statistical inference problems workshop

BIRS, Banff July 16 2006

setting the scene for limits and nuisance parameters

Joel Heinrich
University of Pennsylvania (Physics)

draft version 7.12.608

The charge to group A1 is to select or develop one or more methods to solve problems of the following type. Methods should be ranked on criteria to be determined by the group. The "Commonly proposed methods" and "Typical properties for comparison" listed below are only for illustration; the group must decide for itself what methods and criteria to select.

the problem (quoted from the email sent to participants)

We are performing a counting experiment with the (non-negative integer) number of observed events n being Poisson distributed with mean

$$\mu = \epsilon s + b$$

where s is the parameter of interest for which we wish to set an upper limit ($0 \leq s < s_{\mathrm{U}}$), or a 2-sided interval ($s_{\mathrm{L}} < s < s_{\mathrm{U}}$). s (the "cross section") is the parameter of interest, and in principle can have any real value $0 \leq s < \infty$.

The nuisance parameter ϵ is a factor which converts between n and s in some sense. It must be ≥ 0, and could be > 1. It is either precisely known, or can have an uncertainty (see below).

The nuisance parameter b is the background rate. It must ≥ 0. It is either precisely known, or can have an uncertainty (see below).

When ϵ and b have uncertainties, we may regard them as having been determined in subsidiary counting experiments. The observed numbers of events in these subsidiary experiments is set to give the required uncertainties on ϵ or on b. Another variation of the problem is that we just have Bayesian priors for ϵ and b that are derived from a combination of objective information and personal belief.

Typical values:

$$\epsilon = 1.0 \pm 0.1 \qquad b = 3.0 \pm 0.3 \qquad n = 0, 1, 2, \ldots 20$$

Possible extension:

A 2-channel version of the above, with n, ϵ and b (and the errors on ϵ and b) each having two values, one for each channel, while s is common to the 2 channels i.e. n_1 and n_2 are independent Poisson observables, with means $\epsilon_i s + b_i$. Again it is required to determine an interval for s.

Typical values: divide equally among the channels, e.g.

$$\epsilon_1 = 0.5 \pm 0.1/\sqrt{2} \qquad \epsilon_2 = 0.5 \pm 0.1/\sqrt{2}$$

$$b_1 = 1.5 \pm 0.3/\sqrt{2} \qquad b_2 = 1.5 \pm 0.3/\sqrt{2}$$

where the subsidiary measurements are also divided between the 2 channels.

Commonly proposed methods in High Energy physics:

- Bayes: Prior for s = uniform; $1/\sqrt{s}$; $1/s$
  Prior for ϵ and for b subsidiary measurements = uniform

- Profile likelihood

- Modified profile likelihood

- Feldman-Cousins, with some fix for nuisance parameters

- Fully frequentist, with some ordering rule

Typical properties for comparison:

- Coverage vs s at $(\epsilon, b) = (1,3)$, (1.1,3), (0.9,3), (1,3.3), (1,2.7).

- Bayesian credibility for intervals.

- Interval length values for n distribution (Median and quartiles).

- Behavior as a function of b, given $n = 0$ and $n = 3$.

Why is this problem still unresolved?—A brief introduction.

The physicists present here can generally recommend, but not enforce, statistical procedures. Some procedures, although statistically sound, have nevertheless been unpopular in HEP.

For example, some frequentist methods can produce empty intervals (or single point intervals). For a frequentist, this makes perfect sense, but very few physicists would ever consider publishing such a result. Such procedures don't get adopted, or worse, are replaced after an undesirable result actually occurs.

The Unified method of F&C, which takes the decision of whether to quote a one or two sided interval away from the physicist, has also proved difficult to sell in some cases. The recent emphasis on 5σ discovery may magnify the problem in the future, as physicists become more reluctant to publish a 2-sided interval for a result with lower significance.

Because the frequentist/Bayesian dispute continues unabated in HEP, ideally, a procedure should satisfy both sides. Investigations of the typical coverage of various methods are now available, but without a clearly defined worst case, these may not convince the skeptic. Investigations of the Bayesian credibility of non-Bayesian methods are absent, leaving Bayesians unsatisfied.

Subjective informative priors for the parameter of interest are very unpopular. We have no confidence in our own opinion, for example, of the mass of the Higgs particle, nor in anyone else's opinion. Therefore, even subjective priors are invariably uninformative for the parameter of interest. But priors for some nuisance parameters are, in some cases, both subjective and informative—a problem for frequentists.

Philosophical opposition from some quarters is still present against hybrid or mixed methods. Objective Bayesian approaches, for example, provoke questions like "How am I to interpret such intervals?" "What should I tell my students that they mean?"

Fully frequentist methods that rely on the Neyman construction have, in the past, been considered unworkable in more than a few dimensions. Progress is being made in this area, but more work is clearly required.

Asymptotic solutions that are only well behaved at large n are unappealing, as situations where $n \leq 2$ are not uncommon.

Situations where the observed quantity is an integer tend to cause problems for methods that enforce a strict coverage requirement. In the A1 problem, such methods tend to yield larger intervals, for example, when the uncertainty on the background parameter b is zero, than when it is small but finite. Such behavior is considered to be objectionable by most physicists, who would like to be rewarded for their efforts in understanding their background by obtaining a smaller interval, not a larger one.

A related issue is the choice of the ensemble of experiments over which the coverage is calculated. Traditionally, in calculating the coverage, one fixes the time over which the data is taken (really the integrated luminosity) in the frequentist repetitions. But typically, both the running time and the final number of events n are unspecified at the beginning of a real life experiment. The issue of how closely the frequentist repetitions must match the intention of the experimenters, for the calculated coverage to be meaningful, has received little attention.

Including a mixture of running times in the coverage ensemble would increase the minimum coverage of a method, and permit smaller intervals while maintaining strict coverage.

The issue prompts a debate of the relative merits of "strict coverage" vs "average coverage".

Although flat priors are historically the most common in HEP, some caution must be exercised. While a flat prior for the parameter s is considered to be "conservative" for upper limits (one sided intervals), this is not true for 2 sided intervals or lower limits. Difficulties arrise when a flat prior for s is combined with a Gaussian (normal) prior for ϵ. In multi channel versions of the problem, flat priors for the subsidiary experiments cause trouble—a well known problem for priors flat in large numbers of dimensions.

There are various methods for selecting priors that are better behaved (the objective Bayesian approach). Here, rather than originating from "personal belief", the prior is selected to give good frequentist coverage properties, or minimal influence on the posterior.

The objective Bayesian approach, however, sacrifices some Bayesian purity, as different priors may be selected, depending on the frequentist ensemble (a stopping rule dependence), thus violating the likelihood principle.

Conclusion

A discussion of the benefits and defects of various proposed methods could continue indefinitely. I think that the HEP community will eventually converge on a single solution, or at least a small number of solutions, but this convergence still seems years away. We hope that progress can be accelerated using the input of the statisticians at this workshop.