uc3m | Universidad Carlos III de Madrid

Master in Big Data Analytics
2021-2022

Statistics for Data Analysis. Final Project

“Statistical study of real data in pregnancy”

Miguel Zabaleta Sarasa

Index

1. Introduction

One of the areas in which statistics is most used and useful is medicine. With my father being a gynecologist, I realized I had a great opportunity to work with real data from this field which would come from a reliable source. Additionally, I knew that the results that I would get from this data would be consistent with reality, and very different from a demo sample data.

In particular, this data consisted of more than 40,000 observations which considered many variables involved in the pregnancy of women. For instance, there were variables such as gender and weight of the newborn and the maternal age; with these three variables in mind, I researched and found the following, which led me to set the **weight** as the central variable in my study:

The weight of the newborn is a variable of extreme importance to analyze perinatal mortality. Even though there are different neonatal weight tables, there are significant differences among the different populations, especially due to the differences among ethnicities, maternal weight, nutritional and sanitary condition of the population. For this reason, it is recommended to create personalized tables of neonatal weight for each population. Newborns with low weight for their gestational age are associated with a higher perinatal 'morbid-mortality'. Big newborns for their gestational age present more difficult births, higher caesarean rate and higher morbidity for the mother and the fetus.

It is known that the newborn's weight varies according to the fetal gender (boys have a higher weight than girls) and depending on the mother's risk of diabetes (diabetic mother's children tend to be heavier at moment of birth).

Therefore, in the descriptive analysis we will try to test if this hypothesis is true, and given the appropriate results, determine by statistical inference if our hypothesis is correct. …

We could raise a few clinical questions in the analysis of our sample data:

- Do newborn boys have a higher average weight than girls?
- Are children from mothers with higher risk of diabetes heavier than children from mothers with lower risk?

Thus, the inference chapter will try to confirm if we can say that the observations made in the descriptive analysis are true or not, both as far as the differences between groups (male and female) and regarding the influence between the other variables and the weight of the newborn.

2. Definition of variables

- EDAD: maternal age (truncated years)
- PARIDAD: number of previous children
- ABORTOS: number of previous abortions
- DIAS_PARTO: gestational age (in days)
- GLUCOSA: glucose test in pregnancy (mg/dl)

- HORAS_BOLSA: time of ruptured membranes
- MINUTOS_EXPULSIVO: time of expulsive (how long until baby is born after dilatation has finished) in minutes
- PESO_RN: weight of newborn (grams)
- SEXO: gender of newborn (0 girl, 1 boy)
- PESO_PLACENTA: weight of placenta (grams)
- PH_EXPULSIVO: pH at the expulsive (moment of birth)
- APGAR_1: Apgar test after one minute (neonatal vitality test). From 0 to 10
- APGAR_5: Apgar test after five minutes (neonatal vitality test). From 0 to 10
- PH_ARTERIA: pH at umbilical cord artery
- PH_VENA: pH at umbilical cord vein

3. Descriptive analysis

Before beginning the descriptive analysis as such, we have to perform a small preprocessing study. This means firstly removing the variables that will not be used in this study and removing the observations which have missing values in the remaining variables. Secondly, it is important to take into account the variable *'DIAS_PARTO'*. Observations that have abnormal values in this variable will not be considered in this study, since they greatly influence the weight of the newborn, so it wouldn't make sense to study those cases. Therefore, only observations with **[37-42) weeks of pregnancy** will be considered.

Next, when looking for outliers, it will be possibly interesting to manually input them in our final data set, since they represent valuable information that should not be discarded if one of our goals is to determine the most appropriate distribution fit for all variables. To be able to do this, an *id* variable will be included in the beginning to be able to identify the desired rows.

With this in mind and the previous filters applied, we will separate between two groups, boys and girls, select a random sample without replacement of **size 300 for each group** (fixing the seed to get consistent results), and finally include the considered outliers to its belonging group.

After this, we will begin with the descriptive analysis of these observations, considered as a random independent sample belonging to a population. The described hypothesis will be kept in mind throughout all the study.

3.1 Preprocessing

We will first start by taking the healthy cases that we described before (37-42 weeks of pregnancy). We saw that 20 observations had missing values in this variable; after removing those, we saved the resulting dataframe and proceeded with filtering among healthy observations. Then, we only took into account the variables that seemed most interesting for this study, which are: *'EDAD', 'DIAS_PARTO', 'PARIDAD', 'GLUCOSA', 'PESO_RN', 'SEXO'*.

The final data is obtained by removing all observations with missing values from this dataframe. Since only 6648 out of 33071 observations have missing values, and we will only be sampling 300 observations anyway, there is no problem in removing these records.

Now, we start with the outliers' detection. Observations that seem of special interest for getting the underlying distribution will be saved and input in the resampling in case they weren't selected.

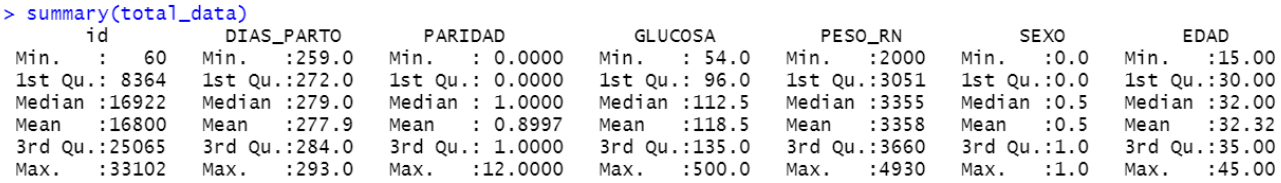
For variable *'EDAD'*, the boxplot considers as outliers all observations with less than 20 and more than 44 years old. It does not make sense to remove those values since they account for a representative part of the population. It also doesn't make sense to remove outliers from *'DIAS_PARTO'* since we are already considering the healthy cases. As for the number of previous children (*'PARIDAD'*), we see that this variable has values that goes from 0 up to 12. This seems a little bit ridiculous and maybe it is due to an error in codification. To see if this is the case, we will look at how many observations there are with more than 6 previous children - there are only 18 observations in total, so we will consider these observations as representing **valuable information**, which will influence the relationship between this and the rest of the variables. This information was also tested with the source of the data, confirming that if there were few of them, this data would be correct.

As for the weight of newborn, we see that it has 3 very distinct outliers in the upper end. Sorting this variable in descending order, we see that the first three values correspond with a supposed weight of 40, 33 and 31 kilos. This is obviously impossible; it is due to an error in the recollecting of the information and therefore these observations will be removed. These ids do not coincide with the other outliers that we want to input, so we can remove these and add the others freely.

The same goes for the variable *'GLUCOSE'*, which has values above 1000 that will be removed (questioned with source of data). We finally see that no values in the ids of the outliers to be added coincide with its respective dataframe, so we can add them freely.

3.2 Descriptive statistics

We will first look at the histograms of the different variables to get an intuition of which distribution they could belong to. But first, a quick summary of some of the basic statistics of these variables:

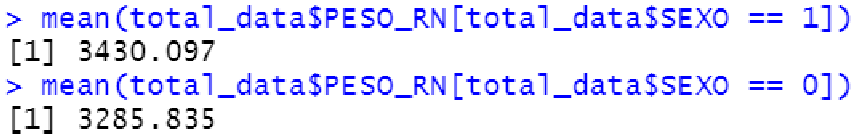


```
> summary(total_data)
       id           DIAS_PARTO        PARIDAD           GLUCOSA          PESO_RN          SEXO             EDAD      
 Min.   :   60   Min.   :259.0   Min.   : 0.0000   Min.   : 54.0   Min.   :2000   Min.   :0.0   Min.   :15.00  
 1st Qu.: 8364   1st Qu.:272.0   1st Qu.: 0.0000   1st Qu.: 96.0   1st Qu.:3051   1st Qu.:0.0   1st Qu.:30.00  
 Median :16922   Median :279.0   Median : 1.0000   Median :112.5   Median :3355   Median :0.5   Median :32.00  
 Mean   :16800   Mean   :277.9   Mean   : 0.8997   Mean   :118.5   Mean   :3358   Mean   :0.5   Mean   :32.32  
 3rd Qu.:25065   3rd Qu.:284.0   3rd Qu.: 1.0000   3rd Qu.:135.0   3rd Qu.:3660   3rd Qu.:1.0   3rd Qu.:35.00  
 Max.   :33102   Max.   :293.0   Max.   :12.0000   Max.   :500.0   Max.   :4930   Max.   :1.0   Max.   :45.00  
```

The worth noticing results that this table is giving us is that *'GLUCOSA'*, *'PESO_RN'* and *'EDAD'*, which we could expect to be Gaussian, in case they were, we could also infer that they would most likely be **symmetric**, since their respective means and medians are very similar. Now we will begin with the study of the underlying distributions.

As we can see in Figure 1, in the number of previous children it seems like the **Poisson** could be a good fit to this variable. In Figure 2, we can see a clear **Gaussian** distribution underlying the variable weight. Examining its boxplot (Figure 3), we also see a clear **symmetry**. In the next chapter we will see if in fact we can confirm whether these hypotheses are false, or whether we don't have enough evidences to say the contrary.

Moving along with the 'weight' variable, since it seems like it would be possible to be a symmetric Gaussian distribution, it is appropriate to now compare whether it seems like there are significant differences in the weight between boys and girls. Considering Figure 4, we see that the medians are a bit different, and regarding the corresponding means:



```
> mean(total_data$PESO_RN[total_data$SEXO == 1])
[1] 3430.097
> mean(total_data$PESO_RN[total_data$SEXO == 0])
[1] 3285.835
```

We see that there is a 200-gram difference between the groups. We cannot yet conclude if these are considerable differences, so this hypothesis will be tested in the inference chapter.

We continue now with variable 'Age', where we could expect a **Gaussian** distribution. Considering Figure 5, we see that this is probably true. In this case, we could maybe appreciate a slight left-skewed distribution, but if we look more closely, we see that the left tail is elongated due to bars that correspond with very few cases of the sample. So, if we were to eliminate these observations, the histogram would represent a much more evident symmetry; therefore, we will not assume a left-skewed distribution. The skewness coefficient obtained was -0.3; as it is quite close to 0, it indicates that it is in fact **symmetric**, corresponding with the observation given by the summary statistics table.

Next, considering the amount of glucose in blood, according to Figure 6, we can presume that it is possible that a **Gamma** distribution could be a good fit to this variable. However, by applying a **log transformation**, we can see that this variable is also quite **Gaussian** (Figure 9). We will fit both models (Gamma and Log-Normal) to the variable in the next chapter and decide which is better.

Finally, we study the variable that measures the duration of pregnancy. At first, we observe the histogram (Figure 7), where we can deduce a **Gaussian** distribution. As we have only considered healthy observations ([37-42) weeks of pregnancy), we know that this variable is bounded, so we will try to fit a Beta distribution with the scaled variable (values from 0 to 1). The result is given in Figure 8. From this graph we can infer that a **Beta** distribution could be a good fit. We will test if both the Gaussian and the Beta are good distributions, and decide which is better.

To conclude this chapter, we will look at the scatter plots of the continuous variables against the weight, to see if there exists a correlation between them that could be tested later on. We will also try to implement appropriate transformations to get better results.

We tried plotting the weight against the mentioned variables, and also did all the combinations including the log and the root transformation, and no relevant results were found. We mostly found that there was no correlation, apart from the plots with *'DIAS_PARTO'*, which showed a bit more relationship between the variables. Some of these graphs are included in Figure 10. We thought that maybe the gender was affecting the correlation, so we tried to do all combinations differentiating between boys and girls, but we obtained similar results. The code for all these plots and correlations is included in the appropriate file.

Now that we have concluded with the descriptive analysis, we will try to fit the best possible model (distribution) for each of the previously described variables, so that we can appropriately make the necessary statistical assumptions to make our hypothesis tests.

4. Model fitting

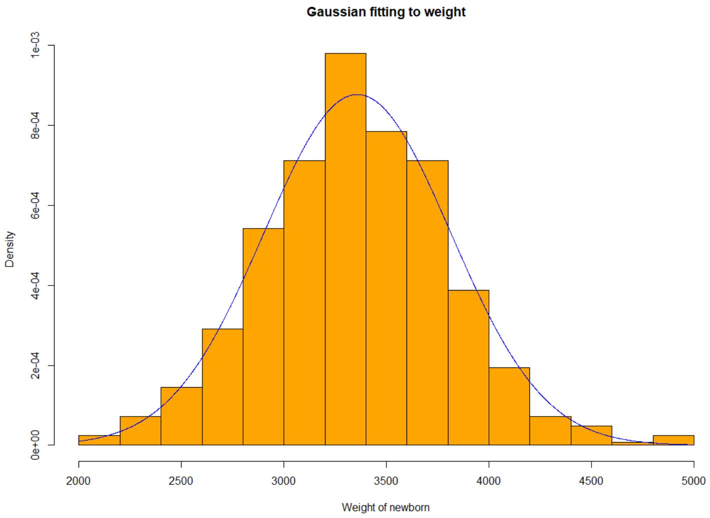
The approach we will take for this part of the study is to first find statistical evidences that our main variable (weight of newborn) follows a Gaussian distribution. This will allow us to make better inferences in the latter chapter. Also, we will try to analyze the distribution of variables which did not have one clear underlying model, as we saw in the last chapter. That is, to study:

- if the weight of the newborn (our main variable) follows a Gaussian distribution,
- if the number of previous children follows a Poisson distribution,
- if for the variable 'glucose', it is better to fit a Gamma or a Log-Normal model,
- if the number of days in pregnancy follows a Gaussian or a Beta distribution (if it is scaled)

Both the *mme* and the *mle* methods of estimation will be used, selecting the best model using the AIC criteria. After choosing the best possible distribution for each variable, we will also try to confirm these statements via the Kolmogorov-Smirnov test, which can be applied to test if you have statistical evidence to affirm that your sample does not follow a certain distribution (in case you reject the null hypothesis).

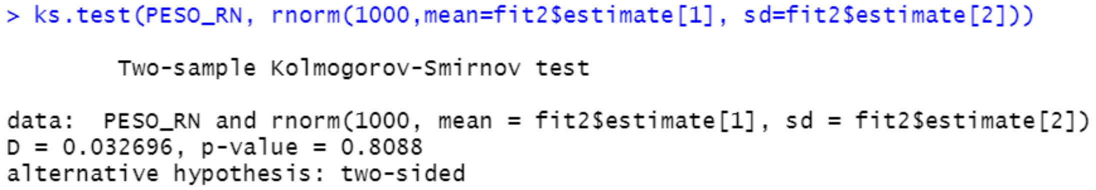
Gaussian fitting to weight

Beginning with the weight, both the *mme* and *mle* provided the same estimation of parameters. Below is the corresponding histogram, with the model's lines included.



We can appreciate a very clear **Gaussian** distribution. The *ecdf* plot was also very well adjusted (included in Figure 17).

We now apply the Kolmogorov-Smirnov test, comparing our distribution with one given by 1000 random observations that follow a Gaussian distribution with the parameters that we got with the *mle* method. The results are the following:



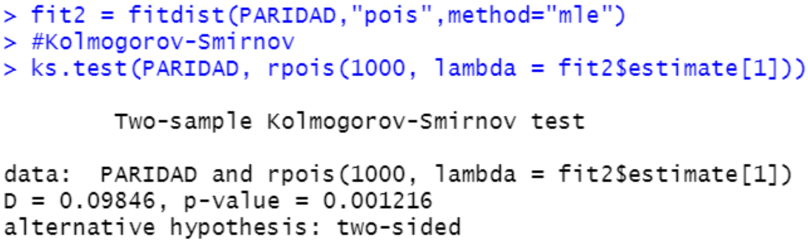
```
> ks.test(PESO_RN, rnorm(1000,mean=fit2$estimate[1], sd=fit2$estimate[2]))

	Two-sample Kolmogorov-Smirnov test

data:  PESO_RN and rnorm(1000, mean = fit2$estimate[1], sd = fit2$estimate[2])
D = 0.032696, p-value = 0.8088
alternative hypothesis: two-sided
```

We can conclude that we do not have statistical evidences to deny that our variable follows a Gaussian distribution with parameters given by the *mle* method (that is, we cannot reject the null hypothesis).

Poisson fitting to number of previous children

For studying if the *'PARIDAD'* variable follows a Poisson distribution, we fitted a model using the *mme* and *mle* methods, and there was no difference between both methods, they gave the same AIC. Looking at the respective plot and *ecdf* (figures 11, 12), we can conclude that, according to our descriptive study, the number of previous children seems to follows a Poisson distribution**.**

However, once we compute the K-S test, we see that the obtained p-value is 0.001, which is smaller than 0.05.



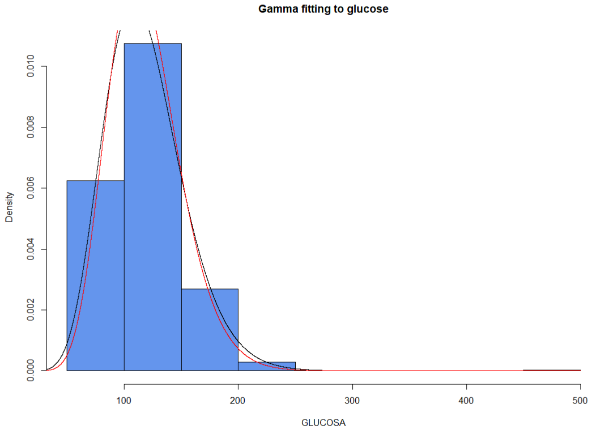
```
> fit2 = fitdist(PARIDAD,"pois",method="mle")
> #Kolmogorov-Smirnov
> ks.test(PARIDAD, rpois(1000, lambda = fit2$estimate[1]))

	Two-sample Kolmogorov-Smirnov test

data:  PARIDAD and rpois(1000, lambda = fit2$estimate[1])
D = 0.09846, p-value = 0.001216
alternative hypothesis: two-sided
```

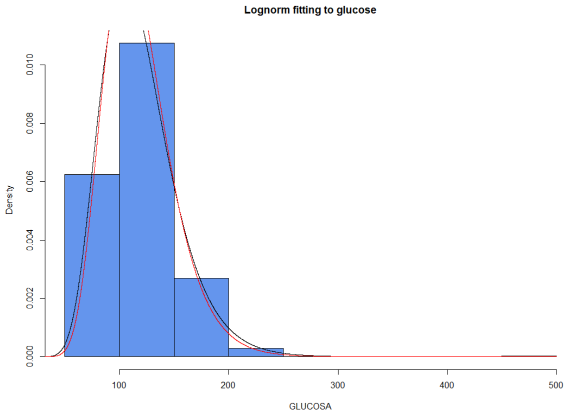
Therefore, according to this test we can say that we reject the null hypothesis with a significance level of 0.05, that is, that our variable follows a Poisson distribution with lambda given by the *mle* method.

For this variable, the descriptive and the analytic approaches seem to give different results (since the plots were very well fitted). Therefore, we cannot conclude about the distribution of this variable.

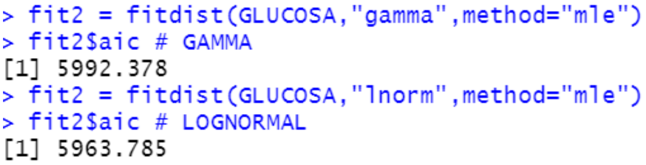
Gamma and Log-Normal fitting to glucose

Now we will examine the amount of glucose in blood. We will start by showing the histograms along with the lines for the Gamma and Log-Normal distributions:


Gamma fitting to glucose

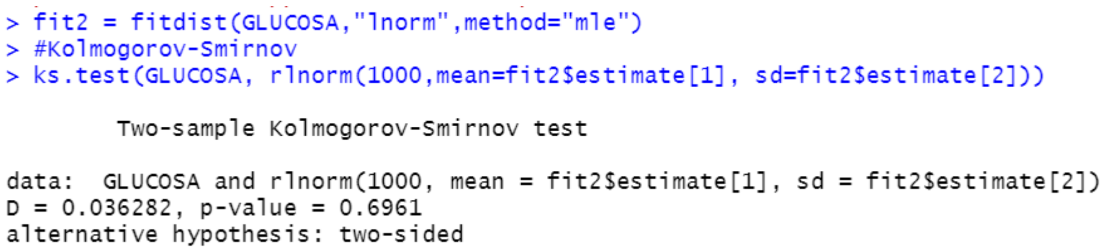

Lognorm fitting to glucose

We can see that there doesn't seem to be a lot of differences between both histograms, both seem like a good fitting model. The respective *ecdf* plots can be found in Figure 13 and Figure 14 respectively, which also indicate that they are both good models. To conclude which distribution is the best fitting, all four AIC were obtained and by selecting the smallest one we are able to conclude that the amount of glucose in blood **follows a Log-Normal distribution**. The specific results were the following:



```
> fit2 = fitdist(GLUCOSA,"gamma",method="mle")
> fit2$aic # GAMMA
[1] 5992.378
> fit2 = fitdist(GLUCOSA,"lnorm",method="mle")
> fit2$aic # LOGNORMAL
[1] 5963.785
```

The Kolmogorov-Smirnov test gives the following result:



```
> fit2 = fitdist(GLUCOSA,"lnorm",method="mle")
> #Kolmogorov-Smirnov
> ks.test(GLUCOSA, rlnorm(1000,mean=fit2$estimate[1], sd=fit2$estimate[2]))

        Two-sample Kolmogorov-Smirnov test

data:  GLUCOSA and rlnorm(1000, mean = fit2$estimate[1], sd = fit2$estimate[2])
D = 0.036282, p-value = 0.6961
alternative hypothesis: two-sided
```

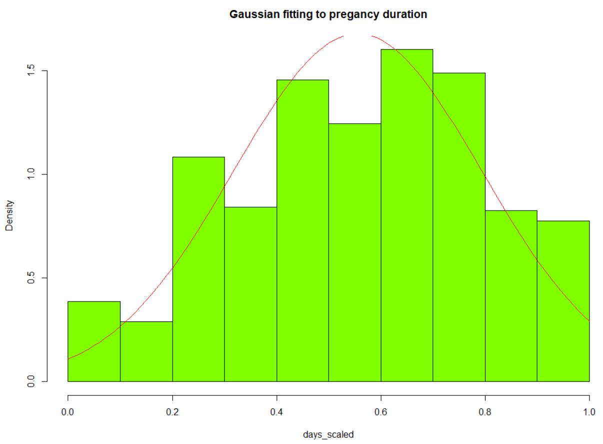
As we can see, p-value is 0.6961, which is greater than 0.05; therefore, we cannot reject the null hypothesis that the amount of glucose in blood follows a Log-Normal distribution, with parameters given by the *mle* method.

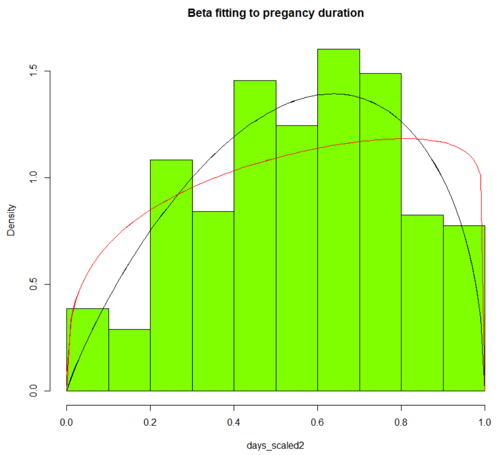
Beta and Gaussian fitting to days in pregnancy

The last variable that we will analyze is the number of days in pregnancy. Let us recall that as we know this variable is bounded, our aim was to fit a Beta and a Gaussian distribution to the data, given the resulting histogram. For doing this, we decided to apply both models to the scaled data (not just to the Beta) so that is easier to compare their respective AIC.

We also want to note that for the Beta model, the R function was not able to estimate its parameters when using the *mle* method, so we adjusted its values equal to 0 to 0.0001 and values equal to 1 to 0.9999.

We now compare both histograms (for the Gaussian, *mle* and *mme* offered same results):





We can see that they could both be good options for fitting this variable. In the image to the right, we can see that the estimation obtained by the *mle* (red line) seems a bit more worse-adjusted. This could be due to the number of bars selected in the plot, or because the *mle* fails to find a good model in this particular case, given that there are values close to 0 and close to 1.

We will conclude our observations with the AIC coefficients, which show that in fact the Beta distribution obtained by the *mle* is the best fit, as it has the lowest value out of the 4 models (-26.23). To help us determine whether it is the underlying distribution, we perform the K-S test, which gives us a p-value smaller than 0.05. Therefore, we reject the null hypothesis, and affirm that (at least according to this test) this scaled variable does not follow a Beta distribution with parameters given by the *mle* method.

5. Statistical inference of one variable

Now that we have concluded with the model fitting of our variables of most interest, and with such deciding results, it's time to take advantage of our resulting distributions and make some interesting inferences.

In the case of the inference for one variable, we decided to show the confidence interval for our principal variable (weight of newborn) and also test if our data coincides with the population data shown in the internet. In particular, we will test if the average weight of newborns is equal to a certain number, depending on the week of pregnancy. Sources for this data can be found in the References [1][2].

As for the hypothesis that have to be applied for constructing the confidence interval, in our case we can confidently affirm to know the distribution of the statistic 'mean of the weight'. This is because we have shown that it is very likely that the weight is Gaussian. Therefore, assuming that the population variance is unknown, we know that under these conditions, the mean of the weight will follow a t-student distribution with $n-1$ degrees of freedom. So, the *t.test()* function in R will do a perfect job of constructing our confidence interval. The result is that a 95% confidence interval for the population mean weight is **(3321.992, 3393.94)**.

Now, to test the hypothesis that depending on the week of pregnancy, the average weight of newborns is equal to a certain number, the conceptual procedure will be a bit different.

We cannot assume that the distribution of the weight during a certain week follows a Gaussian distribution. Therefore, we will make use of the Central Limit Theorem, which tells us that the mean converges to a Gaussian distribution as the sample size goes to infinity. For applying this theorem, we need that each of our variables (weight per week, between [37-42)) has more than 30 sample size, which they all do.

Constructing the according hypothesis tests, these are the results:

Week	Value to be tested	Statistic (t)	p-value	Result
37	2976	-0.594	0.55	Don't reject H_0
38	3158	0.463	0.64	Don't reject H_0
39	3314	0.105	0.91	Don't reject H_0
40	3425	0.653	0.51	Don't reject H_0
41	3508	1.464	0.14	Don't reject H_0

The conclusion is that we have not found statistical evidence to affirm that the average weight for each week in our data is different from the value of the population.

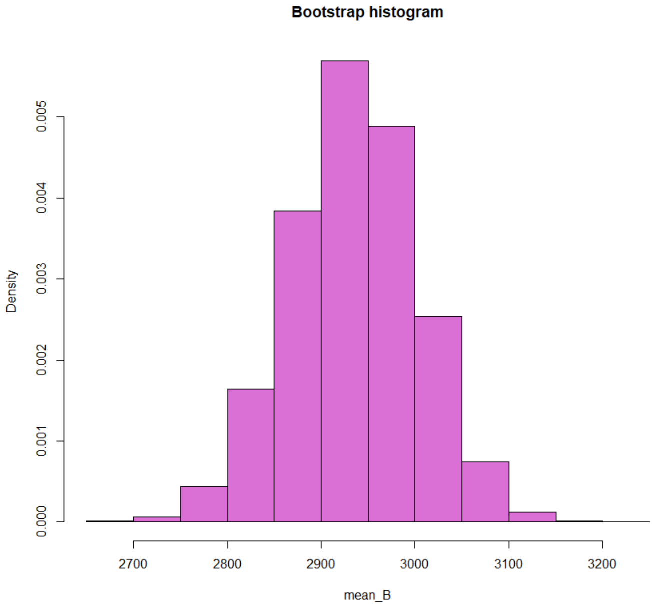
Additional analysis: Bootstrap estimators

Additionally, as an exercise we will apply the bootstrap technique to obtain more robust estimators of the population mean. I want to be clear that the bootstrap technique should not be applied to get better estimators of our population's parameters, but instead to study the precision of an estimator by getting its distribution.

The bootstrap is applied in the following way:

We first set the bootstrap size to 10.000, this means that we will get 10.000 samples with replacement of our variables (weight on each week). For each sample, we save the its mean to a new vector (*mean_B*) which will represent the distribution of our statistic. With this, the final robust estimation of our mean will be the mean of this vector. I also included a small *'if'* condition to test if the initial estimator was closer to the population mean or if the robust estimator is closer.

As a note, this is the histogram of our first bootstrap:



It is clearly Gaussian, so we have confirmed the result that states that if *X* is Gaussian, the sample mean will also be Gaussian.

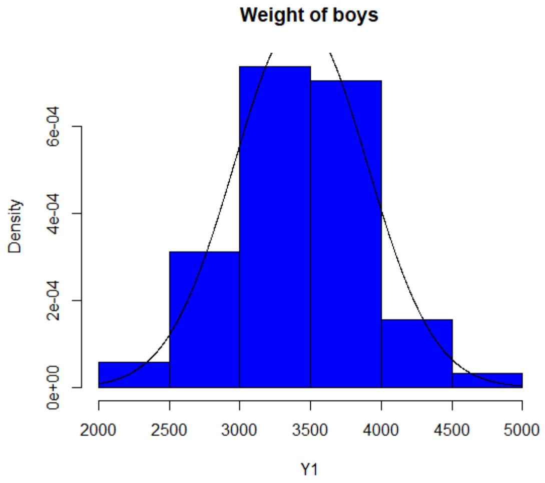
In the five different weeks that we studied, 3/5 times the bootstrap estimator was closer to the population mean value.

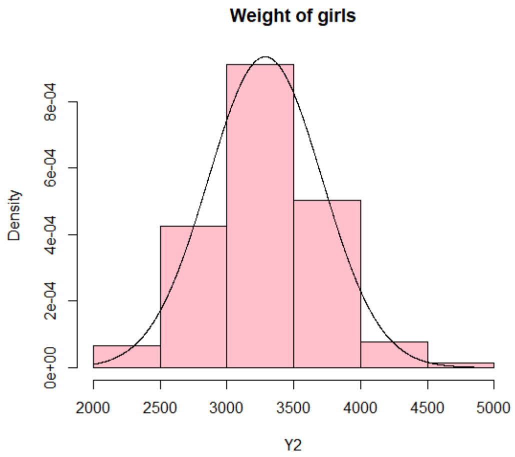
6. Statistical inference of two variables

As for the inference of two variables, this is the chapter where I wanted to get the most interesting results, given the valuable information in my data. One of the most apparent questions that we could try to answer is: Are there differences in the weight between boys and girls? Also, I found that supposedly, women that have diabetes tend to give birth to heavier babies. We will also try to give answer to this question with our data, with some previous research about the commonly used threshold of glucose in blood for defining a woman with diabetes. Lastly, we will analyze the dependence between two quantitative variables: the weight and the amount of glucose. *CON RIESGO DE DIABETES*

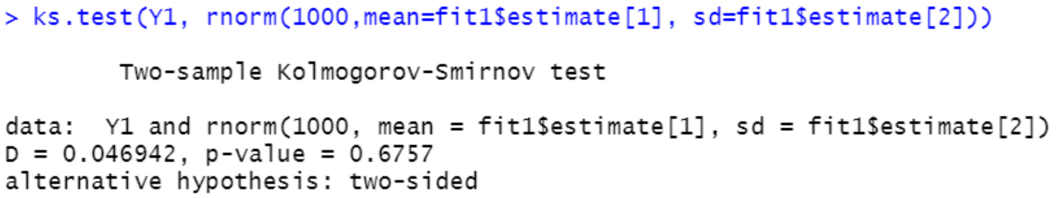
Differences in weight depending on gender

Again, we know that the weight is Gaussian but we do not know if the weight depending on gender is Gaussian. We will try to find if this is true to make the appropriate assumptions in our hypothesis test. From their histograms alone we can sense a Gaussian distribution, so we estimate the parameters of this distribution with the *mle* method, and plot the line that it should follow onto its histogram. These are the results:

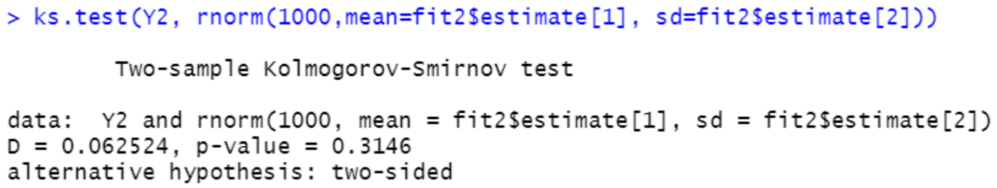




Their *ecdf* plots are also very promising, which can be found in figure 18. Finally, we perform a K-S test on both samples to get an analytic perspective and their p-values are very large, so we cannot find statistical evidences that confirm that these variables don't follow a Gaussian distribution. Therefore, we will treat them as if they do follow a Gaussian distribution. These were the results obtained:

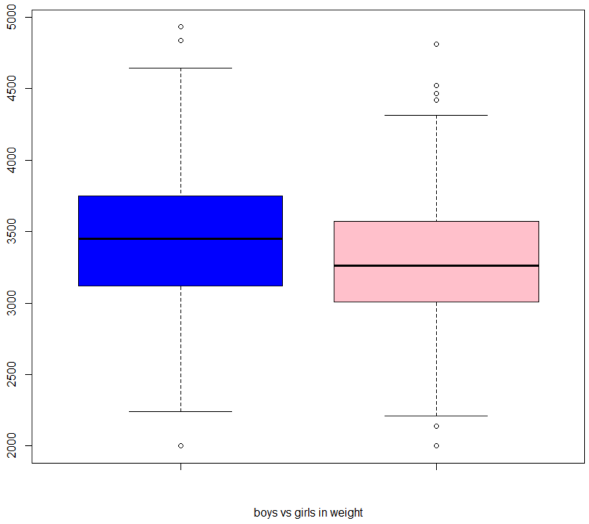


```
> ks.test(Y1, rnorm(1000,mean=fit1$estimate[1], sd=fit1$estimate[2]))

        Two-sample Kolmogorov-Smirnov test

data:  Y1 and rnorm(1000, mean = fit1$estimate[1], sd = fit1$estimate[2])
D = 0.046942, p-value = 0.6757
alternative hypothesis: two-sided
```



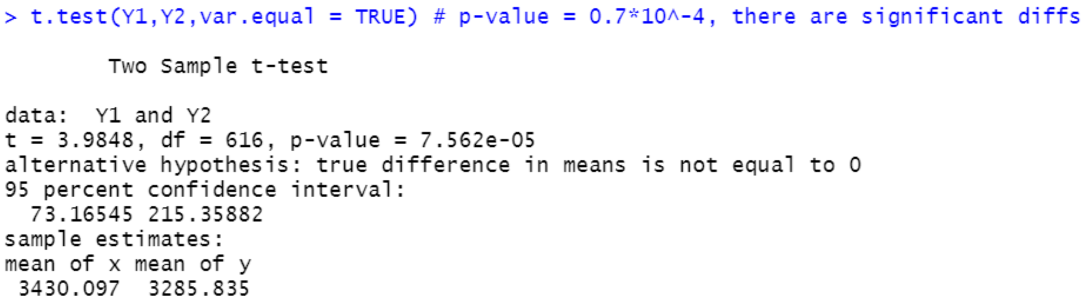
```
> ks.test(Y2, rnorm(1000,mean=fit2$estimate[1], sd=fit2$estimate[2]))

	Two-sample Kolmogorov-Smirnov test

data:  Y2 and rnorm(1000, mean = fit2$estimate[1], sd = fit2$estimate[2])
D = 0.062524, p-value = 0.3146
alternative hypothesis: two-sided
```

Taking a quick look at their respective boxplots:



We can presume a significant difference in weight. To test this hypothesis, we will first make a test that will tell us if their variances are different or not. When we do that, we obtain a p-value > 0.05, so we will treat them as if they had equal variances. Now, as we have found evidences that both variables are very likely to be Gaussian, we know that the distribution of the differences of the mean will follow a t-student distribution, so we can perform this test with these assumptions being correct.

By doing so, we obtain the following results:



```
> t.test(Y1,Y2,var.equal = TRUE) # p-value = 0.7*10^-4, there are significant diffs

	Two Sample t-test

data:  Y1 and Y2
t = 3.9848, df = 616, p-value = 7.562e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  73.16545 215.35882
sample estimates:
mean of x mean of y 
 3430.097  3285.835 
```

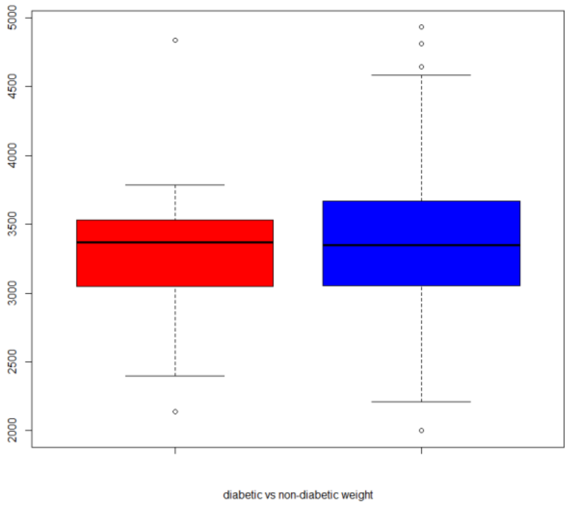
With a value of the statistic of 3.984, we obtain a p-value smaller than 0.05, and therefore we reject the null hypothesis that the means are equal. In other words, we can conclude that **there are significant differences in the average weight between boys and girls**. Another way of looking at this is to look at the confidence interval; as it doesn't contain the value 0, the null hypothesis is rejected.

Difference in weight between diabetic and non-diabetic mothers

Now we will study the difference in weight between the group of mothers who have diabetic levels of glucose and the group that doesn't (reference for this observation can be found at [3]).

To differ between these two groups, we will make sure that the selected amount of glucose in blood is high enough to be considered a 'diabetic level', but also that we have enough samples so that the Central Limit Theorem can be applied.

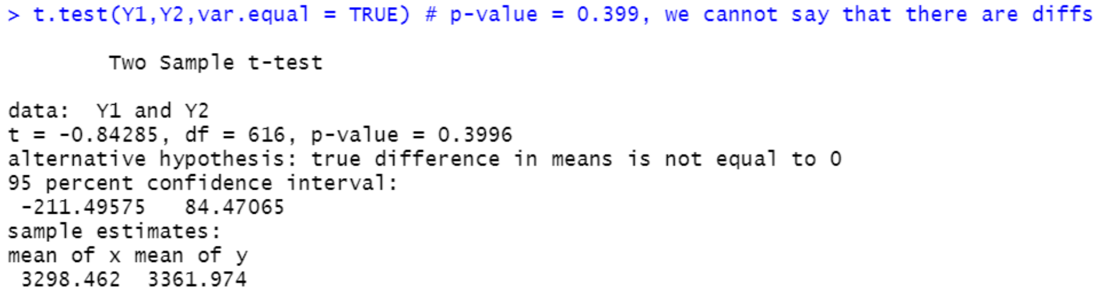
Normally, glucose levels of 140 mg or above is considered diabetic ([4]), so we will take 170 mg to be sure that the group that we get is in fact diabetic. We have 39 observations in such group, so the Central Limit Theorem can be applied. In case we were wrong, and some observations in the 'normal' group were in fact diabetic, we would still have many more non-diabetic observations in that group, so the distribution would not be affected. In any case, this is still a pretty good delimitation of groups.



Comparing their boxplots, we can see a much less evident difference in the weight, compared to the boxplot differentiating between gender.

Once again, we do a test to see if the variances of the populations are the same. We get a p-value greater than 0.05, so we will suppose both populations have the same variance.

We are now in disposition of testing if the difference in average weight is significant. Applying the CLT, we know that the distribution of the difference of the means will follow a Gaussian distribution. Therefore, we can construct the confidence interval in the following way:



```
> t.test(Y1,Y2,var.equal = TRUE) # p-value = 0.399, we cannot say that there are diffs

        Two Sample t-test

data:  Y1 and Y2
t = -0.84285, df = 616, p-value = 0.3996
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -211.49575   84.47065
sample estimates:
mean of x mean of y 
 3298.462  3361.974 
```

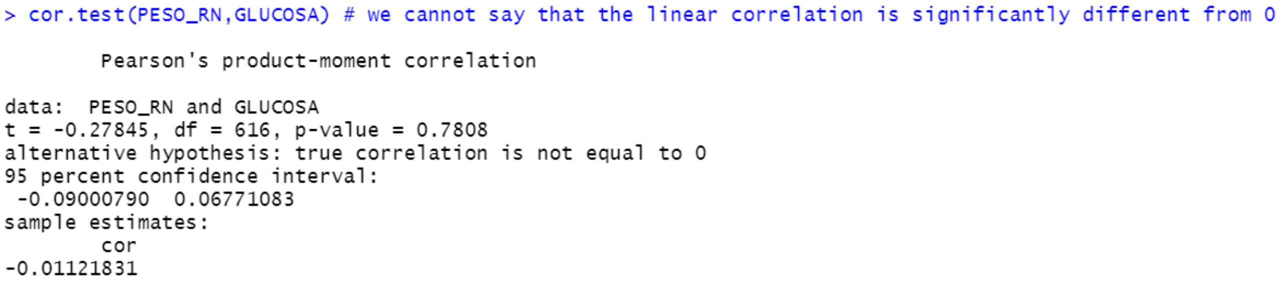
We see that the 0 belongs to our 95% confidence interval which means that we cannot say that there are significant differences in the average weight depending on the diabetic condition of the mother, which is surprising, as it is **contrary to what is known by the science**.

A related question that one could ask oneself is the following: Given that there doesn't seem to be significant differences in the weight depending on very high levels of glucose or regular ones, is the variable 'glucose' related to the weight at all? In other words, is there a significant enough dependence between both variables? That is the question we will try to give answer to next.

Dependence between weight and amount of glucose

For this final comparison, we will assume that the bivariate distribution of the variables 'weight' and 'glucose' is a bivariate normal distribution (we showed that the marginal distributions where most likely Gaussian and Log-Normal, respectively).

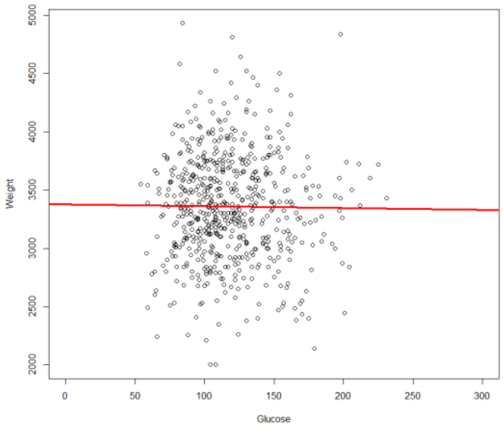
Then, performing a correlation test, we obtain the following results:



```
> cor.test(PESO_RN,GLUCOSA) # we cannot say that the linear correlation is significantly different from 0

	Pearson's product-moment correlation

data:  PESO_RN and GLUCOSA
t = -0.27845, df = 616, p-value = 0.7808
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.09000790  0.06771083
sample estimates:
        cor 
-0.01121831 
```

Therefore, we cannot say that such correlation is significant. In fact, we can see that the sample correlation is very close to being 0. In other words, we have not found statistical evidences to affirm that the linear dependence between 'weight' and 'glucose' is significant.

Additionally, if we look at the scatter plot of both variables and draw a line corresponding to the result given by a linear regression:



We can see by the shape of the plot that in fact it seems like the glucose alone is not related at all to the weight (if we tried to predict the weight in terms of the glucose, we would obtain very bad results).

7. Conclusions

In this study, we were able to successfully answer many of the clinical questions we asked ourselves. As for the study of the underlying distribution of our variables, in some cases we found strong statistical evidences for a certain distribution (both in the descriptive and analytical approach), but in other cases, these approaches seemed to be contrary to each other.

According to the **best model** we could find, such results include:

- Strong likelihood that the weight is Gaussian
- Number of previous children showed contradictions between the descriptive and the Kolmogorov Smirnov test
- Strong likelihood that the glucose is Log-Normal
- Days in pregnancy showed signs of being a Beta (since it was bounded), but the Kolmogorov-Smirnov test denied this hypothesis

Now, as far as the **inferences** we made, these are our final conclusions:

- Regarding the average weight for each week, our sample doesn't seem to have significant differences with the known population values
- We have found statistical evidences that in fact newborn boys are heavier than girls
- Contrary to what is known, we have not been able to find statistical differences in the newborn weight among mothers with high and low risk of diabetes
- Additionally, we have not found a significant dependence between the glucose and the weight (assuming bivariate Gaussian distribution)

With regard to possible future extensions that could be make on this work, I think one good approach would be to take advantage of the valuable information that this data contains. We have shown that many of the things that are known according to studies among the true population are likely to be present in our data. Therefore, a good decision would be to **keep studying this data** and try to test **many different hypotheses** regarding the rest of the variables which were not shown here (number of previous abortions, Apgar test, pH, …) and see if we get different results from the population.

As a final note, I would like to express my satisfaction with this project, I was able to enjoy it very much and appreciate both the statistics and the value of working with excellent data. I hope this enjoyment was carried throughout this work.

References

[1] Population tables regarding neonatal weights (in Spain). https://www.menarini.es/aviso-legal/509-salud/areas-terapeuticas/ginecologia/3073-tablas-espanolas-de-pesos-neonatales.html

[2] Population table. Average weight for each week. https://www.menarini.es/images/ginecologia/tablas-espanolas-neonatos-embarazo-unico-sin-dif-percentilar.pdf

[3] Evidence for differences in weight among mothers with high risk of diabetes. https://evidence.nihr.ac.uk/alert/fewer-large-babies-are-born-to-pregnant-woman-with-type-1-diabetes-if-their-glucose-was-monitored-continuously/#:~:text=Women%20with%20type%201%20diabetes%20are%20at%20higher%20risk%20of,need%20special%20care%20when%20born.

[4] High glucose levels for women in pregnancy. https://www.healthline.com/health/gestational-diabetes#diagnosis

Appendix

Descriptive analysis

Histograms

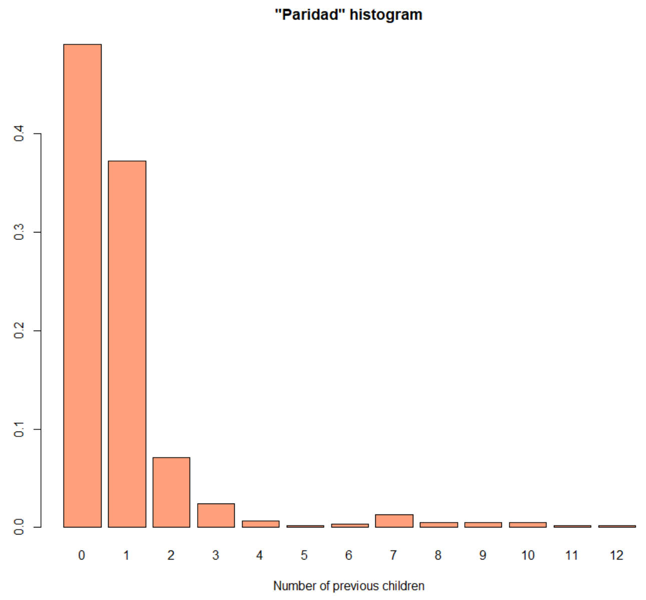


Figure 1: 'Paridad' histogram

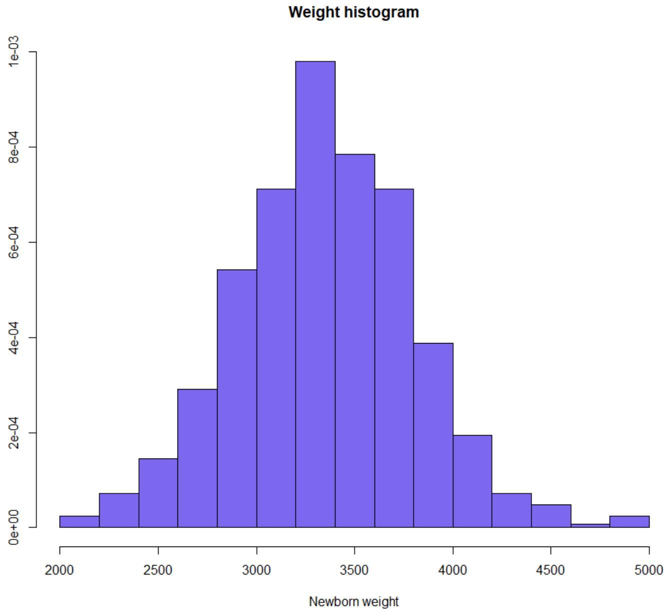


Figure 2: Weight histogram

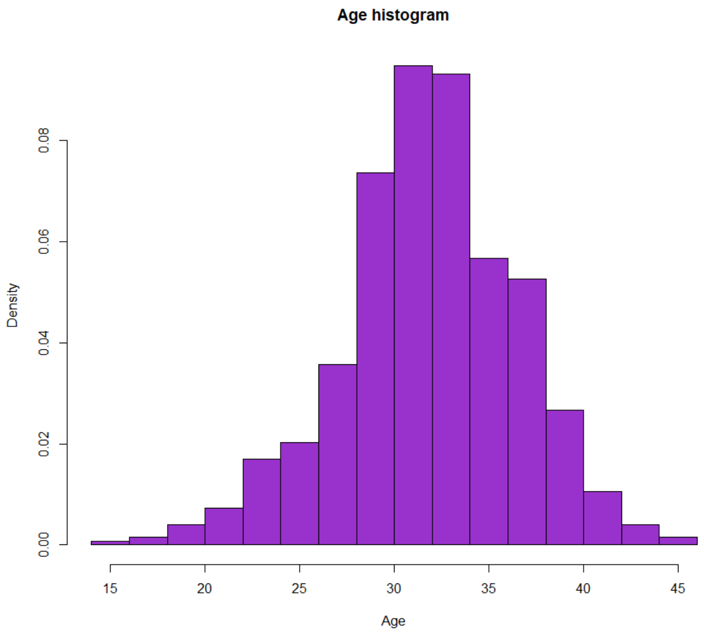


Figure 5: Age histogram

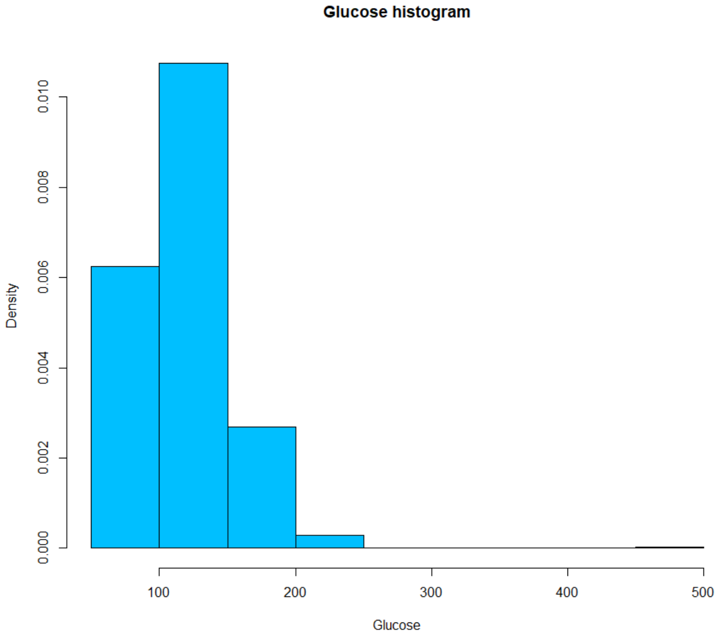


Figure 6: Glucose histogram

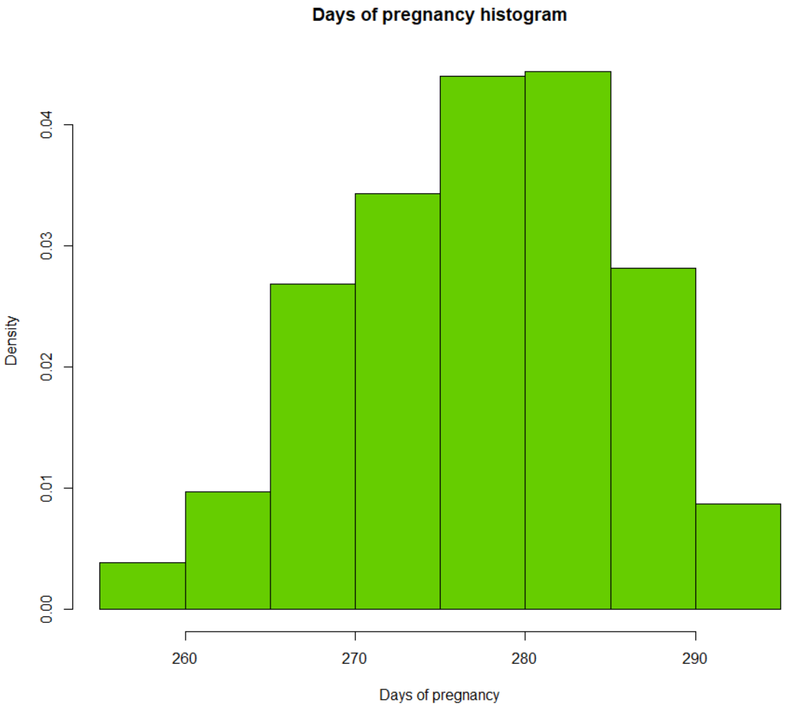


Figure 7: Pregnancy duration histogram

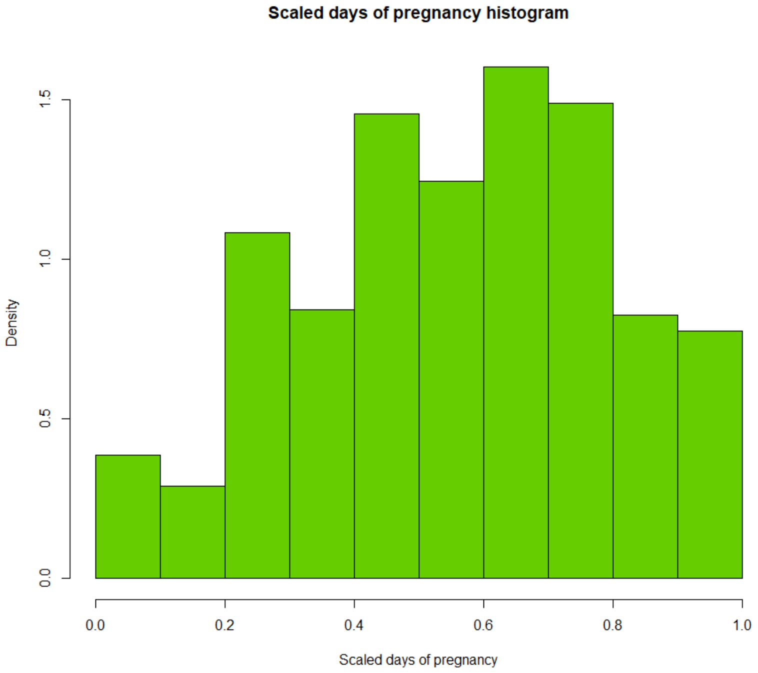


Figure 8: Scaled pregnancy duration histogram

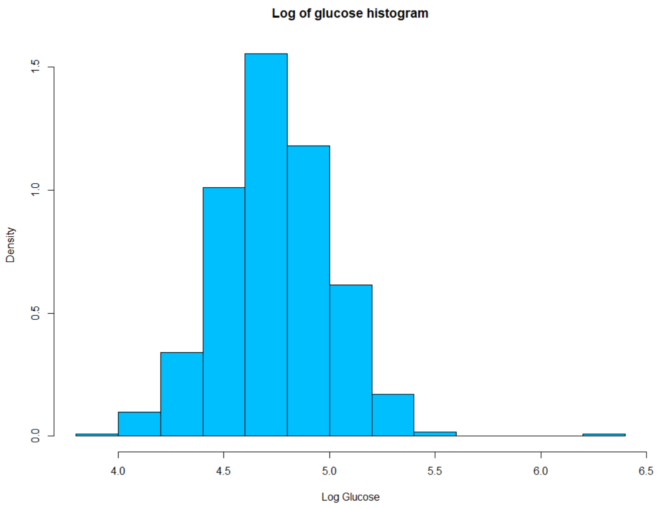


Figure 9: Log of glucose histogram

Boxplots

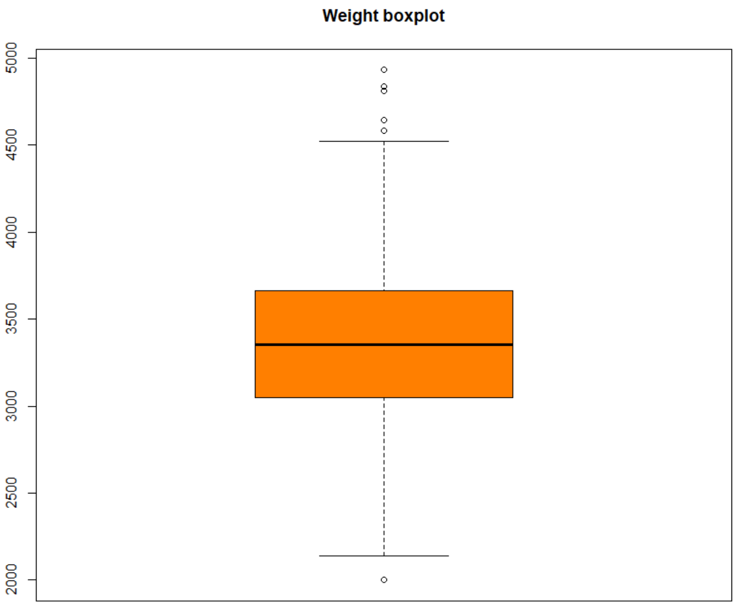


Figure 3: Weight boxplot

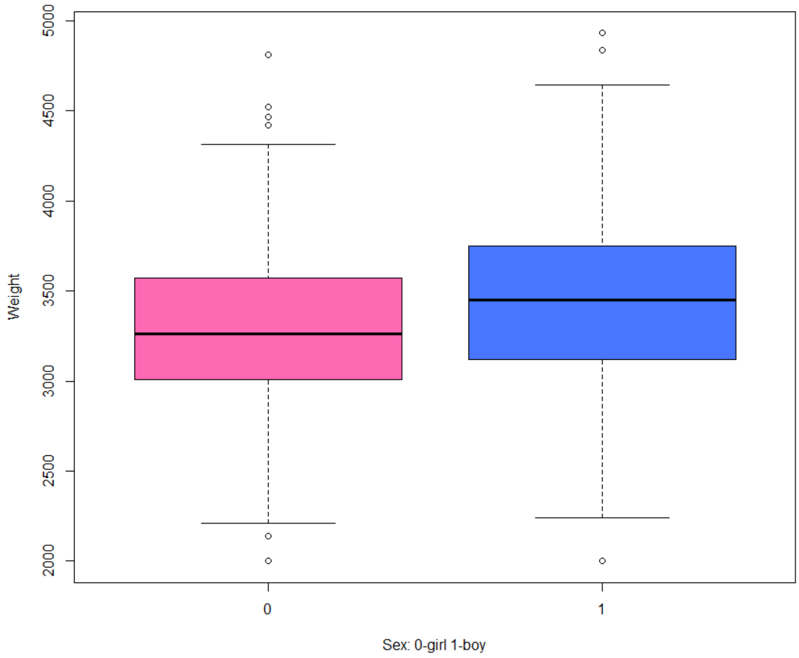


Figure 4: Sex vs Weight boxplot

Scatter plots

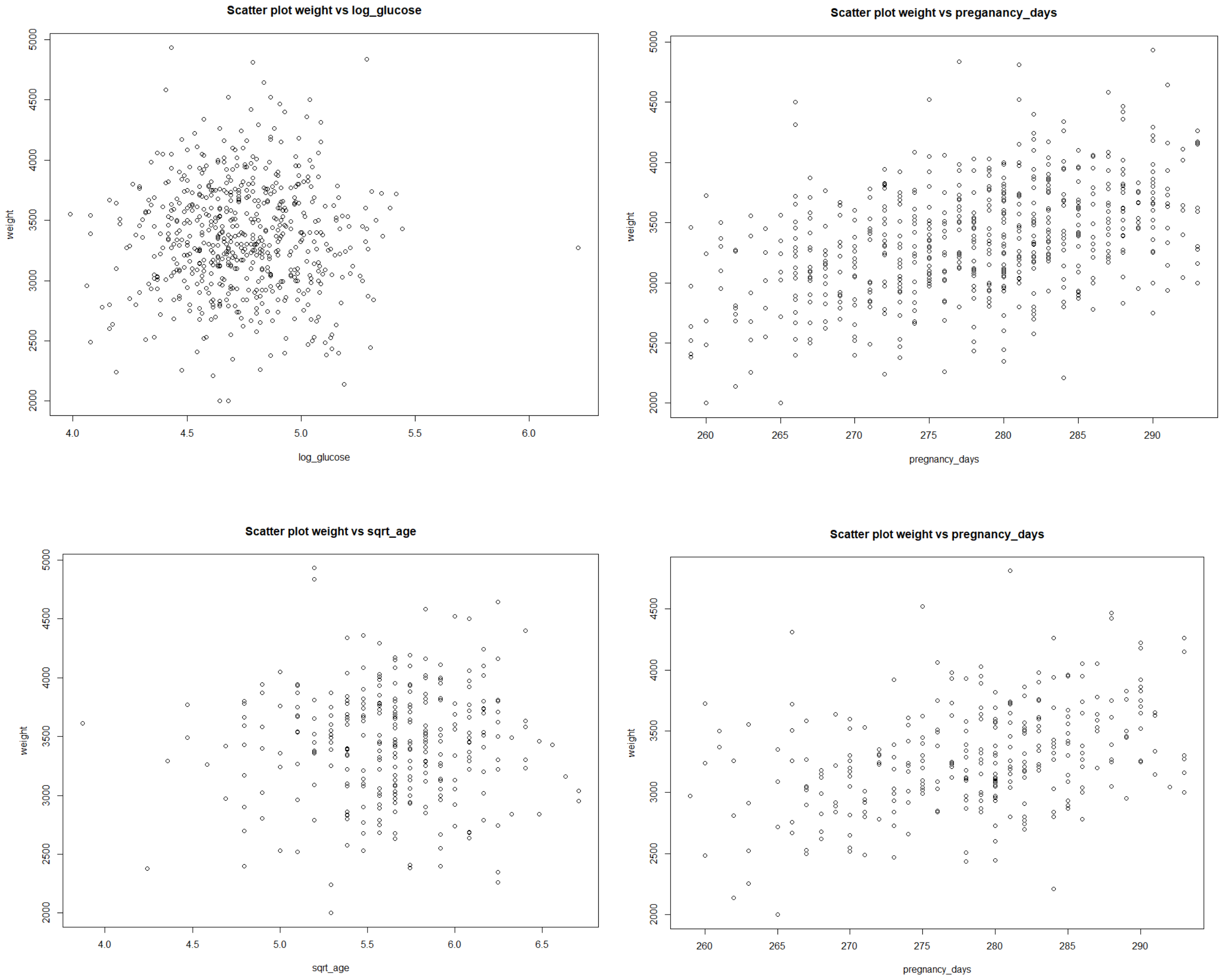


Figure 10: Scatter plots

Model fitting

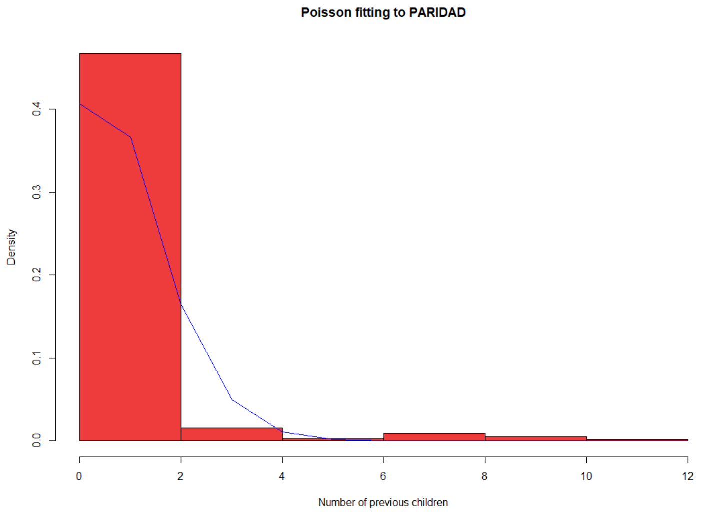


Figure 11: Poisson fitting hist

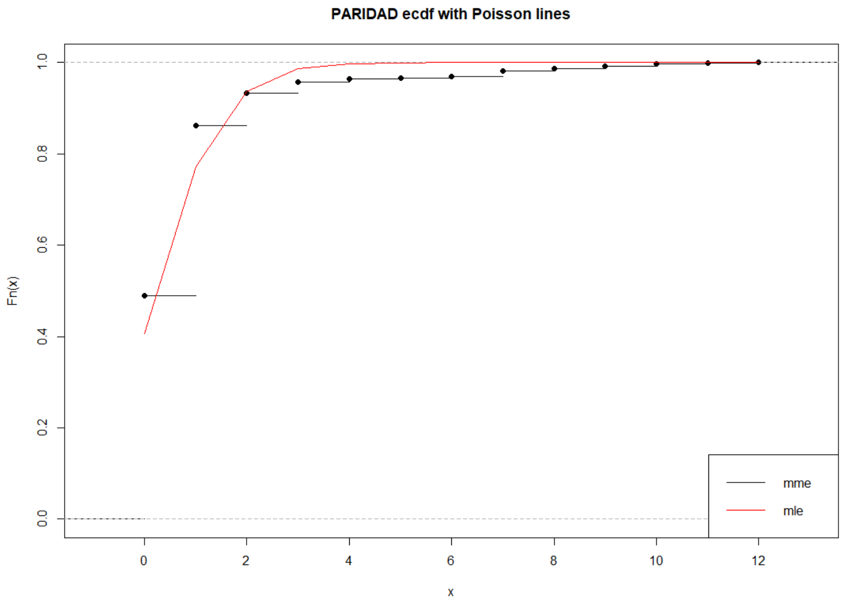


Figure 12: Poisson fitting *ecdf*

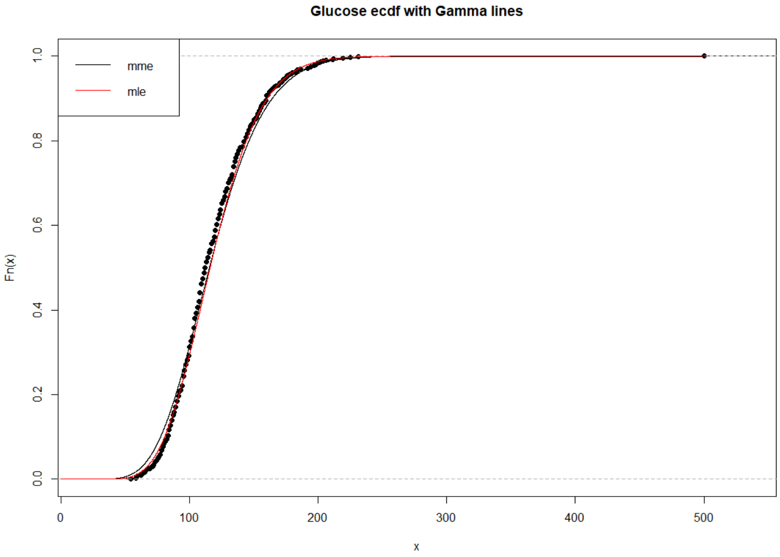


Figure 13: Gamma fitting *ecdf*

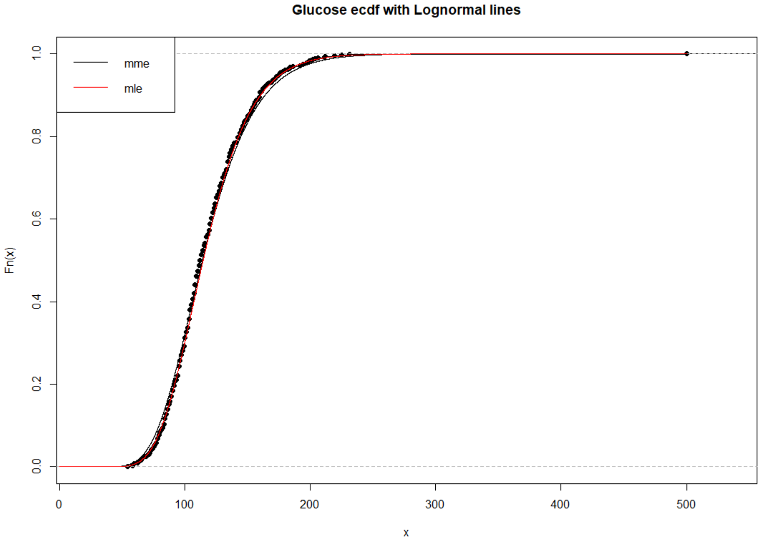


Figure 14: Lognormal fitting *ecdf*

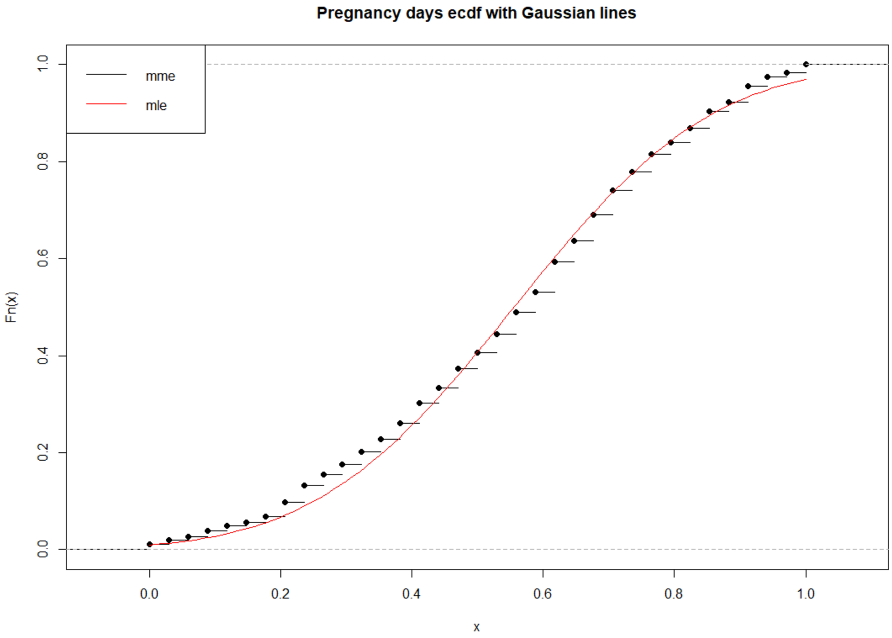


Figure 15: Gaussian fitting *ecdf*

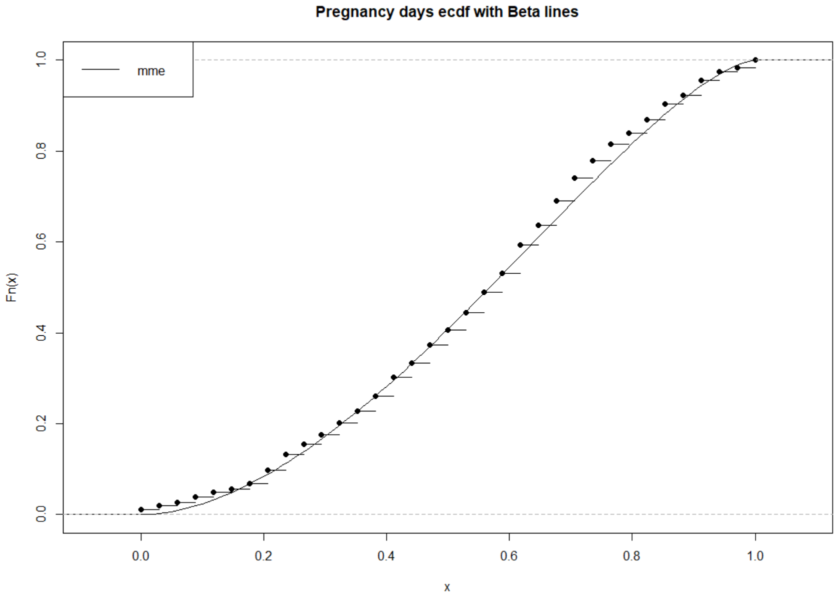


Figure 16: Beta fitting *ecdf*

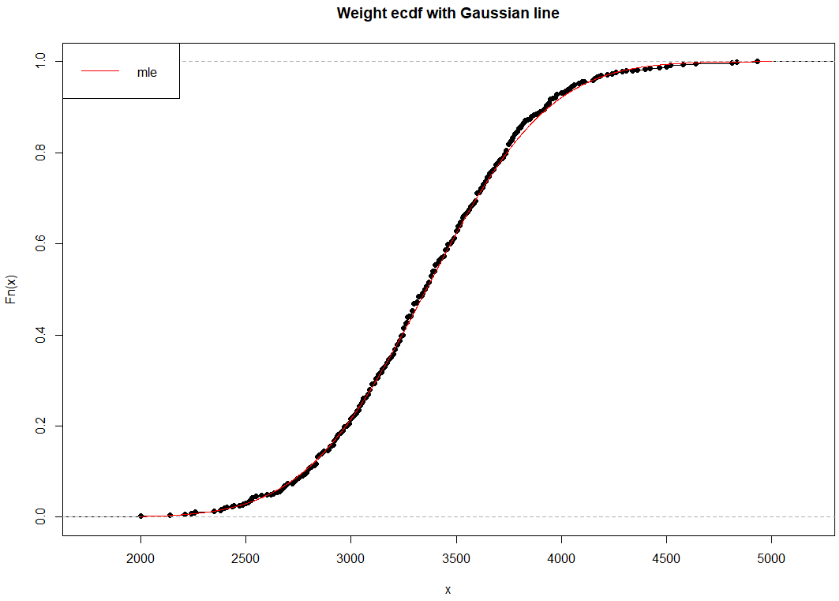


Figure 17: Gaussian fitting *ecdf*

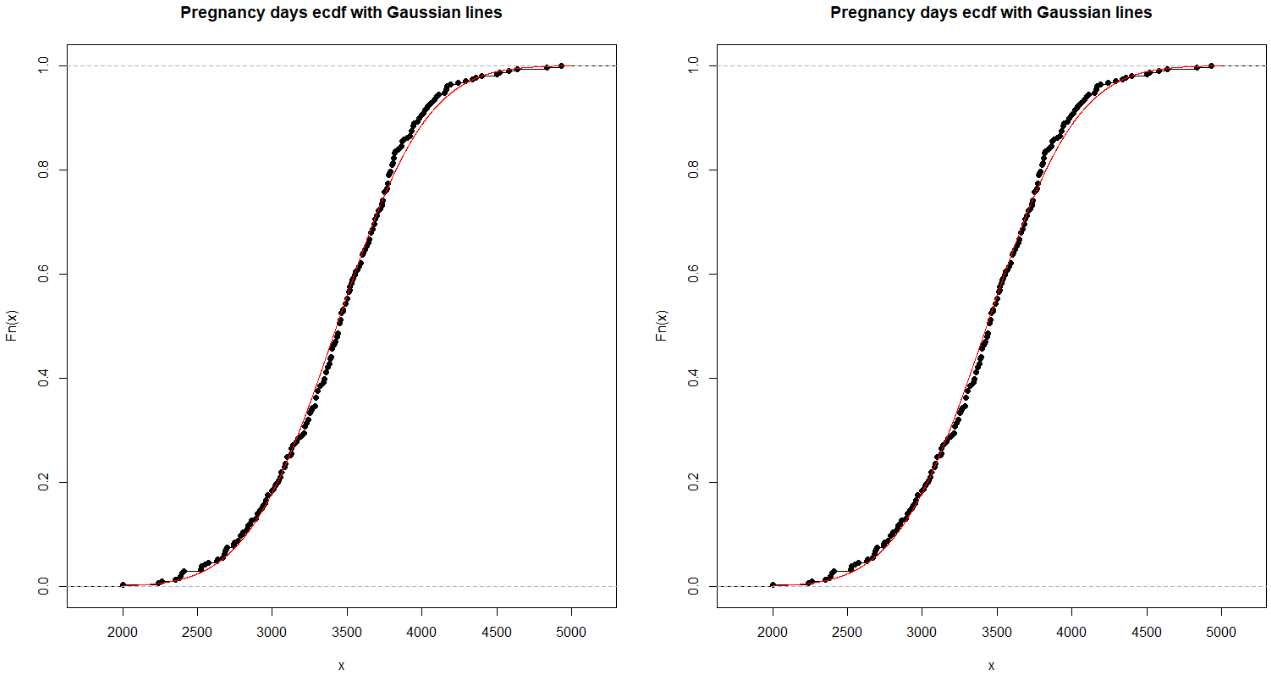


Figure 18: Gaussian fit to weight depending on sex