Kirchhoff-Insitut für Physik
Physikalisches Insitut

Winter semester 2013-14
KIP CIP Pool (1.401)

Exercises for Statistical Methods in Particle Physics

http://www.physi.uni-heidelberg.de/~nberger/teaching/ws13/statistics/statistics.php

Dr. Niklaus Berger (nberger@physi.uni-heidelberg.de)
Dr. Oleg Brandt (obrandt@kip.uni-heidelberg.de)

Exercise 8: Goodness of fit tests

2. December 2013
Hand-in solutions by 14:00, 8. December 2013

Please send your solutions to obrandt@kip.uni-heidelberg.de by 8.12.2013, 14:00, punctually. Make sure that you use *SMIPP:Exercise08* as subject line. If plots are requested, please include print statements to produce pdf files in your code, and provide the plots separately. Please add comments to your source code explaining the steps. Test macros and programs before sending them off...

In experimental particle physics, one is often confronted with the question: does the simulation of signal and background describe the data adequately? Two most commonly used tests for this is the χ^2/D.o.F. based goodness-of-fit test, and the Kolmogorov-Smirnov test. While these tests can be very helpful and provide a solid criterion to decide about the quality of the signal and background predictions, such tests have to be taken with a pinch of salt, and one has always to keep in mind that scientific judgement by-eye is at least equally important.

1 The limitations of the χ^2/D.o.F. test

Assume we measure a quantity y which, from first principle considerations, is known to be constant versus variations in another quantity x, and obtain values as shown in Tab. 1.
Generate a plot of x versus y with `TGraphErrors` and make a χ^2-fit (`ex_8_1.pdf`). Calculate χ^2/D.o.F. with D.o.F=number of degrees of freedom (= number of available measurements minus number of free parameters of the model) using the methods `GetChisquare()` and `GetNDF()`. This will yield a χ^2/D.o.F. which is very close to unity. Also obtain the χ^2-probability, i.e. the p-value of the null hypothesis, which in this case is "the data is described adequately by a constant fit". The p-value is defined as the integral of the corresponding χ^2/D.o.F.-distribution from the measured χ^2/D.o.F. value to $+\infty$, and is thus the probability to obtain a similar χ^2/D.o.F. value or higher, i.e. it is a measure for the probability of the outcome, all this assuming the null hypothesis. The χ^2-probability can be obtained via `TMath::Prob()` with the χ^2 and D.o.F. parameters. Also the χ^2-probability will be quite high.
By eye one can tell that the above is not exactly a good fit. This demonstrates that just having a good χ^2/D.o.F. is not enough to be convinced of the goodness of a fit, and one's scientific judgement has always to remain alert (`ex_8_1.C`).

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

Tabelle 1: Measurements of y versus x.

2 Goodness-of-fit for data and MC comparisons

In particle physics, especially in searches for new particles, the main problem is that there can be relatively small signals over a large background. However, before this is done, one needs to convince oneself that the signal and background simulations describe the model adequately.
Download `ex_8_2_input.root` from the course webpage. You will find three histograms of the invariant mass of two particles (to be precise photons): `hdat` for data; and Monte Carlo simulations of the signal (`hsig`) and background (`hbgr`). Let us assume that theorists have predicted that the new particle has a mass around 125 GeV...
Plot the data distribution as data points, together with the histograms of signal and background stack on top of each other (`THStack`). Note that for simplicity in this part of the exercise the signal and background predictions are already scaled to the integrated luminosity in data. Is the data adequately described by the simulations (by eye)? Provide the χ^2, D.o.F., and the χ^2-probability for the consistency of the signal+background histogram with data, which can be done using the `TH1::Chi2Test()` method. Similarly, provide the Kolmogorov-Smirnov probability (`TH1::KolmogorovTest()`). What does it tell you? Print all values on the plot where you compare the signal+background prediction (in different colors) with the data points (`ex_8_2_sigbgr.pdf`).
Similarly, provide an analogous plot and statistical benchmarks where you consider only the background (`ex_8_2_bgr.pdf`). Can you draw any conclusions?
As a final step, imagine the signal model was different, and the predicted mass was at 140 GeV (`hsig140`). Provide the analogous plots and statistical figures as above. What do they tell you? (`ex_8_2.C`).

3 Fit of signal+background models to data

Now that we have convinced ourselves in Problem 8.2 that the data is described well by the signal and background simulations, we will try to find the signal amidst background by a combined fitting technique.
To start, download `ex_8_3_input.root` from the course webpage, which contains three histograms: `hdat`, `hsig`, and `hbgr` (note that this time the signal and background predictions are *not* scaled to the same integrated luminosity as the data!).
First, one needs to parametrise the background. This can be done well using an exponential distribution. Try to fit the background by using an exponential, and display the fit result together with the histogram (`ex_8_3_bgr.pdf`).
Second, the signal needs to be parameterised, which in this case can be well approximated by a Gaussian, and not a Breit-Wigner, as the resolution is dominated by the detector. Again, display the results of the fit in the histogram (`ex_8_3_sig.pdf`).
After extracting the parameters describing the signal and background, one can focus on the data distribution. The data is fitted with the sum of the expected signal distribution together with the background distribution. Note that both need to be scaled appropriately, as the size of the generated MC simulations is typically much larger than the size (integrated luminosity) of the data sample. Extract the number of signal particles by making this combined fit (`ex_8_3_dat.pdf`). What is the width and the mass of the signal particle? What is the signal fraction you find in data? How do the integrated luminosities of signal and background compare to the integrated luminosity in data? How many standard deviations is the observed signal from zero?
Provide also the χ^2, D.o.F., and the χ^2-probability for the signal+background fit to data (`ex_8_3.C`).

4 Extending the PMT from Problem 7 to 12 dynodes

As you have probably noticed, the number of photoelectrons at the last dynode for a 6-stage PMT from Problem 7 is with $o(1000)$ rather low, while $o(10000)$ and ideally $o(100000)$ or more more would be necessary for the electronic read-out and further signal processing. Therefore, we will extend the PMT in the configuration of Problem 7.2 (i.e. $\nu_1 = 6.0$, $\nu_{i!=1} = 3.0$) to the case of 12 dynodes ($\nu_{i!=1} = 3.0$) in total. You may have noticed that the computation time for Problem 7 was driven by the last stage due to exponential multiplication of the electron cascade. Therefore, it would be unpractical to try to simulate n_i for $i > 6$ using the straight forward method. Instead, to simulate $i = 7$ please take the output for $n_6 = n_{\text{out}}$ of Problem 7.2 for sufficiently large statistics (10000 should is a good ballpark number) filled into an *appropriately binned and ranged* histogram using the example solution code to Problem 7.2, and generate random numbers that follow this distribution. Proceed recursively to $i = 8$ in a similar manner, etc. Please provide the same output histograms as already requested in Problem 7.