Linguistic Data: Quantitative Analysis and Visualisation

Student's t-test in R

Ilya Schurov, Olga Lyashevskaya, George Moroz, Alla Tambovtseva

02 February 2018

Let us load data on a phonological research (Clarett, 2017):

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

The research is dedicated to the relationship between vowel duration in Icelandic language and phonological features of following consonants. Five native speakers were asked to read some texts aloud, their speech was recorded and then the duration of different sounds was measured.

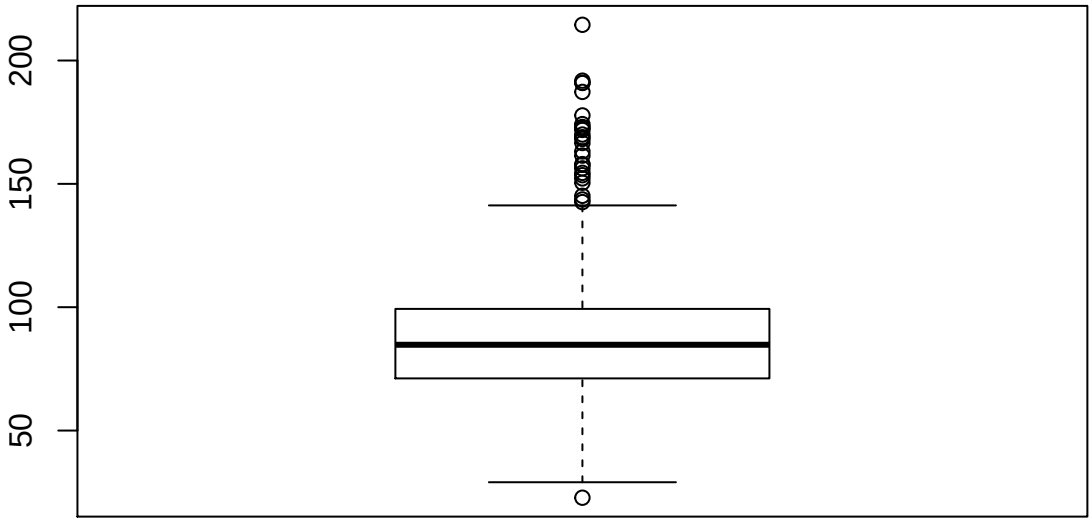
Key variables:

- `word.dur` is word duration in ms;
- `vowel.dur` is vowel duration in ms;
- `cons1` is a consonant type (aspirated, non-aspirated, voiced, voiceless, fricative, etc);
- `height` is a height of a vowel (low, mid, high, diphthong);
- `roundness` is roundness of a vowel (round or unrounded);
- `manner` is a manner of articulation (stop or plosive, nasal, lateral, rhotic);
- `place` is a place of articulation (coronal, labial, velar);
- `aspiration` shows whether it is aspirated or non-aspirated group;
- `syllables` is a number of syllable in a word (one - mono, two - di);
- `syl_structure` is a syllable structure (e.g. cvcc).

Why Icelandic language? In this language we can observe the aspiration effect: vowel duration is different for cases when the following consonant is aspirated and when it is non-aspirated.

Now let's look at the distribution of vowel duration. Instead of a histogram, we will create a different graph called *boxplot* or *box and whiskers plot*.

```
# general boxplot
boxplot(df$vowel.dur)
```



A thick line in the middle of a box corresponds to the median of `vowel.dur` and box borders are lower and upper quartiles. The whiskers show the bounds for typical values in our sample. So, points in this graph that are above the upper whisker and below the lower one are non-typical values or *outliers*. How these whiskers are plotted? There are two options.

1. If there are no outliers at all, the endpoint of a lower whisker corresponds to the minimum value in a sample and the endpoint of an upper whisker corresponds to the maximum value in a sample.
2. If there are outliers on both sides (too small and too large values), the endpoint of a lower whisker is $Q1 - 1.5 \times IRQ$ and the endpoint of an upper whisker is $Q3 + 1.5 \times IRQ$ where IRQ is an interquartile range calculated as follows: $IRQ = Q3 - Q1$.

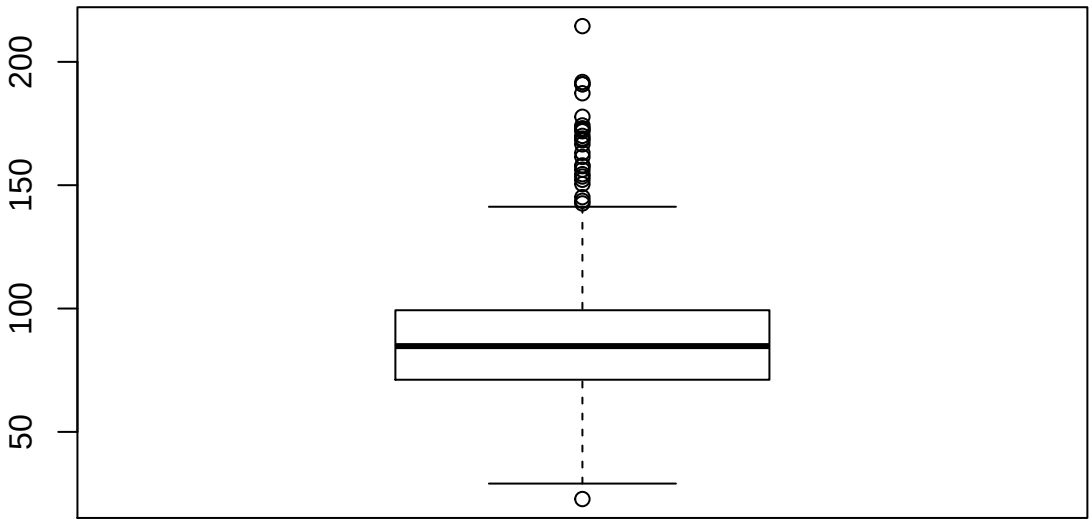
If there are outliers only in a lower partof a sample (too small values), endpoints of whiskers are $[Q1 - 1.5 \times IRQ;$ max]. If there are outliers only in an upper part of a sample (too large values), endpoints of whiskers are [min; $Q3 + 1.5 \times IRQ$].

It is not convenient to count outliers from graph, especially when there are a lot of them. Besides, one point in a graph can correspond to several observations that are equal to each other (e.g. one outlier equal to 2 in a graph can correspond to three repeated values 2, 2, 2). However, we can count outliers by referring to a vector of them:

```
# outliers
boxplot(df$vowel.dur)$out
```

```
##  [1] 166.56157 154.46745 142.60356 157.72465 161.46133 174.22994 153.51177
##  [8] 145.15489 156.27620 191.84090 214.47959 187.27791 173.07899 171.86013
## [15] 190.91578 177.74094 152.20238 168.91030 163.13828 168.19379 169.99201
## [22] 143.67952 190.92719  22.77991 172.58633 158.00735 150.45959
```

```
# number of outliers
length(boxplot(df$vowel.dur)$out)
```




```
## [1] 27
```

Now let's compare distributions of vowel duration for aspirated and non-aspirated cases. First, calculate number of observasions in each group.

```
table(df$apriration)
```

```
## < table of extent 0 >
```

Choose two subsamples (words with aspirated and non-aspirated consonants after vowels):

```r
# choose two subsamples
asp <- df[df$aspiration == 'yes',]
nasp <- df[df$apriration == 'no',]
```

Get the summary for each group:

```r
# summary for aspirated and non-aspirated cases
summary(asp$vowel.dur)
```

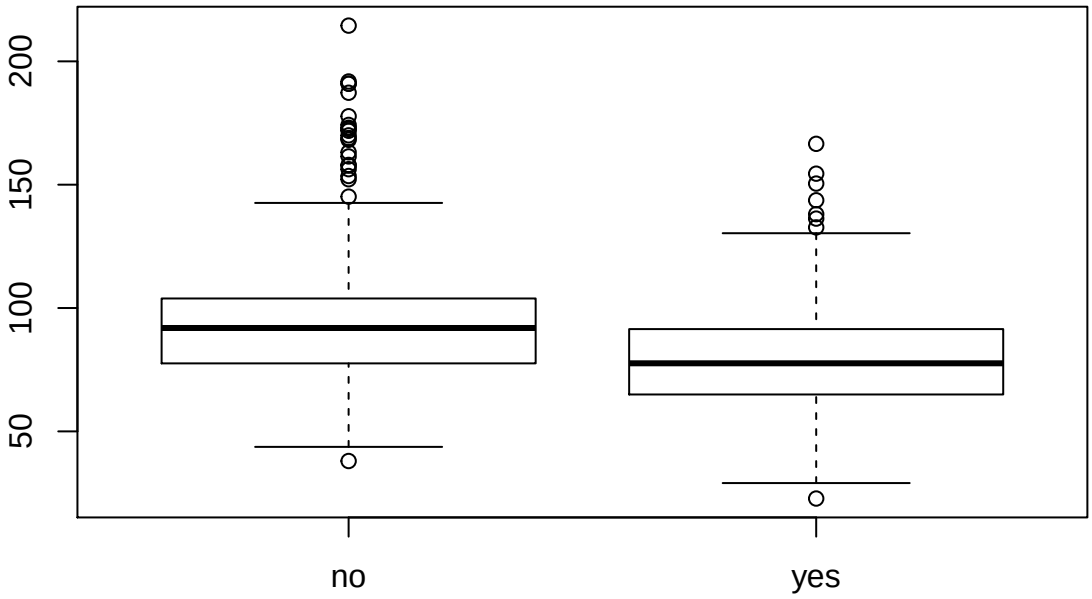
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.78   64.96   77.60   78.76   91.46  166.56
```

```r
summary(nasp$vowel.dur)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##
```

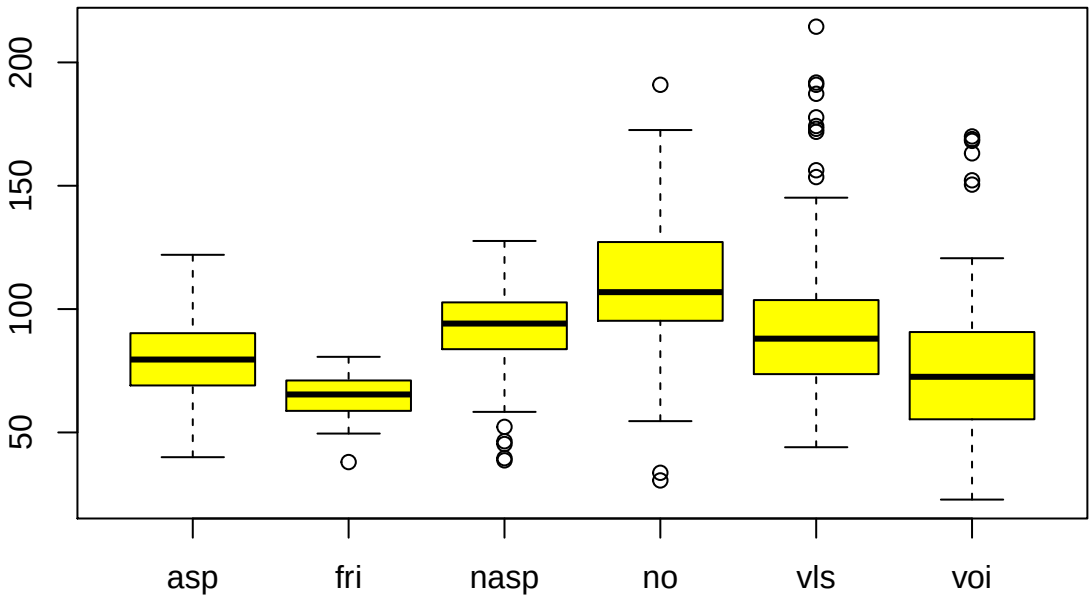
And create box plots by group, one for aspirated cases, another for non-aspirated cases, and both plots in the same graph:

```r
# boxplot for groups
# after ~ here goes a grouping variable
boxplot(df$vowel.dur ~ df$aspiration)
```



We can plot a more interesting picture, boxplots for every type of consonants:

```r
# more interesting - boxplot for all groups
boxplot(df$vowel.dur ~ df$cons1, col="yellow")
```



As was suggested at the seminar, to make our comparisons of vowel duration more correct, we should choose a certain type of vowels and therefore limit the scope of our small research.

```
# correct - make sure we work with the same type of a consonant
asp <- df[df$aspiration == 'yes' & df$height == 'mid', ]
nasp <- df[df$aspiration == 'no' & df$height == 'mid', ]
```

Let's get summaries again:

```
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
```

Question: judging by the output, can we conclude that vowel duration for aspirated and non-aspirated cases differs significantly?

Calculate the number of observations in each group:

```
nrow(asp)
```

```
## [1] 156
```

```
nrow(nasp)
```

```
## [1] 174
```

Now let's proceed to formal testing and perform a two sample Student's t-test:

```
# formal testing
# just list two variables inside
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
```

Note: here it is a Welch Two Sample t-test, but do not mind, the logic of this test is the same, a different formula (with more realistic, less strict assumptions) is used.

Question: what conclusion should we make based on the output obtained?

We can take a one-sided alternative as well, so our null and alternative hypotheses will be:

$$H_0 : \mu_1 = \mu_2 H_1 : \mu_1 < \mu_2.$$

Now we can add an option `alternative="less` so as to set the direction of the alternative:

```
# H1: mu_aps < mu_nasp
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
```