Math 34920
Spring term 2006

Bayesian Modelling B

by Peter Green (University of Bristol, P.J.Green@bristol.ac.uk).

- modelling complex data structures
- conditional independence and graphical models
- exchangeability and hierarchical models
- Markov chain Monte Carlo algorithms
- demonstrations using **WinBUGS**

Relation to other units

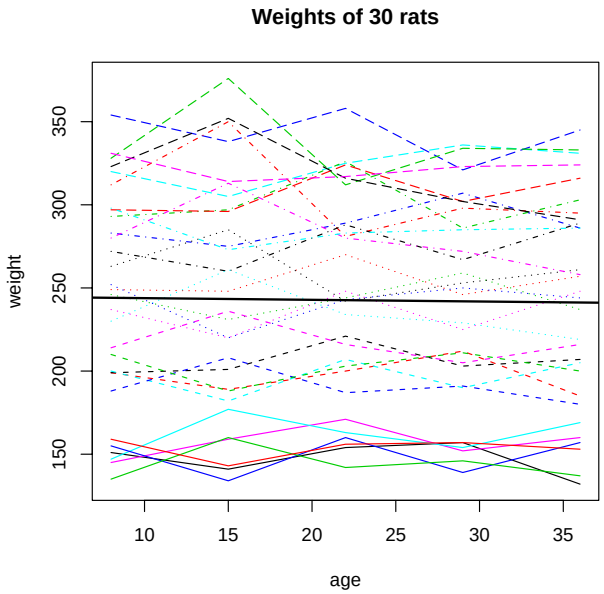
This unit leads on from level 3 *Bayesian modelling A*, and makes some use of ideas about Markov chains covered in level 2 *Applied probability*.

It was given by Christophe Andrieu to a different syllabus (but with some overlap) in 2003/04 and 2004/05.

Motivation: data sets and basic ideas

- hierarchical data
- Bayesian statistics
- complexity

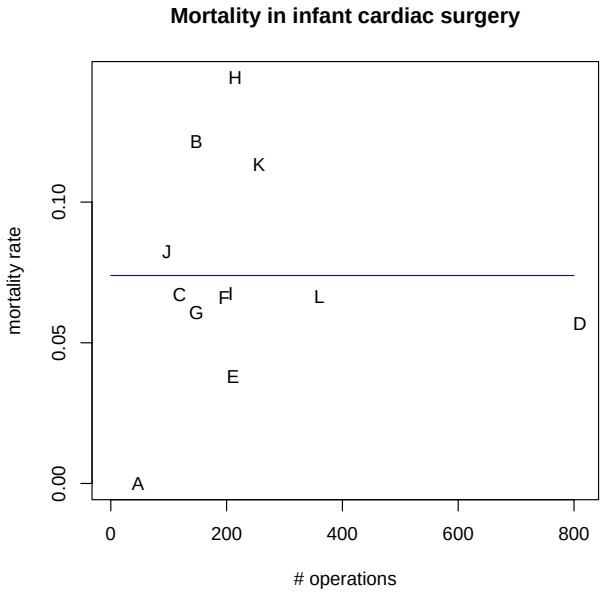
0. Motivation: data sets



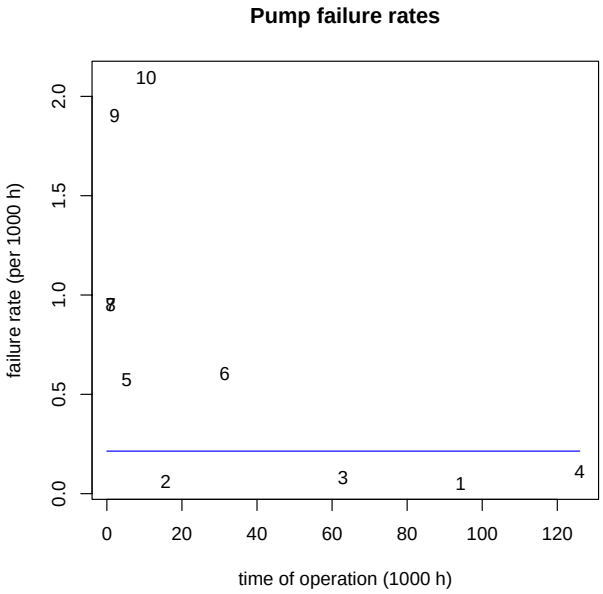
30 young rats are weighed weekly for 5 weeks. Is their growth linear? Or curving 'downwards'? How would you predict the weight in the week for a given rat? Or for the rat?

In 12 hospitals carrying out cardiac surgery on babies, the numbers of operations performed and mortality rates are recorded. What are the

hospitals? Are the differences more than can be attributable to ? What rate do you expect in the hospital? Or in the 12th hospital, in a different year?



Failure rates of pumps in nuclear power plants, running for different operation times.



Key ideas of Bayesian statistics

- All unknowns treated as variables, and all uncertainties measured using
- Inferences are statements (given data)
- Bayes theorem is used to 'turn round' $P\{\ \ |\text{parameters}\}$ to get $P\{\text{parameters}|\text{data}\}$
- Conjugacy is convenient but only found in simple problems
- Bayes allows sequential updating, prediction...
- More to specify: you need a

Lessons from the 'surgical' data set

In the 12 hospitals, the 'raw' mortality rates vary between (hospital A) and =0.1442 (H); the aggregated rate is 208/2814= . What are the 'true' rates in hospitals A and H?

Non-Bayesian answer 1. Assume that in hospital i, the number of deaths $Y_i \sim \text{Bin}(n_i, p_i)$, independently. The maximum likelihood estimator of p_i is $Y_i/n_i = 0$ for A and for H.

Non-Bayesian answer 2. Assume that in hospital i, the number of deaths $Y_i \sim \text{Bin}(n_i, p)$. The maximum likelihood estimator of p is $(\sum_i Y_i)/(\sum_i n_i) =$, which applies to both A and H.

Could the p_i all be equal? If p is 0.0739, the chance that Y_{H} is as big or bigger than 31 is . So, !

Bayesian answer 1. Assume in addition that *a priori*, $p_i \sim \text{Beta}(\alpha, \beta)$ where α and β are say 4 and 46. (This gives a mean and variance for the Beta distribution roughly comparable to the sample mean and variance of the raw mortality rates). Then the posterior mean of p_i is $(Y_i + \alpha)/(n_i + \alpha + \beta)$ = for A and H. (See Statistics 2 for this theory).

Bayesian answer 2. Making a similar prior assumption on p, we get the posterior mean .

Which is best? Note that the Bayesian estimates are 'shrunk' towards the prior mean $\alpha/(\alpha+\beta) =$, to an extent depending on the 'denominator' n_i or n. This eliminates ridiculous conclusions like $p_{\mathrm{A}} =$. However, it is still the case that only the data from hospital is used in estimating . Surely the other hospitals' data carries information too? (For example, suppose that was missing: would you be able to guess its value better after having observed the other data?)

Hierarchical models

The methods in this course will allow us to do better, because we will be able to assume in advance that the true mortality rates across the hospitals are (because the circumstances, patients, doctors, ... are different), but (because the operations, disease, ... are the same). The effect we will see is that the raw estimates are shrunk *towards each other*.

To do this, we need to deal with more than two sorts of variable – the parameters and data of ordinary Bayesian models. The hospitals problem has of uncertainty – the hazard of this type of operation, the variability between hospitals, and chance factors in an individual patients' operation. Such models are called *hierarchical*.

What else do hierarchical models address?

Real data about real systems are complex: classic statistical methods are not enough. Among the features that real data might have that we could begin to handle after this unit are:

measures,

- heterogeneity between individuals,
- explanatory variables at individual and group level,
- measurement errors, instruments,
- missing data,informative censoring,

or temporal structure.

1. Conditional independence

As a simple example to introduce ideas, suppose you take a fair coin from your pocket and toss it 10 times, and get heads. What is the chance that the next toss gives a head? Of course, there is still a chance of getting a head or a tail, and the probability of the next toss being a head is independent of the fact that you have already observed 10 heads.

Now suppose you have two coins in your pocket: coin A is biased in favour of heads, and coin B is biased to the same extent in favour of . You pick a coin at random and get a head on each of 10 tosses. What is the chance now of getting a head on the next toss?

If you knew which coin had been selected, this probability would be either 0.8 (coin A) or 0.2 (coin B) and would again be of the number of previous heads. However, the coin is unknown, and this induces a dependency between the outcomes of the first 10 tosses and the toss. Having observed 10 heads you have a strong suspicion that you have picked coin , which in turn increases your belief that you will get a head on the next toss.

The physical probability of getting a head on the next toss is unaffected by the outcomes of previous tosses; it is only our about this probability that depends on the previous outcomes. This can all be formalised using .

Notation for distributions

In this course we will have to represent distributions of many variables, and conventional notation can be cumbersome. So we will make use of abbreviations. For a random variable X taking values x, instead of the probability mass function or the probability density function in the discrete and continuous cases, respectively, we will simply write $p(x)$. In general we will blur the distinction between and .

Similarly we write $p(x, y)$, $p(y|x)$, etc.

Conditional independence

We say x and y are independent if

$$p(x, y) =$$

(understood to mean that the equality holds for all values of x and y). The intuitive meaning is that observing tells you nothing about the value of (and vice versa). We can write this: x y.

We say x and y are conditionally independent given z if

$$p(x, y|z) = \qquad .$$

This means that if you already know , observing x tells you nothing about the value of y. In symbols: $x \perp y \mid z$.

Example. If x is the number of heads in the first 10 throws, y the number in the next throw (0/1), and θ is the probability of getting a head on any single throw, then x and y are

, but

- conditionally given .

$$p(x|\theta) = \binom{10}{x} \quad (1-\theta)^{10-x}; \quad p(y|\theta) = \theta^y (1-\theta)^{1-y}$$

Properties of conditional independence

We can show that conditional independence satisfies various properties, which are quite intuitive:

1. If $x \perp y \mid z$ then .
2. If $x \perp y \mid z$ and then $x \perp u \mid z$.
3. If $x \perp y \mid z$ and $u = g(y)$ then $x \perp y \mid$.
4. If $x \perp y \mid z$ and $x \perp w \mid (y, z)$ then $x \perp \mid z$.
5. *Under extra conditions,* if $x \perp y \mid (z, w)$ and $x \perp z \mid (y, w)$ then $x \perp \mid w$.

Proof of (4). We have (a) $p(x, y|z) = p(x|z)p(y|z)$ and (b) $p(x, w|y, z) = p(x|y, z)p(w|y, z)$. Thus

$$
\begin{aligned}
p(x, y, w|z) &= p(x, y|z)p(w|x, y, z) \quad \text{[always true]} \\
&= p(x|z)p(y|z) \times p(w|x, y, z) \quad \text{[by]} \\
&= p(x|z)p(y|z) \times p(x, w|y, z)/p(x|y, z) \quad \text{[always true]} \\
&= p(x|z)p(y|z) \times \quad \text{[by (b)]} \\
&= p(x|z)p(y, w|z) \quad \text{[always true], as required.}
\end{aligned}
$$

Conditional independence axioms

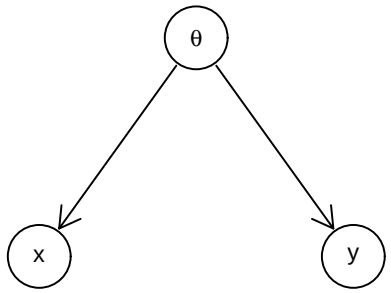
We will not pursue this question in this course, but it is possible to build an entire theory of information, without direct reference to probability, through a concept of conditional independence *defined* by using properties (1)–(5) as .

2. Graphical modelling

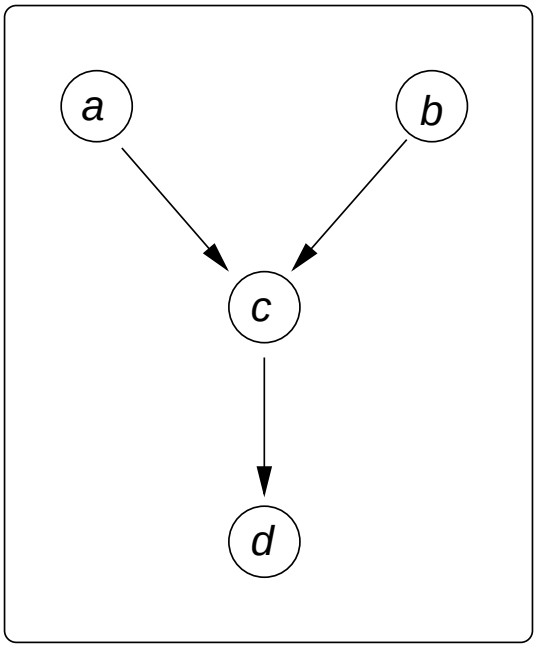
Graphical modelling is concerned with representing the conditional independence relations among a collection of random variables graphically. A graph is a diagram consisting of (called nodes or vertices), joined by (called arcs or edges), which may or may not have indicating direction. In graphical modelling, the vertices represent random variables, and the edges collectively indicate the conditional independence properties.

There are several kinds of graph used in this subject, but we will concentrate on *directed acyclic graphs* (, that is graphs in which all edges are directed, and there are no (directed) loops (in particular, no edges from a vertex to).

Coin-tossing example. This graph represents , the conditional independence of x and y given θ. We have $p(x, y|\theta) = p(x|\theta)p(y|\theta)$, so $p(\theta, x, y) = \quad p(x|\theta)p(y|\theta)$.



This graph is drawn to symbolise the statement that $p(a, b, c, d) = p(a)p(b)p(c|a, b)p(d|c)$. Since it is always true that $p(a, b, c, d) = p(a)p(b|a) \times p(d|a, b, c)$, it indicates that we are assuming and $d \perp (a, b) \mid c$.



It is always true that

$$p(x_1, x_2, \ldots, x_n) = p(x_1) \prod_{i=2}^{n} p(x_i | x_1, x_2, \ldots, \quad)$$

but it is useful and relevant to build models in which not all of $\{x_1, x_2, \ldots, x_{i-1}\}$ are involved in the condition in the ith factor.

The DAG corresponds to a way of factorising $p(x_1, x_2, \ldots, x_n)$: we draw arrows to indicate which variables are needed in each factor: e.g. if the 'd' factor is $p(d|c)$ not $p(d|a, b, c)$ it means we do have arrows from a or b to .

Thus absence of arrows indicates conditional independence.

Parents and children

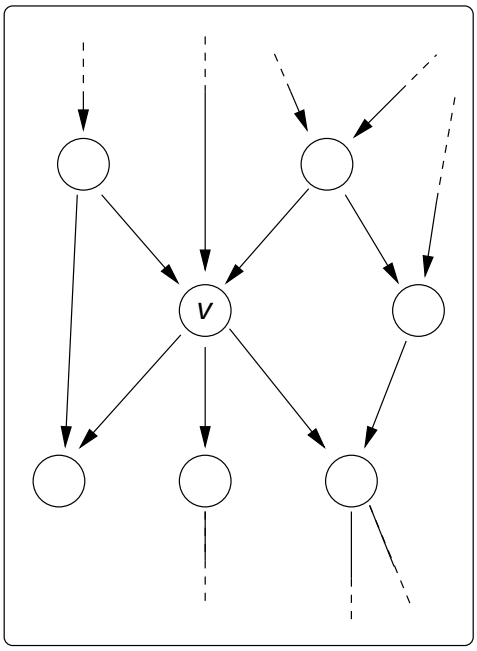
One important application of DAGs is to models of the genes of related individuals. A standard assumption is that, given its parents' genotypes, an individual's genotype is independent of those of the grandparents (and of brothers, sisters, aunts, ...)

We borrow the terminology more generally: the variables from which an arrow leads to variable x are the *parents* of X, denoted . Those x for which $\in \mathrm{pa}(x)$ are the *children* of y. Variables with no parents are called .

With a vector x of variables indexed by v, we will often write $\mathrm{pa}(v)$ rather than $\mathrm{pa}(x_v)$ for simplicity. We also write for the sub-vector consisting of the components indexed by $v \in A$, so $x_{\mathrm{pa}(v)}$ stands for $\{x_w : \in \mathrm{pa}(x_v)\}$.

A more general directed acyclic graph (DAG). It symbolises the fact that the joint distribution of all variables factorises:

$= \prod_v p(x_v | x_{\mathrm{pa}(v)})$.



Example: mortality in infant cardiac surgery

We might assume:

- Deaths in hospital i are $\sim \mathrm{Bin}(n_i, \theta_i)$, independently
- Rates are independent $\sim \mathrm{Beta}(\alpha, \beta)$
- α and β are independent .

The joint distribution of all variables is then

$$p(\alpha, \beta, \theta_1, \theta_2, \ldots, \theta_{12}, y_1, y_2, \ldots, y_{12}) = p(\alpha)p(\beta) \prod_{i=1}^{12} \prod_{i=1}^{12} p(y_i|\theta_i)$$

(we don't usually bother to mention quantities like $\{n_i\}$).

DAGs and full conditionals

In a collection of random variables, the *conditional* (distribution) of any variable means the conditional distribution of that variable given *all* others.

In the coin-tossing example, we might be interested in $p(\theta|x,y)$, $p(x|\theta,y)$ or $p(y|\theta,x)$ – these are all full conditionals.

In the general case, if x_v is one component of a vector x of variables, we write for the sub-vector of all the remaining variables, other than x_v, that is $x_{j:j\neq v}$. So the full conditional for x_v is $p(x_v|$ $)$.

Note that $p(x_v|x_{-v}) = p(x_v, x_{-v})/p(x_{-v}) =$ $/p(x_{-v})$ and that, *regarded as a function of* x_v *alone,* this is simply proportional to the joint distribution $p(x)$.

If the joint distribution is written as a product of factors, then the full conditional for any variable x_v is proportional to the product of just those factors that include .

But in a DAG, we have such a product of factors:

$$p(x) = \prod_v p(\quad |x_{\mathrm{pa}(v)})$$

so that the full conditional satisfies:

$$\propto p(x_v|x_{\mathrm{pa}(v)}) \times \prod_{w:} \quad p(x_w|x_{\mathrm{pa}(w)}),$$

simply the terms in the DAG factorisation corresponding to v itself, and its children (=$\{w : v \in \mathrm{pa}(w)\}$).

Example: mortality in infant cardiac surgery, ctd.

$$p(\theta_1|\alpha,\beta,\theta_2,\ldots,\theta_{12},y_1,y_2,\ldots,y_{12}) \propto$$
$$\propto \theta_1^{\alpha-1}(1-\theta_1)^{\beta-1}\theta_1^{y_1}(1-\theta_1)^{n_1-y_1}$$
$$\text{i.e. } \theta_1|\cdots \sim \qquad (\alpha+y_1,\beta+n_1-y_1)$$

$$p(\alpha|\beta,\theta_1,\theta_2,\ldots,\theta_{12},y_1,y_2,\ldots,y_{12}) \propto$$

$$\propto e^{-\alpha}\prod_{i=1}^{12}\left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)}\theta_i^{\alpha-1}\right]$$

Extended families: the Markov blanket

So which variables do you actually need to know the values of to calculate the full conditional ? Since this is

$$\propto \qquad \times \prod_{w:} \quad p(x_w|x_{\mathrm{pa}(w)}),$$

the answer is: , $\{w : v \in \mathrm{pa}(w)\}$ and $\{\mathrm{pa}(w) : v \in \mathrm{pa}(w)\}$ – i.e. the and '*spouses*' of x_v. We call this set of variables the blanket of x_v.

Later this will be important in algorithms to fit DAG models.

The Markov property for DAGs

Earlier we commented in a genetics context that, given its parents' genotypes, an individual genotype was independent of certain others.

This is an example of a , analogous to the property in a Markov chain that $x_{n+1} \perp (x_{n-1}, x_{n-2}, \ldots) \mid$.

The general Markov property for a DAG is that each individual variable is independent of its , given its parents, i.e.

$$x_v \perp x_{\mathrm{nd}(v)} \mid \quad .$$

The non-descendants are all the variables, except for x_v, its parents, and its descendants (those w for which v is a parent of a parent of ... of w).

Proof

Partition the variables into 4 disjoint subsets: $\mathrm{nd}(v)$ and the descendants of v, denoted $\mathrm{de}(v)$. We know $p(x) = \prod_w p(x_w | x_{\mathrm{pa}(w)})$: write this as the product of 4 products:

$$p(x_v|x_{\mathrm{pa}(v)}) \prod_{w \in} p(x_w|x_{\mathrm{pa}(w)}) \prod_{w \in} p(x_w|x_{\mathrm{pa}(w)}) \prod_{w \in} p(x_w|x_{\mathrm{pa}(w)})$$

Marginalise (sum or integrate) out all the variables in – the 4th term becomes . Now note that or its descendants cannot appear in any of the factors in the term (by the acyclicity of a DAG) or term (by definition of non-descendant). So the 2nd and 3rd terms are functions only of $x_{\mathrm{pa}(v)}$ and $x_{\mathrm{nd}(v)}$. Thus $p(x_v, x_{\mathrm{pa}(v)}, x_{\mathrm{nd}(v)})$ is a product of functions of $(x_v, x_{\mathrm{pa}(v)})$ and of $(x_{\mathrm{pa}(v)}, x_{\mathrm{nd}(v)})$ (no term involving both and). This means $x_v \perp x_{\mathrm{nd}(v)} \mid x_{\mathrm{pa}(v)}$, as required. (See sheet 1, question 3).

More on graphical modelling

Directed acyclic graphs are a natural representation of the way we **usually** *specify* a statistical model (directionally, disease $\rightarrow$, past $\rightarrow$, parameters $\rightarrow$), but

- sometimes (e.g. spatial models) there is *no natural direction*;
- in *understanding associations* between variables implied by a model, however specified, directions confuse;
- and these associations represent the full conditionals needed in *setting up MCMC* methods.

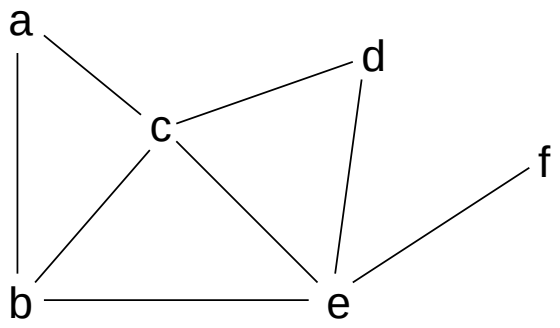
Conditional independence graph: draw an (undirected) edge between variables a and b if they are **not** conditionally independent given all other variables.

Markov properties

The Markov property is familiar from temporal stochastic processes, where we learn that it may be expressed in several equivalent ways. For variables located on an arbitrary graph, the situation is more subtle: can distinguish 4 related properties, each capturing an aspect of Markovness.

pairs of variables are conditionally independent given the rest (see definition of graph).

Local Conditional only on variables (neighbours), each variable is independent of all others (this simplifies full conditionals).



$$P: \quad \perp \quad |(a,b,d,e) \qquad L: \quad \perp (a,b,f)|$$

$$F \quad : \quad p(a,b,c,d,e,f) = \\ \psi_1(a,b,c)\psi_2(b,c,e)\psi_3(c,d,e)$$

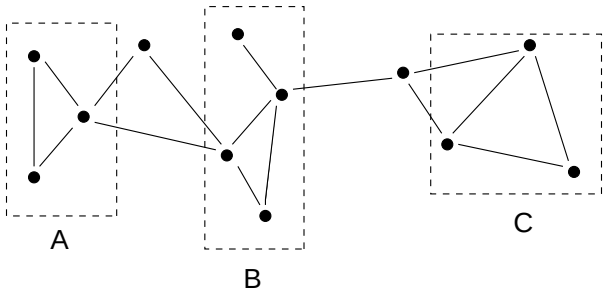
Global Any subsets of variables separated by a third are conditionally independent given the values of the subset.

Factorisation The joint distribution factorises as a product of functions on (=maximal complete subgraphs).

Mathematically, the interesting thing is that these are , although $F \Rightarrow G \Rightarrow L \Rightarrow P$ always.

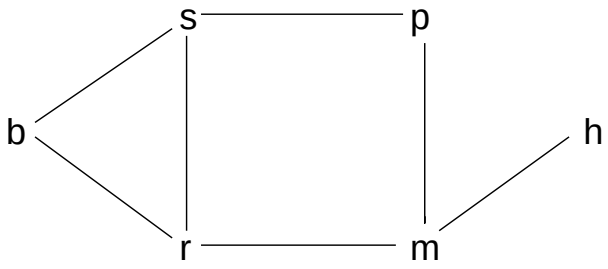
But for most (statistical) purposes, the important thing is that they are often ; a necessary and sufficient condition is that property 5 on slide 17 holds.

This result includes the "Hammersley-Clifford theorem" (Markov random field = Gibbs distribution, $L = F$).



G : $\perp$ $\mid$

Example: using CI graphs to summarise dependencies.
Edwards and Havranek (*Biometrics*, 1985) summarise their conclusions from a contingency table like this:



b	blood pressure > 140?	s	smoking?
p	strenuous physical work?	m	strenuous mental work?
r	ratio of α and β lipoproteins > 3?		
h	family history of coronary heart disease?		

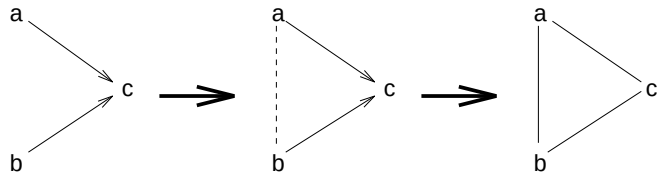
Reading off all conditional dependencies from a model specified as a DAG

(to understand associations, or to help construct a MCMC method)

Need to "moralise" the graph:

parents

2. Drop directions



So the neighbours of a variable (that is, its) are again seen to be its parents, children and spouses.

3. Bayesian hierarchical modelling

How to make inference on multiple parameters { , . . . , } measured on I units (persons, centres, areas, ...) *which are related or connected by the structure of the problem ?*

We can identify three different assumptions:

1. **Identical parameters:** All the θ's are identical, in which case all the data can be and the individual units ignored.

2. **Independent parameters:** All the θ's are entirely unrelated, in which case the results from each unit can be analysed (for example using a fully specified prior distribution within each unit)

   $\rightarrow$ individual estimates of θ_i are likely to be (unless very large sample sizes)

3. **Exchangeable parameters:** The θ's are assumed to be 'similar' in the sense that the 'labels' convey

Exchangeability and de Finetti's theorem

An infinite sequence of 0/1 random variables $x_1, x_2, \ldots$ is called (infinitely) if any finite subset has a joint distribution that is the same whatever the order in which the variables are written. E.g. $p(x_4, x_7, x_9) = p(x_7, x_9, x_4)$.

If the variables are independent Bernoulli(θ), they are obviously exchangeable. This remains true if is random (as in the coin-tossing example, with two biased coins), since e.g.

$$p(x_4, x_7, x_9) = \qquad \theta^{x_4}(1-\theta)^{1-x_4}\theta^{x_7}(1-\theta)^{1-x_7}\theta^{x_9}(1-\theta)^{1-x_9}$$

(in the case θ has a continuous distribution), and this obviously only depends on , not the order they appear.

The remarkable thing is that the converse of this is true – the way to get infinitely exchangeable 0/1 random variables is by Bernoulli trials with a fixed or random θ. This is (a form of) de Finetti's theorem. There are more general versions of the theorem, not just for 0/1 variables.

It gives mathematical support for using hierarchical models: if your prior beliefs about a set of parameters (e.g. the hospital mortality rates $\{\theta_i\}$) are exchangeable (really just a assumption), then without loss of generality you can model them as i.i.d. from some distribution given , and then make random.

$$p(\theta_1, \theta_2, \ldots, \theta_I) = \prod_{i=1}^{I} p(\theta_i|\phi)$$

Implications of de Finetti's theorem

Under broad conditions an assumption of exchangeable units is mathematically to assuming the θ's are drawn at random from some population distribution, just as in a traditional "random effects" model.

We do **not** have to assume that the units (e.g. hospitals) are actually drawn randomly from any population: it is our about the $\{\theta_i\}$ that are exchangeable.

Exchangeability

'Exchangeability' is a formal expression of the idea that we find no systematic reason to distinguish the individual random variables $\theta_1, ..., \theta_I$ – a *judgement* that they are 'similar' but not identical.

Shrinkage and hierarchical models

Suppose in each unit we observe a response x_i assumed to have a Normal likelihood

$$\sim \mathrm{N}(\quad, \tau_i^2)$$

Unit means θ_i are assumed to be exchangeable, and to have a Normal distribution

$$\theta_i \sim$$

where μ and σ^2 are 'hyper-parameters', for the moment assumed , as are τ_i^2.

After observing x_i, Bayes' theorem gives (see Statistics 2 or BMA)

$$\sim \mathrm{N}(w_i\mu + (1-w_i)x_i, (1-w_i)\tau_i^2)$$

where $= \tau_i^2/(\tau_i^2+\sigma^2) \in (0,1)$ is the weight given to the prior mean.

A Bayesian model therefore leads to inferences for each θ_i giving intervals that are than in the non-Bayesian approach, but towards the prior mean response. w_i controls both the 'shrinkage', and the reduction in the width of the interval: it depends on precision of the individual unit i relative to the variability between units. When $\{\tau_i^2\}$ are also given a prior, the same principles apply, although the solution is less explicit.

In a hierarchical model, μ and σ^2 are , and the effect of this is more complicated again, and *best seen numerically*; the amount of shrinkage is not determined in advance – it is discovered from the (an automatic consequence of Bayes' theorem). μ will also be shrunk towards the data in its posterior distribution, so that the are now shrunk towards a "typical" x value.

More analysis of the questions raised by ‘surgical’ example

Our initial Bayesian models revealed difficulty with the assumption that there was a common mortality rate in every hospital; we asked:

- Does this model adequately describe the random variation in outcomes for each hospital?
- Are the hospital failure rates more variable than our model assumes?

and concluded ’ ’ and ’ ’, respectively.

Some general reasons for excess variation in response

- Individual i.e. systematic differences between units which are not attributable to random variation
  - this concept is often termed *frailty* in survival analysis
  - for binary / count data this is often termed *overdispersion*

  response measurements from the same unit tend to be *correlated*

  ⇒ 2 responses from the same unit will be more alike than 2 responses from different units

  ⇒ variation in responses is not completely random
- Failure to a relevant explanatory variable

  measurement of relevant explanatory variables

Modelling the excess variation

Perhaps we could modify our beta-binomial model to allow for a *different* failure probability, for each hospital i:

$(y_i \mid \theta_i) \sim$ Binomial$(n_i,$ $)$ where

Interpretation:

- $\{\theta_i\}$, the 'true' surgical failure rate in the hospitals are viewed as a from a common *population distribution*

  $\Rightarrow$ hospital failure rates are assumed to be **similar** but not

  – Beta(a, b) prior describes the distribution of surgical failure rates amongst the 'population' of hospitals

How would you specify values for ?

Approximate 'empirical Bayes' approach

- Calculate crude failure rates
- Calculate the observed mean and variance of the 12 values y_i/n_i
- Solve for to obtain a beta distribution with this mean and variance
- Using Beta( ) as a prior, apply Bayes theorem to obtain posteriors for true failure rates θ_i, $p(\theta_i|\widehat{a}, \widehat{b}, \boldsymbol{y})$

Potential problems with this approach:

- We are using the data :

  to estimate the prior

  to estimate θ_i for each hospital

  $\Rightarrow$ precision of our inference

- Using any point estimate for a and b ignores some posterior about the population distribution of the θ_i's

Full hierarchical Bayes approach

- Assume a *joint probability model* for the entire set of parameters $(\theta_1, \theta_2, \ldots, \theta_I, \quad)$
  – requires us to assign known prior distributions to a, b, e.g.

$$a \sim \text{Exponential}(1) \quad \text{and} \quad b \sim \text{Exponential}(1)$$

- Apply Bayes theorem to calculate the joint posterior distribution of all the unknown quantities .

Level 1: $\sim$ Binomial(n_i, θ_i), independently for each i

Level 2: $\sim$ Beta(a, b), independently for each i

Level 3: Prior for

Advantages of this approach

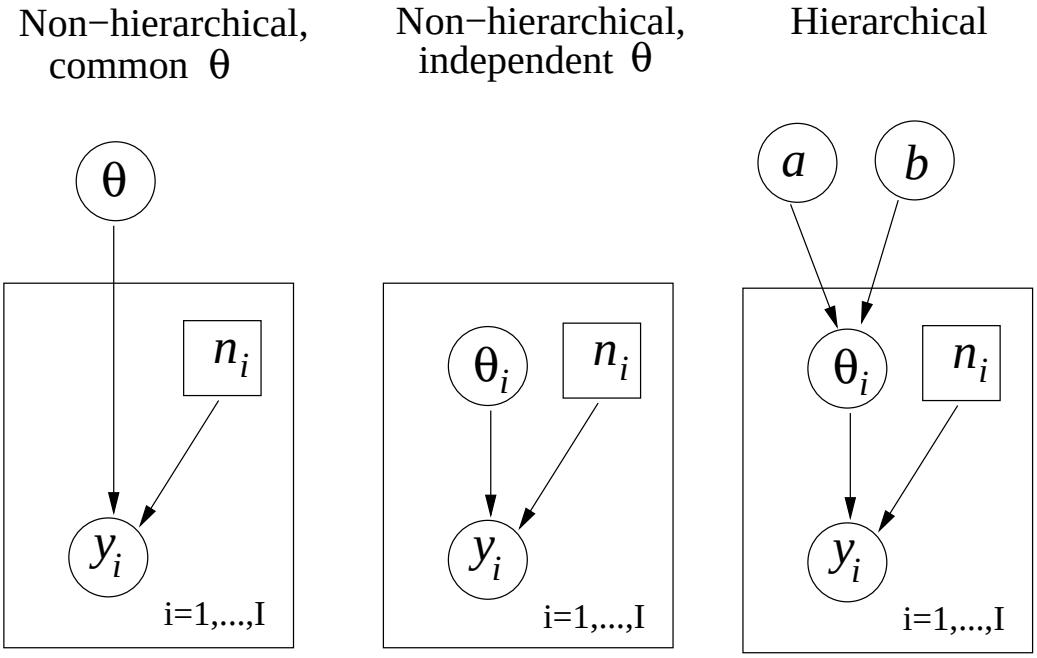
The posterior distribution for each θ_i

from the likelihood contributions for hospitals, via their joint influence on the estimate of the unknown population (prior) parameters a and b

- reflects our uncertainty about the true values of a and b

Such models are also called *Random effects* or *Multilevel* models.

Graphical models (DAGs) for surgical example



Alternative prior choice for the surgical rates

- Beta prior is convenient choice in non-hierarchical models due to conjugacy, but no need to be restricted here
- More natural to work within generalised linear (mixed) model framework for logit$(\theta_i) =$ (e.g. extends to allow explanatory variables).

Likelihood (Level 1)	y_i	$\sim$	Binomial(n_i, θ_i)
Exchangeable rates (Level 2)	logit θ_i	$\sim$	
Hyperpriors (level 3)	μ	$\sim$	Uniform or Normal
	σ^2	$\sim$	'Vague' Gamma (see later)

Summary: why hierarchical?

Many interlinked arguments to favour the use of hierarchical models:

- by the problem in layers, able to structural judgments on observables, on parameters and subjective information

  of hyperparameter choice $\rightarrow$ "robustify" the inference
- natural structure for expressing , prior correlations, ... in a plausible way (see next lectures)
- through and borrowing of , parameter estimates are

4. Computational Bayesian Inference

- Bayesian inference centres around the distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$, which is fully by the prior, likelihood and data
- We can usually write down the functional form of the posterior distributions required for Bayesian inference
- For most models, these functions are complex and high dimensional:
  - Difficult to work with them
- What do we actually need to be able to do?
  (cf. maximum likelihood, for which we need to).

Bayesian inference needs integration

$$p(\theta_1|\boldsymbol{y}) = \frac{p(\phi)\prod_{i=1}^{I}[p(\theta_i|\phi)p(y_i|\theta_i)]}{\int\int\cdots\int p(\phi)\prod_{i=1}^{I}[p(\theta_i|\phi)p(y_i|\theta_i)]d\phi d\theta_2\cdots d\theta_I}$$

$$E(\theta_1|\boldsymbol{y}) = \quad \theta_1 p(\theta_1|\boldsymbol{y})d\theta_1$$

$$P\{\theta_1 > 4.65|\boldsymbol{y}) = \int_{4.65}^{\infty} p(\theta_1|\boldsymbol{y})$$

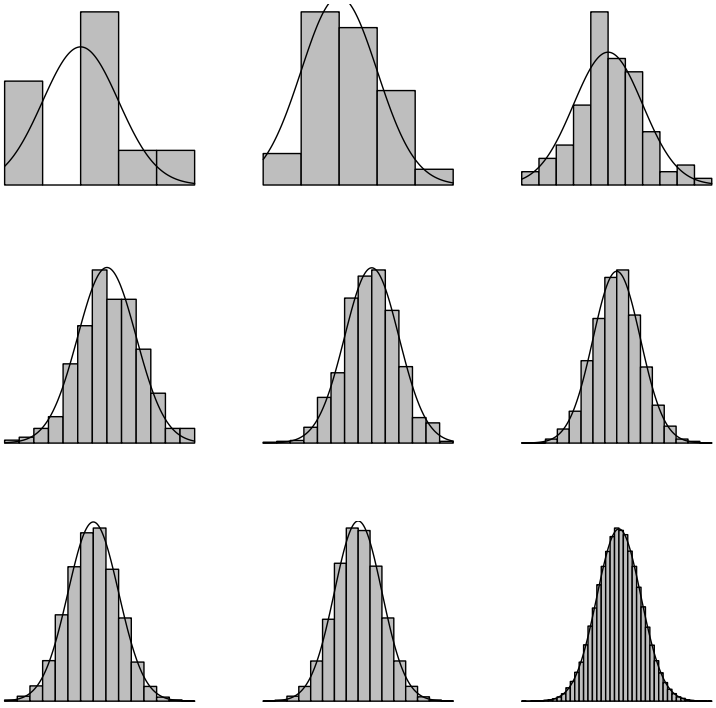
There are approximation methods, and methods for integration, but these prove to be to use in statistical modelling (role of data, many models to entertain).

The alternative to exact methods or classical approximations: simulation

- Instead, it is often straightforward to realisations from the required posterior distributions, that is to use methods
- Posterior summaries (e.g.) are easily obtained by simple data summaries of the simulated values
- Ordinary Monte Carlo methods are not usually practicable for realistically complex models, but Markov chain Monte Carlo ( ) methods are much more , to set up, and (with care) reliable

Overview of Simulation Methods for Bayesian Inference

- Imagine generating a random sample of values from a probability distribution (e.g. normal);
- Construct a from the sample;
- If the sample is large enough, histogram can provide virtually information about the distribution from which these samples were drawn:
  - Mean, variance, percentiles of sample ≈ mean, variance, percentiles of .
- Monte Carlo methods enable us to generate samples from the posterior distributions of model parameters:
  - Samples can be summarised to estimate properties (e.g. mean, variance, percentiles) of the posterior distribution.



Monte Carlo Integration

Suppose we can draw samples from a distribution for $\boldsymbol{\theta}$, *i.e.*

$$\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, ..., \boldsymbol{\theta}^{(N)} \sim$$

Then

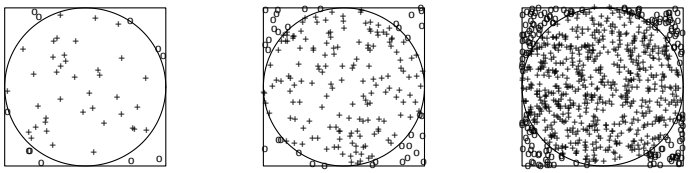
$$E(g(\boldsymbol{\theta})) = \int g(\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$$

$$\approx \qquad\qquad = \bar{g}_N$$

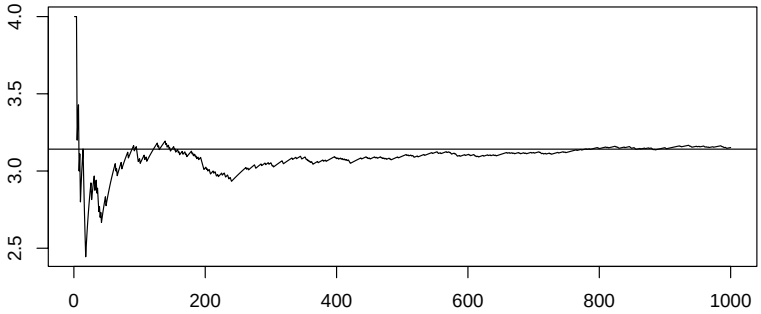
this is so-called "crude" Monte Carlo integration.

There are theorems (the law of large numbers and generalisations) which prove convergence in the limit as – even if the sample is dependent (we need this for *Markov chain* Monte Carlo).

Dropping points at random into a square.



$4\times$ cumulative proportion lying in circle.



Extracting information from a simulated sample

Assume we have a sample $(\theta_k^{(1)}, \theta_k^{(2)}, \ldots, \theta_k^{(N)})$ from $\pi(\theta_k)$:

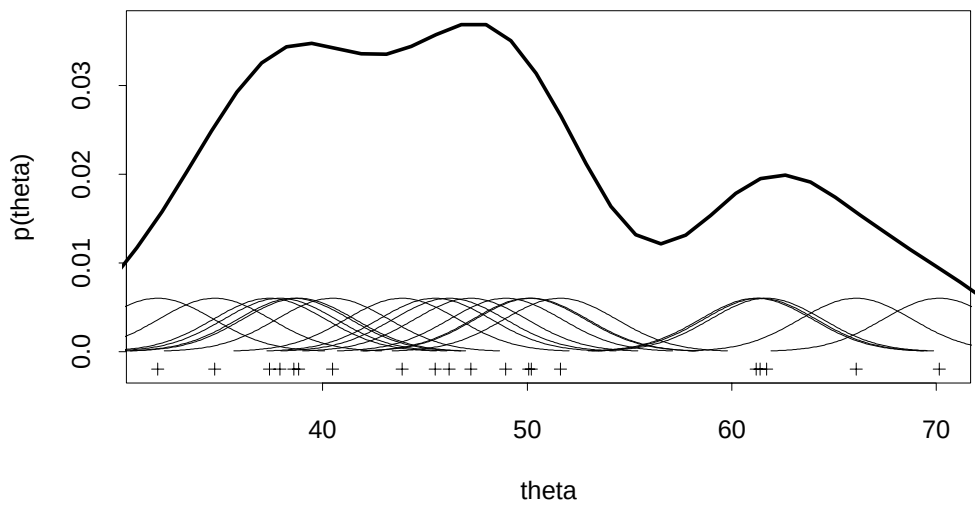
Expectation:

$$E_\pi(\theta_k) \approx \frac{1}{N} \sum_{t=1}^{N}$$

Quantile:

$$P_\pi(\theta_k \leq x) \approx \frac{1}{N} \#\{t : \qquad \leq x\}$$

Kernel density estimates: $\widehat{\pi}(\theta_k) \approx \frac{1}{N}\sum_{t=1}^{N} h(\theta_k;\quad)$



Monte Carlo integration more generally

Suppose we want to evaluate an integral $= \int_0^1 g(\theta)d\theta$. The crude Monte Carlo approach above says we can simulate i.i.d. $(0,1)$ random numbers $\theta^{(1)}, \theta^{(2)}, \ldots$ and estimate by $(1/N)\sum_{t=1}^{N} g(\theta^{(t)})$. The Law of Large Numbers (LoLN) (see Statistics 2) says that

$$N^{-1}\sum_{t=1}^{N} g(\theta^{(t)}) \rightarrow$$

as $N \rightarrow \infty$. Also, the Central Limit Theorem (CLT) says that if $\sigma^2 = \text{var}(g(\theta)) < \infty$ then

$$N^{-1}\sum_{t=1}^{N} g(\theta^{(t)}) \approx \text{N}(\quad,\quad /N) \qquad \text{as } N \rightarrow \infty.$$

Another approach, "hit-or-miss" Monte Carlo, is motivated by the dropping-points-into-a-square example. Enclose the graph of $g(\theta)$ in a rectangle $[0,1] \times [0, M]$ (suppose $g \geq 0$). Then draw *pairs* $(\theta^{(t)}, u^{(t)})$, with both components i.i.d. Uniform$(0,1)$. What is the probability that $\quad < g(\theta^{(t)})/M$? Answer: $\int \int I[u < g(\theta)/M] du\, d\theta = \quad$. Thus $I[u^{(t)} < g(\theta^{(t)})/M]$ are 0/1 variables with expectation G/M, so the LoLN tells us that

$$N^{-1} \sum_{t=1}^{N} \{M \times I[u^{(t)} < g(\theta^{(t)})/M]\} \to \qquad \text{as } N \to \infty.$$

What about the CLT for *this* estimator?

$$N^{-1} \sum_{t=1}^{N} \{M \times I[u^{(t)} < g(\theta^{(t)})/M]\} \approx \text{N}(\quad, \tilde{\sigma}^2/N) \qquad \text{as } N \to \infty,$$

where $\tilde{\sigma}^2 = \text{var}(\{M \times I[u < g(\theta)/M]\})$.

But $\tilde{\sigma}^2 = M^2(G/M)(1 - G/M) =$, while
$\sigma^2 = E(g(\theta)^2) - \quad = \int_0^1 g(\theta)^2 d\theta - G^2$
$M \int_0^1 g(\theta) d\theta - G^2 = G(M - G)$. So crude MC beats hit-or-miss.

Research into simulation methods is mostly about finding ways to

.

Integration of functions over $[0,1]$ is not challenging; they key thing is that all of the points above apply when $\boldsymbol{\theta}$ is a (possibly high-dimensional) vector: you still get unbiased estimates, with standard deviations of order : other numerical integration methods get worse as dimension increases.