9 Correlation

The Pearson product-moment correlation coefficient is probably the single most widely used statistic for summarizing the relationship between two variables. Statistical significance and caveats of interpretation of the correlation coefficient as applied to time series are topics of this lesson. Under certain assumptions, the statistical significance of a correlation coefficient depends on just the sample size, defined as the number of independent observations. If time series are autocorrelated, an "effective" sample size, lower than the actual sample size, should be used when evaluating significance. Transient or spurious relationships can yield significant correlation during some periods and not others. The time variation of strength of linear correlation can be examined with plots of correlation computed for a sliding window. But if many correlation coefficients are evaluated simultaneously, confidence intervals should be adjusted ("Bonferroni adjustment") to compensate for the increased likelihood of observing some high correlations when no relationship exists. Interpretation of sliding correlations can be also be complicated by time variations of mean and variance of the series, as the sliding correlation reflects covariation in terms of standardized departures from means in the time window of interest, which may differ from the long-term means. Finally, it should be emphasized that the Pearson correlation coefficient measures strength of *linear* relationship. Scatterplots are useful for checking whether the relationship is linear.

9.1 Definition of correlation coefficient

The Pearson correlation coefficient goes by various names: "product-moment" correlation coefficient, "simple" correlation coefficient, "ordinary" correlation coefficient, or just "correlation coefficient." The correlation coefficient is the average product of departures of two variables from their respective means divided by the product of the standard deviations of those variables. The equation for the correlation coefficient is

$$r = \frac{\left[1/(N-1)\right]\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\left\{\left[1/(N-1)\right]\sum_{i=1}^{N}(x_i-\bar{x})^2\right\}^{1/2}\left\{\left[1/(N-1)\right]\sum_{i=1}^{N}(y_i-\bar{y})^2\right\}^{1/2}} = \frac{\mathrm{E}\left[(x_t-\bar{x})(y_t-\bar{y})\right]}{\left\{\mathrm{E}\left[(x_t-\bar{x})^2\right]\mathrm{E}\left[(y_t-\bar{y})^2\right]\right\}^{1/2}}$$

$$= \frac{\mathrm{cov}(x,y)}{s_x s_y} = \frac{1}{(N-1)}\sum_{i=1}^{N} z_{i,x} z_{i,y} \qquad (1)$$

where $x_i, i = 1,...,N$ and $y_i, i = 1,...,N$ are time series of length N with sample means $\bar{x}$ and $\bar{y}$, and standard deviations s_x and s_y; "E" denotes expected value; "cov" denotes covariance; and $z_{i,x}$, $z_{i,y}$ are standardized anomalies (Wilks 1995). Standardized anomalies are variables scaled to zero mean and unit standard deviation by subtracting the sample mean and dividing by the sample standard deviation. Standardized anomalies are also called "Z-scores." Equation (1) indicates that the correlation coefficient is approximately equal to the average product of standardized departures of the two variables from their respective means. Linear scaling of the original variables, x_t and y_t, for example by adding a constant or multiplying by a constant, will have no effect on the correlation.

9.2 Statistical significance of correlation coefficient

Significance of an individual sample *r*

The significance of a sample correlation, *r*, depends on the sample size and also on the size of *r*. Significance can be tested with a *t*-test (Snedecor and Cochran 1989; Haan 2002). The assumptions are:

1. The samples, *x* and *y*, are drawn from that a bivariate normal population
2. The samples are random samples from the population
3. The population correlation coefficient is zero: $\rho = 0$

If these assumptions are satisfied, the statistic

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

follows a *t*-distribution with *N*-2 degrees of freedom, where *N* is the sample size. The null and alternative hypotheses for a test (two-sided) are:

H0: $\rho = 0$

H1: $\rho \neq 0$

To apply the test, the steps are:

1) Compute the statistic *T*
2) Decide on an α-level (e.g., $\alpha = 0.05$ for a 95% confidence interval)
3) Compare the computed *T* with the $1-\alpha/2$ probability point of the cdf of Student's *t*-distribution[1]
4) If the absolute value of *T* is greater than the probability point from (3), reject H0

Example. The correlation between a tree-ring index and annual precipitation is *r*=0.48. The sample size is *N*=18. The computed test statistic is *T*=2.1886. A table of the cdf of the *t*-distribution shows that the probability of exceeding this value of *T* by chance is 0.0219. Because this probability is less than 0.025, we reject the null hypothesis of $\rho = 0$ at the $\alpha = 0.05$ level.

Background. The test statistic *T* comes from an analysis of variance for a bivariate regression of *y* on *x*. Details can be found in Panofsky (1958) and Snedecor and Cochrane (1989), and a short summary is presented here. In regression, the total sum-of-squares (SS) of *y* can be split into two parts: SS for the regression estimates and SS for the residuals from the regression line. The regression model is

$$y_t = a + bx_t + e_t, \tag{3}$$

where *a* is the regression constant, *b* is the slope of the regression line, and e_t is the residual. The partitioning of SS is described by

$$\sum_N (y-\bar{y})^2 = \sum_N (\hat{y}-\bar{y})^2 + \sum_N (y-\hat{y})^2 \tag{4}$$

[1] In MATLAB the probability point can be found alternatively by calling function **`tcdf`** or by using **`disttool`**. One can also refer to tables of Student's t-distribution, which are available in many statistics texts.

where $\bar{y}$ is the sample mean of *y*, $\hat{y}$ is the prediction of *y* from the regression, and the summations are over the *N* observations. The terms in (4) are, from left to right: the total SS, the SS due to regression, and the residual SS. Each SS term has an associated "degrees of freedom." The total-SS has $N-1$ degrees of freedom, because one degree of freedom has been used in computing the mean from the sample data. The $N-1$ degrees of freedom are split into two parts: the regression estimate has 1 degree of freedom and the residuals have $N-2$ degrees of freedom. The partitioning of SS and degrees of freedom is summarized in Table 1.

Table 1. Analysis of variance table for linear regression

Source	SS	df	MS	F
Total	NS_y^2	$N-1$	$(N/N-1)S_y^2$	
Regression	$Nb^2S_x^2$	1	$Nb^2S_x^2$	$(N-2)b^2S_x^2/S_e^2$
Residuals	NS_e^2	$N-2$	$NS_e^2/(N-2)$	

The s^2 terms in Table 1 are variances. For example, s_y^2 is the variance of *y* and Ns_y^2 is the corresponding SS. The next-to-last column is the "mean square", which is the quotient of the SS and its degrees of freedom. The last column, labeled "F", is the ratio of mean square of regression and mean square of residuals. This ratio, by statistical theory, is a statistic whose significance can be tested with an F-distribution with degrees of freedom 1 and *N*-2. This is a test of the significance of the slope of regression.

It is shown below that the F-statistic in the last column of Table 2 is identically the square of the T statistic in equation (2). Two substitutions are necessary, relying on the relationship

$$r = b\frac{s_x}{s_y} \tag{5}$$

of the correlation coefficient $(x \text{ vs } y)$ to the linear regression slope *b* and the standard deviations of *x* and *y*; and the relationship

$$s_e^2 = \left(1-r^2\right)s_y^2 \tag{6}$$

of the error variance (from regression of *y* on *x*) to the variance of *y* and the correlation between *x* and *y*.

First re-write the F-statistic term from the Table 1 and make the substitution for s_e^2 from (6)

$$F = \frac{(N-2)b^2s_x^2}{s_e^2} = \frac{(N-2)}{\left(1-r^2\right)}\frac{b^2s_x^2}{s_y^2} \tag{7}$$

Next substitute from (5)

$$F = \frac{(N-2)}{\left(1-r^2\right)}\frac{b^2s_x^2}{s_y^2} = \frac{(N-2)}{\left(1-r^2\right)}r^2 \tag{8}$$

Comparing (8) with (2), we see that

$$F = T^2 \tag{9}$$

If F follows an F-distribution with 1 and (N-2) degrees of freedom,

$$T = \sqrt{F} \tag{10}$$

follows a *t*-distribution with *N*-2 degrees of freedom. A test for the significance of the correlation between *x* and *y* is therefore equivalent to a test for the significance of the slope in a linear regression of *y* on *x*.

Approximation with large sample size

If the sample size is large (say $N>25$), an approximate confidence interval can be derived to assess whether a sample *r* is significantly different from zero. The sample correlation coefficient for *N* observations on two variables *x* and *y* is normally distributed with mean $\mu = 0$ and standard deviation $\sigma_r = 1/\sqrt{N-2}$ under the following assumptions:

- The populations yielding the samples x and y are normal
- The sample size *N* is "large"
- The population coefficient is zero ($\rho = 0$)
- The *N* pairs of observations are a *random* sample

If these assumptions are satisfied, a confidence band for *r* can be computed from the probability points of the standard normal distribution (Panofsky and Brier 1958; Chatfield 2004). Approximately 95% of a normally distributed variable should fall within ± 1.96 standard deviations of the population mean. For a population mean of zero and standard deviation $\sigma_r = 1/\sqrt{N-2}$, a 95% confidence interval for *r* is therefore

$$\frac{-1.96}{\sqrt{N-2}} \text{ to } \frac{+1.96}{\sqrt{\text{N-2}}}. \tag{11}$$

For example, if the sample size is $N = 200$, the 95% confidence interval is

$$\frac{-1.96}{\sqrt{200-2}} \text{ to } \frac{+1.96}{\sqrt{\text{200-2}}} \quad = \quad -0.1393 \quad \text{to} \quad +0.1393. \tag{12}$$

The confidence interval (11) applies to a two-tailed test with null and alternative hypotheses

H0: correlation coefficient is zero
H1: correlation coefficient is "different than" zero

A computed correlation coefficient greater in absolute magnitude than 0.1393 is "significantly different than zero" at the 0.05 α-level, which corresponds to the 95% significance level. In other words, the null hypothesis of zero correlation is rejected at $\alpha = 0.05$

Because the normal distribution is symmetrical, the α-threshold a two-tailed test is equivalent to the $\alpha / 2$ threshold for a one-tailed test. The hypotheses for the one-tailed test, where positive correlations of main interest, are

H0: correlation coefficient is less than or equal to zero
H1: correlation coefficient is "greater than" zero

In the example above, a computed correlation of $r = 0.15$ would indicate rejection of the null hypothesis at the 0.05 level for the two-tailed test and rejection of the null hypothesis at the 0.025

level for the one-tailed test. Whether to use a one-tailed or two-tailed test depends on the context of the problem. If only positive correlations (or only negative) seem plausible from the physical relationship in question, the one-tailed test is appropriate. Otherwise the two-tailed test should be used.

As a second example, consider a computed correlation of r=0.48 between a tree-ring index and a precipitation series, based on N=102 years of data. From equation (11), we compute that r must be larger than $1.96/\sqrt{100}$, or greater than 0.196 for significance at 95% in a two-tailed test. The computed r greatly exceeds this threshold. The conclusion is therefore to reject H0 at $\alpha = 0.05$.

If the sample size were only N=18 years, the normal approximation would yield a required threshold of r=1.96/4=0.49 for significance at 95%. Because the sample correlation, r=0.48, is smaller than the threshold, the conclusion using the large-sample approximation would therefore be "do not reject H0 at $\alpha = 0.05$". But recall that the same computed r for the same sample size resulted in rejection of the null hypothesis of $\rho = 0$ by the t-test of equation (2), which applies for small samples. We should not have used the large-sample approximation with such a small sample size. The comparison does illustrate that inappropriately using the large-sample approximation is conservative in the sense that the chance of a Type-I error (rejecting H0 when H0 is false) is not increased. On the other hand, inappropriate use of the large-sample approximation increases the likelihood of a type-II error – failing to reject the null hypothesis when the null hypothesis is false.

Significance testing when population correlation is nonzero

Sometimes the question is not whether a sample correlation coefficient is significantly different from zero, but whether one sample non-zero correlation coefficient is significantly different than another non-zero correlation coefficient. Or the question might be whether the sample indicates a population correlation different than some specified non-zero correlation (e.g., H0: ρ=0.5). If the population correlation is nonzero, the sample correlation cannot be regarded as a sample from a theoretical normal distribution with mean zero and standard deviation given by equation (15). Moreover, if the population correlation is nonzero, the distribution of the sample correlations about the population correlation is asymmetrical.

A solution is provided by the Fisher transformation, which converts the sample correlation r into another quantity, z, that is distributed almost normally (Snedecor and Cochran 1989, p. 188). The transformation is:

$$z = \frac{1}{2}\left[\ln(1+r) - \ln(1-r)\right], \tag{13}$$

where r is the sample correlation coefficient and z the transformed coefficient. The standard error of z is given by

$$\sigma_z = \frac{1}{\sqrt{N-3}} \tag{14}$$

where, as before, N is the sample size (adjusted for persistence if necessary). The probability points of the normal distribution can then be used to estimate the confidence band for z. For example, the $\alpha = 0.05$ confidence band is $z \pm 1.96\sigma_z$.

This method can be used to determine whether two correlation coefficients for different samples are significantly different from one another. Consider, for example, correlations between tree-ring indices and precipitation for different sub-periods:

$$r_1 = 0.65, \quad \text{40-year period 1901-1940}$$
$$r_2 = 0.80, \quad \text{60-year period 1941-2000}$$

One notion might be that changing climate has resulting in a stronger physical linkage between climate variations and tree growth from the early period to the later. A reasonable null hypothesis then might be that the two independent values r_1 and r_2 are estimates of the same population correlation ρ. To test the hypothesis, we first convert the correlations to the transformed variable using equation (13):

$$z_1 = 0.7753$$
$$z_2 = 1.0986.$$

We then compute the difference of the transformed correlations as

$$D = z_2 - z_1$$
$$= 0.3233,$$

and make use of the relationship that the variance of a difference of two random variables is equal to the sum of the individual variances to compute the variance of D:

$$\sigma_D^2 = \sigma_{z_1}^2 + \sigma_{z_2}^2$$
$$= \frac{1}{\sqrt{40-3}} + \frac{1}{\sqrt{60-3}}$$
$$= 0.0446.$$

The standard error of D is

$$\sigma_D = \sqrt{\sigma_D^2} = 0.2111.$$

The quantity D follows a normal distribution, so that the 95% confidence interval of D is

$$D \pm 1.96\sigma_D = D \pm 0.4138.$$

Because this confidence interval includes zero, the null hypothesis cannot be rejected at $\alpha = 0.05$.

Random samples

The significance tests described above assume random samples. This assumption is necessary because we are testing for a relationship that supposedly is characteristic of the unknown population. This assumption can be voided for many possible reasons. For example, in testing for a relationship between temperature and precipitation, the samples may have been drawn from a period of anomalous atmospheric circulation that favored a particularly strong or weak relationship. Any conclusion on significance would apply only under similar circumstances of atmospheric circulation. In a sense, the "population" must be redefined to be more restrictive.

Effective sample size

It is assumed that each observation for the correlation analysis represents an independent piece of information, such that a sample size of *N* represents *N* distinct pieces of information. In other words, the observations are assumed to be independent of one another. But observations in a time series are seldom independent, especially for time series derived from physical systems characterized by storage – for example, the annual flow of a river. Autocorrelation in physical systems is often reflected by a nonzero first-order autocorrelation coefficient. For a time series of

length N , autocorrelation essentially reduces the number of independent observations below N. Recall that the standard deviation of the correlation coefficient depends on N through

$$\sigma_r = 1/\sqrt{N-2} \tag{15}$$

For autocorrelated time series, equation (15) should be altered to use an adjusted, or "effective" sample size. Decreasing the number of independent observations has the effect of increasing σ_r , which in turn widens the confidence bands. A sample correlation that might otherwise be significantly different than zero might not be so if the series are autocorrelated. Many other statistical tests also assume independence of observations. Below are equations for computing the effective sample size. More information on these adjustments can be found elsewhere (WMO 1966; Dawdy and Matalas 1964). The equations are derived based on the assumption that the autocorrelation in the series is *first-order* autocorrelation (dependence on lag-1 only). In other words, the governing process is *first-order autoregressive*, or *Markov*. Computation of the effective sample size requires only the sample size and first-order sample autocorrelation coefficient. Equations are given below for the effective sample size of a time series, and the effective sample size for computation of a correlation coefficient between two autocorrelated time series.

Effective sample size – univariate adjustment (Mitchell et al. 1966)

$$N' = N\frac{(1-r_1)}{(1+r_1)} \quad , \tag{16}$$

where N is the sample size, N' is the effective samples size, and r_1 is the first-order autocorrelation coefficient. For example, a series with a sample size of 100 years and a first-order autocorrelation of 0.50 has an adjusted sample size of

$$N' = 100\frac{(1-0.5)}{(1+0.5)} \quad = \quad 100\frac{0.5}{1.5} \quad \approx \quad 33 \text{ years.} \tag{17}$$

Effective sample size for correlation coefficient (Dawdy and Matalas 1964)

$$N' = N\frac{(1-r_{1,x}r_{1,y})}{(1+r_{1,x}r_{1,y})}, \tag{18}$$

where $r_{1,x}$ and $r_{1,y}$ are the first-order autocorrelation coefficients of time series x and y , and the other variables are defined as in (16). The adjustment depends on the product of the first-order autocorrelation coefficients in such a way that if either series has zero first-order autocorrelation, the effective sample size equals the original sample size (no reduction in effective samples size). As an example of the adjustment and the effect on assessment of significance of *r*, consider two time series of length 200 years with first-order autocorrelations of 0.50 and 0.40:

$$N' = N\frac{\left[1-(0.50)(0.40)\right]}{\left[1+(0.50)(0.40)\right]} \quad = \quad 200\frac{[1-0.20]}{[1+0.20]} \quad = \quad 200\frac{[0.80]}{[1.20]} \quad = \quad 200(.667) \approx 133. \tag{19}$$

The 95% confidence interval without the adjustment of sample size is approximately $0 \pm \left(1.96/\sqrt{N-2}\right) \;=\; 0 \pm \left(1.96/\sqrt{198}\right) = \pm 0.139$, while the 95% interval with adjustment is $0 \pm \left(1.96/\sqrt{131}\right) = \pm 0.171$. Note from equation (18) that if the lag-1 autocorrelation of either series is zero, the factors in parentheses reduce to 1.0, and the effective sample size equals the original sample size. Because of random sampling variability, it is unlikely that the lag-1 autocorrelation for any observed time series will be exactly zero, even if the generating process (unknown) has no autocorrelation. It is a good idea therefore to first test the first-order autocorrelation coefficient(s) of the individual series for statistical significance before deciding to use an adjusted, or "effective" sample size, for testing the hypothesis of significant correlation between the two series.

Bonferonni adjustment

To a customer at the Olive Garden, "Bonferroni" might bring up an image of bland pasta smothered with tomato sauce. To a dendrochronologist, "Bonferroni" at once brings to mind the famous Italian mathematician Carlo Emilio Bonferroni (1892-1960), after whom a well-known adjustment of confidence intervals for multiple significance testing is named.

The "Bonferroni adjustment" is an adjustment made to confidence levels when multiple statistical tests are evaluated simultaneously. If pairs of independent time series are randomly drawn several times and each time a sample correlation is computed and tested for significance at $\alpha = 0.05$ using the approximate confidence interval described earlier (see equation (11)), the probability of getting at least one "significant" correlation by chance alone is actually greater than 0.05. The confidence should be widened to account for the increased probability (Snedecor and Cochran 1989, p. 116, 167).

The algorithm for the Bonferroni adjustment can be derived from the binomial distribution, which gives the probability of a least one success in k trials given the probability of success in a single trial. If the probability of one success (one test result being significant) is α', the probability of any one test being non-significant is $(1-\alpha')$, and if the test is repeated k times, the probability that none of the test results is significant is

$$P(\text{none signif.}) = (1-\alpha')^k. \tag{20}$$

For any given number of tests and desired probability that no tests results are significant, we can solve (20) for α'. In solving the equation, we can make use of the fact that, for α' small,

$$(1-\alpha')^k \cong 1 - k\alpha'. \tag{21}$$

Substituting (21) into (20) gives

$$P(\text{none significant}) = 1 - k\alpha' \tag{22}$$

Now the probability of one or more tests being significant is simple 1 minus the probability that none of the tests are significant, so that

$$1 - P(\text{one or more tests significant}) = 1 - k\alpha' \tag{23}$$

Equation (23) can be written

$$1 - \alpha = 1 - k\alpha' \tag{24}$$

where α is the alpha-level for a simultaneous confidence interval, or the maximum acceptable probability that one or more of the tests is significant. We can regard any specific desired probability that none of the tests is significant as an alpha-level for a simultaneous confidence interval, Solving (24) for α' gives

$$\alpha' = \alpha / k. \tag{25}$$

The numerator on the right-hand side can be regarded as the α – level for the simultaneous confidence interval. Thus the adjustment amounts to dividing the desired alpha-level by the number of test statistics evaluated. For significance of correlation coefficients, if the number of correlations tested is 10, and we want a 95% confidence level ($\alpha = 0.025$ for 2-tailed test), the α -level for the Bonferroni-adjusted confidence interval is $\alpha' = \alpha/k = 0.025/10 = 0.0025$. Lowering α widens the confidence interval, making it less likely that a particular level of sample correlation is significantly different than zero.

The least objectionable use of a Bonferroni adjustment appears to be when the null hypothesis is unequivocally the "universal" null hypothesis. If k correlations are computed, the universal null hypothesis is "none of the k population correlations is different than zero." The Bonferroni adjustment reduces the probability of a type-I error. (A type-I error is rejection of the null hypothesis when it is true.) It is argued, however, that in inappropriate application, where the null hypothesis is not really the universal null hypothesis, the Bonferroni adjustment raises the probability of a type-II error (not rejecting the null hypothesis when it is false). Some authors, moreover, argue that the Bonferroni adjustment should be avoided completely because it discourages scientists from pursuing leads that may turn out to be wrong (e.g., Rothman 1990). The Bonferroni adjustment is introduced here as an option for adjustment of confidence interval for evaluating a series of time-windowed correlation coefficients (see below).

9.3 Interpretation of correlation coefficient

Some caveats to interpretation of the correlation coefficient are

- Relationship in question should be linear
- Correlation does not imply causality
- Correlation does not address lagged relationships
- Statistical significance may not imply practical significance

Linearity

The correlation coefficient between time series x_t and y_t summarizes the strength of linear relationship, as reflected by closeness of fit of a least-squares straight line of *y* on *x*. If a relationship is curvilinear, the correlation coefficient does not measure the true strength of relationship. Moreover, even for a linear relationship, the correlation coefficient might not measure the strength of relationship over the full range of the variables. For example, a moderately high correlation for tree-ring index against precipitation might result from a strong relationship in drier than normal years and a weak relationship in wetter than normal years.

Scatterplots can help identify serious departures from linearity. A plot for all of the data can be examined for curvature or fanning out of the cloud of points toward the ends of the axes. Separate scatterplots can also be made for observations over different *x* ranges or *y* ranges to quantify change in strength of relationship with *x* or *y*.

An example of such scatterplots is given in Figure 9.1 for tree-ring indices at two northern California sites. Recall from equation (1) that the correlation coefficient is determined by the product of departures from the means of the two variables. Departures of the same sign (either + or -) contribute to positive correlation. Departures of opposite sign contribute to negative correlation. Accordingly, points in quadrants I and III of Figure 9.1 (upper left) contribute to

positive correlation, and points in quadrants II and IV to negative correlation. The orientation of the cloud of points is from lower left to upper right, meaning that points are predominantly in quadrants I and III, as reflected by the positive correlation coefficient.

The remaining three plots in Figure 9.1 are scatterplots and straight line fits to subsets of observations determined by the x-values (Boles Creek). For example, at lower left are shown points for observations with the lowest third (lowest tercile) of Boles Creek indices. Similarly, the observations for the middle tercile are at upper right, and for the highest tercile at lower right. A significant positive relationship between the indices at the two sites is indicated when Boles Creek index is low and when Boles Creek index is high. No relationship is evident for intermediate values of the Boles Creek index (upper right).

The scatterplot for the entire data (upper left) shows a significant correlation even after adjustment of the sample size for persistence. The original sample size is 199 years (1800-1998), while the effective sample size is 123 years. The difference in sample sizes comes from both x and y having significant lag-1 autocorrelation. The lag-1 autocorrelation coefficients are $r_{1,x} = 0.38$ and $r_{1,y} = 0.62$. Substitution of these values into equation (18) yields the effective sample size of 123.

Note that no adjustment of sample size has been made for the tercile subset analyses in Figure 9.1. Each subset has approximately 1/3 of the 199 observations, and the sample size used for the significance testing has not been adjusted downward. No adjustment is made because the data for the subsets are not successive observations arranged in time. Depending on the spacing of the subsets observations, the observations may be more or less random, or independent of one another, even though the original series is autocorrelated.

What can be done if the scatterplot of y on x is nonlinear? If the plan is to apply the data in methods assuming linear relationships (e.g., multiple linear regression), one option is to transform either or both variables to straighten the scatterplot. Methods of transformation are described by Hoaglin et al. (1983). Two simple transformations that sometimes work are the log transform and power transform. If the original variable is x and the transformed variable y, these transformations are:

- Log-transform $v = \log_{10}(y)$
  - compresses scale at high end of distribution; useful on y when scatterplot of y on x shows increasing scatter with increasing x
  - in hydrology, frequently used to transform discharge data to normality
- Power transformation $v = y^p$ (Hoaglin et al. 1983)
  - most often used are square-root transform $(p = 0.5)$ and squaring $(p = 2)$; square root transform has similar effect to log-transform
  - p is usually restricted to positive values
  - if p is negative, transformation $v = -y^p$ preferred

Note that the "z-score" transformation, or transformation to standardized anomalies from the mean, will have no effect of correlation, as the correlation coefficient itself can be written in terms of standardized anomalies (see equation (1)). The z-score transformation is useful, however, to remove differences of level and scale in time series before plotting.

Causality

A significant correlation coefficient between x and y does not necessarily imply that variations in x "cause" variations in y, or vice versa. Sometimes the correlation results simply from both x and y being associated with a third variable that might have some causative influence on

both x and y. For example, a positive correlation between tree-growth variations in Arizona and Colorado in no way implies that tree-growth in one region affects tree-growth in the other, but merely that trees in both regions are influenced by a third variable – precipitation – which happens to be spatially coherent across the Southwest.

Lagged relationships

If two correlated time series, x and y, are shifted so that they are offset in time, the correlation may decrease or disappear. Likewise, the physical system may build in a natural lag such that simply aligning the time series does not capture the strength of relationship. Or the influence of one variable may be spread over several observations of the other variable. In such cases, simple correlation analysis can be misleading, and methods using the cross-correlation function are more appropriate (Box and Jenkins 1976).

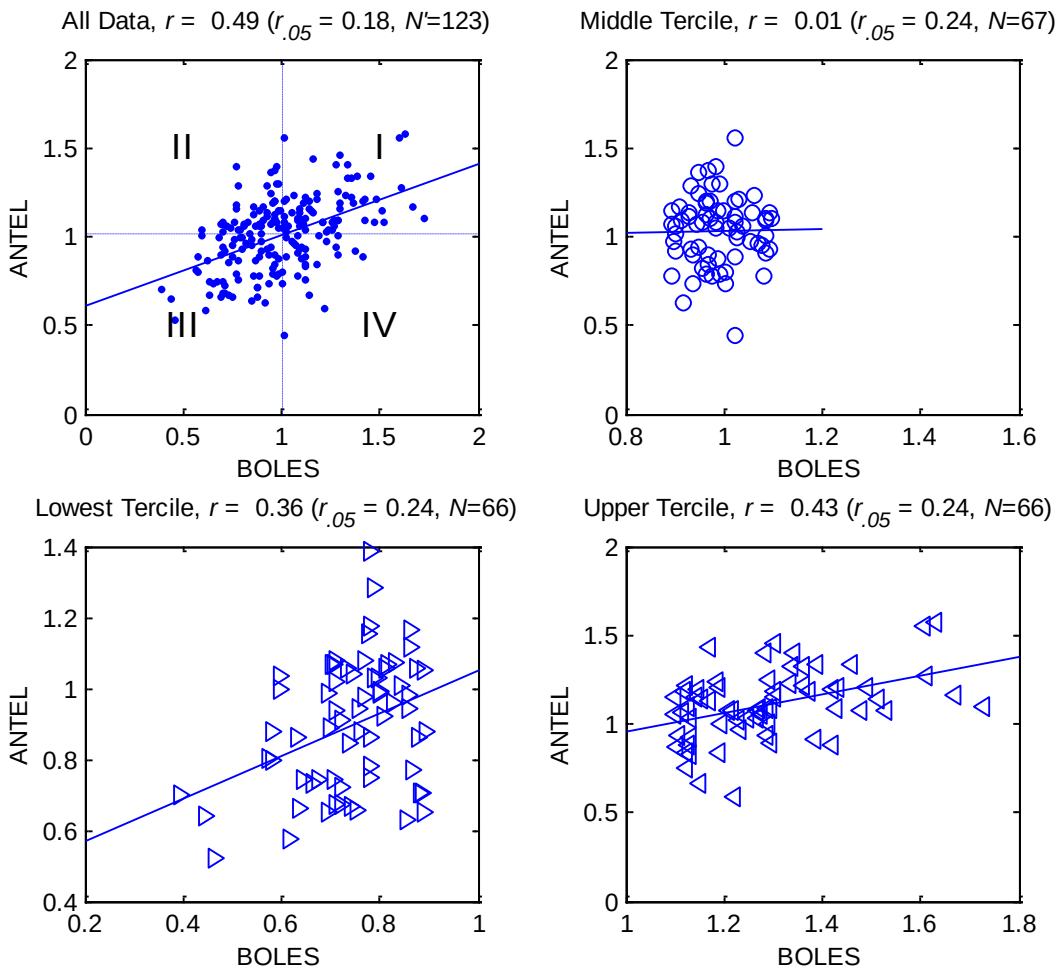

Figure 9.1. Scatterplots of Antelope Lake tree-ring index against Boles Creek tree-ring index, 1800-1998. Upper left: scatter plot for all data, with best-fit straight line (solid), and dashed lines at the sample means. Upper right: scatterplot for observations in middle tercile of Boles Creek data. Lower left: scatterplot for observations in lower tercile of Boles Creek data. Lower right: scatterplot for observations in upper tercile of Boles Creek data. Annotated are the sample correlation coefficient (*r*), the threshold correlation for statitistical significance at α =0.05 level for two-tailed test $(r_{.05})$, and the sample size (N) or effective sample size (N') for computing significance.

Practical significance

The distinction should be made between statistical significance and practical significance. A statistically significant correlation need not be practically significant. When the objective us to infer variations in y from variations in x, interest is often in how much of the variance of y is predictable by some model (e.g. regression) with x as a predictor. For simple linear regression, the decimal fraction of variance predictable by such a model is given by the squared correlation coefficient. The size of the correlation coefficient is therefore critical. But the statistical significance of a correlation coefficient is a function of not only the size of correlation but the length of analysis period. For the same level of correlation, the longer the time series, the greater

the statistical significance. The practical significance, however, remains the same. For example, consider two time series of length 10,000 years correlated at $r = 0.03$. The correlation is "significant" at the 0.05 alpha level, but less than $1/10$ of a percent of the variance of one variable can be explained by a linear regression on the other.

9.4 Stability of correlation

Changes of strength of correlation with time can be shown graphically by a plot of correlation coefficients computed in a sliding window. For example, for a 3000-year-long series, correlations can be computed and plotted for the 30 distinct 100-year periods. Such a plot can be useful for qualitatively evaluating the time pattern of change in strength of relationship. Two basic statistical questions that can be addressed with such a plot are:

1. Is the relationship statistically significant for <u>any</u> of the periods?
2. For <u>which</u> periods is the relationship significant?

These questions can be answered by comparing the computed correlations with the threshold correlation for some selected α – level (e.g., $\alpha = 0.05$). The threshold level of correlation for statistical significance can be computed using the probability points of the normal distribution, as described above. Two considerations in computing the threshold level are (1) reduction of effective sample size due to autocorrelation in the two series, and (2) Bonferroni adjustment of the confidence interval to account for multiple, or simultaneous, evaluation of statistics. Methods for the effective sample size and Bonferroni adjustment have already been covered.

For the question of whether the relationship is statistically significant for <u>any</u> of the periods, the appropriate null hypothesis is the "universal" null hypothesis: the population correlation coefficient is zero for all sub-periods. Rejection indicates that the population correlation is nonzero for at least one period. The Bonferroni adjustment is essential in this application. For example, with 30 distinct sample correlations evaluated, the appropriate α – level for 95% significance is approximately 0.05/30=0.0017.

For the question of <u>which</u> periods the relationship is significant for, the Bonferroni-adjusted confidence interval can also be used, but with the drawback that the risk of a type-II error (failing to reject a false null hypothesis) is increased. An alternative is to specify a sub-period of interest a-prior and compute and test the correlation for just that period. Then no Bonferroni adjustment is necessary, the confidence interval is narrower, and a smaller (in absolute magnitude) correlation is required for significance.

An example of sliding, or windowed, correlation plots for the Boles Creek and Antelope Lake tree-ring series is shown in Figure 9.2. The series and period of analysis are the same as for the example discussed previously (Figure 9.1). The selected window has a width of 50 years and is offset by 50 years (no overlap). The plotted correlations indicate that the universal null hypothesis should be rejected at $\alpha = 0.05$. In other words, the correlation is judged significant for at least one of the sub-periods. In fact, the correlation for the middle period $(r = 0.75)$ greatly exceeds the threshold.

When evaluating sliding correlations, it is important to also consider variations in the sub-sample means and standard deviations of the individual series. (The sub-sample means for the example above are plotted in Figure 9, middle.) Because the correlation coefficients are computed from scaled departures from the sub-sample means (recall equation (1)), the correlation might be low for a period in which the sub-period means of both variables are anomalously high or low and the variation around those means anomalously small. A conclusion of "no relationship" could be misleading in this case. For example if two tree-ring series plunge to extremely low levels over

an entire 50-year period and the variability within that period is small, the correlation for the sub-period might be close to zero, but the "relationship" between the series is strong -- at least in the context of the longer record.

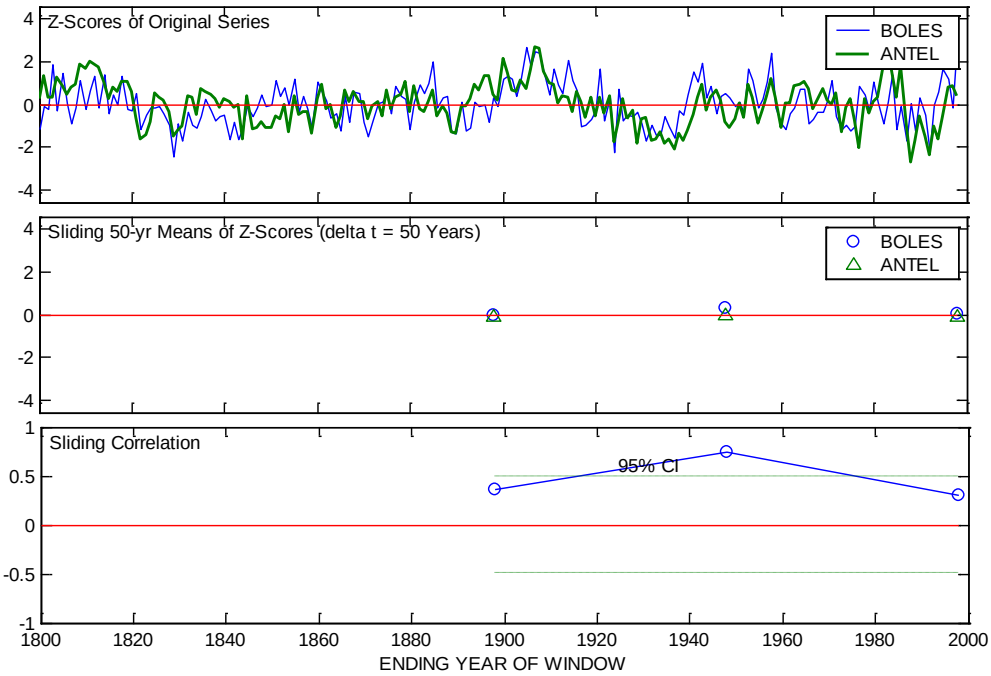

Figure 9.2. Sliding correlations for Boles Creek and Antelope Lake tree-ring indices for analysis period 1800-1998. Correlations for 50-year window plotted at the end of each windowed period. confidence interval adjusted for autocorrelation and multiple testing (Bonferroni adjustment). Top: time series plots of series transformed to "z-scores", or standardized anomalies. Middle: sample means of the annual z-scores for 50-year sub-periods. Bottom: sample correlations, with 95% confidence interval.

To answer the question whether the relationship between two variables is stronger in one time period than another, one must test for the significance of *difference* in correlation. For this question, the section "Significance testing when population correlation is nonzero" (see above) applies, and it is necessary to transform the sample correlations before significance testing. Random sampling variability will always result in some chance differences in correlation from one time period to another. If two time series are strongly correlated, windowed correlations are consequently expected to wander to some extent purely as a result of stochastic behavior. Gershunov et al. (2001) emphasize this point, and suggest that bootstrap methods be applied to evaluate significance of fluctuations in windowed correlations.

9.5 References

Box, G.E.P., and Jenkins, G.M., 1976, Time series analysis: forecasting and control: San Francisco, Holden Day, p. 575 pp.

Chatfield, C., 2004, The analysis of time series, an introduction, sixth edition: New York, Chapman & Hall/CRC.

Dawdy, D.R., and Matalas, N.C., 1964, Statistical and probability analysis of hydrologic data, part III: Analysis of variance, covariance and time series, *in* Ven Te Chow, ed., Handbook of applied hydrology, a compendium of water-resources technology: New York, McGraw-Hill Book Company, p. 8.68-8.90.

Gershunov A., Schneider N. and Barnett T., 2001, Low-frequency modulation of the ENSO–Indian Monsoon rainfall relationship: signal or noise?: J. of Climate 14, 2486-2492.

Haan, C.T., 2002, Statistical methods in Hydrology, second edition: Ames, Iowa, Iowa State University Press.

Hoaglin, D.C., Mosteller, F., and Tukey, J.W., 1983, Understanding Robust and Exploratory Data Analysis: John Wiley & Sons, Inc., New York, p. 447.

Panofsky, H.A., and Brier, G.W., 1958, Some applications of statistics to meteorology: The Pennsylvania State University Press, 224 p.

Mitchell, J.M., Jr., Dzerdzeevskii, B., Flohn, H., Hofmeyr, W.L., Lamb, H.H., Rao, K.N., and Wallén, C.C., 1966, Climatic change: Technicall Note No. 79, report of a working group of the Commission for Climatology; WMO No. 195 TP 100: Geneva, Switzerland, World Meteorological Organizaton, 81 p.

Rothman, K. J. (1990), No adjustments are needed for multiple comparisons, Epidemiology, 1 (1), 43–46.

Snedecor, G.W., and Cochran, William G., 1989, Statistical methods, eighth edition, Iowa State University Press, Ames, Iowa, 803 pp.

Wilks, D.S., 1995, Statistical methods in the atmospheric sciences: Academic Press, 467 p.