Appendix A

95% CL limits and 5 σ discoveries

A.1 Estimators of significance

Several methods exist to quantify the statistical "significance" of an expected signal at future experiments. Following the conventions in high energy physics, the term significance usually means the "number of standard deviations" an observed signal is above expected background fluctuations. It is understood implicitly that S should follow a Gaussian distribution with a standard deviation of one. In statistics, the determination of the sensitivity is a typical problem of hypothesis testing, aiming at the discrimination between a null-hypothesis H_0 stating that only background and no signal is present, and a alternative hypothesis H_1, which states the presence of a signal on top of the background. The "significance level" is the probability to find a value of a suitably constructed test statistic beyond a certain prespecified critical value, beyond which the validity of H_1 is assumed. The significance level has to be converted into an equivalent number of Gaussian sigmas to arrive at the common terminology of a high-energy physicist.

Since a signal is usually searched for in many bins of a distribution, and in many channels, a very high value of the significance of a local excess of events must be demanded before an observed "peak" found somewhere in some distribution can be claimed to be an observation of a signal. If the position of the signal peak is not known a-priori and treated as a free parameter in searches for new physics, the probability of background fluctuations is much higher. This is quantified in a case study in section A.2 below, and this aspect will need careful consideration in the near future before first data taking at the LHC. The general, somewhat arbitrary convention is that the value of S of a local signal excess should exceed five, meaning that the significance level, or the corresponding one-sided Gaussian probability that a local fluctuation of the background mimics a signal, is $2.9{\cdot}10^{-7}$.

Here, the recommendations for the procedures to be used for the studies presented in this document are summarised. The aim of many of these studies is the prediction of the average expected sensitivity to the observation of a new signal in a future experiment. The real experiment might be lucky, i.e. observe a higher significance than the average expectation, or a downward fluctuation of the expected signal could lead to a lower observed significance. The proposed methods have been checked in a large number of pseudo-experiments using Monte Carlo simulation in order to investigate whether the probability of a background fluctuation having produced the claimed significance of the discovery is properly described.

Counting methods use the number of signal events, s, and the number of background events, b, observed in some signal region to define the significance S. These event numbers can be turned into a significance, S_{cP}, by using either the Poisson distribution for small numbers of

events, or, in the high-statistics limit, the Gaussian distribution, leading to

$$S_{c1} = \frac{s}{\sqrt{b}}. \tag{A.1}$$

The significance may also be obtained from the ratio of the likelihoods, $\mathcal{L}_1$ and $\mathcal{L}_0$, belonging to the hypothesis H_0 and H_1,

$$S_L = \sqrt{2 \ln Q}, \; with \; Q = \frac{\mathcal{L}_0}{\mathcal{L}_1}. \tag{A.2}$$

This approach is theoretically well founded and is applicable also to the simple approach of the counting method, leading to

$$S_{cL} = \sqrt{2 \left((s+b) \, \ln (1 + \frac{s}{b}) - s \right)}, \tag{A.3}$$

which follows directly from the Poisson distribution . In the Gaussian limit of large numbers s and b, S_{cL} becomes equivalent to S_{c1}. The likelihood approach can be extended to include the full shapes of the signal and background distributions for the hypothesis H_0 and H_1, and the likelihood may be obtained from binned or unbinned likelihood fits of the background-only and the background-plus-signal hypotheses to the observed distributions of events.

Another estimator,

$$S_{c12} = 2 \, (\sqrt{s+b} - \sqrt{b}) \tag{A.4}$$

has been suggested in literature [78, 762]. The formula for S_{c12} is strictly only valid in the Gaussian limit, but tabulated values exist for small statistics.

The presence of systematic errors deserves some special care. Two cases must be separated clearly:
a) If the background and signal contributions can be determined from the data, *e.g.* by extrapolating the background level into the signal region from sidebands, systematic errors may be irrelevant, and the systematic errors only influence our ability to predict the average expected sensitivity. In this case, simple propagation of the theoretical errors on s and b applied to the above formulae for the various significances is all that is needed.
b) If systematic errors on the background will affect the determination of the signal in the real experiment, *e.g.* because an absolute prediction of the background level or a prediction of the background shape are needed, the theoretical uncertainty must be taken into account when estimating the sensitivity. This can be done by numerical convolution of the Poisson distribution, or the Gaussian distribution in the high-statistics limit, with the probability density function of the theoretical uncertainty. Numerical convolutions of the Poisson distribution with a theoretical error of a Gaussian shape, leading to a variant of S_{cP} including systematic errors, were used for this document [679]. Numerical convolutions of the Poisson distribution with a systematic error of a Gaussian shape, leading to a variant of S_{cP} including systematic errors, were used for this document. The program ScPf [679] computes the significance by Monte Carlo integration with the assumption of an additional Gaussian uncertainty Δb on b. The significance can be approximated by an extension of S_{c12}:

$$S_{c12s} = 2 \, (\sqrt{s+b} - \sqrt{b}) \, \frac{b}{b + \Delta b^2}. \tag{A.5}$$

In the Gaussian limit it leads to

$$S_{c1} = s / \sqrt{b + \Delta b^2}. \tag{A.6}$$

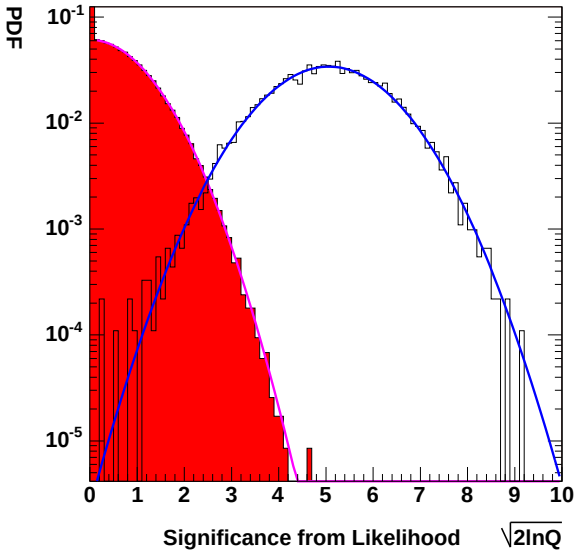

Figure A.1: Probability density functions of the estimator of significance S_L for small statistics (11 signal events over a background of 1.5 events). Filled histogram : pure background sample from 200 000 toy experiments, open histogram: background plus signal from 10 000 toy experiments. Gaussian fits are overlayed; the distribution of S_L for the background-only sample has a mean of -0.004 and a width of σ=1.0 , the background-plus-signal sample has a width of 1.1 .

The most crucial point in this context is a realistic description of the probability density function of the systematic theoretical uncertainty, which can be anything ranging from a flat distribution between $b \pm \Delta b$ to a pathological distribution with a significant non-Gaussian tail, but, in practice, is hardly ever known precisely.

The distribution of a significance estimator S in a series of experiments, its probability density function ("*pdf*"), is of prime importance for the calculation of discovery probabilities in the presence of a real signal, or of fake probabilities due to fluctuations of the background. In the large-statistics limit, the likelihood-based significance estimators are expected to follow a χ^2-distribution with a number of degrees of freedom given by the difference in the number of free parameters between the alternative hypothesis and the null hypothesis [102]. When testing for the presence of a signal on top of background at a fixed peak position, $2 \ln Q = {S_L}^2$ is expected to follow a χ^2 distribution with one degree of freedom, *i.e.* a standard Gaussian distribution. All of the above estimators have been tested in a large number of toy experiments, see *e.g.* References [51, 99, 101]. In particular the likelihood based estimators were found to be well-behaved, *i.e.* the distribution of the values of significance followed the expected behaviour already at moderate statistics, as is shown for one example in Figure A.1. Good scaling with the square root of the integrated luminosity was also observed in these studies. On the other hand, the estimator S_{c1} cannot be considered a useful measure of significance at low statistics.

A quantitative comparison as a function of the number of background events for fixed values of $s/\sqrt{b}$ of the various estimators discussed above is shown in Figure A.2. S_{cL} and S_{cP} are found to agree very well, while S_{c12} tends to slightly underestimate the significance, a result which was also verified in the above Monte Carlo studies with large samples of toy

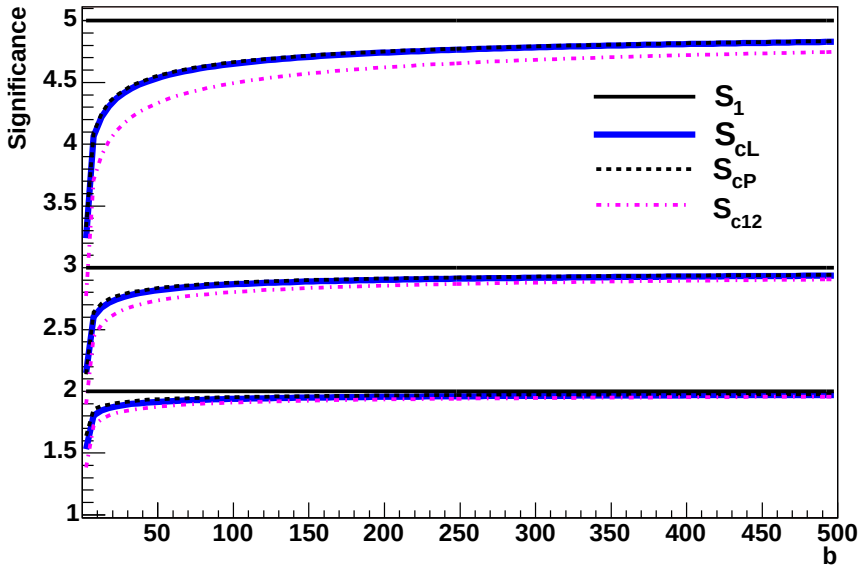

Figure A.2: Comparison of the various significance estimators as a function of the number of background events, b. The number of signal events was taken as $s = S_{c1}\sqrt{b}$, hence the constant black lines represent the value of S_{c1}. As can be seen, S_{cP} and S_{cL} agree perfectly, while S_{12} leads to slightly smaller values of significance. S_1 significantly overestimates the significance at small event numbers.

experiments. While S_{cL} and S_{cP} remain valid independent of the value of b, the simpler estimator S_{c1} can only be used for background levels larger than 50 events.

A.2 On the true significance of a local excess of events

In searching for new phenomena in a wide range of possible signal hypotheses (*e.g.* , a narrow resonance of unknown mass over a broad range background), a special care must be exercised in evaluating the true significance of observing a local excess of events. In past, this fact was given substantial scrutiny by statisticians (*e.g.* , [763, 764]) and physicists alike (*e.g.* , [765–769]). The purpose of this Appendix is to quantify a possible scope of this effect on an example of a search for the Standard Model Higgs boson in the $H \to ZZ^{(*)} \to 4\mu$ decay channel. As the case study, we chose a counting experiment approach widely used in this volume.

The dashed line in Fig. A.3 shows the expected 4μ invariant mass distribution for background at $\mathcal{L} = 30\,\text{fb}^{-1}$ after applying all the $m_{4\mu}$-dependent analysis cuts described in Sec. 3.1. Using this distribution, we played out $\sim 10^8$ pseudo-experiments; an example is shown in Fig. A.3. For each pseudo-experiment, we slid a *signal region window* across the spectrum looking for a local event excess over the expectation. The size of the window $\Delta m = w(m_{4\mu})$ was optimised and fixed *a priori* (about $\pm 2\sigma$) to give close to the best significance for a resonance with a width corresponding to the experimental SM Higgs boson width $\sigma(m_{4\mu})$. The step of probing different values of $m_{4\mu}$ was "infinitesimally" small (0.05 GeV/c^2) in comparison to the Higgs boson width of more than 1 GeV/c^2. The scanning was performed

in *a priori* defined range of 115-600 GeV/c^2.

We used a significance estimator $S_{cL} = \mathrm{sign}(s)\sqrt{2\,n_o \ln(1+s/b) - 2\,s}$, where b is the expected number of background events, n_o is the number of observed events, and the signal is defined as $s = n_o - b$). This estimator, based on the Log-Likelihood Ratio, is known to follow very closely the true Poisson significance, only slightly over-estimating it in the limit of small statistics [51]. Figure A.4 presents the results of such a scan for the pseudo-experiment shown in Fig. A.3. The maximum value of S_{cL}, S_{max}, and the corresponding mass of a "Higgs boson candidate" obtained in each pseudo-experiment were retained for further statistical studies.

After performing 10^8 pseudo-experiments, the differential probability density function for S_{max} and its corresponding cumulative probability function $P(S_{max} > S)$ (Fig. A.5) were calculated. From Fig. A.5, one can see that the frequency of observing some large values of S_{cL} (solid line) is much higher than its naive interpretation might imply (dashed line). If desired, the actual probability can be converted to the true significance. The result of such "renormalisation" is presented in Fig. A.6. One can clearly see that the required de-rating of significance is not negligible; in fact, it is larger than the effect of including all theoretical and instrumental systematic errors for this channel (see Sec. 3.1). More details on the various aspects of these studies can be found in [51].

There are ways of reducing the effect. A more detailed analysis of the shape of the $m_{4\mu}$ distribution will help somewhat. Using the predicted number of signal events $s = s_{theory}$ in the significance estimator to begin with and, then, for validating the statistical consistency of an excess $n_o - b$ with the expectation s_{theory} will reduce the effect further. One can also use a non-flat prior on the Higgs mass as it comes out from the precision electroweak measurements. Whether one will be able to bring the effect to a negligible level by using all these additional constraints on the signal hypotheses is yet to be seen. The purpose of this Appendix is not to give the final quantitative answer, but rather to assert that these studies must become an integral part of all future search analyses when multiple signal hypotheses are tried.

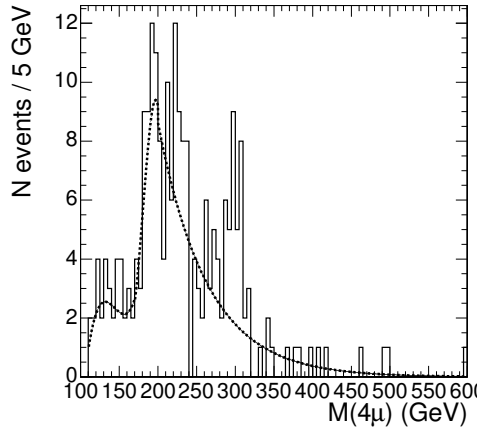

Figure A.3: The background *pdf* and an example of one pseudo-experiment with a statistical fluctuation appearing just like a signal.

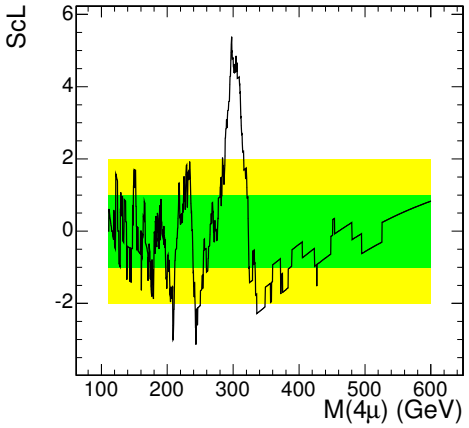

Figure A.4: Profile of the S_{cL} scan corresponding to the pseudo-experiment example shown on the left. Green (inner) and yellow (outer) bands denote $\pm 1\sigma$ and $\pm 2\,\sigma$ intervals. Spikes that can be seen are due to events coming in or dropping off the trial-window, a feature of low-statistics searches.

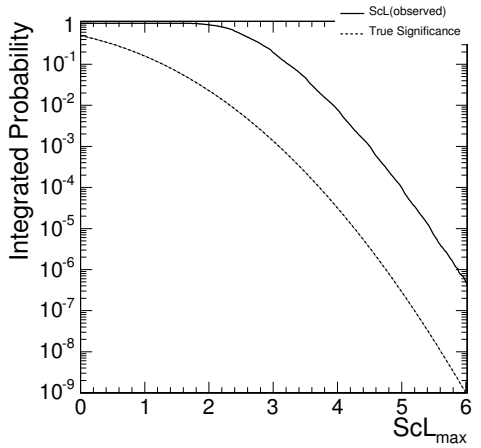

Figure A.5: S_{cL} cumulative probability density function.

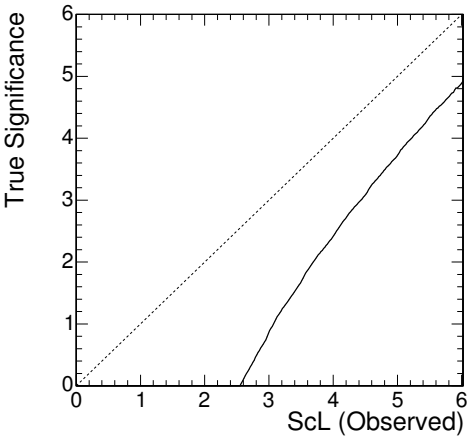

Figure A.6: Local significance "renormalisation" from an observed value to the true significance with a proper probabilistic interpretation.