The University of Newcastle

Kerrie Mengersen

Introduction to Bayesian Modelling - 1

Armidale 2004

INTRODUCTION TO BAYESIAN MODELLING

- Basics of Bayesian Inference
- Markov chain Monte Carlo
- Introduction to BUGS
- Case studies: mixture models, meta-analysis
- Convergence diagnostics

Acknowledgements

Extracts of these notes are taken from:

- Gelman, Carlin, Stern, Rubin (1995) Bayesian Data Analysis. Chapman&Hall
- Congdon (2002) Bayesian Statistical Modelling. Wiley
- Robert, C.P. (2000) The Bayesian Choice Springer-Verlag
- Casella, G. and Robert, C.P. (2001, 2004) Monte Carlo Methods. Springer-Verlag
- On-line notes by David Madigan

Other texts

Bayes and Empirical Bayes Methods for Data Analysis , *Bradley P.Carlin and Thomas A. Louis. London, U.K.:Chapman & Hall/CRC, 2000.*

Statistics: A Bayesian Perspective *by Donald A. Berry. Belmont, Mass.: Wadsworth Publishing Company, 1996.*

Bayesian Inference in Statistical Analysis , *George E. P. Box and George C. Tiao. New York, N.Y.: John Wiley and Sons, Inc., 1973.*

Bayesian Statistics: An Introduction (2nd edition) , *Peter M. Lee. New York, N.Y.: John Wiley and Sons, Inc., 1997*

Markov Chain Monte Carlo in Practice. *Gilks, W., Richardson, S., Speigelhalter, D. Chapman and Hall, 1995*

A biased set of references!

- Mengersen, K.L. (2004) Markov chain Monte Carlo: An Update. *Encyclopedia of Biostatistics. To appear.*
- Marin, J.M., Mengersen, K.L. and Robert, C.P (2004) Bayesian Modelling and Inference on Mixtures of Distributions. *Edited book by Dipak Dey. To appear.*
- Wolpert, R., Mengersen, K. (2004) Adjusted likelihoods for synthesizing empirical evidence from studies that differ in quality and design: effects of environmental tobacco smoke. *Statistical Science*. To appear.
- Mengersen, K.L. and Robert, C.P. (2003) Population Markov Chain Monte Carlo: the Pinball Sampler. *Bayesian Statistics 7.* Editors, J.O. Berger, A.P. Dawid, A.F.M. Smith, Oxford University Press.
- Casella, G, Mengersen, K, Robert, C P and Titterington, D M. (2002) Perfect Slice Samplers for Mixtures of Distributions. *J.Roy.Statist.Soc.B*. 64(4), 777-790
- Mengersen, K. and Robert, C.P. (1999) MCMC convergence diagnostics: a revieWWW. In Bayesian Statistics VI. Editors J.O. Berger, A.P. Dawid, A.F.M. Smith, Oxford University Press, pp. 415-440.
- J. Besag, P. J. Green, D. Higdon and K. Mengersen (1995) Bayesian computation and stochastic systems, *Statistical Science*, **10**, 3-41. With discussion (41-59) and rejoinder (59-66).

Let's start at the beginning

Recall probability statements:

$P(A)$ is probability of event A

$P(AB)$ is joint probability of events A and B

$P(A|B)$ is probability of A conditional on B

$P(AB) = P(A|B)P(B) = P(B|A)P(A)$

So, $P(A|B) = P(B|A)P(A) / P(B)$

Think of: $A=\theta$ (unknown), $B=y$ (known 'data')

So $P(\theta|y) = P(y|\theta)\ p(\theta) / p(y)$

This is Bayes' Rule!

Example

Human chromosomes: males XY, females XX

Haemophilia exhibits X-chromosome-linked recessive inheritance, so a male who inherits the gene on the X chromosome is affected but a female who carries the gene on only one of the two X chromosomes is unaffected. The disease is usually fatal for women who inherit two such genes, and this is very rare, since the frequency of occurrence of the gene is low in human populations.

Source: BDA

Example: the prior distribution

A woman has an affected brother, which implies that her mother must be a carrier of the haemophilia gene with one ‘good’ and one ‘bad’ haemophilia gene.

Her father is not affected.

Thus the woman has a 50-50 chance of having the gene.

Unknown quantity of interest: whether the woman is a carrier of the gene ($\theta=1$) or not ($\theta=0$).

Based on the information provided so far, the prior distribution for the unknown θ is

$$\Pr(\theta=1) = \Pr(\theta=1) = 1/2$$

Example: model and likelihood

We need some data: the woman has two sons, neither of whom is affected.

Let $y_i = 1$ *or* 0 denote affected/unaffected son.

The outcomes of the two sons are *exchangeable* and, conditional on the unknown θ, are independent: we assume the sons are not identical twins.

→likelihood function

$$\Pr(y_1=0, y_2=0|\theta=1) = (0.5)(0.5) = 0.25$$

$$\Pr(y_1=0, y_2=0|\theta=0) = (1)(1) = 1$$

(OK, there is a nonzero probability due to mutation but we will ignore this)

Example: posterior distribution

$$Pr(\theta=1|y) = p(y|\theta=1)p(\theta=1) / p(y)$$

$$p(y) = p(y|\theta=1)p(\theta=1) + p(y|\theta=0)p(\theta=0)$$

$$= \Sigma p(y|\theta)p(\theta)$$

So:

$$Pr(\theta=1|y) = (0.25)(0.5) / \{(0.25)(0.5)+(1.0)(0.5)\}$$

$$= 0.125 / 0.625 = 0.20$$

In terms of odds:

Prior odds of woman being a carrier is 0.5/0.5=1.

Likelihood ratio based on information about unaffected sons is 0.25/1 = 0.25

So posterior odds are 0.25 × 1 = 0.25.

Converting back to a probability: 0.25/(1+0.25) = 0.2

Example: adding more data

Suppose the woman has a third son, who is also unaffected.

The entire calculation does not need to be redone: we can use the previous posterior distribution as the new prior distribution to obtain:

$$\begin{aligned} Pr(\theta=1|y) &= (0.5)(0.2) \,/\, \{ (0.5)(0.2)+(1)(0.8)\} \\ &= 0.111 \end{aligned}$$

Example: your turn!

- Following from the last slide, if the third son *is* affected, show that the posterior probability of the woman becoming a carrier is 1 (again ignoring the possibility of a mutation).
- Going back to the information from the first two sons, what happens if the prior probability that the woman is a carrier is 0.3? What about 0.9?
- What are the posterior odds corresponding to your calculations?

So this is Bayesian Modelling: Posterior $\propto$ Prior $\times$ Data

- **Likelihood for data y given unobserved θ: p(y|θ)**

  θ can be parameters, missing data, latent variables etc

- **Prior for θ: p(θ)**
- **Want posterior distribution of θ: p(θ|y)**

$$
\begin{aligned}
p(\theta|Y) &= \frac{p(Y\,|\,\theta)\,p(\theta)}{p(Y)} \\
&= \frac{p(Y\,|\,\theta)\,p(\theta)}{\int p(Y\,|\,\theta)\,p(\theta)\,d\theta} \\
&\propto p(Y\,|\,\theta)\,p(\theta)
\end{aligned}
$$

Bayesian Prediction: similar logic

Before the data *y* are considered, the distribution of the unknown but observable *y* is

$$p(y) = \int p(y,\theta)\, d\theta = \int p(\theta)\, p(y|\theta)\, d\theta$$

(marginal distribution, prior predictive dist.)

After the data *y* have been observed, we can predict an unknown value *y*⁻, from the same process (posterior predictive distribution)

$$\begin{aligned} p(y^{-}|y) &= \int p(y^{-},\theta|y)\, d\theta \\ &= \int p(y^{-}|\theta,y)\, p(\theta|y)\, d\theta \\ &= \int p(y^{-}|\theta)\, p(\theta|y)\, d\theta \end{aligned}$$

(because y and y- are conditionally independent given θ)

Example: what is the probability of surgical failure (death) after cardiac surgery on babies?

Sample size (n=148) , deaths (y=8)

Example: Estimating a proportion

- ***Data***: *y* successes from *n* independent trials, eg 18 ‘successes’ out of 148 animals in an animal experiment
- ***Unobserved***: *θ*: proportion of successes
- ***Likelihood***: *p(y | θ)* has Binomial distribution

$$p(y \mid \theta) = \binom{n}{y} \theta^{y} (1-\theta)^{y}$$

- ***Prior***: assume we ‘know nothing’ about θ, so we set a uniform prior θ~U[0,1]
- ***Posterior***:

$$p(\theta \mid y) = \binom{n}{y} \theta^{y} (1-\theta)^{y} \propto \theta^{y} (1-\theta)^{y}$$

- ***Form of posterior***: *θ|y ~ Beta(y+1, n-y+1)*

The Beta Distribution

- Continuous distribution on [0,1]
- $\theta \sim \mathrm{Beta}(\alpha,\beta)$; α,β continuous; $\alpha>0$, $\beta>0$

- $p(\theta) \propto K\, \theta^{\alpha-1} (1-\theta)^{\beta-1}$

  $K = \text{constant} = \Gamma(\alpha+\beta) / \Gamma(\alpha)\Gamma(\beta)$

  (Γ is a mathematical function)
- $E(\theta) = \alpha/(\alpha+\beta)$ **("unbiased est.")**

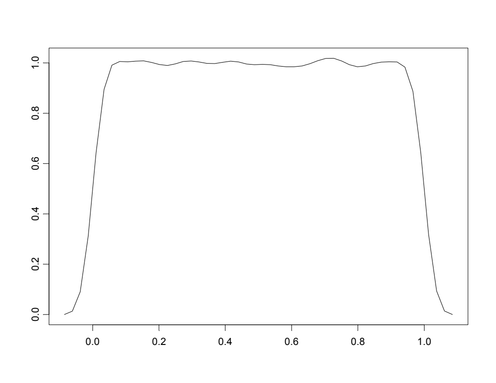
  $\mathrm{Var}(\theta) = \alpha\beta / \{(\alpha+\beta)^2(\alpha+\beta+1)\}$

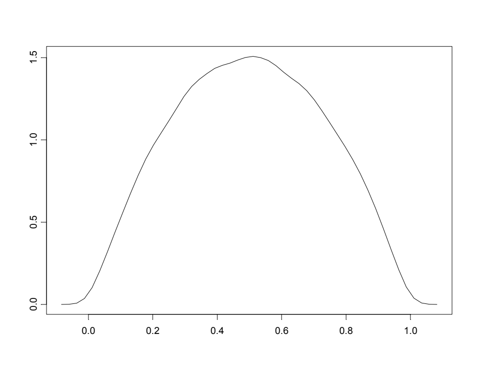
  $\mathrm{mode}(\theta) = (\alpha-1) / (\alpha+\beta-2)$ **("MLE")**
- ***What are the posterior expected value, variance and mode for our example?***

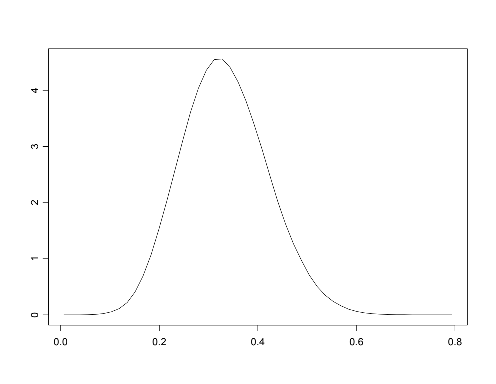
Check out the Beta

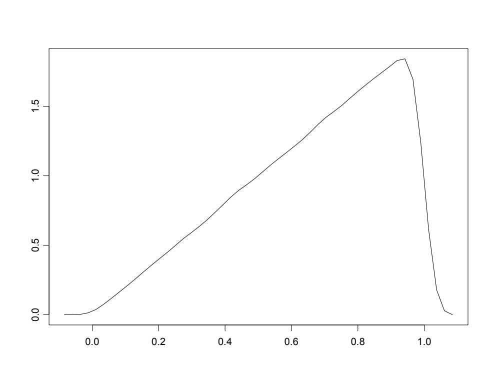
- Match the plots to the distributions. What are the posterior means, modes and variances?

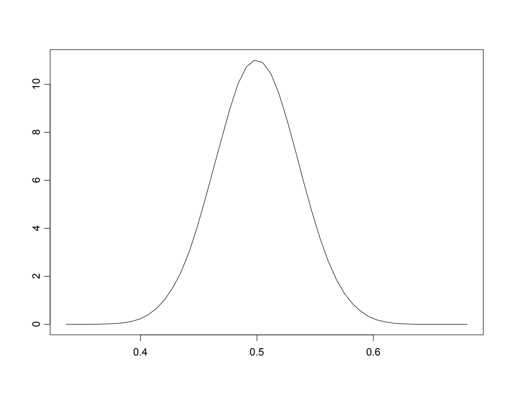
Beta(1,1)   Beta(2,2), Beta(100,100), Beta(2,1), Beta(10,20)











What about prediction?

- With a uniform prior, the prior predictive distribution can be evaluated explicitly: all possible values of *y* are equally likely, *a priori*.
- What about the outcome of one new trail, rather than a set of *n* new trials?
- Let y^- be result of new trial, exchangeable with the first *n*.

$$\begin{aligned}\Pr(y^- = 1|y) &= \int \Pr(y^- = 1|\theta, y)p(\theta|y)d\theta \\ &= \int \theta p(\theta|y)d\theta \\ &= E(\theta|y) = (y+1)/(n+2)\end{aligned}$$

What is this for our example?

Posterior as a compromise

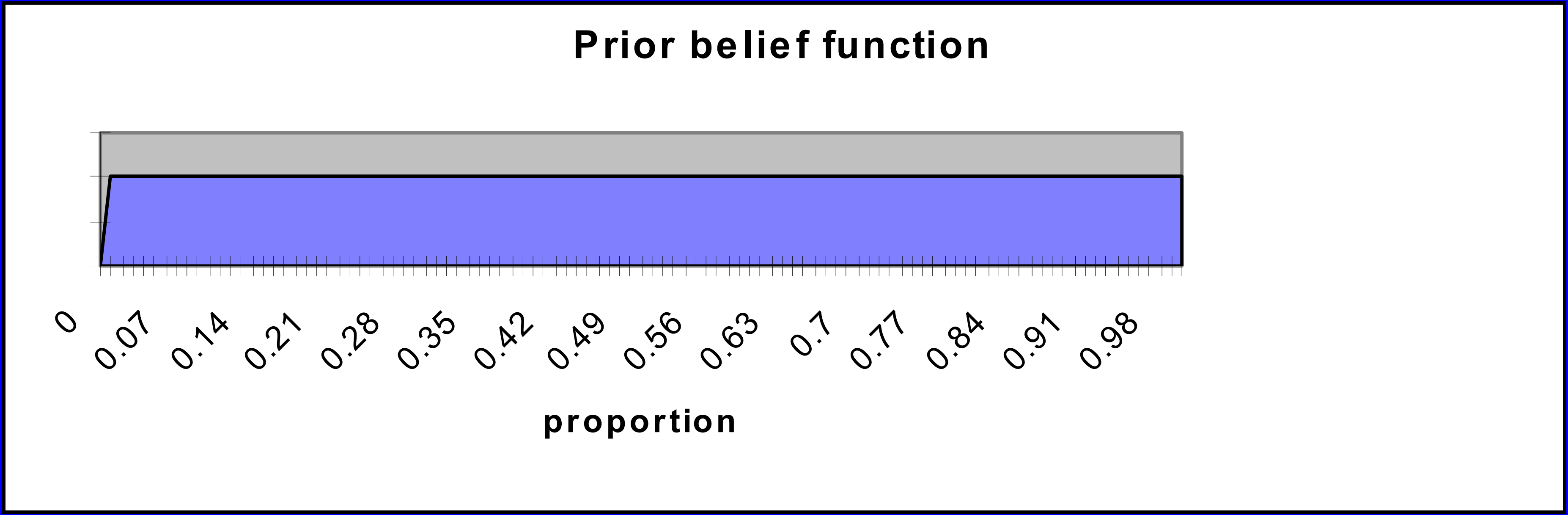
- Look at the mean and variance of θ:

  $E(\theta) = E(E(\theta|y))$; $\mathrm{var}(\theta) = E(\mathrm{var}(\theta|y)) + \mathrm{var}(E(\theta|y))$
- The posterior variance is on average smaller than the prior variance, by an amount that depends on the variation in posterior means over the distribution of possible data. The greater this variation, the more the potential for reducing uncertainty wrt θ.
- *The posterior mean is a compromise between the prior mean and the sample proportion.*

  Confirm this in our example.

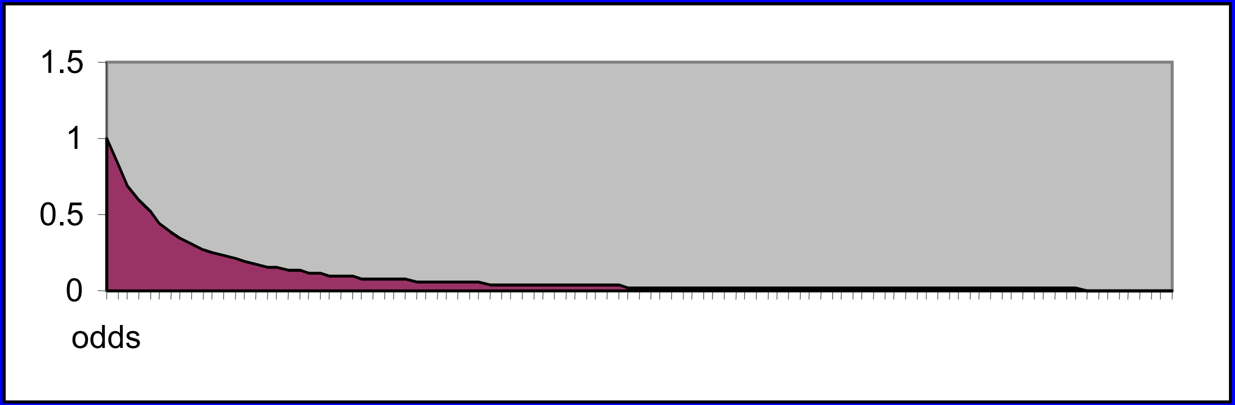
  What happens as the size of the data sample increases?
- General feature of Bayesian inference: posterior distribution is centred at a point that represents a compromise between the prior information and the data, and the compromise is increasingly controlled by the data as the sample size increases.

Does a flat prior mean that I know nothing?



Effect of reparametrisation

Change from θ (proportion) to the odds (O) where O = θ /(1- θ)
Now O ranges from 0 to infinity
A flat prior for θ gives a different picture for the odds



And similarly a flat prior for the odds would give
a different picture for θ

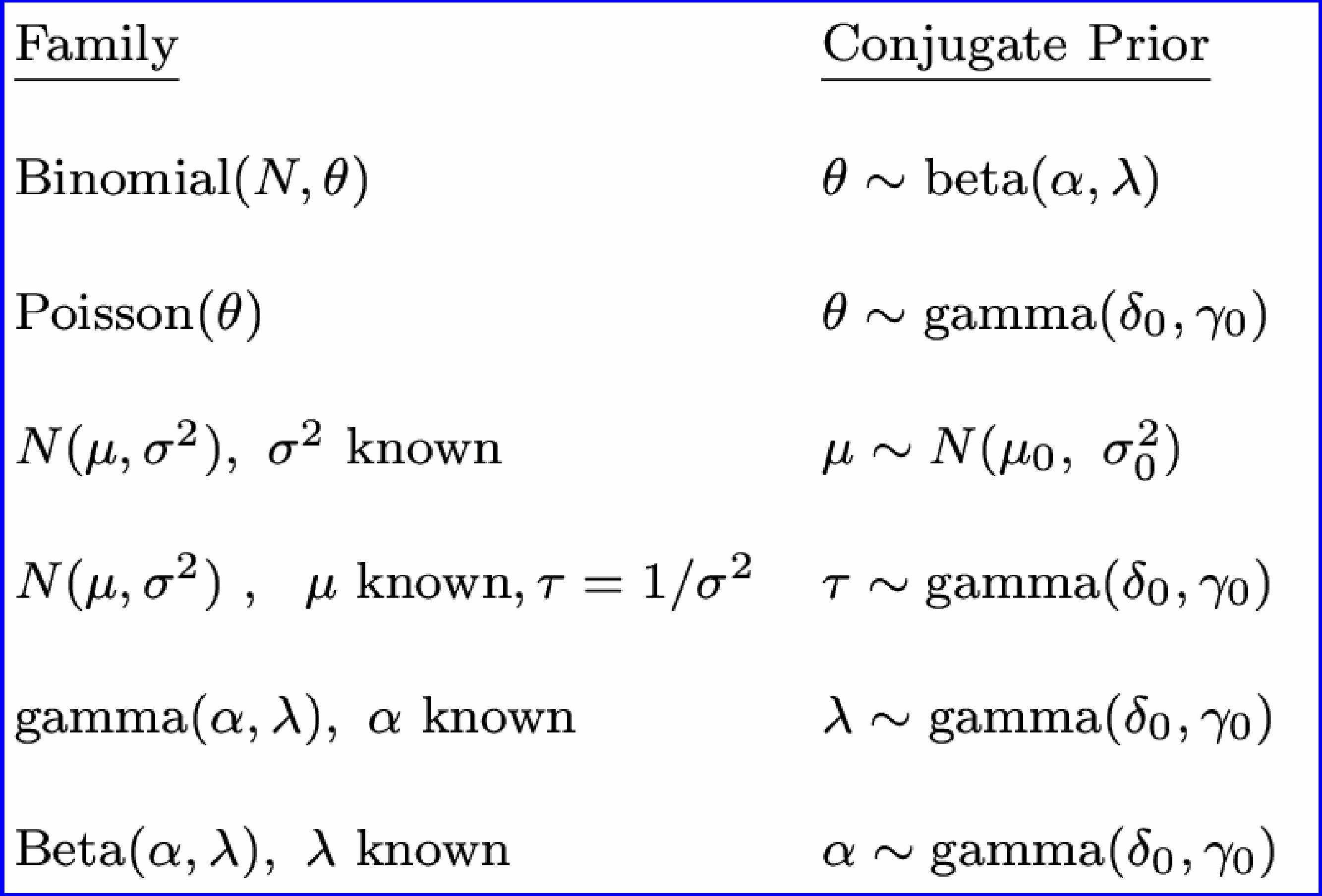
“ignorance about θ” does not imply “ignorance about O”.

The notion of “prior ignorance” may be untenable.

Conjugate priors

- It might be reasonable to expect the posterior distribution to be of the same form as the prior distribution. This is the principle of *conjugacy*
- A conjugate prior for a Binomial likelihood is a Beta distribution: the posterior is then also a Beta distribution

Conjugate priors



Family	Conjugate Prior
Binomial(N,θ)	$\theta \sim \text{beta}(\alpha,\lambda)$
Poisson(θ)	$\theta \sim \text{gamma}(\delta_0,\gamma_0)$
$N(\mu,\sigma^2),\ \sigma^2$ known	$\mu \sim N(\mu_0,\ \sigma_0^2)$
$N(\mu,\sigma^2)\ ,\quad \mu$ known, $\tau = 1/\sigma^2$	$\tau \sim \text{gamma}(\delta_0,\gamma_0)$
gamma$(\alpha,\lambda),\ \alpha$ known	$\lambda \sim \text{gamma}(\delta_0,\gamma_0)$
Beta$(\alpha,\lambda),\ \lambda$ known	$\alpha \sim \text{gamma}(\delta_0,\gamma_0)$

Back to the hospital example

- ***Likelihood*:**

$$p(y|\theta)=\binom{n}{y}\theta^{y}(1-\theta)^{y}\propto\theta^{y}(1-\theta)^{y}$$

- ***Prior*:**

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

- ***Posterior*:**

$$\begin{aligned}p(\theta\,|\,y) &\propto p(y\,|\,\theta)\,p(\theta)\\ &\equiv \theta^{y}(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1}\\ &\propto \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}\end{aligned}$$

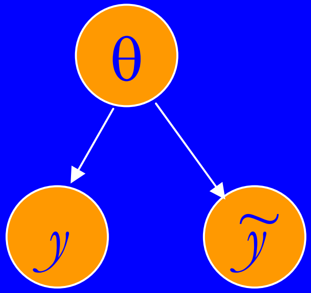
$$\to Beta(y+\alpha,n-y+\beta)$$

$$mean = E(\theta\,|\,y)=\left(\frac{y+\alpha}{n+\alpha+\beta}\right)$$

Prediction

“Posterior Predictive Density” of a future observation

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta$$



binomial example, *n*=20, *x*=12, *a*=1, *b*=1

$$p(\tilde{y} = 1|y) = \int \theta \frac{\Gamma(22)}{\Gamma(13)\Gamma(9)} \theta^{12}(1-\theta)^{8} d\theta = E[\theta|y] = \frac{13}{22}$$

What is the posterior mean for our example, with a prior:

Beta(1,1) *Beta(2,1)* *Beta(100,100)?*

Example: Building the hierarchy

θ is the probability of tumour in a population of female lab rats that receive a zero dose of a drug (control).

- 70 Previous experiments:

  0/20 0/20 0/20 0/19 1/18 16/52 etc
- Current experiment:

  4/14 (4 out of 14 developed the tumour)

Reference: Gelman et al, Bayesian Data Analysis

Rat model:

- Data: Assume a binomial model for the number of tumours, given θ.

Data from experiments j=1,..,J, J=71

$y_j \sim \mathrm{Bin}(n_j,\theta_j)$

- Priors: For convenience, choose priors θ~Beta(α,β) . We don’t know α, β so we’ll put a prior on these as well: p(α,β)
- Joint posterior:

$$p(\theta,\alpha,\beta \mid y) \propto p(\alpha,\beta)\,p(\theta \mid \alpha,\beta)\,p(y \mid \theta,\alpha,\beta)$$

$$\propto p(\alpha,\beta)\prod_{j=1}^{J}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta_j^{\alpha-1}(1-\theta_j)^{\beta-1}\prod_{j=1}^{J}\theta_j^{y_i}(1-\theta_j)^{n_j-y_j}$$

One study: Univariate Normal Model

Assume $y = N(\theta, \sigma^2)$; σ^2 *is* ***known variance***.

- Likelihood: $p(y|\theta) = (\sqrt{2}\pi\sigma)^{-1}\, e^{-.5(y-\theta)^2/\sigma^2}$
- Prior: $\theta \sim N(\mu_0, \tau_0^2)$; μ_0, τ_0^2 specified
- Posterior:

$$p(\theta|y) \propto \exp(-.5\, [(y-\theta)^2/\sigma^2 + (\theta-\mu_0)^2/\tau_0^2]$$

$$[\theta|y] \sim \mathrm{N}(\mu_1, \tau_1^2)$$

$$\mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2} y}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}}$$

$$\frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}$$

In terms of *precisions*

- Likelihood:

  $$y \sim Normal(\theta,\ \nu);\ \nu^2\ known\ \underline{precision},\ \nu^2 = 1/\sigma^2$$

- Prior:

  $$\theta \sim Normal(\mu_0,\ \omega_0);\ \mu_0,\ \omega_0\ specified\ values$$

- Posterior: Normal with mean

  $$\mathrm{E}(\theta\,|y) = (\ \mu_0\,\nu^2 + \mathrm{y}\ \omega_0\ )\ /\ (\nu^2 + \omega_0)$$

  $$\mathrm{Var}(\theta\,|y) = \nu^2 + \omega_0$$

- **Suppose that y = 2 and ν^2 = 1.**
  What happens to the posterior mean and variance as the prior changes?

Normal model, n iid observations

$$p(\theta \mid y) \qquad \propto p(\theta)p(y \mid \theta) = p(\theta)\prod_{i=1}^{n} p(y \mid \theta)$$

$$\propto \exp\left(-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right)\prod_{i=1}^{n}\exp\left(-\frac{1}{2\sigma^2}(y_i-\theta\ )^2\right)$$

$$\propto \exp\left(-\frac{1}{2}\left[\frac{1}{\tau_0^2}(\theta-\mu_0)^2+\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right]\right)$$

Simplify : posterior depends on *y* only through the sample mean

$$\theta \mid \bar{y} \sim N(\mu_n, \tau_n)$$

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}} \qquad \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2}+\frac{n}{\sigma^2}$$

What happens as the sample size *n* changes?

Normal model, unknown variance

- Likelihood:
  $y_1,\ldots,y_n \sim Normal(\theta, \sigma^2)$; θ known, σ^2 unknown variance
- Conjugate Prior:
  $\sigma^2 \sim Inverse\ Gamma\ IG(v_0,\sigma_0^2)$
  (v_0 = d.f.; σ_0^2 = scale; equivalent to v_0 observations with average squared deviation σ_0^2)

$$p(\sigma^2) \propto (\sigma^2)^{-(\nu_0+1)} e^{-\sigma_0^2/\sigma^2}$$

- Posterior: $\sigma^2|y \sim$ Inverse-Chi-squared

$$\sigma^2 \mid y \sim Inv-\chi^2\left(\nu_0+n, \frac{\nu_0\sigma_0^2+n\,\nu}{\nu_0+n}\right), \nu = \frac{1}{n}\sum_{i=1}^{n}(y_i-\theta)^2$$

In terms of precisions:

- Likelihood:

  $y \sim Normal(\theta,\ \tau^2);$ θ known, τ unknown precision

- Conjugate Prior:

  $\tau^2 \sim Gamma\ distribution\ Ga(v_0, \tau_0)$

- Posterior: $\tau^2 | y \sim Gamma$

Normal, mean and var unknown

- Prior: $\mu|\sigma^2 \sim \mathrm{N}(\mu_0, \sigma^2/\kappa_0)$
  $\sigma^2 \sim \text{Inv-}\chi^2(\nu_0,\sigma_0{}^2)$

$$p(\mu,\sigma^2) \propto \sigma^{-1}(\sigma^2)^{-(\nu_0/2+1)} \exp\left(-\frac{1}{2\sigma^2}\left[\nu_0\sigma_0^2 + \kappa_0(\mu_0-\mu)^2\right]\right)$$

- Posterior: $p(\mu,\sigma^2|y) = \text{N-Inv-}\chi^2(\mu_n,\sigma_n{}^2/\kappa_n;\ \nu_n,\sigma_n{}^2)$

$$\mu_n = \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\bar{y}$$

$$\kappa_n = \kappa_0 + n; \quad \nu_n = \nu_0 + n$$

$$\nu_n\sigma_n^2 = \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0+n}(\bar{y}-\mu_0)^2$$

Sample from the joint distribution:

1. Sample from the marginal posterior distribution for σ^2

$$\sigma^2 \mid y \sim Inv-\chi^2(v_n, \sigma_n^2)$$

2. Sample from the conditional posterior distribution for μ, given σ^2

$$\mu \mid \sigma^2, y \sim N(\mu_n, \sigma^2 / \kappa_n)$$

(Note: the Gamma distribution)

- For interest, if θ~Gamma(α,β), with parameters α,β>0, then

$$p(\theta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}, \qquad \theta > 0$$

$$E(\theta) = \alpha / \beta$$

$$Var(\theta) = \alpha / \beta^2$$

$$Mode(\theta) = (\alpha - 1) / \beta \qquad for \quad \alpha \geq 1$$

Complex models

- Repeated measures (irregular spacing)
- individual heterogeneity (frailty models)
- covariates at individual and group level
- errors in measuring responses
- errors in measuring covariates
- multiple instruments
- informative censoring
- binary, ordinal, response measures
- missing data
- spatial structure (disease mapping)
- familial aggregation

Historical note

In 17th and early 18th century, focus was on the ‘pre-data’ question: given θ, what are the various possible outcomes of the random variable y? Bayes and Laplace received independent credit as the first to invert the probability statement and obtain probability statements about θ, given observed y.

In his famous paper in 1763 (unpublished in his lifetime) Bayes sought $\Pr(\theta \in \theta_1,\theta_2)|y)$; his solution was based on a physical analogy of a probability space to a billiard table.

1. (Prior) A ball W is randomly thrown (uniform). Its position on the table is θ.
2. (Likelihood) A ball is randomly thrown *n* times. The value of *y* is the number of times it lands to the right of W.

Bayes then obtained $\Pr(\theta \in \theta_1,\theta_2)|y) = \int \mathrm{Bin}(..)/p(y)$ and $p(y)=1/(n+1)$, showing all possible values of *y* are equally likely *a priori*.

The numerator is an incomplete beta integral with no closed-form expression for large values of y. This presented difficulties for Bayes.

Laplace, independently ‘discovered’ Bayes’ theorem, and developed new analytic tools for computing integrals. He expanded the function around its maximum and evaluated the integral using a normal approximation. Based on 241945 girls and 251527 boys born in Paris from 1745 to 1770, he was ‘morally certain’ that $\theta<0.5$ for the probability that a birth is female. (He obtained $\Pr(\theta>0.5|y)=1.15\text{x}10^{-42}$)

Source: BDA

Thomas Bayes

Imagine

Imagine you're a Bayesian
It's easy if you try,
You just adopt a prior,
And the data updates π.
Statistics is so simple
With subjective probabilityyyyy -- ah-ah! ah ah...

Now imagine you're a frequentist,
Worrying about what might have been,
Spending your whole lifetime
Analyzing data you've never seen.
And if you want an interval,
You'll need a pivotal quantityyyyy -- ah-ah! ah ah...

You may say I sound like Nozer --
But I'm not the only one:
Every four years we all get together,
To talk, drink beer, and lie in the sun.