GSSI - statistical and software tools for data analysis (M. Agostini)

Guido Fantini

Contents

1 Coverage

Let us consider a toy analysis where the distribution of the observable quantity x in both the signal and the background hypothesis is known. The observable is defined in the range $(-10, 10)$ and its probability density function in the two cases is

$$f_s(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{where} \quad \mu = 0 \;\; \sigma = 1$$

$$f_b(x) = \frac{1}{20}$$

A frequentist analysis is performed for different combinations of *true* signal and background expected counts in order to extract a confidence interval for the signal at 90%C.L.. A coverage test is performed in each expected background condition as a function of the expected signal count and an overcoverage is observed when the *true* number of signal events is below the sensitivity defined by the background.

1.1 Procedure

Let us define the *true* expected number of signal (background) events as $S(B)$. Two background conditions were simulated: $B = 1$ (low background) and $B = 100$ (high background). For each of these conditions a set of *true* S values was defined and the following procedure was performed:

- two random numbers N_s and N_b were extracted according to Poisson distributions of parameter S and B respectively (i.e. $Poisson(N_s|S)$ and $Poisson(N_b|B)$). Being a random process, even when the expected number of signal and background events is known exactly, the actual number of signal and background realizations fluctuates;

- N_s events (i.e. realizations of the random variable x in the signal hypothesis) were generated from the f_s distribution and added to an histogram;

- N_b events were generated from the f_b distribution and added to the same histogram;

- the histogram was fit with the distribution $f(x|S, B) = Sf_s + Bf_b$ where S and B are parameters. A binned likelihood was built in order to account for the poissonian fluctuations of the bin contents $\mathcal{L}(S, B) \doteq \prod_{bins} Poisson(counts_{bin}| \int_{bin} f(x|S, B)dx)$ and maximized globally (as a function of both S and B) in order to find the best fit number of signal and background events as maximum likelihood estimators $\hat{S}$ and $\hat{B}$;

- assuming the Wilks theorem applies (i.e. in the large sample limit) and the null model (the one defined by $\hat{S}$ and $\hat{B}$) is true, the log-likelihood ratio test statistics

$$t = -2log\frac{\mathcal{L}(S, \hat{\hat{B}})}{\mathcal{L}(\hat{S}, \hat{B})}$$

  where $\hat{\hat{B}}$ is the value of B that maximizes the likelihood as a function of B only, has a chi-square distribution with one degree of freedom. This means that, once a confidence level is

set, we can define an interval as the union of the S values for which the test statistics is below some threshold t_{max} such that

$$\int_{t_{max}}^{\infty} f_t(t)dt = 1 - \text{C.L.}$$

When the test statistics is below threshold, the hypothesis test between the null hypothesis ($\hat{B}$ and $\hat{S}$) and the alternate hypothesis (S and $\hat{\hat{B}}$) would accept the null hypothesis. Since we assume the Wilks' approximation holds, the threshold equals the $1 - \text{C.L.}$ quantile of the $f(\chi^2, \nu = 1)$ distribution

$$t_{max} = 2.7 \quad C.L. = 90\%$$

A 90% confidence interval was extracted according to this approximation.

- the above procedure from the poisson generation of the number of events from the *true* value to the extraction of the confidence interval was repeated $N_{try} = 10^4$ times;
- for each repetition a check was performed whether the *true* value of the signal events was contained in the confidence interval. The ratio between the number of successes and the number of trials yield an estimate of the coverage of the interval obtained with this method;
- an estimation of the uncertainty on the coverage estimation was performed. The coverage estimator

$$\hat{c} = \frac{n_{inside}}{N_{try}}$$

is given by the ratio of a binomial variable n_{inside} and a constant. Hence

$$\sigma(\hat{c}) = \frac{1}{N_{try}}\sigma(n_{inside}) = \frac{\sqrt{c(1-c)}}{\sqrt{N_{try}}}$$

1.2 Results

The results of the procedure described in the previous section are reported below in Fig. 1 and 2. An overcoverage is observed when the number of *true* signal events is below the sensitivity. A very rough order-of-magnitude-estimate of the sensitivity can be obtained computing the square root of the number of expected background events in the "signal region" defined as $\pm 1\sigma$ of the signal distribution i.e. $\sqrt{B/6}$. In the low background case the sensitivity is up to $S \sim \sqrt{1/6} \sim 0.4$ and the problem is not visible (see Fig. 1). In the high background case instead the sensitivity is up to $S \sim \sqrt{100/6} \sim 4$ and this is the reason why for $S = 1$ and $S = 0.1$ an overcoverage is observed (see Fig. 2).

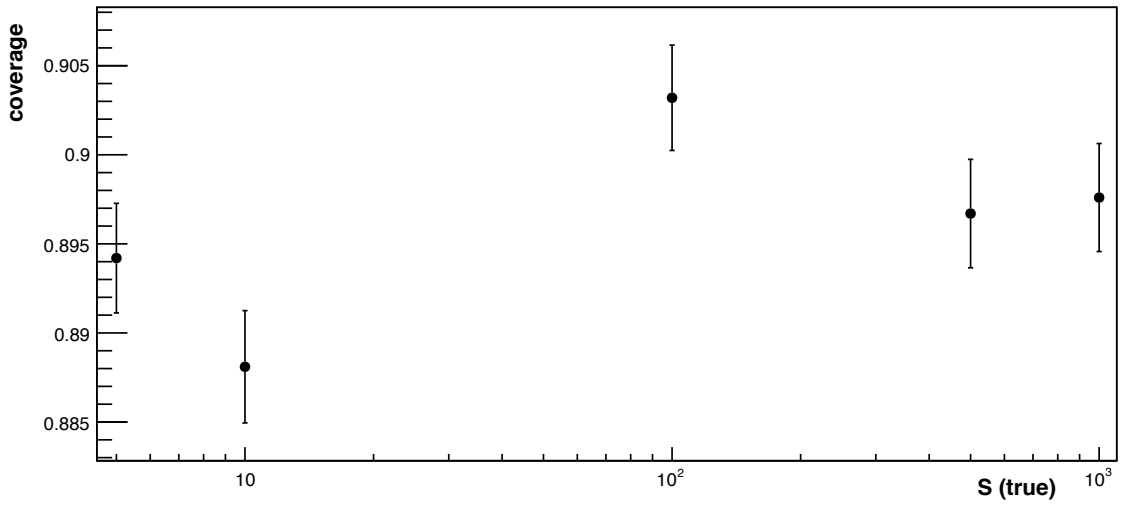

Figure 1: Coverage computation in the low background ($B = 1$) case. $N_{try} = 10^4$ data samples generated.

2 The distribution of the log-likelihood ratio

jhabddjh

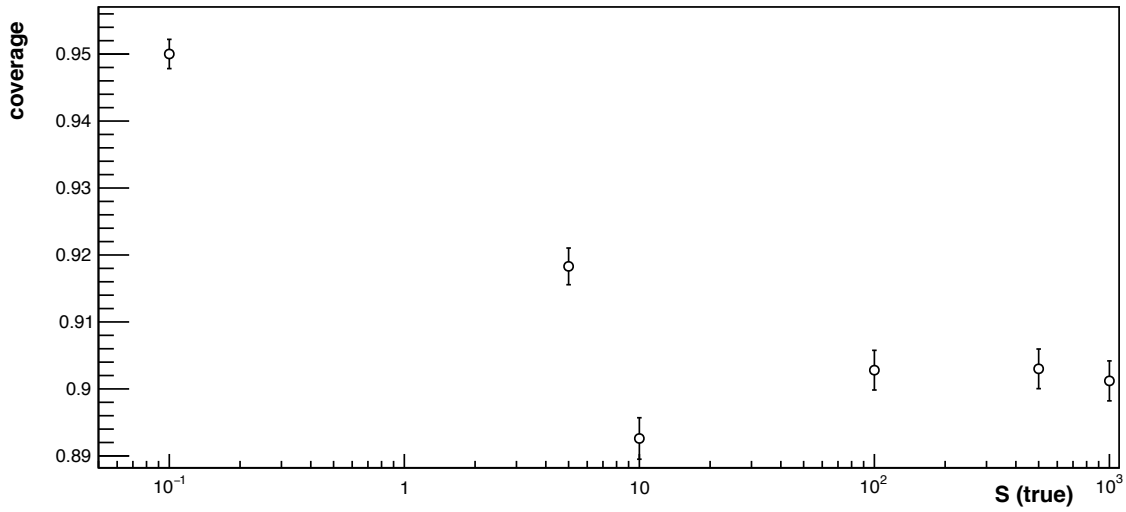

Figure 2: Coverage computation in the high background ($B = 100$) case. $N_{try} = 10^4$ data samples generated.

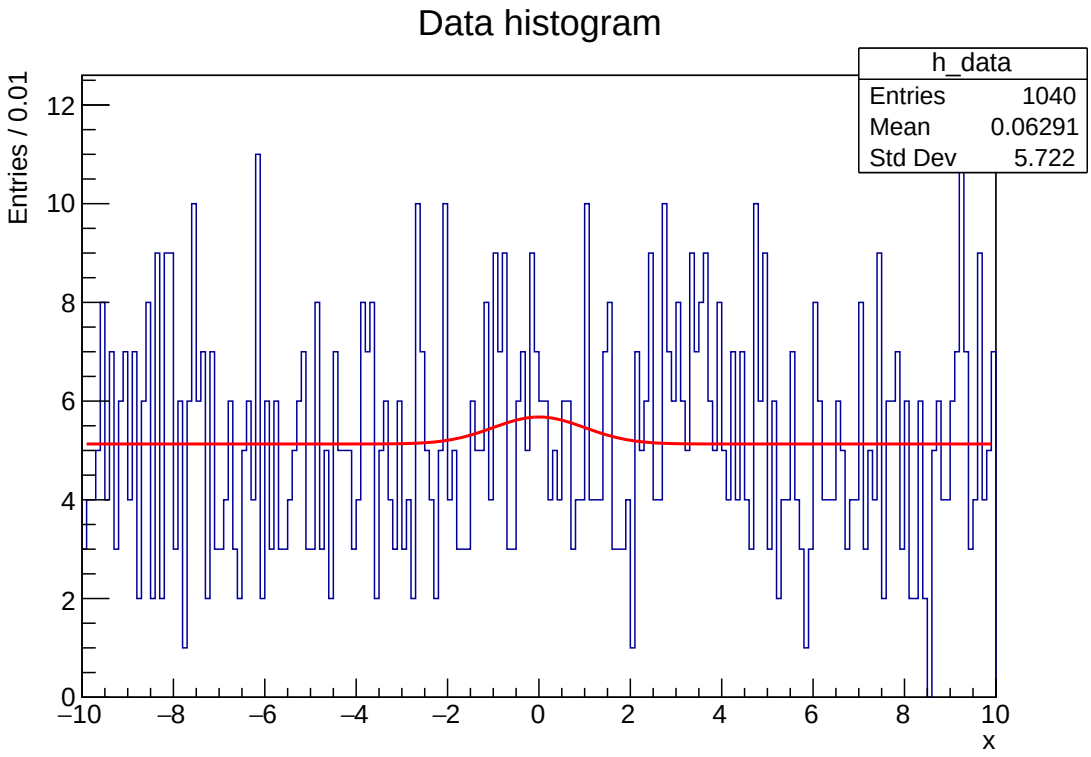

Figure 3: Example of event generation (blue) and best fit function (red).

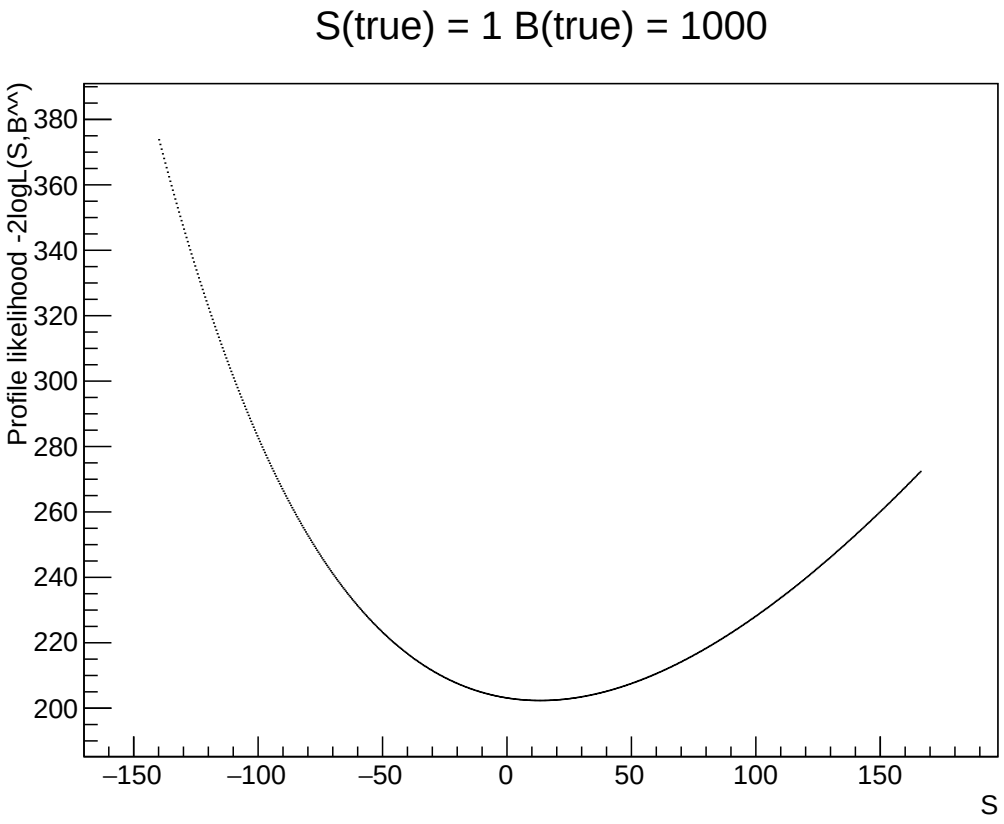

Figure 4: Example of profile likelihood $-2log(\mathcal{L}(S, \hat{\hat{B}}))$ as a function of the alternate signal S.