Exercises for the lecture „Moderne Methoden der Datenanalyse"

Prof. Dr. S. Hansmann-Menzemer, M. Schiller
Physikalisches Institut der Universität Heidelberg

June 29 2010

Exercise 8: Goodness of Fits

How can we judge if a fit describes the fitted data well? A simple example will show that a good $\chi^2/n.d.f.$ is not always sufficient for that. However, there are other possibilities to guaranty the goodness of a fit. This can especially be a problem if we have a large background and look for a small signal. Do we really find a signal or is it just a statistical fluctuation of the background?

Exercise 8.1

Assume we measure something that is constant. e.g.:

x	1	2	3	4	5	6	7	8
y	2	2	2	2	4	4	4	4
σ_y	1.06	1.06	1.06	1.06	1.06	1.06	1.06	1.06

Table 1: *Measurements of a constant parameter.*

Make a plot of x versus y with `TGraphErrors` and make a χ^2-fit. Calculate $\chi^2/n.d.f.$ with n.d.f=number of degrees of freedom (= number of available measurements minus number of free parameters) using the functions `GetChisquare` and `GetNDF`. You will yield a $\chi^2/n.d.f.$ very close to 1. However, one can tell that this is not exactly a good fit. It demonstrates that just having a good $\chi^2/n.d.f.$ is not enough to be convinced of the goodness of a fit.

Exercise 8.2

Two simple checks of fits are very common: the pull distribution and the sign check. The pull distribution is a distribution of

$$\frac{y_i - f(x_i)}{\sigma_i}$$

where (x_i, y_i) are the data points with σ_i the error on y_i and f is the fit function. For a good fit, this should yield a Gaussian with a mean of 0 and a σ of 1.

The sign check depends on the sign of $y_i - f(x_i)$. Here one calculates the sum of the sign changes. When the sign between two consequitive x_i values remains the same, +1 is added, otherwise -1 is added. When the points are randomly distributed around the model prediction, the sum should be close to 0.

Generate 10,000 random numbers according to a Gaussian distribution with a mean of 0 and a sigma of 1.4. Store the values in a histogram with 500 bins between -5 and 5. Calculate the sign sum and plot the pull distribution. Skip empty bins.

Exercise 8.3

In particle physics, especially in searches for new particles, the main problem is that there can be relatively small signals over a large background. We will try to find a small signal amidst background by a combined fitting technique.

Download `ex83.root` from the webpage. You will find there a histogram of the invariant mass of two particles. There are also a Monte Carlo simulation of the background. Let us assume that theorists have predicted that the new particle has a mass around 14.

First, one need to parametrise the background. This can be done well using a Landau distribution. Try to fit the background by using a Landau.

After extracting the parameters describing the background, one can focus on the signal distribution. The signal is fitted with the sum of the expected signal distribution (here a Gaussian around 14) together with the background parametrisation. The luminosity in the background Monte Carlo sample is twice the luminosity in the data. Try to extract the number of particles found by making this combined fit. What is the width and the mass of the particle? How many standard deviations is the observed signal from zero?

Exercise 8.4

A measure of how well a theory fits the data is to form the following quantity,

$$\Delta = \sum_{i=1}^{N_{bins}} \frac{(N_i^{data} - N_i^{theory})^2}{\sigma_i^2}$$

For cases where the errors are approximately Gaussian in nature or, equivalently, where the contents of each bin are greater than about 10, Δ is termed χ^2 and is distributed according to the following analytical probability density function (p.d.f),

$$f(\chi^2; n) = \frac{2^{-n/2}}{\Gamma(n/2)} \chi^{n-2} e^{-\chi^2/2}$$

where n = the number of degrees of freedom (n.d.f).

The 'goodness-of-fit' is then best expressed by forming the probability of having a data set that, by chance, fits the theory as well (or worse) than the data set you have, namely,

$$P = \int_{\Delta}^{\infty} f(\Delta') d\Delta'$$

The file `ex84_data.dat` contains a histogram of data values: The first two columns are the bin boundaries, i.e. the values at the low and high edge of the bin, and the third column gives the number of entries n_i in bin i. We will treat these entries as Poisson random variables ie they are assumed to derive from a Poisson distribution of mean n_i and hence have a standard deviation of $\sqrt{n_i}$. The files `ex84_theory1.dat` and `ex84_theory2.dat` are theoretical predictions of the data and in exactly the same format. Note that lines beginning with a character '!' are comments and should be ignored.

- (a) Download the template. It contains a function `readFile` which reads in the data of one file, filling arrays with bin edges and contents. It also produces a histogram of the data read.
- (b) Determine the Δ value of the data set for each of the two theories.
- (c) Because the data set contains many bins with only a few or no entries, one does not expect Δ to follow the χ^2-distribution. Write a Monte Carlo program to determine the actual p.d.f. for the two hypotheses `ex84_theory1.dat` and `ex84_theory2.dat` (hint: generate N_i^{data} from a random Poisson distribution with mean equal to the value of the theory prediction in that bin. You can generate them via `x=gRandom->Poisson(nu)`. Compute Δ and record the value in a histogram. Repeat the experiment many times, e.g. 10^6.)
- (d) What are the P-values for the two theories? Which theory best describes the data? Use the histogram generated in (c) to obtain the actual p.d.f.
- (e*) Compare the actual p.d.f.s obtained with the χ^2 distribution for the same n.d.f. (N.B. Use `TMath::Gamma()`, with argument ($n_{dof}/2.$) of type double.). What would the P-values be if one assumes a χ^2 distribution?