

# Statistical & Mathematical Foundations

## Lecture 1: Probability, Hypothesis tests



STEAM Academy  
CERN  
6,7 July 2026

<https://indico.cern.ch/event/1699443/>



Glen Cowan  
Physics Department  
Royal Holloway, University of London  
[g.cowan@rhul.ac.uk](mailto:g.cowan@rhul.ac.uk)  
[www.pp.rhul.ac.uk/~cowan](http://www.pp.rhul.ac.uk/~cowan)

# Outline

→ Lecture 1:

Introduction

Probability

Hypothesis tests

Lecture 2:

Parameter estimation

Confidence limits

Systematic uncertainties

Almost everything is a subset of the University of London course:

[http://www.pp.rhul.ac.uk/~cowan/stat\\_course.html](http://www.pp.rhul.ac.uk/~cowan/stat_course.html)

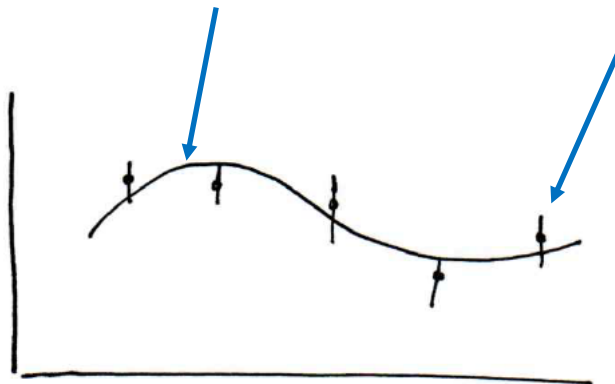
# Theory ↔ Statistics ↔ Experiment

Theory (model, hypothesis):

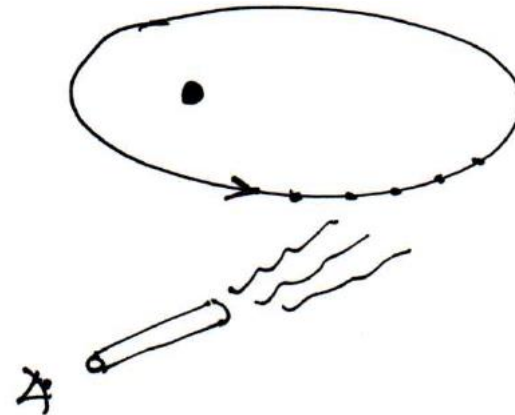
$$F = -G \frac{m_1 m_2}{r^2}, \dots$$

+ response of measurement apparatus

= model prediction



Experiment (observation):



data

Uncertainty enters on many levels

→ quantify with **probability**

# A quick review of probability

Frequentist ( $A$  = outcome of repeatable observation)

$$P(A) = \lim_{n \rightarrow \infty} \frac{\text{outcome is in } A}{n}$$

Subjective ( $A$  = hypothesis)

$P(A)$  = degree of belief that  $A$  is true

Conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

E.g. rolling a die,  
outcome  $n = 1, 2, \dots, 6$ :

$$P(n \leq 3 | n \text{ even}) = \frac{P((n \leq 3) \cap n \text{ even})}{P(n \text{ even})} = \frac{1/6}{3/6} = \frac{1}{3}$$

$A$  and  $B$  are independent iff:

$$P(A \cap B) = P(A)P(B)$$

I.e. if  $A, B$  independent, then

$$P(A|B) = \frac{P(A)P(B)}{P(B)} = P(A)$$

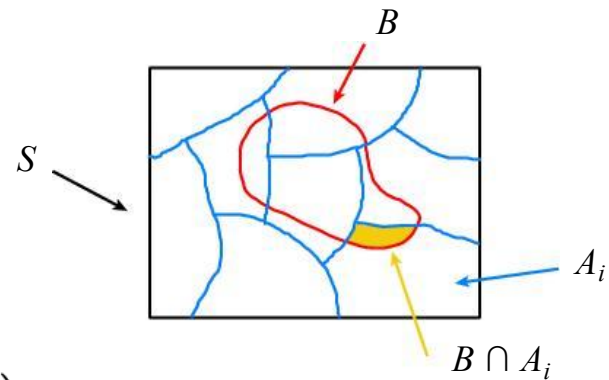
# Bayes' theorem

Use definition of conditional probability and  $P(A \cap B) = P(B \cap A)$

$$\rightarrow P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (\text{Bayes' theorem})$$

If set of all outcomes  $S = \cup_i A_i$   
with  $A_i$  disjoint, then law of total  
probability for  $P(B)$  says

$$P(B) = \sum_i P(B \cap A_i) = \sum_i P(B|A_i)P(A_i)$$



so that Bayes' theorem becomes  $P(A|B) = \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)}$

Bayes' theorem holds regardless of how probability is interpreted (frequency, degree of belief...).

# Frequentist Statistics – general philosophy

In frequentist statistics, probabilities are associated only with the data, i.e., outcomes of repeatable observations (shorthand:  $x$ ).

Probability = limiting frequency

Probabilities such as

$P$  (string theory is true),

$P(0.117 < \alpha_s < 0.119)$ ,

$P$  (Whitmer wins in 2028),

etc. are either 0 or 1, but we don't know which.

The tools of frequentist statistics tell us what to expect, under the assumption of certain probabilities, about hypothetical repeated observations.

Preferred theories (models, hypotheses, ...) are those that predict a high probability for data “like” the data observed.

# Bayesian Statistics – general philosophy

In Bayesian statistics, use subjective probability for hypotheses:

probability of the data assuming hypothesis  $H$  (the likelihood)

prior probability, i.e., before seeing the data

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

posterior probability, i.e., after seeing the data

normalization involves sum over all possible hypotheses

Bayes' theorem has an “if-then” character: **if** your prior probabilities were  $\pi(H)$ , **then** it says how these probabilities should change in the light of the data.

No general prescription for priors (subjective!)

# Hypothesis, likelihood

Suppose the entire result of an experiment (set of measurements) is a collection of numbers  $x$ .

A (simple) hypothesis is a rule that assigns a probability to each possible data outcome:

$$P(\mathbf{x}|H) = \text{the likelihood of } H$$

Often we deal with a family of hypotheses labeled by one or more undetermined parameters (a composite hypothesis):

$$P(\mathbf{x}|\boldsymbol{\theta}) = L(\boldsymbol{\theta}) = \text{the “likelihood function”}$$

Note:

- 1) For the likelihood we treat the data  $x$  as fixed.
- 2) The likelihood function  $L(\boldsymbol{\theta})$  is not a pdf for  $\boldsymbol{\theta}$ .

# Frequentist hypothesis tests

Suppose a measurement produces data  $\mathbf{x}$ ; consider a hypothesis  $H_0$  we want to test and alternative  $H_1$

$H_0, H_1$  specify probability for  $\mathbf{x}$ :  $P(\mathbf{x}|H_0), P(\mathbf{x}|H_1)$

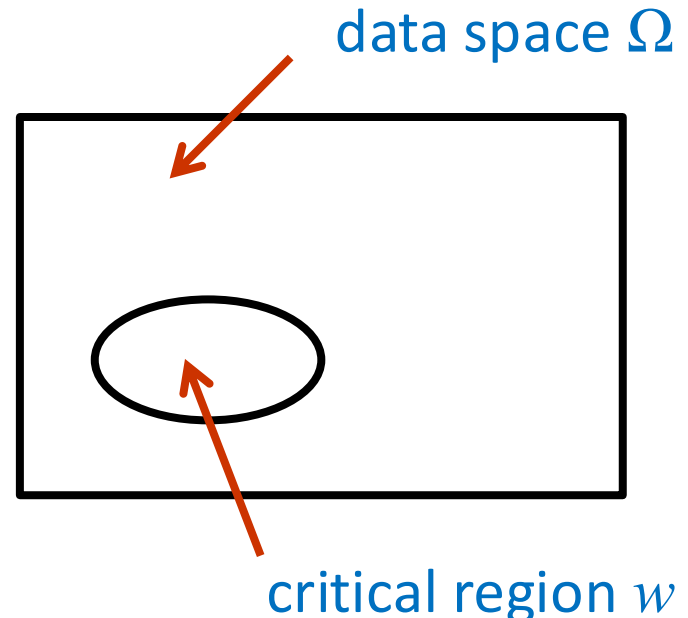
A test of  $H_0$  is defined by specifying a critical region  $w$  of the data space such that there is no more than some (small) probability  $\alpha$ , assuming  $H_0$  is correct, to observe the data there, i.e.,

$$P(\mathbf{x} \in w \mid H_0) \leq \alpha$$

Need inequality if data are discrete.

$\alpha$  is called the size or significance level of the test.

If  $\mathbf{x}$  is observed in the critical region, reject  $H_0$ .

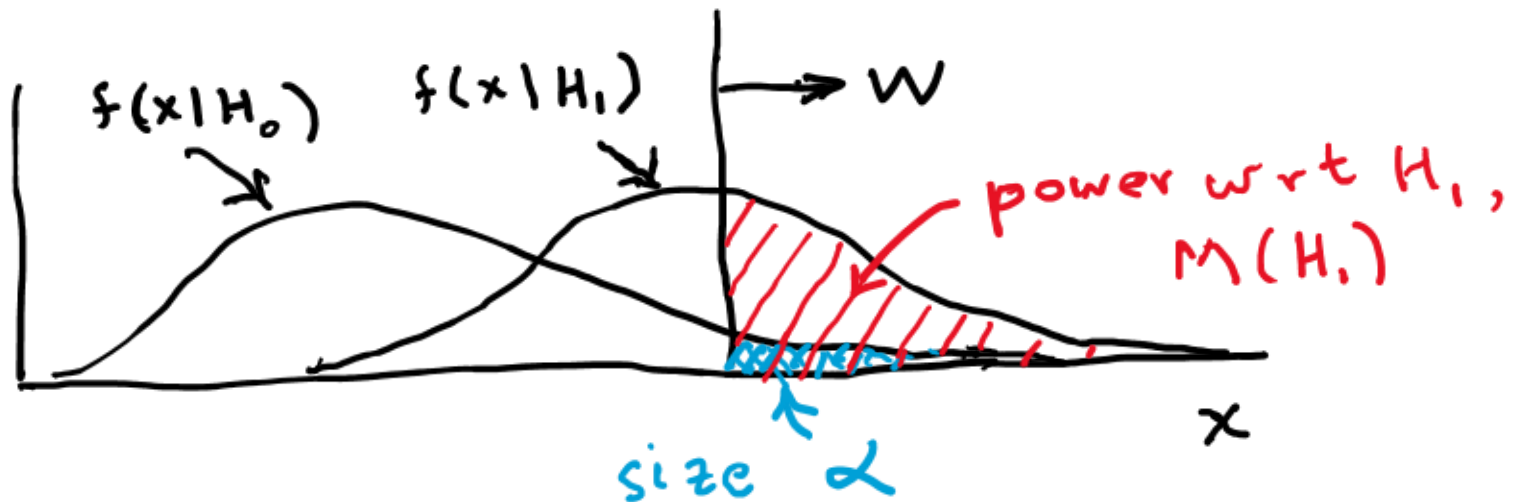


## Definition of a test (2)

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

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 ( $\alpha$ ) to be found if  $H_0$  is true, but high if  $H_1$  is true:



# 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  $\epsilon_b = \int_W f(\mathbf{x}|H_0) d\mathbf{x} = \alpha$

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

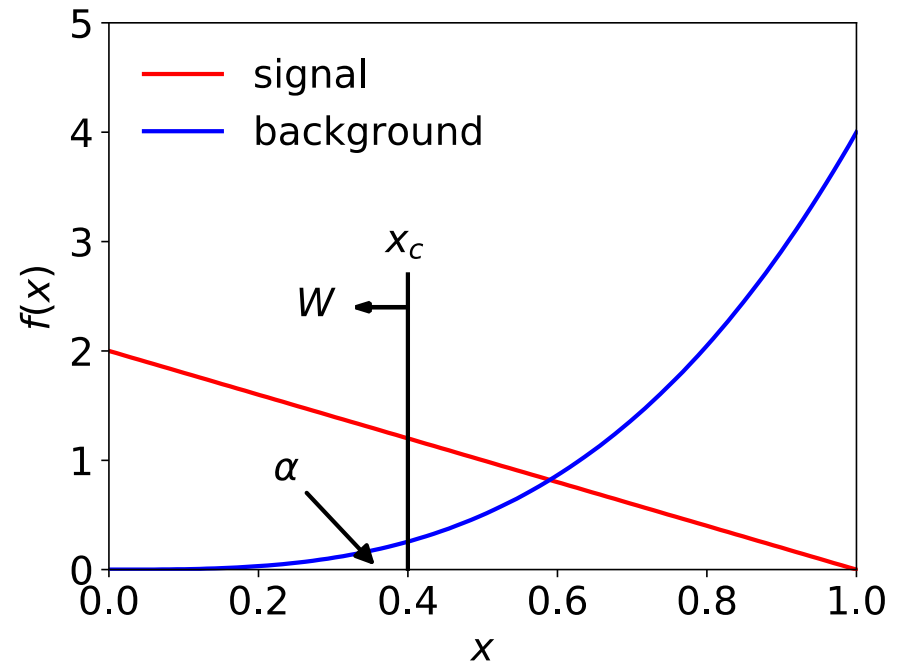
# Example of a test for classification

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

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

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

with  $0 \leq x \leq 1$ .



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:  $W = \{x : x \leq x_c\}$ , where  $x_c$  is a constant that we choose to give a test with the desired size  $\alpha$ .

## Classification example (2)

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

$$\alpha = P(x \leq x_c | b) = \int_0^{x_c} f(x|b) dx = \frac{4x^4}{4} \Big|_0^{x_c} = x_c^4$$

and therefore  $x_c = \alpha^{1/4} = 0.1$

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

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

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:

$$\pi_s = 0.001$$

$$\pi_b = 0.999$$

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

$$\begin{aligned} P(s|x \leq x_c) &= \frac{P(x \leq x_c|s)\pi_s}{P(x \leq x_c|s)\pi_s + P(x \leq x_c|b)\pi_b} \\ &= 0.655 \end{aligned}$$

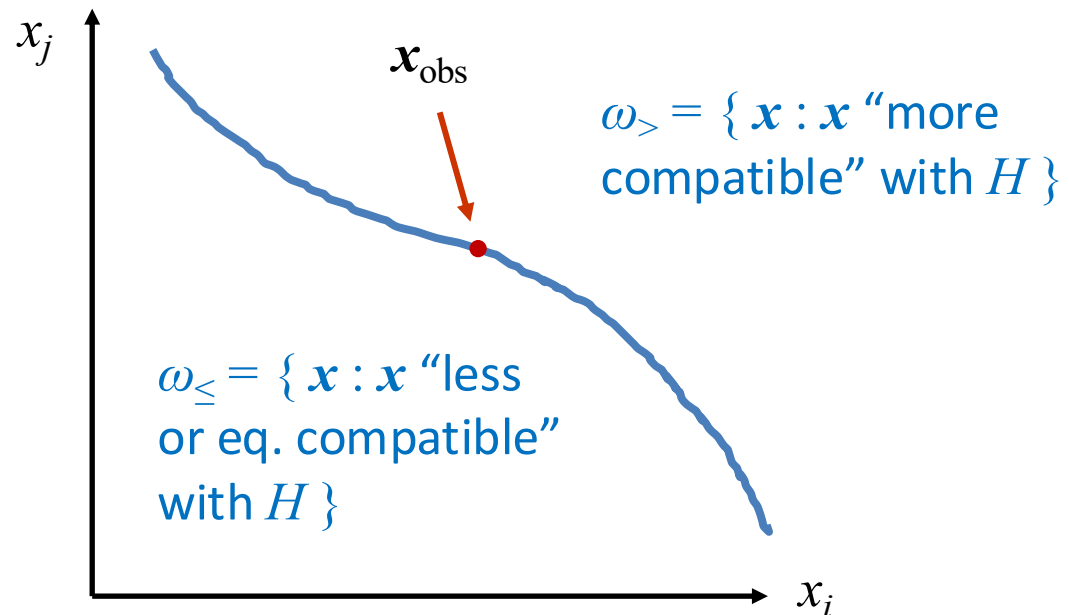
# Testing significance / goodness-of-fit

Suppose hypothesis  $H$  predicts pdf  $f(\mathbf{x}|H)$  for a set of observations  $\mathbf{x} = (x_1, \dots, x_n)$ .

We observe a single point in this space:  $\mathbf{x}_{\text{obs}}$ .

How can we quantify the level of compatibility between the data and the predictions of  $H$ ?

Decide what part of the data space represents equal or less compatibility with  $H$  than does the point  $\mathbf{x}_{\text{obs}}$ . (Not unique!)



# $p$ -values

Express level of compatibility between data and hypothesis (sometimes ‘goodness-of-fit’) by giving the  $p$ -value for  $H$ :

$$p = P(\mathbf{x} \in \omega_{\leq}(\mathbf{x}_{\text{obs}}) | H)$$

- = probability, under assumption of  $H$ , to observe data with equal or lesser compatibility with  $H$  relative to the data we got.
- = probability, under assumption of  $H$ , to observe data as discrepant with  $H$  as the data we got or more so.

Basic idea: if there is only a very small probability to find data with even worse (or equal) compatibility, then  $H$  is “disfavoured by the data”.

If the  $p$ -value is below a user-defined threshold  $\alpha$  (e.g. 0.05) then  $H$  is rejected (equivalent to hypothesis test of size  $\alpha$  as seen earlier).



## $p$ -value of $H$ is not $P(H)$

The  $p$ -value of  $H$  is not the probability that  $H$  is true!

In frequentist statistics we don't talk about  $P(H)$  (unless  $H$  represents a repeatable observation).

If we do define  $P(H)$ , e.g., in Bayesian statistics as a degree of belief, then we need to use Bayes' theorem to obtain

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

where  $\pi(H)$  is the prior probability for  $H$ .

For now stick with the frequentist approach;  
result is  $p$ -value, regrettably easy to misinterpret as  $P(H)$ .

# The Poisson counting experiment

Suppose we do a counting experiment and observe  $n$  events.

Events could be from *signal* process or from *background* – we only count the total number.

Poisson model:

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

$s$  = mean (i.e., expected) # of signal events

$b$  = mean # of background events

Goal is to make inference about  $s$ , e.g.,

test  $s = 0$  (rejecting  $H_0 \approx$  “discovery of signal process”)

test all non-zero  $s$  (values not rejected = confidence interval)

In both cases need to ask what is relevant alternative hypothesis.

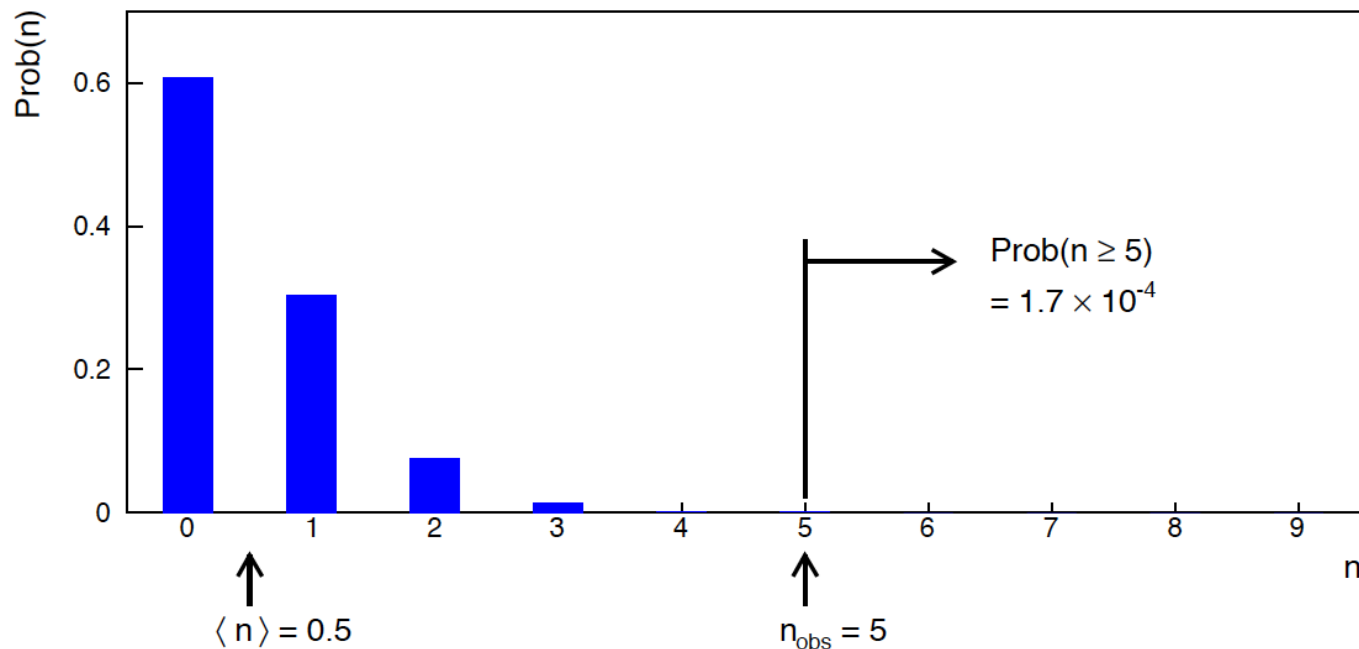
# Poisson counting experiment: discovery $p$ -value

Suppose  $b = 0.5$  (known), and we observe  $n_{\text{obs}} = 5$ .

Should we claim evidence for a new discovery?

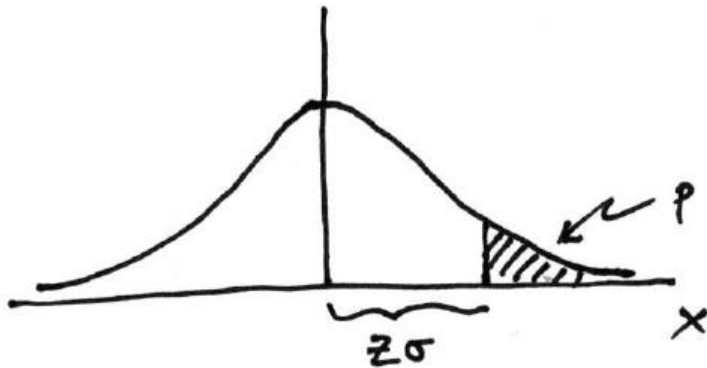
Give  $p$ -value for hypothesis  $s = 0$ , suppose relevant alt. is  $s > 0$ .

$$\begin{aligned} p\text{-value} &= P(n \geq 5; b = 0.5, s = 0) \\ &= 1.7 \times 10^{-4} \neq P(s = 0)! \end{aligned}$$



# Significance from $p$ -value

Often define significance  $Z$  as the number of standard deviations that a Gaussian variable would fluctuate in one direction to give the same  $p$ -value.



$$p = \int_Z^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1 - \Phi(Z)$$

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

in ROOT:

```
p = 1 - TMath::Freq(Z)
Z = TMath::NormQuantile(1-p)
```

in python (scipy.stats):

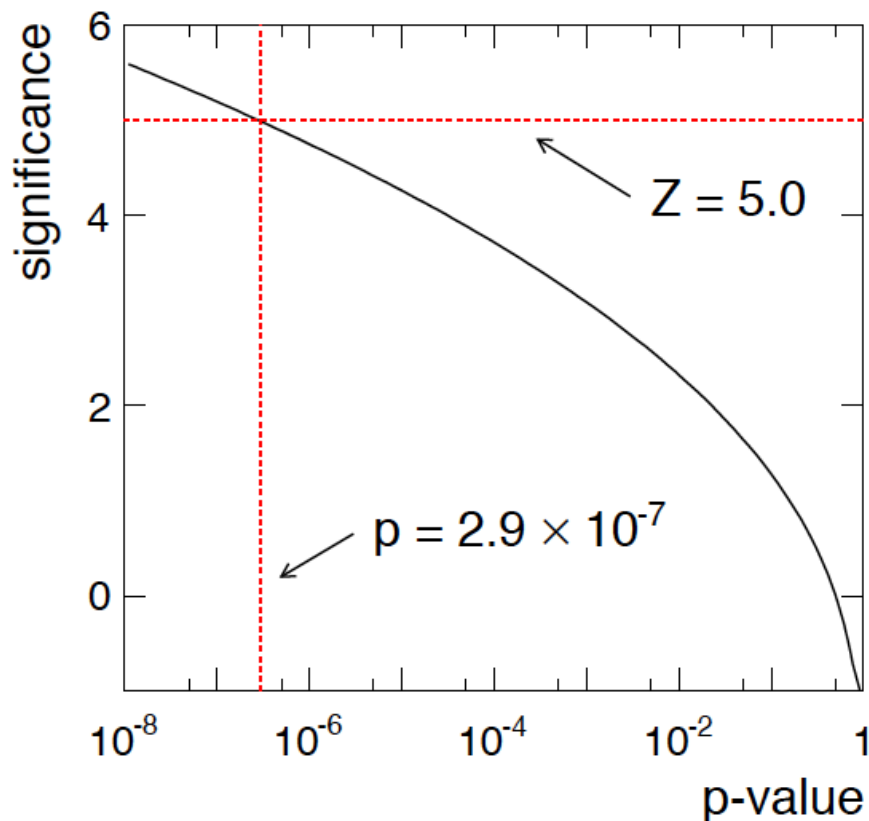
```
p = 1 - norm.cdf(Z) = norm.sf(Z)
Z = norm.ppf(1-p)
```

Result  $Z$  is a “number of sigmas”. Note this does not mean that the original data was Gaussian distributed.

# Poisson counting experiment: discovery significance

Equivalent significance for  $p = 1.7 \times 10^{-4}$ :  $Z = \Phi^{-1}(1 - p) = 3.6$

Often claim discovery if  $Z > 5$  ( $p < 2.9 \times 10^{-7}$ , i.e., a “5-sigma effect”)



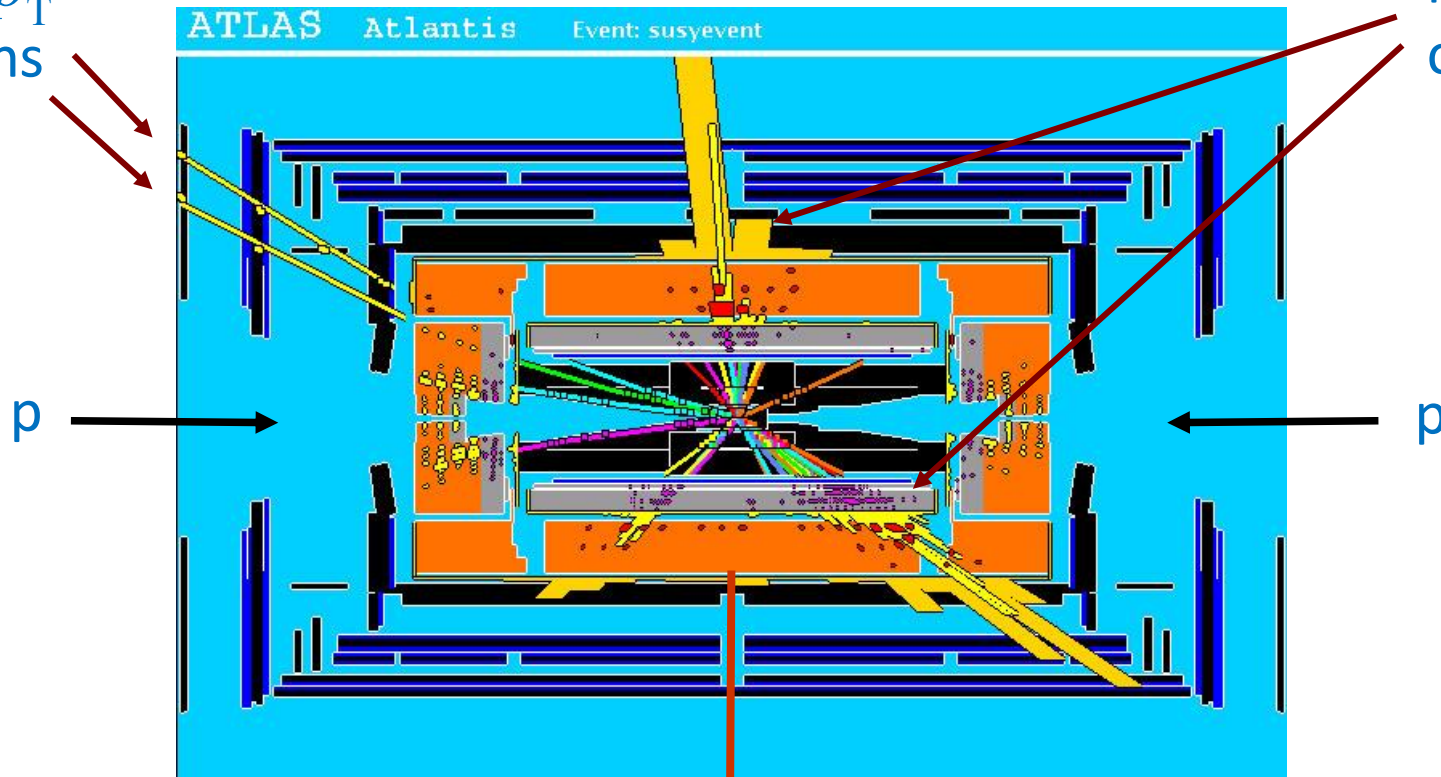
In fact this tradition should be revisited:  $p$ -value intended to quantify probability of a signal-like fluctuation assuming background only; not intended to cover, e.g., hidden systematics, plausibility signal model, compatibility of data with signal, “look-elsewhere effect” ( $\sim$ multiple testing), etc.

# Particle Physics context for a hypothesis test

A simulated SUSY event (“signal”):

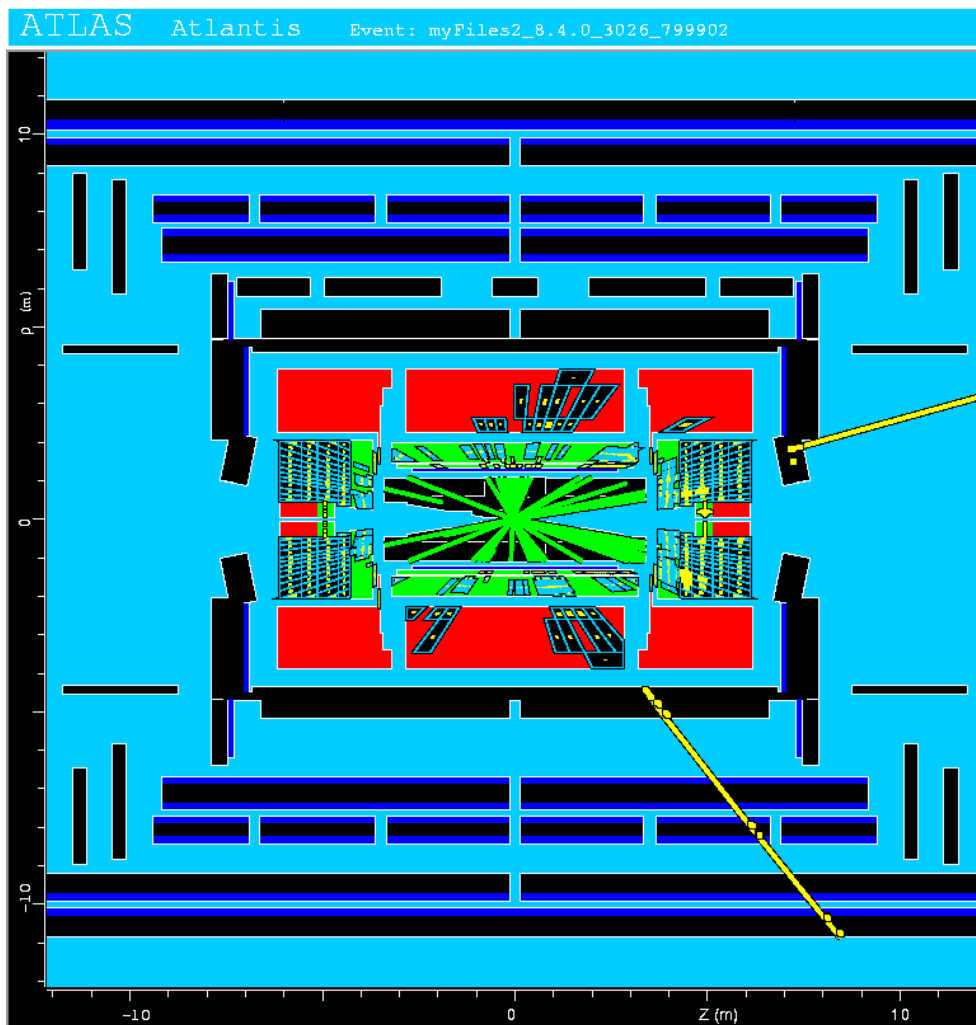
high  $p_T$   
muons

high  $p_T$  jets  
of hadrons



missing transverse energy

# Background events



This event from Standard Model  $t\bar{t}$  production also has high  $p_T$  jets and muons, and some missing transverse energy.

→ can easily mimic a signal event.

# Classification of proton-proton collisions

Proton-proton collisions can be considered to come in two classes:

signal (the kind of event we're looking for,  $y = 1$ )

background (the kind that mimics signal,  $y = 0$ )

For each collision (event), we measure a collection of features:

$x_1 =$  energy of muon

$x_2 =$  angle between jets

$x_3 =$  total jet energy

$x_4 =$  missing transverse energy

$x_5 =$  invariant mass of muon pair

$x_6 = \dots$

The real events don't come with true class labels, but computer-simulated events do. So we can have a set of simulated events that consist of a feature vector  $\mathbf{x}$  and true class label  $y$  (0 for background, 1 for signal):

$$(\mathbf{x}, y)_1, (\mathbf{x}, y)_2, \dots, (\mathbf{x}, y)_N$$

The simulated events are called “training data”.

# Distributions of the features

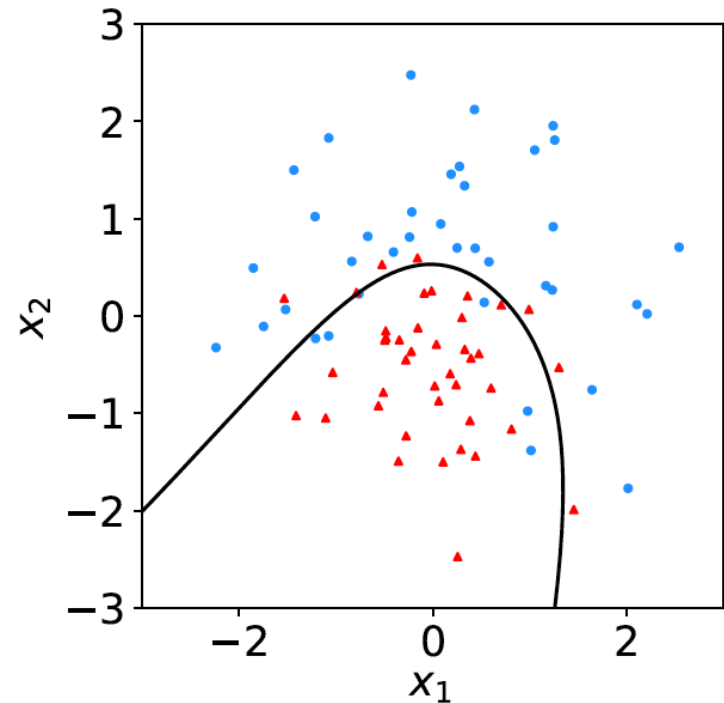
If we consider only two features  $\mathbf{x} = (x_1, x_2)$ , we can display the results in a scatter plot (red:  $y = 0$ , blue:  $y = 1$ ).

For real events, the dots are black (true type is not known).

For each real event test the hypothesis that it is background.

(Related to this: test that a sample of events is *all* background.)

The test's critical region is defined by a “**decision boundary**” – without knowing the event type, we can classify them by seeing where their measured features lie relative to the boundary.



# Decision function, test statistic

A surface in an  $n$ -dimensional space can be described by

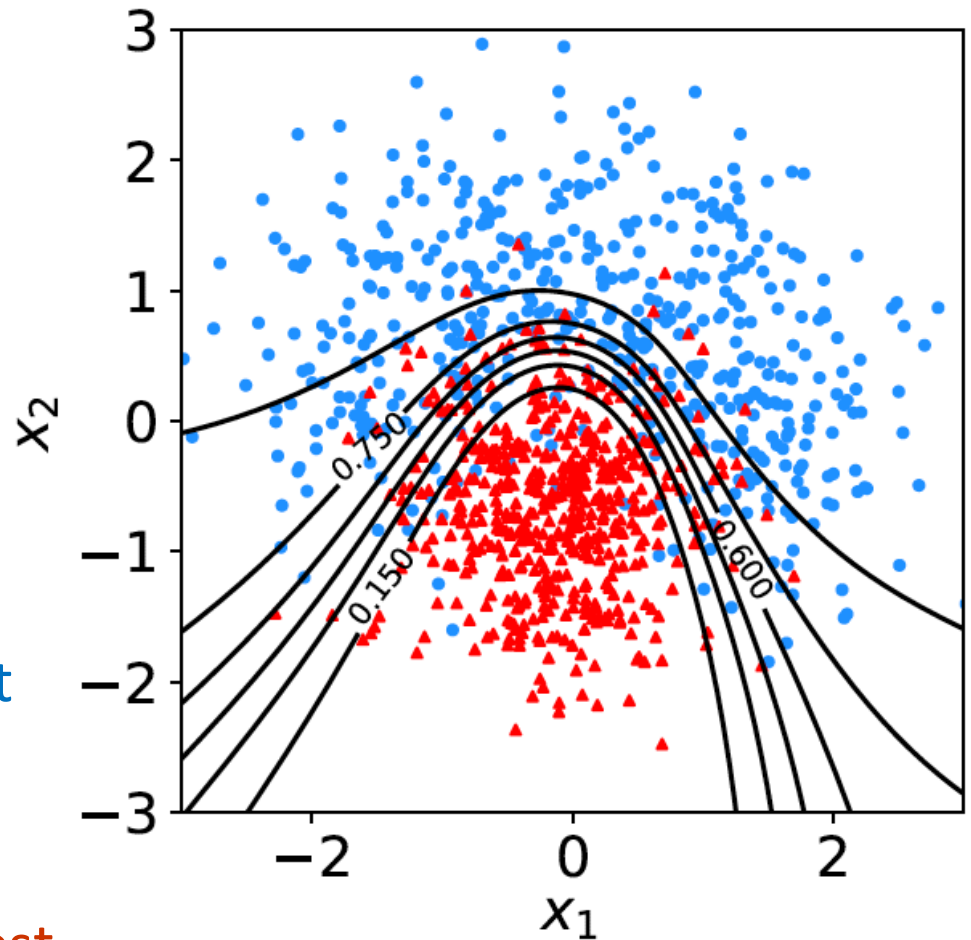
$$t(x_1, \dots, x_n) = t_c$$

scalar  
function

constant

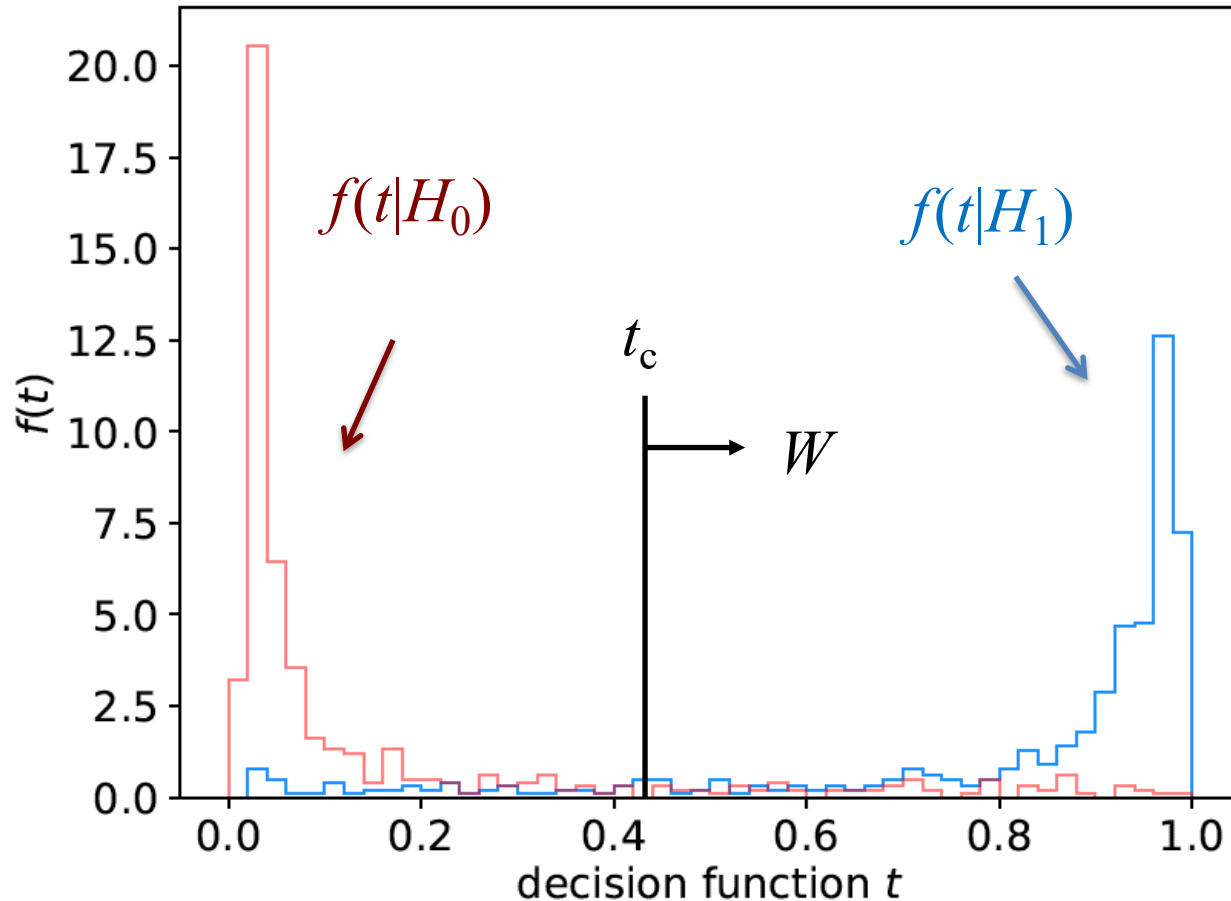
Different values of the constant  $t_c$  result in a family of surfaces.

Problem is reduced to finding the best **decision function** or **test statistic**  $t(\mathbf{x})$ .



# Distribution of $t(\mathbf{x})$

By forming a test statistic  $t(\mathbf{x})$ , the boundary of the critical region in the  $n$ -dimensional  $\mathbf{x}$ -space is determined by a single value  $t_c$ .

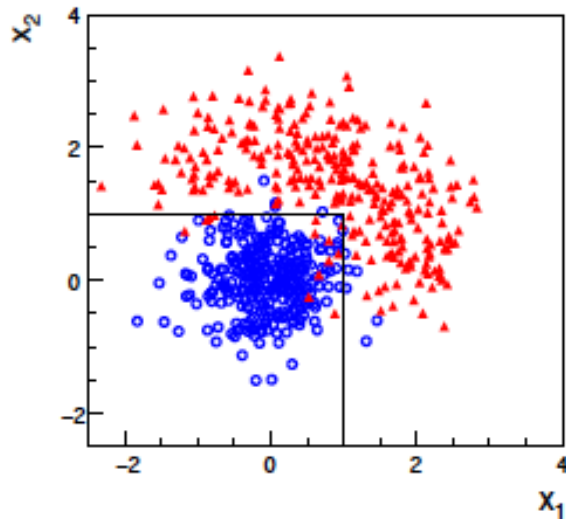


# Types of decision boundaries

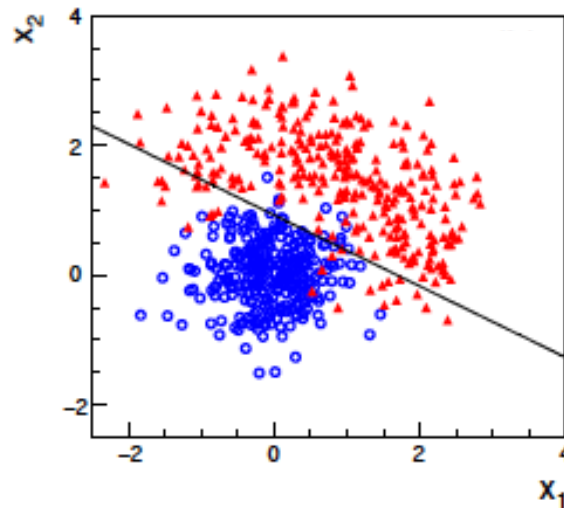
So what is the optimal boundary for the critical region, i.e., what is the optimal test statistic  $t(\mathbf{x})$ ?

First find best  $t(\mathbf{x})$ , later address issue of optimal size of test.

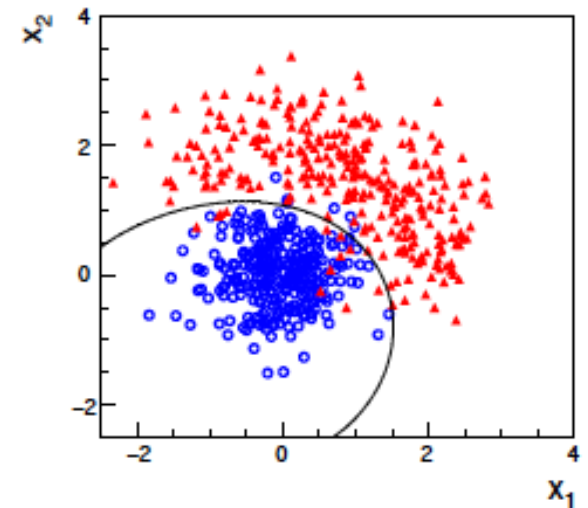
Remember  $\mathbf{x}$ -space can have many dimensions.



“cuts”



linear



non-linear

# Test statistic based on likelihood ratio

How can we choose a test's critical region in an 'optimal way', in particular if the data space is multidimensional?

Neyman-Pearson lemma states:

For a test of  $H_0$  of size  $\alpha$ , to get the highest power with respect to the alternative  $H_1$  we need for all  $\mathbf{x}$  in the critical region  $W$

"likelihood ratio (LR)"  $\longrightarrow \frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)} \geq c_\alpha$

inside  $W$  and  $\leq c_\alpha$  outside, where  $c_\alpha$  is a constant chosen to give a test of the desired size.

Equivalently, optimal scalar test statistic is

$$t(\mathbf{x}) = \frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)}$$

N.B. any monotonic function of this is leads to the same test.

# Proof of Neyman-Pearson Lemma

Consider a critical region  $W$  and suppose the LR satisfies the criterion of the Neyman-Pearson lemma:

$$P(\mathbf{x}|H_1)/P(\mathbf{x}|H_0) \geq c_\alpha \text{ for all } \mathbf{x} \text{ in } W,$$
$$P(\mathbf{x}|H_1)/P(\mathbf{x}|H_0) \leq c_\alpha \text{ for all } \mathbf{x} \text{ not in } W.$$

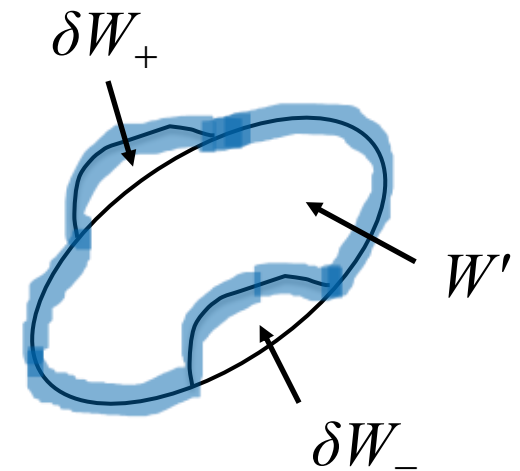
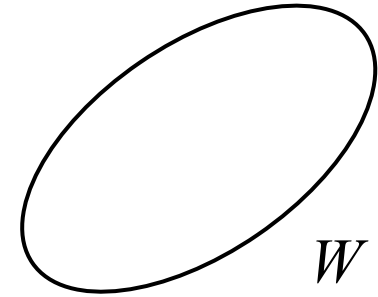
Try to change this into a different critical region  $W'$  retaining the same size  $\alpha$ , i.e.,

$$P(\mathbf{x} \in W'|H_0) = P(\mathbf{x} \in W|H_0) = \alpha$$

To do so add a part  $\delta W_+$ , but to keep the size  $\alpha$ , we need to remove a part  $\delta W_-$ , i.e.,

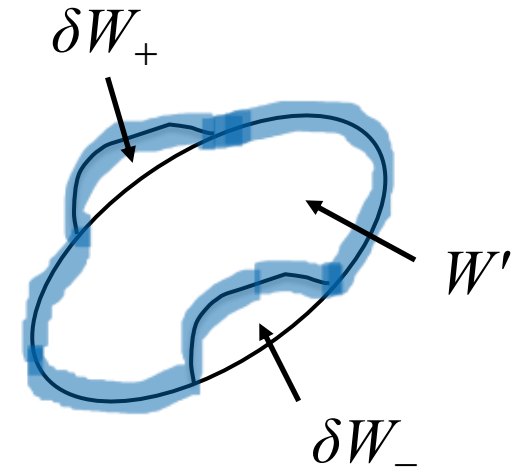
$$W \rightarrow W' = W + \delta W_+ - \delta W_-$$

$$P(\mathbf{x} \in \delta W_+|H_0) = P(\mathbf{x} \in \delta W_-|H_0)$$



## Proof of Neyman-Pearson Lemma (2)

But we are supposing the LR is higher for all  $\mathbf{x}$  in  $\delta W_-$  removed than for the  $\mathbf{x}$  in  $\delta W_+$  added, and therefore



$$P(\mathbf{x} \in \delta W_+ | H_1) \leq P(\mathbf{x} \in \delta W_+ | H_0) c_\alpha$$

$$P(\mathbf{x} \in \delta W_- | H_1) \geq P(\mathbf{x} \in \delta W_- | H_0) c_\alpha$$

The right-hand sides are equal and therefore

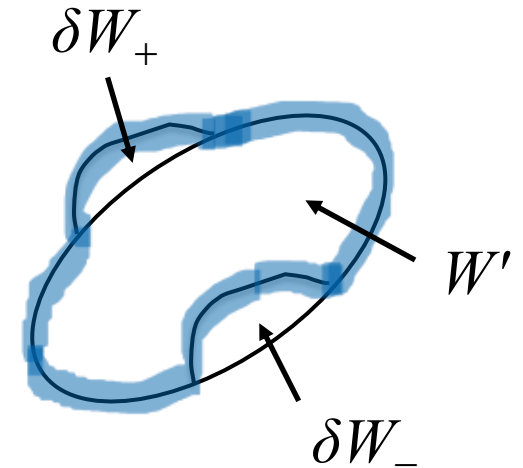
$$P(\mathbf{x} \in \delta W_+ | H_1) \leq P(\mathbf{x} \in \delta W_- | H_1)$$

# Proof of Neyman-Pearson Lemma (3)

We have

$$W \cup W' = W \cup \delta W_+ = W' \cup \delta W_-$$

Note  $W$  and  $\delta W_+$  are disjoint, and  $W'$  and  $\delta W_-$  are disjoint, so by Kolmogorov's 3<sup>rd</sup> axiom,



$$P(\mathbf{x} \in W') + P(\mathbf{x} \in \delta W_-) = P(\mathbf{x} \in W) + P(\mathbf{x} \in \delta W_+)$$

Therefore

$$P(\mathbf{x} \in W' | H_1) = P(\mathbf{x} \in W | H_1) + \underbrace{P(\mathbf{x} \in \delta W_+ | H_1) - P(\mathbf{x} \in \delta W_- | H_1)}_{\leq 0}$$

# Proof of Neyman-Pearson Lemma (4)

And therefore

$$P(\mathbf{x} \in W' | H_1) \leq P(\mathbf{x} \in W | H_1)$$

i.e. the deformed critical region  $W'$  cannot have higher power than the original one that satisfied the LR criterion of the Neyman-Pearson lemma.

# Neyman-Pearson doesn't usually help

We usually don't have explicit formulae for the pdfs  $f(\mathbf{x}|s)$ ,  $f(\mathbf{x}|b)$ , so for a given  $\mathbf{x}$  we can't evaluate the likelihood ratio

$$t(\mathbf{x}) = \frac{f(\mathbf{x}|s)}{f(\mathbf{x}|b)}$$

Instead we may have Monte Carlo models for signal and background processes, so we can produce simulated data:

generate  $\mathbf{x} \sim f(\mathbf{x}|s)$   $\rightarrow$   $\mathbf{x}_1, \dots, \mathbf{x}_N$

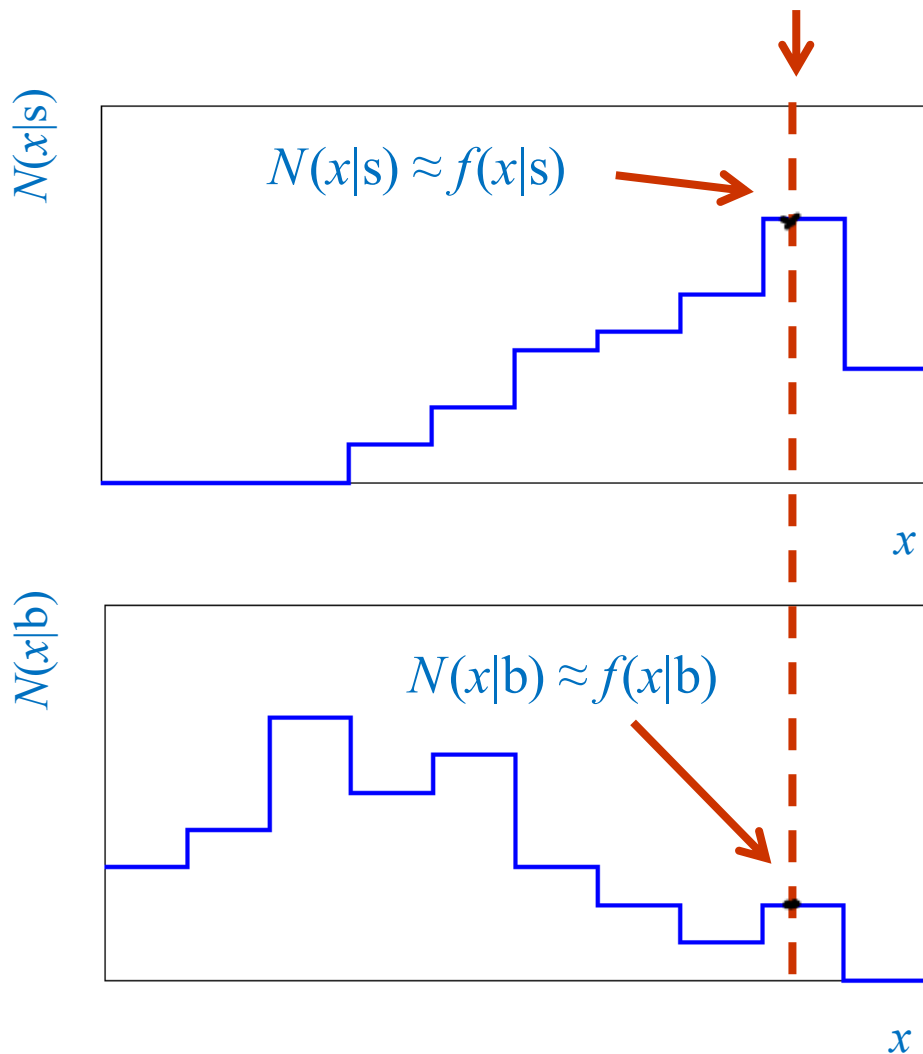
generate  $\mathbf{x} \sim f(\mathbf{x}|b)$   $\rightarrow$   $\mathbf{x}_1, \dots, \mathbf{x}_N$

This gives samples of “training data” with events of known type.

- Use these to construct a statistic that is as close as possible to the optimal likelihood ratio ( $\rightarrow$  Machine Learning).

# Approximate LR from histograms

Want  $t(x) = f(x|s)/f(x|b)$  for  $x$  here



One possibility is to generate MC data and construct histograms for both signal and background.

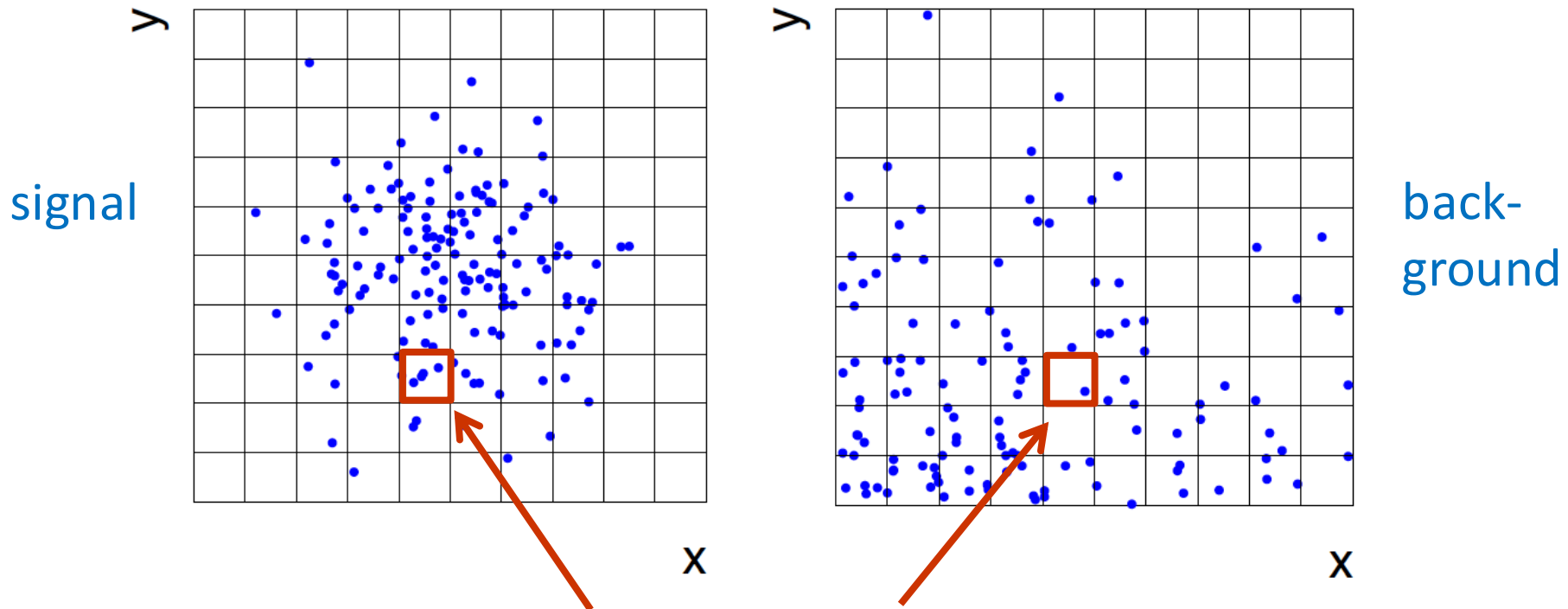
Use (normalized) histogram values to approximate LR:

$$t(x) \approx \frac{N(x|s)}{N(x|b)}$$

Can work well for single variable.

# Approximate LR from 2D-histograms

Suppose problem has 2 variables. Try using 2-D histograms:



Approximate pdfs using  $N(x,y|s)$ ,  $N(x,y|b)$  in corresponding cells.

But if we want  $M$  bins for each variable, then in  $n$ -dimensions we have  $M^n$  cells; can't generate enough training data to populate.

→ Histogram method usually not usable for  $n > 1$  dimension.

# Strategies for multivariate analysis

Neyman-Pearson lemma gives optimal answer, but cannot be used directly, because we usually don't have  $f(\mathbf{x}|s)$ ,  $f(\mathbf{x}|b)$ .

Histogram method with  $M$  bins for  $n$  variables requires that we estimate  $M^n$  parameters (the values of the pdfs in each cell), so this is rarely practical.

A compromise solution is to assume a certain functional form for the test statistic  $t(\mathbf{x})$  with fewer parameters; determine them (using MC) to give best separation between signal and background.

Alternatively, try to estimate the probability densities  $f(\mathbf{x}|s)$  and  $f(\mathbf{x}|b)$  (with something better than histograms) and use the estimated pdfs to construct an approximate likelihood ratio.

# Multivariate methods (Machine Learning)

Many new (and some old) methods:

Fisher discriminant

(Deep) Neural Networks

Kernel density methods

Support Vector Machines

Decision trees

Boosting

Bagging

More in the lectures on Machine Learning

# Extra slides

# Some statistics books, papers, etc.

G. Cowan, *Statistical Data Analysis*, Clarendon, Oxford, 1998

R.J. Barlow, *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences*, Wiley, 1989

Ilya Narsky and Frank C. Porter, *Statistical Analysis Techniques in Particle Physics*, Wiley, 2014.

Luca Lista, *Statistical Methods for Data Analysis in Particle Physics*, Springer, 2017.

L. Lyons, *Statistics for Nuclear and Particle Physics*, CUP, 1986

F. James., *Statistical and Computational Methods in Experimental Physics*, 2nd ed., World Scientific, 2006

S. Brandt, *Statistical and Computational Methods in Data Analysis*, Springer, New York, 1998.

S. Navas et al. (Particle Data Group), Phys. Rev. D 110, 030001 (2024); [pdg.lbl.gov](https://pdg.lbl.gov) sections on probability, statistics, MC.