter(alpha=.2 0.3 );
 var x y;
run;

ods graphics off;
ods html close;
```



6. CONCLUSION

The paper discusses Correlation Coefficient and its Confidence Intervals, and how to normalize its distribution through Fisher Transformation. SAS codes are provided here to get the results in a quick and easy way.

CONTACT INFORMATION

Zaizai Lu
zz_lu@hotmail.com
AstraZeneca Pharmaceuticals
Wilmington, Delaware