

# Statistical Data Analysis 2022/23

## 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”).

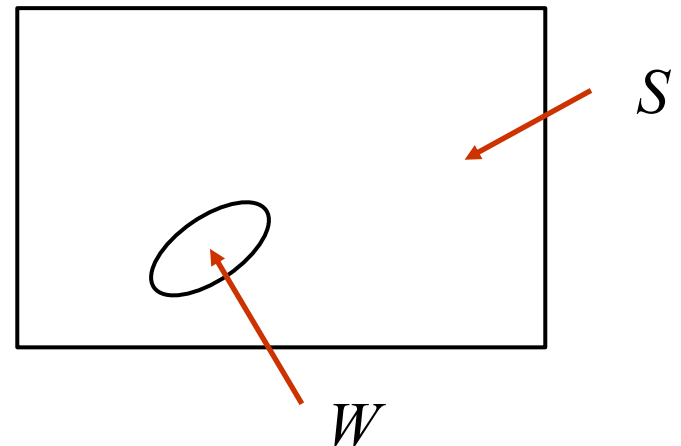
Consider a simple hypothesis  $H_0$  (the “null”) and an alternative  $H_1$ .

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  $\alpha$ , assuming  $H_0$  is correct, to observe the data there, i.e.,

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

$\alpha$  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  $H_0$ .

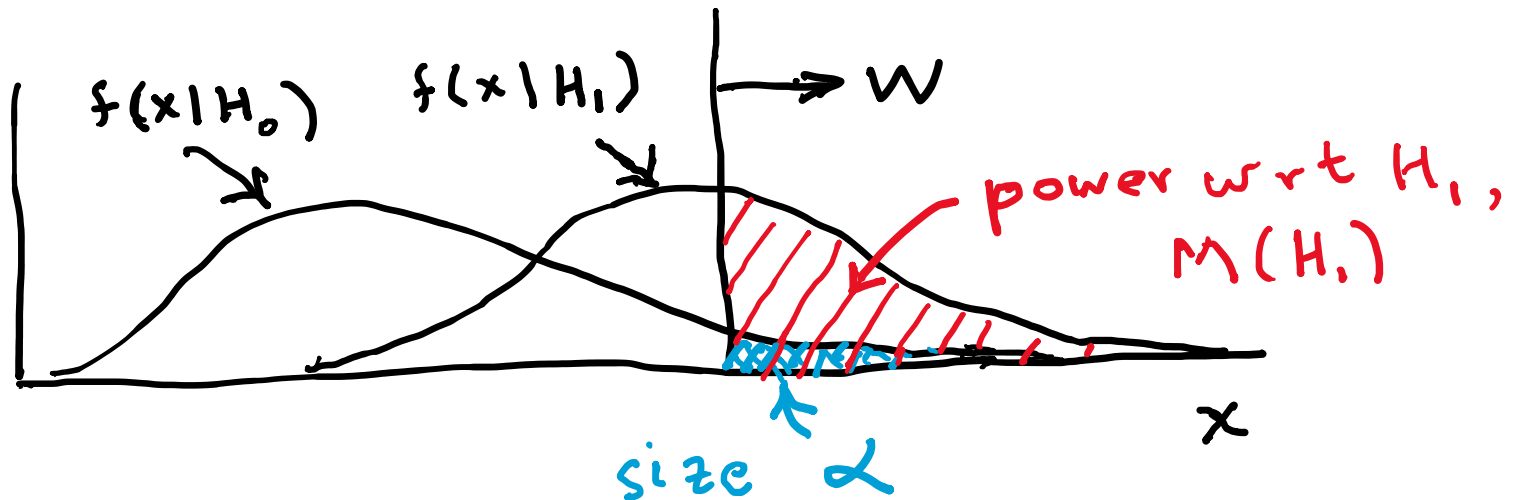


## Definition of a test (2)

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

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:

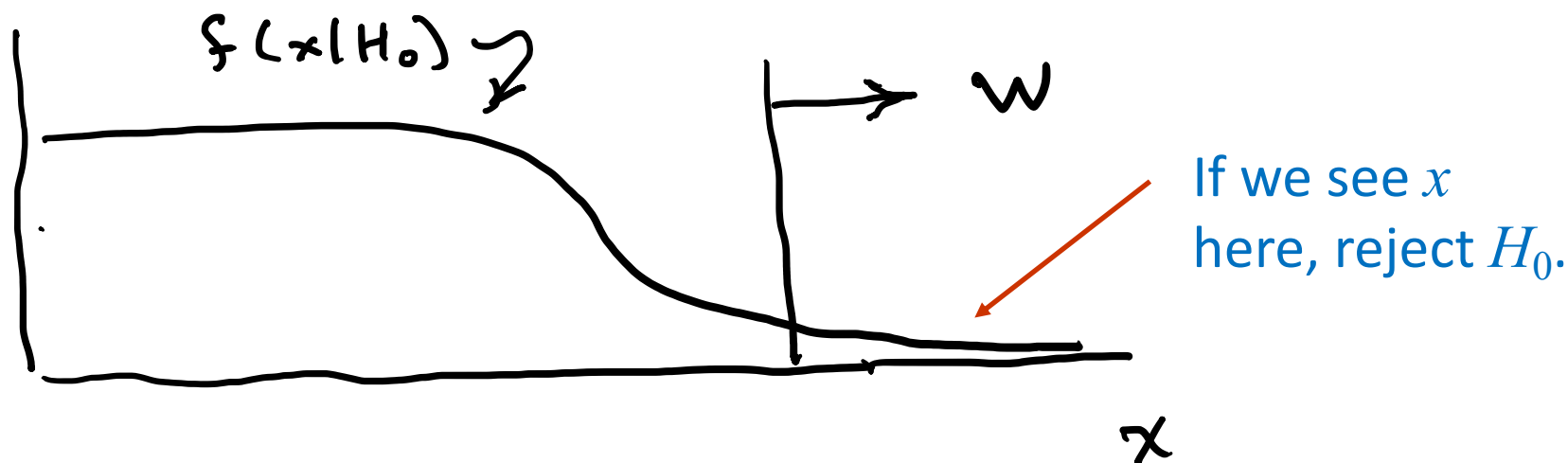


# Obvious where to put $W$ ?

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

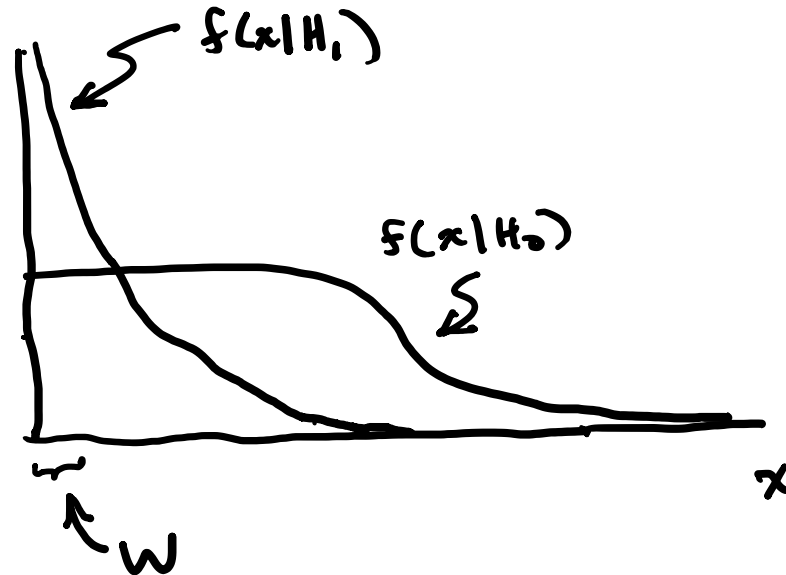
Fisher held that one could test a hypothesis  $H_0$  without reference to an alternative.

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  $x$ :



# 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



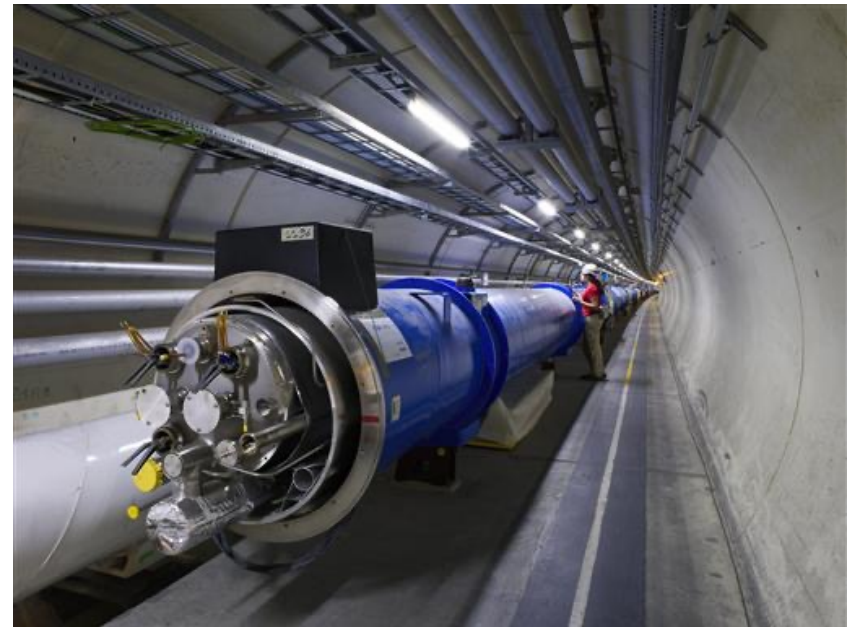
Counter-rotating proton beams  
in 27 km circumference ring

pp centre-of-mass energy 14 TeV

Detectors at 4 pp collision points:

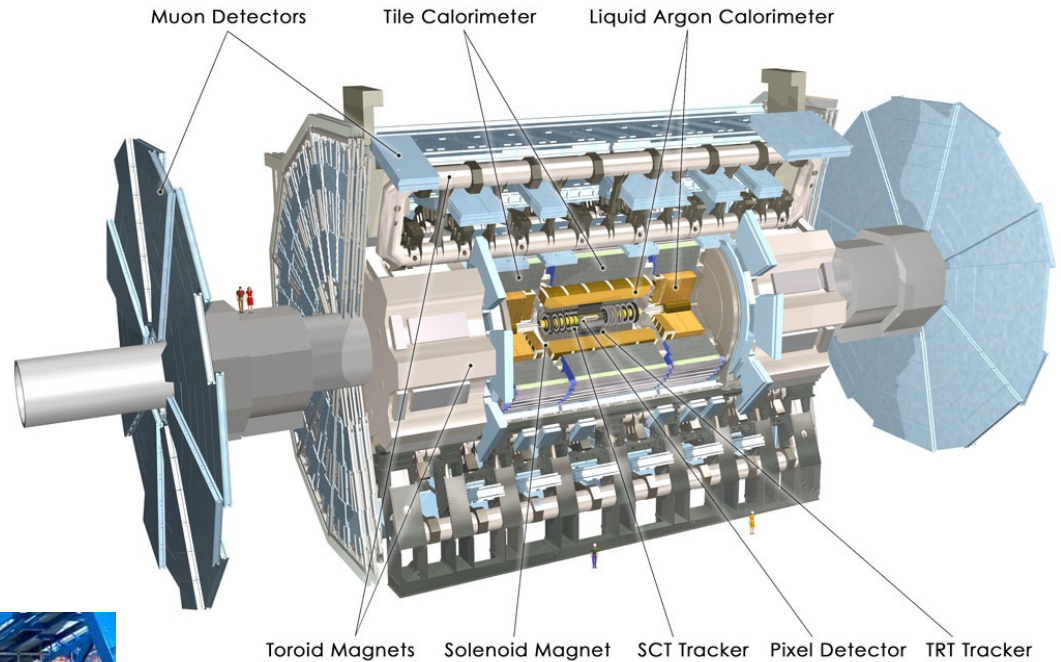
- ATLAS
- CMS
- LHCb (b physics)
- ALICE (heavy ion physics)

general purpose

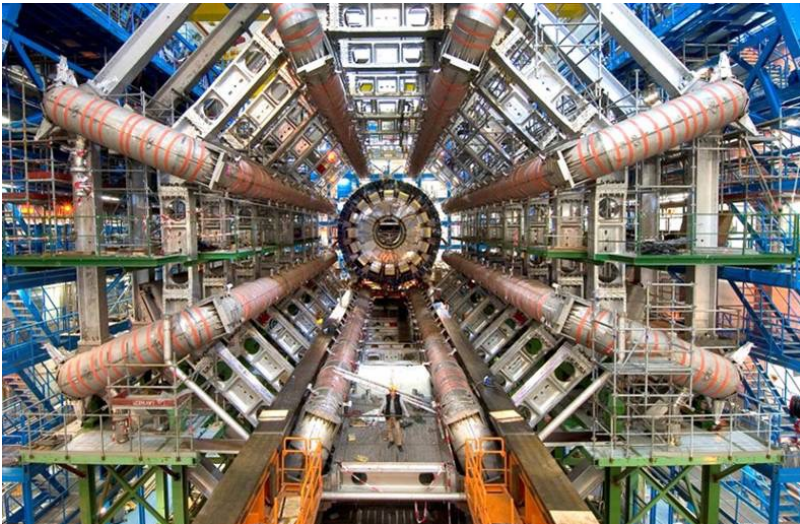


# The ATLAS detector

3000 physicists  
38 countries  
183 universities/labs



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

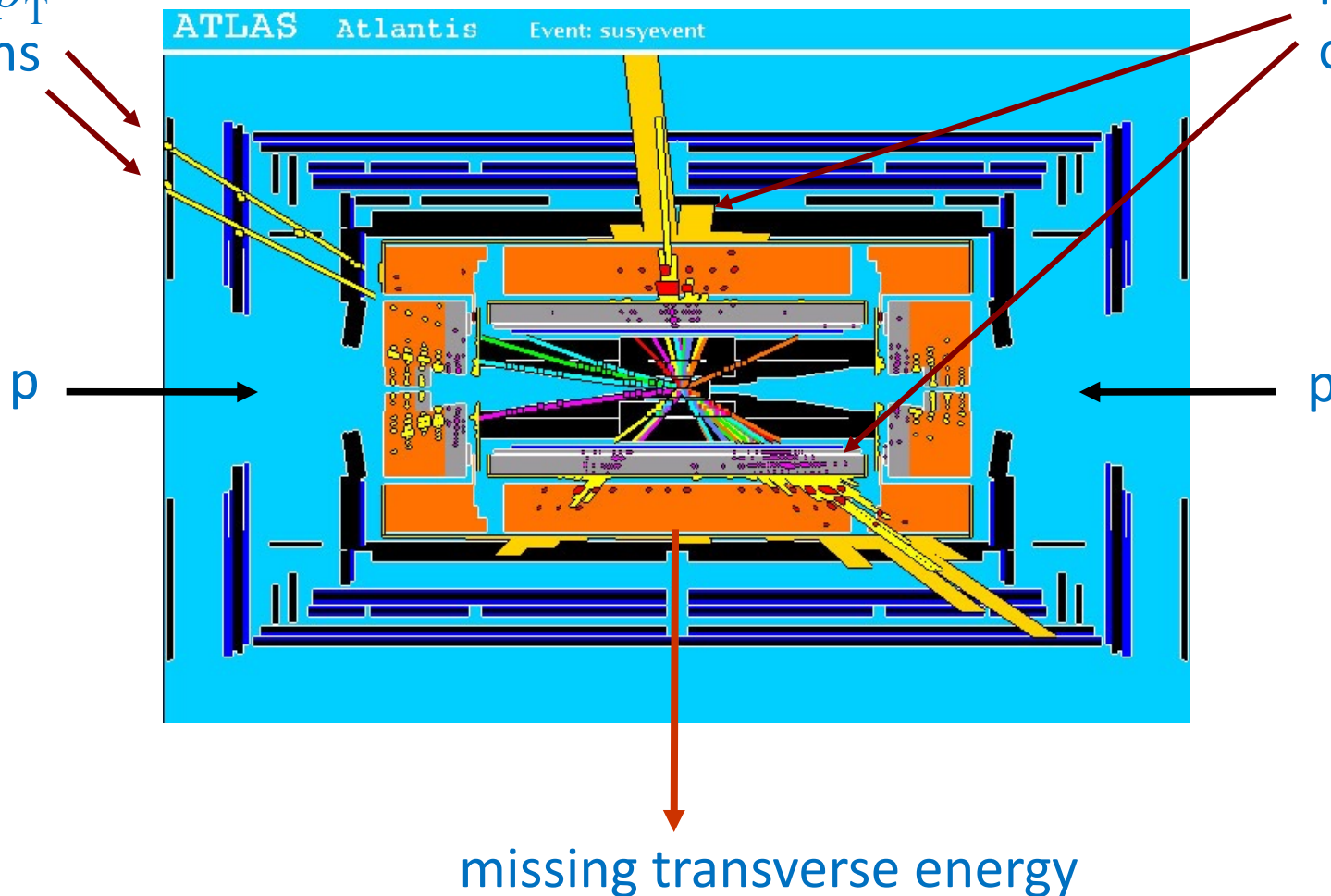




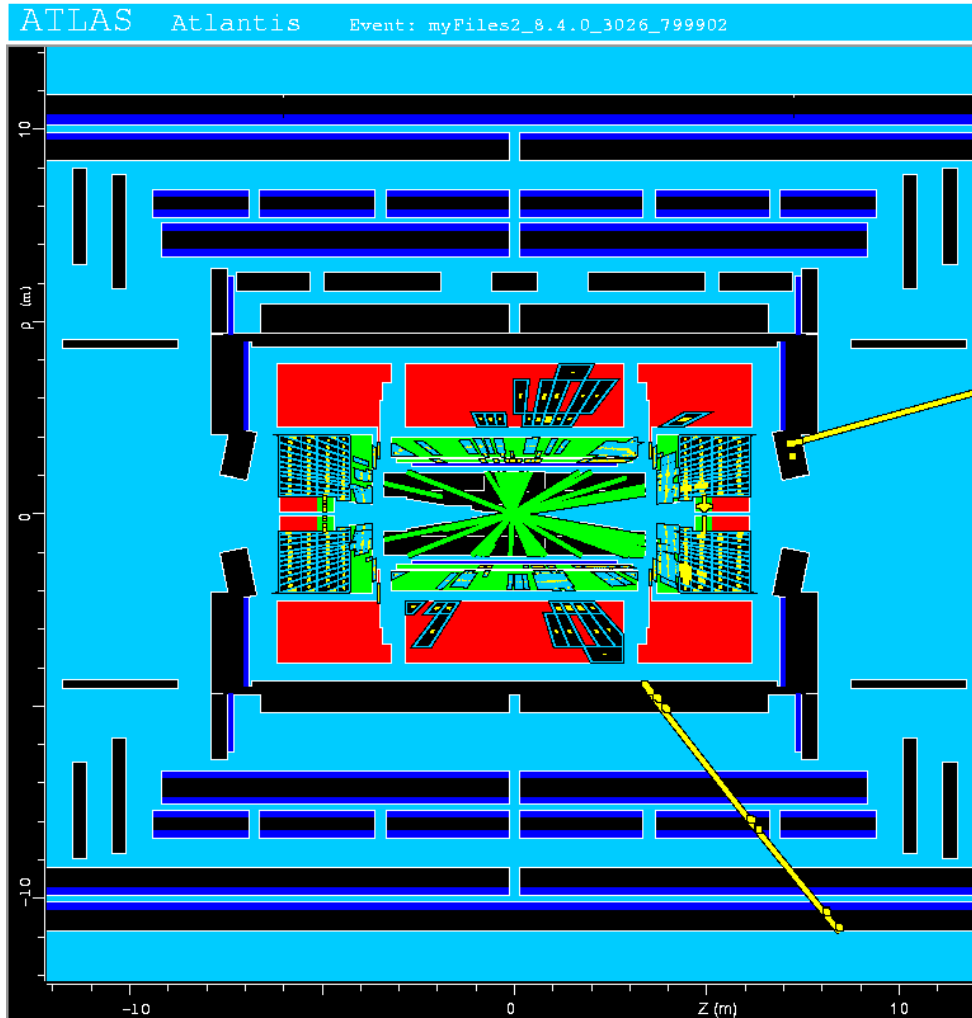
# A simulated SUSY event

high  $p_T$   
muons

high  $p_T$  jets  
of hadrons



# Background events



This event from Standard Model  $t\bar{t}b\bar{b}$  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  $\varepsilon_b = \int_W f(\mathbf{x}|H_0) d\mathbf{x} = \alpha$

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

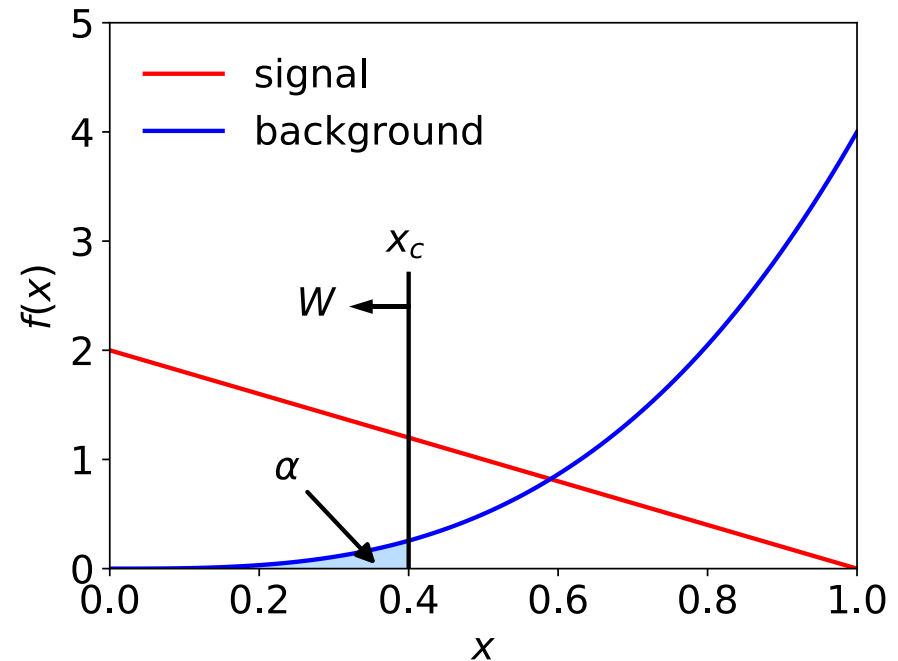
# Example of a test for classification

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:

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

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:

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

## Classification example (4)

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

$$\begin{aligned} P(b|x) &= \frac{f(x|b)\pi_b}{f(x|b)\pi_b + f(x|s)\pi_s} \\ &= \frac{4x^3\pi_b}{4x^3\pi_b + 2(1-x)\pi_s} \\ &= 0.689 \end{aligned}$$

(Here nothing to do with the test using  $x \leq x_c$ , just an illustration of Bayes' theorem.)

# Statistical Data Analysis

## Lecture 4-3

- Hypothesis test for classification
- Test statistic to define critical region
- Neyman-Pearson lemma

# Classifying fish

You scoop up fish which are of two types:

Sea  
Bass



Cod

You examine the fish with automatic sensors and for each one you measure a set of **features**:

$x_1$  = length

$x_2$  = width

$x_3$  = weight

$x_4$  = area of fins

$x_5$  = mean spectral reflectance

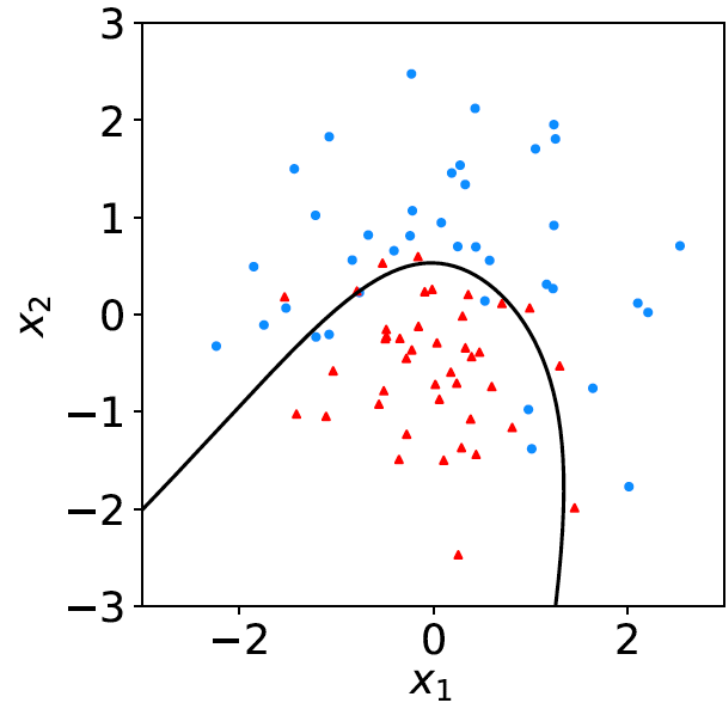
$x_6$  = ...

These constitute the “**feature vector**”  $\mathbf{x} = (x_1, \dots, x_n)$ .

In addition you hire a fish expert to identify the “true class label”  $y = 0$  or  $1$  (i.e.,  $0$  = sea bass,  $1$  = cod) for each fish. We thus obtain “**training data**”:  $(\mathbf{x}, y)_1, (\mathbf{x}, y)_2, \dots, (\mathbf{x}, y)_N$ .

# 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:  $y = 1$ ).



Goal is to determine a decision boundary, so that, without the help of the fish expert, we can classify new fish by seeing where their measured features lie relative to the boundary.

Same idea in multi-dimensional feature space, but cannot represent as 2-D plot. Decision boundary is  $n$ -dim. hypersurface.

# Decision function, test statistic

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

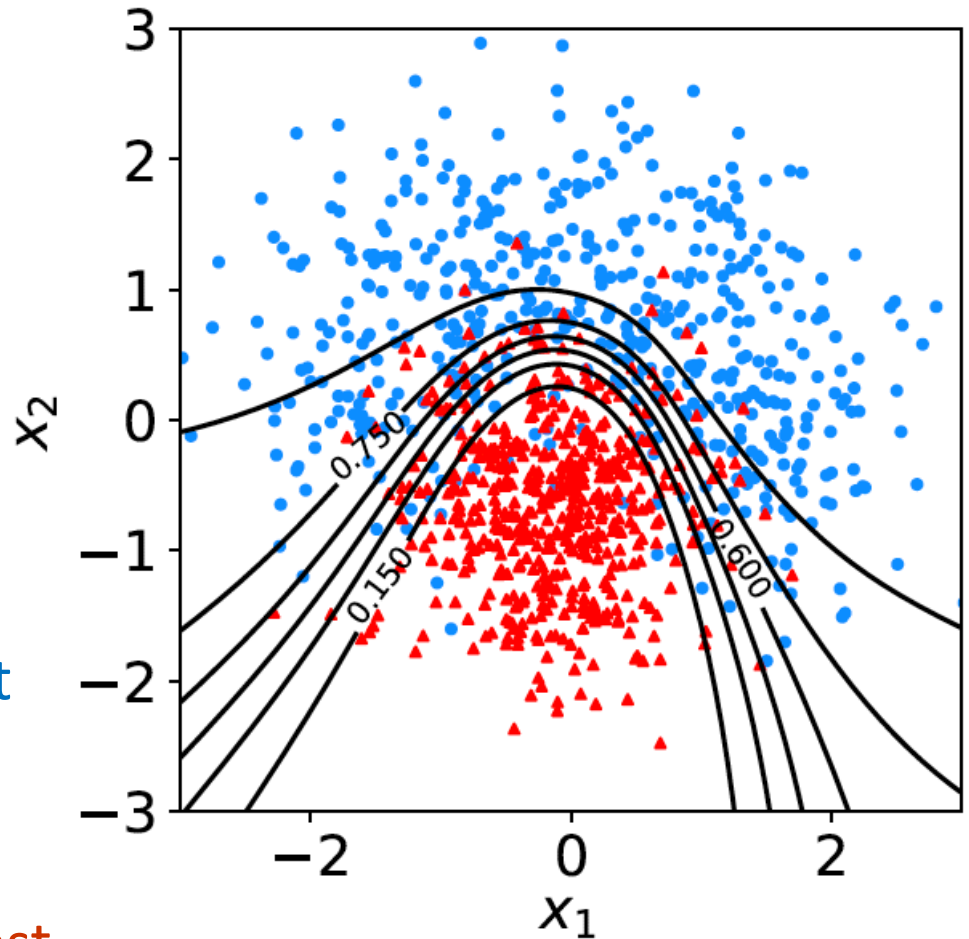
$$t(x_1, \dots, x_n) = t_c$$

scalar  
function

constant

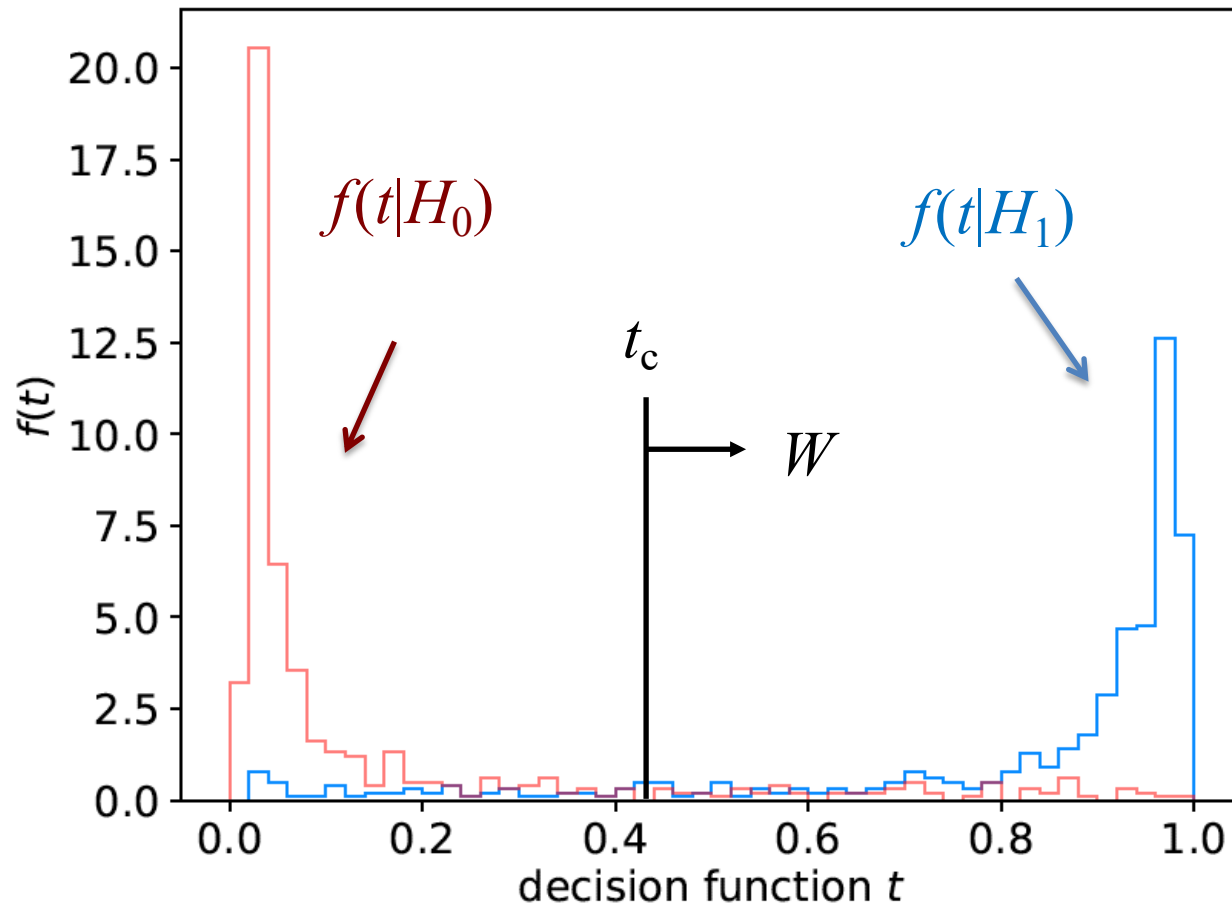
Different values of the constant  $t_c$  result in a family of surfaces.

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  $t_c$ .



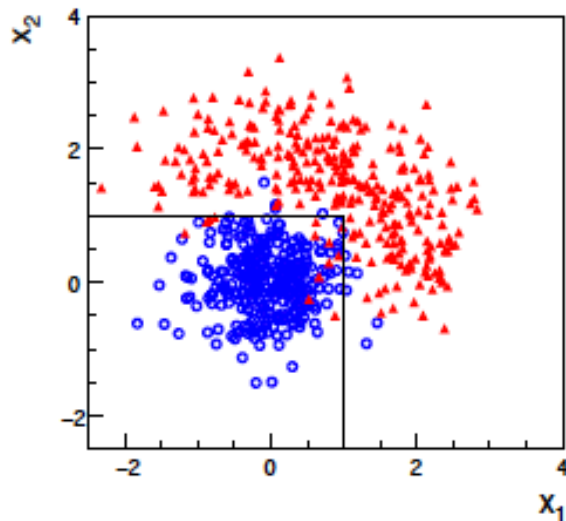


# Types of decision boundaries

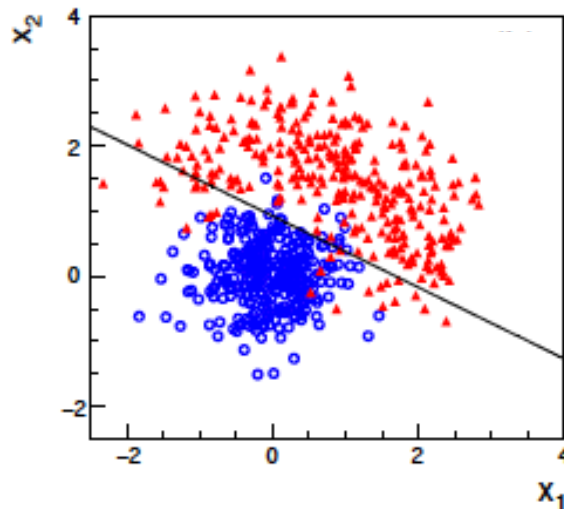
So what is the optimal boundary for the critical region, i.e., what is the optimal test statistic  $t(\mathbf{x})$ ?

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

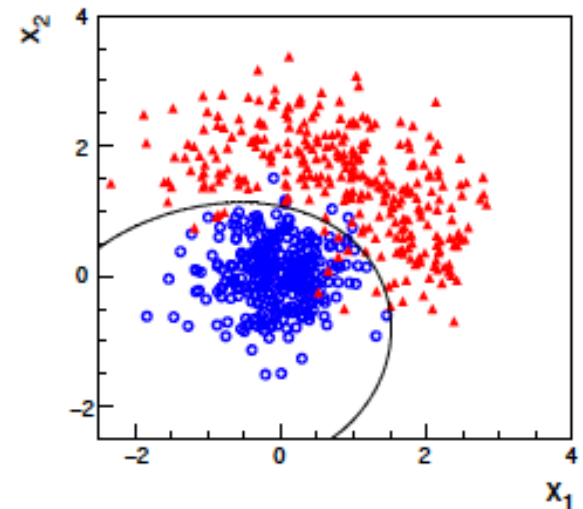
Remember  $\mathbf{x}$ -space can have many dimensions.



“cuts”



linear




non-linear

# Test statistic based on likelihood ratio

How can we choose a test's critical region in an 'optimal way'?

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

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

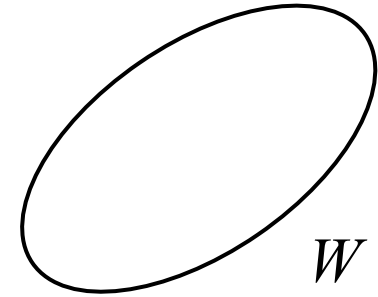
$$t(\mathbf{x}) = \frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)}$$

N.B. any monotonic function of this 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:

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



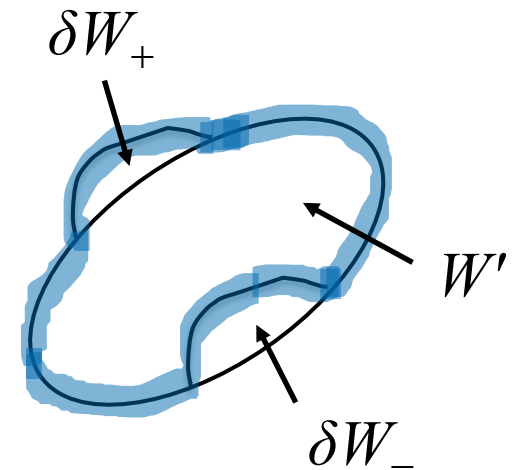
Try to change this into a different critical region  $W'$  retaining the same size  $\alpha$ , i.e.,

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

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)

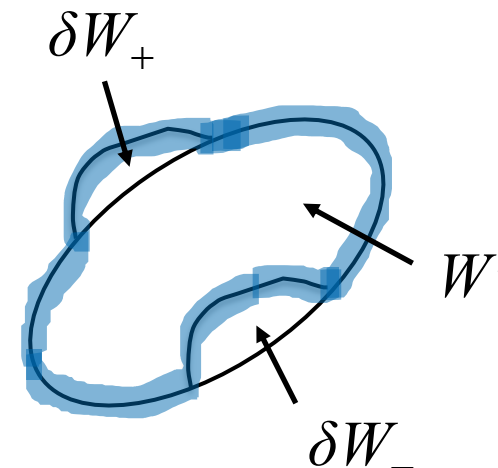
But we are supposing the LR is higher for all  $\mathbf{x}$  in  $\delta W_-$  removed than for the  $\mathbf{x}$  in  $\delta W_+$  added, and therefore

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

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

The right-hand sides are equal and therefore

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

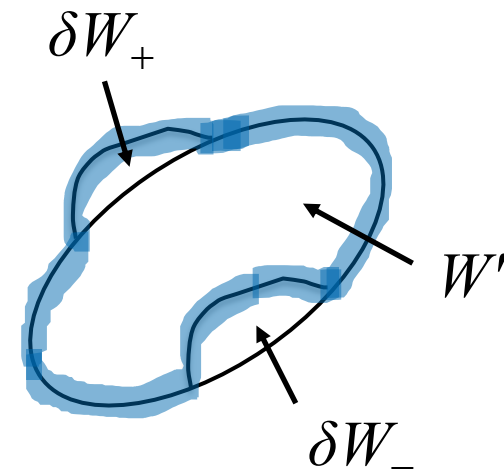


# Proof of Neyman-Pearson Lemma (3)

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.

# Statistical Data Analysis

## Lecture 4-4

- Why the Neyman-Pearson lemma usually doesn't help us
- Strategies for multivariate analysis
- Linear discriminant analysis

# 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.

- Can be expensive (1 fully simulated LHC event  $\sim$  1 CPU minute).



# How is it we don't have $f(\mathbf{x}|H)$ ?

In a Monte Carlo simulation of a complex process, the fundamental hypothesis does not predict the pdf for the finally measured variables  $\mathbf{x}$  but rather for some intermediate set of "latent" variables, say,  $\mathbf{z}_1$ .

So in step 1 we sample  $\mathbf{z}_1 \sim f(\mathbf{z}_1|H)$ , followed by many further intermediate steps:

$$\mathbf{z}_2 \sim f(\mathbf{z}_2|\mathbf{z}_1)$$

$$\mathbf{z}_3 \sim f(\mathbf{z}_3|\mathbf{z}_2)$$

$$\vdots$$

$$\mathbf{x} \sim f(\mathbf{x}|\mathbf{z}_n)$$

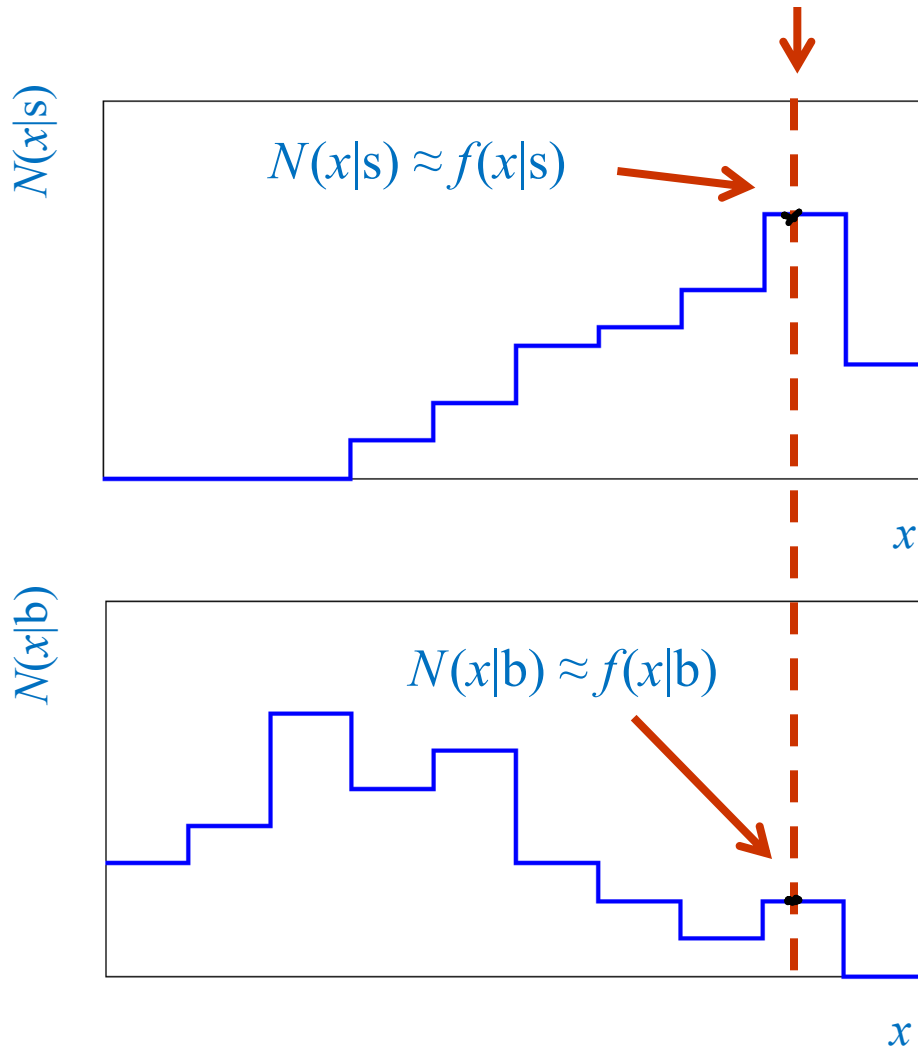
See, e.g., Kyle Cranmer, Johann Brehmer, Gilles Louppe, *The frontier of simulation-based inference*, arXiv:1911.01429 [stat.ML], PNAS doi.org/10.1073/pnas.1912789117

So even though  $H$  is fully defined and we can generate  $\mathbf{x}$  according to it, the formula for  $f(\mathbf{x}|H)$  is an enormous integral that we cannot compute:

$$f(\mathbf{x}|H) = \int \cdots \int d\mathbf{z}_1 \cdots d\mathbf{z}_n f(\mathbf{x}|\mathbf{z}_n) f(\mathbf{z}_n|\mathbf{z}_{n-1}) \cdots f(\mathbf{z}_1|\mathbf{z}_1) f(\mathbf{z}_1|H)$$

# 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.

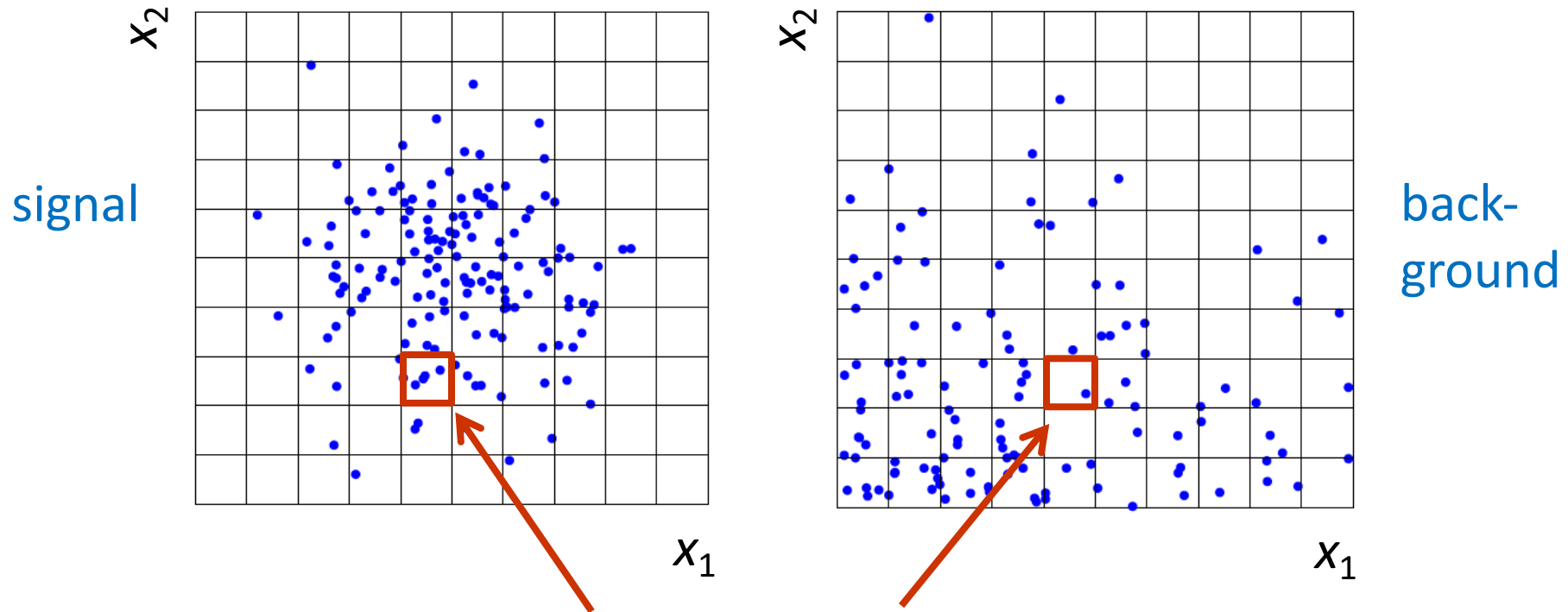
Use (normalized) histogram values to approximate LR:

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

Can work well for single variable.

# Approximate LR from 2D-histograms

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



Approximate pdfs using  $N(x_1, x_2|s)$ ,  $N(x_1, x_2|b)$  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  $f(\mathbf{x}|\mathbf{s})$ ,  $f(\mathbf{x}|\mathbf{b})$ .

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  $f(\mathbf{x}|\mathbf{s})$  and  $f(\mathbf{x}|\mathbf{b})$  (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

- Neural networks

- Kernel density methods

- Support Vector Machines

- Decision trees

  - Boosting

  - Bagging

# Resources on multivariate methods

C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006

T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, 2<sup>nd</sup> ed., Springer, 2009

Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, *An Introduction to Statistical Learning*, Springer, 2017, [faculty.marshall.usc.edu/gareth-james/ISL/](http://faculty.marshall.usc.edu/gareth-james/ISL/)

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

朱永生（编著），实验数据多元统计分析，科学出版社，北京，2009。

# Software

Rapidly growing area of development – two important resources:

## scikit-learn

Python-based tools for Machine Learning

**scikit-learn.org**

Large user community

.

## TMVA, Höcker, Stelzer, Tegenfeldt, Voss, Voss, physics/0703039

From **tmva.sourceforge.net**, also distributed with ROOT

Variety of classifiers

Good manual, widely used in HEP

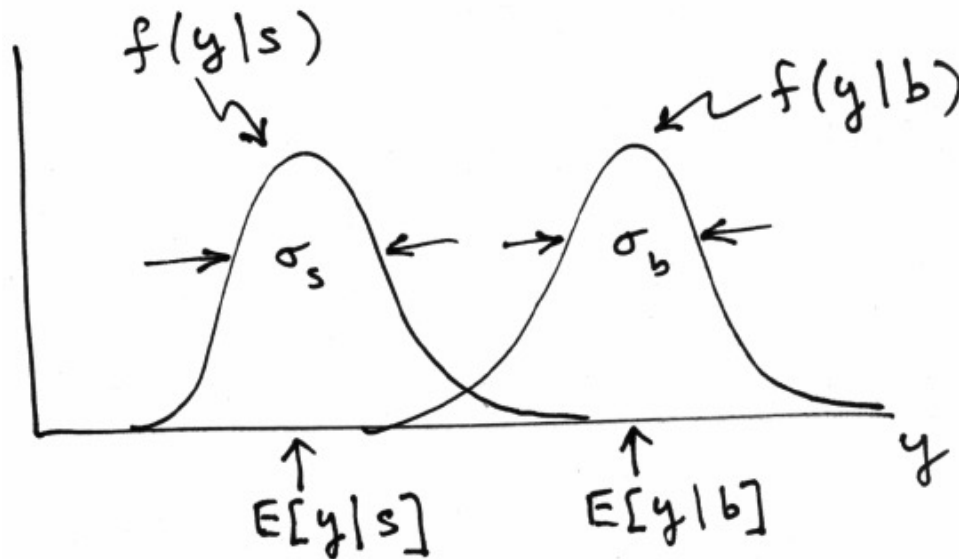
.

# Linear test statistic

Suppose there are  $n$  input variables:  $\mathbf{x} = (x_1, \dots, x_n)$ .

Consider a linear function: 
$$y(\mathbf{x}) = \sum_{i=1}^n w_i x_i$$

For a given choice of the coefficients  $\mathbf{w} = (w_1, \dots, w_n)$  we will get pdfs  $f(y|s)$  and  $f(y|b)$  :





# Linear test statistic

Fisher: to get large difference between means and small widths for  $f(y|s)$  and  $f(y|b)$ , maximize the difference squared of the expectation values divided by the sum of the variances:

$$J(\mathbf{w}) = \frac{(E[y|s] - E[y|b])^2}{V[y|s] + V[y|b]}$$

Setting  $\partial J / \partial w_i = 0$  gives:

$$\mathbf{w} \propto W^{-1}(\mu_b - \mu_s)$$

$$W_{ij} = \text{cov}[x_i, x_j|s] + \text{cov}[x_i, x_j|b]$$

$$\mu_{i,s} = E[x_i|s], \quad \mu_{i,b} = E[x_i|b]$$

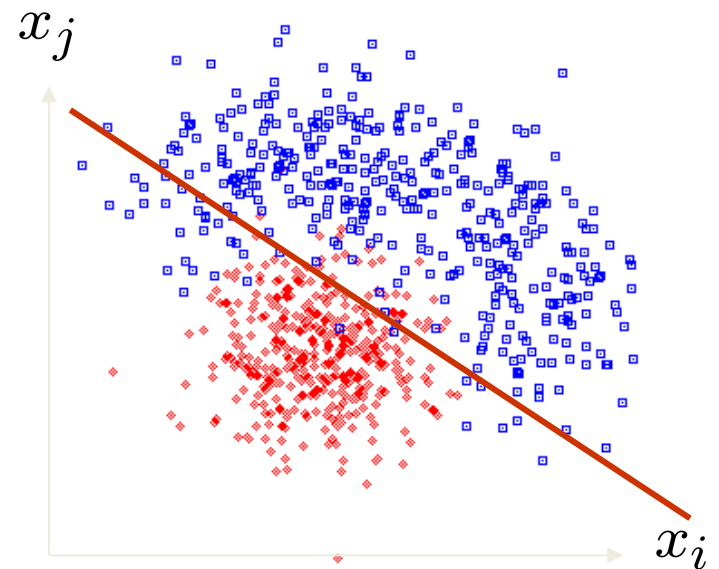
# The Fisher discriminant

The resulting coefficients  $w_i$  define a Fisher discriminant.

Coefficients defined up to multiplicative constant; can also add arbitrary offset, i.e., usually define test statistic as

$$y(\mathbf{x}) = w_0 + \sum_{i=1}^n w_i x_i$$

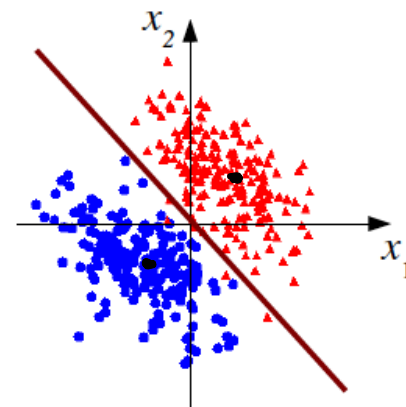
Boundaries of the test's critical region are surfaces of constant  $y(\mathbf{x})$ , here linear (hyperplanes):



# Fisher discriminant for Gaussian data

Suppose the pdfs of the input variables,  $f(\mathbf{x}|\mathbf{s})$  and  $f(\mathbf{x}|\mathbf{b})$ , are both multivariate Gaussians with same covariance but different means:

$$\begin{aligned} f(\mathbf{x}|\mathbf{s}) &= \text{Gauss}(\boldsymbol{\mu}_{\mathbf{s}}, V) \\ f(\mathbf{x}|\mathbf{b}) &= \text{Gauss}(\boldsymbol{\mu}_{\mathbf{b}}, V) \end{aligned} \quad \begin{array}{l} \leftarrow \text{Same covariance} \\ \leftarrow V_{ij} = \text{cov}[x_i, x_j] \end{array}$$



In this case it can be shown that the Fisher discriminant is

$$y(\mathbf{x}) \sim \ln \frac{f(\mathbf{x}|\mathbf{s})}{f(\mathbf{x}|\mathbf{b})}$$

i.e., it is a monotonic function of the likelihood ratio and thus leads to the same critical region. So in this case the Fisher discriminant provides an optimal statistical test.

# Extra slides

# Choosing a critical region

To construct a test of a hypothesis  $H_0$ , we can ask what are the relevant alternatives for which one would like to have a high power.

Maximize power wrt  $H_1$  = maximize probability to reject  $H_0$  if  $H_1$  is true.

Often such a test has a high power not only with respect to a specific point alternative but for a class of alternatives.

E.g., using a measurement  $x \sim \text{Gauss}(\mu, \sigma)$  we may test

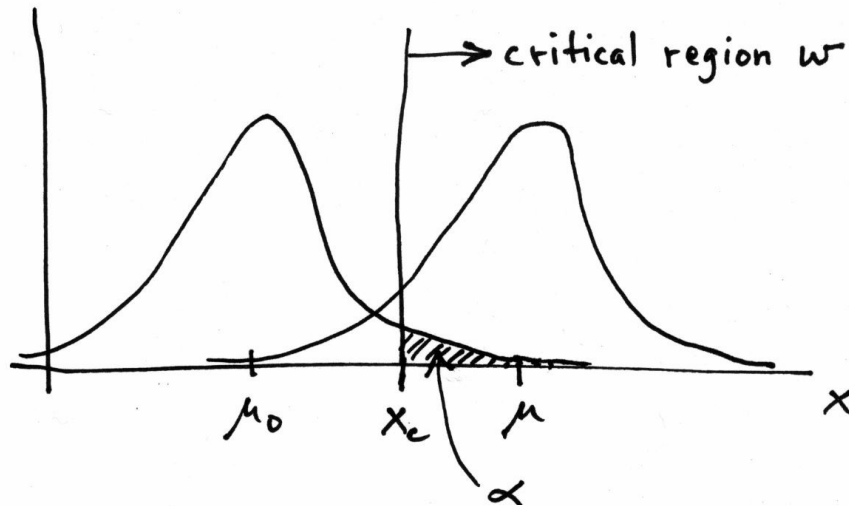
$H_0 : \mu = \mu_0$  versus the composite alternative  $H_1 : \mu > \mu_0$

We get the highest power with respect to any  $\mu > \mu_0$  by taking the critical region  $x \geq x_c$  where the cut-off  $x_c$  is determined by the significance level such that

$$\alpha = P(x \geq x_c | \mu_0).$$

# Test of $\mu = \mu_0$ vs. $\mu > \mu_0$ with $x \sim \text{Gauss}(\mu, \sigma)$

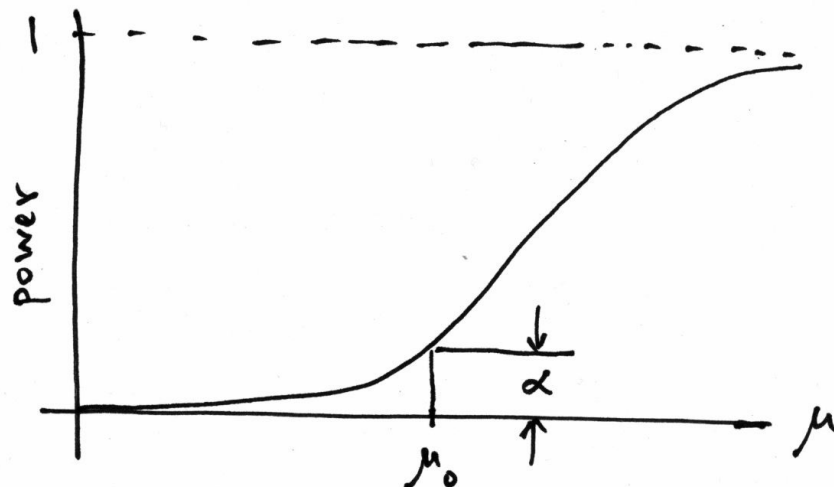
Standard Gaussian  
cumulative distribution



$$\alpha = 1 - \Phi\left(\frac{x_c - \mu_0}{\sigma}\right)$$

$$x_c = \mu_0 + \sigma \Phi^{-1}(1 - \alpha)$$

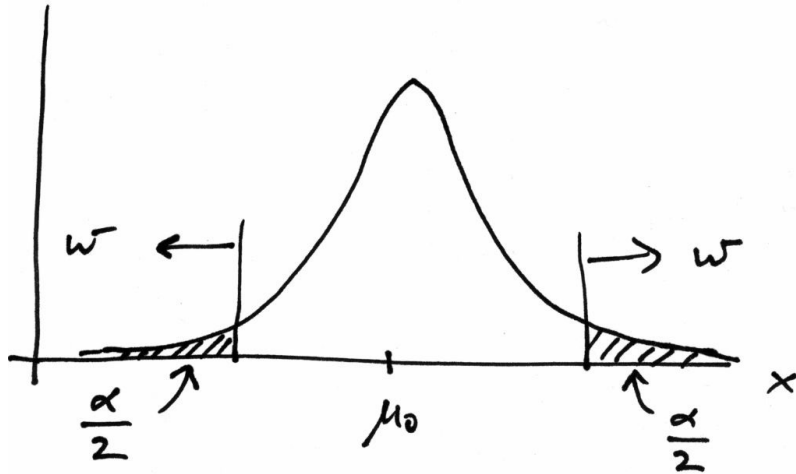
Standard Gaussian quantile



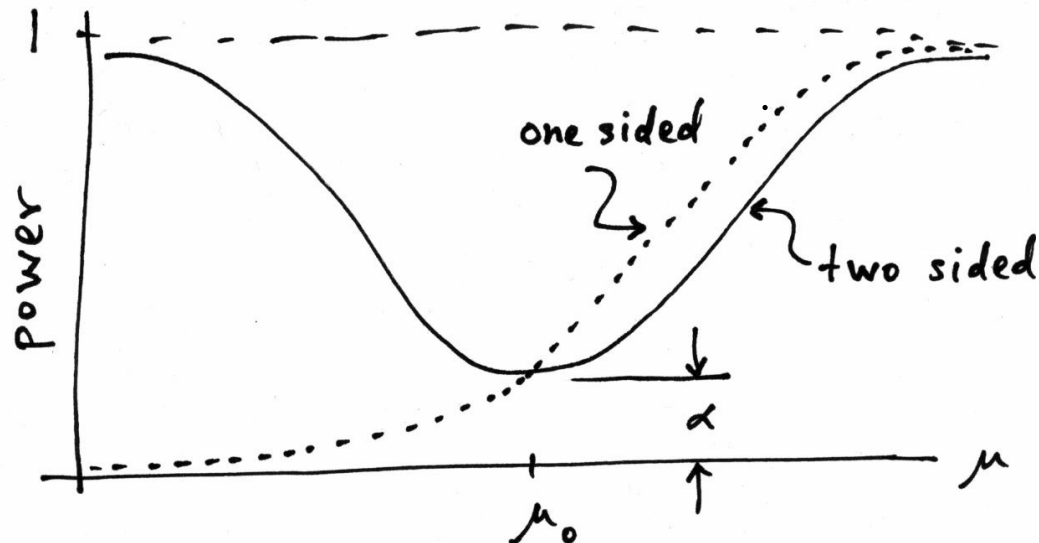
$$\text{power} = 1 - \beta = P(x > x_c | \mu) =$$

$$1 - \Phi\left(\frac{\mu_0 - \mu}{\sigma} + \Phi^{-1}(1 - \alpha)\right)$$

# Choice of critical region based on power (3)



But we might consider  $\mu < \mu_0$  as well as  $\mu > \mu_0$  to be viable alternatives, and choose the critical region to contain both high and low  $x$  (a two-sided test).



New critical region now gives reasonable power for  $\mu < \mu_0$ , but less power for  $\mu > \mu_0$  than the original one-sided test.

# No such thing as a model-independent test

In general we cannot find a single critical region that gives the maximum power for all possible alternatives (no “Uniformly Most Powerful” test).

In HEP we often try to construct a test of

$H_0$  : Standard Model (or “background only”, etc.)

such that we have a well specified “false discovery rate”,

$\alpha$  = Probability to reject  $H_0$  if it is true,

and high power with respect to some interesting alternative,

$H_1$  : SUSY,  $Z'$ , etc.

But there is no such thing as a “model independent” test. Any statistical test will inevitably have high power with respect to some alternatives and less power with respect to others.