STA623 - Bayesian Data Analysis - Practical 4 (Solutions)

Marc Henrion

8 September 2022

Practical 4

Notation

- X, Y, Z - random variables
- x, y, z - measured / observed values
- $\bar{X}$, $\bar{Y}, \bar{Z}$ - sample mean estimators for X, Y, Z
- $\bar{x}$, $\bar{y}, \bar{z}$ - sample mean estimates of X, Y, Z
- $\hat{T}$, $\hat{t}$ - given a statistic T, estimator and estimate of T
- $P(A)$ - probability of an event A occuring
- $f_X(.)$, $f_Y(.), f_Z(.)$ - probability mass / density functions of X, Y, Z; sometimes $p_X(.)$ etc. rather than $f_X(.)$
- p(.) - used as a shorthand notation for pmfs / pdfs if the use of this is unambiguous (i.e. it is clear which is the random variable)
- $X \sim F$ - X distributed according to distribution function F
- $E[X]$, $E[Y]$, $E[Z]$, $E[T]$ - the expectation of X, Y, Z, T respectively

Exercise 1

Let's revisit Exercise 5 from Practical 1&2.

We had 2 groups of women and we compared the number of children born to each women in the 2 groups. For each group we assumed a Poisson sampling model: $Y_{i,j} \sim \text{Pois}(\theta_i), i = 1, \ldots, n_j, j = 1, 2$ and we found that the posterior distributions were:

1. Women without college degree: $\theta_1 \sim \Gamma(219, 112)$
2. Women with college degree: $\theta_2 \sim \Gamma(68, 45)$

We had computed $P(\theta_1 > \theta_2 | n_1, n_2, \sum_i y_{i,1}, \sum_i y_{i,2}) = 0.97$.

Use the Monte Carlo method to compute

$$P(\tilde{Y}_1 > \tilde{Y}_2 | n_1, n_2, \sum_i y_{i,1}, \sum_i y_{i,2})$$

For the group of women without college degree, remember that we found that the posterior predictive distribution was a negative binomial:

$$\tilde{Y}_1|n_1, \sum_i y_{i,1} \sim \text{NegBin}(219, 112/113)$$

Compare this distribution with the empirical distribution of the raw data:

no. children per mother	number of mothers
0	20
1	19
2	38
3	20
4	10
5	2
6	2

Let $\mathbf{y} = (y_{1,1}, \ldots, y_{n_1,1})$. Define $t(\mathbf{y})$ as the ratio of 2's in $\mathbf{y}$ to the number of $1's$. In this dataset we observe $t(\mathbf{y}) = 38/19 = 2$. Use the posterior predictive distribution for $\tilde{Y}_1|n_1, \sum_i y_{i,1}$ and the Monte Carlo method to compute $P(t(\tilde{Y}) \geq 2|n_1, \sum_i y_{i,1})$. What is your conclusion?

Exercise 1 (Solution)

In this exercise we had:

$$a = 2, b = 1, n1 = 111, \sum_i y_{i,1} = 217, n_2 = 44, \sum_i y_{i,2} = 66$$

We found that the posterior distributions for θ_1 and θ_2 were $\Gamma(a + \sum_i y_{i,j}, b + n_j), j = 1, 2$ respectively.

To compute $P(\tilde{Y}_1 > \tilde{Y}_2|...)$ we need to, for $j = 1, 2$

$$\begin{cases} \text{sample } \theta_j^{(1)} \sim \Gamma(a + \sum_i y_{i,j}, b + n_j), \text{ sample } \tilde{y}_j^{(1)} \sim \text{Pois}\left(\theta_j^{(1)}\right) \\ \text{sample } \theta_j^{(2)} \sim \Gamma(a + \sum_i y_{i,j}, b + n_j), \text{ sample } \tilde{y}_j^{(2)} \sim \text{Pois}\left(\theta_j^{(2)}\right) \\ \ldots \\ \text{sample } \theta_j^{(S)} \sim \Gamma(a + \sum_i y_{i,j}, b + n_j), \text{ sample } \tilde{y}_j^{(S)} \sim \text{Pois}\left(\theta_j^{(S)}\right) \end{cases}$$

The `R` code for this is quite short:

```
a<-2; b<-1
n1<-111; sy1<-217
n2<-44; sy2<-66
S<-1e6

s.theta1<-rgamma(n=S,a+sy1,b+n1)
s.theta2<-rgamma(n=S,a+sy2,b+n2)

s.y1<-rpois(n=S,lambda=s.theta1)
s.y2<-rpois(n=S,lambda=s.theta2)
```

Finally we need to compute $\frac{1}{S}\sum_s I(\tilde{y}_1^{(s)} > \tilde{y}_2^{(s)})$, where $I(.)$ is the indicator function, to approximate $P(\tilde{Y}_1 > \tilde{Y}_2|...)$. (This means that $I(TRUE) = 1$ and $I(FALSE) = 0$.)

```
sum(s.y1>s.y2)/S
## [1] 0.482014
```

There is an even quicker way. Remember we saw that

$$\tilde{Y}_{n_1+1,1}|n_1 = 111, \sum_i y_{i,1} = 217 \sim \text{NegBin}(219, \frac{112}{113})$$

$$\tilde{Y}_{n_2+1,2}|n_2 = 44, \sum_i y_{i,2} = 66 \sim \text{NegBin}(68, \frac{45}{46})$$

So we could directly sample from these:

```
s.y1<-rnbinom(n=S,size=219,prob=112/113)
s.y2<-rnbinom(n=S,size=68,prob=45/46)
sum(s.y1>s.y2)/S
```

```
## [1] 0.482357
```

While we have a lot of evidence that the parameters satisfy $\theta_1 > \theta_2$, the overlap between the distributions is significant, and only about 48% of the time, a random woman in the first group has more children than a random woman in the second group.

We now focus only on the first group, the women without a college degree.

Let's compare the empirical (i.e. observed) and the posterior predictive distributions.

First we need to code up the data:

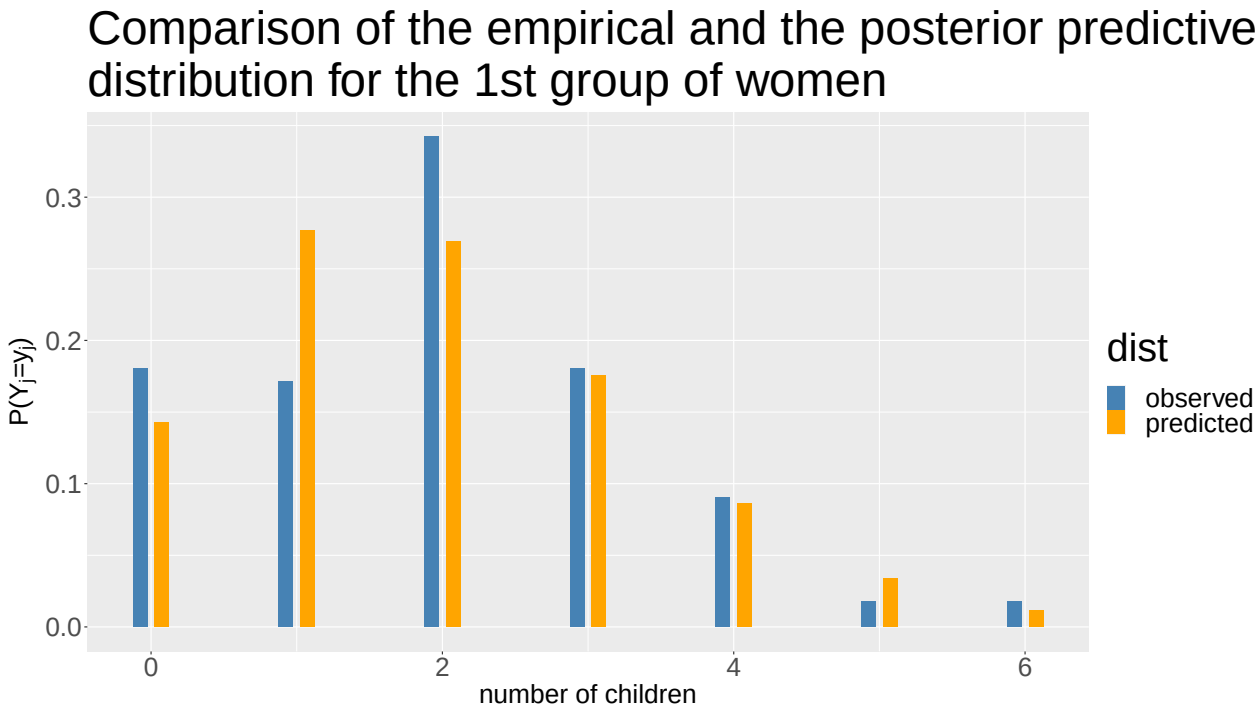
```
freqMatObs<-data.frame(numChild=0:6,propMother=c(20,19,38,20,10,2,2)/111,dist=rep("observed",7))
freqMatPred<-data.frame(numChild=0:6,propMother=dnbinom(0:6,size=219,prob=112/113),dist=rep("predicted"

# Alternatively you can compute estimates for the posterior predictive probabilities from the sampled y
# for(i in 1:nrow(freqMatPred)){
#   freqMatPred$propMother[i]<-sum(s.y1==freqMatPred$numChild[i])/S
# }
freqMat<-rbind(freqMatObs,freqMatPred)
```

Then we can compare these distributions:

```
library(ggplot2)

ggplot(data=freqMat,mapping=aes(x=numChild,y=propMother,fill=dist),
       col=c("steelblue","greenyellow"),cex.lab=2.3,cex.axis=2.3) +
  geom_bar(position=position_dodge(width=0.3),stat="identity",width=0.2) +
  scale_fill_manual(values=c("steelblue","orange")) +
  xlab("number of children") +
  ylab(expression(paste(sep="","P(",Y[j],"=",y[j],")"))) +
  theme(axis.text=element_text(size=24),
        axis.title=element_text(size=24),
        title=element_text(size=36),
        legend.text=element_text(size=24)) +
  ggtitle("Comparison of the empirical and the posterior predictive\ndistribution for the 1st group of
```



We see that for $Y_1 = 2$, the empirical and the posterior predictive distributions are quite different: while in the observed data, twice as many women had 2 children than had 1 child, according to the posterior predictive distribution, we would expect slighly more women to have 1 child than women who have 2.

Let $\mathbf{Y} = (Y_{1,1}, Y_{1,2}, ..., Y_{1,n_1})$. Let $t(\mathbf{Y})$ be the ratio of the number of women with 2 children to the number of women with 1 child in $\mathbf{Y}$.

We see that $t(\mathbf{y}_{obs} = 38/19 = 2)$.

Let's compute $P(t(\tilde{\mathbf{Y}}) > 2|y_{1,1}, \ldots, y_{n_1,1})$. For this we need to, for $s = 1, \ldots, S$:

1. sample $\theta_{1,i}^{(s)} \sim_{iid} p(\theta_1|y_{1,1}, \ldots, y_{n_1,1}) = \Gamma(219, 112)$ for $i = 1, \ldots, n_1$
2. sample $\tilde{\mathbf{Y}}_1^{(s)} = (\tilde{Y}_{1,1}^{(s)}, \ldots, \tilde{Y}_{1,n_1}^{(s)})$ by sampling independently $\tilde{Y}_{1,i} \sim_{\text{iid}} \text{Pois}\left(\theta_{1,i}^{(s)}\right)$
3. compute $t^{(s)} = t\left(\tilde{\mathbf{Y}}_1^{(s)}\right)$

Note that steps 1. and 2. above could be replaced by a single step, directly sampling from the posterior predictive distribution: $\tilde{\mathbf{Y}}_1^{(s)} = (\tilde{y}_{1,1}^{(s)}, \ldots, \tilde{y}_{n_1,1}^{(s)}) \sim_{iid} \text{NegBin}(219, 112/113)$.

In R this becomes:

```
tVect<-rep(NA,S)
for(s in 1:S){
  theta1<-rgamma(n=n1,a+sy1,b+n1)
  s.y1<-rpois(n1,lambda=theta1)
  #s.y1<-rnbinom(n1,size=219,prob=112/113) # alternatively sample directly from NegBin posterior predic
  tVect[s]<-sum(s.y1==2)/sum(s.y1==1)
}

sum(tVect>=2)/S
```

```
## [1] 0.004042
```

We conclude it is very unlikely to observe this ratio of 2's to 1's under the assumed Poisson sampling model. Our sampling model is likely wrong and we should try to identify a more suitable model.

[end of STA623 BDA Practical 4]