Analysis Techniques
Probability & Inference

Harrison B. Prosper

NEPPSR 2007

Exercise 587:
Prove this

Outline

- Introduction
- Descriptive Statistics
- Probability
- Inference

Introduction – 1

- 1600s
  - Pascal, Bernoulli, …
- 1700s
  - Thomas Bayes (1763)
  - Pierre Simon Laplace (1774)
- 1800s
  - George Boole (1854)
- 1900s
  - Pearson, Fisher, Neyman, Jeffreys, Jaynes, Kendall, Stuart, Kolmogorov…

To Be Good Or Not To Be

In 1670 Pascal applied probabilistic reasoning to the following interesting hypotheses

G	God exists
~G	God does not exist

the following two actions

P	Lead a pious life
W	Lead a worldly life

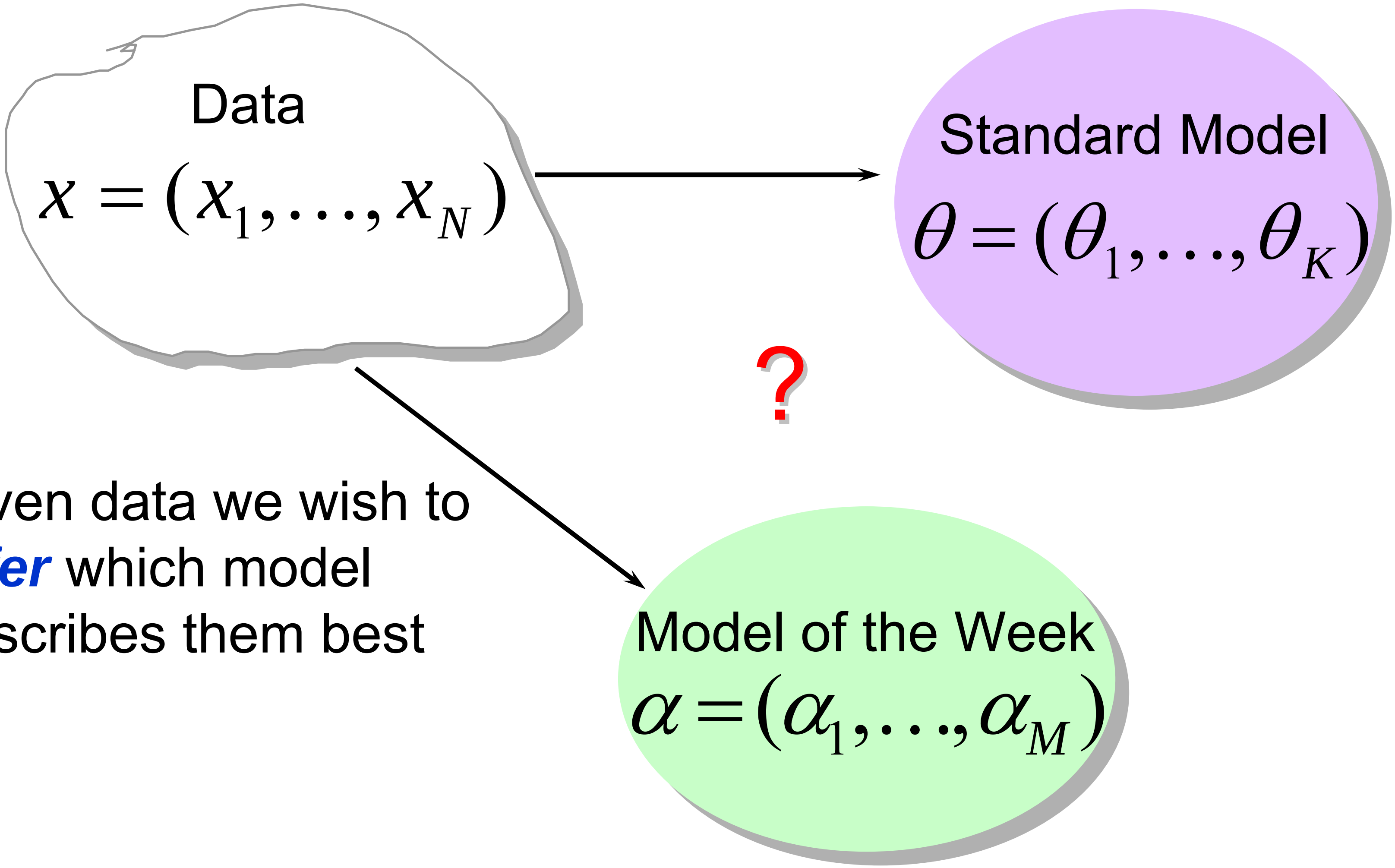
and assigned payoffs (utilities) to each hypothesis / action pair.

To Be Good Or Not To Be

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

If *your* **Pr**(God) > 0, however small, then your expected payoff from being pious >> expected payoff from being worldly. So if you believe in God, even if only on Sundays, the rational course of action is to live a saintly life!

Introduction – 2



Given data we wish to ***infer*** which model describes them best

Descriptive Statistics

Descriptive Statistics – 1

Definition: A **statistic** is any function of the data **X**.

Given a sample **X** = $x_1, x_2, \ldots x_N$ it is of interest to compute **statistics** such as the **sample average**

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

and the **sample variance**

$$S^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^2$$

Descriptive Statistics – 2

Consider an **ensemble** of similar experiments. They could be the results of simulations. In general, the statistics will vary from one experiment to another.

In developing analyses it is good practice to study **ensemble averages**, denoted <…>, of relevant statistics; e.g.,

$$< \overline{x} > = < \frac{1}{N} \sum_{i=1}^{N} x_i >$$

$$= \frac{1}{N} \sum_{i=1}^{N} < x_i >$$

Descriptive Statistics – 3

Ensemble Average	$< x >$
Mean	μ
Error	$\varepsilon = x - \mu$
Bias	$b = < x > - \mu$
Variance	$V = < (x - < x >)^2 >$
	$= < x^2 > - < x >^2$

Descriptive Statistics – 4

Mean Square Error (MSE)

$$\text{MSE} = \langle (x-\mu)^2 \rangle$$
$$= V + b^2$$

Exercise 1: Show this

The **MSE** is the most widely used measure of **closeness** of an ensemble of statistics {**x**} to the **true value** μ

The **root mean square** (RMS) is simply

$$\text{RMS} = \sqrt{\text{MSE}}$$

Descriptive Statistics – 5

Usually, each term in the sum is the same $$< \bar{x} > = \frac{1}{N} \sum_{i=1}^{N} < x_i >$$

Consequently, $$< \bar{x} > = \frac{1}{N} \sum_{i=1}^{N} < x > = < x >$$

Descriptive Statistics – 6

Consider the ensemble average of the **sample variance**

$$< S^2 > = < \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2 >$$

$$= < \frac{1}{N}\sum_{i=1}^{N}x_i^2 - \frac{2}{N}\sum_{i=1}^{N}x_i\overline{x} + \frac{1}{N}\sum_{i=1}^{N}\overline{x}^2 >$$

$$= \frac{1}{N}\sum_{i=1}^{N}< x_i^2 > - < \overline{x}^2 >$$

$$= < x^2 > - < \overline{x}^2 >$$

Descriptive Statistics – 7

The ensemble average of the sample variance is

$$< S^2 > = < x^2 > - < \overline{x}^2 >$$

$$= < x^2 > - \frac{< x^2 >}{N} - \frac{N-1}{N} < x >^2$$

$$= V - \frac{V}{N}$$

We have a **negative bias**

Exercise 2: Show this

Descriptive Statistics – 8

Finally, consider the variance of the sample average

$$< \Delta\overline{x}^2 > = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} < \Delta x_i \Delta x_j >$$

$$= \frac{1}{N^2}\left(\sum_{i=1}^{N} < \Delta x_i^2 > + \sum_{i=1}^{N}\sum_{j\neq i}^{N} < \Delta x_i \Delta x_j >\right)$$

where

$\Delta\overline{x} \equiv \overline{x} - < x >$ and $\Delta x_i \equiv x_i - < x >$

Descriptive Statistics – 9

Suppose that the data are correlated as follows

$$< \Delta x_i \Delta x_j > = \rho V$$

We find that

$$< \Delta\bar{x}^2 > = \frac{1}{N^2}\left(\sum_{i=1}^{N} < \Delta x_i^2 > + \sum_{i=1}^{N} \sum_{j \neq i}^{N} < \Delta x_i \Delta x_j > \right)$$

$$= \frac{V}{N}\left(1 + (N-1)\rho \right)$$

Descriptive Statistics – Summary

The **sample average**
is an unbiased estimate
of the ensemble average

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

The **sample variance**
is a biased estimate
of the ensemble variance

$$S^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2$$

The variance of the sample
average decreases like **1/N**
until we reach a limit imposed
by the degree of correlation in the data

$$V_{\overline{x}} = \frac{V}{N}\left(1 + (N-1)\rho\right)$$

Probability

Probability – 1

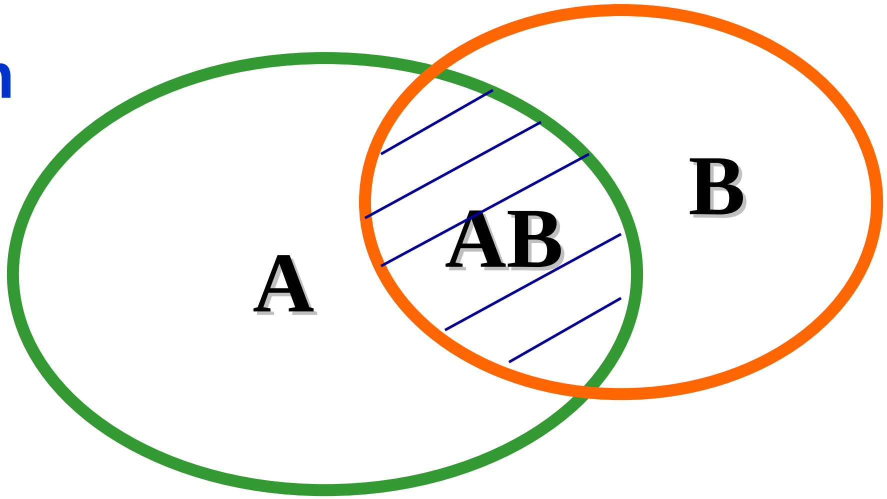
Probability is a **function** with range **[0,1]** defined on sets

Consider the sets ***A***, ***B***, ***A+B*** and ***AB***

To each assign the numbers **P(A), P(B), P(A+B)** and **P(AB)**

The rules of probability specify how these numbers are related.

Venn Diagram



Probability – 2

Theorem

$$P(A+B) = P(A) + P(B) - P(AB)$$

A and B are mutually exclusive if

$$P(AB) = 0$$

A and B are exhaustive if

$$P(A) + P(B) = 1$$

Exercise 3: Prove theorem

Probability – 3

Let A and B be sets of **propositions**, for example,

A = It is a baby

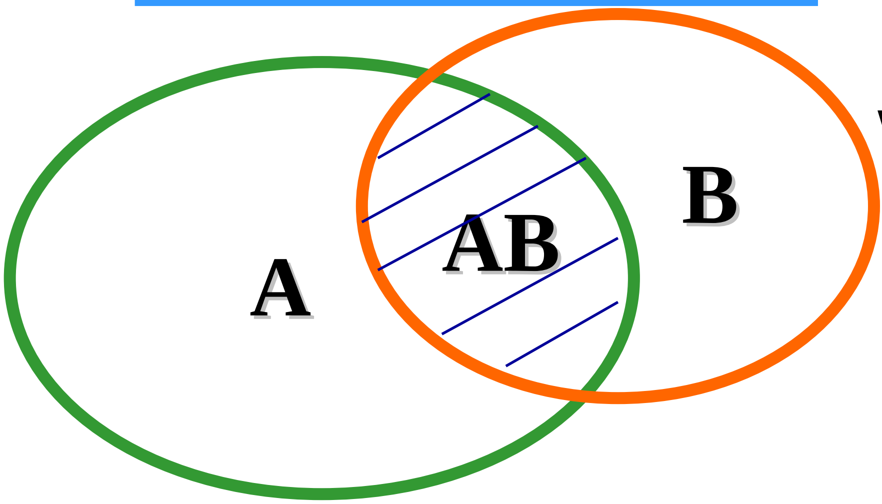
B = It vomits spontaneously

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*** restriction.

P(A|B) is the probability of A when we ***restrict*** to the proposition B



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

Bayes' Theorem – 1

From
we deduce immediately
Bayes' Theorem:

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

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

Bayes' Theorem – 2

Let B_1 and B_2 be ***exhaustive*** propositions. Consider AB_1, AB_2. We can write

$$P(AB_1) = P(B_1|A)\,P(A) \qquad (1)$$

$$P(AB_2) = P(B_2|A)\,P(A) \qquad (2)$$

Now add Eq.(1) and Eq.(2)

$$P(AB_1) + P(AB_2) = [P(B_1|A)+P(B_2|A)]\,P(A)$$
$$= P(A)$$

The summation over exhaustive propositions is called **marginalization**. It is an extremely important operation.

Bayes' Theorem – 3

Bayes' Theorem for propositions **A**, $\mathbf{B_k}$ can be written

$$P(B_k \mid A) = \frac{P(A \mid B_k)P(B_k)}{\sum_n P(A \mid B_n)P(B_n)}$$

Note that $$\sum_k P(B_k \mid A) = 1$$

Exercise 4: Prove this form of Bayes' Theorem

But What Exactly is Probability ?

Probability can be *interpreted* as a **degree of belief**
Probability can be *interpreted* as a **relative frequency**

Contrast the statements

a) There is a 20% chance of rain on 13 August, 2007

b) There is a 20% chance of rain on Mondays

Statement a) says how much one ***believes*** or is invited to ***believe*** it will rain today.

Statement b) states the ***relative frequency*** with which it rains on Mondays.

Distributions and Densities – 1

If X can assume a set of values, then Pr(X) is called a **probability distribution function**.

X can be *discrete* or *continuous*.

If X is continuous, we can define

$$p(X) \equiv \frac{d\,\Pr(X)}{dX}$$

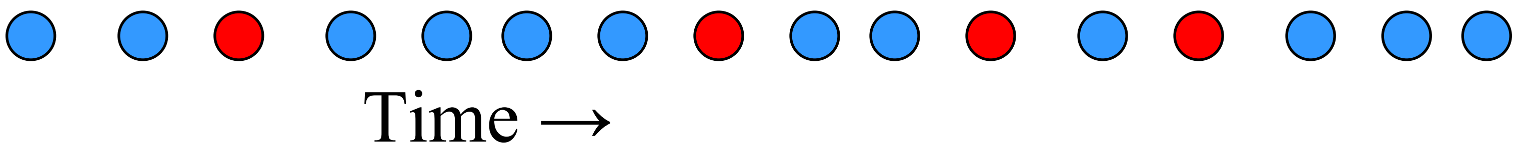
as the **probability density function**. Note: probabilities, being pure numbers, are *dimensionless*, whereas densities have dimensions of 1/X

Common Distributions and Densities

$Uniform\,(x)$	1
$Binomial\,(k,n,p)$	$\binom{n}{k} p^k (1-p)^{n-k}$
$Poisson\,(k,a)$	$a^k \exp(-a)/k!$
$Gaussian\,(x,\mu,\sigma)$	$\exp(-(x-\mu)^2/2\sigma^2)/\sigma\sqrt{2\pi}$
$Chisq\,(x,n)$	$x^{n/2-1}\exp(-x/2)/2^{n/2}\Gamma(n/2)$
$Gamma\,(x,a,b)$	$x^{b-1}a^b \exp(-ax)/\Gamma(b)$
$Exp\,(x,a)$	$a\exp(-ax)$

The Binomial Distribution – 1

A **Bernoulli** trial has two outcomes: ***S*** = success or ***F*** = failure.
Example: Each collision between protons at the LHC will be a Bernoulli trial in which something interesting happens (***S***) or does not (***F***).

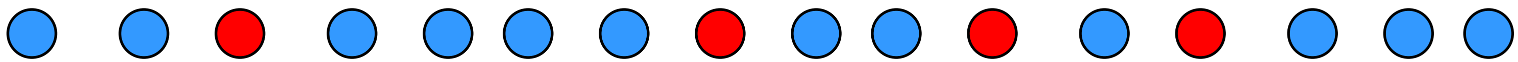


Let $p = P(S)$ be the probability of a success (a **red** spot), assumed to be the ***same at each trial***. Since ***S*** and ***F*** are ***exhaustive***, the probability of a failure is $1 - p$. For a given order ***O*** of ***N*** trails, the probability $P(K,O|N)$ of *exactly* ***K*** successes, and $N - K$ failures is

$$P(K,O \mid N) = p^K (1-p)^{N-K}$$

The Binomial Distribution – 2

If the order **O** of successes and failures is irrelevant, we can eliminate the order from the problem by ***marginalizing*** over all possible orders



Time →

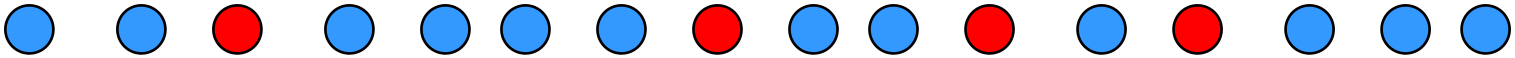
$$P(K \mid N) = \sum_{O} P(K, O \mid N) = \sum_{O} p^{K}(1-p)^{N-K}$$

This yields the **binomial distribution**

$$K \sim Binomial(K, p, N) \equiv \binom{N}{K} p^{K}(1-p)^{N-K}$$

X ~ means “X is distributed as”

The Poisson Distribution



Time →

We expect ***a*** = ***p N***, where ***a*** is the mean number of successes and ***N*** the number of trials. When the probability ***p*** is very small, we can take the limit

$p \to 0$ and $N \to \infty$, such that ***a*** is *constant*,

Binomial$(k, N, p) \to$ ***Poisson***(k, a).

The Poisson distribution is general regarded as a good model of a **counting experiment**

Exercise 5: Show that *Binomial* → *Poisson*, in this limit

Inference

Inference – 1

Here is a very general inference procedure:

a) Compute Pr(Data|Model)

b) Compute Pr(Model|Data) using Bayes' theorem:

$$\Pr(\text{Model}|\text{Data}) = \Pr(\text{Data}|\text{Model})\ \Pr(\text{Model})/\Pr(\text{Data})$$

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 – 2

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

posterior density | likelihood | prior density

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

θ are the **parameters of interest**

marginalization

$$p(\theta \mid x) = \int_{\lambda} p(\theta,\lambda \mid x)\,d\lambda$$

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

Inference – 3

Model Selection (hypothesis testing)

posterior evidence prior

$$P(m \mid x) = \frac{p(x \mid m)\, P(m)}{p(x)}$$

The **evidence** for model **m** is defined by

$$p(x \mid m) = \int p(x \mid \theta_m, \lambda_m, m)\, \pi(\theta_m, \lambda_m \mid m)\, d\theta_m d\lambda_m$$

Inference – 4

posterior odds | Bayes factor | prior odds

$$\frac{P(m \mid x)}{P(n \mid x)} = \left(\frac{p(x \mid m)}{p(x \mid n)} \right) \frac{P(m)}{P(n)}$$

The **Bayes factor** can be used to choose between two competing models m and n.

It can also be used to optimize analyses….

An Example – 1

Model

$$n = s + b$$

Prior information

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

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

Likelihood

$$P(N \mid s, b) = Poisson(N, s + b)$$

s is the mean signal count

b is the mean background count

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

Datum

$$N$$

An Example – 2

Apply Bayes' theorem:

posterior likelihood prior

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

$\pi(s,b)$ is the prior density for *s* and *b*

It *encodes* somehow our prior knowledge of the signal and background means.

The encoding is *difficult* and *controversial*.

An Example – 3

First factor the prior

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

Define the **marginal likelihood**

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

And write the posterior density for the signal as

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

An Example – 4

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 certain cuts.

We assume that the probability for the count ***B*** is given by P(*B*|λ) = *Poisson*(B, λ), where λ is the (unknown) mean count of the Monte Carlo background. We can make an inference about λ by applying Bayes' theorem to the Monte Carlo background experiment

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

An Example – 5

The background prior density…

Assume a prior of the form $\pi(\lambda) = \lambda^p$. The case p = 0, is called the **flat prior**. Using the flat prior, we find

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

Assume that the mean background count ***b*** in the actual experiment is related to the mean count λ in the Monte Carlo experiment via ***b* = *k*** λ, where ***k*** is an accurately known scale factor, for example, the ratio of the data and Monte Carlo integrated luminosities. The background can be estimated as follows

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

An Example – 6

The background prior density…

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

$\pi(b) = p(\lambda|B)$, since $b = k\lambda$.

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

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

An Example – 7

The calculation of the marginal likelihood can be done

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

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

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

Exercise 6: Give a full derivation of this result

And Finally

The signal prior density
We know it is positive and finite! It is far from clear how to translate this prior knowledge into a prior density $\pi($**s**$)$.

We shall simply adopt a flat prior for the signal $\pi(s) = 1$ as a matter of **convention**.

Exercise 7: Derive a formula for $p(s|N)$ and plot the posterior density for $N = 5$, $B = 20$, $k = 0.1$.