Statistical Methods for High Energy Physics

Statistics

Saeid Paktinat

School of Particles and accelerators
IPM, Tehran

Third National Workshop on
Detectors and Calculation Methods in Particle Physics

Azar 26-28, 1387

Modified slides of Glen Cowan `http://www.pp.rhul.ac.uk/~cowan`

Outline

Lecture 1 (Probability)

- Probability
- Random variables, probability densities, etc.
- Brief catalogue of probability densities
- The Monte Carlo method

Lecture 2 (Statistics)

- Statistical tests
- Fisher discriminants, etc.
- Significance and goodness-of-fit tests
- Parameter estimation
- Maximum likelihood and least squares
- Interval estimation (setting limits)

Statistical tests (in a particle physics context)

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

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

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.

E.g. call H_0 the null (background) hypothesis (the event type we know already exists); H_1, H_2, … are alternative (signal) hypotheses.

Selecting events

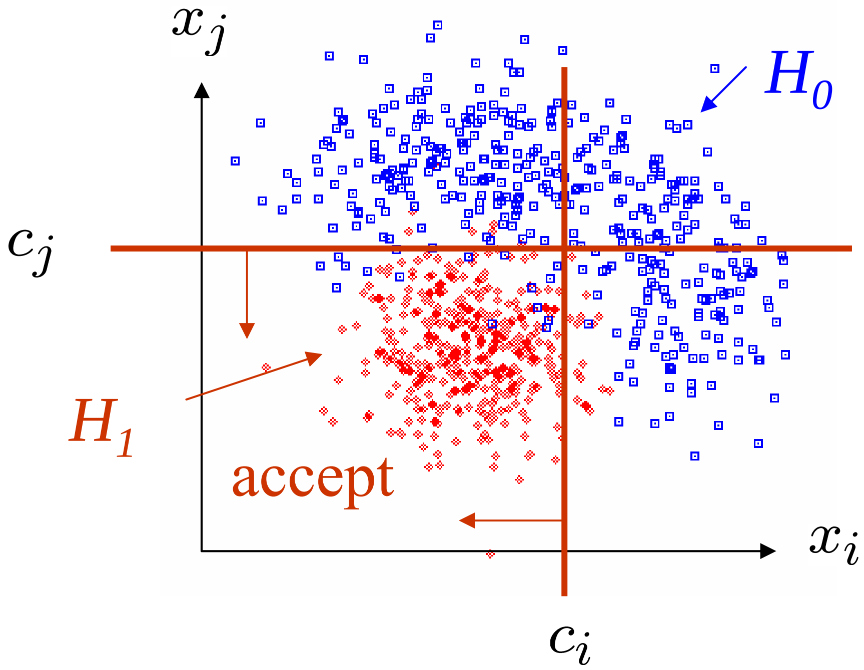
Suppose we have a data sample with two kinds of events, corresponding to hypotheses H_0 and H_1 and we want to select those of type H_1.

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

Perhaps select events with 'cuts':

$$x_i < c_i$$

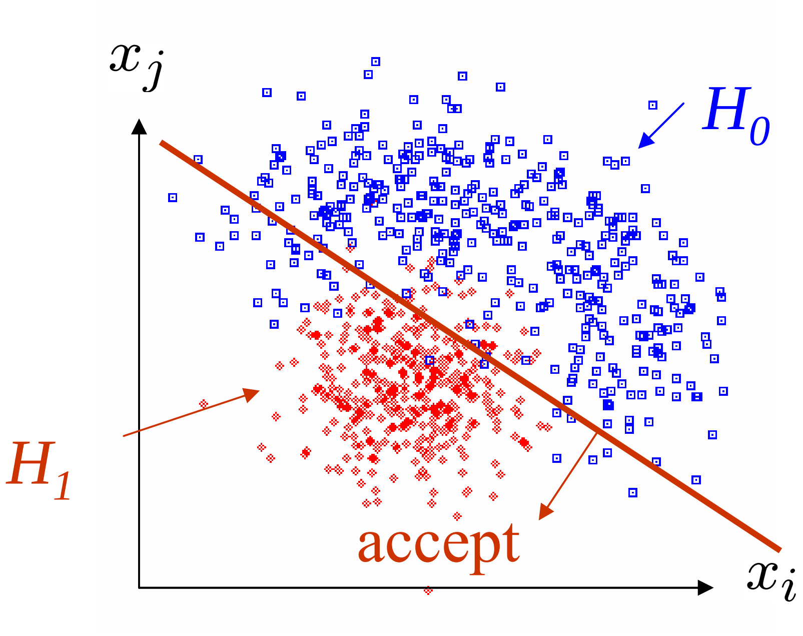
$$x_j < c_j$$



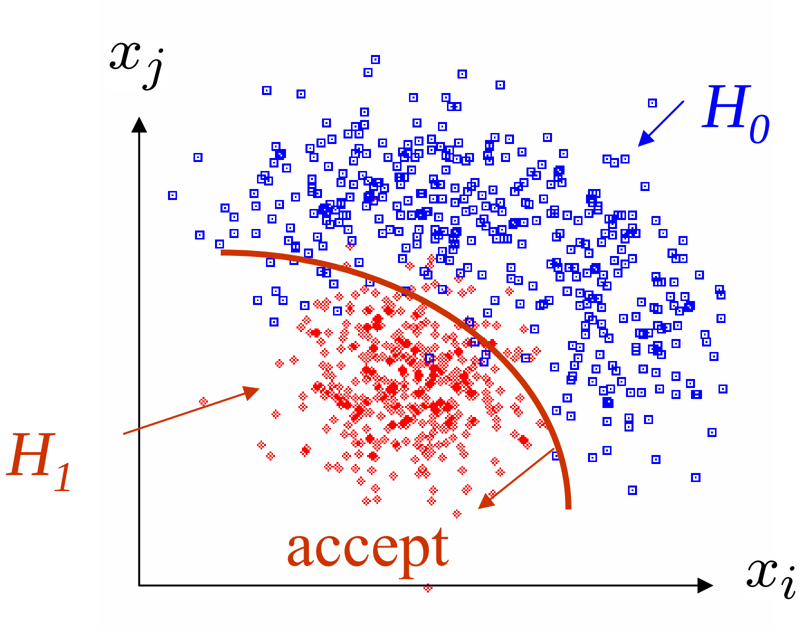
Other ways to select events

Or maybe use some other sort of decision boundary:

linear



or nonlinear



How can we do this in an ‘optimal’ way?

Test statistics

Construct a ‘test statistic’ of lower dimension (e.g. scalar)

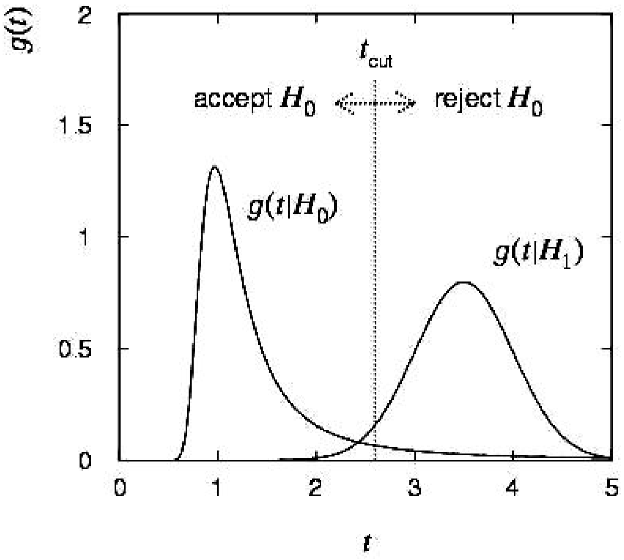
$$t(x_1, \ldots, x_n)$$

Try to compactify data without losing ability to discriminate between hypotheses.

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

Decision boundary is now a single ‘cut’ on *t*.

This effectively divides the sample space into two regions, where we accept or reject H_0.

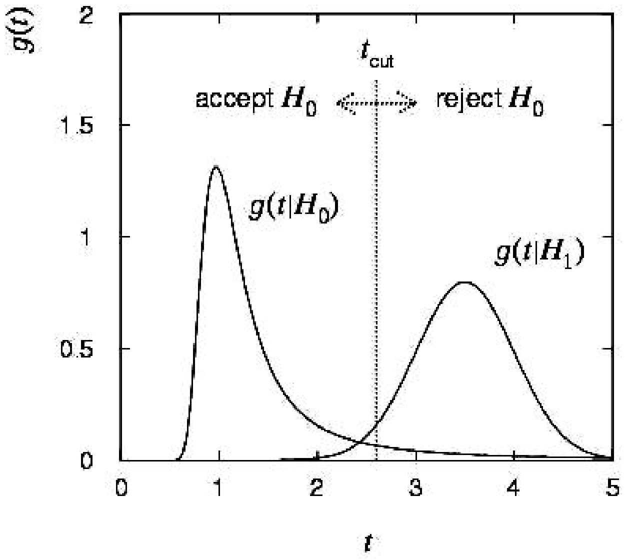


Significance level and power of a test

Probability to reject H_0 if it is true (error of the 1st kind):

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

(significance level)



Probability to accept H_0 if H_1 is true (error of the 2nd kind):

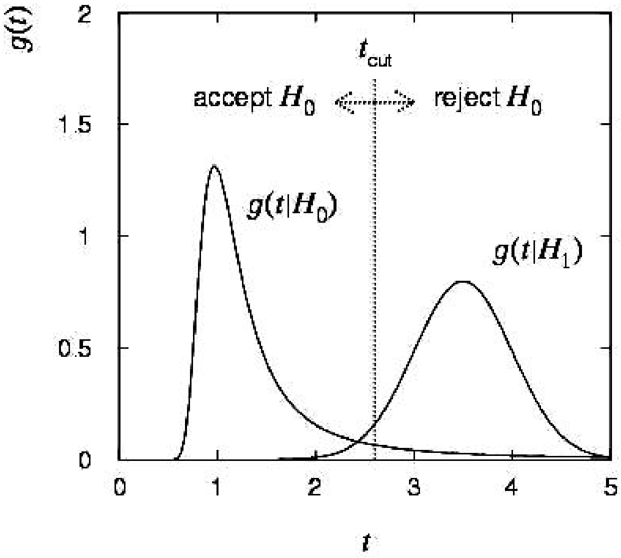
$$\beta = \int_{-\infty}^{t_{\text{cut}}} g(t|H_1)\,dt$$

$(1 - \beta = \text{power})$

Efficiency of event selection

Probability to accept an event which is signal (signal efficiency):

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



Probability to accept an event which is background (background efficiency):

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

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 events with $t > t_{cut}$. 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:

$$
\begin{aligned}
P(\mathrm{s}|t > t_{\mathrm{cut}}) &= \frac{P(t > t_{\mathrm{cut}}|\mathrm{s})\pi_{\mathrm{s}}}{P(t > t_{\mathrm{cut}}|\mathrm{s})\pi_{\mathrm{s}} + P(t > t_{\mathrm{cut}}|\mathrm{b})\pi_{\mathrm{b}}} \\
&= \frac{\varepsilon_{\mathrm{s}}\pi_{\mathrm{s}}}{\varepsilon_{\mathrm{s}}\pi_{\mathrm{s}} + \varepsilon_{\mathrm{b}}\pi_{\mathrm{b}}}
\end{aligned}
$$

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

Constructing a test statistic

How can we select events in an ‘optimal way’?

Neyman-Pearson lemma (proof in Brandt Ch. 8) states:

To get the lowest ε_{b} for a given ε_{s} (highest power for a given significance level), choose acceptance region such that

$$\frac{f(\vec{x}|\mathrm{s})}{f(\vec{x}|\mathrm{b})} > c$$

where c is a constant which determines ε_{s}.

Equivalently, optimal scalar test statistic is $\boxed{t(\vec{x}) = \frac{f(\vec{x}|\mathrm{s})}{f(\vec{x}|\mathrm{b})}}$

Why Neyman-Pearson doesn’t always help

The problem is that we usually don’t have explicit formulae for the pdfs $f(\vec{x}|\mathsf{s}),\ f(\vec{x}|\mathsf{b})$.

Instead we may have Monte Carlo models for signal and background processes, so we can produce simulated data, and enter each event into an *n*-dimensional histogram.

Use e.g. *M* bins for each of the *n* dimensions, total of M^n cells.

But *n* is potentially large, → prohibitively large number of cells to populate with Monte Carlo data.

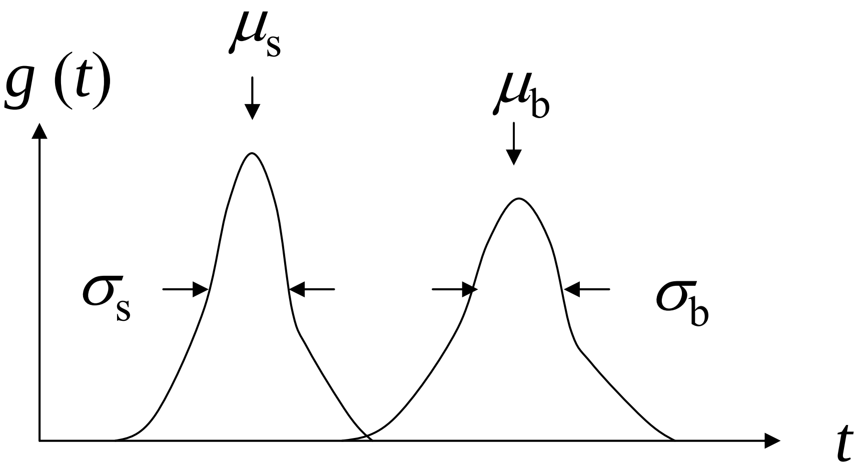
Compromise: make Ansatz for form of test statistic $t(\vec{x})$ with fewer parameters; determine them (e.g. using MC) to give best discrimination between signal and background.

Linear test statistic

Ansatz: $$t(\vec{x}) = \sum_{i=1}^{n} a_i x_i$$

Choose the parameters $a_1, ..., a_n$ so that the pdfs $g(t|\mathrm{s}),\ g(t|\mathrm{b})$ have maximum ‘separation’. We want:

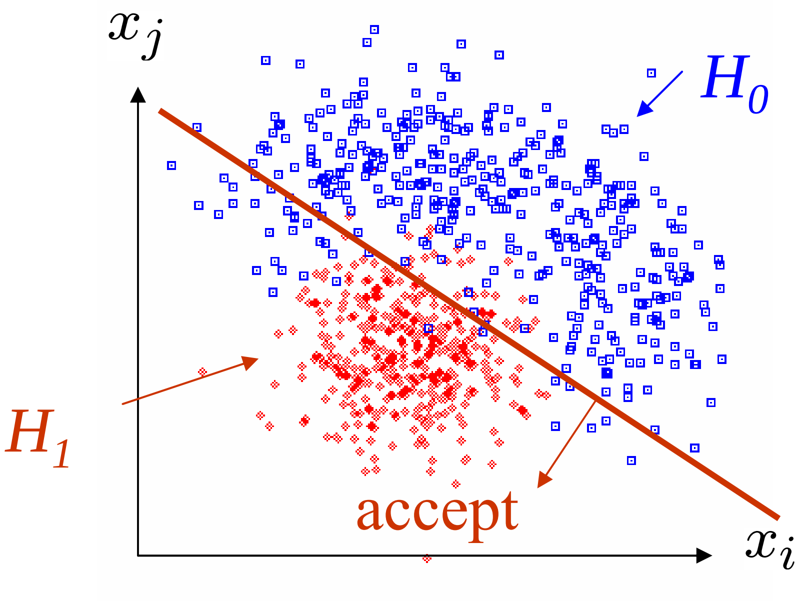
large distance between mean values, small widths



→ Fisher: maximize $$J(\vec{a}) = \frac{(\mu_\mathrm{s} - \mu_\mathrm{b})^2}{\sigma_\mathrm{s}^2 + \sigma_\mathrm{b}^2}$$

Fisher discriminant

Using this definition of separation gives a Fisher discriminant.



Corresponds to a linear decision boundary.

Equivalent to Neyman-Pearson if the signal and background pdfs are multivariate Gaussian with equal covariances; otherwise not optimal, but still often a simple, practical solution.

Nonlinear test statistics

The optimal decision boundary may not be a hyperplane,

→ nonlinear test statistic $t(\vec{x})$

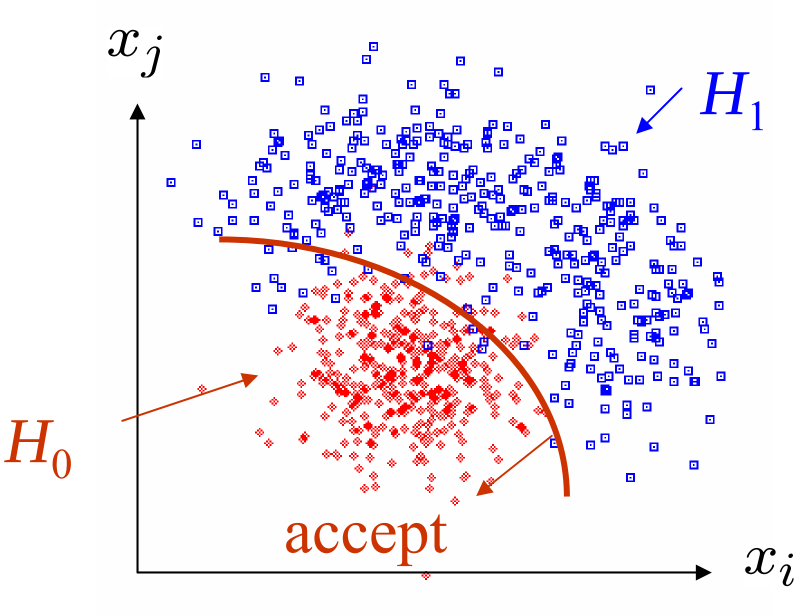
Multivariate statistical methods are a Big Industry:

Neural Networks,

Support Vector Machines,

Kernel density estimation,

Boosted decision trees, ...



New software for HEP, e.g.,

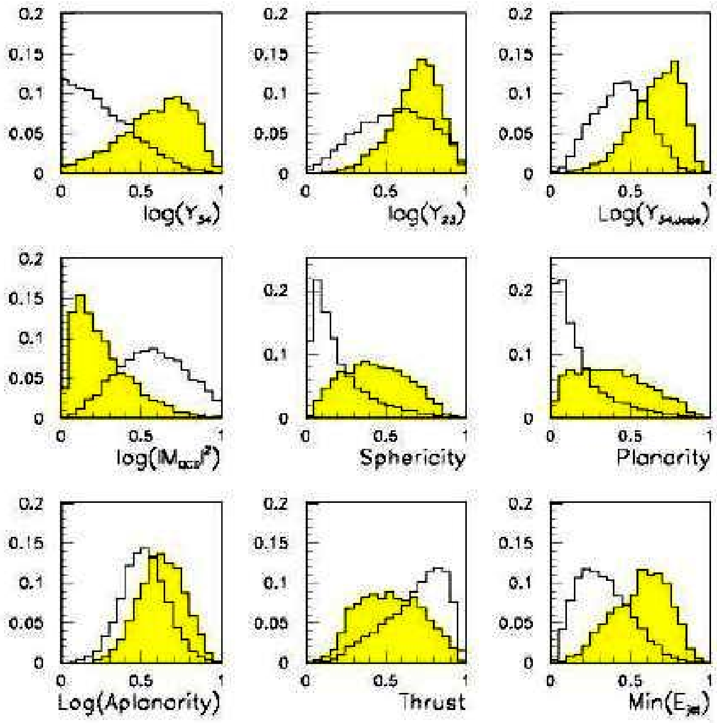
TMVA , Höcker, Stelzer, Tegenfeldt, Voss, Voss, physics/0703039

StatPatternRecognition, I. Narsky, physics/0507143

Neural network example from LEP II

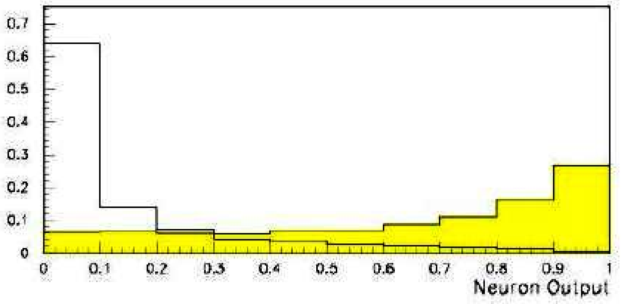
Signal: $e^+e^- \to W^+W^-$ (often 4 well separated hadron jets)

Background: $e^+e^- \to qqgg$ (4 less well separated hadron jets)



← input variables based on jet structure, event shape, ...
none by itself gives much separation.

Neural network output does better...



(Garrido, Juste and Martinez, ALEPH 96-144)

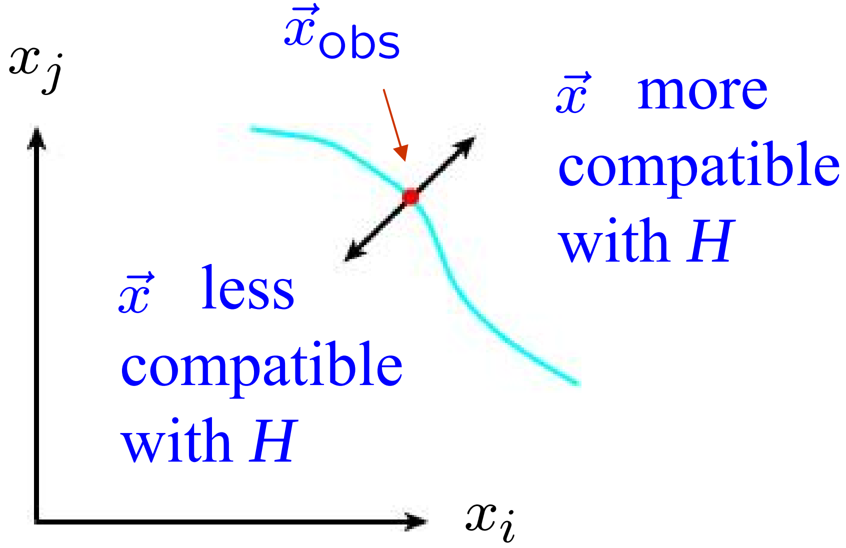
Testing significance/goodness-of-fit

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: $\vec{x}_{\mathrm{obs}}$

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



$\vec{x}$ more compatible with *H*

$\vec{x}$ less compatible with *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).

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*.

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 = $s + b$:

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

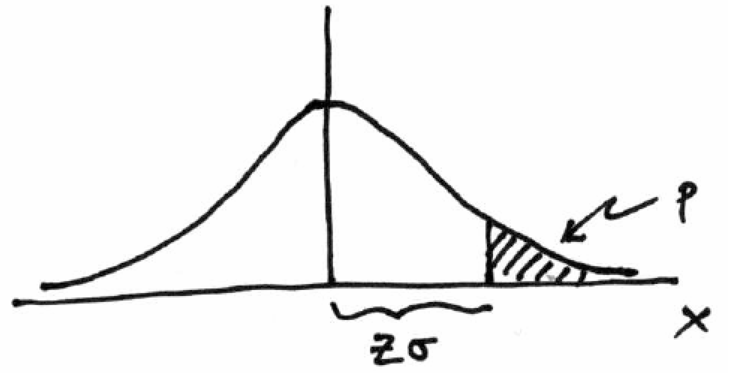
Suppose $b = 0.5$, and we observe $n_{obs} = 5$. Should we claim evidence for a new discovery?

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

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

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.



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

TMath::Prob

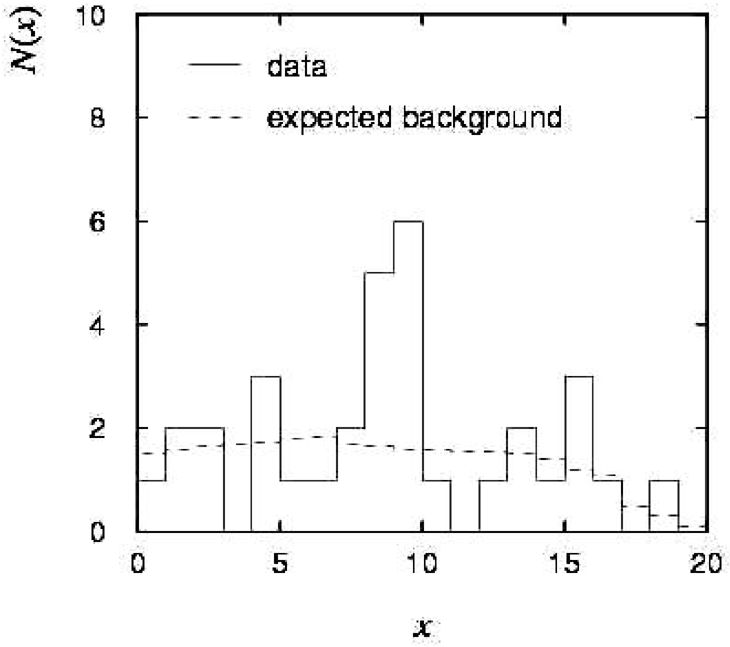
$$Z = \Phi^{-1}(1-p)$$

TMath::NormQuantile

E.g. $Z = 5$ (a ‘5 sigma effect’) means $p = 2.87 \times 10^{-7}$

The significance of a peak

Suppose we measure a value *x* for each event and find:



Each bin (observed) is a Poisson r.v., means are given by dashed lines.

In the two bins with the peak, 11 entries found with *b* = 3.2. The *p*-value for the *s* = 0 hypothesis is:

$$P(n \geq 11; b = 3.2, s = 0) = 5.0 \times 10^{-4}$$

The significance of a peak (2)

But... did we know where to look for the peak?

→ give $P(n \geq 11)$ in any 2 adjacent bins

Is the observed width consistent with the expected *x* resolution?

→ take *x* window several times the expected resolution

How many bins × distributions have we looked at?

→ look at a thousand of them, you’ll find a 10^{-3} effect

Did we adjust the cuts to ‘enhance’ the peak?

→ freeze cuts, repeat analysis with new data

How about the bins to the sides of the peak... (too low!)

Should we publish????

When to publish

HEP folklore: claim discovery when *p*-value of background only hypothesis is 2.87×10^{-7}, corresponding to 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^0D^0 mixing	~0.05
Higgs	$\sim 10^{-7}$ (?)
Life on Mars	$\sim 10^{-10}$
Astrology	$\sim 10^{-20}$

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: $\vec{x} = (x_1, \ldots, x_n)$

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

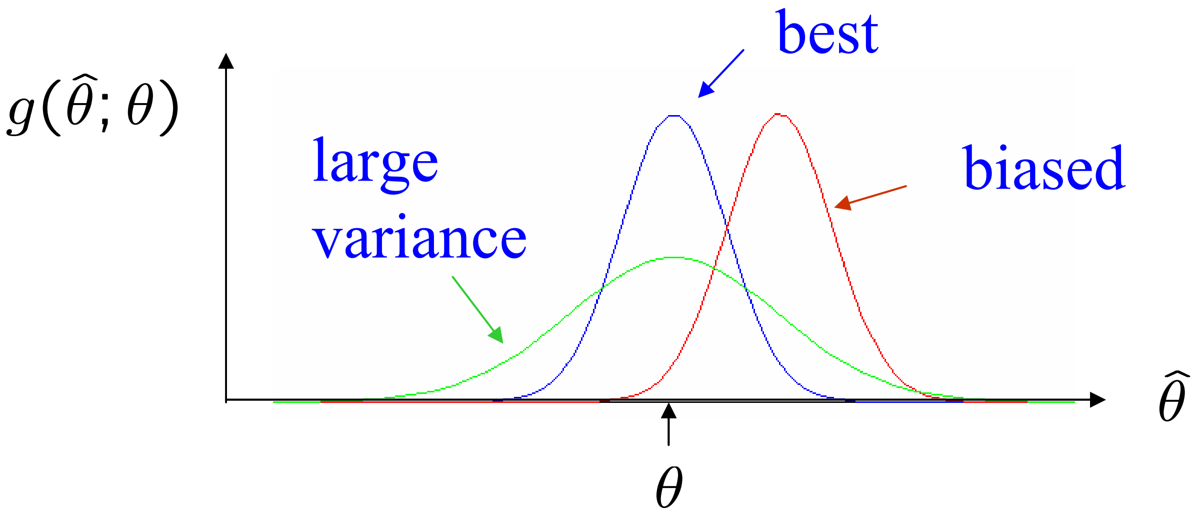
$$\boxed{\hat{\theta}(\vec{x})}$$

← estimator written with a hat

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

Properties of estimators

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



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

→ 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: $\mu = E[x]$

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

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

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

An estimator for the variance

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

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

We find:

$b = E[\widehat{\sigma^2}] - \sigma^2 = 0$ (factor of *n*–1 makes this so)

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

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

The likelihood function

Suppose the outcome of an experiment is: $x_1, \ldots, x_n$, which is modeled as a sample from a joint pdf with parameter(s) θ:

$$f(x_1, \ldots, x_n; \theta)$$

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

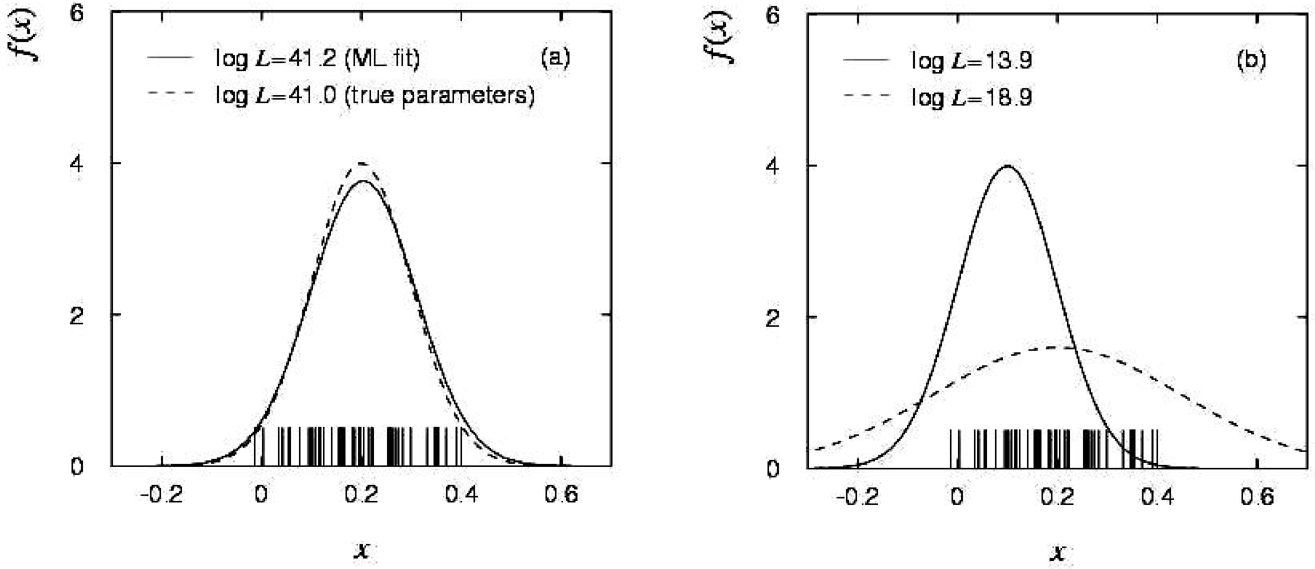
$$\boxed{L(\theta) = f(x_1, \ldots, x_n; \theta)} \qquad (x_i \text{ constant})$$

If the x_i are independent observations of $x \sim f(x;\theta)$, then,

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

Maximum likelihood estimators

If the hypothesized θ is close to the true value, then we expect a high probability to get data like what 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 data, $t_1, \ldots, t_n$

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

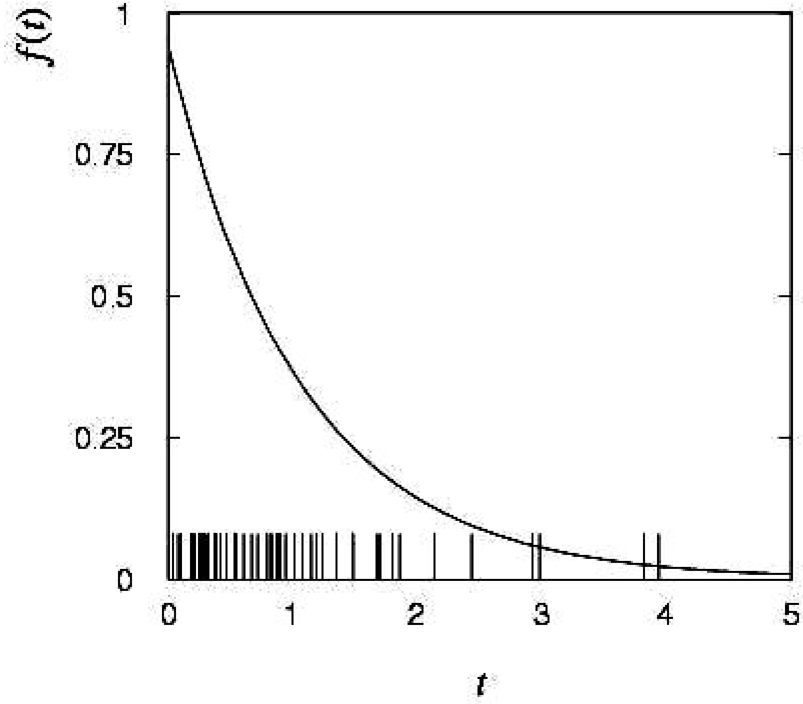
The value of τ for which $L(\tau)$ is maximum also gives the maximum value of its logarithm (the log-likelihood function):

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

ML example: parameter of exponential pdf (2)

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

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



Monte Carlo test:

generate 50 values

using $\tau = 1$:

We find the ML estimate:

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

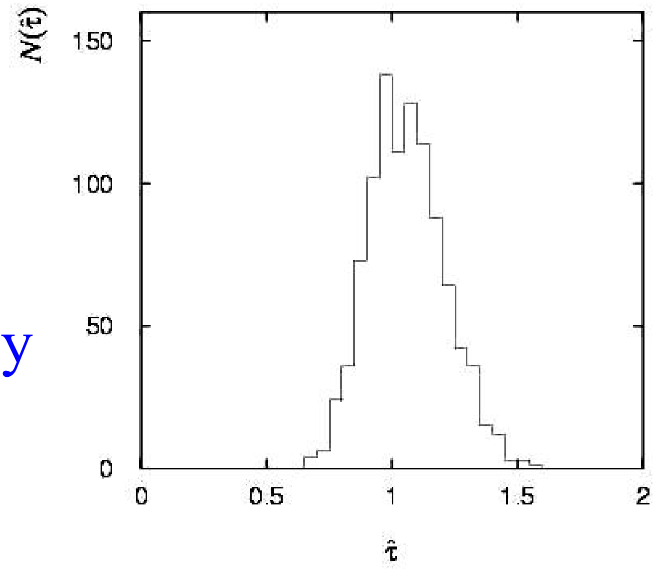
Variance of estimators: Monte Carlo method

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

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

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

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



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

Variance of estimators: graphical method

Expand ln L (θ) about its maximum:

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

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

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

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

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

Example of variance by graphical method

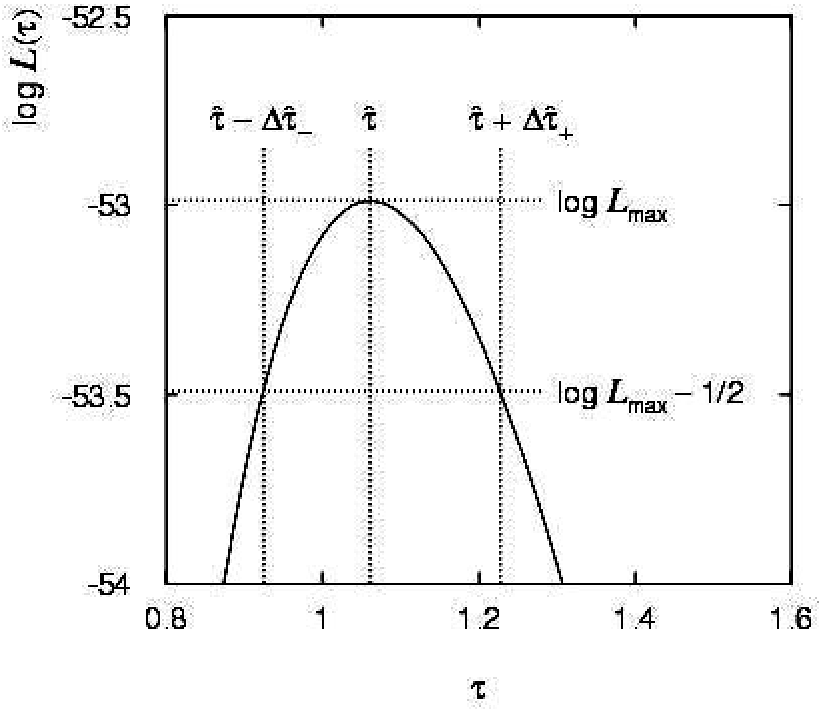
ML example with exponential:

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

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

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

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

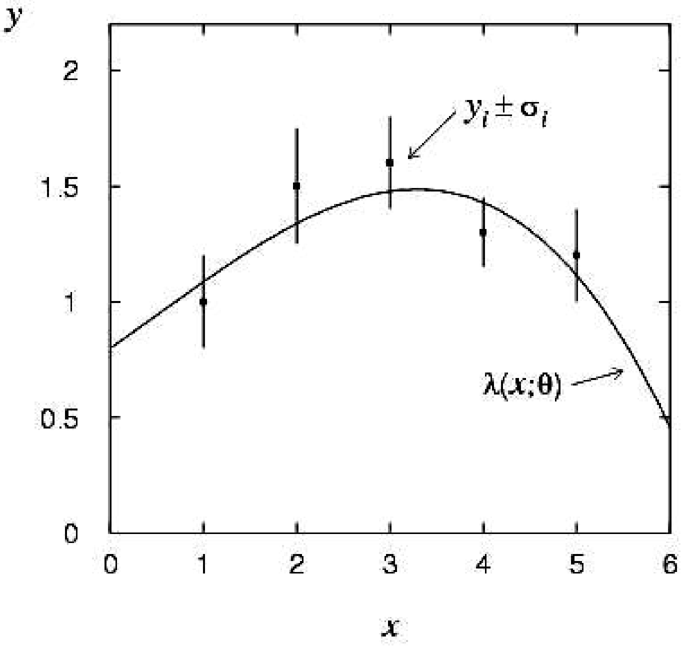


Not quite parabolic ln *L* since finite sample size (n = 50).

The method of least squares

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

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



Assume known values of the control variable $x_1, \ldots, x_N$ 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

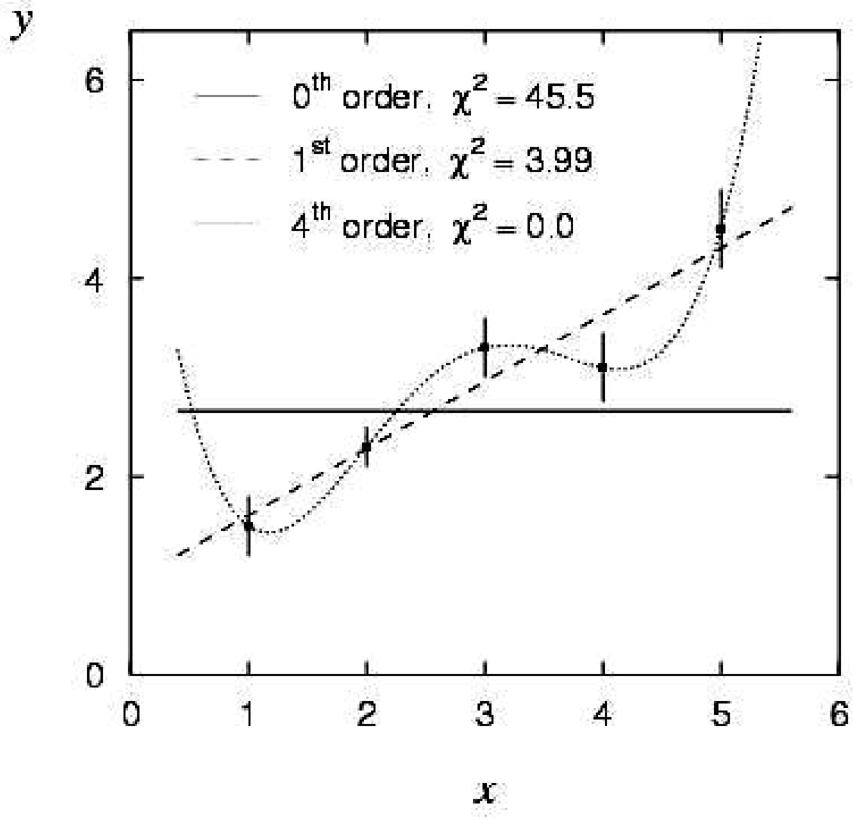
$$\chi^2(\theta) = \sum_{i=1}^{N}\frac{(y_i - \lambda(x_i;\theta))^2}{\sigma_i^2}$$

Minimum of this quantity defines the least squares estimator $\hat{\theta}$.

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

Example of least squares fit

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



Variance of LS estimators

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

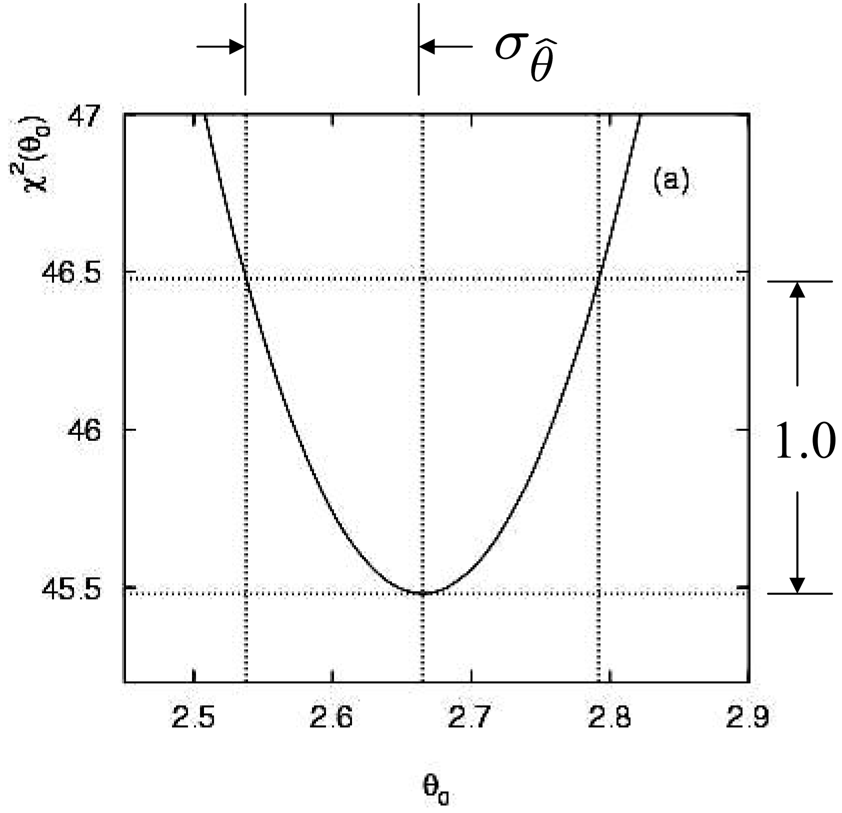
$$\chi^2(\theta) = -2\ln L(\theta)$$

and so

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

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

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



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:

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

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 $t = \chi^2{}_{\min}$ follows the chi-square pdf,

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

where the number of degrees of freedom is

$n_{\rm d}$ = 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 $\chi^2_{\rm min} \approx n_d$ the fit is 'good'.

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

This is the probability of obtaining a $\chi^2_{\rm min}$ 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),

$$\chi^2_{\rm min} = 3.99\,, \qquad n_{\rm d} = 5-2 = 3\,, \qquad p = 0.263$$

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

$$\chi^2_{\rm min} = 45.5\,, \qquad n_{\rm d} = 5-1 = 4\,, \qquad p = 3.1\times 10^{-9}$$

Setting limits

Frequentist intervals (limits) for a parameter *s* can be found by defining a test of the hypothesized value *s* (do this for all *s*):

Specify values of the data *n* that are ‘disfavoured’ by *s* (critical region) such that $P(n \text{ in critical region}) \leq \gamma$ for a prespecified γ, e.g., 0.05 or 0.1.

(Because of discrete data, need inequality here.)

If *n* is observed in the critical region, reject the value *s*.

Now invert the test to define a confidence interval as:

set of *s* values that would not be rejected in a test of size γ (confidence level is $1 - \gamma$).

The interval will cover the true value of *s* with probability $\geq 1 - \gamma$.

Setting limits

Consider again the case of finding $n = n_s + n_b$ events where

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

are Poisson r.v.s with means s, b, and thus $n = n_s + n_b$ is also Poisson with mean = $s + b$. Assume b is known.

Suppose we are searching for evidence of the signal process, but the number of events found is roughly equal to the expected number of background events, e.g., $b = 4.6$ and we observe $n_{\mathrm{obs}} = 5$ events.

The evidence for the presence of signal events is not statistically significant,

→ set upper limit on the parameter s.

Example of an upper limit

Find the hypothetical value of s such that there is a given small probability, say, $\gamma = 0.05$, to find as few events as we did or less:

$$\gamma = P(n \leq n_{\mathrm{obs}}; s, b) = \sum_{n=0}^{n_{\mathrm{obs}}} \frac{(s+b)^n}{n!} e^{-(s+b)}$$

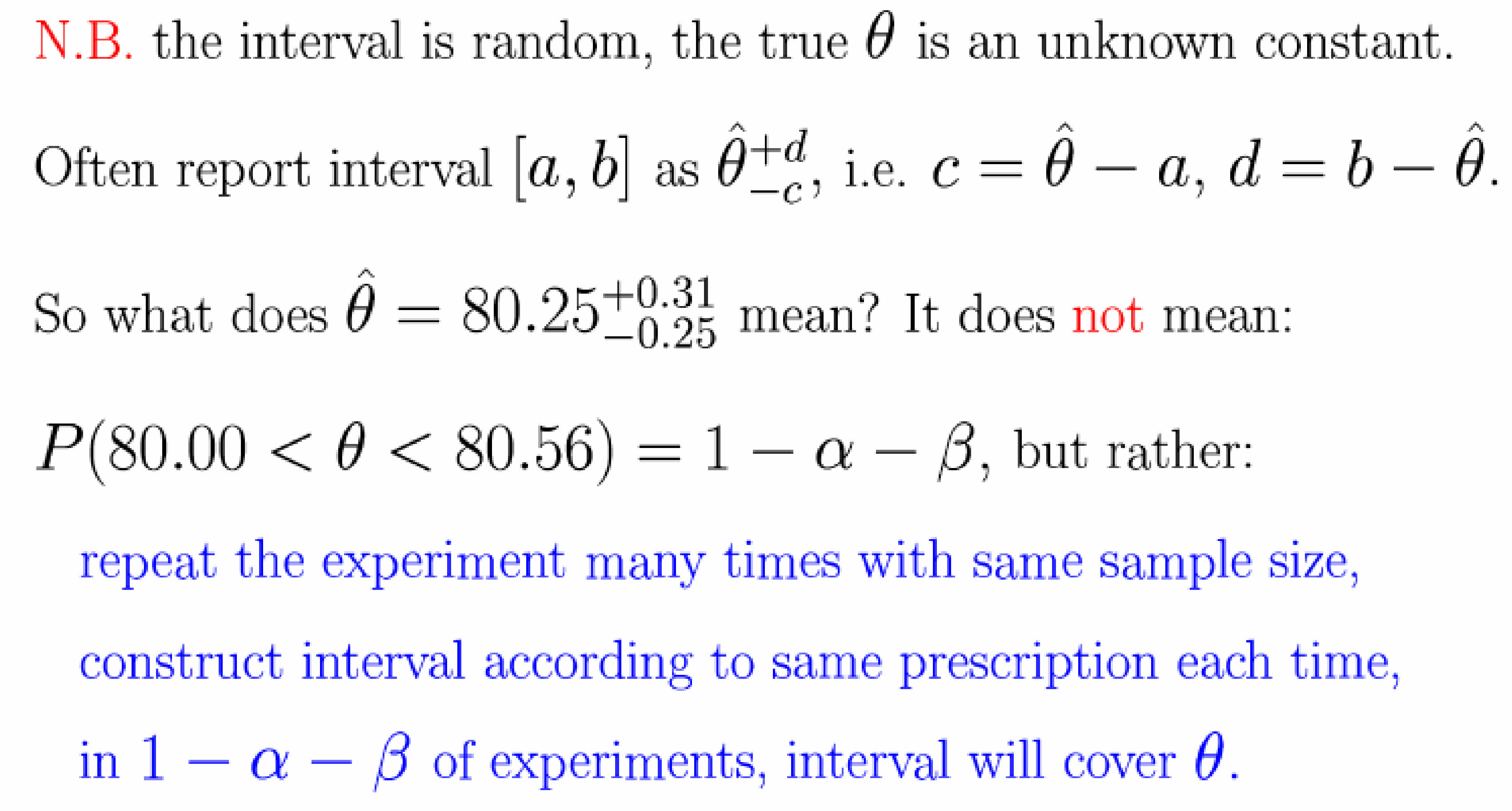
Solve numerically for $s = s_{\mathrm{up}}$, this gives an upper limit on s at a confidence level of $1-\gamma$.

Example: suppose $b = 0$ and we find $n_{\mathrm{obs}} = 0$. For $1-\gamma = 0.95$,

$$\gamma = P(n = 0; s, b = 0) = e^{-s} \quad \rightarrow \quad s_{\mathrm{up}} = -\ln\gamma \approx 3.00$$

The interval $[0, s_{\mathrm{up}}]$ is an example of a confidence interval, designed to cover the true value of s with a probability $1-\gamma$.

Meaning of a confidence interval



N.B. the interval is random, the true θ is an unknown constant.

Often report interval $[a, b]$ as $\hat{\theta}^{+d}_{-c}$, i.e. $c = \hat{\theta} - a$, $d = b - \hat{\theta}$.

So what does $\hat{\theta} = 80.25^{+0.31}_{-0.25}$ mean? It does not mean:

$P(80.00 < \theta < 80.56) = 1 - \alpha - \beta$, but rather:

repeat the experiment many times with same sample size,

construct interval according to same prescription each time,

in $1 - \alpha - \beta$ of experiments, interval will cover θ.

Calculating Poisson parameter limits

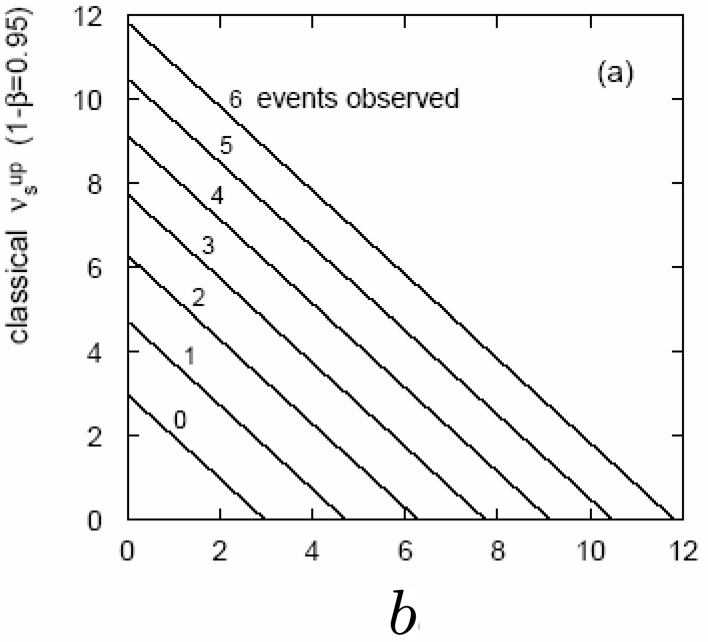
To solve for s_{lo}, s_{up}, can exploit relation to χ^2 distribution:

$$s_{\mathrm{lo}} = \frac{1}{2} F^{-1}_{\chi^2}(\alpha; 2n) - b$$

Quantile of χ^2 distribution
`TMath::ChisquareQuantile`

$$s_{\mathrm{up}} = \frac{1}{2} F^{-1}_{\chi^2}(1-\beta; 2(n+1)) - b$$

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



Many subtle issues here – see e.g. CERN (2000) and Fermilab (2001) confidence limit workshops and PHYSTAT conferences.

Back up slides

Variance of estimators from information inequality

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

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

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

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

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

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