

# Topics in Statistical Data Analysis for HEP

## Lecture 1: Bayesian Methods



CERN-JINR European School  
of High Energy Physics

Bautzen, 14–27 June 2009

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: 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

Setting limits / making a discovery

## Lecture #2: Multivariate methods for HEP

Event selection as a statistical test

Neyman-Pearson lemma and likelihood ratio test

Some multivariate classifiers:

Boosted Decision Trees

Support Vector Machines

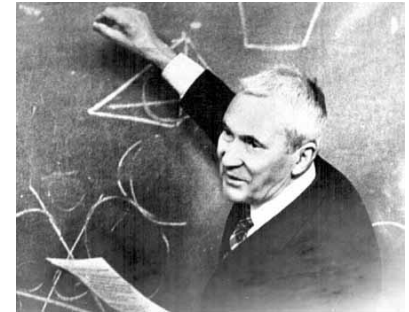
# A definition of probability

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

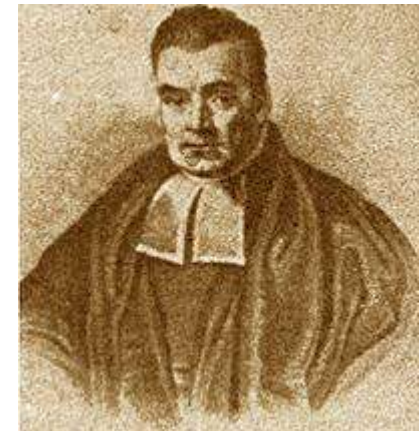
From the definition of conditional probability we have

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

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'.

# 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.)

# Example: fitting a straight line

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

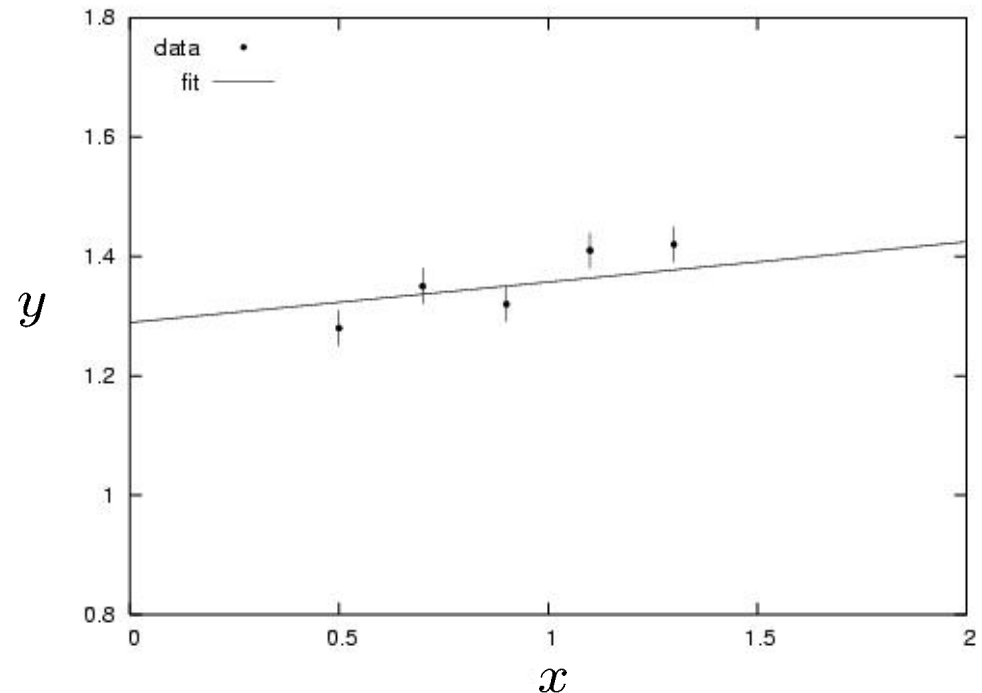
Model: measured  $y_i$  independent, Gaussian:  $y_i \sim N(\mu(x_i), \sigma_i^2)$

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

assume  $x_i$  and  $\sigma_i$  known.

Goal: estimate  $\theta_0$

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





# Frequentist approach

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

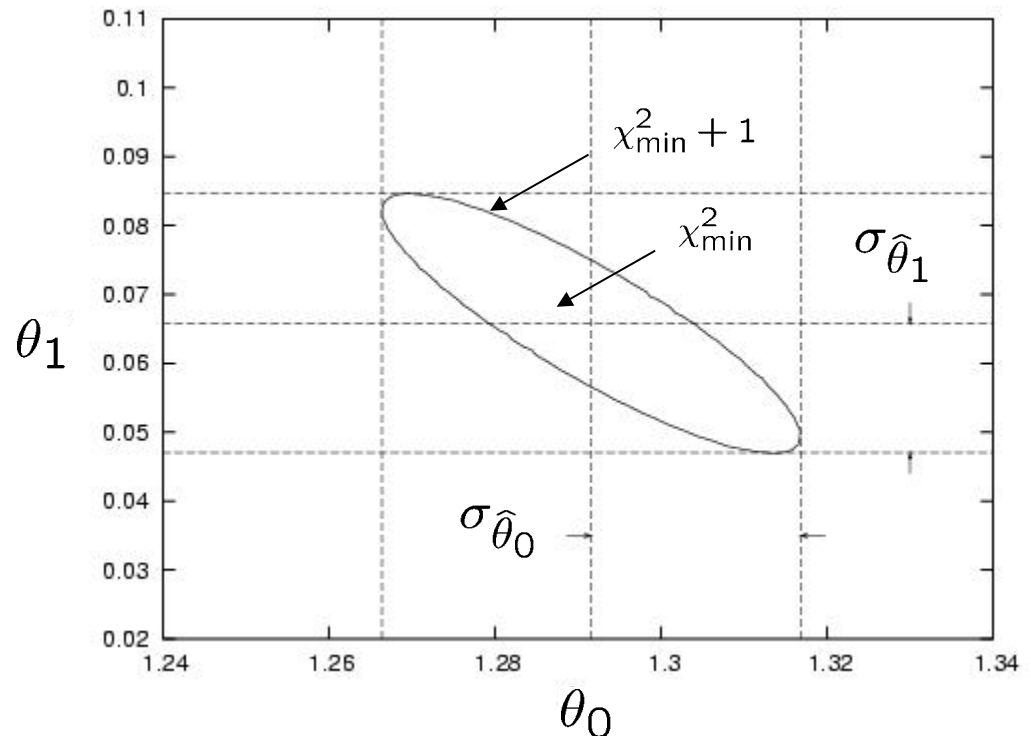
$$\chi^2(\theta_0, \theta_1) = -2 \ln L(\theta_0, \theta_1) + \text{const} = \sum_{i=1}^n \frac{(y_i - \mu(x_i; \theta_0, \theta_1))^2}{\sigma_i^2}.$$

Standard deviations from  
tangent lines to contour

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

Correlation between

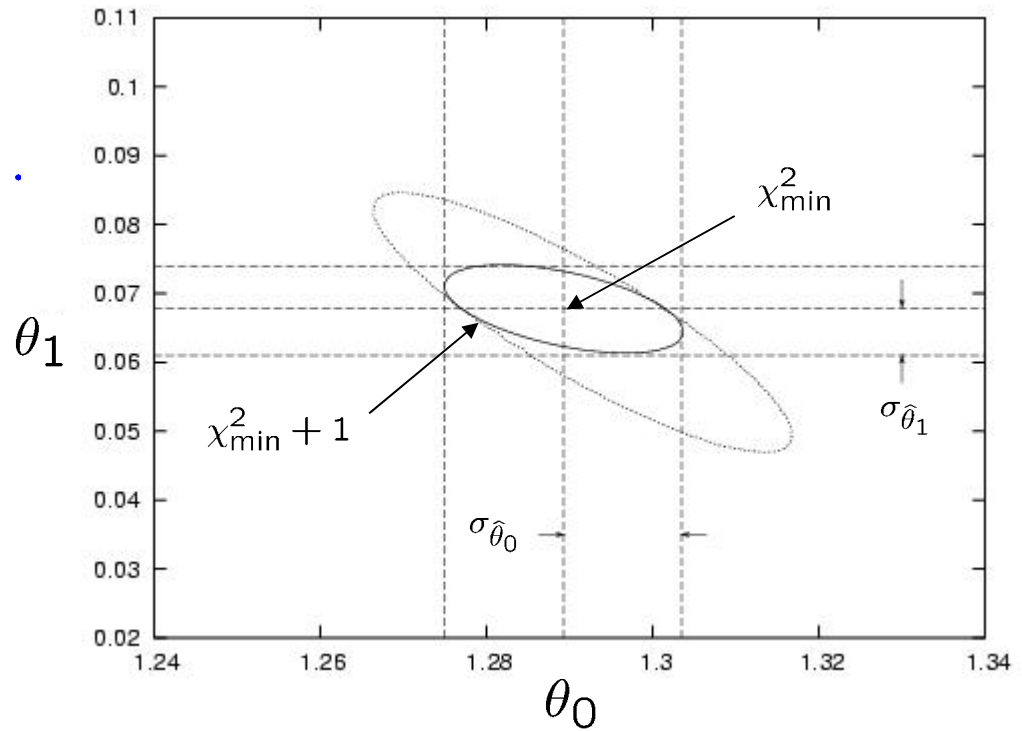
$\hat{\theta}_0, \hat{\theta}_1$  causes errors  
to increase.



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

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

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.,

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

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

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

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  $\propto$  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.

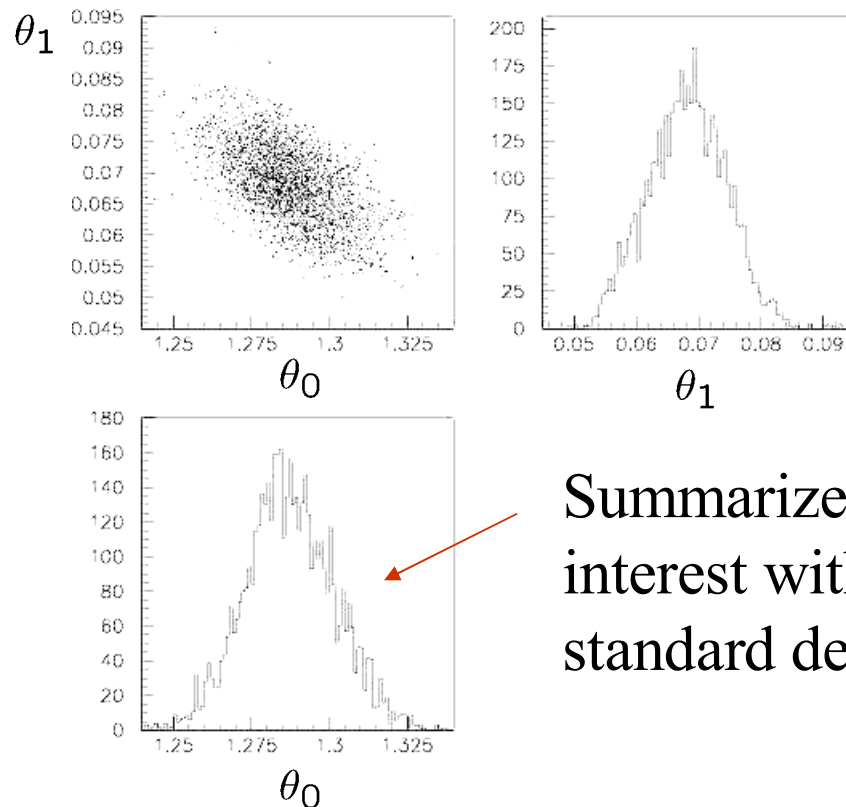
MCMC (e.g., Metropolis-Hastings algorithm) generates  
**correlated** sequence of random numbers:

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.



# 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.

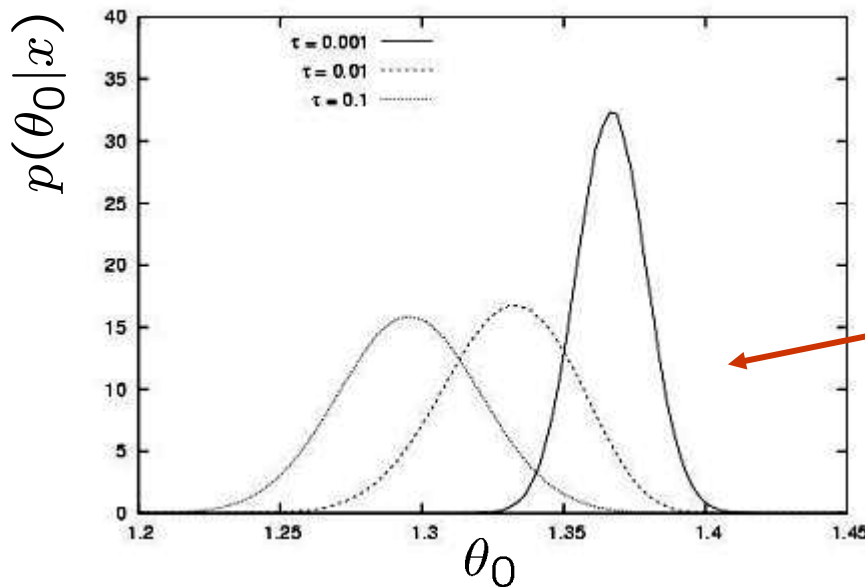
If you think it's converged, try starting from a different point and see if the result is similar.

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

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  $\mathbf{b}$ :

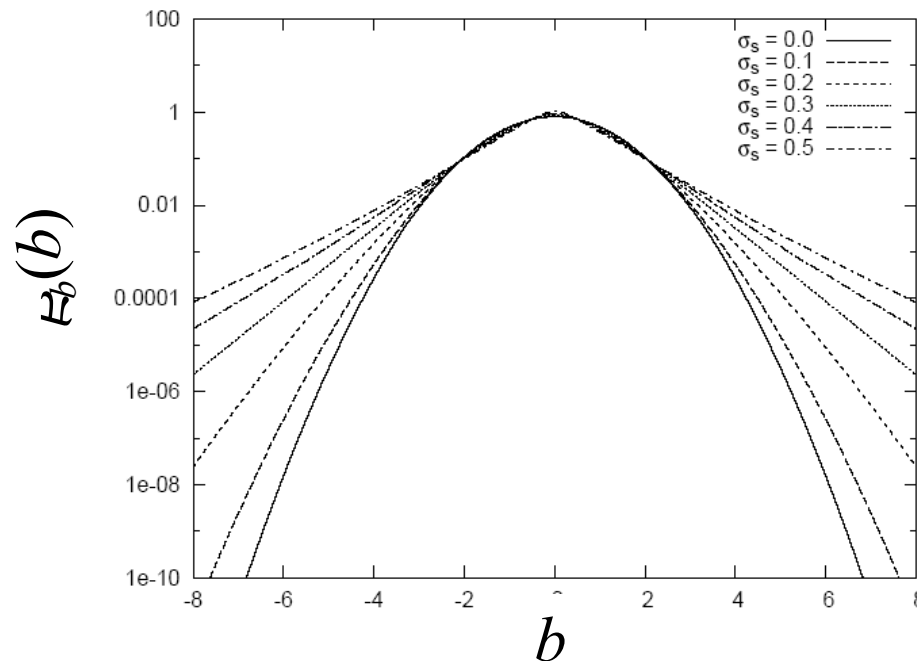
$$p(\theta|\vec{y}) = \int p(\theta, \vec{b}|\vec{y}) d\vec{b}$$

→ 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:

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



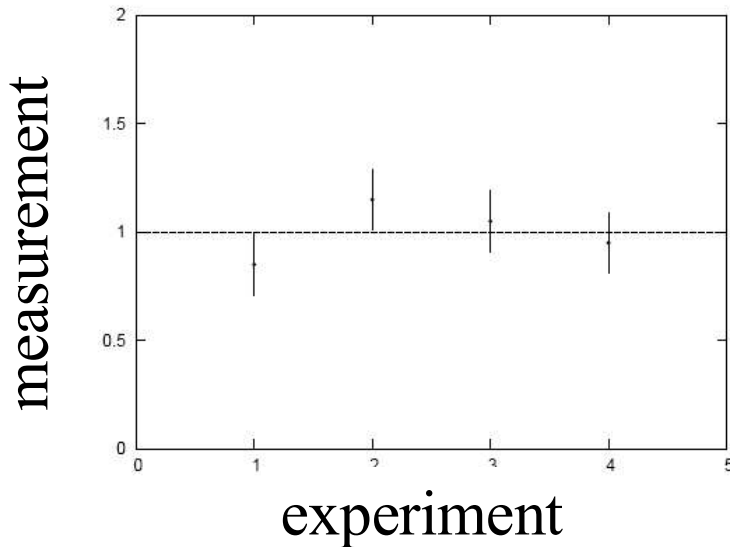
Represents 'error on the error'; standard deviation of  $\pi_s(s)$  is  $\sigma_s$ .

# A simple test

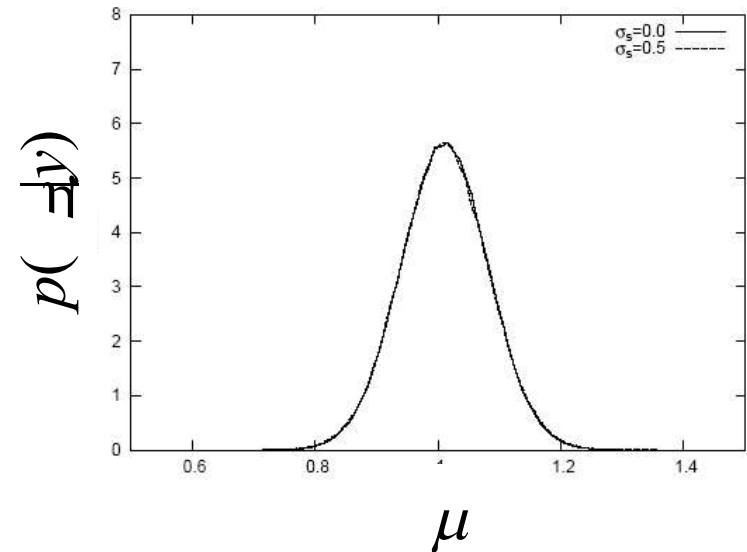
Suppose fit effectively averages four measurements.

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:

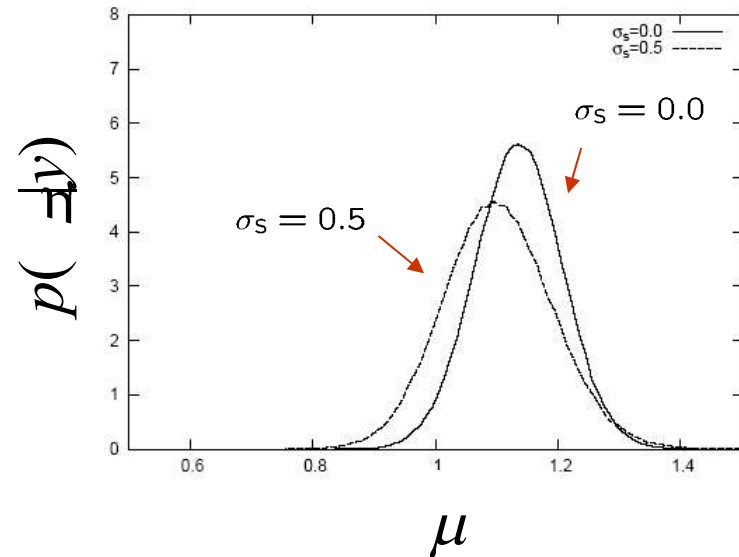
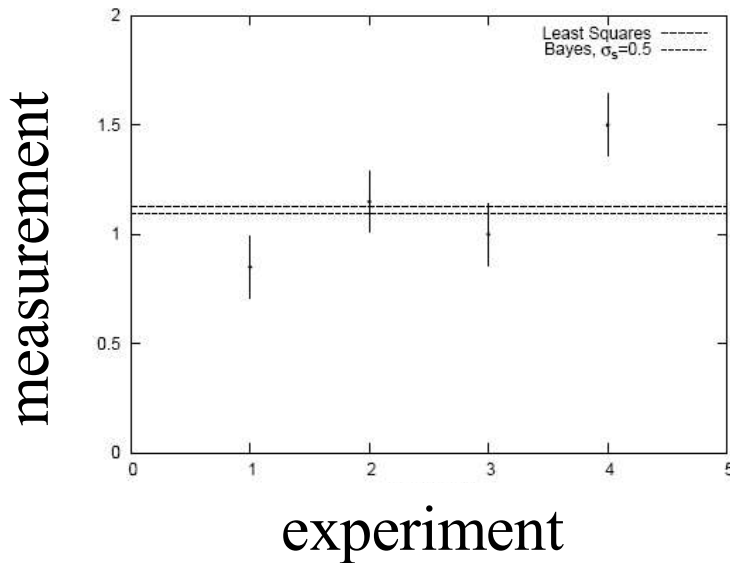
$$\sigma_s = 0.0 : \quad \hat{\mu} = 1.000 \pm 0.071$$

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

# Simple test with inconsistent data

Case #2: there is an outlier

Posterior  $p(\mu|y)$ :



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

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

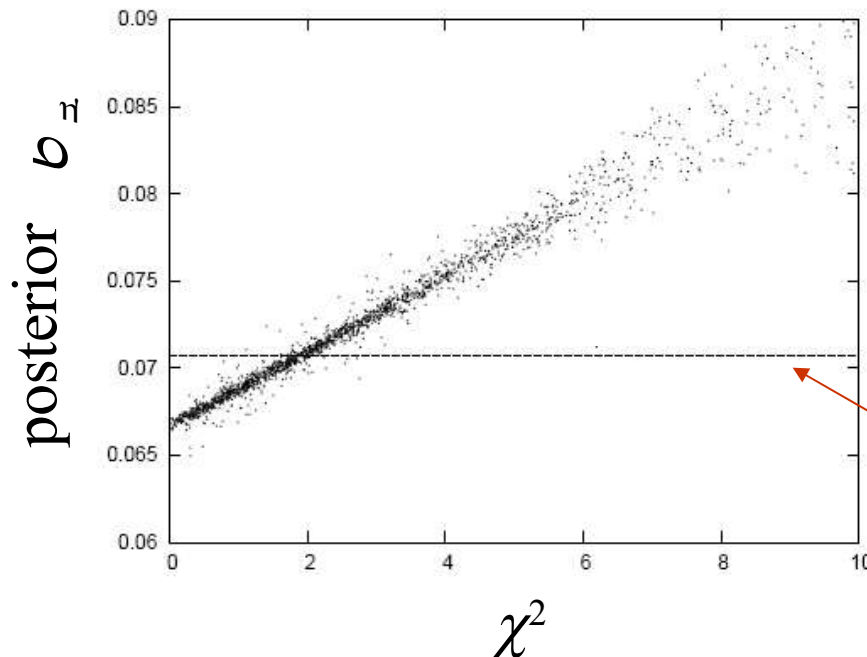
→ Bayesian fit less sensitive to outlier.

(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



# The Bayesian approach to limits

In Bayesian statistics need to start with ‘prior pdf’  $\pi(\theta)$ , this reflects degree of belief about  $\theta$  before doing the experiment.

Bayes’ theorem tells how our beliefs should be updated in light of the data  $x$ :

$$p(\theta|x) = \frac{L(x|\theta)\pi(\theta)}{\int L(x|\theta')\pi(\theta') d\theta'} \propto L(x|\theta)\pi(\theta)$$

Integrate posterior pdf  $p(\theta | x)$  to give interval with any desired probability content.

For e.g. Poisson parameter 95% CL upper limit from

$$0.95 = \int_{-\infty}^{\text{sup}} p(s|n) ds$$

# Analytic formulae for limits

There are a number of papers describing Bayesian limits for a variety of standard scenarios

Several conventional priors

Systematics on efficiency, background

Combination of channels

and (semi-)analytic formulae and software are provided.

Joel Heinrich, *Bayesian limit software: multi-channel with correlated backgrounds and efficiencies*, CDF/MEMO/STATISTICS/PUBLIC/7587 (2005).

Joel Heinrich et al., *Interval estimation in the presence of nuisance parameters. 1. Bayesian approach*, CDF/MEMO/STATISTICS/PUBLIC/7117, physics/0409129 (2004).

Luc Demortier, *A Fully Bayesian Computation of Upper Limits for Poisson Processes*, CDF/MEMO/STATISTICS/PUBLIC/5928 (2004).

But for more general cases we need to use numerical methods (e.g. L.D. uses importance sampling).

# Example: Poisson data with background

Count  $n$  events, e.g., in fixed time or integrated luminosity.

$s$  = expected number of signal events

$b$  = expected number of background events

$$n \sim \text{Poisson}(s+b): \quad P(n; s, b) = \frac{(s + b)^n}{n!} e^{-(s+b)}$$

Sometimes  $b$  known, other times it is in some way uncertain.

Goal: measure or place limits on  $s$ , taking into consideration the uncertainty in  $b$ .

Widely discussed in HEP community, see e.g. proceedings of PHYSTAT meetings, Durham, Fermilab, CERN workshops...

# Bayesian prior for Poisson parameter

Include knowledge that  $s \geq 0$  by setting prior  $\pi(s) = 0$  for  $s < 0$ .

Often try to reflect ‘prior ignorance’ with e.g.

$$\pi(s) = \begin{cases} 1 & s \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Not normalized but this is OK as long as  $L(s)$  dies off for large  $s$ .

Not invariant under change of parameter — if we had used instead a flat prior for, say, the mass of the Higgs boson, this would imply a non-flat prior for the expected number of Higgs events.

Doesn't really reflect a reasonable degree of belief, but often used as a point of reference;

or viewed as a recipe for producing an interval whose frequentist properties can be studied (coverage will depend on true  $s$ ).

# Bayesian interval with flat prior for $s$

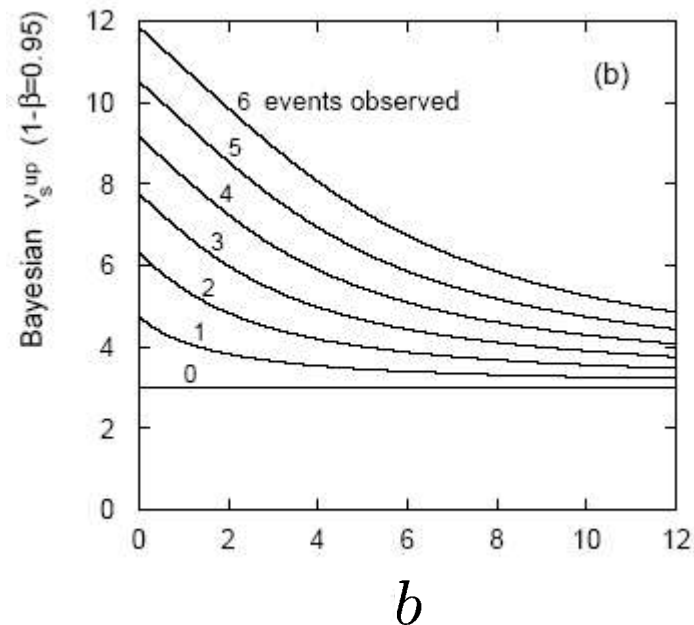
Solve numerically to find limit  $s_{\text{up}}$ .

For special case  $b = 0$ , Bayesian upper limit with flat prior numerically same as classical case ('coincidence').

Otherwise Bayesian limit is everywhere greater than classical ('conservative').

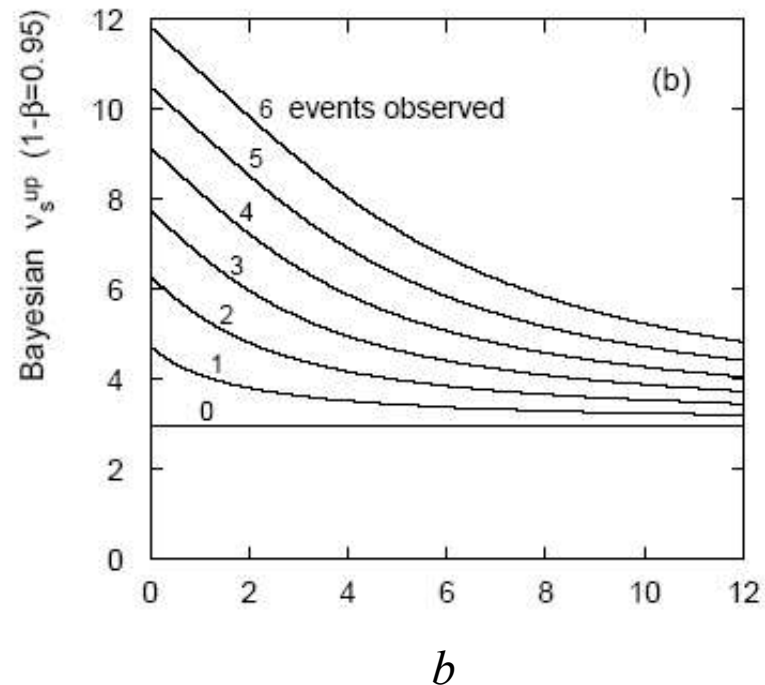
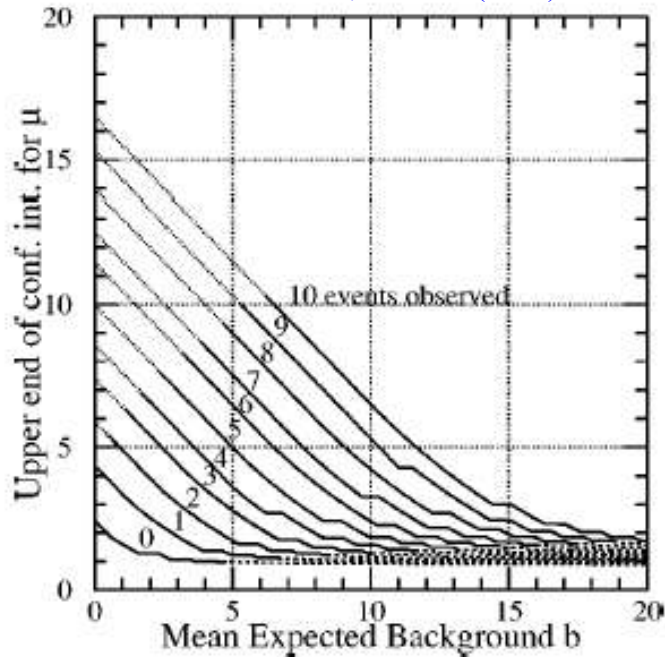
Never goes negative.

Doesn't depend on  $b$  if  $n = 0$ .



# Upper limit versus $b$

Feldman & Cousins, PRD 57 (1998) 3873



If  $n = 0$  observed, should upper limit depend on  $b$ ?

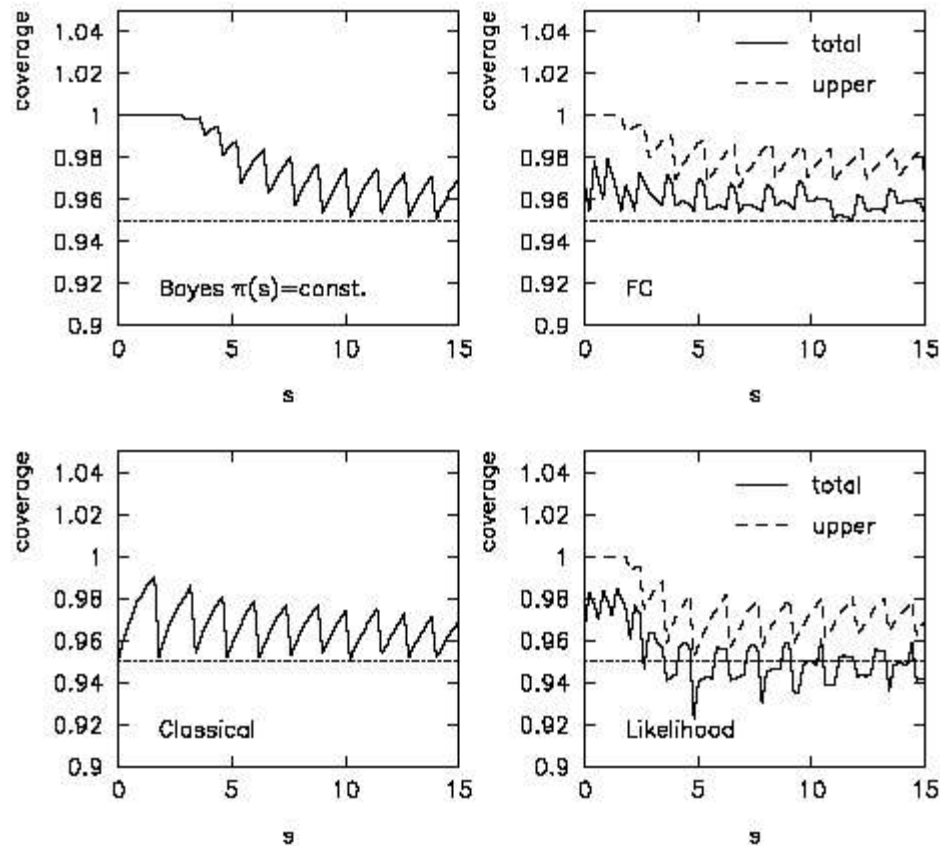
Classical: yes

Bayesian: no

FC: yes

# Coverage probability of confidence intervals

Because of discreteness of Poisson data, probability for interval to include true value in general  $>$  confidence level ('over-coverage')



# Bayesian limits with uncertainty on $b$

Uncertainty on  $b$  goes into the prior, e.g.,

$$\pi(s, b) = \pi_s(s)\pi_b(b) \quad (\text{or include correlations as appropriate})$$

$$\pi_s(s) = \text{const}, \quad \sim 1/s, \dots$$

$$\pi_b(b) = \frac{1}{\sqrt{2\pi}\sigma_b} e^{-(b-b_{\text{meas}})^2/2\sigma_b^2} \quad (\text{or whatever})$$

Put this into Bayes' theorem,

$$p(s, b|n) \propto L(n|s, b)\pi(s, b)$$

Marginalize over  $b$ , then use  $p(s|n)$  to find intervals for  $s$  with any desired probability content.

Controversial part here is prior for signal  $\pi_s(s)$   
(treatment of nuisance parameters is easy).



# Discussion on limits

Different sorts of limits answer different questions.

A frequentist confidence interval does not (necessarily) answer, “What do we believe the parameter’s value is?”

Coverage — nice, but crucial?

Look at sensitivity, e.g.,  $E[s_{\text{up}} | s = 0]$ .

Consider also:

politics, need for consensus/conventions;  
convenience and ability to combine results, ...

For any result, consumer will compute (mentally or otherwise):

$$p(\theta|\text{result}) \propto L(\text{result}|\theta)\pi(\theta)$$

Need likelihood (or summary thereof).

consumer’s prior



# Frequentist discovery, $p$ -values

To discover e.g. the Higgs, try to reject the background-only (null) hypothesis ( $H_0$ ).

Define a statistic  $t$  whose value reflects compatibility of data with  $H_0$ .

$p$ -value = Prob(data with  $\leq$  compatibility with  $H_0$  when compared to the data we got |  $H_0$  )

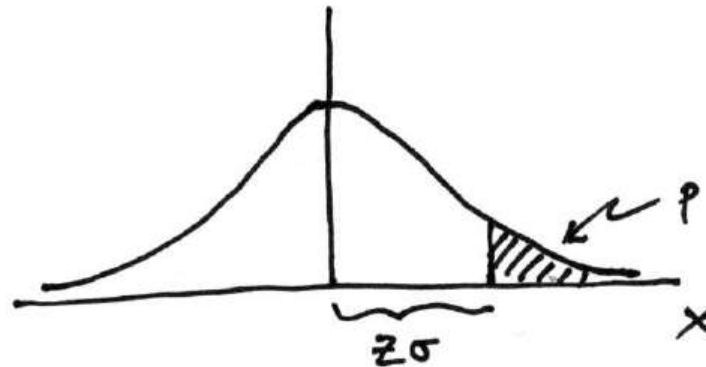
For example, if high values of  $t$  mean less compatibility,

$$p = \int_t^{\infty} f(t'|H_0) dt' .$$

If  $p$ -value comes out small, then this is evidence against the background-only hypothesis  $\rightarrow$  discovery made!

# Significance from $p$ -value

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



$$p = \int_Z^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1 - \Phi(Z) \quad \text{TMath::Prob}$$

$$Z = \Phi^{-1}(1 - p) \quad \text{TMath::NormQuantile}$$

# When to publish

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

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

<u>phenomenon</u>	<u>reasonable <math>p</math>-value for discovery</u>
D <sup>0</sup> D <sup>0</sup> mixing	~0.05
Higgs	~ 10 <sup>-7</sup> (?)
Life on Mars	~10 <sup>-10</sup>
Astrology	~10 <sup>-20</sup>

# Bayesian model selection ('discovery')

The probability of hypothesis  $H_0$  relative to its complementary alternative  $H_1$  is often given by the posterior odds:

no Higgs

$$\frac{P(H_0|x)}{P(H_1|x)} = \frac{P(x|H_0)}{P(x|H_1)} \times \frac{\pi(H_0)}{\pi(H_1)}$$

Higgs

Bayes factor  $B_{01}$

prior odds

The Bayes factor is regarded as measuring the weight of evidence of the data in support of  $H_0$  over  $H_1$ .

Interchangeably use  $B_{10} = 1/B_{01}$

# Assessing Bayes factors

One can use the Bayes factor much like a  $p$ -value (or  $Z$  value).

There is an “established” scale, analogous to our  $5\sigma$  rule:

$B_{10}$	Evidence against $H_0$
1 to 3	Not worth more than a bare mention
3 to 20	Positive
20 to 150	Strong
> 150	Very strong

Kass and Raftery, *Bayes Factors*, J. Am Stat. Assoc 90 (1995) 773.

Will this be adopted in HEP?

# Rewriting the Bayes factor

Suppose we have models  $H_i$ ,  $i = 0, 1, \dots$ ,

each with a likelihood  $p(x|H_i, \vec{\theta}_i)$

and a prior pdf for its internal parameters  $\pi_i(\vec{\theta}_i)$

so that the full prior is  $\pi(H_i, \vec{\theta}_i) = p_i \pi_i(\vec{\theta}_i)$

where  $p_i = P(H_i)$  is the overall prior probability for  $H_i$ .

The Bayes factor comparing  $H_i$  and  $H_j$  can be written

$$B_{ij} = \frac{P(H_i|\vec{x})}{P(H_i)} \bigg/ \frac{P(H_j|\vec{x})}{P(H_j)}$$

# Bayes factors independent of $P(H_i)$

For  $B_{ij}$  we need the posterior probabilities marginalized over all of the internal parameters of the models:

$$\begin{aligned} P(H_i|\vec{x}) &= \int P(H_i, \vec{\theta}_i|\vec{x}) d\vec{\theta}_i \\ &= \frac{\int L(\vec{x}|H_i, \vec{\theta}_i) p_i \pi_i(\vec{\theta}_i) d\vec{\theta}_i}{P(x)} \end{aligned}$$

Use Bayes theorem

So therefore the Bayes factor is

$$B_{ij} = \frac{\int L(\vec{x}|H_i, \vec{\theta}_i) \pi_i(\vec{\theta}_i) d\vec{\theta}_i}{\int L(\vec{x}|H_j, \vec{\theta}_j) \pi_j(\vec{\theta}_j) d\vec{\theta}_j}$$

Ratio of marginal likelihoods

The prior probabilities  $p_i = P(H_i)$  cancel.



# Numerical determination of Bayes factors

Both numerator and denominator of  $B_{ij}$  are of the form

$$m = \int L(\vec{x}|\vec{\theta})\pi(\vec{\theta}) d\vec{\theta} \quad \longleftarrow \text{‘marginal likelihood’}$$

Various ways to compute these, e.g., using sampling of the posterior pdf (which we can do with MCMC).

Harmonic Mean (and improvements)

Importance sampling

Parallel tempering ( $\sim$ thermodynamic integration)

...

See e.g. Kass and Raftery, *Bayes Factors*, J. Am. Stat. Assoc. 90 (1995) 773-795.

# Harmonic mean estimator

E.g., consider only one model and write Bayes theorem as:

$$\frac{\pi(\boldsymbol{\theta})}{m} = \frac{p(\boldsymbol{\theta}|\mathbf{x})}{L(\mathbf{x}|\boldsymbol{\theta})}$$

$\pi(\boldsymbol{\theta})$  is normalized to unity so integrate both sides,

$$m^{-1} = \int \frac{1}{L(\mathbf{x}|\boldsymbol{\theta})} p(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta} = E_p[1/L]$$

posterior  
expectation



Therefore sample  $\boldsymbol{\theta}$  from the posterior via MCMC and estimate  $m$  with one over the average of  $1/L$  (the harmonic mean of  $L$ ).

M.A. Newton and A.E. Raftery, *Approximate Bayesian Inference by the Weighted Likelihood Bootstrap*, Journal of the Royal Statistical Society B 56 (1994) 3-48.

# Improvements to harmonic mean estimator

The harmonic mean estimator is numerically very unstable; formally infinite variance (!). Gelfand & Dey propose variant:

Rearrange Bayes thm; multiply both sides by arbitrary pdf  $f(\boldsymbol{\theta})$ :

$$\frac{f(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{x})}{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})} = \frac{f(\boldsymbol{\theta})}{m}$$

Integrate over  $\boldsymbol{\theta}$ :  $m^{-1} = \int \frac{f(\boldsymbol{\theta})}{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})} p(\boldsymbol{\theta}|\mathbf{x}) = E_p \left[ \frac{f(\boldsymbol{\theta})}{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})} \right]$

Improved convergence if tails of  $f(\boldsymbol{\theta})$  fall off faster than  $L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$

Note harmonic mean estimator is special case  $f(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$ .

A.E. Gelfand and D.K. Dey, *Bayesian model choice: asymptotics and exact calculations*, Journal of the Royal Statistical Society B 56 (1994) 501-514.

# Importance sampling

Need pdf  $f(\boldsymbol{\theta})$  which we can evaluate at arbitrary  $\boldsymbol{\theta}$  and also sample with MC.

The marginal likelihood can be written

$$m = \int \frac{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{\theta})} f(\boldsymbol{\theta}) d\boldsymbol{\theta} = E_f \left[ \frac{L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\boldsymbol{\theta})} \right]$$

Best convergence when  $f(\boldsymbol{\theta})$  approximates shape of  $L(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ .

Use for  $f(\boldsymbol{\theta})$  e.g. multivariate Gaussian with mean and covariance estimated from posterior (e.g. with MINUIT).

# 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

# Setting limits on Poisson parameter

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  $n_{\text{obs}} = 5$  events.

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

→ set upper limit on the parameter  $s$ .



# Upper limit for Poisson parameter

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:

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

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

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

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

# Calculating Poisson parameter limits

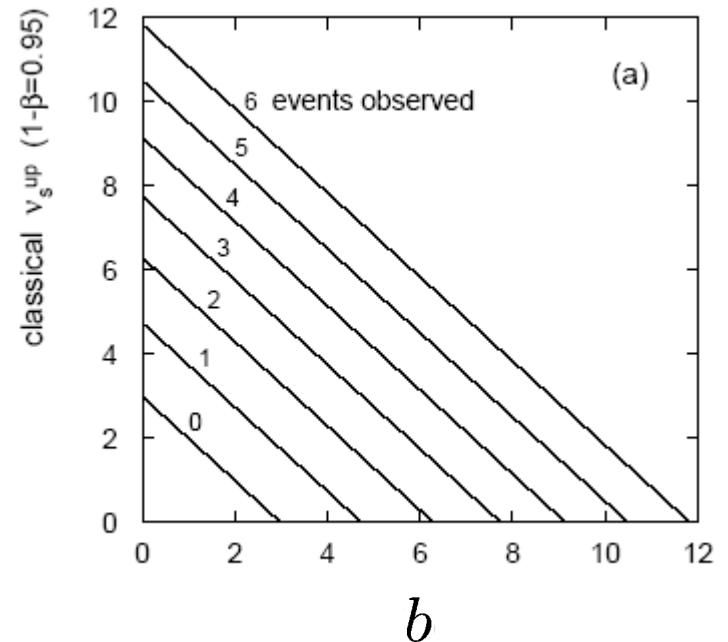
To solve for  $s_{\text{lo}}$ ,  $s_{\text{up}}$ , can exploit relation to  $\chi^2$  distribution:

$$s_{\text{lo}} = \frac{1}{2} F_{\chi^2}^{-1}(\alpha; 2n) - b$$

Quantile of  $\chi^2$  distribution

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

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



# Limits near a physical boundary

Suppose e.g.  $b = 2.5$  and we observe  $n = 0$ .

If we choose  $CL = 0.9$ , we find from the formula for  $s_{\text{up}}$

$$s_{\text{up}} = -0.197 \quad (CL = 0.90)$$

Physicist:

We already knew  $s \geq 0$  before we started; can't use negative upper limit to report result of expensive experiment!

Statistician:

The interval is designed to cover the true value only 90% of the time — this was clearly not one of those times.

Not uncommon dilemma when limit of parameter is close to a physical boundary, cf.  $m_\nu$  estimated using  $E^2 - p^2$ .

# Expected limit for $s = 0$

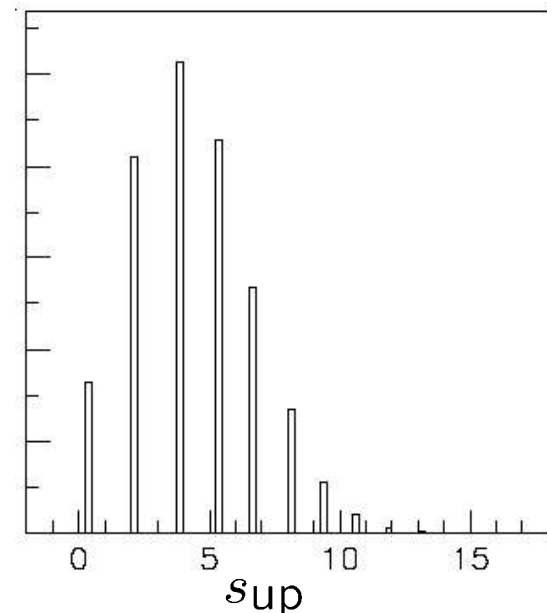
Physicist: I should have used  $CL = 0.95$  — then  $s_{up} = 0.496$

Even better: for  $CL = 0.917923$  we get  $s_{up} = 10^{-4}$ !

Reality check: with  $b = 2.5$ , typical Poisson fluctuation in  $n$  is at least  $\sqrt{2.5} = 1.6$ . How can the limit be so low?

Look at the mean limit for the no-signal hypothesis ( $s = 0$ ) (sensitivity).

Distribution of 95% CL limits with  $b = 2.5, s = 0$ .  
Mean upper limit = 4.44



# The Bayesian approach

In Bayesian statistics need to start with ‘prior pdf’  $\pi(\theta)$ , this reflects degree of belief about  $\theta$  before doing the experiment.

Bayes’ theorem tells how our beliefs should be updated in light of the data  $x$ :

$$p(\theta|x) = \frac{L(x|\theta)\pi(\theta)}{\int L(x|\theta')\pi(\theta') d\theta'} \propto L(x|\theta)\pi(\theta)$$

Integrate posterior pdf  $p(\theta | x)$  to give interval with any desired probability content.

For e.g. Poisson parameter 95% CL upper limit from

$$0.95 = \int_{-\infty}^{\text{sup}} p(s|n) ds$$

# Bayesian prior for Poisson parameter

Include knowledge that  $s \geq 0$  by setting prior  $\pi(s) = 0$  for  $s < 0$ .

Often try to reflect ‘prior ignorance’ with e.g.

$$\pi(s) = \begin{cases} 1 & s \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Not normalized but this is OK as long as  $L(s)$  dies off for large  $s$ .

Not invariant under change of parameter — if we had used instead a flat prior for, say, the mass of the Higgs boson, this would imply a non-flat prior for the expected number of Higgs events.

Doesn’t really reflect a reasonable degree of belief, but often used as a point of reference;

or viewed as a recipe for producing an interval whose frequentist properties can be studied (coverage will depend on true  $s$ ).

# Bayesian interval with flat prior for $s$

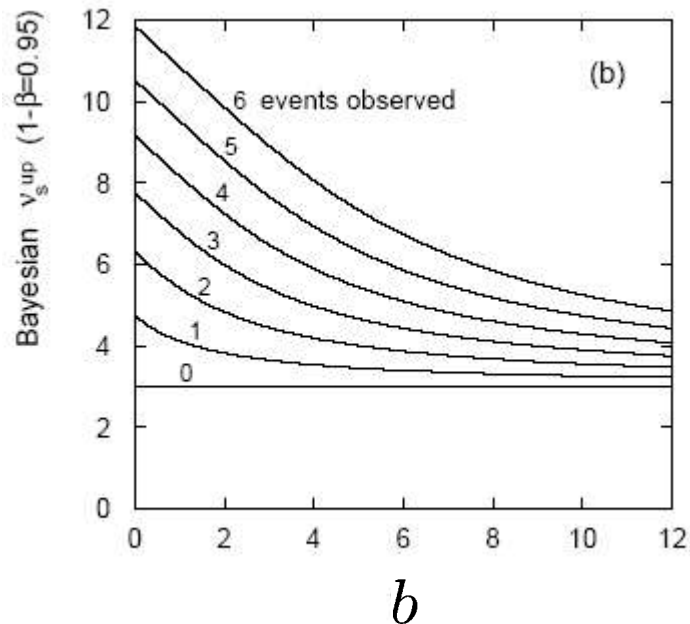
Solve numerically to find limit  $s_{\text{up}}$ .

For special case  $b = 0$ , Bayesian upper limit with flat prior numerically same as classical case ('coincidence').

Otherwise Bayesian limit is everywhere greater than classical ('conservative').

Never goes negative.

Doesn't depend on  $b$  if  $n = 0$ .



# Likelihood ratio limits (Feldman-Cousins)

Define likelihood ratio for hypothesized parameter value  $s$ :

$$l(s) = \frac{L(n|s, b)}{L(n|\hat{s}, b)} \quad \text{where} \quad \hat{s} = \begin{cases} n - b & n \geq b, \\ 0 & \text{otherwise} \end{cases}$$

Here  $\hat{s}$  is the ML estimator, note  $0 \leq l(s) \leq 1$ .

Critical region defined by low values of likelihood ratio.

Resulting intervals can be one- or two-sided (depending on  $n$ ).

(Re)discovered for HEP by Feldman and Cousins,  
Phys. Rev. D 57 (1998) 3873.



# More on intervals from LR test (Feldman-Cousins)

Caveat with coverage: suppose we find  $n \gg b$ .

Usually one then quotes a measurement:  $\hat{s} = n - b$ ,  $\hat{\sigma}_{\hat{s}} = \sqrt{n}$

If, however,  $n$  isn't large enough to claim discovery, one sets a limit on  $s$ .

FC pointed out that if this decision is made based on  $n$ , then the actual coverage probability of the interval can be less than the stated confidence level ('flip-flopping').

FC intervals remove this, providing a smooth transition from 1- to 2-sided intervals, depending on  $n$ .

But, suppose FC gives e.g.  $0.1 < s < 5$  at 90% CL,  $p$ -value of  $s=0$  still substantial. Part of upper-limit 'wasted'?