SYST 664: HW Assignment 3

Jake Whalen

February 12, 2018

Contents

[Problem 1 1](#_Toc506125109)

[Part A 1](#_Toc506125110)

[Part B 2](#_Toc506125111)

[Part C 2](#_Toc506125112)

[Part D 3](#_Toc506125113)

[Part E 3](#_Toc506125114)

[Problem 2 3](#_Toc506125115)

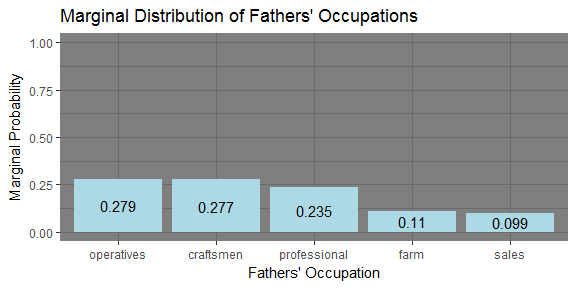
[Part A 3](#_Toc506125116)

[Part B 6](#_Toc506125117)

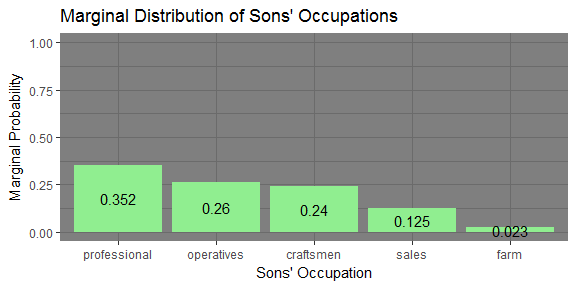
[Part C 8](#_Toc506125118)

# Problem 1

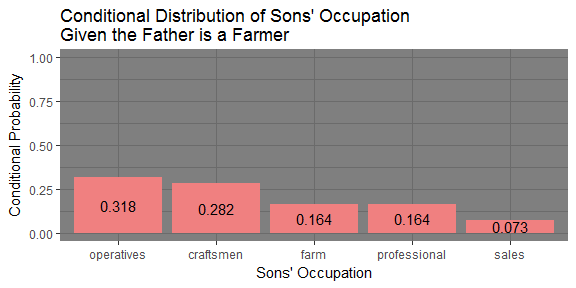
### Part A



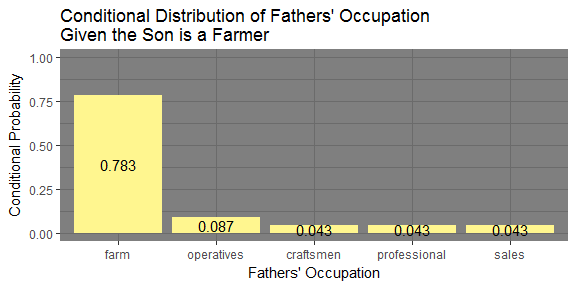
### Part B



### Part C



### Part D



### Part E

##### What do the results in parts a-d say about changes in farming in the population from which these data are drawn?

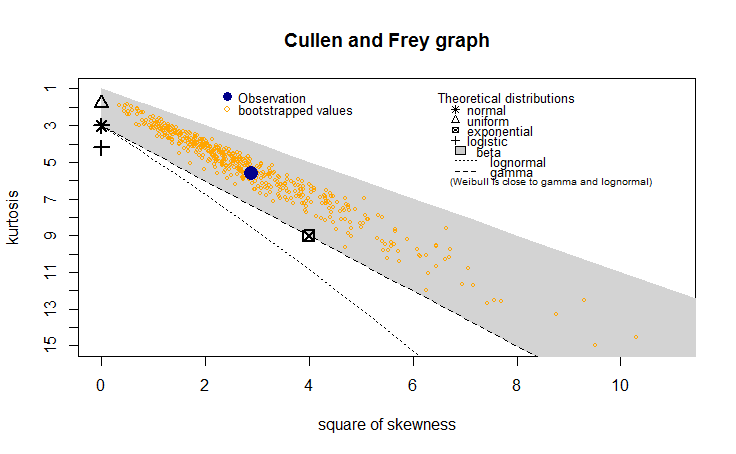
Marginal probability of farming being an individuals occupation drops from fathers (0.11) to sons (0.023). If the father was a farmer then the son has a 16.4% probability of also taking up the farming occupation. If the son was a farmer then the father was a farmer with a 78.3% probability. Thus while farming may not be passed down at a high rate between these generations it also is not an occupation that the younger generation is flocking to. If the son is a farmer it is likely their father was a farmer.

# Problem 2

### Part A

##### Do you think an exponential distribution provides a good model for the interarrival times?

First I wanted to explore a variety of distributions to see if any fit the data better then an exponential. The chart below shows the kurtosis and squared skewness of the inter-arrival sample plotted as a blue point (“Observation”). I also conducted bootstrap sampling using 500 observations to see a range of possible observations from the data.



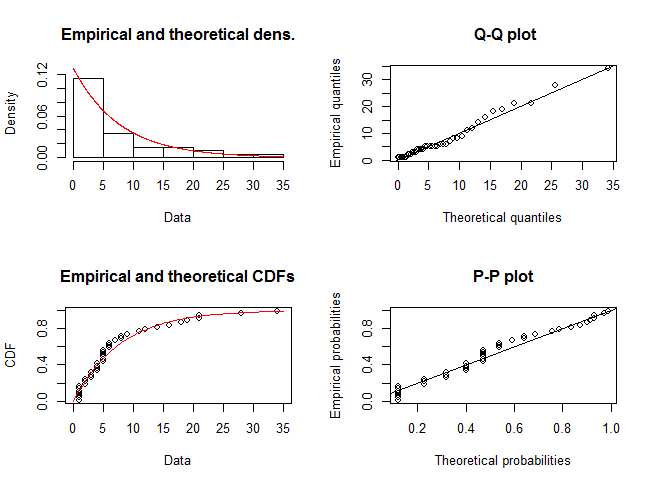
Distribution Exploration

From the graph the possible distributions include the Gamma, and Exponential. The beta distribution was ruled out because the data was not between 0 and 1.

First I fit an exponential distribution to the data.

fit.exp <- fitdist(interArrival, "exp")

The plot below helps to visualize the fit for the exponential distribution.

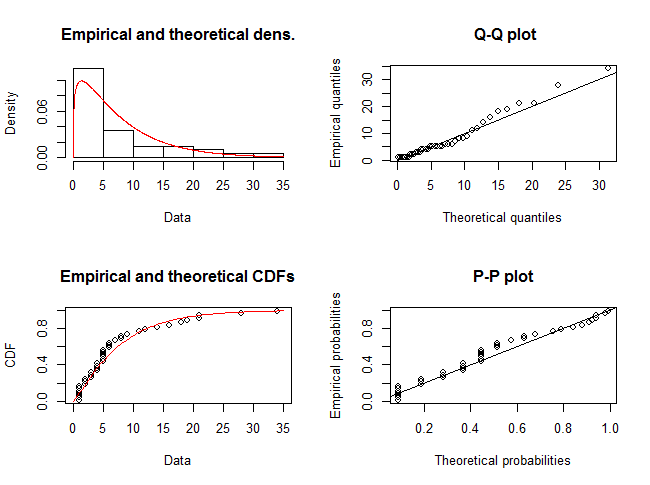


Exponential Distribution

Next I fit a gamma distribution to the data.

fit.gam <- fitdist(interArrival, "gamma")

The plot below helps to visualize the fit for the gamma distribution.



Gamma Distribution

Both distributions look good when fitted on the data and plotted above. The QQ-Plot for the exponential looks slightly better. Not enough to make a claim however that the exponential fits better then Gamma. Also by inspecting the fits AIC and BIC the exponentials’ are lower then the Gammas’.

summary(fit.exp)

## Fitting of the distribution ' exp ' by maximum likelihood   
## Parameters :   
## estimate Std. Error  
## rate 0.1282051 0.02026978  
## Loglikelihood: -122.1649 AIC: 246.3299 BIC: 248.0188

summary(fit.gam)

## Fitting of the distribution ' gamma ' by maximum likelihood   
## Parameters :   
## estimate Std. Error  
## shape 1.201064 0.2402127  
## rate 0.153967 0.0379670  
## Loglikelihood: -121.7653 AIC: 247.5306 BIC: 250.9083   
## Correlation matrix:  
## shape rate  
## shape 1.000 0.811  
## rate 0.811 1.000

After inspecting the plots above and summary of the fits it is safe to say the exponential distribution provides a good model for the inter-arrival times.

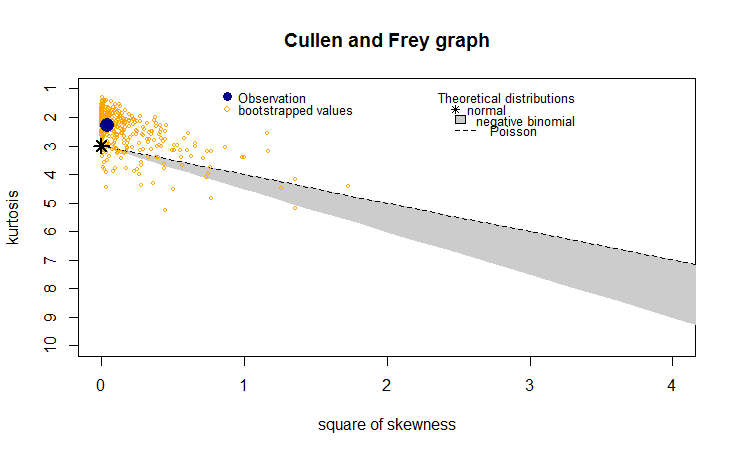
### Part B

##### Do you think a Poisson distribution provides a good model for the count data?

First I generated the count data from the inter-arrival data.

car.df <-  
 data.frame('iat' = interArrival) %>%  
 mutate(actualT = cumsum(iat),   
 tmp=actualT/15,   
 group = ceiling(tmp))  
  
# Compare empirical and theoretical distributions for car arrivals  
Which15Sec <- car.df$group # Unit in which each car arrived  
CarsBy15Sec <- tabulate(Which15Sec) # Counts of cars arrived in each unit

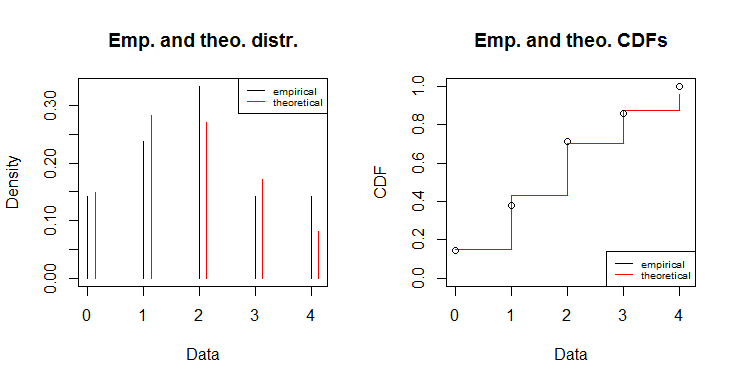
Then I wanted to explore a variety of distributions to see if any looked like possible fits other than Poisson. The chart below shows the kurtosis and squared skewness of the count data sample plotted as a blue point (“Observation”).



Distribution Exploration

It looks like the Poisson and normal may be the best distributions to model the data. To begin analyzing the best model I fit a Poisson distribution on the count data.

fit.pois <- fitdist(CarsBy15Sec, "pois")

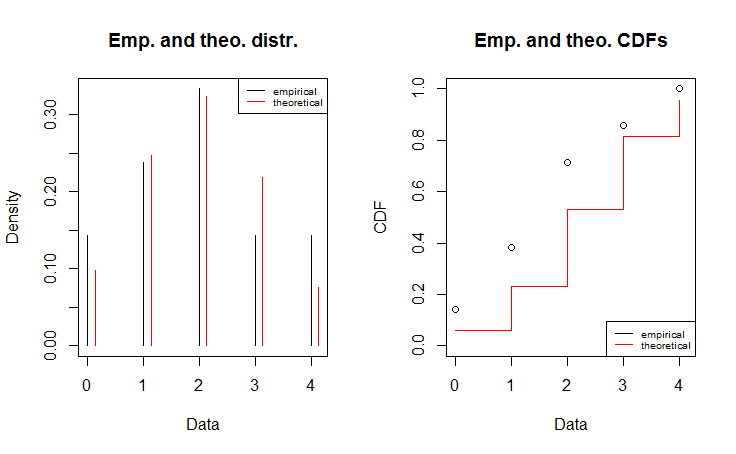


Poisson Distribution

The plot above shows the observations against the theoretical values for the Poisson distribution. Both the pdf and cdf empirical values follow the theoretical values well.

Next I fit the normal distribution to see if there was another distribution that might model the count data well.

fit.norm <- fitdist(CarsBy15Sec, 'norm', discrete = TRUE)



Normal Distribution

The Poisson model seems to fit the data better in both plots. The CDF plot of the normal looks like a much worse fit then the Poisson does. I also inspected the summaries of the fits. The Poisson distribution has the lower AIC and BIC.

summary(fit.pois)

## Fitting of the distribution ' pois ' by maximum likelihood   
## Parameters :   
## estimate Std. Error  
## lambda 1.904762 0.3011692  
## Loglikelihood: -33.98719 AIC: 69.97438 BIC: 71.0189

summary(fit.norm)

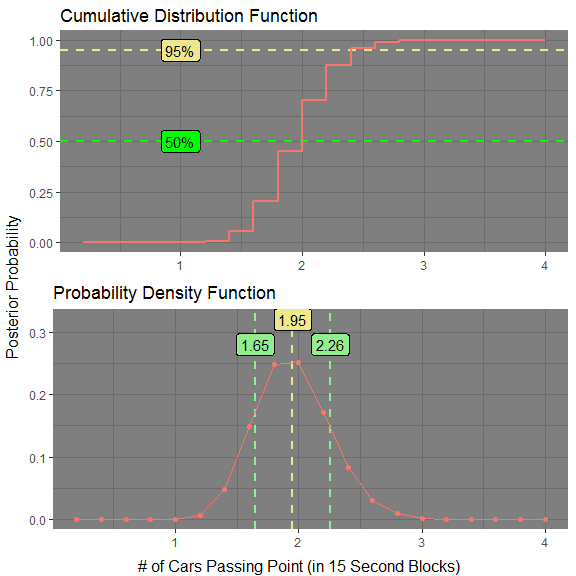
## Fitting of the distribution ' norm ' by maximum likelihood   
## Parameters :   
## estimate Std. Error  
## mean 1.904762 0.2685711  
## sd 1.230747 0.1899079  
## Loglikelihood: -34.15776 AIC: 72.31553 BIC: 74.40457   
## Correlation matrix:  
## mean sd  
## mean 1 0  
## sd 0 1

After inspecting the plots above and summary of the fits it is safe to say the Poisson distribution provides a good model for the count data.

### Part C

##### Find the posterior mean, standard deviation, median and 95th percentile of lambda given the observations.

The plots below show the posterior mean, standard deviation, median and 95th percentile.

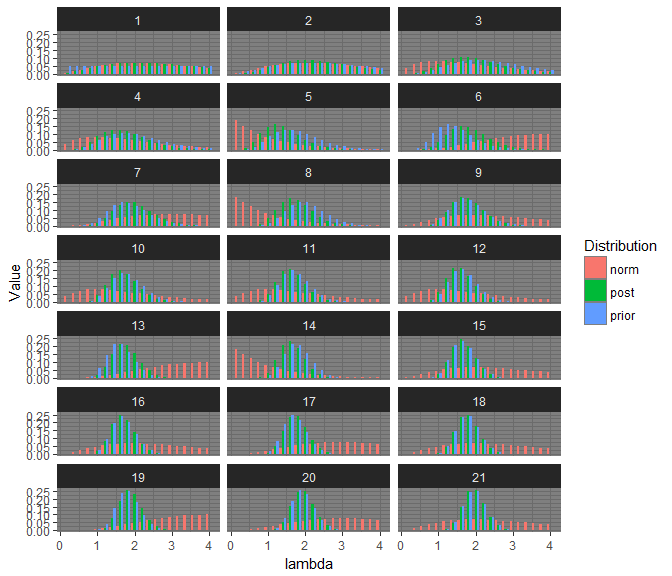


The actual values are as follows:

* Mean: 1.952
* Standard Deviation: 0.305
* Median: 2.0
* 95th Percentile: 2.4

##### Describe what your results mean in terms of traffic on this motorway.

The plot below shows how the posterior inter-arrival distribution changes with each successive unit of 15 seconds of data collected. The posterior distribution in green is altered less and less as the amount of data collected grows. The mean of the posterior begins to settle in on 1.952.



Based on what we know, the data suggests that every 15 seconds there will be about 2 (+/- 0.305) cars passing this point. Without knowing the time of day or area this data was collected it is tough to say much more about the highway traffic and whether this was higher or lower then the daily average.