1. Probability or frequency of occurrence:

Suppose that any random player's chance of winning the lottery is 1 in 1,000. But you go to a psychic whose predictions are 90% accurate (in every sense: 10% false positives, 10% false negatives). The psychic tells you that you will win this week. **What is the probability you will win?**
(this is a much nicer form of the rare-cancer-diagnosis question in the book!)

To solve it, fill in the tables with joint and marginal *probability* distributions of lottery outcome and psychic prediction. I strongly recommending counting at least 10,000 individuals as integers for definiteness in the upper left corner, then simply putting another integer in the denominator in lower right corner of each box to make the results a true "probability." Remember, probability distributions (joint and marginal) add up to 1. I have done "Marginal by W/L" for you.

Joint and marginal distributions:	WIN	LOSE	Marginal by told
TOLD WILL WIN			
TOLD WILL LOSE			
Marginal by W/L →	10 /10000	9990 /10000	

(The question is: Of the people told they will win, how many really will win?)

2. Simple univariate statistics.

Denote the average (mean) over a set of values with an overbar $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ and deviations from the mean using a prime $y' = y - \bar{y}$, as usual. For this problem we always use 1/N rather than 1/(N-1) in these statistics.

a. Confirm using the dataset y={-5, 4, 10} that $\overline{y'} = 0$ and that $\overline{y^2} = \bar{y}^2 + \overline{y'^2}$.

b. What are the mean and standard deviation of y?

c. Let x = {5, 4, -10}. What is its standard deviation?

d. What is cov(x,y), the covariance between x and y? Please express it with sums and fractions, there is no need to use a calculator.

e. What is the correlation coefficient cor(x,y)? Again, I am looking for a sum showing the construction of the answer, not a number from a calculator.

f. What is the regression coefficient of y on x (slope R of the line y = Rx + resid1)?

g. What is the regression coefficient of x on y (slope S of x = Sy + resid2)?

h. What fraction of the variance of y does the term Rx explain?

i. What fraction of the variance of x does the term Sy explain?

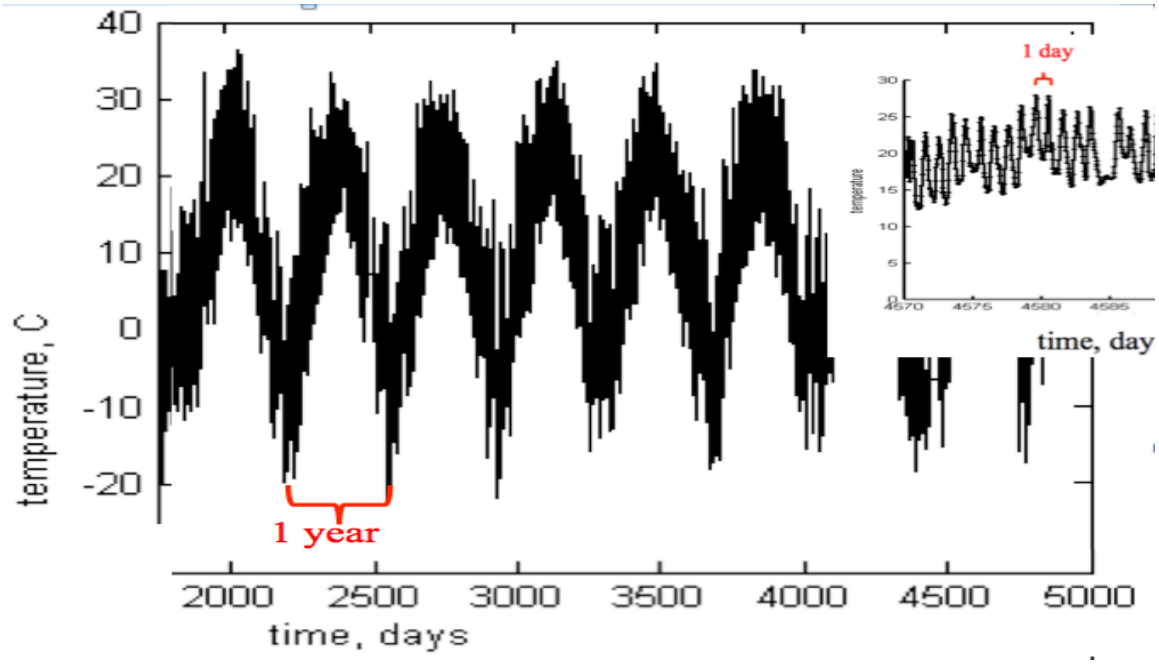
3. Variance and its orthogonal components: trends and Fourier harmonics

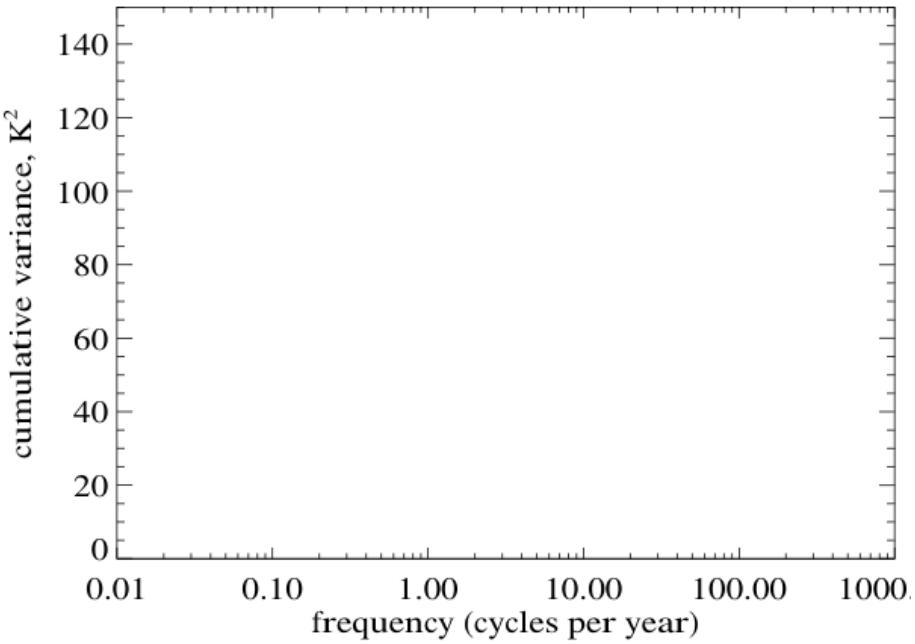
a. What is the variance of a linear trend with slope m in a finite time series? For definiteness, consider the mean square of y'=mx' on the x interval [-1,1]. Don't forget to divide by 2 (the length of the interval) when you evaluate the overbar in $\overline{y'^2}$ (because the bar is an average, not just an integral).

b. Q: What is the variance of a sine or cosine curve with amplitude A, averaged over infinity or a whole number of periods? Hint: the answer is the same for sine and cosine, and $\sin^2 + \cos^2 = 1$. Average both sides of this equation over a cycle $[-\pi,\pi]$ and solve.

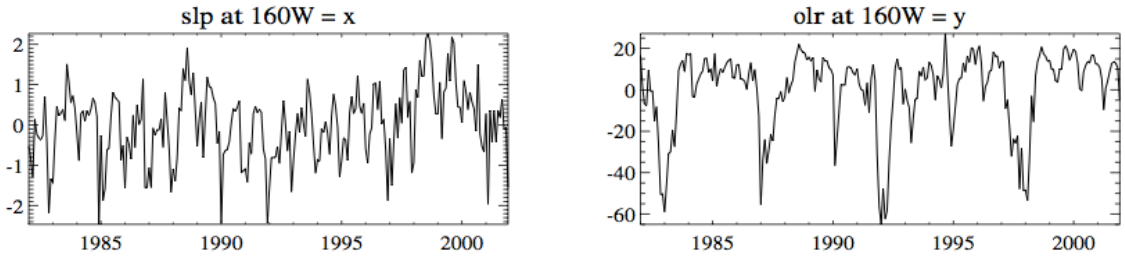
c. Consider the answers to parts a. and b. above, in light of the fact that $\sin(x) \sim x$ for small x. Which has a larger variance, a trend Ax or a cycle $A\sin(x)$? Can you explain why, perhaps by considering a sketch or graph of x^2 and $\sin^2(x)$ over the interval $[-\pi,\pi]$ which obeys $\sin(x) \sim x$ at small x?

4. Sketch a cumulative power spectrum for this time series. Its total variance is about 130 K^2, which should make sense based the typical magnitude of its excursions from the mean.

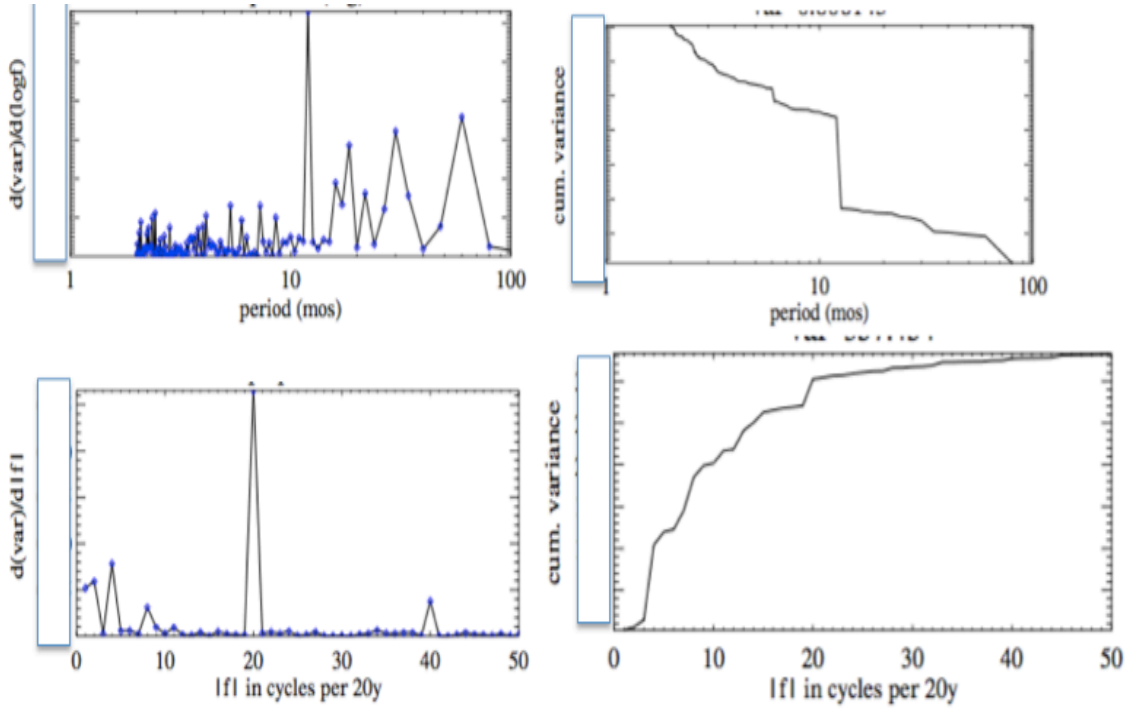




5. SLP (mb) and OLR (Wm^{-2}) time series from our 20-year equatorial monthly time-longitude dataset are plotted below, for one longitude (160W).



Identify which time series corresponds to each of these 4 spectra plots. (Two copies are given: one for scratch work, and one for your final answers). Two are spectral density, two are cumulative variance. Estimate a rough plot range (the covered-up numbers) on each plot's vertical axis. Indicate the units on each plot's vertical axis.



6. Interpreting multiple regression

Background: ENSO cycles in the central Pacific involve warm phases when SST is anomalously warm, air converges more than usual, and rainfall is enhanced; interspersed with cold phases when the opposite anomalies occur.

Student A decides to "explain" or "predict" precipitation anomalies associated with ENSO cycles as resulting from SST anomalies, using a simple univariate regression:
$P'(t) = R\ T'(t)$ *+resid1*
and derives the usual least-squares regression coefficient, R = corr(T',P') σ_P/σ_T.

Student B prefers to "explain" or "predict" P as a combination of thermodynamics and dynamics, using multiple regression on SST and wind divergence d:
$P'(t) = m_1\ T'(t) + m_2\ d'(t)$ *+resid2.*

a. Express Student B's problem in a **d = Gm** form as used in the book.

$$\begin{bmatrix} \\ \\ \\ \\ \end{bmatrix} = \begin{bmatrix} \\ \\ \\ \\ \end{bmatrix} \begin{bmatrix} \\ \end{bmatrix} + resid$$

b. The best estimate of **m** (minimizing the RMS of the residual) is $(\mathbf{G^TG})^{-1}\ \mathbf{G^Td}$. Use words to try to explain the meaning or information content in the factors $(\mathbf{G^TG})^{-1}$ and $\mathbf{G^Td}$ in this case. You may use "similarity" or "projection" or "covariance" to describe what $\mathbf{A^TB}$ measures about two time series (column vectors) **A** and **B**.

c. Based on the mutually positive associations noted in the Background information, might there be some challenge in solving for **m**? Why, and how would it manifest mathematically in your calculation? How might Student B address the issue?

d. The following equality is not true as written. Use the Chain Rule for P(T,d), just as we do to relate Total and Eulerian time derivatives in fluid dynamics, to complete the equation. Which derivatives in the equation correspond to R and m_1?

$$\frac{dP}{dT} = \frac{\partial P}{\partial T}\Big|_d$$

7. Consider the following filter matrices, $\mathbf{M_1}$, $\mathbf{M_2}$:

$$\begin{bmatrix} g(t) \\ \downarrow \\ \downarrow \\ \downarrow \\ \cdots \end{bmatrix} = \begin{bmatrix} ? & & & & & \\ -1 & 0 & 1 & & 0 & \\ & -1 & 0 & 1 & & \\ & & -1 & 0 & 1 & \\ & 0 & & -1 & 0 & 1 \\ & & & & & ? \end{bmatrix} \begin{bmatrix} f(t) \\ \downarrow \\ \downarrow \\ \downarrow \\ \cdots \end{bmatrix} \qquad \mathbf{g = M_1 f}$$

$$\begin{bmatrix} g(t) \\ \downarrow \\ \downarrow \\ \downarrow \\ \cdots \end{bmatrix} = 1/3 \begin{bmatrix} 2 & 1 & & & & \\ 1 & 1 & 1 & & 0 & \\ & 1 & 1 & 1 & & \\ & & 1 & 1 & 1 & \\ & 0 & & 1 & 1 & 1 \\ & & & & 1 & 2 \end{bmatrix} \begin{bmatrix} f(t) \\ \downarrow \\ \downarrow \\ \downarrow \\ \cdots \end{bmatrix} \qquad \mathbf{g = M_2 f}$$

a. What names would we give these, based on what each one does to the discrete time series f(t) = column vector **f**? Please ignore the problem of deciding what to do at the end points (the ? and 2 matrix entries) and just consider structure in the middle.