# Assignment-1

# SIT741 – Statistical Data Analysis

Shantanu Gupta
Deakin University
Student Id: 218200234

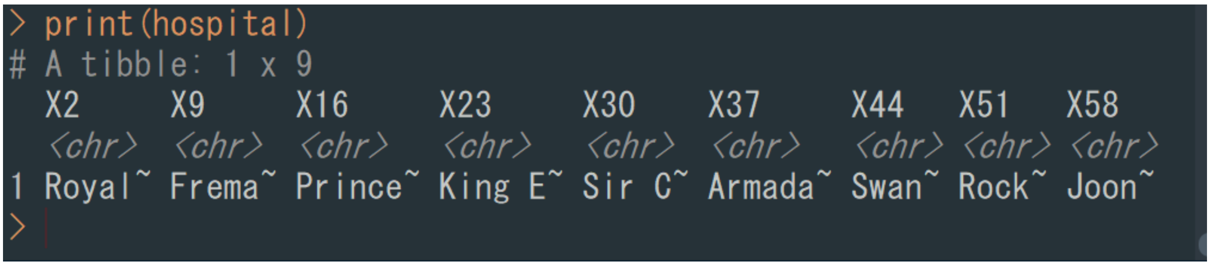
# Task1:

## Task 1.2

1. There are total 367 rows in the dataset. First row contains the hospital name. Second row contains the attributes. Third row contains the actual data. There are 365 observations, 9 different hospital and 8 different attributes.
2. Data types are present in the dataset are numerical, character.
3. The data includes patient attendances, admissions and admissions by five different Triage levels from 1 to 5 for 1 complete financial year from 1-July-2013 to 30 June2014.
4. Attendances - The number of patients recorded as arriving at a public emergency department. An emergency attendance is recorded where a patient is registered in any manner in one of the electronic data collection systems. Attendance counts may include patients who are dead on Arrival (DOA) or those who did not wait to be seen. Admissions - The number of patients who are subsequently admitted to the hospital for care and/or treatment. The patient may be missing a triage category, or may have a triage other than 1 to 5, or may not be clerically registered.
5. Triage categories are allocated to each patient based on an assessment of their presenting conditions, generally by the triage nurse, with triage 1 being the most urgent and triage 5 being the least urgent. (Triage 1: Resuscitation-immediate, within seconds; Triage 2: Emergency- within 10 minutes; Triage 3: Urgent- within 30 minutes; Triage 4: Semi-urgent- within 60 minutes; Triage 5: Non-urgent - within 120 minutes) Triaged, indicated by a code of 1, 2, 3, 4 or 5.

# Task2:

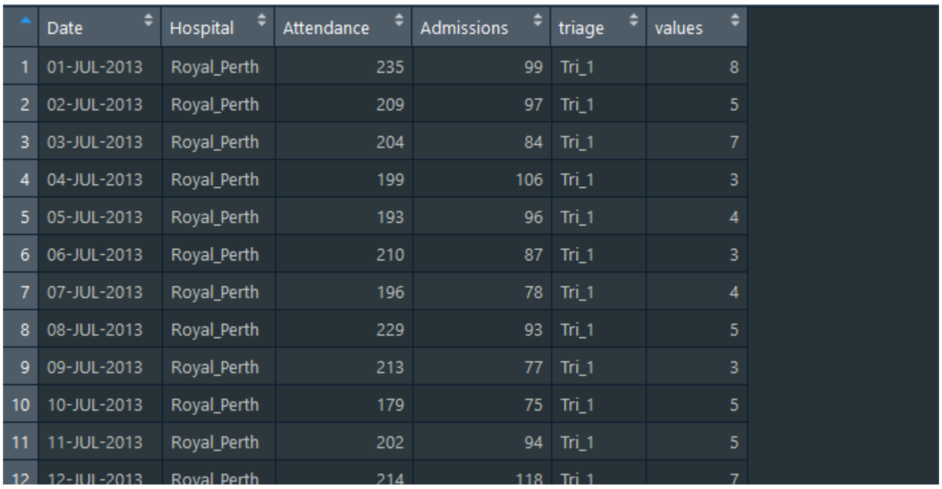
## Task 2.1



```
> print(hospital)
# A tibble: 1 x 9
  X2     X9     X16     X23     X30    X37     X44   X51   X58
  <chr>  <chr>  <chr>   <chr>   <chr>  <chr>   <chr> <chr> <chr>
1 Royal~ Frema~ Prince~ King E~ Sir C~ Armada~ Swan~ Rock~ Joon~
>
```

## Task2.2

1. All 3 question the answer is yes.



|  | Date | Hospital | Attendance | Admissions | triage | values |
|---|---|---|---|---|---|---|
| 1 | 01-JUL-2013 | Royal_Perth | 235 | 99 | Tri_1 | 8 |
| 2 | 02-JUL-2013 | Royal_Perth | 209 | 97 | Tri_1 | 5 |
| 3 | 03-JUL-2013 | Royal_Perth | 204 | 84 | Tri_1 | 7 |
| 4 | 04-JUL-2013 | Royal_Perth | 199 | 106 | Tri_1 | 3 |
| 5 | 05-JUL-2013 | Royal_Perth | 193 | 96 | Tri_1 | 4 |
| 6 | 06-JUL-2013 | Royal_Perth | 210 | 87 | Tri_1 | 3 |
| 7 | 07-JUL-2013 | Royal_Perth | 196 | 78 | Tri_1 | 4 |
| 8 | 08-JUL-2013 | Royal_Perth | 229 | 93 | Tri_1 | 5 |
| 9 | 09-JUL-2013 | Royal_Perth | 213 | 77 | Tri_1 | 3 |
| 10 | 10-JUL-2013 | Royal_Perth | 179 | 75 | Tri_1 | 5 |
| 11 | 11-JUL-2013 | Royal_Perth | 202 | 94 | Tri_1 | 5 |
| 12 | 12-JUL-2013 | Royal_Perth | 214 | 118 | Tri_1 | 7 |

2. I have used gather only 1 time. Didn’t require to use spread function anytime in my dataset.

gather (hospital,triage,values,Tri_1:Tri_5)

Value-It simply a frequency that I wanted to merge on the basis of last column.

Triage-key

3. No, some of the variables are not in correct datatype.

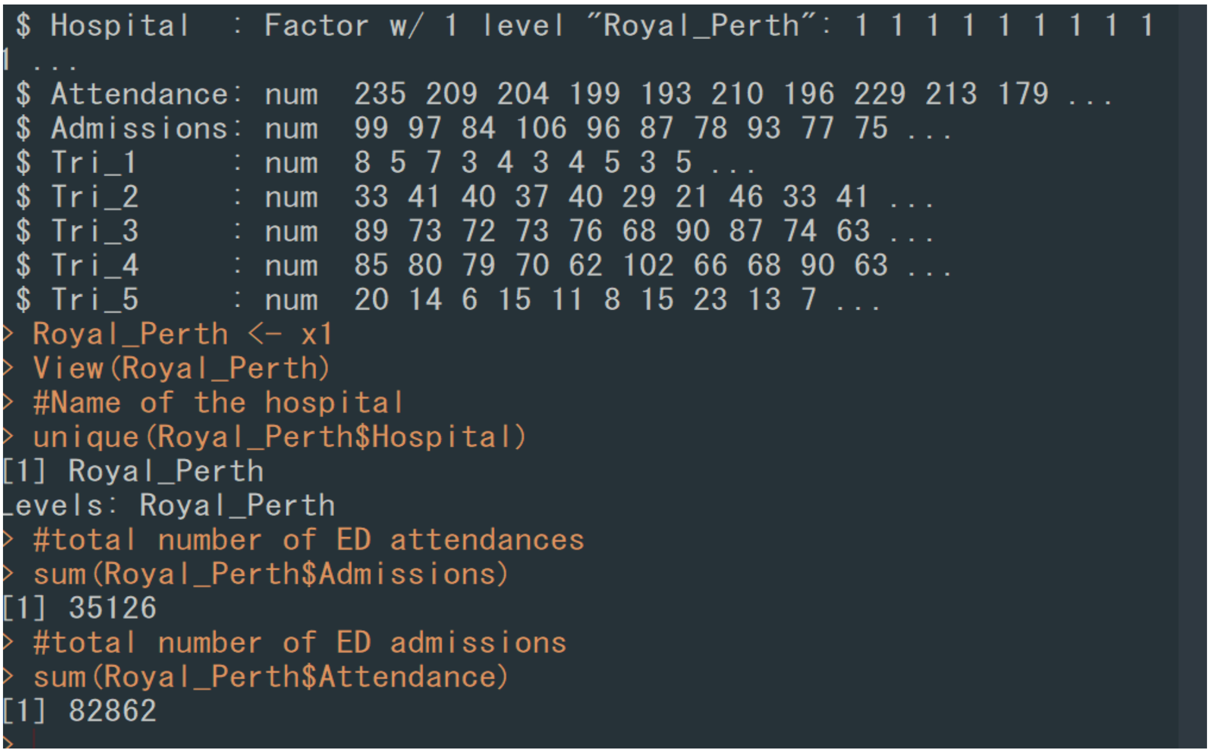
Date is in the form of character. We need to change it in the POSIXCT format to make a time-series plot. Sometimes we need to convert character to numerical.

4.

I have fixed the missing value by each hospital, after that I combine a big hospital data frame in which all hospitals are contained. We fixed all the missing values by mean imputation method. In every hospital there is a missing values in Triage_1 and Triage_5.

### Task 3.1

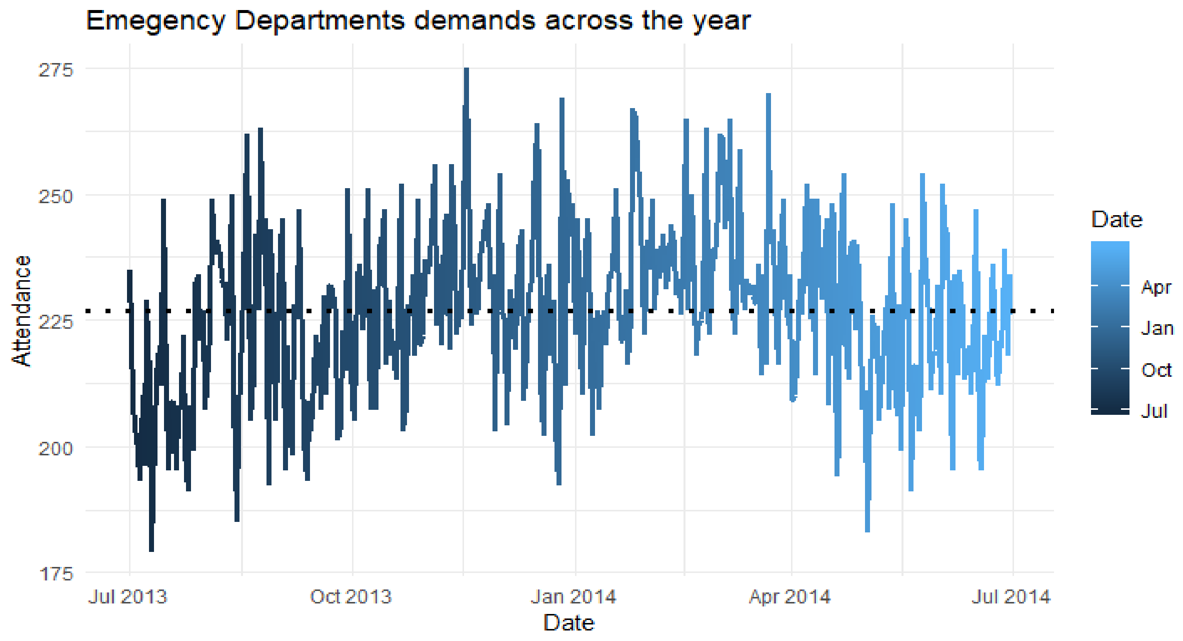
Hospital Selected-Royal Perth Hospital



```
 $ Hospital  : Factor w/ 1 level "Royal_Perth": 1 1 1 1 1 1 1 1 1
1 ...
 $ Attendance: num  235 209 204 199 193 210 196 229 213 179 ...
 $ Admissions: num  99 97 84 106 96 87 78 93 77 75 ...
 $ Tri_1     : num  8 5 7 3 4 3 4 5 3 5 ...
 $ Tri_2     : num  33 41 40 37 40 29 21 46 33 41 ...
 $ Tri_3     : num  89 73 72 73 76 68 90 87 74 63 ...
 $ Tri_4     : num  85 80 79 70 62 102 66 68 90 63 ...
 $ Tri_5     : num  20 14 6 15 11 8 15 23 13 7 ...
> Royal_Perth <- x1
> View(Royal_Perth)
> #Name of the hospital
> unique(Royal_Perth$Hospital)
[1] Royal_Perth
Levels: Royal_Perth
> #total number of ED attendances
> sum(Royal_Perth$Admissions)
[1] 35126
> #total number of ED admissions
> sum(Royal_Perth$Attendance)
[1] 82862
>
```

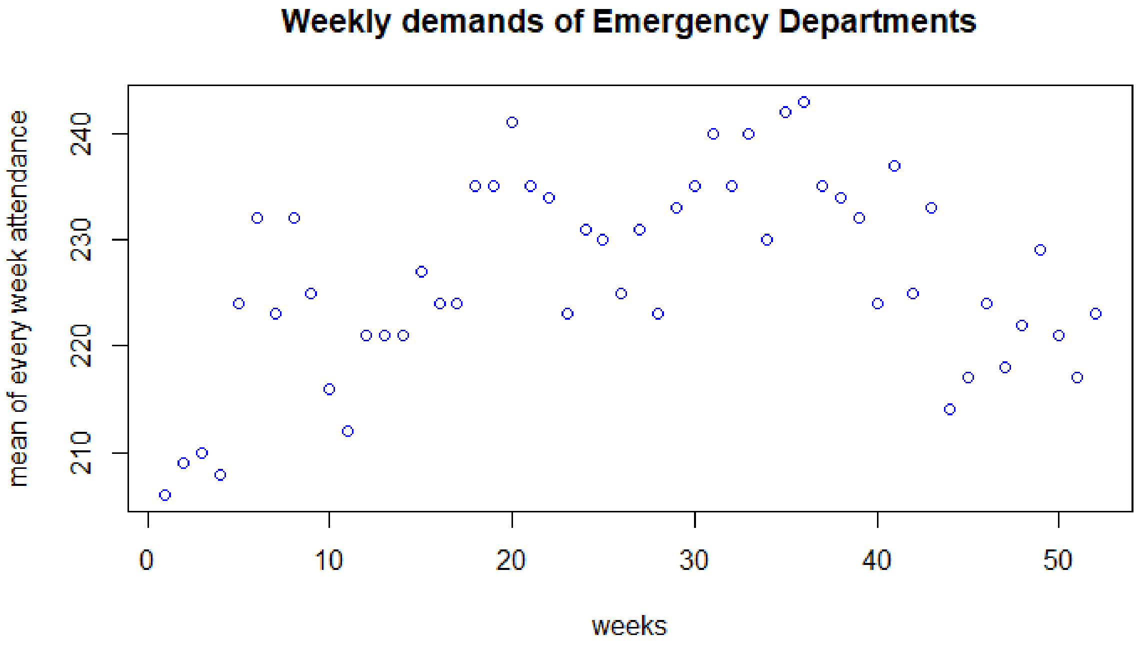
### Task 3.2

Most of the demands are constant every month. There is a surge of demands in the month of December to February. Some of the days demand is less than average i.e.227 but some days it is higher. It possess oscillating nature.

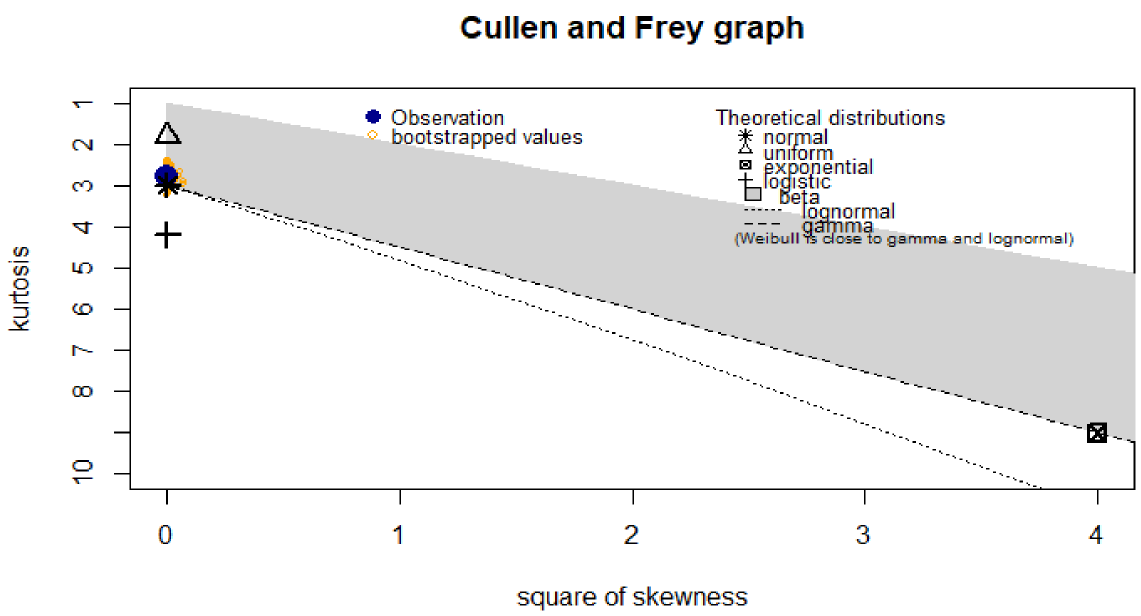


## Task 3.3

Initially and ending the demand is less. There is a lot of demands between weeks 20 and week 40.In the middle it is high. It increases as we go from winter to spring and summer session. After that it’s declining. In the month of summer demand is more and is very high. So we need large number of ambulance at that time. In the month of winter & autumn season demand is less.



## Task 3.4

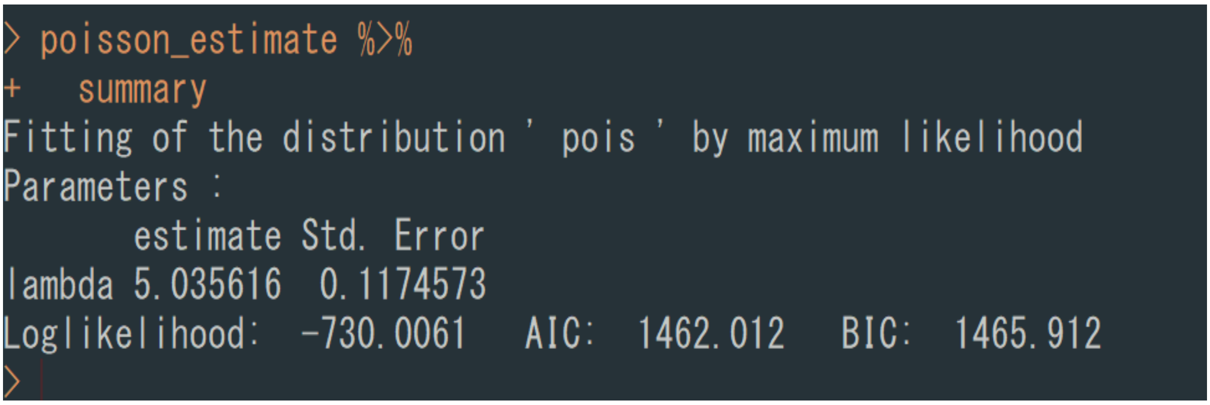


Normal distribution is best for attendance column.

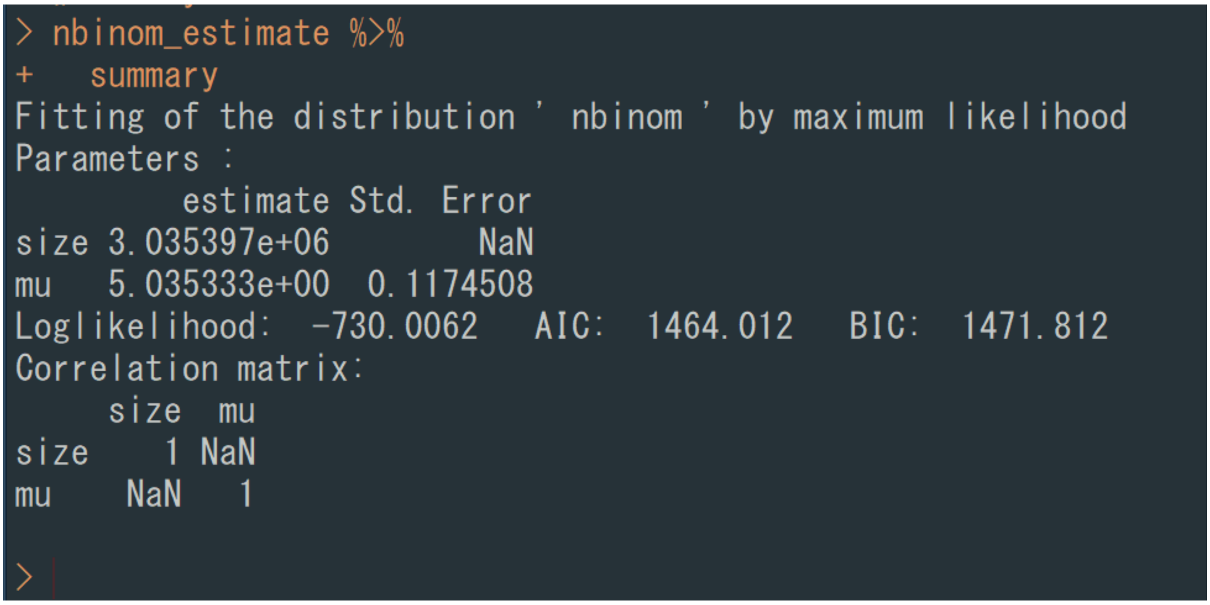
Tri_1, Tri_2, Tri_3, Tri_4, Tri_5 variables can be used for Poisson distribution.

## Task 4

### Task 4.2



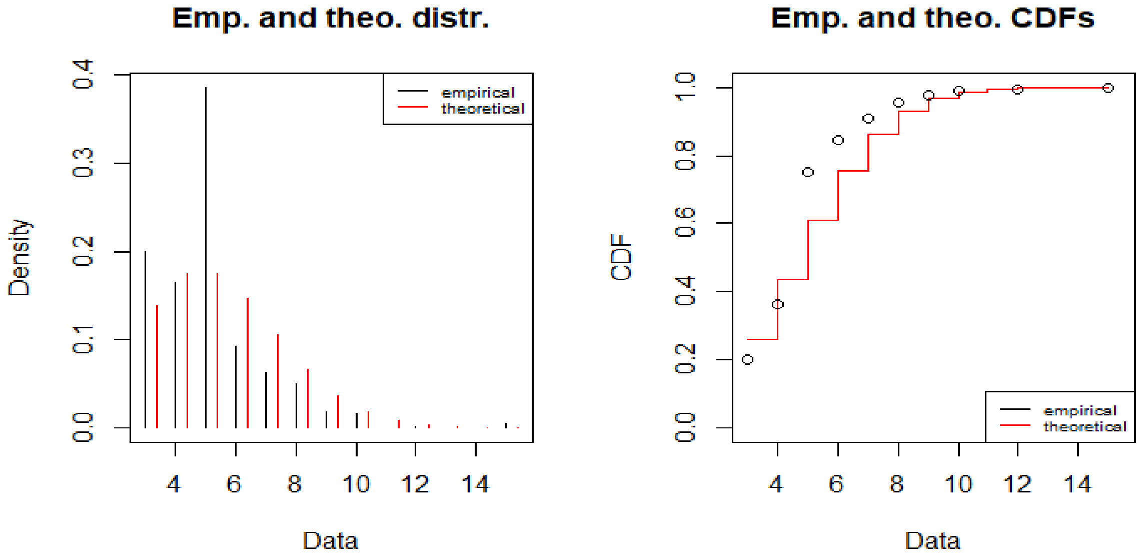
```
> poisson_estimate %>%
+   summary
Fitting of the distribution ' pois ' by maximum likelihood
Parameters :
        estimate Std. Error
lambda 5.035616  0.1174573
Loglikelihood:  -730.0061   AIC:  1462.012   BIC:  1465.912
>
```

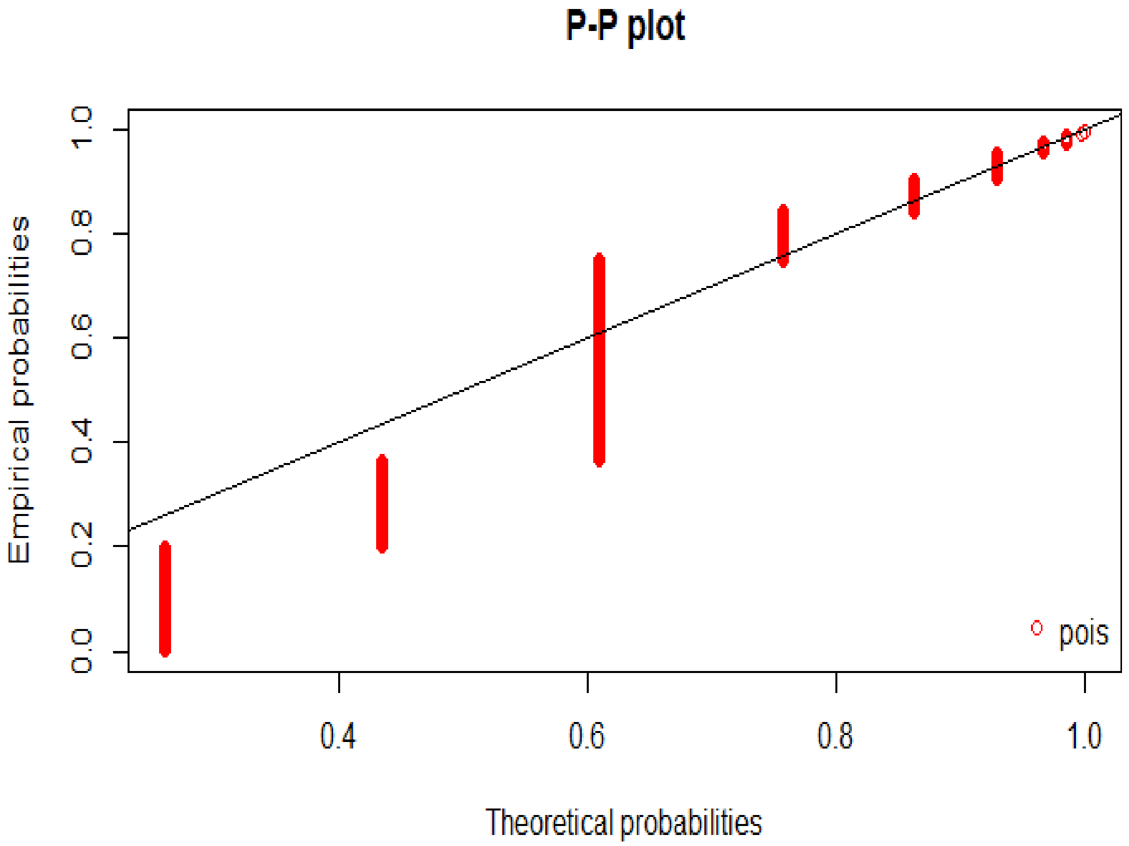


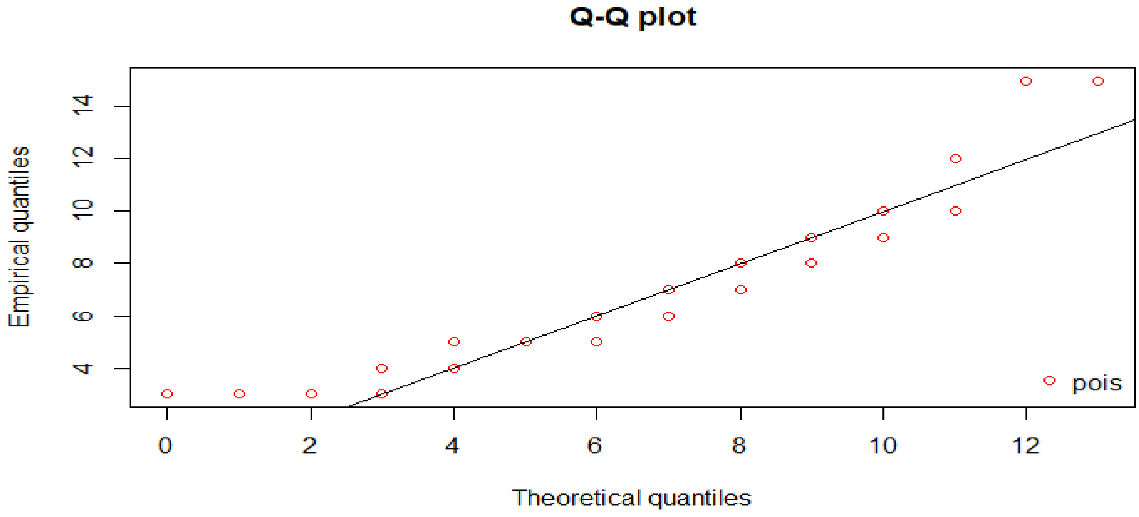
```
> nbinom_estimate %>%
+   summary
Fitting of the distribution ' nbinom ' by maximum likelihood
Parameters :
         estimate Std. Error
size 3.035397e+06        NaN
mu   5.035333e+00  0.1174508
Loglikelihood:  -730.0062   AIC:  1464.012   BIC:  1471.812
Correlation matrix:
     size  mu
size    1 NaN
mu    NaN   1

>
```

Log–likelihood should be a maximum. The AIC and BIC should be minimum because these are the errors. By seeing these two distributions values, I would say poisson is the better distribution than negative binomial distribution. Moreover the value of both distribution are very close to each other but in terms of number poisson is somewhat better than the negative binomial distribution. In other words, I have found that the percentage of this distribution over Triage_1 is 50% with the margin of error is +/-11.74%.95 percent of the time the data would match this results.




P-P plot
Empirical probabilities
0.0
0.2
0.4
0.6
0.8
1.0
0.4
0.6
0.8
1.0
Theoretical probabilities
pois


Q-Q plot
Empirical quantiles
4
6
8
10
12
14
0
2
4
6
8
10
12
Theoretical quantiles
pois

**Task 5:**

**Maximum Likelihood Estimation:**

The **maximum likelihood** (ML) estimate of a parameter is the value of that parameter under which your actual observed data are most likely, relative to any other possible values of the parameter.

**Pros:**

- When sample size n is large (n>30), MLE is unbiased, consistent, normally distributed, and efficient (“regularity conditions”) “Efficient” means it produces the minimum MSE than other methods including Method of Moments
- More useful in statistical inference.

**Cons:**

- MLE can be highly biased for small samples.
- Sometimes, MLE has no closed-form solution.
- MLE can be sensitive to starting values, which might not give a global optimum.

**Method of Moment Estimation:**

- Also commonly used to fit for parametric estimation
- No closed form formulas for this estimator, use numerical optimization techniques to find solution
- Finding the value of the parameter that matches the first theoretical raw moments of the parametric distribution to the corresponding empirical raw moments.

**Pros:**

- Easy to compute and always work .The method often provides estimators when other methods fail to do so or when estimators are hard to obtain.
- MME is consistent.

**Cons:**

- They are usually not the “best estimators” available. By best, we mean most efficient, i.e., achieving minimum square error.
- Sometimes it may be meaningless.

**Quantile Matching Estimation**

- Matching theoretical quantities of the parametric distribution for some specified probabilities against the empirical quantities.
- Empirical quantities are computed on observation x1: xn using the quantile function of the stats package.

**Cons:**

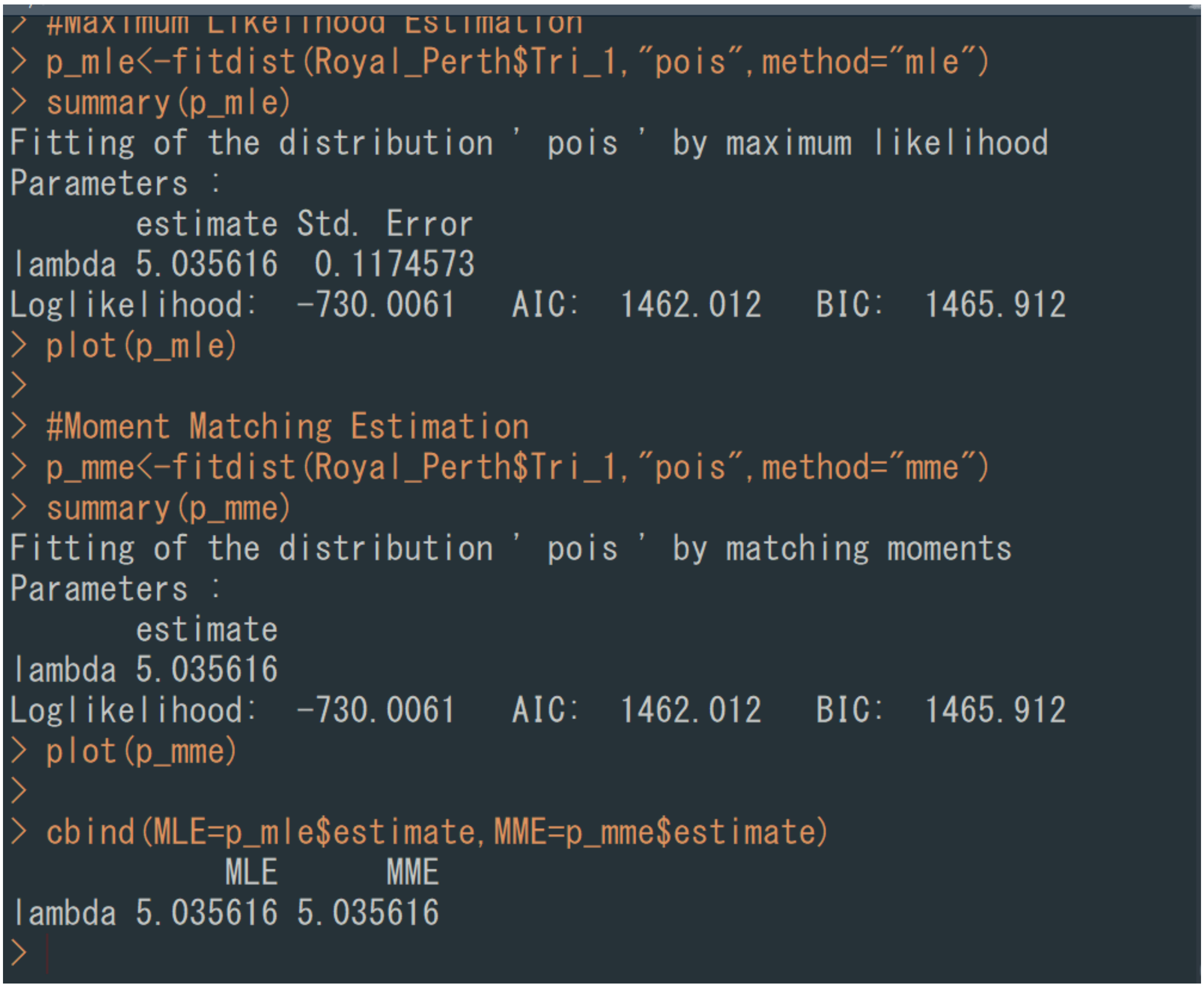
- Additional probability argument need to provide.

**Minimum Distance Estimation**

- Difference distance measures are used like Kolmogorov-Smirnov, Anderson-darling, Cramer-von mises distance
- Squared differences between candidate and empirical distribution function

**Cons:**

- Additional Goodness of fit parameter
- Not suitable for discrete distributions



```
> #Maximum Likelihood Estimation
> p_mle<-fitdist(Royal_Perth$Tri_1,"pois",method="mle")
> summary(p_mle)
Fitting of the distribution ' pois ' by maximum likelihood
Parameters :
       estimate Std. Error
lambda 5.035616  0.1174573
Loglikelihood:  -730.0061   AIC:  1462.012   BIC:  1465.912
> plot(p_mle)
>
> #Moment Matching Estimation
> p_mme<-fitdist(Royal_Perth$Tri_1,"pois",method="mme")
> summary(p_mme)
Fitting of the distribution ' pois ' by matching moments
Parameters :
       estimate
lambda 5.035616
Loglikelihood:  -730.0061   AIC:  1462.012   BIC:  1465.912
> plot(p_mme)
>
> cbind(MLE=p_mle$estimate,MME=p_mme$estimate)
            MLE      MME
lambda 5.035616 5.035616
>
```

## Task 6:

- I don't find any issues in the dataset that is breaching the privacy of an individual.
- One think that I should recommend in the dataset to use dummy names for the hospital name. Sometimes, data science results tarnished the hospital good image
- There are bunch of tools which are available in the market that make your data fully anonymous. To make data anonymous one very popular technique in the market is K-anonymity. K-anonymity might be described as a "hiding in the crowd" guarantee: if each individual is part of a larger group, then any of the records in this group could correspond to a single person. One such tool that I want to discuss is:

**ARX-Data Anonymization Tool**

- Open source software for anonymizing sensitive personal data
- Developed in close cooperation between the Chair for Biomedical Informatics, the Chair for IT Security and the Chair for Database Systems at TUM, Germany
- Download Free software from official website
- Tool transforms datasets into syntactic, statistical and semantic privacy models that mitigate attacks leading to privacy breaches.
- Supports all privacy model, data transformation model and quality model
- Good Documentation
- This tool also provides built-in data import facilities for relational databases (MS SQL, DB2, SQLite, MySQL), MS Excel and CSV files.

This software used in variety of applications like Commercial Big Data Analytics platform, Research Projects, Data Sharing, Training purpose. Department of Western Australia look into this software and publish dataset to public through this software. So, no breach of privacy and security concerns are there.

## Task 7:

1. Me and my 3 friends work together to complete this assignment. Basically, I learn these things from them:
    - General doubts regarding dataset
    - How to use loop for weekly time series plot
    - Learn some new commands that I don’t know
    - Separating different hospitals data frame

We worked collaboratively and whenever we stuck in some problem we discuss with each other and try to find the best optimal solution for that problem. If we did not find any solution, then we try to google it or website like Stack overflow, R-pubs, R-bloggers will come handy on that time.

2. I am an average student. I am expecting above 20 marks in my assignment.