CIND 123 - Data Analytics: Basic Methods

Assignment 2 (10%)

[Alyzeh Jiwani]

[D20, 501106857]

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 $(1, 2, 3, \ldots, 20)$.

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

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

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(ggplot2)
```

Question 1

The Titanic Passenger Survival Data Set 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 (https://biostat.app.vumc.org/wiki/pub/Main/DataSets/titanic3.csv)in several formats. store the Titanic Data Set `titanic_train` using the following commands.

```
titanic_train <- read.csv('https://biostat.app.vumc.org/wiki/pub/Main/DataSets/titanic3.csv')
```

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

```
titanicSubset <- data.frame(Sex = titanic_train$sex, Age=titanic_train$age, Cabin=titanic_train$cabin, S
head(titanicSubset)
```

```
##      Sex   Age   Cabin Survived
## 1 female 29.00      B5        1
## 2   male  0.92 C22 C26        1
## 3 female  2.00 C22 C26        0
## 4   male 30.00 C22 C26        0
## 5 female 25.00 C22 C26        0
## 6   male 48.00     E12        1
```

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

```
sex_surv<-aggregate(Survived~Sex,data = titanicSubset, sum)
sex_surv
```

```
##      Sex Survived
## 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`.

```
titanic_train %>% count(pclass)
```

```
##   pclass   n
## 1      1 323
## 2      2 277
## 3      3 709
```

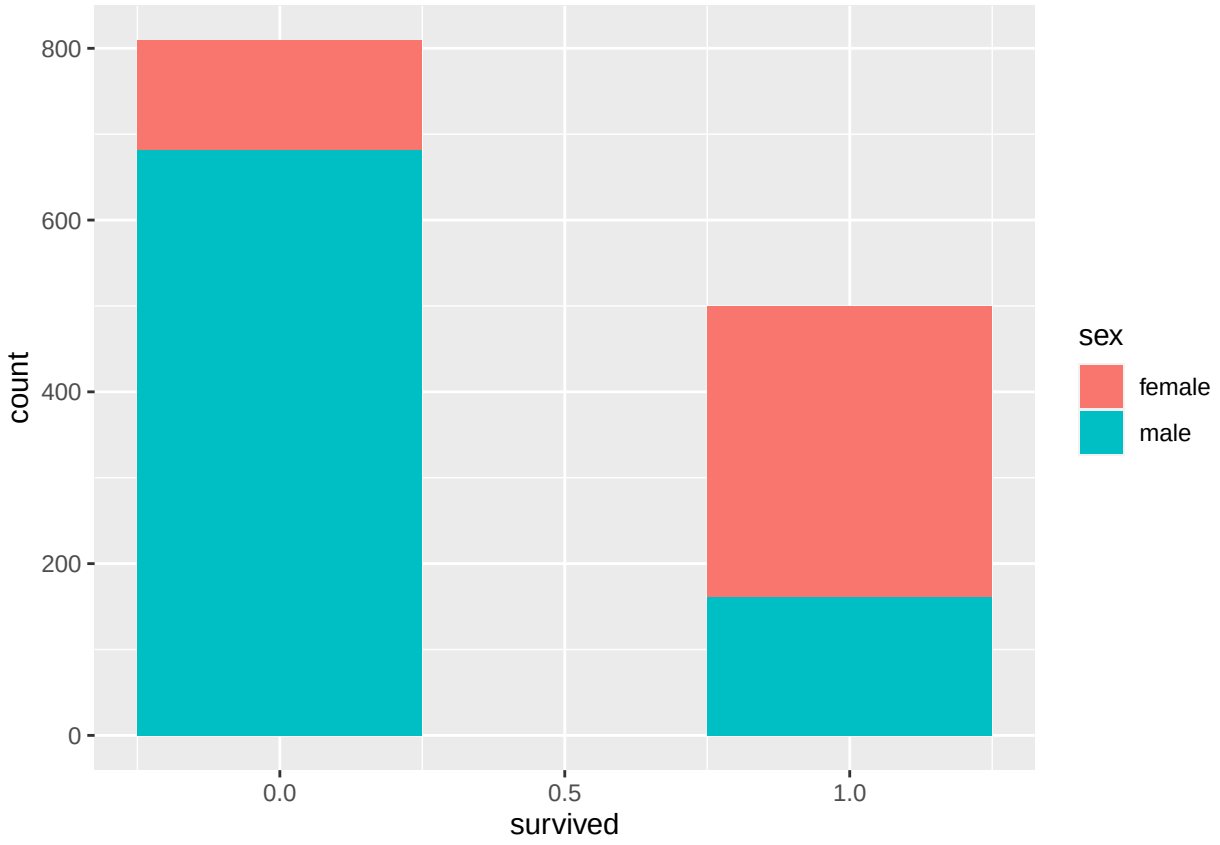
d) Answer the following graphically (using visualization):

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

Hint: You can use ggplot2

From the first we can see that the survival rate for females was much higher than that of males. Average age of people on board the titanic was 30, slight right skew (a larger proprtion of passengers were under the age of 30)

```
titanic_train%>%ggplot(aes(x=survived, fill = sex))+
  geom_bar(width=0.5)
```



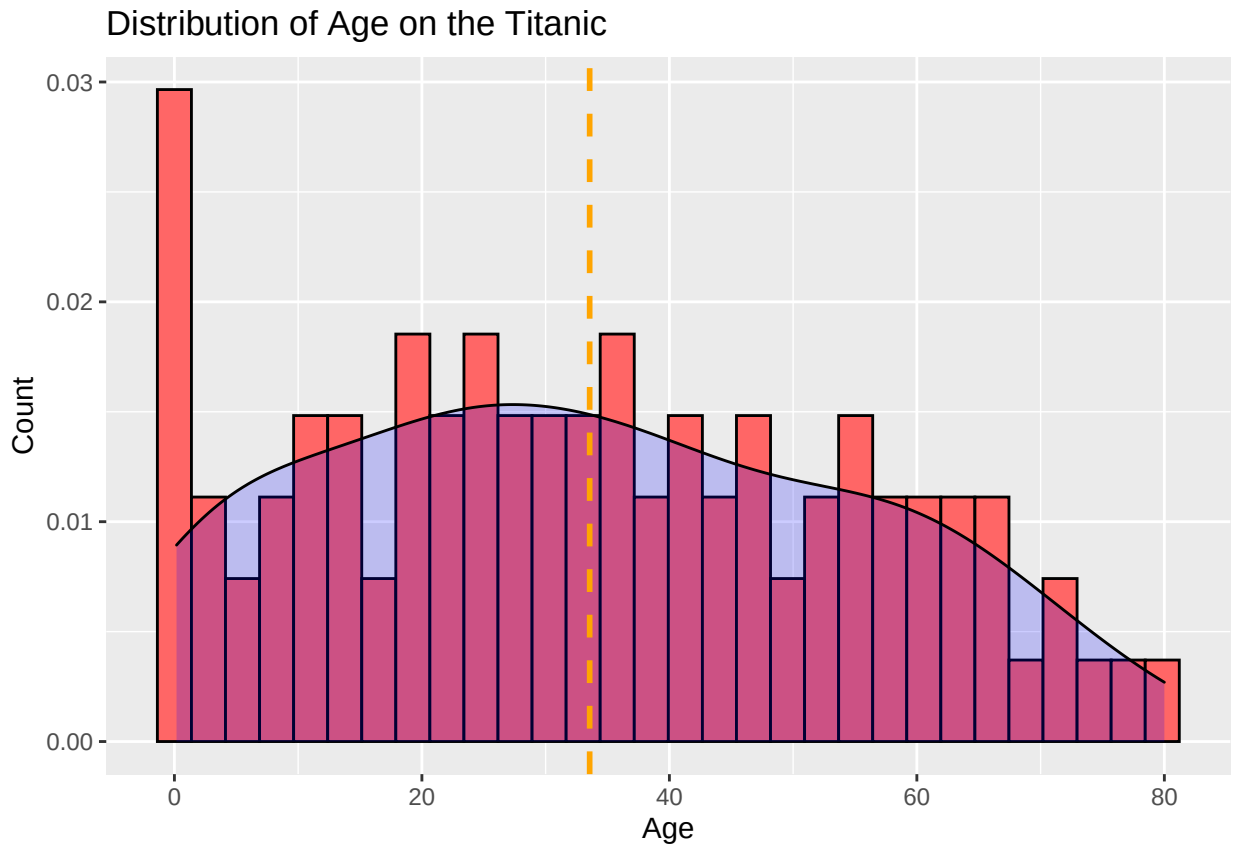

```
age_d <- titanic_train%>% count(age)
age_dist<-ggplot(data = age_d[1:98,],aes(x=age))+
  geom_histogram(aes(y=..density..),bin=30, colour= 'black', fill = '#FF6666')+
```

```
  geom_density(alpha=.2, fill = 'blue')+
  labs(title='Distribution of Age on the Titanic', y='Count', x='Age')+
  geom_vline(aes(xintercept=mean(age_d[1:98,]$age)), colour='orange', linetype='dashed', size = 1)
```

```
## Warning: Ignoring unknown parameters: bin
```

```
age_dist
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



e)Use the `for` loop and `if` control statements to list the women's names, age 34 or more that embarked from S (Southampton), on the Titanic.

```
lis <-list()
d<-0
for(i in 1:nrow(titanic_train)){
  if (!is.na(titanic_train$age[i])){
    if(titanic_train$sex[i]=='female'){
      if(titanic_train$age[i] >= 34){
        d<-d+1
        lis[d] <-titanic_train$name[i]
      }
    }
  }
```

```
}
head(lis)
```

```
## [[1]]
## [1] "Andrews, Miss. Kornelia Theodosia"
##
## [[2]]
## [1] "Appleton, Mrs. Edward Dale (Charlotte Lamson)"
##
## [[3]]
## [1] "Baxter, Mrs. James (Helene DeLaudeniere Chaput)"
##
## [[4]]
## [1] "Beckwith, Mrs. Richard Leonard (Sallie Monypeny)"
##
## [[5]]
## [1] "Bidois, Miss. Rosalie"
##
## [[6]]
## [1] "Bissette, Miss. Amelia"
```

```
length(lis)
```

```
## [1] 129
```

Question 2

A study was conducted on GRE test takers to evaluate the success conditions. The success rate is 25%. A sample of 30 test takers is selected for the study. Use the binomial distribution to calculate the followings:

a) The probability that 10 test takers fail the GRE test:

```
dbinom(x=20, size=30,prob=0.25)
```

```
## [1] 1.538811e-06
```

b) The probability of getting at least five test takers succeed in the test

```
#p(at least 5 pass)=(p>=5)=p(>4)=1-p(<=4)
1-pbinom(4,size=30,prob=0.25)
```

```
## [1] 0.9021304
```

c) The probability of 25 or less fail the test

This is the same thing as finding the probability in part b.

```
#p(25 or less fail) = p(5 or more pass)=same probability as above
pbinom(25, size = 30, prob=0.75)
```

```
## [1] 0.9021304
```

Question 3

In a shipment of 100 tiles in a box, history shows that the probability of one tile in a box is defective is 0.2

a) Use the Binomial approximation to calculate the probability that more than 20 tiles are defective?

```
#P(>20) = 1-p(<=20)
1-pbinom(20, size=100, prob=0.2)
```

```
## [1] 0.4405384
```

b) Use the Poisson approximation to calculate the probability that at most 20 tiles are defective?

```
#lamda = 0.2*100=20
#at most 20 = <=20
ppois(20, lambda = 20)
```

```
## [1] 0.5590926
```

c) Use the binomial approximation to calculate the probability that at most 20 tiles are defective?

```
#p(<=20)
pbinom(20, size=100, prob=0.2)
```

```
## [1] 0.5594616
```

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

We can see from our results from parts b and c that to 3 significant figures, the poisson distributions approximation is the same as that of the binomial. This is represented in our first graph where we can see that both curves are almost layered exactly on top of each other, with near misses of less than 0.015. This is likely to be a result of the value of n being so large

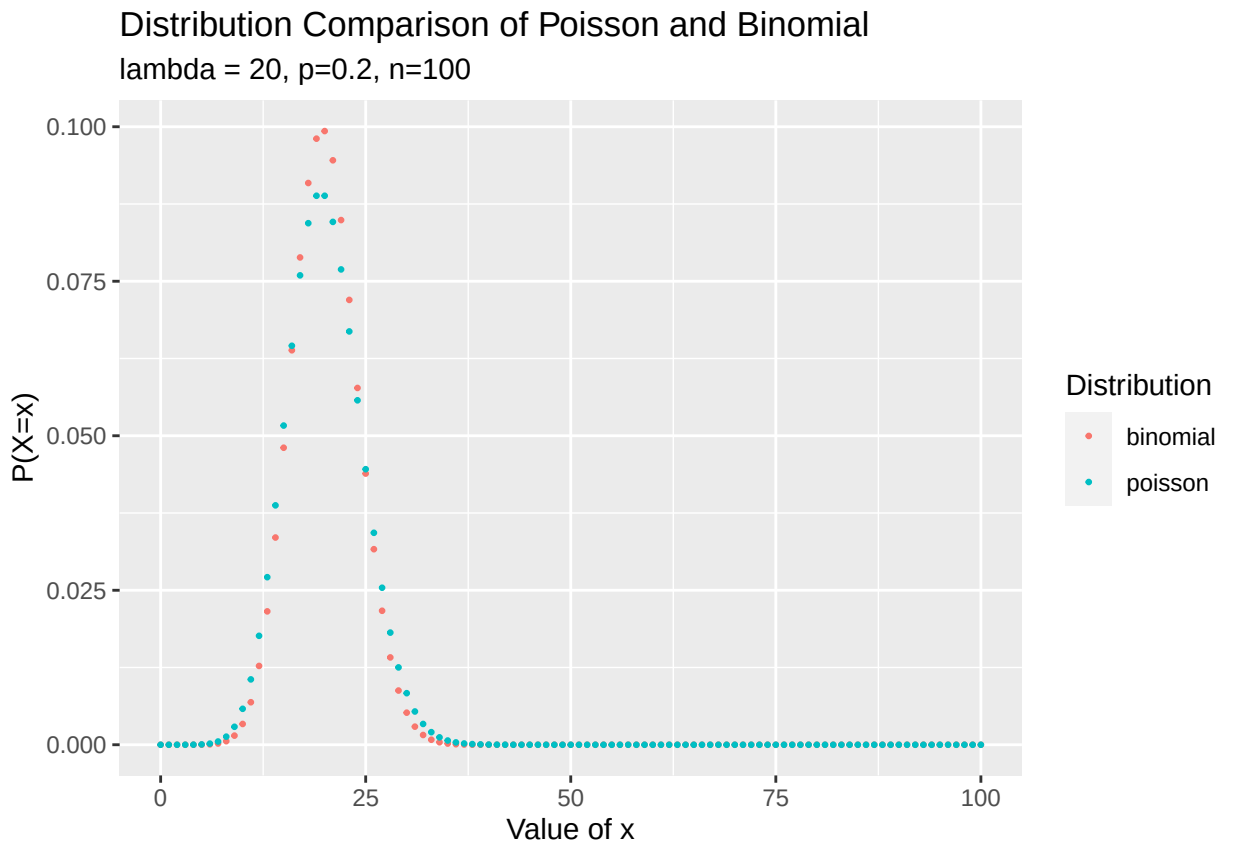
Our second graph compares our results in Q2 with a poisson model. we can see that as n -> 100, the possion approximation to a binomial distribution becomes more accurate as the the gap between the poisson and binomial distributions is much larger at certain points, at about 0.025 apart.

```
bi_dist<-dbinom(seq(0,100), size = 100, prob = 0.2)
pois_dist<-dpois(seq(0,100), lambda=20)
bin_df <- data.frame(Distribution= c(rep('binomial', times=101)), x=seq(0,100), Probability = bi_dist)
pois_df <- data.frame(Distribution = c(rep('poisson', times=101)),x=seq(0,100), Probability=pois_dist)

dist_compare <-rbind(bin_df,pois_df)
head(dist_compare)
```

```
##   Distribution x  Probability
## 1     binomial 0 2.037036e-10
## 2     binomial 1 5.092590e-09
## 3     binomial 2 6.302080e-08
## 4     binomial 3 5.146699e-07
## 5     binomial 4 3.120186e-06
## 6     binomial 5 1.497689e-05
```

```
ggplot(dist_compare, aes(x=x, y = Probability, colour=Distribution))+
  geom_point(size=0.5)+
  labs(title='Distribution Comparison of Poisson and Binomial', subtitle='lambda = 20, p=0.2, n=100', x=
```

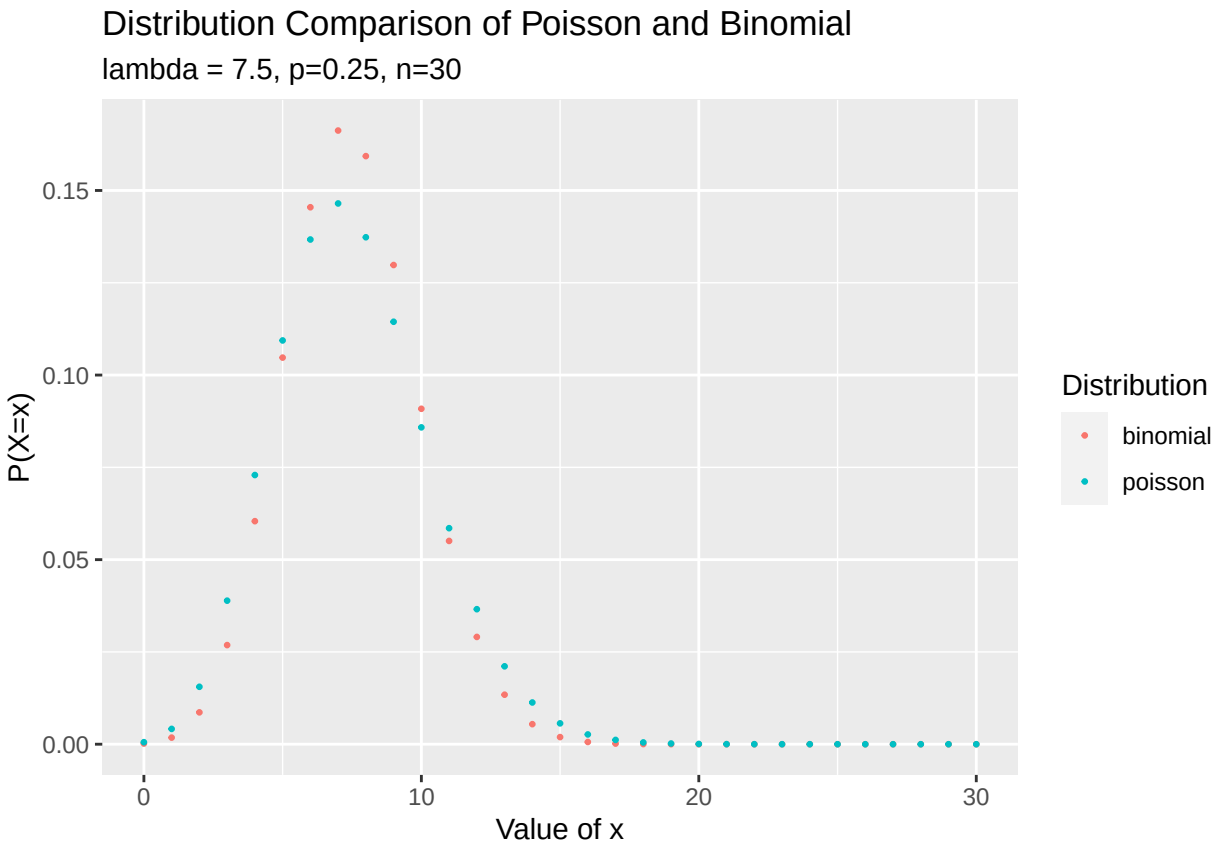



```
bd<-dbinom(seq(0,30), size = 30, prob = 0.25)
pd <- dpois(seq(0,30), lambda = 7.5)
bdf<- data.frame(Distribution= c(rep('binomial', times=31)), x=seq(0,30), Probability = bd)
pdf <- data.frame(Distribution = c(rep('poisson', times=31)),x=seq(0,30), Probability=pd)

d_compare <-rbind(bdf,pdf)
head(d_compare)
```

```
##   Distribution x  Probability
## 1     binomial 0 0.0001785821
## 2     binomial 1 0.0017858209
## 3     binomial 2 0.0086314677
## 4     binomial 3 0.0268534550
## 5     binomial 4 0.0604202738
## 6     binomial 5 0.1047284747
```

```
ggplot(d_compare, aes(x=x, y = Probability, colour=Distribution))+
  geom_point(size=0.5)+
  labs(title='Distribution Comparison of Poisson and Binomial', subtitle='lambda = 7.5, p=0.25, n=30', x
```


Distribution Comparison of Poisson and Binomial
lambda = 7.5, p=0.25, n=30
P(X=x)
0.15
0.10
0.05
0.00
0
10
20
30
Value of x
Distribution
binomial
poisson

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, mean = 16, sd = 3, lower.tail = TRUE)-pnorm(14.4, mean = 16, sd = 3, lower.tail = TRUE)
```

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

b) is greater than 21.8

```
1-pnorm(21.8, mean = 16, sd = 3, lower.tail = TRUE)
```

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

c) is less than or equal to 10.5

```
pnorm(10.5, mean=16, sd=3, lower.tail=TRUE)
```

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

d) is less than 13 or greater than 19

```
pnorm(13, mean=16,sd=3, lower.tail = TRUE) + pnorm(19, mean=16,sd=3, lower.tail=FALSE)
```

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

END of Assignment #2.