Bayesian Statistics - Level 5M

Craig Alexander & Wei Zhang

School of Mathematics and Statistics, University of Glasgow

Summer, 2021

(*Lecture notes prepared by Dr Vlad Vyshmirsky*)

Lecture 1

1.1 Aims and intended learning outcomes

Aims

- To introduce students to the main ideas of modern Bayesian statistics.
- To demonstrate how prior distributions are updated to posterior distributions in simple statistical models.
- To illustrate the formulation and analysis of hierarchical models in WinBugs.

Intended learning outcomes

By the end of this course students will be able to:

- describe the rules for updating prior distributions in the presence of data, and for calculating posterior predictive distributions;
- derive posterior distributions corresponding to simple low-dimensional statistical models, typically (Level M: but not exclusively) with conjugate priors;
- describe and compute various summaries of the posterior distribution, including posterior mean, MAP estimate, posterior standard deviation and credible regions (including HPDRs) and the predictive distribution;
- explain different approaches to the choice of prior distribution;
- explain the role of hyperparameters in Bayesian inference and introduce them appropriately into statistical model and use the empirical Bayes approach for their determination;
- explain the use of independent simulation techniques for posterior sampling (Level M: and apply them in simple contexts using R);
- formulate and analyze simple hierarchical models using Gibbs sampling in WinBugs;
- (Level M) describe and apply simple checks of mixing, and explain when mixing is likely to be poor;
- (Level M) explain the role of decision theory in Bayesian analysis, formulate the decision process mathematically, and prove simple results.

Syllabus

- Main ideas of Bayesian inference: prior, likelihood, posterior, predictive; specification of prior distribution, uninformative versus informative priors, proper versus improper priors, conjugate priors; summaries of the posterior distribution; empirical Bayes methods
- Bayesian inference with conjugate priors: Binomial, Normal, Poisson, Exponential, Multinomial models, others as time permits

- Use of direct simulation for Bayesian inference
- Hierarchical models: formulation and analysis using Gibbs sampling in WinBugs/R, with several real data examples in practical labs
- (Level M): basics of decision theory

Adopted textbooks

- Hoff, P. D. (2009). *A First Course in Bayesian Statistical Methods*, Springer. Download free (from a U. Glasgow computer) from the library site (chapter by chapter) as PDF files.
- Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004 or 2014). *Bayesian Data Analysis*, 2nd or 3rd edn, Chapman & Hall/CRC.

Some other relevant books

Classics:

Lindley, D. V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint*, 2 vols., Cambridge.

DeGroot, M. H. (1970). *Optimal Statistical Decisions*, McGraw-Hill.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Economics*, Wiley.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*, Addison-Wesley.

Introductory:

Lee, P. M. (1989). *Bayesian Statistics: An Introduction*, Oxford.

Berry, D. A. (1996). *Statistics: A Bayesian Perspective*, Duxbury.

Leonard, T. and Hsu, J. S. J. (1999). *Bayesian Methods*, Cambridge.

Sivia, D. S., Skilling, J. (2006). *Data Analysis: A Bayesian Tutorial*, 2nd edn, Oxford.

Bolstad, W. M. (2007). *Introduction to Bayesian Statistics*, 2nd edn, Wiley.

Theoretical emphasis:

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd edn, Springer.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*, Wiley.

Robert, C. P. (2001). *The Bayesian Choice: From Decision-theoretic Foundations to Computational Implementation*, Springer.

O'Hagan, A. (1994). *Kendall's Advanced Theory of Statistics, Vol. 2B, Bayesian Inference*, Arnold.

Applied emphasis:

Carlin, B. P. and Louis, T. A. (2009). *Bayes Methods for Data Analysis*, 3rd edn, Chapman & Hall/CRC.

Press, S. J. (2003). *Subjective and Objective Bayesian Statistics: Principles, Models and Applications*, Wiley.

Congdon, P. (2006). *Bayesian Statistical Modelling*, 2nd edn, Wiley.

Congdon, P. (2003). *Applied Bayesian Modelling*, Wiley.

Congdon, P. (2005). *Bayesian Models for Categorical Data*, Wiley.

Computational emphasis:

Gamerman, D., Lopes, H. F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, 2nd edn, Chapman & Hall/CRC.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (eds.) (1996). *Markov Chain Monte Carlo in Practice*, Chapman & Hall.

Marin, J.-M. and Robert, C. P. (2007). *Bayesian core: A Practical Approach to Computational Bayesian Statistics*, Springer.

Specialized:

Rossi, P. E., Allenby, G., McCulloch, R. (2005). *Bayesian Statistics and Marketing*, Wiley.

Geweke, J. (2005). *Contemporary Bayesian Econometrics and Statistics*, Wiley.

Lynch, S. M. (2007). *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*, Springer.

Gill, J. (2008). *Bayesian Methods: A Social and Behavioral Sciences Approach*, Chapman & Hall/CRC.

Moyé, L. A. (2008). *Elementary Bayesian Biostatistics*, Chapman & Hall.

1.2 Why Bayesian statistics?

- More flexibility in modelling
- More intuitive interpretation of inferences
  - Suppose $Y_i \sim$ i.i.d. $N(\mu, \sigma^2)$, μ unknown, σ known.
  - Know that
    $$(a, b) = (\overline{y} \pm 1.96\sigma/\sqrt{n})$$
    is a 95% C.I. for μ.
  - What does it mean?
  - A probability statement about $\overline{y}$, *not* μ:
    95% of the *random* intervals (a, b) contains the true unknown *fixed* μ.
  - However, users of statistics tend to interpret a 95% C.I. as meaning:
    with probability 0.95 μ lies in the *fixed* interval (a, b) computed using the data at hand.
  - This is indeed the correct interpretation of a Bayesian C.I. (although we call it something else!)
- Can easily account for additional available information, besides that in the data.
  - This may come from previous studies, or from an accepted theory in the field of study.
  - For instance, restrictions on the parameter space, such as $\mu > 0$ in the previous example.

1.3 Main features of Bayesian statistics

- Uncertainty (information) about unknown quantities (unobserved data *and* parameters) can be represented using probability
  - Compare with classical approach:

$$y \sim p(y|\theta), \ y \text{ observed data}, \ \theta \text{ parameter}$$

  - Bayes: build a *joint* probability model for data y and parameter θ

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

  - Need to specify a *prior* distribution $p(\theta)$.
- Observing data, changes the information about a parameter according to

$$p(\theta) \longrightarrow p(\theta|y)$$

$$\text{prior} \longrightarrow \text{posterior}$$

  - The transformation is performed using Bayes' theorem:

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

  - Simpler form of the updating rule using odds:

$$\frac{p(\theta_1|y)}{p(\theta_2|y)} = \frac{p(\theta_1)p(y|\theta_1)/p(y)}{p(\theta_2)p(y|\theta_2)/p(y)} = \frac{p(\theta_1)}{p(\theta_2)} \times \frac{p(y|\theta_1)}{p(y|\theta_2)}$$

    i.e.

$$\text{posterior odds} = \text{prior odds} \times \text{likelihood ratio}$$

- Prediction of new data $\tilde{y}$:

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

  Compare with the plug-in classical prediction $p(\tilde{y}|\hat{\theta})$, where $\hat{\theta}$ is an estimate of θ.

1.4 Overview of the course

Introduce main ideas of Bayesian statistics

- Prior, likelihood, posterior, predictive
- How to specify a prior?
  - informative vs uninformative

  - proper vs improper
  - conjugate priors
  - empirical Bayes
- How to summarize the posterior?
  - posterior mean, posterior mode (MAP), credible intervals
  - multi-dimensional parameters
- How to make predictions?

Models with conjugate priors

Binomial, Poisson, Exponential, Normal, Multinomial, others if time allows

Stochastic simulation

- ... has made Bayesian statistics practically feasible, besides models with conjugate priors
- Direct simulation from the posterior (seldom possible)
- Gibbs sampling

Hierarchical models

- Formulation
- Gibbs sampling estimation with WinBugs

A review exercise

A clinical test detects a certain disease with some probability. What is the probability that a person has the disease, if the test is positive?

To be more precise, let:

$$\mathrm{D} = \{\text{disease is present}\} \qquad \mathrm{D}^c = \{\text{disease is not present}\}$$

$$\mathrm{T} = \{\text{clinical test is positive}\} \qquad \mathrm{T}^c = \{\text{clinical test is negative}\}$$

Suppose it is known that

$$\Pr(\mathrm{T}|\mathrm{D}^c) = \Pr(\text{false positive}) = \alpha \qquad \Pr(\mathrm{T}^c|\mathrm{D}^c) = \Pr(\text{true negative}) = 1-\alpha$$
$$\Pr(\mathrm{T}|\mathrm{D}) = \Pr(\text{true positive}) = 1-\beta \qquad \Pr(\mathrm{T}^c|\mathrm{D}) = \Pr(\text{false negative}) = \beta$$

Terminological aside: A false positive is also called a *type I error*, while a false negative is a *type II error*. The probability $1-\alpha$ of a true negative is also called *specificity*, while the probability $1-\beta$ of a true positive is called *sensitivity*.

In order to answer the original question, one also needs to know the *prevalence* of the disease in the population, i.e., the proportion γ of the population affected by the disease:

$$\Pr(\mathrm{D}) = \gamma \qquad \Pr(\mathrm{D}^c) = 1-\gamma$$

Then, the probability that a randomly selected person from the population is affected, given that he/she tested positive, can be computed using Bayes Theorem:

$$\Pr(D|T) = \frac{\Pr(T|D)\Pr(D)}{\Pr(T|D)\Pr(D) + \Pr(T|D^c)\Pr(D^c)}.$$

Assume that $\alpha = 1/20$, $\beta = 1/10$ and $\gamma = 1/1000$. Using the formula above compute $\Pr(D|T)$. What is $\Pr(D|T^c)$? What are the odds of disease before taking the test? What are the odds after a positive test? After a negative test?

Lecture 2

2.1 A bit of history

- Rev. Thomas Bayes (1763) "An Essay Towards Solving a Problem in the Doctrine of Chances"
- Pierre Simeon Laplace (1774, 1812)
- General use of "inverse probability" method during the 19th century
- Criticisms in the 2nd half of the 19th century (e.g., G. Boole, J. Venn), especially of the "indifference principle" for prior specification
- ...
- Much more on this in Fienberg (2006) *When Did Bayesian Inference Become "Bayesian"?*, Bayesian Analysis, Vol.1, no.1, `http://ba.stat.cmu.edu/`
- "Classical" approach is more recent
  - R. A. Fisher in 1920's (likelihood, sufficiency, efficiency, significance testing)
  - J. Neyman and E. Pearson in the 1930's (hypothesis testing)

2.2 What is probability?

- An open philosophical issue, not likely resolved any time soon
- Two interpretations you have probably seen:
  - Symmetry of elementary outcomes $\rightarrow$ equiprobability;
    ratio of number of favourable to number of possible cases
  - Relative frequency in a long sequence of identical independent trials ("frequentist" interpretation)
  - Both only applicable in special situations
- Subjective interpretation in terms of betting odds (B. de Finetti)
  - $\Pr(A)$ is the amount *you* are willing to pay for a lottery ticket that pays £1 if event A occurs
  - Applies to non-reproducible events, e.g.,
    * Scotland winning *next year*'s Six Nations
    * Finding, in 2025, a *Poecile atricapilus,* the black-capped chickadee, living wild in Antarctica
  - It is *your* probability of A and it depends on the information available to *you*.
    B. de Finetti in the Preface to his *Theory of Probability*:

> "My thesis, paradoxically, and a little provocatively, but nonetheless genuinely, is simply this:
>
> PROBABILITY DOES NOT EXIST.
>
> The abandonment of superstitious beliefs about the existence of Phlogiston, the Cosmic Ether, Absolute Space and Time, ..., or Fairies and Witches, was an essential step along the road to scientific thinking. Probability, too, if regarded as something endowed with some kind of objective existence, is no less a misleading misconception, an illusory attempt to exteriorize or materialize our true probabilistic beliefs."

  - Coherency $\rightarrow$ axioms of probability theory

- Why?
  - Bayesian approach treats unknowns (parameters, unobserved data) as random variables.
  - Subjective probability is one way, adopted by many Bayesians, to justify doing so.
- Another (my favourite!) approach (H. Jeffreys, R. Cox, E. T. Jaynes): probability as logic of plausible reasoning.
  - The rules of probability extend ordinary ("Boolean") logic, where statements are known to be either true or false (1, 0), to inductive logic, where statements are true or false, but we don't know which.
  - Probability then is a scale used to describe how strongly, based on specific information, we believe a statement to be true
  - It is *objective* in the sense that anyone with the same knowledge should assign the same probabilities
  - Simple requirements for such a scale (like, if we know the probability that A is true, then we must know the probability that A is false) imply (via some cute functional equations) that probabilities must obey the usually probability rules (product rule, sum rule)
  - The following quote is reported at the beginning of both Jeffreys' and Jaynes' books:

    > "They say that Understanding ought to work by the rules of right reason. These rules are, or ought to be, contained in Logic; but the actual science of logic is conversant at present only with things either certain, impossible, or entirely doubtful, none of which (fortunately) we have to reason on. Therefore the true logic for this world is the calculus of Probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind." J. Clerk Maxwell

- More on interpretations of probability in FCBSM[1] §1.1, §1.2 and §2.1 and BDA2[2] or BDA3 [3] §1.5: read one or both!
- Criticism of Bayesian analysis:
  - subjective element in $p(\theta)$
- However,
  - can compare several analyses using different priors (sensitivity analysis)
  - choice of likelihood is also subjective and, often, more influential

2.3 Some probability theory results (BDA2/3 §1.8; FCBSM §2.2–2.6)

- marginal pdf (or pmf)
$$p(u) = \int p(u, v)\, dv$$
- conditional pdf
$$p(u|v) = \frac{p(u, v)}{p(v)}$$
- factorization
$$p(u, v) = p(u)p(v|u)$$
more generally, for a k-dim random vector $\mathbf{u}$
$$p(\mathbf{u}) = p(u_1)p(u_2|u_1)p(u_3|u_1, u_2) \times \cdots \times p(u_k|u_1, \ldots, u_{k-1})$$
- Some properties of $E(u)$ and $Var(u)$
$$\begin{aligned} E(u) &= \iint u p(u, v)\, du\, dv \\ &= \iint u p(u|v)\, du\, p(v)\, dv \\ &= \int E(u|v)p(v)\, dv \\ &= E\left[E(u|v)\right] \end{aligned}$$
One can also prove
$$Var(u) = E\left[Var(u|v)\right] + Var\left[E(u|v)\right]$$
These also hold for random vectors $\mathbf{u}, \mathbf{v}$.
- Transformations
  - u has pdf $p_u(u)$, $v = f(u)$, with f 1-1 and continuously differentiable. Then
$$p_v(v) = p_u(f^{-1}(v)) \left|\frac{du}{dv}\right|$$

[1] Hoff, P. D. (2009). *A First Course in Bayesian Statistical Methods*, Springer. Download free (from a U. Glasgow computer) from the library site (chapter by chapter) as PDF files.

[2] Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004). *Bayesian Data Analysis*, 2nd edn, Chapman & Hall/CRC.

[3] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2014). *Bayesian Data Analysis*, 3rd edn, Chapman & Hall/CRC.

– If $\mathbf{u}, \boldsymbol{\nu}$ are random vectors

$$p_{\boldsymbol{\nu}}(\boldsymbol{\nu}) = p_{\mathbf{u}}(f^{-1}(\boldsymbol{\nu}))\,|\det(J)|$$

where $J = ((\partial u_i/\partial \nu_j))$ is the Jacobian matrix of partial derivatives

2.4 Simulation (BDA2/3 §1.9; FCBSM §4.1–4.2)

- Multi-dimensional parameter problem $\theta = (\theta_1, \ldots, \theta_k)$ – posterior: $p(\theta|y)$
- Interested in the marginal distribution of θ_i:

$$p(\theta_i|y) = \int p(\theta|y)\, d\theta_1 \ldots d\theta_{i-1}\, d\theta_{i+1} \ldots d\theta_k$$

In general a complicated multi-dim. integral:

  - closed form not available
  - for moderate k, even numerical evaluation is problematic

- However, suppose we can sample from the posterior $p(\theta|y)$
  - let the sample be $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(L)}$
  - Then

$$\theta_i^{(1)}, \theta_i^{(2)}, \ldots, \theta_i^{(L)}$$

is a sample from the marginal posterior $p(\theta_i|y)$

  - Can estimate, e.g.
    * $E(\theta_i|y)$ using the sample mean $\bar{\theta}_i = \frac{1}{L}\sum_{j=1}^{L} \theta_i^{(j)}$
    * m s.t. $\Pr(\theta_i < m) = 0.5$ using the sample median of $\theta_i^{(j)}$'s
    * a central 95% posterior interval (a, b) for θ_i

$$\Pr(a < \theta_i < b) = 0.95$$

using $(\hat{a}, \hat{b})$ where

$$\begin{aligned} \hat{a} &= \text{0.025-quantile of the } \theta_i^{(j)}\text{'s} \\ \hat{b} &= \text{0.975-quantile of the } \theta_i^{(j)}\text{'s} \end{aligned}$$

For $L = 10000$, $\hat{a}$ is the 250-th value and $\hat{b}$ is the 9751-th value of the *sorted* $\theta_i^{(j)}$'s

- Suppose we are interested in some function $\psi = g(\theta)$
  - Quite general, includes:

$$\begin{aligned} &= \theta_i \\ &= I_A(\theta) \qquad \text{– Here } E[\psi|y] = \Pr(\theta \in A|y) \end{aligned}$$

- Need $p(\psi|y)$. Could
  - transform from $(\theta_1, \theta_2, \ldots, \theta_k)$ to $(\psi, \theta_2, \ldots, \theta_k)$, say

- use Jacobians to find $p(\psi, \theta_2, \ldots, \theta_k|y)$
- integrate wrt $\theta_2, \ldots, \theta_k$

Troublesome, if at all possible!

However,

- can produce a sample from the marginal posterior $p(\psi|y)$, simply by transforming the original sample $\{\theta^{(j)}, j = 1, \ldots, L\}$:

$$\psi^{(1)} = g(\theta^{(1)}), \quad \psi^{(2)} = g(\theta^{(2)}), \quad \ldots, \quad \psi^{(L)} = g(\theta^{(L)})$$

- Summaries of $p(\psi|y)$, such as mean, median, posterior intervals are then easily estimated from the sample of $\psi^{(j)}$'s

Homework 1

Problem 1 [BDA2/3, Exercise 1.12.1]

Conditional probability: suppose that if $\theta = 1$, then y has a normal distribution with mean 1 and standard deviation σ, and if $\theta = 2$, then y has a normal distribution with mean 2 and standard deviation σ. Also, suppose $\Pr(\theta = 1) = 0.5$ and $\Pr(\theta = 2) = 0.5$.

(a) For $\sigma = 2$, write the formula for the marginal probability density for y and sketch it.

(b) What is $\Pr(\theta = 1|y = 1)$, again supposing $\sigma = 2$?

(c) Describe how the posterior density of θ changes in shape as σ is increased and as it is decreased.

Problem 2 [BDA2/3, Exercise 1.12.6]

Conditional probability: approximately 1/125 of all births are fraternal twins and 1/300 of births are identical twins. Elvis Presley had a twin brother (who died at birth). What is the probability that Elvis was an identical twin? (You may approximate the probability of a boy or girl birth as $\frac{1}{2}$.)

Problem 3

Suppose that $\theta \sim \mathrm{Exp}(1)$ and that $\psi = g(\theta) = \dfrac{\sqrt{\theta} \cdot \log \theta}{1 + \theta}$. Using the `R` function `rexp`, draw a sample of 100000 from the distribution of ψ. Compute simulation estimates of $\mathrm{E}(\psi)$, $\mathrm{Var}(\psi)$, $\Pr(\psi > 0)$ and of the 0.025 and 0.975 quantiles of the distribution of ψ.

Lecture 3: Inference for a binomial proportion

3.1 Analysis using a Uniform prior (FCBSM §3.1; BDA2/3 §2.1, §2.3)

- $y_1, \ldots, y_n | \theta \sim$ i.i.d. $\mathrm{Ber}(\theta)$
- $y = \sum_{i=1}^{n} y_i$: total number of "successes" in n exchangeable (the order in which the successes and failures happen doesn't matter) trials
- *Likelihood.* $y \sim \mathrm{Bin}(n, \theta)$

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

- *Prior.* With no prior information, may want to use $\theta \sim \mathrm{Un}(0,1)$ $(= \mathrm{Be}(1,1))$

$$p(\theta) = 1 \qquad 0 < \theta < 1$$

- *Posterior.*

$$\begin{aligned} p(\theta|y) &\propto p(\theta) p(y|\theta) \\ &\propto \theta^y (1-\theta)^{n-y} \qquad 0 < \theta < 1 \\ &\propto \mathrm{Be}(y+1, n-y+1) \quad \text{density} \end{aligned}$$

  Plots, summaries are readily available in R

Example

We toss $n = 15$ times a coin with $\Pr(\text{Heads}) = \theta$. The outcomes of the tosses are FFFFFFSFFFFFFSF, where S and F denote success (heads) and failure (tails), respectively. The plots in Fig. 1 display the $\mathrm{Be}(y+1, n-y+1)$ posterior distribution of θ, as the result of each toss becomes available. The R code used to produce the plots is reported in the appendix. After observing 2 successes and 13 failures, the posterior distribution of θ, with a $\mathrm{Un}(0,1)$ prior, is $\mathrm{Be}(3, 14)$.

A plot of the density can be produced in R using the function `dbeta`. A central posterior interval can be computed using the inverse cdf function (`qbeta` in R) (Fig. 2, top).

One can also sample from the posterior, using the R function `rbeta`, make a histogram of the draws and compute posterior summaries using the simulated draws. See Fig. 2 (bottom). Details of the R code are in the appendix.

Is the coin fair?

- Let's address this question by computing $\Pr(\theta \geqslant 0.5|y)$. Using the $\mathrm{Be}(3, 14)$ cdf:

```
> 1 - pbeta(q=0.5, shape1=3, shape2=14)
[1] 0.002090454
```

- Using simulation: let $\psi = I_A(\theta)$, with $A = [0.5, 1)$ so that $\mathrm{E}[\psi|y] = \Pr(\theta \geqslant 0.5|y)$. This can be estimated using the sample of 10000 from $\mathrm{Be}(3, 14)$:

```
> psi <- draws >= 0.5
> mean(psi)
[1] 0.0017
```

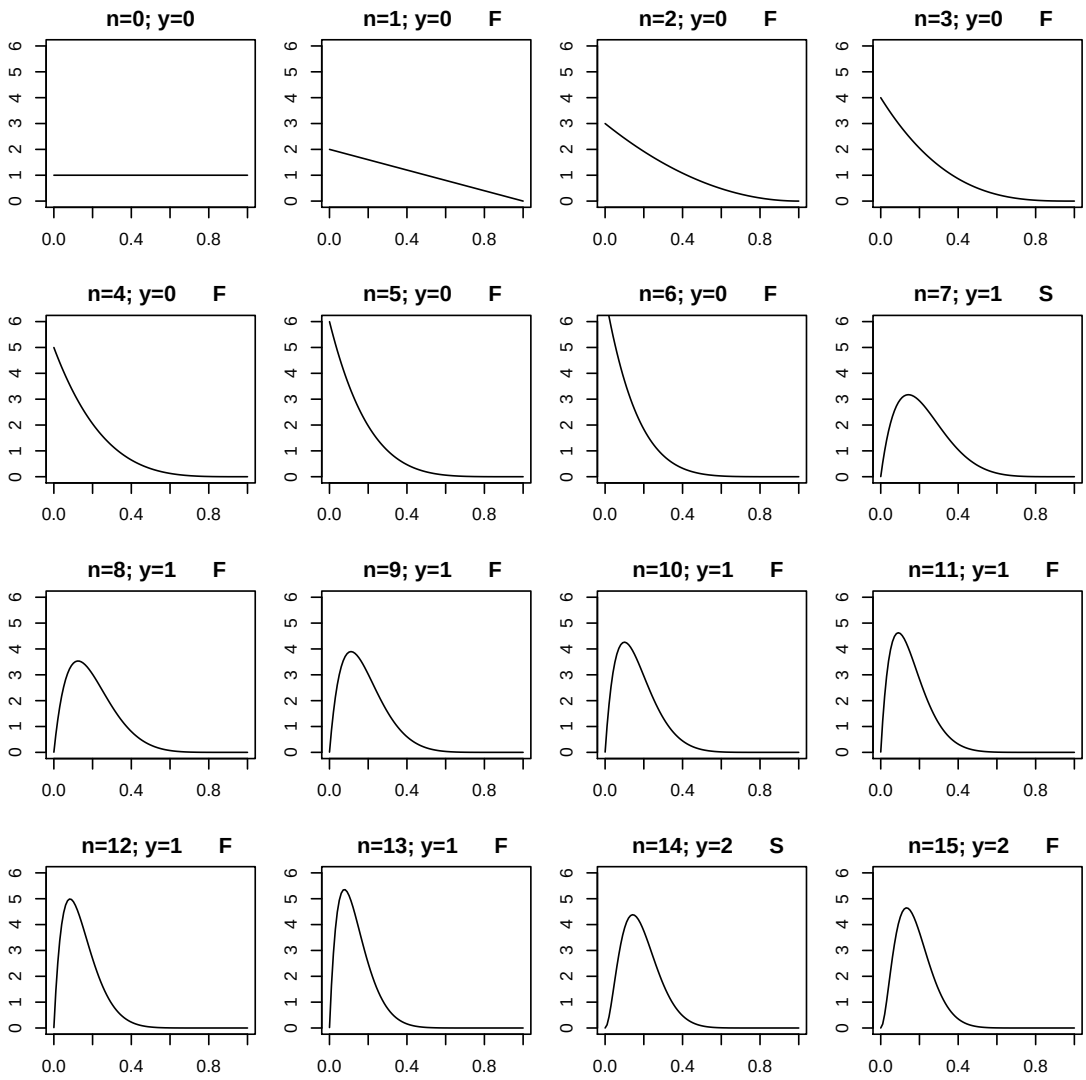


Figure 1: $\mathrm{Be}(y+1, n-y+1)$ posterior distribution of θ in a $\mathrm{Bin}(n, \theta)$ model, with a $\mathrm{Un}(0,1)$ prior. Top left panel is the prior, other panels display successive posterior densities as data becomes available, toss by toss. Outcome of the trial (F=failure, S=success) indicated in the title.

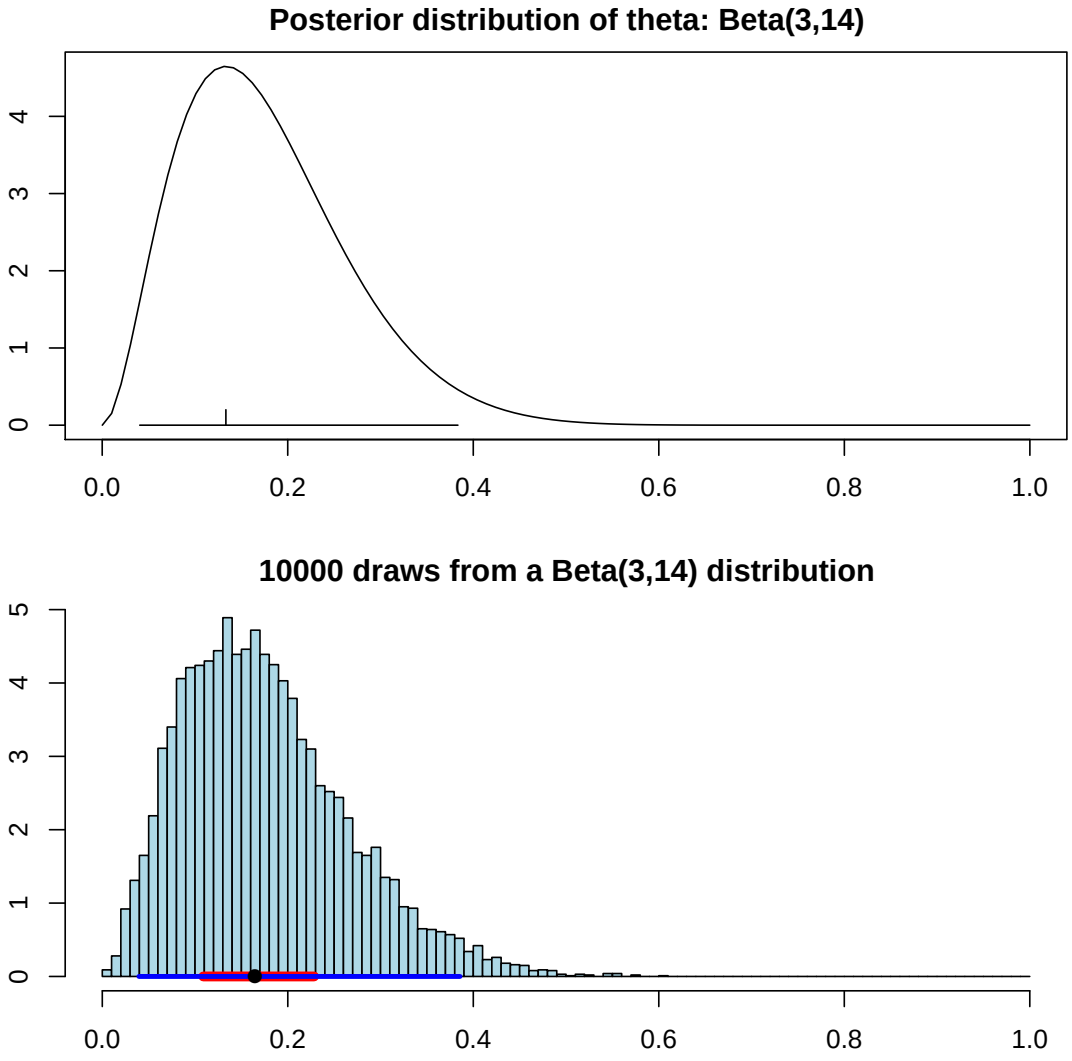


Figure 2: $Be(3, 14)$ posterior distribution of θ, after $y = 2$ successes in $n = 15$ trials. In the top panel, the horizontal line denotes the central 95% posterior interval, the mark is the mode of the distribution. In the bottom panel, the lines are the simulation 95% and 50% central intervals, the mark is the median of the simulation sample.

Prior Prediction

- How many successes does the model predicts, before seeing any data?

$$\begin{aligned} p(y) &= \int_0^1 p(y|\theta)p(\theta)\,d\theta \\ &= \int_0^1 \binom{n}{y}\theta^y(1-\theta)^{n-y}\,d\theta \\ &= \binom{n}{y}B(y+1, n-y+1) \\ &= \frac{n!}{y!(n-y)!}\frac{y!(n-y)!}{(n+1)!} \qquad \text{since } B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} \\ &= \frac{1}{n+1} \qquad y = 0, 1, \ldots, n \end{aligned}$$

- Justification adduced by Bayes for using $p(\theta) = 1$.

Posterior Prediction

- $\tilde{y}$: outcome of a future trial, exchangeable with the data y, i.e.,

$$Pr(\tilde{y} = 1|\theta, y) = Pr(\tilde{y} = 1|\theta) = \theta$$

- Then

$$\begin{aligned} p(\tilde{y} = 1|y) &= \int_0^1 p(\tilde{y} = 1, \theta|y)\,d\theta \\ &= \int_0^1 p(\tilde{y} = 1|\theta, y)\,p(\theta|y)\,d\theta \\ &= \int_0^1 \theta\,p(\theta|y)\,d\theta \\ &= E(\theta|y) = \frac{y+1}{n+2} \end{aligned}$$

- Laplace's law of succession, about which J. M. Keynes remarks:

> "No other formula in the alchemy of logic has exerted more astonishing powers. For it has established the existence of God from the premiss of total ignorance; and it has measured with numerical precision the probability that the sun will rise to-morrow." (*A Treatise on Probability*, 1921, p. 82)

3.2 Analysis with an informative prior (BDA2/3 §2.4; FCBSM §3.1 (p.37)

For Binomial data, Beta Prior ⇒ Beta Posterior

- Suppose we assume that a priori $\theta \sim Be(\alpha, \beta)$

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \qquad 0 < \theta < 1$$

  - Special case: $\alpha = \beta = 1$ yields $Un(0, 1)$

- Then

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

- *Conjugacy:* The posterior is of the same parametric form as the prior
- Beta is the natural conjugate prior for binomial data
  - information in $Be(\alpha, \beta)$ prior is equivalent to $\alpha+\beta$ observations
  - α, β are hyperparameters, completely specify the prior
  - might be determined by assigning the mean m and variance v of $p(\theta)$, then solving

$$m = \frac{\alpha}{\alpha+\beta} \qquad v = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Example (cont.)

Re-analyze the coin toss data using two other priors:

- $Be(2.625, 2.625)$: mean is 0.5, s.d. is 0.2
- $Be(0.5, 0.5)$: Jeffreys' prior – more on this later

See Fig. 3.

Exercise

We have seen that if a priori $\theta \sim Un(0,1)$, then the prior predictive distribution is $p(y) = 1/(n+1)$, $y = 0, 1, \ldots, n$. Find the prior predictive distribution when a priori $\theta \sim Be(\alpha, \beta)$.

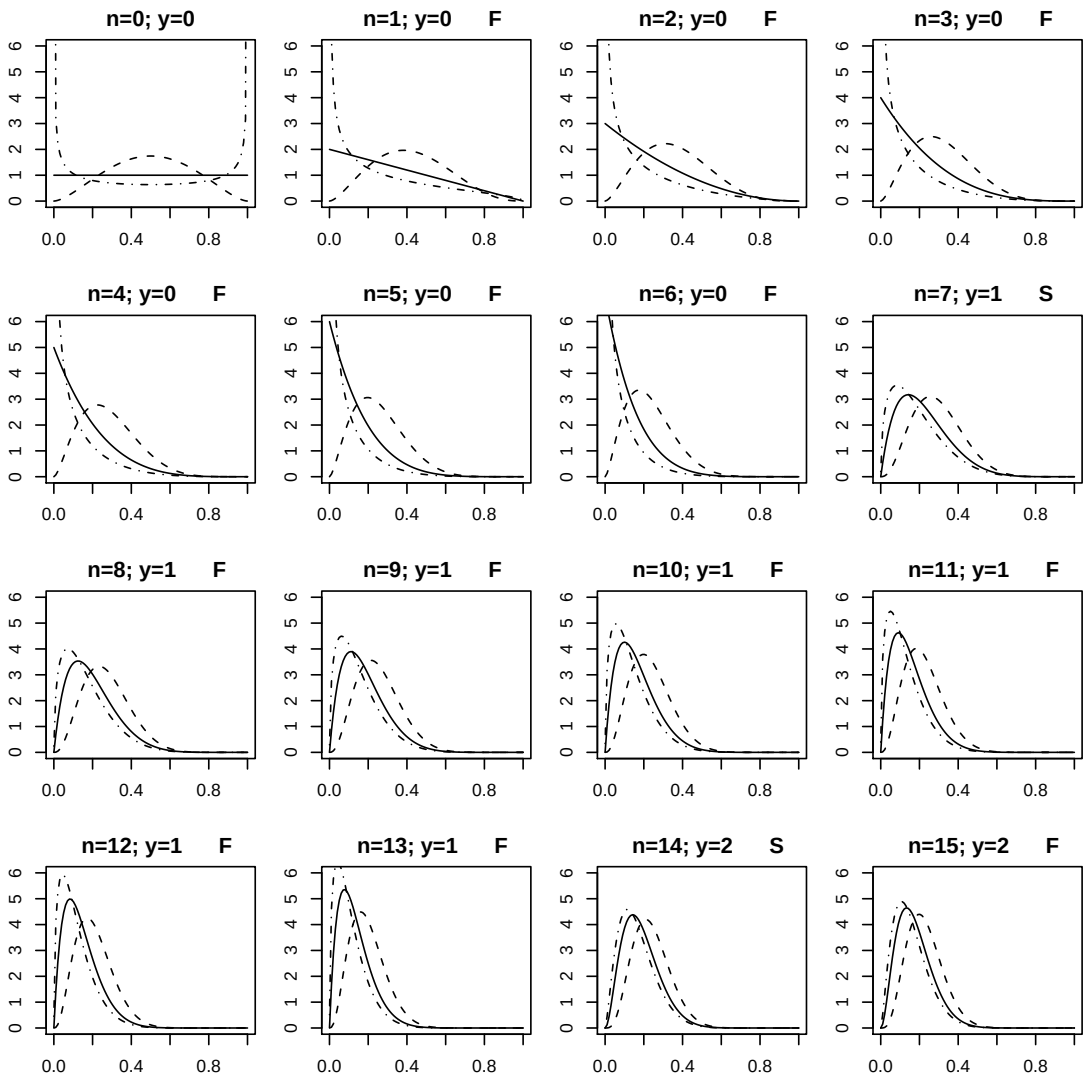


Figure 3: $\mathrm{Be}(\alpha + y, \beta + n - y)$ posterior distributions of θ in a $\mathrm{Bin}(n, \theta)$ model, corresponding to the $\mathrm{Be}(1, 1) = \mathrm{Un}(0, 1)$ prior and two additional priors. The broken lines correspond to a $\mathrm{Be}(2.625, 2.625)$ prior, having mean 0.5 and s.d. 0.2. The dotted lines correspond to the Jeffreys' prior $\mathrm{Be}(0.5, 0.5)$.

Appendix: R code for the Binomial data example

```
yi <- c(0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0)
n <- length(yi)
theta <- seq(0.0001, 0.9999, length=100)

# Analysis with the Un(0,1) prior
# prior 1:  Un(0,1)
par(mar=c(3,2,2,1)+0.1)
alpha1 <- beta1 <- 1
par(mfrow=c(4,4))
y <- 0
dens <- dbeta(theta, alpha1, beta1)
plot(theta, dens, type="l", ylim=c(0,6), ylab="density")
title(main=paste("n=", 0, "; y=", 0, sep=""))
for (i in (1:n))
{
  y <- y + yi[i]
  dens <- dbeta(theta, alpha1+y, beta1+i-y)
  plot(theta, dens, type="l", ylim=c(0,6), ylab="density")
  if(yi[i] == 1) out <- "S" else out <- "F"
  title(main=paste("  n=", i, "; y=", y, "      ", out, sep=""))
}

# Exact 95% Posterior interval in (lower, upper)
par(mfrow=c(2,1), mar=c(3,2,2,1)+0.1)
y <- sum(yi)
alphaU <- alpha1 + y
betaU <- beta1 + n - y
dens <- dbeta(theta, alphaU, betaU)
plot(theta, dens, type="l", ylab="density")
title(main=paste("Posterior distribution of theta: Beta(",
        alphaU, ",", betaU, ")", sep=""))
modtheta <- (alphaU - 1) / (alphaU + betaU - 2)
lines(rep(modtheta, 2), c(0, 0.2))
lower <- qbeta(p=0.025, alphaU, betaU)
upper <- qbeta(p=0.975, alphaU, betaU)
lines(c(lower, upper), rep(0, 2))
#> c(lower, upper)
#[1] 0.04047373 0.38347624

# Summaries of the posterior using simulation
set.seed(432)                     # set seed to be able to reproduce results
nsamp <- 10000
draws <- rbeta(nsamp, alphaU, betaU)
hist(draws, breaks=seq(0, 1, by=0.01), prob=T, main="")
title(main=paste(nsamp, " draws from a Beta(",
        alphaU, ",", betaU, ") distribution", sep=""))
sorted <- sort(draws)
lower95 <- sorted[nsamp * 0.025]
upper95 <- sorted[nsamp * 0.975 + 1]
lower50 <- sorted[nsamp * 0.25]
upper50 <- sorted[nsamp * 0.75 + 1]
med <- median(sorted)
```

```
lines(c(lower95, upper95), rep(0, 2), lwd=2)
lines(c(lower50, upper50), rep(0, 2), lwd=4)
points(med, 0, pch=16, cex=1)
#> c(lower95, upper95)
#[1] 0.03951703 0.38566722

# Comparison of the analyses with the 3 priors
# prior 2: alpha, beta s.t. E(theta)=0.5, sqrt(Var(theta)) = 0.2
alpha2 <- beta2 <- 2.625
# prior 3: Jeffrey's prior
alpha3 <- beta3 <- 0.5
par(mfrow=c(4,4), mar=c(3,2,2,1)+0.1)
y <- 0
dens <- cbind(dbeta(theta, alpha1, beta1),
              dbeta(theta, alpha2, beta2),
              dbeta(theta, alpha3, beta3))
matplot(theta, dens, type="l", lty=c(1,2,4), col=1, ylim=c(0,6),
        ylab="density")
title(main=paste("n=", 0, "; y=", 0, sep=""))
for (i in (1:n))
{
  y <- y + yi[i]
  dens <- cbind(dbeta(theta, alpha1+y, beta1+i-y),
                dbeta(theta, alpha2+y, beta2+i-y),
                dbeta(theta, alpha3+y, beta3+i-y))
  matplot(theta, dens, type="l", lty=c(1,2,4), col=1, ylim=c(0,6),
          ylab="density")
  if(yi[i] == 1) out <- "S" else out <- "F"
  title(main=paste("  n=", i, "; y=", y, "      ", out, sep=""))
}
```

Lecture 4: Inference for a binomial proportion

4.1 Analysis with an informative prior (cont.) [BDA2/3 §2.4; FCBSM §3.1 (p.37)]

We have seen that if

- likelihood: $y|\theta \sim Bin(n, \theta)$ and
- prior: $\theta \sim Be(\alpha, \beta)$, then
- posterior: $\theta|y \sim Be(\alpha + y, \beta + n - y)$.

Posterior summaries

- Posterior mode or MAP (maximum a posteriori) estimate

$$\theta_{MAP} = \frac{\alpha + y - 1}{\alpha + \beta + n - 2}$$

  - Easily obtained by differentiating $\log p(\theta|y)$ w.r.t. θ
  - reduces to MLE when $\alpha = \beta = 1$

- Posterior mean (from a property of the beta distribution)

$$E(\theta|y) = \frac{\alpha + y}{\alpha + \beta + n} = \frac{\frac{\alpha}{n} + \frac{y}{n}}{\frac{\alpha + \beta}{n} + 1}$$

  as $n \to \infty$, $E(\theta|y) \approx \frac{y}{n} = \hat{\theta}$ (MLE)

- One can also prove (we don't do it):
  - an asymptotic formula for the posterior variance

$$Var(\theta|y) \approx \frac{\hat{\theta}(1 - \hat{\theta})}{n} \qquad \text{as } n \to \infty$$

  - and a Bayesian C.L.T.:

$$\left. \frac{\theta - E(\theta|y)}{Var(\theta|y)^{1/2}} \right| y \to N(0, 1)$$

  - Hence, for large n, the posterior distribution $p(\theta|y)$ is approximately normal and centred at the MLE.

Non-conjugate priors

Conjugate priors:

- easy to update into posteriors
- easy to display/summarize

Prior information *may* not be representable by a conjugate distribution, e.g.

- bimodal prior on θ

- you believe a coin is biased, but do not know which direction

You can:

- try approximating the prior using a *mixture* of conjugate distributions. For an example, see Appendix 2.
- choose any prior $p(\theta)$ on a fine grid of θ
  - evaluate $p(\theta)p(y|\theta)$ on the grid
  - plot it, after rescaling so that area is 1
  - sample from it, after normalizing
  - an example in BDA2 §2.5, p. 45–46; BDA3 §2.5, p. 38–39
  - only practical for 1-dim. and 2-dim. problems

Another prior for binomial data – working with logits

- Rewrite $\mathrm{Bin}(n, \theta)$ as an exponential family distribution

$$\begin{aligned} p(y|\theta) &\propto \theta^y (1-\theta)^{n-y} \\ &= (1-\theta)^n \exp\left\{ y \log \frac{\theta}{1-\theta} \right\} \\ &= d(\theta) S(y) \cdot \exp\{T(y) c(\theta)\} \quad \text{(general form)} \end{aligned}$$

  $\phi = \log \frac{\theta}{1-\theta} = \mathrm{logit}(\theta)$: *natural* parameter of the $\mathrm{Bin}(n, \theta)$
- logit transform makes binomial likelihood $p(y|\theta)$ more symmetric
- $\{\text{draws from } p(\phi|y)\} \longleftrightarrow \{\text{draws from } p(\theta|y)\}$ using

$$\phi = \log \frac{\theta}{1-\theta} \qquad \theta = \frac{e^\phi}{1+e^\phi}$$

- May want to specify a prior in terms of ϕ, rather than θ, e.g.

$$\phi \sim N(\mu, \sigma^2)$$

  From the change of variables formula, the prior pdf on θ induced by the normal prior on ϕ is

$$p(\theta) = p(\phi) \left| \frac{d\phi}{d\theta} \right| = N\left(\log \frac{\theta}{1-\theta}; \mu, \sigma^2 \right) \frac{1}{\theta(1-\theta)}$$

  Some examples, with $\mu = 0$ and various values of σ are displayed in Fig. 4.

Appendix 1. R code for priors on θ using the logits

```
mu <- 0
sigma <- c(0.1, 0.5, 1, 1.5, 2, 2.2, 2.5, 3, 4)

ptheta <- function(theta, mu, sigma){
  dnorm(log(theta/(1-theta)), mu, sigma) / (theta*(1-theta))
}
```

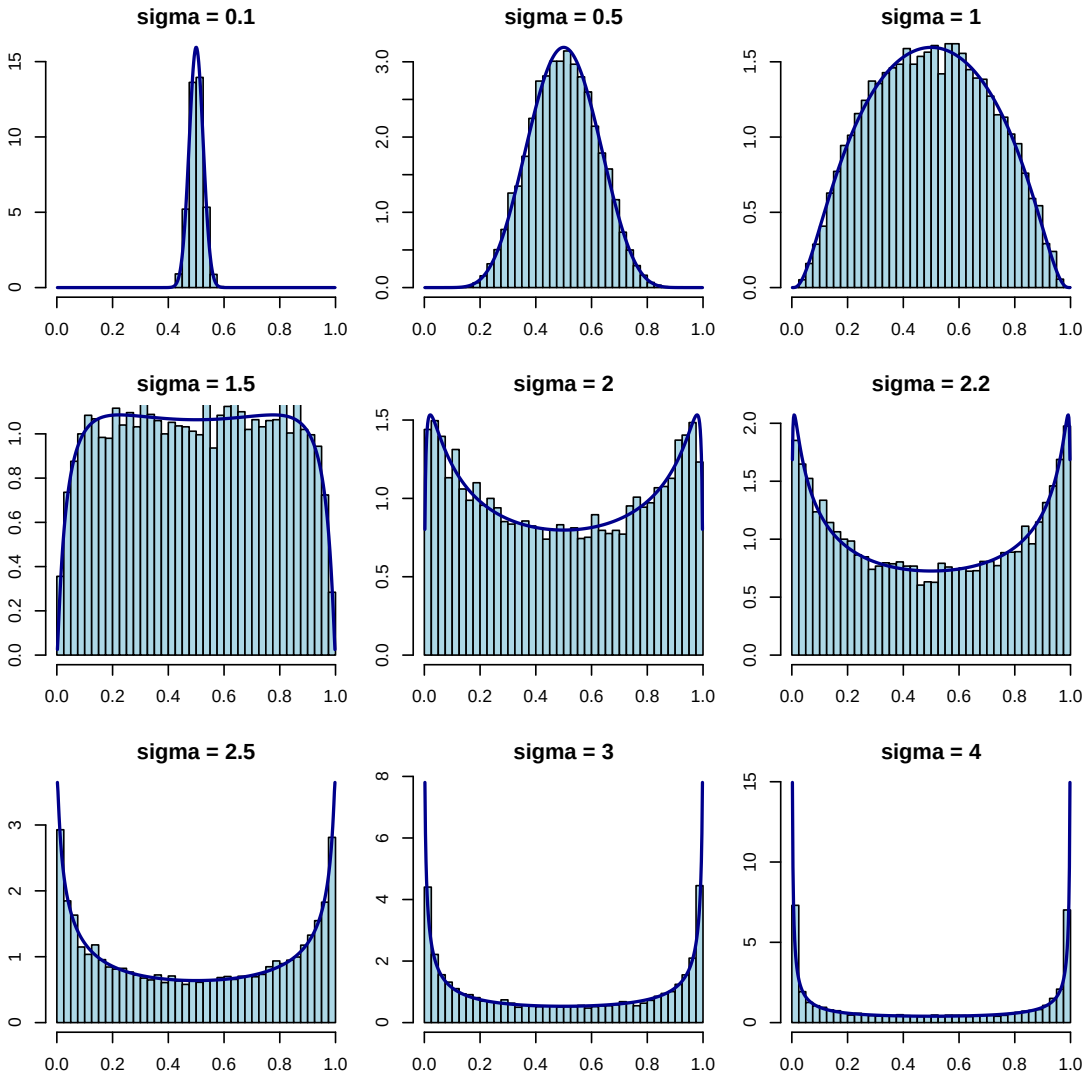


Figure 4: Priors on θ induced by a $N(\mu, \sigma^2)$ prior on $\phi = \text{logit}(\theta)$, for $\mu = 0$ and several values of σ. Both the density and 10,000 draws from the induced prior are displayed.

```
theta <- seq(0.002, 0.998, length=500)

par(mar=c(3,2,2,1)+0.1)
par(mfrow=c(3,3))

for (i in (1:length(sigma))) {
  dens <- ptheta(theta, mu, sigma[i])
  phidraws <- rnorm(10000, mu, sigma[i])
  thetadraws <- exp(phidraws) / (1+exp(phidraws))
  hist(thetadraws, breaks=seq(0, 1, by=0.025), prob=T, xlab="theta",
       ylab="Prior on theta", ylim=c(0, max(dens)), main="", col="lightblue")
  lines(theta, dens, col="darkblue", lwd=2)
  title(main=paste("sigma = ", sigma[i], sep=""), cex=0.4)
}
```

Appendix 2. Example: binomial data with a "mixture of betas" prior

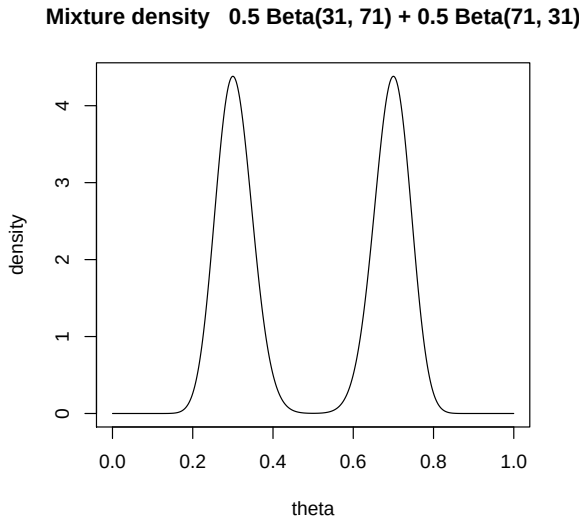
You have flipped a certain coin 100 times in the past, recording 30 successes, but have now forgotten whether "success" was heads or tails. Let the probability of the coin falling heads on a single trial be θ, and assign it a $\mathrm{Un}(0,1)$ prior.

(a) What is the posterior distribution of θ?

Let H = Heads and T = Tails and denote by D the information from the original 100 flips. Define a random variable s which describes what "success" meant in those flips. Since D tells you nothing about s, you can let $\mathrm{p}(s = \mathrm{H}|\mathrm{D}) = \mathrm{p}(s = \mathrm{T}|\mathrm{D}) = 0.5$. Then the posterior of θ is

$$\begin{aligned} \mathrm{p}(\theta|\mathrm{D}) &= \mathrm{p}(\theta|\mathrm{D}, s = \mathrm{H})\ \mathrm{p}(s = \mathrm{H}|\mathrm{D}) + \mathrm{p}(\theta|\mathrm{D}, s = \mathrm{T})\ \mathrm{p}(s = \mathrm{T}|\mathrm{D}) \\ &= \tfrac{1}{2}\mathrm{Be}(31, 71) + \tfrac{1}{2}\mathrm{Be}(71, 31) \end{aligned} \tag{1}$$

This mixture density is plotted below:



Posterior distribution of θ with a $\mathrm{Un}(0,1)$ prior, given only that in 100 flips either 30 heads or 30 tails occurred.

The R code used to produce the plot:

```
theta <- seq(0.0001, 0.9999, length=500)
postD <- 0.5 * dbeta(theta, 31, 71) + 0.5 * dbeta(theta, 71, 31)
plot(theta, postD, type="l", ylab= "density",
     main="Mixture density   0.5 Beta(31, 71) + 0.5 Beta(71, 31)")
```

(b) Now you are allowed $n = 10$ more flips, resulting in $y = 7$ heads. What is your updated posterior of θ?

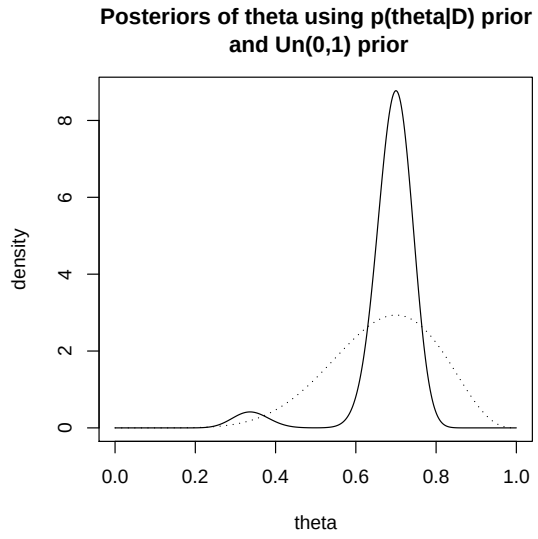
Regard the posterior $p(\theta|D)$ as the prior distribution of θ immediately before the 10 flips experiment. Then, the posterior after the additional 10 flips is

$$\begin{aligned} p(\theta|y, D) &\propto \theta^y(1-\theta)^{n-y}p(\theta|D) \\ &= \theta^7(1-\theta)^3\left[\tfrac{1}{2}Be(31,71) + \tfrac{1}{2}Be(71,31)\right] \end{aligned}$$

This density can be easily evaluated on a grid of θ values, then renormalized:

```
posty <- theta^7 * (1-theta)^3 * postD
posty <- posty / mean(posty)
plot(theta, posty, type="l", ylab="density",
     main="Posteriors of theta using p(theta|D) prior and Un(0,1) prior")
lines(theta, dbeta(theta, 8, 4), lty=3)
```

The resulting density is displayed as a solid line in the plot below:



Posterior distribution of θ *using the correct prior* $p(\theta|D)$ *(solid line) and using a* $Un(0,1)$ *prior (broken line).*

For comparison, the plot also displays, as a broken line, the $Be(8,4)$ posterior distribution resulting from a $Un(0,1)$ prior and the likelihood of the last 10 flips only. Disregarding the (partial) information from the original 100 flips leads to a much less accurate inference.

The posterior $p(\theta|y, D)$ has been evaluated numerically. However, one can show that, like the prior $p(\theta|D)$, it is a mixture of two beta distributions. To obtain

the exact solution, let's restate the problem in more general terms, assuming that the prior is a generic mixture of two betas:

$$p(\theta|D) = \rho Be(\alpha_1, \beta_1) + (1-\rho)Be(\alpha_2, \beta_2),$$

where the mixing weight $\rho \in (0,1)$. Then

$$\begin{aligned}
p(\theta|y,D) &\propto \theta^y(1-\theta)^{n-y}\left[\rho Be(\alpha_1,\beta_1) + (1-\rho)Be(\alpha_2,\beta_2)\right] \\
&= \theta^y(1-\theta)^{n-y}\left[\rho\frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)}\theta^{\alpha_1-1}(1-\theta)^{\beta_1-1}\right. \\
&\qquad\left. +(1-\rho)\frac{\Gamma(\alpha_2+\beta_2)}{\Gamma(\alpha_2)\Gamma(\beta_2)}\theta^{\alpha_2-1}(1-\theta)^{\beta_2-1}\right] \\
&= \rho\frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)}\theta^{\alpha_1+y-1}(1-\theta)^{\beta_1+n-y-1} \\
&\quad + (1-\rho)\frac{\Gamma(\alpha_2+\beta_2)}{\Gamma(\alpha_2)\Gamma(\beta_2)}\theta^{\alpha_2+y-1}(1-\theta)^{\beta_2+n-y-1} \\
&\propto \frac{\rho_1}{\rho_1+\rho_2}Be(\alpha_1+y,\beta_1+n-y) + \frac{\rho_2}{\rho_1+\rho_2}Be(\alpha_2+y,\beta_2+n-y),
\end{aligned}$$

where

$$\begin{aligned}
\rho_1 &= \rho\frac{\Gamma(\alpha_1+\beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)}\frac{\Gamma(\alpha_1+y)\Gamma(\beta_1+n-y)}{\Gamma(\alpha_1+\beta_1+n)}, \\
\rho_2 &= (1-\rho)\frac{\Gamma(\alpha_2+\beta_2)}{\Gamma(\alpha_2)\Gamma(\beta_2)}\frac{\Gamma(\alpha_2+y)\Gamma(\beta_2+n-y)}{\Gamma(\alpha_2+\beta_2+n)}.
\end{aligned}$$

In our example, $\rho = 0.5,\ \alpha_1 = \beta_2 = 31,\ \alpha_2 = \beta_1 = 71,$ so that

$$\frac{\rho_1}{\rho_2} = \frac{\Gamma(31+7)\Gamma(71+3)}{\Gamma(71+7)\Gamma(31+3)} = \frac{37\cdot 36\cdot 35\cdot 34}{77\cdot 76\cdot 75\cdot 74} \approx 0.0488.$$

In R:

```
exp(lgamma(38) + lgamma(74) - lgamma(78) - lgamma(34))
[1] 0.04880383
```

Therefore $\rho_2/(\rho_1+\rho_2) = 0.9535$ and the posterior is

$$p(\theta|y,D) = 0.0465\ Be(38,74) + 0.9535\ Be(78,34) \tag{2}$$

Notice how the prior distribution (1) gets updated into the posterior (2): not only the shape parameters of the two beta distributions in the mixture change (according to the beta–binomial updating rule), but also there is a large shift in the mixture weights.

Homework 2

Problem 1 [BDA2/3, Exercise 2.11.1]

Posterior inference: suppose there is Beta(4,4) prior distribution on the probability θ that a coin will yield a head when spun in a specified manner. The coin is independently spun ten times, and heads appear fewer than 3 times. You are not told how many heads were seen, only that the number is less than 3. Calculate your exact posterior density (up to a proportionality constant) for θ and sketch it.

Problem 2 [BDA2/3, Exercise 2.11.2]

Predictive distributions: consider two coins, C_1 and C_2, with the following characteristics: $\Pr(\text{heads}|C_1) = 0.6$ and $\Pr(\text{heads}|C_2) = 0.4$. Choose one of the coins at random and imagine spinning it repeatedly. Given that the first two spins from the chosen coin are tails, what is the expectation of the number of additional spins until a head shows up?

Problem 3

Refer to the example 'binomial data with a "mixture of betas" prior'. Show that the weights in the posterior distribution $p(\theta|y, D)$ given in (2) are $\Pr[s = H|y, D]$ and $\Pr[s = T|y, D]$.

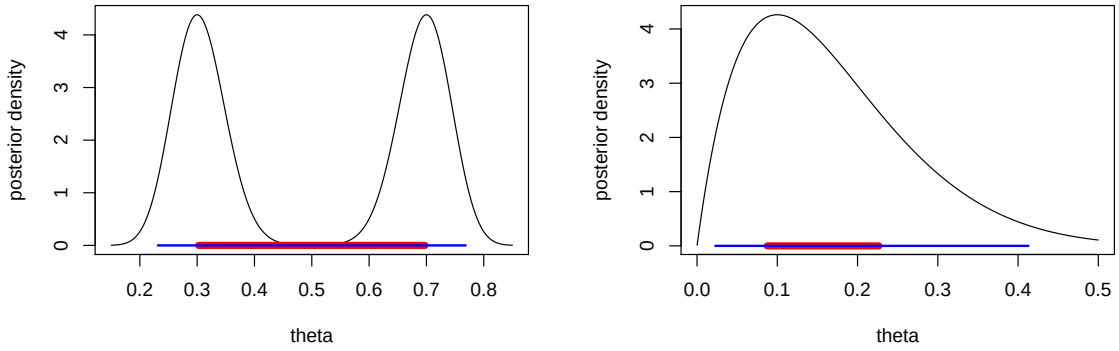
Problem 4

In a sequence of $n = 20$ i.i.d. Bernoulli trials with probability of success θ, you observe $y = 16$ successes and $n-y = 4$ failures. Assuming a $Un(0, 1)$ prior on θ, either compute or estimate by simulation (a) $\Pr[0.5 < \theta < 0.75|y]$, (b) the mean and median of the posterior distribution of θ, (c) a central 95% posterior interval for θ.
Repeat the exercise assuming that a priori θ has a Beta distribution with mean 0.6 and standard deviation 0.2.

Lecture 5

5.1 Central posterior intervals and HPDRs

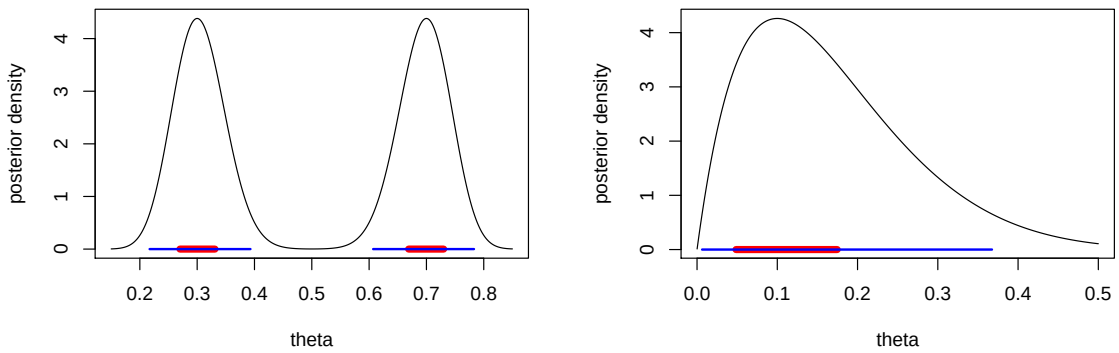
- Central posterior intervals: may be inadequate for
  - multimodal or highly-skewed posterior distributions



50% (thick line) and 95% (thin line) central posterior intervals for the distribution $0.5\text{Be}(71, 31) + 0.5\text{Be}(31, 71)$ *on the left hand side, and for the distribution* $\text{Be}(2, 10)$ *on the right hand side*

- In these cases, best to report the plot of $p(\theta|y)$
- Alternatively, use a Highest Posterior Density Region (HPDR)

  A is a 95% HPDR for θ if
  - $\Pr[\theta \in A|y] = 0.95$ and
  - for all $\theta_1 \in A$ and $\theta_2 \notin A$ one has $p(\theta_1|y) \geqslant p(\theta_2|y)$



50% (thick line) and 95% (thin line) HPDRs for the distribution $0.5\text{Be}(71, 31) + 0.5\text{Be}(31, 71)$ *on the left hand side, and for the distribution* $\text{Be}(2, 10)$ *on the right hand side*

- Central posterior intervals are
  - easier to compute than HPDR's
  - invariant to 1-1 transformations

Some review exercises (Lectures 3 and 4)

Suppose you perform ten independent Bernoulli trials and observe eight successes. Let θ denote the probability of success on each trial.

(a) Compute the MAP and posterior expectation of θ using the following three prior distributions:

- Un(0,1)
- Be(0.5, 0.5)
- Be(α, β) with hyperparameter values such that both the prior mean and the prior standard deviation are equal to $1/3$.

(b) Using the Be(0.5, 0.5) prior, and after observing the data, what probability would you attach to the event that the eleven-th trial will result in a success?

(c) You believe that Beta priors are not flexible enough and prefer to specify a prior on θ by means of a normal prior on the logits:

$$\phi = \text{logit}(\theta) \qquad \phi \sim N(\mu, \sigma^2).$$

You want the induced prior on θ to satisfy the following conditions:

$$\Pr[\theta > 0.7] = 1/2 \qquad \text{and} \qquad \Pr[\theta < 0.9] = 0.975$$

Determine the values of μ and σ required.

Lecture 6: Normal data (BDA2 §2.6-2.7, BDA3 §2.5-2.6; FCBSM §5.1-5.3)

Three cases:

- Mean unknown, variance known: BDA2 §2.6; BDA3 §2.5; FCBSM §5.2
- Mean known, variance unknown: BDA2 §2.7; BDA3 §2.6
- Both mean and variance unknown: BDA2 §3.2-3.4; BDA3 §3.2-3.3; FCBSM §5.3

Today:

- Mean unknown, variance known
  - Single observation case
  - Several observations

6.1 Mean unknown, variance known

Single observation case

- Model:

$$\begin{aligned} y|\theta &\sim N(\theta, \sigma^2) \qquad \sigma \text{ known} \\ \theta &\sim N(\mu_0, \tau_0^2) \qquad \mu_0, \tau_0 \text{ fixed constants} \end{aligned}$$

- Posterior:

$$\begin{aligned} p(\theta|y) &\propto p(\theta)p(y|\theta) \\ &\propto \exp\left\{-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right\}\exp\left\{-\frac{1}{2\sigma^2}(y-\theta)^2\right\} \\ &= \exp\left\{-\frac{1}{2}\left[\frac{(\theta-\mu_0)^2}{\tau_0^2}+\frac{(y-\theta)^2}{\sigma^2}\right]\right\} \\ &\quad \cdots \text{this gap is a homework problem} \cdots \\ &\propto \exp\left\{-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2,\right\} \end{aligned}$$

  where

$$\frac{1}{\tau_1^2} = \frac{1}{\tau_0^2}+\frac{1}{\sigma^2} \qquad \mu_1 = \frac{\mu_0/\tau_0^2 + y/\sigma^2}{1/\tau_0^2+1/\sigma^2}$$

  Therefore

$$\theta|y \sim N(\mu_1, \tau_1^2),$$

  which shows that the prior is conjugate

- Notice how the hyperparameters μ_0, τ_0^2 are updated after observing y:

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

$$\text{posterior precision} = \text{prior precision} + \text{data precision}$$

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

$$\text{posterior mean} = \text{weighted average of prior mean and data}$$

- Posterior predictive distribution - $\tilde{y}$: future observation

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

Since

$$\begin{aligned} \theta|y &\sim N(\mu_1, \tau_1^2) \\ \tilde{y}|\theta, y &\sim N(\theta, \sigma^2) \qquad \text{indep. of } y \end{aligned}$$

one has

$$(\tilde{y}, \theta)|y \sim \text{Bivariate Normal}$$

which implies

$$\tilde{y}|y \sim \text{Normal}$$

Need only determine the mean and variance of this normal

  - mean:

$$\begin{aligned} E[\tilde{y}|y] &= E[E(\tilde{y}|\theta, y)|y] \\ &= E[\theta|y] \\ &= \mu_1 \end{aligned}$$

  - variance:

$$\begin{aligned} Var(\tilde{y}|y) &= E[Var(\tilde{y}|\theta, y)|y] + Var(E[\tilde{y}|\theta, y]|y) \\ &= E[\sigma^2|y] + Var(\theta|y) \\ &= \sigma^2 + \tau_1^2 \end{aligned}$$

In summary

$$\tilde{y}|y \sim N(\mu_1, \sigma^2 + \tau_1^2)$$

Several observations

- Model:

$$\begin{aligned} y_1, \ldots, y_n|\theta &\sim \text{i.i.d. } N(\theta, \sigma^2) \qquad \sigma \text{ known} \\ \theta &\sim N(\mu_0, \tau_0^2) \qquad \mu_0, \tau_0 \text{ fixed constants} \end{aligned}$$

- Posterior:

$$\begin{aligned} p(\theta|y) &\propto \exp\left\{-\frac{1}{2\tau_0^2}(\theta - \mu_0)^2\right\} \prod_{i=1}^{n} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \theta)^2\right\} \\ &= \exp\left\{-\frac{1}{2}\left[\frac{1}{\tau_0^2}(\theta - \mu_0)^2 + \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \theta)^2\right]\right\} \\ &= \exp\left\{-\frac{1}{2}\left[\frac{1}{\tau_0^2}(\theta - \mu_0)^2 + \frac{1}{\sigma^2}\left(\sum y_i^2 - 2\theta\sum y_i + n\theta^2\right)\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[\frac{1}{\tau_0^2}(\theta - \mu_0)^2 + \frac{1}{\sigma^2}\left(n\theta^2 - 2n\theta\overline{y}\right)\right]\right\} \end{aligned}$$

Posterior $p(\theta|y)$ involves y through $\bar{y}$ only, i.e., $p(\theta|y) = p(\theta|\bar{y})$.

- Use results for case of single observation
- together with $\bar{y} \sim N(\theta, \sigma^2/n)$, to yield

$$\theta|y \sim N(\mu_n, \tau_n^2)$$

where

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

- Note the rule for updating the hyperparameters:

$$\begin{aligned} \text{posterior precision} &= \text{prior precision} + \text{precision of sample mean} \\ \text{posterior mean} &= \text{weighted average (prior mean, sample mean)} \end{aligned}$$

- Unless n is very small (or $\tau_0 \ll \sigma$)

$$\tau_n^2 \approx \frac{\sigma^2}{n} \qquad \text{and} \qquad \mu_n \approx \bar{y}$$

i.e.,

$$\theta|y \approx N(\bar{y}, \sigma^2/n)$$

A routine example

In a sample of $n = 16$ observations from the $N(\theta, 4)$ distribution, the sample mean is $\bar{y} = 3$. Write down the posterior distribution of θ and the posterior predictive distribution of a future observation $\tilde{y}$, assuming that a priori $\theta \sim N(1, \tau_0^2)$ with τ_0 taking on three different values: 1, 10 and 100.

Homework 2b

Problem 1 [BDA2/3, Exercise 2.11.8]

Normal distribution with unknown mean: a random sample of n students is drawn from a large population, and their weights are measured. The average weight of the n sampled students is $\bar{y} = 150$ pounds. Assume the weights in the population are normally distributed with unknown mean θ and known standard deviation 20 pounds. Suppose your prior distribution for θ is normal with mean 180 and standard deviation 40.

(a) Give your posterior distribution for θ. (Your answer will be a function of n.)

(b) A new student is sampled at random from the same population and has a weight of $\tilde{y}$ pounds. Give a posterior predictive distribution for $\tilde{y}$. (Your answer will still be a function of n.)

(c) For $n = 10$, give a 95% posterior interval for θ and a 95% posterior predictive interval for $\tilde{y}$.

(d) Do the same for $n = 100$.

Problem 2

Complete the derivation of $p(\theta|y)$ (the gap on page 31) in the normal data problem with mean θ unknown and variance σ^2 known, by proving that

$$\exp\left\{-\frac{1}{2}\left[\frac{(y-\theta)^2}{\sigma^2}+\frac{(\theta-\mu_0)^2}{\tau_0^2}\right]\right\} \propto \exp\left\{-\frac{1}{2}\frac{(\theta-\mu_1)^2}{\tau_1^2}\right\}$$

Lecture 7: Normal data (continued)

Three cases:

- Mean unknown, variance known: BDA2 §2.6; BDA3 §2.5; FCBSM §5.2
- Mean known, variance unknown: BDA2 §2.7; BDA3 §2.6
- Both mean and variance unknown: BDA2 §3.2-3.4; BDA3 §3.2-3.3; FCBSM §5.3

Today:

- Mean known, variance unknown
  - Two parameterizations for the prior on the variance
- Both mean and variance unknown
  - The conjugate prior: μ and σ^2 a priori dependent
  - A semi-conjugate independence prior

7.1 Mean known, variance unknown

- Model:

$$\begin{aligned} y_1, \ldots, y_n|\sigma^2 &\sim \text{i.i.d. } N(\theta, \sigma^2) && \theta \text{ known} \\ \sigma^2 &\sim \text{Inv-gamma}(\alpha, \beta) && \alpha, \beta \text{ fixed constants} \end{aligned}$$

  - *Aside:*
    If $X \sim \text{Gamma}(\alpha, \beta)$ and $Z = 1/X$, we say: $Z \sim \text{Inv-gamma}(\alpha, \beta)$

$$p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x > 0 \qquad\qquad p(z) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{-(\alpha+1)} e^{-\beta/z}, \quad z > 0$$

- Posterior:

$$\begin{aligned} p(\sigma^2|y) &\propto p(\sigma^2) p(y|\sigma^2) \\ &\propto (\sigma^2)^{-(\alpha+1)} e^{-\beta/\sigma^2} \cdot \sigma^{-n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \theta)^2 \right\} \end{aligned}$$

  Let

$$v = \frac{1}{n} \sum_{i=1}^{n} (y_i - \theta)^2,$$

  to re-write posterior as

$$p(\sigma^2|y) \propto (\sigma^2)^{-(\alpha + \frac{n}{2} + 1)} \exp\left\{ -\frac{1}{\sigma^2} \left[\beta + \frac{nv}{2} \right] \right\}$$

  Hence

$$\sigma^2|y \sim \text{Inv-gamma}\left(\alpha + \frac{n}{2}, \beta + \frac{nv}{2} \right)$$

  Again, this proves that the prior is conjugate

A more intuitive parameterization

Definition: We say that Z has a scaled inverse χ^2 distribution with d.o.f. ν and scale s^2, and write $Z \sim \text{Inv-}\chi^2(\nu, s^2)$, if and only if $Z \sim \text{Inv-gamma}(\nu/2, \nu s^2/2)$.

- Then we can restate the prior on σ^2 as

$$\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2)$$

  with $\nu_0 = 2\alpha$ and $\sigma_0^2 = 2\beta/\nu_0 = \beta/\alpha$

- Similarly the posterior becomes

$$\sigma^2|y \sim \text{Inv-}\chi^2\left(\nu_0 + n, \frac{\nu_0\sigma_0^2 + n\nu}{\nu_0 + n}\right)$$

- This parameterization is more intuitive:
  - n observations increase the prior d.o.f. ν_0 by n.
  - The prior scale σ_0^2 is replaced by a weighted average of the prior scale and the sample variance.
  - Can think of the prior as incorporating information equivalent to ν_0 observations with variance σ_0^2.

7.2 Both mean μ and variance σ^2 unknown

We will consider two priors:

The Conjugate Prior

- Model:

$$\begin{aligned} y_i|\mu, \sigma^2 &\sim \text{i.i.d. } N(\mu, \sigma^2) \qquad i = 1, \ldots, n \\ \mu|\sigma^2 &\sim N\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right) \\ \sigma^2 &\sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2) \end{aligned}$$

- $p(\mu|\sigma^2)$ involves σ^2: so μ and σ^2 are a priori dependent
  - prior precision on μ specified in terms of units of data precision
  - Reasonable only if prior information can be regarded as equivalent to that in a certain number of observations

- Perhaps useful to consider that
  - prior on $\mu|\sigma^2$ induces the prior below on the scaled mean μ/σ:

$$\frac{\mu}{\sigma}\Big|\, \sigma^2 \sim N\left(\frac{\mu_0}{\sigma}, \frac{1}{\kappa_0}\right)$$

- Interpretation of the prior:
  - information on σ^2 is equivalent to that in ν_0 obs. with variance σ_0^2

- information on μ is equivalent to that in κ_0 obs. with mean μ_0 and variance σ^2

- One can show that the posterior satisfies

$$\begin{aligned} \mu|\sigma^2, y &\sim N\left(\mu_n, \frac{\sigma^2}{\kappa_n}\right) \\ \sigma^2|y &\sim \text{Inv-}\chi^2(\nu_n, \sigma_n^2) \end{aligned}$$

where

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

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

and

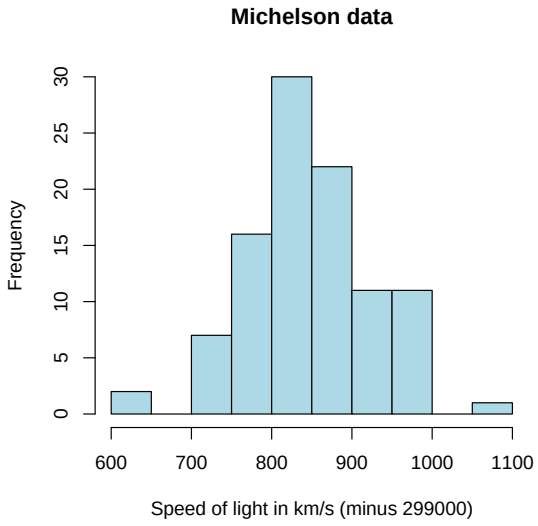
$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

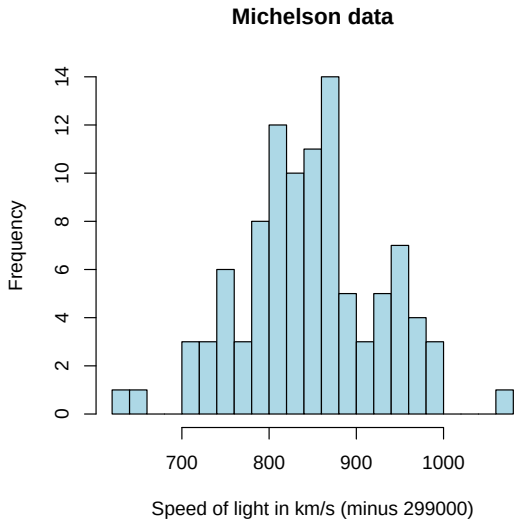
- To generate from the joint posterior $p(\mu, \sigma^2|y)$:

1. Draw a sample $\left\{\sigma^2_{(j)}\right\}_{j=1,N}$ from the Inv-$\chi^2(\nu_n, \sigma_n^2)$ marginal posterior distribution of σ^2
2. Substitute the draws $\left\{\sigma^2_{(j)}\right\}_{j=1,N}$ for σ^2 in the conditional posterior of μ, $p(\mu|\sigma^2, y) = N(\mu|\mu_n, \sigma^2/\kappa_n)$, then sample from it.

Example: Michelson data on speed of light

- Measurements of the speed of light in air, made by Michelson in the summer of 1879
- Values are in km/s, minus 299000
- Currently accepted value, on this scale of measurement, is 734.5[4]

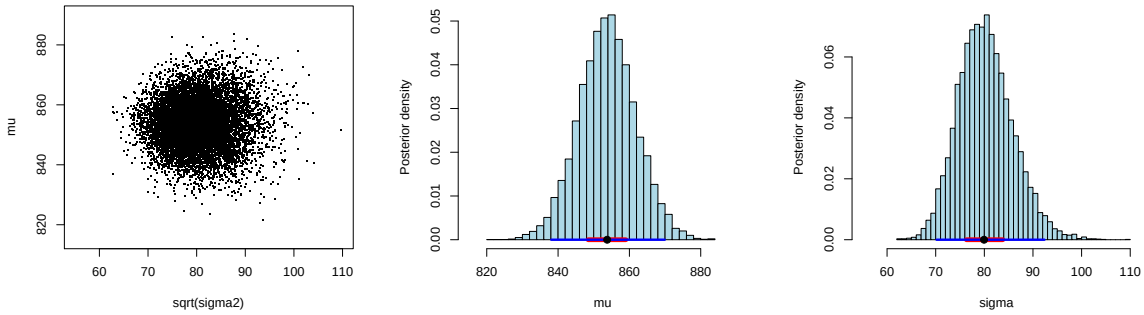




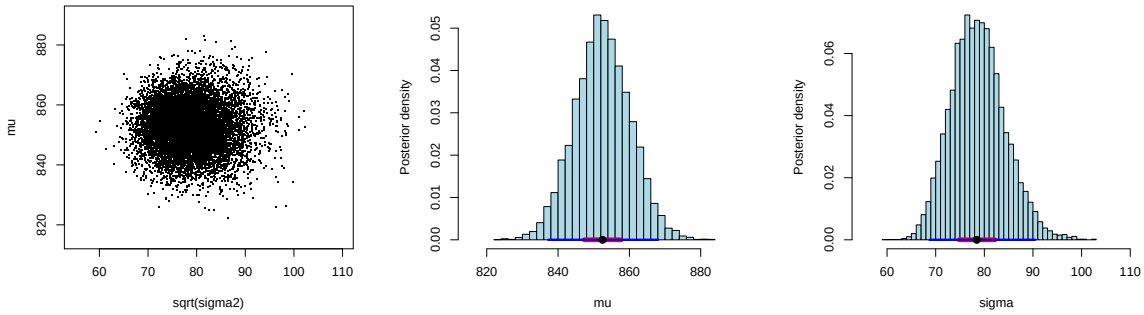
Summary statistics:

[4] In fact, now, the speed of light *in vacuum* is a constant, defined to 299,792.458 km/s. Combined with the standard unit of time, this effectively defines the metre.

n	$\bar{y}$	s
100	852.4	79.01



Conjugate prior with $\nu_0 = 1$, $\sigma_0 = 1$, $\mu_0 = 1000$, $\kappa_0 = 1$. *Left: 10000 draws from the joint posterior of* μ *and* σ*; Centre and Right: 10000 draws from the marginal posterior distributions of* μ *and* σ.



Conjugate prior with $\nu_0 = 1$, $\sigma_0 = 1$, $\mu_0 = 1$, $\kappa_0 = 10^{-6}$. *Left: 10000 draws from the joint posterior of* μ *and* σ*; Centre and Right: 10000 draws from the marginal posterior distributions of* μ *and* σ.

- Marginal posterior $p(\mu|y)$ of μ:

$$\left.\frac{\mu - \mu_n}{\sqrt{\sigma_n^2/\kappa_n}}\right| y \sim t_{\nu_n}$$

- Posterior predictive distribution

$$\begin{aligned} p(\tilde{y}|y) &= \int p(\tilde{y}|\mu, \sigma^2, y) p(\mu, \sigma^2|y)\, d\mu\, d\sigma^2 \\ &= \int N(\tilde{y}|\mu, \sigma^2) p(\mu, \sigma^2|y)\, d\mu\, d\sigma^2 \end{aligned}$$

  - this is also t_{ν_n} (with certain location and scale)

- To draw from $p(\tilde{y}|y)$:
  - draw from the joint posterior of μ and σ^2 as shown above
  - then plug in the normal distribution of $\tilde{y}$ and draw from it

Semi-conjugate prior (μ and σ^2 independent)

- Semi-conjugate prior:

$$\begin{aligned} \mu|\sigma^2 &\sim N(\mu_0, \tau_0^2) \\ \sigma^2 &\sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2) \end{aligned}$$

  where, μ and σ^2 are now a priori independent

- The joint posterior of μ and σ^2 is no longer in closed form. However, we can specify it as

$$p(\mu, \sigma^2|y) = p(\mu|\sigma^2, y)p(\sigma^2|y)$$

- Conditional posterior of μ:

$$\mu|\sigma^2, y \sim N(\mu_n, \tau_n^2)$$

  where

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

  (as in the case with unknown mean and known variance)

- Marginal posterior of σ^2:

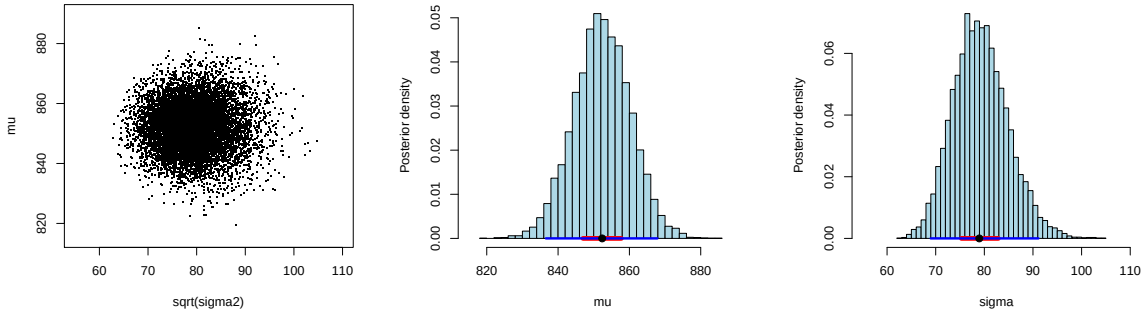
$$\begin{aligned} p(\sigma^2|y) &= \frac{p(\mu, \sigma^2|y)}{p(\mu|\sigma^2, y)} \\ &\propto \frac{p(\mu)p(\sigma^2)p(y|\mu, \sigma^2)}{p(\mu|\sigma^2, y)} \\ &= \frac{N(\mu|\mu_0, \tau_0^2) \cdot \text{Inv-}\chi^2(\sigma^2|\nu_0, \sigma_0^2) \cdot \prod_{i=1}^n N(y_i|\mu, \sigma^2)}{N(\mu|\mu_n, \tau_n^2)} \end{aligned}$$

  - can be evaluated on a grid of σ^2 values

  Note:

  - although RHS contains μ, LHS does not, so μ cancels out in the RHS
  - For computation, best to use $\mu = \mu_n$
  - Should remember that μ_n and τ_n^2 depend on σ: need to re-compute them *for each value* on the grid

Michelson data with semi-conjugate prior



Left: 10000 draws from the joint posterior of μ *and* σ*; centre and right: 10000 draws from the marginal posteriors of* μ *and* σ*. Semi-conjugate prior with* $\nu_0 = 1$, $\sigma_0 = 1$, $\mu_0 = 1$, $\tau_0 = 10000$.

Appendix: R code for the Michelson data example

```
library(MASS)
y <- michelson$Speed

hist(y, col="lightblue", xlab="Speed of light in km/s (minus 299000)",
     main = "Michelson data")
hist(y, 20, col="lightblue", xlab="Speed of light in km/s (minus 299000)",
     main = "Michelson data")

rinvchi2 <- function(n, nu, sigma2) {
  return(1/rgamma(n, shape=nu/2, rate=nu*sigma2/2))
}

histpost <- function(draws, ...)
{
  quantiles <- quantile(draws, probs=c(0.025, 0.25, 0.5, 0.75, 0.975))
  hist(draws, col="lightblue", ...)
  lines(quantiles[c(2,4)], rep(0, 2), lwd=6, col="red")
  lines(quantiles[c(1,5)], rep(0, 2), lwd=3, col="blue")
  points(quantiles[3], 0, pch=16, cex=1.2)
  return(invisible(quantiles))
}

#
# Conjugate prior
#

normalconj <- function(nsamp, y, nu0, sigma0, mu0, kappa0)
{
  n <- length(y);    ybar <- mean(y);     s2 <- var(y)
  kappan <- kappa0 + n
  mun <- (kappa0 * mu0 + n * ybar) / kappan
  nun <- nu0 + n
  sigman <- (nu0 * sigma0^2 + (n-1)*s2 +
             (kappa0 * n / kappan) * (ybar - mu0)^2 ) / nun
  sigman <- sqrt(sigman)
  sigma2.draws <- rinvchi2(nsamp, nun, sigman^2)
  mu.draws <- rnorm(nsamp, mun, sqrt(sigma2.draws/kappan))
  return(invisible(list(mu=mu.draws, sigma2=sigma2.draws)))
}

set.seed(123)
out <- normalconj(nsamp=10000, y=y, nu0=1, sigma0=1, mu0=1000, kappa0=1)
attach(out)
plot(sqrt(sigma2), mu, pch=".", xlim=c(55, 110), ylim=c(815, 890))
histpost(mu, 40, prob=TRUE, xlim=c(815, 890),xlab="mu",
         ylab="Posterior density", main="")
histpost(sqrt(sigma2), 40, xlim=c(55, 110), prob=TRUE, xlab="sigma",
         ylab="Posterior density", main="")
detach(out)

out <- normalconj(nsamp=10000, y=y, nu0=1, sigma0=1, mu0=1, kappa0=1e-6)
attach(out)
```

```
plot(sqrt(sigma2), mu, pch=".", xlim=c(55, 110), ylim=c(815, 890))
histpost(mu, 40, prob=TRUE, xlim=c(815, 890), xlab="mu",
         ylab="Posterior density", main="")
histpost(sqrt(sigma2), 40, xlim=c(55, 110), prob=TRUE, xlab="sigma",
         ylab="Posterior density", main="")
detach(out)

#
# Semi-conjugate prior
#

normalindep <- function(nsamp, y, nu0, sigma0, mu0, tau0,
                        lowsig2, hisig2, ngrid=200)
{
  n <- length(y);     ybar <- mean(y);      s2 <- var(y)
  sigma2.grd <- seq(lowsig2, hisig2, length=ngrid)
  taun2 <- 1 / ((1/tau0^2) + (n/sigma2.grd))
  mun <- ((mu0 /(tau0^2)) + (n/sigma2.grd)*ybar) * taun2
  mu <- mun
  lpostsig2 <- ( log(dnorm(mu, mu0, tau0)) - log(dnorm(mu, mun, sqrt(taun2)))
                - ((nu0/2) + 1) * log(sigma2.grd)
                - 0.5 * nu0 * sigma0^2 / sigma2.grd)
  for (i in (1:n))
    lpostsig2 <- lpostsig2 + log(dnorm(y[i], mu, sqrt(sigma2.grd)))

  draws <- sample(sigma2.grd, size=nsamp, replace=TRUE,
                  prob=exp(lpostsig2 - max(lpostsig2)))
  sig2inc <- sigma2.grd[2] - sigma2.grd[1]
  sigma2.draws <- draws + runif(nsamp, -sig2inc/2, sig2inc/2)

  taun2 <- 1 / ((1/tau0^2) + (n/sigma2.draws))
  mun <- ((mu0 /(tau0^2)) + (n/sigma2.draws)*ybar) * taun2
  mu.draws <- rnorm(nsamp, mun, sqrt(taun2))
  return(invisible(list(mu=mu.draws, sigma2=sigma2.draws)))
}

out <- normalindep(nsamp=10000, y, nu0=1, sigma0=1, mu0=1, tau0=10000,
                   lowsig2=2500, hisig2=12000, ngrid=200)

attach(out)
plot(sqrt(sigma2), mu, pch=".", xlim=c(55, 110), ylim=c(815, 890))
histpost(mu, 40, prob=TRUE, xlim=c(815, 890),xlab="mu",
         ylab="Posterior density", main="")
histpost(sqrt(sigma2), 40, xlim=c(55, 110), prob=TRUE, xlab="sigma",
         ylab="Posterior density", main="")
detach(out)
```

Lecture 8: Several binomial experiments (BDA2/3 §5.1–5.3)

8.1 The Model

$$y_j \sim \text{Bin}(n_j, \theta_j) \qquad j = 1, \ldots, J$$

- If we believe that θ_js are completely unrelated, we can do a separate analysis on each:
  - set $p(\theta_j)$, e.g. $\text{Be}(\alpha_j, \beta_j)$ with α_j, β_j fixed constants, possibly $\alpha_j = \alpha$, $\beta_j = \beta$, all j
  - obtain $p(\theta_j|y_j)$, as for J separate one binomial data problems
- Often the θ_js are related:
  - knowing one of the θs gives *some* information about the others

  Examples:
  - $\theta_j = \Pr$[thumbtack j landing point up]
    where thumbtacks $j = 1, \ldots, J$ are all from the same box
  - $\theta_j = \Pr$[drug j has a certain effect]
    where drugs $j = 1, \ldots, J$ belong to the same family of chemical compounds
  - $\theta_j =$ mortality rate due to a certain disease in city j
  - $\theta_j = \Pr$[tumor in lab rat j receiving a dose of a certain drug]

  Can regard θ_js as a sample from a common population of θs. This is related to what one does in random effects models.

Example: Rat tumor data

Rat tumor data: BDA2 p. 118; BDA3 p. 102:

- Current experiment: 4 rats with tumor / 14 total rats
- Results of other 70 preceding experiments are available

```
# Rat tumor data - BDA2 Chp. 5, p. 118, BDA3 p. 102
> y
 [1]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  1  1  1  1  1  1  1  3  2  2
[26]  2  2  2  2  2  2  2  1  5  2  5  2  7  7  3  3  2  9 10  4  4  4  4  4  4
[51]  4 10  4  4  4  5 11 12  5  5  6  5  6  6  6  6 16 15 15  9  4
> n
 [1] 20 20 20 20 20 20 20 19 19 19 19 18 18 17 20 20 20 20 19 19 18 18 27 25 24
[26] 23 20 20 20 20 20 20 10 49 19 46 17 49 47 20 20 13 48 50 20 20 20 20 20 20
[51] 20 48 19 19 19 22 46 49 20 20 23 19 22 20 20 20 52 46 47 24 14
```

We could analyse the 14 rats in current experiment, as though the preceding experiments hadn't happened (Fig. 5).

Can we use historical data to help improve inference about θ_{71}?

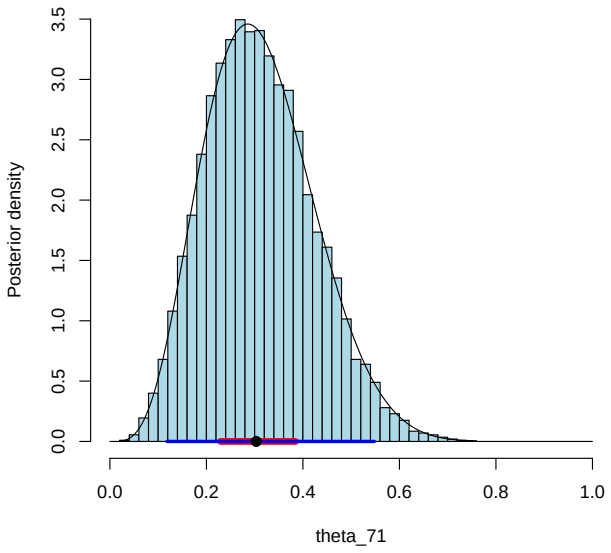


Figure 5: Rat tumor example: 10,000 draws from the posterior distribution of θ_{71} using the $\mathrm{Un}(0,1)$ prior

Empirical Bayes approach

Can form a $\mathrm{Be}(\alpha,\beta)$ prior on θ_{71} by matching its mean and variance to sample mean m and the sample variance v of the raw rates y_j/n_j, $j = 1,\ldots,70$:

$$m = \frac{\alpha}{\alpha+\beta} \qquad v = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

- Solving for α and β yields:

$$\alpha = \frac{(1-m)m^2}{v} - m \qquad \beta = \alpha\left(\frac{1}{m} - 1\right)$$

- With $m = 0.136$ and $\sqrt{v} = 0.103$, this yields

$$\alpha = 1.4 \qquad \beta = 8.6$$

- Posterior on RHS of Fig. 6.
- Called *Empirical Bayes* approach

Empirical Bayes posterior is

- shrunk towards overall mean
- less variable (we've learnt something from earlier experiments)

8.2 Bayesian model with prior on hyperparameters

Another way to account for the dependence between θs:

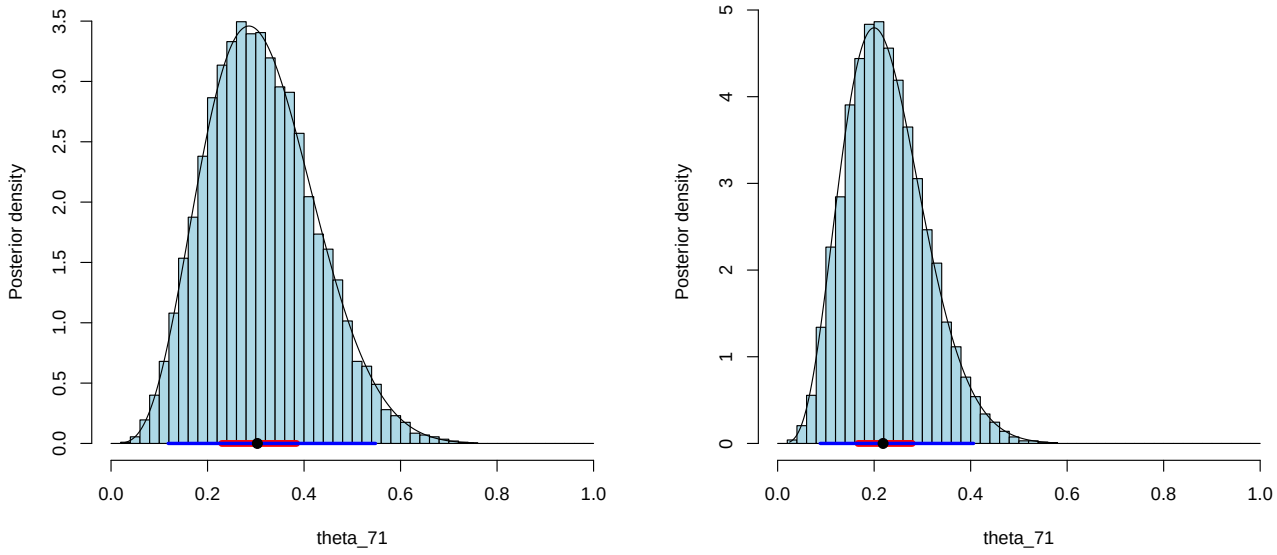


Figure 6: Rat tumor example. Left hand side: 10,000 draws from the posterior distribution of θ_{71} using a $\mathrm{Un}(0,1)$ prior. Right hand side: 10,000 draws from the posterior distribution of θ_{71} with a $\mathrm{Be}(1.4, 8.6)$ prior, obtained by matching the first two central moments to the mean and variance of raw death rates y_i/n_i, $i = 1, \ldots, 70$ (Empirical Bayes approach).

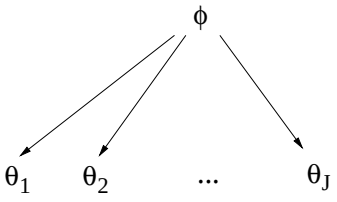
- assume they are conditionally independent given some hyperparameter ϕ (a parameter of the "population" of success rates):

$$\theta_j|\phi \overset{\text{i.i.d.}}{\sim} p(\theta_j|\phi) \qquad j = 1, \ldots, J$$

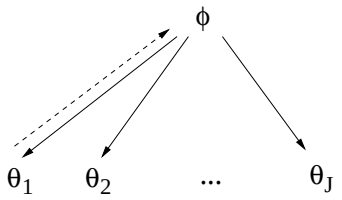
- *and* place a prior on ϕ too:

$$\phi \sim p(\phi)$$

- Graphically:



- If ϕ is known, θ_js are independent
- However, ϕ is unknown:
  - if we are told the value of θ_1, say, this will affect our beliefs about ϕ and, through it, our beliefs about the other θs

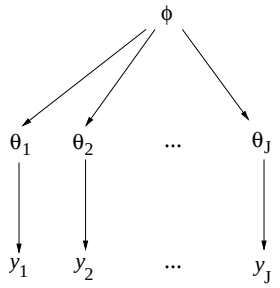


- This model is called *hierarchical* because the prior has two levels:
  (i) on the θs given ϕ, and (ii) on ϕ.

- Now the parameters consists of ϕ and all the θ_js (which I'll symbolize by θ). Posterior of interest is

$$\begin{aligned} p(\theta, \phi|y) &\propto p(\theta, \phi)p(y|\theta, \phi) \\ &= p(\phi)p(\theta|\phi)p(y|\theta) \\ &= p(\phi)\left[\prod_{j=1}^{J} p(\theta_j|\phi)p(y_j|\theta_j)\right] \end{aligned}$$

  since, given θ_j, y_j is independent of ϕ and the other θs:



- The marginal posterior of ϕ is

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

  Typically, difficult to compute!

  Can:

  - simulate both ϕ and θ from joint posterior. This requires a clever trick (coming later).
  - But for some models (with the distribution of $\theta_j|\phi$ conjugate to the distribution of $y_j|\theta_j$), can use

$$p(\phi|y) = \frac{p(\theta, \phi|y)}{p(\theta|\phi, y)} \tag{3}$$

    Next we use this approach for the binomial model with $Be(\alpha, \beta)$ prior and a hyperprior on $\phi = (\alpha, \beta)$

Back to the beta-binomial model!

$$\begin{aligned} y_j|\theta,\alpha,\beta &\sim \text{Bin}(n_j,\theta_j) \quad \text{independently} \\ \theta_j|\alpha,\beta &\sim \text{Be}(\alpha,\beta) \quad \text{independently} \\ (\alpha,\beta) &\sim p(\alpha,\beta) \end{aligned}$$

- Then, joint posterior is

$$\begin{aligned} p(\theta,\alpha,\beta|y) &\propto p(\alpha,\beta)p(\theta|\alpha,\beta)p(y|\theta,\alpha,\beta) \\ &\propto p(\alpha,\beta)\left[\prod_{j=1}^{J}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta_j^{\alpha-1}(1-\theta_j)^{\beta-1}\right]\times \\ &\qquad \times\left[\prod_{j=1}^{J}\theta_j^{y_j}(1-\theta_j)^{n_j-y_j}\right] \end{aligned} \tag{4}$$

- Conditionally on (α,β), the θ_js are a posteriori independent:

$$\theta_j|\alpha,\beta,\theta_{-j},y \sim \text{Be}(\alpha+y_j,\beta+n_j-y_j)$$

so

$$p(\theta|\alpha,\beta,y)=\prod_{j=1}^{J}\frac{\Gamma(\alpha+\beta+n_j)}{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)}\theta_j^{\alpha+y_j-1}(1-\theta_j)^{\beta+n_j-y_j-1} \tag{5}$$

- Substituting (4) and (5) in (3) with $\phi=(\alpha,\beta)$, one obtains

$$\begin{aligned} p(\alpha,\beta|y) &= \frac{p(\theta,\alpha,\beta|y)}{p(\theta|\alpha,\beta,y)} \\ &\propto p(\alpha,\beta)\prod_{j=1}^{J}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)}{\Gamma(\alpha+\beta+n_j)} \end{aligned}$$

  - Once $p(\alpha,\beta)$ is chosen, this can be evaluated on a fine grid of (α,β)
  - Then normalize it and sample from it, to obtain a sample of (α,β)s

- Then, sampling from $p(\theta,\alpha,\beta|y)$ can be done in stages, as follows:
  - Obtain a sample of (α,β) pairs from $p(\alpha,\beta|y)$ as just explained
  - plug these draws in the posteriors $\theta_j|\alpha,\beta,y \sim \text{Be}(\alpha+y_j,\beta+n_j-y_j)$
  - sample from these distributions, obtaining draws from marginal posteriors of θ_js
  - This works because

$$\begin{aligned} p(\theta,\phi|y) &= p(\theta|\phi,y)p(\phi|y) \\ \text{(A)} &= \text{(B)} \quad \text{(C)} \end{aligned}$$

    To sample from (A): sample from (C), plug in (B) and sample from it.
    Draws from (A) are (θ,ϕ) pairs. If only keep θs, these are draws from $p(\theta|y)$

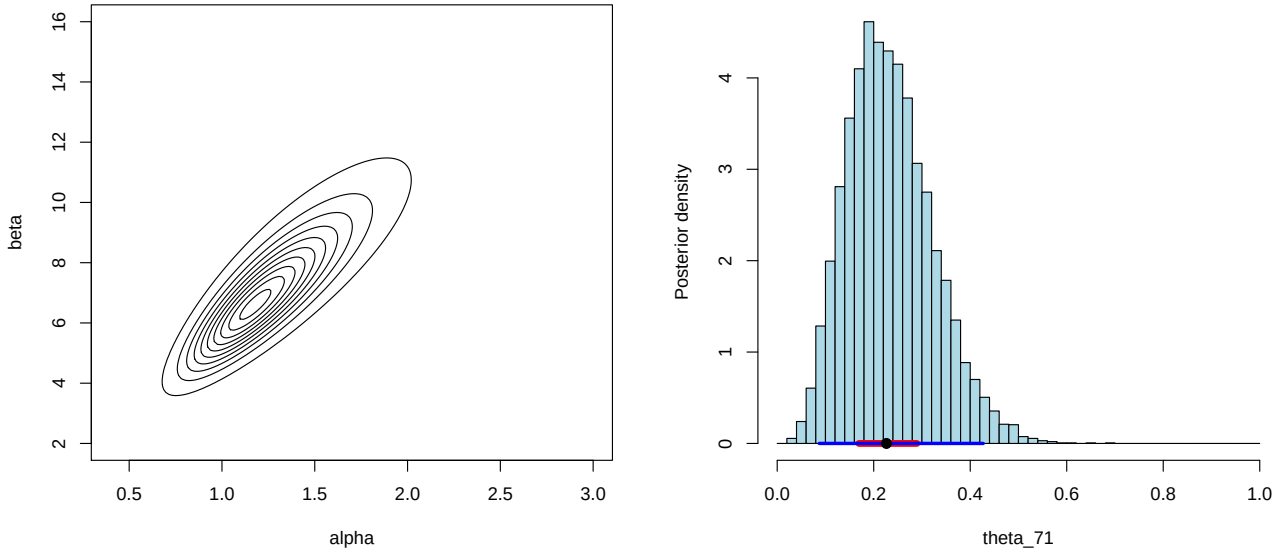


Figure 7: Left hand side: contour plot of the joint posterior distribution of the hyper-parameters $p(\alpha, \beta|y)$, evaluated on a grid, using independent $Exp(1)$ priors on α and β. Right: histogram of 10,000 draws from the marginal posterior distribution of θ_{71}, using the Bayesian hierarchical model with $Exp(1)$ priors on α and β.

Back to Rat tumor example

- As prior on (α, β), decided to use

$$\alpha \sim Exp(\lambda) \qquad \beta \sim Exp(\lambda) \qquad \lambda = 1 \qquad \text{independently}$$

  i.e.

$$p(\alpha, \beta) = \exp\{-(\alpha + \beta)\}$$

- Seemed to make intuitive sense, by recalling interpretation of α, β as "prior numbers of successes and failures"

- BDA prefer to use

$$p(\alpha, \beta) \propto (\alpha + \beta)^{-5/2}.$$

  See BDA2 p. 128 or BDA3 p. 110-111 for a discussion of this *diffuse* prior

- Fig. 7 shows posterior contours of (α, β) and posterior density of θ_{71}.

Lecture 9: Non-informative priors (BDA2 §2.9; BDA3 §2.10)

9.1 Non-informative priors

Bayesian methods require the specification of a prior distribution

- Since
$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$
the prior, in principle, can be influential
- Mainly a problem with θ of medium/high dimension
  - prior *may* be very informative about some function of θ (even if it's not clearly informative about each individual component of θ)
- Hence the search for priors that play a minimal role, that *let the data speak for themselves*
  - Called: vague, flat, diffuse, reference, non-informative priors

9.2 Proper and improper priors

- $p(\theta)$ is a proper prior if it integrates to 1
- if $\int p(\theta)\,d\theta = k < \infty$, can renormalize, dividing by k
- if $\int p(\theta)\,d\theta$ is not finite we say that $p(\theta)$ is an improper prior
- improper priors can be used, as long as the resulting *posterior* is proper

Example: Binomial data with conjugate Beta prior

$$y|\theta \sim \text{Bin}(n, \theta) \qquad \theta \sim \text{Be}(\alpha, \beta)$$

- recall interpretation of α and β as *prior numbers of successes and failures*
- can think of using $\alpha = \beta = 0$ to be non-informative
- This corresponds to $p(\theta) \propto \theta^{-1}(1-\theta)^{-1}$, whose integral on $(0, 1)$ is not defined:
$$\begin{aligned}\int_0^1 \frac{1}{\theta(1-\theta)}\,d\theta &= \int_0^{1/2} \frac{1}{\theta(1-\theta)}\,d\theta + \int_{1/2}^1 \frac{1}{\theta(1-\theta)}\,d\theta \\ &> \int_0^{1/2} \frac{1}{\theta}\,d\theta + \int_{1/2}^1 \frac{1}{1-\theta}\,d\theta \\ &= 2\int_0^{1/2} \frac{1}{\theta}\,d\theta = 2\,[\log\theta]_0^{1/2}\end{aligned}$$
- So, $\text{Be}(0, 0)$ is an improper prior.
- Note however, that letting $\alpha = \beta = 0$ in the $\text{Be}(\alpha+y, \beta+n-y)$ posterior results in a posterior $\text{Be}(y, n-y)$ which is well defined, as long as $y > 0$ and $n - y > 0$.

Safe approach:

1. Obtain the posterior under a proper (conjugate) prior

2. Let hyperparameters tend to the values corresponding to the improper prior ($\alpha = \beta = 0$ in previous example)

3. If resulting posterior is proper, can use the improper prior

Example: Normal data with known mean and unknown variance

$$y_i|\sigma^2 \sim N(\theta, \sigma^2) \qquad i = 1, \ldots, n \qquad \text{independently}$$

$$\sigma^2 \sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2) \qquad \sigma^2|y \sim \text{Inv-}\chi^2\left(\nu_0 + n, \frac{\nu_0\sigma_0^2 + n\nu}{\nu_0 + n}\right)$$

- ν_0 is the number of "equivalent observations" in the prior
- can think of letting $\nu_0 \to 0$, to be non-informative:

$$\begin{aligned} \text{Inv-}\chi^2(\nu_0, \sigma_0^2) &= \text{Inv-gamma}(\nu_0/2, \nu_0\sigma_0^2/2) \\ &\to \text{Inv-gamma}(0, 0) \qquad \text{as } \nu_0 \to 0 \end{aligned}$$

  which has p.d.f. $p(\sigma^2) \propto (\sigma^2)^{-1}$, with no finite integral over $(0, \infty)$

- So, Inv-$\chi^2(\nu_0, \sigma_0^2)$ with $\nu_0 = 0$ is an improper prior
- However, letting $\nu_0 = 0$ in the formulae for the parameters of the Inv-χ^2 posterior of σ^2 yields the Inv-$\chi^2(n, \nu)$ distribution, which is proper.

Note:
In the previous examples an improper prior resulted in a proper posterior

- This is *not* always the case
- An improper prior carries for the user the burden of *proving* that the posterior is proper

9.3 Jeffreys' principle

- If we "know nothing" about θ, then we "know nothing" about any 1-1 function $\phi = g(\theta)$
- Hence, a formal rule to assign non-informative priors should lead to the same recipe for θ and for ϕ, call these $p(\theta)$ and $p(\phi)$.
- On the other hand, it must be that

$$p(\phi) = p(\theta)\left|\frac{d\theta}{d\phi}\right|$$

- These considerations lead to Jeffreys' principle:

$$p(\theta) = [J(\theta)]^{1/2}$$

  where $J(\theta)$ is Fisher's information:

$$J(\theta) = E\left[\left\{\frac{d}{d\theta}\log p(y|\theta)\right\}^2 \middle| \theta\right] = -E\left[\frac{d^2}{d\theta^2}\log p(y|\theta)\middle| \theta\right]$$

- **Lemma**: Let $\phi = h(\theta)$ be a 1-1, continuously differentiable transformation of θ. Then

$$J(\phi) = J(\theta)\left|\frac{d\theta}{d\phi}\right|^2$$

- **Proof**

$$\log p(y|\phi) = \log p(y|\theta) \qquad \theta = h^{-1}(\phi)$$

Differentiating with respect to ϕ gives

$$\frac{d}{d\phi}\log p(y|\phi) = \frac{d}{d\theta}\log p(y|\theta)\cdot\frac{d\theta}{d\phi}$$

Now square both sides and take expectation w.r.t. y (conditional on ϕ on lhs, conditional on $\theta = h^{-1}(\phi)$ on rhs)

$$E\left[\left(\frac{d}{d\phi}\log p(y|\phi)\right)^2\middle|\phi\right] = E\left[\left(\frac{d}{d\theta}\log p(y|\theta)\right)^2\middle|\theta\right]\left[\frac{d\theta}{d\phi}\right]^2$$

□

- If we use the rule $p(\theta) \propto \sqrt{J(\theta)}$, then the change of variable formula and the Lemma above yield

$$\begin{aligned} p(\phi) &= p(\theta)\left|\frac{d\theta}{d\phi}\right| \\ &\propto \sqrt{J(\theta)}\left|\frac{d\theta}{d\phi}\right| \\ &= \sqrt{J(\phi)}\left|\frac{d\theta}{d\phi}\right|^{-1}\left|\frac{d\theta}{d\phi}\right| \\ &= \sqrt{J(\phi)} \end{aligned}$$

So, the same rule $p(\phi) \propto \sqrt{J(\phi)}$ applies to ϕ too

Example: Normal data – Jeffreys' prior for σ^2

Normal data with known mean θ and unknown variance σ^2. For simplicity, let $\psi = \sigma^2$. Then

$$\begin{aligned} p(y|\psi) &\propto \psi^{-n/2}\exp\left\{-\frac{1}{2\psi}\sum_i (y_i-\theta)^2\right\} \\ \log p(y|\psi) &= \text{const.} - \frac{n}{2}\log\psi - \frac{1}{2\psi}\sum_i (y_i-\theta)^2 \\ \frac{d}{d\psi}\log p(y|\psi) &= -\frac{n}{2}\frac{1}{\psi} + \frac{1}{2}\sum_i (y_i-\theta)^2\frac{1}{\psi^2} \\ \frac{d^2}{d\psi^2}\log p(y|\psi) &= \frac{n}{2}\frac{1}{\psi^2} - \sum_i (y_i-\theta)^2\frac{1}{\psi^3} \\ E\left[\frac{d^2}{d\psi^2}\log p(y|\psi)\middle|\psi\right] &= \frac{n}{2}\frac{1}{\psi^2} - \frac{1}{\psi^3}n\psi \;=\; -\frac{n}{2}\frac{1}{\psi^2} \end{aligned}$$

- Therefore, $J(\psi) = \frac{n}{2}\frac{1}{\psi^2}$, and since $\psi = \sigma^2$, $J(\sigma^2) = \frac{n}{2}\frac{1}{\sigma^4}$.

- Hence the Jeffreys' prior is

$$p(\sigma^2) \propto \frac{1}{\sigma^2}$$

Homework 3

Problem 1

In the case of Normal data with known variance σ^2, one can try to be non-informative by letting the prior variance $\tau_0^2 \to \infty$. What is the corresponding form of the posterior distribution of the mean θ? Is it a proper posterior?

Problem 2

Find the Jeffreys' prior for the mean θ in the case of Normal data with known variance.

Problem 3

Find the Jeffreys' prior for the probability of success θ in the case of Binomial data.

Lecture 10: A Normal hierarchical model (BDA2/3 §5.4, 5.5, BDA2 §11.7, BDA3 §11.6; FCBSM §8.3)

J independent experiments: normal observations with known variance σ^2

$$y_{ij}|\theta \sim N(\theta_j, \sigma^2) \qquad i = 1, \ldots, n_j \qquad j = 1, \ldots, J$$

- Aim: estimate the means θ_js
  - Could use unpooled sample means
  $$\overline{y}_{\cdot j} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$$
  - If n_js are very small and/or θ_js are very similar, could use pooled estimate
  $$\overline{y}_{\cdot\cdot} = \frac{\sum_{j=1}^{J} n_j \overline{y}_{\cdot j}}{\sum_{j=1}^{J} n_j}$$
  - Classical approach: decide which to use based on the result of the ANOVA F-test of $H_0 : \theta_1 = \cdots = \theta_J$.
  - Bayesian hierarchical model: estimates of θ_js as weighted combinations of pooled and unpooled estimates

    sample means $\overline{y}_{\cdot j}$s are shrunk towards overall mean $\overline{y}_{\cdot\cdot}$

10.1 A Normal hierarchical model

Let's complete the model, by adding a two-stage prior distribution on the means θ_j:

$$\begin{aligned}
\overline{y}_{\cdot j}|\theta &\sim N(\theta_j, \sigma_j^2) \qquad \sigma_j^2 = \frac{\sigma^2}{n_j} \text{ known} \qquad j = 1, \ldots, J \qquad \text{indep.} \\
\theta_j|\mu, \tau^2 &\sim N(\mu, \tau^2) \qquad j = 1, \ldots, J \qquad \text{indep.} \\
p(\mu|\tau^2) &\propto 1 \\
p(\tau) &\propto 1 \qquad [\Rightarrow p(\tau^2) \propto \tau^{-1} = (\tau^2)^{-1/2}]
\end{aligned}$$

- Cannot use improper prior $p(\tau^2) \propto \frac{1}{\tau^2}$: it yields an improper posterior
- $p(\tau^2) \propto \frac{1}{\tau}$ is fine, as long as $J \geqslant 3$ (BDA2 Exercise 5.8; BDA3 Exercise 5.10)

Joint posterior:

$$p(\mu, \tau^2, \theta|y) \propto (\tau^2)^{-1/2} \prod_{j=1}^{J} N(\theta_j|\mu, \tau^2) N(\overline{y}_{\cdot j}|\theta_j, \sigma_j^2)$$

Gibbs Sampling

We will use Gibbs sampling to produce a sample from the joint posterior

- More discussion of Gibbs sampling later in the course

- For this problem the algorithm is:

  1. Set initial values $\mu_{(0)}$, $\tau^2_{(0)}$, $\theta_{(0)}$
  2. For t in (1:N) {
     generate $\mu_{(t)}$ from $p(\mu|\tau^2_{(t-1)}, \theta_{(t-1)}, y)$
     generate $\tau^2_{(t)}$ from $p(\tau^2|\mu_{(t)}, \theta_{(t-1)}, y)$
     generate $\theta_{(t)}$ from $p(\theta|\mu_{(t)}, \tau^2_{(t)}, y)$
     }

- This procedure defines a Markov chain
- After some burn-in period (discarded), the vectors $(\mu_{(t)}, \tau^2_{(t)}, \theta_{(t)})$ can be regarded (approximately) as draws from the joint posterior.
- WinBUGS, formerly BUGS: Bayesian Analysis Using Gibbs Sampling

Full conditional distributions

To run the Gibbs sampler we need the distribution of each variable conditional on all other variables: *full conditional distributions*.

Begin by rewriting the joint density of all variables

$$\begin{aligned} p(\mu, \tau^2, \theta, y) &\propto p(\tau^2)p(\mu|\tau)p(\theta|\mu, \tau)p(y|\mu, \tau, \theta) \\ &\propto (\tau^2)^{-1/2} \prod_{j=1}^{J} \left[(2\pi\tau^2)^{-1/2} \exp\left\{ -\frac{1}{2}\frac{(\theta_j - \mu)^2}{\tau^2} \right\} \times \right. \\ &\qquad \left. \times (2\pi\sigma_j^2)^{-1/2} \exp\left\{ -\frac{1}{2}\frac{(\overline{y}_{\cdot j} - \theta_j)^2}{\sigma_j^2} \right\} \right] \\ &\propto (\tau^2)^{-(J+1)/2} \exp\left\{ -\frac{1}{2\tau^2}\sum_{j=1}^{J}(\theta_j - \mu)^2 - \frac{1}{2}\sum_{j=1}^{J}\frac{(\overline{y}_{\cdot j} - \theta_j)^2}{\sigma_j^2} \right\} \end{aligned}$$

To obtain the full conditionals of each variable, simply keep the terms in the joint density that involve that variable.

Full conditional of μ

$$\begin{aligned} p(\mu|\tau^2, \theta, y) &\propto p(\mu, \tau^2, \theta, y) \\ &\propto \exp\left\{ -\frac{1}{2\tau^2}\sum_{j=1}^{J}(\theta_j - \mu)^2 \right\} \\ &= \exp\left\{ -\frac{1}{2\tau^2}\left(\sum_j \theta_j^2 + J\mu^2 - 2\mu J\hat{\mu} \right) \right\}, \qquad \hat{\mu} = \frac{1}{J}\sum_{j=1}^{J}\theta_j \\ &\propto \exp\left\{ -\frac{1}{2\tau^2}\left(J\mu^2 - 2\mu J\hat{\mu} \right) \right\} \\ &\propto \exp\left\{ -\frac{J}{2\tau^2}(\mu - \hat{\mu})^2 \right\} \\ &\propto N\left(\mu \middle| \hat{\mu}, \frac{\tau^2}{J} \right) \end{aligned}$$

Full conditional of τ^2

$$\begin{aligned}
p(\tau^2|\mu,\theta,y) &\propto p(\mu,\tau^2,\theta,y) \\
&\propto (\tau^2)^{-(J+1)/2}\exp\left\{-\frac{1}{2\tau^2}\sum_{j=1}^{J}(\theta_j-\mu)^2\right\} \\
&\propto \text{Inv-gamma}\left(\tau^2\left|\frac{J-1}{2},\frac{1}{2}\sum_{j=1}^{J}(\theta_j-\mu)^2\right.\right) \\
&= \text{Inv-}\chi^2\left(\tau^2\left|J-1,\frac{1}{J-1}\sum_{j=1}^{J}(\theta_j-\mu)^2\right.\right)
\end{aligned}$$

Full conditional of θ

$$\begin{aligned}
p(\theta|\mu,\tau^2,y) &\propto p(\mu,\tau^2,\theta,y) \\
&\propto \exp\left\{-\frac{1}{2\tau^2}\sum_{j=1}^{J}(\theta_j-\mu)^2-\frac{1}{2}\sum_{j=1}^{J}\frac{(\overline{y}_{\cdot j}-\theta_j)^2}{\sigma_j^2}\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\sum_{j=1}^{J}\left(\frac{\theta_j^2+\mu^2-2\mu\theta_j}{\tau^2}+\frac{\overline{y}_{\cdot j}^2+\theta_j^2-2\theta_j\overline{y}_{\cdot j}}{\sigma_j^2}\right)\right]\right\} \\
&\propto \prod_{j=1}^{J}\exp\left\{-\frac{1}{2}\left[\frac{\theta_j^2-2\mu\theta_j}{\tau^2}+\frac{\theta_j^2-2\theta_j\overline{y}_{\cdot j}}{\sigma_j^2}\right]\right\}
\end{aligned}$$

Now the expression in square brackets is

$$\begin{aligned}
[\,\cdot\,] &= \frac{1}{\tau^2\sigma_j^2}\left[\left(\sigma_j^2+\tau^2\right)\theta_j^2-2\left(\mu\sigma_j^2+\overline{y}_{\cdot j}\tau^2\right)\theta_j\right] \\
&= \frac{\sigma_j^2+\tau^2}{\tau^2\sigma_j^2}\left[\theta_j^2-2\,\frac{\mu\sigma_j^2+\overline{y}_{\cdot j}\tau^2}{\sigma_j^2+\tau^2}\theta_j\right] \\
&= \frac{1}{V_{\theta_j}}\left[\theta_j^2-2\hat{\theta}_j\theta_j\right]
\end{aligned}$$

where

$$\begin{aligned}
V_{\theta_j}^{-1} &= \frac{\sigma_j^2+\tau^2}{\tau^2\sigma_j^2}=\frac{1}{\tau^2}+\frac{1}{\sigma_j^2} \\
\hat{\theta}_j &= \frac{\mu\sigma_j^2+\overline{y}_{\cdot j}\tau^2}{\sigma_j^2+\tau^2}=\frac{\mu\sigma_j^2+\overline{y}_{\cdot j}\tau^2}{\tau^2\sigma_j^2\left(\frac{1}{\tau^2}+\frac{1}{\sigma_j^2}\right)}=\frac{\mu/\tau^2+\overline{y}_{\cdot j}/\sigma_j^2}{1/\tau^2+1/\sigma_j^2}
\end{aligned}$$

Therefore

$$\begin{aligned}
p(\theta|\mu,\tau^2,y) &\propto \prod_{j=1}^{J}\exp\left\{-\frac{1}{2}\left[\frac{1}{V_{\theta_j}}\left(\theta_j^2-2\hat{\theta}_j\theta_j\right)\right]\right\} \\
&\propto \prod_{j=1}^{J}\exp\left\{-\frac{1}{2V_{\theta_j}}(\theta_j-\hat{\theta}_j)^2\right\} \\
&\propto \prod_{j=1}^{J}N(\theta_j|\hat{\theta}_j,V_{\theta_j})
\end{aligned}$$

i.e. the components of θ have independent normal full conditional distributions:

$$\theta_j|\mu, \tau^2, \theta_{-j}, y \sim N(\theta_j|\hat{\theta}_j, V_{\theta_j})$$

Notes

Can easily make changes to use a proper prior on (μ, τ^2)

- if $\tau^2 \sim$ Inv-χ^2 and $\mu|\tau^2 \sim N$, full conditionals of τ^2 and μ remain Inv-χ^2 and N (FCBSM §8.3).
- e.g. $\tau^2 \sim$ Inv-$\chi^2(\nu, \sigma_\tau^2)$ results in

$$p(\tau^2|\mu, \theta, y) = \text{Inv-}\chi^2\left(\tau^2 \left| J + \nu, \frac{\sum_{j=1}^{J}(\theta_j - \mu)^2 + \nu\sigma_\tau^2}{J + \nu}\right.\right)$$

Can easily allow for unknown σ^2

- with an Inv-χ^2 or even an improper $\propto 1/\sigma^2$ prior, full conditional of σ^2 is Inv-χ^2 (BDA2 p. 301, BDA3 p. 289; FCBSM p. 135)

Example

Read BDA2 §5.5 pp. 138–145 or BDA3 §5.5 pp. 119–124. Part of next week's lab will be devoted to that example.

Lecture 11: The Poisson and Exponential Models (BDA2 §2.7, BDA3 §2.6; FCBSM §3.2)

11.1 The Poisson model

Arises in the study of *counts* data

- e.g. in epidemiological studies of the incidence of diseases

$$y_i|\theta \overset{\text{i.i.d.}}{\sim} \text{Poi}(\theta) \qquad i = 1, \ldots, n$$

Likelihood, Prior, Posterior

- Likelihood:

$$\begin{aligned} p(y|\theta) &= \prod_{i=1}^{n} \frac{\theta^{y_i}}{y_i!} e^{-\theta} \\ &\propto \theta^{\sum_{i=1}^{n} y_i} e^{-n\theta} \end{aligned}$$

  - proportional to a Gamma density: it suggests that the natural conjugate prior is Gamma

- With a $\text{Ga}(\alpha, \beta)$ prior

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

  the posterior is

$$\begin{aligned} p(\theta|y) &\propto p(\theta)p(y|\theta) \\ &\propto \theta^{\alpha+\sum_i y_i - 1} e^{-(\beta+n)\theta} \\ &\propto \text{Ga}\left(\theta \,\middle|\, \alpha + \sum_{i=1}^{n} y_i, \beta + n\right) \end{aligned}$$

- Interpretation of hyperparameters α, β:

  - information in the prior is equivalent to that in β observations with total count equal to $\alpha - 1$

Posterior predictive of new observation $\tilde{y}$

- The prior predictive (also called *marginal likelihood*) is

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

- Similarly, the posterior predictive of $\tilde{y}$ is

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

  Now

  - $p(\tilde{y}|\theta, y) = p(\tilde{y}|\theta) = \text{Poi}(\tilde{y}|\theta)$

- $p(\theta|y) = \text{Ga}(\theta|\alpha + n\bar{y}, \beta + n)$
- $p(\theta|\tilde{y}, y) = \text{Ga}(\theta|\alpha + n\bar{y} + \tilde{y}, \beta + n + 1)$

- Hence

$$\begin{aligned} p(\tilde{y}|y) &= \frac{\text{Poi}(\tilde{y}|\theta)\text{Ga}(\theta|\alpha + n\bar{y}, \beta + n)}{\text{Ga}(\theta|\alpha + n\bar{y} + \tilde{y}, \beta + n + 1)} \\ &= \frac{\frac{\theta^{\tilde{y}} e^{-\theta}}{\tilde{y}!} \frac{(\beta+n)^{\alpha+n\bar{y}}}{\Gamma(\alpha+n\bar{y})} \theta^{\alpha+n\bar{y}-1} e^{-(\beta+n)\theta}}{\frac{(\beta+n+1)^{\alpha+n\bar{y}+\tilde{y}}}{\Gamma(\alpha+n\bar{y}+\tilde{y})} \theta^{\alpha+n\bar{y}+\tilde{y}-1} e^{-(\beta+n+1)\theta}} \\ &= \frac{\Gamma(\alpha+n\bar{y}+\tilde{y})}{\Gamma(\tilde{y}+1)\Gamma(\alpha+n\bar{y})} \left(\frac{\beta+n}{\beta+n+1}\right)^{\alpha+n\bar{y}} \left(\frac{1}{\beta+n+1}\right)^{\tilde{y}} \end{aligned}$$

  Therefore, assuming that α is integer,

$$\alpha + n\bar{y} + \tilde{y}|y \sim \text{Neg-Bin}\left(\alpha + n\bar{y}, \frac{1}{\beta+n+1}\right)$$

- Prior predictive of a single observation:
  - obtained from $p(\tilde{y}|y)$ by letting $n = 0$, $\bar{y} = 0$ and $\tilde{y} = y$:

$$\alpha + y \sim \text{Neg-Bin}(\alpha, 1/(\beta+1))$$

- Note that BDA uses a different parameterisation of the negative binomial distribution

An extension of the Poisson model

Suppose that y_i is the number of cases of a rare disease in city i.

- Can regard y_is as Poisson r.v.s, but should account, at very least, for different size of cities
- Standard assumption:

$$y_i|\theta, x_i \sim \text{Poi}(x_i\theta)$$

  - x_i: known explanatory variable (e.g. population of city i)—this is called the *exposure*
  - θ: unknown *rate* parameter

- Likelihood

$$\begin{aligned} p(y|\theta, x) &= \prod_{i=1}^{n} \frac{(x_i\theta)^{y_i}}{y_i!} e^{-x_i\theta} \\ &\propto \theta^{\sum_i y_i} e^{-\theta\sum_i x_i} \end{aligned}$$

- With a $\text{Ga}(\alpha, \beta)$ prior, the posterior is

$$\begin{aligned} p(\theta|y, x) &\propto \theta^{\alpha-1} e^{-\beta\theta} \theta^{\sum_i y_i} e^{-\theta\sum_i x_i} \\ &\propto \text{Ga}\left(\theta \middle| \alpha + \sum_i y_i, \beta + \sum_i x_i\right) \end{aligned}$$

11.2 Exponential model

Used to model 'waiting times' until some event

$$y_i|\theta \overset{\text{i.i.d.}}{\sim} \text{Exp}(\theta) \qquad i = 1, \ldots, n$$

Likelihood, Prior, Posterior

- Likelihood

$$\begin{aligned} p(y|\theta) &= \prod_{i=1}^{n} \theta e^{-\theta y_i} \\ &= \theta^n e^{-\theta \sum_i y_i} \end{aligned}$$

  - suggests Gamma is the natural conjugate family

- With a $\text{Ga}(\alpha, \beta)$ prior, the posterior is

$$\begin{aligned} p(\theta|y) &\propto \theta^{\alpha-1} e^{-\beta\theta} \cdot \theta^n e^{-\theta \sum_i y_i} \\ &= \theta^{\alpha-1+n} e^{-(\beta+\sum_i y_i)\theta} \\ &\propto \text{Ga}\left(\theta \,\middle|\, \alpha + n, \beta + \sum_{i=1}^{n} y_i\right) \end{aligned}$$

- Interpretation of hyperparameters α, β:

  - information in the prior is equivalent to that in $\alpha - 1$ observations with a total waiting time of β

Example: Leukaemia data

- Times of remission (in weeks) of leukaemia patients

Sample 0 (drug 6-MP)	6*, 6, 6, 6, 7, 9*, 10*, 10, 11*, 13, 16, 17*, 19*, 20*, 22, 23, 25*, 32*, 32*, 34*, 35*
Sample 1 (control)	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

- Data set of Gehan (1965), as reported in Cox and Oakes (1984)
- Half the patients were randomly allocated to be treated with the drug 6-mercaptopurine, the other half served as controls
- Asterisks denote censored remission times
- For the time being, we concentrate on the data in Sample 1 (control)

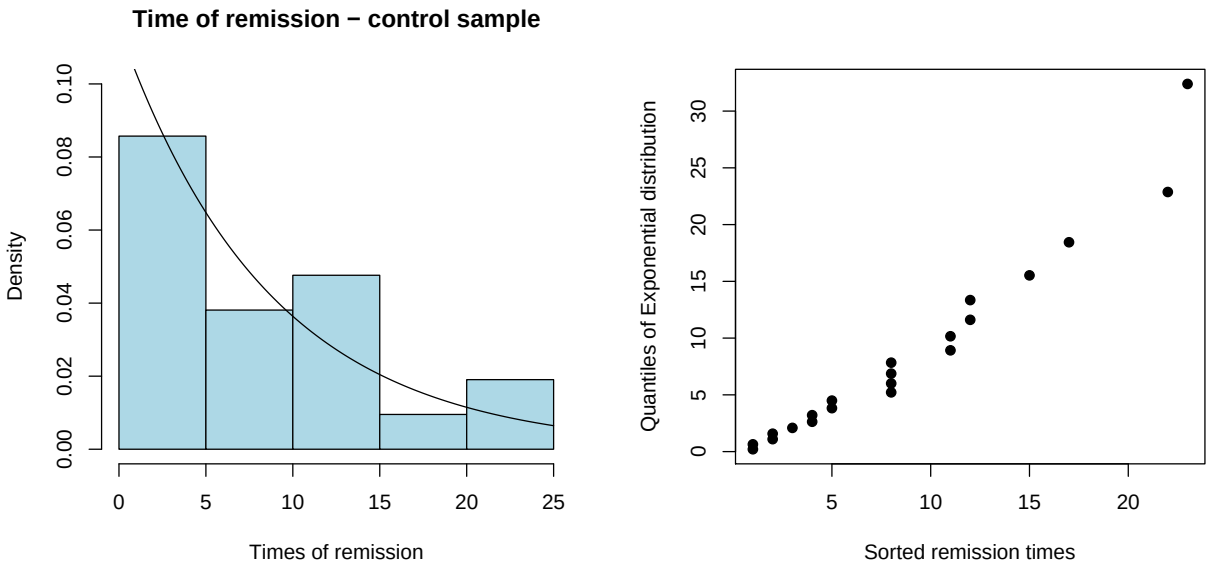


Figure 8: Leukaemia data, Control sample. LHS: Histogram of the data, with superimposed the Exp(1/8.67) density. RHS: Quantile-quantile plot.

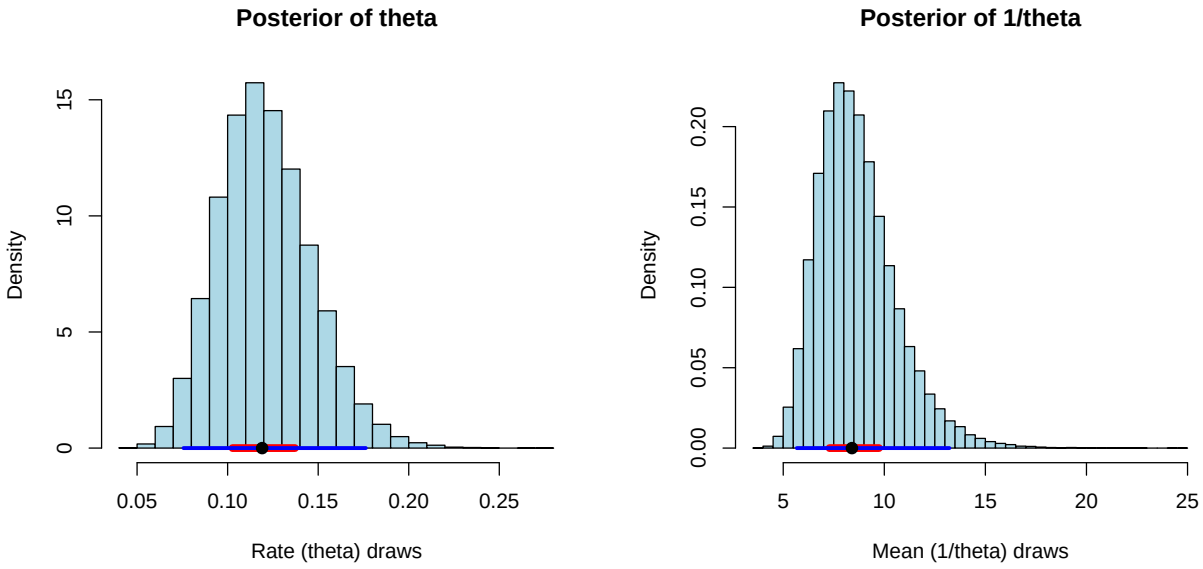


Figure 9: Leukaemia data, Control sample. LHS: Histogram of 100,000 draws from the $\mathrm{Ga}(1+n, \sum_i y_i)$ posterior distribution of the rate parameter θ_C, with a non-informative uniform prior ($\mathrm{Ga}(1, 0)$). RHS: Histogram of 100,000 draws from the posterior distribution of the mean $1/\theta_C$.

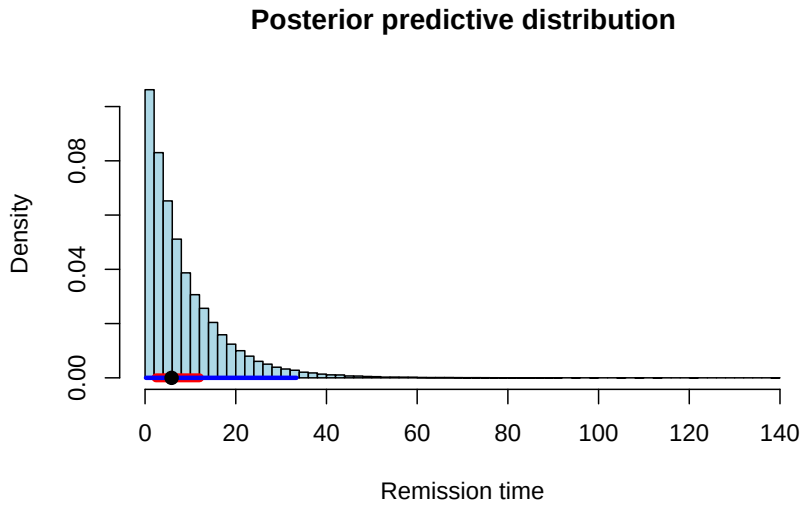


Figure 10: Histogram of 100,000 draws from the posterior predictive distribution of a future remission time in the control group. Can you think of some unsatisfactory feature of this plot?

Appendix: R code for the Leukaemia data example

```
remistime <- c(1,1,2,2,3,4,4,5,5,8,8,8,8,11,11,12,12,15,17,22,23)

hist(remistime, xlab="Times of remission", col="lightblue",
     main="Time of remission - control sample",
     prob=TRUE, ylim=c(0, 0.1))
time <- seq(0.5, 25, length=200)
lines(time, dexp(time, 1/mean(remistime)))

quantiles <- qexp(seq(0.5/21, 20.5/21, by=1/21), 1/mean(remistime))
plot(sort(remistime), quantiles, pch=16,
     xlab="Sorted remission times",
     ylab="Quantiles of Exponential distribution")

n <- length(remistime)
sumy <- sum(remistime)
alpha <- 1
beta <- 0
nsamp <- 100000
set.seed(321)

theta.draws <- rgamma(nsamp, alpha + n, beta + sumy)

histpost <- function(draws, ...)
{
  quantiles <- quantile(draws, probs=c(0.025, 0.25, 0.5, 0.75, 0.975))
  hist(draws, col="lightblue", ...)
  lines(quantiles[c(2,4)], rep(0, 2), lwd=6, col="red")
  lines(quantiles[c(1,5)], rep(0, 2), lwd=3, col="blue")
  points(quantiles[3], 0, pch=16, cex=1.2)
```

```
  return(invisible(quantiles))
}

# histogram of draws from posterior of theta
histpost(theta.draws, breaks=30,
         prob=TRUE, main="Posterior of theta",
         xlab="Rate (theta) draws", ylab="Density")

# histogram of draws from posterior of 1/theta
histpost(1/theta.draws, breaks=30, prob=TRUE, main="Posterior of 1/theta",
         xlab="Mean (1/theta) draws", ylab="Density")

# histogram of draws from the posterior predictive
postpred.draws <- rexp(nsamp, theta.draws)
histpost(postpred.draws, breaks=seq(0, max(postpred.draws)+2, by=2),
         prob=TRUE, main="Posterior predictive distribution",
         xlab="Remission time", ylab="Density")
```

Homework 4

Problem 1 (BDA2, p. 70, BDA3, p. 59, Exercise 2.11.13)

Worldwide airline fatalities, 1976–1985. Death rate is passenger deaths per 100 million passenger miles. Source: Statistical Abstract of the United States.

Year	Fatal accidents	Passenger deaths	Death rate
1976	24	734	0.19
1977	25	516	0.12
1978	31	754	0.15
1979	31	877	0.16
1980	22	814	0.14
1981	21	362	0.06
1982	26	764	0.13
1983	20	809	0.13
1984	16	223	0.03
1985	22	1066	0.15

The table gives the number of fatal accidents and deaths on scheduled airline flights per year over a ten-year period. We use these data as a numerical example for fitting discrete data models.

(a) Assume that the numbers of fatal accidents in each year are independent with a Poisson(θ) distribution. Set a prior distribution for θ and determine the posterior distribution based on the data from 1976 through 1985. Under this model, give a 95% predictive interval for the number of fatal accidents in 1986. You can used the normal approximation to the gamma and Poisson or compute using simulation.

(b) Assume that the numbers of fatal accidents in each year follow independent Poisson distributions with a constant rate and an exposure in each year proportional to the number of passenger miles flown. Set a prior distribution for θ and determine the posterior distribution based on the data for 1976–1985. (Estimate the number of passenger miles flown in each year by dividing the appropriate columns of the table and ignoring round-off errors.) Give a 95% predictive interval for the number of fatal accidents in 1986 under the assumption that 8×10^{11} passenger miles are flown that year.

(c) Repeat (a) above, replacing 'fatal accidents' with 'passenger deaths.'

(d) Repeat (b) above, replacing 'fatal accidents' with 'passenger deaths.'

(e) In which of the cases (a)–(d) above does the Poisson model seem more or less reasonable? Why? Discuss based on general principles, without specific reference to the numbers in the table.

Incidentally, in 1986, there were 22 fatal accidents, 546 passenger deaths, and a death rate of 0.06 per 100 million miles flown.

Problem 2 (BDA2, p. 72, Exercise 2.11.21c, BDA3, p. 61, Exercise 2.11.19c)

Exponential model with conjugate prior distribution. The length of life of a light bulb manufactured by a certain process has an exponential distribution with unknown rate θ. Suppose the prior distribution of θ is a gamma distribution with coefficient of variation 0.5. (The *coefficient of variation* is defined as the standard deviation divided by the mean.) A random sample of light bulbs is to be tested and the lifetime of each obtained. If the coefficient of variation of the distribution of θ is to be reduced to 0.1, how many light bulbs need to be tested?

Problem 3 (BDA2, p. 72, Exercise 2.11.22a,b, BDA3, p. 61, Exercise 2.11.20a,b)

Censored and uncensored data in the exponential model.

(a) Suppose $y|\theta$ is exponentially distributed with rate θ, amd the marginal (prior) distribution of θ is Gamma(α, β). Suppose we observe that $y \geqslant 100$, but do not observe the exact value of y. What is the posterior distribution, $p(\theta|y \geqslant 100)$, as a function of α and β? Write down the posterior mean and variance of θ.

(b) In the above problem, suppose that we are now told that y is exactly 100. Now what are the posterior mean and variance of θ.

Problem 4

First do problem 3a above. Then reconsider the example on leukaemia data and perform an analysis of the drug group, following the outline given for the control group, but accounting for the additional difficulty of the censored times. After obtaining a sample from the posterior of the rate θ_D in the drug group, estimate the probability $\Pr(\theta_D > \theta_C)$ and look at the posterior distribution of other quantities of interest, such as the ratio between the mean remission time in the two groups, etc.

Lecture 12: The Multinomial model (BDA2 §3.5, BDA3 §3.4)

12.1 Multinomial model

Observe result of n trials in an experiment with k possible outcomes

- assume that all trials are independent and that in each

$$\Pr[\text{outcome } j|\theta] = \theta_j \qquad j = 1, \ldots, k$$

- let $y = (y_1, \ldots, y_k)$ with

$$y_j = \text{number of times outcome } j \text{ occurred in the } n \text{ trials}$$

$$[\text{clearly } y_1 + \cdots + y_k = n]$$

- Then $y \sim \text{Mult}(\theta)$, with mass function

$$p(y|\theta) = \frac{n!}{y_1! \cdots y_k!}\, \theta_1^{y_1} \cdots \theta_k^{y_k} \qquad \theta_j > 0 \;\forall j, \qquad \sum_{j=1}^{k} \theta_j = 1$$

- Although θ has k components, there really are only $k-1$ parameters, because of the sum to 1 constraint
- Special case: $k = 2$ yields the binomial model with

$$\theta_1 = \theta \qquad \text{and} \qquad \theta_2 = 1 - \theta$$

Likelihood, Prior, Posterior

- Likelihood

$$p(y|\theta) \propto \prod_{j=1}^{k} \theta_j^{y_j}$$

- Prior
  - *Definition*: We say that θ has a Dirichlet$(\alpha_1, \ldots, \alpha_k)$ distribution if its p.d.f. is

$$p(\theta) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)}\, \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1} \qquad \theta_j > 0, \qquad \sum_{j=1}^{k} \theta_j = 1$$

  - Note the similarity with the multinomial likelihood!
  - Using $\alpha_j = 1$ for all j yields a uniform prior density
- Posterior:

$$\begin{aligned} p(\theta|y) &\propto p(\theta)p(y|\theta) \\ &\propto \prod_{j=1}^{k} \theta_j^{\alpha_j - 1} \times \prod_{j=1}^{k} \theta_j^{y_j} \\ &\propto \prod_{j=1}^{k} \theta_j^{\alpha_j + y_j - 1} \\ &\propto \text{Dir}(\theta|\alpha_1 + y_1, \ldots, \alpha_k + y_k) \end{aligned}$$

  which proves that the Dirichlet distribution is the natural conjugate prior for multinomial data.

Some properties of the Dirichlet distribution

- If $k = 2$, it reduces to the Beta:

$$\begin{aligned} \text{Dir}(\theta_1, \theta_2 | \alpha_1, \alpha_2) &\propto \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \qquad \text{and since } \theta_1 + \theta_2 = 1 \\ &= \theta_1^{\alpha_1 - 1} (1 - \theta_1)^{\alpha_2 - 1} \\ &\propto \text{Be}(\theta_1 | \alpha_1, \alpha_2) \end{aligned}$$

- Suppose $X_j \sim \text{Ga}(\alpha_j, \beta)$, $j = 1, \ldots, k$, independently.

  Let

$$Z = \sum_{j=1}^{k} X_j \qquad \text{and} \qquad \theta_j = \frac{X_j}{Z}, \quad j = 1, \ldots, k$$

  Then $\theta = (\theta_1, \ldots, \theta_k)$ and Z are independent,

$$Z \sim \text{Ga}(\sum_{j=1}^{k} \alpha_j, \beta) \qquad \text{and} \qquad \theta \sim \text{Dir}(\alpha_1, \ldots, \alpha_k).$$

  Handy to generate from a Dirichlet, using calls to a Gamma random number generator

- Marginal distributions are Beta:

$$\text{If} \quad \theta \sim \text{Dir}(\alpha_1, \ldots, \alpha_k) \quad \text{then} \quad \theta_j \sim \text{Be}(\alpha_j, \alpha_0 - \alpha_j)$$

  where

$$\alpha_0 = \alpha_1 + \cdots + \alpha_k$$

- Expectations:

$$\text{E}(\theta_j) = \frac{\alpha_j}{\alpha_0}$$

- Variances and covariances:

$$\text{Var}(\theta_j) = \frac{\alpha_j(\alpha_0 - \alpha_j)}{\alpha_0^2(\alpha_0 + 1)} \qquad \text{Cov}(\theta_j, \theta_i) = \frac{-\alpha_j \alpha_i}{\alpha_0^2(\alpha_0 + 1)}$$

12.2 Example: Homework data

- Source: Warton (1997), *British Journal of Education Psychology*, **67**
- Three samples of Australian children from grade 2 (7–8 years), grade 4 (9–10 years) and grade 6 (11–12 years)
- Were asked their ideas about self-regulation with regard to homework:

  "Should you remember to do homework without being reminded?"

- Responses were categorized into the following four types:

  1. *Internal:* e.g.
     - Yes, because it's my job so it's my responsibility to remember
     - Yes, the homework was given to me, so I should remember

2. *Introjection:* e.g.
   - Yes, because I want to save Mum the trouble of reminding me
   - Yes, because Mum would be unhappy if I didn't remember
3. *External:* e.g.
   - Yes, because I can't watch TV until I finish
   - Yes, because I get nagged if I don't remember
4. *Irrelevant/Don't know:*
   - an answer that didn't involve responsibility, or no answer

- The following data were collected.

Type of reason	Grade 2	Grade 4	Grade 6
Internal	11	15	24
Introjection	9	9	4
External	5	6	3
Irrelevant	9	2	1

  - $\theta_2 = (\theta_{21}, \theta_{22}, \theta_{23}, \theta_{24})$: vector of probabilities for Grade 2
  - θ_4 and θ_6 similarly defined

- Assume independent Dirichlet priors on θ_2, θ_4 and θ_6
  - assign all the α hyperparameters equal to 1

- Then the posterior of θ is given by:

$$\begin{aligned} \theta_2|y &\sim \text{Dir}(12, 10, 6, 10) \\ \theta_4|y &\sim \text{Dir}(16, 10, 7, 3) \\ \theta_6|y &\sim \text{Dir}(25, 5, 4, 2) \end{aligned}$$

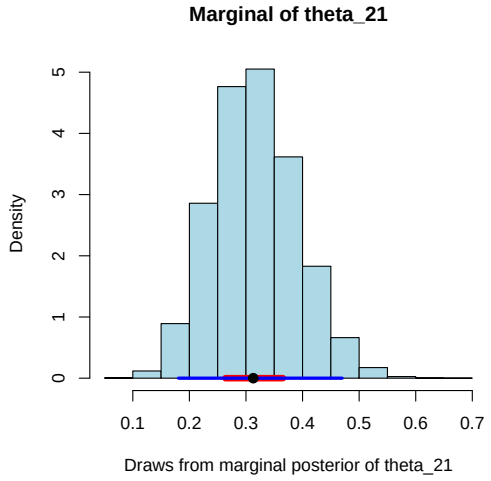
Posterior sampling in R

```
grade2 <- c(11, 9, 5, 9)
grade4 <- c(15, 9, 6, 2)
grade6 <- c(24, 4, 3, 1)

rdiric <- function(n, alpha, sumgam=F) {
# genetates n draws from a Dirichlet(alpha_1, ..., alpha_k)
#      distribution,  returned in a (k,n) matrix p
  k <- length(alpha)
  gam <- rgamma(k*n, rep(alpha,n))
  p <- matrix(gam, k, n)
  psum <- colSums(p)
  p <- sweep(p, MARGIN=2, STATS=psum, FUN="/")
  if(sumgam) return(list(p=p, psum=psum))
  else return(p)
}

alpha <- rep(1, 4)
nsamp <- 100000
theta2 <- rdiric(nsamp, alpha + grade2)
theta4 <- rdiric(nsamp, alpha + grade4)
theta6 <- rdiric(nsamp, alpha + grade6)
```

Marginal distributions and posterior intervals

Histograms and posterior intervals can be readily produced:

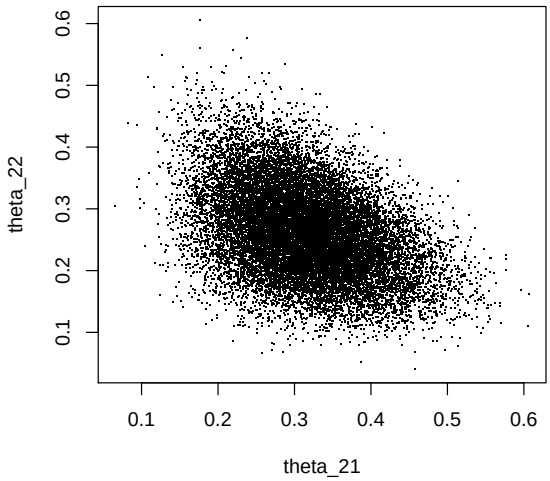
```
> quant <- histpost(theta2[1,], prob=TRUE, main="Marginal of theta_21",
           ylab="Density", xlab="Draws from marginal posterior of theta_21")
> quant
      2.5%     25.0%     50.0%     75.0%     97.5%
 0.1806072 0.2636058 0.3128924 0.3653744 0.4700020
```



Joint distributions

May also be of interest to look at the joint distribution of two probabilities:

```
ind <- seq(1, nsamp, by=5)
plot(theta2[1, ind], theta2[2, ind], pch=".", xlab="theta_21", ylab="theta_22")
```



Posterior probabilities

From the samples we can estimate any quantity of interest. Here are some examples:

- $\Pr\left[\theta_{2,\text{Internal}} > \theta_{2,\text{Introjection}} \middle| y\right]$?

```
> mean(theta2[1,] > theta2[2,])
[1] 0.67041
```

- $\Pr\left[\left|\theta_{2,\text{Internal}} - \theta_{2,\text{Introjection}}\right| < 0.1 \middle| y\right]$?

```
> mean(abs(theta2[1,] - theta2[2,]) < 0.1)
[1] 0.5419
```

- $\Pr\left[\theta_{2,\text{Internal}} > \theta_{2,\text{External}} \middle| y\right]$?

```
> mean(theta2[1,] > theta2[3,])
[1] 0.92895
```

Similarly in grade 4 and grade 6, the probabilities that

- the *Internal* category is more common than the *Introjection* one can be estimated as

```
> mean(theta4[1,] > theta4[2,])
[1] 0.88423
> mean(theta6[1,] > theta6[2,])
[1] 0.99996
```

- the probabilities that the *Internal* category is more common than the *External* one have estimates

```
> mean(theta4[1,] > theta4[3,])
[1] 0.97373
> mean(theta6[1,] > theta6[3,])
[1] 0.99997
```

Can also estimate probabilities for comparisons between grades. For instance,

- the posterior probability that the popularity of the *Internal* response increases with children's age:

$$\Pr\left[\theta_{2,\text{Internal}} < \theta_{4,\text{Internal}} < \theta_{6,\text{Internal}} \middle| y\right]$$

```
> mean(theta2[1,] < theta4[1,] & theta4[1,] < theta6[1,])
[1] 0.85985
```

- or the probability that the *Internal* response increases while the *Introjection* response decrease with age:

$$\Pr\left[\theta_{21} < \theta_{41} < \theta_{61},\ \ \theta_{22} > \theta_{42} > \theta_{62} \middle| y\right]$$

```
> mean(theta2[1,] < theta4[1,] & theta4[1,] < theta6[1,] &
       theta2[2,] > theta4[2,] & theta4[2,] > theta6[2,])
[1] 0.36677
```

- or the probability that (i) the *Internal* response increases with age *and* (ii) the *Introjection* and *External* responses in Grades 2 and 4 are higher than in Grade 6:

$$\Pr[\theta_{21} < \theta_{41} < \theta_{61}, \theta_{22} > \theta_{62}, \theta_{42} > \theta_{62}, \theta_{23} > \theta_{63}, \theta_{43} > \theta_{63} | y]$$

```
> mean(theta2[1,] < theta4[1,] & theta4[1,] < theta6[1,] &
       theta2[2,] > theta6[2,] & theta4[2,] > theta6[2,] &
       theta2[3,] > theta6[3,] & theta4[3,] > theta6[3,])
[1] 0.50813
```

Lecture 13: Direct simulation from the posterior (BDA2 §11.1, BDA3 §10.3)

13.1 Direct simulation

- Feasible in simple problems, such as analyses with conjugate priors (we've used it many times!)
- Sometimes can be done in stages, after factorizing the joint posterior, e.g.

$$p(\theta_1, \theta_2, \theta_3|y) = p(\theta_1|y) \times p(\theta_2|\theta_1, y) \times p(\theta_3|\theta_1, \theta_2, y)$$

1. Draw θ_1^* from $p(\theta_1|y)$
2. Draw θ_2^* from $p(\theta_2|\theta_1^*, y)$
3. Draw θ_3^* from $p(\theta_3|\theta_1^*, \theta_2^*, y)$

   $\Longrightarrow (\theta_1^*, \theta_2^*, \theta_3^*)$ is a draw from $p(\theta_1, \theta_2, \theta_3|y)$

13.2 Grid approximation

- Used it in Rat Tumor data and Michelson data examples
- Only feasible for low-dimensional θ, typically, univariate or bivariate
- In summary:

1. Evaluate $p(\theta|y)$ over a fine grid of θ values:

$$\theta_j = \theta_{lo} + j\delta \qquad j = 0, 1, \ldots, N,$$

   where $\delta = \frac{1}{N}(\theta_{hi} - \theta_{lo})$ is the grid increment
2. Sample from the set of θ_js with weights equal to

$$\frac{p(\theta_j|y)}{\sum_{j=0}^{N} p(\theta_j|y)}$$

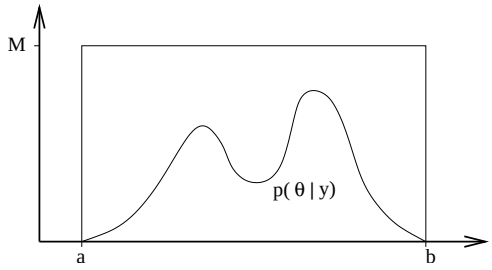
   yielding draws $\theta^{(t)}$, $t = 1, \ldots, T$
3. Randomly displace the draws within each grid interval:

$$\theta^{(t)} \quad \rightarrow \quad \theta^{(t)} + \text{Un}\left(-\frac{\delta}{2}, \frac{\delta}{2}\right)$$

- Works well if grid is fine and there is negligible posterior mass outside its range

13.3 Rejection sampling

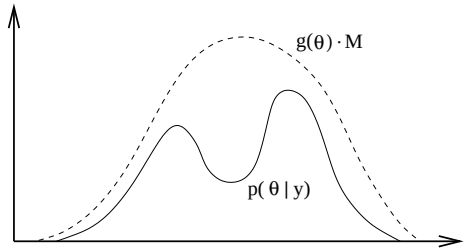
- Suppose that $p(\theta|y)$ has support on an interval:



- Rejection sampling:
  1. draw a random point in the box
  2. if it falls below $p(\theta|y)$ accept it, otherwise reject it and go to 1.
- Can prove that this algorithm yields a draw from $p(\theta|y)$
- Some examples using R ... (code in the Appendix)

Using a non-uniform proposal

- More generally:



- General idea:
  - want to draw from $p(\theta|y)$, known only up to a proportionality constant
  - generate from another distribution $g(\theta)$
    - keep the draw only with some probability
- Requirements on $g(\theta)$:
  - Must be able to generate from $g(\theta)$ easily
  - There exists a constant M such that $\frac{p(\theta|y)}{g(\theta)} \leqslant M$, for all θ.
- Rejection Sampling Algorithm:
  1. Sample θ from $g(\theta)$
  2. Accept θ with probability $\frac{p(\theta|y)}{Mg(\theta)}$; if θ is rejected go to 1.

Example

- Want to draw from

$$p(\theta|y) = K\theta^{\alpha}(1-\theta)^{\beta}(1+\theta)^{\gamma} \qquad 0 < \theta < 1 \qquad \alpha, \beta, \gamma > 0$$

  - Naive approach: $p(\theta|y) \leqslant K \times 2^{\gamma}$, so can use $g(\theta) = 1$ and $M = K \times 2^{\gamma}$
    1. Draw $\theta \sim \text{Un}(0, 1)$
    2. Accept with probability

$$\frac{p(\theta|y)}{K \times 2^{\gamma}} = \theta^{\alpha}(1-\theta)^{\beta}\left(\frac{1+\theta}{2}\right)^{\gamma}$$

  - This can be very inefficient, as bound $K \times 2^{\gamma}$ on $p(\theta|y)$ is very loose

- A more efficient approach:
  - plot $p(\theta|y)/K = \theta^{\alpha}(1-\theta)^{\beta}(1+\theta)^{\gamma}$ on a grid of θ values in $(0, 1)$
  - pick M^* close but larger than maximal ordinate
  - apply algorithm above, but replacing 2^{γ} with M^*, i.e., replace step 2 with following:

    2a. Accept with probability

$$\frac{p(\theta|y)}{K \times M^*} = \frac{\theta^{\alpha}(1-\theta)^{\beta}(1+\theta)^{\gamma}}{M^*}$$

  - Note that this does not require the value of K

Some notes

- $g(\theta) \propto 1$ will work only if $p(\theta|y)$ is relatively flat
  - if it has a very sharp peak, acceptance rate will be very low
- algorithm is self-monitoring:
  - if the acceptance rate is very low, we know that a better proposal distribution $g(\theta)$ is needed

Appendix: an R function to illustrate rejection sampling

```
rejection <- function(n, a, b, M, p)
{
#    illustrates rejection method to draw a sample of n from a distribution
#    with p.d.f. proportional to p with support on [a,b] and max(p) <= M
  while(length(dev.list()) < 2) X11()
  draws <- rep(0, n)
  nprop <- nacc <- 0
  grid <- seq(a, b, length=200)
  plot(grid, p(grid), type="l", ylim=c(0, M))
  while(nacc < n) {
    nprop <- nprop + 1
    x <- runif(1, a, b)
    y <- runif(1, 0, M)
    if (y < p(x)) {
      points(x, y, pch=1, col="green")
      nacc <- nacc + 1
      draws[nacc] <- x
      dev.set(which=dev.next())
      if(nacc > 2) hist(draws[1:nacc], prob=TRUE, 20, col="lightblue")
      dev.set(which=dev.prev())
    }
    else points(x, y, pch=4, col="red")
    if(nprop == 15) locator()
  }
  cat("\n Proposals: ", nprop, "   Acceptances: ", nacc, "\n")
  return(invisible(list(nprop=nprop, draws=draws)))
}

# some examples:
n <- 1000
a <- -3; b <-  3;  M <- 4;       p <- function(x){abs(x)}
res <- rejection(n, a, b, M, p)

a <- -6; b <- 3;   M <- 1;       p <- function(x){abs(sin(x))}
res <- rejection(n, a, b, M, p)

a <- -6; b <- 0.5; M <- 1;       p <- function(x){exp(x)*abs(sin(x))}
res <- rejection(n, a, b, M, p)

a <- -6; b <- 1;   M <- exp(1);  p <- function(x){exp(x)*abs(sin(x))}
res <- rejection(n, a, b, M, p)

a <- 0;  b <- 2;   M <-  1;      p <- function(x){x^10*(x<1) + (x-2)^10*(x>1)}
res <- rejection(n, a, b, M, p)

a <- 0;  b <- 3;   M <- exp(27); p <- function(x){exp(x^3)}
res <- rejection(n, a, b, M, p)

a <- 0;  b <- 1;   M <- 2^2;     p <- function(x){x * (1-x)^3 * (1+x)^2}
res <- rejection(n, a, b, M, p)
```

Lecture 14: Markov Chain Monte Carlo (BDA2/3 Chapter 11; FCBSM Chapters 6 and 10)

14.1 Markov Chain Monte Carlo Methods

Markov Chain Monte Carlo (MCMC) produces a *dependent* sample from the "target" distribution $\pi(\theta) = p(\theta|y)$, the posterior distribution.

The general idea:

- Construct a Markov chain with equilibrium distribution $\pi(\theta)$
- After a "long enough" time, the current state of the Markov chain will be (approximately) a draw from $\pi(\theta)$
- Usually: given a Markov chain transition kernel $P(\theta_{t+1}; \theta_t)$ defined by

$$\Pr[\theta_{t+1} \in A|\theta_t] = \int_A P(\theta_{t+1}; \theta_t)\, d\theta_{t+1}$$

  find the stationary distribution $\pi(\theta)$
- In MCMC: given π, find a P with stationary distribution equal to π
- Need the Markov chain to
  - be *irreducible* (can go from any state to any other state in a finite number of steps)
  - be *aperiodic* (satisfied if chain can remain at current state)
  - satisfy *global balance*:

$$\pi(\theta') = \int P(\theta'; \theta)\pi(\theta)\, d\theta \qquad \text{for all } \theta'$$

  - Easier to work with the stronger *detailed balance* (reversibility) condition:

$$\pi(\theta')P(\theta; \theta') = \pi(\theta)P(\theta'; \theta)$$

How to construct the Markov chain P?

- Two basic approaches:
  - Gibbs Sampling
  - Metropolis–Hastings algorithm (see Advanced Bayesian Methods, Semester 2)

14.2 Gibbs sampling

- Introduced by Geman and Geman (1984); also known as "heat bath"
- Suppose the parameter θ is partitioned as

$$\theta = \{\theta_1, \theta_2, \ldots, \theta_p\},$$

where each θ_i may be multivariate

  - Let the current state of the Markov chain be $\theta^{(t)} = \{\theta_1^{(t)}, \ldots, \theta_p^{(t)}\}$
  - At time $t+1$:
    1. draw $\theta_1^{(t+1)}$ from $\pi(\theta_1|\theta_2^{(t)}, \ldots, \theta_p^{(t)})$
    2. draw $\theta_2^{(t+1)}$ from $\pi(\theta_2|\theta_1^{(t+1)}, \theta_3^{(t)}, \ldots, \theta_p^{(t)})$

       $\vdots$

    p. draw $\theta_p^{(t+1)}$ from $\pi(\theta_p|\theta_1^{(t+1)}, \ldots, \theta_{p-1}^{(t+1)})$
  - This is called a *deterministic sweep* Gibbs sampler
  - Components could also be updated in random order (random sweep)

- Must be able to simulate from $\pi(\theta_i|\theta_{-i})$, the *full conditional* distributions. Here θ_{-i} indicates all the components of θ *except* θ_i
- Faster convergence to equilibrium distribution achieved by grouping θ_is together (*blocking*)

Full conditional distributions

- How to derive the full conditionals $\pi(\theta_i|\theta_{-i})$?
  1. Write the joint density of all quantities $p(\theta, y)$
  2. Only keep the terms that involve θ_i.
     - is the result (proportional to) a "simple" distribution?
- A formal example.
  Suppose that $\theta = \{\alpha, \beta, \gamma, \delta\}$ and that the prior $p(\theta)$ and the likelihood $p(y|\theta)$ satisfy

$$\begin{aligned} p(\theta) &= p(\alpha)\, p(\beta|\alpha)\, p(\gamma|\alpha, \beta)\, p(\delta|\alpha, \beta, \gamma) \\ &= p(\alpha)\, p(\beta)\, p(\gamma|\alpha, \beta)\, p(\delta) \\ p(y|\theta) &= p(y|\gamma, \delta) \end{aligned}$$

  Then

$$p(\theta|y) \propto p(\theta, y) = p(\alpha)\, p(\beta)\, p(\gamma|\alpha, \beta)\, p(\delta)\, p(y|\gamma, \delta)$$

  Full conditional of α:

$$\begin{aligned} p(\alpha|\beta, \gamma, \delta, y) &= \frac{p(\alpha, \beta, \gamma, \delta, y)}{p(\beta, \gamma, \delta, y)} \\ &\propto p(\alpha, \beta, \gamma, \delta, y) \\ &\propto p(\alpha)\, p(\gamma|\alpha, \beta) \end{aligned}$$

In a similar way

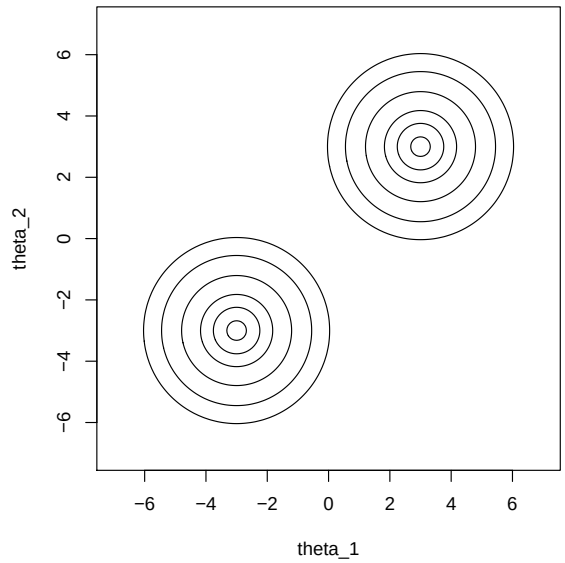
$$\begin{aligned} p(\beta|\cdots) &\propto p(\beta)\, p(\gamma|\alpha,\beta) \\ p(\gamma|\cdots) &\propto p(\gamma|\alpha,\beta)\, p(y|\gamma,\delta) \\ p(\delta|\cdots) &\propto p(\delta)\, p(y|\gamma,\delta) \end{aligned}$$

- This is the procedure we used when deriving the full conditionals for the simple Normal hierarchical model.
- Look again at that section now!

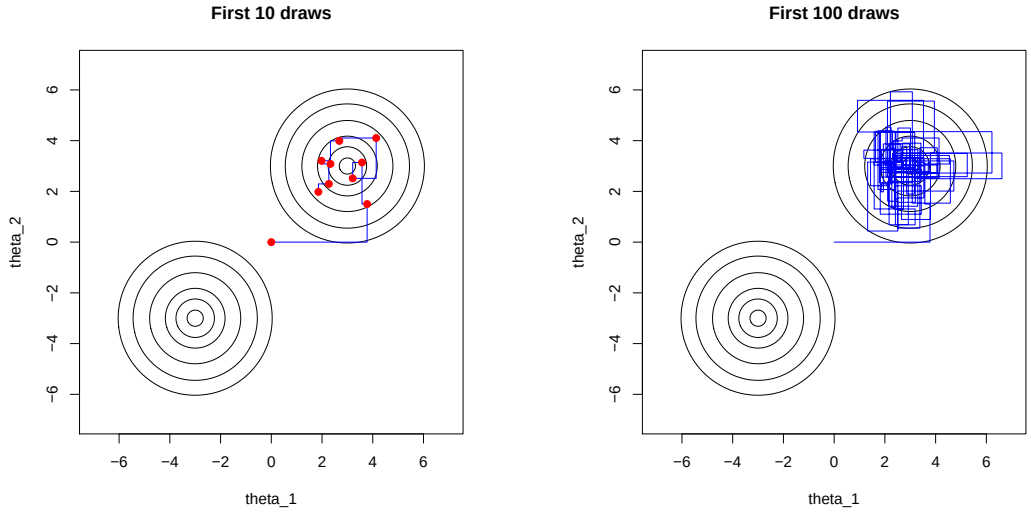
14.3 An example where Gibbs sampling does not work well

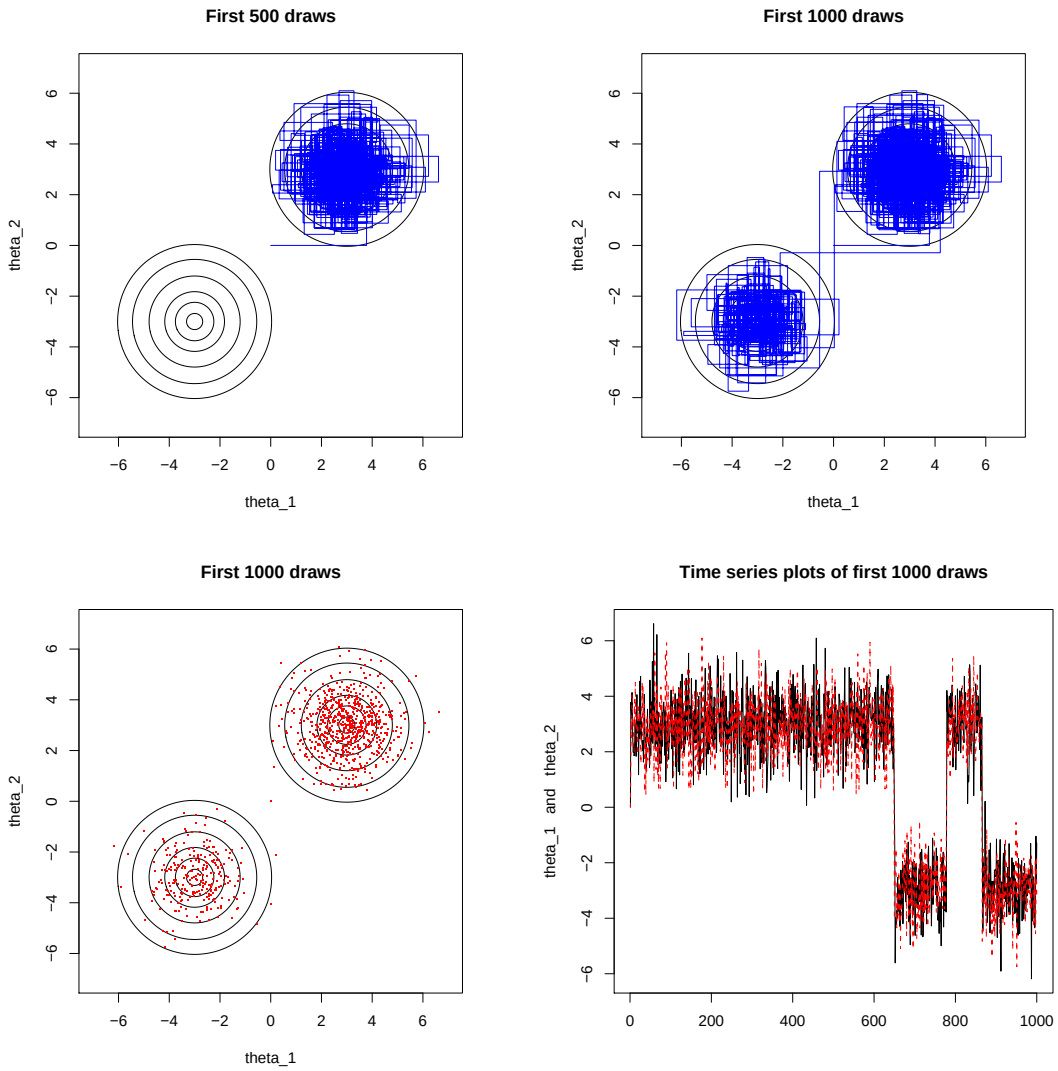
The target distribution is a 50-50 mixture of two bivariate normals distribution:

$$\pi(\theta) = \frac{1}{2} N_2\left\{ \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \middle| \begin{bmatrix} -3 \\ -3 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right\} + \frac{1}{2} N_2\left\{ \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \middle| \begin{bmatrix} 3 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right\}$$



The sampler was started at $\theta_1 = \theta_2 = 0$.





After the first 500 sweeps one mode has not yet been visited! Although the Gibbs sampler eventually moves from one mode to the other, transitions are rather infrequent and the amount of simulation time spent around each mode is, after 1000 sweeps, still quite unbalanced.

Note that the full conditional distributions are mixtures of a $N(-3,1)$ and a $N(3,1)$, but with unequal weights. Denoting by c the component in the mixture, for instance, the full conditional of θ_2 is

$$\begin{aligned} p(\theta_2|\theta_1) &= p(\theta_2|\theta_1, c=1)p(c=1|\theta_1) + p(\theta_2|\theta_1, c=2)p(c=2|\theta_1) \\ &= N(\theta_2|-3,1)p(c=1|\theta_1) + N(\theta_2|3,1)[1-p(c=1|\theta_1)] \end{aligned}$$

where

$$\begin{aligned} p(c=1|\theta_1) &= \frac{p(\theta_1|c=1)p(c=1)}{p(\theta_1|c=1)p(c=1)+p(\theta_1|c=2)p(c=2)} \\ &= \frac{N(\theta_1|-3,1)}{N(\theta_1|-3,1)+N(\theta_1|3,1)} \end{aligned}$$

Appendix: R code for the bivariate mixture example

```
set.seed(234)
nsamp <- 1000
y <- matrix(0, 2, nsamp)
y[,1] <- c(0, 0)
for (i in (2:nsamp)) {
  postc1 <- 1 / (1 + dnorm(y[2, i-1], 3, 1) / dnorm(y[2, i-1], -3, 1) )
  if(runif(1) < postc1) y[1, i] <- rnorm(1, -3, 1)
  else y[1, i] <- rnorm(1, 3, 1)
  postc1 <- 1 / (1 + dnorm(y[1, i], 3, 1) / dnorm(y[1, i], -3, 1) )
  if(runif(1) < postc1) y[2, i] <- rnorm(1, -3, 1)
  else y[2, i] <- rnorm(1, 3, 1)
}

ngrid <- 200
y1.grd <- y2.grd <- seq(-7, 7, length=ngrid)
y12.grd <- expand.grid(y1.grd, y2.grd)
dens <- (0.5 * dnorm(y12.grd[,1], -3, 1) * dnorm(y12.grd[,2], -3, 1) +
         0.5 * dnorm(y12.grd[,1], 3, 1) * dnorm(y12.grd[,2], 3, 1) )
dens <- matrix(dens, ngrid, ngrid)

levs <- max(dens)* c(0.95, 0.75, 0.5, 0.2, 0.05, 0.01)
xl <- "theta_1";    yl <- "theta_2"

oldpar <- par(pty="s")
contour(y1.grd, y2.grd, dens, drawlabels=FALSE, levels=levs, xlab=xl, ylab=yl)

contour(y1.grd, y2.grd, dens, drawlabels=FALSE, levels=levs, xlab=xl, ylab=yl)
title(main="First 10 draws")
lines(t(y[,1:10]), type="s", col="blue")
points(t(y[,1:10]), pch=16, col="red")

contour(y1.grd, y2.grd, dens, drawlabels=FALSE, levels=levs, xlab=xl, ylab=yl)
title(main="First 100 draws")
lines(t(y[,1:100]), type="s", col="blue")

contour(y1.grd, y2.grd, dens, drawlabels=FALSE, levels=levs, xlab=xl, ylab=yl)
title(main="First 500 draws")
lines(t(y[,1:500]), type="s", col="blue")

contour(y1.grd, y2.grd, dens, drawlabels=FALSE, levels=levs, xlab=xl, ylab=yl)
title(main="First 1000 draws")
lines(t(y[,1:1000]), type="s", col="blue")

contour(y1.grd, y2.grd, dens, drawlabels=FALSE, levels=levs, xlab=xl, ylab=yl)
title(main="First 1000 draws")
points(t(y[,1:1000]), pch=".", col="red")
par(oldpar)

matplot(t(y[,1:1000]), type="l", ylab="theta_1   and   theta_2")
title(main="Time series plots of first 1000 draws")
```

Lecture 15: Evaluating integrals by Monte Carlo

15.1 Introduction

- Integration is basic to Bayes, as optimization (maximization) is to maximum likelihood.

  Some instances:

  - Marginal posterior distributions

    $$p(\theta|y) \quad \longrightarrow \quad p(\theta_j|y)$$

  - Marginal (or integrated) likelihoods – prior predictive of observed data

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

    * proportionality constant of $p(\theta|y) = p(\theta)p(y|\theta)/p(y)$
    * plays important role in Bayesian model selection (Bayes Factors: see ABM, Semester 2)

  - Posterior probabilities

    $$p(\theta \in A) = \int_A p(\theta|y)\,d\theta$$

- In general, will be interested in the posterior expectation of some function $\psi(\theta)$ of the parameter:

  $$E[\psi|y] = \int \psi(\theta)p(\theta|y)\,d\theta$$

15.2 Crude Monte Carlo

- Suppose want to compute

  $$\mu = \int_0^1 \phi(\theta)\,d\theta$$

  - Draw $\theta_1, \ldots, \theta_T \overset{\text{i.i.d.}}{\sim} Un(0,1)$
  - let $\phi_t = \phi(\theta_t)$, so that ϕ_ts are i.i.d. r.v.s with

    $$E[\phi_t] = \int_0^1 \phi(\theta_t) \cdot 1\,d\theta_t = \mu$$

  - Consider the estimator

    $$\hat{\mu} = \overline{\phi} = \frac{1}{T}\sum_{t=1}^{T} \phi_t$$

    Then

    $$E[\overline{\phi}] = \mu \qquad Var[\overline{\phi}] = \frac{\sigma^2}{T}$$

    where $\sigma^2 = Var(\phi_t)$. Also, from the Central Limit Theorem,

    $$\overline{\phi} \overset{\text{approx.}}{\sim} N\left(\mu, \frac{\sigma^2}{T}\right)$$

Replacing σ^2 with the sample variance

$$s^2 = \frac{1}{T-1}\sum_{t=1}^{T}(\phi_t - \overline{\phi})^2$$

yields an approximate 95% C.I. for μ:

$$\overline{\phi} \pm 2\frac{s}{\sqrt{T}}$$

- In short: can readily estimate the variability of crude Monte Carlo estimates

- Typically

$$\phi(\theta) = \psi(\theta)p(\theta|y)$$

  - could apply crude Monte Carlo but
    - estimates are inefficient if $p(\theta|y)$ has sharp peaks:
    - $p(\theta|y)$ concentrated $\Longrightarrow$ highly variable ϕ_ts $\Longrightarrow$ highly variable $\overline{\phi}$

- If one can draw from $p(\theta|y)$ things can be improved:

$$\mu = \int \psi(\theta)p(\theta|y)\,d\theta = E[\psi(\theta)|y]$$

- Suggests drawing from $p(\theta|y)$ and averaging the $\psi(\theta_t)$:

$$\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T}\psi(\theta_t) \qquad \text{with } \theta_1, \ldots, \theta_T \overset{\text{i.i.d.}}{\sim} p(\theta|y)$$

  Then previous analysis applies and

$$\hat{\mu} \overset{\text{approx.}}{\sim} N\left(\mu, \frac{\sigma^2}{T}\right) \qquad \text{with } \sigma^2 = \text{Var}(\psi|y)$$

- However
  - often it is infeasible to obtain i.i.d. draws from $p(\theta|y)$;
  - instead, one can make i.i.d. draws from a distribution "close" to $p(\theta|y)$;
  - need to correct for the approximation: this leads to a technique known as "importance sampling"

15.3 Importance sampling (BDA2 §13.3, BDA3 §10.4)

- Suppose we can only draw from a distribution $g(\theta)$. Then, can rewrite $E[\psi(\theta)|y]$ as

$$\begin{aligned}\int \psi(\theta)p(\theta|y)\,d\theta &= \int \psi(\theta)\frac{p(\theta|y)}{g(\theta)}g(\theta)\,d\theta \\ &= E\left[\psi(\theta)\frac{p(\theta|y)}{g(\theta)}\right]\end{aligned}$$

  where, in the last line, E is with respect to the distribution $g(\theta)$.

- Hence, can use the estimate

$$\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T}\psi(\theta_t)w(\theta_t) \qquad \text{with } \theta_1,\ldots,\theta_T \overset{\text{i.i.d}}{\sim} g(\theta)$$

  where $w(\theta) = \dfrac{p(\theta|y)}{g(\theta)}$ are called *importance weights*.

- In practice, one does not divide by T, but by the sum of the weights:

$$\hat{\mu} = \frac{\sum_{t=1}^{T}\psi(\theta_t)w(\theta_t)}{\sum_{t=1}^{T}w(\theta_t)} = \sum_{t=1}^{T}\psi(\theta_t)\frac{w(\theta_t)}{\sum_{\ell=1}^{T}w(\theta_\ell)}$$

  - Advantage: only need to be able to compute $g(\theta)$ and $p(\theta|y)$ up to a proportionality constant

Notes

- precise estimates result if weights $w(\theta) = \dfrac{p(\theta|y)}{g(\theta)}$ do not vary much
- optimal choice: $g(\theta) = p(\theta|y)$ implies constant weights
- $g(\theta)$ should be "not small" wherever $p(\theta|y)$ is "not small".
- tails of $g(\theta)$ should be thicker than those of $p(\theta|y)$: otherwise will occasionally get very large weights
- can look at distribution of sampled importance weights to detect problems
- if importance function $g(\theta)$ was a poor choice, may completely miss a region of high posterior density

A routine example

In the Homework data example discussed in §12.2, we used a sample of $n = 100,000$ draws from the posterior distribution of θ_2 to compute a simulation estimate of the probability

$$\Pr\left[\theta_{2,\text{Internal}} > \theta_{2,\text{Introjection}} \middle| y\right]$$

as

```
> mean(theta2[1,] > theta2[2,])
[1] 0.67041
```

Compute approximate 95% and 99% C.I.s for the probability in question, using the fact that, from the same n draws

```
> var(theta2[1,] > theta2[2,])
[1] 0.2209626
```

Lecture 16: Introduction to (Bayesian) decision theory

16.1 Introduction

- Good book: Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd edn, Springer
- Read Chapter 1 if you can.

We usually do statistical analysis to inform some decision. We perform a clinical trial so as to decide "should this drug go to market". We model the value of a stock in order to decide "should I invest X in it". How to make that decision in a situation where there is uncertainty about the system is what this lecture is about. Bayesian analysis as so far covered spits out a posterior distribution not a decision. Classical hypothesis testing can makes decisions between two hypotheses based on controlling error rates (typically Type I errors), but is silent on what significance level to choose.

The insight of "decision theory" (developed by Wald) is that you can't make a decision without understanding the consequences of that decision: e.g., *just how bad is it I make a particular wrong decision*. Very early it was recognized that expectation played a key rôle in making a decision. For example, in the early development of probability, games of chance were a focus, and the expected financial gain/loss was key to deciding whether to play the game:

$$\text{Expected financial loss} = \text{Cost of bet} - \text{Expected winnings}.$$

However this was not unproblematic.

16.2 St Petersburg paradox

Suppose the game is to toss a coin repeatedly. If the first head occurs on toss n, you win 2^n rubles:

$$\begin{aligned} P(n) &= \left(\frac{1}{2}\right)^n, \\ \mathrm{E}[\text{winnings}] &= \mathrm{E}[2^n] = \sum_{n=1}^{\infty} 2^n \frac{1}{2^n} = \infty. \end{aligned}$$

So you decide the bet is worth taking at any cost! But there is less than a probability of 0.05 of the game going beyond 5 throws and the winnings being more than 32 rubles.)

The resolution of this paradox is that the practical utility of the winnings doesn't increase linearly in the winnings: the practical value of 1000 extra rubles is much less to someone already winning 1,000,000 rubles compared to someone with only a 100 rubles. (Laplace distinguishes "moral expectation" (the practical usefulness) from the raw financial gain.)

16.3 Loss function

All this is formalized by the idea of a *loss function*. Suppose the true state of the system is represented by $\theta \in \{\theta_1, \ldots, \theta_k\}$, e.g. possible values of some parameter in a statistical model of the system. (Note: I've shown the set of possible values as discrete,

but it might be continuous, e.g., the whole real number line. Suppose the possible decisions are $a \in \{a_1, \ldots, a_l\}$ (again potentially continuous). Then the loss function

$$L(\theta, a)$$

represents the loss (financial, practical, waste of energy, moral, ...) in choosing decision a when the true state (parameter) is θ.

Note:

- if θ is a real parameter, then the decision a might trivially be deciding on the value of θ, so a is a real number too, and that is the case we will mostly be thinking about;
- if the possible θs and as are discrete $L(\theta, a)$ is just a matrix L (the 'loss matrix') with elements $L(\theta_i, a_j)$.

	buy a_1	don't buy a_2
bond no default θ_1	-500 (Minus) return on investment	-300 Risk-free (minus) return
bond default θ_2	1000 Loss of investment	-300 Risk-free (minus) return

16.4 Expected loss

In the Bayesian world, θ is not known with certainty so $L(\theta, a)$ can't be evaluated. But θ has a posterior distribution $p(\theta|y)$ for data y. So it makes sense to average loss over the uncertainty in θ, i.e., to take the expectation of $L(\theta, a)$ over $p(\theta|y)$. This is called *Bayesian Expected Loss*:

$$\rho(\pi, a) = \int L(\theta, a) p(\theta|y) d\theta,$$

where π indicates an expectation over the posterior. We can then seek the particular decision a^π which minimizes $\rho(\pi, a)$, called the *Bayes action*.

16.5 Aside

Frequentists need to make decisions too, but they can't calculate an expectation over $p(\theta|y)$, since θ isn't random! What they do is recognize that the decision is a function of the data y (a decision rule), with loss $L\big(\theta, a(y)\big)$. Then they average this loss over $p(y|\theta)$, the likelihood to give *frequentist risk*

$$R\big(\theta, a(y)\big) = \int L\big(\theta, a(y)\big) p(y|\theta) dy.$$

Yet another risk integrates out both y and θ:

$$\int L\big(\theta, a(y)\big) p(y, \theta) dy\, d\theta = \int R\big(\theta, a(y)\big) p(\theta) d\theta;$$

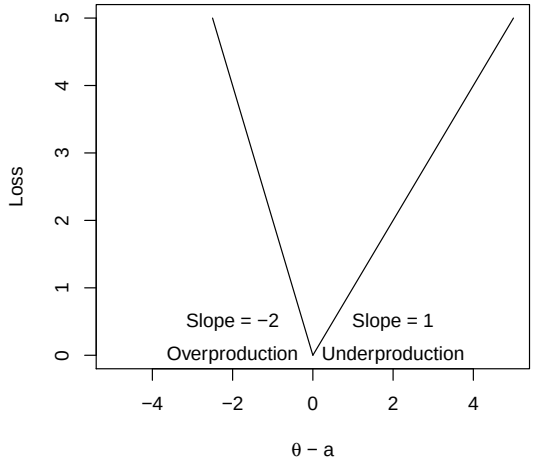
this is called *Bayes risk*.

16.6 Example

Suppose a drug company is investigating whether they should market a new painkiller or not. That depends on what fraction of the market it will take, θ. This is an unknown parameter. The experience with previous drugs suggest a prior

$$\theta \sim \text{Un}[0.1, 0.2].$$

Let's try to make a decision about what θ is. Let's call that decided value, a. So we need a loss function. A conversation with the pharma company suggests:



$$L(\theta, a) = \begin{cases} \theta - a & \theta - a \geqslant 0; \\ 2(a - \theta) & \theta - a < 0. \end{cases}$$

The Bayes expected loss—without collecting any new data about θ so that $p(\theta|y) = p(\theta)$—is

$$\begin{aligned} \rho(\pi, a) &= \int L(\theta, a) p(\theta|y) d\theta = \int L(\theta, a) p(\theta) d\theta \\ &= \int_{0.1}^{0.2} L(\theta, a) \cdot 10 \cdot d\theta \\ &= \begin{cases} 0.15 - a & a \leqslant 0.1; \\ 15a^2 - 4a + 0.3 & 0.1 < a \leqslant 0.2; \\ 2a - 0.3 & a > 0.2. \end{cases} \end{aligned}$$

[CHECK THIS!]

This is minimized at $a^\pi = \frac{2}{15} = 0.1333$ [CHECK THIS TOO!]. This is the Bayes action.

16.7 Particular loss functions

The choice of the loss function depends on the effects of the decision—it needs expert (discipline-specific) input. But we need that input anyway to justify the prior and, even, likelihood. However, some standard off-the-shelf loss functions are often used.

16.7.1 Squared-error loss

$$L(\theta, a) = (\theta - a)^2$$

Theorem 16.1 *The Bayes action is the posterior mean of* θ.

$$\begin{aligned}\rho(\pi, a) &= \int (\theta - a)^2 p(\theta|y) d\theta \\ &= E[(\theta - a)^2|y] \\ &= E[\theta^2|y] - 2aE[\theta|y] + a^2\end{aligned}$$

So $d\rho/da = -2E[\theta|y] + 2a$. Equating to zero and solving yields $a^\pi = E[\theta|y]$ (the second derivative there is positive, so it really is a minimum: CHECK).

Exercise: for the 'weighted' squared-error loss:

$$L(\theta, a) = \omega(\theta)(\theta - a)^2 \qquad \text{for some } \omega(\theta) > 0,$$

show that

$$a^\pi = \frac{E[\omega(\theta)\theta|y]}{E[\omega(\theta)|y]}.$$

16.7.2 Absolute error loss

$$L(\theta, a) = |\theta - a|.$$

Theorem 16.2 *The Bayes action is the posterior median of* θ.

[Below subscripts on probability distributions are there to clarify what the random variable is.]

$$\begin{aligned}\rho(\pi, a) &= \int_{-\infty}^{\infty} |\theta - a| p_\theta(\theta|y) d\theta \\ &= \int_{-\infty}^{a} (a - \theta) p_\theta(\theta|y) d\theta + \int_a^{\infty} (\theta - a) p_\theta(\theta|y) d\theta \\ &= (1) + (2)\end{aligned}$$

For (1) just integrate by parts:

$$\begin{aligned}(1) &= \Big[(a - \theta) P_\theta(\theta|y)\Big]_{-\infty}^{a} + \int_{-\infty}^{a} 1 \cdot P_\theta(\theta|y) d\theta \\ &= 0 + \int_{-\infty}^{a} P_\theta(\theta|y) d\theta,\end{aligned}$$

where $P_\theta(\theta|y)$ is the c.d.f. of $\theta|y$. Hence $d(1)/da = P_\theta(a|y)$.

(2) is a little tricker.

$$\begin{aligned}(2) &= \int_{-a}^{-\infty} (-\phi - a) p_\theta(-\phi|y)(-d\phi) \qquad \text{Chg. of var.: } \theta = -\phi \\ &= -\int_{-\infty}^{-a} (\phi + a) p_\phi(\phi|y) d\phi \qquad \text{by transformation of p.d.f.s} \\ &= -\Big[(\phi + a) P_\phi(\phi|y)\Big]_{-\infty}^{-a} + \int_{-\infty}^{-a} 1 \cdot P_\phi(\phi|y) d\phi \qquad \text{int. by parts} \\ &= 0 + \int_{-\infty}^{-a} P_\phi(\phi|y) d\phi\end{aligned}$$

Hence $d(2)/da = -d(2)/d(-a) = -P_\phi(-a|y) = -Pr(\phi < -a) = -Pr(\theta > a) = -[1 - P_\theta(a|y)] = P_\theta(a|y) - 1$.

Then $d\phi/da = d(1)/da + d(2)/da = 2P_\theta(a|y) - 1$. Equating to zero at $a = a^\pi$ gives

$$P_\theta(a^\pi|y) = \frac{1}{2}.$$

And hence a^π is the posterior median of θ.