Assignment 7

Cheng Zhang cz2532

December 30, 2018

Goals: More practice with simulations. Summarizing data using distributions and estimating parameters.

Firstly, import data

```
setwd("C:/Users/Alan_/Desktop/Semester1 Courses/STAT computing/Week11/assignm
ent 7")
moretti <- read.csv("moretti.csv", as.is = TRUE)
```

1. Assume the variables x1; x2; : : : ; xn are independent and Poisson-distributed Write a function poisLoglik, which takes as inputs a single number lambda and a vector data and returns the log-likelihood of that parameter value on that data. What should the value be when data = c(1, 0, 0, 1, 1) and lambda = 1?

```
poisLoglik <- function(lambda, data){
  sum <- 0
  for (i in data){
    log_value <- log(((lambda^i)*exp(-lambda))/factorial(i))
    sum = sum+log_value
  }
  return(sum)
}

data <- c(1,0,0,1,1)
poisLoglik(lambda=1,data=data)

## [1] -5
```

2. Write a function count new genres which takes in a year, and returns the number of new genres which appeared in that year: 0 if there were no new genres that year, 1 if there was one, 3 if there were three, etc. What should the values be for 1803 and 1850?

```
count_new_genres <- function(year){
  count_new <- sum(moretti$Begin==year)
  return(count_new)
}

count_new_genres(1803)

## [1] 0

count_new_genres(1850)
```

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

3. Create a vector, new genres, which counts the number of new genres which appeared in each year of the data, from 1740 to 1900. What positions in the vector correspond to the years 1803 and 1850? What should those values be? Is that what your vector new genres has for those years?

```
year <- c(1740:1900)
new_genres <- c()
for (i in year) {
  new_genres <- c(new_genres,sum(i==moretti$Begin))
}
```

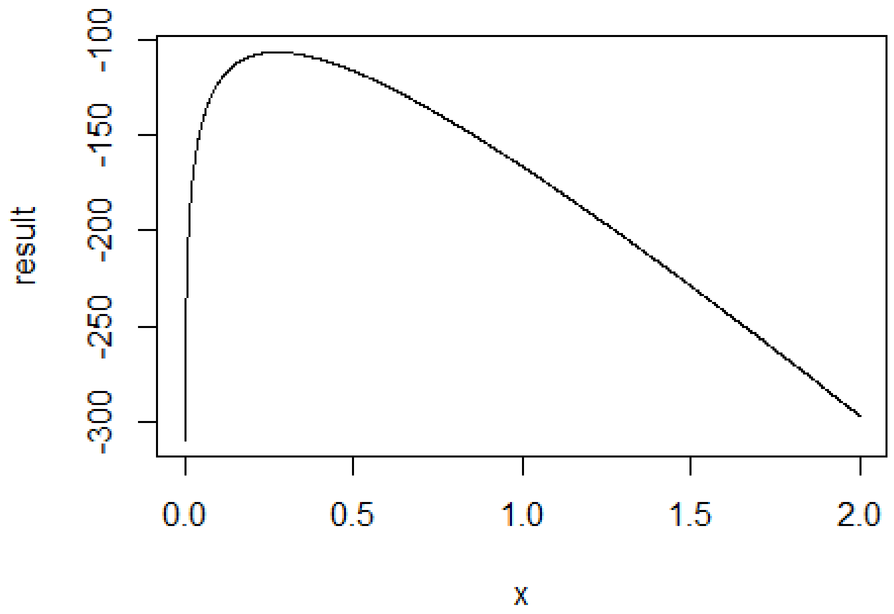
The 1803 is the 64th position in this vector and 1850 is the 110th position in this vector. According to previous question, they should be 0 and 3.

```
new_genres[64]

## [1] 0

new_genres[111]

## [1] 3
```

that is exactly what we get in those position in the vector.

4. Plot poisLoglik as a function of lambda on the new_genres data. (If the maximum is not at lambda = 0.273, you're doing something wrong.)

```
x <- seq(0,2,0.001)
result <- c()
for (i in x){
  result <- c(result,poisLoglik(i, data=new_genres))
}
plot(x,result,type = "l")
```



```
x[which.max(result)]

## [1] 0.273
```

The max value of outputs of the function is reached when lambda equals 0.273. It is the same as the given number

5. Use nlm() to maximize the log likelihood to check the lambda = 0:273 value suggested in the previous question. Hint: you may need to rewrite your function from (i.) with some slight alterations.

```
neg_poisLoglik <- function(lambda, data){
  sum <- 0
  for (i in data){
    log_value <- log(((lambda^i)*exp(-lambda))/factorial(i))
    sum = sum+log_value
  }
  return(-sum)
}

nlm(neg_poisLoglik, 0.2, data = new_genres)$estimate

## [1] 0.2732919
```

the estimated lambda which can maximize the likelihood function is 0.2732919 which is the same as the given number in previous question. So, lambda = 0.273

6. To investigate whether genres appear in bunches or randomly, we look at the spacing between genre births. Create a vector, intergenre intervals, which shows how many

years elapsed between new genres appearing. (If two genres appear in the same year, there should be a 0 in your vector, if three genres appear in the same year your vector should have two zeros, and so on. For example if the years that new genres appear are 1835, 1837, 1838, 1838, 1838 your vector should be 2, 1, 0, 0.) What is the mean of the time intervals between genre appearances? The standard deviation? The ratio of the standard deviation to the mean, called the coefficient of variation? Hint: The diff() function might help you here. Check out ?diff.

```
intergenre_intervals = diff(moretti$Begin)
mean_elapse <- mean(intergenre_intervals)
mean_elapse

## [1] 3.44186

sd_elapse <- sd(intergenre_intervals)
sd_elapse

## [1] 3.705224

coef_of_variation <- sd_elapse/mean_elapse
coef_of_variation

## [1] 1.076518
```

7. For a Poisson process, the coefficient of variation is expected to be around 1. However, that calculation doesn't account for the way Moretti's dates are rounded to the nearest year, or tell us how much the coefficient of variation might fluctuate. We will handle both of these by simulation.

(a) Write a function which takes a vector of numbers, representing how many new genres appear in each year, and returns the vector of the intervals between appearances. Check that your function works by seeing that when it is given new genres, it returns intergenre intervals.

```
get_interval <- function(v){
  posi <- c()
  for (i in c(1:length(v))){
    value <- v[i]
    if(value != 0){
      for (j in c(1:value)){
        posi = c(posi,i)
      }
    }
  }
  interval <- diff(posi)
  return(interval)
}

get_interval(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
```

When it is given new_genres, the output is exactly intergenre_intervals.

(b) Write a function to simulate a Poisson process and calculate the coefficient of variation of its inter-appearance intervals. It should take as arguments the number of years to simulate and the mean number of genres per year. It should return a list, one component of which is the vector of inter-appearance intervals, and the other their coefficient of variation. Run it with 161 years and a mean of 0:273; the mean of the intervals should generally be between 3 and 4.

```
possion_sim <- function(number_years, ave_genre){
  data_points <- rpois(number_years, ave_genre)
  intervals <- get_interval(data_points)
  result <- list(inter_appearence_intervals=intervals,
               coef_var <- sd(intervals)/mean(intervals))
  return(result)
}

outcome <- possion_sim(161,0.273)
mean(outcome[[1]])

## [1] 3.18
```

I tried several times. The mean values are all between 3 and 4

8. Run your simulation 10,000 times, taking the coefficient of variation (only) from each. (This should take less than two minutes to run.) What fraction of simulations runs have a higher coefficient of variation than Moretti's data?

```
r <- numeric(10000)
for(i in c(1:10000)){
  r[i] <- possion_sim(161,0.273)[[2]]
}
sum(r > coef_of_variation)/10000

## [1] 0.2307
```

23.04% of all simulations runs have a higher coefficient of variation than Moretti's data.

9. Explain what this does and does not tell you about the conjecture that genres tend to appear together in burst?

It tells us that the genres not appear in burst. It appears that the coefficient of variation of simulated poission is much smaller than out outcome of the intergenre intervals. The coefficient of variation shows the dispersion of data and the smaller it is, the more concentrate the date will be. So, it shows that the intergenre intervals are more seperated than the simulated poission, the genres not tend to appear together in burst.