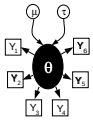
Bayesian Statistical Analysis



David Spiegelhalter
MRC Biostatistics Unit, Cambridge

davids@mrc-bsu.cam.ac.uk

AI and Statistics: January 3rd 1999

With acknowledgements to Nicky Best, Wally Gilks and many others from whom I have taken material.

Tutorial Outline

1. Introduction to Bayesian methods

- Fundamentals of Bayesian inference
- Probability, subjective probability, scoring rules, calibration
- Utilities, decision-making
- Likelihoods, updating beliefs, conjugate priors
- Predictive distributions.
- Using First Bayes software for conjugate Bayesian analysis.

2. Bayesian graphical modelling, MCMC and BUGS

- The problem of 'multiplicity'.
- Hierarchical models with 'unknown' priors.
- Conditional independence, graphical representation of hierarchical models.
- Introduction to MCMC.
- Gibbs sampling, relationship to graph.
- example.
- Complex modelling using WinBUGS.
- Practical issues: parameterisation, initial values, priors.

Some general references - more in lectures

- Books, in order of increasing technicality: Barnett (1982), Lindley (1985), Lee (1989), Gelman et al (1995), Carlin and Louis (1996), and Bernardo and Smith (1994).
- Berry and Stangl is a very good collection on biostatistics.
- An excellent tutorial paper is still Edwards et al (1963).
- Lilford and Braunholtz (1996) and Kadane (1995) are recent non-technical polemics in the medical literature.
- Spiegelhalter et al (1999) is a forthcoming review of Bayesian methods in health technology assessment.
- Jordan (1998) is an excellent collection of papers covering a broad range of topics under the umbrella term 'graphical models'. Buntine (1994, 1996), Neal (1996), Frey (1998) all connect Bayesian methods to neural networks and the AI literature.

INTRODUCTION

Introduction to the Bayesian approach

- The debate between "Bayesian" and "frequentist" statisticians is longstanding and largely philosophical
- There is increasing interest in Bayesian statistical methods in epidemiology, biostatistics, engineering, computer science and so on.
- We shall concentrate on the pragmatic rather than the philosophical

Axioms of probability

Let A, B be events, and $\{A_i, i = 1, 2, 3, ...\}$ be a set of events. The *probability* of $p(A)$ is a number which satisfies:

Axiom 1: $0 \leq p(A) \leq 1$ and $p(A) = 1$ if A is certain.

Axiom 2: If the events A_i are mutually exclusive, $p(\bigcup_i A_i) = \sum_i p(A_i)$

Axiom 3: $p(A \bigcap B) = p(B|A)p(A)$

Bayes' theorem, provable from axioms.

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}.$$

If A_i is a set of mutually exclusive and exhaustive events (*i.e.* $p(\bigcup_i A_i) = \sum_i p(A_i) = 1$), then

$$p(A_i|B) = \frac{p(B|A_i)p(A_i)}{\sum_j p(B|A_j)p(A_j)}.$$

How much does evidence B change our probability for each event A_i?

Example in diagnostic testing

- A new home HIV test is claimed to have "95% sensitivity and 98% specificity".
- In a population with an HIV prevalence of 1/1000, what is the chance that someone testing positive actually has HIV?

Let A be the event that the individual is truly HIV positive, $\overline{A}$ be the event that they are truly HIV negative.

Let B be the event that they test positive.

We want $p(A|B)$.

"95% sensitivity" means that $p(B|A) = .95$.

"98% specificity" means that $p(B|\overline{A}) = .02$.

Now Bayes theorem says

$$p(A|B) = \frac{p(B|A)p(A)}{p(B|A)p(A) + p(B|\overline{A})p(\overline{A})}.$$

Hence $p(A|B) = \frac{.95 \times .001}{.95 \times .001 + .02 \times .999} = .045$.

Thus over 95% of those testing positive will, in fact, not have HIV.

- Our intuition is poor when processing probabilistic evidence
- The vital issue is *how should this test result change our belief that we are HIV positive?*
- Bayes theorem in diagnostic testing is uncontroversial and established.
- The disease prevalence can be thought of as a *'prior'* probability ($p = 0.001$).
- Observing a positive result causes us to modify this probability to $p = 0.045$. This is our *'posterior'* probability of being HIV positive.
- More controversial is the use of Bayes theorem in general statistical analyses, where parameters are the unknown quantities, and their prior distribution needs to be specified

'Classical' statistical inference on parameter θ

- Set up null hypothesis θ_0
- Select test statistics T
- Calculate T for observed data
- Reject H_0 if T extreme, relative to $p(T|H_0)$
- Confidence intervals based on values of θ that cannot be rejected.

Likelihood inference on parameters θ

- Observe data x
- Set up likelihood $p(x|\theta)$
- MLE $\hat{\theta}$ maximises likelihood of θ
- Uncertainty concerning MLE based on curvature of likelihood

Some problems with these approaches.

- θ_0 is generally manifestly untrue (e.g. zero effect)
- P-values misinterpreted, and their importance depends on sample size
- Each analysis carried out in isolation
- Ignores all *external* evidence
- Dependence on stopping rule
- Difficult in complex circumstances
- Multiplicity of analyses: many spatial regions, trials, endpoints, institutions etc.

Bayesian inference

Suppose we have data x, and unknowns θ which might be

- model parameters,
- missing data,
- events we did not observe directly or exactly.

Standard statistical methods suggest a likelihood

$$p(x \mid \theta).$$

In addition we require a prior distribution

$$p(\theta)$$

which expresses our uncertainty about θ **before** taking into account the data.

The posterior distribution $p(\theta \mid x)$ expresses our uncertainty about θ **after** taking into account the data.

Bayes theorem tells us how to calculate this:

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

$$\propto p(\theta)\, p(x \mid \theta).$$

posterior distribution; prior distribution; likelihood

It is not generally necessary to compute the denominator.

Components of Bayesian inference

The prior distribution. Use probability as means of quantifying uncertainty about unknown quantities (variables).

The likelihood. Relate all variables into a 'full probability model'.

The posterior distribution. When observe some variables (the data), use Bayes theorem to obtain conditional probability distributions for unobserved quantities of interest

Utilities If we are willing to quantify the value of different consequences, it is possible to use the posterior probabilities as a basis for decision making.

But what is the interpretation of 'probability'?

Symmetries. Physical property of object: dice, cards, balls in urns.

Frequentist. Long run frequencies of repeatable events.

Logical. Essential property of situation.

Subjective. Personal judgement about unique event.

Subjective probabilities

YOUR probability for an event.

e.g. my probability for *it will rain tomorrow in Cambridge* is 0.4.

Does this mean we can use any number we want?

NO, probabilities MUST 'cohere', *i.e.* obey the laws of probability.

Also, if we want to be taken seriously, our probabilities should have some relationship with reality!

Subjective probability assessment (Chaloner, 1996)

1. *Single events - Probability for A.*

- reference lotteries: compare with willingness to get reward if drawing red ball from urn containing, say, 60 red and 40 black balls.
- relative probability: A is 5 times as likely as *not* A.

2. *Continuous quantities - probability distribution for X.*

- divide plausible range into 3 or 4 equally-likely parts.
- other fractile assessment - *e.g.* x such that $P(X < x) = .05$.
- cumulative distribution assessment - *e.g.* p such that $P(X < 1000) = p$

Known heuristics and biasses in probability assessment. (Tversky, 1994)

Availability. Similar events easily recalled:

"the last treatment worked ".

Representative. How close is event to stereotypic example: over-estimates chance of future success by ignoring base-rate.

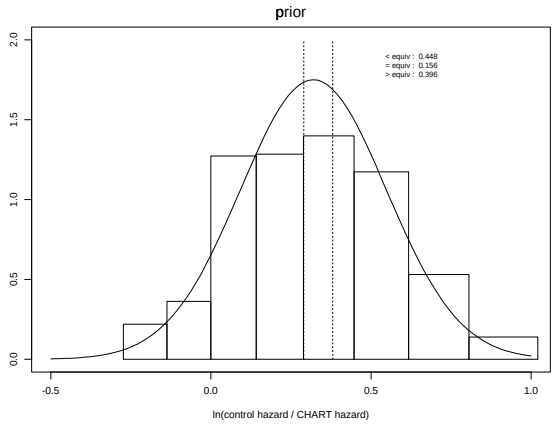
"this is just like the treatments that succeeded in the past!".

Anchoring. Sticking to initial ideas.

"assume all drugs have similar chance of success"

Doc	Range of equiv	– -5	-5–0	0–5	5–10	10–15	15–20	20–25	25–30
		Prior distribution (advantage of CHART over standard in 2-year % disease-free survival)							
1	5-10	10	30	50	10				
2	10-10	10	10	25	25	20	10		
3	5-10			40	40	15	5		
4	10-10			20	40	30	10		
5	10-15			20	20	60			
6	15-20	5	5	10	15	20	25	10	5
7	5-10				10	20	40	30	
8	20-20				10	20	40	20	10
9	10-15					10	50	30	10
mean	10-13	3	5	18	19	22	20	10	3

Summary of results of questionnaires completed by 9 clinicians in MRC CHART Trial for head and neck cancer.



For a clinical trial of the CHART radiotherapy protocol, 9 clinicians assessed prior distributions as histograms. Their average is shown above, transformed to a log(hazard ratio) scale, with a fitted normal distribution (Spiegelhalter etal 1994).

Approaches to conflict between judges; (Genest and Zidek, 1986)

Resolution to consensus:

- automatic averaging (weighting?)
- delphi-technique of iterating to consensus
- joint discussion

Accommodation

Keep multiple opinions in mind but criticise as additional evidence accumulates.

What do we want from 'reasonable' probabilities?

Accuracy = discrimination + calibration

- *discrimination* - probabilities are related to outcome (higher probabilities given to events that occur)
- *calibration* - probabilities mean what they say (*e.g.* events given a probability of 50% occur with frequency around 50%).

Subjective (*e.g.* horse racing) and objective (*e.g.* logistic regression) forecasters aim for both.

Scoring rules measure overall accuracy by penalizing bad predictions (Dawid 1986).

For binary quantities, let p_t be probability given to the outcome that actually occurred. Then alternative 'proper' scoring rules are:

- *Logarithmic:* $-\log p_t$
- *quadratic (Brier):* $(1-p_t)^2$

It is *not* appropriate to use an absolute penalty (1 - p_t). It can be shown this will encourage assessors to exaggerate their confidence.

For example, if it does *not* rain in Cambridge tomorrow, I have assessed probability $p_t = .6$ for the event that occurred, and so would be penalised either $-\log .6 = .51$ or $(1-.6)^2 = .16$ depending on the scoring rule being used (the logarithmic rule is generally given an upper bound, just in case some fool breaks Cromwell's Law and assigns zero probabiity to an event that then occurs).

Assessment of calibration

Let probabilities $p_1, ..., p_I$ be subjectively given to the success of Bernoulli trials $X_1, ..., X_I$. Then under the null hyothesis of perfect calibration,

$$X_i \sim \text{Bernoulli}(p_i)$$

so that a statistic

$$Z = \frac{O-E}{\sqrt{V(O-E)}} = \frac{\sum_{i=1}^{I}(X_i - p_i)}{\sqrt{\sum_{i=1}^{I} p_i(1-p_i)}}$$

is approximately standard normal.

Global calibration: calculated for all assessments.

Local calibration: calculate for, say, $.5 < p_i < .6$.

Rational decision making in the face of uncertainty: Lindley (1985)

Suppose we need to make one of a set of possible decisions $d_1, ..., d_K$.

The value or 'utility' if we decide d_i and θ turns out to be the true 'state of nature' is $U(d_i, \theta)$.

Suppose we have a probability distribution $p(\theta)$ (which may be a posterior distribution having observed some evidence). Then the *expected utility* of deciding d_i is

$$EU(d_i) = \sum_{\theta} U(d_i, \theta) p(\theta).$$

Principle: choose the decision that maximises expected utility.

(Alternative, pessimistic, *minimax* principle used in game theory: choose decision that maximises the lowest utility that could possibly happen, regardless of its probability.)

Examples of Bayesian inference

Bernoulli trials - discrete prior

Assume a drug may have response rate θ of .2, .4, .6 or .8., each of equal prior probability. If we observe a single positive response ($X = 1$), how is our belief revised?

Likelihood, $p(X \mid \theta) = \theta^X (1-\theta)^{(1-X)}$

θ	Prior $p(\theta)$	Likelihood × prior $p(X=1\|\theta)p(\theta)$	Posterior $p(\theta\|X=1)$
.2	.25	.05	.10
.4	.25	.10	.20
.6	.25	.15	.30
.8	.25	.20	.40
$\sum_j$	1.0	.50	1.0

Posterior $p(\theta_i|x) = \frac{p(x|\theta_i)p(\theta_i)}{\sum_j p(x|\theta_j)p(\theta_j)}$

Note: a single positive response makes it four times as likely that the true response rate is 80% rather than 20%.

With a Bayesian approach *prediction* is straightforward. Suppose we wish to predict the outcome of a new observation Z (say), given what we have already observed.

For discrete θ we have

$$p(Z|x) = \sum_j p(Z, \theta_j|x)$$

which, since Z is usually conditionally independent of x given θ, is generally equal to

$$p(Z|x) = \sum_j p(Z|\theta_j) w_j$$

where the $w_j = P(\theta_j|x)$ are 'posterior weights'.

Example: the predictive probability that the next treatment is successful:

$$p(Z = 1|x = 1) = \sum_j \theta_j w_j$$

$= 0.2\times0.1+0.4\times0.2+0.6\times0.3+0.8\times0.4 = 0.6$

Binomial response - discrete prior

If we observe r responses out of n patients, how is our belief revised?

Likelihood, $p(X \mid \theta) \propto \theta^r(1-\theta)^{n-r}$.

Suppose $r = 15, n = 20$

θ_i	Prior $p(\theta_i)$	Likelihood $\times$ prior $\theta_i^r(1-\theta_i)^{n-r}\, p(\theta_i)$ ($\times 10^{-7}$)	Posterior $p(\theta_i \mid X = 1)$
.2	.25	.0	.0
.4	.25	.2	.005
.6	.25	12.0	.298
.8	.25	28.1	.697
$\sum_j$	1.0	40.3	1.0

Binomial response - continuous prior

Data:
r successes from n independent trials

Likelihood:

$$\binom{n}{r}\theta^r(1-\theta)^{n-r} \propto \theta^r(1-\theta)^{n-r}$$

Prior: flexible 'conjugate' beta family

$$\begin{aligned}\theta &\sim \text{Beta}(a,b)\\ &\equiv \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}\\ &\propto \theta^{a-1}(1-\theta)^{b-1}\end{aligned}$$

Posterior:

$$\begin{aligned}p(\theta \mid r,n) &\propto p(r \mid \theta, n)p(\theta)\\ &\equiv \theta^r(1-\theta)^{n-r}\theta^{a-1}(1-\theta)^{b-1}\\ &\propto \theta^{r+a-1}(1-\theta)^{n-r+b-1}\\ &\propto \text{Beta}(r+a,\ n-r+b)\end{aligned}$$

For a beta distribution

$$\begin{aligned}\text{mean } m &= a/(a+b)\\ \text{variance } s^2 &= m(1-m)/(a+b+1)\end{aligned}$$

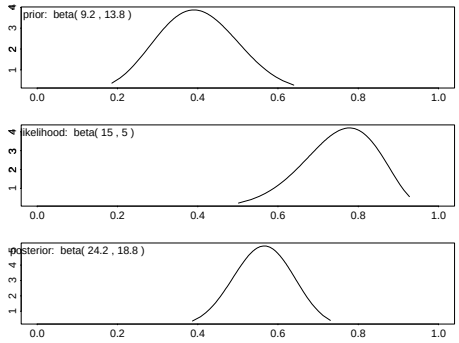
Suppose our prior estimate of the response rate is $m = .4$, with a standard deviation of $s = .1$.

Solving gives $a = 9.2, b = 13.8$.

Convenient to think of this as equivalent to having observed 9.2 successes in $a + b = 23$ patients.

	Prior	Likelihood	Posterior
'Successes'	9.2	15	24.2
'Failures'	13.8	5	18.8

Posterior is a beta distribution with mean $(r+a)/(n+a+b)$



Summarising posterior distributions

Given a posterior distribution $p(\theta|x)$, we may want

- mean, standard deviations, medians etc
- probability of exceeding certain thresholds, $p(\theta > \theta_0|x)$

If θ is multi-dimensional, need techniques for integrating out nuisance parameters (see later).

Predictive distributions

For future data Z, may be interested in predictive distribution $p(Z|x)$, given by

$$p(Z|x) = \int p(Z|x,\theta)\ p(\theta|x)\ d\theta,$$

which by conditiional independence generally simplifies to

$$p(Z|x) = \int p(Z|\theta)\ p(\theta|x)\ d\theta.$$

Independent Normal data

Suppose we have an independent sample

$$x_i \ \sim\ \mathrm{N}(\mu,\sigma^2), \qquad i = 1\ldots n,$$

where μ is unknown, σ^2 known.

Suppose we have a Normal prior distribution:

$$\mu \ \sim\ \mathrm{N}(\nu,\tau^2),$$

where ν and τ^2 are fixed.

Then the posterior distribution for μ is

$$\begin{aligned} p(\mu \mid \mathbf{x}) &\propto p(\mu)\, p(\mathbf{x} \mid \mu) \\ &\propto \exp\left[-\frac{1}{2}\left\{\frac{(\mu-\nu)^2}{\tau^2} + \frac{\sum(x_i-\mu)^2}{\sigma^2}\right\}\right]. \end{aligned}$$

which by completing the square can be seen to have a Gaussian kernel for μ.

So

$$\begin{aligned} p(\mu \mid \mathbf{x}) &= \mathrm{N}\left(\frac{\frac{\nu}{\tau^2} + \frac{n\bar{x}}{\sigma^2}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}},\ \frac{1}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}\right) \\ &= \mathrm{N}\left(w\nu + (1-w)\bar{x},\ \frac{\sigma^2}{n}(1-w)\right) \end{aligned}$$

where

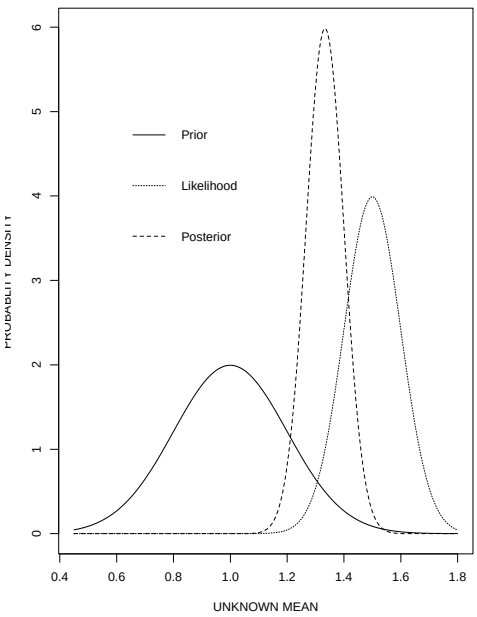
$$w = \frac{\frac{1}{\tau^2}}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}} = \frac{\sigma^2/n}{\sigma^2/n + \tau^2}.$$

Thus the posterior mean is a weighted average of prior and sample mean, weighted inversely to the prior and sample variances.

Note that if we write our prior as $\mathrm{N}(\nu, \sigma^2/m)$ (that is put $\tau^2 = \sigma^2/m$) then we obtain

$$p(\mu \mid \mathbf{x}) = \mathrm{N}\left(\frac{\bar{x}n}{n+m} + \frac{\nu m}{n+m}, \frac{\sigma^2}{n+m}\right)$$

Hence m may be viewed as a 'prior sample size'.



In the frequentist setting, the MLE $\hat{\mu} = \bar{x}$ and

$$p(\hat{\mu} \mid \mu) \;=\; \mathrm{N}(\mu, n^{-1}\sigma^2).$$

So

- the posterior mean is a compromise between the prior mean and the MLE
- the posterior s.d. is less than each of the prior s.d. and the s.e.(MLE)

and as $n \to \infty$,

- the posterior mean $\to$ the MLE
- the posterior s.d. $\to$ the s.e.(MLE)
- the 95% credible interval $\to$ the 95% confidence interval
- the posterior does not depend on the prior.

These observations are generally true, when the MLE exists and is unique.

When the posterior is in the same family as the prior then we have what is known as *conjugacy*. Examples include:

Prior	Likelihood	Posterior
Normal	Normal	Normal
Beta	Binomial	Beta
Dirichlet	Multinomial	Dirichlet
Gamma	Poisson	Gamma

There is a general theory of conjugate priors for exponential families (Bernardo and Smith, 1994).

Unfortunately conjugate priors do not exist for all likelihoods ... (and should we restrict ourselves anyway???)

Example : clinical trial

- Chemotherapy (Levamisole + 5FU) versus control in treatment of bowel cancer (Spiegelhalter et al, 1994)
- Observe n deaths in total
- Log-rank test statistic is

$$\mathrm{LRT}_n \;=\; O_{\mathrm{Lev5FU}} - E_{\mathrm{Lev5FU}}$$

- Asymptotically ($m \to \infty$):

$$x_n \;=\; \frac{4 \times \mathrm{LRT}_n}{n} \;\to\; \mathrm{Normal}(\theta, \sigma_n^2)$$

where θ = log hazard ratio and $\sigma_n^2 = \frac{4}{n}$

- Data from trial gave

$$n \;=\; 192$$
$$x_n \;=\; \frac{4 \times \mathrm{LRT}_n}{n} \;=\; 0.40$$

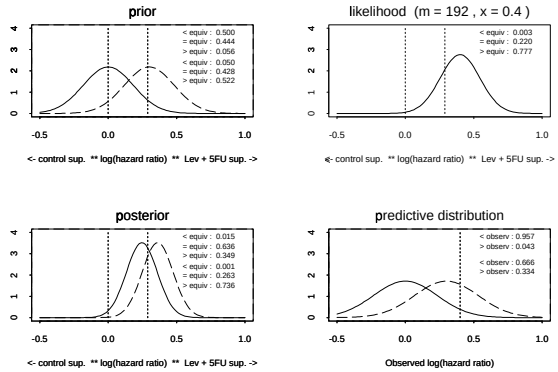
Spiegelhalter et al (1994). consider two priors for the log hazard ratio θ:

- a *sceptical* prior with mean 0 and sd .183. This specifies a 5% chance that $\theta > .30$, which expresses the opinion that very large benefits are implausible.
- an *enthusiastic* prior with mean .30 and sd .183. This specifies a 50% chance that $\theta > .30$.

Fleming and Watelet (1989) state that the treatment might only be clinically worthwhile of $\theta > .29$, and hence the interval (0 , .29) could be considered a range of clinical equivalence.

For each prior

- what is the posterior distribution for θ?
- what is the chance that $\theta > 0$?
- what is the chance that the treatment is clinically worthwhile?



For Lev + 5FU *vs* control, (a) prior, (b) likelihood, (c) posterior, and (d) predictive distribution $p(x) = \int p(x|\theta)\, p(\theta)\, d\theta$, showing the sceptical and enthusiastic analyses, and subsequent probabilities of lying below, within, or above the range of equivalence.

Priors

Where does the prior come from ?

- it is in principle subjective
- it might be elicited from experts
- it might be more convincing to be based on existing data, *e.g.* a meta-analysis. But the assumed relevance is still a subjective judgement.
- there are various "objective" approaches.
- conjugacy is useful but no longer necessary
- 'archetypal' priors expressing scepticism and enthusiasm are useful.
- it might be assumed unknown and 'estimated' using repeated similar situations (see later on hierarchical models).

There have been many attempts to make Bayesian methods objective, by defining automatic priors:

Uniform priors

(Bayes 1763 (Bayes, 1763); Laplace, 1776)

Set $p(\theta) \propto 1$

- This may be improper ($\int p(\theta) d\theta \neq 1$).
- The posterior will still usually be proper.
- Inference is based on the likelihood $p(x \mid \theta)$.
- It is not really objective, since a flat prior on θ does not correspond to a flat prior on $\phi = g(\theta)$.
- For r successes in n trials, predicts next success with probability $(r+1)/(n+2)$ (Laplace's law of succession)

'Non-informative' priors

Also known as *vague, ignorance, flat, diffuse* and so on

- Usually there exists a prior that reproduces the classical results
- Often improper
- Usually quite unrealistic
- However, always useful to report likelihood-based results as arising from a 'reference' prior.
- There is a theory of 'reference' priors designed to maximise information in an experiment

Public reporting of studies

Kass and Greenhouse (1989) ;
Spiegelhalter, Freedman and Parmar (1994)

A community of priors:

- *Reference:* "non-informative"
- *Clinical :* genuine clinical opinion
- *Sceptical :* expression that large differences are unlikely, used to control early positive stopping
- *Enthusiastic :* expression of confidence in new therapy, used to control early negative stopping.

General idea: by sensitivity analysis to a community of priors, one can assess whether the current results will be convincing to a broad spectrum of opinion.

Review: The good bits (in theory)

- Only need probability theory as basis for inference.
- No need to worry about statistical 'principles' such as: unbiassedness, efficiency, sufficiency, ancillarity, consistency, asymptotics and so on.
- Only need probability theory as basis for inference.
- No need to worry about significance tests, stopping rules, p-values.
- Models can be complex as reality demands.
- Integrates with planning decisions.
- Makes explicit and accountable what is usually implicit and hidden.
- Clarifies discussions and disagreements.
- Tells us what we want to know: *how should this piece of evidence change what we currently believe?*

The difficult bits (in practice)

- Inferences need to be justified to an outside world (reviewers, regulatory bodies, the public and so on): in particular
  - Where did the prior come from?
  - Is the model for the data appropriate (standard diagnostics)?
  - If making decisions, whose utilities?
- Although huge progress has been made, computational problems can still be considerable.
- Cannot overcome basic deficiences in design.
- Standards are needed for Bayesnian analysis and reporting.

BAYESIAN GRAPHICAL MODELS AND MCMC

Summary

- The problem of 'multiplicity'.
- Hierarchical models with 'unknown' priors.
- Conditional independence, graphical representation of hierarchical models.
- Introduction to MCMC.
- Gibbs sampling, relationship to graph.
- BUGS and CODA: Surgical example.
- Practical issues: parameterisation, initial values, priors.

'Multiplicity'

- Often we carry out many 'similar analyses'
  - pharmacokinetic models
  - meta-analysis
  - between-centre variability
  - growth curves
  - subsets
  - repeated looks at data
- Classical approach:
  - try to retain overall Type I error by some adjustment, e.g. Bonferroni
  - gives larger p-values, wider confidence intervals
  - in sequential analysis we should adjust estimate too

Bayesian approach to 'multiplicity'

- We are interested in making inferences on many parameters $\theta_1, ..., \theta_k$ measured on k 'units' (subsets, centres, trials, looks at the data etc).
- We are willing to assume the θ's are 'similar'.
- Integrate all analyses into a single model by assuming a hierarchical model (*aka* multilevel, random effects, random coefficient etc).
- Essentially equivalent to assuming or *estimating* a common prior.
- Both approaches can lead to similar conservatism
  - Classical: wider intervals
  - Bayesian: narrower intervals, but *biased* towards mean response

How can we formalise 'similarity'?

By the assumption of *exchangeability* (Bernardo and Smith, 1994, p169).

A finite sequence of variables $x_1, ..., x_n$ is judged to be (finitely) exchangeable if for any permutation $\pi(1), .., \pi(n)$,

$$p(x_1, ..., x_n) = p(x_{\pi(1)}, .., x_{\pi(n)}).$$

i.e. our joint opinion is unaffected by the labels, in that we have no reason to think the units are systematically different.

A sequence is infinitely exchangeable if all sub-sequences are finitely exchangeable.

A Classic Representation Theorem (de Finetti, 1930)

If $x_1, x_2, ..$ is an infinitely exchangeable sequence of 0-1 random variables, there exists a distribution q such that

$$p(x_1, ..., x_n) = \int_0^1 \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} q(\theta) d\theta.$$

i.e. as if there were some true unknown θ with prior distribution $q(\theta)$, under which $x_1, ..., x_n$ are independent Bernoulli trials.

Can also prove more general representation theorem (Bernardo and Smith, 1994, p 177).

So if we are willing to assume our $\theta_1, ..., \theta_k$ are exchangeable ("similar"), this is mathematically equivalent to assuming they are drawn at random from some population distribution. Note that -

- There does not need to be any actual sampling - perhaps only k units exist.
- This is a *judgement*: if there are known reasons to expect the units to be systematically different then those reasons need to be modelled.

Basis idea for Bayesian multiplicity:

- Assume a model for the overall population, with parameter μ.
- Model unit deviations θ_k, with prior mean zero and variance τ^2
- τ^2 may either be
  - assessed to have a particular value (subjective prior)
  - estimated using marginal maximum likelihood (empirical Bayes)
  - given a (usually minimally informative) prior distribution (full Bayes)

Example from Louis (1991)

Assume

$$\begin{aligned} y_k &\sim N(\theta_k, \sigma_k^2) \\ \theta_k &\sim N(\mu, \tau^2) \end{aligned}$$

Bayes theorem says

$$\begin{aligned} \hat{\theta}_k^{EB} = E(\theta_k|y_k) &= B_k\mu + (1 - B_k)y_k \\ V(\theta_k|y_k) &= (1 - B_k)\sigma_k^2 \end{aligned}$$

where $B_k = \sigma_k^2/(\sigma_k^2 + \tau^2)$. Assuming complete independence, rather than exchangeability, is equivalent to assuming $\tau^2 \to \infty, B_k \to 0$, and gives

$$\begin{aligned} E(\theta_k|y_k) &= y_k \\ V(\theta_k|y_k) &= \sigma_k^2 \end{aligned}$$

So $1 - B_k$ controls the 'shrinkage' of the estimate towards μ, and the reduction in the width of the interval for θ_k.

Can estimate μ, τ^2 by maximum marginal likelihood (cf Dersimonian and Laird, random effects meta-analysis).

Morris (1983) shows that EB estimators have lower expected summed mean-square error ($\sum(\hat{\theta_k} - \theta_k)^2$) than $\hat{\theta_k} = y_k$ ('Stein effect').

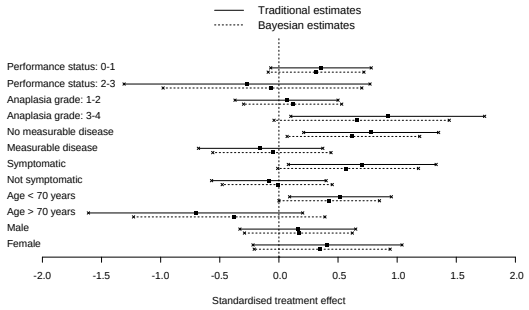
Automatically deals with regression to the mean.

Breslow (1990) on multiplicity

- Individual estimates of θ_i's are unreliable.
- Classical methods ignore 'relatedness' of θ_i's.
- Gives examples of (empirical) Bayes methods for
  - Longitudinal analysis
  - Small area estimation (disease mapping)
  - Relative risks in a case-control study.
  - Multiple tumour sites in a toxicology experiment

Bayes theorem for subset analysis (Dixon and Simon, 1992)

- Clinical trial in advanced colorectal cancer.
- Traditional 'treatment by subgroup' interactions: 4 of 12 intervals exclude zero.
- Subgroup-specific deviations from overall treatment effect have prior distribution centred at zero but with an unknown variability
- Pulled towards each other, due to prior scepticism about substantial subgroup-by-treatment interaction effects.
- Only one 95% interval now excludes zero: the subgroup with non-measurable metastatic disease.
- Generalisable to any number of subsets, a unified means of both providing estimates and tests of hypotheses.



Traditional and Bayesian estimates of standardised treatment effects in a cancer clinical trial. The Bayesian estimates are pulled towards the overall treatment effect by a degree determined by the empirical heterogeneity of the subset results.

Graphical Models

Graphical models are

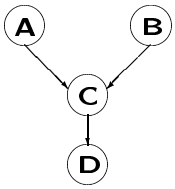
- a very general class,
- based on conditional independence (Dawid, 1979),
- represented using graphs (Whittaker, 1990).

They are also called *conditional independence* models.

Use graphs to:

- break complex models into simple components,
- communicate essential structure
- provide basis for computation

- *Directed Acyclic Graph* (DAG)
  - particular type of graph built from a sequence of conditional independence assumptions



- *Directed Local Markov property*

$$v \perp\!\!\!\perp \text{non-descendants}[v] \mid \text{parents}[v]$$
$$D \perp\!\!\!\perp A, B \mid C$$

- *Factorisation of joint distribution*

$$p(V) = \prod_{v \epsilon V} p(v \mid \text{parents}[v])$$
$$= p(A)\, p(B)\, p(C \mid A, B)\, p(D \mid C)$$

Graphs used in:

- Probabilistic expert systems (Bayesian networks),
  - Nodes are potentially observable
  - Propagate evidence using exact methods
- Classical graphical modelling.
  - Nodes are sets of multivariate data
  - Repeated structure over data bases
  - Parameter estimation by ML
- Latent variable modelling (including neural networks)
  - Nodes are sets of multivariate data and latent variables
  - Repeated structure over data bases
  - Parameter estimation by EM
- Bayesian graphical modelling.
  - Nodes are data, latent variables and parameters
  - Repeated structures over data and parameters ('plates')
  - Parameter estimations by exact, aanalytic approximations, and MCMC