Statistical Methods for Particle Physics

Lecture 2: Introduction to Multivariate Methods

http://benasque.org/2019tae/

CENTRO DE CIENCIAS DE BENASQUE PEDRO PASCUAL

TAE 2019
Benasque, Spain
8-21 Sept 2019

ROYAL HOLLOWAY UNIVERSITY OF LONDON

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

Outline

Lecture 1: Introduction and review of fundamentals

Probability, random variables, pdfs

Parameter estimation, maximum likelihood

Introduction to statistical tests

→ Lecture 2: More on statistical tests

Discovery, limits

Bayesian limits

Lecture 3: Framework for full analysis

Nuisance parameters and systematic uncertainties

Tests from profile likelihood ratio

Lecture 4: Further topics

More parameter estimation, Bayesian methods

Experimental sensitivity

Statistical tests for event selection

Suppose the result of a measurement for an individual event is a collection of numbers $\vec{x} = (x_1, \ldots, x_n)$

x_1 = number of muons,

x_2 = mean p_T of jets,

x_3 = missing energy, ...

$\vec{x}$ follows some n-dimensional joint pdf, which depends on the type of event produced, i.e., was it

$$pp \to t\bar{t} \ , \quad pp \to \tilde{g}\tilde{g} \ , \ldots$$

For each reaction we consider we will have a hypothesis for the pdf of $\boldsymbol{x}$, e.g., $p(\boldsymbol{x}|\mathrm{b})$, $p(\boldsymbol{x}|\mathrm{s})$

E.g. here call H_0 the background hypothesis (the event type we want to reject); H_1 is signal hypothesis (the type we want).

Selecting events

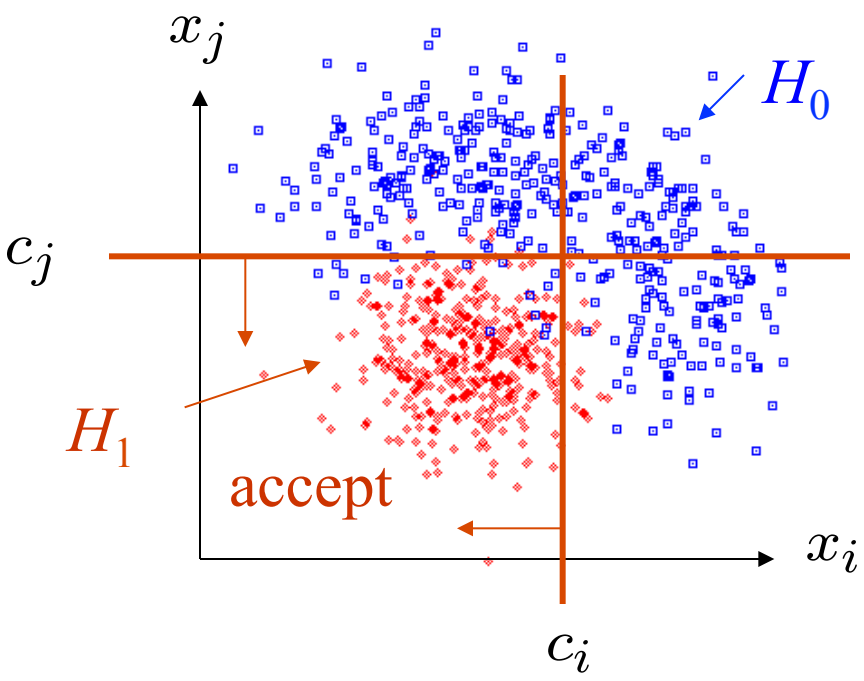
Suppose we have a data sample with two kinds of events, corresponding to hypotheses H_0 and H_1 and we want to select those of type H_1.

Each event is a point in $\vec{x}$ space. What 'decision boundary' should we use to accept/reject events as belonging to event types H_0 or H_1?

Perhaps select events with 'cuts':

$$x_i < c_i$$

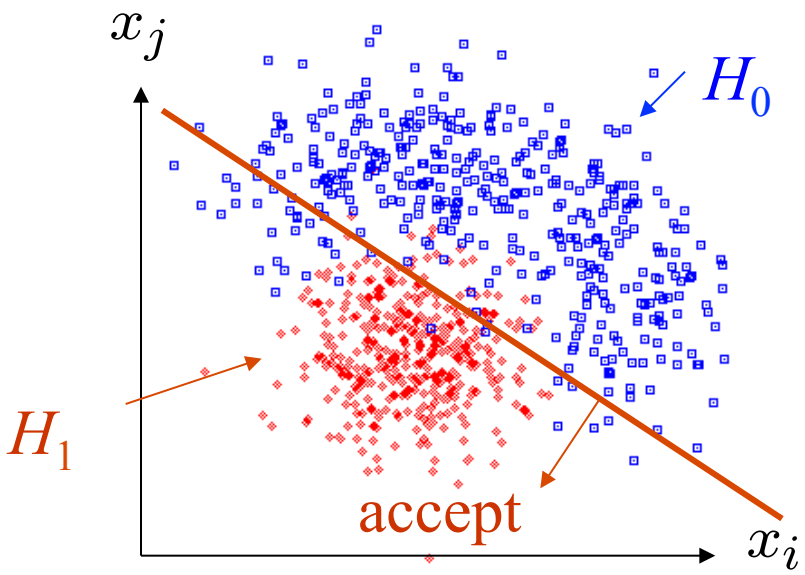
$$x_j < c_j$$



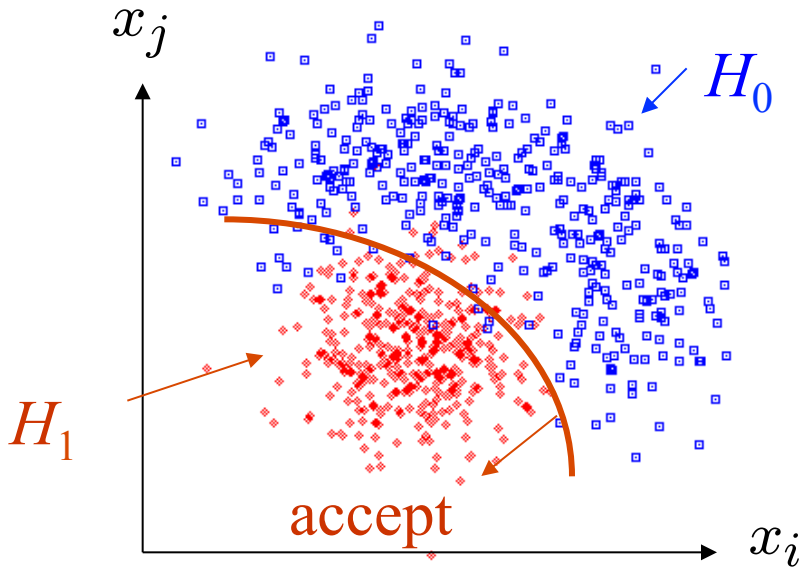
Other ways to select events

Or maybe use some other sort of decision boundary:

linear



or nonlinear



How can we do this in an 'optimal' way?

Test statistics

The boundary of the critical region for an n-dimensional data space $\boldsymbol{x} = (x_1,..., x_n)$ can be defined by an equation of the form

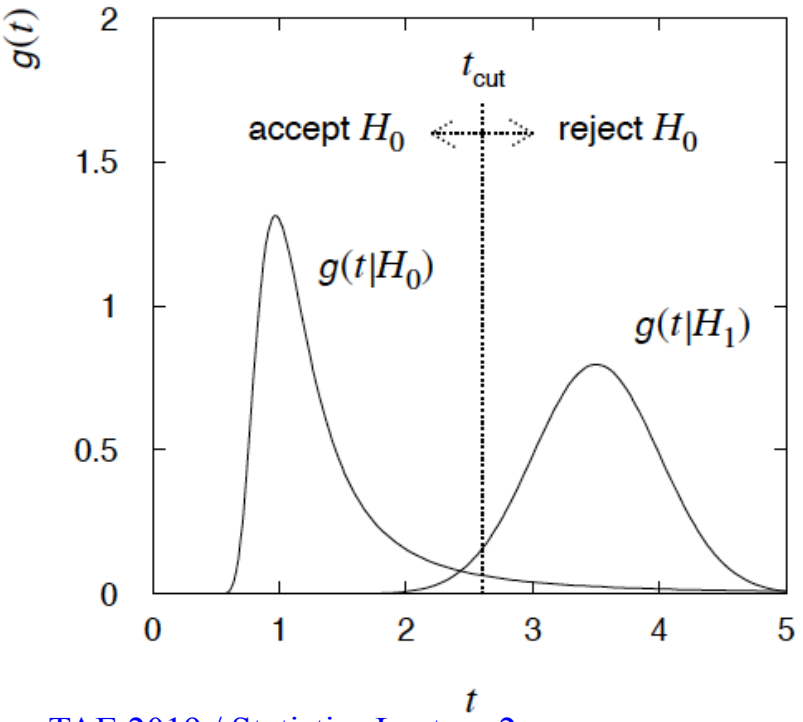
$$t(x_1, \ldots, x_n) = t_{\text{cut}}$$

where $t(x_1,\ldots, x_n)$ is a scalar test statistic.

We can work out the pdfs $g(t|H_0),\ g(t|H_1),\ \ldots$

Decision boundary is now a single 'cut' on t, defining the critical region.

So for an n-dimensional problem we have a corresponding 1-d problem.



Test statistic based on likelihood ratio

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

Neyman-Pearson lemma states:

To get the highest power for a given significance level in a test of H_0, (background) versus H_1, (signal) the critical region should have

$$\frac{f(\mathbf{x}|H_1)}{f(\mathbf{x}|H_0)} > c$$

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

Equivalently, optimal scalar test statistic is

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

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

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:

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

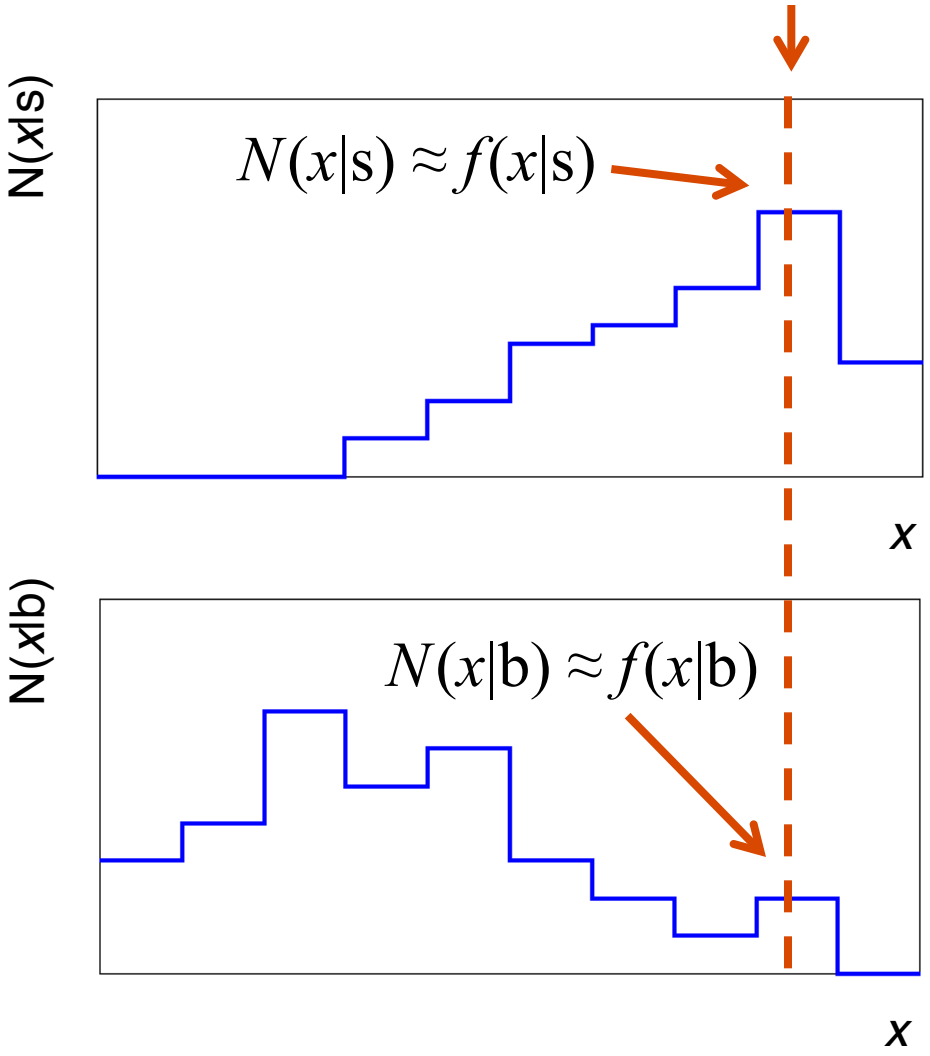
$$\text{generate } \boldsymbol{x} \sim f(\boldsymbol{x}|\mathrm{b}) \quad \rightarrow \quad \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$$

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

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

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.

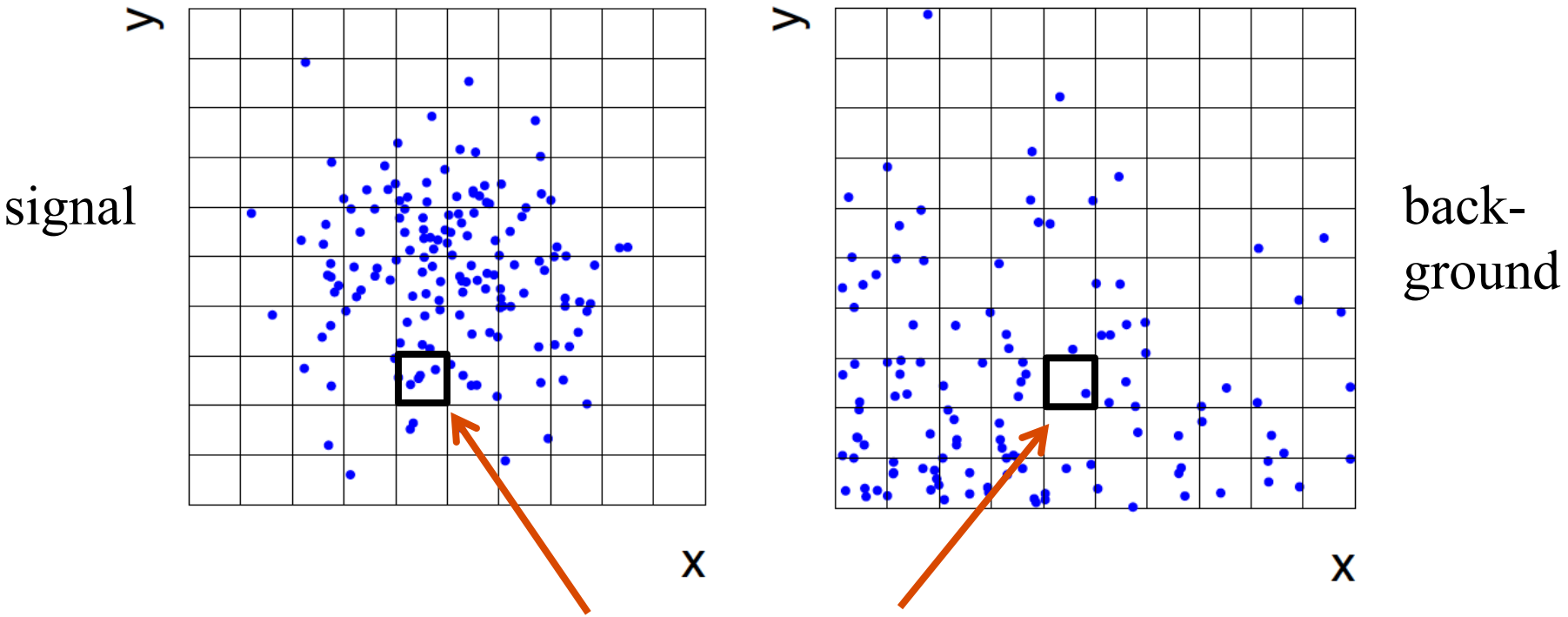
Use (normalized) histogram values to approximate LR:

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

Can work well for single variable.

Approximate LR from 2D-histograms

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



Approximate pdfs using $N(x,y|\mathrm{s})$, $N(x,y|\mathrm{b})$ 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

Many new (and some old) methods esp. from Machine Learning:

- Fisher discriminant
- (Deep) neural networks
- Kernel density methods
- Support Vector Machines
- Decision trees
  - Boosting
  - Bagging

This is a large topic -- see e.g. lectures by Stefano Carrazza or

http://www.pp.rhul.ac.uk/~cowan/stat/stat_2.pdf (from around p 38)

and references therein.

Testing significance / goodness-of-fit

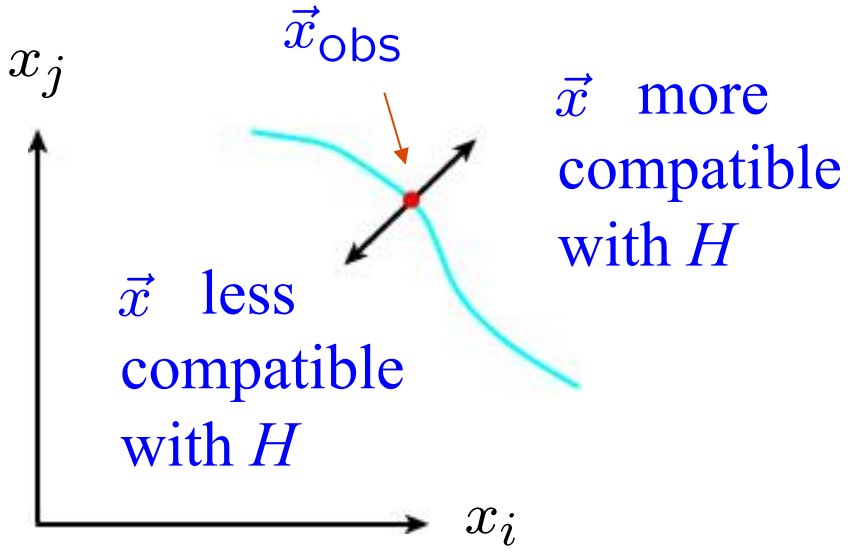
Suppose hypothesis H predicts pdf $f(\vec{x}|H)$ for a set of observations $\vec{x} = (x_1, \ldots, x_n)$.

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

What can we say about the validity of H in light of the data?

Decide what part of the data space represents less compatibility with H than does the point $\vec{x}_{\text{obs}}$.

This region therefore has greater compatibility with some alternative H'.



p-values

Express 'goodness-of-fit' by giving the *p*-value for *H*:

p = probability, under assumption of *H*, to observe data with equal or lesser compatibility with *H* relative to the data we got.

This is not the probability that *H* is true!

In frequentist statistics we don't talk about *P*(*H*) (unless *H* represents a repeatable observation). In Bayesian statistics we do; use Bayes' theorem to obtain

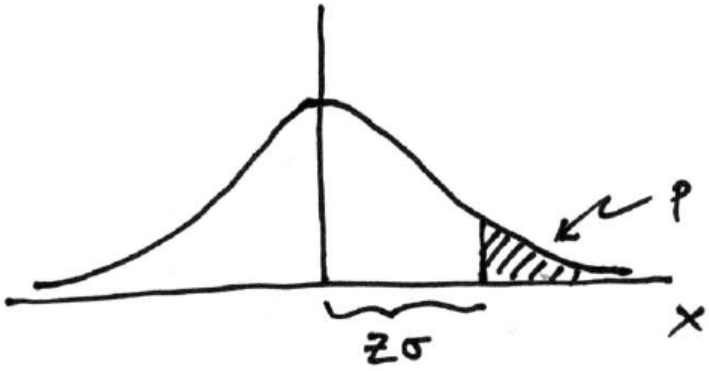
$$P(H|\vec{x}) = \frac{P(\vec{x}|H)\pi(H)}{\int P(\vec{x}|H)\pi(H)\,dH}$$

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*).

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.



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

$$Z = \Phi^{-1}(1-p)$$ `TMath::NormQuantile`

Test statistics and p-values

Consider a parameter μ proportional to rate of signal process.

Often define a function of the data (test statistic) q_μ that reflects level of agreement between the data and the hypothesized value μ.

Usually define q_μ so that higher values increasingly incompatibility with the data (more compatible with a relevant alternative).

We can define critical region of test of μ by $q_\mu \geq$ const., or equivalently define the p-value of μ as:

$$p_\mu = \int_{q_{\mu,\mathrm{obs}}}^{\infty} f(q_\mu|\mu)\, dq_\mu$$

observed value of q_μ

pdf of q_μ assuming μ

Equivalent formulation of test: reject μ if $p_\mu < \alpha$.

Confidence interval from inversion of a test

Carry out a test of size α for all values of μ.

The values that are not rejected constitute a *confidence interval* for μ at confidence level CL = $1 - \alpha$.

> The confidence interval will by construction contain the true value of μ with probability of at least $1 - \alpha$.

The interval will cover the true value of μ with probability $\geq 1 - \alpha$.

Equivalently, the parameter values in the confidence interval have p-values of at least α.

To find edge of interval (the "limit"), set $p_\mu = \alpha$ and solve for μ.

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

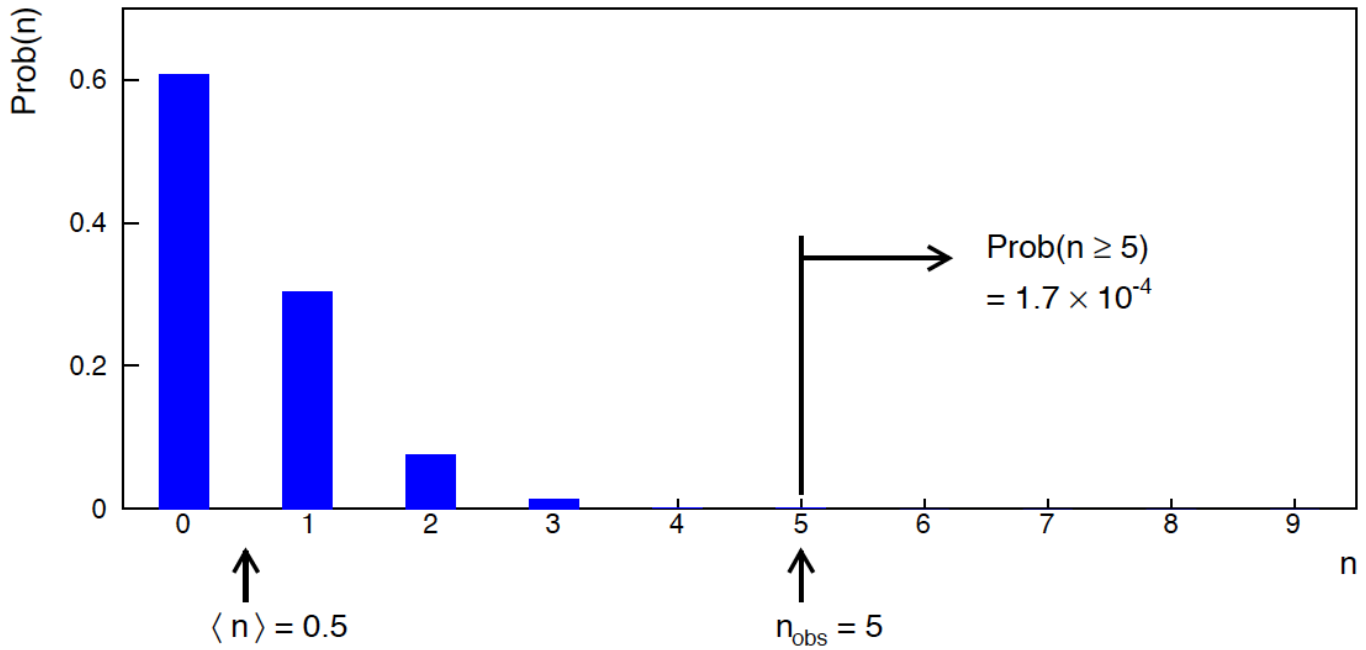
Suppose $b = 0.5$ (known), and we observe $n_{\text{obs}} = 5$.

Should we claim evidence for a new discovery?

Take n itself as the test statistic, p-value for hypothesis $s = 0$ is

$$p\text{-value} = P(n \geq 5; b = 0.5, s = 0)$$

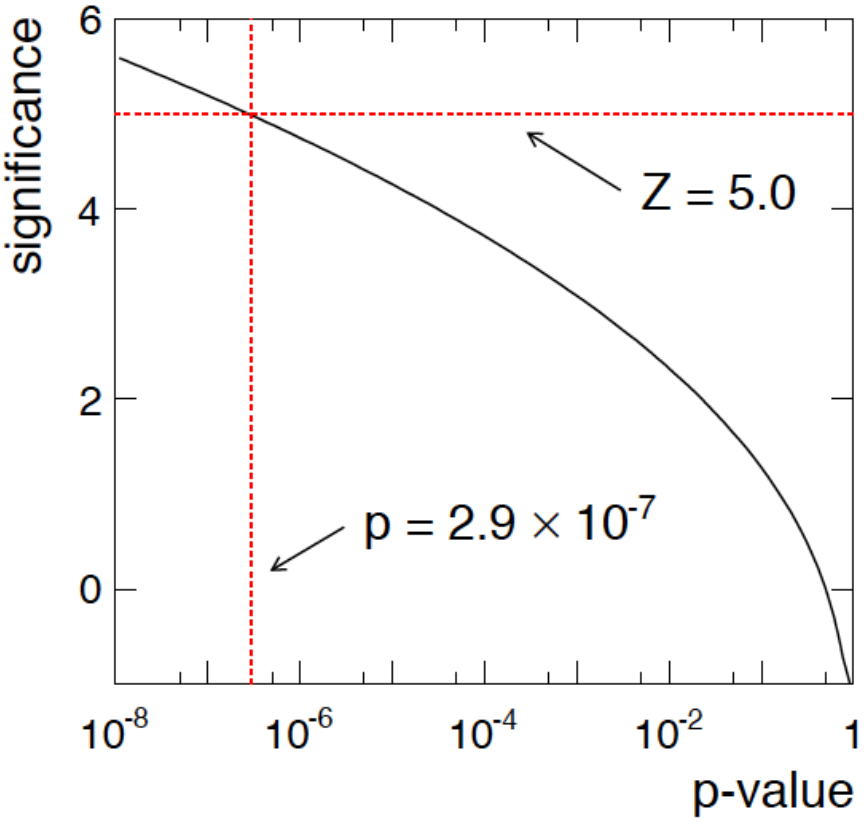
$$= 1.7 \times 10^{-4} \neq P(s = 0)!$$



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" (~multiple testing), etc.

Frequentist upper limit on Poisson parameter

Consider again the case of observing $n \sim \text{Poisson}(s + b)$.

Suppose $b = 4.5$, $n_{\text{obs}} = 5$. Find upper limit on s at 95% CL.

Relevant alternative is $s = 0$ (critical region at low n)

p-value of hypothesized s is $P(n \leq n_{\text{obs}}; s, b)$

Upper limit s_{up} at CL = $1 - \alpha$ found by solving $p_s = \alpha$ for s:

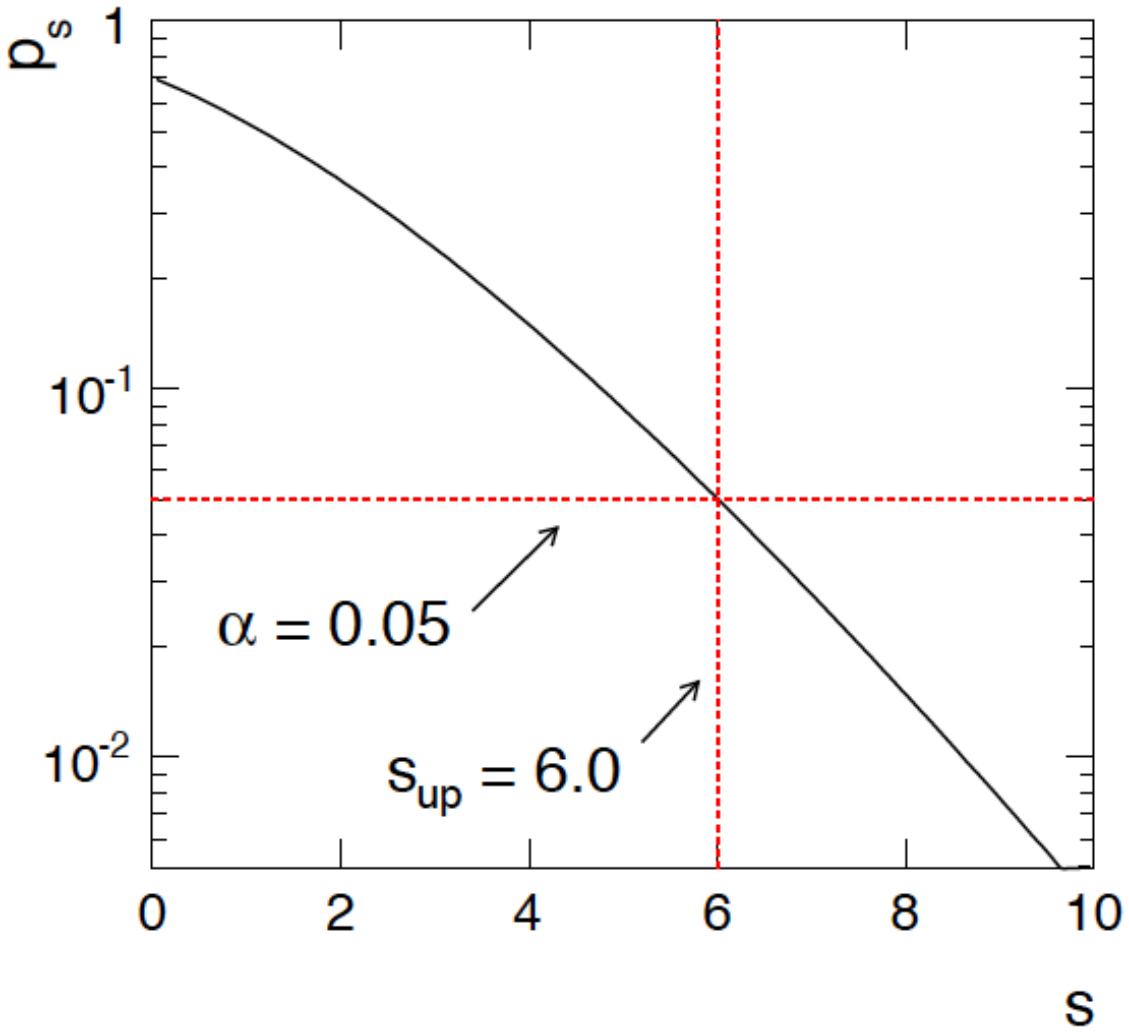
$$\alpha = P(n \leq n_{\text{obs}}; s_{\text{up}}, b) = \sum_{n=0}^{n_{\text{obs}}} \frac{(s_{\text{up}} + b)^n}{n!} e^{-(s_{\text{up}}+b)}$$

$$s_{\text{up}} = \frac{1}{2} F^{-1}_{\chi^2}(1 - \alpha; 2(n_{\text{obs}} + 1)) - b$$

$$= \frac{1}{2} F^{-1}_{\chi^2}(0.95; 2(5 + 1)) - 4.5 = 6.0$$

Frequentist upper limit on Poisson parameter

Upper limit s_{up} at CL = 1 – α found from $p_s = \alpha$.

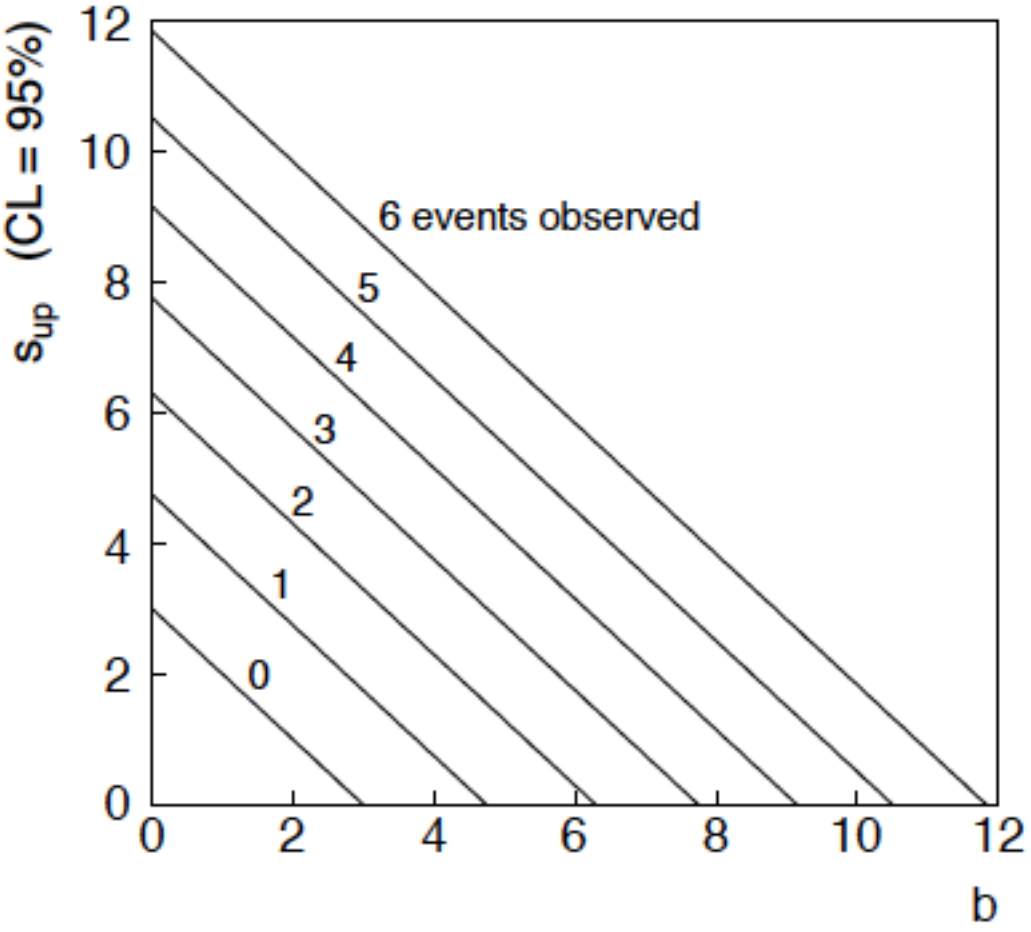


$n_{obs} = 5$,

$b = 4.5$

$n \sim \text{Poisson}(s+b)$: frequentist upper limit on s

For low fluctuation of n formula can give negative result for s_{up}; i.e. confidence interval is empty.



Limits near a physical boundary

Suppose e.g. $b = 2.5$ and we observe $n = 0$.

If we choose CL = 0.9, we find from the formula for s_{up}

$$s_{\text{up}} = -0.197 \quad (\text{CL} = 0.90)$$

Physicist:

We already knew $s \geq 0$ before we started; can't use negative upper limit to report result of expensive experiment!

Statistician:

The interval is designed to cover the true value only 90% of the time — this was clearly not one of those times.

Not uncommon dilemma when testing parameter values for which one has very little experimental sensitivity, e.g., very small s.

Expected limit for $s = 0$

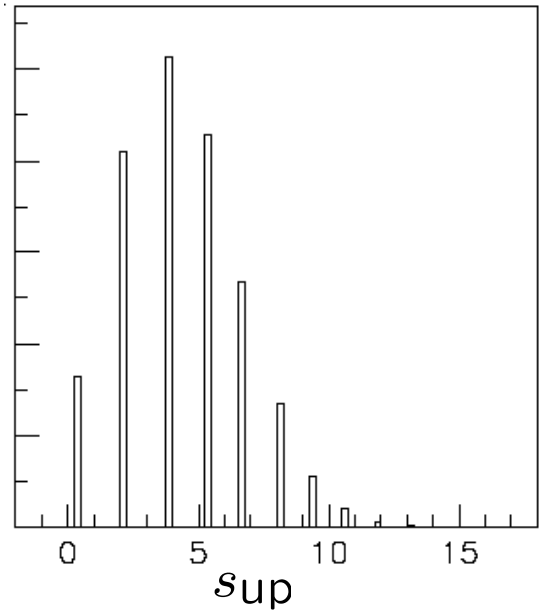
Physicist: I should have used CL = 0.95 — then $s_{up} = 0.496$

Even better: for CL = 0.917923 we get $s_{up} = 10^{-4}$!

Reality check: with $b = 2.5$, typical Poisson fluctuation in n is at least $\sqrt{2.5} = 1.6$. How can the limit be so low?

Look at the mean limit for the no-signal hypothesis ($s = 0$) (sensitivity).

Distribution of 95% CL limits with $b = 2.5$, $s = 0$.
Mean upper limit = 4.44



The Bayesian approach to limits

In Bayesian statistics need to start with 'prior pdf' $\pi(\theta)$, this reflects degree of belief about θ before doing the experiment.

Bayes' theorem tells how our beliefs should be updated in light of the data x:

$$p(\theta|x) = \frac{L(x|\theta)\pi(\theta)}{\int L(x|\theta')\pi(\theta')\,d\theta'} \propto L(x|\theta)\pi(\theta)$$

Integrate posterior pdf $p(\theta|\, x)$ to give interval with any desired probability content.

For e.g. $n \sim$ Poisson(s+b), 95% CL upper limit on s from

$$0.95 = \int_{-\infty}^{s\text{up}} p(s|n)\,ds$$

Bayesian prior for Poisson parameter

Include knowledge that $s \geq 0$ by setting prior $\pi(s) = 0$ for $s < 0$.

Could try to reflect 'prior ignorance' with e.g.

$$\pi(s) = \begin{cases} 1 & s \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Not normalized but this is OK as long as $L(s)$ dies off for large s.

Not invariant under change of parameter — if we had used instead a flat prior for, say, the mass of the Higgs boson, this would imply a non-flat prior for the expected number of Higgs events.

Doesn't really reflect a reasonable degree of belief, but often used as a point of reference;

or viewed as a recipe for producing an interval whose frequentist properties can be studied (coverage will depend on true s).

Bayesian interval with flat prior for s

Solve to find limit s_{up}:

$$s_{\text{up}} = \frac{1}{2} F_{\chi^2}^{-1}[p, 2(n+1)] - b$$

where

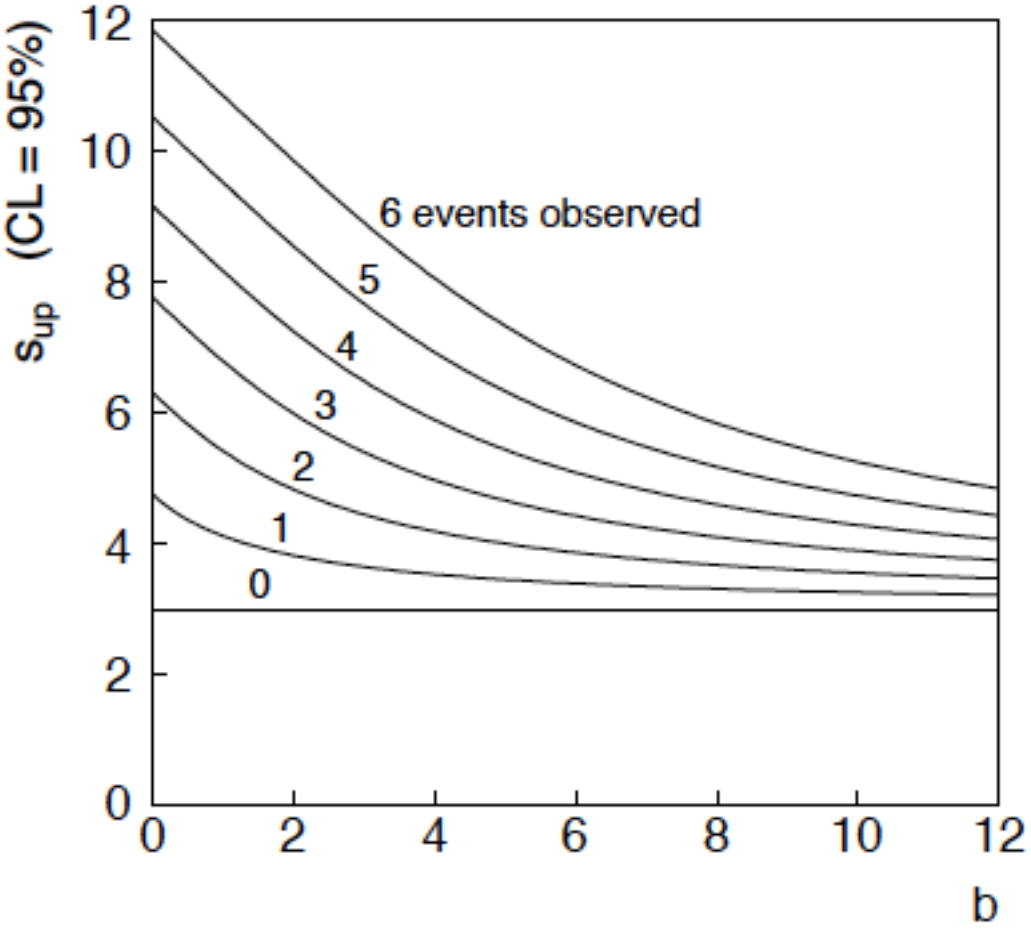
$$p = 1 - \alpha \left(1 - F_{\chi^2}[2b, 2(n+1)]\right)$$

For special case $b = 0$, Bayesian upper limit with flat prior numerically same as one-sided frequentist case ('coincidence').

Bayesian interval with flat prior for s

For $b > 0$ Bayesian limit is everywhere greater than the (one sided) frequentist upper limit.

Never goes negative. Doesn't depend on b if $n = 0$.



Extra slides