# Survmeth 895 Midterm

*David (Daiwei) Zhang*

*March 14, 2017*

## Problem 1(a)

We use the model that the number of patients experiencing pain follows a binomial distribution, with the rate parameter $p$ different for each of the six cells. Moreover, we use Jeffrey's prior Beta(0.5, 0.5) for $p$. Thus we can simulate $p$ for each cell and find the proportion of patients expeiriencing pain for each anesthetic in a large population by weighting $p$'s for both strata.

```
set.seed(3)
strata.num <- 2
anes.num <- 3
pain <- matrix(c(11, 9, 8, 4, 6, 5),
               strata.num, anes.num)
samp.size <- matrix(c(23, 29, 24, 31, 29, 37),
                    strata.num, anes.num)
nopain <- samp.size - pain
beta.a <- matrix(0.5, strata.num, anes.num)
beta.b <- beta.a
beta.a <- beta.a + pain
beta.b <- beta.b + nopain
strata.prop <- c(0.3, 0.7)
sim.num <- 1e4
pain.risk <- matrix(0, sim.num, anes.num)
for(j in 1:anes.num){
    pain.risk[,j] <- rbeta(sim.num, beta.a[1,j], beta.b[1,j]) * strata.prop[1] +
      rbeta(sim.num, beta.a[2,j], beta.b[2,j]) * strata.prop[2]
}
anesA.risk <- pain.risk[,1]
mean(anesA.risk)
```

```
## [1] 0.3642747
```

```
quantile(anesA.risk, c(0.025, 0.975))
```

```
##      2.5%     97.5%
## 0.2417011 0.5008753
```

## Problem 1(b)

```
anesB.risk.relat <- pain.risk[,2] / anesA.risk
anesC.risk.relat <- pain.risk[,3] / anesA.risk
mean(anesB.risk.relat)
```

```
## [1] 0.5712334
```

```
quantile(anesB.risk.relat, c(0, 0.95))
```

```
##        0%       95%
## 0.1391089 0.9085585
```

```r
mean(anesB.risk.relat < 1)
```

```
## [1] 0.9756
```

```r
mean(anesC.risk.relat)
```

```
## [1] 0.4727007
```

```r
quantile(anesC.risk.relat, c(0, 0.95))
```

```
##        0%       95%
## 0.1105922 0.7661633
```

```r
mean(anesC.risk.relat < 1)
```

```
## [1] 0.9937
```

We can see that Anesthetic B has a 0.9756 probability of having a lower risk of pain than Anesthetic A. Anesthetic C has a 0.9937 probability of having a lower risk of pain than Anesthetic A. Thus both Anesthetic B and C are preferred over A. This choice is also supported by the means and credibility intervals.

## Problem 2(a)(b)

We use the model that the number of respondents for Protocol A follows a binomial distribution. Moreover, we use Jeffrey's prior Beta(0.5, 0.5) for the rate parameter $p$. Then the posterior is a Beta-Binomial combination. As for the cost, we use a Poisson model for the cost of each subject, with the parameter $\lambda$ different for respondents and non-respondents. Furthermore, we use the Jeffrey's prior $\pi(\lambda) \propto \lambda^{-1/2}$. Then $\pi(\lambda|\mathbf{x})$ follows a Gamma($\sum x_i - \frac{1}{2}, \frac{1}{n}$) distribution. Thus we can simulate $p$ and use $p$ to draw whether a subject is respondent. Then we use this information to draw the cost for this subject from the appropriate Poisson distribution. Finally, we add up all the cost to get the total.

```r
proto <- function(resp, cost){
  surv.num <- 5000
  noresp <- 1 - resp
  resp.num <- sum(resp)
  noresp.num <- sum(noresp)
  beta.par <- c(0.5, 0.5)
  beta.par <- beta.par + c(resp.num, noresp.num)
  sim.num <- 5e3
  p <- rbeta(sim.num, beta.par[1], beta.par[2])
  surv.resp <- matrix(rbinom(sim.num * surv.num, 1, p), sim.num, surv.num)
  surv.resp.num <- rowSums(surv.resp)
  surv.noresp.num <- surv.num - surv.resp.num
  cost.resp <- cost[resp == 1]
  cost.noresp <- cost[resp == 0]
  gamma.par.resp <- c(sum(cost.resp) - 0.5, 1/resp.num)
  gamma.par.noresp <- c(sum(cost.noresp) - 0.5, 1/noresp.num)
  lambda.resp <- rgamma(sim.num, gamma.par.resp[1], scale = gamma.par.resp[2])
  lambda.noresp <- rgamma(sim.num, gamma.par.noresp[1], scale = gamma.par.noresp[2])
  lambda <- lambda.resp * (surv.resp == 1) + lambda.noresp * (surv.resp == 0)
  surv.cost <- matrix(rpois(sim.num * surv.num, lambda), sim.num, surv.num)
  surv.cost.total <- rowSums(surv.cost)
  return(list(
    surv.resp.num = surv.resp.num,
    surv.cost.total = surv.cost.total
  ))
```

```
}
resp.A <- c(1,0,0,1,0,1,0,1,1,0,0,1,1,1,0,1,1,1,0,1,1,1,1,1,0,
            0,1,1,1,1,1,1,1,0,0,0,1,1,0,1,1,1,1,1,1,0,1,1,1,1)
cost.A <- c(32,30,36,38,48,44,32,33,27,41,30,46,29,27,35,33,47,36,29,31,34,47,39,29,32,
            42,31,43,35,30,33,36,29,31,35,42,31,26,36,32,32,32,31,36,31,37,37,47,31,29)
proto.A <- proto(resp.A, cost.A)
mean(proto.A$surv.resp.num)
```

```
## [1] 3481.82
```

```
quantile(proto.A$surv.resp.num, c(0.025, 0.975))
```

```
##    2.5%   97.5%
## 2833.00 4072.05
```

```
mean(proto.A$surv.cost.total)
```

```
## [1] 173840.3
```

```
quantile(proto.A$surv.cost.total, c(0.025, 0.975))
```

```
##     2.5%    97.5%
## 165666.9 182253.4
```

## Problem 2(c)

```
resp.B <- c(0,0,0,1,0,1,1,1,1,1,1,1,1,1,1,0,1,1,1,1,1,1,1,1,1,
            1,1,1,1,0,1,1,0,1,1,1,1,0,1,1,1,1,1,1,1,1,1,1,0,0)
cost.B <- c(38,38,43,37,51,46,47,53,47,47,35,47,45,37,55,45,48,40,45,47,40,47,36,37,47,
            46,43,33,44,39,46,40,42,40,46,35,41,41,49,43,34,47,48,42,44,38,40,37,42,41)
proto.B <- proto(resp.B, cost.B)
mean(proto.B$surv.resp.num)
```

```
## [1] 3968.984
```

```
quantile(proto.B$surv.resp.num, c(0.025, 0.975))
```

```
##     2.5%    97.5%
## 3357.000 4473.025
```

```
mean(proto.B$surv.cost.total)
```

```
## [1] 213822.7
```

```
quantile(proto.B$surv.cost.total, c(0.025, 0.975))
```

```
##     2.5%    97.5%
## 204822.1 223405.0
```

## Problem 2(d)

```
cost.each.A <- proto.A$surv.cost.total / proto.A$surv.resp.num
cost.each.B <- proto.B$surv.cost.total / proto.B$surv.resp.num
cost.each.diff <- cost.each.B - cost.each.A
mean(cost.each.diff)
```

```
## [1] 3.786987
```

```r
quantile(cost.each.diff, c(0.025, 0.975))
```

```
##      2.5%     97.5%
## -9.660469 16.691991
```

```r
mean(cost.each.diff > 0)
```

```
## [1] 0.737
```

Then Protocol A is recommended since it is more likely to have a lower cost per respondent than Protocal B.

## Problem 4

Let $Y_{INC,i}$ be the number of accounts with errors in the sample (size: $k$) in month $i$, and let $Y_{EXC,i}$ be the number of accounts with erros in the non-sampled accounts (size: $K-k$) in month $i$. Moreover, let $Y_i = Y_{INC,i} + Y_{EXC,i}$. We are interested in $\sum_{i=1}^{k} Y_i$. Our model is

$$
\begin{aligned}
Y_i &\sim Binom(K, \theta_i) \\
Y_{INC_i} &\sim Binom(k, \theta_i) \\
\theta_i &\sim Beta(0.5, 0.5) \\
\theta_i | Y_{INC,i} &\sim Beta(Y_{INC,i} + 0.5, Y_{EXC,i} + 0.5) \\
Y_{EXC,i} &\sim Binom(K-k, \theta_i)
\end{aligned}
$$

Thus for the simulation, we can first draw $\theta_i$ and use it to draw $Y_{EXC,i}$. Then we can obtain $Y_i = Y_{INC,i} + Y_{EXC,i}$.

```r
k <- 400
K <- 20000
n <- 24
y.inc.err <- c(0, 0, 1, 1, 2, 4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7, 8, 9, 9, 10, 10, 13, 14, 17)
y.inc.noerr <- k - y.inc.err
y.inc <- rbind(y.inc.err, y.inc.noerr)
beta.par <- matrix(0.5, 2, n)
beta.par <- beta.par + y.inc
sim.num <- 1e4
theta <- matrix(0, sim.num, n)
y.exc.err <- matrix(0, sim.num, n)
for (i in 1:n){
  theta[,i] <- rbeta(sim.num, beta.par[1, i], beta.par[2,i])
  y.exc.err[,i] <- rbinom(sim.num, K-k, theta[,i])
  y.err <- sweep(y.exc.err, 2, y.inc.err, "+")
}
y.total.err <- rowSums(y.err)
mean(y.total.err)
```

```
## [1] 8265.859
```

```r
quantile(y.total.err, c(0.025, 0.975))
```

```
##     2.5%    97.5%
## 7092.000 9547.025
```

As for simulating the number of accounts with errors for the next two-year period, we can use the model $\theta_i \overset{iid}{\sim} Beta(a,b)$ for $1 \leq i \leq 24$. We use some non-informative prior $p(a,b)$ for $a$ and $b$ and update it with the $\theta_i$ that we have already simulated to obtain the posterior $p(a,b|\theta) \propto \mathbf{p(a,b)p}(\theta|\mathbf{a},\mathbf{b})$. Then we simulate $\theta_i$ for $1 \leq i \leq 24$ from the posterior of $(a,b)$ and simulate the number of accounts with errors for each $\theta_i$ with the binomial distribution. Finally, we add up the number of accounts with erros for all the 24 months to get the total.