A taxonomy of χ^2

In statistics reference manuals, χ^2 variables appear all over the place, but when we're skimming for a quick answer (e.g. how big should error bars be on spectral energy plots?), we're often left unsure what these χ^2 variables really mean and whether one form of χ^2 has any relationship to another.

Here's a quick inventory of χ^2 usage:

(a) **Least-square fitting:** If you least-squares fit data $\mathbf{y}$ to a function $\hat{\mathbf{y}} = a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + ...$, then you expect that the misfit between $\mathbf{y}$ and $\hat{\mathbf{y}}$ should be consistent with the uncertainties σ in your data $\mathbf{y}$. *Numerical Recipes* describes this succinctly. We can define a measure of the misfit:

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - \hat{y}_i}{\sigma_i} \right)^2 . \tag{1}$$

If we assume that the errors are normally distributed, then χ^2 is a sum of squares of normally distributed quantities, and it has a known distribution. Assuming that our misfit matches our data, χ^2 should equal $N - M$ where M is the number of terms used in the least-squares fit, and $N - M$ is a measure of the total degrees of freedom in the system. In least squares fitting, χ^2 is a measure of the goodness of fit. Formally, the χ^2 statistic tells us if the model fit and data uncertainties are consistent: if χ^2 is very small or very big, that could mean that our estimates of σ are inaccurate or that we're fitting the wrong model to our data. In some desperate cases, people assume χ^2 to be $N - M$ and use that to estimate σ.

(b) **Evaluating probability density functions.** A second common use of χ^2 statistics is to decide whether observed data have a probability density function (pdf) that is consistent with either a theoretical pdf or another observed pdf. In this case,

$$\chi^2 = \sum_{i=1}^{m} \frac{(N_i - n_i)^2}{n_i}, \tag{2}$$

where m is the number of bins, N_i is the number of observations, and n_i is the theoretical number of observations in a bin. The χ^2 probability function tests the probability that the observed and theoretical data in the bin should deviate substantially, and it depends on both the value of χ^2 and the number of degrees of freedom (here the number of bins m or else $m - 1$, depending on the normalization of n_i.)

(c) **Error bars for spectra.** When we compute spectra, we Fourier transform our record x, then square it, average, over multiple realizations, and then seek error bars for this squared averaged quantity. Like the measure of least squares fit, the spectral energy has a χ^2 distribution. The best estimate of our spectrum $\hat{\Gamma}(\omega)$ is $\Gamma(\omega)/2m\chi^2(2m)$, where we have used m data segments. Since this has a χ^2 distribution, we use an error estimate that we discussed previously:

$$P(\chi^2_{2m,1-\alpha/2} < \nu \hat{f}(\omega)/f(\omega) < \chi^2_{2m,\alpha/2}) = 1 - \alpha \tag{3}$$

so if we want to find a 95% significance level, we set α to 0.05.

(d) **Uncertainty estimates for wavelets.** Wavelets are analogous to spectra, except that they aren't produced by summing results from multiple segments, so it's sometimes hard to tell how to obtain usable error bars. Torrence and Campo (BAMS, 1998) suggest using a χ^2 variable with two degrees of freedom to determine significance limits for wavelets. You could imagine using this information to supply an error map for your wavelet power spectrum, but that would be hard to interpret. Instead, a standard procedure is to generate a background red noise auto-regressive process and then seek features that are significantly more energetic than predicted by the red noise process.

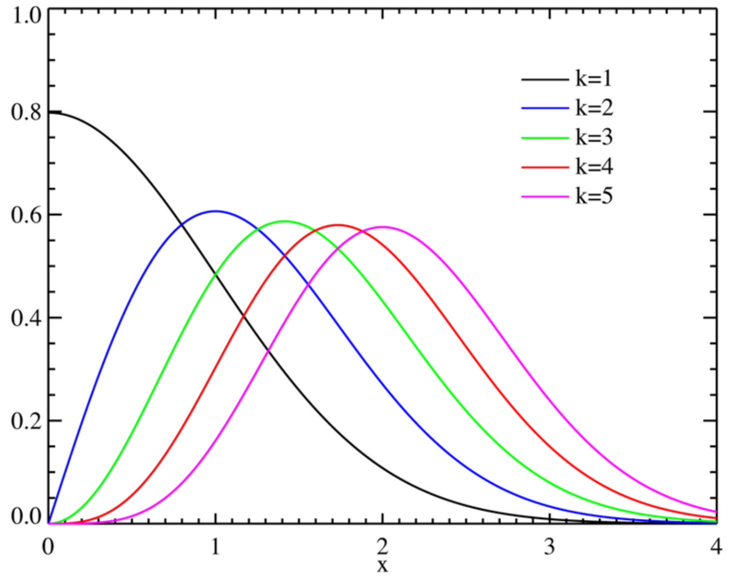

Figure 1: Probability density function of χ^2 distribution as a function of k, from http://en.wikipedia.org/wiki/Chi_distribution.

So what does χ^2 really mean and how do we calculate it?

Simply put, the χ^2 distribution is the probability density function for the squares of summed, normally distributed random variables. Thus, if x_i are Gaussian, then $Q = \sum_{i=1}^{k} x_i^2$ has a χ^2 distribution that depends on k, which is a measure of the number of degrees of freedom. As an example, if you measure normally distributed zonal and meridional wind velocities, u and v, then wind speed squared $(u^2 + v^2)$ will have a χ^2 distribution with two degrees of freedom ($k = 2$). (Wind speed itself, $\sqrt{u^2 + v^2}$, will have a Rayleigh distribution, or equivalently a Weibull distribution with a shape parameter $l = 2$.) If you increase k for the χ^2 distribution, you increase the number of variables that you are summing, and because of the central limit theorem, you expect that for high k, the χ^2 distribution should converge gradually to a progressively more Gaussian-like distribution, as illustrated Figure 1 (taken from Wikipedia), though the distribution is skewed with a long positive tail.

The χ^2 distribution and related terms can be computed relatively easily using the incomplete gamma functions. Usually we need the cumulative distribution in order to obtain the probability that the observed χ^2 value should be as large as it is purely by chance. *Numerical Recipes* has a clear discussion of the formulations, and all of the relevant functions exist in Matlab.