Analysis

Data Preparation

```r
voice <- read.csv("../../data/gender/balanced_train.csv")
head(voice)
```

```
##   meanfreq       sd   median      Q25      Q75         skew   sp.ent      sfm
## 1 8.153891 8.570102 8.002904 7.328629 9.291727 -0.199356530 3.369166 2.584677
## 2 7.846562 8.423659 7.689777 7.146453 9.074857 -0.007415137 3.253375 2.378387
## 3 7.637648 8.369293 7.497563 6.976269 8.985667 -0.016312126 3.214666 2.115166
## 4 7.542351 8.426862 7.100093 6.743659 8.928714 -0.054684730 3.160715 2.180566
## 5 7.681082 8.358729 7.607353 7.034379 8.995721 -0.124070090 3.151379 2.457708
## 6 7.584942 8.456333 6.927504 6.782198 8.959107 -0.167355900 3.146243 2.412977
##    meanfun gender
## 1 3.817343   male
## 2 3.183698   male
## 3 3.052549   male
## 4 2.337924   male
## 5 2.251824   male
## 6 2.393773   male
```

```r
male_data <- voice[voice$gender == "male", ]
female_data <- voice[voice$gender == "female", ]
head(male_data)
```

```
##   meanfreq       sd   median      Q25      Q75         skew   sp.ent      sfm
## 1 8.153891 8.570102 8.002904 7.328629 9.291727 -0.199356530 3.369166 2.584677
## 2 7.846562 8.423659 7.689777 7.146453 9.074857 -0.007415137 3.253375 2.378387
## 3 7.637648 8.369293 7.497563 6.976269 8.985667 -0.016312126 3.214666 2.115166
## 4 7.542351 8.426862 7.100093 6.743659 8.928714 -0.054684730 3.160715 2.180566
## 5 7.681082 8.358729 7.607353 7.034379 8.995721 -0.124070090 3.151379 2.457708
## 6 7.584942 8.456333 6.927504 6.782198 8.959107 -0.167355900 3.146243 2.412977
##    meanfun gender
## 1 3.817343   male
## 2 3.183698   male
## 3 3.052549   male
## 4 2.337924   male
## 5 2.251824   male
## 6 2.393773   male
```

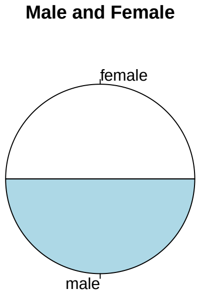
```r
head(female_data)
```

```
##     meanfreq       sd   median      Q25      Q75       skew   sp.ent      sfm
## 726 8.166690 8.581521 8.005468 7.456112 9.171794  0.1165690 3.291912 2.286142
## 727 8.340455 8.599874 8.367010 7.550676 9.296616  0.0456077 3.355465 2.494345
## 728 8.056571 8.546849 7.867097 7.333423 9.110954 -0.1614750 3.272155 2.239860
## 729 8.267929 8.654711 7.734899 7.329590 9.392445  0.1786457 3.301512 2.239053
## 730 7.695245 8.401543 7.485312 6.902178 9.032689  0.2401945 3.160920 2.571130
## 731 8.058577 8.526025 7.825418 7.218233 9.224732  0.2416200 3.262145 2.565457
##      meanfun gender
## 726 3.564867 female
## 727 3.819150 female
## 728 3.372699 female
## 729 3.160863 female
## 730 2.638777 female
## 731 3.231355 female
```

Data Analysis

Gender

```
pie(table(voice$gender), main = "Male and Female")
```



```
visualize_data <- function(column) {
  # return(male_data[column])
  hist(male_data[[column]], xlab = column, col = MALE_COLOR, prob = TRUE, breaks = 80, border = "white"
  hist(female_data[[column]], xlab = column, col = FEMALE_COLOR, prob = TRUE, add = TRUE, breaks = 80, 

  # Calculate and plot KDE for male data
  male_density <- density(male_data[[column]])
  lines(male_density, col = "blue", lwd = 2)

  # Calculate and plot KDE for female data
  female_density <- density(female_data[[column]])
  lines(female_density, col = "red", lwd = 2)

  legend("topright", legend = c("Male", "Female"), col = c("blue", "red"), lwd = 2, fill = c(MALE_COLOR
}
```

```
pdf("../../report/graphs/gender/visualizations.pdf", width = 13, height = 10)
par(mfrow = c(3, 3))

column_names <- colnames(voice)[-which(colnames(voice) == "gender")]
for (column in column_names) {
  visualize_data(column)
}

dev.off()
```

```
## pdf
##   2
```

```r
pdf("../../report/graphs/gender/qq_plot_male.pdf", width = 10, height = 10)
par(mfrow = c(3, 3))

column_names <- colnames(voice)[-which(colnames(voice) == "gender")]
for (column in column_names) {
  qqnorm(voice[voice$gender == "male", column], main = column)
  qqline(voice[voice$gender == "male", column], col = "red")
}

dev.off()
```

```
## pdf
##   2
```

```r
pdf("../../report/graphs/gender/qq_plot_female.pdf", width = 10, height = 10)
par(mfrow = c(3, 3))

column_names <- colnames(voice)[-which(colnames(voice) == "gender")]
for (column in column_names) {
  qqnorm(voice[voice$gender == "female", column], main = column)
  qqline(voice[voice$gender == "female", column], col = "red")
}

dev.off()
```

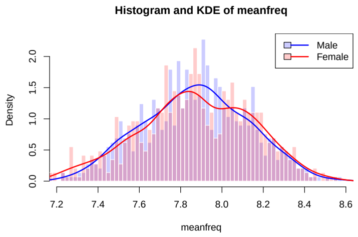
```
## pdf
##   2
```

Mean Frequency

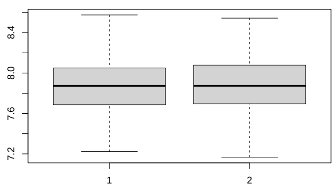
Visualizing the data

We first visualize the data by plotting the histogram

```r
visualize_data("meanfreq")
```



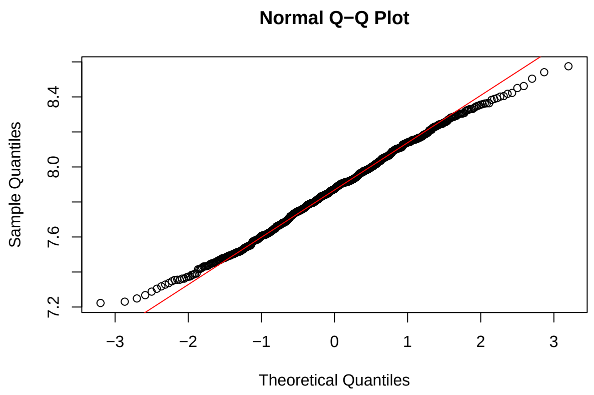
```r
variable <- "meanfreq"
boxplot(male_data[[variable]], female_data[[variable]])
```



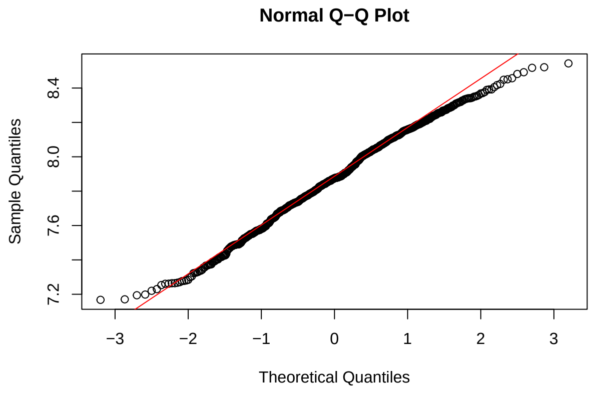
QQ-plot

We then plot the QQ-plot to check for normality

```r
qqnorm(male_data$meanfreq)
qqline(male_data$meanfreq, col = "red")
```



```r
qqnorm(female_data$meanfreq)
qqline(female_data$meanfreq, col = "red")
```



```r
shapiro.test(male_data$meanfreq)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  male_data$meanfreq
## W = 0.99583, p-value = 0.04956
```

```
shapiro.test(female_data$meanfreq)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  female_data$meanfreq
## W = 0.99301, p-value = 0.0018
```

Based on the QQ-plot, we can see that the data is normally distributed. We would therefore use the F test to compare the variance of the data

F-test

```
var.test(male_data$meanfreq, female_data$meanfreq)
```

```
##
##  F test to compare two variances
##
## data:  male_data$meanfreq and female_data$meanfreq
## F = 0.86113, num df = 724, denom df = 724, p-value = 0.04445
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7443040 0.9962966
## sample estimates:
## ratio of variances
##          0.8611315
```

Since the p-value is less than 0.05, we reject the null hypothesis that the variance of the data is the same, we would therefore use the two sample t-test with unequal variance

Two Sample T-test

```
t.test(male_data$meanfreq, female_data$meanfreq, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  male_data$meanfreq and female_data$meanfreq
## t = -0.19049, df = 1440, p-value = 0.849
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.02973444  0.02447059
## sample estimates:
## mean of x mean of y
##  7.870237  7.872869
```

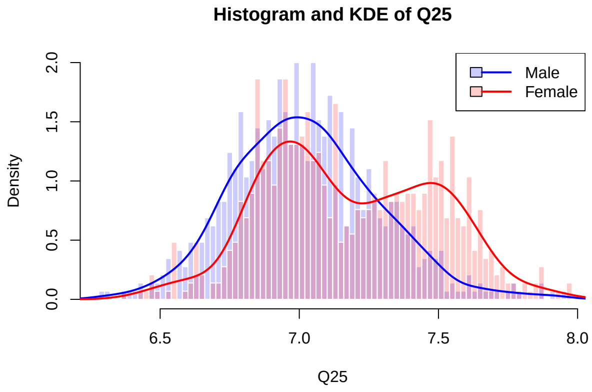
Since the p-value is greater than 0.05, we do not reject the null hypothesis that the mean of the data is the same

Q25

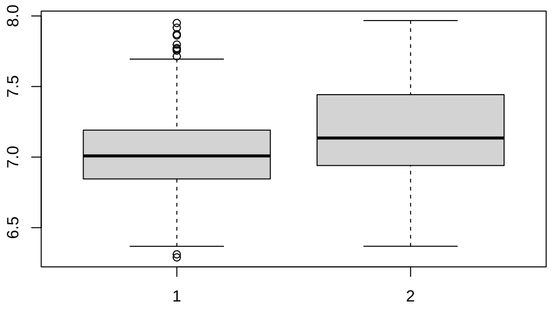
Visualizing the data

We first visualize the data by plotting the histogram

```r
visualize_data("Q25")
```



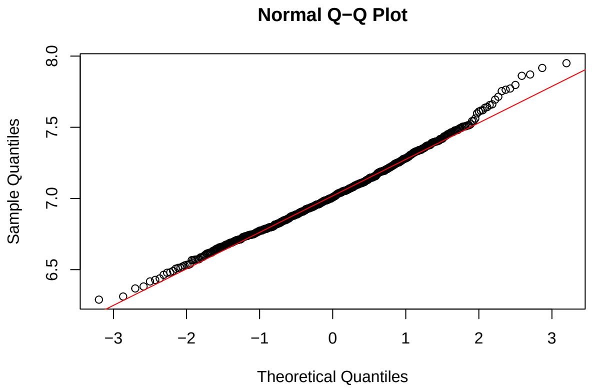
```r
variable <- "Q25"
boxplot(male_data[[variable]], female_data[[variable]])
```



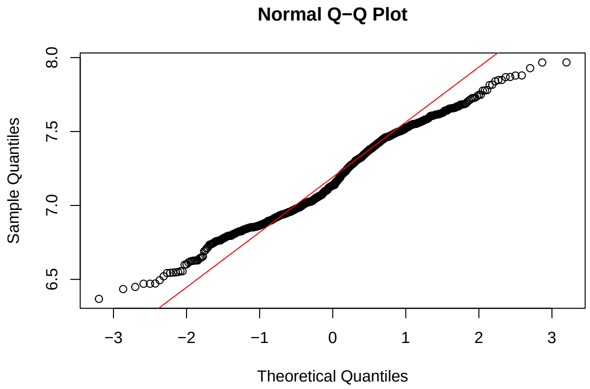
QQ-plot

We then plot the QQ-plot to check for normality

```r
qqnorm(male_data$Q25)
qqline(male_data$Q25, col = "red")
```



```r
qqnorm(female_data$Q25)
qqline(female_data$Q25, col = "red")
```



Based on the QQ-plot, we can see that the data is not normally distributed. We would therefore use the Wilcoxon test to compare the mean of the data

Wilcoxon Test

```r
wilcox.test(male_data$Q25, female_data$Q25, alt = "less")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  male_data$Q25 and female_data$Q25
## W = 188279, p-value < 2.2e-16
## alternative hypothesis: true location shift is less than 0
```

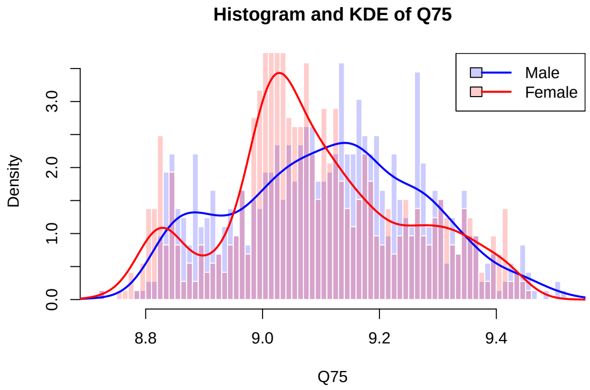
The p-value is less than 0.05, we reject the null hypothesis that the mean of the data is the same

Q75

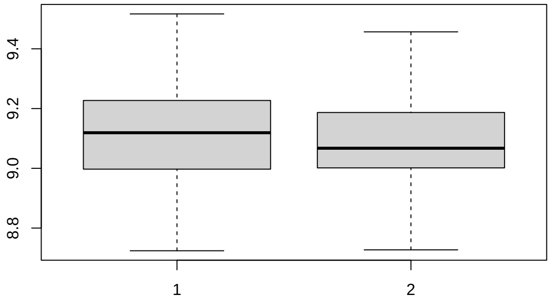
Visualizing the data

We first visualize the data by plotting the histogram

```r
visualize_data("Q75")
```



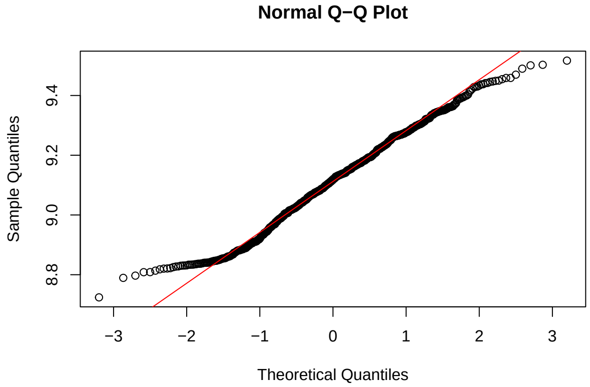
```
variable <- "Q75"
boxplot(male_data[[variable]], female_data[[variable]])
```



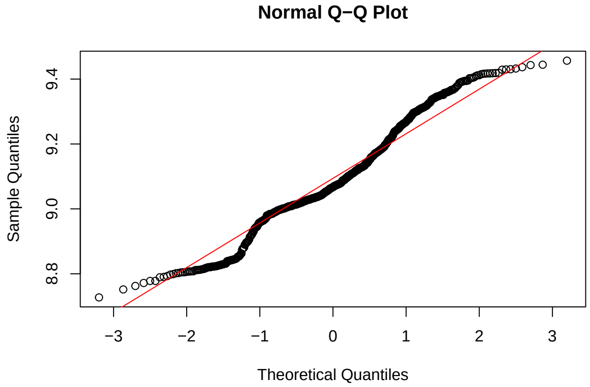
QQ-plot

We then plot the QQ-plot to check for normality

```
qqnorm(male_data$Q75)
qqline(male_data$Q75, col = "red")
```



```
qqnorm(female_data$Q75)
qqline(female_data$Q75, col = "red")
```



Based on the QQ-plot, we can see that the data is not normally distributed. We would therefore use the Wilcoxon test to compare the mean of the data

Wilcoxon Test

```
wilcox.test(male_data$Q75, female_data$Q75, alt = "less")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  male_data$Q75 and female_data$Q75
## W = 288203, p-value = 0.9993
## alternative hypothesis: true location shift is less than 0
```

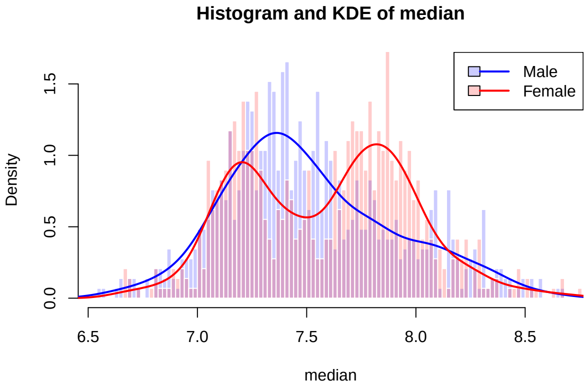
The p-value is more than 0.05, we do not reject the null hypothesis that the mean of the data is the same

median

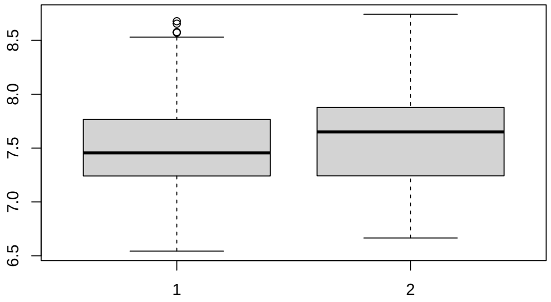
Visualizing the data

We first visualize the data by plotting the histogram

```
visualize_data("median")
```



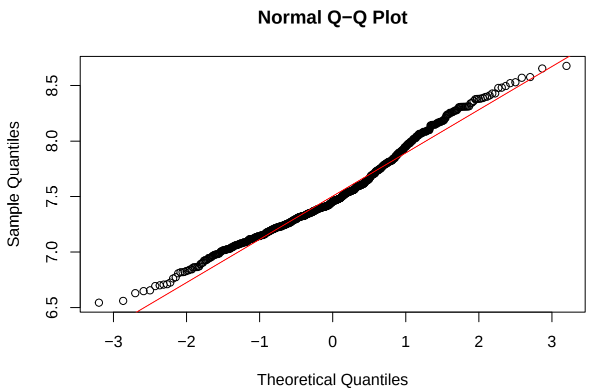
```
variable <- "median"
boxplot(male_data[[variable]], female_data[[variable]])
```



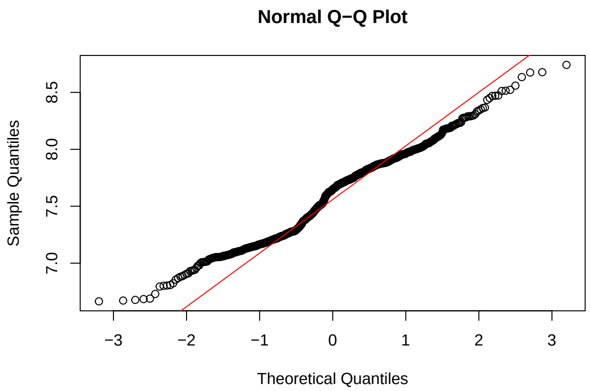
QQ-plot

We then plot the QQ-plot to check for normality

```
qqnorm(male_data$median)
qqline(male_data$median, col = "red")
```



```
qqnorm(female_data$median)
qqline(female_data$median, col = "red")
```



```
shapiro.test(male_data$median)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  male_data$median
## W = 0.97806, p-value = 5.807e-09
```

```
shapiro.test(female_data$median)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  female_data$median
## W = 0.97858, p-value = 8.175e-09
```

From the result, we can see that “median” by gender is not normally distributed. We would therefore use the Wilcoxon test to compare the mean of the data.

Wilcoxon Test

```
wilcox.test(male_data$median, female_data$median, alt = "less")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  male_data$median and female_data$median
## W = 232722, p-value = 8.022e-05
## alternative hypothesis: true location shift is less than 0
```

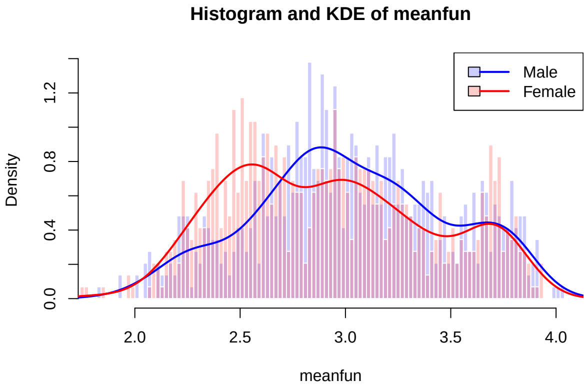
This suggests that there is evidence to support a lower median frequency in the male group compared to the female group.

meanfun

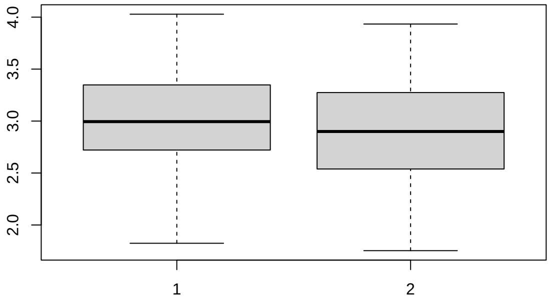
Visualizing the data

We first visualize the data by plotting the histogram

```
visualize_data("meanfun")
```



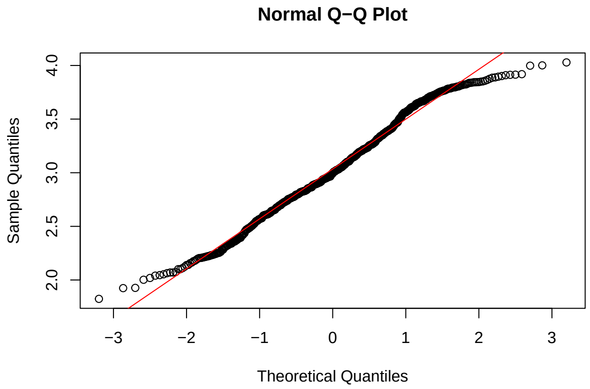
```
variable <- "meanfun"
boxplot(male_data[[variable]], female_data[[variable]])
```



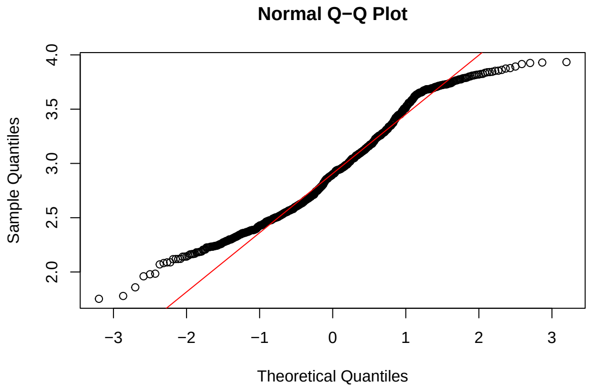
QQ-plot

We then plot the QQ-plot & conduct shapiro test to check for normality

```
qqnorm(male_data$meanfun)
qqline(male_data$meanfun, col = "red")
```



```
qqnorm(female_data$meanfun)
qqline(female_data$meanfun, col = "red")
```



```
shapiro.test(male_data$meanfun)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  male_data$meanfun
## W = 0.98693, p-value = 4.502e-06
```

```
shapiro.test(female_data$meanfun)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  female_data$meanfun
## W = 0.96974, p-value = 4.265e-11
```

From the result, we can see that “meanfun” by gender is not normally distributed. We would therefore use the Wilcoxon test to compare the mean of the data.

Wilcoxon Test

```
wilcox.test(male_data$meanfun, female_data$meanfun, alt = "less")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  male_data$meanfun and female_data$meanfun
## W = 295693, p-value = 1
## alternative hypothesis: true location shift is less than 0
```

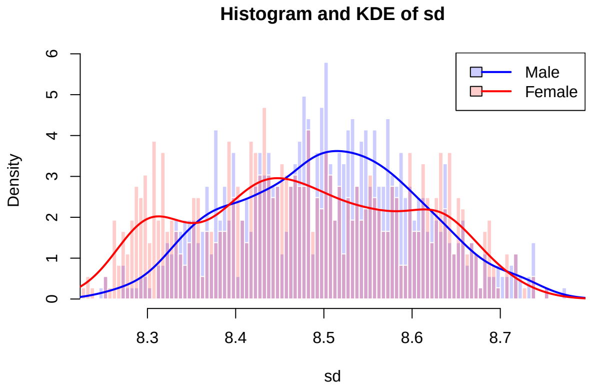
This suggests that there is evidence to support a lower mean fundamental frequency in the male group compared to the female group.

sd

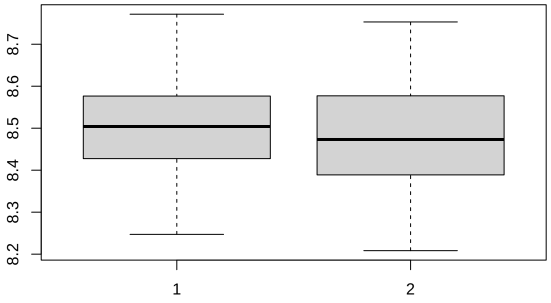
Visualizing the data

We first visualize the data by plotting the histogram

```
visualize_data("sd")
```



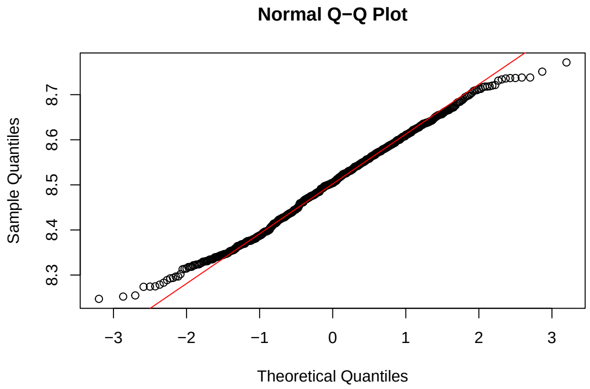
```
variable <- "sd"
boxplot(male_data[[variable]], female_data[[variable]])
```



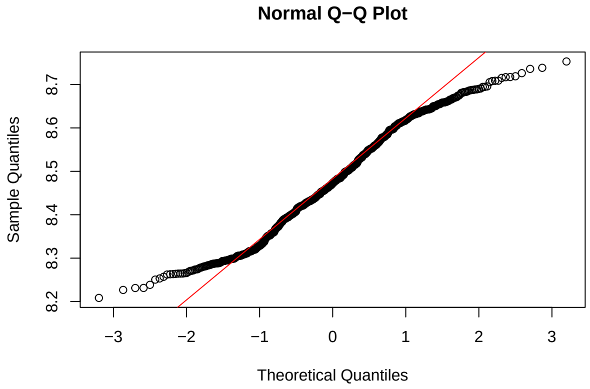
QQ-plot

We then plot the QQ-plot & conduct shapiro test to check for normality

```r
qqnorm(male_data$sd)
qqline(male_data$sd, col = "red")
```



```r
qqnorm(female_data$sd)
qqline(female_data$sd, col = "red")
```



```r
shapiro.test(male_data$sd)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  male_data$sd
## W = 0.9935, p-value = 0.003144
```

```r
shapiro.test(female_data$sd)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  female_data$sd
## W = 0.97639, p-value = 2.002e-09
```

From the result, we can see that “sd” by gender is not normally distributed. We would therefore use the Wilcoxon test to compare the mean of the data.

Wilcoxon Test

```
wilcox.test(male_data$sd, female_data$sd, alt = "greater")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  male_data$sd and female_data$sd
## W = 296638, p-value = 1.104e-05
## alternative hypothesis: true location shift is greater than 0
```

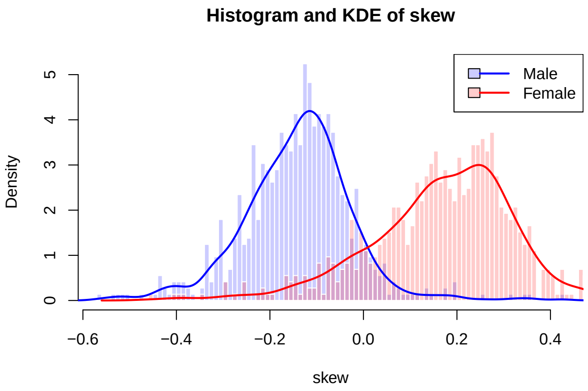
This suggests that there is evidence to support a higher standard deviation in the male group compared to the female group.

skew

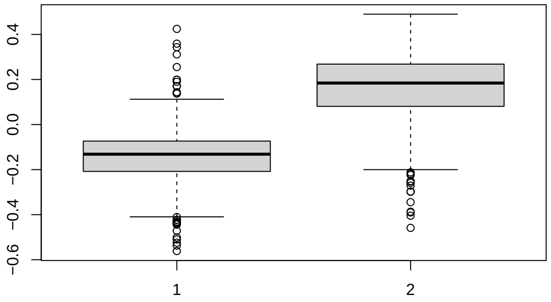
Visualizing the data

We first visualize the data by plotting the histogram

```
visualize_data("skew")
```



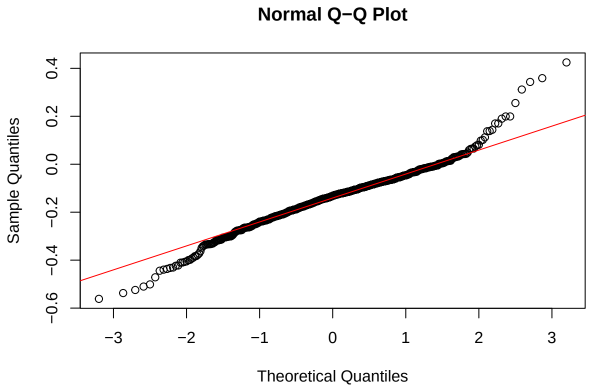
```
variable <- "skew"
boxplot(male_data[[variable]], female_data[[variable]])
```



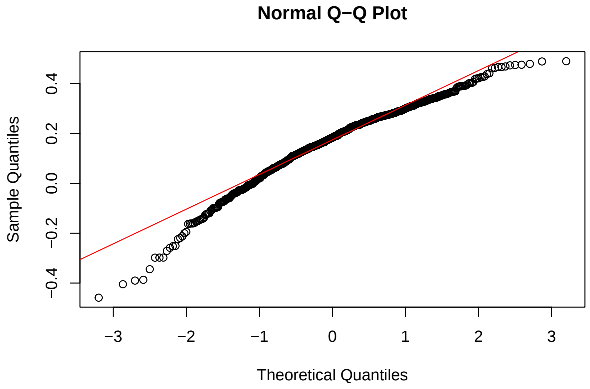
QQ-plot

We then plot the QQ-plot & conduct shapiro test to check for normality

```
qqnorm(male_data$skew)
qqline(male_data$skew, col = "red")
```



```
qqnorm(female_data$skew)
qqline(female_data$skew, col = "red")
```



```
shapiro.test(male_data$skew)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  male_data$skew
## W = 0.96829, p-value = 1.985e-11
```

```
shapiro.test(female_data$skew)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  female_data$skew
## W = 0.96591, p-value = 5.928e-12
```

From the result, we can see that “skew” by gender is not normally distributed. We would therefore use the Wilcoxon test to compare the mean of the data.

Wilcoxon Test

```
wilcox.test(male_data$skew, female_data$skew, alt = "two.sided")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  male_data$skew and female_data$skew
## W = 31736, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

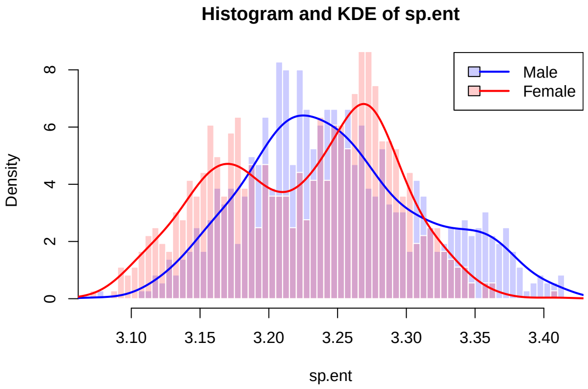
This suggests that there is evidence to support a significant difference of skewness in the male group compared to the female group.

sp.ent

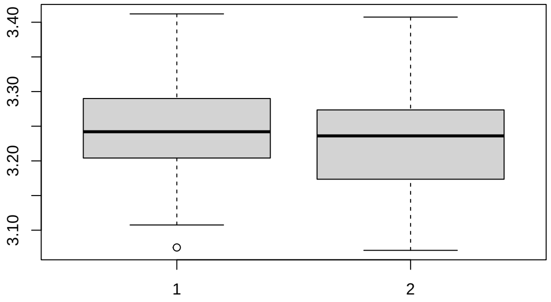
Visualizing the data

We first visualize the data by plotting the histogram

```
visualize_data("sp.ent")
```



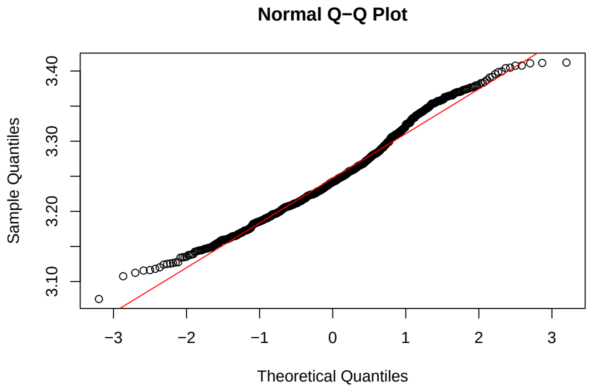
```
variable <- "sp.ent"
boxplot(male_data[[variable]], female_data[[variable]])
```



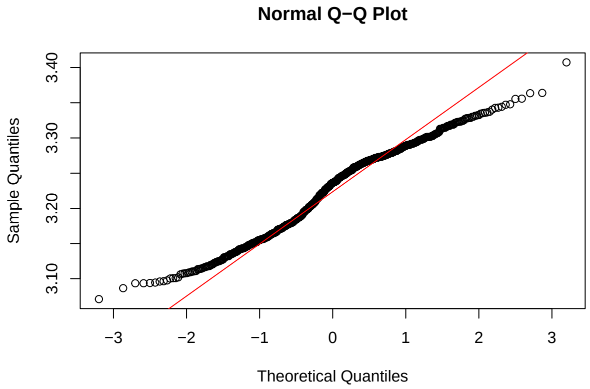
QQ-plot

We then plot the QQ-plot & conduct shapiro test to check for normality

```r
qqnorm(male_data$sp.ent)
qqline(male_data$sp.ent, col = "red")
```



```r
qqnorm(female_data$sp.ent)
qqline(female_data$sp.ent, col = "red")
```



```r
shapiro.test(male_data$sp.ent)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  male_data$sp.ent
## W = 0.98299, p-value = 1.849e-07
```

```r
shapiro.test(female_data$sp.ent)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  female_data$sp.ent
## W = 0.97513, p-value = 9.192e-10
```

From the result, we can see that “sp.ent” by gender is not normally distributed. We would therefore use the Wilcoxon test to compare the mean of the data.

Wilcoxon Test

```
wilcox.test(male_data$sp.ent, female_data$sp.ent, alt = "greater")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  male_data$sp.ent and female_data$sp.ent
## W = 305124, p-value = 5.564e-08
## alternative hypothesis: true location shift is greater than 0
```

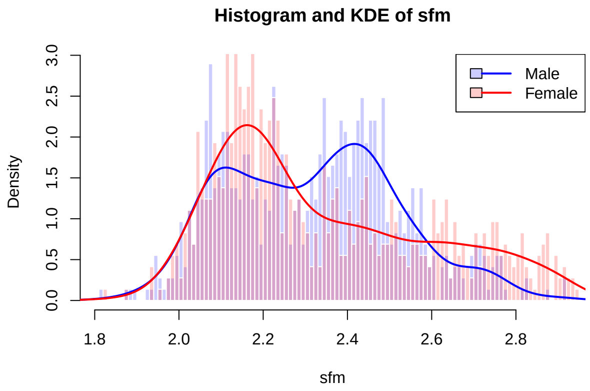
This suggests that there is evidence to support a higher specetral entropy in the male group compared to the female group.

sfm

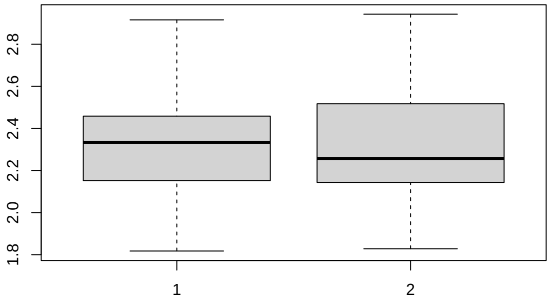
Visualizing the data

We first visualize the data by plotting the histogram

```
visualize_data("sfm")
```



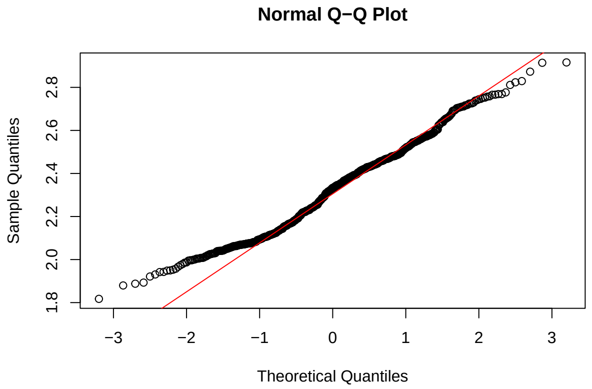
```
variable <- "sfm"
boxplot(male_data[[variable]], female_data[[variable]])
```



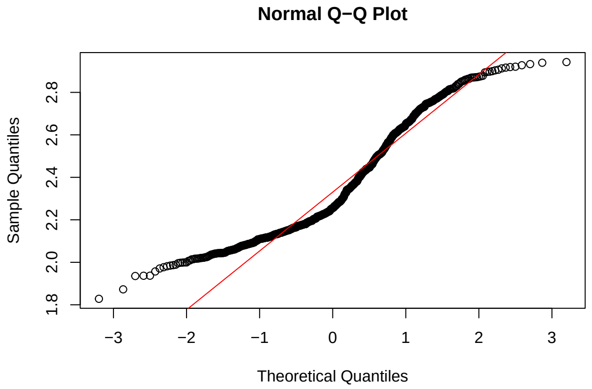
QQ-plot

We then plot the QQ-plot & conduct shapiro test to check for normality

```r
qqnorm(male_data$sfm)
qqline(male_data$sfm, col = "red")
```



```r
qqnorm(female_data$sfm)
qqline(female_data$sfm, col = "red")
```



```r
shapiro.test(male_data$sfm)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  male_data$sfm
## W = 0.98353, p-value = 2.8e-07
```

```r
shapiro.test(female_data$sfm)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  female_data$sfm
## W = 0.92723, p-value < 2.2e-16
```

From the result, we can see that "sfm" by gender is not normally distributed. We would therefore use the Wilcoxon test to compare the mean of the data.

Wilcoxon Test

```
wilcox.test(male_data$sfm, female_data$sfm, alt = "two.sided")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  male_data$sfm and female_data$sfm
## W = 261118, p-value = 0.8317
## alternative hypothesis: true location shift is not equal to 0
```

This suggests that there is NO evidence to support a different sfm in the male group compared to the female group.