Data, Models & Uncertainty in the Natural Sciences

Problem Set 3

As seen in class, this scaled ratio of the *unbiased sample variance* $\hat{\sigma}^2$ (for a sample size N) of a Gaussian random variable, to the unknown *true population variance* σ^2, is a variable distributed according to the χ^2 distribution (`chi2pdf`) where $N-1$ is the number of *degrees of freedom* (*Bendat & Piersol*, p. 94):

$$(N-1)\frac{\hat{\sigma}^2}{\sigma^2} = \frac{\sum_{i=1}^{N}(x_i - \hat{\mu})^2}{\sigma^2} \sim \chi^2_{N-1}. \tag{1}$$

Here's a **different** example, of *another* a "statistic", X^2, based on *observed* frequencies, f_i, in a histogram of N observations grouped into K bins, and *predicted* frequencies, F_i, from an assumed parent distribution. This particular one is **theorized** to be distributed as χ^2_{K-p} (*Bendat & Piersol*, pp 103–104):

$$X^2 = \sum_{i=1}^{K} \frac{(f_i - F_i)^2}{F_i} \sim \chi^2_{K-p}, \tag{2}$$

where the number of degrees of freedom is the number of frequency bins, K, reduced by p, the number of independent linear constraints imposed on the data. **For example:** for a frequency histogram for which you know that $\sum_i^K f_i = N$, tested against a two-parameter distribution such as the Gaussian, $p = 3$, and the number of degrees of freedom is $K-3$. For eq. (2) to hold, the predicted frequencies must be "large", i.e. $F_i \gg 1$.

Eq. (2) is the basis of the *Pearson* χ^2 *test*: we *accept* the hypothesis that the samples are drawn from the tested parent distribution if the **probability of exceeding the observed value** X^2 is *greater* than some α (e.g., 0.05), which is called the *level of significance* (as in 5%, which yields "95% confidence"). In this **one-sided** test, observing *this much "deviation" — or more* has to be rather *unlikely* for us to *reject* the hypothesis.

1. Draw sets of N numbers from a probability distribution of your choice. You might want to use `random`. For each set, make a histogram (use `hist` to calculate the f_i, maybe `bar` to plot them), perform a one-sided χ^2 test as described above (using `cdf` to calculate F_i), for them being drawn from their respective *known* distributions. You be the "oracle" — you know everything. Is your test "working" for you?

2. Draw sets of N numbers from an arbitrary probability distribution. For each set, make a histogram, perform a one-sided χ^2 test as above, but now for them being drawn from the *wrong* distribution (wrong family, and/or wrong parameters). Report on how your choices influence the behavior.

Is the above procedure a good test? Is it easy to tweak the parameters (N, K, bin spacings, α, the types and parameters of your test distributions, etc.) such that you reliably accept the right model, and reject the wrong model, when you know the truth? Do 3 and 4 a few times, e.g. M times, make a table of your parameter choices, and list the proportion of right/wrong rejections/acceptances for your choices (including α).

I note that, since the expected value of a χ^2_n variable with n degrees of freedom is equal to n, often the *reduced* χ^2 distribution is used, χ^2_n/n, which has an expectation of 1. We will deal with this later when we evaluate the goodness-of-fit of solutions to linear systems.