Building Central Confidence Intervals with the Profile Construction Method

John Kelley

June 1, 2008

1 Introduction

In this document we describe a frequentist method of defining central confidence intervals which incorporate systematic errors. This method, the *profile construction* method, is an extension by G. Feldman of the frequentist approach described in his paper with R. Cousins [1]. Originally described (albeit very tersely) in [2], it has only recently been applied to physics analyses.

2 The Canonical Frequentist Approach

We review the canonical frequentist approach to the construction of central confidence intervals, as described in [1].

2.1 Likelihood Ratio

We first define a test statistic to compare our observables x for various hypotheses, characterized by physics parameters θ_r. For a binned distribution, a natural choice arises from the Poisson probability (or likelihood)

$$P(x|\theta_r) = \prod_{i=1}^{N} e^{-\mu_i} \frac{\mu_i^{n_i}}{n_i!} \ , \tag{1}$$

where we form a product over the N bins of our observable(s) x, and in each bin of the data we see n_i counts on an expected μ_i for the hypothesis we are testing with parameters θ_r. At this point it is conventional to switch to the negative logarithm (the *log likelihood*):

$$\mathcal{L}(\theta_r) = -2 \ln P = 2 \sum_{i=1}^{N} \left(\mu_i - n_i \ln \mu_i + \ln n_i! \right) \ . \tag{2}$$

We will come back to the additional factor of 2.

To compare the probabilties of two hypotheses H_1 and H_2 of generating our observed data, we take the *likelihood ratio* (or, working with the logarithm, the difference):

$$\mathcal{L}(\theta_{r1}) - \mathcal{L}(\theta_{r2}) = 2\sum_{i=1}^{N}\left(\mu_{1,i} - \mu_{2,i} + n_i \ln\frac{\mu_{i,2}}{\mu_{i,1}}\right) \ . \tag{3}$$

where hypothesis H_1 with parameters θ_{r1} gives us an expected count $\mu_{1,i}$, and hypothesis H_2 with parameters θ_{r2} gives us an expected count $\mu_{2,i}$, and again we have observed n_i counts in a given bin. Using this, our test statistic compares the hypothesis at a point θ_r to the hypothesis that fits the data the best. Specifically, in the physics parameter space θ_r, the test statistic is the difference of the log likelihood at this point to the best-fit hypothesis with parameters $\hat{\theta}_r$ ($\mathcal{L}$ is minimized[1] by $\hat{\theta}_r$):

$$\Delta\mathcal{L}(\theta_r) = \mathcal{L}(\theta_r) - \mathcal{L}(\hat{\theta}_r) \ . \tag{4}$$

The additional factor of 2 added in equation 2 arises because in the Gaussian regime, $\Delta\mathcal{L}$ so defined approaches a χ^2 distribution with degrees of freedom equal to the dimension of θ_r (Wilks' Theorem).

2.2 Confidence Intervals

At this point, we wish to examine all the physically allowed hypotheses by iterating over the space θ_r, and determine which are allowed given our observation x. It is not uncommon at this point to use Wilks' Theorem and define confidence intervals using a χ^2 distribution. Specifically, one calculates $\Delta\mathcal{L}$ at every point θ_r, and for a given confidence level (CL) α, the allowed region is the set

$$\{\theta_r\}_\alpha = \{\theta_r \mid \Delta\mathcal{L}(\theta_r) < \chi^2(\alpha, \dim\theta_r)\} \ . \tag{5}$$

For two parameters and a 90% confidence level, we would allow the region where $\Delta\mathcal{L} < 4.61$. This is known as the *global scan* method.

As demonstrated in [1], the global scan method has several disadvantages when the likelihood varies in a complicated way over the parameter space. In particular, $\Delta\mathcal{L}$ can deviate from the simple χ^2 distribution by a significant amount if, for example, one of the parameters is extended into a region which has little effect on the observables. In this case, the effective dimensionality of θ_r is reduced and the χ^2 used has too many degrees of freedom. In this case we prefer a frequentist approach to define the confidence intervals which takes this and other issues into account to achieve proper coverage.

[1] By minimizing the negative log likelihood, we maximize the probability.

Specifically, at *each point* in the parameter space θ_r, we perform a number of Monte Carlo experiments where we sample from the parent distribution $\{x \mid \theta_r\}$ and then calculate the likelihood ratio $\Delta\mathcal{L}_i$ for the experiment. This set of $\{\Delta\mathcal{L}_i\}$ allows us to see how our test statistic behaves under statisical variations only. Then to define our confidence intervals at CL α, we find the critical value $\Delta\mathcal{L}_{\rm crit}$ such that

$$\left(\int_0^{\Delta\mathcal{L}_{\rm crit}} \Delta\mathcal{L}_i\right) \Big/ \left(\int_0^{\infty} \Delta\mathcal{L}_i\right) = \alpha \ , \tag{6}$$

and our acceptance region is the set $\{\theta_r\}$ where $\Delta\mathcal{L}_{\rm data}(\theta_r) < \Delta\mathcal{L}_{\rm crit}(\theta_r)$. In this way we have used the likelihood ratio as an *ordering principle* to sort the possibilities into increasing statistical significance. We also point out that the exclusion region at CL α is simply the complement of this set, as acceptance / exclusion is just defined by which side of the critical value one is on.

3 Incorporating Systematic Errors

Unfortunately, the above procedure does not incorporate any kind of systematic errors. In statistical terms, a systematic error can be treated as a *nuisance parameter*: a parameter which one must know to calculate the expected signal, but the value of which is not important to the result. The likelihood depends now on both physics parameters θ_r and nuisance parameters θ_s, but one needs to "project out" any confidence intervals into only the θ_r space.

The key to this procedure is to use an approximation for the likelihood ratio that, in a sense, uses the worst-case values for the nuisance parameters θ_s – the values which make the data fit the hypothesis the best at that point θ_r. Mathematically, we find the best values for θ_s in both the numerator and the denominator of the likelihood ratio:

$$\Delta\mathcal{L}_p(\theta_r) = \mathcal{L}(\theta_r, \hat{\hat{\theta}}_s) - \mathcal{L}(\hat{\theta}_r, \hat{\theta}_s) \ , \tag{7}$$

where we have globally minimized the second term, and we have conditionally minimized the first term, keeping θ_r fixed but varying the nuisance parameters to find $\hat{\hat{\theta}}_s$. This statistic is called the *profile likelihood*.

The profile likelihood is used in combination with the χ^2 approximation in the "MINOS" method in the MINUIT suite and is also explored in some detail by Rolke *et al.* in [3, 4]. To extend the Feldman-Cousins frequentist construction to the profile likelihood, we follow the method suggested by Feldman [5]: we perform Monte Carlo experiments as before, but instead of iterating through the entire (θ_r, θ_s) space, at each point in the physics parameter space θ_r we fix θ_s to its best-fit value from the *data*, $\hat{\hat{\theta}}_s$. Then we recalculate the profile likelihood for the experiment as defined in

equation 7. As before, this gives us a set of likelihood ratios $\{\Delta\mathcal{L}_{p,i}\}$ with which we can define the critical value for a CL α which depends only on θ_r.

To summarize, we describe the procedure step-by-step:

1. The test statistic / ordering principle is the profile likelihood $\Delta\mathcal{L}_p$ as defined in eq. 7.

2. The profile likelihood for the data is calculated at each point θ_r, with the numerator being a conditional minimum at $(\theta_r, \hat{\hat{\theta}}_s)$, and the denominator the global minimum at some $(\hat{\theta}_r, \hat{\theta}_s)$.

3. For each point θ_r, we perform a number of Monte Carlo experiments in which we sample from the parent distribution $\{x \mid \theta_r, \hat{\hat{\theta}}_{s,\mathrm{data}}\}$, then we recalculate the profile likelihood for each experiment.

4. For a CL α, at each point we find the critical value $\Delta\mathcal{L}_{p,\mathrm{crit}}(\theta_r)$ using eq. 6, and this point is in the allowed region if $\Delta\mathcal{L}_{p,\mathrm{data}}(\theta_r) < \Delta\mathcal{L}_{p,\mathrm{crit}}(\theta_r)$.

4 Discussion

We note that the problem of incoporating systematic errors into confidence intervals is still an area of active research: for a survey of recent approaches, including hybrid Bayesian-frequentist methods not discussed here, see [6]. Two fully frequentist constructions (not using the profile likelihood approximation) have been employed in test cases by by G. Punzi [7] and K. Cranmer [8], but there is not a general consensus on an ordering principle. For further information, we refer the reader to the discussion by Cranmer in [9][2].

References

[1] G. J. Feldman and R. D. Cousins, Phys. Rev. **D57**, 873 (1998), arXiv:physics/9711021.

[2] A. Stuart, K. Ord, and S. Arnold, *Kendall's Advanced Theory of Statistics*, Volume 2A, 6th ed. (London:Arnold 1999) and earlier editions.

[3] W. A. Rolke and A. M. Lopez, Nucl. Instrum. Meth. A **458**, 745 (2001), arXiv:hep-ph/0005187.

[4] W. A. Rolke, A. M. Lopez, and J. Conrad, arXiv:physics/0403059.

[5] G. Feldman, Fermilab colloquium, `http://www.hepl.harvard.edu/~feldman/Journeys.pdf`.

[2] We also note this as the origin of the term "profile construction" to describe this method.

[6] R. Cousins, Proceedings of PHYSTAT 05, Oxford, U.K. (2005), http://www.physics.ox.ac.uk/phystat05/proceedings/default.htm.

[7] G. Punzi, Proceedings of PHYSTAT 05, Oxford, U.K. (2005), arXiv:physics/0511202.

[8] K. Cranmer, Proceedings of PHYSTAT 03, Palo Alto, U.S.A. (2003), arXiv:physics/0310108.

[9] K. Cranmer, Proceedings of PHYSTAT 05, Oxford, U.K. (2005), arXiv:physics/0511028.