

## Invisible Statistics Problem

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 ( $b$ ),

$$f(x|s) = 3(1-x)^2, \quad (1)$$

$$f(x|b) = 3x^2. \quad (2)$$

**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  $x_{\text{cut}}$ . Find the value of  $x_{\text{cut}}$  such that the probability to reject the background hypothesis (i.e., accept as signal) if it is background is  $\alpha = 0.05$ . (The value  $\alpha$  is the *size* or significance level of the test.)

**1(b)** For the value of  $x_{\text{cut}}$  that you find, what is the probability to accept an event with  $x < x_{\text{cut}}$  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  $b_{\text{tot}} = 100$  and for a given signal model one expects  $s_{\text{tot}} = 10$  signal events. Find the expected numbers of events  $s$  and  $b$  of signal and background events that will satisfy  $x < x_{\text{cut}}$  using the value of  $x_{\text{cut}} = 0.1$ .

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

$$\pi_s = \frac{s_{\text{tot}}}{s_{\text{tot}} + b_{\text{tot}}} = 0.09, \quad (3)$$

$$\pi_b = \frac{b_{\text{tot}}}{s_{\text{tot}} + b_{\text{tot}}} = 0.91. \quad (4)$$

Based on these values, what is the probability for an event to be signal given that one finds  $x < x_{\text{cut}}$ . (Recall Bayes' theorem or consult [arXiv:1307.2487](https://arxiv.org/abs/1307.2487).)

**1(e)** Now suppose we do the experiment and observe  $n_{\text{obs}}$  events in the search region  $x < x_{\text{cut}}$ . 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  $x < x_{\text{cut}}$  can be modeled as following a Poisson distribution with a mean value of  $s + b$ . That is, the probability to find  $n$  events is

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

Suppose for a certain  $x_{\text{cut}}$  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  $n \geq n_{\text{obs}}$ .

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

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

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

where  $\Phi$  is the standard cumulative Gaussian distribution and  $\Phi^{-1}$  is its inverse (the standard Gaussian quantile). For more information see Sec. 10 of [arXiv:1307.2487](https://arxiv.org/abs/1307.2487). You will need the quantiles of the chi-square and Gaussian distributions, 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

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

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  $x_{\text{cut}}$  that maximizes the median significance.

**1(g)** Now suppose that for each event we do not simply count the events having  $x$  in a certain region but we design a test that exploits each measured value in the entire range  $0 \leq x \leq 1$ . The data thus consist of the number  $n$  of events, which follows a Poisson distribution with mean of  $s + b$ , and the  $n$  values  $x_1, \dots, x_n$ .

We can define a test statistic to test the background-only hypothesis that is a monotonic function of the likelihood ratio  $L_{s+b}/L_b$ ,

$$q = -2 \sum_{i=1}^n \left[ 1 + \frac{s_{\text{tot}}}{b_{\text{tot}}} \frac{f(x_i | s)}{f(x_i | b)} \right] \quad (9)$$

The motivation for this statistic is described further in Sec. 5.1 of [arXiv:1307.2487](https://arxiv.org/abs/1307.2487).

From <http://www.pp.rhul.ac.uk/~cowan/stat/invisibles/mc/> download the program `invisibleMC.cc` and the makefile. Build and run the program. This will produce histograms of  $q$  under the  $s + b$  hypothesis, and also a histogram of  $q$  (called `h_q_sb`) and it will find the median  $q$ ,  $\text{med}[q | s + b]$ .

You should add code in analogy with this that generates data according to the background-only ( $s = 0$ ) hypothesis. Generate  $10^7$  experiments and count how many have  $q < \text{med}[q | s + b]$ . The fraction with  $q < \text{med}[q | s + b]$  is the median  $p$ -value of the background-only hypothesis. Find this and from it find the median significance  $Z$  (the sensitivity). Compare to the values you found from Eq. (8).