

Statistics for Particle Physicists

Lecture 1



<https://pages.lip.pt/data-science/school/>



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

Outline

→ Lecture 1: Introduction

Probability

Hypothesis tests

From hypothesis tests to machine learning

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

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.

R.L. Workman et al. (Particle Data Group), *Prog. Theor. Exp. Phys.* 083C01 (2022); pdg.lsb1.gov sections on probability, statistics, MC.

Theory ↔ Statistics ↔ Experiment

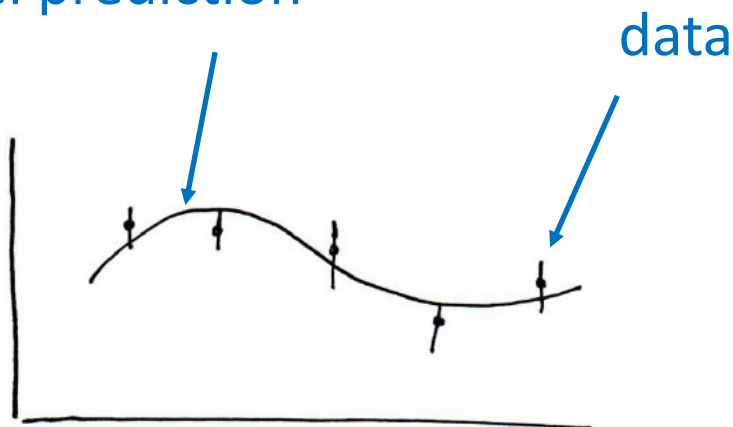
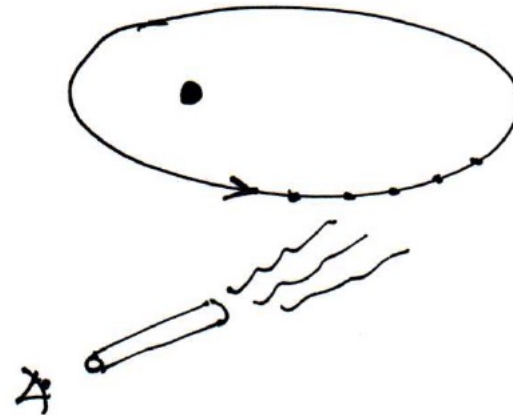
Theory (model, hypothesis):

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

+ response of measurement apparatus

= model prediction

Experiment (observation):



Uncertainty enters on many levels

→ quantify with **probability**

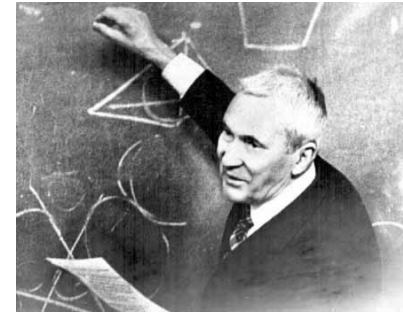
A definition of probability

Consider a set S with subsets A, B, \dots

For all $A \subset S, P(A) \geq 0$

$$P(S) = 1$$

If $A \cap B = \emptyset, P(A \cup B) = P(A) + P(B)$



Kolmogorov
axioms (1933)

Also define conditional
probability of A given B :

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

Subsets A, B independent if: $P(A \cap B) = P(A)P(B)$

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

Interpretation of Probability

- I. Relative frequency (\rightarrow “frequentist statistics”)
 A, B, \dots are outcomes of a repeatable experiment

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

- II. Subjective probability (\rightarrow “Bayesian statistics”)
 A, B, \dots are hypotheses (statements that are true or false)

$$P(A) = \text{degree of belief that } A \text{ is true}$$

- Both interpretations consistent with Kolmogorov axioms.
- In particle physics frequency interpretation often most useful, but subjective probability can provide more natural treatment of non-repeatable phenomena: systematic uncertainties, probability that magnetic monopoles exist,...

Bayes' theorem

From the definition of conditional probability we have

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

but $P(A \cap B) = P(B \cap A)$, so

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

Bayes' theorem



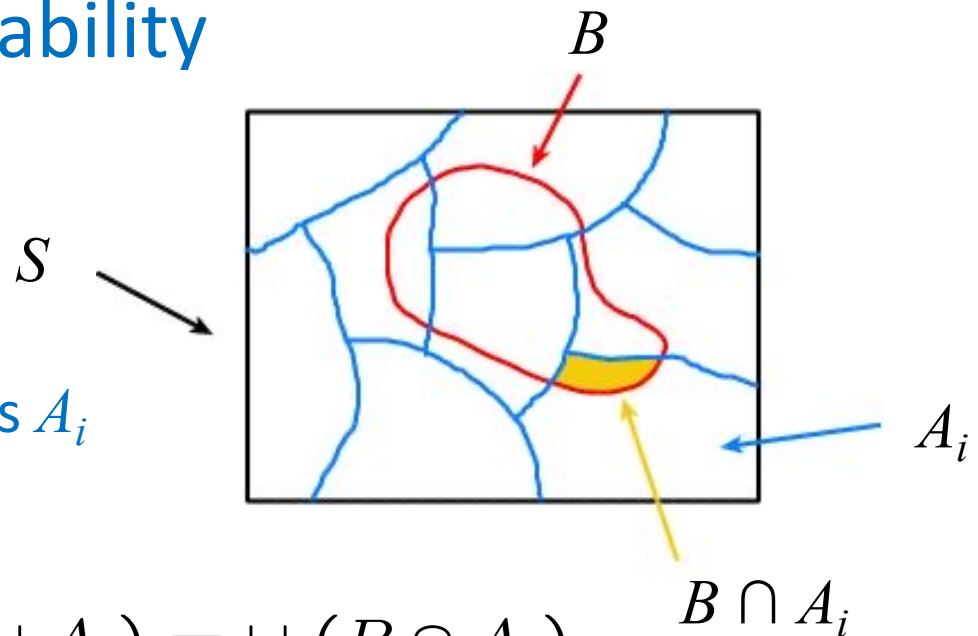
First published (posthumously) by the Reverend Thomas Bayes (1702–1761)

An essay towards solving a problem in the doctrine of chances, Philos. Trans. R. Soc. **53** (1763) 370

The law of total probability

Consider a subset B of the sample space S ,

divided into disjoint subsets A_i such that $\cup_i A_i = S$,



$$\rightarrow B = B \cap S = B \cap (\cup_i A_i) = \cup_i (B \cap A_i),$$

$$\rightarrow P(B) = P(\cup_i (B \cap A_i)) = \sum_i P(B \cap A_i)$$

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

Bayes' theorem becomes

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

An example using Bayes' theorem

Suppose the probability (for anyone) to have a disease D is:

$$P(D) = 0.001$$
$$P(\text{no D}) = 0.999$$

← prior probabilities, i.e.,
before any test carried out

Consider a test for the disease: result is + or –

$$P(+|D) = 0.98$$
$$P(-|D) = 0.02$$

← probabilities to (in)correctly
identify a person with the disease

$$P(+|\text{no D}) = 0.03$$
$$P(-|\text{no D}) = 0.97$$

← probabilities to (in)correctly
identify a healthy person

Suppose your result is +. How worried should you be?

Bayes' theorem example (cont.)

The probability to have the disease given a + result is

$$\begin{aligned} p(\mathbf{D}|+) &= \frac{P(+|\mathbf{D})P(\mathbf{D})}{P(+|\mathbf{D})P(\mathbf{D}) + P(+|\text{no } \mathbf{D})P(\text{no } \mathbf{D})} \\ &= \frac{0.98 \times 0.001}{0.98 \times 0.001 + 0.03 \times 0.999} \\ &= 0.032 \quad \leftarrow \text{posterior probability} \end{aligned}$$

i.e. you're probably OK!

Your viewpoint: my degree of belief that I have the disease is 3.2%.

Your doctor's viewpoint: 3.2% of people like this have the disease.

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 (Biden wins in 2024),

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!)

Random variables and probability density functions

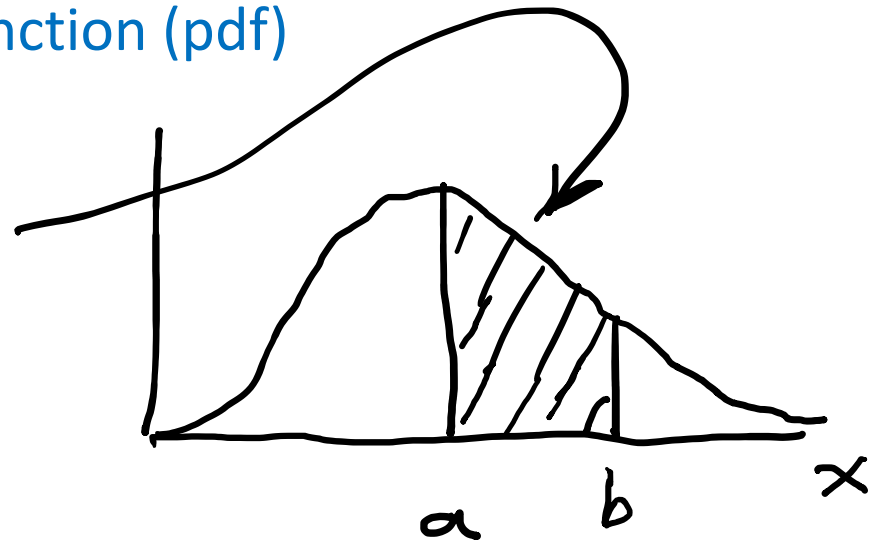
A random variable is a numerical characteristic assigned to an element of the sample space; can be discrete or continuous.

Suppose outcome of experiment is continuous value x

$$P(x \text{ found in } [x, x + dx]) = f(x) dx$$

→ $f(x)$ = probability density function (pdf)

$$\int_a^b f(x) dx = P(a \leq x \leq b)$$



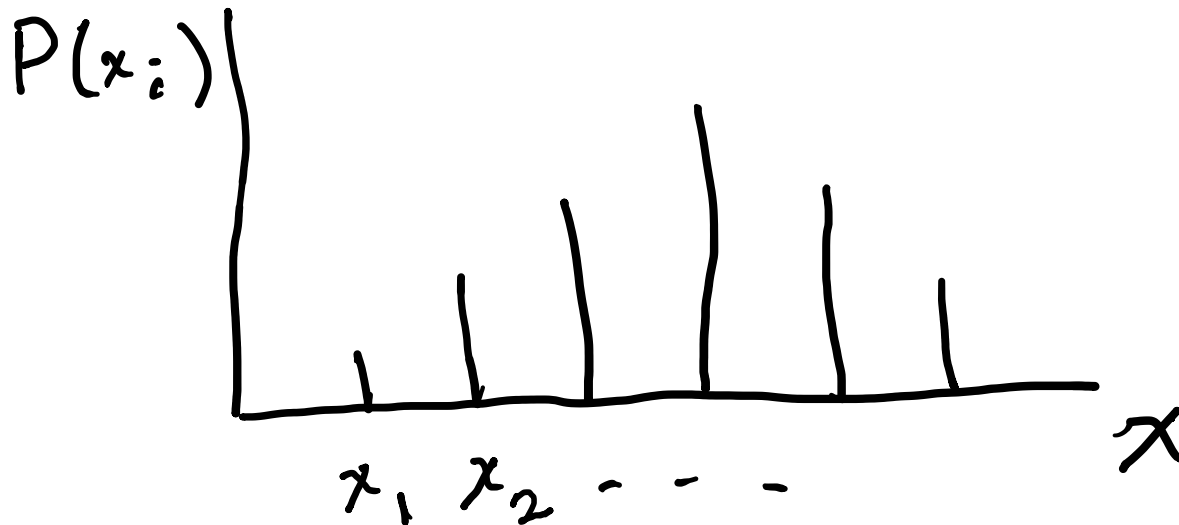
$$\int_{-\infty}^{\infty} f(x) dx = 1 \quad x \text{ must be somewhere}$$

Probability mass function

For discrete outcome x_i with e.g. $i = 1, 2, \dots$ we have

$$P(x_i) = p_i \quad \text{probability (mass) function}$$

$$\sum_i P(x_i) = 1 \quad x \text{ must take on one of its possible values}$$



Some related functions

Cumulative distribution $F(x) = \int_{-\infty}^x f(x') dx'$

Joint pdf, e.g., $f(x,y)$

$$P(x \in [x, x + dx] \text{ and } y \in [y, y + dy]) = f(x, y) dx dy$$

Marginal pdfs $f_x(x) = \int f(x, y) dy$ $f_y(y) = \int f(x, y) dx$

Conditional pdfs $f(x|y) = \frac{f(x, y)}{f_y(y)}$ $f(y|x) = \frac{f(x, y)}{f_x(x)}$

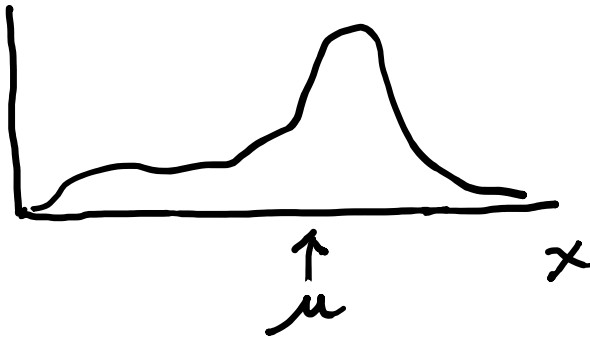
Bayes' theorem: $f(x|y) = \frac{f(y|x)f_x(x)}{f_y(y)}$

Expectation values

Consider continuous r.v. x with pdf $f(x)$.

Define expectation (mean) value as $E[x] = \int x f(x) dx$

Notation (often): $E[x] = \mu \sim$ “centre of gravity” of pdf.



For discrete r.v.s, replace integral by sum: $E[x] = \sum_{x_i \in S} x_i P(x_i)$

For a function $y(x)$ with pdf $g(y)$,

$$E[y] = \int y g(y) dy = \int y(x) f(x) dx \quad (\text{equivalent})$$

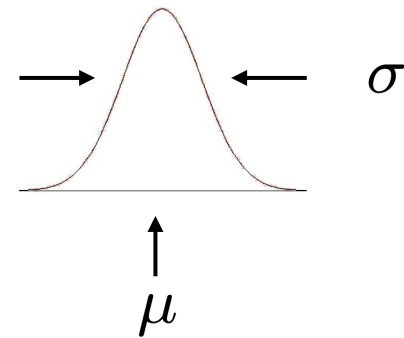
Variance, standard deviation

Variance: $V[x] = E[x^2] - \mu^2 = E[(x - \mu)^2]$

Notation: $V[x] = \sigma^2$

Standard deviation: $\sigma = \sqrt{\sigma^2}$

$\sigma \sim$ width of pdf, same units as x .



Relation between σ and other measures of width, e.g., Full Width at Half Max (FWHM) depend on the pdf, e.g., FWHM = 2.35σ for Gaussian.

Covariance and correlation

Define covariance $\text{cov}[x,y]$ (also use matrix notation V_{xy}) as

$$\text{COV}[x, y] = E[xy] - \mu_x\mu_y = E[(x - \mu_x)(y - \mu_y)]$$

Correlation coefficient (dimensionless) defined as

$$\rho_{xy} = \frac{\text{COV}[x, y]}{\sigma_x\sigma_y} \quad \text{Can show } -1 \leq \rho \leq 1.$$

If x, y , independent, i.e., $f(x, y) = f_x(x)f_y(y)$

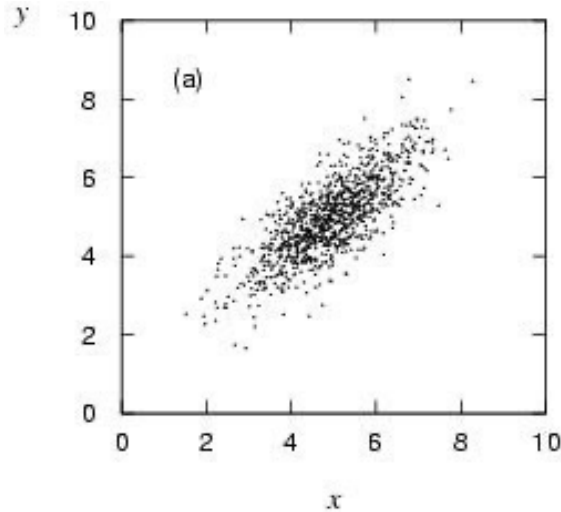
$$E[xy] = \int \int xy f(x, y) dx dy = \mu_x\mu_y$$

$$\rightarrow \text{COV}[x, y] = 0$$

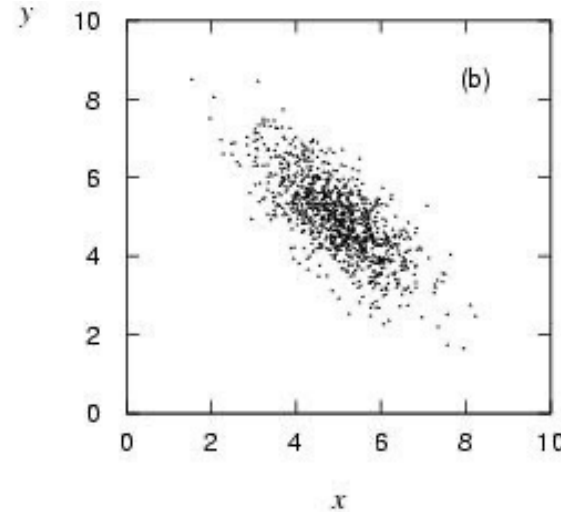
N.B. converse not always true.

Correlation (cont.)

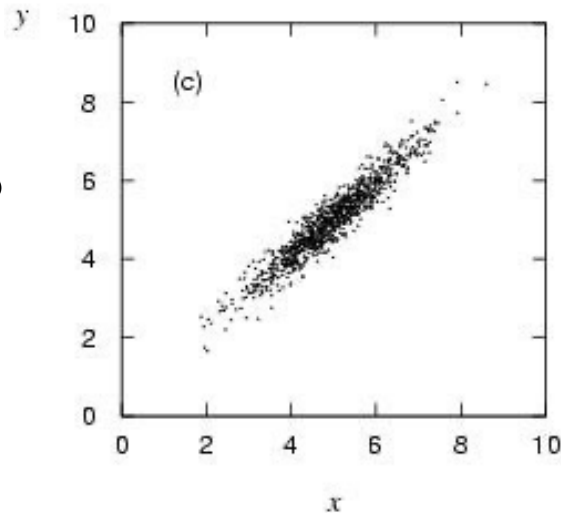
$$\rho = 0.75$$



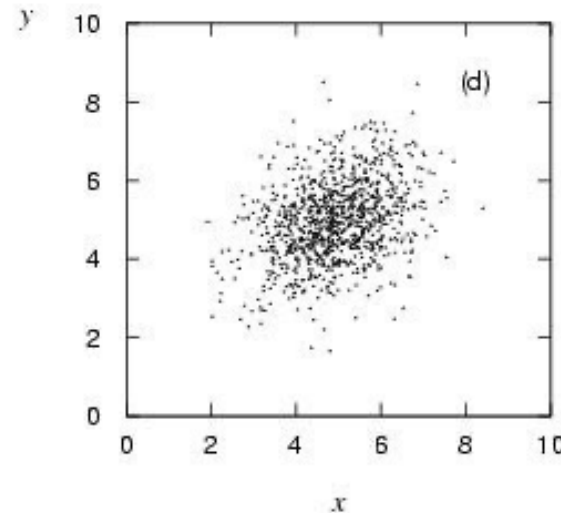
$$\rho = -0.75$$



$$\rho = 0.95$$



$$\rho = 0.25$$



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

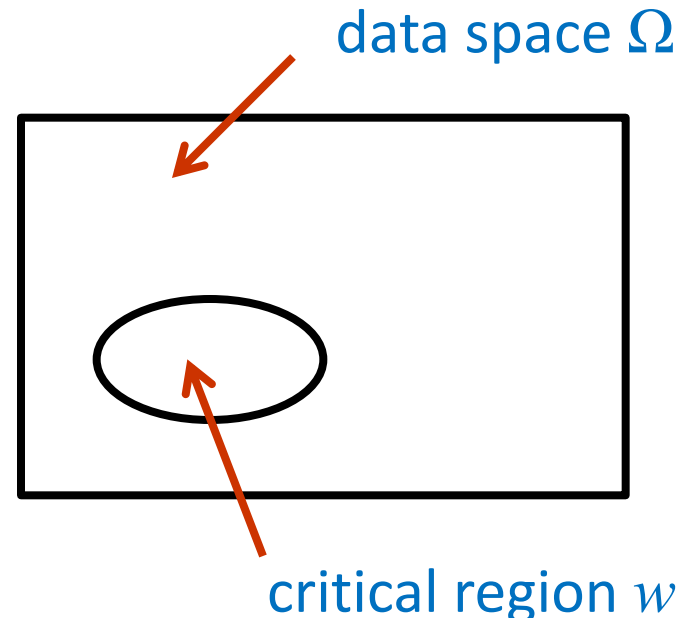
A test of H_0 is defined by specifying a critical region w of the data space such that there is no more than some (small) probability α , assuming H_0 is correct, to observe the data there, i.e.,

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

Need inequality if data are discrete.

α is called the size or significance level of the test.

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

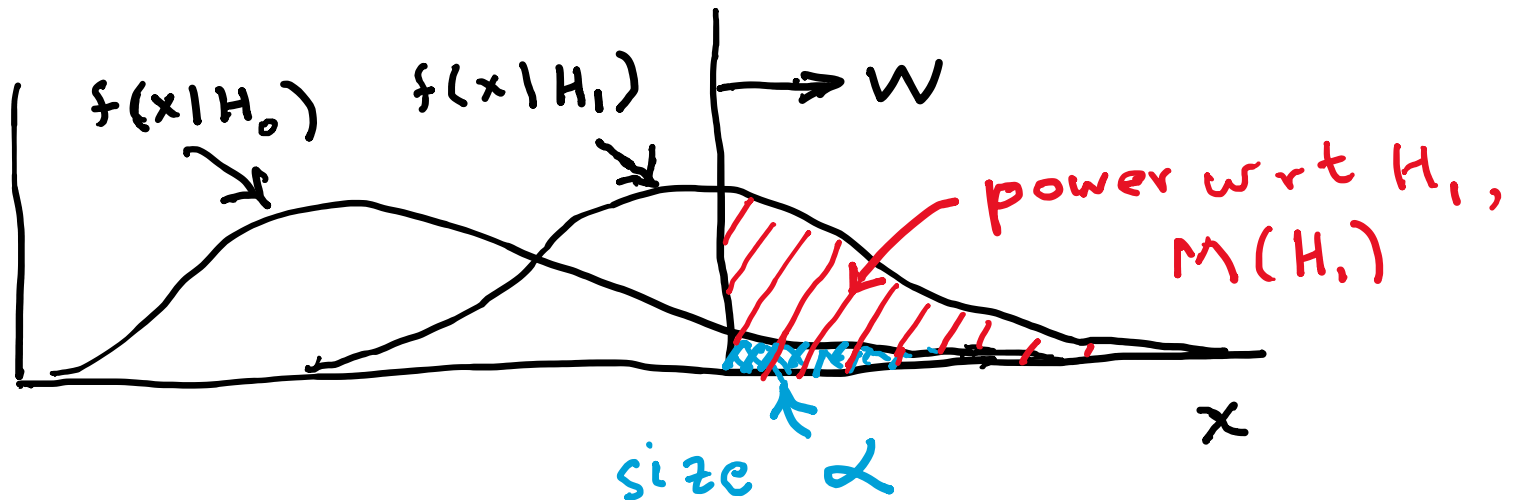


Definition of a test (2)

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



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

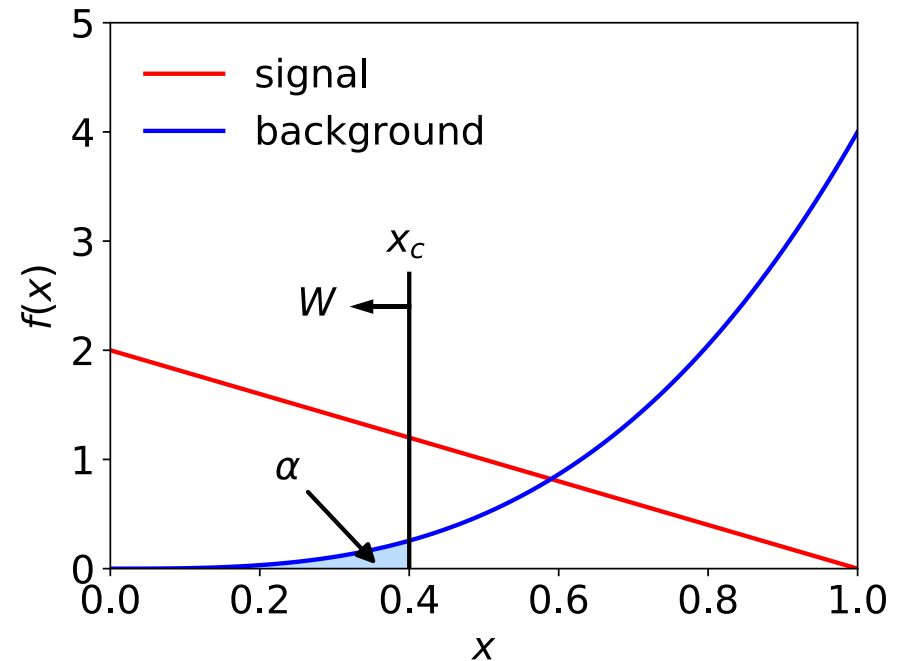
Example of a test for classification

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 α .

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:

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

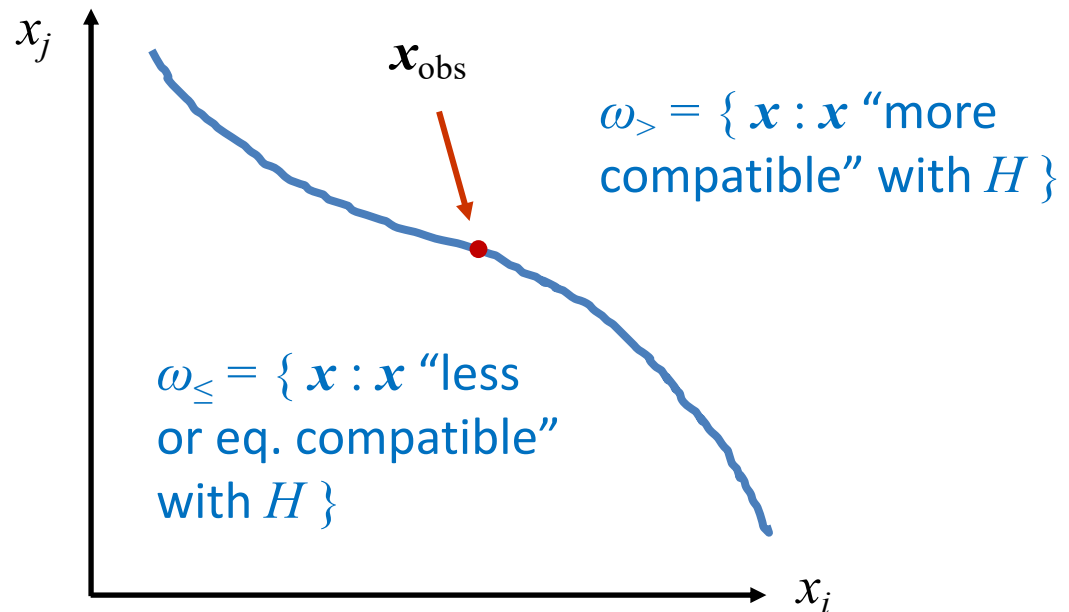
Testing significance / goodness-of-fit

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}_{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}_{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 α (e.g. 0.05) then H is rejected (equivalent to hypothesis test of size α 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.

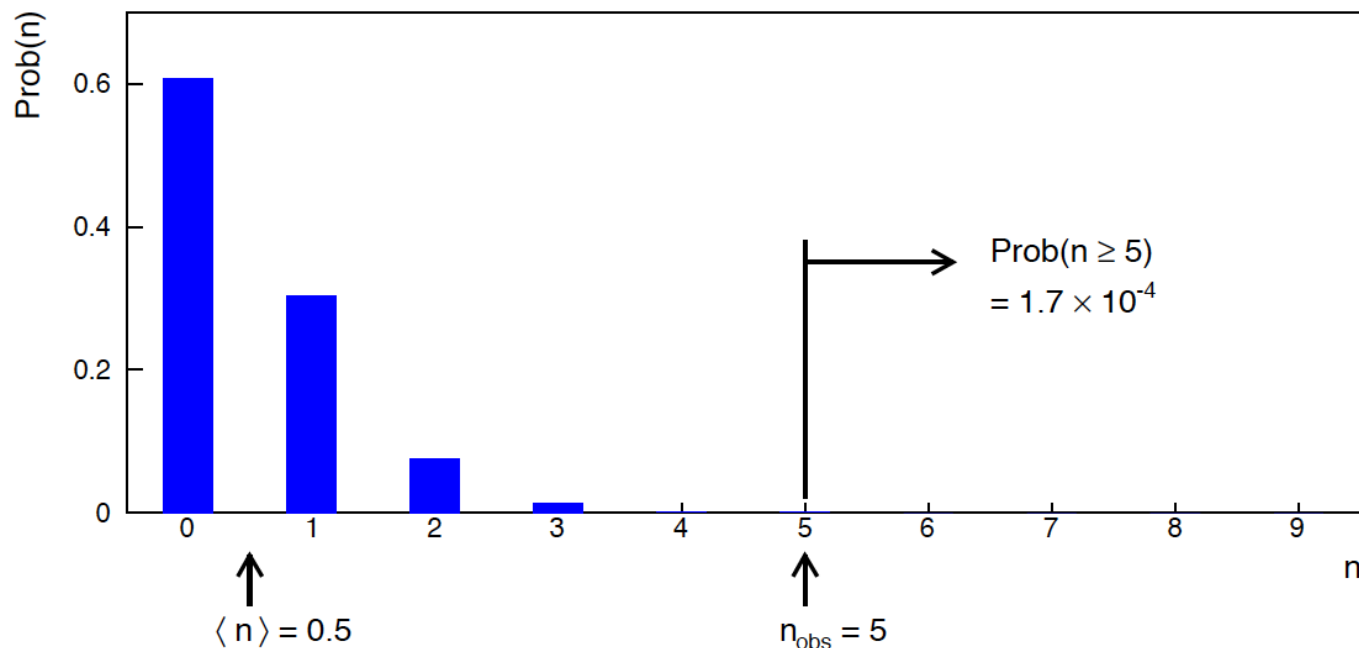
Poisson counting experiment: discovery p -value

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

Should we claim evidence for a new discovery?

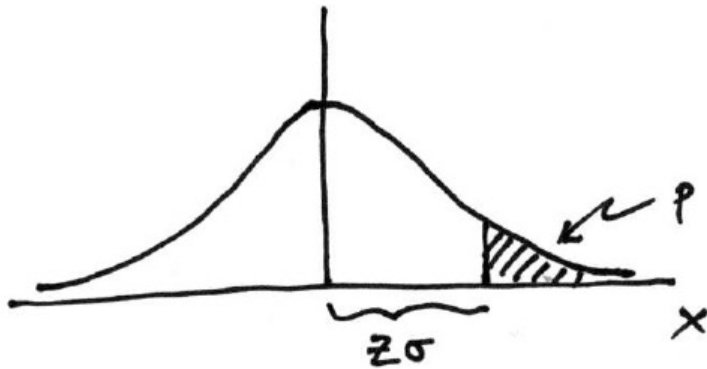
Give p -value for hypothesis $s = 0$:

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



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)
```

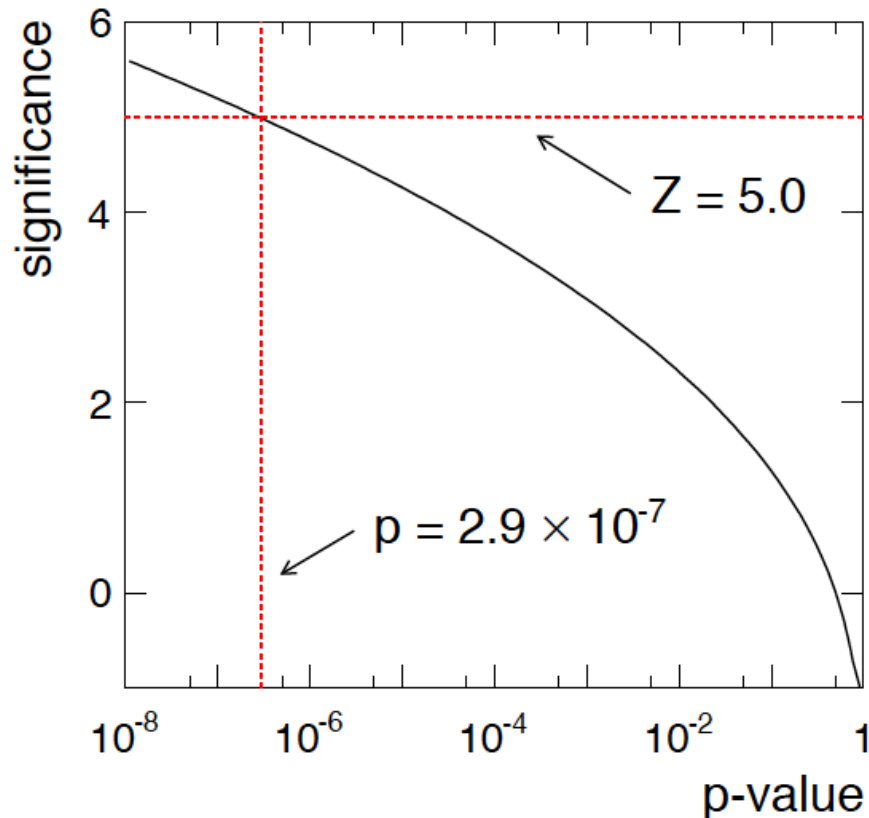
```
Z = norm.ppf(1-p)
```

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”)



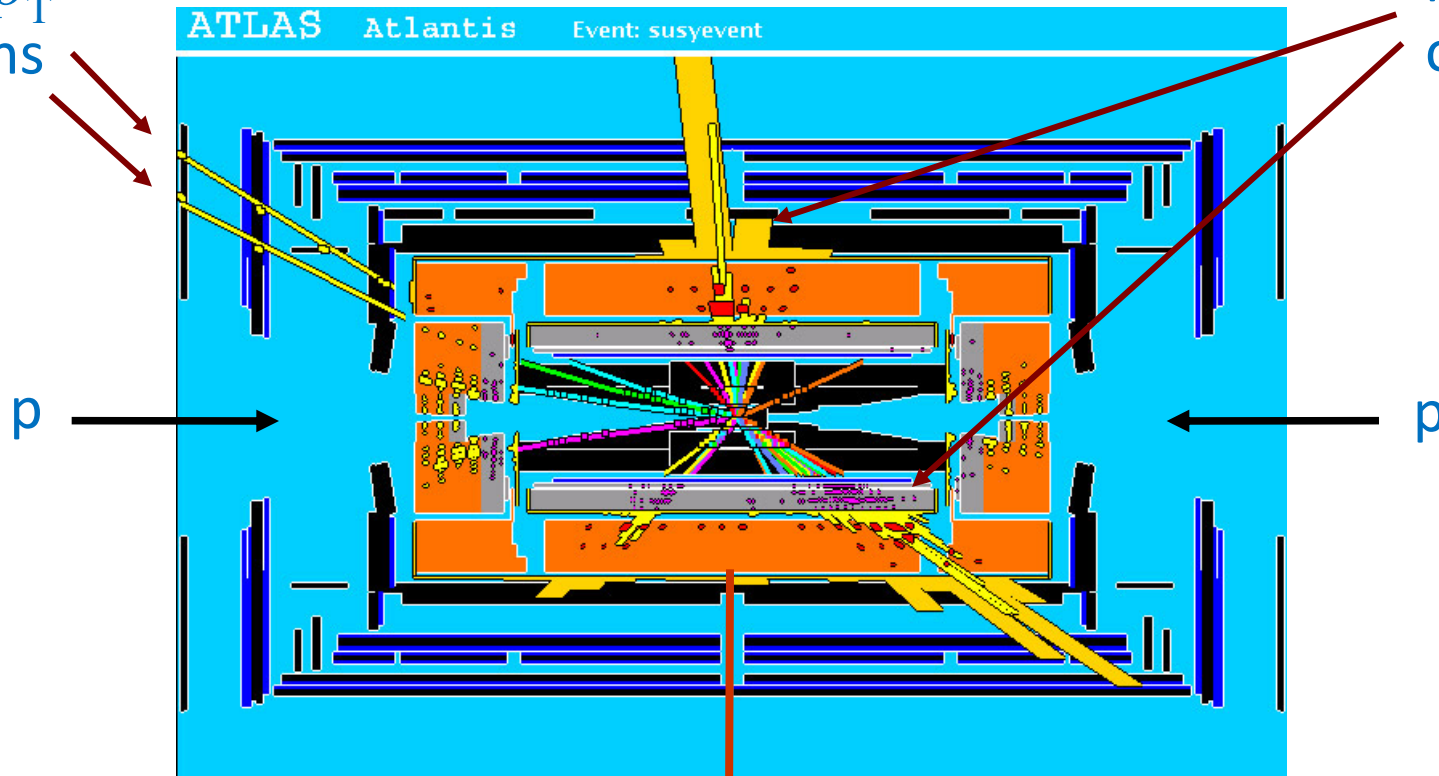
In fact this tradition should be revisited: p -value intended to quantify probability of a signal-like fluctuation assuming background only; not intended to cover, e.g., hidden systematics, plausibility signal model, compatibility of data with signal, “look-elsewhere effect” (\sim multiple testing),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

Particle Physics context for a hypothesis test

A simulated SUSY event (“signal”):

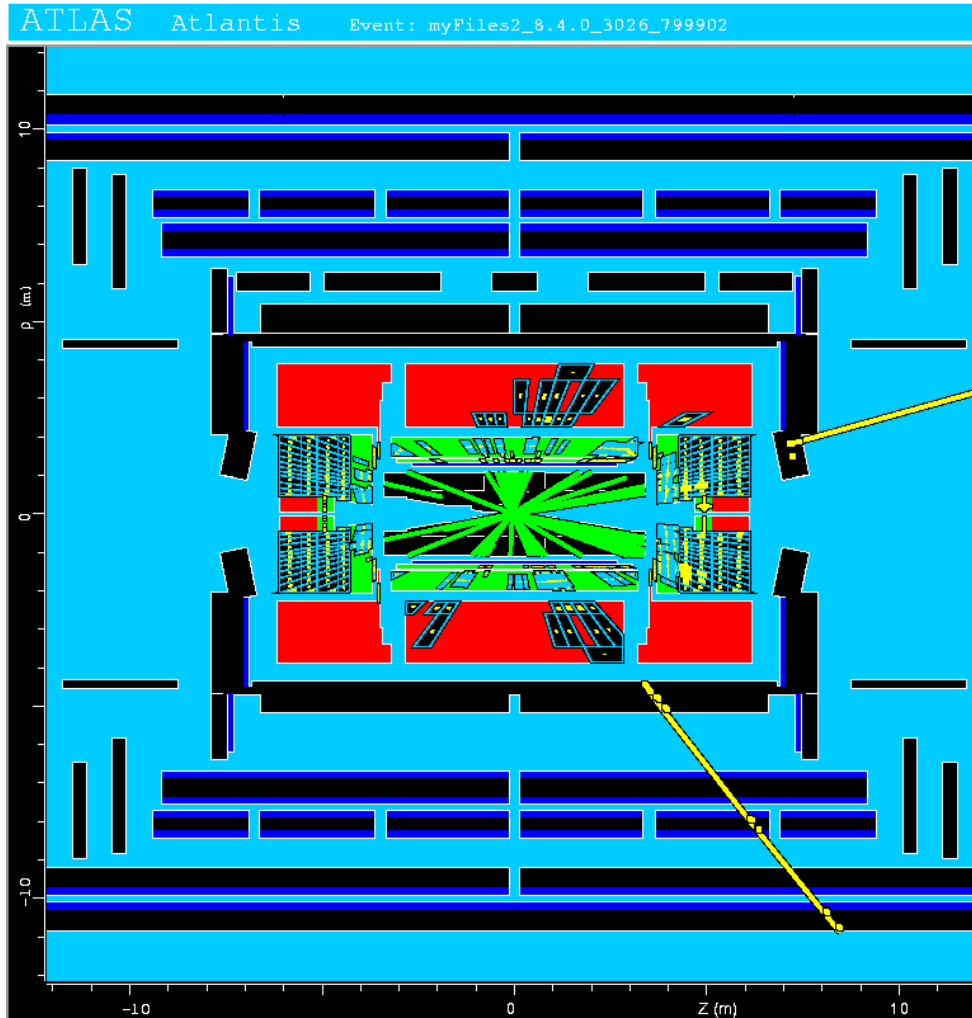
high p_T
muons

high p_T jets
of hadrons



missing transverse energy

Background events



This event from Standard Model $t\bar{t}b\bar{b}$ 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 = ...

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

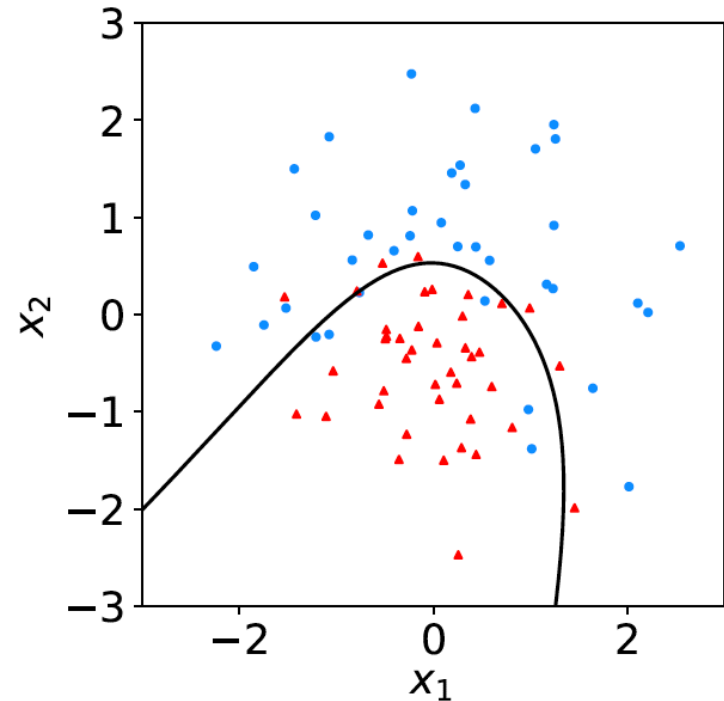
If we consider only two features $\mathbf{x} = (x_1, x_2)$, we can display the results in a scatter plot (red: $y = 0$, blue: $y = 1$).

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

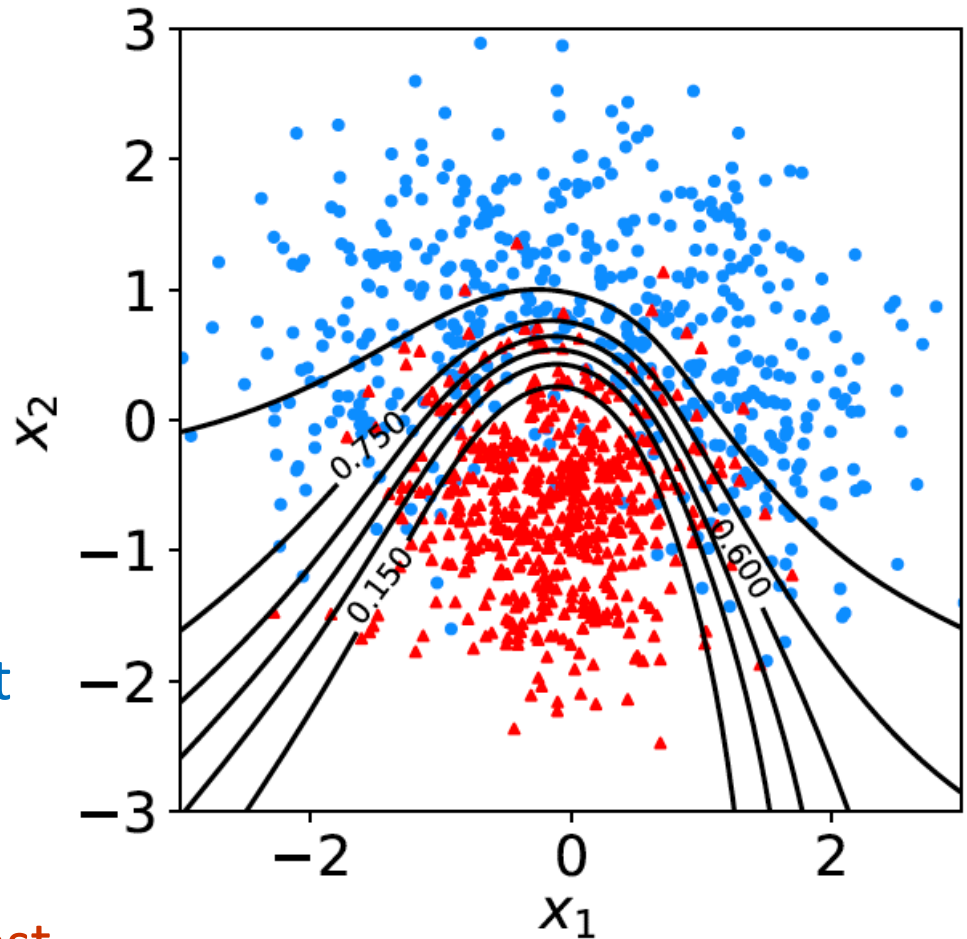
$$t(x_1, \dots, x_n) = t_c$$

scalar
function

constant

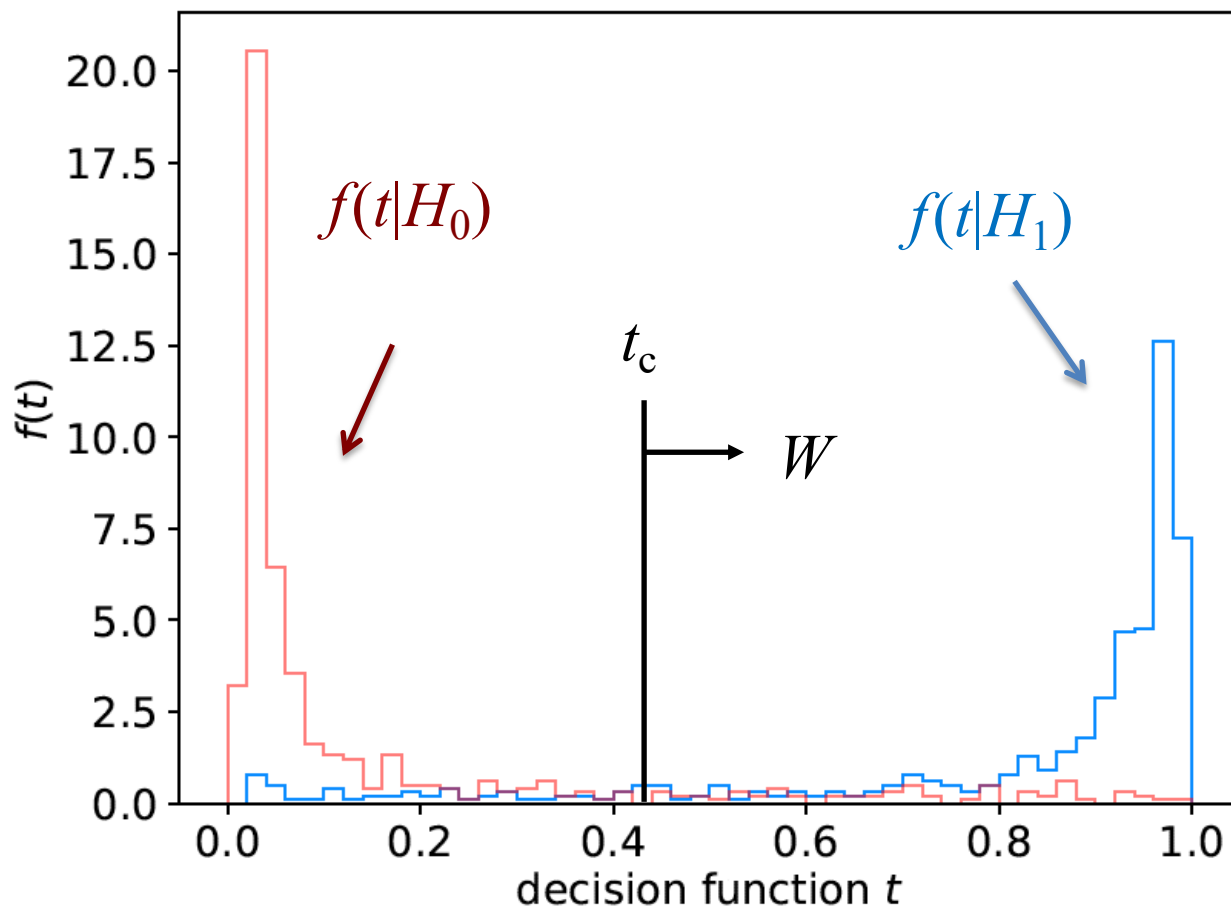
Different values of the constant t_c result in a family of surfaces.

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 .

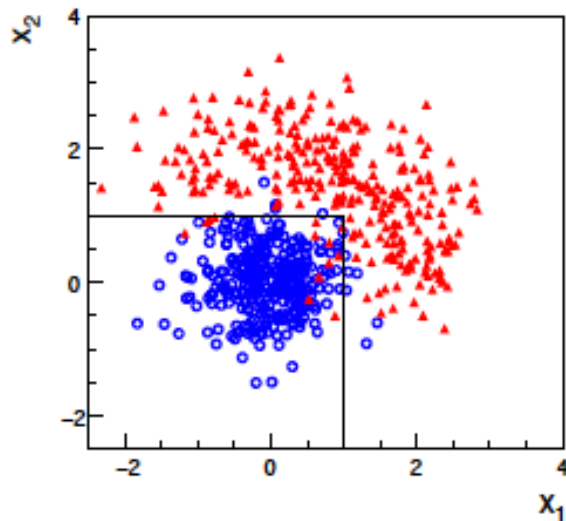


Types of decision boundaries

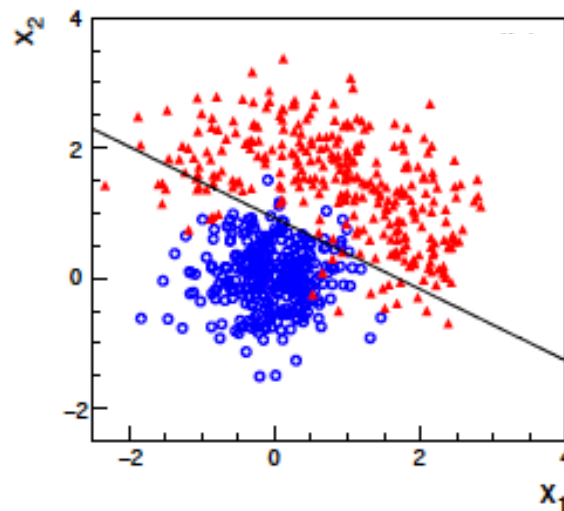
So what is the optimal boundary for the critical region, i.e., what is the optimal test statistic $t(\mathbf{x})$?

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

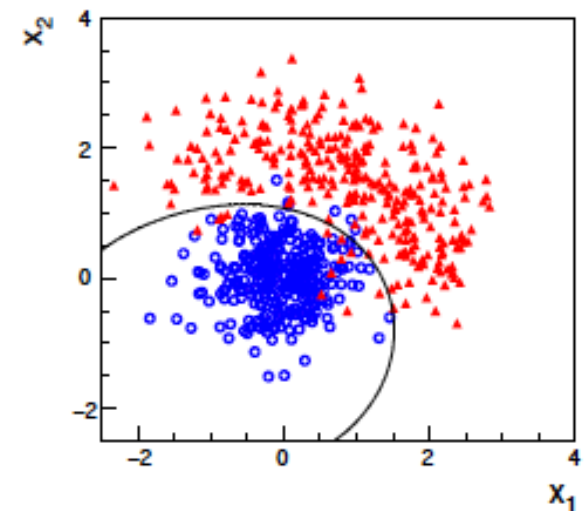
Remember \mathbf{x} -space can have many dimensions.



“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 α , 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_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 leads to the same test.

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 $\mathbf{x}_1, \dots, \mathbf{x}_N$

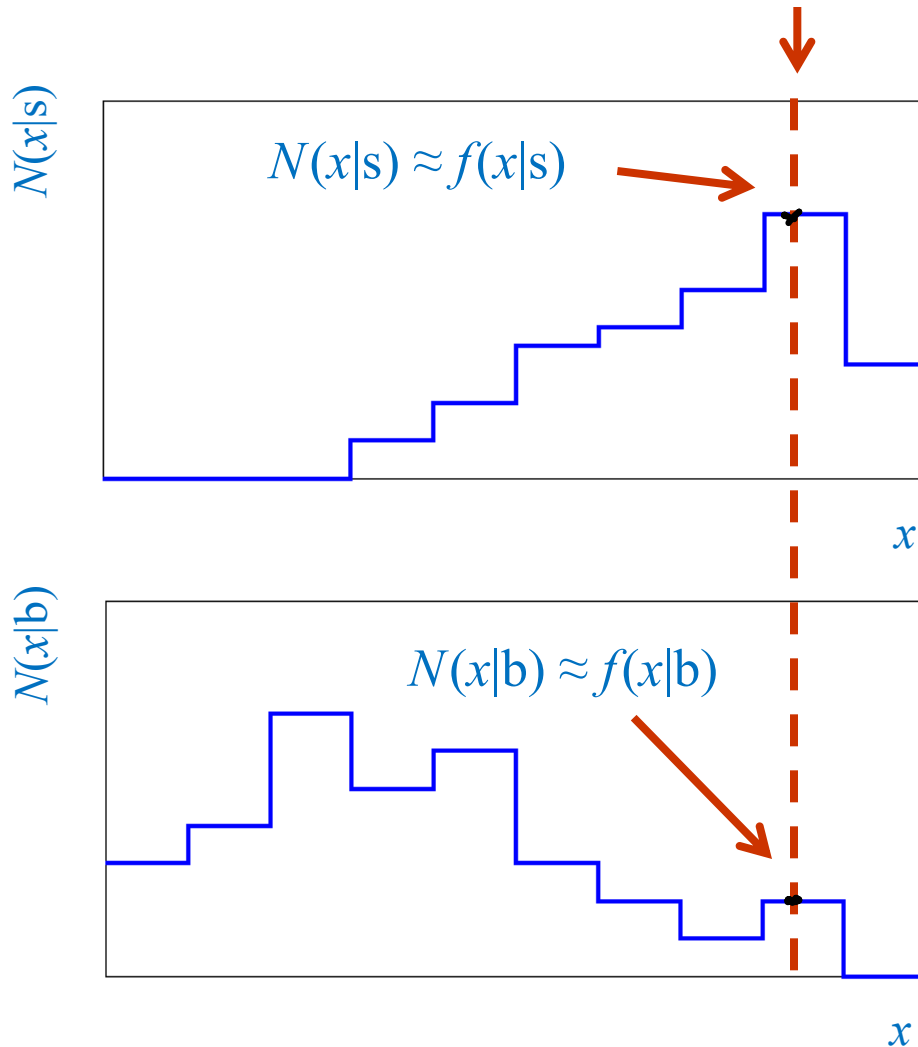
generate $\mathbf{x} \sim f(\mathbf{x}|b)$ \rightarrow $\mathbf{x}_1, \dots, \mathbf{x}_N$

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.

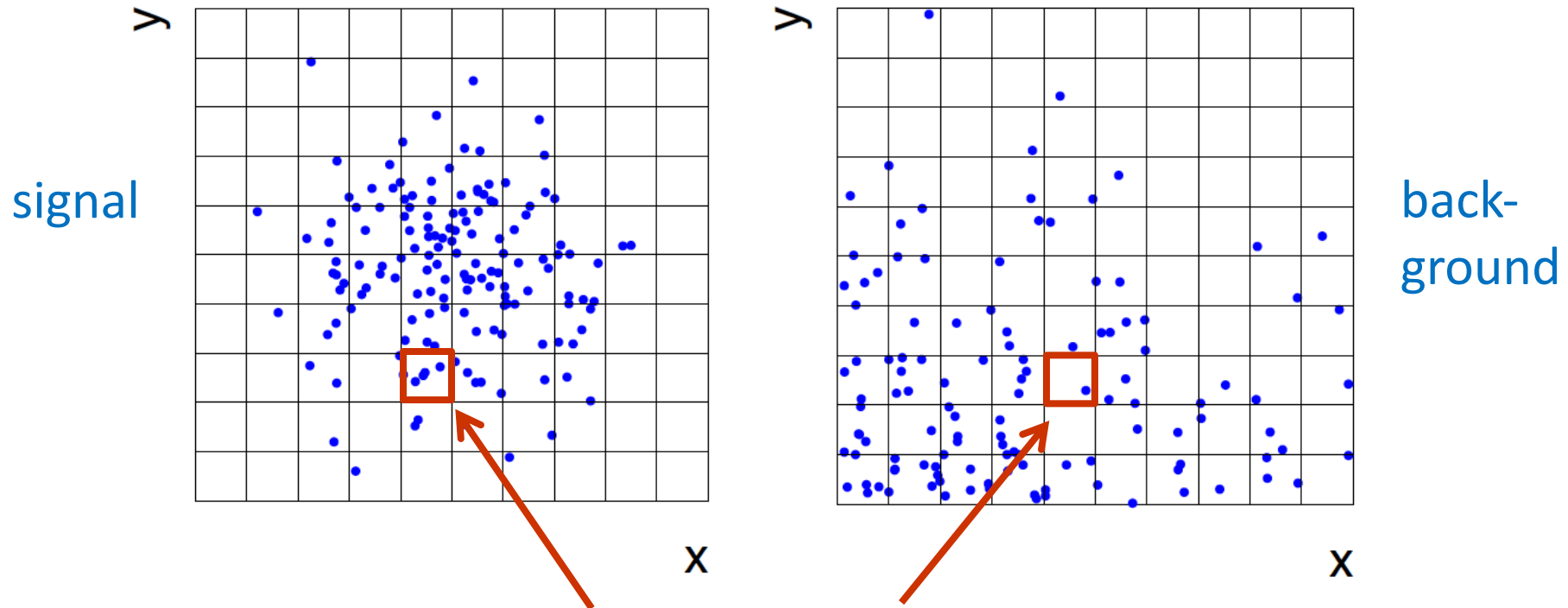
Use (normalized) histogram values to approximate LR:

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

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

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

(Deep) Neural Networks

Kernel density methods

Support Vector Machines

Decision trees

Boosting

Bagging

→ Lectures by Márcio Ferreira

Extra slides

Some distributions

<u>Distribution/pdf</u>	<u>Example use in Particle Physics</u>
Binomial	Branching ratio
Multinomial	Histogram with fixed N
Poisson	Number of events found
Uniform	Monte Carlo method
Exponential	Decay time
Gaussian	Measurement error
Chi-square	Goodness-of-fit
Cauchy	Mass of resonance
Landau	Ionization energy loss
Beta	Prior pdf for efficiency
Gamma	Sum of exponential variables
Student's t	Resolution function with adjustable tails

Binomial distribution

Consider N independent experiments (Bernoulli trials):

outcome of each is ‘success’ or ‘failure’,
probability of success on any given trial is p .

Define discrete r.v. n = number of successes ($0 \leq n \leq N$).

Probability of a specific outcome (in order), e.g. ‘ssfsf’ is

$$pp(1-p)p(1-p) = p^n(1-p)^{N-n}$$

But order not important; there are $\frac{N!}{n!(N-n)!}$

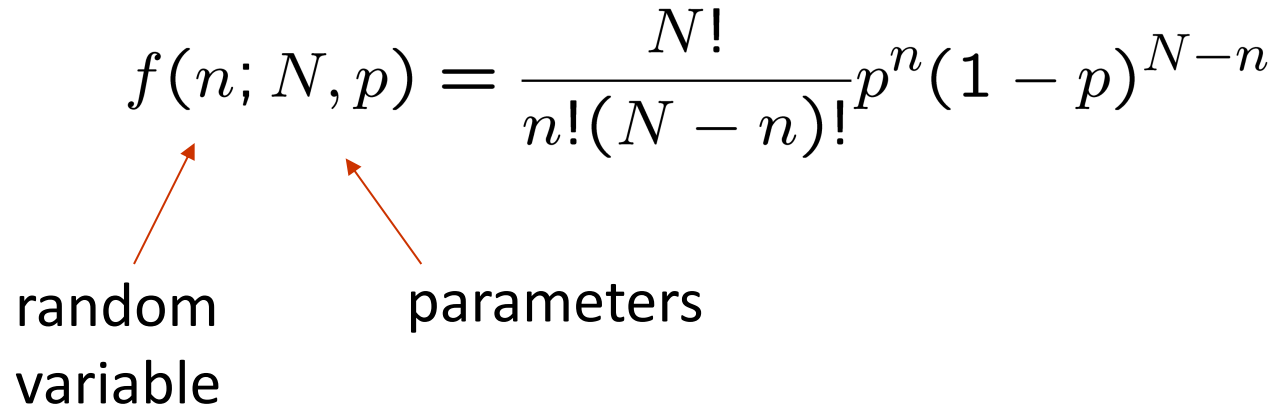
ways (permutations) to get n successes in N trials, total probability for n is sum of probabilities for each permutation.

Binomial distribution (2)

The binomial distribution is therefore

$$f(n; N, p) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$

random variable parameters



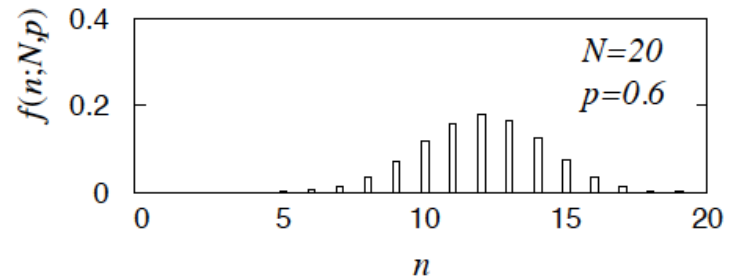
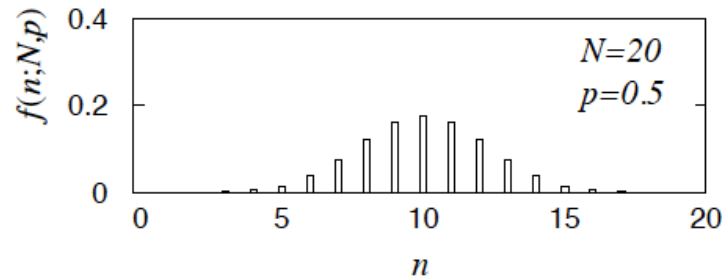
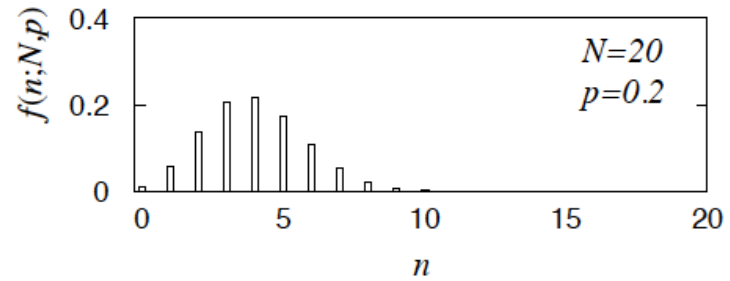
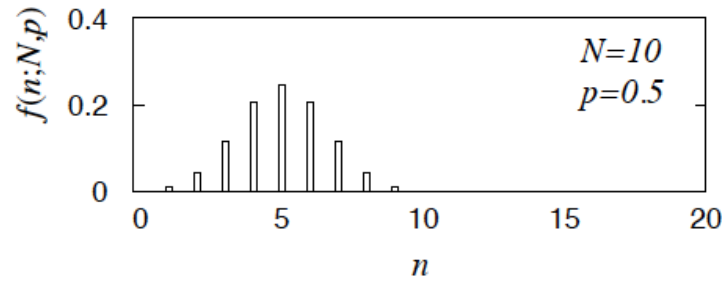
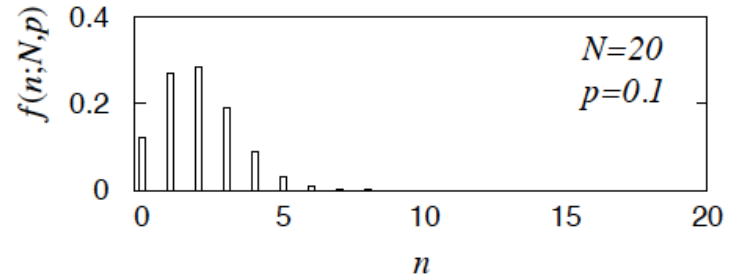
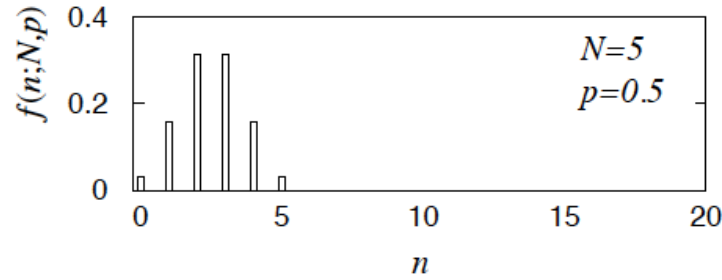
For the expectation value and variance we find:

$$E[n] = \sum_{n=0}^N n f(n; N, p) = Np$$

$$V[n] = E[n^2] - (E[n])^2 = Np(1-p)$$

Binomial distribution (3)

Binomial distribution for several values of the parameters:



Example: observe N decays of W^\pm , the number n of which are $W \rightarrow \mu\nu$ is a binomial r.v., p = branching ratio.

Multinomial distribution

Like binomial but now m outcomes instead of two, probabilities are

$$\vec{p} = (p_1, \dots, p_m), \quad \text{with} \quad \sum_{i=1}^m p_i = 1 .$$

For N trials we want the probability to obtain:

$$\begin{aligned} n_1 &\text{ of outcome 1,} \\ n_2 &\text{ of outcome 2,} \\ &\vdots \\ n_m &\text{ of outcome } m. \end{aligned}$$

This is the multinomial distribution for $\vec{n} = (n_1, \dots, n_m)$

$$f(\vec{n}; N, \vec{p}) = \frac{N!}{n_1! n_2! \dots n_m!} p_1^{n_1} p_2^{n_2} \dots p_m^{n_m}$$

Multinomial distribution (2)

Now consider outcome i as ‘success’, all others as ‘failure’.

→ all n_i individually binomial with parameters N, p_i

$$E[n_i] = Np_i, \quad V[n_i] = Np_i(1 - p_i) \quad \text{for all } i$$

One can also find the covariance to be

$$V_{ij} = Np_i(\delta_{ij} - p_j)$$

Example: $\vec{n} = (n_1, \dots, n_m)$ represents a histogram with m bins, N total entries, all entries independent.

Poisson distribution

Consider binomial n in the limit

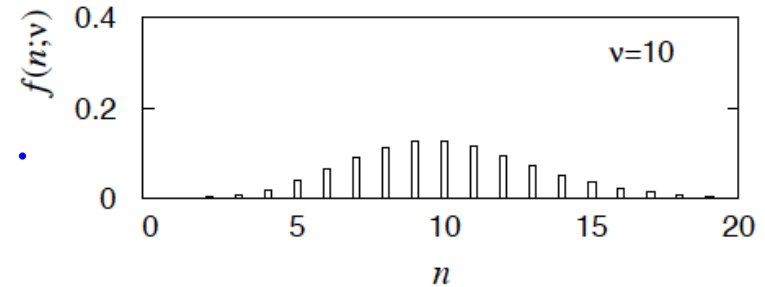
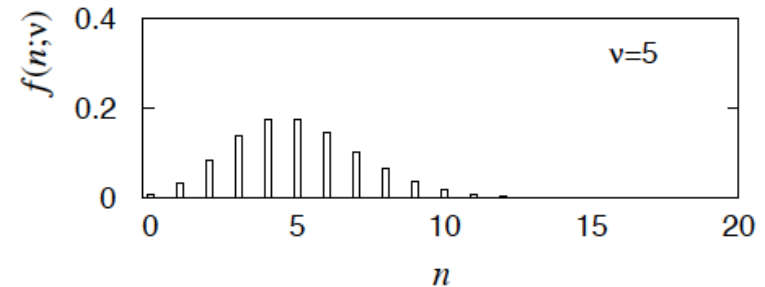
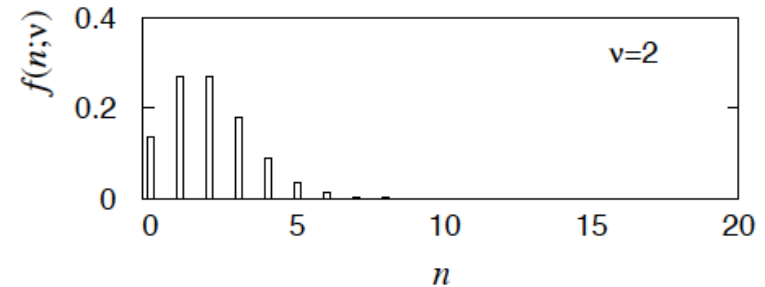
$$N \rightarrow \infty, \quad p \rightarrow 0, \quad E[n] = Np \rightarrow \nu .$$

→ n follows the Poisson distribution:

$$f(n; \nu) = \frac{\nu^n}{n!} e^{-\nu} \quad (n \geq 0)$$

$$E[n] = \nu, \quad V[n] = \nu .$$

Example: number of scattering events n with cross section σ found for a fixed integrated luminosity, with $\nu = \sigma \int L dt$.

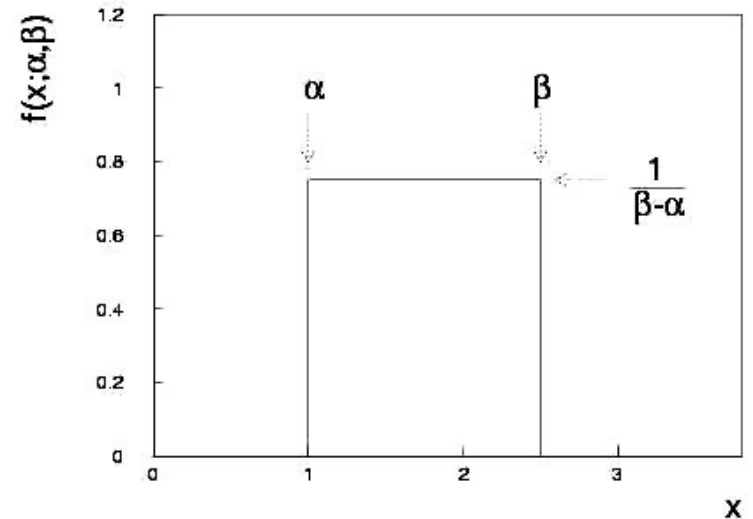


Uniform distribution

$$f(x; \alpha, \beta) = \begin{cases} \frac{1}{\beta - \alpha} & \alpha \leq x \leq \beta \\ 0 & \text{otherwise} \end{cases}$$

$$E[x] = \frac{1}{2}(\alpha + \beta)$$

$$V[x] = \frac{1}{12}(\beta - \alpha)^2$$



Notation: x follows a uniform distribution between α and β

write as: $x \sim U[\alpha, \beta]$

Uniform distribution (2)

Very often used with $\alpha = 0, \beta = 1$ (e.g., Monte Carlo method).

For any r.v. x with pdf $f(x)$, cumulative distribution $F(x)$, the function $y = F(x)$ is uniform in $[0,1]$:

$$g(y) = f(x) \left| \frac{dx}{dy} \right| = \frac{f(x)}{|dy/dx|}$$
$$= \frac{f(x)}{|dF/dx|} = \frac{f(x)}{f(x)} = 1, \quad 0 \leq y \leq 1$$

because $f(x) = dF/dx = dy/dx$

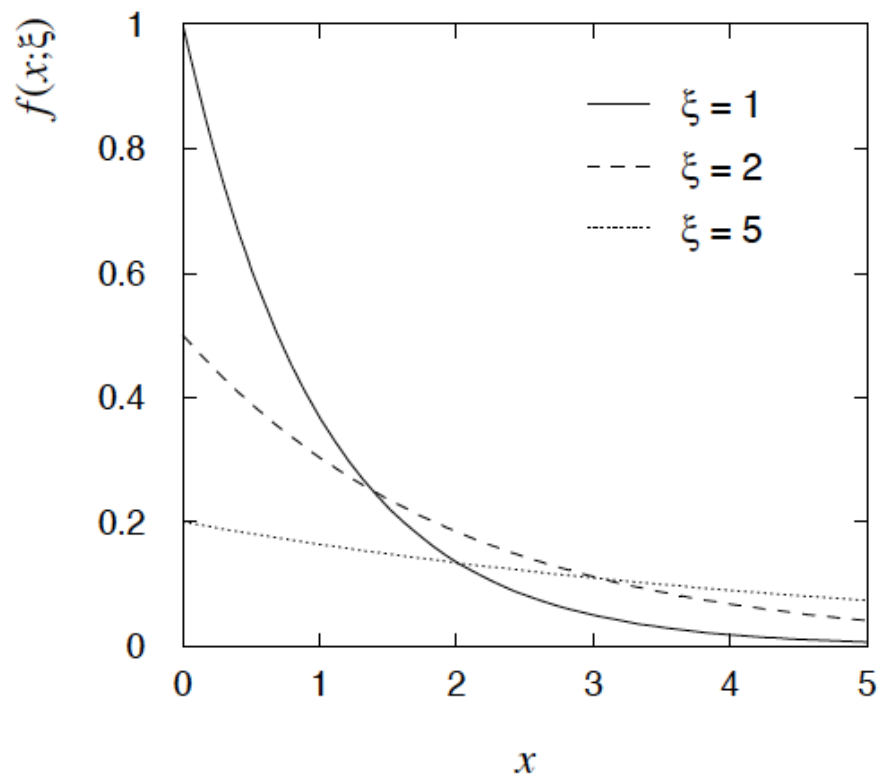
Exponential distribution

The exponential pdf for the continuous r.v. x is defined by:

$$f(x; \xi) = \begin{cases} \frac{1}{\xi} e^{-x/\xi} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$E[x] = \xi$$

$$V[x] = \xi^2$$



Exponential distribution (2)

Example: proper decay time t of an unstable particle

$$f(t; \tau) = \frac{1}{\tau} e^{-t/\tau} \quad (\tau = \text{mean lifetime})$$

Lack of memory (unique to exponential): $f(t - t_0 | t \geq t_0) = f(t)$

Question for discussion:

A cosmic ray muon is created 30 km high in the atmosphere, travels to sea level and is stopped in a block of scintillator, giving a start signal at t_0 . At a time t it decays to an electron giving a stop signal. What is distribution of the difference between stop and start times, i.e., the pdf of $t - t_0$ given $t > t_0$?

Gaussian (normal) distribution

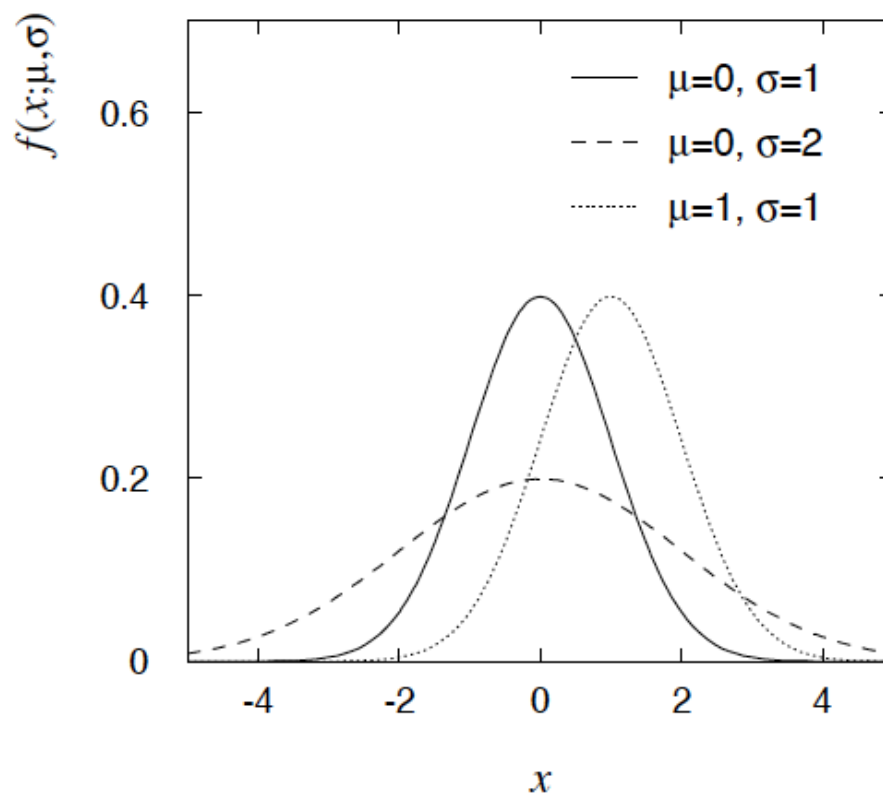
The Gaussian (normal) pdf for a continuous r.v. x is defined by:

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

$$E[x] = \mu$$

$$V[x] = \sigma^2$$

N.B. often μ , σ^2 denote mean, variance of any r.v., not only Gaussian.



Standardized random variables

If a random variable y has pdf $f(y)$ with mean μ and std. dev. σ , then the *standardized* variable

$$x = \frac{y - \mu}{\sigma} \quad \text{has the pdf} \quad g(x) = f(y(x)) \left| \frac{dy}{dx} \right| = \sigma f(\mu + \sigma x)$$

has mean of zero and standard deviation of 1.

Often work with the *standard* Gaussian distribution ($\mu = 0$, $\sigma = 1$) using notation:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad \Phi(x) = \int_{-\infty}^x \varphi(x') dx'$$

Then e.g. $y = \mu + \sigma x$ follows

$$f(y) = \frac{1}{\sigma} \varphi\left(\frac{y - \mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-\mu)^2/2\sigma^2}$$

Multivariate Gaussian distribution

Multivariate Gaussian pdf for the vector $\vec{x} = (x_1, \dots, x_n)$:

$$f(\vec{x}; \vec{\mu}, V) = \frac{1}{(2\pi)^{n/2} |V|^{1/2}} \exp \left[-\frac{1}{2} (\vec{x} - \vec{\mu})^T V^{-1} (\vec{x} - \vec{\mu}) \right]$$

\vec{x} , $\vec{\mu}$ are column vectors, \vec{x}^T , $\vec{\mu}^T$ are transpose (row) vectors,

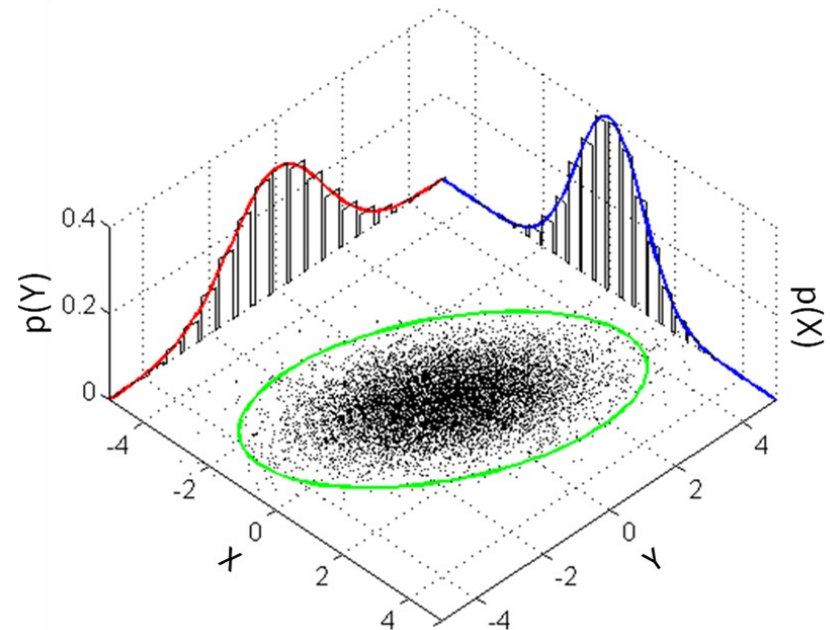
$$E[x_i] = \mu_i, \quad \text{COV}[x_i, x_j] = V_{ij} .$$

Marginal pdf of each x_i is Gaussian with mean μ_i , standard deviation $\sigma_i = \sqrt{V_{ii}}$.

Two-dimensional Gaussian distribution

$$f(x_1, x_2, ; \mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \\ \times \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 - 2\rho \left(\frac{x_1 - \mu_1}{\sigma_1} \right) \left(\frac{x_2 - \mu_2}{\sigma_2} \right) \right] \right\}$$

where $\rho = \text{cov}[x_1, x_2]/(\sigma_1\sigma_2)$
is the correlation coefficient.



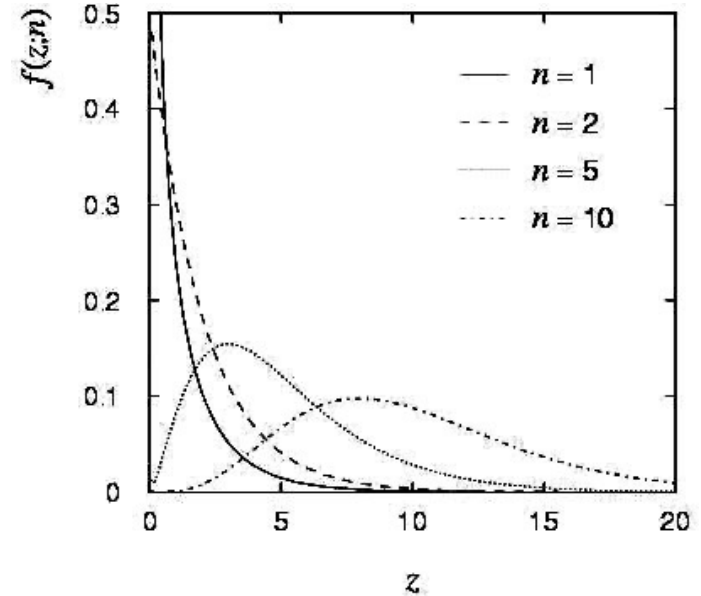
Chi-square (χ^2) distribution

The chi-square pdf for the continuous r.v. z ($z \geq 0$) is defined by

$$f(z; n) = \frac{1}{2^{n/2} \Gamma(n/2)} z^{n/2-1} e^{-z/2}$$

$n = 1, 2, \dots$ = number of 'degrees of freedom' (dof)

$$E[z] = n, \quad V[z] = 2n .$$



For independent Gaussian x_i , $i = 1, \dots, n$, means μ_i , variances σ_i^2 ,

$$z = \sum_{i=1}^n \frac{(x_i - \mu_i)^2}{\sigma_i^2} \quad \text{follows } \chi^2 \text{ pdf with } n \text{ dof.}$$

Example: goodness-of-fit test variable especially in conjunction with method of least squares.

Cauchy (Breit-Wigner) distribution

The Breit-Wigner pdf for the continuous r.v. x is defined by

$$f(x; \Gamma, x_0) = \frac{1}{\pi} \frac{\Gamma/2}{\Gamma^2/4 + (x - x_0)^2}$$

($\Gamma = 2, x_0 = 0$ is the Cauchy pdf.)

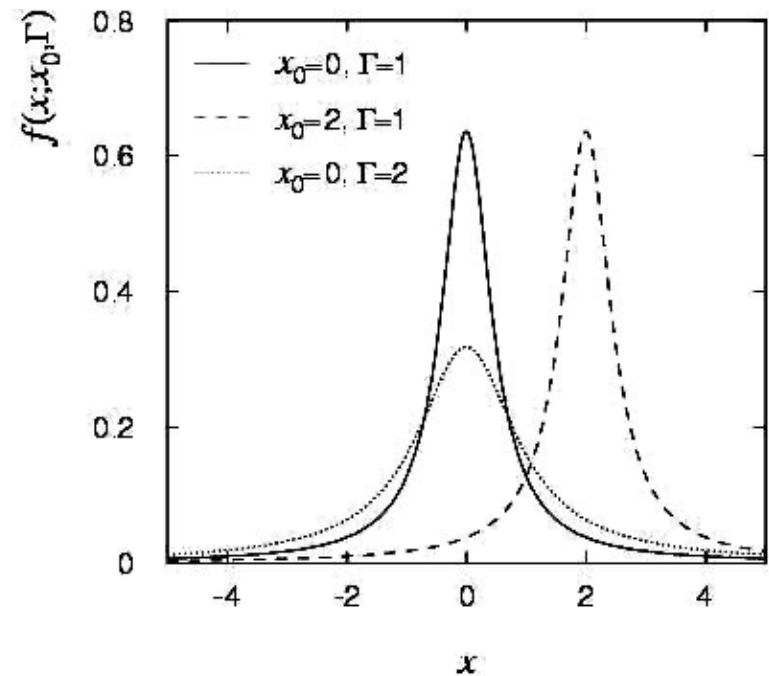
$E[x]$ not well defined, $V[x] \rightarrow \infty$.

$x_0 = \text{mode}$ (most probable value)

$\Gamma = \text{full width at half maximum}$

Example: mass of resonance particle, e.g. ρ, K^*, ϕ^0, \dots

$\Gamma = \text{decay rate}$ (inverse of mean lifetime)



Landau distribution

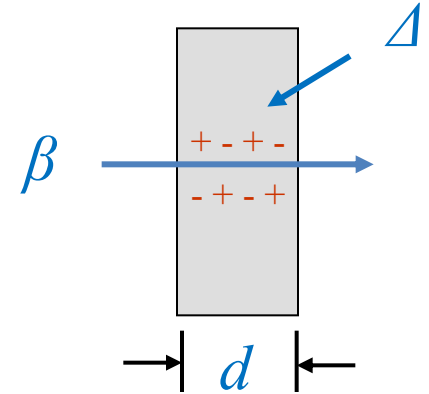
For a charged particle with $\beta = v/c$ traversing a layer of matter of thickness d , the energy loss Δ follows the Landau pdf:

$$f(\Delta; \beta) = \frac{1}{\xi} \phi(\lambda) ,$$

$$\phi(\lambda) = \frac{1}{\pi} \int_0^\infty \exp(-u \ln u - \lambda u) \sin \pi u \, du ,$$

$$\lambda = \frac{1}{\xi} \left[\Delta - \xi \left(\ln \frac{\xi}{\epsilon'} + 1 - \gamma_E \right) \right] ,$$

$$\xi = \frac{2\pi N_A e^4 z^2 \rho \sum Z}{m_e c^2 \sum A} \frac{d}{\beta^2} , \quad \epsilon' = \frac{I^2 \exp \beta^2}{2m_e c^2 \beta^2 \gamma^2} .$$



L. Landau, J. Phys. USSR **8** (1944) 201; see also

W. Allison and J. Cobb, Ann. Rev. Nucl. Part. Sci. **30** (1980) 253.

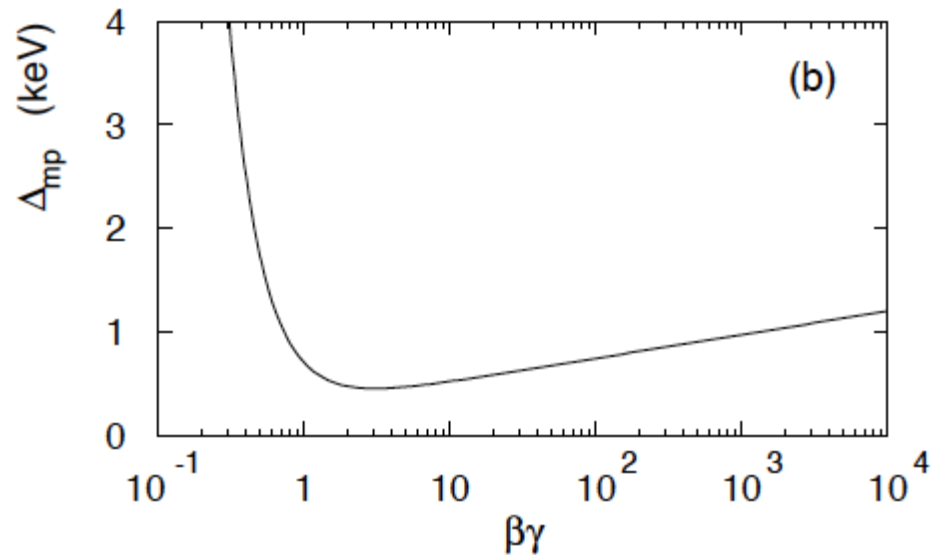
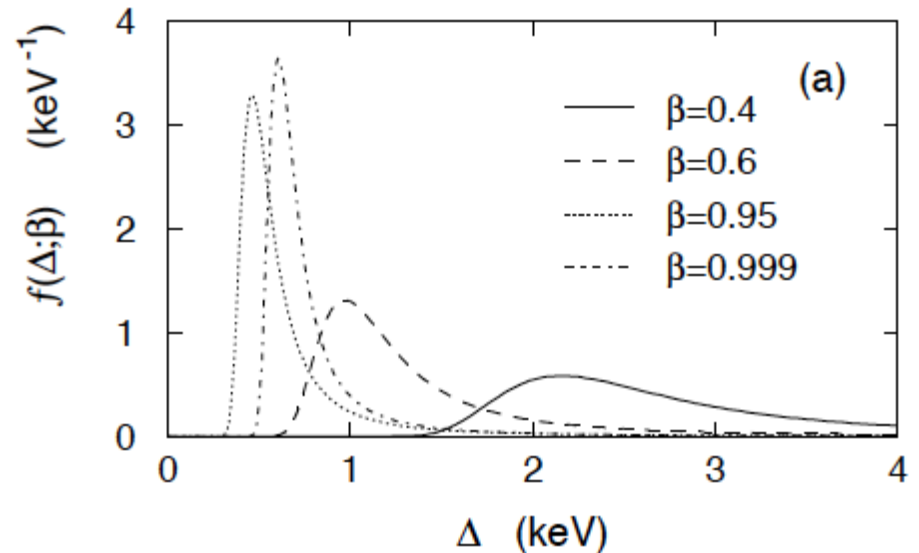
Landau distribution (2)

Long 'Landau tail'

→ all moments ∞

Mode (most probable value) sensitive to β ,

→ particle i.d.



Beta distribution

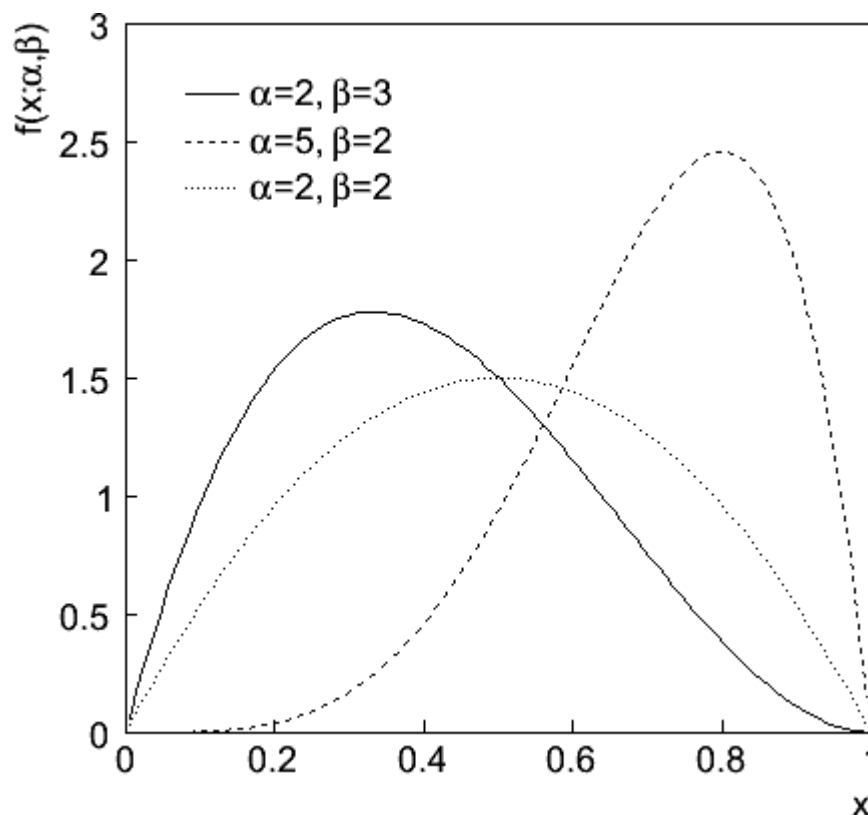
$$f(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1 - x)^{\beta-1}$$

$$E[x] = \frac{\alpha}{\alpha + \beta}$$

$$V[x] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

Often used to represent pdf of continuous r.v. nonzero only between finite limits, e.g.,

$$y = a_0 + a_1x, \quad a_0 \leq y \leq a_0 + a_1$$



Gamma distribution

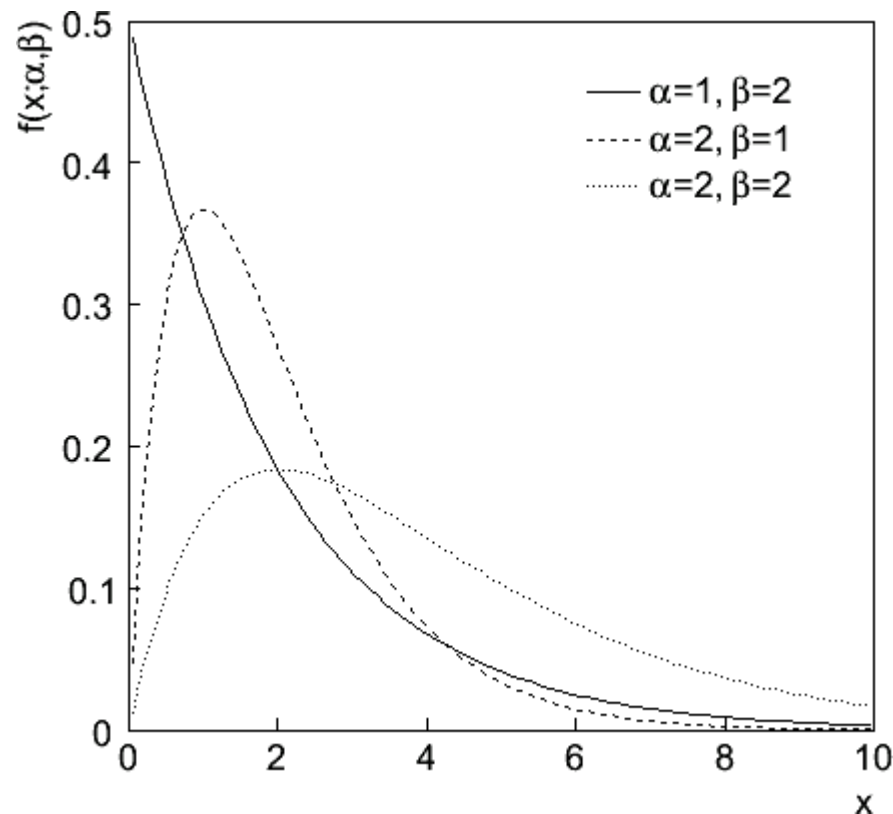
$$f(x; \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$$

$$E[x] = \alpha\beta$$

$$V[x] = \alpha\beta^2$$

Often used to represent pdf of continuous r.v. nonzero only in $[0, \infty]$.

Also e.g. sum of n exponential r.v.s or time until n th event in Poisson process \sim Gamma



Student's t distribution

$$f(x; \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi} \Gamma(\nu/2)} \left(1 + \frac{x^2}{\nu}\right)^{-\left(\frac{\nu+1}{2}\right)}$$

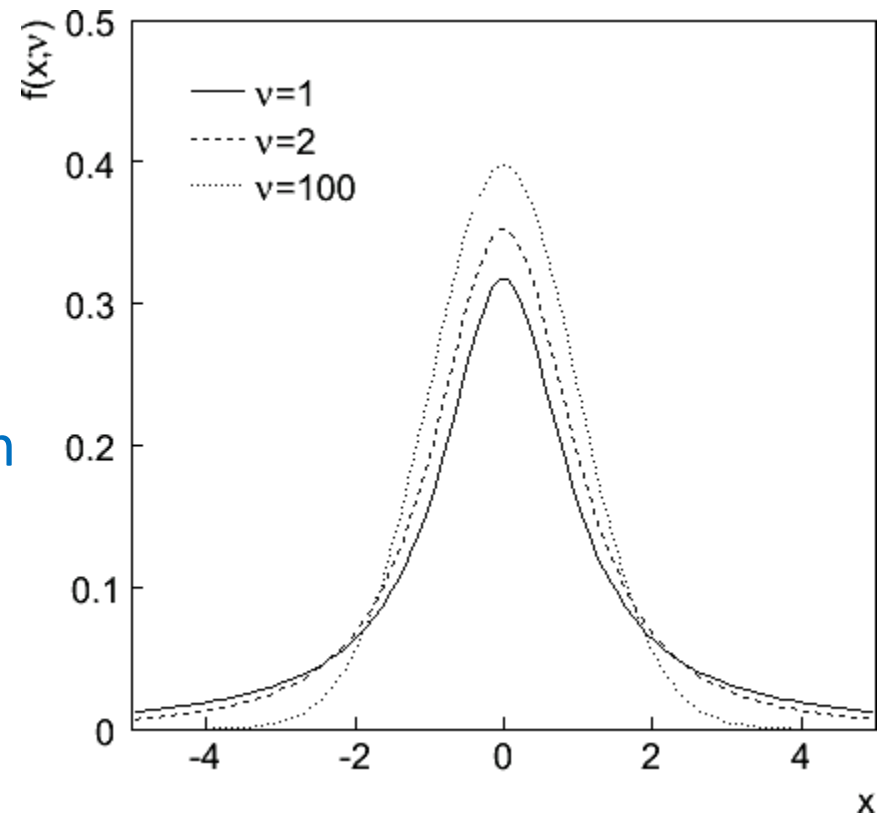
$$E[x] = 0 \quad (\nu > 1)$$

$$V[x] = \frac{\nu}{\nu - 2} \quad (\nu > 2)$$

ν = number of degrees of freedom
(not necessarily integer)

$\nu = 1$ gives Cauchy,

$\nu \rightarrow \infty$ gives Gaussian.



Student's t distribution (2)

If $x \sim$ Gaussian with $\mu = 0$, $\sigma^2 = 1$, and

$z \sim \chi^2$ with n degrees of freedom, then

$t = x / (z/n)^{1/2}$ follows Student's t with $\nu = n$.

This arises in problems where one forms the ratio of a sample mean to the sample standard deviation of Gaussian r.v.s.

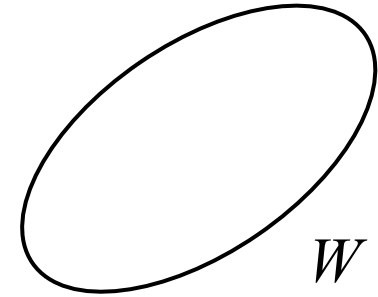
The Student's t provides a bell-shaped pdf with adjustable tails, ranging from those of a Gaussian, which fall off very quickly, ($\nu \rightarrow \infty$, but in fact already very Gauss-like for $\nu =$ two dozen), to the very long-tailed Cauchy ($\nu = 1$).

Developed in 1908 by William Gosset, who worked under the pseudonym "Student" for the Guinness Brewery.

Proof of Neyman-Pearson Lemma

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

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



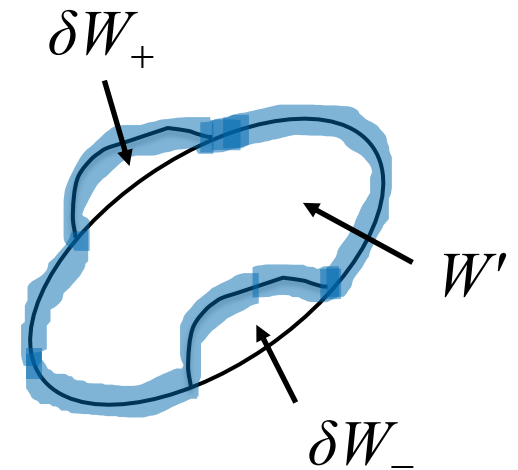
Try to change this into a different critical region W' retaining the same size α , i.e.,

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

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

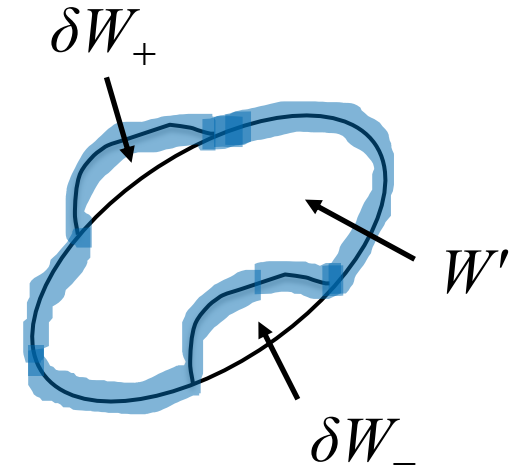
$$W \rightarrow W' = W + \delta W_+ - \delta W_-$$

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



Proof of Neyman-Pearson Lemma (2)

But we are supposing the LR is higher for all \mathbf{x} in δW_- removed than for the \mathbf{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

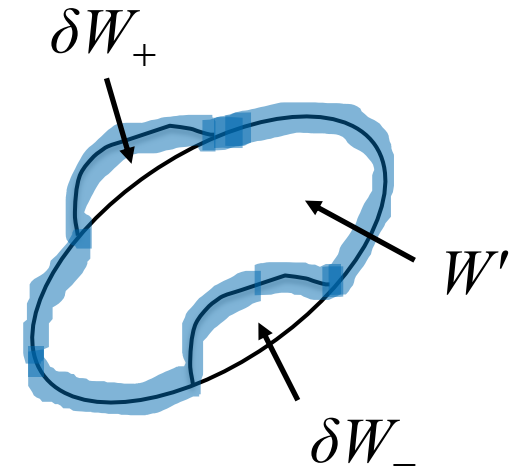
$$P(\mathbf{x} \in \delta W_+ | H_1) \leq P(\mathbf{x} \in \delta W_- | H_1)$$

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

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

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.