ST4234: Bayesian Statistics

Tutorial 5 Solution, AY 19/20

Solutions

1. (a)

$$\begin{aligned}
p(p_N, p_S|y_N, y_S) &\propto p(y_N, y_S|p_N, p_S)p(p_N, p_S) \\
&\propto p(y_N|p_N)p(y_S|p_S) \\
&\quad \text{[because } y_N \text{ and } y_S \text{ are independent and } p(p_n, p_S) \propto 1] \\
&\propto \underbrace{p_N^{y_N}(1-p_N)^{n_N-y_N}}_{\text{terms in } p_N \text{ only}} \underbrace{p_S^{y_S}(1-p_S)^{n_S-y_S}}_{\text{terms in } p_S \text{ only}}.
\end{aligned}$$

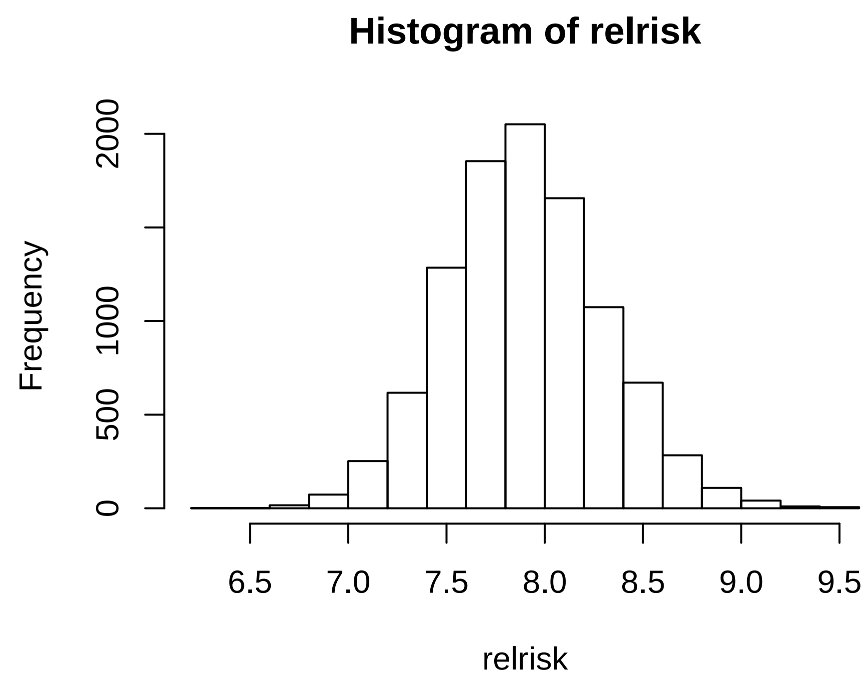
Therefore $p(p_N, p_S|y_S, y_N) = p(p_N|y_N)p(p_S|y_S)$ where
$p_N|y_N \sim \text{Beta}(y_N + 1, n_N - y_N + 1) = \text{Beta}(1602, 162528)$ and
$p_S|y_S \sim \text{Beta}(y_S + 1, n_S - y_S + 1) = \text{Beta}(511, 412369)$.
Hence p_N and p_S have independent beta posterior distributions.

(b) (i) The R-code below keys in the data and generates 10000 values from the joint posterior distribution of (p_N, p_S). Since p_N and p_S have independent beta posterior distributions, we can generate samples from $p(p_S|y_S)$ and $p(p_N|y_N)$ independently.

```r
yN <- 1601
nN <- 162527 + 1601
yS <- 510
nS <- 412368 + 510
S <- 10000
set.seed(1)
pN_draws <- rbeta(S,yN+1, nN-yN+1)
pS_draws <- rbeta(S,yS+1, nS-yS+1)
```

The R-code below constructs a histogram of the relative risk p_N/p_S and computes a 95% quantile-based interval estimate of this relative risk.

```r
relrisk <- pN_draws/pS_draws
hist(relrisk)
```



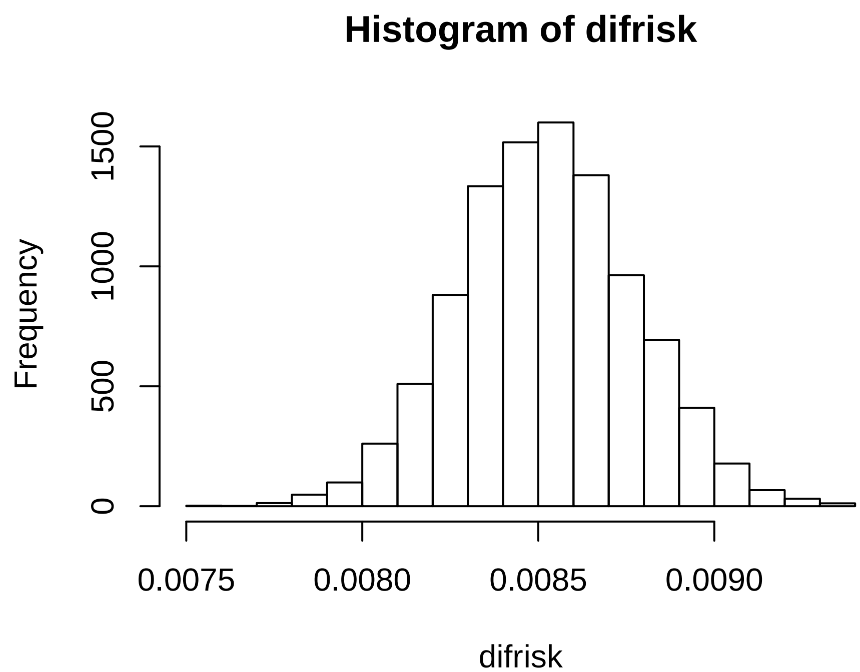

```r
quantile(relrisk,c(0.025,0.975))
```

```
##     2.5%    97.5%
## 7.153585 8.723457
```

The 95% confidence interval for the relative risk is (7.154, 8.723).

(ii) The R-code below constructs a histogram of the difference in risks $p_N - p_S$, which is shown in the figure below. Note that all the mass of the histogram is on positive values.

```r
difrisk <- pN_draws - pS_draws
hist(difrisk)
```



An estimate of the posterior probability that the difference in risks exceeds 0 is 1.

```r
mean(difrisk>0)

## [1] 1
```

2. (a) Let Y denote the amount of time students from a high school spent on studying or homework during an exam period. Assume $Y \sim \text{N}(\theta, \sigma^2)$ and a conjugate prior, $\theta \sim \text{N}(\mu_0, \sigma^2/n_0)$, $\sigma^2 \sim \text{Inv-Gamma}(\nu_0/2, S_0/2)$ where $\{\mu_0 = 5, n_0 = 1, \nu_0 = 2, S_0 = 8\}$. Then the joint posterior distribution is given by

$$p(\theta, \sigma^2|\boldsymbol{y}) = p(\theta|\sigma^2, \boldsymbol{y})p(\sigma^2|\boldsymbol{y}),$$

where

$$\theta|\sigma^2, \boldsymbol{y} \sim \mathrm{N}\left(\mu_1, \frac{\sigma^2}{n_1}\right), \qquad \sigma^2|\boldsymbol{y} \sim \text{Inv-Gamma}\left(\frac{\nu_1}{2}, \frac{S_1}{2}\right),$$

$$\mu_1 = \frac{n\bar{y} + n_0\mu_0}{n + n_0}, \qquad \nu_1 = \nu_0 + n,$$

$$n_1 = n + n_0, \qquad S_1 = S_0 + S + \frac{nn_0(\bar{y} - \mu_0)^2}{n + n_0}.$$

The values of μ_1, n_1, ν_1, S_1 are computed below.

```
school1 <- read.table("datasets/school1.txt",header=FALSE)
school2 <- read.table("datasets/school2.txt",header=FALSE)
school3 <- read.table("datasets/school3.txt",header=FALSE)

y <- NULL # create a list where elements can have different lengths
y[[1]] <- school1$V1
y[[2]] <- school2$V1
y[[3]] <- school3$V1

ymean <- rep(0,3)
n <- rep(0,3)
S <- rep(0,3) # sum of squares
for (i in 1:3){
  ymean[i] <- mean(y[[i]])
  n[i] <- length(y[[i]])
  S[i] <- var(y[[i]])*(n[i]-1)
}

mu0 <- 5 # prior parameters
n0 <- 1
nu0 <- 2
S0 <- 8


n
```

```
## [1] 25 23 20

(n1 <- n + n0)

## [1] 26 24 21

(mu1 <- (n*ymean + n0*mu0)/n1)

## [1] 9.292308 6.948750 7.812381

(nu1 <- nu0 + n)

## [1] 27 25 22

(S1 <- S0 + S + n*n0*(ymean-mu0)^2/n1)

## [1] 389.4737 454.4549 288.0564
```

We can generate samples from the joint posterior distribution $p(\theta, \sigma^2|\boldsymbol{y})$ by first simulating $\sigma^{2(1)}, \ldots, \sigma^{2(M)} \overset{\text{i.i.d.}}{\sim} p(\sigma^2|\boldsymbol{y})$ and then $\theta^{(m)} \sim p(\theta|\sigma^{2(m)}, \boldsymbol{y})$ for $m = 1, \ldots, M$. The posterior means and standard deviations can be computed based on the Monte Carlo samples $\{\theta^{(1)}, \ldots, \theta^{(M)}\}$ and $\{\sigma^{(1)}, \ldots, \sigma^{(M)}\}$. Using $M = 100000$, we have the following estimates.

	Posterior mean
θ_1	9.29
θ_2	6.95
θ_3	7.81
σ_1	3.91
σ_2	4.40
σ_3	3.75

```
require(invgamma)

## Loading required package:  invgamma

M <- 100000
```

```r
sigma <- matrix(0,3,M)
theta <- matrix(0,3,M)
Ytilde <- matrix(0,3,M)
means.sigma <- rep(0,3)
means.theta <- rep(0,3)

set.seed(1)
for (i in 1:3){
  sigma[i,] <- sqrt(rinvgamma(M, nu1[i]/2, S1[i]/2))
  theta[i,] <- rnorm(M, mu1[i], sigma[i,]/sqrt(n1[i]))
  Ytilde[i,] <- rnorm(M, theta[i,], sigma[i,])
  means.sigma[i] <- mean(sigma[i,])
  means.theta[i] <- mean(theta[i,])
}
means.theta

## [1] 9.292139 6.950262 7.813906

means.sigma

## [1] 3.907001 4.398306 3.751229
```

Alternatively, you can also use formulas from the normal-inverse gamma distribution to calculate the posterior means and standard deviations.

(b) The posterior probability that $\theta_3 < \theta_2 < \theta_1$ can be estimated using the Monte Carlo samples $\{\theta_j^{(1)}, \ldots, \theta_j^{(M)}\}$ for $j = 1, 2, 3$, as $\frac{1}{M}\sum_{m=1}^{M} \mathbb{1}\{\theta_3^{(m)} < \theta_2^{(m)} < \theta_1^{(m)}\}$. The estimate of this posterior probability is 0.218.

```r
mean((theta[3,] < theta[2,]) & (theta[2,] < theta[1,]))

## [1] 0.21843
```

(c) Using the Monte Carlo samples $\{\theta_i^{(1)}, \ldots, \theta_i^{(M)}\}$ and $\{\sigma_i^{2(1)}, \ldots, \sigma_i^{2(M)}\}$, we can generate samples from the posterior predictive distribution of school i using $\tilde{Y}_i^{(m)} \sim$

$\mathrm{N}(\theta_i^{(m)}, \sigma_i^{2(m)})$ for $m = 1, \ldots, M$. The posterior probability that $\tilde{Y}_3 < \tilde{Y}_2 < \tilde{Y}_1$ estimate is 0.200.

```
mean((Ytilde[3,] < Ytilde[2,]) & (Ytilde[2,] < Ytilde[1,]))

## [1] 0.19995
```

(d) The posterior probability that θ_1 is bigger than both θ_2 and θ_3 is 0.889, and the posterior probability that $\tilde{Y}_1$ is bigger than both $\tilde{Y}_2$ and $\tilde{Y}_3$ is 0.467.

```
mean((theta[1,] > theta[2,]) & (theta[1,] > theta[3,]))

## [1] 0.88926

mean((Ytilde[1,] > Ytilde[2,]) & (Ytilde[1,] > Ytilde[3,]))

## [1] 0.46714
```