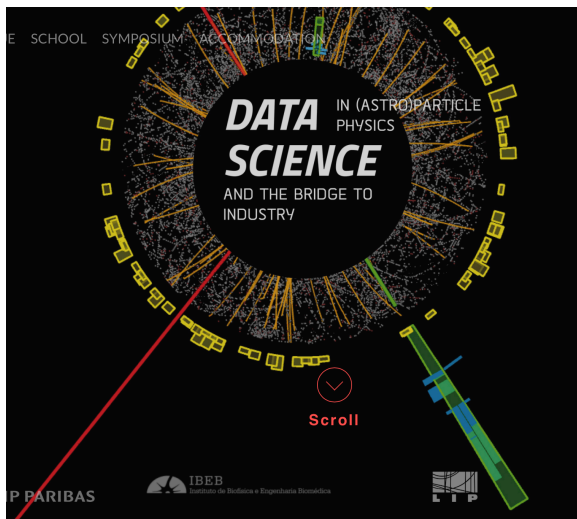


Statistical Methods for Particle Physics

<http://www.lip.pt/data-science-2018/>

<http://www.pp.rhul.ac.uk/~cowan/stat/lip18/>



School on Data Science in (Astro)particle Physics LIP Lisboa, 12-14 March, 2018

Glen Cowan

Physics Department

Royal Holloway, University of London

g.cowan@rhul.ac.uk

www.pp.rhul.ac.uk/~cowan

Outline

Probability, random variables, pdfs

Statistical tests

Comments on multivariate methods (brief)

p -values

Testing the background-only hypothesis: discovery

Parameter estimation

- maximum likelihood

- least squares

- Bayesian parameter estimation

Testing signal hypotheses: setting limits

Experimental sensitivity

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.

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

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

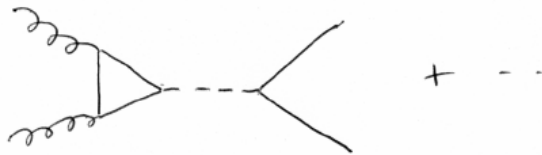
S. Brandt, *Statistical and Computational Methods in Data Analysis*, Springer, New York, 1998 (with program library on CD)

C. Patrignani et al. (Particle Data Group), Chin. Phys. C, 40, 100001 (2016); see also **pdg.lbl.gov** sections on probability, statistics, Monte Carlo

Theory \leftrightarrow Statistics \leftrightarrow Experiment

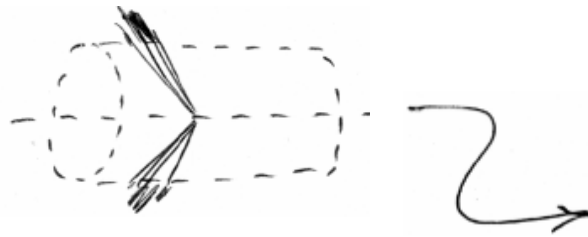
Theory (model, hypothesis):

$$\mathcal{L} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + i \bar{\Psi} \not{D} \Psi + \dots$$

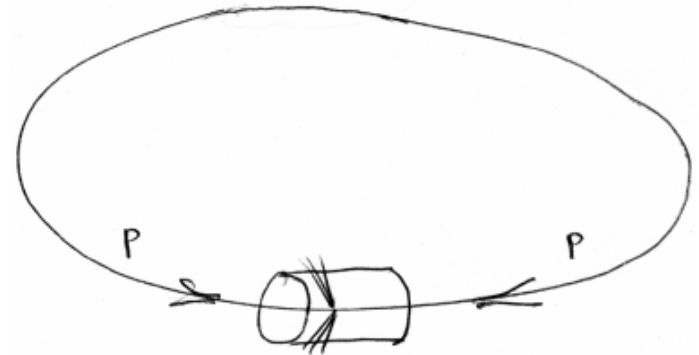


$$\sigma = \frac{G_F \alpha_s^2 m_H^2}{288 \sqrt{2\pi}} \times \dots$$

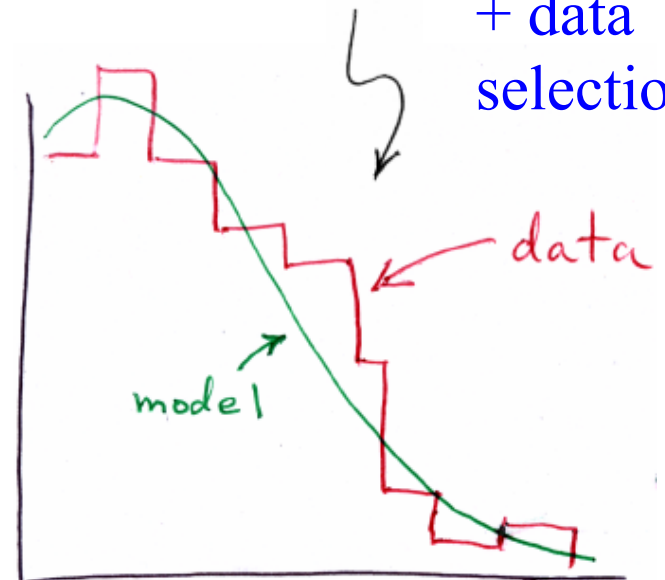
+ simulation
of detector
and cuts



Experiment:



+ data
selection



Data analysis in particle physics

Observe events (e.g., pp collisions) and for each, measure a set of characteristics:

particle momenta, number of muons, energy of jets,...

Compare observed distributions of these characteristics to predictions of theory. From this, we want to:

Estimate the free parameters of the theory: $m_H = 125.4$

Quantify the uncertainty in the estimates: $\pm 0.4 \text{ GeV}$

Assess how well a given theory stands in agreement with the observed data:

0^+ good, 2^+ bad

To do this we need a clear definition of **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

A, B, \dots are outcomes of a repeatable experiment

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

cf. quantum mechanics, particle scattering, radioactive decay...

II. Subjective probability

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 Higgs boson exists,...

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)

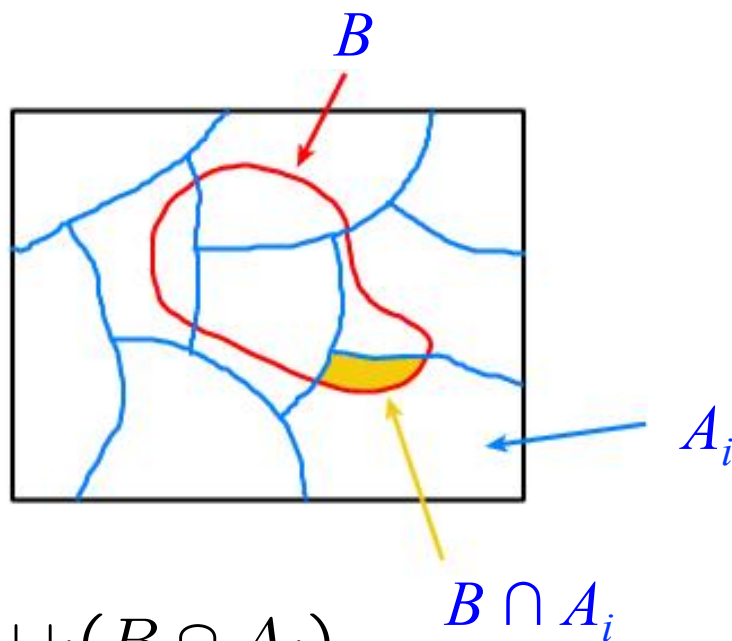
*An essay towards solving a problem in the
doctrine of chances*, Philos. Trans. R. Soc. **53**
(1763) 370; reprinted in Biometrika, **45** (1958) 293.

The law of total probability

Consider a subset B of the sample space S ,

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

S



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

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

$$\rightarrow P(B) = \sum_i P(B|A_i)P(A_i) \quad \text{law of total probability}$$

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:

$$\begin{aligned}P(D) &= 0.001 \\P(\text{no } D) &= 0.999\end{aligned}$$

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

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

$$\begin{aligned}P(+|D) &= 0.98 \\P(-|D) &= 0.02\end{aligned}$$

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

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

← 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

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

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: \vec{x}).

Probability = limiting frequency

Probabilities such as

P (Higgs boson exists),

$P(0.117 < \alpha_s < 0.121)$,

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.

A hypothesis is preferred if the data are found in a region of high predicted probability (i.e., where an alternative hypothesis predicts lower probability).

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)

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

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

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

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

Other types of probability densities

Outcome of experiment characterized by several values,
e.g. an n -component vector, (x_1, \dots, x_n)

→ joint pdf $f(x_1, \dots, x_n)$

Sometimes we want only pdf of some (or one) of the components

→ marginal pdf $f_1(x_1) = \int \cdots \int f(x_1, \dots, x_n) dx_2 \cdots dx_n$

x_1, x_2 independent if $f(x_1, x_2) = f_1(x_1)f_2(x_2)$

Sometimes we want to consider some components as constant

→ conditional pdf $g(x_1|x_2) = \frac{f(x_1, x_2)}{f_2(x_2)}$

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

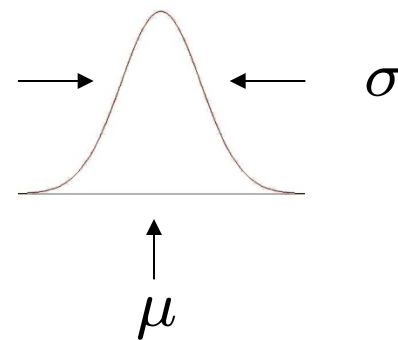
$$E[y] = \int y g(y) dy = \int y(x) f(x) dx \quad (\text{equivalent})$$

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

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

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

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



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

$$\rho_{xy} = \frac{\text{COV}[x, y]}{\sigma_x \sigma_y}$$

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

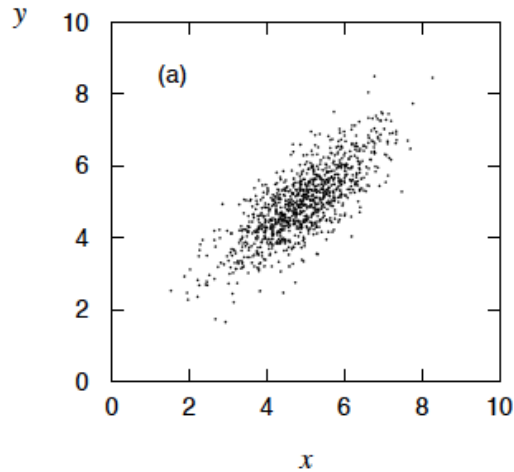
$$E[xy] = \int \int xy f(x, y) dx dy = \mu_x \mu_y$$

→ $\text{COV}[x, y] = 0$ x and y , ‘uncorrelated’

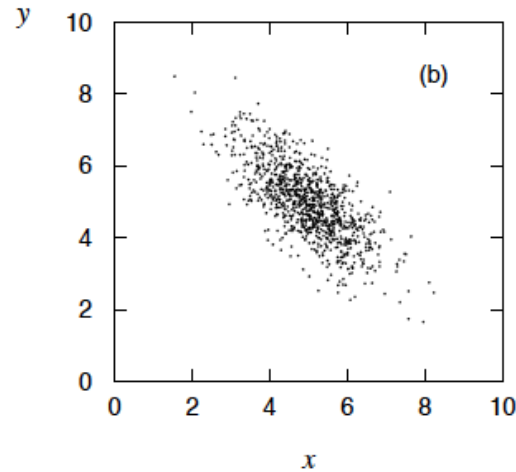
N.B. converse not always true.

Correlation (cont.)

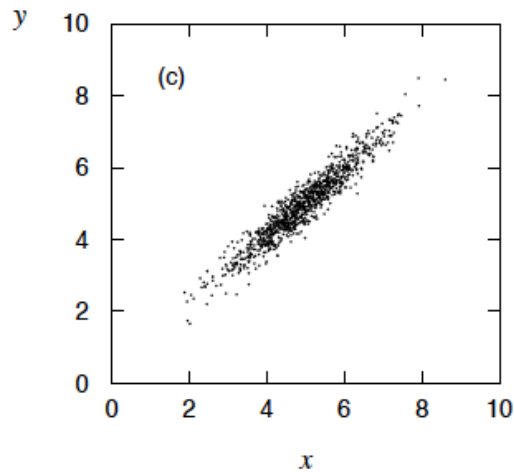
$$\rho = 0.75$$



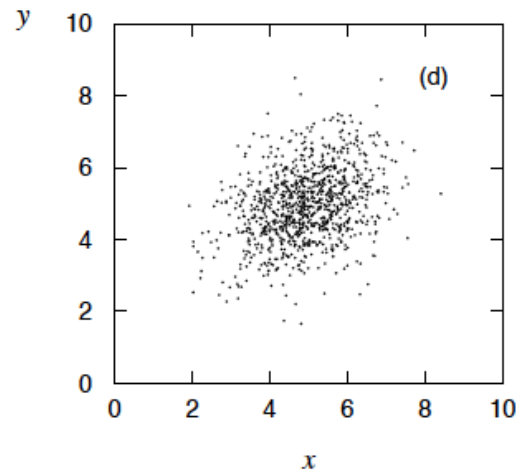
$$\rho = -0.75$$



$$\rho = 0.95$$



$$\rho = 0.25$$



Hypotheses

A hypothesis H specifies the probability for the data, i.e., the outcome of the observation, here symbolically: x .

x could be uni-/multivariate, continuous or discrete.

E.g. write $x \sim f(x|H)$.

x could represent e.g. observation of a single particle, a single event, or an entire “experiment”.

Possible values of x form the sample space S (or “data space”).

Simple (or “point”) hypothesis: $f(x|H)$ completely specified.

Composite hypothesis: H contains unspecified parameter(s).

The probability for x given H is also called the likelihood of the hypothesis, written $L(x|H)$.

Frequentist hypothesis tests

Consider a hypothesis H_0 and alternative H_1 .

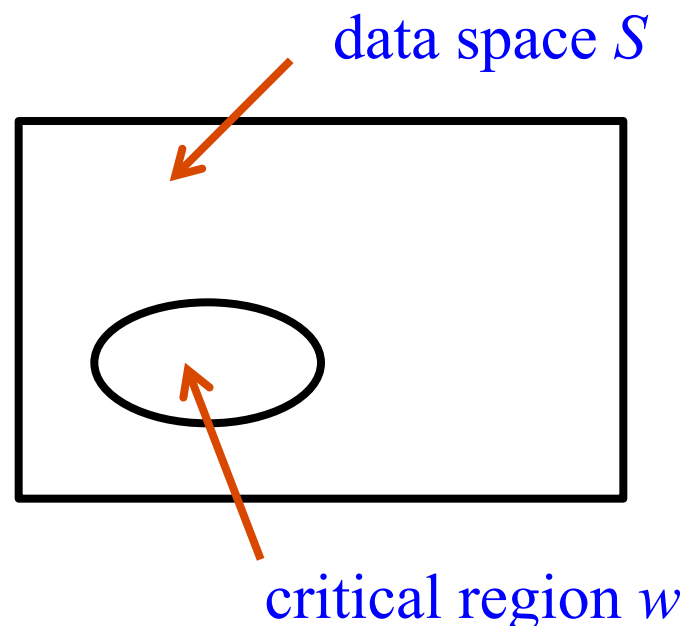
A **test** of H_0 is defined by specifying a **critical region** w of the data space such that there is no more than some (small) probability α , assuming H_0 is correct, to observe the data there, i.e.,

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

Need inequality if data are discrete.

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

If x is observed in the critical region, reject H_0 .

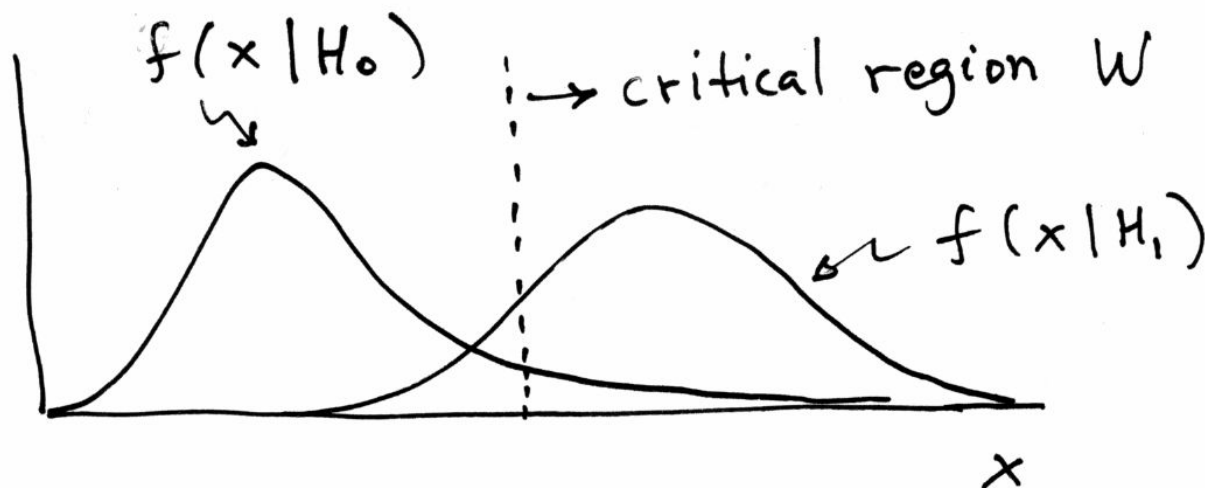


Definition of a test (2)

But in general there are an infinite number of possible critical regions that give the same significance level α .

So the choice of the critical region for a test of H_0 needs to take into account the alternative hypothesis H_1 .

Roughly speaking, place the critical region where there is a low probability to be found if H_0 is true, but high if H_1 is true:



Type-I, Type-II errors

Rejecting the hypothesis H_0 when it is true is a Type-I error.

The maximum probability for this is the size of the test:

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

But we might also accept H_0 when it is false, and an alternative H_1 is true.

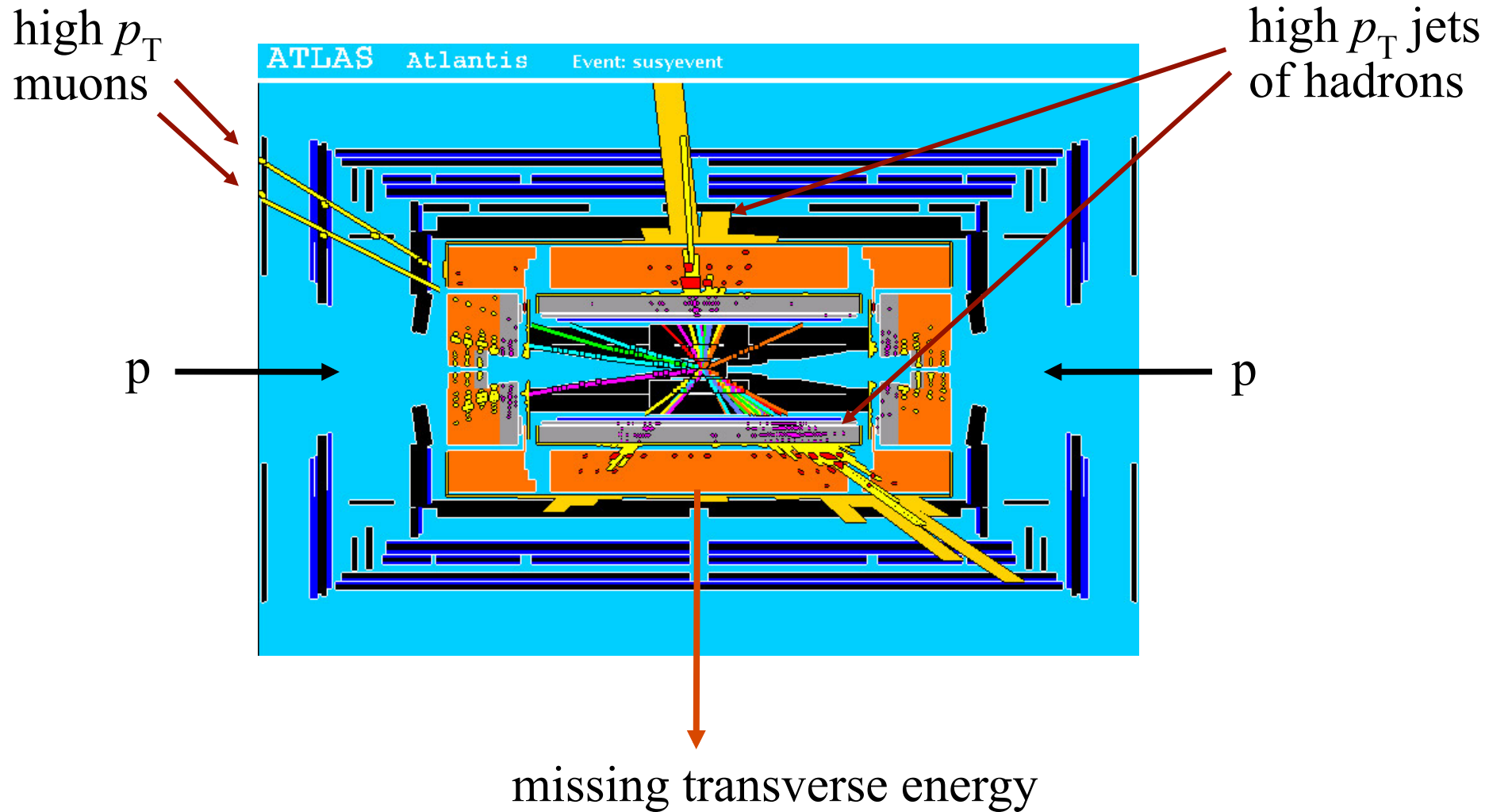
This is called a Type-II error, and occurs with probability

$$P(x \in S - W \mid H_1) = \beta$$

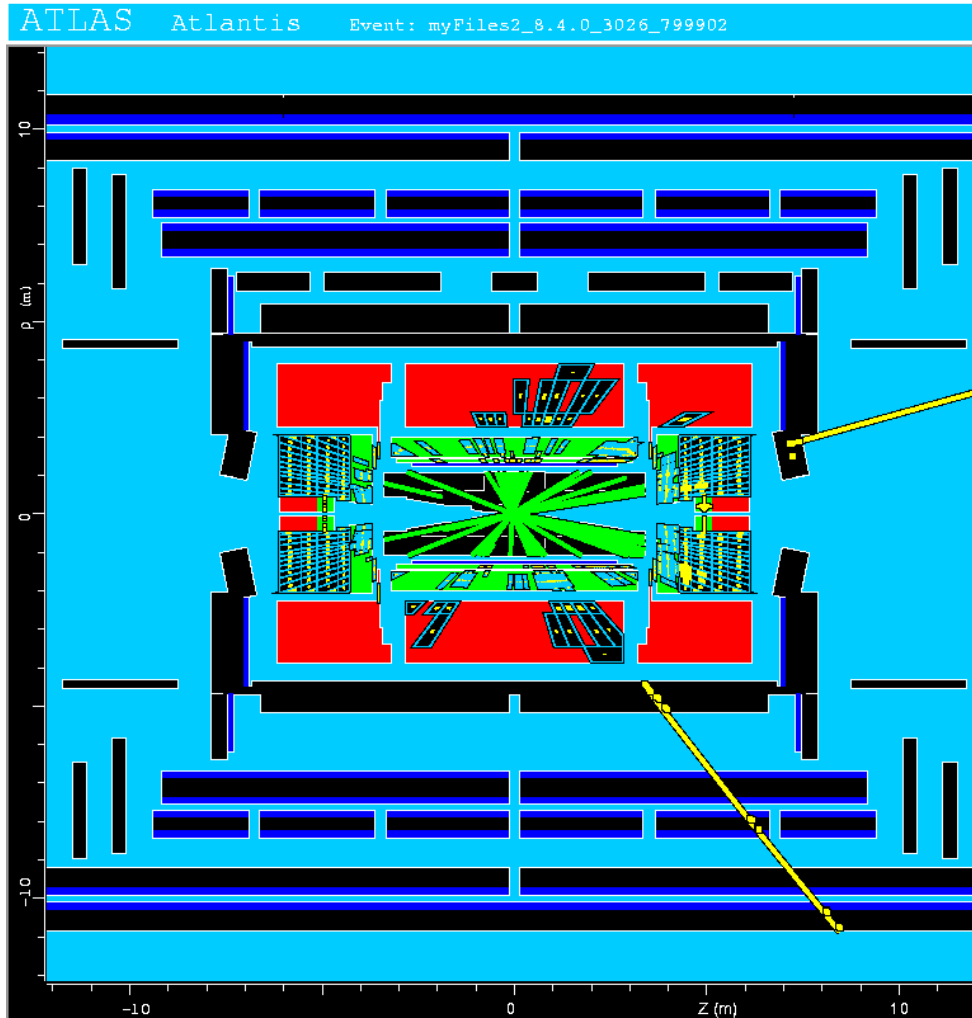
One minus this is called the power of the test with respect to the alternative H_1 :

$$\text{Power} = 1 - \beta$$

A simulated SUSY event



Background events



This event from Standard Model $t\bar{t}$ production also has high p_T jets and muons, and some missing transverse energy.

→ can easily mimic a SUSY event.

Physics context of a statistical test

Event Selection: the event types in question are both known to exist.

Example: separation of different particle types (electron vs muon) or known event types (ttbar vs QCD multijet).

E.g. test H_0 : event is background vs. H_1 : event is signal.

Use selected events for further study.

Search for New Physics: the null hypothesis is

H_0 : all events correspond to Standard Model (background only),

and the alternative is

H_1 : events include a type whose existence is not yet established (signal plus background)

Many subtle issues here, mainly related to the high standard of proof required to establish presence of a new phenomenon. The optimal statistical test for a search is closely related to that used for event selection.

Statistical tests for event selection

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

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

$$pp \rightarrow t\bar{t} , \quad pp \rightarrow \tilde{g}\tilde{g} , \dots$$

For each reaction we consider we will have a hypothesis for the pdf of \vec{x} , e.g., $f(\vec{x}|H_0)$, $f(\vec{x}|H_1)$, etc.

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

Selecting events

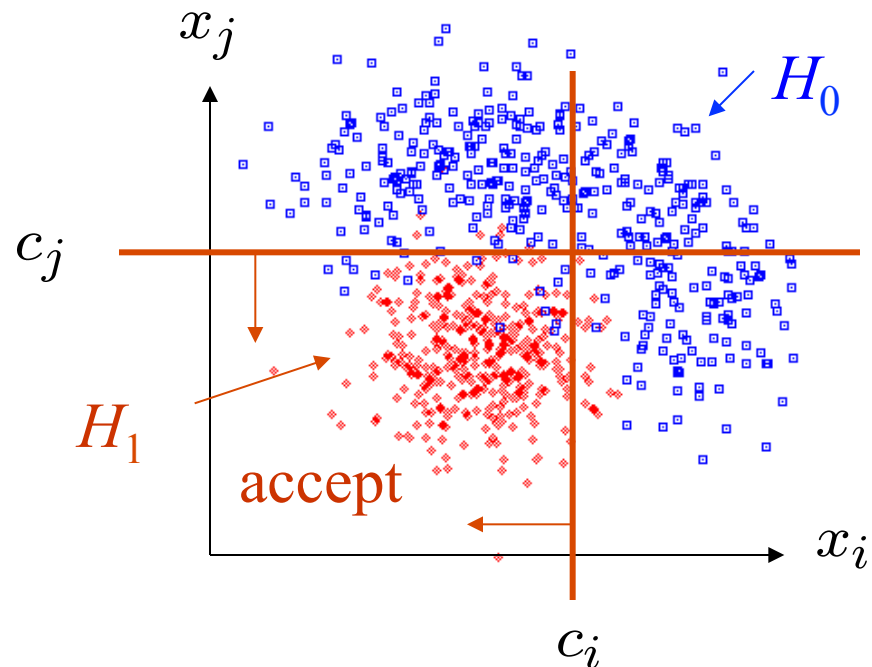
Suppose we have a data sample with two kinds of events, corresponding to hypotheses H_0 and H_1 and we want to select those of type H_1 .

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’:

$$x_i < c_i$$

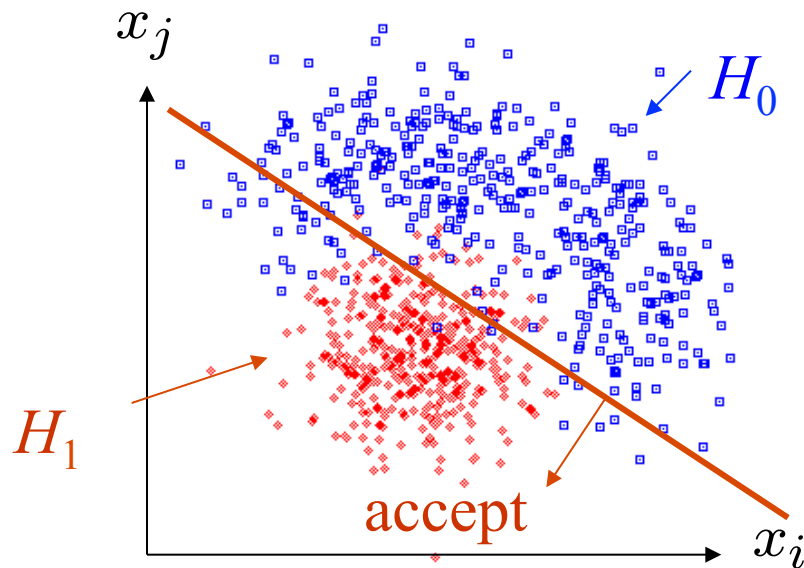
$$x_j < c_j$$



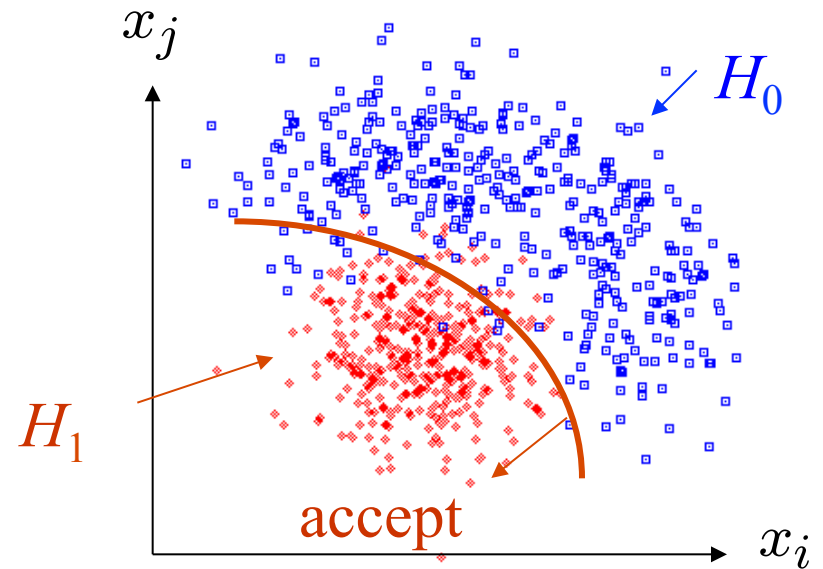
Other ways to select events

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 $\mathbf{x} = (x_1, \dots, x_n)$ can be defined by an equation of the form

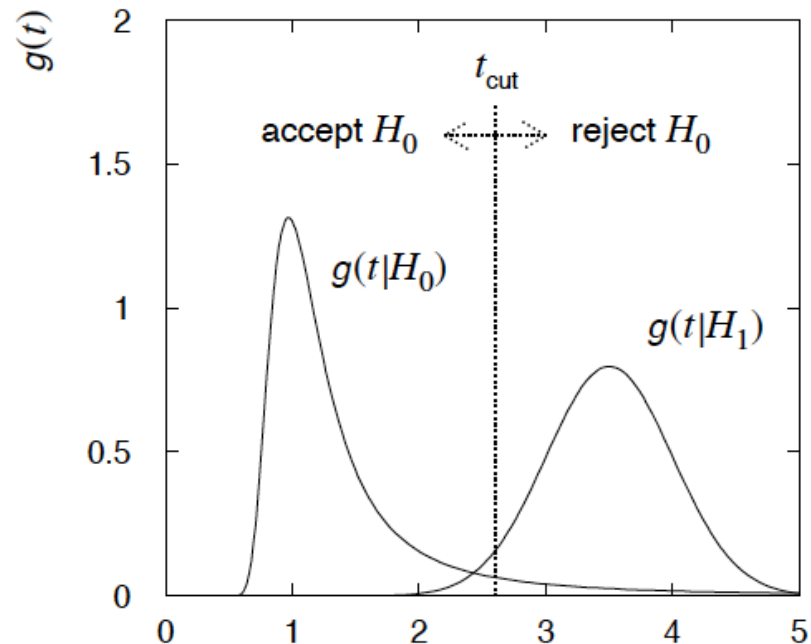
$$t(x_1, \dots, x_n) = t_{\text{cut}}$$

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

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

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 leads to the same test.

Classification viewed as a statistical test

Probability to reject H_0 if true (type I error): $\alpha = \int_W f(\mathbf{x}|H_0)d\mathbf{x}$

α = size of test, significance level, false discovery rate

Probability to accept H_0 if H_1 true (type II error) $\beta = \int_{\bar{W}} f(\mathbf{x}|H_1)d\mathbf{x}$

$1 - \beta$ = power of test with respect to H_1

Equivalently if e.g. H_0 = background, H_1 = signal, use efficiencies:

$$\varepsilon_b = \int_W f(\mathbf{x}|H_0) = \alpha$$

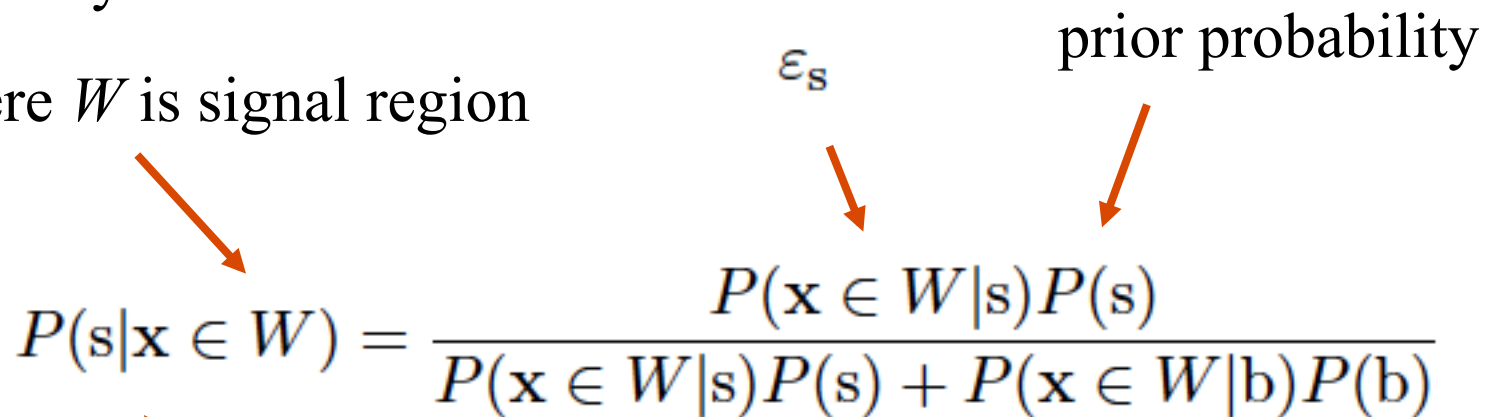
$$\varepsilon_s = \int_W f(\mathbf{x}|H_1) = 1 - \beta = \text{power}$$

Purity / misclassification rate

Consider the probability that an event of signal (s) type classified correctly (i.e., the event selection purity),

Use Bayes' theorem:

Here W is signal region

$$P(s|\mathbf{x} \in W) = \frac{P(\mathbf{x} \in W|s)P(s)}{P(\mathbf{x} \in W|s)P(s) + P(\mathbf{x} \in W|b)P(b)}$$


posterior probability = signal purity
= 1 – signal misclassification rate

Note purity depends on the prior probability for an event to be signal or background as well as on s/b efficiencies.

Neyman-Pearson doesn't usually help

We usually don't have explicit formulae for the pdfs $f(\mathbf{x}|s)$, $f(\mathbf{x}|b)$, so for a given \mathbf{x} we can't evaluate the likelihood ratio

$$t(\mathbf{x}) = \frac{f(\mathbf{x}|s)}{f(\mathbf{x}|b)}$$

Instead we may have Monte Carlo models for signal and background processes, so we can produce simulated data:

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

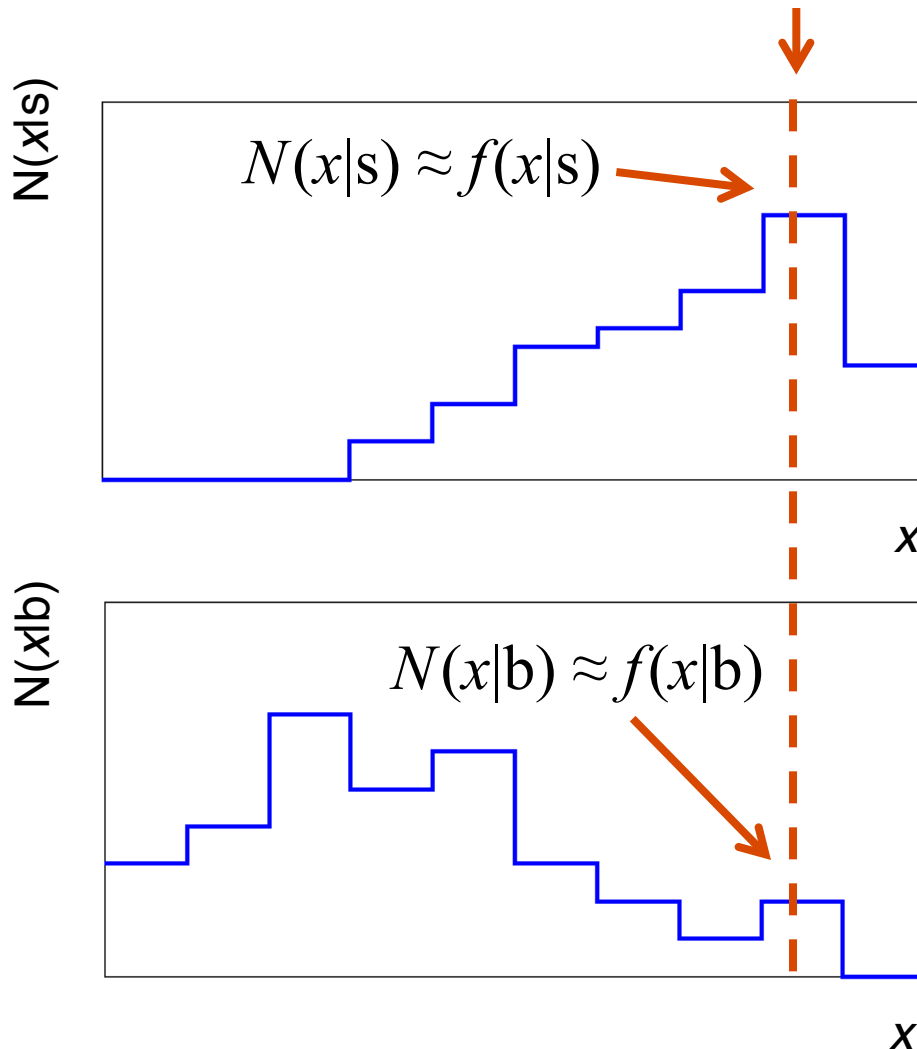
generate $\mathbf{x} \sim f(\mathbf{x}|b)$ $\rightarrow \mathbf{x}_1, \dots, \mathbf{x}_N$

This gives samples of “training data” with events of known type.

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

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.

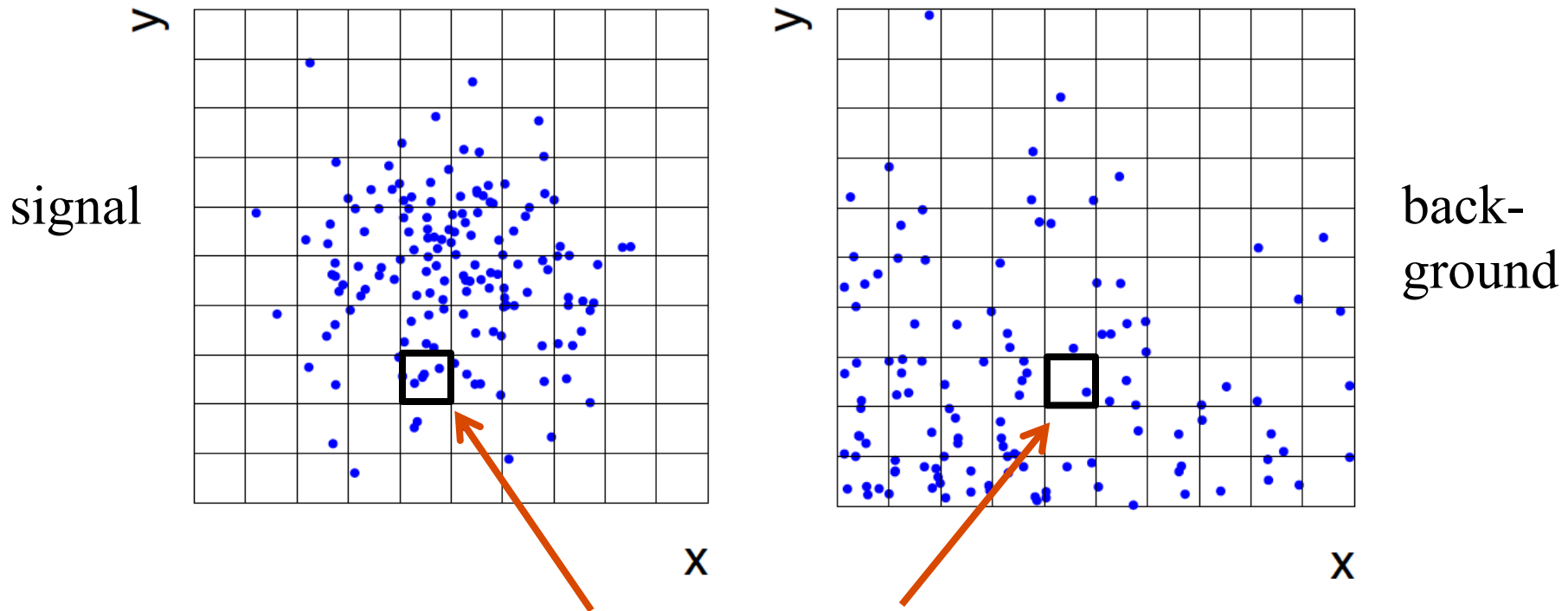
Use (normalized) histogram values to approximate LR:

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

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

Many new (and some old) methods:

Fisher discriminant

(Deep) neural networks

Kernel density methods

Support Vector Machines

Decision trees

Boosting

Bagging

Much more on this in the lectures by Lorenzo Moneta

Resources on multivariate methods

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

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

R. Duda, P. Hart, D. Stork, Pattern Classification, 2nd ed., Wiley, 2001

A. Webb, Statistical Pattern Recognition, 2nd ed., Wiley, 2002.

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

朱永生（编著），实验数据多元统计分析，科学出版社，北京，2009。

Software

Rapidly growing area of development – two important resources:

TMVA, Höcker, Stelzer, Tegenfeldt, Voss, Voss, [physics/0703039](#)

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

Variety of classifiers

Good manual, widely used in HEP

scikit-learn

Python-based tools for Machine Learning

`scikit-learn.org`

Large user community

Testing significance / goodness-of-fit

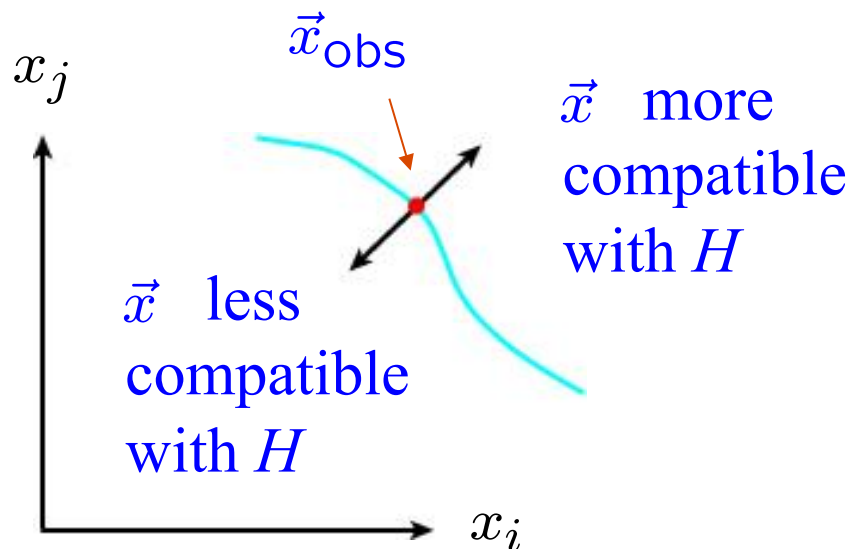
Suppose hypothesis H predicts pdf $f(\vec{x}|H)$ for a set of observations $\vec{x} = (x_1, \dots, x_n)$.

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

Note – “less compatible with H ” means “more compatible 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 $P(H)$.

Distribution of the p -value

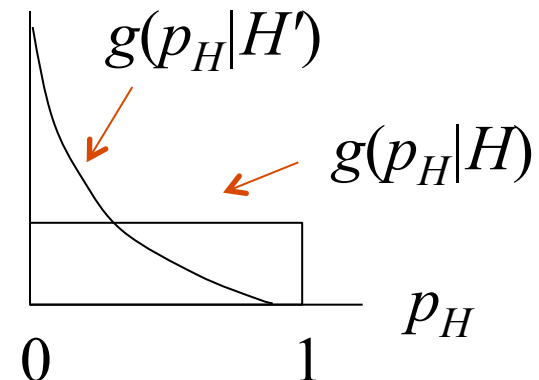
The p -value is a function of the data, and is thus itself a random variable with a given distribution. Suppose the p -value of H is found from a test statistic $t(\mathbf{x})$ as

$$p_H = \int_t^\infty f(t'|H)dt'$$

The pdf of p_H under assumption of H is

$$g(p_H|H) = \frac{f(t|H)}{|\partial p_H / \partial t|} = \frac{f(t|H)}{f(t|H)} = 1 \quad (0 \leq p_H \leq 1)$$

In general for continuous data, under assumption of H , $p_H \sim \text{Uniform}[0,1]$ and is concentrated toward zero for Some class of relevant alternatives.



Using a p -value to define test of H_0

One can show the distribution of the p -value of H , under assumption of H , is uniform in $[0,1]$.

So the probability to find the p -value of H_0 , p_0 , less than α is

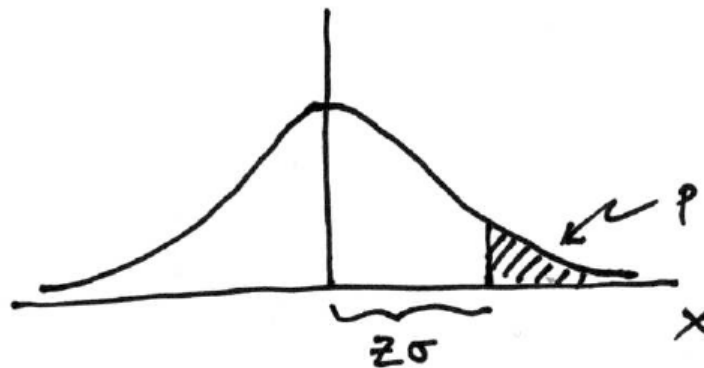
$$P(p_0 \leq \alpha | H_0) = \alpha$$

We can define the critical region of a test of H_0 with size α as the set of data space where $p_0 \leq \alpha$.

Formally the p -value relates only to H_0 , but the resulting test will have a given power with respect to a given alternative H_1 .

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.



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

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

E.g. $Z = 5$ (a “5 sigma effect”) corresponds to $p = 2.9 \times 10^{-7}$.

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.

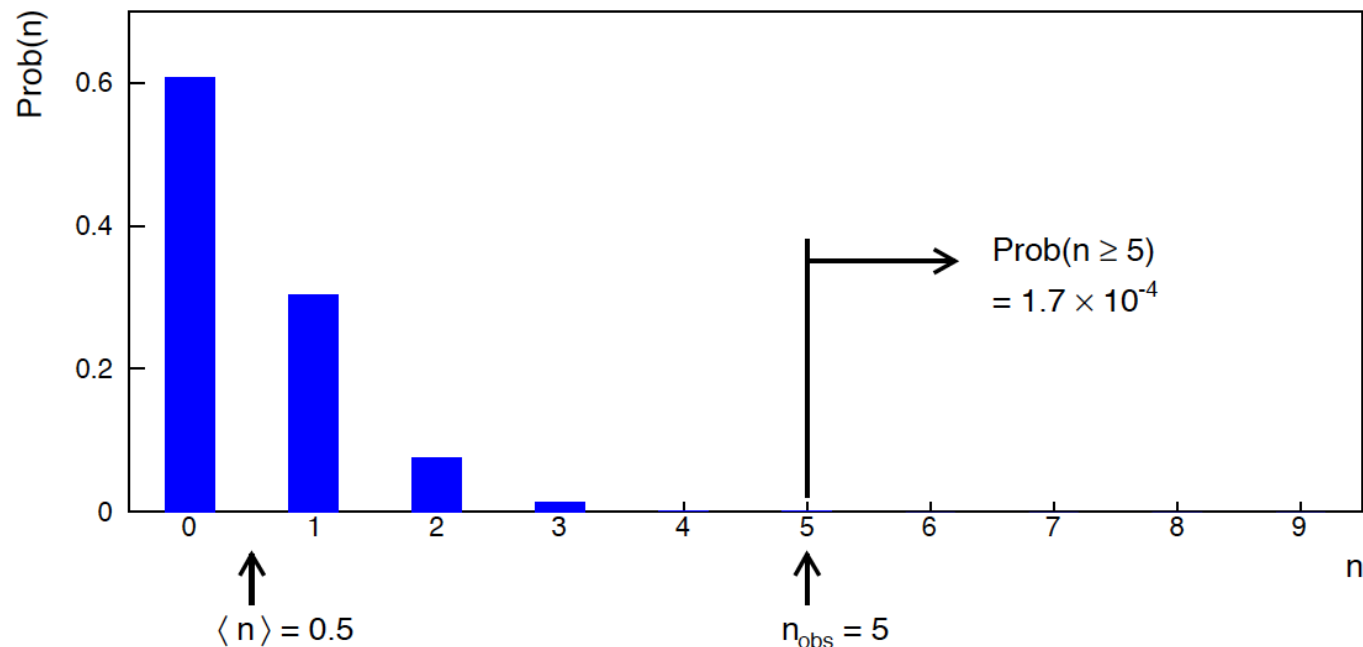
Poisson counting experiment: discovery p -value

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 $s = 0$:

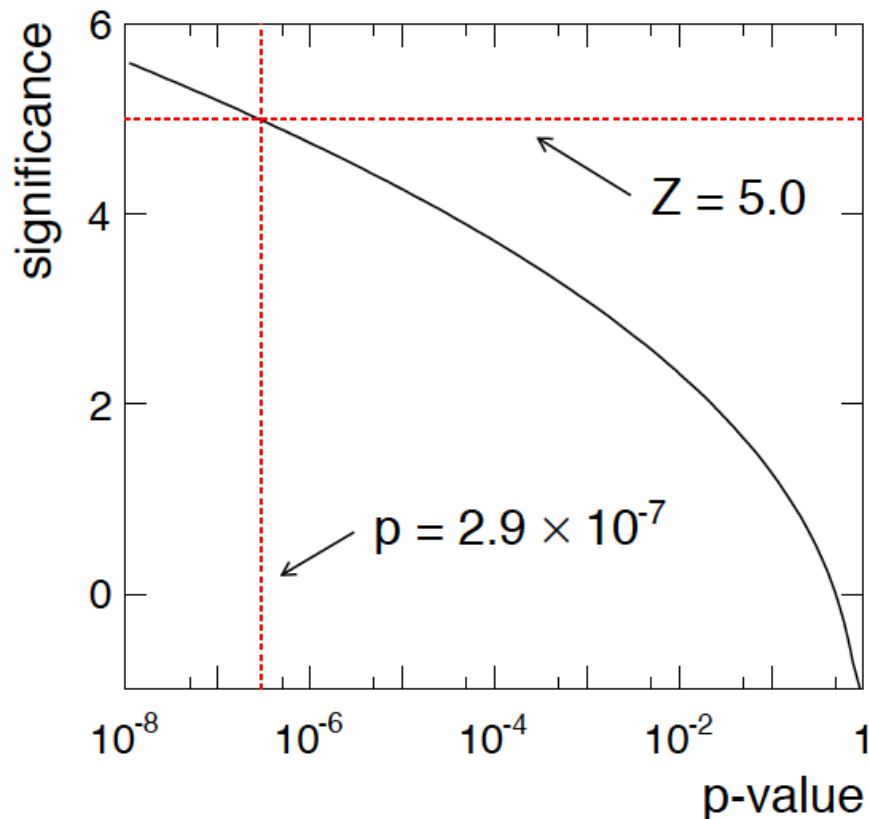
$$\begin{aligned} p\text{-value} &= P(n \geq 5; b = 0.5, s = 0) \\ &= 1.7 \times 10^{-4} \neq P(s = 0)! \end{aligned}$$



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”)



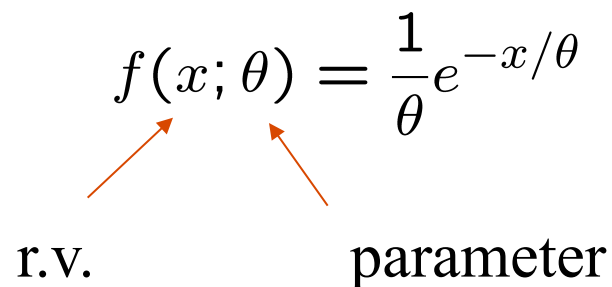
In fact this tradition should be revisited: p -value intended to quantify probability of a signal-like fluctuation assuming background only; not intended to cover, e.g., hidden systematics, plausibility signal model, compatibility of data with signal, “look-elsewhere effect” (~multiple testing), etc.

Parameter estimation

The parameters of a pdf are constants that characterize its shape, e.g.

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

r.v. parameter



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

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

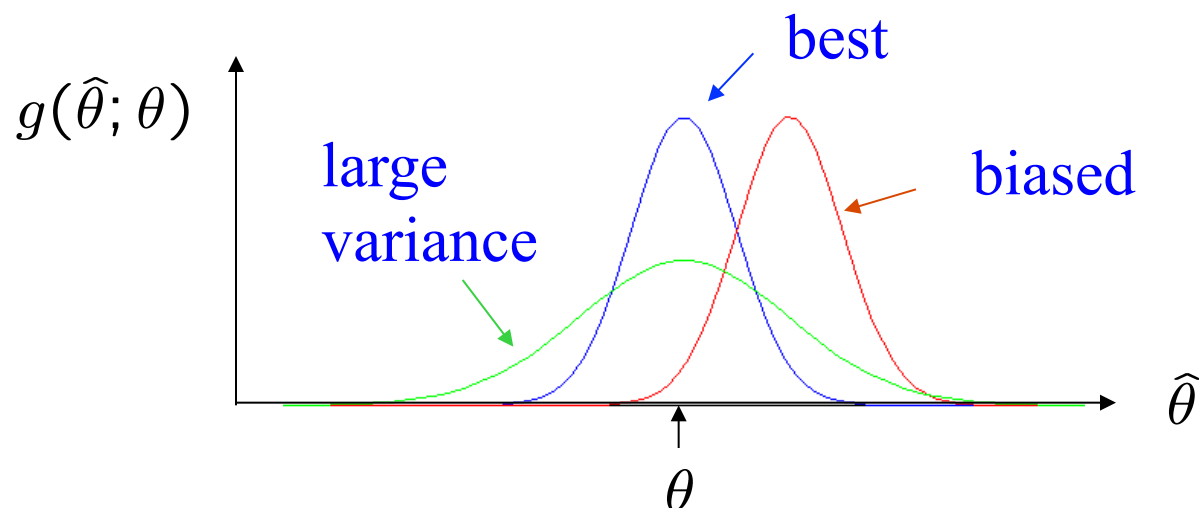
$\hat{\theta}(\vec{x})$

← estimator written with a hat

Sometimes we say ‘estimator’ for the function of x_1, \dots, 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): $b = E[\hat{\theta}] - \theta$

→ average of repeated measurements should tend to true value.

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

→ small bias & variance are in general conflicting criteria

An estimator for the mean (expectation value)

Parameter: $\mu = E[x] = \langle x \rangle = \int_{-\infty}^{\infty} x f(x) dx$

Estimator: $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i \equiv \bar{x}$ ('sample mean')

We find: $b = E[\hat{\mu}] - \mu = 0$

$$V[\hat{\mu}] = \frac{\sigma^2}{n} \quad \left(\sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}} \right)$$

An estimator for the variance

Parameter: $\sigma^2 = V[x] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$

Estimator: $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \equiv s^2$ ('sample variance')

We find:

$$b = E[\hat{\sigma}^2] - \sigma^2 = 0 \quad (\text{factor of } n-1 \text{ makes this so})$$

$$V[\hat{\sigma}^2] = \frac{1}{n} \left(\mu_4 - \frac{n-3}{n-1} \mu_2 \right), \quad \text{where}$$

$$\mu_k = \int (x - \mu)^k f(x) dx$$

The likelihood function

Suppose the entire result of an experiment (set of measurements) is a collection of numbers \mathbf{x} , and suppose the joint pdf for the data \mathbf{x} is a function that depends on a set of parameters θ :

$$P(\mathbf{x}|\theta)$$

Now evaluate this function with the data obtained and regard it as a function of the parameter(s). This is the **likelihood function**:

$$L(\theta) = P(\mathbf{x}|\theta)$$

(\mathbf{x} constant)

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

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

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

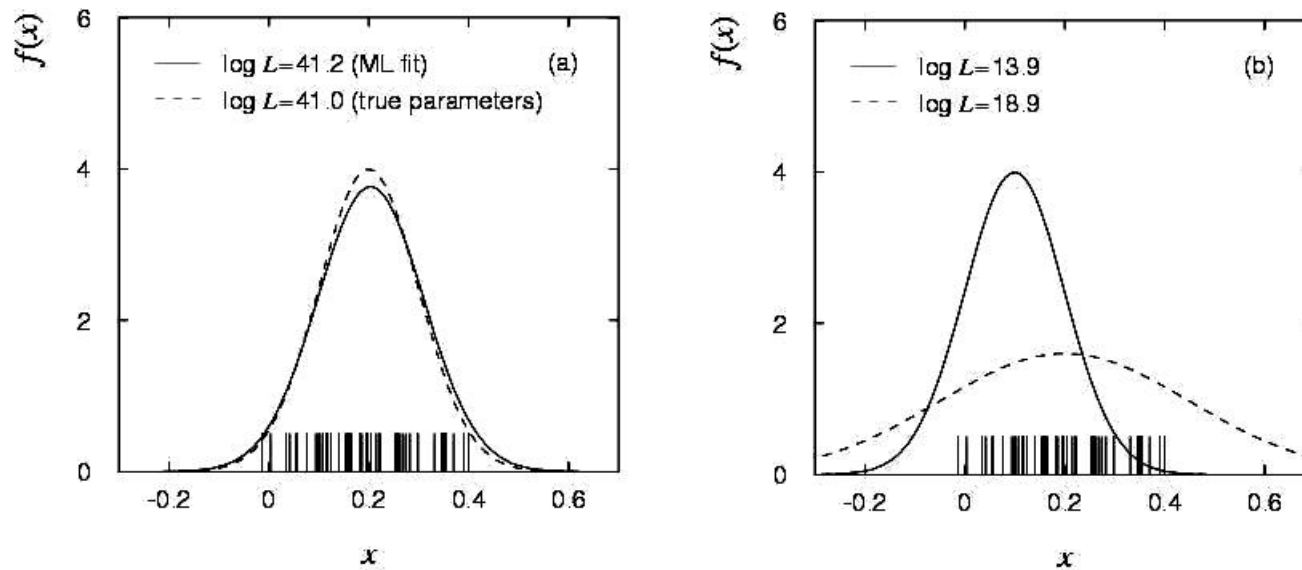
$$f(x_1, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i; \theta)$$

In this case the likelihood function is

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

Maximum likelihood estimators

If the hypothesized θ is close to the true value, then we expect a high probability to get data like that which we actually found.



So we define the maximum likelihood (ML) estimator(s) to be the parameter value(s) for which the likelihood is maximum.

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

ML 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, \dots, 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):

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

ML example: parameter of exponential pdf (2)

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

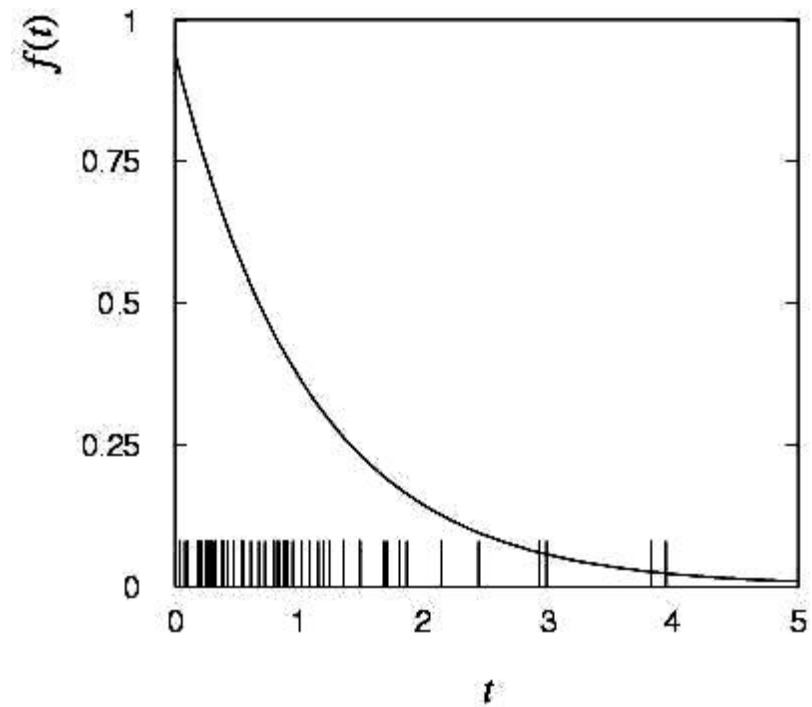
$$\rightarrow \hat{\tau} = \frac{1}{n} \sum_{i=1}^n t_i$$

Monte Carlo test:

generate 50 values
using $\tau = 1$:

We find the ML estimate:

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



ML 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 ML estimator $\hat{\tau} = \frac{1}{n} \sum_{i=1}^n t_i$ we therefore find

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

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

Functions of ML estimators

Suppose we had written the exponential pdf as $f(t; \lambda) = \lambda e^{-\lambda t}$, i.e., we use $\lambda = 1/\tau$. What is the ML estimator for λ ?

For a function (with unique inverse) $\lambda(\tau)$ of a parameter τ , it doesn't matter whether we express L as a function of λ or τ .

The ML estimator of a function $\lambda(\tau)$ is simply $\hat{\lambda} = \lambda(\hat{\tau})$

So for the decay constant we have $\hat{\lambda} = \frac{1}{\hat{\tau}} = \left(\frac{1}{n} \sum_{i=1}^n t_i \right)^{-1}$.

Caveat: $\hat{\lambda}$ is biased, even though $\hat{\tau}$ is unbiased.

Can show $E[\hat{\lambda}] = \lambda \frac{n}{n-1}$. (bias $\rightarrow 0$ for $n \rightarrow \infty$)

Example of ML: parameters of Gaussian pdf

Consider independent x_1, \dots, x_n , with $x_i \sim \text{Gaussian}(\mu, \sigma^2)$

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

The log-likelihood function is

$$\begin{aligned} \ln L(\mu, \sigma^2) &= \sum_{i=1}^n \ln f(x_i; \mu, \sigma^2) \\ &= \sum_{i=1}^n \left(\ln \frac{1}{\sqrt{2\pi}} + \frac{1}{2} \ln \frac{1}{\sigma^2} - \frac{(x_i - \mu)^2}{2\sigma^2} \right) . \end{aligned}$$

Example of ML: parameters of Gaussian pdf (2)

Set derivatives with respect to μ , σ^2 to zero and solve,

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2.$$

We already know that the estimator for μ is unbiased.

But we find, however, $E[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2$, so ML estimator for σ^2 has a bias, but $b \rightarrow 0$ for $n \rightarrow \infty$. Recall, however, that

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \hat{\mu})^2$$

is an unbiased estimator for σ^2 .

Variance of estimators: Monte Carlo method

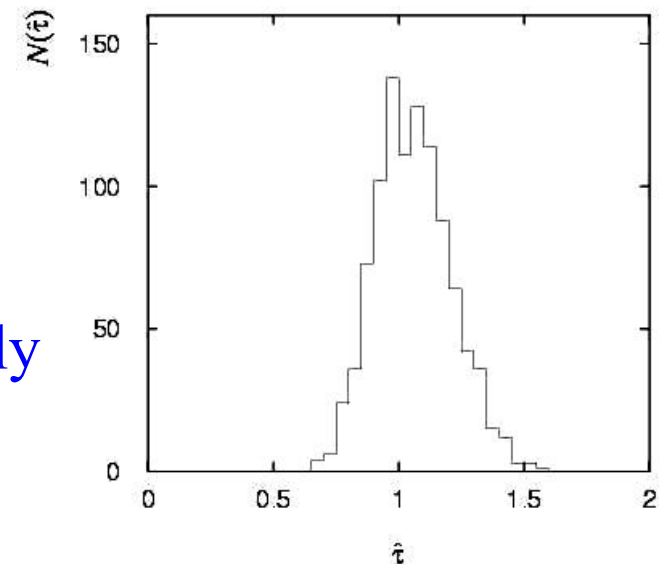
Having estimated our parameter we now need to report its ‘statistical error’, i.e., how widely distributed would estimates be if we were to repeat the entire measurement many times.

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:

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


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):

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

 Minimum Variance Bound (MVB)
($b = E[\hat{\theta}] - \theta$)

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

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

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

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

Variance of estimators: graphical method

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

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

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

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

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

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

Example of variance by graphical method

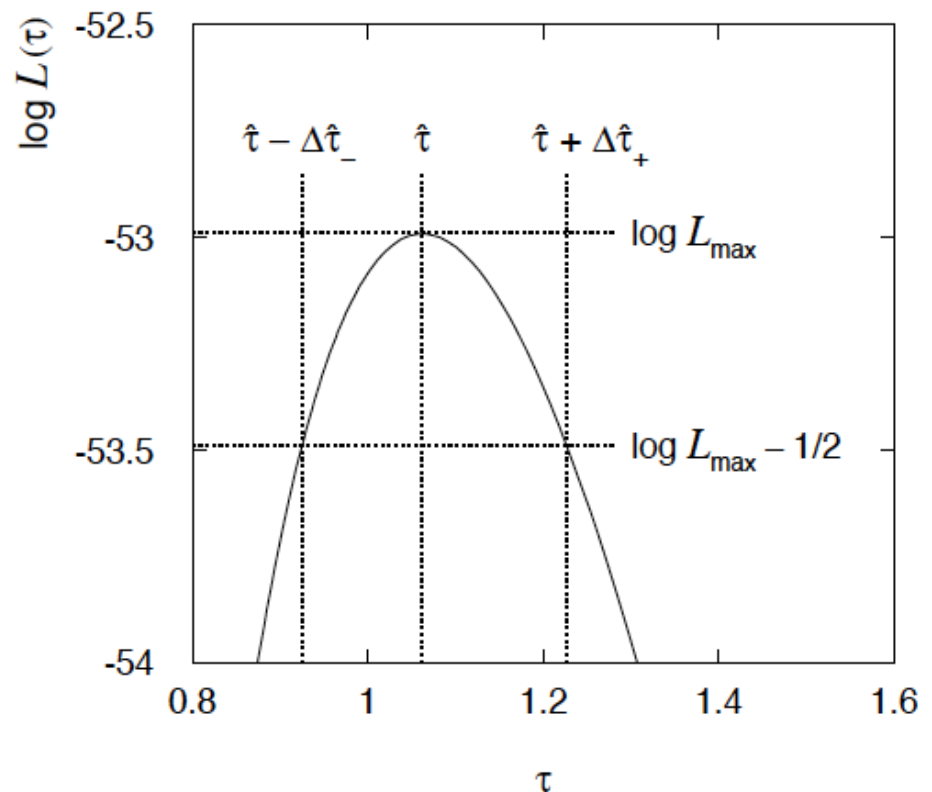
ML example with exponential:

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

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

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

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



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

Information inequality for n parameters

Suppose we have estimated n parameters $\vec{\theta} = (\theta_1, \dots, \theta_n)$.

The (inverse) minimum variance bound is given by the Fisher information matrix:

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

The information inequality then states that $V - I^{-1}$ is a positive semi-definite matrix, where $V_{ij} = \text{cov}[\hat{\theta}_i, \hat{\theta}_j]$. Therefore

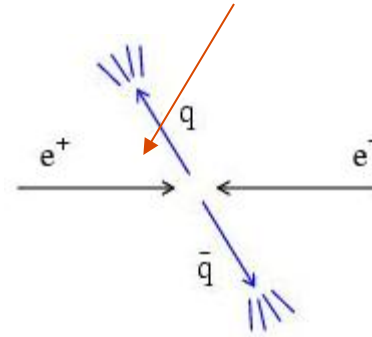
$$V[\hat{\theta}_i] \geq (I^{-1})_{ii}$$

Often use I^{-1} as an approximation for covariance matrix, estimate using e.g. matrix of 2nd derivatives at maximum of L .

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 always 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.

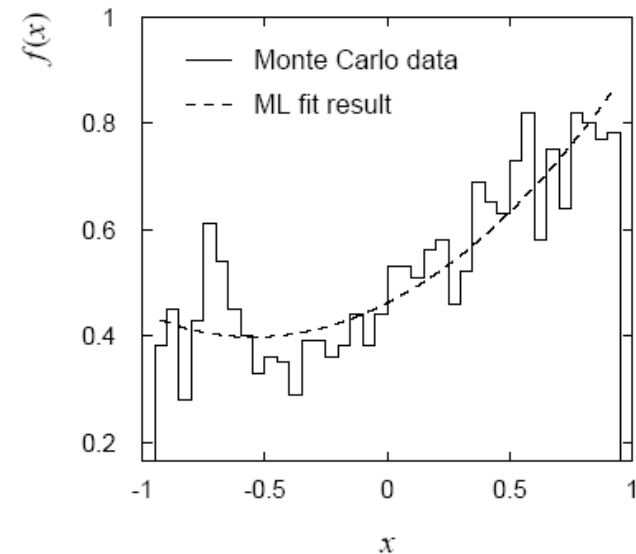
Example of ML with 2 parameters: fit result

Finding maximum of $\ln L(\alpha, \beta)$ numerically (**MINUIT**) 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).



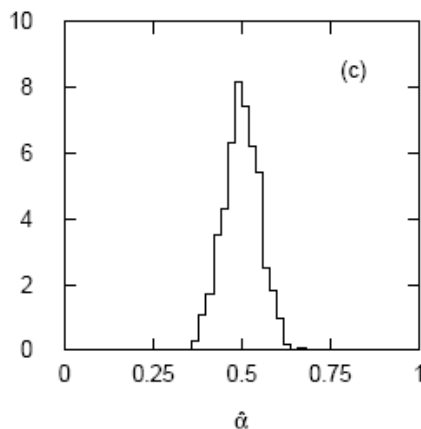
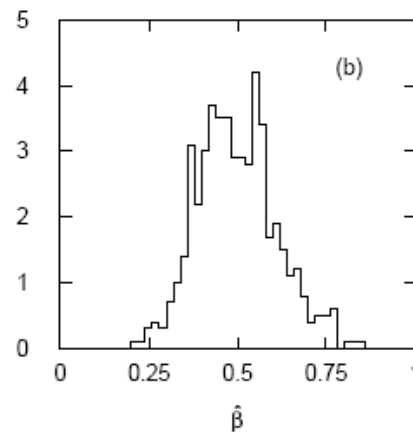
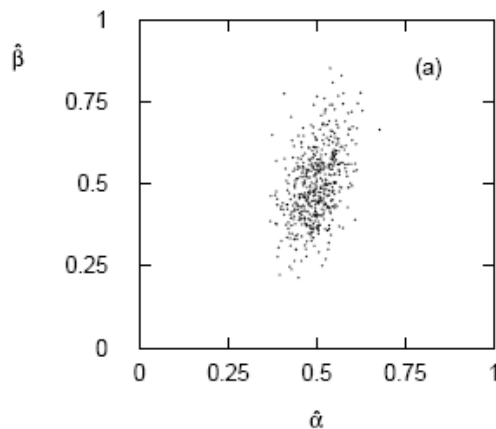
(Co)variances from $(\widehat{V}^{-1})_{ij} = -\left. \frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j} \right|_{\vec{\theta}=\vec{\hat{\theta}}}$ (**MINUIT** routine **HESSE**)

$$\hat{\sigma}_{\hat{\alpha}} = 0.052 \quad \text{cov}[\hat{\alpha}, \hat{\beta}] = 0.0026$$

$$\hat{\sigma}_{\hat{\beta}} = 0.11 \quad r = 0.46$$

Two-parameter fit: MC study

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



$$\overline{\hat{\alpha}} = 0.499$$

$$s_{\hat{\alpha}} = 0.051$$

$$\overline{\hat{\beta}} = 0.498$$

$$s_{\hat{\beta}} = 0.111$$

$$\widehat{\text{cov}}[\hat{\alpha}, \hat{\beta}] = 0.0024$$

$$r = 0.42$$

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

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

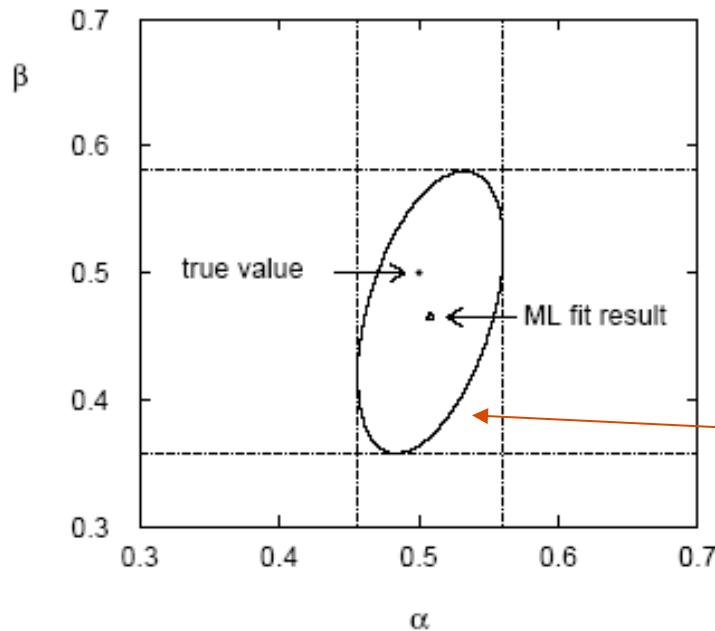
For large n , $\ln L$ takes on quadratic form near maximum:

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

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 α, β 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: $\tan 2\phi = \frac{2\rho\sigma_{\hat{\alpha}}\sigma_{\hat{\beta}}}{\sigma_{\hat{\alpha}}^2 - \sigma_{\hat{\beta}}^2}$

Correlations between estimators result in an increase in their standard deviations (statistical errors).

ML with binned data

Often put data into a histogram: $\vec{n} = (n_1, \dots, n_N)$, $n_{\text{tot}} = \sum_{i=1}^N n_i$

Hypothesis is $\vec{\nu} = (\nu_1, \dots, \nu_N)$, $\nu_{\text{tot}} = \sum_{i=1}^N \nu_i$ where

$$\nu_i(\vec{\theta}) = \nu_{\text{tot}} \int_{\text{bin } i} f(x; \vec{\theta}) dx$$

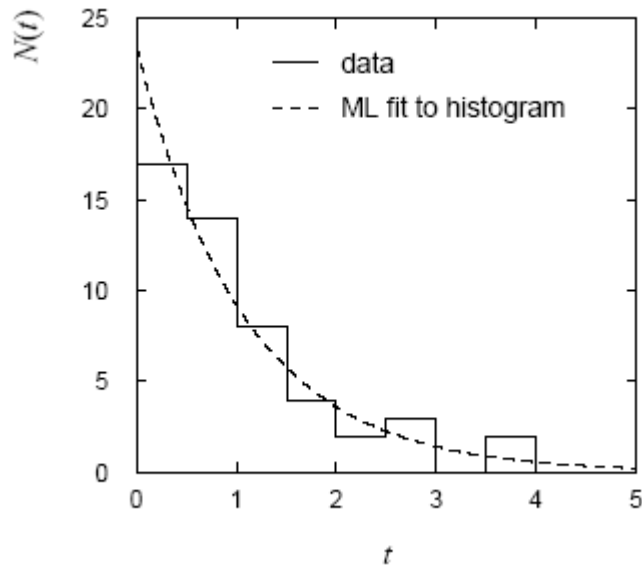
If we model the data as multinomial (n_{tot} constant),

$$f(\vec{n}; \vec{\nu}) = \frac{n_{\text{tot}}!}{n_1! \dots n_N!} \left(\frac{\nu_1}{n_{\text{tot}}} \right)^{n_1} \dots \left(\frac{\nu_N}{n_{\text{tot}}} \right)^{n_N}$$

then the log-likelihood function is: $\ln L(\vec{\theta}) = \sum_{i=1}^N n_i \ln \nu_i(\vec{\theta}) + C$

ML example with binned data

Previous example with exponential, now put data into histogram:



$$\hat{\tau} = 1.07 \pm 0.17$$

(1.06 ± 0.15 for unbinned
ML with same sample)

Limit of zero bin width \rightarrow usual unbinned ML.

If n_i treated as Poisson, we get extended log-likelihood:

$$\ln L(\nu_{\text{tot}}, \vec{\theta}) = -\nu_{\text{tot}} + \sum_{i=1}^N n_i \ln \nu_i(\nu_{\text{tot}}, \vec{\theta}) + C$$


Relationship between ML and Bayesian estimators

In Bayesian statistics, both θ and \mathbf{x} are random variables:

$$L(\theta) = L(\vec{x}|\theta) = f_{\text{joint}}(\vec{x}|\theta)$$

Recall the Bayesian method:

Use subjective probability for hypotheses (θ);
before experiment, knowledge summarized by prior pdf $\pi(\theta)$;
use Bayes' theorem to update prior in light of data:


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

Posterior pdf (conditional pdf for θ given \mathbf{x})

ML and Bayesian estimators (2)

Purist Bayesian: $p(\theta | x)$ contains all knowledge about θ .

Pragmatist Bayesian: $p(\theta | x)$ could be a complicated function,

→ summarize using an estimator $\hat{\theta}_{\text{Bayes}}$

Take mode of $p(\theta | x)$, (could also use e.g. expectation value)

What do we use for $\pi(\theta)$? No golden rule (subjective!), often represent ‘prior ignorance’ by $\pi(\theta) = \text{constant}$, in which case

$$\hat{\theta}_{\text{Bayes}} = \hat{\theta}_{\text{ML}}$$

But... we could have used a different parameter, e.g., $\lambda = 1/\theta$, and if prior $\pi_{\theta}(\theta)$ is constant, then $\pi_{\lambda}(\lambda) = \pi_{\theta}(\theta(\lambda)) |d\theta/d\lambda|$ is not!

‘Complete prior ignorance’ is not well defined.

Priors from formal rules

Because of difficulties in encoding a vague degree of belief in a prior, one often attempts to derive the prior from formal rules, e.g., to satisfy certain invariance principles or to provide maximum information gain for a certain set of measurements.

Often called “objective priors”

Form basis of Objective Bayesian Statistics

The priors do not reflect a degree of belief (but might represent possible extreme cases).

In a Subjective Bayesian analysis, using objective priors can be an important part of the sensitivity analysis.

Priors from formal rules (cont.)

In Objective Bayesian analysis, can use the intervals in a frequentist way, i.e., regard Bayes' theorem as a recipe to produce an interval with certain coverage properties. For a review see:

Robert E. Kass and Larry Wasserman, *The Selection of Prior Distributions by Formal Rules*, J. Am. Stat. Assoc., Vol. 91, No. 435, pp. 1343-1370 (1996).

Formal priors have not been widely used in HEP, but there is recent interest in this direction; see e.g.

L. Demortier, S. Jain and H. Prosper, *Reference priors for high energy physics*, arxiv:1002.1111 (Feb 2010)

Jeffreys' prior

According to *Jeffreys' rule*, take prior according to

$$\pi(\boldsymbol{\theta}) \propto \sqrt{\det(I(\boldsymbol{\theta}))}$$

where

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

is the Fisher information matrix.

One can show that this leads to inference that is invariant under a transformation of parameters.

For a Gaussian mean, the Jeffreys' prior is constant; for a Poisson mean μ it is proportional to $1/\sqrt{\mu}$.

“Invariance of inference” with Jeffreys’ prior

Suppose we have a parameter θ , to which we assign a prior $\pi_\theta(\theta)$.

An experiment gives data x , modeled by $L(\theta) = P(x|\theta)$.

Bayes’ theorem then tells us the posterior for θ :

$$P(\theta|x) \propto P(x|\theta)\pi_\theta(\theta)$$

Now consider a function $\eta(\theta)$, and we want the posterior $P(\eta|x)$.

This must follow from the usual rules of transformation of random variables:

$$P(\eta|x) = P(\theta(\eta)|x) \left| \frac{d\theta}{d\eta} \right|$$

“Invariance of inference” with Jeffreys’ prior (2)

Alternatively, we could have just starting with η as the parameter in our model, and written down a prior pdf $\pi_\eta(\eta)$.

Using it, we express the likelihood as $L(\eta) = P(x|\eta)$ and write Bayes’ theorem as

$$P(\eta|x) \propto P(x|\eta)\pi_\eta(\eta)$$

If the priors really express our degree of belief, then they must be related by the usual laws of probability $\pi_\eta(\eta) = \pi_\theta(\theta(\eta)) |d\theta/d\eta|$, and in this way the two approaches lead to the same result.

But if we choose the priors according to “formal rules”, then this is not guaranteed. For the Jeffrey’s prior, however, it does work!

Using $\pi_\theta(\theta) \propto \sqrt{I(\theta)}$ and transforming to find $P(\eta|x)$ leads to the same as using $\pi_\eta(\eta) \propto \sqrt{I(\eta)}$ directly with Bayes’ theorem.

Jeffreys' prior for Poisson mean

Suppose $n \sim \text{Poisson}(\mu)$. To find the Jeffreys' prior for μ ,

$$L(n|\mu) = \frac{\mu^n}{n!} e^{-\mu} \qquad \frac{\partial^2 \ln L}{\partial \mu^2} = -\frac{n}{\mu^2}$$

$$I = -E \left[\frac{\partial^2 \ln L}{\partial \mu^2} \right] = \frac{E[n]}{\mu^2} = \frac{1}{\mu}$$

$$\pi(\mu) \propto \sqrt{I(\mu)} = \frac{1}{\sqrt{\mu}}$$

So e.g. for $\mu = s + b$, this means the prior $\pi(s) \sim 1/\sqrt{s + b}$, which depends on b . But this is not designed as a degree of belief about s .

The method of least squares

Suppose we measure N values, y_1, \dots, y_N , assumed to be independent Gaussian r.v.s with

$$E[y_i] = \lambda(x_i; \theta) .$$

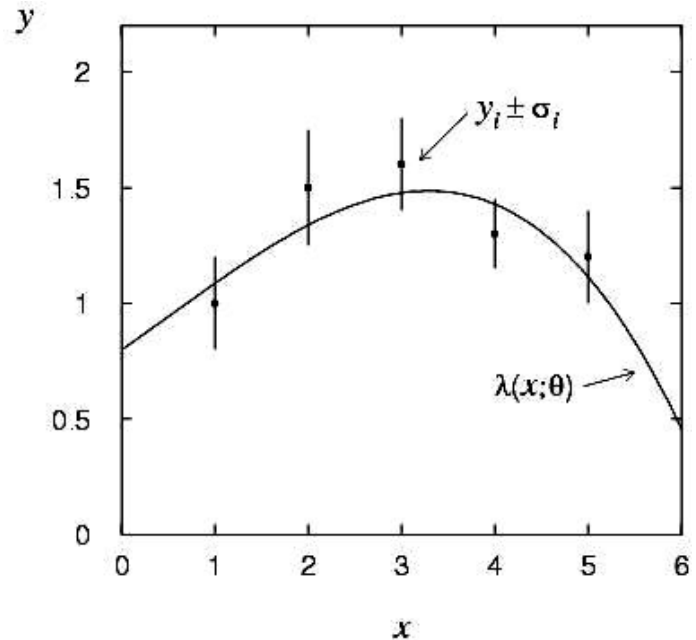
Assume known values of the control variable x_1, \dots, x_N and known variances

$$V[y_i] = \sigma_i^2 .$$

We want to estimate θ , i.e., fit the curve to the data points.

The likelihood function is

$$L(\theta) = \prod_{i=1}^N f(y_i; \theta) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi}\sigma_i} \exp \left[-\frac{(y_i - \lambda(x_i; \theta))^2}{2\sigma_i^2} \right]$$



The method of least squares (2)

The log-likelihood function is therefore

$$\ln L(\theta) = -\frac{1}{2} \sum_{i=1}^N \frac{(y_i - \lambda(x_i; \theta))^2}{\sigma_i^2} + \text{terms not depending on } \theta$$

So maximizing the likelihood is equivalent to minimizing

$$\chi^2(\theta) = \sum_{i=1}^N \frac{(y_i - \lambda(x_i; \theta))^2}{\sigma_i^2}$$

Minimum defines the least squares (LS) estimator $\hat{\theta}$.

Very often measurement errors are \sim Gaussian and so ML and LS are essentially the same.

Often minimize χ^2 numerically (e.g. program **MINUIT**).

LS with correlated measurements

If the y_i follow a multivariate Gaussian, covariance matrix V ,

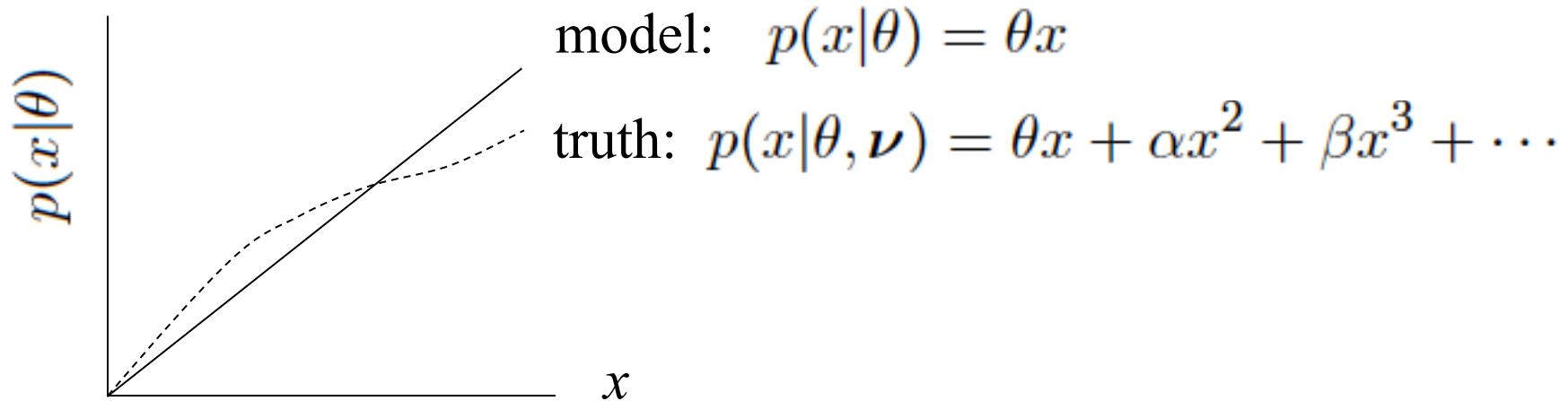
$$g(\vec{y}, \vec{\lambda}, V) = \frac{1}{(2\pi)^{N/2} |V|^{1/2}} \exp \left[-\frac{1}{2} (\vec{y} - \vec{\lambda})^T V^{-1} (\vec{y} - \vec{\lambda}) \right]$$

Then maximizing the likelihood is equivalent to minimizing

$$\chi^2(\vec{\theta}) = \sum_{i,j=1}^N (y_i - \lambda(x_i; \vec{\theta})) (V^{-1})_{ij} (y_j - \lambda(x_j; \vec{\theta}))$$

Systematic uncertainties and nuisance parameters

In general our model of the data is not perfect:



Can improve model by including additional adjustable parameters.

$$p(x|\theta) \rightarrow p(x|\theta, \nu)$$

Nuisance parameter \leftrightarrow 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).

Example: fitting a straight line

Data: (x_i, y_i, σ_i) , $i = 1, \dots, n$.

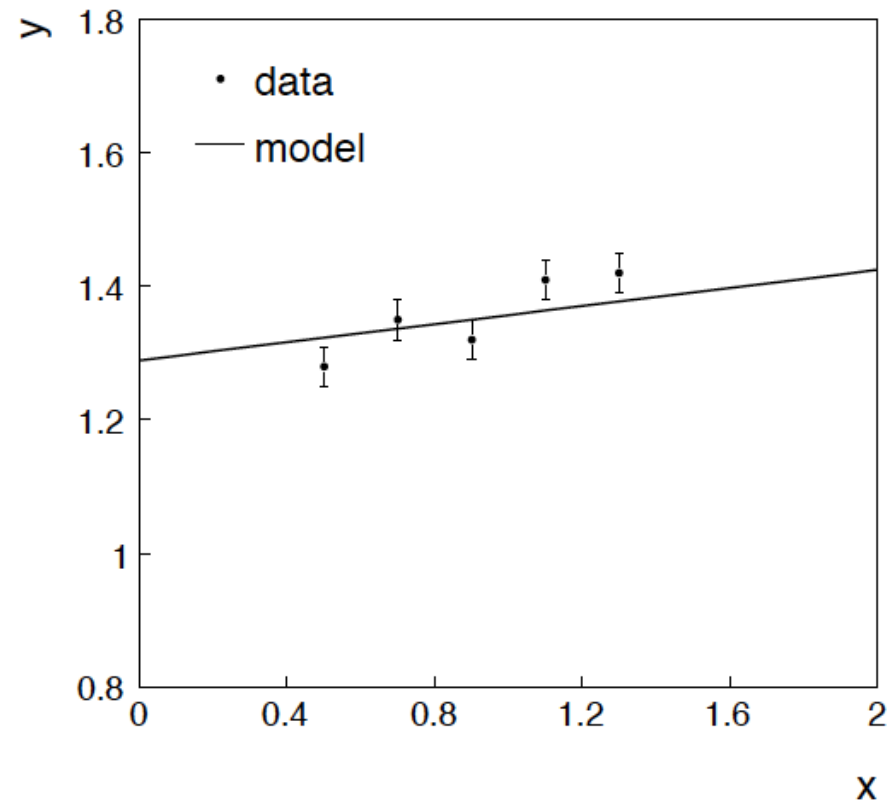
Model: y_i independent and all follow $y_i \sim \text{Gauss}(\mu(x_i), \sigma_i)$

$$\mu(x; \theta_0, \theta_1) = \theta_0 + \theta_1 x,$$

assume x_i and σ_i known.

Goal: estimate θ_0

Here suppose we don't care about θ_1 (example of a “nuisance parameter”)



Maximum likelihood fit with Gaussian data

In this example, the y_i are assumed independent, so the likelihood function is a product of Gaussians:

$$L(\theta_0, \theta_1) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_i} \exp \left[-\frac{1}{2} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} \right],$$

Maximizing the likelihood is here equivalent to minimizing

$$\chi^2(\theta_0, \theta_1) = -2 \ln L(\theta_0, \theta_1) + \text{const} = \sum_{i=1}^n \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2}.$$

i.e., for Gaussian data, ML same as Method of Least Squares (LS)

θ_1 known a priori

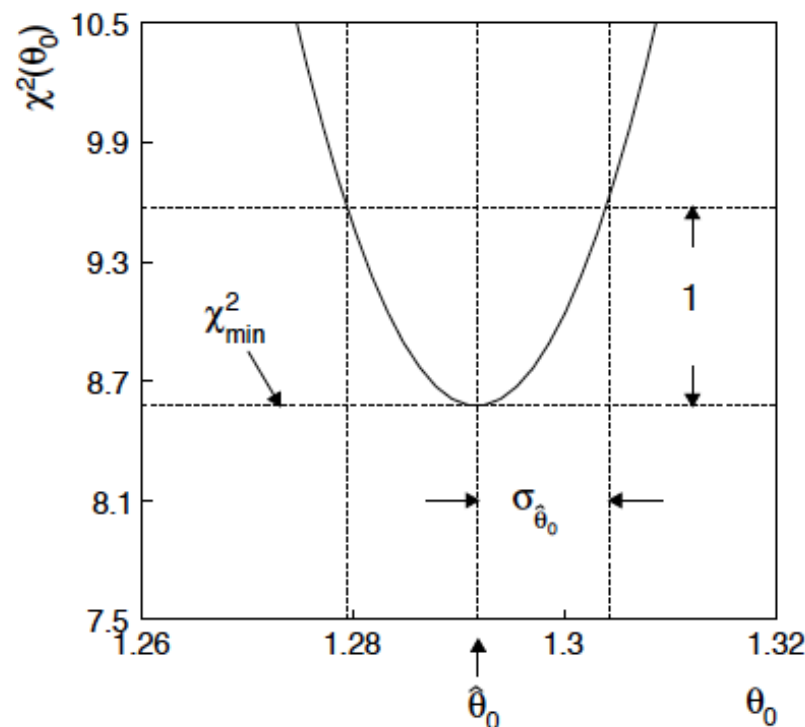
$$L(\theta_0) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_i} \exp \left[-\frac{1}{2} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} \right] .$$

$$\chi^2(\theta_0) = -2 \ln L(\theta_0) + \text{const} = \sum_{i=1}^n \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} .$$

For Gaussian y_i , ML same as LS

Minimize $\chi^2 \rightarrow$ estimator $\hat{\theta}_0$.

Come up one unit from χ^2_{\min}
to find $\sigma_{\hat{\theta}_0}$.



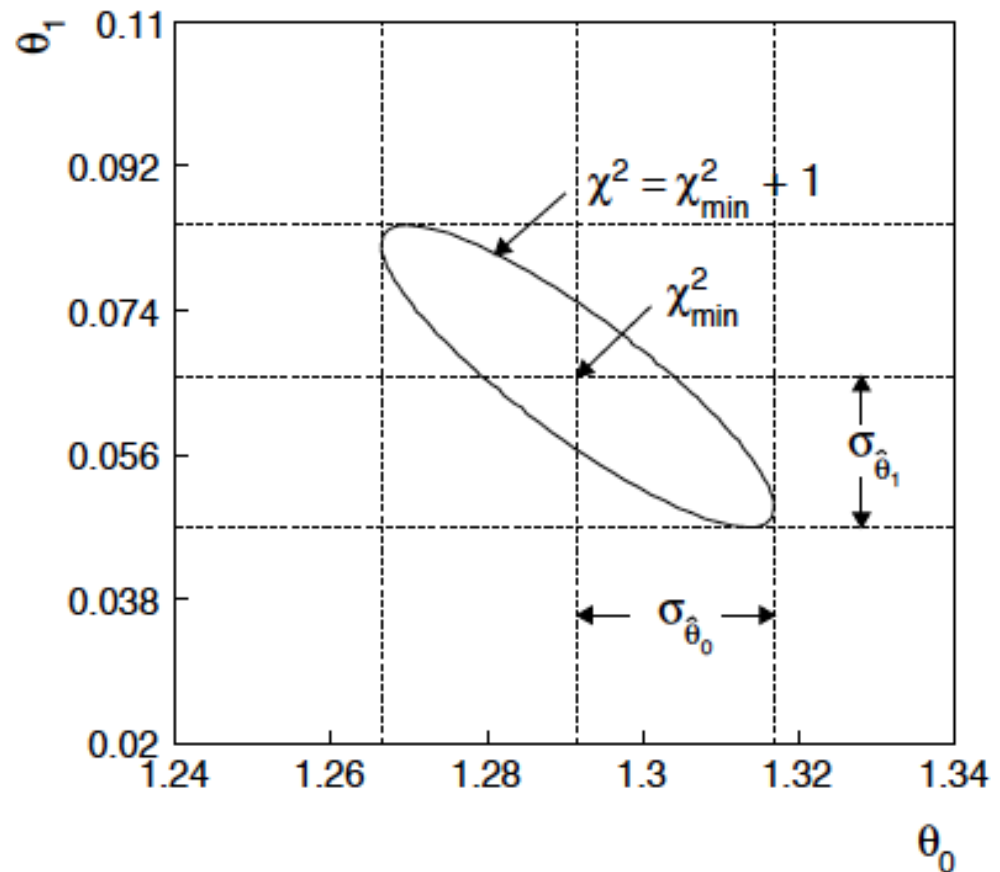
ML (or LS) fit of θ_0 and θ_1

$$\chi^2(\theta_0, \theta_1) = -2 \ln L(\theta_0, \theta_1) + \text{const} = \sum_{i=1}^n \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} .$$

Standard deviations from
tangent lines to contour

$$\chi^2 = \chi_{\min}^2 + 1 .$$

Correlation between
 $\hat{\theta}_0$, $\hat{\theta}_1$ causes errors
to increase.

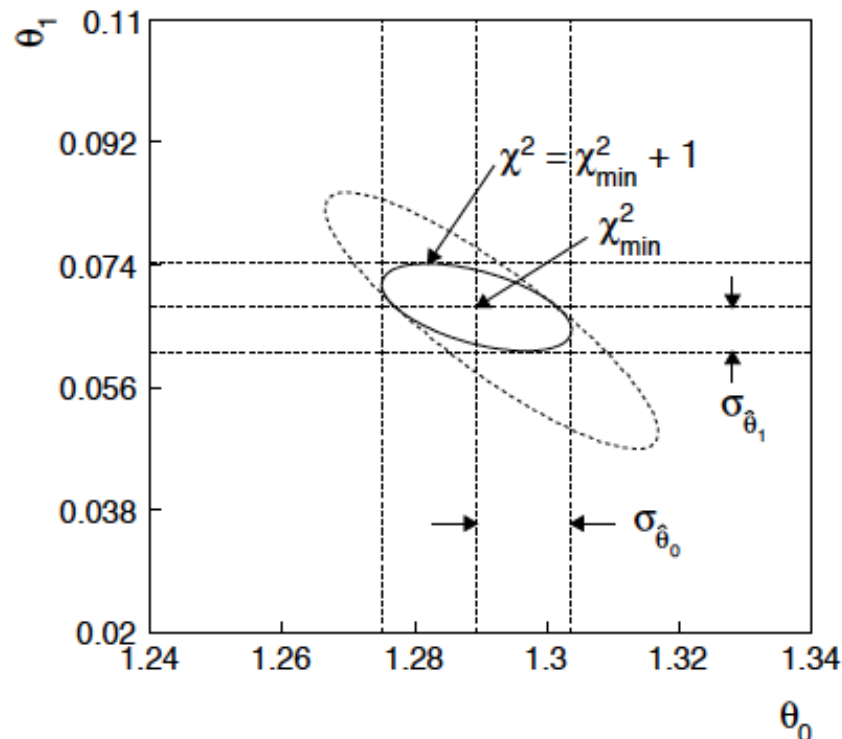


If we have a measurement $t_1 \sim \text{Gauss}(\theta_1, \sigma_{t_1})$

$$L(\theta_0, \theta_1) = \frac{1}{\sqrt{2\pi}\sigma_t} e^{-(t_1 - \theta_1)^2 / 2\sigma_{t_1}^2} \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_i} \exp \left[-\frac{1}{2} \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} \right]$$

$$\chi^2(\theta_0, \theta_1) = \sum_{i=1}^n \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} + \frac{(t_1 - \theta_1)^2}{\sigma_{t_1}^2}$$

The information on θ_1
improves accuracy of $\hat{\theta}_0$.



The Bayesian approach

In Bayesian statistics we can associate a probability with a hypothesis, e.g., a parameter value θ .

Interpret probability of θ as ‘degree of belief’ (subjective).

Need to start with ‘prior pdf’ $\pi(\theta)$, this reflects degree of belief about θ before doing the experiment.

Our experiment has data x , \rightarrow likelihood function $L(x|\theta)$.

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

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

Posterior pdf $p(\theta|x)$ contains all our knowledge about θ .

Bayesian method

We need to associate prior probabilities with θ_0 and θ_1 , e.g.,

$$\begin{aligned}\pi(\theta_0, \theta_1) &= \pi_0(\theta_0) \pi_1(\theta_1) \\ \pi_0(\theta_0) &= \text{const.} \\ \pi_1(\theta_1) &= \frac{1}{\sqrt{2\pi}\sigma_{t_1}} e^{-(\theta_1 - t_1)^2 / 2\sigma_{t_1}^2}\end{aligned}$$

‘non-informative’, in any case much broader than $L(\theta_0)$

← based on previous measurement

Putting this into Bayes’ theorem gives:

$$p(\theta_0, \theta_1 | \vec{y}) \propto \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(y_i - \mu(x_i; \theta_0, \theta_1))^2 / 2\sigma_i^2} \pi_0 \frac{1}{\sqrt{2\pi}\sigma_{t_1}} e^{-(\theta_1 - t_1)^2 / 2\sigma_{t_1}^2}$$

posterior \propto likelihood \times prior

Bayesian method (continued)

We then integrate (marginalize) $p(\theta_0, \theta_1 | x)$ to find $p(\theta_0 | x)$:

$$p(\theta_0 | x) = \int p(\theta_0, \theta_1 | x) d\theta_1 .$$

In this example we can do the integral (rare). We find

$$p(\theta_0 | x) = \frac{1}{\sqrt{2\pi}\sigma_{\theta_0}} e^{-(\theta_0 - \hat{\theta}_0)^2 / 2\sigma_{\theta_0}^2} \quad \text{with}$$

$$\hat{\theta}_0 = \text{same as ML estimator}$$

$$\sigma_{\theta_0} = \sigma_{\hat{\theta}_0} \text{ (same as before)}$$

Usually need numerical methods (e.g. Markov Chain Monte Carlo) to do integral.

Digression: marginalization with MCMC

Bayesian computations involve integrals like

$$p(\theta_0|x) = \int p(\theta_0, \theta_1|x) d\theta_1 .$$

often high dimensionality and impossible in closed form,
also impossible with ‘normal’ acceptance-rejection Monte Carlo.

Markov Chain Monte Carlo (MCMC) has revolutionized
Bayesian computation.




MCMC (e.g., Metropolis-Hastings algorithm) generates
correlated sequence of random numbers:

cannot use for many applications, e.g., detector MC;
effective stat. error greater than if all values independent .

Basic idea: sample multidimensional $\vec{\theta}$,
look, e.g., only at distribution of parameters of interest.

MCMC basics: Metropolis-Hastings algorithm

Goal: given an n -dimensional pdf $p(\vec{\theta})$,
generate a sequence of points $\vec{\theta}_1, \vec{\theta}_2, \vec{\theta}_3, \dots$

- 1) Start at some point $\vec{\theta}_0$
- 2) Generate $\vec{\theta} \sim q(\vec{\theta}; \vec{\theta}_0)$  Proposal density $q(\vec{\theta}; \vec{\theta}_0)$
e.g. Gaussian centred
about $\vec{\theta}_0$
- 3) Form Hastings test ratio $\alpha = \min \left[1, \frac{p(\vec{\theta})q(\vec{\theta}_0; \vec{\theta})}{p(\vec{\theta}_0)q(\vec{\theta}; \vec{\theta}_0)} \right]$
- 4) Generate $u \sim \text{Uniform}[0, 1]$
- 5) If $u \leq \alpha$, $\vec{\theta}_1 = \vec{\theta}$,  move to proposed point
else $\vec{\theta}_1 = \vec{\theta}_0$  old point repeated
- 6) Iterate

Metropolis-Hastings (continued)

This rule produces a *correlated* sequence of points (note how each new point depends on the previous one).

For our purposes this correlation is not fatal, but statistical errors larger than if points were independent.

The proposal density can be (almost) anything, but choose so as to minimize autocorrelation. Often take proposal density symmetric: $q(\vec{\theta}; \vec{\theta}_0) = q(\vec{\theta}_0; \vec{\theta})$

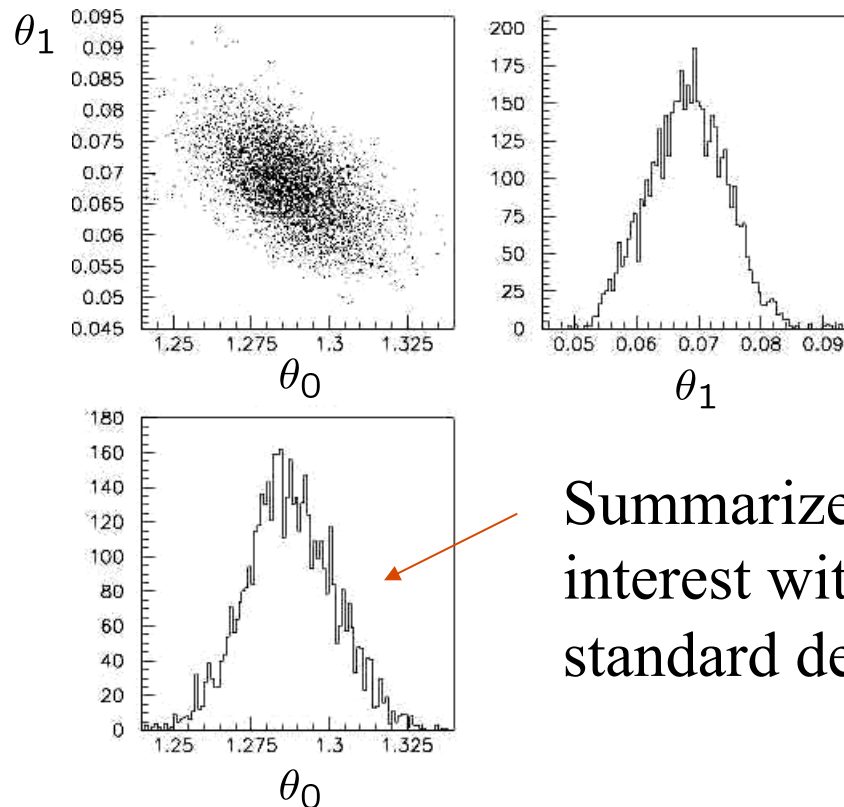
Test ratio is (*Metropolis-Hastings*): $\alpha = \min \left[1, \frac{p(\vec{\theta})}{p(\vec{\theta}_0)} \right]$

I.e. if the proposed step is to a point of higher $p(\vec{\theta})$, take it; if not, only take the step with probability $p(\vec{\theta})/p(\vec{\theta}_0)$.

If proposed step rejected, hop in place.

Example: posterior pdf from MCMC

Sample the posterior pdf from previous example with MCMC:



Summarize pdf of parameter of interest with, e.g., mean, median, standard deviation, etc.

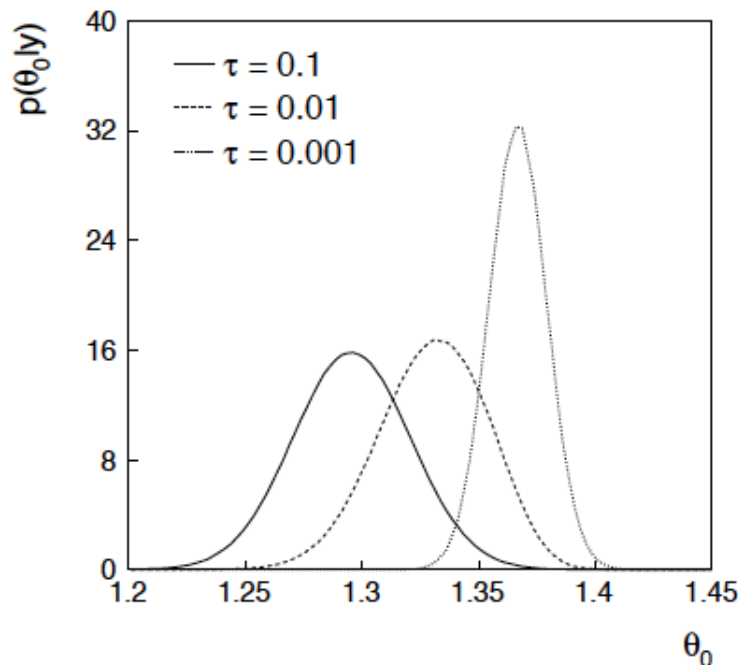
Although numerical values of answer here same as in frequentist case, interpretation is different (sometimes unimportant?)

Bayesian method with alternative priors

Suppose we don't have a previous measurement of θ_1 but rather, e.g., a theorist says it should be positive and not too much greater than 0.1 "or so", i.e., something like

$$\pi_1(\theta_1) = \frac{1}{\tau} e^{-\theta_1/\tau}, \quad \theta_1 \geq 0, \quad \tau = 0.1.$$

From this we obtain (numerically) the posterior pdf for θ_0 :



This summarizes all knowledge about θ_0 .

Look also at result from variety of priors.

Comments on parameter estimation

Estimation of parameters is usually the “easy” part of statistics:

Frequentist: maximize the likelihood.

Bayesian: find posterior pdf and summarize (e.g. mode).

Standard tools for quantifying precision of estimates:

Variance of estimators, confidence intervals,...

But there are many potential stumbling blocks:

bias versus variance trade-off (how many parameters to fit?);

goodness of fit (usually only for LS or binned data);

choice of prior for Bayesian approach;

unexpected behaviour in LS averages with correlations,...

Interval estimation — introduction

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

Desirable properties of such an interval may include:

- communicate objectively the result of the experiment;
- have a given probability of containing the true parameter;
- provide information needed to draw conclusions about the parameter possibly incorporating stated prior beliefs.

Often use \pm the estimated standard deviation of the estimator.

In some cases, however, this is not adequate:

- estimate near a physical boundary,
e.g., an observed event rate consistent with zero.

We will look briefly at Frequentist and Bayesian intervals.

Confidence intervals by inverting a test

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 $P(\text{data in critical region}) \leq \alpha$ 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 would **not** be rejected in a test of size α (confidence level is $1 - \alpha$).

The interval will cover the true value of θ 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 $p_\theta = \alpha$ and solve for θ .

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.

When testing s values to find upper limit, relevant alternative is $s = 0$ (or lower s), so critical region at low n and p -value of hypothesized s is $P(n \leq n_{\text{obs}}; s, b)$.

Upper limit s_{up} at $\text{CL} = 1 - \alpha$ from setting $\alpha = p_s$ and solving for s :

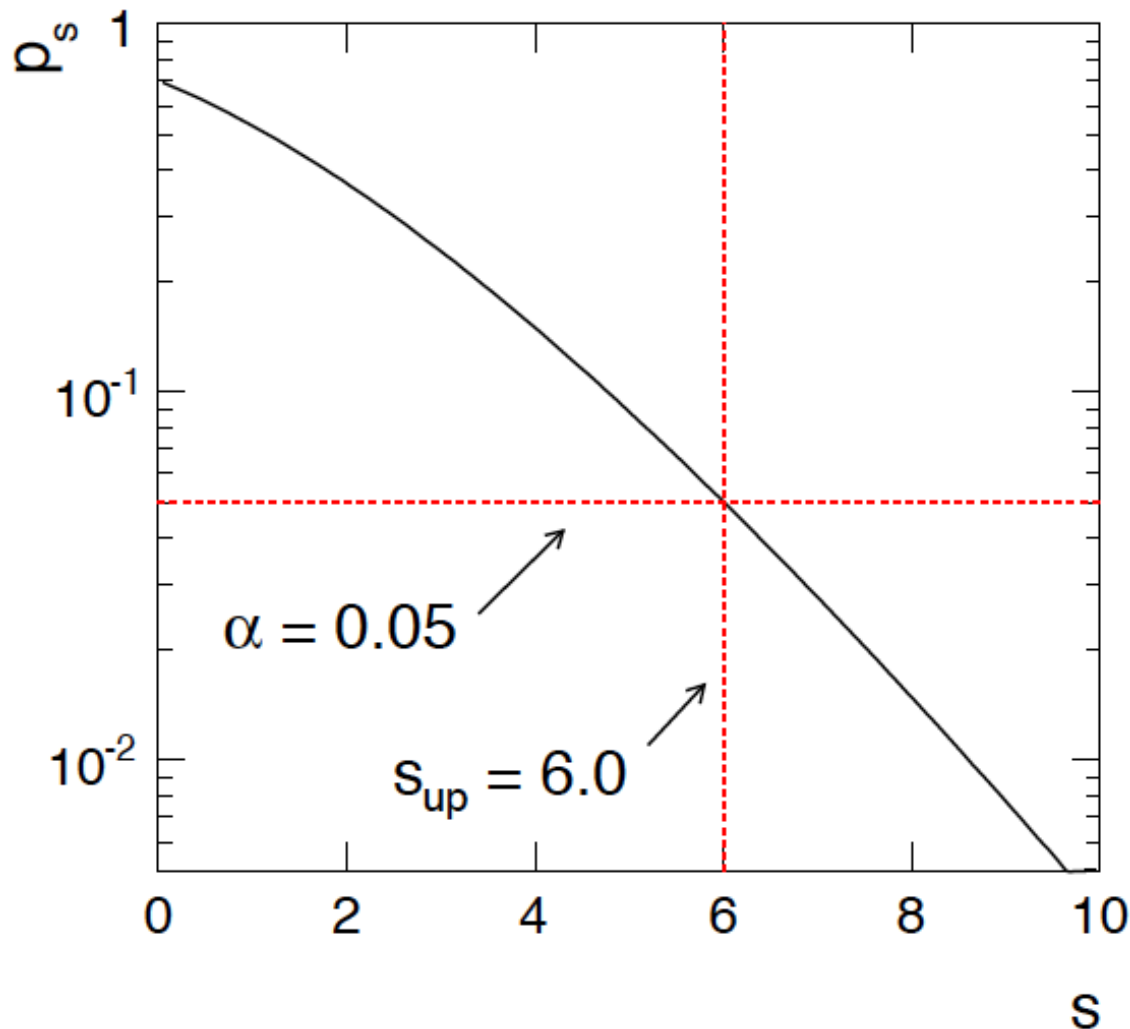
$$\alpha = P(n \leq n_{\text{obs}}; s_{\text{up}}, b) = \sum_{n=0}^{n_{\text{obs}}} \frac{(s_{\text{up}} + b)^n}{n!} e^{-(s_{\text{up}} + b)}$$

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

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

Frequentist upper limit on Poisson parameter

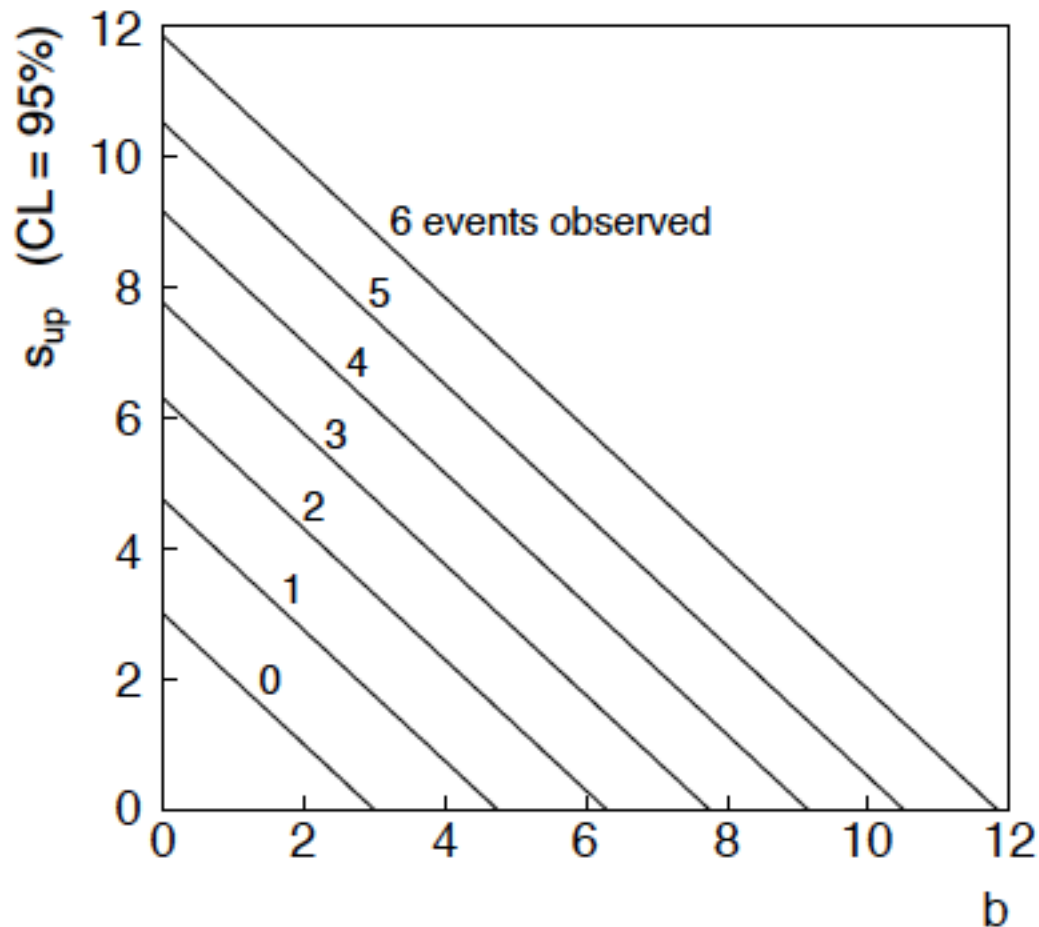
Upper limit s_{up} at $\text{CL} = 1 - \alpha$ found from $p_s = \alpha$.



$$n_{\text{obs}} = 5,$$
$$b = 4.5$$

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

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

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 $s = 0$

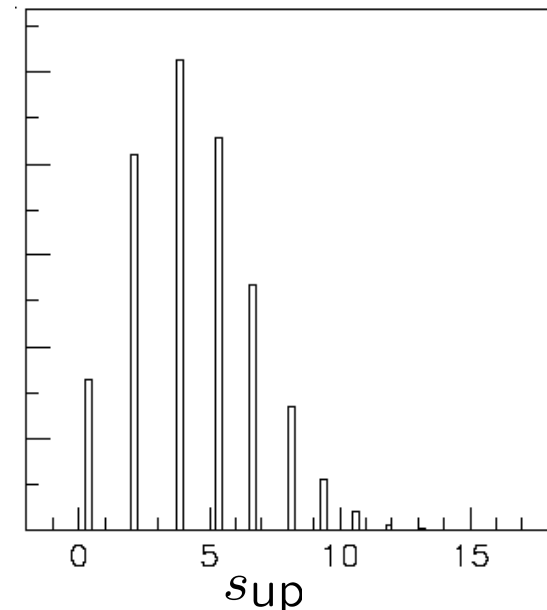
Physicist: I should have used $CL = 0.95$ — then $s_{\text{up}} = 0.496$

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

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

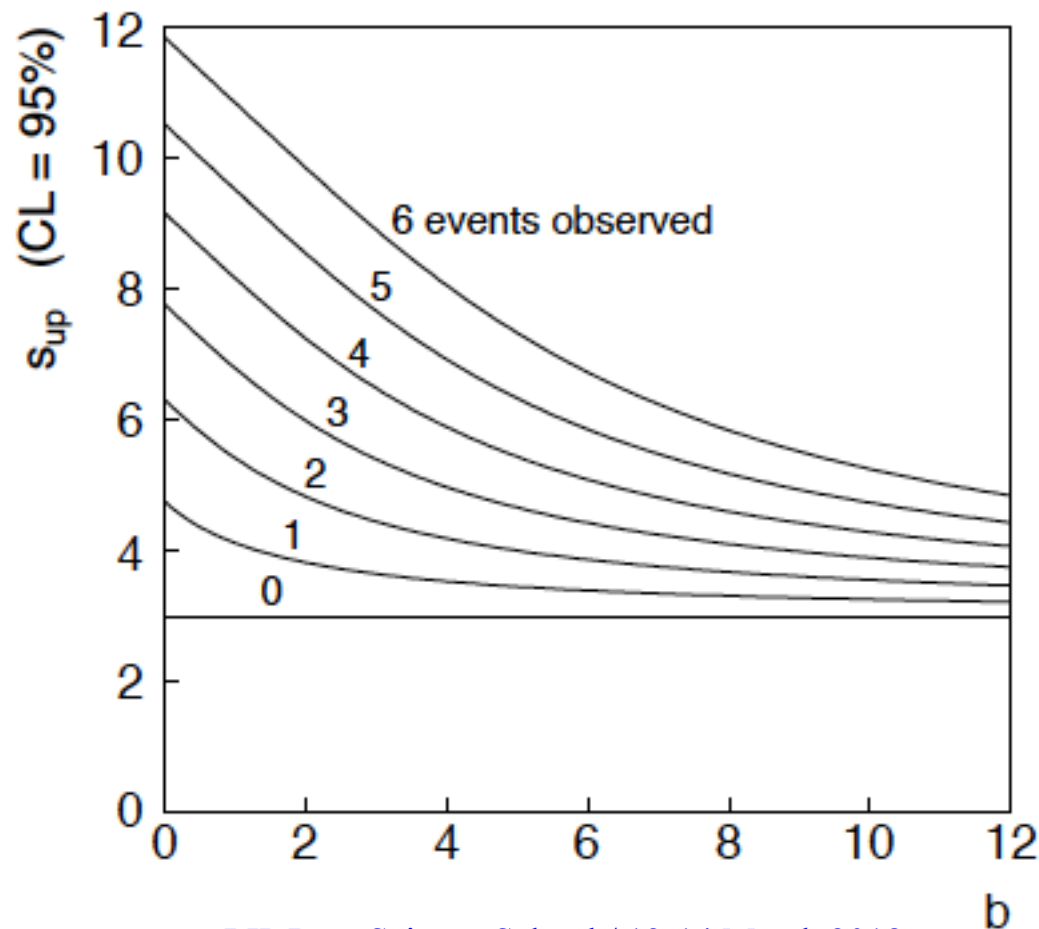
$$p = 1 - \alpha \left(1 - F_{\chi^2} [2b, 2(n+1)] \right)$$

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 $n = 0$.



Priors from formal rules

Because of difficulties in encoding a vague degree of belief in a prior, one often attempts to derive the prior from formal rules, e.g., to satisfy certain invariance principles or to provide maximum information gain for a certain set of measurements.

Often called “objective priors”

Form basis of Objective Bayesian Statistics

The priors do not reflect a degree of belief (but might represent possible extreme cases).

In Objective Bayesian analysis, can use the intervals in a frequentist way, i.e., regard Bayes’ theorem as a recipe to produce an interval with certain coverage properties.

Priors from formal rules (cont.)

For a review of priors obtained by formal rules see, e.g.,

Robert E. Kass and Larry Wasserman, *The Selection of Prior Distributions by Formal Rules*, J. Am. Stat. Assoc., Vol. 91, No. 435, pp. 1343-1370 (1996).

Formal priors have not been widely used in HEP, but there is recent interest in this direction, especially the reference priors of Bernardo and Berger; see e.g.

L. Demortier, S. Jain and H. Prosper, *Reference priors for high energy physics*, Phys. Rev. D 82 (2010) 034002, arXiv:1002.1111.

D. Casadei, *Reference analysis of the signal + background model in counting experiments*, JINST 7 (2012) 01012; arXiv:1108.4270.

Approximate confidence intervals/regions from the likelihood function

Suppose we test parameter value(s) $\theta = (\theta_1, \dots, \theta_n)$ using the ratio

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

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

$$t_\theta = -2 \ln \lambda(\theta)$$

so higher t_θ means worse agreement between θ and the data.

p -value of θ therefore

$$p_\theta = \int_{t_{\theta, \text{obs}}}^{\infty} f(t_\theta | \theta) dt_\theta$$

 need pdf

Confidence region from Wilks' theorem

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

$$f(t_{\theta}|\theta) \sim \chi_n^2$$

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

Assuming this holds, the p -value is

$$p_{\theta} = 1 - F_{\chi_n^2}(t_{\theta}) \quad \text{where} \quad F_{\chi_n^2}(t_{\theta}) \equiv \int_0^{t_{\theta}} f_{\chi_n^2}(t'_{\theta}) t'_{\theta}$$

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

$$t_{\theta} = -2 \ln \frac{L(\theta)}{L(\hat{\theta})} = F_{\chi_n^2}^{-1}(1 - \alpha)$$

Confidence region from Wilks' theorem (cont.)

i.e., boundary of confidence region in θ space is where

$$\ln L(\theta) = \ln L(\hat{\theta}) - \frac{1}{2} F_{\chi_n^2}^{-1}(1 - \alpha)$$

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

$$F_{\chi_1^2}^{-1}(0.683) = 1$$

and so the 68.3% confidence level interval is determined by

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

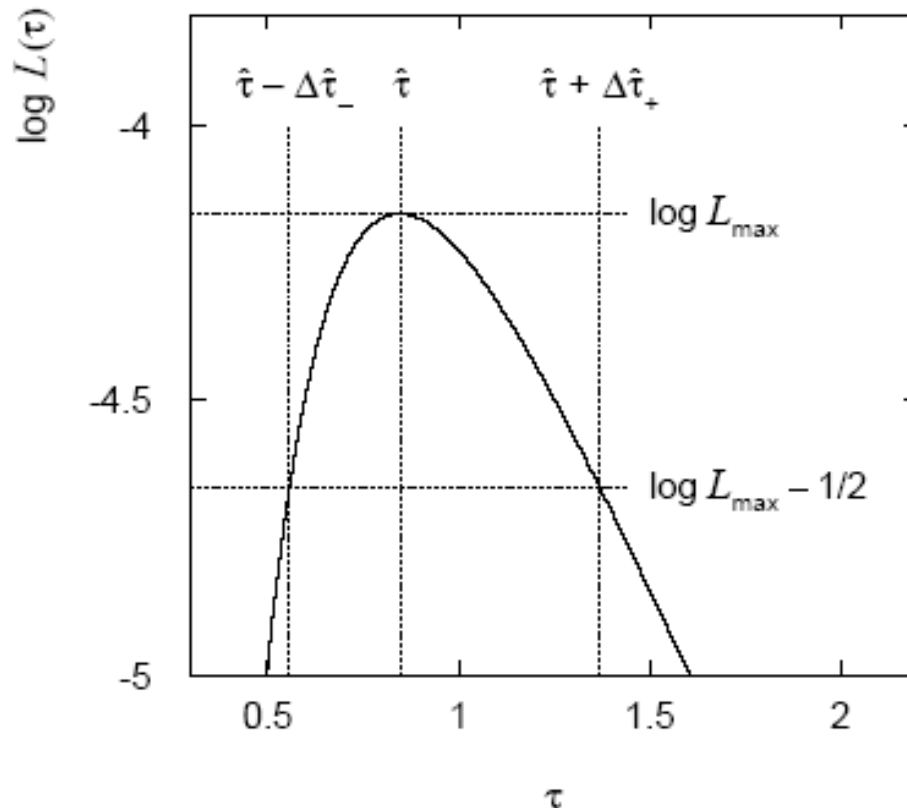
Same as recipe for finding the estimator's standard deviation, i.e.,

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

Example of interval from $\ln L$

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

Exponential example, now with only 5 events:



Parameter estimate and
approximate 68.3% CL
confidence interval:

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

Multiparameter case

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

$$Q_\alpha = F_{\chi_n^2}^{-1}(1 - \alpha)$$

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

Multiparameter case (cont.)

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

$1 - \alpha$	\bar{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

Prototype search analysis

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

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

Assume the n_i are Poisson distributed with expectation values

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

strength parameter

where

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

signal

background

Prototype analysis (II)

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

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

Assume the m_i are Poisson distributed with expectation values

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

 nuisance parameters ($\boldsymbol{\theta}_s, \boldsymbol{\theta}_b, b_{\text{tot}}$)

Likelihood function is

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

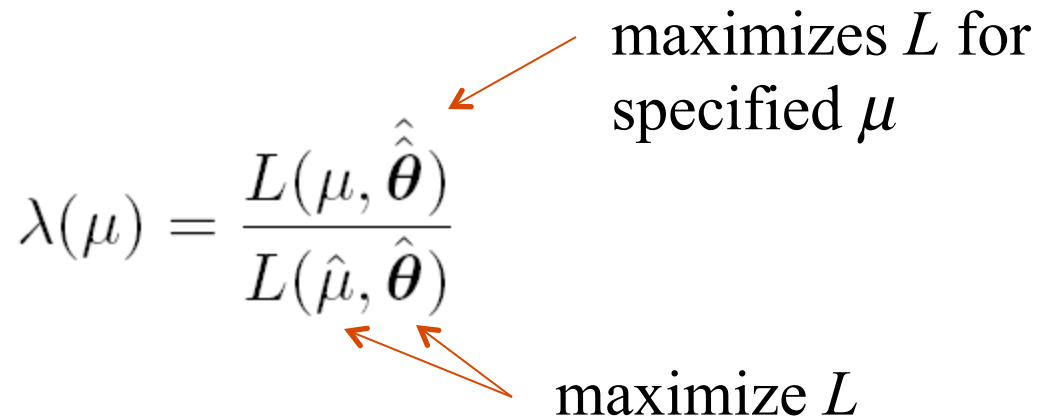
The profile likelihood ratio

Base significance test on the profile likelihood ratio:

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

maximizes L for specified μ

maximize L



The likelihood ratio of point hypotheses gives optimum test (Neyman-Pearson lemma). In practice the profile LR is near-optimal.

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

Test statistic for discovery

Try to reject background-only ($\mu=0$) hypothesis using

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

i.e. here only regard upward fluctuation of data as evidence against the background-only hypothesis.

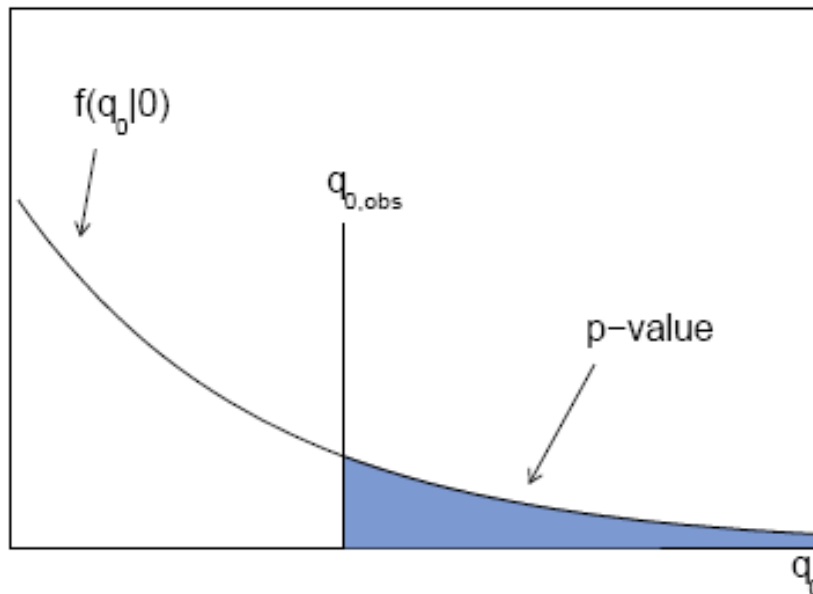
Note that even though here physically $\mu \geq 0$, we allow $\hat{\mu}$ to be negative. In large sample limit its distribution becomes Gaussian, and this will allow us to write down simple expressions for distributions of our test statistics.

p-value for discovery

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

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

use e.g. asymptotic formula



From *p*-value get equivalent significance,

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

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 .

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

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

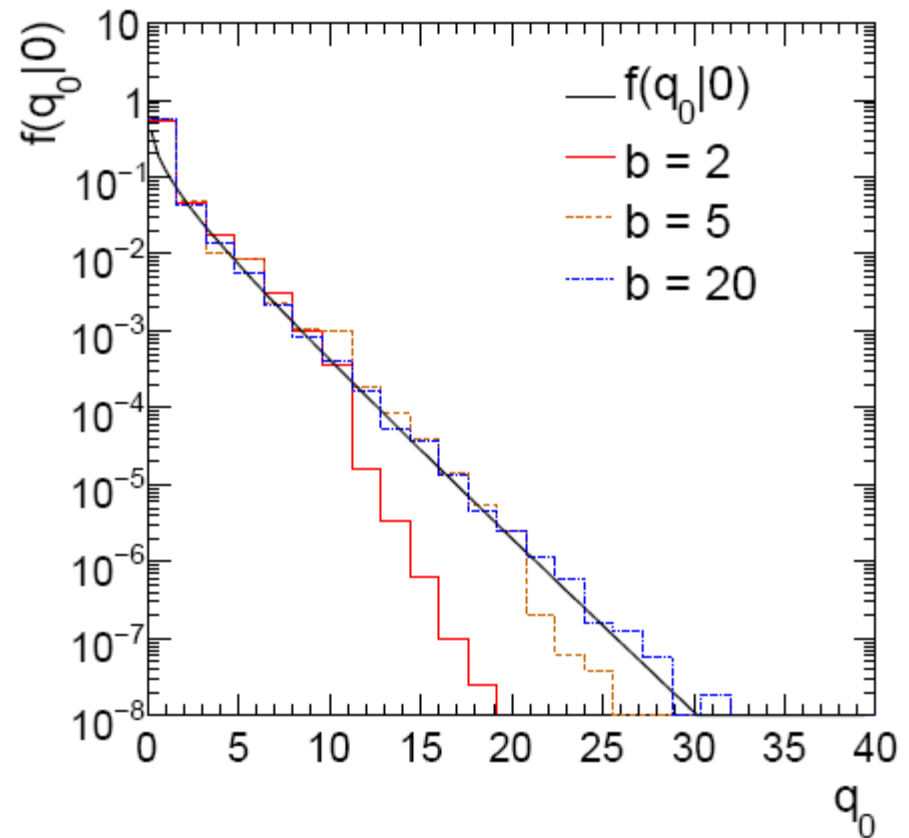
Monte Carlo test of asymptotic formula

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

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

Here take $\tau = 1$.

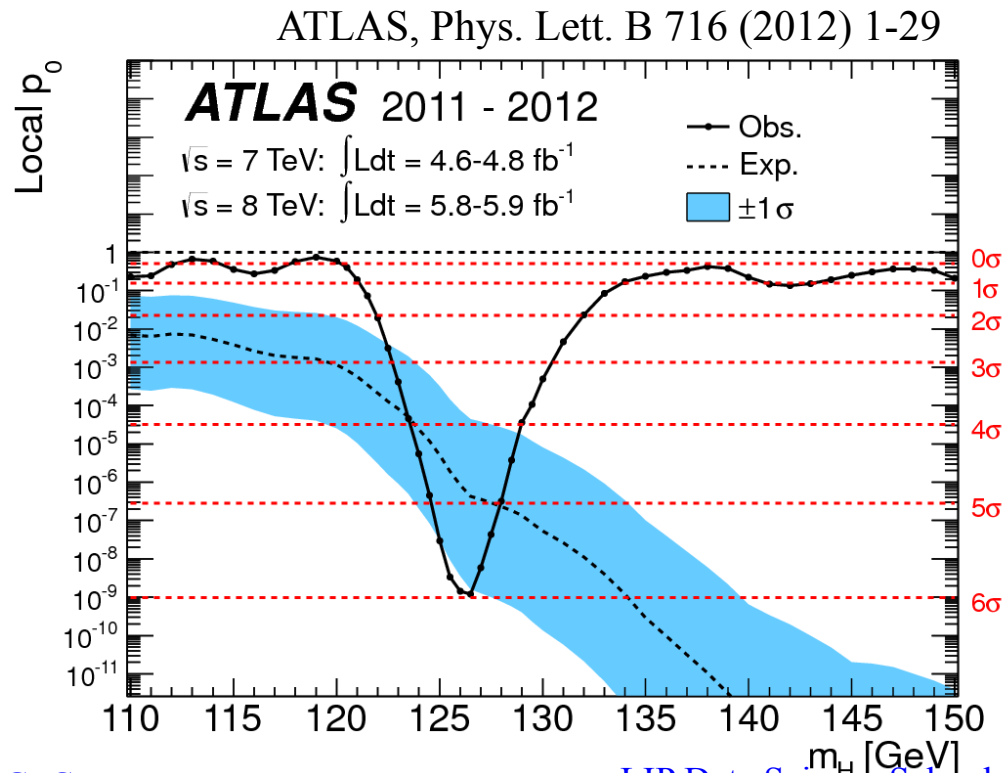
Asymptotic formula is
good approximation to 5σ
level ($q_0 = 25$) already for
 $b \sim 20$.



Example of discovery: 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 correction for the Look-Elsewhere Effect.

The “Expected” (dashed) curve gives the median p_0 under assumption of the SM Higgs ($\mu = 1$) at each m_H .



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

Return to interval estimation

Suppose a model contains a parameter μ ; we want to know which values are consistent with the data and which are disfavoured.

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 probability that the true value of μ will be rejected is not greater than α , so by construction the confidence interval will contain the true value of μ with probability $\geq 1 - \alpha$.

The interval depends on the choice of the test (critical region).

If the test is formulated in terms of a p -value, p_μ , then the confidence interval represents those values of μ for which $p_\mu > \alpha$.

To find the end points of the interval, set $p_\mu = \alpha$ and solve for μ .

Test statistic for upper limits

cf. Cowan, Cranmer, Gross, Vitells, arXiv:1007.1727, EPJC 71 (2011) 1554.

For purposes of setting an upper limit on μ one can use

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

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

Large sample approximation:

$$p_\mu = 1 - \Phi(\sqrt{q_\mu})$$

95% CL upper limit on μ is highest value for which p -value is not less than 0.05.

Monte Carlo test of asymptotic formulae

Consider again $n \sim \text{Poisson}(\mu s + b)$, $m \sim \text{Poisson}(\tau b)$

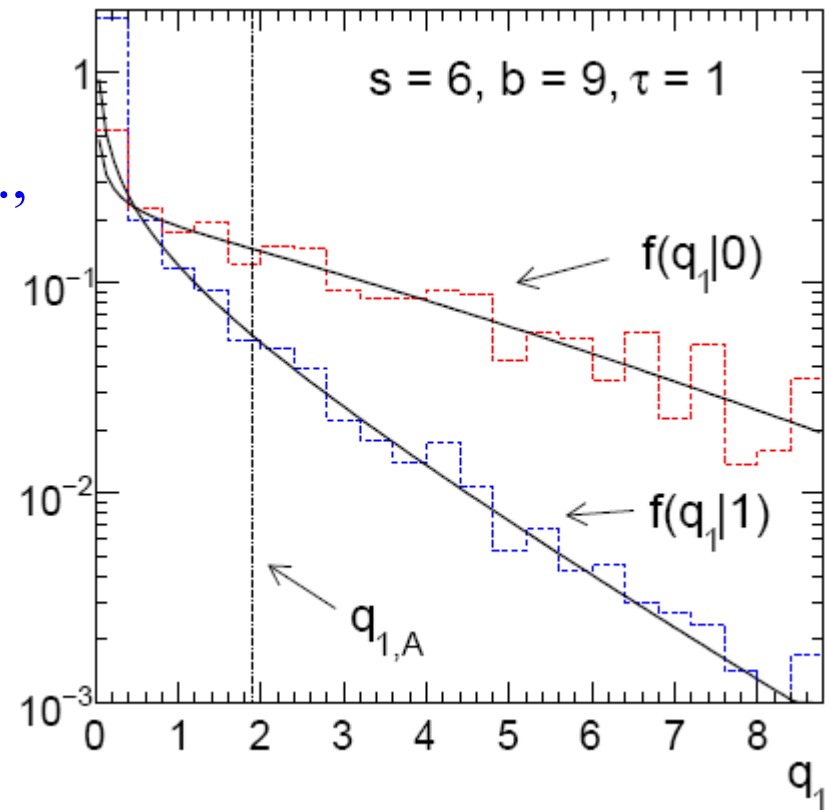
Use q_μ to find p -value of hypothesized μ values.

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”.

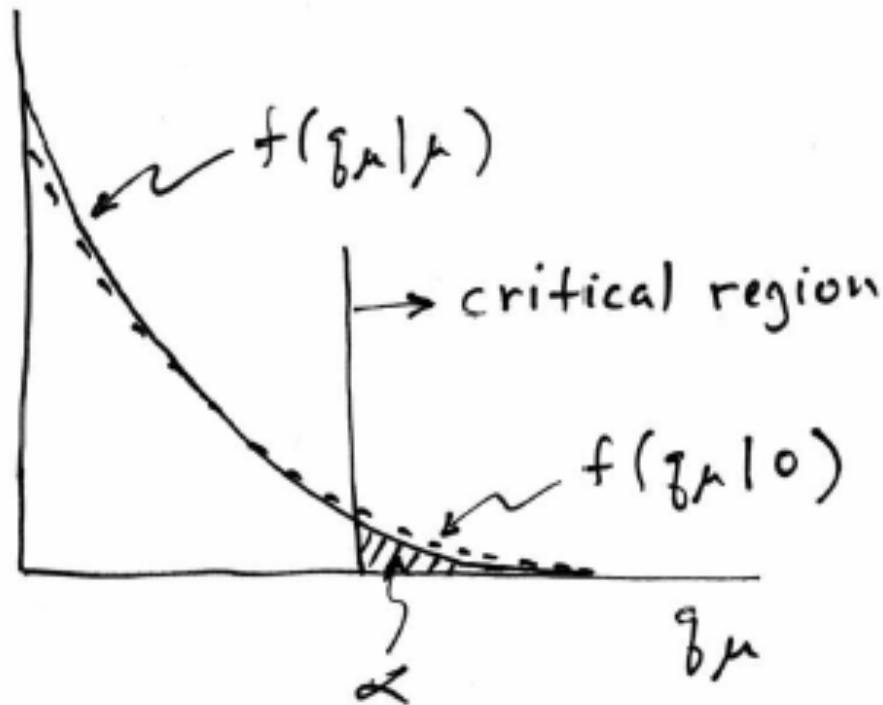
Here asymptotic formulae good for $s = 6$, $b = 9$.



Low sensitivity to μ

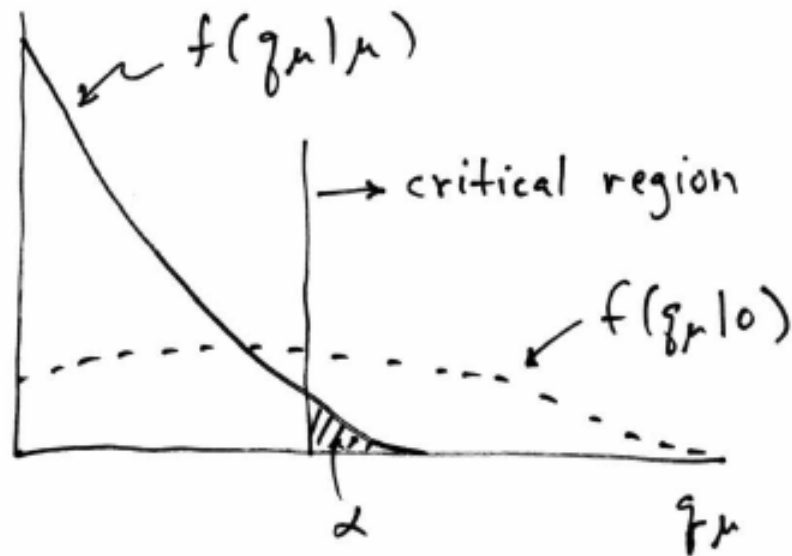
It can be that the effect of a given hypothesized μ is very small relative to the background-only ($\mu = 0$) prediction.

This means that the distributions $f(q_\mu|\mu)$ and $f(q_\mu|0)$ will be almost the same:



Having sufficient sensitivity

In contrast, having sensitivity to μ means that the distributions $f(q_\mu|\mu)$ and $f(q_\mu|0)$ are more separated:

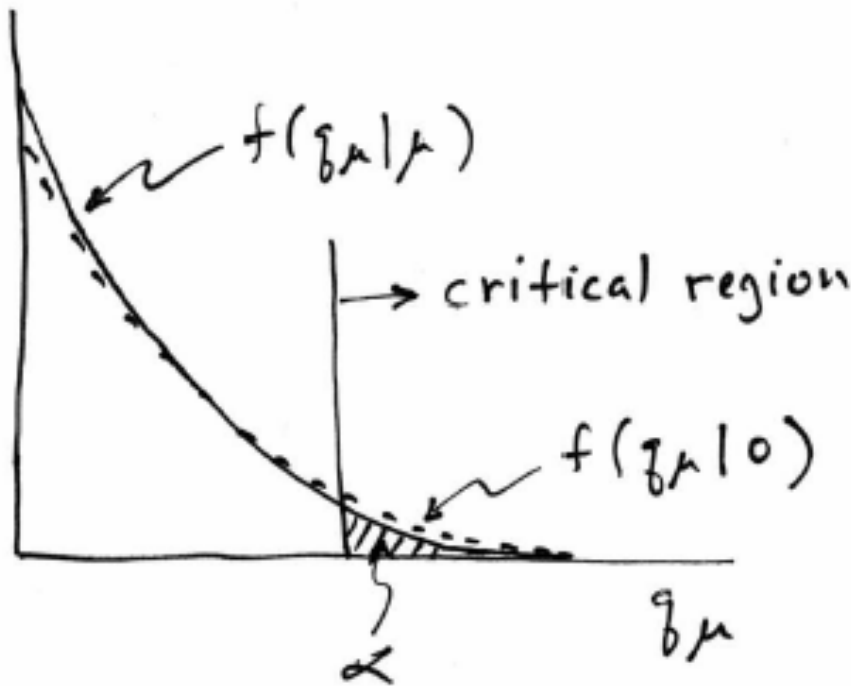


That is, the power (probability to reject μ if $\mu = 0$) is substantially higher than α . Use this power as a measure of the sensitivity.

Spurious exclusion

Consider again the case of low sensitivity. By construction the probability to reject μ if μ is true is α (e.g., 5%).

And the probability to reject μ if $\mu = 0$ (the power) is only slightly greater than α .



This means that with probability of around $\alpha = 5\%$ (slightly higher), one excludes hypotheses to which one has essentially no sensitivity (e.g., $m_H = 1000$ TeV).

“Spurious exclusion”

Ways of addressing spurious exclusion

The problem of excluding parameter values to which one has no sensitivity known for a long time; see e.g.,

Virgil L. Highland, *Estimation of Upper Limits from Experimental Data*, July 1986, Revised February 1987, Temple University Report C00-3539-38.

In the 1990s this was re-examined for the LEP Higgs search by Alex Read and others

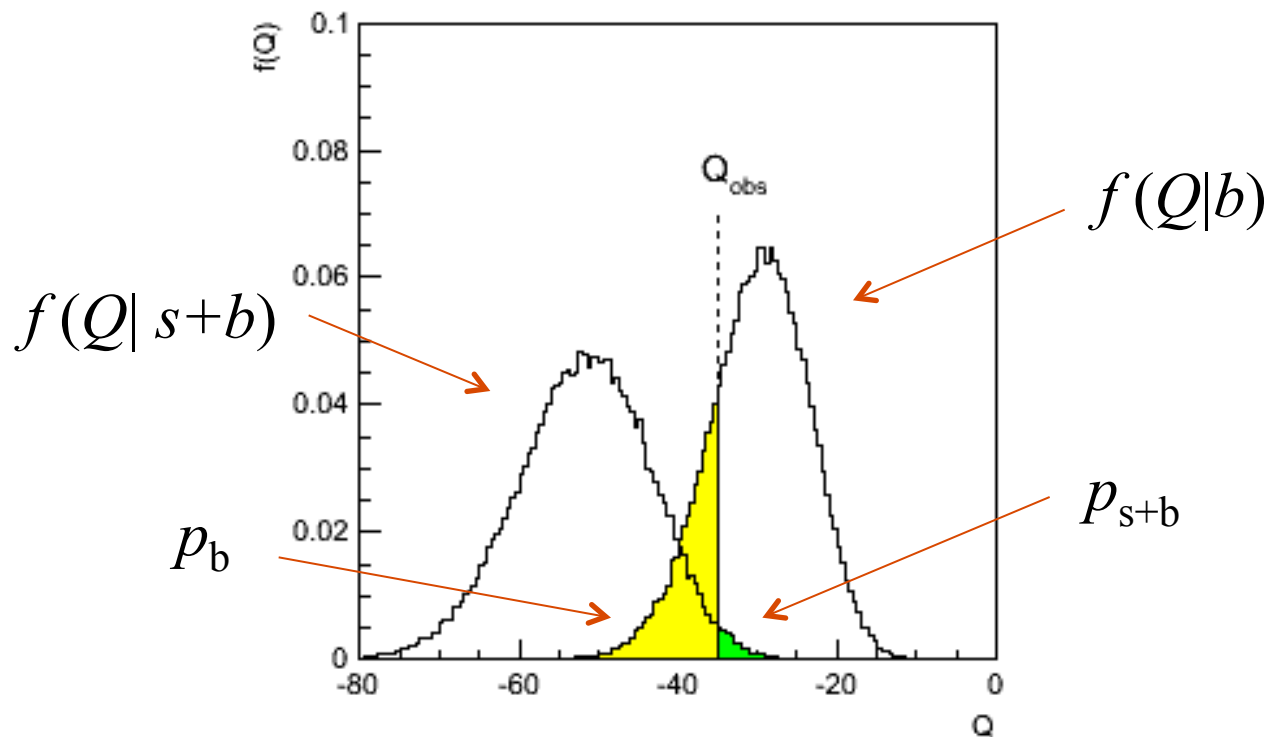
T. Junk, Nucl. Instrum. Methods Phys. Res., Sec. A **434**, 435 (1999); A.L. Read, J. Phys. G **28**, 2693 (2002).

and led to the “ CL_s ” procedure for upper limits.

Unified intervals also effectively reduce spurious exclusion by the particular choice of critical region.

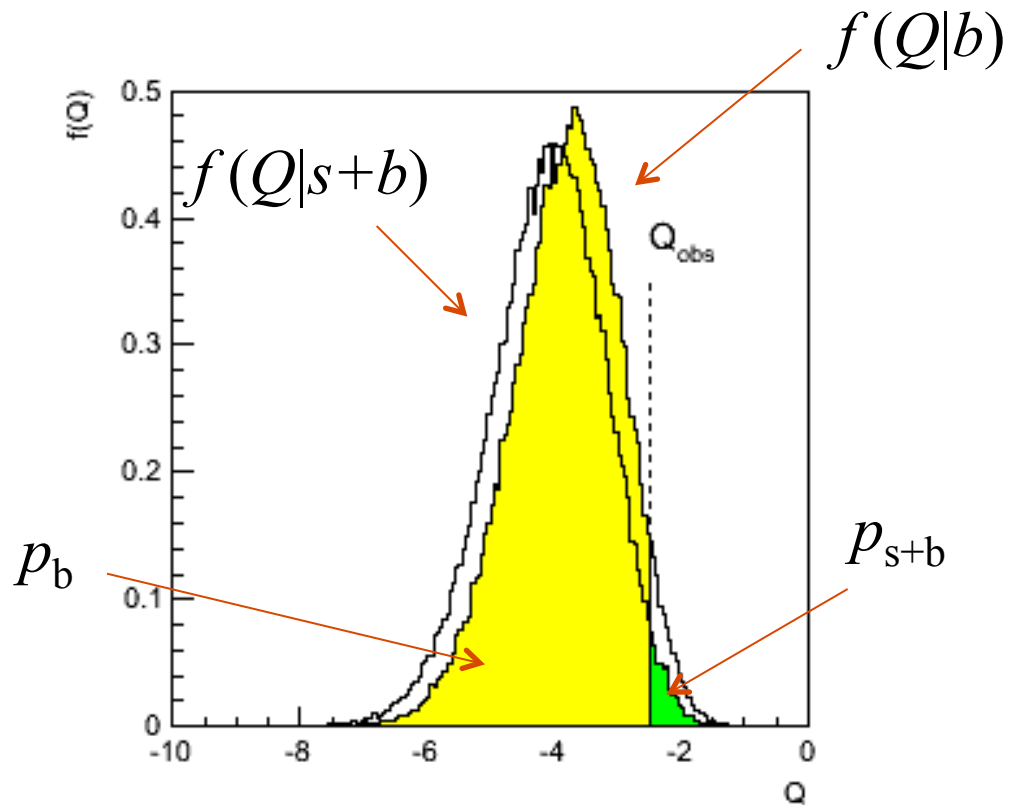
The CL_s procedure

In the usual formulation of CL_s , one tests both the $\mu = 0$ (b) and $\mu > 0$ ($\mu s+b$) hypotheses with the same statistic $Q = -2\ln L_{s+b}/L_b$:



The CL_s procedure (2)

As before, “low sensitivity” means the distributions of Q under b and $s+b$ are very close:



The CL_s procedure (3)

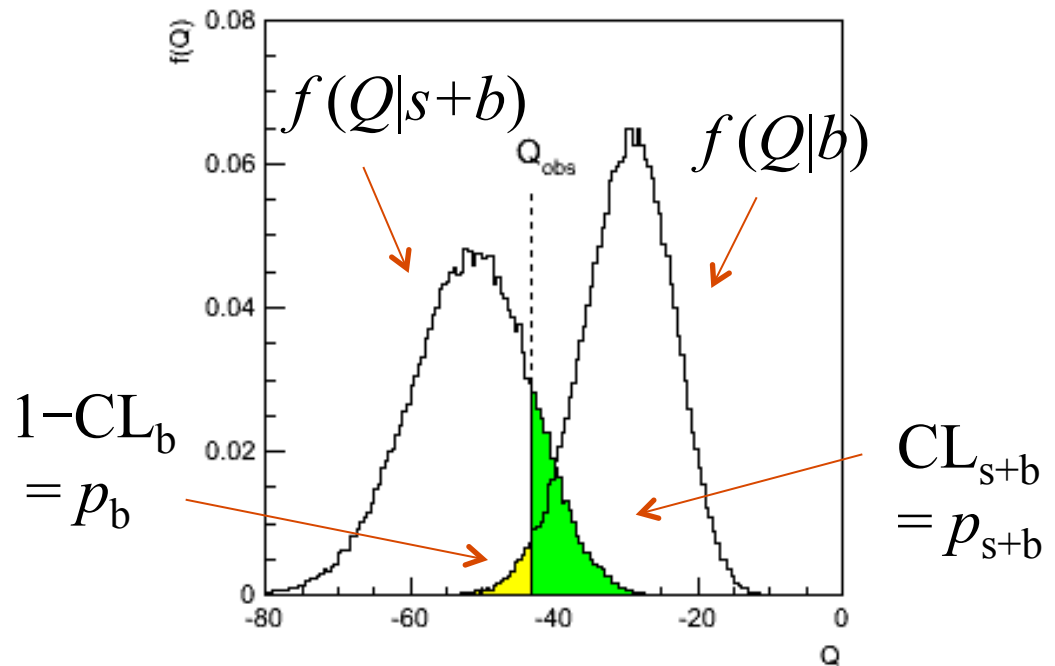
The CL_s solution (A. Read et al.) is to base the test not on the usual p -value (CL_{s+b}), but rather to divide this by CL_b (\sim one minus the p -value of the b -only hypothesis), i.e.,

Define:

$$CL_s = \frac{CL_{s+b}}{CL_b} \\ = \frac{p_{s+b}}{1 - p_b}$$

Reject $s+b$
hypothesis if:

$$CL_s \leq \alpha$$



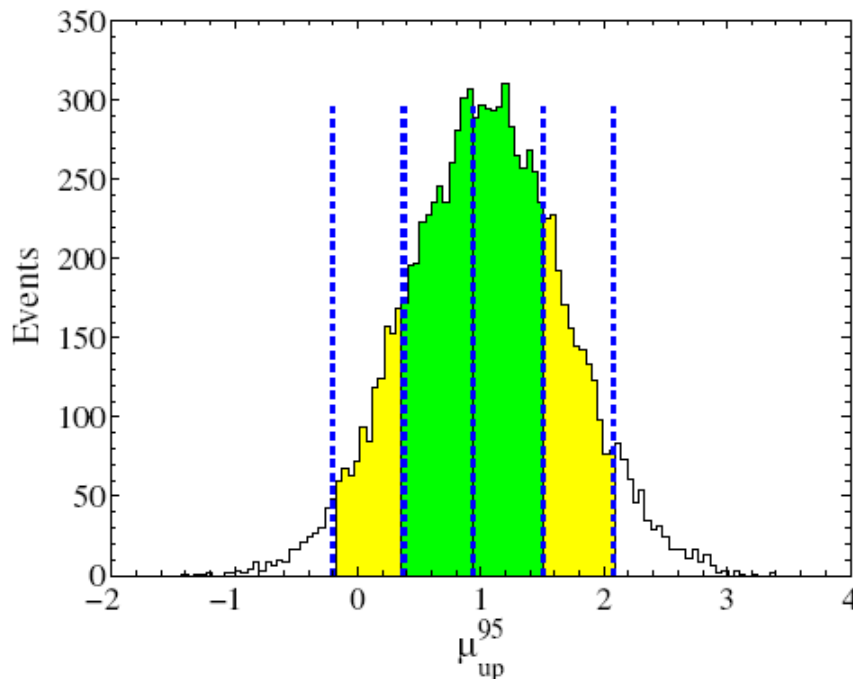
Increases “effective” p -value when the two distributions become close (prevents exclusion if sensitivity is low).

Setting upper limits on $\mu = \sigma/\sigma_{\text{SM}}$

Carry out the CLs procedure for the parameter $\mu = \sigma/\sigma_{\text{SM}}$, resulting in an upper limit μ_{up} .

In, e.g., a Higgs search, this is done for each value of m_{H} .

At a given value of m_{H} , we have an observed value of μ_{up} , and we can also find the distribution $f(\mu_{\text{up}}|0)$:



$\pm 1\sigma$ (green) and $\pm 2\sigma$ (yellow) bands from toy MC;

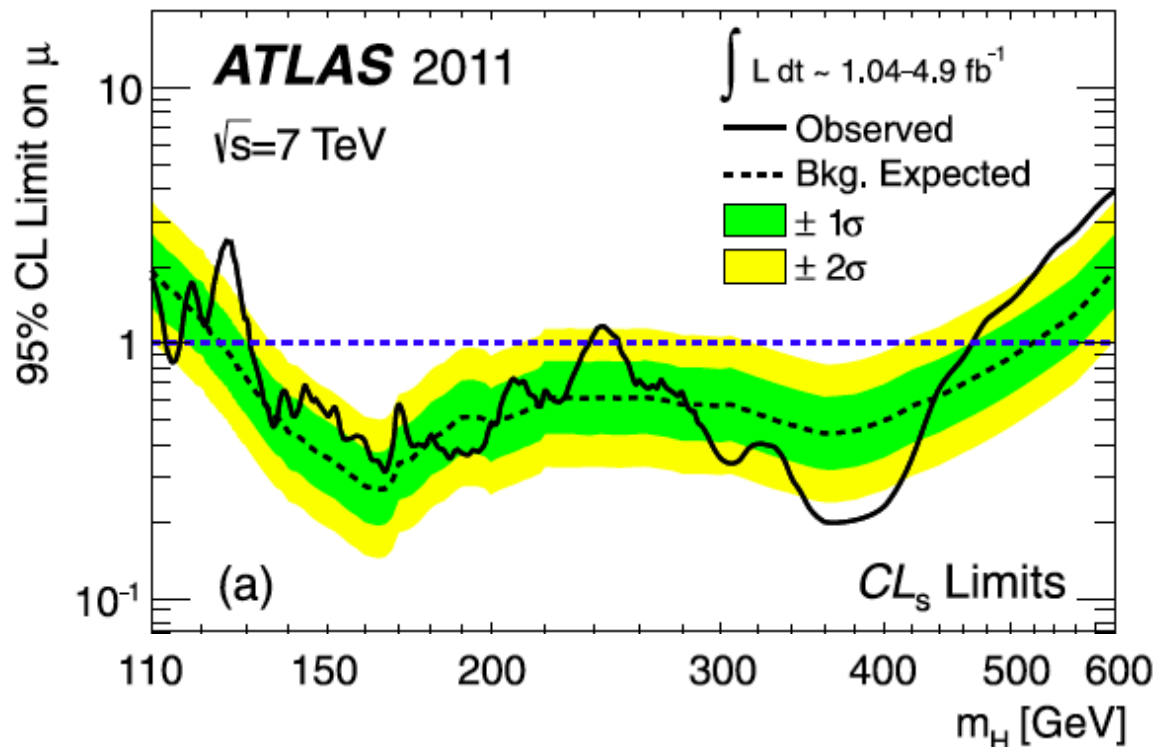
Vertical lines from asymptotic formulae.

How to read the green and yellow limit plots

For every value of m_H , find the CLs 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 $\pm 1\sigma$ (2σ) regions of this distribution.



ATLAS, Phys. Lett.
B 710 (2012) 49-66

Expected discovery significance for counting experiment with background uncertainty

I. Discovery sensitivity for counting experiment with b known:

(a) $\frac{s}{\sqrt{b}}$

(b) Profile likelihood ratio test & Asimov: $\sqrt{2 \left((s+b) \ln \left(1 + \frac{s}{b} \right) - s \right)}$

II. Discovery sensitivity with uncertainty in b , σ_b :

(a) $\frac{s}{\sqrt{b + \sigma_b^2}}$

(b) Profile likelihood ratio test & Asimov:

$$\left[2 \left((s+b) \ln \left[\frac{(s+b)(b + \sigma_b^2)}{b^2 + (s+b)\sigma_b^2} \right] - \frac{b^2}{\sigma_b^2} \ln \left[1 + \frac{\sigma_b^2 s}{b(b + \sigma_b^2)} \right] \right) \right]^{1/2}$$

Counting experiment with known background

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

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

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 $s + b$, $n \rightarrow x \sim \text{Gaussian}(\mu, \sigma)$, $\mu = s + b$, $\sigma = \sqrt{s + b}$.

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

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

Significance for rejecting $s = 0$ is therefore

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

Expected (median) significance assuming signal rate s is

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

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:

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

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

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

Approximate Poisson significance (continued)

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

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

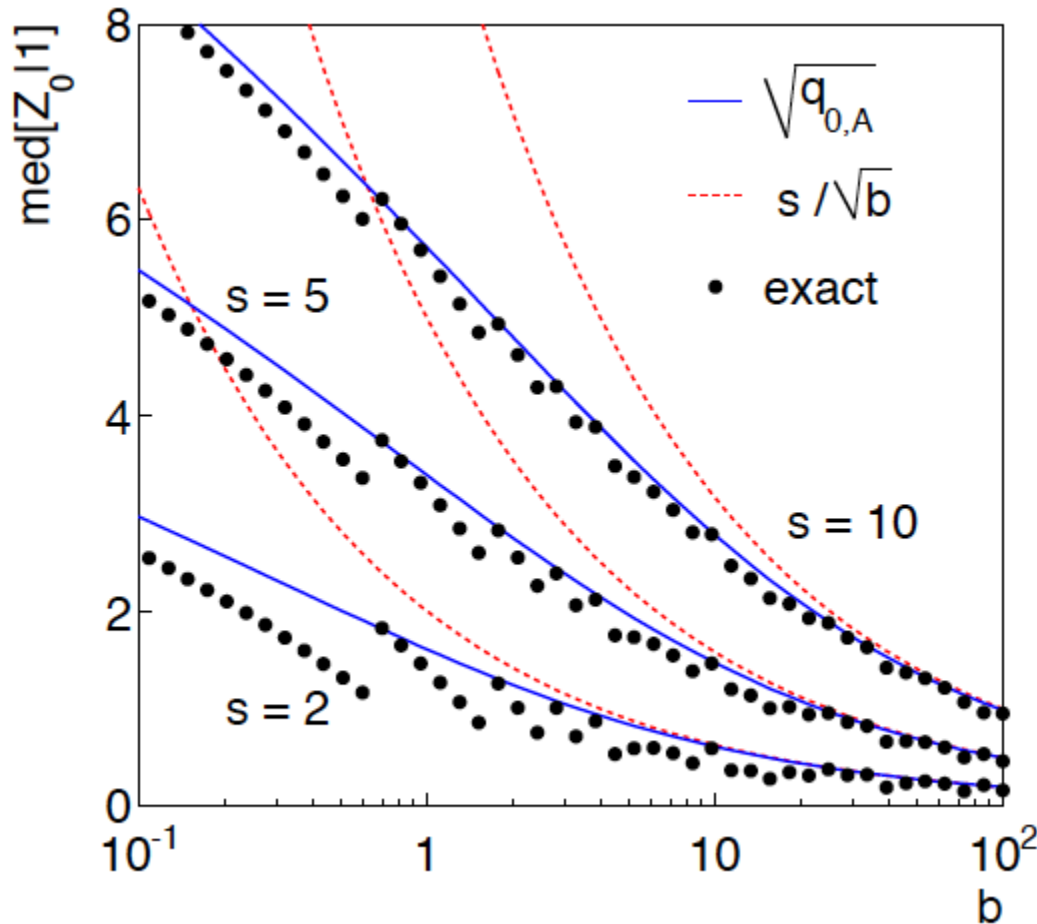
To find $\text{median}[Z|s]$, let $n \rightarrow s + b$ (i.e., the Asimov data set):

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

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

$n \sim \text{Poisson}(s+b)$, 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 $\sqrt{q_{0,A}}$ good approx.
for broad range of s, b .

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

Extending s/\sqrt{b} to case where b uncertain

The intuitive explanation of s/\sqrt{b} is that it compares the signal, s , to the standard deviation of n assuming no signal, \sqrt{b} .

Now suppose the value of b is uncertain, characterized by a standard deviation σ_b .

A reasonable guess is to replace \sqrt{b} by the quadratic sum of \sqrt{b} and σ_b , i.e.,

$$\text{med}[Z|s] = \frac{s}{\sqrt{b + \sigma_b^2}}$$

This has been used to optimize some analyses e.g. where σ_b cannot be neglected.

Profile likelihood with b uncertain

This is the well studied “on/off” problem: Cranmer 2005; Cousins, Linnemann, and Tucker 2008; Li and Ma 1983,...

Measure two Poisson distributed values:

$n \sim \text{Poisson}(s+b)$ (primary or “search” measurement)

$m \sim \text{Poisson}(\tau b)$ (control measurement, τ known)

The likelihood function is

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

Use this to construct profile likelihood ratio (b is nuisance parameter):

$$\lambda(0) = \frac{L(0, \hat{\hat{b}}(0))}{L(\hat{s}, \hat{b})}$$

Ingredients for profile likelihood ratio

To construct profile likelihood ratio from this need estimators:

$$\hat{s} = n - m/\tau ,$$

$$\hat{b} = m/\tau ,$$

$$\hat{\hat{b}}(s) = \frac{n + m - (1 + \tau)s + \sqrt{(n + m - (1 + \tau)s)^2 + 4(1 + \tau)sm}}{2(1 + \tau)} .$$

and in particular to test for discovery ($s = 0$),

$$\hat{\hat{b}}(0) = \frac{n + m}{1 + \tau}$$

Asymptotic significance

Use profile likelihood ratio for q_0 , and then from this get discovery significance using asymptotic approximation (Wilks' theorem):

$$Z = \sqrt{q_0};$$
$$= \left[-2 \left(n \ln \left[\frac{n+m}{(1+\tau)n} \right] + m \ln \left[\frac{\tau(n+m)}{(1+\tau)m} \right] \right) \right]^{1/2}$$

for $n > \hat{b}$ and $Z = 0$ otherwise.

Essentially same as in:

Robert D. Cousins, James T. Linnemann and Jordan Tucker, NIM A 595 (2008) 480–501; arXiv:physics/0702156.

Tipei Li and Yuqian Ma, Astrophysical Journal 272 (1983) 317–324.

Asimov approximation for median significance

To get median discovery significance, replace n , m by their expectation values assuming background-plus-signal model:

$$n \rightarrow s + b$$

$$m \rightarrow \tau b$$

$$Z_A = \left[-2 \left((s + b) \ln \left[\frac{s + (1 + \tau)b}{(1 + \tau)(s + b)} \right] + \tau b \ln \left[1 + \frac{s}{(1 + \tau)b} \right] \right) \right]^{1/2}$$

Or use the variance of $\hat{b} = m/\tau$, $V[\hat{b}] \equiv \sigma_b^2 = \frac{b}{\tau}$, to eliminate τ :

$$Z_A = \left[2 \left((s + b) \ln \left[\frac{(s + b)(b + \sigma_b^2)}{b^2 + (s + b)\sigma_b^2} \right] - \frac{b^2}{\sigma_b^2} \ln \left[1 + \frac{\sigma_b^2 s}{b(b + \sigma_b^2)} \right] \right) \right]^{1/2}$$

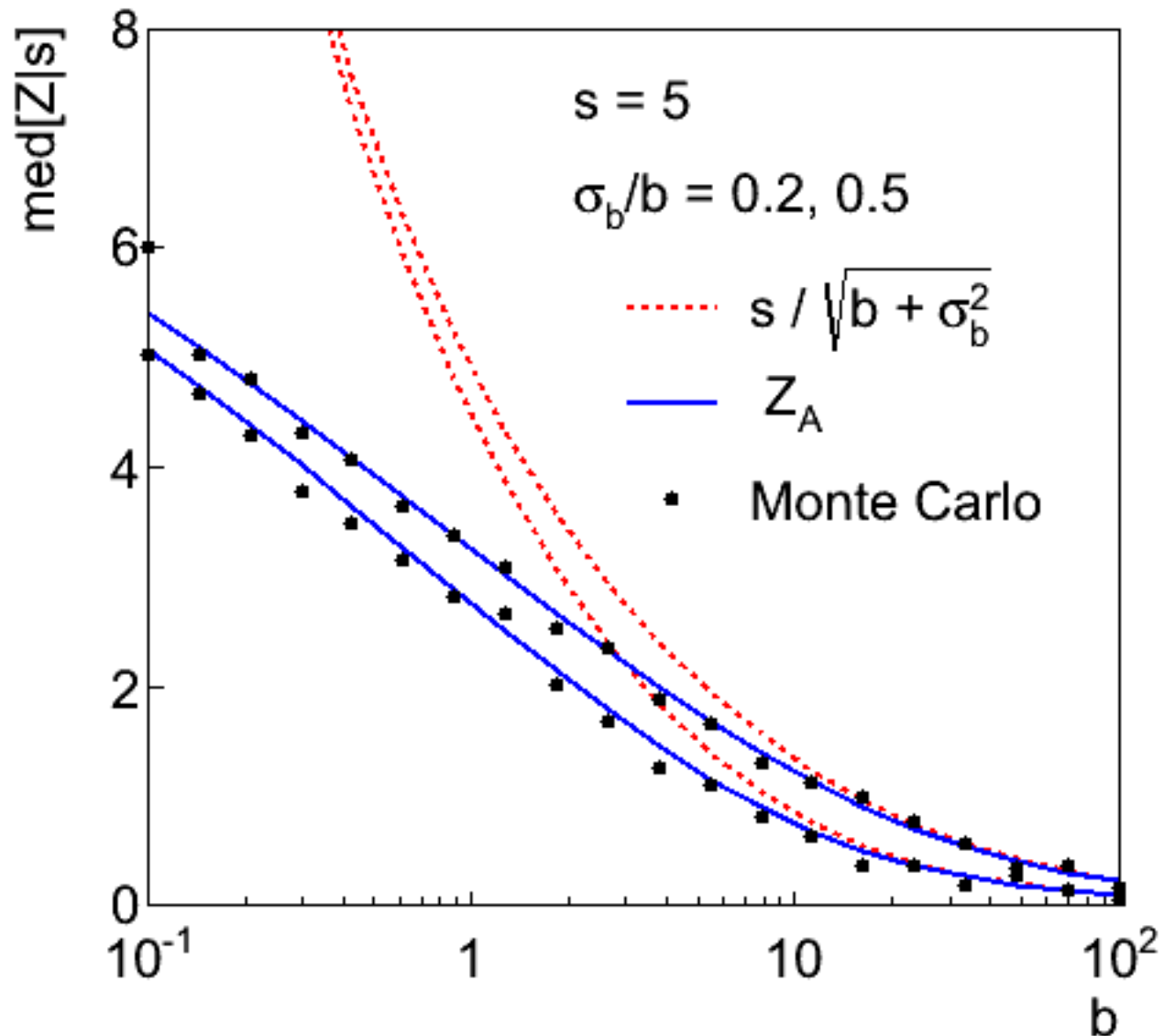
Limiting cases

Expanding the Asimov formula in powers of s/b and σ_b^2/b ($= 1/\tau$) gives

$$Z_A = \frac{s}{\sqrt{b + \sigma_b^2}} \left(1 + \mathcal{O}(s/b) + \mathcal{O}(\sigma_b^2/b) \right)$$

So the “intuitive” formula can be justified as a limiting case of the significance from the profile likelihood ratio test evaluated with the Asimov data set.

Testing the formulae: $s = 5$



Using sensitivity to optimize a cut

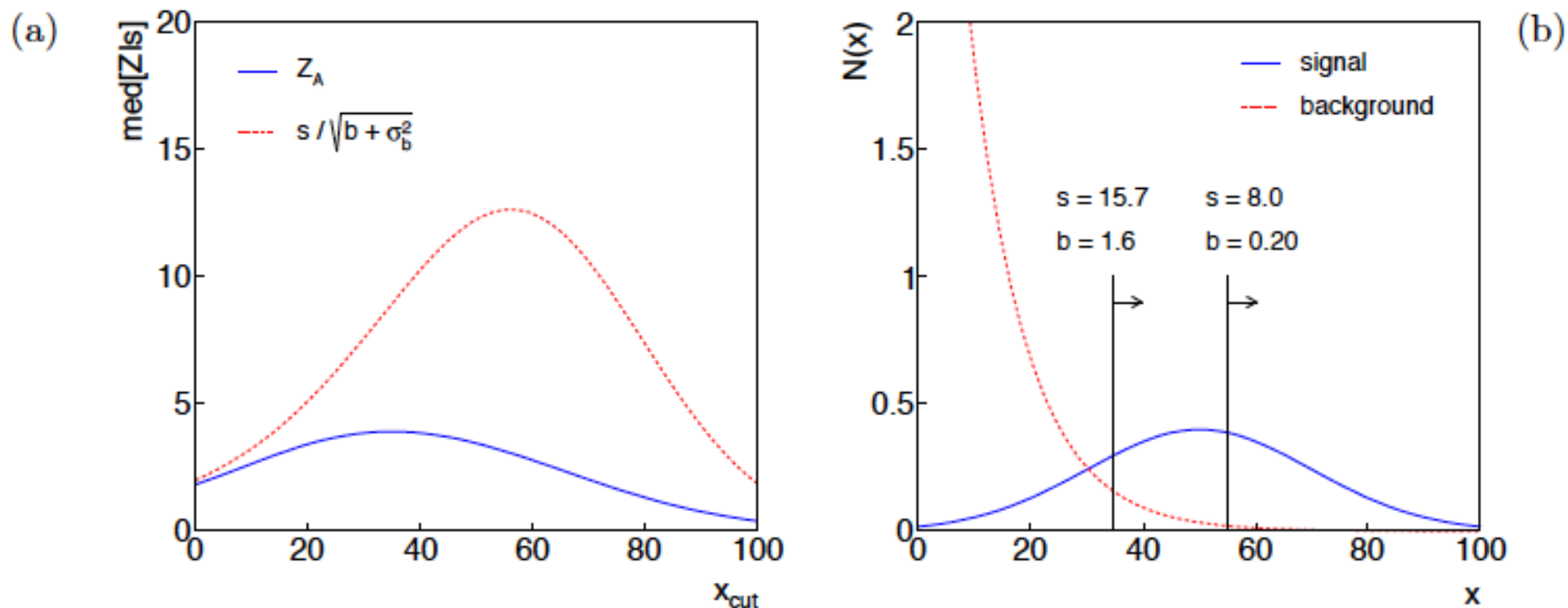


Figure 1: (a) The expected significance as a function of the cut value x_{cut} ; (b) the distributions of signal and background with the optimal cut value indicated.

Summary on discovery sensitivity

Simple formula for expected discovery significance based on profile likelihood ratio test and Asimov approximation:

$$Z_A = \left[2 \left((s + b) \ln \left[\frac{(s + b)(b + \sigma_b^2)}{b^2 + (s + b)\sigma_b^2} \right] - \frac{b^2}{\sigma_b^2} \ln \left[1 + \frac{\sigma_b^2 s}{b(b + \sigma_b^2)} \right] \right) \right]^{1/2}$$

For large b , all formulae OK.

For small b , s/\sqrt{b} and $s/\sqrt{(b+\sigma_b^2)}$ overestimate the significance.

Could be important in optimization of searches with low background.

Formula maybe also OK if model is not simple on/off experiment, e.g., several background control measurements (checking this).

Finally

Three days only enough for a brief introduction to:

- Statistical tests for discovery and limits

- Parameter estimation

- Experimental sensitivity

Many other important topics:

- Bayesian methods, MCMC

- Unfolding

- The look-elsewhere effect, etc., etc.

Final thought: once the basic formalism is understood, most of the work focuses on building the model, i.e., writing down the likelihood, e.g., $P(\mathbf{x}|\theta)$, and including in it enough parameters to adequately describe the data (true for both Bayesian and frequentist approaches).

Extra slides

Some distributions

<u>Distribution/pdf</u>	<u>Example use in HEP</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

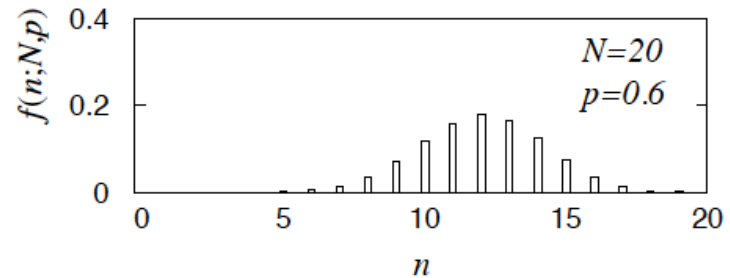
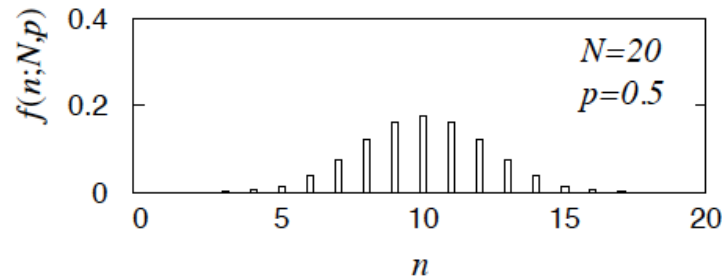
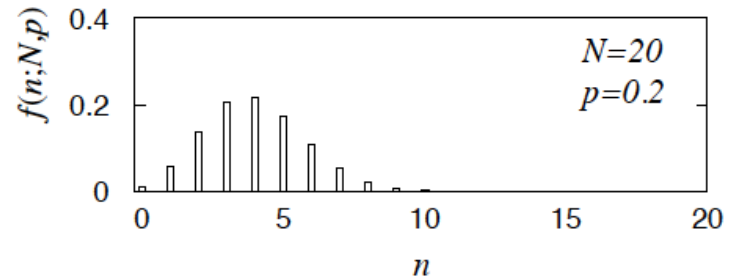
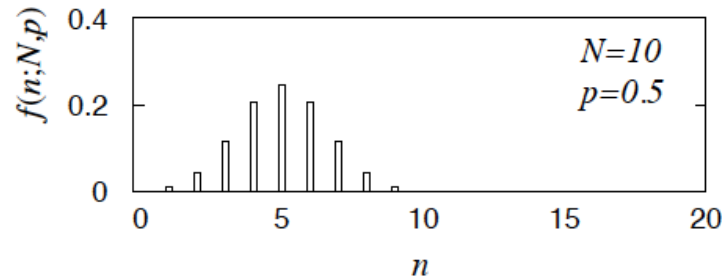
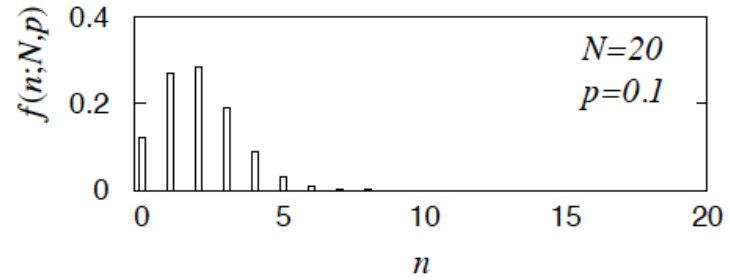
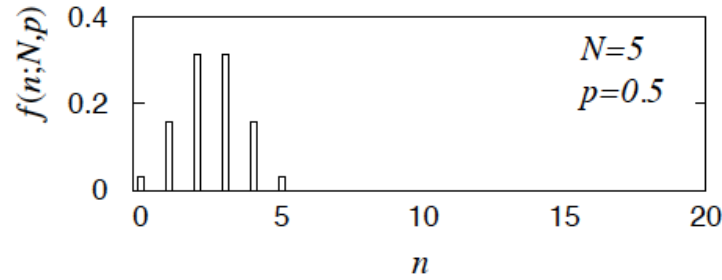
For the expectation value and variance we find:

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

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

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:

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

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

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

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

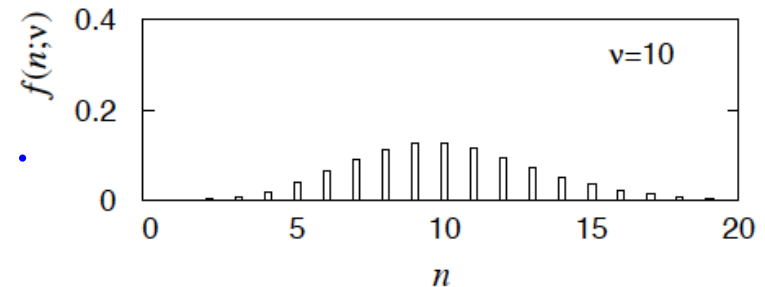
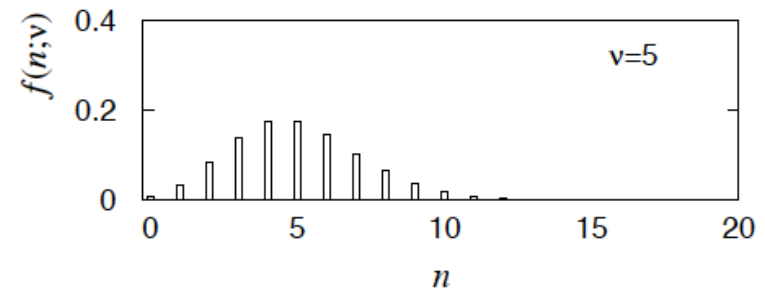
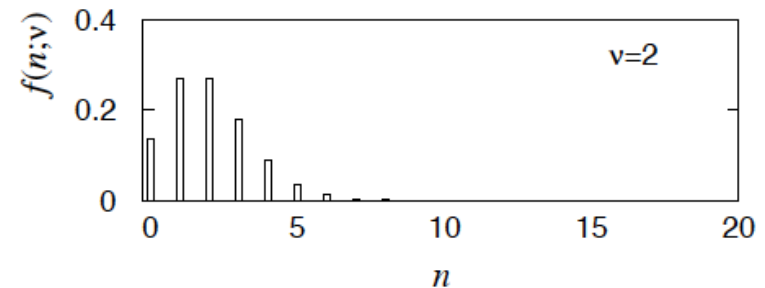
$$N \rightarrow \infty, \quad p \rightarrow 0, \quad E[n] = Np \rightarrow \nu .$$

→ 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 $\nu = \sigma \int L dt$.



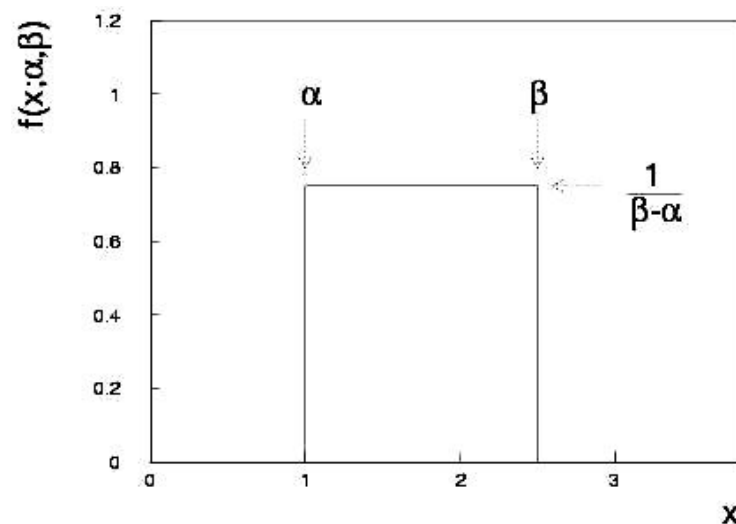
Uniform distribution

Consider a continuous r.v. x with $-\infty < x < \infty$. Uniform pdf is:

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

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

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



N.B. For any r.v. x with cumulative distribution $F(x)$, $y = F(x)$ is uniform in $[0, 1]$.

Example: for $\pi^0 \rightarrow \gamma\gamma$, E_γ is uniform in $[E_{\min}, E_{\max}]$, with

$$E_{\min} = \frac{1}{2}E_\pi(1 - \beta), \quad E_{\max} = \frac{1}{2}E_\pi(1 + \beta)$$

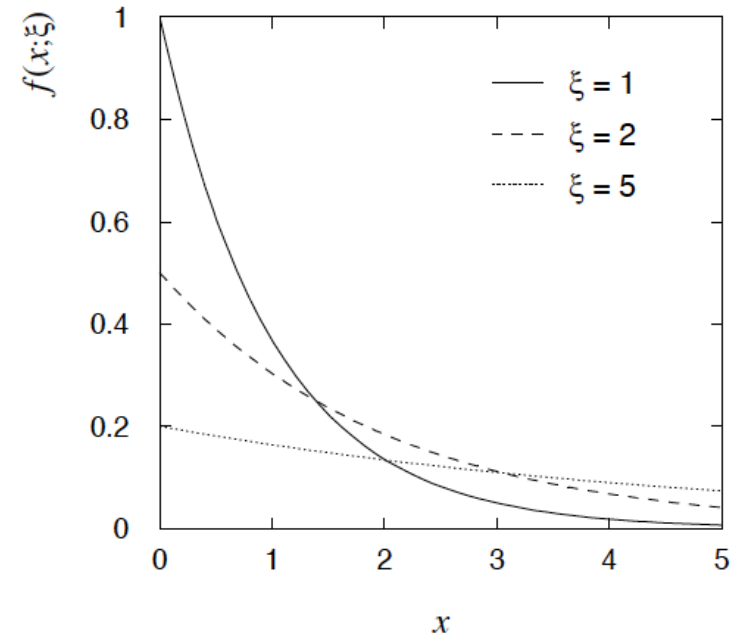
Exponential distribution

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

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

$$E[x] = \xi$$

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



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): $f(t - t_0 | t \geq t_0) = f(t)$

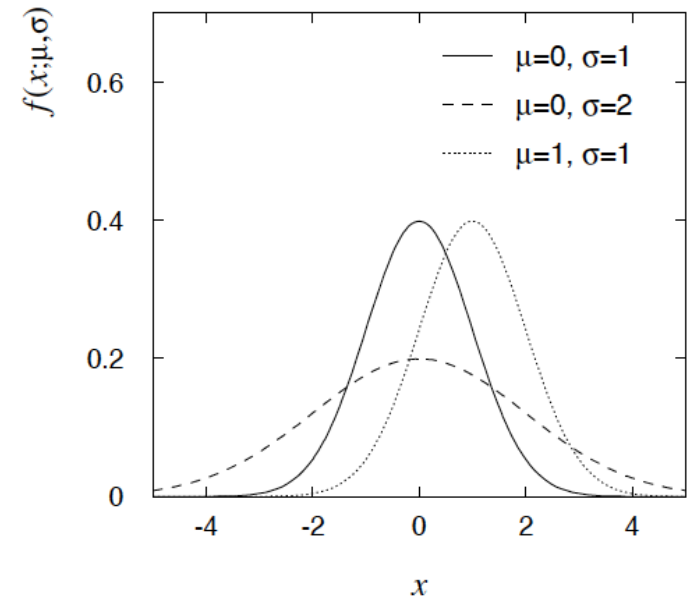
Gaussian distribution

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

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

$$E[x] = \mu \quad (\text{N.B. often } \mu, \sigma^2 \text{ denote mean, variance of any}$$

$$V[x] = \sigma^2 \quad \text{r.v., not only Gaussian.})$$



Special case: $\mu = 0, \sigma^2 = 1$ ('standard Gaussian'):

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

If $y \sim \text{Gaussian with } \mu, \sigma^2$, then $x = (y - \mu) / \sigma$ follows $\varphi(x)$.

Gaussian pdf and the Central Limit Theorem

The Gaussian pdf is so useful because almost any random variable that is a sum of a large number of small contributions follows it. This follows from the Central Limit Theorem:

For n independent r.v.s x_i with finite variances σ_i^2 , otherwise arbitrary pdfs, consider the sum

$$y = \sum_{i=1}^n x_i$$

In the limit $n \rightarrow \infty$, y is a Gaussian r.v. with

$$E[y] = \sum_{i=1}^n \mu_i \quad V[y] = \sum_{i=1}^n \sigma_i^2$$

Measurement errors are often the sum of many contributions, so frequently measured values can be treated as Gaussian r.v.s.

Central Limit Theorem (2)

The CLT can be proved using characteristic functions (Fourier transforms), see, e.g., SDA Chapter 10.

For finite n , the theorem is approximately valid to the extent that the fluctuation of the sum is not dominated by one (or few) terms.



Beware of measurement errors with non-Gaussian tails.

Good example: velocity component v_x of air molecules.

OK example: total deflection due to multiple Coulomb scattering.
(Rare large angle deflections give non-Gaussian tail.)

Bad example: energy loss of charged particle traversing thin gas layer. (Rare collisions make up large fraction of energy loss, cf. Landau pdf.)

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,

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

For $n = 2$ this is

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

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

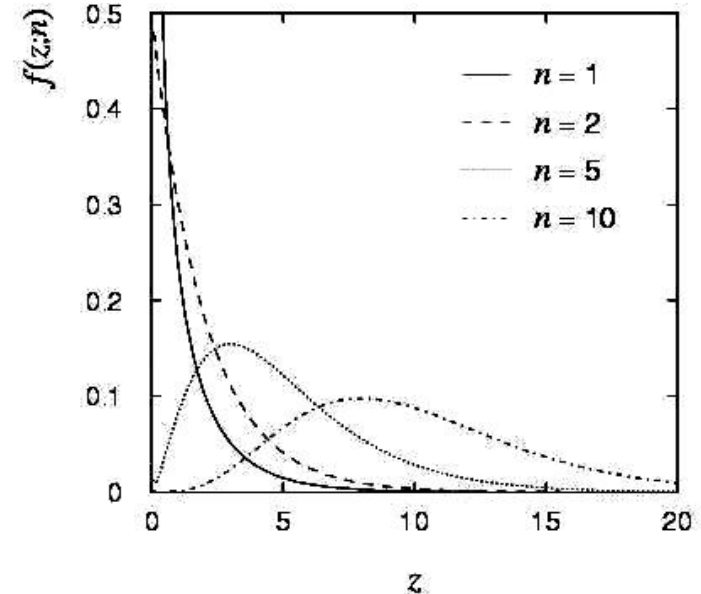
Chi-square (χ^2) distribution

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

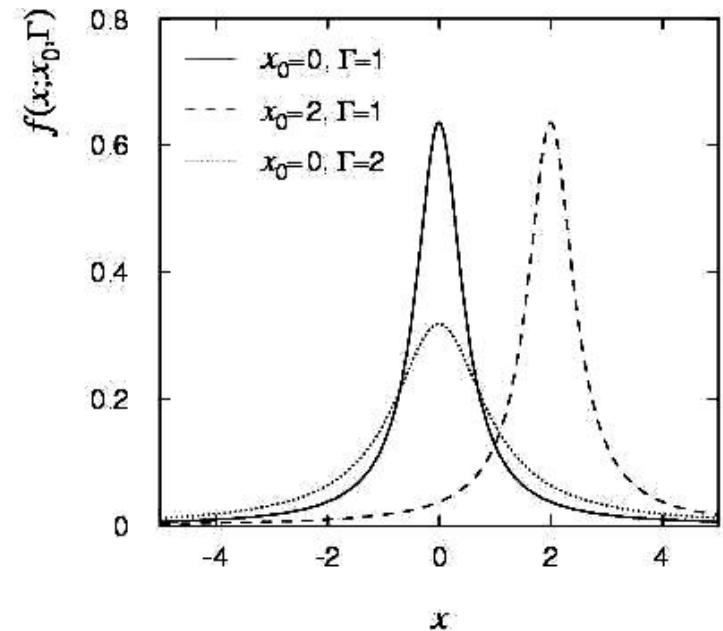
$$f(x; \Gamma, x_0) = \frac{1}{\pi} \frac{\Gamma/2}{\Gamma^2/4 + (x - x_0)^2}$$

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

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

x_0 = 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

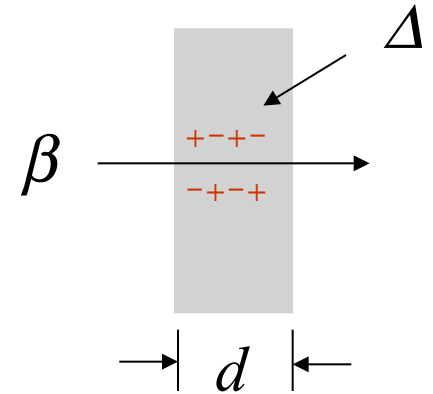
For a charged particle with $\beta = v/c$ traversing a layer of matter of thickness d , the energy loss Δ follows the Landau pdf:

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

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

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

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



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

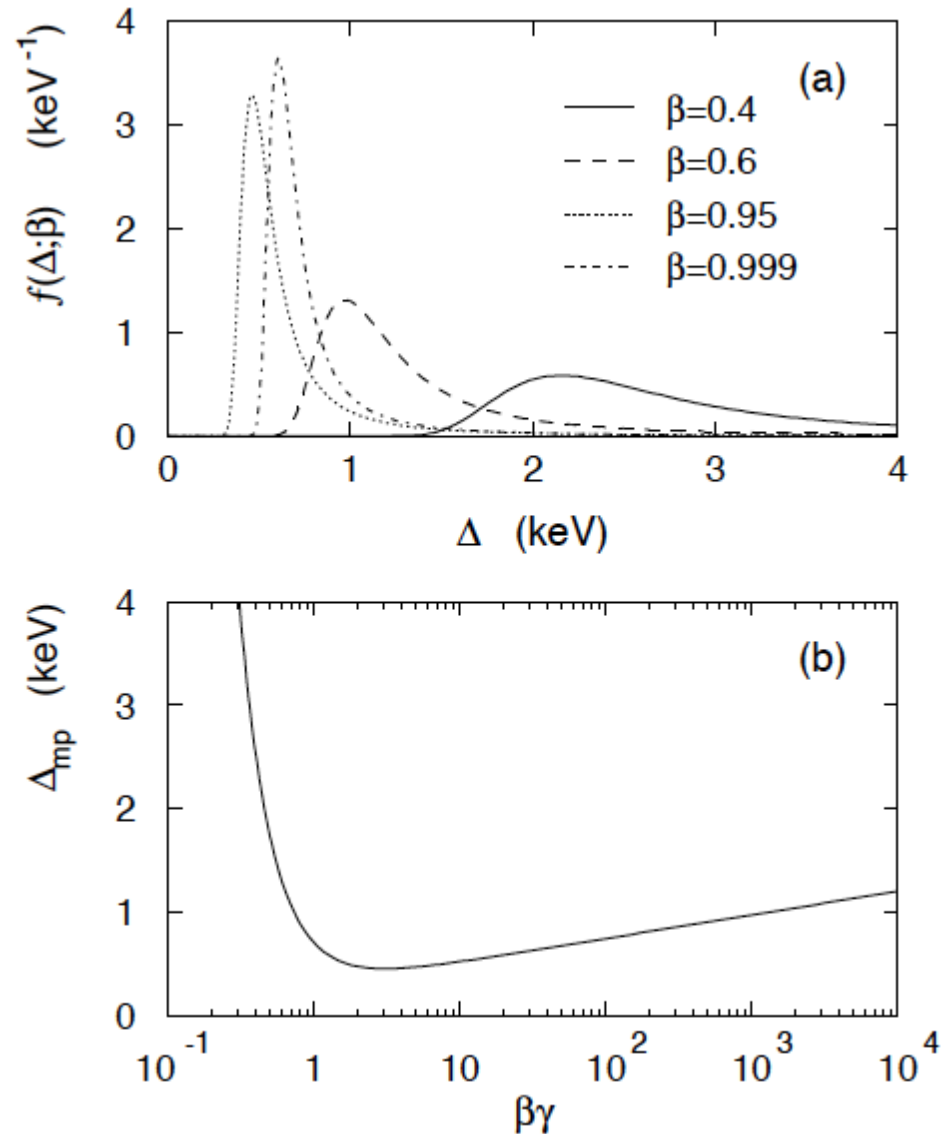
Landau distribution (2)

Long ‘Landau tail’

→ all moments ∞

Mode (most probable value) sensitive to β ,

→ particle i.d.



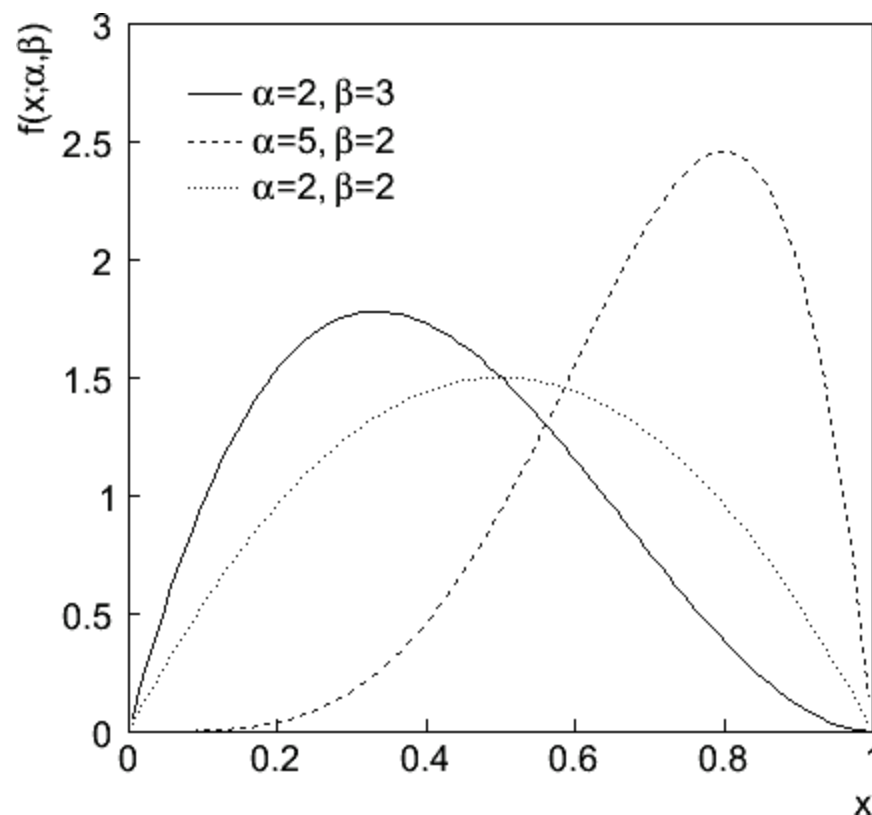
Beta distribution

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

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

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

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



Gamma distribution

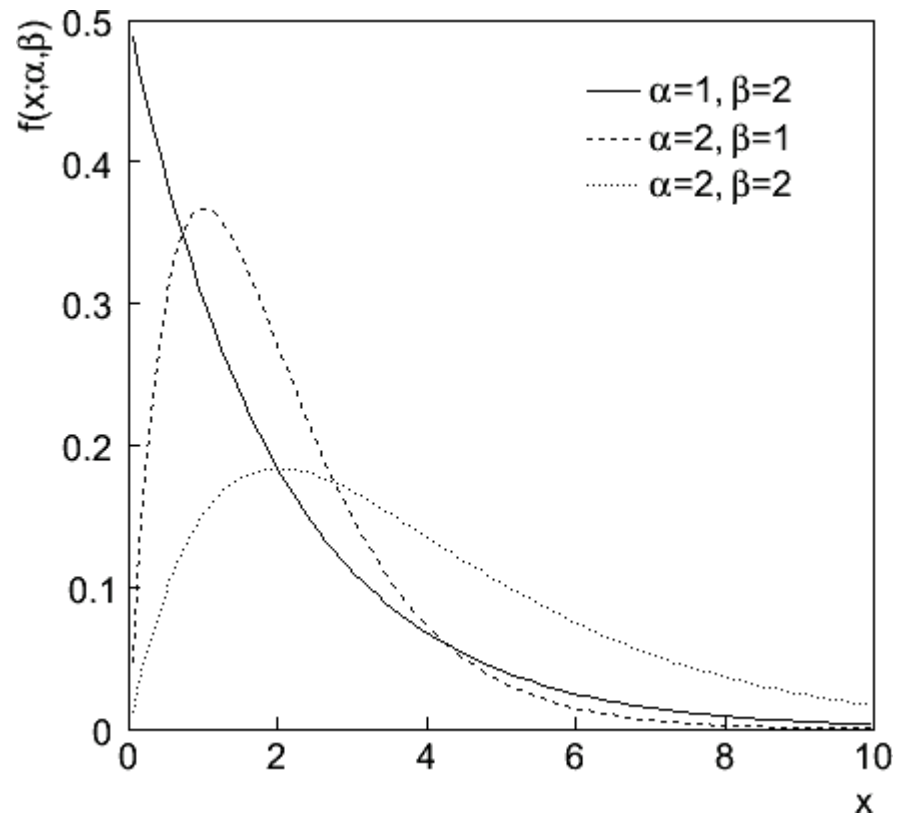
$$f(x; \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$$

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

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

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

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

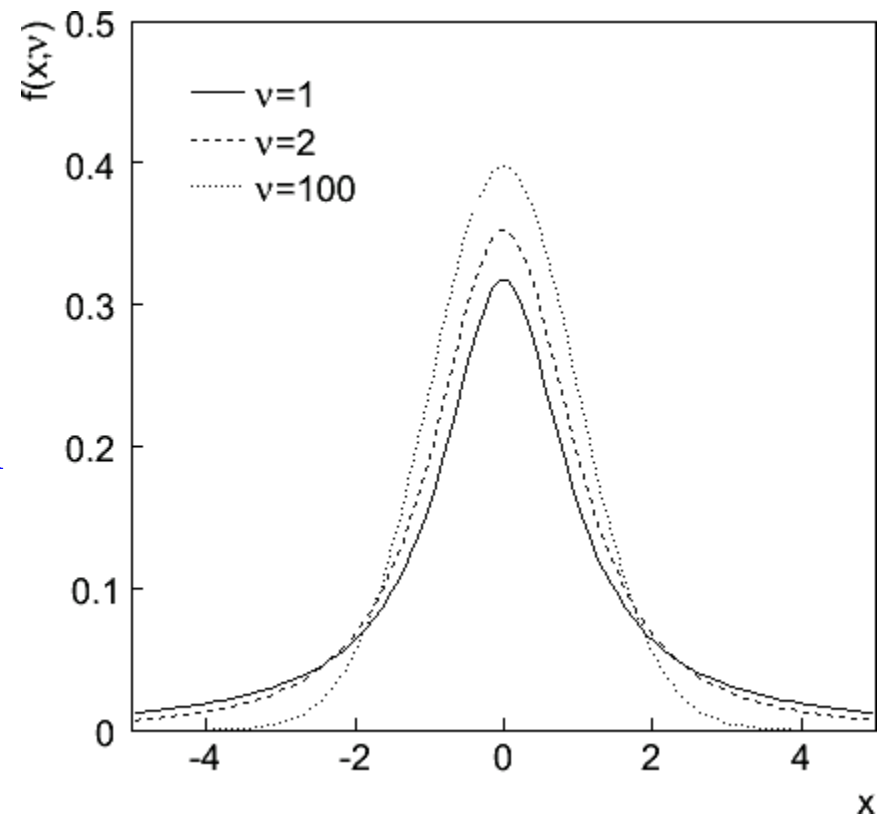
$$E[x] = 0 \quad (\nu > 1)$$

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

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

$\nu = 1$ gives Cauchy,

$\nu \rightarrow \infty$ gives Gaussian.



Goodness of fit from the likelihood ratio

Suppose we model data using a likelihood $L(\boldsymbol{\mu})$ that depends on N parameters $\boldsymbol{\mu} = (\mu_1, \dots, \mu_N)$. Define the statistic

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

Value of $t_{\boldsymbol{\mu}}$ reflects agreement between hypothesized $\boldsymbol{\mu}$ and the data.

Good agreement means $\hat{\boldsymbol{\mu}} \approx \boldsymbol{\mu}$, so $t_{\boldsymbol{\mu}}$ is small;

Larger $t_{\boldsymbol{\mu}}$ means less compatibility between data and $\boldsymbol{\mu}$.

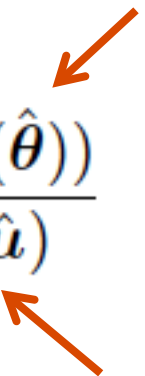
Quantify “goodness of fit” with p -value: $p_{\boldsymbol{\mu}} = \int_{t_{\boldsymbol{\mu},\text{obs}}}^{\infty} f(t_{\boldsymbol{\mu}}|\boldsymbol{\mu}) dt_{\boldsymbol{\mu}}$

Likelihood ratio (2)

Now suppose the parameters $\mu = (\mu_1, \dots, \mu_N)$ can be determined by another set of parameters $\theta = (\theta_1, \dots, \theta_M)$, with $M < N$.

E.g. in LS fit, use $\mu_i = \mu(x_i; \theta)$ where x is a control variable.

Define the statistic

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


fit M parameters

fit N parameters

Use q_μ to test hypothesized functional form of $\mu(x; \theta)$.

To get p -value, need pdf $f(q_\mu | \mu)$.

Wilks' Theorem (1938)

Wilks' Theorem: if the hypothesized parameters $\mu = (\mu_1, \dots, \mu_N)$ are true then in the large sample limit (and provided certain conditions are satisfied) t_μ and q_μ follow chi-square distributions.

For case with $\mu = (\mu_1, \dots, \mu_N)$ fixed in numerator:

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

$$f(t_\mu | \mu) \sim \chi_N^2$$

Or if M parameters adjusted in numerator,

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

$$f(q_\mu | \mu) \sim \chi_{N-M}^2$$

degrees of
freedom



S.S. Wilks, *The large-sample distribution of the likelihood ratio for testing composite hypotheses*, Ann. Math. Statist. **9** (1938) 60-2.

Goodness of fit with Gaussian data

Suppose the data are N independent Gaussian distributed values:

$$y_i \sim \text{Gauss}(\mu_i, \sigma_i), \quad i = 1, \dots, N$$

want to estimate

known

Likelihood:

$$L(\mu) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(y_i - \mu_i)^2 / 2\sigma_i^2}$$

Log-likelihood:

$$\ln L(\mu) = -\frac{1}{2} \sum_{i=1}^N \frac{(y_i - \mu_i)^2}{\sigma_i^2} + C$$

ML estimators:

$$\hat{\mu}_i = y_i \quad i = 1, \dots, N$$

Likelihood ratios for Gaussian data

The goodness-of-fit statistics become

$$t_{\mu} = -2 \ln \frac{L(\mu)}{L(\hat{\mu})} = \sum_{i=1}^N \frac{(y_i - \mu_i)^2}{\sigma_i^2} \quad f(t_{\mu} | \mu) \sim \chi_N^2$$

$$q_{\mu} = -2 \ln \frac{L(\mu(\hat{\theta}))}{L(\hat{\mu})} = \sum_{i=1}^N \frac{(y_i - \mu_i(\hat{\theta}))^2}{\sigma_i^2} \quad f(q_{\mu} | \mu) \sim \chi_{N-M}^2$$

So Wilks' theorem formally states the well-known property of the minimized chi-squared from an LS fit.

Likelihood ratio for Poisson data

Suppose the data are a set of values $\mathbf{n} = (n_1, \dots, n_N)$, e.g., the numbers of events in a histogram with N bins.

Assume $n_i \sim \text{Poisson}(\nu_i)$, $i = 1, \dots, N$, all independent.

Goal is to estimate $\boldsymbol{\nu} = (\nu_1, \dots, \nu_N)$.

Likelihood:
$$L(\boldsymbol{\nu}) = \prod_{i=1}^N \frac{\nu_i^{n_i}}{n_i!} e^{-\nu_i}$$

Log-likelihood:
$$\ln L(\boldsymbol{\nu}) = \sum_{i=1}^N [n_i \ln \nu_i - \nu_i] + C$$

ML estimators:
$$\hat{\nu}_i = n_i, \quad i = 1, \dots, N$$

Goodness of fit with Poisson data

The likelihood ratio statistic (all parameters fixed in numerator):

$$\begin{aligned}t_{\boldsymbol{\nu}} &= -2 \ln \frac{L(\boldsymbol{\nu})}{L(\hat{\boldsymbol{\nu}})} \\&= -2 \sum_{i=1}^N \left[n_i \ln \frac{\nu_i}{\hat{\nu}_i} - \nu_i + \hat{\nu}_i \right] \\&= -2 \sum_{i=1}^N \left[n_i \ln \frac{\nu_i}{n_i} - \nu_i + n_i \right]\end{aligned}$$

Wilks' theorem: $f(t_{\boldsymbol{\nu}}|\boldsymbol{\nu}) \sim \chi_N^2$

Goodness of fit with Poisson data (2)

Or with M fitted parameters in numerator:

$$q_{\nu} = -2 \ln \frac{L(\nu(\hat{\theta}))}{L(\hat{\nu})} = -2 \sum_{i=1}^N \left[n_i \ln \frac{\nu_i(\hat{\theta})}{n_i} - \nu_i(\hat{\theta}) + n_i \right]$$

Wilks' theorem: $f(q_{\nu}|\nu) \sim \chi_{N-M}^2$

Use t_{μ} , q_{μ} to quantify goodness of fit (p -value).

Sampling distribution from Wilks' theorem (chi-square).

Exact in large sample limit; in practice good approximation for surprisingly small n_i (\sim several).

Goodness of fit with multinomial data

Similar if data $\mathbf{n} = (n_1, \dots, n_N)$ follow multinomial distribution:

$$P(\mathbf{n}|\mathbf{p}, n_{\text{tot}}) = \frac{n_{\text{tot}}!}{n_1! n_2! \dots n_N!} p_1^{n_1} p_2^{n_2} \dots p_N^{n_N}$$

E.g. histogram with N bins but fix: $n_{\text{tot}} = \sum_{i=1}^N n_i$

$$\text{Log-likelihood: } \ln L(\boldsymbol{\nu}) = \sum_{i=1}^N n_i \ln \frac{\nu_i}{n_{\text{tot}}} + C \quad (\nu_i = p_i n_{\text{tot}})$$

ML estimators: $\hat{\nu}_i = n_i$ (Only $N-1$ independent; one is n_{tot} minus sum of rest.)

Goodness of fit with multinomial data (2)

The likelihood ratio statistics become:

$$t_{\nu} = -2 \sum_{i=1}^N n_i \ln \frac{\nu_i}{n_i} \qquad f(t_{\nu} | \nu) \sim \chi_{N-1}^2$$
$$q_{\nu} = -2 \sum_{i=1}^N n_i \ln \frac{\nu_i(\hat{\theta})}{n_i} \qquad f(q_{\nu} | \nu) \sim \chi_{N-M-1}^2$$

One less degree of freedom than in Poisson case because effectively only $N-1$ parameters fitted in denominator.

Estimators and g.o.f. all at once

Evaluate numerators with θ (not its estimator):

$$\chi_P^2(\theta) = -2 \sum_{i=1}^N \left[n_i \ln \frac{\nu_i(\theta)}{n_i} - \nu_i(\theta) + n_i \right] \quad (\text{Poisson})$$

$$\chi_M^2(\theta) = -2 \sum_{i=1}^N n_i \ln \frac{\nu_i(\theta)}{n_i} \quad (\text{Multinomial})$$

These are equal to the corresponding $-2 \ln L(\theta)$, so minimizing them gives the usual ML estimators for θ .

The minimized value gives the statistic q_μ , so we get goodness-of-fit for free.

Steve Baker and Robert D. Cousins, *Clarification of the use of the chi-square and likelihood functions in fits to histograms*, NIM **221** (1984) 437.

Linear LS problem

LS has particularly simple properties if $\lambda(x; \vec{\theta})$ linear in $\vec{\theta}$:

$$\lambda(x; \vec{\theta}) = \sum_{j=1}^m a_j(x) \theta_j$$

where $a_j(x)$ are any linearly independent functions of x .

Matrix notation: let $A_{ij} = a_j(x_i)$,

$$\begin{aligned} \chi^2(\vec{\theta}) &= (\vec{y} - \vec{\lambda})^T V^{-1} (\vec{y} - \vec{\lambda}) \\ &= (\vec{y} - A\vec{\theta})^T V^{-1} (\vec{y} - A\vec{\theta}) \end{aligned}$$

Linear LS problem (2)

Set derivatives with respect to θ_i to zero,

$$\nabla \chi^2 = -2(A^T V^{-1} \vec{y} - A^T V^{-1} A \vec{\theta}) = 0$$

Solve to get the LS estimators,

$$\hat{\vec{\theta}} = (A^T V^{-1} A)^{-1} A^T V^{-1} \vec{y} \equiv B \vec{y}$$

N.B. estimators $\hat{\theta}_i$ are linear functions of the measurements y_i .

Linear LS problem (3)

Error propagation (exact for linear problem) for $U_{ij} = \text{cov}[\hat{\theta}_i, \hat{\theta}_j]$:

$$U = B V B^T = (A^T V^{-1} A)^{-1}$$

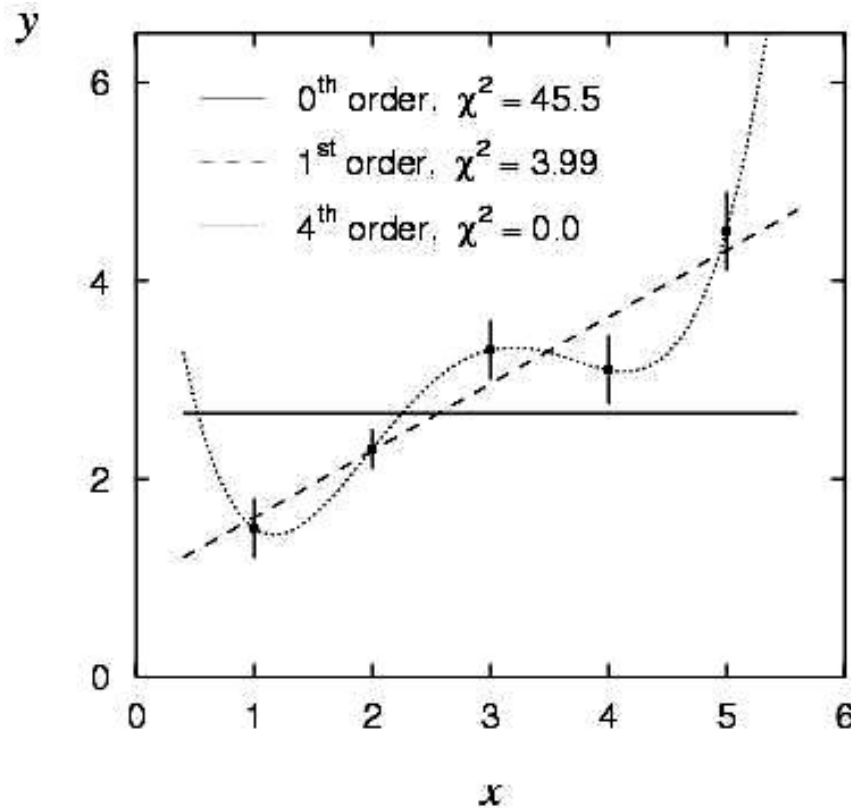
Equivalently, use

$$(U^{-1})_{ij} = \frac{1}{2} \left[\frac{\partial^2 \chi^2}{\partial \theta_i \partial \theta_j} \right]_{\vec{\theta} = \vec{\hat{\theta}}}$$

Equals MVB if y_i Gaussian)

Example of least squares fit

Fit a polynomial of order p : $\lambda(x; \theta_0, \dots, \theta_p) = \sum_{n=0}^p \theta_n x^n$



Variance of LS estimators

In most cases of interest we obtain the variance in a manner similar to ML. E.g. for data \sim Gaussian we have

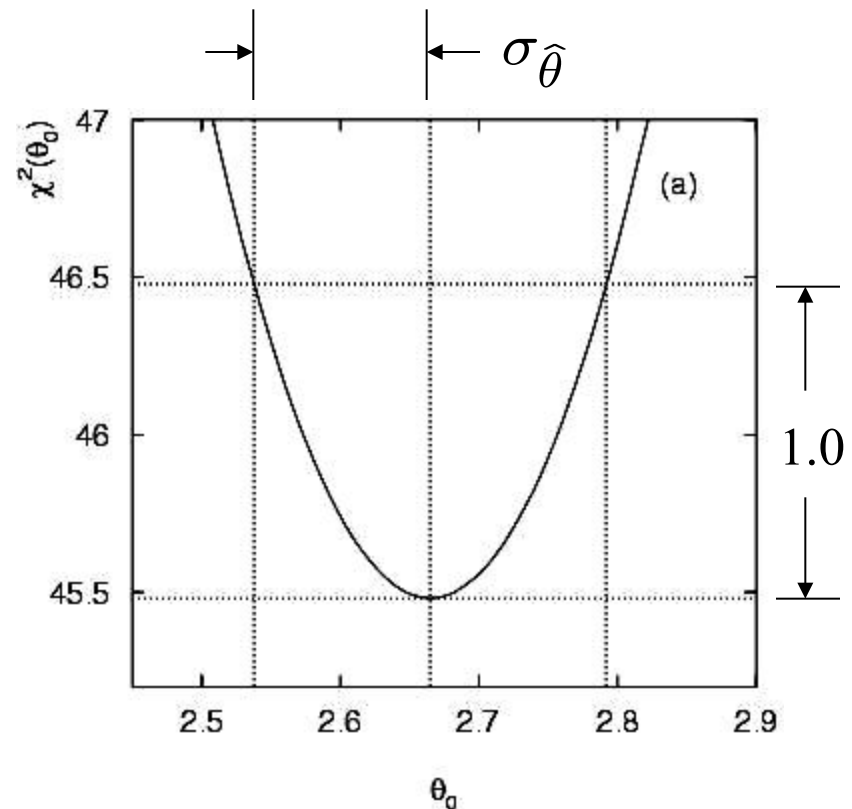
$$\chi^2(\theta) = -2 \ln L(\theta) + C$$

and so

$$\hat{\sigma}_{\hat{\theta}}^2 \approx 2 \left[\frac{\partial^2 \chi^2}{\partial \theta^2} \right]_{\theta=\hat{\theta}}^{-1}$$

or for the graphical method we take the values of θ where

$$\chi^2(\theta) = \chi_{\min}^2 + 1$$



Two-parameter LS fit

2-parameter case (line with nonzero slope):

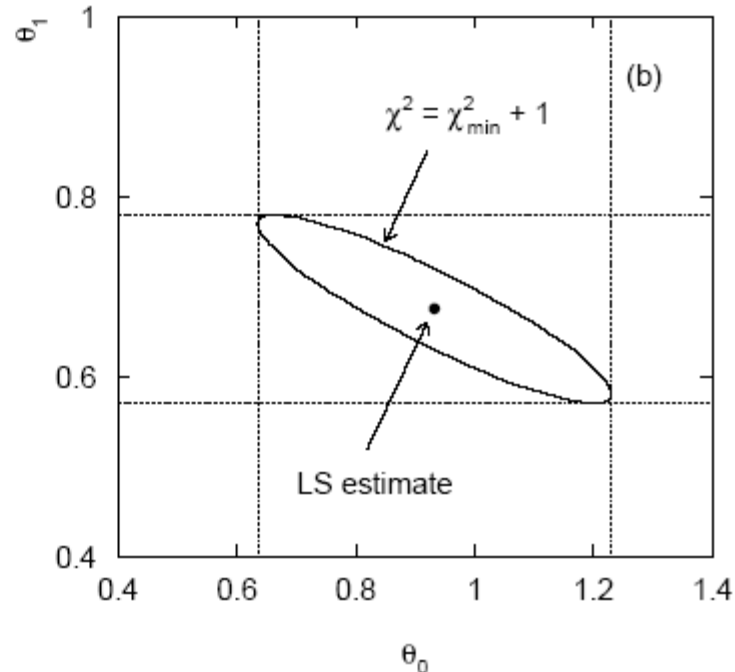
$$\hat{\theta}_0 = 0.93 \pm 0.30,$$

$$\hat{\theta}_1 = 0.68 \pm 0.10$$

$$\widehat{\text{cov}}[\hat{\theta}_0, \hat{\theta}_1] = -0.028$$

$$r = -0.90$$

$$\chi^2 = 3.99$$



Tangent lines $\rightarrow \sigma_{\hat{\theta}_0}, \sigma_{\hat{\theta}_1}$.

Angle of ellipse \rightarrow correlation (same as for ML)

Goodness-of-fit with least squares

The value of the χ^2 at its minimum is a measure of the level of agreement between the data and fitted curve:

$$\chi_{\min}^2 = \sum_{i=1}^N \frac{(y_i - \lambda(x_i; \hat{\theta}))^2}{\sigma_i^2}$$

It can therefore be employed as a goodness-of-fit statistic to test the hypothesized functional form $\lambda(x; \theta)$.

We can show that if the hypothesis is correct, then the statistic $t = \chi_{\min}^2$ follows the chi-square pdf,

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

where the number of degrees of freedom is

n_d = number of data points – number of fitted parameters

Goodness-of-fit with least squares (2)

The chi-square pdf has an expectation value equal to the number of degrees of freedom, so if $\chi^2_{\min} \approx n_d$ the fit is ‘good’.

More generally, find the p -value:
$$p = \int_{\chi^2_{\min}}^{\infty} f(t; n_d) dt$$

This is the probability of obtaining a χ^2_{\min} as high as the one we got, or higher, if the hypothesis is correct.

E.g. for the previous example with 1st order polynomial (line),

$$\chi^2_{\min} = 3.99, \quad n_d = 5 - 2 = 3, \quad p = 0.263$$

whereas for the 0th order polynomial (horizontal line),

$$\chi^2_{\min} = 45.5, \quad n_d = 5 - 1 = 4, \quad p = 3.1 \times 10^{-9}$$

Goodness-of-fit vs. statistical errors

Small statistical error does not mean a good fit (nor vice versa).

Curvature of χ^2 near its minimum \rightarrow statistical errors ($\sigma_{\hat{\theta}}$)

Value of χ^2_{\min} \rightarrow goodness-of-fit

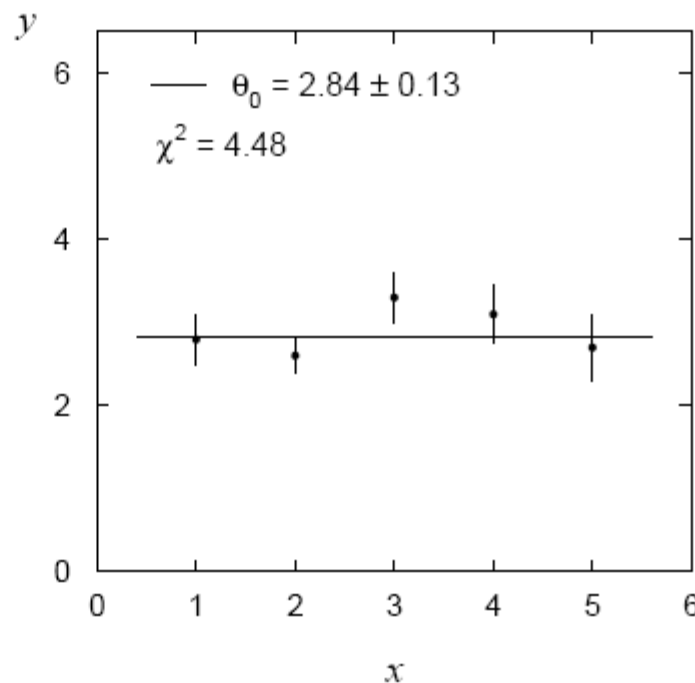
Horizontal line fit, move the data points, keep errors on points same:

$$\hat{\theta}_0 = 2.84 \pm 0.13$$

$$\chi^2_{\min} = 4.48$$

Variance same as before,

now χ^2_{\min} 'good'.



Goodness-of-fit vs. stat. errors (2)

→ $\chi^2(\theta_0)$ shifted down, same curvature as before.

Variance of estimator (statistical error) tells us:

if experiment repeated many times, how wide is the distribution of the estimates $\hat{\theta}$. (Doesn't tell us whether hypothesis correct.)

P -value tells us:

if hypothesis is correct and experiment repeated many times, what fraction will give equal or worse agreement between data and hypothesis according to the statistic χ^2_{\min} .

Low P -value → hypothesis may be wrong → **systematic error**.

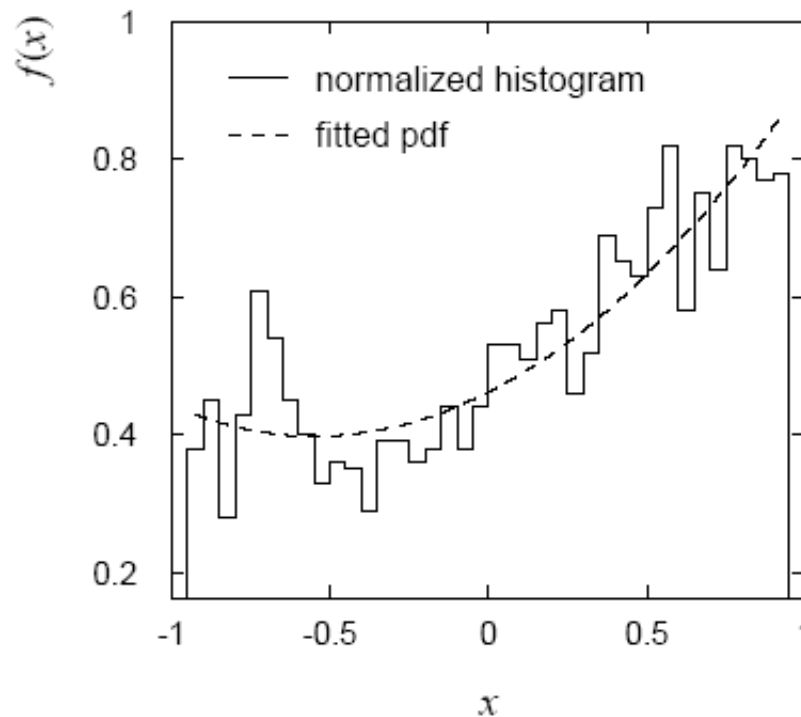
LS with binned data

Histogram:

N bins, n entries.

Hypothesized pdf:

$$f(x; \vec{\theta})$$



We have

y_i = number of entries in bin i ,

$$\lambda_i(\vec{\theta}) = n \int_{x_i^{\min}}^{x_i^{\max}} f(x; \vec{\theta}) dx = np_i(\vec{\theta})$$

LS with binned data (2)

LS fit: minimize

$$\chi^2(\vec{\theta}) = \sum_{i=1}^N \frac{(y_i - \lambda_i(\vec{\theta}))^2}{\sigma_i^2}$$

where $\sigma_i^2 = V[y_i]$, here not known a priori.

Treat the y_i as Poisson r.v.s, in place of true variance take either

$$\sigma_i^2 = \lambda_i(\vec{\theta}) \quad (\text{LS method})$$

$$\sigma_i^2 = y_i \quad (\text{Modified LS method})$$

MLS sometimes easier computationally, but χ^2_{\min} no longer follows chi-square pdf (or is undefined) if some bins have few (or no) entries.

LS with binned data — normalization

Do **not** ‘fit the normalization’:

$$\lambda_i(\vec{\theta}, \nu) = \nu \int_{x_i^{\min}}^{x_i^{\max}} f(x; \vec{\theta}) dx = \nu p_i(\vec{\theta})$$

i.e. introduce adjustable ν , fit along with $\vec{\theta}$.

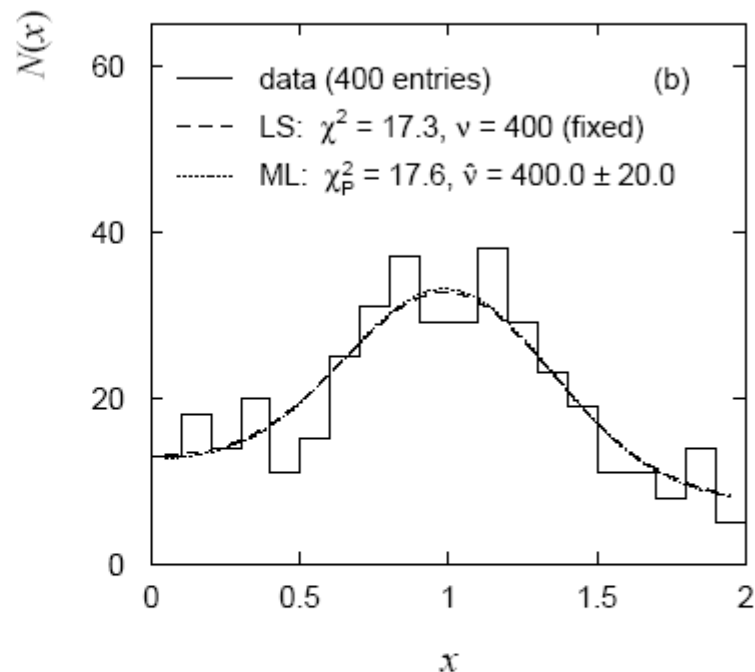
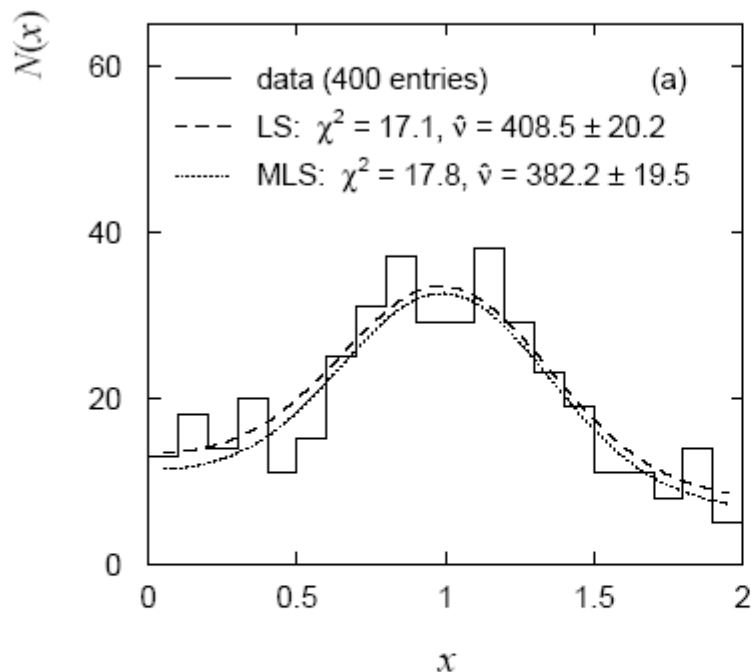
$\hat{\nu}$ is a bad estimator for n (which we know, anyway!)

$$\hat{\nu}_{\text{LS}} = n + \frac{\chi_{\min}^2}{2}$$

$$\hat{\nu}_{\text{MLS}} = n - \chi_{\min}^2$$

LS normalization example

Example with $n = 400$ entries, $N = 20$ bins:



Expect χ^2_{\min} around $N - m$,

→ relative error in \hat{v} large when N large, n small

Either get n directly from data for LS (or better, use ML).

Using LS to combine measurements

Use LS to obtain weighted average of N measurements of λ :

y_i = result of measurement i , $i = 1, \dots, N$;

$\sigma_i^2 = V[y_i]$, assume known;

λ = true value (plays role of θ).

For uncorrelated y_i , minimize

$$\chi^2(\lambda) = \sum_{i=1}^N \frac{(y_i - \lambda)^2}{\sigma_i^2},$$

Set $\frac{\partial \chi^2}{\partial \lambda} = 0$ and solve,

$$\rightarrow \hat{\lambda} = \frac{\sum_{i=1}^N y_i / \sigma_i^2}{\sum_{j=1}^N 1 / \sigma_j^2} \qquad V[\hat{\lambda}] = \frac{1}{\sum_{i=1}^N 1 / \sigma_i^2}$$

Combining correlated measurements with LS

If $\text{cov}[y_i, y_j] = V_{ij}$, minimize

$$\chi^2(\lambda) = \sum_{i,j=1}^N (y_i - \lambda)(V^{-1})_{ij}(y_j - \lambda),$$

$$\rightarrow \hat{\lambda} = \sum_{i=1}^N w_i y_i, \quad w_i = \frac{\sum_{j=1}^N (V^{-1})_{ij}}{\sum_{k,l=1}^N (V^{-1})_{kl}}$$

$$V[\hat{\lambda}] = \sum_{i,j=1}^N w_i V_{ij} w_j$$

LS $\hat{\lambda}$ has zero bias, minimum variance (Gauss–Markov theorem).

Example: averaging two correlated measurements

Suppose we have y_1 , y_2 , and $V = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$

$$\rightarrow \hat{\lambda} = wy_1 + (1 - w)y_2, \quad w = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$$

$$V[\hat{\lambda}] = \frac{(1 - \rho^2)\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} = \sigma^2$$

The increase in inverse variance due to 2nd measurement is

$$\frac{1}{\sigma^2} - \frac{1}{\sigma_1^2} = \frac{1}{1 - \rho^2} \left(\frac{\rho}{\sigma_1} - \frac{1}{\sigma_2} \right)^2 > 0$$

\rightarrow 2nd measurement can only help.

Negative weights in LS average

If $\rho > \sigma_1/\sigma_2$, $\rightarrow w < 0$,

\rightarrow weighted average is not between y_1 and y_2 (!?)

Cannot happen if correlation due to common data, but possible for shared random effect; very unreliable if e.g. ρ , σ_1 , σ_2 incorrect.

See example in SDA Section 7.6.1 with two measurements at same temperature using two rulers, different thermal expansion coefficients: average is outside the two measurements; used to improve estimate of temperature.

G. Cowan, *Statistical Data Analysis*, Oxford University Press, 1998.