

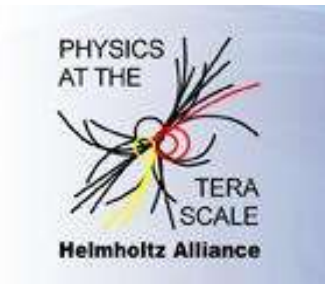
Frequently Bayesian

The role of probability in data analysis

Terascale Statistics School

DESY, Hamburg

30 September, 2008



Glen Cowan

Physics Department

Royal Holloway, University of London

g.cowan@rhul.ac.uk

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



Outline

Tuesday:

The Bayesian method

Bayesian assessment of uncertainties

Bayesian computation: MCMC

Wednesday:

Bayesian limits

Bayesian model selection ("discovery")

Outlook for Bayesian methods in HEP

Statistical data analysis at the terascale

High stakes



"5 sigma"

"4 sigma"



and expensive experiments, so we should make sure the data analysis doesn't waste information.

Specific challenges for LHC analyses include

Huge data volume

Generally cannot trust MC prediction of backgrounds; need to use data (control samples, sidebands...)

Lots of theory uncertainties, e.g., parton densities

People looking in many places ("look-elsewhere effect")

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**

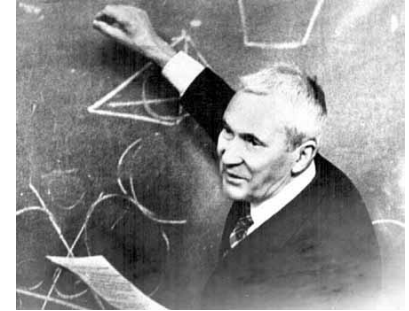
A definition of probability

Consider a set S with subsets A, B, \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

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

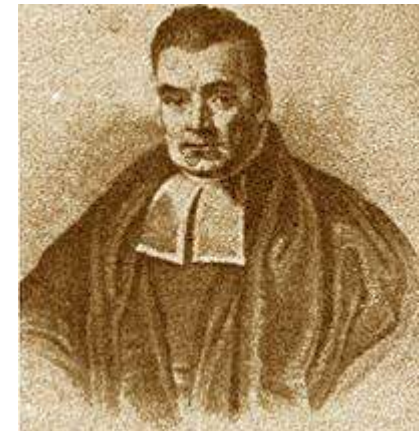
but $P(A \cap B) = P(B \cap A)$, so

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

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.

Bayes' theorem



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:

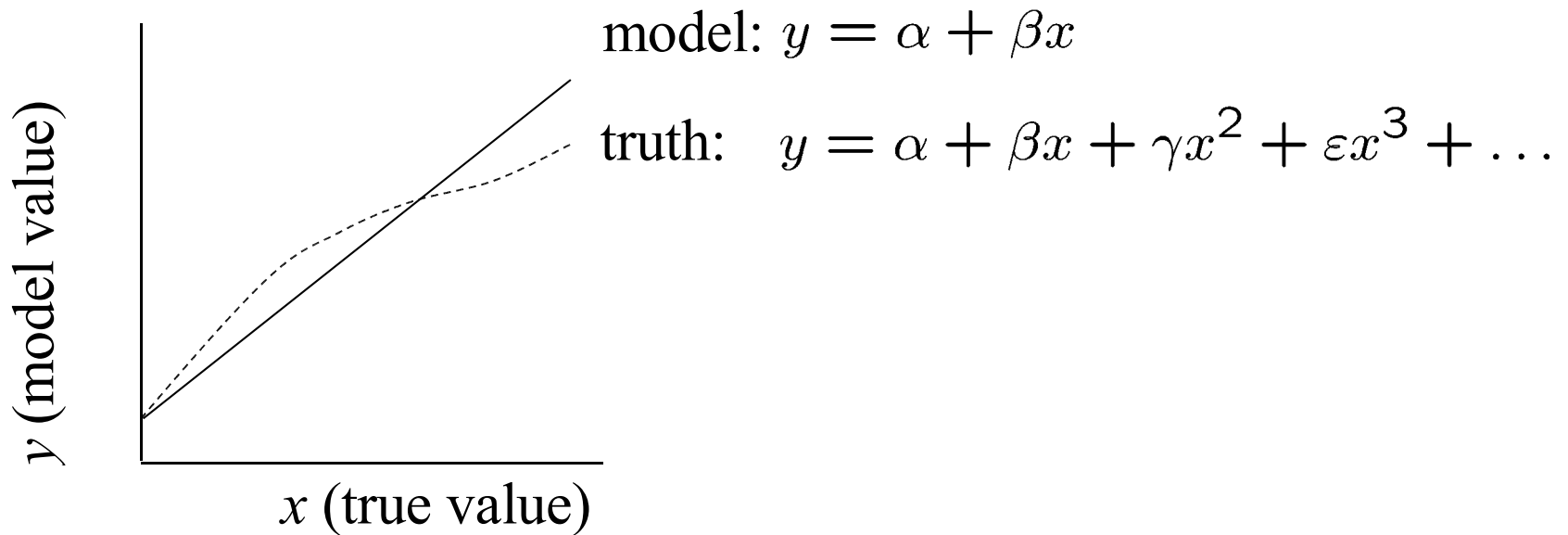
What is the uncertainty in my result due to uncertainty in my assumptions, e.g.,

model (theoretical) uncertainty;
modelling 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, e.g., 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 can be made to approximate better the truth by including more free parameters.

systematic uncertainty \leftrightarrow nuisance parameters

Example: fitting a straight line

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

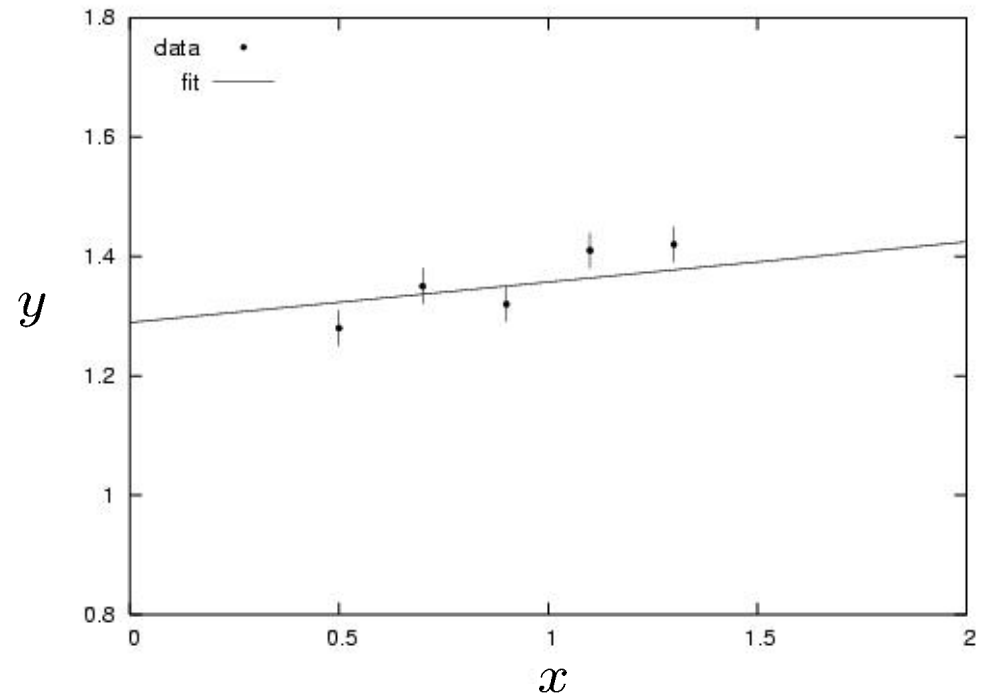
Model: measured y_i independent, Gaussian: $y_i \sim N(\mu(x_i), \sigma_i^2)$

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

assume x_i and σ_i known.

Goal: estimate θ_0

(don't care about θ_1).



Frequentist approach

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

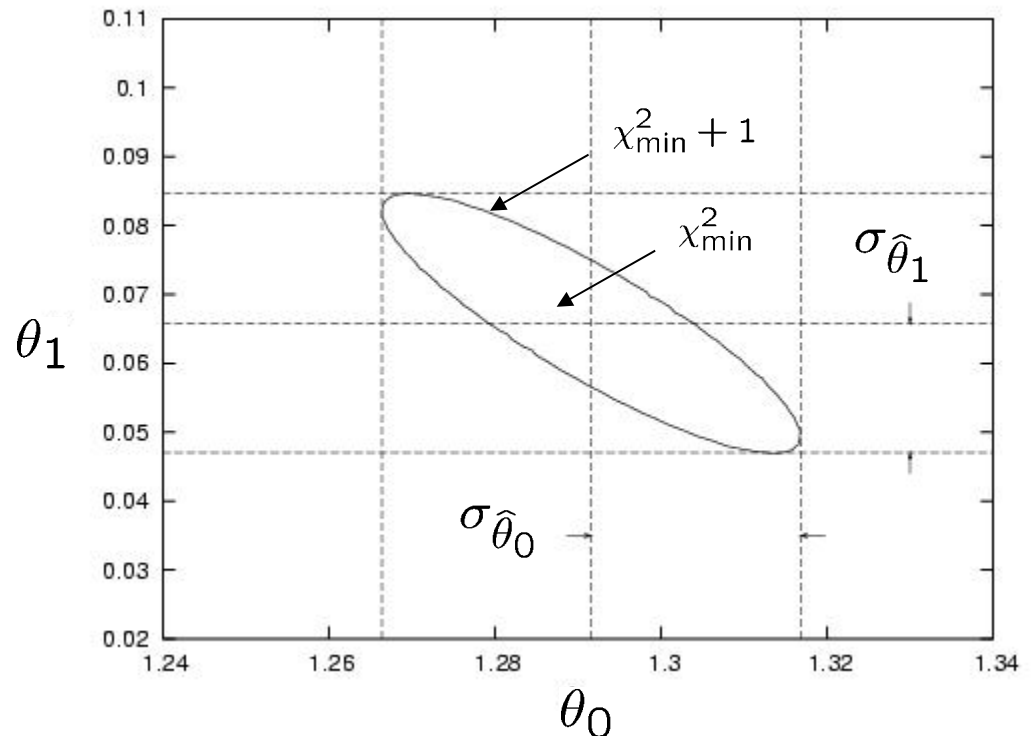
$$\chi^2(\theta_0, \theta_1) = -2 \ln L(\theta_0, \theta_1) + \text{const} = \sum_{i=1}^n \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2}.$$

Standard deviations from
tangent lines to contour

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

Correlation between

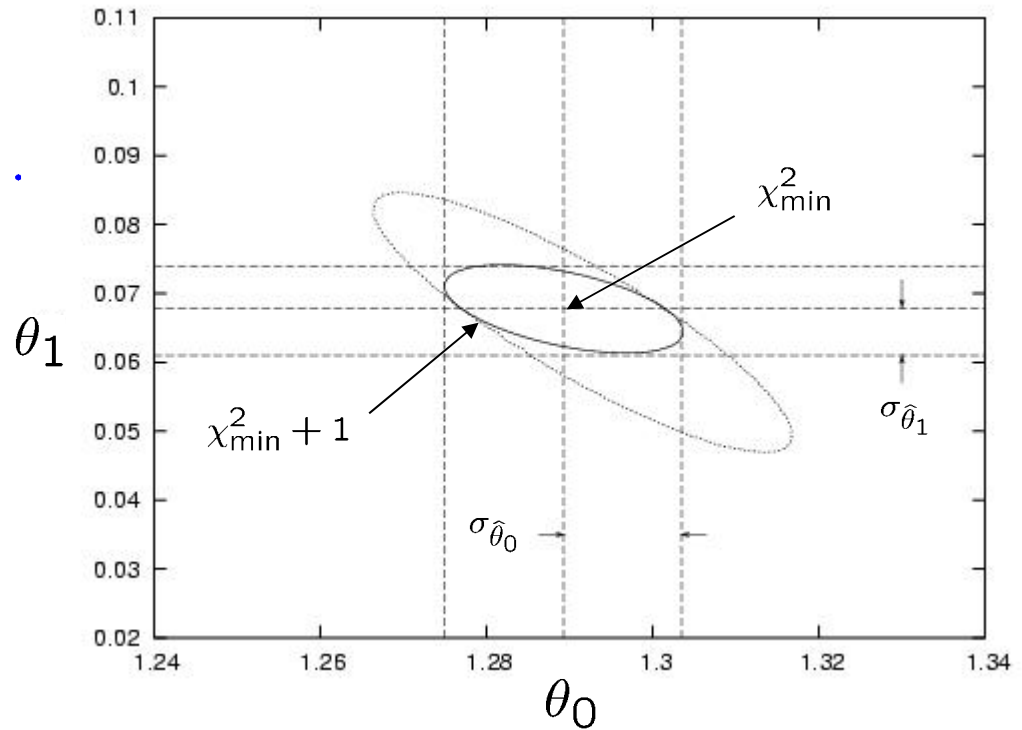
$\hat{\theta}_0, \hat{\theta}_1$ causes errors
to increase.



Frequentist case with a measurement t_1 of θ_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 θ_1
improves accuracy of $\hat{\theta}_0$.



Bayesian method

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

$$\pi(\theta_0, \theta_1) = \pi_0(\theta_0) \pi_1(\theta_1) \quad \text{reflects 'prior ignorance', in any}$$

$$\pi_0(\theta_0) = \text{const.} \quad \leftarrow \text{case much broader than } L(\theta_0)$$

$$\pi_1(\theta_1) = \frac{1}{\sqrt{2\pi}\sigma_{t_1}} e^{-(\theta_1 - t_1)^2 / 2\sigma_{t_1}^2} \quad \leftarrow \text{based on previous measurement}$$

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 \ominus

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

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

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

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

Ability to marginalize over nuisance parameters is an important feature of Bayesian statistics.

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.

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.

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.

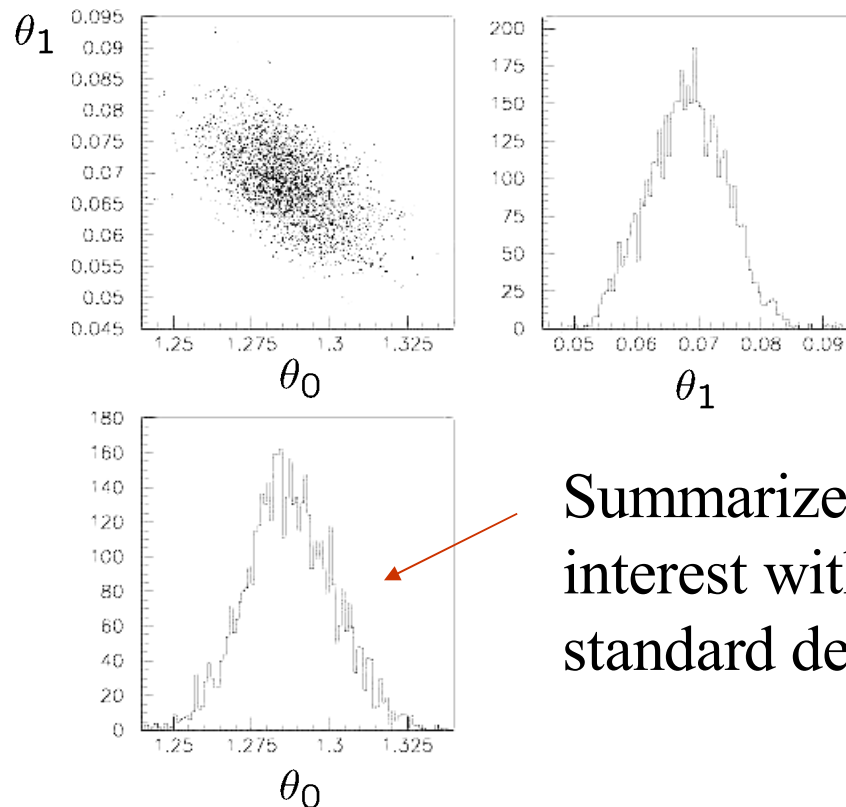
Look at trace plots, autocorrelation.

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.

Example: posterior pdf from MCMC

Sample the posterior pdf from previous example with MCMC:



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

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

Bayesian method with vague prior

Suppose we don't have a previous measurement of θ_1 but rather some vague information, e.g., a theorist tells us:

$\theta_1 \geq 0$ (essentially certain);

θ_1 should have order of magnitude less than 0.1 'or so'.

Under pressure, the theorist sketches the following prior:

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

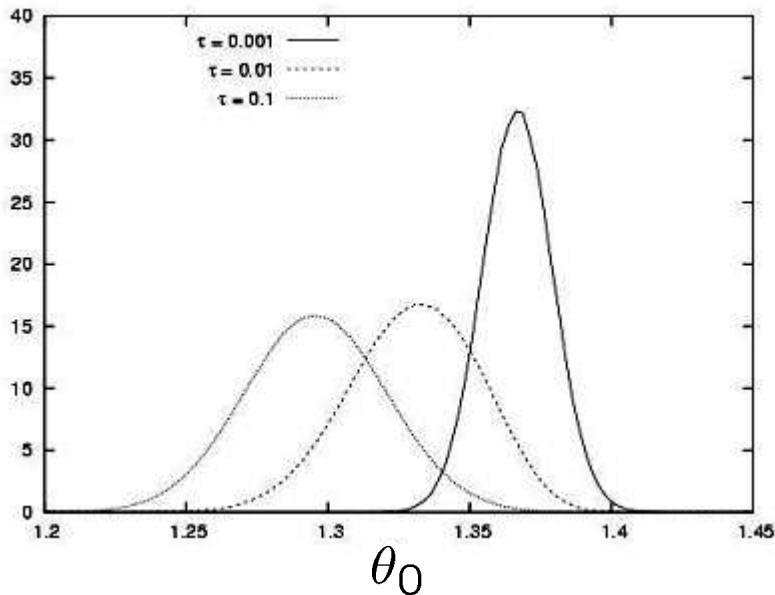
From this we will obtain posterior probabilities for θ_0 (next slide).

We do not need to get the theorist to 'commit' to this prior; final result has 'if-then' character.

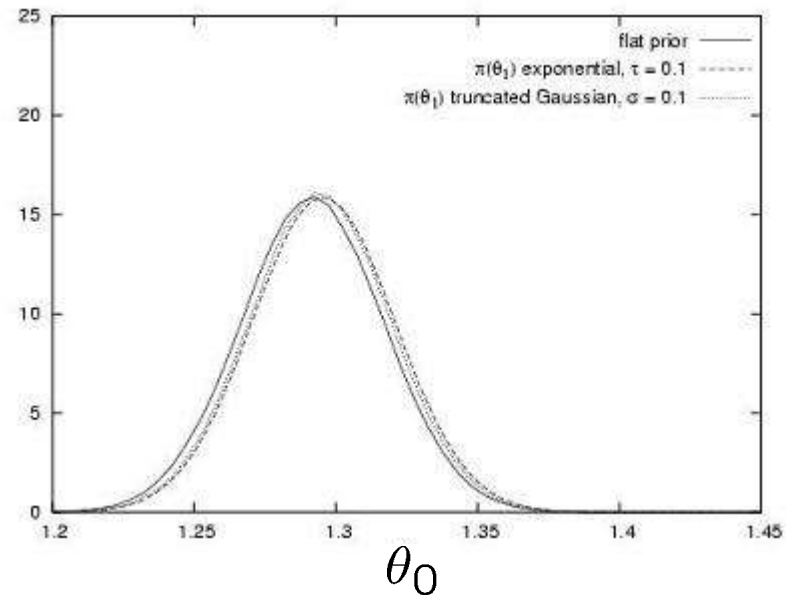
Sensitivity to prior

Vary $\pi(\theta)$ to explore how extreme your prior beliefs would have to be to justify various conclusions (sensitivity analysis).

Try exponential with different mean values...



Try different functional forms...



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 θ , integrate (marginalize) over 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,
 σ_θ same as from $\chi^2 = \chi^2_{\text{min}} + 1$. (Back where we started!)

The error on the error

Some systematic errors are well determined

Error from finite Monte Carlo sample

Some are less obvious

Do analysis in n ‘equally valid’ ways and extract systematic error from ‘spread’ in results.

Some are educated guesses

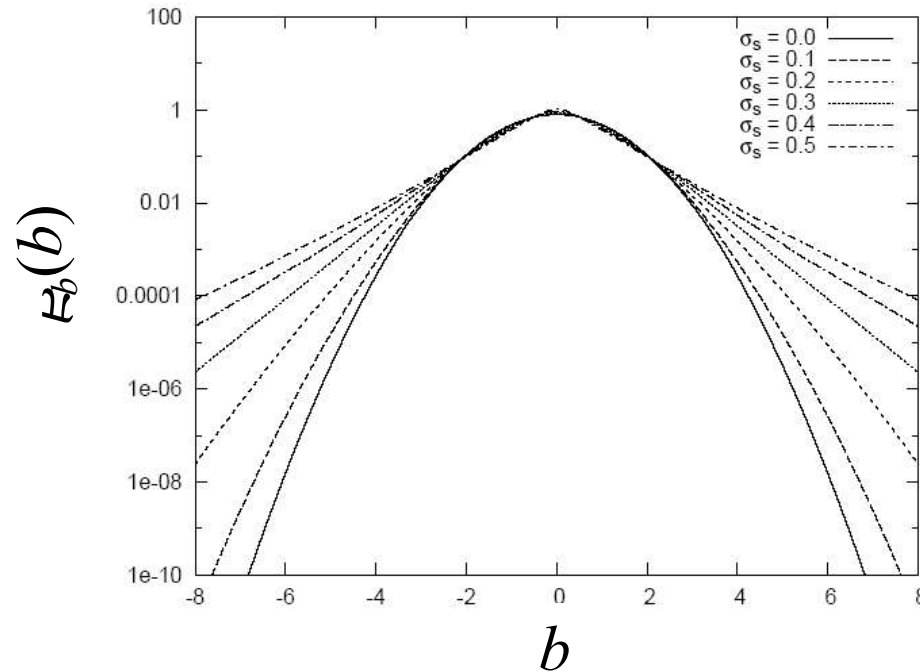
Guess possible size of missing terms in perturbation series;
vary renormalization scale $(\mu/2 < Q < 2\mu ?)$

Can we incorporate the ‘error on the error’?

(cf. G. D’Agostini 1999; Dose & von der Linden 1999)

A prior for bias $\pi_b(b)$ 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 σ_s .

Gaussian ($\sigma_s = 0$) $P(|b| > 4\sigma_{\text{sys}}) = 6.3 \times 10^{-5}$

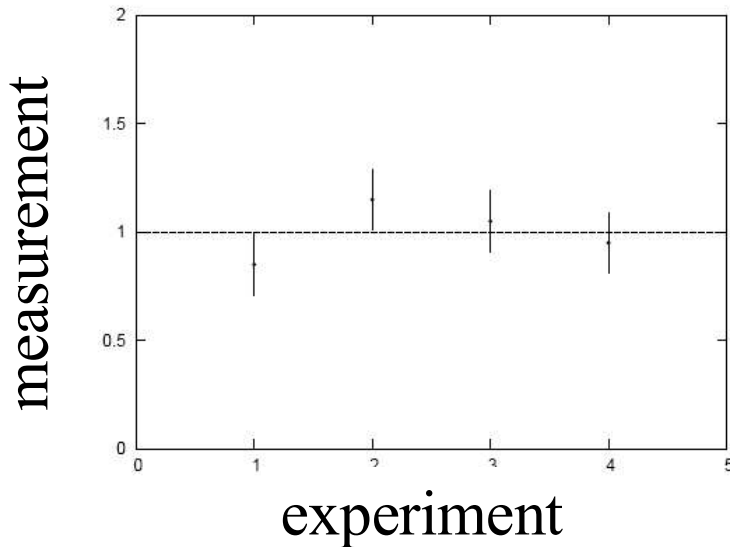
$\sigma_s = 0.5$ $P(|b| > 4\sigma_{\text{sys}}) = 6.5 \times 10^{-3}$

A simple test

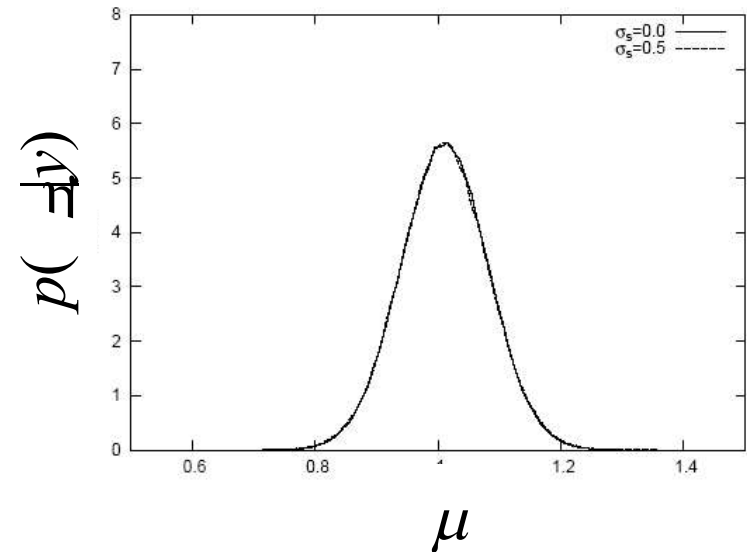
Suppose fit effectively averages four measurements.

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

Case #1: data appear compatible



Posterior $p(\mu|y)$:



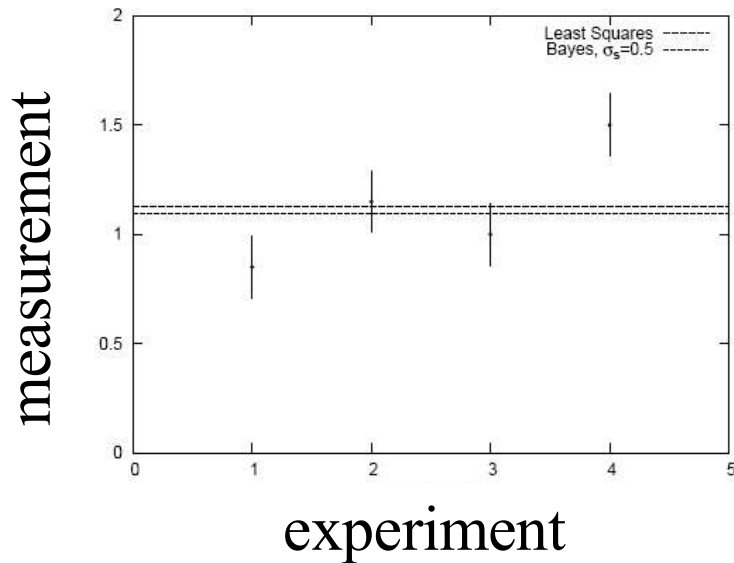
Usually summarize posterior $p(\mu|y)$
with mode and standard deviation:

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

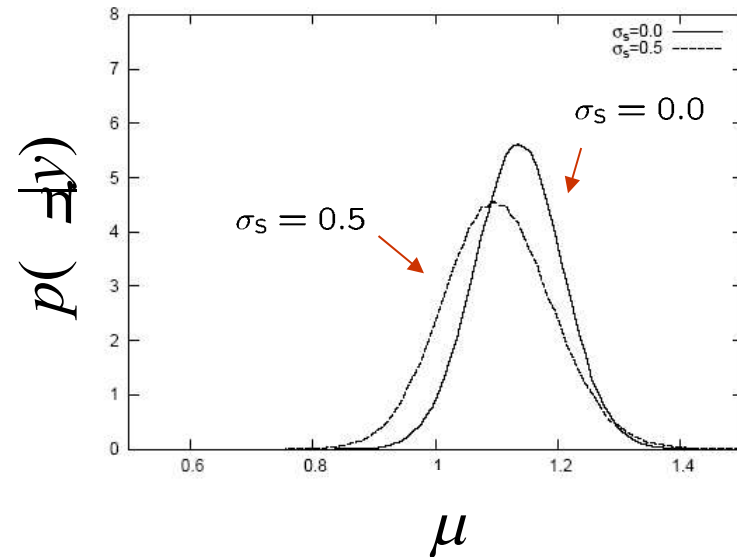
$$\sigma_s = 0.5 : \quad \hat{\mu} = 1.000 \pm 0.072$$

Simple test with inconsistent data

Case #2: there is an outlier



Posterior $p(\mu|y)$:



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

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

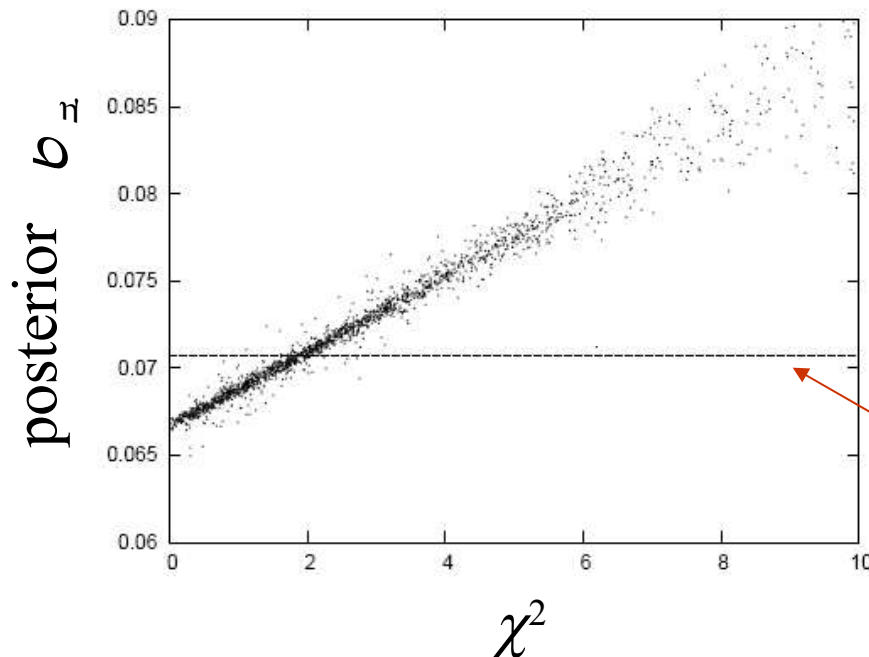
→ Bayesian fit less sensitive to outlier.

→ Error now connected to goodness-of-fit.

Goodness-of-fit vs. size of error

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

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



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

σ_μ 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

Uncertainty from parametrization of PDFs

Try e.g. $xf(x) = ax^b(1-x)^c(1+d\sqrt{x}+ex)$ (MRST)

or $xf(x) = ax^b(1-x)^ce^{d\cdot x}(1+e^ex)^f$ (CTEQ)

The form should be flexible enough to describe the data;
frequentist analysis has to decide how many parameters are justified.

In a Bayesian analysis we can insert as many parameters as we want, but constrain them with priors.

Suppose e.g. based on a theoretical bias for things not too bumpy,
that a certain parametrization ‘should hold to 2%’.

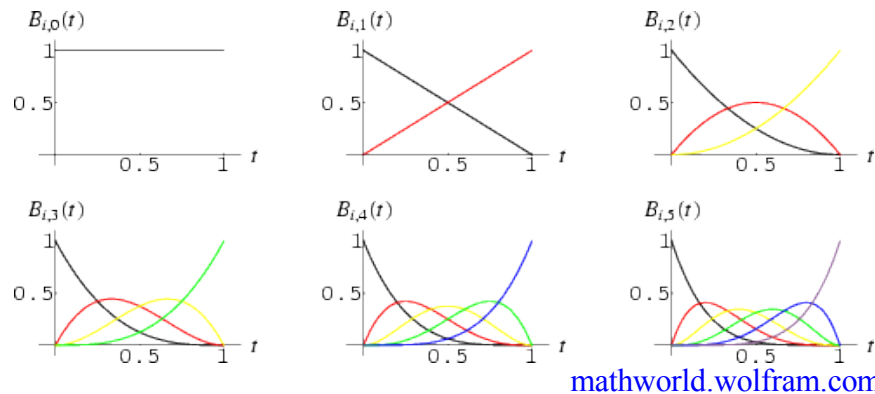
How to translate this into a set of prior probabilities?

Residual function

Try e.g. $xf(x) = ax^b(1-x)^c(1+\dots) + r(x)$ ← ‘residual function’

where $r(x)$ is something very flexible, e.g., superposition of

Bernstein polynomials, coefficients ν_i : $r(x) = \sum_i \nu_i B_i(x)$



$$B_{i,n} = \binom{n}{i} x^i (1-x)^{n-i}$$

Assign priors for the ν_i centred around 0, width chosen to reflect the uncertainty in $xf(x)$ (e.g. a couple of percent).

→ Ongoing effort.