PHP2530: Bayesian Statistical Methods

Homework 2

Antonella Basso

March 11, 2022

Problem 1: (Ch.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 Table 3.2. 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 3.2:

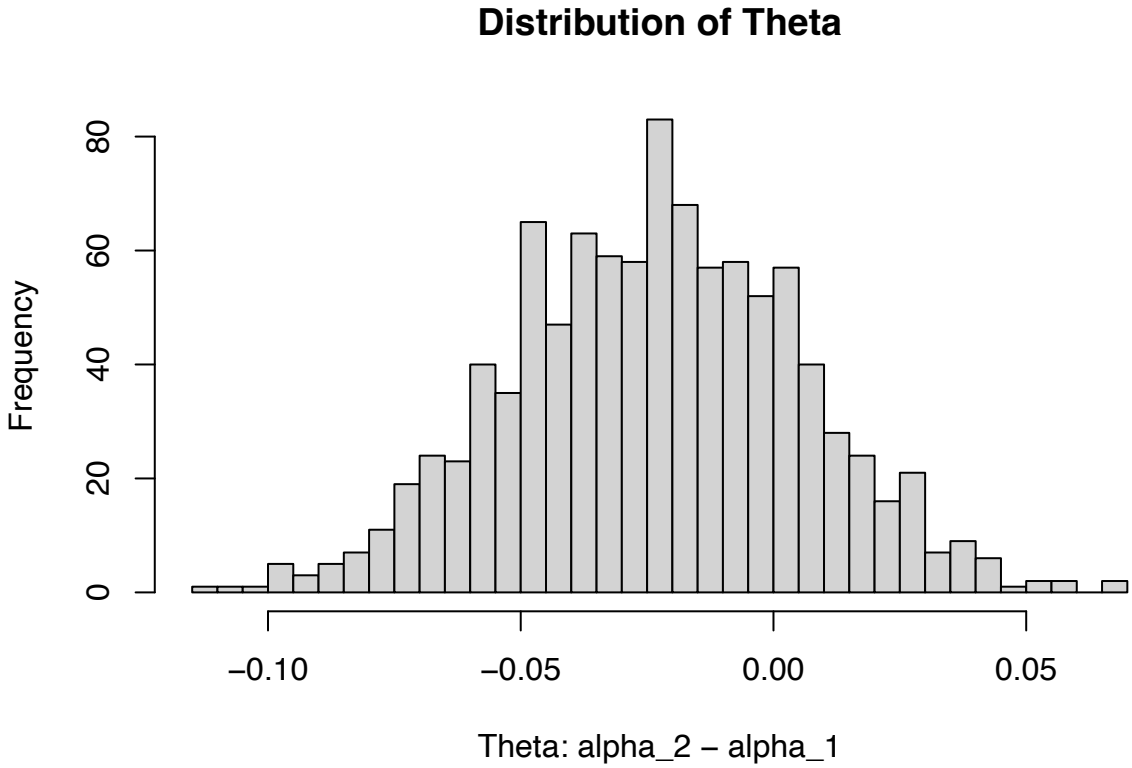
Survey	Bush	Dukakis	No Opinion/Other	Total
pre-debate	294	307	38	639
post-debate	289	332	19	639

Solution

Let $\theta = \alpha_2 - \alpha_1$, such that $\alpha_1 = \frac{p_{11}}{p_{11}+p_{12}}$ and $\alpha_2 = \frac{p_{21}}{p_{21}+p_{22}}$, given that $p_{11}, p_{12}|y \sim \text{Dir}(295, 308)$ and $p_{21}, p_{22}|y \sim \text{Dir}(290, 333)$. Simulating 1,000 values from each distribution for p_{1i} and p_{2i} (for $i = 1, 2$), we obtain the following frequency distribution for θ:

```
################################## QUESTION 1 ########################################
set.seed(47)

# Parameter Vectors
pre <- c(295, 308, 39) #j=1: pre-debate survey
post <- c(290, 333, 20) #j=2: post-debate survey

# Simulating Data
dir_1 <- rdirichlet(1000, pre) #j=1: pre-debate survey
dir_2 <- rdirichlet(1000, post) #j=2: post-debate survey

# Computing Alpha Values: `theta_b/(theta_b+theta_d))`
alpha_1 <- dir_1[, 1]/(dir_1[, 1]+dir_1[, 2])
alpha_2 <- dir_2[, 1]/(dir_2[, 1]+dir_2[, 2])

# Computing Parameter of Interest: `alpha_2-alpha_1`
alpha_diff <- alpha_2 - alpha_1
```

```r
# Histogram of `alpha_2-alpha_1`
hist(alpha_diff, breaks=30,
     main="Distribution of Theta",
     xlab="Theta: alpha_2 - alpha_1")
```




```r
# Posterior Probability: Proportion of `alpha_2-alpha_1` > 0
sum(alpha_diff > 0)/length(alpha_diff)
```

The posterior probability that there was a shift toward Bush is $\approx 21.5\%$.

Problem 2: (Ch.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 $(\mu_c, \mu_t, \log\sigma_c, \log\sigma_t)$.

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 $\mu_t - \mu_c$.

The problem of estimating two normal means with unknown ratio of variances is called the Behrens–Fisher problem.

Solution

Sampling Distribution 1: Control

- $y_{ci} \sim N(\mu_c, \sigma_c^2)$: measurement for the i^{th} control chicken, where $i = 1, 2, ..., 32$
- $\bar{y}_c = 1.013$: sample mean
- $s_c^2 = 0.24^2$: sample variance
- μ_c: mean (parameter of interest)
- σ_c^2: variance

Sampling Distribution 2: Treatment

- $y_{ti} \sim N(\mu_t, \sigma_t^2)$: measurement for the i^{th} treatment chicken, where $i = 1, 2, ..., 36$
- $\bar{y}_t = 1.173$: sample mean
- $s_t^2 = 0.2^2$: sample variance
- μ_t: mean (parameter of interest)
- σ_t^2: variance

Parameters:

- $\mu = [\mu_c, \mu_t]$, known $\bar{y} = [\bar{y}_c, \bar{y}_t] = [1.013, 1.173]$
- $\sigma^2 = [\sigma_c^2, \sigma_t^2]$, known $s^2 = [s_c^2, s_t^2] = [0.24^2, 0.2^2]$

a) Uninformative Uniform Priors:

$$\mu_c, \sigma_c \sim U[\mu_c, \log\sigma_c] \Rightarrow p(\mu_c, \sigma_c) \propto 1 \Rightarrow p(\mu_c, \sigma_c^2) \propto \frac{1}{\sigma_c^2}$$

$$\mu_t, \sigma_t \sim U[\mu_t, \log\sigma_t] \Rightarrow p(\mu_t, \sigma_t) \propto 1 \Rightarrow p(\mu_t, \sigma_t^2) \propto \frac{1}{\sigma_t^2}$$

Sampling Distributions (Likelihoods):

$$y_c|\mu_c, \sigma_c^2 \sim N(\mu_c, \sigma_c^2)]$$

$$y_t|\mu_t, \sigma_t^2 \sim N(\mu_t, \sigma_t^2)$$

Joint Posteriors:

$$p(\mu_c, \sigma_c^2|y_c) \propto \frac{1}{\sigma_c^{n+2}} \cdot e^{-\frac{1}{2\sigma_c^2}\sum_{i=1}^{n}(y_{ci}-\mu_c)^2} = \frac{1}{\sigma_c^{n+2}} \cdot e^{-\frac{1}{2\sigma_c^2}[(n-1)s_c^2+n(\bar{y}_c-\mu_c)^2]}$$

$$p(\mu_t, \sigma_t^2|y_t) \propto \frac{1}{\sigma_t^{n+2}} \cdot e^{-\frac{1}{2\sigma_t^2}\sum_{i=1}^{n}(y_{ti}-\mu_t)^2} = \frac{1}{\sigma_t^{n+2}} \cdot e^{-\frac{1}{2\sigma_t^2}[(n-1)s_t^2+n(\bar{y}_t-\mu_t)^2]}$$

Marginal Posteriors:

$$p(\mu_c|y_c) = \int_0^\infty p(\mu_c, \sigma_c^2|y_c)d\sigma_c^2 \propto \left[1 + \frac{n(\mu_c - \bar{y}_c)^2}{(n-1)s_c^2}\right]^{-n/2} \Rightarrow \mu_c|y_c \sim t_{n-1}(\bar{y}_c, s_c^2/n)$$

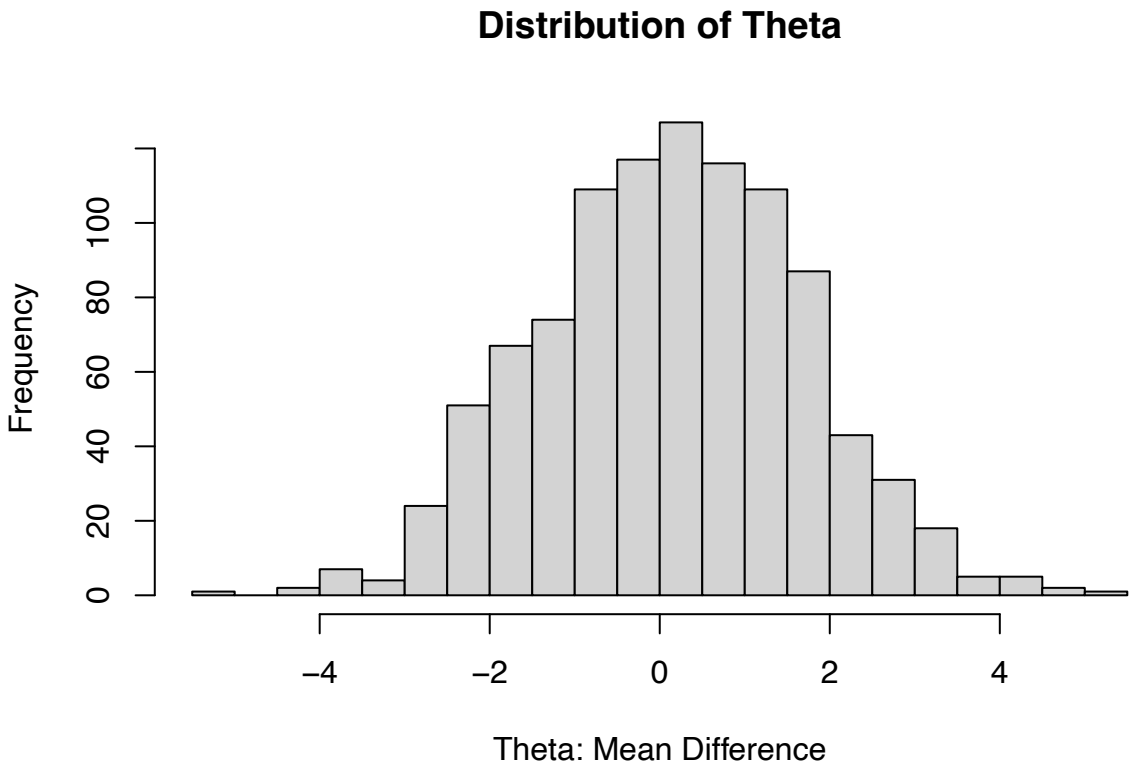
$$p(\mu_t|y_t) = \int_0^\infty p(\mu_t, \sigma_t^2|y_t)d\sigma_t^2 \propto \left[1 + \frac{n(\mu_t - \bar{y}_t)^2}{(n-1)s_t^2}\right]^{-n/2} \Rightarrow \mu_t|y_t \sim t_{n-1}(\bar{y}_t, s_t^2/n)$$

$$\text{Control: } n = 32, \bar{y}_c = 1.013, s_c^2 = 0.24^2 \Rightarrow \mu_c|y_c \sim t_{31}(1.013, 0.24^2/32) \Rightarrow \left.\frac{\mu_c - 1.013}{0.24/\sqrt{32}}\right| y_c \sim t_{31}$$

$$\text{Treatment: } n = 36, \bar{y}_t = 1.173, s_t^2 = 0.2^2 \Rightarrow \mu_t|y_t \sim t_{35}(1.173, 0.2^2/36) \Rightarrow \left.\frac{\mu_t - 1.173}{0.2/\sqrt{36}}\right| y_t \sim t_{35}$$

b) Let $\theta = \mu_t - \mu_c$. Simulating 1,000 values from each marginal distribution for the mean (defined above), we obtain the following frequency distribution for θ:

```
################################# QUESTION 2 ######################################
set.seed(47)

# Control Group Mean
t_c <- 1.013+(0.24/sqrt(32)) #t-statistic (ncp)
mu_c <- rt(1000, 31, t_c) #same as t_c*rt(1000, 31)

# Treatment Group Mean
t_t <- 1.173+(0.2/sqrt(36)) #t-statistic (ncp)
mu_t <- rt(1000, 35, t_t) #same as t_t*rt(1000, 35)

# Theta: Mean Difference (Parameter of Interest)
theta <- mu_t-mu_c

# Distribution of Theta
hist(theta, breaks=30,
     main="Distribution of Theta",
     xlab="Theta: Mean Difference")
```




```
#95% CI
CI_95 <- t.test(theta, conf.level=0.95)
c(CI_95$conf.int[1], CI_95$conf.int[2])
```

```
## [1] 0.05668616 0.24787034
```

A 95% credible (posterior) interval for $\theta = \mu_t - \mu_c$ is given by $[0.057, 0.248]$.

Problem 3: (Ch.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.

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

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

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

Solution

a) Uninformative Uniform Prior:

$$\mu, \sigma \sim \text{U}[\mu, \log\sigma] \Rightarrow p(\mu, \sigma) \propto 1 \Rightarrow p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$$

Sampling Distribution (Likelihood):

$$y|\mu, \sigma^2 \sim \text{N}(\mu, \sigma^2)]$$

Joint Posterior:

$$p(\mu, \sigma^2|y) \propto \frac{1}{\sigma_c^{n+2}} \cdot e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2} = \frac{1}{\sigma^{n+2}} \cdot e^{-\frac{1}{2\sigma^2}[(n-1)s_c^2+n(\bar{y}-\mu)^2]}$$

$$p(\mu, \sigma^2|y) = p(\mu|\sigma^2, y) \cdot p(\sigma^2|y) = \text{N}(\bar{y}, \sigma^2/n) \cdot \text{Inv-}\chi^2(n-1, s^2)$$

Marginal Posteriors:

$$p(\mu|y) = \int_0^\infty p(\mu, \sigma^2|y)d\sigma^2 \propto \left[1 + \frac{n(\mu-\bar{y})^2}{(n-1)s^2}\right]^{-n/2} \Rightarrow \mu|y \sim t_{n-1}(\bar{y}, s^2/n)$$

$$p(\sigma^2|y) = \int_0^\infty p(\mu, \sigma^2|y)d\mu \propto (\sigma^2)^{-(n+1)/2} e^{-\frac{(n-1)s^2}{2\sigma^2}} \Rightarrow \sigma^2|y \sim \text{Inv-}\chi^2(n-1, s^2)$$

```
############################## QUESTION 3: A ###################################

obs <- c(10, 10, 12, 11, 9) # Data
n <- length(obs) # n
y_bar <- mean(obs) # Sample Mean
s2 <- var(obs) # Sample Variance

c(n, y_bar, s2)
```

```
## [1]  5.0 10.4  1.3
```

Given $n = 5, \bar{y} = 10.4, s^2 = 1.3$:

$$\mu|y \sim t_4(10.4, 1.3/5)$$
$$\sigma^2|y \sim \text{Inv-}\chi^2(4, 1.3)$$

b) Assuming the same uninformative prior and sampling distributions (from part (a) above), the posterior distribution for (μ, σ^2) is given by:

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

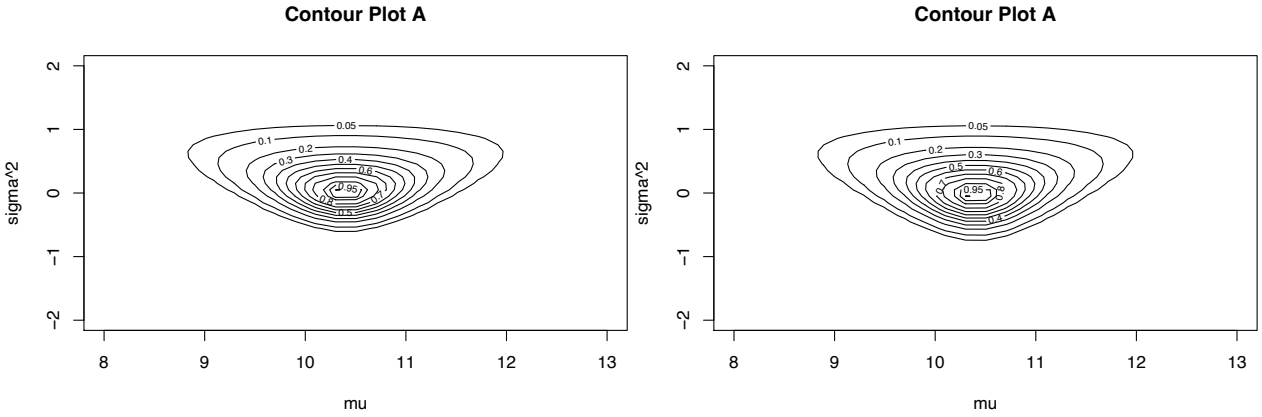
However, since y_i are rounded measurements, the true (unrounded) value of y_i lies between $y_{i-1} + 0.5$ and $y_i + 0.5$. That is, $y_i \pm 0.5$. Thus, the correct posterior distribution for (μ, σ^2) treating the measurements as rounded is given by:

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

c) Comparing the contour plots for both calculated posteriors, we notice that they are almost identical. Similarly, both posterior means come very close (especially, with regards to their means). However, we see that the variance for the posterior for which y_i are treated as unrounded (part (a)) is much larger than that for which we account for the fact that y_i are rounded measurements (part (b)), both in terms of their means and standard deviations. This is likely due to the fact that there is generally more uncertainty regarding the population in part (a).

```
################################ QUESTION 3: C ####################################
# Comparing Contour Plots

set.seed(47)
library(LearnBayes)
library(LaplacesDemon)

# Joint Posterior A (treating measurements as unrounded)
jpost_a <- function(mu, sig2, obs){
  total <-  0
  for (i in 1:length(obs)){
    total <- total + log(dnorm(obs[i], mu, sig2))
  }
  return(total)
}

# Joint Posterior B (treating measurements as rounded)
jpost_b <- function(mu, sig2, obs){
  total <- 0
  for (i in 1:length(obs)){
    total <- total +
      log(pnorm(obs[i] + 0.5, mu, sig2) - pnorm(obs[i] - 0.5, mu, sig2))
  }
  return(total)
}

mu_list <- seq(0, 20, length=100)
sig2_grid <- seq(-5, 5, length=100)

# Contour Plot A
log_dens_a <- outer(X=mu_list, Y=exp(sig2_grid), FUN=jpost_a, obs)
dens_a <- exp(log_dens_a-max(log_dens_a))
contour(mu_list, sig2_grid, dens_a, levels=c(0.05, 0.95, seq(0, 1, 0.1)),
        main="Contour Plot A", xlab="mu", ylab="sigma^2",
        xlim=c(8, 13), ylim=c(-2, 2))
```

```r
# Contour Plot B
log_dens_b <- outer(X=mu_list, Y=exp(sig2_grid), FUN=jpost_b, obs)
dens_b <- exp(log_dens_b-max(log_dens_b))
contour(mu_list, sig2_grid, dens_b, levels=c(0.05, 0.95, seq(0, 1, 0.1)),
        main="Contour Plot A", xlab="mu", ylab="sigma^2",
        xlim=c(8,13), ylim=c(-2,2))
```




```r
############################## QUESTION 3: C ###################################
# Comparing Means and Variances

set.seed(47)

# Marginal Posterior Variance A
sig2_a <- (n-1) * s2/rinvchisq(1000, n-1, scale=s2)
c("posterior variance A:", mean(sig2_a), sd(sig2_a))

# Conditional Posterior Mean A
mu_a <- rnorm(1000, y_bar, sqrt(sig2_a)/sqrt(n))
c("posterior mean A:", mean(mu_a), sd(mu_a))

# Sampling Posterior Mean B
mu_dist <- apply(dens_b, 1, sum)
mu_dist_sample <- sample(x=1:100, size=1000, replace=T, prob=mu_dist)
mu_dist_sample_list <- mu_list[mu_dist_sample]
c("posterior mean B:", mean(mu_dist_sample_list), sd(mu_dist_sample_list))

# Sampling Posterior Variance B
sig2_sample <- numeric(length=1000)
for (i in 1:1000){
  sig2_sample[i] <- exp(sample(sig2_grid, 1, prob=dens_b[mu_dist_sample[i],]))
}
c("posterior variance B:", mean(sig2_sample), sd(sig2_sample))
```

```
## [1] "posterior mean A:" "10.4361121571659"  "0.89319022924946"
## [1] "posterior mean B:" "10.4018181818182"  "0.672911491565593"
## [1] "posterior variance A:" "3.88230272116312"      "2.82831087866404"
## [1] "posterior variance B:" "1.36898234951694"      "0.74795079312028"
```

d) Having drawn simulations from the posterior distribution of z_i, where $i = 1, 2, ..., 5$, the posterior mean of $(z_1 - z_2)^2$ is ≈ 0.156.

```
############################## QUESTION 3: D ####################################

set.seed(47)

# Simulation for Posterior Distribution of z
z_i <- matrix(0, 1000, 5)
for (i in 1:5){
  z_1 <- pnorm(obs[i]-.5, mu_dist_sample_list, sig2_sample)
  z_2 <- pnorm(obs[i]+.5, mu_dist_sample_list, sig2_sample)
  z_i[,i] <- qnorm(z_1 + runif(1000)*(z_2-z_1), mu_dist_sample_list, sig2_sample)}
mean((z_i[,1]-z_i[,2])^2)
```

```
## [1] 0.1555351
```

Problem 4: (Ch.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. Table 3.3 displays the number of bicycles and other vehicles recorded in the study. For this problem, restrict your attention to the data on residential streets.

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.

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

c) Determine the posterior distribution for the parameters in your model and draw 1,000 simulations from the posterior distribution. (Hint:θ_y and θ_z are independent in the posterior distribution, so they can be simulated independently.)

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.

Table 3.3:

Type of Street	Bike Route?	Counts of bicycles/other vehicles
Residential	yes	16/58, 9/90, 10/48, 13/57, 19/103, 20/57, 18/86, 17/112, 35/273, 55/64
Residential	no	12/113, 1/18, 2/14, 4/44, 9/208, 7/67, 9/29, 8/154
Fairly busy	yes	8/29, 35/415, 31/425, 19/42, 38/180, 47/675, 44/620, 44/437, 29/47, 18/462
Fairly busy	no	10/557, 43/1258, 5/499, 14/601, 58/1163, 15/700, 0/90, 47/1093, 51/1459, 32/1086
Busy	yes	60/1545, 51/1499, 58/1598, 59/503, 53/407, 68/1494, 68/1558, 60/1706, 71/476, 63/752
Busy	no	8/1248, 9/1246, 6/1596, 9/1765, 19/1290, 61/2498, 31/2346, 75/3101, 14/1918, 25/2318

Solution

```
################################## QUESTION 4 ######################################

# Computing y_i and z_i

# raw ordered counts of observed bicycle and other vehicle traffic
bikes <- c(16, 9, 10, 13, 19, 20, 18, 17, 35, 55, 12, 1, 2, 4, 9, 7, 9, 8)
other <- c(58, 90, 48, 57, 103, 57, 86, 112, 273, 64, 113, 18, 14, 44, 208, 67, 29, 154)

# whether each count comes from a bike route or not (ordered)
bike_route <- c(rep(1, 10), rep(0, 8)) #1=yes, 0=no

# data frame to compute y and z vectors from columns
traffic <- data.frame(bike_route=bike_route,
                      bikes=bikes,
                      other=other)

# y and z vectors (as counts)
y_c <- traffic$bikes[traffic$bike_route==1] #bikes in bike route
z_c <- traffic$bikes[traffic$bike_route==0] #bikes in non-bike route

y_other <- traffic$other[traffic$bike_route==1] #other vehicles in bike route
z_other <- traffic$other[traffic$bike_route==0] #other vehicles in non-bike route

# y and z vectors (as proportions)
y_p <- y_c/(y_c+y_other)
z_p <- z_c/(z_c+z_other)
```

a) Models for y and z:

Sampling Distributions:

$$y_i \overset{iid}{\sim} \text{Bin}(n_i, \theta_y) \text{ for } i = 1, 2, ..., 10$$

$$z_i \overset{iid}{\sim} \text{Bin}(n_i, \theta_z) \text{ for } i = 1, 2, ..., 8$$

Note*: This model assumes that y_i and z_i are the number of successes (bikes in bike lanes and non-bike lanes, respectively) at points i. To model them as proportions, such that each can take a value between 0 and 1 (with θ instead being the corresponding number of successes), we may consider doing a transformation of variables and use a Jeffrey's prior, or modeling each with a Beta distribution (or Direchlet if we model them together) and reparametarize using $\frac{\alpha}{\alpha+\beta}$ for the mean (for which we could use a noninformative uniform prior) and $\alpha + \beta$ as a rough measure of the precision (for which we could use a Gamma prior).

b) Uninformative Priors:

$$\theta_y \sim \text{Beta}(1, 1)$$

$$\theta_z \sim \text{Beta}(1, 1)$$

Alternatively,

$$\theta_y, \theta_z \sim \text{Dir}(1)$$

c) Posteriors:

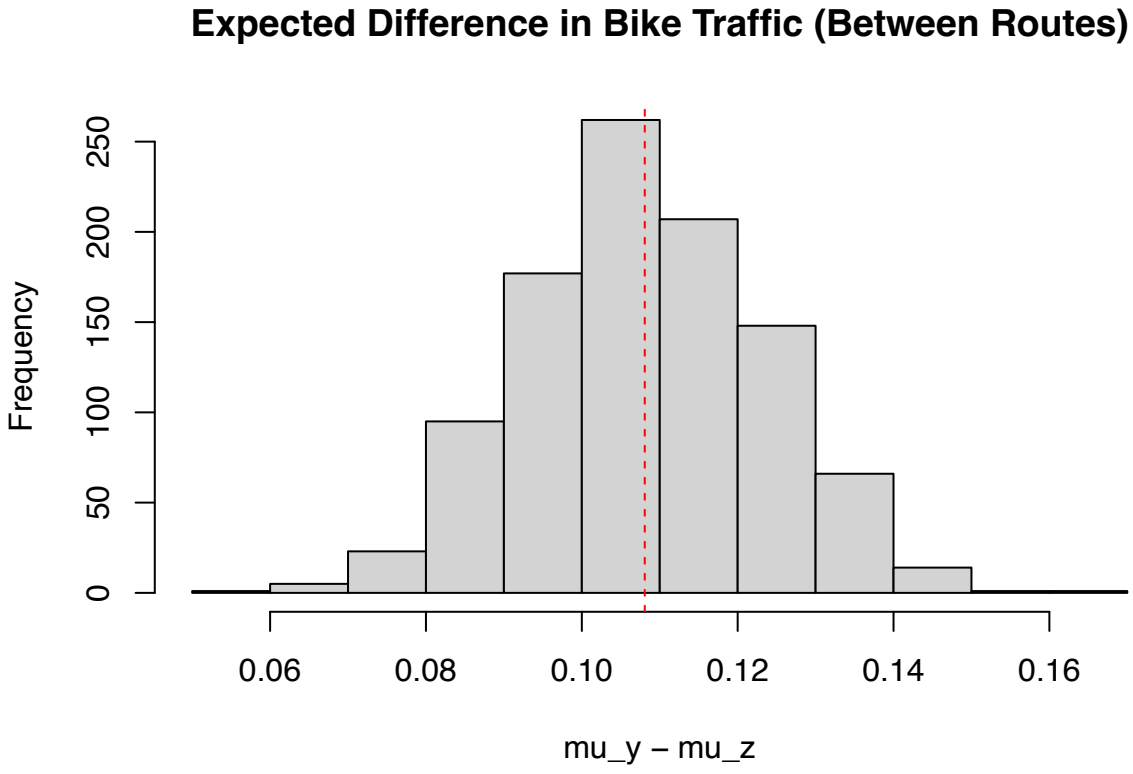
$$\theta_y|y_i \sim \text{Beta}\left(\theta_y \middle| 1+\sum_{i=1}^{n} y_i, 1+n-\sum_{i=1}^{n} y_i\right)$$

$$\theta_z|z_i \sim \text{Beta}\left(\theta_z \middle| 1+\sum_{i=1}^{n} z_i, 1+n-\sum_{i=1}^{n} z_i\right)$$

```
################################ QUESTION 4: C ####################################

set.seed(47)

# Simulating Posterior
n_sample <- 1000
y_post <- rbeta(n_sample, 1+sum(y_c), 1+sum(y_c+y_other)-sum(y_c))
z_post <- rbeta(n_sample, 1+sum(z_c), 1+sum(z_c+z_other)-sum(z_c))
```

d) Using the posterior simulations from part (c), a histogram of the expected difference in bike traffic between residential streets with and without bike lanes, $\mu_y - \mu_z$, is given below.

```
################################ QUESTION 4: D ####################################

set.seed(47)

# Expected Difference in Bike Traffic (Between Routes)
exp_diff <- y_post - z_post
exp_diff_mean <- mean(exp_diff)
hist(exp_diff,
     main="Expected Difference in Bike Traffic (Between Routes)",
     xlab="mu_y - mu_z")
abline(v=exp_diff_mean, lty=2, col="red")
```



Problem 5: (Ch.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.

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

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).

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

d) Check that the posterior density is proper.

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

f) Plot the contours and take 1,000 draws from the joint posterior density of (α, β).

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

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

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

Model for t:

$$y_i \sim \text{Poisson}(\theta_i), \text{ where } \theta_i = \alpha + \beta t_i$$

a) Possible Noninformative Priors:

1. Noninformative Unifrom: $p(\alpha, \beta) \propto 1$
2. Jeffrey's Prior

Since there is more than one parameter, a diffuse uniform prior might be a better alternative to Jeffrey's prior.

b) Given that $\theta_i = \alpha + \beta t_i$, we could use linear regression to estimate values for α and β from the data. Moreover, since linear regression assumes that intercept and slope coefficients are normally distributed, it would make sense to construct a normal informative on these grounds. Specifically, we could model our prior beliefs about α and β through independent normal distributions, using their computed estimates and standard errors as their means and standard deviations, respectively.

```
############################### QUESTION 5: B ####################################

# Regression

# Fatal Accidents Data (t_i, y_i)
fa <- data.frame(t_i=1:10, y_i=c(24, 25, 31, 31, 22, 21, 26, 20, 16, 22))

# Linear Regression Model
lin_reg <- lm(y_i ~ t_i, fa)

# Estimates to Parameterize Prior
summary(lin_reg)$coefficients
```

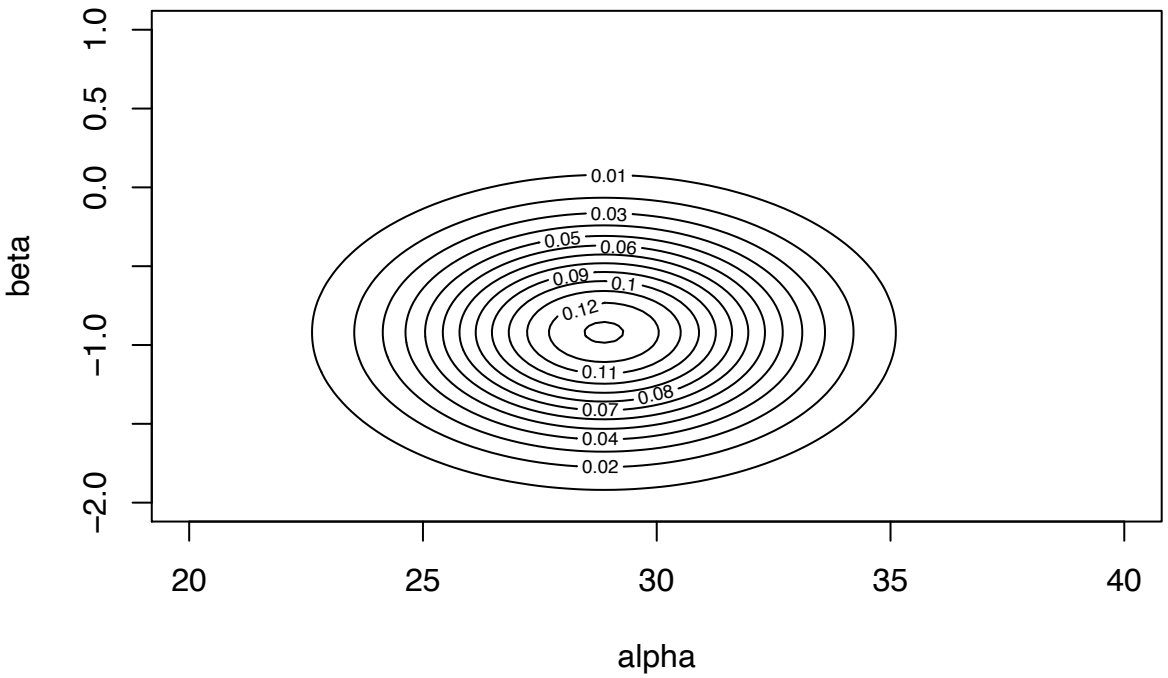
```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 28.8666667  2.7494168 10.499196 5.893425e-06
## t_i         -0.9212121  0.4431086 -2.078976 7.122850e-02
```

Informative Priors:

$$\alpha \sim \mathrm{N}(28.87, 2.75)$$

$$\beta \sim \mathrm{N}(-0.92, 0.44)$$

```
############################## QUESTION 5: B ###################################

# Contour Plot

# Informative Prior
inf_prior <- function(alpha, beta){
  return(dnorm(alpha, 28.87, 2.75)*dnorm(beta, -0.92, 0.44)) #
}

# Alpha/Beta Matrix
ab <- matrix(0, 100, 100)
alpha_grid <- seq(20, 40, length=100) #with mean falling approx in between
beta_grid <- seq(-2, 1, length=100) #with mean falling approx in between

# Contour Plot
for (i in 1:100){
  for (j in 1:100){
    ab[i,j] <- inf_prior(alpha_grid[i], beta_grid[j])
  }
}

contour(alpha_grid, beta_grid, ab, xlab="alpha", ylab="beta", xlim=c(20, 40))
```



c) Uninformative Prior:

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

Likelihood:

$$p(y|\alpha, \beta) \propto \prod_{i=1}^{10} (\alpha + \beta t_i)^{y_i} e^{-(\alpha+\beta t_i)}$$

Posterior:

$$p(\alpha, \beta|y) \propto p(\alpha, \beta) \cdot p(y|\alpha, \beta) \propto 1 \cdot e^{-(n\alpha+\beta\sum_{i=1}^{10} t_i)} \prod_{i=1}^{10} (\alpha + \beta t_i)^{y_i}$$

Sufficient Statistics:

$$(y_i, t_i), \text{ for } i = 1, 2, ..., 10$$

d) The posterior density, $p(\alpha, \beta|y)$, is proper if it integrates to a finite quantity. That is,

$$\int \int p(\alpha, \beta|y) d\alpha d\beta < \infty$$

Since $p(\alpha, \beta|y)$ has an exponential term raised to a growing negative powe, this term will decrease quicker than the product of polynomials will grow. Thus, the integral must converge to some positive value, meaning that $p(\alpha, \beta|y)$ is necessarily proper.

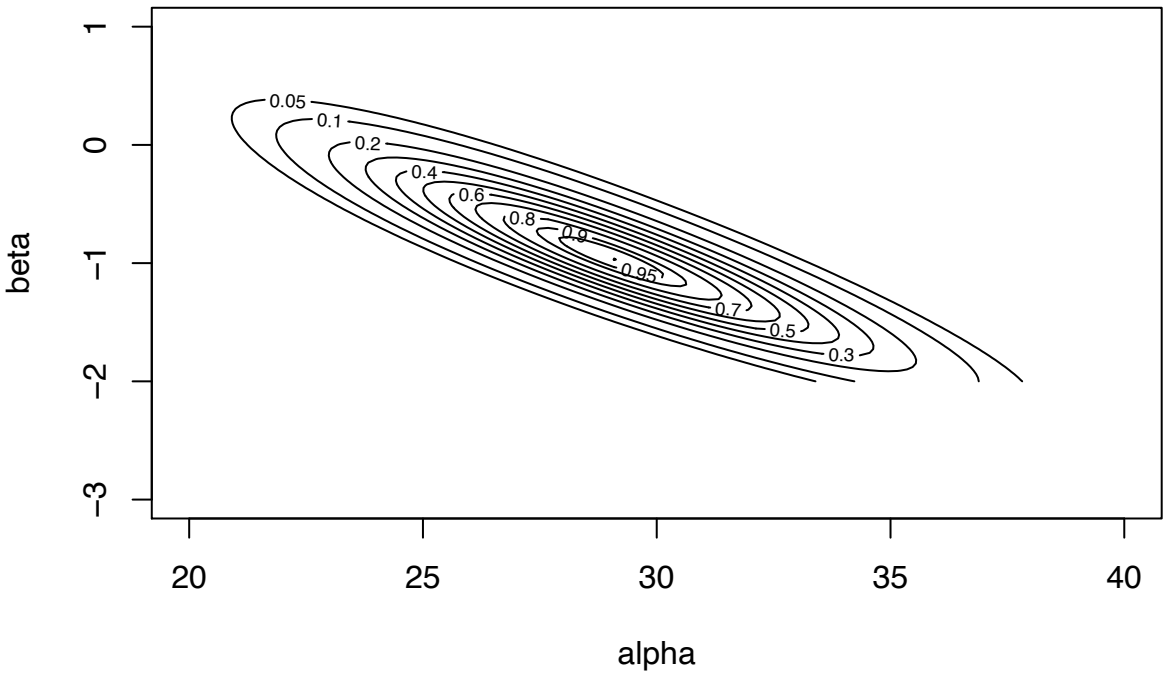
e) Done in part (b)

$$\alpha \sim \text{N}(28.87, 2.75) \Rightarrow \alpha \approx 29.87$$
$$\beta \sim \text{N}(-0.92, 0.44) \Rightarrow \beta \approx -0.92$$

f) Posterior:

$$p(\alpha, \beta|y) \propto p(\alpha, \beta) \cdot p(y|\alpha, \beta) \propto e^{-(n\alpha+\beta\sum_{i=1}^{10} t_i)} \prod_{i=1}^{10} (\alpha + \beta t_i)^{y_i}$$

```
############################### QUESTION 5: F ####################################

# Contour Plot

# Posterior
post <- function(alpha, beta){
  dens <- 0
  for (i in fa$t_i){
    dens <- dens - (alpha+beta*i) +
      fa$y_i[i]*log((alpha+beta*i)) -
      log(factorial(fa$y_i[i])) #log of density
  }
  return(dens)
}

# Contour Plot
for (i in 1:100){
  for (j in 1:100){
    ab[i,j] <- post(alpha_grid[i], beta_grid[j])
  }
}
ab2 <- exp(ab-max(ab))
contour(alpha_grid, beta_grid, ab2, xlab="alpha", ylab="beta",
        xlim = c(20, 40), ylim=c(-3, 1), levels=c(0.05,0.95, seq(0,1,0.1)))
```



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

```
################################## QUESTION 5: G #####################################

set.seed(47)

ab_vec <- c(ab2) #vector of values

# Sampling
samples <- sample(length(alpha_grid)*length(beta_grid),
                      length(alpha_grid),
                      replace=T, prob=ab_vec)

alpha <- rep(0, length(alpha_grid))
beta <- rep(0, length(beta_grid))

for (i in 1:100){
  j <- samples[i]%/%100
  k <- samples[i]%%100
  j <- j + 1
  if (k==0){k=100; j=j-1}
  alpha[i]=alpha_grid[k]
  beta[i]=beta_grid[j]
}

# Histogram of Expected Fatal Accidents in 1986
hist(alpha+11*beta,
     main="Histogram of Expected Fatal Accidents in 1986",
     xlab="alpha + (1986 x beta)")
```

h) A 95% predictive interval for the number of fatal accidents in 1986 is $\approx [10, 30]$.

```
################################ QUESTION 5: H ####################################

set.seed(47)

hat_y <- rpois(1000, alpha+beta*11)
quantile(hat_y, c(0.025, 0.975))
```

```
##  2.5% 97.5%
##    10    30
```

i) The posterior distributions obtained through the hypothetical informative and noninformative priors differ in that the former assumes independence between α and β, yielding very similar and symmetric results (observed in the plots above), while the latter, in not making this assumption, produces less "uniform" outcomes that reflect the (perhaps more accurate) dependency between the two parameters.