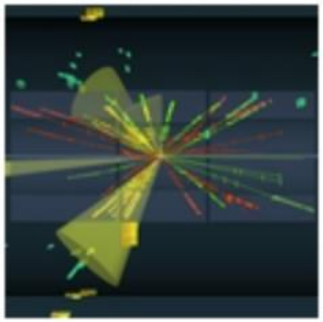
Statistics for Particle Physics
Lecture 1



Taller de Altas Energías
Benasque, Spain
2 September 2024

http://benasque.org/2024tae/

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

Outline

→ Monday 11:30: Introduction
Probability
Hypothesis tests

Monday 12:30: Parameter estimation
Confidence limits
Systematic uncertainties
Experimental sensitivity

Monday 16:30: Tutorial on parameter estimation

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

http://www.pp.rhul.ac.uk/~cowan/stat_course.html

Theory ↔ Statistics ↔ Experiment

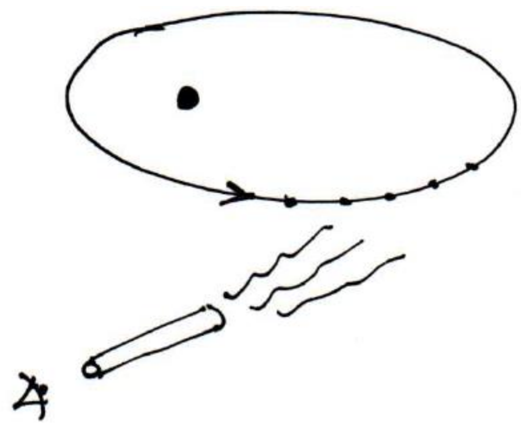
Theory (model, hypothesis):

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

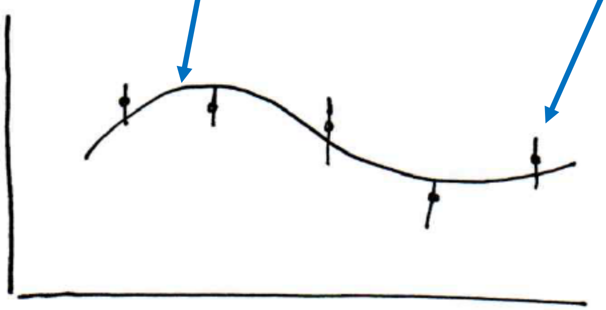
+ response of measurement apparatus

= model prediction

Experiment (observation):



data



Uncertainty enters on many levels

→ quantify with probability

A quick review of probability

Frequentist (A = outcome of repeatable observation)

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

Subjective (A = hypothesis)

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

Conditional probability:

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

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

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

A and B are independent iff:

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

I.e. if A, B independent, then

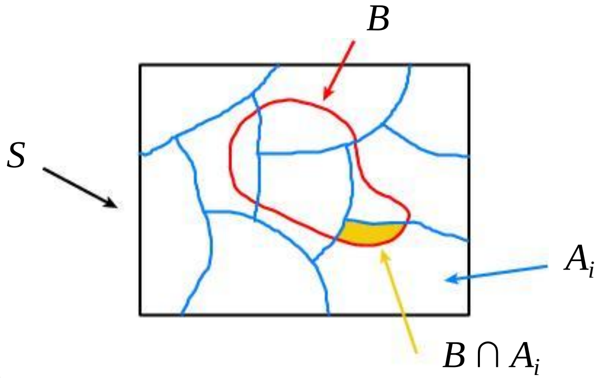
$$P(A|B) = \frac{P(A)P(B)}{P(B)} = P(A)$$

Bayes' theorem

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

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

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



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

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

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

Frequentist Statistics – general philosophy

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

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!)

Hypothesis, likelihood

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

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

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

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

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

Note:

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

Frequentist hypothesis tests

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

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

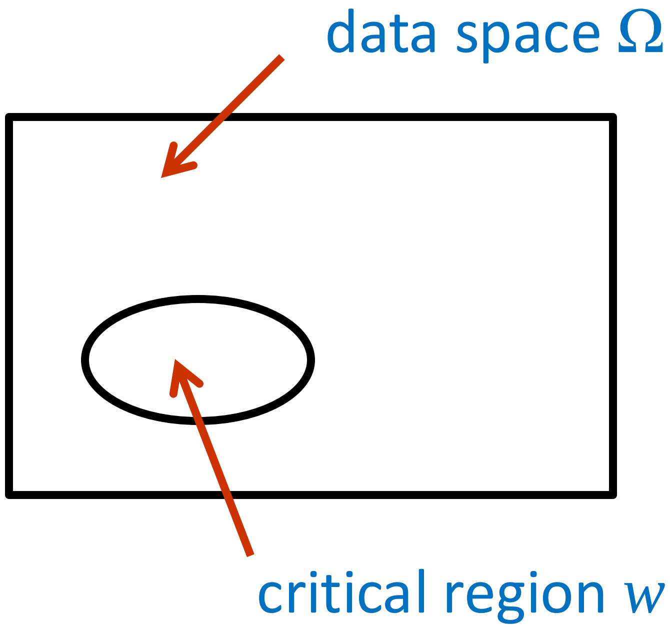
A test of H_0 is defined by specifying a critical region w of the data space such that there is no more than some (small) probability α, assuming H_0 is correct, to observe the data there, i.e.,

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

Need inequality if data are discrete.

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

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

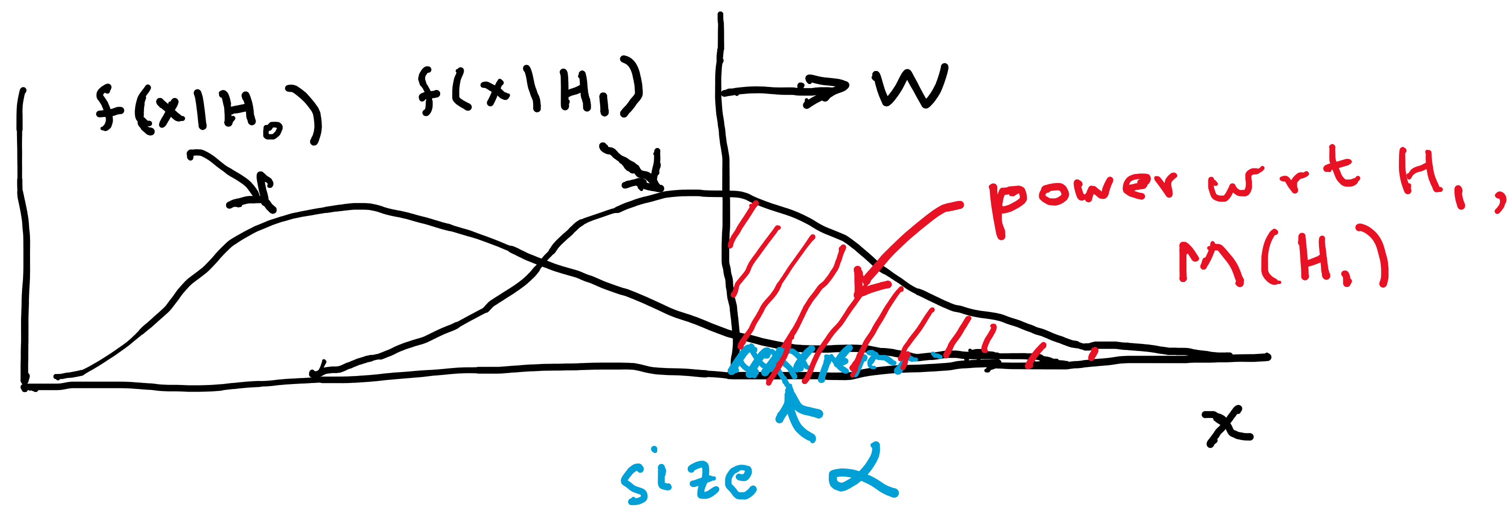


Definition of a test (2)

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

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

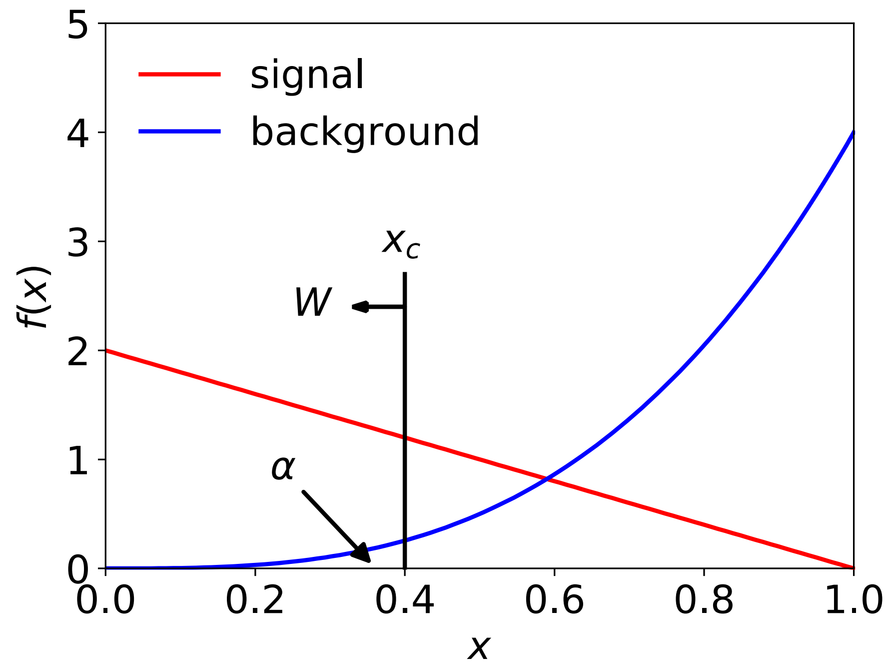
Example of a test for classification

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

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

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

with $0 \le x \le 1$.



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:

$$\alpha = P(x \le x_{\rm c}|{\rm b}) = \int_0^{x_{\rm c}} f(x|b)\,dx = \left.\frac{4x^4}{4}\right|_0^{x_{\rm c}} = x_{\rm c}^4$$

and therefore $x_{\rm c} = \alpha^{1/4} = 0.1$

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

$$M = P(x \le x_{\rm c}|{\rm s}) = \int_0^{x_{\rm c}} f(x|{\rm s})\,dx = 2x_{\rm c} - x_{\rm c}^2 = 0.19$$

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:

$$P(\mathrm{s}|x \leq x_{\mathrm{c}}) = \frac{P(x \leq x_{\mathrm{c}}|\mathrm{s})\pi_{\mathrm{s}}}{P(x \leq x_{\mathrm{c}}|\mathrm{s})\pi_{\mathrm{s}} + P(x \leq x_{\mathrm{c}}|\mathrm{b})\pi_{\mathrm{b}}}$$

$$= 0.655$$

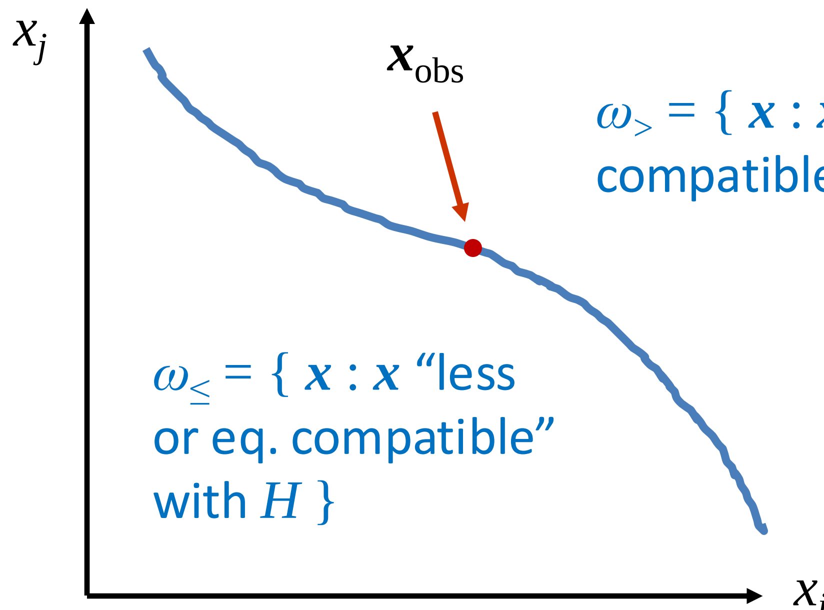
Testing significance / goodness-of-fit

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

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

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 $\boldsymbol{x}_{\mathrm{obs}}$. (Not unique!)



p-values

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

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

= 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.

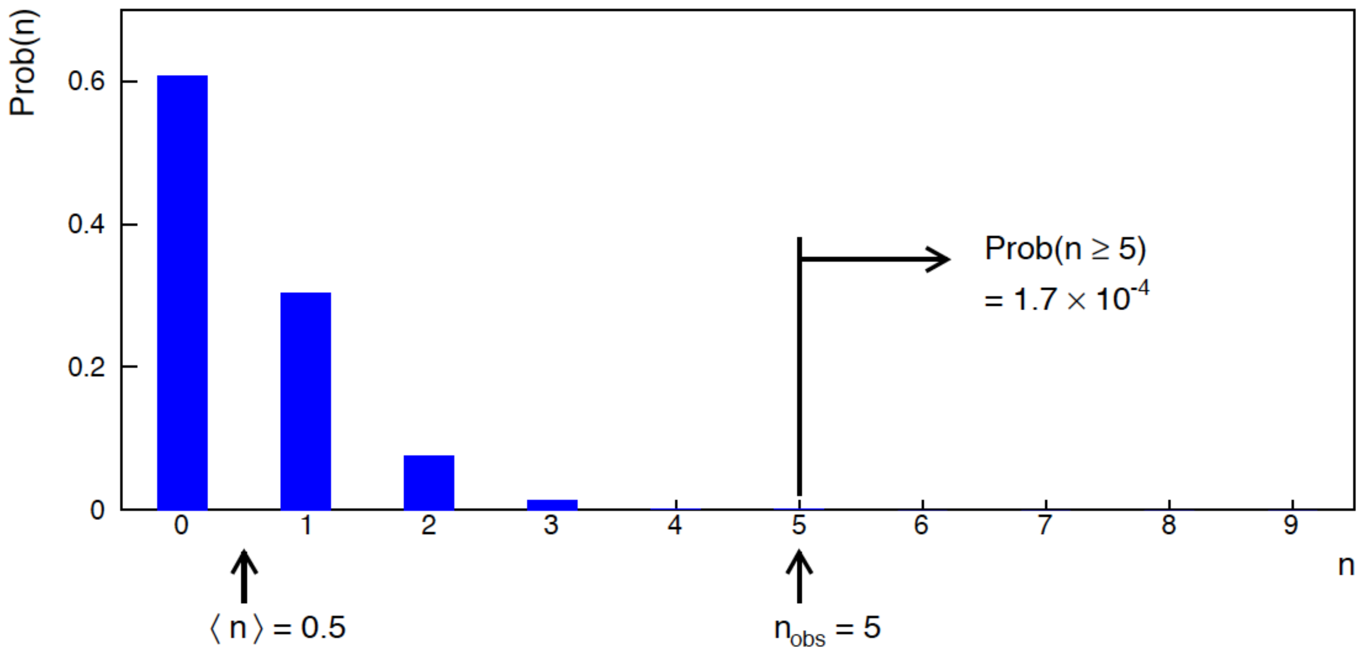
Poisson counting experiment: discovery p-value

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

Should we claim evidence for a new discovery?

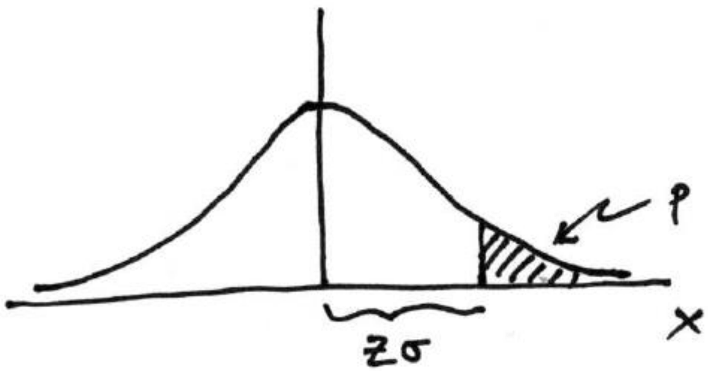
Give p-value for hypothesis $s = 0$, suppose relevant alt. is $s > 0$.

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



Significance from p-value

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



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

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

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

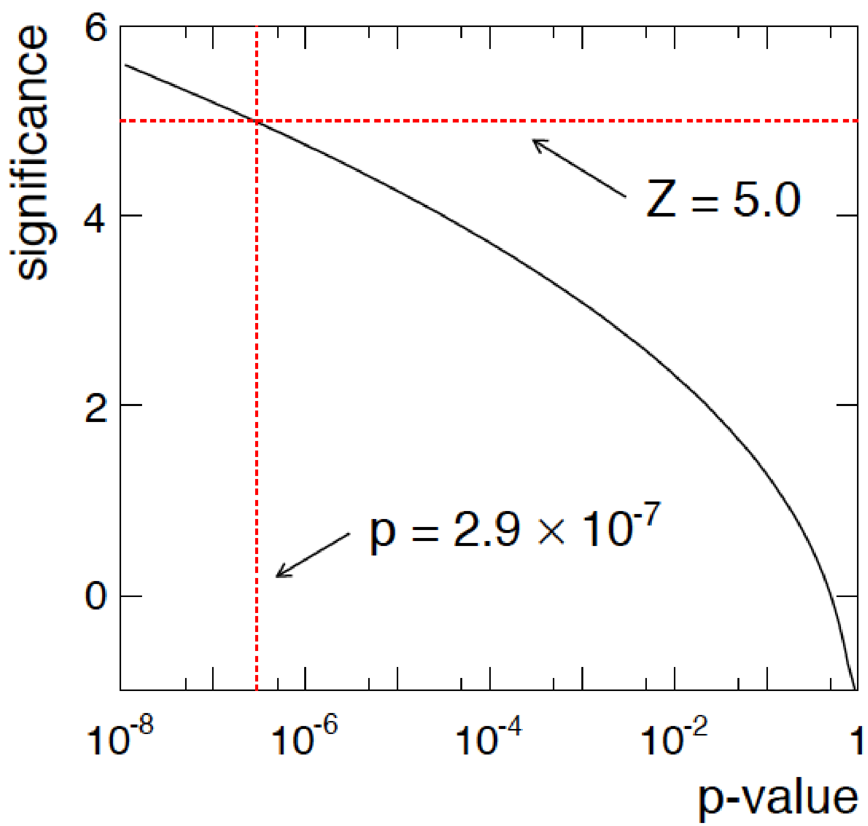
```
in python (scipy.stats):
p = 1 - norm.cdf(Z) = norm.sf(Z)
Z = norm.ppf(1-p)
```

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

Poisson counting experiment: discovery significance

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

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



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

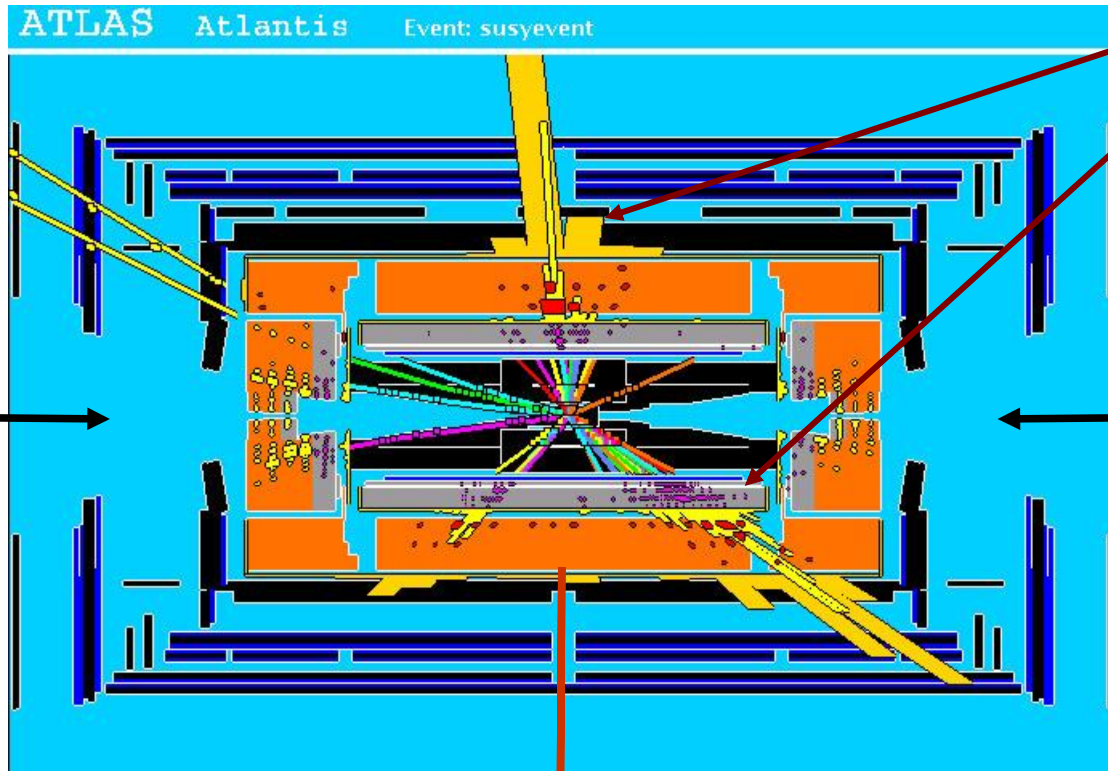
Particle Physics context for a hypothesis test

A simulated SUSY event ("signal"):

high p_{T} muons

high p_{T} jets of hadrons

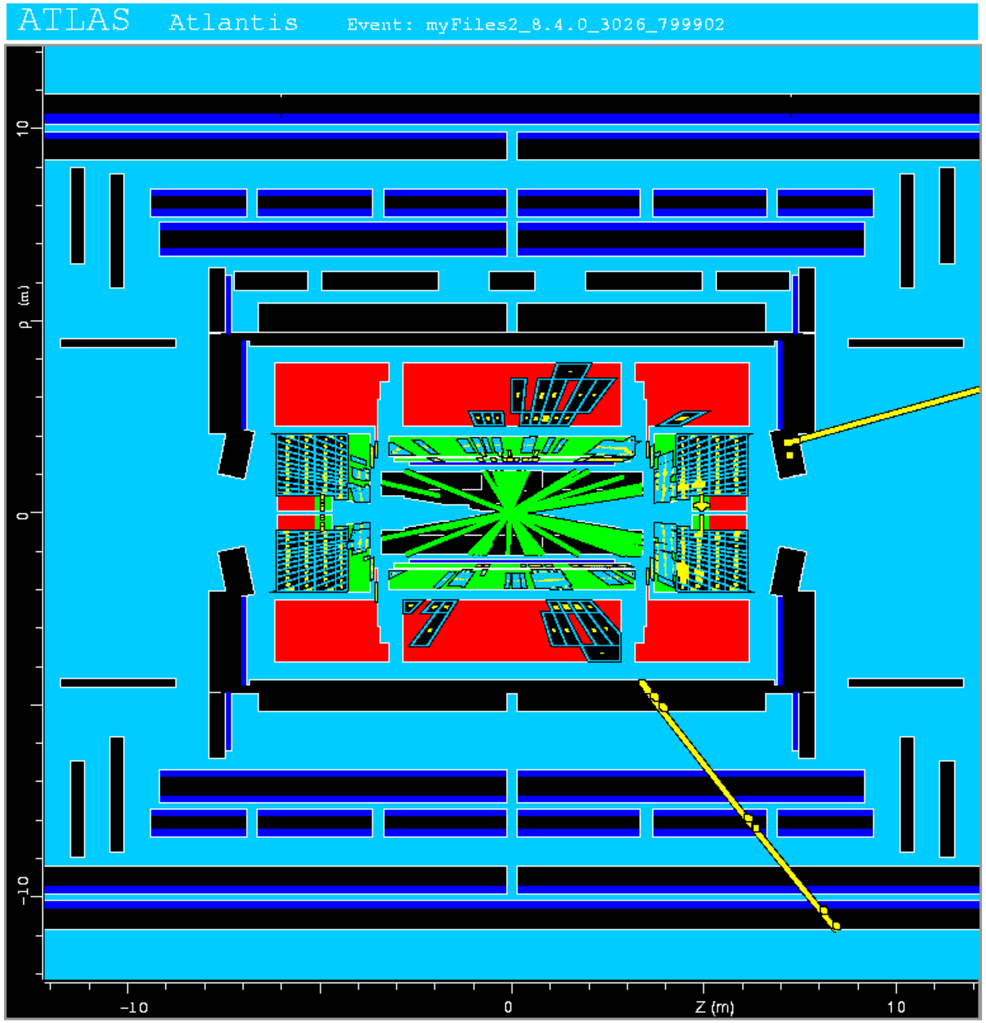
p



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 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 = ...

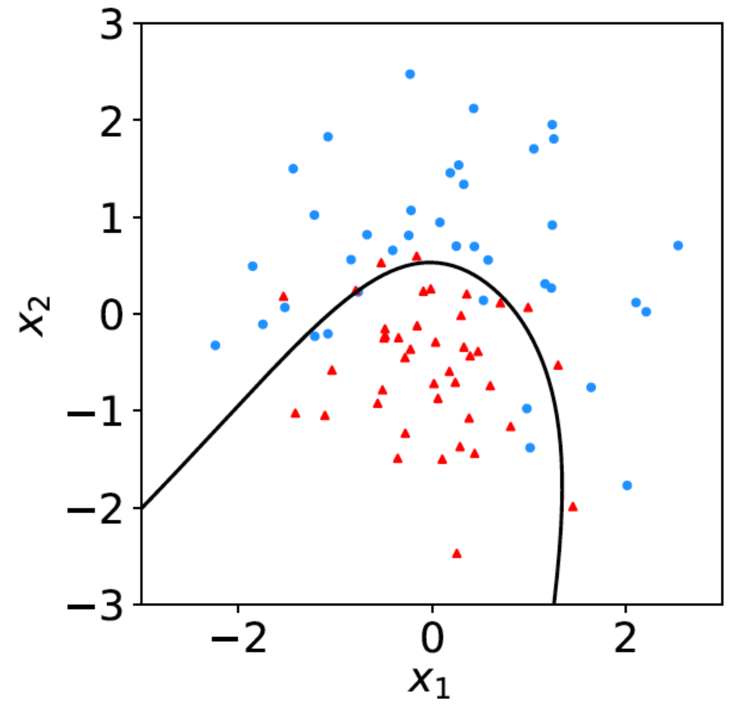
The real events don't come with true class labels, but computer-simulated events do. So we can have a set of simulated events that consist of a feature vector $\boldsymbol{x}$ and true class label y (0 for background, 1 for signal):

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

The simulated events are called "training data".

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: $y = 1$).



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.

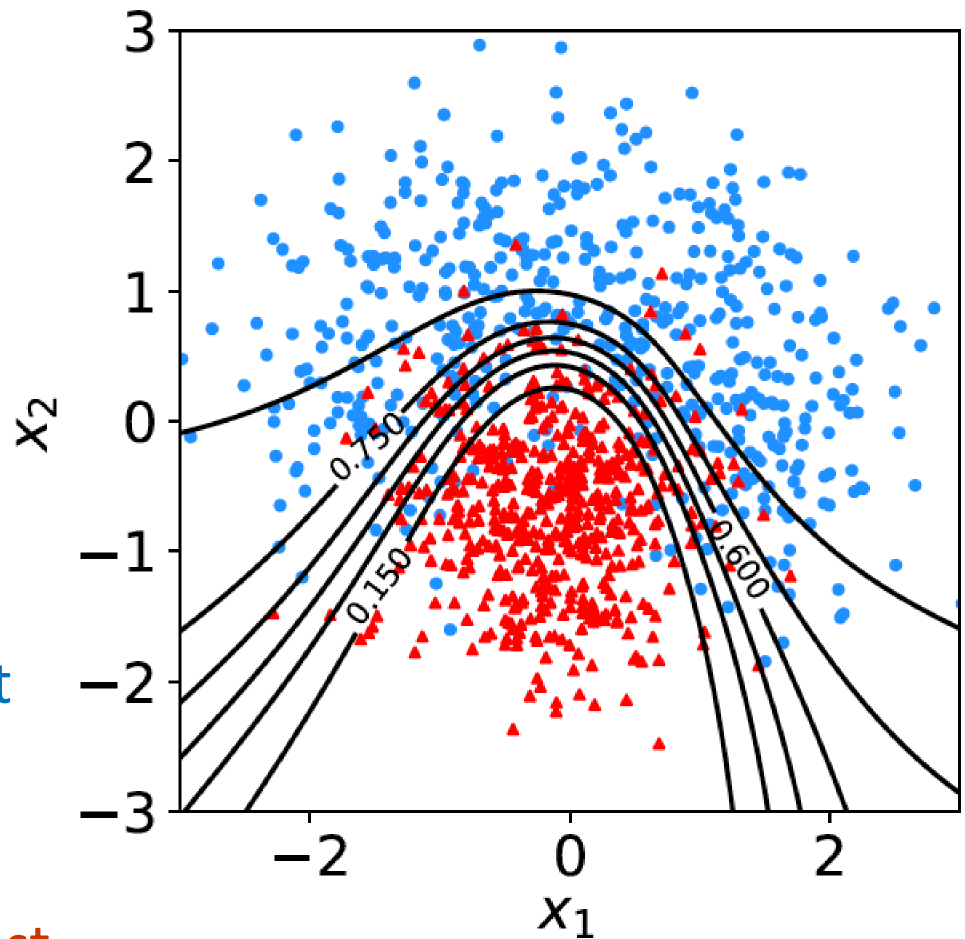
Decision function, test statistic

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

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

scalar function

constant

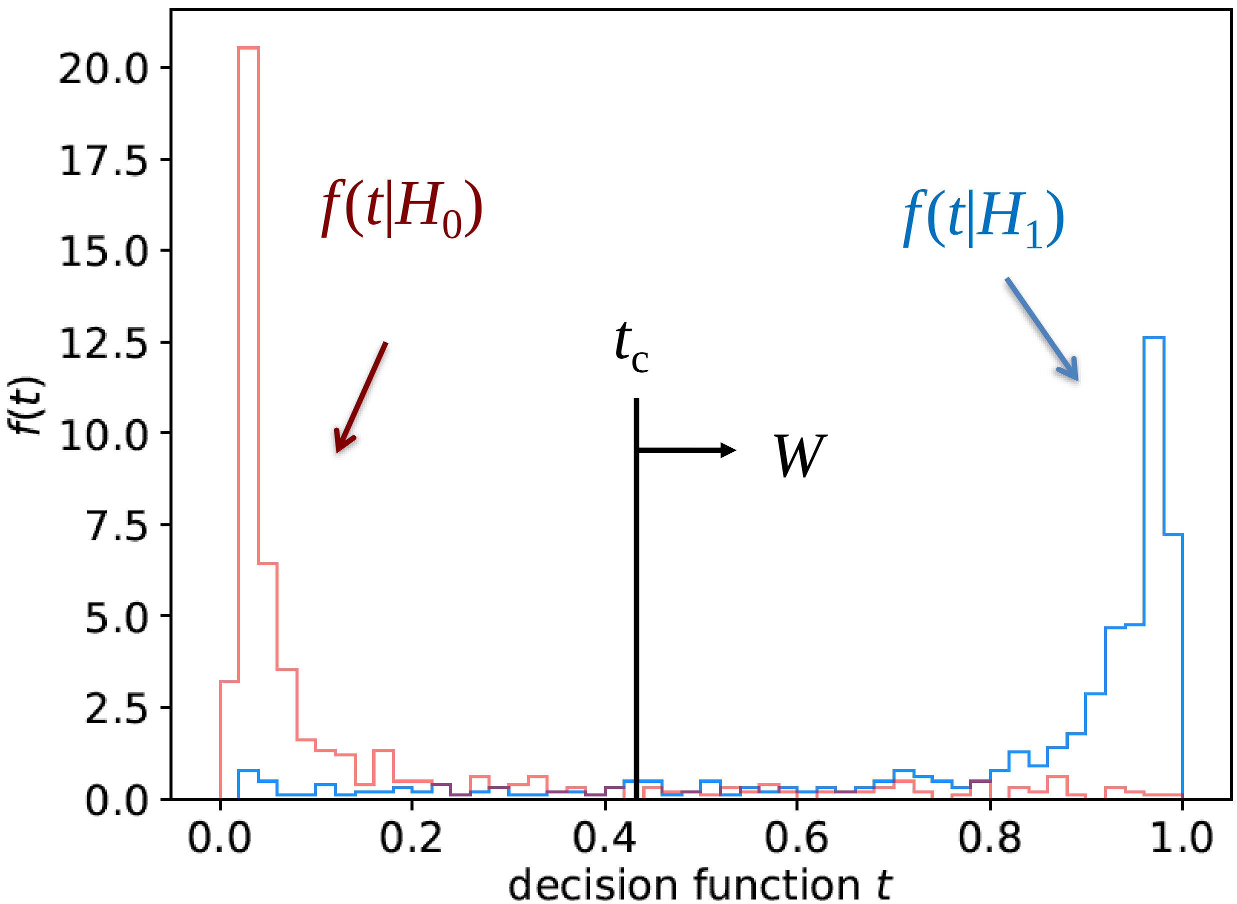


Different values of the constant t_c result in a family of surfaces.

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.

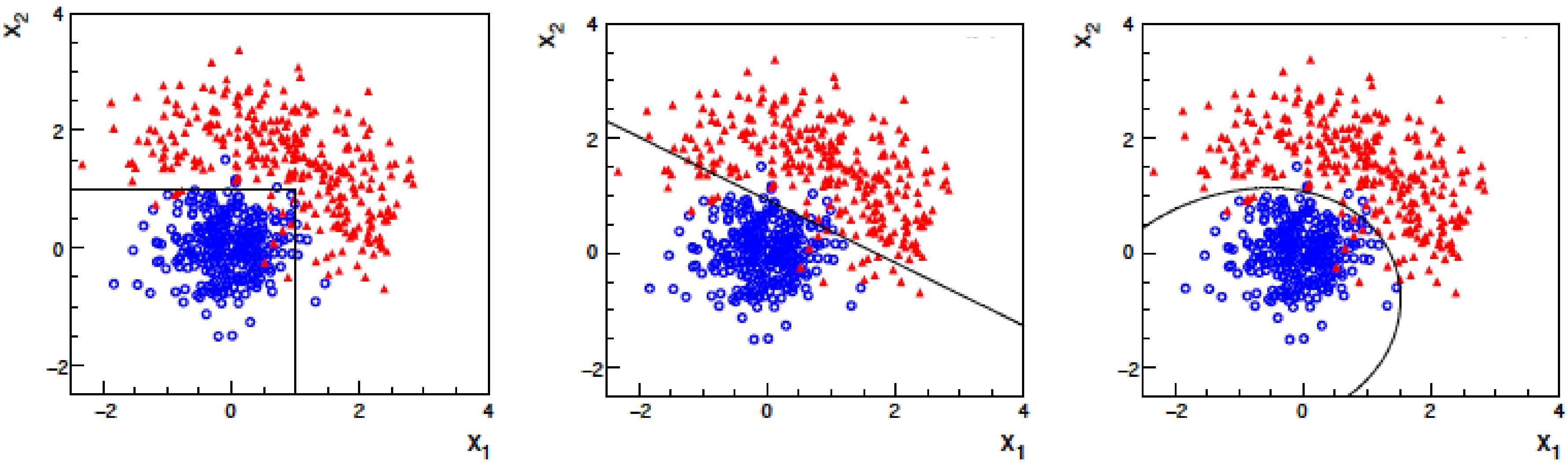


Types of decision boundaries

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', 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 $\boldsymbol{x}$ in the critical region W

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

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

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

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

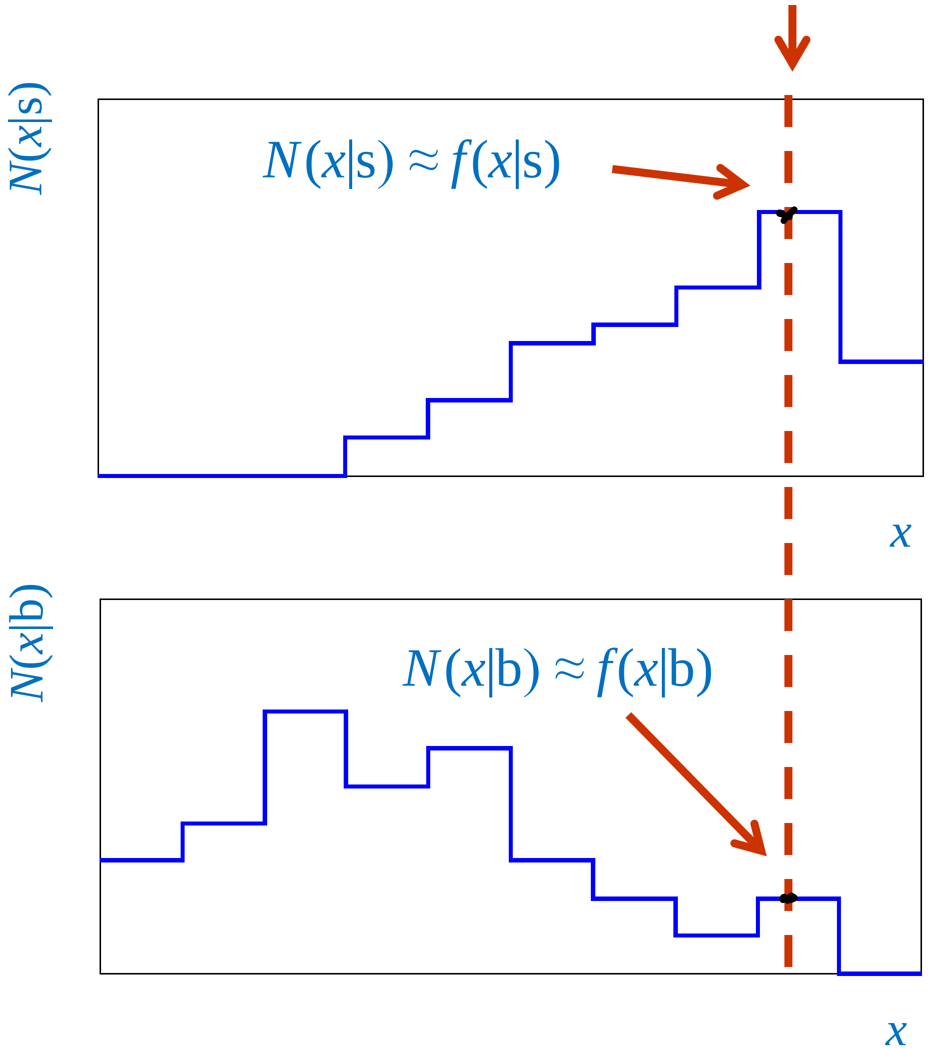
generate $\boldsymbol{x} \sim f(\boldsymbol{x}|\mathrm{b}) \quad \rightarrow \quad \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$

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

- 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.

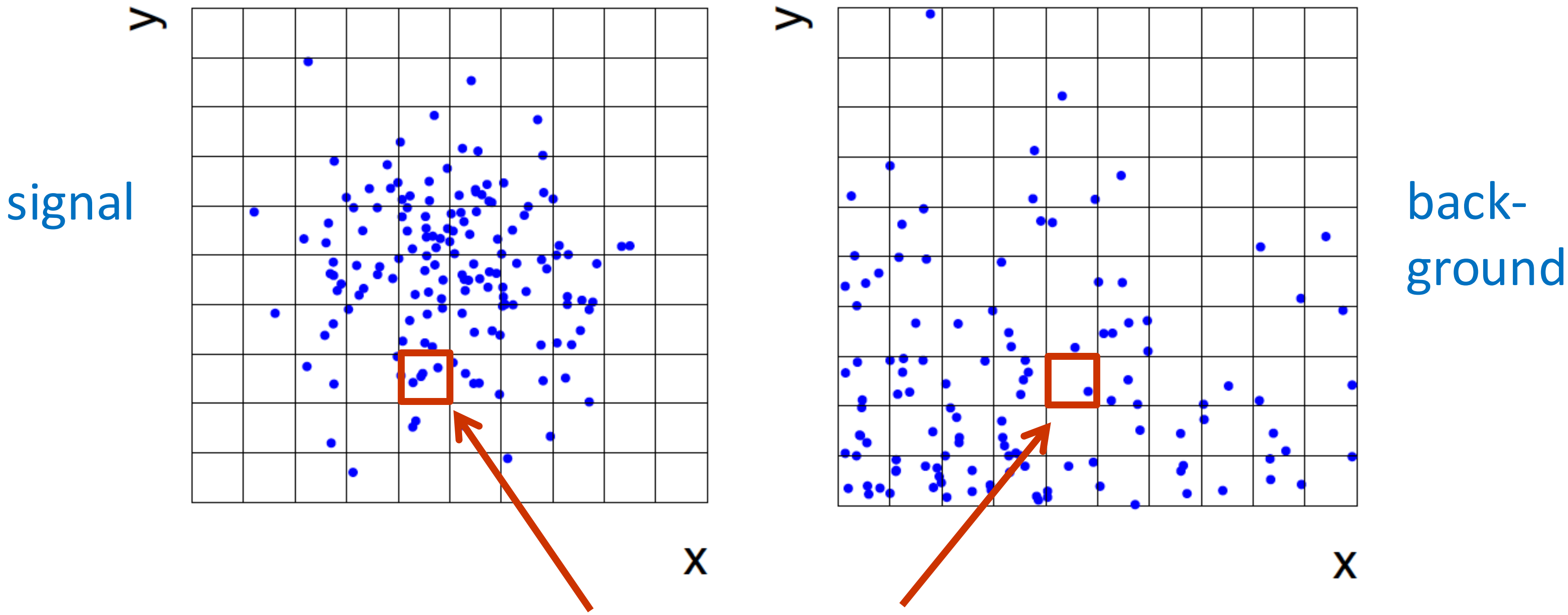
Use (normalized) histogram values to approximate LR:

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

Can work well for single variable.

Approximate LR from 2D-histograms

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



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

$\rightarrow$ 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
- (Deep) Neural Networks
- Kernel density methods
- Support Vector Machines
- Decision trees
  - Boosting
  - Bagging

More in the lectures by Arantza Oyanguren

Extra slides

Some statistics books, papers, etc.

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

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

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

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

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

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

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

S. Navas et al. (Particle Data Group), Phys. Rev. D 110, 030001 (2024); **`pdg.lbl.gov`** sections on probability, statistics, MC.

Some distributions

Distribution/pdf	Example use in Particle Physics
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 $\dfrac{N!}{n!(N-n)!}$

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

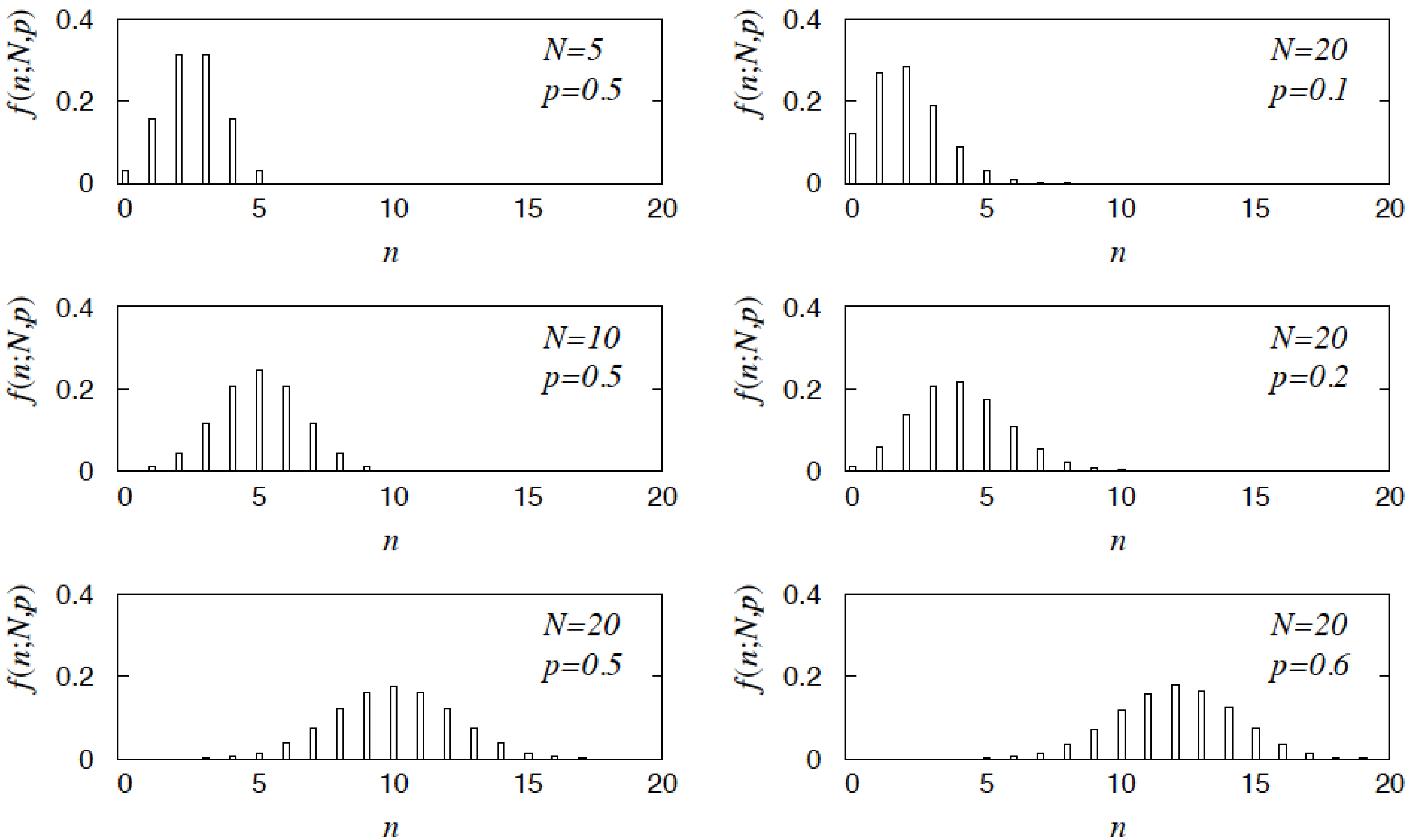
For the expectation value and variance we find:

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

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

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

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

For N trials we want the probability to obtain:

n_1 of outcome 1,
n_2 of outcome 2,
$\vdots$
n_m of outcome m.

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

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

Multinomial distribution (2)

Now consider outcome i as 'success', all others as 'failure'.

$\rightarrow$ 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: $\vec{n} = (n_1, \ldots, n_m)$ represents a histogram

with m bins, N total entries, all entries independent.

Poisson distribution

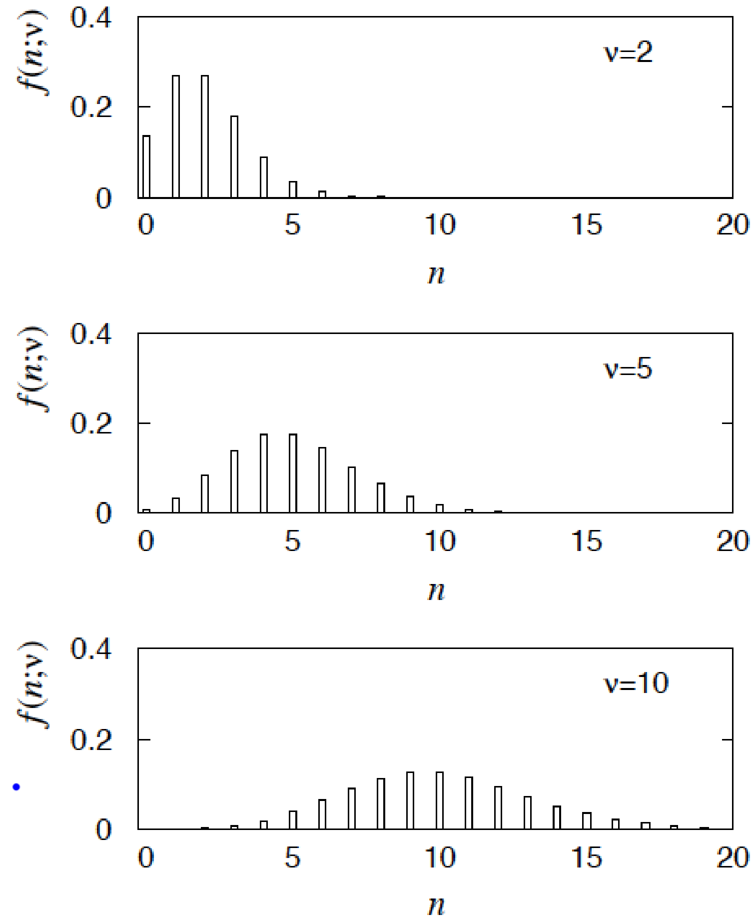
Consider binomial n in the limit

$$N \to \infty, \qquad p \to 0, \qquad E[n] = Np \to \nu \ .$$

$\rightarrow$ n follows the Poisson distribution:

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

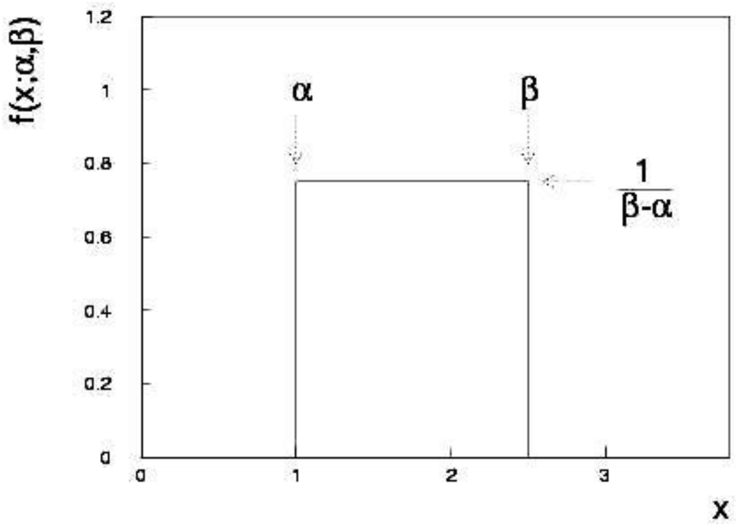
$$E[n] = \nu \, , \qquad V[n] = \nu \ .$$



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

Uniform distribution

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



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

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

Notation: x follows a uniform distribution between α and β

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

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

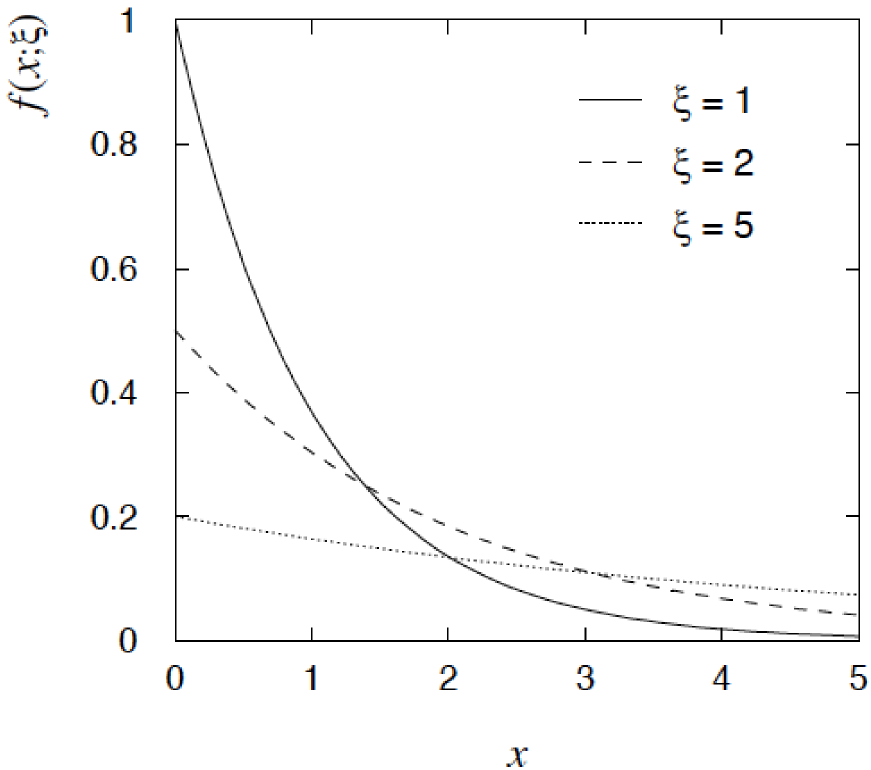
Exponential distribution

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

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

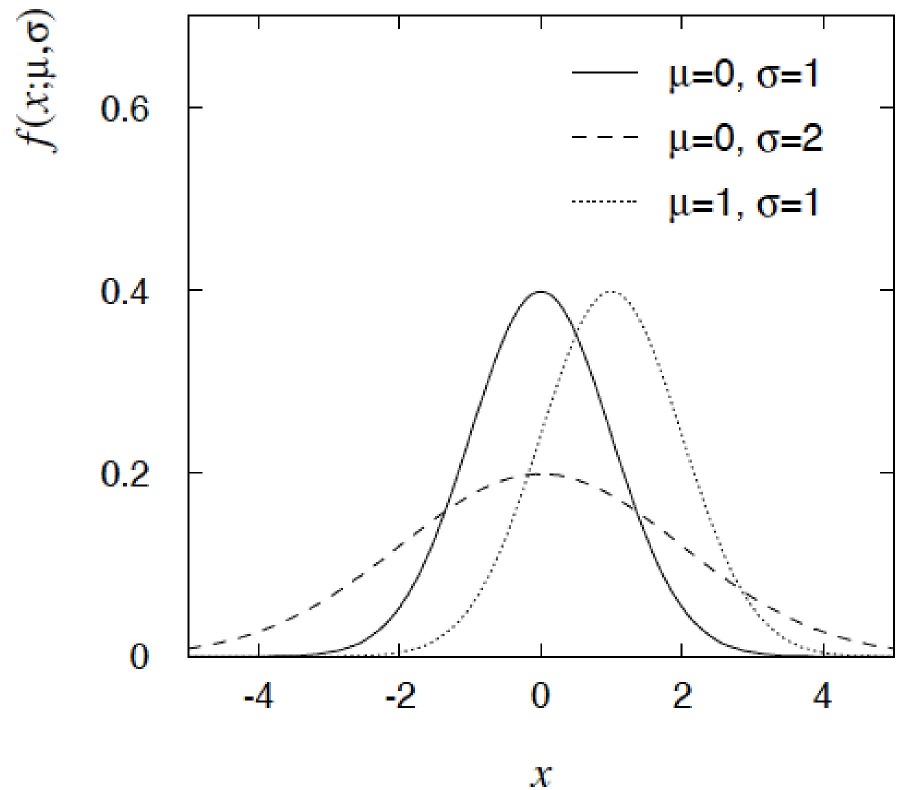
The Gaussian (normal) pdf for a continuous r.v. x is defined by:

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

$$E[x] = \mu$$

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

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 (μ = 0. σ = 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 $\vec{x} = (x_1, \ldots, x_n)$:

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

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

$$E[x_i] = \mu_i, \ , \qquad \mathrm{cov}[x_i, x_j] = V_{ij} \ .$$

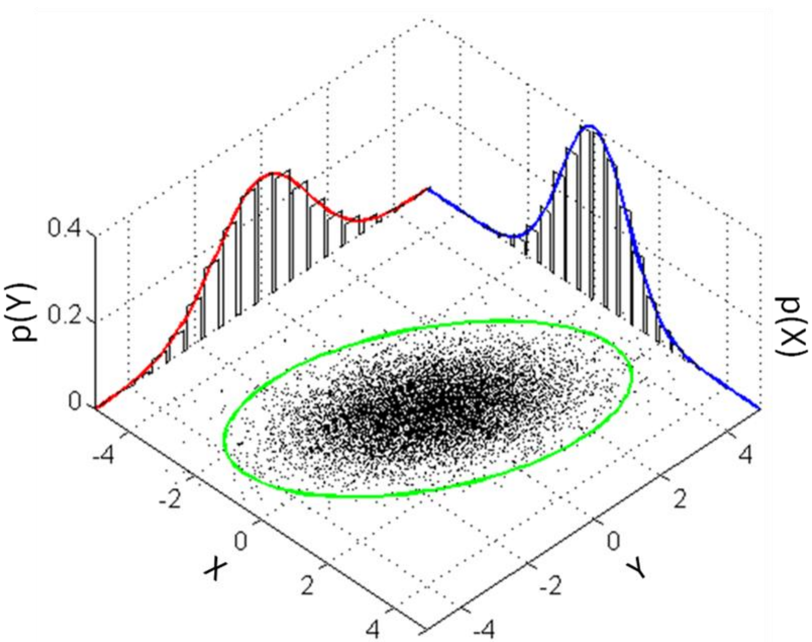
Marginal pdf of each x_i is Gaussian with mean μ_i, standard deviation $\sigma_i = \sqrt{V_{ii}}$.

Two-dimensional Gaussian distribution

$$f(x_1, x_2, ; \mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$$

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

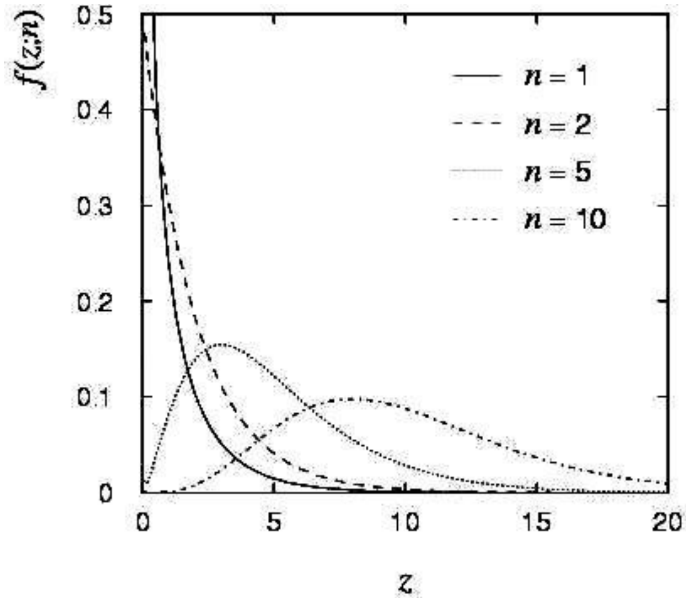
where $\rho = \mathrm{cov}[x_1, x_2]/(\sigma_1 \sigma_2)$
is the correlation coefficient.



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, ... = number of 'degrees of freedom' (dof)

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

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

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

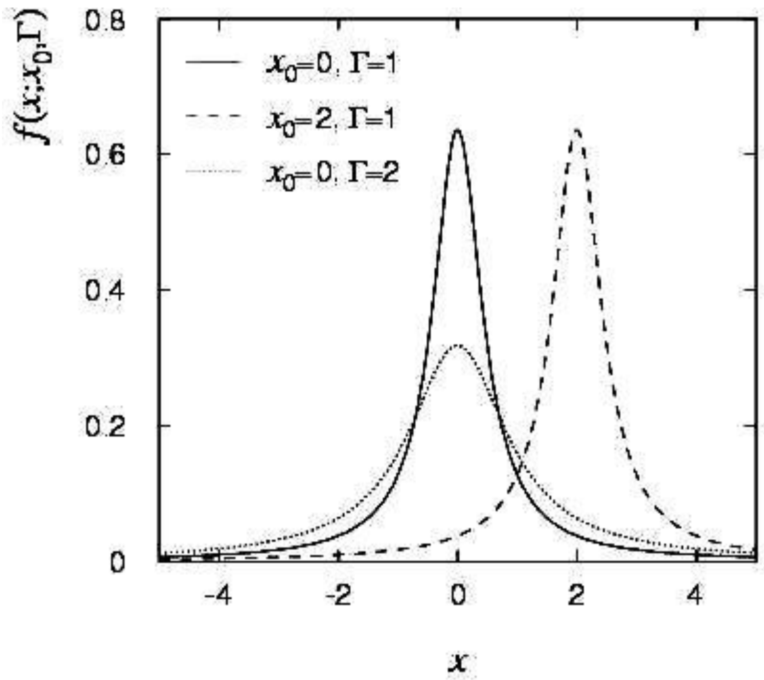
Example: goodness-of-fit test variable especially in conjunction with method of least squares.

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, $V[x] \to \infty$.

x_0 = mode (most probable value)

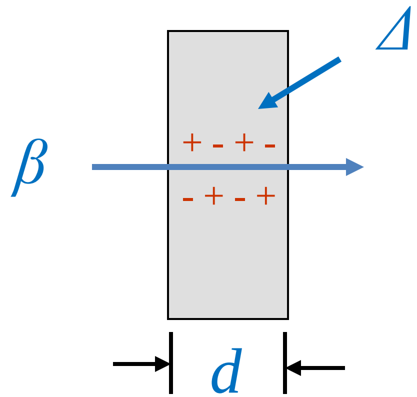
Γ = full width at half maximum

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

Γ = 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:



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

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

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

$$\xi = \frac{2\pi N_{\mathrm{A}} e^4 z^2 \rho \sum Z}{m_{\mathrm{e}} c^2 \sum A}\frac{d}{\beta^2} \,, \qquad \epsilon' = \frac{I^2 \exp \beta^2}{2 m_{\mathrm{e}} c^2 \beta^2 \gamma^2} \,.$$

L. Landau, J. Phys. USSR **8** (1944) 201; see also
W. Allison and J. Cobb, Ann. Rev. Nucl. Part. Sci. **30** (1980) 253.

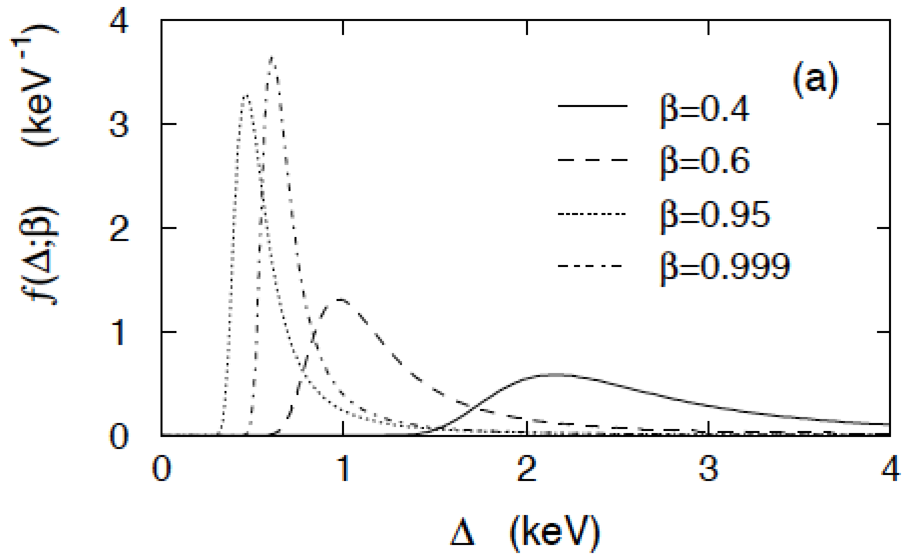
Landau distribution (2)

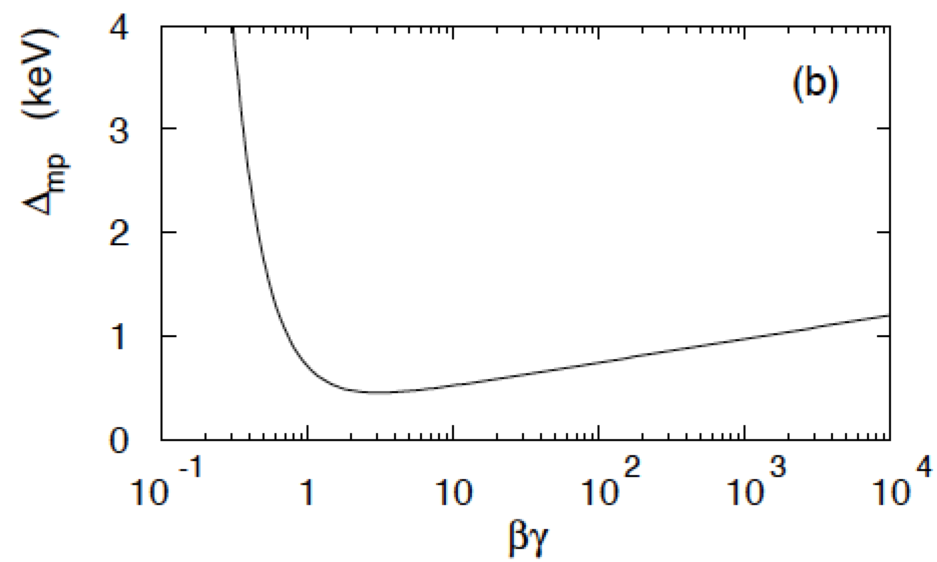
Long ‘Landau tail’

→ all moments ∞

Mode (most probable value) sensitive to β ,

→ particle i.d.



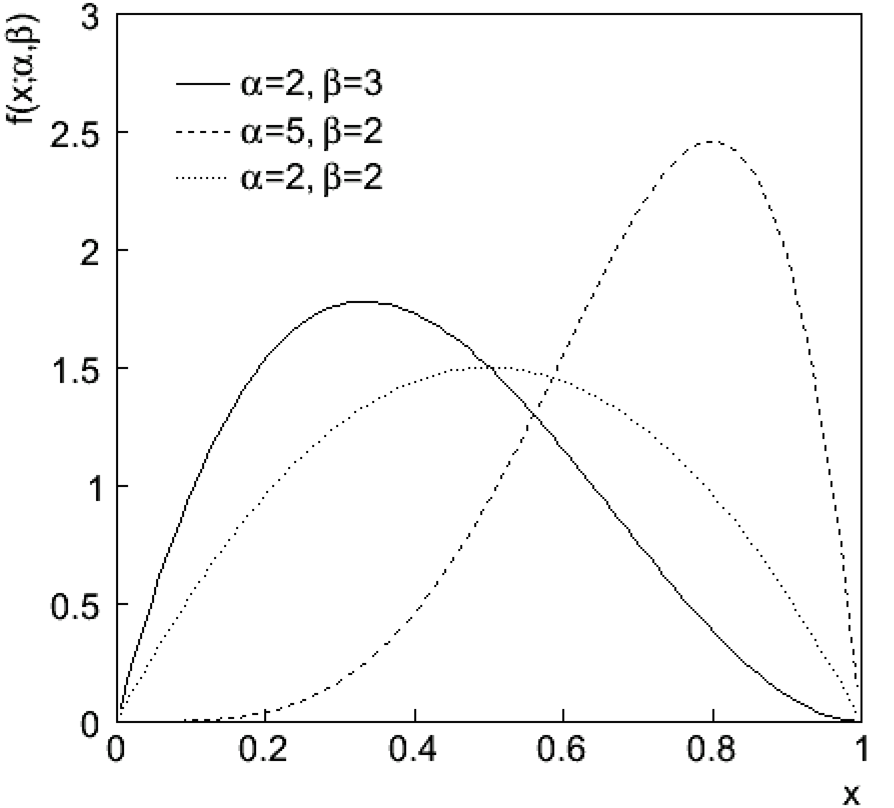


Beta distribution

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

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

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



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

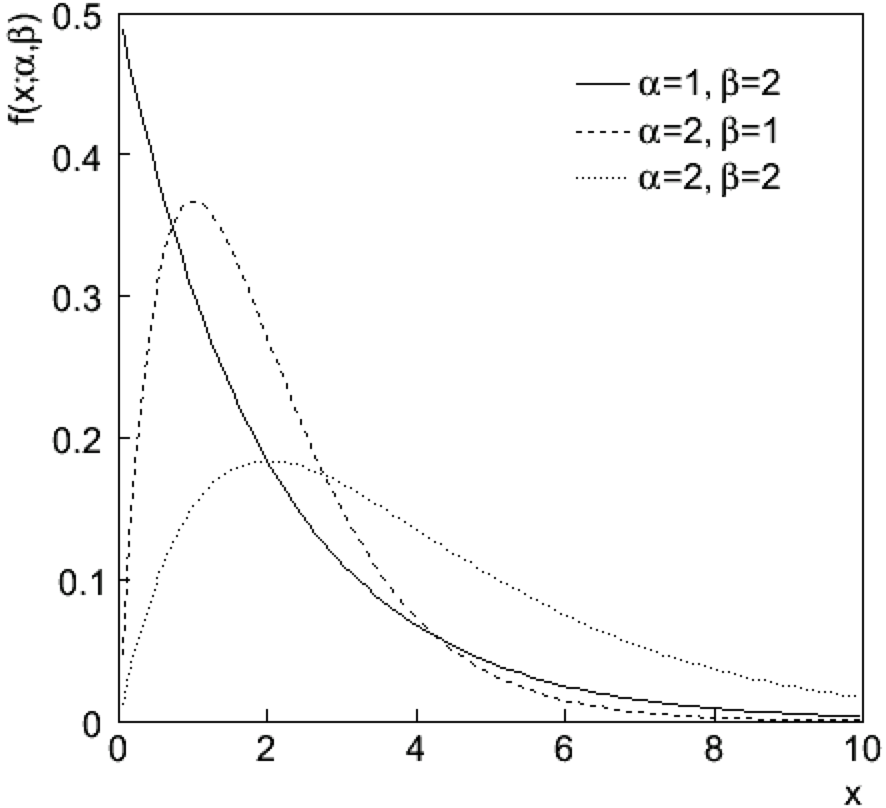
$y = a_0 + a_1 x, \quad a_0 \le y \le a_0 + a_1$

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 nth event in Poisson process ~ Gamma

Student's *t* distribution

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

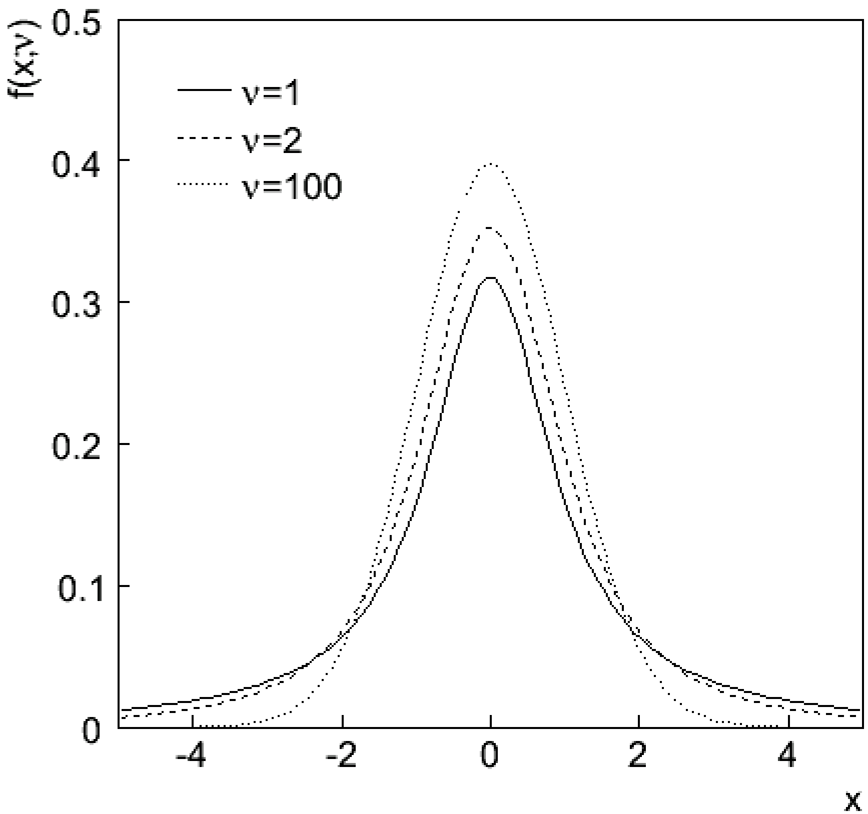
$$E[x] = 0 \quad (\nu > 1)$$

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

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

$\nu = 1$ gives Cauchy,

$\nu \longrightarrow \infty$ 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 $\nu = n$.

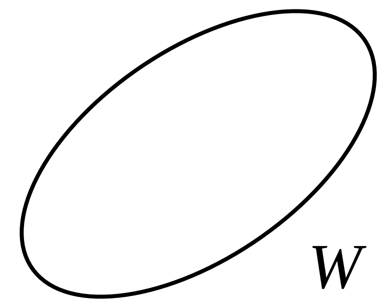
This arises in problems where one forms the ratio of a sample mean to the sample standard deviation of Gaussian r.v.s.

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:



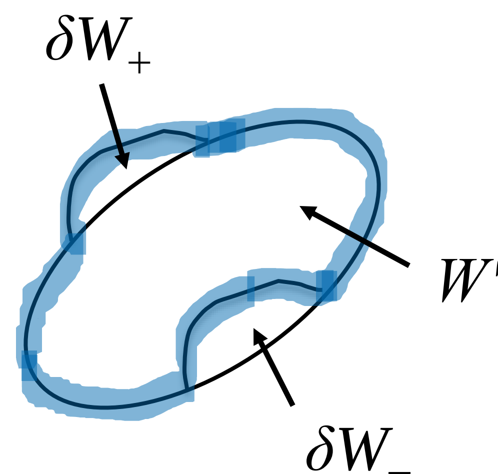
$$P(\mathbf{x}|H_1)/P(\mathbf{x}|H_0) \geq c_\alpha \quad \text{for all } \mathbf{x} \text{ in } W,$$

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

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.,

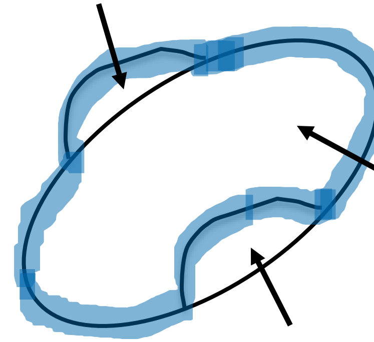


$$W \rightarrow W' = W + \delta W_+ - \delta W_-$$

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

Proof of Neyman-Pearson Lemma (2)

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



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

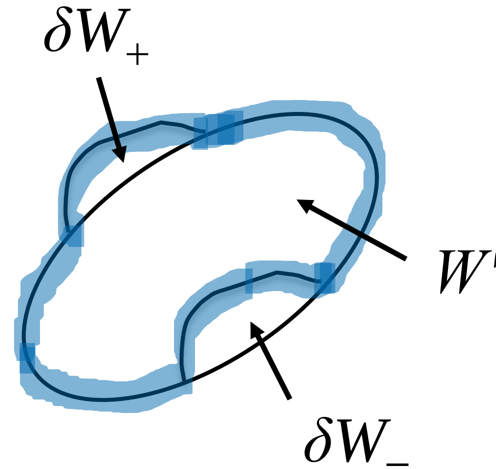
$$P(\mathbf{x} \in \delta W_- | H_1) \geq P(\mathbf{x} \in \delta W_- | H_0) c_\alpha$$

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

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

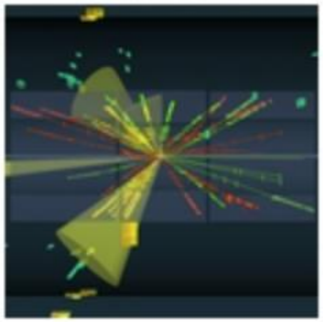
Proof of Neyman-Pearson Lemma (4)

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.

Statistics for Particle Physics
Lecture 2



CENTRO DE CIENCIAS DE BENASQUE PEDRO PASCUAL

Taller de Altas Energías
Benasque, Spain
2 September 2024

http://benasque.org/2024tae/

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

Monday 9:00:	Introduction
	Probability
	Hypothesis tests
→ Tuesday 9:00:	Parameter estimation
	Confidence limits
	Systematic uncertainties
	Experimental sensitivity
Tuesday 15:30:	Tutorial on parameter estimation

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

http://www.pp.rhul.ac.uk/~cowan/stat_course.html

Parameter estimation

The parameters of a pdf are any constants that characterize it,

$$f(x;\theta) = \frac{1}{\theta} e^{-x/\theta}$$

r.v. parameter

i.e., θ indexes a set of hypotheses.

Suppose we have a sample of observed values: $\boldsymbol{x} = (x_1, \ldots, x_n)$

We want to find some function of the data to estimate the parameter(s):

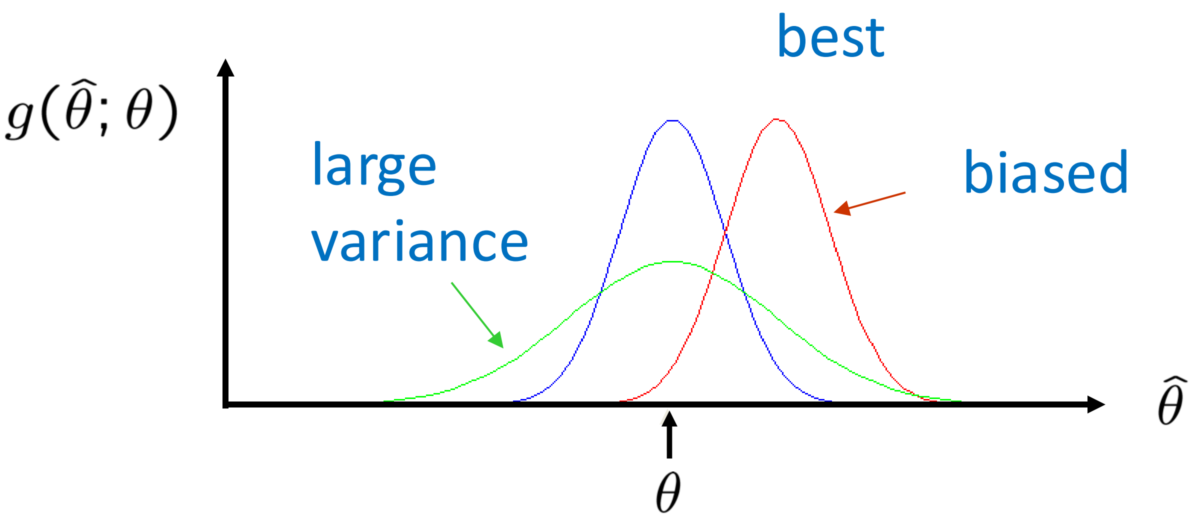
$$\boxed{\widehat{\theta}(\vec{x})}$$

$\leftarrow$ estimator written with a hat

Sometimes we say 'estimator' for the function of $x_1, \ldots, x_n$;
'estimate' for the value of the estimator with a particular data set.

Properties of estimators

If we were to repeat the entire measurement, the estimates from each would follow a pdf:



We want small (or zero) bias (systematic error): $b = E[\hat{\theta}] - \theta$

$\rightarrow$ average of repeated measurements should tend to true value.

And we want a small variance (statistical error): $V[\hat{\theta}]$

$\rightarrow$ small bias & variance are in general conflicting criteria

The likelihood function for i.i.d.* data

* i.i.d. = independent and identically distributed

Consider n independent observations of x: $x_1, \ldots, x_n$, where x follows $f(x; \theta)$. The joint pdf for the whole data sample is:

$$f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

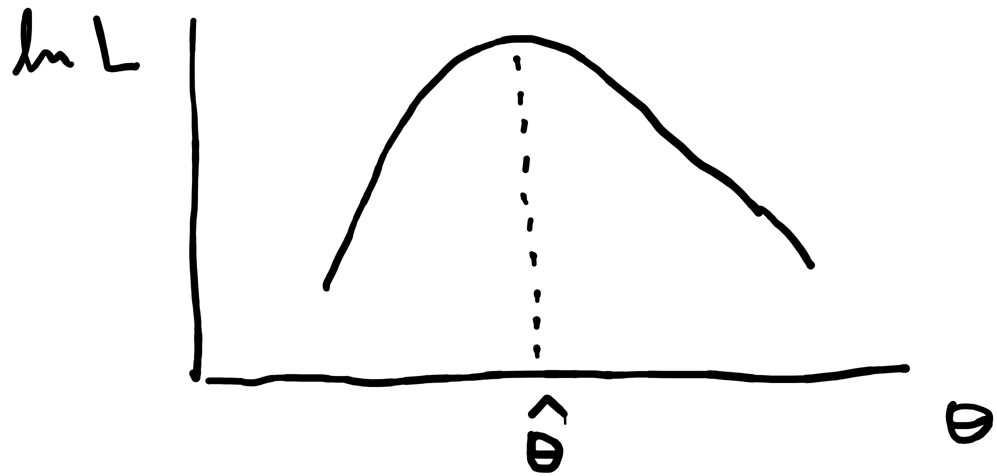
In this case the likelihood function is

$$L(\vec{\theta}) = \prod_{i=1}^{n} f(x_i; \vec{\theta}) \qquad (x_i \text{ constant})$$

Maximum Likelihood Estimators (MLEs)

We *define* the maximum likelihood estimators or MLEs to be the parameter values for which the likelihood is maximum.

Maximizing L equivalent to maximizing $\log L$



$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\boldsymbol{\theta})$$

Could have multiple maxima (take highest).

MLEs not guaranteed to have any ‘optimal’ properties, (but in practice they’re very good).

MLE example: parameter of exponential pdf

Consider exponential pdf, $f(t;\tau) = \frac{1}{\tau} e^{-t/\tau}$

and suppose we have i.i.d. data, $t_1, \ldots, t_n$

The likelihood function is $L(\tau) = \prod_{i=1}^{n} \frac{1}{\tau} e^{-t_i/\tau}$

The value of τ for which $L(\tau)$ is maximum also gives the maximum value of its logarithm (the log-likelihood function):

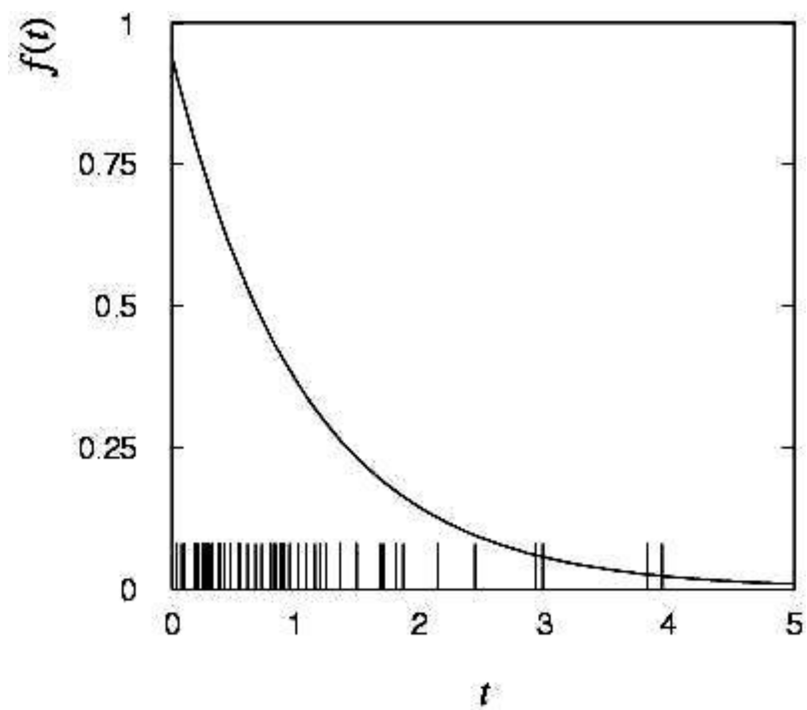
$$\ln L(\tau) = \sum_{i=1}^{n} \ln f(t_i;\tau) = \sum_{i=1}^{n} \left(\ln \frac{1}{\tau} - \frac{t_i}{\tau} \right)$$

MLE example: parameter of exponential pdf (2)

Find its maximum by setting $$\frac{\partial \ln L(\tau)}{\partial \tau} = 0 \,,$$

$$\rightarrow \quad \hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i$$

Monte Carlo test:
    generate 50 values
    using $\tau = 1$:



We find the ML estimate:

$$\hat{\tau} = 1.062$$

MLE example: parameter of exponential pdf (3)

For the exponential distribution one has for mean, variance:

$$E[t] = \int_0^\infty t \, \frac{1}{\tau} e^{-t/\tau} \, dt = \tau$$

$$V[t] = \int_0^\infty (t-\tau)^2 \, \frac{1}{\tau} e^{-t/\tau} \, dt = \tau^2$$

For the MLE $\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i$ we therefore find

$$E[\hat{\tau}] = E\left[\frac{1}{n}\sum_{i=1}^{n} t_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[t_i] = \tau \quad \longrightarrow \quad b = E[\hat{\tau}] - \tau = 0$$

$$V[\hat{\tau}] = V\left[\frac{1}{n}\sum_{i=1}^{n} t_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} V[t_i] = \frac{\tau^2}{n} \quad \longrightarrow \quad \sigma_{\hat{\tau}} = \frac{\tau}{\sqrt{n}}$$

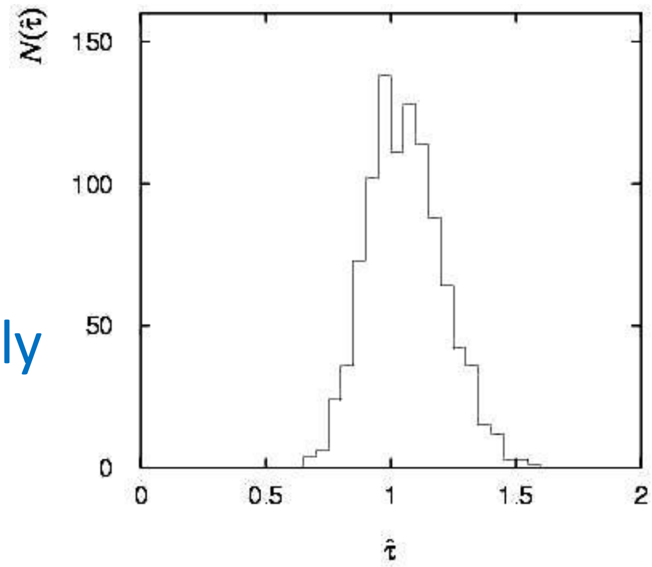
Variance of estimators: Monte Carlo method

Having estimated our parameter we now need to report its 'statistical error', i.e., how widely distributed would estimates be if we were to repeat the entire measurement many times.

One way to do this would be to simulate the entire experiment many times with a Monte Carlo program (use ML estimate for MC).

For exponential example, from sample variance of estimates we find:

$$\widehat{\sigma}_{\widehat{\tau}} = 0.151$$



Note distribution of estimates is roughly Gaussian − (almost) always true for ML in large sample limit.

Variance of estimators from information inequality

The information inequality (RCF) sets a lower bound on the variance of any estimator (not only ML):

Minimum Variance Bound (MVB)

$$V[\widehat{\theta}] \geq \left(1 + \frac{\partial b}{\partial \theta}\right)^2 \Big/ E\left[-\frac{\partial^2 \ln L}{\partial \theta^2}\right]$$

$$(b = E[\widehat{\theta}] - \theta)$$

Often the bias b is small, and equality either holds exactly or is a good approximation (e.g. large data sample limit). Then,

$$V[\widehat{\theta}] \approx -1 \Big/ E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]$$

Estimate this using the 2nd derivative of $\ln L$ at its maximum:

$$\widehat{V}[\widehat{\theta}] = -\left(\frac{\partial^2 \ln L}{\partial \theta^2}\right)^{-1}\Bigg|_{\theta=\widehat{\theta}}$$

MVB for MLE of exponential parameter

Find $$\mathrm{MVB} = -\left(1 + \frac{\partial b}{\partial \tau}\right)^2 \Big/ E\left[\frac{\partial^2 \ln L}{\partial \tau^2}\right]$$

We found for the exponential parameter the MLE $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} t_i$

and we showed $b = 0$, hence $\partial b/\partial\tau = 0$.

We find $$\frac{\partial^2 \ln L}{\partial \tau^2} = \sum_{i=1}^{n}\left(\frac{1}{\tau^2} - \frac{2t_i}{\tau^3}\right)$$

and since $E[t_i] = \tau$ for all i, $E\left[\frac{\partial^2 \ln L}{\partial \tau^2}\right] = -\frac{n}{\tau^2}$,

and therefore $\mathrm{MVB} = \frac{\tau^2}{n} = V[\hat{\tau}]$. (Here MLE is "efficient").

Variance of estimators: graphical method

Expand $\ln L(\theta)$ about its maximum:

$$\ln L(\theta) = \ln L(\widehat{\theta}) + \left[\frac{\partial \ln L}{\partial \theta}\right]_{\theta=\widehat{\theta}} (\theta - \widehat{\theta}) + \frac{1}{2!}\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]_{\theta=\widehat{\theta}} (\theta - \widehat{\theta})^2 + \ldots$$

First term is $\ln L_{\max}$, second term is zero, for third term use information inequality (assume equality):

$$\ln L(\theta) \approx \ln L_{\max} - \frac{(\theta - \widehat{\theta})^2}{2\widehat{\sigma^2}_{\widehat{\theta}}}$$

i.e., $$\ln L(\widehat{\theta} \pm \widehat{\sigma}_{\widehat{\theta}}) \approx \ln L_{\max} - \frac{1}{2}$$

$\rightarrow$ to get $\widehat{\sigma}_{\widehat{\theta}}$, change θ away from $\widehat{\theta}$ until $\ln L$ decreases by 1/2.

Example of variance by graphical method

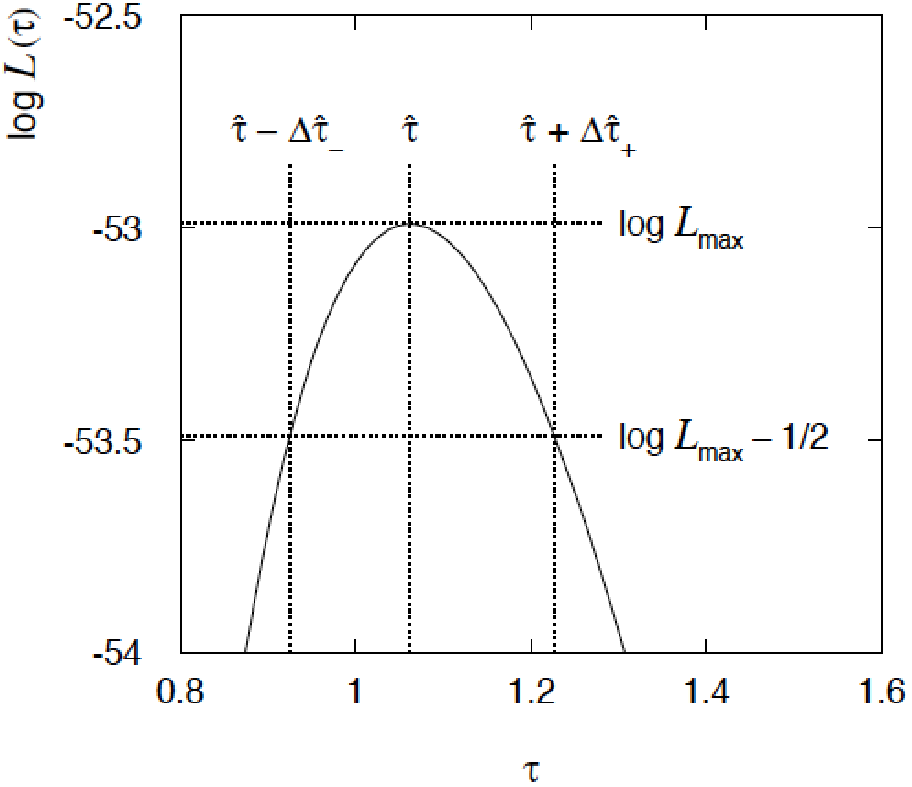
ML example with exponential:

$$\hat{\tau} = 1.062$$

$$\Delta\hat{\tau}_{-} = 0.137$$

$$\Delta\hat{\tau}_{+} = 0.165$$

$$\hat{\sigma}_{\hat{\tau}} \approx \Delta\hat{\tau}_{-} \approx \Delta\hat{\tau}_{+} \approx 0.15$$



Not quite parabolic $\ln L$ since finite sample size (n = 50).

Confidence intervals by inverting a test

In addition to a 'point estimate' of a parameter we should report an interval reflecting its statistical uncertainty.

Confidence intervals for a parameter θ can be found by defining a test of the hypothesized value θ (do this for all θ):

Specify values of the data that are 'disfavoured' by θ (critical region) such that $P(\text{data in critical region} \mid \theta) \leq \alpha$ for a prespecified α, e.g., 0.05 or 0.1.

If data observed in the critical region, reject the value θ.

Now invert the test to define a confidence interval as:

set of θ values that are not rejected in a test of size α (confidence level CL is $1-\alpha$).

Relation between confidence interval and p-value

Equivalently we can consider a significance test for each hypothesized value of θ, resulting in a p-value, p_θ.

If $p_\theta \leq \alpha$, then we reject θ.

The confidence interval at CL $= 1 - \alpha$ consists of those values of θ that are not rejected.

E.g. an upper limit on θ is the greatest value for which $p_\theta > \alpha$.

In practice find by setting $p_\theta = \alpha$ and solve for θ.

For a multidimensional parameter space $\boldsymbol{\theta} = (\theta_1, \ldots \theta_M)$ *use* same idea – result is a confidence "region" with boundary determined by $p_{\boldsymbol{\theta}} = \alpha$.

Coverage probability of confidence interval

If the true value of θ is rejected, then it's not in the confidence interval. The probability for this is by construction (equality for continuous data):

$$P(\text{reject } \theta | \theta) \leq \alpha = \text{type-I error rate}$$

Therefore, the probability for the interval to contain or "cover" θ is

$$P(\text{conf. interval "covers" } \theta | \theta) \geq 1 - \alpha$$

This assumes that the set of θ values considered includes the true value, i.e., it assumes the composite hypothesis $P(\boldsymbol{x}|H,\theta)$.

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 $P(n \le n_{\text{obs}}; s, b)$

Upper limit s_{up} at $\text{CL} = 1 - \alpha$ found from

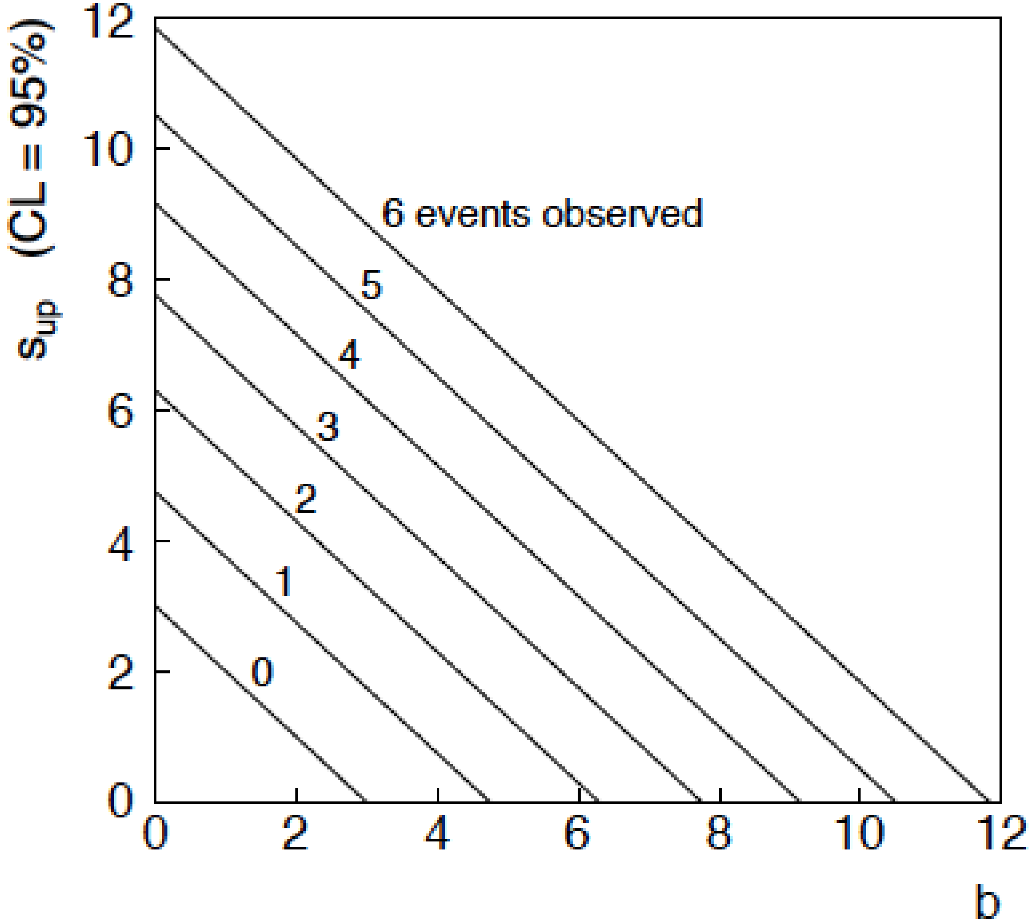
$$\alpha = P(n \le n_{\text{obs}}; s_{\text{up}}, b) = \sum_{n=0}^{n_{\text{obs}}} \frac{(s_{\text{up}} + b)^n}{n!} e^{-(s_{\text{up}}+b)}$$

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

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

$n \sim$ 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; all values of $s \geq 0$ have $p_s \leq \alpha$.



Limits near a boundary of the parameter space

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

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

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

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 $s = 0$

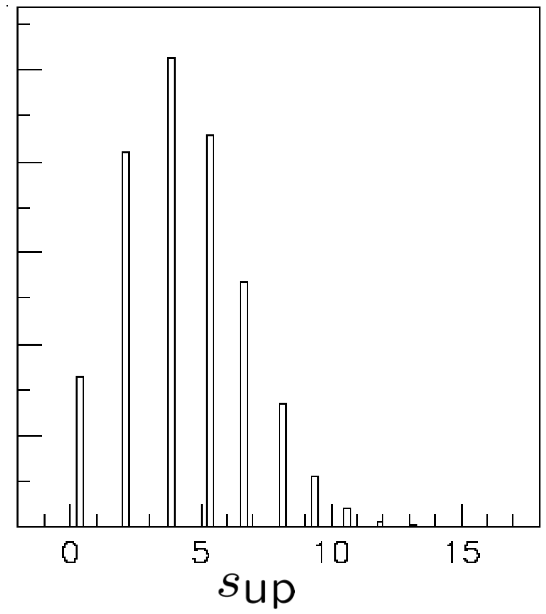
Physicist: I should have used CL = 0.95 — then $s_{\mathrm{up}} = 0.496$

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

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



moved to tutorial

Approximate confidence intervals/regions from the likelihood function

Suppose we test parameter value(s) $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$ using the ratio

$$\lambda(\boldsymbol{\theta}) = \frac{L(\boldsymbol{\theta})}{L(\hat{\boldsymbol{\theta}})} \qquad 0 \leq \lambda(\boldsymbol{\theta}) \leq 1$$

Lower $\lambda(\boldsymbol{\theta})$ means worse agreement between data and hypothesized $\boldsymbol{\theta}$. Equivalently, usually define

$$t_{\boldsymbol{\theta}} = -2 \ln \lambda(\boldsymbol{\theta})$$

so higher t_θ means worse agreement between $\boldsymbol{\theta}$ and the data.

p-value of $\boldsymbol{\theta}$ therefore $\quad p_{\boldsymbol{\theta}} = \int_{t_{\boldsymbol{\theta},\mathrm{obs}}}^{\infty} f(t_{\boldsymbol{\theta}}|\boldsymbol{\theta})\, dt_{\boldsymbol{\theta}}$

need pdf

moved to tutorial

Confidence region from Wilks' theorem

Wilks' theorem says (in large-sample limit and provided certain conditions hold...)

$$f(t_{\boldsymbol{\theta}}|\boldsymbol{\theta}) \sim \chi^2_n$$

chi-square dist. with # d.o.f. = # of components in $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$.

Assuming this holds, the p-value is

$$p_{\boldsymbol{\theta}} = 1 - F_{\chi^2_n}(t_{\boldsymbol{\theta}})$$

← set equal to α

To find boundary of confidence region set $p_{\theta} = \alpha$ and solve for t_{θ}:

$$t_{\boldsymbol{\theta}} = F^{-1}_{\chi^2_n}(1-\alpha)$$

Recall also

$$t_{\boldsymbol{\theta}} = -2 \ln \frac{L(\boldsymbol{\theta})}{L(\hat{\boldsymbol{\theta}})}$$

moved to tutorial

Confidence region from Wilks' theorem (cont.)

i.e., boundary of confidence region in $\boldsymbol{\theta}$ space is where

$$\ln L(\boldsymbol{\theta}) = \ln L(\hat{\boldsymbol{\theta}}) - \tfrac{1}{2} F^{-1}_{\chi^2_n}(1-\alpha)$$

For example, for $1 - \alpha = 68.3\%$ and $n = 1$ parameter,

$$F^{-1}_{\chi^2_1}(0.683) = 1$$

and so the 68.3% confidence level interval is determined by

$$\ln L(\theta) = \ln L(\hat{\theta}) - \frac{1}{2}$$

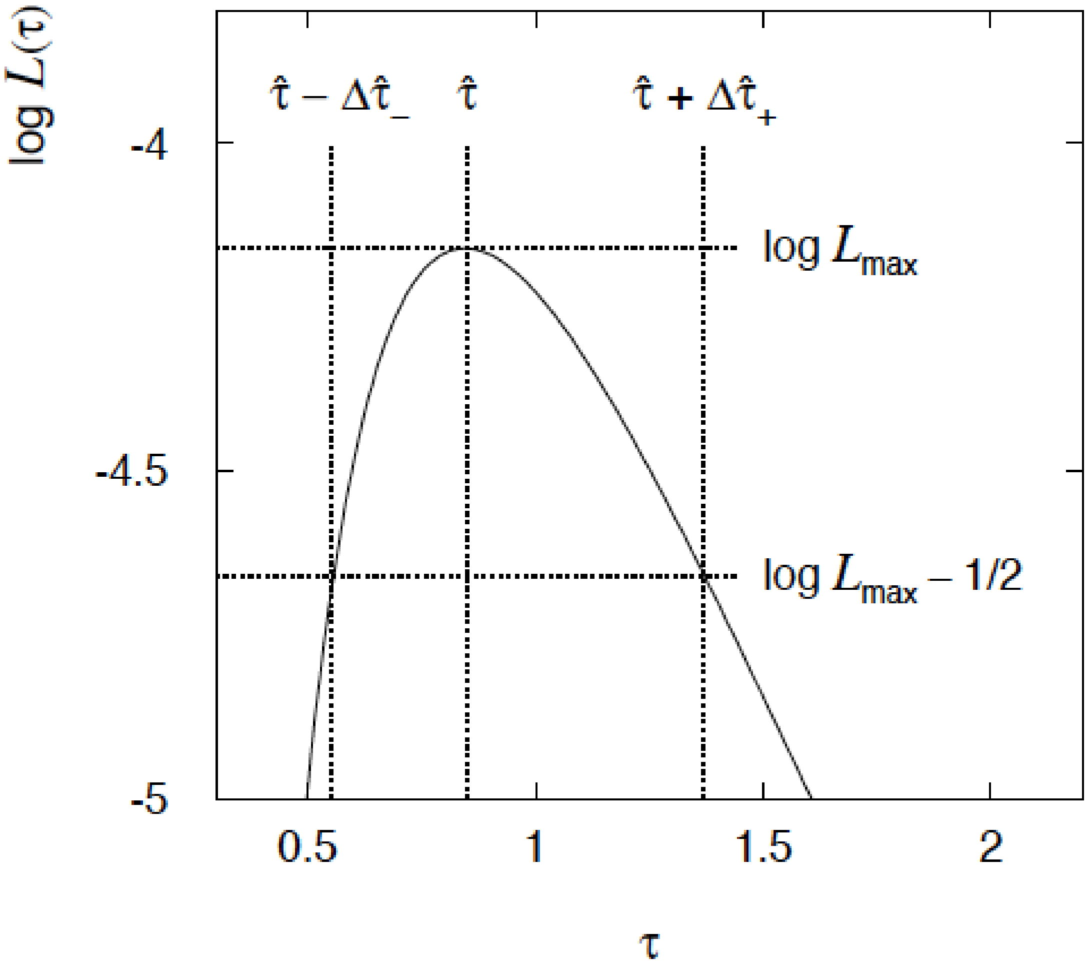
Same as recipe for finding the estimator's standard deviation, i.e.,

$[\hat{\theta} - \sigma_{\hat{\theta}}, \hat{\theta} + \sigma_{\hat{\theta}}]$ is a 68.3% CL confidence interval.

moved to tutorial

Example of interval from $\ln L(\theta)$

For n=1 parameter, CL = 0.683, $Q_\alpha = 1$.



Our exponential example, now with only $n = 5$ events.

Can report ML estimate with approx. confidence interval from $\ln L_{\max} - 1/2$ as "asymmetric error bar":

$$\hat{\tau} = 0.85^{+0.52}_{-0.30}$$

moved to tutorial

Multiparameter case

For increasing number of parameters, CL $= 1 - \alpha$ decreases for confidence region determined by a given

$$Q_\alpha = F^{-1}_{\chi^2_n}(1-\alpha)$$

Q_α	$1-\alpha$				
	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
1.0	0.683	0.393	0.199	0.090	0.037
2.0	0.843	0.632	0.428	0.264	0.151
4.0	0.954	0.865	0.739	0.594	0.451
9.0	0.997	0.989	0.971	0.939	0.891

moved to tutorial

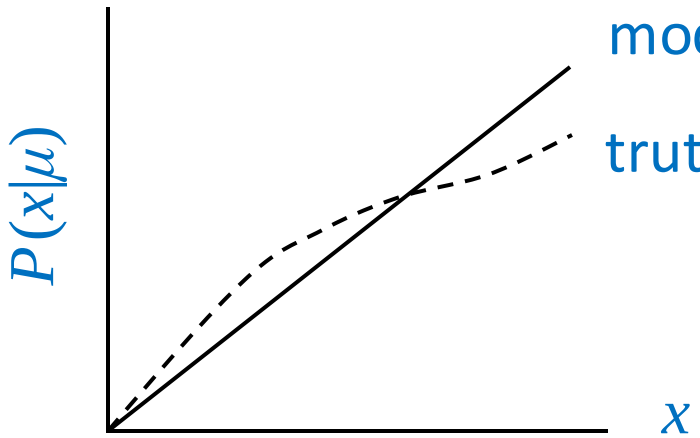
Multiparameter case (cont.)

Equivalently, Q_α increases with n for a given CL = $1 - \alpha$.

$1-\alpha$	Q_α				
	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.683	1.00	2.30	3.53	4.72	5.89
0.90	2.71	4.61	6.25	7.78	9.24
0.95	3.84	5.99	7.82	9.49	11.1
0.99	6.63	9.21	11.3	13.3	15.1

Systematic uncertainties and nuisance parameters

In general, our model of the data is not perfect:



model: $P(x|\mu) = \mu x$

truth: $P(x|\mu) = \mu x + \alpha x^2 + \beta x^3 + \cdots$

Can improve model by including additional adjustable parameters.

$$P(x|\mu) \to P(x|\mu, \boldsymbol{\theta})$$

Nuisance parameter ↔ systematic uncertainty. Some point in the parameter space of the enlarged model should be "true".

Presence of nuisance parameter decreases sensitivity of analysis to the parameter of interest (e.g., increases variance of estimate).

Profile Likelihood

Suppose we have a likelihood $L(\boldsymbol{\mu},\boldsymbol{\theta}) = P(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\theta})$ with N parameters of interest $\boldsymbol{\mu} = (\mu_1,\ldots, \mu_N)$ and M nuisance parameters $\boldsymbol{\theta} = (\theta_1,\ldots, \theta_M)$. The "profiled" (or "constrained") values of $\boldsymbol{\theta}$ are:

$$\hat{\hat{\boldsymbol{\theta}}}(\boldsymbol{\mu}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\boldsymbol{\mu}, \boldsymbol{\theta})$$

and the profile likelihood is: $L_{\mathrm{p}}(\boldsymbol{\mu}) = L(\boldsymbol{\mu}, \hat{\hat{\boldsymbol{\theta}}})$

The profile likelihood depends only on the parameters of interest; the nuisance parameters are replaced by their profiled values.

The profile likelihood can be used to obtain confidence intervals/regions for the parameters of interest in the same way as one would for all of the parameters from the full likelihood.

Profile Likelihood Ratio – Wilks theorem

Goal is to test/reject regions of $\boldsymbol{\mu}$ space (param. of interest).

Rejecting a point $\boldsymbol{\mu}$ should mean $p_{\boldsymbol{\mu}} \leq \alpha$ for all possible values of the nuisance parameters $\boldsymbol{\theta}$.

Test $\boldsymbol{\mu}$ using the "profile likelihood ratio": $\lambda(\boldsymbol{\mu}) = \dfrac{L(\boldsymbol{\mu}, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\theta}})}$

Let $t_{\boldsymbol{\mu}} = -2\ln\lambda(\boldsymbol{\mu})$. Wilks' theorem says in large-sample limit:

$$t_{\boldsymbol{\mu}} \sim \text{chi-square}(N)$$

where the number of degrees of freedom is the number of parameters of interest (components of $\boldsymbol{\mu}$). So p-value for $\boldsymbol{\mu}$ is

$$p_{\boldsymbol{\mu}} = \int_{t_{\mu,\mathrm{obs}}}^{\infty} f(t_{\boldsymbol{\mu}} | \boldsymbol{\mu}, \boldsymbol{\theta})\, dt_{\boldsymbol{\mu}} = 1 - F_{\chi^2_N}(t_{\boldsymbol{\mu},\mathrm{obs}})$$

Profile Likelihood Ratio – Wilks theorem (2)

If we have a large enough data sample to justify use of the asymptotic chi-square pdf, then if $\boldsymbol{\mu}$ is rejected, it is rejected for any values of the nuisance parameters.

The recipe to get confidence regions/intervals for the parameters of interest at CL = 1 – α is thus the same as before, simply use the profile likelihood:

$$\ln L_{\rm p}(\boldsymbol{\mu}) = \ln L_{\max} - \frac{1}{2} F^{-1}_{\chi^2_N}(1-\alpha)$$

where the number of degrees of freedom N for the chi-square quantile is equal to the number of parameters of interest.

If the large-sample limit is not justified, then use e.g. Monte Carlo to get distribution of $t_{\boldsymbol{\mu}}$.

Finally

Two lectures only enough for a brief introduction to:

- Parameter estimation
- Hypothesis tests ($\rightarrow$ path to Machine Learning)
- Limits (confidence intervals/regions)
- Systematics (nuisance parameters)

No time for many other interesting topics:

- Experimental sensitivity
- Bayesian parameter estimation

Final thought: once the basic formalism is fixed, most of the work focuses on writing down the likelihood, e.g., $P(\boldsymbol{x}|\boldsymbol{\theta})$, and including in it enough parameters to adequately describe the data (true for both Bayesian and frequentist approaches) so often best to invest most of your time with it.

Extra slides

Information inequality for N parameters

Suppose we have estimated N parameters $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_N)$

The *Fisher information matrix* is

$$I_{ij} = -E\left[\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right] = -\int \frac{\partial^2 \ln P(\mathbf{x}|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j} P(\mathbf{x}|\boldsymbol{\theta})\, d\mathbf{x}$$

and the covariance matrix of estimators $\hat{\boldsymbol{\theta}}$ is $V_{ij} = \mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j]$

The information inequality states that the matrix

$$M_{ij} = V_{ij} - \sum_{k,l} \left(\delta_{ik} + \frac{\partial b_i}{\partial\theta_k}\right) I^{-1}_{kl} \left(\delta_{lj} + \frac{\partial b_l}{\partial\theta_j}\right)$$

is positive semi-definite:

$\boldsymbol{z}^{\mathrm{T}} M \boldsymbol{z} \geq 0$ for all $\boldsymbol{z} \neq 0$, diagonal elements ≥ 0

Information inequality for N parameters (2)

In practice the inequality is ~always used in the large-sample limit:

bias $\rightarrow$ 0

inequality $\rightarrow$ equality, i.e, $M = 0$, and therefore $V^{-1} = I$

That is, $$V_{ij}^{-1} = -E\left[\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right]$$

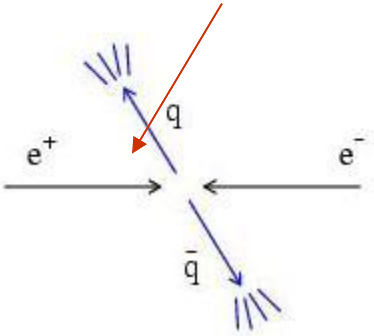
This can be estimated from data using $$\widehat{V}_{ij}^{-1} = -\left.\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right|_{\hat{\boldsymbol{\theta}}}$$

Find the matrix V^{-1} numerically (or with automatic differentiation), then invert to get the covariance matrix of the estimators

$$\widehat{V}_{ij} = \widehat{\mathrm{cov}}[\hat{\theta}_i, \hat{\theta}_j]$$

Example of ML with 2 parameters

Consider a scattering angle distribution with $x = \cos\theta$,



$$f(x;\alpha,\beta) = \frac{1+\alpha x+\beta x^2}{2+2\beta/3}$$

or if $x_{\min} < x < x_{\max}$, need to normalize so that

$$\int_{x_{\min}}^{x_{\max}} f(x;\alpha,\beta)\,dx = 1\ .$$

Example: $\alpha = 0.5$, $\beta = 0.5$, $x_{\min} = -0.95$, $x_{\max} = 0.95$, generate $n = 2000$ events with Monte Carlo.

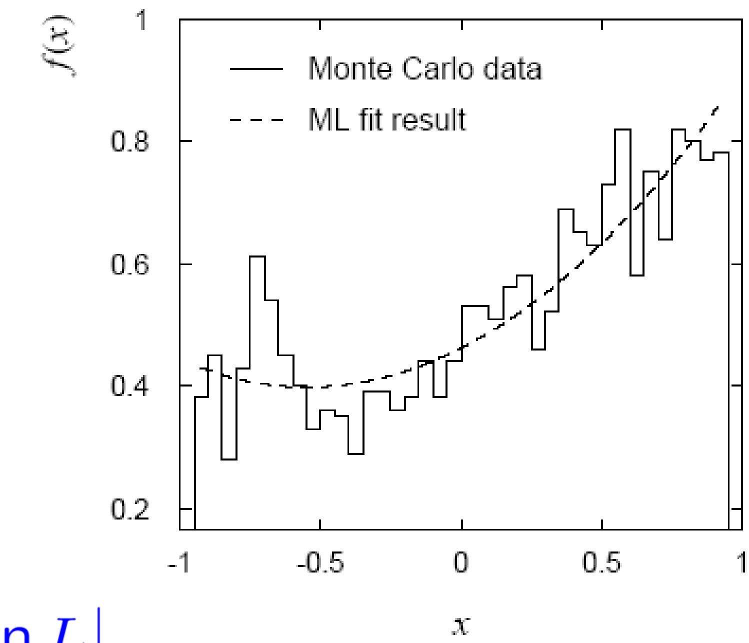
$$\ln L(\alpha,\beta) = \sum_{i=1}^{n} \ln f(x_i;\alpha,\beta)$$

need to find maximum numerically

Example of ML with 2 parameters: fit result

Finding maximum of $\ln L(\alpha, \beta)$ numerically gives

$$\hat{\alpha} = 0.508$$
$$\hat{\beta} = 0.47$$



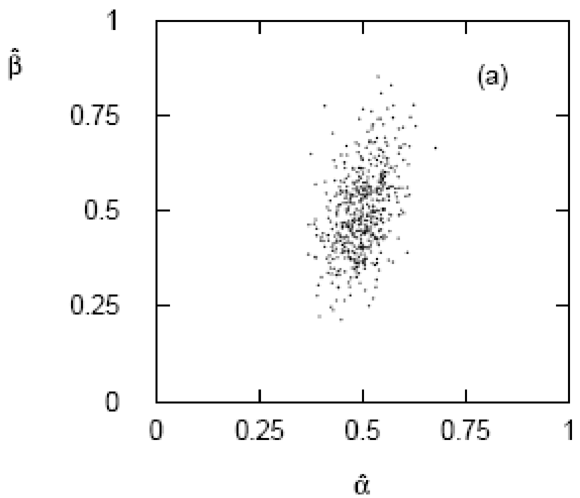
N.B. No binning of data for fit,
but can compare to histogram for
goodness-of-fit (e.g. 'visual' or χ^2).

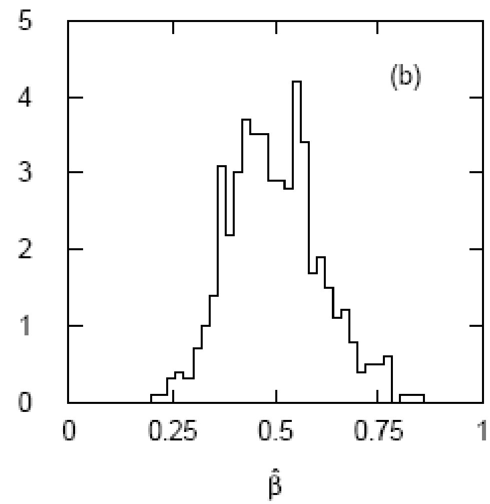
(Co)variances from $\displaystyle (\widehat{V^{-1}})_{ij} = - \left. \frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j} \right|_{\vec{\theta} = \hat{\vec{\theta}}}$

$$\hat{\sigma}_{\hat{\alpha}} = 0.052 \qquad \mathrm{cov}[\hat{\alpha}, \hat{\beta}] = 0.0026$$
$$\hat{\sigma}_{\hat{\beta}} = 0.11 \qquad r = 0.46$$ = correlation coefficient

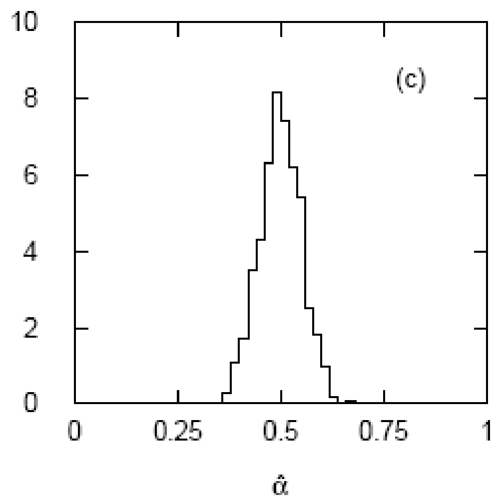
Two-parameter fit: MC study

Repeat ML fit with 500 experiments, all with $n = 2000$ events:





$$\overline{\hat{\alpha}} = 0.499$$
$$s_{\hat{\alpha}} = 0.051$$
$$\overline{\hat{\beta}} = 0.498$$
$$s_{\hat{\beta}} = 0.111$$
$$\widehat{\mathrm{cov}}[\hat{\alpha}, \hat{\beta}] = 0.0024$$
$$r = 0.42$$



Estimates average to ~true values;
(Co)variances close to previous estimates;
marginal pdfs approximately Gaussian.

Multiparameter graphical method for variances

Expand $\ln L(\boldsymbol{\theta})$ to 2nd order about MLE:

$$\ln L(\boldsymbol{\theta}) \approx \ln L(\hat{\boldsymbol{\theta}}) + \sum_i \left.\frac{\partial \ln L}{\partial \theta_i}\right|_{\hat{\boldsymbol{\theta}}} (\theta_i - \hat{\theta}_i) + \frac{1}{2!} \sum_{i,j} \left.\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}\right|_{\hat{\boldsymbol{\theta}}} (\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j)$$

$\ln L_{\max}$

zero

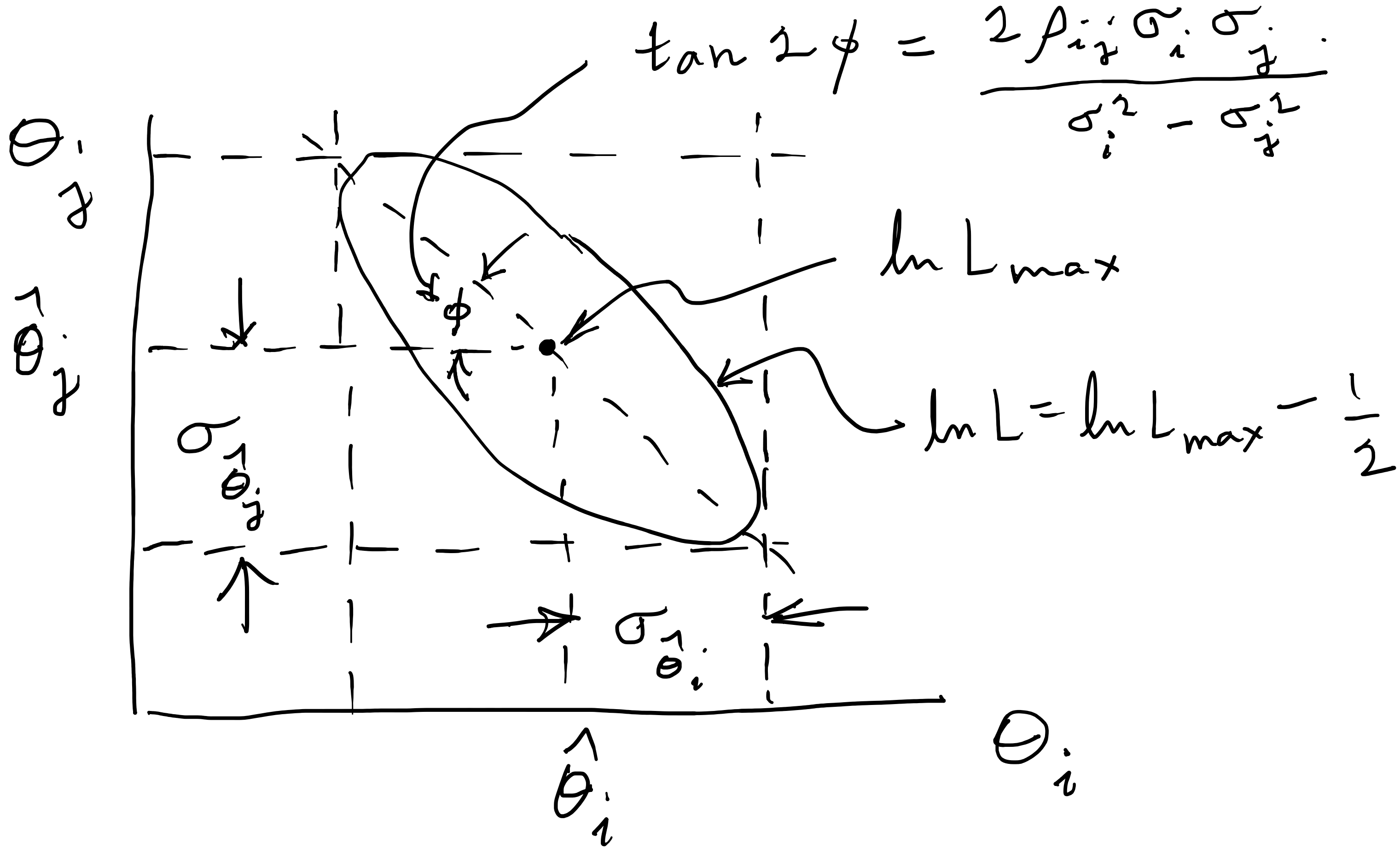
relate to covariance matrix of MLEs using information (in)equality.

Result: $$\ln L(\boldsymbol{\theta}) = \ln L_{\max} - \frac{1}{2} \sum_{i,j} (\theta_i - \hat{\theta}_i) V_{ij}^{-1} (\theta_j - \hat{\theta}_j)$$

So the surface $\ln L(\boldsymbol{\theta}) = \ln L_{\max} - \frac{1}{2}$ corresponds to

$(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T V^{-1} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) = 1$, which is the equation of a (hyper-) ellipse.

Multiparameter graphical method (2)



Distance from MLE to tangent planes gives standard deviations.

The $\ln L_{\max} - 1/2$ contour for two parameters

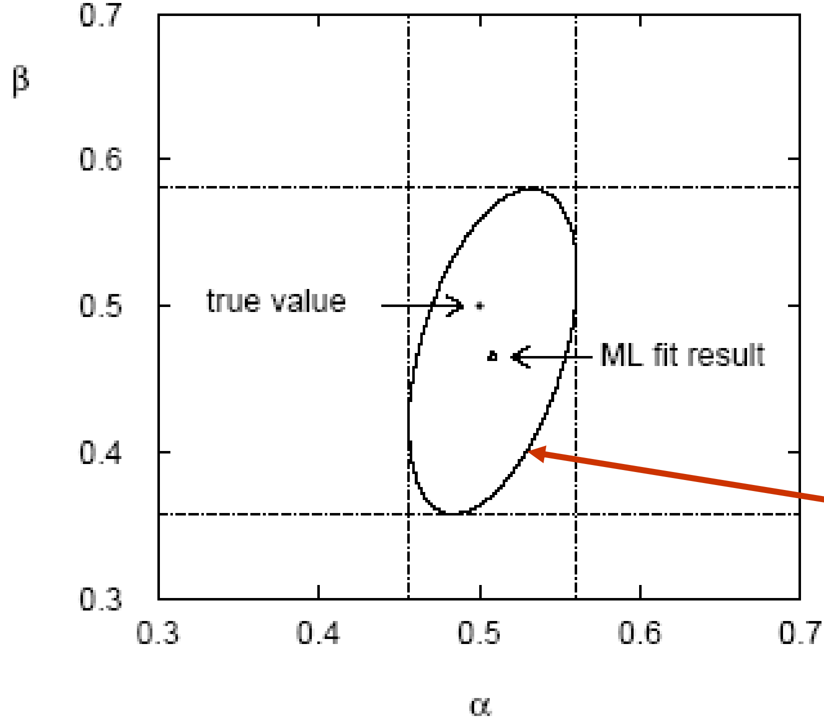
For large n, $\ln L$ takes on quadratic form near maximum:

$$\ln L(\alpha,\beta) \approx \ln L_{\max}$$
$$-\frac{1}{2(1-\rho^2)}\left[\left(\frac{\alpha-\hat{\alpha}}{\sigma_{\hat{\alpha}}}\right)^2 + \left(\frac{\beta-\hat{\beta}}{\sigma_{\hat{\beta}}}\right)^2 - 2\rho\left(\frac{\alpha-\hat{\alpha}}{\sigma_{\hat{\alpha}}}\right)\left(\frac{\beta-\hat{\beta}}{\sigma_{\hat{\beta}}}\right)\right]$$

The contour $\ln L(\alpha,\beta) = \ln L_{\max} - 1/2$ is an ellipse:

$$\frac{1}{(1-\rho^2)}\left[\left(\frac{\alpha-\hat{\alpha}}{\sigma_{\hat{\alpha}}}\right)^2 + \left(\frac{\beta-\hat{\beta}}{\sigma_{\hat{\beta}}}\right)^2 - 2\rho\left(\frac{\alpha-\hat{\alpha}}{\sigma_{\hat{\alpha}}}\right)\left(\frac{\beta-\hat{\beta}}{\sigma_{\hat{\beta}}}\right)\right] = 1$$

(Co)variances from ln *L* contour



The 〈, ® plane for the first MC data set

$$\ln L(\alpha, \beta) = \ln L_{\max} - 1/2$$

→ Tangent lines to contours give standard deviations.

→ Angle of ellipse φ related to correlation: $\tan 2\phi = \dfrac{2\rho\sigma_{\hat{\alpha}}\sigma_{\hat{\beta}}}{\sigma_{\hat{\alpha}}^2 - \sigma\hat{\beta}^2}$

Prototype search analysis

Search for signal in a region of phase space; result is histogram of some variable x giving numbers:

$$\mathbf{n} = (n_1, \ldots, n_N)$$

Assume the n_i are Poisson distributed with expectation values

$$E[n_i] = \mu s_i + b_i$$

strength parameter

where

$$s_i = s_{\rm tot} \int_{\mathrm{bin}\, i} f_s(x; \boldsymbol{\theta}_s)\, dx \,, \qquad b_i = b_{\rm tot} \int_{\mathrm{bin}\, i} f_b(x; \boldsymbol{\theta}_b)\, dx \,.$$

signal

background

Prototype analysis (II)

Often also have a subsidiary measurement that constrains some of the background and/or shape parameters:

$$\mathbf{m} = (m_1, \ldots, m_M)$$

Assume the m_i are Poisson distributed with expectation values

$$E[m_i] = u_i(\boldsymbol{\theta})$$

nuisance parameters ($\boldsymbol{\theta}_{\rm s}$, $\boldsymbol{\theta}_{\rm b}$,$b_{\rm tot}$)

Likelihood function is

$$L(\mu, \boldsymbol{\theta}) = \prod_{j=1}^{N} \frac{(\mu s_j + b_j)^{n_j}}{n_j!} e^{-(\mu s_j + b_j)} \quad \prod_{k=1}^{M} \frac{u_k^{m_k}}{m_k!} e^{-u_k}$$

The profile likelihood ratio

Base significance test on the profile likelihood ratio:

maximizes L for specified μ

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}$$

maximize L

Define critical region of test of μ by the region of data space that gives the lowest values of $\lambda(\mu)$.

Important advantage of profile LR is that its distribution becomes independent of nuisance parameters in large sample limit.

Test statistic for discovery

Suppose relevant alternative to background-only ($\mu = 0$) is $\mu \geq 0$.

So take critical region for test of $\mu = 0$ corresponding to high q_0 and $\hat{\mu} > 0$ (data characteristic for $\mu \geq 0$).

That is, to test background-only hypothesis define statistic

$$q_0 = \begin{cases} -2\ln\lambda(0) & \hat{\mu} \geq 0 \\ 0 & \hat{\mu} < 0 \end{cases}$$

i.e. here only large (positive) observed signal strength is evidence against the background-only hypothesis.

Note that even though here physically $\mu \geq 0$, we allow $\hat{\mu}$ to be negative. In large sample limit its distribution becomes Gaussian, and this will allow us to write down simple expressions for distributions of our test statistics.

Distribution of q_0 in large-sample limit

Assuming approximations valid in the large sample (asymptotic) limit, we can write down the full distribution of q_0 as

$$f(q_0|\mu') = \left(1 - \Phi\left(\frac{\mu'}{\sigma}\right)\right)\delta(q_0) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_0}}\exp\left[-\frac{1}{2}\left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right)^2\right]$$

The special case $\mu' = 0$ is a "half chi-square" distribution:

$$f(q_0|0) = \frac{1}{2}\delta(q_0) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_0}}e^{-q_0/2}$$

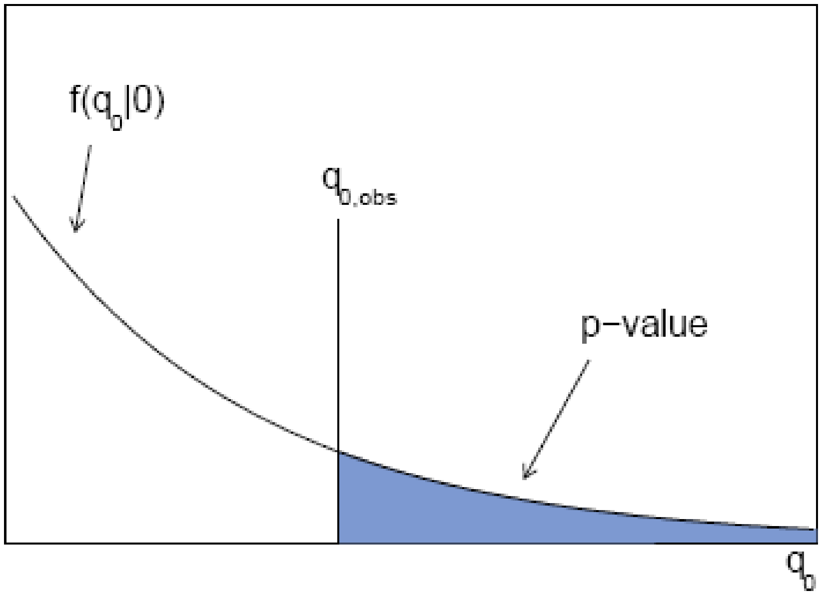
In large sample limit, $f(q_0|0)$ independent of nuisance parameters; $f(q_0|\mu')$ depends on nuisance parameters through σ.

p-value for discovery

Large q_0 means increasing incompatibility between the data and hypothesis, therefore p-value for an observed $q_{0,\mathrm{obs}}$ is

$$p_0 = \int_{q_{0,\mathrm{obs}}}^{\infty} f(q_0|0)\, dq_0$$

use e.g. asymptotic formula



From p-value get equivalent significance,

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

Cumulative distribution of q_0, significance

From the pdf, the cumulative distribution of q_0 is found to be

$$F(q_0|\mu') = \Phi\left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right)$$

The special case $\mu' = 0$ is

$$F(q_0|0) = \Phi\left(\sqrt{q_0}\right)$$

The *p*-value of the $\mu = 0$ hypothesis is

$$p_0 = 1 - F(q_0|0)$$

Therefore the discovery significance Z is simply

$$\boxed{Z = \Phi^{-1}(1 - p_0) = \sqrt{q_0}}$$

Monte Carlo test of asymptotic formula

$$n \sim \text{Poisson}(\mu s + b)$$

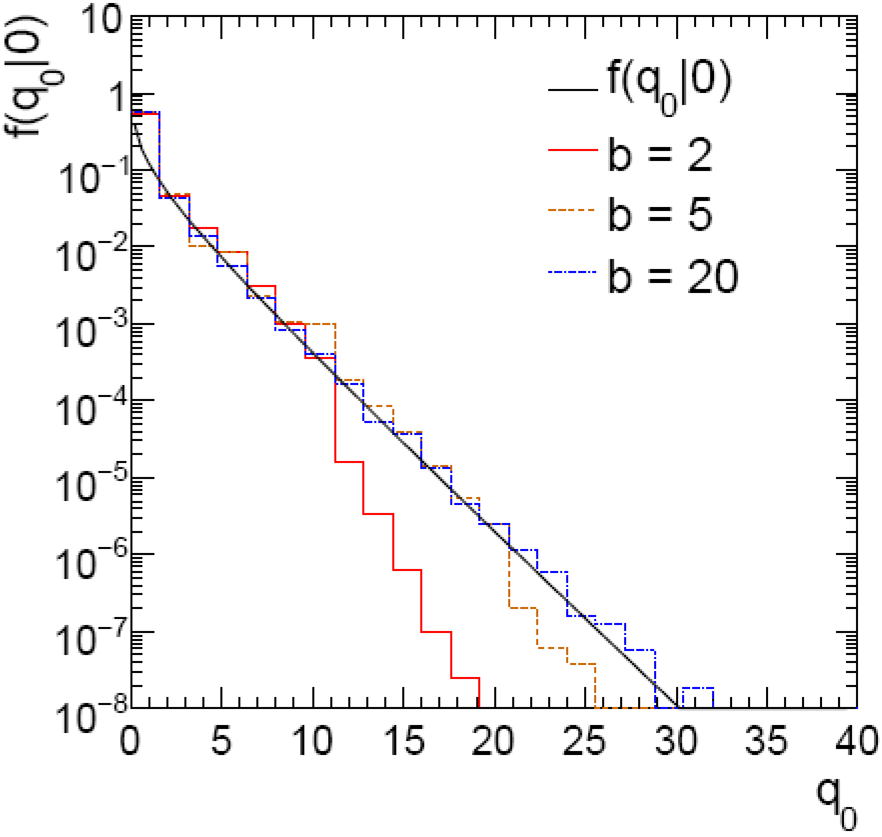
$$m \sim \text{Poisson}(\tau b)$$

μ = param. of interest

b = nuisance parameter

Here take s known, $\tau = 1$.

Asymptotic formula is good approximation to 5σ level ($q_0 = 25$) already for $b \sim 20$.

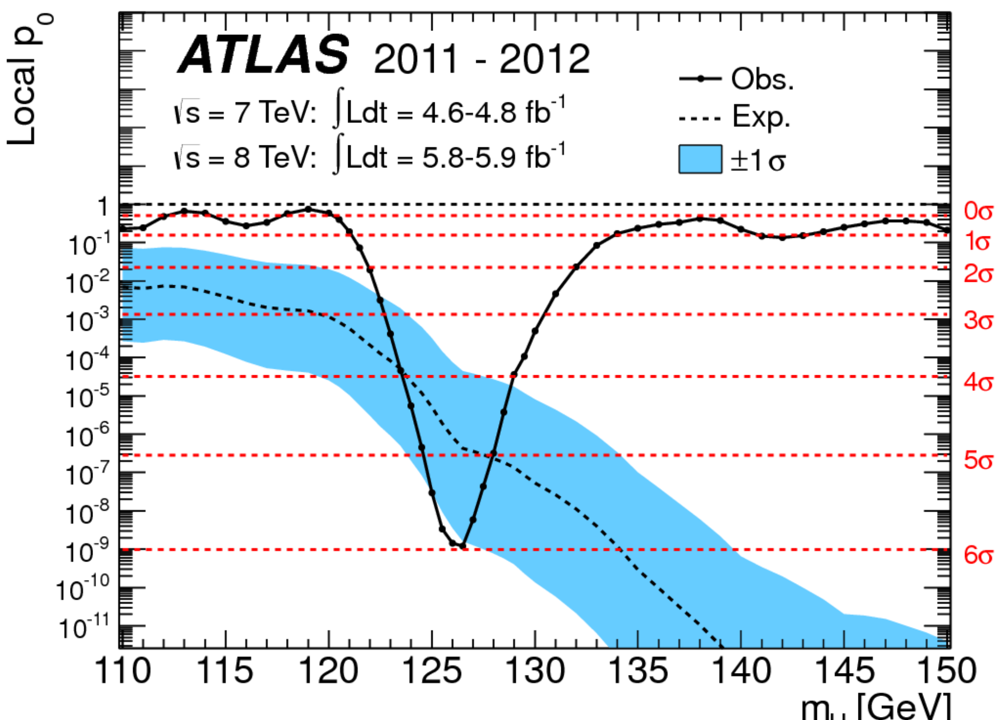


How to read the p_0 plot

The "local" p_0 means the p-value of the background-only hypothesis obtained from the test of $\mu = 0$ at each individual m_{H}, without any correct for the Look-Elsewhere Effect.

The "Expected" (dashed) curve gives the median p_0 under assumption of the SM Higgs (μ = 1) at each m_{H}.

ATLAS, Phys. Lett. B 716 (2012) 1-29



The blue band gives the width of the distribution ($\pm 1\sigma$) of significances under assumption of the SM Higgs.

Test statistic for upper limits

For purposes of setting an upper limit on μ use

$$q_\mu = \begin{cases} -2\ln\lambda(\mu) & \hat{\mu} \le \mu \\ 0 & \hat{\mu} > \mu \end{cases} \qquad \text{where} \qquad \lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}$$

I.e. when setting an upper limit, an upwards fluctuation of the data is not taken to mean incompatibility with the hypothesized μ :

From observed q_μ find p-value: $$p_\mu = \int_{q_{\mu,\mathrm{obs}}}^{\infty} f(q_\mu|\mu)\, dq_\mu$$

Large sample approximation: $$\boxed{p_\mu = 1 - \Phi\left(\sqrt{q_\mu}\right)}$$

To find upper limit at CL = 1–α, set $p_\mu = \alpha$ and solve for μ.

Monte Carlo test of asymptotic formulae

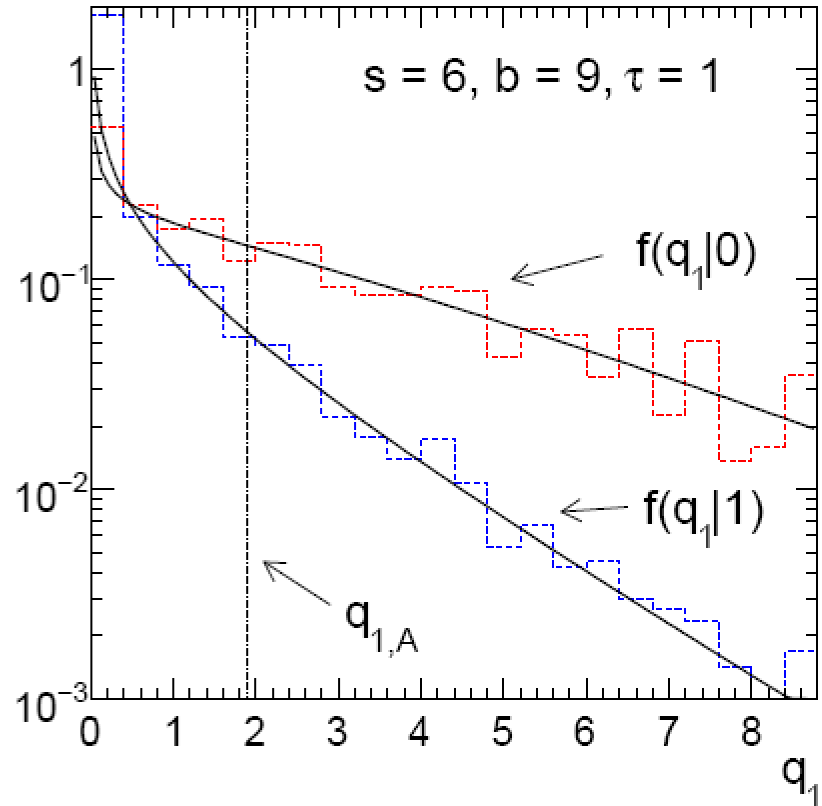
Consider again $n \sim$ Poisson($\mu s + b$), $m \sim$ Poisson(τb)
Use q_μ to find p-value of hypothesized μ values.

E.g. $f(q_1|1)$ for p-value of $\mu = 1$.

Typically interested in 95% CL, i.e., p-value threshold = 0.05, i.e., $q_1 = 2.69$ or $Z_1 = \sqrt{q_1} = 1.64$.

Median[$q_1|0$] gives "exclusion sensitivity".

Here asymptotic formulae good for $s = 6$, $b = 9$.



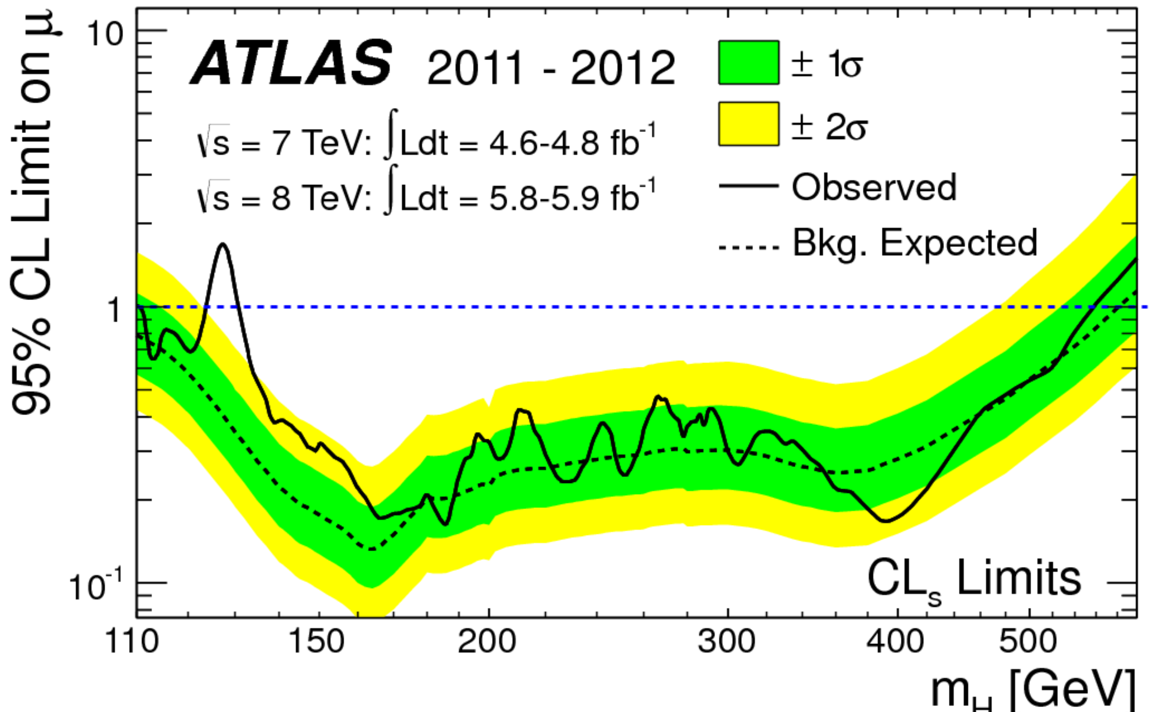
How to read the green and yellow limit plots

For every value of m_{H}, find the upper limit on μ.

Also for each m_{H}, determine the distribution of upper limits μ_{up} one would obtain under the hypothesis of $\mu = 0$.

The dashed curve is the median μ_{up}, and the green (yellow) bands give the $\pm$ 1σ (2σ) regions of this distribution.

ATLAS, Phys. Lett. B 716 (2012) 1-29



Sensitivity for Poisson counting experiment

Count a number of events $n \sim$ Poisson($s+b$), where

s = expected number of events from signal,

b = expected number of background events.

To test for discovery of signal compute p-value of $s = 0$ hypothesis,

$$p = P(n \geq n_{\rm obs}|b) = \sum_{n=n_{\rm obs}}^{\infty} \frac{b^n}{n!} e^{-b} = 1 - F_{\chi^2}(2b; 2n_{\rm obs})$$

Usually convert to equivalent significance: $Z = \Phi^{-1}(1-p)$
where Φ is the standard Gaussian cumulative distribution, e.g.,
$Z > 5$ (a 5 sigma effect) means $p < 2.9 \times 10^{-7}$.

To characterize sensitivity to discovery, give expected (mean or median) Z under assumption of a given s.

$s/\sqrt{b}$ for expected discovery significance

For large $s + b$, $n \to x \sim$ Gaussian(μ,σ) , $\mu = s + b$, $\sigma = \sqrt{(s + b)}$.

For observed value x_{obs}, p-value of $s = 0$ is Prob($x > x_{obs} \mid s = 0$),:

$$p_0 = 1 - \Phi\left(\frac{x_{\rm obs} - b}{\sqrt{b}}\right)$$

Significance for rejecting $s = 0$ is therefore

$$Z_0 = \Phi^{-1}(1 - p_0) = \frac{x_{\rm obs} - b}{\sqrt{b}}$$

Expected (median) significance assuming signal rate s is

$$\mathrm{median}[Z_0|s + b] = \frac{s}{\sqrt{b}}$$

Better approximation for significance

Poisson likelihood for parameter *s* is

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

For now no nuisance params.

To test for discovery use profile likelihood ratio:

$$q_0 = \begin{cases} -2 \ln \lambda(0) & \hat{s} \geq 0 \,, \\ 0 & \hat{s} < 0 \,. \end{cases}$$

$$\lambda(s) = \frac{L(s, \hat{\hat{\theta}}(s))}{L(\hat{s}, \hat{\theta})}$$

So the likelihood ratio statistic for testing $s = 0$ is

$$q_0 = -2 \ln \frac{L(0)}{L(\hat{s})} = 2 \left(n \ln \frac{n}{b} + b - n \right) \quad \text{for } n > b, \ 0 \text{ otherwise}$$

Approximate Poisson significance (continued)

For sufficiently large $s + b$, (use Wilks' theorem),

$$Z = \sqrt{2\left(n \ln \frac{n}{b} + b - n\right)} \quad \text{for } n > b \text{ and } Z = 0 \text{ otherwise.}$$

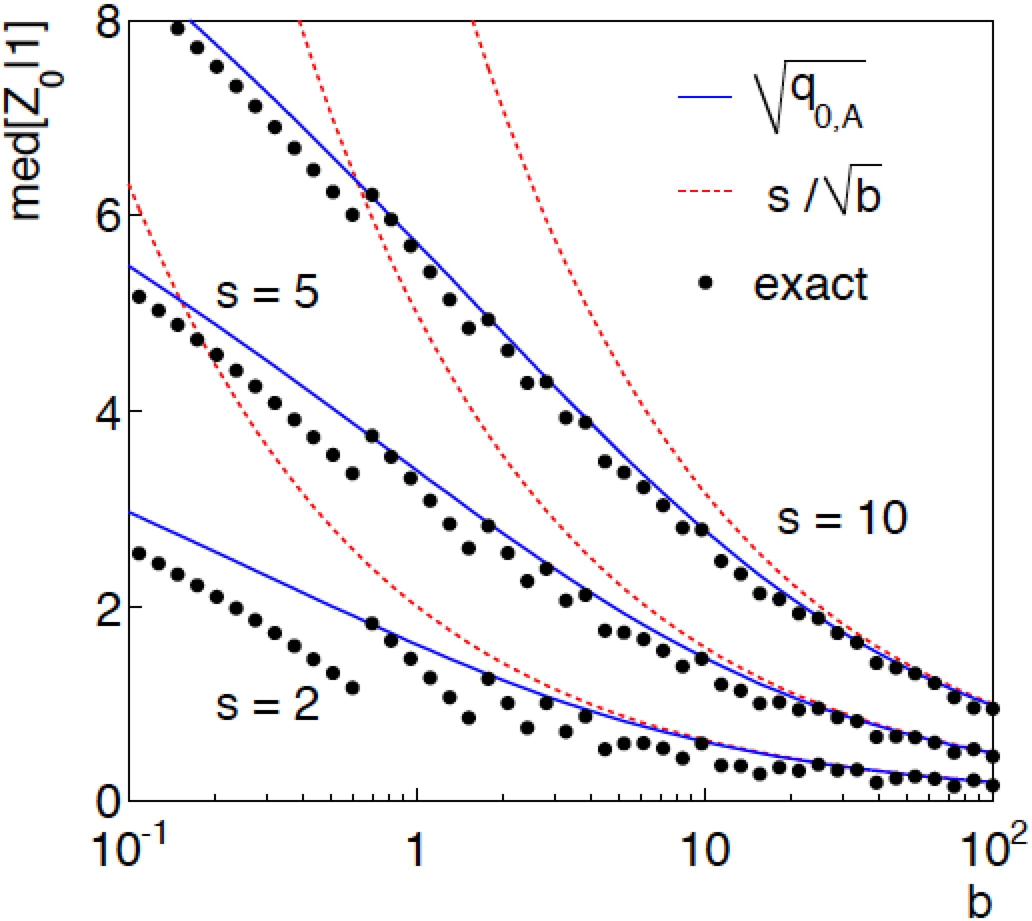
To find median[$Z|s$], let $n \rightarrow s + b$ (i.e., the Asimov data set):

$$Z_{\rm A} = \sqrt{2\left((s+b)\ln\left(1 + \frac{s}{b}\right) - s\right)}$$

This reduces to $s/\sqrt{b}$ for $s << b$.

n ~ Poisson(s+b), median significance, assuming s, of the hypothesis s = 0

CCGV, EPJC 71 (2011) 1554, arXiv:1007.1727



"Exact" values from MC, jumps due to discrete data.

Asimov $\sqrt{q_{0,\mathrm{A}}}$ good approx. for broad range of s, b.

$s/\sqrt{b}$ only good for $s \ll b$.