FERMILAB-PUB-16-732-E
Accepted for Publication
This paper appears as an appendix in the book
"Discovery of the Higgs Boson," World Scientific, 2016
Editors: A. Nisati and V. Sharma.
ISBN 9814425877

Appendix B

Statistical Methods

Thomas R. Junk[1], Andrey Korytov[2], and Alexander L. Read[3]
[1] *Fermilab, Batavia, Illinois, 60510, USA*
[2] *University of Florida, Gainesville, FL 32611 USA*
[3] *University of Oslo, Postboks 1048 Blindern, 0316 Oslo, Norway*

B.1. Introduction

This Appendix summarises the statistical methods which are used to set limits on possible signals (Sec. B.2), to compute the significances of observed excesses of events (Sec. B.3), and to measure signal model parameters (Sec. B.4). While these techniques are by now standard, having been put to use in a broad variety of scientific and other applications over the last century, there is always more than one valid technique which can be used to interpret experimental data in order to produce the results. Often, even within a single publication, a variety of techniques may be chosen in order to optimise the sensitivity or to best incorporate the effects of systematic uncertainty on the model predictions. For a comprehensive overview, see, for example, Refs. [1–5].

Henceforth, the expected event yields for the nominal signal (the Standard Model Higgs boson in this book) will be generically denoted as s, and the background predictions as b. Depending on the context, these will stand for event counts in one or multiple bins (e.g., s_i) or for unbinned probability density functions of some observables, whichever approach is used in an analysis. The notations b and $s+b$ will be also used to represent symbolically the background-only and the signal+background hypotheses, respectively; $\mu s+b$ will stand for a signal+background hypothesis, in which all nominal signal event yields are scaled by μ. Predictions for the signal and the background yields are subject to multiple uncertainties that are handled by introducing a set of nuisance parameters $\theta=(\theta_1, \ldots \theta_n)$, so that

signal and background expectations become functions of the nuisance parameters: $s(\boldsymbol{\theta})$ and $b(\boldsymbol{\theta})$. Nuisance parameters may affect either the overall event rates, the shapes of distributions, or both. A set of observed events collected by the experimental apparatus, again in a binned or unbinned format, is referred to as the "data" or the "observation". The term "pseudo-data" refers to simulated experimental outcomes, or pseudo-observations.

B.2. Limits

The Bayesian method and the classical frequentist method, the latter with a number of modifications, are two statistical approaches commonly used in high energy physics for characterising the results of a search for a possible signal. In the absence of an observed signal, the upper limit on the signal strength is the primary result of the search.

The frequentist methods set limits that are characterised by a confidence level (CL). A statement that a signal of strength μ is excluded at $1-\beta$ CL (where β is some small number, usually set at 5%) is expected to imply that if a signal is truly present at the quoted signal strength, then in repetitions of the experiment, a fraction of at most β of them will falsely exclude it. A limit-setting procedure that satisfies this requirement for the error rate is said to have proper coverage. A limit-setting procedure with a larger error rate than stated is said to undercover, and a test with a smaller error rate is said to overcover. The false exclusion of a signal in its presence is known as a Type II error.*

Bayesian results are characterised by a credibility level, which is also abbreviated CL. At a CL of $1-\beta$, the integral of the "belief" probability density function of the signal event rate over values greater than the limit is β. No claim is made regarding the coverage of Bayesian methods, although in practice they tend to overcover when flat priors on signal strength are used (see Sec. B.2.1).

Limits can and should be set even in the case that an excess of events is observed. Doing so is a condition for proper coverage and eliminates the flip-flop hazard of quoting limits only when no signal is observed. Limits quoted when an excess is observed also have important physical interpretations as they exclude signal strengths stronger than observed, which may test interesting models which predict anomalously large cross sections, branching ratios, or may have kinematic properties that enhance the signal

*A Type I error, to be discussed in Sec. B.3, refers to a false claim of a signal that is not actually present.

acceptance.

In addition to reporting the exclusion confidence level for a fully specified model (e.g., the Standard Model (SM) Higgs boson of a given m_H), null results of a search targeting a specific signal production mechanism and a particular decay mode can be reported as approximately model-independent limits on the signal cross section times the branching ratio ($\sigma \times \mathrm{BR}$) for the decay mode targeted by the analysis. Less model dependence is induced by setting limits on the cross section times the branching ratio times the experimental acceptance ($\sigma \times \mathrm{BR} \times \mathcal{A}$). However, neither is perfect. The former explicitly depends on assumptions made on the fraction of a signal cross section in the phase space not covered by an experiment. The latter does not introduce such dependencies, but, in order to allow theorists to calculate the signal cross section within experimental acceptance $\mathcal{A}$, one has to provide a model of that acceptance, the exact definition of which may be too complicated for a practical use. If an analysis is based on the distribution of a discriminating observable then, in any case its results (limits) can be interpreted only in models that yield the same shape as the signal for which they were derived. Therefore, extrapolating results of an analysis to make a statement about a signal that has different kinematic properties from the one assumed in a given analysis is not trivial and, in general, requires additional dedicated studies.

In a combination of multiple analyses sensitive to different signal production mechanisms and different decay modes, presenting results in a form of limits on $\sigma \times \mathrm{BR}$ or $\sigma \times \mathrm{BR} \times \mathcal{A}$ is impossible. The customary alternative is to set limits on a common signal strength modifier μ that is taken to change event yields in each (production)$\times$(decay) mode by exactly the same scale. The Standard Model Higgs boson is said to be excluded at, say, 95%CL, when the 95% CL limit on μ becomes smaller than one. In the next sub-sections, we will follow this convention and discuss limits on the common signal strength modifier μ.

B.2.1. *Bayesian approach*

In the Bayesian approach, a degree of belief is assigned to each value of μ, as a probability density function for μ. Bayes's theorem is then invoked to calculate the impact of the experimental data to update the prior probability density $\pi(\mu)$ to obtain the posterior probability density $L(\mu)$:

$$L(\mu) = \frac{1}{C} \int_{\boldsymbol{\theta}} p(\mathrm{data} \,|\, \mu\, s(\boldsymbol{\theta}) + b(\boldsymbol{\theta}))\ \rho(\boldsymbol{\theta})\ \pi(\mu)\ d\boldsymbol{\theta}, \tag{B.1}$$

where $p(\text{data}\,|\,\mu\,s+b)$ is the probability to observe the data as seen in an experiment assuming the $\mu\,s+b$ hypothesis. The function $\rho(\boldsymbol{\theta})$ is the density function describing our prior belief in the values of the nuisance parameters which affect the predicted signal and background event yields and distributions, and is typically a product of prior densities for each of the θ_i. Popular functional choices for individual nuisance parameter prior densities $\rho(\theta_i)$ are: Gaussian (often truncated so that all signal and background predictions are non-negative), log-normal, Gamma, or flat (either constrained, a so-called box distribution, or not). The function $\pi(\mu)$ is the prior probability density for the signal strength, and is commonly taken to be uniform for $\mu \geq 0$ and zero for $\mu < 0$. Other priors are possible, but have hardly ever been used in high energy physics. The constant C is set by requiring that $\int L(\mu)d\mu = 1$.

The probability $p(\text{data}\,|\,\mu\,s+b)$ can be expressed as the product of Poisson probabilities for the number of observed (or simulated) events (n_k) in each bin (k) given the expected event rates per bin $\mu\,s_k + b_k$:

$$p(\text{data}\,|\,\mu\,s+b) \;=\; \prod_k \frac{(\mu\,s_k+b_k)^{n_k}}{n_k!}\,e^{-(\mu\,s_k+b_k)} \tag{B.2}$$

$$\;=\; e^{-(\mu\,S+B)}\,\prod_k \frac{(\mu\,s_k+b_k)^{n_k}}{n_k!}\,, \tag{B.3}$$

where $\mu\,S+B = \mu\sum_k s_k + \sum_k b_k$ is the total expected event rate. An unbinned approach to data can be thought of as a binned analysis in the limit of infinitely narrow bins in some observable x, which in general can be multi-dimensional. In this case, the function $p(\text{data}\,|\,\mu\,s+b)$, up to an irrelevant constant factor, becomes:

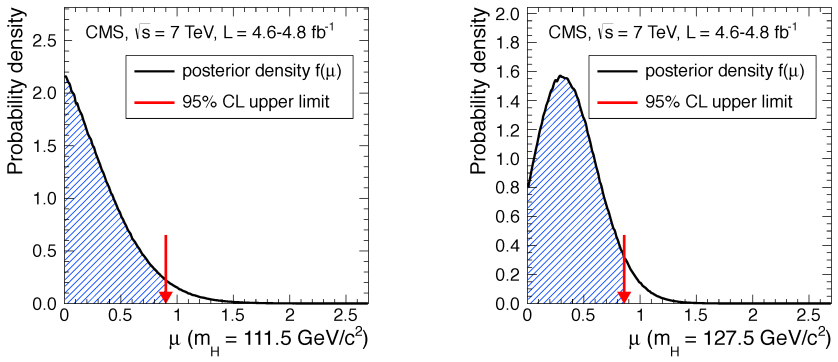
$$p(\text{data}\,|\,\mu\,s+b) \;\sim\; e^{-(\mu\,S+B)}\,\prod_i \mathcal{P}(\,x_i\,|\,\mu\,s+b\,), \tag{B.4}$$

where index i runs over all events, and $\mathcal{P}(\,x_i\,|\,\mu\,s+b\,)$ is an event density function of x such that the expected event rate in the vicinity of a given value of x is predicted as $\mathcal{P}(\,x\,|\,\mu\,s+b\,)dx$.

Integration over nuisance parameters in Eq. (B.1) is known as marginalisation. Marginalising the nuisance parameters sums the credibility of the parameter of interest over all possible values of the nuisance parameters, including the effects of systematic uncertainty, even far from the central predictions $\tilde{\boldsymbol{\theta}}$. This integration step also serves to constrain the values of the nuisance parameters in situ because the kernel of the integral is large for nuisance parameter values that fit the data well and is vanishingly small

for nuisance parameter values that fit the data poorly. The inclusion in the combination of data sets that constrain nuisance parameters helps improve the sensitivity of the Bayesian limits in much the same way that fits to nuisance parameter values improve the sensitivity of CL_s limits as discussed in Sec. B.2.2.2. The integrals over the space of nuisance parameters are often performed using Markov Chain Monte Carlo methods, such as the Metropolis-Hastings algorithm.[6] A benefit of this procedure is that posterior credibility density distributions for the nuisance parameters can be calculated alongside that for the signal, and inspecting these is an important validation step of the analysis. If one or more nuisance parameters are pulled multiple sigmas from their central values, or if the posterior uncertainties are unusually small for one or more nuisance parameters, this behavior ought to be investigated and explained. It is also useful to know if a nuisance parameter is driven against a boundary in its prior distribution.

Figure B.1 gives examples of Bayesian posterior probability densities $L(\mu)$ for experimental situations without or with an event excess. The distinction is whether the maximum of the posterior probability density is reached at zero signal strength or at a positive value.



(a) (b)

Fig. B.1. Examples of the Bayesian posterior probability density $L(\mu)$ for cases (a) without and (b) with an event excess observed. The 95% CL limit $\mu_{95\%\mathrm{CL}}$ is defined such that the integral of the shaded area for $0 \leq \mu \leq \mu_{95\%\mathrm{CL}}$ equals 0.95. Note that $\mu = 1$ is excluded in both cases. The plots are taken from Ref. [7].

The Bayesian one-sided 95% CL upper limit on μ is extracted using the

following equation:

$$\int_0^{\mu_{95\%\mathrm{CL}}} L(\mu)\, d\mu \;=\; 0.95. \tag{B.5}$$

This equation implies that one decides a priori to set limits only on high values of signal strength μ, even in a situation when a large excess of events is observed and small values of μ become just as unlikely as high values. Defining Bayesian credibility regions with an upper and a lower bound is performed with the same posterior probability density function; the procedure is described in Sec. B.4.

The experimental sensitivity is characterised by the limits expected to be set in the absence of a signal. These are computed by simulating many repeated runs of the experiment under the assumption of the background-only hypothesis, and computing an observed limit for each set of pseudo-data. Pseudo-data are simulated according to the Poisson distribution, assuming event rates $b(\boldsymbol{\theta})$. The Bayesian approach is to fluctuate the values of the nuisance parameters $\boldsymbol{\theta}$ for each pseudo-experiment according to their priors $\rho(\boldsymbol{\theta})$, so that the expected limit distribution is summed over all values the nuisance parameters can take, according to how much credibility they have. This is an important step when computing the sensitivity of an experiment that has not yet run and only highly uncertain *a priori* predictions are available for the signal and background yields. After the experiment has collected data, more is known about the expected backgrounds and the signal efficiencies, and the sensitivity may be updated. Alternatively, one may be tempted to take a seemingly conservative approach by setting the nuisance parameter values to the values that correspond to the least predicted sensitivity. One must be careful however, as nuisance parameter values set to increase the background prediction weaken the sensitivity, but strengthen the observed limit (for a given observed event count, the larger the assumed background is, the less room is left for a possible signal).

Since the distribution of expected limits is typically asymmetrical, the sensitivity is quoted as the median expected limit. Typically the distribution of expected limits is indicated by showing intervals containing 68% and 95% of the integral of the distribution, centered on the median. The quantiles are thus 0.025, 0.16, 0.5, 0.84, and 0.975. When these intervals are indicated on a plot of observed and expected limits versus m_H, they are usually shown with colored bands (e.g., see Fig. B.3).

The Bayesian method, in addition to the frequentist methods described in the next section, was frequently used to quantify results of Higgs boson

searches at the Tevatron. Its usage at LEP and the LHC was less prominent.

B.2.2. *Frequentist approach and its modifications*

B.2.2.1. *Classical frequentist*

The classical frequentist approach is formulated for the case of no systematic uncertainties and begins by defining a test statistic q_μ designed to discriminate signal-like from background-like events. The test statistic summarises all signal-vs-background discriminating information in one number. By the Neyman-Pearson lemma,[8] the ratio of likelihoods Q is the most powerful discriminator. For a number of practical reasons, the actual quantity used is a logarithm of the ratio, or more accurately, $-2\ln Q$:

$$q_\mu \;=\; -2\,\ln\frac{p(\text{data}\,|\,\mu\,s+b)}{p(\text{data}\,|\,b)}. \tag{B.6}$$

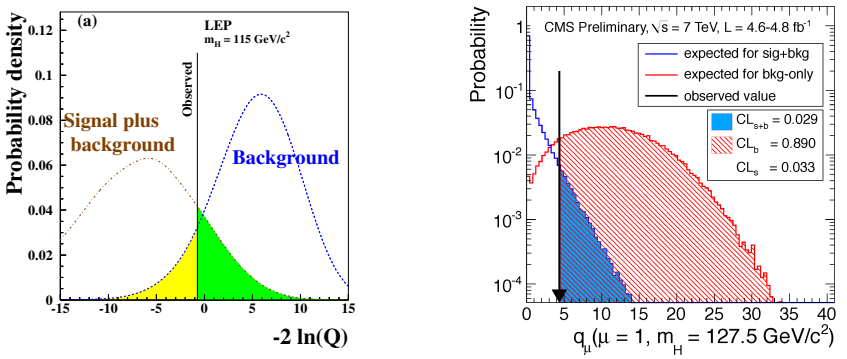
Modulo the modifications associated with handling systematic uncertainties, this is the test statistic used in quantifying null Higgs boson search results at LEP and the Tevatron in the frequentist paradigm context. In LEP papers, this test statistic was referred to as $-2\ln Q$, and in Tevatron papers, it was denoted LLR. There is another definition of the test statistic that has taken a prominent role at the LHC:

$$q_\mu \;=\; -2\,\ln\frac{p(\text{data}\,|\,\mu\,s+b)}{p(\text{data}\,|\hat{\mu}\,s+b)}, \quad \text{with a constraint: } 0\le\hat{\mu}\le\mu, \tag{B.7}$$

where $\hat{\mu}$ maximises the likelihood $p(\text{data}\,|\,\mu\,s+b)$. The advantage of this test statistic is that its distribution can be approximated by asymptotic formulae based on the theorems of Wilks and Wald, as derived in Ref. [9]. The upper bound on $\hat{\mu}$ ($\hat{\mu}\le\mu$) is needed when one desires to set one-sided limits only on high values of signal strength μ, even if an excess of events is observed.

Having chosen the test statistic q_μ, its distributions are constructed under the signal+background and background-only hypotheses by means of generating toy pseudo-observations according to the very same Poisson probabilities $p(\text{data}\,|\,\text{rate})$. Figure B.2 shows examples of distributions of the test statistics $-2\ln Q$ and q_μ defined by Eqs. (B.6) and (B.7) for the hypotheses of signal+background ($\mu=1$) and background-only ($\mu=0$).

For the test statistic defined by Eq. (B.6), experimental outcomes with $q_\mu > 0$ are more likely to appear under the background-only hypothesis than under the background+signal assumption. Assuming the signal+background hypothesis, the smaller number of observed events, the



(a) (b)

Fig. B.2. Examples of distributions of the test statistic q_μ defined by (a) Eq. (B.6) and (b) Eq. (B.7) for the signal+background and background-only hypotheses. Plots (a) and (b) are taken from Refs. [10] and [7], respectively.

larger value of the test statistic is. For the test statistic defined by Eq. (B.7), the test statistic is always positive definite; the smaller number of observed events, the larger value of the test statistic is.

Using these distributions, one can then evaluate the probability for the observed value $q_\mu^{\rm data}$ to be as or less compatible with the background+signal hypothesis. Such a probability, $P(q_\mu \geq q_\mu^{\rm data} \,|\, \mu\, s + b)$, is denoted as $\mathrm{CL_{s+b}}$. These probabilities correspond to the green and blue areas in Fig. B.2 (a) and ig. B.2 (b), respectively. In the classical frequentist approach, one says that the signal is excluded at, say, 95% CL, if $\mathrm{CL_{s+b}} = 0.05$.

However, such a definition has a pitfall: by taking the signal strength equal to zero, one expects, by construction, that $\mathrm{CL_{s+b}} \leq 0.05$ with a 5% chance in the background-only hypothesis — hence, 5% of all searches will end up excluding a signal of zero strength. In these cases, what has actually been observed is a downward fluctuation of the background. The exclusion of a zero-strength signal is certainly a questionable physics-wise result, even though proper mathematical coverage is guaranteed by the method. The problem with excluding a signal of zero strength is that an experiment cannot possibly test for the presence or absence of such a signal, and thus should not make a statement about it. To prevent, at least partially, the inference of a signal in presence of such downward fluctuations, a number of solutions have been suggested.

B.2.2.2. *Modifications of the classical frequentist method*

A method of constructing unified (i.e. one/two-sided) confidence intervals was suggested in Ref. [11] by Feldman and Cousins (FC). In this method, confidence intervals are constructed using ranking of experimental outcomes based on the value of the likelihood-ratio test statistic:

$$q_\mu \;=\; -2\,\ln\frac{p(\text{data}\,|\,\mu\,s+b)}{p(\text{data}\,|\,\hat{\mu}\,s+b)}, \quad \text{with a constraint: } 0\le\hat{\mu}, \tag{B.8}$$

where $\hat{\mu}$ maximises the likelihood $p(\text{data}\,|\,\mu\,s+b)$. Such construction automatically protects the limits on signal strength from the undesired effects of downward fluctuations of background, preserves coverage, and does not suffer from undercoverage due to having to make flip-flop decisions between reporting one-sided upper limits (no excess) and two-sided intervals when a considerable excess of events is observed. One, however, faces a conundrum: the FC method starts giving a lower limit on signal strength μ for excesses not yet significant enough for claiming a discovery (see Sec. B.3). To avoid the "inconvenience" of giving a statistical interpretation of reporting a lower limit on signal strength, while not claiming an observation of a signal, one can choose to report upper limits only—the price is overcoverage for the cases in which an excess of events is observed.

At the time of LEP, the so-called modified frequentist approach was introduced with the same goal to "protect" against too-strong statements made about vanishingly weak signals when downward fluctuations occur in the observed data.[12–14] In this method, in addition to probability $\text{CL}_{\text{s+b}} = P(q_\mu \ge q_\mu^{\text{data}}\,|\,\mu s+b)$, one also calculates $\text{CL}_{\text{b}} = P(q_\mu \ge q_\mu^{\text{data}}\,|\,b)$, by simulating pseudo-data for assuming the background-only hypothesis, and, then calculating the quantity CL_{s} as the ratio of these two probabilities:

$$\text{CL}_{\text{s}} \;=\; \frac{\text{CL}_{\text{s+b}}}{\text{CL}_{\text{b}}}. \tag{B.9}$$

The method does not prescribe the test statistic to be used. In the modified frequentist approach, it is this value, CL_{s}, that is required to be less than or equal to 0.05 in order to declare the 95% CL exclusion. By construction, the CL_{s}-based limits are one-sided. For $\mu = 0$, $\text{CL}_{\text{s}} \equiv 1$; hence, $\mu = 0$ cannot be excluded, regardless of how low the observed event count is. The price of the protection from background downward fluctuations is a gradual increase in the overcoverage as one observes fewer and fewer events. For an observation right on the top of the background-only expectation ($\text{CL}_{\text{b}} \sim 0.5$), CL_{s} is twice as large as $\text{CL}_{\text{s+b}}$.

Between the two modifications, Feldman-Cousins and CL_s, the latter was most frequently used at LEP, the Tevatron, and the LHC. However, there were distinct variations of the CL_s method, stemming form the differences in the choice of the test statistic and in the methods used to incorporate systematic uncertainties.

B.2.2.3. *Introducing systematic uncertainties*

Systematic uncertainties on the predicted signal and background rates, $s(\boldsymbol{\theta})$ and $b(\boldsymbol{\theta})$, are introduced via modifications to the test statistic itself and/or the way pseudo-data are generated. In the following, the prior densities for the nuisance $\boldsymbol{\theta}$ will be written as $\rho(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})$, where $\tilde{\boldsymbol{\theta}}$ is the "nominal" best-guess value of the nuisance parameter.

At LEP, the test statistic given by Eq. (B.6) was used; it was always evaluated at the nominal values of the signal and background rates, i.e. at $s(\tilde{\boldsymbol{\theta}})$ and $b(\tilde{\boldsymbol{\theta}})$. The effect of systematic uncertainties was then introduced via modifying $s(\boldsymbol{\theta})$ and $b(\boldsymbol{\theta})$ before each pseudo-data set was generated by drawing random numbers from the $\rho(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})$ distributions. This method was first introduced to the field by Cousins and Highland[15] and is now known as a hybrid Bayesian-frequentist method, since the treatment of nuisance parameters in this case is explicitly Bayesian.

At the Tevatron, the hybrid Bayesian-frequentist approach to generating the pseudo-data remained the same as at LEP, but the test statistic given by Eq. (B.6) was redefined in order to improve the sensitivity in the face of large systematic uncertainties. The Poisson-like likelihoods were extended to include the nuisance parameter densities $\rho(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})$

$$\mathcal{L}(\text{data}\,|\,\mu,\boldsymbol{\theta}) = p(\,\text{data}\,|\,\mu\cdot s(\boldsymbol{\theta}) + b(\boldsymbol{\theta})\,)\cdot\rho(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}}) \tag{B.10}$$

Before taking the ratio, both the numerator and the denominator likelihoods were maximised with respect to nuisance parameters. The test statistic then takes the following form:

$$q_\mu \;=\; -2\,\ln\frac{\mathcal{L}(\text{data}\,|\,\mu,\hat{\boldsymbol{\theta}}_\mu)}{\mathcal{L}(\text{data}\,|\,0,\hat{\boldsymbol{\theta}}_0)}, \tag{B.11}$$

where $\hat{\boldsymbol{\theta}}_\mu$ and $\hat{\boldsymbol{\theta}}_0$ are maximum likelihood estimators for the signal+background hypothesis (with the signal strength factor μ) and for the background-only hypothesis ($\mu = 0$).

At the LHC, the ATLAS and CMS experiments started to use the profile likelihood test statistic given by Eq. (B.7), which was further modified

to incorporate systematic uncertainties in the definition of likelihoods, as described below. The overall treatment of systematic uncertainties was conceptually different with respect to that used by LEP and the Tevatron; it was brought to be closer to the frequentist treatment of data fluctuations. First, following the Bayesian paradigm, systematic uncertainty densities $\rho(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})$ were reinterpreted as posteriors of some measurements of $\tilde{\boldsymbol{\theta}}$, either real (e.g., measurements in control regions) or imaginary (e.g., uncertainties on theoretical cross sections):

$$\rho(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}}) \sim p(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta}) \cdot \pi_\theta(\boldsymbol{\theta}), \tag{B.12}$$

where priors $\pi_\theta(\boldsymbol{\theta})$ were assumed to be flat. Second, these initial best-guess values $\tilde{\boldsymbol{\theta}}$ were treated on par with data in construction of the likelihoods and in generation of pseudo-data. The likelihood was as follows:

$$\mathcal{L}(\text{data}, \tilde{\boldsymbol{\theta}} \,|\, \mu, \boldsymbol{\theta}) = p(\,\text{data}\,|\,\mu \cdot s(\boldsymbol{\theta}) + b(\boldsymbol{\theta})\,) \cdot p(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta}), \tag{B.13}$$

which formally coincided with that used by the Tevatron. Then, as in the case of the Tevatron, to take advantage of data constraining a priori uncertainties, the test statistic was defined with the numerator and denominator likelihoods maximised:

$$q_\mu \;=\; -2\,\ln \frac{\mathcal{L}(\text{data}, \tilde{\boldsymbol{\theta}} \,|\, \mu, \hat{\boldsymbol{\theta}}_\mu)}{\mathcal{L}(\text{data}, \tilde{\boldsymbol{\theta}} \,|\, \hat{\mu}, \hat{\boldsymbol{\theta}})}, \quad 0 \le \hat{\mu} \le \mu, \tag{B.14}$$

where the pair of parameters $\hat{\mu}$ and $\hat{\boldsymbol{\theta}}$ gives the global maximum of the likelihood. Finally, the treatment of nuisance parameters in generation of pseudo-observations is where the approach taken by LHC was very different from that used at LEP and the Tevatron. Using the best values of the nuisance parameters for the background-only and for the signal+backhround hypotheses ($\boldsymbol{\theta_0}$ and $\boldsymbol{\theta_\mu}$, respectively), the pseudo-data and pseudo-measurements of $\tilde{\boldsymbol{\theta}}$ were generated. In other words, instead of using a Bayesian-frequentist hybrid method, the nuisance parameters were treated in a nearly pure frequentist way.

B.2.2.4. *Summary of the frequentist approaches used in the Higgs boson search at LEP, the Tevatron, and the LHC*

For comparison purposes, the differences in the CL_s method, as used at LEP, the Tevatron, and the LHC, are summarised in Table B.1 below. The LEP prescription does not allow one to take full advantage of the constraints imposed on the nuisance parameters by the data used in the

statistical analysis. The Tevatron and LHC versions of CL_s, though constructed differently, in practice give nearly identical results. The benefit of the LHC-type CL_s is that it uses a test statistic with useful asymptotic properties, as shown in Fig. B.3. Also, the sampling distributions of the test statistic can be built following the pure frequentist language.

Table B.1. Comparison of CL_s definitions as used at LEP, Tevatron, and LHC.

Collider	Test statistic	Profiled?	Test statistic sampling
LEP	$q_\mu = -2\ln\frac{\mathcal{L}(\text{data}\|\mu,\tilde{\theta})}{\mathcal{L}(\text{data}\|0,\tilde{\theta})}$	no	Bayesian-frequentist hybrid
Tevatron	$q_\mu = -2\ln\frac{\mathcal{L}(\text{data}\|\mu,\hat{\theta}_\mu)}{\mathcal{L}(\text{data}\|0,\hat{\theta}_0)}$	yes	Bayesian-frequentist hybrid
LHC	$q_\mu = -2\ln\frac{\mathcal{L}(\text{data}\|\mu,\hat{\theta}_\mu)}{\mathcal{L}(\text{data}\|\hat{\mu},\hat{\theta})}$ $(0 \le \hat{\mu} \le \mu)$	yes	frequentist

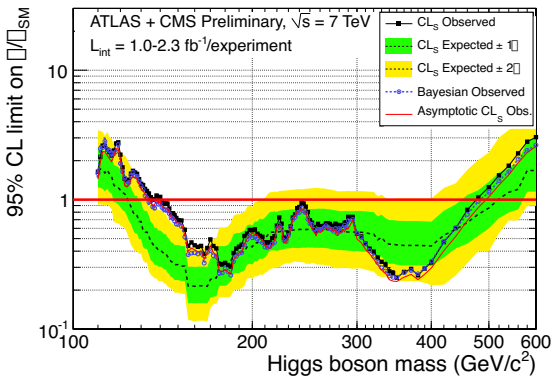


Fig. B.3. An example of limits on signal strength μ as obtained by generating pseudo-observations (exact) and via asymptotic formula (approximate). Also shown are limits obtained with the Bayesian technique. The results of the three calculations are very similar in the full Higgs boson mass range. The plot is taken from Ref. [16].

In all cases, the sensitivity of the experiment is given by the median limit expected to be set in the absence of a signal, and in nearly all cases,

the intervals containing 68% and 95% of the limits expected in pseudo-data are shown, centered on the median.

B.3. Significance of an excess of events

B.3.1. *Quantifying an excess of events for a given model*

In the case of observing an excess of events, characterisation of it begins with evaluating the p-value, i.e. the probability for background alone to yield an outcome as signal-like as observed. This can be done by generating background-only pseudo-data and building up the corresponding probability distribution for the test statistic of choice.

The four test statistics given in Eqs. (B.6), (B.11), (B.8), and (B.14) can be used. The first two compare the models $\mu = 0$ with $\mu = 1$, while the profile likelihood ratio used by the LHC is constructed for $\mu = 0$ and $\hat{\mu}$, where $\hat{\mu}$ is either unconstrained or constrained to be positive, which makes no difference to the tail of the distribution:

$$q_0 \;=\; -2\,\ln \frac{\mathcal{L}(\text{data}\,|\,0, \hat{\boldsymbol{\theta}}_0)}{\mathcal{L}(\text{data}\,|\,\hat{\mu}, \hat{\boldsymbol{\theta}})}. \tag{B.15}$$

For the first two test statistics, Eqs. (B.6) and (B.11), observations with a large excess of events would form a left-hand tail (see Fig. B.2), while the profile likelihood test statistic would stretch to the right as shown in Fig. B.4.

The p-value, i.e. the probability of getting an observation as or less compatible as seen in data for the background-only hypothesis, is then defined as $P(q_1 \leq q_1^{\text{data}})$ for the test statistics given by Eqs. (B.6) and (B.11), and $P(q_0 \geq q_0^{data})$ for the profile likelihood test statistic given by Eq. (B.15).

In addition to the p-value, the significance Z, commonly described as the number of standard deviations, is reported. A significance Z of three standard deviations is the customary criterion for "evidence", while a significance of five standard deviations is the commonly accepted criterion for "observation" of a new particle or process. Two conventions have been used to compute Z from the p-value (one-sided or two-sided normal distribution tail probability):

$$p = \int_Z^\infty \frac{1}{\sqrt{2\pi}}\,\exp(-x^2/2)\,dx, \tag{B.16}$$

$$p = \int_{-\infty}^{-Z} \frac{1}{\sqrt{2\pi}}\,\exp(-x^2/2)\,dx \;+\; \int_Z^\infty \frac{1}{\sqrt{2\pi}}\,\exp(-x^2/2)\,dx. \tag{B.17}$$

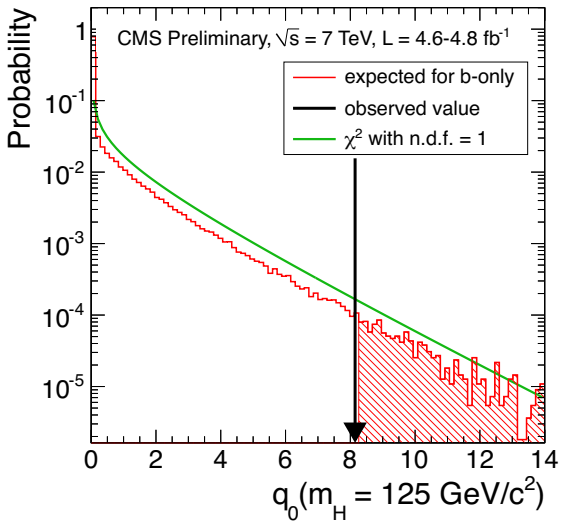


Fig. B.4. Example of a distribution of the profile likelihood test statistic q_0 (Eq. (B.15)). The shaded area represents the p-value, or the probability $P(q_0 \geq q_0^{\rm data})$. The solid green curve shows the asymptotic χ^2-distribution for one degree of freedom. The plot is taken from Ref. [7].

The two-sided convention of Eq. B.17 was used by the four LEP collaborations, while the one-sided convention of Eq. B.16 was used by the two Tevatron collaborations and the two LHC collaborations. In the one-sided convention, the Type-I error rate, that is, the probability that the null hypothesis will generate an experimental outcome that elicits a false claim of evidence ($Z > 3$), is approximately 0.00135, and the Type-I error rate for observation ($Z > 5$) is approximately 2.87×10^{-7}. For the two-sided convention, the Type-I error rates are twice as large.

In the asymptotic regime the profile likelihood test statistic (Eq. B.15) has the very attractive property of being distributed as a half χ^2 for one degree of freedom, which allows one to approximately estimate the significance, Z, as defined by Eq. (B.16) from the following formula:

$$Z \approx \sqrt{q_0^{\rm data}}. \tag{B.18}$$

The asymptotic approximation gives very satisfactory results for significance estimations even when one is far from the asymptotic regime (e.g., if one observes a few events, while the expected background rate is less than one event).

B.3.2. *Look-elsewhere effect*

In the Higgs boson search, the experimental collaborations scan over Higgs boson mass hypotheses and look for the one giving the minimum local p-value $p^{\min}_{\text{local}}$ (corresponding to local significance Z_{local}), which describes the probability of a background fluctuation for that particular Higgs boson mass hypothesis, as shown in Fig. B.5 (b). The probability to find a fluctuation with a local p-value lower or equal to the observed $p^{\min}_{\text{local}}$ anywhere in the explored mass range is referred to as the global p-value, p_{global}:

$$p_{\text{global}} = \mathrm{P}(p_0 \leq p^{\min}_{\text{local}} \,|\, b), \tag{B.19}$$

The fact that the global p-value can be significantly larger than $p^{\min}_{\text{local}}$ is often referred to as the look-elsewhere effect (LEE). The global significance (and global p-value) of the observed excess can be evaluated in this case by generating pseudo-datasets, which, however, becomes too CPU-intensive and not practical for very small p-values. Therefore, the method suggested in Ref. [18] was used. The relationship between the global and local p-values is given by:

$$p_{\text{global}} = p^{\min}_{\text{local}} + C \cdot e^{-Z^2_{\text{local}}/2}. \tag{B.20}$$

Assuming one can simulate correlations in data selected for different Higgs boson mass hypotheses, the constant C can be found by generating a relatively small set of pseudo-data and then use it to evaluate the global p-value corresponding to the value $p^{\min}_{\text{local}}$ observed in the experiment.

For a very wide mass range, the constant C can be evaluated directly from the data[19] by counting the number N_{up} of times that $\hat{\mu}(m_H)$ crosses the line $\mu = 0$ in the upwards direction, as shown in Fig. B.5 (c), and setting $C = N_{\text{up}}$.

B.3.3. *Discovery sensitivity*

In analogy to the procedure to compute the sensitivity of the experiment using the median expected limit, the discovery sensitivity is quantified using the median expected p-value assuming the presence of a signal. This is often quoted as the median expected Z value. Sensitivities are often shown without correction for the LEE, as curves on plots of median expected p-values as functions of the Higgs boson mass (e.g, see Fig. B.5 (b)).

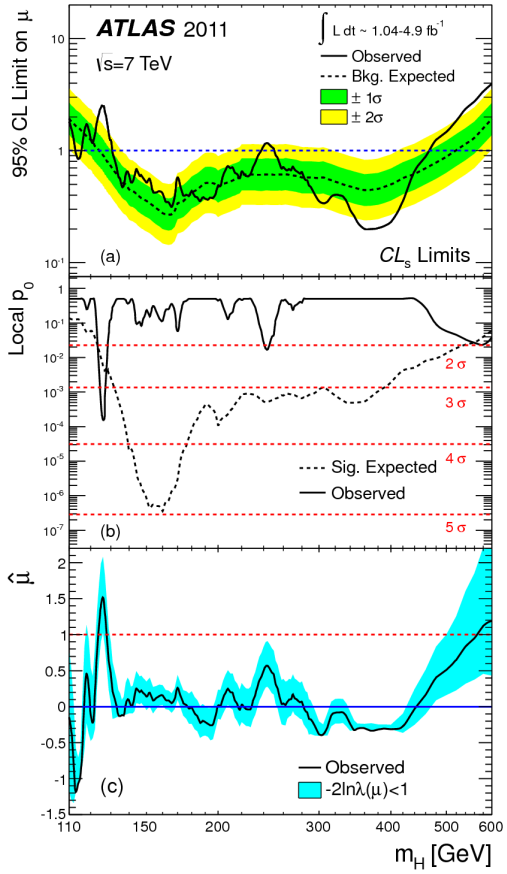


Fig. B.5. (a) Observed and expected limits on signal strength vs hypothesized Higgs boson mass. (b) Observed p-value vs hypothesized Higgs boson mass. (c) Best-fit μ as a function of a hypothesized Higgs boson mass. The number of times μ crosses 0 from negative to the positive value is called the number of upcrossings. The number of upcrossings can be used to evaluate the look-elsewhere effect directly from data as described in the text. The plot is taken from Ref. [17].

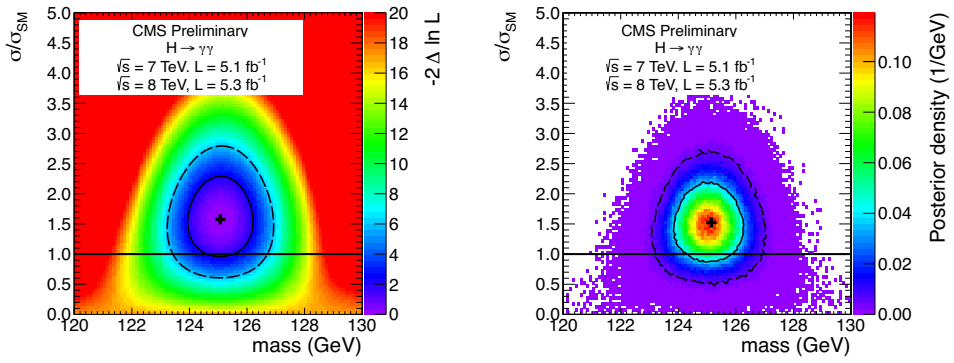
B.4. Extracting signal model parameters

Signal model parameters $\boldsymbol{a}$ (the signal strength modifier μ can be one of them) are evaluated from a scan of the profile likelihood ratio $q(\boldsymbol{a})$:

$$q(\boldsymbol{a}) = -2 \ln \frac{\mathcal{L}(\text{obs} \,|\, s(\boldsymbol{a}) + b, \hat{\boldsymbol{\theta}}_a)}{\mathcal{L}(\text{obs} \,|\, s(\hat{\boldsymbol{a}}) + b, \hat{\boldsymbol{\theta}})}, \tag{B.21}$$

The parameter values $\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{\theta}}$ that maximise the likelihood, $\mathcal{L}(\text{obs}\,|\,s(\hat{\boldsymbol{a}}) + b, \hat{\boldsymbol{\theta}}) = \mathcal{L}_{\max}$, are called the best-fit set. The 68% (95%) CL on a single parameter of interest a is evaluated from $q(a) = 1$ (3.84) with all other unconstrained model parameters treated in the same way as the nuisance parameters. The 2D 68% (95%) CL contours for pairs of parameters are derived from $q(a_1, a_2) = 2.3$ (6.0), as shown in Fig. B.6 (a) for a pair of parameters of interest $(m_H; \mu)$. One should keep in mind that boundaries of 2D confidence regions projected on either parameter axis are not identical to the 1D confidence interval for that parameter.

Alternatively, model parameters can be extracted using the Bayesian technique. For example, the posterior probability density $L(\boldsymbol{a})$ is computed by marginalising over the nuisance parameters, usually using a uniform prior density for the parameters of interest $\boldsymbol{a}$. The best-fit values $\hat{\boldsymbol{a}}$ are those which maximise $L(\boldsymbol{a})$, and the 68% (95%) CL region is the smallest-area region that contains 68% (95%) of the integral of the posterior density. Figure B.6 (b) shows the Bayesian posterior density $L(\boldsymbol{a})$ and 68% (95%) CL contours for the same datasets used for the profile likelihood scan presented in Fig. B.6 (a). The CL contours on both plots are remarkably similar. As is the case with limits, the marginalisation of the nuisance parameters explores the behavior of the likelihood function for all values of the nuisance parameters and not just those near the maximum.



(a) (b)

Fig. B.6. Examples of (a) 2D profile likelihood scan and (b) Bayesian posterior likelihood function. Solid (dashed) lines indicate 68% (95%) CL intervals. The plots are taken from Ref. [7].

The measurement sensitivity is quantified by computing the distribution of expected measurement uncertainties in pseudo-data drawn from models with known values of the parameters of interest. Usually the uncertainty obtained in the data fit is compared with the distribution of expected uncertainties in the process of validating a result in order to determine if the experiment is much luckier or unluckier than expected, and the median expected uncertainty is the figure of merit used to optimise the analysis. An observed uncertainty that is very different from what is expected requires investigation and explanation. The expected measurement uncertainty is also of value when averaging parameters, as the observed uncertainty is often correlated with the value of the parameter being measured, and this can bias weighted averages. Optimising simultaneous measurements of two or more parameters leaves more choices for figures of merit to use.

References

1. L. Lyons, *Statistics for Nuclear and Particle Physicists*. Cambridge University Press (1986).
2. B. P. Roe, *Probability and Statistics in Experimental Physics*. Springer (1992).
3. R. J. Barlow, *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences*. John Wiley & Sons (1993).
4. G. Cowan, *Statistical data analysis*. Oxford University Press (1998).
5. F. James, *Statistical methods in experimental physics*. World Scientific Publishing Company, Inc. (2006).
6. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, Equation of State Calculations by Fast Computing Machines, *J.Chem.Phys.* **21**, 1087 (1953). doi: 10.1063/1.1699114.
7. G. Petrucciani, *Observation of a new state in the search for the Higgs boson at CMS*. Springer (2013). URL `http://link.springer.com/book/10.1007%2F978-88-7642-482-3`.
8. J. Neyman and E. Pearson, On the problem of the most efficient tests of statistical hypotheses, *Phil. Trans. of the Royal Soc. of London A*. **31**, 289 (1933).
9. G. Cowan, K. Cranmer, E. Gross, and O. Vitells, Asymptotic formulae for likelihood-based tests of new physics, *Eur.Phys.J.* **C71**, 1554 (2011). doi: 10.1140/epjc/s10052-011-1554-0.
10. LEP Working Group for Higgs boson searches, ALEPH Collaboration, DELPHI Collaboration, L3 Collaboration, OPAL Collaboration, Search for the Standard Model Higgs boson at LEP, *Phys.Lett.* **B565**, 61–75 (2003). doi: 10.1016/S0370-2693(03)00614-2.
11. G. J. Feldman and R. D. Cousins, A Unified approach to the classical sta-

tistical analysis of small signals, *Phys.Rev.* **D57**, 3873–3889 (1998). doi: 10.1103/PhysRevD.57.3873.

12. T. Junk, Confidence level computation for combining searches with small statistics, *Nucl.Instrum.Meth.* **A434**, 435–443 (1999). doi: 10.1016/S0168-9002(99)00498-2.
13. A. Read, Modified frequentist analysis of search results (The CL_s method). In eds. F. James, Y. Perrin, and L. Lyons, *Workshop on confidence limits, CERN, Geneva, Switzerland, 17-18 Jan 2000: Proceedings* (2000). URL `https://cds.cern.ch/record/451614`.
14. A. L. Read, Presentation of search results: The CL_s technique, *J. Phys.* **G28**, 2693–2704 (2002). doi: 10.1088/0954-3899/28/10/313.
15. R. D. Cousins and V. L. Highland, Incorporating systematic uncertainties into an upper limit, *Nucl.Instrum.Meth.* **A320**, 331–335 (1992). doi: 10.1016/0168-9002(92)90794-5. Revised version.
16. ATLAS and CMS Collaborations, Combined Standard Model Higgs boson searches with up to 2.3 inverse femtobarns of pp collision data at sqrt(s)=7 TeV at the LHC, *ATLAS-CONF-2011-157, CMS-PAS-HIG-11-023* (2011). URL `http://cds.cern.ch/record/1399599,https://cds.cern.ch/record/1399607`.
17. ATLAS Collaboration, Combined search for the Standard Model Higgs boson using up to 4.9 fb^{-1} of *pp* collision data at $\sqrt{s} = 7$ TeV with the ATLAS detector at the LHC, *Phys.Lett.* **B710**, 49–66 (2012). doi: 10.1016/j.physletb.2012.02.044.
18. E. Gross and O. Vitells, Trial factors for the look elsewhere effect in high energy physics, *The European Physical Journal C - Particles and Fields*. **70**, 525–530 (2010). ISSN 1434-6044. URL `http://dx.doi.org/10.1140/epjc/s10052-010-1470-8`. 10.1140/epjc/s10052-010-1470-8.
19. ATLAS and CMS Collaborations, Procedure for the LHC Higgs boson search combination in summer 2011 (2011). URL `https://cds.cern.ch/record/1379837`.