2. Probability

Review of lecture 1

Science is decision - we must work out how to decide

Decision is by comparing – with a 'clean' sample or a hypothesis

Statistics are properties of the data, nothing else

Astronomers start from problem areas – we must use best practice

We must be familiar with (a) non-parametric statistics
(b) different measurement scales

Science is done in a process which can be defined, and each stage requires statistical familiarity

We need stats because of decision, self-defense .....

Concepts of probability and probability distributions will be heavily involved

The literature....

……that you need is mostly in

Numerical Recipes…The Art of Scientific Computing
Press, Teukolsky, Vetterling, Flannery
Cambridge University Press 2007

If you are a serious physical scientist, you should think about buying this book.

Consider also

Data Analysis....A Bayesian Tutorial 2nd ed., D. S. Sivia, CUP 2006

Bayesian Logical Data Analysis for the Physical Sciences, Phil Gregory (UBC), CUP 2005

Modern Statistical Methods for Astronomy...with R Applications, Eric Feigelson and Jogesh Babu, CUP 2012

Probability is essential for us....

(1) Astronomical measurements are subject to random measurement error and we need to have a common language of expression. If we quote an error, what is the unspoken assumption about it?

(2) The inability to do experiments on our subject matter leads us to draw conclusions by contrasting properties of controlled samples. They are usually 'too small', leading to 'statistical error'.

Example: `the distributions of luminosity in X-ray-selected Type I and Type II objects differ at the 95 per cent level of significance.'

Very often the strength of this conclusion is:
- dominated by the number of objects in the sample
- unaffected by observational error.

So: probability + conditionality + independence + Bayes' Theorem + prior + posterior probabilities.

What is probability?

From games of chance involving cards or dice: probabilities are often considered as a kind of limiting case of a frequency…….
`Obviously' probabilities of certain events are

number of favourable events / total number of events

Example: throwing a six with one roll of the dice (die) is 'obviously' 1/6.

Laplace - the "**Principle of Indifference**" (**PoI**):
Assign equal probabilities to events unless we have any information distinguishing them.

Example: probability of one spot = x, two spots =x, etc
With convention that probability of a certain event ≡ 1.0, $6x = 1$

But we cannot *define* probability by this kind of ratio. We have had to assume that each face of the die is equally probable – thus the definition of probability becomes circular.

Probability - frequentist

Sometimes we estimate probabilities from data.

Example: The probability of our precious observing run being clouded out is estimated by :

number of cloudy nights last year / 365

but two issues:

1. limited data – ten years' worth of data would give a different, more accurate result? Do we know this?
2. identifying `equally likely' cases e.g more likely to be cloudy in winter? So what is winter then? A set of nights equally likely to be very cloudy?

It is common to define probabilities as empirical statements about frequencies, in the limit of large numbers of cases.

This definition must be circular because selecting the data depends on knowing which cases are equally likely.

This **'frequentist' approach** is sometimes the only way; but the risks must be recognized.

What is probability?

Probability is a numerical formalization of our degree or intensity of belief.

Example: in throwing dice, x measures the strength of our belief that any face will turn up.

One person's degree of belief is another person's certainty, depending on what is known.

If our probabilities turn out to be wrong, the deficiency is in what we know, not in the definition.

But two people with the same information must arrive at the same probabilities. This constraint, properly expressed, is enough to develop a theory of probability mathematically identical to the one often interpreted in frequentist terms.

Probability theory

Formalizing 'measure of belief' => deduction of a useful set of properties of probability.

If A, B and C are three events and we wish to have some measure of how strongly we think each is likely to happen, then for consistent reasoning we should at least apply the rule (Cox 1946):
If A is more likely than B, and B is more likely than C, then A is more likely than C.

This is sufficient to put constraints on the probability function which are identical to the Kolmogorov axioms of probability:

- Any random event A has a probability prob(A) between 0 and 1.
- The sure event has prob(A)=1.
- If A and B are exclusive, then prob(A or B) = prob(A) + prob(B).

The Kolmogorov axioms are a sufficient foundation for the entire development of mathematical probability theory, the apparatus for manipulating probabilities once we have assigned them.

Example: naked-eye supernovae

< 1987, 4 naked-eye supernovae had been recorded in 10 centuries.
What, before 1987, was the probability of a bright supernova happening in the 20th century?

There are three possible answers.

(1) Probability is **meaningless** in this context. This is physics, deterministic, and timing can be calculated. They are not random events.

(2) **Frequentist** point of view: best estimate of the probability is 4/10, although it is obviously not very well determined. (Assumes equally likely to be reported throughout ten centuries - some degree of belief about detection efficiency will have to be made explicit in this kind of probability assignment.)

(3) We could try an **a priori** assignment. We might know
- the stellar mass function,
- the fate and lifetime as a function of mass,
- the stellar birth rate, and
- detection efficiency.

From this we could calculate the mean number of supernovae expected in 1987, and we would put some error bars around this number to reflect unknowns……

Example: naked-eye supernovae 2

…….The belief-measure structure is more complicated in this detailed model. The model deals in populations, not individual stars, and assumes that certain groups of stars can be identified which are equally likely to explode at a certain time.

Suppose now that we sight supernova 1987A. Is the probability of there being a supernova later in the 20th century affected by this event?

(1) would say No -- one supernovae does not affect another.
(2) in which the probability reflects what we know, would revise the probability upward to 5/10.
(3) might need to adjust some aspects of its models in the light of fresh data; predicted probabilities would change.

Probabilities reflect what we know -- they are not things with an existence all of their own. Even if we could define `random events' (Approach 1), we should not regard the probabilities as being properties of supernovae.

Conditionality and Independence

Two events **A** and **B** are said to be **independent** if the probability of one is unaffected by (what we may know about) the other. From the Kolmogorov axioms

$$\mathrm{prob}(\mathrm{A\ and\ B}) = \mathrm{prob}(\mathrm{A})\mathrm{prob}(\mathrm{B})$$

Sometimes independence does not hold, so that we would also like to know the **conditional probability**: the probability of **A**, given that we know **B**. The definition is:

$$\mathrm{prob}(\mathrm{A} \mid \mathrm{B}) = \frac{\mathrm{prob}(\mathrm{A\ and\ B})}{\mathrm{prob}(\mathrm{B})}$$

If **A** and **B** are independent, knowing that **B** has happened should not affect our beliefs about the probability of **A**. Hence prob(**A/B**) =prob(**A**) and the definition reduces to **prob(A and B)=prob(A)prob(B)** again.

If there are several possibilities for event **B** (label them $\mathbf{B_1}$, $\mathbf{B_2}$ ….) then we have that

$$\mathrm{prob}(A) = \sum_i \mathrm{prob}(A \mid B_i)\mathrm{prob}(B_i)$$

A might be a cosmological parameter of interest, while the **B**s are not of interest. Knowing the probabilities prob($\mathbf{B_i}$) we get rid of these *nuisance parameters* by a summation (or integration); this is called **marginalization.**

Conditional probability - example

1. Two QSOs of different redshift are beside each other on the sky. Remarkable!
Calculate probability: it is **conditional** on having noticed this at the start.
Thus prob(**A/A**) = 1, consistent with our measure of belief in something we know.

2. Now calculate probability of finding a galaxy and a QSO within **r** of each other.
We search the solid angle Ω and have already found ς_G and ς_Q. We need:

$$\text{prob}(\text{G in field and Q within } r) = \text{prob}(\text{Q within } r \mid \text{G in field})\text{prob}(\text{G in field})$$

Assumes probabilities are independent – and this is what we want to test.
Without resorting to models:

$$\text{prob}(\text{G in field}) = \varsigma_G \Omega$$

and

$$\text{prob}(\text{Q within } r) = \pi r^2 \varsigma_Q.$$

So we get

$$\text{prob}(\text{G in field and Q within } r) = \varsigma_G \varsigma_Q \Omega \pi r^2.$$

…..symmetrical in QSO and galaxy surface densities – we could search first for a galaxy or for a QSO. Note strong dependence on search area – specify this **before** the experiment!

Bayes' Theorem

$$\text{prob}(\text{B} \mid \text{A}) = \frac{\text{prob}(\text{A} \mid \text{B})\text{prob}(\text{B})}{\text{prob}(\text{A})}$$

The event **A** (the data), follow **B**
Prob(**A**) is the **normalizing** factor
Prob(**B**) is the **prior probability**, to be modified by experience (namely the data **A**)
Prob(**A**|**B**) is the **likelihood**
Prob(**B**|**A**) is the **posterior probability**, the answer, the subsequent state of belief

An innocent mathematical identity – but ***its interpretation or application has momentous consequences*** for analysis of data, experimentation.

Notice also the affinity with **maximum likelihood** analysis - we'll come to this later.

What does this mean? I

Example: the famous and classical urn calculation, M white balls, N red balls. What's the probability of drawing 3 red and 2 white out of the M+N? This is a counting problem, PoI, etc, and we can count.

But this is not what we want to know! This is the wrong sum!

We do not want to know the probability of drawing a certain number of each colour.

What we want is the inverse probability calculation: we have data, ie we have a certain number drawn, say 2 white, 3 red – and we want to infer the population properties of the urn.

This is generally true in astronomy: we want to solve the inverse problem - we have a small sample and we wish to infer details about the population from which it was drawn.

What does this mean? 2

Example: N red, M white in an urn, total N+M=10.

If I make 5 draws (T=5) and get 3 red and 2 white, how many reds are in the urn?

So: from Bayes

$$\text{prob(contents of urn | data)} \propto \text{prob(data|contents of urn)} \times \text{prob(contents of urn)}$$

We can deal with the terms on the RHS. We take as a model for the probability of red = N/[N+M], assuming no funny business.

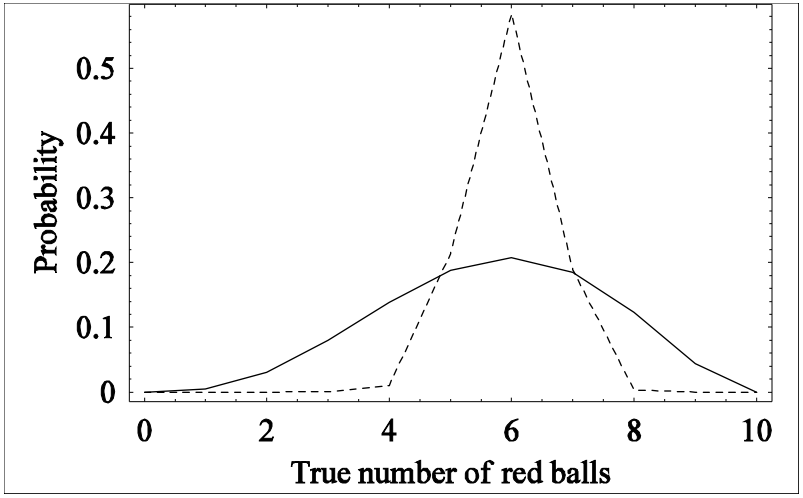
What does this mean? 3

The likelihood term = $\binom{T}{R}\left(\frac{N}{N+M}\right)^{R}\left(\frac{M}{N+M}\right)^{T-R}$

This is just the binomial distribution, which we meet next week – the old question of **n** successes out of **N** trials, given a fixed success rate of **ρ**.

What about the prior? Let's take it as uniformly likely between 0 and N+M.

So we can calculate the (un-normalized) posterior probability:



Binomial coefficient

$$\binom{x}{y} = \frac{x!}{y!\,(x-y)!}$$

Here's the results for our original 5 tries (3 reds) and 50 tries (30 reds).

What does this mean? 4

Unsurprising? Common sense? Maybe….but consider….

1. We now can describe our state of belief about the contents of the urn in physical or mathematical terms.

We believe on the basis of data that there are 6 reds, but – in the case of the 5 tries, there could be as few as 2 and as many as 10.

The probability of the urn containing 3 reds or less is 11 per cent, etc.

2. We have answered our scientific question: we have made an inference about the contents on the basis of data.

Bayes' theorem allows us to make inferences from the data, rather than compute the data we would get if we happened to know all relevant information.

What does this mean? 5

Again:

Bayes' theorem allows us to make inferences from the data, rather than compute the data we would get if we happened to know all relevant information.

Example: data from two populations – different means? Most books show you how to calculate the data you'd get if you have populations with different means. That's **not** what we asked!

We want to know, given the data, what is the probability/belief state of our model.

3. Note use of prior information – we assigned probabilities to **N** to reflect what we know. 'Prior' suggests 'before' but means **'what we know apart from the data'.** In the current example we used a uniform prior – and got a not unexpected result.

Priors can **change anticipated results** in violent and dramatic ways.

More on priors 1

Example: Sometimes for a prior we even need a probability of a probability:

Supernova rate per century: call this **ρ**.

Our data are 4 supernova in 10 centuries.

Our prior on **ρ** is uniform between 0 and 1; we know nothing.

A suitable model for **prob(data|ρ)** is the binomial distribution again, because in any century we either get a supernova or we do not.

The posterior probability is then

$$\text{prob}(\rho \mid \text{data}) \propto \left(\begin{array}{c} 10 \\ 4 \end{array} \right) \rho^4 (1-\rho)^6 \times \text{prior on } \rho$$

More on priors 2

Following Bayes and Laplace, take the prior uniform in the range 0 to 1. Then, to normalize the posterior probability properly, set

$$\int_0^1 \text{prob}(\rho \mid \text{data}) d\rho = 1$$

resulting in the normalizing constant

$$\int_0^1 \binom{10}{4} \rho^4 (1-\rho)^6 d\rho$$

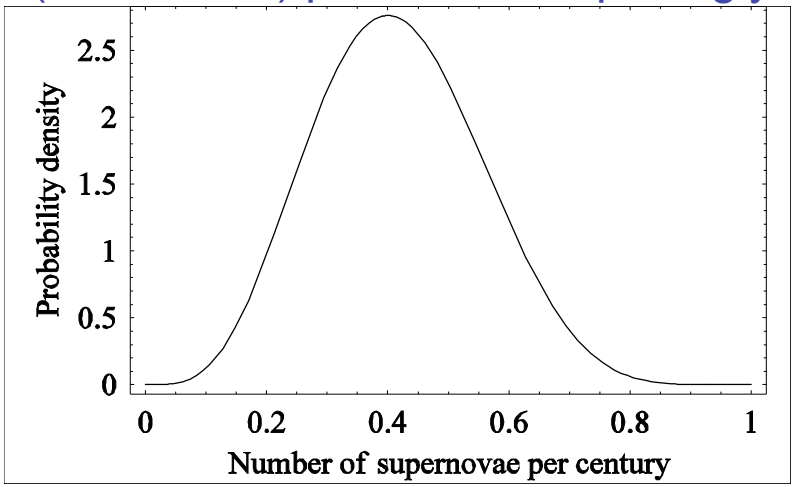
which is

$$\frac{\Gamma(10)\Gamma(4)}{\Gamma(14)} = B[5,7]$$

with ***B*** the (tabulated) beta function.

Our distribution (n=4, m=10) peaks - unsurprisingly - at 4/10, as shown.

.



More on priors 3

In general, for **n** supernova in **m** centuries, the distribution is

$$\text{prob}(\rho|\text{data}) = \frac{\rho^n(1-\rho)^{m-n}}{B[n+1, m-n+1]}$$

As sample size increases => narrower distribution, better definition of peak posterior probability.

'The law of large numbers' - a converging estimate.

More on priors 4

Key step – ascribing a **probability distribution to ρ**, in itself a probability.

- makes no sense in **frequentist** approach.
- makes no sense in any interpretation of probabilities as **objective.**

Assignment of a prior probability is generally much more difficult than this! It is the assignment of priors that really stokes the heat of debate between Bayesians and Frequentists.

Jeffreys, Jaynes discuss the uniform prior on **ρ** as being far too agnostic.
They reach other possibilities :

$$\mathrm{prob}(\rho) = \frac{1}{\rho(1-\rho)}$$

or $$\mathrm{prob}(\rho) = \frac{1}{\sqrt{\rho(1-\rho)}}$$ the `Haldane' prior

Some obvious priors like uniform **-∞** to **+∞** are not normalizable!

A common prior for a scale factor **σ** is **uniform in log σ** (Jeffrey's prior)

The **Maximum entropy principle** provides one way of determining a prior.

Yet another use of our few supernovae...

... to illustrate how 'prior' = knowledge

Example: Finally, the few supernovae can illustrate the use of Bayes' Theorem as a method of induction.

Assume we establish our posterior distribution at the end of the 19th century, so that it is

$$\rho^4(1-\rho)^6/B[5,7]$$

as shown earlier. At this stage, our data are 4 supernovae in 10 centuries. At the end of the 20th century, we take this as our prior.

Available new data consist of one supernova, so that the likelihood is simply the probability of observing exactly one event of probability **ρ**, namely **ρ**. The updated posterior distribution is

$$\text{prob}(\rho \mid \text{data}) = \frac{\rho^5(1-\rho)^6}{B[6,7]}$$

which peaks at **ρ** =5/11 as we might expect.

How to describe the posterior distribution? I

(1) The **peak of the posterior probability distribution** is one way amongst many of characterizing the distribution by a single number.

(2) The **posterior mean** is another choice, defined by

$$< \rho >= \int_0^1 \rho \, \mathrm{prob}(\rho|\mathrm{data}) d\rho$$

If we have had **N** successes and **M** failures, the posterior mean is given by a famous result called **Laplace's Rule of Succession**:

$$< \rho >= \frac{N+1}{N+M+2}$$

Example: For our SNs, at the end of the 19th century Laplace's rule would give 5/12 as an estimate of the probability of a supernova during the 20th century. This differs from the 4/10 derived from the peak of the posterior probability.

How to describe the posterior distribution? 2

Unless posterior distributions are very narrow, attempting to characterize them by a single number is misleading.

(3) A central measure plus a width is of course better, but such distributions are often asymmetrical with a long tail (or two).

It all depends as ever on what is to be done with the answer, which in turn depends on having a carefully-posed question in the first place.