PHP 2530: BAYESIAN STATISTICAL METHODS
HOMEWORK II SOLUTIONS

NICK LEWIS

Problem 1: BDA 3rd Ed. 3.2

Comparison of two multinomial observations: on September 25, 1988, the evening of a presidential campaign debate, ABC News conducted a survey of registered voters in the United States; 639 persons were polled before the debate, and 639 different persons were polled after. The results are displayed in the table below. Assume the surveys are independent simple random samples from the population of registered voters. Model the data with two different multinomial distributions. For j = 1, 2, let α_j be the proportion of voters who preferred Bush, out of those who had a preference for either Bush or Dukakis at the time of survey j. Plot a histogram of the posterior density for $\alpha_2 - \alpha_1$. What is the posterior probability that there was a shift toward Bush?

Table 1. Number of respondents in each preference category from ABC News pre– and post–surveys in 1988.

Survey	Bush	Dukakis	No Opinions/other	Total
pre–debate	294	307	38	639
post–debate	288	332	19	639

Solution

Define $\mathbf{y_j} = (y_{j1}, y_{j2}, y_{j3})$ to be a vector at time j, where $j = \{pre-, post-\}$, of the number of respondents in each preference category.

$$\begin{aligned} \mathbf{y_j} \mid \boldsymbol{\theta_j} &\sim Multinomial(639, \theta_{j1}, \theta_{j2}, \theta_{j3}) \\ \boldsymbol{\theta}_j &\sim Dirichlet(\gamma_1, \gamma_2, \gamma_3) \end{aligned}$$

Our goal is to estimate α_j, the proportion of voters who preferred Bush to Dukakis out of those who had a preference for either Bush or Dukakis at the time of the survey. In other words, our estimands are $\alpha_j = \frac{\theta_{j1}}{\theta_{j1}+\theta_{j2}}$.

We can simplify this problem greatly by setting $\gamma_1 = \gamma_2 = \gamma_3 = 1$ which is essentially a uniform prior. As the dirichlet prior is a conjugate prior, we find that the posterior distribution is simply:

$$\boldsymbol{\theta}_j|\mathbf{y_j} \sim Dirichlet(y_{j1} + 1, y_{j2} + 1, y_{j3} + 1)$$

In the context of this problem, this provides us with the following two distributions:

$$\begin{aligned} \boldsymbol{\theta}_{pre}|\mathbf{y_{pre}} &\sim Dirichlet(295, 308, 39) \\ \boldsymbol{\theta}_{post}|\mathbf{y_{post}} &\sim Dirichlet(289, 333, 20) \end{aligned}$$

At this point we could stop and simply sample from $\boldsymbol{\theta}_j|\mathbf{y_j}$ to calculate α_j, but we can go further by finding the posterior distribution of α_j.

Recall that since θ_j is the parameter in the multinomial distribution, it is constrained by the summation $\sum_{k=1}^{3} \theta_{jk} = 1$. This means our posterior function takes the form

$$\begin{aligned} p(\boldsymbol{\theta}_j|\mathbf{y_j}) &\propto \prod_{k=1}^{3} \theta_{jk}^{y_{jk}} \\ &\propto \theta_{j1}^{y_{j1}} \theta_{j2}^{y_{j2}} (1-\theta_{j1}-\theta_{j2})^{y_{j3}} \end{aligned}$$

Letting $\alpha_j = \frac{\theta_{j1}}{\theta_{j1}+\theta_{j2}}$, and $\beta_j = \theta_{j1} + \theta_{j2}$ we find the jacobian to be

$$J(\alpha, \beta) = \left| \begin{pmatrix} \beta & \alpha \\ -\beta & 1-\alpha \end{pmatrix} \right| = 1$$

$$\begin{aligned} p(\alpha, \beta|\mathbf{y_j}) &\propto |J(\alpha, \beta)| p(\boldsymbol{\theta}_j|\mathbf{y_j}) \\ &= (\alpha\beta)^{y_{j1}} (\beta - \alpha\beta)^{y_{j2}} (1-\beta)^{y_{j3}} \\ &= \alpha^{y_{j1}} (1-\alpha)^{y_{j2}} \beta^{y_{j1}+y_{j2}} (1-\beta)^{y_{j3}} \end{aligned}$$

We can see that $p(\alpha, \beta|y)$ breaks own into two independent beta distributions. We can then sample from α directly and bypass the need to calculate the fraction

$$\begin{aligned} \alpha|\mathbf{y_j} &\sim Beta(y_{j1}+1, y_{j2}+1) \\ \beta|\mathbf{y_j} &\sim Beta(y_{j1}+y_{j2}+1, y_{j3}+1) \end{aligned}$$

$$\begin{aligned} \alpha_{pre}|\mathbf{y_{pre}} &\sim Beta(295, 308) \\ \alpha_{post}|\mathbf{y_{post}} &\sim Beta(289, 333) \end{aligned}$$

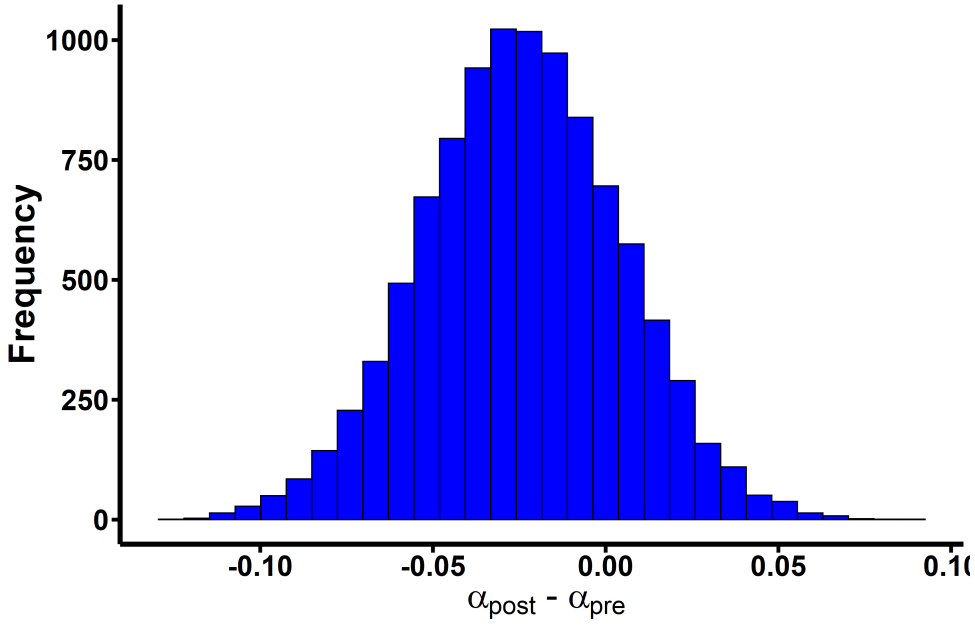


Figure 1. Histogram of the Difference in Proportions of Respondents who preferred Bush over Dukakis, pre- and post-debate.

A shift toward Bush would be indicated by $\alpha_2 - \alpha_1 > 0$. Our simulation provides us with 1,872 out of 10,000 draws indicating a shift towards Bush, or a $\approx 20\%$ posterior probability of a shift toward Bush.

Problem 2: BDA 3rd Ed. 3.3

Estimation from two independent experiments: an experiment was performed on the effects of magnetic fields on the flow of calcium out of chicken brains. Two groups of chickens were involved: a control group of 32 chickens and an exposed group of 36 chickens. One measurement was taken on each chicken, and the purpose of the experiment was to measure the average flow μ_c in untreated (control) chickens and the average flow μ_t in treated chickens. The 32 measurements on the control group had a sample mean of 1.013 and a sample standard deviation of 0.24. The 36 measurements on the treatment group had a sample mean of 1.173 and a sample standard deviation of 0.20.

(a) Assuming the control measurements were taken at random from a normal distribution with mean μ_c and variance σ_c^2, what is the posterior distribution of μ_c? Similarly, use the treatment group measurements to determine the marginal posterior distribution of μ_t. Assume a uniform prior distribution on $(\mu_c, \mu_t, log(\sigma_c), log(\sigma_t))$.

Solution:

We will solve this problem for $(\mu_t, log(\sigma_t))$. The results will hold for $(\mu_c, log(\sigma_c))$. Given that both are generated by a normal distribution, we see that our likelihood function is simply:

$$p(y \mid \mu_t, \mu_c, \sigma_t, \sigma_c) = \Big(\prod_{j=1}^{32} p(y_{c,j} \mid \mu_c, \sigma_c^2)\Big)\Big(\prod_{k=1}^{36} p(y_{t,k} \mid \mu_t, \sigma_t^2)\Big)$$

Before we continue, note that the treatment and control groups are separated in the likelihood. This will indicate that the posterior will break into two independent posteriors. To do so though, we will convert the uniform prior $p(\mu_t, \mu_c, log(\sigma_t), log(\sigma_c))$ to $p(\mu_t, \mu_c, \sigma_t^2, \sigma_c^2)$. A simple change of variables yields a diagonal jacobian matrix whose determinant is $\frac{1}{\sigma_t^2}\frac{1}{\sigma_c^2}$. This gives us that $p(\mu_t, \mu_c, \sigma_t^2, \sigma_c^2) \propto \frac{1}{\sigma_t^2}\frac{1}{\sigma_c^2}$ yielding a posterior density:

$$\begin{aligned} p(\mu_t, \mu_c, \sigma_t^2, \sigma_c^2 \mid y) &\propto p(y \mid \mu_t, \mu_c, \sigma_t^2, \sigma_c^2) p(\mu_t, \mu_c, \sigma_t^2, \sigma_c^2) \\ &\propto \Big(\prod_{j=1}^{32} p(y_{c,j} \mid \mu_c, \sigma_c^2)\Big)\Big(\prod_{k=1}^{36} p(y_{t,k} \mid \mu_t, \sigma_t^2)\Big)\frac{1}{\sigma_t^2}\frac{1}{\sigma_c^2} \end{aligned}$$

The details for the integration are in the BDA book, but the end result we yield are two marginal posteriors for the means of the form $t_{n-1}(\bar{y}, s^2/n)$:

$$\begin{aligned} \mu_c \mid y &\sim t_{31}(1.013, (0.24)^2/32 = 0.0018) \\ \mu_t \mid y &\sim t_{35}(1.173, (0.20)^2/36 = 0.0011) \end{aligned}$$

(b) What is the posterior distribution for the difference, $\mu_t - \mu_c$? To get this, you may sample from the independent t distributions you obtained in part (a) above. Plot a histogram of your samples and give an approximate 95% posterior interval for $\mu_t - \mu_c$.

Solution:

A t-distribution is a location scale family, which means we can write out the marginal distribution as

$$\frac{\mu - \bar{y}}{s/\sqrt{n}} \mid y \sim t_{n-1}$$
$$t_{n-1}(\bar{y}, s^2/n) = \frac{s}{\sqrt{n}} t_{n-1} + \bar{y}$$

This is particularly useful since the sampling function for a t distribution in R and python doesn't account for noncentrality. The histogram from the resulting draws is below:

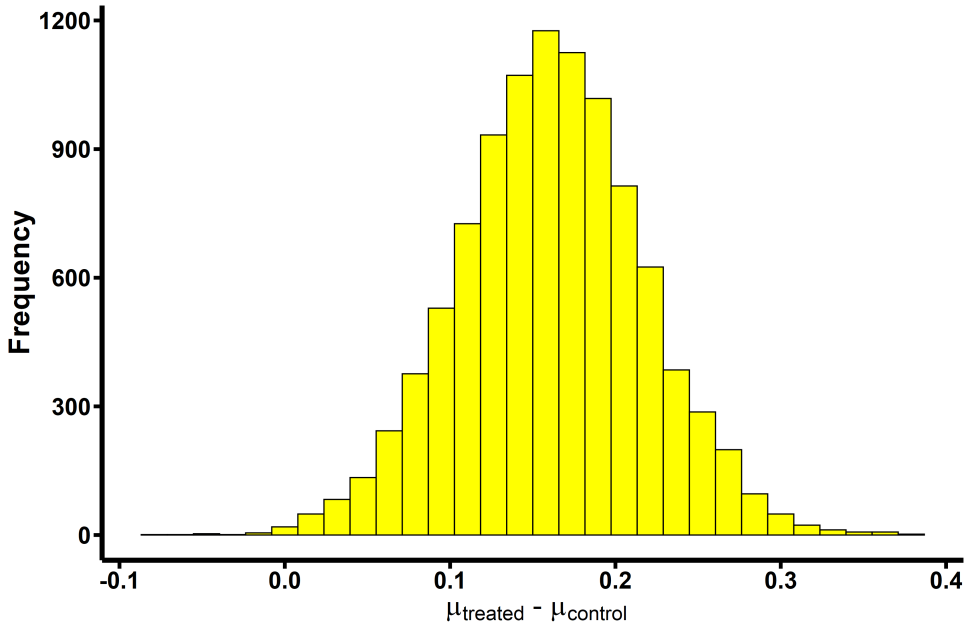


Figure 2. Histogram of Difference in Means between Treatment Group and Control Group

Our 95% Posterior Interval for $\mu_t - \mu_c$ is $[0.049, 0.269]$

Problem 3: BDA 3rd Ed. 3.5

Rounded data: it is a common problem for measurements to be observed in rounded form (for a review, see Heitjan, 1989). For a simple example, suppose we weigh an object five times and measure weights, rounded to the nearest pound, of 10, 10, 12, 11, 9. Assume the unrounded measurements are normally distributed with a noninformative prior distribution on the mean μ and variance σ^2.

(a) Give the posterior distribution for (μ, σ^2) obtained by pretending that the observations are exact unrounded measurements.

Solution:

Assuming our data is not rounded, we consider it to be generated from a normal distribution.

$$y_j \mid \mu, \sigma^2 \sim N(\mu, \sigma^2)$$
$$p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$$

Our posterior then takes the form

$$
\begin{aligned}
p(\mu, \sigma^2 \mid y) &\propto \frac{1}{\sigma^2}\Big(\prod_{j=1}^{N} p(y_j \mid \mu, \sigma^2)\Big) \\
&= \frac{1}{\sigma^2}\Big(\prod_{j=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_j-\mu)^2}{2\sigma^2}})\Big) \\
&\propto \Big(\frac{1}{\sigma^2}\Big)^{\frac{N}{2}+1} e^{-\sum_{j=1}^{N}\frac{(y_j-\mu)^2}{2\sigma^2}}
\end{aligned}
$$

The calculations are presented in Chapter 3 Section 3 of the Bayesian Data Analysis, 3rd Edition, text, but the marginal posteriors come out to be:

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

For the purposes of this exercise, using either σ or $log(\sigma)$ is fine. For those who choose to use $log(\sigma)$, here is a slight validation of why the posterior is of the form it appears to be. Note that $\sigma = e^{log(\sigma)}$ so when one performs the change of variables from σ^2 to $log(\sigma)$, the posterior takes the following form:

$$
\begin{aligned}
p(\mu, log(\sigma) \mid y) &\propto \Big(\prod_{j=1}^{N} p(y_j \mid \mu, (e^{log(\sigma)})^2)\Big) p(\mu, log(\sigma)) \\
&= \Big(\prod_{j=1}^{N} \frac{1}{\sqrt{2\pi(e^{log(\sigma)})^2}} e^{-\frac{(y_j-\mu)^2}{2(e^{log(\sigma)})^2}})\Big) \\
&\propto \Big(\frac{1}{e^{log(\sigma)}}\Big)^{N} e^{-\sum_{j=1}^{N}\frac{(y_j-\mu)^2}{2(e^{log(\sigma)})^2}}
\end{aligned}
$$

You will see in my code for both (a) and (b) that the functions I use begin by making this transformation. Now you are aware as of why.

(b) Give the correct posterior distribution for (μ, σ^2) treating the measurements as rounded.

Solution:

Assuming our data is rounded now we have to approach the problem slightly differently. Call z_j to be the true value of the measurement. The first question one ought to ask is "what is y_j rounded from?" If one were to round to the nearest whole integer, there are two options; it was either rounded up or down. If it was rounded down, then $z_j < y_j + 0.5$. If it was rounded up, then $y - 0.5 \leqslant z_j$. In other words, we can be assured that the true values z_j lies in the following interval $[y_j - 0.5, y_j + 0.5)$. Therefore the pdf associated with z_j is simply the cdf described below.

$$
\begin{aligned}
p(y_j \mid \mu, \sigma^2) &= \int_{y_j-0.5}^{y_j+0.5} p(z_j \mid \mu, \sigma^2) dz_j \\
&= \Phi_{\mu,\sigma}(y_j - 0.5 \leqslant z_j < y_j + 0.5) \\
&= \Phi(\frac{y_j + 0.5 - \mu}{\sigma}) - \Phi(\frac{y_j - 0.5 - \mu}{\sigma})
\end{aligned}
$$

Choosing the prior $p(log(\sigma))\propto 1$ We then find that posterior is of the form :

$$p(\mu, log(\sigma) \mid y)\propto\Big(\prod_{j=1}^{N}\Phi(\frac{y_j+0.5-\mu}{e^{log(\sigma)}})-\Phi(\frac{y_j-0.5-\mu}{e^{log(\sigma)}})\Big)$$

(c) How do the incorrect and correct posterior distributions differ? Compare means, variances, and contour plots.

Solutions:

We will start this problem off by comparing the contour plots of the two posterior distributions.

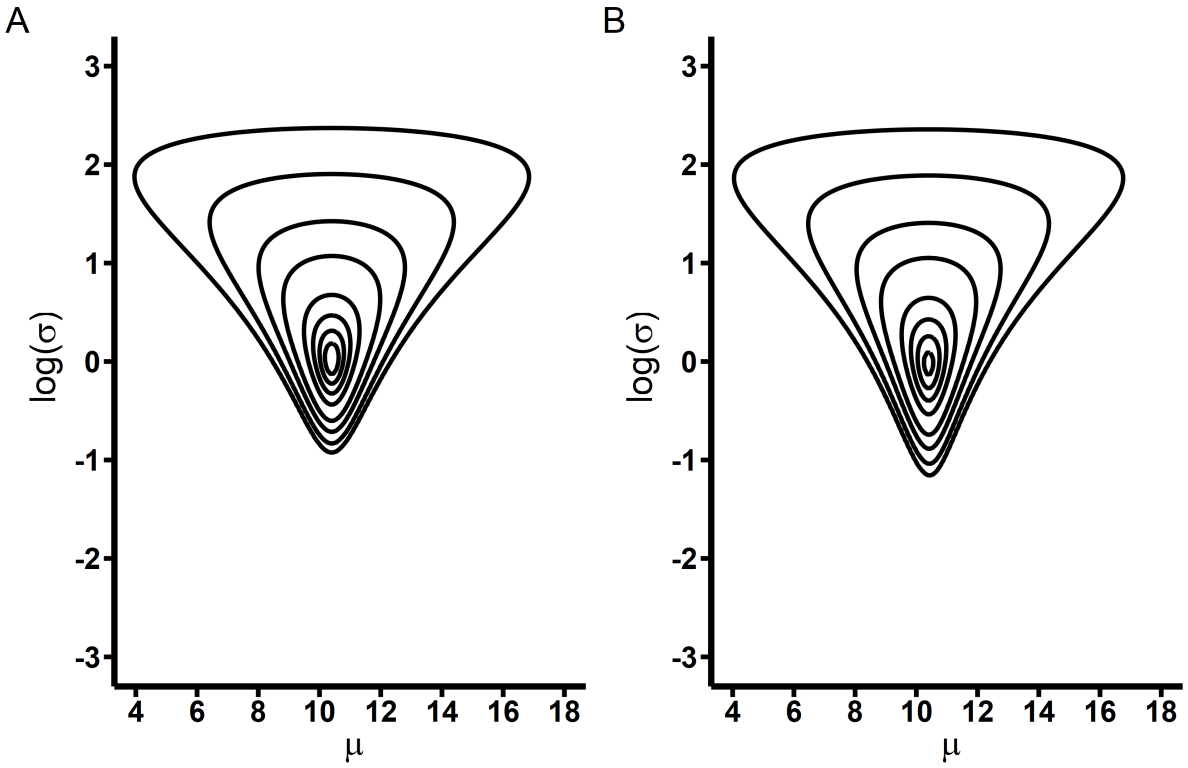


Figure 3. Plot of the Posterior Distributions when we consider the data to be unrounded (A), and rounded (B). The contour lines are based upon the 0.01%, 0.10%, 1%, 5%, 25%, 50%, 75%, and 95% quantiles.

The two posteriors look remarkably similar when looking at the contour plots. The most notable difference is that the Rounded Posterior appears to be more concentrated than the Unrounded.

	Mean	Variance	2.5	50%	95%
Unrounded Posterior					
μ	10.41	0.52	9.03	10.41	11.87
σ	1.41	0.53	0.68	1.23	3.21
Rounded Posterior					
μ	10.41	0.48	9.02	10.36	11.79
σ	1.36	0.53	0.61	1.18	3.19

In terms of summary statistics, the two posteriors have almost negligible differences in their parameters suggesting that accounting for rounding provides no meaningful change.

(d) Let $z = (z_1, \ldots, z_5)$ be the original, unrounded measurements corresponding to the five observations above. Draw simulations from the posterior distribution of z. Compute the posterior mean of $(z_1 - z_2)^2$.

Solutions:

This problem asks us to find the posterior mean of the difference of the square of the original unrounded measurements z_1 and z_2. The reason these are chosen is that both measurements were close enough to either be rounded up or rounded down to 10. This is essentially finding the mean squared error.

We assume our measurements come from a likelihood which is the difference of two Normal CDF's. In order to generate draws from this distribution, we need to use the inverse-cdf method which is described in the paragraph below.

We want to generate samples $X \sim P_X$, where P_X is the probability distribution for X and $F_X(x)$ is the cdf. However, sampling from this distribution P_X is difficult. We first note that F_X is a monotonically increasing, right continuous function bounded between 0 and 1. Given that the function is bounded in the unit interval, if we assume for $U \sim Unif(0,1)$ that $U = F_X(X)$, we can easily find samples X by taking the inverse of the cdf as it is quite easy to sample from the uniform distribution. In other words, $F_X^{-1}(U) = X$.

In the context of this problem, each true measurement $z_j \mid \mu, \sigma$ is generated from the difference of two normal cdfs. The scheme to get samples of z_j is as follows:

(1) Plug in values of (μ, σ) into the likelihood $p(z_j \mid \mu, \sigma) = \Phi(\frac{y_j+0.5-\mu}{\sigma}) - \Phi(\frac{y_j-0.5-\mu}{\sigma})$. This will provide a vector of probabilities for z_j

(2) Note that for an interval [a,b], where $a < b$, $F_X : [a,b] \to [F_X(a), F_X(b)]$. By letting $a = y_j - 0.5$, and $b = y_j + 0.5$, we can find all possible values that y_j could have been rounded up or down from by using the inverse cdf method

(3) The inverse cdf method boils down to $F_X^{-1}([F_X(a) + (F_X(b) - F_X(a)) * U) = [a,b]$, where $U \sim Unif(0,1)$

Using the above scheme, we find that the posterior mean $(z_1 - z_2)^2 \approx 0.16$

Problem 4: BDA 3rd Ed. 3.8

Analysis of proportions: a survey was done of bicycle and other vehicular traffic in the neighborhood of the campus of the University of California, Berkeley, in the spring of 1993. Sixty city blocks were selected at random; each block was observed for one hour, and the numbers of bicycles and other vehicles traveling along that block were recorded. The sampling was stratified into six types of city blocks: busy, fairly busy, and residential streets, with and without bike routes, with ten blocks measured in each stratum. The table below displays the number of bicycles and other vehicles recorded in the study. For this problem, restrict your attention to the first four rows of the table: the data on residential streets.

Table 2. Counts of bicycles and other vehicles in one hour in each of 10 city blocks in each of six categories. (The data for two of the residential blocks were lost.) For example, the first block had 16 bicycles and 58 other vehicles, the second had 9 bicycles and 90 other vehicles, and so on. Streets were classified as 'residential,' 'fairly busy,' or 'busy' before the data were gathered.

Bike Route	Proportion of Bicycles	No Bike Route	Proportion of Bicycles
y_1	16/(16+58)	z_1	12/(12+113)
y_2	9/(9+90)	z_2	1/(1+18)
y_3	10/(10+48)	z_3	2/(2+14)
y_4	13/(13+57)	z_4	4/(4+44)
y_5	19/(19+103)	z_5	9/(9+208)
y_6	20/(20+57)	z_6	7/(7+67)
y_7	18/(18+86)	z_7	9/(9+29)
y_8	17/(17+112)	z_8	8/(8+154)
y_9	35/(35+273)		
y_{10}	55/(55+64)		

(a) Let $y_1, ..., y_{10}$ and $z_1, ..., z_8$ be the observed proportion of traffic that was on bicycles in the residential streets with bike lanes and with no bike lanes, respectively (so $y_1 = 16/(16+58)$ and $z_1 = 12/(12+113)$, for example). Set up a model so that the $y_i's$ are independent and identically distributed given parameters θ_y and the $z_i's$ are independent and identically distributed given parameters θ_z.

Solution:

There are multiple ways to approach this problem. I will present three methods throughout this solution. You can find code for each method in the appendix.

Approach 1

The first approach involves recognizing that the proportions lie strictly between 0 and 1. Therefore a natural way to model the data is to use a beta distribution.

$$y_j|\alpha_y, \beta_y \sim Beta(\alpha_y, \beta_y) \text{ for } j = 1, 2, \ldots, 10$$
$$z_k|\alpha_z, \beta_z \sim Beta(\alpha_z, \beta_z) \quad \text{for } k = 1, 2, \ldots, 8$$

One might notice that this approach leaves us without the parameter of interest, θ_y and θ_z. We can simply say $\theta_y = \{\alpha_y, \beta_y\}$, and $\theta_z = \{\alpha_z, \beta_z\}$.

Approach 2

The second approach involves looking at the number of bicycles and the total number of observed traffic rather than the proportion. Define $y_j = \frac{b_j^y}{b_j^y + v_j^y}$ where b_j^y is the number of bicycles on street j, and v_j^y the number of non-bicycle vehicles. Call $n_j^y = b_j^y + v_j^y$ to be the total number of vehicles seen on street j. It is similarly defined for z.

The reason we do this is that working with the number of bicycles rather than the proportion allows us much more leeway. There are more probability distributions that deal with count data. For this approach we make the assumption that the total number of vehicles seen is fixed, but the number of bicycles is not. We can then model $b_j^y \mid \theta_y \sim Bin(n_j^y, \theta_y)$.

$$b_j^y \mid \theta_y \sim Bin(n_j^y, \theta_y) \text{ for } j = 1, 2, \ldots, 10$$
$$b_k^z \mid \theta_z \sim Bin(m_k^z, \theta_z) \text{ for } k = 1, 2, \ldots, 8$$

Approach 3

The third approach is similar in the setup of the second approach, except now we model both b_j^y and v_j^y. We no longer assume that we observe a fixed amount of vehicles in total, but rather we count the number of bicycles and non-bicycles that we see. In this regard the Poisson distribution becomes a natural formulation for the resulting models described below. The model for residential streets with bike lanes is:

$$b_j^y | \theta_y^b \sim Poisson(\theta_y^b) \text{ if } j = 1, 2, \ldots, 10$$
$$v_j^y | \theta_y^v \sim Poisson(\theta_y^v) \text{ if } j = 1, 2, \ldots, 10$$

The model for residential streets without bike lanes is

$$b_j^z | \theta_z^b \sim Poisson(\theta_z) \text{ if } k = 1, 2, \ldots, 8$$
$$v_j^z | \theta_z^v \sim Poisson(\theta_z^v) \text{ if } k = 1, 2, \ldots, 8$$

For this problem we introduce four new parameters to estimate, θ_y^b, θ_y^v, θ_z^b and θ_z^v. These parameters are simply the rates in which those vehicles occur on the streets, and we can then estimate θ_y and θ_z as:

$$\theta_y = \frac{\theta_y^b}{\theta_y^b + \theta_y^v}$$
$$\theta_z = \frac{\theta_z^b}{\theta_z^b + \theta_z^v}$$

Of course though, for what I will do throughout this problem, you will see the calculation of θ_y and θ_z is unnecessary.

(b) Set up a prior distribution that is independent in θ_y and θ_z.

Solution:

Approach 1: Beta Distribution

For this problem, I won't be using an improper prior. In this approach we assume that the data follows a beta distribution with parameters (α_y, β_y). A prior on these two parameters needs to be restricted to positive values since those are the only ones that make the beta distribution proper. Therefore I propose we use weakly informative priors via a uniform distribution on both α_y and β_y.

$$p(\alpha_y, \beta_y) \propto I_{\alpha_y \in [\epsilon, 100], \beta_y \in [\epsilon, 100]}$$
$$p(\alpha_z, \beta_z) \propto I_{\alpha_z \in [\epsilon, 100], \beta_z \in [\epsilon, 100]}$$

These grids will more or less allow us some leeway in searching around the distribution space. For both models we choose $\epsilon = 0.001$.

Approach 2: Binomial Distribution

For ease of calculation we will assume that our priors follow a beta distribution. We choose our priors in a similar fashion to the schematic laid out in the first approach by using a Beta which is centered around 0.50, but with sufficient spread.

$$\theta_y \sim Beta(5,5)$$
$$\theta_z \sim Beta(5,5)$$

Another valid approach would be to use a hierarchical model where we put priors on the parameters of $\theta_y \mid \alpha_y, \beta_y$, but this approach appears in a later chapter so I will restrict this analysis to the simple case.

Approach 3: Poisson Distribution

The third approach follows similar logic as in the first and second. We will use Gamma Prior's due to their conjugacy. Given the form of the kernel of the Gamma distribution, we opt to set the rate parameter equal to 1. This is because a large rate parameter will cause the mass to be concentrated in too narrow of an area as for large values, the mass will practically be zero. For these priors, since I have to model vehicles as well, I opt to assume that out of 100 recorded vehicles on residential streets with bike lanes, 15 are bicycles while the other 85 are motorized vehicles. This yields the following four priors.

$$\theta_y^b \sim Gamma(15,1)$$
$$\theta_y^v \sim Gamma(85,1)$$
$$\theta_z^b \sim Gamma(15,1)$$
$$\theta_z^v \sim Gamma(85,1)$$

(c) Determine the posterior distribution for the parameters in your model and draw 1000 simulations from the posterior distribution. (Hint: θ_y and θ_z are independent in the posterior distribution, so they can be simulated independently.)

Solution:

Approach 1: Beta Distribution

The two posteriors for θ_y and θ_z are of the following form:

$$p(\alpha_y, \beta_y|y) \propto \prod_{j=1}^{10} y_j^{\alpha_y - 1}(1-y_j)^{\beta_y - 1} I_{\alpha_y \in [\epsilon, 100], \beta_y \in [\epsilon, 100]}$$
$$p(\alpha_z, \beta_z|z) \propto \prod_{k=1}^{8} z_k^{\alpha_z - 1}(1-z_k)^{\beta_z - 1} I_{\alpha_z \in [\epsilon, 100], \beta_z \in [\epsilon, 100]}$$

The posteriors are plotted below

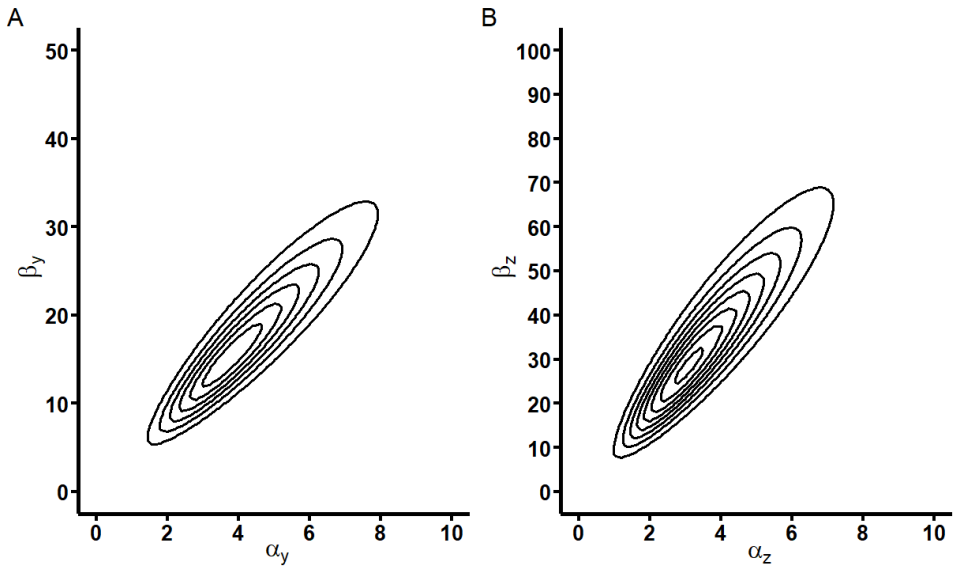


Figure 4. Posterior Plots for Beta Model. (A) is the Posterior Plot for θ_y, (B) is the Posterior Plot for θ_z

Approach 2: Binomial Distribution

The posteriors for the Binomial Approach are written below along with their plots.

$$\theta_y|b^y \sim Beta(5 + \sum_{j=1}^{10} b_j^y, 5 + \sum_{j=1}^{10} n_j - b_j^y)$$

$$\theta_z|b^z \sim Beta(5 + \sum_{k=1}^{8} b_k^z, 5 + \sum_{k=1}^{8} m_k - b_k^z)$$

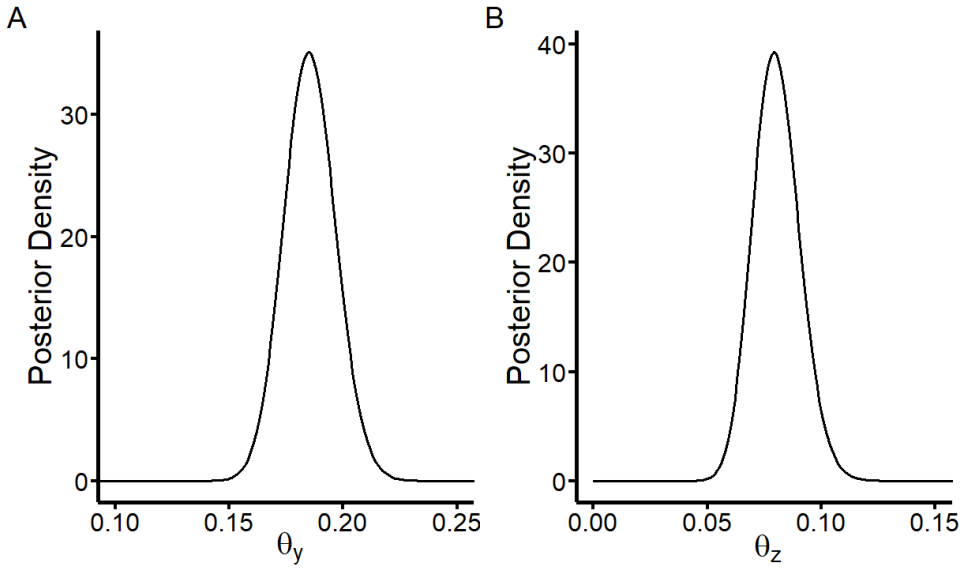


Figure 5. Posterior Plots for Binomial Model. (A) is the Posterior Plot for θ_y, (B) is the Posterior Plot for θ_z

Approach 3: Poisson Distribution

The posterior distributions and posterior plots are listred below for the Gamma Likelihood Approach.

$$\theta_y^b \sim Gamma(15 + \sum_{j=1}^{10} b_j^y, 1 + 10)$$

$$\theta_y^v \sim Gamma(85 + \sum_{j=1}^{10} v_j^y, 1 + 10)$$

$$\theta_z^b \sim Gamma(15 + \sum_{k=1}^{8} b_k^z, 1 + 8)$$

$$\theta_z^v \sim Gamma(85 + \sum_{k=1}^{8} v_k^z, 1 + 8)$$

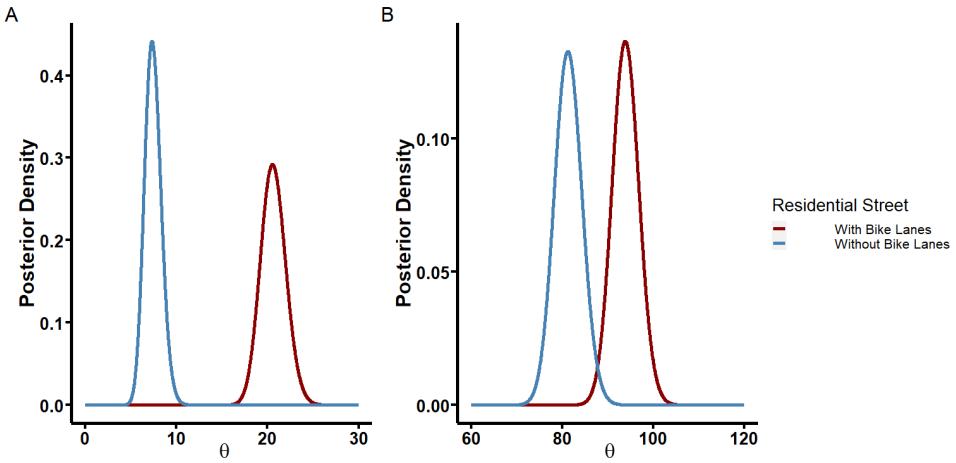


Figure 6. Posterior Plots for the Gamma Model. (A) is the Posterior Plot for θ_y, (B) is the Posterior Plot for θ_z

(d) Let $\mu_y = E(y_i|\theta_y)$ be the mean of the distribution of the $y_i's$; μ_y will be a function of θ_y. Similarly, define μ_z. Using your posterior simulations from (c), plot a histogram of the posterior simulations of $\mu_y - \mu_z$, the expected difference in proportions in bicycle traffic on residential streets with and without bike lanes.

Solution:

Here I exclaim that I disagree with what the problem is asking, or at least how it is asked. What we want is the expected difference in proportions, which in my opinion, is asking for the posterior predictive draws. Therefore the histograms seen below are generated by sampling posterior predictive draws. No matter which method is chosen, one should find that the mean difference in proportions in bicycle traffic on residential streets with and without bike lanes $\approx 10\%$.

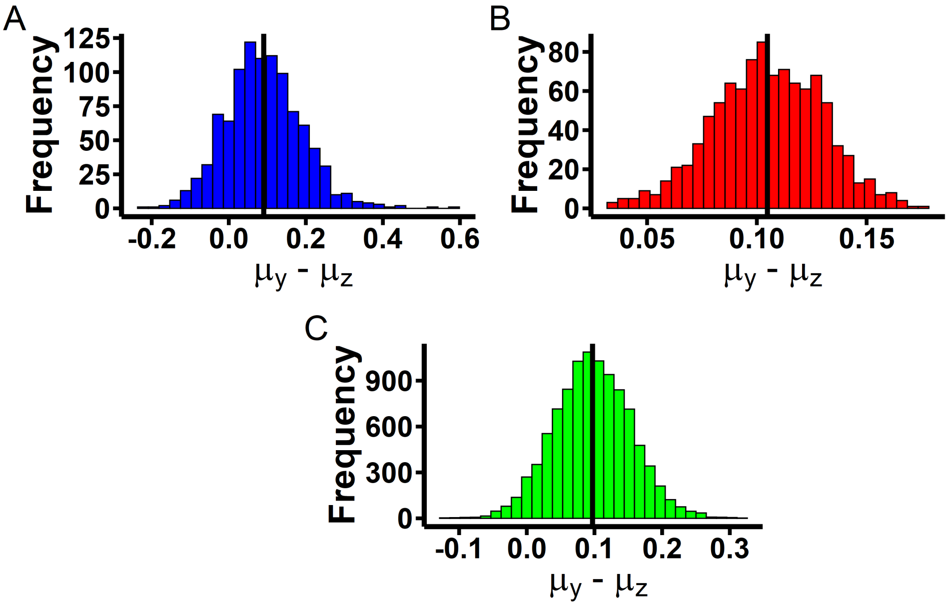


Figure 7. Posterior Predictive Draws for the Expected Differences in proportions. (A) is the Beta Likelihood Model, (B) is the Binomial Likelihood Model, and (C) is the Poisson Likelihood Model

Problem 5: BDA 3rd Ed. 3.12

Poisson regression model: expand the model of Exercise 2.13(a) by assuming that the number of fatal accidents in year t follows a Poisson distribution with mean $\alpha + \beta t$. You will estimate α and β, following the example of the analysis in Section 3.7. The table is provided below for reference

Table 3. Worldwide airline fatalities, 1976–1985. Death rate is passenger deaths per 100 million passenger miles. Source: Statistical Abstract of the United States.

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

(a) Discuss various choices for a ‘noninformative’ prior for (α, β). Choose one.

Solution:

Given that we assume $y_j|\alpha, \beta \sim Poisson(\alpha + \beta t)$, we want to choose a prior such that our rate parameter $\alpha + \beta t > 0$. One possibility would be to assign the flat prior $p(\alpha, \beta) \propto \prod_{j=1}^{N} 1_{\{\alpha+\beta t_j>0\}}$. This prior ensures that for an time point t_j, that our draws of $\alpha + \beta t_j$ will be positive.

Another noninformative prior that we could use would be Jeffrey's Prior, $p(\alpha,\beta)\propto\sqrt{detI(\theta)}$. A quick calculation shows this prior takes the form:

$$p(\alpha,\beta)\propto\sqrt{\sum_{j=1}^{N}\frac{1}{\alpha+\beta t_j}\sum_{j=1}^{N}\frac{t_j^2}{\alpha+\beta t_j}-\Big(\sum_{j=1}^{N}\frac{t_j}{\alpha+\beta t_j}\Big)^2}$$

This is quite controversial to use for multiple parameters, however.

Lastly, one more informative prior could be a "truncated" normal prior where $(\alpha,\beta)\sim N(0,\sigma^2 I)\prod_{j=1}^{N}1_{\{\alpha+\beta t_j>0\}}$, where σ^2 is an incredibly large variance. We put the indicators on this multivariate normal to truncate it to values that produce positive rates.

I will choose the first prior presented, $\prod_{j=1}^{N}1_{\{\alpha+\beta t_j>0\}}$. Regardless of the noninformative prior that is chosen, we will reparameterize $t_j\rightarrow t_j-1975$ so our time is now years since 1975 rather than years.

(b) Discuss what would be a realistic informative prior distribution for (α,β). Sketch its contours and then put it aside. Do parts (c)–(h) of this problem using your noninformative prior distribution from (a).

Solution:

A realistic Informative Prior could assume independence between the two parameters and allow

$$\alpha\sim Gamma(50,1)$$
$$\beta\sim N(\mu=0,\sigma^2=0.5)$$

This prior restricts the range of β by centering it at zero and giving it an extremely small variance, and provides a large range for α by keeping it near 50. This prior also has the added benefit of keeping α always positive and sufficiently large. Personally, I don't have a good working knowledge of fatal aircraft accidents so I wouldn't exactly call this informative. The contours for this prior are below.

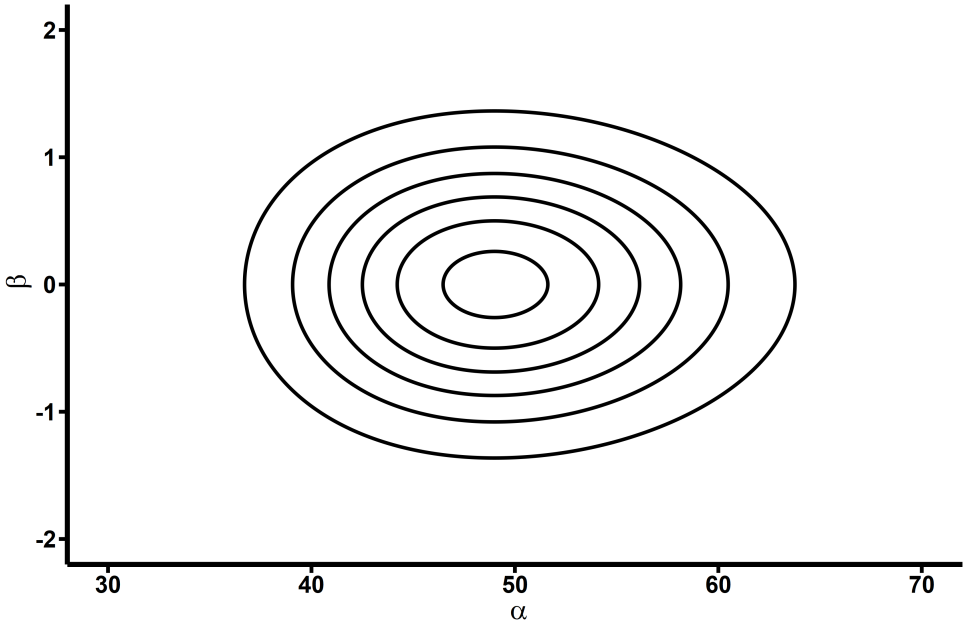


Figure 8. Contours of Our Informative Multivariate Normal Prior

(c) Write the posterior density for (α, β). What are the sufficient statistics?

Solution:

The posterior density for (α, β) is proportional to

$$p(\alpha, \beta|y, t) \propto \prod_{j=1}^{N} 1_{\{\alpha+\beta t_j > 0\}} (\alpha + \beta t_j)^{y_j} e^{-\alpha-\beta t_j}$$

To find the sufficient statistics, recall that the Fisher-Neyman Factorization Theorem states that for a probability density function $f_\theta(x)$, where $X \sim P_\theta$, T(X) is a sufficient statistic for θ iff $f_\theta(x) = h(x)g_\theta(T(x))$, where $g_\theta(T(x))$ is the pdf for $T(x) \mid \theta$. As one may notice, the above pdf is incredibly difficult to factor. We do however have one choice in sufficient statistics which leads to no data reduction, the order statistics. Define $\{(y_{(j)}, t_{(j)})\}_{j=1}^{N}$ to be the order statistics, where $(y_{(j)}, t_{(j)})$ is the number of fatal accidents and year corresponding to the j-th largest year. One can then notice that

$$\begin{aligned} p(\alpha, \beta|y, t) &\propto \prod_{j=1}^{N} 1_{\{\alpha+\beta t_j > 0\}} (\alpha + \beta t_j)^{y_j} e^{-\alpha-\beta t_j} \\ &= \prod_{j=1}^{N} 1_{\{\alpha+\beta t_{(j)} > 0\}} (\alpha + \beta t_{(j)})^{y_{(j)}} e^{-\alpha-\beta t_{(j)}} \end{aligned}$$

so the order statistics are a sufficient statistic for the parameters under this prior distribution.

(d) Check that the posterior density is proper.

Solution:

Choosing our prior as $p(\alpha, \beta) \propto \prod_{j=1}^{N} 1_{\{\alpha+\beta t_j > 0\}}$, we find the posterior density to be proportional to

$$p(\alpha, \beta \mid y, t) \propto \prod_{j=1}^{N} 1_{\{\alpha+\beta t_j > 0\}} (\alpha + \beta t_j)^{y_j} e^{-\alpha-\beta t_j}$$

Now we wish to show that this posterior is proper which is to say it's integral over the domain is finite.

$$\begin{aligned} \iint_{\mathbb{R}^2} p(\alpha, \beta|y, t)\, d\alpha\, d\beta &\propto \iint_{\mathbb{R}^2} \prod_{j=1}^{N} 1_{\{\alpha+\beta t_j > 0\}} (\alpha + \beta t_j)^{y_j} e^{-\alpha-\beta t_j}\, d\alpha\, d\beta \\ &= \iint_{\Omega} \prod_{j=1}^{N} (\alpha + \beta t_j)^{y_j} e^{-\alpha-\beta t_j}\, d\alpha\, d\beta \end{aligned}$$

Where $\Omega = \cap_{j=1}^{N} \{(\alpha, \beta) \in \mathbb{R}^2 \mid \alpha + \beta t_j > 0\}$. This integral looks quite nasty, but thankfully there is an extremely efficient way to deal with this posterior. Firstly, notice that Ω boils down to the intersection of the line segment corresponding to the smallest time point and the largest time point. We can see this graphically below in Figure 9.

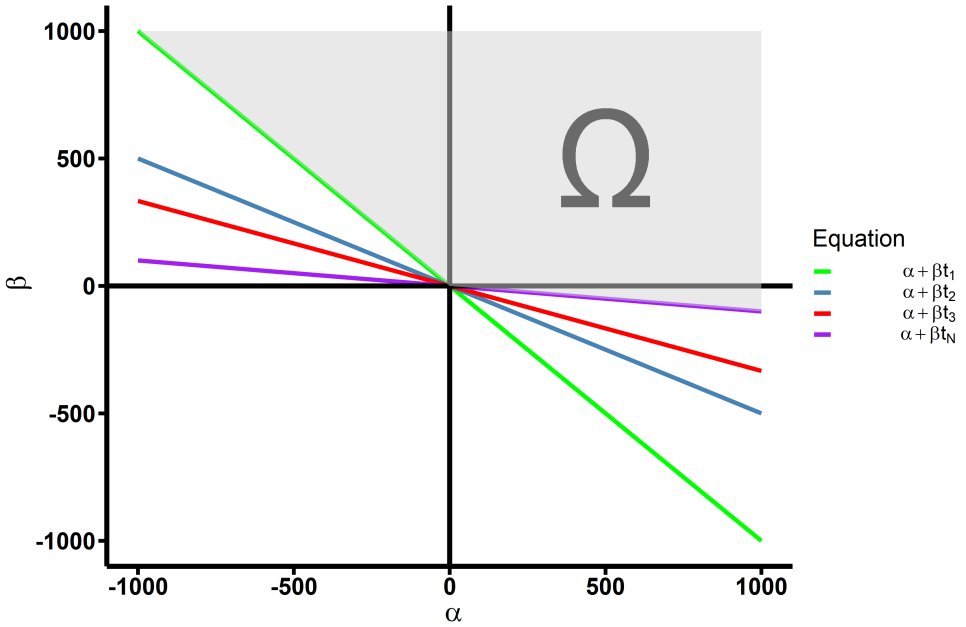


Figure 9. Region of Integration for Problem 3.12d. The gray region denotes the set $\Omega = \{(\alpha, \beta) \in \mathbb{R}^2 \mid \alpha + \beta t_1 > 0, \alpha + \beta t_N > 0\}$

Now that we have our domain of integration, the next thing to notice is that each component of the posterior, $(\alpha + \beta t_j)^{y_j} e^{-\alpha - \beta t_j}$ is bounded. Define $C_j = \sup_{(\alpha,\beta)\in\Omega}(\alpha + \beta t_j)^{y_j} e^{-\alpha-\beta t_j}$ and define $C = max_{1 \leqslant j \leqslant N}\{C_1, \ldots, C_N\}$. We can then find an upper bound for the posterior via:

$$\iint_{\Omega} \prod_{j=1}^{N} (\alpha + \beta t_j)^{y_j} e^{-\alpha-\beta t_j} \, d\alpha \, d\beta \leqslant C^8 \iint_{\Omega} (\alpha + \beta t_1)^{y_1} e^{-\alpha-\beta t_1} (\alpha + \beta t_N)^{y_N} e^{-\alpha-\beta t_N} \, d\alpha \, d\beta$$

We use the smallest time point t_1 and the largest time point t_N since their lines define the domain. Using a change of variables by setting $u = \alpha + \beta t_1$ and $v = \alpha + \beta t_N$, we find that which yields a Jacobian of the form:

$$J(u, v) = \left| \begin{pmatrix} \frac{-t_N}{t_1 - t_N} & \frac{t_1}{t_1 - t_N} \\ \frac{1}{t_1 - t_N} & -\frac{1}{t_1 - t_N} \end{pmatrix} \right| = \left| \frac{1}{t_1 - t_N} \right|$$

which yields the following integral

$$\begin{aligned} C^8 \iint_{\Omega} (\alpha + \beta t_1)^{y_1} e^{-\alpha-\beta t_1} (\alpha + \beta t_N)^{y_N} e^{-\alpha-\beta t_N} \, d\alpha \, d\beta &= \frac{C^8}{|t_1 - t_N|} \int_0^{\infty} \int_0^{\infty} u^{y_1} e^{-u} v^{y_N} e^{-v} du dv \\ &= \frac{C^8}{|t_1 - t_N|} \int_0^{\infty} u^{y_1} e^{-u} du \int_0^{\infty} v^{y_N} e^{-v} dv \end{aligned}$$

Both of these integrals are finite as they are just the kernels of a $Gamma(y_1 + 1, 1)$ distribution and $Gamma(y_N + 1, 1)$ distribution.

(e) Calculate crude estimates and uncertainties for (α, β) using linear regression.

Solution:

This is relatively straightforward. The estimates and corresponding standard errors can be calculated in R and Python.

Table 4. Crude Estimates for α and β

	Estimate	Standard Error	t value	p value
$\hat{\alpha}$	28.8	2.75	10.5	5.89×10^{-6}
$\hat{\beta}$	-0.92	0.44	-2.08	0.07

The covariance matrix is

$$Cov(\hat{\alpha}, \hat{\beta}) = \begin{pmatrix} 7.56 & -1.08 \\ -1.08 & 0.196 \end{pmatrix}$$

(f) Plot the contours and take 1000 draws from the joint posterior density of (α, β).

Solution:

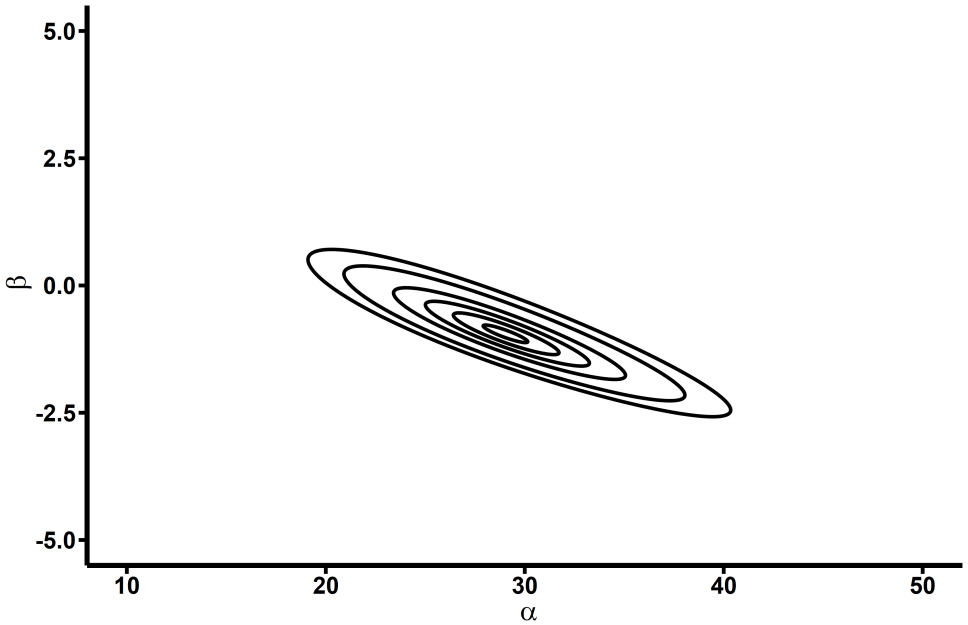


Figure 10. Contour of the Posterior distribution described in part (c).

An interesting thing to note here is that this problem performs a Poisson Regression using the identity link rather than the canonical link, the natural logarithm.

Under the uniform prior chosen here, the MLE of a Poisson regression is simply the MAP (Maximum a Posteriori) of this Posterior distribution.

$(\alpha_{MAP}, \beta_{MAP}) = \arg\max_{(\alpha,\beta)\in\Omega} p(\alpha, \beta|y, t) = \arg\max_{(\alpha,\beta)\in\Omega} p(y \mid \alpha, \beta, t)p(\alpha, \beta) = \arg\max_{(\alpha,\beta)\in\Omega} p(y \mid \alpha, \beta, t) = (\alpha_{MLE}, \beta_{MLE})$

(g) Using your samples of (α, β), plot a histogram of the posterior density for the expected number of fatal accidents in 1986, $\alpha + 1986\beta$.

Solution:

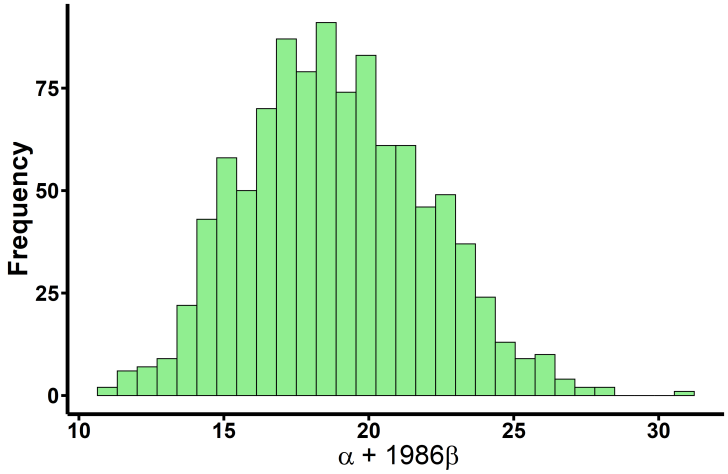


Figure 11. Histogram of the Expected Value of the Poisson Distribution for the year 1986.

It should be noted that the histogram here is calculated using our re-parameterization, $\alpha + \beta * (t - 1975)$.

(h) Create simulation draws and obtain a 95% predictive interval for the number of fatal accidents in 1986.

Solution:

To create simulation draws and obtain a 95% predictive interval, we simply plug our posterior draws from part (f) into the Poisson distribution $Poisson(\alpha + 1986\beta)$. This results in a 95% Predictive Interval of $[10, 31]$.

(i) How does your hypothetical informative prior distribution in (b) differ from the posterior distribution in (f) and (g), obtained from the noninformative prior distribution and the data? If they disagree, discuss.

Solution:

Below is a plot of the informative prior and the posterior obtained from the noninformative prior.

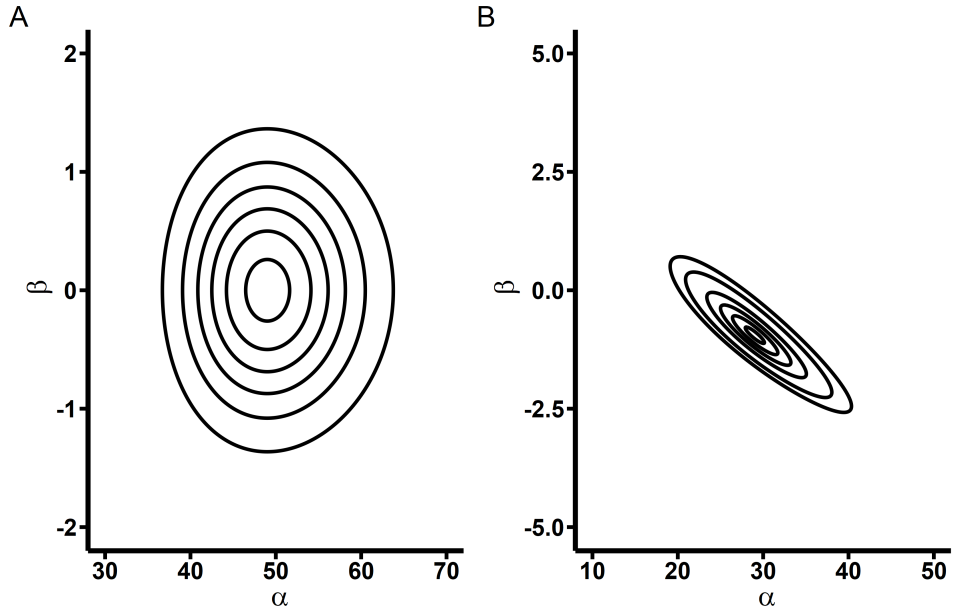


Figure 12. (A) Informative Prior. (B) Posterior Distribution obtained from the Non-Informative Prior

They do not exactly agree as the β is symmetric for the informative prior, while it is heavily skewed towards negative values in the posterior. The posterior is essentially dominated by the data which leads to the difference when comparing the two contours.

Problem 6: BDA 3rd Ed. 3.14

Improper prior and proper posterior distributions: prove that the posterior density (3.15) for the bioassay example has a finite integral over the range $(\alpha, \beta) \in (-\infty, \infty) \times (-\infty, \infty)$. Recall the posterior is of the form:

$$p(\alpha, \beta | y, n, x) \propto p(\alpha, \beta) \prod_{j=1}^{N} [logit^{-1}(\alpha + \beta x_j)]^{y_j} [1 - logit^{-1}(\alpha + \beta x_j)]^{n_j - y_j}$$

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

Solution:

For the Bioassay example detailed in chapter 3, recall we're given the following data:

Dose, x_j log g/ml	Number of animals, n_j	Number of deaths, y_j
-0.86	5	0
-0.30	5	1
-0.05	5	3
0.73	5	5

While it is not necessary to use this data in proving the properness of the posterior, it will be an interesting aside at the end once we prove sufficient conditions for the posterior to be proper.

Assume for the problem that we have $0 < y_j < n_j$ for at least two indices $j, k \in \{1, 2, 3, \ldots, N\}$ and $x_j \neq x_k$. The most important thing to recognize is that $logit^{-1}(\alpha + \beta x_j) = \frac{e^{\alpha + \beta x_j}}{1 + e^{\alpha + \beta x_j}}$ is between 0 and 1 $\forall (\alpha, \beta) \in \mathbb{R}^2$. For shorthand, so I don't have to constantly write that monstrosity, call $\gamma_j(\alpha, \beta) = logit^{-1}(\alpha + \beta x_j)$. We wish to demonstrate that

$$\iint_{\mathbb{R}^2} p(\alpha,\beta|y,n,x)\,d\alpha\,d\beta \propto \iint_{\mathbb{R}^2} \prod_{j=1}^{N} [\gamma_j(\alpha,\beta)]^{y_j}[1-\gamma_j(\alpha,\beta)]^{n_j-y_j}\,d\alpha\,d\beta$$
$$< \infty$$

First, note that $\forall j \in \{1,2,\ldots,N\}$, $0 < \gamma_j(1-\gamma_j) < 1$, which further implies that

$$\prod_{j=1}^{M} [\gamma_j(\alpha,\beta)]^{y_j}[1-\gamma_j(\alpha,\beta)]^{n_j-y_j}\,d\alpha\,d\beta < \prod_{k\in S} [\gamma_k(\alpha,\beta)]^{y_k}[1-\gamma_k(\alpha,\beta)]^{n_k-y_k}$$

where S is a subset of $\{1,2,\ldots,N\}$. The simplest example would be for S to simply be a single element. However if S has a cardinality of 1, the unormalized posterior on the right is improper.

$$\iint_{\mathbb{R}^2} [logit^{-1}(\alpha+\beta x_j)]^{y_j}[1-logit^{-1}(\alpha+\beta x_j)]^{n_j-y_j}\,d\alpha\,d\beta \propto \iint_{\mathbb{R}^2} [\gamma_j(\alpha,\beta)]^{y_j}[1-\gamma_j(\alpha,\beta)]^{n_j-y_j}\,d\alpha\,d\beta$$

Using a u-substitution of the form $u = \gamma_j(\alpha,\beta) = logit^{-1}(\alpha+\beta x_j)$ and $v = \beta$, one can turn the integral into a simpler form (I didn't write out the calculations here, but they follow from simply copying the methodology for the two case below).

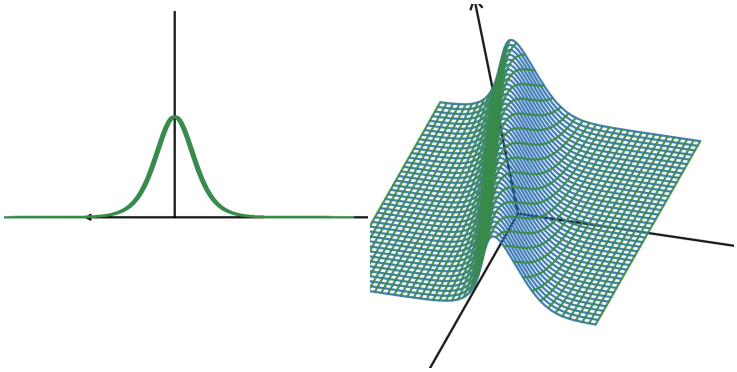
$$\iint_{\mathbb{R}^2} [logit^{-1}(\alpha+\beta x_j)]^{y_j}[1-logit^{-1}(\alpha+\beta x_j)]^{n_j-y_j}\,d\alpha\,d\beta \propto \int_{\mathbb{R}}\int_0^1 u^{y_j}(1-u)^{n_j-y_j}\,du\,dv$$

One quickly recognizes that the integral wrt u is simply a Beta Kernel, and is finite so long as $0 < y_j < n_j$. Assuming this is true, the integral just reduces down to:

$$\int_{\mathbb{R}}\int_0^1 u^{y_j}(1-u)^{n_j-y_j}\,du\,dv = \int_{\mathbb{R}} B(y_j, n_j-y_j)\,dv$$

And now one can visualize where the problem arises. We're taking the integral of a constant over the entire real line which is infinite, so using only one observation as an upper bound puts us back at square one.

Now that we have shown that one observation is improbable which makes sense heuristically. Think about it in terms of the 1D case. Given $f(x) = \frac{(e^x)^{y_j}}{(1+e^x)^{n_j}}$, where $0 < y_j < n_j$, the integral over the real line is finite as you are simply integrating over a small hump and the mass everywhere else is essentially zero. Now extending this to two dimensions, we have; $g(\alpha,\beta) = \frac{(e^{\alpha+\beta x_j})^{y_j}}{(1+e^{\alpha+\beta x_j})^{n_j}}$, where $0 < y_j < n_j$, and where x_j is a fixed value. We're now in three dimensions. There are infinitely many combinations of (α,β) that will make $\alpha+\beta x_j = c$, where $c \in \mathbb{R}$ so $f(c) = f(\alpha+\beta x_j)$. We can see this means that each point on the original graph gets extended linearly over a 3D volume. The plots below illustrate the idea I'm trying to convey. That in the 1D case now gets extended in the 2D case. That hump now becomes an infinitely long tunnel. Another, possibly simpler way to think about this is that in the 1D case you integrate the area below the entrance of an infinitely long tunnel. In the 2D case you integrate the volume of the entire infinitely long tunnel.



(a) $f(x) = \frac{(e^x)^{y_j}}{(1+e^x)^{n_j}}$ **(b)** $g(\alpha,\beta) = \frac{(e^{\alpha+\beta x_j})^{y_j}}{(1+e^{\alpha+\beta x_j})^{n_j}}$

Since using one likelihood failed us, we will now use two.Let $S = \{j,k\}$, where $1 \leqslant j,k \leqslant N$.

$$\iint_{\mathbb{R}^2} [\gamma_j(\alpha,\beta)]^{y_j}[1-\gamma_j(\alpha,\beta)]^{n_j-y_j}[\gamma_k(\alpha,\beta)]^{y_k}[1-\gamma_k(\alpha,\beta)]^{n_k-y_k}\,d\alpha\,d\beta$$

We will begin by performing a change of variables. The problem beforehand with a change of variables had to with the fact that we had N lines of the form $\alpha+\beta x_i$ making it impossible to transform the un-normalized posterior into a nice form. Now that we only have two, we can do a change of variables from (α,β) to (u,v). Define

$$u = logit^{-1}(\alpha+\beta x_j)$$
$$v = logit^{-1}(\alpha+\beta x_k)$$

Some simple algebra yields the jacobian for the transformation.

$$J(\alpha,\beta) = |\begin{pmatrix} \frac{-x_k}{x_j-x_k}\frac{1}{u(1-u)} & \frac{x_j}{x_j-x_k}\frac{1}{v(1-v)} \\ \frac{1}{x_j-x_k}\frac{1}{u(1-u)} & \frac{-1}{x_j-x_k}\frac{1}{v(1-v)} \end{pmatrix}|$$

which gives $J \propto \frac{1}{u(1-u)}\frac{1}{v(1-v)}$, Haldane's Prior (Recall problem 7 from the first homework. You had to prove this was the uniform prior for the natural parameter of the Binomial). This simplifies the integral substantially

$$\iint_{\mathbb{R}^2} [\gamma_j(\alpha,\beta)]^{y_j}[1-\gamma_j(\alpha,\beta)]^{n_j-y_j}[\gamma_k(\alpha,\beta)]^{y_k}[1-\gamma_k(\alpha,\beta)]^{n_k-y_k}\,d\alpha\,d\beta =$$

$$= \int_0^1\int_0^1 u^{y_j}(1-u)^{n_j-y_j}v^{y_k}(1-v)^{n_k-y_k}\frac{1}{u(1-u)}\frac{1}{v(1-v)}dudv$$

$$= \int_0^1 v^{y_k-1}(1-v)^{n_k-y_k-1}dv\int_0^1 u^{y_j-1}(1-u)^{n_j-y_j-1}du$$

$$= B(y_j,n_j-y_j)B(y_k,n_k-y_k)$$

This provides us with a finite bound for the posterior distribution so under the conditions that there exists at least two data points such that $x_j \neq x_k$, $0 < y_j < n_j$ and $0 < y_k < n_k$, the posterior is proper. Applying this to the data in the Bioassay example, we can see that the data will give rise to a proper posterior.

Problem 7: BDA 3rd Ed. 3.15

Joint distributions: The autoregressive time-series model $y_1, y_2, \ldots$ with mean level 0, autocorrelation 0.8, residual standard deviation 1, and normal errors can be written as

$$(y_t|y_{t-1}, y_{t-2}, \dots) \sim \mathcal{N}(0.8y_{t-1}, 1) \quad \forall t$$

(a) Prove that the distribution of y_t, given the observations at all other integer time points t, depends only on y_{t-1} and y_{t+1}.

Solution:

For this problem consider the number of observations to be finite. That is to say our data can be written as $y = (y_1, y_2, \dots, y_T)$. Now let t be a fixed integer such that $1 < t < T$. The gist of this problem is that we want to show

$$p(y_t|\{y_j\}_{j\neq t}^T) = p(y_t \mid y_{t+1}, y_{t-1})$$

Here, $\{y_j\}_{j\neq t}^T$ is shorthand notation for all time data points not including y_t. Via Bayes Theorem,

$$\begin{aligned} p(y_t|y_1, y_2, \dots, y_{t-1}, y_{t+1}, \dots, y_T) &= \frac{p(y_1, y_2, \dots, y_{t-1}, y_t, y_{t+1}, \dots, y_T)}{p(y_1, y_2, \dots, y_{t-1}, y_{t+1}, \dots, y_T)} \\ &\propto p(y_1, y_2, \dots, y_{t-1}, y_t, y_{t+1}, \dots, y_T) \end{aligned}$$

The proportionality follows from the fact that the denominator doesn't include y_t so it can be treated as a constant. Now, from probability, we know for any two events A,B that $P(A, B) = P(A \mid B)P(B)$. Applying this formula recursively yields

$$\begin{aligned} p(y_t|y_1, y_2, \dots, y_{t-1}, y_{t+1}, \dots, y_T) &\propto p(y_1, y_2, \dots, y_{t-1}, y_t, y_{t+1}, \dots, y_T) \\ &= p(y_T \mid \{y_j\}_{j=1}^{T-1}) p(y_{T-1} \mid \{y_j\}_{j=1}^{T-2}) p(y_{T-2} \mid \{y_j\}_{j=1}^{T-3}) \cdot \dots \cdot p(y_2 \mid y_1) p(y_1) \end{aligned}$$

This expression can be greatly simplified by noticing that the problem states that the present data conditional on all past data is equivalent to the present data conditional on the data at the previous time point. In other words, $p(y_t \mid \{y_j\}_{j=1}^{t-1}) = p(y_t \mid y_{t-1})$. Using this, our expression simplifies down to

$$\begin{aligned} p(y_t|y_1, y_2, \dots, y_{t-1}, y_{t+1}, \dots, y_T) &\propto \Big(\prod_{j=1}^{T} p(y_{j+1} \mid y_j) \Big) p(y_1) \\ &\propto p(y_t \mid y_{t-1}) p(y_{t+1} \mid y_t) \end{aligned}$$

The second line follows from the fact that only the previous and future time step contain information regarding y_t.

(b) What is the distribution of y_t given y_{t-1} and y_{t+1}?

Solution:

Part a give us that $p(y_t \mid y_{t-1}, y_{t+1}) \propto p(y_t \mid y_{t-1}) p(y_{t+1} \mid y_t)$. For a more formal derivation,

$$\begin{aligned} p(y_t \mid y_{t+1}, y_{t-1}) &\propto p(y_{t+1}, y_{t-1} \mid y_t) p(y_t) \\ &= p(y_{t+1} \mid y_{t-1}, y_t) p(y_{t-1} \mid y_t) p(y_t) \\ &= p(y_{t+1} \mid y_t) p(y_{t-1} \mid y_t) p(y_t) \\ &\propto p(y_{t+1} \mid y_t) p(y_t \mid y_{t-1}) \end{aligned}$$

Now given this expression, we know that both of the distributions on the right are normals. However, we cannot just use the formula from Problem 6 in the previous homework since we have one distribution where y_t is in the conditional. Instead, we do some quick algebraic manipulation. We can express the distribution of $y_t \mid y_{t+1}$ as

$$\begin{aligned} y_{t+1} &= 0.8y_t + \epsilon_t \\ y_t &= \frac{1}{0.8}y_{t+1} + \frac{1}{0.8}\epsilon_t \\ y_t \mid y_{t+1} &\sim N(\frac{10}{8}y_{t+1}, \frac{100}{64}) \end{aligned}$$

Where $\epsilon_t \sim N(0,1)$. This technically isn't the proper way to derive the conditional distribution. The proper/rigorous way to do this would be to actually use the densities and complete the square. Step 2 is only justified by the fact that ϵ_t is a symmetric random variable centered at 0 so $-\epsilon_t \sim N(0,1)$. In fact, $p(y_t \mid y_{t+1}) \propto p(y_{t+1} \mid y_t)$. This gives us

$$p(y_t \mid y_{t+1}, y_{t-1}) \propto p(y_t \mid y_{t+1})p(y_t \mid y_{t-1})$$

Now our conditional is in a form where the two densities are functions of y_t so using the formula we find that

$$\begin{aligned} E(y_t \mid y_{t+1}, y_{t-1}) &= \frac{\frac{4}{5}y_{t-1} + \frac{64}{100}\frac{5}{4}y_{t+1}}{1 + \frac{64}{100}} \\ Var(y_t \mid y_{t+1}, y_{t-1}) &= \frac{1}{1 + \frac{64}{100}} \end{aligned}$$

So finally we get that

$$y_t \mid y_{t+1}, y_{t-1} \sim N(\frac{20}{41}(y_{t-1} + y_{t+1}), \frac{100}{164})$$