Statistical Data Analysis 2023/24
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").

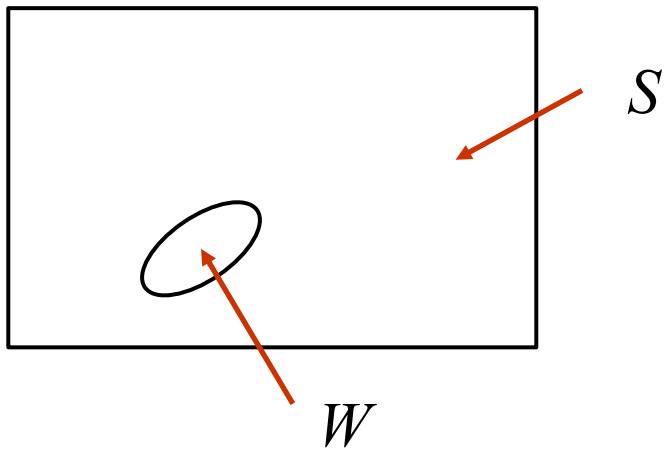
Consider a simple hypothesis H_0 (the "null") and an alternative H_1.

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.,

$$P(x \in W \,|\, H_0) \leq \alpha$$

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.

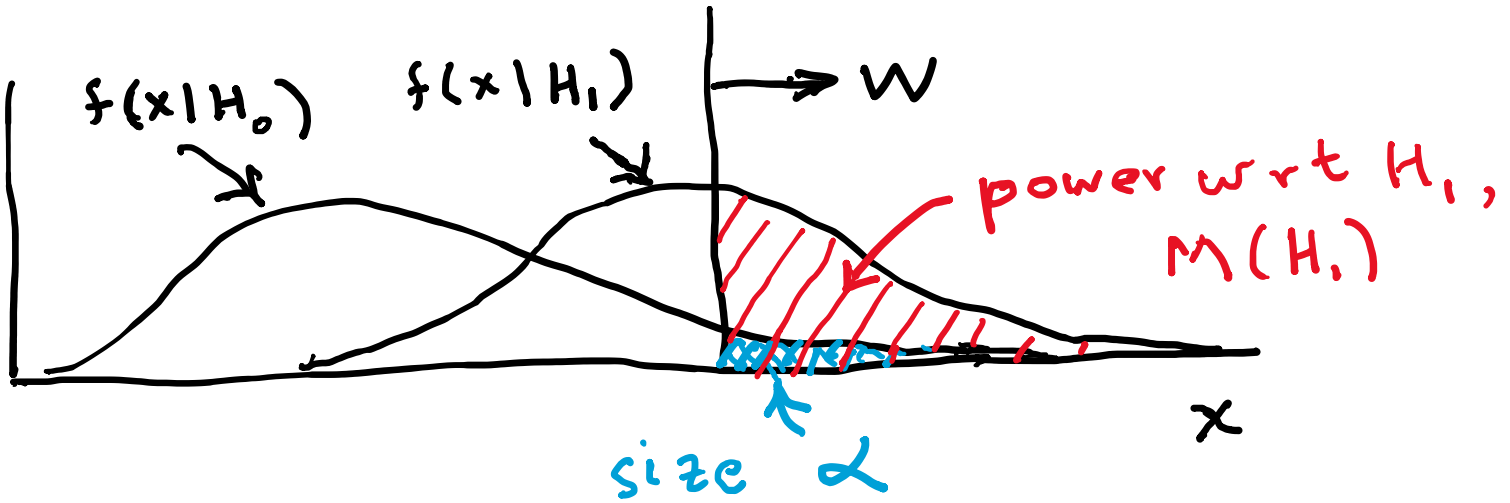


Definition of a test (2)

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:

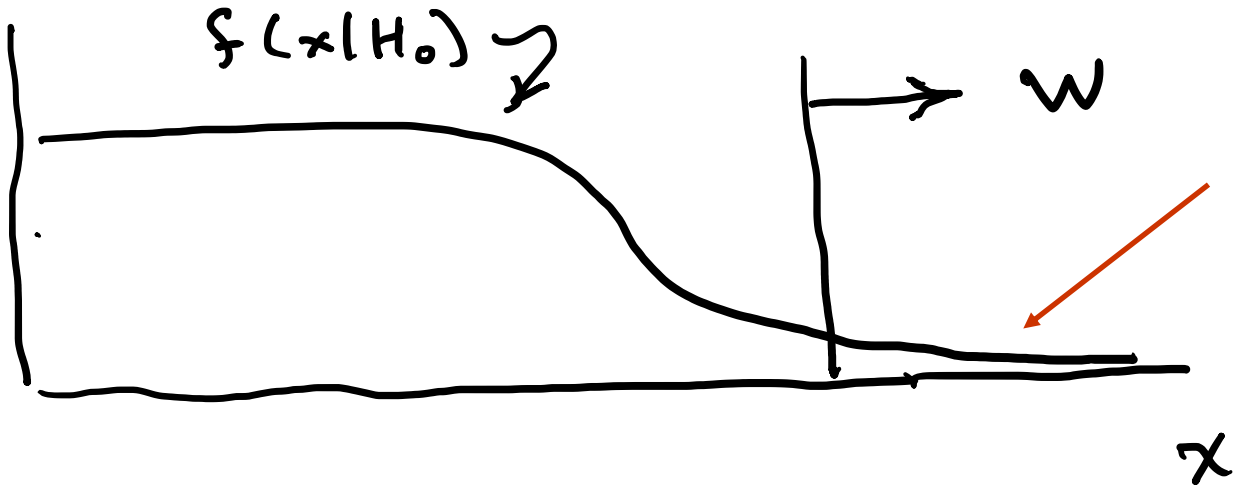


Obvious where to put W?

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.

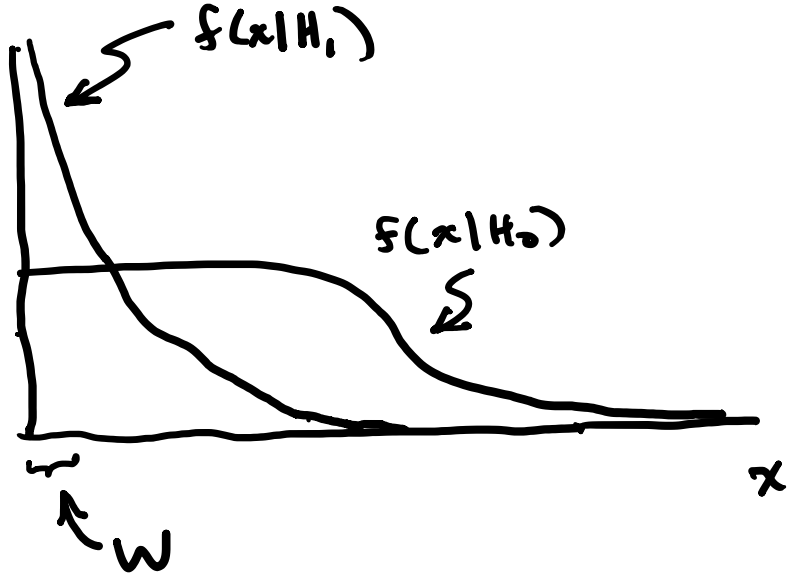
Suppose, e.g., H_0 predicts that x (suppose positive) usually comes out low. High values of x are less characteristic of H_0, so if a high value is observed, we should reject H_0, i.e., we put W at high x:



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:

$$P(x \in W \,|\, H_0) \leq \alpha$$

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

$$P(x \in \mathrm{S} - W \,|\, H_1) = \beta$$

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.

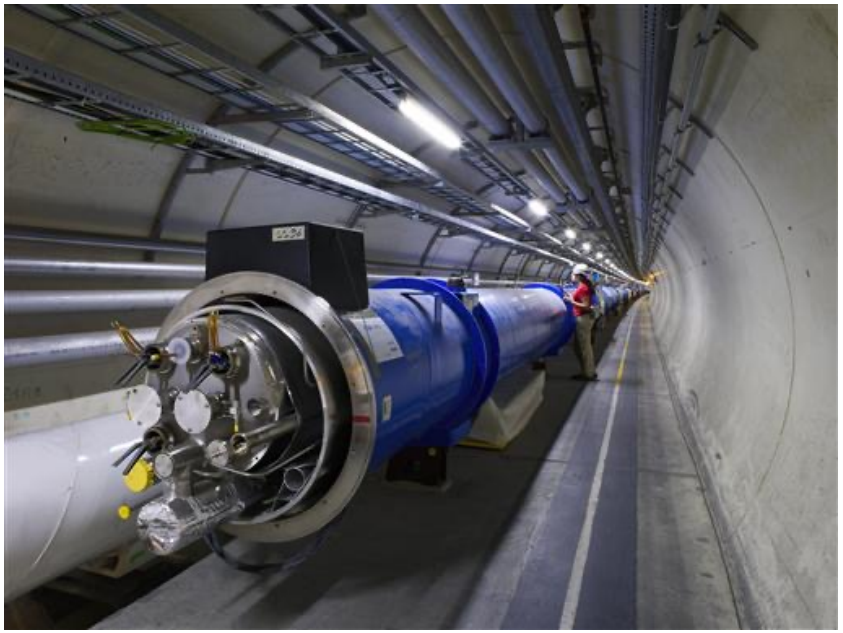
Statistical Data Analysis
Lecture 4-2

- 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:

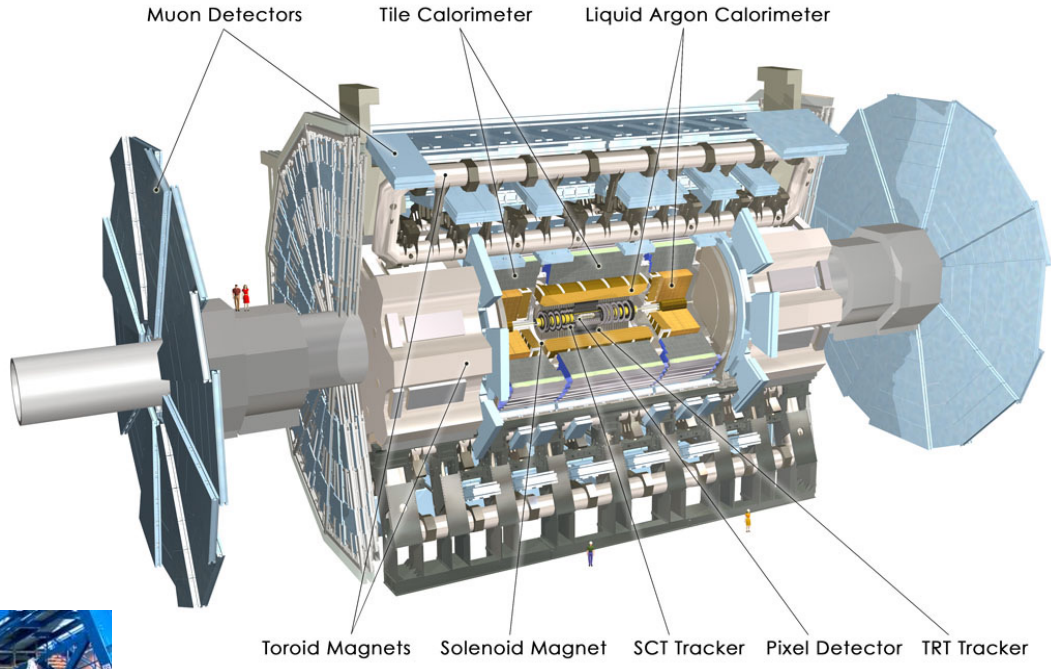
ATLAS
CMS
} general purpose

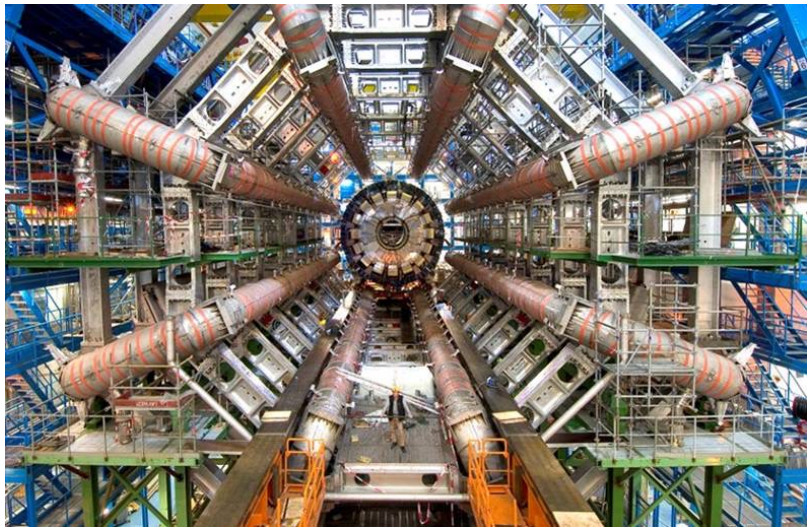
LHCb (b physics)

ALICE (heavy ion physics)

The ATLAS detector

3000 physicists
38 countries
183 universities/labs



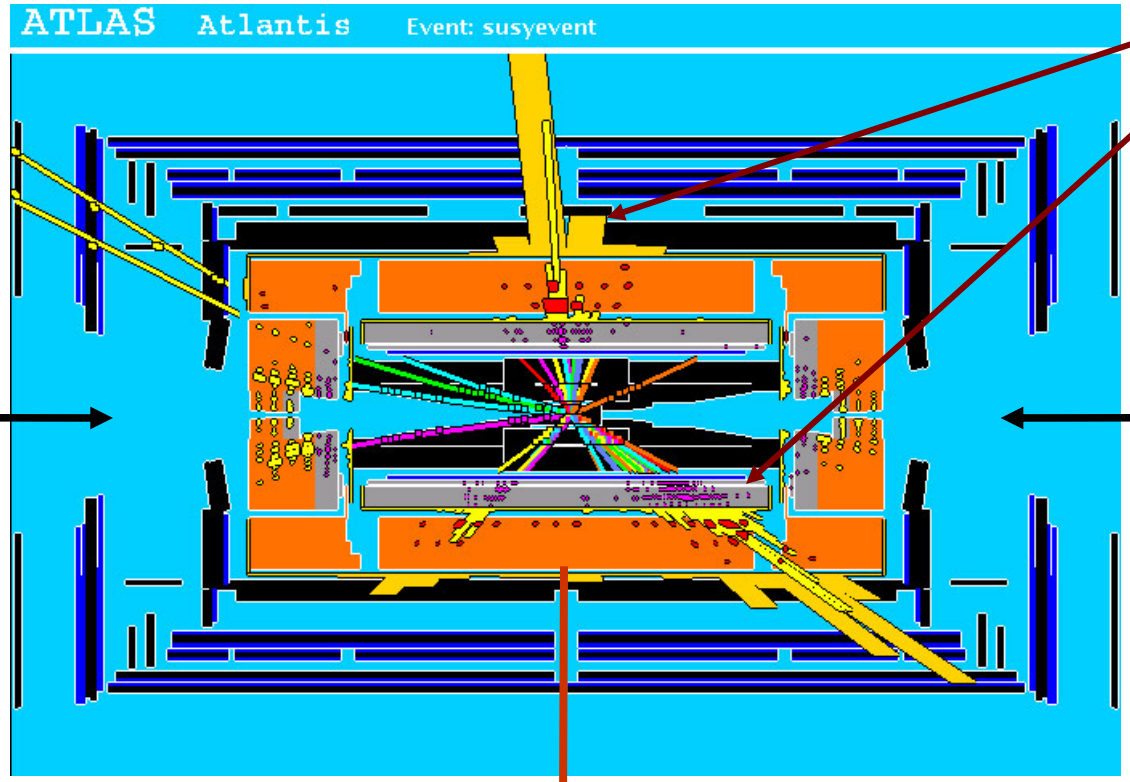


25 m diameter
46 m length
7000 tonnes
~10^8 electronic channels

A simulated SUSY event

high p_T muons

high p_T jets of hadrons

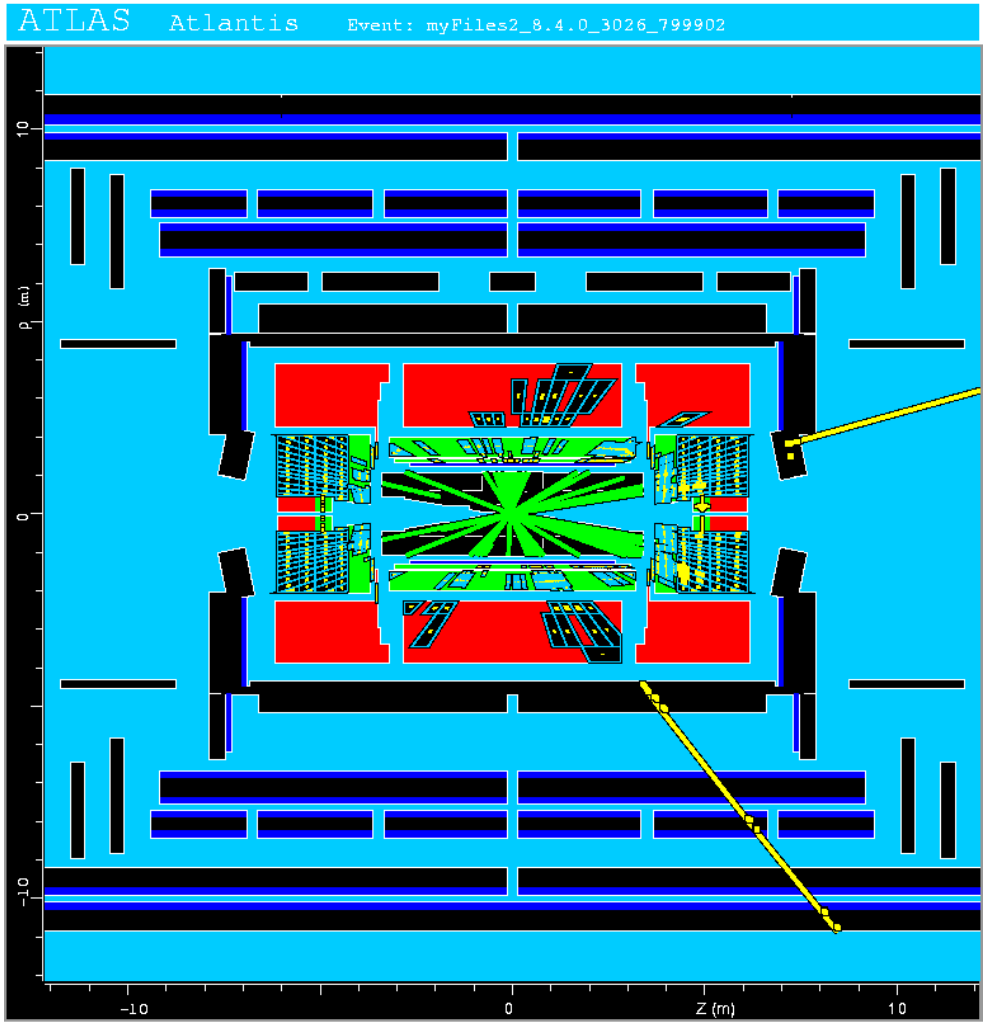


p

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

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

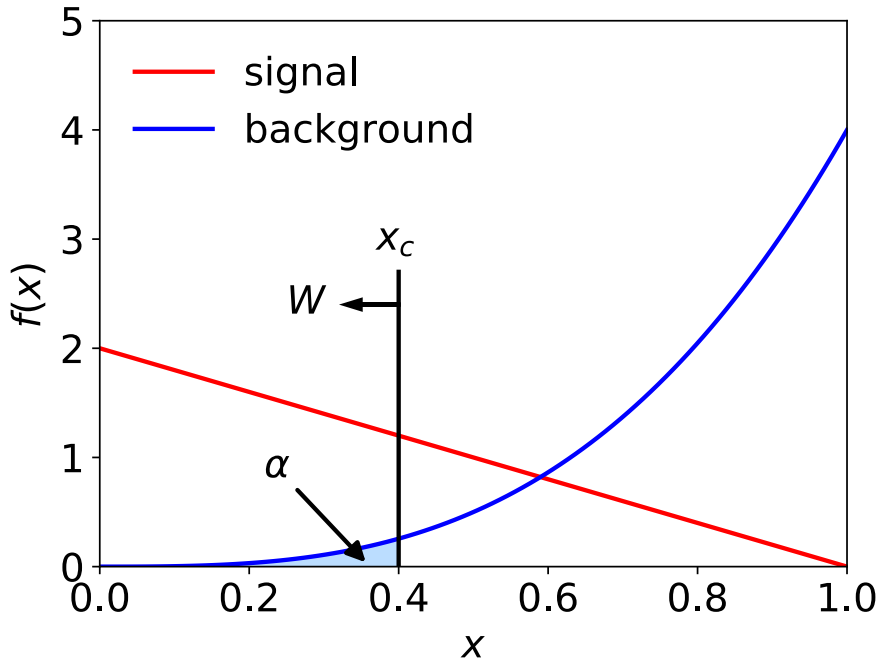
Example of a test for classification

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: $W = \{x : x \le x_c\}$, where x_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:

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

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

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

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:

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

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.)

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:

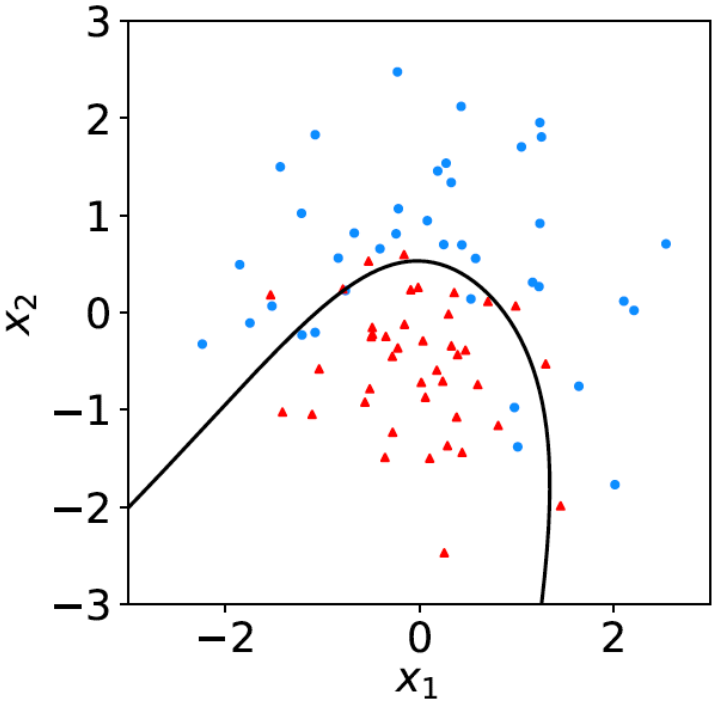
x_1 = length
x_2 = width
x_3 = weight
x_4 = area of fins
x_5 = mean spectral reflectance
x_6 = ...

These constitute the "feature vector" $\boldsymbol{x} = (x_1, \ldots, x_n)$.

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": $(\boldsymbol{x}, y)_1, (\boldsymbol{x}, y)_2, \ldots, (\boldsymbol{x}, y)_N$.

Distributions of the features

If we consider only two features $\boldsymbol{x} = (x_1, x_2)$, 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.

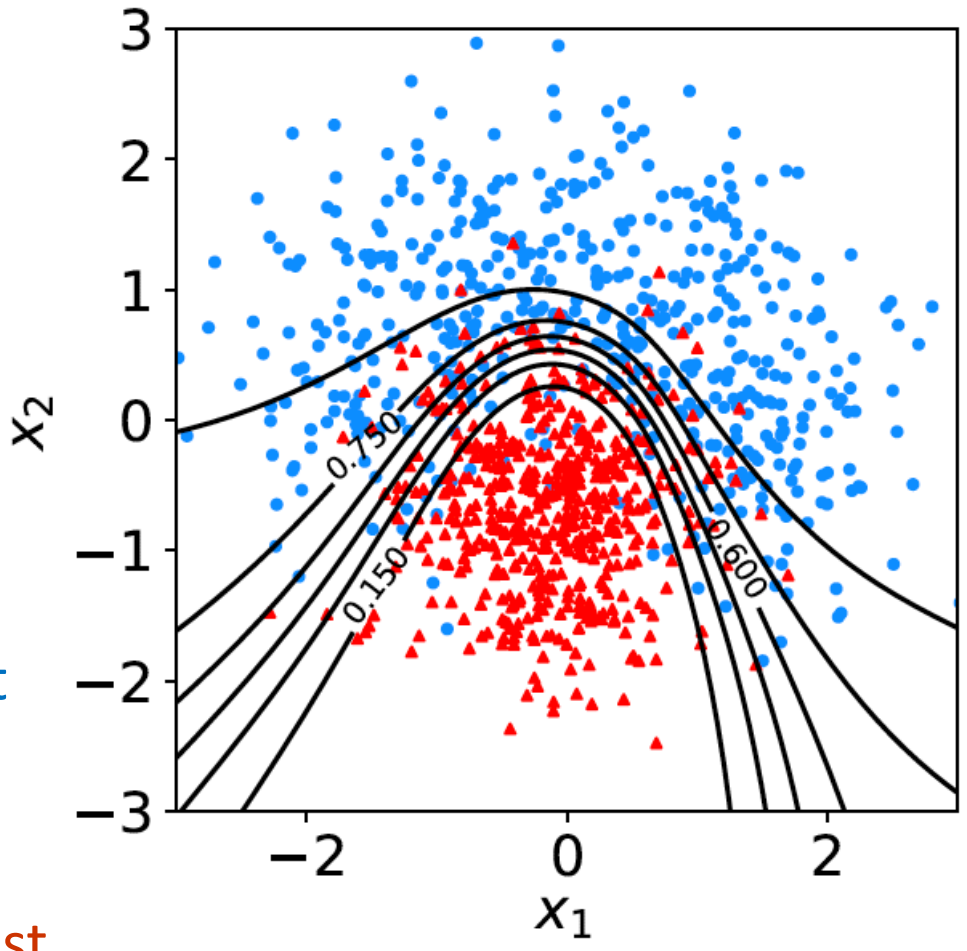
Decision function, test statistic

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

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

scalar function

constant

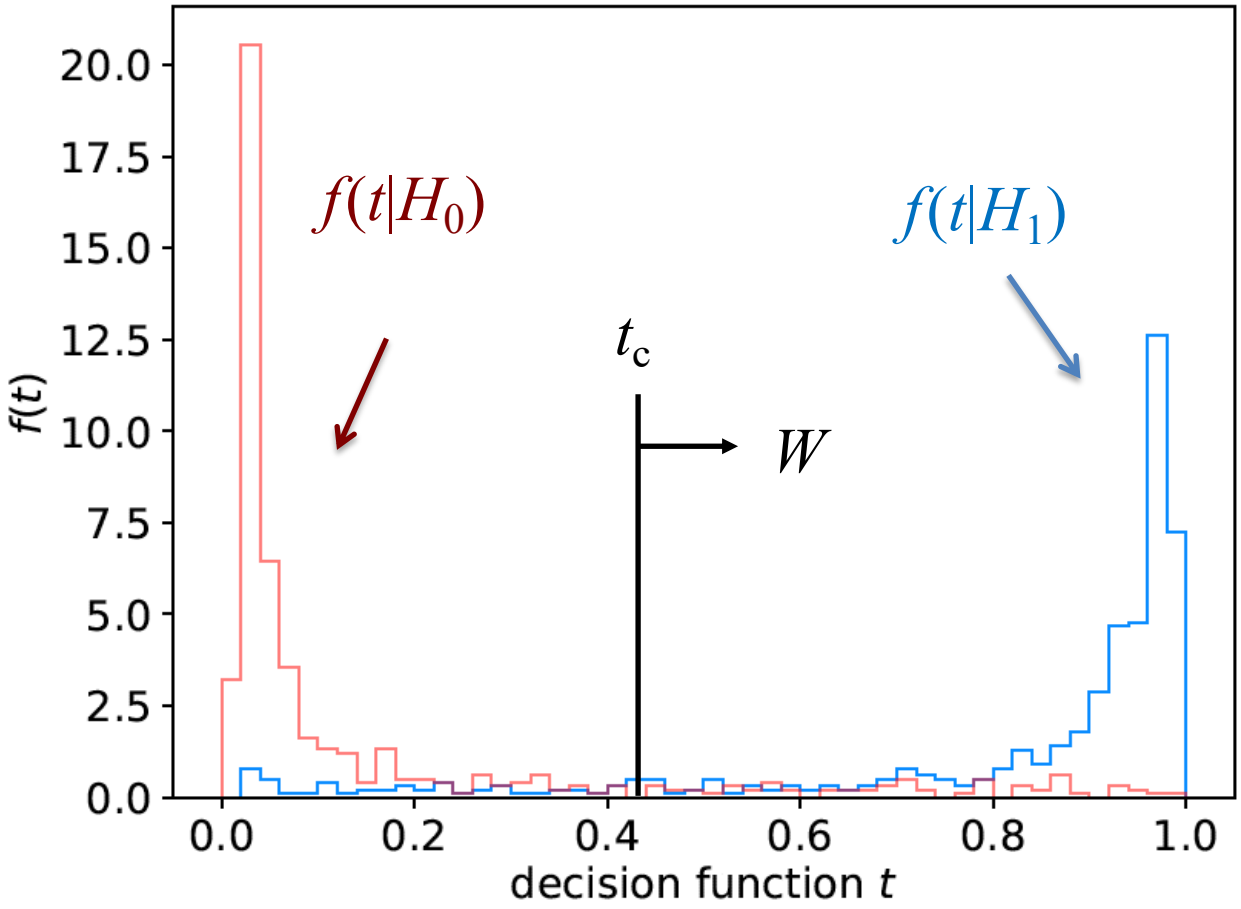


Different values of the constant t_c result in a family of surfaces.

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

Distribution of $t(\boldsymbol{x})$

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

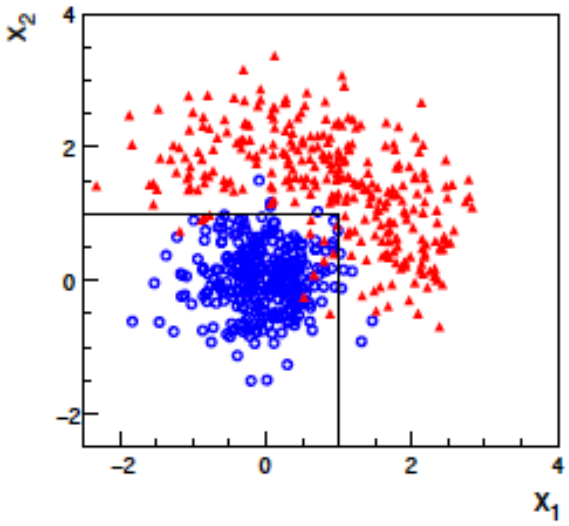


Types of decision boundaries

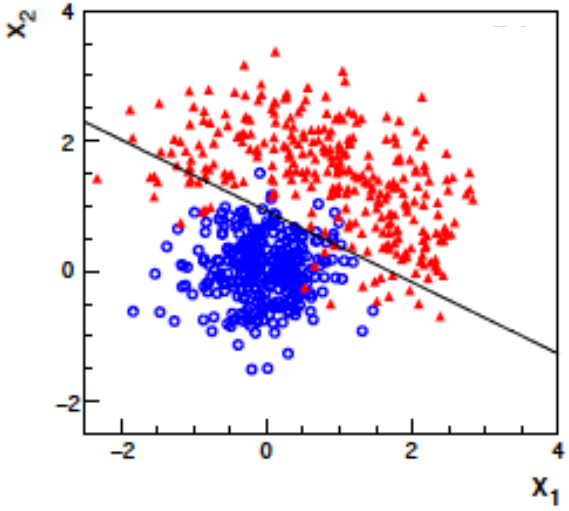
So what is the optimal boundary for the critical region, i.e., what is the optimal test statistic $t(\boldsymbol{x})$?

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

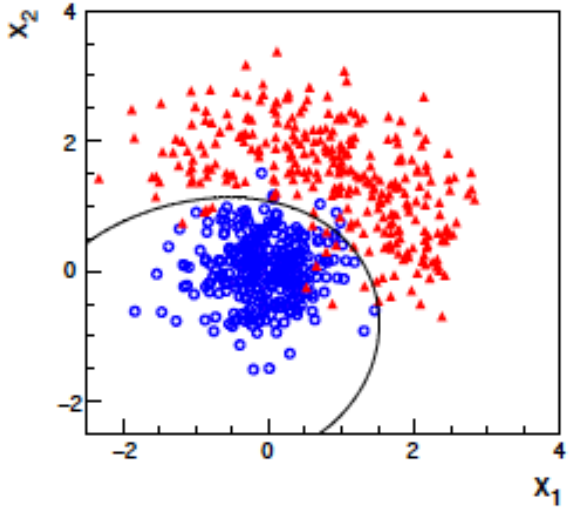
Remember $\boldsymbol{x}$-space can have many dimensions.



"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 α, to get the highest power with respect to the alternative H_1 we need for all $\boldsymbol{x}$ 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_α is a constant chosen to give a test of the desired size.

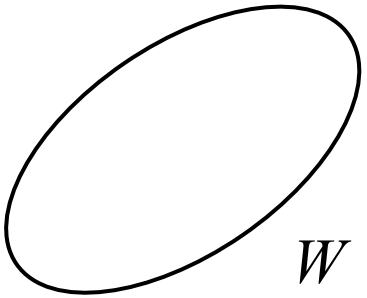
Equivalently, optimal scalar test statistic is $$t(\mathbf{x}) = \frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)}$$

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:

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



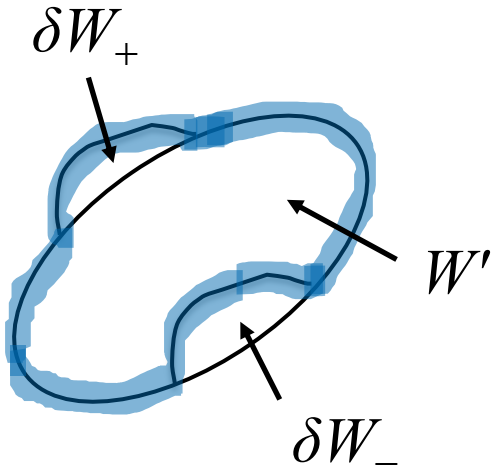
Try to change this into a different critical region W' retaining the same size α, i.e.,

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

To do so add a part δW_+, but to keep the size α, we need to remove a part δW_-, i.e.,

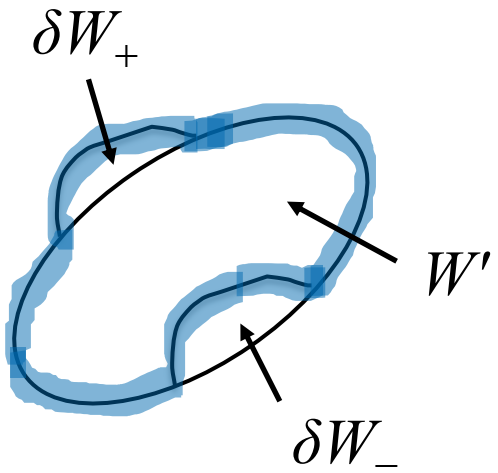
$$W \to W' = W + \delta W_+ - \delta W_-$$

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



Proof of Neyman-Pearson Lemma (2)

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



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

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

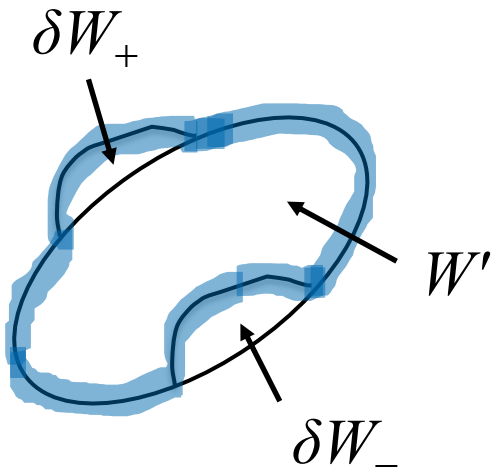
The right-hand sides are equal and therefore

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

Proof of Neyman-Pearson Lemma (3)

We have

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



Note W and δW_+ are disjoint, and W' and δW_- are disjoint, so by Kolmogorov's 3rd axiom,

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

Therefore

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

≥ 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 $f(\boldsymbol{x}|\mathrm{s}), f(\boldsymbol{x}|\mathrm{b})$, so for a given $\boldsymbol{x}$ 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 $\boldsymbol{x} \sim f(\boldsymbol{x}|\mathrm{s}) \quad \rightarrow \quad \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$

generate $\boldsymbol{x} \sim f(\boldsymbol{x}|\mathrm{b}) \quad \rightarrow \quad \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$

This gives samples of "training data" with events of known type.

- Can be expensive (1 fully simulated LHC event ~ 1 CPU minute).

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

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

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

$$\boldsymbol{z}_2 \sim f(\boldsymbol{z}_2|\boldsymbol{z}_1)$$
$$\boldsymbol{z}_3 \sim f(\boldsymbol{z}_3|\boldsymbol{z}_2)$$
$$\vdots$$
$$\boldsymbol{x} \sim f(\boldsymbol{x}|\boldsymbol{z}_n)$$

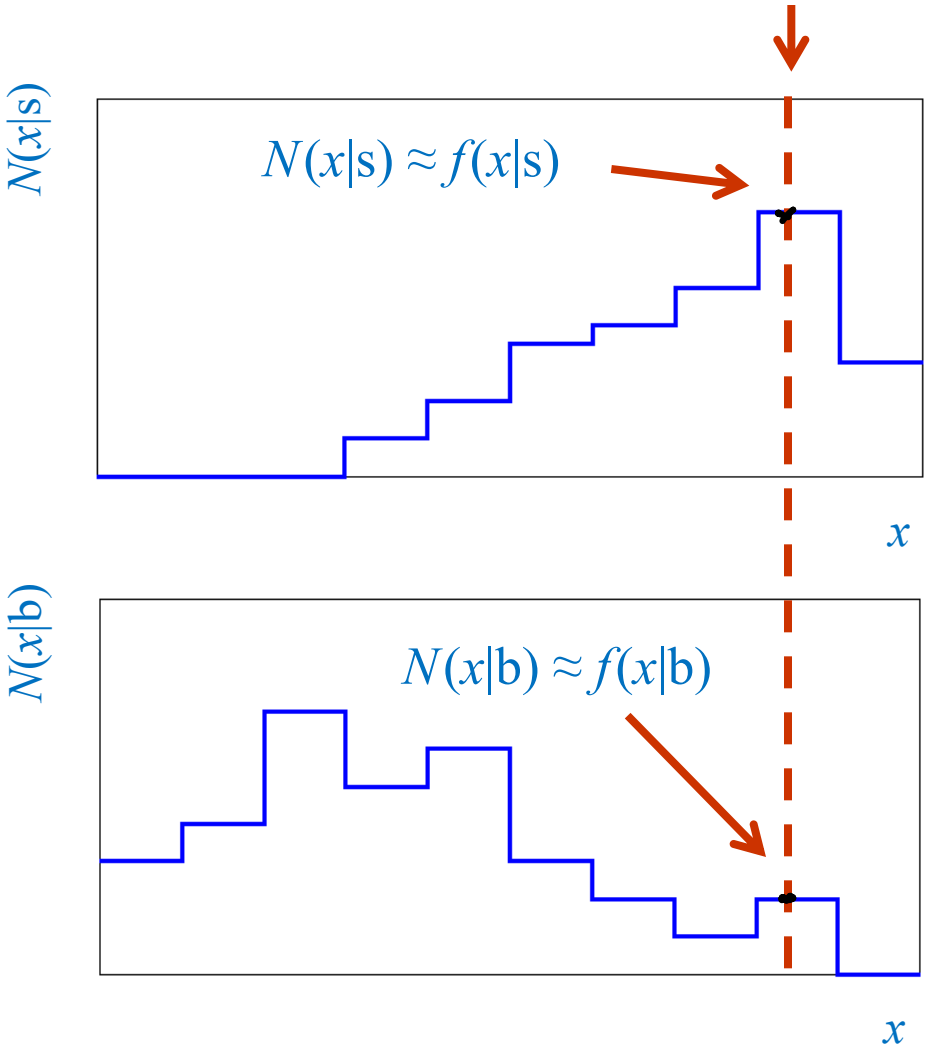
See, e.g., Kyle Cranmer, Johann Brehmer, Gilles Louppe, *The frontier of simulation-based inference*, arXiv:1911.01429 [stat.ML], PNAS doi.org/10.1073/pnas.1912789117

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

$$f(\mathbf{x}|H) = \int \cdots \int d\mathbf{z}_1 \cdots d\mathbf{z}_n f(\mathbf{x}|\mathbf{z}_n)\, f(\mathbf{z}_n|\mathbf{z}_{n-1}) \cdots f(\mathbf{z}_2|\mathbf{z}_1)\, f(\mathbf{z}_1|H)$$

Approximate LR from histograms

Want $t(x) = f(x|\mathrm{s})/f(x|\mathrm{b})$ 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:

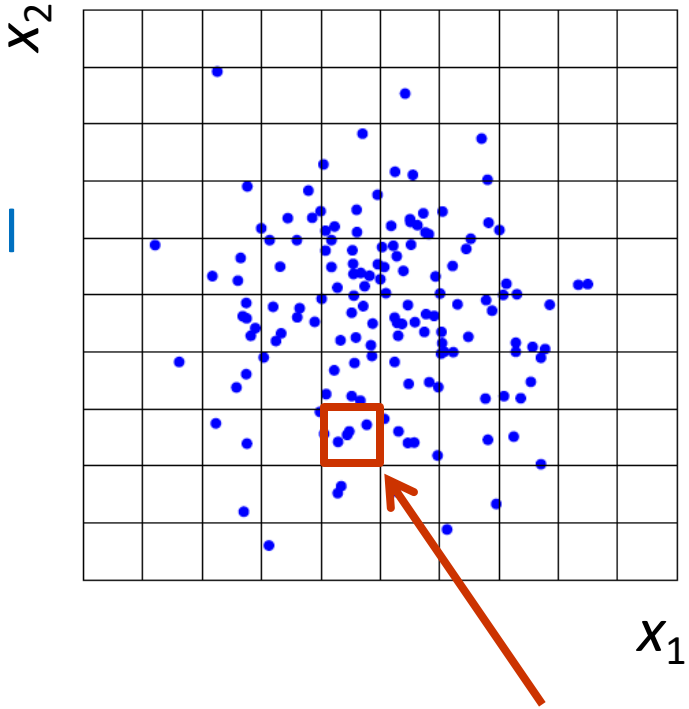
$$t(x) \approx \frac{N(x|s)}{N(x|b)}$$

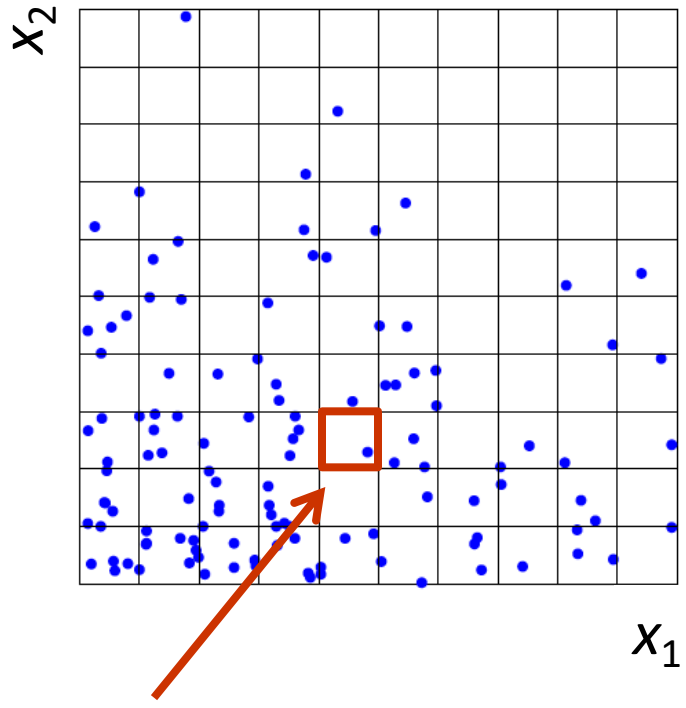
Can work well for single variable.

Approximate LR from 2D-histograms

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

signal





back-ground

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

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 $t(\boldsymbol{x})$ with fewer parameters; determine them (using MC) to give best separation between signal and background.

Alternatively, try to estimate the probability densities $f(\boldsymbol{x}|\mathrm{s})$ and $f(\boldsymbol{x}|\mathrm{b})$ (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, 2nd ed., Springer, 2009

Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, *An Introduction to Statistical Learning*, Springer, 2017, **https://www.statlearning.com/**

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

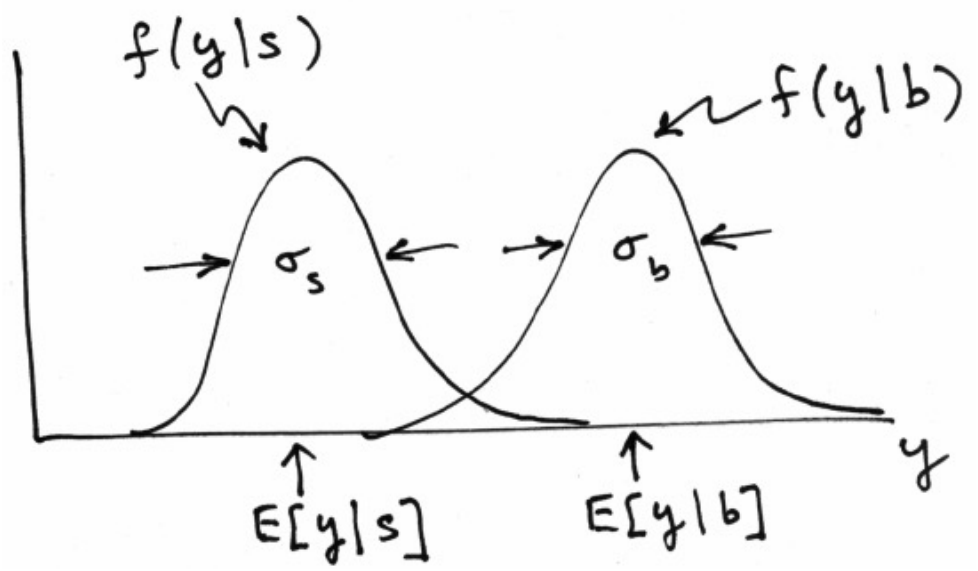
Good manual, widely used in HEP

Linear test statistic

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

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

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



Linear test statistic

Fisher: to get large difference between means and small widths for $f(y|\mathrm{s})$ and $f(y|\mathrm{b})$, 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:

$$\mathbf{w} \propto W^{-1}(\boldsymbol{\mu}_\mathrm{b} - \boldsymbol{\mu}_\mathrm{s})$$

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

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

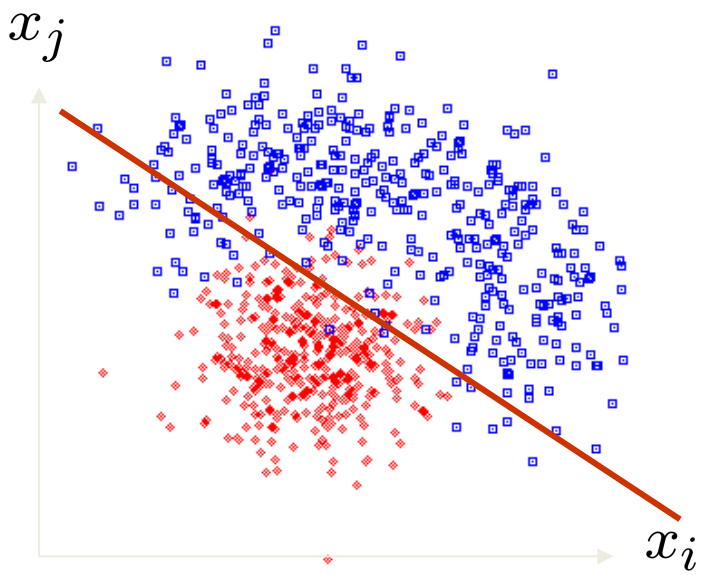
The Fisher discriminant

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 $y(\boldsymbol{x})$, here linear (hyperplanes):



Fisher discriminant for Gaussian data

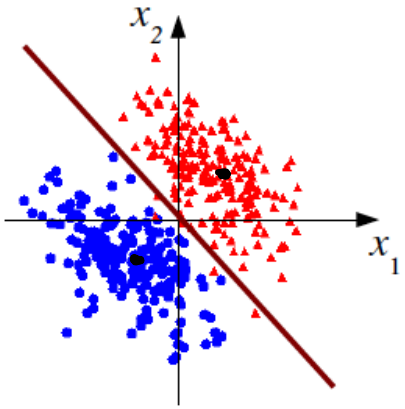
Suppose the pdfs of the input variables, $f(\boldsymbol{x}|\mathrm{s})$ and $f(\boldsymbol{x}|\mathrm{b})$, are both multivariate Gaussians with same covariance but different means:

$$f(\boldsymbol{x}|\mathrm{s}) = \mathrm{Gauss}(\boldsymbol{\mu}_\mathrm{s}, V)$$

$$f(\boldsymbol{x}|\mathrm{b}) = \mathrm{Gauss}(\boldsymbol{\mu}_\mathrm{b}, V)$$

Same covariance

$$V_{ij} = \mathrm{cov}[x_i, x_j]$$



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

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

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.

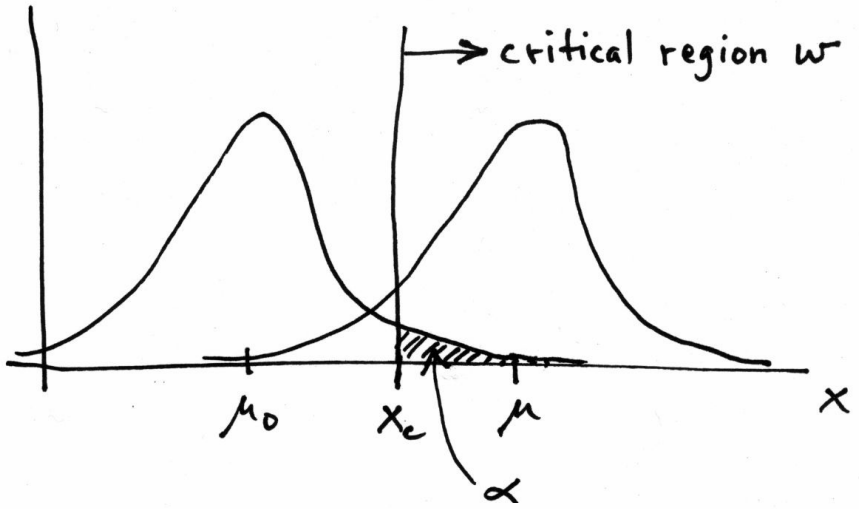
Often such a test has a high power not only with respect to a specific point alternative but for a class of alternatives.
E.g., using a measurement $x \sim \text{Gauss}\,(\mu, \sigma)$ we may test

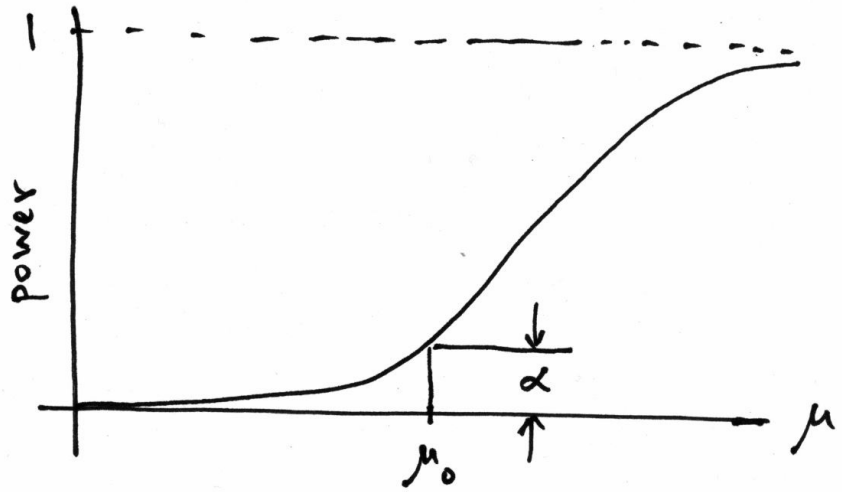
$H_0 : \mu = \mu_0$ versus the composite alternative $H_1 : \mu > \mu_0$

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

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

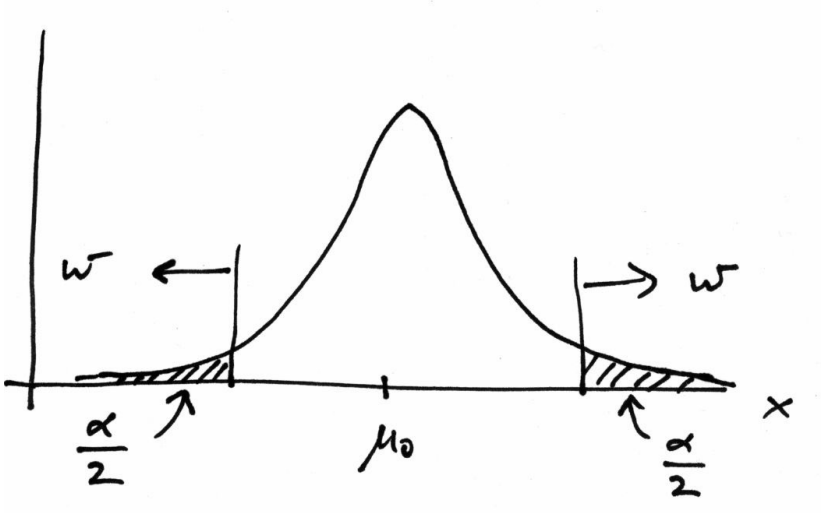
$$x_c = \mu_0 + \sigma\Phi^{-1}(1-\alpha)$$

Standard Gaussian quantile

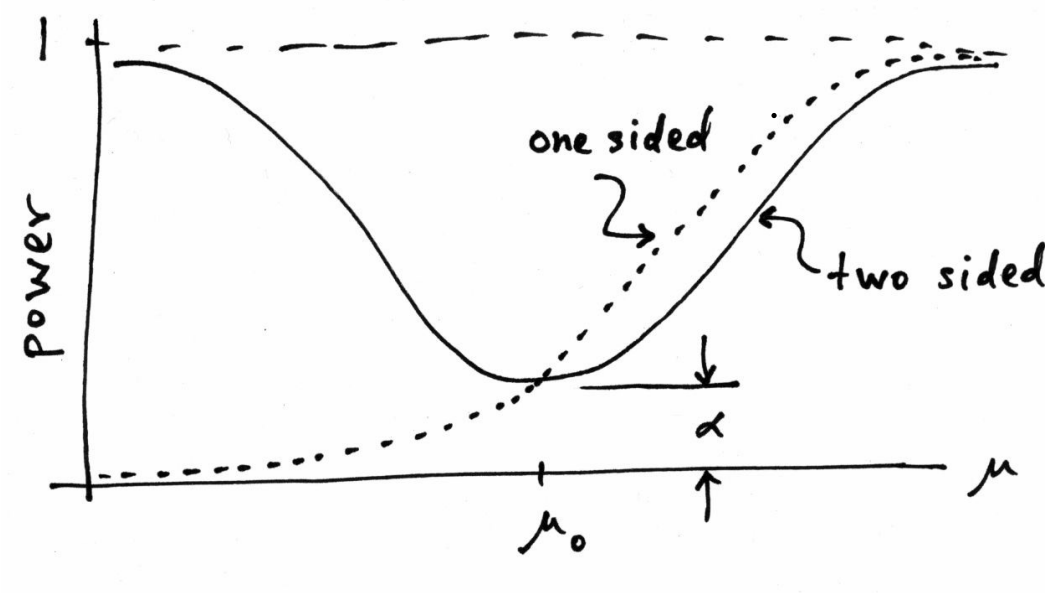
$$\text{power} = 1 - \beta = P(x > x_c|\mu) =$$

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

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",

α = 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.