Solutions to some exercises from *Bayesian Data Analysis*, third edition, by Gelman, Carlin, Stern, and Rubin

22 Aug 2014

These solutions are in progress. For more information on either the solutions or the book (published by CRC), check the website, `http://www.stat.columbia.edu/~gelman/book/`

For each graph and some other computations, we include the code used to create it using the S computer language. The S commands are set off from the text and appear in `typewriter font`.

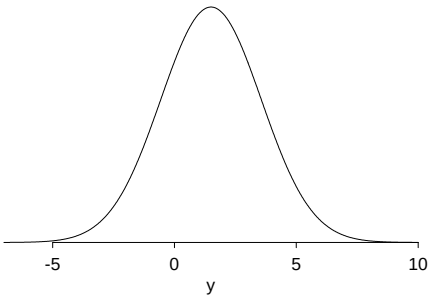
If you find any mistakes, please notify us by e-mailing to `gelman@stat.columbia.edu`. Thank you very much.

 We also thank Jiangtao Du for help in preparing some of these solutions and Ewan Cameron, Rob Creecy, Xin Feng, Lei Guo, Yi Lu, Pejman Mohammadi, Fei Shi, Dwight Sunada, Ken Williams, Corey Yanovsky, and Peng Yu for finding mistakes.

We have complete (or essentially complete) solutions for the following exercises:
Chapter 1: 1, 2, 3, 4, 5, 6
Chapter 2: 1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 16, 17, 20
Chapter 3: 1, 2, 3, 5, 9, 10
Chapter 4: 2, 3, 4, 6, 7, 9, 11, 13
Chapter 5: 3, 4, 5, 7, 8, 9, 10, 11, 12
Chapter 6: 1, 5, 6, 7
Chapter 8: 1, 2, 7, 15
Chapter 10: 4
Chapter 11: 1
Chapter 13: 7, 8
Chapter 14: 1, 3, 4, 7
Chapter 17: 1

1.1a.

$$\begin{aligned} p(y) &= \Pr(\theta=1)p(y|\theta=1) + \Pr(\theta=2)p(y|\theta=2) \\ &= 0.5\mathrm{N}(y|1,2^2) + 0.5\mathrm{N}(y|2,2^2). \end{aligned}$$




```
y <- seq(-7,10,.02)
dens <- 0.5*dnorm(y,1,2) + 0.5*dnorm(y,2,2)
plot (y, dens, ylim=c(0,1.1*max(dens)),
  type="l", xlab="y", ylab="", xaxs="i",
  yaxs="i", yaxt="n", bty="n", cex=2)
```

1.1b.

$$\begin{aligned} \Pr(\theta=1|y=1) &= \frac{p(\theta=1\,\&\,y=1)}{p(\theta=1\,\&\,y=1)+p(\theta=2\,\&\,y=1)} \\ &= \frac{\Pr(\theta=1)p(y=1|\theta=1)}{\Pr(\theta=1)p(y=1|\theta=1)+\Pr(\theta=2)p(y=1|\theta=2)} \end{aligned}$$

$$
\begin{aligned}
&= \frac{0.5\mathrm{N}(1|1,2^2)}{0.5\mathrm{N}(1|1,2^2)+0.5\mathrm{N}(1|2,2^2)}\\
&= 0.53.
\end{aligned}
$$

1.1c. As $\sigma \to \infty$, the posterior density for θ approaches the prior (the data contain no information): $\Pr(\theta = 1|y = 1) \to \frac{1}{2}$. As $\sigma \to 0$, the posterior density for θ becomes concentrated at 1: $\Pr(\theta = 1|y = 1) \to 1$.

1.2. (1.8): For each component u_i, the univariate result (1.8) states that $\mathrm{E}(u_i) = \mathrm{E}(\mathrm{E}(u_i|v))$; thus, $\mathrm{E}(u) = \mathrm{E}(\mathrm{E}(u|v))$, componentwise.

(1.9): For diagonal elements of $\mathrm{var}(u)$, the univariate result (1.9) states that $\mathrm{var}(u_i) = E(\mathrm{var}(u_i|v))+ \mathrm{var}(E(u_i|v))$. For off-diagonal elements,

$$
\begin{aligned}
&\mathrm{E}[\mathrm{cov}(u_i,u_j|v)] + \mathrm{cov}[\mathrm{E}(u_i|v),\mathrm{E}(u_j|v)]\\
&\quad = \mathrm{E}[\mathrm{E}(u_iu_j|v) - \mathrm{E}(u_i|v)\mathrm{E}(u_j|v)] + \mathrm{E}[\mathrm{E}(u_i|v)\mathrm{E}(u_j|v)] - \mathrm{E}[\mathrm{E}(u_i|v)]\mathrm{E}[\mathrm{E}(u_j|v)]\\
&\quad = \mathrm{E}(u_iu_j) - \mathrm{E}[\mathrm{E}(u_i|v)\mathrm{E}(u_j|v)] + \mathrm{E}[\mathrm{E}(u_i|v)\mathrm{E}(u_j|v)] - \mathrm{E}[\mathrm{E}(u_i|v)]\mathrm{E}[\mathrm{E}(u_j|v)]\\
&\quad = \mathrm{E}(u_iu_j) - \mathrm{E}(u_i)\mathrm{E}(u_j) = \mathrm{cov}(u_i,u_j).
\end{aligned}
$$

1.3. Note: We will use "Xx" to indicate all heterozygotes (written as "Xx or xX" in the Exercise).

$$
\begin{aligned}
&\Pr(\text{child is Xx}\,|\,\text{child has brown eyes \& parents have brown eyes})\\
&\quad = \frac{0\cdot(1-p)^4 + \frac{1}{2}\cdot 4p(1-p)^3 + \frac{1}{2}\cdot 4p^2(1-p)^2}{1\cdot(1-p)^4 + 1\cdot 4p(1-p)^3 + \frac{3}{4}\cdot 4p^2(1-p)^2}\\
&\quad = \frac{2p(1-p)+2p^2}{(1-p)^2+4p(1-p)+3p^2}\\
&\quad = \frac{2p}{1+2p}.
\end{aligned}
$$

To figure out the probability that Judy is a heterozygote, use the above posterior probability as a prior probability for a new calculation that includes the additional information that her n children are brown-eyed (with the father Xx):

$$
\Pr(\text{Judy is Xx}\,|\,n\text{ children all have brown eyes \& all previous information}) = \frac{\frac{2p}{1+2p}\cdot\left(\frac{3}{4}\right)^n}{\frac{2p}{1+2p}\cdot\left(\frac{3}{4}\right)^n + \frac{1}{1+2p}\cdot 1}.
$$

Given that Judy's children are all brown-eyed, her grandchild has blue eyes only if Judy's child is Xx. We compute this probability, recalling that we know the child is brown-eyed and we know Judy's spouse is a heterozygote:

$$
\begin{aligned}
&\Pr(\text{Judy's child is Xx}\,|\,\text{all the given information})\\
&\quad = \Pr((\text{Judy is Xx \& Judy's child is Xx})\ \textit{or}\ (\text{Judy is XX \& Judy's child is Xx})\,|\,\text{all the given information})\\
&\quad = \frac{\frac{2p}{1+2p}\cdot\left(\frac{3}{4}\right)^n}{\frac{2p}{1+2p}\cdot\left(\frac{3}{4}\right)^n + \frac{1}{1+2p}}\left(\frac{2}{3}\right) + \frac{\frac{1}{1+2p}}{\frac{2p}{1+2p}\cdot\left(\frac{3}{4}\right)^n + \frac{1}{1+2p}}\left(\frac{1}{2}\right).
\end{aligned}
$$

Given that Judy's child is Xx, the probability of the grandchild having blue eyes is 0, 1/4, or 1/2, if Judy's child's spouse is XX, Xx, or xx, respectively. Given random mating, these events have probability $(1-p)^2$, $2p(1-p)$, and p^2, respectively, and so

$$
\Pr(\text{Grandchild is xx}\,|\,\text{all the given information})
$$

$$= \frac{\frac{2}{3}\frac{2p}{1+2p}\cdot\left(\frac{3}{4}\right)^n + \frac{1}{2}\frac{1}{1+2p}}{\frac{2p}{1+2p}\cdot\left(\frac{3}{4}\right)^n + \frac{1}{1+2p}}\left(\frac{1}{4}2p(1-p) + \frac{1}{2}p^2\right)$$
$$= \frac{\frac{2}{3}\frac{2p}{1+2p}\cdot\left(\frac{3}{4}\right)^n + \frac{1}{2}\frac{1}{1+2p}}{\frac{2p}{1+2p}\cdot\left(\frac{3}{4}\right)^n + \frac{1}{1+2p}}\left(\frac{1}{2}p\right).$$

1.4a. Use relative frequencies: $\Pr(A|B) = \frac{\#\text{ of cases of } A \text{ and } B}{\#\text{ of cases of B}}$.

$$\Pr(\text{favorite wins} \,|\, \text{point spread=8}) = \frac{8}{12} = 0.67$$
$$\Pr(\text{favorite wins by at least 8} \,|\, \text{point spread=8}) = \frac{5}{12} = 0.42$$
$$\Pr(\text{favorite wins by at least 8} \,|\, \text{point spread=8 \& favorite wins}) = \frac{5}{8} = 0.63.$$

1.4b. Use the normal approximation for $d = (\text{score differential} - \text{point spread})$: $d \sim \text{N}(0, 13.86^2)$. Note: "favorite wins" means "score differential > 0"; "favorite wins by at least 8" means "score differential ≥ 8."

$$\Pr(\text{favorite wins} \,|\, \text{point spread=8}) = \Phi\left(\frac{8.5}{13.86}\right) = 0.730$$
$$\Pr(\text{favorite wins by at least 8} \,|\, \text{point spread=8}) = \Phi\left(\frac{0.5}{13.86}\right) = 0.514$$
$$\Pr(\text{favorite wins by at least 8} \,|\, \text{point spread=8 \& favorite wins}) = \frac{0.514}{0.730} = 0.70.$$

Note: the values of 0.5 and 8.5 in the above calculations are corrections for the discreteness of scores (the score differential must be an integer). The notation Φ is used for the normal cumulative distribution function.

1.5a. There are many possible answers to this question. One possibility goes as follows. We know that most Congressional elections are contested by two candidates, and that each candidate typically receives between 30% and 70% of the vote. For a given Congressional election, let n be the total number of votes cast and y be the number received by the candidate from the Democratic party. If we assume (as a first approximation, and with no specific knowledge of this election), that y/n is uniformly distributed between 30% and 70%, then

$$\Pr(\text{election is tied}|n) = \Pr(y = n/2) = \begin{cases} \frac{1}{0.4n} & \text{if } n \text{ is even} \\ 0 & \text{if } n \text{ is odd} \end{cases}.$$

If we assume that n is about 200,000, with a 1/2 chance of being even, then this approximation gives $\Pr(\text{election is tied}) \approx \frac{1}{160{,}000}$.

A national election has 435 individual elections, and so the probability of at least one of them being tied, in this analysis, is (assuming independence, since we have no specific knowledge about the elections),

$$\Pr(\text{at least one election is tied}) = 1 - \left(1 - \frac{1}{160{,}000}\right)^{435} \approx \frac{435}{160{,}000} \approx 1/370.$$

A common mistake here is to assume an overly-precise model such as $y \sim \text{Bin}(n, 1/2)$. As in the football point spreads example, it is important to estimate probabilities based on observed outcomes rather than constructing them from a theoretical model. This is relevant even in an example such as this one, where almost no information is available. (In this example, using a binomial model implies that almost all elections are extremely close, which is not true in reality.)

1.5b. An empirical estimate of the probability that an election will be decided within 100 votes is 49/20,597. The event that an election is tied is $(y = n/2)$ or, equivalently, $|2y-n| = 0$; and the event that an election is decided within 100 votes is $|y-(n-y)| \leq 100$ or, equivalently, $|2y-n| \leq 100$. Now, $(2y-n)$ is a random variable that can take on integer values. Given that n is so large (at least 50,000), and that each voter votes without knowing the outcome of the election, it seems that the distribution of $(2y-n)$ should be nearly exactly uniform near 0. Then $\Pr(|2y-n| = 0) = \frac{1}{201}\Pr(|2y-n| \leq 100)$, and we estimate the probability that an election is tied as $\frac{1}{201}\frac{49}{20{,}597}$. As in 1.5a, the probability that any of 435 elections will be tied is then approximately $435\frac{1}{201}\frac{49}{20{,}597} \approx 1/190$.

(We did not make use of the fact that 6 elections were decided by fewer than 10 votes, because it seems reasonable to assume a uniform distribution over the scale of 100 votes, on which more information is available.)

1.6. First determine the unconditional probabilities:

$$\begin{aligned}\Pr(\text{identical twins \& twin brother}) &= \Pr(\text{identical twins})\Pr(\text{both boys} \mid \text{identical twins}) = \frac{1}{2}\cdot\frac{1}{300}\\ \Pr(\text{fraternal twins \& twin brother}) &= \Pr(\text{fraternal twins})\Pr(\text{both boys} \mid \text{fraternal twins}) = \frac{1}{4}\cdot\frac{1}{125}.\end{aligned}$$

The conditional probability that Elvis was an identical twin is

$$\begin{aligned}\Pr(\text{identical twins} \mid \text{twin brother}) &= \frac{\Pr(\text{identical twins \& twin brother})}{\Pr(\text{twin brother})}\\ &= \frac{\frac{1}{2}\cdot\frac{1}{300}}{\frac{1}{2}\cdot\frac{1}{300}+\frac{1}{4}\cdot\frac{1}{125}}\\ &= \frac{5}{11}.\end{aligned}$$

2.1. Prior density:

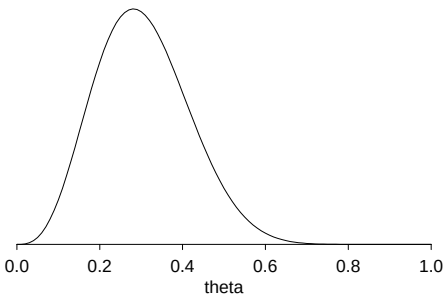
$$p(\theta) \propto \theta^3(1-\theta)^3.$$

Likelihood:

$$\begin{aligned}\Pr(\text{data}|\theta) &= \binom{10}{0}(1-\theta)^{10} + \binom{10}{1}\theta(1-\theta)^9 + \binom{10}{2}\theta^2(1-\theta)^8\\ &= (1-\theta)^{10} + 10\theta(1-\theta)^9 + 45\theta^2(1-\theta)^8.\end{aligned}$$

Posterior density:

$$p(\theta|\text{data}) \propto \theta^3(1-\theta)^{13} + 10\theta^4(1-\theta)^{12} + 45\theta^5(1-\theta)^{11}.$$




```
theta <- seq(0,1,.01)
dens <- theta^3*(1-theta)^13 + 10*theta^4*(1-theta)^12 +
  45*theta^5*(1-theta)^11
plot (theta, dens, ylim=c(0,1.1*max(dens)),
  type="l", xlab="theta", ylab="", xaxs="i",
  yaxs="i", yaxt="n", bty="n", cex=2)
```

2.2. If we knew the coin that was chosen, then the problem would be simple: if a coin has probability π of landing heads, and N is the number of additional spins required until a head, then

$$\mathrm{E}(N|\pi) = 1 \cdot \pi + 2 \cdot (1-\pi)\pi + 3 \cdot (1-\pi)^2\pi + \cdots = 1/\pi.$$

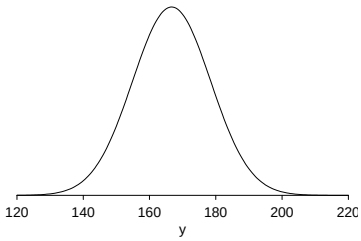
Let TT denote the event that the first two spins are tails, and let C be the coin that was chosen. By Bayes' rule,

$$\begin{aligned}\Pr(C = C_1|TT) &= \frac{\Pr(C = C_1)\Pr(TT|C = C_1)}{\Pr(C = C_1)\Pr(TT|C = C_1) + \Pr(C = C_2)\Pr(TT|C = C_2)} \\ &= \frac{0.5(0.4)^2}{0.5(0.4)^2 + 0.5(0.6)^2} = \frac{16}{52}.\end{aligned}$$

The posterior expectation of N is then

$$\begin{aligned}\mathrm{E}(N|TT) &= \mathrm{E}[\mathrm{E}(N|TT, C)|TT] \\ &= \Pr(C = C_1|TT)\mathrm{E}(N|C = C_1, TT) + \Pr(C = C_2|TT)E(N|C = C_2, TT) \\ &= \frac{16}{52}\frac{1}{0.6} + \frac{36}{52}\frac{1}{0.4} = 2.24.\end{aligned}$$

2.3a. $\mathrm{E}(y) = 1000(\frac{1}{6}) = 166.7$, and $\mathrm{sd}(y) = \sqrt{1000(\frac{1}{6})(\frac{5}{6})} = 11.8$. Normal approximation:




```
y <- seq(120,220,.5)
dens <- dnorm (y, 1000*(1/6), sqrt(1000*(1/6)*(5/6)))
plot (y, dens, ylim=c(0,1.1*max(dens)),
  type="l", xlab="y", ylab="", xaxs="i",
  yaxs="i", yaxt="n", bty="n", cex=2)
```

2.3b. From normal approximation:

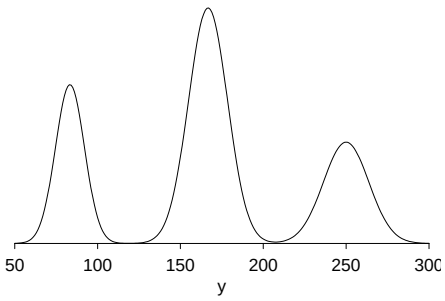
$$\begin{aligned}5\% \text{ point is } & 166.7 - 1.65(11.8) = 147.2 \\ 25\% \text{ point is } & 166.7 - 0.67(11.8) = 158.8 \\ 50\% \text{ point is } & 166.7 \\ 75\% \text{ point is } & 166.7 + 0.67(11.8) = 174.6 \\ 95\% \text{ point is } & 166.7 + 1.65(11.8) = 186.1\end{aligned}$$

Since y is discrete, round off to the nearest integer: 147, 159, 167, 175, 186.

2.4a.

$$\begin{aligned}y \,|\, \theta = \frac{1}{12} &\text{ has mean } 83.3 \text{ and sd } 8.7 \\ y \,|\, \theta = \frac{1}{6} &\text{ has mean } 166.7 \text{ and sd } 11.8 \\ y \,|\, \theta = \frac{1}{4} &\text{ has mean } 250 \text{ and sd } 13.7.\end{aligned}$$

The distribution for y is a mixture of the three conditional distributions:




```
y <- seq(50,300,1)
dens <- function (x, theta){
  dnorm (x, 1000*theta, sqrt(1000*theta*(1-theta)))}
dens.mix <- 0.25*dens(y,1/12) + 0.5*dens(y,1/6) +
  0.25*dens(y,1/4)
plot (y, dens.mix, ylim=c(0,1.1*max(dens.mix)),
  type="l", xlab="y", ylab="", xaxs="i",
  yaxs="i", yaxt="n", bty="n", cex=2)
```

2.4b. Because the three humps of the distribution have very little overlap, $\frac{1}{4}$ of the distribution of y is in the first hump, $\frac{1}{2}$ is in the second hump, and $\frac{1}{4}$ is in the third hump.

The 5% point of $p(y)$ is the 20% point of the first hump ($p(y \,|\, \theta = \frac{1}{12})$): $83.3 - (0.84)8.7 = 75.9$, round to 76. (-0.84 is the 20% point of the standard normal distribution.)

The 25% point of $p(y)$ is between the first and second humps (approximately 120, from the graph).

The 50% point of $p(y)$ is at the middle of the second hump: 166.7, round to 167.

The 75% point of $p(y)$ is between the second and third humps (approximately 205 or 210, from the graph).

The 95% point of $p(y)$ is the 80% point of the first hump: $250 + (0.84)13.7 = 261.5$, round to 262.

2.5a.

$$\begin{aligned}
\Pr(y=k) &= \int_0^1 \Pr(y=k \,|\, \theta) d\theta \\
&= \int_0^1 \binom{n}{k} \theta^k (1-\theta)^{n-k} d\theta \qquad (1) \\
&= \binom{n}{k} \frac{\Gamma(k+1)\Gamma(n-k+1)}{\Gamma(n+2)} \qquad (2) \\
&= \frac{1}{n+1}. \qquad (3)
\end{aligned}$$

To go from (1) to (2), use the identity $\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$; that is, the beta density has an integral of 1. To go from (2) to (3), use the fact that $\Gamma(x) = (x-1)!$.

2.5b. Posterior mean is $\frac{\alpha+y}{\alpha+\beta+n}$. To show that it lies between $\frac{\alpha}{\alpha+\beta}$ and $\frac{y}{n}$, we will write it as $\frac{\alpha+y}{\alpha+\beta+n} = \lambda \frac{\alpha}{\alpha+\beta} + (1-\lambda)\frac{y}{n}$, and show that $\lambda \in (0,1)$. To do this, solve for λ:

$$\begin{aligned}
\frac{\alpha+y}{\alpha+\beta+n} &= \frac{y}{n} + \lambda\left(\frac{\alpha}{\alpha+\beta} - \frac{y}{n}\right) \\
\frac{\alpha+y}{\alpha+\beta+n} - \frac{y}{n} &= \lambda\left(\frac{\alpha}{\alpha+\beta} - \frac{y}{n}\right) \\
\frac{n\alpha - \alpha y - \beta y}{(\alpha+\beta+n)n} &= \lambda\left(\frac{n\alpha - \alpha y - \beta y}{(\alpha+\beta)n}\right) \\
\lambda &= \frac{\alpha+\beta}{\alpha+\beta+n},
\end{aligned}$$

which is always between 0 and 1. So the posterior mean is a weighted average of the prior mean and the data.

2.5c. Uniform prior distribution: $\alpha = \beta = 1$. Prior variance is $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{1}{12}$.

$$\begin{aligned} \text{Posterior variance} &= \frac{(1+y)(1+n-y)}{(2+n)^2(3+n)} \\ &= \left(\frac{1+y}{2+n}\right)\left(\frac{1+n-y}{2+n}\right)\left(\frac{1}{3+n}\right). \end{aligned} \tag{4}$$

The first two factors in (4) are two numbers that sum to 1, so their product is at most $\frac{1}{4}$. And, since $n \geq 1$, the third factor is less than $\frac{1}{3}$. So the product of all three factors is less than $\frac{1}{12}$.

2.5d. There is an infinity of possible correct solutions to this exercise. For large n, the posterior variance is definitely lower, so if this is going to happen, it will be for small n. Try $n = 1$ and $y = 1$ (1 success in 1 try). Playing around with low values of α and β, we find: if $\alpha = 1, \beta = 5$, then prior variance is 0.0198, and posterior variance is 0.0255.

2.7a. The binomial can be put in the form of an exponential family with (using the notation of Section 2.4) $f(y) = \binom{n}{y}$, $g(\theta) = (1-\theta)^n$, $u(y) = y$ and natural parameter $\phi(\theta) = \log(\frac{\theta}{1-\theta})$. A uniform prior density on $\phi(\theta)$, $p(\phi) \propto 1$ on the entire real line, can be transformed to give the prior density for $\theta = e^\phi/(1+e^\phi)$:

$$q(\theta) = p\left(\frac{e^\phi}{1+e^\phi}\right)\left|\frac{d}{d\theta}\log\left(\frac{\theta}{1-\theta}\right)\right| \propto \theta^{-1}(1-\theta)^{-1}.$$

2.7b. If $y = 0$ then $p(\theta|y) \propto \theta^{-1}(1-\theta)^{n-1}$ which has infinite integral over any interval near $\theta = 0$. Similarly for $y = n$ at $\theta = 1$.

2.8a.

$$\theta|y \sim \mathrm{N}\left(\frac{\frac{1}{40^2}180 + \frac{n}{20^2}150}{\frac{1}{40^2}+\frac{n}{20^2}}, \frac{1}{\frac{1}{40^2}+\frac{n}{20^2}}\right)$$

2.8b.

$$\tilde{y}|y \sim \mathrm{N}\left(\frac{\frac{1}{40^2}180 + \frac{n}{20^2}150}{\frac{1}{40^2}+\frac{n}{20^2}}, \frac{1}{\frac{1}{40^2}+\frac{n}{20^2}} + 20^2\right)$$

2.8c.

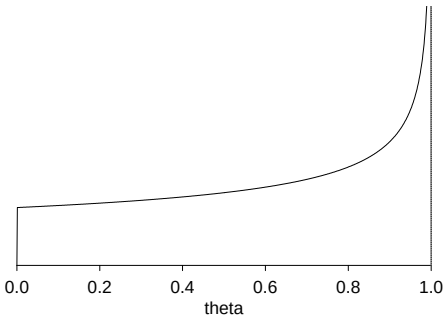
95% posterior interval for $\theta \,|\, \overline{y} = 150, n = 10$: $\quad 150.7 \pm 1.96(6.25) = [138, 163]$
95% posterior interval for $\tilde{y} \,|\, \overline{y} = 150, n = 10$: $\quad 150.7 \pm 1.96(20.95) = [110, 192]$

2.8d.

95% posterior interval for $\theta \,|\, \overline{y} = 150, n = 100$: $\quad [146, 154]$
95% posterior interval for $\tilde{y} \,|\, \overline{y} = 150, n = 100$: $\quad [111, 189]$

2.9a. From (A.3) on p. 583:

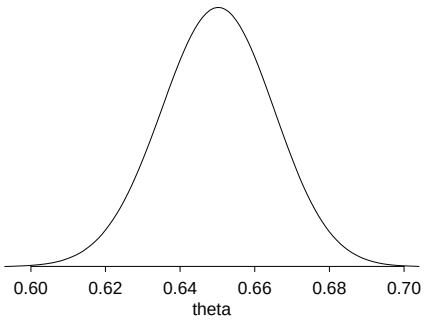
$$\begin{aligned}\alpha+\beta &= \frac{\mathrm{E}(\theta)(1-\mathrm{E}(\theta))}{\mathrm{var}(\theta)} - 1 = 1.67\\ \alpha &= (\alpha+\beta)\mathrm{E}(\theta) = 1\\ \beta &= (\alpha+\beta)(1-\mathrm{E}(\theta)) = 0.67\end{aligned}$$




```
theta <- seq(0,1,.001)
dens <- dbeta(theta,1,.67)
plot (theta, dens, xlim=c(0,1), ylim=c(0,3),
  type="l", xlab="theta", ylab="", xaxs="i",
  yaxs="i", yaxt="n", bty="n", cex=2)
lines (c(1,1),c(0,3),col=0)
lines (c(1,1),c(0,3),lty=3)
```

The density blows up at $\theta = 1$ but has a finite integral.

2.9b. $n = 1000, y = 650$. Posterior distribution is $p(\theta|y) = \mathrm{Beta}(\alpha{+}650, \beta{+}350) = \mathrm{Beta}(651, 350.67)$. The data dominate the prior distribution. $\mathrm{E}(\theta|y) = 0.6499$, $\mathrm{sd}(\theta|y) = 0.015$.




```
theta <- seq(0,1,.001)
dens <- dbeta(theta,651,350.67)
cond <- dens/max(dens) > 0.001
plot (theta[cond], dens[cond],
  type="l", xlab="theta", ylab="", xaxs="i",
  yaxs="i", yaxt="n", bty="n", cex=2)
```

2.10a.

$$\begin{aligned}p(\mathrm{data}|N) &= \begin{cases}\frac{1}{N} & \text{if } N \geq 203\\ 0 & \text{otherwise}\end{cases}\\ p(N|\mathrm{data}) &\propto p(N)p(\mathrm{data}|N)\\ &= \frac{1}{N}(0.01)(0.99)^{N-1} \text{ for } N \geq 203\\ &\propto \frac{1}{N}(0.99)^N \text{ for } N \geq 203.\end{aligned}$$

2.10b.

$$p(N|\mathrm{data}) = c\frac{1}{N}(0.99)^N.$$

We need to compute the normalizing constant, c. $\sum_N p(N|\mathrm{data}) = 1$, so

$$\frac{1}{c} = \sum_{N=203}^{\infty} \frac{1}{N}(0.99)^N.$$

This sum can be computed analytically (as $\sum_{N=0}^{\infty} \frac{1}{N}(0.99)^N - \sum_{N=0}^{202} \frac{1}{N}(0.99)^N$), but it is easier to do the computation numerically on the computer (the numerical method is also more general and can be applied even if the prior distribution does not have a simple, analytically-summable form).

$$\begin{aligned}
\text{Approximation on the computer: } \sum_{N=203}^{1000} \frac{1}{N}(0.99)^N &= 0.04658 \\
\text{Error in the approximation: } \sum_{N=1001}^{\infty} \frac{1}{N}(0.99)^N &< \frac{1}{1001} \sum_{N=1001}^{\infty} (0.99)^N \\
&= \frac{1}{1001} \frac{(0.99)^{1001}}{1-0.99} \\
&= 4.3 \times 10^{-6} \text{ (very minor).}
\end{aligned}$$

So $\frac{1}{c} = 0.04658$ and $c = 21.47$ (to a good approximation).

$$\begin{aligned}
\mathrm{E}(N|\text{data}) &= \sum_{N=203}^{\infty} N p(N|\text{data}) \\
&= c \sum_{N=203}^{\infty} (0.99)^N \\
&= 21.47 \frac{(0.99)^{203}}{1-0.99} \\
&= 279.1 \\
\mathrm{sd}(N|\text{data}) &= \sqrt{\sum_{N=203}^{\infty} (N-279.1)^2 c \frac{1}{N}(0.99)^N} \\
&\approx \sqrt{\sum_{N=203}^{1000} (N-279.1)^2 21.47 \frac{1}{N}(0.99)^N} \\
&= 79.6.
\end{aligned}$$

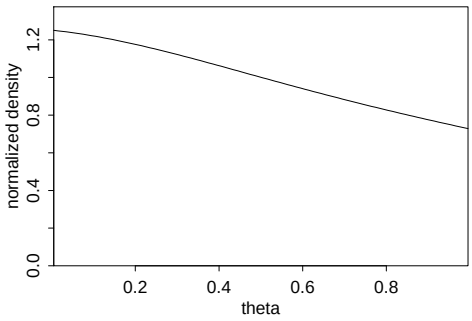
2.10c. Many possible solutions here (see Jeffreys, 1961, Lee, 1989, and Jaynes, 2003). One idea that does *not* work is the improper discrete uniform prior density on N: $p(N) \propto 1$. This density leads to an improper *posterior* density: $p(N) \propto \frac{1}{N}$, for $N \geq 203$. ($\sum_{N=203}^{\infty}(1/N) = \infty$.) The prior density $p(N) \propto 1/N$ is improper, but leads to a proper prior density, because $\sum_N 1/N^2$ is convergent.

Note also that:

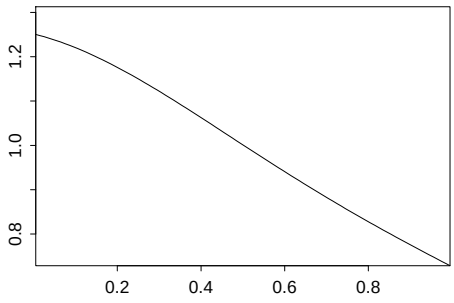
- If more than one data point is available (that is, if more than one cable car number is observed), then the posterior distribution is proper under all the above prior densities.
- With only one data point, perhaps it would not make much sense in practice to use a noninformative prior distribution here.

2.11a. Note: the solution to this exercise, as given, uses data values from an eariler edition of the book. The code should still work as long as the data are updated.

Here is the code:




```
dens <- function (y, th){
  dens0 <- NULL
  for (i in 1:length(th))
    dens0 <- c(dens0, prod (dcauchy (y, th[i], 1)))
  dens0}
y <- c(-2, -1, 0, 1.5, 2.5)
step <- .01
theta <- seq(step/2, 1-step/2, step)
dens.unnorm <- dens(y,theta)
dens.norm <- dens.unnorm/(step*sum(dens.unnorm))
plot (theta, dens.norm, ylim=c(0,1.1*max(dens.norm)),
  type="l", xlab="theta", ylab="normalized density",
  xaxs="i", yaxs="i", cex=2)
```

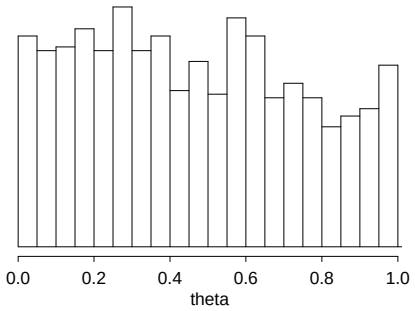


[a common mistake]

Note: a common error here is to forget to scale the y-axis from zero, thus yielding a plot as shown to the left. This is *incorrect* because it misleadingly implies that the density goes to zero at $\theta = 1$. When plotting densities, the y-axis must extend to zero!

2.11b. Note: the solution to this exercise, as given, uses data values from an eariler edition of the book. The code should still work as long as the data are updated.

Here is the code:

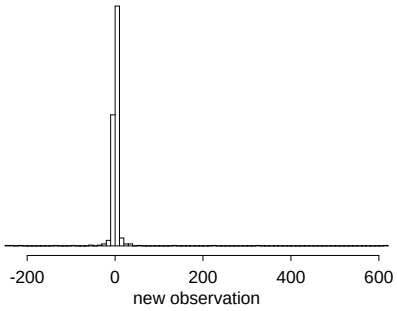



```
thetas <- sample (theta, 1000, step*dens.norm,
  replace=TRUE)
hist (thetas, xlab="theta", yaxt="n",
  breaks=seq(0,1,.05), cex=2)
```

The histogram is jagged because there are only 1000 simulation draws.

2.11c. Note: the solution to this exercise, as given, uses data values from an eariler edition of the book. The code should still work as long as the data are updated.

Here is the code:




```
y6 <- rcauchy (length(thetas), thetas, 1)
hist (y6, xlab="new observation", yaxt="n",
  nclass=100, cex=2)
```

Draws from a Cauchy distribution (or, in this case, a mixture of Cauchy distributions) do not fit well onto a histogram. Compare to Figure 4.2 from the book.

2.12. The Poisson density function is $p(y|\theta) = \theta^y e^{-\theta}/y!$, and so $J(\theta) = \mathrm{E}(-d^2 \log p(y|\theta)/d\theta^2|\theta) = \mathrm{E}(y/\theta^2) = 1/\theta$. This corresponds to an (improper) gamma density with $a = 1/2$ and $b = 0$.

2.13a. Let y_i = number of fatal accidents in year i, for $i = 1, \ldots, 10$, and θ = expected number of accidents in a year. The model for the data is $y_i|\theta \sim \mathrm{Poisson}(\theta)$.

Use the conjugate family of distributions for convenience. If the prior distribution for θ is $\mathrm{Gamma}(\alpha, \beta)$, then the posterior distribution is $\mathrm{Gamma}(\alpha + 10\overline{y}, \beta + 10)$.

Assume a noninformative prior distribution: $(\alpha, \beta) = (0, 0)$—this should be ok since we have enough information here: $n = 10$. Then the posterior distribution is $\theta|y \sim \mathrm{Gamma}(238, 10)$. Let $\tilde{y}$ be the number of fatal accidents in 1986. Given θ, the predictive distribution for $\tilde{y}$ is $\mathrm{Poisson}(\theta)$.

Here are two methods of obtaining a 95% posterior interval for $\tilde{y}$:

- **Simulation.** Draw θ from $p(\theta|y)$ and $\tilde{y}$ from $p(\tilde{y}|\theta)$:

  Computed 95% interval is $[14, 35]$.

  ```
  theta <- rgamma(1000,238)/10
  y1986 <- rpois(1000,theta)
  print (sort(y1986)[c(25,976)])
  ```

- **Normal approximation.** From gamma distribution, $\mathrm{E}(\theta|y) = 238/10 = 23.8$, $\mathrm{sd}(\theta|y) = \sqrt{238}/10 = 1.54$. From Poisson distribution, $\mathrm{E}(\tilde{y}|\theta) = \theta$, $\mathrm{sd}(\tilde{y}|\theta) = \sqrt{\theta}$.

  From (1.6) and (1.7), the mean and variance of the posterior predictive distribution for $\tilde{y}$ are:

$$\begin{aligned}
\mathrm{E}(\tilde{y}|y) &= \mathrm{E}(\mathrm{E}(\tilde{y}|\theta, y)|y) \\
&= \mathrm{E}(\theta|y) = 23.8 \\
\mathrm{var}(\tilde{y}|y) &= \mathrm{E}(\mathrm{var}(\tilde{y}|\theta, y)|y) + \mathrm{var}(\mathrm{E}(\tilde{y}|\theta, y)|y) \\
&= \mathrm{E}(\theta|y) + \mathrm{var}(\theta|y) \\
&= 26.2 = 5.12^2.
\end{aligned}$$

  Normal approximation to $p(\tilde{y}|y)$ gives a 95% interval for $\tilde{y}$ of $[23.8 \pm 1.96(5.12)] = [13.8, 33.8]$. But $\tilde{y}$ must be an integer, so interval containing at least 95% becomes $[13, 34]$.

2.13b. Estimated numbers of passenger miles in each year: for 1976, $(734/0.19)$(100 million miles) $= 3.863 \times 10^{11}$ miles; for 1977, $(516/0.12)$(100 million miles) $= 4.300 \times 10^{11}$ miles; and so forth:

Year	Estimated number of passenger miles
1976	3.863×10^{11}
1977	4.300×10^{11}
1978	5.027×10^{11}
1979	5.481×10^{11}
1980	5.814×10^{11}
1981	6.033×10^{11}
1982	5.877×10^{11}
1983	6.223×10^{11}
1984	7.433×10^{11}
1985	7.106×10^{11}

Let x_i=number of passenger miles flown in year i and θ=expected accident rate per passenger mile. The model for the data is $y_i|x_i, \theta \sim \text{Poisson}(x_i\theta)$.

Again use Gamma$(0, 0)$ prior distribution for θ. Then the posterior distribution for θ is

$$y|\theta \sim \text{Gamma}(10\overline{y}, 10\overline{x}) = \text{Gamma}(238, 5.716 \times 10^{12}).$$

Given θ, the predictive distribution for $\tilde{y}$ is Poisson$(\tilde{x}\theta) = \text{Poisson}(8 \times 10^{11}\theta)$.

Here are two methods of obtaining a 95% posterior interval for $\tilde{y}$:

- **Simulation.** Draw θ from $p(\theta|y)$ and $\tilde{y}$ from $p(\tilde{y}|\tilde{x}, \theta)$:

  Computed 95% interval is $[22, 47]$.

  ```
  theta <- rgamma(1000,238)/5.716e12
  y1986 <- rpois(1000,theta*8e11)
  print (sort(y1986)[c(25,976)])
  ```

  (Note: your answer may differ slightly due to simulation variability. Given the wide range of the posterior interval, there is no practical reason to do a huge number of simulations in order to estimate the endpoints of the interval more precisely.)

- **Normal approximation.** From gamma distribution, $\text{E}(\theta|y) = 238/(5.716 \times 10^{12}) = 4.132 \times 10^{-11}$, $\text{sd}(\theta|y) = \sqrt{238}/(5.716 \times 10^{12}) = 0.270 \times 10^{-11}$, From Poisson distribution, $\text{E}(\tilde{y}|\theta) = (8 \times 10^{11})\theta$, $\text{sd}(\tilde{y}|\tilde{x}, \theta) = \sqrt{(8 \times 10^{11})\theta}$.

  From (1.6) and (1.7), the mean and variance of the posterior predictive distribution for $\tilde{y}$ are:

$$\begin{aligned}
\text{E}(\tilde{y}|y) &= \text{E}(\text{E}(\tilde{y}|\theta, y)|y) \\
&= \text{E}((8 \times 10^{11}\theta|y) = (8 \times 10^{11})(4.164 \times 10^{-11}) = 33.3 \\
\text{var}(\tilde{y}|y) &= \text{E}(\text{var}(\tilde{y}|\theta, y)|y) + \text{var}(\text{E}(\tilde{y}|\theta, y)|y) \\
&= \text{E}((8 \times 10^{11})\theta|y) + \text{var}((8 \times 10^{11})\theta|y) \\
&= (8 \times 10^{11})(4.164 \times 10^{-11}) + (8 \times 10^{11})^2(0.270 \times 10^{-11})^2 \\
&= 38.0 = 6.2^2.
\end{aligned}$$

  Normal approximation to $p(\tilde{y}|y)$ gives a 95% interval for $\tilde{y}$ of $[33.3 \pm 1.96(6.2)] = [21.1, 45.5]$. But $\tilde{y}$ must be an integer, so interval containing at least 95% becomes $[21, 46]$.

The 95% interval for (b) is higher than for (a) because the model in (a) has constant expected accident rate per year, whereas the model in (b) has expected accident rate that is proportional to passenger miles, which are increasing over time.

2.13c. Repeat analysis from (a), replacing 238 by 6919, the total number of deaths in the data. From 1000 simulation draws: 95% posterior interval is $[638, 750]$ deaths.

2.13d. Repeat analysis from (b), replacing 238 by 6919. From 1000 simulation draws: 95% posterior interval is $[900, 1035]$ deaths.

2.13e. Just based on general knowledge, without specific reference to the data in Table 2.2, the Poisson model seems more reasonable with rate proportional to passenger miles (models (b) and (d)), because if more miles are flown in a year, you would expect more accidents. On the other hand, if airline safety is improving, we would expect the accident rate to decline over time, so maybe the expected number of accidents would remain roughly constant (as the number of passenger miles flown is gradually increasing). The right thing to do here is to put a time trend in the model, as in Exercise 3.12.

Also just based on general knowledge, we would expect accidents to be independent. But passenger deaths are *not independent*—they occur in clusters when airplanes crash. So the Poisson model should be more reasonable for accidents (models (a) and (b)) than for total deaths. What if you were interested in total deaths? Then it might be reasonable to set up a *compound model*: a Poisson distribution for accidents, and then another distribution for deaths given accidents.

In addition, these models can be checked by comparing to data (see Exercise 6.2).

2.16a.

$$\begin{aligned} p(y) &= \int p(y|\theta)p(\theta)d\theta \\ &= \int_0^1 \binom{n}{y}\theta^y(1-\theta)^{n-y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta \\ &= \frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}d\theta \\ &= \frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(n+\alpha+\beta)}, \end{aligned}$$

which is the beta-binomial density (see Appendix A).

2.16b. We are interested only in seeing whether $p(y)$ varies with y, so we need only look at the factors in $p(y)$ that depend on y; thus, we investigate under what circumstances the quantity $\Gamma(a+y)\Gamma(b+n-y)/\Gamma(y+1)\Gamma(n-y+1)$ changes with y.

The preceding expression evaluates to 1 if $a = b = 1$; hence $p(y)$ is constant in y if $a = b = 1$.

On the other hand, suppose that $p(y)$ is constant in y. Then, in particular, $p(0) = p(n)$ and $p(0) = p(1)$. The first equality implies that $\Gamma(a)\Gamma(b+n)/\Gamma(1)\Gamma(n+1) = \Gamma(a+n)\Gamma(b)/\Gamma(n+1)\Gamma(1)$, so that $\Gamma(a)\Gamma(b+n) = \Gamma(a+n)\Gamma(b)$.

Recalling the formula $\Gamma(t) = (t-1)\Gamma(t-1)$, we must have $\Gamma(a)\Gamma(b)(b+n-1)...(b+1)(b) = \Gamma(a)\Gamma(b)(a+n-1)...(a+1)(a)$, which implies that $(b+n-1)...(b+1)(b) = (a+n-1)...(a+1)(a)$. Since each term in each product is positive, it follows that $a = b$.

Continuing, $p(0) = p(1)$ implies $\Gamma(a)\Gamma(b+n)/\Gamma(1)\Gamma(n+1) = \Gamma(a+1)\Gamma(b+n-1)/\Gamma(2)\Gamma(n)$; again using $\Gamma(t) = (t-1)\Gamma(t-1)$, we have $b+n-1 = na$. Since $a = b$, this reduces to $a+n-1 = na$, whose only solution is $a = 1$.

Therefore $a = b = 1$ is a necessary as well as sufficient condition for $p(y)$ to be constant in y.

2.17a. Let $u = \sigma^2$, so that $\sigma = \sqrt{u}$. Then $p(\sigma^2) = p(u) = p(\sqrt{u})d\sqrt{u}/du = p(\sigma)(1/2)u^{-1/2} = (1/2)p(\sigma)/\sigma$, which is proportional to $1/\sigma^2$ if $p(\sigma) \propto \sigma^{-1}$.

2.17b. Proof by contradiction: The posterior density $p(\sigma|\text{data})$ is proportional to $\sigma^{-1-n}\exp(-c/\sigma^2)$, which may be written as $(\sigma^2)^{-1/2-n/2}\exp(-c/\sigma^2)$. The posterior density $p(\sigma^2|\text{data})$ is proportional to $(\sigma^2)^{-1-n/2}\exp(-c/\sigma^2)$. In the preceding expressions, we have defined $c = nv/2$, and we assume this quantity to be positive.

Let $(\sqrt{a}, \sqrt{b})$ be the 95% interval of highest density for $p(\sigma|\text{data})$. (It can be shown using calculus that the density function is unimodal: there is only one value of σ for which $dp(\sigma|\text{data})/d\sigma = 0$. Therefore the 95% region of highest density is a single interval.) Then

$$a^{-1/2-n/2}\exp(-c/a) = b^{-1/2-n/2}\exp(-c/b).$$

Equivalently, $(-1/2 - n/2)\log a - c/a = (-1/2 - n/2)\log b - c/b$.

Now, if (a, b) were the 95% interval of highest density for $p(\sigma^2|\text{data})$, then

$$a^{-1-n/2}\exp(-c/a) = b^{-1-n/2}\exp(-c/b).$$

That is, $(-1 - n/2)\log a - c/a = (-1 - n/2)\log b - c/b$. Combining the two equations, $1/2\log a = 1/2\log b$, so that $a = b$, in which case $[a, b]$ cannot be a 95% interval, and we have a contradiction.

2.20a.

$$\begin{aligned} p(\theta \,|\, y \geq 100) &\propto p(y \geq 100 \,|\, \theta)p(\theta) \\ &\propto \exp(-100\theta)\theta^{\alpha-1}\exp(-\beta\theta) \\ p(\theta \,|\, y \geq 100) &= \text{Gamma}(\theta \,|\, \alpha, \beta + 100) \end{aligned}$$

The posterior mean and variance of θ are $\frac{\alpha}{\beta+100}$ and $\frac{\alpha}{(\beta+100)^2}$, respectively.

2.20b.

$$\begin{aligned} p(\theta \,|\, y = 100) &\propto p(y = 100 \,|\, \theta)p(\theta) \\ &\propto \theta\exp(-100\,\theta)\theta^{\alpha-1}\exp(-\beta\theta) \\ p(\theta \,|\, y = 100) &= \text{Gamma}(\theta \,|\, \alpha + 1, \beta + 100) \end{aligned}$$

The posterior mean and variance of θ are $\frac{\alpha+1}{\beta+100}$ and $\frac{\alpha+1}{(\beta+100)^2}$, respectively.

2.20c. Identity (2.8) says *on average*, the variance of θ decreases given more information: in this case,

$$\text{E}(\text{var}(\theta|y) \,|\, y \geq 100) \leq \text{var}(\theta \,|\, y \geq 100). \tag{5}$$

Plugging in $y = 100$ to get $\text{var}(\theta|y = 100)$ is not the same as averaging over the distribution of $y \,|\, y \geq 100$ on the left side of (5).

3.1a Label the prior distribution $p(\theta)$ as Dirichlet$(a_1, \ldots, a_n)$. Then the posterior distribution is $p(\theta|y) = \text{Dirichlet}(y_1 + a_1, \ldots, y_n + a_n)$. From the properties of the Dirichlet distribution (see Appendix A), the marginal posterior distribution of $(\theta_1, \theta_2, 1 - \theta_1 - \theta_2)$ is also Dirichlet:

$$p(\theta_1, \theta_2|y) \propto \theta_1^{y_1+a_1-1}\theta_2^{y_2+a_2-1}(1-\theta_1-\theta_2)^{y_{\text{rest}}+a_{\text{rest}}-1}, \quad \text{where } y_{\text{rest}} = y_3+\ldots+y_J,\ a_{\text{rest}} = a_3+\ldots+a_J.$$

(This can be proved using mathematical induction, by first integrating out θ_n, then θ_{n-1}, and so forth, until only θ_1 and θ_2 remain.)

Now do a change of variables to $(\alpha, \beta) = (\frac{\theta_1}{\theta_1+\theta_2}, \theta_1 + \theta_2)$. The Jacobian of this transformation is $|1/\beta|$, so the transformed density is

$$\begin{aligned} p(\alpha, \beta|y) &\propto \beta(\alpha\beta)^{y_1+a_1-1}((1-\alpha)\beta)^{y_2+a_2-1}(1-\beta)^{y_{\text{rest}}+a_{\text{rest}}-1} \\ &= \alpha^{y_1+a_1-1}(1-\alpha)^{y_2+a_2-1}\beta^{y_1+y_2+a_1+a_2-1}(1-\beta)^{y_{\text{rest}}+a_{\text{rest}}-1} \\ &\propto \text{Beta}(\alpha|y_1+a_1, y_2+a_2)\text{Beta}(\beta|y_1+y_2+a_1+a_2, y_{\text{rest}}+a_{\text{rest}}). \end{aligned}$$

Since the posterior density divides into separate factors for α and β, they are independent, and, as shown above, $\alpha|y \sim \text{Beta}(y_1+a_1, y_2+a_2)$.

3.1b. The $\text{Beta}(y_1+a_1, y_2+a_2)$ posterior distribution can also be derived from a $\text{Beta}(a_1, a_2)$ prior distribution and a binomial observation y_1 with sample size $y_1 + y_2$.

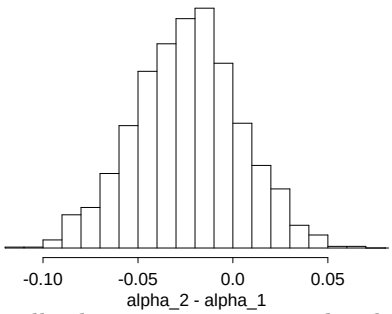
3.2. Assume independent uniform prior distributions on the multinomial parameters. Then the posterior distributions are independent multinomial:

$$\begin{aligned} (\pi_1, \pi_2, \pi_3)|y &\sim \text{Dirichlet}(295, 308, 39) \\ (\pi_1^*, \pi_2^*, \pi_3^*)|y &\sim \text{Dirichlet}(289, 333, 20), \end{aligned}$$

and $\alpha_1 = \frac{\pi_1}{\pi_1+\pi_2}$, $\alpha_2 = \frac{\pi_1^*}{\pi_1^*+\pi_2^*}$. From the properties of the Dirichlet distribution (see Exercise 3.1),

$$\begin{aligned} \alpha_1|y &\sim \text{Beta}(295, 308) \\ \alpha_2|y &\sim \text{Beta}(289, 333). \end{aligned}$$

The histogram of 2000 draws from the posterior density of $\alpha_2 - \alpha_1$ is attached. Based on this histogram, the posterior probability that there was a shift toward Bush is 19%.




```
alpha.1 <- rbeta (2000, 295, 308)
alpha.2 <- rbeta (2000, 289, 333)
dif <- alpha.2 - alpha.1
hist (dif, xlab="alpha_2 - alpha_1", yaxt="n",
  breaks=seq(-.12,.08,.01), cex=2)
print (mean(dif>0))
```

Essentially the same answer may be obtained using normal approximations for the distributions of α_2 and α_1 with means and standard deviations computed from the relevant beta distributions.

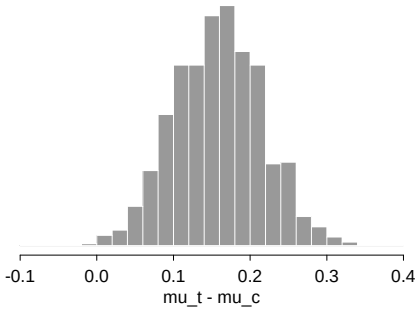
3.3a. Data distribution is $p(y|\mu_c, \mu_t, \sigma_c, \sigma_t) = \prod_{i=1}^{32} \text{N}(y_{c\,i}|\mu_c, \sigma_c^2) \prod_{i=1}^{36} \text{N}(y_{t\,i}|\mu_t, \sigma_t^2)$. Posterior distribution is

$$\begin{aligned} p(\mu_c, \mu_t, \log\sigma_c, \log\sigma_t|y) &= p(\mu_c, \mu_t, \log\sigma_c, \log\sigma_t)p(y|\mu_c, \mu_t, \log\sigma_c, \log\sigma_t) \\ &= \underbrace{\prod_{i=1}^{32} \text{N}(y_{c\,i}|\mu_c, \sigma_c^2)}_{p(\mu_c, \log\sigma_c|y)} \underbrace{\prod_{i=1}^{36} \text{N}(y_{t\,i}|\mu_t, \sigma_t^2)}_{p(\mu_t, \log\sigma_t|y)}. \end{aligned}$$

The posterior density factors, so (μ_c, σ_c) are independent of (μ_t, σ_t) in the posterior distribution. So, under this model, we can analyze the two experiments separately. Using the results from Section 3.2, the marginal posterior distributions for μ_c and μ_t are:

$$\begin{aligned}\mu_c|y &\sim t_{31}(1.013, \underbrace{0.24^2/32}_{0.0424^2}) \\ \mu_t|y &\sim t_{35}(1.173, \underbrace{0.20^2/36}_{0.0333^2}).\end{aligned}$$

3.3b The histogram of 1000 draws from the posterior density of $\mu_t - \mu_c$ is attached. Based on this histogram, a 95% posterior interval for the average treatment effect is $[0.05, 0.27]$.




```
mu.c <- 1.013 + (0.24/sqrt(32))*rt(1000,31)
mu.t <- 1.173 + (0.20/sqrt(36))*rt(1000,35)
dif <- mu.t - mu.c
hist (dif, xlab="mu_t - mu_c", yaxt="n",
  breaks=seq(-.1,.4,.02), cex=2)
print (sort(dif)[c(25,976)])
```

3.5a. If the observations are treated as exact measurements then the theory of Section 3.2 applies. Then $p(\sigma^2|y) \sim \text{Inv-}\chi^2(n-1, s^2)$ and $p(\mu|\sigma^2, y) \sim N(\bar{y}, \sigma^2/n)$ with $n = 5, \bar{y} = 10.4, s^2 = 1.3$.

3.5b. The posterior distribution assuming the noninformative prior distribution $p(\mu, \sigma^2) \propto 1/\sigma^2$ is

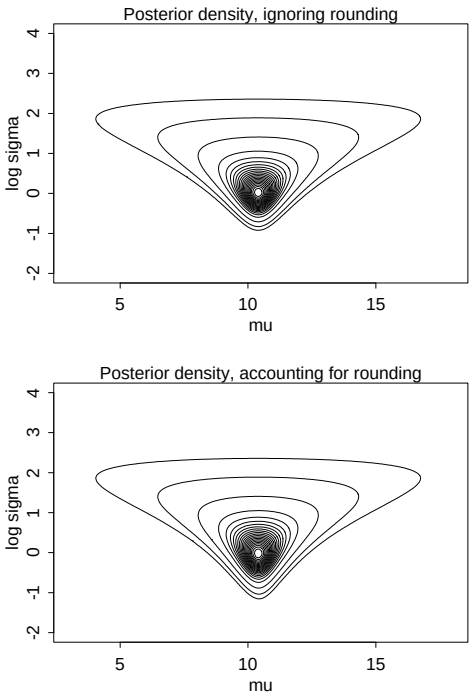
$$p(\mu, \sigma^2|y) \propto \frac{1}{\sigma^2} \prod_{i=1}^{n} \left(\Phi\left(\frac{y_i + 0.5 - \mu}{\sigma}\right) - \Phi\left(\frac{y_i - 0.5 - \mu}{\sigma}\right)\right),$$

where Φ is the standard normal cumulative distribution function.

3.5c. We computed contour plots on a grid on the $(\mu, \log\sigma)$ scale (levels are 0.0001, 0.001, 0.01, 0.05–0.95). Draws from the posterior density in 3.5a can be obtained directly as described in Section 3.2. Because we are working on the scale of $\log\sigma$ rather than σ^2, we express the prior density on that scale (that is, $p(\log\sigma) \propto 1$ rather than $p(\sigma^2) \propto \sigma^{-2}$). Draws from the posterior density in 3.5b were obtained by sampling from the grid approximation (again using the $(\mu, \log\sigma)$ scale).

	mean	sd	2.5%	25%	50%	75%	97.5%
Ignoring rounding							
μ	10.4	0.74	8.9	10.1	10.4	10.8	11.8
σ	1.4	0.74	0.68	0.98	1.2	1.6	3.3
Accounting for rounding							
μ	10.5	0.69	9.2	10.1	10.4	10.8	11.8
σ	1.4	0.73	0.61	0.90	1.2	1.6	3.3

The only notable difference is in the lower quantiles of σ.




```
post.a <- function(mu,sd,y){
  ldens <- 0
  for (i in 1:length(y)) ldens <- ldens +
    log(dnorm(y[i],mu,sd))
  ldens}
post.b <- function(mu,sd,y){
  ldens <- 0
  for (i in 1:length(y)) ldens <- ldens +
    log(pnorm(y[i]+0.5,mu,sd) - pnorm(y[i]-0.5,mu,sd))
  ldens}
summ <- function(x){c(mean(x),sqrt(var(x)),
  quantile(x, c(.025,.25,.5,.75,.975)))}

nsim <- 2000
y <- c(10,10,12,11,9)
n <- length(y)
ybar <- mean(y)
s2 <- sum((y-mean(y))^2)/(n-1)
mugrid <- seq(3,18,length=200)
logsdgrid <- seq(-2,4,length=200)
contours <- c(.0001,.001,.01,seq(.05,.95,.05))
logdens <- outer (mugrid, exp(logsdgrid), post.a, y)
dens <- exp(logdens - max(logdens))
contour (mugrid, logsdgrid, dens, levels=contours,
  xlab="mu", ylab="log sigma", labex=0, cex=2)
mtext ("Posterior density, ignoring rounding", 3)
sd <- sqrt((n-1)*s2/rchisq(nsim,4))
mu <- rnorm(nsim,ybar,sd/sqrt(n))
print (rbind (summ(mu),summ(sd)))

logdens <- outer (mugrid, exp(logsdgrid), post.b, y)
dens <- exp(logdens - max(logdens))
contour (mugrid, logsdgrid, dens, levels=contours,
  xlab="mu", ylab="log sigma", labex=0, cex=2)
mtext ("Posterior density, accounting for rounding",
  cex=2, 3)
dens.mu <- apply(dens,1,sum)
muindex <- sample (1:length(mugrid), nsim, replace=T,
  prob=dens.mu)
mu <- mugrid[muindex]
sd <- rep (NA,nsim)
for (i in (1:nsim)) sd[i] <- exp (sample
  (logsdgrid, 1, prob=dens[muindex[i],]))
print (rbind (summ(mu),summ(sd)))
```

3.5d. Given μ and σ, the conditional distribution of z_i is $\mathrm{N}(\mu, \sigma^2)$, truncated to fall in the range $(y_i - 0.5, y_i + 0.5)$. So, for each of our posterior draws of (μ, σ), we draw a vector $(z_1, \ldots, z_5)$ from independent truncated normal distributions using the inverse-cdf method.

The posterior mean of $(z_1 - z_2)^2$ is 0.16.

```
z <- matrix (NA, nsim, length(y))
for (i in 1:length(y)){
  lower <- pnorm (y[i]-.5, mu, sd)
  upper <- pnorm (y[i]+.5, mu, sd)
  z[,i] <- qnorm (lower + runif(nsim)*(upper-lower), mu, sd)}
mean ((z[,1]-z[,2])^2)
```

3.9.

$$\begin{aligned}
p(\mu,\sigma^2|y) &\propto p(y|\mu,\sigma^2)p(\mu,\sigma^2) \\
&\propto (\sigma^2)^{-n/2}\exp\left(-\frac{(n-1)s^2+n(\mu-\bar{y})^2}{2\sigma^2}\right)\sigma^{-1}(\sigma^2)^{-(\nu_0/2+1)}\exp\left(-\frac{\nu_0\sigma_0^2+\kappa_0(\mu-\mu_0)^2}{2\sigma^2}\right) \\
&\propto \sigma^{-1}(\sigma^2)^{-((\nu_0+n)/2+1)}\exp\left(-\frac{\nu_0\sigma_0^2+(n-1)s^2+\frac{n\kappa_0(\bar{y}-\mu_0)^2}{n+\kappa_0}+(n+\kappa_0)(\mu-\frac{\mu_0\kappa_0+n\bar{y}}{n+\kappa_0})^2}{2\sigma^2}\right)
\end{aligned}$$

And so,

$$\mu,\sigma^2|y \sim \text{N-Inv-}\chi^2\left(\frac{\mu_0\kappa_0+n\bar{y}}{n+\kappa_0},\frac{\sigma_n^2}{n+\kappa_0};n+\nu_0,\sigma_n^2\right),$$

where

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

3.10. From equation (3.4), $p(\sigma_j^2|y)\propto(\sigma_j^2)^{-n/2-1/2}\exp(-(n-1)s^2/2\sigma_j^2)$ for each j. Thus $p(1/\sigma_j^2|y)\propto(\sigma_j^2)^2(1/\sigma_j^2)^{n/2+1/2}\exp(-(n-1)s^2/2\sigma_j^2)=(1/\sigma_j^2)^{n/2-3/2}\exp(-(n-1)s^2/2\sigma_j^2)$.

Since the proportionality is in σ_j^2, it is also correct to say that the latter object is proportional to $((n-1)s^2/\sigma_j^2)^{n/2-3/2}\exp(-(n-1)s^2/2\sigma_j^2)=((n-1)s^2/\sigma_j^2)^{(n-1)/2-1}\exp(-(n-1)s^2/2\sigma_j^2)$, which implies that $(n-1)s^2/\sigma_j^2$ has a χ^2_{n-1} distribution.

The independence assumptions imply that $(n_1-1)s_1^2/\sigma_1^2$ and $(n_2-1)s_2^2/\sigma_2^2$ are independent χ^2 with n_1-1 and n_2-1 degrees of freedom, respectively, and so we invoke the well-known result that the quotient of two independent χ^2 random variables, each divided by their degrees of freedom, has the F distribution. In particular, $\frac{s_1^2/\sigma_1^2}{s_2^2/(\sigma_2)^2}$ has the F_{n_1-1,n_2-1} distribution. A trivial rearrangement yields the desired result.

4.2.

$$\begin{aligned}
p(y_i|\theta) &\propto (\text{logit}^{-1}(\alpha+\beta x_i))^{y_i}(1-\text{logit}^{-1}(\alpha+\beta x_i))^{n_i-y_i} \\
l_i=\log p(y_i|\theta) &= \text{constant }+y_i\log(\text{logit}^{-1}(\alpha+\beta x_i))+(n_i-y_i)\log(1-\text{logit}^{-1}(\alpha+\beta x_i)) \\
\frac{d^2l_i}{d\alpha^2} &= -\frac{n_i\exp(\alpha+\beta x_i)}{(1+\exp(\alpha+\beta x_i))^2} \\
\frac{d^2l_i}{d\alpha d\beta} &= -\frac{n_ix_i\exp(\alpha+\beta x_i)}{(1+\exp(\alpha+\beta x_i))^2} \\
\frac{d^2l_i}{d\beta^2} &= -\frac{n_ix_i^2\exp(\alpha+\beta x_i)}{(1+\exp(\alpha+\beta x_i))^2}
\end{aligned}$$

The prior density on (α,β) is uniform, so $\log p(\theta|y)=\text{constant }+\sum_{i=1}^4 l_i$, and

$$I(\hat{\theta})=\left.\begin{pmatrix}\sum_{i=1}^4\frac{n_i\exp(\alpha+\beta x_i)}{(1+\exp(\alpha+\beta x_i))^2} & \sum_{i=1}^4\frac{n_ix_i\exp(\alpha+\beta x_i)}{(1+\exp(\alpha+\beta x_i))^2}\\ \sum_{i=1}^4\frac{n_ix_i\exp(\alpha+\beta x_i)}{(1+\exp(\alpha+\beta x_i))^2} & \sum_{i=1}^4\frac{n_ix_i^2\exp(\alpha+\beta x_i)}{(1+\exp(\alpha+\beta x_i))^2}\end{pmatrix}\right|_{(\alpha,\beta)=(\hat{\alpha},\hat{\beta})},$$

where $\hat{\theta}=(\hat{\alpha},\hat{\beta})$ is the posterior mode. Denoting I as $\begin{pmatrix}a & c\\ c & b\end{pmatrix}$, the normal approximation variances of α and β are the diagonal elements of I^{-1}: $\frac{b}{ab-c^2}$, and $\frac{a}{ab-c^2}$, respectively.

4.3. Let $\theta = \text{LD50}$; we also introduce $\nu = \beta$, in anticipation of a change of coordinates. Formula (4.1) suggests that the (asymptotic) posterior median and mode should coincide, and the (asymptotic) posterior standard variance should be the inverse of observed information, evaluated at the posterior mode.

With some effort (or using the "generalized linear model" function in a statistics package), it is possible to obtain decent numerical estimates for the posterior mode and standard deviation associated with this data set; we will take these values as proxies for their asymptotic analogues. Here is one way to proceed:

Observe that $p(\theta|y) = \int p(\theta, \nu|y)d\nu = \int p(\alpha, \beta|y)|\nu|d\nu$, where $\alpha = -\theta\nu$ and $\beta = \nu$ in the last integral, and $|\nu|$ is the Jacobian associated with the change in coordinates. (An expression for $p(\alpha, \beta|y)$ is given as (3.16).)

Fixing θ equal to a value around -0.12, which appears to be near the posterior mode (see Figure 3.4), we may compute $\int p(\alpha, \beta|y)|\nu|d\nu$ numerically (up to a proportionality constant that does not depend on θ). The region of integration is infinite, but the integrand decays rapidly after ν passes 50, so that it suffices to estimate the integral assuming that ν runs from 0 to, say, 70.

This procedure can be repeated for several values of θ near -0.12. The values may be compared directly to find the posterior mode for θ. To three decimal places, we obtain -0.114.

One can fix a small value of h, such as $h = 0.002$, and compute $d^2/d\theta^2 \log p(\theta|y)$, evaluated at θ equal to the posterior mode, by the expression $[\log p(-0.114+h|y) - 2\log p(-0.114|y) + \log p(-0.114-h|y)]/h^2$ (see (13.2)).

The negative of of the preceding quantity, taken to the -0.5 power, is our estimate for the posterior standard deviation. We get 0.096, which appears quite reasonable when compared to Figure 3.4, but which seems smaller than what we might guess from Figure 4.2(b). As remarked on page 87, however, the actual posterior distribution of LD50 has less dispersion than what is suggested by Figure 4.2.

4.4. In the limit as $n \to \infty$, the posterior variance approaches zero—that is, the posterior distribution becomes concentrated near a single point. Any one-to-one continuous transformation on the real numbers is locally linear in the neighborhood of that point.

4.6. We assume for simplicity that the posterior distribution is continuous and that needed moments exist. The proofs can be easily adapted to discontinuous distributions.

4.6a. The calculus below shows that $L(a|y)$ has zero derivative at $a = \text{E}(\theta|y)$. Also the second derivative is positive so this is a minimizing value.

$$\frac{d}{da}\text{E}(L(a|y)) = \frac{d}{da}\int (\theta - a)^2 p(\theta|y)d\theta = -2\int (\theta - a)p(\theta|y)d\theta = -2(\text{E}(\theta|y) - a) = 0 \text{ if } a = \text{E}(\theta|y).$$

4.6b. We can apply the argument from 4.6c with $k_o = k_1 = 1$.

4.6c. The calculus below shows that $L(a|y)$ has zero derivative at any a for which $\int_{-\infty}^{a} p(\theta|y)d\theta = k_0/(k_0 + k_1)$. Once again the positive second derivative indicates that it is a minimizing value.

$$\begin{aligned}
\frac{d}{da}\text{E}(L(a|y)) &= \frac{d}{da}\left(\int_{-\infty}^{a} k_1(a - \theta)p(\theta|y)d\theta + \int_{a}^{\infty} k_0(\theta - a)p(\theta|y)d\theta\right) \\
&= k_1\int_{-\infty}^{a} p(\theta|y)d\theta - k_0\int_{a}^{\infty} p(\theta|y)d\theta \\
&= (k_1 + k_0)\int_{-\infty}^{a} p(\theta|y)d\theta - k_0.
\end{aligned}$$

4.7. Denote the posterior mean by $m(y) = \mathrm{E}(\theta|y)$ and consider $m(y)$ as an estimator of θ. Unbiasedness implies that $\mathrm{E}(m(y)|\theta) = \theta$. In this case, the marginal expectation of $\theta m(y)$ is $E(\theta m(y)) = \mathrm{E}[\mathrm{E}(\theta m(y)|\theta)] = \mathrm{E}[\theta^2]$. But we can also write $\mathrm{E}(\theta m(y)) = \mathrm{E}[\mathrm{E}(\theta m(y)|y)] = \mathrm{E}[m(y)^2]$. It follows that $\mathrm{E}[(m(y) - \theta)^2] = 0$. This can only hold in degenerate problems for which $m(y) = \theta$ with probability 1.

4.9. Our goal is to show that, for sufficiently large sigma, the "Bayes estimate" (the posterior mean of θ based on the prior density $p(\theta) = 1$ in $[0, 1]$) has lower mean squared error than the maximum likelihood estimate, for any value of $\theta \in [0, 1]$.

The maximum likelihood estimate, restricted to the interval $[0, 1]$, takes value 0 with probability $\Phi(-c/\sigma)$ and takes value 1 with probability $1-\Phi((1-c)/\sigma)$; these are just the probabilities (given σ and θ) that y is less than 0 or greater than 1, respectively. For very large σ, these probabilities both approach $\Phi(0) = 1/2$. Thus, for very large σ, the mean squared error of the maximum likelihood estimate is approximately $1/2[(1 - \theta)^2 + \theta^2] = 1/2 - \theta + \theta^2$. (We use the notation Φ for the unit normal cumulative distribution function.)

On the other hand, $p(\theta|y)$ is proportional to the density function $\mathrm{N}(\theta|y, \sigma^2)$ for $\theta \in [0, 1]$. The posterior mean of θ, $\mathrm{E}(\theta|y)$, is $\int_0^1 \theta \mathrm{N}(\theta|y, \sigma^2)d\theta / \int_0^1 \mathrm{N}(\theta|y, \sigma)d\theta$. For very large σ, $\mathrm{N}(\theta|y, \sigma^2)$ is approximately constant over small ranges of θ (e.g., over $[0, 1]$). So the Bayes estimate is close to $\int_0^1 \theta d\theta = 1/2$. (This works because, since the true value of θ is assumed to lie in $[0, 1]$, the observed y will almost certainly lie within a few standard deviations of $1/2$.) Hence, for large σ, the mean squared error of the posterior mean is about $(\theta - 1/2)^2 = 1/4 - \theta + \theta^2$.

The difference in (asymptotic, as $\sigma \to \infty$) mean squared errors is independent of the true value of θ. Also, for large σ the maximum likelihood estimate generally chooses 0 or 1, each with probability almost $1/2$, whereas the Bayes estimate chooses $1/2$ with probability almost 1.

4.11. This is a tricky problem.

For this to work out, given the form of $\hat{\theta}$, the prior distribution should be a mixture of a spike at $\theta = 0$ and a flat prior distribution for $\theta \neq 0$. It's easy to get confused with degenerate and noninformative prior distributions, so let's write it as

$$p(\theta) = \lambda \mathrm{N}(\theta|0, \tau_1^2) + (1 - \lambda)\mathrm{N}(\theta|0, \tau_2^2),$$

work through the algebra, and then take the limit $\tau_1 \to 0$ (so that $\hat{\theta} = 0$ is a possibility) and $\tau_2 \to \infty$ (so that $\hat{\theta} = \overline{y}$ is a possibility).

We'll have to figure out the appropriate limit for λ by examining the posterior distribution:

$$\begin{aligned}
p(\theta|y) &\propto p(\theta)\mathrm{N}(\overline{y}|\theta, \sigma^2/n) \\
&\propto \lambda \mathrm{N}(\theta|0, \tau_1^2)\mathrm{N}(\overline{y}|\theta, \sigma^2/n) + (1 - \lambda)\mathrm{N}(\theta|0, \tau_2^2)\mathrm{N}(\overline{y}|\theta, \sigma^2/n) \\
&\propto \lambda \mathrm{N}(\overline{y}|0, \tau_1^2 + \sigma^2/n)\, \mathrm{N}\left(\theta \left| \frac{\frac{n}{\sigma^2}\overline{y}^2}{\frac{1}{\tau_1^2} + \frac{n}{\sigma^2}}, \frac{1}{\frac{1}{\tau_1^2} + \frac{n}{\sigma^2}}\right.\right) + \\
&\quad (1 - \lambda)\mathrm{N}(\overline{y}|0, \tau_2^2 + \sigma^2/n)\, \mathrm{N}\left(\theta \left| \frac{\frac{n}{\sigma^2}\overline{y}^2}{\frac{1}{\tau_2^2} + \frac{n}{\sigma^2}}, \frac{1}{\frac{1}{\tau_2^2} + \frac{n}{\sigma^2}}\right.\right)
\end{aligned}$$

In this last step, what we have done is, for each term, we have replaced the factorization $p(\theta)p(y|\theta)$ by the factorization $p(y)p(\theta|y)$, which allows us to separate out the factors that depend on θ.

The result is a mixture of two normal densities in θ. In the limit $\tau_1 \to 0$ and $\tau_2 \to \infty$, this is

$$p(\theta|y) = \lambda \mathrm{N}(\overline{y}|0, \sigma^2/n)\mathrm{N}(\theta|0, \tau_1^2) + (1 - \lambda)\mathrm{N}(\overline{y}|0, \tau_2^2)\mathrm{N}(\theta|\overline{y}, \sigma^2/n).$$

The estimate $\hat{\theta}$ cannot be the posterior *mean* (that would not make sense, since, as defined, $\hat{\theta}$ is a discontinuous function of y). Since the two normal densities have much different variances, it

would also not make sense to use a posterior *mode* estimate. A more reasonable estimate (if a point estimate must be used) is the mode of the hump of the posterior distribution that has *greater mass*. That is,

$$\begin{aligned} \text{Set } \hat{\theta} = 0 \text{ if:} \qquad \lambda \mathrm{N}(\overline{y}|0, \sigma^2/n) &> (1-\lambda)\mathrm{N}(\overline{y}|0, \tau_2^2) \\ \lambda \frac{1}{\sqrt{2\pi/n}\sigma} \exp\left(-\frac{1}{2}\frac{n}{\sigma^2}y^2\right) &> (1-\lambda)\frac{1}{\sqrt{2\pi}\tau_2}, \end{aligned} \tag{6}$$

and set $\hat{\theta} = \overline{y}$ otherwise.

For condition (6) to be equivalent to "if $\overline{y} < 1.96\sigma/\sqrt{n}$," as specified, we must have

$$\begin{aligned} 0.146\lambda \frac{1}{\sqrt{2\pi/n}\sigma} &= (1-\lambda)\frac{1}{\sqrt{2\pi}\tau_2} \\ \frac{\lambda}{1-\lambda} &= \frac{\sigma}{0.146\sqrt{n}\tau_2}. \end{aligned}$$

Since we are considering the limit $\tau_2 \to \infty$, this means that $\lambda \to 0$. That would be acceptable (a kind of improper prior distribution), but the more serious problem here is that the limiting value for $\lambda\tau_2$ depends on n, and thus the prior distribution for θ depends on n. A prior distribution cannot depend on the data, so there is *no* prior distribution for which the given estimate $\hat{\theta}$ is a reasonable posterior summary.

4.13. There are many possible answers here; for example:

(a) With a weak prior distribution, why bother with Bayesian inference? Because it is a direct way to get inferences for quantities of interest. For instance, in the examples of Sections 3.5 and 3.7, the Bayesian inference is straightforward. Why *not* use Bayesian inference in these examples? Is there a method that gives better results?

(b) With a strong prior distribution, why collect new data? This question misses the point, because any strong prior distribution must come from lots of previous data (as in the example of the sex ratio of births in Chapter 2) or a theoretical model, or some combination of the two. If enough previous data have been collected to form a precise inference, then new data may be collected in order to see whether the model generalizes to other situations. For example, in the placenta previa example of Section 2.4, we cannot really use the data from all previous births in Germany to construct an extremely strong prior distribution—we have to be open to the possibility that the new data will have a different rate of female births.

(c) With something in between, does Bayesianism "duck the issue"? The "issue" here is: what should the relative weights be for the prior and data? Bayesian inference sets these weights for different cases based on the availability of data, as is explained for the cancer rate example in Section 2.7.

The issue is further clarified in Chapter 5. In each of the three examples in Chapter 5, there are multiple experiments, each run under slightly different conditions and with a relatively small sample size. Using an informative—but not extremely strong—prior distribution allows us to get better inferences for the result of each individual experiment. The Bayesian method does not "duck the issue" because the relative weights given to the prior distribution and the data are determined by the data themselves.

5.3a. The following results are based on a different set of simulations than contained in Section 5.5. Two noteworthy differences: the support of τ for simulation purposes was increased to $[0, 40]$,

and the number of draws from the posterior distribution is 1000. Based on the 1000 posterior simulations, we obtain

		Pr(better than school)							
School	Pr(best)	A	B	C	D	E	F	G	H
A	0.25	–	0.64	0.67	0.66	0.73	0.69	0.53	0.61
B	0.10	0.36	–	0.55	0.53	0.62	0.61	0.37	0.49
C	0.10	0.33	0.45	–	0.46	0.58	0.53	0.36	0.45
D	0.09	0.34	0.47	0.54	–	0.61	0.58	0.37	0.47
E	0.05	0.27	0.38	0.42	0.39	–	0.48	0.28	0.38
F	0.08	0.31	0.39	0.47	0.42	0.52	–	0.31	0.40
G	0.21	0.47	0.63	0.64	0.63	0.72	0.69	–	0.60
H	0.12	0.39	0.51	0.55	0.53	0.62	0.60	0.40	–

5.3b. In the model with τ set to ∞, the school effects θ_j are independent in their posterior distribution with $\theta_j|y \sim N(y_j, \sigma_j^2)$. It follows that $\Pr(\theta_i > \theta_j|y) = \Phi((y_i - y_j)/\sqrt{\sigma_i^2 + \sigma_j^2})$. The probability that θ_i is the largest of the school effects can be expressed as a single integral given below.

$$\Pr(\theta_i \text{ is the largest}) = \int_{-\infty}^{\infty} \prod_{j \neq i} \Phi\left(\frac{\theta_i - y_j}{\sigma_j}\right) \phi(\theta_i|y_i, \sigma_i) d\theta_i$$

This integral can be evaluated numerically (results given below) or estimated by simulating school effects from their independent normal posterior distributions. The results are provided below.

		Pr(better than school)							
School	Pr(best)	A	B	C	D	E	F	G	H
A	0.556	–	0.87	0.92	0.88	0.95	0.93	0.72	0.76
B	0.034	0.13	–	0.71	0.53	0.73	0.68	0.24	0.42
C	0.028	0.08	0.29	–	0.31	0.46	0.43	0.14	0.27
D	0.034	0.12	0.47	0.69	–	0.70	0.65	0.23	0.40
E	0.004	0.05	0.27	0.54	0.30	–	0.47	0.09	0.26
F	0.013	0.07	0.32	0.57	0.35	0.53	–	0.13	0.29
G	0.170	0.28	0.76	0.86	0.77	0.91	0.87	–	0.61
H	0.162	0.24	0.58	0.73	0.60	0.74	0.71	0.39	–

5.3c. The model with τ set to ∞ has more extreme probabilities. For example, in the first column, the probability that School A is the best increases from 0.25 to 0.56. It is also true in the pairwise comparisons. For example, the probability that School A's program is better than School E under the full hierarchical model is 0.73, whereas it is 0.95 under the $\tau = \infty$ model. The more conservative answer under the full hierarchical model reflects the evidence in the data that the coaching programs appear fairly similar in effectiveness. Also, noteworthy is that the preferred school in a pair can change, so that School E is better than School C when τ is set to ∞, whereas School C is better than School E when averaging over the posterior distribution of τ. This effect occurs only because the standard errors, σ_j, differ.

5.3d. If $\tau = 0$ then all of the school effects are the same. Thus no school is better or worse than any other.

5.4a. Yes, they are exchangeable. The joint distribution is

$$p(\theta_1, \ldots, \theta_{2J}) = \binom{2J}{J}^{-1} \sum_p \left(\prod_{j=1}^{J} \mathrm{N}(\theta_{p(j)}|1, 1) \prod_{j=J+1}^{2J} \mathrm{N}(\theta_{p(j)}| - 1, 1) \right), \tag{7}$$

where the sum is over all permutations p of $(1, \ldots, 2J)$. The density (7) is obviously invariant to permutations of the indexes $(1, \ldots, 2J)$.

5.4b. Pick any i, j. The covariance of θ_i, θ_j is *negative*. You can see this because if θ_i is large, then it probably comes from the $\mathrm{N}(1,1)$ distribution, which means that it is more likely than not that θ_j comes from the $\mathrm{N}(-1,1)$ distribution (because we know that exactly half of the $2J$ parameters come from each of the two distributions), which means that θ_j will probably be negative. Conversely, if θ_i is negative, then θ_j is most likely positive.

Then, by Exercise 5.5, $p(\theta_1, \ldots, \theta_{2J})$ cannot be written as a mixture of iid components.

The above argument can be made formal and rigorous by defining $\phi_1, \ldots, \phi_{2J}$, where half of the ϕ_j's are 1 and half are -1, and then setting $\theta_j|\phi_j \sim \mathrm{N}(\phi_j, 1)$. It's easy to show first that $\mathrm{cov}(\phi_i, \phi_j) < 0$, and then that $\mathrm{cov}(\theta_i, \theta_j) < 0$ also.

5.4c. In the limit as $J \to \infty$, the negative correlation between θ_i and θ_j approaches zero, and the joint distribution approaches iid. To put it another way, as $J \to \infty$, the distinction disappears between (1) independently assigning each θ_j to one of two groups, and (2) picking exactly half of the θ_j's for each group.

5.5. Let $\mu(\phi) = \mathrm{E}(\theta_j|\phi)$. From (1.9) on page 21 (also see Exercise 1.2),

$$\begin{aligned}
\mathrm{cov}(\theta_i, \theta_j) &= \mathrm{E}(\mathrm{cov}(\theta_i, \theta_j|\phi)) + \mathrm{cov}(\mathrm{E}(\theta_i|\phi), \mathrm{E}(\theta_j|\phi)) \\
&= 0 + \mathrm{cov}(\mu(\phi), \mu(\phi)) \\
&= \mathrm{var}(\mu(\phi)) \\
&\geq 0.
\end{aligned}$$

5.7a. We want to find $\mathrm{E}(y)$ and $\mathrm{var}(y)$, where $y|\theta \sim \mathrm{Poisson}(\theta)$ and $\theta \sim \mathrm{Gamma}(\alpha, \beta)$. From (2.7), $\mathrm{E}(y) = \mathrm{E}(\mathrm{E}(y|\theta))$; from properties of the Poisson and Gamma distributions, we have $\mathrm{E}(y|\theta) = \theta$ so that $\mathrm{E}(\mathrm{E}(y|\theta)) = \mathrm{E}(\theta) = \alpha/\beta$.

Similarly, by formula (2.8),

$$\mathrm{var}(y) = \mathrm{E}(\mathrm{var}(y|\theta)) + \mathrm{var}(\mathrm{E}(y|\theta)) = \mathrm{E}(\theta) + \mathrm{var}(\theta) = \alpha/\beta + \alpha/\beta^2 = \alpha/\beta^2(\beta + 1).$$

5.7b. This part is a little bit trickier because we have to think about what to condition on in applying formulas (2.7) and (2.8). Some reflection (and, perhaps, a glance at formula (3.3)) will lead to the choice of conditioning on σ^2.

Throughout these computations, n, s, and $\bar{y}$ are essentially treated like constants.

From (2.7) and (3.3),

$$\mathrm{E}(\sqrt{n}(\mu - \bar{y})/s|y) = \mathrm{E}(\mathrm{E}(\sqrt{n}(\mu - \bar{y})/s|\sigma, y)|y) = \mathrm{E}((\sqrt{n}/s)\mathrm{E}(\mu - \bar{y}|\sigma, y)|y) = \mathrm{E}((\sqrt{n}/s) \cdot 0|y) = \mathrm{E}(0|y) = 0.$$

Obviously we must have $n > 1$ for s to be defined. But in order for the expectation to exist, we must have $n > 2$. You can deduce this from inspecting the formula for $p(\mu|y)$ at the top of page 69: the exponent must be less than negative one for the quantity to be integrable, which is true if and only if $n > 2$.

Similarly, we can compute from (2.8), (3.3), and (3.5) that

$$\begin{aligned}
\mathrm{var}(\sqrt{n}(\mu - \bar{y})/s|y) &= \mathrm{var}(\mathrm{E}(\sqrt{n}(\mu - \bar{y})/s|\sigma, y)|y) + \mathrm{E}(\mathrm{var}(\sqrt{n}(\mu - \bar{y})/s|\sigma, y)|y) \\
&= \mathrm{var}(0|y) + \mathrm{E}((n/s^2)\mathrm{var}(\mu|\sigma, y)|y) \\
&= \mathrm{E}((n/s^2)\sigma^2/n|y) \\
&= \mathrm{E}(\sigma^2|y)/s^2 = \frac{n-1}{n-3}.
\end{aligned}$$

For this to work, we need $n > 3$.

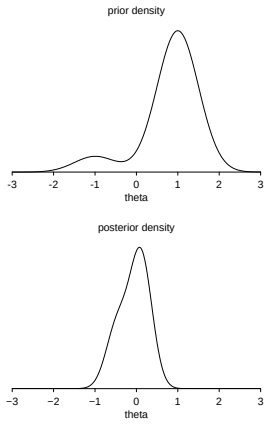
5.8. Let $p_m(\theta|y)$ denote the posterior density of θ corresponding to the prior density $p_m(\theta)$. That is, for each m, $p_m(\theta|y) = p_m(\theta)p(y|\theta)/p_m(y)$, where $p_m(y)$ is the prior predictive density.

If $p(\theta) = \sum_m \lambda_m p_m(\theta)$, then the posterior density of θ is proportional to $\sum_m \lambda_m p_m(\theta)p(y|\theta) = \sum_m \lambda_m p_m(y)p_m(\theta|y)$: this is a mixture of the posterior densities $p_m(\theta|y)$ with weights proportional to $\lambda_m p_m(y)$. Since each $p_m(\theta)$ is conjugate for the model for y given θ, the preceding computation demonstrates that the class of finite mixture prior densities is also conjugate.

Consider an example: $p_1(\theta) \sim \mathrm{N}(1, 0.5^2)$, $p_2(\theta) \sim \mathrm{N}(-1, 0.5^2)$, and suppose that $\lambda_1 = 0.9$ amd $\lambda_2 = 0.1$. (The exact choices of λ_1 and λ_2 are not important. What is important is how and why the posterior mixture weights are different than the prior weights.) We know that $p_1(\theta|y) \sim \mathrm{N}(1.5/14, 1/14)$ and that $p_2(\theta|y) \sim \mathrm{N}(-6.5/14, 1/14)$; see, e.g., formulas (2.9) and (2.10). We also know from the last paragraph that $p(\theta|y)$ will be a weighted sum of these conditional posterior densities with weights $\lambda_m p_m(y)/\sum_k \lambda_k p_k(y)$ for $m = 1, 2$.

You can compute $p_1(y)$ and $p_2(y)$, using convenient properties of normal distributions: $p_1(y) = \mathrm{N}(-0.25|1, 0.5^2 + 1/10) = 0.072$, and $p_2(y) = \mathrm{N}(-0.25| - 1, 0.5^2 + 1/10) = 0.302$.

So the weights for $p_1(\theta|y)$ and $p_2(\theta|y)$ are not 0.9 and 0.1 but are, rather, $\frac{0.9 \cdot 0.072}{0.9 \cdot 0.072 + 0.1 \cdot 0.302} = 0.68$ and $\frac{0.1 \cdot 0.302}{0.9 \cdot 0.072 + 0.1 \cdot 0.302} = 0.32$.




```
theta <- seq(-3,3,.01)
prior <- c (0.9, 0.1)
dens <- prior[1]*dnorm(theta,1,0.5) +
  prior[2]*dnorm(theta,-1,0.5)
plot (theta, dens, ylim=c(0,1.1*max(dens)),
  type="l", xlab="theta", ylab="", xaxs="i",
  yaxs="i", yaxt="n", bty="n", cex=2)
mtext ("prior density", cex=2, 3)

marg <- dnorm(-.25,c(1,-1),sqrt(c(0.5,0.5)^2+1/10))
posterior <- prior*marg/sum(prior*marg)

dens <- posterior[1]*dnorm(theta,1.5/14,sqrt(1/14)) +
  posterior[2]*dnorm(theta,-6.5/14,sqrt(1/14))
plot (theta, dens, ylim=c(0,1.1*max(dens)),
  type="l", xlab="theta", ylab="", xaxs="i",
  yaxs="i", yaxt="n", bty="n", cex=2)
mtext ("posterior density", cex=2, 3)
```

5.9a. Consider the limit $(\alpha + \beta) \to \infty$ with α/β fixed at any nonzero value. The likelihood (see equation (5.8)) is

$$\begin{aligned} p(y|\alpha,\beta) &\propto \prod_{j=1}^{J} \frac{[\alpha \cdots (\alpha + y_j - 1)][\beta \cdots (\beta + n_j - y_j - 1)]}{(\alpha+\beta) \cdots (\alpha+\beta+n_j-1)} \qquad (8) \\ &\approx \prod_{j=1}^{J} \frac{\alpha^{y_j}\beta^{n_j - y_j}}{(\alpha+\beta)^{n_j}} \\ &= \prod_{j=1}^{J} \left(\frac{\alpha}{\alpha+\beta}\right)^{y_j} \left(\frac{\beta}{\alpha+\beta}\right)^{n_j - y_j}, \end{aligned}$$

which is a constant (if we are considering y, n, and α/β to be fixed), so the prior density determines whether the posterior density has a finite integral in this limit. A uniform prior density on $\log(\alpha+\beta)$ has an infinite integral in this limit, and so the posterior density does also in this case.

5.9b. The Jacobian of the transformation is

$$\left| \begin{array}{cc} \frac{\beta}{(\alpha+\beta)^2} & -\frac{\alpha}{(\alpha+\beta)^2} \\ -\frac{1}{2}(\alpha+\beta)^{-3/2} & -\frac{1}{2}(\alpha+\beta)^{-3/2} \end{array} \right| = \text{constant} \cdot (\alpha+\beta)^{-5/2}.$$

5.9c. There are 4 limits to consider:

1. $\alpha \to 0$ with $\alpha + \beta$ fixed
2. $\beta \to 0$ with $\alpha + \beta$ fixed
3. $\alpha + \beta \to 0$ with α/β fixed
4. $\alpha + \beta \to \infty$ with α/β fixed

As in Exercise 5.9a, we work with expression (5.8). We have to show that the integral is finite in each of these limits.

Let J_0 be the number of experiments with $y_j > 0$, J_1 be the number of experiments with $y_j < n_j$, and J_{01} be the number of experiments with $0 < y_j < n_j$.

1. For $\alpha \to 0$, proving convergence is trivial. All the factors in the likelihood (8) go to constants except α^{J_0}, so the likelihood goes to 0 (if $J_0 > 0$) or a constant (if $J_0 = 0$). The prior distribution is a constant as α goes to 0. So whether $J_0 = 0$ or $J_0 > 0$, the integral of the posterior density in this limit is finite.

2. For $\beta \to 0$, same proof.

3. First transform to $(\frac{\alpha}{\alpha+\beta}, \alpha+\beta)$, so we only have to worry about the limit in one dimension. Multiplying the prior distribution (5.9) by the Jacobian yields

$$p\left(\frac{\alpha}{\alpha+\beta}, \alpha+\beta\right) \propto (\alpha+\beta)^{-3/2}$$

   For $\alpha + \beta \to 0$, the likelihood looks like

$$p(y|\alpha,\beta) \propto (\alpha+\beta)^{-J_{01}},$$

   ignoring all factors such as $\Gamma(\alpha+\beta+n_j)$ and $\frac{\alpha}{\alpha+\beta}$ that are constant in this limit. So the posterior density in this parameterization is

$$p\left(\frac{\alpha}{\alpha+\beta}, \alpha+\beta \,\middle|\, y\right) \propto (\alpha+\beta)^{-3/2}(\alpha+\beta)^{J_{01}}. \tag{9}$$

   The function x^c has a finite integral as $x \to 0$ if $c > -1$, so (9) has a finite integral if $J_{01} > \frac{1}{2}$.

   Note: the statistical reasoning here is that if $J_{01} = 0$ (so that all the data are of "0 successes out of n_j" or "n_j successes out of n_j"), then it is still possible that $\alpha = \beta = 0$ (corresponding to all the θ_j's being 0 or 1), and we have to worry about the infinite integral of the improper prior distribution in the limit of $(\alpha+\beta) \to 0$. If $J_{01} > 0$, the likelihood gives the information that $(\alpha+\beta) = 0$ is not possible. In any case, values of α and β near 0 do not make much sense in the context of the problem (modeling rat tumor rates), and it might make sense to just constrain $\alpha \geq 1$ and $\beta \geq 1$.

4. For $\alpha + \beta \to \infty$, the likelihood is constant, and so we just need to show that the prior density has a finite integral (see the solution to Exercise 5.9a). As above,

$$p\left(\frac{\alpha}{\alpha+\beta}, \alpha+\beta\right) \propto (\alpha+\beta)^{-3/2},$$

   which indeed has a finite integral as $(\alpha+\beta) \to \infty$.

5.10. We first note that, since $p(\mu|\tau, y)$ and $p(\theta|\mu, \tau, y)$ have proper distributions, the joint posterior density $p(\theta, \mu, \tau|y)$ is proper if and only if the marginal posterior density $p(\tau|y)$ from (5.21) is proper—that is, has a finite integral for τ from 0 to ∞.

5.10a. Everything multiplying $p(\tau)$ in (5.21) approaches a nonzero constant limit as τ tends to zero; call that limit $C(y)$. Thus the behavior of the posterior density near $\tau = 0$ is determined by the prior density. The function $p(\tau) \propto 1/\tau$ is not integrable for any small interval including $\tau = 0$, and so it leads to a nonintegrable posterior density. (We can make this more formal: for any $\delta > 0$ we can identify an interval including zero on which $p(\tau|y) \geq p(\tau)(C - \delta)$.)

5.10b. The argument from 5.10a shows that if $p(\tau) \propto 1$ then the posterior density is integrable near zero. We need to examine the behavior as $\tau \to \infty$ and find an upper bound that is integrable. The exponential term is clearly less than or equal to 1. We can rewrite the remaining terms as $(\sum_{j=1}^{J}[\prod_{k\neq j}(\sigma_k^2 + \tau^2)])^{-1/2}$. For $\tau > 1$ we make this quantity bigger by dropping all of the σ^2 to yield $(J\tau^{2(J-1)})^{-1/2}$. An upper bound on $p(\tau|y)$ for τ large is $p(\tau)J^{-1/2}/\tau^{J-1}$. When $p(\tau) \propto 1$, this upper bound is integrable if $J > 2$, and so $p(\tau|y)$ is integrable if $J > 2$.

5.10c. There are several reasonable options here. One approach is to abandon the hierarchical model and just fit the two schools with independent noninformative prior distributions (as in Exercise 3.3 but with the variances known).

Another approach would be to continue with the uniform prior distribution on τ and try to analytically work through the difficulties with the resulting improper posterior distribution. In this case, the posterior distribution has all its mass near $\tau = \infty$, meaning that no shrinkage will occur in the analysis (see (5.17) and (5.20)). This analysis is in fact identical to the first approach of analyzing the schools independently.

If the analysis based on noninformative prior distributions is not precise enough, it might be worthwhile assigning an informative prior distribution to τ (or perhaps to (μ, τ)) based on outside knowledge such as analyses of earlier coaching experiments in other schools. However, one would have to be careful here: with data on only two schools, inferences would be sensitive to prior assumptions that would be hard to check from the data at hand.

5.11a. $p(\theta, \mu, \tau|y)$ is proportional to $p(\theta, \mu, \tau)p(y|\theta, \mu, \tau)$, where we may note that the latter density is really just $p(y|\theta)$. Since $p(\theta, \mu, \tau) = p(\theta|\mu, \tau)p(\mu, \tau)$,

$$p(\theta, \mu, \tau|y) \propto p(\mu, \tau) \prod_{j=1}^{J} \left[\theta_j^{-1}(1-\theta_j)^{-1}\tau^{-1} \exp(-\frac{1}{2}(\text{logit}(\theta_j) - \mu)^2/\tau^2)\right] \prod_{j=1}^{J} \left[\theta_j^{y_j}(1-\theta_j)^{n_j - y_j}\right].$$

(The factor $\theta_j^{-1}(1-\theta_j)^{-1}$ is $d(\text{logit}(\theta_j))/d\theta_j$; we need this since we want to have $p(\theta|\mu, \tau)$ rather than $p(\text{logit}(\theta)|\mu, \tau)$.)

5.11b. Even though we can look at each of the J integrals individually—the integrand separates into independent factors—there is no obvious analytic technique which permits evaluation of these integrals. In particular, we cannot recognize a multiple of a familiar density function inside the integrals. One might try to simplify matters with a substitution like $u_j = \text{logit}(\theta_j)$ or $v_j = \theta_j/(1-\theta_j)$, but neither substitution turns out to be helpful.

5.11c. In order for expression (5.5) to be useful, we would have to know $p(\theta|\mu, \tau, y)$. Knowing it up to proportionality in θ is insufficient because our goal is to use this density to find another density that depends on μ and τ. In the rat tumor example (see equation (5.7)), the conjugacy of the beta

distribution allowed us to write down the relevant "constant" of proportionality, which appears in equation (5.8).

5.12 Following the hint we apply (2.7) and (2.8).

$$\begin{aligned}
\mathrm{E}(\theta_j|\tau,y) &= \mathrm{E}[\mathrm{E}(\theta_j|\mu,\tau,y)\,|\,\tau,y] = \mathrm{E}\left[\left.\frac{\frac{1}{\sigma_j^2}y_j+\frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2}+\frac{1}{\tau^2}}\right|\tau,y\right] = \frac{\frac{1}{\sigma_j^2}y_j+\frac{1}{\tau^2}\hat{\mu}}{\frac{1}{\sigma_j^2}+\frac{1}{\tau^2}} \\
\mathrm{var}(\theta_j|\tau,y) &= \mathrm{E}[\mathrm{var}(\theta_j|\mu,\tau,y)\,|\,\tau,y] + \mathrm{var}[\mathrm{E}(\theta_j|\mu,\tau,y)\,|\,\tau,y] = \frac{1}{\frac{1}{\sigma_j^2}+\frac{1}{\tau^2}} + \left(\frac{\frac{1}{\tau^2}}{\frac{1}{\sigma_j^2}+\frac{1}{\tau^2}}\right)^2 V_\mu,
\end{aligned}$$

where expressions for $\hat{\mu} = \mathrm{E}(\mu|\tau,y)$ and $V_\mu = \mathrm{var}(\mu|\tau,y)$ are given in (5.20) in Section 5.4.

6.1a. Under the model that assumes identical effects in all eight schools ($\tau = 0$), the posterior distribution for the common θ is $\mathrm{N}(7.7, 4.1^2)$; see page 119–120. To simulate the posterior predictive distribution of y^{rep} we first simulate $\theta \sim \mathrm{N}(7.7, 4.1^2)$ and then simulate $y_j^{\mathrm{rep}} \sim \mathrm{N}(\theta, \sigma_j^2)$ for $j = 1, \ldots, 8$. The observed order statistics would be compared to the distribution of the order statistics obtained from the posterior predictive simulations. The expected order statistics provided in the problem (from a $\mathrm{N}(8, 13^2)$ reference distribution) are approximately the mean order statistics we would find in the replications. (The mean in the replications would differ because the school standard errors vary and because the value 13 does not incorporate the variability in the common θ.)

Just by comparing the observed order statistics to the means provided, it is clear that the identical-schools model *does* fit this aspect of the data. Based on 1000 draws from the posterior predictive distribution we find that observed order statistics correspond to upper-tail probabilities ranging from 0.20 to 0.62, which further supports the conclusion.

6.1b. It is easy to imagine that we might be interested in choosing some schools for further study or research on coaching programs. Under the identical-effects model, all schools would be equally good for such follow-up studies. However, most observers would wish to continue experimenting in School A rather than School C.

The decision to reject the identical-effects model was based on the notion that the posterior distribution it implies (all θ_j's the same) does not make substantive sense.

6.5.

1. A "hypothesis of no difference" probably refers to a hypothesis that $\theta = 0$ or $\theta_1 = \theta_2$ in some model. In this case, we would just summarize our uncertainty about the parameters using the posterior distribution, rather than "testing" the hypothesis $\theta = 0$ or $\theta_1 = \theta_2$ (see, for example, Exercises 3.2–3.4). The issue is not whether θ is exactly zero, but rather what we know about θ.

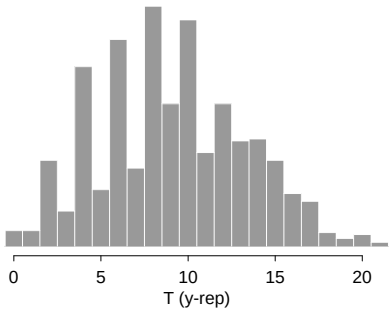
2. One can also think of a model itself as a hypothesis, in which case the model checks described in this chapter are hypothesis tests. In this case, the issue is not whether a model is exactly true, but rather what are the areas where the model (a) does not fit the data, and (b) is sensitive to questionable assumptions. The outcome of model checking is not "rejection" or "acceptance," but rather an exploration of the weak points of the model. With more data, we will be able to find more weak points. It is up to the modeler to decide whether an area in which the model does not fit the data is important or not (see page 362 for an example).

6.6a. Under the new protocol,

$$p(y, n|\theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} 1_{(y_n=0)} 1_{\left(\sum_{i=1}^{n-1}(1-y_i)=12\right)}$$

For the given data, this is just $\theta^7(1-\theta)^{13}$, which is the likelihood for these data under the "stop after 20 trials" rule. So if the prior distribution for θ is unchanged, then the posterior distribution is unchanged.

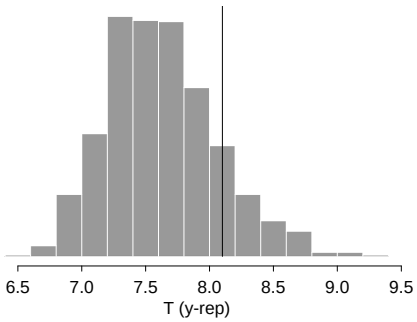
6.6b. Here is the code:




```
test <- NULL
for (i in 1:1000){
  theta <- rbeta (1,8,14)
  y.rep <- rbinom (1,1,theta)
  while (sum(y.rep==0) < 13)
    y.rep <- c(y.rep, rbinom(1,1,theta))
  n.rep <- length(y.rep)
  test <- c(test,
    sum (y.rep[2:n.rep] != y.rep[1:(n.rep-1)]))}
hist (test, xlab="T (y-rep)", yaxt="n",
  breaks=seq(-.5, max(test)+.5), cex=2)
```

The posterior predictive distribution differs from that of Figure 6.4 in two notable ways. First, there is more spread, which makes sense since n is now a random variable. Second, $T(y^{\text{rep}})$ is much more likely to be even, especially for low values of T, which is a peculiarity of this particular sequential data collection rule.

6.7a. For very large A, the posterior distribution for θ is approximately $\text{N}(5.1, 1/100)$. The posterior predictive distribution is $y^{\text{rep}}|y \sim \text{N}(5.1, (1/100)\mathbf{1}\mathbf{1}^T + I)$, where $\mathbf{1}\mathbf{1}^T$ is the 100×100 square matrix all of whose elements are 1. We compute the distribution of $T(y^{\text{rep}})$ by simulation. The posterior predictive p-value is estimated as the proportion of simulations of $T(y^{\text{rep}})$ that exceed $T(y) = 8.1$; this is 14%.




```
test <- NULL
for (i in 1:1000){
  theta <- rnorm (1,5.1,1/sqrt(100))
  y.rep <- rnorm (100,theta,1)
  test <- c(test,max(abs(y.rep)))}
postscript ("fig6.6a.ps", horizontal=TRUE)
par (mar=c(5,5,4,1)+.1)
hist (test, xlab="T (y-rep)", yaxt="n",
  nclass=20, cex=2)
lines (rep(8.1,2), c(0,1000))

print (mean(test>8.1))
```

6.7b. For very large A, the prior predictive distribution for $T(y^{\text{rep}}) = \max_i |y_i^{\text{rep}}|$ will extend from 0 out to the neighborhood of A. Thus, if A is large, the prior predictive p-value for $T(y) = 8.1$ will be approximately 1.0.

6.7c. Since the prior distribution of θ is essentially noninformative (for these data), the data almost completely determine the posterior distribution. If the data fit the model reasonably well, then it makes sense that the posterior predictive distribution is consistent with the data. On the other

hand, we would not expect a diffuse prior distribution to be consistent with any observed data set. In particular, any finite dataset will be unusually close to 0 compared to an improper distribution that extends out to ∞.

8.1a. "Randomization implies exchangeability" is not necessarily true because if you have a randomized design but covariates x are also available, and you think the covariates might be related to y, it makes sense to model (x, y) together. For example, if the cow experiment in Section 7.4 had really been conducted using a single complete randomization, it would still be a good idea to model milk production conditional on age, lactation, and initial weight. Setting up a conditional model will yield more precise inferences for the distribution of y.

"Randomization implies exchangeability" has a kernel of truth because, under randomization, it is acceptable to use an exchangeable model: no additional variables are needed for an adequate summary of the data.

8.1b. "Randomization is required for exchangeability" is not necessarily true because if you have no other information on a set of units but the measurements y, then you *must* model them exchangeably. For example, no randomization was used in many of the examples used in this book (the football games, the speed of light measurements, the selection of the 8 schools, etc.)—but with no other information available, exchangeable models are the only possibility.

"Randomization is required for exchangeability" has a kernel of truth because if data are collected nonrandomly based on additional information, and there is information available to distinguish the units that affected the data collection procedure (as in the cow experiment), then it is not appropriate to use an exchangeable model on the data y alone.

8.1c. "Randomization implies ignorability" is not necessarily true for two reasons:

1. Even under randomization, data collection can be affected by other factors. For example, what if treatments are assigned randomly, but then, in addition, all values of $y > 200$ are censored. Then the censoring should be in the model; the data collection cannot be ignored.

2. If θ (the parameters for the data model) and ϕ (the parameters for the data collection model) are not distinct, then the design is not ignorable: even if the data are collected randomly, the inclusion pattern provides information about ϕ, and thus θ (see variation 2 in Section 8.7).

"Randomization implies ignorability" has a kernel of truth because if the above two issues do not arise, then a randomized design is ignorable—because under most methods of randomization, $p(I|\phi, y)$ does not depend on y_{mis}, so the data are missing at random. Missing at random and distinct parameters imply ignorability (see page 202).

8.1d. "Randomization is required for ignorability" is not necessarily true because many deterministic data collection designs have the property thet inclusion does not depend on y_{mis} (and thus, since a deterministic design has no parameters, the design is ignorable). For example: designs in which all the data are used (for example, the football games, the speed of light experiment), designs chosen in advance (for example, laying out two agricultural treatments in an $AABBAABBAABB$ pattern), designs that have no clear rule but are based entirely on observed covariates (for example, the cow experiment), and designs based on observed outcomes (for example, sequential designs, which are ignorable conditional on the order of the units and the observed outcomes).

"Randomization is required for ignorability" has a kernel of truth because if a nonrandomized design is used, with data collection depending on the (possibly unobserved) value of y or on an unobserved variable, then the pattern of data collected is, in general, informative, and the data collection is nonignorable. For example, consider a medical experiment in which treatment A is more

likely to be given to patients that "look sick," but this measure of "looking sick" is not quantified so that it is not available for the analysis. This is a censoring-like problem and is nonignorable.

8.1e. "Ignorability implies exchangeability" is not necessarily true for the same reason as the answer to Exercise 8.1a. Even if you are allowed to use an exchangeable model on y, you might still do better by modeling (x, y), even if x is not used in the data collection.

"Ignorability implies exchangeability" has a kernel of truth because, if the data collection is ignorable given y_{obs}, so that it does not depend on other variables x, then it is not necessary to include x in the analysis: one can apply an exchangeable model to y.

8.1f. "Ignorability is required for exchangeability" is not necessarily true: consider censored data collection (for example, variation 3 in Section 8.7). The design is nonignorable, and the inclusion pattern must be included in the model, but the model is still exchangeable in the $y_{\text{obs}\,i}$'s.

"Ignorability is required for exchangeability" has a kernel of truth because if the data collection is nonignorable given y_{obs}, and additional information x is available that is informative with respect to the data collection, then x should be included in the model because $p(\theta|x, y_{\text{obs}}, I) \neq p(\theta|y_{\text{obs}})$. This is an exchangeable model on (x, y), not y.

8.2. Many possibilities: for example, in the analysis of the rat tumor experiments in Chapter 5, we are assuming that the sample sizes, n_j, provide no information about the parameters θ_j—this is like the assumption of *distinct parameters* required for ignorability. But one could imagine situations in which this is not true—for example, if more data are collected in labs where higher tumor rates are expected. This would be relevant for the parameter of interest, y_{71}, because n_{71} has the relatively low value of 14. One way the analysis could reflect this possibility is to model θ_j conditional on n_j. To do this would be awkward using the beta distribution on the θ_j's; it might be easier to set up a logistic regression model—basically, a hierarchical version of the model in Section 3.7. This would take a bit of computational and modeling effort—before doing so, it would make sense to plot the estimated θ_j's vs. the n_j's to see if there seems to be any relation between the variables.

8.7a. We can write the posterior distribution of μ, σ as,

$$\begin{aligned} \sigma^2|y^{\text{obs}} &\sim \text{Inv-}\chi^2(n-1, s^2) \\ \mu|\sigma^2, y^{\text{obs}} &\sim \text{N}(\bar{y}^{\text{obs}}, \sigma^2/n), \end{aligned}$$

where $\bar{y}^{\text{obs}}$ and s^2 are the sample mean and variance of the n data points y_i^{obs}.

The predictive distribution of the mean of the missing data is,

$$\bar{y}^{\text{mis}}|\mu, \sigma^2, y^{\text{obs}} \sim \text{N}(\mu, \sigma^2/(N-n)).$$

Averaging over μ yields,

$$\bar{y}^{\text{mis}}|\sigma^2, y^{\text{obs}} \sim \text{N}(\bar{y}^{\text{obs}}, (1/n + 1/(N-n))\sigma^2).$$

Since σ^2 has an inverse-χ^2 posterior distribution, averaging over it yields a t distribution (as on p. 76):

$$\bar{y}^{\text{mis}}|y^{\text{obs}} \sim t_{n-1}(\bar{y}^{\text{obs}}, (1/n + 1/(N-n))s^2).$$

The complete-data mean is $\bar{y} = n/N\bar{y}^{\text{obs}} + (N-n)/N\bar{y}^{\text{mis}}$, which is a linear transformation of $\bar{y}^{\text{mis}}$ (trating n, N, and $\bar{y}^{\text{obs}}$ as constants, since they are all observed) and so it has a t posterior distribution also:

$$\begin{aligned} \bar{y}|y^{\text{obs}} &\sim t_{n-1}(\bar{y}^{\text{obs}}, s^2(1/n + 1/(N-n))(N-n)^2/N^2) \\ &= t_{n-1}(\bar{y}^{\text{obs}}, s^2(N-n)/(nN)) \\ &= t_{n-1}(\bar{y}^{\text{obs}}, s^2(1/n - 1/N)). \end{aligned}$$

8.7b. The t distribution with given location and scale parameters tends to the corresponding normal distribution as the degrees of freedom go to infinity:

Since $(1 + z/w)^w \to \exp(z)$ as $w \to \infty$, the expression $(1/\sigma)(1 + (1/\nu)((\theta - \mu)/\sigma)^2)^{-(\nu+1)/2}$ approaches $(1/\sigma)\exp(-\frac{1}{2}((\theta - \mu)/\sigma)^2)$ as $\nu \to \infty$.

8.15a. The assignment for the current unit depends on previous treatment assignments. Thus, the time sequence of the assignments must be recorded. This can be expressed as a covariate as t_i for each observation i, where $t_i = 1, 2, \ldots$. A time covariate giving the actual times (not just the time order) of the observations would be acceptable also (that is, the design would be ignorable given this covariate) and, in practice, might make it easier to model the outcomes, given t.

8.15b. If the outcome is $y = (y_1, \ldots, y_n)$, we would model it as $(y|x, \theta)$, along with a prior distribution, $p(\theta|x)$, where θ describes all the parameters in the model and x are the covariates, which must include the treatment assignment and the time variable (and can also include any other covariates of interest, such as sex, age, medical history, etc.).

8.15c. Under this design, inferences are sensitive to dependence of y on t (conditional on θ and the other variables in x). If such dependence is large, it needs to be modeled, or the inferences will not be appropriate. Under this design, t plays a role similar to a blocking variable or the pre-treatment covariates in the milk production example in Section 8.4. For example, suppose that $\mathrm{E}(y|x, \theta)$ has a linear trend in t but that this dependence is not modeled (that is, suppose that a model is fit ignoring t). Then posterior means of treatment effects will tend to be reasonable, but posterior standard deviations will be too large, because this design yields treatment assignments that, compared to complete randomization, tend to be more balanced for t.

8.15d. Under the proposed design and alternative (ii), you must include the times t as one of the covariates in your model; under alternative (i), it is not necessary to include t. This is an advantage of alternative (i), since it reduces the data collection requirements. On the other hand, it could be considered an advantage of the proposed design and alternative (ii), since it forces you to use this information, which may very well lead to more relevant inferences (if, for example, treatment efficacy increases with t).

Suppose that, under all three designs, you will model the data using the same set of covariates x (that include treatment assignment and t). Of the three designs, alternative (ii) is most balanced for t, alternative (i) is least balanced (on average), and the proposed design falls in between. So, assuming smooth changes in the dependence of y on t (conditional on x and θ), we would expect data collected under alternative (ii) to yield the most precise inferences and under alternative (i) the least precise. If for some reason there was an alternating trend (for example, if the patients are assigned to two doctors, and the doctors tend to alternate), then alternative (ii) could become the least precise.

Another potential difficulty in alternative (ii) is posterior predictive model checking (or, in classical terms, hypothesis testing). Since this alternative has only two possible treatment vectors, it is basically impossible to check the fit of the model under replications of the treatment under the same patients. (It would be, however, possible to test the fit of the model as it would apply to additional patients.)

10.4a. Suppose that θ is drawn from the density proportional to $g(\theta)$, and U is a random Uniform$(0, 1)$ draw. Then we can express the cumulative distribution function of draws accepted by rejection sam-

pling as

$$\begin{aligned}\Pr(\theta \le \theta^* | \theta \text{ is accepted}) &= \frac{\Pr(\theta \le \theta^* \text{ and } U \le \frac{p(\theta|y)}{Mg(\theta)})}{\Pr(U \le \frac{p(\theta|y)}{Mg(\theta)})} \\ &= \frac{\int_{-\infty}^{\theta^*} \int_0^{p(\theta|y)/(Mg(\theta))} g(\theta) du d\theta}{\int_{-\infty}^{\infty} \int_0^{p(\theta|y)/(Mg(\theta))} g(\theta) du d\theta} \\ &= \frac{\frac{1}{M} \int_{-\infty}^{\theta^*} p(\theta|y) d\theta}{\frac{1}{M}},\end{aligned}$$

which is the cumulative density function for $p(\theta|y)$.

A similar argument works in higher dimensions.

10.4b. The above proof requires that (i) $g(\theta) > 0$ wherever $p(\theta) > 0$ (otherwise the final integral will not cover the support of p) and that (ii) $p(\theta)/Mg(\theta) \le 1$ always (otherwise the integral over u will have the range of $[0, 1]$ rather than $[0, p/Mg]$, and the transition to the last line of the equation will not be valid). If p/g is unbounded, than one or both of (i) and (ii) will be violated.

11.1. As described on p. 279: consider any two points θ_a and θ_b, at iteration t labeled so that $p(\theta_b|y)J_t(\theta_a|\theta_b) \ge p(\theta_a|y)J_t(\theta_b|\theta_a)$. To show that the posterior distribution is a stationary distribution, suppose that θ^{t-1} is a draw from the posterior distribution. Then the unconditional probability density of a transition from θ_a to θ_b is

$$p(\theta^{t-1} = \theta_a, \theta^t = \theta_b) = p(\theta_a|y)J_t(\theta_b|\theta_a),$$

where the acceptance probability is 1 because of our labeling of a and b so that the ratio of importance ratios is at least 1. The unconditional probability density of a transition from θ_b to θ_a is

$$p(\theta^t = \theta_a, \theta^{t-1} = \theta_b) = p(\theta_b|y)J_t(\theta_a|\theta_b) \left(\frac{p(\theta_a|y)/J_t(\theta_a|\theta_b)}{p(\theta_b|y)/J_t(\theta_b|\theta_a)} \right),$$

which is the same as the unconditional probability in the other direction. Since their joint distribution is symmetric, θ^t and θ^{t-1} have the same marginal distributions, and so $p(\theta|y)$ is a stationary distribution of the Markov chain.

As with the Metropolis algorithm, the stationary distribution is unique if the Markov chain is irreducible, aperiodic, and not transient.

13.7. We find the conditional posterior distribution of $\gamma = \log \sigma$ by transforming the inverse-χ^2.

$$p(\gamma|\theta, \mu, \tau, y) \propto 2e^{2\gamma}(e^{-2\gamma})^{n/2+1} e^{-n\hat{\sigma}^2/2e^{2\gamma}}.$$

Maximizing the log of this density (which is equivalent to maximizing the density) yields $\hat{\gamma} = \log \hat{\sigma}$. The argument for $\log \tau$ is similar.

13.8. The joint posterior density is given by the formula on page 288. If $\theta_j = \mu$ for all j, then the expression simplifies to

$$\begin{aligned}p(\theta, \mu, \log \sigma, \log \tau|y) &\propto \tau(\sqrt{2\pi}\tau)^{-J} \prod_{j=1}^{J} \prod_{i=1}^{n_j} \mathrm{N}(y_{ij}|\theta_j, \sigma^2) \\ &= (2\pi)^{-J/2} \tau^{-(J-1)} \prod_{j=1}^{J} \prod_{i=1}^{n_j} \mathrm{N}(y_{ij}|\theta_j, \sigma^2).\end{aligned}$$

This is just $\tau^{-(J-1)}$ multiplied by a function of the other parameters and the data. In this case, $J = 4$, so as $\tau \to 0$ with $\theta_j = \mu$ for all j, the joint posterior density is proportional to τ^{-3}, which approaches infinity.

14.1a. We fit a linear regression with a uniform prior density on $(\beta, \log \sigma)$. We include coefficients for the basement indicator and for three county indicators, and no constant term. (If data on more counties were available, or if we were interested in forecasting for additional counties, then we would fit a hierarchical regression at the county level, as discussed in Chapter 15.)

	Posterior quantiles		
Quantity of interest	25%	50%	75%
geometric mean for Blue Earth (no basement), $\exp(\beta_2)$	4.1	5.0	6.5
geometric mean for Blue Earth County (basement), $\exp(\beta_1 + \beta_2)$	6.1	7.1	8.2
geometric mean for Clay County (no basement), $\exp(\beta_3)$	3.8	4.7	5.8
geometric mean for Clay County (basement), $\exp(\beta_1 + \beta_3)$	5.6	6.5	7.6
geometric mean for Goodhue County (no basement), $\exp(\beta_4)$	3.9	4.9	6.2
geometric mean for Goodhue County (basement), $\exp(\beta_1 + \beta_4)$	5.8	6.8	7.9
factor for basement vs. no basement, $\exp(\beta_1)$	1.1	1.4	1.7
geometric sd of predictions, $\exp(\sigma)$	2.1	2.2	2.4

All estimates in the table come from the R code below: applying the normal regression model to the logarithms of the data, drawing 1000 simulations of (β, σ), and then computing quantiles of the quantities of interest.

Nontechnical summary: The first six lines of the above table gives stimates and uncertainties for the average county radon level (separating into houses without and with basements). These range from about 4 to 8 pCi/L. Within a county, the houses with basements have, on average, radon levels between 10% and 70% higher than houses without basements. One can expect the radon measurement in any given house to vary by about a factor of 2.2 from the mean of houses of the same "basement status" in the same county.

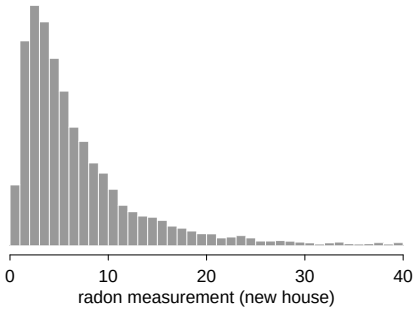
```
make.indicators <- function (x){
  ux <- unique(x)
  mat1 <- matrix (x, nrow=length(x), ncol=length(ux))
  mat2 <- matrix (ux, nrow=length(x), ncol=length(ux), byrow=TRUE)
  (mat1==mat2)*1}
y.1 <- c (5.0, 13.0, 7.2, 6.8, 12.8, 5.8, 9.5, 6.0, 3.8, 14.3, 1.8, 6.9, 4.7, 9.5)
y.2 <- c (0.9, 12.9, 2.6, 3.5, 26.6, 1.5, 13.0, 8.8, 19.5, 2.5, 9.0, 13.1, 3.6, 6.9)
y.3 <- c (14.3, 6.9, 7.6, 9.8, 2.6, 43.5, 4.9, 3.5, 4.8, 5.6, 3.5, 3.9, 6.7)
basement.1 <- c(1,1,1,1,1,0,1,1,1,0,1,1,1,1)
basement.2 <- c(0,1,1,0,1,1,1,1,1,0,1,1,1,0)
basement.3 <- c(1,0,1,0,1,1,1,1,1,1,1,1,1,1)
counties <- rep(1:3,c(length(y.1),length(y.2),length(y.3)))
y <- c(y.1,y.2,y.3)
x <- cbind (c(basement.1,basement.2,basement.3), make.indicators(counties))
ls.out <- lsfit (x, log(y), intercept=F)
lsd <- ls.diag (ls.out)

nsim <- 10000
n <- nrow(x)
k <- ncol(x)
sigma <- rep (NA, nsim)
beta <- array (NA, c(nsim, k))
for (i in 1:nsim){
  sigma[i] <- lsd$std.dev*sqrt((n-k)/rchisq(1,n-k))
```

```
  beta[i,] <- ls.out$coef + (sigma[i]/lsd$std.dev)*lsd$std.err*t(chol(lsd$corr))%*%rnorm(k)}
output <- exp (cbind (beta[,2], beta[,1]+beta[,2], beta[,3],
  beta[,1] + beta[,3], beta[,4], beta[,1] + beta[,4], beta[,1], sigma))
for (i in 1:ncol(output)) print (round(quantile(output[,i],c(.25,.5,.75)),1))
```

14.1b. Let $\tilde{y}$ be the radon measurement in the new house, $\tilde{b}$ be the variable that equals 1 if the house has a basement, and $\tilde{c}$ be the county where the house is located. We are given $\tilde{c}$. Then we can draw from the posterior predictive distribution of $\tilde{y}$ in four steps:

1. Draw (β, σ^2) from the posterior distribution as in Exercise 8.1a.
2. Draw $\tilde{b}$ from the posterior distribution $p(b|\tilde{c})$.
3. Draw $\log \tilde{y}$ from the conditional posterior distribution $p(\log y|\tilde{b}, \tilde{c})$ from the regression model in Exercise 8.1a ($\mathrm{N}(\beta_3, \sigma^2)$ if $b = 0$ or $\mathrm{N}(\beta_1 + \beta_3, \sigma^2)$ if $b = 1$).
4. Compute $\tilde{y} = \exp(\log \tilde{y})$.

To perform step 2 above, we need a model for $b|c$. Let θ_c be the proportion of houses in Blue Earth County that have basements. For simplicity, we shall assume a binomial model (equivalent to assuming simple random sampling of houses within each county, and a large population of houses relative to the sample size), so that $b_i|c_i \sim \text{Bernoulli}, (\theta_c)$, for each house i in this county. We also, for simplicity, assume a uniform prior distribution on θ_c. The posterior distribution for θ_c is then $\text{Beta}(3, 13)$. So we draw $\tilde{b}$ by first drawing $\theta_c \sim \text{Beta}(3, 13)$, then drawing $\tilde{b} \sim \text{Bernoulli}(\theta_c)$.




```
theta <- rbeta (nsim, 3, 13)
b <- rbinom (nsim, 1, theta)
logy.rep <- rnorm (nsim, beta[,3] + b*beta[,1], sigma)
y.rep <- exp(logy.rep)
print (round(quantile(y.rep,c(.025,.25,.5,.75,.975)),1))
hist (y.rep[y.rep<40], yaxt="n", breaks=0:40,
  xlab="radon measurement (new house)", cex=2)
```

The 2.5%, 25%, 50%, 75%, and 97.5% posterior quantiles for $\tilde{y}$ are 0.8, 2.8, 5.1, 9.0, and 33.3. To make the histogram readable, we omit the approximately 2% of simulated values that are greater than 40.

14.3. Assuming $p(\beta|\sigma) = p(\beta)$ to be uniform, we have $p(\beta|\sigma, y)$ proportional to $p(y|\beta, \sigma)$, that is, to $\exp(-\frac{1}{2}(y - X\beta)^t(y - X\beta)/\sigma^2)$. This posterior density is an exponential of a quadratic form in β and hence is normal.

To show: the mean is $\hat{\beta} = (X^tX)^{-1}X^ty$ and the variance is $(X^tX)^{-1}\sigma^2$.

It suffices to demonstrate that the following two quantities are equal (up to additive terms not depending on β, which correspond to multiplicative constants in the posterior density function): $(y - X\beta)^t(y - X\beta)$ and $(\beta - \hat{\beta})^tX^tX(\beta - \hat{\beta})$.

Expanding the quadratics, $(y - X\beta)^t(y - X\beta) = -\beta^tX^ty - y^tX\beta + \beta^tX^tX\beta + \text{constant}$, and $(\beta - \hat{\beta})^tX^tX(\beta - \hat{\beta}) = \beta^tX^tX\beta - y^tX(X^tX)^{-1}(X^tX)\beta - \beta^t(X^tX)(X^tX)^{-1}X^ty + \text{constant} = \beta^tX^tX\beta - y^tX\beta - \beta^tX^ty + \text{constant}$, which establishes (14.4) and (14.5).

14.4. $p(\sigma^2|y) = p(\beta,\sigma^2|y)/p(\beta|\sigma^2,y)$. Assuming a prior density $p(\beta,\sigma^2) = p(\beta)p(\sigma^2) = \sigma^{-2}$ and fixing β at the value $\hat{\beta}$ on the right side of this equation, we get something proportional to

$$\frac{\sigma^{-(n+2)}\exp(-\frac{1}{2}(y-X\hat{\beta})^t(y-X\hat{\beta})/\sigma^2)}{(\det((X^tX)^{-1}\sigma^2))^{-1/2}\exp(0)}. \tag{10}$$

Note that $\det((X^tX)^{-1}\sigma^2) = \sigma^{2k}\det((X^tX)^{-1})$, since X^tX has dimension k, and this expression is σ^{2k} times a constant. Thus (10) is proportional in σ to $\sigma^{-(n-k+2)}\exp(-\frac{1}{2}\sigma^{-2}(y-X\hat{\beta})^t(y-X\hat{\beta})/\sigma^2)$.

Comparing this result to the scaled inverse-χ^2 density in Appendix A reveals that our posterior density is, in fact, inverse-$\chi^2(n-k,s^2)$.

14.7. First write the joint distribution of everything, $p(\beta,\log\sigma,y,\tilde{y}) = \mathrm{N}(y|\beta,\sigma^2I)\mathrm{N}(\tilde{y}|\beta,\sigma^2I)$. Now consider y and σ to be constants (that is, condition on them). The joint density is then a quadratic function in $(\beta,\tilde{y})$. So $(\beta,\tilde{y})$ jointly have a normal distribution, which means that the marginal distribution of $\tilde{y}$ is also normal (see Appendix A).

17.1a. Conditional on the parameters $\mu_1,\tau_1,\lambda_1,\mu_2,\tau_2,\lambda_2$ (which are assumed known for this problem), the posterior distribution is

$$p(\theta|y) \propto \prod_{j=1}^{8}\sum_{m=1}^{2}\lambda_m\mathrm{N}(\theta_j|\mu_m,\tau_m^2)\mathrm{N}(y_j|\theta_j,\sigma_j^2).$$

We use the fact that

$$\mathrm{N}(\theta|\mu,\tau^2)\mathrm{N}(y|\theta,\sigma^2) = \mathrm{N}(y|\mu,\sigma^2+\tau^2)\mathrm{N}\left(\theta\left|\frac{\frac{1}{\tau^2}\mu+\frac{1}{\sigma^2}y}{\frac{1}{\tau^2}+\frac{1}{\sigma^2}},\frac{1}{\frac{1}{\tau^2}+\frac{1}{\sigma^2}}\right.\right).$$

(This is based on the identity $p(\theta)p(y|\theta) = p(y,\theta) = p(y)p(\theta|y)$ for the normal model.) We can then write the posterior distribution for the mixture model as

$$p(\theta|y) \propto \prod_{j=1}^{8}\sum_{m=1}^{2}\lambda_{mj}^*\mathrm{N}\left(\theta_j\left|\frac{\frac{1}{\tau_m^2}\mu_m+\frac{1}{\sigma_j^2}y_j}{\frac{1}{\tau_m^2}+\frac{1}{\sigma_j^2}},\frac{1}{\frac{1}{\tau_m^2}+\frac{1}{\sigma_j^2}}\right.\right),$$

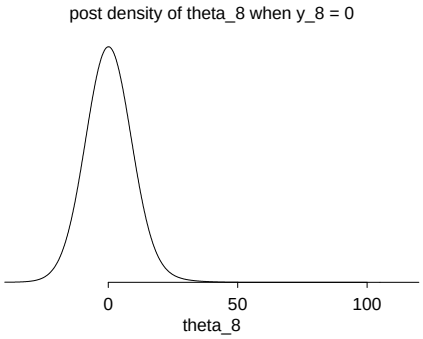
where

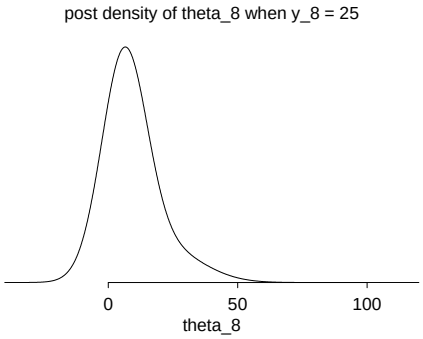
$$\lambda_{mj}^* = \frac{\lambda_m\mathrm{N}(y_j|\mu_m,\sigma_j^2+\tau_m^2)}{\lambda_1\mathrm{N}(y_j|\mu_1,\sigma_j^2+\tau_1^2)+\lambda_2\mathrm{N}(y_j|\mu_2,\sigma_j^2+\tau_2^2)} \quad \text{for } m=1,2.$$

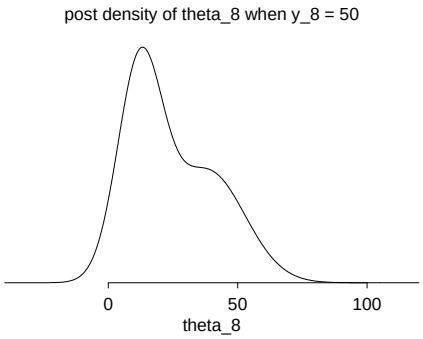
For the given parameter values and the data in Table 5.2, the posterior distributions are independent and are as follows:

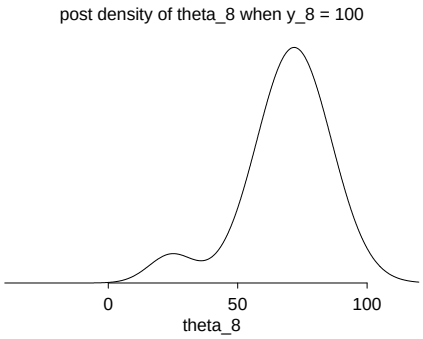
$$\begin{aligned}
\theta_1|y &\sim 0.70\,\mathrm{N}(8.8,8.3^2) + 0.30\,\mathrm{N}(24.9,12.8^2)\\
\theta_2|y &\sim 0.90\,\mathrm{N}(3.9,7.1^2) + 0.10\,\mathrm{N}(8.9,9.4^2)\\
\theta_3|y &\sim 0.94\,\mathrm{N}(-0.8,8.5^2) + 0.06\,\mathrm{N}(2.5,13.7^2)\\
\theta_4|y &\sim 0.91\,\mathrm{N}(3.1,7.4^2) + 0.09\,\mathrm{N}(8.1,10.1^2)\\
\theta_5|y &\sim 0.97\,\mathrm{N}(-0.3,6.8^2) + 0.03\,\mathrm{N}(1.3,8.8^2)\\
\theta_6|y &\sim 0.95\,\mathrm{N}(0.3,7.5^2) + 0.05\,\mathrm{N}(3.1,10.4^2)\\
\theta_7|y &\sim 0.71\,\mathrm{N}(8.7,7.2^2) + 0.29\,\mathrm{N}(17.6,9.6^2)\\
\theta_8|y &\sim 0.88\,\mathrm{N}(3.0,8.7^2) + 0.12\,\mathrm{N}(13.1,14.4^2)
\end{aligned}$$

17.1b. The effect of the two-component mixture model is to shrink toward μ_1 when y_j is low; to shrink toward μ_2 when y_j is high; and to look like a mixture for intermediate values of y_j.










```
lambda0 <- c(.9,.1)
mu0 <- c(0,15)
tau0 <- c(10,25)
y <- c(28.39, 7.94, -2.75, 6.82, -0.64, 0.63, 18.01, 12.16)
sigma <- c(14.9, 10.2, 16.3, 11.0, 9.4, 11.4, 10.4, 17.6)
J <- length(y)
M <- length(lambda0)
lambda <- array (NA, c(J,M))
mu <- array (NA, c(J,M))
tau <- array (NA, c(J,M))
for (m in 1:M){
  lambda[,m] <- lambda0[m]*
    dnorm(y,mu0[m],sqrt(sigma^2+tau0[m]))
  tau[,m] <- sqrt(1/(1/tau0[m]^2 + 1/sigma^2))
  mu[,m] <- (mu0[m]/tau0[m]^2 + y/sigma^2)*tau[,m]^2}
for (j in 1:J)
  lambda[j,] <- lambda[j,]/sum(lambda[j,])
print (cbind(lambda[,1],mu[,1],tau[,1],
  lambda[,2],mu[,2],tau[,2]))

possible <- c(0,25,50,100)
for (i in 1:length(possible)){
  y[8] <- possible[i]
  for (m in 1:M){
    lambda[,m] <- lambda0[m]*
      dnorm(y,mu0[m],sqrt(sigma^2+tau0[m]))
    tau[,m] <- sqrt(1/(1/tau0[m]^2 + 1/sigma^2))
    mu[,m] <- (mu0[m]/tau0[m]^2 + y/sigma^2)*tau[,m]^2}
  for (j in 1:J)
    lambda[j,] <- lambda[j,]/sum(lambda[j,])
  theta <- seq(-40,120,.1)
  dens <- lambda[8,1]*dnorm(theta,mu[8,1],tau[8,1]) +
    lambda[8,2]*dnorm(theta,mu[8,2],tau[8,2])
  plot (theta, dens, ylim=c(0,1.1*max(dens)),
    type="l", xlab="theta_8", ylab="", xaxs="i",
    yaxs="i", yaxt="n", bty="n", cex=2)
  mtext (paste("Post density of theta_8 when",
    "y_8 =", possible[i]), cex=2, 3)}
```