Problems for INFN School of Statistics
(G. Cowan; parameter estimation)

Exercise 1: This exercise provides an introduction to the class `TMinuit`, used in `ROOT` for function minimization. In this example we will use `TMinuit` to carry out a Maximum Likelihood fit where we minimize the quantity $-2 \ln L$. For more information on TMminuit see

`root.cern.ch/root/html/TMinuit.html`

First we will generate some data using a simple Monte Carlo program. Download, build and test the program `makeData` from the course website. `makeData` generates values according to an exponential distribution

$$f(x;\xi) = \frac{1}{\xi} e^{-x/\xi} \tag{1}$$

and writes the values to a file. It also produces a histogram of the values. Use this to generate and store, e.g., $n = 50$ values of x.

In a separate directory, download and build the program `expFit` from the course website. This program reads in the file of individual values provided by `makeData` and does a maximum likelihood fit of the parameter of the exponential pdf. Run both programs and make sure you understand what they are doing.

Now modify `makeData` so that it generates values according to the pdf

$$f(x;\alpha,\xi_1,\xi_2) = \alpha \frac{1}{\xi_1} e^{-x/\xi_1} + (1-\alpha) \frac{1}{\xi_2} e^{-x/\xi_2}; \;, \tag{2}$$

with $\alpha = 0.2$, $\xi_1 = 1.0$ and $\xi_2 = 5$. To do this, first generate a random number r uniform in $[0, 1]$. If $r < \alpha$, then generate x according to an exponential with mean ξ_1, otherwise use ξ_2. Run the program and save 200 individual values to a text file.

Now modify the program `expFit` so that it reads in the values and carries out an ML fit of the parameters α, ξ_1 and ξ_2. You will have to supply start values and "step sizes" for the parameters. Choose start values not too far (say, within a factor of two) to the true values used in `makeData`. For the step sizes you can take, e.g., 0.1.

Try running the program with the minimum and maximum values (in the arrays `minVal` and `maxVal`) set equal to zero; this is equivalent to having no bounds on the parameters. If the program runs into a region of parameter space that it shouldn't, e.g., $\xi_1 < 0$, then you can place appropriate bounds on the parameter values. In the end it is best to see if you can rerun the fit with improved guesses for start values but without any bounds on the parameters.

Modify the program so it makes a reasonable plot of the fit (extend the limit of the horizontal axis as appropriate). Find the ML estimators and their covariance matrix using the routines `mnpout` and `mnemat`. Determine as well the matrix of correlation coefficients.

2 For this problem go back to the original exponential fit with a single parameter. modify the program `makeData` so that it stores the values in a histogram with $N = 10$ bins from $0 \leq x \leq 5$.

Modify the program `expFit.cc` so that it reads in the histogram (you can use a `TH1D` object). The data for the are now the numbers of entries in each of the N bins, $\mathbf{n} = (n_1, \ldots, n_N)$.

To do an ML fit of the parameter ξ to the histogram, calculate the expectation value for each bin as

$$\nu_i(\xi) = n_{\text{tot}} \int_{x_{i,\min}}^{x_{i,\max}} f(x;\xi)\, dx = n_{\text{tot}} (e^{-x_{i,\min}/\xi} - e^{-x_{i,\max}/\xi}) \ . \tag{3}$$

If we treat the numbers of entries n_i as following a multinomial distribution, then the likelihood function is

$$L(\xi) = \frac{n_{\text{tot}}!}{n_1! n_2! \ldots n_N!} p_1^{n_1} p_2^{n_2} \ldots p_N^{n_N} \ , \tag{4}$$

where $p_i = \nu_i / n_{\text{tot}}$. Rewriting in terms of the ν_i gives the log-likelihood function (up to a constant)

$$\ln L(\boldsymbol{\nu}) = \sum_{i=1}^{N} n_i \ln \frac{\nu_i}{n_{\text{tot}}} \ . \tag{5}$$

This differs only by a constant (i.e., not depending on ξ) from the quantity

$$\chi^2_{\text{M}}(\xi) = -2 \sum_{i=1}^{N} n_i \ln \frac{\nu_i(\xi)}{n_i} \tag{6}$$

Rewrite the function `expFit` so that it minimizes the function $\chi^2_{\text{M}}(\xi)$.

Assuming the statistic $t = \chi^2_{\text{M}}(\hat{\xi})$ follows a chi-square distribution for $N - 2$ degrees of freedom (N bins minus one constraint for n_{tot} minus one fitted parameter), find the p-value to assess the goodness of fit from

$$p = \int_{t_{\text{obs}}}^{\infty} f_{\chi^2_{N-2}}(t)\, dt \tag{7}$$

(You can get the integral of the chi-square distribution from the root function `TMath::Prob`.)