PROCEEDINGS OF SCIENCE

Pseudosignificances as figures of merit: a systematic study and exploration of Bayesian solutions

Pietro Vischia[*]
[†]

Institut de recherche en Mathématique et Physique,
Université catholique de Louvain
E-mail: pietro.vischia@cern.ch

Optimization problems in HEP often involve maximizing a measure of how sensitive is a given analysis to an hypothesis with respect to another hypothesis; the latter is referred to as *null* hypothesis and in a frequentist framework is tested against the former, which is referred to as *alternative* hypothesis.

In most cases, it is desirable to fully compute the expected frequentist significance, accounting for all sources of systematic uncertainty and interpreting the result as the real sensitivity of the analysis to the effect sought. Sometimes, however, either computational or conceptual reasons can favour the use of different or approximate figures of merit, often collectively called "pseudosignificances", which exhibit different properties depending on the relationship between the hypotheses being tested.

This work will review the most common definitions of sensitivity (pseudosignificances), and compare them with the fully frequentist significances computed in toy analyses spanning a spectrum of typical HEP use cases. A connection will be made with the concept of Bayes Factor, and evidence values from Bayesian significance tests will be studied and evaluated in the same toy cases, attempting to build an improved approximate condition-aspecific figure of merit. Finally, an attempt will be made at transporting to the typical HEP cases a Bayesian solutions to the on-off problem developed in an astrophysics context.

XIII Quark Confinement and the Hadron Spectrum - Confinement2018
31 July - 6 August 2018
Maynooth University, Ireland

[*] Speaker.

[†] P.V. is funded by a UCLouvain FSR grant (BE)

1. Introduction

Statistics is all about answering questions. In High Energy Physics (HEP), oftentimes the question is: if I design the experiment in this or that way, will we be able to observe the phenomenon we seek, assuming the phenomenon itself occurs? The question is usually generalized in the context of experiment design as an optimization problem: which are the experimental settings that maximize my ability of observing the phenomenon I seek, assuming the phenomenon itself occurs? In statistics it is common practice to frame such problems as *hypothesis testing* problems; a *null* hypothesis , usually taken as the well established best-theory-so-far, is tested against an *alternative* hypothesis. The sought phenomenon does not necessarily consist in new, unobserved physics; in general, it just corresponds to a different hypothesis than the well-established one. In HEP, the most common tests involve a null hypothesis consisting in event counts originated by well-known physics processes (*background*); the alternative consists in event counts originated by a sum of the counts from well-known physics processes and the counts from an additional process (*signal*), to form the *signal-plus-background* hypothesis. In the following we use the notation S for the signal counts and B for the background counts.

In this framework, the optimization of experiment design aims to predict to a good degree of approximation the *sensitivity* of the experiment, and to compare the expected sensitivity of various experiments (or configurations of the same one) in order to decide the best one to pursue.

A suitable definition of sensitivity is needed. For problems of optimizing experiment design, the most common definition of sensitivity is that of estimated median significance; this is derived from the general calculations of the significance associated to an experiment, by replacing the observed counts with the expectations of the input models. The expected counts, when used to replace the observed ones, are commonly know as the *Asimov dataset*; the procedure is better described in 2.1. A peculiar alternative definition of sensitivity is strictly linked to maximizing the probability of getting a predefined value of significance, assuming the signal is present, as described by Punzi [1].

2. Framing the problem

The classical framework of hypothesis testing will be employed in this Paper. The null hypothesis, denoted H_0, is taken to be the model that represent the best of our knowledge; the alternative hypothesis, denoted H_1, usually represents a model that introduces some new physics process on top of the predictions from H_0, but in general just represents the model that describes the process whose chance of observing we want to maximize.

The two hypotheses are parameterized, often without loss of generality, as nested models depending on some parameter (or vector thereof) θ; the notation H_0 and H_1 will be then maintaned to implicitly mean $H_0 = H(\theta = 0)$ and $H_1 = H(\theta = \theta_i)$, with $\theta_i \neq 0$.

In the common case of Poisson counts, we express the models in terms of the expected signal counts S and of the expected background counts B, hence $H_0 = Pois(B)$ and $H_1 = Pois(S+B)$. The parameter that makes the hypothesis nesting explicit is then the expected signal count (zero for H_0, non-zero for H_1).

A criterion to reject the null hypothesis is then to reject H_0 if the observed count lies in a *critical region* defined by a (desirably small) probability α of rejecting the null hypothesis conditioned to the null hypothesis being true. The formal definition for this probability is $\alpha = P(reject\ H_0|H_0)$. The *power* of the test is defined as the probability β of correcly rejecting H_0 conditioned to the alternative hypothesis being true. The procedure is illustrated in figure 1.

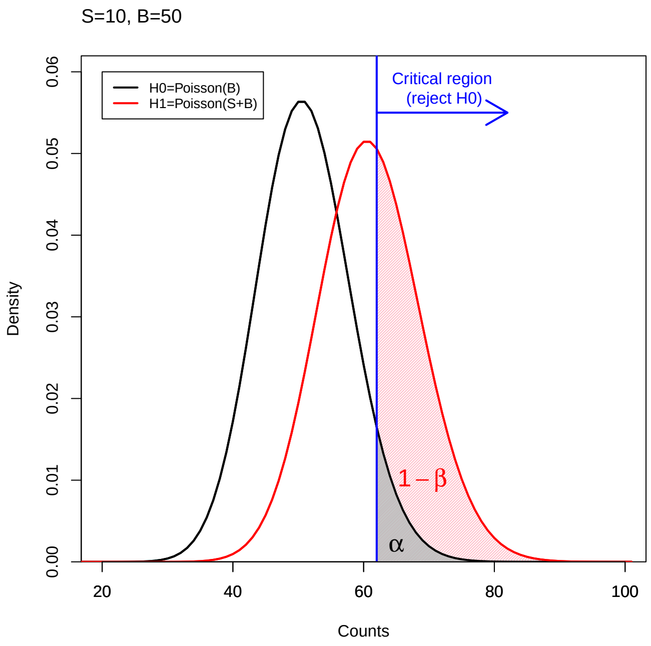

Figure 1: An example of hypothesis testing.

2.1 The Asimov dataset

When using the formulas described in Section 3 as figures of merit for the optimization of an analysis, it is important to not use the observed data in their evaluation. This is achieved by substituting the number of observed events—in all the formulas that depend on it—with a data set based only on the available simulated models. In order for such a data set to be representative (and thus, in this case, useful for estimating the expected significance to be maximized), it is commonly defined as the data set such that using it to evaluate estimators for all the parameters of the model, one would obtain the true parameter values. In practice, this is shown to correspond to the data set corresponding to the expectations computed from a very large simulated sample; such a dataset is known in literature as the *Asimov dataset* [2]. It is important to note that while the statistical errors due to the limited statistics of simulated samples are usually accounted for (e.g. as nuisance parameters in maximum likelihood fits), they are suppressed for the computation of the Asimov data set (the underlying idea is that in the limit of a very large sample these fluctuations would be negligible).

3. Review of figures of merit in literature

The typical HEP use case illustrated in Fig. 1 involves the comparison between a known, oftentimes large, background count and a slightly larger count that accounts for both the background and a (usually small) signal. Any uncertainty in the background count would reduce the chances of observing a signal, and thus an elementary definition of significance involves comparing the signal counts S with the statistical uncertainty in the background count B; in case of Poisson counts, the expression is then $Z_{sb} := \frac{S}{\sqrt{B}}$. This definition corresponds to the bare minimum needed for observing a signal, and for this reason is often used in order to optimize an analysis for setting limits on the signal production cross section (a quantity proportional to the counts). It is worth to note that this expression can be derived directly from Poisson calculations in the limit of large average background, where the Poisson distribution for the background can be approximated with a Gaussian distribution. This expression for the sensitivity breaks down (diverges) for low background counts. Figure 2 (left) details its behaviour for a variety of combinations of signal and background counts.

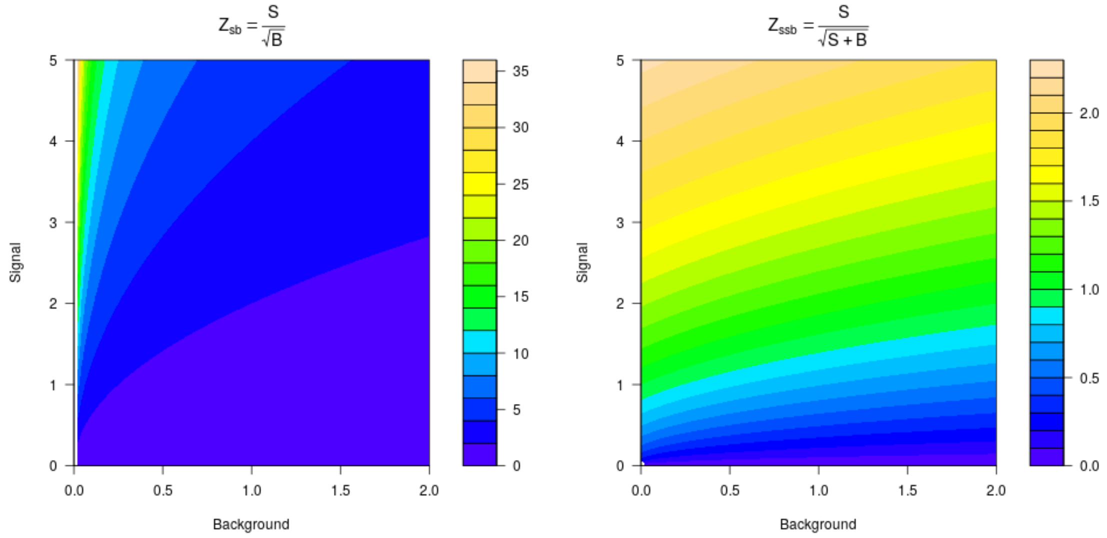

Figure 2: The pseudosignificances Z_{sb} (left) and Z_{ssb} (right), computed in a grid of values for the expected number of signal (S) and background (B) events. The grid scan for Z_{sb} is zoomed into the low-background region, where the expression diverges.

In case the existence of a signal has already been ascertained in past experiments, the optimization is performed to yield the best cross section measurement possible, and a reasonable figure of merit compares the signal with the overall statistical uncertainty in the joint S+B Poisson count, yielding a sensitivity $Z_{ssb} := \frac{S}{\sqrt{S+B}}$. A nice perk of this definition is that the denominator does not break down for small background counts; however, if the background counts are affected by a systematic uncertainty, this simple formula can significantly overestimate the proper significance, as it will be shown later. Figure 2 (right) details the behaviour of this figure of merit for a variety of combinations of signal and background counts.

In order to avoid the problems at low background counts of Z_{sb}, Byutikov and Krashnikov [3] introduced a formula based on the difference between the statistical uncertainties in the $S+B$ and in the B only counts, $S_{12} := 2\left[\sqrt{S+B} - \sqrt{B}\right]$. The same formula has been studied by Bartsch and Quast [4] under the name Q. This formula breaks down neither for small S nor small B counts, as illustrated in Fig. 3 (left). In the limit of both high signal and background counts, some colleagues argued in private correspondence that $\lim_{S\to\infty, B\to\infty} Z_{12} = \frac{S}{\sqrt{B}}$, but Fig. 3 (right) suggests otherwise; in fact, the two formulas converge to each other in the limit of high background and low signal counts. An interesting extension is the case in which the systematic uncertainty is not zero, but this approach is left for future studies.

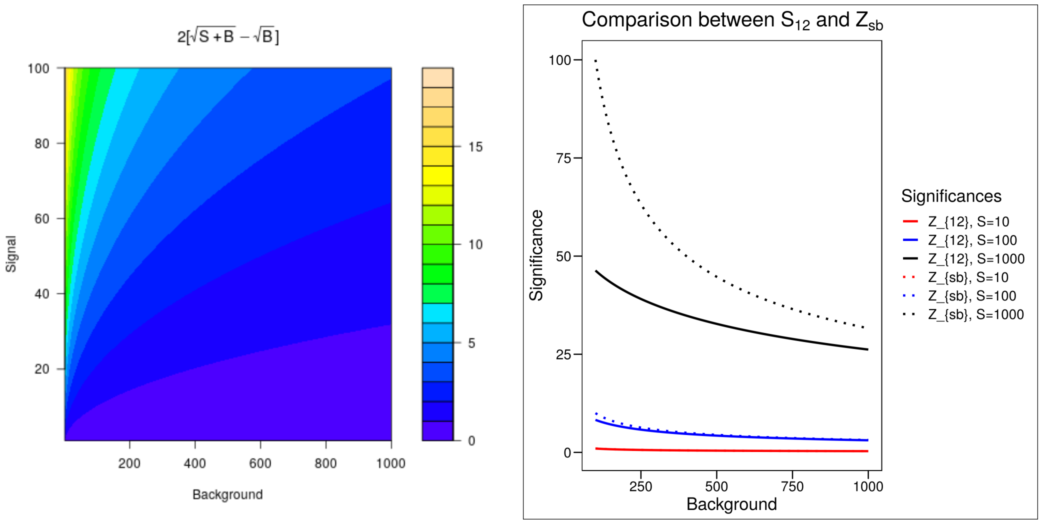

Figure 3: The pseudosignificance Z_{12} (left) computed in a grid of values for the expected number of signal (S) and background (B) events; a comparison (right) with the simple Z_{sb} expression outlines the evident shortcomings of the latter.

A way of naïvely accounting for systematic uncertainties in the background counts consists in comparing the signal count with the overall uncertainty in the background count, approximated as a quadratic sum of the background statistical and systematic uncertainties, $Z_{berr} := \frac{S}{\sqrt{B+\Delta B^2}}$, where ΔB is the systematic uncertainty in the background count. The behaviour of the formula for various signal and background counts is shown in Fig. 4 (left). This formula represents an immediate extension of Z_{sb}, to which in fact it converges in the limit $\Delta B \to 0$, as illustrated in Fig. 4 (right). Since it accounts for the background overall uncertainty, Z_{berr} is often used to optimize for a discovery. It is interesting to note that Z_{berr} diverges significantly from Z_{sb} already for a systematic error on B of 1–5%, in line with the previous consideration on Z_{sb}.

Probably the most used pseudosignificances in HEP, apart from computation power considerations, is the log-likelihood ratio. Wilks theorem is used to derive confidence intervals at a predetermined confidence level. In the archetypical HEP problem, Wilks theorem is actually not satisfied, because H_0 lies at the boundary of the allowed values for S (unless H_1 does consist in a modified model that predicts deficit of events with respect to H_0, which can happen in HEP [5]); however, it has been shown [2] that the asymptotic properties for the Type I errors are unaffected. Consequently, it is still possible to obtain a meaningful expression for the significance by plugging in the appropriate likelihood function. Li and Ma [6] have found a solution valid for Poisson counts

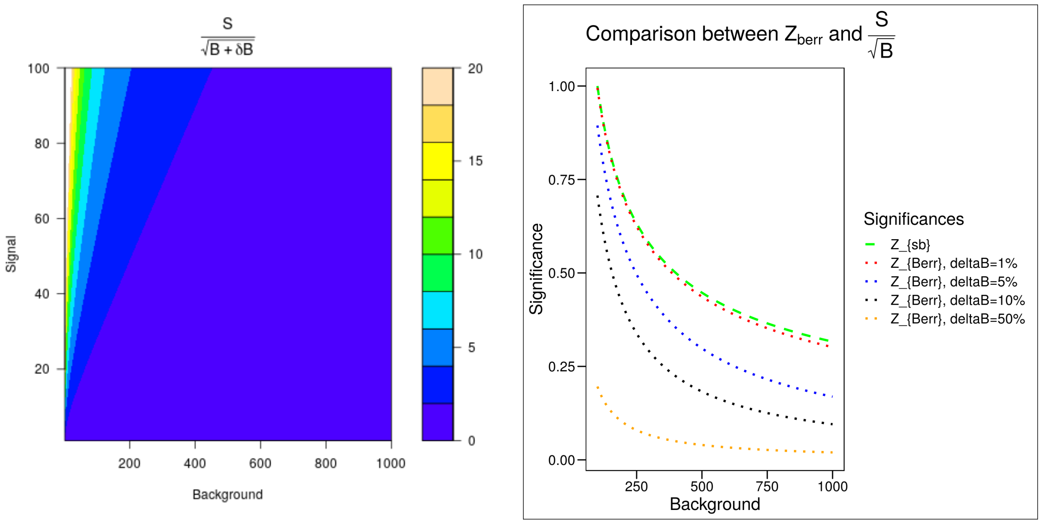

Figure 4: The pseudosignificance Z_{berr} (left) computed in a grid of values for the expected number of signal (S) and background (B) events; a comparison (right) with the simple Z_{sb} expression outlines the extent to which ignoring the systematic error on the background yields is a good approximation.

in the so-called *on/off* problem. In HEP jargon, the problem is framed in terms of Poisson counts in a given region (*on* region) where the background counts are measured in a sideband (*off* region). The Li and Ma formula is expressed in terms of the counts in the *on* and in the *off* regions as:

$$Z_{PL} := \sqrt{2}\sqrt{n_{\text{on}}\ln\frac{n_{\text{on}}(1+\tau)}{n_{\text{tot}}} + n_{\text{off}}\ln\frac{n_{\text{off}}(1+\tau)}{n_{\text{tot}}\tau}} \tag{3.1}$$

where the transfer factor τ is defined as the ratio between the background counts in the *off* and *on* regions, $\tau := \frac{n_{b,\text{off}}}{n_{b,\text{on}}}$. In this paper, the transfer factor is taken to be unity, and the problem is reframed in terms of background yields in the signal region. The dependence on τ will be studied in an ongoing study of broader scope. Figure 5 (left) illustrates the behaviour of Z_{PL} for various sets of signal and background counts.

When the systematic uncertainty in the background counts can be considered negligible, then an expression for the significance can be obtained from an approximation of the Cowan-Cranmer-Gross-Vitells asymptotic formula [2] for known B:

$$\sqrt{q_{0,A}} := \sqrt{2((S+B)\ln(1+\frac{S}{B}) - S} \tag{3.2}$$

The formula is illustrated in Fig. 5 (right) for various values of S and B counts. This expression can be easily expanded in powers of $\ln(\frac{S}{B})$, yielding $\sqrt{q_{0,A}} = \frac{S}{\sqrt{B}}(1+\mathscr{O}(\frac{S}{B}))$. This highlights the fact that the simple formula Z_{sb} is a good approximation of the exact significance only in the case of $S << B$; unfortunately, in literature Z_{sb} has been often thought to be useful for the general case of large $S+B$, hence leading to possibly catastrophic failures when $S \sim B$. Figure 6 (left) illustrates the extent to which the two formulas diverge. Figure 6 (right) shows a comparison of $\sqrt{q_{0,A}}$ with the Li–Ma expression Z_{PL}, showing that convergence is achieved for higher background counts,

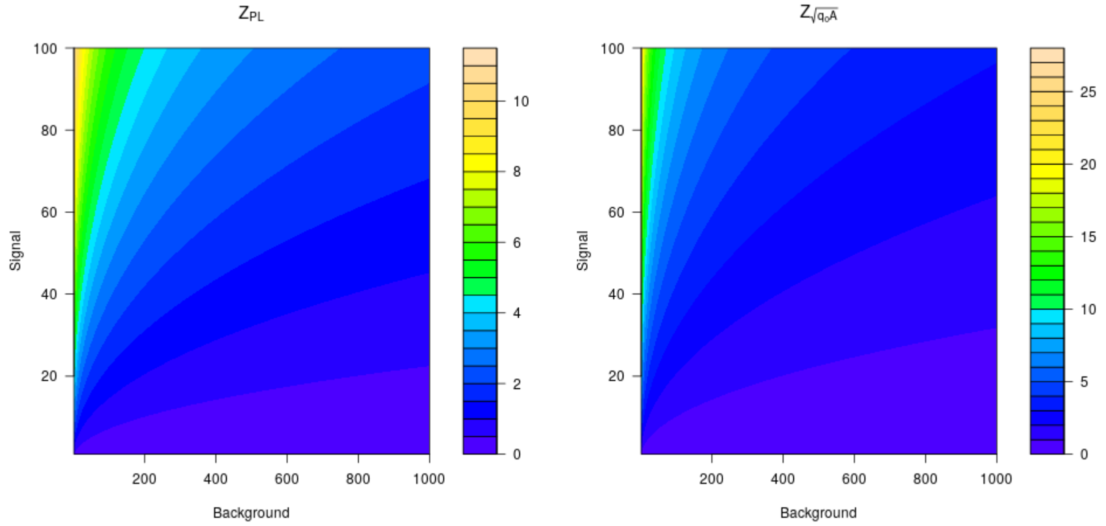

Figure 5: The pseudosignificance Z_{PL} (left) computed in a grid of values for the expected number of signal (S) and background (B) events; a comparison (right) with the asymptotic expression in case of background known with negligible uncertainty, $\sqrt{q_{0,A}}$, outlines the extent to which the two expressions approximate each other.

as expected because of the asymptotic nature of $\sqrt{q_{0,A}}$. It can also be shown that the asymptotic expression does not converge well to Z_{PL} for values of the significance larger than $\sim 10\sigma$; this is probably because the approximations made in both formulas entail a suboptimal modelling of the tail probabilities. In a broader-scope study in preparation, I will explore the inclusion of higher-order approximations [7, 8, 9] to the asymptotic expression for the likelihood ratio.

A fully frequentist solution to the on/off problem can be derived by rephrasing it in terms of the conditional binomial probability for the on/off events to be divided as observed [10], resulting in the formula:

$$Z_{Bi} := \sqrt{2}\mathrm{erf}^{-1}(1-2\frac{B(\frac{\mu_{\mathrm{on}}}{\mu_{\mathrm{tot}}}, n_{\mathrm{on}}, 1+n_{\mathrm{off}})}{B(n_{\mathrm{on}}, 1+n_{\mathrm{off}})}) \tag{3.3}$$

The expression is here computed for $\mu_{\mathrm{on}} = S+B$, $\mu_{\mathrm{off}} = B$, corresponding to assuming a unity transfer factor $\tau = \frac{\mu_{\mathrm{off}}}{\mu_b} = 1$, without loss of generality; the transfer function somehow encodes the increase of the uncertainty in the background counts due to estimating them in a sideband region. An ongoing study with broader scope will examine the dependence of such expressions on the transfer factor. Since we are interested in the expected significances for optimization purposes, as before the observed counts N_{obs} are substituted by the Asimov data set in both *on* and *off* regions for obtaining the expected significance. Fig. 7 illustrates this important expression for the significance. The formula is reportedly good for optimizing cuts, with the caveat that a handful signal events should survive the cuts for the formula to hold. A practical threshold is indicated in Ref.[10] to be around 5 events.

The introduction of such a formula has spanned different communities, having been introduced

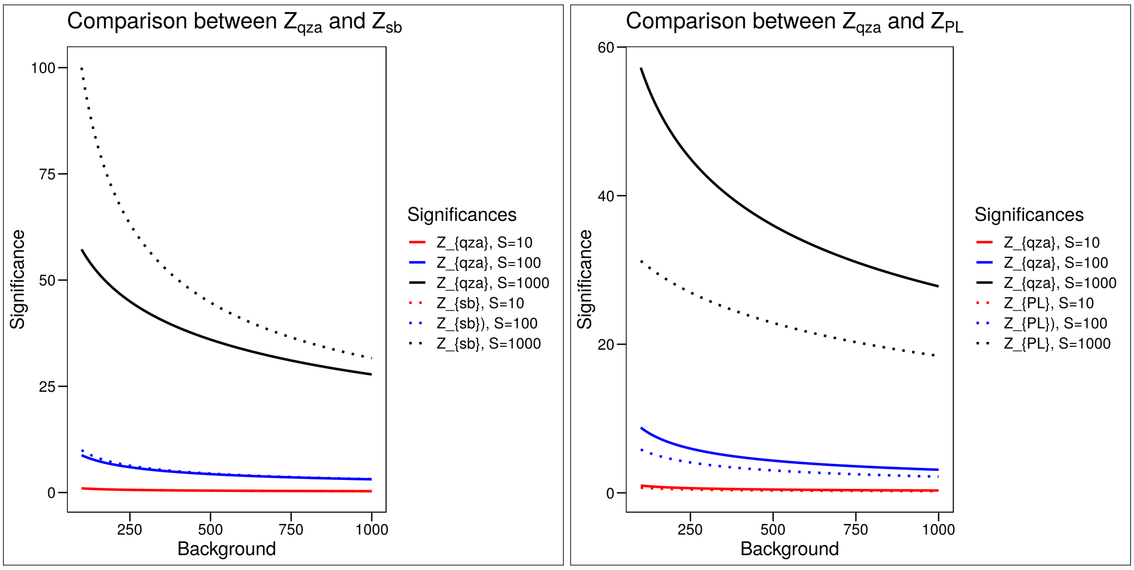

Figure 6: The pseudosignificance $\sqrt{q_{0,A}}$ (left) computed in a grid of values for the expected number of signal (S) and background (B) events; a comparison (right) with the simple Z_{sb} expression outlines the shortcomings of the latter.

in statistics by Przyborowski and Wilenski [11], in HEP by James and Roos [12], and in Gamma Ray Astronomy (GRA) by Gehrels [13]. An in-depth review of Z_{bi} can be examined in Ref. [10], where a hybrid recipe, involving Bayesian-style averaging and frequentist tail-integral calculations, has also been derived under the name Z_{Γ}. The latter formula has been analytically demonstrated to be equivalent to Z_{Bi}; because of that, the latter is used in this paper as a placeholder for both computations.

3.1 The Punzi significance formulas

A frequentist criterion for the definition of sensitivity of an experiment has been given by Punzi [1]. The Poisson counts for the two hypotheses is written down explicitly, $H_0 = Pois(B)$ and $H_1 = Pois(S+B)$. Type I and II error rates are then parameterized as $\alpha = P(reject\ H_0|H_0)$ and $1-\beta = P(reject\ H_0|H_1)$, and a confidence level CL for the limits in case of no discovery is chosen as a reference. The Z-scores needed to obtain a one-sided Gaussian test at significances α and β are then denoted as a and b, and an expression for the minimum significance to reach the desired probability is derived:

$$S_{min} := \frac{b^2}{2} + a\sqrt{B} + \frac{b}{2}\sqrt{b^2 + 4a\sqrt{B} + 4B} \tag{3.4}$$

The expression is illustrated in Fig. 8 (left) for various background counts and significance levels. It is important to note that, contrary to all the pseudosignificances previously examined, this definition does not depend on the expected number of signal event. As a consequence, this expression can be—and often is—used in optimization problems when the signal model is not defined a priori. The arbitrariness in the choice of the parameters a and b (or equivalently of α

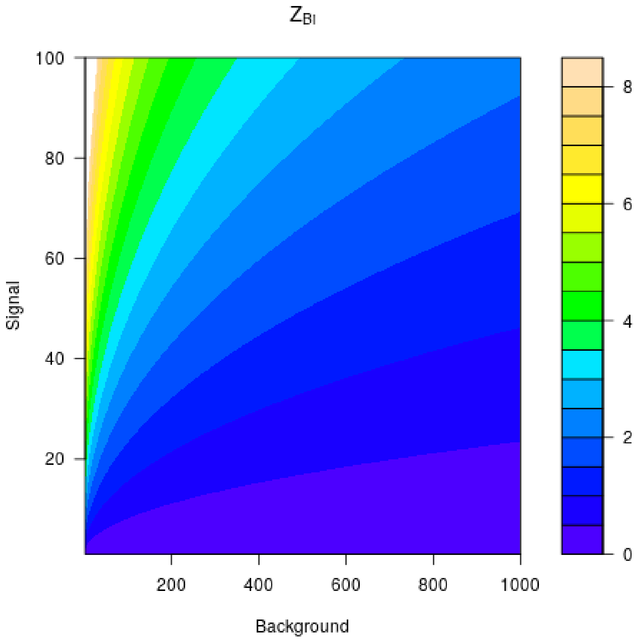

Figure 7: The pseudosignificance Z_{Bi}, computed in a grid of values for the expected number of signal (S) and background (B) events.

and β) is sometimes criticized by orthodox exponents of the frequentist school, and will not be discussed here.

A further refinement to this formula stems from empirically accounting for the differences between the Poisson and Gaussian integral tails, yielding the improved formula illustrated in Fig. 8 (right):

$$S_{min}^{improved} := \frac{a^2}{8} + \frac{9b^2}{13} + (S_{min} - \frac{b^2}{2}) \tag{3.5}$$

For increasing background counts, the regular and the improved Punzi formulas maintain a certain difference between them, as illustrated in Fig. 9 and explained by the fact that the Poisson and Gaussian tail integrals are non-negligibly different even in the high-counts regime.

A fair summary for this pseudosignificance expression is that it stems from an attempt at finding a figure of merit suitable for analysis optimization in the absence of a signal model (or in cases in which there are multiple signal models that one does not want to prioritize one over the other); it is rooted in a frequentist approach and expressed as a minimum number of signal events needed to reach a given significance, i.e. it is naturally cast in an optimization framework as a problem of maximization of selection efficiency. Because of this, a comparison with other expressions can be done only when such expressions can be cast in the same framework as a ratio with a numerator that can depend on the signal counts (that is then subject to the maximization) and a denominator that cannot depend on the signal counts. Punzi in his paper shows an example comparison with Z_{sb} and Z_{ssb}; further comparisons will be pursued in a study with broader scope.

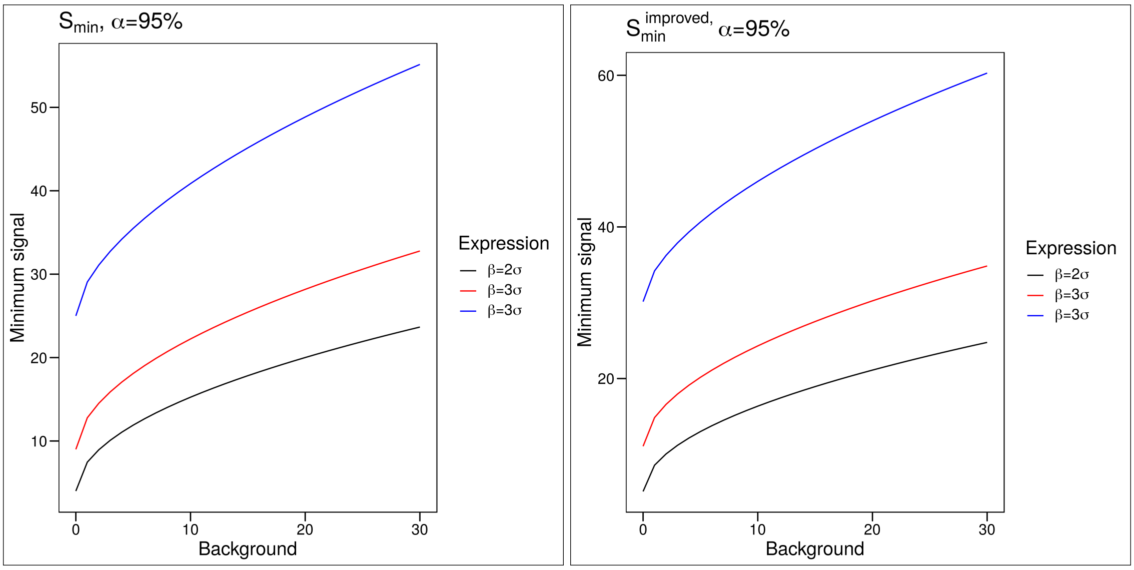

Figure 8: The minimum number of events needed to reach the desired power and significance, according to the basic Punzi formula S_{min} (left) and to the improved expression $S_{min}^{improved}$ (right).

3.2 Bayesian evidence values

The Li–Ma problem can be recast in a Bayesian framework, as shown by an application in GRA by Ref. [14]. A *Bayesian Z-score* is proposed as a function of the odds B_{01} of the background model over the signal model as:

$$S_{bayes} := \sqrt{2}\mathrm{erf}^{-1}(1 - B_{01}), \tag{3.6}$$

where B_{01} is defined as a function of N_{on}, N_{off}, and α defined as the ratio of exposures for the *on* and *off* region. The odds of an hypothesis against another are commonly called *Bayes factor*, and represent the standard Bayesian way of expressing favour towards an hypothesis in comparison to another one, rather than the frequentist procedure of relying on the error rates. The full expression for B_{01} can be found in Ref. [14], and relies on the hypergeometric function ${}_2F_1(a,b;c;z)$; issues in the numerical convergence of the implementation—provided in Ref [14]—of such function for large values of its arguments forced us to restrict the phase space study to relatively low signal and background yields. Such an issue has not been noted by the original paper, likely because the field of GRA is usually characterized by very low (compared with the typical HEP case) signal and background yields. Figure 10 (left) shows the result of the phase space scan, whereas Fig. 10 (right) outlines the properties of Z_{bayes} with respect with the frequentist $\sqrt{q_{0,A}}$. It is very clear that a real comparison between this pseudosignificance and classic expressions can—if at all—be done only in the low counts regime, but the behaviour at higher counts is radically different and a source of concern. I think more studies are needed before considering Eq. 3.6 for an analysis of real physics cases.

Another aspect to be considered is that the Z_{bayes} formula originates from a simple analogy with the concept of Z-scores, as outlined by Eq. 3.6. A more rigorous approach, such as the one

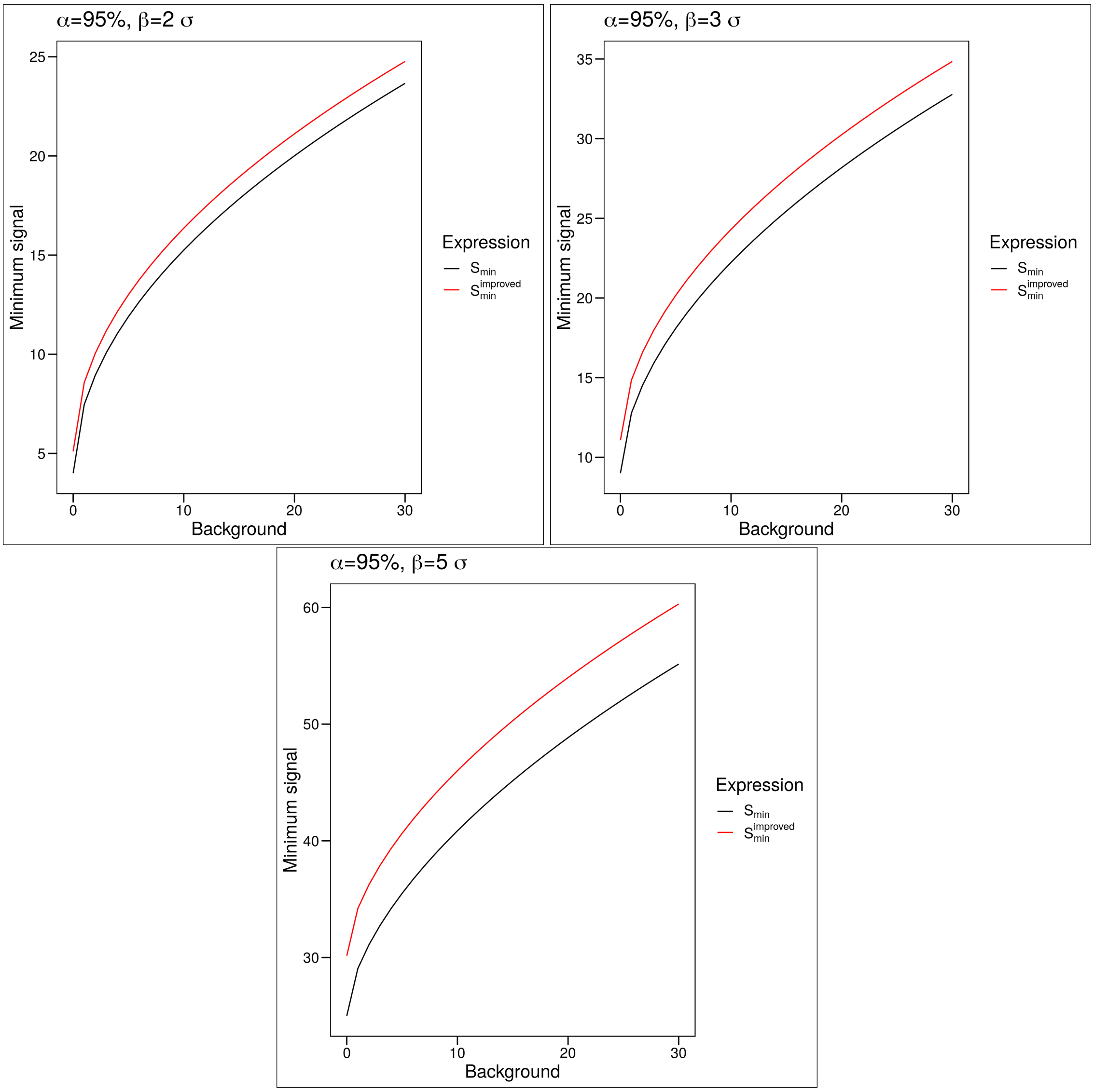

Figure 9: Comparison between the minimum number of events needed to reach the desired significance in the basic (black lines) and improved (red lines) Punzi formulas, for a fixed significance level and three different powers.

proposed by Ref. [15], will be pursued in future studies.

4. Summary and future work

Various definitions of sensitivity for a counting experiment in the typical High Energy Physics case have been studied. The fully-frequentist definition of expected significance has been taken as a reference, and compared to various simplified formulas, obtained by means of different approximation assumptions; the range of validity of such approximations have been investigated and highlighted. A Bayesian figure of merit, proposed for Gamma Ray Astronomy, is found problem-

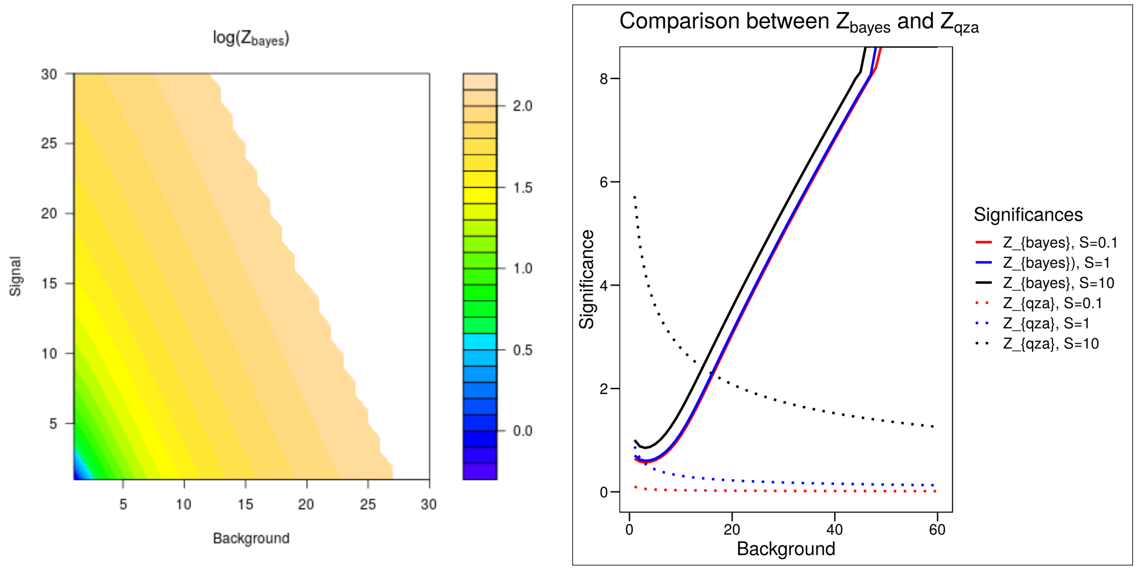

Figure 10: The pseudosignificance Z_{bayes}, computed in a grid of values for the expected number of signal (S) and background (B) events; a comparison (right) with the frequentist expression $\sqrt{q_{0,A}}$.

atic in that it has an undesirable behaviour for any set of yields typical of HEP problems. Another Bayesian definition is found in literature to give exactly the same results as the fully-frequentist solution in the asymptotic regime.

For the time being, we recommend avoiding oversimplified expressions of significance; analyzers should preferentially stick to the profile likelihood ratio method, whose properties are well-established and that is guaranteed to be optimal thanks to the Neyman-Pearson lemma. One Bayesian solution has been shown to be equivalent and interchangeable with the full frequentist solution, but more studies would be needed to explore alternative solutions.

A few questions are left for a later study of larger scope, that is in preparation by the author. Such questions are: the interpretation of the tunable parameters in the Punzi formulas and a fair comparison with other figures of merit; the dependence on several pseudosignificances on the transfer factor τ in the classical on/off problem; the inclusion, in several of the examined formulas, of a parametric dependence on systematic uncertainty in the background counts; the exploration of higher order corrections to the likelihood in case of low counts; the implementation of a pure Bayesian evidence formula, and its extension to include systematic errors on the background yields.

Acknowledgments

This paper has been inspired by several interesting and fruitful discussions I have participated to within the Statistics Committee of the CMS Collaboration; this is a partial attempt at answering in a broad and systematic way to several misconceptions that arise in the HEP community; I hope to provide a deeper picture in a broader publication (in preparation).

I would like to thank the IRMP of Université catholique de Louvain for providing a rich, free, and stimulating environment that really stimulates research and goes a long way towards the ideal research environment outlined by Lindley in the introduction to his book *Understanding Uncertainty*.

Finally, I would like to especially thank my beautiful wife for her constant support and encouragement, and for marrying me.

References

[1] G. Punzi, *Sensitivity of searches for new signals and its optimization*, *eConf* **C030908** (2003) MODT002 [physics/0308063].

[2] G. Cowan, K. Cranmer, E. Gross and O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, *Eur. Phys. J.* **C71** (2011) 1554 [1007.1727].

[3] S. I. Bityukov and N. V. Krasnikov, *New physics discovery potential in future experiments*, *Mod. Phys. Lett.* **A13** (1998) 3235 [physics/9811025].

[4] V. Bartsch and G. Quast, *Expected Signal Observability at Future Experiments*, .

[5] CMS collaboration, CMS Collaboration, *Updated search for a light charged higgs boson in top quark decays in pp collisions at $\sqrt{s} = 7$ TeV*, CMS Physics Analysis Summary CMS-PAS-HIG-12-052, 2012.

[6] T.-P. Li and Y.-Q. Ma, *Analysis methods for results in gamma-ray astronomy*, *The Astrophysical Journal* **272** (1983) 317.

[7] *Properties of sufficiency and statistical tests*, *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* **160** (1937) 268 [http://rspa.royalsocietypublishing.org/content/160/901/268.full.pdf].

[8] I. M. Skovgaard, *Likelihood asymptotics*, *Scand. J. Statist.* **28** (2001) 3.

[9] A. R. Brazzale and A. C. Davison, *Accurate parametric inference for small samples*, *Statist. Sci.* **23** (2008) 465.

[10] R. D. Cousins, J. T. Linnemann and J. Tucker, *Evaluation of three methods for calculating statistical significance when incorporating a systematic uncertainty into a test of the background-only hypothesis for a Poisson process*, *Nuclear Instruments and Methods in Physics Research A* **595** (2008) 480 [physics/0702156].

[11] J. Przyborowski and H. Wilenski, *Homogeneity of results in testing samples from poisson series with an application to testing clover seed for dodder*, *Biometrika* **31** (240) 31 [/oup/backfile/content_public/journal/biomet/31/3-4/10.1093/biomet/31.3-4.313/2/31-

[12] F. James and M. Roos, *Errors on ratios of small numbers of events*, *Nuclear Physics B* **172** (1980) 475.

[13] N. Gehrels, *Confidence limits for small numbers of events in astrophysical data*, *Astrophysical Journal* **303** (1986) 336.

[14] M. L. Ahnen, *On the On-Off Problem: An Objective Bayesan Analysis*, *PoS* **ICRC2015** (2016) 701 [1508.05855].

[15] S. Gillessen and H. L. Harney, *Significance in gamma-ray astronomy - The Li & Ma problem in Bayesian statistics*, *Astron. Astrophys.* **430** (2005) 355 [astro-ph/0411660].