Stat 570 Final

Dongyang Wang

2022-12-03

Q1

a

During each cycle, the likelihood can be calculated by multiplying together the probability $p(1-p)^{t-1}$ of pregnancy for Y_t conceptions (to the power of Y_t), and for the last case when we calculate the likelihood for the number of women that have not conceived by cycle N, the probability that they do not conceive for N cycles is $(1-p)^N$, and there are Y_{N+1} of them.

Therefore, we obtain the likelihood function as follows:
$L(p) = f(t=1)f(t=2)\ldots f(t=N)f(t>N) = (p(1-p)^{1-1})^{Y_1}(p(1-p)^{2-1})^{Y_2}\ldots(p(1-p)^{N-1})^{Y_N}((1-p)^N)^{Y_{N+1}} = [\Pi_{t=1}^N(p(1-p)^{t-1})^{Y_t}]((1-p)^N)^{Y_{N+1}}$

b

Log likelihood $l(p) = \log p \sum_{t=1}^N Y_t + \log(1-p)\sum_{t=1}^N Y_t(t-1) + NY_{N+1}\log(1-p)$

Taking derivative wrt p,

$\frac{1}{p}\sum_{t=1}^N Y_t - \frac{1}{1-p}\sum_{t=1}^N Y_t(t-1) - \frac{1}{1-p}NY_{N+1} = 0$

Solving, we have $\hat{p} = \frac{\sum_{t=1}^N Y_t}{\sum_{t=1}^N (Y_t t) + NY_{N+1}}$

c

Taking derivative of the score function, we have $-\frac{1}{p^2}\sum_{t=1}^N Y_t - \frac{1}{(1-p)^2}\sum_{t=1}^N Y_t(t-1) - \frac{1}{(1-p)^2}NY_{N+1}$

Therefore, the observed FIN is $\frac{1}{p^2}\sum_{t=1}^N Y_t + \frac{1}{(1-p)^2}\sum_{t=1}^N Y_t(t-1) + \frac{1}{(1-p)^2}NY_{N+1}$ And theoretically it would be $\frac{N}{p^3} + \frac{N(t-1)}{(1-p)^2 p} + \frac{N}{(1-p)}$

Also, the asymptotic variance is $\frac{1}{FIN} = \frac{1}{\frac{1}{p^2}\sum_{t=1}^N Y_t + \frac{1}{(1-p)2}\sum_{t=1}^N Y_t(t-1) + \frac{1}{(1-p)2}NY_{N+1}}$

d

```
##                          MLE     MLE sd  CI lower  CI upper
## Smoker result    0.2203791 0.01957709 0.1820088 0.2587495
## Nonsmoker result 0.3289382 0.01148539 0.3064273 0.3514492
```

e

$P(Y|p) \propto [\Pi_{t=1}^N(p(1-p)^{t-1})^{Y_t}]((1-p)^N)^{Y_{N+1}}p^a(1-p)^b = p^{\Sigma_{t=1}^N Y_t + a}(1-p)^{\Sigma_{t=1}^N (t-1)Y_t + NY_{N+1} + b} \sim Beta(\Sigma_{t=1}^N Y_t + a, \Sigma_{t=1}^N (t-1)Y_t + NY_{N+1} + b)$

Prior mean pm is $\frac{a}{a+b}$

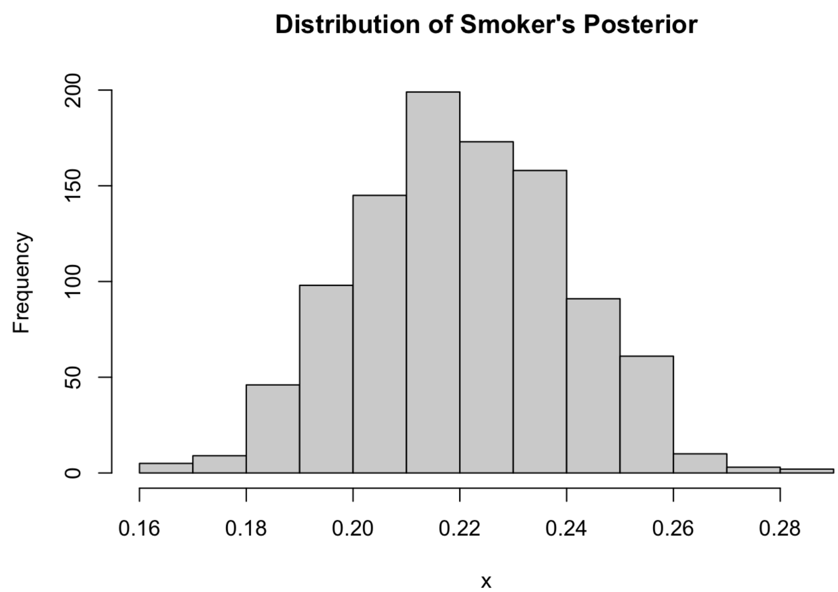
The posterior mean is $\frac{\Sigma_{t=1}^N Y_t + a}{a + \Sigma_{t=1}^N tY_t + NY_{N+1} + b} = \frac{\Sigma_{t=1}^N tY_t + NY_{N+1}}{a + \Sigma_{t=1}^N tY_t + NY_{N+1} + b}\hat{p} + \frac{a+b}{a + \Sigma_{t=1}^N tY_t + NY_{N+1} + b} * pm$

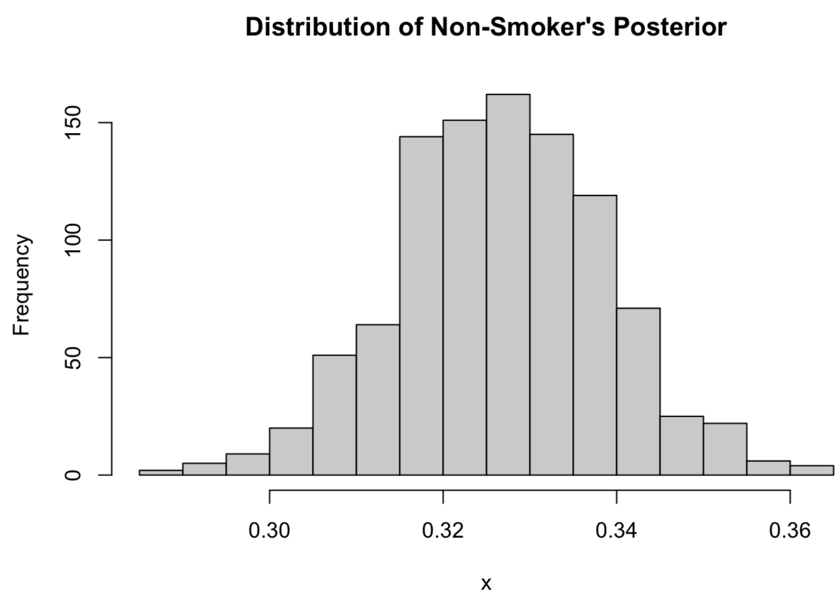
f

$\mu = \frac{a}{a+b}$ and $\sigma^2 = \frac{ab}{(a+b)^2(a+b+1)}$.

Writing $a = \frac{\mu}{1-\mu}b$ we obtain $b = \frac{\mu - \frac{\sigma^2}{1-\mu}}{\frac{\sigma^2}{(1-\mu)^2}}$ so we have $a = \frac{\mu}{1-\mu}b = \frac{\mu}{1-\mu}\frac{\mu - \frac{\sigma^2}{1-\mu}}{\frac{\sigma^2}{(1-\mu)^2}}$

g






```
##                   Posterior a Posterior b Posterior mean Posterior sd  CI lower
## Smoker result           97.8       348.2      0.2192825   0.01957020 0.1809256
## Nonsmoker result       478.8       986.2      0.3268259   0.01225053 0.3028153
##                    CI upper
## Smoker result    0.2576394
## Nonsmoker result 0.3508365
```

Q2

a

By definition,
$P(T = t) = E(p(1-p)^{t-1}) = E((1-p)^{t-1}) - E((1-p)^{t-1}) + E(p(1-p)^{t-1}) = E((1-p)^{t-1}) - E((1-p)^{t-1} - p(1-p)^{t-1}) = E((1-p)^{t-1}) - E((1-p)^{t})$

Also, from the equation above, $P(T \leq t) = P(T = 1) + P(T = 2) + \ldots + P(T = t) = \Sigma_{i=1}^{t} E((1-p)^{i-1}) - E((1-p)^{i}) = 1 - E((1-p)^{t})$ and therefore $P(T > t) = E((1-p)^{t})$

b

$P(T = t|\alpha, \beta) = E(p(1-p)^{t-1}) = \int_0^1 p(1-p)^{t-1} \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)} dp = \frac{1}{B(\alpha,\beta)} \int_0^1 p^{\alpha}(1-p)^{t+\beta-2}\, dp = \frac{B(\alpha+1,\beta+t-1)}{B(\alpha,\beta)} = \frac{\alpha\Gamma(\alpha+\beta)\Gamma(\beta+t-1)}{\Gamma(\beta)\Gamma(\beta+\alpha+t)}$

$P(T > t|\alpha,\beta) = E((1-p)^t) = \int_0^1 (1-p)^t \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)} dp = \frac{B(\alpha,\beta+t)}{B(\alpha,\beta)} = \frac{\Gamma(\alpha+\beta)\Gamma(\beta+t)}{\Gamma(\beta)\Gamma(\beta+\alpha+t)}$

c

Substituting the original parts of the probability in 1(a) for results in 2(a), we switch $L(p) = \{\Pi_{i=1}^{N} [p(1-p)^{t-1}]^{Y_t}\}[(1-p)^N]^{Y_{N+1}}$ to

$L(\alpha,\beta) = \{\Pi_{i=1}^{N} [\frac{\alpha\Gamma(\alpha+\beta)\Gamma(\beta+t-1)}{\Gamma(\beta)\Gamma(\beta+\alpha+t)}]^{Y_t}\}[\frac{\Gamma(\alpha+\beta)\Gamma(\beta+N)}{\Gamma(\beta)\Gamma(\beta+\alpha+N)}]^{Y_{N+1}}$

d

Calculating the MLE, we obtain $l(\alpha,\beta) = -Y_{N+1} \log B(\alpha,\beta) + Y_{N+1} \log B(\alpha,\beta+N) + \sum_{t=1}^{N} Y_t \log B(\alpha+1,\beta+t-1) - \sum_{t=1}^{N} Y_t \log B(\alpha,\beta)$ which is equivalent to $l(\alpha,\beta) = -N \log B(\alpha,\beta) + Y_{N+1} \log B(\alpha,\beta+N) + \sum_{t=1}^{N} Y_t \log B(\alpha+1,\beta+t-1)$

To write in gamma form, we have

$l(\alpha,\beta) = -n(\log\Gamma(\alpha) + \log\Gamma(\beta) - \log\Gamma(\alpha+\beta)) + Y_{N+1}(\log\Gamma(\alpha) + \log\Gamma(\beta+N) - \log\Gamma(\alpha+\beta+N)) + \sum_{t=1}^{N} Y_t(\log\Gamma(\alpha+1) + \log\Gamma(\beta+t-1) - \log\Gamma$
where $n = \sum_{t=1}^{N+1} Y_t$

Then we use the optim function to solve for answers.

```
## [1]  2.022436e+00 -5.551115e-17
```

```
## [1]  1.664322e+00 -5.551115e-17
```

```
##            alpha MLE      beta MLE
## Smoker      2.022436 -5.551115e-17
## Nonsmoker   1.664322 -5.551115e-17
```

Q3

a

$L(p) = \{\Pi_{i=1}^{N} [p(1-p)^{t-1}]^{Y_t}\}[(1-p)^N]^{Y_{N+1}} = \Pi_{i=1}^{N} p^{Y_t}(1-p)^{(t-1)Y_t+Y_{N+1}} \propto \Pi_{t=1}^{N} \binom{tY_t+Y_{N+1}}{Y_t} p^{Y_t}(1-p)^{(t-1)Y_t+Y_{N+1}}$ which follows a product of the binomial distribution.

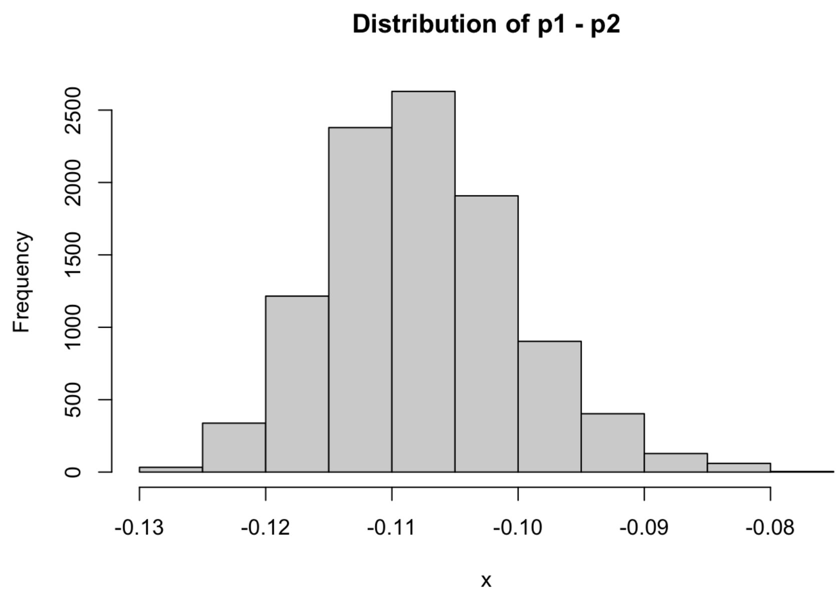
b

The log likelihood is almost the same as calculated in Q1. $l(p) = c + \log p \sum_{t=1}^{N} Y_t + \log(1-p)\sum_{t=1}^{N} Y_t(t-1) + NY_{N+1}\log(1-p)$ where $c = \log \Pi_{t=1}^{N} \binom{tY_t+Y_{N+1}}{Y_t}$. Therefore, after taking derivative wrt p, the result of MLE of p will remain the same. Same logic applies for the fisher information as well as asymptotic variance. Therefore, as discussed in Q1, the stats are as follows.

```
##                        MLE     MLE sd  CI lower  CI upper
## Smoker result    0.2203791 0.01957709 0.1820088 0.2587495
## Nonsmoker result 0.3289382 0.01148539 0.3064273 0.3514492
```

c

Since the likelihood function of this problem only differs from the likelihood function from Q1, we know that posterior distribution has form $P(Y|p) \propto [\Pi_{t=1}^{N}(p(1-p)^{t-1})^{Y_t}]((1-p)^N)^{Y_{N+1}} p^a(1-p)^b = p^{\Sigma_{t=1}^{N} Y_t+a}(1-p)^{\Sigma_{t=1}^{N}(t-1)Y_t+NY_{N+1}+b} \sim Beta(\Sigma_{t=1}^{N} Y_t + a, \Sigma_{t=1}^{N}(t-1)Y_t + NY_{N+1} + b)$.
In this case, we have prior $Beta(1,1)$ so we plug in $a = 1, b = 1$.



d

```
## [1] 10000
```

The probability is 100%.

e

The predicted probability is as follows:

```
##                 [,1]      [,2]      [,3]      [,4]       [,5]       [,6]
## Smokers    0.2900000 0.1600000 0.1700000 0.0400000 0.03000000 0.09000000
## Nonsmokers 0.4074074 0.2201646 0.1131687 0.0781893 0.03703704 0.04526749
##                  [,7]       [,8]       [,9]      [,10]      [,11]      [,12]
## Smokers    0.04000000 0.05000000 0.01000000 0.01000000 0.01000000 0.03000000
## Nonsmokers 0.01440329 0.01851852 0.01028807 0.00617284 0.01234568 0.01234568
##                 [,13]
## Smokers    0.07000000
## Nonsmokers 0.02469136
```

Appendix

Q1

```
time = seq(1,13,1)
smoker = c(29, 16,17,4,3,9,4,5,1,1,1,3,7)
nonsmoker = c(198,107,55,38,18,22,7,9,5,3,6,6,12)

part1_1 = smoker[c(1:12)]
part2_1 = smoker[c(1:12)]*time[c(1:12)]
part3_1 = length(smoker) * smoker[13]

mle1 = sum(part1_1)/(sum(part2_1) + part3_1)

fin1 = 1/mle1^2 * sum(part1_1) + 1/(1-mle1)^2 * sum(part2_1) + 1/(1-mle1)^2 * part3_1
var1 = 1/fin1
sd1 = sqrt(var1)
ci1 = c(mle1 - sqrt(var1) * qnorm(0.975), mle1 + sqrt(var1) * qnorm(0.975))

part1_2 = nonsmoker[c(1:12)]
part2_2 = nonsmoker[c(1:12)]*time[c(1:12)]
part3_2 = length(nonsmoker) * nonsmoker[13]

mle2 = sum(part1_2)/(sum(part2_2) + part3_2)

fin2 = 1/mle2^2 * sum(part1_2) + 1/(1-mle2)^2 * sum(part2_2) + 1/(1-mle2)^2 * part3_2
var2 = 1/fin2
sd2 = sqrt(var2)
ci2 = c(mle2 - sqrt(var2) * qnorm(0.975), mle2 + sqrt(var2) * qnorm(0.975))

smoker_res = c(mle1, sd1,ci1)
nonsmoker_res = c(mle2, sd2,ci2)

res_1d = rbind(smoker_res, nonsmoker_res)
rownames(res_1d) = c("Smoker result", "Nonsmoker result")
colnames(res_1d) = c("MLE", "MLE sd", "CI lower", "CI upper")
res_1d

mu = 0.2
sigma = 0.08

b = (mu - sigma^2/(1-mu))/(sigma^2/(1-mu)^2)
a = mu/(1-mu) * b

#sanity check
#a/(a+b)
#sqrt(a*b/((a+b)^2*(a+b+1)))

#post dist
post1_a = sum(part1_1) + a
post1_b = sum(part2_1) + part3_1 + b -sum(part1_1)
post2_a = sum(part1_2) + a
post2_b = sum(part2_2) + part3_2 + b -sum(part1_2)

#post mean
post1_mean = (sum(part1_1) + a)/(sum(part2_1) + part3_1 + a + b)
post2_mean = (sum(part1_2) + a)/(sum(part2_2) + part3_2 + a + b)

#post sd
post1_sd = sqrt(post1_a * post1_b/((post1_a+post1_b)^2*(post1_a+post1_b+1)))
post2_sd = sqrt(post2_a * post2_b/((post2_a+post2_b)^2*(post2_a+post2_b+1)))

#ci
ci1 = c(post1_mean - post1_sd * qnorm(0.975), post1_mean + post1_sd * qnorm(0.975))
ci2 = c(post2_mean - post2_sd * qnorm(0.975), post2_mean + post2_sd * qnorm(0.975))

#hist
hist(rbeta(1000, post1_a, post1_b), xlab ='x', main = "Distribution of Smoker's Posterior")
plot.new()
hist(rbeta(1000, post2_a, post2_b), xlab ='x', main = "Distribution of Non-Smoker's Posterior")

#make table
par(mfrow = c(2, 1))
post_1 = c(post1_a, post1_b, post1_mean, post1_sd, ci1[1], ci1[2])
post_2 = c(post2_a, post2_b, post2_mean, post2_sd, ci2[1], ci2[2])
```

```
res_1 = rbind(post_1, post_2)
colnames(res_1) = c("Posterior a", "Posterior b", "Posterior mean", "Posterior sd", "CI lower", "CI upper")
rownames(res_1) = c("Smoker result", "Nonsmoker result")
res_1
```

Q2

```
getLogL <- function(parameters){
  a = parameters[1]
  b = parameters[2]
  part1 = 0
  for (i in 1:12){
    part1 = part1 + smoker[i] *(lgamma(a+1) + lgamma(b+i -1) - lgamma(a+b+i))
  }
  result = -sum(smoker) * ( lgamma(a) + lgamma(b) - lgamma(a+b) )+
    smoker[13] * (lgamma(a) + lgamma(b+12) - lgamma(a+b+12)) + part1
  return(result)
}

ab1 <- optim(par=c(1,1),fn=getLogL)$par
ab1

getLogL2 <- function(parameters){
  a = parameters[1]
  b = parameters[2]
  part1 = 0
  for (i in 1:12){
    part1 = part1 + nonsmoker[i] *(lgamma(a+1) + lgamma(b+i -1) - lgamma(a+b+i))
  }
  result = -sum(nonsmoker) * ( lgamma(a) + lgamma(b) - lgamma(a+b) )+
    nonsmoker[13] * (lgamma(a) + lgamma(b+12) - lgamma(a+b+12)) + part1
  return(result)
}

ab2 <- optim(par=c(1,1),fn=getLogL2)$par
ab2

res_2d = rbind(ab1, ab2)
colnames(res_2d) = c("alpha MLE", "beta MLE")
rownames(res_2d) = c("Smoker", "Nonsmoker")
res_2d
```

Q3

```
post_a1 = sum(part1_1) + 1
post_b1 = sum(part2_1) + part3_1 + 1 -sum(part1_1)
post_a2 = sum(part1_2) + 1
post_b2 = sum(part2_2) + part3_2 + 1 -sum(part1_2)

post_1_sorted = sort(rbeta(10000, post_a1, post_b1))
post_2_sorted = sort(rbeta(10000, post_a2, post_b2))

hist(post_1_sorted - post_2_sorted, xlab ='x', main = "Distribution of p1 - p2")

set.seed(42)
res_3d1 = rbeta(10000, post_a1, post_b1)
res_3d2 = rbeta(10000, post_a2, post_b2)
sum(res_3d1<res_3d2)
```