MAT-51706
Bayesian Methods

Antti Penttinen
University of Jyväskylä

Robert Piché
Tampere University of Technology

2010

Bayesian statistical methods are widely used in many science and engineering areas including machine intelligence, expert systems, medical imaging, pattern recognition, decision theory, data compression and coding, estimation and prediction, bioinformatics, and data mining.

These course notes present the basic principles of Bayesian statistics. The first sections explain how to estimate parameters for simple standard statistical models (normal, binomial, Poisson, exponential), using both analytical formulas and the free WinBUGS data modelling software. This software is then used to explore multivariate hierarchical problems that arise in real applications. Advanced topics include decision theory, missing data, change point detection, model selection, and MCMC computational algorithms.

Students are assumed to have knowledge of basic probability. A standard introductory course in statistics is useful but not necessary. Additional course materials (exercises, recorded lectures, model exams) are available at `http://math.tut.fi/~piche/bayes`

Contents

1 Introduction **1**
- 1.1 Who was Thomas Bayes? . . . 1
- 1.2 The Fall and Rise of the Bayesians . . . 1

2 Probability **2**
- 2.1 Probability as a measure of belief . . . 2
- 2.2 How to assign probability . . . 3
- 2.3 Data changes probability . . . 3
- 2.4 Odds . . . 4
- 2.5 Independence . . . 4
- 2.6 Bayes's formula . . . 5

3 Normal Data **6**

4 Posteriors, Priors, and Predictive Distributions **7**
- 4.1 Using the Posterior . . . 8
- 4.2 Bayesian Data Analysis with Normal Models . . . 9
- 4.3 Using Bayes's Formula Sequentially . . . 10
- 4.4 Predictive Distributions . . . 11

5 Single-Parameter Models **13**
- 5.1 Estimating the mean of a normal likelihood . . . 13
- 5.2 Estimating the probability parameter in a binomial model . . . 15
- 5.3 Poisson model for count data . . . 18
- 5.4 Exponential model for lifetime data . . . 22
- 5.5 Estimating the variance of a normal model . . . 24

6 Jeffreys's prior **25**

7 Some General Principles **27**
- 7.1 Ancillarity and Sufficiency . . . 27
- 7.2 Likelihood Principle and Stopping Rules . . . 28

8 Hypothesis Testing **29**

9 Simple Multiparameter Models **31**
- 9.1 Two-parameter normal model . . . 32
- 9.2 Comparing two normal populations . . . 37
- 9.3 Multinomial model . . . 40

10 The modal approximation and Laplace's method **42**

11 Hierarchical Models and Regression Models **45**
- 11.1 DAGs . . . 45
- 11.2 Hierarchical normal model . . . 45
- 11.3 Linear regression . . . 46
- 11.4 Autoregressive model of time series . . . 48

11.5 Logistic regression 50
11.6 Change point detection 51

12 MCMC **52**
12.1 Markov chains 52
12.2 Gibbs sampler 55

13 Model comparison **59**
13.1 Bayes factors 59
13.2 Deviance Information Criterion (DIC) 63

14 Decision Theory **65**
14.1 The Bayesian choice 65
14.2 Loss functions for point estimation 68
14.3 Decision Rules and the Value of an Observation 69

15 Exact marginalisation **71**
15.1 Change Point Detection 71
15.2 Multivariate normal linear model with a parameter 72
15.3 Spectrum Analysis 73
15.4 Autoregressive model of time series 75
15.5 Regularisation 76

1 Introduction

1.1 Who was Thomas Bayes?

The reverend Thomas Bayes (1702–1761) was an English presbyterian minister whose mathematical writings earned him a place as fellow of the Royal Society of London. His friend Richard Price found a manuscript among Bayes's effects after his death and had it published:

> BAYES, T. (1763) An Essay towards solving a Problem in the Doctrine of Chances, *Philosophical Transactions of the Royal Society* **53**, 370–418.

Bayes studied the following "inverse probability" (i.e. inference) problem: given the results of independent trials, $y_1, y_2, \ldots, y_n \overset{\text{(say)}}{=} 0, 1, 1, 0, 1, \ldots, 0$, what is the probability of success? Probability theory (started in the 1650's by Pascal and Fermat) was able to solve the "direct problem": given θ, the number of successes $s = \sum_{i=1}^{n} y_i$ has the distribution $s \,|\, \theta \sim \text{Binomial}(n, \theta)$, that is, $p(s \,|\, \theta) = \binom{n}{s} \theta^s (1-\theta)^{n-s}$. Bayes's solution to the inverse problem, which was also independently discovered in 1774 by Laplace, was:

1. specify a "prior" distribution $p(\theta)$ (Bayes and Laplace used $p(\theta) \equiv 1$);
2. calculate the "posterior" density

$$p(\theta \,|\, s) = \frac{p(\theta) p(s \,|\, \theta)}{\int_0^1 p(\theta) p(s \,|\, \theta) \, \mathrm{d}\theta};$$

3. use the posterior to answer specific questions about θ, for example,

$$\mathsf{P}(a < \theta < b \,|\, s) = \int_a^b p(\theta \,|\, s) \, \mathrm{d}\theta.$$

Because of this influential paper, Thomas Bayes's name is commemorated in the terms *Bayes's formula* (law, theorem) and *Bayesian statistics*.

1.2 The Fall and Rise of the Bayesians

Although the Bayesian approach to inference was extensively developed by Laplace and others starting in the late eighteenth century, in the twentieth century a completely different approach to inference, sometimes called Frequentist statistics, was developed and eventually came to dominate the field. This schism within the discipline of statistics has generated a great deal of polemical writing on both sides.

One of the historical drawbacks of Bayesian methods is that inference formulas exist only for relatively simple models. This is no longer a limitation since the development in the 1990's of effective computer algorithms and software that

allow the analysis of even very complex models. This development has led to a dramatic rise in the number of successful applications of Bayesian statistical methods in all areas of science and technology that continues to this day.

In this course we make extensive use of the free data modelling software `WinBUGS`. We explain the Markov Chain Monte Carlo (MCMC) computation method used by `WinBUGS` only in the later part of the course, after you've acquired experience in setting up statistical models.

2 Probability

2.1 Probability as a measure of belief

In Bayesian statistics, you use probability to represent degrees of belief (plausibility, confidence, credibility, certainty). The probability $\mathsf{P}(E\,|\,H)$ is a number that measures your belief in the truth of event E given the knowledge that H is true. It can reasonably be expected to obey the following axioms:

P1 $\mathsf{P}(E\,|\,H) \geq 0$

P2 $\mathsf{P}(H\,|\,H) = 1$

P3 $\mathsf{P}(E \cup F\,|\,H) = \mathsf{P}(E\,|\,H) + \mathsf{P}(F\,|\,H)$ when $E \cap F \cap H = \emptyset$

P4 $\mathsf{P}(E\,|\,F \cap H)\mathsf{P}(F\,|\,H) = \mathsf{P}(E \cap F\,|\,H)$

A sounder mathematical basis is obtained by strengthening axiom P3 to the not so intuitive axiom

P3* $\mathsf{P}(\cup_n E_n\,|\,H) = \sum_n \mathsf{P}(E_n\,|\,H)$ for countable $\{E_1, E_2, \ldots\}$ that are pairwise-disjoint given H, that is, $E_i \cap E_j \cap H = \emptyset$ whenever $i \neq j$.

From the axioms a number of results can be deduced; these too are intuitively reasonable properties for a system of plausible reasoning.

- If your information H implies that E is certainly true, then $\mathsf{P}(E\,|\,H) = 1$. To show this, first note that

$$\mathsf{P}(E\,|\,H) = \mathsf{P}(E\,|\,H \cap H) \cdot 1 \overset{P2}{=} \mathsf{P}(E\,|\,H \cap H)\mathsf{P}(H\,|\,H) \overset{P4}{=} P(E \cap H\,|\,H). \quad (1)$$

  Then $H \subseteq E$ implies $\mathsf{P}(E\,|\,H) \overset{(1)}{=} \mathsf{P}(E \cap H\,|\,H) \overset{H \subseteq E}{=} \mathsf{P}(H\,|\,H) \overset{P2}{=} 1$.

- The degree of belief $\mathsf{P}(E\,|\,H)$ is a number between 0 and 1. To show this, first note that

$$\begin{aligned} 1 &\overset{P2}{=} \mathsf{P}(H\,|\,H) = \mathsf{P}((H \setminus E) \cup (E \cap H)\,|\,H) \overset{P3}{=} \mathsf{P}(H \setminus E\,|\,H) + \mathsf{P}(E \cap H\,|\,H) \\ &\overset{(1)}{=} \mathsf{P}(H \setminus E\,|\,H) + \mathsf{P}(E\,|\,H). \end{aligned} \quad (2)$$

  Then $\mathsf{P}(E\,|\,H) \leq 1$ follows from the fact that $\mathsf{P}(H \setminus E\,|\,H) \overset{P1}{\geq} 0$, and $\mathsf{P}(E\,|\,H) \geq 0$ follows from P1.

- If your information H implies that E is certainly false then $\mathsf{P}(E\,|\,H) = 0$.

- If E implies F, given H (that is, if $E \cap H \subseteq F \cap H$) then $\mathsf{P}(E\,|\,H) \leq \mathsf{P}(F\,|\,H)$.

Note that in Bayesian theory there is no such thing as unconditional probability: all probabilities are conditional. However, in a given context where all probabilities are conditional on some generally accepted state of knowledge Ω, and all events under consideration are subsets of Ω, it is convenient to suppress Ω and write, for example, $\mathsf{P}(E)$ in place of $\mathsf{P}(E\,|\,\Omega)$.

2.2 How to assign probability

One way to determine (elicit) your degree of belief is to invite you to make a bet. Betting an amount M on event E at odds ω means that

- you lose (pay) M if E turns out to be false, and

- you win (receive) $\omega \cdot M$ if E turns out to be true.

If you believe strongly in E, then you are willing to accept small odds, whereas you will insist on large odds if you consider E to be doubtful. You consider the odds to be *fair* if, in your estimation, there is no advantage in betting for or against the proposition. Fair odds given a state of knowledge H is denoted $\tilde{\omega}(H)$ and satisfies

$$\mathsf{P}(E\,|\,H) \cdot \tilde{\omega}(H) M = \mathsf{P}(\bar{E}\,|\,H) \cdot M, \tag{3}$$

where $\bar{E} = \Omega \setminus E$ is the complement of E. Solving (3) gives $\mathsf{P}(E\,|\,H) = \frac{1}{1+\tilde{\omega}(H)}$. Thus, at least in theory, the value of $\mathsf{P}(E\,|\,H)$ can be deduced from what you consider to be a fair bet. In practice, however, people are not mathematically consistent in their evaluations of probabilities!

A more familiar way to determine probability is to use symmetry principles to determine "equally probable" events. For example, if we known that an urn contains r red balls and k black balls, then we assign the probability of $E =$ "a red ball is drawn" as

$$\mathsf{P}(E\,|\,(r,k)) = \frac{r}{r+k}.$$

when we have no information that would favour any ball over any other ball. Note that this does not assume any physical "randomness" (whatever that means) in the drawing process, it only models the symmetry that exists in our state of knowledge (and ignorance).

2.3 Data changes probability

Suppose that after assigning a value to $\mathsf{P}(E\,|\,H)$, you obtain new information F. Then your degree of belief in E is updated to $\mathsf{P}(E\,|\,(F \cap H))$ (henceforth written as $\mathsf{P}(E\,|\,F,H)$), which can differ from your earlier degree of belief.

Example: Inferring bias Suppose you have a lapel pin with a convex face and a flat back. When the pin is spun on a flat surface, it comes to rest with the face upwards (this outcome is denoted $r = 0$) or back upwards ($r = 1$). Given this information (H), and before performing any experiment, you might believe that $\mathsf{P}(r = 1 \,|\, H)$ is, say, $\frac{1}{2}$. Then you perform the experiment ten times and obtain the results

$$r_{1:10} = [0, 0, 0, 0, 0, 1, 1, 0, 0, 1]$$

Because of the small number of outcomes with $r = 1$, you may think it reasonable to update your belief such that $\mathsf{P}(r = 1 \,|\, r_{1:10}, H)$ is a number that is smaller than $\frac{1}{2}$. We'll see later how to do this update.

2.4 Odds

The odds on E against F given H are defined as the ratio

$$\frac{\mathsf{P}(E \,|\, H)}{\mathsf{P}(F \,|\, H)} \text{ to } 1,$$

or equivalently

$$\mathsf{P}(E \,|\, H) \text{ to } \mathsf{P}(F \,|\, H).$$

In some contexts, this is a more natural concept than probability, and can be easier to define. For example, in the example in §2.3, the prior odds on $r = 1$ against $r = 0$ are 1 to 1. Odds are widely used in betting, decision theory, and risk analysis, and we'll encounter them in chapter 8 in the context of hypothesis testing.

2.5 Independence

Events E and F are said to be *conditionally independent given* H if

$$\mathsf{P}(E \cap F \,|\, H) = \mathsf{P}(E \,|\, H) \cdot \mathsf{P}(F \,|\, H).$$

In case the events are independent given a common state of knowledge Ω, we say the events are independent and write

$$\mathsf{P}(E \cap F) = \mathsf{P}(E) \cdot \mathsf{P}(F).$$

By axiom P4, equivalent characterisations of independence are $\mathsf{P}(E \,|\, F) = \mathsf{P}(E)$ (when $\mathsf{P}(F) \neq 0$), $\mathsf{P}(E \,|\, \bar{F}) = \mathsf{P}(E)$ (when $\mathsf{P}(\bar{F}) \neq 0$), $\mathsf{P}(F \,|\, E) = \mathsf{P}(F)$ (when $\mathsf{P}(E) \neq 0$), and $\mathsf{P}(F \,|\, \bar{E}) = \mathsf{P}(F)$ (when $\mathsf{P}(\bar{E}) \neq 0$). Loosely speaking, events E and F are independent if knowledge of one of the events does not change your degree of belief in the other. In other words, you learn nothing about E from F, and vice versa.

A set of events is said to be *mutually independent* given H if

$$\mathsf{P}(E_{i_1} \cap E_{i_2} \cap \cdots \cap E_{i_k} \,|\, H) = \mathsf{P}(E_{i_1} \,|\, H)\mathsf{P}(E_{i_2} \,|\, H) \cdots \mathsf{P}(E_{i_k} \,|\, H)$$

for any finite subset of them.

Example: Balls and Urns This example brings out some subtleties about independence. Consider an urn with r red balls and k black balls, from which we draw balls one at a time, with replacement. Let $E =$ "the first drawn ball is red" and $F =$ "the second drawn ball is red".

If we know the proportion of red balls $\theta = r/(r+k)$, then (following classical probability theory) we have $\mathsf{P}(F \mid \theta) = \theta$ and $\mathsf{P}(F \mid E, \theta) = \theta$, that is, events E and F are *conditionally independent given* θ.

If we don't know θ, then we could treat it (i.e. *model* it) as a random variable. Then E and F are not independent, because the result of the first draw provides information about the proportion of red balls in the urn, and this can change our state of belief about the result of the second draw. In particular, we have

$$\mathsf{P}(E) = \int_0^1 g(\theta) \underbrace{\mathsf{P}(E \mid \theta)}_{\theta} \, \mathrm{d}\theta = \mathsf{E}(\theta),$$

where g is the density function of θ, and

$$\mathsf{P}(E \cap F) = \int_0^1 g(\theta) \underbrace{\mathsf{P}(E \cap F \mid \theta)}_{\mathsf{P}(E \mid \theta)\mathsf{P}(F \mid \theta)} \, \mathrm{d}\theta = \mathsf{E}(\theta^2) = \mathsf{V}(\theta) + \mathsf{E}(\theta)^2,$$

and so

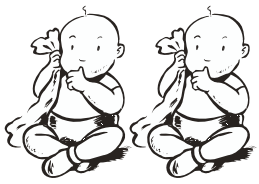
$$\mathsf{P}(F \mid E) = \frac{\mathsf{P}(E \cap F)}{\mathsf{P}(E)} = \frac{\mathsf{V}(\theta) + \mathsf{E}(\theta)^2}{\mathsf{E}(\theta)} = \frac{\mathsf{V}(\theta)}{\mathsf{E}(\theta)} + \mathsf{P}(F) \geq \mathsf{P}(F).$$

Thus, the more prior uncertainty you have about θ, the more you learn about F from E, and you learn nothing if and only if $\mathsf{V}(\theta) = 0$.

2.6 Bayes's formula

Let $H_1, H_2, \ldots$ be a partition of Ω, that is, $H_i \cap H_j = \emptyset$ for all $i \neq j$, and $\cup_n H_n = \Omega$. Then for any event E we have (using axioms P4 and P3*) the *total probability formula*

$$\mathsf{P}(E) = \sum_n \mathsf{P}(E \mid H_n)\mathsf{P}(H_n).$$



Example: Probability of twin girls[1] Suppose that girls are equally likely to be born as boys. When twins are born, what is the probability that both babies are girls? The answer, surprisingly enough, is not $\frac{1}{4}$. Here's why.

Twins can be identical (monozygotic, denoted M) or fraternal (dizygotic, D). Identical twins are always of the same sex. In light of this, a reasonable model (where we denote GB for twins of different sex, GG for twin girls, and BB for twin boys) for the birth of a pair of twins is

$$\begin{aligned} \mathsf{P}(GG \mid M) = \mathsf{P}(BB \mid M) &= \tfrac{1}{2}, \quad \mathsf{P}(GB \mid M) = 0 \\ \mathsf{P}(GG \mid D) = \mathsf{P}(BB \mid D) &= \tfrac{1}{4}, \quad \mathsf{P}(GB \mid D) = \tfrac{1}{2} \end{aligned}$$

[1] taken from Peter M. Lee, *Bayesian Statistics: An Introduction*, 2nd ed., 1997.

Then the total probability formula gives

$$\begin{aligned} \mathsf{P}(GG) &= \mathsf{P}(GG\,|\,M)\mathsf{P}(M) + \mathsf{P}(GG\,|\,D)\mathsf{P}(D) \\ &= \tfrac{1}{2}\mathsf{P}(M) + \tfrac{1}{4}(1-\mathsf{P}(M)) \\ &= \tfrac{1}{4}(\mathsf{P}(M)+1) \\ &> \tfrac{1}{4} \end{aligned}$$

because $\mathsf{P}(M) > 0$. ■

Because

$$\mathsf{P}(H_n\,|\,E)\mathsf{P}(E) = \mathsf{P}(E \cap H_n) = \mathsf{P}(H_n)\mathsf{P}(E\,|\,H_n),$$

then for $\mathsf{P}(E) \neq 0$ we have *Bayes's formula*

$$\mathsf{P}(H_n\,|\,E) = \frac{\mathsf{P}(H_n)\mathsf{P}(E\,|\,H_n)}{\sum_k \mathsf{P}(E\,|\,H_k)\mathsf{P}(H_k)}.$$

Bayes's formula can be written concisely as

$$\underbrace{\mathsf{P}(H_n\,|\,E)}_{\text{posterior}} \propto \underbrace{\mathsf{P}(H_n)}_{\text{prior}}\underbrace{\mathsf{P}(E\,|\,H_n)}_{\text{likelihood}},$$

with the constant of proportionality being $1/\mathsf{P}(E)$.

Bayes's formula leads directly to the following *model comparison* formula for the odds in favour of H_j against H_k given E:

$$\underbrace{\frac{\mathsf{P}(H_j\,|\,E)}{\mathsf{P}(H_k\,|\,E)}}_{\text{posterior odds}} = \underbrace{\frac{\mathsf{P}(H_j)}{\mathsf{P}(H_k)}}_{\text{prior odds}} \times \underbrace{\frac{\mathsf{P}(E\,|\,H_j)}{\mathsf{P}(E\,|\,H_k)}}_{\text{Bayes ratio}}.$$

3 Normal Data

A simple model that relates a random variable θ and real-valued observations $y_1, \ldots, y_n$ is the following:

- the observations are mutually independent conditional on θ,
- the observations are identically normally distributed with mean θ, that is, $y_i\,|\,\theta \sim \text{Normal}(\theta, v)$, where the variance v is a known constant.

With this model, the joint distribution of the observations given θ has the pdf

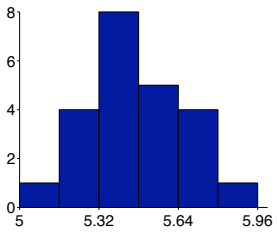
$$p(y_{1:n}\,|\,\theta) = \left(\frac{1}{2\pi v}\right)^{n/2} \mathrm{e}^{-\frac{1}{2v}\sum_{i=1}^{n}(y_i-\theta)^2}. \tag{4}$$

We shall often refer to this pdf as the *likelihood*, even though some statisticians reserve this name for the function $\theta \mapsto p(y_{1:n}\,|\,\theta)$, whose values are denoted $\text{lhd}(\theta; y_{1:n})$.

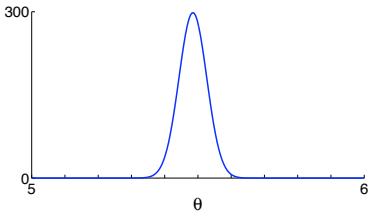
Notice that $p(y_1, \ldots, y_n\,|\,\theta) = p(y_{i_1}, \ldots, y_{i_n}\,|\,\theta)$ for any permutation of the order of the observations; a sequence of random variables with this property is said to be *exchangeable*.

Example 3.1: Cavendish's data The English physicist Henry Cavendish performed experiments in 1798 to measure the specific density of the earth. The results of 23 experiments are

```
5.36   5.29   5.58   5.65   5.57
5.53   5.62   5.29   5.44   5.34
5.79   5.10   5.27   5.39   5.42
5.47   5.63   5.34   5.46   5.30
5.78   5.68   5.85
```



Using the simple normal model described above, the function $\text{lhd}(\theta; y_{1:n})$ when the assumed variance is $v = 0.04$ is:



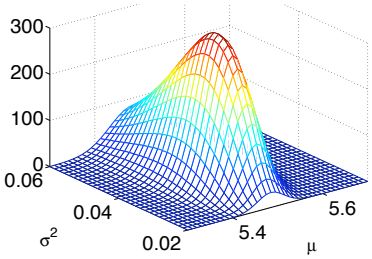
In the above model we've assumed for simplicity that the variance is known. A slightly more complicated model has two parameters $\theta = (\mu, \sigma^2)$:

- the observations are mutually independent conditional on (μ, σ^2),
- $y_i \,|\, \mu, \sigma^2 \sim \text{Normal}(\mu, \sigma^2)$.

With this model, the joint distribution of the data given (μ, σ^2) has the density

$$p(y_{1:n} \,|\, \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \mathrm{e}^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2}.$$

Here's a plot of $\text{lhd}(\mu, \sigma^2; y_{1:n})$ for the Cavendish data.



4 Posteriors, Priors, and Predictive Distributions

Our basic inference tool is Bayes's theorem in the form

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

where $p(y \,|\, \theta)$ is the data model, $p(\theta)$ is the probability density describing our state of knowledge about θ before we've received observation values (the *prior* density), and $p(\theta \,|\, y)$ describes our state of knowledge taking account of the observations (the *posterior* density).

4.1 Using the Posterior

Compared to the inference machinery of classical statistics, with its p-values, confidence intervals, significance levels, bias, etc., Bayesian inference is straightforward: the inference result is the posterior, which is a probability distribution. Various descriptive statistical tools are available to allow you to study interesting aspects of the posterior distribution.

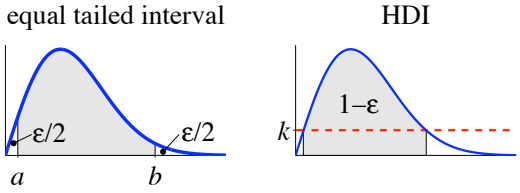
Plots You can plot the posterior density of $\theta = [\theta_1, \ldots, \theta_p]$ if $p = 1$ or $p = 2$, while for $p \geq 3$ you can plot densities of marginal distributions.

Credibility regions A credibility (or Bayesian confidence) region is a subset C_ε in parameter space such that

$$\mathsf{P}(\theta \in C_\varepsilon \,|\, y) = 1 - \varepsilon, \tag{5}$$

where ε is a predetermined value (typically 5%) that is considered to be an acceptable level of error. This definition does not specify C_ε uniquely. In the case of a single parameter, the most common ways of determining a credibility region (credibility interval, in this case) are to seek either

- the shortest interval,
- an equal tailed interval, that is, an interval $[a,b]$ such that $\mathsf{P}(\theta \leq a \,|\, y) = \frac{\varepsilon}{2}$ and $\mathsf{P}(\theta \geq b \,|\, y) = \frac{\varepsilon}{2}$, or



- the highest density interval (HDI), which is the set of θ points where posterior density values are higher than at points outside the region, that is,

  $$C_\varepsilon = \{\theta : p(\theta \,|\, y) \geq k\},$$

  where k is determined by the constraint (5).

The HDI definition is readily generalised to HDR (highest density region) for multiple parameters.

Hypothesis testing The probability that a hypothesis such as $H : \theta > 0$ is true[2] is, at least conceptually, easily computed from the posterior:

$$\mathsf{P}(H \,|\, y) = \mathsf{P}(\theta > 0 \,|\, y) = \int_0^\infty p(\theta \,|\, y)\, \mathrm{d}\theta.$$

[2] Orthodox "frequentist" statistics has something called a p-value that is often mistakenly interpreted by students as a probability. It is defined as "the maximum probability, consistent with H being true, that evidence against H as least as strong as that provided by the data would occur by chance".

Point estimates The posterior is a complete description of your state of knowledge about θ, so in this sense the distribution *is* the estimate, but in practical situations you often want to summarize this information using a single number for each parameter. Popular alternatives are:

- the mode, $\arg\max_\theta p(\theta \mid y)$,
- the median, $\arg\min_t \mathsf{E}(|\theta - t| \mid y)$,
- the mean, $\mathsf{E}(\theta \mid y) = \int \theta\, p(\theta \mid y)\, \mathrm{d}\theta = \arg\min_t \mathsf{E}((\theta - t)^2 \mid y)$.

You can also augment this summary with some measure of dispersion, such as the posterior's variance. The posterior mode is also known as the *maximum a posteriori* (MAP) estimator.

4.2 Bayesian Data Analysis with Normal Models

Consider the one-parameter normal data model presented in section 3, in which the observations are assumed to be mutually independent conditional on the θ and identically distributed with $y_i \mid \theta \sim \text{Normal}(\theta, v)$, where the variance v is a known constant. With these assumptions, the distribution $p(y_{1:n} \mid \theta)$ is given by (4), which can be written

$$p(y_{1:n} \mid \theta) \propto \exp\left(-\frac{1}{2v}\sum_{i=1}^{n}(y_i - \theta)^2\right).$$

What prior do we choose? A convenient choice (because it gives integrals that we can solve in closed form!) is a normal distribution $\theta \sim \text{Normal}(m_0, w_0)$, with m_0 and w_0 chosen such that the prior distribution is a sufficiently accurate representation of our state of knowledge. Then we have

$$p(\theta) = \frac{1}{\sqrt{2\pi w_0}} \mathrm{e}^{-\frac{(\theta - m_0)^2}{2w_0}} \propto \exp\left(-\frac{1}{2w_0}(\theta - m_0)^2\right).$$

By Bayes's law, the posterior is

$$p(\theta \mid y) \propto p(\theta) p(y \mid \theta) \propto \exp\left(-\frac{1}{2v}\sum_{i=1}^{n}(y_i - \theta)^2 - \frac{1}{2w_0}(\theta - m_0)^2\right) = \mathrm{e}^{-\frac{1}{2}Q},$$

where (using completion of squares)

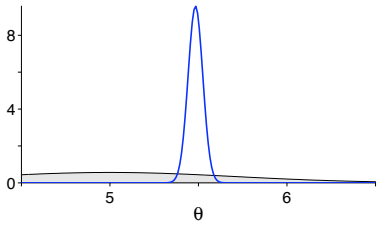
$$Q = \left(\frac{1}{w_0} + \frac{n}{v}\right)\left(\theta - \frac{\frac{m_0}{w_0} + \frac{n\bar{y}}{v}}{\frac{1}{w_0} + \frac{n}{v}}\right)^2 + \text{constant}.$$

and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ denotes the sample mean. Thus the posterior is $\theta \mid y \sim \text{Normal}(m_n, w_n)$ with

$$m_n = \frac{\frac{m_0}{w_0} + \frac{n\bar{y}}{v}}{\frac{1}{w_0} + \frac{n}{v}}, \quad w_n = \frac{1}{\frac{1}{w_0} + \frac{n}{v}}.$$

We are able to find a closed-form solution in this case because we chose a prior of a certain form (a *conjugate* prior) that facilitates symbolic manipulations in Bayesian analysis. For other priors the computations (normalisation factor $p(y) = \int p(\theta) p(y \mid \theta)\, \mathrm{d}\theta$, credibility interval, etc.) generally have to be done using approximations or numerical methods.

Example: Cavendish's data (continued) Knowing that rocks have specific densities between 2.5 (granite) and 7.5 (lead ore), let's choose a prior that is relatively broad and centred at 5, say $\theta \sim \text{Normal}(5, 0.5)$ (shown grey-filled on the right). If we assume a data model with conditionally independent observations that are normally distributed $y_i \sim \text{Normal}(\theta, v)$ with $v = 0.04$, we obtain a posterior $\theta \,|\, y \sim \text{Normal}(m_{23}, w_{23})$ with



$$m_{23} = \frac{\frac{5}{0.5} + \frac{23 \cdot 5.485}{0.04}}{\frac{1}{0.5} + \frac{23}{0.04}} = 5.483, \quad w_{23} = \frac{1}{\frac{1}{0.5} + \frac{23}{0.04}} = (0.0416)^2 = 0.00173.$$

From this posterior (the unfilled curve) we obtain the 95% equal-tail credibility interval $(5.4015, 5.5647)$.

Here's a WinBUGS[3] model of this example:

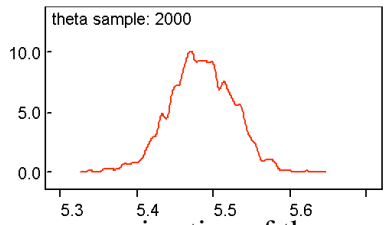
```
model {
   for (i in 1:n) { y[i] ~ dnorm(theta,25) }
   theta ~ dnorm(5,2)
}
```

Here `dnorm` means a normal distribution and its two arguments are the mean and the *precision* (the reciprocal of variance). The precision values we entered are $25 = 1/0.04$ and $2 = 1/0.5$.

The data are coded as

```
list( y=c(5.36,5.29,5.58,5.65,5.57,5.53,5.62,5.29,5.44,5.34,
          5.79,5.10,5.27,5.39,5.42,5.47,5.63,5.34,5.46,5.30,
          5.78,5.68,5.85),  n=23 )
```

and the simulation initial value is generated by pressing the `gen inits` button. The simulation is then run for 2000 steps; the resulting statistics agree well with the analytical results:



node	mean	sd	2.5%	median	97.5%
theta	5.484	0.0414	5.402	5.484	5.565

The smoothed histogram of the simulated theta values is an approximation of the posterior distribution.

4.3 Using Bayes's Formula Sequentially

Suppose you have two observations y_1 and y_2 that are conditionally independent given θ, that is, $p(y_1, y_2 \,|\, \theta) = p_1(y_1 \,|\, \theta) p_2(y_2 \,|\, \theta)$. The posterior is then

$$p(\theta \,|\, y_1, y_2) \propto p(\theta) p(y_1, y_2 \,|\, \theta) = \underbrace{p(\theta) p_1(y_1 \,|\, \theta)}_{\propto p(\theta \,|\, y_1)} p_2(y_2 \,|\, \theta).$$

[3] Download it from `http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml` Watch the demo at `http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/winbugsthemovie.html`

From this formula we see that the new posterior $p(\theta \,|\, y_1, y_2)$ can be found in two steps:

1. Use the first observation to update the prior $p(\theta)$ via Bayes's formula to $p(\theta \,|\, y_1) \propto p(\theta) p_1(y_1 \,|\, \theta)$, then

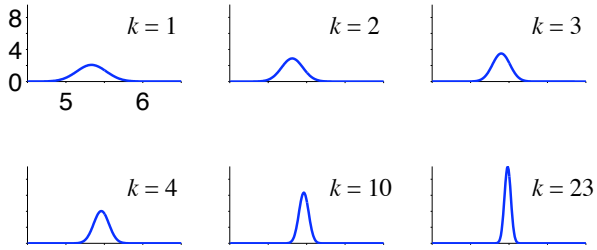
2. use the second observation to update $p(\theta \,|\, y_1)$ via $p(\theta \,|\, y_1, y_2) \propto p(\theta \,|\, y_1) p_2(y_2 \,|\, \theta)$.

This idea can obviously be extended to process a data sequence, recursively updating the posterior one observation at a time. Also, the observations can be processed in any order.

In particular, for conditionally independent normal data $y_i \sim \text{Normal}(\theta, v)$ and prior $\theta \sim \text{Normal}(m_0, w_0)$, the posterior distribution for the first k observations is $\theta \,|\, y_1, \ldots, y_k \sim \text{Normal}(m_k, w_k)$, with the posterior mean and variance obtained via the recursion

$$w_k = \frac{1}{\frac{1}{w_{k-1}} + \frac{1}{v}}, \quad m_k = \left(\frac{m_{k-1}}{w_{k-1}} + \frac{y_k}{v} \right) w_k.$$

Example: Cavendish's data (continued) Here is how the posterior pdf $p(\theta \,|\, y_{1:k})$ evolves as the dataset is processed sequentially:



4.4 Predictive Distributions

Before an observation (scalar or vector) y is measured or received, it is an unknown quantity — let's denote it $\tilde{y}$. Its distribution is called the *prior predictive distribution* or *marginal distribution of data*. The density can be computed from the likelihood and the prior:

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

The predictive distribution is defined in the data space $\mathscr{Y}$, as opposed to the parameter space Θ, which is where the prior and posterior distributions are defined. The prior predictive distribution can be used to assess the validity of the model: if the prior predictive density "looks wrong", you need to re-examine your prior and likelihood!

In a similar fashion, after observations $y_1, \ldots, y_n$ are received and processed, the next observation is an unknown quantity, which we also denote $\tilde{y}$. Its distribution is called the *posterior predictive distribution* and the density can be computed from

$$p(\tilde{y} \,|\, y_{1:n}) = \int p(\tilde{y} \,|\, \theta, y_{1:n}) p(\theta \,|\, y_{1:n}) \, \mathrm{d}\theta.$$

If $\tilde{y}$ is independent of $y_{1:n}$ given θ, then the formula simplifies to

$$p(\tilde{y}\,|\,y_{1:n}) = \int p(\tilde{y}\,|\,\theta)p(\theta\,|\,y_{1:n})\,\mathrm{d}\theta.$$

In particular, when the data are modelled as conditionally mutually independent $y_i\,|\,\theta \sim \text{Normal}(\theta, v)$ and the prior is $\theta \sim \text{Normal}(m_0, v_0)$, the posterior predictive distribution for a new observation (given n observations) can be found by the following trick. Noting that $\tilde{y} = (\tilde{y}-\theta)+\theta$ is the sum of normally distributed random variables and that $(\tilde{y}-\theta)\,|\,\theta, y_{1:n} \sim \text{Normal}(0, v)$ and $\theta\,|\,y_{1:n} \sim \text{Normal}(m_n, w_n)$ are independent given $y_{1:n}$, we deduce that $\tilde{y}\,|\,y_{1:n}$ is normally distributed with

$$\begin{aligned}
\mathsf{E}(\tilde{y}\,|\,y_{1:n}) &= \mathsf{E}(\tilde{y}-\theta\,|\,\theta, y_{1:n}) + \mathsf{E}(\theta\,|\,y_{1:n}) = 0 + m_n = m_n \\
\mathsf{V}(\tilde{y}\,|\,y_{1:n}) &= \mathsf{V}(\tilde{y}-\theta\,|\,\theta, y_{1:n}) + \mathsf{V}(\theta\,|\,y_{1:n}) = v + w_n,
\end{aligned}$$

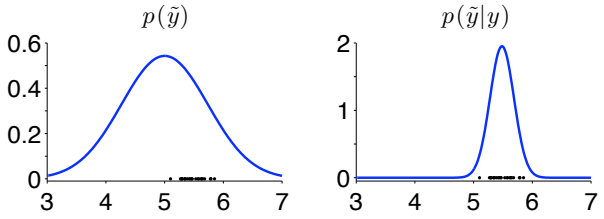
that is, $\tilde{y}\,|\,y_{1:n} \sim \text{Normal}(m_n, v+w_n)$. This result holds also for $n=0$, that is, the prior predictive distribution is $\tilde{y} \sim \text{Normal}(m_0, v+w_0)$.

The above formulas can alternatively be derived using the formulas

$$\begin{aligned}
\mathsf{E}(\tilde{y}\,|\,y_{1:n}) &= \mathsf{E}_{\theta|y_{1:n}}(\mathsf{E}(\tilde{y}\,|\,\theta, y_{1:n})), \\
\mathsf{V}(\tilde{y}\,|\,y_{1:n}) &= \mathsf{E}_{\theta\,|\,y_{1:n}}(\mathsf{V}(\tilde{y}\,|\,\theta, y_{1:n})) + \mathsf{V}_{\theta\,|\,y_{1:n}}(\mathsf{E}(\tilde{y}\,|\,\theta, y_{1:n}))
\end{aligned}$$

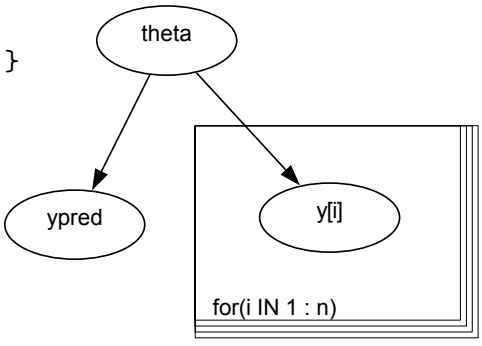
Details are left as an exercise.

Example: Cavendish's data (continued) Here the prior predictive distribution is $\tilde{y} \sim \text{Normal}(5, 0.54)$ and the posterior predictive distribution given 23 observations is $\tilde{y}\,|\,y_{1:23} \sim \text{Normal}(5.483, 0.0417) = \text{Normal}(5.483, (0.2043)^2)$.



The predictive distribution can be computed in WinBUGS with

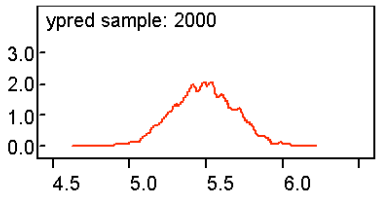
```
model {
   for (i in 1:n) { y[i] ~ dnorm(theta,25) }
   theta ~ dnorm(5,2)
   ypred ~ dnorm(theta,25)
}
```



The diagram on the right is the DAG (directed acyclic graph) representation of the model, drawn with DoodleBUGS. The ellipses denote stochastic nodes (variables that have a probability distribution), and the directed edges (arrows) indicate conditional dependence. Repeated parts of the graph are contained in a loop construct called a "plate".

The results after 2000 simulation steps are

node	mean	sd	2.5%	median	97.5%
theta	5.483	0.04124	5.402	5.484	5.564
ypred	5.484	0.2055	5.093	5.486	5.88



5 Single-Parameter Models

In this section we look at some standard one-parameter models whose posterior and predictive distributions can be found in closed form. This is accomplished by choosing priors that are *conjugate* to the likelihood. A family $\mathscr{C}$ of distributions is said to be conjugate to a likelihood distribution if, for every prior chosen from $\mathscr{C}$, the posterior also belongs to $\mathscr{C}$. In practice, conjugate families are parametrised, and by choosing appropriate parameter values you can usually obtain a distribution that is an acceptable model of your prior state of knowledge.

The conjugate families and inference solutions that will be presented in this section are summarised below.

$y_i \mid \theta \sim$	$\theta \sim$	$\theta \mid y_{1:n} \sim$	$\tilde{y} \mid y_{1:n} \sim$
$\text{Normal}(\theta, v)$	$\text{Normal}(m_0, w_0)$	$\text{Normal}(m_n, w_n)$	$\text{Normal}(m_n, v + w_n)$
$\text{Binomial}(1, \theta)$	$\text{Beta}(\alpha, \beta)$	$\text{Beta}(\alpha + s, \beta + n - s)$	$\text{Binomial}(1, \frac{\alpha+s}{\alpha+\beta+n})$
$\text{Poisson}(\theta)$	$\text{Gamma}(\alpha, \beta)$	$\text{Gamma}(\alpha + s, \beta + n)$	$\text{NegBin}(\alpha + s, \beta + n)$
$\text{Exp}(\theta)$	$\text{Gamma}(\alpha, \beta)$	$\text{Gamma}(\alpha + n, \beta + s)$	
$\text{Normal}(m, \theta)$	$\text{InvGam}(\alpha, \beta)$	$\text{InvGam}(\alpha + \frac{n}{2}, \beta + \frac{n}{2} s_0^2)$	

$s = \sum_{i=1}^n y_i$, $\bar{y} = s/n$, $\frac{1}{w_n} = \frac{1}{w_0} + \frac{1}{v/n}$, $m_n = (\frac{m_0}{w_0} + \frac{\bar{y}}{v/n}) w_n$, $s_0^2 = \frac{1}{n} \sum_{i=1}^n (y_i - m)^2$

The properties of the distributions are summarised below.

$x \sim$	$p(x)$	$x \in$	$\mathsf{E}(x)$	$\text{mode}(x)$	$\mathsf{V}(x)$
$\text{Normal}(\mu, \sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$\mathbb{R}$	μ	μ	σ^2
$\text{Binomial}(n, p)$	$\binom{n}{x} p^x (1-p)^{n-x}$	$\{0, 1, \ldots, n\}$	np	$\lfloor (n+1)p \rfloor$	$np(1-p)$
$\text{Beta}(\alpha, \beta)$	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$	$[0, 1]$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha-1}{\alpha+\beta-2}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
$\text{Poisson}(\lambda)$	$\frac{1}{x!} \lambda^x \mathrm{e}^{-\lambda}$	$\{0, 1, \ldots\}$	λ	$\lfloor \lambda \rfloor$	λ
$\text{Gamma}(\alpha, \beta)$	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \mathrm{e}^{-\beta x}$	$(0, \infty)$	$\frac{\alpha}{\beta}$	$\frac{\alpha-1}{\beta}$	$\frac{\alpha}{\beta^2}$
$\text{Exp}(\lambda)$	$\lambda \mathrm{e}^{-\lambda x}$	$(0, \infty)$	$\frac{1}{\lambda}$	0	$\frac{1}{\lambda^2}$
$\text{NegBin}(\alpha, \beta)$	$\binom{x+\alpha-1}{\alpha-1} (\frac{\beta}{\beta+1})^\alpha (\frac{1}{\beta+1})^x$	$\{0, 1, \ldots\}$	$\frac{\alpha}{\beta}$		$\frac{\alpha}{\beta^2}(\beta+1)$
$\text{InvGam}(\alpha, \beta)$	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} \mathrm{e}^{-\beta/x}$	$(0, \infty)$	$\frac{\beta}{\alpha-1}$	$\frac{\beta}{\alpha+1}$	$\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$

$\mu \in \mathbb{R}$, $\sigma > 0$, $\alpha > 0$, $\beta > 0$, $\lambda > 0$, $n \in \{1, 2, \ldots\}$, $p \in [0, 1]$

5.1 Estimating the mean of a normal likelihood

This is the situation that was considered in sections 3 and 4: we have real-valued observations $y_1, \ldots, y_n$ that are assumed to be mutually independent given θ and identically normally distributed with unknown mean θ and known variance v. The likelihood is thus

$$p(y_{1:n} \mid \theta) = \left(\frac{1}{2\pi v}\right)^{n/2} \mathrm{e}^{-\frac{1}{2v}\sum_{i=1}^n (y_i - \theta)^2}. \tag{6}$$

For a normal prior $\theta \sim \text{Normal}(m_0, w_0)$, we showed earlier that the posterior is $\theta \,|\, y \sim \text{Normal}(m_n, w_n)$ with

$$m_n = \frac{\frac{m_0}{w_0} + \frac{n\bar{y}}{v}}{\frac{1}{w_0} + \frac{n}{v}}, \quad w_n = \frac{1}{\frac{1}{w_0} + \frac{n}{v}}.$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. The posterior belongs to the same family of distributions as the prior, that is, the family of normal distributions is conjugate to the likelihood (6).

The posterior mean can be written as the weighted average of the prior mean m_0 and the sample mean $\bar{y}$:

$$m_n = \frac{w_n}{w_0} \cdot m_0 + \frac{w_n}{v/n} \cdot \bar{y}$$

The weights reflect the relative importance given to the prior and the observations. The ratio of the weights, $\frac{1}{w_0} : \frac{1}{v/n}$, is the ratio of the precisions (reciprocal variances) of the prior and sample mean.

Another way of writing the posterior mean formula is

$$m_n = \bar{y} - (\bar{y} - m_0)\frac{v/n}{w_0 + v/n},$$

which shows "shrinkage" as the posterior mean m_n is pulled away from the sample mean $\bar{y}$ toward the prior mean m_0 ().

These alternative formulas indicate how the posterior mean is a compromise between prior beliefs and observations. We see that as the number of observations grows, $m_n \to \bar{y}$, regardless of the prior mean and variance. Thus, given enough data, the choice of prior becomes irrelevant.

When the prior is diffuse relative to the precision of the observations (e.g. a normal prior with $w_0 \gg \frac{v}{n}$), then $m_n \approx \bar{y}$ and $w_n \approx \frac{v}{n}$, and the 95% credibility interval can be approximated as $\bar{y} \pm 1.96\sqrt{\frac{v}{n}}$.

If prior information about the location of the mean is rather weak, one can consider using a constant prior

$$p(\theta) \propto \kappa \quad \text{(a constant)}.$$

This of course does not describe a probability density: a proper density satisfies $\int_{-\infty}^{\infty} p(\theta)\,\mathrm{d}\theta = 1$, and with a constant prior this integral diverges for $\kappa > 0$ and is equal to zero for $\kappa = 0$. However, this *improper* prior can sometimes be a convenient shortcut: substituting it into Bayes's rule produces the posterior

$$p(\theta \,|\, y) \propto \kappa p(y \,|\, \theta) \propto \exp\left(-\frac{1}{2v}\sum_{i=1}^{n}(y_i - \theta)^2\right)$$

which is a proper density function, because the integral

$$\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2v}\sum_{i=1}^{n}(y_i - \theta)^2\right)\mathrm{d}\theta$$

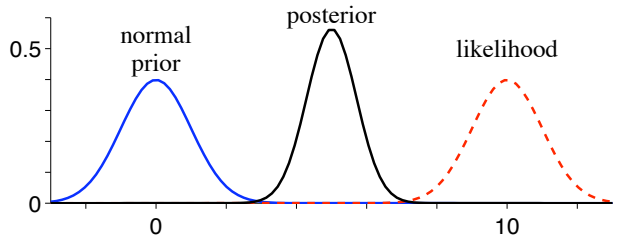
is convergent. The posterior that we obtain is

$$\theta \,|\, y \sim \text{Normal}(\bar{y}, \frac{v}{n}),$$

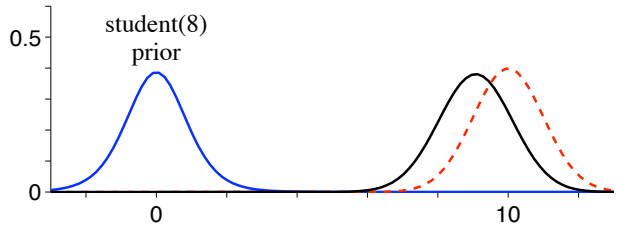
which the same distribution as the limiting case $w_0 \to \infty$ of a posterior that is based on the proper prior $\theta \sim \text{Normal}(m_0, w_0)$. One should in general however be careful about using improper priors, because they can in some cases lead to improper posteriors. Also, they

can lead to difficulties in hypothesis testing (and decision theory in general). Improper priors cannot be used in WinBUGS.

Situations can arise where there is a serious conflict between the observations and the prior, that is, the likelihood is small whenever the prior is large, and vice versa. In such a situation, the product of the prior and likelihood is small for all θ, and the shape of the posterior is determined by the tails of the prior and likelihood. When prior and likelihood conflict, the posterior can be quite sensitive to the shapes of the tails. For example, if the prior is $\theta \sim \text{Normal}(0,1)$ and the likelihood is $y \,|\, \theta \sim \text{Normal}(\theta,1)$, then the observation $y = 10$ produces a posterior whose mean is halfway between the observation and the prior's zero mean:



Replacing the standard normal prior by an 8-degree of freedom Student-t distribution[4] with the same mean and variance causes a dramatic shift in the posterior toward the likelihood:



Generally, in case of conflict the posterior will be dominated by the distribution with the lighter tail.

The Student-t distribution is often used as a heavy-tailed alternative to the normal distribution for both the likelihood (because extreme values typically show up in real data more than the normal distribution predicts) and for the prior (to ensure that observations will dominate in information-conflict situations). Summary values (mean, variance, HDI, etc.) for heavy-tailed model posteriors usually have to be computed numerically.

5.2 Estimating the probability parameter in a binomial model

A *Bernoulli trial* has two possible results, conventionally termed "success" and "failure", with the probability of success denoted θ. The result of an observation is represented as $y_i = 1$ for success and $y_i = 0$ for failure. The problem (first considered by Bayes and Laplace) is to infer θ from a set of such observations.

The likelihood pmf for a single observation is

$$p(y_i \,|\, \theta) = \left\{ \begin{array}{ll} \theta & (y_i = 1) \\ 1-\theta & (y_i = 0) \end{array} \right\} = \theta^{y_i}(1-\theta)^{1-y_i} \quad (y_i \in \{0,1\}),$$

[4] The pdf for the Student-t distribution with ν degrees of freedom $x \sim \text{t}_\nu(\mu,\sigma^2)$ is $p(x) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi\sigma^2}}\left(1+\frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right)^{-(\nu+1)/2}$, and $\mathsf{E}(x) = \mathsf{mode}(x) = \mu$ for $\nu > 1$, $\mathsf{V}(x) = \frac{\nu}{\nu-2}\sigma^2$ for $\nu > 2$. The standard Student-t distribution $\text{t}_\nu(0,1)$ is denoted t_ν; $x \sim \text{t}_\nu(\mu,\sigma^2)$ implies $(x-\mu)/\sigma \sim \text{t}_\nu$. The distribution is named for the pseudonym of its author.

where 0^0 is taken to be equal to 1. The likelihood pmf of a sequence $y_1, \ldots, y_n$ that is assumed to be mutually independent given θ is then

$$p(y_{1:n} \,|\, \theta) = \prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i} = \theta^s (1-\theta)^{n-s}, \tag{7}$$

where $s = \sum_{i=1}^{n} y_i$ is the number of successes. Note that s is the only feature of the observations that appears in the likelihood, and the inference problem could equally well be stated with s treated as the observation, in which case the likelihood distribution would be $s \,|\, \theta \sim \text{Binomial}(n, \theta)$.

Computations are facilitated by taking as prior a beta distribution $\text{Beta}(\alpha, \beta)$ whose parameters α, β are chosen so as to give an acceptable approximation of your prior state of knowledge. The beta distribution has pdf

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \quad (\theta \in [0,1])$$

where Γ is the gamma function[5] and the parameters α and β are positive. The uniform distribution corresponds to the case $\alpha = \beta = 1$. The distribution is named after the normalisation factor of its pdf, the beta function

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

The mean, variance and mode of the $\text{Beta}(\alpha, \beta)$ distribution are

$$\mathsf{E}(\theta) = \frac{\alpha}{\alpha+\beta}, \quad \mathsf{V}(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}, \quad \mathsf{mode}(\theta) = \frac{\alpha-1}{\alpha+\beta-2}.$$

The clever way to compute the mean is

$$\begin{aligned}
\mathsf{E}(\theta) &= \int_0^1 \theta \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \,\mathrm{d}\theta \\
&= \frac{\Gamma(\alpha+1)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\alpha+\beta+1)} \int_0^1 \frac{\Gamma(\alpha+\beta+1)}{\Gamma(\alpha+1)\Gamma(\beta)} \theta^{(\alpha+1)-1}(1-\theta)^{\beta-1} \,\mathrm{d}\theta \\
&= \frac{\alpha\Gamma(\alpha)}{\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha+\beta)}{(\alpha+\beta)\Gamma(\alpha+\beta)} \cdot 1 \\
&= \frac{\alpha}{\alpha+\beta}.
\end{aligned}$$

The integral in the second line is equal to 1 because it is the integral of the pdf of $\text{Beta}(\alpha+1, \beta)$. A similar trick can be used to compute the variance — details are left as an exercise.

In practical situations it may be easier for you to specify the prior's mean $\mathsf{E}(\theta) = m_0$ and variance $\mathsf{V}(\theta) = w_0$, then use the formulas

$$\alpha = \frac{m_0(m_0 - m_0^2 - w_0)}{w_0}, \quad \beta = \frac{m_0 - m_0^2 - w_0}{w_0(1-m_0)}$$

to define the beta distribution parameters.

With a likelihood $p(y_{1:n} \,|\, \theta) = \theta^s(1-\theta)^{n-s}$ and prior $p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$, the posterior is given by Bayes's formula as

$$p(\theta \,|\, y_{1:n}) \propto \theta^{\alpha+s-1}\theta^{\beta+n-s-1},$$

[5] $\Gamma(z) = \int_0^\infty t^{z-1}\mathrm{e}^{-t}\,\mathrm{d}t$ satisfies the recursion $\Gamma(z) = (z-1)\Gamma(z-1)$ with $\Gamma(1) = 1$, so $\Gamma(n) = (n-1)!$ for $n \in \{1, 2, \ldots\}$. Stirling's formula is $\Gamma(z) \approx \sqrt{2\pi}\mathrm{e}^z z^{z-\frac{1}{2}}$.

that is, $\theta \,|\, y_{1:n} \sim \text{Beta}(\alpha + s, \beta + n - s)$. Notice that the posterior belongs to the same family of distributions as the prior, that is, the family of beta distributions is conjugate to the likelihood (7). The α and β in the prior are updated to $\alpha + s$ and $\beta + n - s$, and the summary statistics are updated similarly, for example the posterior mean and the posterior mode (MAP estimate) are

$$\mathsf{E}(\theta \,|\, y_{1:n}) = \frac{\alpha + s}{\alpha + \beta + n}, \quad \text{mode}(\theta \,|\, y_{1:n}) = \frac{\alpha + s - 1}{\alpha + \beta + n - 2}.$$

As $n \to \infty$, the mean and mode both tend toward $\bar{y} = s/n$. Also, with large n one can use the normal approximation to define an approximate 95% credibility interval as

$$\mathsf{E}(\theta \,|\, y) \pm 1.96\sqrt{\mathsf{V}(\theta \,|\, y)}.$$

The predictive posterior distribution $\tilde{y} \,|\, y_{1:n}$ for a single observation has the pmf specified by

$$\mathsf{P}(\tilde{y} = 1 \,|\, y_{1:n}) = \int_0^1 \underbrace{\mathsf{P}(\tilde{y} = 1 \,|\, \theta)}_{=\theta} p(\theta | y_{1:n}) \,\mathrm{d}\theta = \mathsf{E}(\theta \,|\, y_{1:n}) = \frac{\alpha + s}{\alpha + \beta + n}.$$

As the number of observations grows, the predicted probability that a trial will be a "success" tends toward

$$\lim_{n \to \infty} \frac{\alpha + s}{\alpha + \beta + n} = \frac{s}{n} = \bar{y},$$

that is, the mean number of successes in the observation set, and the influence of the prior disappears as $n \to \infty$.

The predictive posterior distribution for the number of successes $\tilde{s}$ in m observations has the pmf

$$\begin{aligned} p(\tilde{s} \,|\, y_{1:n}) &= \int_0^1 p(\tilde{s} \,|\, \theta) p(\theta \,|\, y_{1:n}) \,\mathrm{d}\theta \\ &= \int_0^1 \binom{m}{\tilde{s}} \theta^{\tilde{s}} (1-\theta)^{m-\tilde{s}} \frac{\theta^{\alpha+s-1}(1-\theta)^{\beta+n-s-1}}{B(\alpha + s, \beta + n - s)} \,\mathrm{d}\theta \\ &= \binom{m}{\tilde{s}} \frac{B(\alpha + s + \tilde{s}, \beta + n + m - s - \tilde{s})}{B(\alpha + s, \beta + n - s)}. \end{aligned}$$

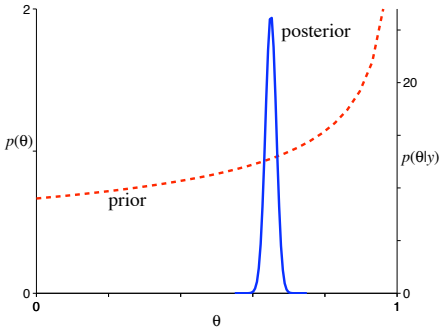
This pmf is called a beta-binomial distribution — the predictive posterior is *not* a binomial distribution! The predictive posterior mean is

$$\mathsf{E}(\tilde{s} \,|\, y_{1:n}) = \mathsf{E}(\tilde{y}_1 + \cdots + \tilde{y}_m \,|\, y_{1:n}) = \mathsf{E}(\tilde{y}_1 \,|\, y_{1:n}) + \cdots + \mathsf{E}(\tilde{y}_m \,|\, y_{1:n}) = m \cdot \frac{\alpha + s}{\alpha + \beta + n}.$$

Example: Opinion survey An opinion survey is conducted to determine the proportion θ of the population that is in favour of a certain policy. After some discussion with various experts, you determine that the prior belief has $\mathsf{E}(\theta) > 0.5$, but with a lot of uncertainty. The results of the survey are that, out of $n = 1000$ respondents, $s = 650$ were in favour of the policy. What do you conclude?

You decide on a prior distribution with $\mathsf{E}(\theta) = 0.6$ and $\mathsf{V}(\theta) = 0.3^2$, which corresponds to a beta distribution with $\alpha = 1$ and $\beta = \frac{2}{3}$, that is,

$$p(\theta) \propto (1-\theta)^{-1/3} \quad \theta \in [0, 1].$$



The survey results give you a posterior distribution with density $\theta \,|\, y_{1:n} \sim \text{Beta}(651, 350.667)$, whose mean and variance are

$$\mathsf{E}(\theta \,|\, y_{1:n}) = 0.6499, \quad \mathsf{V}(\theta \,|\, y_{1:n}) = 0.0151^2.$$

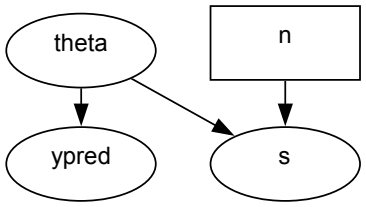
The 95% credibility interval found using the normal approximation is

$$0.6499 \pm 1.96 \cdot 0.0151 = (0.620, 0.679),$$

which agrees (to three decimals) with the 95% credibility interval computed using the inverse beta cdf. The probability that the 1001st respondent will be in favour is $\mathsf{P}(\tilde{y} = 1 \,|\, y_{1:n}) = \mathsf{E}(\theta \,|\, y_{1:n}) = 0.6499$.

In the following WinBUGS model, we base the inference on the number of successes observed, and use the likelihood $s \,|\, \theta \sim \text{Binomial}(n, \theta)$. The rectangle in the DAG denotes a constant.

```
model {
   s ~ dbin(theta,n)
   theta ~ dbeta(1,0.667)
   ypred ~ dbin(theta,1)
}
```

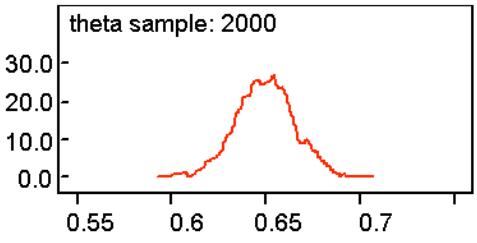


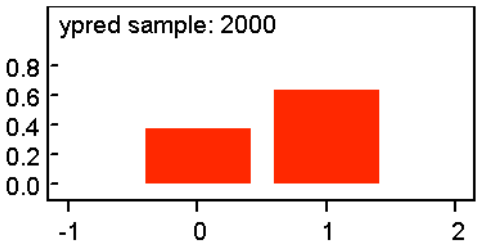
The data are entered as

```
list(s=650,n=1000)
```

The results after 2000 simulation steps are

node	mean	sd	2.5%	median	97.5%
theta	0.6497	0.01548	0.6185	0.6499	0.6802
ypred	0.6335				





5.3 Poisson model for count data

Let $\#I$ denote the number of occurrences of some phenomenon that are observed in an interval I (of time, usually). For example, $\#I$ could be the number of traffic accidents on a given stretch of highway, the number of particles emitted in the radioactive decay of an isotope sample, the number of outbreaks of a given disease in a given city... The number y of occurrences per unit time is often modelled as $y \,|\, \theta \sim \text{Poisson}(\theta)$, which has the pmf $\mathsf{P}(\#(t_0, t_0+1] = y \,|\, \theta) = \frac{(\theta)^y}{y!} \mathrm{e}^{-\theta}$ $(y \in \{0, 1, 2, \ldots\})$.

The Poisson model can be derived as follows. Assume that the events are relatively rare and occur at a constant rate θ, that is,

$$\mathsf{P}(\#(t, t+h] = 1 \,|\, \theta) = \theta h + o(h), \quad \mathsf{P}(\#(t, t+h] \geq 2 \,|\, \theta) = o(h),$$

where $o(h)$ means $\lim_{h\to 0}\frac{o(h)}{h} = 0$. Assume also that the numbers of occurences in distinct intervals are independent given θ. Letting $P_k(t) := \mathsf{P}(\#(0,t] = k \,|\, \theta)$, we have

$$\begin{aligned} P_0(t+h) &= \mathsf{P}(\#(0,t] = 0 \;\&\; \#(t,t+h] = 0 \,|\, \theta) \\ &= \mathsf{P}(\#(t,t+h] = 0 \,|\, \#(0,t] = 0, \theta)\mathsf{P}(\#(0,t] = 0 \,|\, \theta) \\ &= (1-\theta h + o(h))P_0(t). \end{aligned}$$

Letting $h \to 0$ gives the differential equation $P_0'(t) = -\theta P_0(t)$, which with the initial condition $P_0(0) = 1$ has the solution $P_0(t) = \mathrm{e}^{-\theta t}$. Similary, for $k > 0$ we have

$$\begin{aligned} P_k(t+h) &= \mathsf{P}(\#(0,t] = k \;\&\; \#(t,t+h] = 0 \,|\, \theta) \\ &\quad + \mathsf{P}(\#(0,t] = k-1 \;\&\; \#(t,t+h] = 1 \,|\, \theta) \\ &= (1-\theta h + o(h))P_k(t) + (\theta h + o(h))P_{k-1}(t), \end{aligned}$$

which in the limit $h \to 0$ gives the differential equations

$$P_k'(t) = -\theta P_k(t) + \theta P_{k-1}(t).$$

Solving these with the initial conditions $P_k(0) = 0$ $(k > 0)$ gives

$$P_k(t) = \frac{(\theta t)^k}{k!}\mathrm{e}^{-\theta t}, \quad (k \in \{0,1,2,\ldots\}),$$

which for $t = 1$ is the Poisson pmf.

Thus, a Poisson-distributed random variable $y \,|\, \theta \sim \text{Poisson}(\theta)$ has the pmf

$$\mathsf{P}(y = k \,|\, \theta) = \frac{\theta^k}{k!}\mathrm{e}^{-\theta}, \quad (k \in \{0,1,2,\ldots\})$$

and the summary statistics

$$\mathsf{E}(y \,|\, \theta) = \theta, \quad \mathsf{V}(y \,|\, \theta) = \theta.$$

The likelihood pmf of a sequence $y_1, \ldots, y_n$ of Poisson-distributed counts on unit-length intervals, assumed to be mutually independent conditional on θ, is

$$p(y_{1:n} \,|\, \theta) = \prod_{i=1}^{n} \frac{(\theta)^{y_i}}{y_i!}\mathrm{e}^{-\theta} \propto \theta^s \mathrm{e}^{-n\theta}$$

where $s = \sum_{i=1}^n y_i$.

The conjugate prior for the Poisson distribution is the Gamma(α, β) distribution, which has the pdf

$$p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}\mathrm{e}^{-\beta\theta} \quad (\theta > 0).$$

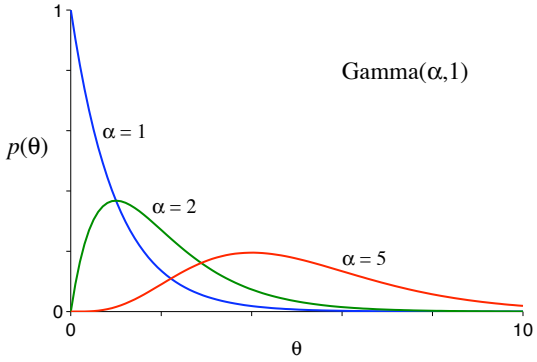
The distribution gets its name from the normalisation factor of its pdf. The mean, variance and mode of $\theta \sim \text{Gamma}(\alpha, \beta)$ are

$$\mathsf{E}(\theta) = \frac{\alpha}{\beta}, \quad \mathsf{V}(\theta) = \frac{\alpha}{\beta^2}, \quad \mathsf{mode}(\theta) = \frac{\alpha - 1}{\beta}.$$

The formula for the mean can be derived as follows:

$$\begin{aligned} \mathsf{E}(\theta) &= \int_0^\infty \theta \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}\mathrm{e}^{-\beta\theta}\,\mathrm{d}\theta \\ &= \frac{\Gamma(\alpha+1)}{\beta\Gamma(\alpha)}\int_0^\infty \frac{\beta^{\alpha+1}}{\Gamma(\alpha+1)}\theta^{(\alpha+1)-1}\mathrm{e}^{-\beta\theta}\,\mathrm{d}\theta \\ &= \frac{\alpha\Gamma(\alpha)}{\beta\Gamma(\alpha)} \cdot 1 = \frac{\alpha}{\beta}. \end{aligned}$$

The parameter $\beta > 0$ is a scaling factor (note that some tables and software use $1/\beta$ instead of β to specify the gamma distribution); the parameter $\alpha > 0$ determines the shape:



With the likelihood pdf $p(y_{1:n}\,|\,\theta) \propto \theta^s \mathrm{e}^{-n\theta}$ and the prior pdf $p(\theta) \propto \theta^{\alpha-1}\mathrm{e}^{-\beta\theta}$, Bayes's formula gives the posterior pdf

$$p(\theta\,|\,y_{1:n}) \propto \theta^{\alpha+s-1}\mathrm{e}^{-(\beta+n)\theta},$$

that is, $\theta\,|\,y_{1:n} \sim \mathrm{Gamma}(\alpha+s, \beta+n)$. The α and β parameters in the prior's gamma distribution are thus updated to $\alpha+s$ and $\beta+n$ in the posterior's gamma distribution. The summary statistics are updated similarly, in particular the posterior mean and posterior mode (MAP estimate) are

$$\mathsf{E}(\theta\,|\,y_{1:n}) = \frac{\alpha+s}{\beta+n}, \quad \mathsf{mode}(\theta\,|\,y_{1:n}) = \frac{\alpha+s-1}{\beta+n}.$$

As $n \to \infty$, both the posterior mean and posterior mode tend to $\bar{y} = s/n$.

The prior predictive distribution (marginal distribution of data) has the pmf

$$\begin{aligned}
\mathsf{P}(\tilde{y}=k) &= \int_0^\infty \mathsf{P}(\tilde{y}=k\,|\,\theta)p(\theta)\,\mathrm{d}\theta = \int_0^\infty \frac{\theta^k}{k!}\mathrm{e}^{-\theta}\frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}\mathrm{e}^{-\beta\theta} \\
&= \frac{\Gamma(\alpha+k)\beta^\alpha}{k!(\beta+1)^{\alpha+k}\Gamma(\alpha)} \underbrace{\int_0^\infty \frac{(\beta+1)^{\alpha+k}}{\Gamma(\alpha+k)}\theta^{\alpha+k-1}\mathrm{e}^{-(\beta+1)\theta}\,\mathrm{d}\theta}_{=1} \\
&= \frac{(\alpha+k-1)(\alpha+k-2)\cdots\alpha\Gamma(\alpha)}{\Gamma(\alpha)k!}\left(\frac{\beta}{\beta+1}\right)^\alpha\left(\frac{1}{\beta+1}\right)^k \\
&= \binom{\alpha+k-1}{\alpha-1}\left(\frac{\beta}{\beta+1}\right)^\alpha\left(\frac{1}{\beta+1}\right)^k.
\end{aligned}$$

This is the pmf of the *negative binomial* distribution. The summary statistics of $\tilde{y} \sim \mathrm{NegBin}(\alpha,\beta)$ are

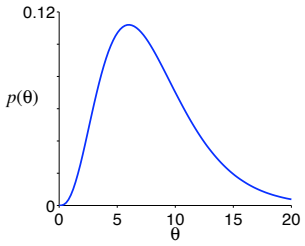
$$\mathsf{E}(\tilde{y}) = \frac{\alpha}{\beta}, \quad \mathsf{V}(\tilde{y}) = \frac{\alpha}{\beta^2} + \frac{\alpha}{\beta}.$$

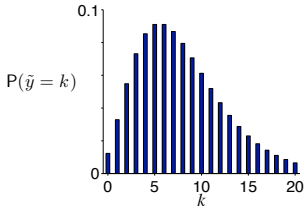
The negative binomial distribution also happens to model the number of Bernoulli failures occurring before the αth success when the probability of success is $p = \frac{\beta}{\beta+1}$. For this reason, many software packages (including Matlab, R and WinBUGS) use p instead of β as the second parameter to specify the negative binomial distribution.

The posterior predictive distribution can be derived similarly as the prior predictive, and is

$$\tilde{y}\,|\,y_{1:n} \sim \mathrm{NegBin}(\alpha+s, \beta+n).$$

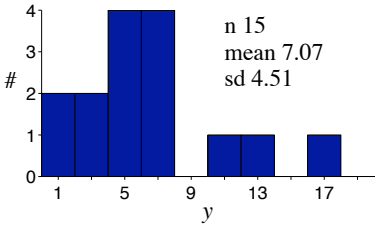
Example: moose counts A region is divided into equal-area (100 km^2) squares and the number of moose spotted in each square is recorded. The prior distribution is $\theta \sim \text{Gamma}(4, 0.5)$, which corresponds to the prior predictive pmf $\tilde{y} \sim \text{NegBin}(4, 0.5)$.





On a certain day the following moose counts are collected from an aerial survey of 15 squares:

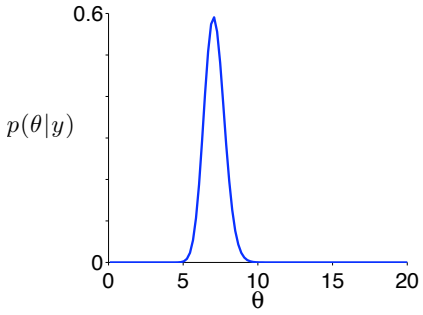
5	7	7	12	2
14	7	8	5	6
18	6	4	1	4

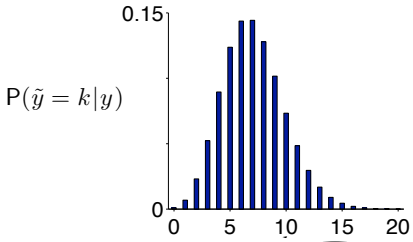


The posterior distribution for the rate (i.e. number of moose per 100 km^2) is $\theta \,|\, y_{1:15} \sim \text{Gamma}(110, 15.5)$, for which

$$\mathsf{E}(\theta \,|\, y_{1:15}) = 7.0968, \quad \mathsf{V}(\theta \,|\, y_{1:15}) = 0.6767^2, \quad \mathsf{mode}(\theta \,|\, y_{1:15}) = 7.0323$$

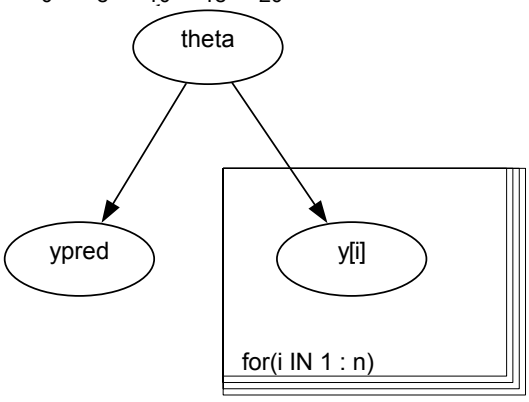
and the 95% credibility interval is $[5.83, 8.48]$. (The normal approximation gives the interval $[5.77, 8.42]$.) The predictive posterior distribution is $\tilde{y} \,|\, y_{1:n} \sim \text{NegBin}(110, 15.5)$.





A WinBUGS model for this problem is

```
model {
  for (i in 1:n) { y[i]~ dpois(theta) }
  theta ~ dgamma(4,0.5)
  ypred ~ dpois(theta)
}
```

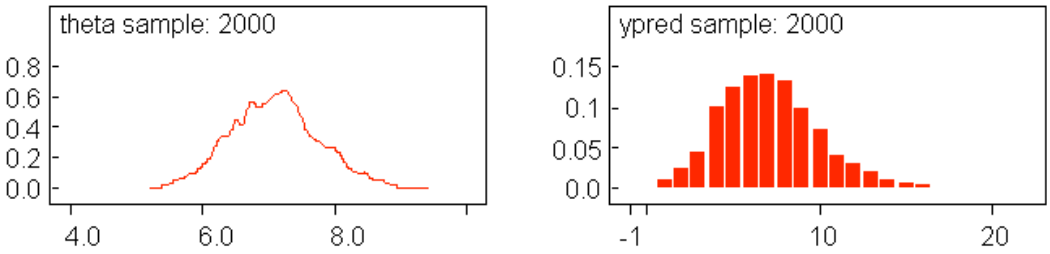


The data are entered as

```
list( y=c(5,7,7,12,2,14,7,8,5,6,18,6,4,1,4), n=15)
```

The results are

node	mean	sd	2.5%	median	97.5%
theta	7.107	0.6608	5.85	7.101	8.482
ypred	7.098	2.838	2.0	7.0	13.0



■

A more general Poisson model can be used for counts of occurrences in intervals of different sizes. The model is

$$y_i \,|\, \theta \sim \text{Poisson}(\theta t_i)$$

where the t_i are known positive values, sometimes called *exposures*. Assuming as usual that the counts are mutually independent given θ, the likelihood is

$$p(y_{1:n} \,|\, \theta) \propto \theta^s \mathrm{e}^{-\theta T}$$

where $s = \sum_{i=1}^n y_i$ and $T = \sum_{i=1}^n t_i$. With the conjugate prior $\theta \sim \text{Gamma}(\alpha, \beta)$, the posterior is

$$\theta \,|\, y_{1:n} \sim \text{Gamma}(\alpha + s, \beta + T),$$

with

$$\mathsf{E}(\theta \,|\, y_{1:n}) = \frac{\alpha + s}{\beta + T}, \quad \mathsf{mode}(\theta \,|\, y_{1:n}) = \frac{\alpha + s - 1}{\beta + T}.$$

As $n \to \infty$, both the posterior mean and posterior mode tend towards s/T.

5.4 Exponential model for lifetime data

Consider a non-negative random variable y used to model intervals such as the time-to-failure of machine components or a patient's survival time. In such applications, it is typical to specify the probability distribution using a hazard function, from which the cdf and pdf can be deduced (and vice versa).

The hazard function is defined by

$$h(t)\,\mathrm{d}t = \mathsf{P}(\underbrace{t < y \le t + \mathrm{d}t}_{\text{fail in } (t, t+\mathrm{d}t]} \,|\, \underbrace{t < y}_{\text{OK at } t}) = \frac{\mathsf{P}(t < y \le t + \mathrm{d}t)}{\mathsf{P}(t < y)} = \frac{p(t)\,\mathrm{d}t}{S(t)},$$

where p is the pdf of y and $S(t) := \mathsf{P}(t < y)$ is called the *reliability function*. Now, because $p(t) = -S'(t)$, we have the differential equation $h(t) = -\frac{S'(t)}{S(t)}$ with initial condition $S(0) = 1$, which can be solved to give

$$S(t) = \mathrm{e}^{-\int_0^t h(\tau)\,\mathrm{d}\tau}.$$

In particular, for constant hazard $h(t) = \theta$ the reliability is $S(t) = \mathrm{e}^{-\theta t}$ and the density is the *exponential distribution* pdf

$$p(t) = \theta \mathrm{e}^{-\theta t}.$$

Suppose a component has worked without failure for s time units. Then according to the constant-hazard model, the probability that it will survive at least t time units more is

$$\mathsf{P}(y > s + t \,|\, y > s) \quad = \quad \frac{\mathsf{P}(y > s \;\&\; y > s + t)}{\mathsf{P}(y > s)} = \frac{\mathsf{P}(y > s + t)}{\mathsf{P}(y > s)} = \frac{\mathrm{e}^{-\theta(t+s)}}{\mathrm{e}^{-\theta s}} = \mathrm{e}^{-\theta t},$$

which is the same probability as for a new component! This is the "lack-of-memory" or "no-aging" property of the constant-hazard model.

For an exponentially distributed random variable $y \,|\, \theta \sim \text{Exp}(\theta)$ the mean and variance are

$$\mathsf{E}(y \,|\, \theta) = \frac{1}{\theta}, \quad \mathsf{V}(y \,|\, \theta) = \frac{1}{\theta^2}.$$

The exponential distribution also models the durations (waiting times) between consecutive Poisson-distributed occurrences.

For exponentially-distributed samples $y_i \,|\, \theta \sim \text{Exp}(\theta)$ that are mutually independent given θ, the likelihood is

$$p(y_{1:n} \,|\, \theta) = \prod_{i=1}^{n} \theta \mathrm{e}^{-\theta y_i} = \theta^n \mathrm{e}^{-\theta s}$$

where $s = \sum_{i=1}^{n} y_i$. Using the conjugate prior $\theta \sim \text{Gamma}(\alpha, \beta)$, the posterior pdf is

$$p(\theta \,|\, y_{1:n}) \propto \theta^{\alpha-1} \mathrm{e}^{-\beta\theta} \cdot \theta^n \mathrm{e}^{-\theta s} = \theta^{\alpha+n-1} \mathrm{e}^{-(\beta+s)\theta},$$

that is, $\theta \,|\, y_{1:n} \sim \text{Gamma}(\alpha+n, \beta+s)$, for which

$$\mathsf{E}(\theta \,|\, y_{1:n}) = \frac{\alpha+n}{\beta+s}, \quad \mathsf{mode}(\theta \,|\, y_{1:n}) = \frac{\alpha+n-1}{\beta+s}, \quad \mathsf{V}(\theta \,|\, y_{1:n}) = \frac{\alpha+n}{(\beta+s)^2}.$$

It often happens that lifetime or survival studies are ended before all the subjects have failed or died. Then, in addition to k observations $y_1, \ldots, y_k \in [0, L]$, we have $n-k$ samples whose lifetimes are known to be $y_j > L$, but are otherwise unknown. This is called a *censored* data set. The censored observations can be modelled as Bernoulli trials with "success" ($z_j = 1$), corresponding to $y_j > L$, having the probability

$$\mathsf{P}(y_j > L \,|\, \theta) = \mathrm{e}^{-\theta L}.$$

The likelihood of the censored data is

$$p(y_{1:k}, z_{1:n-k} \,|\, \theta) = \prod_{i=1}^{k} \theta \mathrm{e}^{-\theta y_i} \cdot \prod_{j=1}^{n-k} \mathrm{e}^{-\theta L} = \theta^k \mathrm{e}^{-\theta(s_k + (n-k)L)},$$

where $s_k = \sum_{i=1}^{k} y_i$. With the conjugate prior $\theta \sim \text{Gamma}(\alpha, \beta)$, the posterior pdf is

$$p(\theta \,|\, y_{1:k}, z_{1:n-k}) \propto \theta^{\alpha-1} \mathrm{e}^{-\beta\theta} \cdot \theta^k \mathrm{e}^{-\theta(s_k+(n-k)L)} = \theta^{\alpha+k-1} \mathrm{e}^{-(\beta+s_k+(n-k)L)\theta},$$

that is, $\theta \,|\, y_{1:k}, z_{1:n-k} \sim \text{Gamma}(\alpha+k, \beta+s_k+(n-k)L)$.

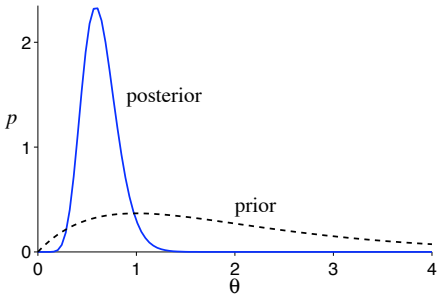
Example: Censored lifetime data In a two-year survival study of 15 cancer patients, the lifetimes (in years) are

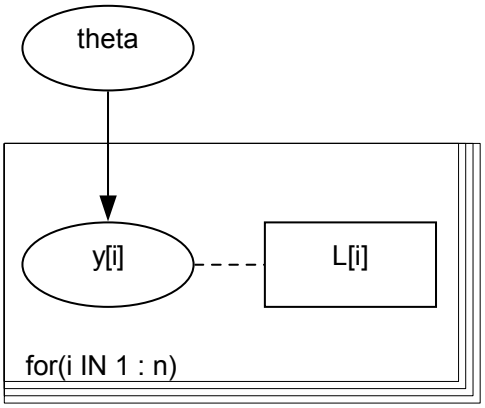
```
1.54,0.70,1.23,0.82,0.99,1.33,0.38,0.99,1.97,1.10,0.40
```

and 4 patients are still alive at the end of the study. Assuming mutually independent $y_i \,|\, \theta \sim \text{Exp}(\theta)$ conditional on θ, and choosing the prior $\theta \sim \text{Gamma}(2, 1)$, we obtain the posterior

$$\theta \,|\, y_{1:11}, z_{1:4} \sim \text{Gamma}(2+11, 1+11.45+4 \cdot 2) = \text{Gamma}(13, 20.45)$$

which has mean 0.636, variance $(0.176)^2$, and 95% credibility interval $(0.338, 1.025)$. The normal approximation has 95% credibility interval $(0.290, 0.981)$.





A WinBUGS model for this problem is

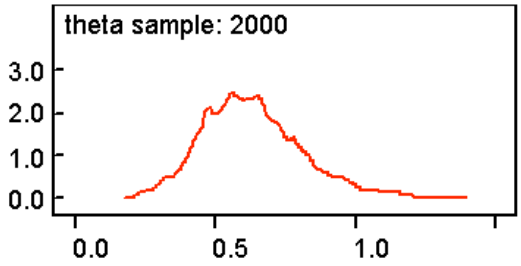
```
model {
  theta ~ dgamma(2,1)
  for (i in 1:n) {y[i] ~ dexp(theta)I(L[i],)}
}
```

Censoring is represented by appending the `I(lower,upper)` modifier to the distribution specification. The data is entered as

```
list(y=c(1.54,0.70,1.23,0.82,0.99,1.33,0.38,0.99,1.97,1.10,0.40,
   NA,NA,NA,NA), n=15, L=c(0,0,0,0,0,0,0,0,0,0,0,2,2,2,2))
```

where the censored observations are represented by `NA`. The results after 2000 simulation steps are

node	mean	sd	2.5%	median	97.5%
`theta`	0.6361	0.1789	0.3343	0.6225	1.045



5.5 Estimating the variance of a normal model

Suppose we have real-valued observations $y_1, \ldots, y_n$ that are mutually independent given ϕ and identically normally distributed with known mean m and unknown variance ϕ. The likelihood is then

$$p(y_{1:n} \mid \phi) = \left(\frac{1}{2\pi\phi}\right)^{n/2} \mathrm{e}^{-\frac{n}{2\phi}s_0^2}$$

where $s_0^2 := \frac{1}{n}\sum_{i=1}^{n}(y_i - m)^2$.

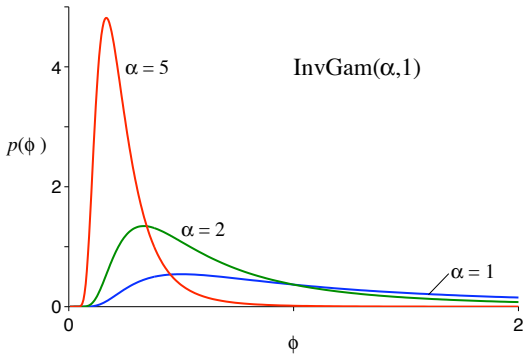
The conjugate prior for this likelihood is the inverse gamma distribution $\mathrm{InvGam}(\alpha, \beta)$, which has the pdf

$$p(\phi) = \frac{\beta^\alpha}{\Gamma(\alpha)}\phi^{-(\alpha+1)}\mathrm{e}^{-\beta/\phi} \quad (\phi > 0).$$

The name of this distribution comes from the fact that if $\tau \sim \mathrm{Gamma}(\alpha, \beta)$ and $\phi = \frac{1}{\tau}$ then $\phi \sim \mathrm{InvGam}(\alpha, \beta)$, as the reader can (should!) verify. The mean, variance and mode of $\phi \sim \mathrm{InvGam}(\alpha, \beta)$ are

$$\mathsf{E}(\phi) = \frac{\beta}{\alpha - 1} \quad (\alpha > 1); \quad \mathsf{V}(\phi) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)} \quad (\alpha > 2); \quad \mathsf{mode}(\phi) = \frac{\beta}{\alpha+1}.$$

The inverse gamma distribution parameter $\beta > 0$ is a scaling factor; the parameter $\alpha > 0$ determines the shape:



With the likelihood pdf $p(y_{1:n} \,|\, \phi) \propto \phi^{-n/2} \mathrm{e}^{-\frac{n}{2\phi} s_0^2}$ and the prior pdf $p(\phi) \propto \phi^{-(\alpha+1)} \mathrm{e}^{-\beta/\phi}$, Bayes's formula gives the posterior pdf

$$p(\phi \,|\, y_{1:n}) \propto \phi^{-(\alpha+\frac{n}{2}+1)} \mathrm{e}^{-(\beta+\frac{n}{2}s_0^2)/\phi},$$

that is, $\phi \,|\, y_{1:n} \sim \mathrm{InvGam}(\alpha + \frac{n}{2}, \beta + \frac{n}{2}s_0^2)$. The α and β parameters in the prior's inverse gamma distribution are thus updated to $\alpha + \frac{n}{2}$ and $\beta + \frac{n}{2}s_0^2$ in the posterior's inverse gamma distribution. The summary statistics are updated similarly, in particular the posterior mean and posterior mode (MAP estimate) are

$$\mathsf{E}(\phi \,|\, y_{1:n}) = \frac{\beta + \frac{n}{2}s_0^2}{\alpha + \frac{n}{2} - 1}, \quad \mathsf{mode}(\phi \,|\, y_{1:n}) = \frac{\beta + \frac{n}{2}s_0^2}{\alpha + \frac{n}{2} + 1}.$$

As $n \to \infty$, both the posterior mean and posterior mode tend to $s_0^2 = \frac{1}{n}\sum_{i=1}^n (y_i - m)^2$.

An alternative derivation of these results can be obtained by considering the unknown parameter to be the precision $\tau = \frac{1}{v}$. Then the likelihood has pdf $p(y_{1:n} \,|\, \tau) \propto \tau^{n/2} \mathrm{e}^{-\frac{n}{2}\tau s_0^2}$, the conjugate prior is $\tau \sim \mathrm{Gamma}(\alpha, \beta)$ with pdf $p(\tau) \propto \tau^{\alpha-1} \mathrm{e}^{-\beta\tau}$, and Bayes's formula gives the posterior pdf

$$p(\tau \,|\, y_{1:n}) \propto \tau^{\alpha-1+\frac{n}{2}} \mathrm{e}^{-(\beta+\frac{n}{2}s_0^2)\tau},$$

that is, $\tau \,|\, y_{1:n} \sim \mathrm{Gamma}(\alpha + \frac{n}{2}, \beta + \frac{n}{2}s_0^2)$.

6 Jeffreys's prior

In their solution for the problem of estimating the "probability of success" θ from a set of observations of Bernoulli trials, Bayes and Laplace used a uniform distribution as the prior for θ, that is, $\theta \sim \mathrm{Beta}(1,1)$. Intuitively, the uniform distribution appears to be a natural choice to model a state of complete *ignorance*: you have no prior preference for any value of θ because all values are equally likely.

There is, however, a flaw in this way of thinking. If you are completely ignorant about θ, then clearly you are completely ignorant about any function of θ. But if θ has a uniform density, the density for $\psi = h(\theta)$ is not generally uniform! For example, if $\psi = \theta^2$ and $\theta \propto 1$, then $p(\psi) \propto \psi^{-\frac{1}{2}}$. Using the uniform density as a general 'ignorance prior' is therefore not consistent with the change-of-variables rule for parameter transformations.

Jeffreys proposed the following general rule for selecting a prior pdf to represent ignorance:

$$p(\theta) \propto \sqrt{J(\theta)}, \quad \text{where } J(\theta) = -\mathsf{E}\left(\frac{\partial^2 \log p(y \,|\, \theta)}{\partial \theta^2} \,|\, \theta \right)$$

is the *Fisher information*. This rule is invariant to smooth one-to-one transformations $\psi = h(\theta)$, in the sense that $p(\psi) = \sqrt{J(\psi)}$ and $p(\theta) = \sqrt{J(\theta)}$ are consistent with the change-of-variables rule $p(\psi) = p(\theta)/h'(\theta)$, because

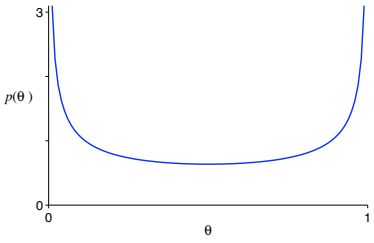
$$\begin{aligned} J(\psi) &= -\mathsf{E}\left(\frac{\partial^2 \log p(y \,|\, \psi)}{\partial \psi^2} \,|\, \psi\right) \\ &= -\mathsf{E}\left(\frac{\partial^2 \log p(y \,|\, \theta)}{\partial \theta^2} \left(\frac{1}{h'}\right)^2 \,|\, \theta\right) \\ &= J(\theta)\left(\frac{1}{h'}\right)^2. \end{aligned}$$

In the case of the binomial data model we have $p(y_i \,|\, \theta) = \theta^{y_i}(1-\theta)^{1-y_i}$ and $\mathsf{E}(y_i \,|\, \theta) = \theta$. Then

$$\begin{aligned} \log p(y_i \,|\, \theta) &= y_i \log\theta + (1-y_i)\log(1-\theta), \\ \frac{\partial \log p(y_i \,|\, \theta)}{\partial \theta} &= \frac{y_i}{\theta} - \frac{1-y_i}{1-\theta}, \\ \frac{\partial^2 \log p(y_i \,|\, \theta)}{\partial \theta^2} &= -\frac{y_i}{\theta^2} - \frac{1-y_i}{(1-\theta)^2}, \\ -\mathsf{E}\left(\frac{\partial^2 \log p(y_i \,|\, \theta)}{\partial \theta^2} \,|\, \theta\right) &= \frac{\theta}{\theta^2} + \frac{1-\theta}{(1-\theta)^2} = \frac{1}{\theta(1-\theta)}. \end{aligned}$$

The Jeffreys prior pdf for the binomial model is therefore

$$p(\theta) \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}},$$



that is, $\theta \sim \text{Beta}(\frac{1}{2}, \frac{1}{2})$. The density is higher at the ends of the interval $[0,1]$ than in the middle, which may come as a surprise. The corresponding posterior (assuming observations $y_1, \ldots, y_n$ to be conditionally independent given θ) is $\text{Beta}(\frac{1}{2} + \sum y_i, \frac{1}{2} + n - \sum y_i)$.

Jeffreys priors for commonly encountered data models are:

$y_i \,\|\, \theta \sim$	$p(\theta) \propto$	$\theta \,\|\, y_{1:n}$
$\text{Normal}(\theta, v)$	1	$\text{Normal}(\frac{1}{n}\sum y_i, \frac{v}{n})$
$\text{Binomial}(1, \theta)$	$\theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}$	$\text{Beta}(\frac{1}{2} + \sum y_i, \frac{1}{2} + n - \sum y_i)$
$\text{Poisson}(\theta)$	$\theta^{-\frac{1}{2}}$	$\text{Gamma}(\frac{1}{2} + \sum y_i, n)$
$\text{Exp}(\theta)$	θ^{-1}	$\text{Gamma}(n, \sum y_i)$
$\text{Normal}(m, \theta)$	θ^{-1}	$\text{InvGam}(\frac{n}{2}, \frac{1}{2}\sum (y_i - m)^2)$

Although it has some attractive properties, there are several reasons why Jeffreys's rule should not be your default way to choose a prior:

- It is not defined for all data models (the Fisher information expectation integral may diverge), and often produces improper priors.
- It is difficult to apply in problems with several parameters.
- It is based on the likelihood, rather than being based entirely on prior information. Thus it is not a Bayesian procedure and it leads to a logical inconsistency called "violation of the likelihood principle"; this is explained in the next chapter.

7 Some General Principles

7.1 Ancillarity and Sufficiency

In Frequentist statistics the theory of ancillararity and sufficiency is a tricky but essential element of inference. In Bayesian inference, this theory is not needed to do inference, so this section can be skipped without loss of continuity with the rest of the text.

Any data y that is independent of the parameter θ is said to be *ancillary*. For ancillary data we have $p(\theta\,|\,y) = p(\theta)$, that is, the posterior probability distribution (state of knowledge) is the same as the prior probability distribution (state of knowledge): ancillary data tells us nothing (directly) about θ.

Ancillarity might give information indirectly, however. For example, if $y = (y_1, y_2)$, then

$$p(\theta\,|\,y) = p(\theta\,|\,y_1, y_2) \propto p(\theta\,|\,y_1)p(y_2\,|\,y_1, \theta),$$

If y_1 is ancillary, then this reduces to

$$p(\theta\,|\,y) = p(\theta)p(y_2\,|\,y_1, \theta).$$

Thus, if y_1 affects the distribution of y_2 given θ, then in this indirect way it can have an influence on the posterior distribution.

Example: Urn and Die An urn contains an unknown number θ of balls, all of them red. A die is tossed with the result $y_1 \in \{1, 2, \ldots, 6\}$, and y_1 black balls are added to the urn. Obviously, y_1 is ancillary: it does not change our degree of belief (whatever it may be) about θ (the number of red balls).

Next, the urn is shaken, and a ball is drawn: we set $y_2 = 1$ if the drawn ball is red, $y_2 = 0$ otherwise. Then we have

$$\mathsf{P}(y_2 = 1\,|\,y_1, \theta) = \frac{\theta}{\theta + y_1}.$$

Inference about θ thus uses the observation y_1, but the mechanism that produced y_1 can be ignored: it doesn't matter, for example, whether the die is loaded. ■

In our study of various basic single-parameter models in chapter 5, we found that, in many cases, the only feature of the observations $y_1, \ldots, y_n$ that was needed to define the likelihood was the sum $s = \sum_i y_i$. Such a function of the data is called a *sufficient statistic*. Let's explain this concept more precisely.

In general, for data partitioned as $y = (x, z)$, x is said to be sufficient if any of the following conditions holds:

(a) $p(\theta\,|\,y)$ does not depend on z;

(b) $p(z\,|\,x, \theta) = p(z\,|\,x)$, that is, z is ancillary given x;

(c) $p(y\,|\,\theta) = q_1(x, \theta)q_2(x, z)$.

The conditions are equivalent: condition (a) implies $p(\theta\,|\,x, z) = p(\theta\,|\,x)$, which means that z and θ are independent given x, which implies (b); condition (b) implies

$$p(y\,|\,\theta) = p(x, z\,|\,\theta) = p(x\,|\,\theta)p(z\,|\,x, \theta) = \underbrace{p(x\,|\,\theta)}_{q_1(x,\theta)}\underbrace{p(z\,|\,x)}_{q_2(x,z)}$$

which is condition (c); and condition (c) implies $p(\theta \,|\, y) \propto q_1(x,\theta)p(\theta)$, so (c) implies (a). If x is sufficient then so is $h(x)$ for any one-to-one function h.

For example, for $y_i \sim \text{Normal}(\theta, v)$ with $y_1, \ldots, y_n$ independent given θ, the likelihood is

$$p(y_{1:n} \,|\, \theta) = \left(\frac{1}{2\pi v}\right)^{n/2} \mathrm{e}^{-\frac{1}{2v}\sum_{i=1}^n (y_i-\theta)^2} = \underbrace{\left(\frac{1}{2\pi v}\right)^{n/2} \mathrm{e}^{-\frac{n}{2v}(\theta-\bar{y})^2}}_{q_1(\theta,\bar{y})} \underbrace{\mathrm{e}^{-\frac{1}{2v}\sum_{i=1}^n (y_i-\bar{y})^2}}_{q_2(\bar{y},y_{1:n})}$$

and so $\bar{y}$ is sufficient, and thus so is $s = n\bar{y}$. The observations y_i are ancillary given the sample mean or given the sum.

Another example is $y_i \sim \text{Binomial}(1, \theta)$ with $y_1, \ldots, y_n$ independent given θ; the likelihood is

$$p(y_{1:n} \,|\, \theta) = \prod_{i=1}^n \theta^{y_i}(1-\theta)^{1-y_i} = \theta^s(1-\theta)^{n-s}$$

and so s is sufficient.

7.2 Likelihood Principle and Stopping Rules

The *likelihood principle* is the precept that all information *from observed data* y that is relevant to inferences about θ is found in the likelihood $p(y \,|\, \theta)$ (up to a multiplicative factor that does not depend on θ). This fundamental rule of logical consistency is satisfied automatically by Bayesian inference, because proportional likelihoods with the same priors lead to the same posteriors.

Suppose we conduct an experiment (experiment A) in which we count the number s of successes in a fixed predetermined number n of Bernoulli trials, with the observations assumed mutually independent given the probability of success θ. Then, as we saw in section 5.2, the likelihood pdf is

$$p(s \,|\, \theta, n) = \binom{n}{s} \theta^s(1-\theta)^{n-s}, \tag{8}$$

that is, $s \,|\, \theta, n \sim \text{Binomial}(n, \theta)$.

Now consider experiment B, which consists of counting the number of trials n needed to get a fixed predetermined number s of successes. There are $\binom{n-1}{s-1}$ possible sequences having s successes and $n-s$ failures, because the last observation must be a success. Thus, the likelihood for the number of trials is

$$p(n \,|\, \theta, s) = \binom{n-1}{s-1} \theta^s(1-\theta)^{n-s}, \tag{9}$$

that is, $(n-s) \,|\, \theta, s \sim \text{NegBin}(s, \frac{\theta}{1-\theta})$.

If the number of successes and failures is the same in both experiments, the likelihood functions (8) and (9) are proportional — their ratio is n/s. Thus, by the likelihood principle, the influence *of the data* on inference about θ should be the same for both experiments.

In Bayesian inference, when we learn the results of experiment A, our state of knowledge about θ gets updated to the posterior distribution with pdf

$$p(\theta \,|\, s, n) \propto p(\theta)\theta^s(1-\theta)^{n-s}. \tag{10}$$

When we learn the results of experiment B, if our prior state of knowledge about θ is the same then the posterior is also (10). The Bayesian inferences about θ from the two experiments are the same, in agreement with the Likelihood Principle.[6]

A useful consequence of the likelihood principle is the *stopping rule principle*, which states that the information gained from a sequential experiment does not depend on the rule adopted to terminate the data collection, provided only that this stopping rule does not depend on θ (*noninformative stopping*). For example, suppose you plan out to carry out a sequence of N Bernoulli trials, but the experiment is interrupted by some unrelated unexpected event — say the client calls and demands an interim report — leaving you with $n < N$ observations. As a Bayesian data analyst, you can proceed to analyse the results using the likelihood (8), the same as if it had been your intention all along to collect n observations.

As another example, consider experiment C, which consists of collecting independent Bernoulli samples, stopping as soon as the number of successes equals the number of failures, or as soon as N samples have been collected, whichever comes first. N is fixed to be very large, say $N = 10^6$, but finite, to ensure that the experiment can't run forever. Given the number of trials n and number of successes s from this experiment, you can proceed using the likelihood

$$p(s,n \,|\, \theta,N) = \binom{n}{s} \theta^s (1-\theta)^{n-s},$$

and obtain exactly the same inference results about θ as if the data had come from experiment A or B.[7]

As mentioned earlier, Jeffreys's rule also conflicts with the likelihood principle. To illustrate, consider experiment B. From (9) we have

$$\begin{aligned}
\log p(n \,|\, \theta) &= s \log \theta + (n-s) \log(1-\theta) + \log \binom{n-1}{s-1}, \\
\frac{\partial \log p(n \,|\, \theta)}{\partial \theta} &= \frac{s}{\theta} - \frac{n-s}{1-\theta}, \\
\frac{\partial^2 \log p(n \,|\, \theta)}{\partial \theta^2} &= -\frac{s}{\theta^2} - \frac{n-s}{(1-\theta)^2}, \\
-\mathsf{E}\left(\frac{\partial^2 \log p(x \,|\, \theta)}{\partial \theta^2} \,|\, \theta \right) &= \frac{s}{\theta^2(1-\theta)} \quad [\text{using } \mathsf{E}(n-s \,|\, \theta, s) = \frac{s(1-\theta)}{\theta}],
\end{aligned}$$

and so the Jeffreys prior is $p(\theta) \propto \theta^{-1}(1-\theta)^{-\frac{1}{2}}$, that is, $\theta \sim \text{Beta}(0, \frac{1}{2})$, an improper distribution. The Jeffreys prior for experiment A (calculated in section 6) is $\theta \sim \text{Beta}(\frac{1}{2}, \frac{1}{2})$. Because the two Jeffreys priors differ and the likelihoods are proportional when both experiments produce the same number of successes and failures, the posterior distributions will differ.

8 Hypothesis Testing

As pointed out in section 4.1, Bayesian hypothesis testing is straightforward. For a *hypothesis* of the form $H_i : \theta \in \Theta_i$, where Θ_i is a subset of the parameter space Θ, we can

[6] The Frequentist hypothesis tests for the experiments A and B are different and generally give different inferences — a violation of the likelihood principle.

[7] Frequentist statistics does not work this way.

compute the prior probability

$$\pi_i = \mathsf{P}(H_i) = \mathsf{P}(\theta \in \Theta_i)$$

and the posterior probability

$$p_i = \mathsf{P}(H_i \,|\, y) = \mathsf{P}(\theta \in \Theta_i \,|\, y).$$

Often, there are only two hypotheses, the "null hypothesis" H_0 and its logical negation H_1 : $\theta \notin \Theta_0$, called the "alternative hypothesis". The hypothesis with the highest probability can be chosen as the "best" hypothesis; a more sophisticated choice can be made using Decision Theory (to be discussed in section 14).

Two hypotheses can be compared using odds. The posterior odds in favour of H_0 against H_1 given data y are given by the ratio

$$\frac{p_0}{p_1} = \frac{\mathsf{P}(H_0 \,|\, y)}{\mathsf{P}(H_1 \,|\, y)} = \underbrace{\frac{\mathsf{P}(H_0)}{\mathsf{P}(H_1)}}_{\pi_0/\pi_1} \times \underbrace{\frac{\mathsf{P}(y \,|\, H_0)}{\mathsf{P}(y \,|\, H_1)}}_{B}.$$

The number B, called the *Bayes factor*, tells us how much the data alters our prior belief. In general, the Bayes factor depends on the prior:

$$B = \frac{\mathsf{P}(H_0 \,|\, y)/\pi_0}{\mathsf{P}(H_1 \,|\, y)/\pi_1} = \frac{\int_{\Theta_0} p(\theta \,|\, y)\,\mathrm{d}\theta/\pi_0}{\int_{\Theta_1} p(\theta \,|\, y)\,\mathrm{d}\theta/\pi_1} = \frac{\int_{\Theta_0} p(y \,|\, \theta) p(\theta)/\pi_0\,\mathrm{d}\theta}{\int_{\Theta_1} p(y \,|\, \theta) p(\theta)/\pi_1\,\mathrm{d}\theta}.$$

However, when the parameter space has only two elements, the Bayes factor is the *likelihood ratio*

$$B = \frac{p(y \,|\, \theta_0)}{p(y \,|\, \theta_1)}.$$

which does not depend on the choice of the prior. This interpretation applies in the following example.

Example: Transmission of hemophilia The human X chromosome carries a gene that is essential for normal clotting of the blood. The defect in this gene that is responsible for the blood disease *hemophilia* is recessive: no disease develops in a woman at least one of whose X chromosomes has a normal gene. However, a man whose X chromosome has the defective gene develops the disease. As a result, hemophilia occurs almost exclusively in males who inherit the gene from non-hemophiliac mothers. Great Britain's Queen Victoria (pictured here) carried the hemophilia gene, and it was transmitted through her daughters to many of the royal houses of Europe.

Question. Alice has a brother with hemophilia, but neither she, her parents, nor her two sons (aged 5 and 8) have the disease. What is the probability that she is carrying the hemophilia gene?

Solution. Let H_0 : Alice does not carry the hemophilia gene, and H_1: she does. The X chromosome that Alice inherited from her father does not have the defective gene, because he's healthy. We know that Alice's mother has one X chrosomome with the defective gene, because Alice's brother is sick[8] and her mother is healthy. The X chromosome that Alice inherited from her mother could be the good one or the bad one; let's take

[8] For simplicity, we neglect the fact that hemophilia can also develop spontaneously as a result of a mutation.

$\mathsf{P}(H_0) = \mathsf{P}(H_1) = \frac{1}{2}$ as our prior, that is, we assume prior odds to be 1 to 1. Let Y denote the fact that Alice's two sons are healthy. Because the sons are not identical twins, we can assume

$$\mathsf{P}(Y \,|\, H_1) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}, \quad \mathsf{P}(Y \,|\, H_0) = 1.$$

The posterior probability is then

$$\mathsf{P}(H_1 \,|\, Y) = \frac{\mathsf{P}(Y \,|\, H_1)\mathsf{P}(H_1)}{\mathsf{P}(Y \,|\, H_1)\mathsf{P}(H_1) + \mathsf{P}(Y \,|\, H_0)\mathsf{P}(H_0)} = \frac{\frac{1}{4} \cdot \frac{1}{2}}{\frac{1}{4} \cdot \frac{1}{2} + 1 \cdot \frac{1}{2}} = \frac{1}{5},$$

and $\mathsf{P}(H_0 \,|\, Y) = 1 - \frac{1}{5} = \frac{4}{5}$. The posterior odds in favour of H_0 against H_1 are 4 to 1, much improved (Bayes factor $B = 4$) compared to the prior odds. ■

A *one-sided hypothesis* for a continuous parameter has a form such as $H_0 : \theta \leq \theta_0$, where θ_0 is a given constant. This could represent a statement such as "the new fertilizer doesn't improve yields". After you compute the posterior probability

$$p_0 = \mathsf{P}(H_0 \,|\, y) = \mathsf{P}(\theta \leq \theta_0 \,|\, y) = \int_{-\infty}^{\theta_0} p(\theta \,|\, y)\, \mathrm{d}\theta,$$

you can make straightforward statements such as "The probability that $\theta \leq \theta_0$ is p_0", or "The probability that $\theta > \theta_0$ is $1 - p_0$".

Some practitioners like to mimic Frequentist procedures and choose beforehand a "significance level" α (say, $\alpha = 5\%$), and then if $p_0 < \alpha$ they "reject H_0 (and accept H_1) at the α level of significance". This is all rather convoluted, however, and as we shall see in section 14, a systematic approach should be based on Decision Theory. Simply reporting the probability value p_0 is more direct and informative, and usually suffices.

A *two-sided hypothesis* of the form $H_0 : \theta = \theta_0$ might be used to model statements such as "the new fertilizer doesn't change yields." In Bayesian theory, such a "sharp" hypothesis test with a continuous prior pdf is pointless because it is always false:

$$\mathsf{P}(\theta = \theta_0 \,|\, y) = \int_{\theta_0}^{\theta_0} p(\theta \,|\, y)\, \mathrm{d}\theta = \int_{\theta_0}^{\theta_0} p(y \,|\, \theta) p(\theta)\, \mathrm{d}\theta = 0.$$

Thus it would seem that the question is not a sensible one. It nevertheless arises fairly often, for example when trying to decide whether to add or remove terms to a regression model. How, then, can one deal with such a hypothesis?

- One could test whether θ_0 lies in some credibility interval C_ε. However, this isn't a Bayesian hypothesis test.
- A Bayesian hypothesis test consists of assigning a prior probability π_0 to the hypothesis $H_0 : \theta = \theta_0$, yielding a prior that is a mixture of discrete pmf and continuous pdf.

Hypothesis testing is discussed further in section 13.

9 Simple Multiparameter Models

Often, even though one may need many parameters to define a model, one is only interested in a few of them. For example, in a normal model with unknown mean and variance $y_i \,|\, \mu, \sigma^2 \sim \text{Normal}(\mu, \sigma^2)$, one is usually interested only in the mean μ. The uninteresting parameters are called *nuisance* parameters, and they can simply be integrated

out of the posterior to obtain marginal pdfs of the parameters of interest. Thus, denoting $\theta = (\theta_1, \theta_2)$, where θ_1 is the vector of parameters of interest and θ_2 is the vector of 'nuisance' parameters, we have

$$p(\theta_1 \,|\, y) = \int p(\theta \,|\, y)\, \mathrm{d}\theta_2.$$

This marginalisation integral can also be written as

$$p(\theta_1 \,|\, y) = \int p(\theta_1 \,|\, \theta_2, y) p(\theta_2 \,|\, y)\, \mathrm{d}\theta_2, \tag{11}$$

which expresses $\theta_1 \,|\, y$ as a mixture of conditional posterior distributions given the nuisance parameters, weighted by the posterior density of the nuisance parameters. This explains why the posterior pdf for the parameters of interest is generally more diffuse than $p(\theta_1 \,|\, \theta_2, y)$ for any given θ_2.

9.1 Two-parameter normal model

Consider a normal model with unknown mean and variance, that is, $y_i \,|\, \mu, \sigma^2 \sim \text{Normal}(\mu, \sigma^2)$ with $y = y_1, \ldots, y_n$ conditionally independent given μ, σ^2. The likelihood is then

$$p(y \,|\, \mu, \sigma^2) \propto \sigma^{-n} \mathrm{e}^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\mu)^2} = \sigma^{-n} \mathrm{e}^{-\frac{n(\bar{y}-\mu)^2 + \sum_{i=1}^n (y_i-\bar{y})^2}{2\sigma^2}} = \sigma^{-n} \mathrm{e}^{-\frac{n(\bar{y}-\mu)^2 + (n-1)s^2}{2\sigma^2}},$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ is the sample mean and $s^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2$ is the sample variance.

We assume the following prior information:

- μ and σ^2 are independent.
- $p(\mu) \propto 1$. This improper distribution expresses indifference about the *location* of the mean, because a translation of the origin $\mu' = \mu + c$ gives the same prior.
- $p(\sigma) \propto \frac{1}{\sigma}$. This improper distribution expresses indifference about the *scale* of the standard deviation, because a scaling $\sigma' = c\sigma$ gives the same prior $p(\sigma') = p(\sigma)\, \mathrm{d}\sigma / \mathrm{d}\sigma' \propto \frac{1}{\sigma'}$. Equivalently, we can say that this improper distribution expresses indifference about the location of $\log(\sigma)$, because a flat prior $p(\log(\sigma)) \propto 1$ corresponds to $p(\sigma) = \frac{\mathrm{d}\log(\sigma)}{\mathrm{d}\sigma} p(\log(\sigma)) \propto \frac{1}{\sigma}$. The corresponding prior distribution for the variance is $p(\sigma^2) \propto \frac{1}{\sigma^2}$.

With this prior and likelihood, the joint posterior pdf is

$$p(\mu, \sigma^2 \,|\, y) \propto (\sigma^2)^{-(\frac{n}{2}+1)} \mathrm{e}^{-\frac{(n-1)s^2 + n(\bar{y}-\mu)^2}{2\sigma^2}}, \tag{12}$$

The posterior mode can be found as follows. The mode's μ value is $\bar{y}$ because of the symmetry about $\mu = \bar{y}$. Then, denoting $v = \log\sigma$ and $A = (n-1)s^2 + n(\bar{y}-\mu)^2$, we have

$$\log p(\mu, \sigma^2 \,|\, y) = -(n+2)v - \tfrac{1}{2} A \mathrm{e}^{-2v}$$

Differentiating this with respect to v, equating to zero, and solving gives

$$\mathrm{e}^{-2v} = \frac{n+2}{A}.$$

Substituting $\mathrm{e}^v = \sigma$ and $\mu = \bar{y}$ gives

$$\mathsf{mode}(\mu, \sigma^2 \,|\, y) = (\bar{y}, \frac{n-1}{n+2} s^2). \tag{13}$$

The marginal posterior pdf of σ^2 is obtained by integrating over μ:

$$\begin{aligned} p(\sigma^2\,|\,y) &= \int_{-\infty}^{\infty} p(\mu,\sigma^2\,|\,y)\,\mathrm{d}\mu \\ &= (\sigma^2)^{-(\frac{n}{2}+1)}\mathrm{e}^{-\frac{(n-1)s^2}{2\sigma^2}} \underbrace{\int_{-\infty}^{\infty} \mathrm{e}^{-\frac{n(\mu-\bar{y})^2}{2\sigma^2}}\,\mathrm{d}\mu}_{\sqrt{2\pi\sigma^2/n}} \\ &\propto (\sigma^2)^{-(\frac{n+1}{2})}\mathrm{e}^{-\frac{(n-1)s^2}{2\sigma^2}}, \end{aligned}$$

that is, $\sigma^2\,|\,y \sim \text{InvGam}(\frac{n-1}{2}, \frac{n-1}{2}s^2)$, for which

$$\mathsf{E}(\sigma^2\,|\,y) = \frac{n-1}{n-3}s^2, \quad \mathsf{mode}(\sigma^2\,|\,y) = \frac{n-1}{n+1}s^2, \quad \mathsf{V}(\sigma^2\,|\,y) = \frac{2(n-1)^2s^4}{(n-3)^2(n-5)}.$$

Notice that the mode of the marginal posterior is different from (a bit larger than) the joint posterior mode's σ^2 value given in (13).

The following general result will be useful.

Lemma 1 *If $x\,|\,w \sim \text{Normal}(0,w)$ and $w \sim \text{InvGam}(\frac{m}{2}, \frac{m}{2}S^2)$ then $\frac{x}{S} \sim \mathrm{t}_m$, a standard Student-t distribution with m degrees of freedom.*

Proof: The marginal pdf is

$$\begin{aligned} p(x) &= \int_0^{\infty} p(x\,|\,w)p(w)\,\mathrm{d}w \propto \int w^{-\frac{1}{2}}\mathrm{e}^{-\frac{x^2}{2w}} w^{-(\frac{m}{2}+1)}\mathrm{e}^{-\frac{mS^2}{2w}}\,\mathrm{d}w \\ &= \left(\frac{x^2+mS^2}{2}\right)^{-(m+1)/2} \int z^{-(m+3)/2}\mathrm{e}^{-1/z}\,\mathrm{d}z \quad [\text{ where } z = \frac{2w}{x^2+mS^2} \\ &\propto (x^2+mS^2)^{-(m+1)/2} \quad [\text{ integral of an inverse-gamma pdf} \\ &\propto \left(1+\frac{x^2}{mS^2}\right)^{-(m+1)/2}. \ \square \end{aligned}$$

From (12) we have $\mu\,|\,\sigma^2, y \sim \text{Normal}(\bar{y}, \frac{\sigma^2}{n})$, so that $\frac{\mu-\bar{y}}{1/\sqrt{n}}\,|\,\sigma^2, y \sim \text{Normal}(0,\sigma^2)$. Then by Lemma 1 we have $\frac{\mu-\bar{y}}{s/\sqrt{n}}\,|\,y \sim \mathrm{t}_{n-1}$, that is, $\mu\,|\,y \sim \mathrm{t}_{n-1}(\bar{y}, \frac{s^2}{n})$, and

$$\mathsf{E}(\mu\,|\,y) = \bar{y}, \quad \mathsf{mode}(\mu\,|\,y) = \bar{y}, \quad \mathsf{V}(\mu\,|\,y) = \frac{n-1}{n-3}\frac{s^2}{n}.$$

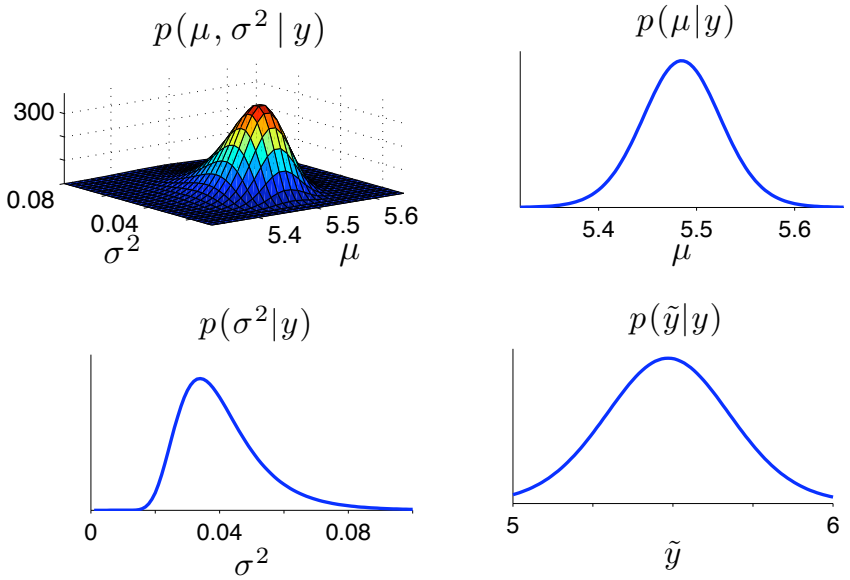
The Student-t distribution roughly resembles a Normal distribution but has heavier tails. This marginal posterior distribution has the same mean as that of the posterior $\mu\,|\,y \sim \text{Normal}(\bar{y}, \frac{v}{n})$ that we found in section 5.1 for the one-parameter normal model with known variance v and uniform prior $p(\mu) \propto 1$.

Next, we find the posterior predictive distribution. The model is $\tilde{y}\,|\,\mu,\sigma^2 \sim \text{Normal}(\mu,\sigma^2)$ (conditionally independent of y given μ,σ^2). Because $\tilde{y}-\mu\,|\,\sigma^2, y \sim \text{Normal}(0,\sigma^2)$ and $\mu\,|\,\sigma^2, y \sim \text{Normal}(\bar{y}, \frac{\sigma^2}{n})$ are independent given σ^2, y, we have $\tilde{y}\,|\,\sigma^2, y \sim \text{Normal}(\bar{y}, (1+\frac{1}{n})\sigma^2)$, that is, $\frac{\tilde{y}-\bar{y}}{(1+\frac{1}{n})^{1/2}}\,|\,\sigma^2, y \sim \text{Normal}(0,\sigma^2)$. Then by Lemma 1, we obtain $\frac{\tilde{y}-\bar{y}}{(1+\frac{1}{n})^{1/2}s} \sim \mathrm{t}_{n-1}$, that is, $\tilde{y}\,|\,y \sim \mathrm{t}_{n-1}(\bar{y}, (1+\frac{1}{n})s^2)$, for which

$$\mathsf{E}(\tilde{y}\,|\,y) = \bar{y}, \quad \mathsf{mode}(\tilde{y}\,|\,y) = \bar{y}, \quad \mathsf{V}(\tilde{y}\,|\,y) = \frac{n-1}{n-3}(1+\frac{1}{n})s^2.$$

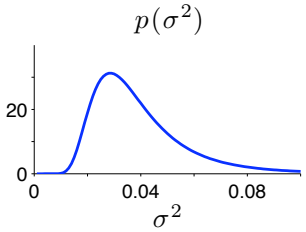
Formulas can also be derived for proper conjugate prior distributions, but we do not present these here. We proceed instead to look at how a two-parameter normal model can be analysed using numerical simulation.

Example: Two-parameter normal model for Cavendish's data We have $n = 23$, $\bar{y} = 5.4848$, and $s^2 = (0.1924)^2 = 0.0370$. With the prior $p(\mu, \sigma^2) = 1/\sigma^2$, the posterior pdf's are:



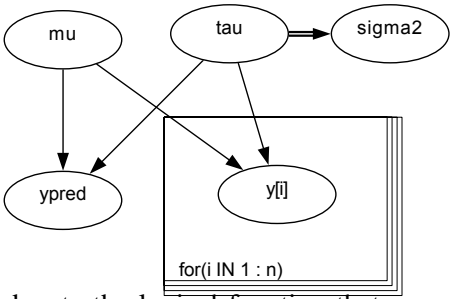
$$\text{mode}(\mu, \sigma^2 \,|\, y) = (5.4848, 0.0326),$$
$$\text{E}(\mu \,|\, y) = 5.4848, \quad \text{mode}(\mu \,|\, y) = 5.4848, \quad \text{V}(\mu \,|\, y) = 0.0018,$$
$$\text{E}(\sigma^2 \,|\, y) = 0.0407, \quad \text{mode}(\sigma^2 \,|\, y) = 0.0339, \quad \text{V}(\sigma^2 \,|\, y) = 1.84 \cdot 10^{-4},$$
$$\text{E}(\tilde{y} \,|\, y) = 5.4848, \quad \text{mode}(\tilde{y} \,|\, y) = 5.4848, \quad \text{V}(\tilde{y} \,|\, y) = 0.0425.$$

For a WinBUGS model, we assume μ and σ^2 to be independent a priori. As in section 4.2, we choose $\mu \sim \text{Normal}(5, 0.5)$. Our prior for σ^2 is based on the judgement that $\sigma^2 \approx 0.04 \pm 0.02$. Assuming $\sigma^2 \sim \text{InvGam}(\alpha, \beta)$ and solving $\text{E}(\sigma^2) = 0.04$ and $\text{V}(\sigma^2) = 0.02^2$ for α and β, we obtain $\sigma^2 \sim \text{InvGam}(6, 0.2)$. The corresponding prior distribution for the precision $\tau = 1/\sigma^2$ is $\tau \sim \text{Gamma}(6, 0.2)$.



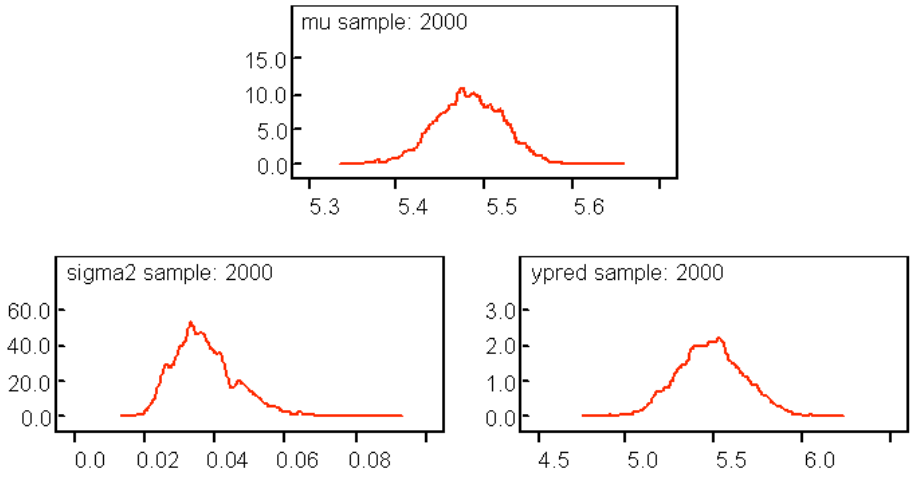
The WinBUGS model is

```
model {
   for (i in 1:n) { y[i] ~ dnorm(mu,tau) }
   mu ~ dnorm(5,2)
   tau ~ dgamma(6,0.2)
   sigma2 <- 1/tau
   ypred ~ dnorm(mu,tau)
}
```



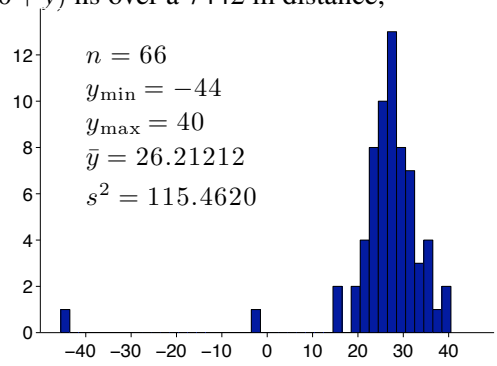
The double link in the DAG and the "<-" in the code denote the logical function that specifies σ^2 as a deterministic function of τ. The results after 2000 simulation steps are

node	mean	sd	2.5%	median	97.5%
mu	5.483	0.04028	5.403	5.483	5.56
sigma2	0.03767	0.0097	0.023	0.036	0.060
ypred	5.483	0.195	5.116	5.482	5.869

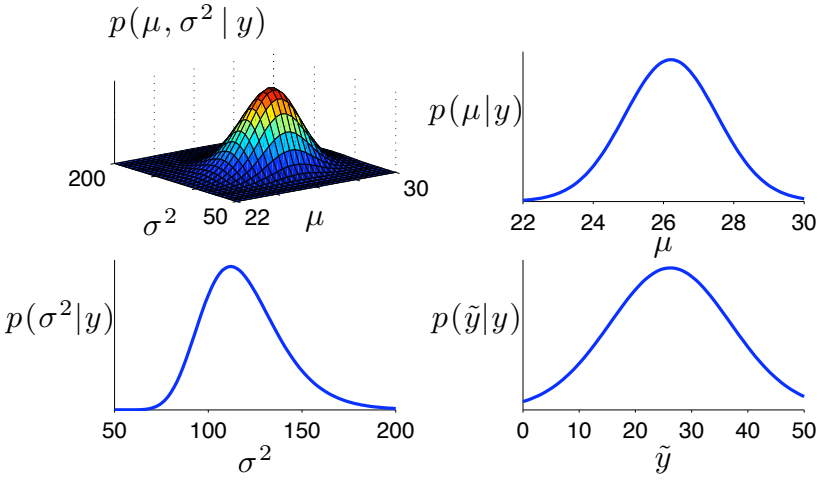


Example: Robust Models for Newcomb's data The American astronomer and mathematician Simon Newcomb performed experiments in 1882 to measure the speed of light. He repeatedly measured the travel time $t = (24800 + y)$ ns over a 7442 m distance, obtaining the following results:

28	26	33	24	34	-44	27	16	40	-2
29	22	24	21	25	30	23	29	31	19
24	20	36	32	36	28	25	21	28	29
37	25	28	26	30	32	36	26	30	22
36	23	27	27	28	27	31	27	26	33
26	32	32	24	39	28	24	25	32	25
29	27	28	29	16	23				



Assuming $y_i \,|\, \mu, \sigma^2 \sim \text{Normal}(\mu, \sigma^2)$ with $y = y_1, \ldots, y_n$ conditionally independent given μ, σ^2, and the noninformative prior $p(\mu, \sigma^2) \propto 1/\sigma^2$, the posterior pdf's are:



and the summarising statistics are

$$\text{mode}(\mu, \sigma^2 \,|\, y) = (26.2121, 110.3681),$$
$$\text{E}(\mu \,|\, y) = 26.2121, \quad \text{mode}(\mu \,|\, y) = 26.2121, \quad \text{V}(\mu \,|\, y) = 1.8050,$$
$$\text{E}(\sigma^2 \,|\, y) = 119.1275, \quad \text{mode}(\sigma^2 \,|\, y) = 112.0154, \quad \text{V}(\sigma^2 \,|\, y) = 116.3226,$$
$$\text{E}(\tilde{y} \,|\, y) = 26.2121, \quad \text{mode}(\tilde{y} \,|\, y) = 26.2121, \quad \text{V}(\tilde{y} \,|\, y) = 120.9324.$$

Because the standard t_{65} distribution has 95% of its probability in the interval $[-1.997, 1.997]$, a 95% credibility interval for μ is $\bar{y} \pm 1.997 s/\sqrt{66} = [23.57, 28.85]$. Notice the lack of resemblance between the predictive posterior density and the histogram of actual measurements. This indicates that the normal model is apparently not a good description of the variation in this data, which contains two obvious "outliers".

One approach to analysing normally-distributed data that is "corrupted" with outliers is to model the data as being a mixture of "good" and "bad" observations. A good observation is assumed to be Normal(μ, σ^2) and to occur with probability $p_1 = 1 - \varepsilon$, while a bad observation is assumed to be Normal$(\mu, k_2\sigma^2)$ and to occur with probability $p_2 = \varepsilon$, where $k_2 \gg 1$.

A WinBUGS implementation of a mixture model uses a stochastic index r_i, a categorical random variable with prior pmf

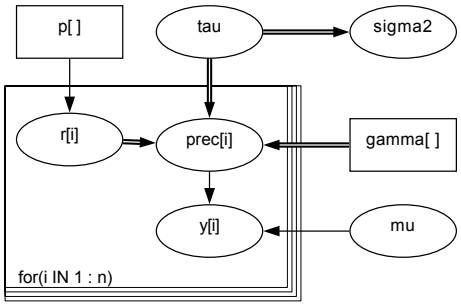
$$\mathsf{P}(r_i = 1) = 1 - \varepsilon, \quad \mathsf{P}(r_i = 2) = \varepsilon$$

The likelihood for the mixture model is

$$y_i \,|\, \mu, \sigma^2, r_i \sim \text{Normal}(\mu, k_{r_i}\sigma^2)$$

where $k_1 = 1$ and $k_2 \gg 1$. Here's a WinBUGS implementation of a mixture model for the Newcomb data:

```
model {
   k[1]<-1; k[2]<-10
   p[1]<-0.95; p[2]<-0.05
   for( i in 1 : n ) {
      y[i] ~ dnorm(mu,prec[i])
      r[i] ~ dcat(p[ ])
      prec[i] <- tau / k[r[i]]
   }
   tau ~ dgamma(0.1,0.1)
   mu ~ dnorm(25,0.01)
   sigma2 <- 1 / tau
}
```

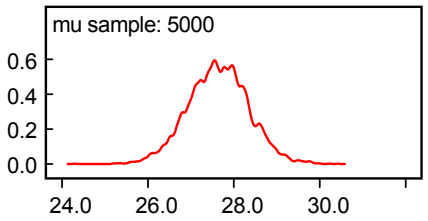


The Newcomb data are entered as

```
list(y=c(28,26,33,24,34,-44,27,16,40,-2,29,22,24,21,25, 30,23,29,31,19,
         24,20,36,32,36, 28,25,21,28,29,37,25,28,26,30, 32,36,26,30,22,
         36,23,27,27,28, 27,31,27,26,33,26,32,32,24,39, 28,24,25,32,25,
         29,27,28,29,16, 23), n=66)
```

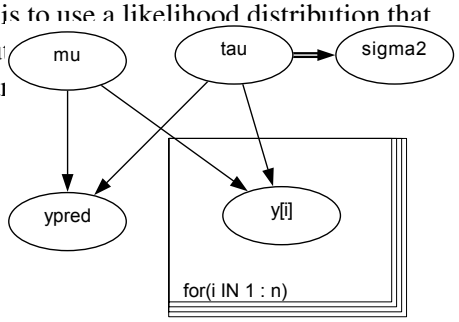
The results after 5000 simulation steps are

node	mean	sd	2.5%	median	97.5%
mu	27.64	0.711	26.24	27.64	29.02
sigma2	29.13	5.824	19.94	29.48	41.75



We see that with this mixture model the effect of the outliers is greatly reduced, compared to the normal model, and we obtain $\mu \,|\, y$ with larger mean and less dispersion than the pure normal model considered earlier. Also, the "good" data's variance $\sigma^2 \,|\, y$ is much smaller than the variance inferred using the pure normal model.

Another way to deal with outlier-corrupted data is to use a likelihood distribution that has "heavier" tails, for example a Student-t distribu



WinBUGS model using $\eta = 4$ degrees of freedom a

```
model {
   for (i in 1:n) { y[i] ~ dt(mu,tau,4)
   mu ~ dnorm(0,0.001)
   tau ~ dgamma(0.001,0.001)
   sigma2 <- 1/tau
}
```

The Monte Carlo simulation starting values are set to

```
list(mu=0,tau=0.01)
```

The results after 5000 simulation steps are similar to those of the mixture model:

node	mean	sd	2.5%	median	97.5%
mu	27.48	0.668	26.11	27.48	28.81
sigma2	27.4	4.676	18.09	27.42	33.8

9.2 Comparing two normal populations

Consider two sets of measurements:

$$x_i \,|\, \lambda, \phi \sim \text{Normal}(\lambda, \phi), \quad y_i \,|\, \mu, \psi \sim \text{Normal}(\mu, \psi),$$

with $x_1, \ldots, x_m, y_1, \ldots, y_n$ mutually independent given λ, μ, ϕ, ψ. The parameter of interest is the difference of the means, $\delta = \lambda - \mu$. We consider three cases, in increasing order of difficulty.

Variances known

Assuming noninformative "flat" prior $p(\lambda, \mu) \propto 1$, we obtain the posteriors

$$\lambda \,|\, x, y \sim \text{Normal}(\bar{x}, \frac{\phi}{m}), \quad \mu \,|\, x, y \sim \text{Normal}(\bar{y}, \frac{\psi}{n}),$$

with $\lambda \,|\, x, y$ and $\mu \,|\, x, y$ independent given x, y. The posterior for the difference $\delta = \lambda - \mu$ is then

$$\delta \,|\, x, y \sim \text{Normal}(\bar{x} - \bar{y}, \frac{\phi}{m} + \frac{\psi}{n}).$$

This solution can easily be generalised to models with proper conjugate (i.e. normal) priors, as in section 4.2.

Variances unknown but equal

Assuming $\phi = \psi$ and the noninformative prior $p(\lambda, \mu, \phi) \propto 1 \cdot 1 \cdot \frac{1}{\phi}$, we have

$$\begin{aligned} p(\lambda, \mu, \phi \,|\, x, y) &\propto p(x, y \,|\, \lambda, \mu, \phi) p(\lambda, \mu, \phi) \\ &\propto \phi^{-m/2} \mathrm{e}^{-\frac{m(\bar{x}-\lambda)^2 + (m-1)s_x^2}{2\phi}} \phi^{-n/2} \mathrm{e}^{-\frac{n(\bar{y}-\mu)^2 + (n-1)s_y^2}{2\phi}} \phi^{-1} \\ &\propto \phi^{-(m+n)/2} \mathrm{e}^{-\frac{(m+n-2)s^2}{2\phi}} \times \phi^{-1/2} \mathrm{e}^{-\frac{m(\bar{x}-\lambda)^2}{2\phi}} \times \phi^{-1/2} \mathrm{e}^{-\frac{n(\bar{y}-\mu)^2}{2\phi}}, \end{aligned}$$

where $s^2 = \frac{(m-1)s_x^2+(n-1)s_y^2}{m+n-2}$. Marginalising this over λ, μ gives

$$p(\phi \,|\, x, y) = \iint p(\lambda, \mu, \phi \,|\, x, y)\, \mathrm{d}\lambda\, \mathrm{d}\mu \propto \phi^{-(m+n)/2} \mathrm{e}^{-\frac{(m+n-2)s^2}{2\phi}},$$

that is, $\phi \,|\, x, y \sim \text{InvGam}(\frac{m+n-2}{2}, \frac{m+n-2}{2}s^2)$.

Now, because

$$\begin{aligned} p(\lambda, \mu \,|\, \phi, x, y) &\propto p(x, y \,|\, \lambda, \mu, \phi) \underbrace{p(\lambda, \mu \,|\, \phi)}_{\propto 1} \\ &\propto p(x \,|\, \lambda, \phi) p(y \,|\, \mu, \phi) \\ &\propto \phi^{-m/2} \mathrm{e}^{-\frac{m(\bar{x}-\lambda)^2+(m-1)s_x^2}{2\phi}} \phi^{-n/2} \mathrm{e}^{-\frac{n(\bar{y}-\mu)^2+(n-1)s_y^2}{2\phi}}. \end{aligned}$$

it follows that $\lambda \,|\, \phi, x, y \sim \text{Normal}(\bar{x}, \frac{\phi}{m})$ and $\mu \,|\, \phi, x, y \sim \text{Normal}(\bar{y}, \frac{\phi}{n})$ are independent given ϕ, x, y, and so $\delta \,|\, \phi, x, y \sim \text{Normal}(\bar{x}-\bar{y}, (\frac{1}{m}+\frac{1}{n})\phi)$, that is, $\frac{\delta-(\bar{x}-\bar{y})}{(\frac{1}{m}+\frac{1}{n})^{1/2}} \,|\, \phi, x, y \sim \text{Normal}(0, \phi)$. Then, by Lemma 1, we obtain $\frac{\delta-(\bar{x}-\bar{y})}{(\frac{1}{m}+\frac{1}{n})^{1/2}s} \,|\, x, y \sim \mathrm{t}_{m+n-2}$, that is,

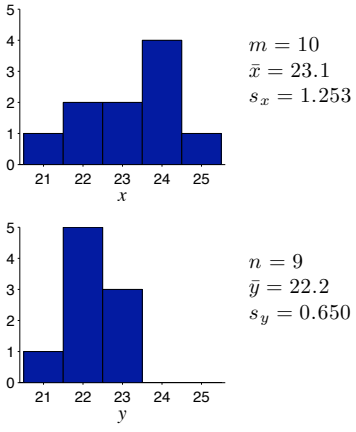
$$\delta \,|\, x, y \sim \mathrm{t}_{m+n-2}(\bar{x}-\bar{y}, (\tfrac{1}{m}+\tfrac{1}{n})s^2).$$

Variances unknown

In this case $\delta \,|\, x, y$ cannot be expressed in terms of any of the standard statistical distributions, but is readily found using numerical simulation.

Example: Cuckoo eggs Cuckoo (*Cuculus canorus*) eggs found in $m = 10$ dunnock (*Prunella modularis*) nests have the following diameters in mm (denoted x_i):

```
22.0, 23.9, 20.9, 23.8, 25.0,
24.0, 21.7, 23.8, 22.8, 23.1
```



The diameters (y_i) of cuckoo eggs found in $n = 9$ sedge warbler (*Acrocephalus schoenobaenus*) nests are

```
23.2, 22.0, 22.2, 21.2, 21.6,
21.9, 22.0, 22.9, 22.8
```

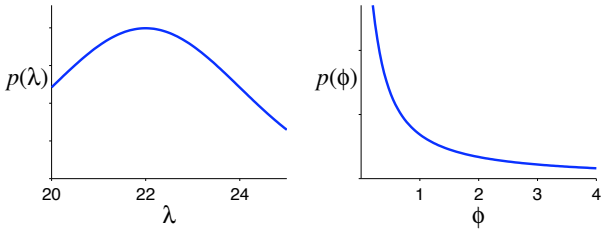
We assume that the data come from normally distributed populations with unknown means and variances,

$$x_i \,|\, \lambda, \phi \sim \text{Normal}(\lambda, \phi), \quad y_i \,|\, \mu, \psi \sim \text{Normal}(\mu, \psi),$$

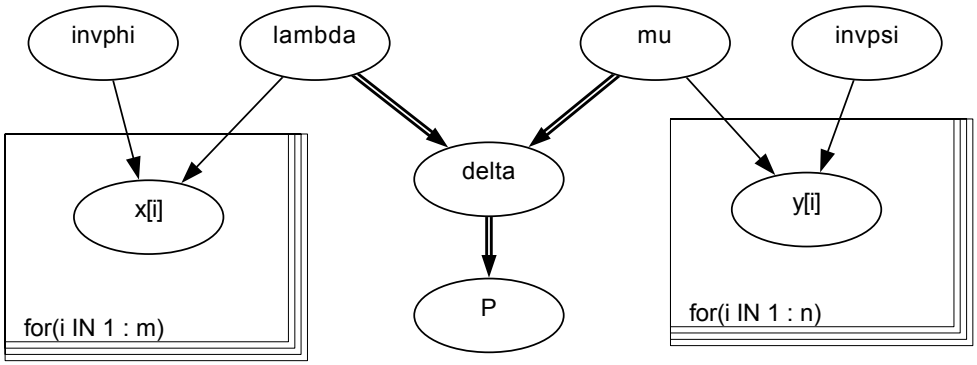
with $x_1, \ldots, x_m, y_1, \ldots, y_n$ mutually independent given λ, μ, ϕ, ψ. The parameter of interest is the difference of the means, $\delta = \lambda - \mu$. In particular, we are interested in the hypothesis that the difference is greater than zero, i.e. do cuckoos lay bigger eggs in the nests of dunnocks than in the nests of sedge warblers?

We assume proper, relatively vague priors

$$\lambda \sim \text{Normal}(22, 4),\ \mu \sim \text{Normal}(22, 4),\ \phi \sim \text{InvGam}(0.1, 0.1),\ \psi \sim \text{InvGam}(0.1, 0.1).$$



The WinBUGS model is




```
model {
   for (i in 1:m) { x[i] ~ dnorm(lambda,invphi) }
   for (i in 1:n) { y[i] ~ dnorm(mu,invpsi) }
   lambda ~ dnorm(22,0.25)
   mu ~ dnorm(22,0.25)
   invphi ~ dgamma(0.1,0.1)
   invpsi ~ dgamma(0.1,0.1)
   delta <- lambda-mu
   P <- step(delta-0)
}
```

Here we use the function `step`, which equals 1 if its argument is ≥ 0 and which equals 0 otherwise, to compute

$$\mathrm{P}(\delta \geq 0 \,|\, x, y) = \int_0^\infty p(\delta' \,|\, x, y)\, \mathrm{d}\delta'.$$

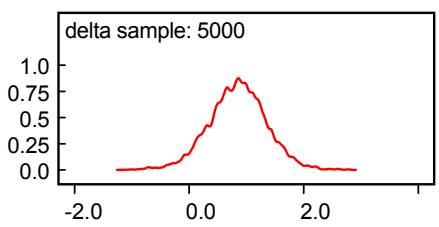
Also, because WinBUGS uses the reciprocal of variance to specify the normal distribution, the model has the variables `invphi` and `invpsi` for $\frac{1}{\phi}$ and $\frac{1}{\psi}$.

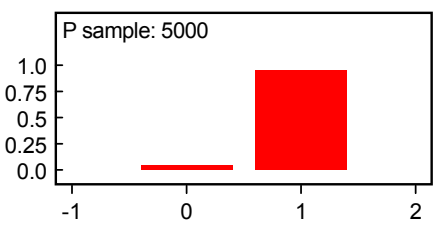
The data is entered as

```
list( x=c(22.0,23.9,20.9,23.8,25.0,24.0,21.7,23.8,22.8,23.1),m=10,
        y=c(23.2,22.0,22.2,21.2,21.6,21.9,22.0,22.9,22.8),n=9)
```

The results after 5000 simulation steps are

node	mean	sd	2.5%	median	97.5%
`delta`	0.8489	0.5021	-0.1599	0.8584	1.826
`P`	0.9542				





The posterior probability of the hypothesis $\delta \geq 0$ is over 95%, that is, the odds are over 20 to 1 in favour of the hypothesis against its alternative.

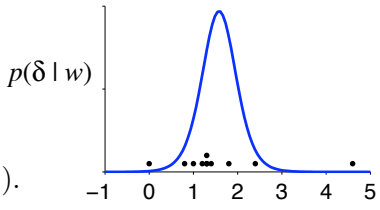
Example: Paired observations The following table is data on the extra hours of sleep gained by $n = 10$ insomnia patients who at different times were given treatment A and treatment B.

patient i	1	2	3	4	5	6	7	8	9	10
gain x_i with A	1.9	0.8	1.1	0.1	-0.1	4.4	5.5	1.6	4.6	3.4
gain y_i with B	0.7	-1.6	-0.2	-1.2	-0.1	3.4	3.7	0.8	0	2.0
$w_i = x_i - y_i$	1.2	2.4	1.3	1.3	0	1.0	1.8	0.8	4.6	1.4

Suppose we are interested in the difference between the effects of the two treatments. The model for comparing two normal populations should *not* be used to analyse this data, because the measurements are not independent: the responses of a single patient to different treatments can be expected to be more similar than the responses from two different patients. In this case, a "paired observations" experimental design is a good way to detect the difference between the treatment effects.

The results can be analysed using a model of the form $w_i \,|\, \delta, \varphi \sim \text{Normal}(\delta, \varphi)$, assumed mutually independent given δ, φ. The sufficient statistics are $\bar{w} = 1.58$ and $s^2 = \frac{1}{n-1}\sum(w_i - \bar{w})^2 = 1.513$. Assuming a noninformative prior $p(\delta, \varphi) \propto \frac{1}{\varphi}$ as in section 9.1, we obtain the marginal posterior

$$\delta \,|\, w \sim \text{t}_{n-1}(\bar{w}, s^2/n) = \text{t}_9(1.58, 0.1513).$$



In particular, the posterior probability for the hypothesis that treatment A is more effective than treatment B is

$$\text{P}(\delta > 0 \,|\, w) = \int_0^\infty p(\delta' \,|\, w)\, \text{d}\delta' = 0.9986,$$

that is, the odds are over 700 to 1 in favour of the hypothesis against the alternative.

9.3 Multinomial model

The multinomial model is a generalisation of the binomial model (section 5.2): instead of two possible results, an observation can have k possible outcomes, $x_i \in \{X_1, X_2, \ldots, X_k\}$, with corresponding probabilities $\theta = [\theta_1, \ldots, \theta_k]$, $\sum_{j=1}^k \theta_j = 1$. The pmf for a single observation x_i is

$$p(x_i \,|\, \theta) = \text{P}(x_i = X_j \,|\, \theta) = \theta_j \quad (j \in \{1, \ldots, k\}).$$

The likelihood pmf of a sequence $x_1, \ldots, x_n$ whose elements are assumed to be mutually independent given θ is then

$$p(x_{1:n} \,|\, \theta) = \theta_1^{y_1} \ldots \theta_k^{y_k},$$

where y_j is the number of elements of x whose value is X_j. (Note that $y_j \in \{0, \ldots, n\}$ and $\sum_{j=1}^k y_j = n$.) The likelihood pmf for the sequence $y = y_1, \ldots, y_k$ (a sufficient statistic) is

$$p(y \,|\, \theta) = \frac{n!}{y_1! y_2! \cdots y_k!} \theta_1^{y_1} \ldots \theta_k^{y_k} \propto \prod_{j=1}^k \theta_j^{y_j},$$

and the distribution is denoted $y \,|\, \theta \sim \text{Multinomial}(\theta_1, \cdots, \theta_n)$. In the case $k = 2$ this is the binomial distribution.

The conjugate prior for the multinomial likelihood is the Dirichlet distribution $\theta \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k)$ with positive parameters $\alpha_1, \ldots, \alpha_k$, whose density is

$$p(\theta) \propto \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1} \quad (\theta_j \geq 0, \sum_{j=1}^{k} \theta_j = 1).$$

Summarising statistics of the Dirichlet distribution are

$$\begin{aligned} \mathsf{E}(\theta_j) &= \frac{\alpha_j}{A}, \quad \mathsf{mode}(\theta_j) = \frac{\alpha_j - 1}{A - k}, \\ \mathsf{V}(\theta_j) &= \frac{\alpha_j(A - \alpha_j)}{A^2(A+1)}, \quad \mathsf{cov}(\theta_j, \theta_l) = \frac{-\alpha_j \alpha_l}{A^2(A+1)} \ (j \neq l) \end{aligned}$$

where $A = \sum \alpha_j$. A small value of A corresponds to a prior that is relatively "noninformative". Note that $\theta \sim \text{Dirichlet}(\alpha_1, \alpha_2)$ implies $\theta_1 \sim \text{Beta}(\alpha_1, \alpha_2)$, that the marginal distributions of $\theta \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k)$ are $\theta_j \sim \text{Beta}(\alpha_j, A - \alpha_j)$, and that a density that is uniform over the simplex $\{\theta : \theta_1, \ldots, \theta_k \geq 0, \sum_{j=1}^{k} \theta_j = 1\}$ is obtained with $\alpha_1 = \cdots = \alpha_k = 1$.

With likelihood $y \,|\, \theta \sim \text{Multinomial}(\theta_1, \cdots, \theta_k)$ and prior $\theta \sim \text{Dirichlet}(\alpha_{1:k})$, the posterior pdf is

$$p(\theta \,|\, y) \propto \prod_{j=1}^{k} \theta_j^{y_j} \cdot \prod_{j=1}^{k} \theta_j^{\alpha_j - 1} = \prod_{j=1}^{k} \theta_j^{\alpha_j + y_j - 1},$$

that is, $\theta \,|\, y \sim \text{Dirichlet}(\alpha_1 + y_1, \ldots, \alpha_k + y_k)$. Summarising statistics of the posterior distribution are

$$\begin{aligned} \mathsf{E}(\theta_j \,|\, y) &= \frac{\alpha_j + y_j}{A + n}, \quad \mathsf{mode}(\theta_j \,|\, y) = \frac{\alpha_j + y_j - 1}{A + n - k}, \\ \mathsf{V}(\theta_j) &= \frac{(\alpha_j + y_j)(A + n - \alpha_j - y_j)}{(A+n)^2(A+n+1)}, \quad \mathsf{cov}(\theta_j, \theta_l) = \frac{-(\alpha_j + y_j)(\alpha_l + y_l)}{(A+n)^2(A+n+1)} \ (j \neq l) \end{aligned}$$

Example: opinion survey with six choices In September 2001, 1962 Finnish adults were interviewed and reported their support for political parties as follows:

party	SDP	Kesk	Kok	Vihr	Vas	other
#	471	453	396	243	177	222
%	24.0	23.1	20.2	12.4	9.0	11.2

With the uniform prior $\theta \sim \text{Dirichlet}(1, 1, 1, 1, 1, 1)$, we obtain the posterior

$$\theta \,|\, y \sim \text{Dirichlet}(472, 454, 397, 244, 178, 223).$$

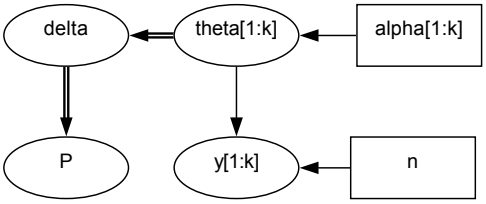
The marginals have the following summaries.

	SDP	Kesk	Kok	Vihr	Vas	other
$E(\theta_i \,|\, y)$	0.240	0.231	0.202	0.124	0.090	0.113
$1.96\sqrt{\mathsf{V}(\theta_i \,|\, y)}$	0.019	0.019	0.018	0.015	0.013	0.014

Thus, for example, the 95% credibility interval for SDP support (using the normal approximation) is $(24.0 \pm 1.9)\% = [22.1, 25.9]\%$.

The hypothesis that SDP support is higher than Keskusta support can be investigated with the following WinBUGS model.

```
model {
   y[1:k] ~ dmulti(theta[1:k],n)
   theta[1:k] ~ ddirch(alpha[1:k])
   delta <- theta[1]-theta[2]
   P <- step(delta)
}
```

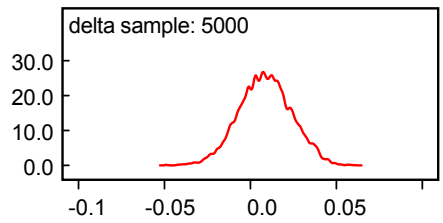


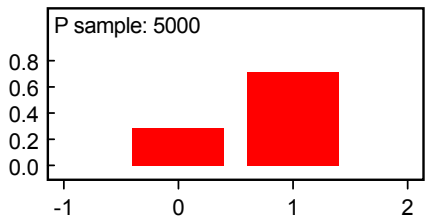
The input data is

```
list(k=6,alpha=c(1,1,1,1,1,1),y=c(471,453,396,243,177,222),n=1962)
```

The simulation results are

node	mean	sd	2.5%	median	97.5%
delta	0.008791	0.01529	-0.02177	0.008713	0.03855
P	0.7164				





The results indicate that $\mathsf{P}(\theta_{\text{sdp}} > \theta_{\text{kesk}} \,|\, y) \approx 0.72$, that is, the posterior odds that SDP has more support than Keskusta are roughly 5 to 2.

10 The modal approximation and Laplace's method

In many of the examples presented earlier in these notes, we have indicated how a posterior distribution can be approximated by a normal distribution that is based on matching the moments (mean and variance), that is,

$$\theta \,|\, y \dot{\sim} \text{Normal}(\mathsf{E}(\theta \,|\, y), \mathsf{V}(\theta \,|\, y)). \tag{14}$$

This approximation gives the convenient formula $\mathsf{E}(\theta \,|\, y) \pm 1.96\sqrt{\mathsf{V}(\theta \,|\, y)}$ for the 95% credibility interval, which was used in the following examples:

- opinion survey (binomial model, §5.2)
- moose counts (Poisson model, §5.3)
- lifetime data (Exponential model, §5.4)

The normal approximation can be expected to be accurate if the distribution has a single sharp peak and is not too skewed.

Moment-matching requires integrals (mean and variance); here's an alternative approximation that is based on derivatives. Let $f(\theta)$ be a nonnegative unimodal function with mode $\hat{\theta}$. The quadratic Taylor approximation of $\log f(\theta)$ about $\hat{\theta}$ is

$$\log f(\theta) \approx \log f(\hat{\theta}) - \tfrac{1}{2}(\theta - \hat{\theta})^T Q(\theta - \hat{\theta}),$$

where

$$Q_{ij} = -\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f(\theta)\right]_{\theta=\hat{\theta}}.$$

Taking exponentials of both sides gives

$$f(\theta) \approx f(\hat{\theta}) \mathrm{e}^{-\frac{1}{2}(\theta-\hat{\theta})^T Q(\theta-\hat{\theta})}. \tag{15}$$

Integrating (15) gives

$$\int f(\theta)\,\mathrm{d}\theta \approx f(\hat{\theta}) \int \mathrm{e}^{-\frac{1}{2}(\theta-\hat{\theta})^T Q(\theta-\hat{\theta})}\,\mathrm{d}\theta = \frac{f(\hat{\theta})}{\sqrt{\det(Q/(2\pi))}}. \tag{16}$$

This approximate integration formula for unimodal sharp-peaked nonnegative functions is known as *Laplace's method*.

In the case where $f(\theta) = p(\theta)p(y\,|\,\theta)$ is an unnormalised posterior density, (15) gives the *modal approximation* of the posterior as

$$\theta\,|\,y \mathrel{\dot\sim} \text{Normal}(\hat{\theta}, Q^{-1}). \tag{17}$$

The Laplace approximation of the normalising constant $\int p(\theta)p(y\,|\,\theta)\,\mathrm{d}\theta$ of the unnormalised posterior (called the *evidence* of the statistical model) is given by (16).

Example: Approximating a gamma distribution For $\theta \sim \text{Gamma}(\alpha, \beta)$, we have

$$p(\theta) \propto \underbrace{\theta^{\alpha-1}\mathrm{e}^{-\beta\theta}}_{f(\theta)}.$$

Using the tabulated formulas for the mean and variance of a gamma distribution, we obtain the moment-matching normal approximation $\theta \mathrel{\dot\sim} \text{Normal}(\frac{\alpha}{\beta}, \frac{\alpha}{\beta^2})$.

To find the modal approximation, we compute

$$\begin{aligned}
\log f &= (\alpha-1)\log\theta - \beta\theta \\
\frac{\mathrm{d}}{\mathrm{d}\theta}\log f &= \frac{\alpha-1}{\theta} - \beta \\
\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f &= -\frac{\alpha-1}{\theta^2}
\end{aligned}$$

The mode is found by solving $\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(\theta) = 0$, yielding $\hat{\theta} = \frac{\alpha-1}{\beta}$. Then $Q = \frac{\alpha-1}{\hat{\theta}^2} = \frac{\beta^2}{\alpha-1}$. The modal approximation (17) is thus $\theta \mathrel{\dot\sim} \text{Normal}(\frac{\alpha-1}{\beta}, \frac{\alpha-1}{\beta^2})$, which is close to the moment-matching approximation when $\beta \gg 1$.

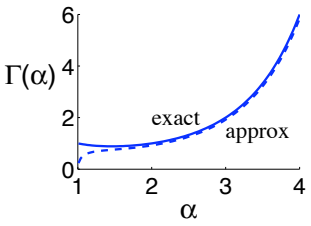
Laplace's approximation of the normalisation factor is

$$\int_0^\infty \theta^{\alpha-1}\mathrm{e}^{-\beta\theta}\,\mathrm{d}\theta \approx \frac{\hat{\theta}^{\alpha-1}\mathrm{e}^{-\beta\hat{\theta}}}{\sqrt{Q/(2\pi)}} = \sqrt{\frac{2\pi(\alpha-1)}{\beta^2}}\left(\frac{\alpha-1}{\beta}\right)^{\alpha-1}\mathrm{e}^{-\alpha+1}$$

Using the fact that the exact normalisation factor is $\Gamma(\alpha)/\beta^\alpha$, we arrive at the following approximation formula for the Gamma function:

$$\Gamma(\alpha) \approx \sqrt{2\pi}(\alpha-1)^{\alpha-\frac{1}{2}}\mathrm{e}^{-\alpha+1}.$$

This approximation is reasonably accurate for $\alpha > 2$.



Example: Two-parameter normal model In §9.1, the posterior for a normal model with unknown mean and variance and flat prior on μ and $\log\sigma$ was derived as

$$p(\mu,\sigma^2\,|\,y) \propto (\sigma^2)^{-(\frac{n}{2}+1)}\,\mathrm{e}^{-\dfrac{(n-1)s^2+n(\bar{y}-\mu)^2}{2\sigma^2}},$$

where $\bar{y}=\frac{1}{n}\sum_{i=1}^n y_i$ and $s^2=\frac{1}{n-1}\sum_{i=1}^n(y_i-\bar{y})^2$. Denoting $v=\log\sigma$, the density as a function of (μ,v) is

$$p(\mu,v\,|\,y) \propto \underbrace{\mathrm{e}^{-nv-\dfrac{(n-1)s^2+n(\bar{y}-\mu)^2}{2\mathrm{e}^{2v}}}}_{f(\mu,v)},$$

We have

$$\log f = -nv-\frac{(n-1)s^2+n(\bar{y}-\mu)^2}{2\mathrm{e}^{2v}}$$

The first-order derivatives (i.e. components of the gradient vector) are

$$\begin{aligned}\frac{\partial}{\partial\mu}\log f &= \frac{n(\bar{y}-\mu)}{\mathrm{e}^{2v}}\\ \frac{\partial}{\partial v}\log f &= -n+\frac{(n-1)s^2+n(\bar{y}-\mu)^2}{\mathrm{e}^{2v}}.\end{aligned}$$

Equating the gradient to zero and solving gives the mode[9]

$$(\hat{\mu},\hat{v}) = \mathsf{mode}(\mu,v\,|\,y) = \left(\bar{y},\ \tfrac{1}{2}\log(\frac{n-1}{n}s^2)\right).$$

The second-order derivatives are

$$\begin{aligned}\frac{\partial^2}{\partial\mu^2}\log f &= -\frac{n}{\mathrm{e}^{2v}}\\ \frac{\partial^2}{\partial\mu\partial v}\log f &= -\frac{2n(\bar{y}-\mu)}{\mathrm{e}^{2v}}\\ \frac{\partial^2}{\partial v^2}\log f &= -2\frac{(n-1)s^2+n(\bar{y}-\mu)^2}{\mathrm{e}^{2v}},\end{aligned}$$

so that $Q=\begin{pmatrix}\frac{n^2}{(n-1)s^2} & 0\\ 0 & 2n\end{pmatrix}$, and the modal approximation of the distribution is

$$\mu,\log\sigma\,|\,y \,\dot\sim\, \mathrm{Normal}\left(\left(\bar{y},\ \tfrac{1}{2}\log(\frac{n-1}{n}s^2)\right),\begin{pmatrix}(n-1)s^2/n^2 & 0\\ 0 & 1/(2n)\end{pmatrix}\right).$$

Note that μ and $\log\sigma$ are not conditionally independent given y, but in the modal approximation they are. The approximate marginal posterior distribution of μ is

$$\mu\,|\,y \,\dot\sim\, \mathrm{Normal}(\bar{y},(n-1)s^2/n^2).$$

For large n this agrees well with the exact marginal posterior $\mu\,|\,y\sim \mathrm{t}_{n-1}(\bar{y},\frac{s^2}{n})$ found in §9.1, for which

$$\mathsf{E}(\mu\,|\,y)=\bar{y},\quad \mathsf{mode}(\mu\,|\,y)=\bar{y},\quad \mathsf{V}(\mu\,|\,y)=\frac{(n-1)s^2}{n(n-3)}.$$

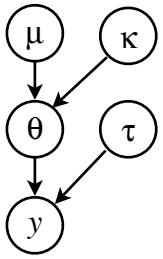
[9] This mode differs from $\mathsf{mode}(\mu,\sigma^2\,|\,y)=(\bar{y},\frac{n-1}{n+2}s^2)$ derived in §9.1 because of the change of variables $v=\log\sigma$.

11 Hierarchical Models and Regression Models

11.1 DAGs

So far we have seen relatively simple statistical models, in which a few quantities of interest are described by standard probability distributions having one or two parameters. More realistic models of complex phenomena will naturally involve many quantities, and the distributions' parameters can themselves be treated as unknown quantities with distributions that have their own parameters, called *hyperparameters*. Models with hyperparameters (and hyperhyperparameters and ...) are called *hierarchical models*.

Directed acyclic graphs (DAGs) are useful for representing Bayesian models, especially complex hierarchical models. The DAG shows how the joint probability distribution can be factored into a product of conditional distributions, because the absence of an arrow between two nodes implies that they are conditionally independent given all other nodes that precede either of them in the graph. For example, the DAG on the right corresponds to a joint density of the form



$$p(y,\theta,\mu,\kappa,\tau) = p(y\,|\,\theta,\tau)p(\theta\,|\,\mu,\kappa)p(\tau)p(\mu)p(\kappa)$$

11.2 Hierarchical normal model

Suppose we have data from K groups, with n_i independent observations from each group. Assume the observations are conditionally independent given θ_i, normally distributed with mean θ_i and known variance v, that is,

$$y_{ij}\,|\,\theta_i \sim \text{Normal}(\theta_i, v) \quad j \in \{1,\ldots,n_i\}, i \in \{1,\ldots,K\}$$

These might for example be

- the responses of patients to K treatments,
- national matriculation exam grades obtained by students from K schools,
- points scored in games played by K baseball teams...

A simple nonhierarchical approach would be to model each group separately, each with their own a-priori independent parameters θ_i. Then the likelihood for each θ_i is

$$\bar{y}_i\,|\,\theta_i \sim \text{Normal}(\theta_i, s_i^2),$$

where $\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$ and $s_i^2 = v/n_i$.

An alternative approach would be to consider that all the observations are estimating a common effect. Then we would pool all the measurements into a single set of observations assumed to come from the same distribution with common parameter θ.

Both approaches have drawbacks: the first neglects the effect that is common to all groups, while the second neglects the effects that are specific to each group. The following hierarchical model allows us to combine information without assuming that all the θ_i are equal. Assume the θ_i are normally distributed with common mean μ and precision κ:

$$\theta_i\,|\,\theta,\kappa \sim \text{Normal}(\mu, \kappa^{-1})$$

The θ_i are assumed to be conditionally independent given the hyperparameters μ and κ.

Example: baseball scores The average number of runs (i.e. points) per game scored by 7 American Baseball League teams in the 1993 season are listed as follows.

i	1	2	3	4	5	6	7
$\bar{y}_i$	5.549	5.228	5.154	5.068	4.877	4.852	4.79
s_i	0.266	0.257	0.254	0.252	0.246	0.245	0.243

What is the probability that team 1 is better than team 2?

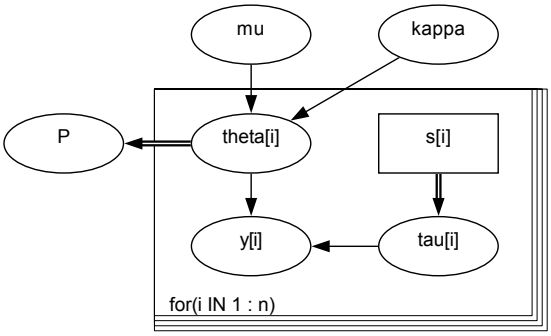
A nonhierarchical model with a-priori independent parameters would give

$$\theta_1 - \theta_2 \,|\, \text{data} \sim \text{Normal}(\bar{y}_1 - \bar{y}_2, s_1^2 + s_2^2) = \text{Normal}(0.321, 0.37)$$

from which we compute $\mathsf{P}(\theta_1 > \theta_2 \,|\, \text{data}) = 0.807$.

A hierarchical model is

```
model {
  for( i in 1 : n ) {
    theta[i] ~ dnorm(mu,kappa)
    y[i] ~ dnorm(theta[i],tau[i])
    tau[i] <- 1 / (s[i] * s[i])
  }
  mu ~ dnorm(0,0.001)
  kappa ~ dgamma( 0.1, 0.1)
  P <- step(theta[1] - theta[2])
}
```



The data are entered as

```
list(n=7,y=c(5.549,5.228,5.154,5.068,4.877,4.852,4.79),
     s=c(0.266,0.257,0.254,0.252,0.246,0.245,0.243))
```

The results after 5000 simulation steps are

node	mean	sd	2.5%	median	97.5%
theta[1]	5.349	0.2257	4.936	5.341	5.829
theta[2]	5.16	0.2121	4.746	5.158	5.585
theta[3]	5.119	0.2031	4.735	5.113	5.524
theta[4]	5.064	0.2024	4.662	5.066	5.462
theta[5]	4.95	0.2027	4.541	4.955	5.34
theta[6]	4.933	0.2033	4.526	4.93	5.325
theta[7]	4.893	0.2042	4.479	4.897	5.278
P	0.7376				

Notice that the posterior means are more closely grouped together than the observations: $\mathsf{E}(\theta_1 \,|\, \text{data}) = 5.349$ is lower than $\bar{y}_1 = 5.549$ and $\mathsf{E}(\theta_7 \,|\, \text{data}) = 4.893$ is higher than $\bar{y}_1 = 4.79$. The posterior probability that $\theta_1 > \theta_2$ is 0.7376, less than what was obtained with the non-hierarchical model.

11.3 Linear regression

The linear regression model

$$y_i = c_1 + c_2 x_i + \varepsilon_i, \quad \varepsilon_i \,|\, \sigma^2 \sim \text{Normal}(0, \sigma^2), \quad i \in \{1, \ldots, n\}$$

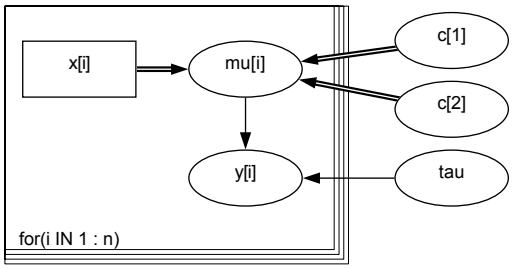
can also be written as

$$y_i \,|\, c_1, c_2, \sigma^2 \sim \text{Normal}(c_1 + c_2 x_i, \sigma^2)$$

There are three unknown parameters: the noise variance σ^2 and the regression coefficients c_1, c_2. Just like for the two-parameter normal model, a closed-form solution is possible when conjugate priors are chosen. We won't present this solution here; instead, we go directly to WinBUGS for the solution.

Example: grades We are interested in seeing how well the number of points x_i that a student earns by doing weekly homework problems will predict the student's exam grade y_i. Here is a WinBUGS model for fitting a linear regression model for data from a Bayesian statistics course.

```
model {
  for( i in 1 : n ) {
    mu[i] <- c[1] + c[2] * x[i]
    y[i] ~ dnorm(mu[i],tau)
  }
  c[1] ~ dnorm(5,0.1)
  c[2] ~ dnorm(2,0.1)
  tau ~ dgamma( 0.1, 0.1)
}
```



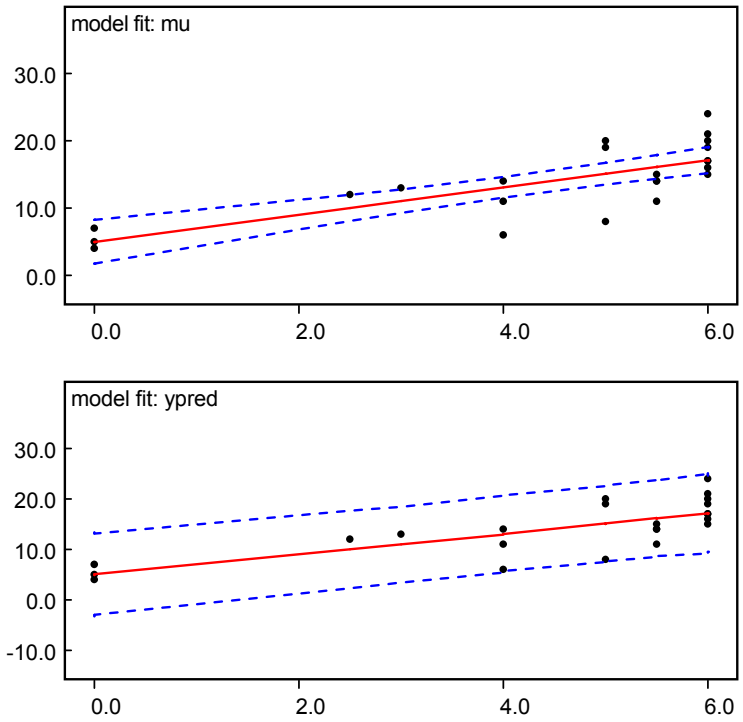
The data are entered as

```
list(n=23,x=c( 6,5,6,5.5,5.5,6,6,6,0,6,3,5,
               4,6,4,5,5.5,5.5,2.5,4,0,6,0),
          y=c(15,20,16,15,14,21,24,17,7,17,13,19,
              14,19,11,8,11,14,12,6,5,20,4) )
```

The results after 5000 simulation steps are

node	mean	sd	2.5%	median	97.5%
c[1]	4.966	1.635	1.764	4.96	8.257
c[2]	2.027	0.3437	1.355	2.026	2.711

Here are plots of the data and of the means and 95% confidence intervals of $c_1 + c_2 x \,|\, y$ and of the predictive posterior $\tilde{y} \,|\, y$.



To produce the above plots, make sure `mu` and `ypred` are selected as nodes in the Sample Monitor Tool; this tool's window is opened by selecting menu item Inference/Samples. After running the simulation, open the Comparison Tool window (menu item Inference/Compare) and enter `mu` or `ypred` in the *node* space, `y` in the *other* space, and `x` in the *axis* space. Finally, press the `model fit` button.

11.4 Autoregressive model of time series

The 1-state autoregressive model of a time series is

$$y_i - m = a \cdot (y_{i-1} - m) + e_i \quad (i = 1, 2, \ldots, N)$$

where the $e_i \sim \text{Normal}(0, \sigma^2)$ are independent. The parameter $a \in (-1, 1)$ specifies how strongly the consecutive mean-centred observations $y_i - m$ are correlated. The other unknown parameters in the model are m (the mean), y_0 (the initial state), and σ^2 (the noise variance). The AR(1) model can also be written

$$y_i \,|\, a, m, y_0, \sigma^2 \sim \text{Normal}(\underbrace{m + a \cdot (y_{i-1} - m)}_{\mu_i}, \sigma^2)$$

Here μ_i represents the model's "one-step" forecast of y_i, given y_0, a, m, and the past observations $y_1, \ldots, y_{i-1}$.

Example: earthquakes Here is a WinBUGS model to fit an AR(1) model to the time series of the number of earthquakes of intensity ≥ 7 Richter in the years 1900–1998.

```
model {
  m ~ dnorm(0,0.01)
  y0 ~ dnorm(m,0.01)
  mu[1] <- a*(y0-m) + m
  y[1] ~ dnorm(mu[1],tau)
  t[1] <- 1900
```

```
 for (i in 2:N) {
    mu[i] <- a*(y[i-1]-m) + m
    y[i] ~ dnorm(mu[i],tau)
    t[i] <- 1899 + i
    ypred[i] ~ dnorm(mu[i],tau)
  }
  tau ~ dgamma(0.01,0.01)
  a ~ dnorm(0.5,5)
  sigma2 <- 1/tau
}
```

The data is

```
list(y=c(13,14,8,10,16,26,32,27,18,32,36,24,22,23,22,18,25,21,21,14,
8,11,14,23,18,17,19,20,22,19,13,26,13,14,22,24,21,22,26,21,
23,24,27,41,31,27,35,26,28,36,39,21,17,22,17,19,15,34,10,15,
22,18,15,20,15,22,19,16,30,27,29,23,20,16,21,21,25,16,18,15,
18,14,10,15,8,15, 6,11, 8, 7,13,10,23,16,15,25,22,20,16),N=99)
```

Because the priors are so vague, WinBUGS does not generate suitable starting values for the Monte Carlo simulation, so these must be provided, as follows. Add a line

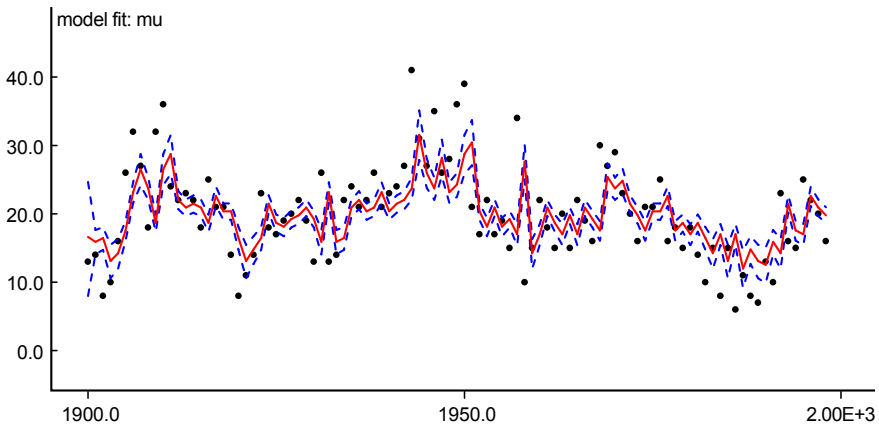
```
list(a=0.5,y0=13,m=20,tau=0.03)
```

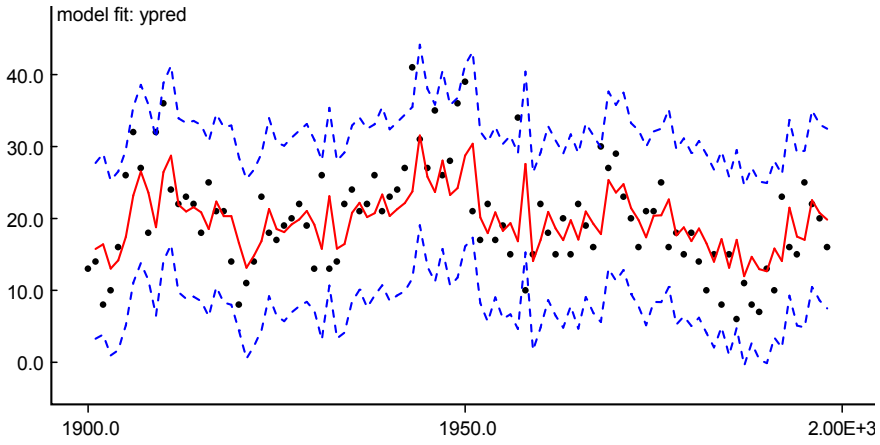
to the model file, double-click the word `list`, and press `load inits` after you've compiled the model. Then press `gen inits` to generate initial values for the predictive posteriors `ypred`.

Results after 5000 simulation steps are

node	mean	sd	2.5%	median	97.5%
a	0.5583	0.08629	0.3933	0.5557	0.7274
m	19.46	1.509	16.23	19.53	22.21
y0	14.39	7.481	-0.2498	14.39	28.91
sigma2	38.34	5.464	28.88	37.85	50.23

The one-step forecasts $\mu_i \,|\, y_{1:n}$ and predictive posteriors $\tilde{y}_i \,|\, y_{1:n}$ can be plotted using the Comparison Tool.





11.5 Logistic regression

Example: rats In a series of experiments, n_i lab rats are each given an injection of a substance at concentration X_i (in g/ml); shortly afterwards, y_i rats die. Letting θ_i be the mortality rate for $x_i = \log(X_i)$, the number of deaths can be modelled as

$$y_i \,|\, \theta_i \sim \text{Binomial}(n_i, \theta_i).$$

x_i	n_i	y_i
-0.863	5	0
-0.296	5	1
-0.053	5	3
0.727	5	5

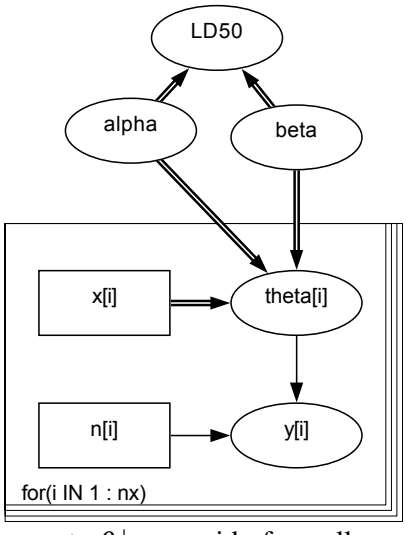
The relation between mortality rate and log-dosage is modelled as

$$\underbrace{\text{logit}(\theta_i)}_{\log \frac{\theta_i}{1-\theta_i}} = \alpha + \beta x_i$$

In this kind of study, a parameter of interest is $x_{\text{LD50}} = -\alpha/\beta$, the log-dosage corresponding to 50% mortality rate, that is, $\text{logit}(0.5) = \alpha + \beta x_{\text{LD50}}$.

Here's a WinBUGS model.

```
model {
  for (i in 1:nx) {
    logit(theta[i]) <- alpha + beta*x[i]
    y[i] ~ dbin(theta[i],n[i])
  }
  alpha ~ dnorm(0.0,0.001)
  beta ~ dnorm(0.0,0.001)
  LD50 <- -alpha/beta
  for ( i in 1:21 ) {
    xx[i] <- -1+2*(i-1)/20
    logit(tt[i]) <- alpha + beta*xx[i]
  }
}
```



The last few lines of the code (not shown in the DAG) compute $\theta \,|\, y$ on grid of equally spaced x values, in order to produce a smooth plot.

The data are entered as

```
list(y=c(0,1,3,5), n=c(5,5,5,5),
  x=c(-0.863,-0.296,-0.053,0.727), nx=4)
```

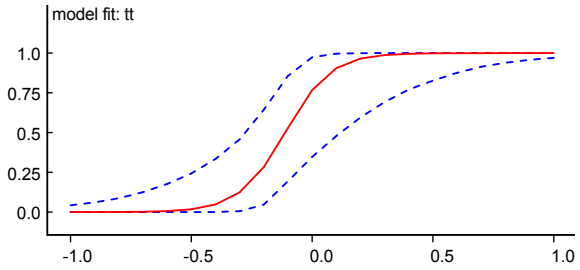
and the simulation initial values are set as

```
list(alpha=0,beta=1)
```

The results after 5000 simulation steps are

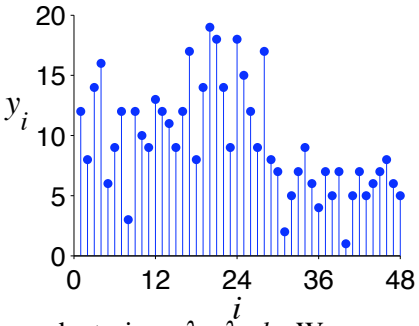
node	mean	sd	2.5%	median	97.5%
alpha	1.274	1.067	-0.6333	1.193	3.607
beta	11.38	5.56	3.463	10.44	24.78
LD50	-0.1052	0.09512	-0.2738	-0.1109	0.1196

The Comparison Tool is used to plot `tt` vs. `xx`. The plot shows the mortality rate θ as a function of the log-dosage x, with 95% credibility intervals.



11.6 Change point detection

A production unit produces items at a rate of λ_1 units per hour. At time k a component is replaced and the production changes to λ_2 units per hour. Given a sequence of n hourly production counts, we wish to determine the production rates and the change point k.



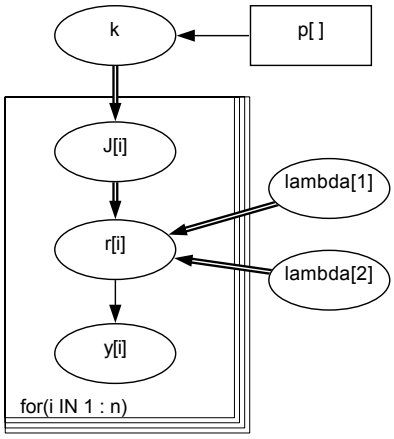
We model the production counts as $y_i \,|\, \lambda_1, \lambda_2, k \sim$ Poisson(r_i) with

$$r_i = \begin{cases} \lambda_1 & i \in \{1, 2, \ldots, k\} \\ \lambda_2 & i \in \{k+1, \ldots, n\} \end{cases}$$

and assume the counts to be conditionally mutually independent given λ_1, λ_2, k. We assume prior distributions $\lambda_1 \sim \text{Gamma}(\alpha_1, \beta_1)$, $\lambda_2 \sim \text{Gamma}(\alpha_2, \beta_2)$, and a uniform prior pmf for k.

Here's a WinBUGS model.

```
model {
  for( i in 1 : n ) {
    J[i] <- 1 + step(i - k - 0.5)
    r[i] <- lambda[J[i]]
    y[i] ~ dpois(r[i])
    p[i] <- 1/n
  }
  k ~ dcat(p[ ])
  lambda[1] ~ dgamma(alpha[1],beta[1])
  lambda[2] ~ dgamma(alpha[2],beta[2])
}
```



The data are entered as

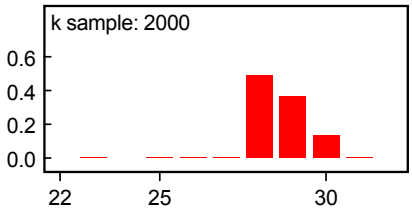
```
list(y=c(12,8,14,16,6,9,12,3,12,10,9,13,12,11,9,12,17,8,14,19,18,
   14,9,18,15,12,9,17,8,7,2,5,7,9,6,4,7,5,7,1,5,7,5,6,7,8,6,5), n=48,
   alpha=c(0.1,0.1),beta=c(0.1,0.1))
```

The simulation initial values are set as

```
list(k=24,lambda=c(10,10))
```

The results after 2000 simulation steps are

node	mean	sd	2.5%	median	97.5%
`lambda[1]`	11.94	0.6642	10.67	11.92	13.29
`lambda[2]`	5.78	0.5549	4.722	5.761	6.915



12 MCMC

As we have seen, a *formula* for the posterior density, up to a scaling factor, is relatively easy to obtain via Bayes's rule

$$p(\theta \,|\, y) \propto p(\theta) p(y \,|\, \theta).$$

With the formula one can compute the value of the unscaled density at any point in the parameter space. To make useful inferences, however, one needs to do things like find summary statistics (mean, median, variance, credibility regions) and compute hypothesis probabilities. For simple models with conjugate priors these results can be obtained using algebraic manipulations with standard statistical functions, as we did in §5. If such closed-form solutions are not possible, then in one-dimensional or two-dimensional parameter spaces the unscaled posterior density can be plotted and standard numerical quadrature algorithms can be used to compute means, variances, and other expectation integrals. For high-dimensional parameter spaces, however, such computations are challenging and require specialised algorithms.

In this section, we look at algorithms that produce sets of *random samples* from the posterior distribution. From a set $\{\theta^1, \ldots, \theta^N\}$ of such samples, it is straightforward to compute expectations using the (frequentist!) estimator

$$\mathsf{E}(h(\theta) \,|\, y) = \int h(\theta) p(\theta \,|\, y) \, \mathrm{d}\theta \approx \frac{1}{N} \sum_{t=1}^{N} h(\theta^t). \tag{18}$$

Other summary statistics, such as the median and credibility intervals, can similarly be estimated directly from the samples.

Numerical algorithms that use random samples are known as Monte Carlo methods. Monte Carlo methods that generate samples using a Markov chain are called MCMC methods. In these notes we focus on the MCMC method called the Gibbs sampler, the main method used by WinBUGS (Windows program for Bayesian analysis Using Gibbs Sampling).

12.1 Markov chains

Consider a sequence θ^t, $t \in \{0, 1, 2, \ldots\}$ of random variables that can take on a finite number of values, say $\theta^t \in \{1, \ldots, p\}$. The sequence is a (homogenous) *Markov chain* if, for all $t \geq 1$, the pmf of the tth state conditional on past states depends only on the previous state and not on the index t or on older states. In other words, there exists a function $T(\cdot, \cdot)$ (the *transition probability*) such that

$$\mathsf{P}(\theta^t = x_t \,|\, \theta^0 = x_0, \theta^1 = x_1, \ldots, \theta^{t-1} = x_{t-1}) = \mathsf{P}(\theta^t = x_t \,|\, \theta^{t-1} = x_{t-1}) = T(x_t, x_{t-1}).$$

The transitional probability and the initial distribution $\pi_0(x) := \mathsf{P}(\theta^0 = x)$ suffice to define the joint distribution of all the variables in a Markov chain:

$$\begin{aligned}
\mathsf{P}(\theta^0 = x_0,\, \theta^1 = x_1) &= \mathsf{P}(\theta^1 = x_1 \,|\, \theta^0 = x_0)\mathsf{P}(\theta^0 = x_0) \\
&= T(x_1, x_0)\pi_0(x_0) \\
\mathsf{P}(\theta^0 = x_0,\, \theta^1 = x_1,\, \theta^2 = x_2) &= \mathsf{P}(\theta^2 = x_2 \,|\, \theta^0 = x_0,\, \theta^1 = x_1)\mathsf{P}(\theta^0 = x_0,\, \theta^1 = x_1) \\
&= T(x_2, x_1)T(x_1, x_0)\pi_0(x_0) \\
&\vdots \qquad \vdots \\
\mathsf{P}(\theta^0 = x_0, \ldots, \theta^t = x_t) &= T(x_t, x_{t-1}) \cdots T(x_1, x_0)\pi_0(x_0)
\end{aligned}$$

A probability distribution π is said to be a *stationary* (or invariant or equilibrium) distribution of a Markov chain if it satisfies the equation

$$\pi(x') = \sum_{x=1}^{p} T(x', x)\pi(x). \tag{19}$$

The name comes from the fact that if θ^{t-1} has the distribution π, then

$$\mathsf{P}(\theta^t = x') = \sum_{x=1}^{p} \mathsf{P}(\theta^t = x', \theta^{t-1} = x) = \sum_{x=1}^{p} T(x', x) \underbrace{\mathsf{P}(\theta^{t-1} = x)}_{=\pi(x)} = \pi(x'),$$

that is, θ^t has the same distribution, and thus, so does every subsequent state $\theta^{t+1}, \theta^{t+2}, \ldots$.

Under mild conditions, it can be shown that the stationary distribution is unique and that, for any initial π_0, the sequence of marginal distributions $\mathsf{P}(\theta^t = x)$ converges to $\pi(x)$ as $t \to \infty$.

Example: a one-dimensional random walk Consider a 6-valued Markov chain whose transition probability is described by the matrix

$$[T(\cdot,\cdot)] = \frac{1}{2}\begin{bmatrix}
1 & 1 & & & & & \\
1 & 0 & 1 & & & & \\
 & 1 & 0 & 1 & & & \\
 & & & 1 & 0 & 1 & \\
 & & & & 1 & 0 & 1 \\
 & & & & & 1 & 1
\end{bmatrix}.$$

This chain can be described in terms of a random walk:

- If the state at some time is $\theta \in \{2, 3, 4, 5\}$, then the subsequent state is $\theta + 1$ or $\theta - 1$ with equal probability.
- If the state is $\theta = 1$ then the subsequent state is 1 or 2 with equal probability.
- If the state is $\theta = 6$ then the subsequent state is 5 or 6 with equal probability.

Suppose the initial state is $\theta^0 = 3$, so that the initial pmf is

$$p(\theta^0) = [0,\, 0,\, 1,\, 0,\, 0,\, 0]$$

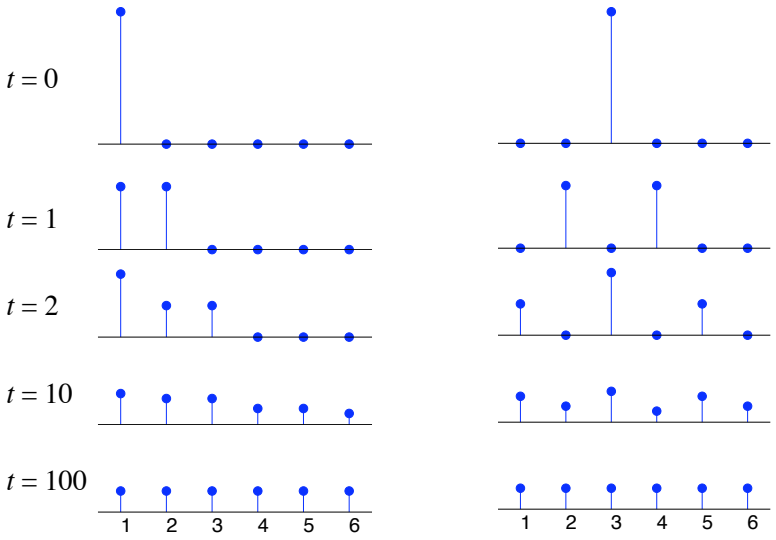
Then the pmf's of the subsequent states are

$$\begin{aligned}
p(\theta^1) &= [0, \tfrac{1}{2}, 0, \tfrac{1}{2}, 0, 0] \\
p(\theta^2) &= [\tfrac{1}{4}, 0, \tfrac{1}{2}, 0, \tfrac{1}{4}, 0]
\end{aligned}$$

and so on. If the initial state is $\theta^0 = 1$, then the sequence of pmf's of the subsequent states are

$$\begin{aligned} p(\theta^0) &= [1,0,0,0,0,0] \\ p(\theta^1) &= [\tfrac{1}{2},\tfrac{1}{2},0,0,0,0] \\ p(\theta^2) &= [\tfrac{1}{2},\tfrac{1}{4},\tfrac{1}{4},0,0,0], \end{aligned}$$

and so on. These sequences are illustrated below:



We see that both sequences converge towards the pmf $\pi = [\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6}]$, which is the stationary distribution of this chain, as can be verified by substitution into (19).

An MCMC algorithm to sample from the equilibrium distribution π consists of simulating a random walk: starting from some arbitrary state, choose the next state at random according to the transition probability. Continue to move in this way from state to state for a large number of steps. After a sufficient number of "warmup" simulation steps (which are usually discarded), the sequence of states $\{\theta^1,\ldots,\theta^N\}$ of this random walk can be considered as samples from (approximately) the stationary distribution π. One can then compute expectations for the distribution π using sums:

$$\mathsf{E}(h(\theta)) = \int h(\theta)\pi(\theta)\,\mathrm{d}\theta \approx \frac{1}{N}\sum_{t=1}^{N} h(\theta^t).$$

Here is pseudocode for an MCMC algorithm that finds N samples from the equilibrium distribution of the Markov chain introduced at the beginning of this example. The first t_0 states in the random walk are discarded.

initialise θ to some arbitrary value in $\{1,\ldots,6\}$, say $\theta \leftarrow 3$
for i from 1 to $t_0 + N$ do
    with equal probability, either:
        if $\theta > 1$, decrement θ by 1
    or
        if $\theta < 6$, increment θ by 1
    end choice
    if $i > t_0$, $\theta^{i-t_0} \leftarrow \theta$
end do

12.2 Gibbs sampler

The Gibbs sampler is used to produce samples from a posterior distribution $p(\theta \,|\, y)$ with multidimensional parameter vector θ. The samples are produced by a random walk in a Markov chain that has stationary distribution $p(\theta \,|\, y)$. In each step of this algorithm, only one component of θ^t is changed: it is replaced by a draw from the one-dimensional distribution that is obtained when all the other components are kept fixed.

In the following pseudocode of the algorithm, the notation θ_{-i} denotes the vector θ with the ith component removed, that is, $\theta_{-i} = [\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_d]$.

$\theta^0 \leftarrow$ some vector in the parameter space
for t from 1 to N do
    choose a dimension $i_t \in \{1, \ldots, d\}$ at random (with pmf $[r_1, \ldots, r_d]$, say)
    $\theta^t_{i_t} \leftarrow$ a sample drawn from $p(\theta_{i_t} \,|\, \theta^{t-1}_{-i_t}, y)$
    $\theta^t_{-i_t} \leftarrow \theta^{t-1}_{-i_t}$
end do

Often, the algorithm is implemented with the updates performed by cycling through the indices i, instead of choosing indices in random order. However, the proof that $p(\theta \,|\, y)$ is a stationary distribution of the Gibbs sampler's Markov chain is simpler when the indices are chosen in random order. Here's the proof.

Consider what happens when θ^{t-1} is drawn from the distribution $p(\theta \,|\, y)$. The probability of transition from θ to θ' via an update of the ith component is

$$\begin{aligned} \mathsf{P}(\theta^{t-1} = \theta, \theta^t = \theta', i_t = i \,|\, y) &= \mathsf{P}(\theta^t = \theta' \,|\, \theta^{t-1} = \theta, i_t = i, y)\mathsf{P}(\theta^{t-1} = \theta, i_t = i \,|\, y) \\ &= \begin{cases} r_i p(\theta \,|\, y) p(\theta'_i \,|\, \theta_{-i}, y) & \text{if } \theta_{-i} = \theta'_{-i} \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Thus, the probability of transition from θ to θ' is

$$\mathsf{P}(\theta^{t-1} = \theta, \theta^t = \theta' \,|\, y) = \sum_{i=1}^{d} r_i p(\theta \,|\, y) p(\theta'_i \,|\, \theta_{-i}, y) \chi(\theta_{-i} = \theta'_{-i}),$$

where $\chi(\text{FALSE}) = 0$ and $\chi(\text{TRUE}) = 1$. Similarly, the probability of transition from θ' to θ is

$$\mathsf{P}(\theta^{t-1} = \theta', \theta^t = \theta \,|\, y) = \sum_{i=1}^{d} r_i p(\theta' \,|\, y) p(\theta_i \,|\, \theta'_{-i}, y) \chi(\theta_{-i} = \theta'_{-i})$$

These transition probabilities are equal, because

$$p(\theta'_i \,|\, \theta_{-i}, y) = p(\theta'_i \,|\, \theta'_{-i}, y) = \frac{p(\theta' \,|\, y)}{p(\theta'_{-i} \,|\, y)}$$

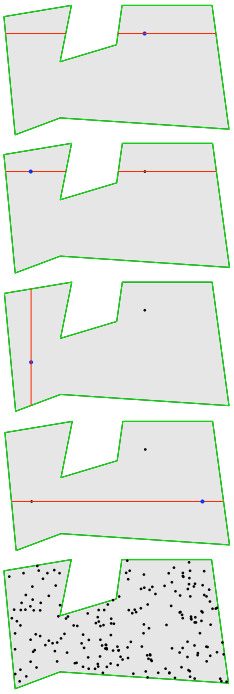
and

$$p(\theta_i \,|\, \theta'_{-i}, y) = p(\theta_i \,|\, \theta_{-i}, y) = \frac{p(\theta \,|\, y)}{p(\theta_{-i} \,|\, y)} = \frac{p(\theta \,|\, y)}{p(\theta'_{-i} \,|\, y)}$$

when $\theta_{-i} = \theta'_{-i}$. Because their joint distribution is symmetric, $\theta^{t-1} \,|\, y$ and $\theta^t \,|\, y$ have the same marginal distributions. It follows that the distribution of $\theta \,|\, y$ is a stationary distribution of this Markov chain.

Example: Uniformly distributed points inside a polygon Here's a Gibbs algorithm to sample from this distribution:

1. Choose a point inside the polygon and draw a horizontal line through it.
2. Choose a new point uniformly at random along the portion of the horizontal line that lies inside the polygon.
3. Draw a vertical line through the new point and choose a new point uniformly at random along the portion of the line that lies inside the polygon.
4. Draw a horizontal line through the new point and choose a new point uniformly at random along the portion of the line that lies inside the polygon.
5. Repeat steps 3 and 4 as many times as desired. Retain every second point.



Example: Two-parameter normal model We saw in §5.1 that the posterior for a normal model with known variance σ^2 and unknown mean μ with flat prior is

$$\mu \,|\, y, \sigma^2 \sim \text{Normal}(\bar{y}, \frac{\sigma^2}{n})$$

In §5.5 we saw that the posterior for known mean μ and unknown variance σ^2 with prior $p(\sigma^2) \propto \sigma^{-2}$ is

$$\sigma^2 \,|\, y, \mu \sim \text{InvGam}(\tfrac{n}{2}, \tfrac{1}{2}\textstyle\sum_{i=1}^{n}(y_i - \mu)^2)$$

Thus, a (cyclic-order) Gibbs sampler to generate samples from the normal model with unknown mean and variance and prior $p(\mu, \sigma^2) \propto \sigma^{-2}$ is

initialise μ_0 and σ_0^2

for t from 1 to N do

$\mu_t \leftarrow$ a sample drawn from $\text{Normal}(\bar{y}, \frac{\sigma_{t-1}^2}{n})$

$\sigma_t^2 \leftarrow$ a sample drawn from $\text{InvGam}(\frac{n}{2}, \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu_t)^2)$

end do

Here's a Matlab script that uses the above algorithm to generate 200 samples for a two-parameter normal model of Cavendish's data.

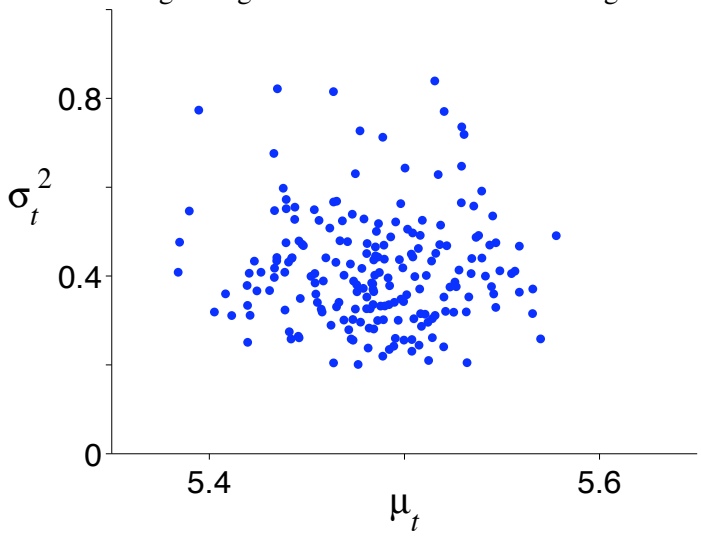
```
% data
y = 5+[36 29 58 65 57 53 62 29 44 34 79 10 ...
       27 39 42 47 63 34 46 30 78 68 85]/100;
n = length(y);  ybar = mean(y);

% number of MCMC samples to generate
N = 200;
```

```
% initial values of mu and sigma2
mu = 0; sigma2 = 1;

% set the random number generator seeds (for repeatability)
randn('state',0);  rand('state',0);

% allocate memory for the MCMC samples
mu = repmat(mu,1,N);
sigma2 = repmat(sigma2,1,N);

% simulation loop
for t = 2:N
   mu(t) = normrnd(ybar,sqrt(sigma2(t-1)/n));
   tau = gamrnd(n/2,2/sum((y-mu(t)).^2));
   sigma2(t) = 1/tau;
end

% look at the results
t0=1;
ii = t0+1:N;       % discard t0 samples
stats = [mean(mu(ii)), var(mu(ii)), mean(sigma2(ii)), var(sigma2(ii))];
plot(mu,sigma2,'.')
```

The sample means and variances provide the following estimates of the summary statistics of the posterior marginals:

$$\mathsf{E}(\mu\,|\,y) \approx 5.485,\ \mathsf{V}(\mu\,|\,y) \approx 1.55\cdot 10^{-3},\ \mathsf{E}(\sigma^2\,|\,y) \approx 0.0415,\ \mathsf{V}(\sigma^2\,|\,y) \approx 1.58\cdot 10^{-4}$$

These values are in good agreement with the exact values given in §9.1.



Example: Change point detection The posterior density for the change-point detection example presented in §11.5 is

$$\begin{aligned} p(\lambda_1,\lambda_2,k\,|\,y) &\propto \prod_{i=1}^{n} p(\lambda_1)p(\lambda_2)p(k)p(y_i\,|\,\lambda_1,\lambda_2,k) \\ &\propto \lambda_1^{\alpha_1-1}\mathrm{e}^{-\beta_1\lambda_1}\cdot\lambda_2^{\alpha_2-1}\mathrm{e}^{-\beta_2\lambda_2}\cdot 1\cdot\lambda_1^{s_k}\mathrm{e}^{-k\lambda_1}\cdot\lambda_2^{s_n-s_k}\mathrm{e}^{-(n-k)\lambda_2} \end{aligned}$$

where $s_k = \sum_{i=1}^{k} y_i$. The update probability densities for the rates λ_1 and λ_2 are therefore

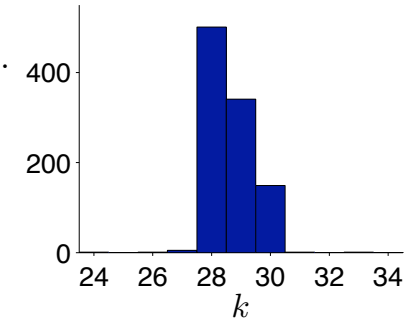
$$\begin{aligned} p(\lambda_1 \,|\, \lambda_2, k, y) &\propto \lambda_1^{\alpha_1 - 1 + s_k} \mathrm{e}^{-(\beta_1 + k)\lambda_1} \\ p(\lambda_2 \,|\, \lambda_1, k, y) &\propto \lambda_2^{\alpha_2 - 1 + s_n - s_k} \mathrm{e}^{-(\beta_1 + n - k)\lambda_2} \end{aligned}$$

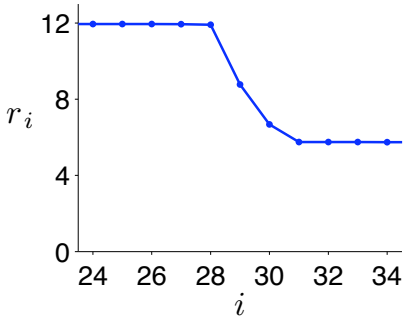
that is, $\lambda_1 \,|\, \lambda_2, k, y \sim \text{Gamma}(\alpha_1 + s_k, \beta_1 + k)$ and $\lambda_2 \,|\, \lambda_1, k, y \sim \text{Gamma}(\alpha_2 + s_n - s_k, \beta_2 + n - k)$. The update distribution for the change point k has the pmf

$$\mathsf{P}(k^t = k \,|\, \lambda_1 = x_1, \lambda_2 = x_2, y = y_{1:n}) \propto (x_1/x_2)^{s_k} \mathrm{e}^{-k(x_1 - x_2)}$$

Here's a Matlab script that generates 1000 samples from the posterior distribution using a Gibbs sampler, and plots the histogram of change-point values k.

```
% Data
y=[12,8,14,16,6,9,12,3,12,10,9,13,12,11,9,12,17,8,14,19,18,14,...
    9,18,15,12,9,17,8,7,2,5,7,9,6,4,7,5,7,1,5,7,5,6,7,8,6,5];
n=length(y);
alpha=[0.1,0.1];  beta=[0.1,0.1];   % parameters of lambda prior
k=24; lambda=[10;10];               % initial values of stoch. variables
rand('state',0); randn('state',0);  % random generator seeds
N=1000;                             % number of update steps
s=cumsum(y);                        % vector of cumulative sums

% allocate memory for the samples
k=repmat(k,1,N); lambda=repmat(lambda,1,N); r=zeros(n,N);

% simulation loop
for t=2:N
    % draw lambda(1) and lambda(2) from gamma pdf's
    A=[alpha(1)+s(k(t-1)); alpha(2)+s(n)-s(k(t-1))];
    B=1./[beta(1)+k(t-1);beta(2)+n-k(t-1)];
    lambda(:,t)=gamrnd(A,B);
    % construct the pmf for k
    % using logs to avoid overflow/underflow
    kk=1:n-1;
    logp = s(kk)*log(lambda(1,t)/lambda(2,t)) ...
              - kk*(lambda(1,t)-lambda(2,t));
    p=exp(logp-max(logp));
    p=p/norm(p,1);
    % draw k ~ categorical(p)
    [notused,k(t)]=histc(rand,[0 cumsum(p)]);
    % rate parameters of y(i) ~ Pois(r(i))
    r(1:k(t),t)=lambda(1,t);
    r(k(t)+1:n,t)=lambda(2,t);
end

% plot a histogram of the k values
figure(1), hist(k,24:34)

% plot the mean rate
figure(2), plot(24:34,mean(r(24:34,:),2),'.-')
```





13 Model comparison

The Bayesian data analysis technique requires that you model the statistical data analysis problem by specifying a prior distribution and a likelihood distribution. However, any model is an approximation:

> *Essentially, all models are wrong, but some are useful.* G. E. P. Box

Thus, any statistical data analysis includes consideration of model adequacy, model sensitivity, and alternative models. The development of practical tools for these tasks is a very active research area in Bayesian statistics. In this section we only scratch the surface of this aspect of modelling, and present two widely-used model comparison approaches: Bayes factors and DIC.

13.1 Bayes factors

Suppose you have two alternative models,

$$\begin{array}{lll} \text{model } \mathscr{M}_1\text{:} & \text{likelihood } p_1(y\,|\,\theta_1), & \text{prior } p_1(\theta_1) \\ \text{model } \mathscr{M}_2\text{:} & \text{likelihood } p_2(y\,|\,\theta_2), & \text{prior } p_2(\theta_2) \end{array}$$

where the parameter vectors θ_1 and θ_2 may have different numbers of components. A fully Bayesian approach to coping with your uncertainty about which model produced the data is to construct a single all-encompassing "supermodel" with a model index $m \in \{1,2\}$ as one of the parameters to be inferred from the data. Let $\pi_1 = \mathsf{P}(\text{model} = \mathscr{M}_1)$ and $\pi_2 = \mathsf{P}(\text{model} = \mathscr{M}_2)$ denote your prior probabilities (degrees of belief) in the models, with $\pi_1 + \pi_2 = 1$. Then Bayes's rule gives

$$\mathsf{P}(\text{model} = \mathscr{M}_m\,|\,y) \propto \mathsf{P}(\text{data} = y\,|\,\text{model} = \mathscr{M}_m)\pi_m$$

where

$$\mathsf{P}(\text{data} = y\,|\,\text{model} = \mathscr{M}_m) = \int p_m(y\,|\,\theta_m)p_m(\theta_m)\,\mathrm{d}\theta_m$$

is called the *evidence* for model $\mathscr{M}_m$. Note that this is the normalizing constant in Bayes's formula when doing inference about θ_m using the mth model alone.

The posterior odds in favour of model $\mathscr{M}_1$ against model $\mathscr{M}_2$ given the data y are given by the ratio

$$\frac{\mathsf{P}(\text{model} = \mathscr{M}_1\,|\,y)}{\mathsf{P}(\text{model} = \mathscr{M}_2\,|\,y)} = \frac{\pi_1}{\pi_2} \times \underbrace{\frac{\int p_1(y\,|\,\theta_1)p_1(\theta_1)\,\mathrm{d}\theta_1}{\int p_2(y\,|\,\theta_2)p_2(\theta_2)\,\mathrm{d}\theta_2}}_{B_{12}}$$

where the number B_{12} (the ratio of evidences) is called the *Bayes factor* for model $\mathscr{M}_1$ against model $\mathscr{M}_2$. Recall that the Bayes factor was discussed earlier in §8 in the context of hypothesis testing.

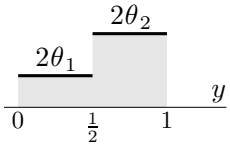
It is straightforward to generalise the above-described technique to compare any finite set of models. Because $B_{ij} = B_{ik}B_{kj}$, models can be ordered consistently based on pairwise comparisons using Bayes factors.

Example: density estimation Consider independent real-valued samples $y_{1:5} = [0.3, 0.6, 0.7, 0.8, 0.9]$ drawn from some pdf. According to model $\mathscr{M}_1$, $y_i \sim \text{Uniform}(0,1)$; this model has no parameters. The evidence for $\mathscr{M}_1$ is

$$\mathsf{P}(\text{data} = y\,|\,\text{model} = \mathscr{M}_1) = p_1(y_1)p_1(y_2)\cdots p_1(y_5) = 1\cdot 1\cdots 1 = 1.$$

According to model $\mathscr{M}_2$, the density is piecewise constant with two pieces,

$$p_2(y_i \,|\, \theta_{1:2}) = \begin{cases} 2\theta_1 & 0 \le y_i < \frac{1}{2} \\ 2\theta_2 & \frac{1}{2} \le y_i < 1 \\ 0 & \text{otherwise} \end{cases}$$



This model has effectively one parameter, because $\theta_1 + \theta_2 = 1$. The prior distribution is taken to be $\theta \sim \text{Dirichlet}(1,1)$, so that $p_2(\theta) = 1$ (uniform on the θ-simplex $0 \le \theta_1, \theta_2 \le 1,\ \theta_1 + \theta_2 = 1$). The posterior is

$$p_2(\theta \,|\, y) \propto p_2(\theta) p_2(y \,|\, \theta) \propto \theta_1 \theta_2^4$$

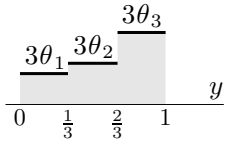
so that $\theta \,|\, y \sim \text{Dirichlet}(2,5)$, for which $\mathsf{E}(\theta_1 \,|\, y) = \frac{2}{7} \approx 0.29$, $\mathsf{E}(\theta_2 \,|\, y) = \frac{5}{7} \approx 0.71$. The evidence for $\mathscr{M}_2$ is

$$\int p_2(\theta) p_2(y \,|\, \theta) \,\mathrm{d}\theta = \int 1 \cdot (2\theta_1)(2\theta_2)^4 \,\mathrm{d}\theta = 32 \frac{\Gamma(2)\Gamma(5)}{\Gamma(7)} = \frac{16}{15}.$$

If the prior beliefs in the two models $\mathscr{M}_2$ and $\mathscr{M}_3$ are equal (i.e. $\pi_1 = \pi_2$), then the Bayes factor is $B_{21} = \frac{16}{15} \approx 1.07$, that is, the odds are only very slightly in favour of the more complex model.

Consider a more complex model $\mathscr{M}_3$, in which the density is piecewise constant with three pieces,

$$p_2(y_i \,|\, \theta_{1:3}) = \begin{cases} 3\theta_1 & 0 \le y_i < \frac{1}{3} \\ 3\theta_2 & \frac{1}{3} \le y_i < \frac{2}{3} \\ 3\theta_3 & \frac{2}{3} \le y_i < 1 \\ 0 & \text{otherwise} \end{cases}$$



This model has effectively two parameters, because $\theta_1 + \theta_2 + \theta_3 = 1$. The prior distribution is taken to be $\theta \sim \text{Dirichlet}(1,1,1)$, so that $p_3(\theta) = 2$ (uniform on the θ-simplex). The posterior is

$$p_3(\theta \,|\, y) \propto p_3(\theta) p_3(y \,|\, \theta) \propto \theta_1 \theta_2 \theta_3^3$$

so that $\theta \,|\, y \sim \text{Dirichlet}(2,2,4)$, for which $\mathsf{E}(\theta_1 \,|\, y) = \frac{2}{8} = 0.25$, $\mathsf{E}(\theta_2 \,|\, y) = \frac{2}{8} = 0.25$, and $\mathsf{E}(\theta_3 \,|\, y) = \frac{4}{8} = 0.5$. The evidence for $\mathscr{M}_3$ is

$$\int p_3(\theta) p_3(y \,|\, \theta) \,\mathrm{d}\theta = \int 2(3\theta_1)(3\theta_2)(3\theta_3)^3 \,\mathrm{d}\theta = 2 \cdot 3^5 \cdot \frac{\Gamma(2)\Gamma(2)\Gamma(4)}{\Gamma(8)} = \frac{81}{140}.$$

If the prior beliefs in the two models $\mathscr{M}_2$ and $\mathscr{M}_3$ are equal (i.e. $\pi_2 = \pi_3$), then the Bayes factor is $B_{23} = \frac{16/15}{81/140} \approx 1.8436$, that is, the odds somewhat favour the simpler model. ■

Often, model comparison indicators are used to guide *model choice*: the best model is retained, and the poorer models are thrown out. Model choice is a decision, and a Bayesian framework for making decisions will be presented in section 14. Note, however, that throwing out alternative models is not a strictly Bayesian approach to inference and prediction. In situations where are M alternative models, a fully Bayesian approach would be to use the full posterior distribution $p(m, \theta_{\mathscr{M}_1}, \ldots \theta_{\mathscr{M}_M} \,|\, y)$. This approach, called *model averaging*, is advanced by its proponents as a "robust" (with respect to model uncertainty) statistical technique.

We have seen in earlier examples that, especially when there is a lot of data, the choice of different vague priors has almost no influence on the results of an inference

analysis, and that one can often use an improper prior. In the model comparison context, however, the situation is different. The evidence decreases when priors are made more vague; improper ("infinitely vague") priors have zero evidence. The Bayes factor thus tends to favour models with less vague priors: ignorance is penalized. This is illustrated in the following example.

Example: evidence for a normal model Suppose $y_i \,|\, \mu, v \sim \text{Normal}(\mu, v)$ (normally conditionally independently distributed given μ and v), with $i \in \{1, 2, \ldots, n\}$, $\mu \sim \text{Uniform}(a, b)$ and v known. The evidence for this model is

$$\begin{aligned} \int p(\mu) p(y_{1:n} \,|\, \mu) \, \mathrm{d}\mu &= \int_a^b \frac{1}{(b-a)(2\pi v)^{n/2}} \mathrm{e}^{-\dfrac{S^2 + n(\bar{y}-\mu)^2}{2v}} \, \mathrm{d}\mu \\ &= \frac{\mathrm{e}^{-\dfrac{S^2}{2v}}}{(b-a)(2\pi v)^{(n-1)/2} n^{1/2}} \int_a^b \frac{\mathrm{e}^{-\dfrac{(\bar{y}-\mu)^2}{2v/n}}}{\sqrt{2\pi v/n}} \, \mathrm{d}\mu, \end{aligned}$$

where $S^2 = \sum_{i=1}^n (y_i - \bar{y})^2$ and $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$. As $b \to \infty$ and $a \to -\infty$, the integral is ≈ 1, and so the evidence for the model $\to 0$. Thus, a vague prior (= large value of $b - a$) tends to be penalised in a Bayesian model comparison, and the evidence for the normal model with the improper prior $p(\mu) \propto 1$ is zero — according to the Bayes factor, any model with proper prior is infinitely better! ■

The following examples illustrate how the Bayes factor penalises models with too many parameters. Because of this property, Bayesian model comparison is a concrete implementation of the precept known as *Occam's razor*, named after the 13th century philosopher William of Ockham, whereby simple models that fit the data should be preferred over more complicated models that fit the data equally well. This precept appears in Machine Learning textbooks as the warning to avoid *overfitting*: overly-complex models generalize poorly (i.e. they have poor fit to out-of-sample data).

Example: density estimation (continued) The more complex model $\mathscr{M}_3$ fits better than $\mathscr{M}_2$, in the sense that the maximum likelihood value is greater. Here are the calculations. Because the prior is flat, the maximum likelihood for $\mathscr{M}_2$ occurs at the posterior mode $\hat{\theta} = \text{mode}(\theta \,|\, y) = (\frac{1}{5}, \frac{4}{5})$, at which the likelihood is

$$p(y_{1:5} \,|\, \hat{\theta}) = 2^5 \cdot (\frac{1}{5})(\frac{4}{5})^4 \approx 2.62.$$

Similarly, the posterior mode of $\mathscr{M}_3$ is $\hat{\theta} = \text{mode}(\theta \,|\, y) = (\frac{1}{5}, \frac{1}{5}, \frac{3}{5})$, at which the likelihood is

$$p(y_{1:5} \,|\, \hat{\theta}) = 3^5 \cdot (\frac{1}{5})(\frac{1}{5})(\frac{3}{5})^3 \approx 4.98.$$

However, as we saw earlier, model $\mathscr{M}_3$ has a somewhat smaller evidence value than model $\mathscr{M}_2$, and the model comparison based on the Bayes factor favours the simpler model. ■

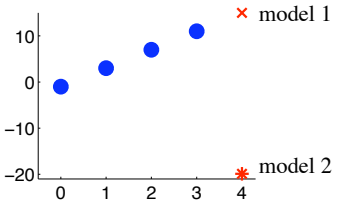
Example[10]**: The next integer in a sequence.** Consider the task of predicting the next integer y_4 in the sequence

$$y_0 = -1,\ y_1 = 3,\ y_2 = 7,\ y_3 = 11$$

Two alternative models are

$\mathcal{M}_1$: the sequence is an arithmetic progression, that is, $y_0 = a_1,\ y_{n+1} = y_n + b$
$\mathcal{M}_2$: the sequence is generated by $y_0 = a_2,\ y_{n+1} = cy_n^3 + dy_n^2 + e$

Fitting the data to the first model gives the parameters $(a_1, b) = (-1, 4)$ and the prediction $y_4 = 15$. The second model is fit by $(a_2, c, d, e) = (-1, \frac{-1}{11}, \frac{9}{11}, \frac{23}{11})$, and the prediction is $y_4 = -19.9$. Both models fit the data equally well (i.e. perfectly). Which model is more plausible?



For the model $\mathcal{M}_1$, because the parameter pair $(a_1, b) = (-1, 4)$ is the only one that fits the data, the likelihood pmf is

$$p_1(y_{0:3} \,|\, a_1, b) = \begin{cases} 1 & \text{if } a_1 = -1, b = 4 \\ 0 & \text{otherwise} \end{cases}$$

If we assume a-priori that a_1 and b can be integers between -50 and 50, then the evidence for model $\mathcal{M}_1$ is

$$\sum_{a_1, b} p_1(y \,|\, a_1, b) p(a_1, b) = \frac{1}{101} \times \frac{1}{101} \approx 0.0001.$$

For the model $\mathcal{M}_2$, the likelihood pmf is $p(y_{0:3} \,|\, a_2, c, d, e) = 1$ if

$$(a_2, c, d, e) \in \{-1\} \times \left\{ \frac{-1}{11}, \frac{-2}{22}, \frac{-3}{33}, \frac{-4}{44} \right\} \times \left\{ \frac{9}{11}, \frac{18}{22}, \frac{27}{33}, \frac{36}{44} \right\} \times \left\{ \frac{23}{11}, \frac{46}{22} \right\},$$

and zero otherwise. We assume that the initial value of the sequence (denoted here as a_2) can be an integer between -50 and 50, and assume that the parameters c, d, e can be rational numbers with numerator between -50 and 50 and denominator between 1 and 50; for simplicity we assume that all (unsimplified) rational numbers in this range are equally likely. Then the evidence for model $\mathcal{M}_2$ is

$$\sum_{a_2, c, d, e} p_2(y \,|\, a_2, c, d, e) p(a_2, c, d, e) = \frac{1 \times 4 \times 4 \times 2}{101 \times (101 \cdot 50)^3} \approx 2.5 \cdot 10^{-12}.$$

The Bayes factor is thus $B_{12} \approx 0.0001 / 2.5 \cdot 10^{-12} = 40 \cdot 10^6$, so that, even if our prior probabilities π_1 and π_2 were equal, the odds in favour of $\mathcal{M}_1$ against $\mathcal{M}_2$ are about forty million to one. ■

The Bayes factor is a sound approach to model comparison, but it is difficult to compute, and many statisticians are not happy with its heavy penalization of vague priors and its total rejection of models with improper priors. The following section presents a popular alternative (albeit heuristic) approach to model comparison.

[10] from *Information Theory, Inference, and Learning Algorithms* by David J. C. MacKay, Cambridge University Press, 2003, full text at `http://www.inference.phy.cam.ac.uk/mackay/itila/`

13.2 Deviance Information Criterion (DIC)

The *deviance* is defined as

$$D(\theta) = -2\log p(y_{1:n} \,|\, \theta).$$

In the case of conditionally independent data with $y_i \,|\, \theta \sim \text{Normal}(\theta_i, v)$, the deviance is

$$\begin{aligned} D(\theta) &= -2\log\left((2\pi v)^{-n/2} \mathrm{e}^{-\frac{1}{2v}\sum_{i=1}^{n}(y_i-\theta)^2}\right) \\ &= n\log(2\pi v) + \frac{1}{v}\sum_{i=1}^{n}(y_i - \mathsf{E}(y_i \,|\, \theta))^2, \end{aligned}$$

that is, the sum of squares of standardized residuals plus a constant. Deviance can thus be considered to be a measure of poorness of fit (i.e. larger values indicate poorer fit).

The posterior mean deviance

$$\bar{D} = \mathsf{E}(D(\theta) \,|\, y) = \int D(\theta) p(\theta \,|\, y)\, \mathrm{d}\theta$$

has been suggested as a criterion for comparing models, but it has generally been judged to insufficiently penalize model complexity. The Deviance Information Criterion (DIC) is a modification proposed by Spiegelhalter et al. in 2002 that adds a term that penalizes complexity. It is defined by the formula

$$\text{DIC} = \bar{D} + \underbrace{\bar{D} - \hat{D}}_{p_D},$$

where $\hat{D} = D(\mathsf{E}(\theta \,|\, y))$. The number p_D is called the 'effective number of parameters', although it is not an integer and in some models does not correspond at all to the number of parameters, and can even be negative. The DIC is easy to compute in MCMC models, and WinBUGS has menu items for DIC computation.

In contrast to the Bayes factor, the absolute size of DIC is not relevant: only differences in DIC are important. However, as the WinBUGS documentation says,

> It is difficult to say what would constitute an important difference in DIC. Very roughly, differences of more than 10 might definitely rule out the model with the higher DIC, differences between 5 and 10 are substantial, but if the difference in DIC is, say, less than 5, and the models make very different inferences, then it could be misleading just to report the model with the lowest DIC.

Example: density estimation (continued) A WinBUGS model for $\mathscr{M}_2$ is

```
model {
   for (i in 1:n) {
      z[i] <- trunc(np*y[i]+1)
      z[i] ~ dcat(theta[ ])
   }
   for (j in 1:np)  { a[j] <- 1 }
   theta[ 1:np ] ~ ddirch(a[1:np])
}
```

The data are entered as

```
list(y=c(0.3,0.6,0.7,0.8,0.9),n=5,np=2)
```

The model for $\mathscr{M}_3$ is the same, but with the data value `np=3`. The DIC values for the two models, based on 5000 simulations each, are

model	$\bar{D}$	$\hat{D}$	p_D	DIC
$\mathscr{M}_2$	5.854	5.203	0.651	6.505
$\mathscr{M}_3$	10.918	9.705	1.214	12.132

The DIC difference is $\text{DIC}_2 - \text{DIC}_3 \approx 5.6$, which indicates that the simpler model is "substantially" better than the more complex model. This is in agreement with the model comparison result found earlier using Bayes factors. ■

Example: Homework and exam grades (revisited) This example from §11.3 presented a regression model of exam grades y_i with homework grades x_i as the explanatory factor. An alternative to the linear regression model is the second-order polynomial model

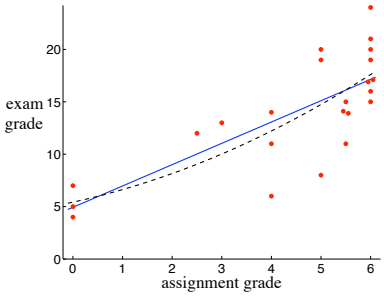
$$y_i = c_1 + c_2 x_i + c_3 x_i^2 + \varepsilon_i, \quad \varepsilon_i \,|\, \sigma^2 \sim \text{Normal}(0, \sigma^2), \quad i \in \{1, \ldots, n\}$$

A WinBUGS model of the alternative model is

```
model {
  for( i in 1 : n ) {
    mu[i] <- c[1] + c[2]*x[i] + c[3]*x[i]*x[i]
    y[i] ~ dnorm(mu[i],tau)
  }
  c[1] ~ dnorm(12,0.01)
  c[2] ~ dnorm(2,0.01)
  c[3] ~ dnorm(0,0.01)
  tau ~ dgamma( 0.1, 0.1)
}
```

The data are the same as in §11.3. After 5000 simulation steps the results are

node	mean	sd	2.5%	median	97.5%
`c[1]`	5.426	1.734	2.035	5.427	8.831
`c[2]`	1.032	1.278	-1.419	1.04	3.568
`c[3]`	0.1662	0.2065	-0.2388	0.165	0.5774
`tau`	0.07803	0.02422	0.03746	0.07619	0.1312



The dashed curve shows $\mathsf{E}(c_1 \,|\, y) + \mathsf{E}(c_2 \,|\, y)x + \mathsf{E}(c_3 \,|\, y)x^2$; the solid line is the linear regression model computed earlier. The DIC values for the two models are

$\mu_i =$	$\bar{D}$	$\hat{D}$	p_D	DIC
$c_1 + c_2 x_i$	125.071	122.284	2.787	127.858
$c_1 + c_2 x_i + c_3 x_i^2$	125.176	121.516	3.660	128.836

Observe that the p_D values are roughly equal to the numbers of model parameters (3 and 4, respectively), and that the linear regression model is "better" than the quadratic regression model, albeit only by a small margin ($\text{DIC}_2 - \text{DIC}_1 \approx 1$).

14 Decision Theory

14.1 The Bayesian choice

The basic idea of decision theory is simple: you have to choose an action a from a set of alternatives $\mathscr{A}$. You have a real-valued *loss function* $L(a,\theta)$ that specifies the cost incurred; it is a function of your choice a and of the "state of nature" θ, a quantity or quantities about which you have some uncertainly. The best choice is the action that minimizes the expected loss

$$\mathsf{E}(L(a,\theta)).$$

Equivalently, you could define a *utility* (gain, reward, preference) function $U = -L$ and choose the action that maximises the expected utility $\mathsf{E}(U(a,\theta))$.

When you receive new information y, the best choice is then the action that minimizes the posterior mean loss $\mathsf{E}(L(a,\theta)\,|\,y)$. The following examples illustrate how the posterior best choice can differ from the prior best choice.

Example: What if it rains? You are thinking of going out for a walk, and you need to decide whether to go and whether to take your umbrella. Suppose your cost function is

	θ_1 (rain)	θ_2 (no rain)
a_1 (stay home)	4	4
a_2 (go out, don't take umbrella)	5	0
a_3 (go out, take umbrella)	2	5

and that your state of belief about the weather for the rest of the day is modelled by the probability mass function

$$\mathsf{P}(\theta=\theta_1)=\frac{1}{2},\quad \mathsf{P}(\theta=\theta_2)=\frac{1}{2}.$$

What do you decide?

Suppose you then read the newspaper's weather forecast (y), which promises rain ($y=y_1$) for today. Your model for the accuracy of newspaper's weather forecast is

$$\mathsf{P}(y=y_1\,|\,\theta=\theta_1)=0.8,\quad \mathsf{P}(y\neq y_1\,|\,\theta=\theta_2)=0.9.$$

What do you decide now?

Solution. Before you read the weather forecast, your expected loss for choice a_3 is

$$\mathsf{E}(L(a_3,\theta))=\sum_{i=1}^{2}L(a_3,\theta_i)\mathsf{P}(\theta=\theta_i)=2\cdot\frac{1}{2}+5\cdot\frac{1}{2}=3.5,$$

and similarly $\mathsf{E}(L(a_1,\theta))=4$ and $\mathsf{E}(L(a_2,\theta))=2.5$. Then, the decision that minimizes the expected loss is to choose a_2 (i.e., go out without your umbrella).

After reading the forecast, your state of belief about the weather for the rest of the day is updated by Bayes's formula:

$$\begin{aligned}\mathsf{P}(\theta=\theta_1\,|\,y=y_1) &\propto \mathsf{P}(\theta=\theta_1)\mathsf{P}(y=y_1\,|\,\theta=\theta_1)=0.5\cdot0.8=0.4,\\ \mathsf{P}(\theta=\theta_2\,|\,y=y_1) &\propto \mathsf{P}(\theta=\theta_2)\mathsf{P}(y=y_1\,|\,\theta=\theta_2)=0.5\cdot0.1=0.05,\end{aligned}$$

where the proportionality factor is $1/\mathsf{P}(y=y_1)$. The expected losses are now

$$\begin{aligned}\mathsf{E}(L(a_1,\theta)\,|\,y_1) &\propto 4\cdot0.4+4\cdot0.05=1.8\\ \mathsf{E}(L(a_2,\theta)\,|\,y_1) &\propto 5\cdot0.4+0\cdot0.05=2.0\\ \mathsf{E}(L(a_3,\theta)\,|\,y_1) &\propto 2\cdot0.4+5\cdot0.05=1.05\end{aligned}$$

and so now the best decision is a_3 (go out with your umbrella).

In the case the newspaper had instead forecast "no rain" (y_2), your updated state of belief would have been

$$\mathsf{P}(\theta = \theta_1 \,|\, y = y_2) \propto 0.5 \cdot 0.2 = 0.1, \quad \mathsf{P}(\theta = \theta_2 \,|\, y = y_2) \propto 0.5 \cdot 0.9 = 0.45$$

and the expected losses would have been

$$\begin{aligned} \mathsf{E}(L(a_1, \theta) \,|\, y_2) &\propto 4 \cdot 0.1 + 4 \cdot 0.45 = 2.2 \\ \mathsf{E}(L(a_2, \theta) \,|\, y_2) &\propto 5 \cdot 0.1 + 0 \cdot 0.45 = 0.5 \\ \mathsf{E}(L(a_3, \theta) \,|\, y_2) &\propto 2 \cdot 0.1 + 5 \cdot 0.45 = 2.45, \end{aligned}$$

and the best decision, based on this information, would have been a_2.

Example: Monty Hall and the three doors You are a contestant in the TV game show *Let's Make a Deal* hosted by Monty Hall. After successfully completing various tasks, you arrive at the part of the show where he offers you a choice: "Door number one, door number two, or door number three?" You know that there is a valuable prize behind one door, and no prize behind the other doors, but of course you don't know which door has the prize. Which door should you choose?

After you've chosen a door, and as he *always* does at this stage of the show, Monty does not yet open your door. Instead, he opens one of the remaining doors to reveal that it does *not* have the prize behind it. He then offers you a new choice: do you stick with your initial choice, or do you switch to the other closed door?

Solution. Let θ_i denote "the prize is behind door i" and a_i denote "choose door i". When you are shown three closed doors, by symmetry (i.e. you have no reason to prefer any door over another), your prior pmf is uniform:

$$p(\theta_1) = p(\theta_2) = p(\theta_3) = \frac{1}{3}.$$

With the utility function

$$U(a, \theta_i) = \chi(a = a_i) = \begin{cases} 1 & a = a_i \\ 0 & a \neq a_i \end{cases},$$

the expected utility for a_1 is

$$\mathsf{E}(U(a_1, \theta)) = \sum_{i=1}^{3} U(a_1, \theta_i) p(\theta_i) = 1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} = \frac{1}{3},$$

and similarly $\mathsf{E}(U(a_2, \theta)) = \frac{1}{3}$ and $\mathsf{E}(U(a_3, \theta)) = \frac{1}{3}$. Thus, all three options are a-priori equally good.

Let's say you chose door 1. If the prize is behind door 1 then, according to the rules of the game, Monty could have chosen either of the other doors $y \in \{y_2, y_3\}$ to open before offering you the stick-or-switch option. By symmetry, your knowledge of which door he would choose in this case is

$$p(y_2 \,|\, \theta_1) = \frac{1}{2}, \quad p(y_3 \,|\, \theta_1) = \frac{1}{2}.$$

If the prize is not behind door 1, then Monty was obliged to open the remaining door that does not have the prize:

$$\begin{aligned} p(y_2 \,|\, \theta_2) = 0, \quad p(y_3 \,|\, \theta_2) = 1 \\ p(y_2 \,|\, \theta_3) = 1, \quad p(y_3 \,|\, \theta_3) = 0. \end{aligned}$$

Suppose Monty opened door 3. Then, by Bayes's theorem, your state of belief about the location of the prize is

$$\begin{aligned} p(\theta_1 \,|\, y_3) \propto p(y_3 \,|\, \theta_1) p(\theta_1) &= \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6} \\ p(\theta_2 \,|\, y_3) \propto p(y_3 \,|\, \theta_2) p(\theta_2) &= 1 \cdot \frac{1}{3} = \frac{1}{3} \\ p(\theta_3 \,|\, y_3) \propto p(y_3 \,|\, \theta_3) p(\theta_3) &= 0 \cdot \frac{1}{3} = 0, \end{aligned}$$

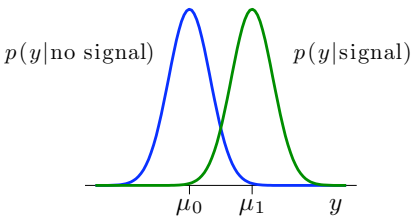
where the normalising constant is $1/p(y_3)$. The expected utilities are now

$$\mathsf{E}(U(a_1, \theta) \,|\, y_3) \propto \frac{1}{6}, \quad \mathsf{E}(U(a_2, \theta) \,|\, y_3) \propto \frac{1}{3}, \quad \mathsf{E}(U(a_3, \theta) \,|\, y_3) = 0,$$

and so you should choose to switch to door number 2. Similarly, if Monty had opened door 2, you should switch to door 3.

Example: Signal detection A transmitter either emits a signal ($\theta = 1$) or does not ($\theta = 0$). A receiver measures a corresponding voltage level μ_1 or μ_0 with $\mu_0 < \mu_1$. The measurement is corrupted by zero-mean gaussian noise:

$$y = \mu_\theta + \varepsilon, \quad \varepsilon \sim \text{Normal}(0, \sigma^2),$$



that is, $y \,|\, \theta \sim \text{Normal}(\mu_\theta, \sigma^2)$. The decision is whether to report $a = 1$: the signal is present, or $a = 0$: the signal is absent.

Solution. To solve this problem you need the prior probabilities $\pi_1 := \mathsf{P}(\theta = 1)$ and $\pi_0 := \mathsf{P}(\theta = 0) = 1 - \pi_1$, and a loss function $L(a, \theta)$, which for this problem is specified by four constants:

	$\theta = 0$	$\theta = 1$
$a = 0$	L_{00}	L_{01}
$a = 1$	L_{10}	L_{11}

Here L_{01} is the penalty for a *miss* (you fail to report a signal that is sent) and L_{10} is the penalty for a *false alarm* (you report a signal when none is sent). Normally, one has $L_{10} > L_{00}$ and $L_{01} > L_{11}$, that is, the penalties for making erroneous decisions are greater than the penalties for being correct.

Before the receiver voltage measurement value is available, the expected losses corresponding to the two options are

$$\mathsf{E}(L(0, \theta)) = L_{00}\pi_0 + L_{01}\pi_1, \quad \mathsf{E}(L(1, \theta)) = L_{10}\pi_0 + L_{11}\pi_1.$$

The best no-data decision is to choose $a = 1$ if $\mathsf{E}(L(0, \theta)) > \mathsf{E}(L(1, \theta))$, that is, if

$$\frac{(L_{01} - L_{11})\pi_1}{(L_{10} - L_{00})\pi_0} > 1,$$

and choose $a = 0$ otherwise; the corresponding loss is $\min(L_{00}\pi_0 + L_{01}\pi_1, L_{10}\pi_0 + L_{11}\pi_1)$.

Now consider the decision when the measurement y is available. By Bayes's rule, we have

$$\begin{aligned} \mathsf{P}(\theta = 0 \,|\, y) &\propto \mathsf{P}(\theta = 0) p(y \,|\, \theta = 0) \propto \pi_0 \mathrm{e}^{-(y-\mu_0)^2/(2\sigma^2)} \\ \mathsf{P}(\theta = 1 \,|\, y) &\propto \mathsf{P}(\theta = 1) p(y \,|\, \theta = 1) \propto \pi_0 \mathrm{e}^{-(y-\mu_1)^2/(2\sigma^2)} \end{aligned}$$

The expected loss for decisions $a = 0$ and $a = 1$ are

$$\begin{aligned} \mathsf{E}(L(0,\theta) \,|\, y) &= L_{00}\mathsf{P}(\theta = 0 \,|\, y) + L_{01}\mathsf{P}(\theta = 1 \,|\, y) \propto L_{00}\pi_0 \mathrm{e}^{-\frac{(y-\mu_0)^2}{2\sigma^2}} + L_{01}\pi_1 \mathrm{e}^{-\frac{(y-\mu_1)^2}{2\sigma^2}}, \\ \mathsf{E}(L(1,\theta) \,|\, y) &= L_{10}\mathsf{P}(\theta = 0 \,|\, y) + L_{11}\mathsf{P}(\theta = 1 \,|\, y) \propto L_{10}\pi_0 \mathrm{e}^{-\frac{(y-\mu_0)^2}{2\sigma^2}} + L_{11}\pi_1 \mathrm{e}^{-\frac{(y-\mu_1)^2}{2\sigma^2}}. \end{aligned}$$

The best decision is to choose $a = 1$ whenever $\mathsf{E}(L(0,\theta) \,|\, y) > \mathsf{E}(L(1,\theta) \,|\, y)$. This condition is equivalent to

$$y > \frac{\mu_1 + \mu_0}{2} - \frac{\ln(\frac{\pi_1(L_{01}-L_{11})}{\pi_0(L_{10}-L_{00})})}{(\mu_1 - \mu_0)/\sigma^2}.$$

The quantity on the right of the inequality is called the *threshold level*: the optimal decision is to report the presence of a signal if the voltage exceeds this value, and to report the absence of a signal if the voltage is below it.

14.2 Loss functions for point estimation

Point estimation means choosing a real value a that is in some sense a good approximation of θ. (To keep notation simple, the discussion here is based on θ, but all the results apply equally well to $\theta \,|\, y$.) The task of making the choice can be formulated as a decision problem. We now show that the familiar summary statistics *mean*, *median*, and *mode* are optimal decisions that correspond to certain loss functions.

Quadratic-error loss This is $L(a,\theta) = (\theta - a)^2$, for which the expected loss is

$$\begin{aligned} \mathsf{E}(L(a,\theta)) &= \int (\theta - a)^2 p(\theta)\,\mathrm{d}\theta = \int (\theta - \mathsf{E}(\theta) + \mathsf{E}(\theta) - a)^2 p(\theta)\,\mathrm{d}\theta \\ &= \int (\theta - \mathsf{E}(\theta))^2 p(\theta)\,\mathrm{d}\theta + (\mathsf{E}(\theta) - a)^2 \\ &\quad + 2(\mathsf{E}(\theta) - a) \int (\theta - \mathsf{E}(\theta))\, p(\theta)\,\mathrm{d}\theta \\ &= \mathsf{V}(\theta) + (\mathsf{E}(\theta) - a)^2, \end{aligned}$$

which is minimized when $a = \mathsf{E}(\theta)$. Thus the *mean* is the point estimate corresponding to a quadratic loss function.

Absolute-error loss This is $L(a,\theta) = |\theta - a|$, for which the point estimate is $\mathsf{median}(\theta)$. To show this, let m denote $\mathsf{median}(\theta)$, assumed for simplicity to be unique. Then $\mathsf{P}(\theta \le m) = \mathsf{P}(\theta \ge m) = \frac{1}{2}$, and for any $a > m$, we have

$$L(m,\theta) - L(a,\theta) = |\theta - m| - |\theta - a| = \begin{cases} m - a & \text{if } \theta \le m \\ 2\theta - (m + a) & \text{if } m < \theta < a \\ a - m & \text{if } \theta \ge a \end{cases}$$

Now, because $m < \theta < a \Rightarrow 2\theta < 2a \Rightarrow 2\theta - a < a \Rightarrow 2\theta - (m+a) < a - m$, we obtain

$$L(m,\theta) - L(a,\theta) \le \begin{cases} m - a & \text{if } \theta \le m \\ a - m & \text{if } m < \theta \end{cases}$$

Taking expectations gives

$$\mathsf{E}(L(m,\theta)) - \mathsf{E}(L(a,\theta)) \leq (m-a)\underbrace{\int_{-\infty}^{m} p(\theta)\,\mathrm{d}\theta}_{1/2} + (a-m)\underbrace{\int_{m}^{\infty} p(\theta)\,\mathrm{d}\theta}_{1/2} = 0.$$

Similarly, one can show that $\mathsf{E}(L(m,\theta)) - \mathsf{E}(L(a,\theta)) \leq 0$ also when $a < m$. Thus, m minimises the absolute-error loss.

Perfectionist's loss The loss function

$$L(a,\theta) = -\delta(a-\theta)$$

considers perfect, error-free estimates to be infinitely more valuable than erroneous estimates. The expected loss is

$$\mathsf{E}(L(a,\theta)) = \int -\delta(\theta-a)p(\theta)\,\mathrm{d}\theta = -p(a),$$

which is minimized by the point estimate $a = \mathsf{mode}(\theta)$. The perfectionist's loss function thus corresponds to the decision to choose the *most probable* value.

Although it is gratifying to have decision-theoretical derivations for familiar summary statistics, keep the following cautions in mind:

- The quadratic-error, absolute-error, or perfectionist's loss functions may not be appropriate models for your specific decision problem.
- If one is mainly interested in inference, without any clear need for decision, then the decision theory machinery is superfluous: the posterior distribution is itself a complete description of the state of belief.

14.3 Decision Rules and the Value of an Observation

To simplify notation in the following, we suppress the conditioning on y, writing simply θ in place of $\theta \,|\, y$, and focus on how the Bayesian choice depends on a potential or future observation $\tilde{y}$.

A *decision rule* is a function $d : \mathscr{Y} \to \mathscr{A}$, that is, it is a strategy for choosing an action a given some $\tilde{y}$. The *Bayes risk* of a decision rule is determined by taking expectation of the loss over both θ and $\tilde{y}$:

$$r(d) = \iint L(d(\tilde{y}),\theta)p(\tilde{y},\theta)\,\mathrm{d}\tilde{y}\,\mathrm{d}\theta$$

If you are working with utilities rather than losses, you can use instead the Bayes "safety"

$$s(d) = \iint U(d(\tilde{y}),\theta)p(\tilde{y},\theta)\,\mathrm{d}\tilde{y}\,\mathrm{d}\theta.$$

The *Bayes decision rule* (denoted d^*) corresponds to making the Bayesian choice

$$d^*(\tilde{y}) = \arg\min_a \mathsf{E}(L(a,\theta)\,|\,\tilde{y})$$

for every $\tilde{y} \in \mathscr{Y}$. No other decision rule gives a smaller Bayes risk, because

$$\begin{aligned} r(d) - r(d^*) &= \iint \big(L(d(\tilde{y}),\theta) - L(d^*(\tilde{y}),\theta)\big)p(\tilde{y},\theta)\,\mathrm{d}\tilde{y}\,\mathrm{d}\theta \\ &= \int \left(\int L(d(\tilde{y}),\theta)p(\theta\,|\,\tilde{y})\,\mathrm{d}\theta - \int L(d^*(\tilde{y}),\theta)p(\theta\,|\,\tilde{y})\,\mathrm{d}\theta\right) p(\tilde{y})\,\mathrm{d}\tilde{y} \\ &= \int \underbrace{\big(\mathsf{E}(L(d(\tilde{y}),\theta)\,|\,\tilde{y}) - \mathsf{E}(L(d^*(\tilde{y}),\theta)\,|\,\tilde{y})\big)}_{\geq 0} p(\tilde{y})\,\mathrm{d}\tilde{y} \geq 0. \end{aligned}$$

The *prior value* of a potential observation is the difference between the minimum Bayes risk $r(d^*)$ and the expected loss $\min_a \mathsf{E}(L(a,\theta))$ of the best "no-data" decision.

Example: What if it rains? (continued) We found earlier that the expected loss of the optimal decision that is taken before reading the weather forecast (that is, the decision a_2) is 2.5.

The Bayes decision rule was found to be

$$d^*(\tilde{y}) = \begin{cases} a_3 & \text{if } \tilde{y} = y_1 \\ a_2 & \text{if } \tilde{y} = y_2 \end{cases}$$

The corresponding Bayes risk is

$$\begin{aligned} r(d^*) &= \sum_i \mathsf{E}(L(d^*(y_i),\theta)\,|\,y_i)\mathsf{P}(\tilde{y}=y_i) \\ &= \underbrace{\mathsf{E}(L(a_3,\theta)\,|\,y_1)}_{=1.05/\mathsf{P}(\tilde{y}=y_1)}\mathsf{P}(\tilde{y}=y_1) + \underbrace{\mathsf{E}(L(a_2,\theta)\,|\,y_2)}_{=0.5/\mathsf{P}(\tilde{y}=y_2)}\mathsf{P}(\tilde{y}=y_2) \\ &= 1.55 \end{aligned}$$

Thus, the prior value of the weather forecast's information to you is $2.5 - 1.55 = 0.95$.

Example: Monty Hall and the three doors (continued) We found earlier that the expected utility for the optimal decision (which is to choose any one of the doors) before Monty opens a door is $\frac{1}{3}$.

If 1 denotes the door you initially choose and $\tilde{y} \in \{2,3\}$ denotes the door that Monty opens, the Bayes decision rule is

$$d^*(\tilde{y}) = \begin{cases} a_2 & \text{if } \tilde{y} = 3 \\ a_3 & \text{if } \tilde{y} = 2 \end{cases}$$

The corresponding Bayes safety is

$$\begin{aligned} s(d^*) &= \sum \mathsf{E}(U(d^*(y_i),\theta)\,|\,y_i)p(y_i) \\ &= \underbrace{\mathsf{E}(U(d^*(y_2),\theta)\,|\,y_2)}_{=(1/3)/p(y_2)}p(y_2) + \underbrace{\mathsf{E}(U(d^*(y_3),\theta)\,|\,y_3)}_{=(1/3)/p(y_3)}p(y_3) = \frac{2}{3} \end{aligned}$$

Thus, the prior value of the information that can be obtained by the opening of a non-winning door is $\frac{2}{3} - \frac{1}{3} = \frac{1}{3}$.

Example: Signal detection (continued) We found earlier that the expected loss for the optimal decision in the absence of a voltage observation is $\min(L_{00}\pi_0 + L_{01}\pi_1, L_{10}\pi_0 + L_{11}\pi_1)$. In particular, for perfectionist's loss and equal priors $\pi_0 = \pi_1 = \frac{1}{2}$, the expected loss is $\frac{1}{2}$.

We also found that the Bayes decision rule is $d^*(\tilde{y}) = \chi(\tilde{y} > T)$, where the threshold level is

$$T = \frac{\mu_1 + \mu_0}{2} - \frac{\ln(\frac{\pi_1(L_{01}-L_{11})}{\pi_0(L_{10}-L_{00})})}{(\mu_1 - \mu_0)/\sigma^2}.$$

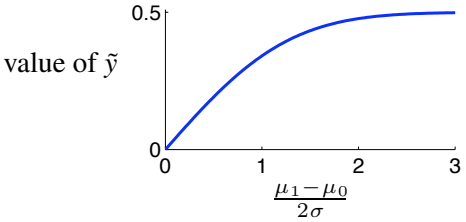
The corresponding Bayes risk is

$$\begin{aligned} r(d^*) &= \int_{-\infty}^{T} \mathsf{E}(L(0,\theta)\,|\,\tilde{y})p(\tilde{y})\,\mathrm{d}\tilde{y} + \int_{T}^{\infty} \mathsf{E}(L(1,\theta)\,|\,\tilde{y})p(\tilde{y})\,\mathrm{d}\tilde{y} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{T} \left(L_{00}\pi_0 \mathrm{e}^{-\frac{(\tilde{y}-\mu_0)^2}{2\sigma^2}} + L_{01}\pi_1 \mathrm{e}^{-\frac{(\tilde{y}-\mu_1)^2}{2\sigma^2}} \right) \mathrm{d}\tilde{y} \\ &\quad + \frac{1}{\sqrt{2\pi\sigma^2}} \int_{T}^{\infty} \left(L_{10}\pi_0 \mathrm{e}^{-\frac{(\tilde{y}-\mu_0)^2}{2\sigma^2}} + L_{11}\pi_1 \mathrm{e}^{-\frac{(\tilde{y}-\mu_1)^2}{2\sigma^2}} \right) \mathrm{d}\tilde{y} \\ &= L_{00}\pi_0 \Phi\left(\frac{T-\mu_0}{\sigma}\right) + L_{01}\pi_1 \Phi\left(\frac{T-\mu_1}{\sigma}\right) \\ &\quad + L_{10}\pi_0 \left(1 - \Phi\left(\frac{T-\mu_0}{\sigma}\right)\right) + L_{11}\pi_1 \left(1 - \Phi\left(\frac{T-\mu_1}{\sigma}\right)\right), \end{aligned}$$

where $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-t^2/2}\,\mathrm{d}t$ is the standard normal cdf. In particular, in the case of perfectionist's loss and equal priors, the threshold is $T = \frac{\mu_0+\mu_1}{2}$ and the Bayes risk is

$$r(d^*) = \frac{1}{2}\Phi\left(\frac{\mu_0-\mu_1}{2\sigma}\right) + \frac{1}{2}\left(1 - \Phi\left(\frac{\mu_1-\mu_0}{2\sigma}\right)\right) = 1 - \Phi\left(\frac{\mu_1-\mu_0}{2\sigma}\right).$$

The prior value of the voltage observation is then $\Phi\left(\frac{\mu_1-\mu_0}{2\sigma}\right) - \frac{1}{2}$:



15 Exact marginalisation

Recall that the marginal posterior distribution of some parameter of interest can always be obtained by integrating out the remaining parameters. In this section we show how these integrals can sometimes be done in closed form, yielding expressions for the pdf of the parameter of interest.

This section requires knowledge of matrix algebra.

15.1 Change Point Detection

Consider the change point detection problem from §11.5. Denoting $s_k = \sum_{i=1}^{k} y_i$, the posterior is

$$p(\lambda_1, \lambda_2, k\,|\,y) \propto \lambda_1^{\alpha_1-1+s_k} \mathrm{e}^{-(\beta_1+k)\lambda_1} \cdot \lambda_2^{\alpha_2-1+s_n-s_k} \mathrm{e}^{-(\beta_2+n-k)\lambda_2}$$

The rate parameter λ_1 is eliminated using the Gamma distribution's normalisation:

$$\int_0^{\infty} \lambda_1^{\alpha_1-1+s_k} \mathrm{e}^{-(\beta_1+k)\lambda_1}\,\mathrm{d}\lambda_1 = \frac{\Gamma(\alpha_1+s_k)}{(\beta_1+k)^{\alpha_1+s_k}}$$

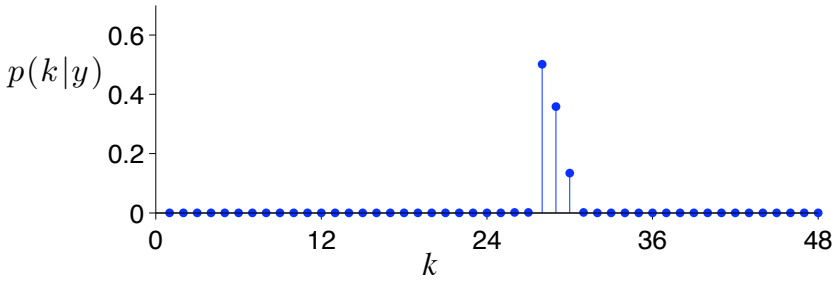
and similarly for λ_2, leaving

$$p(k\,|\,y) \propto \frac{\Gamma(\alpha_1+s_k)}{(\beta_1+k)^{\alpha_1+s_k}} \cdot \frac{\Gamma(\alpha_2+s_n-s_k)}{(\beta_1+n-k)^{\alpha_2+s_n-s_k}}$$

Dividing the the expression on the right by its sum over $k = 1, 2, \ldots, 48$ gives $p(k\,|\,y)$, the exact marginal posterior pmf of the change point k.

Here is Matlab code to compute and plot $p(k\,|\,y)$ for the example's data.

```
alpha=[.1 .1];   beta=[.1 .1];  % parameters of prior
y=[12,8,14,16,6,9,12,3,12,10,9,13,12,11,9,12,17,8,14,19,18,...
14,9,18,15,12,9,17,8,7,2,5,7,9,6,4,7,5,7,1,5,7,5,6,7,8,6,5];
n=length(y);
sk=cumsum(y);   % vector of cumulative sums
sn=sum(y);
k=1:n;          % vector 1,2,...,n
logp=gammaln(alpha(1)+sk) + gammaln(alpha(2)+sn-sk) ...
    - (alpha(1)+sk).*log(beta(1)+k) - (alpha(2)+sn-sk).*log(beta(2)+n-k);
logp=logp-max(logp);      % rescale p to avoid overflow or underflow
p=exp(logp); p=p/sum(p);  % normalise the pmf
stem(k,p)                 % plot the pmf
```



The above model assumes that there is one change point. A competing, simpler model assumes that there is no change point:

$$y_i\,|\,\lambda \sim \text{Poisson}(\lambda), \quad \lambda \sim \text{Gamma}(\alpha, \beta)$$

Denoting this simpler model $\mathscr{M}_0$ and the original model $\mathscr{M}_1$, the Bayes factor for $\mathscr{M}_1$ against $\mathscr{M}_0$ is

$$B_{10} = \sum_{k=1}^{48} \frac{1}{48} \cdot \frac{\beta_1^{\alpha_1}\beta_2^{\alpha_2}}{\beta^{\alpha}} \cdot \frac{\Gamma(\alpha_1 + s_k)\Gamma(\alpha_2 + s_n - s_k)}{\Gamma(\alpha + s_n)} \cdot \frac{(\beta + n)^{\alpha + s_n}}{(\beta_1 + k)^{\alpha_1 + s_k}(\beta_2 + n - k)^{\alpha_2 + s_n - s_k}}$$

which for this data is $B_{10} = 4.5 \cdot 10^7$. Thus, there is very strong evidence for the existence of a change point.

15.2 Multivariate normal linear model with a parameter

Consider the observation model

$$\mathbf{y}\,|\,a, \mathbf{b}, \sigma^2 \sim \text{Normal}(\mathbf{Xb}, \sigma^2\mathbf{P}^{-1}),$$

where $\mathbf{X}$ is an $n \times k$ matrix with $n > k$, $\mathbf{P}$ is an $n \times n$ symmetric positive definite matrix, and the multivariate normal distribution's density is

$$p(\mathbf{y}\,|\,a, \mathbf{b}, \sigma^2) = (2\pi\sigma^2)^{-n/2}(\det\mathbf{P})^{1/2}\mathrm{e}^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{Xb})^T\mathbf{P}(\mathbf{y}-\mathbf{Xb})}.$$

This model says that observations are equal to a linear function of the unknown parameters $b_1, \ldots, b_k$, to which is added a zero-mean gaussian noise with covariance $\sigma^2\mathbf{P}^{-1}$. Here we consider the case when one or both of $\mathbf{X}$ and $\mathbf{P}$ contain an unknown parameter a.

With the prior

$$p(a,\mathbf{b},\sigma^2) \propto p(a)\sigma^{-2}$$

the posterior is

$$p(a,\mathbf{b},\sigma^2\,|\,\mathbf{y}) \propto p(a)(\det\mathbf{P})^{1/2}(\sigma^2)^{-1-\frac{n}{2}}\mathrm{e}^{-\frac{1}{2\sigma^2}Q}$$

where

$$\begin{aligned} Q &= (\mathbf{y}-\mathbf{X}\mathbf{b})^T\mathbf{P}(\mathbf{y}-\mathbf{X}\mathbf{b}) \\ &= \mathbf{y}^T\mathbf{P}\mathbf{y}-\hat{\mathbf{b}}^T\mathbf{X}^T\mathbf{P}\mathbf{X}\hat{\mathbf{b}}+(\mathbf{b}-\hat{\mathbf{b}})^T\mathbf{X}^T\mathbf{P}\mathbf{X}(\mathbf{b}-\hat{\mathbf{b}}) \end{aligned}$$

with $\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{P}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{P}\mathbf{y}$.

Eliminate $\mathbf{b}$ via the marginalisation integral

$$\begin{aligned} p(a,\sigma^2\,|\,\mathbf{y}) &= \int p(a,\mathbf{b},\sigma^2\,|\,\mathbf{y})\,\mathrm{d}\mathbf{b} \\ &\propto p(a)\left(\frac{\det\mathbf{P}}{\det\mathbf{X}^T\mathbf{P}\mathbf{X}}\right)^{1/2}(\sigma^2)^{-1-\frac{n}{2}+\frac{k}{2}}\mathrm{e}^{-\frac{1}{2\sigma^2}(\mathbf{y}^T\mathbf{P}\mathbf{y}-\hat{\mathbf{b}}^T\mathbf{X}^T\mathbf{P}\mathbf{X}\hat{\mathbf{b}})} \end{aligned}$$

Eliminate σ^2 via the marginalisation integral

$$\begin{aligned} p(a\,|\,\mathbf{y}) &= \int p(a,\sigma^2\,|\,\mathbf{y})\,\mathrm{d}\sigma^2 \\ &\propto p(a)\left(\frac{\det\mathbf{P}}{\det\mathbf{X}^T\mathbf{P}\mathbf{X}}\right)^{1/2}\left(\mathbf{y}^T\mathbf{P}\mathbf{y}-\hat{\mathbf{b}}^T\mathbf{X}^T\mathbf{P}\mathbf{X}\hat{\mathbf{b}}\right)^{\frac{k-n}{2}} \end{aligned} \tag{20}$$

The MAP estimate of a,

$$\hat{a} = \arg\max_a p(a\,|\,\mathbf{y}),$$

can be found by plotting the logarithm of the expression in (20). If $a\,|\,\mathbf{y}$ has small dispersion, i.e. if its pdf is a narrow spike, then we can approximate the distributions of the other parameters by taking $a = \hat{a}$. In this way we obtain the approximation

$$p(\mathbf{b},\sigma^2\,|\,\mathbf{y}) \doteq (\sigma^2)^{-1-\frac{n}{2}}\mathrm{e}^{-\frac{(\mathbf{y}-\tilde{\mathbf{X}}\mathbf{b})^T\tilde{\mathbf{P}}(\mathbf{y}-\tilde{\mathbf{X}}\mathbf{b})}{2\sigma^2}}$$

where $\tilde{\mathbf{X}} = \mathbf{X}(\hat{a})$ and $\tilde{\mathbf{P}} = \mathbf{P}(\hat{a})$. Marginalising out σ^2 gives

$$\begin{aligned} p(\mathbf{b}\,|\,\mathbf{y}) &\doteq \left((\mathbf{y}-\tilde{\mathbf{X}}\mathbf{b})^T\tilde{\mathbf{P}}(\mathbf{y}-\tilde{\mathbf{X}}\mathbf{b})\right)^{n/2} \\ &= \left(\mathbf{y}^T\tilde{\mathbf{P}}\mathbf{y}-\tilde{\hat{\mathbf{b}}}^T\tilde{\mathbf{X}}^T\tilde{\mathbf{P}}\tilde{\mathbf{X}}\tilde{\hat{\mathbf{b}}}+(\mathbf{b}-\tilde{\hat{\mathbf{b}}})^T\tilde{\mathbf{X}}^T\tilde{\mathbf{P}}\tilde{\mathbf{X}}(\mathbf{b}-\tilde{\hat{\mathbf{b}}})\right)^{-n/2} \end{aligned}$$

with $\tilde{\hat{\mathbf{b}}} = (\tilde{\mathbf{X}}^T\tilde{\mathbf{P}}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{P}}\mathbf{y}$. This is a multivariate Student-t distribution; $\tilde{\hat{\mathbf{b}}}$ is its mean and its mode.

15.3 Spectrum Analysis

A well-established heuristic technique for detecting periodicity and estimating its frequency is to find the maximum of the *periodogram*. The following presentation, based on Bretthorst's book[11], shows that this procedure can be given a Bayesian interpretation.

[11] G. Larry Bretthorst, *Bayesian Spectrum Analysis and Parameter Estimation*, 1988, full text at `http://bayes.wustl.edu/glb/book.pdf`

Here the objective is to identify the frequency of a single stationary harmonic signal given a noisy time series. Assume the model

$$y_i \,|\, \omega, b_1, b_2, \sigma^2 \sim \text{Normal}(f(t_i), \sigma^2)$$

where the signal is

$$f(t) = b_1 \cos(\omega t) + b_2 \sin(\omega t)$$

and the n sampling instants are equally-spaced in $t \in [-\frac{1}{2}T, \frac{1}{2}T]$ with sampling period $\Delta = T/(n-1)$:

$$t_i = \frac{(i-1)T}{n-1} - \frac{T}{2} \quad (i \in \{1, 2, \ldots, n\})$$

Assuming conditionally independent samples, the observation model is

$$\mathbf{y} \,|\, \omega, \mathbf{b}, \sigma^2 \sim \text{Normal}(\mathbf{X}\mathbf{b}, \sigma^2 \mathbf{I})$$

where

$$\mathbf{X} = [\cos(\omega \mathbf{t}) \;\; \sin(\omega \mathbf{t})] = \begin{bmatrix} \cos(\omega t_1) & \sin(\omega t_1) \\ \cos(\omega t_2) & \sin(\omega t_2) \\ \vdots & \vdots \\ \cos(\omega t_n) & \sin(\omega t_n) \end{bmatrix}$$

This is a special case of the model in §15.1, so the marginal posterior is given by (20) with $a \to \omega$ and $\mathbf{P} \to \mathbf{I}$. Using the fact that

$$\mathbf{X}^T \mathbf{X} = \begin{bmatrix} c & 0 \\ 0 & s \end{bmatrix}$$

with

$$c = \frac{n}{2} + \frac{\sin(n\omega\Delta)}{2\sin(\omega\Delta)}, \quad s = \frac{n}{2} - \frac{\sin(n\omega\Delta)}{2\sin(\omega\Delta)},$$

and denoting

$$R = \sum_{i=1}^{n} y_i \cos(\omega t_i), \quad I = \sum_{i=1}^{n} y_i \sin(\omega t_i),$$

formula (20) can be written as

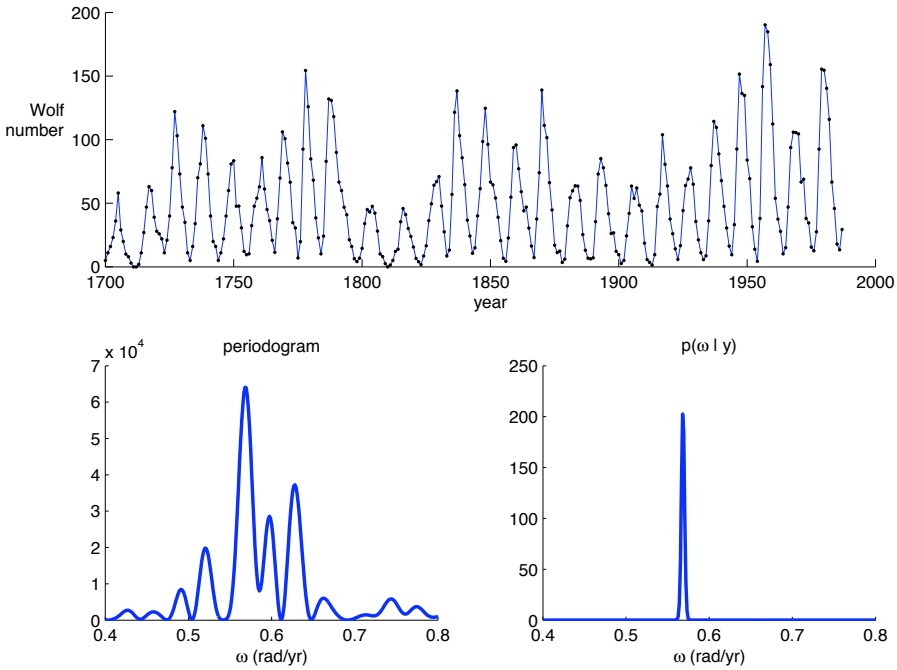
$$p(\omega \,|\, \mathbf{y}) \propto \frac{p(\omega)}{\sqrt{cs}} \left(\|\mathbf{y}\|^2 - \frac{R^2}{c} - \frac{I^2}{s} \right)^{1-\frac{n}{2}}.$$

Assuming a flat prior $p(\omega) \propto 1$, and noting that $c \approx \frac{n}{2}$ and $s \approx \frac{n}{2}$ for large n, the MAP estimate of the frequency is approximately

$$\hat{\omega} = \arg\max_{\omega} \underbrace{\frac{R^2}{n} + \frac{I^2}{n}}_{\text{periodogram}}.$$

Thus the estimate obtained by the periodogram method is pretty much the same as the posterior mode $\hat{\omega}$ found by the Bayesian method. However, in addition to a point estimate of the frequency, the Bayesian method provides the posterior distribution $p(\omega \,|\, \mathbf{y})$, from which information about the *accuracy* of the estimate, such as credibility intervals, can be obtained.

Example: Sunspots We analyse the famous time series[12] of 288 "Wolf numbers", which are measures of annual sunspot activity. Assuming a flat prior $p(\omega) \propto 1$, and using numerical quadrature to determine the normalisation constant, the pdf for $\omega \,|\, \mathbf{y}$ can be plotted to scale. The pdf has a maximum of 208.2 at $\hat{\omega} \approx 0.5684$, which corresponds to a period of 11.05 years. A normal pdf with the same maximum value has standard deviation $\frac{1}{208.2\sqrt{2\pi}} = 0.0019$, so a 95% credibility interval of $\omega \,|\, \mathbf{y}$ is $0.5684 \pm 1.95 \cdot 0.0019 = [0.5647, 0.5722]$. Thus, the period is determined to within about ± 27 days.



15.4 Autoregressive model of time series

Consider the first-order autoregressive model for a zero-mean time series

$$y_i = a y_{i-1} + e_i, \quad e_i \sim \text{Normal}(0, \sigma^2)$$

Denoting

$$\mathbf{u} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \mathbf{D} = \begin{bmatrix} 0 & & & & \\ 1 & 0 & & & \\ & 1 & 0 & & \\ & & \ddots & \ddots & \\ & & & 1 & 0 \end{bmatrix}$$

the AR(1) model can be written in the form

$$\mathbf{y} \,|\, a, y_0, \sigma^2 \sim \text{Normal}(\mathbf{X} y_0, \sigma^2 \mathbf{P}^{-1})$$

with

$$\mathbf{P} = (\mathbf{I} - a\mathbf{D})^T (\mathbf{I} - a\mathbf{D}), \quad \mathbf{X} = a(\mathbf{I} - a\mathbf{D})^{-1}\mathbf{u}.$$

It follows that

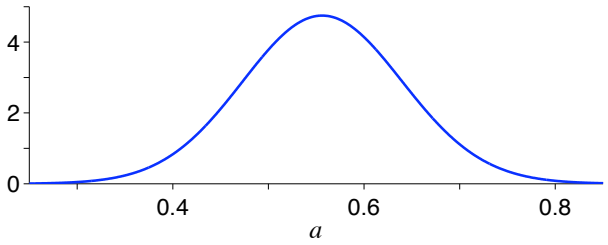
$$\det \mathbf{P} = 1, \quad \mathbf{X}^T \mathbf{P} \mathbf{X} = a^2, \quad \hat{y}_0 = \frac{y_1}{a},$$

[12] The data can be loaded into your Matlab session using `load sunspot.dat`

and so, for this problem, formula (20) can be written as

$$p(a\,|\,\mathbf{y}) \propto \frac{p(a)}{a}\left(\|\mathbf{y}-a\mathbf{D}\mathbf{y}\|^2 - y_1^2\right)^{\frac{1-n}{2}}$$

Example: earthquakes Consider the earthquake data in §11.4, from which $\bar{y}$ has been subtracted to make it zero-mean. Assuming a flat prior $p(a) \propto 1$, and using numerical quadrature to determine the normalisation constant, the pdf for $a\,|\,\mathbf{y}$ can be plotted to scale:



The maximum pdf value is 4.63 and occurs at $\hat{a} = 0.53$. A normal pdf with the same maximum value has standard deviation $\frac{1}{4.63\sqrt{2\pi}} = 0.086$, so an approximate 95% credibility interval of $a\,|\,\mathbf{y}$ is $0.53 \pm 1.95 \cdot 0.086 = [0.36, 0.70]$.

15.5 Regularisation

The least-squares method has been successfully used for hundreds of years in a wide variety of applications. There are, however, many applications where the method fails miserably because the estimate is too sensitive to noise in the data. Such difficulties typically arise in problems where $\mathbf{b}$ is high-dimensional, such as image restoration and tomography. Tikhonov regularisation (also called Ridge Regression) is a technique that has been introduced to cope with ill-conditioned least squares problems. Here we present a Bayesian derivation of this technique, together with a Bayesian approach (due to S. Gull) to estimating the "regularisation parameter".

Our starting point is the linear gaussian observation model from §15.2,

$$\mathbf{y}\,|\,\mathbf{b}, \sigma^2 \sim \text{Normal}(\mathbf{X}\mathbf{b}, \sigma^2\mathbf{P}^{-1}) \tag{21}$$

Now we assume that $\mathbf{X}$ and $\mathbf{P}$ are known, so that the likelihood pdf may be written as

$$p(\mathbf{y}\,|\,\mathbf{b}, \sigma^2) \propto (\sigma^2)^{-\frac{n}{2}} \mathrm{e}^{-\frac{1}{2\sigma^2}\mathbf{b}^T\mathbf{X}^T\mathbf{P}\mathbf{X}\mathbf{b}}.$$

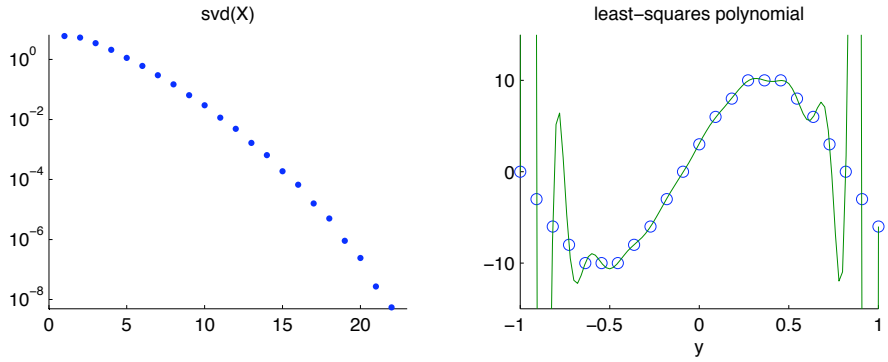
We saw in §15.2 that if we assume a flat prior for $\mathbf{b}$ and a Jeffreys prior for σ^2, we obtain the MAP estimate $\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{P}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{P}\mathbf{y}$, which coincides with the classic weighted least-squares estimate. Here is a simple example where the least squares method doesn't work.

Curve-fitting Suppose we want to construct a polynomial approximation $\sum_{j=1}^{k} b_j t^{k-j}$ of a smooth function $f(t)$, given noisy samples $y_i = f(t_i) + e_i$ at $t_1, \ldots, t_n$. Assuming the noises e_i are mutually independent and gaussian with zero mean and variance σ^2, the model is given by (21) with $\mathbf{P} = \mathbf{I}$ and

$$\mathbf{X} = \begin{bmatrix} t_1^{k-1} & \cdots & t_1^2 & t_1 & 1 \\ t_2^{k-1} & \cdots & t_2^2 & t_2 & 1 \\ \vdots & & \vdots & \vdots & \vdots \\ t_n^{k-1} & \cdots & t_n^2 & t_n & 1 \end{bmatrix}.$$

Here we consider the function $f(t) = 10\sin((1+11t)\pi/10)$ and $y_i = \text{round}(f(t_i))$, that is, the "noise" is actually caused by rounding the real values to integers. Taking $n = 23$ equally spaced points in the interval $[-1,1]$ and $k = n-1$, we obtain the least-squares fitted polynomial's coefficients

$$b_1 = 8.6895 \cdot 10^6,\ b_2 = 9.1934 \cdot 10^4,\ \ldots,\ b_{22} = 3.0819$$



The singular values of $\mathbf{X}$ span a wide range of values: the ratio of the largest to the smallest is about 10^9. This indicates that $\mathbf{X}$ is ill-conditioned, and that the least-squares solution will be sensitive to noise. Indeed, the coefficients found using the clean (i.e. not rounded) data are

$$b_1 = 6.2767 \cdot 10^{-8},\ b_2 = 6.1664 \cdot 10^{-8},\ \ldots,\ b_{22} = 3.0902$$

that is, the coefficients of the high-degree terms are quite different from those found with the noisy data! We also see that the polynomial fits the observations very well, but that it oscillates wildly, especially near the ends of the interval. □

In order to cope with ill-conditioned $\mathbf{X}$, we use the conjugate prior

$$p(\mathbf{b}\,|\,\lambda,\sigma^2) \propto (\lambda/\sigma^2)^{\frac{r}{2}} \mathrm{e}^{-\frac{\lambda}{2\sigma^2}\|\mathbf{L}\mathbf{b}\|^2}.$$

where $\mathbf{L}$ has rank r. This prior is improper if $r < k$; components of $\mathbf{b}$ that are in the null space of $\mathbf{L}$ have a flat prior, while other components are "penalised", in the sense that large values are less probable than small values. The hyperparameter λ, which is sometimes called the *regularisation parameter*, determines the strength of the penalisation.

The posterior is then

$$p(\mathbf{b},\lambda,\sigma^2\,|\,\mathbf{y}) \propto p(\lambda,\sigma^2)\lambda^{\frac{r}{2}}(\sigma^2)^{-\frac{n+r}{2}}\mathrm{e}^{-\frac{1}{2\sigma^2}Q}$$

where

$$\begin{aligned} Q &= (\mathbf{y}-\mathbf{X}\mathbf{b})^T\mathbf{P}(\mathbf{y}-\mathbf{X}\mathbf{b}) + \lambda\mathbf{b}^T\mathbf{L}^T\mathbf{L}\mathbf{b} \\ &= \mathbf{y}^T\mathbf{P}\mathbf{y} - \hat{\mathbf{b}}^T(\mathbf{X}^T\mathbf{P}\mathbf{X}+\lambda\mathbf{L}^T\mathbf{L})\hat{\mathbf{b}} + (\mathbf{b}-\hat{\mathbf{b}})^T(\mathbf{X}^T\mathbf{P}\mathbf{X}+\lambda\mathbf{L}^T\mathbf{L})(\mathbf{b}-\hat{\mathbf{b}}) \end{aligned}$$

with

$$\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{P}\mathbf{X}+\lambda\mathbf{L}^T\mathbf{L})^{-1}\mathbf{X}^T\mathbf{P}\mathbf{y}.$$

Eliminating $\mathbf{b}$ by marginalisation gives

$$\begin{aligned} p(\lambda,\sigma^2\,|\,\mathbf{y}) &= \int p(\mathbf{b},\lambda,\sigma^2\,|\,\mathbf{y})\,\mathrm{d}\mathbf{b} \\ &\propto \frac{p(\lambda,\sigma^2)\lambda^{\frac{r}{2}}(\sigma^2)^{\frac{k-n-r}{2}}}{\sqrt{\det(\mathbf{X}^T\mathbf{P}\mathbf{X}+\lambda\mathbf{L}^T\mathbf{L})}}\mathrm{e}^{-\frac{\mathbf{y}^T\mathbf{P}\mathbf{y}-\hat{\mathbf{b}}^T(\mathbf{X}^T\mathbf{P}\mathbf{X}+\lambda\mathbf{L}^T\mathbf{L})\hat{\mathbf{b}}}{2\sigma^2}} \end{aligned}$$

Assuming now a flat prior $p(\lambda, \sigma^2) \propto 1$, we eliminate σ^2 by marginalisation and obtain the posterior distribution

$$\begin{aligned} p(\lambda \,|\, \mathbf{y}) &= \int p(\lambda, \sigma^2 \,|\, \mathbf{y}) \, \mathrm{d}\sigma^2 \\ &\propto \lambda^{\frac{r}{2}} \left(\det(\mathbf{X}^T\mathbf{P}\mathbf{X} + \lambda \mathbf{L}^T\mathbf{L}) \right)^{-1/2} \left(\mathbf{y}^T\mathbf{P}\mathbf{y} - \hat{\mathbf{b}}^T (\mathbf{X}^T\mathbf{P}\mathbf{X} + \lambda \mathbf{L}^T\mathbf{L}) \hat{\mathbf{b}} \right)^{1 + \frac{k-n-r}{2}} \end{aligned}$$

The MAP estimate of λ,

$$\hat{\lambda} = \arg\max_{\lambda} p(\lambda \,|\, \mathbf{y}),$$

can be found by plotting. We can then approximate the distributions of the other parameters by taking $\lambda = \hat{\lambda}$. In this way we obtain the approximate MAP estimate

$$\hat{\mathbf{b}} \doteq (\mathbf{X}^T\mathbf{P}\mathbf{X} + \hat{\lambda}\mathbf{L}^T\mathbf{L})^{-1}\mathbf{X}^T\mathbf{P}\mathbf{y}$$

Curve-fitting (continued) Applying the above results to the curve-fitting problem with $\mathbf{L} = \mathbf{I}$, we find the MAP estimate of the regularisation parameter to be $\hat{\lambda} = 0.0017$. The coefficients corresponding to this value of λ are

$$b_1 = 4.4431, \; b_2 = 1.1616, \; \ldots, \; b_{22} = 3.0897$$

We see that regularisation has considerably reduced the amplitudes of the higher-degree terms' coefficients, and that the oscillation between data points has been eliminated: