Practical Statistics for High Energy Physics

Eilam Gross
eilam.gross@weizmann.ac.il
Weizmann Institute of Science, Rehovot, Israel

Abstract
In these lecture notes the frequentist methods used in the Higgs search, discovery and measurement are reviewed. The idea is that the reader will be able to understand what lies beneath the surface of the results and the plots shown in the experiments publications. Though the results shown are mainly from ATLAS and CMS, the methods and the lessons can be propagated to other fields such as Astro-Particles and fixed target experiments.

Keywords
CERN report; ESHEP; statistics; Data Analysis.

1 Introduction

These lecture notes are based on statistics lectures I gave in the European CERN school for High Energy Physics, 2015. They contain material published mainly in the following two papers: "Asymptotic formulae for likelihood-based tests of new physics" by Cowan, Cranmer, Gross and Vitells [1] and "Trial factors or the look elsewhere effect in high energy physics" by Gross and Vitells [2]. The frequentist approach used in the Higgs search, discovery and measurement are reviewed. Examples from real data analysis are given to clarify the methods.

2 The Search for the Higgs Boson

From Wikipedia: On 4 July 2012, the discovery of a new particle with a mass between 125 and 127 GeV/c2 was announced; physicists suspected that it was the Higgs boson. Since then, the particle has been shown to behave, interact, and decay in many of the ways predicted by the Standard Model.

High Energy Physicists (HEP) rely on a hypothesis: The Standard Model. This model relies on the existence of the 2012 discovery of the Higgs Boson. The minimal content of the Standard Model includes the Higgs Boson, the Quarks, the Leptons and the force mediating Bosons including the photons, gluons, W and Z. However, the Standard Model suffers from some problems, e.g. the hierarchy and naturality problems that are solved by various extensions of the Model and include other particles that are yet to be discovered. The challenge of HEP is to generate tons of data and to develop powerful analyses to tell if the data indeed contains evidence for new particles. Once the new particle, such as the 2012 scalar, has been discovered, the next step would have been to measure its mass, and confirm that it has the expected properties of the Higgs Boson (Spin, CP). Perhaps it is not the expected Standard Model Higgs Boson, but a member of a family of Scalar Bosons, the rest, yet to be discovered.

The statistical challenge is obvious: to tell in the most powerful way, and to the best of our current scientific knowledge, if, in our data, there is new physics, beyond what is already known. In that sense, what is already known is the background to what we search, which is treated as the signal. The complexity of the apparatus and the physics (both signal and background) suffer from large systematic errors that should be taken care of in a correct statistical way.

Though the Higgs Boson has been already discovered, in these lecture notes, for pedagogic reasons, it is assumed, that, the so-called Standard Model, contains no Higgs Boson, serve as the background to the signal, which is the Higgs Boson. The Higgs Boson cannot exist without the Standard Model, so

there are two nested hypotheses tested against each other. The Standard Model (denoted by b for background) and the Standard Model containing a Higgs Boson with a mass m_H, i.e. the signal+background, denoted by $s(m_H) + b$.

3 Essential Terminology

3.1 A Tale of Two Hypotheses

From Wikipedia: A hypothesis (plural hypotheses) is a proposed explanation for a phenomenon. For a hypothesis to be a scientific hypothesis, the scientific method requires that one can test it. Scientists generally base scientific hypotheses on previous observations that cannot satisfactorily be explained with the available scientific theories.

The expected signal and background are determined by the corresponding cross sections, luminosity delivered by the accelerator and the detectors response (efficiency and geometrical acceptance). $s(m_H)$ is given by

$$s(m_H) = L \cdot \sigma_{SM}(m_H) \cdot \epsilon \cdot A. \tag{1}$$

Where L is the luminosity delivered by the accelerator, $\sigma_{SM}(m_H)$ is the Standard Model (SM) production cross section of the Higgs Boson, and ϵ and A are the efficiency and geometrical acceptance of the detector. For simplicity, let's assume a counting experiment and let n be the number of observed events, then

$$n = \mu s(m_H) + b. \tag{2}$$

b is the expected background, and μ is the signal strength given by

$$\mu = \frac{\sigma_{obs}}{\sigma_{SM}}. \tag{3}$$

There are therefore two hypotheses. One is the background only (b), and the other is the $\mu s(m_H) + b$ hypothesis, i.e., a Higgs Boson with a strength μ on top of the background. For a Standard Model Higgs Boson, we expect to measure $\mu = 1.0$. The background only hypothesis is denoted by H_0 while H_μ is the Higgs Boson hypothesis with H_1 being the SM Higgs Boson hypothesis.

3.2 Testing an Hypothesis

From Wikipedia: A statistical hypothesis test is a method of statistical inference. Commonly, two statistical data sets are compared, or a data set obtained by sampling is compared against a synthetic data set from an idealized model. A hypothesis is proposed for the statistical relationship between the two data sets, and this is compared as an alternative to an idealized null hypothesis that proposes no relationship between two data sets. The comparison is deemed statistically significant if the relationship between the data sets would be an unlikely realization of the null hypothesis according to a threshold probabilityÑthe significance level. Hypothesis tests are used in determining what outcomes of a study would lead to a rejection of the null hypothesis for a pre-specified level of significance.

The first step in any hypothesis testing is to identify and state the relevant null, H_{null} and alternative H_{alt} hypotheses.The next step is to define a test statistic, q, under the null hypothesis (the tested hypothesis). We then compute from the observations the observed value q_{obs} of the test statistic q. Finally, decide (based on q_{obs}) to either fail to reject the null hypothesis or reject it in favour of an alternative hypothesis.

3.3 Discovery and Exclusion in a Nut Shell

To establish a discovery we define the null hypothesis as the background only hypothesis, $H_{null} = H_0$, and test it. We either fail to reject it or manage to reject it in favour of the alternative hypothesis, $H_{alt} = H_\mu$. Rejection of the null H_0 hypothesis at the level of 5σ (see 3.5) is considered a discovery.

Defining the null hypothesis as $H_{null} = H_\mu$ enables the exclusion of the signal. For example, if we define the null hypothesis as the Standard Model Higgs with a mass m_H, $H_{null} = H_1$, testing and rejecting this hypothesis at the 95% Confidnece Level (see 3.5) is considered an exclusion of the Standard Model Higgs with a mass m_H.

3.4 A Test Statistic

As defined in Wikipedia: A hypothesis test is typically specified in terms of a test statistic, considered as a numerical summary of a data-set that reduces the data to one value that can be used to perform the hypothesis test. In general, a test statistic is selected or defined in such a way as to quantify, within observed data, behaviours that would distinguish the null from the alternative hypothesis, where such an alternative is prescribed, or that would characterise the null hypothesis if there is no explicitly stated alternative hypothesis, which often occurs when performing a measurement.

One example for using a test statistic is the discovery of the Higgs, when the data of Billions of Collisions is summarised in one number which determines if LHC rejected the background only hypothesis in favour of the Higgs Boson with a mass m_H or not.

There are many ways to define a test statistic based on the nature of the required test. Test statistics for discovery or exclusion are commonly based on Likelihood ratios.

Note that the likelihood is a function of the data, i.e.

$$L(H_0) = Prob(x|H_0) \tag{4}$$

where x is the data.

Before classifying the test statistics in a formal way, let us take a simplified approach. The two most common test statistics in High Energy Physics are the Neyman-Pearson (NP) and Profile Likelihood (PL). The NP test statistic given by

$$q^{NP} = -2ln\frac{L(H_0)}{L(H_1)}. \tag{5}$$

$L(H_0)$ and $L(H_1)$ are the likelihoods of the null (b) and alternative ($s(m_H) + b$) hypotheses. Note that inverting the roles of the null and alternative hypotheses, simply swap the sign of the NP test statistic. The PL test statistic depends on the tested hypothesis and for a simple counting experiment (see Equation 2), when testing the b-only hypothesis, H_0, the test statistic is given by

$$q_0 = -2ln\frac{L(b)}{L(\hat{\mu}s(m_H) + b)}. \tag{6}$$

$\hat{\mu}$ is the Maximul Likelihood Estimators (MLE) of μ. In this simplified example b is assumed to be known. The probability distribution function (PDF) of both test statistics under the null $f(q^{NP}|b)$,$f(q_0|b)$ and the alternative $f(q^{NP}|s(m_H) + b)$,$f(q_0|s(m_H) + b)$ hypotheses are shown in Figure 1.

3.5 What is the p-value

As defined in Wikipedia: An important property of a test statistic is that its sampling distribution under the null hypothesis must be calculable, either exactly or approximately, which allows p-values to be calculated.

The observed $p - value$ is a measure of the incompatibility of the data with the tested hypothesis. It is the probability, under assumption of the null hypothesis H_{null}, of finding data of equal or greater incompatibility with the predictions of H_{null}. This is clearly illustrated in Figure 1 for the PL test statistic by the light blue area (right plot). Here H_0 is the tested null hypothesis (b only) and the $p - value$ is given by

$$p = \int_{q_{0,obs}}^{\infty} f(q_0|b)dq_0. \tag{7}$$

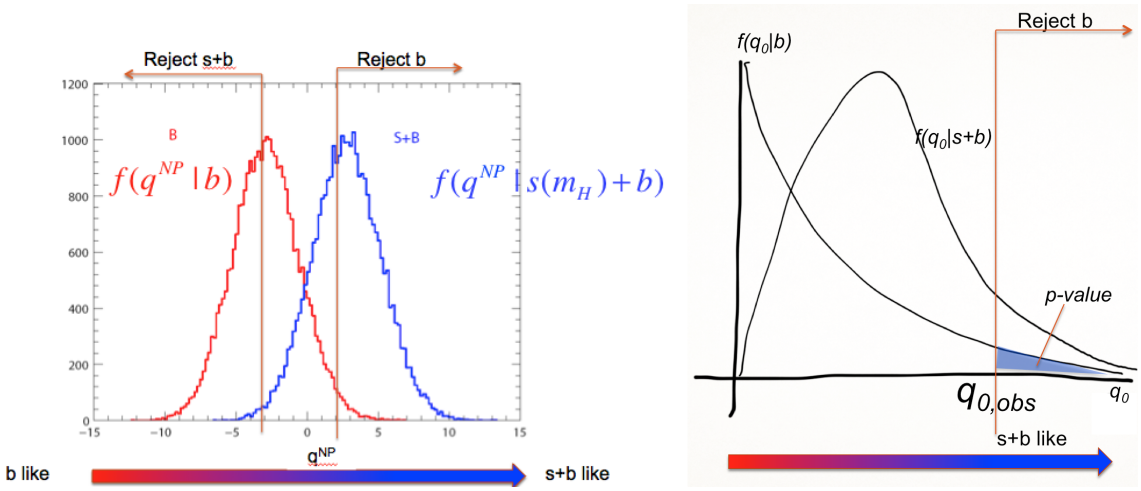

Fig. 1: The pdf of the Neyman-Pearson q^{NP} (left) and PL (Profile-Likelihood), q_0 (right) test statistics, under the null (b) and alternative ($s(m_H) + b$) hypotheses.

One can regard the hypothesis as excluded if its p-value is observed below a specified threshold (usually denoted by α).

Now, depending on the nature of the statistical test, one considers a one-sided or two-sided p-value. When performing a measurement, any deviation above or below the mean is drawing our attention and might serve an indication of some anomaly or new physics. Here we consider a two sided p-value. However, when trying to reject an hypothesis while performing searches, one usually considers only one-sided tail probabilities. When the null hypothesis is the b-only hypothesis, downward fluctuations of the background, are not considered as an evidence against the background. Likewise, when deriving a limit, upward fluctuations of the hypothesised signal are not considered as an evidence against the signal. In both cases only one-sided tail probabilities are considered.

In particle physics, when performing searches, one usually converts the p-value into an equivalent significance, Z defined such that a Gaussian distributed variable, which is found Z standard deviations above its mean, has an upper-tail probability equal to p (Figure 2). That is,

$$Z = \Phi^{-1}(1 - p)\,, \tag{8}$$

where Φ^{-1} is the quantile (inverse of the cumulative distribution) of the standard Gaussian. For a signal process such as the Higgs boson, the particle physics community has a tendency to regard rejection of the background hypothesis with a significance of at least $Z = 5$, as an appropriate level to constitute a discovery. This corresponds to $p = 2.87 \times 10^{-7}$. For purposes of excluding a signal hypothesis, a threshold p-value of 0.05 (i.e., 95% confidence level) is often used, which corresponds to $Z = 1.64$. This should not be confused with a 1.96σ fluctuation of a Gaussian variable that gives 0.05 for the two-sided tail area.

Note that, for a sufficiently large data sample, one would obtain a p-value of 0.5 for data in perfect agreement with the expected background. With the definition of Z given above, this gives $Z = 0$.

3.6 Expected Significance and the Asimov Data Set

As defined in Wikipedia: The use of a single representative individual to stand in for the entire population can help in evaluating the sensitivity of a statistical method. Franchise, a science fiction short story by Isaac Asimov, was cited as the inspiration of the term "Asimov data set", where an ensemble of simulated experiments can be replaced by a single representative one.

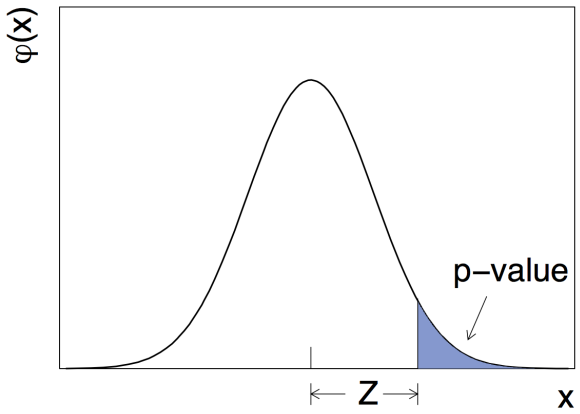

Fig. 2: The relationship between a p-value and a significance of Z sigma.

It is often useful to quantify the sensitivity of an experiment by reporting the expected significance one would obtain with a given measurement under the assumption of various hypotheses. For example, the sensitivity to discovery of a given signal process H_1 could be characterized by the expectation value, under the assumption of H_1, of the value of Z obtained from a test of H_0. This would not be the same as the Z obtained using Eq. (8) with the expectation of the p-value, however, because the relation between Z and p is nonlinear. The median Z and p will, however, satisfy Eq. (8) because this is a monotonic relation. Therefore we take the term 'expected significance' to refer to the median.

In the Standard Model there is only one Higgs Boson with well defined couplings. To find the discovery sensitivity of an experiment, one needs to generate one ensemble of experiments containing the Higgs Boson at the tested mass. However, if one goes beyond the Standard Model, e.g., supersymmetric models, one faces a multi-dimensional parameter space where the Higgs Boson's couplings, and hence its production cross section and decay properties (both related to the signal strength) vary as a function of the parameters. For each point in parameter space one needs to estimate the experiment's discovery sensitivity. One faces the need to generate an enormous number of ensembles of experiments and evaluate the median sensitivity for each ensemble.

In [1] it was shown that one can replace each ensemble of the alternate-hypothesis experiments with one data set that represents the typical experiment. This "Asimov" data set delivers the desired median sensitivity. Hence, one is exempted from the need to perform an ensemble of experiments for each set of parameters.

The Asimov data set is constructed such that when one uses it to evaluate the estimators for all parameters, one obtains the true parameter values.

As intuitively used for years till proven at [1], the Asimov data set can trivially be constructed from the true parameters values. For example, in a counting experiment (see Eq. 2) the Asimov data set corresponding to the H_1 hypothesis is $n_A = s + b$. and the one correspond to the H_0 hypothesis is $n_A = b$. As strange as it reads, the Asimov data set is not necessarily an integer.

3.7 Nuisance Parameters.

From Wikipedia: In statistics, a nuisance parameter is any parameter which is not of immediate interest but which must be accounted for in the analysis of those parameters which are of interest.

A widely used procedure to establish discovery (or exclusion) in particle physics is based on a frequentist significance test using a likelihood ratio as a test statistic. In addition to parameters of interest such as the rate (cross section) of the signal process, the signal and background models will contain in

general *nuisance parameters* whose values are not taken as known *a priori* but rather must be fitted from the data.

It is assumed that the parametric model is sufficiently flexible so that for some value of the parameters it can be regarded as true. The additional flexibility introduced to parametrise systematic effects results, as it should, in a loss in sensitivity. To the degree that the model is not able to reflect the truth accurately, an additional systematic uncertainty will be present that is not quantified by the statistical method presented here.

Here, nuisance parameters are denoted by θ.The likelihood is then a function of the parameter of interest, say, μ. Then $L = L(\mu, \theta)$. When testing H_μ, the Profile Likelihood test statistic in the presence of nuisance parameters, become

$$q_\mu = -2ln\frac{L(\mu, \hat{\hat{\theta}}_\mu)}{L(\hat{\mu}, \hat{\theta})}. \tag{9}$$

μ is the parameter of interest, θ represent the nuisance parameters (including b). A hat stands for the MLE (Maximum Likelihood Estimator) while a double hat is the constrained MLE, i.e. the MLE of θ, fixing μ. It is common to say that θ is profiled.

3.8 Confidence Interval, Confidence Level and Coverage.

From Wikipedia: A confidence interval (CI) is a type of interval estimate of a population parameter. It is an observed interval (i.e., it is calculated from the observations), in principle different from sample to sample, that frequently includes the value of an unobservable parameter of interest if the experiment is repeated. How frequently the observed interval contains the (true) parameter is determined by the confidence level... Whereas two-sided confidence limits form a confidence interval, their one-sided counterparts are referred to as lower or upper confidence bounds.

Say, the result of a measurement is given by $\mu = 1.1 \pm 0.3$. This means that the Confidence Interval, CI, is $\mu = [0.8, 1.4]$ at the 68% Confidence Level (CL). I.e., in an ensemble of repeated experiments, each producing a CI, 68% of the Confidence Intervals contain the unknown true value of the parameter of interest μ.

There are many ways to derive a CI at a given CL. If, the method produces a CI that contains the true value of the parameter of interest (p.o.i) more than the CL (e.g. in our example, more than 68%), the method is said to over-cover, and is considered conservative. If, however, the CI contains the true value of the p.o.i. less than the claimed Confidence Level, the method is considered to under-cover, which means, one cannot trust the CL, and the true CL might be lower than the claimed one.

3.9 Upper Limits and Confidence Levels.

If one deduces that the CI of μ contains $\mu = 0$, i.e. $\mu = [0, \mu_{up}]$ at the 95% CL, then one says that $\mu < \mu_{up}$ at the 95% CL. This means that in an ensemble of experiments, 95% of the intervals contain the true value of μ including $\mu = 0$.

If $\mu < 1$ at the 95% CL, and μ is given by Eq. 3, i.e.

$$\mu = \frac{\sigma_{obs}(m_H)}{\sigma_{SM}(m_H)} < 1 \tag{10}$$

one concludes that $\sigma_{obs}(m_H) < \sigma_{SM}(m_H)$, i.e. a SM Higgs with a mass m_H is excluded at the 95% CL.

3.10 The Neyman Pearson Lemma.

Wikipedia: In statistics, the Neyman Pearson lemma, named after Jerzy Neyman and Egon Pearson,

states that when performing a hypothesis test between two simple hypotheses H_{null} and H_{alt}, the likelihood-ratio test which rejects H_{null} in favour of H_{alt} is the most powerful test at (a given) significance level...

When we reject the null hypothesis H_{null} based on a very small p-value, we also take a risk. We might be wrong (this is referred to as a type I error, see section 3.11). The null hypothesis can still be true and the p-value is a measure for this risk. The p-value can therefore be interpreted as the false-positive rate and it satisfies

$$p \leq Prob(reject\ H_{null}|H_{null} = TRUE) \tag{11}$$

However, if while rejecting the null hypothesis, the probability for the alternative hypothesis to be true is small.... the test statistic is probably not doing its job, i.e. it is not powerful. The power of a test is therefore related to the probability that $H_{alt} = TRUE$ while rejecting H_{null}, i.e.

$$POWER = Prob(rejectH_{null}|H_{alt} = TRUE). \tag{12}$$

Neyman and Pearson showed [3], that (in the absence of nuisance parameters) the most powerful test statistic is the likelihood ratio defined in Eq. 5.

3.11 Type I & Type II Errors, the Modified Frequentist p-value, or, the CLs Technique.

Wikipedia: CLs (from Confidence Levels) is a statistical method for setting upper limits (also called exclusion limits) on model parameters, a particular form of interval estimation used for parameters that can take only non-negative values..... .....it differs from standard confidence intervals in that the stated confidence level of the interval is not equal to its coverage probability. The reason for this deviation is that standard upper limits based on a most powerful test necessarily produce empty intervals with some fixed probability when the parameter value is zero, and this property is considered undesirable by most physicists and statisticians.

For the sake of clarity let us define now type I and type II errors. Type I error is the probability to reject the null hypothesis, when the null hypothesis is true. This is referred to as "False Positive". It is usually denoted by α, i.e. $\alpha = Prob(rejectH_{null}|H_{null} = TRUE)$. Type II error, referred to as "False Negative", is when we accept the null hypothesis, when the alternative hypothesis is true. It is usually denoted by β. $\beta = Prob(AcceptH_{null}|Hnull = FALSE) = Prob(AcceptH_{null}|H_{alt} = TRUE)$. Quoting Birnbaum [4]: *A concept of statistical evidence is not plausible unless it finds strong evidence for H_{alt} against H_{null}, with small probability α when H_{null} is true, and with much larger probability $(1 - \beta)$) when H_{alt} is true.* $1 - \beta = Prob(rejectH_{null}|H_{alt} = TRUE)$ is defined as the power of the statistical test. Since rejecting H_{null} is accepting H_{alt} by definition, we find

$$POWER = 1 - \beta = Prob(acceptH_{alt}|H_{alt} = TRUE) = 1 - Prob(rejectH_{alt}|H_{alt} = TRUE). \tag{13}$$

Let $H_{null} = H_{s+b}$, i.e. the $s + b$ hypothesis, then, given an observation, H_{s+b} is rejected if the p-value$= p_{s+b} \leq \alpha$. At the threshold we find

$$p_{s+b} = Prob(reject\ H_{s+b}|H_{s+b} = TRUE). \tag{14}$$

with a power (Equation 13) of

$$Power = 1 - p_b. \tag{15}$$

A situation occurs when the power is very small and the experiment has no sensitivity to reject with high power the $s+b$ hypothesis, because it almost rejects the b-only hypothesis as well, as seen in Figure 3. A way out, was suggested by the CL_s technique [5] which is based on Birnbaum [4]. Birnbaum suggested in 1962 that the the $\{p - value\}/\{power\}$ should be used as a measure of the strength of statistical evidence provided by significance tests, rather than the $p - value$ alone. This translates into using a modified $p - value$

$$p'_{s+b} = \frac{p_{s+b}}{1 - p_b} \tag{16}$$

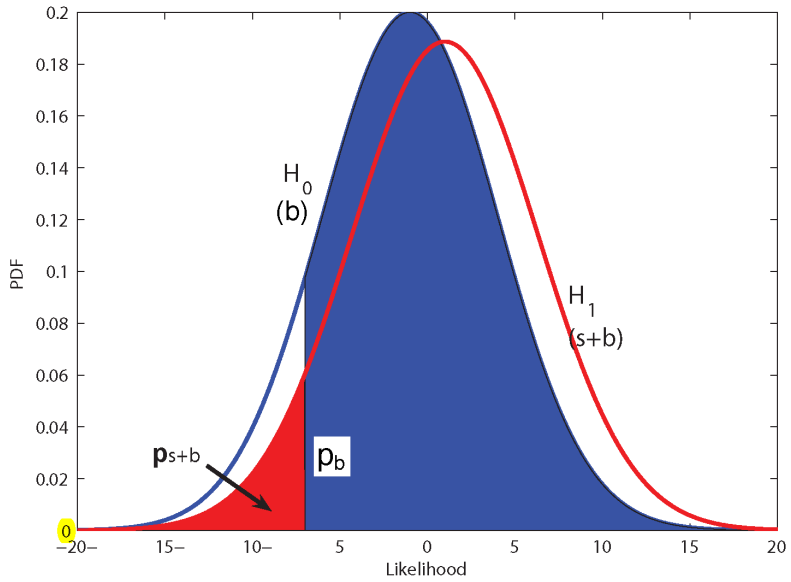

Fig. 3: An illustration showing the reasoning of the CL_s method. In this situation a signal+background hypothesis might be rejected though the experiment has no sensitivity to observe that particular signal.

Equation 16 can also be interpreted as a normalised p-value, where p_{s+b} is normalised to the acceptance probability of H_b. Obviously if, while rejecting H_{s+b} one does not accept H_b , one does not have a sensitivity to exclude the $s + b$ hypothesis.

$$p'_{s+b} = \frac{Prob(reject\ H_{s+b}|H_{s+b} = TRUE)}{Prob(accept\ H_b|H_b = TRUE)} \tag{17}$$

The CL_s method 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 the physicists point of view! Here is what happens: One uses the Neyman-Pearson likelihood ratio as a test statistics. When the expected signal is very low the two pdf are almost overlapping (see Figure 3). The background might fluctuate down resulting in a very small p_{s+b}. As a result we are tempted to exclude the signal hypothesis. However, it is not the signal hypothesis s, that is excluded, but the signal+background hypothesis $s + b$. It is the small expected signal $s << s + b$ that is leading to a false exclusion. To protect against such an inference one uses the modified $p - value$ (Eq. 16) as a criterion for taking a decision of rejecting the signal hypothesis.

As a result, for heavy Higgses with low cross section, where the experiment lacks sensitivity, the false exclusion rate is too low and the method over-covers. This is conservative because it avoids excluding when there is no sensitivity. When the signal cross section is high (light m_H), the coverage is close to full.

3.12 Feldman-Cousins: Ensuring Coverage by Neyman Construction.

Wikipedia: Neyman construction is a frequentist method to construct an interval at a confidence level $CL\%$, that if we repeat the experiment many times the interval will contain the true value a fraction $CL\%$ of the time, this way, one guarantees full coverage by construction.

As said, 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, $f(s_m|s)$, 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 $[s_l, s_h](s) = \{s_m | \int_{s_l}^{s_h} f(s_m|s) ds_m = 68\%\}$ (Figure 4). Even in the simplest case where f is a Gaussian, there is an ambiguity in the choice of the integration boundaries, 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 (i.e. which measurements should be considered within the integration boundaries and which should stay out). The construction of the acceptance intervals for all s forms a belt from which one can easily get the corresponding (e.g.) 68% confidence interval $[s_d, s_u](s_o)$, given one measurement s_o via inversion (Figure 4).

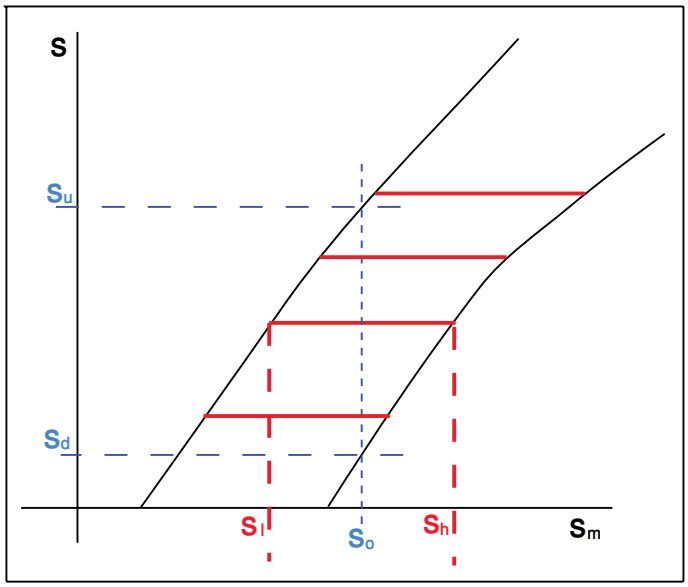

Fig. 4: An illustration showing the Neyman belt. The horizontal lines are the acceptance intervals in the measured parameter space s_m for a given possible true s, $[s_l, s_h](s)$. Given an observation s_o one can construct the confidence interval $[s_d, s_u]$ via inversion, as indicated in the Figure.

3.12.1 The Feldman-Cousins Method

The full Neyman construction was introduced to HEP by Feldman and Cousins [6]. The test statistic is the likelihood ratio $q(s) = \frac{L(s+b)}{L(\hat{s}+b)}$ where $\hat{s}$ is the MLE of s (in $L(\hat{s}+b)$) under the constraint that s is physically allowed (i.e. positive). To construct a 68% acceptance 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 (see Fig. 4). 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...).

One can clearly see in Fig. 4 that depending on the observation, s_o, one gets either a one sided bound, or a two sided interval.

A noted 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 [7] or in the

LHC way, i.e. using the Profile Likelihood [1] as described above.

4 Classification of Test Statistics.

Depending on the nature of the test, one can classify the various test statistics, all based on Likelihood ratios, where the nuisance parameters are profiled (e.g. Eq. 9). The classification is based on [1] and is shown in Table 1.

Table 1: Classification of Test Statistics

Test Stat.	**Purpose**	**Expression**	**LR**
q_0	discovery of positive signal	$q_0 = \begin{cases} -2\ln\lambda(0) & \hat{\mu} \geq 0 \\ 0 & \hat{\mu} < 0 \end{cases}$	$\lambda(0) = \frac{L(0,\hat{\hat{\vec{\theta}}})}{L(\hat{\mu},\hat{\vec{\theta}})}$
t_μ	2-sided measurement	$t_\mu = -2ln\lambda(\mu)$	$\lambda(\mu) = \frac{L(\mu,\hat{\hat{\vec{\theta}}})}{L(\hat{\mu},\hat{\vec{\theta}})}$
$\tilde{t}_\mu$	avoid negative signal (Feldman-Cousins)	$\tilde{t}_\mu = -2ln\tilde{\lambda}(\mu)$	$\tilde{\lambda}(\mu) = \begin{cases} \frac{L(\mu,\hat{\hat{\vec{\theta}}}(\mu))}{L(\hat{\mu},\hat{\vec{\theta}})} & \hat{\mu} \geq 0 \\ \frac{L(\mu,\hat{\hat{\vec{\theta}}}(\mu))}{L(0,\hat{\hat{\vec{\theta}}}(0))} & \hat{\mu} < 0 \end{cases}$
q_μ	exclusion	$q_\mu = \begin{cases} -2\ln\lambda(\mu) & \hat{\mu} \leq \mu \\ 0 & \hat{\mu} > \mu \end{cases}$	
$\tilde{q}_\mu$	exclusion of positive signal	$\tilde{q}_\mu = \begin{cases} -2\ln\frac{L(\mu,\hat{\hat{\vec{\theta}}}(\mu))}{L(0,\hat{\hat{\vec{\theta}}}(0))} & \hat{\mu} < 0\ , \\ -2\ln\frac{L(\mu,\hat{\hat{\vec{\theta}}}(\mu))}{L(\hat{\mu},\hat{\vec{\theta}})} & 0 \leq \hat{\mu} \leq \mu \\ 0 & \hat{\mu} > \mu \end{cases}$	

5 Asymptotic Formulae

Wikipedia: In mathematics and statistics, an asymptotic distribution is a distribution that is in a sense the "limiting" distribution of a sequence of distributions. One of the main uses of the idea of an asymptotic distribution is in providing approximations to the cumulative distribution functions of statistical estimators.

The frequentist approach of statistics requires the knowledge of the probability distribution functions (PDFs) of the test statistic under the null and alternative hypotheses. These PDFs are used to find both the significance for a specific data set and the expected significance. However, obtaining these PDFs, can involve Monte Carlo generations that are computationally expensive. Ref [1] developed the asymptotic formulae based on results due to Wilks [8] and Wald [9] by which one can obtain both the significance for given data as well as the full sampling distribution of the significance under the hypothesis of different signal models, all without recourse to Monte Carlo. In this way one can find, for example, the median significance and also a measure of how much one would expect this to vary as a result of statistical fluctuations in the data. Obtaining the same things with Monte Carlo is sometimes impossible. One LHC collision might take $o(10mins)$ to generate, and one needs over 10^7 events to calculate a 5σ tail of a PDF. Moreover, the test statistics involve heavy duty fits which also take time. Combining ATLAS and CMS results in over 4000 Nuisance Parameters. Repeated fits of that many parameters result

often in failure fits. Some we are not even aware of. It could be that the PDF generated by toys is subject to unknown failure of fits and is not reliable for $p-value$ calculations. In most cases, the number of events involved is satisfying the condition for the asymptotic approximation to work.

All of the asymptotic approximations of the PDFs of the test statistics shown in Table 1 have been calculated under the null and alternative hypotheses [1]. There is no point in reproducing them all here. Three common uses are for exclusion, discovery and measurement.

5.1 Exclusion

For exclusion one can either use q_μ or $\tilde{q}_\mu$ (Table 1) as a test statistic. In numerical examples we have found that the difference between the two tests is negligible, but use of q_μ leads to important simplifications. Furthermore, in the context of the asymptotic approximation, the two statistics are equivalent. That is, assuming the approximations below, q_μ can be expressed as a monotonic function of $\tilde{q}_\mu$ and thus they lead to the same results. We will therefore recommend the use of q_μ for the derivation of exclusion.

Using the asymptoric formulae of [1] we find that $f(q_\mu|\mu)$ distributes as a half-chi-square:

$$f(q_\mu|\mu) = \frac{1}{2}\delta(q_\mu) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_\mu}}e^{-q_\mu/2} \,. \tag{18}$$

It is therefore recommended to verify that $f(q_\mu|\mu) \sim \chi^2_1$. This is usually the case, in particular when combining channels.

The cumulative distribution is

$$F(q_\mu|\mu) = \Phi\left(\sqrt{q_\mu}\right) \,. \tag{19}$$

5.1.1 The $p-value$

The p-value of the hypothesized μ is

$$p_\mu = 1 - F(q_\mu|\mu) = 1 - \Phi\left(\sqrt{q_\mu}\right) \tag{20}$$

and therefore the corresponding significance is

$$Z_\mu = \Phi^{-1}(1-p_\mu) = \sqrt{q_\mu} \,. \tag{21}$$

If the p-value is found below a specified threshold α (often one takes $\alpha = 0.05$), then the value of μ is said to be excluded at a confidence level (CL) of $1-\alpha$. The upper limit on μ is the largest μ with $p_\mu \leq \alpha$. Here this can be obtained simply by setting $p_\mu = \alpha$ and solving for μ. One finds

$$\mu_{\rm up} = \hat{\mu} + \sigma\Phi^{-1}(1-\alpha) \,. \tag{22}$$

For example, $\alpha = 0.05$ gives $\Phi^{-1}(1-\alpha) = 1.64$. Any point μ_0 satisfying $\mu_0 \leq \mu_{\rm up}$ is excluded at the $100(1-\alpha)\%$ Confidence Level. (for $\alpha = 0.05$ the 95% Confidence Interval does not contain $\mu = \mu_0$). Also as noted above, σ depends in general on the hypothesized μ. Thus in practice one may find the upper limit numerically as the value of μ for which $p_\mu = \alpha$.

5.1.2 Expected Limit and Error Bands

To find the expected limit, one should plug in the Asimov data which represents the alternative hypothesis, which in this case is the expected background (with no fluctuations). The signal strength is set to zero (in a simple counting experiment $n = b$). One then gets $q_{\mu,A}$ and the corresponding μ^{med}_{up} is given by solving $q_{\mu^{med}_{up},A} = 1.64^2$ (for $\alpha = 0.05$). The error bands are given by

$$\mu_{up+N} = \sigma(\Phi^{-1}(1-\alpha) + N) \tag{23}$$

with

$$\sigma^2 = \frac{\mu^2}{q_{\mu,A}} \tag{24}$$

μ can be taken as μ_{up}^{med} in the calculation of σ.

5.2 Expected Limit and Error Bands a-la "(CL_s)"

To avoid setting limits when the experiment is not sensitive to the signal, one might use the modified p-value defined above, "p'_{s+b}"

$$p'_{s+b} = \frac{p_{s+b}}{1-p_b} \tag{25}$$

We find

$$p'_{\mu} = \frac{1-\Phi(\sqrt{q_\mu})}{\Phi(\sqrt{q_{\mu,A}} - \sqrt{q_\mu})} \tag{26}$$

The median and expected error bands will therefore be

$$\mu_{up+N} = \sigma(\Phi^{-1}(1-\alpha\Phi(N)) + N) \tag{27}$$

with

$$\sigma^2 = \frac{\mu^2}{q_{\mu,A}} \tag{28}$$

To get the 95% expected upper limit, set $\alpha = 0.05$. μ can be taken as μ_{up}^{med} in the calculation of σ..
Note that for $N = 0$ we find the median limit

$$\mu_{up}^{med} = \sigma\Phi^{-1}(1-0.5\alpha) \tag{29}$$

The expected μ and the expectation for error band N is shown in Figure 5. one can clearly see the shrinkage of the error band, $\mu_{up+N\sigma} - \mu_{up+(N-1)\sigma}$, when $N \to -\infty$

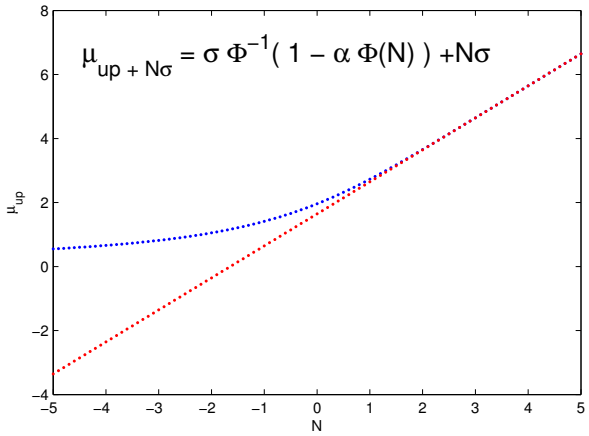

Fig. 5: $\mu_{up+N\sigma}$ as a function of N (in units of σ). Red is based on p_{s+b} blue is based on p'_{s+b} (CL_s).

5.3 Example from the Higgs Boson Search

Figure 6 taken from [10] shows μ_{up} as a function of m_H at one of the stages of the Higgs search. The

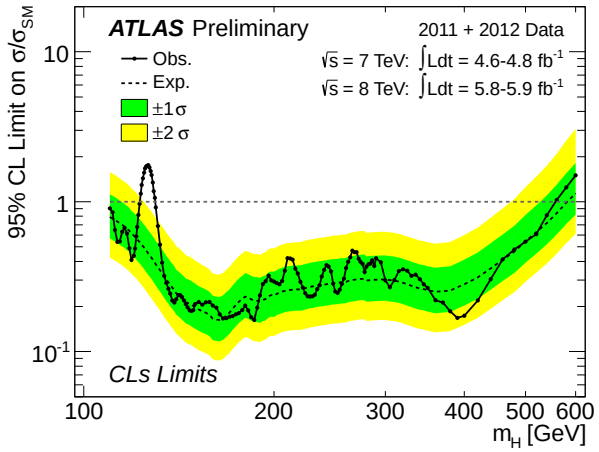

Fig. 6: The observed (full line) and expected (dashed line) 95% CL combined upper limits on the SM Higgs boson signal strength (μ_{up}) in the full mass range m_H considered in this analysis. The dashed curves show the median expected limit in the absence of a signal and the green and yellow bands indicate the corresponding 68% and 95% intervals.

mass range where $\mu_{up}(m_H) \leq 1$ is where a SM Higgs Boson with a mass m_H is excluded. Obviously one cannot exclude the Higgs around $m_H = 125$ GeV, where a real signal is being built up with luminosity $\mu_{up} > 1$. The median expected is given by the dashed line (following Equation 29 with $\alpha = 0.05$). The error bands are derived using Equation 27, with $N = \pm 1$ (Green) and $N = \pm 2$ (yellow).

Figure 7 taken from the same reference, shows p'_{s+b} (labeled in the Figure as CL_s), as a function of m_H. Mass regions where $p'_{s+b} \leq 0.05$ are excluded at, at least, the 95% CL.

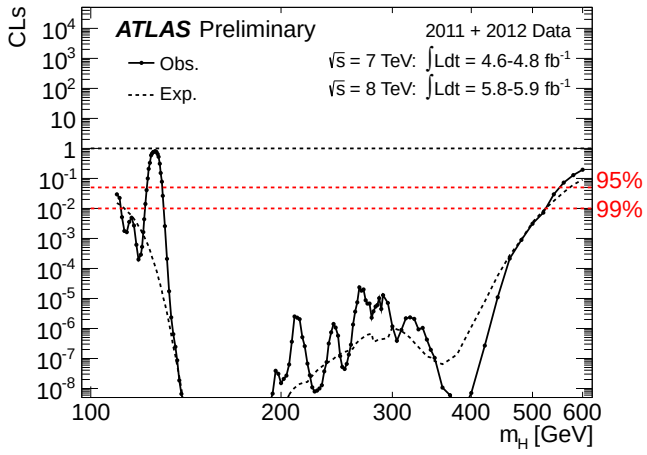

Fig. 7: The value of the combined CL_s (p'_{s+b}), testing the Standard Model Higgs boson hypothesis, as a function of m_H in the full mass range of this analysis. The expected CL_s is shown in the dashed curves. The regions with $CL_s < 0.05$ are excluded at least at 95% CL. The 95% and 99% CL values are indicated as dashed horizontal lines.

5.4 Measurement

Let the statistic be $t_\mu = -2\ln\lambda(\mu)$ (Table 1) as the basis of the statistical test of a hypothesized value of μ. This could be a test of $\mu = 0$ for purposes of establishing existence of a signal process, or non-zero values of μ for purposes of obtaining a confidence interval. In the asymptotic regime the pdf of t_μ distributes like a χ^2 with one degree of freedom, under the H_μ hypothesis.

$$f(t_\mu|\mu) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{t_\mu}}e^{-t_\mu/2} \; . \tag{30}$$

To measure μ, one scans the test statistics, finds $\hat{\mu}$ and σ^{up}, σ^{lo} by substituting $t_\mu = 1$. The 68% Confidence Interval of μ is then estimated to be $[\hat{\mu} - \sigma^{lo}, \hat{\mu} + \sigma^{up}]$. If one wants to estimate with how many standard deviations a specific value of μ, e.g. $\mu = 0.$ is unlikely, one calculates $\sqrt{t_0}$.

To get the expcted μ one repeats the above procedure, calculating t_μ with the Asimov data set, for which $\hat{\mu} = \mu$.

A formulation of the asymptotic properties of t_μ is given in [1].

5.5 Discovery

To establish a discovery one tries to reject the background only hypothesis. We use the q_o test statistics (Table 1). Since we do not want downward fluctuations of the background to serve as an evidence against the background we define the test statistics such that $q_0 = 0$ if $\hat{\mu} < 0$. The test statistic is therefore given by (Table 1):

$$q_0 = \begin{cases} -2\ln\frac{L(0)}{L(\hat{\mu})} & \hat{\mu} \geq 0, \\ 0 & \hat{\mu} < 0 \; , \end{cases} \tag{31}$$

Under the background only hypothesis, H_0, q_0 is asymptotically distributed as half a chi squared with one degree of freedom, i.e.

$$f(q_0|0) = \frac{1}{2}\delta(q_0) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_0}}e^{-q_0/2} \; . \tag{32}$$

The significance of the observation is given by

$$Z_0 = \Phi^{-1}(1 - p_0) = \sqrt{q_0} \; . \tag{33}$$

The p_0 value can easily be calculated using

$$p_0 = 1 - F(q_0|0) \; , \tag{34}$$

where

$$F(q_0|0) = \Phi\Big(\sqrt{q_0}\Big) \; . \tag{35}$$

A significance of 3σ is considered as an observation, while a significance exceeding 5σ is regarded as a discovery. The reason for using such a large number to establish a discovery is because of the Look Elsewhere Effect, discussed in section 7.

5.6 Discovery Example

In Figure 10 we show the $p - value$ as a function of the mass, taken from the ATLAS discovery conference note [10]. Both, the $p - value$ and its corresponding significance are indicated. One clearly sees an

upward fluctuation of the background (downward fluctuation in $p-value$) around a mass of 125 GeV. The fluctuation is at the level of 5σ. For other masses the $p-value$ fluctuates around 0.5, meaning a significance of 0σ. The expected $p-value$ is given by the dashed line. One can clearly see that only around $m_H = 125$ GeV, the expected and the observed $p-value$ are similar, indicating a signal strength $\mu \sim 1$, as can clearly be seen in Figure 11.

5.6.1 Significance in a nut-shell.

Many people use a thumbnail formula $Z = \frac{s}{\sqrt{b}}$ to estimated the significance of an apparent signal. s represents here $n-b$, where b is the expected background, and n is the number of observed events.

Using the profile likelihood formalism we can get a much more accurate estimation for the apparent observed significance [1].

If we regard b as known, the data consist only of n and thus the likelihood function is

$$L(\mu) = \frac{(\mu s + b)^n}{n!} e^{-(\mu s + b)} \ , \tag{36}$$

The test statistic for discovery q_0 can be written

$$q_0 = \begin{cases} -2 \ln \frac{L(0)}{L(\hat{\mu})} & \hat{\mu} \geq 0, \\ 0 & \hat{\mu} < 0 \ , \end{cases} \tag{37}$$

where $\hat{\mu} = n - b$. For sufficiently large b we can use the asymptotic formula [1] to obtain

$$Z_0 = \sqrt{q_0} = \begin{cases} \sqrt{2\left(n \ln \frac{n}{b} + b - n\right)} & \hat{\mu} \geq 0, \\ 0 & \hat{\mu} < 0. \end{cases} \tag{38}$$

To approximate the median significance assuming the nominal signal hypothesis ($\mu = 1$) we replace n by the Asimov value $s + b$ to obtain

$$\text{med}[Z_0|1] = \sqrt{q_{0,\text{A}}} = \sqrt{2\left((s+b)\ln(1+s/b) - s\right)} \ . \tag{39}$$

Expanding the logarithm in s/b one finds

$$\text{med}[Z_0|1] = \frac{s}{\sqrt{b}} \left(1 + \mathcal{O}(s/b)\right) \ . \tag{40}$$

Although $Z_0 \approx s/\sqrt{b}$ has been widely used for cases where $s + b$ is large, one sees here that this final approximation is strictly valid only for $s \ll b$. We therefore recommend to use Eq. 39 to estimate a significance in a nut shell. It is much more accurate.

6 Testing an hypothesis with boundaries.

In [6] Feldman and Cousins derive the test statistics with the physical condition, namely, the true value of μ must be positive, i.e. $\mu > 0$. In [1] the $\tilde{t}_\mu$ test statistic is introduced (see Table 1) in order to avoid a negative non-physical signal. As a result, depends on the observation, a two sided (measurement) or one sided (limit) Confidence Interval is obtained. This is the equivalence of the Feldman-Cousins test statistic with the advantage of taking care of the nuisance parameters. The original Feldman-Cousins test statistic is not considering systematics. In [1] the asymptotic formula of $\tilde{t}_\mu$ is derived. In a later paper [11] the same authors improve the test statistic by taking into account two sided boundaries. This

is the case, for example when one wants to measure or set limits on the measurement of a Branching Ratio, which must be $0 < BR < 1$ by definition. The revised $\tilde{t}_\mu$ is defined by

$$\tilde{t}_\mu = \begin{cases} -2\ln \frac{L(\mu,\hat{\hat{\theta}}(\mu))}{L(\mu_-,\hat{\hat{\theta}}(\mu_-))} & \hat{\mu} \leq \mu_- \\ -2\ln \frac{L(\mu,\hat{\hat{\theta}}(\mu))}{L(\hat{\mu},\hat{\theta})} & \mu_- < \hat{\mu} < \mu_+ \\ -2\ln \frac{L(\mu,\hat{\hat{\theta}}(\mu))}{L(\mu_+,\hat{\hat{\theta}}(\mu_+))} & \hat{\mu} \geq \mu_+ \ , \end{cases} \tag{41}$$

$\hat{\hat{\theta}}$ represent the nuisance parameters, $\hat{\hat{\theta}}(\mu)$ is the conditional maximum likelihood estimate of θ given μ. μ_- and μ_+ are the physical boundaries. The Feldman-Cousins test statistic is retrieved for $\mu_- = 0$ and no upper boundary, μ_+. The asymptotic formulas are derived in [11].

6.1 Pull

The pull of a nuisance parameter θ, with an expectation θ_0 is defined as:

$$pull(\theta) = \frac{\hat{\theta} - \theta_0}{\sigma_\theta} \tag{42}$$

the pull quantifies how far from its expected value we had to "pull" the parameter while finding the MLE. A healthy situation is when the pull average is zero with a standard deviation close to 1, if this is not the case, further investigation is required. The expected value of a nuisance parameter and its assumed standard deviation will be based on an auxiliary measurement or MC studies.

6.2 Impact

the impact of a nuisance parameter is defined as:

$$impact(\theta) = \Delta\mu^{\pm} = \hat{\hat{\mu}}_{\theta_0 \pm \sigma_\theta} - \hat{\mu} \tag{43}$$

where $\hat{\hat{\mu}}_{\theta_0 \pm \sigma}$ is the MLE of μ when we profile every parameter except θ, and set the value of θ to its expectation value plus or minus one standard deviation. The impact gives a measure of how much our parameter of interest varies as we change the nuisance parameter. Obviously not all nuisance parameters are equally important, so a nuisance parameter with low impact may be possibly discarded (or "pruned") to simplify the fit procedure.

6.3 Example of pull and impact

To illustrate the use of impact and pull, consider a simple counting experiment which measures n events, with $n = \mu \cdot \mathrm{s} \cdot \mathrm{A} \cdot \epsilon + b$, where s is the number of signal events, μ is the p.o.i and A (acceptance) ϵ (efficiency) and b (background) are nuisance parameters with gaussian distributions.
The likelihood is given by:

$$L(\mu, A, \epsilon, b) = \frac{(\mu s A \epsilon + b)^n}{n!} exp\left(-(\mu s A \epsilon + b)\right) exp\left(-\frac{(b - b_{obs})^2}{\sigma_b}\right) exp\left(-\frac{(A - A_{obs})^2}{\sigma_A}\right) exp\left(-\frac{(\epsilon - \epsilon_{obs})^2}{\sigma_\epsilon}\right) \tag{44}$$

For each nuisance parameter, there is an "observed" value which could come from some auxiliary measurement. In this simplified case all nuisance parameters are measured by their MLEs, i.e. ($\hat{\theta} = \theta_{obs}$). We assume the "true" value of the parameters are known to be θ_0.
The pulls are calculated straightforward from equation 42. The impact is calculated with the test statistic $t_\mu(\epsilon) = -2ln\frac{L(\hat{\hat{\mu}},\hat{\hat{A}},\epsilon,\hat{\hat{b}})}{L(\hat{\mu},\hat{A},\hat{\epsilon},\hat{b})}$ (for the nuisance parameter ϵ), with double hat indicating that the fit is constrained to ϵ, as was described above. Table 2 shows the values of the parameters used in the toy

calculation. The measured value for n, was picked from a poisson distribution with expectation value of $n_{exp} = \mu \cdot \mathrm{s} \cdot \mathrm{A} \cdot \epsilon + b$ (the true, Asimov, values) and ϵ_{obs} A_{obs} and b_{obs} were picked from gaussian distributions.

Figure 8 shows a typical overlay plot of pull and impact (right plot for Asimov and left plot for some toy data set). Note the different x-axis (top for the impact, bottom for the pull). Figure 9 shows in more detail the calculation of the impact - it shows the scan of $t_\mu(\epsilon)$, $\hat{\hat{\mu}}(\epsilon)$ and the procedure leading from $\hat{\epsilon} \pm \sigma_\epsilon$ points to the Impact range (right plot for Asimov and left plot for some toy data set).

Parameter	Asimov	Measured
s	90	-
n	131.5	132
μ	1	1.4
ϵ	0.5	0.465
σ_ϵ	0.05	-
A	0.7	0.487
σ_A	0.2	-
b	100	103.21
σ_b	10	-

Table 2: Parameters for toy experiment

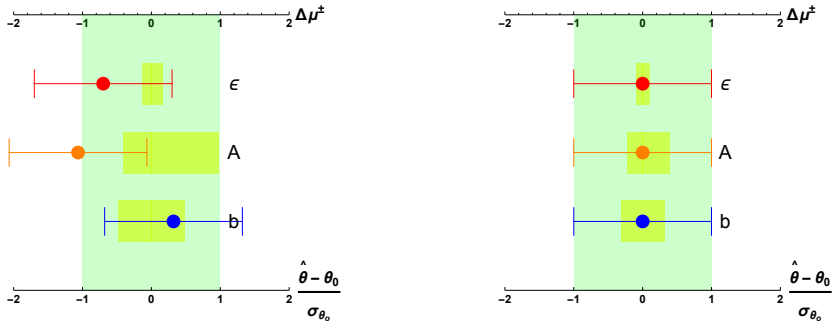

Fig. 8: Impact and pull for the three nuisance parameters (right plot for Asimov and left plot for some toy data set). The yellow rectangles show the impact range (upper x-axis) and the coloured dots show the pull (lower x-axis) with one σ error bars

7 The Look Elsewhere Effect (LEE).

Wikipedia: The look-elsewhere effect is a phenomenon in the statistical analysis of scientific experiments, particularly in complex particle physics experiments, where an apparently statistically significant observation may have actually arisen by chance because of the size of the parameter space to be searched.Once the possibility of look-elsewhere error in an analysis is acknowledged, it can be compensated for by careful application of standard mathematical techniques [2].

7.1 The LEE with one parameter (m) undefined under the null hypothesis.

When searching for a new resonance somewhere in a possible mass range, the significance of observing a local excess of events must take into account the probability of observing such an excess *anywhere* in

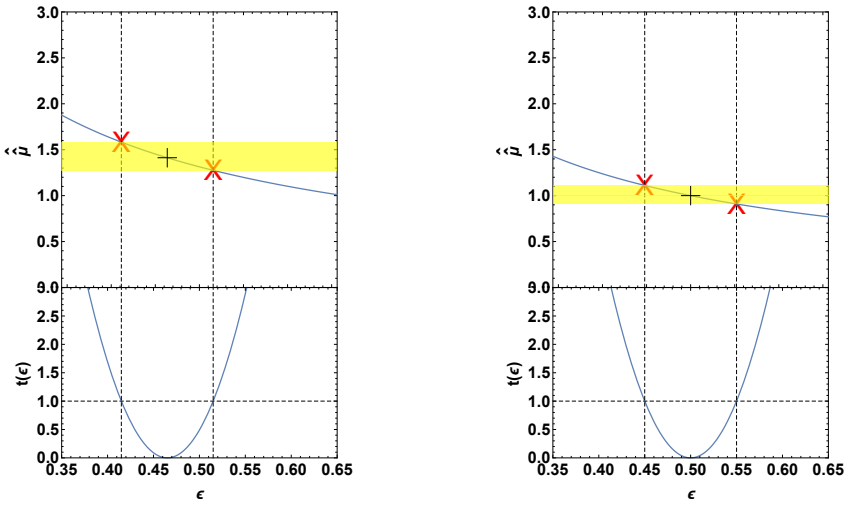

Fig. 9: Calculation of the impact of the nuisance parameter ϵ (right plot for Asimov and left plot for some toy data set). The upper plot shows the MLE of μ when profiling all parameters except ϵ (the blue curve) and the red X's show the point where $\hat{\hat{\mu}}(\epsilon)$ intersects with the $\hat{\epsilon} \pm \sigma_\epsilon$ points (the dashed vertical lines), which marks the end points of the impact. The bottom plot shows the scan of the test statistic $t_\mu(\epsilon) = -2ln\frac{L(\hat{\hat{\mu}},\hat{\hat{A}},\epsilon,\hat{\hat{b}})}{L(\hat{\mu},\hat{A},\hat{\epsilon},\hat{b})}$ and shows that the $\hat{\epsilon} \pm \sigma_\epsilon$ points correspond to $min(t_\mu(\epsilon)) \pm 1$

the range. This is the so called "look elsewhere effect". The effect can be quantified in terms of a trial factor, which is the ratio between the probability of observing the excess at some fixed mass point (local $p-value$), to the probability of observing it anywhere in the range (global $p-value$). The question we try to answer with a $p-value$ is *What is the probability of observing an excess anywhere in the search range"*. For years it was a common knowledge that in order to convert the local probability into a global probability one has to apply a trial factor which is simply the number of possible independent search regions, i.e. $trial\ \# = \frac{p_{float}}{p_{fix}} = \frac{search\ range}{mass\ resolution}$. In [2] it was shown that an important factor was missing from this rule of thumb estimation. The trial number is linearly dependent on the local significance. This can be intuitively understood by the possibility of havung a look elsewhere effect within the independent search range, where the number of possibilities peak can arrange itself is proportional to the significance. The trial number is therefore asymptotically (for small $p-values$, i.e. large significance) given by

$$trial\# \approx 1 + \sqrt{\frac{\pi}{2}}\mathcal{N}Z_{fix} \tag{45}$$

where $\mathcal{N}$ is the number of independent search regions.

The trial factor is thus asymptotically linear with both the effective number of independent regions, and with the fixed-mass significance.

The number of independent search region is not a trivial quantity. The resolution might not be well defined and is usually depending on the mass. We applied the formula obtained by Davies [12] for an hypothesis testing when a nuisance parameter (the mass) is known only under the alternative hypothesis. The mass is not defined under the null (background only) hypothesis.

Let $q_0(m,\theta)$ be the discovery test statistics (following Equation 31). m is undefined under the null hypothesis ($\mu = 0$). Nevertheless, there is a dependence of q_0 on the mass through the denominator.

$$q_0(m) = \begin{cases} -2\ln\frac{L(0)}{L(\hat{\mu},m)} & \hat{\mu} \geq 0, \\ 0 & \hat{\mu} < 0\ , \end{cases} \tag{46}$$

Given some data set, we scan $q_0(m)$ and find the maximal one (smallest $p-value$ over all possible masses). We define it as

$$\hat{q}_0 \equiv max_m[q_0(m)] = q_0(\hat{m}) \tag{47}$$

Since for any given m, $q_0(\hat{m} \geq q_0(m)$, the global $p-value$, $p_{global} \geq p_{local}$. Hence, the trial number is always greater or equal to one, $Trial\# \geq 1$. We find that for high local significance (at the tail of the pdf distributions), the following relation exists between the global and local $p-value$:

$$P(q_0(\hat{m}) > u) \approx \frac{1}{2}P(\chi^2_1 > u) + \mathcal{N}P(\chi^2_2 > u) \tag{48}$$

where in the tail $u \to \infty$. $\mathcal{N}$ is the number of independent search regions. To obtain this we find the average number of upcrossings at a level $u = Z^2$, n_u, i.e. $E[n_u] = \mathcal{N}e^{-u/2}$.

Since we are interested to know the global significance for high level , normally $u = Z^2 > 16$, the number of upcrossings is very small and one needs to generate expensive toys to estimate $E[n_u]$. One then renormalize the upcrossings level. Let us pick a low level u_0 where the number of upcrossings is relatively large and the statistical error on the estimation is therefore small (normally one picks $u_0 = 0$ or $u_0 = 0.5^2$). We find $E(n_{u_0}) = \mathcal{N}e^{-u_0/2}$ and therefore

$$E(n_u) = E(n_{u_0})e^{\frac{u_0-u}{2}} \tag{49}$$

Finally we find that the answer to the question: *What is the probability to have a fluctuation with a significance bigger than $Z = \sqrt{u}$ all over a given mass range?* is given by

$$P_{global}(u) \approx p_{local}(u) + E(n_{u_0})e^{\frac{u_0-u}{2}} \tag{50}$$

where u_0 is some low reference level, where the estimation of the number of upcrossings $E(n_{u_0})$ is easy and fast.

To illustrate it let us look at a real example from the Higgs Boson search and discovery. In the following Figures we show the p_0 (Figure 10) and the signal strength μ (Figure 11) as a function of the Higgs mass. The plots are taken from the ATLAS discovery conference note [10]. $Z = 0$ corresponds

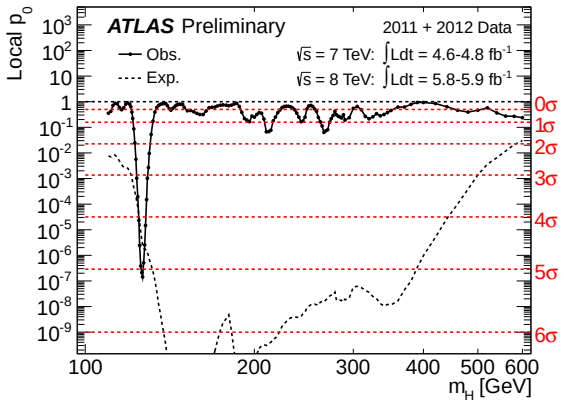

Fig. 10: The local probability p_0 for a background-only experiment to be more signal-like than the observation in the full mass range of this analysis as a function of m_H. The dashed curves show the median expected local p_0 under the hypothesis of a Standard Model Higgs boson production signal at that mass. The horizontal dashed lines indicate the p-values corresponding to significances of 1σ to 6σ.

to either $p_0 = 0.5$ or $\hat{\mu} = 0$. Se we have to count the number of up-crossings at 0σ. We should have performed a few Momte Carlo experiments and count the average number of up-crossings at $u = 0$. But

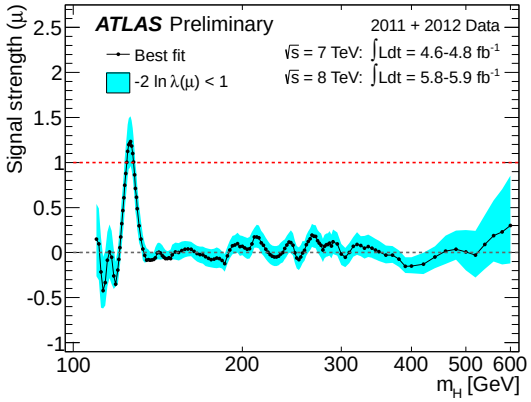

Fig. 11: The combined best-fit signal strength $\hat{\mu}$ as a function of the Higgs boson mass hypothesis in the full mass range of this analysis

this seems to be not practical when we combine all the channels. Instead we could simply take the data itself and count $n_{u_0} = 9 \pm 3$. This accuracy is sufficient for the estimation of the trial number. Following Equation 50, substituting $u_0 = 0$ and $u = 5^2 = 25$, we find

$$p_{global} = O(10^{-7}) + 9 \times e^{-25/2} = 3.3 \times 10^{-5} \tag{51}$$

The trial number is about $trial\# \approx \frac{10^{-5}}{10^{-7}} \approx 100$ and it reduces the significance from 5σ to 4σ.

7.2 The LEE with two parameters (m, Γ) undefined under the null hypothesis.

In cases where there are two parameters undefined under the null hypothesis, such as mass (m) and width (Γ) the Look Elsewhere Effect is broader. Ref [13] solved the case for a multi-dimensional search.

Suppose we would like to estimate the global significance of some observed excess. When allowing both the mass and the width float, we observe that the highest significance of $Z\sigma$ occurs for some specific mass and width. This observation corresponds to a local background fluctuation with a p-value of p_{local}. However, any fluctuation at any mass and width in the 2D search plane of m and Γ would have drawn our attention. The increased probability to observe a fluctuation of $Z\sigma$ or more anywhere in the mass-width plane $A = (m, \Gamma)$ (LEE) is given by the global p-value, p_{global}. The local p-value is based on scanning the $q_0(m, \Gamma)$ test statistic, $q_0(m, \Gamma)$ given by

$$q_0(m, \Gamma) = -2 \log \frac{L(0, m, \Gamma, \hat{\hat{\theta}})}{L(\hat{\mu}, \hat{m}, \hat{\Gamma}, \hat{\theta})}. \tag{52}$$

The distribution of the maximum local significance $u = Z^2 = \max_{m,\Gamma} q_0(m, \Gamma)$ was studied in [13]. The global p-value is given by

$$p_{global} \approx E[\phi(A_u)] = p_{local} + e^{-u/2}(N_1 + \sqrt{u} N_2) \tag{53}$$

where N_1 and N_2 are coefficients that are estimated by calculating the average Euler characteristic of the plane A. To solve for N_1 and N_2, it is convenient to set two reference levels u_0 and u_1, find the Euler characteristics for each level, and solve the consequent system of two linear equations. In a 2D manifold with closed islands, some with holes, each disconnected full island takes the value $+1$. Each hole contributes -1. In that sense a full round shape has the Euler characteristic of $+1$. If you dig a hole in it, its Euler characteristics becomes $+1 - 1 = 0$ (Figure 12).

An example can be taken from the search for di-photon in ATLAS [14]. In Figure 13 one sees the 2D $(m_X, \Gamma_X/m_X)$ plane. The manifold A_u is obtained by slicing this plane at a level $u = Z^2$. The Euler characteristic is the number of "disconnected" islands in that slice.

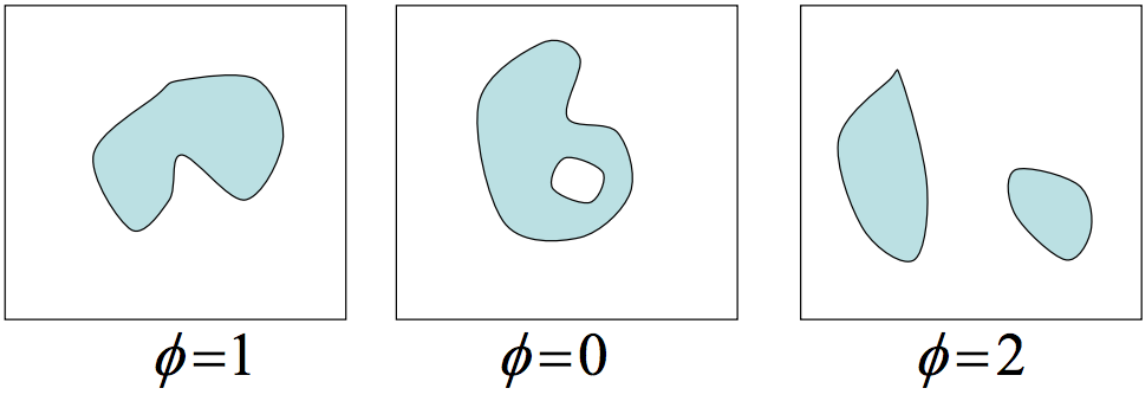

Fig. 12: Illustration of the Euler characteristic of some 2-dimensional manifold.

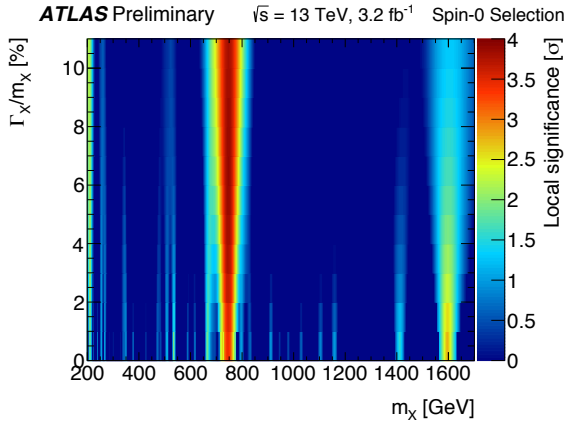

Fig. 13: The 2D $(m_X, \Gamma_X/m_X)$ plane. The colors are the significance Z, where the level u is given by $u = Z^2$

Acknowledgements

The author would like to thank Jonathan Shlomi for helping to produce some of the Figures shown and for his careful reading of the manuscript. His comments were very useful. The author would also like to thank the organisers of ESPHEP, Nick Ellis and Martijn Mulders for the wonderful opportunity to prepare these lectures.

References

[1] G. Cowan, K. Cranmer, E. Gross and O. Vitells, Eur. Phys. J. C **71** (2011) 1554 [Eur. Phys. J. C **73** (2013) 2501] doi:10.1140/epjc/s10052-011-1554-0, 10.1140/epjc/s10052-013-2501-z [arXiv:1007.1727 [physics.data-an]].

[2] E. Gross and O. Vitells, Eur. Phys. J. C **70**, 525 (2010) doi:10.1140/epjc/s10052-010-1470-8 [arXiv:1005.1891 [physics.data-an]].

[3] Neyman, Jerzy; Pearson, Egon S. (1933). "On the Problem of the Most Efficient Tests of Statistical Hypotheses". Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 231 (694Ð706): 289Ð337. Bibcode:1933RSPTA.231..289N. doi:10.1098/rsta.1933.0009. JSTOR 91247.

[4] Birnbaum, Allan (1962). "On the foundations of statistical inference". Journal of the American Statistical Association 57 (298): 269Ð326. doi:10.2307/2281640. JSTOR 2281640. MR 0138176.

[5] 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**

[6] G. J. Feldman and R. D. Cousins, Phys. Rev. D **57** (1998) 3873 doi:10.1103/PhysRevD.57.3873 [physics/9711021 [physics.data-an]].

[7] R. D. Cousins and V. L. Highland, "Incorporating systematic uncertainties into an upper limit," Nuclear Instruments and Methods A.320 (1992) 331-335

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

[9] 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.

[10] [ATLAS Collaboration], "Observation of an Excess of Events in the Search for the Standard Model Higgs boson with the ATLAS detector at the LHC," ATLAS-CONF-2012-093.

[11] G. Cowan, K. Cranmer, E. Gross and O. Vitells, "Asymptotic distribution for two-sided tests with lower and upper boundaries on the parameter of interest," arXiv:1210.6948 [physics.data-an].

[12] R. B. Davies, *Hypothesis testing when a nuisance parameter is present only under the alternative*, Biometrika **74** (1987), 33-43.

[13] O. Vitells and E. Gross, "Estimating the significance of a signal in a multi-dimensional search," Astropart. Phys. **35**, 230 (2011) doi:10.1016/j.astropartphys.2011.08.005 [arXiv:1105.4355 [astro-ph.IM]].

[14] The ATLAS collaboration, "Search for resonances in diphoton events with the ATLAS detector at $\sqrt{s}$ = 13 TeV," ATLAS-CONF-2016-018.