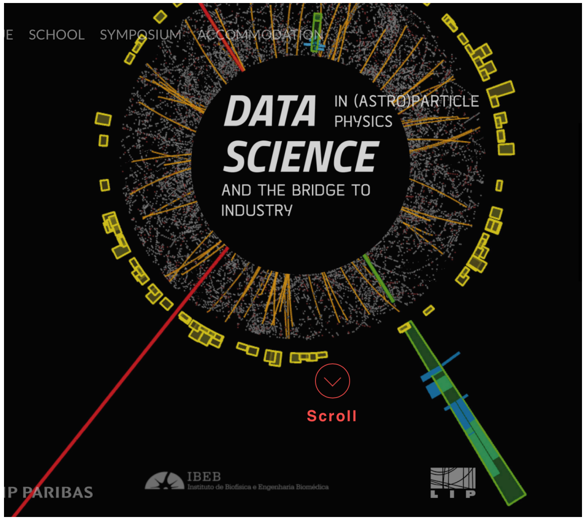
Statistical Methods for Particle Physics

Problem 2: statistical test for discovery

www.lip.pt/data-science-2018/
www.pp.rhul.ac.uk/~cowan/stat/lip18/prob_sheet_2.pdf



School on Data Science in (Astro)particle Physics
LIP Lisboa, 12-14 March, 2018

Glen Cowan, RHUL Physics

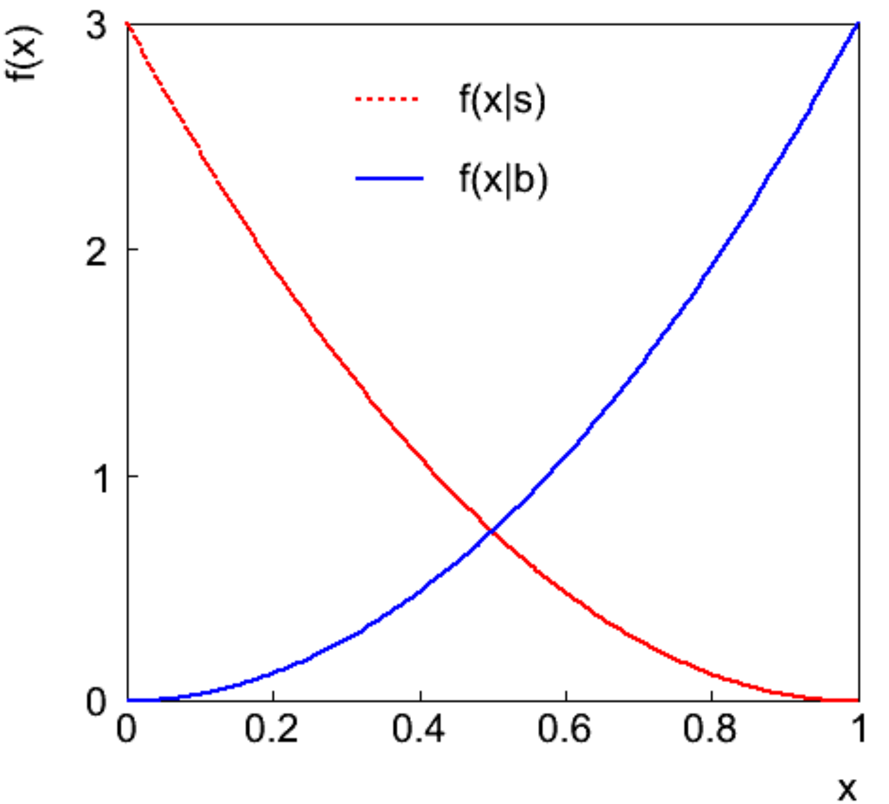
Problem 2 – discovering a small signal

Materials at **`www.pp.rhul.ac.uk/~cowan/stat/invisibles/`**

Problem concerns searching for a signal such as Dark Matter by counting events. Suppose signal/background events are characterized by a variable x ($0 \leq x \leq 1$):

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

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



As a first step, test the background hypothesis for each event: if $x < x_{cut}$, reject background hypothesis.

Testing the outcome of the full experiment

In the full experiment we will find n events in the signal region ($x < x_{\rm cut}$), and we can model this with a Poisson distribution:

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

Suppose total expected events in $0 \le x \le 1$ are $b_{\rm tot} = 100$, $s_{\rm tot} = 10$; expected in $x < x_{\rm cut}$ are s, b.

Suppose for a given $x_{\rm cut}$, $b = 0.5$ and we observe $n_{\rm obs} = 3$ events. Find the p-value of the hypothesis that $s = 0$:

$$p = P(n \ge n_{\rm obs}|s=0,b) = \sum_{n=n_{\rm obs}}^{\infty} \frac{b^n}{n!} e^{-b} = 1 - \sum_{n=0}^{n_{\rm obs}-1} \frac{b^n}{n!} e^{-b}$$

and the corresponding significance: $Z = \Phi^{-1}(1-p)$

Experimental sensitivity

To characterize the experimental sensitivity we can give the median, assuming s and b both present, of the significance of a test of $s = 0$. For $s \ll b$ this can be approximated by

$$\mathrm{med}[Z_b|s+b] = s/\sqrt{b}$$

A better approximation is:

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

Try this for $x_{\mathrm{cut}} = 0.1$ and if you have time, write a small program to maximize the median Z with respect to x_{cut}.

We will discuss these formulae in later lectures, including methods for treating uncertainty in b.

Using the x values

Instead of just counting events with $x < x_{cut}$, we can define a statistic that takes into account all the values of x. I.e. the data are: $n, x_1, .., x_n$. Later we will discuss ways of doing this with the likelihood ratio L_{s+b}/L_b, which leads to the statistic

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

Using **www.pp.rhul.ac.uk/~cowan/stat/invisibles/mc/invisibleMC.cc** find the distribution of this statistic under the "b" and "$s+b$" hypotheses.

From these find the median, assuming the $s+b$ hypothesis, of the significance of the b (i.e., $s = 0$) hypothesis. Compare with result from the experiment based only on counting n events.