CIND 123 - Data Analytics: Basic Methods

Rui Zhang

Assignment 2 (10%)

Rui Zhang

DHA 500736315

Instructions

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. Review this website for more details on using R Markdown http://rmarkdown.rstudio.com.

Use RStudio for this assignment. Complete the assignment by inserting your R code wherever you see the string "#INSERT YOUR ANSWER HERE".

When you click the **Knit** button, a document (PDF, Word, or HTML format) will be generated that includes both the assignment content as well as the output of any embedded R code chunks.

Submit **both** the rmd and generated output files. Failing to submit both files will be subject to mark deduction.

Sample Question and Solution

Use `seq()` to create the vector $(1, 2, 3, \ldots, 20)$.

```r
seq(1,20)
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
```

Question 1

The Titanic Passenger Survival DataSet provides information on the fate of passengers on the fatal maiden voyage of the ocean liner "Titanic." The dataset is available from the Department of Biostatistics at the Vanderbilt University School of Medicine (http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/titanic3.csv)in several formats. store the Titanic DataSet `titanic_train` using the following commands.

```
library(titanic)
titanicDataset <- read.csv(file = "http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/titanic3.csv
summary(titanicDataset)
```

```
##      pclass         survived         name               sex           
##  Min.   :1.000   Min.   :0.000   Length:1309        Length:1309       
##  1st Qu.:2.000   1st Qu.:0.000   Class :character   Class :character  
##  Median :3.000   Median :0.000   Mode  :character   Mode  :character  
##  Mean   :2.295   Mean   :0.382                                        
##  3rd Qu.:3.000   3rd Qu.:1.000                                        
##  Max.   :3.000   Max.   :1.000                                        
##                                                                       
##       age            sibsp            parch          ticket          
##  Min.   : 0.17   Min.   :0.0000   Min.   :0.000   Length:1309       
##  1st Qu.:21.00   1st Qu.:0.0000   1st Qu.:0.000   Class :character  
##  Median :28.00   Median :0.0000   Median :0.000   Mode  :character  
##  Mean   :29.88   Mean   :0.4989   Mean   :0.385                     
##  3rd Qu.:39.00   3rd Qu.:1.0000   3rd Qu.:0.000                     
##  Max.   :80.00   Max.   :8.0000   Max.   :9.000                     
##  NA's   :263                                                        
##       fare            cabin             embarked             boat          
##  Min.   :  0.000   Length:1309        Length:1309        Length:1309       
##  1st Qu.:  7.896   Class :character   Class :character   Class :character  
##  Median : 14.454   Mode  :character   Mode  :character   Mode  :character  
##  Mean   : 33.295                                                           
##  3rd Qu.: 31.275                                                           
##  Max.   :512.329                                                           
##  NA's   :1                                                                 
##       body        home.dest        
##  Min.   :  1.0   Length:1309       
##  1st Qu.: 72.0   Class :character  
##  Median :155.0   Mode  :character  
##  Mean   :160.8                     
##  3rd Qu.:256.0                     
##  Max.   :328.0                     
##  NA's   :1188
```

a) Extract the columns `sex`, `age`, `cabin` and `survived` into a new data frame of the name 'titanicSubset'.

```
titanicSubset <- titanicDataset[c("sex","age","cabin","survived")]
summary(titanicSubset)
```

```
##      sex                 age            cabin              survived    
##  Length:1309        Min.   : 0.17   Length:1309        Min.   :0.000  
##  Class :character   1st Qu.:21.00   Class :character   1st Qu.:0.000
##  Mode  :character   Median :28.00   Mode  :character   Median :0.000
##                     Mean   :29.88                      Mean   :0.382
##                     3rd Qu.:39.00                      3rd Qu.:1.000
##                     Max.   :80.00                      Max.   :1.000
##                     NA’s   :263
```

b) Use the aggregate() function to display the total number of survivors grouped by `sex`

```
aggregate(titanicDataset$survived, by=list(titanicDataset$sex), FUN=sum, na.rm=TRUE)
```

```
##   Group.1   x
## 1  female 339
## 2    male 161
```

c) Use the count() function in `dplyr` package to display the total number of passengers within each Ticket Class `Pclass`.

```
#install.packages("dplyr")
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
count(titanicDataset, pclass)
```

```
## # A tibble: 3 x 2
##   pclass     n
##    <int> <int>
## 1      1   323
## 2      2   277
## 3      3   709
```

d) Answer the following graphically:

1. What was the survival rates for females and males?
2. What was the age distribution on the Titanic?

```
sum(titanicDataset[titanicDataset$sex=="male",]$survived, na.rm=T)/nrow(titanicDataset[titanicDataset$s
```

```
## [1] 0.1909846
```

```
sum(titanicDataset[titanicDataset$sex=="female",]$survived, na.rm=T)/nrow(titanicDataset[titanicDataset
```

```
## [1] 0.7274678
```

```
summary(titanicDataset$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA’s
##    0.17   21.00   28.00   29.88   39.00   80.00     263
```

e)Use the `for` loop and `if` control statements to list the children’s names, aged 14 or under, on the Titanic.

```
titanicDataset.nona = titanicDataset[!is.na(titanicDataset$age),]
for (i in 1:nrow(titanicDataset.nona)) {
    if (titanicDataset.nona[i,]$age <= 14) {
      print(titanicDataset.nona[i,]$name)
    }
}
```

```
## [1] "Allison, Master. Hudson Trevor"
## [1] "Allison, Miss. Helen Loraine"
## [1] "Carter, Master. William Thornton II"
## [1] "Carter, Miss. Lucile Polk"
## [1] "Dodge, Master. Washington"
## [1] "Ryerson, Master. John Borie"
## [1] "Spedden, Master. Robert Douglas"
## [1] "Becker, Master. Richard F"
## [1] "Becker, Miss. Marion Louise"
## [1] "Becker, Miss. Ruth Elizabeth"
## [1] "Caldwell, Master. Alden Gates"
## [1] "Collyer, Miss. Marjorie \"Lottie\""
## [1] "Davies, Master. John Morgan Jr"
## [1] "Drew, Master. Marshall Brines"
## [1] "Hamalainen, Master. Viljo"
## [1] "Harper, Miss. Annie Jessie \"Nina\""
## [1] "Hart, Miss. Eva Miriam"
## [1] "Laroche, Miss. Louise"
## [1] "Laroche, Miss. Simonne Marie Anne Andree"
## [1] "Mallet, Master. Andre"
## [1] "Mellinger, Miss. Madeleine Violet"
## [1] "Nasser, Mrs. Nicholas (Adele Achem)"
## [1] "Navratil, Master. Edmond Roger"
## [1] "Navratil, Master. Michel M"
## [1] "Quick, Miss. Phyllis May"
## [1] "Quick, Miss. Winifred Vera"
## [1] "Richards, Master. George Sibley"
## [1] "Richards, Master. William Rowe"
## [1] "Sweet, Mr. George Frederick"
## [1] "Watt, Miss. Bertha J"
## [1] "Wells, Master. Ralph Lester"
## [1] "Wells, Miss. Joan"
## [1] "West, Miss. Barbara J"
## [1] "West, Miss. Constance Mirium"
## [1] "Abbott, Master. Eugene Joseph"
## [1] "Aks, Master. Philip Frank"
## [1] "Andersson, Master. Sigvard Harald Elias"
## [1] "Andersson, Miss. Ebba Iris Alfrida"
## [1] "Andersson, Miss. Ellis Anna Maria"
## [1] "Andersson, Miss. Ingeborg Constanzia"
## [1] "Andersson, Miss. Sigrid Elisabeth"
## [1] "Asplund, Master. Carl Edgar"
## [1] "Asplund, Master. Clarence Gustaf Hugo"
## [1] "Asplund, Master. Edvin Rojj Felix"
## [1] "Asplund, Master. Filip Oscar"
## [1] "Asplund, Miss. Lillian Gertrud"
## [1] "Ayoub, Miss. Banoura"
## [1] "Baclini, Miss. Eugenie"
## [1] "Baclini, Miss. Helene Barbara"
## [1] "Baclini, Miss. Marie Catherine"
## [1] "Boulos, Master. Akar"
## [1] "Boulos, Miss. Nourelain"
## [1] "Coutts, Master. Eden Leslie \"Neville\""
## [1] "Coutts, Master. William Loch \"William\""
## [1] "Danbom, Master. Gilbert Sigvard Emanuel"
## [1] "Dean, Master. Bertram Vere"
## [1] "Dean, Miss. Elizabeth Gladys \"Millvina\""
## [1] "Emanuel, Miss. Virginia Ethel"
## [1] "Ford, Miss. Robina Maggie \"Ruby\""
## [1] "Goldsmith, Master. Frank John William \"Frankie\""
## [1] "Goodwin, Master. Harold Victor"
## [1] "Goodwin, Master. Sidney Leonard"
## [1] "Goodwin, Master. William Frederick"
## [1] "Goodwin, Miss. Jessie Allis"
## [1] "Goodwin, Mr. Charles Edward"
## [1] "Hassan, Mr. Houssein G N"
## [1] "Hirvonen, Miss. Hildur E"
## [1] "Johnson, Master. Harold Theodor"
## [1] "Johnson, Miss. Eleanor Ileen"
## [1] "Karun, Miss. Manca"
## [1] "Kink-Heilmann, Miss. Luise Gretchen"
## [1] "Klasen, Miss. Gertrud Emilia"
## [1] "Moor, Master. Meier"
## [1] "Nakid, Miss. Maria (\"Mary\")"
## [1] "Nicola-Yarred, Master. Elias"
## [1] "Nicola-Yarred, Miss. Jamila"
## [1] "Olsen, Master. Artur Karl"
## [1] "Palsson, Master. Gosta Leonard"
## [1] "Palsson, Master. Paul Folke"
## [1] "Palsson, Miss. Stina Viola"
## [1] "Palsson, Miss. Torborg Danira"
## [1] "Panula, Master. Eino Viljami"
## [1] "Panula, Master. Juha Niilo"
## [1] "Panula, Master. Urho Abraham"
## [1] "Panula, Mr. Jaako Arnold"
## [1] "Peacock, Master. Alfred Edward"
## [1] "Peacock, Miss. Treasteall"
## [1] "Rice, Master. Albert"
## [1] "Rice, Master. Arthur"
## [1] "Rice, Master. Eric"
## [1] "Rice, Master. Eugene"
## [1] "Rice, Master. George Hugh"
## [1] "Rosblom, Miss. Salli Helena"
## [1] "Sandstrom, Miss. Beatrice Irene"
## [1] "Sandstrom, Miss. Marguerite Rut"
## [1] "Skoog, Master. Harald"
## [1] "Skoog, Master. Karl Thorsten"
## [1] "Skoog, Miss. Mabel"
## [1] "Skoog, Miss. Margit Elizabeth"
## [1] "Strom, Miss. Telma Matilda"
## [1] "Svensson, Mr. Johan Cervin"
## [1] "Thomas, Master. Assad Alexander"
## [1] "Touma, Master. Georges Youssef"
## [1] "Touma, Miss. Maria Youssef"
## [1] "van Billiard, Master. Walter John"
## [1] "Van Impe, Miss. Catharina"
## [1] "Vestrom, Miss. Hulda Amanda Adolfina"
```

Question 2

In an experiment of rolling 10 dice simultaneously. Use the binomial distribution to calculate the followings:

a) The probability of getting six 6's

```
dbinom(6,10,1/6)
```

```
## [1] 0.002170635
```

b) The probability of getting six, seven, or eight 3's

```
sum(dbinom(c(6,7,8),10,1/6))
```

```
## [1] 0.002437313
```

c) The probability of getting six even numbers

```
sum(dbinom(c(0,2,4,6,8,10),10,1/6))
```

```
## [1] 0.5086708
```

Question 3

In a shipment of 20 engines, history shows that the probability of any one engine proving unsatisfactory is 0.1

a) Use the Binomial approximation to calculate the probability that at least three engines are defective?

```
1 - pbinom(2,20,0.1)
```

```
## [1] 0.3230732
```

b) Use the Poisson approximation to calculate the probability that at least three engines are defective?

```
ppois(2,0.1*20,lower.tail = F)
```

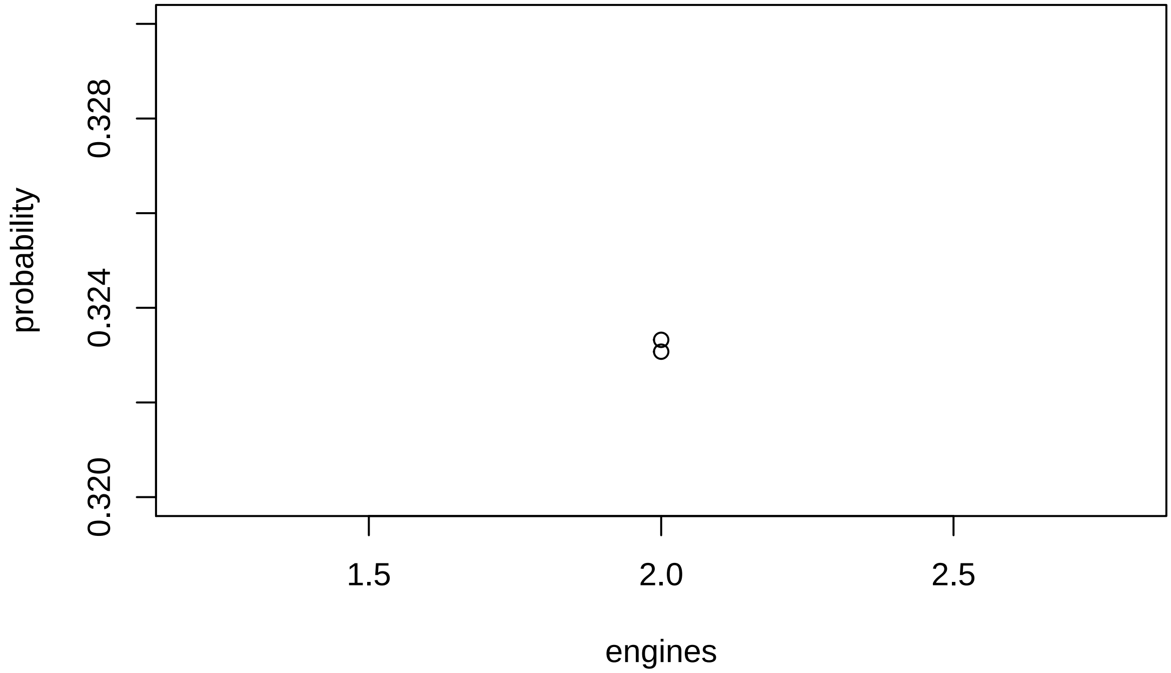
```
## [1] 0.3233236
```

c) Compare the results of parts a and b, then illustrate graphically on how well the Poisson probability distribution approximates the Binomial probability distribution.

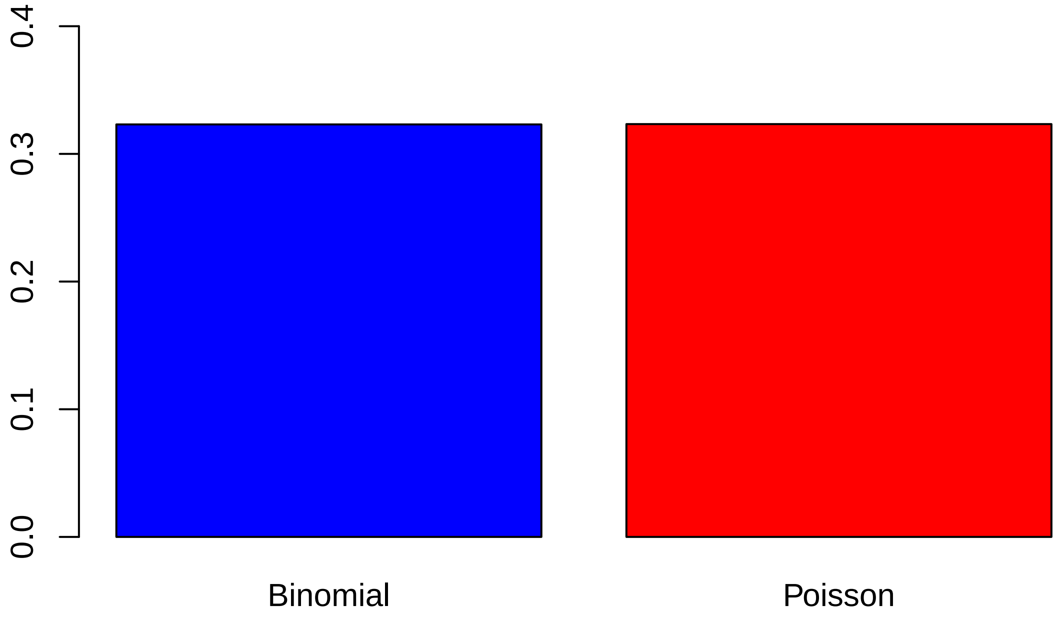
```
abs(1 - pbinom(2,20,0.1) - ppois(2,0.1*20,lower.tail = F))/(1 - pbinom(2,20,0.1))
```

```
## [1] 0.0007750225
```

```
plot(c(2,2),c(1 - pbinom(2,20,0.1),ppois(2,0.1*20,lower.tail = F)),xlab = "engines", ylab="probability"
```



```
barplot(c(1 - pbinom(2,20,0.1),ppois(2,0.1*20,lower.tail = F)),ylim=c(0,0.4), names.arg=c("Binomial", "
```



```
# Poisson probability are very close to Binomial probability, the relative error is 0.0007750225
```

Question 4

Write a script in R to compute the following probabilities of a normal random variable with mean 16 and variance 9

a) lies between 14.4 and 20.3 (inclusive)

```
pnorm(20.3,16,3)-pnorm(14.4,16,3)
```

```
## [1] 0.6272173
```

b) is greater than 21.8

```
pnorm(21.8,16,3,lower.tail = F)
```

```
## [1] 0.02659757
```

c) is less or equal to 10.5

```
pnorm(10.5,16,3)
```

```
## [1] 0.03337651
```

d) is less than 13 or greater than 19

```
pnorm(13,16,3) + pnorm(19,16,3,lower.tail = F)
```

```
## [1] 0.3173105
```

END of Assignment #2.