Aachen Online Statistics School

GDC Lecture 1: Bayesian probability and credible intervals

RWTH Aachen (online)
13-17 March 2023

https://indico.desy.de/event/37562/

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

Outline of GDC lectures

→ Tue. 14.3	Probability (Bayes vs. Frequentist) Bayesian parameter and interval estimation
Wed. 15.3	Frequentist confidence regions and intervals
Thu. 16.3	Python software for frequentist and Bayesian confidence regions.
Fri. 17.3	Searches and discoveries using likelihoods

A quick review of probability

Frequentist (A = outcome of repeatable observation)

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

Subjective (A = hypothesis)

$$P(A) = \text{degree of belief that } A \text{ is true}$$

Conditional probability:

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

E.g. rolling a die,
outcome n = 1,2,...,6:

$$P(n \leq 3|n \text{ even}) = \frac{P((n \leq 3) \cap n \text{ even})}{P(n \text{ even})} = \frac{1/6}{3/6} = \frac{1}{3}$$

A and B are independent iff:

$$P(A \cap B) = P(A)P(B)$$

I.e. if A, B independent, then

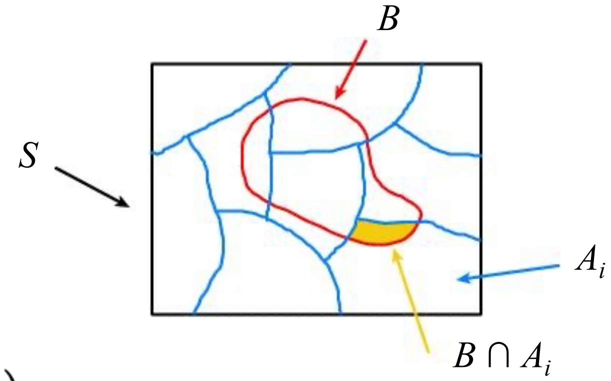
$$P(A|B) = \frac{P(A)P(B)}{P(B)} = P(A)$$

Bayes' theorem

Use definition of conditional probability and $P(A \cap B) = P(B \cap A)$

$$\rightarrow \quad P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$ (Bayes' theorem)

If set of all outcomes $S = \cup_i A_i$ with A_i disjoint, then law of total probability for $P(B)$ says



$$P(B) = \sum_i P(B \cap A_i) = \sum_i P(B|A_i)P(A_i)$$

so that Bayes' theorem becomes $$P(A|B) = \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)}$$

Bayes' theorem holds regardless of how probability is interpreted (frequency, degree of belief...).

Frequentist Statistics – general philosophy

In frequentist statistics, probabilities are associated only with the data, i.e., outcomes of repeatable observations (shorthand: $\boldsymbol{x}$).

Probability = limiting frequency

Probabilities such as

P (string theory is true),
P ($0.117 < \alpha_s < 0.119$),
P (Biden wins in 2024),

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.

Preferred theories (models, hypotheses, ...) are those that predict a high probability for data “like” the data observed.

Bayesian Statistics – general philosophy

In Bayesian statistics, use subjective probability 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

Bayes' theorem has an "if-then" character: If your prior probabilities were $\pi(H)$, then it says how these probabilities should change in the light of the data.

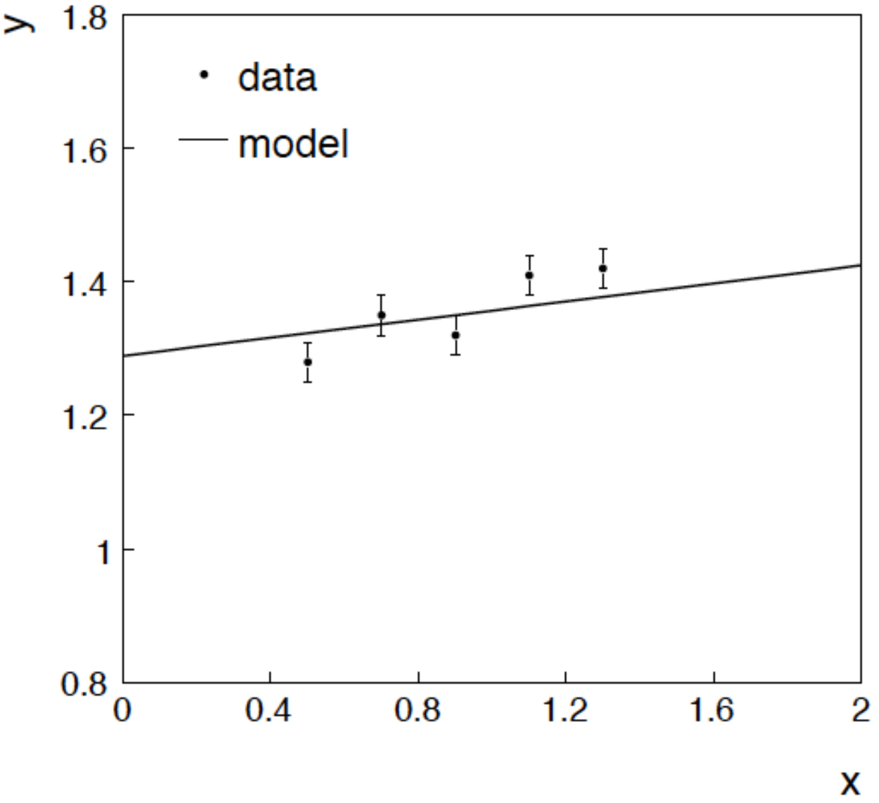
No general prescription for priors (subjective!)

Bayesian parameter estimation
Example: fitting a straight line

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

Model: y_i independent and all follow $y_i \sim$ Gauss($\mu(x_i), \sigma_i$)

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



assume x_i and σ_i known.

Goal: estimate θ_0

Here suppose we don't care about θ_1 (example of a "nuisance parameter")

Connection with Maximum Likelihood and Least Squares

Both the Bayesian a Frequentist approaches require the **likelihood**, P(data|parameters) $= P(\boldsymbol{y}|\boldsymbol{\theta}) = L(\theta_0, \theta_1)$.

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)

Probability in the Bayesian approach

In Bayesian statistics we can associate a probability with a hypothesis, e.g., a parameter $\boldsymbol{\theta}$.

Interpret probability of $\boldsymbol{\theta}$ as 'degree of belief' (subjective).

Need to start with 'prior pdf' $\pi(\boldsymbol{\theta})$, this reflects degree of belief about $\boldsymbol{\theta}$ before doing the experiment.

Our experiment has data $\boldsymbol{y}$, → likelihood $P(\boldsymbol{y}|\boldsymbol{\theta})$.

Bayes' theorem tells how our beliefs should be updated in light of the data $\boldsymbol{y}$:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{P(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int P(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \propto P(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$$

Posterior pdf $p(\boldsymbol{\theta}|\boldsymbol{y})$ contains all our knowledge about $\boldsymbol{\theta}$.

Assigning prior probabilities

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

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

← suppose knowledge of θ_0 has no influence on knowledge of θ_1

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

← 'non-informative', in any case much broader than $L(\theta_0)$

Suppose we have an independent control measurement of θ_1, e.g., $t_1 \sim \text{Gauss}(\theta_1, \sigma_{t1})$. Take "prior" to mean after t_1, before $\boldsymbol{y}$:

$$\pi_1(\theta_1) = p(\theta_1 | t_1) \propto p(t_1 | \theta_1)\pi_{\text{Ur}}(\theta_1) = \frac{1}{\sqrt{2\pi}\sigma_t} e^{-(t_1-\theta_1)^2/2\sigma_t^2} \times \text{const.}$$

prior after t_1, before $\boldsymbol{y}$

Ur = "primordial" prior

Likelihood for control measurement t_1

Posterior for parameters from Bayes' theorem

Putting the ingredients 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 ∝ likelihood × prior

Note here the likelihood only reflects the measurements $\boldsymbol{y}$.

The information from the control measurement t_1 has been put into the prior for θ_1.

We would get the same result using the likelihood $P(\boldsymbol{y},t|\theta_0,\theta_1)$ and the constant "Ur-prior" for θ_1.

Marginalizing the posterior pdf

We then integrate (marginalize) $p(\theta_0, \theta_1|\mathbf{y})$ to find $p(\theta_0|\mathbf{y})$:

$$p(\theta_0|\mathbf{y}) = \int p(\theta_0, \theta_1|\mathbf{y})\, d\theta_1$$

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

$$p(\theta_0|\mathbf{y}) = \frac{1}{\sqrt{2\pi}\sigma_{\theta_0}} e^{-(\theta_0 - \hat{\theta}_0)^2/2\sigma_{\theta_0^2}}$$

$\hat{\theta}_0 =$ same as from maximum likelihood

$\sigma_{\theta_0} = \sigma_{\hat{\theta}_0}$ (same as for maximum likelihood)

For this example, numbers come out same as in frequentist approach, but interpretation different.

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.

→ Use Markov Chain Monte Carlo (MCMC)

Bayesian Analysis Toolkit: https://github.com/bat/BAT.jl

BAT
BAYESIAN ANALYSIS TOOLKIT

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 all values independent .

Basic idea: sample multidimensional $\boldsymbol{\theta}$ but look only at distribution of parameters of interest.

MCMC basics: Metropolis-Hastings algorithm

Goal: given an n-dimensional pdf $p(\boldsymbol{\theta})$ up to a proportionality constant, generate a sequence of points $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$, $\boldsymbol{\theta}_3$,...

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

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

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

3) Form 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).

Still works if $p(\boldsymbol{\theta})$ is known only as a proportionality, which is usually what we have from Bayes' theorem: $p(\boldsymbol{\theta}|\boldsymbol{x}) \propto p(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

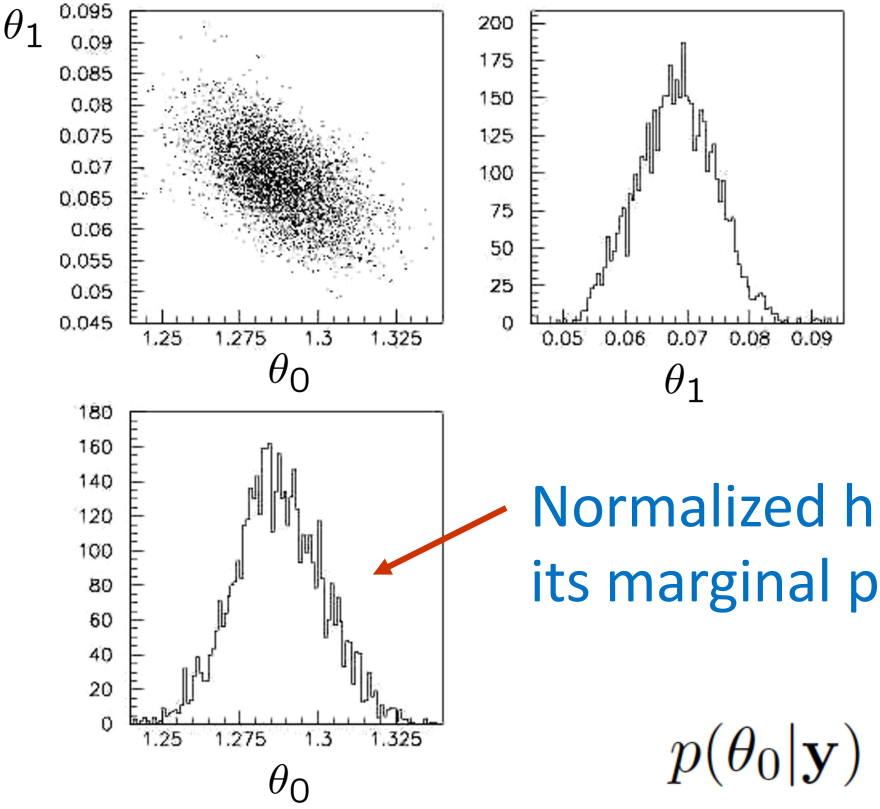
The proposal density can be (almost) anything, but choose so as to minimize autocorrelation. Often take proposal density symmetric: $q(\boldsymbol{\theta}; \boldsymbol{\theta}_0) = q(\boldsymbol{\theta}_0; \boldsymbol{\theta})$

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 $p(\boldsymbol{\theta})$, take it; if not, only take the step with probability $p(\boldsymbol{\theta})/p(\boldsymbol{\theta}_0)$. If proposed step rejected, repeat the current point.

Example: posterior pdf from MCMC

Sample the posterior pdf from previous example with MCMC:



Normalized histogram of θ_0 gives its marginal posterior pdf:

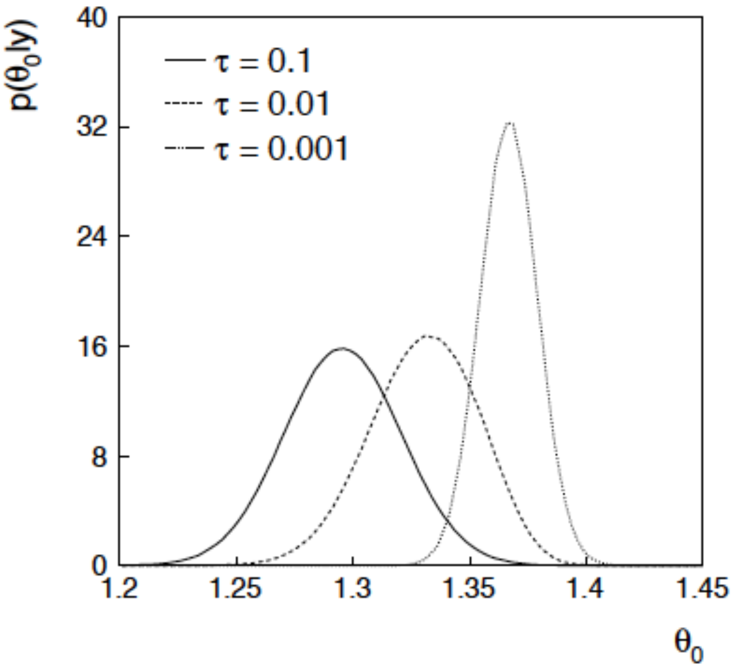
$$p(\theta_0|\mathbf{y}) = \int p(\theta_0, \theta_1|\mathbf{y})\, d\theta_1$$

Bayesian method with alternative priors

Suppose we don't have a previous measurement of θ_1 but rather, an "expert" 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 θ_0:



This summarizes all knowledge about θ_0.

Look also at result from variety of priors.

The Poisson counting experiment

Suppose we do a counting experiment and observe n events.

Events could be from *signal* process or from *background* – we only count the total number.

Poisson model:

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

s = mean (i.e., expected) # of signal events

b = mean # of background events (suppose known)

Goal is to make inference about s, e.g., given an observed n, set an upper limit on s.

The Bayesian approach to limits

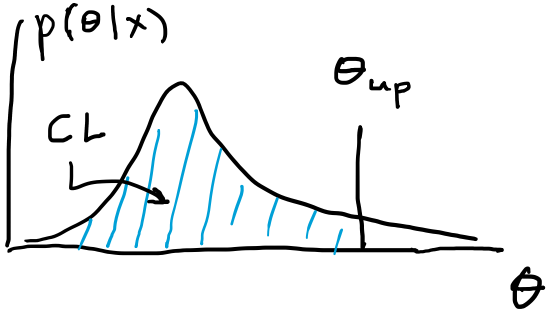
In Bayesian statistics need to start with 'prior pdf' $\pi(\theta)$, this reflects degree of belief about θ before doing the experiment.

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

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

Normalize posterior pdf $p(\theta|x)$ to unity, then integrate to give interval with any desired probability content (or "credibility level" CL or 1 – α), e.g., CL = 95%:

$$\mathrm{CL} = 1 - \alpha = \int_{-\infty}^{\theta_{\mathrm{up}}} p(\theta|x)\,d\theta$$



Bayesian prior for Poisson parameter

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

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

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

Not normalized; can be OK provided $p(n|s)$ dies off quickly for large s.

Not invariant under change of parameter — if we had used instead a flat prior for a nonlinear function of s, then this would imply a non-flat prior for s.

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 (e.g., coverage probability, which will depend on true s).

Bayesian upper limit with flat prior for *s*

Put Poisson likelihood and flat prior into Bayes' theorem:

$$p(s|n) \propto p(n|s)\pi(s) = \frac{(s+b)^n}{n!}e^{-(s+b)} \times 1\,, \quad s \geq 0$$

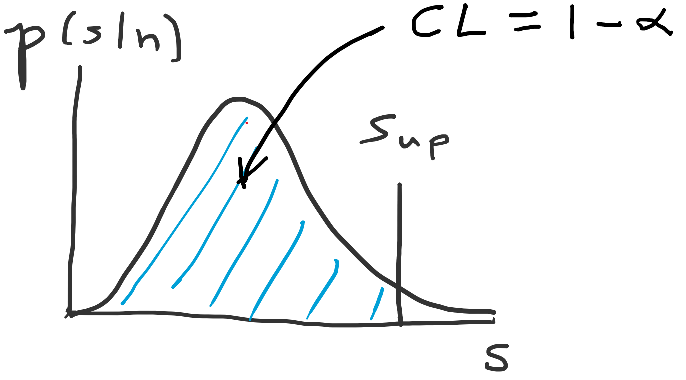
Normalize to unit area:

$$p(s|n) = \frac{(s+b)^n e^{-(s+b)}}{\Gamma(b, n+1)}$$

upper incomplete gamma function

Upper limit $s_{\rm up}$ determined by

$$1 - \alpha = \int_0^{s_{\rm up}} p(s|n)\, ds$$



Bayesian interval with flat prior for s

Solve to find limit s_{up}:

quantile of chi-square distribution

$$s_{\text{up}} = \frac{1}{2} F^{-1}_{\chi^2}[p, 2(n+1)] - b$$

where

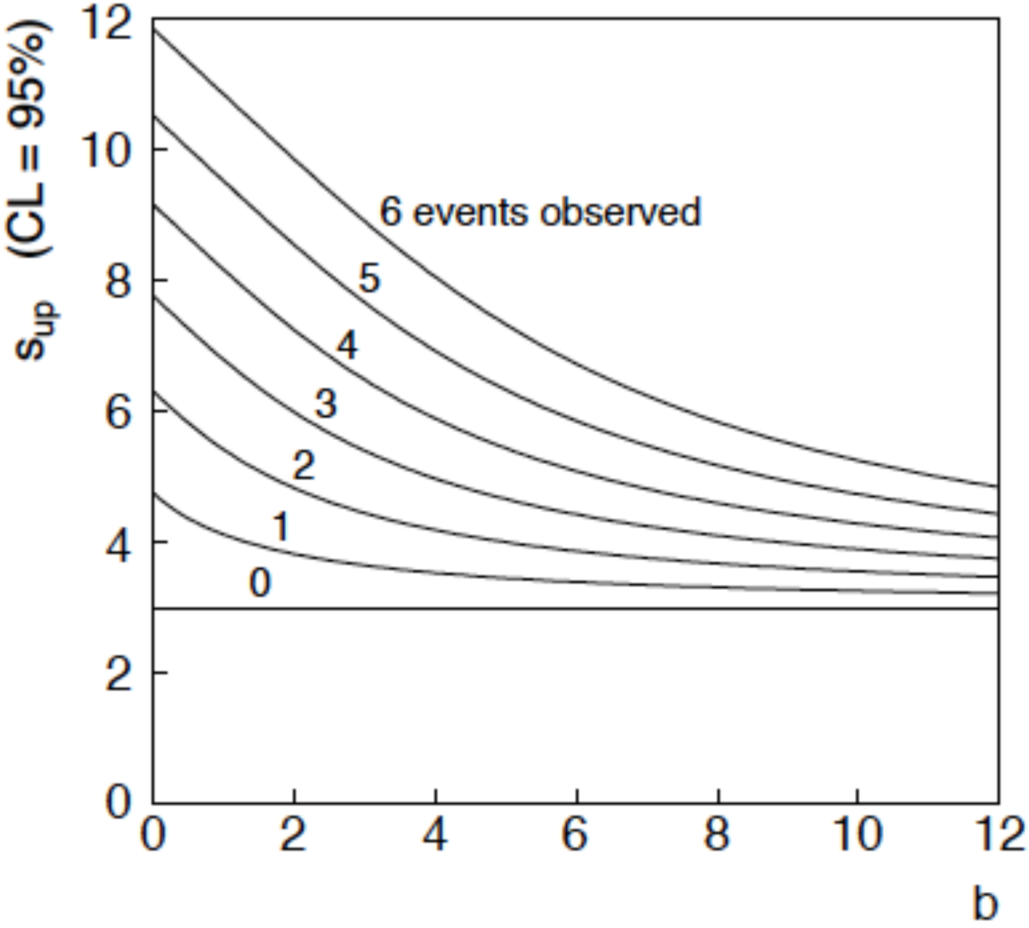
$$p = 1 - \alpha \left(1 - F_{\chi^2}[2b, 2(n+1)]\right)$$

For special case $b = 0$, Bayesian upper limit with flat prior numerically same as one-sided frequentist case (‘coincidence’).

Bayesian interval with flat prior for *s*

For $b > 0$ Bayesian limit is everywhere greater than the (one sided) frequentist upper limit.

Never goes negative. Doesn't depend on b if $n = 0$.



Priors from formal rules

We took the prior for a Poisson mean to be constant to reflect a lack of prior knowledge and noted this was not invariant under change of parameter.

Because of difficulties in encoding a vague degree of belief in a prior, one often attempts to derive the prior from formal rules, e.g., to satisfy certain invariance principles or to provide maximum information gain for a certain set of measurements.

Often called “objective priors”
Form basis of Objective Bayesian Statistics

The priors do not reflect a degree of belief (but might represent possible extreme cases).

In Objective Bayesian analysis, can use the intervals in a frequentist way, i.e., regard Bayes’ theorem as a recipe to produce an interval with a given coverage probability.

Jeffreys prior

According to *Jeffreys' rule*, take prior according to

$$\pi(\boldsymbol{\theta}) \propto \sqrt{\det(I(\boldsymbol{\theta}))}$$

where

$$I_{ij}(\boldsymbol{\theta}) = -E\left[\frac{\partial^2 \ln P(\mathbf{x}|\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right] = -\int \frac{\partial^2 \ln P(\mathbf{x}|\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j} P(\mathbf{x}|\boldsymbol{\theta})\,d\mathbf{x}$$

is the Fisher information matrix.

One can show that this leads to inference that is invariant under a transformation of parameters in the following sense:

Start with the Jeffreys prior for θ: $\pi_\theta(\theta) \sim \sqrt{}(\det I(\theta))$

Use it in Bayes' theorem to find:

$$P(\theta|\mathbf{x}) \propto P(\mathbf{x}|\theta)\pi_\theta(\theta)$$

Jeffreys prior (2)

Now consider a function $\eta(\theta)$. The posterior for η is

$$P(\eta|\mathbf{x}) = P(\theta|\mathbf{x}) \left| \frac{d\theta}{d\eta} \right|$$

Alternatively, start with η and use its Jeffreys' prior:

$$\pi_\eta(\eta) \propto \sqrt{\det I(\eta)}$$

Use this in Bayes' theorem: $P(\eta|\mathbf{x}) \propto P(\mathbf{x}|\eta)\pi_\eta(\eta)$

One can show that Jeffreys' prior results in the same $P(\eta|x)$ in both cases. For details (single-parameter case) see:

http://www.pp.rhul.ac.uk/~cowan/stat/notes/JeffreysInvariance.pdf

Jeffreys prior for Poisson mean

Suppose $n \sim$ Poisson(μ). To find the Jeffreys' prior for μ,

$$P(n|\mu) = \frac{\mu^n}{n!} e^{-\mu} \qquad \frac{\partial^2 \ln P(n|\mu)}{\partial \mu^2} = -\frac{n}{\mu^2}$$

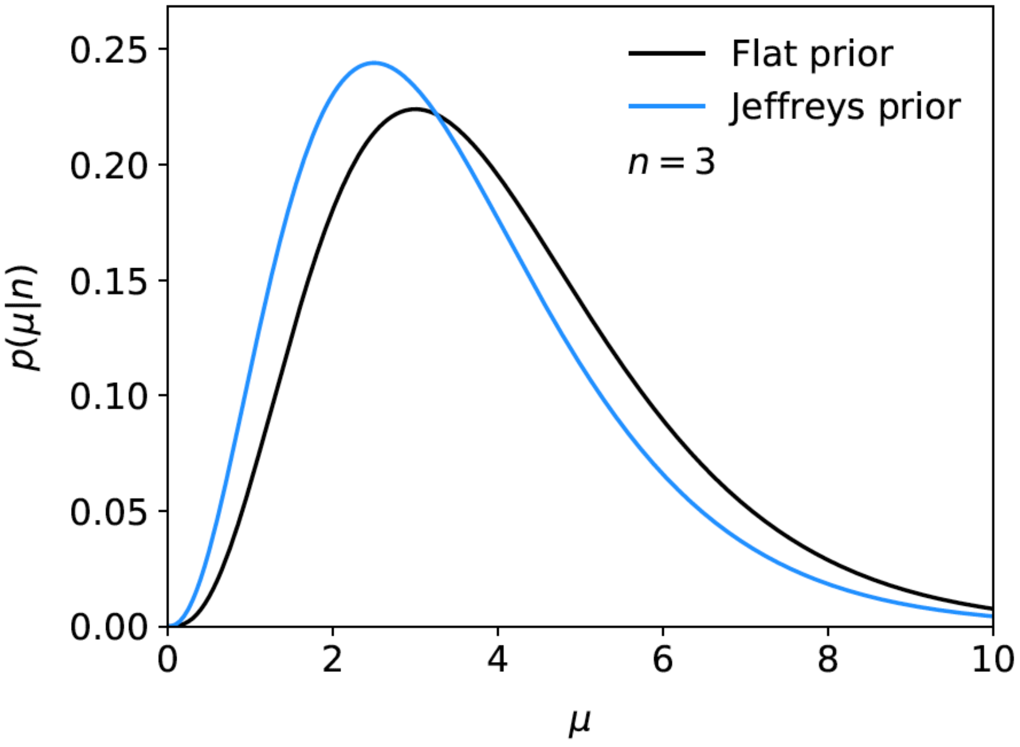
$$I = -E\left[\frac{\partial^2 \ln P(n|\mu)}{\partial \mu^2}\right] = \frac{E[n]}{\mu^2} = \frac{1}{\mu}$$

$$\pi(\mu) \propto \sqrt{I(\mu)} = \frac{1}{\sqrt{\mu}}$$

So e.g. for $\mu = s + b$, this means the prior $\pi(s) \sim 1/\sqrt{(s+b)}$, which depends on *b*. But this is not designed as a degree of belief about s.

Posterior pdf for Poisson mean

From Bayes' theorem, $p(\mu|n) \propto \mu^n e^{-\mu} \pi(\mu)$



Flat, $\pi(\mu) = \text{const.}$

$$p(\mu|n) = \frac{\mu^n e^{-\mu}}{\Gamma(n+1)}$$

Jeffreys, $\pi(\mu) \sim 1/\sqrt{\mu}$

$$p(\mu|n) = \frac{\mu^{n-\frac{1}{2}} e^{-\mu}}{\Gamma(n+\frac{1}{2})}$$

In both cases, posterior is special case of gamma distribution.

Upper limit for Poisson mean

To find upper limit at CL = 1–α, solve

$$1 - \alpha = \int_0^{\mu_{\rm up}} p(\mu | n) \, d\mu$$

Jeffreys prior: $\mu_{\rm up} = P^{-1}(n + \frac{1}{2}, 1 - \alpha)$ = 7.03

Flat prior: $\mu_{\rm up} = P^{-1}(n + 1, 1 - \alpha)$ = 7.75

n=3,
CL=0.95

where P^{-1} is the inverse of the normalized lower incomplete gamma function (see scipy.special)

$$P(a, \mu_{\rm up}) = \frac{1}{\Gamma(a)} \int_0^{\mu_{\rm up}} \mu^{a-1} e^{-\mu} \, d\mu$$

Summing up...

Bayesian methods allow one to associate a probability with a hypothesis, e.g., a hypothesized parameter.

The final result consists of the posterior probability for the hypothesis (or parameter) given the observed data (or a summary statistic obtained from it, e.g., a limit).

Requires one to specify prior probabilities.

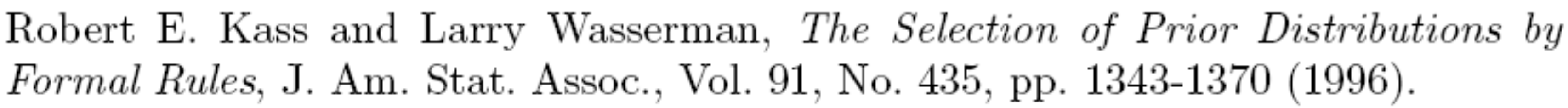
Bayesian computation involves integrals over parameter space (MCMC).

Sometimes (often?) use Bayesian methods to obtain a result, then "forget" its Bayesian origins and exploit its frequentist properties.

Extra slides

Priors from formal rules (cont.)

For a review of priors obtained by formal rules see, e.g.,



Robert E. Kass and Larry Wasserman, *The Selection of Prior Distributions by Formal Rules*, J. Am. Stat. Assoc., Vol. 91, No. 435, pp. 1343-1370 (1996).

Formal priors have not been widely used in Particle Physics, but there has been interest in this direction, especially the reference priors of Bernardo and Berger; see e.g.

L. Demortier, S. Jain and H. Prosper, *Reference priors for high energy physics*, Phys. Rev. D 82 (2010) 034002, arXiv:1002.1111.

D. Casadei, *Reference analysis of the signal + background model in counting experiments*, JINST 7 (2012) 01012; arXiv:1108.4270.