ADDITIONAL COMMENTS ON METHODS FOR SETTING CONFIDENCE LIMITS *

Robert D. Cousins †
Department of Physics and Astronomy, University of California,
Los Angeles, California 90095, U.S.A.

1. INTRODUCTION

My talk at the Fermilab Confidence Limits Workshop was largely a repeat of my talk at the CERN CLW two months earlier. In this writeup, I add a few comments I made a Fermilab which were not in the writeup of my CERN talk. My CERN talk, and all other talks to which I refer in this writeup, can be found by following the links from http://cern.web.cern.ch/CERN/Divisions/EP/Events/CLW/. Several of my added remarks at Fermilab were simply to draw attention to talks at the CERN CLW which were not at the Fermilab CLW. The third section below is expanded since the CLW, in order to try to resolve some confusion in my talk.

2. FURTHER EXAMPLES AND CRITICISMS OF LIKELIHOOD-RATIO ORDERING IN THE UNIFIED APPROACH

At the CERN CLW, an interesting talk was given by B. P. Kerševan, applying the unified approach to "Estimation of Confidence Intervals in Measurements of Trilinear Gauge Boson Couplings". This is a non-trivial example, for which the authors were unhappy with classical central intervals and with intervals using the "maximum log-likelihood – 1/2" recipe. The unified approach provided non-simply-connected intervals which were sensible to them.

In a published example [1], the CDF collaboration made a measurement of the standard model CP-violation parameter $\sin 2\beta$ in $B \to \psi/JK_S^0$ decay. This is an interesting example because the physical constraint is two-sided: $-1 \le \sin 2\beta \le 1$. CDF reported an unconstrained measured value of $\sin 2\beta = 0.79 \pm 0.39(\mathrm{stat}) \pm 0.16(\mathrm{syst})$. With the unified approach, they found that $0 < \sin 2\beta < 1$ at 93% C.L. With a prior pdf uniform in $\sin 2\beta$, a Bayesian calculation yielded $P(\sin 2\beta > 0) = 95\%$. Finally, they calculated that if $\sin 2\beta = 0$ and $\sigma = 0.44$ (Gaussian), then the probability of measuring $\sin 2\beta > 0.79$ is 3.6%.

G. Zech's paper at the CERN CLW contains some interesting criticisms of the unified approach. Zech finds the Feldman-Cousins (FC) intervals "intolerable" in the Poisson case with background when zero events are observed. While I would suggest caution in interpreting some of his statements, I found two of his examples to be particularly interesting. The first is a 2D example with Gaussian resolution in x and y, and a physical boundary in y. The second is an "earthquake time predictor" which has a single unknown parameter on which both the mean and the width of the distribution depend.

3. HOW TO USE THE KNOWLEDGE THAT YOU KNOW $N_b \le n$?

This part of my talk was added the night of the first day of the Fermilab CLW in an attempt to elaborate on a discussion earlier that day regarding the Poisson-with-background problem. Since the workshop, I have looked yet again at the relevant papers, and discovered that I had not understood the relationships among the equations of Zech's 1989 paper [2], Highland's Comment on it [3], and Roe and Woodroofe's 1999 paper (RW) [4]. In particular, I did not realize that the papers of Zech and RW were closely related, to the extent that Zech's main formula appears in RW's paper (with a different derivation). Therefore, rather than repeating the confused statements I made at the workshop, I hope to clarify the situation here.

* Part of talk at Fermilab Workshop on Confidence Limits, 28 March 2000
† Email address: cousins@physics.ucla.edu

I adopt the notation of RW, since one of the authors is a professional statistician. In that regard, I note that statisticians use the word "distribution" in a different manner than many physicists (and nearly all astronomers). To statisticians, it is not synonymous with "probability density function" pdf, but rather it is the integral of the pdf up to some value. Here we are dealing with the discrete Poisson case,

$$p_\mu(k) = \mu^k e^{-\mu}/k!, \tag{1}$$

for non-negative integer k and mean μ. Physicists call this a "probability" and RW call it a "density", but Woodroofe tells me he actually prefers (probability) "mass function". The distribution function, denoted with upper case P, is

$$P_\mu(k) = p_\mu(0) + \cdots + p_\mu(k). \tag{2}$$

Also, statisticians use upper case for random observables, and corresponding lower case for a particular observed value. For the problem at hand, let

$\lambda =$ unknown signal mean

$b =$ known background mean

$N =$ random total number of events (signal and background)

$n =$ an observed total number of events (signal and background)

N_b = random number of background events.

n_b = an observed number of background events.

Then, for example, N follows the probability $p_{b+\lambda}(N)$, while N_b follows the probability $p_b(N_b)$.

Now suppose that, in your experiment, you see n total events. This gives information about N_b, most obviously (but not limited to) the fact that $N_b \leq n$. How shall we express what we know about N_b?

Virgil Highland [3], in commenting on a paper by Günter Zech (see below) asserted that the relevant thing to compute is the conditional probability of N_b, given that you obtained n total events (signal and background). In other words, imagine an ensemble of experiments identical to yours (i.e., signal and background events sampled from the same unknown λ and known b as for your experiment). Select the subset containing those experiments which had the same total events n as you obtained. Then what is the probability of N_b within this subset?

This is a conditional probability, and Highland correctly computed it using Bayes' Theorem (the non-controversial version for sets, $P(A|B) = P(B|A)P(A)/P(B)$). We denote this conditional probability by the somewhat tortured notation $p_b(N_b|n)$, and it is then

$$p_b(N_b|n) = p_{b+\lambda}(n|N_b)\; p_b(N_b) \Big/ p_{b+\lambda}(n). \tag{3}$$

The first p on the right-hand-side (rhs) is also $p_\lambda(n - N_b)$, so the entire rhs can be computed using a subroutine for unconditional Poisson p's. Highland's paper gives numerical examples of this computation, which clarify the calculation.

Said yet another way, Highland computed the conditional probability of N_b, given that each experimenter *knows* precisely that $N_b < n$.

Highland was commenting on Zech's 1989 paper [2], which computes a different probability, one that in contrast is the conditional probability that $N_b < n$, *whether or not each experimenter knows that* $N_b < n$. In Zech's words, (which I have converted to the present notation):

"In an experiment where b background events are expected and n events are found, $p_b(N_b)$ no longer corresponds to our improved knowledge of the background distributions. Since N_b can only take the numbers $N_b \leq n$, it has to be renormalized to the new range of N_b:

$$p'_b(N_b) = p_b(N_b) \Big/ P_b(n)." \tag{4}$$

To obtain Zech's ensemble, again imagine an ensemble of experiments identical to yours (i.e., signal and background events sampled from the same unknown λ and known b as for your experiment). This time select the subset containing those experiments which had n_b less than or equal to the particular number n you obtained in your experiment. (This is possible only in a MC calculation.) Eqn. 4 gives the probability of N_b within this subset.

Everyone agrees that Highland's and Zech's probabilities are different. (Highland gives a numerical example.) In addition, I believe that Zech's wording is unfortunate: I think that Highland's formula is the one which truly "corresponds to our improved knowledge of the background distributions", an opinion which reflects my opinions of the ensembles. Note that $p_b(N_b|n)$ depends on λ while $p'_b(N_b)$ does not; this has been used an an argument in favor of each.

One point I want to emphasize, and to which I return below, is a practical difference between the Zech and Highland ensembles. Given a real set of actually performed experiments, it is possible to to point to those experiments which are in Highland's ensemble (they are the ones with n events). It is impossible to point to the ones in Zech's ensemble: since signal and background events are indistinguishable, we cannot be sure of all the experiments with $n_b < n$. Of course, in a Monte Carlo (MC) simulation, one can distinguish between signal and background events. So Zech's ensemble conditions on a number known in the MC but not in real experiments.

Continuing with Zech's ensemble (for which Eqn. 4 is valid) note that the selection of the ensemble did not affect the probability of obtaining k signal events, which is still the independent Poisson probability $p_\lambda(k)$. Now consider any number $N \leq n$ of total events. Within Zech's ensemble, the probability $w(N)$ of obtaining N is the sum over all possible ways of obtaining N:

$$
\begin{aligned}
w(N) &= \sum_{k=0}^{N} p_\lambda(k)\, p'_b(N-k) &(5)\\
&= \sum_{k=0}^{N} p_\lambda(k)\, \frac{p_b(N-k)}{P_b(n)} &(6)\\
&= \frac{1}{P_b(n)} \sum_{k=0}^{N} p_\lambda(k)\, p_b(N-k) &(7)\\
&= \frac{p_{b+\lambda}(N)}{P_b(n)}, &(8)
\end{aligned}
$$

where the last step uses the fact that two independent Poisson variates have a sum which is also a Poisson variate, with mean equal to the sum of the means. Highland disputed this formula for $w(N)$, because it makes no sense for Highland's ensemble, within which $N = n$ always. I agreed with him, but it is finally clear to me that it makes sense in Zech's ensemble. By summing $w(N)$ from $N = 0$ to n, we have, in Zech's words:

"The probability $D_{b,\lambda}(n)$ of observing n events or less, knowing that the background is n events or less for a signal λ, is then...

$$D_{b,\lambda}(n) = \frac{P_{b+\lambda}(n)}{P_b(n)}. \text{''} \tag{9}$$

This is algebraically equivalent to the "Helene formula" [5] of the "old PDG", which Helene derived in a completely different way using Bayesian techniques. I say "algebraically", since conceptually, they are different equations, with different definitions of probability. Thus I think I will call this equation "Helene's equation" when it means degree of belief, and call it "Zech's equation" when it means relative frequency, as in the above calculation with his ensemble.

A decade later, Roe and Woodroofe [4] proposed that a conditional probability be included in the unified approach. For a long time, I did not recognize that their conditioning was equivalent to Zech's ensemble. But in fact, Eqn. (7) in Roe and Woodroofe's paper is identical to Zech's equation!

Here is RW's derivation. They let $q^n_{b,\lambda}(k)$ be the conditional probability of exactly k events total given at most n background events. Then they assert that

$$q^n_{b,\lambda}(k) = \begin{cases} p_{b+\lambda}(k)/P_b(n) & \text{if } k \le n \\ \sum_{j=0}^{n} p_b(j) p_\lambda(k-j)/P_b(n), & \text{if } k > n. \end{cases} \tag{10}$$

In 1999, the first thing I did when I saw RW's paper was check that this was consistent with Bayes' Theorem. We can write (stretching the notation)

$$q^n_{b,\lambda}(k) \;\equiv\; p_{b+\lambda}(k|N_b \le n) \;=\; p_b(N_b \le n|k)\, p_{b+\lambda}(k) \Big/ p_b(N_b \le n). \tag{11}$$

The denominator is the distribution $P_b(n)$. For $k \le n$, the first term on the r.h.s. is 1. For $k > n$, the numerator is perhaps most easily seen to be the joint probability of "$N_b \le n$ AND k". Thus one obtains Eqn. 10.

If we sum $q^n_{b,\lambda}(k)$ for $k \le n$, then the numerator is the distribution $P_{b+\lambda}(n)$ and we obtain, in RW's words, "the conditional probability of at most n events total, given at most n background events". This is Eqn. 9 above, and RW's verbal description agrees with Zech also. RW follow with a explanation of the "not obvious" fact that $D_{b,\lambda}(n)$ is a distribution.

Thus, RW and Zech completely agree on Eqn. 9 for $D_{b,\lambda}(n)$ and on the verbal description of it. In the past, I saw similarities in the techniques of RW and Highland, and apparent disparity in the techniques of Zech and RW, so that I missed the clear fact that RW and Zech calculated the same thing; I am sorry for the confusion this created. Discussions with John Conway helped to point me in the right direction. I assume that now everyone can agree on the equivalence of the Zech and RW derivations and on the words describing the result.

The confidence-belt method RW proposed was not based on $D_{b,\lambda}(n)$; they mention it in a section after their main result, which is a modification to the unified approach using $q^n_{b,\lambda}(k)$. [In the Fermilab CLW, B. Roe's talk advocated a new method not based on modifying the unified approach.]

I now return to the practical difference in the Highland and Zech conditional ensembles. One of Highland's comments would now seem to apply to both RW and Zech: that for an ensemble of experiments which have N_b less than some constant, "it is difficult to see what physical experiment[s] this would correspond to, since background and signal are indistinguishable." This comment is related to the novel aspect of the Zech/RW conditioning. In standard conditioning on an ancillary statistic, it is known whether or not two experiments are in each other's ensemble: if they have the same value of the ancillary statistic, they are, otherwise not. With the Zech/RW conditioning on an inequality, it can be unknown if completed Experiment B is in the ensemble used to calculate coverage for Experiment A. (As A. Read was careful to point out in his talk at the CERN CLW, only in the Monte Carlo simulation does one know this.) This would seem to make it hard to verify conditional coverage experimentally. It would be interesting to find a precedent for this type of conditioning in the professional statistics literature.

In summary, I think we can all finally agree that Zech derived (and RW rederived) an equation algebraically equivalent to the Helene formula, but using the "frequency" definition of probability. We can also all agree that this calculation does not correspond to "frequentist confidence intervals" for an unconditional ensemble. (For upper limits, they always overcover.) Nor does it correspond with an subset which can be identified in a set of real (not MC) experiments. Zech correctly points out that he never made claims to the contrary. I think there is still disagreement about how useful Zech's conditioning is, but I think that much more is clear now. I am very grateful to Günter Zech for patiently discussing these issues with me many times until I finally understood his point of view, and most recently for crucially helpful comments on this writeup.

4. CONCLUSION

The conclusion of my Fermilab talk echoed my CERN talk. In spite of major differences, there are a number of things I believe we can agree on. Some of them may appear trivial to experts, but nonetheless I see papers by authors who are unaware of the agreement about them. G. Zech reminded me after the CERN CLW that I should have included something about difficulties with nuisance parameters and systematic errors in the frequentist approach. Gary Feldman's talk at this workshop describes some of our work (mostly his) in this area. (See also Wolfgang Rolke's talk on the same topic.)

REFERENCES

[1] T. Affolder, et al., (CDF collaboration), Phys. Rev. **D61** 072005 (2000).

[2] G. Zech, Nucl. Instr. and Meth. **A277** 608 (1989).

[3] V. Highland, Nucl. Instr. and Meth. **A398** 429 (1997), followed by reply by G. Zech.

[4] B.P. Roe and M.B. Woodroofe, Phys. Rev. **D60** 053009 (1999). At this Fermilab workshop, B. Roe described a new method.

[5] O. Helene, Nucl. Instr. and Meth. **212** 319 (1983).