PHP2530: Bayesian Statistical Methods

Homework 3

Antonella Basso

July 9, 2022

Problem 1: (Ch.5.3)

Hierarchical Models and Multiple Comparisons:

a. Reproduce the computations in Section 5.5 for the educational testing example. Use the posterior simulations to estimate:

   i. for each school j, the probability that its coaching program is the best of the eight; and

   ii. for each pair of schools, j and k, the probability that the coaching program in school j is better than that in school k.

b. Repeat (a), but for the simpler model with τ set to ∞ (that is, separate estimation for the eight schools). In this case, the probabilities (ii) can be computed analytically.

c. Discuss how the answers in (a) and (b) differ.

d. In the model with τ set to 0, the probabilities (i) and (ii) have degenerate values; what are they?

Table 5.2:

School	y_j	σ_j
A	28	15
B	8	10
C	-3	16
D	7	11
E	-1	9
F	1	11
G	18	10
H	12	18

Solution

a. Let $y_j = \bar{y}_{\cdot j}$ and $\sigma_j^2 = \sigma^2/n_j$ be the j^{th} school's estimated coaching effect[1] on SAT-V scores and its corresponding sampling variance, for $j = 1, 2, ..., 8$, where n_j is the sample size in school j. Moreover, let θ_j's be the true population coaching effects in the eight schools, which are assumed to be drawn from a normal distribution with hyperparameters (μ, τ). Thus assuming that all y_j's are obtained through independent experiments and have approximately normal sampling distributions, for which σ_j^2's are known, we compute the posterior distribution of the parameters θ_j using the normal hierarchical model, presented in Section 5.4, as follows:

[1] Where $\bar{y}_{\cdot j} = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}$ is the sample mean of the j^{th} group (or school) and is a sufficient statistic for this model.

(Sampling) Likelihood:

$$y_j|\theta_j \sim \mathrm{N}(\theta_j, \sigma_j^2), \quad \text{for } j = 1, 2, ..., 8$$

Prior:

$$p(\theta_1, ..., \theta_8|\mu, \tau) = \prod_{j=1}^{8} \mathrm{N}(\theta_j|\mu, \tau^2)$$

$$p(\theta_1, \theta_2, ..., \theta_8) = \int \prod_{j=1}^{8} [\mathrm{N}(\theta_j|\mu, \tau^2)] p(\mu, \tau) d(\mu, \tau)$$

Noninformative (Uniform) Hyperprior:

$$p(\mu, \tau) = p(\mu|\tau)p(\tau) \propto 1$$

Joint Posterior:

$$p(\theta, \mu, \tau|y) \propto p(\mu, \tau)p(\theta|\mu, \tau)p(y|\theta)$$
$$\propto p(\mu, \tau) \prod_{j=1}^{8} \mathrm{N}(\theta_j|\mu, \tau^2) \prod_{j=1}^{8} \mathrm{N}(y_j|\theta_j, \sigma_j^2)$$

Conditional Posterior of θ_j:

$$\theta_j|\mu, \tau, y_j \sim \mathrm{N}(\hat{\theta}_j, V_j), \quad \text{for } j = 1, 2, ..., 8$$
$$\text{where} \;\; \hat{\theta}_j = \frac{\frac{1}{\sigma_j^2} y_j + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}, \; V_j = \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}$$

Marginal Posterior of (μ, τ):

$$p(\mu, \tau|y) = p(\mu|\tau, y)p(\tau|y) \propto p(\mu, \tau)p(y|\mu, \tau)$$
$$\propto p(\mu, \tau) \prod_{j=1}^{8} \mathrm{N}(y_j|\mu, \sigma_j^2 + \tau^2)$$

Posterior of μ given τ:

$$\mu|\tau, y \sim \mathrm{N}(\hat{\mu}, V_\mu), \quad \text{where} \;\; \hat{\mu} = \frac{\sum_{j=1}^{8} \frac{1}{\sigma_j^2+\tau^2} y_j}{\sum_{j=1}^{8} \frac{1}{\sigma_j^2+\tau^2}}, \; V_\mu^{-1} = \sum_{j=1}^{8} \frac{1}{\sigma_j^2 + \tau^2}$$

Marginal Posterior of τ:

$$p(\tau|y) = \frac{p(\mu, \tau|y)}{p(\mu|\tau, y)}$$
$$\propto \left. \frac{p(\tau) \prod_{j=1}^{8} \mathrm{N}(y_j|\mu, \sigma_j^2 + \tau^2)}{\mathrm{N}(\mu|\hat{\mu}, V_\mu)} \right|_{\mu=\hat{\mu}} = \frac{p(\tau) \prod_{j=1}^{8} \mathrm{N}(y_j|\hat{\mu}, \sigma_j^2 + \tau^2)}{\mathrm{N}(\hat{\mu}|\hat{\mu}, V_\mu)}$$
$$\propto p(\tau) V_\mu^{1/2} \prod_{j=1}^{8} (\sigma_j^2 + \tau^2)^{-1/2} \exp\left(- \frac{(y_j - \hat{\mu})^2}{2(\sigma_j^2 + \tau^2)} \right)$$

Assuming a uniform prior, $p(\tau) \propto 1$, on τ, we obtain the following graph of the marginal posterior density and approximate CDF of τ (in range [0, 30]), via simulation, shown in Figures 1 and 2 below.

Figure 1: Density of Marginal Posterior Tau

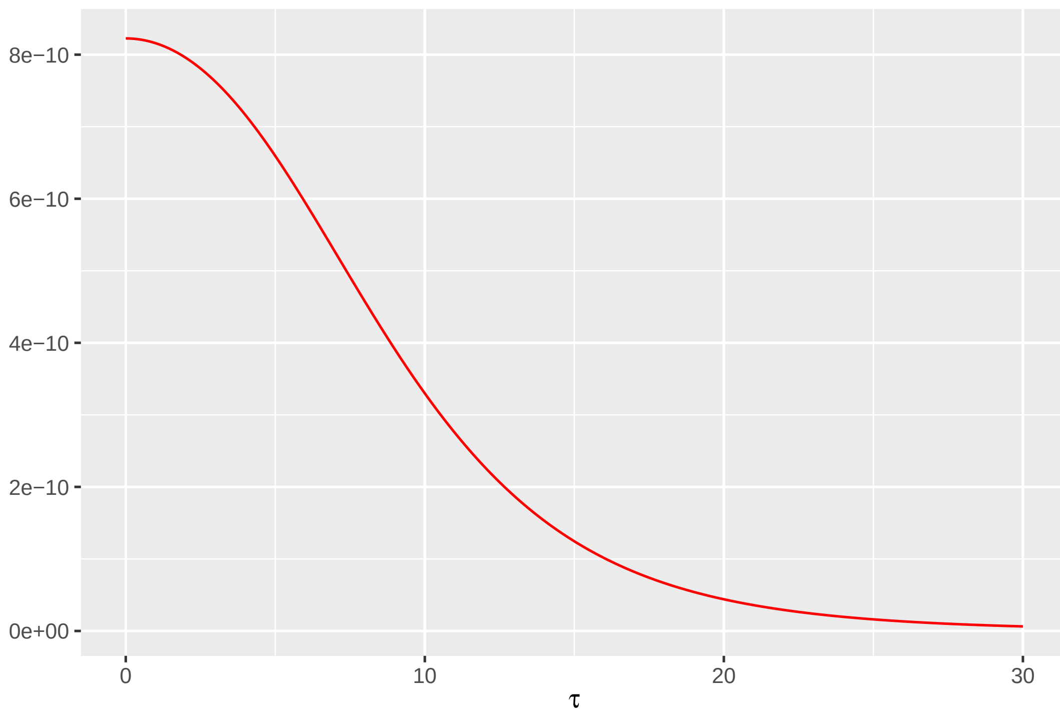
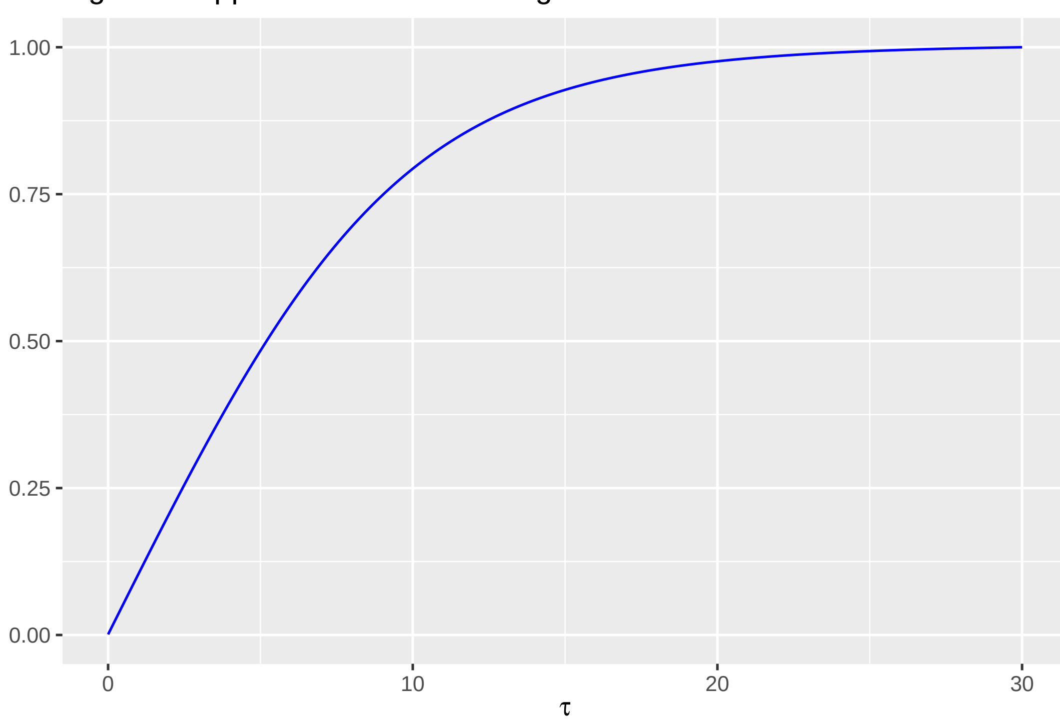


Figure 2: Approximate CDF of Marginal Posterior Tau



To obtain the disired probabilities (i) and (ii), we first follow these steps in a simulation-based approach for approximating our parameters of interest:

1. Sample n values[2] of τ from its marginal posterior distribution[3], $p(\tau|y)$, by appeal to its approximate CDF, F_τ, through the inverse transform (sampling) method for discrete distributions (mentioned in Section 1.9) as follows.
   a. Sample n values from $U[0,1]$.
   b. For each draw, u, find the value of $\tau = \tau^*$ in the CDF for which $u = F_\tau(\tau^*)$, or equally, for which $\tau^* = F_\tau^{-1}(u)$. That is, $\tau^* = mq$, where q is the index for which $\sum_{i=1}^{q} P(\tau = \tau_i) \leq u$, and $m = max\{\tau\}/n$ is the interval[4] used for τ.
2. Conditional on τ, sample n values of μ from $\mu|\tau, y \sim \mathrm{N}(\hat{\mu}, V_\mu)$. Namely, for each draw of τ from (1), τ_x, compute the values $\hat{\mu}_x$ and V_{μ_x}, and draw a single μ_x from $\mathrm{N}(\hat{\mu}_x, V_{\mu_x})$. This should result in n total pairs (μ, τ).
3. Conditional on each (μ, τ) pair, sample n values for each $theta_j$ from its corresponding conditional posterior $\theta_j|\mu, \tau, y_j \sim \mathrm{N}(\hat{\theta}_j, V_j)$. In other words, for each simulated pair, (μ_x, τ_x), compute the values $\hat{\theta}_{j_x}$ and V_{j_x}, and draw a single θ_{j_x} from $\mathrm{N}(\hat{\theta}_{j_x}, V_{j_x})$. This should result in n total draws of θ_j for each j=1,2,. . . ,8.

Using our simulated posterior values of τ, μ, and θ_j, we now estimate probabilities (i) and (ii) as follows.

i. Estimating the probability that school j's coaching program is the best of the eight is tantamount to estimating the probability that school j has the greatest coaching effect, θ_j. Thus, for each n-sample of θ_j we compute,

$$P(\theta_j = max\{\theta_1, \theta_2, ..., \theta_8\}) = \frac{1}{n}\sum_{i=1}^{n}\big[\theta_{ij} = max\{\theta_{i1}, \theta_{i2}, ..., \theta_{i8}\}\big].$$

That is, for each θ_j, we determine the proportion of times (out of the n draws) that it was the largest of all eight coaching effects. Multiplying each resulting value by 100, we obtain the following eight probabilities in the form of percentages.

Table 2: Probability (%) of Having the Best Coaching Program

School	Probability (%)
A	24.53
B	10.67
C	8.13
D	9.63
E	5.33
F	6.93
G	20.20
H	14.57

[2]For this simulation we use $n = 3,000$.

[3]Given the support of this distribution, it is safe to restrict our sample to values of τ between 0 and 30. Moreover, in using increments of size 0.01 to estimate τ's marginal posterior density, we limit our sample further to values of $\tau \in \{0, 0.01, 0.02, ..., 30\}$.

[4]For this simulation, $m = 30/3,000 = 0.01$.

ii. For each pair of schools, j and k, the estimated probability that the coaching program in school j is better than that in school k can be expressed as the probability that school j's coaching effect, θ_j is greater than the coaching effect in school k, θ_k. Namely,

$$P(\theta_j > \theta_k) = \frac{1}{n}\sum_{i=1}^{n} \theta_{ij} > \theta_{ik}, \quad \text{for } \{j, k \in \{1, 2, ..., 8\} |\ j \neq k\}.$$

In a similar approach to (i), that is, we calculate for each (θ_j, θ_k) pair, the proportion of times (of all n draws) that θ_j was greater than θ_k. Multiplying each resulting value by 100, we obtain the following $\frac{8!}{(8-2)!} = 56$ probabilities in the form of percentages.

Table 3: Probability (%) of Having a Better Coaching Program than School k

	School k							
j	A	B	C	D	E	F	G	H
A	—	12.87	21.17	14.00	23.33	20.13	6.70	12.67
B	16.70	—	25.10	17.73	27.90	24.83	8.97	15.90
C	19.80	20.63	—	22.20	34.23	29.83	11.90	19.83
D	23.37	26.00	35.13	—	39.73	35.97	15.67	24.63
E	27.70	31.90	40.37	32.47	—	42.10	20.23	30.00
F	31.83	38.17	47.10	38.40	53.23	—	25.60	35.53
G	35.73	44.60	53.43	44.93	60.23	54.00	—	41.47
H	40.57	52.13	59.03	51.63	65.90	60.60	38.13	—

b. Evaluating the normal hierarchical model from part (a) in the limit as $\tau \to \infty$, we obtain the following conditional posterior distribution for θ_j,

$$\theta_j | \mu, \tau, y_j \sim \mathrm{N}(\hat{\theta}_j, V_j), \quad \text{for } j = 1, 2, ..., 8,$$

$$\text{where } \hat{\theta}_j = \lim_{\tau\to\infty} \frac{\frac{1}{\sigma_j^2} y_j + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} = y_j \text{ and } V_j = \lim_{\tau\to\infty} \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} = \sigma_j^2.$$

That is, under this new assumption, each θ_j can be estimated separately from their observed sample means and variances. Namely, from a normal distribution with parameters y_j and σ_j^2. Following the same simulation-based procedure to compute probabilities (i) and (ii) in the previous section, we now obtain the following.

Table 4: Probability (%) of Having the Best Coaching Program

School	Probability (%)
A	52.27
B	3.43
C	2.50
D	3.97
E	0.37
F	1.33
G	17.30
H	18.83

Table 5: Probability (%) of Having a Better Coaching Program than School k

	School k							
j	A	B	C	D	E	F	G	H
A	—	25.37	59.07	29.67	58.67	51.20	4.77	25.47
B	4.50	—	61.30	32.73	63.37	55.10	5.77	27.40
C	5.30	31.90	—	35.33	67.43	58.23	7.03	29.30
D	5.73	35.17	65.57	—	71.53	61.30	8.50	31.27
E	6.47	38.57	68.03	41.73	—	65.30	10.13	33.27
F	7.40	42.70	70.53	45.17	78.30	—	12.03	35.07
G	8.73	46.73	73.03	48.93	81.23	71.17	—	36.97
H	10.10	50.73	74.73	52.40	83.77	73.87	15.80	—

Note:

Asssuming infinite between-group variance.

c. Evidently, the probabilities from part (a) displayed in Tables 2 and 3 are far less extreme compared to those obtained in part (b), which appear to overemphasize the differences between observed coaching effects. Particularly, this change in probabilities from part (a) to part (b) can be attributed to the fact that when we allow the hyperparameter, τ, to take on an infinite value, our posterior estimates are pulled away from one another in the directions of their distinct sample statistics, resulting in values for (i) and (ii) that are similar to what we would obtain from the observed data. That is, in more practical terms, when we assume a population distribution in which variability is considerably large, we allow these effects/parameters to be radically different from each other, and hence, we increase our posterior uncertainty about the true value of each θ_j.

d. Probabilities (i) and (ii) given by Tables 6 and 7 below, which were computed (using the same simulation-based approach from parts (a) and (b)) under the assumption that $\tau = 0$, evidently take on degenerate values. Specifically, we notice in Table 6 that each $P(\theta_j = max\{\theta_1, \theta_2, ..., \theta_8\}) \approx 1/8$. That is, conditional on $\tau = 0$, all schools have approximately the same probability of having the largest coaching effect. Similarly, we notice in Table 7 that $P(\theta_j > \theta_k) \approx 1/2$, meaning that for any pair of schools $(j, k) \in \{1, 2, ..., 8\}|_{j \neq k}$ there is roughly a 50% chance that the effect in school j is greater than that in school k, under the assumption that no variability exists between coaching effect estimates. Evaluating the normal hierarchical model from part (a) in the limit as $\tau \to 0$ we see that each coaching effect estimate, θ_j, eventually reduces to the pooled estimate, $\hat{\theta} = \bar{y}_{..}$. This would indicate that observations, y_j, must come from the same normal distribution and produce independent estimates of the same quantity, resulting in the degeneracy we see in the tables below. Thus, we can infer under this assumption that coaching effects in all eight schools are equal.

Table 6: Probability (%) of Having the Best Coaching Program

School	Probability (%)
A	11.53
B	13.03
C	12.83
D	11.87
E	12.10
F	12.23
G	12.03
H	14.37

Table 7: Probability (%) of Having a Better Coaching Program than School k

j	School k A	B	C	D	E	F	G	H
A	—	35.10	35.23	34.17	34.77	35.73	33.57	32.77
B	38.03	—	37.73	35.93	36.70	37.93	36.20	34.70
C	40.00	39.20	—	38.23	38.93	40.10	38.33	36.70
D	42.17	41.67	42.50	—	41.23	42.17	40.60	39.03
E	43.97	44.40	44.47	42.60	—	44.73	43.10	41.23
F	46.30	46.47	46.93	44.87	45.50	—	45.60	44.03
G	48.47	49.03	49.43	46.97	47.87	49.47	—	46.33
H	51.23	51.33	51.77	49.67	50.37	52.20	51.33	—

Note:

Asssuming 0 between-group variance.

Posterior Estimates Under Different Assumptions for Tau

Table 8: Coaching Effect Estimates

School	Observed	Complete Pooling	Partial-Pooling	No-Pooling
A	28	7.2	11	27.55
B	8	7.54	7.77	7.9
C	-3	7.7	6.08	-3
D	7	7.66	7.47	6.97
E	-1	7.66	5.16	-1.02
F	1	7.46	6.02	0.84
G	18	7.69	10.44	17.99
H	12	8.39	8.6	12.75
Variance:	109.07	0.11	4.44	107.83

Figure 3.a: Coaching Effect Posterior Estimates, Complete Pooling

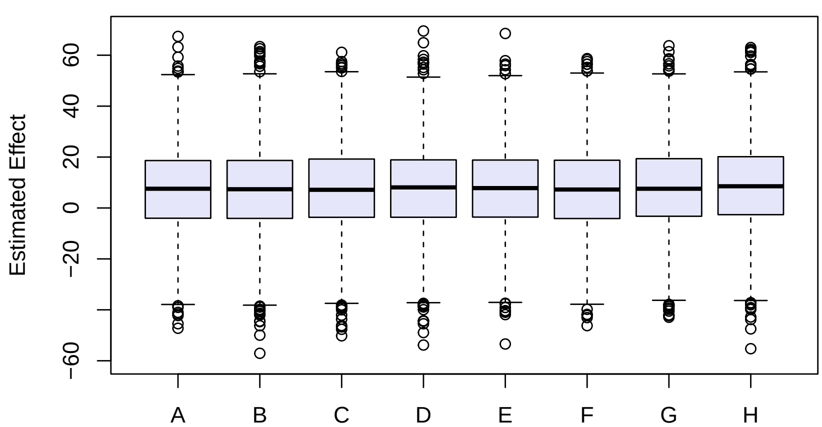


Figure 3.b: Coaching Effect Posterior Estimates, Partial−Pooling

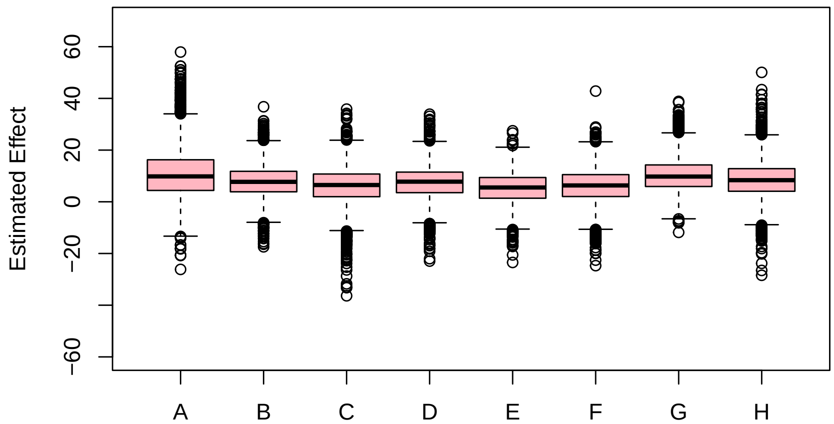
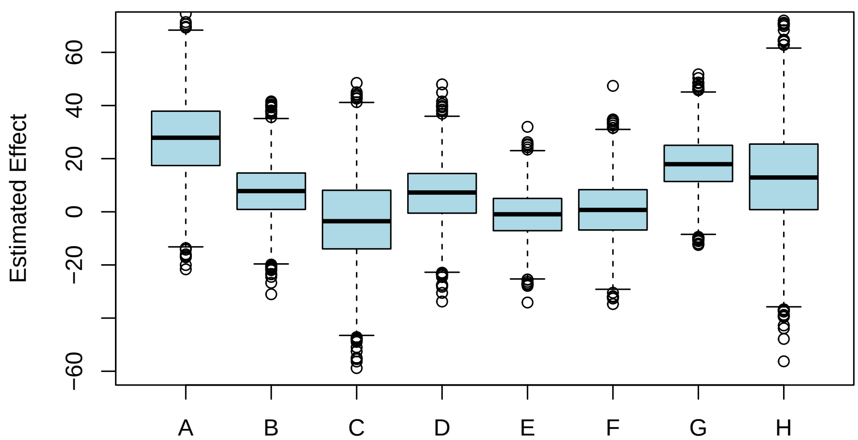


Figure 3.c: Coaching Effect Posterior Estimates, No−Pooling



Problem 2: (Ch.5.5)

Mixtures of independent distributions: Suppose the distribution of $\theta = (\theta_1, ..., \theta_J)$ can be written as a mixture of independent and identically distributed components:

$$p(\theta) = \int \prod_{j=1}^{J} p(\theta_j|\phi)p(\phi)d\phi.$$

Prove the covariances, $Cov(\theta_i, \theta_j)$, are all non-negative.

Solution

Intuition: If all elements of θ are independent and identically distributed (iid) given (hyper)parameter(s), ϕ, then, it is safe to treat θ as an iid random variable which can take on any value in $\{\theta_1, ..., \theta_J\}$ (denoting J independent groups/experiments). From this, it would follow that they must have equal conditional expectations (namely, $\mathrm{E}[\theta_{i\in J}|\phi] = \mathrm{E}[\theta_{j\neq i\in J}|\phi] = \mathrm{E}[\theta|\phi]$), and hence, equal expectations ($\mathrm{E}[\theta_{i\in J}] = \mathrm{E}[\theta_{j\neq i\in J}]$) by the Law of Total Expectation. This being the case, it would make sense to extend a similar thinking to their covariances. That is, if all elements of θ are estimating the same quantity, then it seems reasonable for their covariances to reduce to the variance in θ, which necessarily must be greater than 0.

Proof: Provided that $\theta_i|\phi$ are iid, it must be the case that,

$$\forall_{\theta_i}\forall_{\theta_j} \in \theta|_{i\neq j},\ \mathrm{E}[\theta_i|\phi] = \mathrm{E}[\theta_j|\phi].$$

Thus, by the Law of Total Covariance (an extension of the Law of Total Variance (or Eve's Law) outlined in Section 1.8), it follows that,

$$\begin{aligned}
Cov(\theta_i, \theta_j) &= \mathrm{E}\big[Cov(\theta_i, \theta_j|\phi)\big] + Cov\big(\mathrm{E}[\theta_i|\phi], \mathrm{E}[\theta_j|\phi]\big) \\
&= \mathrm{E}\big[\mathrm{E}[\theta_i, \theta_j|\phi] - \mathrm{E}[\theta_i|\phi]\mathrm{E}[\theta_j|\phi]\big] + Cov\big(\mathrm{E}[\theta|\phi], \mathrm{E}[\theta|\phi]\big) \\
&= \mathrm{E}\big[\mathrm{E}[\theta_i|\phi]\mathrm{E}[\theta_j|\phi] - \mathrm{E}[\theta_i|\phi]\mathrm{E}[\theta_j|\phi]\big] + V\big(\mathrm{E}[\theta|\phi]\big) \\
&= \mathrm{E}[0] + V\big(\mathrm{E}[\theta|\phi]\big) \\
&= V\big(\mathrm{E}[\theta|\phi]\big).
\end{aligned}$$

Noticeably, this is a special case of the Law of Total Covariance in which $\mathrm{E}[\theta_i|\phi] = \mathrm{E}[\theta_j|\phi]$ and hence, yields the variance of $\mathrm{E}[\theta|\phi]$, proving our initial intuition.[5] Thus, we conclude that the covariance of any pair of iid parameters $(\theta_i, \theta_j)|_{i\neq j} \in \theta$ is simply the total variance of $\mathrm{E}[\theta|\phi]$.

Problem 3: (Ch.5.11)

Nonconjugate hierarchical models: Suppose that in the rat tumor example, we wish to use a normal population distribution on the log-odds scale: $\mathrm{logit}(\theta_j) \sim \mathrm{N}(\mu, \tau^2)$, for $j = 1, ..., J$. As in Section 5.3, you will assign a noninformative prior distribution to the hyperparameters and perform a full Bayesian analysis.

a. Write the joint posterior density, $p(\theta, \mu, \tau|y)$.
b. Show that the integral (5.4) has no closed-form expression.
c. Why is expression (5.5) no help for this problem?

In practice, we can solve this problem by normal approximation, importance sampling, and Markov chain simulation, as described in Part III.

Solution

a. Assuming a noninformative hyperprior, $p(\mu, \tau) \propto 1$, and given a likelihood of $y_j \sim Bin(n_j, \theta_j)$ and a logit-normal distribution[6] for each θ_j, $j = 1, 2, ..., J$, we obtain the following joint posterior for all parameters,

$$\begin{aligned}
p(\theta, \mu, \tau|y) &= p(\mu, \tau)p(\theta|\mu, \tau)p(y|\theta, \mu, \tau) \\
&= p(\mu, \tau) \prod_{j=1}^{J} \frac{1}{\theta_j(1-\theta_j)} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left[-\frac{\big(\mathrm{logit}(\theta_j) - \mu\big)^2}{2\tau^2}\right] \prod_{j=1}^{J} \theta_j^{y_j}(1-\theta_j)^{n_j - y_j} \\
&= p(\mu, \tau) \prod_{j=1}^{J} \theta_j^{y_j - 1}(1-\theta_j)^{n_j - y_j - 1} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left[-\frac{\big(\mathrm{logit}(\theta_j) - \mu\big)^2}{2\tau^2}\right] \\
&\propto \frac{1}{\tau^J} \prod_{j=1}^{J} \theta_j^{y_j - 1}(1-\theta_j)^{n_j - y_j - 1} \exp\left[-\frac{\big(\mathrm{logit}(\theta_j) - \mu\big)^2}{2\tau^2}\right].
\end{aligned}$$

b. The integral of the joint posterior above over θ, which can be expressed as a product of J independent integrals corresponding to each of the parameters, θ_j, gives us the following marginal posterior of the hyperparameters,

[5] Note that this proof makes use of the definition of covariance, properties of expectation and variance/covariance, as well as independence assumption which allows us to write $\mathrm{E}[\theta_i, \theta_j|\phi]$ as the product $\mathrm{E}[\theta_i|\phi]\mathrm{E}[\theta_j|\phi]$.

[6] Where each $\theta_j \sim$ Logit-Normal(μ, τ^2), provided that $\mathrm{logit}(\theta_j) \sim \mathrm{N}(\mu, \tau^2)$, for $j = 1, 2, ..., J$.

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

Notably, this integral has no closed-form solution. That is, after inspection, we find that it can't be expressed in a way that allows us to evaluate it analytically. First, we may notice that the integrand has no further simplifications. Secondly, in attempting u-substitution and integration by parts, we see that neither computation strategy is successful in that they result in other (perhaps even more) complicated functions which cannot be antidifferentiated. In the case of u-substitution, where the most sensible choice for u_j is $\text{logit}(\theta_j)$ (such that $du_j = \theta_j^{-1}(1-\theta_j)^{-1}d\theta_j$), we see that,

$$
\begin{aligned}
p(\mu,\tau|y) &\propto \frac{1}{\tau^J}\prod_{j=1}^{J}\int \theta_j^{y_j}(1-\theta_j)^{n_j-y_j}\exp\left[-\frac{(u-\mu)^2}{2\tau^2}\right]du_j \\
&\propto \frac{1}{\tau^J}\prod_{j=1}^{J}\int (1-\theta_j)^{n_j}\left(\frac{\theta_j}{1-\theta_j}\right)^{y_j}\exp\left[-\frac{(u-\mu)^2}{2\tau^2}\right]du_j \\
&\propto \frac{1}{\tau^J}\prod_{j=1}^{J}\int (1-\theta_j)^{n_j}\left[\exp\left[\log\left(\frac{\theta_j}{1-\theta_j}\right)\right]\right]^{y_j}\exp\left[-\frac{(u-\mu)^2}{2\tau^2}\right]du_j \\
&\propto \frac{1}{\tau^J}\prod_{j=1}^{J}\int (1-\theta_j)^{n_j}\exp\Big[y_j\ \text{logit}(\theta_j)\Big]\exp\left[-\frac{(u-\mu)^2}{2\tau^2}\right]du_j \\
&\propto \frac{1}{\tau^J}\prod_{j=1}^{J}\int (1-\theta_j)^{n_j}\exp\left[y_j u_j-\frac{(u-\mu)^2}{2\tau^2}\right]du_j,
\end{aligned}
$$

which can be further algebraically manipulated, but ultimately has no integrable from. Similarly, the approach of integration by parts results in a much more complex expression, which does not allow for analytical integration. Thus, given that our marginal posterior of the hyperparameters is not integrable, it follows that one would have to appeal to approximation or simulation-based techniques to show that it is a proper pdf.

c. Section 5.3 mentions the use of the conditional probability formula (Equation 5.5) shown below to compute the marginal posterior of a vector of hyperparameters, ϕ, algebraically, rather than using an integration method.

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

Despite the convenience it provides when applied to various (standard) conjugate models however, it is of no use in our case (for deriving $p(\mu,\tau|y)$), given the fact that the denominator, namely $p(\theta|\mu,\tau,y)$, must be of a known form. That is, to serve as a proportionality constant, which is a necessary condition provided that this density, like the one we seek to find, is a function of both μ and τ for fixed y, it would have to have a direct known and integrable form. Notice that,

$$
p(\theta|\mu,\tau,y) = \frac{p(\theta,\mu,\tau|y)}{p(\mu,\tau|y)}.
$$

Having the same requirements as $p(\mu,\tau|y)$ regarding a normalizing factor, this presupposes that to obtain $p(\theta|\mu,\tau,y)$, which is needed to compute $p(\mu,\tau|y)$ in this way, the desired marginal posterior itself must have a closed form. And, having shown in part (b) that this is not the case, it follows that $p(\theta|\mu,\tau,y)$ can't be known, which prevents us from being able to make use of this approach to obtain $p(\mu,\tau|y)$.

Problem 4: (Ch.5.14)

Hierarchical Poisson model: Consider the dataset in the previous problem, but suppose only the total amount of traffic at each location is observed.

a. Set up a model in which the total number of vehicles observed at each location j follows a Poisson distribution with parameter θ_j, the "true" rate of traffic per hour at that location. Assign a gamma population distribution for the parameters θ_j and a noninformative hyperprior distribution. Write down the joint posterior distribution.

b. Compute the marginal posterior density of the hyperparameters and plot its contours. Simulate random draws from the posterior distribution of the hyperparameters and make a scatterplot of the simulation draws.

c. Is the posterior density integrable? Answer analytically by examining the joint posterior density at the limits or empirically by examining the plots of the marginal posterior density above.

d. If the posterior density is not integrable, alter it and repeat the previous two steps.

e. Draw samples from the joint posterior distribution of the parameters and hyperparameters, by analogy to the method used in the hierarchical binomial model.

Table 3.3: Counts of Bicycles and Other Vehicles (Restricted to Residential Streets)

Type of Street	Bike Route?	Counts of Bicycles/Other Vehicles
Residential	yes	16/58, 9/90, 10/48, 13/56, 19/103, 20/57, 17/112, 35/273, 35/273, 55/64
Residential	no	12/113, 1/18, 2/14, 4/44, 9/208, 7/67, 9/29, 8/154

Solution

a. Let y_j be the number of bicycles in the j^{th} residential bike route, and let n_j be the corresponding observed count of all vehicles, such that $\bar{y}_j = y_j/n_j$ is the observed bike traffic rate (proportion of vehicles observed that were bikes). Similarly, let z_k be the observed bike count for the k^{th} residential non-bike route, where n_k denotes its total observed vehicle count and $\bar{z}_k = z_k/n_k$ the observed bike traffic rate. With $j = 1, 2, ..., 10$ and $k = 1, 2, ..., 8$, we rewrite the information from the table above as follows.

Table 9: Observed Vehicle Counts and Traffic Rates in Residential Bike Routes

j	n_j	$\bar{y}_j$
1	74	0.21621622
2	99	0.09090909
3	58	0.17241379
4	70	0.18571429
5	112	0.15573770
6	77	0.25974026
7	104	0.17307692
8	129	0.13178295
9	308	0.11363636
10	119	0.46218487

Table 10: Observed Vehicle Counts and Traffic Rates in Residential Non-Bike Routes

k	n_k	z_k
1	125	0.09600000
2	19	0.05263158
3	16	0.12500000
4	48	0.08333333
5	217	0.04147465
6	74	0.09459459
7	38	0.23684211
8	162	0.04938272

Restricting our attention only to bike routes, let θ_j be the "true" hourly bike traffic rate at the j^{th} location. Assuming a Gamma prior density for these parameters with a noninformative hyperprior, we obtain the following model (and joint posterior), given the data follows a Poisson distribution.

Likelihood:

$$y_j|\theta_j \sim \text{Poisson}(\theta_j), \quad \text{for } j = 1, 2, ..., 10$$

Prior:

$$\theta_j|\alpha, \beta \sim \text{Gamma}(\alpha, \beta), \quad \text{for } j = 1, 2, ..., 10$$

Noninformative Hyperprior:

$$p(\alpha, \beta) \propto 1$$

Joint Posterior:

$$\begin{aligned} p(\theta, \alpha, \beta|y) &\propto p(\alpha, \beta) p(\theta|\alpha, \beta) p(y|\theta) \\ &\propto p(\alpha, \beta) \prod_{j=1}^{10} \text{Gamma}(\theta_j|\alpha, \beta) \prod_{j=1}^{10} \text{Poisson}(y_j|\theta_j) \\ &\propto p(\alpha, \beta) \prod_{j=1}^{10} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta_j^{\alpha-1} e^{-\beta\theta_j} \prod_{j=1}^{10} \frac{\theta_j^{y_j} e^{-\theta_j}}{y_j!} \end{aligned}$$

b. Given the conjugacy in our model and properties of exchangeability, we can use the conditional probability formula to obtain the marginal posterior of the hyperparameters as follows. From Equation 5.5,

$$p(\alpha, \beta|y) = \frac{p(\theta, \alpha, \beta|y)}{p(\theta|\alpha, \beta, y)}, \text{ where } \ p(\theta|\alpha, \beta, y) \propto \prod_{j=1}^{10} \text{Gamma}(\theta_j|\alpha + y_j, \beta + 1).$$

Marginal Posterior of (α, β):

$$\begin{aligned} p(\alpha, \beta|y) &\propto p(\alpha, \beta) \prod_{j=1}^{10} \frac{\frac{\beta^\alpha}{\Gamma(\alpha)} \theta_j^{\alpha-1} e^{-\beta\theta_j} \frac{\theta_j^{y_j} e^{-\theta_j}}{y_j!}}{\frac{(\beta+1)^{\alpha+y_j}}{\Gamma(\alpha+y_j)} \theta_j^{\alpha+y_j-1} e^{-(\beta+1)\theta_j}} \\ &\propto \prod_{j=1}^{10} \frac{\Gamma(\alpha + y_j)}{y_j!\Gamma(\alpha)} \frac{\beta^\alpha}{(\beta+1)^{\alpha+y_j}} \\ &\propto \prod_{j=1}^{10} \frac{\Gamma(\alpha + y_j)}{y_j!\Gamma(\alpha)} \left(\frac{\beta}{\beta+1}\right)^\alpha \left(\frac{1}{\beta+1}\right)^{y_j} \\ &\propto \prod_{j=1}^{10} \text{Negative-Binomial}(y_j|r = \alpha, p = 1/(\beta+1)) \end{aligned}$$

As in Section 5.3, we use $p(\alpha, \beta) \propto (\alpha + \beta)^{-5/2}$ as the uninformative hyperprior for the alpha and beta hyperparameters (Equation 5.9), and obtain the following marginal posterior contour (Figure 4.a) and scatterplot of marginal posterior draws.

Figure 4.a: Marginal Posterior Contour of (α, β)

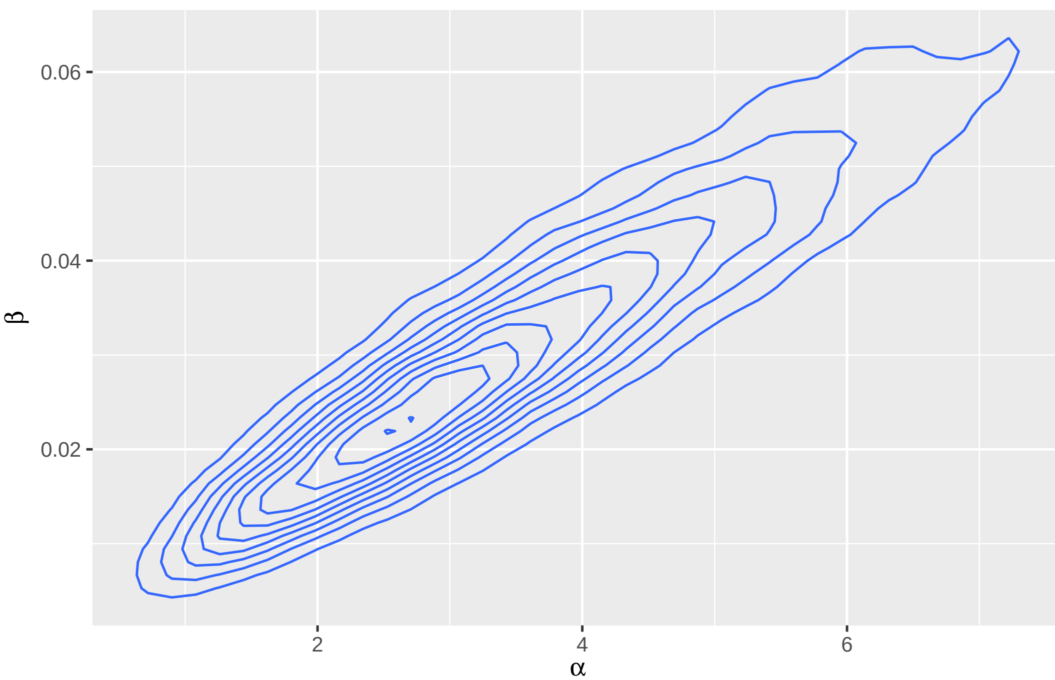
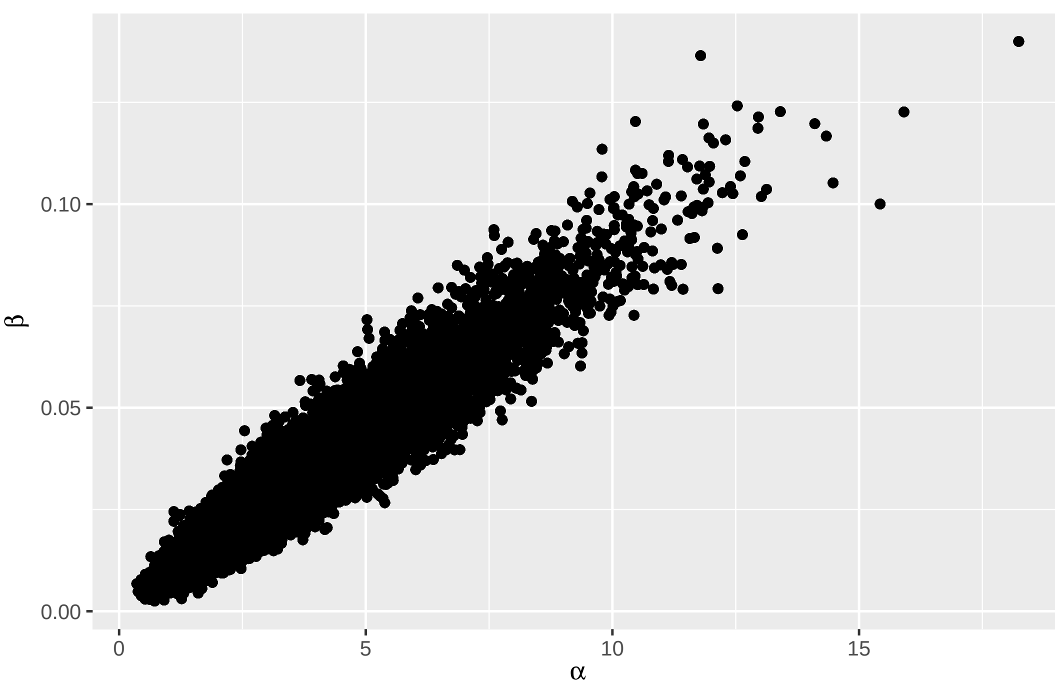


Figure 4.b: Marginal Posterior Draws of (α, β)



c. From Figures 4.a and 4.b, we see that the joint density of α and β is centered around ($\alpha \approx 2.5, \beta \approx 0.025$) and that $\beta \approx \alpha/100$ for α values between 0 and 15. That is, these figures show that roughly $\alpha \in (0, 15]$ and $\beta \in (0, 0.15]$. Given these limits as α and β tend towards 0 and ∞, which are not exact, but show that $0 < \alpha < \infty$ and $0 < \beta < \infty$, we conclude that the marginal posterior density of hyperparameters (α, β) must be integrable.

d. It is safe to assume that our choice of hyperprior distribution for this model, successfully yielded an integrable posterior. However, as discussed in the text, most other noninformative hyperpriors will result in improper (and hence, non-integrable) posterior densities.

e. Drawing samples from the joint posterior distribution of parameters θ_j and hyperparameters (α, β), we obtain the following summary statistics.

Table 12: Summary Statistics of Parameter Posterior Draws

	Mean	Mean SE	SD	2.5%	25%	50%	75%	97.5%
alpha	3.80	0.0192	1.85	1.18	2.47	3.46	4.80	8.25
beta	0.03	0.0002	0.02	0.01	0.02	0.03	0.04	0.07
theta[1]	75.17	0.0648	8.57	59.25	69.33	74.85	80.73	92.97
theta[2]	99.44	0.0723	9.77	81.27	92.71	99.13	105.85	119.50
theta[3]	59.80	0.0587	7.64	45.75	54.45	59.51	64.84	75.62
theta[4]	71.49	0.0633	8.44	55.92	65.57	71.19	77.07	88.86
theta[5]	121.62	0.0822	11.01	101.29	113.89	121.27	128.89	144.15
theta[6]	78.36	0.0656	8.63	62.18	72.47	78.09	83.94	96.17
theta[7]	104.25	0.0757	10.05	85.45	97.38	103.99	110.88	124.91
theta[8]	128.46	0.0816	11.08	107.21	120.89	128.14	135.63	151.12
theta[9]	301.44	0.1339	17.45	268.05	289.68	300.99	313.10	336.52
theta[10]	118.83	0.0774	10.70	98.84	111.53	118.50	125.79	140.70

Problem 5: (Ch.6.2)

Model checking: In Exercise 2.13, the counts of airline fatalities in 1976–1985 were fitted to four different Poisson models.

a. For each of the models, set up posterior predictive test quantities to check the following assumptions.

   i. Independent Poisson distributions.

   ii. No trend over time.

b. For each of the models, use simulations from the posterior predictive distributions to measure the discrepancies. Display the discrepancies graphically and give p-values.

c. Do the results of the posterior predictive checks agree with your answers in Exercise 2.13 (e)?

Solution

Problem 6: (Ch.6.7)

Prior vs. posterior predictive checks (from Gelman, Meng, and Stern, 1996): Consider 100 observations, $y_1, ..., y_n$, modeled as independent samples from a $\mathrm{N}(\theta, 1)$ distribution with a diffuse prior distribution, say, $p(\theta) = \frac{1}{2A}$ for $\theta \in [-A, A]$ with some extremely large value of A, such as 10^5. We wish to check the model using, as a test statistic, $T(y) = \max_i |y_i|$: is the maximum absolute observed value consistent with the normal model? Consider a dataset in which $\bar{y} = 5.1$ and $T(y) = 8.1$.

a. What is the posterior predictive distribution for y^{rep}? Make a histogram for the posterior predictive distribution of $T(y^{\text{rep}})$ and give the posterior predictive p-value for the observation $T(y) = 8.1$.

b. The prior predictive distribution is $p(y^{\text{rep}}) = \int p(y^{\text{rep}}|\theta)p(\theta)d\theta$. (Compare to equation (6.1)). What is the prior predictive distribution for y^{rep} in this example? Roughly sketch the prior predictive distribution of $T(y^{\text{rep}})$ and give the approximate prior predictive p-value for the observation $T(y) = 8.1$.

c. Your answers for (a) and (b) should show that the data are consistent with the posterior predictive but not the prior predictive distribution. Does this make sense? Explain.

Solution

a. **Model**[7]:

$$p(\theta) = \frac{1}{2A} \quad \text{for } \theta \in [-A, A] y_j|\theta \sim \mathrm{N}(\theta, 1) \quad \text{for } j = 1, 2, ..., n$$

$$\begin{aligned}
p(\theta|y) &= p(y|\theta)p(\theta) \\
&= \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y_j-\theta)^2}{2}} I_{\theta\in[-A,A]} \\
&\propto e^{-\frac{\sum_{j=1}^{n}(y_j-\theta)^2}{2}} I_{\theta\in[-A,A]} \\
&\propto e^{-\frac{(n-1)s^2+n(\bar{y}-\theta)^2}{2}} I_{\theta\in[-A,A]} \\
&\propto e^{-\frac{(n-1)s^2}{2}} e^{-\frac{n(\bar{y}-\theta)^2}{2}} I_{\theta\in[-A,A]} \\
&\propto e^{-\frac{(\theta-\bar{y})^2}{2\frac{1}{n}}} I_{\theta\in[-A,A]} \\
&\propto \mathrm{N}(\bar{y}, 1/n) I_{\theta\in[-A,A]}
\end{aligned}$$

Posterior Predictive:

$$\begin{aligned}
p(y^{\text{rep}}|y) &= \int_{\Theta} p(y^{\text{rep}}|\theta)p(\theta|y)d\theta \\
&\propto \int_{\Theta} e^{-\frac{(y^{\text{rep}}-\theta)^2}{2}} e^{-\frac{(\theta-\bar{y})^2}{2\frac{1}{n}}} I_{\theta\in[-A,A]} d\theta
\end{aligned}$$

From the properties of normal distributions outlined in Section 2.5, we gether that the posterior predictive density is normal with the following mean and variance.

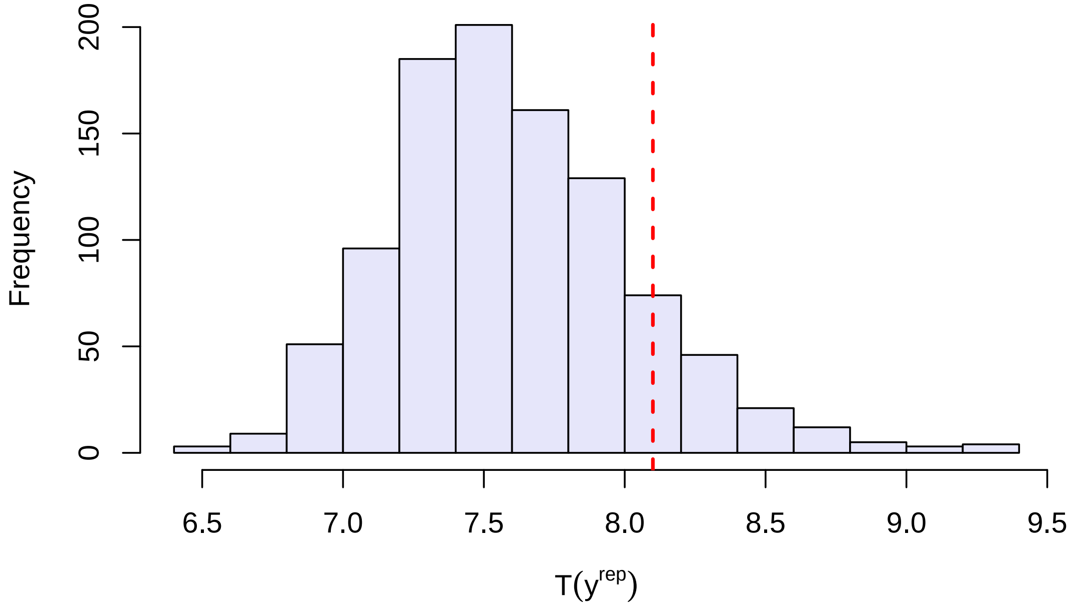
$$\mathrm{E}[y^{\text{rep}}|y] = \mathrm{E}\big[\mathrm{E}[y^{\text{rep}}|\theta, y]|y\big] = \mathrm{E}[\theta|y] = \bar{y}$$

$$\begin{aligned}
Var(y^{\text{rep}}|y) &= \mathrm{E}\big[Var(y^{\text{rep}}|\theta, y)|y\big] + Var(\mathrm{E}[y^{\text{rep}}|\theta, y]|y) \\
&= \mathrm{E}[\sigma^2|y] + Var(\theta|y) \\
&= \sigma^2 + \frac{1}{n}
\end{aligned}$$

[7] The properties outlined in the following source were used to derive this posterior: https://www.cs.ubc.ca/~murphyk/Papers/bayesGauss.pdf.

Thus, with $\bar{y} = 5.1$, $\sigma^2 = 1$, and $n = 100$, it follows that $y^{\text{rep}}|y \sim \text{N}(5.1, 1/100 + 1)$. Drawing $N = 1,000$ samples of size $n = 100$ from this posterior predictive distribution and calculating each sample's test statistic, we obtain the histogram of $T(y^{\text{rep}}) = \max_i |y_i^{\text{rep}}|$, given by Figure 5 below. Moreover, given $p_B = p(T(y^{\text{rep}}) \geq T(y)|y) \approx \frac{1}{N}\sum_{i=1}^{N} I_{T(y_i^{\text{rep}}) \geq T(y)}$, the posterior predictive p-value for $T(y) = 8.1$ is thus ≈ 0.126. This indicates that approximately 12.6% of the samples produced by our model have maximum values that exceed the one observed in the data.

Figure 5: Histogram of Posterior Predictive $T(y^{\text{rep}})$



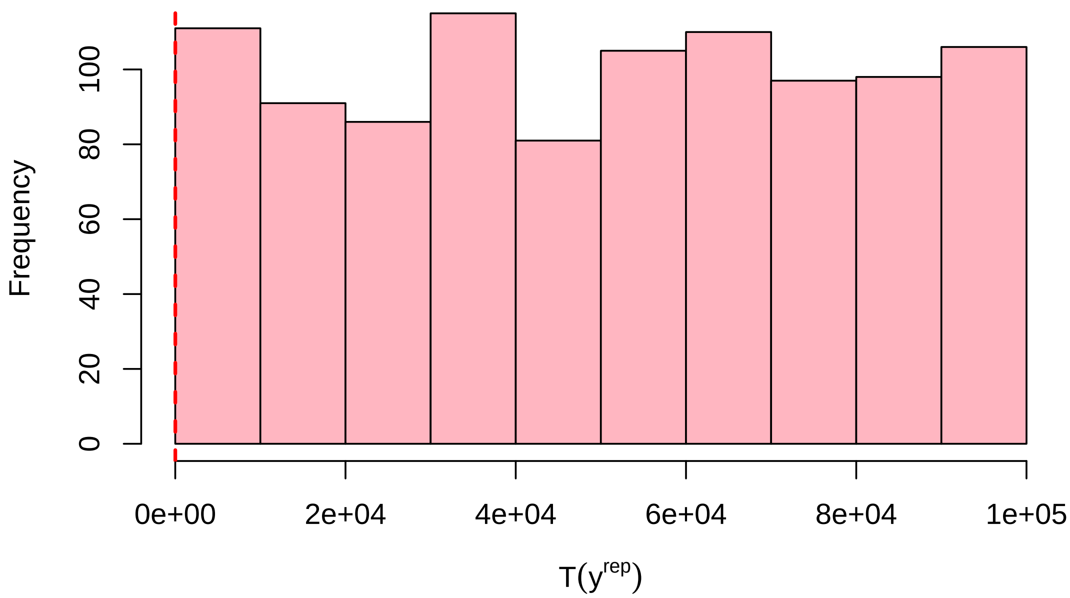
b. **Prior Predictive**:

$$
\begin{aligned}
p(y^{\text{rep}}) &= \int_{\Theta} p(y^{\text{rep}}|\theta) p(\theta) d\theta \\
&\propto \int_{\Theta} e^{-\frac{(y^{\text{rep}}-\theta)^2}{2}} I_{\theta \in [-A, A]} d\theta \\
&\propto \sqrt{\frac{\pi}{2}} \frac{1}{2A} \text{erf}\Big(\frac{\theta - y^{\text{rep}}}{\sqrt{2}}\Big) \\
&\propto \frac{1}{2A}\Big[\Phi\big(\theta - y^{\text{rep}}\big) - \Phi\big(-(\theta - y^{\text{rep}})\big)\Big]
\end{aligned}
$$

Given that $\lim_{\theta \to \infty} \Big[\Phi\big(\theta - y^{\text{rep}}\big) - \Phi\big(-(\theta - y^{\text{rep}})\big)\Big] \approx 1$, it follows that, for large values of $\theta \in [-A, A]$, $p(y^{\text{rep}}) \approx \frac{1}{2A}$.

While the posterior predictive, $p(y^{\text{rep}}|y)$ (Equation 6.2), averages over the parameter space with regards the posterior distribution, the prior predictive, $p(y^{\text{rep}})$, does so with regards to the prior distribution of θ. Thus, unlike the posterior predictive, the prior predictive distribution reflects our lack of knowledge regarding θ in the replication of new y. That is, values of y in the posterior predictive are generated based on our inferences about θ after having observed the data, while values of y in the prior predictive are generated according to what was known about θ before the data was observed. For this reason, drawing N samples of size n with uniformly distributed values of θ in range $[-A, A]$, will yield values of $T(y^{\text{rep}})$ that are similarly distributed between 0 and A, as shown in Figure 6 below.[8] Consequently, given the size of A relative to $T(y) = 8.1$, we obtain a p-value of ≈ 1, indicating that almost all of our simulated values of $T(y^{\text{rep}}) \geq T(y)$ (since necessarily, $T(y^{\text{rep}}) \geq 0$).

Figure 6: Histagram of Prior Predictive $T(y^{\text{rep}})$



c. Indeed it makes sense for the distribution obtained in part (a) to be far more representative of the observed data than the prior predictive distribution found in part (b). As mentioned previously, this can be attributed to the fact that the parameter used to simulate y in the posterior predictive, is itself largely determined by the data observed. On the other hand, the prior predictive, models y after values of θ that reflect our lack of knowledge about the parameter prior to observing the data. Had we used a more informative prior however, we likely would've obtained a less ambiguous prior predictive distribution of $T(y^{\text{rep}})$ with a p-value closer to the desired 0.5.

Problem 7: (Ch.6.9)

Model checking: Check the assumed model fitted to the rat tumor data in Section 5.3. Define some test quantities that might be of scientific interest, and compare them to their posterior predictive distributions.

Solution

Model: Let y_j be the number of rats with tumors out of n_j rats in experiment j, with known n_j, for $j = 1, 2, ..., 71$. Moreover, let parameter θ_j be the probability of developing a tumor (or the true proportion of rats with tumors) in the j^{th} experiment, with all θ_j's assumed to be independently sampled from a Beta distribution with hyperparameters (α, β). Using the same noninformative hyperprior discussed in Section 5.3 for this data, we obtain the following model.

$$y_j|\theta_j \sim \text{Binomial}(n_j, \theta_j) \quad \text{for } j = 1, 2, ..., 71$$
$$\theta_j|\alpha, \beta \sim \text{Beta}(\alpha, \beta) \quad \text{for } j = 1, 2, ..., 71$$
$$p(\alpha, \beta) \propto (\alpha, \beta)^{-5/2}$$

[8] Namely, $p(T(y^{\text{rep}})) \approx \frac{1}{A}$, since this test statistic is $\max_i|y_i^{\text{rep}}|$ rather than $\max_i(y_i^{\text{rep}})$.

Joint Posterior:

$$\begin{aligned} p(\theta,\alpha,\beta|y) &\propto p(\alpha,\beta)p(\theta|\alpha,\beta)p(y|\theta,\alpha,\beta) \\ &\propto p(\alpha,\beta)\prod_{j=1}^{71}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta_j^{\alpha-1}(1-\theta_j)^{\beta-1}\prod_{j=1}^{71}\theta_j^{y_j}(1-\theta_j)^{n_j-y_j} \end{aligned}$$

Conditional Posterior:

$$\begin{aligned} p(\theta|\alpha,\beta,y) &= p(\alpha,\beta)\prod_{j=1}^{71}\frac{\Gamma(\alpha+\beta+n_j)}{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)}\theta_j^{\alpha+y_j-1}(1-\theta_j)^{\beta+n_j-y_j-1} \\ &\propto \text{Beta}(\alpha+y,\beta+n-y) \end{aligned}$$

Marginal Posterior:

$$p(\alpha,\beta|y)\propto p(\alpha,\beta)\prod_{j=1}^{71}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)}{\Gamma(\alpha+\beta+n_j)}$$

To obtain posterior predictive replications, y^{rep}, we follow the simulation process outlined below.

1. Use the marginal posterior distribution of hyperparameters to obtain probabilities, $p(\alpha,\beta|y)$, for all pairwise values of (α,β) in some finite range[9]. Draw $N = 1,000$ pairs, (α_i,β_i), via grid sampling.
2. Draw $N = 1,000$ parameter values, θ_{ij}, from the conditional posterior distribution, $\theta_j|\alpha,\beta,y_j \sim \text{Beta}(\alpha+y_j,\beta+n_j-y_j)$, such that each draw corresponds to a unique sampled pair, (α_i,β_i). Repeat for each experiment $j = 1,2,...,71$ (or pair (y_j,n_j)).
3. Produce $N = 1,000$ posterior predictive replications, y_{ij}^{rep}, from $y_j^{\text{rep}}|\theta_j \sim \text{Binomial}(n_j,\theta_j)$, such that each value corresponds to a unique posterior draw, θ_{ij}. Repeat for each experiment $j = 1,2,...,71$ (or set of θ_j draws).

From this process, we obtain a total $N = 1,000$ simulations of replicated y observations for the $J = 71$ experiments, which we then use to evaluate our model by way of comparing a set of test statistics, $T(y^{\text{rep}})$, to their observed counterparts, $T(y^{\text{obs}})$. Results are given below.

Test Statistics

a. Number of 0's in y:

$$T_A(y) = \sum_{i=1}^{J} I_{y_j=0}$$

$$T_A(y^{\text{obs}}) = 14$$

$$p_B = p\big(T_A(y^{\text{rep}}) \geq T_A(y^{\text{obs}})\big) = \frac{1}{N}\sum_{i=1}^{N} I_{T_A(y_i^{\text{rep}})\geq T_A(y^{\text{obs}})} \approx 0.08$$

b. Mode[10] of y:

$$T_B(y) = k|_{f_k=\max\{f_{\min\{\vec{y}\}},\ldots,f_{\max\{\vec{y}\}}\}}$$

$$T_B(y^{\text{obs}}) = 0$$

$$p_B = p\big(T_B(y^{\text{rep}}) \geq T_B(y^{\text{obs}})\big) = \frac{1}{N}\sum_{i=1}^{N} I_{T_B(y_i^{\text{rep}})\geq T_B(y^{\text{obs}})} \approx 0.79$$

[9]Based on the observed hyperparameter space, we use $\alpha \in (0, 15]$ and $\beta \in (0, 50]$ for this problem.

[10]Where f_k is the frequency of value k, for all unique $k \in \vec{y}$.

c. Mean of y:

$$T_C(y) = \frac{1}{J}\sum_{j=1}^{J} y_j$$

$$T_C(y^{\text{obs}}) \approx 3.76$$

$$p_B = p\big(T_C(y^{\text{rep}}) \geq T_C(y^{\text{obs}})\big) = \frac{1}{N}\sum_{i=1}^{N} I_{T_C(y_i^{\text{rep}}) \geq T_C(y^{\text{obs}})} \approx 0.467$$

d. Maximum of y:

$$T_D(y) = \max_j\{y_1, y_2, ..., y_J\}$$

$$T_D(y^{\text{obs}}) = 16$$

$$p_B = p\big(T_D(y^{\text{rep}}) \geq T_D(y^{\text{obs}})\big) = \frac{1}{N}\sum_{i=1}^{N} I_{T_D(y_i^{\text{rep}}) \geq T_D(y^{\text{obs}})} \approx 0.601$$

e. Mean of $\frac{y}{n}$:

$$T_E(y) = \frac{1}{J}\sum_{j=1}^{J} \frac{y_j}{n_j}$$

$$T_E(y^{\text{obs}}) \approx 0.138$$

$$p_B = p\big(T_E(y^{\text{rep}}) \geq T_E(y^{\text{obs}})\big) = \frac{1}{N}\sum_{i=1}^{N} I_{T_E(y_i^{\text{rep}}) \geq T_E(y^{\text{obs}})} \approx 0.651$$

f. Maximum of $\frac{y}{n}$:

$$T_F(y) = \max_j\{\frac{y_1}{n_1}, \frac{y_2}{n_2}, ..., \frac{y_J}{n_J}\}$$

$$T_F(y^{\text{obs}}) \approx 0.375$$

$$p_B = p\big(T_F(y^{\text{rep}}) \geq T_F(y^{\text{obs}})\big) = \frac{1}{N}\sum_{i=1}^{N} I_{T_F(y_i^{\text{rep}}) \geq T_F(y^{\text{obs}})} \approx 0.875$$

Figure 7.a: Histogram of Posterior Predictive $T_A(y^{\text{rep}})$

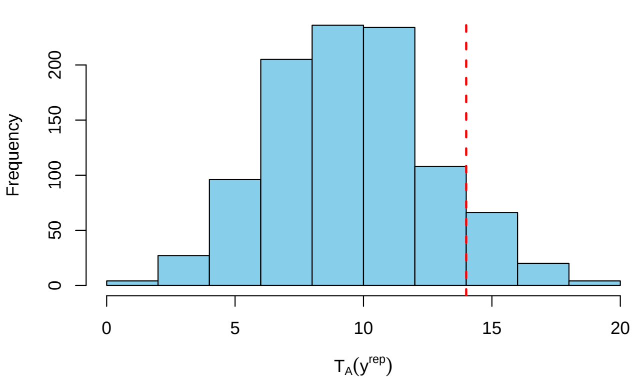
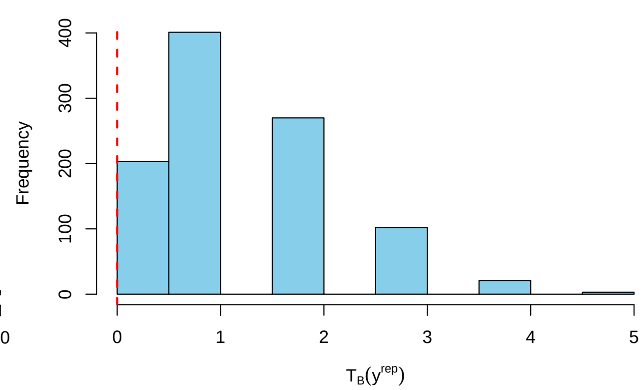


Figure 7.b: Histogram of Posterior Predictive $T_B(y^{\text{rep}})$



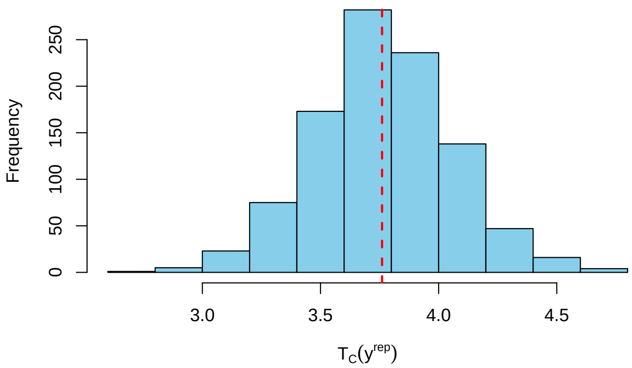


Figure 7.c: Histogram of Posterior Predictive $T_C(y^{rep})$

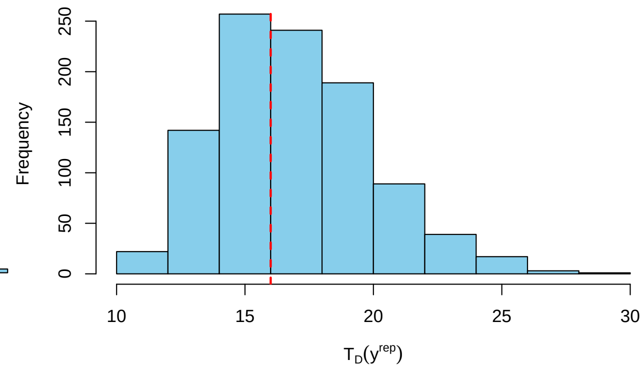


Figure 7.d: Histogram of Posterior Predictive $T_D(y^{rep})$

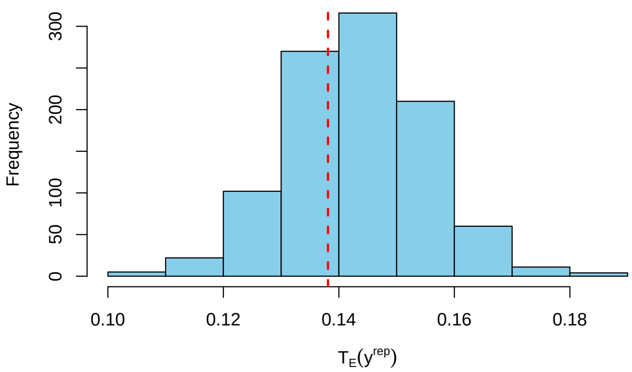


Figure 7.e: Histogram of Posterior Predictive $T_E(y^{rep})$

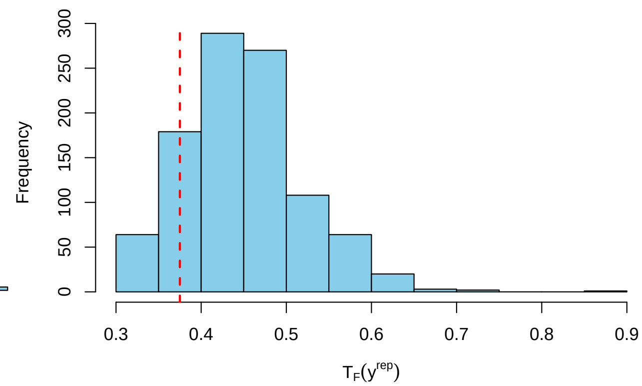


Figure 7.f: Histogram of Posterior Predictive $T_F(y^{rep})$

Problem 8: (Ch.7.5)

Power-transformed normal models: A natural expansion of the family of normal distributions, for all-positive data, is through power transformations, which are used in various contexts, including regression models. For simplicity, consider univariate data $y = (y_1, ..., y_n)$, that we wish to model as independent and identically normally distributed after transformation. Box and Cox (1964) propose the model, $y_i^{(\phi)} \sim \mathrm{N}(\mu, \sigma^2)$, where

$$y_i^{(\phi)} = \begin{cases} (y_i^{(\phi)} - 1)/\phi & \text{for } \phi \neq 0 \\ \log y_i & \quad\text{for } \phi = 0. \end{cases}$$

The parameterization in terms of $y_i^{(\phi)}$ allows a continuous family of power transformations that includes the logarithm as a special case. To perform Bayesian inference, one must set up a prior distribution for the parameters, (μ, σ, ϕ).

a. It seems natural to apply a prior distribution of the form $p(\mu, \log\sigma, \phi) \propto p(\phi)$, where $p(\phi)$ is a prior distribution (perhaps uniform) on ϕ alone. Unfortunately, this prior distribution leads to unreasonable results. Set up a numerical example to show why. (*Hint: Consider what happens when all the data points y_i are multiplied by a constant factor).

(b) Box and Cox (1964) propose a prior distribution that has the form $p(\mu, \sigma, \phi) \propto \dot{y}^{1-\phi} p(\phi)$, where $\dot{y} = (\prod_{i=1}^n y_i)^{1/n}$. Show that this prior distribution eliminates the problem in (a).

(c) Write the marginal posterior density, $p(\phi|y)$, for the model in (b).

(d) Discuss the implications of the fact that the prior distribution in (b) depends on the data.

(e) The power transformation model is used with the understanding that negative values of $y_i^{(\phi)}$ are not possible. Discuss the effect of the implicit truncation on the model.

See Pericchi (1981) and Hinkley and Runger (1984) for further discussion of Bayesian analysis of power transformations.

Solution

Problem 9: (Ch.7.6)

Fitting a power-transformed normal model: Table 7.3 gives short-term radon measurements for a sample of houses in three counties in Minnesota (see Section 9.4 for more on this example). For this problem, ignore the first-floor measurements (those indicated with asterisks in the table).

a. Fit the power-transformed normal model from Exercise 7.5 (b) to the basement measurements in Blue Earth County.
b. Fit the power-transformed normal model to the basement measurements in all three counties, holding the parameter ϕ equal for all three counties but allowing the mean and variance of the normal distribution to vary.
c. Check the fit of the model using posterior predictive simulations.
d. Discuss whether it would be appropriate to simply fit a lognormal model to these data.

Table 7.3: Excluding First-Floor Measurements

County	Radon Measurements (pCi/L)
Blue Earth	5.0, 13.0, 7.2, 6.8, 12.8, 9.5, 6.0, 3.8, 1.8, 6.9, 4.7, 9.5
Clay	12.9, 2.6, 26.6, 1.5, 13.0, 8.8, 19.5, 9.0, 13.1, 3.6
Goodhue	14.3, 7.6, 2.6, 43.5, 4.9, 3.5, 4.8, 5.6, 3.5, 3.9, 6.7

Solution

Code

```r
#### Q1 PART A ####
set.seed(47)

## Observed Data
y_j <- c(28, 8, -3, 7, -1, 1, 18, 12) # coaching effects in schools 1-8 (ordered)
sigma_j <- c(15, 10, 16, 11, 9, 11, 10, 18) # standard error in schools 1-8 (ordered)
sigma2_j <- sigma_j^2 # sampling variance in schools 1-8 (ordered)

## Simulating tau
tau <- seq(0, 30, 0.01) # random uniformly distributed values of tau

tau_posterior_density <- function(y_j, sigma2_j, tau){
  # posterior of tau assuming uniform prior
  inv_V_mu <- sum(1/(sigma2_j+tau^2))
  mu_hat <- sum(y_j/(sigma2_j+tau^2))/inv_V_mu
  a <- inv_V_mu^(-1/2)
  b <- (sigma2_j+tau^2)^(-1/2)
  c <- exp(-((y_j-mu_hat)^2)/(2*(sigma2_j+tau^2)))
  return(a*prod(b*c))
}

tau_posterior <- c() # probabilities for values of tau
for (i in 1:length(tau)){
  tau_i <- tau_posterior_density(y_j=y_j, sigma2_j=sigma2_j, tau[i])
  tau_posterior <- c(tau_posterior, tau_i)
}

tau_posterior_df <- data.frame(tau, tau_posterior)
ggplot(tau_posterior_df, aes(tau, tau_posterior)) +
  geom_line(color="red") +
  labs(title="Figure 1: Density of Marginal Posterior Tau",
       x=TeX(r"($\tau$)"),
       y="")

# inverse transform (cdf) sampling method
# approximating cdf empirically
norm_tau_post = tau_posterior/sum(tau_posterior) # normalizing probability values for tau
tau_post_cdf = cumsum(norm_tau_post) # cumulative sum
cdf_df = data.frame(tau, tau_post_cdf)
ggplot(cdf_df, aes(tau, tau_post_cdf)) +
  geom_line(color="blue") +
  labs(title="Figure 2: Approximate CDF of Marginal Posterior Tau",
       x=TeX(r"($\tau$)"),
       y="")

n <- 3000 # sample size

# tau draws
unif_vals <- runif(n, 0, 1) # uniform values
tau_sample <- c()
for (i in 1:n){
  tau_draw <- sum(tau_post_cdf <= unif_vals[i])*0.01 # sample from cdf
```

```r
  tau_sample <- c(tau_sample, tau_draw)
}

#hist(tau_sample)
```

```r
set.seed(47)

## Simulating mu (from mu|tau,y)

inv_V_mu <- function(sigma2_j, tau){
  sum(1/(sigma2_j+tau^2)) # inverse variance for mu|tau,y
}

mu_hat <- function(y_j, sigma2_j, tau){
  sum(1/(sigma2_j+tau^2)*y_j)/inv_V_mu(sigma2_j, tau) # mean for mu|tau,y
}

# mu draws
mu_sample <- c()
for (i in 1:n){
  mu_draw <- rnorm(n=1,
                   mean=mu_hat(y_j, sigma2_j, tau_sample[i]),
                   sd=inv_V_mu(sigma2_j, tau_sample[i])^(-1/2)) # inv_V ^(-1/2) = sd
  mu_sample <- c(mu_sample, mu_draw)
}

#hist(mu_sample)
```

```r
set.seed(47)

## Simulating theta_j's (from theta|mu,tau,y)

V_j <- function(sigma2_j, tau){
  if(tau==0){
    v_j <- round(1/sum(1/sigma2_j), 1) # posterior variance | tau=0, 16.6
  }
  else {
    v_j <- 1/((1/sigma2_j)+(1/tau^2))
  }
  return(v_j)
}

theta_hat_j <- function(y_j, sigma2_j, mu, tau){
  if(tau==0){
    theta_hat <- round(sum(y_j/sigma2_j)/sum(1/sigma2_j), 1) # pooled estimate, 7.7
  }
  else {
    theta_hat <- ((1/sigma2_j)*y_j+(1/tau^2)*mu)*V_j(sigma2_j, tau)
  }
  return(theta_hat)
}

# pairs of mu and tau
mu_tau_pairs <- cbind(mu_sample, tau_sample)
```

```r
# function to sample theta_j for specified j
theta_j <- function(j){
  theta_j_sample <- c()
  for(i in 1:n){
    theta_j_draw <- rnorm(n=1,
                          mean=theta_hat_j(y_j[j], sigma2_j[j],
                                           mu_tau_pairs[i,1], mu_tau_pairs[i,2]),
                          sd=V_j(sigma2_j[j], mu_tau_pairs[i,2])^(1/2))
    theta_j_sample <- c(theta_j_sample, theta_j_draw)
  }
  return(theta_j_sample)
}

# matrix of all parameter simulations
param_sims <- mu_tau_pairs

# applying function to each j and appending to matrix
for (i in 1:length(y_j)){
  param_sims <- cbind(param_sims, theta_j(i))
  colnames(param_sims)[i+2] <- paste0("theta_", i)
}

param_sims_df <- as.data.frame(param_sims)
head(param_sims_df)
```

```r
## (i) Probability that school j has the largest coaching effect

max_theta_j <- apply(param_sims_df[,3:10], 1, which.max) # max theta_j for each draw
table(max_theta_j) # counts

max_theta_j_prob <- round(table(max_theta_j)/n, 4)*100 # probabilities (in percentages)

max_theta_j_prob <- as.data.frame(max_theta_j_prob)
colnames(max_theta_j_prob) <- c("School", "Probability (%)")
max_theta_j_prob[,1] <- LETTERS[1:8]
```

```r
## (ii) Probability that school j has a larger effect than school k

theta_j_df <- param_sims_df[,3:10] # theta_j's df

# function to compare a column/school j to remaining columns/schools (k)
# returns the probabilities that school j's effect > each school k's
prob_comp <- function(theta_j_df, col_j){
  new_df <- theta_j_df[,-col_j]
  greater_than_probs <- c()
  for (k in 1:ncol(new_df)){
    greater_than_prob <- round(sum(col_j > new_df[,k])/nrow(new_df), 4)*100 # probability
    greater_than_probs <- c(greater_than_probs, greater_than_prob)
  }
  return(greater_than_probs)
}

# compiling data frame of probabilities
```

```r
theta_j_gt_prob <- c(NA, prob_comp(theta_j_df, 1)) # function applied to school A

# applying function to schools B-H
for (i in 2:ncol(theta_j_df)){
  probs <- prob_comp(theta_j_df, i)
  row_probs <- c(probs[1:(i-1)], NA, probs[-c(1:(i-1))])
  theta_j_gt_prob <- rbind(theta_j_gt_prob, row_probs)
}

theta_j_gt_prob <- as.data.frame(theta_j_gt_prob)
theta_j_gt_prob <- cbind(LETTERS[1:8], theta_j_gt_prob)
colnames(theta_j_gt_prob) <- c("j", LETTERS[1:8])
row.names(theta_j_gt_prob) <- 1:8
```

```r
#### Q1 PART B ####
set.seed(47)

## Simulating theta_j's (from theta|mu,tau,y, with tau=infty)

# function to sample unpooled theta_j for specified j
unpooled_theta_j <- function(j){
  unpooled_theta_j_sample <- c()
  for(i in 1:n){
    unpooled_theta_j_draw <- rnorm(n=1, mean=y_j[j], sd=sigma_j[j])
    unpooled_theta_j_sample <- c(unpooled_theta_j_sample, unpooled_theta_j_draw)
  }
  return(unpooled_theta_j_sample)
}

# matrix of all parameter simulations (unpooled)
param_sims_b <- mu_tau_pairs

# applying function to each j and appending to matrix
for (i in 1:length(y_j)){
  param_sims_b <- cbind(param_sims_b, unpooled_theta_j(i))
  colnames(param_sims_b)[i+2] <- paste0("theta_", i)
}

param_sims_b_df <- as.data.frame(param_sims_b)

## (i) Probability that school j has the largest coaching effect (given tau=infty)

max_unpooled_theta_j <- apply(param_sims_b_df[,3:10], 1, which.max)
table(max_unpooled_theta_j)

max_unpooled_theta_j_prob <- round(table(max_unpooled_theta_j)/n, 4)*100 # probabilities (%)

max_unpooled_theta_j_prob <- as.data.frame(max_unpooled_theta_j_prob)
colnames(max_unpooled_theta_j_prob) <- c("School", "Probability (%)")
max_unpooled_theta_j_prob[,1] <- LETTERS[1:8]
```

```r
## (ii) Probability that school j has a larger effect than school k (given tau=infty)

# compiling data frame of probabilities
```

```r
unpooled_theta_j_df <- param_sims_b_df[,3:10] # theta_j's df
unpooled_theta_j_gt_prob <- c(NA, prob_comp(unpooled_theta_j_df, 1)) # function form (a) applied to sch

# applying function to schools B-H
for (i in 2:ncol(unpooled_theta_j_df)){
  probs <- prob_comp(unpooled_theta_j_df, i)
  row_probs <- c(probs[1:(i-1)], NA, probs[-c(1:(i-1))])
  unpooled_theta_j_gt_prob <- rbind(unpooled_theta_j_gt_prob, row_probs)
}

unpooled_theta_j_gt_prob <- as.data.frame(unpooled_theta_j_gt_prob)
unpooled_theta_j_gt_prob <- cbind(LETTERS[1:8], unpooled_theta_j_gt_prob)
colnames(unpooled_theta_j_gt_prob) <- c("j", LETTERS[1:8])
row.names(unpooled_theta_j_gt_prob) <- 1:8
```

```r
#### Q1 PART D ####
set.seed(47)

## Simulating theta_j's (from theta|mu,tau,y, with tau=0)

# function to sample pooled theta_j
pooled_theta_j <- function(){
  pooled_theta_j_sample <- rnorm(n=n,
                                 mean=round(sum(y_j/sigma2_j)/sum(1/sigma2_j), 1),
                                 sd=round(1/sum(1/sigma2_j), 1))
  return(pooled_theta_j_sample)
}

# matrix of all parameter simulations (pooled)
param_sims_d <- mu_tau_pairs

# applying function to each j and appending to matrix
for (i in 1:length(y_j)){
  param_sims_d <- cbind(param_sims_d, pooled_theta_j())
  colnames(param_sims_d)[i+2] <- paste0("theta_", i)
}

param_sims_d_df <- as.data.frame(param_sims_d)

## (i) Probability that school j has the largest coaching effect (given tau=0)

max_pooled_theta_j <- apply(param_sims_d_df[,3:10], 1, which.max)
table(max_pooled_theta_j)

max_pooled_theta_j_prob <- round(table(max_pooled_theta_j)/n, 4)*100 # probabilities (%)

max_pooled_theta_j_prob <- as.data.frame(max_pooled_theta_j_prob)
colnames(max_pooled_theta_j_prob) <- c("School", "Probability (%)")
max_pooled_theta_j_prob[,1] <- LETTERS[1:8]

## (ii) Probability that school j has a larger effect than school k (given tau=0)

# compiling data frame of probabilities
pooled_theta_j_df <- param_sims_d_df[,3:10] # theta_j's df
```

```r
pooled_theta_j_gt_prob <- c(NA, prob_comp(pooled_theta_j_df, 1)) # function form (a) applied to school

# applying function to schools B-H
for (i in 2:ncol(pooled_theta_j_df)){
  probs <- prob_comp(pooled_theta_j_df, i)
  row_probs <- c(probs[1:(i-1)], NA, probs[-c(1:(i-1))])
  pooled_theta_j_gt_prob <- rbind(pooled_theta_j_gt_prob, row_probs)
}

pooled_theta_j_gt_prob <- as.data.frame(pooled_theta_j_gt_prob)
pooled_theta_j_gt_prob <- cbind(LETTERS[1:8], pooled_theta_j_gt_prob)
colnames(pooled_theta_j_gt_prob) <- c("j", LETTERS[1:8])
row.names(pooled_theta_j_gt_prob) <- 1:8
```

```r
## Comparing Estimates of theta_j

y_j # observed estimates

# posterior estimates from a (partial-pooling)
theta_js_a <- as.vector(round(apply(param_sims_df[,3:10], 2, mean), 2)) # mean
#round(apply(param_sims_df[,3:10], 2, var), 2) # within-group variances
var(round(apply(param_sims_df[,3:10], 2, mean), 2)) # between-group variance

# posterior estimates from b (no-pooling)
theta_js_b <- as.vector(round(apply(param_sims_b_df[,3:10], 2, mean), 2)) # mean
#round(apply(param_sims_b_df[,3:10], 2, var), 2) # within-group variances
var(round(apply(param_sims_b_df[,3:10], 2, mean), 2)) # between-group variance

# posterior estimates from d (complete pooling)
theta_js_d <- as.vector(round(apply(param_sims_d_df[,3:10], 2, mean), 2)) # mean
#round(apply(param_sims_d_df[,3:10], 2, var), 2) # within-group variances
var(round(apply(param_sims_d_df[,3:10], 2, mean), 2)) # between-group variance

theta_j_estimates <- as.data.frame(cbind(LETTERS[1:8],
                                         y_j,
                                         theta_js_d,
                                         theta_js_a,
                                         theta_js_b))

colnames(theta_j_estimates) <- c("School", "Observed",
                                 "Complete Pooling", "Partial-Pooling", "No-Pooling")

# between-group variances
theta_j_estimates <- rbind(theta_j_estimates,
                           c("Variance:",
                             round(var(theta_j_estimates[[2]]), 2),
                             round(var(theta_j_estimates[[3]]), 2),
                             round(var(theta_j_estimates[[4]]), 2),
                             round(var(theta_j_estimates[[5]]), 2)))

theta_j_estimates$School[9] <- cell_spec(theta_j_estimates$School[9], bold=TRUE)
```

```r
## Boxplots for Estimates
```

```r
# complete pooling
boxplot(param_sims_d_df[,3:10],
        names=LETTERS[1:8],
        ylim=c(-60, 70),
        col="lavender")
title(main="Figure 3.a: Coaching Effect Posterior Estimates, Complete Pooling",
      xlab=" ",
      ylab="Estimated Effect")

# partial-pooling
boxplot(param_sims_df[,3:10],
        names=c(LETTERS[1:8]),
        ylim=c(-60, 70),
        col="lightpink")
title(main="Figure 3.b: Coaching Effect Posterior Estimates, Partial-Pooling",
      xlab=" ",
      ylab="Estimated Effect")

# no-pooling
boxplot(param_sims_b_df[,3:10],
        names=LETTERS[1:8],
        ylim=c(-60, 70),
        col="lightblue")
title(main="Figure 3.c: Coaching Effect Posterior Estimates, No-Pooling",
      xlab=" ",
      ylab="Estimated Effect")
```

```r
#### Q4 PART C ####
set.seed(47)

## Observed Data (Residential Street Bike Routes)
y_j <- c(16, 9, 10, 13, 19, 20, 18, 17, 35, 55) # bikes
n_j <- c(16+58, 9+90, 10+48, 13+57, 19+103, 20+57, 18+86, 17+112, 35+273, 55+64) # total
bar_y_j <- y_j/n_j # observed proportions

## Data Frame for Stan
i <- seq(1, 10, 1)
total <- n_j
bike_df <- as.data.frame(cbind(i, total))

# ## Contour Grid
# grid_x <- log(alpha/beta)
# grid_y <- log(alpha+beta)

## Model in Stan
stan_bike_model <- stan_model("/Users/antonellabasso/Desktop/PHP2530/bikes.stan")

## Fitting Model
stan_bike_model_fit <- stan_bike_model %>%
  sampling(data=compose_data(bike_df), warmup=1000, iter=5000) %>% # tidybayes
  recover_types(bike_df) # tidybayes

## Marginal Posterior Draws
stan_bike_model_draws <- stan_bike_model_fit %>%
```

```
  spread_draws(alpha, beta, theta[i]) # tidybayes

## Contour Plot for (alpha, beta)
bike_post_ab <- ggplot(stan_bike_model_draws, aes(alpha, beta))
bike_post_ab +
  stat_density_2d() +
  labs(title=TeX(r"(Figure 4.a: Marginal Posterior Contour of $(\alpha, \beta)$)"),
       x=TeX(r"($\alpha$)"),
       y=TeX(r"($\beta$)"))
# contour(z=kde2d(stan_bike_model_draws$alpha, stan_bike_model_draws$beta)$z)

## Scatterplot of Posterior Draws of (alpha, beta)
bike_post_ab +
  geom_point() +
  labs(title=TeX(r"(Figure 4.b: Marginal Posterior Draws of $(\alpha, \beta)$)"),
       x=TeX(r"($\alpha$)"),
       y=TeX(r"($\beta$)"))
```

```
#### Q4 PART E ####

## Parameter Summary Statistics
# stan_bike_model_fit
param_stats <- round(as.data.frame(summary(stan_bike_model_fit)), 4)[1:12, 1:8]
param_stats[, c(1, 3:8)] <- round(param_stats[, c(1, 3:8)], 2)
colnames(param_stats) <- c("Mean", "Mean SE", "SD",
                           "2.5\\%", "25\\%", "50\\%", "75\\%", "97.5\\%")
```

```
#### Q6 PART A ####
set.seed(47)

n <- 100
bar_y <- 5.1

## Posterior Predictive Draws
t_stat <- rep(0, 1000)
for (i in 1:1000){
  theta_post <- rnorm(1, bar_y, sqrt(1/n))
  y_rep <- rnorm(100, theta_post, 1)
  t_stat[i] <- max(abs(y_rep))
}

## P-value for T(y)=8.1
sum(t_stat>8.1)/length(t_stat) # 0.126

## Histogram of T(y_rep)
hist(t_stat,
     col="lavender",
     main=TeX(r"(Figure 5: Histagram of Posterior Predictive $T(y^{rep})$)"),
     xlab=TeX(r"($T(y^{rep})$)"))
abline(v=8.1, lty=2, col="red", lwd=2)
```

```
#### Q6 PART B ####

set.seed(47)
```

```r
## Prior Predictive Draws
A <- 10^5
t_stat2 <- rep(0, 1000)
for (i in 1:1000){
  theta_prior <- runif(1, -A, A)
  y_rep <- rnorm(100, theta_prior, 1)
  t_stat2[i] <- max(abs(y_rep))
}

## P-value for T(y)=8.1
sum(t_stat2>8.1)/length(t_stat2) # 1

## Histogram of T(y_rep)
hist(t_stat2,
     col="lightpink",
     main=TeX(r"(Figure 6: Histagram of Prior Predictive $T(y^{rep})$)"),
     xlab=TeX(r"($T(y^{rep})$)"))
abline(v=8.1, lty=2, col="red", lwd=2)
```

```r
#### Q7 ####
set.seed(47)

## Observed Data (Rat Tumors)
y_j <- c(0, 0, 0, 0, 0, 0, 1, 1, 2, 2, 1, 5, 3, 2, 9, 4, 10, 4, 6, 5, 6, 0, 0, 0, 0, 1, 1, 2, 2, 2, 2,

n_j <- c(20, 20, 20, 19, 18, 18, 18, 18, 25, 20, 10, 49, 20, 13, 48, 20, 48, 19, 23, 19, 22, 20, 20, 20

bar_y_j <- y_j/n_j # observed means/proportions

## Marginal Posterior of (alpha, beta)
alpha <- seq(0.1, 15, 0.1)
beta <- seq(from=0.1, 50, 0.1)

marginal_post <- function(y=y_j, n=n_j, a, b){
  J <- length(y)
  hyperprior <- (a+b)^(-5/2)

  log_liks <- c()
  for (i in 1:J){
    num <- lgamma(a+b)+lgamma(a+y[i])+lgamma(b+n[i]-y[i])
    denom <- lgamma(a)+lgamma(b)+lgamma(a+b+n[i])
    log_lik <- num-denom
    log_liks <- c(log_liks, log_lik)
  }

  log_post <- log(hyperprior)+sum(log_liks)
  return(log_post)
}

## Posterior Draws of (alpha, beta) - via grid sampling
N <- 1000

alpha_grid <- rep(alpha, each=length(beta))
```

```r
beta_grid <- rep(beta, times=length(alpha))
alpha_beta_combs <- cbind(alpha_grid, beta_grid)

log_probs <- c()
for (i in 1:nrow(alpha_beta_combs)){
    log_probs[i] <- marginal_post(y=y_j,
                                  n=n_j,
                                  a=alpha_beta_combs[i, 1],
                                  b=alpha_beta_combs[i, 2])
}

alpha_beta_probs <- exp(log_probs-max(log_probs)) # posterior, avoiding underflow errors
alpha_beta_combs <- cbind(alpha_beta_combs, alpha_beta_probs)

index_samps <- sample(1:nrow(alpha_beta_combs), N, replace=TRUE, prob=alpha_beta_probs)
#hist(index_samps)

alpha_draws <- c()
beta_draws <- c()
for (i in 1:N){
  index <- index_samps[i]
  alpha_draws[i] <- alpha_beta_combs[index, 1]
  beta_draws[i] <- alpha_beta_combs[index, 2]
}

alpha_beta_draws <- cbind(alpha_draws, beta_draws) # final draws

## Contour Plot for (alpha, beta)
rat_post_ab <- ggplot(as.data.frame(alpha_beta_draws),
                      aes(alpha_draws, beta_draws))
rat_post_ab +
  stat_density_2d() +
  labs(title=TeX(r"(Marginal Posterior Contour of $(\alpha, \beta)$)"),
       x=TeX(r"($\alpha$)"),
       y=TeX(r"($\beta$)"))

## Scatterplot of Posterior Draws of (alpha, beta)
rat_post_ab +
  geom_point() +
  labs(title=TeX(r"(Marginal Posterior Draws of $(\alpha, \beta)$)"),
       x=TeX(r"($\alpha$)"),
       y=TeX(r"($\beta$)"))

## Conditional Posterior Draws of theta | alpha, beta
rat_param_sims <- alpha_beta_draws

for (i in 1:length(y_j)){
  theta_samp_j <- rbeta(N,
                        alpha_beta_draws[, 1]+y_j[i],
                        alpha_beta_draws[, 2]+n_j[i]-y_j[i])
  rat_param_sims <- cbind(rat_param_sims, theta_samp_j)
  colnames(rat_param_sims)[i+2] <- paste0("theta_", i)
}
```

```r
## Posterior Predictive Draws of y_rep | theta
predictive_post <- function(n, theta){
  y_rep_j <- rbinom(N, n, theta_draw)
  return(y_rep_j)
}

y_rep_1 <- rbinom(N, n_j[1], rat_param_sims[, 3])
for (i in 2:length(y_j)){
  y_rep_samp_j <- rbinom(N, n_j[i], rat_param_sims[, i+2])
  y_rep_1 <- cbind(y_rep_1, y_rep_samp_j)
  colnames(y_rep_1)[i] <- paste0("y_rep_", i)
}

rat_y_reps <- as.data.frame(y_rep_1)
```

```r
set.seed(47)

## Test Statistics for y

# (A) Zero Count
t_y_rep_0 <- apply(rat_y_reps, 1, function(x) sum(x==0))
t_y_0 <- sum(y_j==0) # T(y)=14
hist(t_y_rep_0,
     col="skyblue",
     main=TeX(r"(Figure 7.a: Histagram of Posterior Predictive $T_A(y^{rep})$)"),
     xlab=TeX(r"($T_A(y^{rep})$)"))
abline(v=t_y_0, lty=2, col="red", lwd=2)
sum(t_y_rep_0>t_y_0)/length(t_y_rep_0) # p-value=0.08

# (B) Mode
mode <- function(x){
  unique_x <- unique(x)
  mode <- unique_x[which.max(tabulate(match(x, unique_x)))]
  return(as.numeric(mode))
}
t_y_rep_mode <- apply(rat_y_reps, 1, mode)
t_y_mode <- mode(y_j) # T(y)=0
hist(t_y_rep_mode,
     col="skyblue",
     main=TeX(r"(Figure 7.b: Histagram of Posterior Predictive $T_B(y^{rep})$)"),
     xlab=TeX(r"($T_B(y^{rep})$)"))
abline(v=t_y_mode, lty=2, col="red", lwd=2)
sum(t_y_rep_mode>t_y_mode)/length(t_y_rep_mode) # p-value=0.79

# (C) Mean
t_y_rep_mean <- apply(rat_y_reps, 1, mean)
t_y_mean <- mean(y_j) # T(y)=3.760563
hist(t_y_rep_mean,
     col="skyblue",
     main=TeX(r"(Figure 7.c: Histagram of Posterior Predictive $T_C(y^{rep})$)"),
     xlab=TeX(r"($T_C(y^{rep})$)"))
abline(v=t_y_mean, lty=2, col="red", lwd=2)
sum(t_y_rep_mean>t_y_mean)/length(t_y_rep_mean) # p-value=0.467
```

```r
# (D) Maximum
t_y_rep_max <- apply(rat_y_reps, 1, max)
t_y_max <- max(y_j)
hist(t_y_rep_max,
     col="skyblue",
     main=TeX(r"(Figure 7.d: Histagram of Posterior Predictive $T_D(y^{rep})$)"),
     xlab=TeX(r"($T_D(y^{rep})$)"))
abline(v=t_y_max, lty=2, col="red", lwd=2)
sum(t_y_rep_max>t_y_max)/length(t_y_rep_max) # p-value=0.601

## Test Statistics for bar_y = y/n (proportions/means)
rat_bar_y_reps <- as.data.frame(t(apply(rat_y_reps, 1, function(x) x/n_j)))

# (E) Mean
t_bar_y_rep_mean <- apply(rat_bar_y_reps, 1, mean)
t_bar_y_mean <- mean(bar_y_j)
hist(t_bar_y_rep_mean,
     col="skyblue",
     main=TeX(r"(Figure 7.e: Histagram of Posterior Predictive $T_E(y^{rep})$)"),
     xlab=TeX(r"($T_E(y^{rep})$)"))
abline(v=t_bar_y_mean, lty=2, col="red", lwd=2)
sum(t_bar_y_rep_mean>t_bar_y_mean)/length(t_bar_y_rep_mean) # p-value=0.651

# (F) Maximum
t_bar_y_rep_max <- apply(rat_bar_y_reps, 1, max)
t_bar_y_max <- max(bar_y_j)
hist(t_bar_y_rep_max,
     col="skyblue",
     main=TeX(r"(Figure 7.f: Histagram of Posterior Predictive $T_F(y^{rep})$)"),
     xlab=TeX(r"($T_F(y^{rep})$)"))
abline(v=t_bar_y_max, lty=2, col="red", lwd=2)
sum(t_bar_y_rep_max>t_bar_y_max)/length(t_bar_y_rep_max) # p-value=0.875
```