RooStats Exercise

Jochen Ott

January 14–16, 2013

Contents

Statistical Model

The result of almost all analyses in HEP are statistical statements. The statistical inference relies on having a *statistical model*: A statistical model specifies the probability to observe a certain dataset as the function of the real-valued *model parameters*.
The parameters comprise

- *parameter(s) of interest (POI)*: the parameters one wants to make a statistical statement about. In many cases, there is only one POI, such as the cross section of some specific process.
- *nuisance parameters*: all other model parameters, included to account for uncertainties, e.g. background level, jet energy scale, luminosity, ...

Example: Counting Experiment I/II

The prototype model for many analyses is the counting experiment. The dataset is given by the number of observed events n. The probability to observe some dataset n is the Poisson distribution,

$$p(n|r) = \frac{e^{-\mu}\mu^n}{n!}$$

where

$$\mu = r \cdot \mu_s + \mu_b$$

where μ_s and μ_b are the (fixed) signal and background yields. The only model parameter is r.
r is the cross section, in units of the cross section used to compute μ_s.

Example: Counting Experiment II/II

Typically, systematic uncertainties are introduced via *nuisance parameters*. Consider the same counting experiment, but this time, in

$$\mu = r \cdot \mu_s + e^{0.1\delta}\mu_b$$

where δ is a nuisance parameter. To encode prior knowledge, a *Gaussian prior* for δ is used with mean 0 and std.-dev. 1. This corresponds to a lognormal uncertainty of 10% on the background yield; this is what will be used in this exercise later.
Other generalizations:

- multiple bins: Poisson in each bin; model parameters can affect multiple bins simultaneously
- bin number $\to \infty$: probability density function for observables

Statistical Methods

The kind of statistical inference and methods used are typically:

- Cross section limits: CL_s, Bayesian, (profile likelihood)
- "Significance" (hypothesis testing): via tail distribution of a test statistic
- "Measurement" (point and interval estimation): maximum likelihood method, Bayesian, Feldman-Cousins

Note: The statistical model is independent of the method you want to apply, so think of model building and applying the statistical method as two completely separate stages of the analysis.

https://twiki.cern.ch/twiki/bin/view/CMS/SWGuideCMSDataAnalysisSchoolStatisticsFall2012
(cmsdas2013.desy.de → Exercise Twiki → Short Exercises 2013, 11. RooStats)
Rough session plan:

1. CombinedLimit (up to section 5)
2. RooFit and RooStats (sections 5 and Y)
3. RooStats and HistFactory (the rest of Y and Z and W)

Go with your own speed, and ask questions any time you need help.
Please use your school account on the provided machine (nafhh-cmsYY.desy.de).

Contents

Hypothesis Testing

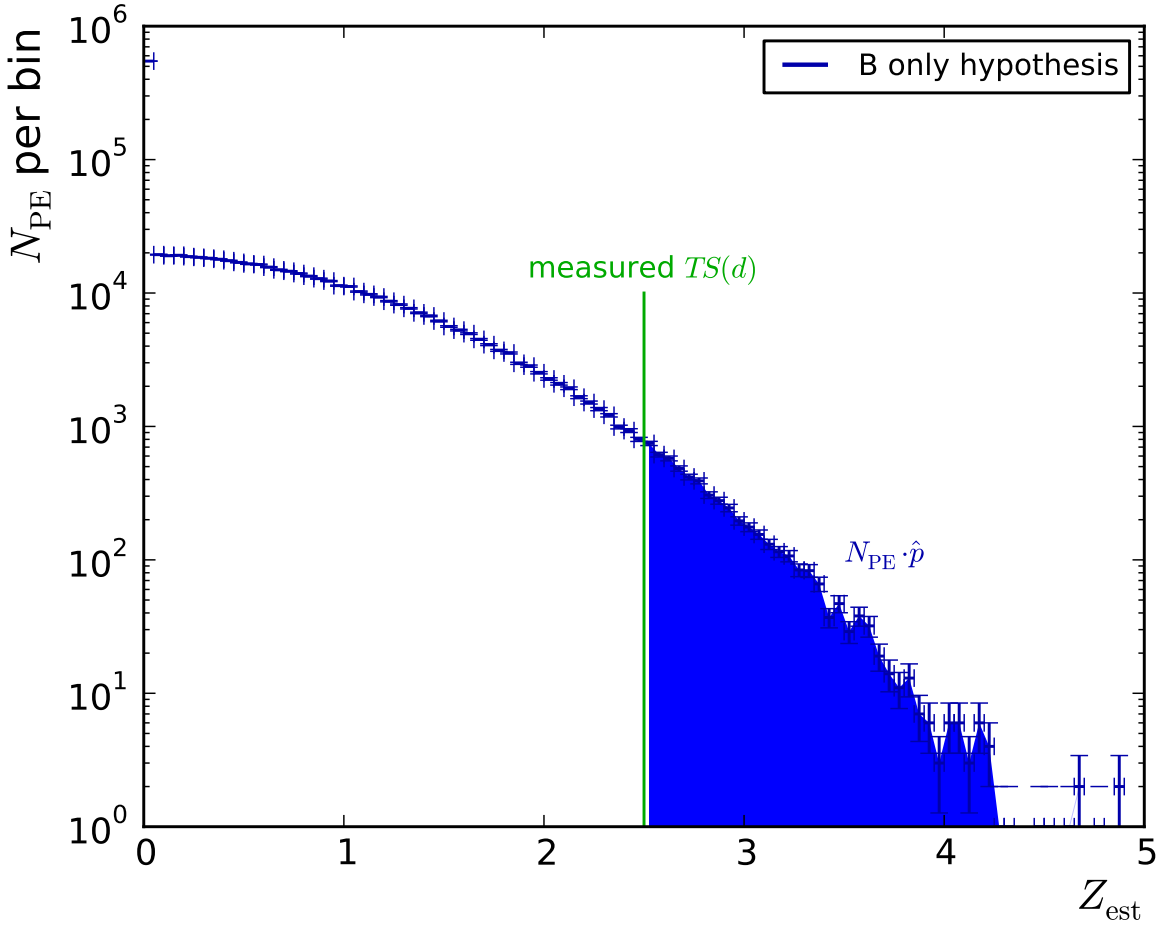
Hypothesis testing is a formal procedure to decide between different *hypotheses*. It tests the "null hypothesis" (only background) against the "alternative hypothesis" (signal + background). The result is a p-value, which is the *probability to observe a dataset at least as "signal-like" as the one observed, assuming the background-only hypothesis is true*.
"Signal-like" above is undefined, and several possibilities exist to reduce the dataset to a single number that meaures how "signal-like" the dataset is. This quantity is called *test statistic*; a common choice is the likelihood ratio:

$$TS = \frac{\max_{r,\nu} L(r, \nu|\text{data})}{\max_{\nu} L(r = 0, \nu|\text{data})}$$

It is large for "signal-like" data and small ($\mathcal{O}(1)$) for background.

Hypothesis Testing with Toys

In practice, the TS distribution for background-only has is measured by throwing toy datasets and calculating TS for each toy dataset; p is the fraction of toys with a value larger than the observed TS value.



Hypothesis Testing: Misc

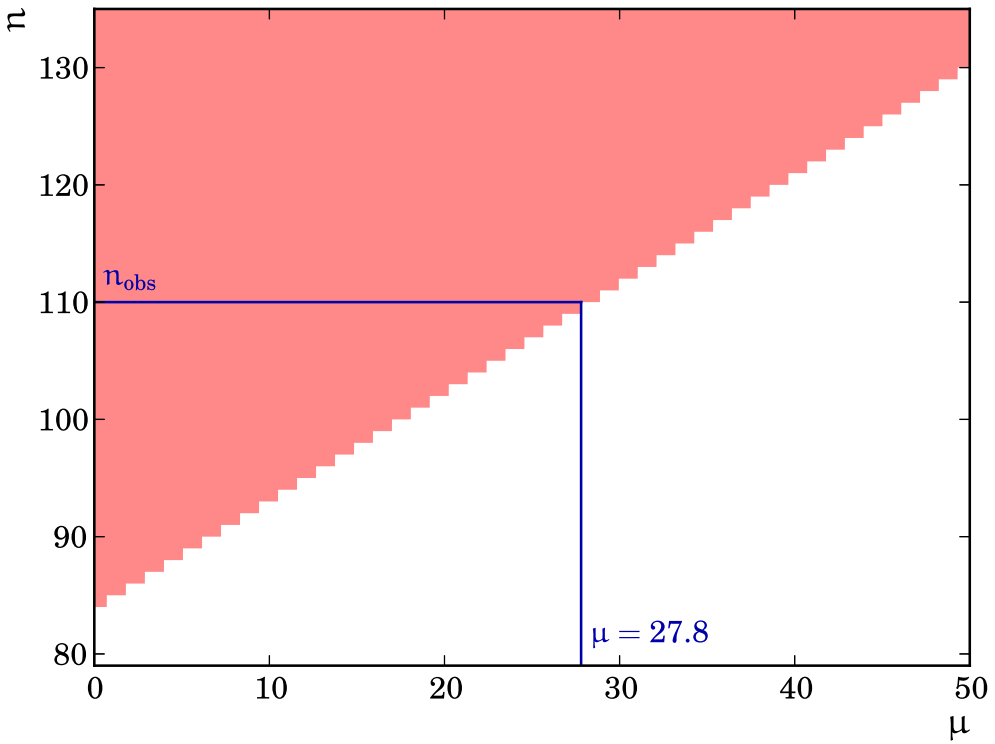
- The test statistic definition is not unique, one can use different variations (see also the "combine" manual).
- The number of toys should be large enough to measure the p-value accurately: If $p = N/N_0$ is small, then the error is approximately $\sqrt{N}$. As a rule of thumb, you usually want $N > 50$.
- The "number of sigma" is the same as a p-value. The only advantage is that you deal with numbers $\mathcal{O}(1)$ instead of 10^{-7}; see the PDG statistics section for the conversion rule and table.
- The p-value is the probability of an *error if the first kind*: reject null although it's true. Apart from that, another important concept is the *power*, which is the probability to reject the null hypothesis if the alternative is true.
- By throwing signal+background toys and evaluating their p-value, one can determine a (band of) "expected significance"

Frequentist Interval Construction

Frequentist interval construction can be seen as hypothesis test inversion: the null hypothesis is $r = r_0$ versus the alternative $r \neq r_0$. The interval for r is the set of values r_0 for which the null hypothesis *cannot* be rejected at some predefined level α (say, $\alpha = 0.05$).
The *confidence level* of such an interval construction is $1 - \alpha$ (95% in this case). It is the *coverage* of the procedure: The coverage is the probability that the true value for r is contained in the interval. In general, this probability is a function of the value for r. A sound interval construction guarantees that the coverage probability corresponds to the claimed confidence level.
In practice, this procedure consists of many hypothesis tests and requires some "scanning" in r_0. The interval construction can be visualized as a belt construction.

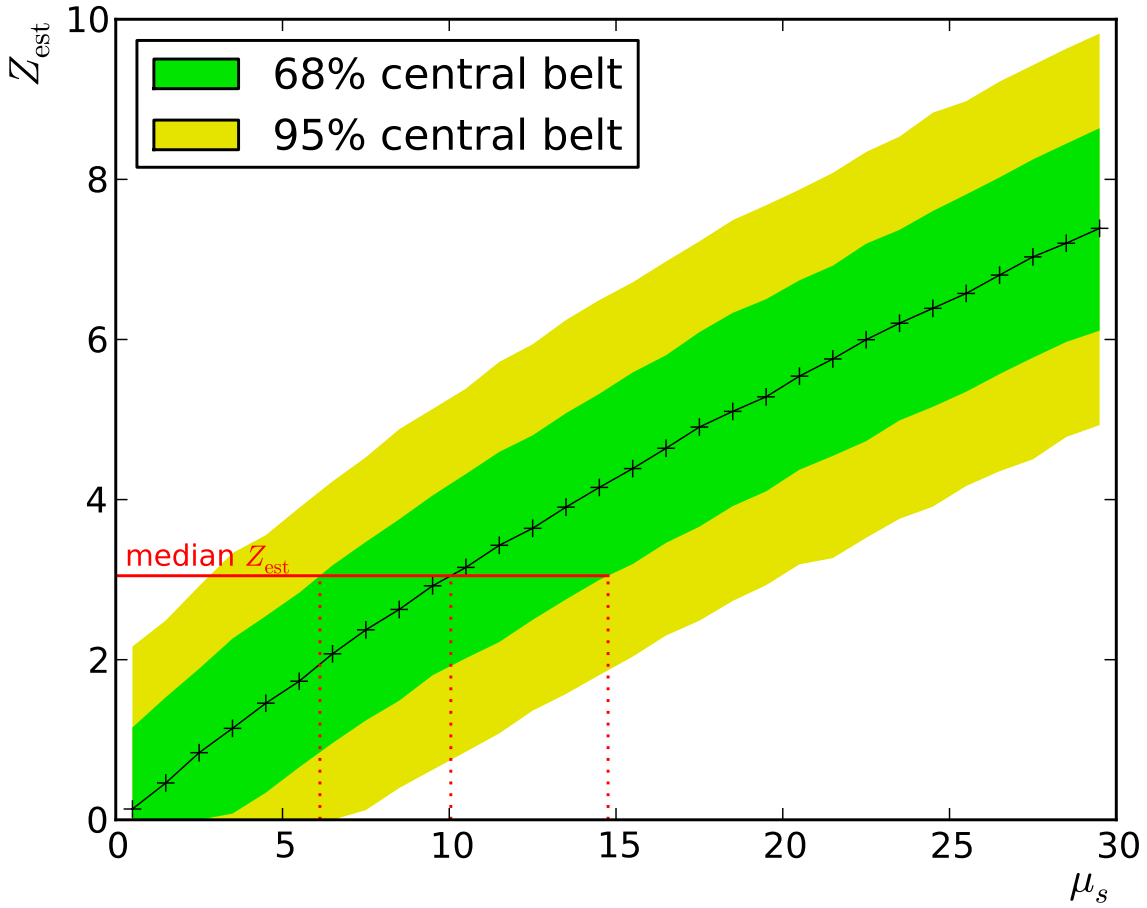
Belt

The belt is a visualization of the non-excluded points in the hypothesis tests, in practice done via toys:



$\rightsquigarrow$ Requires an *Ordering rule* to be well-defined.

Belt II


68% central belt
95% central belt
median $Z_{\rm est}$
$Z_{\rm est}$
μ_s
0
2
4
6
8
10
0
5
10
15
20
25
30

Interval Construction: Misc

- The same caveats as for hypothesis testing apply: the *TS* definition is not unique, ...
- In addition, there is arbitrariness in the *ordering rule*, i.e. of which points in the plane to include in the band (extreme case: upper vs. lower limits).
- The coverage property is guaranteed by construction: pick any value for the true value of r, r_t. The probability to measure a test statistic value which is within the belt is $1 - \alpha$, by construction of the belt. Exactly for those test statistic values, the value r_t will be within the resulting interval. This argument can be repeated for any value r_t. Therefore, the coverage is $1 - \alpha$.
- Sometimes, discreteness of data prohibits exact coverage, in this case it is customary to be "conservative", i.e. to increase the coverage for some points of r_t.

Contents

Statistics recommendations/requirements at CMS

- Methods: For Higgs, EXO and SUSY limits, there is an agreement between ATLAS and CMS to use the CL_s method; Bayesian is allowed in some circumstances. See https://hypernews.cern.ch/HyperNews/CMS/get/physics-announcements/1616.html (Or via the CMS Statistics Committee Twiki Site in "News").
- Tools: There is no requirement about using specific tools, but don't reinvent the wheel unless you have a good reason. To my knowledge, more widely used are RooStats (esp. combine), LandS, theta.

General Advise

- Always try to write down the statistical model in as much detail as possible.
- Some methods make simplifying assumptions, e.g. assume a Gaussian distribution although this might only be valid for large event numbers. Be aware of these limitations; for example: a very small efficiency derived from MC should not be modeled with a Gaussian. If possible, try to write down the "full", correct model.
- Remain skeptical about results of statistical tools, especially "simple" ones such as `roostats_cl95.C`.
  - Repeat it with different random seeds / more toys.
  - Make consistency checks (e.g. coverage test)
  - Use alternative tools
- If in doubt, resist the temptation to guess (it might take a lot time later in the review to fix); read or someone who might know more (e.g. via `hn-cms-statistics@cern.ch`).

Optimizing Analyses

Optimizing analyses for highest sensitivity is non-trivial. Some things to keep in mind:

- Never optimize using data; use MC.
- Try to optimize the actual result. For example, do not use $S/\sqrt{B}$ if systematic uncertainties are large; note that in general, optimizing for discovery is different as optimizing for limits / measurement.
- It can be beneficial to divide the dataset to channels which differ in some aspects important for the sensitivity, such as background level/composition, signal resolution/purity, and then combine the channels (example: see $H \to \gamma\gamma$). However, you usually also have to keep a minimum amount of events (e.g., for background estimation from sideband or from MC).