Statistics for HEP
Hands-on Tutorial #2

Nicolas Berger (LAPP)

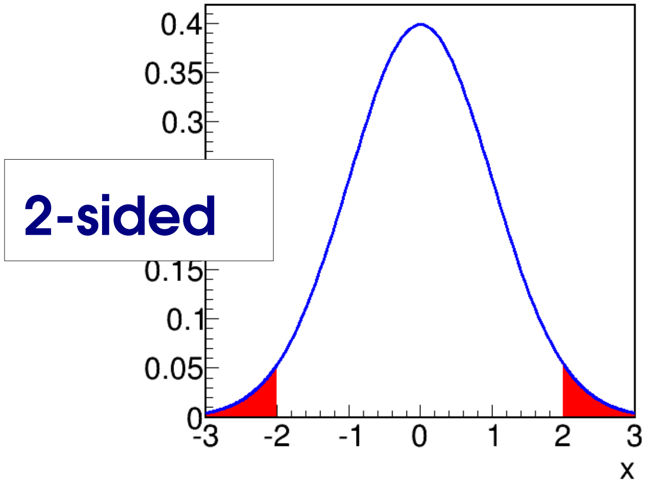
Introduction

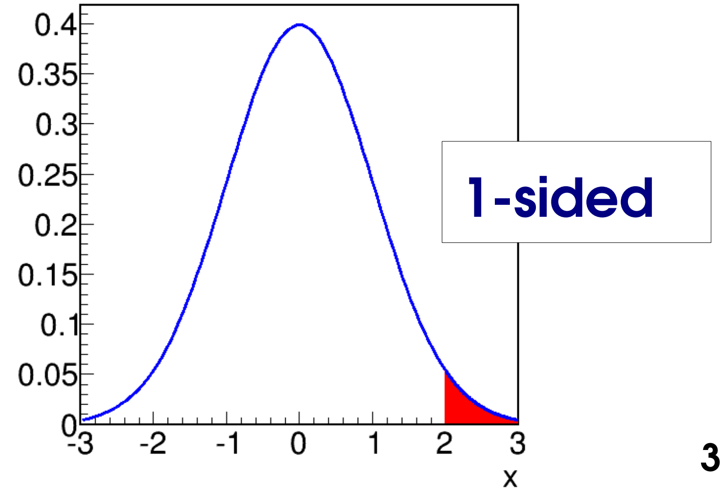
- We will use the same setup as the previous tutorial to compute significances and upper limits
- Two main examples
  - Gaussian S+B measurement
  - $H \rightarrow \gamma\gamma$ – like setup
- Some knowledge of the previous tutorial is assumed.
  - A lot of the code from yesterday's tutorial can be reused.
  - If needed, please have a look at the slides here:

    https://indico.in2p3.fr/event/10777/contribution/33/material/slides/0.pdf
- Please be careful when cut-and-pasting from the slides, as some characters don't seem to carry over properly (instead, copy from the solution macros on the last slides)

Gaussian Quantiles

- The function **`ROOT::Math::gaussian_cdf(x)`** gives the integral of a standard Gaussian (x_0=0, σ=1) from -∞ to x.
- The function **`ROOT::Math::gaussian_quantile(p)`** gives the reverse : the point z such that the integral from -∞ to x is p.
- **Exercise 9:**
  - Find the two-sided p-value for 1σ, 3σ and 5σ
  - Find the number of sigmas corresponding to a two-sided p-value of 10% and 5%
  - Find the number of sigmas corresponding to a 1-sided p-value of 5%
- Reminder :





chi2 Quantiles

- The function **`ROOT::Math::chisquared_cdf(x, 1)`** gives the integral of a standard chi2 distribution from 0 to x.
- The function **`ROOT::Math::chisquared_quantile(p, 1)`** gives the reverse : the point z such that the integral from 0 to x is p.

- **Exercise 10**
  - A variable has a chi2 distribution if it is the square of a Gaussian-distributed variable (like $q_0 \sim s^2$). To check that this is true, find the chi2 values corresponding to 10% and 5% p-values, using e.g.
    `ROOT::Math::chisquared_quantile(0.10, 1)`
  - Check that these chi2 values are the squares of the Gaussian quantities on the previous page.

Gaussian S+B measurement

- If you have 2 variables x and y, you can define their sum a=x+y:
  `RooAddition a("a", "", RooArgList(x,y))`
- **Exercise 11**
  - Set up
    - A variable n=0 (range 9000 - 11000)
    - A variable s=0 (range 0-500)
    - A variable b=10000 (fixed)
    - A variable sigma=100
    - A Gaussian PDF G(n, s+b, sigma)
    - a dataset with 1 event at n=10200
  - Alternatively, the setup is here:
  - Plot the PDF and the data. To be able to see the PDF, use
    `g.plotOn(p, RooFit::Normalization(100))`

Gaussian S+B measurement

- **Exercise 12**
  - Setup the PDF and data as exercise 11
  - Fit the PDF to the data (`g.fitTo(*d)`)
  - Check the central value and error on s (`s.getVal()`, `s.getError()`)
  - Check that these values correspond to what we expect:
    - S = N-B
    - error on s = the value of "sigma" in the model (=100) (68% CI interval : (S – sigma, S + sigma))

Gaussian S+B measurement : Discovery

- **Exercise 13**
  - Setup the same PDF and data as exercise 11
  - Define a NLL variable:

    ```
    RooNLLVar nll("nll", "", g, *d);
    ```
  - Fit the PDF to the data, so that s is set at its best-fit value
  - Get the value of the NLL (`nll.getVal()`) : this is $\lambda(\hat{s})/2$
  - Set s = 0 (`s.setVal(0)`). Get the value of the NLL again: this is now the value for s = 0, i.e. $\lambda(0)/2$
  - Compute $q_0 = \lambda(0) - \lambda(\hat{s})$
  - Compute the significance as $Z = \sqrt{q_0}$
  - Use the values from the previous exercise to compute the significance in the Gaussian approximation, $Z = S/\delta S$, compare to the value above.

Gaussian S+B measurement : Discovery

- **Exercise 14**
  - Run the same code as exercise 13, but with n=10000. Before computing the results, try to predict the values of S, δS, q_0 and Z.
  - Same with n=10500.

Gaussian S+B measurement : Limit

- **Exercise 15**
  - Setup the same PDF as exercises 11-14
  - Use n=10050 as the data
  - Fit the Gaussian to the data, check the values of S and δS.
  - Compute the 95% UL as $s+1.96*\delta s$
  - Note the value of the NLL at the best-fit s, i.e. $\lambda(\hat{s})/2$
  - Set s = 250.
    - Get the new value of the NLL, i.e. $\lambda(250)/2$
    - Compute $q_s = \lambda(s) - \lambda(\hat{s})$ for s=250
    - Compare with 3.84 (see exercise 10 for the value) to figure out if s=250 is rejected
  - Repeat with other values of s to estimate the value corresponding to 95% exclusion.
  - Compare with $s+1.96*\delta s$ formula above

Shape Analysis Discovery and Limits

- **Exercise 16**
  - Setup the shape analysis, as in the previous tutorial, with mH set to constant (`mH.setConstant();`). You can reuse your previous code, or the one here:

http://nberger.web.cern.ch/nberger/IDPASC/Exercises/shape_setup2.C

  - Generate 10000 events with s=150
  - Plot the data and the PDF.
  - Repeat exercises 11-15:
    - Get the best-fit s and its error
    - Estimate the significance ($Z=s/\delta s$) and the 95% upper limit ($s + 1.96\ \delta s$) in the Gaussian approximation
    - Note the NLL at the best fit, compute the NLL at s=0, evaluate the significance from q_0.
    - Compute the NLL at various s values, estimate the limit using q_s values.

Solutions

http://nberger.web.cern.ch/nberger/IDPASC/Exercises/exercise9.C

http://nberger.web.cern.ch/nberger/IDPASC/Exercises/exercise10.C

http://nberger.web.cern.ch/nberger/IDPASC/Exercises/exercise11.C

http://nberger.web.cern.ch/nberger/IDPASC/Exercises/exercise12.C

http://nberger.web.cern.ch/nberger/IDPASC/Exercises/exercise13.C

http://nberger.web.cern.ch/nberger/IDPASC/Exercises/exercise14.C

http://nberger.web.cern.ch/nberger/IDPASC/Exercises/exercise15.C

http://nberger.web.cern.ch/nberger/IDPASC/Exercises/exercise16.C