Fundamental Issues in Statistical Detection of Physical Phenomena

P. S. Shawhan
LIGO Laboratory, California Institute of Technology, Pasadena, CA 91125, USA

Many physics experiments are designed to search for some rare, previously unseen phenomenon which would leave a distinctive event signature in the detector. Generally, there are one or more "background" processes which can mimic the signature, so that detecting the new phenomenon is a matter of observing significantly more events than would be expected from background processes. However, the Bayesian approach to credible interval construction does not, in itself, address the question of whether a given excess should be interpreted as a "detection" of the phenomenon. Unified frequentist (*e.g.* Feldman-Cousins) approaches to confidence interval construction dictate when the interval should exclude zero, but are rarely (if ever) calculated using a high confidence level that would be appropriate for making a detection. The standard quantitative way to judge the significance of an apparent signal (in excess of the expected background) is to calculate the *p*-value for the null hypothesis.

1. PREFACE

The goal of this article is to review some important fundamental issues—philosophical, not technical—which arise when interpreting the results of a search for a physical phenomenon which has not yet been observed. We will discuss these issues, starting at the most basic level, in the context of very familiar, straightforward statistical analysis approaches, and will point out that the question of what constitutes a "detection" is not directly addressed by these approaches.

2. "DETECTION" AND UPPER LIMITS IN PHYSICS EXPERIMENTS

A physics experiment typically produces a large amount of data, which needs to be distilled down to something meaningful. In other words, the outcome of an experiment calls for an interpretation. The usual approach is to discard most of the details in the raw data and construct a simple "statistic" to summarize the data, usually a scalar quantity, such as the number of events satisfying a set of selection criteria. For a given choice of statistic, X, the full information content of the experimental result consists of the observed value of the statistic, x, along with the probability function $P(X|\theta)$ which describes the probability of observing any given value of the statistic as a function of one or more imperfectly known physical parameters, denoted by θ. This information content is completely objective, assuming that the decision about what statistic to construct was made without reference to the experimental data.[1]

The final step in an analysis is to extract a physical interpretation from $P(X|\theta)$ and x, normally to infer favored ranges for the values of the physical parameter(s) θ. This may be done in a frequentist sense, by defining an "acceptance region" (of "likely" experimental outcomes) in $P(X|\theta)$ and using the Neyman construction [1] to calculate the resulting confidence interval given the observed value x. Or it may be done in a Bayesian sense, by folding together one's prior belief about θ with the likelihood function $L(\theta) \equiv P(x|\theta)$ to arrive at a posterior probability density function (pdf), and perhaps then go on to derive a credible interval from that pdf. Either approach to interpretation involves a choice about how to define the interval, as we shall discuss further in the later sections of this article.

Many physics experiments are designed to try to detect a distinct signature in the detector from "new physics", some hypothesized physical phenomenon which has not previously been observed. Current examples include the Higgs boson and gravitational waves; past examples have included the top quark, *CP* violation in *B* mesons, etc. In some cases, there are good theoretical reasons, or indirect information from other experiments, to believe that the effect exists, and there may even be an estimate of its magnitude or event rate.[2] In other cases, the magnitude of the effect is unknown, and may even be unmeasurably small or nonexistent. Generally, there are one or more "background" processes which can mimic the signature in the detector, so that detecting the new phenomenon is a matter of observing significantly more events than would be expected from background processes. Of course, even if the *average* magnitude of the background is known accurately, the statistical analysis must allow for fluctuations.

A common aspect of searches for "new physics" is that physicists generally take a conservative approach (in a sociological sense, not the statistical sense) to claiming a "detection". In other words, they require a high standard of evidence. This is sometimes expressed in terms of an equivalent number of standard deviations for

---

[1] Note that the probability function contains everything which is known about the random aspects of the experiment, regardless of whether a frequentist or a Bayesian approach is to be used to interpret it. Thus, despite occasional claims to the contrary, frequentist and Bayesian analyses are equally dependent on the concept of randomness (sometimes discussed in conceptual terms as an ensemble of identical experiments).

[2] It might seem that the best determination of the event rate or other physical parameters would come from a Bayesian analysis using the theoretical or indirect information in the prior. However, physicists often want to *test* the theory or the consistency of the indirect information, so using that information in the analysis would lead to circular reasoning.

a Gaussian random process, *e.g.* "5 sigma", even when the distribution of the statistic is not Gaussian; the intent is to convey the false detection probability (less than 10^{-6} in this case).

More often than not, these experiments fail to observe clear evidence for the physical effect being looked for. The absence of a significant excess means that the rate or magnitude of the physical effect is unlikely to be very large; this may be expressed quantitatively as an *upper limit* on the event rate or magnitude.[3] Upper limits are typically reported with a 90% or 95% confidence level, depending on conventions established by past experiments in each field of research.

3. DETECTION ISSUES IN A BAYESIAN ANALYSIS

To illustrate some issues which are encountered in a Bayesian analysis, we consider the archetypal "Poisson process with background" case considered, for instance, by Feldman and Cousins [2]. This represents a "counting" experiment, in which the statistic used to summarize the data is the number of events, n, which satisfy a set of selection criteria designed to keep most signal events (if any exist) and reject uninteresting events. If μ is the mean number of signal events expected (an unknown physical parameter, in the range $0 \leq \mu < \infty$) and b is the mean number of background events expected (and is known accurately), then the likelihood function is the Poisson distribution with mean $\mu+b$:

$$L(\mu) = (\mu+b)^n \, e^{-(\mu+b)} / n!$$

Given some prior belief about the relative probabilities of different values of μ, we apply Bayes' theorem to get a posterior probability density function (pdf). For example, Figure 1 shows the posterior pdf if 7 events are observed, assuming b=3 and a constant prior pdf for μ.

A true Bayesian might consider this posterior pdf to be the final product of the analysis, but most physicists, I think, would want to go one step further and summarize the result with a credible interval. There is no objective rule which dictates what sort of credible interval should be constructed; three possibilities are illustrated in Figure 2. Choosing a credible interval which excludes zero is, in essence, a decision to interpret the result as an apparent detection with some degree of confidence. Is that an appropriate choice in this case? The fact that the pdf is distinctly peaked away from zero is certainly suggestive, but how robust is that as an indicator?

[3] In fact, some experiments / analyses, for which detection is unexpected according to theoretical predictions, are optimized so as to minimize the expectation value of the upper limit (assuming that no signal is seen).

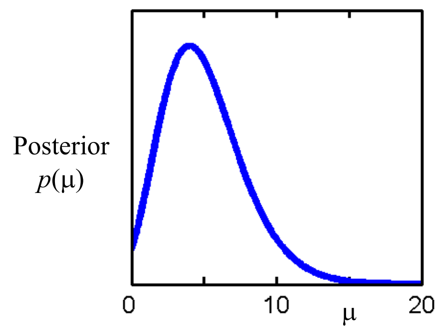

Figure 1: A posterior pdf for the example considered in the text, if n=7.

The peakedness of the posterior pdf depends, in part, on the choice of prior. Figure 3 shows the posterior pdfs for values of n between 6 and 10, for three different priors. Which ones do you think look significant enough that you would be comfortable publishing a paper claiming a detection? How often are you willing to be wrong? In the case of the constant prior, the posterior pdf is noticeably peaked away from zero even for n=6, but it turns out that the background will fluctuate up to 6 or more events 8.4% of the time,[4] so the presence of a peak is not necessarily a reliable indicator. This reflects the fact that a constant prior is too optimistic when searching for a signal which is likely to be small. In fact, if we are completely ignorant about the value of μ (which is a scale parameter in the likelihood), then the Principle of Maximum Entropy [3] suggests that we should use a prior of the form $1/\mu$. In this case, the posterior pdf develops a peak at somewhat higher values of μ, but it is improper for all values of n, so we cannot calculate credible intervals at all! In essence, this prior would lead us to conclude that *any* number of excess events is more likely to be a background fluctuation than to be a real signal.[5] The final prior considered in Figure 3, $1/\mathrm{sqrt}(\mu)$, represents a sort of compromise: it emphasizes small values of μ, but yields integrable posterior pdfs. Still, there is no guidance about what is significant enough to represent a detection, other than by considering the false detection probability (a frequentist concept!).

[4] For reference, the background (3 events on average) will fluctuate up to 7 or more events 3.3% of the time, 8 or more 1.2% of the time, 9 or more 0.4% of the time, and 10 or more 0.1% of the time.

[5] One might be tempted to use a prior of the form $1/(\mu+b)$, in which case a change of variables seems to reduce the problem to the simple Poisson case without background. However, this is not quite true, because the domain of the Poisson mean parameter becomes $[b, \infty)$, not $[0, \infty)$. In any case, this prior is conceptually flawed: one's *prior* belief about a physical parameter cannot depend on the properties of the *present* experiment!

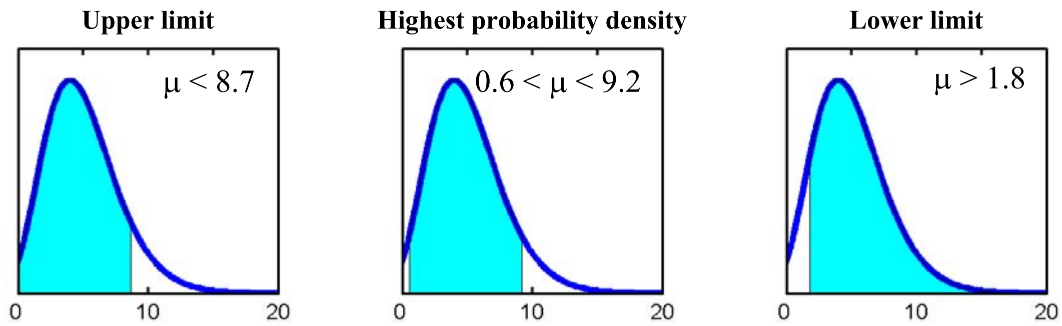

Figure 2: Three possible 90% credible intervals constructed from the posterior pdf shown in Figure 1.

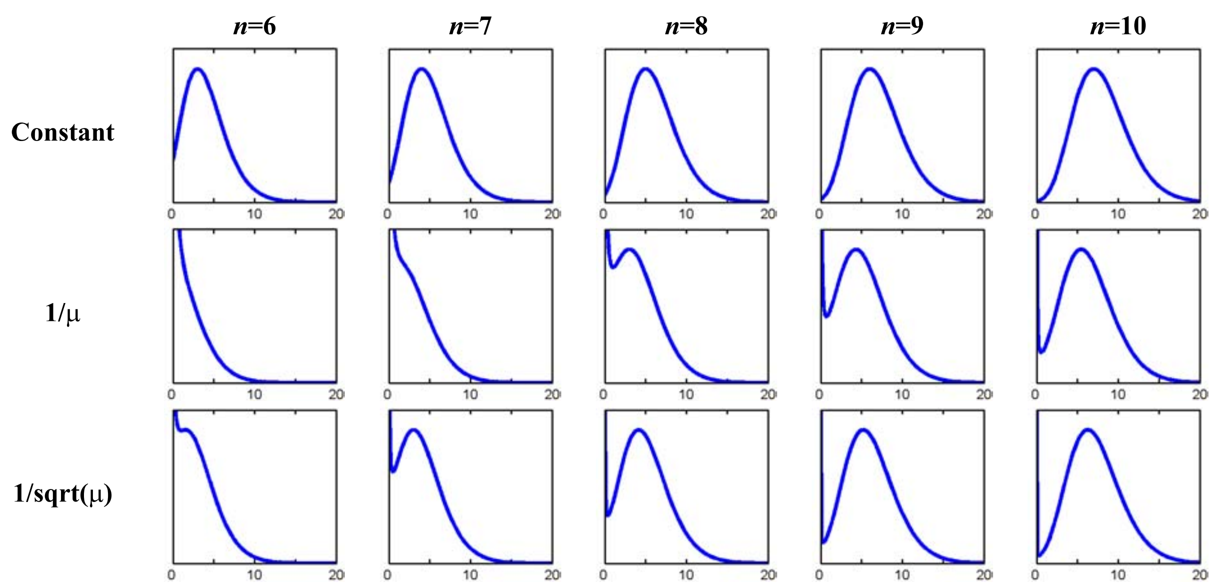

Figure 3: Posterior pdfs for the example considered in the text, for three different functional forms of the prior pdf: constant, $1/\mu$, and $1/\text{sqrt}(\mu)$. Note that in the latter two cases, the posterior pdf diverges as $\mu \to 0$, for all values of n.

4. DETECTION ISSUES IN A UNIFIED FREQUENTIST ANALYSIS

Feldman and Cousins have popularized a "unified" frequentist approach, in which the classical Neyman construction is performed with an alternative ordering principle based on likelihood ratios [2]. This approach (which has a few variations) yields confidence intervals which transition smoothly from one-sided to symmetric two-sided as n increases, maintaining the desired minimum coverage. This may seem to provide a well-defined detection criterion, at the point of the transition to two-sided intervals, but there is a crucial caveat (originally pointed out by Feldman and Cousins): this type of confidence interval is almost always calculated for a 90% confidence level, so an interval which excludes zero does not necessarily represent a detection at the higher confidence level that we want to require. It is, of course, possible to calculate the interval for a higher confidence level (say, 99.9%), but then the upper end of the interval will no longer be analogous to a traditional 90% upper limit, which is generally considered to be a desirable feature. Faced with this situation, some collaborations follow a policy of giving a 90% unified confidence interval (which may be two-sided) and stating separately whether there is a "detection", based on the p-value for the null hypothesis. For example, Figure 4 shows the Feldman-Cousins construction for our Poisson process with background, which yields two-sided intervals for $n \geq 6$, whereas a detection with a p-value of 0.01 or less would require $n \geq 9$.

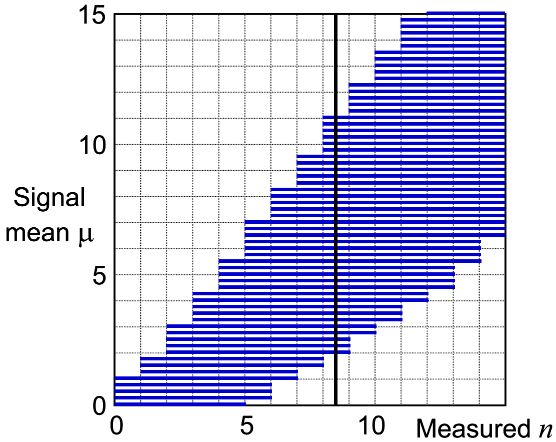

Figure 4: Feldman-Cousins confidence intervals for Poisson process with expected background b=3 (adapted from Figure 6 of Ref. [2]). The horizontal bands are "acceptance regions" for various values of μ. The thick vertical line indicates the mean number of observed events, $n \geq 9$, which would be required to make a detection with a false detection probability less than 1%.

5. SUMMARY

Physicists generally expect an interpretation of the outcome of an experiment, such as a statement about the significance of any excess events observed. In the case of a search for a new phenomenon, a high standard of evidence is required to support a claim of a "detection". Even in the absence of a signal, a Bayesian analysis may occasionally yield a pdf which is peaked away from zero, and does not provide a quantitative measure of significance. A unified frequentist approach could in principle provide a well-defined detection criterion, but is not customarily calculated for an appropriately high confidence level. The best established quantitative approach to evaluate an apparent detection of a new physical phenomenon is to calculate the p-value for the null hypothesis. Of course, human judgment is still required to decide how low the p-value must be to be interpreted as a detection.

References

[1] J. Neyman, Philos. Trans. R. Soc. London **A236**, 333 (1937). Reprinted in *A selection of Early Statistical Papers on J. Neyman* (University of California Press, Berkeley, 1967), pp. 250-289. Also see the discussion in reference [2], below.

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

[3] E. T. Jaynes, IEEE Trans. Syst. Sci. Cybernet. **SSC-4**, 227 (1968).