Note *RW1906*
version 3/2019-11-14

The illusion of coverage probability

Roland Waldi
Institut für Physik, Universität Rostock

The big advantage of frequentist confidence intervals is that their construction procedure requires CL = CP for measurements of continuous quantities x, where CL = confidence level and CP = coverage probability, i. e. the (physical or frequentist) probability that the confidence interval, which is a random interval depending on the measured value x, contains (covers) the true parameter value $\tilde{p}$. While CP(p) as a function of all possible parameters p can be calculated unambiguously for any procedure (algorithm) used for confidence interval construction, a frequentist procedure starts with this probability: intervals $[x_1, x_2]$ are determined for each possible parameter p via

$$P([x_1, x_2]) = \int_{x_1}^{x_2} f(x|p)\,\mathrm{d}x = \mathrm{CL} \tag{1}$$

defining an area in the p, x-plane. Then, a confidence interval for p is obtained by using the boundaries of this area in p-direction, as illustrated in Fig. 1.

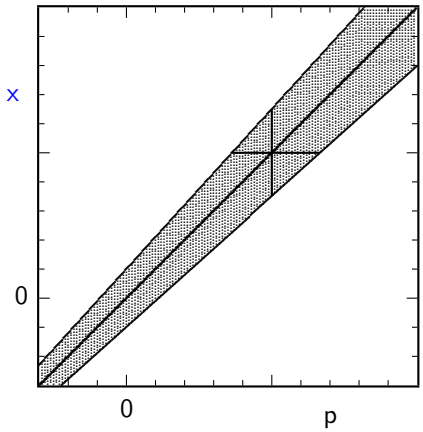

Fig. 1 Confidence region in the p, x-plane: the vertical intervals contain x for any given p with probability CL, the horizontal intervals are the confidence intervals for a measured x-value at this CL.

This procedure suggests that CP = CL, which is a constant probability that the interval includes the true parameter $p = \tilde{p}$. But this is often not the case for two reasons:

1. For discrete measurements we cannot have a constant CP(p) = CL due to the discontinuity of the measured variables $x = n$ (counts), and instead require CL ≤ CP. The Clopper-Pearson algorithm often used for binomially distributed data is known to overcover extremely [1], but also for better algorithms and for Poisson data overcoverage is unavoidable. For a Poisson distribution with parameter p (distribution mean), a frequentist confidence upper limit is constructed in analogy to (1) by finding lower values $n_{\min}$ for counts n with

$$\sum_{n=n_{\min}}^{\infty} P(n) = \sum_{n=n_{\min}}^{\infty} \frac{e^{-p} p^n}{n!} \geq \mathrm{CL}$$

and then for a given n the maximum p for which this n is $n_{\min}(p)$ is used. A coverage probability CP(p) for 90%CL with this procedure is shown in Fig. 2. If there are many single measurements, e. g. histogram bins, with true expectation value $p = \langle n \rangle < 1$, the coverage probability is CP = 100% for any reasonable CL, and CL can no longer be used as an—even approximate—frequency.

2. Also, the situation is different considering actual practice, best illustrated in the case of a Gaussian x with

$$f(x|p) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-p)^2}{2\sigma^2}}$$

where p is limited to nonnegative values, $p \geq 0$. There exist different sensible frequentist algorithms: (a) using the shortest x-intervals, leading to boundaries[1] of $x = p \pm 1.64\sigma$, (b) using lower limits in x to obtain upper limits in p, with $x > p - 1.28\sigma$, and (c) the Feldman-Cousins (FC) boundaries [2].

If each algorithm is used independent of the result x, all three have indeed CP = CL. However, since negative values will yield empty confidence intervals for $x < -1.64\sigma$ for (a) and $x < -1.28\sigma$ for (b), only the FC procedure is a viable alternative of these three with x below -1.64σ. Furthermore, 5σ is the generally accepted observation limit, so below this value typically an upper limit will be given as confidence interval. Since nobody uses procedure (c) for significant positive results, we are faced with a mixture of (a,b,c).

A reasonable way to mix methods that is in agreement with many (most?) experimental papers is to use method (a) if $x > 5\sigma$, method (b) if $0 \leq x < 5\sigma$ and method (c) if $x < 0$. This mixed method is shown in Fig. 3 for 90%CL. It shows regions where CP > CL, in particular for low values of p, and regions where CP < CL, with CP = 85% for regions around $p \approx 4$ to 5σ.

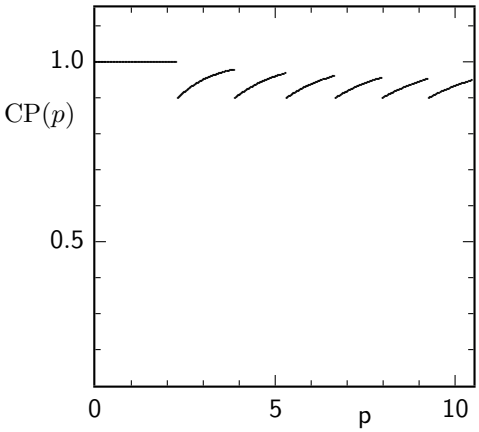

Fig. 2 Coverage probability CP(p) of the upper limit interval $0 \leq p \leq p_{\max}$ at 90%CL for Poisson-distributed data.

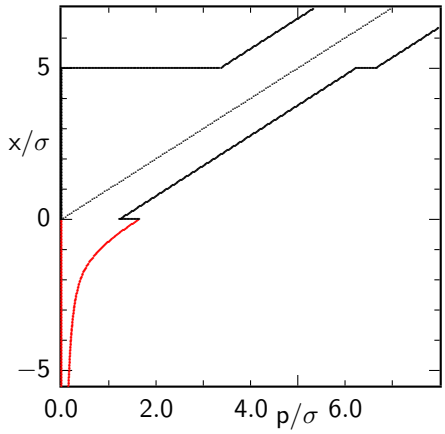

Fig. 3 Frequentist confidence regions for 90%CL using x-dependent methods (a), (b) and (c) described in the text, with CP from 85% to almost 100%.

This means that in practical use a mixed method, where the choice is made based on the measurement x, has both over- and undercoverage. This effect has already been discussed in [2], Fig. 4: But their "flip-flopping physicist" is a reality. A possible way to maintain CP would be to strictly follow Ref. [2] and use this method for any result x. However, this is unrealistic, since it means to give a two-sided interval with a lower limit at 90%CL already for a result of 1.3σ significance (1.7σ at 95%CL).

So, for both reasons (1) and (2) CL $\neq$ CP is unavoidable, and the assumption that CL is the probability that published intervals cover the true parameter value is an illusion.

References

[1] T. E. Sterne, *Some remarks on confidence or fiducial limits*, Biometrika **41**, 275 (1954).

[2] G. J. Feldman, R. D. Cousins, *A Unified approach to the classical statistical analysis of small signals*, Phys. Rev. **D57**, 3873 (1998); updated e-Print `physics/9711021` Dec. 1999.

[1] I use values rounded to three digits, for a strict Gaussian the numbers are 1.28155 and 1.64485 to six digit precision, but to this precision we cannot expect exact Gaussian behaviour anyway.