ETC2420 Assignment

Siming Wu (31511856)

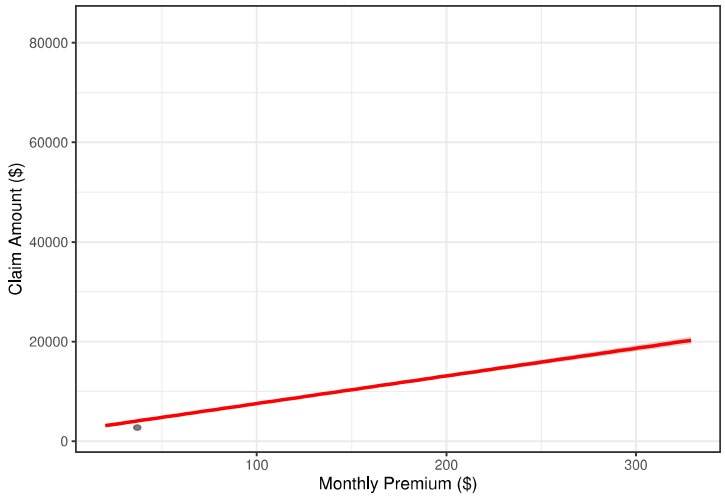
Monday, October 17 2022

Contents

Executive Summary

Task 1

The first part of this task was to investigate whether the size of the claim and monthly premium are correlated. We used R-Studio programming to plot the variables and calculate multiple descriptive statistics on the data.



From the initial scatterplot we could observe that there was a weak positive correlation between the variables, and we added a linear regression line to see this more clearly. The p-value found was extremely close to 0, which is well below the standard threshold of 0.05 and even below 0.01. From this value alone we can say with almost 100% confidence that there is some relationship between the size of the claim and the monthly premium.

To test the strength of the correlation, we calculated the Correlation Coefficient (R) and the Coefficient of Determination (R^2) We found the R value to be 0.33 and the R^2 to be 0.11. This means that 11% of the variation in Claim Amount can be explained by the variation in Monthly Premium. These statistics indicate that the size of the claim and the monthly premium are correlated, albeit a weak correlation.

For the client, this information indicates that they should be vigilant of trends in the combination of large claims for customers with large monthly premiums. The converse is also true, that those with small claim amounts might be more likely to have small monthly premiums. The analysis for these variables may seem pointless from a first glance, but our team recognises the importance of identifying correlations in the data before introducing the main variable, fraud. This ensures that in later analysis, we can accurately recommend advice without the chance that there may be other explanations in the background of our analysis.

The second part of this task was to investigate whether a higher proportion of young men (age < 27) commit fraud, compared with their young female counterparts (Age < 27). Here we completed a hypothesis test with the null and alternative hypotheses as:

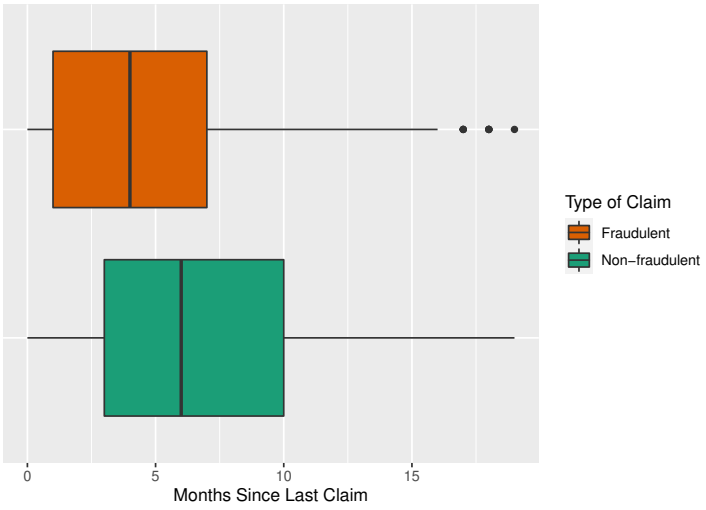
$$H_0 : p_{youngmalefraud} - p_{Youngfemalefraud} = 0 \quad \text{vs.} \quad H_1 : p_{Youngmalefraud} - p_{Youngfemalefraud} > 0,$$

We undertook a t-test, finding a p-value of 0.0114, which is below the standard threshold of 0.05, indicating that there is evidence to reject the null hypothesis in favour of the alternative.

We can also implement a randomisation test to test the hypotheses. From this test, we found a p-value of 0.012, which is also below the standard threshold of 0.05, indicating that there is evidence to reject the null hypothesis in favour of the alternative. For the client, this tells us that they should be aware that young males are more likely to commit insurance fraud than young females. This could be addressed by increasing male monthly premiums to both cover fraud losses and discourage them to commit fraud. They could also increase the frequency of fraudulent claim investigations for males under 27 years of age.

Task 2

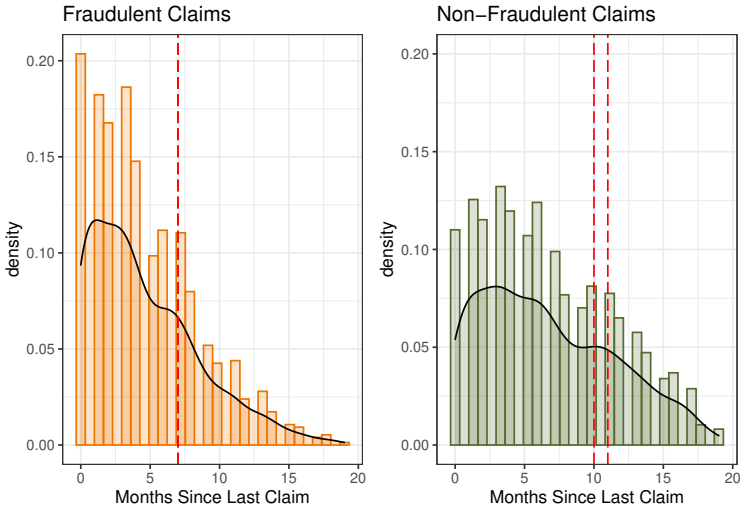
This task asked us to find an interval estimate of the third quartile of months since last claim for young drivers (age < 27) for both fraudulent and non-fraudulent claims. To get an overview of the task we can first created parallel boxplots showcasing the distributions of months since the last claim for our sample.



The third quartile for fraudulent claims was 7 months and for non-fraudulent claims it was 10 months.

We also utilised bootstrapping to calculate an interval estimate for the third quartile. We created 10000 bootstrap samples of the Months Since Last Claim for both fraudulent and non-fraudulent claims. We found that the 95% confidence interval of the third quartile for fraudulent claims was (7, 7), and for non-fraudulent claims it was (10, 11). From these simulations, we can say with 95% confidence that the third quartile of Months Since Last Claim for young fraudulent drivers was 7 months and for non-fraudulent young drivers it was 10-11 months.

We performed further investigation in the general trend of months since last claim for fraudulent and non-fraudulent claims for young drivers.



The plot for fraudulent claims indicates a positively skewed dataset, with the highest frequency occurred at zero months. The plot for non-fraudulent claims indicated a positively skewed dataset as well, although the highest frequency occurred around four months. The plots indicate the claims are more spread out for non-fraudulent and more heavily positively skewed for fraudulent drivers. The histograms and the boxplot strongly evidenced that months since last claim was less for fraudulent claims.

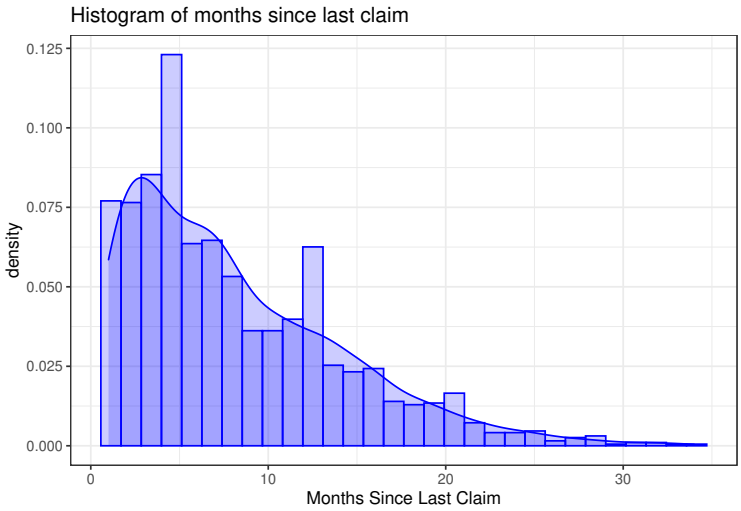
From this, we can advise the client to be vigilant for young drivers who lodge insurance claims frequently. Furthermore, they could use this frequency of claims to act as an indicator of fraudulent behaviour. We can advise to the client that months since last claim may be an indicator of fraudulent behaviour in young drivers.

Our team suspects that young fraudulent drivers lodge insurance claims more often than nonfraudulent drivers to increase their chances of a payout. This analysis is extremely important for the client as it shows that they need to be on top of fraudulent claims as they happen frequently.

Task 3

This task asked us to parameterise the distribution of the months since last claim for fraudulent cases. It is also suggested to use the Weibull distribution.

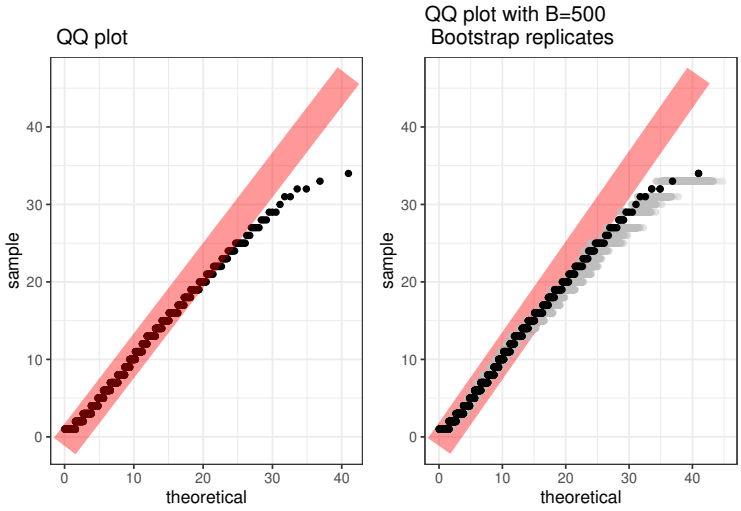
We first plotted the history of months since last claim so get an idea of the shape of the data.



From this histogram we saw that the data looks to be multimodal and positively skewed. Then we fitted the data to the Weibull distribution, with the following parameters:

term	estimate	std.error
shape	1.374384	0.0260066
scale	8.916692	0.1660858

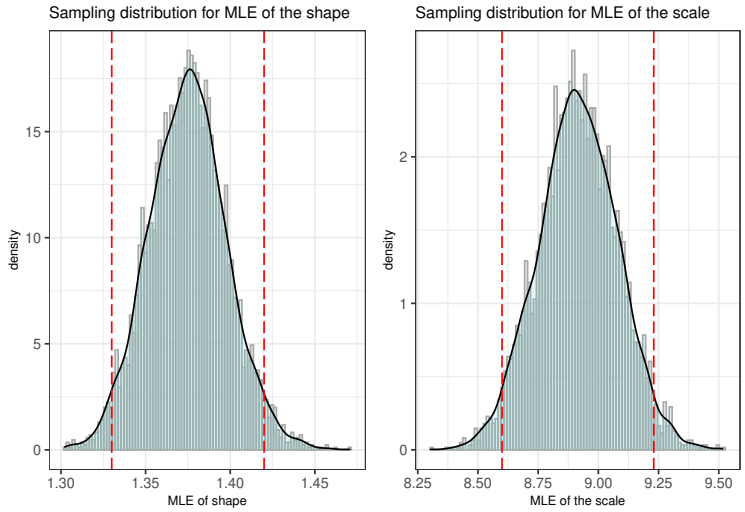
The Weibull distribution is a continuous probability distribution. It is used for particle size distribution, failure analysis, delivery time, extreme value theory and it is one of the best methods to analyse life data. It has two parameters, shape and scale, with the shape being heavily influenced by a change in scale.



When we construct a QQplot, for the most part, the Weibull distribution fits our data well. Some data is outside our "thick" line, but this may just be due to randomness. For larger differences it seems that there is a poorer fit.

We also construct bootstrap QQplots. It seems to suggest the the Weibull distribution will always struggle capturing the larger tail values. Most likely, this is the best that we can do. For the majority of the time, the Weibull fits incredibly well, so it seems like a reasonable choice.

Furthermore, we utilised MLE to show that this distribution described the data well and with high confidence. Finally, we bootstrapped the data to obtain confidence intervals for the parameters of the distribution. The intervals we found were 1.331-1.421 for the shape and 8.596-9.232 for the scale.



Task 4

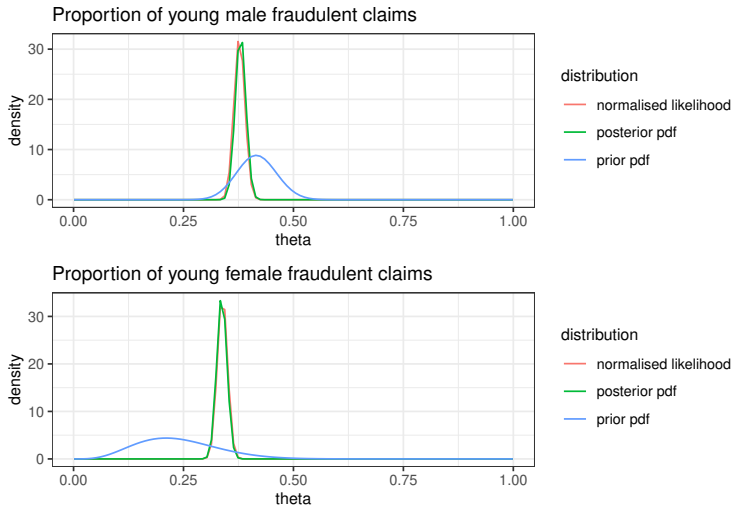
The first part of this task asked us to compare the credible intervals for the proportion of young female driver fraudulent claims with young male driver fraudulent claims. Bayesian analysis comprises mainly of a prior belief that is influenced by information, usually in the form of a distribution, which creates a posterior. The prior belief for young male drivers to commit fraud was 40% and the information was Beta(50,70). The prior belief for young female drivers to commit fraud was 20% and the information was Beta(5,16), although with less certainty. Our team used the prior and information to create posteriors for both categories.

The credible interval for the proportion of young male drive fraudulent claims is (0.3566652, 0.4034407), while the credible interval for the proportion of young female drive fraudulent claims is (0.3147928, 0.3596743)

We found the credibility factor for males was 0.927 and the credibility factor for females was 0.988. The credibility factor indicates how much the posterior was influenced by the information as opposed to the prior. As it was found to be high for both males and females, this tells us that the information greatly affects the posterior.

The next part of this task asked us to obtain and interpret the statistic that minimises squared error loss (and quantify the contribution the data makes to this). As well as the statistic that minimises the absolute error loss. The squared error loss for males was 0.3799 and 0.3373 for females. The absolute error loss for males was 0.6164 and 0.5807 for females.

To help the client visual the relationship between the prior, information and posterior, we included plots for both males and females. From these we recognised that the prior look similar, where they differ is the posterior.

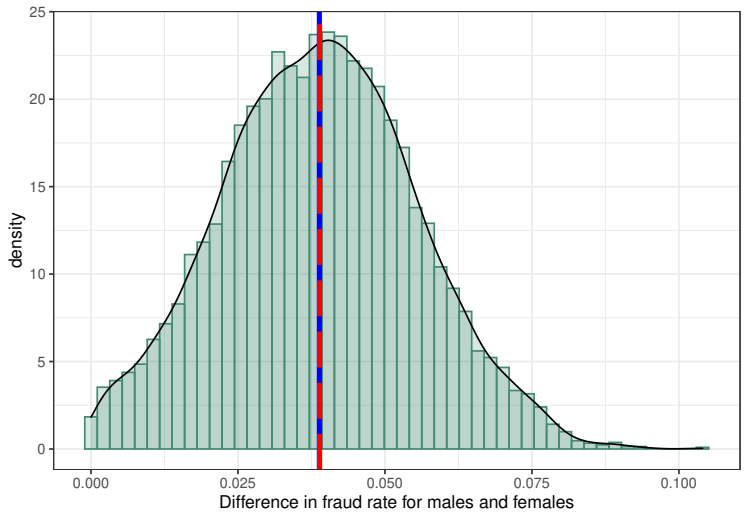


Finally, we are asked to investigate the difference in the fraud rate for the two groups, providing measures of central tendency and an indication of the probability that the difference is within 0.10.

According to our data, the difference in fraud rate for males and females is 0.039. But to test this using a randomisation method, we create 10000 bootstrap samples and calculate the difference in fraud rate for each. Then we can find the mean and median difference in fraud rate between the two groups, as well as the probability that the difference is within 0.10.

Of these 10000 bootstrap samples, the probability of a difference within 0.10 being 0.9998. This means that of the 10000 bootstrap samples we constructed, only 2 had a difference of more than 0.10. Furthermore, the median difference in fraud rate was 0.03883995 and the mean difference was 0.03887778. This indicates that the distribution of the samples is approximately symmetric.

We can then plot this information below, with the red line indicating the median difference, and the blue line showing the mean difference.



This information is valuable for the client as it indicates the probabilities of fraud in these two groups and how this belief can be influenced by new information. More specifically, future data in fraud, insurance, and young people's driving habits in general.

Technical Analysis

Task 1

For this task, we are conducting two tests. 1- The first test is to determine whether the size of the claim and monthly premium are correlated 2- The second test is to see if a higher proportion young male drivers (under 27 years of age) commit fraud than young female drivers.

Test 1

Our results from testing whether the size of the claim and monthly premium are correlated will be helpful for the company to adjust the insurance fees.

For most people, they assume there is a positive relationship between the size of claim and the monthly premium they would like to pay. For example, it is likely that a luxury car owner will have a much higher monthly premium than most other type of car owners.

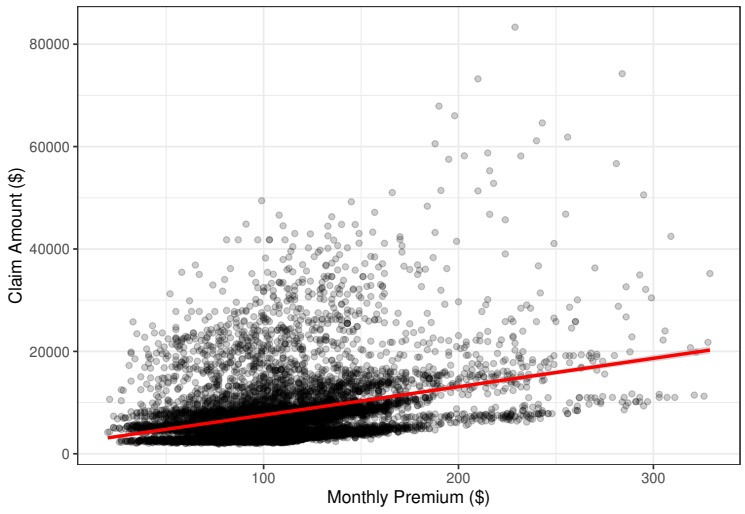
There are several ways to see whether the size of claim and monthly premium their client paid are correlated. First, it is a good idea to plot both variables of concern (size of the claim and monthly premium) against each other on a scatterplot and see if we can identify any relationship between the two.

Before doing so, we must convert Claim Amount to a numerical variable, as it is currently not supplied this way in the CIF.csv file. The following code allows this to be achieved.

```
CIF$Claim.Amount <- as.numeric(gsub(",","", CIF$Claim.Amount))
```

Now we can construct our scatterplot, also fitting a linear regression line to the figure.

```
CIF %>% ggplot(aes(x=Monthly.Premium, y=Claim.Amount)) +
        geom_point(alpha=0.2) +
        geom_smooth(method = "lm", color='red') +
        labs(x = "Monthly Premium ($)", y = "Claim Amount ($)") +
        theme_bw()
```



According to the figure above, it appears as though there may be a positive correlation between the Claim Amount and Monthly Premium. However, it seems to be quite weak. This is only an assumption though, so it is best to find a numerical result that can explain the relationship between these two variables.

To test the strength of the correlation, we calculated the Correlation Coefficient (R) and the Coefficient of Determination (R^2)

```
cor(CIF$Monthly.Premium, CIF$Claim.Amount)
```

```
## [1] 0.3310577
```

Our assumption is confirmed with the result of the Correlation Coefficient. We find the Correlation Coefficient to be +0.33 correlation between the claim amount and the monthly premium - which is a weak positive relationship.

```
CIF_lm <- CIF %>% lm(formula = Claim.Amount ~ Monthly.Premium)
summary(CIF_lm)$r.squared
```

```
## [1] 0.1095992
```

We found the R^2 value to be 0.11. This means that 11% of the variation in claim amount can be explained by the variation in monthly premium. These statistics indicate that the size of the claim and the monthly premium are correlated, albeit a weak correlation.

```
CIF_lm <- CIF %>% lm(formula = Claim.Amount ~ Monthly.Premium)
tidy(CIF_lm) %>% kable()
```

term	estimate	std.error	statistic	p.value
(Intercept)	2001.06673	191.496638	10.44962	0
Monthly.Premium	55.47243	1.654563	33.52694	0

The p-value found was extremely close to 0, which is well below the standard threshold of 0.05 and even below 0.01. From this value alone we can say with almost 100% confidence that there is some relationship between the size of the claim and the monthly premium.

Test 2

The second part of this task was to investigate whether a higher proportion of young men (age < 27) commit fraud, compared with their young female counterparts (Age < 27). Here we completed a hypothesis test with the null and alternative hypotheses as:

$$H_0 : p_{Youngmalefraud} - p_{Youngfemalefraud} = 0 \quad \text{vs.} \quad H_1 : p_{Youngmalefraud} - p_{Youngfemalefraud} > 0,$$

Decision rule: reject H_0 if p-value $< 5\%$ level, in favour on H_1.

```
CIF_Young <- CIF %>% filter(Age<27)

CIF_Yfemales <- CIF_Young %>% filter(Gender == "F") %>% pull(Fraud)
CIF_Yfemales <- as.numeric(CIF_Yfemales)

CIF_Ymales <- CIF_Young %>% filter(Gender == "M") %>% pull(Fraud)
CIF_Ymales <- as.numeric(CIF_Ymales)

ttest <- t.test(x = CIF_Ymales, y = CIF_Yfemales, alternative = 'greater') %>%
  tidy() %>% kable() %>% kable_styling(latex_options= c("scale_down", "HOLD_position"))

ttest
```

estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	method	alternative
0.0385495	0.3770385	0.338489	2.277098	0.0114234	3169.278	0.0106953	Inf	Welch Two Sample t-test	greater

From this test, we found a p-value of 0.0114, which is below the standard threshold of 0.05, indicating that there is evidence to reject the null hypothesis in favour of the alternative.

We can also implement a randomisation test to test the hypotheses. The calculations to implement the randomisation test in this setting are shown in the **R** code chunk below. We want to "break" the association between fraud rate and gender by permuting, or shuffling one of the columns. This gives us an approximate sampling distribution of the difference in proportions under H_0, where there is no association between gender and fraud rate. Since we are "shuffling" our observed data, we must sample **without replacement**. So we should get R *xobs* (proportion difference) values created assuming that there is no association between gender and fraud rate, which we can plot with a histogram. Our p-value will be the number of these simulated proportion differences that are greater than the difference we observed in our data.

```
Fraud <- tibble(gender = c(rep('male', length(CIF_Ymales)),
                           rep('female', length(CIF_Yfemales))),
                claim = c(rep('Fraudulent', sum(CIF_Ymales)),
                          rep('NonFraudulent', length(CIF_Ymales)-sum(CIF_Ymales)),
                          rep('Fraudulent', sum(CIF_Yfemales)),
                          rep('NonFraudulent', length(CIF_Yfemales)-sum(CIF_Yfemales))))
Fraud2 <- Fraud %>%
    group_by(gender, claim) %>%
    tally() %>%
    ungroup() %>%
    pivot_wider(names_from = claim, values_from = n) %>%
    mutate(total = Fraudulent + NonFraudulent) %>%
    mutate(prop = round(Fraudulent/total, digits=3))
Fraud3 <- Fraud2 %>%
  arrange(desc(row_number()))

xobs <- Fraud3$prop[Fraud3$gender=='male'] - Fraud3$prop[Fraud3$gender=='female']
n <- nrow(Fraud)
R <- 1000
Rxobs <- array(dim = R)
set.seed(43)
RFraud <- Fraud
for (r in 1:R) {
  RFraud <- RFraud %>% mutate(gender = sample(RFraud$gender, n, replace=FALSE))
  RFraud2 <- RFraud %>%
    group_by(gender, claim) %>%
    tally() %>%
    ungroup() %>%
    pivot_wider(names_from = claim, values_from = n) %>%
    mutate(total = Fraudulent + NonFraudulent) %>%
    mutate(prop = round(Fraudulent/total, digits=3))
  Rxobs[r] <- RFraud2$prop[Fraud2$gender=='male'] - RFraud2$prop[Fraud2$gender=='female']
}

pval <- sum(Rxobs >= xobs)/R
pval
```

```
## [1] 0.012
```

We will reject the H_0 that there is no difference of proportion on fraud claim between male and female, since the p-value 1.2% < significant level 5%.

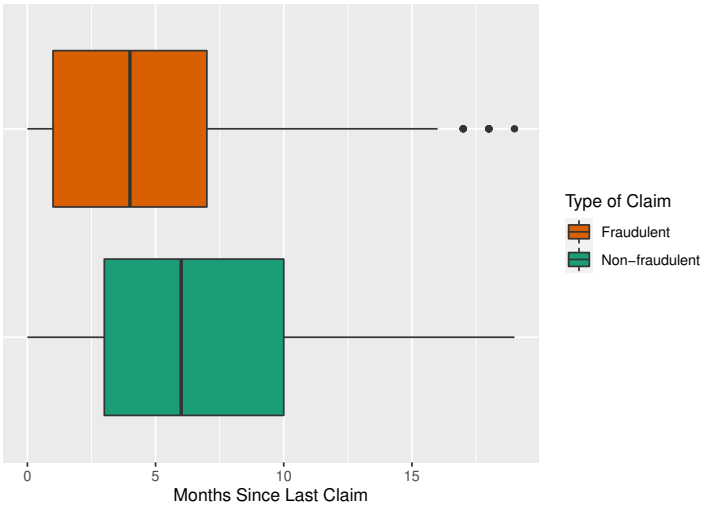
Therefore we will be in favour of our alternative hypothesis proportion of male is higher than female To conclude, the company should set higher conditions to claim for young clients (27< years old), and take slightly more seriously on male clients' claims in order to lower the loss from fraud claims.

Task 2

The client would like an interval estimate of the third quartile of months since last claim for young drivers for both fraudulent and non-fraudulent claims.

Looking on the third quartile for fraudulent and non-fraudulent claims we may find the differences between them. If there is a considerable difference in the third quartile of months since last claims between fraudulent and non-fraudulent claims, this will be a good consideration whether the future claims may be fraudulent or not.

```
ggplot(CIF_Young, aes(y = factor(Fraud), x = Months.Since.Last.Claim, fill=factor(Fraud))) +
    geom_boxplot() +
    theme(axis.text.y=element_blank(), #remove x axis labels
        axis.ticks.y=element_blank())+ #remove x axis ticks
    labs(x = "Months Since Last Claim", y = "") +
    scale_fill_brewer(palette="Dark2", guide = guide_legend(reverse=TRUE),
        name = "Type of Claim", labels = c("Non-fraudulent", "Fraudulent"))
```



Looking at the parallel box plots at the third quartiles, quartile 3 for fraudulent claims is 7 months since last claim, and quartile 3 for non-fraudulent claims is 10 months since last claim. This possibly means that clients with fraudulent claims will issue claims more regularly.

To find an interval estimate of the third quartile for both fraudulent and non-fraudulent claims, we can construct 10000 bootstrap samples of the Months Since Last Claim for both fraudulent and non-fraudulent claims, and then find the 95% confidence interval of the third quartile for both.

```
CIF_YFraudulent <- CIF_Young %>% filter(Fraud == 1)
x_YFraudulent = CIF_YFraudulent$Months.Since.Last.Claim
n_YFraudulent <- nrow(CIF_YFraudulent)
```

```
CIF_YNonFraudulent <- CIF_Young %>% filter(Fraud == 0)
x_YNonFraudulent = CIF_YNonFraudulent$Months.Since.Last.Claim
n_YNonFraudulent <- nrow(CIF_YNonFraudulent)

set.seed(43)
B <- 10000
x_3quartileFRAUD <- rep(NA, B)
x_3quartileNONFRAUD <- rep(NA, B)

for (i in 1:B) {
tempFRAUD <- sample(x_YFraudulent, size = n_YFraudulent, replace = TRUE)
tempNONFRAUD <- sample(x_YNonFraudulent, size = n_YNonFraudulent, replace = TRUE)
x_3quartileFRAUD[i] <- quantile(tempFRAUD, probs = 0.75)
x_3quartileNONFRAUD[i] <- quantile(tempNONFRAUD, probs = 0.75)
}

boot.CI_YFraudulent <- quantile(x_3quartileFRAUD, c(0.025, 0.975))
boot.CI_YNonFraudulent <- quantile(x_3quartileNONFRAUD, c(0.025, 0.975))
```

```
#FRAUDULENT CLAIMS
boot.CI_YFraudulent
```

```
##  2.5% 97.5%
##     7     7
```

```
#NON-FRAUDULENT CLAIMS
boot.CI_YNonFraudulent
```

```
##  2.5% 97.5%
##    10    11
```

We found that the 95% confidence interval of the third quartile for fraudulent claims was (7, 7), and for non-fraudulent claims it was (10, 11). From these simulations, we can say with 95% confidence that the third quartile of Months Since Last Claim for young fraudulent drivers was 7 months and for non-fraudulent young drivers it was 10-11 months.

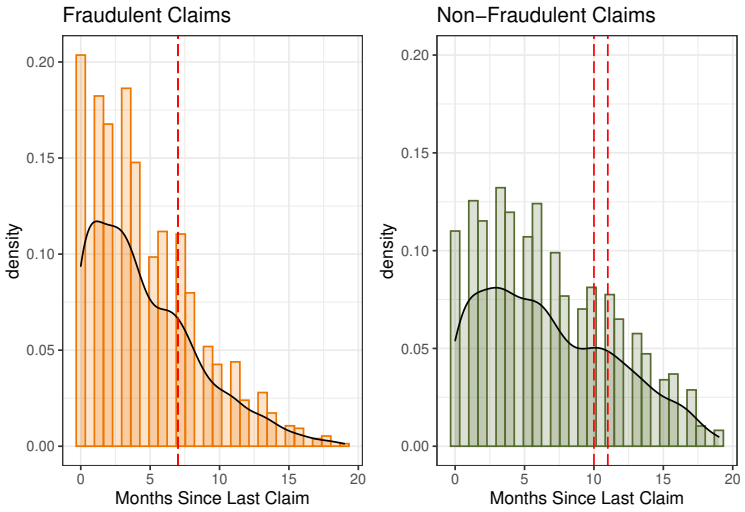
We performed further investigation in the general trend of months since last claim for fraudulent and non-fraudulent claims for young drivers.

To get more information, we can then construct density plotss with these third quartile interval estimates. This will be helpful to seek the distribution of Months Since Last Claim for both Fraudulent and Non-Fraudulent claims.

```
dt_data <- tibble(x = x_YFraudulent)
p1_data <- dt_data %>%
  ggplot(aes(x = x, y = ..density..)) +
  geom_histogram(colour = "darkorange2", fill = "darkorange2", alpha = 0.2) +
  geom_density(colour = "black", fill = "darkorange2", alpha = 0.2) + theme_bw() +
  geom_vline(xintercept = boot.CI_YFraudulent[1], colour = "red", linetype=5) +
  geom_vline(xintercept = boot.CI_YFraudulent[2], colour = "red", linetype=5) +
  labs(x = "Months Since Last Claim") +
  ggtitle("Fraudulent Claims")
```

```
dt_data <- tibble(x = x_YNonFraudulent)
p2_data <- dt_data %>%
  ggplot(aes(x = x, y = ..density..)) +
  geom_histogram(colour = "darkolivegreen", fill = "darkolivegreen", alpha = 0.2) +
  geom_density(colour = "black", fill = "darkolivegreen", alpha = 0.2) + theme_bw() +
  geom_vline(xintercept = boot.CI_YNonFraudulent[1], colour = "red", linetype=5) +
  geom_vline(xintercept = boot.CI_YNonFraudulent[2], colour = "red", linetype=5) +
  labs(x = "Months Since Last Claim") +
  ggtitle("Non-Fraudulent Claims")+
  ylim(0,0.2)

grid.arrange(p1_data, p2_data, ncol = 2)
```



The plot for fraudulent claims indicates a positively skewed dataset, with the highest frequency occurred at zero months. The plot for non-fraudulent claims indicated a positively skewed dataset as well, although the highest frequency occurred around four months. The plots indicate the claims are more spread out for non-fraudulent and more heavily positively skewed for fraudulent drivers.

The histograms and the boxplot strongly evidenced that months since last claim was less for fraudulent claims. From this, we can advise the client to be vigilant for young drivers who lodge insurance claims frequently. Furthermore, they could use this frequency of claims to act as an indicator of fraudulent behaviour. We can advise to the client that months since last claim may be an indicator of fraudulent behaviour in young drivers.

Our team suspects that young fraudulent drivers lodge insurance claims more often than nonfraudulent drivers to increase their chances of a payout. This analysis is extremely important for the client as it shows that they need to be on top of fraudulent claims as they happen frequently.

Task 3

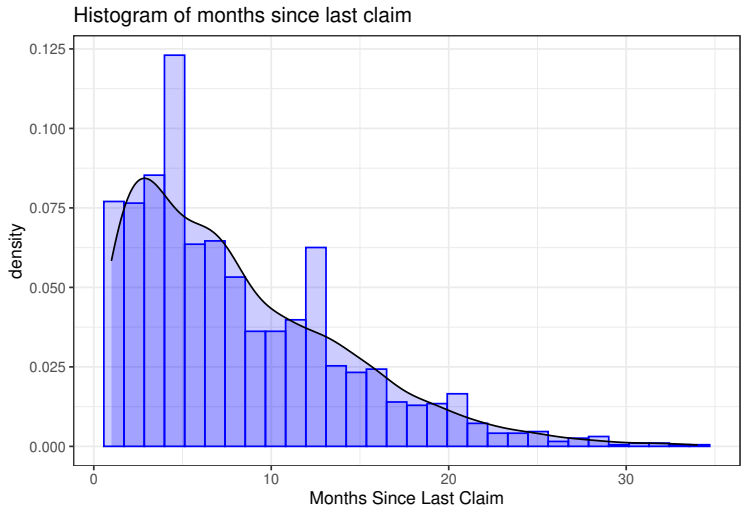
The customer needs a parameterized expression representing the distribution of months since the last claim in the fraud case. Therefore, for task 3, it needs to redefine the dataset first. Otherwise, it is also suggested to use weibull distribution. We attempt to justify why Weibull Distribution can be fit for this distribution by its characteristics.

We use the filter function filter(Fraud == 1) to filter(Months.Since.Last.Claim != 0) because we focus on the distribution of the "months since last claim" and exclude non-zero values. Then we plotted the history

of the months since the last claim to get a sense of the shape of the data, and titled it "Histogram of months since last claim" by using hist plot and density plot from ggplot2.

```
CIF_T3 <- CIF %>% filter(Months.Since.Last.Claim != 0, Fraud == 1)
x <- CIF_T3$Months.Since.Last.Claim
n <- nrow(CIF_T3)
dt <- tibble(id = 1:n, x = x)

p1_pdf <- dt %>% ggplot(aes(x=x, y=..density..)) +
geom_histogram(colour="blue", fill="blue", alpha=0.2) +
geom_density(colour="black", fill="blue", alpha=0.2) +
ggtitle("Histogram of months since last claim") +
xlab("Months Since Last Claim") + theme_bw()
p1_pdf
```



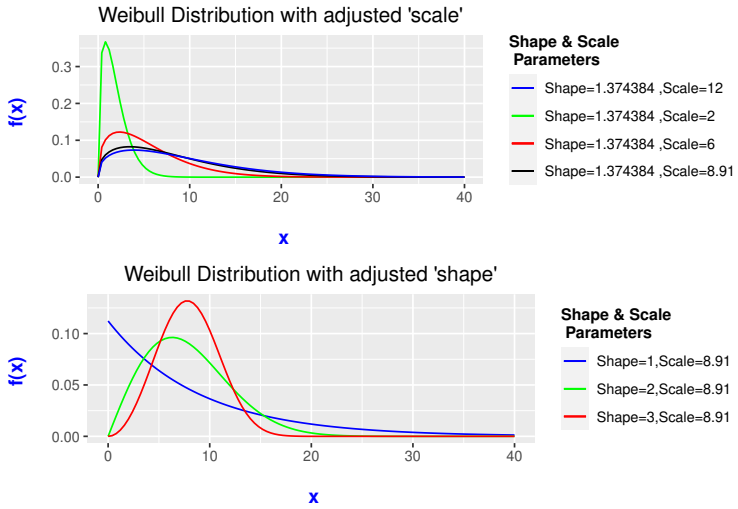
We can use tidy to summarise the parameters of weibull distribution

```
weibull_fit <- fitdistr(x, "weibull")
weibull_fit %>% tidy() %>% kable()
```

term	estimate	std.error
shape	1.374384	0.0260066
scale	8.916692	0.1660858

The shape for our Weibull distribution is 1.374384 and the scale is 8.916691, with the estimated standard errors quite small which indicates a good fit.

The Weibull distribution is a continuous probability distribution. It is used for particle size distribution, failure analysis, delivery time, extreme value theory and it is one of the best methods to analyse life data. It has two parameters, shape and scale, with the shape being heavily influenced by a change in scale. The plots below showcase how the distribution is impacted as the parameters are adjusted.



We attempt to observe to what degree the distribution is approximate to Weibull distribution, so QQ plot is introduced.

The upper part of the following code generates a vector of the same size of dt that follows a Weibull distribution and creates a QQ plot for this dataset to verify that it does indeed follow a normal distribution for this example.

In order to remove the bias, create a QQ plot for bootstrap sampling. The purpose of setseed is to make the sample reproductive. Using the loop for 500 times enables generating 500 samples from x.

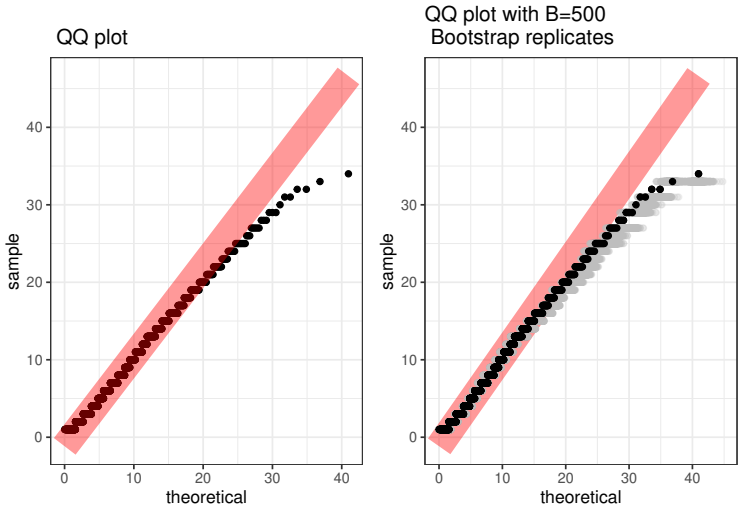
We also construct bootstrap QQplots. It seems to suggest the the Weibull distribution will always struggle capturing the larger tail values. Most likely, this is the best that we can do. For the majority of the time, the Weibull fits incredibly well, so it seems like a reasonable choice.

```
params <- weibull_fit$estimate
p <- ggplot(dt, aes(sample = x)) +
  stat_qq(distribution = qweibull, dparams = params) +
  stat_qq_line(distribution = qweibull,
               dparams = params, color = "red",
               size=8, alpha=0.4) + #increased line size
  theme(aspect.ratio = 1) + theme_bw() +
  ggtitle("\n QQ plot") +
  labs(x = 'theoretical', y='sample')
p1 <- p

set.seed(43)
MLE.x <- weibull_fit$estimate # point estimate
boot.seq <- seq(1,n,1)/n-1/(2*n)
B <- 500
MLE.x_boot <- matrix(rep(NA,2*B), nrow=B, ncol=2)
for(i in 1:B){
  temp <- sample(dt$x, size=n, replace=TRUE)
  dt <- dt %>% mutate(temp=temp)
  MLE.x_boot[i,] <- fitdistr(temp, "weibull")$estimate
  params_boot <- MLE.x_boot[i,]
  p <- p + stat_qq(aes(sample=temp), distribution = qweibull,
                   dparams = params_boot, colour="grey",
                   alpha=0.2)
}
```

```
p <- p + stat_qq(aes(sample=x), distribution = qweibull,
                 dparams = params) +
  ggtitle("QQ plot with B=500 \n Bootstrap replicates")
p2 <- p

grid.arrange(p1, p2, ncol = 2)
```



When we construct a QQplot, the Weibull distribution fits our data well in most cases. Some of the data went beyond our "thick" line, but that could just be due to randomness. For large differences, there seems to be a poor fit.This is the same case for the bootstrap QQplots. It seems to suggest the the Weibull distribution will always struggle capturing the larger tail values. Most likely, this is the best that we can do. For the majority of the time, the Weibull fits incredibly well, so it seems like a reasonable choice.

Furthermore, we utilised MLE to show that this distribution described the data well and with high confidence. We bootstrapped the data to obtain confidence intervals for the parameters of the distribution. The following code is used to bootstrap to estimate the true parameters scale and shape.

```
set.seed(43)
B <- 5000
MLE.x_boot <- matrix(rep(NA,2*B), nrow=B, ncol=2)
for(i in 1:B){
temp <- sample(dt$x, size=n, replace=TRUE)
MLE.x_boot[i,] <- fitdistr(temp, "weibull")$estimate
}
```

```
boot.CI.shape <- quantile(MLE.x_boot[,1], c(0.025, 0.975))
boot.CI.shape
```

```
##     2.5%    97.5%
## 1.330865 1.421220
```

```
boot.CI.scale <- quantile(MLE.x_boot[,2], c(0.025, 0.975))
boot.CI.scale
```

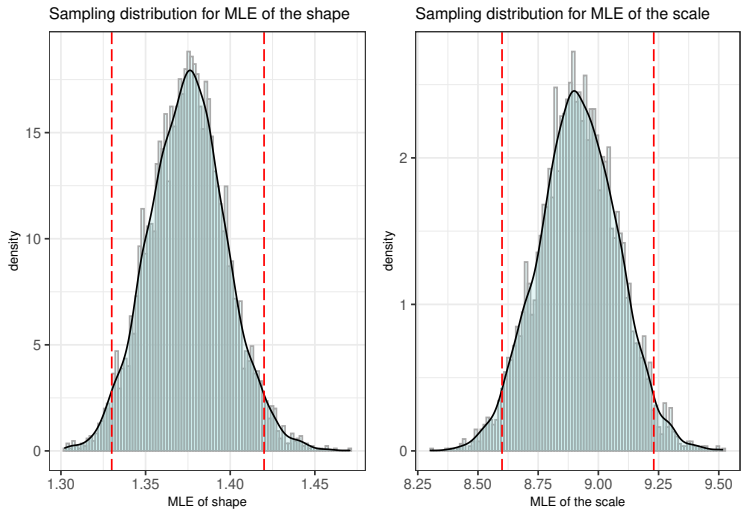
```
##     2.5%    97.5%
## 8.596139 9.232320
```

The intervals we found were 1.331-1.421 for the shape and 8.596-9.232 for the scale.

Use the new function named bootplot.f to plot the density estimate (overlayed histogram and kernel density plot) corresponding to the empirical Bootstrap distribution of parameters, and the position of the 2.5 and 97.5% quantiles of this distribution for both scale and shape with a red vertical line.

```
########### the bootplot.f function ############
bootplot.f<- function(stat.boot, bins=50){
df <- tibble(stat = stat.boot)
CI <- round(quantile(stat.boot, c(0.025, 0.975)), 2)
p <- df %>% ggplot(aes(x=stat, y=..density..)) +
geom_histogram(bins=bins, colour="darkgrey", fill="darkslategray3", alpha=0.2) +
geom_density(fill="darkslategray3", colour="black", alpha=0.2) +
geom_vline(xintercept = CI, colour = "red", linetype=5) +
theme_bw()
return(p)
}
######## end of bootplot.f function #######
```

```
p_MLEboot.shapeCIF_T3 <- bootplot.f(MLE.x_boot[, 1], bins = 100) +
ggtitle("Sampling distribution for MLE of the shape") +
  xlab("MLE of shape") +
  theme(title = element_text(size = 8))
p_MLEboot.scaleCIF_T3 <- bootplot.f(MLE.x_boot[, 2], bins = 100) +
  ggtitle("Sampling distribution for MLE of the scale") +
  xlab("MLE of the scale") +
  theme(title = element_text(size = 8))
grid.arrange(p_MLEboot.shapeCIF_T3, p_MLEboot.scaleCIF_T3, ncol = 2)
```



To sum up, sampling distribution is very close to Weibull distribution in most cases, so we think it is a good choice.

Task 4

As mentioned earlier, the client would like to utilise more Bayesian analysis. They would like you to compare the credible intervals for the proportion of young female driver fraudulent claims with young male driver

fraudulent claims. The client believes strongly that close to 40% of young male drivers commit fraud, suggesting a Beta(50,70) distribution. They believe around 20% of young female drivers commit fraud, but are not as sure about this and suggest a Beta(5,16) distribution.

The first step to perform Bayesian analysis, with the assistance of R-Studio, is to create a function. As the distributions for both males and females are Beta distributions, this function is a Beta function.

```
beta_binomial <- function(n, x, alpha = 1, beta = 1) {
atil <- alpha + x
btil <- beta + n - x
cf <- n/(alpha + beta + n)
out <- list(alpha_tilde = atil, beta_tilde = btil,
credibility_factor = cf)
return(out)
}
```

Now we use this function to determine the posterior for males. We input the information given in the question and get an output in R.

```
n_Ymales <- length(CIF_Ymales)
xobs_Ymales <- sum(CIF_Ymales)
alpha_Ymales <- 50
beta_Ymales <- 70
MALEout <- beta_binomial(n = n_Ymales, x = xobs_Ymales,
                         alpha = alpha_Ymales, beta = beta_Ymales)

MALEout
```

```
## $alpha_tilde
## [1] 628
##
## $beta_tilde
## [1] 1025
##
## $credibility_factor
## [1] 0.9274047
```

The posterior distribution is Beta(628, 1025) with a credibility factor of 0.927.

We can now also calculate the posterior value

```
thetaxx <- seq(0.001, 0.999, length.out = 100)
mean(qbeta(thetaxx, MALEout$alpha_tilde, MALEout$beta_tilde))
```

```
## [1] 0.3799194
```

The prior belief was that 40% of young male drivers submit fraudulent claims. The posterior value is 38%, calculated from the expected value of the posterior Beta distribution. This is only slightly lower than the prior belief, indicating that the client's prior belief is very similar to the posterior.

We then perform the same analysis for female drivers

```r
n_Yfemales <- length(CIF_Yfemales)
xobs_Yfemales <- sum(CIF_Yfemales)
alpha_Yfemales <- 5
beta_Yfemales <- 16
FEMALEout <- beta_binomial(n = n_Yfemales, x = xobs_Yfemales,
                           alpha = alpha_Yfemales, beta = beta_Yfemales)

FEMALEout
```

```
## $alpha_tilde
## [1] 574
##
## $beta_tilde
## [1] 1128
##
## $credibility_factor
## [1] 0.9876616
```

The posterior distribution is Beta(574, 1128) with a credibility factor of 0.988.

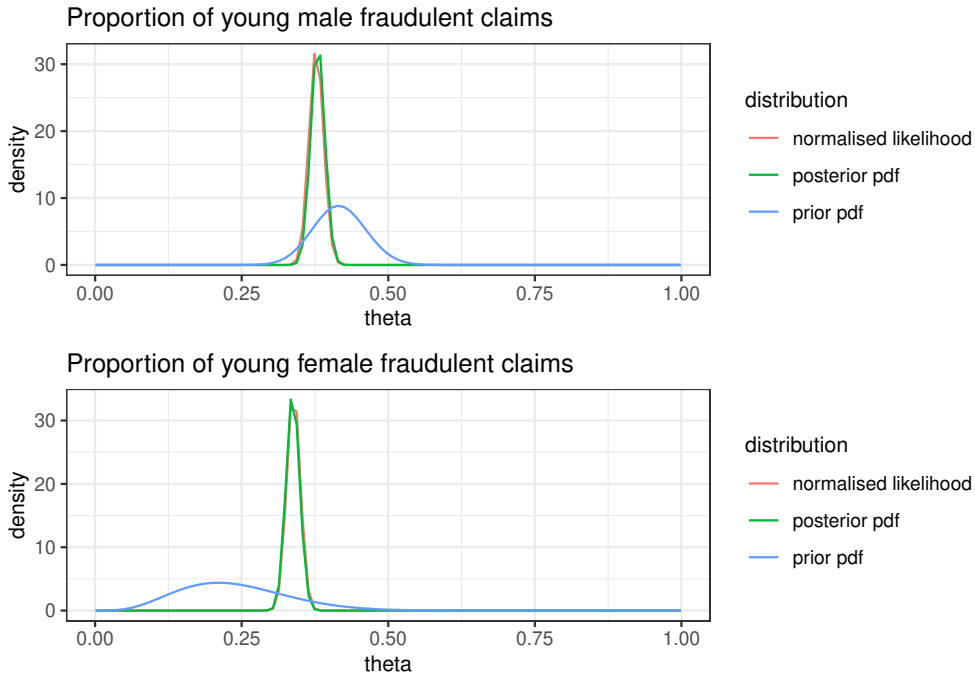
We can now also calculate the posterior value

```r
thetaxx <- seq(0.001, 0.999, length.out = 100)
mean(qbeta(thetaxx, FEMALEout$alpha_tilde, FEMALEout$beta_tilde))
```

```
## [1] 0.3372556
```

The prior belief was that 20% of young female drivers submit fraudulent claims, although the client was not as sure about this as the male %. The posterior value is 34%, much higher than 20%, suggesting that the incidence of fraud in young female drivers is higher than the client originally believed. As both credibility factors are quite high, it suggests that with a different distribution, or the same Beta distribution with different values for alpha and beta, that the posterior value could be different.

To visualise the process of Bayesian analysis, we have created a plot for both the male and female data.



We can see in blue the prior beliefs of 20% and 40% fraud rate. We can see that the information/normalised likelihood and the posterior pdf looks quite similar as both are Beta distributions. We can also see that the density of the prior for males has lower variability than that of females, this may have caused the difference in the distance between the prior and posterior % value.

The second part of this task involves analysis of the MSE and MAE. First we created a function that calculates the mean squared error and mean absolute error. Then we used the posterior data from above to calculate the four statistics.

```
beta_sqe_abe <- function(alpha, beta){

  mean_min_sqe <- alpha/(alpha+beta)
  mean_min_abe <- sqrt(mean_min_sqe)
  out <- list(mean_min_sqe=mean_min_sqe, mean_min_abe=mean_min_abe)
  return(out)
}
```

```
#Male statistics
MALE_posterior_meanvar <- beta_sqe_abe(alpha=MALEout$alpha_tilde,
                                       beta=MALEout$beta_tilde)
MALE_posterior_meanvar
```

```
## $mean_min_sqe
## [1] 0.3799153
##
## $mean_min_abe
## [1] 0.6163727
```

```
#Female statistics
FEMALE_posterior_meanvar <- beta_sqe_abe(alpha=FEMALEout$alpha_tilde,
                                         beta=FEMALEout$beta_tilde)
FEMALE_posterior_meanvar
```

```
## $mean_min_sqe
## [1] 0.3372503
##
## $mean_min_abe
## [1] 0.5807325
```

The statistic that minimises the squared error loss for males is 0.3799, and 0.3373 for females. The statistic that minimises the absolute error loss for males is 0.6164, and 0.5807 for females. The MSE indicates how accurate the model is at predicting, for this data. As both the MSE are reasonably low, we can say that the posterior Beta distributions accurately described the data. We can see that the female distribution has a lower MSE and therefore that posterior distribution was more accurate than the male's. The MSE here is also the mean for each distribution. The MAE is the statistic that minimises the absolute error loss and is usually the median. The MAE for females is lower than males suggesting that the Beta distribution fits that data better than the males, but only slightly.

The last part of this task asks for the difference in the fraud rate for young males and females. We need to provide measures of central tendency and an indication of the probability that the difference is within 0.10. As the fraud variable is a binomial or "dummy" variable, the mean will represent the fraud rate for both young females and young males.

Additionally, we have been asked to investigate the difference in the fraud rate for the two groups, providing them with measures of central tendency and an indication of the probability that the difference is within 0.10.

```
#difference in fraud rate
mean(CIF_Ymales) - mean(CIF_Yfemales)
```

```
## [1] 0.03854949
```

According to our data, the difference in fraud rate for males and females is 0.039.

To test this, we can create 10000 bootstrap samples and calculate the difference in fraud rate for each.

```
set.seed(43)
B <- 10000
nfem = length(CIF_Yfemales)
nmale = length(CIF_Ymales)
mf_diff = array(dim = B)
for(i in 1:B){
CIF_YfemalesBOOT <- sample(CIF_Yfemales, size=nfem, replace=TRUE)
CIF_YmalesBOOT <- sample(CIF_Ymales, size=nmale, replace=TRUE)
mf_diff[i] <- abs(mean(CIF_YmalesBOOT) - mean(CIF_YfemalesBOOT))
}
```

Then we can find the mean and median difference in fraud rate between the two groups, as well as the probability that the difference is within 0.10.

```
#Mean difference
xbar_boot <- mean(mf_diff)
xbar_boot
```

```
## [1] 0.03883995
```

```r
#Median difference
xmed_boot <- median(mf_diff)
xmed_boot
```

```
## [1] 0.03887778
```

```r
#Probability that the difference is within 0.10
proba <- sum(mf_diff <= 0.10)/B
proba
```

```
## [1] 0.9998
```

Of these 10000 bootstrap samples, the probability of a difference within 0.10 being 0.9998. This means that of the 10000 bootstrap samples we constructed, only 2 had a difference of more than 0.10. Furthermore, the median difference in fraud rate was 0.03883995 and the mean difference was 0.03887778.This indicates that the distribution of the samples is approximately symmetric.

We can then plot this information below, with the red line indicating the median difference, and the blue line showing the mean difference.

```r
df <- tibble(stat = mf_diff)

p <- df %>% ggplot(aes(x=stat, y=..density..)) +
    geom_histogram(bins=50, colour="aquamarine4", fill="aquamarine4", alpha=0.2) +
    geom_density(fill="aquamarine4", colour="black", alpha=0.2) +
    geom_vline(xintercept = xbar_boot, colour = "blue", linetype=1, size =1.5) +
    geom_vline(xintercept = xmed_boot, colour = "red", linetype=5, size = 1.5) +
    labs(x = "Difference in fraud rate for males and females") +
    theme_bw()

p
```