Computing and Statistical Data Analysis
Problem sheet #7b

This problem sheet contains an extended version of exercise 7-2.

Exercise 2 (extended version): The number of events n observed in electron-positron collisions having particular kinematic properties can be treated as a Poisson variable with a mean of $s + b$, where s is the expected number of events from the signal process and b is the expected number from background. The likelihood function is therefore

$$L(s) = \frac{(s+b)^n}{n!} e^{-(s+b)} \ .$$

Suppose that $b = 3.9$ (known exactly) and we observe $n = 16$ events.

(a) Compute the p-value for the hypothesis that $s = 0$, i.e., no new process is contributing to the number of events. To sum Poisson probabilities, you can use the relation

$$\sum_{n=0}^{m} P(n;\nu) \ = \ 1 - F_{\chi^2}(2\nu; n_{\rm dof}) \,, \tag{1}$$

where $P(n:\nu)$ is the Poisson probability for n given a mean value ν, and F_{χ^2} is the cumulative χ^2 distribution for $n_{\rm dof} = 2(m+1)$ degrees of freedom. This can be computed using the ROOT routine `TMath::Prob` (which gives one minus $F_{\chi 2}$) or looked up in standard tables. If you have difficulty getting a program to return F_{χ^2}, you can simply carry out the sum of Poisson probabilities explicitly.

(b) Compute the corresponding significance Z.

(c) Find the p-value and significance based on the formula in the lecture where the Poisson distributed quantities are modeled as following a Gaussian distribution. Would one expect the approximation to be valid in this case?

(d) Find the p-value and significance using the formulas from the lecture for the case where one uses the (profile) likelihood ratio test and assumes its distribution is given by the formula valid in the large sample limit.

(e) Write a simple Monte Carlo program to simulate the observation of n under the background-only hypothesis a large number of times and compute the p-value by determining the fraction of experiments that give you as many events as found in the the real experiment (16) or more. Find also the corresonding significance Z.

You can use the program `SimpleMC.cc` as a starting point. To generate Poisson distributed random values, you can use the same `TRandom3` object to call the method `Poisson` (see the `TRandom3` class definition on the web). You should generate enough events so that the p-value is estimated with a reasonble statistical accuracy (say, 10%).

(f) In the same computer program, compute for each event the value of the likelihood ratio statistic

$$q_0 = -2 \ln \frac{L(0)}{L(\hat{s})}$$

for $n > b$ and $q_0 = 0$ otherwise. Check that the histogram of this statistic under the $s = 0$ hypothesis has approximately the expected form: a delta function at zero with weight 1/2, plus a chi-square distribution for one degree of freedom, also with weight 1/2. (Think of a simple way you can check this graphically.)

(g) Suppose we do not yet know n, but we want to know the expected (median) significance (i.e., the sensitivit) with which we would reject the $s = 0$ hypothesis if the true value of s is 13. Estimate the expected significance using both the MC program and also the formulae given in Thursday's lecture.