Eindoprdracht deel 2

Mohammed Al Hor

2023-02-12

3. Now load the actual data into R and transform the data into an appropriate format for analysis using the scripts we will provide. Clean for outliers.Determine the average processing time for each phase (checking and admin) and determine the proportion of parcels sent out in time. Is the KPI target of 90% fulfilled?

```
summary(data$admin_time)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -10.3500   0.1517   0.1667   0.7985   0.3561  31.2167
```

```
summary(data$check_time)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -10.71667   0.08333   0.16667   0.29957   0.29167   7.37778
```

```
mean_check_time
```

```
## [1] 0.360472
```

```
mean_admin_time
```

```
## [1] 0.8573892
```

```
summary(data$admin_time)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -10.3500   0.1517   0.1667   0.7985   0.3561  31.2167
```

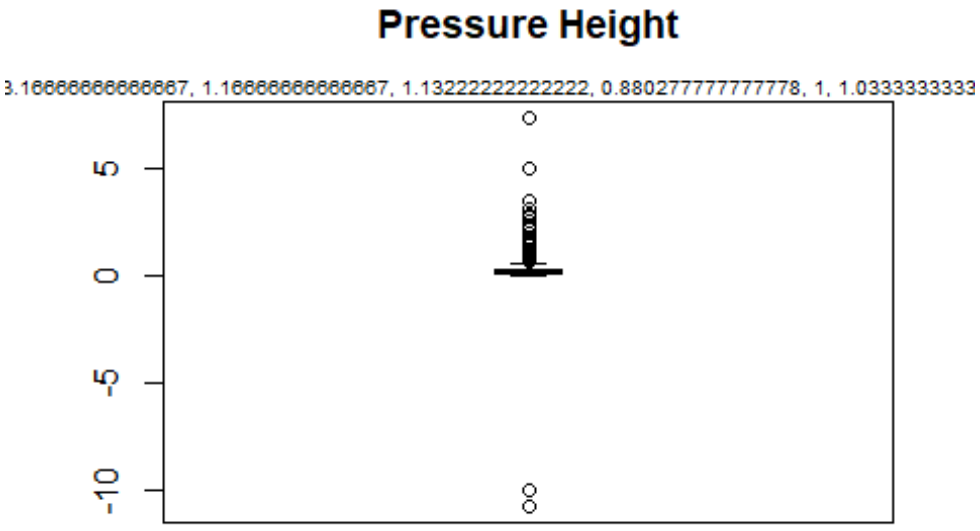
```
summary(data$check_time)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -10.71667   0.08333   0.16667   0.29957   0.29167   7.37778
```

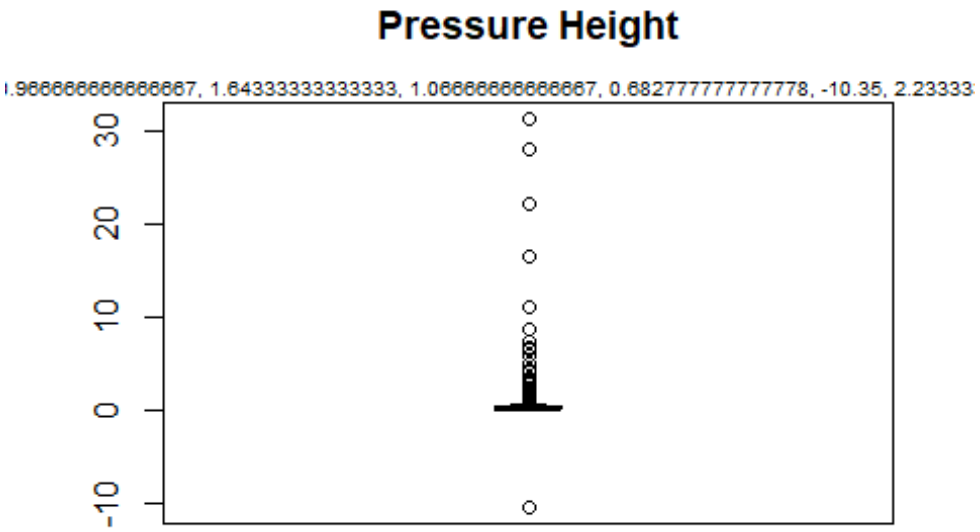
Outlier analyse.

De outliers worden opgespoord en in een vector gezet, deze wordt later gebruikt om die observaties te vervangen met het gemiddelde. Outliers package wordt hiervoor gebruikt.

Boxplot checking time:



Boxplot admin time:



Voor zowel checking time als admin time zien we outliers, deze worden vervangen door het gemiddele. Voor de code, zie toegevoegd r-bestand.

```
mean(df_final$check_time)
```

```
## [1] 0.2697075
```

```
mean(df_final$admin_time)
```

```
## [1] 0.4632331
```

Als laatste behandelen we de vraag of de KPI van 90% is behaald. Hiervoor berekenen we de totale throughput van de pakketjes (check_time + admin_time) en berekenen we de fractie van pakketjes dat binnen de tijd zijn behandeld.

```
## [1] 100
```

Alle pakketjes zijn op tijd en de KPI is dus behaald. Dit is wel berekend op basis van data waar de outliers zijn vervangen door het gemiddelde.

4. (2 points) Determine the utilisation (= fraction of time a worker is busy) of the express workers (between 07h00 and 18h00). Do the same for the admin workers. Om de fractie te berekenen dat een medewerker bezig is met een pakketje berekenen we eerst de totale werktijd. Dit is 6 dagen per week, 11 uur per dag. De periode over de hele dataset is van 3 oktober 2016 tot 30 november 2016. Dit zijn 59 werkdagen.

```
utilization of checker
```

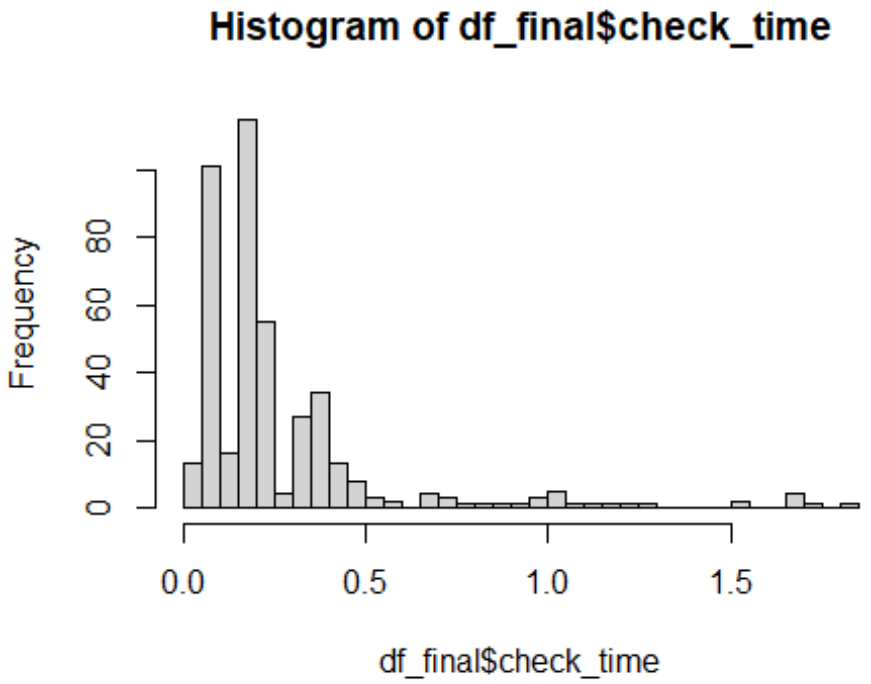
```
## [1] 0.1757878
```

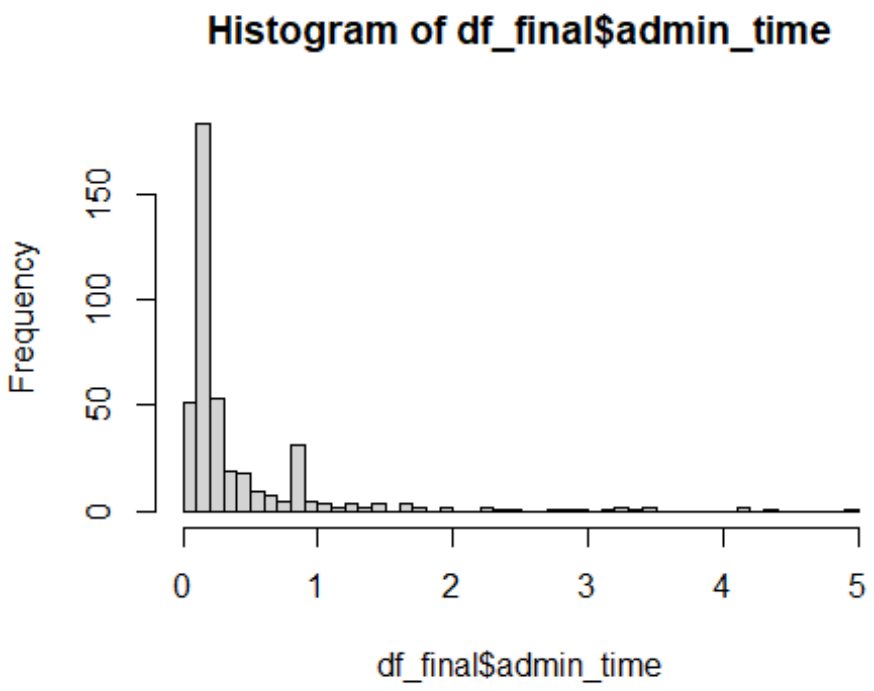
```
Utilization of admisitrator
```

```
## [1] 0.1509612
```

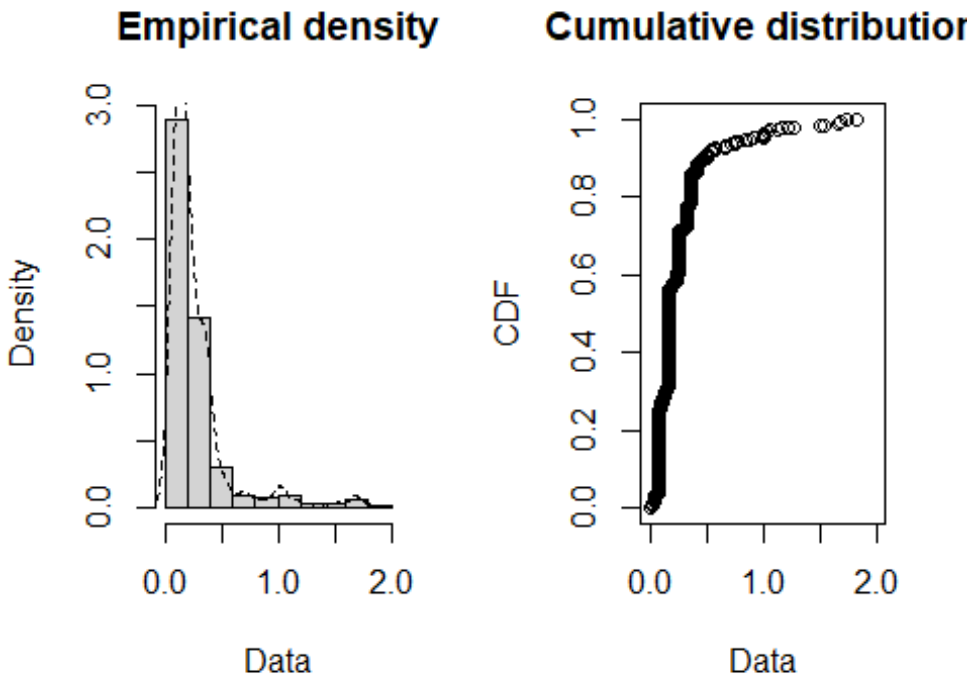
5. (3 points) Determine for each phase (checking and admin) the best fitting distribution (including the fitted parameters) and explain your choice.

Laten we eerst naar wat histogrammen kijken.

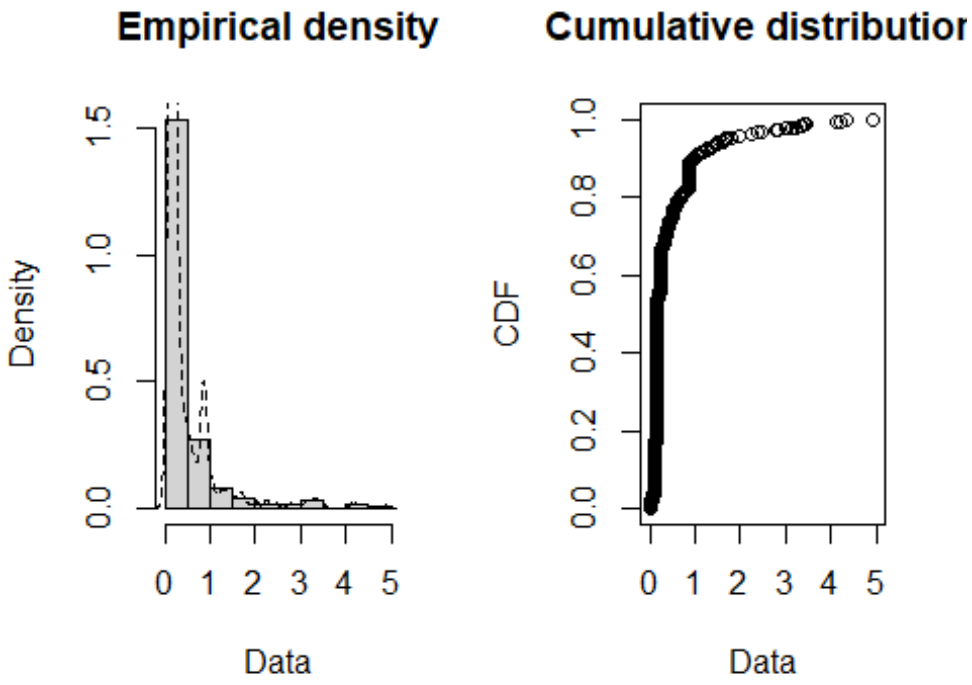




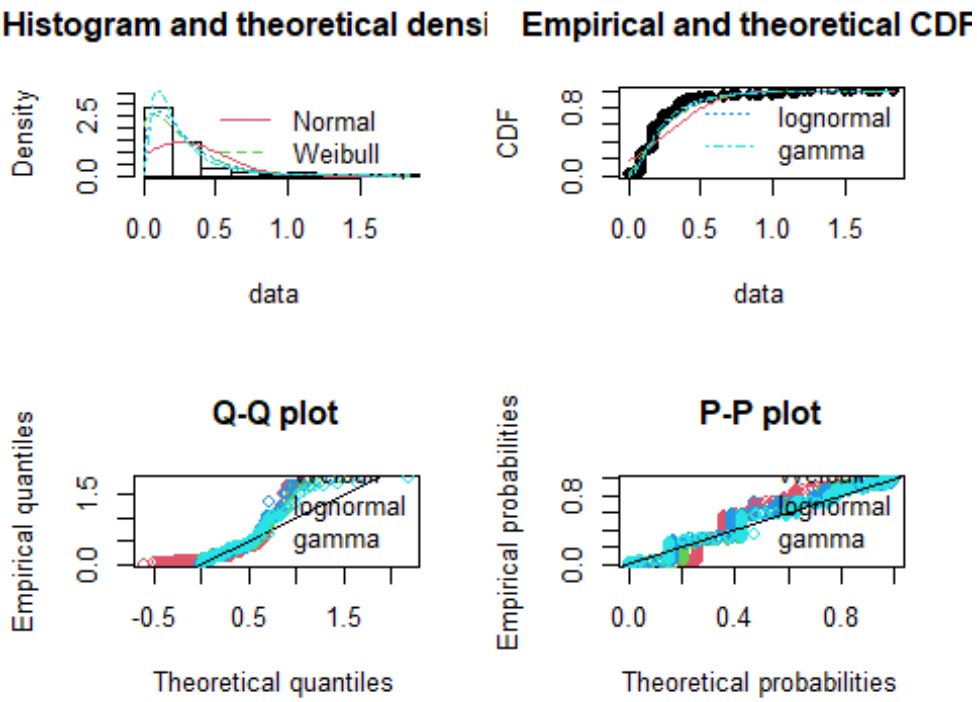

```
# Emperische verdeling & CDF
plotdist(df_final$check_time, histo = TRUE, demp=TRUE)
```



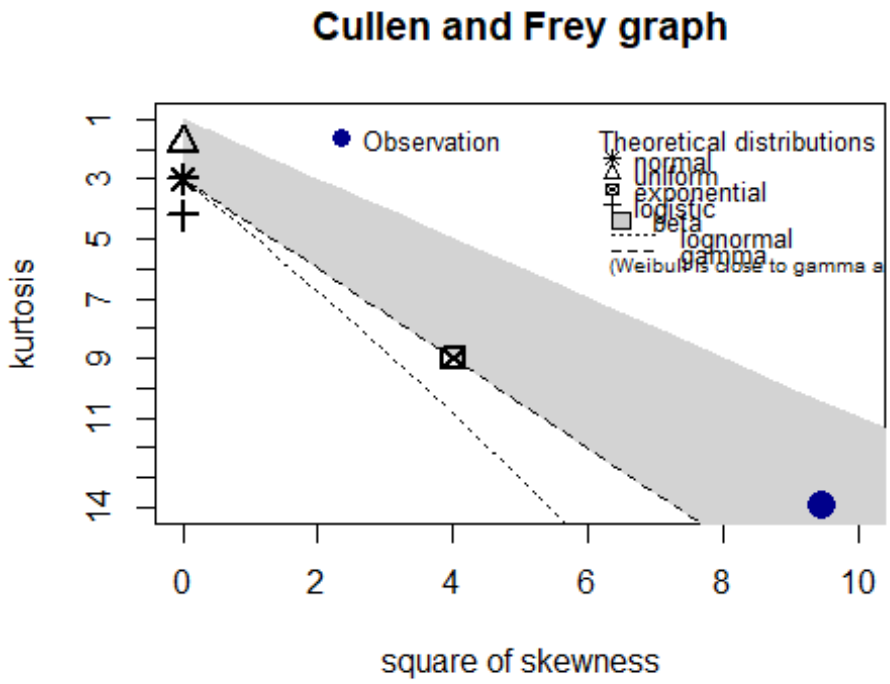

```
plotdist(df_final$admin_time, histo = TRUE, demp=TRUE)
```



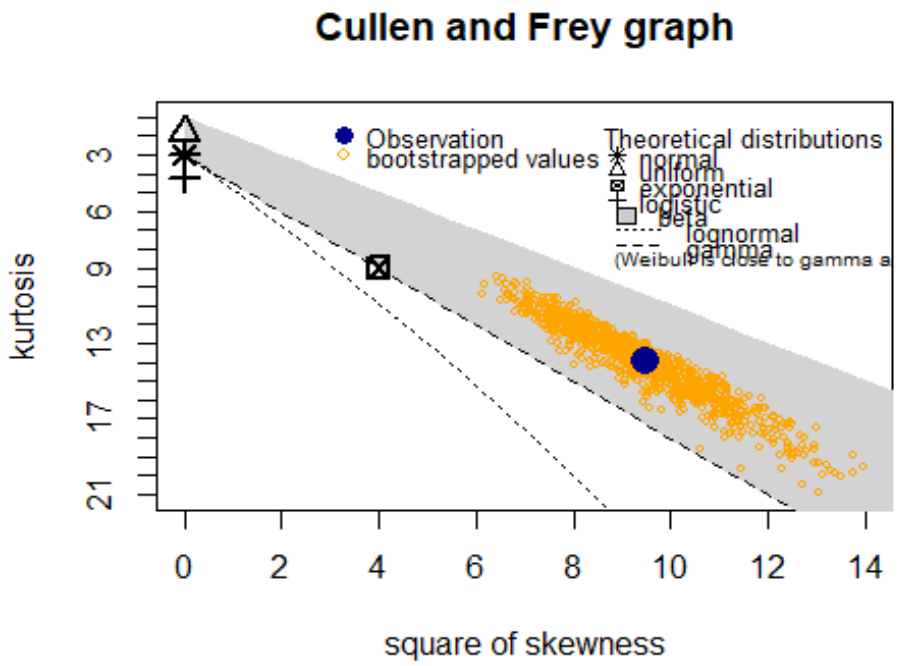
Kijkende naar deze histogrammen, de empirical density en de CDF kunnen we wel stellen dat beide activiteiten niet normaal zijn verdeeld. Laten we een aantal verdelingen proberen en op zoek gaan naar degene met de beste fit. We bekijken de normale, weibull, gamma en de lognormale verdelingen. We doen dit eerst voor checking time.



Aan de hand van deze grafieken kunnen we zien dat de normale verdeling geen hele goede fit heeft. Als we kijken naar de CDF lijken de Gamma, Weibull en Lognormale verdeling het beste te passen. We checken vervolgens de Cullen and Frey graphs voor checking time, wellicht dat we visueel kunnen afleiden wat de beste verdeling is.




```
## summary statistics
## ------
## min:  0.006666667   max:  1.833611
## median:  0.1666667
## mean:  0.2697075
## estimated sd:  0.2844704
## estimated skewness:  3.076956
## estimated kurtosis:  13.95203
```



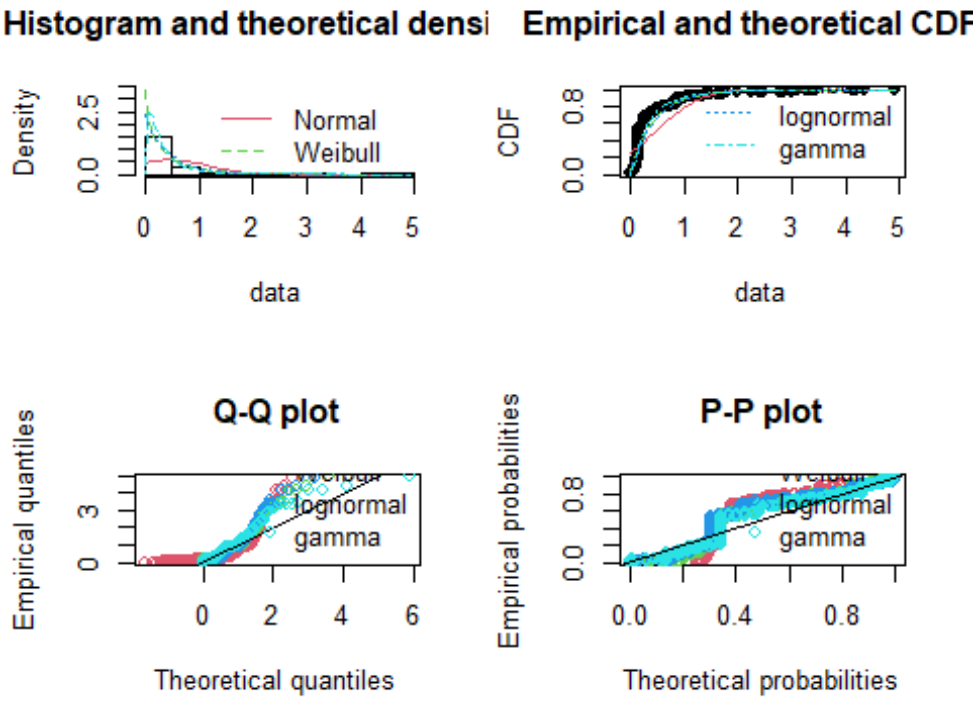

```
## summary statistics
## ------
## min:  0.006666667   max:  1.833611
## median:  0.1666667
## mean:  0.2697075
## estimated sd:  0.2844704
## estimated skewness:  3.076956
## estimated kurtosis:  13.95203
```

Om een definitieve keuze te maken over de verdeling kunnen we kijken naar de AIC (Akaike Information Criterion). De laagste waaarde heeft de beste fit.

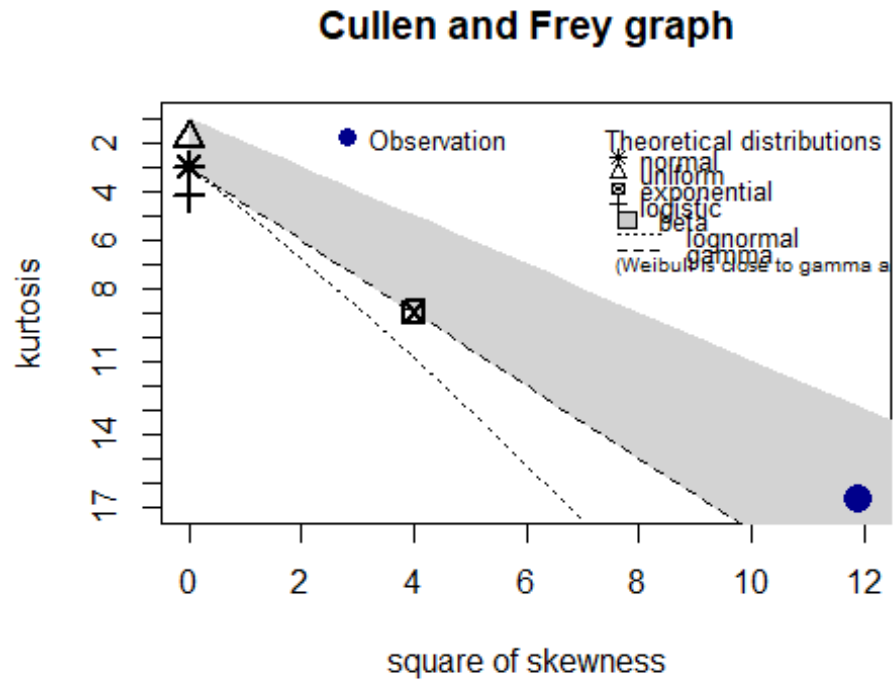
```
## [1] "AIC normal ="     "139.892024143688"

## [1] "AIC weibull ="     "-279.434164829018"

## [1] "AIC gamma ="       "-308.284623453416"

## [1] "AIC lnorm ="       "-383.281910908791"
```

De lognormale verdeling heeft voor checking time de laagste AIC en dus de beste fit. Deze zullen we in het volgende onderdeel gebruiken.

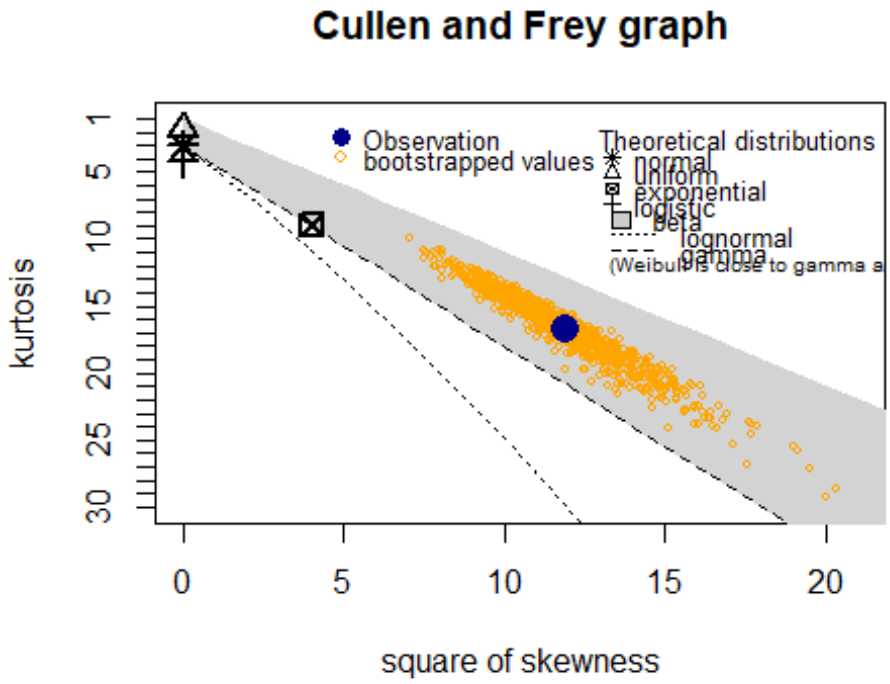
In dit onderdeel doen we hetzelfde voor admin time. We gebruiken wederom dezelfde verdelingen als hiervoor.



Wederom zien we dat de normale verdeling niet geschikt is voor deze data. De Gamma, Weibull en lognormale verdeling komen beter in de buurt. Laten we kijken naar de Cullen en Frey graphs.



```
## summary statistics
## ------
## min:  0.002777778   max:  4.905278
## median:  0.1666667
## mean:  0.4632331
## estimated sd:  0.6909581
## estimated skewness:  3.44835
## estimated kurtosis:  16.6898
```




```
## summary statistics
## ------
## min:  0.002777778   max:  4.905278
## median:  0.1666667
## mean:  0.4632331
## estimated sd:  0.6909581
## estimated skewness:  3.44835
## estimated kurtosis:  16.6898
```


Voor checking time is de Cullen and Frey graph lastiger te interpreteren. Laten we dus wederom een blik werpen op de verschillende AIC waarden en op basis daarvan een keuze maken.

```
## [1] "AIC normal ="     "890.674878111596"

## [1] "AIC weibull ="    "188.612087600184"

## [1] "AIC gamma ="      "198.723886635092"
```

```
## [1] "AIC lnorm ="    "75.09700750941"
```

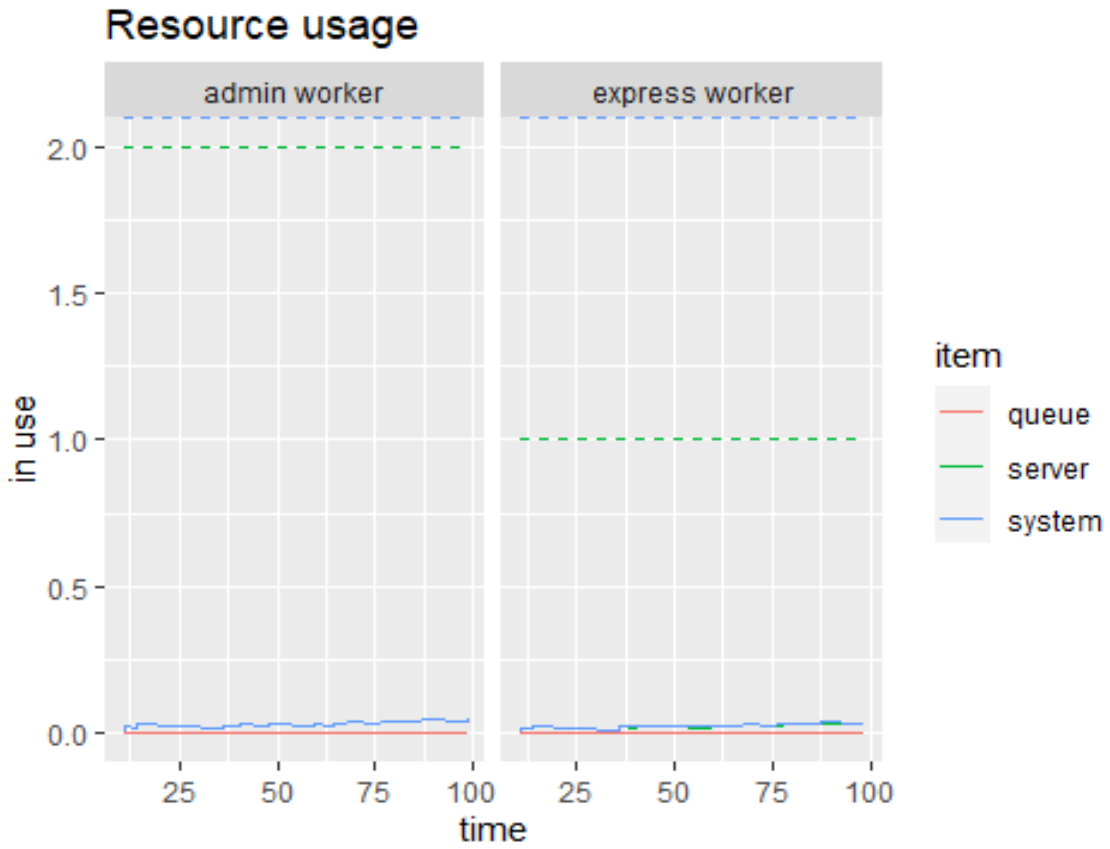
De lognormale verdeling heeft de laagste AIC en dus de beste fit. Dit zullen we gebruiken in de volgende vraag waarin we de simulatie gaan doen. De parameters zijn als volgt:

```
fit_lnn

## Fitting of the distribution ' lnorm ' by maximum likelihood
## Parameters:
##          estimate Std. Error
## meanlog -1.366801 0.05019776
## sdlog    1.032416 0.03549503

fit_ln

## Fitting of the distribution ' lnorm ' by maximum likelihood
## Parameters:
##           estimate Std. Error
## meanlog -1.6535597 0.03889674
## sdlog    0.7999877 0.02750395
```

6. Replace the statistical distributions in the simmer script with the fitted distributions from the previous question. For the arrivals use the exact 'Eind Lossen' time stamps. Run the simulation for 10 working days, and repeat 100 times. Recompute the performance measures from question 1.

```
Hieronder wat beschrijvende informatie over deze run.

##    resource              time           server           queue
##  Length:68          Min.   :10.75   Min.   :0.0000   Min.   :0.00000
##  Class :character   1st Qu.:40.39   1st Qu.:0.0000   1st Qu.:0.00000
##  Mode  :character   Median :67.34   Median :1.0000   Median :0.00000
##                     Mean   :60.84   Mean   :0.5735   Mean   :0.01471
##                     3rd Qu.:78.62   3rd Qu.:1.0000   3rd Qu.:0.00000
##                     Max.   :98.49   Max.   :2.0000   Max.   :1.00000
##     capacity     queue_size      system           limit      replication
##  Min.   :1.0   Min.   :Inf   Min.   :0.0000   Min.   :Inf   Min.   :1
##  1st Qu.:1.0   1st Qu.:Inf   1st Qu.:0.0000   1st Qu.:Inf   1st Qu.:1
##  Median :1.5   Median :Inf   Median :1.0000   Median :Inf   Median :1
##  Mean   :1.5   Mean   :Inf   Mean   :0.5882   Mean   :Inf   Mean   :1
##  3rd Qu.:2.0   3rd Qu.:Inf   3rd Qu.:1.0000   3rd Qu.:Inf   3rd Qu.:1
##  Max.   :2.0   Max.   :Inf   Max.   :2.0000   Max.   :Inf   Max.   :1
```

```
Hieronder een plot van het gebruik van resources (administrator en checker).
```



Het is me helaas niet gelukt om met echte datums te werken voor de tijdsintervallen, zou hier graag feedback op willen ontvangen. Online kon ik moeilijk documentatie vinden voor het simmer package. Desalniettemin, gaan we wat performance metrics uit de vorige vragen berekenen. Laten we eerst kijken naar de checker.

```
## [1] "Mean activity time =" "0.191341860888656"

## [1] "SD activity time =" "0.0071714689874333"

## [1] "Mean wait time ="    "0.00526238685685938"

## [1] "SD wait time ="     "0.0216973768536922"
```

Nu kijken we naar administratie werkers.

```
## [1] "Mean activity time =" "0.258730400293704"

## [1] "SD activity time =" "0.0117450484660652"

## [1] "Mean wait time ="    "0.00526238685685938"

## [1] "SD wait time ="     "0.0216973768536922"
```

Als laatste nog even naar de totale throughput van de pakketjes kijken.

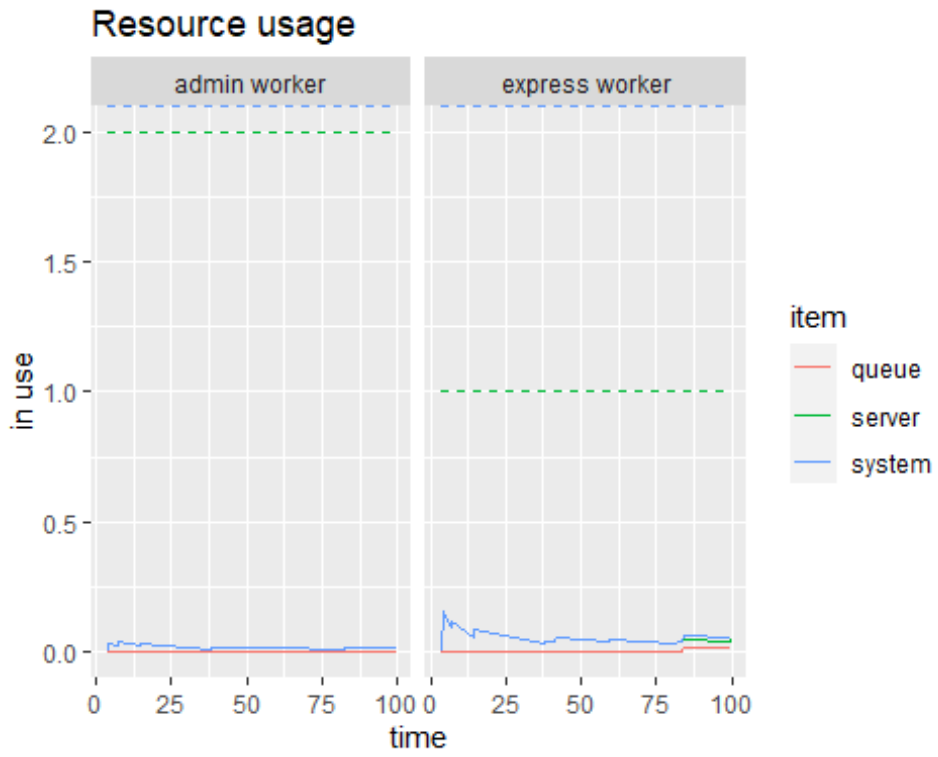
```
## [1] "Mean total throughput =" "0.455334648039219"

## [1] "SD total throughput  =" "0.022421323039129"
```

We zien dat de gemiddelde wachttijd van een checker zeer klein is. Dit komt waarschijnlijk doordat we niet de echte arrival datums van een pakketje gebruiken, maar een fictieve. Ik ben benieuwd wat de uitkomsten zullen zijn als ik dit kan toepassen!

7. Now replace the statistical distributions in the simmer script with the empirical distributions. How do the different simulations compare?

Hieronder wat beschrijvende informatie over deze run.

```
##    resource              time            server           queue        
##  Length:44          Min.   : 3.466   Min.   :0.0000   Min.   :0.00000  
##  Class :character   1st Qu.:14.759   1st Qu.:0.0000   1st Qu.:0.00000  
##  Mode  :character   Median :59.627   Median :1.0000   Median :0.00000  
##                     Mean   :54.534   Mean   :0.5909   Mean   :0.09091  
##                     3rd Qu.:84.359   3rd Qu.:1.0000   3rd Qu.:0.00000  
##                     Max.   :99.879   Max.   :2.0000   Max.   :2.00000  
##     capacity     queue_size      system           limit      replication
##  Min.   :1.0   Min.   :Inf   Min.   :0.0000   Min.   :Inf   Min.   :1  
##  1st Qu.:1.0   1st Qu.:Inf   1st Qu.:0.0000   1st Qu.:Inf   1st Qu.:1  
##  Median :1.5   Median :Inf   Median :1.0000   Median :Inf   Median :1  
##  Mean   :1.5   Mean   :Inf   Mean   :0.6818   Mean   :Inf   Mean   :1  
##  3rd Qu.:2.0   3rd Qu.:Inf   3rd Qu.:1.0000   3rd Qu.:Inf   3rd Qu.:1  
##  Max.   :2.0   Max.   :Inf   Max.   :3.0000   Max.   :Inf   Max.   :1
```




```
## [1] "Mean activity time =" "0.398887888645582"

## [1] "SD activity time =" "0.225498190107898"

## [1] "Mean wait time ="  "0.125145547930934"

## [1] "SD wait time ="    "0.278571358056705"
```

Nu kijken we naar administratie werkers.

```
## [1] "Mean activity time =" "0.149555862119045"

## [1] "SD activity time =" "0"

## [1] "Mean wait time ="  "0.125145547930934"

## [1] "SD wait time ="    "0.278571358056705"
```

Als laatste nog even naar de totale throughput van de pakketjes kijken.

```
## [1] "Mean total throughput =" "0.67358929869556"

## [1] "SD total throughput  =" "0.243695923108491"
```