Search Procedures in High Energy Physics

Luc Demortier

Laboratory of Experimental High-Energy Physics
The Rockefeller University

Abstract

The usual procedure for searching for new phenomena in high energy physics involves a frequentist hypothesis test followed by the construction of an interval for the parameter of interest. This procedure has a couple of well-known flaws: the effect of the test on subsequent inference is ignored, and in some circumstances the size of the reported interval does not properly reflect experimental conditions. Furthermore, proper treatment of nuisance parameters in a frequentist context is nearly always troublesome. For these reasons there has recently been considerable interest in applying the ideas of Bayesian reference analysis to this problem. We describe ongoing work to calculate reference priors for search experiments, via both numerical and analytical methods. We also show how intrinsic interval estimation can provide a very elegant solution to the testing problem in search procedures.

Prepared for O-Bayes09, the Seventh International Workshop on Objective Bayes Methodology, Philadelphia, USA, June 5–9, 2009.

1luc.demortier@rockefeller.edu
Contents

1 Introduction and Motivation ........................................ 3

2 Reference Priors with Partial Information ................. 6
   2.1 The Single-Count Model ..................................... 7
   2.2 Method 1 Applied to the Single-Count Model ............ 7
   2.3 Method 2 Applied to the Single-Count Model ............ 9
   2.4 Marginal Priors and Posteriors .......................... 11
   2.5 Generalizations of the Single-Count Model .............. 12

3 Upper Limits ...................................................... 14

4 Coverage ............................................................ 14

5 Intrinsic intervals .................................................. 15
   5.1 Conditional approach ....................................... 17
   5.2 Marginal approach ......................................... 18

6 Reference Analysis Search Procedures .......................... 18

7 Open Questions .................................................... 22

Acknowledgments ..................................................... 22

Figures ............................................................... 23

References ........................................................... 35
1 Introduction and Motivation

One of the main goals of experimental high energy physics is to discover new particles that may shed light on such basic questions as the origin of mass, the architecture of fundamental forces, the origin of the universe, and the nature of dark matter and dark energy. The principal tools for these investigations are the large particle accelerator centers currently operating at Fermilab (near Chicago, IL), at CERN (near Geneva, Switzerland), and in a few other places. Accelerated particles are collided with each other at very high energies and very high rates, and state-of-the-art detectors are used to search the collision products for evidence of new particles. The great complexity of these experiments has motivated physicists to take a closer look at the statistical aspects of the procedures they use to search for new phenomena and claim discoveries.

In a typical search the new particle of interest, or “signal”, is produced (if it exists) in a very small fraction of collisions; however, its signature in the detecting apparatus is often mimicked by more common particles. Although this background can almost never be completely eliminated, it can be minimized by selecting collision events with appropriate characteristics. The background contamination $\mu$ that remains after event selection must then be quantified and subtracted from the observed event rate in order to obtain the signal rate. The latter is proportional to a quantity of intrinsic physics interest, the so-called signal cross section $\sigma$; the proportionality factor is a product of accelerator dependent parameters and event selection efficiencies collectively referred to as the effective luminosity $\epsilon$. In general the process by which collision events are selected for further study is very stringent and rejects the vast majority of collisions. Thus the final observed event rate obeys Poisson statistics to high accuracy, and the Poisson mean can be written as $\epsilon \sigma + \mu$. In the remainder of this paper we will assume that after event selection we are left with a data sample of $n$ events, and that some information is available about the true values of $\epsilon$ and $\mu$.

For historical reasons physicists tend to favor frequentist approaches to statistical problems, and their standard search procedure takes the following form:

1. Choose three confidence levels $\alpha_1$, $\alpha_2$, and $\alpha_3$.
2. Find a test statistic $T$ that is sensitive to the presence of signal.
3. Compute the $p$ value corresponding to the observed value of $T$.
4. If $p \leq 1 - \alpha_1$, claim discovery and compute an $\alpha_2$ confidence level two-sided interval on the signal cross section $\sigma$.
5. If $p > 1 - \alpha_1$, make no claim, and compute an $\alpha_3$ confidence level upper limit on $\sigma$.

Common confidence level choices are $1 - \alpha_1 = 2.8 \times 10^{-7}$, $\alpha_2 = 0.68$, and $\alpha_3 = 0.95$. The two-sided interval calculated at step 4 provides an estimate of the magnitude of the observed effect. When no discovery is claimed, the upper limit at step 5 is an upper bound on values of the signal cross section $\sigma$ that the experiment cannot convincingly distinguish from zero. In a sense, this bound serves as a measure of the
severity of the test to which the background-only model was subjected \cite{14}: the smaller the upper limit, the smaller a deviation from background the test is able to detect, and the stronger the case is in favor of the background-only model when the test failed to reject it. The upper limit is also an indicator to future experiments of what their sensitivity should be if they want to probe the same phenomenon more deeply.

Statisticians are often surprised by the stringency of the \( p \) value test in the above procedure: in a standard normal distribution, a one-sided tail probability of \( 2.8 \times 10^{-7} \) corresponds approximately to a \( 5\sigma \) deviation from the mean. This high discovery threshold was suggested more than 40 years ago by the physicist A. Rosenfeld \cite{16}, who was concerned by the so-called “look-elsewhere” effect, whereby the probability of a large background fluctuation increases with the number of independent tests that are performed. He estimated the number of histograms that are examined every year by high energy physicists, and with a few simple assumptions concluded that one should expect several spurious discovery claims at the \( 4\sigma \) level per year. This led him to recommend a \( 5\sigma \) threshold.

In a paper widely cited by experimental physicists, Feldman and Cousins (themselves experimental physicists) criticized the above search procedure for yielding intervals that do not have exact frequentist coverage, and in fact undercover, because the \( p \) value test is not incorporated in the frequentist reference ensemble of these intervals \cite{11}. In other words, the decision to report an upper limit or a two-sided interval is based on the result of the test, but the intervals are constructed as if that decision had been made \textit{before} looking at the data. The authors called this practice “flip-flopping”. A second problem they noted is that in some cases the above procedure yields an empty interval, a result that, even though correct from a frequentist point of view, is unsatisfying from a scientific one. The solution they proposed was to replace the standard search procedure by a Neyman interval construction with a likelihood ratio ordering rule. In addition to never producing an empty interval\footnote{Although intervals constructed with a likelihood-ratio ordering rule are never empty, they sometimes have very low Bayesian credibility; see \cite{9} for an example.}, this rule has the property that the resulting interval smoothly transitions from one-sided to two-sided as the observed effect increases in strength. Thus the decision to claim discovery at a given significance level depends on whether or not the corresponding likelihood ratio interval contains the zero-signal value \( \sigma = 0 \), and the choice between reporting an upper limit or a two-sided interval is taken out of the hands of the user and becomes automatic.

Although this method has become quite popular, it does seem to remove the degrees of freedom that the standard procedure offers in the choice of the confidence levels \( \alpha_1 \), \( \alpha_2 \), and \( \alpha_3 \). The Feldman-Cousins construction requires these three \( \alpha \)'s to be equal, resulting either in intervals that are too wide to be informative or in tests that are too lenient to be trustworthy. To address this issue, the authors propose to report more than one interval for each measurement. One could for example report three intervals, each corresponding to one of the confidence levels of the standard procedure. This does not solve all problems however. If the \( \alpha_1 \) confidence interval includes the null value \( \sigma = 0 \), one cannot claim a discovery, but there is no guarantee that the \( \alpha_3 \) confidence interval will be an upper limit, as it should be in order to serve the same necessary
purpose as in the standard procedure. Furthermore, if a consumer down the line adopts
the policy of using the $\alpha_2$ interval every time a discovery is claimed and the $\alpha_3$ interval
otherwise, then he or she falls back into the undercoverage problem of the standard
procedure. Thus, while the Feldman-Cousins procedure is useful for characterizing the
result of a measurement, it does not allow one to follow through with a choice of interval
once a decision has been made regarding discovery, essentially rendering that decision
toothless. One of the authors argues that this last step, the act of deciding about
discovery and drawing further inferences, should be handled via a subjective Bayesian
approach [7], but it is not entirely clear how a set of Feldman-Cousins intervals can
serve as input to such a procedure.

Some statisticians may wonder why the flip-flopping problem wasn’t solved by
adopting a conditional frequentist approach: in principle one could construct post-
test intervals conditionally on the result of the test, simply by appropriately restricting
the sample space and renormalizing the relevant probability densities [9]. The answer,
unfortunately, is that most physicists are not very familiar with conditional frequen-
tism, and all the possible ramifications of such an approach have not been investigated
yet. In addition, this approach does not necessarily solve the empty-interval problem.

The Feldman-Cousins construction has another well known problem [12]: when the
observed number of events is smaller than the expected background, the upper limit on
the signal decreases rapidly as the expected background increases. This would argue
against efforts at optimizing the composition of the dataset, since a larger background
contamination will ensure a more stringent upper limit. The authors respond that when
given two experimental results, one should compare expected upper limits in addition
to the observed ones, the former being a measure of the sensitivity of an experiment.
It is interesting that Bayesian methods do not have this problem because they obey
the likelihood principle. For example, when no events are observed and when the
background $\mu$ and effective luminosity $\epsilon$ are known exactly, then the Bayesian upper
limit on $\sigma$ is independent of $\mu$ if the $\sigma$ prior is flat.

Finally, as with all frequentist methods, there is the problem of nuisance parameters.
In principle it can be handled via Neyman’s construction, but as the number of such
parameters increases this approach quickly becomes unmanageable in practice. Simpler
solutions involve profiling or marginalizing the likelihood with respect to the nuisance
parameters, but the coverage of the resulting intervals is not guaranteed and must be
investigated on a case-by-case basis.

The above discussion suggests that an objective Bayesian approach may provide
a solution to the problems encountered. We propose here to follow the methodol-
gy known as reference analysis and advocated by J. Bernardo, J. Berger, and their
collaborators [5, and references therein]. This approach has several advantages that
should resonate with the concerns of high energy physicists: the inferences it yields
about physics parameters are transformation invariant; it has good frequentist cover-
age properties; it is very general; it is computationally tractable; it avoids the famous
marginalization paradoxes discovered in the 1970’s; and it can (and should) be embed-
ded in a subjective Bayesian framework.

The remainder of this paper is organized as follows. In the next section we describe
briefly the calculation of reference priors when partial information is available about
the nuisance parameters, a common situation in high energy physics. We apply this technology to the so-called single-count model, basically a Poisson pdf with mean \( \mu + \epsilon \sigma \) as described earlier in this introduction. We examine the marginal reference prior and posterior for this model in section 2.4. Ongoing work on numerical calculations dealing with various generalizations of the single-count model is described in section 2.5. We study reference posterior upper limits in section 3 and the frequentist coverage of reference posterior intervals in section 4. Intrinsic intervals for the single-count model are introduced in section 5 and are used to define a Bayesian search procedure in section 6. Finally, some open questions are listed in section 7.

2 Reference Priors with Partial Information

A good description of the reference prior algorithm for any number of nuisance and interest parameters is provided in Ref. [5]. In contrast with the general situation described in that reference, most nuisance parameters encountered in high energy physics come with non-trivial prior information, in the form of auxiliary measurements, Monte Carlo simulations, and/or theoretical beliefs. The incorporation of this partial information in the construction of reference priors is described in Ref. [17]. Suppose that \( \phi \) labels the nuisance parameter(s) and \( \theta \) the parameter of interest. There are two ways to proceed:

Method 1: Assume that we are given a marginal prior \( \pi(\phi) \) for the nuisance parameters; compute the conditional reference prior \( \pi_R(\theta \mid \phi) \) for the interest parameter given a fixed value of \( \phi \); the full prior is then \( \pi(\theta, \phi) = \pi_R(\theta \mid \phi) \pi(\phi) \);

Method 2: Assume that we are given a conditional prior \( \pi(\phi \mid \theta) \) for the nuisance parameter given the interest parameter; marginalize the probability model \( p(x \mid \theta, \phi) \) with respect to \( \phi \) in order to obtain \( p(x \mid \theta) = \int p(x \mid \theta, \phi) \pi(\phi \mid \theta) d\phi \), and compute the reference prior \( \pi_R(\theta) \) for the marginalized model; the full prior is then \( \pi(\theta, \phi) = \pi_R(\theta) \pi(\phi \mid \theta) \).

In many high energy physics measurements there are physics reasons for assuming that the nuisance parameter is independent of the parameter of interest. Information about a detector energy scale for example, is typically determined separately from the measurement of interest, say of a particle mass, and is therefore considered to be a priori independent from one’s information about that particle mass. When an experimenter is willing to make this assumption, he or she can declare that \( \pi(\phi \mid \theta) = \pi(\phi) \) and use Method 2. When this assumption does not seem fully justified, and it is too difficult to elicit the \( \theta \) dependence of \( \pi(\phi \mid \theta) \), then it will seem preferable to use Method 1, which only requires knowledge of the marginal prior \( \pi(\phi) \). When one is unsure of which method to use, one should use both, and treat the results as part of a test of robustness. An important practical advantage of Method 1 is that the conditional reference prior is computed once and for all, for a given model, and can be used with any subjective prior for the nuisance parameters. In contrast, for Method 2 the reference prior must be recomputed every time the subjective priors change.
2.1 The Single-Count Model

As indicated earlier, the basic observable in high energy physics is an event count $N$ (a number of collisions passing a predefined selection procedure) that obeys Poisson statistics. In a very common measurement model the expectation value of $N$ has the form $\epsilon \sigma + \mu$, where $\sigma$ is the cross section of a physics signal process, which we detect with an effective luminosity $\epsilon$, and $\mu$ is a background contamination. Thus, $\sigma$ is the parameter of interest, whereas $\epsilon$ and $\mu$ are nuisance parameters for which we usually have partial information. For physical reasons none of these three parameters can be negative. The likelihood for this model is given by

$$p(n|\sigma, \epsilon, \mu) = \frac{(\epsilon \sigma + \mu)^n}{n!} e^{-\epsilon \sigma - \mu} \quad \text{with} \quad 0 \leq \sigma < \infty \quad \text{and} \quad 0 < \epsilon, \mu < \infty. \quad (2.1)$$

Information about $\epsilon$ and $\mu$ usually comes from a variety of sources, such as auxiliary measurements, Monte Carlo simulations, theoretical calculations, and evidence-based beliefs (for example, some sources of background contributing to $\mu$ may be deemed small enough to ignore, and some physics effects on $\epsilon$ may be believed to be well enough reproduced by the simulation to be reliable “within a factor of 2”). It is therefore natural to represent that information by a subjective prior $\pi(\epsilon, \mu)$. The problem we are facing is that of finding a prior for $\sigma$, about which either little is known or one wishes to pretend that this is so.

2.2 Method 1 Applied to the Single-Count Model

In method 1 [17, section 2.3], we construct the conditional reference prior $\pi_R(\sigma | \epsilon, \mu)$. The first step consists in calculating Jeffreys’ prior for $\sigma$ while holding $\epsilon$ and $\mu$ fixed:

$$\pi_J(\sigma | \epsilon, \mu) \propto \left\{ \mathbb{E} \left[ -\frac{\partial^2}{\partial \sigma^2} \ln p(n | \sigma, \epsilon, \mu) \right] \right\}^{\frac{1}{2}} \propto \frac{\epsilon}{\sqrt{\epsilon \sigma + \mu}}. \quad (2.2)$$

This prior is improper with respect to $\sigma$ however, so that an additional step, known as the “compact support argument,” is required. One starts by choosing a nested sequence $\Lambda_1 \subset \Lambda_2 \subset \cdots$ of compact subsets of the parameter space $\Lambda$ for $(\sigma, \epsilon, \mu)$, such that $\bigcup_i \Lambda_i = \Lambda$ and the integral $K_i(\epsilon, \mu)$ of $\pi_J(\sigma | \epsilon, \mu)$ over $\Omega_i \equiv \{ \sigma : (\sigma, \epsilon, \mu) \in \Lambda_i \}$ is finite. The conditional reference prior for $\sigma$ on $\Omega_i$ is then:

$$\pi_{R,i}(\sigma | \epsilon, \mu) = \frac{\pi_J(\sigma | \epsilon, \mu)}{K_i(\epsilon, \mu)} I_{\Omega_i}(\sigma), \quad (2.3)$$

where $I_{\Omega_i}(\sigma)$ is the indicator function of $\sigma$ in $\Omega_i$. To obtain the conditional reference prior on the whole parameter space, one chooses a fixed point $(\sigma_0, \epsilon_0, \mu_0)$ and takes the limit of the ratio:

$$\pi_R(\sigma | \epsilon, \mu) = \lim_{i \to \infty} \frac{\pi_{R,i}(\sigma | \epsilon, \mu)}{\pi_{R,i}(\sigma_0 | \epsilon_0, \mu_0)}. \quad (2.4)$$

The theory of reference priors does not currently provide guidelines for choosing the compact sets $\Lambda_i$, other than to require that the resulting posterior be proper. In most
cases this choice makes no difference and one is free to base it on considerations of simplicity and convenience. However, we have found that some care is required with the single-count model. Indeed, suppose we choose:

$$\Lambda_i = \left\{ (\sigma, \epsilon, \mu) : \sigma \in [0, u_i], \epsilon \in [0, v_i], \mu \in [0, w_i] \right\},$$  \hspace{1cm} (2.5)

where \{u_i\}, \{v_i\}, and \{w_i\} are increasing sequences of positive constants. If we use these sets, applying eqs. (2.3) and (2.4) to the prior (2.2) yields:

$$\pi_{R,i}(\sigma | \epsilon, \mu) = \frac{1}{K_i(\epsilon, \mu)} \frac{\epsilon}{\sqrt{\epsilon \sigma + \mu}} I_{[0,u_i]}(\sigma),$$  \hspace{1cm} (2.6)

where:

$$K_i(\epsilon, \mu) \equiv \int_{\Omega_i} \frac{\epsilon}{\sqrt{\epsilon \sigma + \mu}} d\sigma = 2 \left[ \sqrt{\epsilon u_i + \mu} - \sqrt{\mu} \right],$$  \hspace{1cm} (2.7)

and therefore:

$$\pi_R(\sigma | \epsilon, \mu) \propto \sqrt{\frac{\epsilon}{\epsilon \sigma + \mu}}.$$  \hspace{1cm} (2.8)

Although this prior is still improper with respect to \sigma, its dependence on \epsilon is different from that of the conditional Jeffreys’ prior (2.2). This demonstrates the potential importance of the compact subset argument. The prior (2.8) has a serious problem however. Suppose that the marginal prior for \epsilon and \mu can be factorized as follows:

$$\pi(\epsilon, \mu) = e^{-\epsilon \sqrt{\epsilon \pi/2}} \pi(\mu).$$  \hspace{1cm} (2.9)

The full posterior density for (\sigma, \epsilon, \mu) is then:

$$p(\sigma, \epsilon, \mu | n) \propto \frac{(\epsilon \sigma + \mu)^n}{n!} e^{-\epsilon \sigma - \mu} \frac{\epsilon}{\sqrt{\epsilon \sigma + \mu}} \frac{e^{-\epsilon}}{\sqrt{\pi \epsilon/2}} \pi(\mu).$$  \hspace{1cm} (2.10)

It is easy to see that this posterior is improper. Indeed, integrating the right-hand side of the above expression over \sigma yields:

$$\frac{\Gamma(n + \frac{1}{2})}{\Gamma(n + 1)} \left[ 1 - P(n + \frac{1}{2}, \mu) \right] \sqrt{\frac{2}{\pi}} \frac{e^{-\epsilon}}{\epsilon} \pi(\mu),$$  \hspace{1cm} (2.11)

with \(P(a, x)\) the incomplete Gamma function. The \epsilon dependence of this expression is \(e^{-\epsilon}/\epsilon\), which is not integrable over the range of \epsilon. The cause of this problem is the choice of compact sets (2.5).

Fortunately it is not difficult to find a sequence of compact sets that lead to a proper posterior. A hint is provided by the fact that the Jeffreys prior (2.2) yields a “density” \(\epsilon \frac{d\sigma}{\sqrt{\epsilon \sigma + \mu}}\) that is invariant under scale transformations \(\epsilon \rightarrow c \epsilon, \sigma \rightarrow \sigma/c\), where \(c\) is constant. Surprisingly, this property is not shared by the prior (2.8), even though the latter also depends on \sigma through the product \(\epsilon \sigma\). This suggests that the compact sets
should be constructed in such a way that they respect the scale invariance of Jeffreys’ prior. Accordingly, we set:

\[ \Lambda_i = \left\{ (\sigma, \epsilon, \mu) : \sigma \in [0, u_i/\epsilon], \epsilon \in [1/v_i, v_i], \mu \in [0, w_i] \right\}, \]  

(2.12)

where \( u_i, v_i, \) and \( w_i \) are as before. Again using eqs. (2.2), (2.3), and (2.4), we now find:

\[ \pi_{R1}(\sigma | \epsilon, \mu) \propto \frac{\epsilon}{\sqrt{\epsilon \sigma + \mu}}, \]  

(2.13)

which is identical to Jeffreys’ prior for this problem and yields well-behaved posteriors. For future use, the subscript \( R1 \) on the left-hand side indicates that this reference prior was obtained with Method 1.

2.3 Method 2 Applied to the Single-Count Model

In contrast with Method 1, Method 2 requires from the start that we specify a subjective prior for the effective integrated luminosity \( \epsilon \) and the background contamination \( \mu \). Furthermore, this specification must be done conditionally on the signal rate \( \sigma \), so that we need an expression for \( \pi(\epsilon, \mu | \sigma) \). Here we will assume that \( \epsilon \) and \( \mu \) are independent of \( \sigma \) and that their prior factorizes as a product of two gamma densities:

\[ \pi(\epsilon, \mu | \sigma) = \pi(\epsilon, \mu) = \frac{a(\alpha) x^{-1/2} e^{-ax} b(b) y^{-1/2} e^{-by}}{\Gamma(x + 1/2) \Gamma(y + 1/2)}, \]  

(2.14)

where \( a, b, x, \) and \( y \) are known constants. There are two ways of interpreting this prior. The first one is appropriate when information about \( \epsilon \) and \( \mu \) comes from one or more non-experimental sources, such as simulation studies and theoretical calculations, and takes the form of a central value plus an uncertainty. Since the \( \epsilon \) and \( \mu \) components of the prior are each modeled by a two-parameter density, one can fix the parameters \( (x, a, y, b) \) by identifying the means of the component distributions with the central values \( (\bar{\epsilon}, \bar{\mu}) \) of the measurements, and their coefficients of variation with the relative uncertainties \( (\delta\epsilon, \delta\mu) \):

\[ \bar{\epsilon} = \frac{x + \frac{1}{2}}{a}, \quad \delta\epsilon = \frac{1}{\sqrt{x + \frac{1}{2}}}, \quad \text{or} \quad x = \frac{1}{\delta\epsilon^2} - \frac{1}{2}, \quad a = \frac{1}{\epsilon \delta\epsilon^2}; \]  

(2.15)

\[ \bar{\mu} = \frac{y + \frac{1}{2}}{b}, \quad \delta\mu = \frac{1}{\sqrt{y + \frac{1}{2}}}, \quad \text{or} \quad y = \frac{1}{\delta\mu^2} - \frac{1}{2}, \quad b = \frac{1}{\mu \delta\mu^2}. \]  

(2.16)

It will then be necessary to check the robustness of the final analysis results to reasonable changes in this procedure. For example, one may want to replace the gamma distribution by a log-normal or truncated Gaussian one, and identify the central value of the measurement with the mode or median instead of the mean.

The second interpretation of prior (2.14) follows from the analysis of two independent, auxiliary Poisson measurements, in which the observed number of events is \( x \) for
the effective luminosity and $y$ for the background. The expected numbers of events in
these auxiliary measurements are $a\epsilon$ and $b\mu$, respectively. For a Poisson likelihood with
mean $a\epsilon$ the standard reference prior coincides with Jeffreys’ prior and is proportional
to $1/\sqrt{\epsilon}$. Given a measurement $x$, the posterior will then be a gamma distribution
with shape parameter $x + 1/2$ and scale parameter $1/a$. A similar result holds for
the background measurement. In this manner the prior (2.14) is obtained as a joint
reference posterior from two auxiliary measurements.

Although interesting, we take the second interpretation with a grain of salt. This
is because it implies that the data we have at our disposal is actually $(n, x, y)$ rather
than just $n$, and that our model is:

$$p(n, x, y \mid \sigma, \epsilon, \mu) = \left(\frac{\mu + \epsilon\sigma}{n!} e^{-\mu - \epsilon\sigma} \frac{(a\epsilon)^x}{x!} \frac{e^{-a\epsilon}}{y!} \right),$$

instead of (2.1). However, the reference priors for these two models are not the same.
Our position here is to assume that prior information is available about $\epsilon$ and $\mu$,
but we do not look too deeply into how it was obtained.

The next step in the application of Method 2 is to marginalize the probability
model (2.1) with respect to $\epsilon$ and $\mu$:

$$p(n \mid \sigma) = \int \int p(n \mid \sigma, \epsilon, \mu) \pi(\epsilon, \mu \mid \sigma) \, d\epsilon \, d\mu,$$

$$= \int \int \frac{(\sigma + \mu)^n}{n!} e^{-\epsilon\sigma - \mu} \frac{a(\epsilon\sigma)^{x-1/2}}{\Gamma(x+1/2)} e^{-a\epsilon} \frac{b(b\mu)^{y-1/2}}{\Gamma(y+1/2)} e^{-b\mu} \, d\epsilon \, d\mu,$$

$$= \left[ a \left( \frac{\sigma}{a + \sigma} \right)^{x+1/2} \left( \frac{b}{b + 1} \right)^{y+1/2} \sum_{k=0}^{n} u_{nk} \left[ \frac{\sigma}{a + \sigma} \right]^k \right],$$

where

$$u_{nk} = \left( x - \frac{1}{2} + k \right) \left( y - \frac{1}{2} + n - k \right) \left[ \frac{1}{b + 1} \right]^{n-k},$$

and we used generalized binomial coefficients:

$$\binom{v}{w} \equiv \frac{\Gamma(v + 1)}{\Gamma(w + 1) \Gamma(v - w + 1)}.$$  

The marginal pdf $p(n \mid \sigma)$ is shown as a function of $\sigma$ for several values of $n$ in Fig. 1.
Aside from being the starting point of the calculation of Method 2 reference priors,
this quantity will also be used to study the behavior of reference posteriors under
measurement replications (section 4), and to compute intrinsic intervals (section 5.2).

Finally, the reference prior algorithm must be applied to the marginalized model
$p(n \mid \sigma)$. Since this model involves a single, continuous parameter, the reference prior
coincides with Jeffreys’ prior; it can be written as:

$$\pi_{R2}(\sigma) \propto \sqrt{\mathbb{E}\left\{ \frac{d}{d\sigma} \ln p(n \mid \sigma) \right\}^2} \propto \sqrt{\sum_{n=0}^{\infty} \frac{[(x + 1/2) S_n^0 - (a/\sigma) S_n^1]^2}{(a + \sigma)^{x+5/2} S_n^0}},$$
with
\[ S^m_n \equiv \sum_{k=0}^{n} k^m u_{nk} \left( \frac{\sigma}{a + \sigma} \right)^k \quad \text{for} \quad m = 0, 1. \] (2.24)

We will use the notation \( \pi_{R2}(\sigma) \) to refer to the marginal reference prior for \( \sigma \) obtained with Method 2.

### 2.4 Marginal Priors and Posteriors

One way to compare the Method 1 and Method 2 reference priors calculated in the previous sections is to examine their \( \sigma \) marginals, since \( \sigma \) is the parameter of interest. For this purpose we will use the nuisance prior introduced in equation (2.14). As the starting point of Method 2 is the marginalized model (2.20), its marginals are already known: the prior is eq. (2.23), and the posterior is simply
\[ \pi_{R2}(\sigma | n) \propto p(n | \sigma) \pi_{R2}(\sigma). \] (2.25)

The normalization of the latter must be obtained numerically. Examples of the Method 2 prior and posterior are shown in Figures 3 and 6 respectively.

Method 1 requires more work. The marginal cross section prior is:
\[ \pi_{R1}(\sigma) \propto \int_{0}^{\infty} \int_{0}^{\infty} \epsilon \, \mu \, \frac{a(\epsilon)^{x-1/2} e^{-ax}}{\Gamma(x + \frac{1}{2})} \frac{b(\mu)^{y-1/2} e^{-b\mu}}{\Gamma(y + \frac{1}{2})}. \] (2.26)

To calculate this integral we first perform the parameter substitution
\[ (\epsilon, \mu) \rightarrow (\nu, \mu) \quad \text{where} \quad \nu \equiv \epsilon/\mu. \] (2.27)

The differential transforms according to \( d\epsilon \, d\mu = \mu \, d\nu \, d\mu \) and the integral becomes:
\[ \pi_{R1}(\sigma) \propto \frac{\Gamma(x + y + \frac{3}{2})}{\Gamma(x + \frac{1}{2}) \Gamma(y + \frac{1}{2})} \int_{0}^{\infty} \frac{d\nu}{\sqrt{\nu \sigma + 1}} \frac{(a\nu)^{x+1/2} (b\nu)^{y+1/2}}{(a\nu + b)^{x+y+3/2}}. \] (2.28)

Recognizing a Gauss hypergeometric function [1] in this expression, we have:
\[ \pi_{R1}(\sigma) \propto {\mbox 2F}_1 \left( \frac{1}{2}, x + \frac{3}{2}; x + y + 2; 1 - \frac{b}{a} \sigma \right). \] (2.29)

This prior is improper; indeed, using standard properties of the hypergeometric function, it can be rewritten as
\[ \pi_{R1}(\sigma) \propto \frac{\sqrt{a}}{b \sigma} {\mbox 2F}_1 \left( \frac{1}{2}, y + \frac{1}{2}; x + y + 2; 1 - \frac{a}{b \sigma} \right). \] (2.30)

Thus, as \( \sigma \rightarrow \infty \), \( \pi_{R1}(\sigma) \) decreases only as \( 1/\sqrt{\sigma} \). Equations (2.29) and (2.30) indicate that \( a \) and \( b \) are simple scaling parameters for the marginal prior, whereas \( x \) and \( y \) are shape parameters. Some examples of this prior are plotted in Figure 2 and compared with Method 2 priors in Figure 4.
The Method 1 marginal posterior for the cross section $\sigma$ is given by:

$$
\pi_{R1}(\sigma | n) \propto \int_0^\infty d\epsilon \int_0^\infty d\mu \frac{e(\epsilon \sigma + \mu)^{n-\frac{1}{2}}}{n!} \frac{a(\epsilon \sigma - \frac{1}{2}) e^{-a\epsilon}}{\Gamma(x + \frac{1}{2})} \frac{b(b\mu)^{y-\frac{1}{2}} e^{-b\mu}}{\Gamma(y + \frac{1}{2})}.
$$

(2.31)

Performing the parameter substitution (2.27), we obtain here:

$$
\pi_{R1}(\sigma | n) \propto \frac{\Gamma(n + x + y + \frac{3}{2})}{n! \Gamma(x + \frac{1}{2}) \Gamma(y + \frac{1}{2})} b^{y+\frac{1}{2}} \int_0^\infty d\nu \frac{(\nu \sigma + 1)^{n-\frac{1}{2}} (a(\nu + \sigma\mu + b + 1))^{x+y+n+\frac{1}{2}}}{(a\nu + \sigma\mu + b + 1)^{x+y+n+\frac{1}{2}}}.
$$

(2.32)

This can also be expressed with the help of a hypergeometric function:

$$
\pi_{R1}(\sigma | n) = K 2F_1\left(\frac{1}{2} - n, x + \frac{3}{2}; x + y + 2, \frac{1-b\sigma/a}{1+\sigma/a}\right),
$$

(2.33)

where $K$ is a normalization constant. To compute it, we first integrate the right-hand side of (2.31) over $\sigma$; the integral over $\epsilon$ is then trivial and that over $\mu$ yields an incomplete beta function. The final result is:

$$
K = \left(\frac{x + y + n + \frac{1}{2}}{n - \frac{1}{2}}\right) \frac{x + \frac{1}{2}}{a} \frac{b^{y+\frac{1}{2}}}{(1+b)^{x+y+n}} \frac{1}{\Gamma(1/a) \Gamma(1+1/a) \Gamma(1+b) \Gamma(1+b + 1)}.
$$

(2.34)

where

$$
I_z(u, v) \equiv \frac{\Gamma(u+v)}{\Gamma(u) \Gamma(v)} \int_0^1 t^{u-1} (1-t)^{v-1} dt.
$$

(2.35)

Examples of the marginal Method-1 posterior are shown in Figure 5. Note the interesting feature that for $n = 0$ the order of the curves labeled $y = 0, y = 1, \text{ and } y = 10$ is inverted at small $\sigma$ with respect to the case $n > 0$ (left-hand side plots).

### 2.5 Generalizations of the Single-Count Model

A straightforward and common generalization of the single-count model is the multi-bin model (e.g. histograms), based on the likelihood:

$$
p(\vec{\epsilon} | \sigma, \vec{\mu}) = \prod_{i=1}^M \frac{\mu_i + \epsilon_i \sigma}{n_i!} e^{-\mu_i - \epsilon_i \sigma}.
$$

(2.36)

To obtain the Method 1 reference prior for this model, we first calculate Jeffreys’ prior:

$$
\pi_J(\sigma | \vec{\epsilon}, \vec{\mu}) = \sqrt{\sum_{i=1}^M \frac{\epsilon_i^2}{\mu_i + \epsilon_i \sigma}}.
$$

(2.37)

Since it is improper, we need to apply the compact support argument described in section 2.2. We note that Jeffreys’ prior density $\pi_J(\sigma | \vec{\epsilon}, \vec{\mu}) d\sigma$ is invariant under the
transformation that maps $\epsilon_i$ to $c\epsilon_i$ and $\sigma$ to $\sigma/c$, with $c$ a constant. Hence we construct compact sets that respect this scale invariance:

$$\Lambda_\ell = \left\{ (\sigma, \vec{\epsilon}, \vec{\mu}) : \sigma \in [0, u_\ell / \epsilon^+] \, , \, \epsilon_i \in [1/v_i \ell, v_i \ell] \, , \, \mu_i \in [0, w_i \ell] \, , \, i = 1, \ldots, M \right\},$$

(2.38)

where $\epsilon^+ \equiv \sum_{i=1}^M \epsilon_i$, and $\{u_\ell\}$, $\{v_i \ell\}$, and $\{w_i \ell\}$ are increasing sequences of constants with respect to the index $\ell$. The remainder of the calculation is similar to that for the single-count model and concludes that here too the Method 1 reference prior equals Jeffreys’ prior. Note that we could have used a different sequence of compact sets here, one that generalizes (2.5) instead of (2.12). Had we done this, we would have obtained a different form for the Method 1 reference prior; we have not checked whether this alternate form leads to improper posteriors, as in the case of the single-count model. Our choice of (2.38) is solely motivated by consistency with (2.12).

We have not attempted to obtain analytical expressions for the marginal prior and posterior of the multi-bin model.

Further generalizations include cases where the $\mu_i$ and $\epsilon_i$ are correlated across bins, as well as unbinned likelihoods. To handle these more complicated situations we are developing numerical algorithms in collaboration with Harrison Prosper and Supriya Jain [10]. For Method 1 the algorithm is as follows:

1. Set $\vec{n}_o$ to the array of observed event numbers.
2. For $i = 1, \ldots, I$:
   3. Generate $(\sigma_i, \vec{\epsilon}_i, \vec{\mu}_i) \sim p(\vec{n}_o | \sigma, \vec{\epsilon}, \vec{\mu}) \pi(\vec{\epsilon}, \vec{\mu})$.
   4. For $j = 1, \ldots, J$:
      5. Generate $\vec{n}_j \sim p(\vec{n} | \sigma_i, \vec{\epsilon}_i, \vec{\mu}_i)$.
      6. Calculate $d^2[-\ln p(\vec{n}_j | \sigma_i, \vec{\epsilon}_i, \vec{\mu}_i)]/d\sigma^2_i$ by numerical differentiation.
5. Average the $J$ values of $d^2[-\ln p(\vec{n} | \sigma_i, \vec{\epsilon}_i, \vec{\mu}_i)]/d\sigma^2_i$ obtained at line 6, and take the square root. This yields a numerical approximation to the conditional Jeffreys’ prior $\pi(\sigma_i | \vec{\epsilon}_i, \vec{\mu}_i)$.
8. Histogram the $\sigma_i$ values generated at line 3, weighing them by $\pi(\sigma_i | \vec{\epsilon}_i, \vec{\mu}_i)/p(\vec{n}_o | \sigma_i, \vec{\epsilon}_i, \vec{\mu}_i)$. This yields $\pi_{R1}(\sigma)$, the $\sigma$-marginal prior.
9. Histogram the $\sigma_i$ values generated at line 3, weighing them by $\pi(\sigma_i | \vec{\epsilon}_i, \vec{\mu}_i)$. This yields $\pi_{R1}(\sigma | \vec{n}_o)$, the $\sigma$-marginal posterior.

The generation step at line 3 is done via a Markov chain Monte Carlo procedure. The particular choice of parent distribution for the generated $(\sigma, \vec{\epsilon}, \vec{\mu})$ triplets is motivated by the desire to obtain weights with reasonably small variance at steps 8 and 9. However, we discovered that the sampling density $p(\vec{n}_0 | \sigma, \vec{\epsilon}, \vec{\mu}) \pi(\vec{\epsilon}, \vec{\mu})$ is not always proper with respect to $(\sigma, \vec{\epsilon}, \vec{\mu})$. When $M = 1$ for example (single-count model), the distribution is improper if $x \leq 1/2$. Propriety can then be restored by multiplying the sampling density by $\epsilon$. Another feature of the above algorithm is that it does not implement the compact support argument. In the cases that we examined, this argument
made no difference, but this may not be true for all problems our code seeks to solve. Unfortunately the current lack of guidelines in the choice of compact sets limits our ability to address this issue in the code.

The algorithm for Method 2 is much simpler, since all it does is calculate Jeffreys’ prior on a marginalized likelihood provided by the user. In addition, it does not require a compact support argument. We will not discuss it further in this paper.

In the following sections we consider only the single-count model.

### 3 Upper Limits

A common way to summarize posterior distributions is via the computation of intervals. Figures 7 and 8 show some 68% and 95% credibility level upper limits and central intervals for the signal cross section $\sigma$ as a function of the observed number of events $N_{\text{obs}}$. These were obtained with the Method 1 reference posterior (2.33). As $N_{\text{obs}}$ increases, the interval boundaries become straight lines. There is no difference between Method 1 and Method 2 on the scale of these figures. Figure 9 provides a closer look at the difference between the two methods when $N_{\text{obs}}$ is small. The difference increases with the credibility level, but goes down to zero as $N_{\text{obs}}$ becomes large. This latter behavior is expected since the likelihood dominates inferences at large $N_{\text{obs}}$.

Figure 10 illustrates the variation of Method 1 posterior upper limits as a function of the prior mean background $\mu$, when the prior relative uncertainty (i.e. the coefficient of variation) $\delta\mu$ on the background is kept constant. These quantities are related to the prior parameters $y$ and $b$ according to eq. (2.16). In these plots the prior effective luminosity $\epsilon$ is assigned a mean of 1 and a relative uncertainty identical to that on the background ($\bar{\epsilon} = 1, \delta\epsilon = \delta\mu$). These plots show that once the mean background exceeds the observed number of events, the upper limits vary little. This is to be contrasted with the physically undesirable behavior of Feldman-Cousins limits, which decrease quickly with the mean background, even when no events are observed (see for example Figure 2 in [12]).

Another interesting feature of Fig. 10 is that for $n = 0$ the upper limit increases with the mean background, whereas for $n > 0$ it decreases. This behavior is directly related to the corresponding inversion of the $y$-labeled posterior densities in Fig. 5, mentioned at the end of section 2.4. It is due to the form of the reference prior for $\sigma$.

A comparison of the left and right panels of Figure 10 also demonstrates how upper limits increase with the uncertainties on background and effective luminosity.

### 4 Coverage

A well known property of subjective Bayesian intervals is that they cover exactly, when the coverage (a function of the parameters) is averaged over the prior. Similarly, frequentist intervals have exact credibility when the latter (a function of the observed data) is averaged over the prior-predictive distribution. This is a straightforward consequence of the law of total probability [15].
For a Bayesian, averaging over subjective priors is very natural since such priors provide proper, data-independent measures over parameter space. The same is not necessarily true for objective priors however. Thus, to calibrate an inference about a parameter with an objective prior, one option for Bayesians is to check how the inference behaves under replication of the measurement, i.e. its pointwise coverage, as opposed to average coverage. For measurements that combine subjective and objective priors, the logical extension of these procedures is to average over the subjective priors and check pointwise coverage with respect to the remaining parameters. Reference [8] writes:

With the rapid advances in computational techniques for Bayesian statistics that exploit the increased computing power now available, researchers are able to adopt more realistic, and usually more complex, models. However, it is then less likely that the statistician will be able to properly elicit prior beliefs about all aspects of the model. Moreover, many parameters may not have a direct interpretation. This suggests that there is a need to develop general robust methods for prior specification that incorporate both subjective and nonsubjective components. In this case, the matching property could be recast as being the approximate equality of the posterior probability of a suitable set and the corresponding frequentist probability averaged over the parameter space with respect to any continuous prior that preserves the subjective element of the specified prior. [pg. 27]

In our case, adopting this procedure means that we will calculate the coverage with respect to the marginalized data pdf \( p(n | \sigma) \); this is equation (2.20) for the case of the single-count model.

We start by calculating posterior upper limits and intervals on \( \sigma \) for an observed number of events \( n \) going from 0 to 200. The sum of \( p(n | \sigma) \) over this range of \( n \) values equals 1 within a fraction of a percent, for \( \sigma \) as high as 25 and for the \( x, y, a, b \) values used in the figures. For a given true value of \( \sigma \), the coverage of a set of intervals is defined as the sum of \( p(n | \sigma) \) over all \( n \) values for which \( \sigma \) is bracketed by the corresponding interval. Figures 11, 12, and 13 show the coverage of the Method 1 and Method 2 intervals plotted in Figures 7 and 8. Due to the large number of discontinuities in the coverage curves, the coverage was evaluated at 1001 equidistant values of \( \sigma \in [0, 25] \) and plotted as a set of unconnected points. As expected, central intervals undercover near \( \sigma = 0 \) since they always exclude that value, and upper limits overcover there since they always include it. The coverage of a 68% upper limit or interval settles much faster around its nominal level than the coverage of a 95% upper limit or interval.

5 Intrinsic intervals

Intrinsic intervals are regions of parameter space with lowest posterior loss, for a special type of loss function [6]. In brief, suppose we have one observation \( x \) from a model \( p(x | \theta) \) and we are interested in the true value of \( \theta \). Let \( \ell(\theta_0, \theta) \) be the loss suffered if the value \( \theta_0 \) is used as a proxy for the unknown true value of \( \theta \) in some application of
our measurement result. The posterior expected loss from using $\theta_0$ is then:

$$l(\theta_0 | x) = \mathbb{E}_{\theta|x} \left[ \ell(\theta_0, \theta) \right] = \int d\theta \, \ell(\theta_0, \theta) \, \pi(\theta | x).$$  \hfill (5.1)$$

The idea is to use the $\theta$ value with lowest expected loss as a point estimate of $\theta$, and regions of lowest expected loss as interval estimates.

Reference analysis proposes to use intrinsic discrepancy to define the loss function. In its general form, the intrinsic discrepancy loss function is:

$$\delta\{\theta_0, \theta\} = \min \left\{ \int dx \, p(x | \theta_0) \ln \frac{p(x | \theta_0)}{p(x | \theta)}, \int dx \, p(x | \theta) \ln \frac{p(x | \theta)}{p(x | \theta_0)} \right\}. \hfill (5.2)$$

If nuisance parameters $\nu$ are present, the definition of $\delta\{\theta_0, \theta\}$ includes an additional minimization:

$$\delta\{\theta_0, (\theta, \nu)\} = \inf_{\nu_0} \left\{ \min \left\{ \int dx \, p(x | \theta_0, \nu_0) \ln \frac{p(x | \theta_0, \nu_0)}{p(x | \theta, \nu)}, \right. \right.$$

$$\left. \left. \int dx \, p(x | \theta, \nu) \ln \frac{p(x | \theta, \nu)}{p(x | \theta_0, \nu_0)} \right\} \right\}. \hfill (5.3)$$

When compared with other loss functions, intrinsic loss has the advantage of being invariant under one-to-one transformations of the parameter(s), under one-to-one transformations of the data, and under reduction of the data by sufficiency.

Unfortunately Ref. [6] does not discuss situations like ours, where (1) the parameter of interest is not identifiable via the model (only a specific combination of interest and nuisance parameters is identifiable), and (2) the nuisance parameters are constrained by subjective priors. If we were to minimize the intrinsic discrepancy loss over the nuisance parameters ($\nu_0$) we would obtain zero, because any difference between $p(x | \theta_0, \nu_0)$ and $p(x | \theta, \nu)$ in the log-likelihood ratios of eq. (5.3) can always be exactly cancelled by an appropriate change in $\nu_0$.

In principle there are two ways of solving this problem, which we shall label “conditional” and “marginal”, respectively.

1. In the conditional approach the loss $\ell((\theta, \nu), (\theta_0, \nu))$ is evaluated from the full model $p(x | \theta, \nu)$ by assuming that the nuisance parameters $\nu$ are exactly known. The expected loss is then obtained by averaging $\ell((\theta, \nu), (\theta_0, \nu))$ over the posterior $\pi(\theta, \nu | x)$.

2. In the marginal approach the loss $\ell(\theta, \theta_0)$ is evaluated from the marginal model $p(x | \theta) = \int d\nu \, p(x | \theta, \nu) \pi(\nu)$ and is therefore independent of nuisance parameters. The expected loss is then obtained by integrating the loss function over the marginal posterior $\pi(\theta | x) = \int d\nu \, \pi(\theta, \nu | x)$.

A possible objection against the conditional approach is that we may not want the loss suffered from using the wrong value of $\theta$ to depend on nuisance parameters that are conceptually independent of $\theta$, and that are, furthermore, unknown.

We now apply these two approaches to our single-count Poisson model with gamma priors.
5.1 Conditional approach

Here we work with a simple Poisson model:

\[ p(n \mid \sigma, \mu, \epsilon) = \frac{(\mu + \epsilon \sigma)^n e^{-\mu - \epsilon \sigma}}{n!}, \]  

(5.4)

where \( \epsilon \) and \( \mu \) are assumed known. The intrinsic loss is:

\[ \delta\{\sigma_0, \sigma\} = \min\{\kappa\{\sigma_0 \mid \sigma\}, \kappa\{\sigma \mid \sigma_0\}\}; \]  

(5.5)

where:

\[ \kappa\{\sigma_0 \mid \sigma\} = \sum_{n=0}^{\infty} \frac{(\mu + \epsilon \sigma)^n e^{-\mu - \epsilon \sigma}}{n!} \ln \frac{(\mu + \epsilon \sigma)^n e^{-\mu - \epsilon \sigma}}{(\mu + \epsilon \sigma_0)^n e^{-\mu - \epsilon \sigma_0}}, \]  

(5.6)

\[ = \sum_{n=0}^{\infty} \frac{(\mu + \epsilon \sigma)^n e^{-\mu - \epsilon \sigma}}{n!} \left[n \ln \frac{\mu + \epsilon \sigma}{\mu + \epsilon \sigma_0} + \epsilon (\sigma_0 - \sigma)\right], \]  

(5.7)

\[ = (\mu + \epsilon \sigma) \ln \frac{\mu + \epsilon \sigma}{\mu + \epsilon \sigma_0} + \epsilon (\sigma_0 - \sigma). \]  

(5.8)

To compute the minimum (5.5), we need to find out for what values of \( \sigma \) we have \( \kappa\{\sigma_0 \mid \sigma\} < \kappa\{\sigma \mid \sigma_0\} \), or:

\[ (\mu + \epsilon \sigma) \ln \frac{\mu + \epsilon \sigma}{\mu + \epsilon \sigma_0} + \epsilon (\sigma_0 - \sigma) < (\mu + \epsilon \sigma_0) \ln \frac{\mu + \epsilon \sigma_0}{\mu + \epsilon \sigma} + \epsilon (\sigma - \sigma_0), \]  

(5.9)

or

\[ g(\sigma) \equiv [2\mu + \epsilon(\sigma + \sigma_0)] \ln \frac{\mu + \epsilon \sigma}{\mu + \epsilon \sigma_0} + 2\epsilon (\sigma_0 - \sigma) < 0. \]  

(5.10)

To figure this out, note that:

\[ g'(\sigma) = \epsilon \ln \frac{\mu + \epsilon \sigma}{\mu + \epsilon \sigma_0} + \frac{\epsilon^2 (\sigma_0 - \sigma)}{\mu + \epsilon \sigma}, \]  

(5.11)

\[ g''(\sigma) = -\frac{\epsilon^3 (\sigma_0 - \sigma)^2}{(\mu + \epsilon \sigma)^2}. \]  

(5.12)

Hence when \( \sigma < \sigma_0, g''(\sigma) < 0 \) so that \( g'(\sigma) \) is decreasing. Combined with the fact that \( g'(\sigma_0) = 0 \), this implies that \( g'(\sigma) > 0 \). Thus, \( g(\sigma) \) increases for \( \sigma < \sigma_0 \), and since it is zero at \( \sigma_0 \), it must be negative for \( \sigma < \sigma_0 \). Therefore, \( \delta\{\sigma_0, \sigma\} = \kappa\{\sigma_0 \mid \sigma\} \) for \( \sigma < \sigma_0 \). By symmetry we have \( \delta\{\sigma, \sigma_0\} = \kappa\{\sigma \mid \sigma_0\} \) for \( \sigma > \sigma_0 \). We conclude that

\[ \delta\{\sigma_0, \sigma \mid \mu, \epsilon\} = \epsilon |\sigma_0 - \sigma| - \left[\mu + \epsilon \min(\sigma, \sigma_0)\right] \ln \frac{\mu + \epsilon \sigma}{\mu + \epsilon \sigma_0}, \]  

(5.13)

where we reintroduced the nuisance parameters \( \mu \) and \( \epsilon \) in the argument list of \( \delta \). The posterior expected intrinsic loss is then

\[ d(\sigma_0 \mid n) = \int d\sigma \int d\mu \int d\epsilon \delta\{\sigma_0, \sigma \mid \mu, \epsilon\} \pi(\sigma, \mu, \epsilon \mid n). \]  

(5.14)

Because of the previously mentioned objection we do not pursue this approach here.


5.2 Marginal approach

Here we use the marginal model $p(n | \sigma)$ of equation (2.20). The intrinsic loss is

$$\delta\{\sigma_0, \sigma\} = \min\left\{ \sum_{n=0}^{\infty} p(n | \sigma_0) \ln \frac{p(n | \sigma_0)}{p(n | \sigma)}, \sum_{n=0}^{\infty} p(n | \sigma) \ln \frac{p(n | \sigma)}{p(n | \sigma_0)} \right\}, \quad (5.15)$$

and the posterior expected intrinsic loss is

$$d(\sigma_0 | n) = \int d\sigma \, \delta\{\sigma_0, \sigma\} \pi(\sigma | n), \quad (5.16)$$

where $\pi(\sigma | n)$ is the marginal posterior, obtained via either Method 1 or Method 2. Neither of the above expressions can be further simplified, but they can be coded numerically.

Figure 14 illustrates the construction of minimal reference posterior expected intrinsic loss intervals for the single-count model. The reference posterior used is that from Method 1 and is shown by the dashed lines. The solid lines indicate the expected loss as a function of $\sigma$. Starting with the point of lowest expected loss, one adds $\sigma$ values with increasing expected loss until the credibility of all the included points equals the desired level. For this figure we set both the background and effective luminosity to $1.0 \pm 0.2$. When the number $N_{obs}$ of observed events equals 1, the left plot shows that the 95, 99, and 99.9% intrinsic intervals are upper limits, whereas they are two-sided when $N_{obs} = 10$ (right plot).

The left-hand plot of Figure 15 shows the boundaries of the 95% credibility intrinsic intervals as a function of $N_{obs}$. For small $N_{obs}$ the intervals are upper limits, and they become two-sided as $N_{obs}$ increases. At large $N_{obs}$ the boundaries follow straight lines. The corresponding frequentist coverage is shown in the right-hand plot. Since intrinsic intervals transition smoothly from one-sided to two-sided, their coverage avoids the extremes of 100% and 0%. Excursions from the nominal level decrease as $\sigma$ increases.

6 Reference Analysis Search Procedures

We now return to the search procedure discussed in the introduction. As we saw there, one of the difficulties in the frequentist approach is the interaction between the hypothesis test and the subsequent interval construction. This problem is absent in the Bayesian approach, which is based on credibility rather than coverage. Using tools from reference analysis, the search procedure becomes then:

1. Choose three credibility levels $\alpha_1$, $\alpha_2$, and $\alpha_3$.

2. Compute the marginal reference posterior for the parameter of interest, say $\sigma$.

3. Compute an $\alpha_1$-credibility intrinsic interval $S$ for $\sigma$.

This terminology is a mouthful; we will try to abbreviate it in the following.
4. If interval $S$ does not include the value $\sigma = 0$, claim discovery and compute an $\alpha_2$-credibility intrinsic interval on $\sigma$.

5. If interval $S$ includes $\sigma = 0$, make no claim and compute an $\alpha_3$-credibility upper limit on $\sigma$.

The motivation for the post-test interval calculations (steps 4 and 5) is the same as in the original procedure. However, in the Bayesian reference analysis framework these calculations avoid the problems that plagued the frequentist approach: flip-flopping, empty intervals, upper limit sensitivity, and treatment of nuisance parameters.

For an example consider Figure 14. Suppose we expect a background of $1.0 \pm 0.2$ and an effective luminosity of $1.0 \pm 0.2$, and set $\alpha_1 = 0.1\%$. If we then observe 1 event, we conclude from the left-hand plot that our observation is consistent with background. On the other hand, if we observe 10 events, the right-hand plot shows that this is significant at the $0.1\%$ level, and we can proceed to the calculation of narrower, more informative two-sided intervals on $\sigma$.

Physicists often like to quote a significance. Within a given class of tests, significance can be defined in general as the discovery threshold of the most stringent test for which the observed data rejects the null hypothesis. The classical formulation of this concept is the $p$ value. Although the latter is quite popular in high energy physics, it cannot be justified from a strictly Bayesian point of view. To remedy this shortcoming, Refs. [3, 4, 5] propose another measure of significance, the Bayes Reference Criterion (BRC). This is essentially the reference posterior expected intrinsic loss evaluated at the null hypothesis $H_0 : \sigma = 0$. Using eq. (5.16) with the Method-1 reference posterior, this yields:

$$
BRC \equiv d(0 \mid n) = \int_0^\infty d\sigma \pi_{R1}(\sigma \mid n) \min\left\{\kappa(\sigma \mid 0), \kappa(0 \mid \sigma)\right\},
$$

(6.1)

where

$$
\kappa(\sigma_1 \mid \sigma_2) \equiv \sum_{n=0}^\infty p(n \mid \sigma_2) \ln \frac{p(n \mid \sigma_2)}{p(n \mid \sigma_1)},
$$

(6.2)

and $p(n \mid \sigma)$ is given by eq. (2.20). The BRC can be interpreted as the minimum posterior expected log-likelihood ratio in favor of the model that generated the data, and it therefore provides its own calibration of the evidence against the null hypothesis: $BRC=\ln(10) \approx 2.3$ is mild evidence, $BRC=\ln(100) \approx 4.6$ is strong, and $BRC=\ln(1000) \approx 6.9$ is decisive. In contrast with $p$ values, the BRC does not need adjusting for dimensionality and sample size. Returning to the example of Fig. 14, we obtain $BRC=6.12$ for $N_{\text{obs}} = 10$ (right-hand plot). This represents strong, but somewhat less than decisive evidence against the background-only hypothesis. It is interesting to compare this result with the standard $p$ value used by physicists, which gives the probability for a background of $1.0 \pm 0.2$ events to fluctuate up to 10. Using a prior-predictive approach to the treatment of the background uncertainty, the result is $p_{PP} = 3.966 \times 10^{-7}$. One can convert this $p$ value into a “number of $\sigma$’s”, or a $Z$ value, i.e. the distance a normal variate is from zero when the probability outside $\pm Z$
equals 2p:

\[ p = \int_{-\infty}^{+\infty} \frac{e^{-x^2/2}}{\sqrt{2\pi}} \, dx, \quad \text{or} \quad Z(p) = \sqrt{2} \text{erf}^{-1}(1 - 2p). \quad (6.3) \]

The above value of \( p_{PP} \) corresponds to \( Z_{PP} = 4.94 \), just short of the \( 5\sigma \) discovery threshold favored by physicists.

There are other Bayesian measures of significance. Following Lindley [13], one can ask for the credibility \( \gamma \) of the widest highest-posterior-density (HPD) interval that does not contain the value \( \sigma = 0 \). A significance can then be defined as \( 1 - \gamma \). Using the right-hand plot of Figure 14 as example, we find that the posterior density at \( \sigma = 0 \) equals \( 9.72 \times 10^{-7} \); the interval of \( \sigma \) values with higher posterior density is \([0.0, 47.41[\) and has a posterior probability \( \gamma = 1 - 3.37 \times 10^{-6} \). The \( Z \) value corresponding to \( 1 - \gamma \) is \( Z_{HPD} = 4.50 \). Another possibility is to use intrinsic intervals instead of HPD ones. The set of \( \sigma \) values with expected loss lower than our observed BRC of 6.12 is the interval \([0.0, 33.0[\), which has a posterior credibility of 99.973%. The significance is therefore 0.027%, corresponding to \( Z_{BRC} = 3.46 \). That these various measures of significance give rather different results for our example should perhaps not be too surprising, since they simply answer different questions:

- **BRC**: What is the reference posterior expected intrinsic loss at \( \sigma = 0 \)?
- **\( Z_{BRC} \)**: What is the posterior probability for a larger expected loss than the one at \( \sigma = 0 \)?
- **\( Z_{HPD} \)**: What is the posterior probability for a smaller posterior density than the one at \( \sigma = 0 \)?
- **\( Z_{PP} \)**: What is the prior-predictive probability for the \( \sigma = 0 \) hypothesis to yield data at least as extreme as observed?

The behavior of BRC, \( Z_{BRC} \), \( Z_{HPD} \), and \( Z_{PP} \) is compared for the single-count model in Table 1. Also listed are the posterior expectation values of the directed Kullback-Leibler divergences:

\[ d_1 = \int_0^\infty d\sigma \, \pi_{R1}(\sigma \mid n) \, \kappa\{0 \mid \sigma\}, \quad (6.4) \]
\[ d_2 = \int_0^\infty d\sigma \, \pi_{R1}(\sigma \mid n) \, \kappa\{\sigma \mid 0\}. \quad (6.5) \]

In principle one could use \( d_1 \) or \( d_2 \) as a measure of significance. Note that BRC is not necessarily equal to the minimum of \( d_1 \) and \( d_2 \), even though this seems to be the case for the examples reported in the table, at least within the numerical accuracy of the calculations. From a simple geometrical argument one expects that \( \text{BRC} \leq \min\{d_1, d_2\} \).

In contrast with the \( p \) value, the Bayesian measures of evidence depend on both the signal model and the background model. However, BRC, \( d_1 \), and \( d_2 \) depend on the signal model only through the relative uncertainty \( \delta \epsilon \) on the effective luminosity.
Table 1: Examples of significance calculations for the single-count model. Columns 2 and 3 show the prior mean ± the prior standard deviation for background and effective luminosity. The other column headers are defined in the text. To help identify trends, line 2 is repeated at lines 5, 8, and 12.

(compare lines 2–4 with 5–7 in Table 1). This is a direct consequence of the fact that the posterior (2.33) and the marginal data model (2.20) only depend on $\sigma$ and $a$ through the ratio $\sigma/a$, and that the $\sigma$ integration in (6.1) goes from 0 to $\infty$. Thus, BRC depends on the signal model only through $x = 1/\delta \epsilon^2 - 1/2$.

It is interesting that $d_1$ and BRC behave oppositely under changes in the signal model: an increase in $\delta \epsilon$ makes BRC smaller and $d_1$ larger. To see which behavior is correct, consider that if our prior belief in our ability to detect signal has large uncertainties (e.g. due to poor knowledge of the effective luminosity), then the evidence provided by the data should be downgraded, making us less willing to claim discovery. Hence only BRC seems to behave correctly.

Lines 8 to 15 in the table illustrate the variation of significance measures with changes in the background model. Not surprisingly, changes in the background mean have more effect than corresponding changes in the background uncertainty. A more puzzling feature is that $d_1$, $Z_{HPD}$, and $Z_{PP}$ are much more sensitive to the background uncertainty than $d_2$, BRC, and $Z_{BRC}$.

The search procedure described at the beginning of this section is partially motivated by physicists’ fondness for the classical complementarity between hypothesis testing and interval construction. In terms of significance measures it corresponds to using $Z_{BRC}$, but this is not equivalent to using BRC (for example, $Z_{BRC}$ depends on both $x$ and $a$, whereas BRC only depends on $x$). Further study is needed to determine which measure of significance best serves the needs of high energy physics research.
7 Open Questions

Reference prior methodology has now been around for about 30 years, during which it has been continuously refined by attempts to apply it in an increasing number of situations. One result of these refinements is the so-called compact support argument, which we discussed in sections 2.2 and 2.5. As we showed, application of this argument is not always trivial, and one would welcome a more deterministic approach to the construction of the required compact sets. This would be especially useful for coding purposes.

Experimental high energy physicists spend a considerable amount of time trying to obtain useable information about the nuisance parameters in their measurements, and they are generally successful in this, to varying degrees. It is therefore somewhat surprising that most of the reference analysis literature deals with nuisance parameters about which nothing is known a priori. One of the few exceptions is Ref. [17]. However, there remain some open issues, as we indicated in section 5, with the treatment of partially known nuisance parameters in the definition of intrinsic loss. We proposed two approaches, labeled conditional and marginal. Most of our work on the single-count model was done with the marginal approach, but we wonder whether a case could be made for using the conditional approach, in spite of the objection mentioned in section 5.

In section 6 we discovered that some care is needed in the selection of a Bayesian test procedure, because the many measures of significance that one can think of tend to disagree rather sharply. References [3, 4, 5] lay out strong theoretical arguments for using the BRC. What is still needed is an argument for selecting an appropriate discovery threshold based on this criterion, and this hinges on a proper understanding of the various effects that physicists are concerned about. For example, does the look-elsewhere effect, which motivates the 5σ threshold on \( p \) values, affect the BRC, and if so, how? Next, we need to build confidence in the behavior of BRC with respect to uncertainties on the background and signal models. The numerical examples discussed in section 6 indicate that this behavior may be correct, but one ought to verify this more generally.

Acknowledgments

I wish to thank Supriya Jain and Harrison Prosper for many meetings and discussions on the topic of reference analysis. Their Monte Carlo calculations helped check some of the analytical results presented here. Thanks also to Harrison for suggesting the marginal approach to intrinsic intervals in section 5.2.
Figure 1: Marginalized likelihood of equation (2.20) as a function of the signal cross section $\sigma$, for several values of the observed number of events $n$. 

\[ x = y = \sigma = b = 1 \]
Figure 2: Method 1 reference prior for the cross section \( \sigma \), for three values of the shape parameter \( y \) of the background prior (left), and for four values of the shape parameter \( x \) of the efficiency prior (right). The priors are normalized by the requirement that \( \pi_{R1}(1) = 1 \).

Figure 3: Marginal Method 2 reference prior for the cross section \( \sigma \), for three values of the shape parameter \( y \) of the background prior (left), and for four values of the shape parameter \( x \) of the efficiency prior (right). The priors are normalized by the requirement that \( \pi_{R2}(1) = 1 \).
Figure 4: Comparison of Method 1 and Method 2 reference priors for various values of $x$ and $y$. Both types of prior are normalized to 1 at $\sigma = a/b$. 
Figure 5: Marginal Method 1 reference posterior for the cross section $\sigma$, for three values of the shape parameter $y$ of the background prior (left), and for three values of the shape parameter $x$ of the efficiency prior (right), as the number of observed events $n$ goes from 0 to 10 (top to bottom).
Figure 6: Method 2 reference posterior for the cross section $\sigma$, for three values of the shape parameter $y$ of the background prior (left), and for three values of the shape parameter $x$ of the efficiency prior (right), as the number of observed events $n$ goes from 0 to 10 (top to bottom).
Figure 7: Upper limits and central intervals at the 68 and 95% credibility levels, calculated from Method 1 posteriors. The prior for the background $\mu$ and the effective luminosity $\epsilon$ is given by eq. (2.14), with $x = y = a = b = 1$; this corresponds to a mean of 1.5 and a coefficient of variation of 82% for both $\mu$ and $\epsilon$. Method 2 intervals are indistinguishable from Method 1 on this scale.
Figure 8: Upper limits and central intervals at the 68 and 95% credibility levels, calculated from Method 1 posteriors. The prior for the background $\mu$ and the effective luminosity $\epsilon$ is given by eq. (2.14), with $x = y = 24.5$ and $a = b = 25.0$; this corresponds to a mean of 1.0 and a coefficient of variation of 20% for both $\mu$ and $\epsilon$. Method 2 intervals are indistinguishable from Method 1 on this scale.
Figure 9: Difference between the Method 1 and Method 2 upper limits at the 68 and 95% credibility levels.

Figure 10: Variation of the Method 1 reference posterior upper limit with mean background for several values of the observed number of events $n$. The relative uncertainty on the background and on the effective luminosity is 20% for the left plot and 50% for the right one.
Figure 11: Method 1 reference posterior: coverage of 68% credibility central intervals (top left), 68% upper limits (top right), 95% central intervals (bottom left), and 95% upper limits (bottom right) for the cross section $\sigma$, as a function of $\sigma$. The coverage is averaged over the prior for the effective luminosity $\epsilon$ and the background $\mu$. This prior is given by eq. (2.14), with $x = y = a = b = 1$. Note the offset zero in the bottom plots. The solid horizontal lines indicate the credibility level of the constructions.
Figure 12: Method 2 reference posterior: coverage of 68% credibility central intervals (top left), 68% upper limits (top right), 95% central intervals (bottom left), and 95% upper limits (bottom right) for the cross section $\sigma$, as a function of $\sigma$. The coverage is averaged over the prior for the effective luminosity $\epsilon$ and the background $\mu$. This prior is given by eq. (2.14), with $x = y = a = b = 1$. Note the offset zero in the bottom plots. The solid horizontal lines indicate the credibility level of the constructions.
Figure 13: Method 1 reference posterior: coverage of 68% credibility central intervals (top left), 68% upper limits (top right), 95% central intervals (bottom left), and 95% upper limits (bottom right) for the cross section $\sigma$, as a function of $\sigma$. The coverage is averaged over the prior for the effective luminosity $\epsilon$ and the background $\mu$. This prior is given by eq. (2.14), with $x = y = 24.5$ and $a = b = 25.0$. Note the offset zero in the bottom plots. The solid horizontal lines indicate the credibility level of the constructions.
Figure 14: Construction of 95%, 99%, and 99.9% credibility intrinsic intervals for a new particle cross section when 1 event (left) or 10 events (right) are observed over a background of $1.0 \pm 0.2$ events and with an effective luminosity of $1.0 \pm 0.2$ (this corresponds to $x = y = 24.5$ and $a = b = 25$). The dashed curves represent the Method 1 reference posteriors (rescaled by a factor of 5 and 50, respectively) and the solid curves show the posterior expected intrinsic loss.

Figure 15: Left: boundaries of minimum posterior expected intrinsic loss intervals with 95% credibility as a function of the observed number of events. Right: frequentist coverage of these intervals as a function of the true value of the cross section $\sigma$ (note the offset zero on the $y$ axis).
References


[7] Cousins, R. D., private communication. 5


