PROCEEDINGS OF SCIENCE

Statistics Challenges in High Energy Physics Search Experiments

Eilam Gross*
The Weizmann Institute of Science, Rehovot, Israel
E-mail: eilam.gross@weizmann.ac.il

Ofer Vitells
The Weizmann Institute of Science, Rehovot, Israel
E-mail: Ofer.Vitells@weizmann.ac.il

Statistical methods used for discovery and exclusion of signal at the LHC and the TEVATRON are described. An emphasis is given to the Look Elsewhere Effect, the CL_s controversy and the equivalence between the Bayesian and frequentist Profile Likelihood when using flat priors. For the Look Elsewhere Effect, formulas are derived that allow the estimation of the effect from the simple fixed mass result without the need to perform complicated Monte Carlo simuations.

13th International Workshop on Advanced Computing and Analysis Techniques in Physics Research, ACAT2010
February 22-27, 2010
Jaipur, India

*Speaker.

1. Introduction and the Statistical Challenge

The Large hadron Collider (LHC) is colliding a huge amount of Protons with each other. The Standard Model describes the interactions in nature and is able to predict the outcome of these collisions. There is only one ingredient of the Standard Model which was not discovered yet and its existence is crucial for the completeness of the model. This is the particle that acquire mass to all fundamental particles in nature, it is called the Higgs Boson. From the statistical point of view there are two hypotheses being tested here. One is the Standard Model without the Higgs Boson (denoted by H_0 and referred to as "background-only"), and the other one is the Standard Model including the Higgs Boson (H_1 referred to as the "signal" or "signal+background" hypothesis). The difficulty in testing the hypotheses is that most of the collisions give rise to a data which is compatible with the H_0 hypothesis. It is only, once per a Billion collisions or so, that the yet undiscovered Higgs Boson is expected (H_1).

The first step in the hypothesis testing is to state the relevant null hypothesis and then try to reject it. Rejecting the H_0 hypothesis in favor of the H_1 hypothesis is considered a discovery. On the other hand, rejecting the the H_1 hypothesis in favor of the H_0 hypothesis is interpreted as excluding the Higgs Boson.

This writeup does not aim to cover the basic definitions and various techniques of statistical hypotheses inference. Those were covered elsewhere [1]. Rather, we prefer to emphasize some statistical issues which are relevant mainly to High-Energy physics. In section 2 we elaborate on the look elsewhere effect which is one of the main issues in high-energy discovery physics. In section 3 we explain the difficulties in statistical high-energy exclusion. Finally in section 4 we show the equivalence between Bayesian and frequentist Profile-Likelihood exclusion.

2. Frequentist Discovery and the Look Elsewhere Effect

The statistical significance that is associated to the observation of new phenomena is usually expressed using a *p-value*, that is, the probability that a similar effect or larger would be seen when the signal does not exist (a situation usually referred to as the null or background-only H_0 hypothesis). A p-value of $2.87 \cdot 10^{-7}$ is traditionally associated with discovery (this is equivalent to a 5 σ one sided effect). It is often the case that one does not *a-priory* know where the signal will appear within some possible range. In that case, the significance calculation must take into account the fact that an excess of events anywhere in the range could equally be considered as a signal. This is known as the “look elsewhere effect” [2]. A straightforward way of quantifying this effect is by simply running many Monte-Carlo simulations of background only experiments, and finding for each one the fluctuation with the largest significance that resembles a signal. While this procedure is simple and gives the correct answer, it is also time and CPU consuming, as one would have to repeat it $\mathscr{O}(10^7)$ times to get the p-value corresponding to a 5σ significance. In [3] the effect was studied to its full scope. Here we briefly review the analysis and its results.

Consider a gaussian signal with a fixed width on top of a background that follows a Raleigh distribution in the range [0,100]. An example pseudo-experiment is shown in Fig. (1).

We assume that the background shape is known but it’s normalization is not, so that it is a free parameter in the fit (i.e. a nuisance parameter), together with the signal location and normalization.

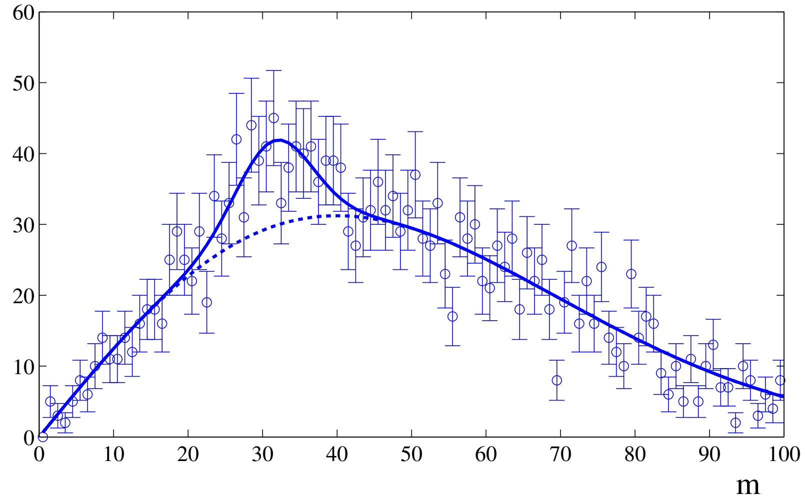

Figure 1: An example pseudo-experiment showing the fit components.

We use a binned profile likelihood ratio as our test statistic, where the number of events in each bin are assumed to be Poisson distributed with an expected value

$$E(n_i) = \mu s_i(m) + \beta b_i \tag{2.1}$$

where μ is the signal strength parameter, $s_i(m)$ corresponds to a gaussian located at a mass m, β is the background normalization and b_i are fixed and given by the Raleigh distribution. For simplicity of notation we will use in the following $\mathbf{s} = \{s_i\}$ and $\mathbf{b} = \{\beta b_i\}$. The hypothesis that no signal exists, or equivalently that $\mu = 0$, will be referred to as the null hypothesis, H_0. $\hat{\mu}$ and $\hat{\mathbf{b}}$ will denote maximum likelihood estimators while $\hat{\hat{\mathbf{b}}}$ will denote the conditional maximum likelihood estimator of the background normalization under the null hypothesis.

In a fixed mass scenario one is only interested in looking for a signal at some specific, pre-defined mass m_0. The test statistic in this case is defined using the likelihood ratio evaluated at the pre-defined mass,

$$t_{fix} = -2ln\frac{\mathscr{L}(\hat{\hat{\mathbf{b}}})}{\mathscr{L}(\hat{\mu}\mathbf{s}(m_0)+\hat{\mathbf{b}})}. \tag{2.2}$$

where $\mathscr{L}$ is the likelihood function. The distribution of the test statistic t_{fix} under the null hypothesis, $f(t_{fix}|H_0)$, is expected to follow a chi-square distribution with one degree of freedom at the large sample limit, due to the well known theorem by Wilks [4] . If the observed test-statistic is $t_{fix,obs}$, the significance of the observation can be expressed via the observed p-value

$$p_{fix} = \int_{t_{fix,obs}} f(t_{fix}|H_0)dt_{fix} \tag{2.3}$$

This p-value is related to the probability to observe a result as or less compatible with the background-only hypothesis. In other words it is the probability that the background will fluctuate *at this mass point*, as or even more than the observed fluctuation. The distribution of t_{fix} under H_0 is shown in Figure 2 (blue full line).

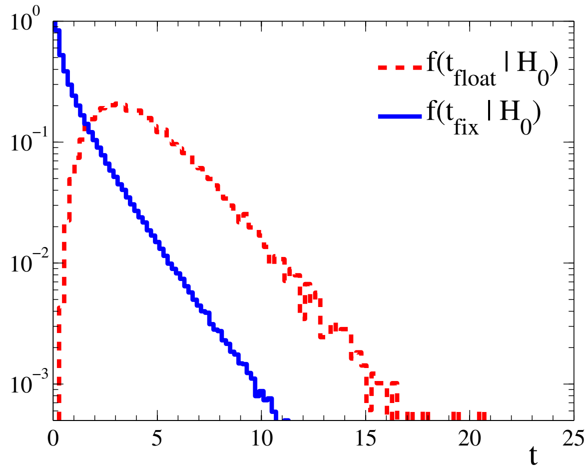

Figure 2: The distributions of the test statistics t_{fix} (blue full line) and t_{float} (red dash line) under the null hypothesis. The distribution of t_{fix} closely follows a χ^2 with one degree of freedom.

If one does not *a-priory* know the location of the signal, any signal-like fluctuation of the background, anywhere in the mass range, could be considered as a discovery. The probability for the background to fluctuate anywhere in the mass range is obviously bigger than its probability to fluctuate at a specific mass point. The ratio between the two probabilities is called the trial factor, i.e.

$$trial\# = \frac{p_{anywhere}}{p_{fix}} \tag{2.4}$$

In order to give precise meaning to $p_{anywhere}$ we must specify a search procedure, or equivalently a test statistic that will be used to measure the compatibility of the data to a signal hypothesis, when the signal location is not known. The most natural procedure would be to scan the entire range, in steps that are sufficiently smaller then the mass resolution, and select the point for which the signal likelihood is the largest, namely that maximizes (2.2). This is tantamount to including the mass as a free parameter over which the likelihood is maximized in a "floating mass" fit. The test statistic would be therefore

$$t_{float} = -2ln \frac{\mathscr{L}(\hat{\hat{b}})}{\mathscr{L}(\hat{\mu}s(\hat{m}) + \hat{b})} \tag{2.5}$$

where $\hat{m}$ is the mass point that globally maximizes the likelihood, i.e. the maximum likelihood estimator of m. The distribution of t_{float} under H_0 is also shown in Figure 2 (dashed red line).

When generating background-only experiments we usually find, as would be expected, that there are several local maxima of the likelihood ratio as a function of the mass m. such an example is shown in Fig. (3).

The average number of local maxima is naturally proportional to the ratio of the mass range to the mass resolution, as shown in Fig.(4).

$$\langle N \rangle \sim \frac{\text{mass range}}{\text{mass resolution}} \tag{2.6}$$

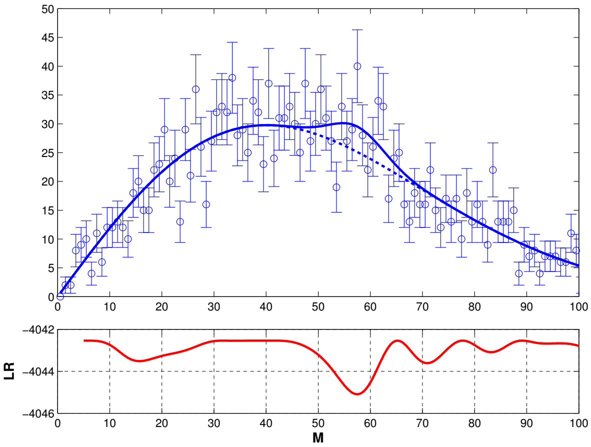

Figure 3: A background-only experiment with a few local maxima (here shown as local minima of the inverse likelihood ratio). The maximum Likelihood occurs around $m = 58$ units.

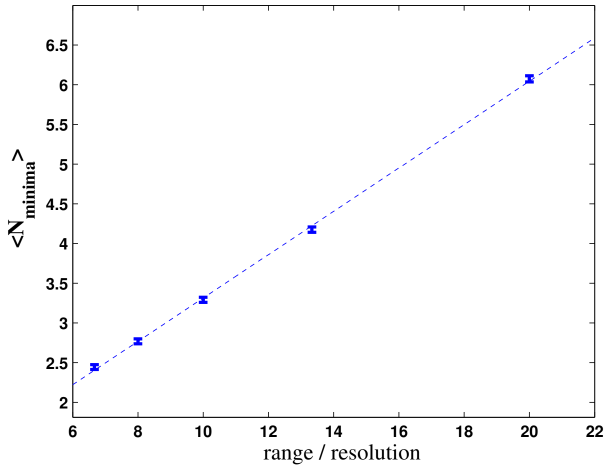

Figure 4: The average number of local maxima as a function of the thumb rule number: mass-range/mass-resolution.

If we divide the mass range to several regions such that each contains a single local maximum, we might expect Wilk's theorem to hold for each region separately. This is because in each region the likelihood function has, by construction, a single local maximum. In that case, the values of the test statistics at the local maxima would distribute as a χ^2 with two degrees of freedom, since now both μ and m can be regarded as parameters of interest. This point, while not rigorously proved, was first demonstrated in [5], where it was shown that the distribution of t_{float} can be reproduced to a very good approximation by taking the maximal of several χ^2 variates. In our case we find similar results. We use this observation as a starting point from which we estimate the trial factor. Denote the values of the test-statistic at the local maxima by $t_{float}^{(i)}$, $i = 1...N$, such that

$$t_{float} = \max_i [t_{float}^{(i)}] \tag{2.7}$$

A straightforward analysis ([3] shows that for small enough values of $p_{\chi^2_2}$ (the χ^2_2 p-value), one can approximate

$$P(t_{float} > t) \simeq p_{\chi^2_2}\langle N\rangle \tag{2.8}$$

Thus the p-value of the floating fit test statistic is approximately equal to the p-value of a χ^2 with two d.o.f, times the average number of local minima. Note that this approximation is valid when

$$p_{\chi^2_2}\langle N\rangle \ll 1 \tag{2.9}$$

that is, as $\langle N\rangle$ becomes large the approximation requires $p_{\chi^2_2}$ to become correspondingly small.

Using the above result, we can easily estimate the trial factor (2.4). We distinguish however between two scenarios for which the definition of the trial factor may be slightly different:
case(a). In this case we have an observed data set with some observed value of t_{float} with corresponding maximum likelihood estimators $\hat{\mu}$ and $\hat{m}$. We wish to estimate the significance of this measurement. The "true" p-value is:

$$p_{float} = \int_{t_{float,obs}} f(t_{float}|H_0)dt_{float} \tag{2.10}$$

while the "local p-value" can be defined as the probability that a background fluctuation *at the observed mass* $\hat{m}$ will give an equal or larger value then $t_{float,obs}$
(note that $t_{float,obs} = t_{fix,obs}(m_0 = \hat{m})$). This probability is

$$p_{fix} = \int_{t_{fix,obs}(\hat{m})} f(t_{fix}|H_0)dt_{fix} \tag{2.11}$$

i.e. this corresponds to the fixed mass scenario, had the pre-defined mass m_0 would have been set equal to $\hat{m}$. The trial factor is defined as the ratio of two above probabilities,

$$trial\#_{observed} = \frac{p_{float}}{p_{fix}} \tag{2.12}$$

The two p-values defined above are shown as the shaded areas in Fig. (5) (left plot).

Using the approximation we obtained for small p-values (high significance), and with $p_{fix} = p_{\chi^2_1}$ from Wilk's theorem, we have

$$trial\#_{observed} \simeq \frac{p_{\chi^2_2}\langle N\rangle}{p_{\chi^2_1}} \tag{2.13}$$

for high significance we can also approximate $p_{\chi^2_1}$ with $\frac{1}{\sqrt{t_{obs}}}\sqrt{\frac{2}{\pi}}e^{-t_{obs}/2}$, while $p_{\chi^2_2}$ is exactly given by $e^{-t_{obs}/2}$. We therefore have

$$trial\#_{observed} \simeq \langle N\rangle\sqrt{\frac{\pi}{2}}\sqrt{t_{obs}} = \langle N\rangle\sqrt{\frac{\pi}{2}}Z_{fix} \tag{2.14}$$

where Z_{fix} is the quantile of a standard gaussian with the same p-value (i.e. number of standard deviations). The trial factor is therefore proportional to the fixed mass significance, and to the average number of local minima.

case(b). Here we want to estimate the experiment expected sensitivity for a discovery, given some signal hypothesis. We have a Monte-Carlo prediction of the *expected* number of background and signal events for some hypothesized mass value m_0, and we wish to estimate the median significance of the experiment, assuming the true mass is equal to m_0. The median p-value is:

$$p_{float,med} = \int_{t_{float,med}} f(t_{float}|H_0)dt_{float} \tag{2.15}$$

where $t_{float,med}$ is the median value of t_{float}. The median of the local p-value is defined as the probability that a background fluctuation at m_0 will give an equal or larger value then the median value of $t_{fix}(m_0)$, i.e. :

$$p_{fix,med} = \int_{t_{fix,med}(m_0)} f(t_{fix}|H_0)dt_{fix} \tag{2.16}$$

and the trial factor is defined as the ratio between the two above probabilities,

$$trial\#_{expected} = \frac{p_{float,med}}{p_{fix,med}} \tag{2.17}$$

The two p-values defined above are shown as the shaded areas in Fig. 5 (right plot).

It can be shown that for high significance, $t_{float,med} \simeq t_{fix,med} + 1$ [6]. Using the same approximations as before we have $p_{float,med} \simeq \langle N \rangle e^{-t_{float,med}/2} = \langle N \rangle e^{-t_{fix,med}/2} \frac{1}{\sqrt{e}}$, therefore

$$trial\#_{expected} \simeq \langle N \rangle \sqrt{\frac{\pi}{2e}} Z_{fix} \tag{2.18}$$

Both trial factors are increasing with the significance as can be seen in Figure 6.

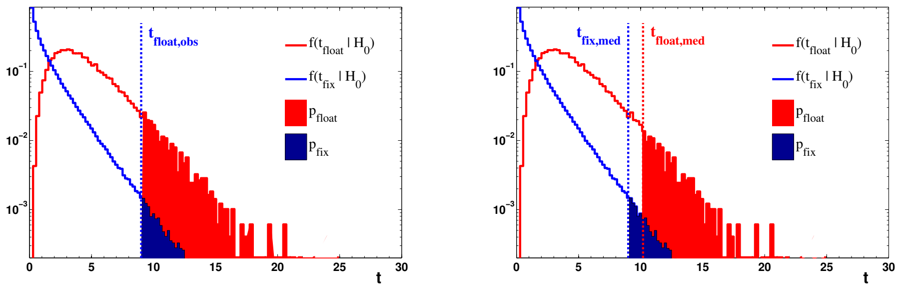

Figure 5: Demonstration of the p-values (shaded areas) used to define the trial factors in case **a** (left) and **b** (right).

We find empirically (Fig. 4) that the relation between the average number of local maxima $\langle N \rangle$ and the thumb-rule number is such that

$$trial\#_{observed} \simeq \frac{1}{3} \frac{range}{resolution} Z_{fix} \tag{2.19}$$

and equivalently

$$trial\#_{expected} \approx \frac{1}{3\sqrt{e}} \frac{range}{resolution} Z_{fix} \tag{2.20}$$

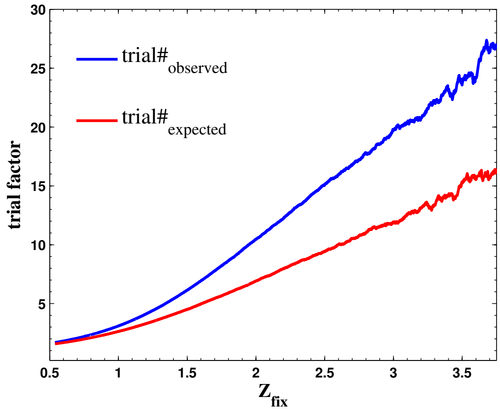

Figure 6: The trial factors as a function of the fixed mass significance Z_{fix}.

and the approximation is found to be good for a significance $\gtrsim 2$.

More complicated scenarios, in which e.g. the mass resolution depends on the mass, may occur. In that case one should not expect the same relation between $\langle N \rangle$ and the thumb-rule number as above (namely, roughly one local minimum per three signal widths). Such cases, are dealt in detail in [3] where the result is still that asymptotically, the trial factor grows linearly with the (fixed mass) significance.

3. Frequentist Exclusion and CL_s

Exclusion of a signal with a strength μ occurs when the signal+background hypothesis H_μ is rejected at the 95% Confidence Level. This means that the observed p-value under H_μ (p_μ) is less than 5%. The procedure could be standard, except for the fact, that, a downward fluctuation in the expected background could lead to exclusion of very weak signals (low cross section) to which the experiment has no sensitivity. The CL_s method for setting upper limits was originally introduced in high energy physics as a generalization of the conditional interval proposed by Zech [8] for the single channel counting experiment. In that case, given an observation of n^{obs} events, the confidence level is defined according to the p-value:

$$P(n < n^{obs}|s+b, n_b < n^{obs}) = \frac{P(n < n^{obs}|s+b)}{P(n_b < n^{obs})} = \frac{P(n < n^{obs}|s+b)}{P(n < n^{obs}|b)} \tag{3.1}$$

where n_b is the (unknown) number of background events in the sample. This is the probability, given $n_b < n^{obs}$, to observe n^{obs} or less events, assuming some signal rate s. Confidence intervals constructed from conditional probabilities as above are referred to as conditional intervals. In the context of setting upper limits on an unknown signal rate, considering such probability seems to be more relevant to the question one is trying to answer, compared to the unconditional one, $P(n < n^{obs}|s+b)$. This is because we do not want the answer to be affected by how unlikely the background fluctuation is, when we *know* that such an unlikely fluctuation has occurred. As a consequence, large downwards fluctuations of the background do not lead the exclusion of very small signals. CL_s was then defined as a “generalization” of the above p-value to more complicated cases, by simply replacing the right-hand side of (3.1) with the corresponding generalized p-values [9]:

$$CL_s = \frac{P(q_\mu > q_\mu^{obs}|\mu)}{P(q_\mu > q_\mu^{obs}|0)} \tag{3.2}$$

Where q_μ is some test statistics corresponding to a hypothesized signal strength μ.

One of the main objections to CL_s is the fact in the transition from (3.1) to (3.2), the frequentist meaning of the construction is lost, namely that it is not clear what is the conditional probability, analogous to the left-hand side of (3.1), that CL_s is equal to (or if such probability exists at all). In general, however, there is no reason why an analogue of the conditional probability could not be constructed to begin with. Such a construction may be or may not be equal to CL_s, but it will retain the original frequentist meaning of (3.1).

As an example for such a construction we consider a general likelihood ratio test statistic in the large sample limit, where the sampling distributions can be approximated by the asymptotic limit of Wald [12]. We then consider the quantity

$$\Delta_\mu = \hat{\mu} - \mu \tag{3.3}$$

where $\hat{\mu}$ is the maximum likelihood estimator of μ (for which the true value is unknown). Since μ is assumed to be positive, Δ_μ is constrained by the data to be

$$\Delta_\mu \le \hat{\mu}^{obs} \tag{3.4}$$

and we define the p-value of the data to be the conditional probability

$$P(q_\mu > q_\mu^{obs}|\mu, \Delta_\mu \le \hat{\mu}^{obs}) = \frac{P(q_\mu > q_\mu^{obs}|\mu)}{P(\Delta_\mu \le \hat{\mu}^{obs})} \tag{3.5}$$

In the limit we are considering, $\hat{\mu}$ is normally distributed around μ and the distribution of Δ_μ is independent of μ. Furthermore q_μ is a monotonically decreasing function of $\hat{\mu}$ [6]. The denominator of (3.5) can be therefore replaced with

$$P(\Delta_\mu \le \hat{\mu}^{obs}) = P(\hat{\mu} < \hat{\mu}^{obs}|0) = P(q_\mu > q_\mu^{obs}|0) \tag{3.6}$$

which leads to the definition (3.2) of CL_s. Therefore when the large sample approximations can be used, CL_s can be interpreted as the frequentist conditional probability (3.5). Practically, the conditional probability can be taken into account by modifying p_μ to

$$p_\mu^{CL_s} = \frac{p_\mu}{1 - p_0} \tag{3.7}$$

4. The Equivalence between Bayesian and Frequentist Exclusion

In the Bayesian approach we assign a degree of belief to the signal and background with priors, $\pi(\mu)$ and $\pi(b)$. Let μ be the signal strength, the posterior probability for μ is given by

$$p(\mu|data) = \frac{\int \mathscr{L}(\mu\mathbf{s} + \mathbf{b})\pi(\mu)\pi(\mathbf{b})d\mathbf{b}}{\int\int \mathscr{L}(\mu\mathbf{s} + \mathbf{b})\pi(\mu)\pi(\mathbf{b})d\mu d\mathbf{b}} \tag{4.1}$$

PoS(ACAT2010)007

To set an upper limit on the signal strength , μ, one calculates the credibility interval $[0,\mu_{95}]$

$$0.95=\int_0^{\mu_{95}} p(\mu|data)d\mu \tag{4.2}$$

Improper flat priors are used in High Energy Physics. For example quoting ref [7]:*Because there is no experimental information on the production cross section for the Higgs Boson, in the Bayesian technique we assign a flat prior to the total number of selected Higgs events*. We do not justify the use of flat priors here. However, if one uses flat priors one finds using the saddle-point approximation

$$p(\mu|data)=\frac{\int\mathcal{L}(\mu\mathbf{s}+\mathbf{b})d\mathbf{b}}{\int\int\mathcal{L}(\mu\mathbf{s}+\mathbf{b})d\mu d\mathbf{b}}=\frac{e^{\ln\mathcal{L}(\mu\mathbf{s}+\hat{\hat{\mathbf{b}}})}}{e^{\ln\mathcal{L}(\hat{\mu}\mathbf{s}+\hat{\mathbf{b}})}}=\frac{\mathcal{L}(\mu\mathbf{s}+\hat{\hat{\mathbf{b}}})}{\mathcal{L}(\hat{\mu}\mathbf{s}+\hat{\mathbf{b}})} \tag{4.3}$$

There is therefore an equivalence between the Bayesian posterior probability and the profile likelihood ratio when using flat priors.

5. Conclusions

Statistical methods used for discovery and exclusion of signal at the LHC and the TEVATRON were described. We showed a full formalism of the Look Elsewhere Effect. Formulas were derived that allow the estimation of the effect from the simple fixed mass result without the need to perform complicated Monte Carlo simuations. We have shown that the CL_s method could be interpreted as a frequentist method. We have also shown that deriving Bayesian upper limits with flat priors is equivalent to using a frequentist Profile Likelihood.

6. Acknowledgements

One of the authors (E.G.) would like to thank Louis Lyons for initiating this talk and providing useful suggestions.

References

[1] *Statistical Issues for Higgs Physics*, Eilam Gross and Ofer Vitells. Proceedings os the EPS-HEP 2009 meeting, Poland, PoS(EPS-HEP 2009)223.

[2] For a general discussion see for example L. Lyons, *Open statistical issues in particle physics*, The Annals of Applied Statistics 2008, Vol. 2, No. 3, 887-915.

[3] *Trial factors for the look elsewhere effect in high energy physics*, E. Gross and O. Vitells, The European Physical Journal C - DOI: 10.1140/epjc/s10052-010-1470-8arXiv:1005.1891

[4] S.S. Wilks, *The large-sample distribution of the likelihood ratio for testing composite hypotheses*, Ann. Math. Statist. **9** (1938) 60-2.

[5] W. Quayle, *Combining chanells with fits*, PHYSTAT-LHC Workshop 2007.

[6] G. Cowan, K. Cranmer, E. Gross and O. Vitells, *Using the Profile Likelihood in Searches for New Physics*, to be published.

[7] *Combined CDF and D0 Upper Limits on Standard Model Higgs-Boson Production with 2.1 - 5.4 fb-1 of Data* arXiv:0911.3930V1

[8] G. Zech, Nucl. Instrum. Methods Phys. Res. A 277, 608 (1989).

[9] A. L. Read, J. Phys. G: Nucl. Part. Phys. 28 2693 (2002).

[10] B. P. Roe and M. B. Woodroofe, Phys. Rev. D 60, 053009 (1999).

[11] R. D. Cousins, Phys. Rev. D 62, 098301 (2000).

[12] A. Wald, *Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large*, Transactions of the American Mathematical Society, Vol. **54**, No. 3 (Nov., 1943), pp. 426-482.