

# Topics in Statistical Data Analysis for HEP

## Lecture 1: Intro / Parameter Estimation / Tests



CERN – Latin-American School  
on High Energy Physics

Natal, Brazil, 3 April 2011



Glen Cowan

Physics Department

Royal Holloway, University of London

[g.cowan@rhul.ac.uk](mailto:g.cowan@rhul.ac.uk)

[www.pp.rhul.ac.uk/~cowan](http://www.pp.rhul.ac.uk/~cowan)

# Outline

→ Lecture 1: Introduction and basic formalism

Probability

Parameter estimation

Statistical tests

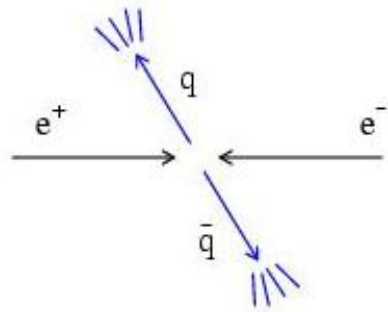
Lecture 2: Statistics for making a discovery

Multivariate methods

Discovery significance and sensitivity

Systematic uncertainties

# Data analysis in particle physics



Observe events of a certain type

Measure characteristics of each event (particle momenta, number of muons, energy of jets,...)

Theories (e.g. SM) predict distributions of these properties up to free parameters, e.g.,  $\alpha$ ,  $G_F$ ,  $M_Z$ ,  $\alpha_s$ ,  $m_H$ , ...

Some tasks of data analysis:

Estimate (measure) the parameters;

Quantify the uncertainty of the parameter estimates;

Test the extent to which the predictions of a theory are in agreement with the data ( $\rightarrow$  presence of New Physics?)

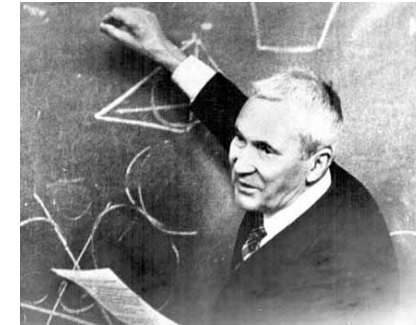
# A definition of probability

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:

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

# 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

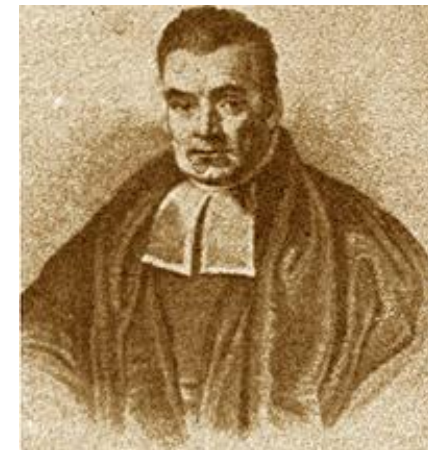
From the definition of conditional probability we have

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

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.

# 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),

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

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.

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

# Frequentist approach to parameter estimation

The parameters of a probability distribution function (pdf) are constants that characterize its shape, e.g.

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

random variable  $\nearrow$   $\nwarrow$  parameter

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

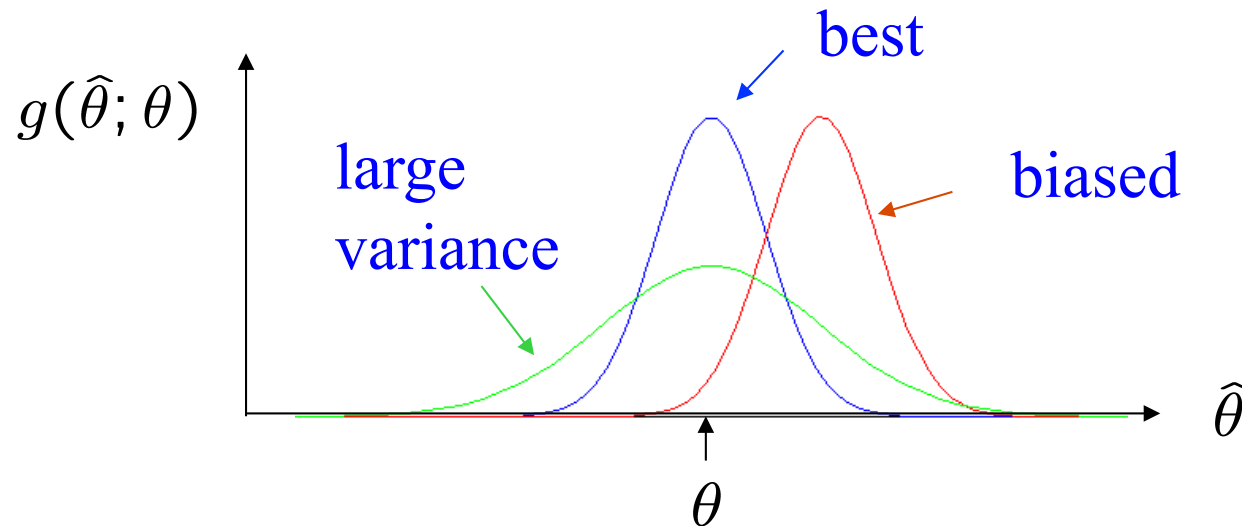
We want to find some function of the data to **estimate** the parameter(s):

$$\hat{\theta}(\vec{x}) \leftarrow \text{estimator written with a hat}$$



# 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

# 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  $\theta$ :

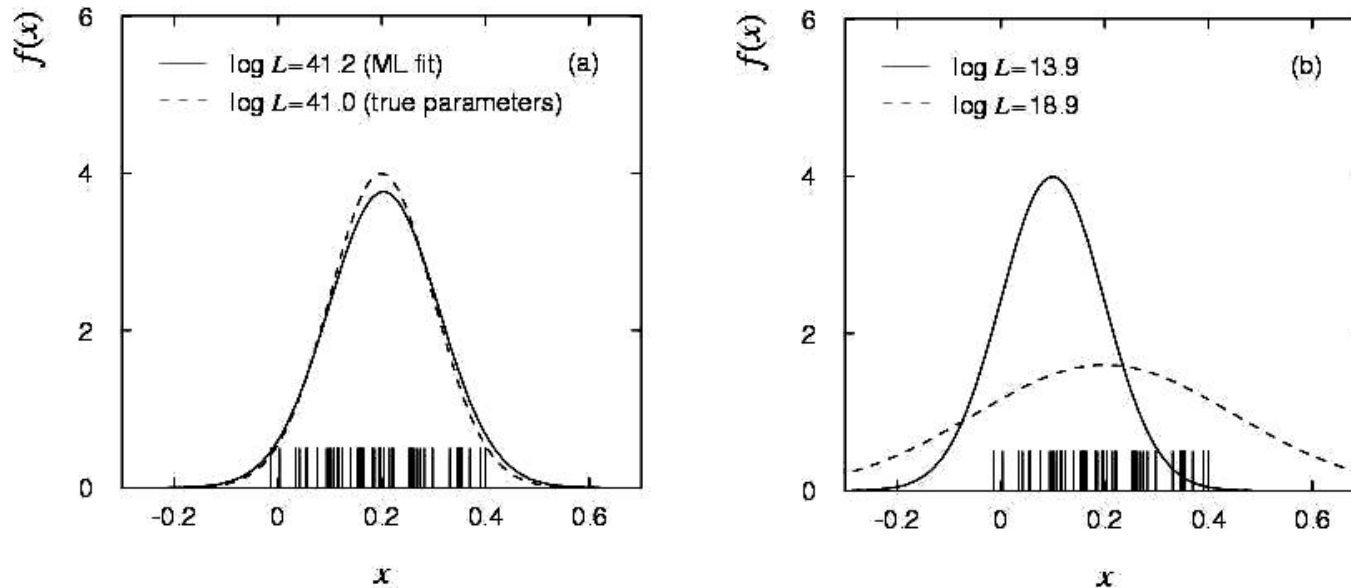
$$f(\vec{x}; \vec{\theta})$$

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

$$L(\vec{\theta}) = f(\vec{x}; \vec{\theta}) \quad (\mathbf{x} \text{ constant})$$

# Maximum likelihood estimators

If the hypothesized  $\theta$  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).

# Example: fitting a straight line

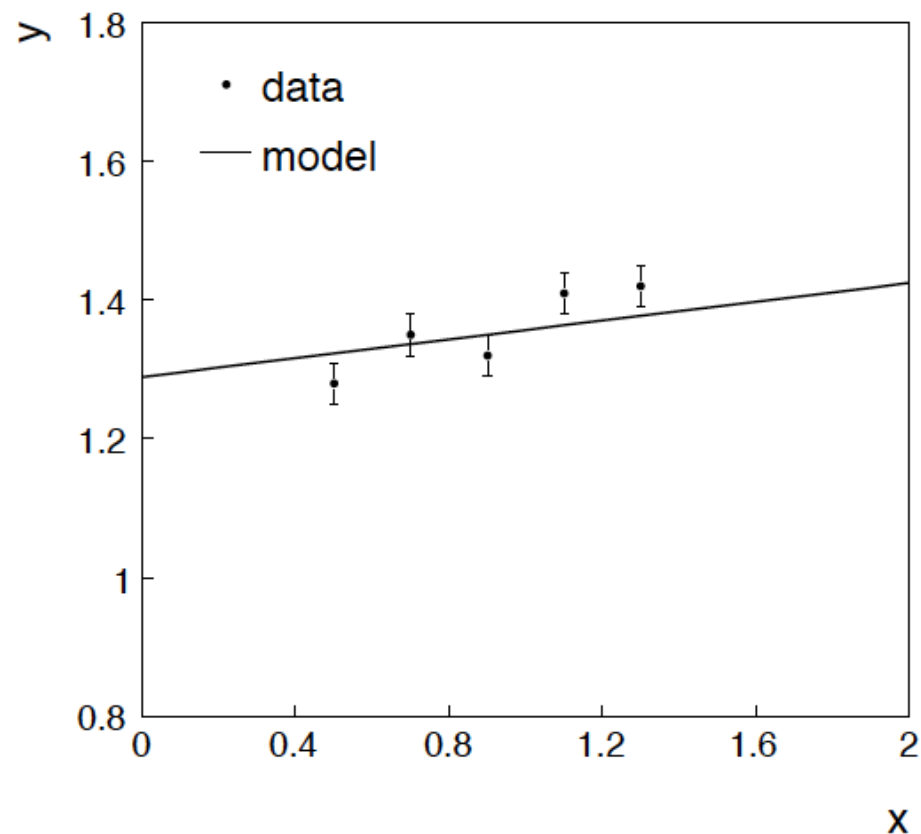
Data:  $(x_i, y_i, \sigma_i), i = 1, \dots, n$ .

Model: measured  $y_i$  independent, Gaussian:  $y_i \sim N(\mu(x_i), \sigma_i^2)$

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

assume  $x_i$  and  $\sigma_i$  known.

Goal: estimate  $\theta_0$   
(don't care about  $\theta_1$ ).



# 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)

## Case #1: $\theta_1$ known a priori

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

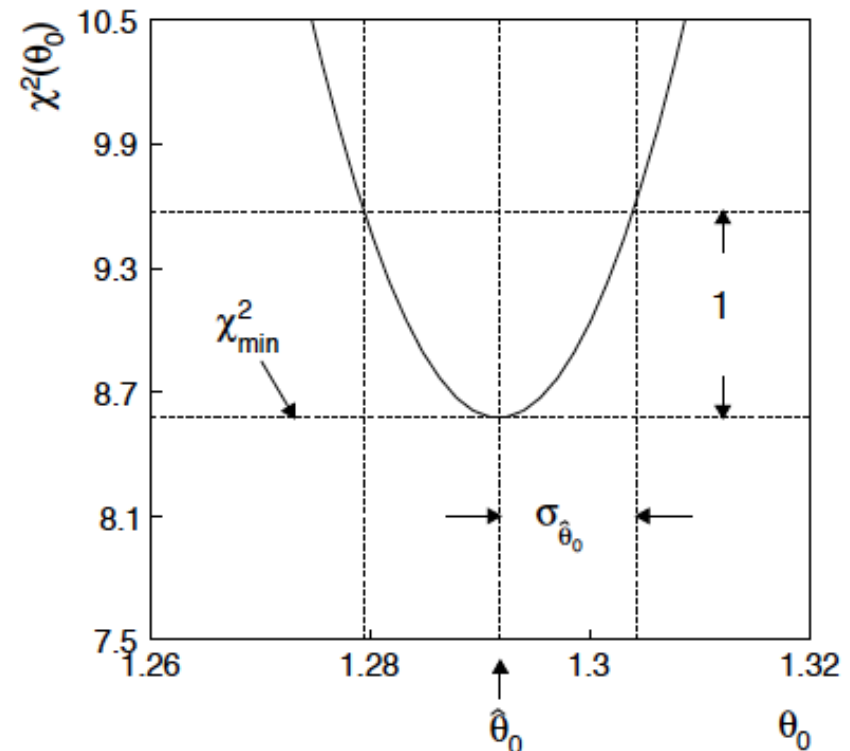
$$\chi^2(\theta_0) = -2 \ln L(\theta_0) + \text{const} = \sum_{i=1}^n \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} .$$

For Gaussian  $y_i$ , ML same as LS

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

Come up one unit from  $\chi_{\min}^2$

to find  $\sigma_{\hat{\theta}_0}$  .



## Case #2: both $\theta_0$ and $\theta_1$ unknown

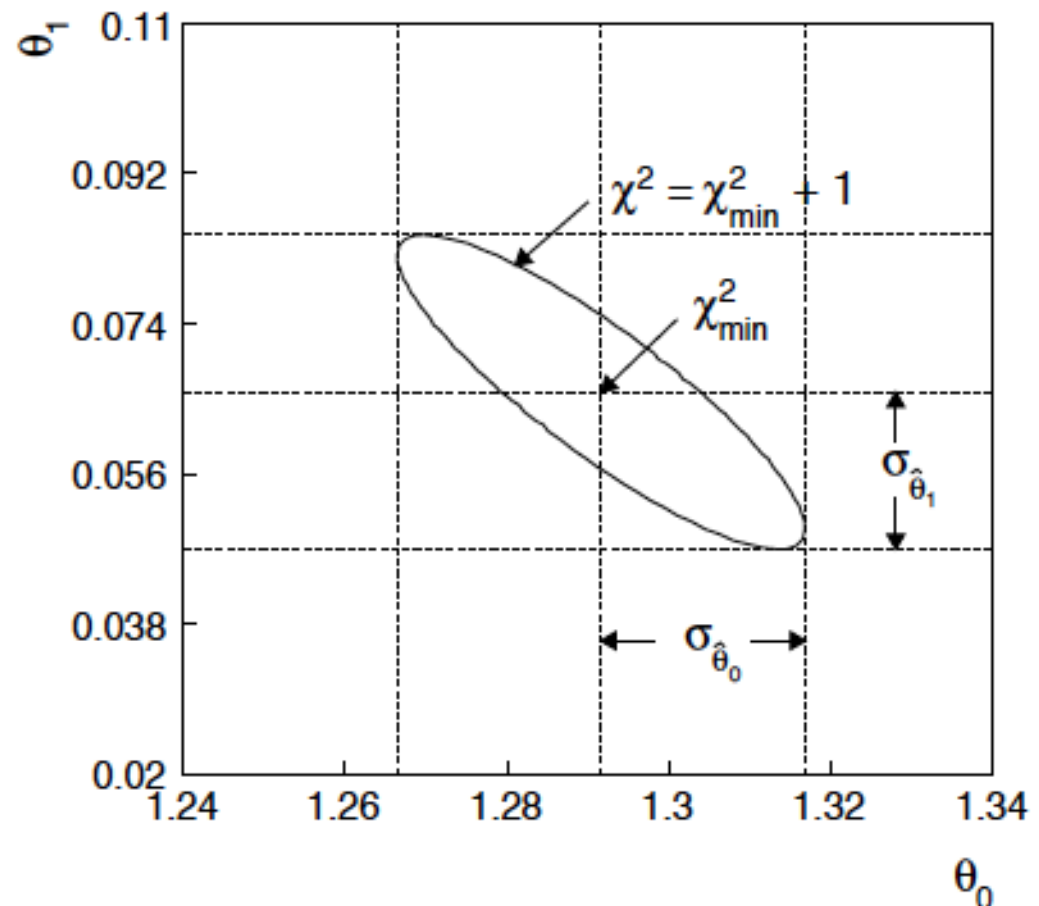
$$\chi^2(\theta_0, \theta_1) = -2 \ln L(\theta_0, \theta_1) + \text{const} = \sum_{i=1}^n \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2} .$$

Standard deviations from  
tangent lines to contour

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

Correlation between

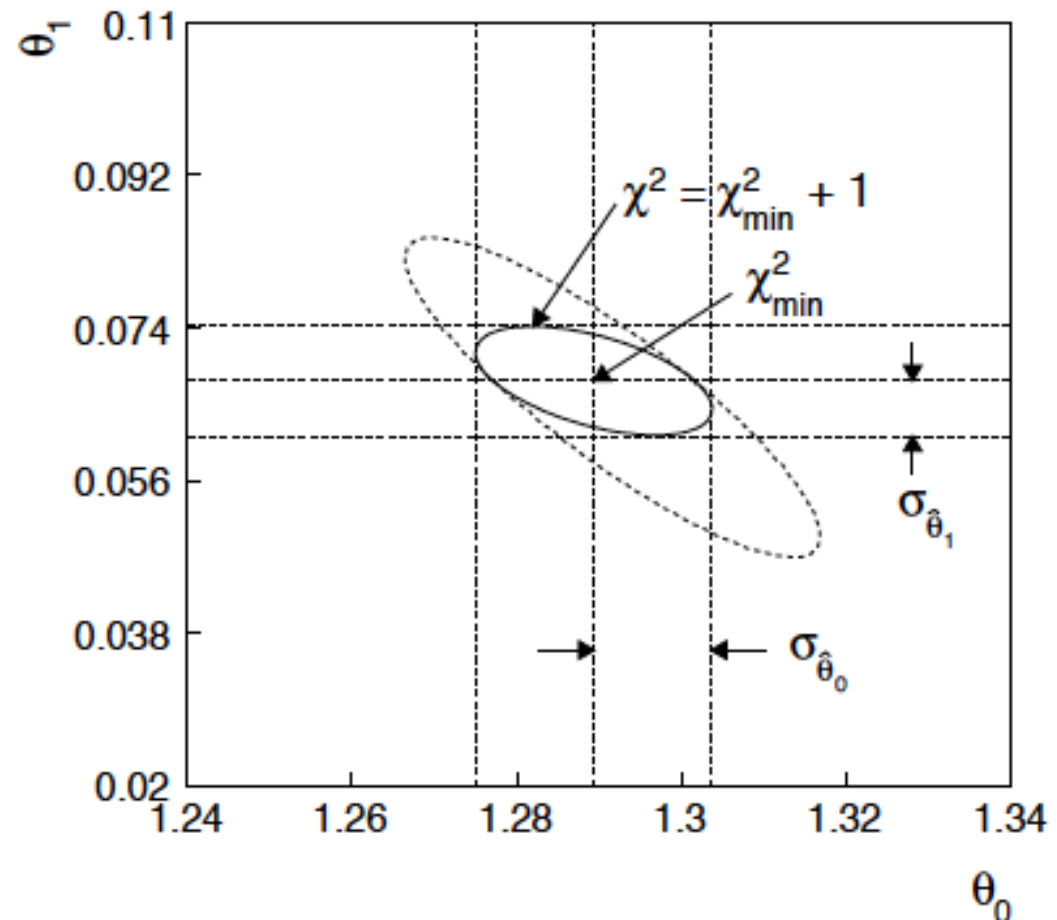
$\hat{\theta}_0, \hat{\theta}_1$  causes errors  
to increase.



## Case #3: we have a measurement $t_1$ of $\theta_1$

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

The information on  $\theta_1$   
improves accuracy of  $\hat{\theta}_0$ .





# 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.)

## Case #4: Bayesian method

We need to associate prior probabilities with  $\theta_0$  and  $\theta_1$ , e.g.,

$$\begin{aligned} \pi(\theta_0, \theta_1) &= \pi_0(\theta_0) \pi_1(\theta_1) && \text{reflects 'prior ignorance', in any} \\ \pi_0(\theta_0) &= \text{const.} && \text{case much broader than } L(\theta_0) \\ \pi_1(\theta_1) &= \frac{1}{\sqrt{2\pi}\sigma_{t_1}} e^{-(\theta_1 - t_1)^2 / 2\sigma_{t_1}^2} && \leftarrow \text{based on previous} \\ &&& \text{measurement} \end{aligned}$$

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.

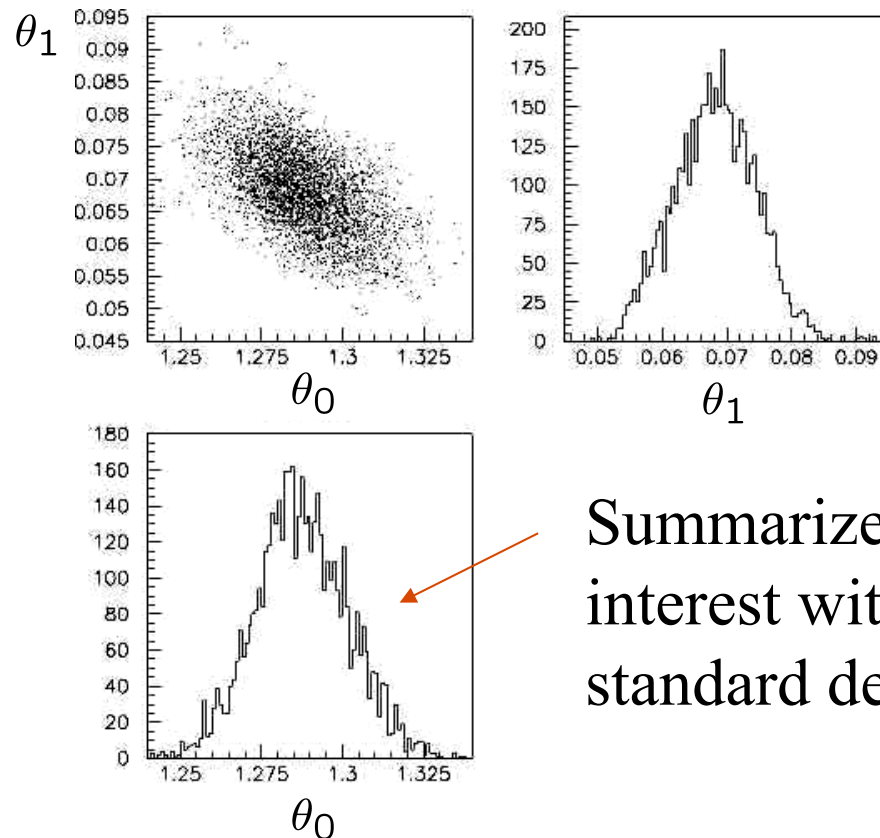
MCMC (e.g., Metropolis-Hastings algorithm) generates  
**correlated** sequence of random numbers:

cannot use for many applications, e.g., detector MC;  
effective stat. error greater than naive  $\sqrt{n}$  .

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

# 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?)

# 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 naive  $\sqrt{n}$  .

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  $p(\vec{\theta})/p(\vec{\theta}_0)$  .

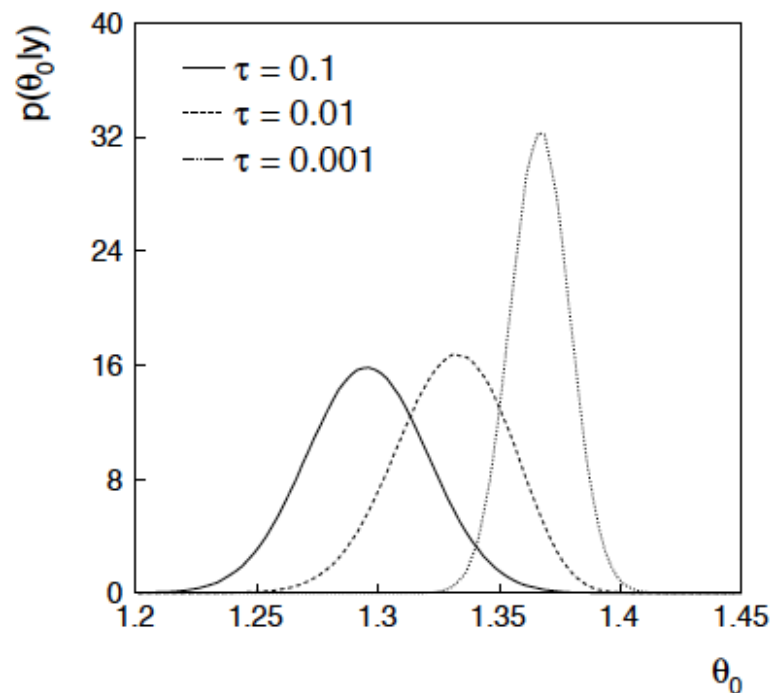
If proposed step rejected, hop in place.

# Bayesian method with alternative priors

Suppose we don't have a previous measurement of  $\theta_1$  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  $\theta_0$ :



This summarizes all knowledge about  $\theta_0$ .

Look also at result from variety of priors.



# A more general fit (symbolic)

Given measurements:  $y_i \pm \sigma_i^{\text{stat}} \pm \sigma_i^{\text{sys}}, \quad i = 1, \dots, n,$

and (usually) covariances:  $V_{ij}^{\text{stat}}, V_{ij}^{\text{sys}}.$

Predicted value:  $\mu(x_i; \theta),$  expectation value  $E[y_i] = \mu(x_i; \theta) + b_i$   
control variable  $\nearrow$  parameters  $\nearrow$  bias  $\nearrow$

Often take:  $V_{ij} = V_{ij}^{\text{stat}} + V_{ij}^{\text{sys}}$

Minimize  $\chi^2(\theta) = (\vec{y} - \vec{\mu}(\theta))^T V^{-1} (\vec{y} - \vec{\mu}(\theta))$

Equivalent to maximizing  $L(\theta) \gg e^{-\chi^2/2},$  i.e., least squares same as maximum likelihood using a Gaussian likelihood function.

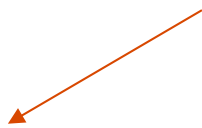
# Its Bayesian equivalent

Take  $L(\vec{y}|\vec{\theta}, \vec{b}) \sim \exp \left[ -\frac{1}{2}(\vec{y} - \vec{\mu}(\theta) - \vec{b})^T V_{\text{stat}}^{-1} (\vec{y} - \vec{\mu}(\theta) - \vec{b}) \right]$

$$\pi_b(\vec{b}) \sim \exp \left[ -\frac{1}{2} \vec{b}^T V_{\text{sys}}^{-1} \vec{b} \right]$$

$$\pi_\theta(\theta) \sim \text{const.}$$

Joint probability  
for all parameters



and use Bayes' theorem:  $p(\theta, \vec{b}|\vec{y}) \propto L(\vec{y}|\theta, \vec{b})\pi_\theta(\theta)\pi_b(\vec{b})$

To get desired probability for  $\theta$ , integrate (marginalize) over  $\mathbf{b}$ :

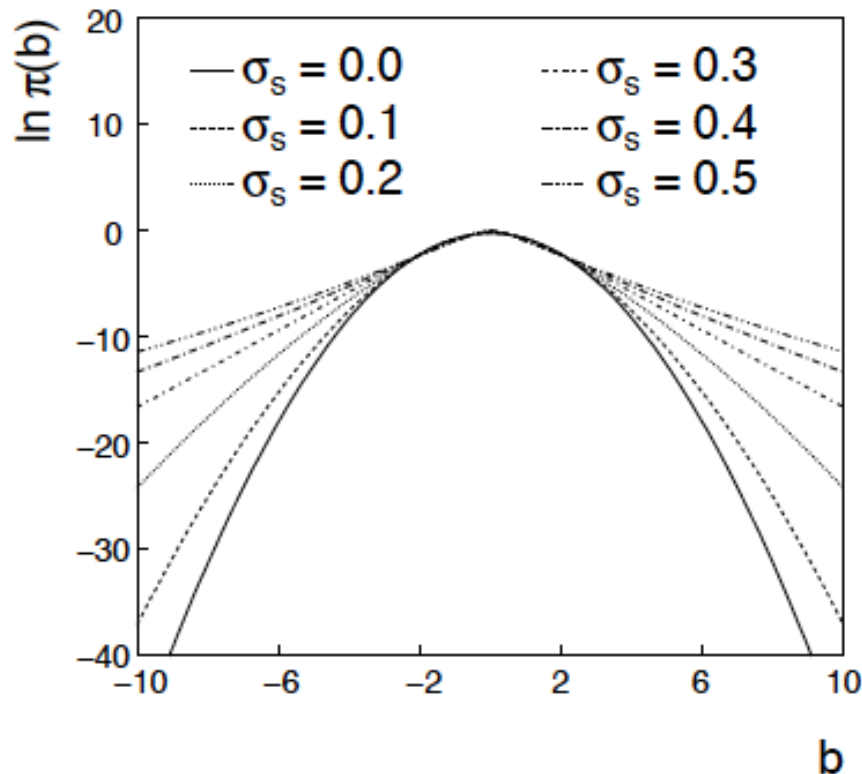
$$p(\theta|\vec{y}) = \int p(\theta, \vec{b}|\vec{y}) d\vec{b}$$

→ Posterior is Gaussian with mode same as least squares estimator,  $\sigma_\theta$  same as from  $\chi^2 = \chi^2_{\text{min}} + 1$ . (Back where we started!)

# Alternative priors for systematic errors

Gaussian prior for the bias  $b$  often not realistic, especially if one considers the "error on the error". Incorporating this can give a prior with longer tails:

$$\pi_b(b_i) = \int \frac{1}{\sqrt{2\pi s_i \sigma_i^{\text{sys}}}} \exp \left[ -\frac{1}{2} \frac{b_i^2}{(s_i \sigma_i^{\text{sys}})^2} \right] \pi_s(s_i) ds_i$$



Represents 'error on the error'; standard deviation of  $\pi_s(s)$  is  $\sigma_s$ .

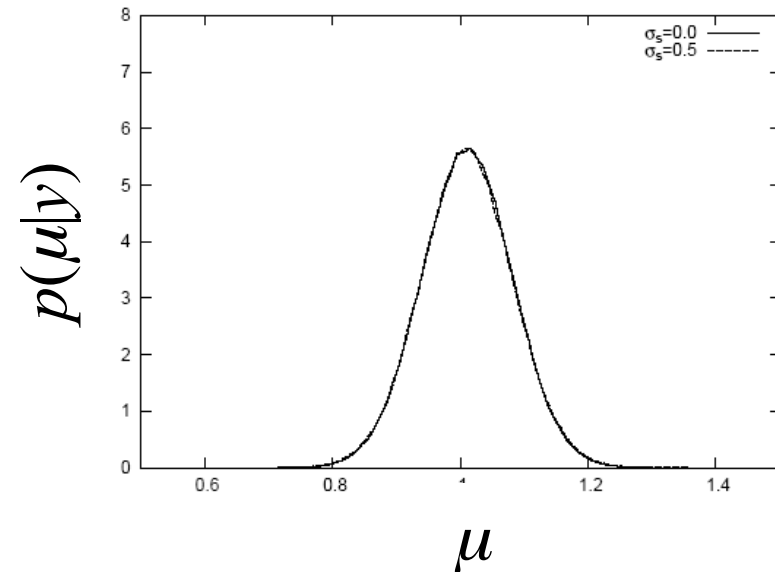
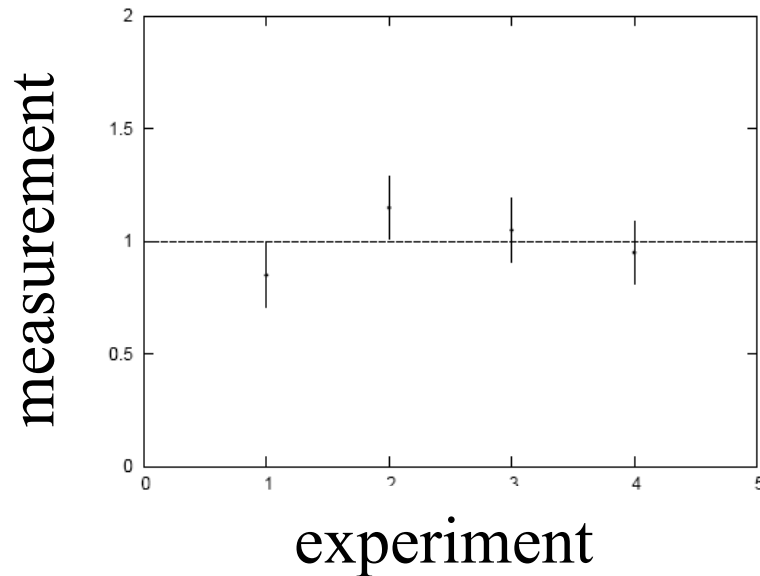
# A simple test

Suppose fit effectively averages four measurements.

Take  $\sigma_{\text{sys}} = \sigma_{\text{stat}} = 0.1$ , uncorrelated.

Case #1: data appear compatible

Posterior  $p(\mu|y)$ :



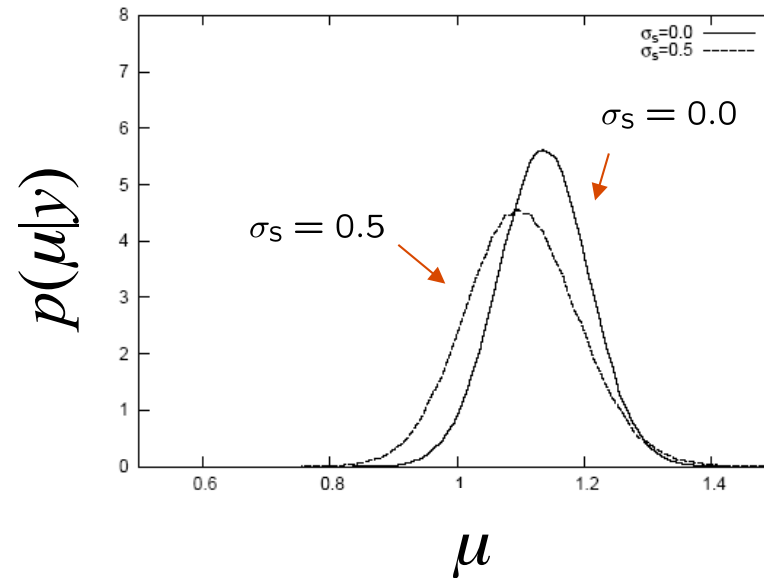
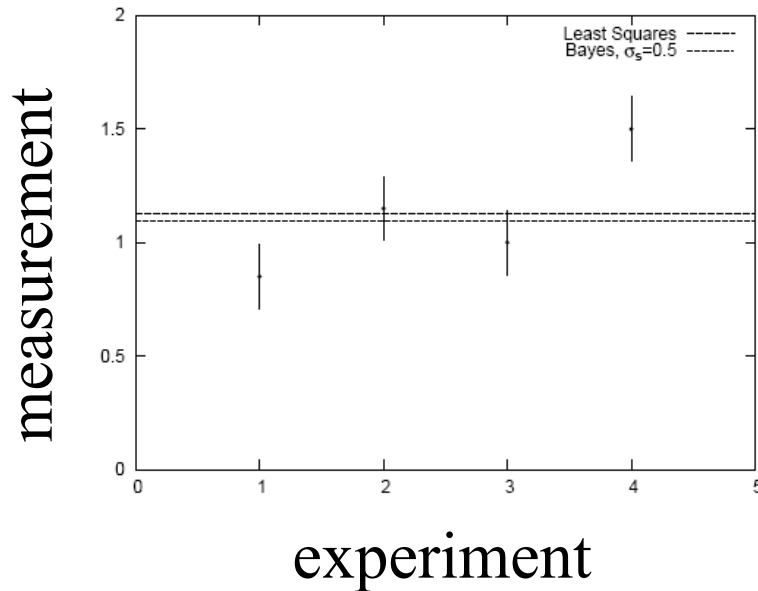
Usually summarize posterior  $p(\mu|y)$   
with mode and standard deviation:

$$\begin{aligned} \sigma_s = 0.0 : & \quad \hat{\mu} = 1.000 \pm 0.071 \\ \sigma_s = 0.5 : & \quad \hat{\mu} = 1.000 \pm 0.072 \end{aligned}$$

# Simple test with inconsistent data

Case #2: there is an outlier

Posterior  $p(\mu|y)$ :



$$\sigma_s = 0.0 : \quad \hat{\mu} = 1.125 \pm 0.071$$

$$\sigma_s = 0.5 : \quad \hat{\mu} = 1.093 \pm 0.089$$

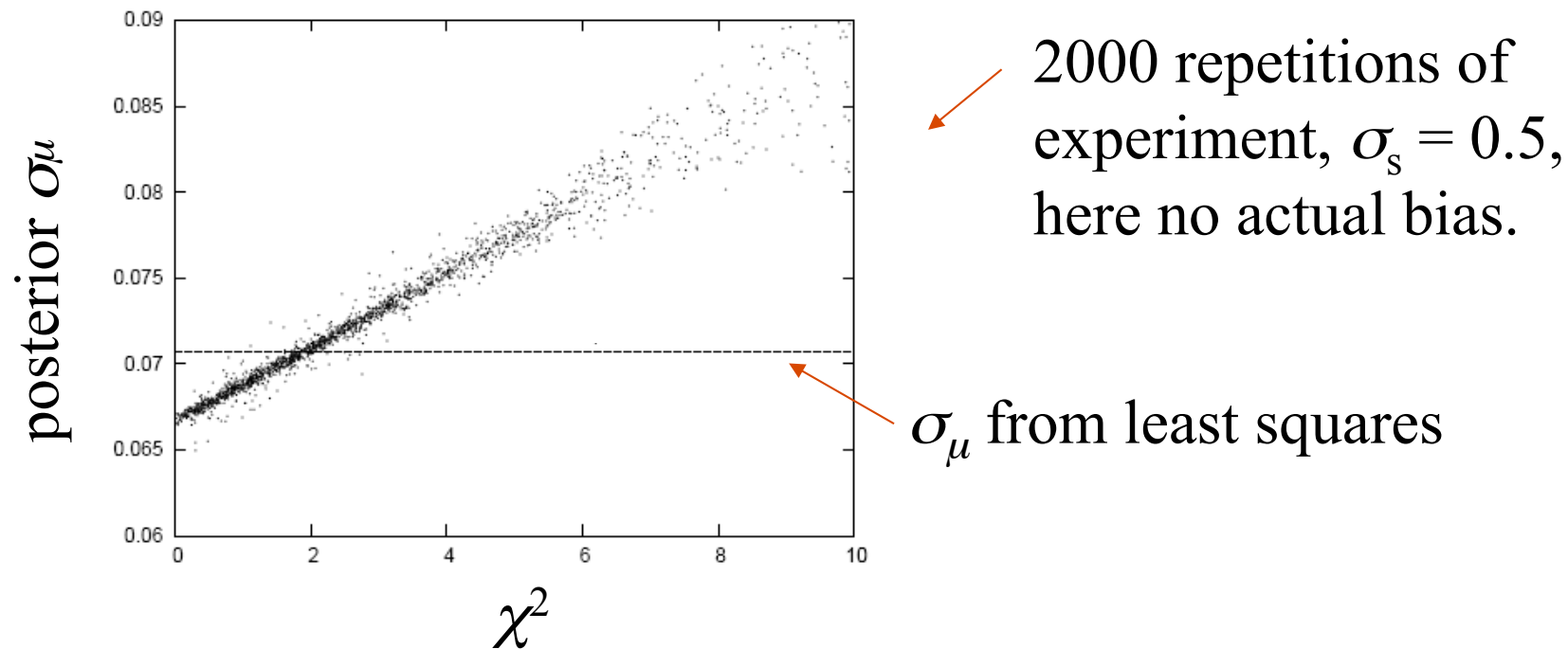
→ Bayesian fit less sensitive to outlier.

(See also D'Agostini 1999; Dose & von der Linden 1999)

# Goodness-of-fit vs. size of error

In LS fit, value of minimized  $\chi^2$  does not affect size of error on fitted parameter.

In Bayesian analysis with non-Gaussian prior for systematics, a high  $\chi^2$  corresponds to a larger error (and vice versa).



# Introduction to hypothesis testing

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  $\alpha$ , assuming  $H_0$  is correct, to observe the data there, i.e.,

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

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

$\alpha$  is called the **size** or **significance level** of the test.

Critical region also called “rejection” region; complement is acceptance region.

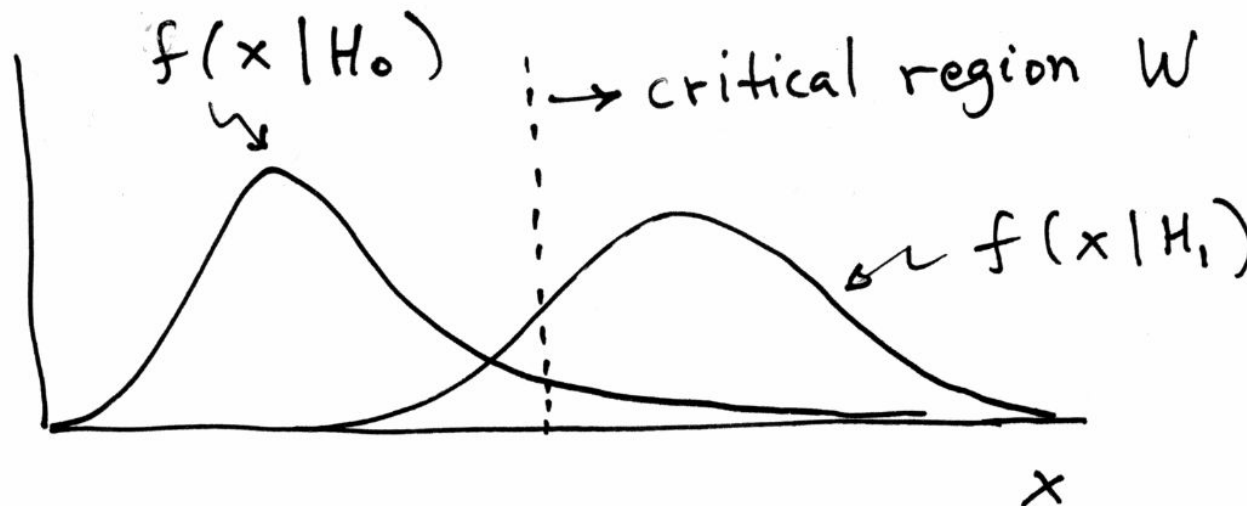


## Definition of a test (2)

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

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.

# 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.

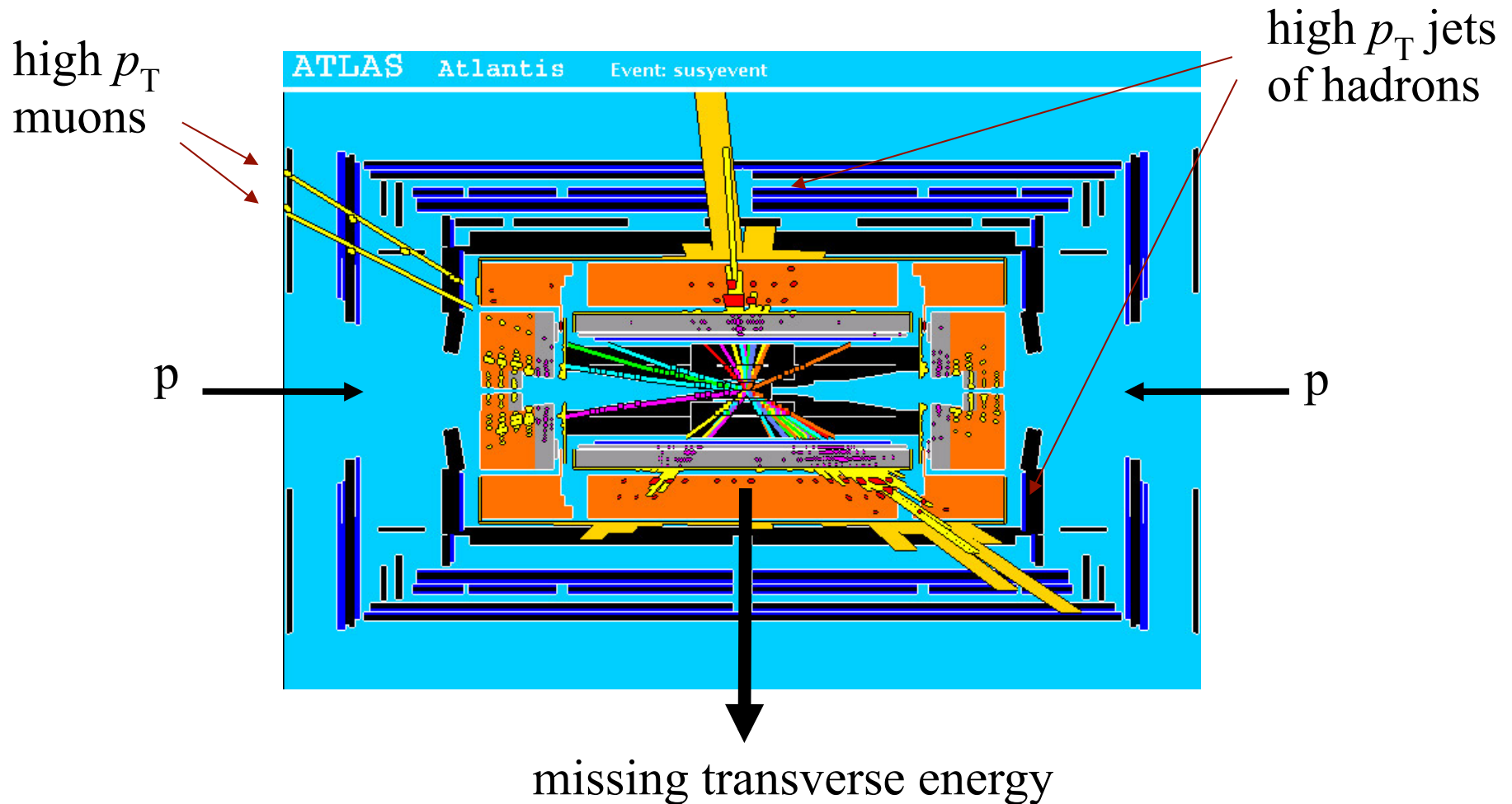
**Search for New Physics:** the null hypothesis  $H_0$  means Standard Model events, and the alternative  $H_1$  means "events of a type whose existence is not yet established" (to establish or exclude the signal model is the goal of the analysis).

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.

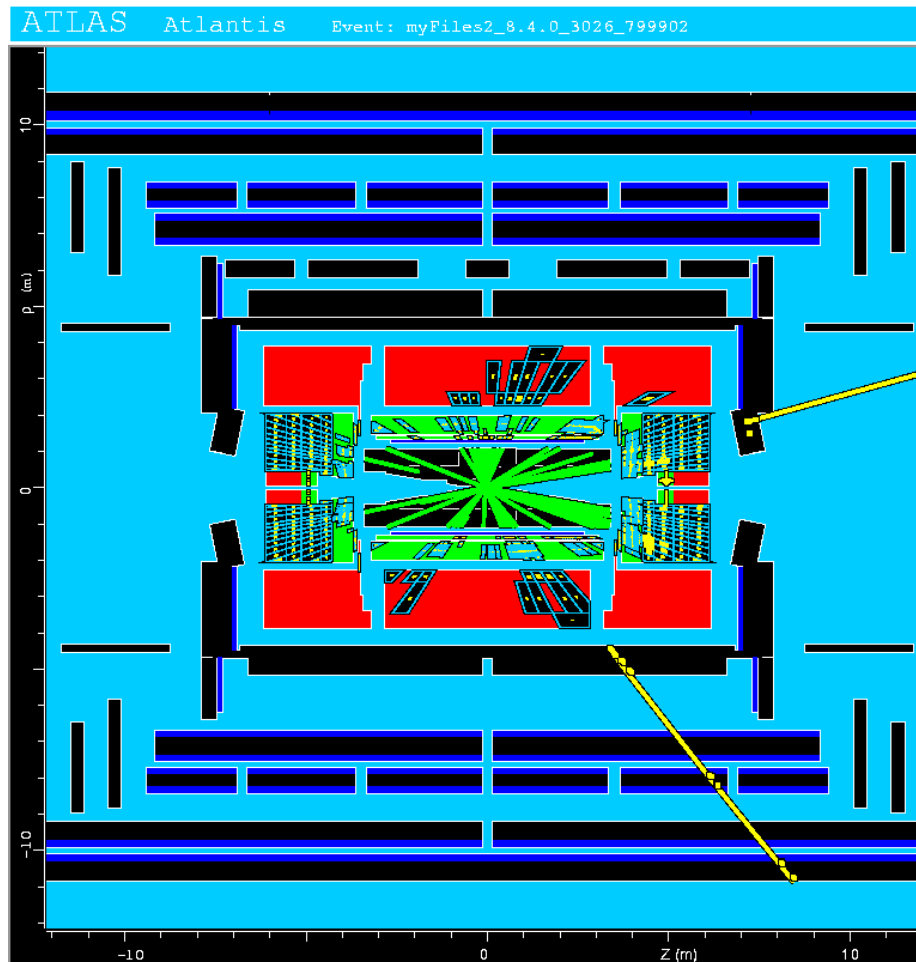
# Suppose we want to discover this...

SUSY event (ATLAS simulation):



# 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  $\vec{x} = (x_1, \dots, x_n)$

$x_1$  = 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.

Often call  $H_0$  the **background** hypothesis (e.g. SM events);  $H_1, H_2, \dots$  are possible **signal** hypotheses.

# Defining a multivariate critical region

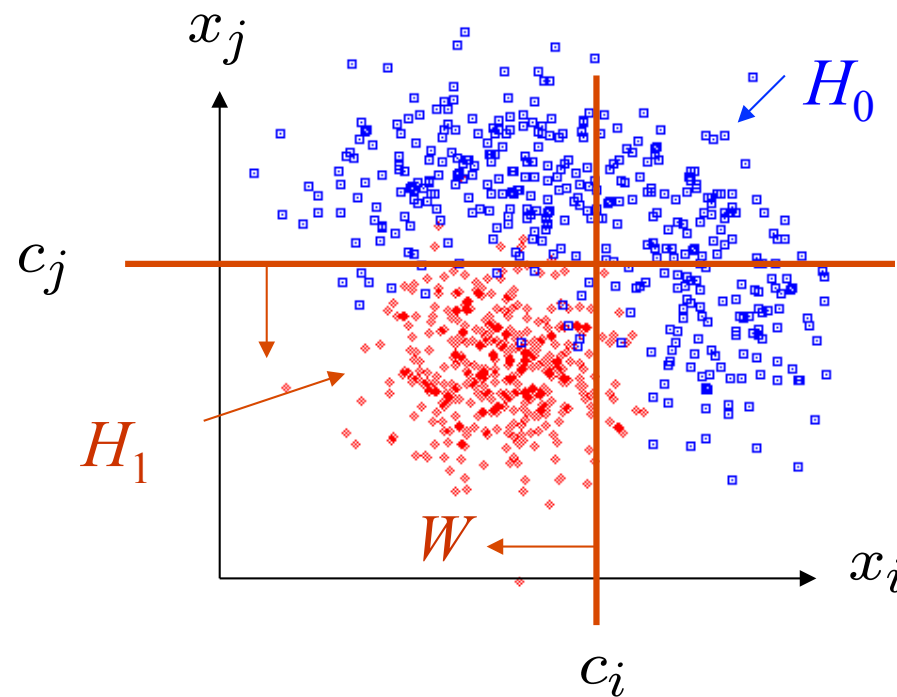
Each event is a point in  $\mathbf{x}$ -space; critical region is now defined by a ‘decision boundary’ in this space.

What is best way to determine the decision boundary?

Perhaps with ‘cuts’:

$$x_i < c_i$$

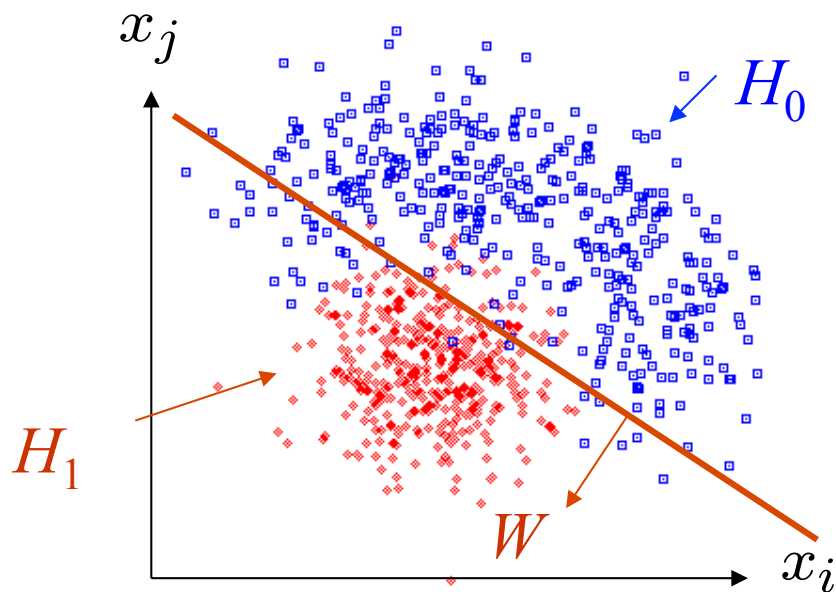
$$x_j < c_j$$



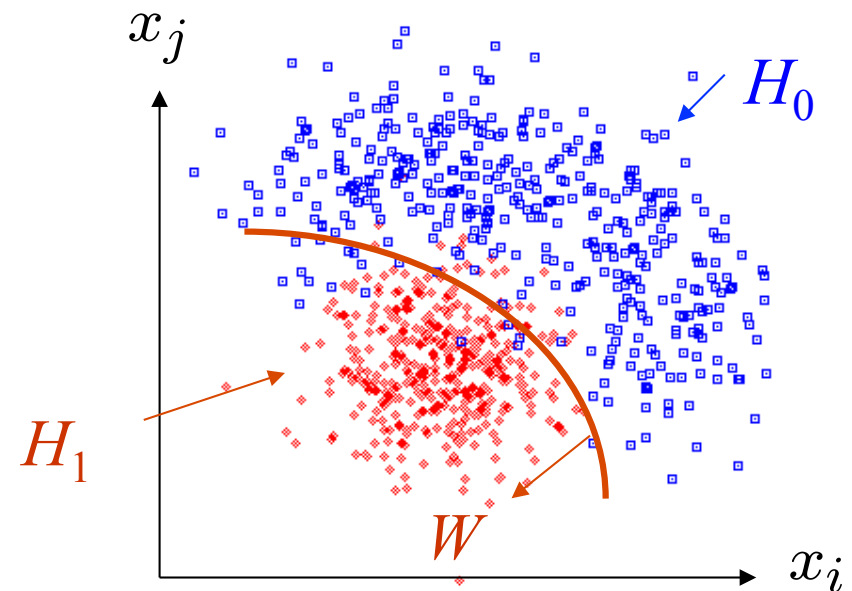
# Other multivariate decision boundaries

Or maybe use some other sort of decision boundary:

linear



or nonlinear





# Test statistics

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

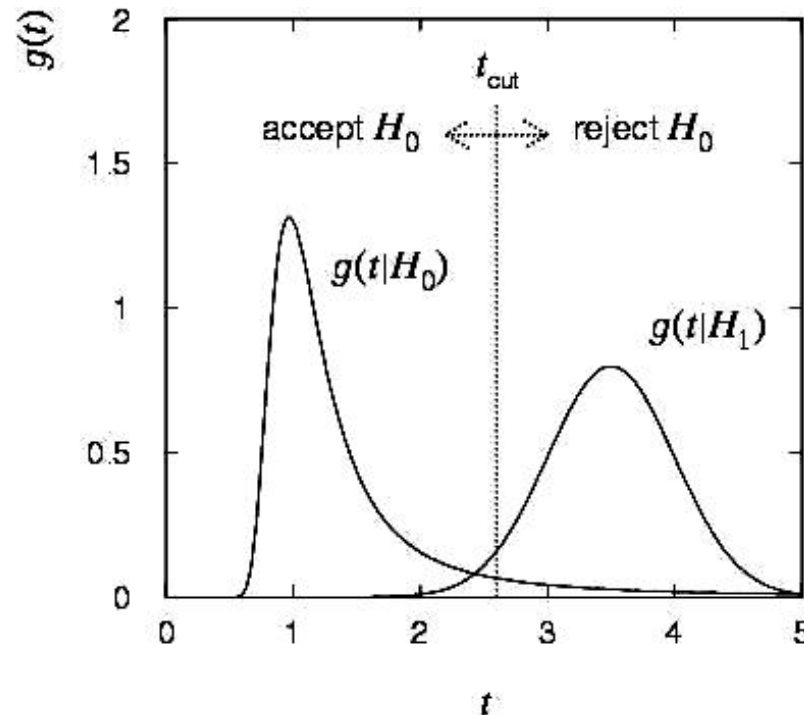
$$t(x_1, \dots, x_n) = t_{\text{cut}}$$

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

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

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):

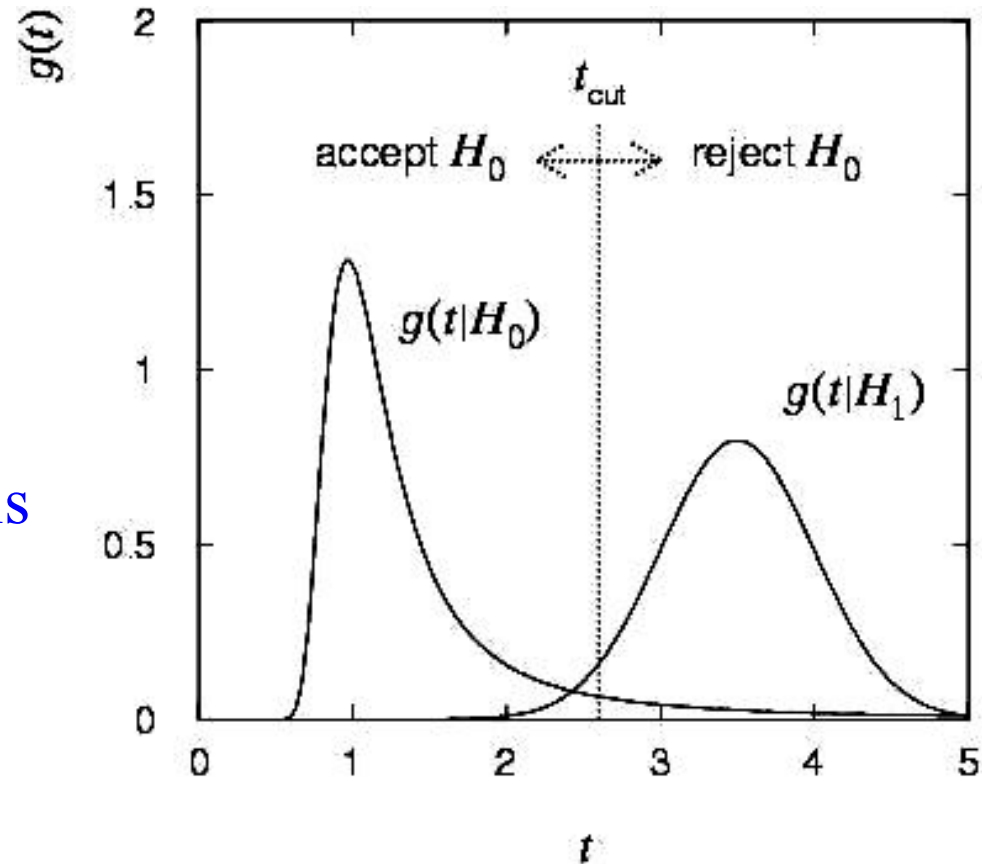
$$\alpha = \int_{t_{\text{cut}}}^{\infty} g(t|H_0) dt$$

(significance level)

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

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

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



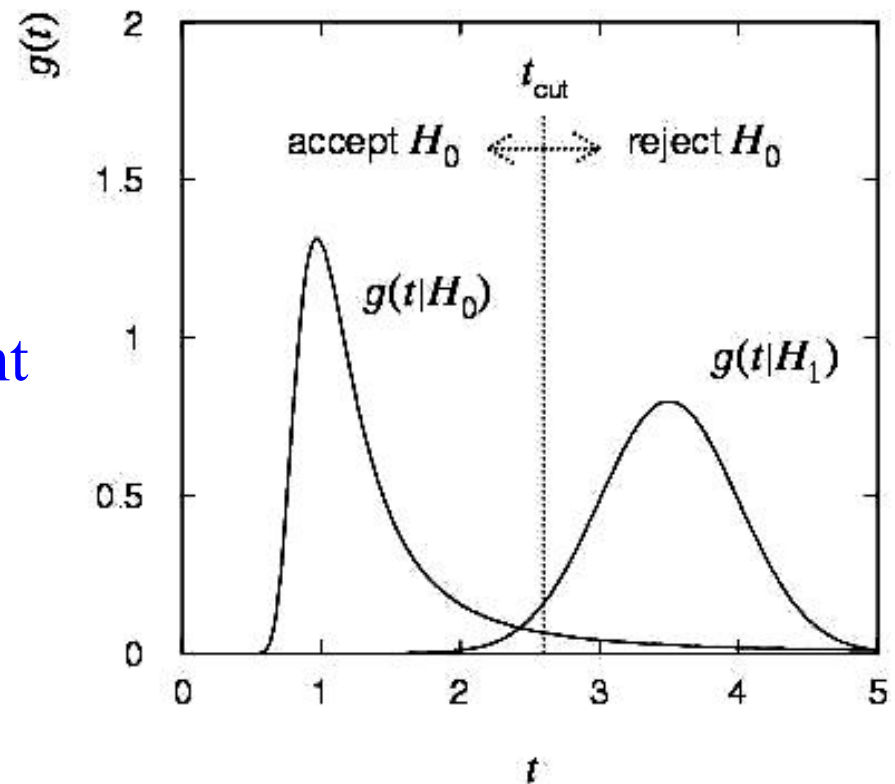
# Signal/background efficiency

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

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

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

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



# Purity of event selection

Suppose only one background type  $b$ ; overall fractions of signal and background events are  $\pi_s$  and  $\pi_b$  (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.

# Summary of lecture 1

Theoretical predictions contain in general adjustable parameters; estimating them from the data is a central task of statistics.

## Frequentist approach:

Probability only assigned to data

Construct functions of data (estimators): ML, LS

## Bayesian approach:

Probability assigned also to parameters (“degree of belief”)

Marginalize over nuisance parameters (MCMC)

## Hypothesis tests

Identify subset of data space (critical region,  $W$ ) disfavoured by hypothesis  $H_0$ , but favoured by alternative(s).

If data are observed in  $W$ , reject  $H_0$

Next lecture: how to set critical region in a multivariate space.

# Extra slides

## Some statistics books, papers, etc.

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

see also [www.pp.rhul.ac.uk/~cowan/sda](http://www.pp.rhul.ac.uk/~cowan/sda)

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

see also [hepwww.ph.man.ac.uk/~roger/book.html](http://hepwww.ph.man.ac.uk/~roger/book.html)

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. Amsler et al. (Particle Data Group), *Review of Particle Physics*, *Physics Letters B* 667 (2008) 1; see also [pdg.lbl.gov](http://pdg.lbl.gov) sections on probability statistics, Monte Carlo

## Some Bayesian references

P. Gregory, *Bayesian Logical Data Analysis for the Physical Sciences*, CUP, 2005

D. Sivia, *Data Analysis: a Bayesian Tutorial*, OUP, 2006

S. Press, *Subjective and Objective Bayesian Statistics: Principles, Models and Applications*, 2nd ed., Wiley, 2003

A. O'Hagan, Kendall's, *Advanced Theory of Statistics, Vol. 2B, Bayesian Inference*, Arnold Publishers, 1994

A. Gelman et al., *Bayesian Data Analysis*, 2nd ed., CRC, 2004

W. Bolstad, *Introduction to Bayesian Statistics*, Wiley, 2004

E.T. Jaynes, *Probability Theory: the Logic of Science*, CUP, 2003