Linguistic Data: Quantitative Analysis and Visualisation

Lab on a Student’s t-test

### Aspiration and vowel duration in Icelandic

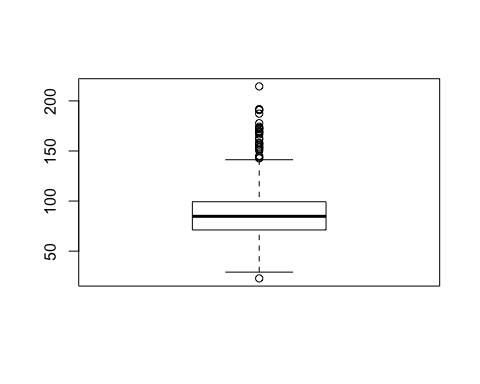
This set is based on (Coretta 2017, [link](https://goo.gl/NrfgJm). This dissertation dealt with the relation between vowel duration and aspiration in consonants. Author carried out a data collection with 5 natives speakers of Icelandic. Then he extracted the duration of vowels followed by aspirated versus non-aspirated consonants. Check out whether vowels before aspirated consonants (like in Icelandic takka ‘key’ [tʰaʰka]) are signiﬁcantly shorter than vowels followed by non-aspirated consonants (like in kagga ‘barrel’ [kʰakka]). [Link](http://math-info.hse.ru/f/2018-19/ling-data/icelandic.csv) to the dataset.

df <- read.csv("http://math-info.hse.ru/f/2018-19/ling-data/icelandic.csv")

### Descriptive statistics

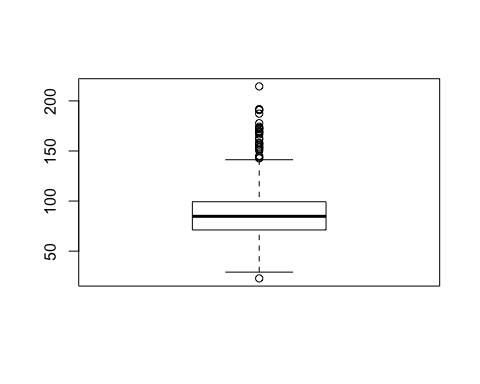
A general boxplot:

boxplot(df$vowel.dur)



Get the number of outliers:

length(boxplot(df$vowel.dur)$out)



## [1] 27

Look at number of observations by groups (aspirated and non-aspirated cases):

table(df$aspriration)

## < table of extent 0 >

Choose two subsamples, one for words where vowels are followed by aspirated consonants and another for non-aspirated consonants.

asp <- df[df$aspiration == 'yes',]  
nasp <- df[df$aspiration == 'no',]

Summary for aspirated and non-aspirated cases:

summary(asp$vowel.dur)

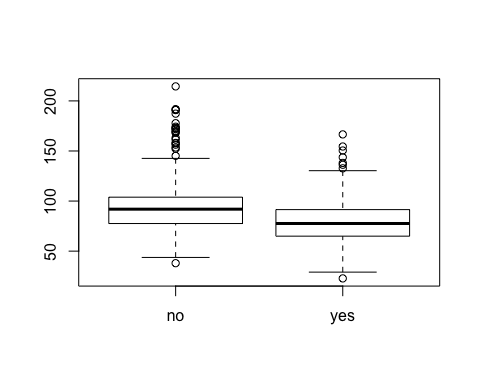
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 22.78 64.96 77.60 78.76 91.46 166.56

summary(nasp$vowel.dur)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 37.98 77.56 91.91 94.69 103.88 214.48

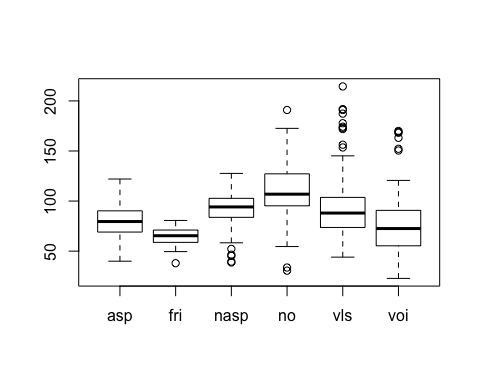
Boxplot by groups:

boxplot(df$vowel.dur ~ df$aspiration)



More interesting - let us create a boxplot by all groups (see the field cons1):

boxplot(df$vowel.dur ~ df$cons1)



You can compare distribution of vowel.dur in asp(irated), fri(cative), nasp(non-aspirated), voi(ced), etc.

We can limit our data to just one type of vowels, say, middle vowels. Therefore, we will work with the same type of a consonant:

asp <- df[df$aspiration == 'yes' & df$height == 'mid', ]  
nasp <- df[df$aspiration == 'no' & df$height == 'mid', ]

Again, here is a summary for a corrected case:

summary(asp$vowel.dur)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 38.67 71.41 81.92 82.65 95.19 150.46

summary(nasp$vowel.dur)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 37.98 80.90 97.97 98.73 110.51 190.93

nrow(asp)

## [1] 156

nrow(nasp)

## [1] 174

### T-test

Let us formulate the null hypothesis, the alternative hypothesis, and apply t-test to our dataset.

t.test(asp$vowel.dur, nasp$vowel.dur)

##   
## Welch Two Sample t-test  
##   
## data: asp$vowel.dur and nasp$vowel.dur  
## t = -6.4869, df = 317.72, p-value = 3.356e-10  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -20.94772 -11.19801  
## sample estimates:  
## mean of x mean of y   
## 82.65371 98.72657

By default, R calculates t.test with regard to the bi-directional alternative hypothesis, such as .

### Unidirectional t-test

H1:

t.test(asp$vowel.dur, nasp$vowel.dur, alternative = "less")

##   
## Welch Two Sample t-test  
##   
## data: asp$vowel.dur and nasp$vowel.dur  
## t = -6.4869, df = 317.72, p-value = 1.678e-10  
## alternative hypothesis: true difference in means is less than 0  
## 95 percent confidence interval:  
## -Inf -11.98542  
## sample estimates:  
## mean of x mean of y   
## 82.65371 98.72657

### Density plots

require(tidyverse)

## Loading required package: tidyverse

## Warning in library(package, lib.loc = lib.loc, character.only = TRUE,  
## logical.return = TRUE, : there is no package called 'tidyverse'

require(dplyr)

## Loading required package: dplyr

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

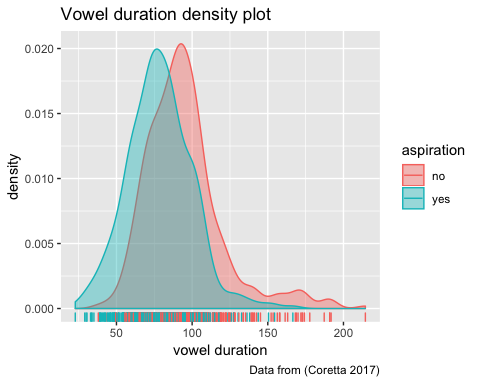
Let’s get a descriptive summary of our data in a dplyr style.

df %>%   
 group\_by(aspiration) %>%  
 summarise(mean = mean(vowel.dur),  
 st.dev = sd(vowel.dur))

## # A tibble: 2 x 3  
## aspiration mean st.dev  
## <fct> <dbl> <dbl>  
## 1 no 94.7 25.9  
## 2 yes 78.8 21.2

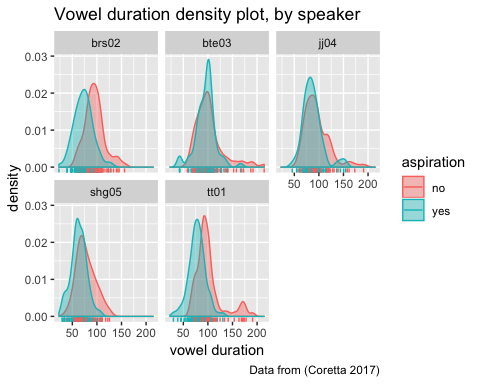
Density plots can be thought of as plots of smoothed histograms.

library(ggplot2)  
df %>%   
 ggplot(aes(vowel.dur, fill = aspiration, color = aspiration))+  
 geom\_density(alpha = 0.4)+  
 geom\_rug()+  
 labs(title = "Vowel duration density plot",  
 caption = "Data from (Coretta 2017)",  
 x = "vowel duration")



Density plot by speaker:

df %>%   
 ggplot(aes(vowel.dur, fill = aspiration, color = aspiration))+  
 geom\_density(alpha = 0.4)+  
 geom\_rug()+  
 facet\_wrap(~speaker)+  
 labs(title = "Vowel duration density plot, by speaker",  
 caption = "Data from (Coretta 2017)",  
 x = "vowel duration")



and descriptive statistics:

df %>%   
 group\_by(aspiration, speaker) %>%  
 summarise(mean = mean(vowel.dur),  
 st.dev = sd(vowel.dur))

## # A tibble: 10 x 4  
## # Groups: aspiration [?]  
## aspiration speaker mean st.dev  
## <fct> <fct> <dbl> <dbl>  
## 1 no brs02 95.3 19.8  
## 2 no bte03 103. 29.4  
## 3 no jj04 95.7 25.1  
## 4 no shg05 77.7 18.9  
## 5 no tt01 101. 26.8  
## 6 yes brs02 72.9 18.9  
## 7 yes bte03 95.4 20.0  
## 8 yes jj04 86.8 20.1  
## 9 yes shg05 63.3 15.7  
## 10 yes tt01 78.3 16.7