Bayesian Statistics in Analysis

Harrison B. Prosper
Florida State University

Workshop on Top Physics:
from the TeVatron to the LHC
October 19, 2007

Outline

- Introduction
- Inference
- Model Selection
- Summary

Introduction

Blaise Pascal
1670

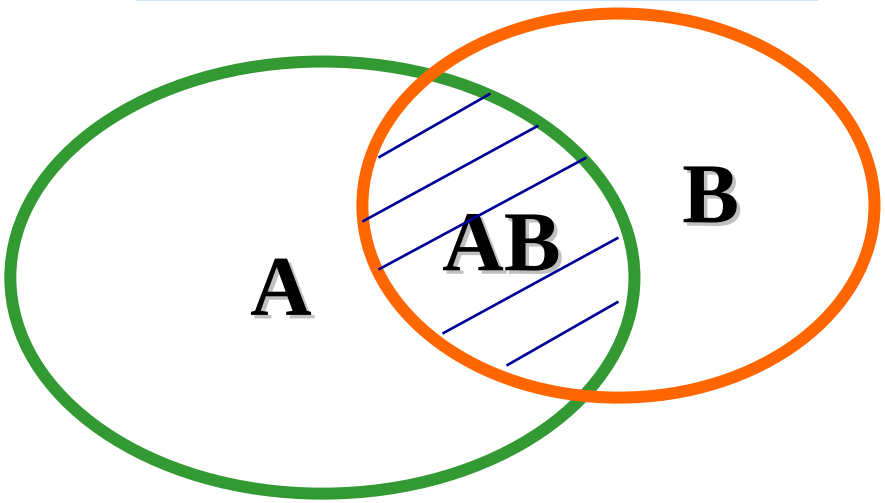
Thomas Bayes
1763

Pierre Simon de Laplace
1812

Introduction

Let $P(A)$ and $P(B)$ be **probabilities**, *assigned* to statements, or events, A and B and let $P(AB)$ be the probability *assigned* to the joint statement AB, then the **conditional probability** of A **given** B is *defined* by

$$P(A \mid B) = \frac{P(AB)}{P(B)}$$



$P(A)$ is the probability of A ***without*** the restriction specified by B.

$P(A|B)$ is the probability of A when we ***restrict*** to the conditions specified by statement B

$$P(B \mid A) = \frac{P(AB)}{P(A)}$$

Introduction

From
we deduce immediately
Bayes' Theorem:

$$P(AB) = P(B \mid A)P(A)$$
$$= P(A \mid B)P(B)$$

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

Bayesian statistics is the application of Bayes' theorem to problems of inference

Inference

Inference

The Bayesian approach to inference is conceptually simple and *always* the same:

Compute

Pr(Data|Model)

Compute

Pr(Model|Data) = Pr(Data|Model) Pr(Model)/Pr(Data)

Pr(Model)	is called the **prior**. It is the probability *assigned* to the Model *irrespective* of the Data
Pr(Data|Model)	is called the **likelihood**
Pr(Model|Data)	is called the **posterior probability**

Inference

In practice, inference is done using the continuous form of Bayes' theorem:

posterior density | likelihood | prior density

$$p(\theta, \lambda \mid D) = \frac{p(D \mid \theta, \lambda)\, \pi(\theta, \lambda)}{\int p(D \mid \theta, \lambda)\, \pi(\theta, \lambda) d\theta d\lambda}$$

θ are the **parameters of interest**

λ denote *all* other parameters of the problem, which are referred to as **nuisance parameters**

marginalization

$$p(\theta \mid D) = \int p(\theta, \lambda \mid D) d\lambda$$

Example – 1

Model

$$n = s + b$$

Prior information

$$\hat{b} \pm \delta b$$

$$0 < s < s_{\max}$$

s is the mean signal count

b is the mean background count

Task: Infer ***s***, given ***N***

Datum

$$D = \{N\}$$

Likelihood

$$P(D \mid s, b) = \text{Poisson}(N, s + b)$$

Example – 1

Apply Bayes' theorem:

posterior likelihood prior

$$p(s,b\,|\,D) = \frac{P(D\,|\,s,b)\,\pi(s,b)}{\iint P(D\,|\,s,b)\,\pi(s,b)\,ds\,db}$$

$\pi(s,b)$ is the prior density for s and b, which *encodes* our prior knowledge of the signal and background means.

The encoding is often *difficult* and can be *controversial*.

Example – 1

First factor the prior

$$\pi(s,b) = \pi(b \mid s)\,\pi(s)$$
$$= \pi(b)\,\pi(s)$$

Define the **marginal likelihood**

$$l(D \mid s) \equiv \int P(D \mid s,b)\,\pi(b)\,db$$

and write the posterior density for the signal as

$$p(s \mid D) = \frac{l(D \mid s)\,\pi(s)}{\int l(D \mid s)\,\pi(s)\,ds}$$

Example – 1

The Background Prior Density

Suppose that the background has been estimated from a Monte Carlo simulation of the background process, yielding ***B*** events that pass the cuts.

Assume that the probability for the count ***B*** is given by $P(B|\lambda) = \text{Poisson}(B, \lambda)$, where λ is the (unknown) mean count of the Monte Carlo sample. We can infer the value of λ by applying Bayes' theorem to the Monte Carlo background experiment

$$p(\lambda \mid B) = \frac{P(B \mid \lambda)\,\pi(\lambda)}{\int P(B \mid \lambda)\,\pi(\lambda)\,d\lambda}$$

Example – 1

The Background Prior Density

Assuming a **flat prior** prior $\pi(\lambda)$ = constant, we find

$$p(\lambda|B) = \text{Gamma}\,(\lambda, 1, B+1) \qquad (= \lambda^B \exp(-\lambda)/B!).$$

Often the mean background count ***b*** in the real experiment is related to the mean count λ in the Monte Carlo experiment linearly, $b = k\,\lambda$, where ***k*** is an accurately known scale factor, for example, the ratio of the data to Monte Carlo integrated luminosities.

The background can be estimated as follows

$$\hat{b} = k\,B, \quad \delta b = k\sqrt{B}$$

Example – 1

The Background Prior Density

The posterior density $p(\lambda|B)$ now serves as the ***prior density*** for the background ***b*** in the real experiment

$$\pi(b) = p(\lambda|B), \text{ where } b = k\lambda.$$

We can write $$l(D\,|\,s) = k\int P(D\,|\,s, k\lambda)\,\pi(k\lambda)\,d\lambda$$

and

$$p(s\,|\,D) = \frac{l(D\,|\,s)\,\pi(s)}{\int l(D\,|\,s)\,\pi(s)\,ds}$$

Example – 1

The calculation of the marginal likelihood yields:

$$l(D \mid s) = \int_{\lambda} P(D \mid s, k\lambda)\, \pi(k\lambda)\, d\lambda$$

$$= \int_0^{\infty} \frac{e^{-(s+k\lambda)} (s+k\lambda)^N}{N!} \frac{e^{-\lambda} \lambda^B}{B!} d\lambda$$

$$= e^{-s} \sum_{r=0}^{N} \frac{s^r}{r!} \frac{k^{N-r}}{(1+k)^{N-r+B+1}} \frac{\Gamma(N-r+B+1)}{(N-r)!B!}$$

Example – 2: Top Mass – Run I

Data partitioned into ***K*** bins and modeled by a sum of ***N*** sources of strength ***p***. The numbers ***A*** are the source distributions for model ***M***. Each *M* corresponds to a different top signal + background model

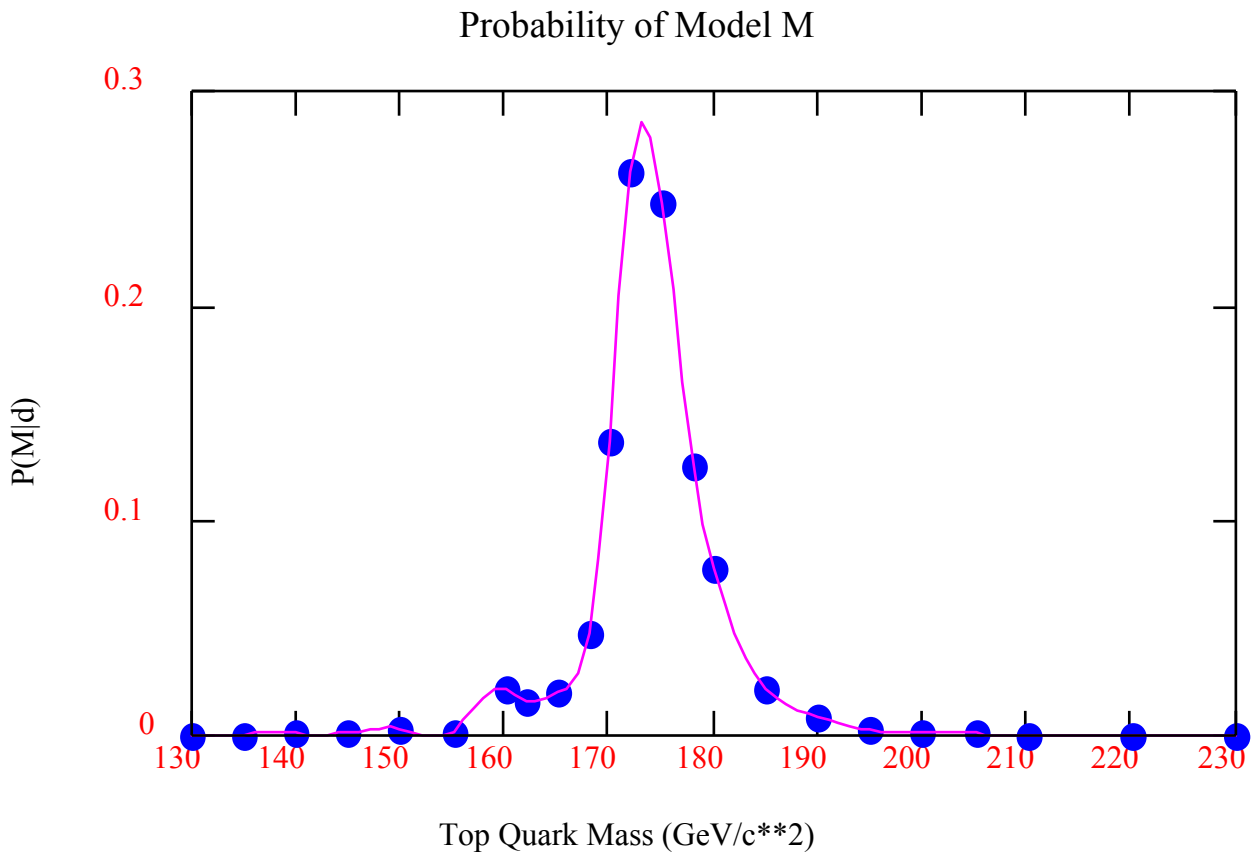
model $$d_i = \sum_{j=1}^{N} p_j a_{ji}$$

likelihood $$P(D \mid a, p, M) = \prod_{i=1}^{K} \exp(-d_i) d^{D_i} \Big/ D_i !$$

prior $$\pi(a, p, M) = \pi(p) \prod_{j=1}^{N} \exp(-a_{ji}) a_{ji}^{A_{ji}} \Big/ A_{ji} !$$

posterior $$P(M \mid D) = \int \cdots \int P(a, p, M \mid D)\, da\, dp$$

Example – 2: Top Mass – Run I



m_{top} = 173.5 ± 4.5 GeV

s = 33 ± 8 events

b = 50.8 ± 8.3 events

To Bin Or Not To Bin

- Binned – Pros
  - Likelihood can be modeled accurately
  - Bins with low counts can be handled exactly
  - Statistical uncertainties handled exactly

- Binned – Cons
  - Information loss can be severe
  - Suffers from the curse of dimensionality

To Bin Or Not To Bin

December 8, 2006 - Binned likelihoods do work!

HIGH-ENERGY PHYSICS

Top quarks go it alone

The first, long-sought evidence for the production
has been reported from a sophisticated analysis o
collisions at the Tevatron.

www.physicstoday.org
physics today
May 2007

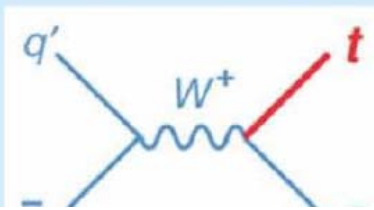
physics update
Suppleme
related to
can be fo
www.phy

SCIENTIFIC AMERICAN

BLACK HOLE BLOWBACK:
Building Galactic Clusters

MARCH 2007
WWW.SCIAM.COM

Unpaired top quarks. Weighing 200 times as much as a proton, the top quark is by far the heaviest elementary particle known. Because the strong nuclear force can't change a quark's flavor, it can produce quarks only in pairs with their antiquarks. The weak force can change flavors. But weak-interaction cross sections are so small that it's almost impossible



PHYSICS

Alone at the Top

CLOSER TO GOD: FERMILAB MAKES SOLO TOP QUARKS BY ALEXANDER HELLEMANS

The world's biggest accelerator, the physics, top quarks may emerge in collisions

To Bin Or Not To Bin

- Un-Binned – Pros
  - No loss of information (in principle)

- Un-Binned – Cons
  - Can be difficult to model likelihood accurately. Requires fitting (either parametric or KDE)
  - Error in likelihood grows approximately linearly with the sample size. So at LHC, large sample sizes could become an issue.

Un-binned Likelihood Functions

Start with the standard binned likelihood over K bins

model $$d_i = a_i\sigma + b_i$$

likelihood $$P(D\,|\,\sigma, a, b) = \prod_{i=1}^{K} \exp(-d_i) d^{D_i} \big/ D_i!$$

$$= \exp(-\sum_{i=1}^{K} d_i) \prod_{i=1}^{K} d^{D_i} \big/ D_i!$$

Un-binned Likelihood Functions

Make the bins smaller and smaller

$$d_i = \int_i d(x)dx \approx [a(x_i)\sigma + b(x_i)]\Delta x_i$$

the likelihood becomes

$$P(D \mid \sigma, A, B) = \exp[-\sum_i \int_i (a(x)\sigma + b(x))dx]$$

where K is *now* the number of events and $a(x)$ and $b(x)$ are the effective luminosity and background densities, respectively, and A and B are their integrals

$$\times \prod_{i=1}^{K} [a(x_i)\sigma + b(x_i)]\Delta x_i$$

$$\propto \exp[-(A\sigma + B)] \prod_{i=1}^{K} [a(x_i)\sigma + b(x_i)]$$

Un-binned Likelihood Functions

The un-binned likelihood function

$$p(D \mid \sigma, A, B) = \exp[-(A\sigma + B)] \prod_{i=1}^{K} [a(x_i)\sigma + b(x_i)]$$

is an example of a *marked* Poisson likelihood. Each event is *marked* by the discriminating variable x_i, which could be multi-dimensional.

The various methods for measuring the top cross section and mass differ in the choice of discriminating variables x.

Un-binned Likelihood Functions

Note: Since the functions $a(x)$ and $b(x)$ have to be modeled, they will depend on sets of modeling parameters α and β, respectively. Therefore, in general, the un-binned likelihood function is

$$p(D \mid \sigma, A, B, \alpha, \beta) = \exp[-(A\sigma + B)] \prod_{i=1}^{K} [a(x_i, \alpha)\sigma + b(x_i, \beta)]$$

which must be combined with a prior density

$$\pi(\sigma, A, B, \alpha, \beta)$$

to compute the posterior density for the cross section

$$p(\sigma \mid D) \propto \int dA \int dB \int d\alpha \int d\beta \; p(D \mid \sigma, A, B, \alpha, \beta)\, \pi(\sigma, A, B, \alpha, \beta)$$

Computing the Un-binned Likelihood Function

If we write $s(x) = a(x)\sigma$, and $S = A\,\sigma$ we can re-write the un-binned likelihood function as

$$p(D \mid S, B) = \exp[-(S+B)] \prod_{i=1}^{K} [s(x_i) + b(x_i)]$$

Since a likelihood function is defined only to within a scaling by a parameter-independent quantity, we are free to scale it by, for example, the *observed* distribution $d(x)$

$$p(D \mid S, B) = \exp[-(S+B)] \prod_{i=1}^{K} \left[\frac{s(x_i) + b(x_i)}{d(x_i)} \right]$$

Computing the Un-binned Likelihood Function

One way to approximate the ratio $[s(x)+ b(x)]/d(x)$ is with a neural network function trained with an admixture of data, signal and background in the ratio 2:1:1.

If the training can be done accurately enough, the network will approximate

$$n(x) = [s(x)+ b(x)]/[\ s(x)+b(x)+d(x)]$$

in which case we can then write

$$p(D\,|\,S,B) = \exp[-(S+B)]\prod_{i=1}^{K}\left[\frac{n(x_i)}{1-n(x_i)}\right]$$

Model Selection

Model Selection

Model selection can also be addressed using Bayes' theorem. It requires computing

posterior evidence prior

$$P(M \mid D) = \frac{p(D \mid M)\, P(M)}{p(D)}$$

where the **evidence** for model ***M*** is defined by

$$p(D \mid M) = \int p(D \mid \theta_M, \lambda_M, M) \times \pi(\theta_M, \lambda_M \mid M)\, d\theta_M\, d\lambda_M$$

Model Selection

posterior odds | Bayes factor | prior odds

$$\frac{P(M \mid D)}{P(N \mid D)} = \left(\frac{p(D \mid M)}{p(D \mid N)} \right) \frac{P(M)}{P(N)}$$

The **Bayes Factor**, B_{MN}, or any one-to-one function thereof, can be used to choose between two competing models M and N, e.g., signal + background versus background only.

However, one must be careful to use *proper* priors.

Model Selection – Example

Consider the following two prototypical models

Model 1 $$P(D \mid s,b) = \text{Poisson}(N, s+b), \quad \pi(s,b)$$

Model 2 $$P(D \mid b) \quad = \text{Poisson}(N, b), \quad \pi(b)$$

The Bayes factor for these models is given by

$$B_{12} = \frac{P(D \mid 1)}{P(D \mid 2)} = \frac{\int \text{Poisson}(N, s+b)\, \pi(s,b)\, ds db}{\int \text{Poisson}(N, b)\, \pi(b)\, db}$$

Model Selection – Example

Calibration of Bayes Factors

Consider the quantity (called the Kullback-Leibler divergence)

$$k(2 \,||\, 1) = \int P(D \,|\, 1) \ln \frac{P(D \,|\, 1)}{P(D \,|\, 2)} \, dD$$

For the simple Poisson models with *known* signal and background, it is easy to show that

$$k(2 \,||\, 1) = -s + (s + b) \ln \left(1 + \frac{s}{b} \right)$$

For s << b, we get $\sqrt{k}(2||1) \approx s/\sqrt{b}$. That is, roughly speaking, for s << b, $\sqrt{\ln B_{12}} \approx s/\sqrt{b}$

Summary

Bayesian statistics is a well-founded and general framework for *thinking* about and *solving* analysis problems, including:

- Analysis design
- Modeling uncertainty
- Parameter estimation
- Interval estimation (limit setting)
- Model selection
- Signal/background discrimination etc.

It well worth learning how to *think* this way!