CITED REFERENCES AND FURTHER READING:

14.5 Linear Correlation

We next turn to measures of association between variables that are ordinal or continuous, rather than nominal. Most widely used is the linear correlation coefficient. For pairs of quantities \((x_i, y_i), i = 1, \ldots, N\), the linear correlation coefficient \(r\) (also called the product-moment correlation coefficient, or Pearson’s \(r\)) is given by the formula

\[
 r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}
\]  

(14.5.1)

where, as usual, \(\bar{x}\) is the mean of the \(x_i\)’s, and \(\bar{y}\) is the mean of the \(y_i\)’s.

The value of \(r\) lies between −1 and 1, inclusive. It takes on a value of 1, termed “complete positive correlation,” when the data points lie on a perfect straight line with positive slope, with \(x\) and \(y\) increasing together. The value 1 holds independent of the magnitude of the slope. If the data points lie on a perfect straight line with negative slope, \(y\) decreasing as \(x\) increases, then \(r\) has the value −1; this is called “complete negative correlation.” A value of \(r\) near zero indicates that the variables \(x\) and \(y\) are uncorrelated.

When a correlation is known to be significant, \(r\) is one conventional way of summarizing its strength. In fact, the value of \(r\) can be translated into a statement about what residuals (root mean square deviations) are to be expected if the data are fitted to a straight line by the least-squares method (see §15.2, especially equations 15.2.13 – 15.2.14). Unfortunately, \(r\) is a rather poor statistic for deciding whether an observed correlation is statistically significant, and/or whether one observed correlation is significantly stronger than another. The reason is that \(r\) is ignorant of the individual distributions of \(x\) and \(y\), so there is no universal way to compute its distribution in the case of the null hypothesis.

About the only general statement that can be made is this: If the null hypothesis is that \(x\) and \(y\) are uncorrelated, and if the distributions for \(x\) and \(y\) each have enough convergent moments (“tails” die off sufficiently rapidly), and if \(N\) is large
(typically > 500), then $r$ is distributed approximately normally, with a mean of zero and a standard deviation of $1/\sqrt{N}$. In that case, the (double-sided) significance of the correlation, that is, the probability that $|r|$ should be larger than its observed value in the null hypothesis, is

$$\text{erfc}\left(\frac{|r| \sqrt{N}}{\sqrt{2}}\right)$$

(14.5.2)

where $\text{erfc}(x)$ is the complementary error function, equation (6.2.8), computed by the routines $\text{erfc}$ or $\text{erfcc}$ of §6.2. A small value of (14.5.2) indicates that the two distributions are significantly correlated. (See expression 14.5.9 below for a more accurate test.)

Most statistics books try to go beyond (14.5.2) and give additional statistical tests that can be made using $r$. In almost all cases, however, these tests are valid only for a very special class of hypotheses, namely that the distributions of $x$ and $y$ jointly form a binormal or two-dimensional Gaussian distribution around their mean values, with joint probability density

$$p(x, y) \, dx \, dy = \text{const} \times \exp\left[-\frac{1}{2}(a_{11}x^2 - 2a_{12}xy + a_{22}y^2)\right] \, dx \, dy$$

(14.5.3)

where $a_{11}$, $a_{12}$, and $a_{22}$ are arbitrary constants. For this distribution $r$ has the value

$$r = -\frac{a_{12}}{\sqrt{a_{11}a_{22}}}$$

(14.5.4)

There are occasions when (14.5.3) may be known to be a good model of the data. There may be other occasions when we are willing to take (14.5.3) as at least a rough and ready guess, since many two-dimensional distributions do resemble a binormal distribution, at least not too far out on their tails. In either situation, we can use (14.5.3) to go beyond (14.5.2) in any of several directions:

First, we can allow for the possibility that the number $N$ of data points is not large. Here, it turns out that the statistic

$$t = r \sqrt{\frac{N - 2}{1 - r^2}}$$

(14.5.5)

is distributed in the null case (of no correlation) like Student’s $t$-distribution with $\nu = N - 2$ degrees of freedom, whose two-sided significance level is given by $1 - A(t|\nu)$ (equation 6.4.7). As $N$ becomes large, this significance and (14.5.2) become asymptotically the same, so that one never does worse by using (14.5.5), even if the binormal assumption is not well substantiated.

Second, when $N$ is only moderately large ($\geq 10$), we can compare whether the difference of two significantly nonzero $r$’s, e.g., from different experiments, is itself significant. In other words, we can quantify whether a change in some control variable significantly alters an existing correlation between two other variables. This is done by using Fisher’s $z$-transformation to associate each measured $r$ with a corresponding $z$,

$$z = \frac{1}{2} \ln \left(\frac{1 + r}{1 - r}\right)$$

(14.5.6)
Then, each $z$ is approximately normally distributed with a mean value

$$
\overline{z} = \frac{1}{2} \left[ \ln \left( \frac{1 + r_{\text{true}}}{1 - r_{\text{true}}} \right) + \frac{r_{\text{true}}}{N - 1} \right]
$$

(14.5.7)

where $r_{\text{true}}$ is the actual or population value of the correlation coefficient, and with a standard deviation

$$
\sigma(z) \approx \frac{1}{\sqrt{N - 3}}
$$

(14.5.8)

Equations (14.5.7) and (14.5.8), when they are valid, give several useful statistical tests. For example, the significance level at which a measured value of $r$ differs from some hypothesized value $r_{\text{true}}$ is given by

$$
erfc \left( \frac{|z - \overline{z}|}{\sqrt{N - 3}} \right)
$$

(14.5.9)

where $z$ and $\overline{z}$ are given by (14.5.6) and (14.5.7), with small values of (14.5.9) indicating a significant difference. (Setting $\overline{z} = 0$ makes expression 14.5.9 a more accurate replacement for expression 14.5.2 above.) Similarly, the significance of a difference between two measured correlation coefficients $r_1$ and $r_2$ is

$$
erfc \left( \frac{|z_1 - z_2|}{\sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}} \right)
$$

(14.5.10)

where $z_1$ and $z_2$ are obtained from $r_1$ and $r_2$ using (14.5.6), and where $N_1$ and $N_2$ are, respectively, the number of data points in the measurement of $r_1$ and $r_2$.

All of the significances above are two-sided. If you wish to disprove the null hypothesis in favor of a one-sided hypothesis, such as that $r_1 > r_2$ (where the sense of the inequality was decided a priori), then (i) if your measured $r_1$ and $r_2$ have the wrong sense, you have failed to demonstrate your one-sided hypothesis, but (ii) if they have the right ordering, you can multiply the significances given above by 0.5, which makes them more significant.

But keep in mind: These interpretations of the $r$ statistic can be completely meaningless if the joint probability distribution of your variables $x$ and $y$ is too different from a binormal distribution.

SUBROUTINE pearsn(x,y,n,r,prob,z)
INTEGER n
REAL prob,r,z,x(n),y(n),TINY
PARAMETER (TINY=1.e-20)
C USES betai
Will regularize the unusual case of complete correlation.
INTEGER j
REAL ax ay df sxx sxy syy t xt yt betai
ax=0.
ay=0.
do n j=1,n
Find the means.
14.6 Nonparametric or Rank Correlation

It is precisely the uncertainty in interpreting the significance of the linear correlation coefficient \( r \) that leads us to the important concepts of nonparametric or rank correlation. As before, we are given \( N \) pairs of measurements \((x_i, y_i)\). Before, difficulties arose because we did not necessarily know the probability distribution function from which the \( x_i \)'s or \( y_i \)'s were drawn.

The key concept of nonparametric correlation is this: If we replace the value of each \( x_i \) by the value of its rank among all the other \( x_j \)'s in the sample, that is, \( 1, 2, 3, \ldots, N \), then the resulting list of numbers will be drawn from a perfectly known distribution function, namely uniformly from the integers between 1 and \( N \), inclusive. Better than uniformly, in fact, since if the \( x_i \)'s are all distinct, then each integer will occur precisely once. If some of the \( x_i \)'s have identical values, it is conventional to assign to all these "ties" the mean of the ranks that they would have had if their values had been slightly different. This midrank will sometimes be an