ADA HW1

Ao Liu, al3472

1.

Let eta **denote the median of a random variable X. Consider testing** $H_0 : \eta = 0$ **against**
$H_a : \eta \neq 0$ **using** $X_1, X_2, ..., X_{25}$**, a random sample of size** $n = 25$ **from the distribution of** X
(a) Let S **denote the sign statistic. Determine the level of the test that rejects** H_0 **is**
$S \geq 16$**.**

```
1  y<-c(10,10,12,11,9)
2  length(y)
```

[1] 5

```
1  mean(y)
```

[1] 10.4

```
1  var(y)
```

[1] 1.3
so we get:

$$p(\sigma^2|y) \sim \text{Inv-}\chi^2(4, 1.3)$$

$$p(\mu|\sigma^2, y) \sim N(10.4, \frac{\sigma^2}{5})$$

Thus, the joint posterior distribution is:

$$P(\mu, \sigma^2|y) \propto p(\sigma^2|y)p(\mu|\sigma^2, y)$$

(b)

If we treat the measurement as rounded, then the true measurement are regarded as missing
values. Let y_i be the true value of the 5 measurements, $i = 1, 2, 3, 4, 5$. Then, according to the
observed rounded measurements, we have:

$$9.5 < y_1 \leq 10.5$$
$$9.5 < y_2 \leq 10.5$$
$$11.5 < y_3 \leq 12.5$$
$$10.5 < y_4 \leq 11.5$$
$$8.5 < y_5 \leq 9.5,$$

the distribution for y_i should be

$$\int_{y_i-5}^{y_i+5} N(t|\mu, \sigma^2)dt$$

Assume a noninformative prior distribution for the parameters:

$$p(\mu, \sigma) \propto (\sigma^2)^{-1}$$

thus, the posterior distribution for the parameters is:

$$p(\mu, \sigma^2|y_1, ..., y_5) \propto \frac{1}{\sigma^2} \prod_{i=1}^{5} \int_{y_i-5}^{y_i+5} N(t|\mu, \sigma^2)dt$$

1

(c)

Firstly we draw the contour plot for the incorrect model:

```
#input here
```

Then we draw the plot for the correct model:

```
post.a <- function(mu,sd,y){
  ldens <- 0
  for (i in 1:length(y)) ldens <- ldens +
      log(dnorm(y[i],mu,sd))
  ldens}
post.b <- function(mu,sd,y){
  ldens <- 0
  for (i in 1:length(y)) ldens <- ldens +
      log(pnorm(y[i]+0.5,mu,sd) - pnorm(y[i]-0.5,mu,sd))
  ldens}
summ <- function(x){c(mean(x),sqrt(var(x)),
                    quantile(x, c(.025,.25,.5,.75,.975)))}
nsim <- 2000
y <- c(10,10,12,11,9)
n <- length(y)
ybar <- mean(y)
s2 <- sum((y-mean(y))^2)/(n-1)
mugrid <- seq(3,18,length=200)
logsdgrid <- seq(-2,4,length=200)
contours <- c(0.0001,.001,.01,seq(.05,.95,.05))
logdens <- outer (mugrid, exp(logsdgrid), post.a, y)
dens <- exp(logdens - max(logdens))
contour (mugrid, logsdgrid, dens, levels=contours,
        xlab="mu", ylab="log(sigma)", labex=0, cex=2)
mtext ("Posterior density, ignoring rounding", 3)
sd <- sqrt((n-1)*s2/rchisq(nsim,4))
mu <- rnorm(nsim,ybar,sd/sqrt(n))
print (rbind (summ(mu),summ(sd)))
```

```
mu<-seq(3,18,length=200)
lgsigma<-seq(-2,4,length=200)
matrix<-matrix(0,nrow=length(mu),ncol=length(lgsigma))
up<-c(10.5,10.5,12.5,11.5,9.5)
low<-c(9.5,9.5,11.5,10.5,8.5)

for (i in 1:length(mu)){
  for (j in 1:length(lgsigma)){
    matrix[i,j]<--2*lgsigma[j]+sum(log(pnorm(up,mu[i],exp(lgsigma[j]))-
                                    pnorm(low,mu[i],exp(lgsigma[j]))))
  }
}
contours <- c(0.000002,0.00001,.0001,.001,.01,seq(.05,.95,.05))
matrix<-matrix-max(matrix)
contour(mu,lgsigma,exp(matrix),levels=contours,
        xlab="mu", ylab="log(sigma)", labex=0, cex=2)
mtext ("Posterior density, not ignoring rounding", 3)
```

According to the plot shown above, the contour plots look similar, there's almost no difference in terms of the means, the difference lies in the variance, where the correct posterior distribution has a lower variance, the wrong posterior has a higher variance.

(d)

Given μ and σ, the conditional distribution of z_i is $N(\mu, \sigma^2)$, truncated to fall in the range $(y_i - 0.5, y_i + 0.5)$. For each of our posterior draws of (μ, σ), we draw a vector $(z_i, ..., z_5)$ from independent truncated normal distributions using the inverse-cdf method. According to the calculation, the posterior mean of $(z_1 - z_2)^2 = 0.16$

2. *Chapter 8 Exercise 5, BDA3, Gelman et al.*

(a)

The table of units by measurements is sketched as follows:

table of units by measurements					
block	number	trt1	trt2	trt3	trt4
1	1				
1	2				
1	3				
1	4				
1	5				
2	1				
2	2				
2	3				
2	4				
2	5				
3	1				
3	2				
3	3				
3	4				
3	5				
4	1				
4	2				
4	3				
4	4				
4	5				
5	1				
5	2				
5	3				
5	4				
5	5				

1. There are 4 trials in each block, but only one treatment observed for each trial, so y_{obs} is the 20 data shown in table 8.6

2. The missing data is the outcomes of the other 3 treatments in each unit

3. x_i is unit i's block membership, $x_i \in$ {1,2,3,4,5}

4. I is a 20×4 matrix, with each column representing a treatment, each of its elements could be described as:
$$I_i^k, i = 1, 2, ..., 20, k = \text{A,B,C,D}$$
which must satisfy:
$$I_i^k = 0 \text{ or } 1, I_i^A + I_i^B + I_i^C + I_i^D = 1$$
what's more, for each j=1,2,3,4,5 and each k $\in$ {A,B,C,D}:
$$\sum_{i:x_i=j} I_i^k = 1$$

4

5. The total number of measurements N = 80

6. There are 20 observed measurement and 60 unobserved measurements in this problem

(b)

$$p(I|x) = \begin{cases} (\frac{1}{4!})^5 & \text{for } I \text{ satisfying the conditions} \\ 0 & \text{o.w.} \end{cases}$$

I is ignorable, known and strongly ignorable. The propensity scores are an adequate summary.

(c)

We use the **Randomized Complete Block** Model (RCB):

1. There are 4 treatments to compare

2. 5 different blocks (groups of similar units)

3. 4 units in each block (20 units)

4. within each block, randomly assign one unit to each treatment

5. let y_i^k be the outcome for unit in block i that got treatment k.

$$y_i^k = \mu + \beta_i + \tau^k + e_{ik}, \ \ e_{ij} \sim \text{i.i.d } N(0, \sigma^2)$$

Thus,

$$\text{E}(y_i^k) = \mu + \beta_i + \tau^k$$

Let $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix}$ be the vector of block effect

$$p(\beta) \propto \begin{cases} 1 & \beta_1 + ... + \beta_5 = 0 \\ 0 & else \end{cases}$$

(d)

If we average over the block effect, then the vector β is 0,

$$E(y^k) = \mu + \tau^k, \ \ k \in \{\text{A,B,C,D}\}$$

3. *Chapter 8 Exercise 7, BDA3, Gelman et al.*

(a)

We can write the posterior distribution of μ, σ as:

$$\sigma^2 | y^{obs} \sim \text{Inv-}\chi^2(n-1, s^2)$$

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

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

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

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

.

averaging over μ yields:

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

.

Since σ^2 has an inverse-χ^2 posterior distribution, averaging over it yields a t distribution:

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

The complete-data mean is $\bar{y} = \frac{n}{N}\bar{y}^{obs} + \frac{N-n}{N}\bar{y}^{mis}$, which is a linear transformation of $\bar{y}^{mis}$ and so it has a t posterior distribution also:

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

$$= t_{n-1}(\bar{y}^{obs}, s^2\frac{(N-n)}{nN})$$

$$= t_{n-1}(\bar{y}^{obs}, s^2(\frac{1}{n} - \frac{1}{N}))$$

4. *Chapter 8 Exercise 11, BDA3, Gelman et al.*

  (a)

Let y be the number of fishes that are caught tagged during the second time, and N be the total number of fishes. According to the question, the fisherman kept on catching fishes until he catches 20 tagged fishes, thus, y follows:

$$p(y|N) = \frac{\binom{100}{19}\binom{N-100}{70}}{\binom{N}{89}} \times \frac{81}{N-89}$$

Let N follow a uninformative uniform distribution, thus, the posterior distribution for N is:

$$p(N|y) \propto p(y|N)p(N)$$
$$\propto p(y|N)$$
$$= p(y|N)$$
$$= \frac{\binom{100}{19}\binom{N-100}{70}}{\binom{N}{89}} \times \frac{81}{N-89}$$

  (b)

Since the fisherman tagged 100 fishes at the first time and during the second time he caught 70 fished with no tag, which means there are at least 170 fishes in the pond. Thus, to check if the posterior distribution is proper, we can focus on the summation of the posterior distribution.

```
1  pny = function(n){
2     choose(100,19)*choose(n-100,70)/choose(n,89)*81/(n-89)
3  }
4  sum(pny(c(170:1000)))
```

[1] 5.260337

$$\sum_{N=170}^{1000} p(y|N) = 5.260337$$

so it's easy to conlude that:

$$\sum_{N=170}^{\infty} p(y|N) < \infty$$

Thus, we arrive at the conclusion that the posterior distribution is proper.

  (c)

We give the following definition:

$$\tilde{y} = \begin{cases} 1, & \text{if the next fish is tagged} \\ 0, & \text{otherwise} \end{cases}$$

Thus, the probability that the next fish caught by the fisherman is tagged is:

$$p(\tilde{y} = 1|y) = \sum_{N=170}^{\infty} p(\tilde{y} = 1|N)p(N|y)$$

$$= \sum_{N=170}^{\infty} \frac{80}{N - 90} p(N|y)$$

(d)

Now the condition has been changed to: of all the 90 fish caught, 15 fish caught are tagged, 70 are untagged the other 5 are uncertain. Now we define the following notation:

$$\begin{cases} y_{mis} : & \text{the tagged fish among the 5 uncertain fish} \\ y_{obs} : & \text{the 15 tagged fish} \end{cases}$$

then we know that

$$y_{mis} \sim \text{Binomial}(5, \frac{15}{85})$$

$$p(N|y_{obs}) = \sum_{k=0}^{5} p(N|y_{obs}, y_{min} = k)p(y_{mis} = k|y)$$

$$= \sum_{k=0}^{5} p(N|5 + k \text{ tagged}, 75 - k \text{ untagged}) \binom{5}{k} (\frac{15}{85})^k (\frac{70}{85})^{5-k}$$

Then we plot the posterior distribution of N:

```
1  pny = function(n,n1,n2){
2    choose(100,n1-1)*choose(n-100,n2)/choose(n,n1+n2-1)*(101-n1)/(n-n1-n2+1)
3  }
4  n = seq(175,1000)
5  result = rep(0,length(n))
6  k = c(0:5)
7  for (i in 1:length(n)){
8    result[i] = sum(pny(n[i],15+k,75-k)*dbinom(k,5,15/85))
9  }
10 plot(result~n,type="l",ylab="posterior distribution")
11 title("Posterior Distribution of N")
```

5. *Chapter 10 Exercise 4, BDA3, Gelman et al.*

(a)

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

$$p(\theta \leq \theta^* | \theta \text{ is accepted}) = \frac{p(\theta \leq \theta^* \text{and } U \leq \frac{p(\theta|y)}{Mg(\theta)})}{p(U \leq \frac{p(\theta|y)}{Mg(\theta)})}$$

$$= \frac{\int_{-\infty}^{\theta^*} \int_0^{\frac{p(\theta|y)}{Mg(\theta)}} g(\theta)dud\theta}{\int_{-\infty}^{\infty} \int_0^{\frac{p(\theta|y)}{Mg(\theta)}} g(\theta)dud\theta}$$

$$= \frac{\frac{1}{M}\int_{-\infty}^{\theta^*} p(\theta|y)d\theta}{\frac{1}{M}}$$

Which is the cumulative density function for $p(\theta|y)$. A similar argument works in higher dimensions.

(b)

The above proof requires that

1. $g(\theta) > 0$ whereever $p(\theta) > 0$

2. $\frac{p(\theta)}{Mg(\theta)} \leq 1$ always

If $\frac{p}{g}$ is unbounded, then one or both of the requirements above will be violated.