Bayesian Methods: Theory and Practice
Lecture 2 – Foundations & Applications

Harrison B. Prosper

Florida State University

CMS Statistics Committee

7 August, 2008

Outline

- Lecture 2 – **Foundations & Applications**
  - The Bayesian Approach – Recap
  - Decisions & Loss
  - Hypothesis Tests
    - Bayes Factors
    - A Single Count
  - Summary

The Bayesian Approach – Recap

Recap

In a Bayesian analysis, the basic elements are *always* the same:

$p(\boldsymbol{D} \mid \omega)$ **probability model / likelihood**

data $\boldsymbol{D}$ model parameters ω.

$\pi(\omega)$ **prior probability density / prior**

$p(\omega \mid \boldsymbol{D})$ **posterior density / posterior**

The model parameters can be continuous, discrete, or both.

Recap

And *always* combined the same way:

posterior likelihood prior

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

and, *in principle*, the restriction to the parameters of interest is *always* done the same way:

marginalization

$$p(\theta \mid D) = \int_{\Phi} p(\theta,\phi \mid D)\,d\phi$$

Decisions & Loss

Decisions and Loss

The posterior density $p(\theta \mid D)$ is the **complete** answer to an inference about the parameter θ.

However, it is often of interest to extract a useful summary of it, such as a **point estimate** θ^* and/or an **interval estimate** (θ_L, θ_U).

Or, perhaps, we wish to decide which of two or more competing models is preferred by the data.

Decisions and Loss

Decision theory provides a general way to model such problems.

To render a decision about the value of θ – implemented as a function ***d*** that returns an **estimate** θ^* of θ – we must quantify how beneficial the decision is to us, via a **utility function** $U(d, \theta)$ that ***we*** specify. A function ***d*** that returns estimates is called an **estimator**.

Or, equivalently, we can specify a **loss function** $L(d, \theta)$ that quantifies what we lose should the decision turn out to have been a bad one.

Pascal's Wager

Pascal was one of the first mathematicians to make explicit use of what we now call utility. In 1670, he considered the following two hypotheses:

God **God exists**

~God **God does not exist**

and the following two actions:

P **Lead a pious life**

W **Lead a worldly life**

and assigned ***utilities*** to each hypothesis/action pair.

Pascal's Wager

	God	~ God
P	$+\infty$ (eternal bliss!)	− (no worldly pleasures)
W	+ (worldly pleasures) $-\infty$ (eternal damnation!)	+ (worldly pleasures)

He argued that if your **Pr**(God) is strictly > 0 then your **expected utility** from being pious is so very much greater than your expected utility from being worldly, that the only rational option for you is to live a saintly life!

Decisions and Loss

In practice, since our knowledge of the parameter θ is encoded in the posterior density $p(\theta \mid D)$, our decisions will be more **robust** if we average (E[*]) the loss $L(d, \theta)$ with respect to $p(\theta \mid D)$

$$\begin{aligned} R(d) &= \mathrm{E}[L(d, \theta)] \\ &= \int L(d, \theta)\, p(\theta \mid D)\, \mathrm{d}\theta \end{aligned}$$

The quantity $R(d)$ is called the **risk function**. *By definition*, the **optimal estimate** of θ is the one that minimizes the risk

$$\theta^* = \arg\min_d R(d)$$

Comments

In general, different loss functions will yield different estimates. Therefore, even with ***exactly the same data*** one should not be surprised to obtain competing results.

☺Moreover, to the degree that the mathematics has been done correctly, none of the competing results is wrong. Each merely enjoys a different set of properties.

Reasonable people can disagree about the results simply because they disagree about what properties are thought most useful. ***Some properties can even be in conflict…***

Comments

Example 4: Consider a loss function $L(d, m)$ to extract a **point estimate** of the Higgs mass, m, from a posterior density $p(m|D)$.

Suppose $L(d, m)$ is invariant in the sense that it yields an estimate m^* of m which, when inserted into the prediction for the Higgs production cross section $\sigma = g(m)$, yields an estimate of the cross section $\sigma^* = g(m^*)$ that is identical to the one we would have obtained had we used the loss function $L(d, \sigma)$.

$L(d, \sigma)$ is the loss function $L(d, m)$ but with m replaced by σ. In general, either m^* or σ^* or both will be **biased**.

Comments

Example 4:

To see this, expand $\sigma^* = g(m^*)$ about the ***true*** Higgs mass m

$$\sigma^* \approx g(m) + (m^* - m)\, g' + \tfrac{1}{2}\,(m^* - m)^2\, g''$$

and average both sides over an **ensemble** of estimates. This gives

$$E[\sigma^*] \approx \sigma + \textbf{bias}\; g' + \tfrac{1}{2}\textbf{mse}\; g'',$$

$$E[\sigma^*] \approx \sigma + \textbf{bias}\; g' + \tfrac{1}{2}[\textbf{bias}^2 + \textbf{variance}]\; g'',$$

where **variance** $= E[m^{*2}] - E[m^*]^2$.

mse: mean square error (note: **rms** $= \sqrt{\textbf{mse}}$)

Comments

Example 4: So, invariance and lack of bias are in conflict.

☺ However, bias should not be considered a problem provided that the estimates are **accurate**, that is, they lie as close to the true value of the parameter as possible. One measure of closeness is the **mean square error**

$$\textbf{mse} = \textbf{bias}^2 + \textbf{variance}$$

Historically, zero bias has been considered to be very important in high energy physics. It could be argued, however, that invariance is the more important property: the ability to use $\sigma^* = g(m^*)$ to go from m^* to σ^* is better because the loss function is the same for both, indeed for ***all***, estimates.

Decisions and Loss

Point Estimation

quadratic loss

$L(d, \theta) = (d - \theta)^2$

Average with respect to $p(\theta|D)$

risk $R_\theta(d) = \mathrm{E}[(d-\theta)^2]$

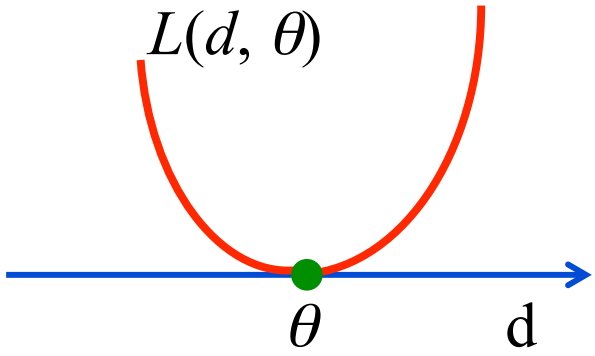
$= \mathrm{E}[d^2] - 2\mathrm{E}[\theta d] + \mathrm{E}[\theta^2]$

$= d^2 - 2\mathrm{E}[\theta]d + \mathrm{E}[\theta^2]$

minimize

$dR/dd = 2d - 2\mathrm{E}[\theta] = 0$

so, $\theta^* = \mathbf{E}[\theta]$



Note: quadratic loss is ***not invariant***. If $\alpha = \mathrm{g}(\theta)$, then

$L(d, \alpha) = (d - \alpha)^2$ gives

$\alpha^* = \mathrm{E}[\alpha] = \mathrm{E}[\mathrm{g}(\theta)]$

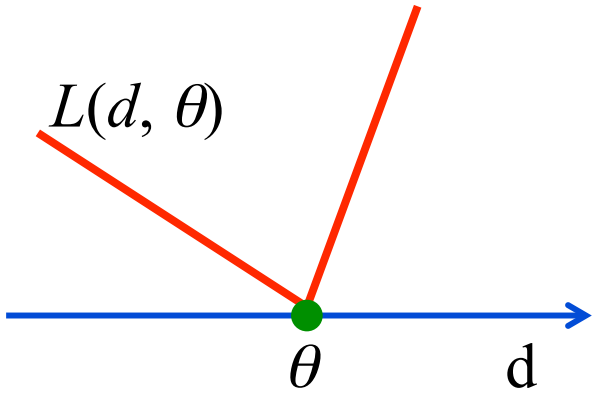
$\neq \mathrm{g}(\theta^*) = \mathrm{g}(\mathrm{E}[\theta])$

Decisions and Loss

Point Estimation

bilinear loss

$$L(d,\theta) = \begin{cases} a(\theta - d), & d < \theta \\ b(d - \theta), & d \geq \theta \end{cases}$$



risk $$R(d) = a\int H(\theta - d)(\theta - d)p(\theta \mid D)d\theta$$

$$+ b\int H(d - \theta)(d - \theta)p(\theta \mid D)d\theta$$

$H(x) = 1$ if $x > 0$ else 0

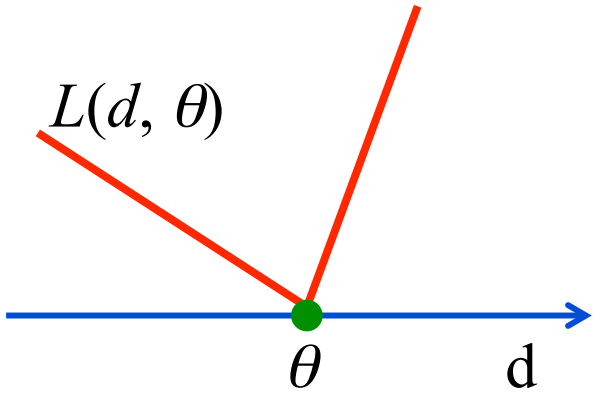
Decisions and Loss

Point Estimation

bilinear loss

The optimal estimate is

$$\theta^* = \arg\min_d R(d)$$



where θ^* is the $a/(a+b)$ **quantile**

$$\int_{\theta^*} p(\theta \mid D) d\theta = a/(a+b)$$

of $p(\theta \mid D)$. If we set $a = b$, θ^* = **median** of $p(\theta \mid D)$

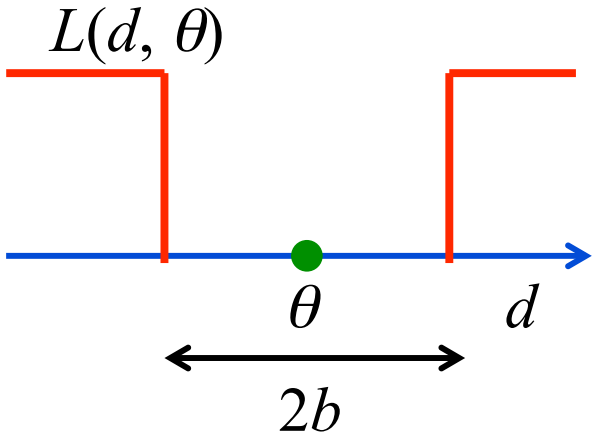
Note: estimates based on quantiles are ***invariant***.

Decisions and Loss

Point Estimation

zero-one loss

$$L(d,\theta) = \begin{cases} 0, & |d-\theta| \le b \\ 1, & |d-\theta| > b \end{cases}$$



Its risk function is

$$R(d) = \int [H(\theta - b - d) + H(d - \theta - b)]\, p(\theta \mid D)\, d\theta$$

and the optimal estimate $\theta^* = \min_d R(d)$ is the solution of

$p(\theta^* + b|D) = p(\theta^* - b|D)$.

In the limit $b \to 0$, one obtains θ^* = **mode** of $p(\theta \mid D)$. The mode is ***not invariant***.

Decisions and Loss

Probability Estimation

Consider a variable θ that can assume one of several values $\theta_1, \ldots \theta_k$.* The problem is to associate a *probability* d_i to each of the possible values of θ. Let $L(d_i) = L(d_i, \theta_i)$ be the corresponding loss function and denote by $p_i = p(\theta_i)$ the *unknown* optimal choice (it could be, for example, the posterior distribution, $p(\theta_i | D)$).

The risk associated with our decisions is given by

$$R(d) = \sum_{i=1}^{k} L(d_i, \theta_i)\, p(\theta_i)$$

Again, we find the optimal estimates by minimizing the risk.

* see O'Hagan, 2.54, p56

Decisions and Loss

Probability Estimation

But, the minimization must obey the constraint $\sum d_i = 1$, which is readily implemented using a Lagrange multiplier:

$$F = \sum_{i=1}^{k} L(d_i) p_i + \lambda \sum_{i=1}^{k} d_i$$

Suppose that our loss function is such that at the absolute minimum of the risk, our estimates of the probabilities are optimal, that is, $d_i = p_i$. In this case, we can write

$$\frac{\partial F}{\partial d_i} = \frac{dL}{dd_i} d_i + \lambda = 0, \quad \rightarrow \quad L(d_i) \propto -\ln d_i$$

We conclude that in the neighborhood of minimum risk, our loss function must be ***logarithmic*** in the density.

Decisions and Loss

Probability Estimation

The argument can be extended to continuous densities, with the risk given by

$$R(d) = -\int p(\theta) \ln d(\theta) d\theta$$

As noted, its minimum occurs when $d(\theta) = p(\theta)$, that is, when the minimum value of the risk is

$$\min_d R = -\int p(\theta) \ln p(\theta) d\theta \equiv H(p)$$

that is, when the risk equals the **entropy** of the density.

Decisions and Loss

Probability Estimation

We can make the risk zero at its minimum by subtracting off the entropy, ***H(p)***. This yields another important quantity, called the **Kullback-Leibler (K-L) divergence**

$$D(p \,||\, d) = R(d) - H(p)$$
$$= \int p(z) \ln \frac{p(z)}{d(z)} dz$$

where the integration can be over the **parameter space** or the **sample space** depending on the application. This widely used, ***non-negative***, measure of the "distance" between two densities ***p*** and ***d*** is zero ***if and only if*** the densities are identical. We shall return to it in Lecture 3.

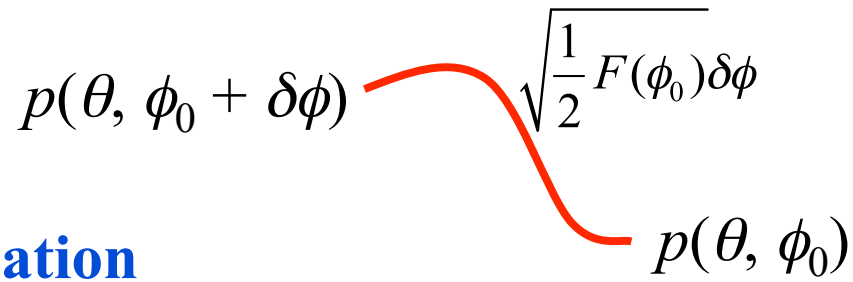
Decisions and Loss

Probability Estimation

The word "distance" is in quotes because $D(p||d) \neq D(d||p)$.

However, if p and d are close together in the sense that $p = p(\theta, \phi_0 + \delta\phi)$ and $d = p(\theta, \phi_0)$, then $D(p||d)$ can be interpreted as a **metric** in the space of probability densities:

$$D(p \,||\, d) \approx \frac{1}{2} F(\phi_0)(\delta\phi)^2$$



where $F(\phi)$ is the **Fisher information**

$$F(\phi) = -\int p(z,\phi) \frac{\partial^2 \ln p(z,\phi)}{\partial \phi^2} dz$$

Again, the integration can be either over the parameter space or the sample space

Comments

Although we have arrived at the concepts of **entropy**, **K-L divergence** and **Fisher information** through our consideration of probability estimation, these quantities play a much broader role in many disciplines and have numerous interesting applications. Here is one.

Paris and his counterpart in ATLAS have urged every group to design analyses optimal for discovery. But what *precisely* do they mean by this? Perhaps they mean the following: ***an analysis is optimal for discovery if it maximizes the "distance" between the background + signal and background-only models, where "distance" is the K-L divergence.***

Example

Consider a simple **counting experiment**, described by

H_0: background-only

$$p(D|b) = \text{Poisson}(D|b)$$

H_1: background+signal

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

where, for simplicity, we assume the mean signal and background counts ***s*** and ***b***, respectively, are known. Our goal is to design an analysis for which the densities ***p(D|b)*** and ***p(D|b, s)*** are as "far apart" as possible, in the sense of the K-L divergence (where the integration is now over the ***sample space***):

$$D(p_{b+s} \,||\, p_b) = \sum_{D=0}^{\infty} p(D \,|\, b+s) \ln \frac{p(D \,|\, b+s)}{p(D \,|\, b)}$$

Example

For this model we get

$$D(p_{b+s} \,||\, p_b) = \sum_{D=0}^{\infty} p(D \,|\, b+s) \ln \frac{p(D \,|\, b+s)}{p(D \,|\, b)}$$

$$= -s + (b+s)\ln(1+s/b)$$

which should look familiar to many of you!

One can get a bit of insight into this expression by considering the "search" limit ***s << b***:

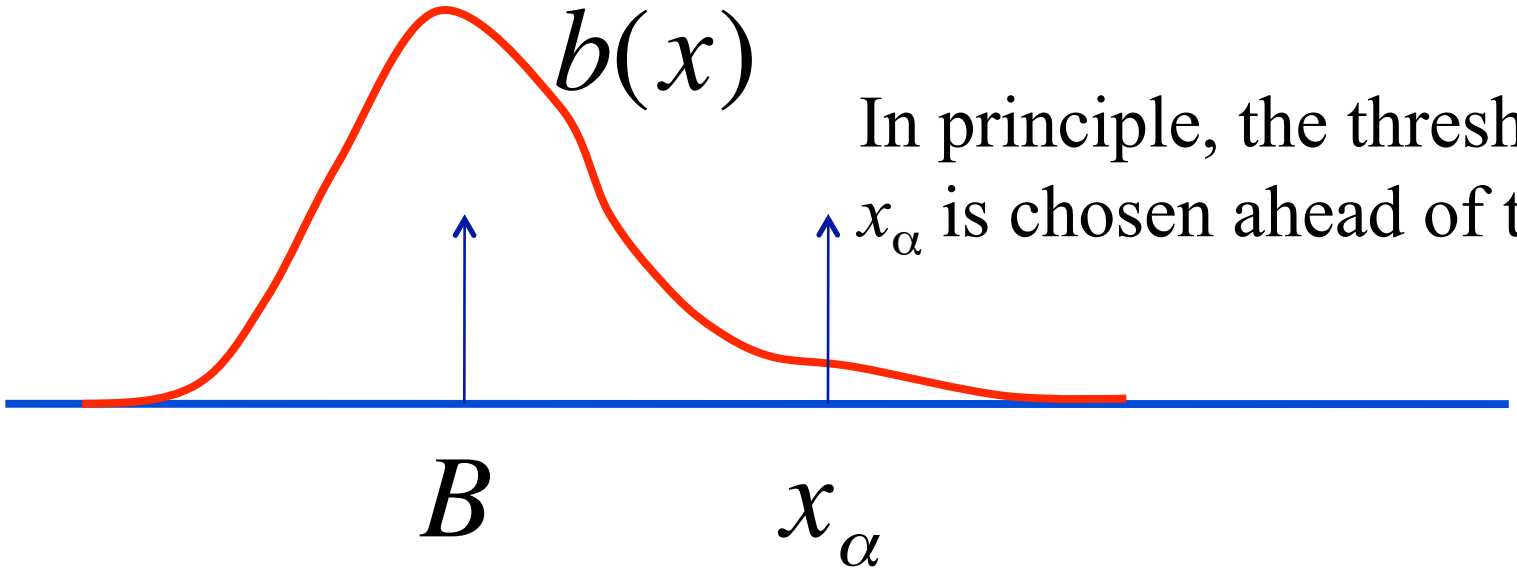
$$D(p_{b+s} \,||\, p_b) \approx -s + (b+s)\left(\frac{s}{b} - \frac{1}{2}\frac{s^2}{b^2} + \ldots\right) = \frac{1}{2}\frac{s^2}{b}$$

This suggests that $\sqrt{[2D(p_{s+b}||p_b)]}$ is a generalization of the well-worn, but oft-abused, discovery favorite $s/\sqrt{b}$.

Hypothesis Testing

In HEP & Biology

Null hypothesis (H_0): background-only



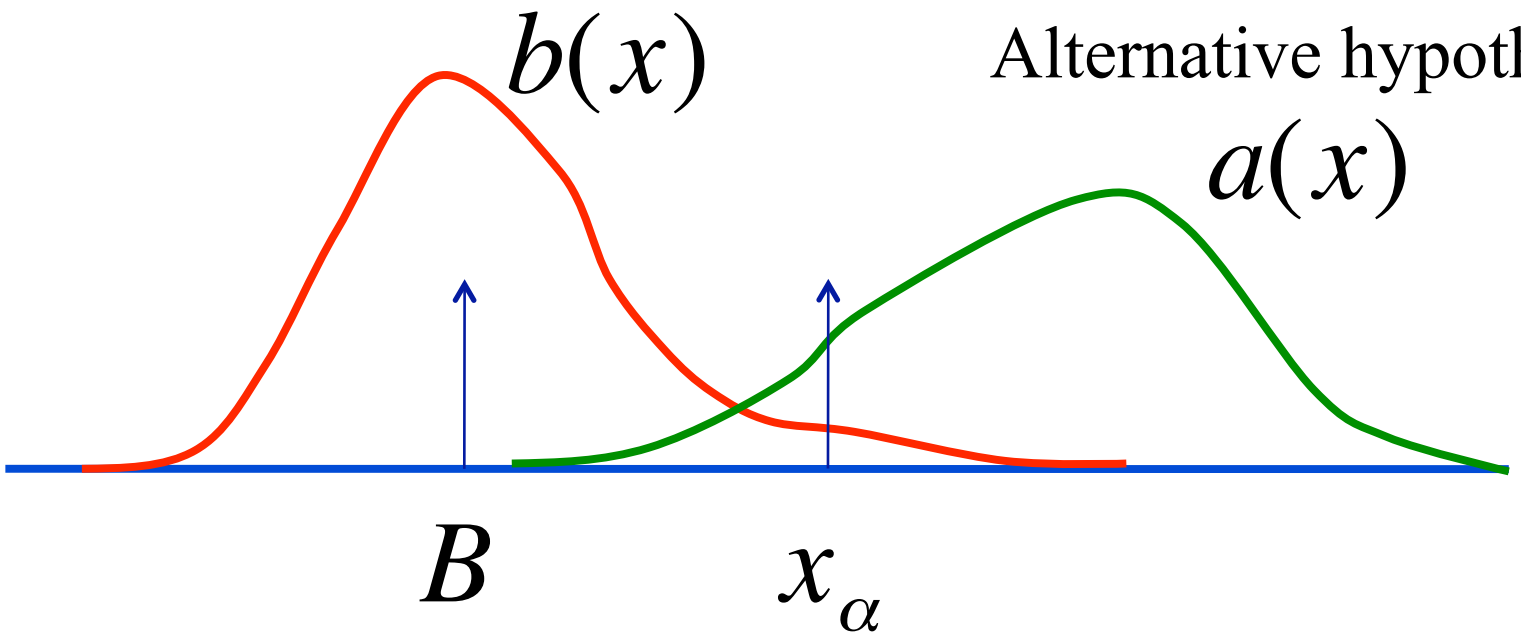
In principle, the threshold x_α is chosen ahead of time

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

significance level

HEP – 2.7×10^{-7}
BIO – 0.05

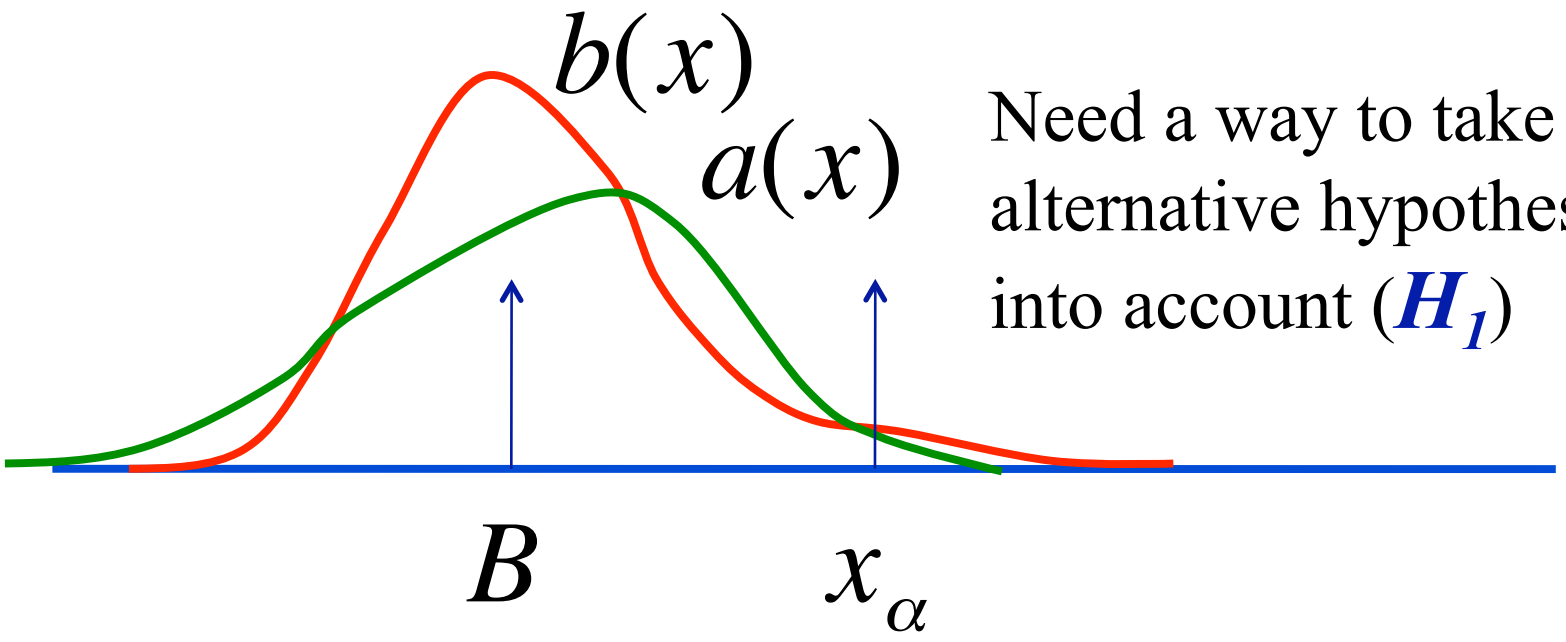
In HEP & Biology



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

HEP – 2.7 x 10^{-7}
BIO – 0.05

In HEP & Biology



Need a way to take alternative hypotheses into account (***H*** $_1$)

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

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

power

Hypothesis Testing
from a Bayesian Viewpoint

Hypothesis Testing

Unlike Fisher's method of hypothesis testing*, in the standard Bayesian approach it is ***always*** necessary to compare one hypothesis, or model, against at least one alternative.

In particular, a **goodness of fit (gof) test** – which determines whether ***to reject an hypothesis*** (called the **null hypothesis**, often denoted by the symbol, H_0) without the explicit specification of alternatives – does not exist in the Bayesian approach.**

*which uses p-values, as in a χ^2-test

**The closest thing to this is a Bayesian test of a null hypothesis against a non-parametric alternative, developed by statistician Jim Berger.

Hypothesis Testing

Conceptually, Bayesian hypothesis testing proceeds in exactly the same way as any other Bayesian calculation: compute the posterior density

posterior likelihood prior

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

and marginalize it with respect to all parameters except those indexing the hypotheses

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

and thereby get your ***pHD***!

Hypothesis Testing

However, it is usually more convenient, and instructive, to arrive at ***p(H|D)*** in stages.

1. Factorize the priors: $\pi(\theta, \phi, H) = \pi(\theta, \phi|H)\ \pi(H)$

2. Then, for each hypothesis, H, compute the function

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

3. Then, compute the **probability of each hypothesis, *H***

$$p(H \mid D) = \frac{p(D \mid H)\pi(H)}{\sum_H p(D \mid H)\pi(H)}$$

Bayes Factors

It is clear, however, that to compute $p(H|D)$, it is necessary to specify the priors $\pi(H)$. Unfortunately, there is unlikely to be a consensus about what their values should be.

So, instead of asking for the probability of an hypothesis, $p(H|D)$, we could ask: how much more probable is one hypothesis H_1 than another H_0?

$$\frac{p(H_1 \mid D)}{p(H_0 \mid D)} = \left[\frac{p(D \mid H_1)}{p(D \mid H_0)}\right] \frac{\pi(H_1)}{\pi(H_0)}$$

The ratio in brackets, called the **Bayes factor**, B_{10}, is the amount by which the **odds** in favor of H_1 has increased, given the observations.

Bayes Factors

From the way it appears in the Bayes factor

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

it is natural to interpret ***p(D | H)*** as the **evidence** for hypothesis ***H***. The larger the evidence the better the match between hypothesis and data.

This is all very elegant, but, Bayes factors come with a serious health warning: they can be very sensitive to the choice of priors $\pi(\theta, \phi|H)$; therefore, the latter should be carefully chosen ***probability*** densities: $\int \pi(\theta, \phi|H) \mathrm{d}\theta\, \mathrm{d}\phi = 1$

A Single Count

To illustrate Bayesian hypothesis testing, we shall use the **nested model*** described earlier

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 model (H_0) is nested if is a special case of another (H_1).

Comments

- ☺ In principle, the Bayesian approach is straightforward because the procedure is always the same: compute the posterior density of the model parameters.

- ☺ Moreover, the words we use to describe uncertainties, **statistical**, **systematic**, **theoretical**, **best guess**, **gut feeling**, etc., are totally irrelevant from a Bayesian viewpoint because all forms of uncertainty are handled in the same way.

In practice, however, a fully Bayesian analysis can be extremely challenging.

Comments

Why? Firstly, because one has to construct a detailed model of *both* the ***likelihood*** and the ***prior*** that captures what we know about a particular problem.

Secondly, because the use of a detailed model may entail a daunting computational challenge.

Thirdly, because no model is perfect, we should in principle cycle through the entire procedure a few times, varying those parts of the model about which we are least certain in order to check the ***robustness*** of our answers.

Even for a counting experiment, there is much to think about!

A Single Count

We need to decide how to factor our priors. Let's try:

$$\pi(b, H_0) = \pi(b|\, H_0)\, \pi(H_0)$$

$$\pi(b, s, H_1) = \pi(b, s|H_1)\, \pi(H_1)$$

$$= \pi(b|\, s, H_1)\, \pi(s|H_1)\, \pi(H_1)$$

Points to note

1. Any joint probability can be factored in different ways and each way is valid.
2. Consequently, the background prior is, in principle, ***conditioned*** on the signal, and *vice versa*. Therefore, it is an ***assumption*** to assert they are not.

A Single Count

Next, compute the **evidence** for each hypothesis

$$p(D|H_0) = \int \text{Poisson}(D|\boldsymbol{b}, H_0)\, \pi(\boldsymbol{b}|\, H_0)\, d\boldsymbol{b}$$

$$p(D|H_1) = \int\int \text{Poisson}(D|\boldsymbol{b}, s, H_1)\, \pi(\boldsymbol{b}|\, s, H_1)\pi(s|H_1)\, d\boldsymbol{b}\, ds$$

$$= \int p(D|s, H_1)\pi(s|H_1)\, ds$$

where $p(D|s, H_1) = \int \text{Poisson}(D|\boldsymbol{b}, s, H_1)\, \pi(\boldsymbol{b}|s, H_1)\, d\boldsymbol{b}$.

However, since H_0 is nested in H_1, its evidence, $p(D|H_0)$, is simply the function $p(D|s, H_1)$, evaluated at $\boldsymbol{s = 0}$, assuming that

$$\pi(\boldsymbol{b}|s, H_1) = \pi(\boldsymbol{b}|\, H_1) = \pi(\boldsymbol{b}|\, H_0) = \pi(\boldsymbol{b})$$

A Single Count

Modeling the Background Prior $\pi(b)$

We assume that the background is to be estimated from a Monte Carlo (MC) simulation that yielded y events passing an appropriate set of cuts, or from sideband data.

If y is an integer that is much smaller than the original sample size it would be reasonable to adopt the probability model $p(y|\, cb) = \text{Poisson}(y|\, cb)$, where c is a known scale factor between the background in the MC sample and that of the real sample.

Bayes' theorem will yield a background prior* of the form

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

*Assuming MC prior ~$1/\sqrt{b}$

A Single Count

Modeling the Background Prior $\pi(b)$

In practice, MC events will be **weighted** because of the need to make the MC distributions match the observed ones better. Then ***y*** will be a **weighted sum**.

Even so, the gamma model $\pi(b) = \text{Gamma}\,(b|\, y{+}\tfrac{1}{2}, \text{c})$,

or perhaps one comprising a mixture of them with appropriate values for ***y***, may still be good enough to model prior beliefs about the background.

Priors that are obtained through thorough inspection of the problem domain are called **subjective priors**. A better name is **evidence-based priors** (Sir David, PhyStat-LHC).

A Single Count

The **Prior Predictive Distribution** is given by the integral:

$$p(D \mid s, H_1) = \int_0^\infty \text{Poisson}(D \mid b+s)\pi(b)db$$

$$= \int_0^\infty \frac{(b+s)^D \exp[-(b+s)]}{D!} \frac{c(cb)^{y-1/2}\exp(-cb)}{\Gamma(y+1/2)} db$$

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

The background-only ***evidence*** is then

$$p(D|\, H_0) = p(D|s=0, H_1)$$

A Single Count

The **Background+Signal Evidence** is given by the integral:

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

$$= \sum_{r=0}^{D} C_r \int_0^\infty \text{Poisson}(D - r \mid s)\pi(s)ds$$

where $C_r = \left(\frac{c}{1+c}\right)^{y+1/2} \frac{1}{(1+c)^r} \frac{\Gamma(y+1/2+r)}{\Gamma(y+1/2)r!}$

and $\pi(s) = \pi(s|H_1)$ is the prior for the ***signal***…

…and this is where things become controversial!

Summary

- **Decision Theory**
  - The basic insight is that optimal decision making entails combining a ***utility function*** or, equivalently, a ***loss function*** with a ***posterior density***. Since loss functions can differ, it is unsurprising that results can differ even when using the same data.

- **Hypothesis Tests**
  - The standard Bayesian approach requires the explicit consideration of at least one alternative hypothesis. Ironically, this was also the opinion of arch-frequentist Jerzy Neyman!

Summary

- **Hypothesis Tests**
  - It is necessary to specify priors for each of hypothesis.
  - In particular, for our simple counting experiment, we need to specify the prior $\pi(s)$ for the signal since it is part of the specification of the background+signal hypothesis.
  - Unfortunately, doing so ***sensibly*** is hard!

More tomorrow!