DocDB 15884 v3 August 12, 2016

Notes on the Inclusion of Nuisance Parameters in the Unified Approach

Gary Feldman

1. Introduction: Frequentist and Bayesian Inference

There are two main approaches to statistical inference, frequentist and Bayesian. A frequentist calculates the probability of getting a particular set of data given a hypothesis, $P(\text{data}\,|\,\text{hypothesis})$. At least in principle, this is a calculable quantity. A Bayesian, on the other hand, attempts to calculate the probability of a hypothesis being true given a set of data, $P(\text{hypothesis}\,|\,\text{data})$. Although this is what we would actually like to know, it is not calculable without some additional information.

Note that $P(A\,|\,B)$ is not in general equal to $P(B\,|\,A)$. The classical example is that the probability of a person being pregnant given that she is a female is perhaps 2%, while the probability that a person is a female, given that she is pregnant is considerably higher. However, the relationship of these two quantities is given by Bayes' theorem, which in its simplest form is $P(A\,|\,B)P(B) = P(B\,|\,A)P(A)$. Thus, for a Bayesian to calculate her desired quantity, she must use Bayes' theorem, which then states

$$P(\text{hypothesis}\,|\,\text{data}) = \frac{P(\text{data}\,|\,\text{hypothesis})P(\text{hypothesis})}{P(\text{data})}. \qquad (1.1)$$

The denominator is just a normalization term, but the potential problem is the probability of the hypothesis, which is called a prior probability, or just "a prior." Unfortunately, the prior is inherently a subjective quantity incorporating all previous information and intuition. For example, I have a bottle of wine bet with Stephen Parke that θ_{23} will be in the lower octant. We clearly have different priors. People have tried to define "unbiased" priors, but these efforts are usually unsuccessful, unless the data completely overwhelms the prior. It is mainly for this reason that Bob Cousins and I called for experimental data to be presented using frequentist inference.[1]

An old chestnut summarizes the situation: "A frequentist uses impeccable logic to calculate something that nobody cares about. A Bayesian calculates what everyone wants to know using assumptions that nobody believes."

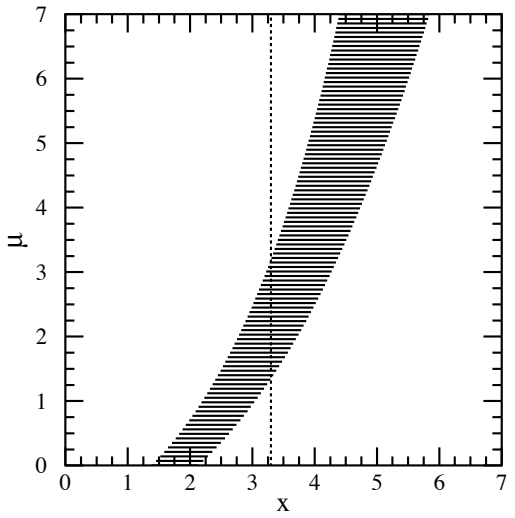
2. Confidence Intervals and Neyman Constructions

The situation for a frequentist is not as bleak as it may seem at this point because she can create something that is very close to what a Bayesian creates without the need of a prior. It is a statement of the form "The true value of the parameter μ lies between μ_a and μ_b with a confidence level *c*." The true value of μ will be within the specified

[1] G. J. Feldman and R. D. Cousins, *Phys. Rev. D* **57**, 3873 (1998).

range, called a "confidence interval," c of the time and outside the specified range $(1 - c)$ of the time.[2] An analysis that produces such a correct statement for all possible values of μ is said to cover. If there is any value of μ for which the probability of its being in the confidence interval is less than c, it is said to undercover, and, conversely if there is any value of μ for which the probability of its being in the confidence interval is greater than c, it is said to overcover.[3] In frequentist analyses, undercoverage is forbidden and overcoverage is allowed only to the extent that it is unavoidable due to discrete variables, nuisance parameters, or similar issues.[4]

In 1937, Jerzy Neyman developed a method for producing such a statement, which we now call a "Neyman construction."[5] Consider a simple example: Suppose that we measure a quantity x and want to deduce limits on the true value of a parameter μ from it. Then, for every possible value of μ, we calculate the probability of each possible outcome of the experiment and designate a set of contiguous possible measurements, x_a to x_b the sum of whose probabilities are equal to c as the "confidence belt" associated with μ.[6] Once the experiment is done, we simply take all the values of μ whose confidence belts include our measured value of x to be the confidence interval for the true value of μ. Coverage is guaranteed by construction for every value of μ. This is illustrated in the figure to the right. The dotted line represents our measured value of x and the confidence interval is given by the intersection of the dotted line and the confidence belt, in this case between about 1.2 and 3.0. Note that the confidence belt is constructed horizontally and the confidence interval is read out vertically.



---

[2] Throughout this note we will refer to parameters such as μ (and λ to be introduced later) as if they are a single parameter. However, the extension to the case in which they are sets of parameters is completely straightforward. In the case of a single parameter μ, we generate a confidence interval. For the case in which μ is a set of n parameters, we generate an n-dimensional confidence contour.

[3] These statements are too terse to give the full meaning to coverage. See section 7 for a more complete discussion of coverage and the procedure to test it.

[4] The analog of confidence intervals in Bayesian inference are usually called "credible intervals." There is no requirement of coverage for credible intervals. However, to convince physicists of their usefulness, when used to analyze data they are often accompanied with both analyses of their sensitivity to priors and their coverage properties.

[5] J. Neyman, *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences* **236** (767): 333–380 (1937).

[6] For the NOνA cases, the theory may not allow only contiguous intervals. I think we can safely interpret Nyman's requirement to mean that the intervals should not be gratuitously noncontiguous.

There is freedom within the Neyman construction to choose which set of measurements to include in the confidence belts, and an ordering principle is required to do this. Several ordering principles are in common use. For example, starting the confidence belt with $x = 0$ produces an ordering for an upper limit, starting the confidence belt after the first (1 – c) of probability produces an ordering for a lower limit, and placing (1 – c)/2 of probability above and below the confidence belt produces a central value ordering. The choice of an ordering principle should be chosen before examining the data to avoid a process we call "flip-flopping," which does not cover.[1]

The above figure, which illustrates central value ordering, shows some undesirable features. For example, if x is measured to be 1.5, the resulting upper limit is extremely small, much below the sensitivity level of the experiment. If x is measured to be 1.2, every value of μ is excluded at confidence level c. If μ is a parameter that must have some value, then this conclusion is clearly wrong. Now, we are allowed, in fact expected, to be wrong $(1 - c)$ of the time, but this is a case in which we know that we are in the $(1 - c)$ fraction of the time. It does not appear to be good policy to spend millions of dollars of taxpayers' money only to announce that the result of the experiment is wrong.

Pathologies of this type are common when the data indicate a value of μ near a physical boundary.

3. The Unified Approach: Likelihood Ratio Ordering

To avoid problems of this type, Bob Cousins and I proposed what we thought was a novel ordering principle.[1] When our paper was in proof, we discovered that our proposal had been in the standard reference book of statistics for almost 40 years;[7] however, as far as we know, no one had ever actually used it. The proposal was to use the likelihood ratio as the ordering principle. We form a rank R,

$$R(\mu) = \frac{\mathcal{L}(x \mid \mu)}{\mathcal{L}(x \mid \hat{\mu})}, \tag{3.1}$$

where $\hat{\mu}$ is the value of μ that maximizes the denominator. We include potential measurements x in the confidence belt for each μ in decreasing rank order until the required sum of probabilities equal to c is obtained. The advantages of the likelihood ratio ordering is that
(1) it avoids pathologies that occur with other orderings near a physical boundary;
(2) it automatically chooses limits or central values, eliminating flip-flopping; and
(3) it reduces to the normal central value limits far from a physical boundary.

[7] M. Kendall and A. Stuart, *The Advanced Theory of Statistics, Volume 2: Inference and Relationship* (Hafner, 1961).

4. The Unified Approach: Procedure When Parameters Are Estimated from Fits to Spectra

This section will discuss the preferred procedure when parameters are estimated from fits to spectra, but will ignore nuisance parameters. In the following section, we will deal with nuisance parameters by removing them. Once the nuisance parameters are removed, this section will be relevant, subject to a modification discussed in Section 6. The ordering of the sections is to allow the introduction of needed terminology and symbols.

The procedure is as follows:

(1) For the measured data, a set of N_i events in each bin, obtain the best fit to the true parameter $\hat{\mu}$, and the number of expected events in each bin, M_i, for $\hat{\mu}$.[8]

(2) (A) For each set of potential values of the unknown parameter μ, obtain the number of expected events in each bin, m_i.[9]

(B) Throw a large number of pseudo-experiments for the appropriate flux. For each pseudo-experiment,

(a) obtain the number of events in each bin, n_i, and the best fit to the parameter μ' and the number of expected events in each bin m_i'.

(b) calculate and tabulate $\Delta\chi^2(\mu)$,[10]

$$\Delta\chi^2(\mu) = 2\sum_i \left[m_i - m_i' + n_i \ln\left(\frac{m_i'}{m_i}\right)\right]. \tag{4.1}$$

(C) Calculate the critical value of $\Delta\chi^2(\mu)$, $\Delta\chi_c^2(\mu)$, such that *c* of the thrown experiments have $\Delta\chi^2(\mu) \le \Delta\chi_c^2(\mu)$, where c is the desired confidence level.[11]

(D) Calculate $\Delta\chi^2(\mu,\hat{\mu})$, the $\Delta\chi^2$ between the data and the hypothesis μ,

$$\Delta\chi^2(\mu,\hat{\mu}) = 2\sum_i \left[m_i - M_i + N_i \ln\left(\frac{M_i}{m_i}\right)\right]. \tag{4.2}$$

If $\Delta\chi^2(\mu,\hat{\mu}) \le \Delta\chi_c^2(\mu)$, then μ is within the confidence interval. Otherwise it is not.

---

[8] The use of expected rather than observed events in this section is not required, but I think it increases robustness and will speed up the removal of nuisance parameters.

[9] The expected numbers of events include both signal and background. The systematic and small statistical uncertainties in the background estimation can be treated as a nuisance parameter.

[10] Note that $\Delta\chi^2(\mu) = -2\ln[R(\mu)]$, so sections 3 and 4 are doing the same thing, although they appear different.

[11] $\Delta\chi_c^2(\mu)$ defines the confidence belt for the parameter μ. We can use these two terms interchangeably.

5. The Unified Approach: Treatment of Nuisance Parameters.

A nuisance parameter is a parameter whose value we are not particularly interested in for the task at hand, but for which we are obliged to provide coverage for all values of the parameter, at least in a frequentist analysis. For example, if we just want to know what we can say about δ_{CP} in the normal ordering, then all of the other physics parameters such as $\sin^2(\theta_{23}), \sin^2(2\theta_{13}), \Delta m^2_{32},$ etc., as well as all of the experimental systematic uncertainties are nuisance parameters. The only practical way to deal with nuisance parameters is to remove them from the likelihood function.

There are two common ways of removing nuisance parameters, a Bayesian procedure known as marginalizing and a frequentist procedure known as profiling. Marginalizing is just averaging over the likelihood function:

$$\mathcal{L}(x \mid \mu, \lambda) \rightarrow \int \mathcal{L}(x \mid \mu, \lambda) \mathrm{d}\lambda = \tilde{\mathcal{L}}(x \mid \mu), \quad [12] \tag{5.1}$$

where λ represents the nuisance parameter. For the procedure in Section 4, this would just be randomly throwing the nuisance parameter in the pseudo-experiments from an appropriate distribution, often Gaussian. The frequentist objection to marginalization is that there is a hidden prior involved. This can be most easily seen in that there is a choice of the variable of integration; it could just as easily be $\mathrm{d}\lambda^2, \mathrm{d}(1/\lambda), \mathrm{d}(\sin^2\lambda),$ or any of an infinite number of other variables. The attraction of this procedure is that it is easy to do. We will return to this option below.

Profiling is the procedure that was suggested by Ref. 7. Equation 3.1 becomes

$$R(\mu, \lambda) = \frac{\mathcal{L}(x \mid \mu, \lambda)}{\mathcal{L}(x \mid \hat{\mu}, \hat{\lambda})} \rightarrow \tilde{R}(\mu) = \frac{\mathcal{L}(x \mid \mu, \hat{\hat{\lambda}})}{\mathcal{L}(x \mid \hat{\mu}, \hat{\lambda})}, \tag{5.2}$$

where $\hat{\mu}$ and $\hat{\lambda}$ maximize the denominator and $\hat{\hat{\lambda}}$ maximizes the numerator for the parameter μ. In the language of Section 4, one simply replaces the nuisance parameter by the value that minimize the $\Delta\chi^2$ between the data and the hypothesis μ, Eq. 4.2, in step 2A. These nuisance parameters stay fixed for step 2B. The idea here is that if we cover for the values of nuisance parameters most favorable to the data, we should cover for all values of the nuisance parameters.[13]

For small Gaussian-distributed uncertainties, marginalization and profiling should give the same results. The issue of the Bayesian prior is not present because the uncertainties are assumed to be Gaussian in the parameter that they are thrown. Thus, I think it is acceptable to randomly throw small systematic uncertainties from a Gaussian distribution in step 2B. However, profiling should be done for the major physics parameters being treated as nuisance parameters and any unusually large systematic uncertainties.

---

[12] The normalization of the likelihood function is not relevant.

[13] See Section 7 for a more detailed discussion.

6. **The Unified Approach: Modification of the Procedure in the Presence of Profiled Nuisance Parameters**

Often we want to display our knowledge of a parameter, or a set of parameters, without regard to the values of other parameters, as in the example cited at the beginning of the previous section. Since we are willing to accept any values of the nuisance parameters, we expect the confidence intervals or contours to be smaller than if we just take the projection of the full multi-dimensional contour on a lower dimensional contour. To do this, we need to remove the effects of the nuisance parameters from each of the thrown experiments, and we can do this by setting them to values that are most favorable to the μ being tested before calculating the $\Delta\chi^2(\mu)$.

Thus, the procedure of Section 4 becomes

(1) For the measured data, a set of N_i events in each bin, obtain the best fit to the true parameters $\{\hat{\mu},\hat{\lambda}\}$,[14] and the number of expected events in each bin, M_i, for $\{\hat{\mu},\hat{\lambda}\}$.

(2) (A) For each set of potential values of the unknown parameter μ, find the values of the nuisance parameter $\hat{\hat{\lambda}}$ that most favors the data. If, for a given value of λ, the numbers of expected events in each bin are $\bar{m}_i$, $\hat{\hat{\lambda}}$ is obtained by minimizing

$$\Delta\chi^2 = 2\sum_i \left[\bar{m}_i - M_i + N_i \ln\left(\frac{M_i}{\bar{m}_i}\right)\right]. \tag{6.1}$$

For $\{\mu,\hat{\hat{\lambda}}\}$ call the number of expected events in each bin, m_i.

(B) Throw a large number of pseudo-experiments for the appropriate flux and m_i. For each pseudo-experiment,

(a) obtain the number of events in each bin, n_i, and the best fit to the parameters $\{\mu',\lambda'\}$.

(b) If, for given values of $\{\mu',\lambda\}$, the numbers of expected events in each bin are $\bar{m}_i'$, then the closest point to $\{\mu,\hat{\hat{\lambda}}\}$ can be found by varying λ to minimize[15]

$$\Delta\chi^2(\mu) = 2\sum_i \left[m_i - \bar{m}_i + n_i \ln\left(\frac{\bar{m}_i}{m_i}\right)\right]. \tag{6.2}$$

(c) Tabulate the minimized value of $\Delta\chi^2(\mu)$.

---

[14] If constraints on the nuisance parameters are desired, appropriate penalty terms can be added to the fits here and below.

[15] If minimizing each pseudo-experiment is too costly in CPU time, the values of μ' can be collected in fine bins and the minimization can be done on the central value of each bin.

(C) Using the minimized values of $\Delta\chi^2(\mu)$, calculate the critical value of $\Delta\chi^2(\mu)$, $\Delta\chi^2_c(\mu)$, such that *c* of the thrown experiments have $\Delta\chi^2(\mu) \le \Delta\chi^2_c(\mu)$, where c is the desired confidence level.

(D) Calculate $\Delta\chi^2(\mu,\hat{\mu})$, the $\Delta\chi^2$ between the data and the hypothesis μ,

$$\Delta\chi^2(\mu,\hat{\mu}) = 2\sum_i \left[m_i - M_i + N_i \ln\left(\frac{M_i}{m_i}\right)\right]. \qquad (6.3)$$

If $\Delta\chi^2(\mu,\hat{\mu}) \le \Delta\chi^2_c(\mu)$, then μ is within the contour or interval. Otherwise it is not.

7. The Unified Approach: The Meaning of Coverage and the Procedure to Test for It

In Section 2 we stated that the goal of a frequentist analysis is often to produce a statement of the form "The true value of the parameter μ lies between μ_a and μ_b with a confidence level *c*." The phrase "…with a confidence level *c*" means that the first part of the sentence is true *c* of the time. If for every value (or set of values) of μ, the statement is true exactly *c* of the time, the procedure that produced the statement is said to "cover." In the absence of discrete variables (e.g., numbers of events) and nuisance parameters, the procedure we have given above, the Neyman construction, produces exact coverage by construction. This is obvious because we have made the confidence belt, the horizontal lines in the figure in Section 2, just the right length to have this property. If we made the confidence belt longer, we would have the statement be true more than *c* of the time, and this would be "overcoverage." Conversely, if we made the confidence belt shorter, we would have "undercoverage."

Now let's consider how the situation changes when we have nuisance parameters. We need to cover all values of μ and λ. Overcoverage is now unavoidable because there will probably be values of λ that are very unlikely to give us the measurements that we obtained. For these values of λ, we need only a short confidence belt or, perhaps, no belt at all. For each value (or set of values) of μ, we will only have one belt, because we do not care about the values of λ. However, we would like that belt to be as short as possible, because that will give us the shortest confidence intervals or most compact confidence contours. Our strategy to do this is given in Section 6. For each μ, we set λ to the value most favorable to the data, and we exactly cover for this point. Since we have covered for the most likely value of λ, we hope that all the less likely values of λ will give us confidence belts that are contained in the one we calculated. If this proves not to be the case, we need to expand the confidence belt to accommodate the rogue λs.[16]

Notes on confirming coverage:

(1) We only need to confirm coverage for the observed data.

[16] Expanding the confidence belt may or may not increase the confidence interval depending on where the observed data cuts the belt. See the figure in Section 2.

(2) We do not need to confirm coverage for values of μ between μ_a and μ_b, because our statement about the true value of μ is automatically true for these values.
(3) For values of μ outside the confidence interval, we do not need to confirm coverage for the value of λ most favorable to data, because we have already shown that these points are outside the confidence interval by construction.
(4) Thus, we only need to search for values of μ outside the confidence interval and values of λ not most favorable to the data, and verify that these $\{\mu,\lambda\}$ points have confidence belts wholly within the one we have constructed for this value of μ. If we find cases where this is not true, then the procedure is to extend the confidence belt to accommodate these rogue values.
(5) With the addition of the procedure for handling rogue λs, we have a procedure that is guaranteed to cover for all cases. Further, by construction, it is guaranteed to yield the most compact confidence intervals possible using this procedure.[17]

[17] There may be other procedures that give more compact confidence intervals, but if the procedure is chosen based on the observed data, this is flip-flopping, which in general does not cover for an ensemble of measurements.