STAT7016 Assignment 1

Rui Qiu

2017-08-10

Problem 1

Construct a Monte Carlo study (that is, use computer simulation) that investigates how the probability of coverage depends on the sample size and true proportion value. In the study, let n be 10, 25, and 100 and let p be 0.05, 0.25, and 0.50. Write an R function that has three inputs, n, p, and the number of Monte Carlo simulations m, and will output the estimate of the exact coverage probability. Implement your function using each combination of n and p and $m = 1000$ simulations. Describe how the actual probability of coverage of the traditional interval depends on the sample size and true proportion value.

Solution

First of all, we claim that the following simulation is based on a binomial distribution, and the confidence interval we are setting up here is a 95% confidence interval. Note that the number of simulation runs we use is 1000.

```
# This function returns a 95% CI for binomial distribution
set.seed(1024)

binom.ci <- function(y,n,p=0.95){
    p.hat <- y/n
    z <- qnorm(1-(1-p)/2)
    return(cbind(p.hat-z*sqrt(p.hat*(1-p.hat)/n),
              p.hat+z*sqrt(p.hat*(1-p.hat)/n)))
}

simulation <- function(n,p,m=1000){
  y <- rbinom(m,n,p)
  ci <- binom.ci(y,n)
  coverage.prob <- mean(ci[,1]<p&p<ci[,2])
  return(coverage.prob)
}

n <- c(10,25,100)
p <- c(0.05,0.25,0.5)

for (i in n){
  for (j in p){
    print(c(i, j, simulation(i,j)))
  }
}
```

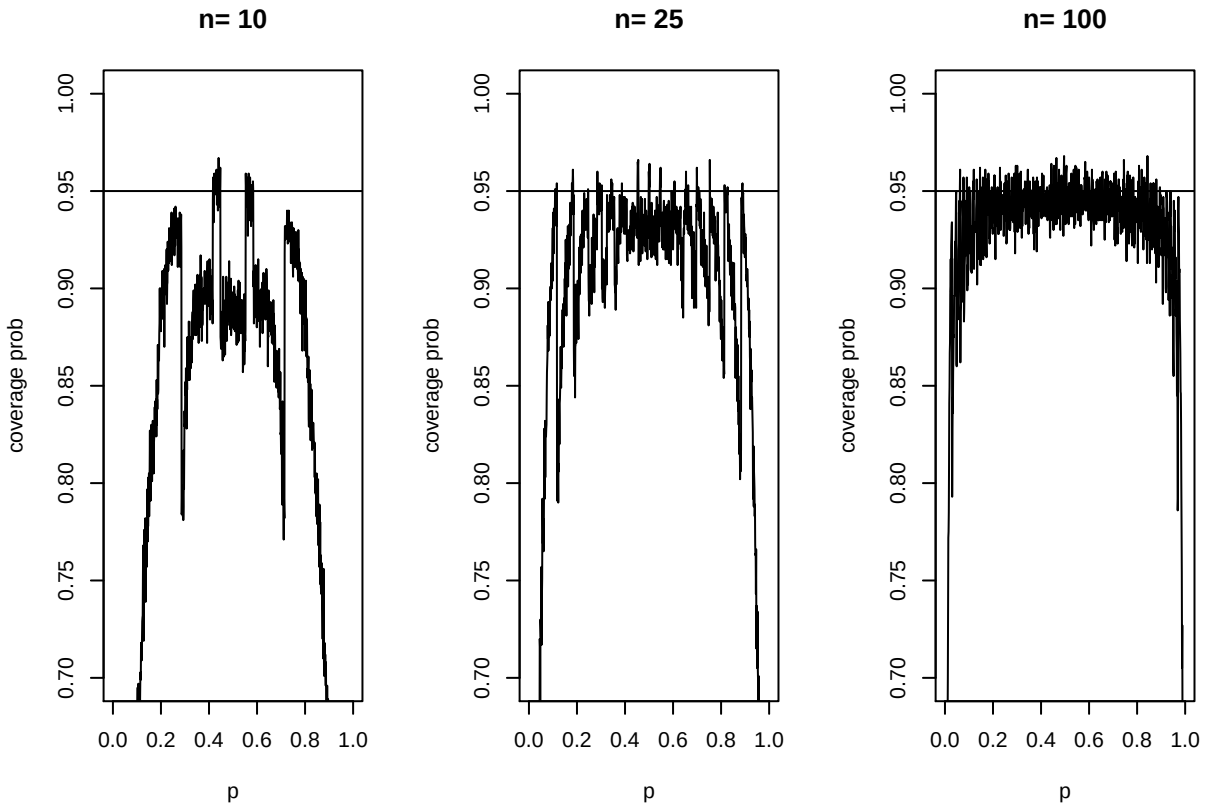
```
## [1] 10.000  0.050  0.405
## [1] 10.000  0.250  0.933
## [1] 10.000  0.500  0.895
## [1] 25.000  0.050  0.712
## [1] 25.000  0.250  0.876
## [1] 25.000  0.500  0.945
```

```
## [1] 100.00   0.05   0.87
## [1] 100.000   0.250   0.942
## [1] 100.000   0.500   0.949
```

So far, we have calculated the required 9 pairs of combinations of sample size n and probability p.

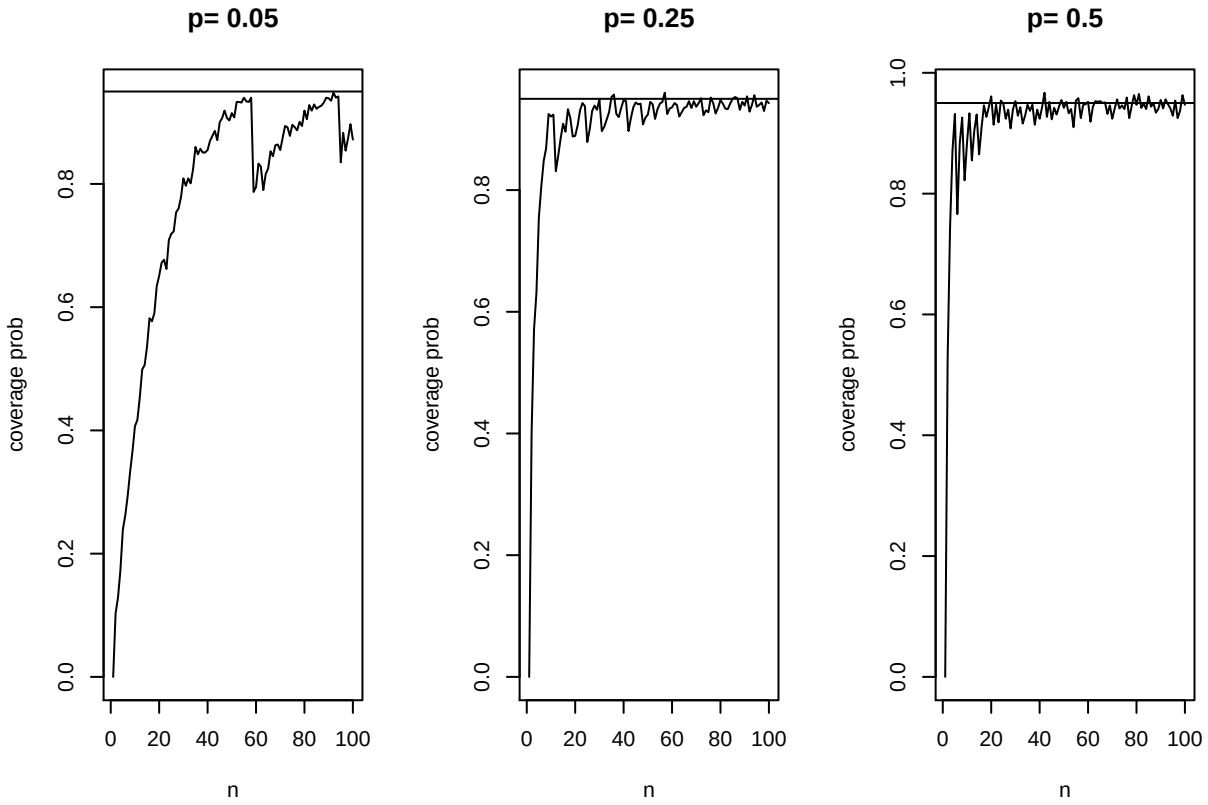
Additionally, we would like to run more simulations to further investigate the pattern within.

```
n.cont <- seq(1,100,1)
p.cont <- seq(0.001,0.999,0.001)

par(mfrow=c(1,3))

# fixed p, different n
for (j in n){
    cvg <- c()
    for (i in 1:length(p.cont)){
        cvg[i] <- simulation(j,p.cont[i])
    }
    plot(p.cont,cvg,type='l',xlab='p',ylab='coverage prob',
         main=paste('n=',j),ylim=c(0.7,1))
    abline(h=0.95)
}
```




```
# fixed n, different p
for (j in p){
  cvg <- c()
  for (i in 1:length(n.cont)){
    cvg[i] <- simulation(n.cont[i],j)
  }
```

```
  plot(n.cont,cvg,type='l',xlab='n',ylab='coverage prob',
       main=paste('p=',j))
  abline(h=0.95)
}
```



According to the returned data, we have 2 general patterns:

- When the true proportion value p is fixed, the larger sample size n is, the actual probability coverage is closer to 95%.
- When the sample size n is fixed, the closer to 0 or 1 true proportion value is, the actual probability coverage is closer to 95%.

Problem 2

A hypothetical study is performed to estimate the effect of a simple training program on basketball free-throw shooting. A random sample of 100 college students is recruited into the study. Each student first shoots 100 free-throws to establish some baseline success probability. Each student then takes 50 practice shots each day for a month. At the end of that time, he or she takes 100 shots for a final measurement. Let θ be the average improvement in success probability. Give three prior distributions for θ, explaining each in a sentence:

(a) A noninformative prior,

(b) A subjective prior based on your best knowledge, and

(c) A weakly informative prior.

Solution

- A noninformative prior: Beta(1,1) distribution would serve as a flat prior since it is not informative to the likelihood function.
- A subjective prior based on my best knowledge: Suppose Beta(20,80) is our prior based on previous study about free-throw shooting percentage conducted in another college, it is then rather informative.
- A weakly informative prior: Beta(2,8) distribution is relatively informative than flat prior but uninformative than Beta(20,80). This might be based on the some individual player studies.

Problem 3

Suppose that there is a Beta(4,4) prior distribution on the probability θ that a coin will yield a "head" when spun in a specified manner. The coin is independently spun 10 times, and "heads" appears fewer than 3 times. You are not told how many heads were seen, only that the number is less than 3 times. Calculate your exact posterior density, posterior mean and posterior variance of θ. Provide a sketch of your posterior density and obtain a 95% posterior interval for θ. Discuss how your results vary if a Beta (20,20) prior is assumed.

Solution

Since the prior follows a Beta(4,4) distribution. We would have:

$$p(\theta) = \frac{1}{B(4,4)}\theta^{4-1}(1-\theta)^{4-1} \propto \theta^3(1-\theta)^3$$

where $B(4,4) = \frac{\Gamma(4)\Gamma(4)}{\Gamma(4+4)}$.

The likelihood function can be expressed explicity as

$$\begin{aligned}\Pr(\text{fewer than 3 heads} \mid \theta) &= \binom{10}{0}\theta^0(1-\theta)^{10} + \binom{10}{1}\theta(1-\theta)^9 + \binom{10}{2}\theta^2(1-\theta)^8 \\ &= (1-\theta)^{10} + 10\theta(1-\theta)^9 + 45\theta^2(1-\theta)^8\end{aligned}$$

Therefore, posterior is

$$p(\theta \mid \text{fewer than 3 heads}) = K \cdot \left(\theta^3(1-\theta)^{13} + 10\theta^4(1-\theta)^{12} + 45\theta^5(1-\theta)^{11}\right) \text{ for some constant } K$$

We want to solve for K

$$\int_0^1 K\left(\theta^3(1-\theta)^{13} + 10\theta^4(1-\theta)^{12} + 45\theta^5(1-\theta)^{11}\right) = 1$$

$$\int_0^1 \theta^3(1-\theta)^{13} + 10\theta^4(1-\theta)^{12} + 45\theta^5(1-\theta)^{11} = \frac{1}{K}$$

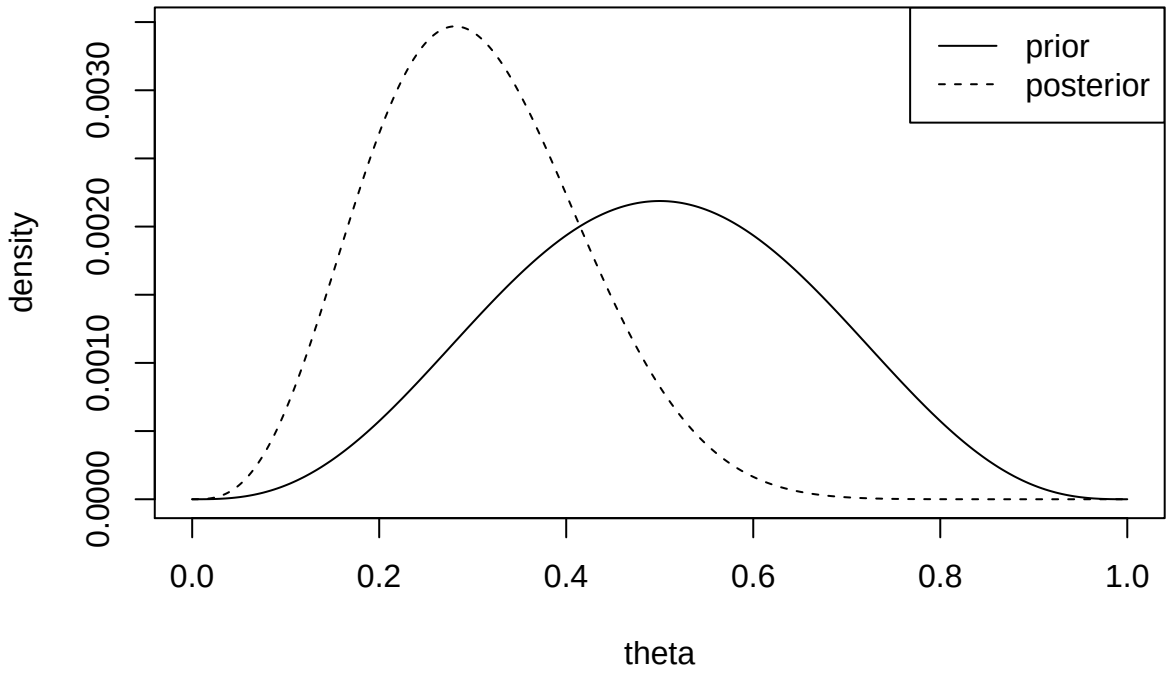
```
integrate(function(x){x^3*(1-x)^13+10*x^4*(1-x)^12+45*x^5*(1-x)^11},0,1)
```

```
## 0.00103426 with absolute error < 1.1e-17
```

```
(K <- 1/0.00103426)
```

```
## [1] 966.8749
```

```
theta <- seq(0,1,0.001)
prior <- dbeta(theta,4,4)/sum(dbeta(theta,4,4))
posterior <- K*prior*(dbinom(0,10,theta)+dbinom(1,10,theta)+dbinom(2,10,theta))
posterior <- posterior/sum(posterior)
plot(theta,posterior,type='l',xlab='theta',ylab='density',cex=1.5,
     lty=2,main="Posterior density plot based on a Beta(4,4) prior")
lines(theta,prior,cex=1.5,lty=1)
legend('topright',c('prior','posterior'),lty=c(1,2))
```



The mean is 0.3046875.

```
sum(posterior*theta)
```

```
## [1] 0.3046875
```

The variance is 0.01247437.

```
sum(posterior*theta^2)-(sum(posterior*theta))^2
```

```
## [1] 0.01247437
```

The posterior interval is about $(0.146, 0.574)$.

```
post.cum <- cumsum(posterior*theta) / sum(posterior*theta)
theta[which(post.cum >= 0.025)[1]]   # lower boundary
```

```
## [1] 0.146
```

```
theta[which(post.cum >= 0.975)[1]-1] # upper boundary
```

```
## [1] 0.574
```

If the prior changes to a Beta(20,20), then

$$p(\theta) = \frac{1}{B(20,20)}\theta^{19}(1-\theta)^{19}$$

The posterior becomes

$$p(\theta \mid \text{fewer than 3 heads}) = K'\left(\theta^{19}(1-\theta)^{29} + 10\theta^{20}(1-\theta)^{28} + 45\theta^{21}(1-\theta)^{27}\right) \text{ for some constant } K'$$

The following procedure would be similar: first we find the constant K' based on the fact that posterior should be proper, then we plot the prior and posterior, calculate mean, variance and posterior interval respectively.

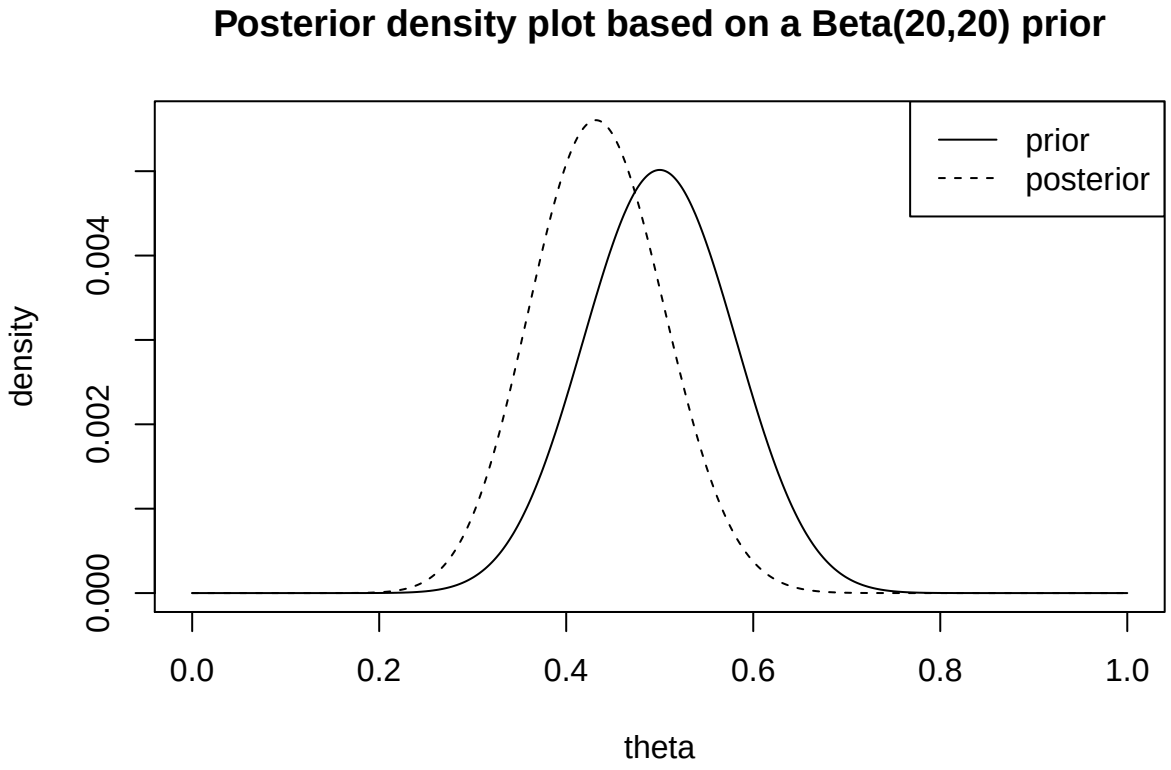
```
integrate(function(x){x^19*(1-x)^29+10*x^20*(1-x)^28+45*x^21*(1-x)^27},0,1)
```

```
## 5.511871e-14 with absolute error < 4.3e-20
```

```
(K2 <- 1/5.511871e-14)
```

```
## [1] 1.814266e+13
```

```
prior2 <- dbeta(theta,20,20)/sum(dbeta(theta,20,20))
posterior2 <- K2*prior2*(dbinom(0,10,theta)+dbinom(1,10,theta)+dbinom(2,10,theta))
posterior2 <- posterior2/sum(posterior2)
plot(theta,posterior2,type='l',xlab='theta',ylab='density',
     cex=1.5,lty=2,main="Posterior density plot based on a Beta(20,20) prior")
lines(theta,prior2,cex=1.5,lty=1)
legend('topright',c('prior','posterior'),lty=c(1,2))
```




```
sum(posterior2*theta)
```

```
## [1] 0.434292
```

```
sum(posterior2*theta^2)-(sum(posterior2*theta))^2 # variance
```

```
## [1] 0.004922443
```

```
post.cum2 <- cumsum(posterior2*theta) / sum(posterior2*theta)
theta[which(post.cum2 >= 0.025)[1]]   # lower boundary
```

```
## [1] 0.312
```

```
theta[which(post.cum2 >= 0.975)[1]-1] # upper boundary
```

```
## [1] 0.582
```

If we compare the calculated posterior statistics above, we notice that

- Beta(4,4) prior has smaller mean (0.3046875), greater variance (0.01247437), and wider posterior interval $(0.146, 0.574)$.
- Beta(20,20) prior has greater mean (0.434292), smaller variance (0.004922443), and narrower posterior interval $(0.312, 0.582)$.

This is because Beta(20,20) is a stronger prior comparing with Beta(4,4), it contains more information, and less variability. Also, the shape of posterior is more 'dominated' by prior Beta(20,20), while the posterior of prior Beta(4,4) looks quite different.

Problem 4

Suppose there are n taxis in Canberra numbered sequentially from 1 to N. You see a taxi at random; it is numbered 159. You wish to estimate N.

(a) Assume your prior distribution on N is geometric with mean 200; that is,

$$p(N) = \frac{1}{200}\left(\frac{199}{200}\right)^{N-1} \quad \text{for } N = 1, 2, \ldots$$

What is your posterior distribution for N?

(b) What are the posterior mean and standard deviation of N? (Sum the infinite series analytically or find an approximation using a computer).

(c) What would be a reasonable "noninformative" prior distribution for N?

Solution

part (a)

Since we already observed a taxi numbered 159, N should be no fewer than this. But we include the case in likelihood function where $N < 159$ anyway:

$$p(\text{randomly see 159} \mid N) = \begin{cases} \frac{1}{N} & \text{if } N \geq 159 \\ 0 & \text{otherwise.} \end{cases}$$

So the posterior is proportional to the product of prior and likelihood:

$$\begin{aligned} p(N \mid \text{randomly see 159}) &\propto \begin{cases} \frac{1}{N} \cdot \frac{1}{200} \cdot \left(\frac{199}{200}\right)^{N-1} & \text{if } N \geq 159 \\ 0 & \text{otherwise.} \end{cases} \\ &\propto \begin{cases} k \cdot \frac{1}{N}\left(\frac{199}{200}\right)^{N} & \text{if } N \geq 159 \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

part (b)

Suppose $p(N \mid \text{randomly see 159}) = k \cdot \frac{1}{N}\left(\frac{199}{200}\right)^{N}$. And we know

$$\begin{aligned} \sum_{N}^{\infty} p(N \mid \text{randomly see 159}) &= 1 \\ \sum_{N}^{\infty} k \frac{1}{N}\left(\frac{199}{200}\right)^{N} &= 1 \\ k \sum_{N}^{\infty} \frac{1}{N}\left(\frac{199}{200}\right)^{N} &= 1 \\ \sum_{N}^{\infty} \frac{1}{N}\left(\frac{199}{200}\right)^{N} &= \frac{1}{k} \end{aligned}$$

All of those are based the precondition that $N \geq 159$, so we can approximate the value of $\frac{1}{k}$ by calculating value of $\sum_{N=159}^{1000} \frac{1}{N} \left(\frac{199}{200}\right)^N$.

```
k.reciprocal <- 0
for (n in 159:1000){
  k.reciprocal <- k.reciprocal + (1/n)*(199/200)^n
}
k.reciprocal
```

```
## [1] 0.3125823
```

```
1/k.reciprocal
```

```
## [1] 3.199158
```

Our approximation shows that $k \approx 3.2$.

$$
\begin{aligned}
E(N \mid \text{randomly see 159}) &= \sum_{N=159}^{\infty} N \cdot p(N \mid \text{randomly see 159}) \\
&= \sum_{N=159}^{\infty} N \cdot k \cdot \frac{1}{N} \left(\frac{199}{200}\right)^N \\
&= k \cdot \sum_{N=159}^{\infty} \left(\frac{199}{200}\right)^N \\
&\approx 3.2 \cdot \frac{\left(\frac{199}{200}\right)^{159}}{1 - \frac{199}{200}} \\
&= 288.4362
\end{aligned}
$$

Similarly, we can compute the standard deviation (computer apporximation also required).

$$
\begin{aligned}
\text{SD}(N \mid \text{randomly see 159}) &= \sqrt{\sum_{N=159}^{\infty} (N - 288.4362)^2 \cdot k \cdot \frac{1}{N} \left(\frac{199}{200}\right)^N} \\
&\approx \sqrt{\sum_{N=159}^{1000} (N - 288.4362)^2 \cdot 3.2 \cdot \frac{1}{N} \left(\frac{199}{200}\right)^N}
\end{aligned}
$$

```
variance <- 0
for (n in 159:1000){
  variance <- variance + (n-288.4362)^2*3.2*(1/n)*(199/200)^n
}
(sd <- sqrt(variance))
```

```
## [1] 132.0843
```

The posterior mean is 288.4362, the posterior standard deviation is 132.0843.

part (c)

Assume the noninformative prior is

$$p(N) = \begin{cases} \frac{1}{m} & N = 1, 2, \ldots, m, \text{ for some large } m \geq 159 \\ 0 & \text{otherwise.} \end{cases}$$

then the new posterior is proportional to $\frac{1}{N}$ only.

So the new posterior mean can be expressed as

$$\frac{m - 159}{\sum_{N=159}^{m} \frac{1}{N}}$$

And the new posterior standard deviation can be expressed as

$$\sqrt{\frac{\frac{m(m+1)}{2} - \frac{159 \times (159+1)}{2}}{\sum_{N=159}^{m} \frac{1}{N}} - \left(\frac{m - 159}{\sum_{N=159}^{m} \frac{1}{N}} \right)^2}$$

Problem 5

Suppose you own a trucking company with a large fleet of trucks. Breakdowns occur randomly in time and the number of breakdowns during an interval of t days is assumed to be Poisson distributed with mean $t\lambda$. The parameter λ is the daily breakdown rate. The possible values for λ are 0.5, 1, 1.5, 2, 2.5, and 3 with respective probabilities 0.1, 0.2, 0.3, 0.2, 0.15, and 0.05. If one observes y breakdowns, then the posterior probability of λ is proportional to

$$g(\lambda)\exp(-t\lambda)(t\lambda)^y$$

where $g(\lambda)$ is the prior probability.

(a) If 12 trucks break down in a six-day period, find the posterior probabilities for the different values of λ.

(b) Find the probability that there are no breakdowns during the next week.

Solution

part (a)

$$p(\lambda \mid y) = \frac{p(y \mid \lambda)p(\lambda)}{\int_\lambda p(y \mid \lambda)p(\lambda)} \propto g(\lambda)e^{-6\lambda}(6\lambda)^{12}$$

we implement the calculation in the following R codes.

```
lambda <- c(0.5,1,1.5,2,2.5,3)
lambda.prob <- c(0.1,0.2,0.3,0.2,0.15,0.05)

truck.post <- c()
for (i in 1:6){
  truck.post[i] <- exp(-6*lambda[i])*(6*lambda[i])^12*lambda.prob[i]
}
truck.post <- truck.post/sum(truck.post)
truck.post
```

```
## [1] 9.021401e-05 3.679430e-02 3.565196e-01 3.735713e-01 2.029885e-01
## [6] 3.003598e-02
```

So the posterior probabilities for each case of λ are:

$$\begin{cases} p(\lambda = 0.5 \mid y) = 9.021401 \times 10^{-5} \\ p(\lambda = 1 \mid y) = 3.679430 \times 10^{-2} \\ p(\lambda = 1.5 \mid y) = 3.565196 \times 10^{-1} \\ p(\lambda = 2 \mid y) = 3.735713 \times 10^{-1} \\ p(\lambda = 2.5 \mid y) = 2.029885 \times 10^{-1} \\ p(\lambda = 3 \mid y) = 3.003598 \times 10^{-2} \end{cases}$$

part (b)

The predictive probability is the sum of conditional probabilities and posterior probabilities which are the results from part (a).

$$
\begin{aligned}
p(\tilde{y} = 0 \mid y_1, \ldots, y_6) &= \int_0^1 p(\tilde{y}, \lambda \mid y_1, \ldots, y_6) d\lambda \\
&= \int_0^1 p(\tilde{y} \mid \lambda) p(\lambda \mid y_1, \ldots, y_6) d\lambda
\end{aligned}
$$

```
pred.post <- 0
for (i in 1:6){
  pred.post <- pred.post + exp(-7*lambda[i])*truck.post[i]
}
pred.post
```

```
## [1] 4.640932e-05
```

So the probability that there are no breakdowns during the next weeks is about 4.64×10^{-5}.

Problem 6

Consider two coins C_1 and C_2, with the following characteristics: $\Pr(\text{heads} \mid C_1) = 0.25$ and $\Pr(\text{heads} \mid C_2) = 0.75$. Choose one of the coins at random and imagine spinning it repeatedly. Given that the first two spins from the chosen coin are tails, what is the expectation of the number of additional spins until a head shows up?

Solution

$$\Pr(C_1 \mid \text{two tails}) = \frac{\Pr(C_1)\Pr(\text{two tails} \mid C_1)}{\Pr(C_1)\Pr(\text{two tails} \mid C_1) + \Pr(C_2)\Pr(\text{two tails} \mid C_2)}$$

Note that $\Pr(C_1) = \Pr(C_2) = 0.5$ since we choose them at random.

$$\begin{aligned}\Pr(\text{two tails} \mid C_1)) &= (1 - \Pr(\text{heads} \mid C_1))^2 = 0.75^2 = 0.5625\\ \Pr(\text{two tails} \mid C_2)) &= (1 - \Pr(\text{heads} \mid C_2))^2 = 0.25^2 = 0.0625\end{aligned}$$

Therefore

$$\begin{aligned}\Pr(C_1 \mid \text{two tails}) &= \frac{0.5 \cdot 0.5625}{0.5 \cdot 0.5625 + 0.5 \cdot 0.0625} = 0.9\\ \Pr(C_2 \mid \text{two tails}) &= 0.1\end{aligned}$$

Hence, the expectation of number of additional spins until a head shows up can be expressed as:

$$\begin{aligned}E(\text{number of spins} \mid \text{two tails}) &= \Pr(C_1 \mid \text{two tails}) \cdot E(\text{number of spins} \mid C_1, \text{two tails})\\ &\quad + \Pr(C_2 \mid \text{two tails}) \cdot E(\text{number of spins} \mid C_2, \text{two tails})\\ &= 0.9 \cdot \frac{1}{0.25} + 0.1 \cdot \frac{1}{0.75}\\ &\approx 3.733\end{aligned}$$

Note that the expected value of a geometric distribution is $\frac{1}{p}$.

So the expected additional number of spins until a head shows up is 3.733.