Bayesian Methods: Theory and Practice
Lecture 3 – Applications

Harrison B. Prosper

Florida State University

CMS Statistics Committee

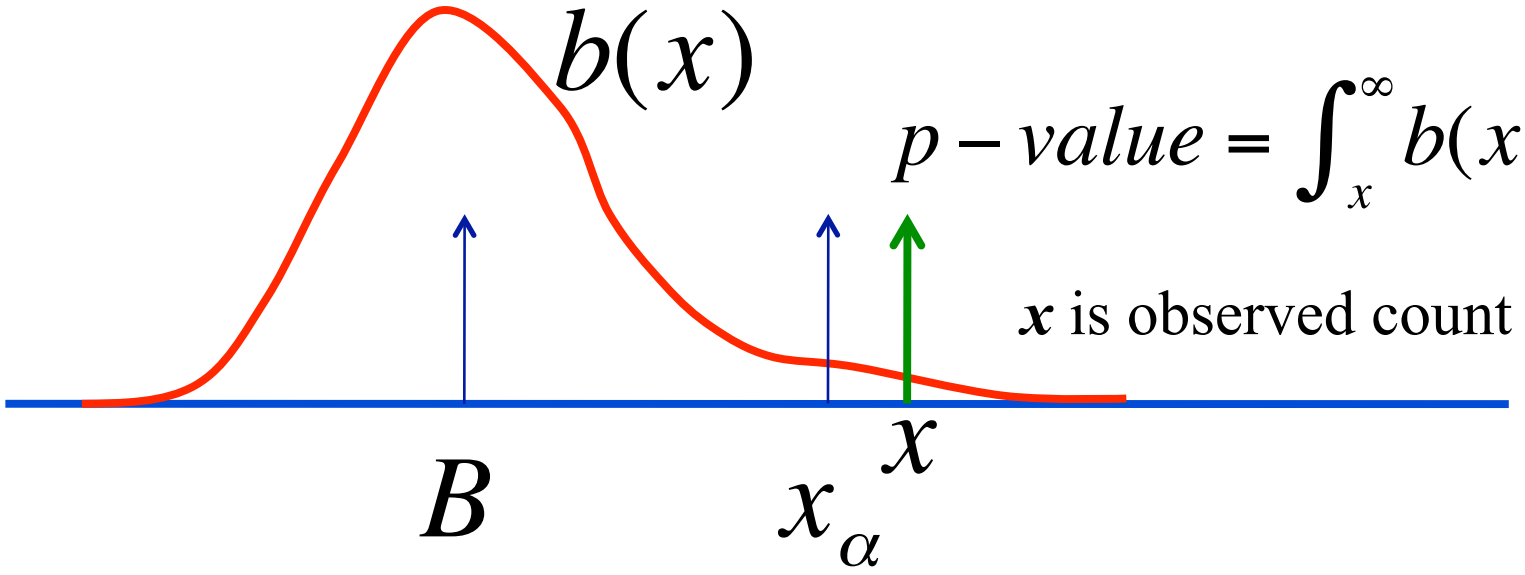
08-08-08

Outline

- Lecture 3 – **Applications**
  - Hypothesis Testing – Recap
  - A Single Count – Recap
  - A Single Count
  - Formal Priors
    - Intrinsic Bayes Factors
    - Flat Priors
    - Jeffreys Priors
    - Reference Priors
  - Reference Prior Project
  - Summary

Hypothesis Testing – Recap

Null hypothesis (H_0): background-only



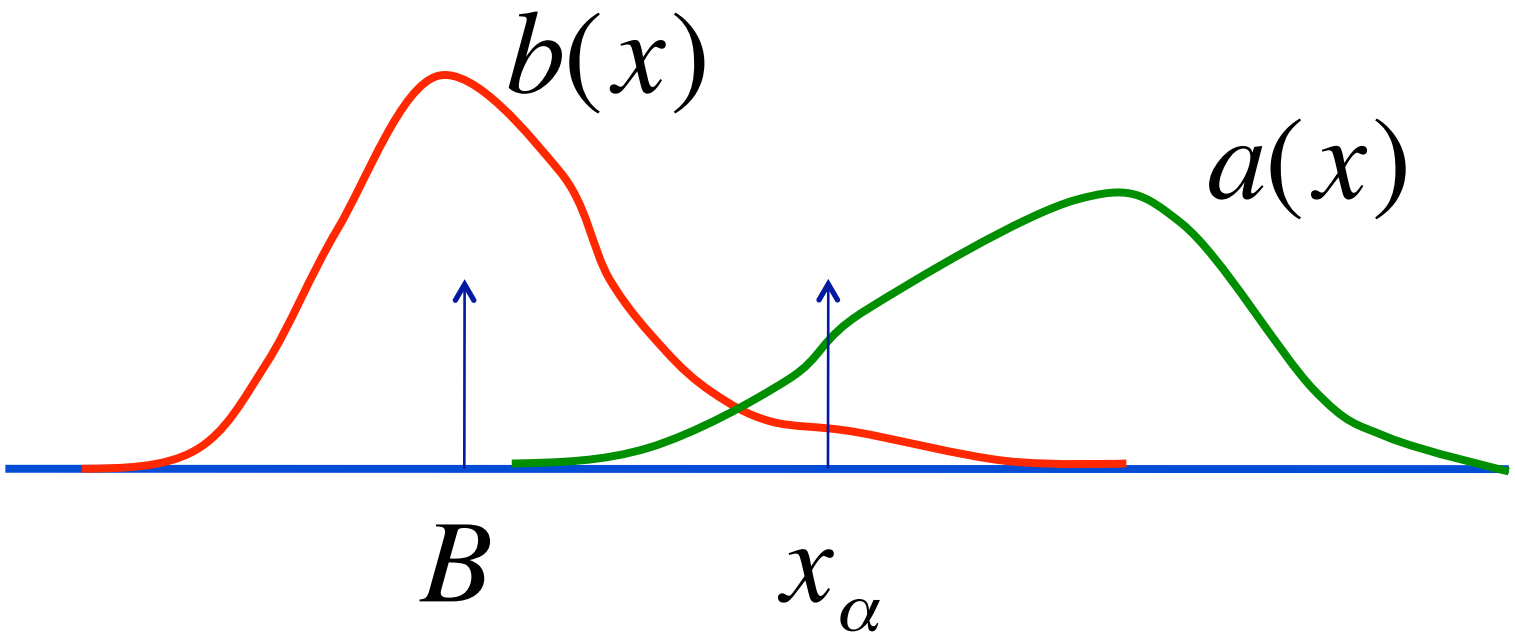
$$\alpha = \int_{x_\alpha}^{\infty} b(x)dx$$

significance level

HEP – **2.7 x 10^{-7}**
BIO – **5.0 x 10^{-2}**

If *p-value* < α declare victory!

Hypothesis Testing – Recap



$$\alpha = \int_{x_\alpha}^{\infty} b(x)dx \qquad p = \int_{x_\alpha}^{\infty} a(x)dx$$

significance level of test ***power*** of test

Hypothesis Testing – Recap

Standard Bayesian method:

1. Factorize the priors: $\pi(\theta, \phi, H) = \pi(\theta, \phi|H)\ \pi(H)$
2. For each hypothesis, H, compute the **evidence**

$$p(D \mid H) = \int\int p(D \mid \theta, \phi, H)\, \pi(\theta, \phi \mid H)\, d\theta\, d\phi$$

3. Compute the **Bayes factors**

$$B_{ij} = \frac{p(D \mid H_i)}{p(D \mid H_j)}$$

4. If, e.g., B_{10}, or some function thereof, say, $\sqrt{(2 \ln B_{10})}$ > agreed-upon threshold, accept H_1, otherwise keep H_0.

A Single Count – Recap

The standard model for a counting experiment:

background-only

$p(D|b, H_0) = \text{Poisson}(D|b, H_0)$ probability model

$\pi(b, H_0)$ prior

background+signal

$p(D|b, s, H_1) = \text{Poisson}(D|b + s)$ probability model

$\pi(b, s, H_1)$ prior

A Single Count – Recap

The **prior predictive density** $p(D|s, H_1)$ for the experiment:

$$p(D \mid s, H_1) = \left(\frac{c}{1+c}\right)^{y+1/2} \sum_{r=0}^{D} \frac{1}{(1+c)^r} \frac{\Gamma(y+1/2+r)}{\Gamma(y+1/2)r!} \text{Poisson}(D-r \mid s)$$

The **evidence** $p(D| H_0)$ for the background-only hypothesis:
$p(D| H_0) = p(D|s=0, H_1)$, that is,

$$p(D \mid H_0) = \left(\frac{c}{1+c}\right)^{y+1/2} \frac{1}{(1+c)^D} \frac{\Gamma(y+1/2+D)}{\Gamma(y+1/2)D!}$$

A Single Count – Recap

The **evidence** for the background+signal hypothesis $p(D|\, H_1)$:

$$p(D \mid H_1) = \int_0^\infty p(D \mid s, H_1)\pi(s \mid H_1)ds$$

where $\pi(s|\, H_1)$ is the prior for the signal, which brings us to yesterday's question:

What should $\pi(s|\, H_1)$ be?

…back to A Single Count

The short answer is:

ideally, it is a well-motivated **evidence-based prior** for which

$$\int_S \pi(s \mid H_1) ds = 1$$

otherwise, it is (or ought to be!) a ***very*** carefully chosen **objective prior**.*

*☺ I prefer the name **formal prior**.

A Single Count

Evidence-Based Prior

To create such a prior for the signal requires that we have *some* idea of what signal we are looking for.

So what is the basic strategy?

"know thine enemy"

The Art of War, Sun Tzu, ~300 BC

then

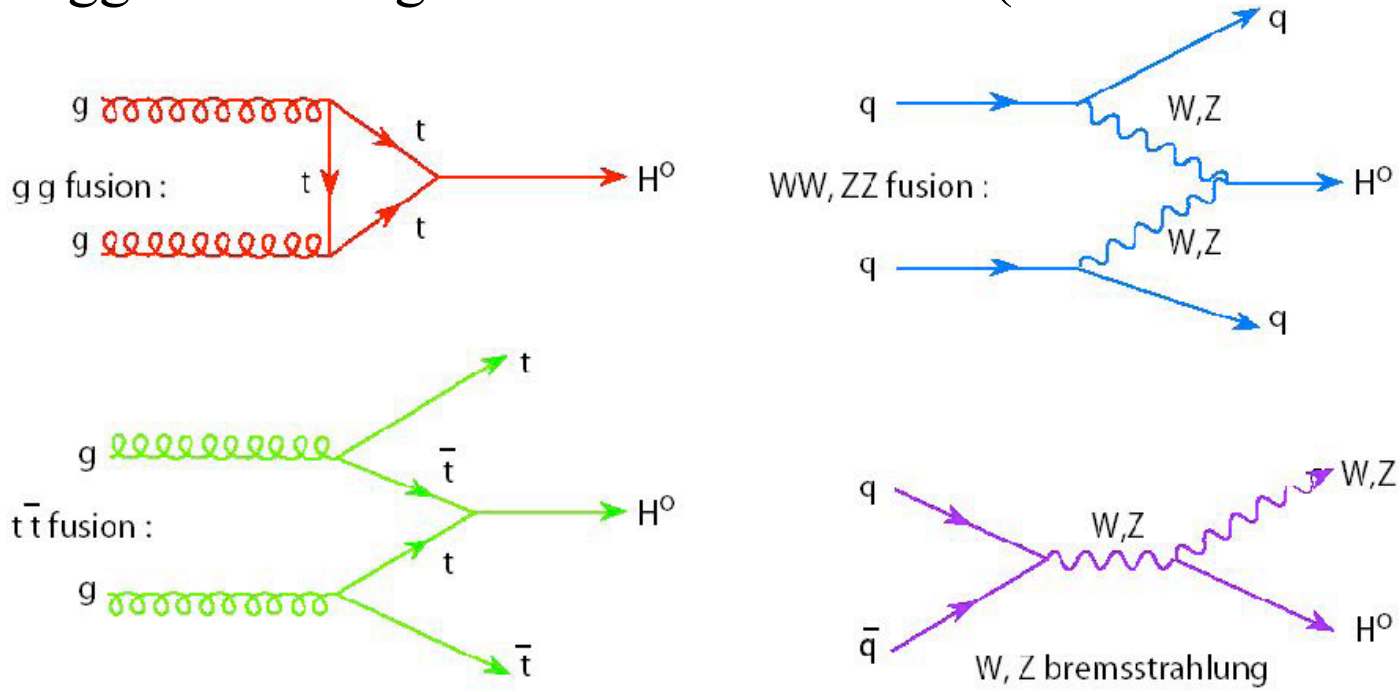
"divide and conquer"

Teenage Mutant Ninja Turtles

A Single Count

Evidence-Based Prior

Consider looking for ***H → WW, WH/ZH → WW***, in the Higgs mass range ***m*** = 155 – 200 GeV (FERMILAB-08-270-E).

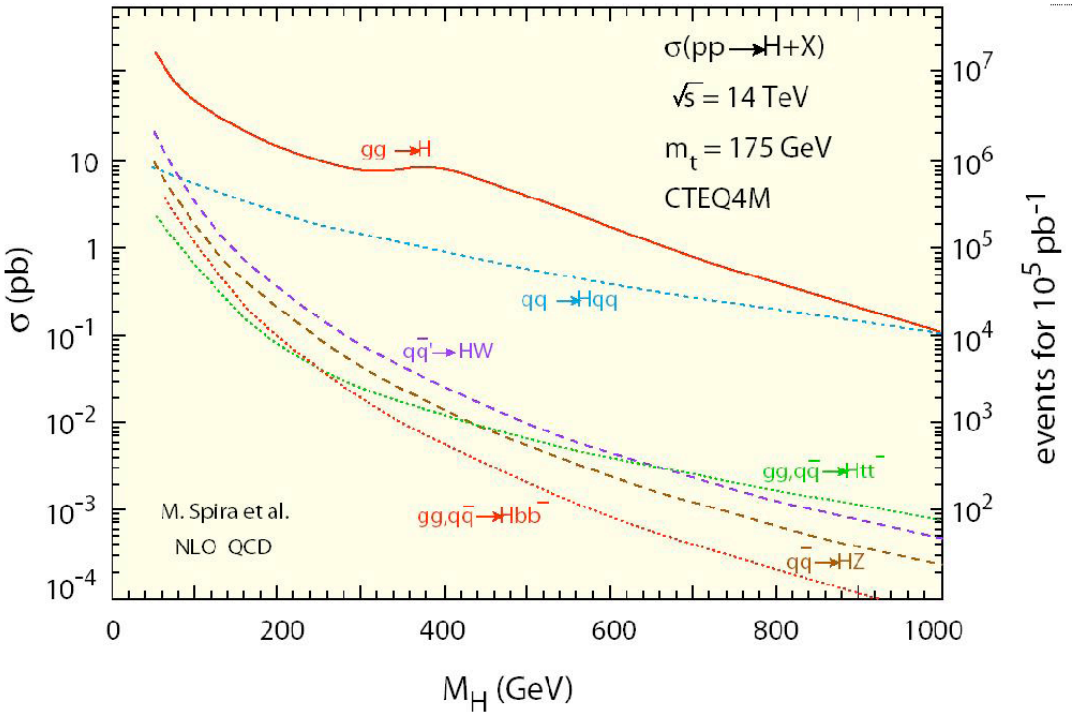


http://www.hep.ph.ic.ac.uk/cms/physics/higgs.html

A Single Count

Evidence-Based Prior

Our ***beliefs*** about the putative Higgs are precise and detailed:



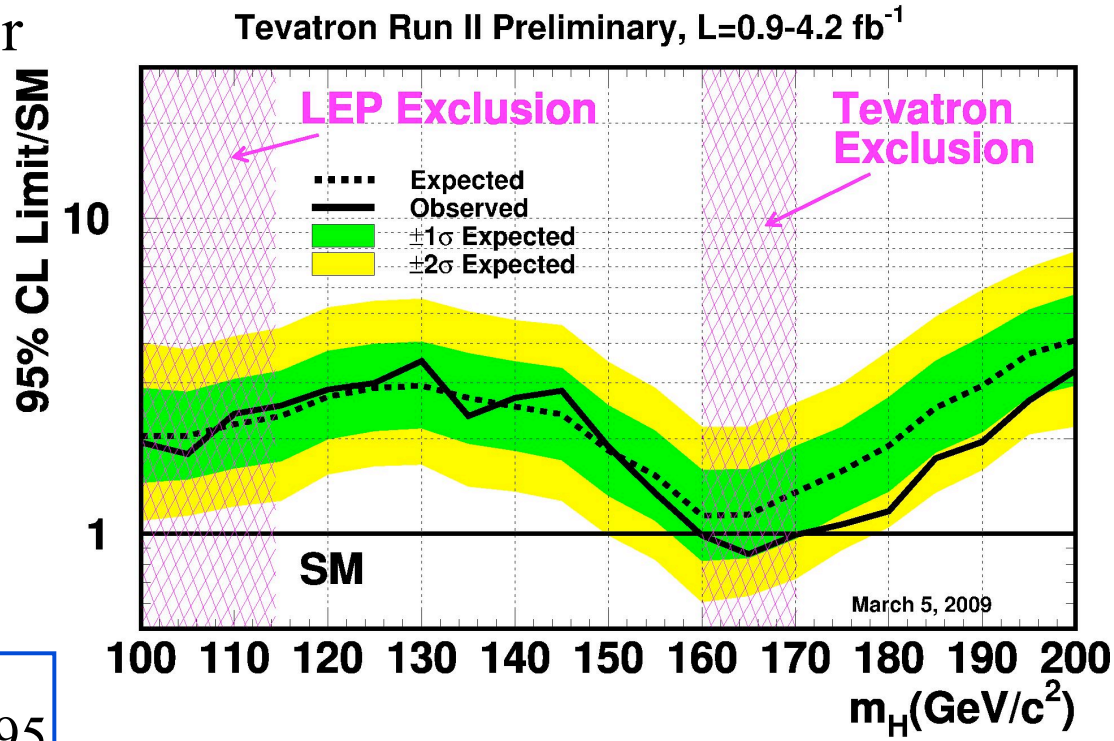
http://www.hep.ph.ic.ac.uk/cms/physics/higgs.html

A Single Count

Evidence-Based Prior

Moreover, there is experimental information to draw upon: the CDF/Dzero posterior density $p(R|D, m)$, where $R = \sigma / \sigma_{SM}(m)$, is the signal cross section relative to the Standard Model prediction, $\sigma_{SM}(m)$.

$$\int_0^{R_U} p(R \mid D) dR = 0.95$$



http://arxiv.org/abs/0903.4001

A Single Count

Evidence-Based Prior

Given this, and if we are willing to interpret H_1 as the Higgs hypothesis ***for a given Higgs mass***, then it would be reasonable to use the following evidence-based prior

$$\pi(s|\, H_1) = p(R \mid D_{\text{RunII}}, m)$$

for the signal, where $R = s \,/\, \varepsilon\, \sigma_{\text{SM}}$ and ε is the effective integrated luminosity, that is, the integrated luminosity times the signal efficiency, for ATLAS or CMS, assumed known.

In practice, of course, we would need to construct an evidence-based prior for ε also!

A Single Count

Evidence-Based Prior

Finally, suppose we can model the signal prior using

$$\begin{aligned}\pi(s|\,H_1) &= p(R\,|\,D_{RunII}, m)\\ &= \text{Gamma}(R|\,\boldsymbol{a}, \sigma_{SM}^{-1})\\ &= \sigma_{SM}^{-1}\, R^{\mathbf{a-1}} \exp(R)/\Gamma(\boldsymbol{a}),\end{aligned}$$

then the **evidence** for $\boldsymbol{H_1}$ is readily calculated:

$$p(D\,|\,H_1) = \left(\frac{c}{1+c}\right)^{y+1/2} \frac{1}{(1+\varepsilon\sigma_{\text{SM}})^a}\frac{1}{\Gamma(a)}$$

$$\sum_{r=0}^{D} \frac{1}{(1+c)^r}\frac{\Gamma(y+1/2+r)}{\Gamma(y+1/2)r!}\left(\frac{\varepsilon\sigma_{\text{SM}}}{1+\varepsilon\sigma_{\text{SM}}}\right)^{D-r}\frac{\Gamma(D-r+a)}{(D-r)!}$$

A Single Count

Evidence-Based Prior

Now that we have computed the evidences $p(D|H_0)$ and $p(D|H_1)$ for the two hypotheses, we can compute the Bayes factor (at a given Higgs mass)

$$B_{10} = p(D|H_1) \,/\, p(D|H_0)$$

and check if it exceeds an agreed-upon LHC discovery threshold.

But, what if we want to interpret H_1 as the Higgs hypothesis ***regardless of mass***? In this case, we might choose to model $\pi(s|\, H_1)$ as a **hierarchical prior**:

$$\pi(s|\, H_1) = \int \pi(s|\, \boldsymbol{m})\, \pi(\boldsymbol{m}|\, H_1)\, d\boldsymbol{m}$$

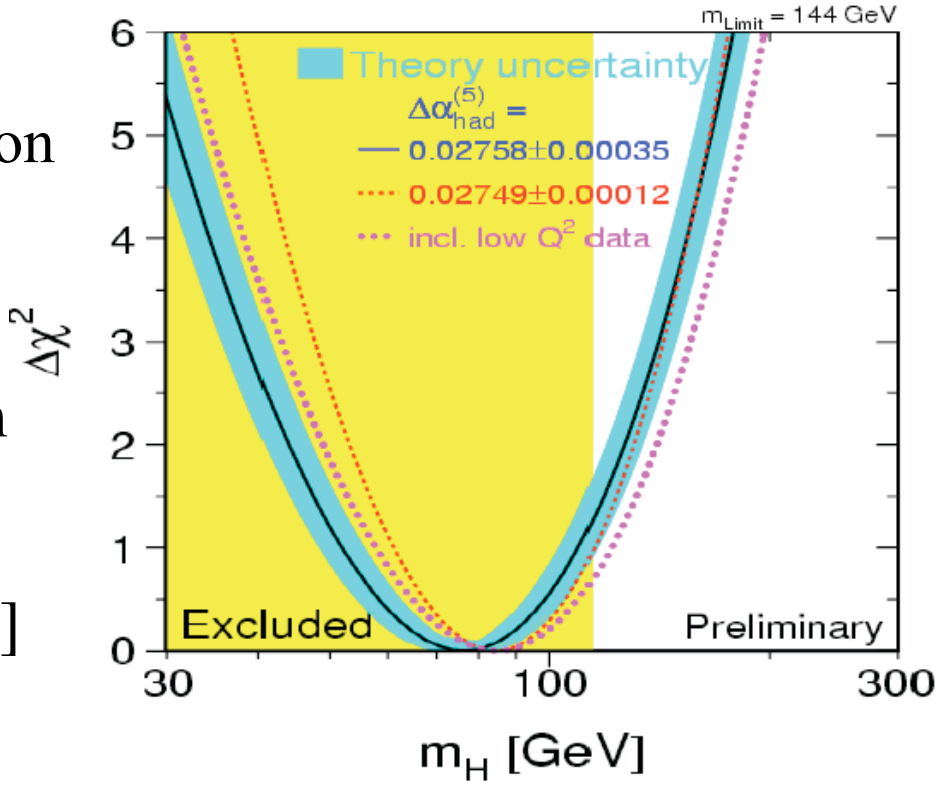
A Single Count

Evidence-Based Prior

This we can do because we have experimental information about the Higgs mass:

A plausible choice for a high mass prior might be

$$\pi(m \mid H_1) \propto \exp[-\Delta\chi^2(m)/2]$$



https://twiki.cern.ch/twiki/bin/view/CMS/HiggsWGConf

Formal Priors

Formal Priors

Suppose in the Higgs search we chose to act as if we are unaware of the beautiful results from LEP and the TeVatron. How might we proceed then to compute Bayes factors?

We would be obliged to produce a prior using a ***formal rule*** of ***our choosing***, but which yields results with good ***properties*** (Berger, Bernardo).

We shall consider four formal rules: one proposed by Jeffreys, one proposed by Berger and Pericchi, one popular in high energy physics and the other proposed by Bernardo.

Jeffreys Priors

The first broadly successful formal prior for ***one***-parameter problems was proposed by **Jeffreys** in the 1930s:

$$\pi(\theta) = \sqrt{F(\theta)}$$

where $F(\theta) = -\int p(x|\theta)\, \partial^2 \ln p(x|\theta)/\partial\theta^2\, dx$ is the Fisher information, where the integration is over the **sample space**.

One of Jeffreys' motivations was to find priors invariant in the following sense. If $z = g(s)$, where g is a one-to-one transformation – and $\pi(z)$ and $\pi(s)$ are calculated using the proposed algorithm, then $\pi(z)dz = \pi(s)ds$ holds, as should be the case for probabilities. For a Poisson distribution, with parameter θ, this algorithm yields $\pi(\theta) = 1/\sqrt{\theta}$

Berger & Pericchi Priors

The **Berger and Pericchi** formal rule:

1. Compute $p(D|H_1) = \int p(D|\, s, H_1)\ \pi_F(s)\ ds$, using a prior $\pi_F(s)$ chosen by a formal rule!
2. Choose the smallest subset of D for which $p(D|H_1) < \infty$
3. Compute $p(s|D, H_1) = p(D|\, s, H_1)\ \pi_F(s) \,/\, p(D|H_1)$
4. Take $\pi(s|\, H_1) = p(s|D, H_1)$ as the signal prior

The idea is to split the data D into a **training sample** to be used to construct the prior $\pi(s|\, H_1)$ and a remaining sample for computing the Bayes factor. The Bayes factor is then ***averaged*** over all possible training samples.

In our Higgs example, we could use simulated data for the training samples.

Berger & Pericchi Priors

A Bayes factor computed with this rule is called an **intrinsic Bayes factor (IBF)**. Berger shows that it yields sensible results.

However, to compute it, we still have the problem of specifying $\pi_F(s)$ from some ***other*** formal rule!

We shall consider two rules:

1. The **flat prior** rule popular in high energy physics
2. The **reference prior** rule of Bernardo

Flat Priors

Consider the value of $p(D|H_1)$ computed using the *flat prior* $\pi_F(s) = 1$. For our Higgs example, it turns out that $D = 0$ is the smallest value for which $p(D|H_1) < \infty$. Indeed, we find

$$p(D \mid H_1) = \int_0^\infty p(D \mid s, H_1)\pi_F(s)ds = \left(\frac{c}{1+c}\right)^{y+1/2}, \quad D = 0$$

Because $p(D=0|H_1)$ is finite, the posterior $p(s|D=0, H_1)$

$$p(s \mid D, H_1) = p(D \mid s, H_1)\pi_F(s) / p(D \mid H_1)$$

$$= \exp(-s), \quad D = 0$$

is **proper**, that is, integrates to one. It can therefore be used as the signal prior $\pi(s| H_1)$ in the calculation of the evidence for hypothesis H_1.

Flat Priors

Setting $\pi(s|H_1) = \exp(-s)$ yields

$$p(D|H_1) = \int_0^\infty p(D|s,H_1)\pi(s|H_1)ds$$

$$= \left(\frac{c}{1+c}\right)^{y+1/2} \sum_{r=0}^{D} \frac{1}{(1+c)^r} \frac{\Gamma(y+1/2+r)}{2^{D-r+1}\Gamma(y+1/2)r!}$$

for the **signal evidence**, which when combined with $p(D|H_0)$ gives the following Bayes factor

$$B_{10} = \frac{p(D|H_1)}{p(D|H_0)}$$

$$= \frac{D!}{\Gamma(y+1/2+D)} \sum_{r=0}^{D} (1+c)^{D-r} \frac{\Gamma(y+1/2+r)}{2^{D-r+1}r!}$$

Flat Priors

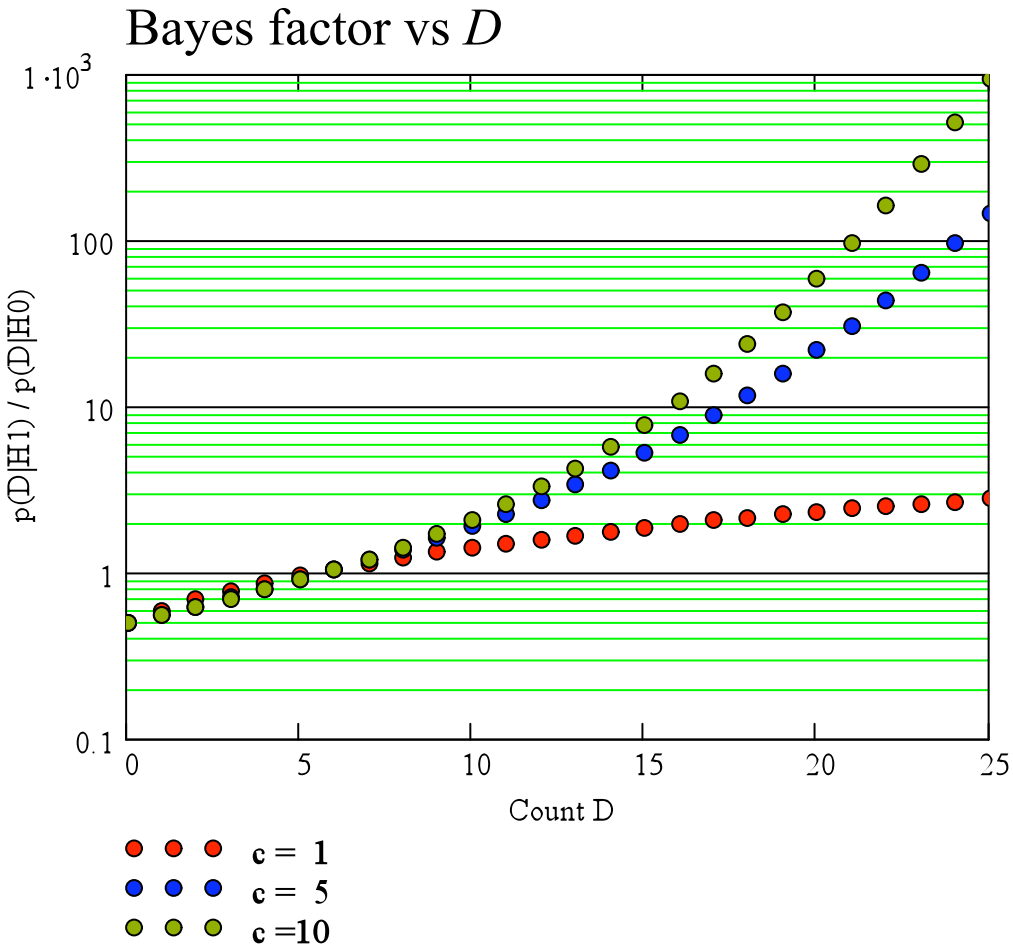
Example:

The figure shows B_{10} as a function of D for

c = 1, 5, 10

y = 5, 25, 50

Note the sensitivity of the Bayes factor to the accuracy with which the background ***b ~ y / c*** is known.



Flat Priors

We have arrived at a seemingly reasonable result. But, consider this. We could have expressed the evidence integral in terms of the variable $z = \ln(s)$, which of course cannot change the value of the integral. Consequently, the Jacobian of the transformation forces the formal prior for z to be $\pi_F(z) = \exp(-z)$.

The unsettling point is that we have not justified why we chose the formal prior to be flat in s rather than in z, or some other variable. Our choice seems arbitrary. That being said, if the event count D is large, as expected at the LHC, the precise form of $\pi_F(s)$ will be less of an issue than when D is small.

Reference Priors

The **Bernardo** formal rule

In 1979, the statistician José Bernardo introduced what proved to be a very successful formal rule for constructing what he called **reference priors**.

His idea was to construct priors which, in a sense to be made precise shortly, contained as little information as possible relative to the probability model. Such priors would be expected to approximate the (impossible!) ideal of "letting the data speak for themselves."

Reference Priors

Reference priors have several desirable properties, including

1. broad applicability
2. invariance, in the sense that $z = g(s)$, implies $\pi(z) = \pi(s)\,|\partial z/\partial s|$, that is, $\pi(z)dz = \pi(s)ds$, where $\pi(z)$ and $\pi(s)$ are reference priors.
3. generation of posterior distributions, which when computed for an **ensemble** of experiments, cluster correctly about the true value of the parameters.
4. avoidance of incoherent inferences, such as **marginalization paradoxes**.

Reference Priors

A **marginalization paradox** is said to occur when the calculation of a posterior can be done in different ways that ought to yield the same answer but do not:

1 $p(D \mid \theta, \phi) \rightarrow p(\theta \mid D)$

$\downarrow$

$p_1(\theta \mid t) \quad t = t(D)$

$\uparrow$

2 $p(t \mid \theta, \phi) = p(t \mid \theta) \rightarrow p_2(\theta \mid t)$

For references priors $p_1(\theta|t) = p_2(\theta|t)$ as it should be.*

*This holds only if one uses different priors

Reference Priors

The reference priors makes use of the notion of **expected intrinsic information**, defined by

$$I_k = \int p(x_k)\, D(p \,||\, \pi_k)\, dx_k$$

that is, it is the expectation of the Kullback-Leibler divergence

$$D(p||\pi_k) = \int p(\theta \,|\, x_k) \ln [p(\theta \,|\, x_k) \,/\, \pi_k(\theta)]\, d\theta$$

between the posterior $p(\theta \,|\, x_k)$ and the prior $\pi_k(\theta)$, where the averaging is with respect to the **marginal density**

$$p(x_k) = \int p(x_k|\theta)\, \pi_k(\theta)\, d\theta$$

I_k measures the amount of information about the value of θ that might be expected from a sequence of k observations $x = x_1,\ldots, x_k$. Note the prior's dependence on k.

Reference Priors

As more and more observations are made, one would expect the amount of information about the value θ to increase.

Reference prior algorithm

Given k observations, I_k is maximized with respect to the prior $\pi_k(\theta)$, thereby maximizing the expected discrepancy between it and the posterior $p(\theta\,|x_k)$.

(Note: if needed, the domain of $\pi_k(\theta)$ may have to be chosen so as to ensure that $\pi_k(\theta)$ is **proper**, that is, integrates to one.)

By definition, the **reference prior**

$$\pi(\theta) = \lim_{k \to \infty} \pi_k(\theta)$$

Reference Priors

This procedure seems, at first, challenging to implement.

Happily, however, Bernardo has given an explicit formula for computing the functions $\pi_k(\theta)$,

$$\pi_k(\theta) = \exp\{\int p(x_k \mid \theta) \ln \left[\frac{p(x_k \mid \theta) h(\theta)}{\int p(x_k \mid \theta) h(\theta) d\theta} \right] dx_k\}$$

where $h(\theta)$ is *any* convenient function, such as $h(\theta) = 1$, and

$p(x_k|\theta) = p(x_1|\theta)\, p(x_2|\theta)\, \ldots$ is the **joint likelihood** for k observations. The functions $\pi_k(\theta)$ are computed for increasing values of k until one obtains convergence.

Reference Priors

CMS Reference Prior Project

Reference priors have not been widely used so far in high energy physics. Consequently, relatively little is known about how they would fair in the problems we face.

However, given their remarkable properties, they are potentially useful to us.

This was the motivation for the **CMS Reference Prior Project** started by Luc Demortier, Supriya Jain and HBP.

We are currently studying the construction of reference priors for the kind of Poisson probability models in common use.

Our goal is to develop sufficient understanding of reference priors so that they can be put to routine use at the LHC.

Reference Prior Project

We are currently studying the following probability model (and its generalization to multiple counts):

background+signal

$$p(D|b,\varepsilon,\sigma) = \text{Poisson}(D|b + \varepsilon\sigma)$$

$$\pi(b,\ \varepsilon,\ \sigma)$$

with **evidence-based priors**

$$\pi(b|y) = \text{Gamma}\,(b|\ \boldsymbol{y}{+}\tfrac{1}{2},\ c) = c(cb)^{\boldsymbol{y}-1/2}\exp(-cb)/\Gamma(\boldsymbol{y}{+}\tfrac{1}{2})$$

$$\pi(\varepsilon|\ x) = \text{Gamma}\,(\varepsilon|\ \boldsymbol{x}{+}\tfrac{1}{2},\ \tau) = \tau(\tau\varepsilon)^{\boldsymbol{x}-1/2}\exp(-\tau\varepsilon)/\Gamma(\boldsymbol{x}{+}\tfrac{1}{2})$$

where $x,\ y,\ \tau$ and c are known constants.

Reference Prior Project

Given the presence of the nuisance parameters, b and ε , there are at two plausible ways one might proceed.

Method 1 (Berger) – factorize the prior as follows

$$\pi(b,\ \varepsilon,\ \sigma) \quad = \pi(\sigma\,|b,\ \varepsilon)\ \pi(b,\ \varepsilon),$$

compute the reference prior $\pi(\sigma\,|b,\ \varepsilon)$ **conditional** on b and ε, then compute the **marginal density**

$$p(D|\ b,\ \varepsilon) = \int p(D|b,\ \varepsilon,\ \sigma)\ \pi(\sigma\,|b,\ \varepsilon)\ d\sigma$$

The **conditional reference prior** $\pi(\sigma\,|b,\ \varepsilon)$ can be computed exactly (Demortier).

Reference Prior Project

Method 2 (Bernardo, Prosper) – factorize the prior as follows

$$\pi(b,\ \varepsilon,\ \sigma) = \pi(b,\ \varepsilon|\ \sigma)\ \pi(\sigma)$$

compute the marginal density

$$p(D|\sigma) = \int p(D|b,\ \varepsilon,\ \sigma)\ \pi(b,\ \varepsilon|\ \sigma)\ db\ d\varepsilon$$

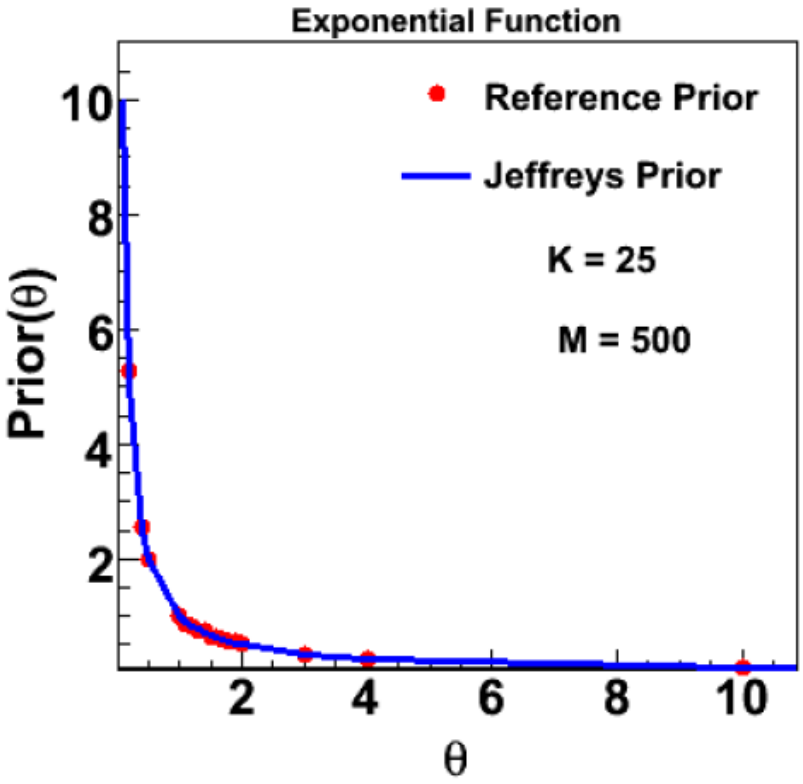
then compute the reference prior $\pi(\sigma)$ for $p(D|\sigma)$, which, in general, must be done numerically (Jain).

The following slides show some preliminary results.

Reference Priors

Test of Numerical Calculation of Reference Priors

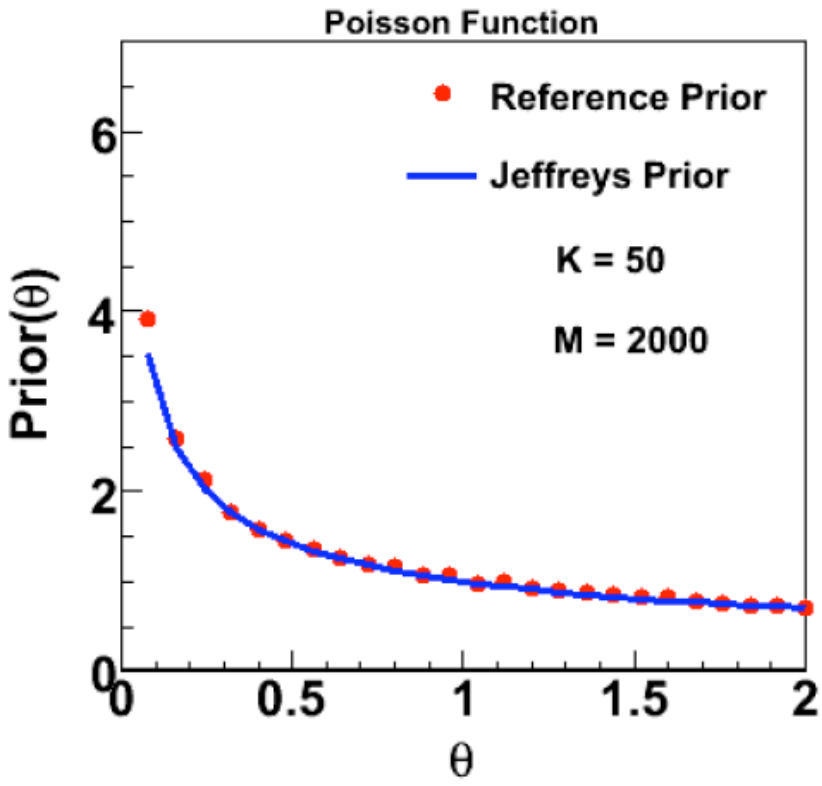
Here we compare our numerical calculation of the reference prior for an **exponential** density with the exact analytical result, $\pi(\theta) \sim 1/\theta$. Bernardo showed that under certain conditions, the prior suggested by **Jeffreys** agrees with the reference prior, as shown in this plot.



Reference Priors

Test of Numerical Calculation of Reference Priors

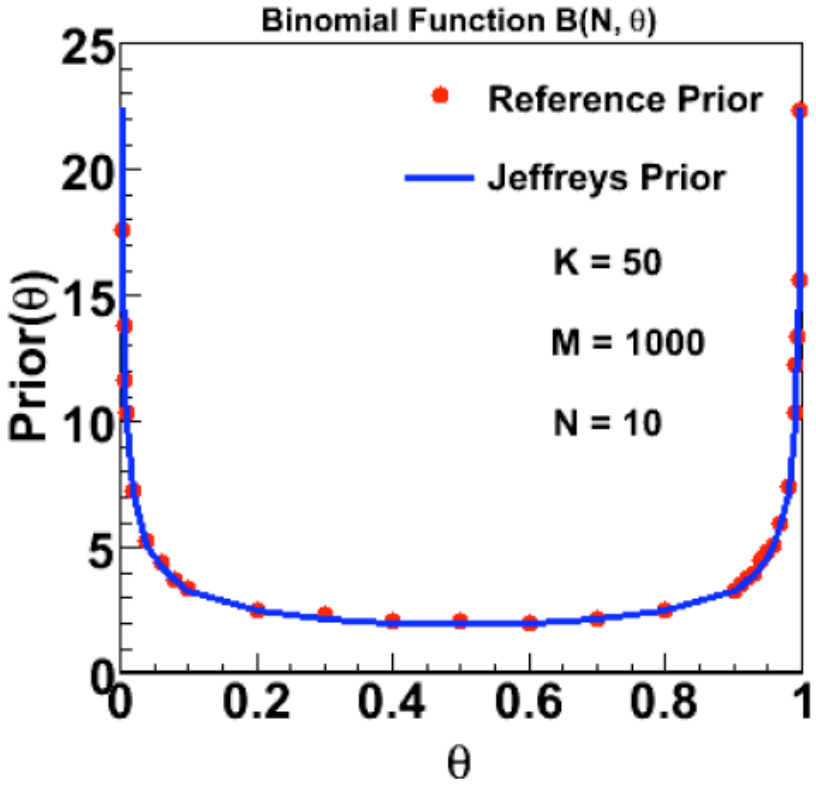
Here we compare our numerical calculation of the reference prior for a **Poisson** distribution with the exact analytical result $\pi(\theta) \sim 1/\sqrt{\theta}$.



Reference Priors

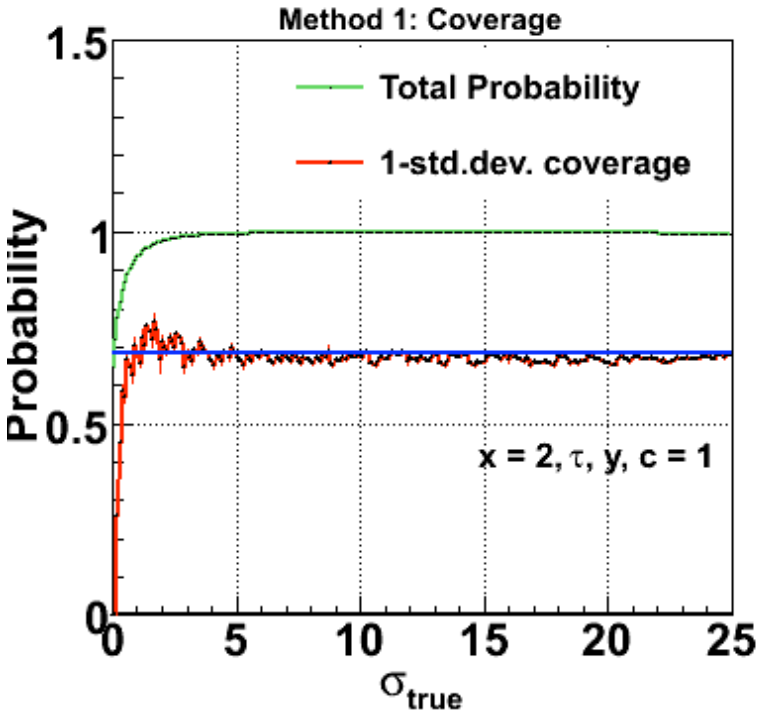
Test of Numerical Calculation of Reference Priors

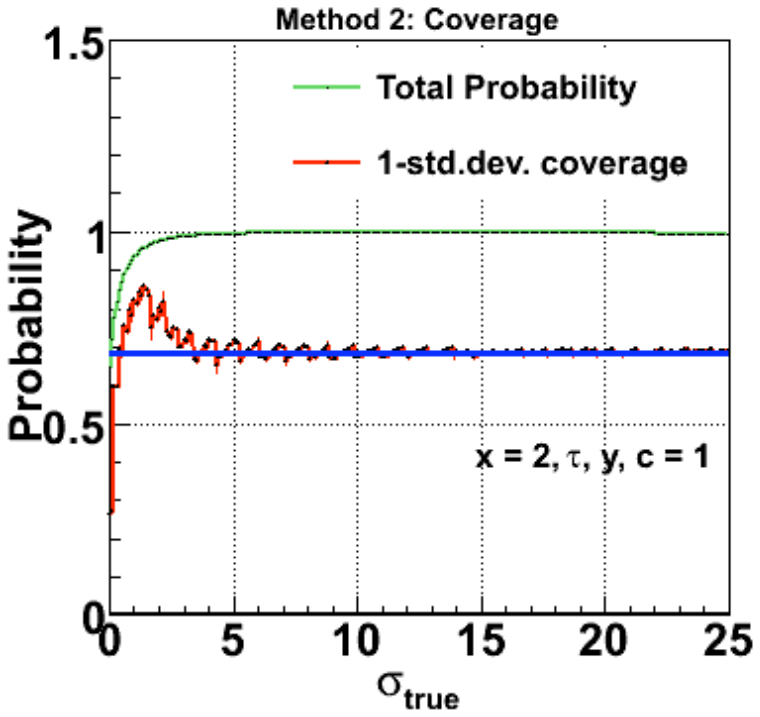
Comparison of our numerical calculation of the reference prior for a **binomial** distribution with the exact result $\pi(\theta) \sim 1/\sqrt{[\theta\,(1-\theta\,)]}$.



Reference Priors

Coverage of intervals





These plot show the coverage probability of 68.3% intervals (***averaged over the priors***) for methods 1 and 2.

Summary

In these lectures I hope to have shown that Bayesian methods are:

1. well-founded,
2. general
3. and powerful.

Moreover, they encourage you to think hard about what you are doing.

However, it is important to understand the methods well in order to minimize the probability of making egregious mistakes…but…

The End

"Have the courage to use your ***own*** understanding!"

Immanuel Kant