Statistical Data Analysis
Problem sheet #3
Due Monday, 2 November 2020

Exercise 1: Consider the exponential p.d.f.,

$$f(x;\xi) = \frac{1}{\xi}\, e^{-x/\xi}\,, \quad x \geq 0\,. \tag{1}$$

1(a) [2 marks] Show that the corresponding cumulative distribution is given by

$$F(x) = 1 - e^{-x/\xi}\,, \quad x \geq 0\,. \tag{2}$$

1(b) [6 marks] Show that the conditional probability to find a value $x < x_0 + x'$ given that $x > x_0$ is equal to the (unconditional) probability to find x less than x', i.e.

$$P(x \leq x_0 + x' | x \geq x_0) = P(x \leq x')\,. \tag{3}$$

Exercise 2: Suppose a beam of particles is known to consist of charged pions and muons. For each particle in the beam we measure a variable t, whose distribution for pions (π) and muons (μ) is

$$f(t|\pi) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(t-\mu_\pi)^2/2\sigma^2}\,, \qquad f(t|\mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(t-\mu_\mu)^2/2\sigma^2}\,,$$

where $\mu_\pi = 0$, $\mu_\mu = 2$ and $\sigma = 1$. For each particle we want to test the hypothesis H_0 that it is a pion against the alternative H_1 that it is a muon. The critical region of the test is given by $t > t_{\rm c}$ where $t_{\rm c}$ is a given constant.

2(a) [6 marks] Suppose we want to have a test of size $\alpha = 0.05$. Illustrate where the critical region lies and what α means on a sketch of the pdfs $f(t|\pi)$ and $f(t|\mu)$ and show that $t_{\rm c}$ is numerically about 1.64.

2(b) [4 marks] Using the value of $t_{\rm c}$ from (a), find the power M of the test with respect to the alternative hypothesis that the particle is a muon. Indicate what the power corresonds to on the sketch of $f(t|\pi)$ and $f(t|\mu)$.

2(c) [4 marks] Suppose a sample of particles is known to consist of 99% pions and 1% muons. What is the purity of the muon sample selected by $t > t_{\rm c}$? Here, purity means the probability to be a muon given that the particle had $t > t_{\rm c}$ (i.e., it was rejected as a pion and thus selected as a muon candidate).

For this exercise you will need the cumulative Gaussian distribution and quantile, available e.g. from the python routines `norm.cdf` and `norm.ppf` from `scipy.stats`, from ROOT routines `TMath::Freq` and `TMath::NormQuantile` or from standard tables.

Exercise 3 (computing warm up [0 marks]): There is nothing to turn in for this exercise – it is just a warm-up exercise to ensure that you have your computing environment set up.

Starting with `simpleMC.py`, `simpleMC.ipynb` or `simpleMC.cc` from the course website, generate 10000 random values uniformly distributed between 0 and 1 and display the result as a histogram with 100 bins. (This is what simpleMC already does; you just need to ensure that you can run it.)