Chemometrics in Spectroscopy

Comparison of Goodness of Fit Statistics for Linear Regression, Part III

The authors continue their discussion of computing confidence limits for the correlation coefficient in developing a calibration for quantitative analysis.

Jerome Workman Jr. and Howard Mark

Continuing our discussion in our two previous columns (1, 2), this month we calculate the confidence limits for the correlation coefficient for a user-selected confidence level. The user selects the test correlation coefficient, the number of samples in the calibration set, and the confidence level. A MathCad worksheet (MathSoft Engineering & Education, Inc., Cambridge, MA), is used to calculate the z-statistic for the lower and upper limits and computes the appropriate correlation for the z-statistic. The upper and lower confidence limits are displayed. The worksheet also contains the tabular calculations for any set of correlation coefficients (given as ρ). A graphic showing the general case entered for the table also is displayed.

For n pairs of values (X, Y) the set of pairs can be interpreted as a subset of the entire population of X and Y values throughout some larger population of samples. For example, X and Y might constitute all possible combinations of an instrument response (Y) and an analyte concentration (X) in a specific solvent matrix. The population correlation coefficient can be referred to as the Greek letter rho (ρ), which is estimated using the correlation coefficient computed for a specific subset of values, designated as (r). It is known that tests of significance can be performed on a measured r to determine if it is significantly different from another r calculated from a different subset of X, Y values. The significance between any specific r calculated from a subset of X, Y values also can be compared to the estimated population correlation for all such possible samples, r. When a hypothesis test is used to calculate whether r is statistically equal to zero, the distribution is approximated using the Student's t distribution. When r is tested to be not equal to zero the use of the Fisher transformation produces a statistic that is distributed normally. This transformation is referred to as Fisher's z transformation (that is, the z-statistic).

The z-statistic for testing a non-zero population correlation is given by equation 1 as $Z_1$ where $e = 2.71828$. A good discussion of this is found in reference 3.

$$Z_1 = 0.5 \cdot \log_e \left( \frac{1+r}{1-r} \right) \quad [1]$$

A more standard form (equation 2) used for computational purposes is:

$$Z_1 = 1.1513 \cdot \log_{10} \left( \frac{1+r}{1-r} \right) \quad [2]$$

The confidence limits for a correlation coefficient for a given number of X, Y pairs (n) at a specified confidence limit is calculated as $Z_2$ (Equation 3).

$$Z_2 = 1.1513 \cdot \log_{10} \left( \frac{1+r}{1+r} \right) \pm z \cdot \left( \frac{1}{\sqrt{n-3}} \right) \quad [3]$$
Chemometrics in Spectroscopy

Note that the $z$-statistic is computed as $z_1$ or is available from standard statistical tables as the Student’s $t$ distribution such as having a confidence limit is shown in Figure 1a (top) as having a $p$ value of 0.91551; this represents the upper confidence limit for the 0.80 correlation example problem. Finally, for the example problem, the correlation confidence limits are from 0.562575 to 0.91551 (that is, from 0.56 to 0.92).

**Testing Correlation for Different-sized Populations**

The following description and corresponding MathCad worksheet allow the user to test whether two correlation coefficients are significantly different. The test statistic for this test is $z = \frac{\hat{r}_2 - \hat{r}_1}{\sqrt{\frac{1}{n_1} - \frac{1}{n_2}}}$ given as equation 8.

A test result of 0 indicates a significant difference between the correlation coefficients; a test result of 1 indicates no significant difference in the correlation coefficients; 1b represents the lower confidence limit for the 0.80 correlation example problem.

For a specific example problem, we calculate 0.563 and 0.920 as the lower and upper confidence limits for the correlation coefficient $(\hat{r})$ by solving for $r$ in the following equation or by using the MathCad worksheet, the user enters the confidence level for the test (for example, 0.95), two comparative correlation coefficients, $r_1$ and $r_2$, and the respective number of paired $(X, Y)$ samples as $n_1$ and $n_2$. The desired confidence level is entered and the corresponding $z$-statistic is computed as:

\[
\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3} = 0.63663 \quad \frac{1}{n_2 - 3} = 0.562575
\]

In summary, for any stated value of the population correlation ($\rho$) the $z$-statistic is denoted as $Z(\rho)$, and the corresponding correlation confidence limits can be determined. For example, the $z$-statistic of 0.6366 corresponding to the lower correlation coefficient confidence limit is shown in Figure 1a as having a $p$ value of 0.562575; this represents the lower confidence limit for the correlation coefficient for this example.

 Likewise for this example, the $z$-statistic of 1.5606 corresponding to the upper correlation coefficient confidence limit is shown in Figure 1b as having a $p$ value of 0.91551; this represents the upper confidence limit for the 0.80 correlation example problem. Finally, for the example problem, the correlation confidence limits are from 0.562575 to 0.91551 (that is, from 0.56 to 0.92).

The test statistic for this problem is given as equation 8.

The null hypothesis test for this problem is stated as follows: are two correlation coefficients $r_1$ and $r_2$ statistically the same (that is, $r_1 = r_2$)? The alternative hypothesis then is $r_1 \neq r_2$. If the absolute value of the test statistic $Z(n)$ is greater than the absolute value of the $z$-statistic, then the null hypothesis is rejected and the alternative hypothesis accepted — there is a significant difference between $r_1$ and $r_2$. If the absolute value of $Z(n)$ is less than the $z$-statistic, then the
null hypothesis is accepted and the alternative hypothesis is rejected; thus a significant difference does not exist between $r_1$ and $r_2$. Let’s look at a standard example again (equation 9).

And $Z(n) = 0.89833$, therefore $Z(n)$, the test statistic, is less than 1.96, the $z$-statistic, and the null hypothesis is accepted — there is not a significant difference between the correlation coefficients.

In a second example, which might be more typical, let us see what happens when $r_1$ is 0.87 and $r_2$ is 0.96, with $n_1$ as 20, and $n_2$ as 25. At a confidence level test of 0.95 we use the above equations for $Z(n)$ and find that there is not a significant difference (for example, $Z(n) = 1.8978$, which is less than 1.96). The use of this statistical test emphasizes the point that comparison of correlation coefficients for small numbers of sample pairs definitely is risky when confidence limits and statistical hypothesis testing are not used. In our experience we have seen analytical techniques and methods accepted or rejected by large research organizations using the “correlation eye-balling” test, where the method is accepted or rejected solely upon a relative comparison of correlation coefficients, without the benefit of computing the confidence limits. This is a somewhat common, but easily preventable, mistake.

References


Next month in Chemometrics in Spectroscopy ...

The authors continue the Comparison of Goodness of Fit Statistics for Linear Regression discussion. Part IV, covering Confidence Limits for Slope and Intercept, shows two sets of equations for calculations of slope and intercept — one as a summation notation set useful for application in MathCad software, and a second set as shown from a reference.