Statistical Methods for HEP
Lecture 1: Introduction

Taller de Altas Energías 2010

Universitat de Barcelona

1-11 September 2010

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

Outline

→ Lecture 1: Introduction and basic formalism
    Probability, statistical tests, parameter estimation.

Lecture 2: Multivariate methods
    General considerations
    Example of a classifier: Boosted Decision Trees

Lecture 3: Discovery and Exclusion limits
    Quantifying significance and sensitivity
    Systematic uncertainties (nuisance parameters)

Quick review of probablility

Frequentist (A = outcome of repeatable observation):

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

Subjective (A = hypothesis): $P(A)$ = degree of belief that A is true

Conditional probability: $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$

Bayes' theorem: $P(A|B) = \dfrac{P(B|A)P(A)}{P(B)} = \dfrac{P(B|A)P(A)}{\Sigma_i P(B|A_i)P(A_i)}$

Frequentist Statistics − general philosophy

In frequentist statistics, probabilities are associated only with the data, i.e., outcomes of repeatable observations.

Probability = limiting frequency

Probabilities such as

P (Higgs boson exists),
P (0.117 < α_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.

The preferred theories (models, hypotheses, ...) are those for which our observations would be considered 'usual'.

Bayesian Statistics − general philosophy

In Bayesian statistics, interpretation of probability extended to degree of belief (subjective probability). Use this 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

Bayesian methods can provide more natural treatment of non-repeatable phenomena:
systematic uncertainties, probability that Higgs boson exists,...

No golden rule for priors ("if-then" character of Bayes' thm.)

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 $L(x|H)$.

Definition of a test

Consider e.g. a simple 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.,

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

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

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

Critical region also called "rejection" region; complement is acceptance region.

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:
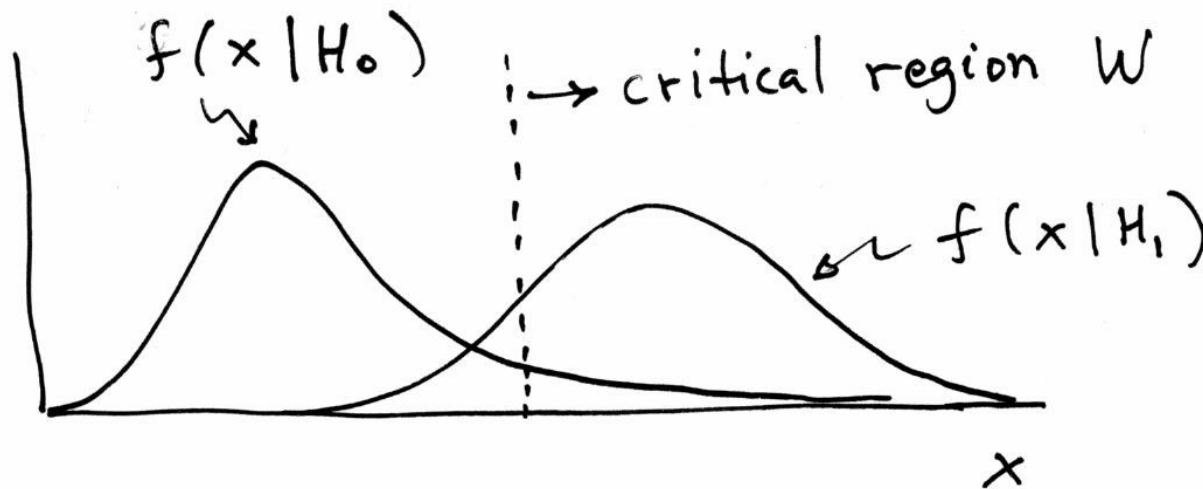
Rejecting a hypothesis

Note that rejecting H_0 is not necessarily equivalent to the statement that we believe it is false and H_1 true. In frequentist statistics only associate probability with outcomes of repeatable observations (the data).

In Bayesian statistics, probability of the hypothesis (degree of belief) would be found using Bayes' theorem:

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

which depends on the prior probability $\pi(H)$.

What makes a frequentist test useful is that we can compute the probability to accept/reject a hypothesis assuming that it is true, or assuming some alternative is true.

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:

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

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).
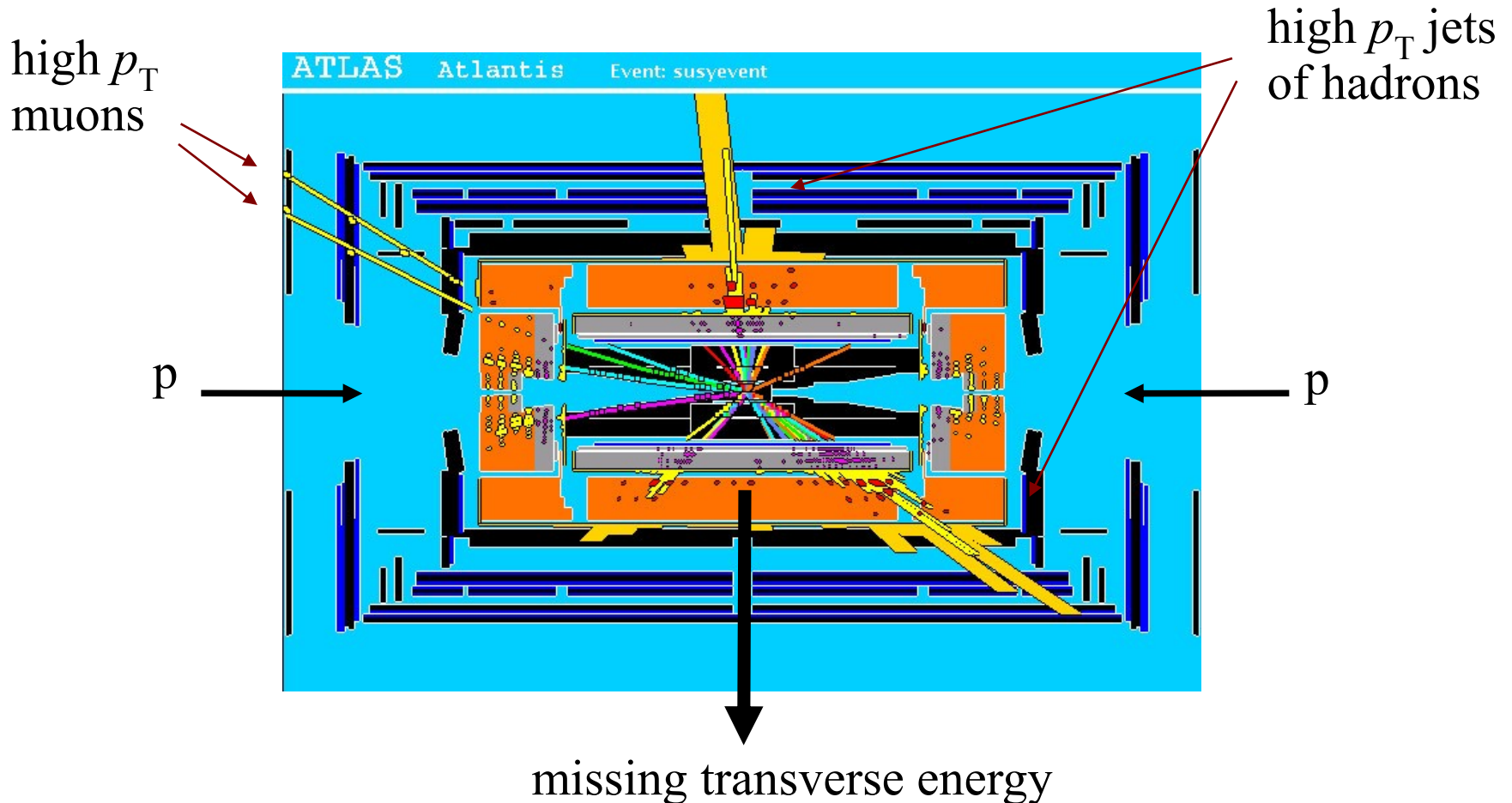Use the selected sample for further study.

Search for New Physics: the null hypothesis H_0 means Standard Model events, and the alternative H_1 means "events of a type whose existence is not yet established" (to establish or exclude the signal model is the goal of the analysis).

Many subtle issues here, mainly related to the heavy burden 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.
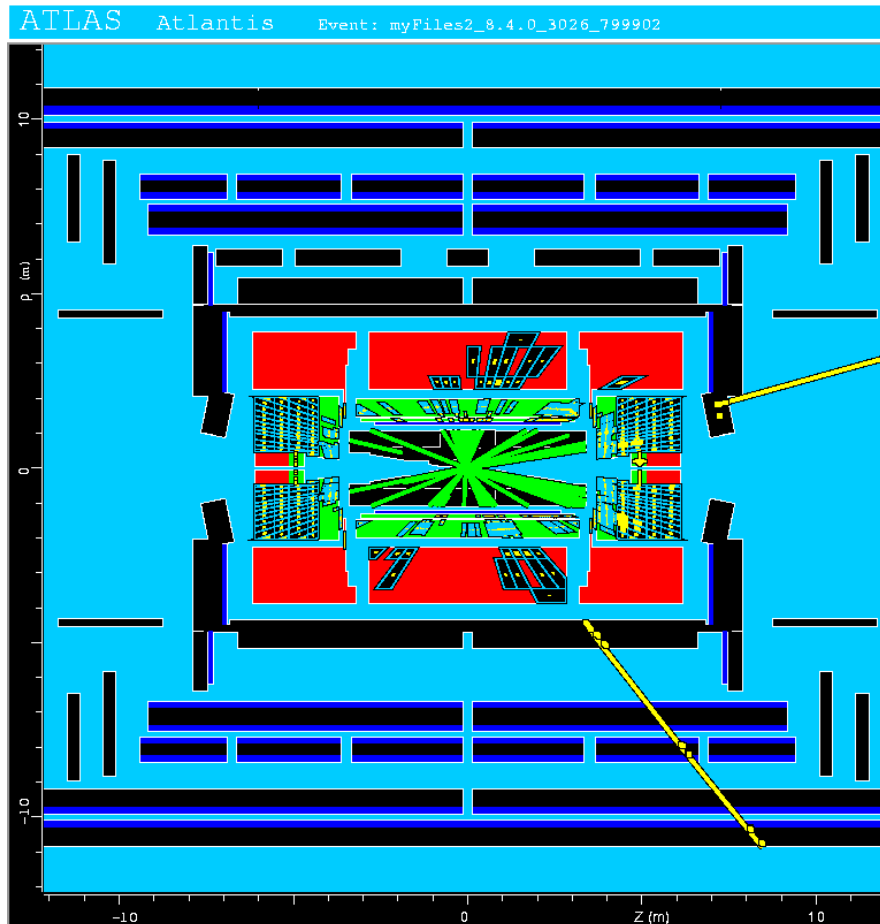
Suppose we want to discover this…

SUSY event (ATLAS simulation):

high p_T muons

high p_T jets of hadrons

ATLAS   Atlantis   Event: susyevent

p

p

missing transverse energy

But we know we'll have lots of this…

ttbar event (ATLAS simulation)

SM event also has high p_T jets and muons, and missing transverse energy.

→ can easily mimic a SUSY event and thus constitutes a background.

Example of a multivariate statistical test

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

x_1 = number of muons,

x_2 = mean p_t of jets,

x_3 = missing energy, ...

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

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

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.

Often call H_0 the background hypothesis (e.g. SM events); $H_1, H_2, \ldots$ are possible signal hypotheses.
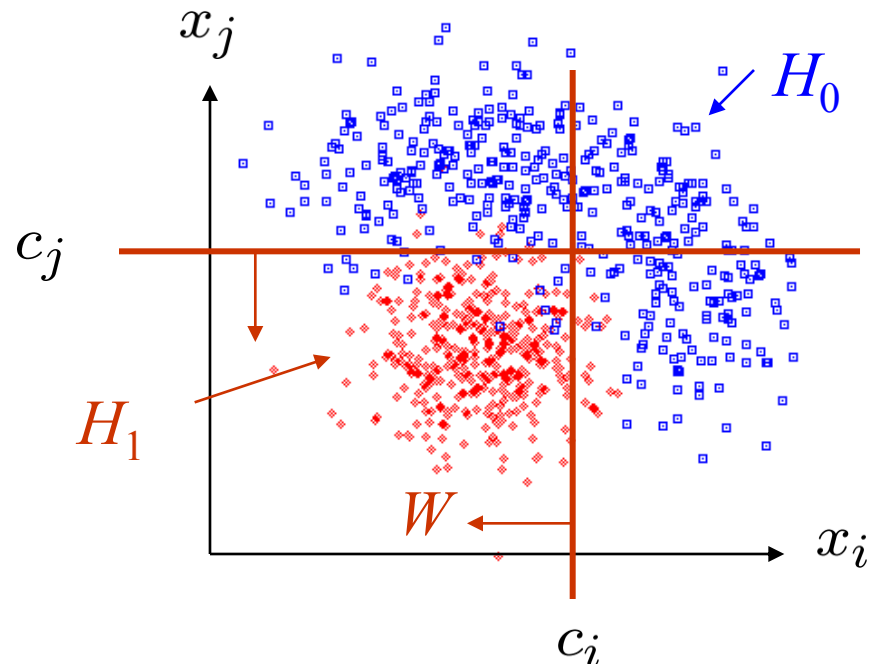
Defining a multivariate critical region

Each event is a point in **x**-space; critical region is now defined by a 'decision boundary' in this space.

What is best way to determine the decision boundary?

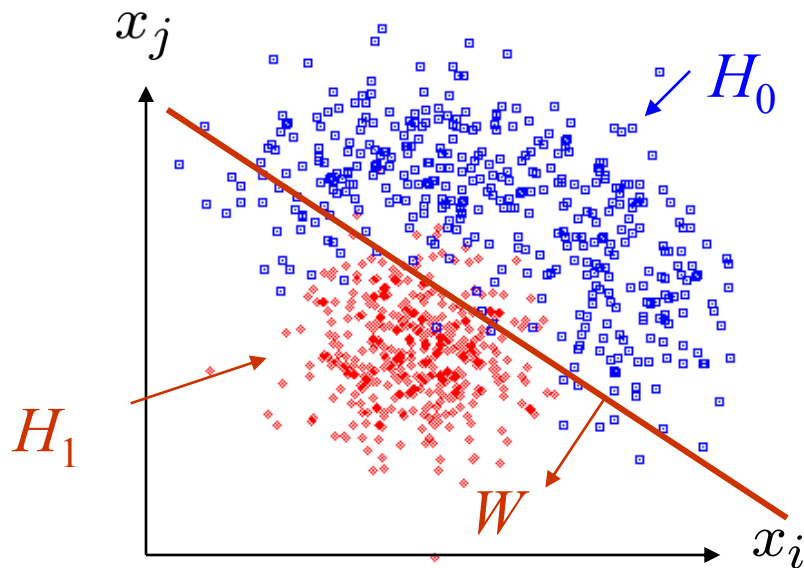Perhaps with 'cuts':

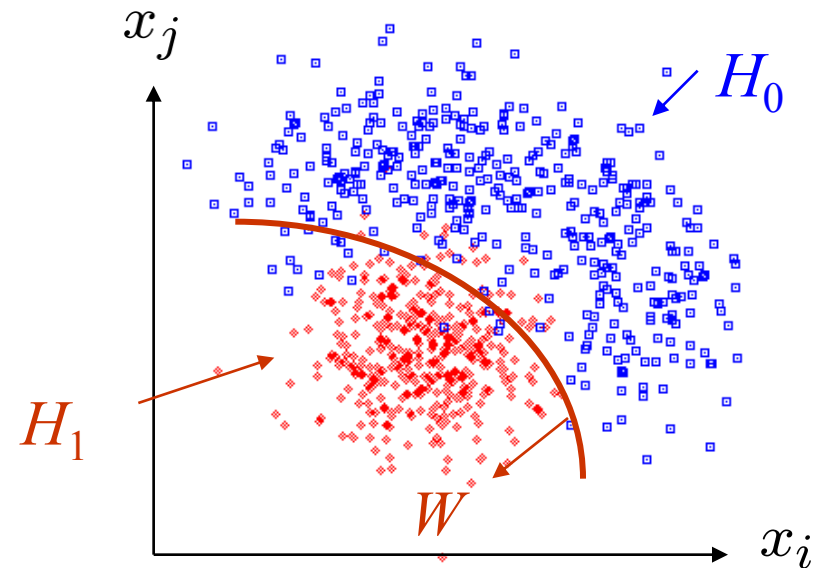$$x_i \;\; < c_i$$

$$x_j \;\; < c_j$$

Other multivariate decision boundaries

Or maybe use some other sort of decision boundary:

linear

or nonlinear

Multivariate methods

Many new (and some old) methods for finding decision boundary:

       Fisher discriminant

       Neural networks

       Kernel density methods

       Support Vector Machines

       Decision trees

            Boosting

            Bagging

New software for HEP, e.g.,

TMVA , Höcker, Stelzer, Tegenfeldt, Voss, Voss, physics/0703039

For more see e.g. references at end of this lecture.

For the rest of these lectures, I will focus on other aspects of tests, e.g., discovery significance and exclusion limits.

Test statistics

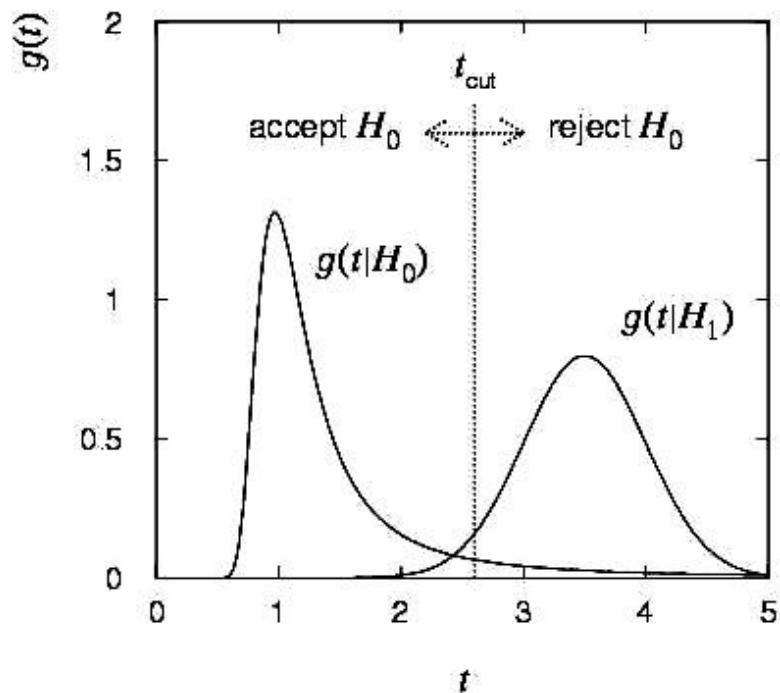The decision boundary can be defined by an equation of the form

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

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

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

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

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

Significance level and power

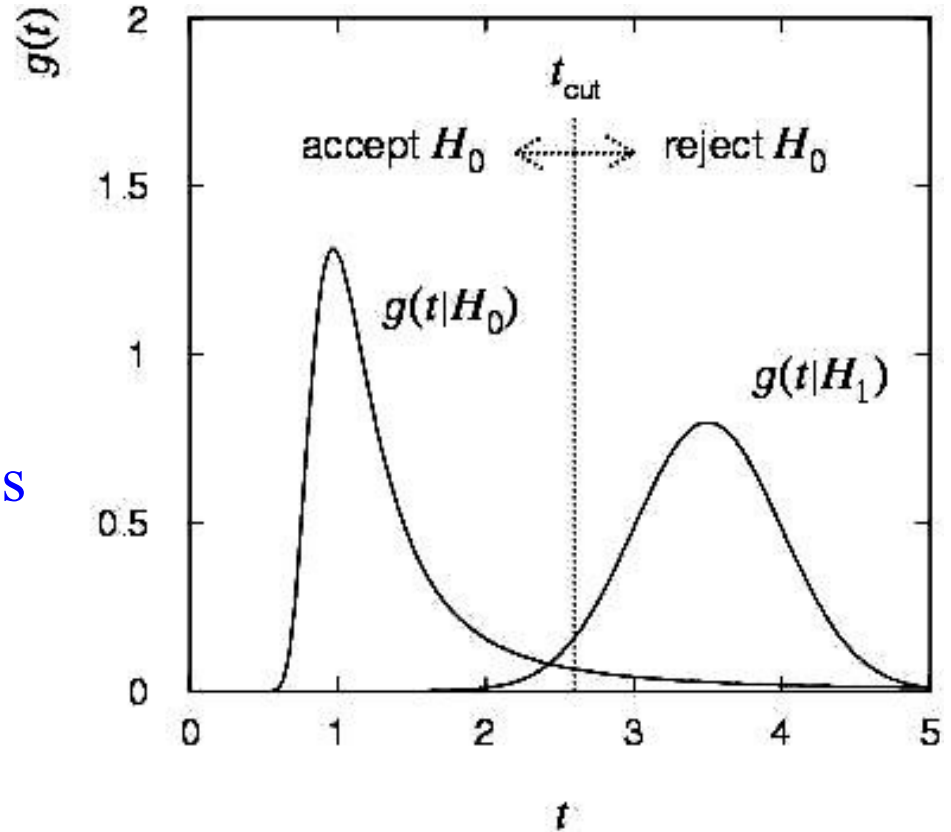Probability to reject H_0 if it is true (type-I error):

$$\alpha = \int_{t_{\text{cut}}}^{\infty} g(t|H_0)\,dt$$

(significance level)

Probability to accept H_0 if H_1 is true (type-II error):

$$\beta = \int_{-\infty}^{t_{\text{cut}}} g(t|H_1)\,dt$$
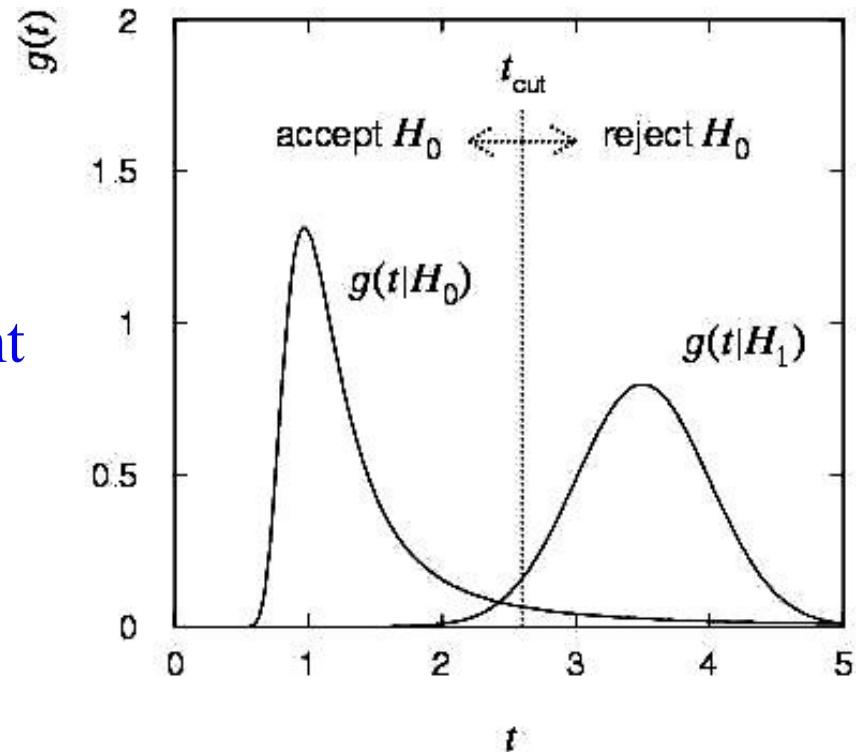
$(1 - \beta = \text{power})$

Signal/background efficiency

Probability to reject background hypothesis for background event (background efficiency):

$$\varepsilon_{\mathrm{b}} = \int_{t_{\mathrm{cut}}}^{\infty} g(t|\mathrm{b}) \, dt = \alpha$$

Probability to accept a signal event as signal (signal efficiency):

$$\varepsilon_{\mathrm{s}} = \int_{t_{\mathrm{cut}}}^{\infty} g(t|\mathrm{s}) \, dt = 1 - \beta$$

Purity of event selection

Suppose only one background type b; overall fractions of signal and background events are π_s and π_b (prior probabilities).

Suppose we select signal events with $t > t_{\text{cut}}$. What is the 'purity' of our selected sample?

Here purity means the probability to be signal given that the event was accepted. Using Bayes' theorem we find:

$$
\begin{aligned}
P(s|t > t_{\text{cut}}) &= \frac{P(t > t_{\text{cut}}|s)\pi_s}{P(t > t_{\text{cut}}|s)\pi_s + P(t > t_{\text{cut}}|b)\pi_b} \\
&= \frac{\varepsilon_s \pi_s}{\varepsilon_s \pi_s + \varepsilon_b \pi_b}
\end{aligned}
$$

So the purity depends on the prior probabilities as well as on the signal and background efficiencies.

Constructing a test statistic

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 H_0, (background) versus H_1, (signal) (highest ε_s for a given ε_b) choose the critical (rejection) region such that

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

where c is a constant which determines the power.

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

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

Proof of Neyman-Pearson lemma

We want to determine the critical region W that maximizes the power

$$1 - \beta = \int_W P(x|H_1)\, dx$$

subject to the constraint

$$\alpha = \int_W P(x|H_0)\, dx$$

First, include in W all points where $P(x|H_0) = 0$, as they contribute nothing to the size, but potentially increase the power.

Proof of Neyman-Pearson lemma (2)

For $P(x|H_0) \neq 0$ we can write the power as

$$1 - \beta = \int_W \frac{P(x|H_1)}{P(x|H_0)} P(x|H_0)\, dx$$

The ratio of $1 - \beta$ to α is therefore

$$\frac{1-\beta}{\alpha} = \frac{\int_W \frac{P(x|H_1)}{P(x|H_0)} P(x|H_0)\, dx}{\int_W P(x|H_0)\, dx}$$

which is the average of the likelihood ratio $P(x|H_1) / P(x|H_0)$ over the critical region W, assuming H_0.

$(1 - \beta) / \alpha$ is thus maximized if W contains the part of the sample space with the largest values of the likelihood ratio.

Testing significance / goodness-of-fit

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

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

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 $\vec{x}_{\mathsf{obs}}$.
(Not unique!)

x_j

$\vec{x}_{\mathsf{obs}}$

$\vec{x}$ more compatible with H

$\vec{x}$ less compatible with H

x_i
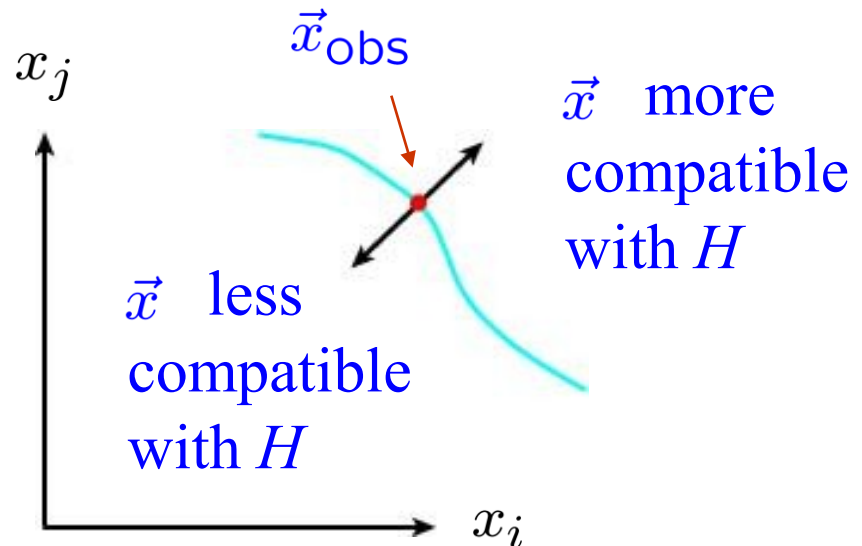
p-values

Express level of agreement between data and *H* with *p*-value:

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

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

In frequentist statistics we don't talk about $P(H)$ (unless *H* represents a repeatable observation). In Bayesian statistics we do; use Bayes' theorem to obtain
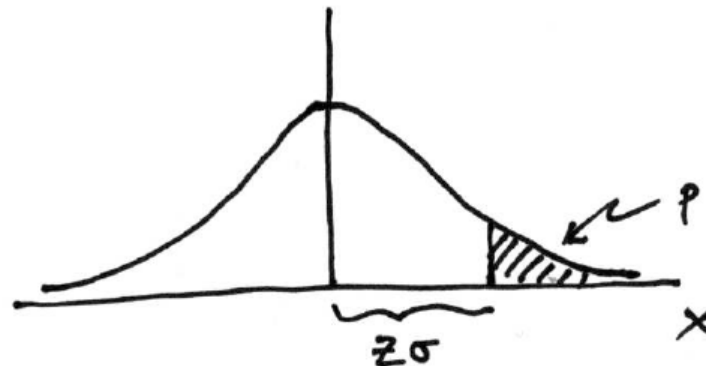
$$P(H|\vec{x}) = \frac{P(\vec{x}|H)\pi(H)}{\int P(\vec{x}|H)\pi(H)\,dH}$$

where $\pi(H)$ is the prior probability for *H*.

For now stick with the frequentist approach; result is *p*-value, regrettably easy to misinterpret as $P(H)$.

Significance from *p*-value

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

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

`1 - TMath::Freq`

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

`TMath::NormQuantile`

The significance of an observed signal

Suppose we observe n events; these can consist of:

n_b events from known processes (background)
n_s events from a new process (signal)

If n_s, n_b are Poisson r.v.s with means s, b, then $n = n_s + n_b$ is also Poisson, mean $= s + b$:

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

Suppose $b = 0.5$, and we observe $n_{obs} = 5$. Should we claim evidence for a new discovery?

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

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

When to publish

HEP folklore is to claim discovery when $p = 2.9 \times 10^{-7}$, corresponding to a significance $Z = 5$.

This is very subjective and really should depend on the prior probability of the phenomenon in question, e.g.,

phenomenon	reasonable p-value for discovery
$D^0 D^0$ mixing	~0.05
Higgs	$\sim 10^{-7}$ (?)
Life on Mars	$\sim 10^{-10}$
Astrology	$\sim 10^{-20}$

One should also consider the degree to which the data are compatible with the new phenomenon, not only the level of disagreement with the null hypothesis; p-value is only first step!

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 $t(\boldsymbol{x})$ as

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

The pdf of p_H under assumption of H is

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

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

$g(p_H|H')$

$g(p_H|H)$

p_H

0   1
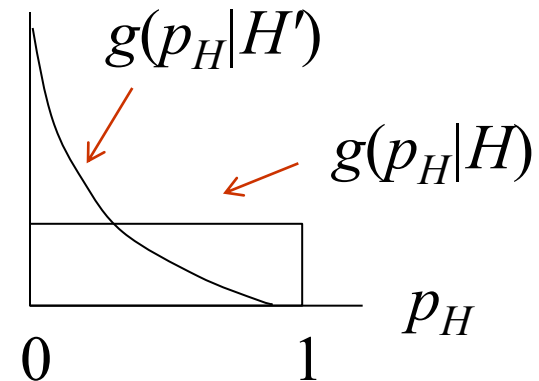
Using a p-value to define test of H_0

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 started by defining critical region in the original data space ($\boldsymbol{x}$), then reformulated this in terms of a scalar test statistic $t(\boldsymbol{x})$.

We can take this one step further and 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.

Quick review of parameter estimation

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

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

random variable   parameter

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

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

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

$\leftarrow$ estimator written with a hat

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

Properties of estimators

If we were to repeat the entire measurement, the estimates from each would follow a pdf:
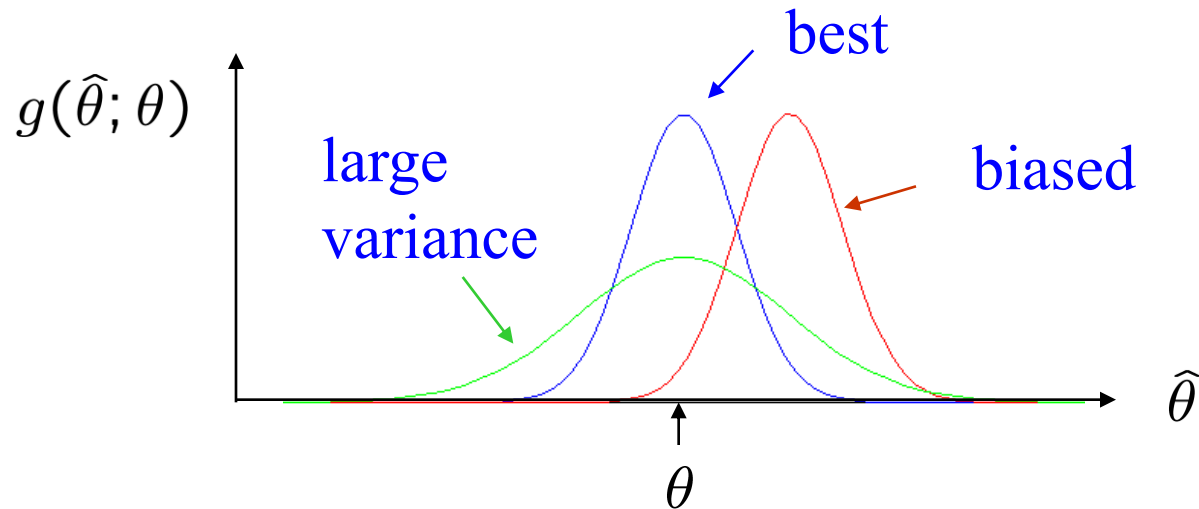
We want small (or zero) bias (systematic error): $\quad b = E[\hat{\theta}] - \theta$

→ average of repeated measurements should tend to true value.

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

→ small bias & variance are in general conflicting criteria

The likelihood function

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

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

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

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

(*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, ..., x_n, where x follows $f(x; \theta)$. The joint pdf for the whole data sample is:
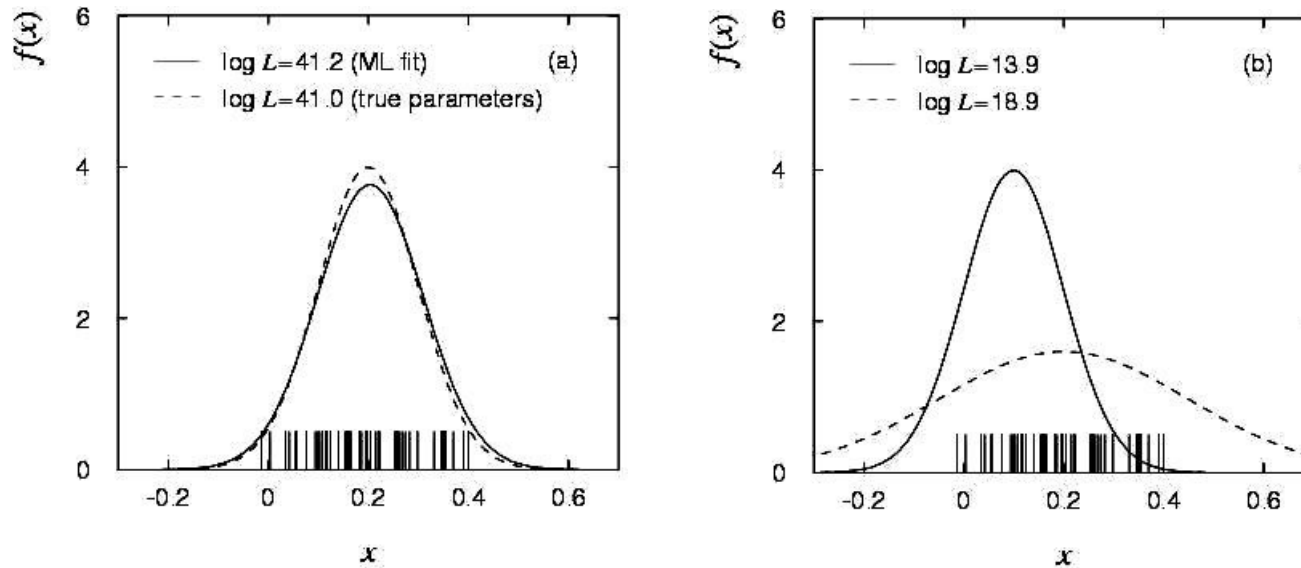
$$f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

In this case the likelihood function is

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

Maximum likelihood estimators

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).

Wrapping up lecture 1

General framework of a statistical test:

Divide data spaced into two regions; depending on where data are then observed, accept or reject hypothesis.

Properties:

significance level (rate of Type-I error)

power (one minus rate of Type-II error)

Significance tests (also for goodness-of-fit):

p-value = probability to see level of incompatibility between data and hypothesis equal to or greater than level found with the actual data.

Parameter estimation

Maximize likelihood function $\rightarrow$ ML estimator

Extra slides

p-value example: testing whether a coin is 'fair'

Probability to observe *n* heads in *N* coin tosses is binomial:

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

Hypothesis *H*: the coin is fair (*p* = 0.5).

Suppose we toss the coin *N* = 20 times and get *n* = 17 heads.

Region of data space with equal or lesser compatibility with *H* relative to *n* = 17 is: *n* = 17, 18, 19, 20, 0, 1, 2, 3. Adding up the probabilities for these values gives:

$$P(n = 0, 1, 2, 3, 17, 18, 19, \text{ or } 20) = 0.0026 \,.$$

i.e. *p* = 0.0026 is the probability of obtaining such a bizarre result (or more so) 'by chance', under the assumption of *H*.

ML example: parameter of exponential pdf

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

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

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

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)

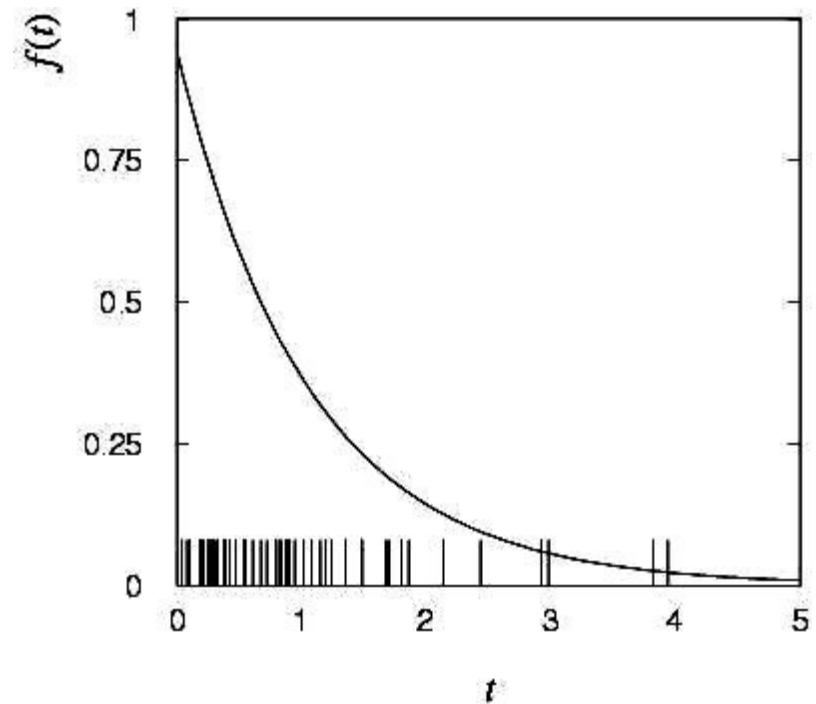Find its maximum by setting $\dfrac{\partial \ln L(\tau)}{\partial \tau} = 0$,

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

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

We find the ML estimate:

$$\widehat{\tau} = 1.062$$

Variance of estimators from information inequality

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

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

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:

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

Information inequality for *n* parameters

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

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

$$I_{ij} = E\left[-\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}\right] = -n \int f(x; \vec{\theta}) \frac{\partial^2 \ln f(x; \vec{\theta})}{\partial \theta_i \partial \theta_j} \, dx$$

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.

Extended ML

Sometimes regard n not as fixed, but as a Poisson r.v., mean ν.

Result of experiment defined as: n, x_1, ..., x_n.

The (extended) likelihood function is:

$$L(\nu, \vec{\theta}) = \frac{\nu^n}{n!} e^{-\nu} \prod_{i=1}^{n} f(x_i; \vec{\theta})$$

Suppose theory gives $\nu = \nu(\boldsymbol{\theta})$, then the log-likelihood is

$$\ln L(\vec{\theta}) = -\nu(\vec{\theta}) + \sum_{i=1}^{n} \ln(\nu(\vec{\theta}) f(x_i; \vec{\theta})) + C$$

where C represents terms not depending on $\boldsymbol{\theta}$.

Extended ML (2)

Example: expected number of events $\nu(\vec{\theta}) = \sigma(\vec{\theta}) \int L\, dt$
where the total cross section $\sigma(\theta)$ is predicted as a function of the parameters of a theory, as is the distribution of a variable x.

Extended ML uses more info → smaller errors for $\widehat{\vec{\theta}}$

Important e.g. for anomalous couplings in $e^+e^- \rightarrow W^+W^-$

If ν does not depend on θ but remains a free parameter, extended ML gives:

$$\hat{\nu} = n$$
$$\hat{\theta} = \text{ same as ML}$$

Extended ML example

Consider two types of events (e.g., signal and background) each of which predict a given pdf for the variable x: $f_s(x)$ and $f_b(x)$.

We observe a mixture of the two event types, signal fraction $= \theta$, expected total number $= \nu$, observed total number $= n$.

Let $\mu_s = \theta\nu$, $\mu_b = (1 - \theta)\nu$, goal is to estimate μ_s, μ_b.

$$f(x; \mu_s, \mu_b) = \frac{\mu_s}{\mu_s + \mu_b} f_s(x) + \frac{\mu_b}{\mu_s + \mu_b} f_b(x)$$

$$P(n; \mu_s, \mu_b) = \frac{(\mu_s + \mu_b)^n}{n!} e^{-(\mu_s + \mu_b)}$$

$$\rightarrow \ \ln L(\mu_s, \mu_b) = -(\mu_s + \mu_b) + \sum_{i=1}^{n} \ln \left[(\mu_s + \mu_b) f(x_i; \mu_s, \mu_b) \right]$$

Extended ML example (2)

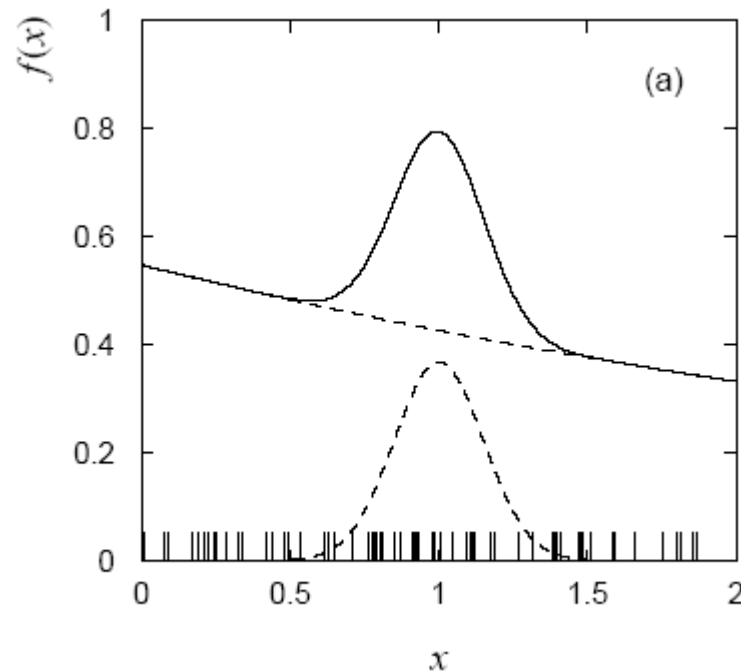Monte Carlo example with combination of exponential and Gaussian:

$$\mu_\text{s} = 6$$

$$\mu_\text{b} = 60$$

Maximize log-likelihood in terms of μ_s and μ_b:

$$\hat{\mu}_\text{s} = 8.7 \pm 5.5$$

$$\hat{\mu}_\text{b} = 54.3 \pm 8.8$$

Here errors reflect total Poisson fluctuation as well as that in proportion of signal/background.
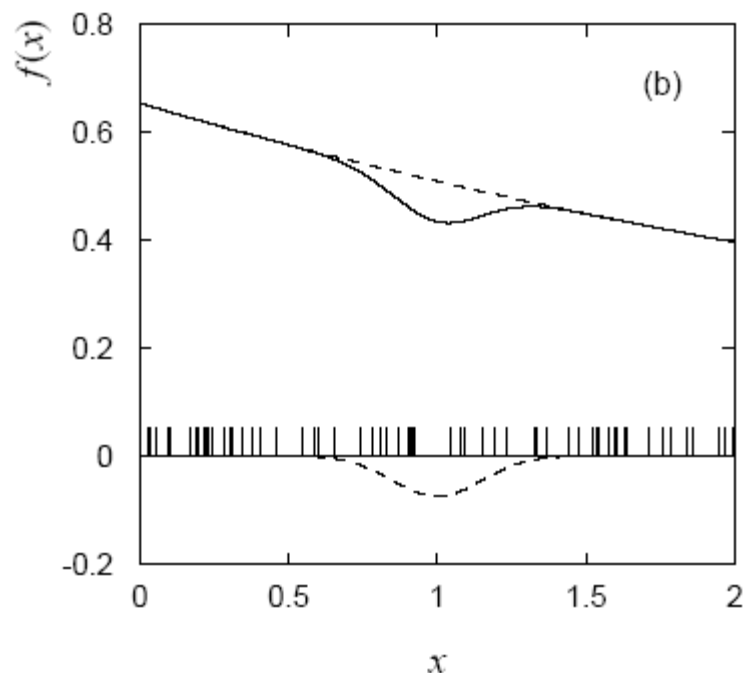
Extended ML example: an unphysical estimate

A downwards fluctuation of data in the peak region can lead to even fewer events than what would be obtained from background alone.

Estimate for μ_s here pushed negative (unphysical).

We can let this happen as long as the (total) pdf stays positive everywhere.

Unphysical estimators (2)

Here the unphysical estimator is unbiased and should nevertheless be reported, since average of a large number of unbiased estimates converges to the true value (cf. PDG).

Repeat entire MC experiment many times, allow unphysical estimates: