LHC Statistics for Pedestrians

Eilam Gross
Weizmann Institute, Rehovot, Israel

Abstract
A pedestrians guide aimed at the LHC laymen statisticians is presented. It is not meant to replace any text book but to help the confused physicist to understand the jargon and methods used by HEP Phystatisticians[1]

1 Introduction

The first Phystat meeting was a workshop at CERN on Confidence Limits followed by a similar workshop at Fermilab. Fred James who organized the meeting with Louis Lyons presented then his personal wish list titled: 'What I would like to see'. Fred wishes that physicists learn the vocabulary of statistics. This pedestrian guide is aimed at the Atlas and CMS physicists who wish to become Phystatisticians so that when ATLAS or CMS publish a combined limit or discovery significance they will know what it is all about.

When interpreting the result of the experiment, there are two alternate questions and one must not confuse between them. Question number one would be: Did I or did I not establish a discovery? Question number two would be: How well does my alternate model describe this discovery? The first question has to do with the goodness of the fit of the observed data to the good and old Standard Model while the second question has to do with hypotheses testing and the derivation of confidence intervals and upper limits. The LHC physics community is not only a mixture of physicists speaking all sorts of languages, from Hebrew and Chinese to English, German and French but who are also refugees of all sorts of experiments each with its preferred statistical method. Physicists educated at LEP advocate the CLs method while some Tevatron physicists prefer Bayesian methods with some of their friends from BaBar and Belle using pure frequentist methods. It seems that the only way out is to do it all... But, in a way, as we will show, conceptually one way leads to another.

But in order to introduce the different methods and compare them a basic lesson in the related statistics jargon is necessary.

2 Test Statistics

A test statistic is a quantity calculated from our sample of data. Its value can be used to estimate how probable is the result that we observe with respect to some null hypothesis. A physicist's intuition will attribute the null hypothesis to the 'background only' hypothesis. Normally it depends on the nature of the problem, but in this write up we will stick to this definition. In this context the value of the test statistic is used to decide whether or not the null hypothesis should be rejected in our hypothesis test.

It is important to note that the observed test statistics is based on our ONE experiment and could be a result of years of data collecting! Normally to conclude anything based on the observed test statistic one needs the pdf of the test statistic. This can sometimes be calculated analytically but can always be generated with toy Monte Carlo experiments.

A consequence of the Neyman-Pearson lema is that if H_0 is the null hypothesis (background only) and H_1 is the alternate hypothesis (say, a Higgs Boson with a mass m) then the most powerful test statistic one can construct (in absence of systematics) is the Likelihood ratio

$$Q(m) = \frac{L(H_1)}{L(H_0)} = \frac{L(s(m)+b)}{L(b)} \tag{1}$$

[1] A Phystatistician is a Physicist who knows his way in Statistics and knows how Kendall's advanced theory of statistics book looks like....

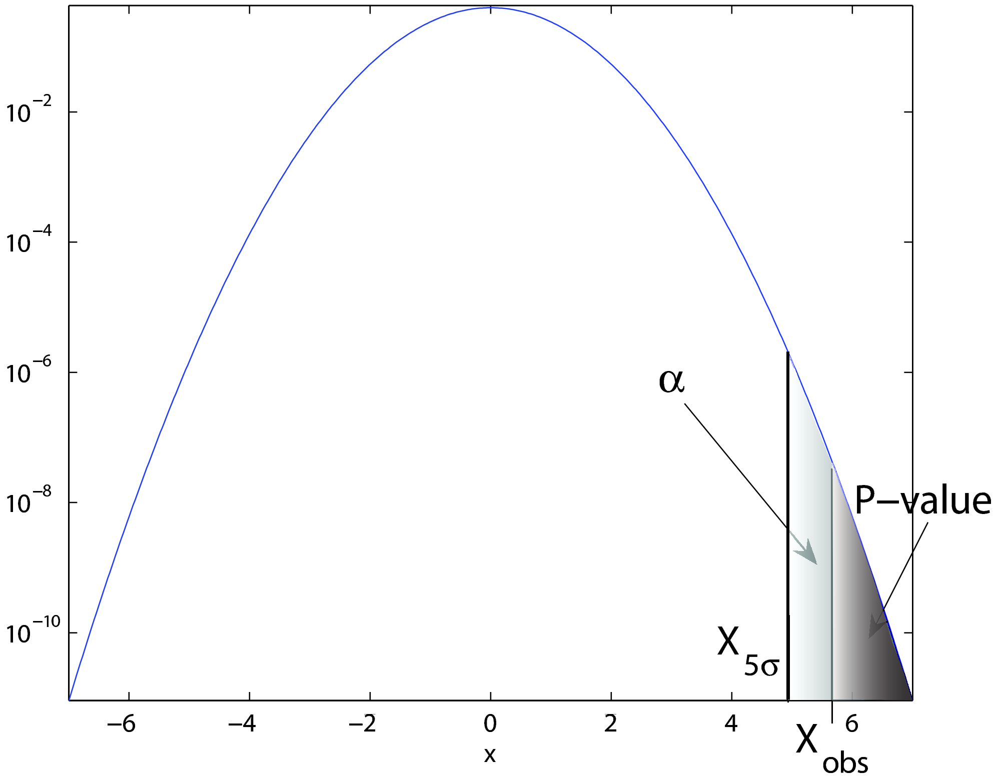

Fig. 1: An illustration showing the control area α and the $p - value$ of a Gaussian distribution. Note, in this example $X_{obs} > X_{5\sigma}$.

In a counting experiment s and b would be the average number of the expected signal and background events and the Likelihoods would be derived from the data using Poisson statistics.

3 p-value

At LEP, trying to discover the Higgs boson, people examined the distribution of the observed $1 - CL_b$ as a function of the hypothesized Higgs mass and looked for troughs.... That might have been the right thing to do but the wrong statistical jargon. A discovery by definition is a deviation from the Standard Model, i.e. the "background only" hypothesis (H_0). Given the pdf of the test statistic for background only experiments, it is common in HEP to announce a discovery if the result is at least 5 σ away from the expectation. Given a pdf $g(x|H_0)$ of the test statistic x, one can define a control area of size α at the tail of the pdf distribution (for this example let us assume that the less probable result is on one side of the distribution only), i.e. $\alpha = \int_{x_{5\sigma}}^{\infty} g(x|H_0)dx$ (Figure 1). If the observed result $x_{obs} > x_{5\sigma}$ then the probability to get a result which is as or less compatible with the background hypothesis is given by $p = \int_{x_{obs}}^{\infty} g(x|H_0)dx$ and it is smaller than α. This probability is called the $p - value$ and a discovery is considered when $p < \alpha$. This means also that the background-only hypothesis is rejected with a probability of $1 - p$.

Historically physicists have the tendency to mix confidence with p-value. In looking for the Higgs, the LEP experiments used the ‘confidence level in the background, $1 - CL_b(m_H)$, where $CL_{b(m_H)}$ is defined as the tail area $CL_b = \int_{-\infty}^{x_{obs}} g(x|H_0)dx$, with the statistic x being the log likelihood ratio for the background plus signal model (i.e. Standard Model with Higgs of mass m_H) as compared with background only (i.e. H_0, the Standard Model with no Higgs in the observable mass range). This implies that in an ensemble of backgound only experiments, a fraction $1 - CL_b$ would be expected to have a larger value than the observed value. The terminology is confusing since $1 - CL_b$ is in fact a p value. The LEP experiments were looking for tiny values of $1 - CL_b$, which would indicate a very large fluctuation of the background (or the presence of a signal), but none was found. The correspondence between the hypothesis test property $1 - p$ and the background confidence estimation, CL_b is further discussed in [1].

4 The Look Elsewhere Effect

The Standard Model predicts a Higgs Boson but not its mass. It can be anywhere up to a few hundreds of GeV. We can specify an hypothesis with a specific Higgs mass but had we observed some possible signal we should take into account that this signal could be a fluctuation which could be observed anywhere in our sensitivity range [2]. Here we change the signal hypothesis from a Higgs with a specific mass m_H to a Higgs with some mass in the observed region. It is not clear how to take these effects into account. One common way is to degrade the observed p-value by multiplying it by the size of the sensitivity region divided by the experimental resolution. A common claim is that the control region for discovery is so small that "who cares".... Another common belief is that the "look elsewhere effect" is the reason for the habit of defining a discovery as a 5σ and not for example 4σ, because even if you quote 5σ your effective significance is lower.

5 Confidence Intervals and Coverage

Assume you have a measurement m_{meas} of m with m_t being the true value of m and suppose you know the pdf $p(m_{meas}|m)$. You use some method to calculate a 90% confidence interval $[m_1, m_2]$. What does it mean?

Most physicists interpret it as if the probability that there is a Higgs Boson with $m_t \in [m_1, m_2]$ is 90%. However, this is totally wrong. If you run a bunch of toy Monte Carlo experiments, each one will yield a different interval. The correct statement is that if there is a Higgs with a mass, m_t, then, in an ensemble of experiments, 90% of the obtained confidence intervals will contain the true value of m, m_t. More on the source of this misconception in section 6.

Subsequent to the above definition of interval is the notion of coverage. The confidence interval is estimated using the physicist preferable method. If in an ensemble of Monte Carlo experiments the true value of m is covered within (e.g.) 90% of the estimated confidence intervals, we claim a coverage. If it occurs less than 90%, the method is claimed to undercover.

Some physicists doubt the importance of coverage. Their claim is that coverage answers the wrong question. What we really want to know, so they claim, is the probability that the Higgs Boson exists and is in the specified mass interval. So there are two possibilities here. Either educate the physicists about the correct meaning of coverage or try to answer the "right" question...

6 Subjective Bayesian

What is the "right" question? It must be: Is there a Higgs Boson? When pronouncing this question, I cannot escape from an immediate association to the question: Is there a God? Can one really answer this question based on the data (earth)? The answer is yes, but with many significant prior assumptions.... each weakens the credibility of the answer.

I believe that the source of the common misconception regarding the interpretation of a confidence interval is that our mind is sometimes acting in a Bayesian manner. We try to deduce something about the Higgs ("asking the right question"), we derive a confidence integral and translates it to our degree of belief that there is a Higgs given the data, i.e. $Prob(m_t \in [m_1, m_2]|data)$.

A model (A Higgs Boson with a mass m) can only be assigned a degree of belief, but not a probability in a frequentist manner (i.e. as a random variable in a repetitive set of experiments).

The relation between the degree of belief and the true probabilities is given by the Bayesian relation

$$Prob(Higgs|data) = \frac{L(data|Higgs)\pi(Higgs)}{Normalization}$$

where $\pi(Higgs)$ is the prior for a Higgs Boson which many times is taken to be uniform in the Higgs mass (or simply 1) without even noticing!

Last comment here; in this approach instead of talking about confidence intervals we talk about credible intervals, where $p(Higgs|data)$ is the credibility of the Higgs given the data.

7 The Likelihood Principle

Bayesian inference obeys by definition the Likelihood Principle (LP). According to this, the Likelihood function $L(\{\theta\})$ contains the full information from the experimental data. A consequence to the LP is that methods that provide different results for a measurement yet have proportional likelihood functions are inconsistent. A nice discussion about the LP can be found in [3].

8 Who is Afraid of Nuisance Parameters?

The answer to the question appearing in the title is nobody, yet everybody.... Nobody, because Nuisance parameters is just the term used by statisticians for what we physicists refer to as systematics. Everybody, because systematics can kill an experimental observation if not under control. The significance of an observation is given in the limit of large numbers as $S/\sqrt{B}$, however, this number is degraded in the presence of a systematic uncertainty Δ on the background and becomes $S/\sqrt{B(1+\Delta^2\cdot B)}$, which in the limit of infinite luminosity (and large B) becomes $\frac{S}{B\cdot\Delta}$. So if there is 10% background uncertainty, one will never reach a 5σ significance if $S/B < 0.5$.

Physicists find difficulties in both classifying and estimating the systematic uncertainties and implementing them in the analysis interpretation. There are systematic errors that reduce with increasing statistics and therefore can be handled, and those that do not. In what follows, we will concentrate on the possible treatment of systematics in the interpretation phase of the analysis.

9 Integrating Out the Systematic Errors

When applied to Bayesian credibilities, integrating out the systematics via marginalization with a prior is a natural thing to do. If we denote by s the Higgs signal, by b the background which has some systematic uncertainty, the equation in section 6 becomes

$$p(s,b|data) = \frac{L(s,b|data)\pi(s,b)}{Normalization}$$

The prior is often assumed to factorize $\pi(s,b) = \pi(s)\pi(b)$ with the signal prior taken to be flat. Hence the background systematics is explicitly included in the background prior. We can then integrate the background systematics via $p(s|data) = \int p(s,b|data)db$.

Integrating the nuisance parameters is also used in the so called Cousins-Highland hybrid-frequentist technique [4]. Here the recipe is given by $p(data,data'|s) = \int p(data|s,b)p(b|data')db$ where the $data$ is used for the main measurement and the $data'$ for the auxiliary measurement of the background (e.g. via a side band). It is to be noted that one can fake an auxiliary measurement in order to apply for example a 5% systematics to the background. The Bayesian nature of this method is apparent by the use of the posterior $p(b|data')$.

10 Priors

A prior, e.g. $\pi(\lambda)$ is interpreted as a description of what we believe about a parameter λ preceding the current experiment. One can distinguish two kinds of priors. Informative priors which are based on some information one has on λ and uninformative priors. When the parameter is that of no-interest (nuisance) an auxiliary measurement might supply a legitimate basis for an informative prior. The Higgs signal, on the other hand, is a parameter of interest. Some would say that all priors of the parameters of interest should be uninformative. I would say that using the lower bound of 115 GeV on the Higgs mass as part of a prior, is hard to argue with..... But note also that choosing a prior is a science by itself. A prior flat in the coupling g is not flat in the cross section $\sigma \sim g^2$. That led to the development of reference priors [5]. Reference priors have a minimal effect (relative to the data) on our prospective final inference. In the simple one dimensional case, with one parameter, the reference prior is reduced to the Jeffry's prior which is metric invariant, i.e. $\int L(data|s,\lambda)\pi(\lambda)d\lambda = \int L(data|s,\lambda)\pi(f(\lambda))df$ and can be easily obtained in an analytic way.

11 Doing Justice with the CL_s Method

In section 3 we defined the confidence level CL_b. In a similar manner one can define the signal+background confidence level CL_{s+b}. But what is the meaning of a signal confidence level? Using the terminology of confidence levels CL_s was defined as $CL_s \equiv \frac{CL_{s+b}}{CL_b}$ [6].

The CL_s method is the most discredited method in HEP statistical inference.The reason is that it lacks a frequentist coverage. However, it lacks it in places where the experiment is insensitive to the expected signal! And this is not necessarily a disadvantage from some physicists point of view! Here is what happens:

One uses the Neyman-Pearson likelihood ratio as a test statistics (see section 2) and construct its pdf for background only and signal+background experiments. When the expected signal is very low the two pdf are almost overlapping (see Figure 2). When the number of observed events fluctuates far below the expected background, both hypotheses $s(m_H), s(m_H)+b$ are not favored, yet, given the low p-value of the $s+b$ hypothesis $p_{s+b} = 3\%$ for example, one might exclude the $s(m_H)+b$ hypothesis and the common physicist will interpret the result as if a Higgs with a mass m_h (e.g. 116 GeV in LEP case) is excluded at the 97% Confidence Level. But this is a false statement. To protect against such an inference one defines a new quantity with an unfortunate name $CL_s = \frac{p_{s+b}}{1-p_b}$. In the limit of a light Higgs mass $CL_s \overset{m_H\downarrow}{\longrightarrow} CL_{s+b}$. As a result the false exclusion rate is too low for heavy Higgs Bosons, i.e. the method undercovers where the experiment lacks sensitivity. However this is conservative because it avoids excluding when there is no sensitivity, while simple usage of the pure frequentist CL_{s+b} could result in an exclusion.

12 Neyman Construction

The Neyman construction is a method of parameter estimation that ensures coverage. One scans over all the possible true values of some parameter s and defines an acceptance interval for each s, based on the known pdf, $g(s_m|s)$, of the measured s_m given a possible true s (there is only ONE unknown true s though). The (e.g.) 68% acceptance interval $[s_l, s_h](s)$ is defined via the integration $[s_l, s_h](s) = \{s_m| \int_{s_l}^{s_h} g(s_m|s)ds_m = 68\%\}$ (Figure 3). Even in the simplest case where g is a Gaussian, there is an ambiguity in the choice of the integration limit, which will lead to two-sided intervals, or one-sided integral bounded from below or above. To sort out the integration limits one needs to specify an ordering rule. The construction of the acceptance intervals for all s turns out to be a belt from which one can easily get the corresponding (e.g.) 68% confidence interval $[s_d, s_u](s_o)$ (see section 5), given one measurement s_o via inversion (Figure 3). Due to space limitations there is no way I can describe here the Neyman construction in the necessary detail. Full descriptions can be found in [7].

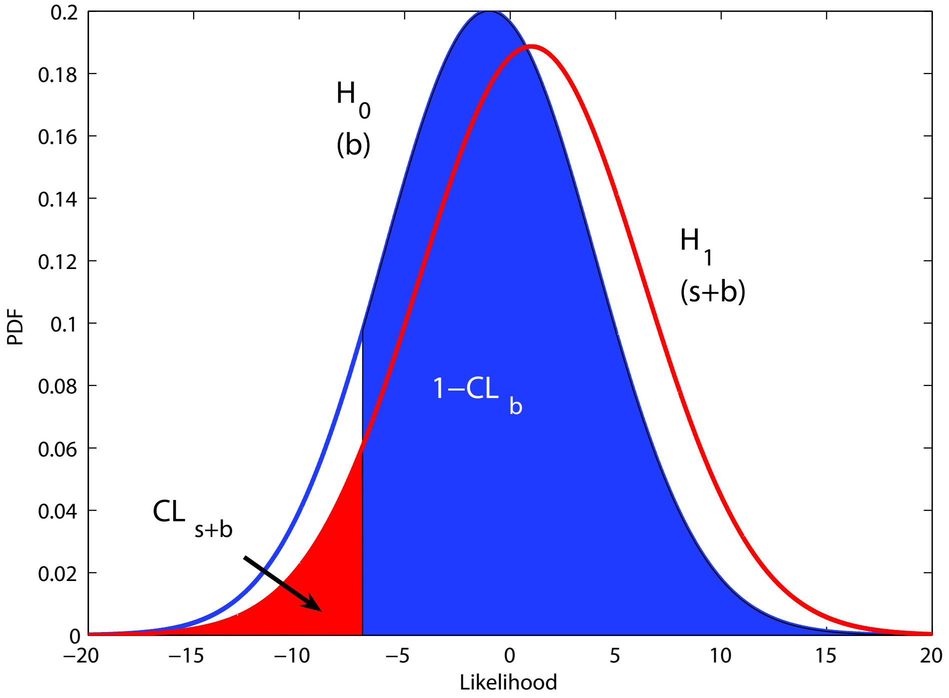

Fig. 2: An illustration showing the reasoning of the CL_s method. In this situation a signal+background hypothesis might be rejected though the experiment has no sensitivity to observe that particular signal.

13 The Feldman-Cousins Method

The full Neyman construction was introduced to HEP by Feldman and Cousins [8]. The test statistic is the likelihood ratio $Q(s) = \frac{L(s+b)}{L(\hat{s}+b)}$ where $\hat{s}$ is the physically allowed mean s that maximizes the Likelihood $L(\hat{s}+b)$. To construct an acceptance 68% interval in the number of observed events, $[n_1, n_2]$, one is using Q as an ordering rule, i.e. $\sum_{n_1}^{n_2} p(n|s,b) \geq 68\%$ where only terms with decreasing order of $Q(n)$ are included in the sum, till the sum exceeds the 68% confidence. When n_o events are observed, one is using this constructed Neyman belt to derive a confidence interval, which, depending on the observation, might be a one-sided or a two-sided interval. This method is therefore called the unified method, because it avoids a flip-flop of the inference (i.e. one decides to flip from a limit to an interval if the result is significant enough...).

The difficulty with this approach is that an experiment with higher expected background which observes no events might set a better upper limit than an experiment with lower or no expected background. This would never occur with the CL_s method.

Another difficulty is that this approach does not incorporate a treatment of nuisance parameters. However, it can either be plugged in "by hand", using the hybrid Cousins and Highland method [9] or a Neyman construction can be performed, as described below.

14 The Profile Likelihood Full Construction Method

Treating the background as a nuisance parameter, one can perform a full Neyman construction with the Feldman-Cousins test statistic used as an order $\ell(s) = \frac{L(s,\hat{\hat{b}})}{L(\hat{s},\hat{b})}$. This is a very cumbersome construction. In this relatively simple example, the construction is done in a 4-dimensional space, the two observables (n, b_m) and the two possible true values (s, b) . For each s the MLE of b is found, $\hat{\hat{b}}(s,n)$. So far only

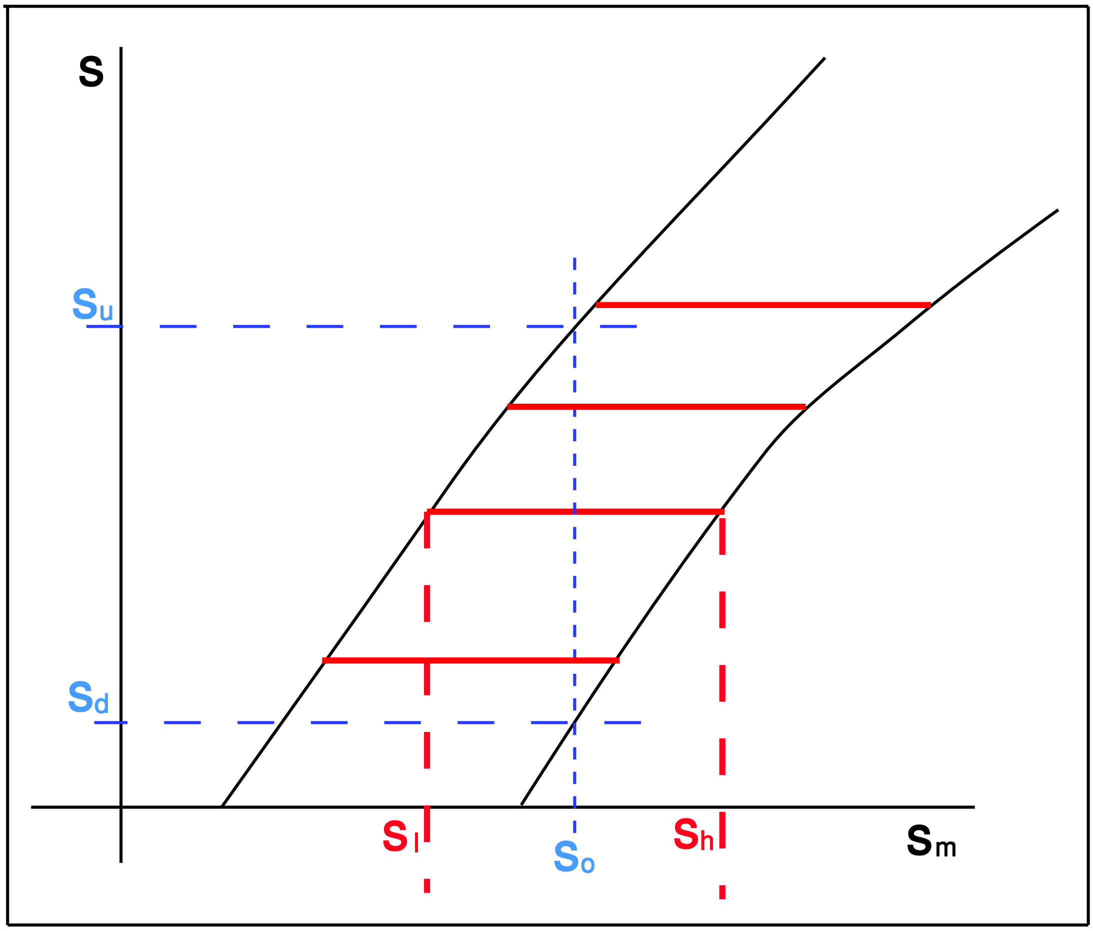

Fig. 3: An illustration showing the Neyman belt. The horizontal lines are the acceptance intervals in the measured parameter space s_m for a given possible true s, $[s_l, s_h](s)$. Given an observation s_o one can construct the confidence interval $[s_d, s_u]$.

low dimensional toy models were fully constructed [10]. To ease the procedure an approximate Neyman construction was suggested [11] by fixing $\hat{\hat{b}}$ to be $\hat{\hat{b}}(s, n_{obs})$. Gary Feldman does not recommend to try the full construction at home for many reasons [12]. One of them is that using a simple Profile Likelihood method works quite well.

15 The Profile Likelihood Method

The simplest way to incorporate systematics into hypothesis inference is the Profile Likelihood. High Energy Physicists are unaware of their familiarity with this method via its implementation in the MINOS process within MINUIT [13].

For simplicity let us define the Profile Likelihood for one channel as $\lambda(s) = \frac{L(s,\hat{\hat{b}}(s))}{L(\hat{s},\hat{b})}$. Here $\hat{\hat{b}}(s)$ is the MLE of b given s and $\hat{s}, \hat{b}$ are the MLE of s and b. When generating experiments, each with data distributed according to $Poisson(n, s+b)$ we find that the pdf of $-2ln\lambda(s)$ is distributed as a $\chi^2(1)$. This is not surprising since in the asymptotic limit the Likelihood function $L(s)$ becomes a Gaussian centered about the ML estimator $\hat{s}$, i.e. $lnL(\hat{s} \pm N\sigma_{\hat{s}}) = lnL_{max} - \frac{N^2}{2}$. The magic of the Profile Likelihood method is that the χ^2 approximation works very well and there is no need for toy Monte Carlo experiments... One can calculate the exclusion or discovery sensitivity or significance in a fraction of a second.

16 Acknowledgements

This work would have never been possible without the patience and amazing support of Bob Cousins, Kyle Cranmer and Glen Cowan. They helped me to become as close as possible to a Phystatistician in two months...

I would also like to thank Alex Read, Bill Quayle, Luc Demortier and my student Ofer Vitells for always being there to answer my questions and requests.

Last but not least, applause to Louis Lyons for keeping the Phystat meetings going. These meetings are the ultimate source of statistics for the High Energy Physicists. Long may these meetings live! Toda Louis.

I am obliged to the Benoziyo center for High Energy Physics for their support of this work. This work was also supported by the Israeli Science Foundation(ISF), by the Minerva Gesellschaft and by the Federal Ministry of Education, Science, Research and Technology (BMBF) within the framework of the German-Israeli Project Cooperation in Future-Oriented Topics(DIP).

References

[1] A. Stuart and J.K. Ord, Kendallś Advanced Theory of Statistics, Vol. 2, Classical Inference and Relationship, 5th Ed. (Oxford University Press, New York, 1991), tests of hypotheses, Table 20, chapter 20.10.

[2] See talks by Alexey Drozdetskiy and Luc Demortier, these proceedings. See also the CMS TDR appendix A.

[3] G. Zech, Confronting classical and Bayesian confidence limits to examples, contribution to the first Phystat meeting (workshop on Confidence Limits) CERN, Jan 2000. arXiv:hep-ex/0004011. See also refrences therein.

[4] R.D. Cousins and V.L. Highland. Incorporating systematic uncertainties into an upper limit. Nucl. Instrum. Meth., A320:331, 1992.

[5] Berger, J. O. and Bernardo, J. M. (1989). Estimating a product of means: Bayesian analysis with reference priors. J. American Statistical Association 84, 200-207; Berger, J. O. and Bernardo, J. M. (1992). On the development of reference priors. Bayesian Statistics 4 (J. M. Bernardo, J. O. Berger, D. V. Lindley and A. F. M. Smith, eds). Oxford: Oxford University Press, 61-77 (with discussion). See also Luc Demortier, Bayesian Reference Analysis for Particle Physics, Phystat05.

[6] Presentation of search results: the CLs technique, A L Read 2002 J. Phys. G: Nucl. Part. Phys. 28 2693-2704, **doi:10.1088/0954-3899/28/10/313**

[7] I would recommend the Statistics books of Fredrick James and Glen Cowan for a full description of the Neyman construction.

[8] Gary J. Feldman and Robert D. Cousins. A unified approach to the classical statistical analysis of small signals. Phys. Rev., D57:38733889, 1998.

[9] J. Conrad et al.,Including systematic uncertainties in confidence interval construction for Poisson statistics, Phys. Rev. D67 (2003) 012002

[10] K. Cranmer, Frequentist hypothesis testing with background uncertainty. PhyStat2003 physics (2003) 0310108 .

[11] Gary Feldman,Multiple Measurements and Parameters in the Unified Approach,Workshop on Confidence Limits Fermilab March 28, 2000

[12] Gary Feldman, Concluding Remarks: Phystat 2005

[13] F.James and M. Roos, Comput.Phys.Commun. 10, 343 (1975);