Part I: Probability

Suppose that a random player's chance of winning the lottery is 1 in a million.
But you go to a psychic, whose predictions are 99% accurate. Really, 99% accurate.

a. The psychic tells you that you will win this week! What is the probability you will win?
b. The psychic tells you that you will lose this week. What is the probability you will win?

Hint: it will help a LOT to consider integers: 100 million people going to equally good psychics and playing the lottery. Just tally them as integers contingency tables (the joint distribution and the marginal distributions). (To make probabilities, you just have to divide each box by the proper total.)

Then you just count heads: How many individuals are told they will win? How many of those do win? Since "you" are typical of this group, this is the answer to a.
Answer a = 99 told and win / (.99x99,999,900 + 99) told yes

Joint and marginal distributions:	WIN	LOSE	Marginal by told
TOLD WILL WIN	99 /100,000,000	.99 x 99,999,900 /100,000,000	.99 x 99,999,900 +99
TOLD WILL LOSE	1	.1 x 99,999,900 = 9,999,990	9,999,991
Marginal by W/L →	100 /100,000,000	99,999,900 /100,000,000	

Your answer is an application of Bayes' theorem
http://en.wikipedia.org/wiki/Bayes%27_theorem

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}.$$

Part II: Mean, deviations, mean square, mean squared deviation (or variance), RMS (or standard deviation), covariance, correlation, regression.

DEFINITIONS:
Let an overbar indicate averaging over a discrete set y of data values y = {y_i}. Let y' indicate the same set of data values with the mean (which is a single, constant number) subtracted from each value.

$$y = \{y_i\}$$
$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$
$$y' = \{(y_i - \bar{y})\} = \{y'_i\}$$

The variance var(y), and its square root the standard deviation or RMS, are:

$$\mathrm{var}(y) = \overline{y^2} = \frac{1}{N}\sum_{i=1}^{N} (y'_i)^2$$
$$\sigma_y = sd(y) = RMS(y) = \mathrm{var}^{1/2}(y)$$

The covariance cov(x,y) has units of x times y, and is given by:

$$\mathrm{cov}(x,y) = \mathrm{cov}(y,x) = \overline{x'y'} = \overline{y'x'}$$

From cov(x,y) we can get the dimensionless correlation coefficient:

$$cor(x,y) = r_{xy} = \frac{\mathrm{cov}(x,y)}{sd(x)sd(y)} = \frac{\overline{x'y'}}{\sigma_x\sigma_y}$$

Note that both of the above quantites are reversible: cor(x,y) = cor(y,x).

REGRESSION:
Now suppose we try to "fit" data to a straight line, with equation y = mx + b, where m is slope and b is intercept. For clarity here, let's remove the mean of both variables so that we just have y' = mx' + residual. Fitting means minimizing the summed square of error or residual ("least squares")

The optimal value of m in our linear fit y'=mx' is the regression slope or regression coefficient, "regression of y ON x", reg(y,x):

$$m_{best} = reg(y,x) = r_{xy}\frac{\sigma_y}{\sigma_x} = {}^{\overline{x'y'}}\!\Big/_{\sigma_x^{\,2}}$$

Note that reg(y,x) is NOT symmetric in x and y: it has units of y divided by x. This is how I remember its form: it is proportional to the correlation, but has the right units built from the only quantity we have with the units: standard deviation.

This would be called "regressing y on x" (fitting a line y'=mx'). It is not the same as regressing x on y (fitting a line x'=wy'), because you are minimizing different squared errors or residuals, with different units, in the two cases. So w is not equal to 1/m!

If you standardize all variables at the outset, $y_{std} = {}^{(y-\bar{y})}\!/_{\sigma_y}$, $x_{std} = {}^{(x-\bar{x})}\!/_{\sigma_x}$,

then w = 1/m, the regression slope (or regression coefficient) is just the correlation coefficient, and regression is symmetric just like correlation.

Now on to *multiple* linear regression, or generalized linear models.

Here we are postulating that a table of data or matrix **d** (think of it like a clothes closet full of hanging column vectors, but often just one) is to be "predicted" or "explained" by a linear combination of "predictors" **G** (another closet full of column vectors, perhaps several!) **d = Gm** is the postulated relationship (and of course there will be some unexplained residual).

In general the columns of **G** are not orthogonal. This case is the whole point of multiple regression, otherwise you could just fit regression slopes for terms on the RHS one by one.

So multiple regression results are an untidy kind of "explaining" or prediction. One cannot say very cleanly that the coefficients m1, m2, m3… each explain some fraction of the variance in **d**. We will have to transform $\mathbf{G}^T\mathbf{G}$ to eigenspace to get an orthogonal decomposition… but then the eigenvectors (EOFs) will turn out to be hard to interpret, since they are complicated linear combinations of the columns of **G**. We will be able to say how much variance each one explains, but not able to give them clear and meaningful names!

Instead, the values of **m**, coefficients multiplying individual predictors (or inputs), are like partial derivatives *with the other predictors held constant,* so they are called "partial regression" coefficients. As in calculus, we are invited to be careless about stating what was held constant, leading to all sorts of confusion!

For example, suppose we were analyzing precipitation P(t) for trends (a coefficient proportional to t). It matters whether we simultaneously allow this decomposition to have a contribution from ENSO (let's call it N(t)), which itself may have a trend. We should write partial derivatives showing what is held constant in each but rarely do. The total trend dP/dt (which we could estimate with single linear regression on t) has 2 parts in this case: 1. the trend at constant N, and 2. the trend in N times the partial correlation of P with N in the 2-variable regression on **G** = **[**N(t) t**]**.

$$\frac{dP}{dt} = \frac{\partial P}{\partial t}|_N + \frac{\partial P}{\partial N}|_t \frac{dN}{dt}$$

Part 3: Another way to use (and think of) matrices: as "filters" of time series.

$$\begin{bmatrix} g(t) \\ \downarrow \\ \downarrow \\ \downarrow \\ \cdots \end{bmatrix} = \begin{bmatrix} 1 & & & & & \\ & 1 & & & 0 & \\ & & 1 & & & \\ & & & 1 & & \\ & 0 & & & \cdots & \\ & & & & & 1 \end{bmatrix} \begin{bmatrix} f(t) \\ \downarrow \\ \downarrow \\ \downarrow \\ \cdots \end{bmatrix}$$

No filter (the identity matrix)

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

Time derivative. Ends need a special rule.

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

Centered smoother. Ends need a special rule.

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

Causal filter. Only the past affects g(t).

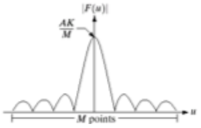
A spike in f would pick out a column of the filter and return it as the column vector result g(t). Results for arbitrary f can be built up as a sequence of such spikes.

I might give you a matrix, and a graph of a time series f(t), and ask you to sketch g(t). Or I give graphs of f(t) and a filtered g(t), and ask you to sketch the matrix M.

Part IV: Fourier analysis

Try to gain a clear understanding of how the Fourier transform maps variance (power) from the physical domain (time or space) to the corresponding spectral domain (frequency or wavenumber).

A time series and its autocorrelation function have the SAME power spectrum. Only the phase spectrum, i.e. the phase of each sinusoidal component, is different. For example, a time series of uncorrelated random values in time has a white power spectrum (equal amplitude of all frequencies, with random phase). Meanwhile the same power spectrum (white), if all the components are phase aligned, gives a localized tall spike (the delta function). The delta function is the *autocovariance function* of random white noise: a white series has perfect correlation at lag 0, and zero correlation at any lag other than zero. Autocovariance is simply the sum of all the Fourier harmonics assigned to the cos() part, i.e. with phase aligned at lag=0.

physical space (time t)	**spectral space (frequency f)**
sin(ft) or cos(ft)	delta function (spike) at frequency f
localized spike δ(t)	white spectrum (flat)
uncorrelated random values	same as above!
wide bump or smoothed spike	red spectrum (reduced high frequencies compared to white noise)
Gaussian	Gaussian
boxcar	sinc() function: 

You should understand how the finite length of a time series appears in the power spectrum (as a lowest possible frequency: one cycle per time series length). You should understand how a finite sampling interval appears (as a highest resolvable frequency, the Nyquist frequency, with one cycle per 2 datapoints – a zigzag).

You should be alert to the existence of *vertical* error bars on spectral peaks in power spectra (from aliasing; or from having too few cycles in the sample so that it might involve the statistics of small numbers, the happenstance of sampling). For example, if you have a decade of hourly data, and find a peak at 5-year period, the peak height (amplitude) is quite uncertain since it comes only from 2 cycles, even though it may be made up of thousands of datapoints. The information content and statistical significance on that may be quite low, despite all the thousands of datapoints.

You should develop opinions about power spectrum plotting conventions. An orthogonal decomposition cries out for a “density function” type graph where area under the curve is the meaningful quantity (variance or power). Cumulative power, culminating in the total variance, is another nice way to plot spectra, since the whiskey spiky quality can make PSD (power spectral density) hard to see. Period is a lot more natural as an x axis than frequency, or (worse yet) radians per second as some purists insist on plotting.