GGI 2017
Problems on Statistics (2)
G. Cowan / RHUL

Exercise 1 Consider the following two pdfs for a continuous random variable x that correspond to two types of events, signal (s) and background (b):

$$f(x|\mathrm{s}) = 2(1-x) ,$$
$$f(x|\mathrm{b}) = 4x^3 ,$$

where $0 \leq x \leq 1$. We want to select events of type s by requiring $x < x_{\mathrm{cut}}$ for a specified value of x_{cut}. This can be viewed as a test of the hypothesis b, whereby an event is selected if the b hypothesis is rejected. Suppose we want a test of size $\alpha = 10^{-4}$ (i.e., a background efficiency of 10^{-4}).

(a) Find x_{cut} and indicate the critical region of the test. Find the power M with respect to the hypothesis s (i.e., the signal efficiency) and evaluate numerically.

(b) Suppose the prior probabilities for events to be of types s and b are $\pi_{\mathrm{s}} = 0.001$ and $\pi_{\mathrm{b}} = 0.999$, respectively. Find the purity of signal events in the selected sample, i.e., the expected fraction of selected events that are of type s. Evaluate numerically.

(c) Suppose an event is observed with $x = 0.1$. Find the probability that the event is of type b and evaluate numerically.

(d) Again for an event with $x = 0.1$, find the p-value for the hypothesis that the event is of type b and evaluate numerically. Describe briefly how to interpret this number and why it is not equal to the probability found in (c).

(e) Suppose in addition to x, for each event we measure a quantity y, and that the joint pdfs for the s and b hypotheses are:

$$f(x,y|\mathrm{s}) = 4(1-x)y ,$$
$$f(x,y|\mathrm{b}) = 8x^3(1-y) .$$

Write down a statistic $t(x,y)$ for testing the hypothesis b which provides the highest power with respect to s for a test of a given size. Justify your answer.

Exercise 2: The number of events observed in high-energy particle collisions having particular kinematic properties can be treated as a Poisson distributed variable. Suppose that for a certain integrated luminosity (i.e. time of data taking at a given beam intensity), $b = 3.9$ events are expected from known processes and $n_{\mathrm{obs}} = 16$ are observed.

2(a) Compute the p-value for the hypothesis that no new signal process is contributing to the number of events. To sum Poisson probabilities, you can use the relation

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

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 $n_{\mathrm{dof}} = 2(m+1)$ 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.

2(b) Find the corresponding equivalent Gaussian significance Z and evaluate numerically.