Fakultät für Physik 18.12.2008
Prof. Dr. M. Feindt, Prof. Dr. G. Quast,
Dr. G. Schott, Dr. T. Kuhr,
D. Piparo, F. Wick

Praktikum zu
Moderne Methoden der Datenanalyse

Exercise 6: Estimation of upper limits and hypothesis testing

We'll consider in this exercise Poisson processes. In those a measurement of the number of detected events is distributed according to the probability function:

$$P(n|\mu_t) = \frac{\mu_t^n e^{-\mu_t}}{n!}, \tag{1}$$

where n is the number of detected events and μ_t the true (or expected) number of events. In the presence of signal and background the expected number of events is $\mu_t = \mu_{t,S} + \mu_{t,B}$.

1 Classical (frequentist) approach

Let us assume the expected number of background events negligible ($\mu_{t,B} = 0$). Using the classical approach, we will compute a 68% confidence interval on $\mu_t = \mu_{t,S}$ if the measured number of event is $n_0 = 3$.

- Find $\mu_1 < n_0$ and $\mu_2 > n_0$ such that:

$$\sum_{n=n_0}^{\infty} P(n|\mu_1) = 0.16, \tag{2}$$
$$\sum_{n=0}^{n_0} P(n|\mu_2) = 0.16. \tag{3}$$

- Compute the 90% confidence level upper limit and lower limit on $\mu_{t,S}$. The strategy consists in using the first formula above alone.
- Compare to table 32.3 in PDG (http://pdg.lbl.gov/2008/reviews/statrpp.pdf).

2 Likelihood approach

The likelihood function for a Poisson process, supposing one single measurement, is:

$$L(n_0|\mu_t) = \frac{\mu_t^{n_0} e^{-\mu_t}}{n_0!}. \quad (4)$$

where n_0 is the number of measured events.

- Draw the $-2\ln L$ curve as function of μ_t, performing a scan over a significative range of values. Where is the minimum $-2\ln L_{min}$ of this curve?
- The 68% confidence level confidence interval boundaries correspond to points where $2 \cdot \Delta \ln L = 2 \cdot (-\ln L + \ln L_{min})$ is 1. Where are they?
- The 90% confidence level upper limit correspond to the point with $\mu_t > n_0$ where $2 \cdot \Delta \ln L$ is 1.28. What is the upper limit in this case?
  To translate a CL into the proper $\Delta \ln L$, you can use ROOT:

  $$2 \cdot \Delta \ln L = \sqrt{2} \cdot TMath :: ErfInverse(2 \cdot CL - 1) \quad (5)$$

  Check that for $CL = 0.90$, $2 \cdot \Delta \ln L = 1.28$. For more details see these lectures `http://www.hef.kun.nl/~wes/stat_course/statist_2002.pdf`, in particular chapter 8.4.

3 Bayesian approach

The Bayesian posterior probability $P(\mu_t|n_0)$ is given by the Bayes theorem:

$$P(\mu_t|n_0) = \frac{L(n_0|\mu_t)\ P(\mu_t)}{\int_{\text{all}\mu_t} L(n_0|\mu_t)\ P(\mu_t) d\mu_t}. \quad (6)$$

$P(\mu_t)$ is called the prior probability on μ_t and describe our prior belief about the distribution of this parameter. We'll try 2 priors:

- $P(\mu_t)$ constant for $\mu_t > 0$ and null otherwise,
- $P(\mu_t)$ proportional to $1/\mu_t$ for $\mu_t > 0$ and null otherwise.

Now:

- Compute and draw the posterior probability in both cases.
- What are the 90% credibility upper and lower limits with this method (with each of the 2 prior distributions)?

Finally: compare the upper and lower limits obtained with the 3 methods.

4 Classical upper limits in presence of background

Now $\mu_{t,B}$ is not anymore negligible.

- Compute with the classical method the 90% confidence level upper limits on $\mu_{t,S}$ as function of $\mu_{t,B}$. The convention is that one substracts the number of background events from the limit on $\mu_{t,S}$ obtained with no background events. Which is the inconvenient of this procedure?
- Make a plot of those limits if $n_0 = 0$, $n_0 = 1$, $n_0 = 2$, ... You can draw in the same canvas one curve for every value of n_0.
- Normalise CL_{SB} by CL_B and make the plot again using CL_S instead of CL_{SB} as it was done in exercise 1. CL_{SB} and CL_B are defined below. CL_{SB} measures the compatibility of the experiment with the signal plus background hypothesis, while CL_B the compatibility with the background only hypothesis.

$$CL_{SB} = \sum_{n=0}^{n_0} P(n|\mu_{t,S} + \mu_{t,B}), \tag{7}$$

$$CL_B = \sum_{n=0}^{n_0} P(n|\mu_{t,B}), \tag{8}$$

$$CL_S = CL_{SB}/CL_B. \tag{9}$$

5 Signal significance

We know the signal is expected to be $\mu_{t,S} = 15$ and the background $\mu_{t,B} = 40$ while the measurement is still $n_0 = 56$.

- What is the probability to measure n_0 or less events if you expect only background?
- What is the probability to measure more than n_0 events if you expect signal and background?
- What is the corresponding significance (i.e. the "number of sigmas" of a Gaussian distribution corresponding to this probability)? You can use the formula $p = \int_s^{+\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}} dt$. ROOT offers you also the function Erf in the TMath namespace.
- One among the many significance estimators is the so called S_{L2} and it has many desirable features. It is defined as:

$$S_{L2} = \sqrt{2 \ln Q}, \; Q = L_{SB}/L_B \tag{10}$$

  where L_{SB} and L_B are the likelihood in the signal+background and background-only hypotheses. This means calculated using the same dataset but with the signal+background and background only models respectively.

6 Application with RooStats (optional)

Since the version 5.21.06, ROOT includes RooStats which is a package to ease the statistical studies. RooStats is based on the RooFit Toolkit for data modelling `http://root.cern.ch/root/doc/RootDoc.html`. A macro written using RooFit and RooStats looks like the one attached below. It performs hypothesis testing for you. The signal model is a Gaussian centred at zero, while the background is flat. The signal yield is 15, while the background yield is 40 events with a Gaussian systematic uncertainity.

```
{
  using namespace RooFit;
  using namespace RooStats;

  double my_sigma = 2;
  double my_sig_yield = 15;
  double my_bkg_yield = 40;
  double my_bkg_yield_uncertainty = my_bkg_yield*.25; // absolute uncertainty
  bool use_systematics=false;

  // build the models for background and signal+background
  RooRealVar x("x","",-3,3);
  RooArgList observables(x); // variables to be generated

  // gaussian signal
  RooRealVar sig_mean("sig_mean","",0);
  RooRealVar sig_sigma("sig_sigma","",my_sigma);
  RooGaussian sig_pdf("sig_pdf","",x,sig_mean,sig_sigma);
  RooRealVar sig_yield("sig_yield","",my_sig_yield,0,(my_sig_yield+my_bkg_yield)*2);

  // flat background (extended PDF)
  RooRealVar bkg_slope("bkg_slope","",0);
  RooPolynomial bkg_pdf("bkg_pdf","",x,bkg_slope);
  RooRealVar bkg_yield("bkg_yield","",my_bkg_yield,0,(my_sig_yield+my_bkg_yield)*2);
  RooExtendPdf bkg_ext_pdf("bkg_ext_pdf","",bkg_pdf,bkg_yield);

  // total sig+bkg (extended PDF)
  RooAddPdf tot_pdf("tot_pdf","",RooArgList(sig_pdf,bkg_pdf),RooArgList(sig_yield,bkg_yield));

  // build the prior PDF on the parameters to be integrated
  // gaussian contraint on the background yield
  RooGaussian bkg_yield_prior("bkg_yield_prior","",bkg_yield,RooConst(bkg_yield.getVal()),RooConst(my_bkg_yield_uncertainty));
  RooArgSet nuisance_parameters(bkg_yield); // variables to be integrated

  //*********************************************************************//

  // run HybridCalculator on those inputs
  HybridCalculator myHybridCalc("myHybridCalc","HybridCalculator example",tot_pdf,bkg_ext_pdf,observables,nuisance_parameters,bkg_yield_prior);

  // here I use the default test statistics: 2*lnQ (optional)
  myHybridCalc.SetTestStatistics(1);

  // run 1000 toys with gaussian prior on the background yield
  HybridResult* myHybridResult = myHybridCalc.Calculate(3000,use_systematics);

  // Set the test statistic value. The value is arbitrary. We focus on the distributions
  myHybridResult->SetDataTestStatistics(0);

  // nice plot of the results
  HybridPlot* myHybridPlot = myHybridResult->GetPlot("myHybridPlot","Plot of results with HybridCalculator",100);
  myHybridPlot->Draw();

}
```

In the end it's just C++, you know how to handle it. To make life easier you can download the macro here: */home/staff/tkuhr/Praktikum/Exercise6*

What happens to the test statistic distributions if:

- The Gaussian is rather flat ($\sigma = 2$)? The Gaussian is narrow ($\sigma = 0.3$)?
- The systematic on the background is 50%?