PHSX 315 Assignment 4

Due by Wed. Feb. 22 at 1 PM

Learning Goals: Standard deviations. Chi-squared and tests of simple hypotheses. Scientific method in action. The truth is out there!

Problem 1 Use the PValues calculator in ClassExamples to calculate (and report) the two-tailed probability for a normally distributed random number to be more than 1.0 σ, 2.0 σ, 3.0 σ, 4.0 σ, 5.0 σ, and 6.0 σ from the mean. These can be conveniently calculated from the tail probability for the chi-squared distribution with one degree-of-freedom (after realizing that the chi-squared distribution is essentially the squared version of the standard normal distribution). These should lead to the usual conclusions that about 68% of measurements should be within one standard deviation, and that about 95% of measurements should be within two standard deviations.

Problem 2

Medieval natural philosophy student, Bill Ockham, is planning to make measurements of Earth's acceleration due to gravity, g, at various heights, h, above Earth's surface. Before starting his measurements, he knows that the local acceleration due to gravity at Earth's surface is precisely known to be 9.800 m/s^2.

He therefore thinks that the most straightforward model for the results of his experiment would be that the acceleration due to gravity should not vary with height, and that the true value of g is 9.800 m/s^2 independent of height. Formally, he specifies an *a priori* hypothesis that $g = 9.800$ m/s^2 for all heights, h where $h \geq 0$. This is his **null hypothesis**. He plans to make 20 measurements at different heights each with measurement uncertainties of 0.05 m/s^2. He has learned in class that the chi-squared χ^2 quantity is a good way to test "goodness-of-fit" and to test hypotheses. With 20 measurements, and a completely specified hypothesis (g is assumed constant at the specified value - so there are no free parameters), there are how many degrees of freedom? In this case there are 20 degrees of freedom (the number of measuremements (20) minus the number of free parameters (0)). A completely specified hypothesis is called a "**simple hypothesis**". Having learned about formal hypothesis testing, and unwilling to accept the standard 5% risk of a false positive (Type I error), he decides to use a **significance level** of 1% in defining the **critical region** of the χ^2_{20} statistic which would be used to reject his null hypothesis. What value of χ^2_{20} corresponds to this 1% significance level?

Included in the file experiment1.dat are the results of his first experimental measurements. The file contains columns with the measurement number, i, height (m), measured g_i (m/s^2),

uncertainty on g_i (m/s^2). The third column is the known statistical uncertainty on the individual measurements, $\sigma(g_i)$. It is assumed that the actual measurements are normally distributed about their true expected value with an underlying standard deviation given by the assigned uncertainty. (Do not confuse true expected value with the expected value for the model under test!).

You should construct the value of the the chi-squared statistic for this data-set,

$$\chi^2_{20} = \sum_{i=1}^{20} \left(\frac{g_i - 9.800}{\sigma(g_i)} \right)^2$$

and report the observed value.

Does the value fall in the critical tail region? If yes, then formally the null hypothesis can be rejected with 99% confidence level, and one has reasonably significant evidence that the null hypothesis is incorrect (ie. false) and one needs an alternative hypothesis (ie. model of nature) to explain the data. If the chi-squared value is less than the critical value, then the test has not yielded a result that is significant at the pre-determined significance level, but it is still useful to report the actual level of significance often denoted the ***p*-value**. This reports the fraction of repetitions of the experiment that would yield a value of χ^2_{20} (the current test-statistic) greater than or equal to the observed value of χ^2_{20} occuring by chance when the null hypothesis is true. If this value is small, but not smaller than 1%, it still suggests that the null hypothesis may in fact be incorrect, but the current experiment may lack sufficient precision to test it sufficiently.

Problem 3

After a few years, Bill has been working on upgrading his experiment, making it better able to measure g, and developing a new technology, (balloon), that allows measurements at larger heights. The data from the new experiment are in experiment2.dat. Again he wants to test the same null hypothesis as in Problem 2. The new experiment has 30 data points so one should recalculate and report the critical region for a chi-squared test at a 1% significance level using the appropriate number of degrees of freedom for this experiment. What is the new value of the test statistic for this experiment. Is it significant at the 1% level? What is the p-value? So should Bill abandon his null hypothesis?

Problem 4

A few years later Bill has figured out some stuff about gravitation and decided to re-analyze his experiment2 data under the new hypothesis that the gravitational acceleration is $g = g_0 = 9.800$ m/s^2 at $h = 0$ corresponding to a distance from the center of the Earth of $R_E = 6378$ km, but that the acceleration due to gravity varies with distance from the center of the Earth in the manner expected from Newton's law of gravitation (depending on $r = R_E + h$). You should be able to express this new model for $g(h) = g_0\, f(h, R_E)$ by absorbing factors of G and M into g_0. So again a simple hypothesis with all parameters specified. You should construct a chi-squared comparing the experiment 2 data with this new model. Does this model fit the data? Is there significant evidence to reject this model

at the 1% level? What is the p-value? Remember that just because data are consistent with a model at a reasonable confidence level does not imply that the model is true.

Finally check again with the experiment1 data whether those data are consistent with your new hypothesis.

Problem 5

(This may take some time for some of you - please submit problems 1–4 by the due date). The concept of test **power** relates to the ability to distinguish hypotheses, and needs both a null hypothesis and a well defined **alternative hypothesis**. We now have two hypotheses, Ockham's original naive one, which we'll call the null hypothesis, and the alternative one based on Newton's universal law of gravitation. We also have two different experiments, differing in quality, number of data-points, and data range. An important question is to assess how well each experiment will correctly lead to the rejection of the null hypothesis when it is indeed false. This is quantified using the test power, the probability of a true positive. Namely in what fraction of repeated experiments where the data are drawn from the model defined by the alternative hypothesis is the result of the experiment significant (where one correctly rejects the null hypothesis at the specified significance level). To calculate this, one way is to set up a set of Monte Carlo toy experiments that replicate the expected distribution of outcomes (when the alternative hypothesis is true) and count the fraction of times that the null hypothesis is rejected. The underlying statistical distribution of the number of successful rejections is again the Binomial distribution. Thus we know that the statistical uncertainty on our measurement of the fraction of rejections is

$$\sigma_p = \sqrt{\frac{p(1-p)}{N_{\text{expts}}}} \quad .$$

When you compare the powers of the two experimental tests of the null hypothesis you should find that the power of the second type of experiment is much greater than the power of the first type of experiment. In practice the test power is often not calculable a priori because the alternative hypothesis has not yet been formulated, or more complicated to calculate because the alternative hypothesis is not a simple hypothesis where all parameters are specified (no free parameters and so no wiggle room).

You should upload to your repository a summary of your findings, a copy of your code, example results from running the code, and appropriate figures. In these cases plots of the data with error bars and super-imposed curves corresponding to the models would be appropriate.