ENHANCING THE PHYSICAL SIGNIFICANCE OF FREQUENTIST CONFIDENCE INTERVALS

Carlo Giunti
INFN, Sezione di Torino, and Dipartimento di Fisica Teorica, Università di Torino, Via P. Giuria 1, I–10125 Torino, Italy

Abstract
It is shown that all the Frequentist methods are equivalent from a statistical point of view, but the physical significance of the confidence intervals depends on the method. The Bayesian Ordering method is presented and confronted with the Unified Approach in the case of a Poisson process with background. Some criticisms to both methods are answered. It is also argued that a general Frequentist method is not needed.

1. INTRODUCTION

In this report I will be concerned mainly with the Frequentist (classical) theory of statistical inference, but I think that it is interesting and useful that I express my opinion on the war between Frequentists and Bayesians. To the question

"Are you Frequentist or Bayesian?"

I answer

"I like statistics."

I think that if one likes statistics, one can appreciate the beauty of both Frequentist and Bayesian theories and the subtleties involved in their formulation and application. I think that both approaches are valid from a statistical as well as physical point of view. Their difference arises from different definitions of probability and their results answer different statistical questions. One can like more one of the two theories, but I think that it is unreasonable to claim that only one of them is correct, as some partisans of that theory claim. These partisans often produce examples in which the other approach is shown to yield misleading or paradoxical results. I think that each theory should be appreciated and used in its limited range of validity, in order to answer the appropriate questions. Finding some example in which one approach fails does not disprove its correctness in many other cases that lie in its range of validity.

My impression is that the Bayesian theory (see, for example, [1]) has a wider range of validity because it can be applied to cases in which the experiment can be done only once or a few times (for example, our thoughts in everyday decisions and judgments seem to follow an approximate Bayesian method). In these cases the Bayesian definition of probability as *degree of belief* seems to me the only one that makes sense and is able to provide meaningful results.

Let me recall that since Galileo an accepted basis of scientific research is the *repeatability of experiments*. This assumption justifies the Frequentist definition of probability as ratio of the number of positive cases and total number of trials in a large ensemble. The concept of *coverage* follows immediately: a $100\alpha\%$ *confidence interval* for a physical quantity μ is an interval that contains (covers) the unknown true value of that quantity with a Frequentist probability α. In other words, a $100\alpha\%$ confidence interval for μ belongs to a set of confidence intervals that can be obtained with a large ensemble of experiments, $100\alpha\%$ of which contain the true value of μ.

2. THE STATISTICAL AND PHYSICAL SIGNIFICANCE OF CONFIDENCE INTERVALS

I think that in order to fully appreciate the meaning and usefulness of Frequentist confidence intervals obtained with Neyman's method [2, 3], it is important to understand that the experiments in the ensemble do not need to be identical, as often stated, or even similar, but can be real, different experiments [2, 4]. One can understand this property in a simple way [5] by considering, for example, two different experiments that measure the same physical quantity μ. The $100\alpha\%$ classical confidence interval obtained from the results of each experiment belongs by construction to a set of confidence intervals which can be obtained with an ensemble of identical experiments and contain the true value of μ with probability α. It is clear that the sum of these two sets of confidence intervals, containing the two confidence intervals obtained in the two different experiments, is still a set of confidence intervals that contain the true value of μ with probability α.

Moreover, for the same reasons it is clear that *the results of different experiments can also be analyzed with different Frequentist methods* [6], *i.e.* methods with correct coverage but different prescriptions for the construction of the confidence belt. This for me is amazing and beautiful: *whatever method you choose you get a result that can be compared meaningfully with the results obtained by different experiments using different methods*! It is important to realize, however, that the choice of the Frequentist method must be done independently of the knowledge of the data (before looking at the data), otherwise the property of coverage is lost, as in the "flip-flop" example in Ref. [7].

This property allow us to solve an apparent paradox that follows from the recent proliferation of proposed Frequentist methods [7, 8, 9, 10, 11, 12]. This proliferation seems to introduce a large degree of subjectivity in the Frequentist approach, supposed to be objective, due to the need to choose one specific prescription for the construction of the confidence belt, among several available with similar properties. From the property above, we see that whatever Frequentist method one chooses, if implemented correctly, the resulting confidence interval can be compared statistically with the confidence intervals of other experiments obtained with other Frequentist methods. Therefore, *the subjective choice of a specific Frequentist method does not have any effect from a statistical point of view*!

Then you should ask me:

Why are you proposing a specific Frequentist method?

The answer lies in *physics*, not statistics. It is well known that the statistical analysis of the same data with different Frequentist methods produce different confidence intervals. This difference is sometimes crucial for the physical interpretation of the result of the experiment (see, for example, [8, 10]). Hence, the physical significance of the confidence intervals obtained with different Frequentist methods is sometimes crucially different. In other words, *the Frequentist method suffers from a degree of subjectivity from a physical, not statistical, point of view*.

3. THE BEAUTY OF THE UNIFIED APPROACH AND ITS PITFALLS

The possibility to apply successfully Frequentist statistics to problematic cases in frontier research has received a fundamental contribution with the proposal of the Unified Approach by Feldman and Cousins [7]. The Unified Approach consists in a clever prescription for the construction of "a classical confidence belt which unifies the treatment of upper confidence limits for null results and two-sided confidence intervals for non-null results".

In the following I will consider the case of a Poisson process with signal μ and known background b. The probability to observe n events is

$$P(n|\mu, b) = \frac{(\mu + b)^n e^{-(\mu+b)}}{n!} \,. \tag{1}$$

The Unified Approach is based on the construction of the acceptance intervals $[n_1(\mu), n_2(\mu)]$ ordering the n's through their rank given by the relative magnitude of the likelihood ratio

$$R(n, \mu, b) = \frac{P(n|\mu, b)}{P(n|\mu_{\rm best}, b)} = \left(\frac{\mu + b}{\mu_{\rm best} + b}\right)^n e^{\mu_{\rm best} - \mu}\,, \tag{2}$$

where $\mu_{\rm best}$ is the maximum likelihood estimate of μ,

$$\mu_{\rm best}(n, b) = {\rm Max}[0, n - b]\,. \tag{3}$$

As a result of this construction the confidence intervals are two-sided (*i.e.* $[\mu_{\rm low}, \mu_{\rm up}]$ with $\mu_{\rm low} > 0$) for $n \gtrsim b$, whereas for $n \lesssim b$ they are upper limits (*i.e.* $\mu_{\rm low} = 0$).

The fact that the confidence intervals are two-sided for $n \gtrsim b$ can be understood by considering $n > b$, that gives $\mu_{\rm best} = n - b$. In this case the likelihood ratio (2) is given by

$$R(n > b, \mu, b) = \left(\frac{\mu + b}{n}\right)^n e^{n-(\mu+b)} = \exp\left\{n\left[1 + \ln(\mu + b) - \ln n\right] - (\mu + b)\right\} \stackrel{n\to\infty}{\longrightarrow} 0\,. \tag{4}$$

This implies that the rank of high values of n is very low and they are excluded from the confidence belt. Therefore, the acceptance intervals $[n_1(\mu), n_2(\mu)]$ are always bounded, *i.e.* $n_2(\mu)$ is finite, and the confidence intervals are two-sided for $n \gtrsim b$, as illustrated in Fig. 1, where the solid lines show the borders of the confidence belt for a background $b = 5$ and a confidence level $\alpha = 0.90$.

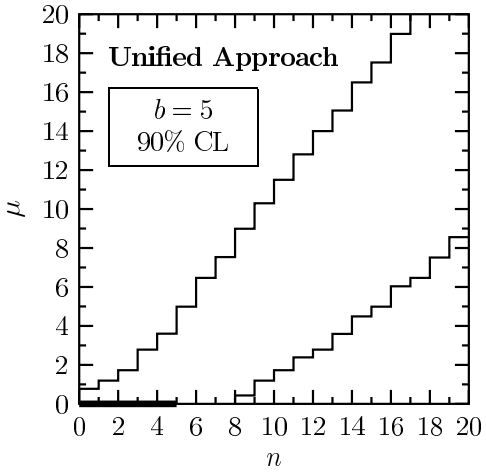

Fig. 1: Confidence belt in the Unified Approach for background $b = 5$ and confidence level $\alpha = 0.90$.

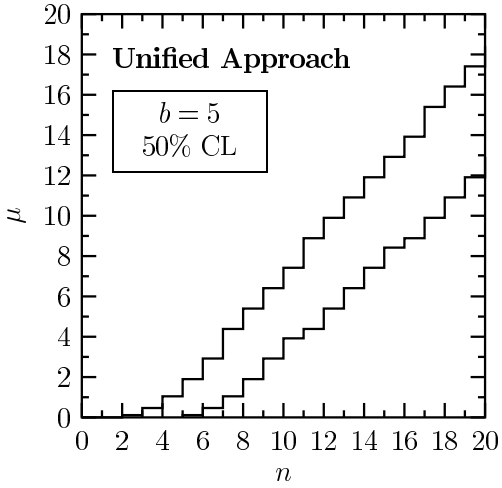

Fig. 2: Confidence belt in the Unified Approach for background $b = 5$ and confidence level $\alpha = 0.50$.

The fact that the confidence intervals are upper limits for $n \lesssim b$ can be understood by considering $n \leq b$, for which we have $\mu_{\rm best} = 0$ and the likelihood ratio that determines the ordering of the n's in the acceptance intervals is given by

$$R(n{\leq}b, \mu, b) = \left(1 + \frac{\mu}{b}\right)^n e^{-\mu}\,. \tag{5}$$

Considering now the acceptance interval for $\mu = 0$, we have $R(n \leq b, \mu = 0, b) = 1$. Therefore, all $n \leq b$ for $\mu = 0$ have highest rank and are guaranteed to lie in the confidence belt. This is illustrated in Fig. 1, where the thick solid segment shows the $n \leq b$ part of the acceptance interval for $\mu = 0$, that must lie in the confidence belt. Since μ is a continuous parameter, also for small values of μ the $n \leq b$

have rank close to the highest one and lie in the confidence belt. Indeed, for $\mu > 0$, the likelihood ratio (2) increases for n going from zero to the largest integer smaller or equal to b and decreases for larger values of n. Hence, the largest integer $n_{\rm hr}$ such that $n_{\rm hr} \leq b$ has highest rank. If μ is sufficiently small all $n \leq b$ have rank close to maximum and are included in the confidence belt if the confidence level is large enough, $\alpha \gtrsim 0.60$. For example, $R(n = 0, \mu, b) > R(n_{\rm hr} + 1, \mu, b)$ for $\mu < (1 + b)e^{-1/(1+b)} - b$. Therefore, the left edge of the confidence belt must change its slope for $n \lesssim b$ and intercept the μ-axis at a positive value of μ, as illustrated in Fig. 1. The value of μ at which the left edge of the confidence belt intercepts the μ-axis, that corresponds to $\mu_{\rm up}(n = 0)$, depends on the value of the background b and on the value of the confidence level α.

However[1], for small values of α the Unified Approach gives zero-width confidence intervals for $n \ll b$, as illustrated in Fig. 2, where I have chosen $b = 5$ and $\alpha = 0.50$. One can see that the segment $n \leq b$ is enclosed in the confidence belt for $\mu = 0$, but for any value of $\mu > 0$ the sum of the probabilities of the n's close to $\mu + b$ is enough to reach the confidence level and low values of n are not included in the confidence belt. Hence, in this case the Unified Approach gives zero-width confidence intervals for $n < 2$.

The unification of the treatments of upper confidence limits for null results and two-sided confidence intervals for non-null results obtained with the Unified Approach is wonderful, but it has been noticed that the upper limits obtained with the Unified Approach for $n < b$ are too stringent (meaningless) from a physical point of view [8, 13]. In other words, although these limits are statistically correct from a Frequentist point of view, they cannot be taken as reliable upper bounds to be used in physical applications.

This problem is illustrated in Fig. 3A, where I plotted the 90% CL upper limit $\mu_{\rm up}$ as a function of b for $n = 0, \ldots, 5$. The solid part of each line shows where $b \geq n$. One can see that for a given n, $\mu_{\rm up}$ decreases rather steeply when b is increased, until a minimum value close to one is reached. The curves have jumps because n is an integer and generally the desired confidence level cannot be obtained exactly, but with some unavoidable overcoverage.

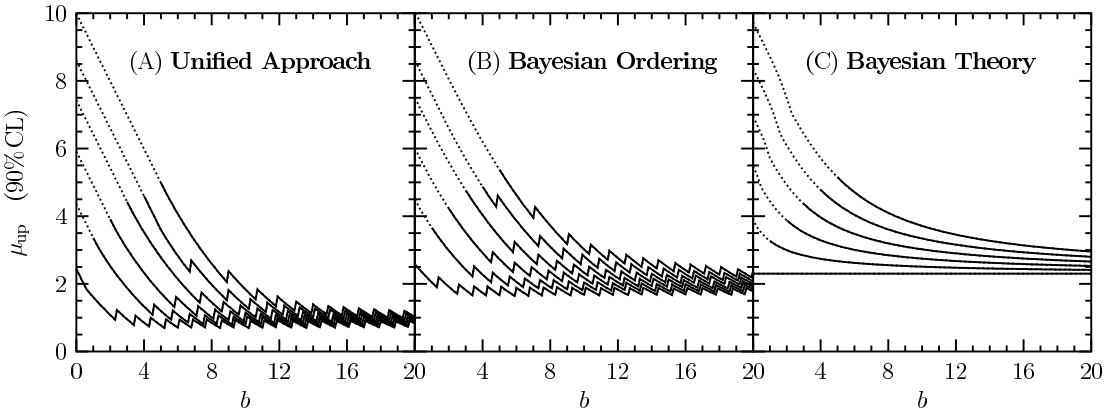

Fig. 3: 90% CL upper limit $\mu_{\rm up}$ as a function of the background b for $n = 0$ (lower lines), . . ., $n = 5$ (upper lines). The solid part of each line shows where $b \geq n$.

Let me emphasize that the problem of obtaining too stringent upper limits for $n < b$ is very serious for a scientist that wants to obtain reliable information from experiment and use this information for other purposes (as input for a theory or another experiment). In the past, researchers bearing the same physical point of view refrained to report empty confidence intervals or very stringent upper limits when $n < b$

[1]Let me emphasize that I discuss this case only for the sake of curiosity. It is pretty obvious that a low value of α is devoid of any practical interest.

was measured. These confidence intervals are correct from a statistical point of view, but useless from a physical point of view. Furthermore, the same reasoning lead to prefer the Unified Approach to central confidence intervals or upper limits, because the non-empty confidence interval obtained when $n < b$ is measured is certainly more significant, from a physical point of view, than an empty one, although they are statistically equivalent, as shown in Section 2..

4. A BRUTAL MODIFICATION OF THE UNIFIED APPROACH

In the Unified Approach $\mu_{\rm best}$ is positive and equal to zero for $n \leq b$. If $\mu_{\rm best}$ is forced to be always bigger than zero, the n's smaller than b have rank higher than in the Unified Approach. As a consequence, the decrease of the upper limit $\mu_{\rm up}$ as b increases is weakened. This is illustrated by a *"Brutally Modified Unified Approach"* (BMUA) in which we take

$$\mu_{\rm best} = {\rm Max}[\mu_{\rm best}^{\rm min}, n - b]\,, \tag{6}$$

where $\mu_{\rm best}^{\rm min}$ is a positive real number.

In Fig. 4 I plotted the confidence belts for $\mu_{\rm best}^{\rm min} = 0$ (solid lines), that corresponds to the Unified Approach, $\mu_{\rm best}^{\rm min} = 1$ (dashed lines) and $\mu_{\rm best}^{\rm min} = 2$ (dotted lines), for $b = 10$. One can see that in the BMUA the upper limits of the confidence intervals are considerably higher than in the Unified Approach. The behavior of $\mu_{\rm up}$ as a function of b for $n = 0$ is shown in Fig. 5, from which it is clear that the decrease of $\mu_{\rm up}$ when b increases is much weaker in the BMUA (dashed and dotted lines) than in the Unified Approach (solid line) and it is almost absent for $\mu_{\rm best}^{\rm min} \gtrsim 2$.

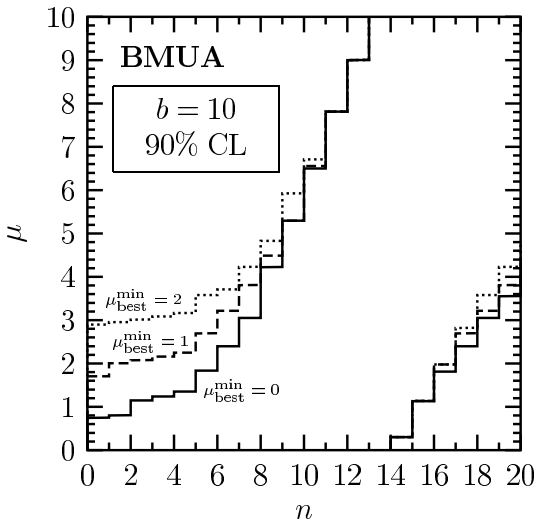

Fig. 4: 90% confidence belts for $b = 10$ in the Unified Approach ($\mu_{\rm best}^{\rm min} = 0$, solid lines) and in the Brutally Modified Unified Approach (BMUA) for $\mu_{\rm best}^{\rm min} = 1$ (dashed lines) and $\mu_{\rm best}^{\rm min} = 2$ (dotted lines).

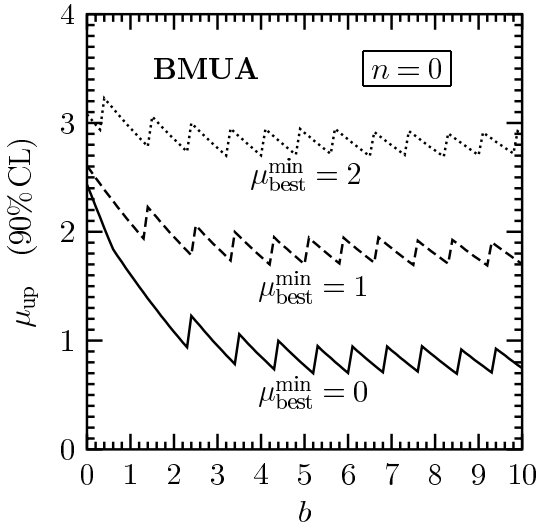

Fig. 5: 90% CL upper limit $\mu_{\rm up}$ as a function of the background b for $n = 0$ in the Unified Approach ($\mu_{\rm best}^{\rm min} = 0$, solid line) and in the BMUA for $\mu_{\rm best}^{\rm min} = 1$ (dashed line) and $\mu_{\rm best}^{\rm min} = 2$ (dotted line).

Let me emphasize that

1. The BMUA is a statistically correct Frequentist method and coverage is satisfied.
2. In the BMUA one obtains upper limits for $n \lesssim b$ and central confidence intervals for $n \gtrsim b$, as in the Unified Approach[2].

[2]For $n \leq b + \mu_{\rm best}^{\rm min}$ we have $\mu_{\rm best} = \mu_{\rm best}^{\rm min}$ and the likelihood ratio (2) becomes

$$R(n \leq b + \mu_{\rm best}^{\rm min}, \mu, b) = \left(\frac{\mu + b}{\mu_{\rm best}^{\rm min} + b}\right)^n e^{\mu_{\rm best}^{\rm min} - \mu}\,. \tag{7}$$

3. The BMUA method is not general (although it can be extended in an obvious way at least to the case of a gaussian variable with a physical boundary).
4. *I am not proposing the BMUA*! (But those that think that the upper limit for $n = 0$ should not depend on b may consider the possibility of using the BMUA with $\mu_{\rm best}^{\rm min} = 2$ instead of resorting to more complicated methods that may even jeopardize the property of coverage[3].)

As shown in Fig. 4, the right edge of the confidence belt in the BMUA is not very different from the one in the Unified Approach. This is due to the fact that adding small values of n with low probability to the acceptance intervals has little effect. Moreover, it is clear that the acceptance interval for $\mu = 0$ is equal for all Frequentist methods with correct coverage that unify the treatment of upper confidence limits and two-sided confidence intervals.

5. BAYESIAN ORDERING

An elegant, natural and general way to obtain automatically $\mu_{\rm best}^{\rm min} > 0$ is given by the *Bayesian Ordering* method [8], in which $\mu_{\rm best}$ is replaced by the Bayesian expectation value for μ, $\mu_{\rm B}$.

Choosing a natural flat prior, the Bayesian expectation value for μ in a Poisson process with background is given by

$$\mu_{\rm B}(n,b) = n + 1 - \left(\sum_{k=0}^{n} \frac{k b^k}{k!}\right)\left(\sum_{k=0}^{n} \frac{b^k}{k!}\right)^{-1} = n + 1 - b\left(\sum_{k=0}^{n-1} \frac{b^k}{k!}\right)\left(\sum_{k=0}^{n} \frac{b^k}{k!}\right)^{-1} . \qquad (8)$$

The obvious inequality $\sum_{k=0}^{n} k\, b^k/k! \leq n \sum_{k=0}^{n} b^k/k!$ implies that $\mu_{\rm B} \geq 1$. Therefore, the reference value for μ in the likelihood ratio

$$R(n,\mu,b) = \frac{P(n|\mu,b)}{P(n|\mu_{\rm B},b)} = \left(\frac{\mu+b}{\mu_{\rm B}+b}\right)^n e^{\mu_{\rm B}-\mu} , \qquad (9)$$

that determines the construction of the acceptance intervals as in the Unified Approach, is bigger or equal than one. As a consequence, the decrease of the upper confidence limit $\mu_{\rm up}$ for a given n when the expected background b increases is significantly weaker than in the Unified Approach, as illustrated in Fig. 3B.

Figure 3C shows $\mu_{\rm up}$ as a function of b in the Bayesian Theory with a flat prior and shortest credibility intervals[4]. One can see that the behavior of $\mu_{\rm up}$ obtained with the Bayesian Ordering method

For $\mu < \mu_{\rm best}^{\rm min}$, we have $(\mu+b)/(\mu_{\rm best}^{\rm min}+b) < 1$ and $R(n \leq b + \mu_{\rm best}^{\rm min}, \mu, b)$ decreases with increasing n. Let us consider now $n > b + \mu_{\rm best}^{\rm min}$, for which $\mu_{\rm best} = n - b$ and the likelihood ratio (2) is given by the expression in Eq. (4). This expression has a maximum for n equal to one of the two integers closest to $\mu + b$. For $\mu < \mu_{\rm best}^{\rm min}$, this integer is the first one in the considered range ($n > b + \mu_{\rm best}^{\rm min}$). Therefore, for sufficiently low values of μ, $\mu < \mu_{\rm best}^{\rm min}$, the likelihood ratio (2) decreases monotonically as n increases. In this case, low values of n have highest ranks and are guaranteed to lie in the confidence belt and the left edge of the confidence belt must change its slope for $n \lesssim \mu_{\rm best}^{\rm min} + b$ and intercept the μ-axis at a positive value of μ, as illustrated in Fig. 4.

[3] By the way, I think that coverage is the most important property of the Frequentist theory. If coverage is not satisfied the results are statistically useless in the contest of Frequentist theory.

[4] In this case the posterior p.d.f. for μ is

$$P(\mu|n,b) = (b+\mu)^n\, e^{-\mu} \left(n! \sum_{k=0}^{n} \frac{b^k}{k!}\right)^{-1} , \qquad (10)$$

and the probability (degree of belief) that the true value of μ lies in the range $[\mu_1, \mu_2]$ is given by

$$P(\mu \in [\mu_1,\mu_2]|n,b) = \left(e^{-\mu_1} \sum_{k=0}^{n} \frac{(b+\mu_1)^k}{k!} - e^{-\mu_2} \sum_{k=0}^{n} \frac{(b+\mu_2)^k}{k!}\right)\left(\sum_{k=0}^{n} \frac{b^k}{k!}\right)^{-1} . \qquad (11)$$

The shortest $100\alpha\%$ credibility intervals $[\mu_{\rm low}, \mu_{\rm up}]$ are obtained by choosing $\mu_{\rm low}$ and $\mu_{\rm up}$ such that $P(\mu \in [\mu_{\rm low}, \mu_{\rm up}]|n,b) = \alpha$ and $P(\mu_{\rm low}|n,b) = P(\mu_{\rm up}|n,b)$ if possible (with $\mu_{\rm low} \geq 0$), or $\mu_{\rm low} = 0$.

is intermediate between those in the Unified Approach and in the Bayesian Theory. Although one must always remember that the statistical meaning of $\mu_{\rm up}$ is different in the two Frequentist methods (Unified Approach and Bayesian Ordering) and in the Bayesian Theory, for scientists using these upper limits it is often irrelevant how they have been obtained. Hence, I think that an approximate agreement between Frequentist and Bayesian results is desirable.

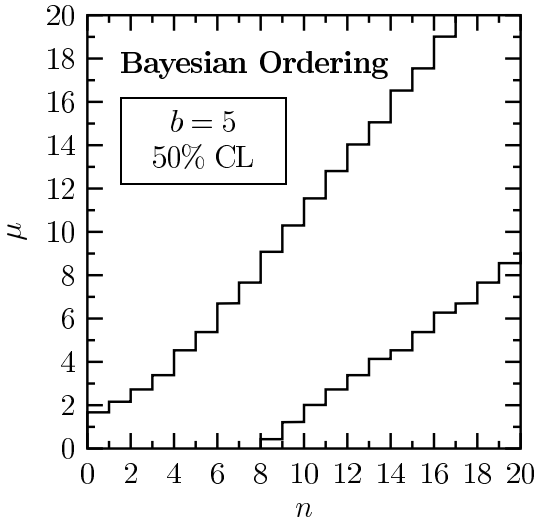

Fig. 6: Confidence belt obtained with the Bayesian Ordering for background $b = 5$ and confidence level $\alpha = 0.90$.

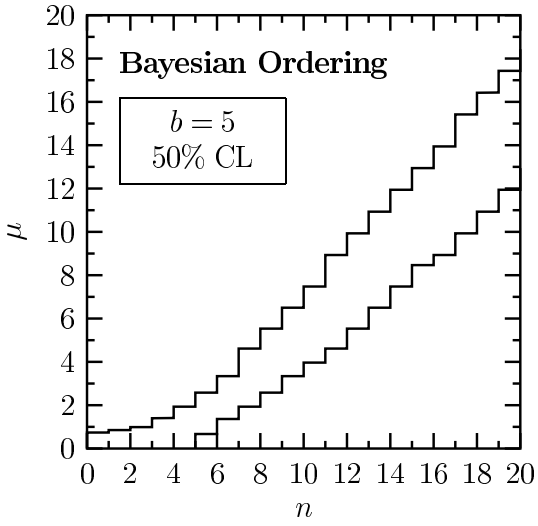

Fig. 7: Confidence belt obtained with the Bayesian Ordering for background $b = 5$ and confidence level $\alpha = 0.50$.

From Eq. (8) one can see that

$$n \gg b \quad \Longrightarrow \quad \mu_{\rm B}(n,b) \simeq n + 1 - b \simeq n\,, \tag{12}$$

$$n \lesssim b\,, \quad b \gg 1 \quad \Longrightarrow \quad \mu_{\rm B}(n,b) \simeq 1\,. \tag{13}$$

Therefore, for $n \gg b$ the confidence belt obtained with the Bayesian Ordering method is similar to that obtained with the Unified Approach. The difference between the two methods show up only for $n \lesssim b$. This is illustrated in Figs. 6 and 7, that must be confronted with the corresponding Figures 1 and 2 in the Unified Approach. Notice that, as shown in Fig. 7, contrary to the Unified Approach, the Bayesian Ordering method gives physically significant (non-zero-width) confidence intervals even for low values of the confidence level α.

6. ANSWERS TO SOME CRITICISMS

Criticism: *Bayesian Ordering is a mixture of Frequentism and Bayesianism. The uncompromising Frequentist cannot accept it.*

No! It is a Frequentist method.

Bayesian theory is only used for the *choice of ordering* in the construction of the acceptance intervals, that in any case is subjective and beyond Frequentism (as, for example, the central interval prescription or the Unified Approach method). The Bayesian method for such a subjective choice is quite natural.

If you belong to the Frequentist Orthodoxy (sort of religion!) and the word "Bayesian" gives you the creeps, you can change the name "Bayesian Ordering" into whatever you like and use its prescription for the construction of the acceptance intervals as a successful recipe.

Criticism: *In the Unified Approach (and maybe Bayesian Ordering?) the upper limit on μ goes to zero for every n as b goes to infinity, so that a low fluctuation of the background entitles to claim a very stringent limit on the signal.*

This is not true!

One can see it[5] doing a calculation of the upper limit for μ as a function of b for large values of b. The result of such a calculation in the Unified Approach is shown in Fig. 8A, where the 90% CL upper limit $\mu_{\rm up}$ is plotted as a function of b in the interval $0 \leq b \leq 200$ for $n = 0$ (solid line), $n = 5$ (dashed line) and $n = 10$ (dotted line). One can see that initially $\mu_{\rm up}$ decreases with increasing b, but it stabilizes to about 0.8 for $b \gg n$, with fluctuations due to the discreteness of n. Figure 8B shows the same plot obtained with the Bayesian Ordering. One can see that initially $\mu_{\rm up}$ decreases with increasing b, but less steeply than in the Unified Approach, and it stabilizes to about 1.8. For comparison, in Fig. 8C I plotted $\mu_{\rm up}$ as a function of b in the Bayesian Theory with a flat prior and shortest credibility intervals. One can see that the behavior of $\mu_{\rm up}$ in the three methods considered in Fig. 8 is rather similar.

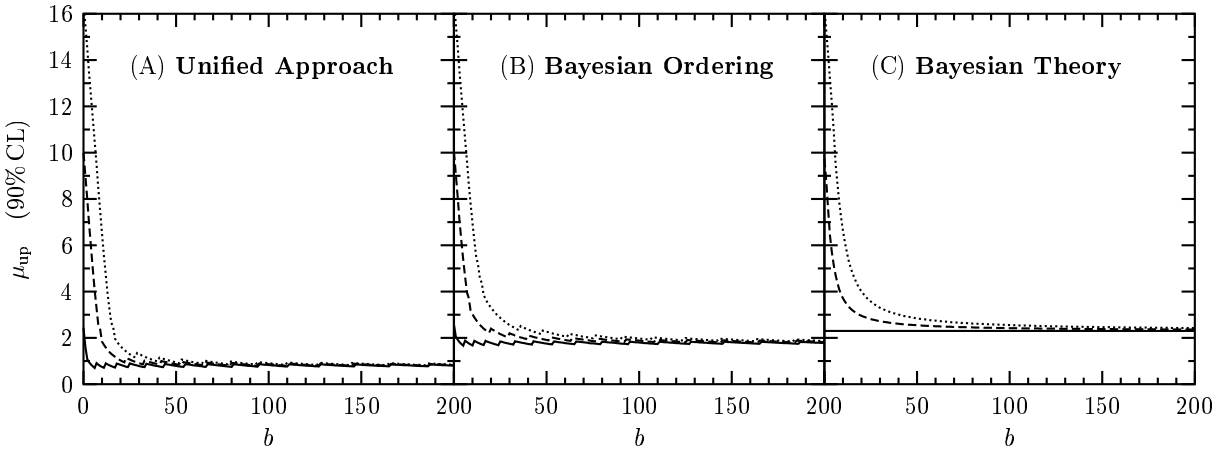

Fig. 8: 90% CL upper limit $\mu_{\rm up}$ as a function of the background b for $n = 0$ (solid lines), $n = 5$ (dashed lines) and $n = 10$ (dotted lines).

Criticism: *For $n = 0$ the upper limit $\mu_{\rm up}$ should be independent of the background b.*

But for $n > 0$ the upper limit $\mu_{\rm up}$ always decreases with increasing b! It is true that for $n = 0$ one is sure that no background event as well as no signal has been observed. But this is just the effect of a low fluctuation of the background that *is present*! Should we built a special theory for $n = 0$? I think that this is not interesting in the Frequentist framework, because I guess that it leads necessarily to a violation of coverage (that could be tolerated, but not welcomed, only if it is overcoverage).

I think that if one is so interested in having an upper limit $\mu_{\rm up}$ independent of the background b for $n = 0$, one better embrace the Bayesian theory (see Fig. 3C, Fig. 8C and Ref. [14]), which, by the way, may present many other attractive qualities (see, for example, [1]).

Criticism: *A (worse) experiment with larger background b should not give a smaller upper limit $\mu_{\rm up}$ for the same number n of observed events.*

But, as shown in Fig. 3, this always happens! Notice that it happens both for $n > b$ (dotted part of lines) and for $n \leq b$ (solid part of lines), in Frequentist methods as well as in the Bayesian Theory (for $n > 0$). As far as I know, nobody questions the decrease of $\mu_{\rm up}$ as b is increased if $n > b$. So why should we question the same behavior when $n \leq b$? The reason for this behavior is simple: the observation of a

[5] In the Unified Approach the likelihood ratio for $n \leq b$ is given by the expression in Eq. (5), that tends to $e^{-\mu}$ for $b \gg n$ and small μ. For $\mu \ll 1$, $e^{-\mu} \simeq 1$ and all $n \ll b$ have rank close to maximum. For $n > b$ the likelihood ratio is given by the expression in Eq. (4). For large values of b, taking into account that $n > b$, we have $1 + \ln(\mu + b) - \ln n \simeq \ln b - \ln n < 0$ and $\mu + b \simeq b$, which imply that $R(n > b, \mu, b) < e^{-b} \stackrel{b \to \infty}{\longrightarrow} 0$. So the rank drops rapidly for $n > b$. Therefore, for small values of μ the n's much smaller than b have highest rank. Since they have also very small probability, they all lie comfortably in the confidence belt, if the confidence level α is sufficiently large ($\alpha \gtrsim 0.60$).

given number n of observed events has the same probability if the background is small and the signal is large or the background is large and the signal is small.

I think that it is physically desirable that and experiment with a larger background do not give a *much smaller* upper limit for the same number of observed events, but a *smaller* upper limit is allowed by *statistical fluctuation*. Indeed,

> upper limits (as confidence intervals, etc.) are statistical quantities that *must fluctuate!*

I think that the current race of experiments to find the most stringent upper limit is bad[6], because it induces people to think that limits are fixed and certain. Instead, everybody should understand that

> a better experiment can sometimes give a worse upper limit because of statistical fluctuations and there is nothing wrong about it!

7. CONCLUSIONS

In this report I have shown that the necessity to choose a specific Frequentist method, among several available, does not introduce any degree of subjectivity from a statistical point of view (Section 2.) [6]. In other words, all Frequentist methods are statistically equivalent.

However, the physical significance of confidence intervals obtained with different methods is different and scientists interested in obtaining reliable and useful information on the characteristics of the real world must worry about this problem. Obtaining empty or very small confidence intervals for a physical quantity as a result of a statistical procedure is useless. Sometimes it is even dangerous to present such results, that lead non-experts in statistics (and sometimes experts too) to false beliefs.

In Section 3. I have discussed some virtues and shortcomings of the Unified Approach [7]. These shortcomings are ameliorated in the Bayesian Ordering method [8], discussed in Section 5., that is natural, relatively easy, and leads to more reliable upper limits.

In conclusion, I would like to emphasize the following considerations:

- One must always remember that, in order to have coverage, the choice of a specific Frequentist method must be done independently of the knowledge of the data.
- Finding some examples in which a method fails does not imply that it should not be adopted in the cases in which it performs well.
- Since all Frequentist methods are statistically equivalent,

  there is no need of a general Frequentist method!

  In each case one can choose the method that works better (basing the judgment on easiness, meaningfulness of limits, etc.). Complicated methods with a wider range of applicability are theoretically interesting, but not attractive in practice.
- Somebody thinks that the physics community should agree on a standard statistical method (see, for example, [15])[7]. In that case, it is clear that this method must be always applicable. But this is not the case, for example, of the Unified Approach, as shown in [16]. Although the Bayesian Ordering method has not been submitted to a similar thorough examination, I doubt that it is generally applicable.

[6] It is surprising that even at the Panel Discussion [15] of this Workshop (full of experts) the statement "the experimenters like to quote the smallest bound they can get away with" was not strongly criticized. What is the purpose of experiments? (A) Give the smallest bound. (B) Give useful and reliable information. If your answer is (A) and you are an experimentalist, I suggest that you stop deceiving us and move to some more rewarding cheating activity.

[7] As a theorist, I find the argument, presented by an experimentalist, that a standard is useful because otherwise one is tempted to analyze the data with the method that gives the desired result quite puzzling. But if I were an experimentalist I would be quite offended by it. Isn't it a denigration of the professional integrity of experimental physicists?

I do not see why experiments that explore different physics and use different experimental techniques should all use the same statistical method (except a possible ignorance of statistics and blind belief of "authorities").
I would recommend that

> instead of wasting time on useless characteristics as generality, *the physics community should worry about the usefulness and credibility of experimental results*.

Acknowledgements

I would like to thank Marco Laveder for fruitful collaboration and many stimulating discussions.

References

[1] G. D'Agostini, CERN Yellow Report 99-03 (available at http://www-zeus.roma1.infn.it/%7Eagostini/prob+stat.html); Am. J. Phys. **67**, 1260 (1999) [arXiv:physics/9908014]; arXiv:physics/9906048.

[2] Philos. Trans. R. Soc. London Sect. A **236**, 333 (1937), reprinted in *A selection of Early Statistical Papers on J. Neyman*, University of California, Berkeley, 1967, p. 250.

[3] W.T. Eadie, D. Drijard, F.E. James, M. Roos and B. Sadoulet, *Statistical Methods in Experimental Physics*, North Holland, Amsterdam, 1971.

[4] R.D. Cousins, Am. J. Phys. **63**, 398 (1995).

[5] C. Giunti, Phys. Rev. D **59**, 113009 (1999) [arXiv:hep-ex/9901015].

[6] C. Giunti and M. Laveder, arXiv:hep-ex/0002020.

[7] G.J. Feldman and R.D. Cousins, Phys. Rev. D **57**, 3873 (1998) [arXiv:physics/9711021].

[8] C. Giunti, Phys. Rev. D **59**, 053001 (1999) [arXiv:hep-ph/9808240].

[9] S. Ciampolillo, Il Nuovo Cimento A **111**, 1415 (1998).

[10] B.P. Roe and M.B. Woodroofe, Phys. Rev. D **60**, 053009 (1999) [arXiv:physics/9812036].

[11] M. Mandelkern and J. Schultz, arXiv:hep-ex/9910041.

[12] G. Punzi, arXiv:hep-ex/9912048.

[13] C. Giunti, in *Summary of the NOW'98 Phenomenology Working Group* [arXiv:hep-ph/9906251].

[14] P. Astone and G. Pizzella, these Proceedings [hep-ex/0002028].

[15] "Panel Discussion" at this this Workshop (http://www.cern.ch/CERN/Divisions/EP/Events/CLW).

[16] G. Zech, these Proceedings (http://www.cern.ch/CERN/Divisions/EP/Events/CLW).

Discussion after talk of Carlo Giunti. Chairman: Jim Linnemann.

Bob Cousins

In the previous talk I showed on the page from Kendall and Stuart and Ord that the likelihood ratio ordering is anything but arbitrary. It is inspired by the Neyman-Pearson lemma which is hardly arbitrary.

C. Giunti

What do you mean - Is that a bible?

R. Cousins

The Neyman-Pearson lemma shows that the likelihood ratio is the optimal way to classify events. We thought about it a lot and we don't completely understand how to leap from there to confidence intervals, but the fact that an ordering based on the Neyman-Pearson lemma is the optimal way to separate out signal from background is not arbitrary.

C. Giunti

No, but that has nothing to do with ordering. So, I understand that the maximum likelihood is a useful quantity, but ...

R. Cousins

It's not the maximum likelihood, it is the likelihood *ratio*.

Gary Feldman

Can I add to that comment? You made the statement that you get coverage in all cases, so it's arbitrary which ordering principle you use. This is not true. The point is that there are two types of errors you can make in statistics. The error of the first type is to reject a true hypothesis. That we do a fixed fraction of the time if we do statistics and that's the confidence level. The error of the second type is to accept a false hypothesis. This is the power of the technique, and generally you strive for the technique which is most powerful. Now, when you have two-sided intervals there is no uniformly most powerful method. However, for one sided intervals there is, and you will notice that in the case where your intervals are one-sided, compared to where they are one-sided in the unified method, that the unified method is more powerful, in other words the intervals are smaller.

C. Giunti

You say that the limit is more stringent. Let's take a specific example. For the Poisson with background, when the number of events is bigger than the background, then the limit is very similar in the two cases. When you go below the background, then the limit in the unified approach is significantly more stringent, but this, from the physical point of view, is not meaningful.

Jacques Bouchez

Maybe Bob can comment also on my question. I wanted to know on which criteria you judge that one ordering method is better than another. In Bob's list of properties of the Feldman-Cousins method, one was: There is an improvement over central intervals. In what sense do you think it's an improvement?

Just because there are fewer null results? Is it not subjective to consider that the null hypothesis is bad or good?

C. Giunti

The choice of the methods is always subjective. Take, for example, the case of the Karmen I experiment which is very well-known. In the Neutrino '98 conference, they used Feldman and Cousins' method and they observed zero events with a background of 2.88, and then with this they claimed that they were in contradiction with the LSND experiment. But people believed that. So if you take it only statistically, it's very good.

J. Bouchez

Is it worse to publish signal lower than minus 2, or -0.0001, which doesn't seem to please you, or lower than one, depending on the method you choose? Is there a criterion to decide what is the best ordering?

C. Giunti

I don't know if there is a best ordering. I am proposing this as an improvement. I am not claiming that it is the best.

J. Bouchez

Why do you think it is an improvement?

C. Giunti

Just what I said. If you measure fewer events than you expect from background, then your limits are not unphysically narrow.

J. Bouchez

And why does Bob think that the Feldman-Cousins method is better than central intervals?

R. Cousins

Let me refer back to this plot from my talk. These are two methods for calculating neutrino oscillations limits which both have exact frequentist coverage, and the point Gary was trying to make was that two methods with exact frequentist coverage are not equivalent if they are very different in rejecting false hypotheses. Any statistician will tell you that you want to minimize both types of errors. There is a trade-off you can argue about, but if two methods both give frequentist coverage, certainly one criterion for preferring one or the other is the power to reject false hypotheses.

C. Giunti

What is the problem with false hypotheses? Here we are giving some range of parameters so there is no false hypothesis.

R. Cousins

Both those methods happen to give an interval that actually covers the true value, and one of them covers a whole lot of values that are not the true value, and the other one covers just a few values that are not the true value.

Don Groom

Maybe I misunderstood you, but I thought you said that the case of zero events wasn't especially different from any other case, and you seem happy that the limits should be dependent on the background. And yet you know for sure that there are no signal events and that's a totally independent fact from whatever the background is. Why doesn't the limit have to be independent of B?

C. Giunti

Well, I am also asking why not? When $N = 0$ it is true that there is no background and no signal observed. But still there is background expected, so in my opinion we could formulate a special theory for $N = 0$, but it would not be general. We should treat all N in the same way. For example here I coloured the curves only when B is bigger than N. So you can see that in both the unified approaches, there is a decrease. In Bayesian ordering the decrease is less steep and the minimum upper limit that one can get is higher. But I think nobody can question, for example, this part of the curves. Indeed, what do you have here ? If you measure a given number of events, the limit will be stronger if the expected background is higher. So this is a natural behaviour, as I said, because there is less room for a true event, and this is true also when $N = 0$. When $N = 0$, if you measure $N = 0$ there is less room for a true event.

Peter Clifford

The likelihood ratio, of course, finds the optimal test for one simple hypothesis against a simple hypothesis. When you put in the denominator the maximum likelihood estimate, you're saying 'well let's construct a test against the specific value of the parameter, let's make that parameter the one which, in a sense, would be the most challenging one to test against'. It's not the likelihood ratio with the maximum likelihood estimate, it's not the uniformly most powerful test, because there isn't a uniformly most powerful test. I just wanted to eliminate that confusion which might have crept in. So, the question is 'what value of the parameter decides the one you're testing?' Should you be trying to design your test to be optimal against. I would tend to support you when saying 'well, that's your choice, and you've chosen a parameter value which is the expected value according to some Bayesian calculation' whereas you could choose a parameter according to a maximum likelihood criterion.

C. Giunti

I make this choice only for physical reasons, not for statistical reasons.