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 $\theta_r$. For a binned distribution, a natural choice arises from the Poisson probability (or likelihood)

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

(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 $\mu_i$ for the hypothesis we are testing with parameters $\theta_r$. At this point it is conventional to switch to the negative logarithm (the log likelihood):

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

(2)

We will come back to the additional factor of 2.
To compare the probabilities 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):

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

where hypothesis $H_1$ with parameters $\theta_{r1}$ gives us an expected count $\mu_{1,i}$, and hypothesis $H_2$ with parameters $\theta_{r2}$ 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 $\theta_r$ to the hypothesis that fits the data the best. Specifically, in the physics parameter space $\theta_r$, the test statistic is the difference of the log likelihood at this point to the best-fit hypothesis with parameters $\hat{\theta}_r$ ($L$ is minimized by $\hat{\theta}_r$):

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

The additional factor of 2 added in equation 2 arises because in the Gaussian regime, $\Delta L$ so defined approaches a $\chi^2$ distribution with degrees of freedom equal to the dimension of $\theta_r$ (Wilks’ Theorem).

### 2.2 Confidence Intervals

At this point, we wish to examine all the physically allowed hypotheses by iterating over the space $\theta_r$, 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 $\chi^2$ distribution. Specifically, one calculates $\Delta L$ at every point $\theta_r$, and for a given confidence level (CL) $\alpha$, the allowed region is the set

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

For two parameters and a 90% confidence level, we would allow the region where $\Delta L < 4.61$. 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, $\Delta L$ can deviate from the simple $\chi^2$ 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 $\theta_r$ is reduced and the $\chi^2$ 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.

---

1By minimizing the negative log likelihood, we maximize the probability.
Specifically, at each point in the parameter space $\theta_r$, we perform a number of Monte Carlo experiments where we sample from the parent distribution $\{x | \theta_r\}$ and then calculate the likelihood ratio $\Delta L_i$ for the experiment. This set of $\{\Delta L_i\}$ allows us to see how our test statistic behaves under statistical variations only. Then to define our confidence intervals at CL $\alpha$, we find the critical value $\Delta L_{\text{crit}}$ such that

$$
\left( \int_0^{\Delta L_{\text{crit}}} \Delta L_i \right) / \left( \int_0^{\infty} \Delta L_i \right) = \alpha ,
$$

and our acceptance region is the set $\{\theta_r\}$ where $\Delta L_{\text{data}}(\theta_r) < \Delta L_{\text{crit}}(\theta_r)$. 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 $\alpha$ 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 $\theta_r$ and nuisance parameters $\theta_s$, but one needs to “project out” any confidence intervals into only the $\theta_r$ 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 $\theta_s$ – the values which make the data fit the hypothesis the best at that point $\theta_r$. Mathematically, we find the best values for $\theta_s$ in both the numerator and the denominator of the likelihood ratio:

$$
\Delta L_p(\theta_r) = L(\theta_r, \hat{\theta}_s) - L(\hat{\theta}_r, \hat{\theta}_s) ,
$$

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

The profile likelihood is used in combination with the $\chi^2$ 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 $(\theta_r, \theta_s)$ space, at each point in the physics parameter space $\theta_r$ we fix $\theta_s$ to its best-fit value from the data, $\hat{\theta}_s$. Then we recalculate the profile likelihood for the experiment as defined in
equation 7. As before, this gives us a set of likelihood ratios \( \{ \Delta L_{p,i} \} \) with which we can define the critical value for a CL \( \alpha \) which depends only on \( \theta_r \).

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

1. The test statistic / ordering principle is the profile likelihood \( \Delta L_p \) as defined in eq. 7.

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

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

4. For a CL \( \alpha \), at each point we find the critical value \( \Delta L_{p,\text{crit}}(\theta_r) \) using eq. 6, and this point is in the allowed region if \( \Delta L_{p,\text{data}}(\theta_r) < \Delta L_{p,\text{crit}}(\theta_r) \).

4 Discussion

We note that the problem of incorporating 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 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


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

