**STAT452: Lab 4**

**Correlation**

Terence Henriod

STAT 452

3/7/2013

**Introduction**

The correlation between two variables is a useful measure of how well observations of one variable might predict another, or their association. However, this is not to say that the two variables are necessarily related; correlation does not imply causation. Correlation is a useful tool to help guide research and point us in the direction of the true relationships. It is also easier to find the correlation between two variables than it is to find their joint distribution.

We will be working with two such measures of correlation: the Pearson correlation coefficient, r, and the Spearman rank order correlation coefficient, ρ.

Pearson’s r is a measure of linear association that applies a transformation to observations that produces their deviation from the mean of each variable, and then measures the deviations of the observations from those means:

[1]

It should be noted that Pearson’s r is meant to measure linear associations only, and may perform poorly when evaluating data that may be correlated in other patterns, like data with a quadratic association, for instance.

While Pearson’s r may not be as robust as other correlation coefficients, it is useful for hypothesis testing purposes. Pearson’s r operates under the assumption that the variables are Normally distributed, so if it is reasonable to assume this is true, we can apply hypothesis testing to r. The hypothesis set is the following:

[2]

.

and the test statistic is computed by the following:

[3]

which, for independent jointly Normal(X, Y) has the Student distribution with (n – 2) degrees of freedom.

Spearman’s ρ is a rank order correlation coefficient, meaning the data is viable if the data exhibit a monotonic increasing association, eliminating the restriction of its use to linear associations. ρ utilizes Pearson’s r, ρ is defined as Pearson’s r for the rank values assigned to the observation values of the sample:

[4]

This makes Spearman’s ρ more robust, although it can’t really be used for hypothesis testing.

The following operations will be performed using four instructor provided samples for the purpose of practicing computation of correlation coefficients and interpreting the resulting values and significance of those coefficients using Minitab.

**Assignments**

1. Decide which samples are from the Normal distribution (you can use the probability plot option with the Anderson-Darling test).



Figure 1: A probability plot of the data from Sample W against a 95% confidence interval for the Normal distribution at a similar scale. The data fits well, and the Anderson-Darling test also suggests that the sample is likely Normal. W is Normally distributed.



Figure 2: A probability plot of the data from Sample X against a 95% confidence interval for the Normal distribution at a similar scale. X has very few data falling outside the confidence interval, and the Anderson-Darling tests presents a P-Value that allows us to reject the hypothesis that X is not Normal. X is Normally distributed.



Figure 3: A probability plot of the data from Sample Y against a 95% confidence interval for the Normal distribution at a similar scale. Visually, we can see that Y does not fit a Normal distribution well, and the P-Value from the Anderson-Darling test is highly significant, suggesting we reject the hypothesis that Y is Normal. Y is not Normally distributed.



Figure 4: A probability plot of the data from Sample Z against a 95% confidence interval for the Normal distribution at a similar scale. Again, Z also fits the graph very poorly, and again, the Anderson-Darling result is significant and suggests we reject that Z is Normal. Z is not Normally Distributed.

1. Perform a visual correlation analysis of the data, decide which pairs of variables (out of 6 possible pairs) have association, and discuss whether the association is monotone or linear.



Figure 5: A matrix plot of all the samples to be considered. This plot gives us a visual display of how the data are associated, with the samples being paired one way above the diagonal, and then the pairs are inverted below the diagonal.

* 1. W, X

There appears to be a linear association between W and X. The correlation does not appear to be tight, but the trend does appear to be monotone, and more precisely, linear.

* 1. W, Y

The association of W and Y appears to be relatively tight. While it is not a linear association, the association certainly does appear to be monotonic and increasing.

* 1. W, Z

There is clearly no association between W and Z, the plot has no discernible trend.

* 1. X, Y

The data looks like it may be loosely associated, and there does appear to be a monotonic increasing non-linear trend. By visual assessment, X and Y do appear to be associated in a monotonic increasing manner, but certainly not a linear one.

* 1. X, Z

Again, as with W and Z, X and Z have no discernible association, the plot displays no discernible trend.

* 1. Y, Z

While there does seem to be some grouping of the data points, the groupings do not seem to indicate any trend, and the groups of outliers seem too large to actually be considered outliers. Y and Z are not associated.

1. Compute Pearson correlation for each pair.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sample | W | X | Y | Z |
| W | 1 | 0.884 | 0.872 | -0.150 |
| X | 0.000 | 1 | 0.791 | -0.051 |
| Y | 0.000 | 0.000 | 1 | 0.009 |
| Z | 0.298 | 0.726 | 0.950 | 1 |

Table 1: A display of computed values for the Pearson r coefficient of correlation for the samples. The numbers above the main diagonal represent the values computed for r, the values below the diagonal are the P-Values that might be used for hypothesis testing concerning the values computed for r. The main diagonal has been lightly highlighted in gray, P-Values that may not be used legitimately have been darkened.

1. Compute significance (P-value) for the Pearson correlation; discuss which P-values can be used for analysis, and which should be dismissed (use the results of [1]).

See Table 1 for P-Values. Because the computation of a P-Value for a Pearson’s r statistic is based on the assumption that both variables being compared are Normally distributed, we are left with only a single pair for which utilizing the P-Value would be a legitimate option: W and X. The rest of the P-Values should all be ignored, they were computed for pairs where the samples were not Normally distributed, and are therefore invalid.

1. Compute Spearman correlation for each pair.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sample | W | X | Y | Z |
| W | 1 | 0.873 | 0.954 | -0.156 |
| X | 0.000 | 1 | 0.853 | -0.097 |
| Y | 0.000 | 0.000 | 1 | -0.176 |
| Z | 0.280 | 0.505 | 0.221 | 1 |

Table 2: The format of this table is similar to that of Table 1, except the values above the diagonal are the computed values for Spearman’s ρ. Because this table represents the results for Spearman’s ρ, it would be illegitimate to use any of the P-Values, so all of them have been darkened to signify that none of these P-Values are applicable.

1. Compare the results of 3)-6), discuss, and make conclusions.
   1. W, X

We can say that samples W and X are linearly associated. First, they have a Spearman rank order coefficient ρ = 0.873, meaning the data are increasing in a corresponding manner. The samples have a Pearson correlation coefficient r = 0.873, verifying that the data correspond in a *linear* increasing manner. Further, because W and X were identified as both being Normally distributed, we can apply hypothesis testing to their r coefficient. With a P-Value of 0.000, it is highly significant, allowing us to confidently reject the null hypothesis that the two samples are uncorrelated. W and X are linearly and increasingly correlated.

* 1. W, Y

Samples W and Y are certainly correlated, with r = 0.872 and ρ = 0.954, we can certainly say that the two are associated in an increasing trend. However, are they associated linearly, or just monotonically? Based on visual assessment and the fact that ρ > r, I would assert that the data is not linearly associated, and Pearson’s r just happened to produce a favorable result, really the data is simply associated in a monotonic increasing manner.

* 1. W, Z

W and Z are likely not associated. The computed coefficients, r = -0.150 and ρ = -0.156, both suggest that the data is weakly correlated in a negative trend, but the relative magnitude of these coefficients is small, and if we could subject them to hypothesis testing, they would likely prove insignificant. Based on the small magnitude of the coefficients and visual assessment, W and Z are not associated.

* 1. X, Y

The results for X and Y were r = 0.791 and ρ = 853. These results agree with visual assessment, as the matrix plot displays an increasing trend as well. Again, like with W and Y, the results suggest that X and Y are associated in a monotonic increasing trend, but not a linear one, for all the same reasons: visually the trend does not appear linear and ρ > r.

* 1. X, Z

X and Z are not likely associated either, with r = -0.051 and ρ = -0.097. Again, the absolute value of these coefficients is similar, but small, and therefore not likely significant, and again, visually there is no discernible trend in the data.

* 1. Y, Z

Like all of the other pairs including Z, there is no significant evidence of association. With r = 0.009 and ρ = -0.176, again, the magnitudes are small and likely insignificant. Further, the fact that the two coefficients disagree on whether or not the data is correlated positively or negatively is further evidence that the values for the coefficients are insignificant. Pearson’s r and Spearman’s ρ don’t measure the same thing and may not always agree as to how well the data is correlated, but they should always agree on the *direction* of the correlation if it the trend is monotone and significant. By visual assessment, the data do not appear to have some non-monotonic trend, so it is not the case that the correlation computations were simply inappropriately applied to the data. The coefficients and visual assessment are in agreement: Y and Z are not associated.

**Conclusion**

As stated in the introduction, various association determination methods were applied, including the computation of Pearson’s r and Spearman’s ρ correlation coefficients. Hypothesis testing was applied, when appropriate, to these coefficients. Computational results were in agreement with visual assessments of the data.

The resulting associations were as follows: W and X exhibit a positive linear association; W and Y, and X and Y have a positive correlation, although it is unlikely that these associations are linear; and finally, any pair involving Z had no indications of association existing.