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?

```
mean_check_time
```

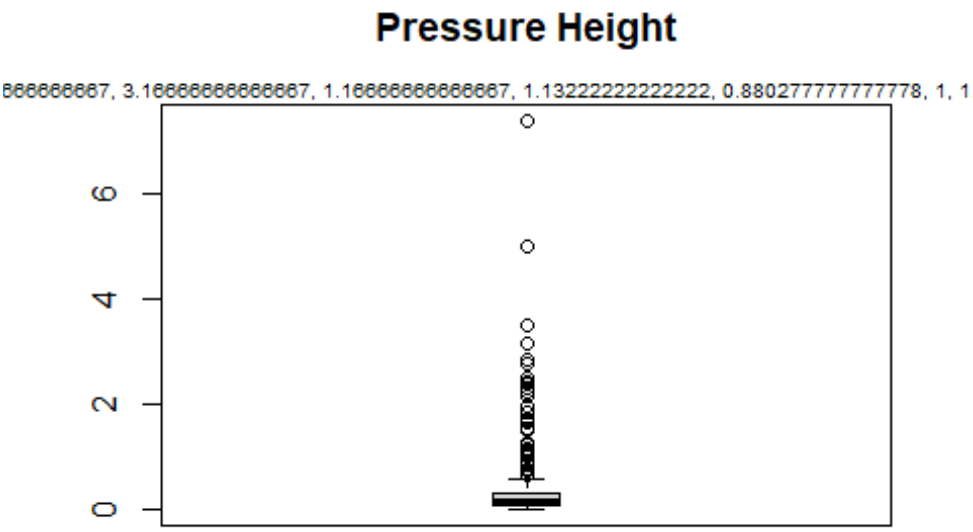
```
0.3584562
```

```
mean_admin_time
```

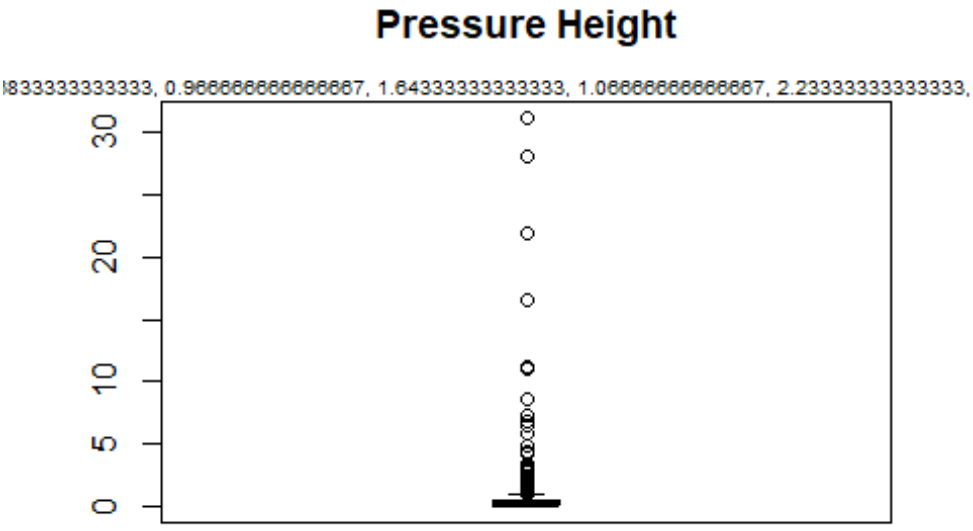
```
0.855021
```

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

Boxplot checking time:



Boxplot admin time:



Voor zowel checking time als admin time zien we outliers, deze worden vervangen door het gemiddele in het volgende code:

```
# In dit onderdeel doen we wat data manipulatie, de waarden die kleiner of ge
lijk aan nul zijn worden vervangen door gemiddelden. Hetzelfde geldt voor de
outliers.
df_final <- data %>%
  mutate(check_time = ifelse(check_time_outliers,mean_check_time, check_time)
,
         admin_time = ifelse(admin_time_outliers,mean_admin_time, admin_time)
) %>%
  mutate(check_time    = ifelse(check_time <= 0, mean_check_time, check_time)
,
         admin_time    = ifelse(admin_time <= 0, mean_admin_time, admin_time)
,
         total_throughput = check_time + admin_time)


mean(df_final$check_time)

## [1] 0.2429038

mean(df_final$admin_time)

## [1] 0.4607978
```

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
```

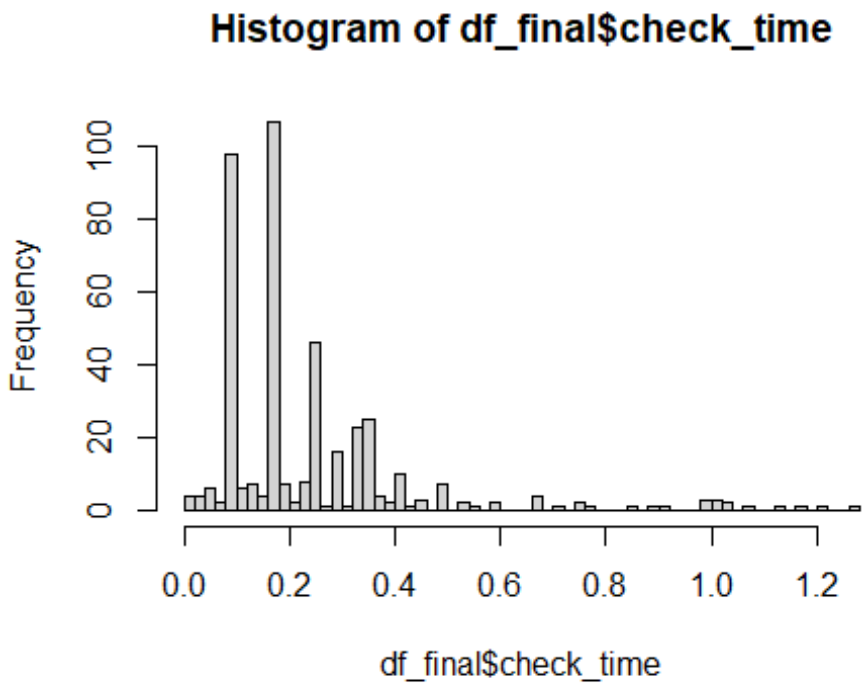
Alle pakketjes zijn op tijd en de KPI is dus behaald. Dit is wel berekend op basis van data waar de outliers zijn vervangen door het gemiddelde.

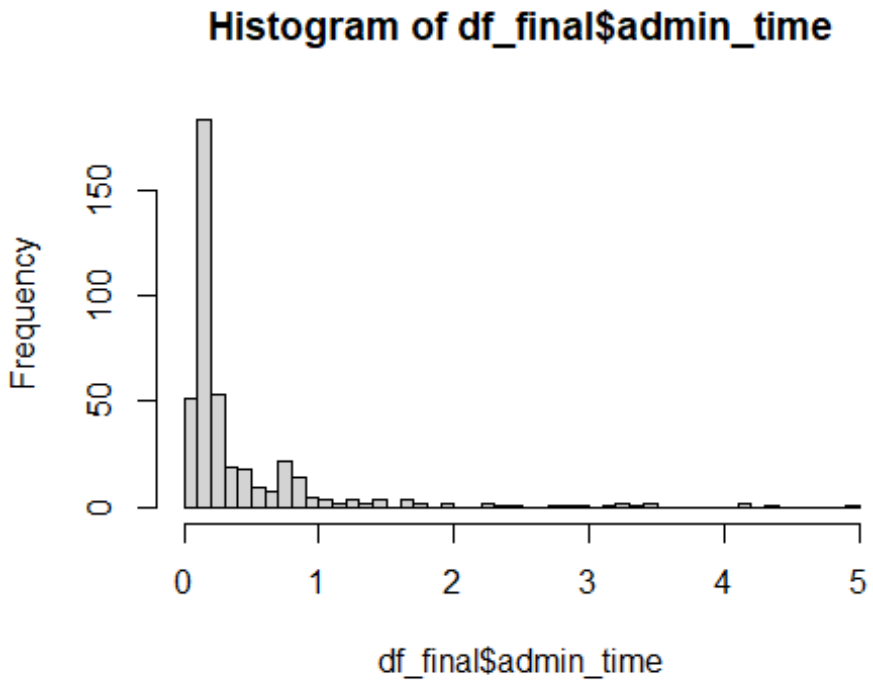
4. (2 points) Determine the utilisation (= fraction of time a worker is busy) of the express workers (between 07h00 and 18h00). Do the same for the admin workers. Om de fractie te berekenen dat een medewerker bezig is met een pakketje berekenen we eerst de totale werktijd. Dit is 6 dagen per week, 11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

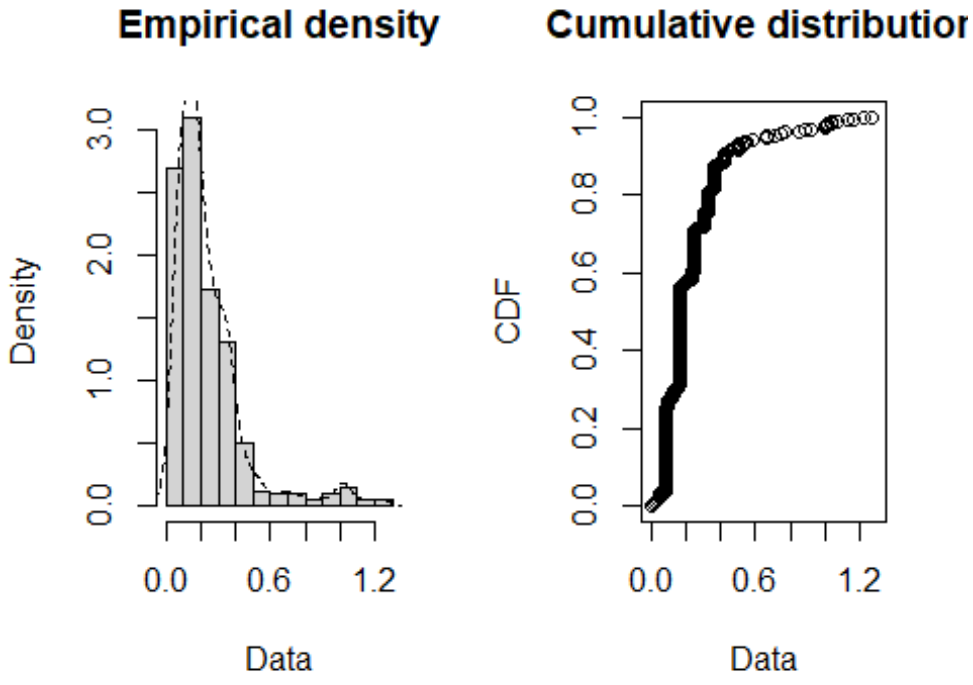
```
utilization of checker

0.1583179

Utilization of administrator

0.1501675
```

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

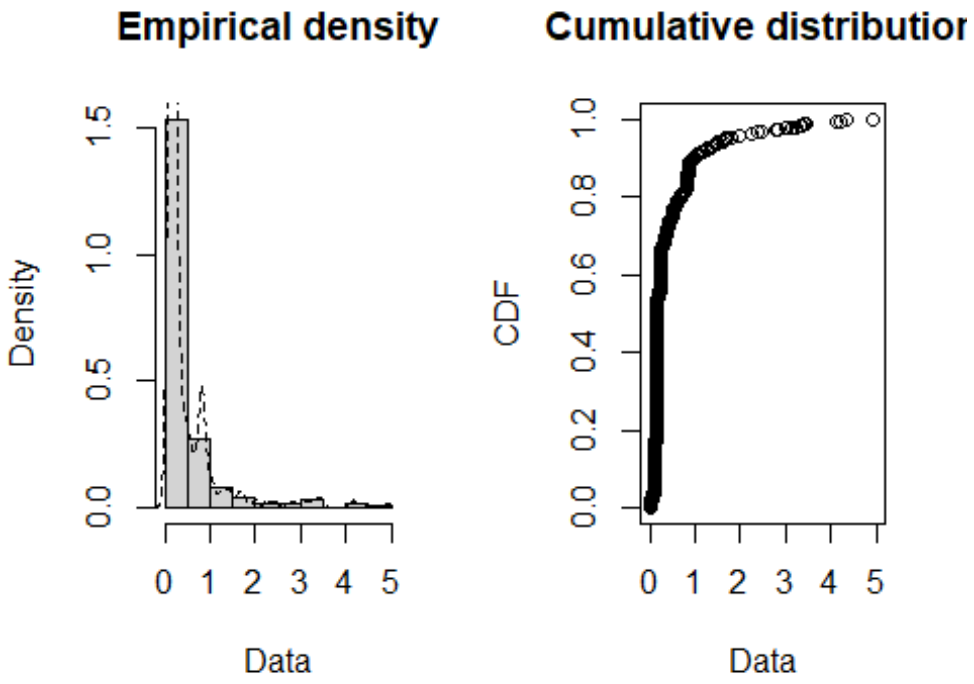
Laten we eerst naar wat histogrammen kijken.





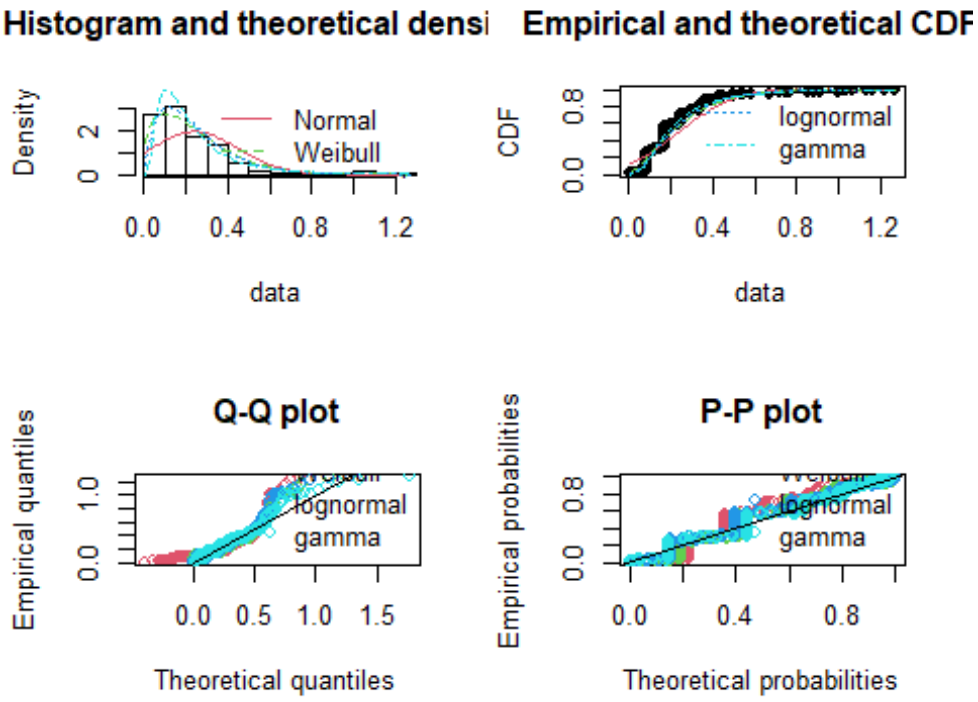



```
plotdist(df_final$admin_time, histo = TRUE, demp=TRUE)
```

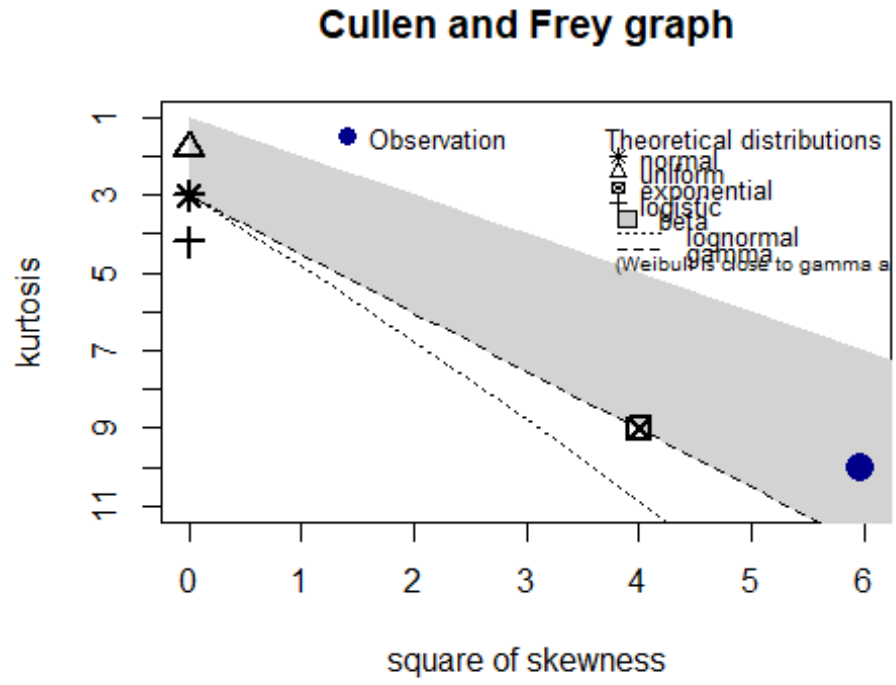


Kijkende naar deze histogrammen, de empirical density en de CDF kunnen we wel stellen dat beide activiteiten niet normaal zijn verdeeld. Laten we een aantal verdelingen proberen en op zoek gaan naar degene met de beste fit. We bekijken de normale, weibull, gamma en de lognormale verdelingen. We doen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

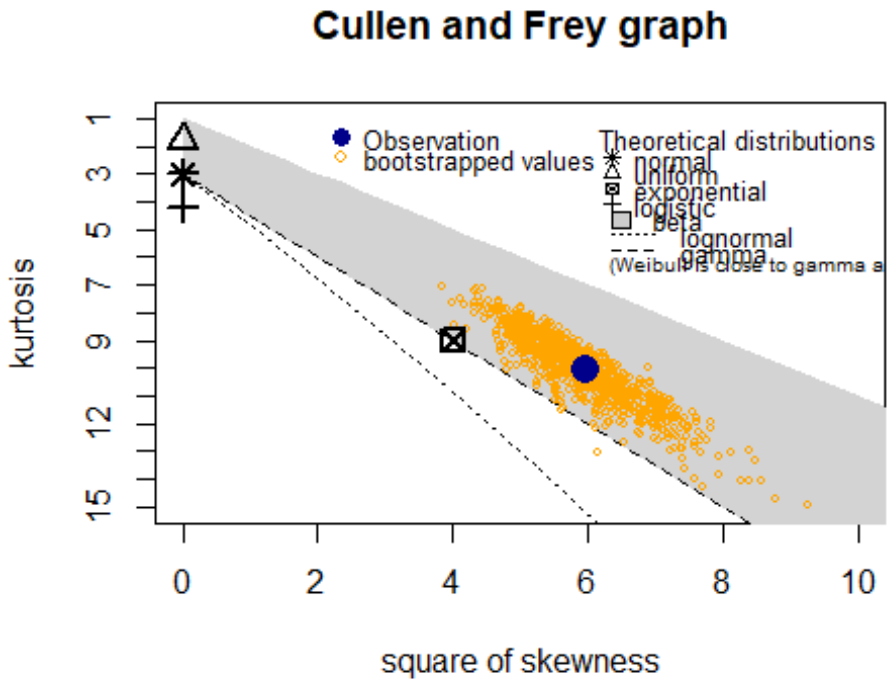
```
# Fit some other distributions
fit_n  <-  fitdist(df_final$check_time, "norm")
fit_w  <- fitdist(df_final$check_time,  "weibull")
fit_g  <- fitdist(df_final$check_time,  "gamma")
fit_ln <- fitdist(df_final$check_time,  "lnorm")

plot.legend <- c("Normal", "Weibull", "lognormal", "gamma")
par(mfrow=c(2,2))
denscomp(list(fit_n, fit_w, fit_g, fit_ln), legendtext = plot.legend)
cdfcomp (list(fit_n, fit_w, fit_g, fit_ln), legendtext = plot.legend)
qqcomp  (list(fit_n, fit_w, fit_g, fit_ln), legendtext = plot.legend)
ppcomp  (list(fit_n, fit_w, fit_g, fit_ln), legendtext = plot.legend)
```



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:  1.266667
## median:  0.1666667
## mean:  0.2429038
## estimated sd:  0.2073224
## estimated skewness:  2.44264
## estimated kurtosis:  10.03937
```



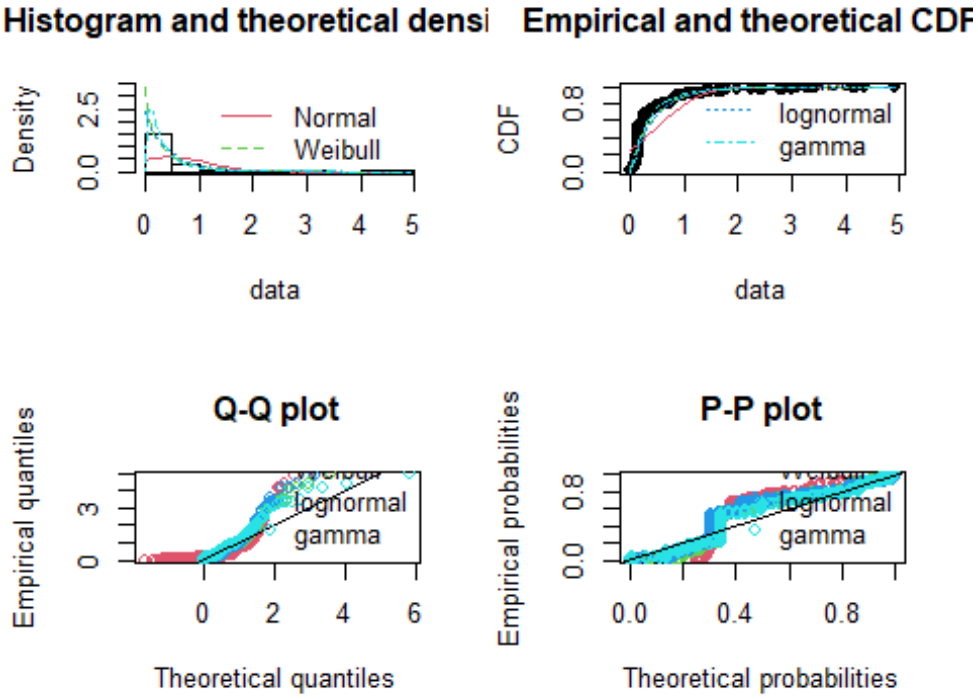

```
## summary statistics
## ------
## min:  0.006666667   max:  1.266667
## median:  0.1666667
## mean:  0.2429038
## estimated sd:  0.2073224
## estimated skewness:  2.44264
## estimated kurtosis:  10.03937
```


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.

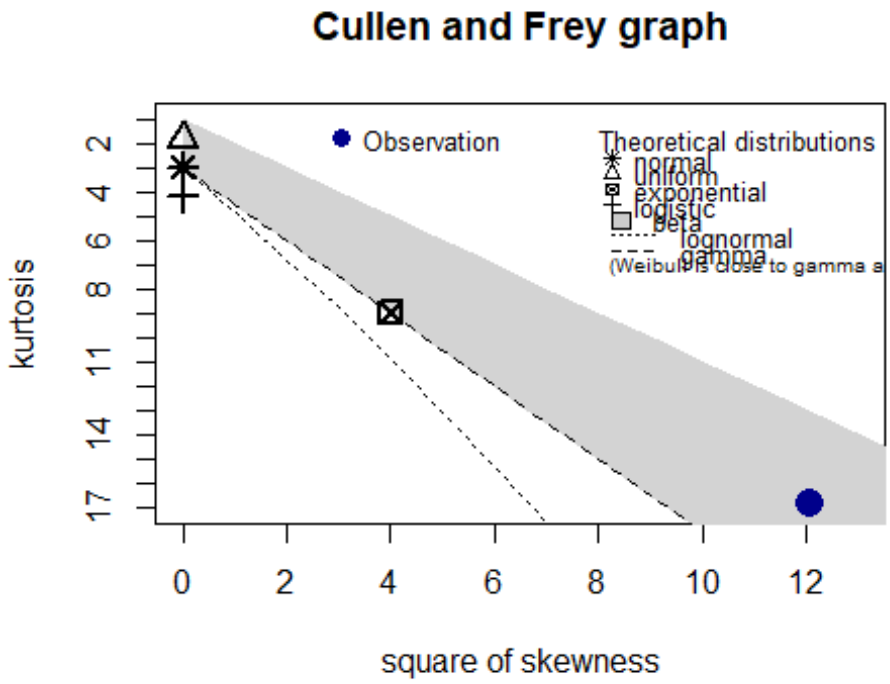
```
"AIC normal ="      "-127.743494529332"

"AIC weibull ="   "-409.5362485545"

"AIC gamma ="       "-442.369612630227"

"AIC lnorm ="       "-476.608877716932"
```

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

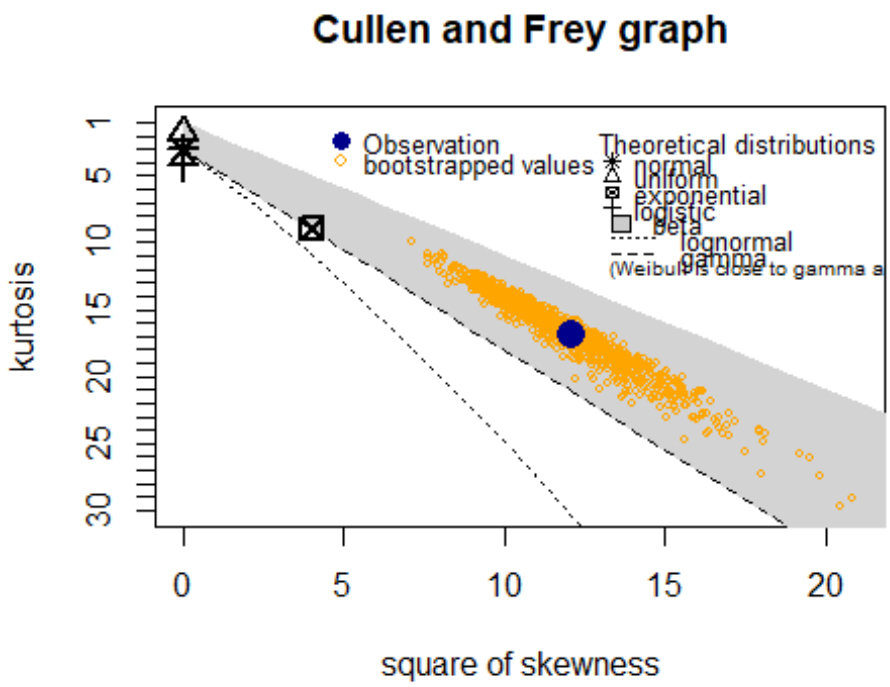
In dit onderdeel doen we hetzelfde voor admin time. We gebruiken wederom dezelfde verdelingen als hiervoor.



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:  0.4607978
## estimated sd:  0.6896614
## estimated skewness:  3.475479
## estimated kurtosis:  16.86317
```




```
## summary statistics
## ------
## min:  0.002777778   max:  4.905278
## median:  0.1666667
## mean:  0.4607978
## estimated sd:  0.6896614
## estimated skewness:  3.475479
## estimated kurtosis:  16.86317
```

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.

```
"AIC normal ="     "889.085696707452"

"AIC weibull ="    "184.350780302785"

"AIC gamma ="      "194.320463282346"

"AIC lnorm ="      "69.8478100541442"
```

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 -1.369741 0.05003415
## sdlog    1.029051 0.03537934
```

```
fit_ln
```

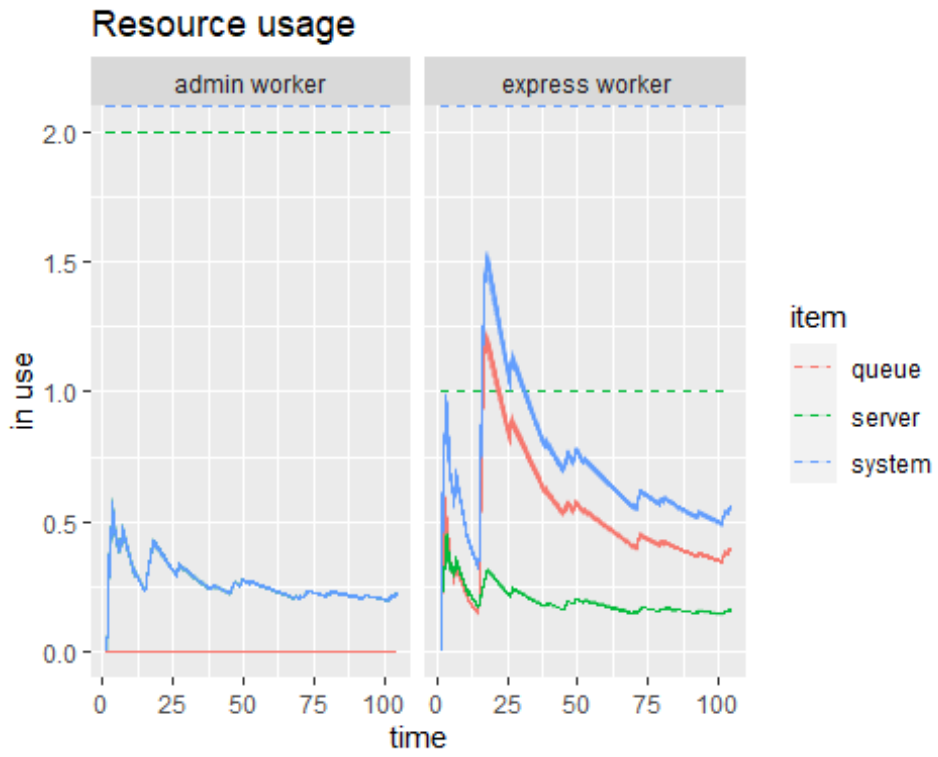
```
## Fitting of the distribution ' lnorm ' by maximum likelihood
## Parameters:
##           estimate Std. Error
## meanlog -1.6888910 0.03608676
## sdlog    0.7421951 0.02551699
```

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.

```
summary(get_mon_resources(env))
```

```
##    resource              time             server          queue
##  Length:36800       Min.   :  1.317   Min.   :0.000   Min.   : 0.000
##  Class :character   1st Qu.: 16.244   1st Qu.:1.000   1st Qu.: 0.000
##  Mode  :character   Median : 46.036   Median :1.000   Median : 0.000
##                     Mean   : 49.132   Mean   :1.047   Mean   : 1.053
##                     3rd Qu.: 79.443   3rd Qu.:1.000   3rd Qu.: 1.000
##                     Max.   :104.638   Max.   :2.000   Max.   :13.000
##     capacity     queue_size      system         limit      replication
##  Min.   :1.0   Min.   :Inf   Min.   : 0.0   Min.   :Inf   Min.   :  1.00
##  1st Qu.:1.0   1st Qu.:Inf   1st Qu.: 1.0   1st Qu.:Inf   1st Qu.: 25.75
##  Median :1.5   Median :Inf   Median : 1.0   Median :Inf   Median : 50.50
##  Mean   :1.5   Mean   :Inf   Mean   : 2.1   Mean   :Inf   Mean   : 50.50
##  3rd Qu.:2.0   3rd Qu.:Inf   3rd Qu.: 2.0   3rd Qu.:Inf   3rd Qu.: 75.25
##  Max.   :2.0   Max.   :Inf   Max.   :14.0   Max.   :Inf   Max.   :100.00
```

```
plot(get_mon_resources(env))
```



Tweede poging is het gelukt, met wat hulp van medestudenten. Datums worden verwerkt in een apart dataframe en vervolgens met behulp van 'add_dataframe' toegevoegd aan de simulatie. Laten we naar wat performance metrieken gaan kijken.

Checker:

```
"Mean activity time =" "0.184756579692523"

"SD activity time ="  "0.00668082464741411"

"Mean wait time ="  "0.444733214930131"

"SD wait time ="    "0.520994190809467"
```

Nu kijken we naar **administratie werkers.**

```
"Mean activity time =" "0.254448346825642"

"SD activity time =" "0.01292963321176"

"Mean wait time ="  "0.444733214930131"

"SD wait time ="    "0.520994190809467"
```

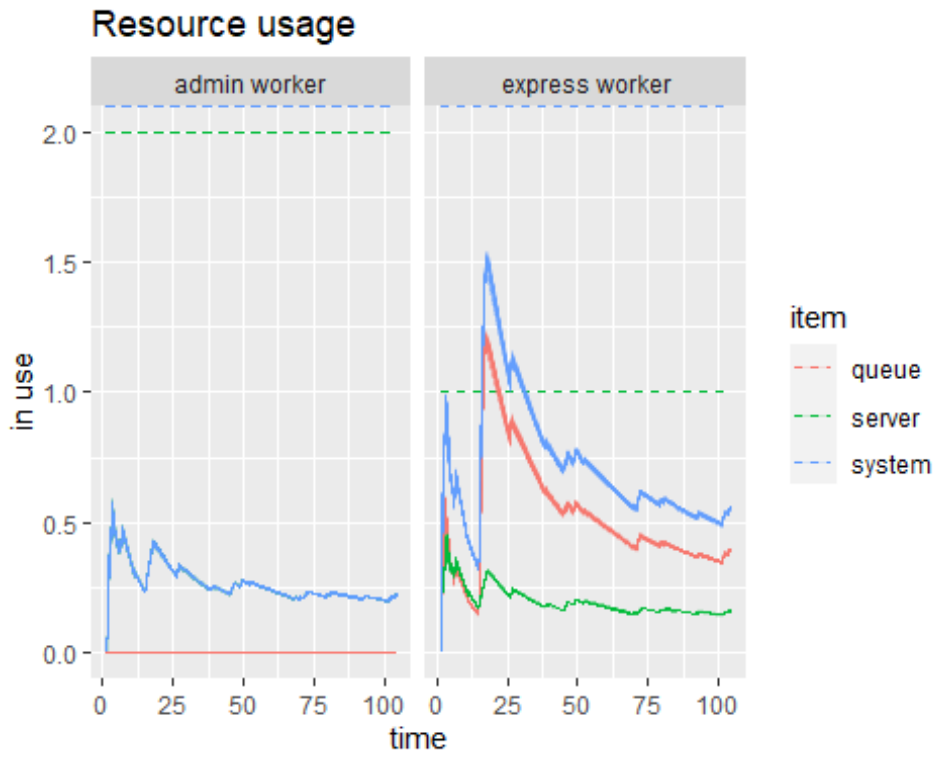
Als laatste nog even naar de totale throughput van de pakketjes kijken.

```
"Mean total throughput =" "0.929470152510589"

"SD total throughput  =" "0.531889826044731"
```

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

```
summary(get_mon_resources(env_emp))
```

```
##    resource              time             server          queue
##  Length:36800       Min.   :  1.317   Min.   :0.000   Min.   : 0.000
##  Class :character   1st Qu.: 16.244   1st Qu.:1.000   1st Qu.: 0.000
##  Mode  :character   Median : 46.036   Median :1.000   Median : 0.000
##                     Mean   : 49.132   Mean   :1.047   Mean   : 1.053
##                     3rd Qu.: 79.443   3rd Qu.:1.000   3rd Qu.: 1.000
##                     Max.   :104.638   Max.   :2.000   Max.   :13.000
##     capacity     queue_size      system         limit      replication
##  Min.   :1.0   Min.   :Inf   Min.   : 0.0   Min.   :Inf   Min.   :  1.00
##  1st Qu.:1.0   1st Qu.:Inf   1st Qu.: 1.0   1st Qu.:Inf   1st Qu.: 25.75
##  Median :1.5   Median :Inf   Median : 1.0   Median :Inf   Median : 50.50
##  Mean   :1.5   Mean   :Inf   Mean   : 2.1   Mean   :Inf   Mean   : 50.50
##  3rd Qu.:2.0   3rd Qu.:Inf   3rd Qu.: 2.0   3rd Qu.:Inf   3rd Qu.: 75.25
##  Max.   :2.0   Max.   :Inf   Max.   :14.0   Max.   :Inf   Max.   :100.00
```



Checker:

```
"Mean activity time =" "0.184756579692523"
"SD activity time ="  "0.00668082464741411"
"Mean wait time ="  "0.444733214930131"
"SD wait time ="    "0.520994190809467"
```

Nu kijken we naar **administratie werkers.**

```
"Mean activity time =" "0.254448346825642"
"SD activity time =" "0.01292963321176"
"Mean wait time ="  "0.444733214930131"
"SD wait time ="    "0.520994190809467"
```

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

```
"Mean total throughput =" "0.929470152510589"
"SD total throughput  =" "0.531889826044731"
```