

Bayesian Data Analysis

Ken Newman & Vanda Inácio

University of Edinburgh



Semester 2, 2017/2018

General information

- ↪ **Lecturers:** Vanda Inácio & Ken Newman.
- ↪ **Email:** vanda.inacio@ed.ac.uk and ken.newman@biostats.ac.uk.
- ↪ **Office:** 4601 (Vanda) and 3617 (Ken), JCMB.
- ↪ **Lectures and location:** Friday, 13:00-15:00, Swann 7.15, odd weeks.
- ↪ **Computer labs:** Friday, 13:00-15:00, KB Centre, level 3, even weeks.

Assessment

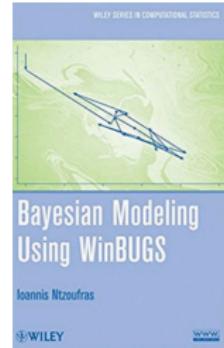
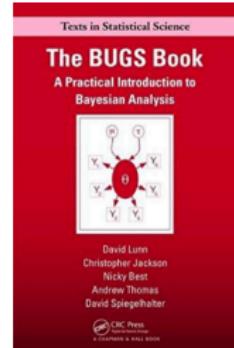
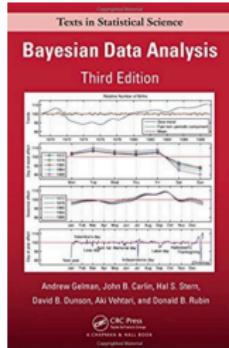
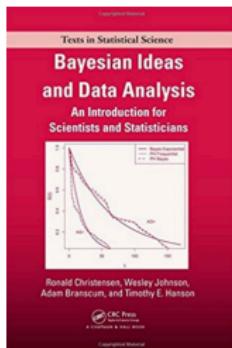
- ↪ 100% coursework: two homework assignments.
- ↪ Homework assignment 1 (50%) will be delivered in week 5. Solutions should be handed in week 7.
- ↪ Homework assignment 2 (50%) will be delivered in week 10. Solutions should be handed in week 12.
- ↪ Both homework assignments are to be done individually and consist of a sample of applied problems requiring the use of statistical software. A report containing all the analyses and conclusions should be delivered.

Scope

- ↪ The goal of this course is to provide practical experience of applying Bayesian analyses to a range of statistical models.
- ↪ The statistical analyses will be conducted using the widely used computer package WinBUGS.
- ↪ Topics:
 - ↪ Brief overview of main Bayesian ideas.
 - ↪ Introduction to WinBUGS.
 - ↪ Linear and generalised linear models (fixed effects).
 - ↪ Hierarchical Bayesian models: linear and generalised linear models with random effects.
 - ↪ Further topics (which might include: models with missing data, measurement error, mixture models).

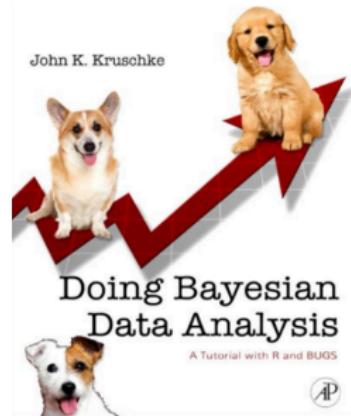
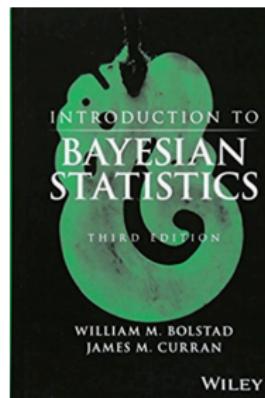
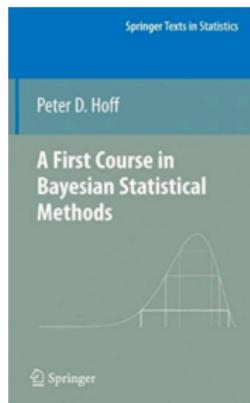
Recommended textbooks

- The course material (slides, worksheets, and other support material) contain all the information needed for the course.
- However, there is a wide variety of books, at a wide range of levels, that cover all or part of the material within this course that may be of interest to supplement the provided material and/or provide additional examples.
- Personally, we like



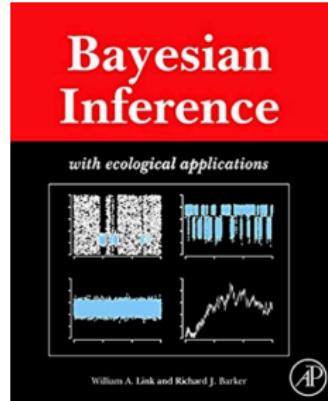
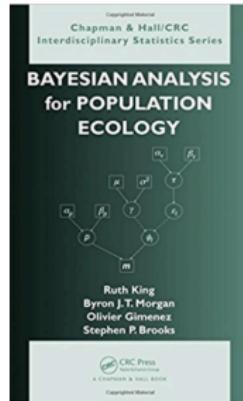
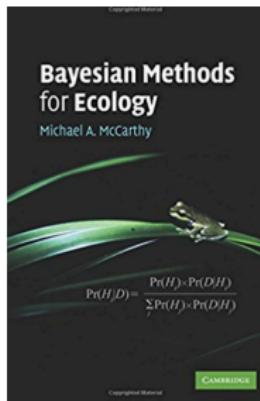
Additional textbooks

Popular introductory texts



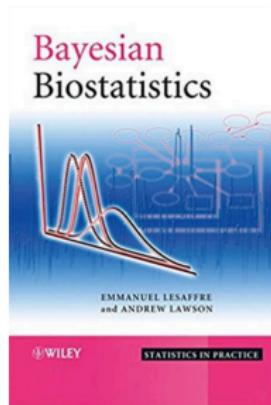
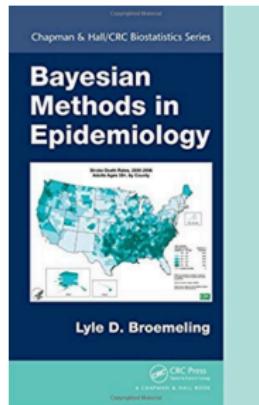
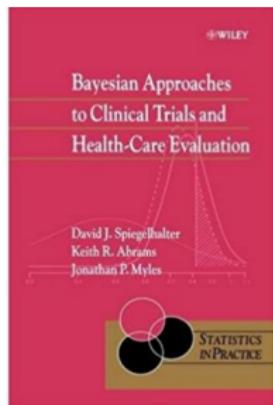
Additional textbooks

Ecological applications (covers from [amazon.co.uk](https://www.amazon.co.uk))



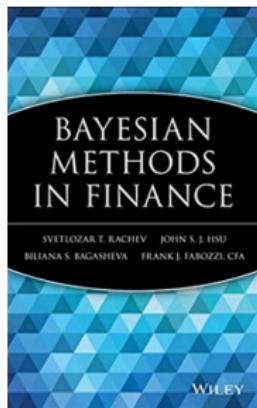
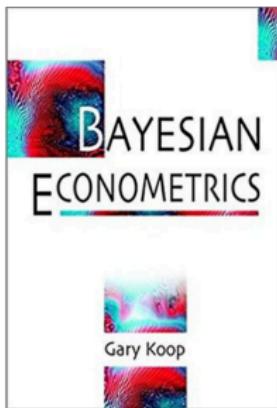
Additional textbooks

Medical/epidemiological applications (covers from amazon.co.uk)



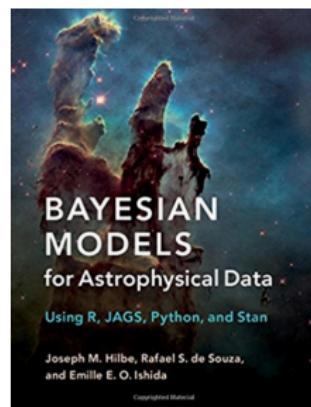
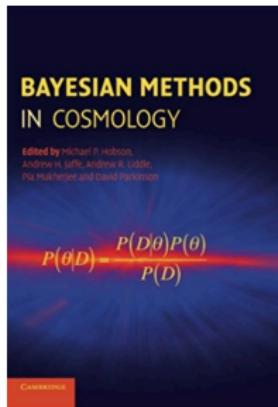
Additional textbooks

Finance/business/econometric applications (covers from [amazon.co.uk](https://www.amazon.co.uk))



Additional textbooks

Astrophysics/cosmology applications (covers from [amazon.co.uk](https://www.amazon.co.uk))



Review of Bayesian inference

- ↪ Bayesian methods have been widely applied in many areas of science (e.g., medicine, finance, ecology, physics, psychology, etc).
- ↪ Motivations for adopting Bayesian approach vary:
 - ↪ natural and coherent way of thinking about science and learning,
 - ↪ pragmatic choice that is suitable for the problem in hand.
- ↪ Spiegelhalter et al. (2004) define a Bayesian approach as
 - 'the explicit use of external information in the design, monitoring, analysis, interpretation, and reporting of a [scientific investigation].'
- ↪ These authors argue that a Bayesian approach is
 - ↪ more flexible in adapting to each unique situation,
 - ↪ more efficient in using all available evidence,
 - ↪ more useful in providing relevant quantitative summaries,than traditional methods.

Review of Bayesian inference

Bayesian vs frequentist statistics

- ↪ The **frequentist** approach can be regarded as a procedure that quantifies uncertainty (p-value, confidence interval, etc) in terms of repeating the process that generated the data many times.
- ↪ Parameters are fixed and unknown, on the data is random.
- ↪ Aims to be objective.
- ↪ Both approaches have pros and cons. When both are applicable they are unlikely to give different answers.
- ↪ **Bayesian** represent their uncertainty about parameters with probability distributions and treat them as random variables.
- ↪ We can thus, under a Bayesian framework, make probability statements about model parameters.
- ↪ This is in contrast with the frequentist framework where probability statements only concern the data.

Review of Bayesian inference

Main components

- ↪ Suppose we have a parameter $\theta \in \Theta$ on which we wish to make inference.
- ↪ The main ‘ingredients’ of Bayesian inference are:
 - ↪ The prior distribution, $p(\theta)$, which represents the initial beliefs concerning the parameter prior to any data being observed.
 - ↪ The likelihood $f(\mathbf{y} | \theta)$, which plays a key role in statistical inference, Bayesian or not. Represents the information contained in the data \mathbf{y} about the parameter θ .
 - ↪ The posterior distribution, $p(\theta | \mathbf{y})$, which updates the prior beliefs with respect to θ , following the data \mathbf{y} being observed.
- ↪ Bayesian learning combines past experience (prior) with new data (likelihood) in a mathematically coherent way (Bayes’ Theorem) to form the current state of knowledge (posterior).

Review of Bayesian inference

Bayes' theorem

- ↪ Remember that Bayes' theorem (continuous version) tells us

$$p(\theta | \mathbf{y}) = \frac{f(\mathbf{y} | \theta)p(\theta)}{f(\mathbf{y})},$$

where $f(\mathbf{y}) = \int_{\Theta} f(\mathbf{y} | \theta)p(\theta)d\theta$ is the marginal distribution data of the data.

- ↪ Since the marginal distribution does not depend on θ (we are integrating it out), we can write

$$p(\theta | \mathbf{y}) \propto f(\mathbf{y} | \theta)p(\theta),$$

i.e., the posterior distribution is proportional to the likelihood times the prior distribution.

Review of Bayesian inference

Summarising posterior distributions

- ↪ All Inference about the parameter of interest θ is based on the posterior distribution.
- ↪ The information contained in the posterior distribution can be summarised in different ways as appropriate to the inference goal, e.g.
 - ↪ Means, standard deviations, medians.
 - ↪ Probability of exceeding a certain threshold, say θ_0 , $\Pr(\theta > \theta_0 | \mathbf{y})$.
 - ↪ Credibility intervals.

Review of Bayesian inference

Decision theory

- ↪ What is the ‘best’ one-number summary of the posterior to be used as an estimator?
- ↪ This depends on the situation, and in particular, on the penalty associated with different types of errors (e.g., maybe overestimation is way worse than underestimation).
- ↪ Decision theory is used to form estimators with good properties.
- ↪ Let $L(\theta, a)$ be the loss associated with using a as the estimate, when the true value is θ . Note that for simplicity we only write a but, in fact, it is $a(\mathbf{y})$.
- ↪ The corresponding Bayes estimator is then chosen to minimise the expectation of the loss function with respect to the posterior distribution, i.e. the posterior expected loss.
- ↪ Mathematically, the Bayes estimate, $\hat{\theta}$, is defined such that

$$\begin{aligned}\hat{\theta} &= \min_{a \in \Theta} \mathbb{E}_{\text{post.}} [L(\theta, a)] \\ &= \min_{a \in \Theta} \left[\int_{\theta \in \Theta} L(\theta, a) p(\theta | \mathbf{y}) d\theta \right].\end{aligned}$$

Review of Bayesian inference

Decision theory

- Three commonly used loss functions and corresponding Bayes estimates are:

Loss	Bayes estimate
$L(\theta, a) = (\theta - a)^2$ (quadratic loss)	$\widehat{\theta} = \mathbf{E}_{\text{post.}}(\theta) = \int_{\theta \in \Theta} \theta p(\theta \mathbf{y}) d\theta$
$L(\theta, a) = \theta - a $ (absolute error loss)	$\widehat{\theta} = \text{median}_{\text{post}}(\theta)$
$L(\theta, a) = I(\theta \neq a)$ (zero/one loss)	$\widehat{\theta} = \arg \max_{\theta} p(\theta \mathbf{y})$

- Appropriate loss functions for hypothesis testing and interval estimation do also exist.

Choice of prior distribution

- ↪ Picking the prior is obviously important and uniquely Bayesian.
- ↪ Some types of priors include (these categories are not mutually exclusive):
 - ↪ Informative/expert priors
 - ↪ Non-informative/vague priors
 - ↪ Conjugate priors
- ↪ It is also important to try several priors in a sensitivity analysis.

Choice of prior distribution

Informative/expert priors

- ↪ A major advantage of the Bayesian approach is the ability to include expert prior information.
- ↪ This can come from either an expert in the field or from past data (literature, pilot study, etc).
- ↪ We call elicitation to the process of extracting prior knowledge in a suitable manner to permit the formulation of a suitable prior distribution that represents the expert/historical information as accurately as possible.
- ↪ Spiegelhalter et al. (2004), sections 5.2, 5.3, and 5.4, contain a good discussion on how priors might be elicited from experts or historical data.

Choice of prior distribution

Informative/expert priors

- ↪ What about if we ask more than one expert and they don't agree?
- ↪ Suppose we are interested in a quantity θ and that expert j recommends prior $\theta \sim N(\mu_j, \sigma_j)$.
- ↪ We could weight the experts using a mixture model

$$p(\theta) = \sum_{j=1}^J \omega_j N(\theta | \mu_j, \sigma_j^2),$$

where ω_j is the weight given to expert j ($\sum_{j=1}^J \omega_j = 1$).

Choice of prior distribution

Non-informative/vague priors

- ↪ What should we do if we do not have any prior information concerning the parameter of interest?
- ↪ Bayes himself suggested that when this is the case, the Uniform prior should be used, so that $p(\theta) = c$, for all θ .
- ↪ In this case we clearly have

$$p(\theta | \mathbf{y}) \propto f(\mathbf{y} | \theta),$$

i.e., the posterior distribution has the same shape as the likelihood function.

- ↪ In this case the mode of the posterior distribution coincides with the MLE.
- ↪ If the parameter space Θ is unbounded, this prior will be improper (regardless the value of c), that is, $\int_{\Theta} p(\theta) d\theta = \infty$. Improper priors can be used but we must check if the resulting posterior is proper (i.e., if it integrates to one).

Choice of prior distribution

Non-informative/vague priors

- ↪ As argued by Raiffa and Schlaiffer (1961), if one is ignorant about θ , one should also be ignorant about θ^2 , and one cannot find a distribution that is uniform on both θ and θ^2 .
- ↪ To see this, suppose that we place a Uniform prior on $\theta \in [0, 1]$, so that $p(\theta) = 1$. The corresponding prior on $\psi = \theta^2$ is $p(\psi) = \frac{1}{2\sqrt{\psi}}$, which is obviously non-uniform on ψ .
- ↪ In practice, priors should be specified on the parameter that the statistician is interested in within the analysis.

Choice of prior distribution

Non-informative/vague priors: Jeffreys' prior

- ↪ Jeffreys (1961) proposed a class of priors that are invariant to transformations.
- ↪ The Jeffrey's prior is $p(\theta) \propto \sqrt{I(\theta)}$, where $I(\theta)$ is the expected Fisher information.
- ↪ Remember that

$$I(\theta) = \mathbf{E} \left(\frac{d}{d\theta} \log f(\mathbf{Y} | \theta) \right)^2,$$

which in regular cases equals to $I(\theta) = -\mathbf{E} \left(\frac{d^2}{d\theta^2} \log f(\mathbf{Y} | \theta) \right)$.

- ↪ For the case of a multivariate parameter vector, say $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^T$, Jeffreys' prior is given by

$$p(\boldsymbol{\theta}) \propto \sqrt{\det I(\boldsymbol{\theta})},$$

with $I(\boldsymbol{\theta}) = -\mathbb{E} \left(\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \log f(\mathbf{Y} | \boldsymbol{\theta}) \right)$.

Choice of prior distribution

Non-informative/vague priors: Jeffreys' prior

- ↪ Let us exemplify the calculation of Jeffreys' prior on binomial data.
- ↪ To do that, let us suppose that $Y \sim \text{Bin}(n, \theta)$, where Y denotes the number of 'successes' out of n trials, and where the probability of success is θ .
- ↪ The likelihood is

$$f(y | \theta) = \binom{n}{y} \theta^y (1 - \theta)^{n-y},$$

and the log likelihood is

$$\log f(y | \theta) = C + y \log \theta + (n - y) \log(1 - \theta).$$

- ↪ The first derivative is

$$\frac{d}{d\theta} \log f(y | \theta) = \frac{y}{\theta} - \frac{n - y}{1 - \theta},$$

and the second derivative is

$$\frac{d^2}{d\theta^2} \log f(y | \theta) = -\frac{y}{\theta^2} - \frac{n - y}{(1 - \theta)^2}.$$

Choice of prior distribution

Non-informative/vague priors: Jeffreys' prior

↪ Thus,

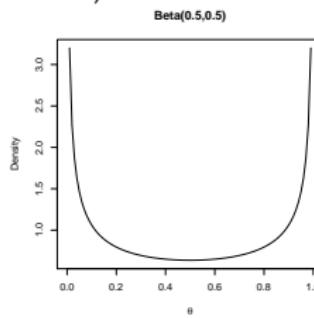
$$I(\theta) = -\mathbb{E} \left(\frac{d^2}{d\theta^2} \log f(Y | \theta) \right) = \frac{1}{\theta^2} \mathbb{E}(Y) + \frac{1}{(1-\theta)^2} (n - \mathbb{E}(Y)).$$

↪ Remember that $Y \sim \text{Bin}(n, \theta)$, implies $E(Y) = n\theta$.

↪ Therefore,

$$I(\theta) = \frac{n}{\theta(1-\theta)}.$$

↪ We can then conclude that Jeffreys's prior is $p(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$, which is a Beta($1/2, 1/2$) distributions and gives greater plausibility to values near 0 and 1 than to values in between (see figure below).



Choice of prior distribution

Non-informative/vague priors: Jeffreys' prior

→ Let us now compute Jeffreys' prior for a normal mean θ (assuming that the variance σ^2 is known).

→ With $Y_1, \dots, Y_n \stackrel{\text{iid}}{\sim} N(\theta, \sigma^2)$, the likelihood is

$$\begin{aligned} f(\mathbf{y} | \theta) &= \prod_{i=1}^n \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{\sigma^2} (y_i - \theta)^2 \right\} \right] \\ &= (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \theta)^2 \right\}, \end{aligned}$$

and the corresponding log likelihood is

$$\log f(\mathbf{y} | \theta) = C - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \theta)^2.$$

Choice of prior distribution

Non-informative/vague priors: Jeffreys' prior

- ↪ The first and second derivatives are, respectively

$$\frac{d}{d\theta} \log f(\mathbf{y} | \theta) = \frac{n}{\sigma^2} (\bar{y} - \theta),$$

and

$$\frac{d^2}{d\theta^2} \log f(\mathbf{y} | \theta) = -\frac{n}{\sigma^2}.$$

- ↪ The expected Fisher information is then $I(\theta) = \frac{n}{\sigma^2}$, which does not depend on θ , thus implying that $p(\theta) \propto 1, \forall \theta$ (\Rightarrow improper prior).

Choice of prior distribution

Non-informative/vague priors: Jeffreys' prior

- ↪ We will now consider the Poisson case. Let $Y_1, \dots, Y_n \stackrel{\text{iid}}{\sim} \text{Poisson}(\theta)$, $\theta > 0$.
- ↪ The likelihood and log likelihood are, respectively, given by

$$\begin{aligned} f(\mathbf{y} \mid \theta) &= \prod_{i=1}^n \left\{ \frac{e^{-\theta} \theta^{y_i}}{y_i!} \right\} \\ &= \frac{e^{-n\theta} \theta^{\sum_{i=1}^n y_i}}{\prod_{i=1}^n y_i!}, \end{aligned}$$

and

$$\log f(\mathbf{y} \mid \theta) = -n\theta + \sum_{i=1}^n y_i \log \theta + C.$$

- ↪ The first derivative is

$$\frac{d}{d\theta} \log f(\mathbf{y} \mid \theta) = -n + \frac{1}{\theta} \sum_{i=1}^n y_i,$$

while the second derivative is

$$\frac{d^2}{d\theta^2} \log f(\mathbf{y} \mid \theta) = -\frac{\sum_{i=1}^n y_i}{\theta^2}.$$

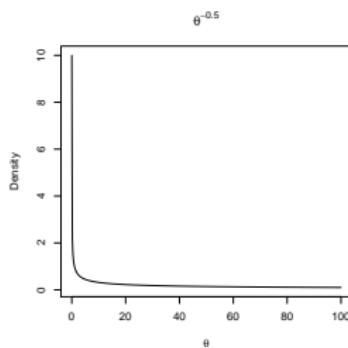
Choice of prior distribution

Non-informative/vague priors: Jeffreys' prior

↪ Thus, the expected Fisher information is

$$I(\theta) = \frac{1}{\theta^2} n \mathbb{E}[Y] = \frac{n}{\theta},$$

implying that $p(\theta) \propto \theta^{-1/2}$.



↪ Jeffreys' prior is improper in this case, but can be approximated by a Gamma distribution with parameters $\alpha = 1/2$ and $\beta \rightarrow 0$.

Choice of prior distribution

Non-informative/vague priors: Jeffreys' prior

- ↪ Jeffreys' prior is objective in that there is no prior tuning. It is a means of constructing a prior in the absence of prior information.
- ↪ Although Jeffreys' prior has the desirable property of being invariant to reparameterisations it can lead to improper priors.
- ↪ Alternative vague or non-informative prior distributions often have a reasonable mean for the distribution, but with a large variance parameter.
- ↪ Several different priors may be considered, each of which may be described to be vague or non-informative, and the sensitivity of the posterior on these priors investigated.

Choice of prior distribution

Conjugate priors

- ↪ A conjugate prior leads to a posterior from the same parametric family as the prior.
- ↪ There are long lists of conjugacies that we should be aware of

https://en.wikipedia.org/wiki/Conjugate_prior

- ↪ Conjugate priors are used often for computational convenience because the posterior has a closed form.
- ↪ In fancier models, conjugate priors facilitate Gibbs sampling which is the easiest Bayesian computational algorithm.

Choice of prior distribution

Conjugate priors: beta-binomial model

↪ Let us suppose $Y \sim \text{Bin}(n, \theta)$, whose likelihood is

$$f(y | \theta) = \binom{n}{y} \theta^y (1 - \theta)^{n-y}.$$

↪ It is mathematically convenient to use a $\text{Beta}(a, b)$ prior distribution for θ because it has a similar form to the binomial likelihood. Its density function is

$$p(\theta) = \frac{1}{B(a, b)} \theta^{a-1} (1 - \theta)^{b-1} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1 - \theta)^{b-1}.$$

↪ If $\theta \sim \text{Beta}(a, b)$, then

$$\mathbb{E} = \frac{a}{a+b},$$

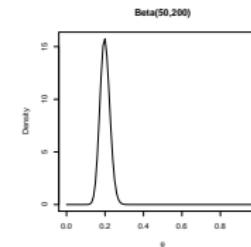
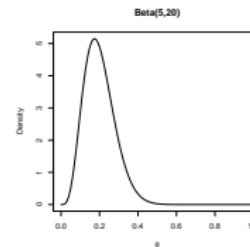
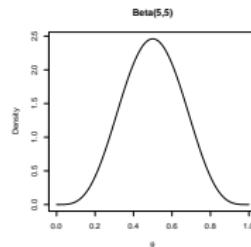
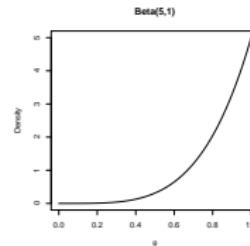
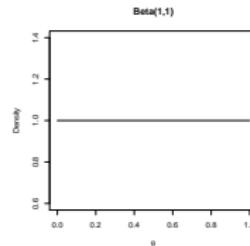
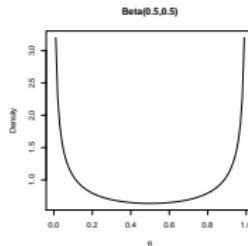
and variance

$$\text{var}(\theta) = \frac{ab}{(a+b)^2(a+b+1)}.$$

Choice of prior distribution

Conjugate priors: beta-binomial model

↪ The beta distribution can take several different shapes.



Choice of prior distribution

Conjugate priors: beta-binomial model

- Combining the beta prior distribution for θ with the binomial likelihood results in the following posterior distribution

$$\begin{aligned} p(\theta | y) &\propto f(y | \theta)p(\theta) \\ &= \binom{n}{y} \theta^y (1-\theta)^{n-y} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1} \\ &\propto \theta^{a+y-1} (1-\theta)^{b+n-y-1}. \end{aligned}$$

- That is, the posterior distribution turns out to be another Beta distribution

$$\theta | y \sim \text{Beta}(a + y, b + n - y)$$

Choice of prior distribution

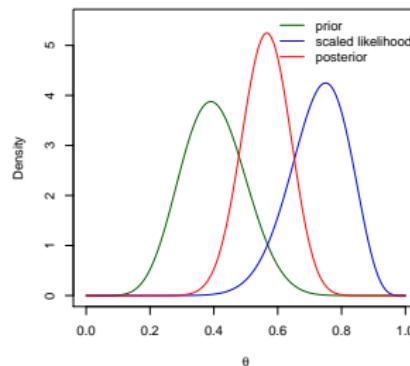
Conjugate priors: beta-binomial model

- ↪ Some comments are in order.
- ↪ The prior mean is $E(\theta) = \frac{a}{a+b}$.
- ↪ The data mean (mle) is y/n .
- ↪ The posterior mean $\mathbb{E}[\theta | y] = \frac{a+y}{a+b+n}$.
- ↪ Can interpret prior information as being equivalent to having observed $a - 1$ successes in $a + b - 2$ prior trials.
- ↪ With fixed a and b , as y and n increase, $\mathbb{E}[\theta | y] \rightarrow \frac{y}{n}$ (the mle), and the variance tends to zero.
- ↪ This is a general phenomenon: as n increases, posterior distribution gets more concentrated and the likelihood dominates the prior.

Choice of prior distribution

Conjugate priors: beta-binomial model

- ↪ Consider a drug to be given for relief of chronic pain.
- ↪ Experience with similar compounds has suggest that response rates, say θ , between 0.2 and 0.6 could be feasible.
- ↪ Interpret this as a distribution with mean 0.4 and standard deviation 0.1.
- ↪ A Beta(9.2, 13.8) distribution has these properties.
- ↪ Suppose we treat $n = 20$ volunteers with the compound and observe $y = 15$ positive responses.
- ↪ The parameters of the Beta distribution are updated to $9.2 + 15 = 24.2$ and $13.8 + 20 - 15 = 18.8$.



Choice of prior distribution

Conjugate priors: beta-binomial model

- ↪ Note that the likelihood, although a function of the parameter, it is not a density and so, in particular, does not integrate to one.
- ↪ In order to plot the likelihood along with the prior and posterior distributions, it is convenient that the three are in the same scale
- ↪ Therefore, we have rescaled the likelihood function so that it integrates to one.
- ↪ See R script `rescaled_likelihood`.

Choice of prior distribution

Conjugate priors: poisson-gamma model

↪ Suppose we have $Y_1, \dots, Y_n \stackrel{\text{iid}}{\sim} \text{Poisson}(\theta)$.

↪ We have already seen that the likelihood is

$$f(\mathbf{y} | \theta) = \prod_{i=1}^n \left\{ \frac{e^{-\theta} \theta^{y_i}}{y_i!} \right\}$$

↪ The kernel of the Poisson likelihood (as a function of θ) has the same form as that of a Gamma(a, b) prior for θ

$$p(\theta) = \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta}.$$

↪ This parameterisation of the Gamma distributions in terms of the shape parameter a and rate parameter b , has mean a/b and variance a/b^2 .

Choice of prior distribution

Conjugate priors: poisson-gamma model

↪ This implies the following posterior

$$\begin{aligned} p(\theta \mid \mathbf{y}) &\propto f(\theta \mid \mathbf{y})p(\theta) \\ &= \left\{ \prod_{i=1}^n \frac{e^{-\theta} \theta^{y_i}}{y_i!} \right\} \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta} \\ &\propto e^{-n\theta} \theta^{\sum_{i=1}^n y_i} \theta^{a-1} e^{-b\theta} \\ &= \theta^{a+\sum_{i=1}^n y_i - 1} e^{-\theta(b+n)} \end{aligned}$$

↪ We recognise this as the kernel of a gamma distribution with parameters $a + n\bar{y}$ and $b + n$, that is

$$\theta \mid \mathbf{y} \sim \text{Gamma}(a + n\bar{y}, b + n).$$

Choice of prior distribution

Conjugate priors: poisson-gamma model

↪ Note that

$$\mathbb{E}(\theta | \mathbf{y}) = \frac{a + n\bar{y}}{b + n} = \bar{y} \left(\frac{n}{n + b} \right) + \frac{a}{b} \left(1 - \frac{n}{n + b} \right).$$

↪ The posterior mean is then a compromise between prior mean a/b and the mle \bar{y} .

Choice of prior distribution

Conjugate priors: normal-normal model

- ↪ Let us now consider $Y_1, \dots, Y_n \stackrel{\text{iid}}{\sim} N(\theta, \sigma^2)$, with σ^2 known. Further, let $\theta \sim N(\mu_0, \sigma_0^2)$.
- ↪ We've already seen the 'simplified' form of the likelihood

$$f(\mathbf{y} | \theta) = (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \theta)^2 \right\}.$$

- ↪ The posterior is

$$\begin{aligned} p(\theta | \mathbf{y}, \sigma^2) &\propto f(\mathbf{y} | \theta)p(\theta) \\ &\propto \exp \left\{ -\frac{1}{2\sigma^2} \sum_i (y_i - \theta)^2 \right\} \exp \left\{ -\frac{1}{2\sigma_0^2} (\theta - \mu_0)^2 \right\} \\ &\propto \exp \left\{ -\frac{1}{2\sigma^2 \sigma_0^2} \left[\theta^2(n\sigma_0^2 + \sigma^2) - 2\theta(n\bar{y}\sigma_0^2 + \mu_0\sigma^2) \right] \right\} \end{aligned}$$

Choice of prior distribution

Conjugate priors: normal-normal model

- ↪ This can be recognised as the kernel of a normal distribution with mean

$$\mu_n = \frac{n\bar{y}\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2} = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}},$$

and variance

$$\sigma_n^2 = \frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2} = \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}.$$

- ↪ The posterior parameters μ_n and σ_n^2 combine the prior parameters μ_0 and σ_0^2 with terms from the data.
- ↪ For instance, notice that

$$\mu_n = \frac{\tau_0}{\tau_0 + n\tau}\mu_0 + \frac{n\tau}{\tau_0 + n\tau}\bar{y},$$

where $\tau_0 = 1/\sigma_0^2$ (prior precision) and $\tau = 1/\sigma^2$ (sampling precision), and so the posterior mean is a weighted average of the prior mean and the sample mean.

Choice of prior distribution

Conjugate priors: normal-gamma model

- ↪ Suppose now— and this is unrealistic, but just to make a point— that θ is known, but the variance σ^2 is unknown.
- ↪ It is often convenient in Bayesian statistics to work with the precision, $\tau = 1/\sigma^2$.
- ↪ Let us take $\tau \sim \text{Gamma}(a, b)$.
- ↪ Then the posterior of τ is

$$\begin{aligned} p(\tau | \mathbf{y}, \theta) &\propto f(\mathbf{y} | \tau)p(\tau) \\ &\propto \tau^{n/2} \exp\left\{-\frac{\tau}{2} \sum_{i=1}^n (y_i - \theta)^2\right\} \tau^{a-1} e^{-b\tau} \\ &= \tau^{a+n/2-1} \exp\left\{-\tau \left(b + \frac{1}{2} \sum_i (y_i - \theta)^2\right)\right\}. \end{aligned}$$

- ↪ That is,

$$\tau | \mathbf{y}, \theta \sim \text{Gamma}\left(a + \frac{n}{2}, b + \frac{1}{2} \sum_i (y_i - \theta)^2\right).$$

Choice of prior distribution

Conjugate priors: normal-gamma model

→ A common choice is to take both a and b very small, then the posterior is approximately

$$\tau \mid \mathbf{y}, \theta \sim \text{Gamma} \left(\frac{n}{2}, \frac{1}{2} \sum_i (y_i - \theta)^2 \right),$$

and so

$$\mathbb{E}[\tau \mid \mathbf{y}, \theta] = \left(\frac{1}{n} \sum_i (y_i - \theta)^2 \right)^{-1},$$

so that the posterior expectation of the precision is (approximately) the sample precision (but with a divisor of n and not $n - 1$).

Choice of prior distribution

Conjugate priors: mixture priors

- ↪ Conjugate priors are convenient, but sometimes they might be not flexible enough.
- ↪ Mixtures of conjugate priors are a good alternative and they are actually quite flexible.
- ↪ And the good news is that mixtures of conjugate priors are also conjugate. A mixture prior is

$$p(\theta) = \sum_{j=1}^J \pi_j p_j(\theta),$$

where p_j are conjugate priors with different hyperparameters and the mixture weights $\pi_j \in [0, 1]$ sum to one.

- ↪ **Question:** If $Y \sim \text{Bin}(n, \theta)$ and $p(\theta) = \pi \text{Beta}(\theta | a_1, b_1) + (1 - \pi) \text{Beta}(\theta | a_2, b_2)$, then $\theta | y$ is ... (see first practical lab.).

Predictive inference

- ↪ We now consider prediction.
- ↪ The prior predictive distribution of \mathbf{y} is

$$f(\mathbf{y}) = \int_{\Theta} f(\mathbf{y} | \theta) p(\theta) d\theta,$$

also called marginal distribution of the data, usually in the frequentist approach.

- ↪ It is useful to check whether the model (likelihood+prior) gives (un)reasonable predictions.
- ↪ The posterior predictive distribution of a future observation z , given the data \mathbf{y} is

$$\begin{aligned} f(z | \mathbf{y}) &= \int_{\Theta} f(z | \theta, \mathbf{y}) p(\theta | \mathbf{y}) d\theta \\ &= \int_{\Theta} f(z | \theta) p(\theta | \mathbf{y}) d\theta, \end{aligned}$$

where the second equality is due to the conditional independence of Y and Z .

- ↪ Notice that the predictive distribution only depends on z and \mathbf{y} .

Predictive inference

- ↪ Let us consider a beta binomial experiment, where $Y \sim \text{Bin}(n, \theta)$ and $\theta \sim \text{Beta}(a, b)$.
- ↪ We have seen that the posterior distribution $\theta | y \sim \text{Beta}(a + y, b + n - y)$.
- ↪ For ease of notation, let $c = a + y$ and $d = b + n - y$.
- ↪ Now let us consider a further random quantity Z for which we judge $Z \sim \text{Bin}(m, \theta)$ and Z and Y are conditionally independent given θ .
- ↪ So, having conducted n trials, we consider conducting further m trials.
- ↪ We seek the predictive distribution of Z (the number of successes in m new trials) given the observed y .

Predictive inference

↪ We have

$$\begin{aligned}f(z \mid y) &= \int_{\Theta} f(z \mid \theta)p(\theta \mid y)d\theta \\&= \int_0^1 \binom{m}{z} \theta^z (1-\theta)^{m-z} \frac{1}{B(c, d)} \theta^{c-1} (1-\theta)^{d-1} d\theta \\&= \binom{m}{z} \frac{1}{B(c, d)} \int_0^1 \theta^{c+z-1} (1-\theta)^{d+m-z-1} d\theta \\&= \binom{m}{z} \frac{B(c+z, d+m-z)}{B(c, d)} \\&= \binom{m}{z} \frac{\Gamma(c+d)}{\Gamma(c)\Gamma(d)} \frac{\Gamma(c+z)\Gamma(d+m-z)}{\Gamma(c+d+m)}.\end{aligned}$$

↪ We say that $Z \mid y$ is the Binomial-Beta distribution with parameters c , d , and m .