Data Analysis by Age on Mean Frequency

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)
```

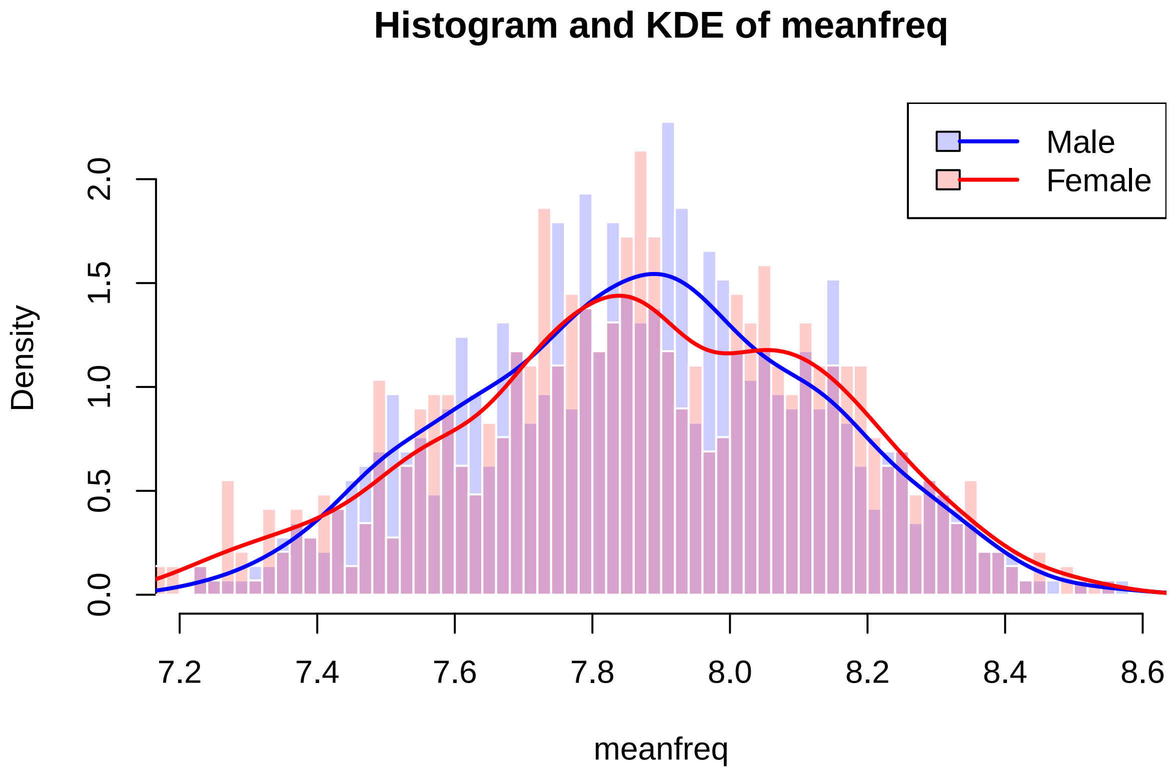
```
##     meanfreq       sd   median      Q25      Q75       skew   sp.ent      sfm
## 726 8.166690 8.581521 8.005468 7.456112 9.171794  0.1165690 3.291912 2.286142
## 727 8.340455 8.599874 8.367010 7.550676 9.296616  0.0456077 3.355465 2.494345
## 728 8.056571 8.546849 7.867097 7.333423 9.110954 -0.1614750 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
```

Visualizing the data

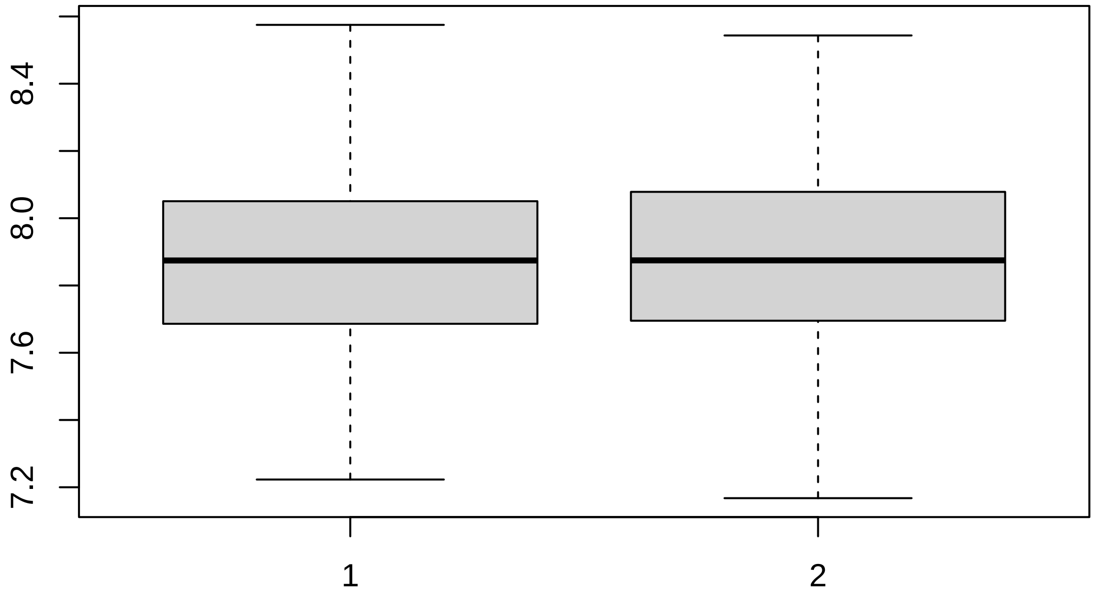
```
visualize_data <- function(column) {
  # return(male_data[column])
  hist(male_data[[column]], xlab = column, col = MALE_COLOR, prob = TRUE, breaks = 80, border = "white"
  hist(female_data[[column]], xlab = column, col = FEMALE_COLOR, prob = TRUE, add = TRUE, breaks = 80,

  # 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
}
```

We first visualize the data by plotting the histogram.

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



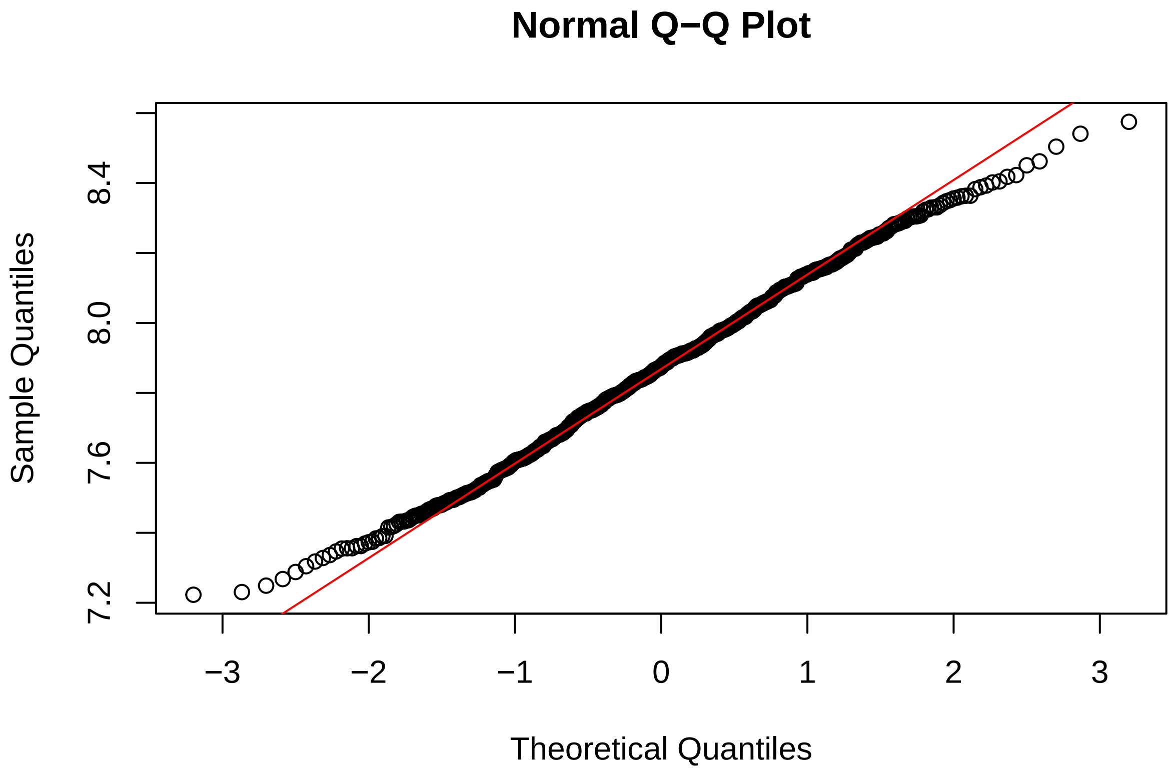
```
variable <- "meanfreq"
boxplot(male_data[[variable]], female_data[[variable]])
```



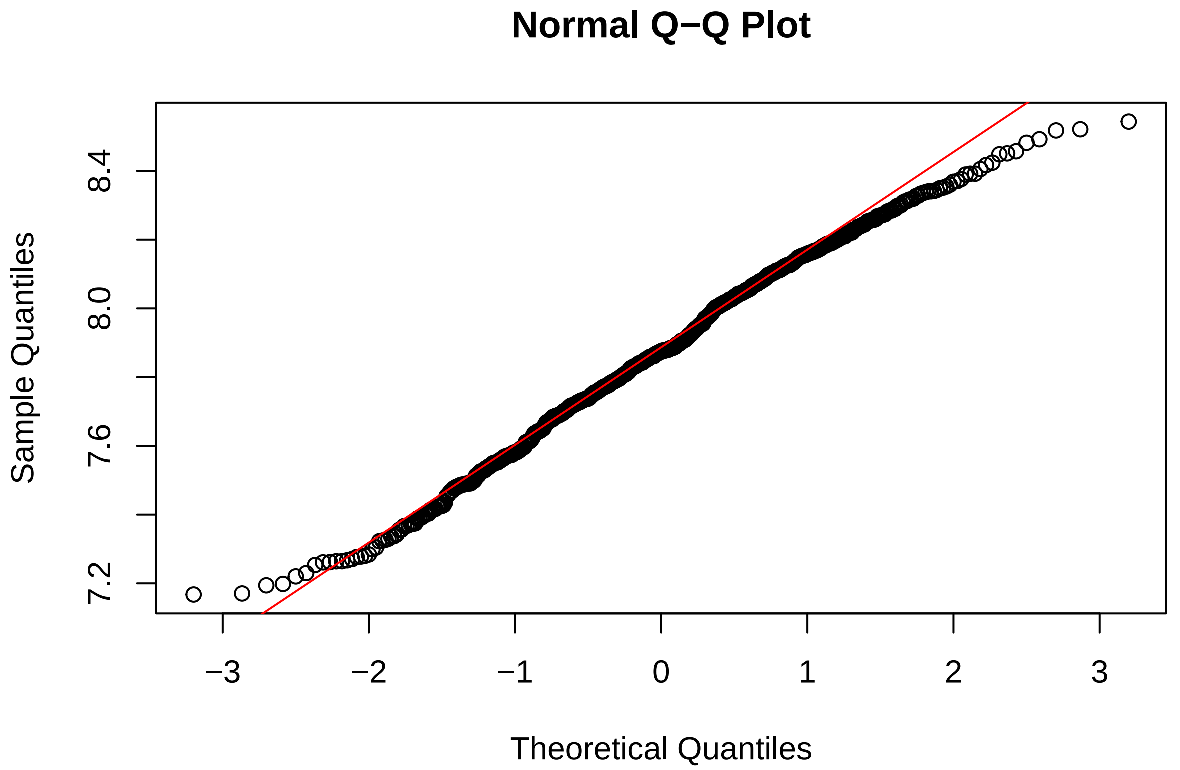
QQ-plot

We then plot the QQ-plot to check for normality

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



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



```
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

```r
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.