Glen Cowan
RHUL Physics
14 December, 2011

Statistics Problems for the NIKHEF Onderzoekschool Subatomaire Fysica

Exercise 1: This is a problem from a recent University of London exam. Part (e) on Monte Carlo will not be in the lectures (but you may know anyway how to do this). The rest of the material is contained in lecture 1.

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

$$\begin{aligned} f(x|\mathrm{s}) &= 3(x-1)^2 \,, \\ f(x|\mathrm{b}) &= 3x^2 \,, \end{aligned}$$

where $0 \leq x \leq 1$. We want to select events of type s by requiring $x < x_{\mathrm{cut}}$, with $x_{\mathrm{cut}} = 0.1$.

(a) Find the efficiencies for selecting signal and background, i.e., the probabilities to accept events of types s and b, and evaluate numerically.

(b) Suppose the prior probabilities for events to be of types s and b are $\pi_{\mathrm{s}} = 0.01$ and $\pi_{\mathrm{b}} = 0.99$, respectively. Find the purity of signal events in the selected sample, i.e., the expected fraction of events with $x < x_{\mathrm{cut}}$ that are of type s and evaluate numerically.

(c) Suppose an event is observed with $x = 0.05$. Find the probability that the event is of type b and evaluate numerically.

(d) Again for an event with $x = 0.05$, 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 comment on why it is not equal to the probability found in (c).

(e) Describe with the aid of a sketch how to generate values of x following using the acceptance-rejection method.

Describe how to generate values of x following using the transformation method, and find the required transformation.

In both cases assume one has available a generator of random numbers uniformly distributed in $[0, 1]$.

(f) 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:

$$\begin{aligned} f(x, y|\mathrm{s}) &= 6(x-1)^2 y \,, \\ f(x, y|\mathrm{b}) &= 6x^2(1-y) \,. \end{aligned}$$

Write down the test statistic $t(x, y)$ which provides the highest signal purity for a given efficiency by selecting events inside a region defined by $t(x, y) = t_{\mathrm{cut}}$, where t_{cut} is a specified constant.

Exercise 2: The number of events n observed in an experiment with a given integrated luminosity can be modeled as a Poisson variable with a mean $s + b$, where s and b are the contributions from signal and background processes, respectively. Suppose $b = 3.9$ events are expected from background processes and $n = 16$ events are observed. Compute the p-value for the hypothesis $s = 0$, i.e., that 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.

Compute the corresponding significance, $Z = \Phi^{-1}(1-p)$. To evaluate the standard normal quantile Φ^{-1} you can use the ROOT routine `TMath::NormQuantile`.

Exercise 3: In the lectures we considered an experiment that measured a number of events n, modeled as following a Poisson distribution with a mean value of $\mu s + b$. Here b is the contribution from background and s represents the mean number of events from the nominal signal model. The goal of the experiment is to determine whether the signal is present, i.e., whether μ is non-zero. The likelihood function is

$$L(\mu) = \frac{(\mu s + b)^n}{n!} e^{-(\mu s + b)}\,. \tag{2}$$

The test statistic for discovery q_0 can be written

$$q_0 = \begin{cases} -2\ln\frac{L(0)}{L(\hat{\mu})} & \hat{\mu} \geq 0, \\ 0 & \hat{\mu} < 0\,, \end{cases} \tag{3}$$

where $\hat{\mu} = (n-b)/s$.

(a) By using the asymptotic relation $Z = \sqrt{q_0}$ from the lecture, show that for $n > b$ the discovery significance Z can be written

$$Z = \sqrt{2\left(n\ln\frac{n}{b} + b - n\right)}\,, \tag{4}$$

and that therefore the median significance assuming the nominal signal model can be approximated by

$$\text{med}[Z_0|1] = \sqrt{2\left((s+b)\ln(1+s/b) - s\right)}\,. \tag{5}$$

(b) By expanding the logarithm show that this reduces to

$$\text{med}[Z_0|1] = \frac{s}{\sqrt{b}}\left(1 + \mathcal{O}(s/b)\right)\,. \tag{6}$$

Although $Z_0 \approx s/\sqrt{b}$ has been widely used for cases where $s + b$ is large, one sees here that this final approximation is strictly valid only for $s \ll b$.