χ^2 Fitting When Overall Normalization is a Fit Parameter

Byron Roe
Department of Physics
University of Michigan
Ann Arbor, MI 48109

Problems in the use of the χ^2 method to fit event histograms, when the normalization is one of the fit parameters appears every few years in experimental studies and has caused some confusion. It appeared in our MiniBooNE experiment in 2015. In Great Britain it was known as Peelle's Pertinent Puzzle (PPP) and this puzzle was also found in 1996 by an Italian physicist who has made significant contributions to the statistics used by physicists.

The puzzle is that in a χ^2 fit, if overall normalization is one of the parameters to be fit, the fitted curve may be seriously high with respect to the data points, sometimes above all of them.

This problem and the solution for it are well known within the statistics community, but, apparently, not well known among some of the physics community. The purpose of this talk is didactic, to explain the cause of the problem and the easy and elegant solution.

The solution is to use maximum likelihood (ML) instead of χ^2. The essential difference between the two approaches is that ML uses the normalization of each term in the χ^2 assuming it is a normal distribution $1/\sqrt{2\pi\sigma^2}$.
The normalization is applied to the theoretical expectation, not to the data so $\sigma^2 = N$ and the normalization changes as the expectation value changes.

We illustrate what goes wrong and and how maximum likelihood fixes the problem in a very simple toy example which illustrates the problem clearly and and is the appropriate physics model for event histograms. We then note how a simple modification to the χ^2 method gives a result almost identical to the ML method.

Consider a simple data set with only two bins. Suppose theory predicts that the expected value for the number of events in the bin 'N' is the same for each bin and that the bins are uncorrelated. Let x_1 and x_2 be the number of events found experimentally in each bin. The variance (σ^2) is N for each bin, $\sigma = \sqrt{N}$

$$\chi^2 = \frac{(N - x_1)^2}{\sigma^2} + \frac{(N - x_2)^2}{\sigma^2}.$$

We want to find the minimum,

$$\chi^2 = \frac{(N - x_1)^2}{\sigma^2} + \frac{(N - x_2)^2}{\sigma^2}.$$

Call Term 1 the derivative of the χ^2 with respect to the numerators of the χ^2. Term 1 =

$$2\frac{(N - x_1 + N - x_2)}{N} = 2(1 - \frac{x_1}{N}) + 2(1 - \frac{x_2}{N}).$$

If we ignore the derivative of the denominator, then Term 1 =0, is solved by

$$N = (x_1 + x_2)/2$$

Call this the naive solution

Call Term 2 the derivative with respect to the denominator of the χ^2:

$$\text{Term } 2 = -\frac{(N - x_1)^2 + (N - x_2)^2}{N^2}.$$

Term 2 is negative and of the order of 1. The only way that Term 1 + Term 2 =0 is for Term 1 to be positive. This means that the χ^2 solution must have N greater than the naive value. Although Term 1 is O(1), x_1/N and x_2/N are O(1/N). N is pulled up as the fit wants to make the fractional errors larger.

Suppose, incorrectly, one had used as a variance the number of data events (x_i) rather than the expected number of events (N). Had the normalization been put into the data, then:

$$\chi^2 = \frac{(N - x_1)^2}{x_1} + \frac{(N - x_2)^2}{x_2}$$

Upon taking derivatives then instead of the desired $N = .5(x_1 + x_2)$ one gets the averages of the inverses:

$$\frac{2}{N} = \frac{1}{x_1} + \frac{1}{x_2}$$

and the fitted curve is low

Next use maximum likelihood for our toy model. The likelihood (L) is the probability density function for the two bins assuming each bin has a normal distribution. (This requires that N is not too small.)

$$L = \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(N-x_1)^2/(2\sigma^2)} e^{-(N-x_2)^2/(2\sigma^2)} .$$

$$L = \frac{1}{2\pi\sigma^2} e^{-\chi^2/2}$$

For $\sigma^2 = N$, the log of the likelihood is:

$$\ln L = -\ln(2\pi) - \ln N - \chi^2/2.$$

Let Term 3 be the derivative of the normalization.

Term 3 = $-1/N$.

The derivative of ln L is

Term 3 - (Term 1)/2 — (Term 2)/2.

Term 3 - (Term 2)/2 =

$$-\frac{1}{N}+\frac{(N-x_1)^2+(N-x_2)^2}{2N^2}$$

$$=\frac{-2N+(N-x_1)^2+(N-x_2)^2}{2N^2}.$$

Since the expectation value of
$E(N - x_1)^2 = E(N - x_2)^2 = N$,
the expectation value of Term 3 —(Term 2)/2 =0.

Assume now that there are n_b bins and there are a total of n_f parameters to be fit including an overall normalization. The expectation value for the minimum χ^2 is $n_b - n_f$. This occurs because, after fitting, the multidimensional normal distribution loses n_f variables.

There is a simple general way to handle this.
Consider n_b bins and n_f fitting parameters, p_j.
Let the expected number of events in bin i be

$$n_i(p_1, p_2, \cdots, p_{n_f})$$

The distribution of experimental events in each bin is taken as approximately normal.
The total number of events in each bin is not fixed.
Choose the set n_i as the basis. The error matrix is diagonal in this basis. We then use the ML method.

Ignoring the 2π constants:

$$\ln L = \sum_{i=1}^{n_b} -\frac{\ln n_i}{2} - \frac{(x_i - n_i)^2}{2n_i} .$$

$$\frac{d \ln L}{dn_i} = \frac{x_i - n_i}{n_i} + \frac{1}{2n_i}\left[\left(\frac{(x_i - n_i)^2}{n_i}\right) - 1\right] .$$

The expectation value for the term in square brackets is zero.

Recall that the expectation refers to the average value over a number of repetitions of the experiment. It is x_i that changes with each experiment not the theoretical expectation, n_i. The expectation value of the term in square brackets will remain zero even if it is multiplied by a complicated function of the p_j fitting parameters.

Ignoring this term leads to:

$$\frac{\partial \ln L}{\partial p_j} = \sum_{i=1}^{n_b} \left(\frac{x_i - n_i}{n_i}\right) \frac{\partial n_i}{\partial p_j}.$$

By expressing the n_i as appropriate functions of the $p_{j'}$ the error matrix can be written in terms of the p_j. However the derivative of the inverse error matrix does not appear in the transform of the above equation.

This result means that one can use a modified χ^2 approach. Use the usual χ^2, but, when derivatives are taken to find the χ^2 minimum, omit the derivatives of the inverse error matrix. The result is almost identical to the result from ML. The modified χ^2 method should be generally used in place of the regular χ^2 method.

In practice, since the differences are not precisely the expectation values for a given experiment, there is a small residual higher order effect, which causes no bias on the average.

Practical Considerations

If you are using one of the standard minimization routines
such as the CERN MINUIT routine then:

DO NOT MINIMIZE CHI-SQUARE!
It will effectively, take the derivative of the denominator.
INSTEAD:
Minimize minus the log of the likelihood which is just as easy
You will then get the correct fitted parameters
HOWEVER:
The errors will not yet be right.

IN ODER TO GET THE ERRORS CORRECT:

The exponential term
in the maximum likelihood is not χ^2/σ^2 ,
but is $\chi^2/(2\sigma^2)$.
In MINUIT the default parameter is set to 1 to give the correct value to go above the minimum for one standard deviation for χ^2.
You need to set ERRORDEF to 0.5
and then you will have the correct errors.

Summary

The problem with using the χ^2 method when one of the parameters is the normalization of the curve is solved by using the maximum likelihood method

A good approximation is to use the χ^2 method, but not take the derivative of the inverse error matrix.

However, if you use a standard minimization routine such as the CERN MINUIT program, you need to minimize, not χ^2 but minus the log of the likelihood.

You then also need to set the error parameter from the value for χ^2 to 1/2 of that value.

If there's time:
New topic: Feldman-Cousin's method

In the last few minutes I will turn to a different topic. Many of you are familiar with the Feldman-Cousins method for constructing confidence belts:
Physical Review Vol.57 Number 7 (1998).
They apply it to a Poisson problem with expected background b. For a possible signal μ and a result x, they investigate the ratio $R = P(\mu\,|\,x)/P(\mu_{best}\,|\,x)$, where "best" means the μ giving the highest probability for that value of x. Next they find a region in x by picking the highest $R's$ until P=90% is reached. Having done this for a series of $\mu's$ they have a confidence belt when this is looked at as a function of x.

They note that if the experimental result is $x < b$, then one should also look at the "sensitivity", the average upper limit which would be obtained by an ensemble of experiments with background b and no true signal. This can indicate a problem qualitatively, but no quantitative suggestion.

They were careful to insist that one only use information before any fit is performed. For example if you make a decision after looking at the data whether to quote upper limits only or to quote upper and lower limits this would cause a bias which they labelled “flip-flopping”.

However, they ignored the fact that one can make some decisions after looking at the data without bias. For example, if there are n events then the background for that set of data cannot be greater than n. Their results are then true compared to the total universe of all possible results, but misleading if extra information is available reducing the universe of results.

As luck would have it, one of the early experiments to which it was applied was the Karmen experiment looking for an anomalous neutrino oscillation appearance of ν_e events which had been reported by the LSND experiment. Karmen expected b=3 and observed x=0. Feldman and Cousins listed a 90% CL as 1.08 events and sensitivity of 4.42. 1.08 was in serious disagreement with the LSND result.

However, for this particular result, if there were zero events observed, then there was zero signal and zero background. From zero signal, all one can conclude is the 90% CL is 2.4. The background is irrelevant.

Together with a statistician M. Woodroofe we used a conditional probability $P(x \mid b \leq x)$ which worked very well for a Poisson discrete problem. However, for the second FC problem, measuring x in a continuous distribution given a positive signal θ with gaussian error Δ, ($x = \theta + \Delta$) our method gave a positive lower limit for all positive observations, as pointed out to us by R. Cousins.

We then suggested a Bayesian approach using credible intervals which worked well for both the Poisson and the continuous problem.

Byron P. Roe and Michael W. Woodroofe, Phys. Rev. D63, 013009 (2000).

For the Bayesian solution, the prior was taken as the interval from 0 to infinity for the signal and then the credible region was arranged to be the shortest possible region for each possible signal μ. A slight modification was made at the end to produce slightly better results. The slides on the next pages show the results before and after the modifications.

On the slides:
RW old is the conditional probability result and
RW2 and RW2 modified are the Bayes results
without and with the final modification which brings
the credible limits closer to the CL for x>b.

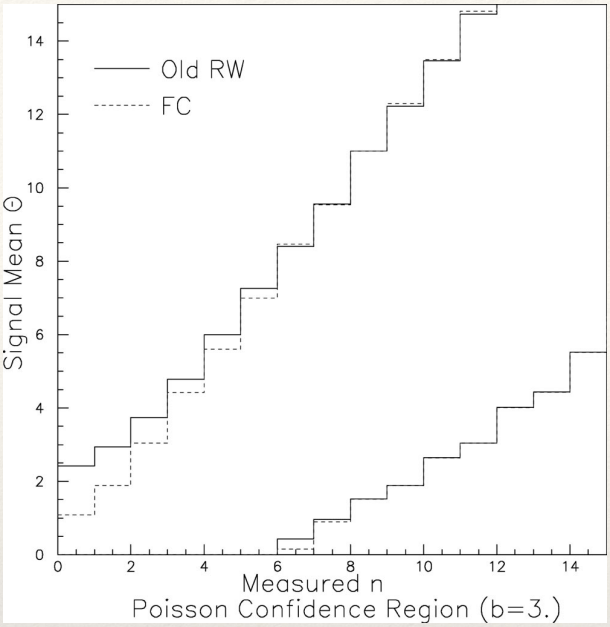

FIG. 1. The 90% C.L. belt for a Poisson probability with $b = 3$, using the old RW procedure (solid line) and the Feldman-Cousins unified procedure (dashed line).

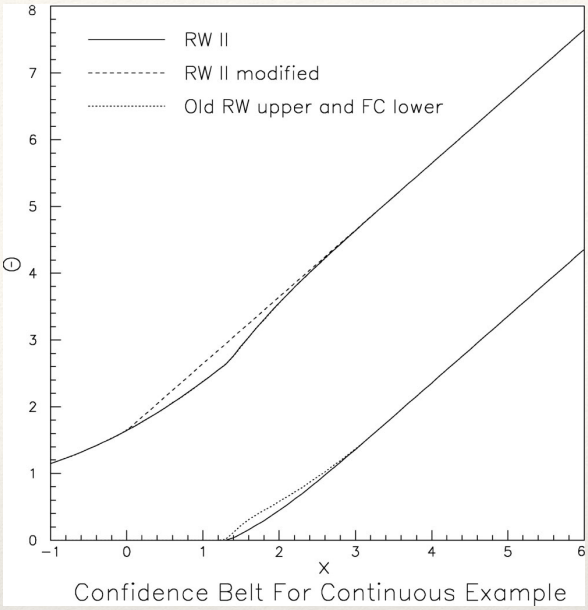

FIG. 4. The 90% C.L. belt using the Bayesian procedure (solid line) and the conservative modification (dashed line) for the continuous example. The old RW upper limit and the F-C unified method lower limit are shown as dotted lines.

Old RW refers to the conditional probability used for the Poisson example. RW and RW2 refer to the Bayesian result for the continuous example without and with the ad hoc addition shown. For RW2, and a confidence limit of 90%, the RW2 credible limit is at least 90% to 3 decimal places. For RW the credible limit is >86%.

Conclusion

The Feldman Cousins method of selecting confidence belts is quite useful, obtaining compact belts and automatically going from one-sided to two-sided limits.
However, it has a problem when faced with experiments obtaining values less than estimated backgrounds.
For the Poisson distribution, the conditional probability $P(\mu \mid b < n)$ solves the problem. For the continuous example a Bayesian analysis is used.