Statistical Data Analysis 2020/21
Lecture Week 4

London Postgraduate Lectures on Particle Physics

University of London MSc/MSci course PH4515

Glen Cowan
Physics Department
Royal Holloway, University of London
g.cowan@rhul.ac.uk
www.pp.rhul.ac.uk/~cowan

Course web page via RHUL moodle (PH4515) and also

www.pp.rhul.ac.uk/~cowan/stat_course.html

Statistical Data Analysis
Lecture 4-1

- Frequentist statistical tests
  - Hypotheses
  - Definition of a test
    - critical region
    - size
    - power
  - Type-I, Type-II errors

Hypotheses

A hypothesis H specifies the probability for the data, i.e., the outcome of the observation, here symbolically: x.

x could be uni-/multivariate, continuous or discrete.

E.g. write $x \sim P(x|H)$.

x could represent e.g. observation of a single object, a single event, or an entire "experiment".

Possible values of x form the sample space S (or "data space").

Simple (or "point") hypothesis: $P(x|H)$ completely specified.

Composite hypothesis: H contains unspecified parameter(s).

$P(x|H)$ is also called the likelihood of the hypothesis H, often written $L(H)$ if we want to emphasize just the dependence on H.

Definition of a test

Goal is to make some statement based on the observed data x about the validity of the possible hypotheses (here, "accept or reject").

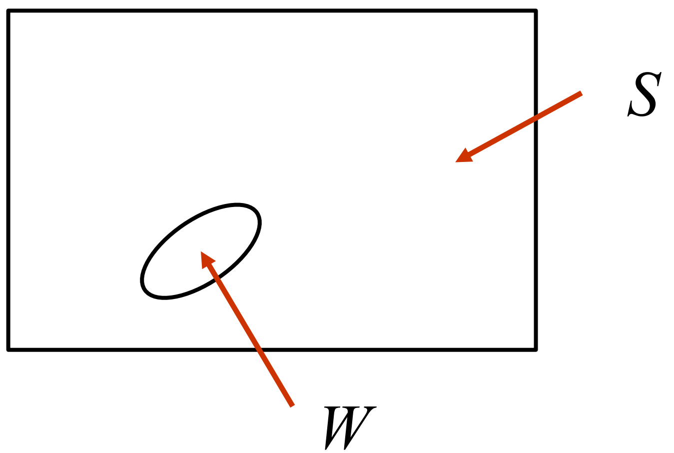
Consider a simple hypothesis H_0 (the "null") and an alternative H_1.

A test of H_0 is defined by specifying a critical region W of the sample (data) space S such that there is no more than some (small) probability α, assuming H_0 is correct, to observe the data there, i.e.,

$$P(x \in W \mid H_0) \leq \alpha$$

α is called the size of the test, prespecified equal to some small value, e.g., 0.05.

If x is observed in the critical region, reject H_0.

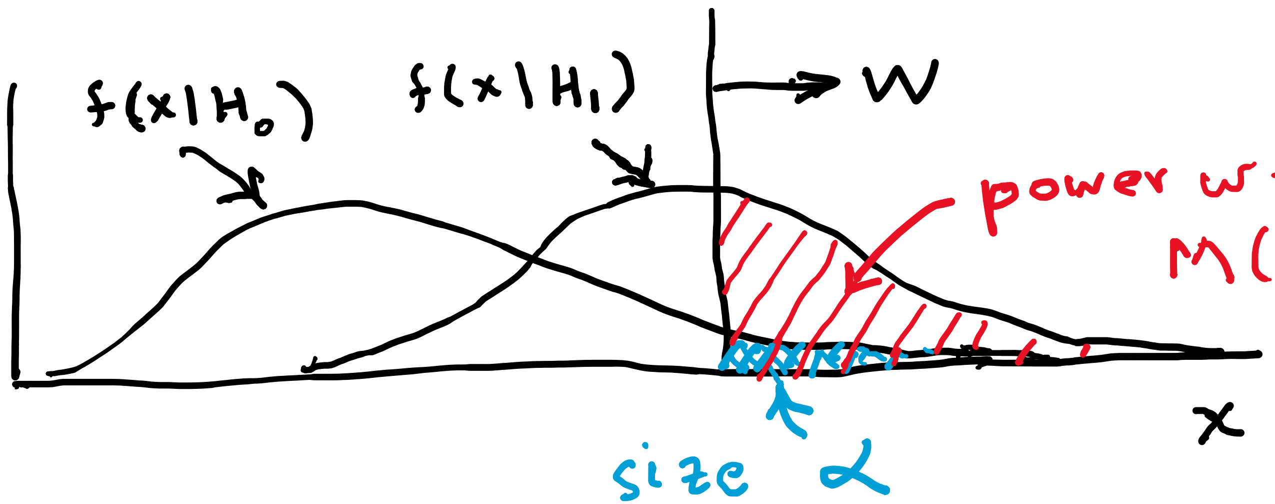


Definition of a test (2)

But in general there are an infinite number of possible critical regions that give the same size α.

Use the alternative hypothesis H_1 to motivate where to place the critical region.

Roughly speaking, place the critical region where there is a low probability (α) to be found if H_0 is true, but high if H_1 is true:

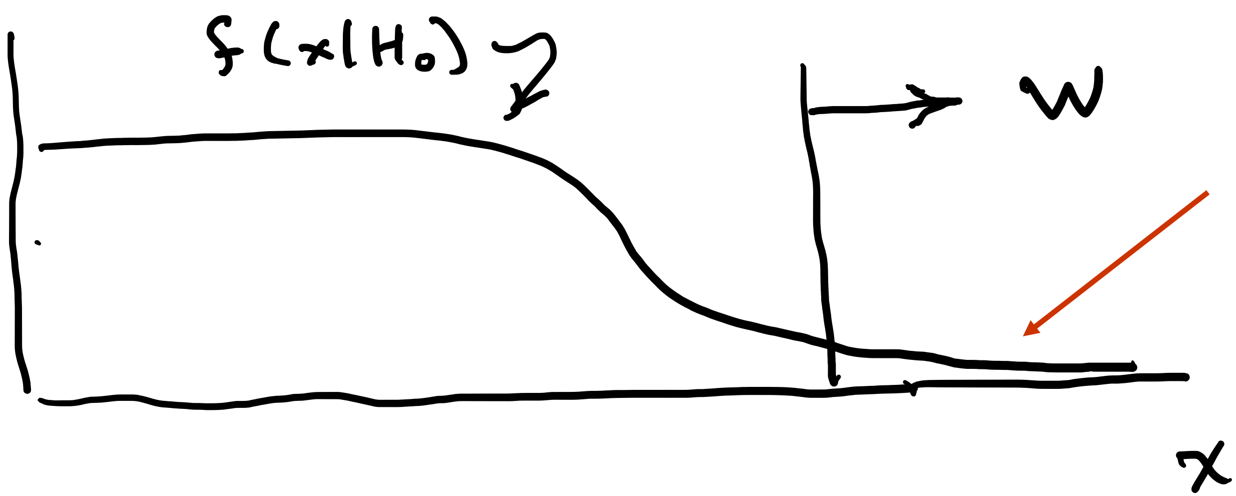


Obvious where to put W?

In the 1930s there were great debates as to the role of the alternative hypothesis.

Fisher held that one could test a hypothesis H_0 without reference to an alternative.

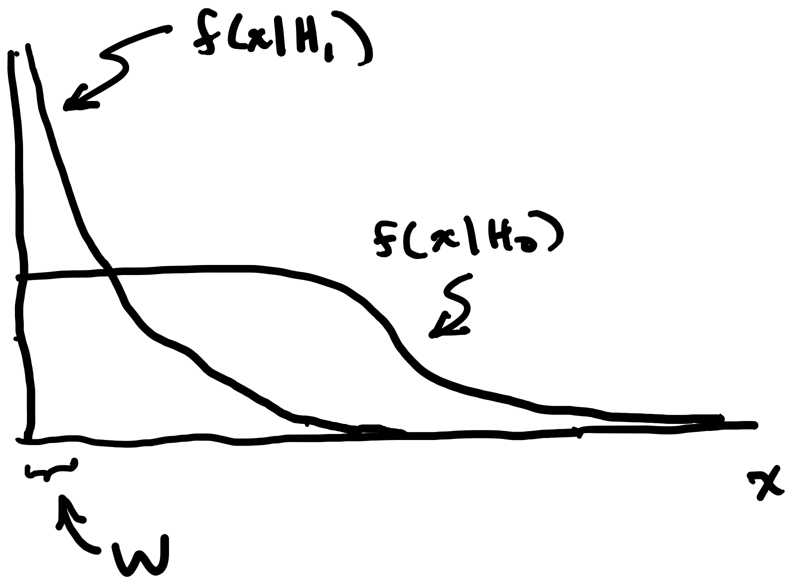
Suppose, e.g., H_0 predicts that x (suppose positive) usually comes out low. High values of x are less characteristic of H_0, so if a high value is observed, we should reject H_0, i.e., we put W at high x:



If we see x here, reject H_0.

Or not so obvious where to put W?

But what if the only relevant alternative to H_0 is H_1 as below:



Here high x is more characteristic of H_0 and not like what we expect from H_1. So better to put W at low x.

Neyman and Pearson argued that "less characteristic of H_0" is well defined only when taken to mean "more characteristic of some relevant alternative H_1".

Type-I, Type-II errors

Rejecting the hypothesis H_0 when it is true is a Type-I error.

The maximum probability for this is the size of the test:

$$P(x \in W \mid H_0) \leq \alpha$$

But we might also accept H_0 when it is false, and an alternative H_1 is true.

This is called a Type-II error, and occurs with probability

$$P(x \in S - W \mid H_1) = \beta$$

One minus this is called the power of the test with respect to the alternative H_1:

$$\text{Power} = 1 - \beta$$

Rejecting a hypothesis

Note that rejecting H_0 is not necessarily equivalent to the statement that we believe it is false and H_1 true. In frequentist statistics only associate probability with outcomes of repeatable observations (the data).

In Bayesian statistics, probability of the hypothesis (degree of belief) would be found using Bayes' theorem:

$$P(H|x) = \frac{P(x|H)\pi(H)}{\int P(x|H)\pi(H)\,dH}$$

which depends on the prior probability $\pi(H)$.

What makes a frequentist test useful is that we can compute the probability to accept/reject a hypothesis assuming that it is true, or assuming some alternative is true.

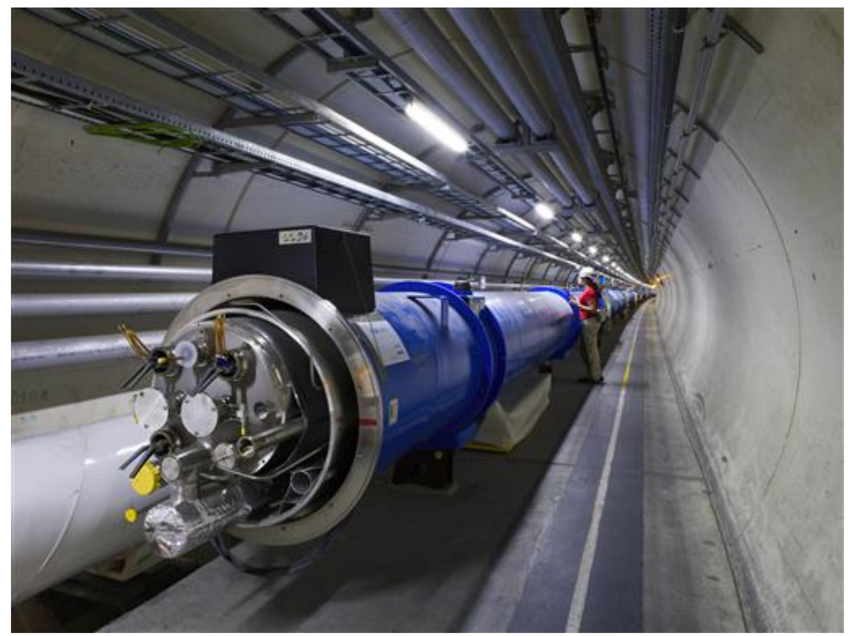
Statistical Data Analysis
Lecture 4-2

- Particle Physics example for statistical tests
- Statistical tests to select objects/events

Example setting for statistical tests: the Large Hadron Collider

Counter-rotating proton beams in 27 km circumference ring

pp centre-of-mass energy 14 TeV

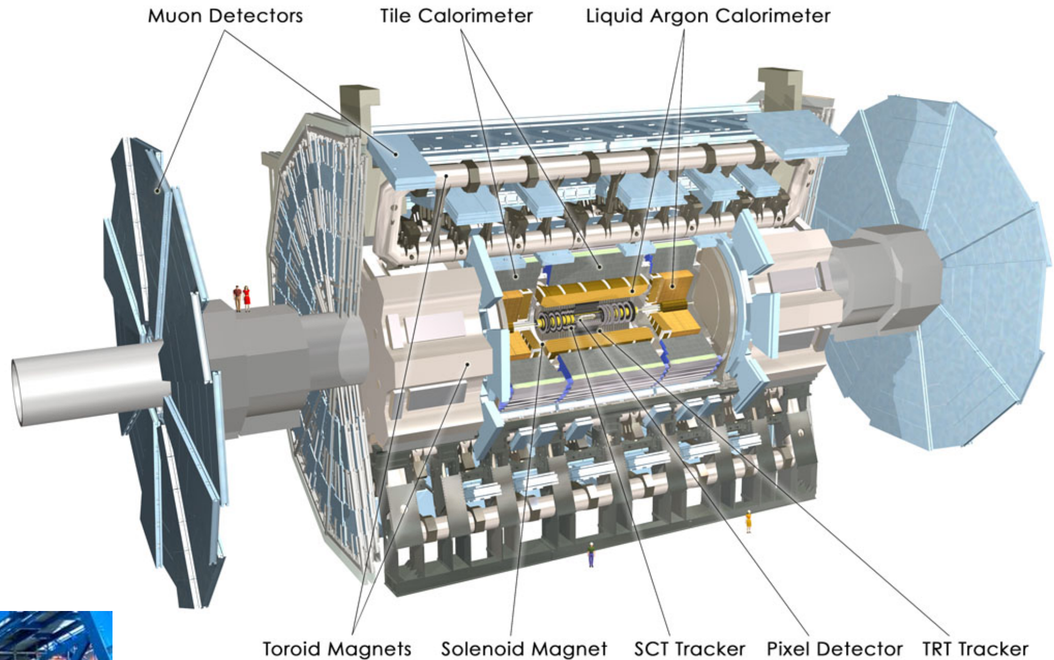


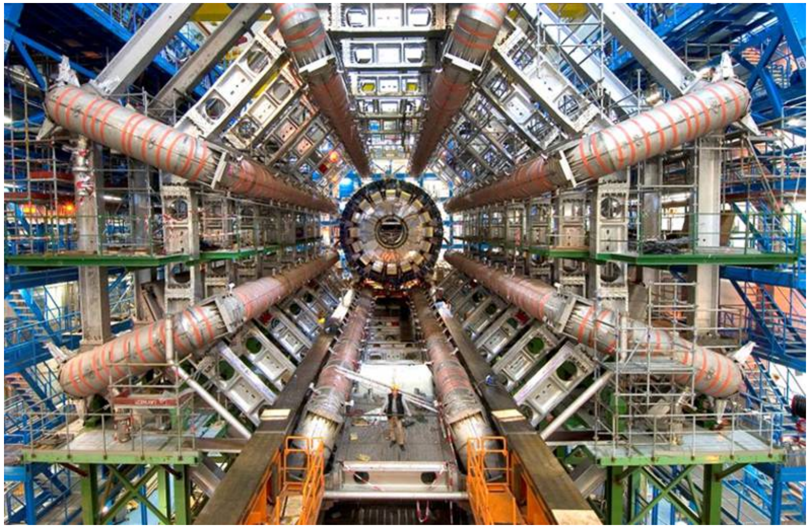
Detectors at 4 pp collision points:

- ATLAS ← general purpose
- CMS ← general purpose
- LHCb (b physics)
- ALICE (heavy ion physics)

The ATLAS detector

3000 physicists
38 countries
183 universities/labs



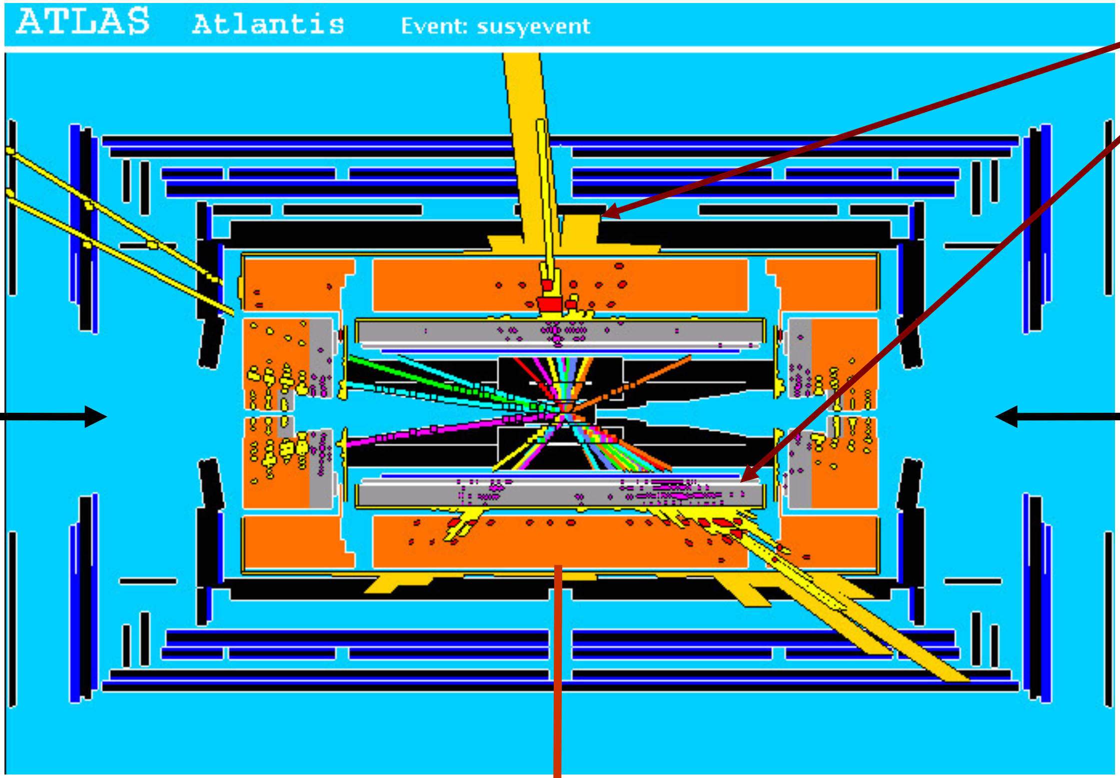


25 m diameter
46 m length
7000 tonnes
$\sim 10^8$ electronic channels

A simulated SUSY event

high p_T muons

high p_T jets of hadrons

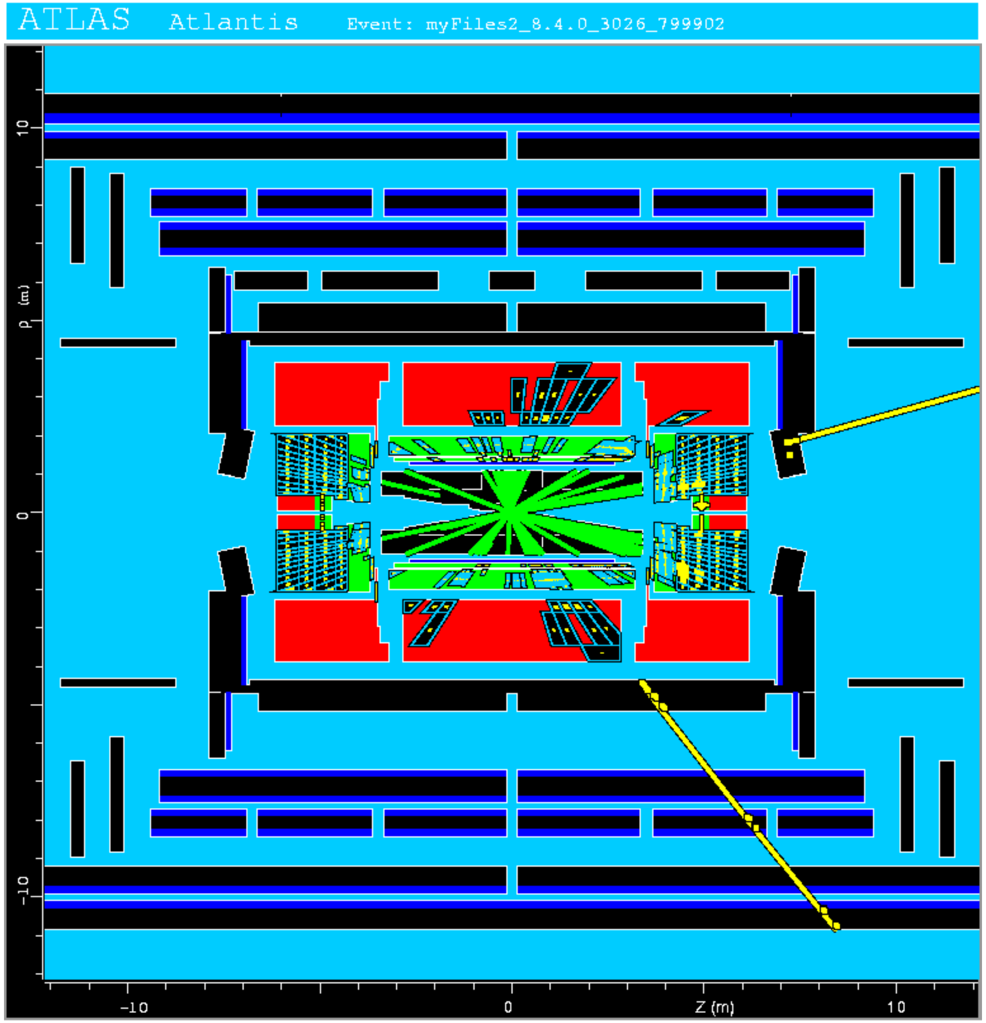


p

p

missing transverse energy

Background events



This event from Standard Model ttbar production also has high p_{T} jets and muons, and some missing transverse energy.

→ can easily mimic a signal event.

Classification viewed as a statistical test

Suppose events come in two possible types:

s (signal) and b (background)

For each event, test hypothesis that it is background, i.e., H_0 = b.

Carry out test on many events, each is either of type s or b, i.e., here the hypothesis is the "true class label", which varies randomly from event to event, so we can assign to it a frequentist probability.

Select events for which where H_0 is rejected as "candidate events of type s". Equivalent Particle Physics terminology:

background efficiency $$\varepsilon_{\rm b} = \int_W f(\mathbf{x}|H_0)\, d\mathbf{x} = \alpha$$

signal efficiency $$\varepsilon_{\rm s} = \int_W f(\mathbf{x}|H_1)\, d\mathbf{x} = 1 - \beta = \text{power}$$

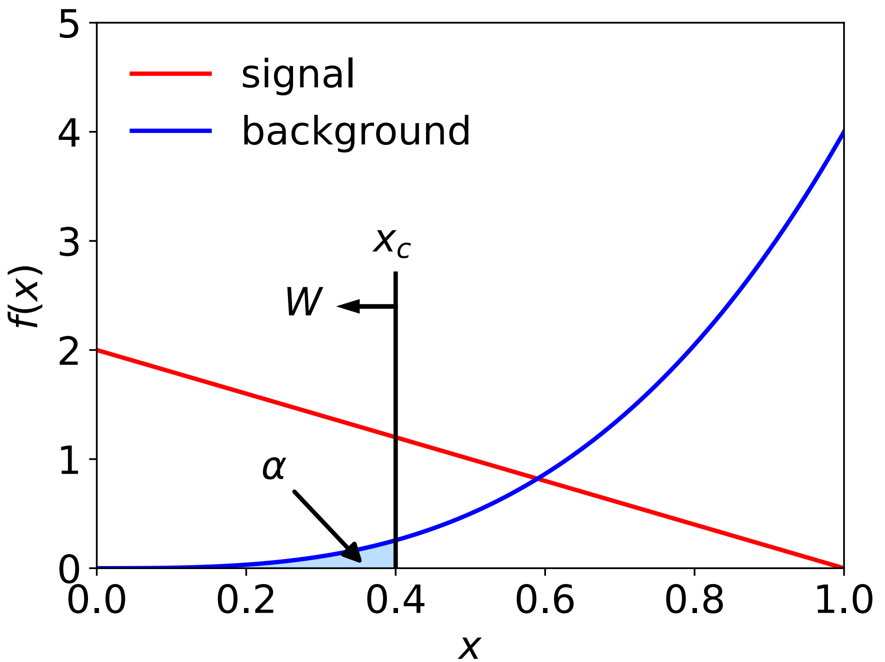
Example of a test for classification

Suppose we can measure for each event a quantity x, where

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

$$f(x|b) = 4x^3$$

with $0 \le x \le 1$.



For each event in a mixture of signal (s) and background (b) test

H_0 : event is of type b

using a critical region W of the form: $W = \{x : x \le x_{\rm c}\}$, where $x_{\rm c}$ is a constant that we choose to give a test with the desired size α.

Classification example (2)

Suppose we want $\alpha = 10^{-4}$. Require:

$$\alpha = P(x \leq x_{\rm c}|{\rm b}) = \int_0^{x_{\rm c}} f(x|b)\,dx = \left.\frac{4x^4}{4}\right|_0^{x_{\rm c}} = x_{\rm c}^4$$

and therefore $x_{\rm c} = \alpha^{1/4} = 0.1$

For this test (i.e. this critical region W), the power with respect to the signal hypothesis (s) is

$$M = P(x \leq x_{\rm c}|{\rm s}) = \int_0^{x_{\rm c}} f(x|{\rm s})\,dx = 2x_{\rm c} - x_{\rm c}^2 = 0.19$$

Note: the optimal size and power is a separate question that will depend on goals of the subsequent analysis.

Classification example (3)

Suppose that the prior probabilities for an event to be of type s or b are:

$$\pi_{\mathrm{s}} = 0.001$$
$$\pi_{\mathrm{b}} = 0.999$$

The "purity" of the selected signal sample (events where b hypothesis rejected) is found using Bayes' theorem:

$$P(\mathrm{s}|x \leq x_{\mathrm{c}}) = \frac{P(x \leq x_{\mathrm{c}}|\mathrm{s})\pi_{\mathrm{s}}}{P(x \leq x_{\mathrm{c}}|\mathrm{s})\pi_{\mathrm{s}} + P(x \leq x_{\mathrm{c}}|\mathrm{b})\pi_{\mathrm{b}}}$$

$$= 0.655$$

Classification example (4)

Suppose an individual event is observed at x = 0.1. What is the probability that this event is background?

$$P(\mathrm{b}|x) = \frac{f(x|\mathrm{b})\pi_{\mathrm{b}}}{f(x|\mathrm{b})\pi_{\mathrm{b}} + f(x|\mathrm{s})\pi_{\mathrm{s}}}$$

$$= \frac{4x^3\pi_{\mathrm{b}}}{4x^3\pi_{\mathrm{b}} + 2(1-x)\pi_{\mathrm{s}}}$$

$$= 0.689$$

(Here nothing to do with the test using $x \leq x_{\mathrm{c}}$, just an illustration of Bayes' theorem.)