What makes a correlation signficant?

Recall from last time that we have defined correlation as:

$$r = \frac{\overline{(x-\overline{x})(y-\overline{y})}}{\sqrt{\overline{(x-\overline{x})^2}\,\overline{(y-\overline{y})^2}}} \tag{1}$$

How do we tell if a correlation is significantly different from zero? Clearly this will depend on the number of elements in x and y. If x and y each have two elements, then r will necessarily be 1 or -1, but that does not indicate any statistical link between x and y.

To evaluate correlation, we assume as a "null hypothesis" that the measurements are completely unrelated. Then we test to see if the null hypothesis is violated, that is if the correlation coefficient r exceeds what we would expect to find if x and y were produced by a random number generator.

When the number of data points N is large, r should have a normal distribution with mean zero and standard deviation $1/\sqrt{N}$. (We can predict the standard deviation, knowing that for $N = 2$, the standard deviation must be 1, and that for an averaged quantity, the standard deviation should scale like $\sigma/\sqrt{N}$.) This means that the PDF for r will be

$$P(r)\,dr = \frac{1}{\sigma_r\sqrt{2\pi}}\exp\left[-\frac{r^2}{2\sigma_r^2}\right]\,dr = \frac{\sqrt{N}}{\sqrt{2\pi}}\exp\left[-\frac{r^2N}{2}\right]\,dr \tag{2}$$

The cumulative distribution function for r,

$$C(r_o) = \int_{-\infty}^{r_o} P(r')\,dr', \tag{3}$$

formally defines the probability of measuring a correlation coefficient r less than r_o. Here we are interested in the absolute value of r, so we'll look at

$$S(r_o) = \int_0^{r_o} 2P(r')\,dr'. \tag{4}$$

The function S is defined as the error function, erf. Here $S(r_o)$ defines the chance that random data should give a value smaller than the observed value r_o. We would like the opposite scenario, the probability that random noise would produce a correlation coefficient exceeding r_o. That comes from the complementary error function, which is $1 -$ erf. So formally, we compute, the probability that the results are just an artifact of analyzing a small number of noisy observations:

$$Pr = \text{erfc}\left(\frac{|r|\sqrt{N}}{\sqrt{2}}\right). \tag{5}$$

If our probability Pr is small, then r is likely to indicate real correlation.

Sometimes this test is a little cumbersome. For a given value of N, we would like to know how big r needs to be to indicate a significant correlation. For this we can invert (5).

$$r_{sig} = \text{erf}^{-1}(s)\sqrt{\frac{2}{N}} \tag{6}$$

where erf^{-1} is the inverse error funtion ("erfinv" in Matlab) and s is a significance level. Often we use $s = 0.95$ to test for cases where there is only a 5% chance that noise could produce equivalent results.

The web site links to a section from *Numerical Recipes* by Press et al., that discusses other tests for significant correlation when N is small, and you can take a look at that.

Often the greater challenge than dealing with small N, is deciding how big N really is. We will address that by looking closely at the autocorrelation.

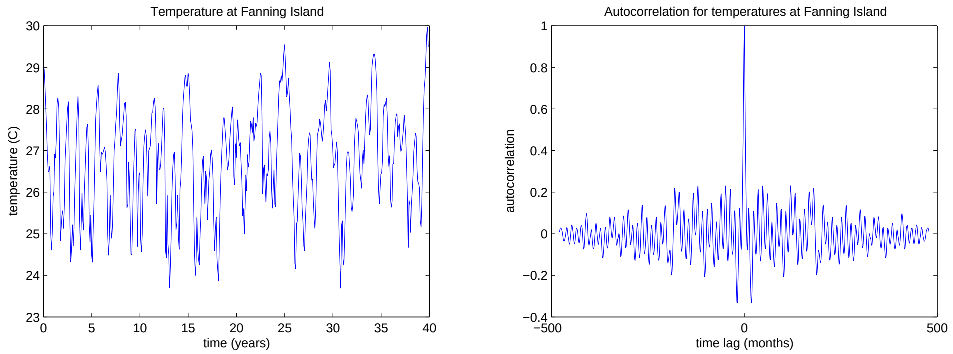

Figure 1: (left) Time series of temperature at Fanning Island (in the tropical Pacific) from a 40-year simulation of the NCAR Community Ocean Model. (right) Autocorrelation for time series, normalized to have maximum correlation of 1 and to attenuate to zero at high temporal lags.

Autocorrelation

Autocorrelation measures how well a series of data x is correlated with itself. If we just compute r for x versus x the result is guaranteed to be one, so not very interesting. We are really interested in the lagged autocorrelation, which measures how rapidly x changes. Thus, we compute

$$C(\Delta t) = \frac{\overline{(x(t)-\overline{x})(x(t+\Delta t)-\overline{x})}}{\overline{(x-\overline{x})^2}}, \tag{7}$$

where the overbar is telling us to average over all measurement times t. The autocorrelation is symmetric about zero, since the average of $\overline{x(t)x(t+\Delta t)}$ is the same as the average $\overline{x(t-\Delta t)x(t)}$.

Figure 1 shows the timeseries and autocorrelation of of temperatures at Fanning Island that we examined in our first computer session. The correlation in Figure 1 goes to zero at large lags because the averages in C are divided by N regardless of the size of the lag. We can also compute an unbiased autocorrelation by dividing by the actual number of data points available at each lag, $N - n$, as shown in the left panel of Figure 2. The right panel of Figure 2 shows an enlargement of the autocorrelation near zero lag.

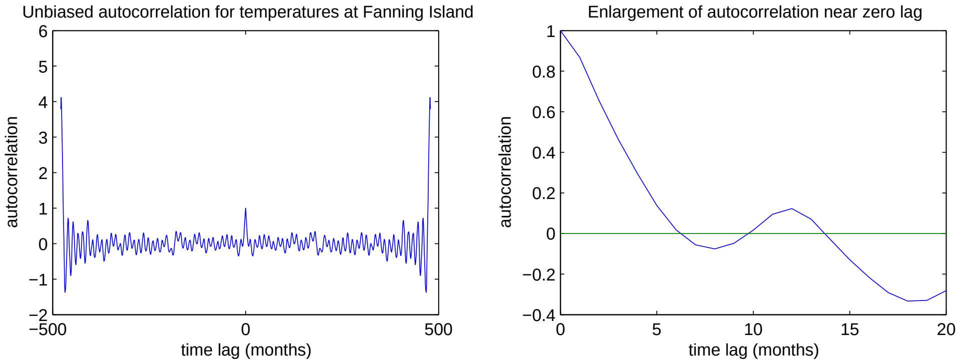

Figure 2: (left) Unbiased autocorrelation for time series in Figure 1. At large lags, autocorrelation estimates are based on fewer estimates and therefore more uncertain (implying possibly large.) (right) Enlargement of (biased) autocorrelation near zero time lag, shows that autocorrelation drops to zero for lags of about 6 months.