# Statistical Methods in Particle Physics
# Lecture 1: Bayesian methods


SUSSP65

St Andrews

16–29 August 2009

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

# Outline

Lecture #1: An introduction to Bayesian statistical methods

Role of probability in data analysis (Frequentist, Bayesian)

A simple fitting problem : Frequentist vs. Bayesian solution

Bayesian computation, Markov Chain Monte Carlo

Lecture #2: Setting limits, making a discovery

Frequentist vs Bayesian approach,

treatment of systematic uncertainties

Lecture #3: Multivariate methods for HEP

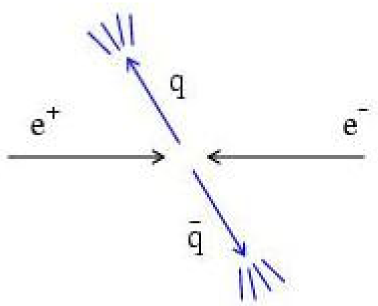
Event selection as a statistical test

Neyman-Pearson lemma and likelihood ratio test

Some multivariate classifiers:

NN, BDT, SVM, ...

# 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?)

# Dealing with uncertainty

In particle physics there are various elements of uncertainty:

theory is not deterministic

quantum mechanics

random measurement errors

present even without quantum effects

things we could know in principle but don't

e.g. from limitations of cost, time, ...

We can quantify the uncertainty using PROBABILITY

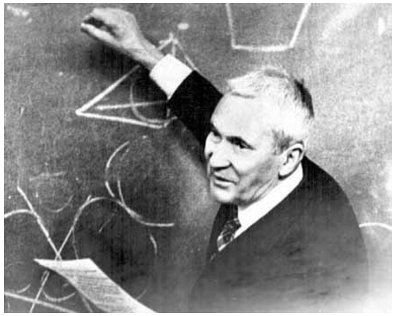
# A definition of probability

Consider a set $S$ with subsets $A$, $B$, ...

$$\text{For all } A \subset S, P(A) \geq 0$$

$$P(S) = 1$$

$$\text{If } A \cap B = \emptyset, P(A \cup B) = P(A) + P(B)$$



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*, ... are outcomes of a repeatable experiment

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

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

## II. Subjective probability

*A*, *B*, ... 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

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

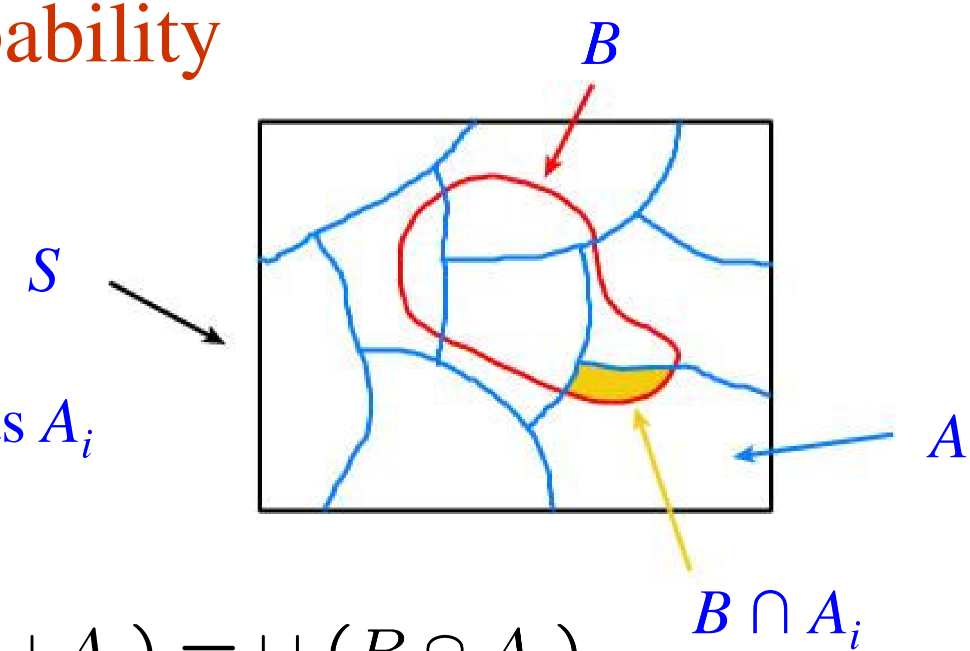
Bayes' theorem

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 $\cup_i A_i = S$,

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

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

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

Bayes' theorem becomes

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

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

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'.

# 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.)

# Statistical vs. systematic errors

## Statistical errors:

How much would the result fluctuate upon repetition of the measurement?

Implies some set of assumptions to define probability of outcome of the measurement.

## Systematic errors:

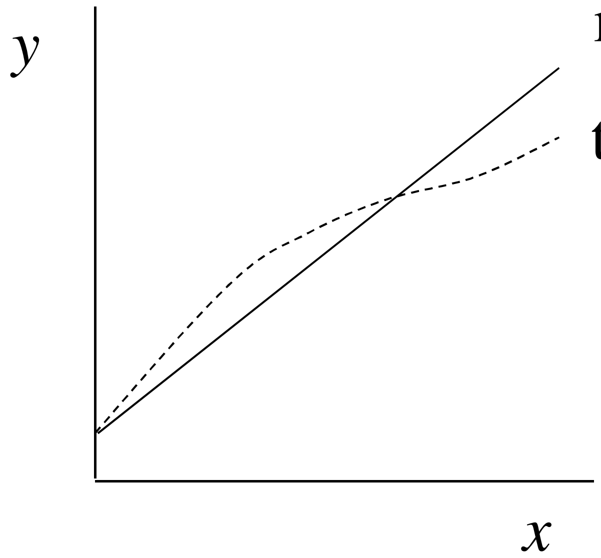
What is the uncertainty in my result due to uncertainty in my assumptions, e.g.,

model (theoretical) uncertainty;
modeling of measurement apparatus.

Usually taken to mean the sources of error do not vary upon repetition of the measurement. Often result from uncertain value of calibration constants, efficiencies, etc.

# Systematic errors and nuisance parameters

Model prediction (including e.g. detector effects) never same as "true prediction" of the theory:




model: $y = \alpha + \beta x$

truth: $y = \alpha + \beta x + \gamma x^2 + \varepsilon x^3 + \ldots$

Model can be made to approximate better the truth by including more free parameters.

systematic uncertainty $\leftrightarrow$ nuisance parameters

# Example: fitting a straight line

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

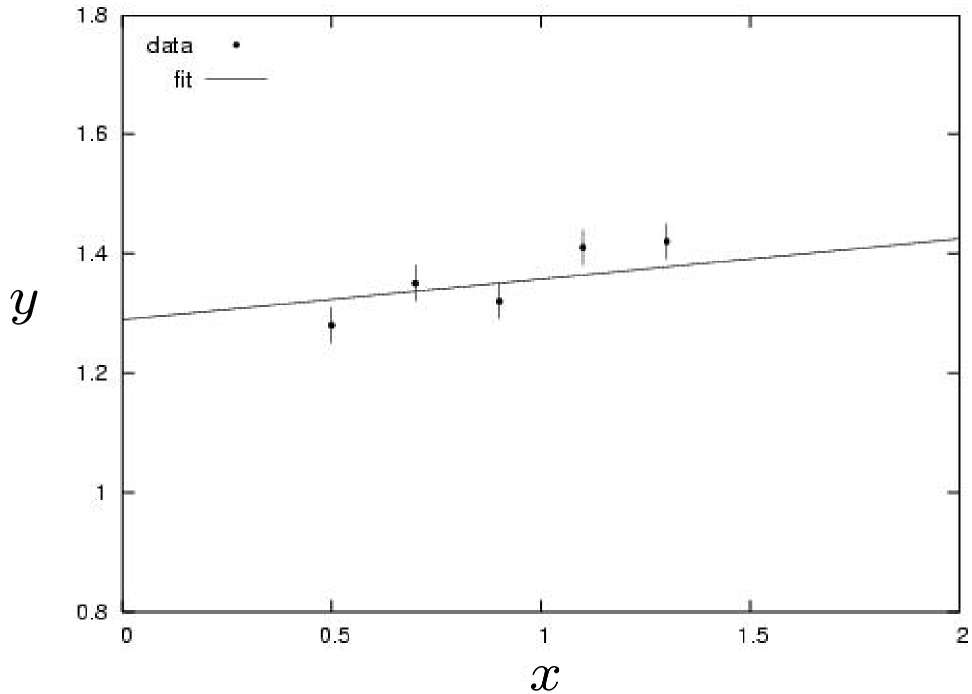
Model: measured $y_i$ independent, Gaussian: $y_i \sim N(\mu(x_i), \sigma_i^2)$

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

assume $x_i$ and $\sigma_i$ known.

Goal: estimate $\theta_0$

(don't care about $\theta_1$).



# Frequentist approach

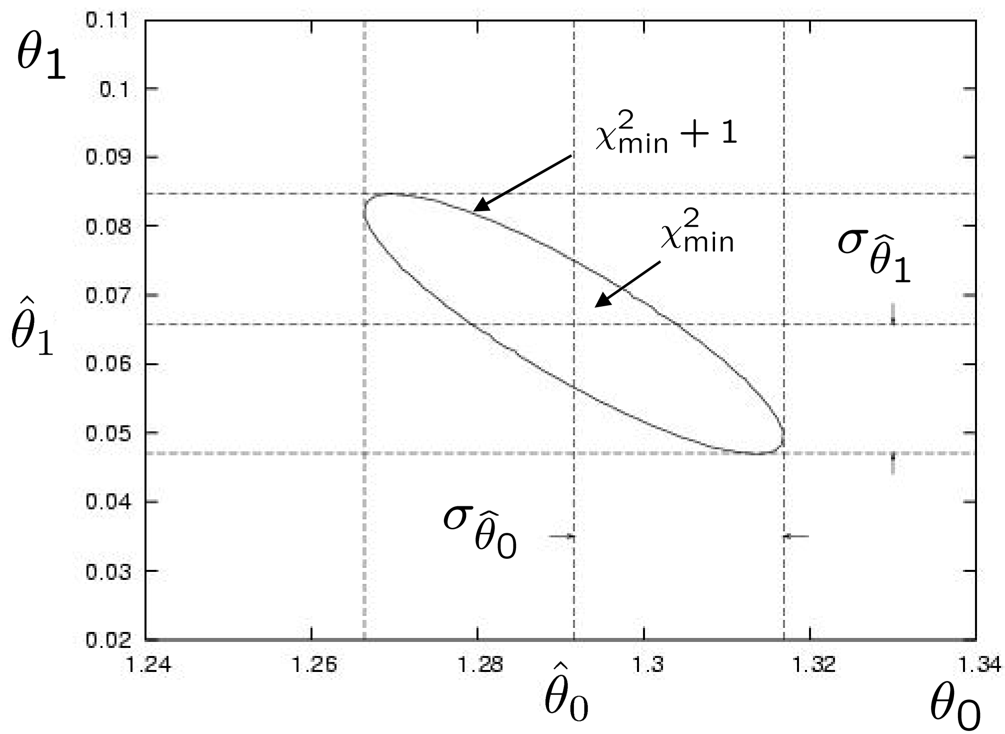
$$L(\theta_0,\theta_1) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\frac{(y_i-\mu(x_i;\theta_0,\theta_1))^2}{\sigma_i^2}\right] ,$$

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

Standard deviations from tangent lines to contour

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

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

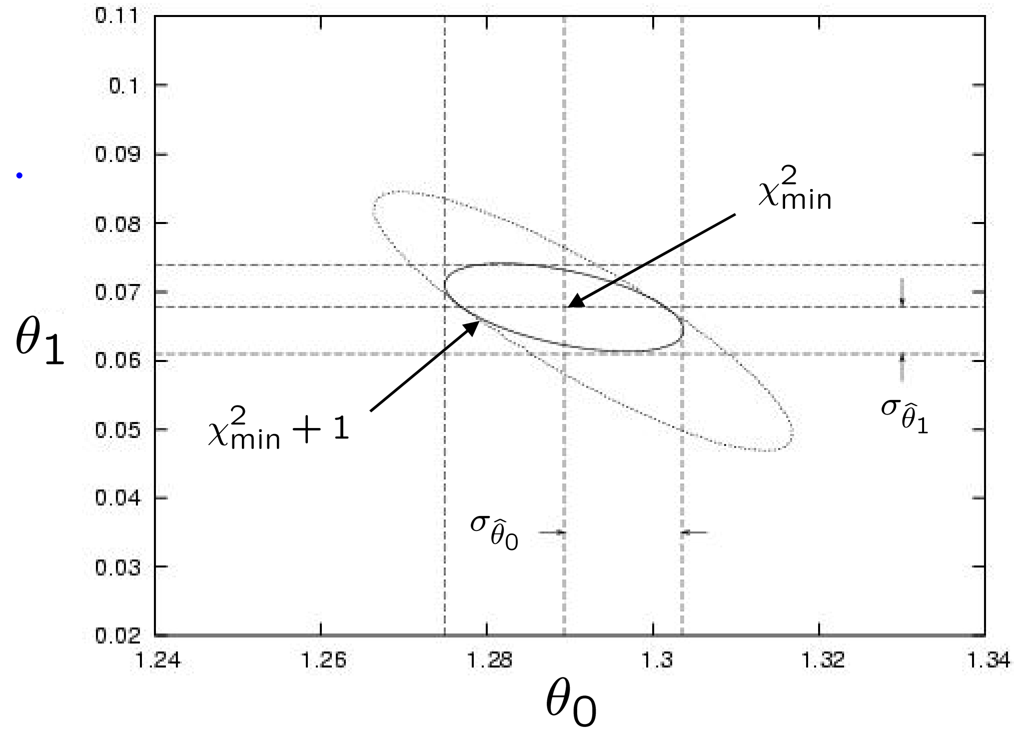


# Frequentist case with a measurement $t_1$ of $\theta_1$

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

The information on $\theta_1$

improves accuracy of $\hat{\theta}_0$ .



# Bayesian method

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

$$\pi(\theta_0, \theta_1) = \pi_0(\theta_0)\,\pi_1(\theta_1)$$

$$\pi_0(\theta_0) = \text{const.}$$

reflects 'prior ignorance', in any case much broader than $L(\theta_0)$

$$\pi_1(\theta_1) = \frac{1}{\sqrt{2\pi}\sigma_{t_1}} e^{-(\theta_1 - t_1)^2/2\sigma_{t_1}^2}$$

← 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) $p(\theta_0, \theta_1 \mid x)$ to find $p(\theta_0 \mid x)$:

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

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

$$\begin{aligned} p(\theta_0|x) &= \frac{1}{\sqrt{2\pi}\sigma_{\theta_0}} e^{-(\theta_0 - \hat{\theta}_0)^2/2\sigma_{\theta_0}^2} \quad \text{with} \\ \hat{\theta}_0 &= \text{same as ML estimator} \\ \sigma_{\theta_0} &= \sigma_{\hat{\theta}_0} \; \text{(same as before)} \end{aligned}$$

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.

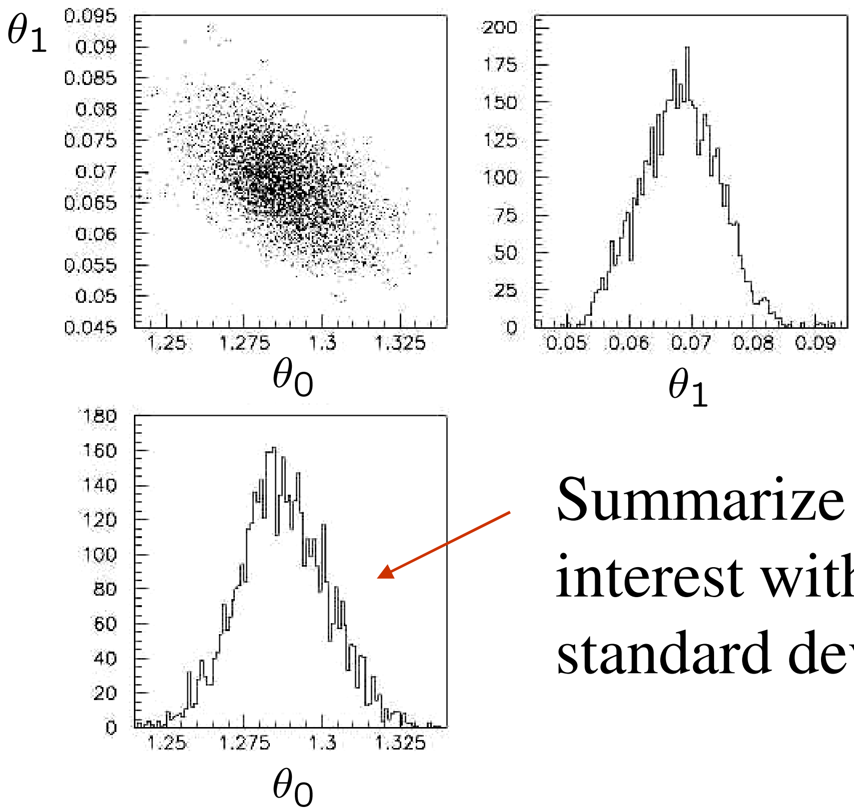
MCMC (e.g., Metropolis-Hastings algorithm) generates correlated sequence of random numbers:

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

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 $\vec{\theta}_1, \vec{\theta}_2, \vec{\theta}_3, \ldots$

Proposal density $q(\vec{\theta}; \vec{\theta}_0)$
e.g. Gaussian centred
about $\vec{\theta}_0$

1) Start at some point $\vec{\theta}_0$

2) Generate $\vec{\theta} \sim q(\vec{\theta}; \vec{\theta}_0)$

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 $u \sim \mathrm{Uniform}[0,1]$

5) If $u \leq \alpha$, $\vec{\theta}_1 = \vec{\theta}$ , $\longleftarrow$ move to proposed point

else $\vec{\theta}_1 = \vec{\theta}_0$ $\longleftarrow$ 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 it would be with uncorrelated points.

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.

# Metropolis-Hastings caveats

Actually one can only prove that the sequence of points follows the desired pdf in the limit where it runs forever.

There may be a “burn-in” period where the sequence does not initially follow $p(\vec{\theta})$ .

Unfortunately there are few useful theorems to tell us when the sequence has converged.

Look at trace plots, autocorrelation.

Check result with different proposal density.

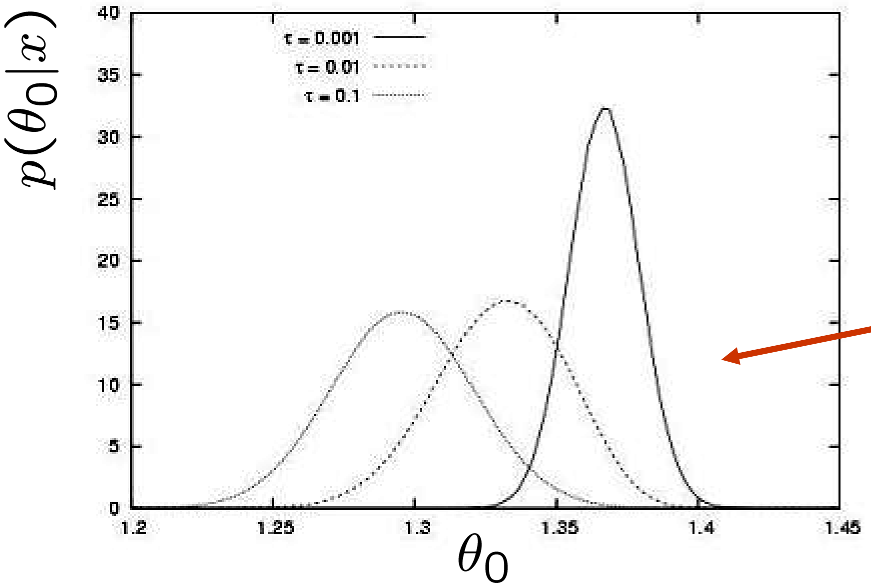
If you think it’s converged, try starting from a different point and see if the result is similar.

# 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: $y_i \pm \sigma_i^{\text{stat}} \pm \sigma_i^{\text{sys}}, \quad i = 1, \ldots, n$ ,

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 parameters bias

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

# Its Bayesian equivalent

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

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

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

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 $\boldsymbol{b}$:

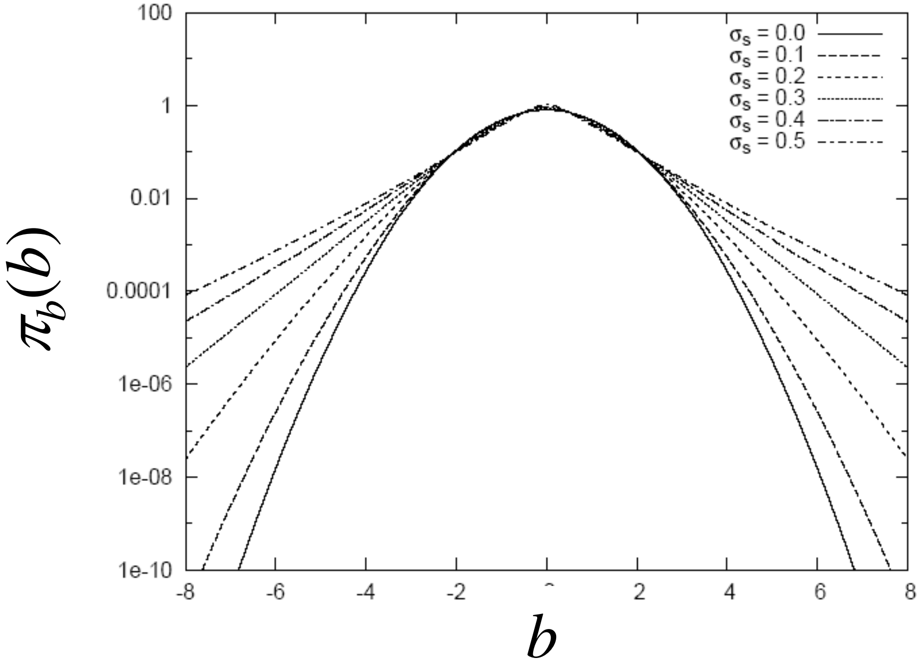
$$p(\theta|\vec{y}) = \int p(\theta,\vec{b}|\vec{y})\,d\vec{b}$$

→ Posterior is Gaussian with mode same as least squares estimator, $\sigma_\theta$ same as from $\chi^2 = \chi^2_{\mathrm{min}} + 1$. (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:

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



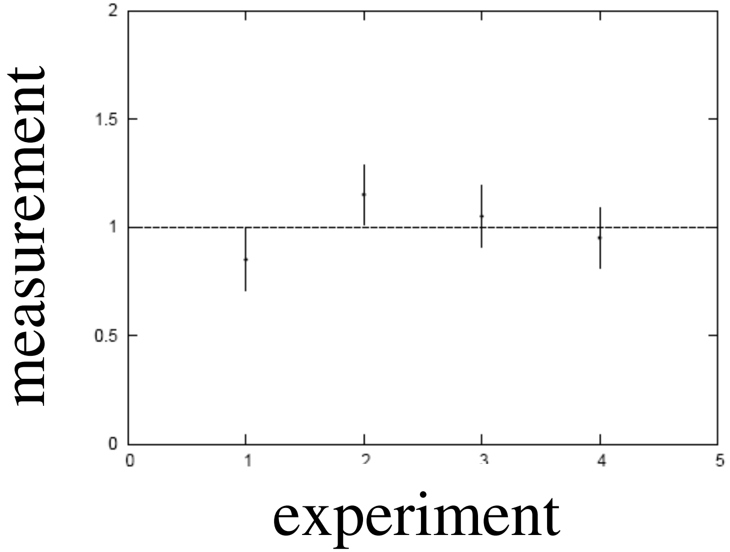

Represents 'error on the error'; standard deviation of $\pi_s(s)$ is $\sigma_s$.

# A simple test

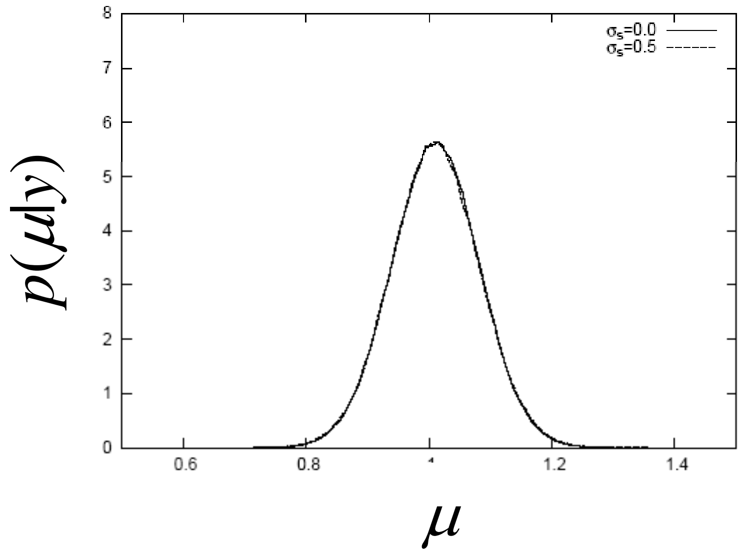
Suppose fit effectively averages four measurements.

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

Case #1: data appear compatible




Posterior $p(\mu|y)$:



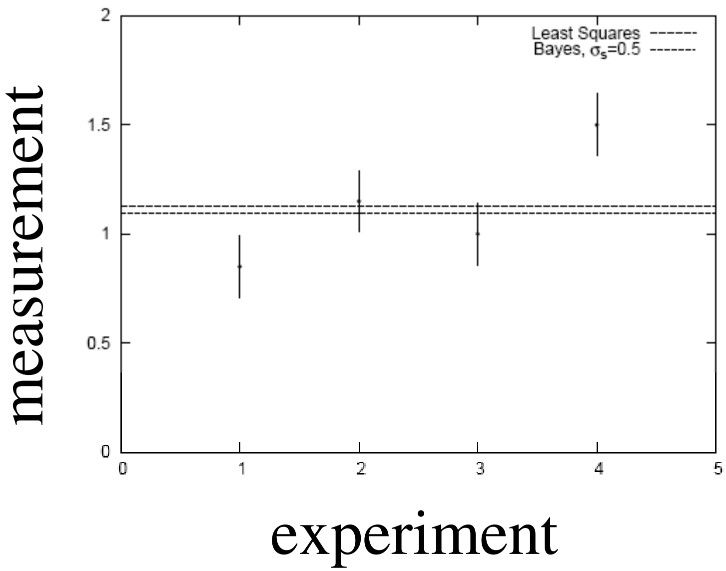

Usually summarize posterior $p(\mu|y)$ with mode and standard deviation:

$$\sigma_s = 0.0: \quad \hat{\mu} = 1.000 \pm 0.071$$

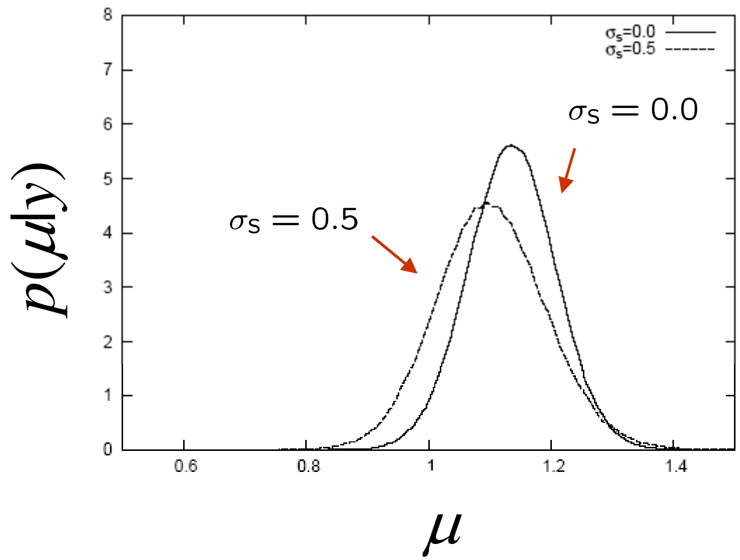
$$\sigma_s = 0.5: \quad \hat{\mu} = 1.000 \pm 0.072$$

# Simple test with inconsistent data

Case #2: there is an outlier




Posterior $p(\mu|y)$:




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

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

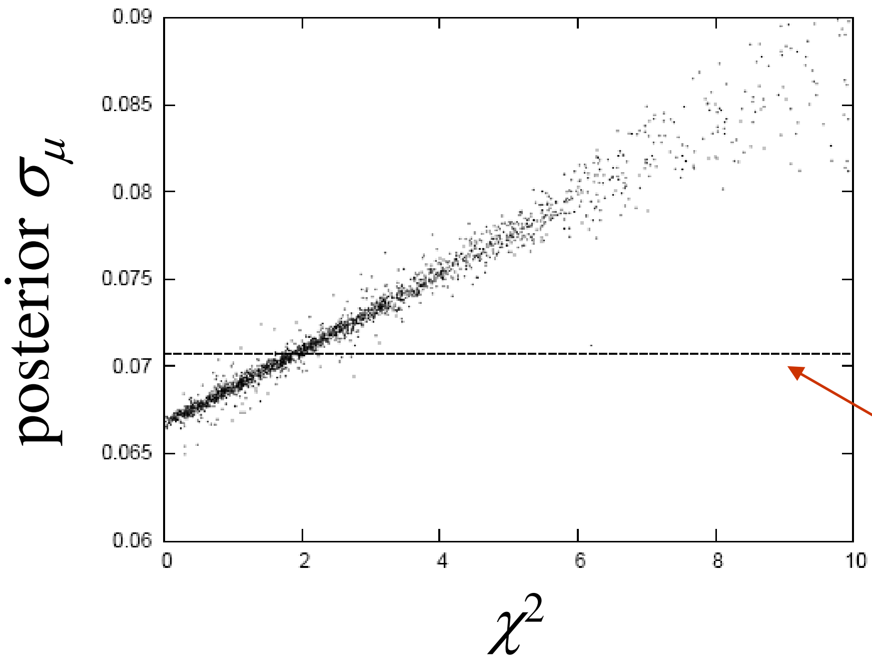
→ Bayesian fit less sensitive to outlier.

(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).




2000 repetitions of experiment, $\sigma_s = 0.5$, here no actual bias.

$\sigma_\mu$ from least squares

# Summary of lecture 1

The distinctive features of Bayesian statistics are:

Subjective probability used for hypotheses (e.g. a parameter).

Bayes' theorem relates the probability of data given *H* (the likelihood) to the posterior probability of *H* given data:

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

Requires prior probability for *H*

Bayesian methods often yield answers that are close (or identical) to those of frequentist statistics, albeit with different interpretation.

This is not the case when the prior information is important relative to that contained in the data.

# Extra slides

# 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