About the proof of the so called *exact classical confidence intervals*. Where is the trick? *

G. D'Agostini
Università "La Sapienza" and INFN, Roma, Italia
(giulio.dagostini@roma1.infn.it, `http://www.roma1.infn.it/~dagos`)

Abstract

In this note I go through the 'proof' of frequentistic confidence intervals and show what it logically implies concerning the value of a physical quantity given an experimental observation (nothing).

"... to emancipate us from the capricious ipse dixit of authority"
(John Henry Newman)

1 Introduction

The construction of frequentistic confidence intervals is the statistical tool adopted in all 'conventional statistics' [1] books and lecture notes in order to provide the result on the value of a physical quantity resulting from a measurement. Though many physicists are aware of the unsuitability of the 'prescription' to handle critical cases that occur in frontier science (and the troubles are also known to the supporters of the prescription: see [2], references therein and follows up), it seems they do not always recognize the real reason the prescription fails.

*Based on a lectures to graduate students at the University of Rome "La Sapienza" (May, 4 2005 and May, 8 2006).

In the past I am always been reluctant to go through the details of the definition of the frequentistic confidence intervals, simply because I thought that, once one has realized that:

- the outcome of these methods is usually misleading (in the sense that the common interpretation differs from what they are supposed to mean, see e.g. Ref. [1]);
- the resulting 'confidence' intervals can often come out absurd (see e.g. Ref. [3], sections 1.7 and 10.7);
- the celebrated 'frequentistic coverage' does not do its job in the important cases of interest in frontier physics (see Ref. [3], section 10.7);

then there is little more to do, apart from looking for something better. I admit I have been naive. In fact I have met too many students and colleagues touched by the above arguments but still not fully convinced, because impressed by the names given to the criticized methods (something that is *classical*, *rigorous* and *exact* cannot be wrong!) and the authority of books and publications that use and recommend them.[1]

The aim of this note is to go through the definition of the confidence interval as one finds it in books and lecture notes[2] and to show why that reasoning yields no new information concerning the value of the unknown quantity of interest and about which one wants to gain some confidence based on experimental observations.

[1] I found a curious analogy between the learning of fundamental concepts of physics and the learning of statistical methods for data analysis, as they are taught in the standard physics curriculum. It is a matter of fact that many concepts of physics are not easy at all to be grasped by students (and not not only by students!). After some time, students assume the habit to learn at least the practical formulas, trusting teachers and text books as far as correctness and deep meaning of the difficult concepts are concerned. When they first learn frequentistic statistics applied to data analysis, presented in the usual shifty way we know, the students tend to act in the same way, thinking the things must be difficult but correct, because granted by the teachers, who in most case just repeat the lesson they learned but tacitly hold the same doubts of the pupils. The small 'Socratic exchange' by George Gabor reported in Ref. [3] might be enlightening.

[2] I avoid to give a particular reference. Everyone can check his/her preferred text and see how confidence intervals are presented.

2 From the probability interval of the estimator $\hat{\theta}$ to the confidence about the true value θ

Here is, essentially, the basic reasoning that introduces confidence intervals.

1. Let θ be the *true value* of the quantity of interest and $\hat{\theta}$ its 'estimate' (an experimental quantity, probabilistically related to θ).

2. We assume to know, somehow,[3] the probability density function (pdf) of $\hat{\theta}$ for any given value θ, i.e. $f(\hat{\theta}\,|\,\theta)$. The knowledge of $f(\hat{\theta}\,|\,\theta)$ allows then us to make probabilistic statements about $\hat{\theta}$, for example that $\hat{\theta}$ will occur in the interval between $\theta - \Delta\theta_1$ and $\theta + \Delta\theta_2$:

$$P(\theta - \Delta\theta_1 < \hat{\theta} < \theta + \Delta\theta_2) \;=\; \alpha\,. \tag{1}$$

   For example, if $f(\hat{\theta}\,|\,\theta)$ is a Gaussian and $\Delta\theta_1 = \Delta\theta_2 = \sigma$, then $\alpha =$ 68.3 % (hereafter it will be rounded to 68%).

3. To establish the confidence interval about θ, it is said, we *invert the probabilistic content of (1)*:

$$\text{from} \quad \left\{ \begin{array}{ccccc} \theta & - & \Delta\theta_1 & < & \hat{\theta} \\ \theta & + & \Delta\theta_2 & > & \hat{\theta} \end{array} \right. \qquad \text{follows} \quad \left\{ \begin{array}{ccccc} \theta & < & \hat{\theta} & + & \Delta\theta_1 \\ \theta & > & \hat{\theta} & - & \Delta\theta_2 \end{array} \right. \tag{2}$$

   thus rewriting (1) as

$$P(\hat{\theta} - \Delta\theta_2 < \theta < \hat{\theta} + \Delta\theta_1) \;=\; \alpha\,. \tag{3}$$

Then, it usually follows some humorous nonsense to explain what (3) is not and what, instead, should be. Some say that *this formal expression does not represent a probabilistic statement about θ, because θ is not a random variable, having a well defined value, although unknown.* Instead, it is said, $\theta_1 = \hat{\theta} - \Delta\theta_2$ and $\theta_2 = \hat{\theta} - \Delta\theta_1$ are random variables. Then, the meaning of (3) is not probability that θ lies inside the interval, but rather probability that the interval $[\theta_1, \theta_2]$ encloses θ, in the sense ('frequentistic coverage') that if *we repeat an infinite number of times the experiment, the true value θ (that is always the same) will be between θ_1 and θ_2 (that change from time to time) in a fraction α of the cases; i.e. in each single measurement the statement $\theta_1 < \theta < \theta_2$ has a probability α of being true.* (And all people of

[3] I skip over the fact that in most cases $f(\hat{\theta}\,|\,\theta)$ is not really obtained by past frequencies of $\hat{\theta}$ for some fixed value of θ, as the frequency based definition of probability would require.

good sense wonder what is the difference between the latter statement and saying that α is the probability that θ is between θ_1 and θ_2.)

Anyhow — and more seriously —, besides all this nonsense, what matters is what finally remains in the mind of those who learn the prescription and how Eq. (3) is used in scientific questions. If a scientist knows $f(\hat{\theta}\,|\,\theta)$ and observes $\hat{\theta}$, then he/she feels authorized by Eq. (3) to *be confident*, with confidence level α, that the unknown value of θ is in the range $\hat{\theta} - \Delta\theta_2 < \theta < \hat{\theta} + \Delta\theta_1$. And it is a matter of fact that, in practice, all users of the prescription consider Eq. (3) as a probabilistic statement about α, i.e. they feel confident that θ is in that interval as he/she is confident to extract a white ball from a box that contains a fraction α of white balls.[4]

3 What do we really learn from Eqs. (1)–(3)?

Let us now go through the details of the previous 'proof' and try to understand what we initially knew and what we know after the 'probabilistic content inversion' provided by Eq. (3). In particular, we need to understand what we have really learned about the unknown true value θ as we went through the steps (1)–(3).

Given the general assumptions, the statement (1) is certainly correct. One would argue whether $f(\hat{\theta}\,|\,\theta)$ is indeed the 'right' pdf, but this is a different story (the issue here is just logic, as we are only interested in logical consistency of the various statements). Much awareness is gained about what is going on in steps (1)–(3), if we rewrite Eq. (1) stating explicitly the basic assumption as a explicit condition in the probabilistic statement, and distinguishing the name of variable from its particular numerical value,

---

[4] What happens is that the poor teacher (see footnote 1) at the end of the day is forced to tell that Eq. (3) is 'in practice' a probabilistic statement about θ, perhaps adding that this is not rigorously correct but, essentially, 'it can be interpreted *as if*'. However, this is not just an understandable imprecision of the 'poor teacher', in conflict between good sense and orthodoxy [1]. For example, we read in Ref. [5] (the authors are influential supporters of the use frequentistic methods in the particle physics community):

> *When the result of a measurement of a physics quantity is published as $R = R_0 \pm \sigma_0$ without further explanation, it simply implied that R is a Gaussian-distributed measurement with mean R_0 and variance σ_0^2. This allows to calculate various confidence intervals of given "probability", i.e. the "probability" P that the true value of R is within a given interval.*

(The quote marks are original and nowhere in the paper is explained why probability is in quote marks.)

indicating the former with a capital letter, as customary:

$$P[\theta - \Delta\theta_1 < \hat{\Theta} < \theta + \Delta\theta_2 \mid f(\hat{\theta}\,|\,\theta)] \;=\; \alpha\,, \tag{4}$$

that, for a Gaussian distribution of $\hat{\Theta}$ [indicated by the shorthand $\hat{\Theta} \sim \mathcal{N}(\theta,\sigma)$] and for $\Delta\theta_1 = \Delta\theta_2 = \sigma$, becomes

$$P[\theta - \sigma < \hat{\Theta} < \theta + \sigma \mid \hat{\Theta} \sim \mathcal{N}(\theta,\sigma)] \;=\; 68\%: \tag{5}$$

if we know the value of θ and the standard deviation of the distribution we can evaluate the probability that $\hat{\Theta}$ shall occur in the interval $[\theta-\sigma,\ \theta+\sigma]$, where 'to know' means that θ and σ have some numeric values, e.g. $\theta = 5$ and $\sigma = 2$. I find particularly enlightening to use, for a while, these particular values of μ and σ, rewriting Eq. (5) as

$$P[5-2 < \hat{\Theta} < 5+2 \mid \hat{\Theta} \sim \mathcal{N}(5,2)] \;=\; 68\%. \tag{6}$$

Obviously, knowing $\hat{\Theta} \sim \mathcal{N}(5,2)$, we can write an infinite number of probabilistic statements. In particular

$$P[\hat{\Theta} < 5-2 \mid \hat{\Theta} \sim \mathcal{N}(5,2)] \;=\; 16\% \tag{7}$$
$$P[\hat{\Theta} > 5+2 \mid \hat{\Theta} \sim \mathcal{N}(5,2)] \;=\; 16\%\,, \tag{8}$$

i.e.

$$P[\hat{\Theta}+2 < 5 \mid \hat{\Theta} \sim \mathcal{N}(5,2)] \;=\; 16\% \tag{9}$$
$$P[\hat{\Theta}-2 > 5 \mid \hat{\Theta} \sim \mathcal{N}(5,2)] \;=\; 16\%\,, \tag{10}$$

from which we have[5]

$$P[\hat{\Theta}-2 < 5 < \hat{\Theta}+2 \mid \hat{\Theta} \sim \mathcal{N}(5,2)] \;=\; 68\%. \tag{11}$$

Obviously, this expression is valid for any <u>known</u> value of θ and σ:

$$P[\hat{\Theta}-\sigma < \theta < \hat{\Theta}+\sigma \mid \hat{\Theta} \sim \mathcal{N}(\theta,\sigma)] \;=\; 68\%\,. \tag{12}$$

[5] I read once in a frequentistic book something like this: "*if you do not trust logic, prove it with a Monte Carlo*". These are the two lines of R code [6] needed to 'prove' by Monte Carlo the equality of Eqs. (6) and (11):

```
x ← rnorm(10000, 5, 2)
length(x[ (5-2) < x & x < (5+2) ]) == length(x[ (x-2) < 5 & 5 < (x+2) ]) .
```

Similarly, for an asymmetric interval, e.g. $\Delta\theta_1 = 2$ and $\Delta\theta_2 = 3$, we have

```
length(x[ (5-2) < x & x < (5+3) ]) == length(x[ (x-3) < 5 & 5 < (x+2) ]) .
```

The fact that we have replaced the numbers 5 and 2 by the generic symbols θ and σ changes nothing about meaning and possible use of (12): *just a rephrasing of a probabilistic statement about* $\hat{\Theta}$*!*

This is also the meaning of the 'probabilistic content inversion' (1)–(3): a simple rephrasing of a probabilistic statement about '$\hat{\theta}$' (that indeed should be $\hat{\Theta}$) under the assumption that we know the value of θ and the pdf of '$\hat{\theta}$' around θ:

$$P[\theta - \Delta\theta_1 < \hat{\Theta} < \theta + \Delta\theta_2 \mid f(\hat{\theta}\,|\,\theta)] \;=\; \alpha \tag{13}$$

equivalent to

$$P[\hat{\Theta} - \Delta\theta_2 < \theta < \hat{\Theta} + \Delta\theta_1 \mid f(\hat{\theta}\,|\,\theta)] \;=\; \alpha \tag{14}$$

Therefore, there is no doubt that (14) follows from (13). The question is that this is true if θ, $\Delta\theta_1$ and $\Delta\theta_2$ are real numbers, whose values, together with the knowledge of $f(\hat{\theta}\,|\,\theta)$, allow us to calculate α. However, rephrasing the probabilistic statement concerning the possible observation $\hat{\Theta}$ given a certain θ does not help us in solving the problem we are really interested in, i.e. to gain knowledge about the true value and to express our level of confidence on it, given the experimental observation $\hat{\theta}$.

What we are really looking for is, indeed, $P(\theta_1 < \Theta < \theta_2 \mid \hat{\theta})$. But this can be only achieved if we are able to write down, given the best knowledge of the physics case, the pdf $f(\theta\,|\,\hat{\theta})$. Pretending to express our confidence about Θ without passing through $f(\theta\,|\,\hat{\theta})$ is pure nonsense, based on a proof that reminds the 'game of the three cards' proposed by con artists in disreputable streets. Now, it is a matter of logic that the only way to go from $f(\hat{\theta}\,|\,\theta)$ to $f(\theta\,|\,\hat{\theta})$ is to make a correct 'probability inversion', following the rules of probability theory, instead of that shameful outrage against logic. The probabilistic tool to perform the task is Bayes' theorem, by which it is possible to establish intervals that contain the true value at a given level of *probability* (meant really by how much we are confident the true value is in a given interval!). It is easy to show that, under well defined conditions that often hold in routine applications, the interval calculated at a given level of probability α_p is equal to 'confidence interval' calculated with a 'confidence level' α_{CL}, if numerically[6] $\alpha_{CL} = \alpha_p$ (see e.g. Ref. [3]). It is not

[6] Those interested in Bayesian/frequentistic comparisons might give a look at Ref. [4]. Personally, as explained in Ref. [3] (footnote 18 of p. 229), I dislike the quantitative comparisons of Bayesian and frequentistic methods to solve the same problem simply because quantitative comparisons assume that we are dealing with homogeneous quantities, while frequentistic CL's and Bayesian probability intervals are as homogeneous as apples and tomatoes are. (This is also the reason I used here two different symbols, α_{CL} and α_p.)

a surprise, then, that the confidence interval prescriptions yield quite often a correct result, but they might also miserably fail, especially in frontier physics applications.

4 Conclusions

The proof of frequentistic 'confidence intervals' is sterile and there is no logical reason why one should attach a 'level of confidence' to the intervals calculated following that prescription. Paraphrasing a sentence of the same author of the opening quote, *they are called confidence intervals by their advocates because they provide confidence in no other way.*

References

[1] G. D'Agostini, *Bayesian Reasoning versus Conventional Statistics in High Energy Physics*, Proc. XVIII International Workshop on Maximum Entropy and Bayesian Methods, Garching (Germany), July 1998, V. Dose et al. eds., Kluwer Academic Publishers, Dordrecht, 1999 [`physics/9811046`].

[2] G. D'Agostini, *Confidence limits: what is the problem? Is there the solution?*, Workshop on Confidence Limits, Geneva, Switzerland, January 2000, CERN Report 2000-005 [`hep-ex/0002055`].

[3] G. D'Agostini, *Bayesian reasoning in data analysis. A critical Introduction*, World Scientific Publishing, 2003.

[4] G. Zech, *Frequentistic and Bayesian confidence limits*, EPJdirect **C12** (2002) 1 [`hep-ex/0106023`].

[5] F. James and M. Roos, *Errors on ratios of small numbers of events* Nucl. Phys. **B172** (1980) 475 [scanned version of preprint at KEK].

[6] The R language, `http://www.r-project.org/`.