HW7

Pan Yijia/yp2424

November 13, 2016

```r
# 1
poisLoglik<-function(lam,data){
  poislog<-sum(dpois(data,lam,log=TRUE))
  return(poislog)
}
x1<-poisLoglik(lam=1,data=c(1,0,0,1,1))
x1
```

```
## [1] -5
```

```r
# The value is -5 when data=c(1,0,0,1,1)
```

```r
# 2
moretti <- read.csv("~/Downloads/moretti.csv")
count_new_genres<-function(x){
new<-moretti[moretti$Begin==x,1]
length(new)
}
count_new_genres(1803)
```

```
## [1] 0
```

```r
count_new_genres(1850)
```

```
## [1] 3
```

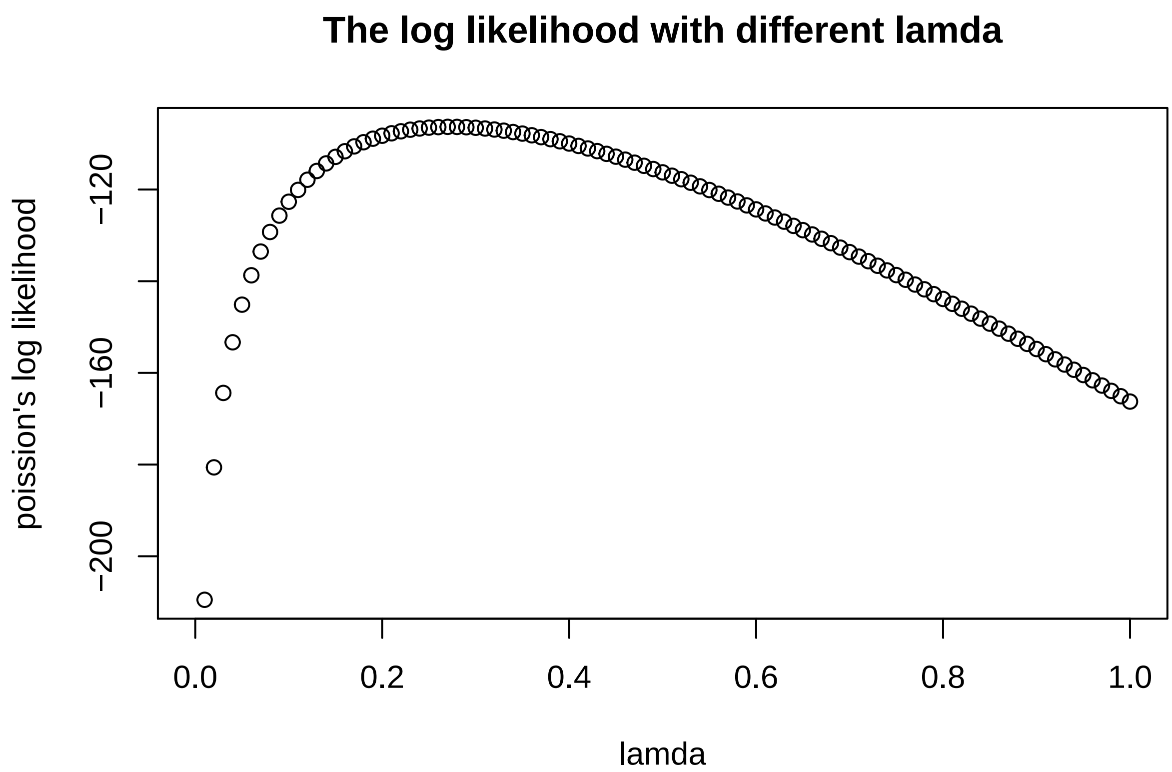
```r
# 3
n<-length(1740:1900)
new_genres<-rep(NA,n)
for(i in 1:n){
  new_genres[i]<-count_new_genres(i+1739)
}
# The 1803-1740+1=64 and 1850-1740+1=111 correspond
# to the years 1803 and 1850. Their value should be 0 and 3.
new_genres[c(64,111)]
```

```
## [1] 0 3
```

```r
# It is what my vector new_genres has for those years.
```

```r
# 4
lamda<-seq(0,1,by=0.01)
n<-length(lamda)
```

```r
pois<-rep(NA,n)
for (i in 1:n){
  pois[i]<-poisLoglik(lamda[i],new_genres)
}
plot(lamda,pois,xlab="lamda",ylab="poission's log likelihood",main="The log likelihood with different la
```




```r
# 5
neg.poisLoglik<-function(lam,data){
  neg.poislog<--sum(dpois(data,lam,log=TRUE))
  return(neg.poislog)
}
nlm(neg.poisLoglik,0.1,new_genres)[1:3]
```

```
## $minimum
## [1] 106.3349
##
## $estimate
## [1] 0.2732914
##
## $gradient
## [1] 3.879563e-06
```

```r
# The maximum is at lamda=0.273
```

```r
# 6
intergenre_intervals<-diff(moretti$Begin)
mean<-mean(intergenre_intervals)
mean
```

```
## [1] 3.44186
```

```
sd<-sd(intergenre_intervals)
sd
```

```
## [1] 3.705224
```

```
coefficient_of_variation<-sd/mean
coefficient_of_variation
```

```
## [1] 1.076518
```

```
# 7 (a)
f1<-function(vector){
 k<-max(vector)
 a<-NULL
 for (i in 1:k){
   a<-sort(c(a,rep(which(vector==i),i)))
 }
 return(diff(a))
}
f1(new_genres)
```

```
##  [1]  8 11  7  2  2  3 16  1  1  9  4  4  6  8  3  1  2  2  0  2  6  1  7
## [24]  0  1  1  1  1  0  0  1  6 11  3  1  0  1  3  8  1  0  3  0
```

```
intergenre_intervals
```

```
##  [1]  8 11  7  2  2  3 16  1  1  9  4  4  6  8  3  1  2  2  0  2  6  1  7
## [24]  0  1  1  1  1  0  0  1  6 11  3  1  0  1  3  8  1  0  3  0
```

```
# 7 (b)
b<-function(x,mean){
  y<-rpois(x,mean)
  inter<-f1(y)
  cov<-sd(inter)/mean(inter)
  return(list(inter,cov))
}
b(161,0.273)
```

```
## [[1]]
##  [1] 16  3  5  9  0  1  4  4  1  2  1  2  2  2  6  0  2  1  4  0  1  0  7
## [24]  5  3 15  1  3  0  6  3  1  1  0  3 14  2  1  5  2  7  7  1  0  3
##
## [[2]]
## [1] 1.109734
```

```
# 8
simu<-rep(NA,100000)
for(i in 1:100000){
  simu[i]<-b(161,0.273)[2]
}
sum(simu>coefficient_of_variation)/100000
```

```
## [1] 0.22638
```

```r
# 9
# Probably not. As we know that, for a possion distribution,
# the coefficient of variation is expected to be around 1.
# If the number of genres which appear in a given year follow
# a Possion distribution,it's coefficient of variation will
# not larger than most coefficient of variation of Possion
# distribution we generated.
```

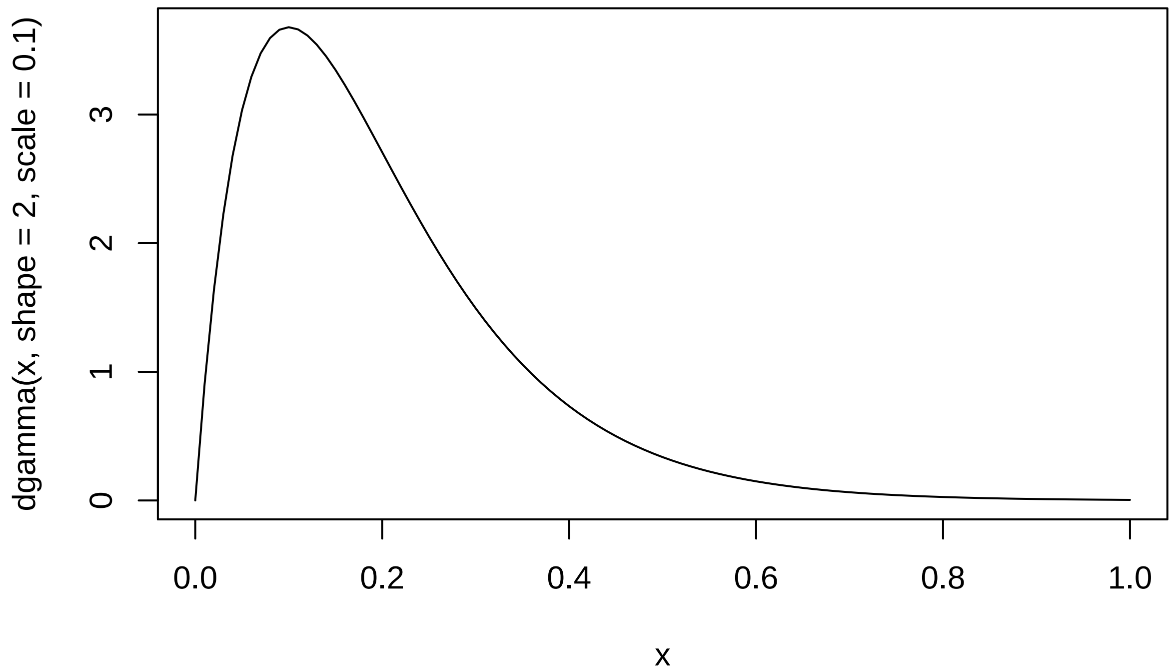
```r
# Part 2
# 10
d<-aggregate(moretti$Name,list(as.factor(moretti$Begin)),length)
names(d)<-c("year","Number")
all.dates <- seq(1740,1900, by=1)
all.dates.frame <- data.frame(list(year=all.dates))
merged.data <- merge(all.dates.frame, d, all.x=T)
merged.data$Number[which(is.na(merged.data$Number))] <- 0
head(merged.data)
```

```
##   year Number
## 1 1740      1
## 2 1741      0
## 3 1742      0
## 4 1743      0
## 5 1744      0
## 6 1745      0
```

```r
df <- split(merged.data, (as.numeric(rownames(merged.data))-1) %/% 10)[-17]
lambda<-NULL
for( i in 1:16){
  lambda[i]<-mean(as.data.frame(df[i])[,2])
}
lambda
```

```
##  [1] 0.2 0.1 0.2 0.2 0.1 0.2 0.3 0.1 0.4 0.5 0.5 0.5 0.1 0.5 0.5 0.0
```

```r
# 11 (a)
curve(dgamma(x,shape=2,scale=0.1))
```




```
# No, it is not.
```

```
# 12 (b)
initial<-function(){
  rgamma(1,shape=2,scale=0.1)
}
initial()
```

```
## [1] 0.09836649
```

```
# 12 (c)
proposal<-function(a){
  y<-rnorm(1,sd=0.001)
  q<-ifelse(a+y>0,a+y,a)
  return(q)
}
```

```
# 12 (d)
posterior<-function(a,data){
  x<-dgamma(a,shape=2,scale=0.1)
  y<-exp(poisLoglik(a,data))
  z<-x*y
  return(z)
}
posterior(0.2,new_genres)
```

```
## [1] 2.571445e-47
```

```r
# 12 (e)
new_genres12<-as.data.frame(df[12])$X11.Number
metrostep<-function(x){
  z<-proposal(x)
  u<-runif(1)
  ratio<-posterior(z,new_genres12)/posterior(x,new_genres12)
  if(u<ratio){
    accepted.val<-z
  }
  else{
    accepted.val<-x
  }
  return(accepted.val)
}
n<-100000
vals<-vector(length=n)
vals[1]<-initial()
for(t in 2:n){
  vals[t]<-metrostep(vals[t-1])
}
mean(vals)
```

```
## [1] 0.2302589
```

```r
sd(vals)
```

```
## [1] 0.06226543
```

```r
mean(vals[-c(1:10000)])
```

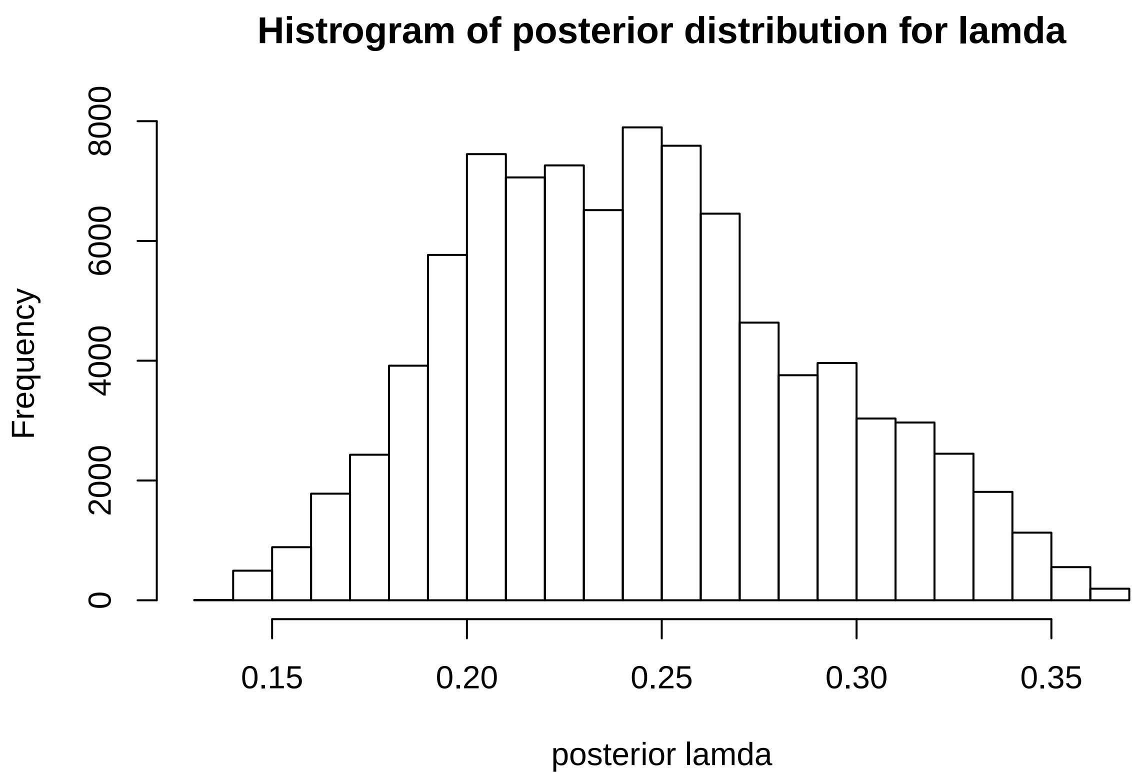
```
## [1] 0.2444446
```

```r
sd(vals[-c(1:10000)])
```

```
## [1] 0.045354
```

```r
# The mean of lamda will be greater and the standard
# deviation will be smaller.
```

```r
# 12 (f)
hist(vals[-c(1:10000)],main="Histrogram of posterior distribution for lamda",xlab="posterior lamda")
```




```
# 1 (a)
proposal<-function(a){
  y<-rnorm(16,sd=0.001)
  q<-vector(length = 16)
for(i in 1:16){
  if((a+y)[i]>0){
    q[i]<-(a+y)[i]
  }
  else
    q[i]<-a[i]
}
  return(q)
}
```

```
# 1 (b)
 posterior<-function(lambda,data){
  a<-1
  df <- as.data.frame(split(data, ceiling(seq_along(data)/10)))

  for(i in 1:length(lambda)){
  a<-a*exp(poisLoglik(lambda[i],df[,i]))*dgamma(lambda[i],2,scale=0.1)

  }
  return(a)
}
posterior(rep(0.2,16),new_genres[1:160])
```

```
## [1] 9.630563e-41
```

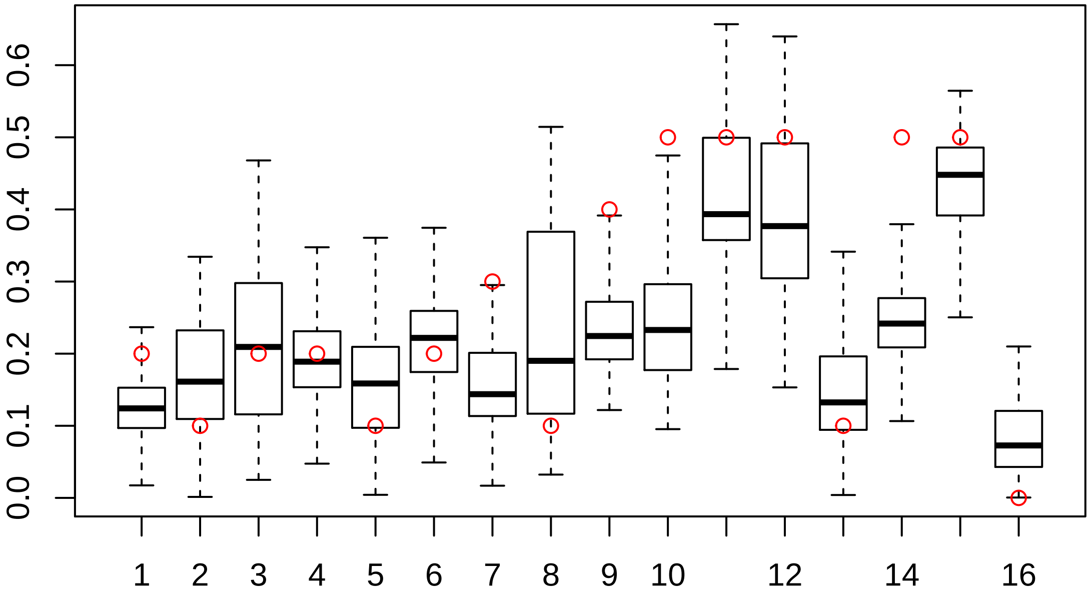
```
# 1 (c)
new_genres <- new_genres[1:160]
  initial <- function() {
    return(rgamma(16, shape = 2, scale = 0.1))
}
  metrostep <- function(x) {
    # x is now a vector of 16 intensity values
    z <- proposal(x)
    u <- runif(1)
    ratio <- posterior(z, new_genres)/posterior(x, new_genres)
    if(u < ratio) {
      accepted.val <- z
    } else {
      accepted.val <- x
}
    return(accepted.val)
  }
  n       <- 100000
  vals    <- matrix(NA, nrow = n, ncol = 16)
  vals[1,] <- initial()
  for (t in 2:n) {
      vals[t, ] <- metrostep(vals[t-1, ])
  }
vals1<-vals[-c(1:10000),]
mean1<-apply(vals,2,mean)
sd1<-apply(vals,2,sd)
mean1
```

```
##  [1] 0.13031353 0.17174815 0.22802270 0.20602774 0.15229413 0.20728379
##  [7] 0.15349283 0.22581594 0.23104831 0.25117334 0.40125447 0.37479741
## [13] 0.14958864 0.23396731 0.43617513 0.08208857
```

```
sd1
```

```
##  [1] 0.04612260 0.07935095 0.11567402 0.07122871 0.07239970 0.06518234
##  [7] 0.05434510 0.13623470 0.06173084 0.09166689 0.11000090 0.11246664
## [13] 0.06831064 0.06272813 0.06355264 0.04683123
```

```
boxplot(vals1,outline=FALSE)
points(lambda,col="red")
```


0.0
0.1
0.2
0.3
0.4
0.5
0.6
1
2
3
4
5
6
7
8
9
10
12
14
16