DRAFT 0.0

Glen Cowan
5 September 2022

Test from likelihood ratio

Suppose we measure the diphoton or dielectron mass distribution and the result is a histogram with N bins, having numbers of entries $\mathbf{n} = (n_1, \ldots, n_N)$. A particular model will predict the expected number of events in each bin i,

$$E[n_i] = \mu s_i + b_i \ , \tag{1}$$

where b_i is the expected number under the background-only (SM) hypothesis and s_i is the expected number from a specific signal model, i.e., a given clockwork model with parameters (k, M_5). The strength parameter μ has been introduced for convenience such that $\mu = 0$ gives the background-only model and $\mu = 1$ is the nominal signal model.

1 Test using the likelihood ratio

A possible way of testing a signal model is take as the test statistic the (log-)likelihood ratio

$$t(\mathbf{n}) = \ln \frac{L(1)}{L(0)} \tag{2}$$

where

$$L(\mu) = P(\mathbf{n}|\mu) = \prod_{i=1}^{N} \frac{(\mu s_i + b_i)^{n_i}}{n_i!} e^{-(\mu s_i + b_i)} \tag{3}$$

is the likelihood for strength parameter μ (and assuming the set of s_i and b_i) assuming that the n_i are independent and Poisson distributed.

Using the Poisson likelihood (3) in the test statistic (2) gives

$$\begin{aligned} t(\mathbf{n}) &= \sum_{i=1}^{N} [n_i \ln(s_i + b_i) - (s_i + b_i) - n_i \ln b_i + b_i] \\ &= \sum_{i=1}^{N} s_i + \sum_{i=1}^{N} n_i \ln\left(1 + \frac{s_i}{b_i}\right) \ . \end{aligned} \tag{4}$$

The first term $\sum_i s_i$ in Eq. (4) is independent of the data and thus can be dropped from the definition of the statistic, which we thus take to be

$$t(\mathbf{n}) = \sum_{i=1}^{N} n_i \ln\left(1 + \frac{s_i}{b_i}\right) \ . \tag{5}$$

This confirms that if the signal and background distributions have the same shape, i.e., s_i/b_i is constant, then the optimal statistic is simply (proportional to) the total number of events observed.

To carry out the test one proceeds as with any other test statistic, i.e., with Monte Carlo one can work out the distributions of t under assumptions of $\mu = 0$ and $\mu = 1$, $f(t|0)$ and $f(t|1)$. The real data give an observed value $t_{\rm obs}$, and the p-values of the models are

$$p_0 = \int_{t_{\rm obs}}^{\infty} f(t|0)\, dt \;, \tag{6}$$

$$p_1 = \int_{-\infty}^{t_{\rm obs}} f(t|1)\, dt \;. \tag{7}$$

Note that this p_1 pertains to the specific point in the parameter space of the signal model used to compute the s_i. The unbinned version of this method is discussed, e.g., in Ref. [1].

2 Test with profile likelihood ratio

To find the p-values in Eq. (6) one needs to use toy Monte Carlo. This is not too difficult but it may also be interesting to use the statistic

$$t_\mu = -2 \ln \frac{L(\mu)}{L(\hat{\mu})} \;, \tag{8}$$

where $\hat{\mu}$ is the maximum-likelihood estimator (MLE) of the strength parameter μ. According to Wilks' theorem [2], in the large-sample limit (which should be an excellent approximation for the clockwork analysis), the statistic t_μ is a chi-square distributed for with one degree of freedom. To the extent that this distribution is a reliable approximation, one does not need any Monte Carlo simulation to obtain the desired p-values, but rather can find this from appropriate integrals of the chi-square distribution, as discussed in Ref. [3]. In general one does need, however, a numerical optimization to find the MLE $\hat{\mu}$.

3 Including systematics

Systematic uncertainties can be included even if the test is carried out using the statistic defined by Eq. (4) with the nominal values of s_i and b_i without systematics. The key is to find the distributions of t by including the systematic uncertainties into the model that generates the data histogram $\mathbf{n}$. Suppose that the data histogram follows a probability distribution $P(\mathbf{n}|\mu, \theta)$ where as above μ is the parameter of interest and θ represents one or more nuisance parameters, for which one has a prior pdf $\pi(\theta)$. The data distribution including systematics is found from

$$P(\mathbf{n}|\mu) = \int P(\mathbf{n}|\mu, \theta)\pi(\theta)\, d\theta \;. \tag{9}$$

As usual one generates data histograms $\mathbf{n}$ by first sampling $\theta \sim \pi(\theta)$ and using this to generate $\mathbf{n} \sim P(\mathbf{n}|\mu, \theta)$. Then one evaluates $t(\mathbf{n})$ and records, e.g., in a histogram. The resulting pdf of t corresponds to whatever hypothesized μ was chosen and it includes systematic uncertainties.

References

[1] Glen Cowan, *Statistics for Searches at the LHC*, Proceedings, 69th Scottish Universities Summer School in Physics, St.Andrews, Scotland, August 19-September 1, 2012, 321-355; e-print: arXiv:1307.2487.

[2] S.S. Wilks, *The large-sample distribution of the likelihood ratio for testing composite hypotheses*, Ann. Math. Statist. **9** (1938) 60-62.

[3] Glen Cowan, Kyle Cranmer, Eilam Gross and Ofer Vitells, Eur. Phys. J. C 71 (2011) 1554.