G. Cowan
4,5 May 2020

UGR Statistics Course Problem Sheet

The purpose of this exercise is to design a statistical test to discover a signal process such as dark matter by counting events in a detector. Suppose the detector can for each event measure a quantity x with $0 \leq x \leq 1$, for which probability density functions (pdfs) are for signal (s) and background (b),

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

$$f(x|b) = 3x^2 . \tag{2}$$

1(a) Suppose for each event we test the hypothesis that it is background. We reject this hypothesis if the observed value of x is less than a specified cut value $x_{\rm cut}$. Find the value of $x_{\rm cut}$ such that the probability to reject the background hypothesis (i.e., accept as signal) if it is background is $\alpha = 0.05$. (The value α is the *size* or significance level of the test.)

1(b) For the value of $x_{\rm cut}$ that you find, what is the probability to reject the background hypothesis (i.e., to accept the event as a candidate signal event) with $x < x_{\rm cut}$ given that it is signal. This is the *power* of the test with respect to the signal hypothesis or equivalently the signal efficiency.

1(c) Suppose that the expected number of background events is $b_{\rm tot} = 100$ and for a given signal model one expects $s_{\rm tot} = 10$ signal events. Find the expected numbers of events s and b of signal and background events that will satisfy $x < x_{\rm cut}$ using the value of $x_{\rm cut} = 0.1$.

1(d) Assuming the numbers from 1(c), the prior probabilities for an event to be signal or background are

$$\pi_s = \frac{s_{\rm tot}}{s_{\rm tot} + b_{\rm tot}} = 0.09 , \tag{3}$$

$$\pi_b = \frac{b_{\rm tot}}{s_{\rm tot} + b_{\rm tot}} = 0.91 . \tag{4}$$

Based on these values, what is the probability for an event to be signal given that one finds $x < x_{\rm cut}$. (Recall Bayes' theorem or consult `arXiv:1307.2487`.)

1(e) Now suppose we do the experiment and observe $n_{\rm obs}$ events in the search region $x < x_{\rm cut}$. We now want to test the hypothesis that $s = 0$ (the background-only hypothesis or "b"), against the alternative that signal is present with $s \neq 0$ (the "$s + b$" hypothesis).

The actual number of events n found in the experiment with $x < x_{\rm cut}$ can be modeled as following a Poisson distribution with a mean value of $s + b$. That is, the probability to find n events is

$$P(n|s,b) = \frac{(s+b)^n}{n!} e^{-(s+b)} . \tag{5}$$

Suppose for a certain x_{cut} one has $b = 0.5$ and we find there $n_{\text{obs}} = 3$ events. The p-value of the background-only hypothesis is the probability, assuming $s = 0$, to find $n \geq n_{\text{obs}}$.

$$p = P(n \geq n_{\text{obs}}|s = 0, b) = \sum_{n=n_{\text{obs}}}^{\infty} \frac{b^n}{n!} e^{-b} = 1 - \sum_{n=0}^{n_{\text{obs}}-1} \frac{b^n}{n!} e^{-b} \ . \tag{6}$$

Find the p-value and from this find the *significance* with which one can reject the $s = 0$ hypothesis, defined as

$$Z = \Phi^{-1}(1 - p) \ , \tag{7}$$

where Φ is the standard cumulative Gaussian distribution and Φ^{-1} is its inverse (the standard Gaussian quantile). For more information see Sec. 10 of `arXiv:1307.2487`. You will need the cumulative chi-square distribution and the quantile of the Gaussian distribution, which from ROOT are available as `1 - TMath::Prob` and `TMath::NormQuantile`.

1(f) The expected (median) significance assuming the $s+b$ hypothesis of the test of the $s = 0$ hypothesis is a measure of sensitivity and this is what one tries to maximize when designing an experiment. It can be approximated with a number of different formulas. For $s \ll b$ one can use $\text{med}[Z_b|s + b] = s/\sqrt{b}$. If $s \ll b$ does not hold, a better approximation is

$$\text{med}[Z_b|s + b] = \sqrt{2\left((s + b)\ln\left(1 + \frac{s}{b}\right) - s\right)} \ . \tag{8}$$

Using Eq. (8), find me median significance for $x_{\text{cut}} = 0.1$. If you have time, try to write a program to find the value of x_{cut} that maximizes the median significance.