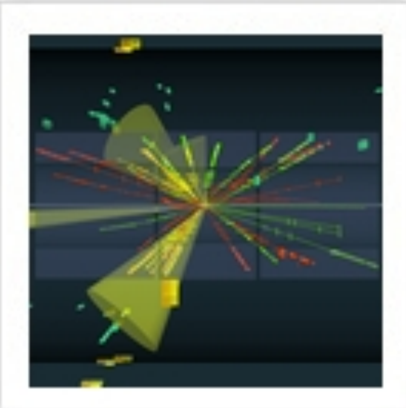


Taller de Altas Energías 2013

Statistics Problems: Solutions

Arely Cortes Gonzalez
arelycg@ifae.es

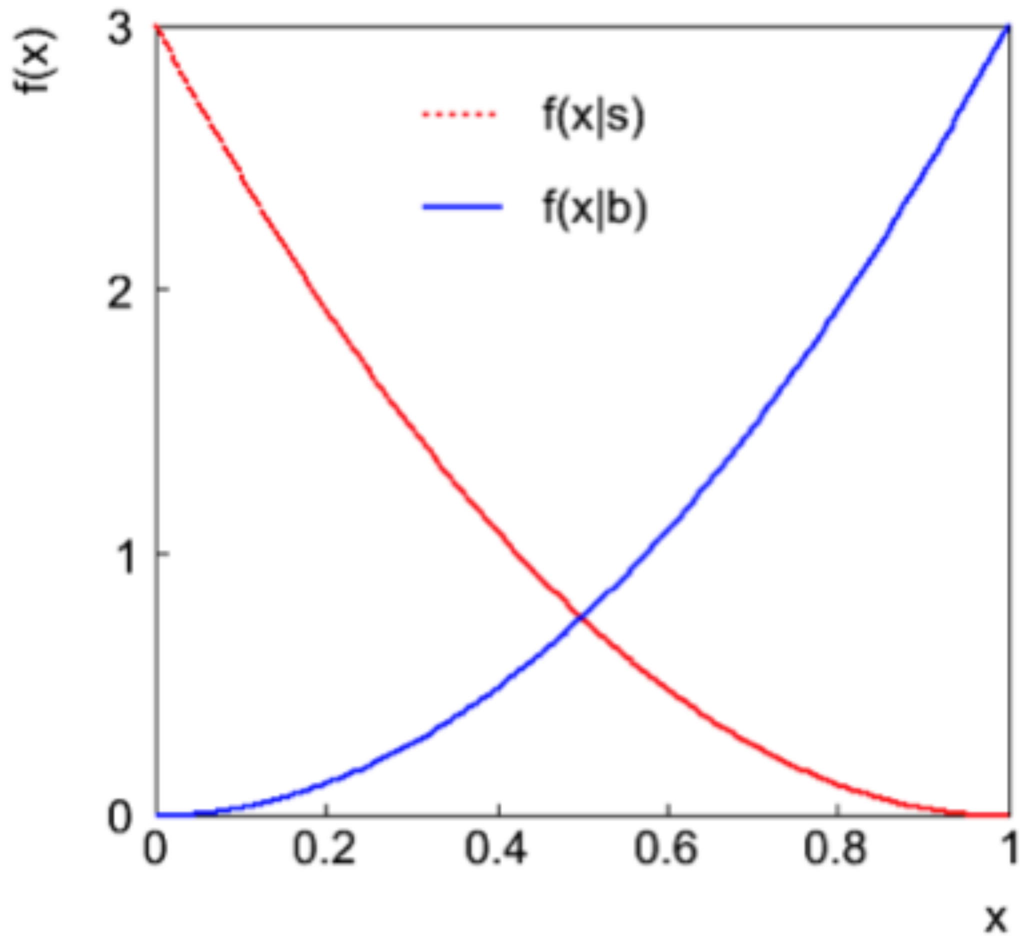
Problem 1: Discovering a small signal

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

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

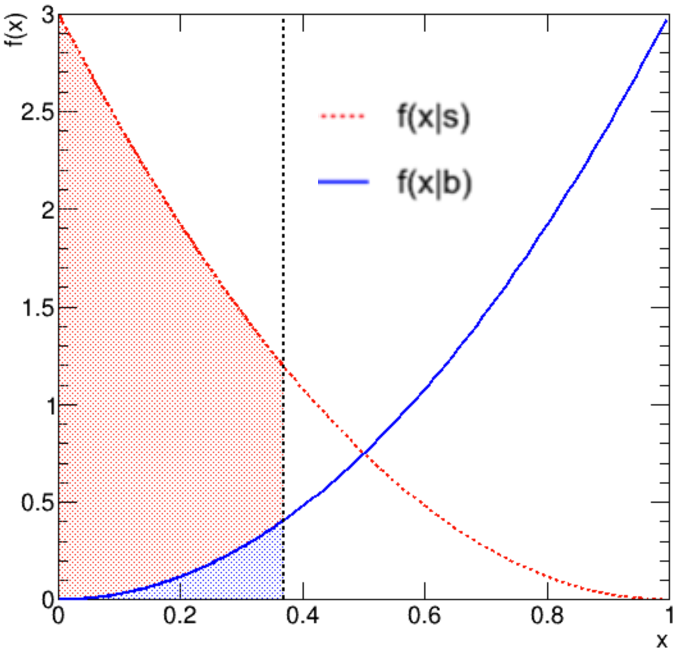
where

$$0 \leq x \leq 1$$



Size and Power of the test

1(a,b) Find the cumulative distribution...



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

$$\int_0^{x_{cut}} f(x|b)dx = \int_0^{x_{cut}} 3x^2 dx = x_{cut}^3$$
$$\int_0^{x_{cut}} f(x|s)dx = \int_0^{x_{cut}} 3(1-x)^2 dx = 1-(1-x_{cut})^3$$

Recall the definition of the *size* of the test:

Rejecting the hypothesis H_0 when it is true $P(x \in W|H_0) \leq \alpha$

$$x_{cut}^3 = \alpha \rightarrow x_{cut} = \alpha^{1/3} = 0.368$$

Power of the test: $P(x \in S - W|H_1) = \beta,\qquad \text{Power} = 1-\beta$

$$1-(1-x_{cut})^3 = 1-(1-0.368)^3 = 0.748$$

Bayes' theorem

1(c) $s_{\rm tot} = 10, b_{\rm tot} = 100, x_{cut} = 0.1$

we found the cumulative probabilities before:

$$\epsilon_b = x_{cut}^3 \qquad \epsilon_s = 1 - (1 - x_{cut})^3$$

$$b = b_{\rm tot} \cdot \epsilon_b = 100 x_{cut}^3 \big|_{x_{cut}=0.1} = 0.10 \qquad s = s_{\rm tot} \cdot \epsilon_s = 10(1 - (1 - x_{cut})^3) \big|_{x_{cut}=0.1} = 2.71$$

1(d) $\pi_s = 0.09, \pi_b = 0.91$

Recall Bayes' theorem

$$P(s|x < x_{cut}) = \frac{P(x < x_{cut}|s)\pi_s}{P(x < x_{cut}|s)\pi_s + P(x < x_{cut}|b)\pi_b} = \frac{\epsilon_s \pi_s}{\epsilon_s \pi_s + \epsilon_b \pi_b}$$

Signal purity $\quad P(s|x < x_{cut}) = 0.964$

p-value, significance

1(e) Experiment → observe n_{obs} events, in the region $x < x_{cut.}$

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

Expected bkg events $b = 0.5 \to b = b_{\text{tot}} \cdot \epsilon_b = b_{\text{tot}} \cdot x^3_{cut} \to x_{cut} = 0.171$

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

$$p = 1 - (1 + b + \frac{b^2}{2})e^{-b} = 0.0144$$

p-value, significance

1(e) Experiment → observe n_{obs} events, in the region $x < x_{cut.}$

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

Expected bkg events $b = 0.5 \rightarrow b = b_{\text{tot}} \cdot \epsilon_b = b_{\text{tot}} \cdot x_{cut}^3 \rightarrow x_{cut} = 0.171$

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

$$p = 1 - (1 + b + \frac{b^2}{2})e^{-b} = 0.0144$$

More events?! → Use identity from Sec. 10 of arXiv:1307.2487, relating the sum of Poisson probabilities to the cumulative χ^2 distribution:

$$\sum_{n=0}^{m} \frac{b^n}{n!} e^{-b} = 1 - F_{\chi^2}(2b; n_{dof}), \text{ with } n_{dof} = 2(m+1) \rightarrow 2n_{obs}$$

$p = F_{\chi^2}(2b; 2n_{obs}) = 1 - \text{TMath::Prob}(2b, 2n_{obs})$ ← Function in ROOT

$$p = 1 - \text{TMath::Prob}(1.0, 6) = 0.0144$$

Significance: $Z = \Phi^{-1}(1-p) = \text{TMath::NormQuantile}(1-p) = 2.19$

Expected (median) significance

1(f) $x_{cut} = 0.1$

For $s \ll b, \text{med}[Z_b|s+b] = s/\sqrt{b},$ otherwise $\text{med}[Z_b|s+b] = \sqrt{2\left((s+b)\ln\left(1+\frac{s}{b}\right) - s\right)}$

We computed s and b for x_{cut}=0.1

$b = b_{\text{tot}} \cdot \epsilon_b = 100 x_{cut}^3 \Big|_{x_{cut}=0.1} = 0.10$ $\qquad s = s_{\text{tot}} \cdot \epsilon_s = 10(1-(1-x_{cut})^3)\Big|_{x_{cut}=0.1} = 2.71$

$$\text{med}[Z_b|s+b] = \sqrt{2\left((s+b)\ln\left(1+\frac{s}{b}\right) - s\right)} = 3.65$$

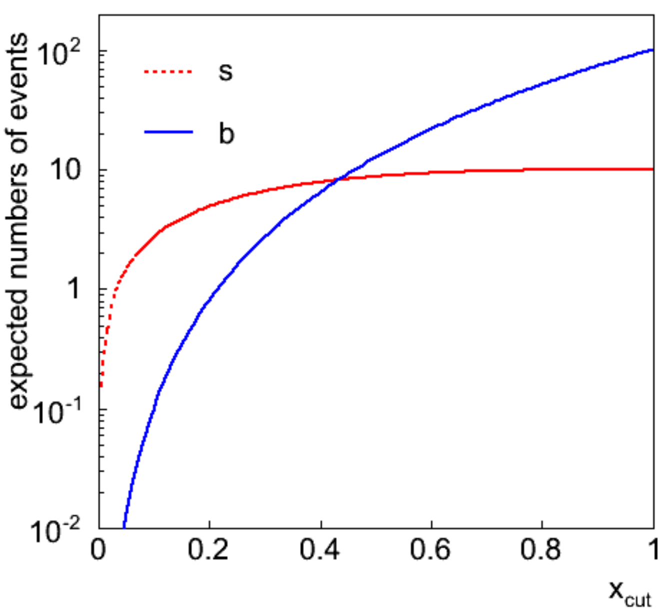
Compare to: $\text{med}[Z_b|s+b] = \frac{s}{\sqrt{b}} = 8.57$

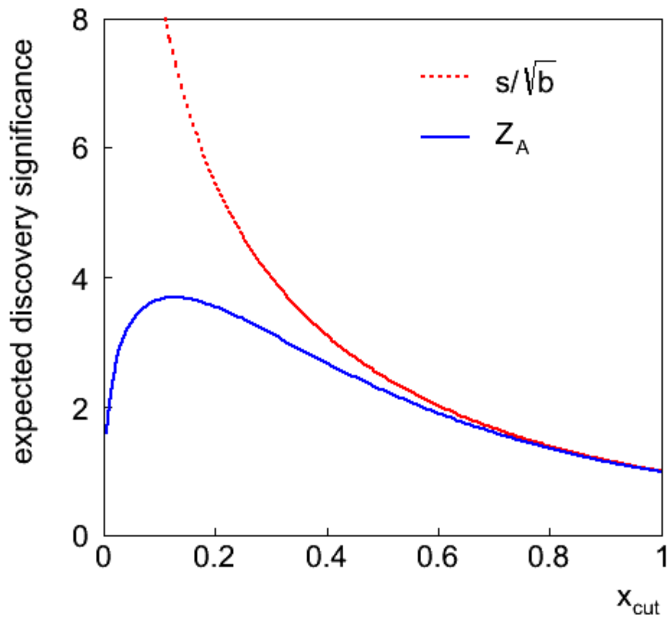
Expected (median) significance

1(f) $$\text{med}[Z_b|s+b] = \sqrt{2\left((s+b)\ln\left(1+\frac{s}{b}\right) - s\right)} = 3.65$$

Compare to: $$\text{med}[Z_b|s+b] = \frac{s}{\sqrt{b}} = 8.57$$

Write a program to compute the s and b at different x_{cut}, and then compute the median significance as a function of x_{cut}.



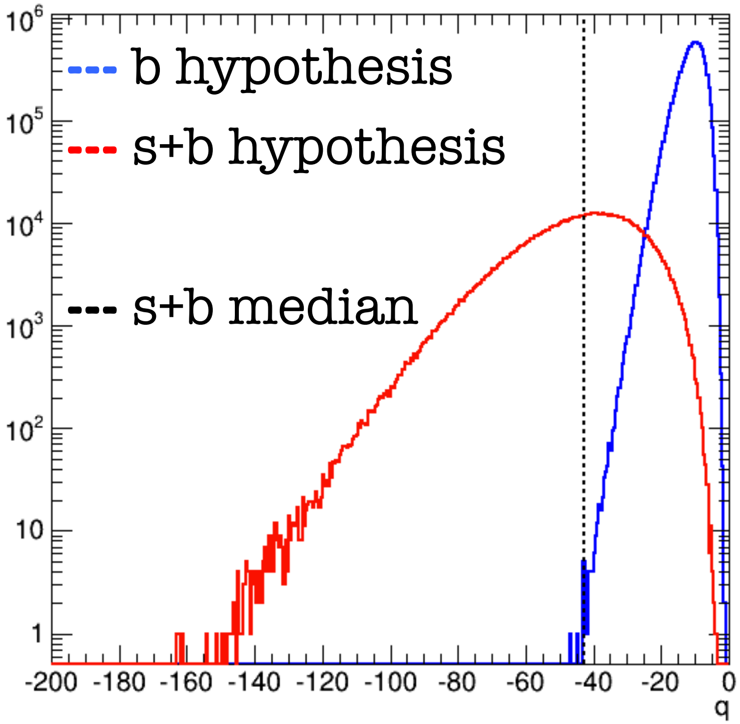


From the plot: $\text{Max med}[Z_b|s+b] = 3.68$ for $x_{cut} = 0.13$

Expected (median) significance

1(g) Design a test that exploits each measured value in the entire range of x. We define a test statistic to test the bkg-only hypothesis that is a monotonic function of the likelihood ratio

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



The code generates 10M experiments. Count number of events found in the b-only test, below the s+b median

$$q < \mathrm{med}[q|s+b] = 7$$

$$p_b = 7 \times 10^{-7}$$

$$\mathrm{med}[Z_b|s+b] = \Phi^{-1}(1-p_b) = 4.83$$

Expected (median) significance

1(g) How to generate s+b hypothesis

$$f(x) = \pi_s f(x|s) + \pi_b f(x|b)$$

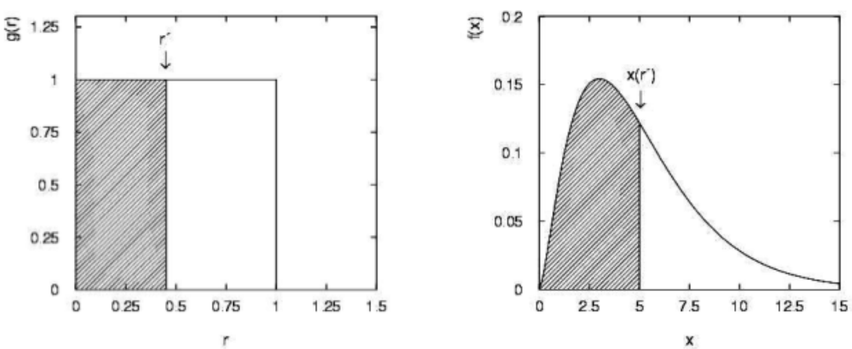
Find $r \sim U[0,1]$

If $(r < \pi_s)$ take $x \sim f(x|s)$

otherwise $x \sim f(x|b)$

← Mixture model

For $x \sim f(x)$ we find the cumulative distribution function and solve for r



Transformation method

$$F(x) = \int_0^x f(x')dx' = r$$

Require: $P(r \leq r') = P(x \leq x(r'))$

i.e. $\int_{-\infty}^{r'} g(r)\,dr = r' = \int_{-\infty}^{x(r')} f(x')\,dx' = F(x(r'))$

$$F(x|s) = 1 - (1-x)^3 \rightarrow x = 1 - r^{1/3}$$

$$F(x|b) = x^3 \rightarrow x = r^{1/3}$$

NB We are generating two independent random numbers here...