SELECTING BETWEEN TWO HYPOTHESES

Louis Lyons
Dept of Physics
Oxford, UK

August 1999

Abstract

We often need to compare data with two different hypotheses to see if either is significantly preferred. The classical statistical approach is to reject one (or both) theories only if its χ^2 probability is unsatisfactorily small. It is pointed out that it may be possible to use a simple modification, which is more efficient in rejecting the wrong hypothesis.

OUNP–99–12

The problem

Assume we have an experimental distribution that we want to use to test two different hypotheses, whose predictions for this experimental distribution are well-defined. For each hypothesis we can construct a χ^2 that tests the consistency of the data with that hypothesis. In order to decide whether a hypothesis can be rejected, it is commonly believed that only the χ^2 probability for that hypothesis is relevant. It is the purpose of this note to show that there is often a test based on the χ^2 difference that is more powerful.

Example: Mean of a set of observations

As a simple example, we consider the case of a set of N measurements of the same quantity, all with the same uncertainty, σ. To find the best value and its error, we first construct

$$S(X) = \sum (x_i - X)^2/\sigma^2 \tag{1}$$

The best value X is given by minimising S with respect to changes in X, while the error on X is determined by finding the change in X which results in S being 1.0 larger than the minimum S_{min}. This gives

$$X = \sum x_i/N \tag{2}$$

$$\sigma(X) = \sigma/\sqrt{N}. \tag{3}$$

Alternatively $\sigma(X)$ can be derived from eqn.(2), simply by propagating through the errors on x_i. This gives the identical answer for $\sigma(X)$.

Paradox?

So if we have 100 different measurements x_i, the average X could correspond to a value of S_{min} of, say, 90. Then the error on X is determined by varying X until the value of S in eqn.(1) increases by 1 unit to 91. i.e. the acceptable values (at the 1 standard deviation level) of X have S values in the range 90 to 91. This is despite the fact that if we chose a much more distant value of X which gave, say, $S = 115$, we would not in general have regarded that as unreasonable for a data set with 99 degrees of freedom. Thus we almost have a paradox:

(A) The value of X which gives $S = 115$ has, according to χ^2 tables, a reasonable probability; but

(B) The 'acceptable' range of X corresponds to S of 90 to 91, and so an X which gives S^2 of 115 is ruled out at the 5σ level.

Resolution of paradox

The resolution of the paradox is as follows. Although in general an S value of 115 is acceptable, in this particular case where we are varying just a single parameter and the data is such that the minimum S is 90, it is not.

We can make this argument more convincing by writing the identity

$$\begin{aligned} S(X) &= \sum (x_i - X)^2/\sigma^2 \\ &= \sum (x_i - \bar{x})^2/\sigma^2 + N(\bar{x} - X)^2/\sigma^2 \end{aligned} \tag{4}$$

This shows that $S(X)$ can be regarded as being made up of two terms. The first is the scatter of the individual x_i about their mean $\bar{x}$. It should have a χ^2 distribution with 99 degrees of freedom. It is independent of the value of X. The second term is the one which does depend on X. It minimises at zero for $X = \bar{x}$. It increases by 1 when

$$X = \bar{x} \pm \sigma/\sqrt{N}$$

This helps resolve the paradox. Our $S_{\rm min}$ value is simply a test of how well the different values of x_i are consistent with each other. This is what determines whether $S_{\rm min}$ is 90, 110 or some vastly different value. But for a given data set, it has a particular value. Then the accuracy with which we know X is determined just by the second term on the right hand side of the identity (4). It is obtained by allowing an increase of only 1.0 in this term, and hence in $S(X)$. This explains why, if $S_{\rm min}$ is 90 for a given data set, then only X values that give S up to 91 are allowed (at the 1 σ level), even though for other data sets an S value of 115 could be acceptable.

I have not produced such a mathematical demonstration for other one-parameter fits, but tend to believe that the same principle applies.

Two separate hypotheses

Now let us consider comparing 2 different hypotheses, that are not simply different values of the same parameter. Thus we could be comparing the distribution in jet energy obtained at a $p\bar{p}$ collider, with the predictions from QCD, based on 2 different sets of proton structure functions. This sounds a different problem from the above. However, it can be recast as a parameter problem by rewriting the predicted distribution as

$$\text{param} \times (\text{Theory 1}) + (\text{1-param}) \times (\text{Theory 2})$$

Thus

$$\begin{aligned} &\text{param} = 1 \text{ is equivalent to Theory 1} \\ \text{and} \quad &\text{param} = 0 \text{ is equivalent to Theory 2,} \end{aligned}$$

so it is not so different from the 'different values of a single parameter' considered above.

Toy example

In order to check out whether the χ^2 difference really does apply to the 'Two Hypotheses' (at least for one specific example), I have performed the following Monte Carlo test:

I generate histograms for some imagined variable x in 100 bins over the range -1 to $+1$, according to one of 2 possible 'hypotheses':

$$\begin{aligned} H1: &\quad 100 * (1 + 0.05x) \text{ events per bin} \\ H2: &\quad 100 * (1 + 0.05\cos(\pi x)) \text{ events per bin} \end{aligned}$$

Then each bin is allowed statistical fluctuations with $\sigma = \sqrt{n}$.

Next the distribution for $H1$ is compared with the theoretical $H1$ and $H2$ (i.e without statistical fluctuations), and a χ^2 value is calculated for each (χ_1^2 and χ_2^2 respectively). This process is repeated for 500 separately generated histograms according to $H1$. What is observed is that:

a) χ_1^2 is distributed as expected. It is centred on 100 and its width is consistent with the expected $\sqrt{200}$. By imposing a cut at 130, we have an efficiency of 97% for accepting the correct hypothesis.

b) χ_2^2 tends to be larger than χ_1^2, because the 'data' is being compared with the wrong theory. However, since neither $H1$ nor $H2$ differs very much from a constant value of 100 events/bin, even the wrong theory is not in much worse agreement with the 'data' than the correct theory. The χ_2^2 distribution is centred on 110, and has a similar width to χ_1^2.

A cut of 130 would result in the wrong hypothesis being regarded as acceptable in 69% of the examples. i.e. Accepting a hypotheses simply on the basis of its χ^2 results in our often being unable to reject the wrong hypothesis.

c) $\chi^2_2 - \chi^2_1$ has a distribution centered on +10 with a σ of around 5. Only 8/500 cases have χ^2_2 less than χ^2_1. This suggests that the χ^2 difference is more useful for separating the hypotheses.

For completeness, 'data' was also generated according to $H2$. This time I observe:

a) χ^2_1 is now centred on 110, because $H1$ is the wrong hypothesis.

b) χ^2_2 is as expected. i.e. centred on 100

c) $\chi^2_2 - \chi^2_1$ has a distribution centred on -10. Only 3/500 cases give a positive value.

This confirms that the χ^2 difference is, for this case, a more powerful method for discriminating between hypotheses. As usual, the method is tunable. i.e.
We can select the hypothesis in every case simply according to which χ^2 is smaller. This results in a small number of incorrect decisions.
OR
We can decide to made decisions only when the magnitude of the χ^2 difference is larger than some number (e.g. 2.0 for our problem). This reduces the fraction of wrong decisions, at the expense of not always making a decision.

Actual selection procedure

There are two basic possibilities for using this approach to try to discriminate between hypotheses:

a) In the classical approach, for each hypothesis $\Delta\chi^2$ is simply a statistic which can be used. Thus if the observed $\Delta\chi^2$ were to be such that it lay very much in the tail of the predicted distribution, then that hypothesis would be rejected. Essentially each hypothesis is tested separately, and the relevant feature of the simulated distribution is the probabability of obtaining a value of $\Delta\chi^2$ equal to the observed one, or more extreme.

b) For a Bayesian, the relative probabilities of the two hypotheses, as determined from the data, are simply given by the relative heights of the two distributions (such as those of the figure), at a value of $\Delta\chi^2$ equal to the experimental one. This is just the likelihood ratio for the two hypotheses, and hence is closely related to the actual value of $\Delta\chi^2$, at least in the limit of large statistics.

As usual for a Bayesian approach, this probability ratio could be multiplied by any a priori probabilities for the two hypotheses. In the absence of any prior knowledge, the two hypotheses could be given equal weight. (Because the hypotheses are distinct, this procedure is less problematic than for the case of discriminating among different values of a continuous parameter e.g. the mass of the electron-neutrino.) In any case, the ratio of the heights of the two curves gives the relative probabilities of the two hypotheses, as determined by the particular experiment alone.

For either a) or b) above, it is a good idea also to check the relevant χ^2_1 or χ^2_2 before deciding to accept a hypothesis. i.e. it is necessary that, as well as the χ^2 difference favouring the hypothesis, its individual χ^2 value must also be satisfactory.

Possible Application to Neutrino Oscillations

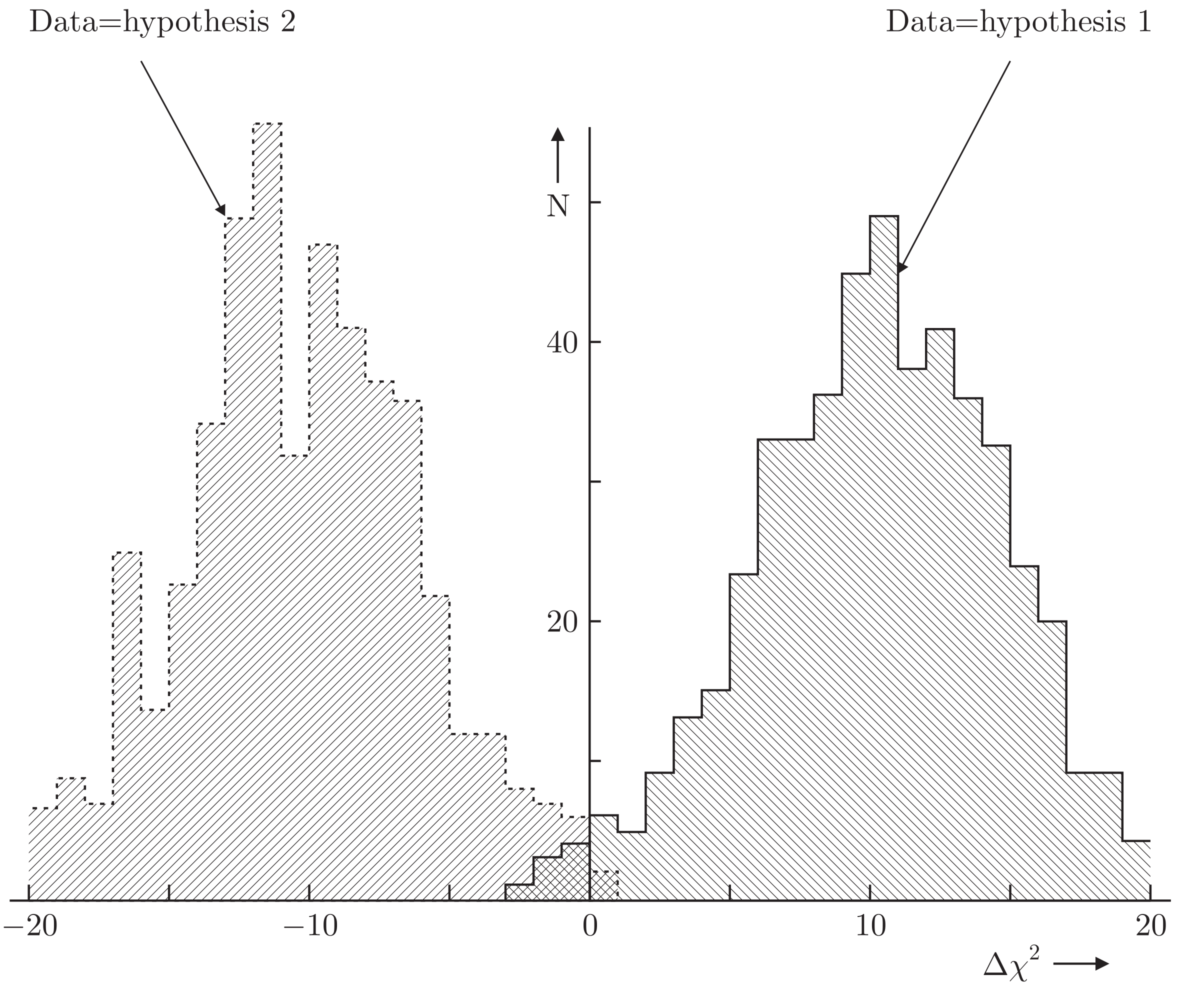
In analysing the rates of observed events in solar neutrino experiments with the various detectors (chlorine, hydrogen and gallium), different solutions are currently possible: either vacuum oscillations, or MSW 'small mixing angle' and 'large mixing angle' solutions. A possible way to distinguish these could be to look at the distribution of the number of events in some variable. e.g. electron energy or day-time/night-time in the hydrogen experiment, or as a function of the earth-sun distance. The traditional way of trying to use these distributions is to look at the χ^2 probabilities for the competing hypotheses. The extension of our earlier examples is that it is worth while doing Monte Carlo studies to test whether a χ^2 difference method might give better sensivity to discriminating these hypotheses.

Conclusion

The χ^2 difference may be a good way to distinguish between alternative theories. This suggestion is not too radical as the χ^2 difference is very closely related to the log-likelihood ratio, which is also used in several methods to select among competing hypotheses.

The performance of this approach can relatively easily be investigated for any particular case by performing Monte Carlo simulations.

I wish to thank Peter Clifford, Bill Murray and Byron Roe for useful conversations.



The distribution of $\Delta\chi^2 = \chi_2^2 - \chi_1^2$, the difference in χ^2 values. Here χ_i^2 is the value of χ^2 for the comparison of the generated distribution with the expectation for hypothesis i; hypothesis 1 is that the distribution is $1 + 0.05x$, while for hypothesis 2 it is $1 + 0.05\cos(\pi x)$. The solid histogram is for 500 different examples of an 'experiment' generated according to hypothesis 1, which is why $\Delta\chi^2$ tends to be positive, while the dashed histogram is for 'experiments' generated according to hypothesis 2. The statistic $\Delta\chi^2$ is better than the individual χ_1^2 and χ_2^2 for discriminating between the hypotheses.