Non-parametric ANOVA

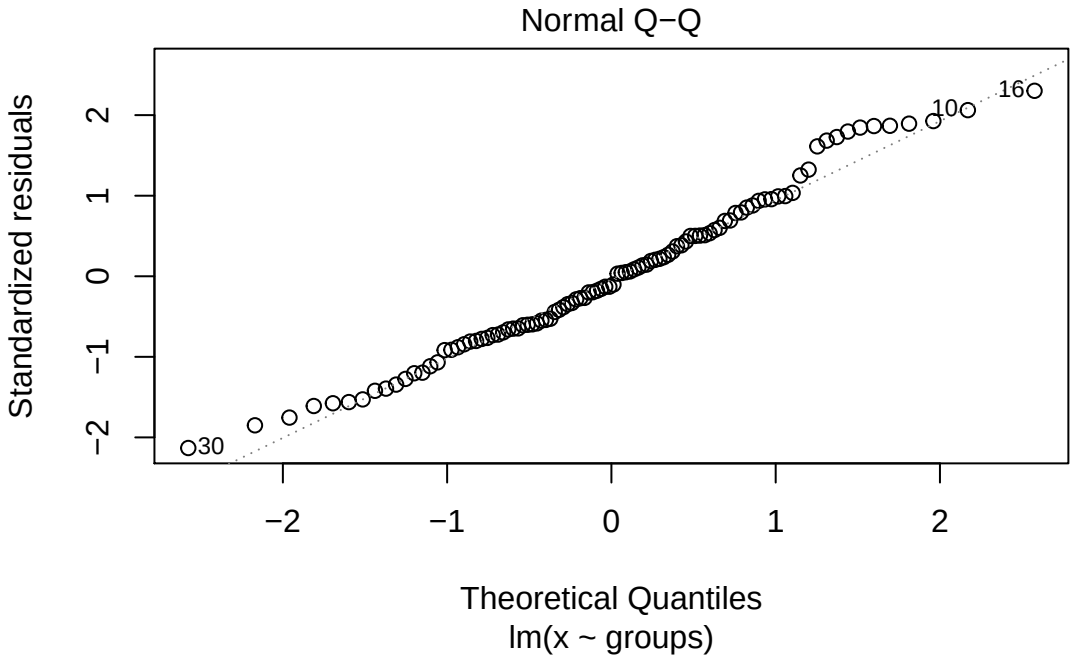
Berti E, Mata JC

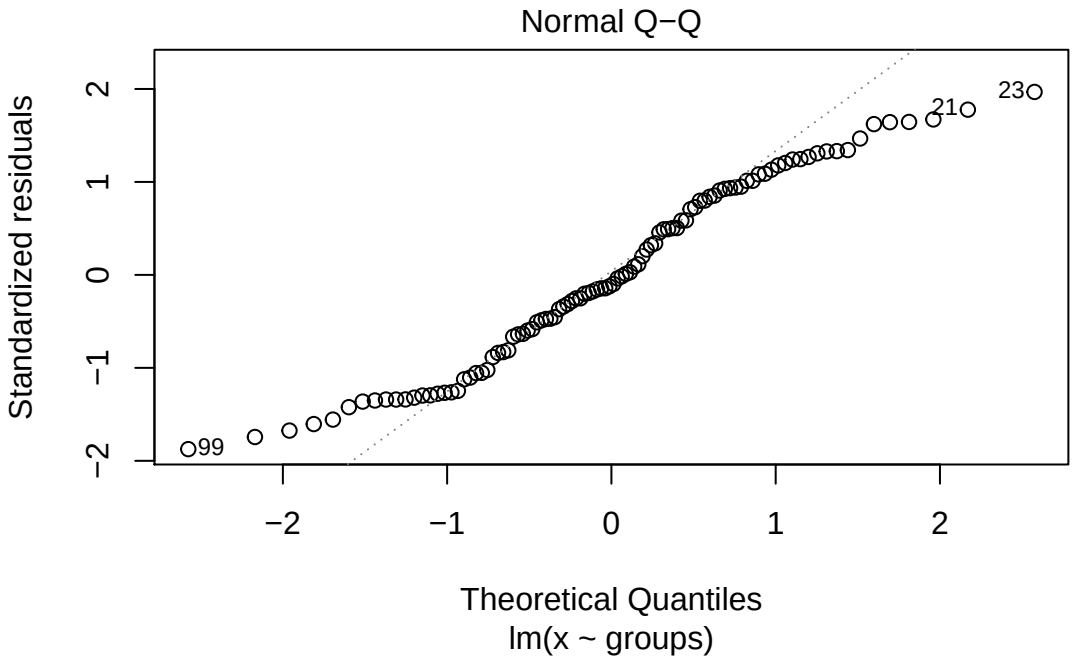
15 January 2019

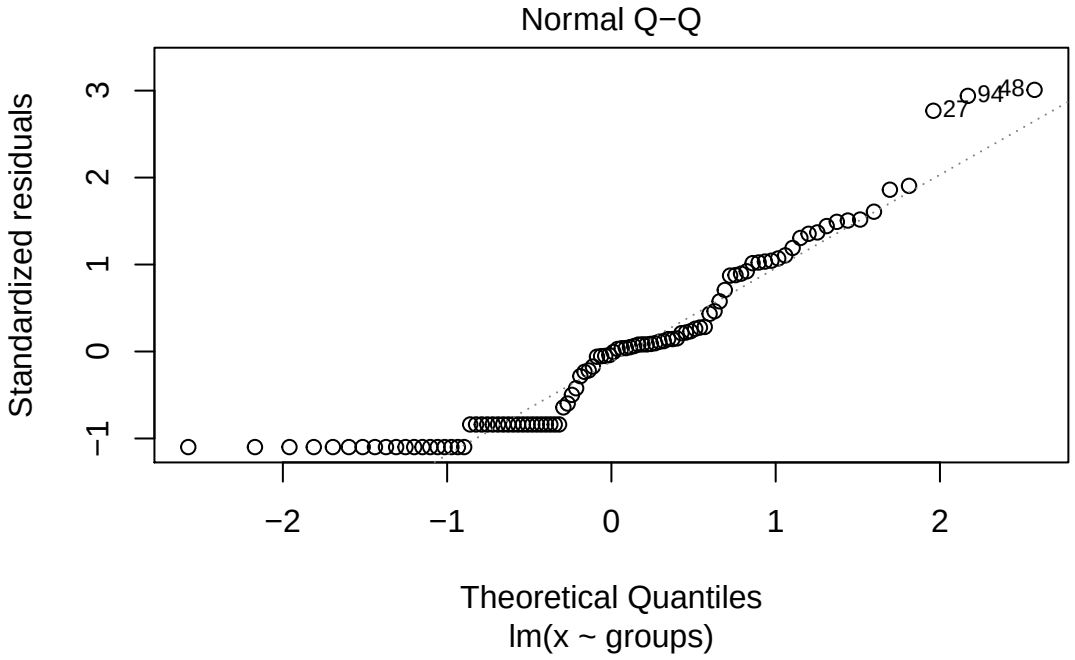
Introduction

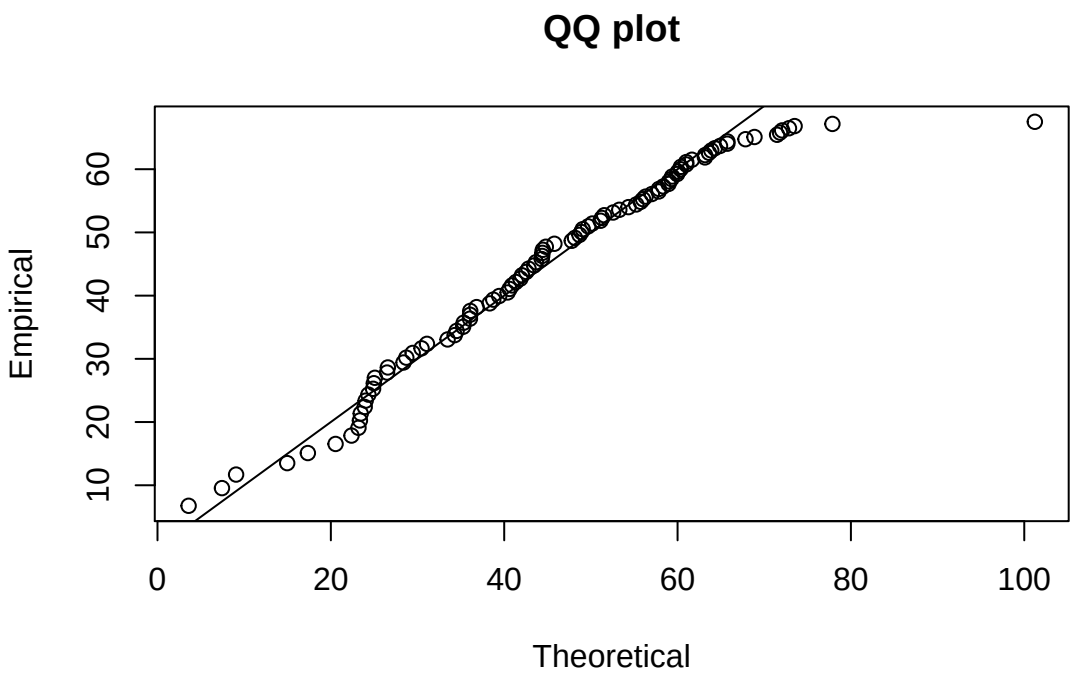
So far, we have encountered data that had normally distributed residuals, and we studied statistical methods to analyze them (Student's t-test, ANOVA). However, what can we do then residuals are not normally distributed? First of all, linear regression analysis is quite robust to non-normality of the residuals. Even if the residuals are not exactly normally distributed, linear regression can give reasonably good results. One way to check if residuals are *more or less* normally distributed is the quantile-quantile plot. If the points in the Q-Q plot fall on the line with intercept = 0 and slope = 1, the residuals are normally distributed.

Question 1: look at the figures below. Would you perform a linear regression analysis on them? Justify your anwer.




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(x ~ groups)
99
21
23
−2
−1
0
1
2


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(x ~ groups)
27
94
48
−1
0
1
2
3



Alternatives to linear regression analysis

Once we established that the residuals are clearly not normally distributed, we have to rely on different statistical approaches that do not assume a particular distribution or estimate parameters. Because they are free of parameters, these methods are usually called non-parametric tests. The most common non-parametric tests are:

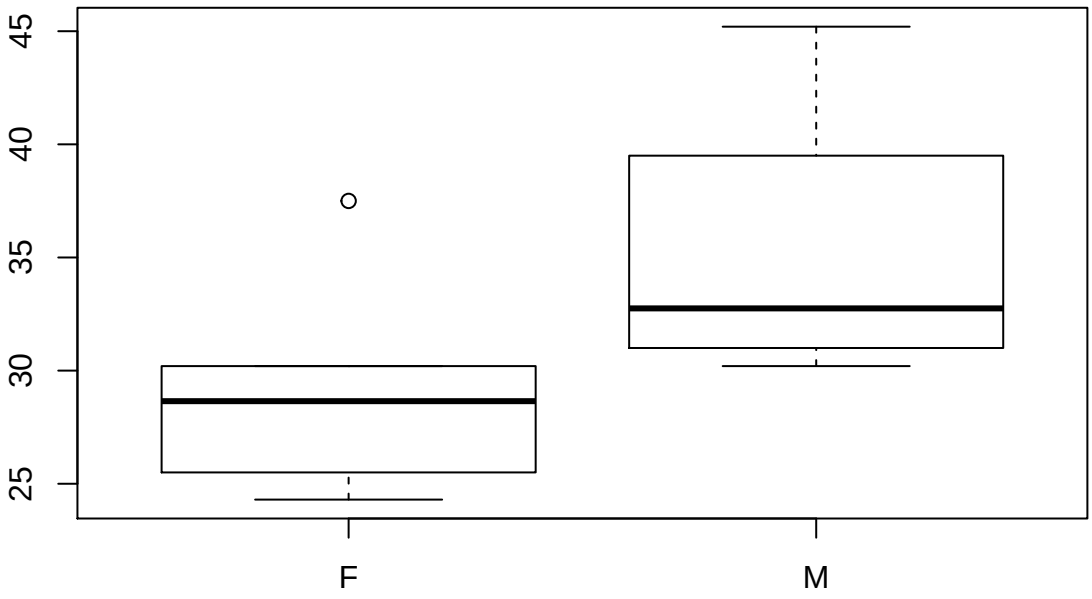
- Kolmogorov-Smirnov test
- Rank-sum test (also called Mann–Whitney U test)
- Signed-rank test (also called Wilcoxon test)
- Kruskal-Wallis test

In the next exercise we will use the rank-sum test (U test) and the Kruskal-Wallis test. **Julia: write here how you want to explain them**

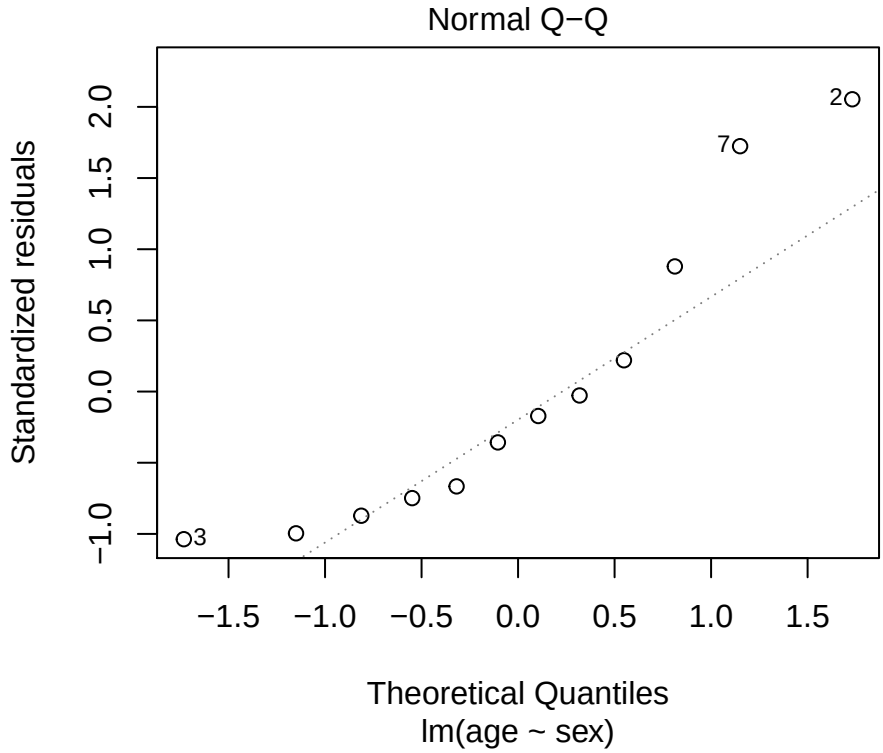
Example 1

We sampled the age of marriage of 12 Danish people, six male and six women, without relation with each other (independent). We want to know if the age at which women and men marry is different.

```
df <- read.csv('marriage_age.csv') #load the data
boxplot(age ~ sex, data = df) #explore the data with a boxplot
```




```
model <- lm(age ~ sex) #create a linear model
plot(model, which = 2) #check residuals
```



Because the residuals have large deviation from the dotted line, we decide to perform the non-parametric rank-sum test.

```
wilcox.test(age ~ sex, data = df, paired = F) # Mann-Whitney U test (rank-sum)
```

```
## Warning in wilcox.test.default(x = c(24.3, 25.5, 28.3, 29, 30.2, 37.5), :
## cannot compute exact p-value with ties
```

```
##
```

```
##  Wilcoxon rank sum test with continuity correction
##
## data:  age by sex
## W = 4.5, p-value = 0.03704
## alternative hypothesis: true location shift is not equal to 0
```

If we had used a parametric linear model, we would have obtained a different result:

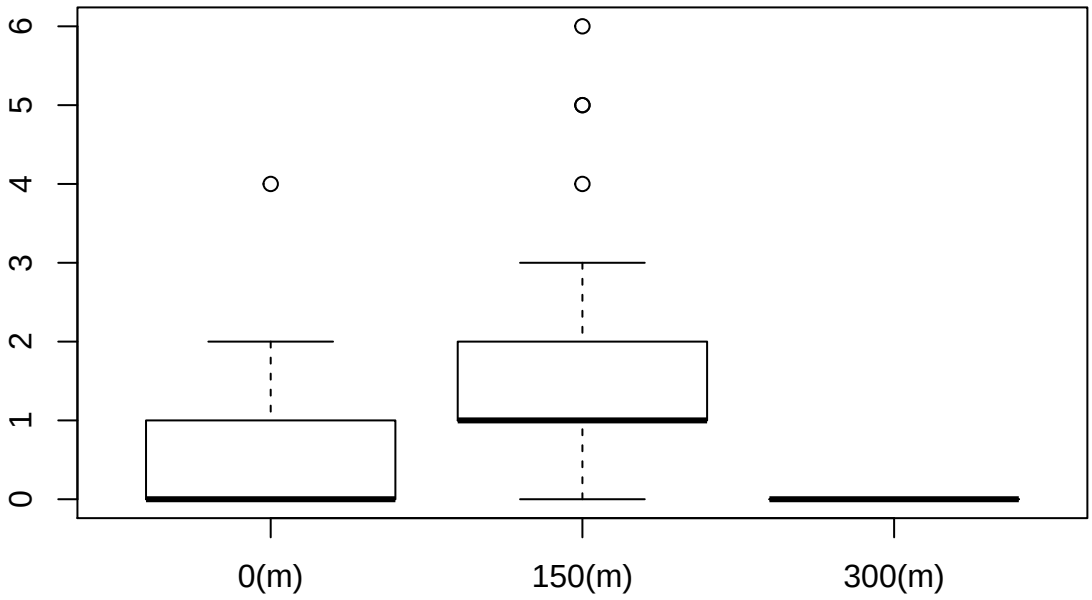
```
t.test(age ~ sex, data = df) #what we would do with normal residuals
```

```
##
##  Welch Two Sample t-test
##
## data:  age by sex
## t = -1.9873, df = 9.4854, p-value = 0.07651
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -12.990047   0.790047
## sample estimates:
## mean in group F mean in group M
##        29.13333        35.23333
```

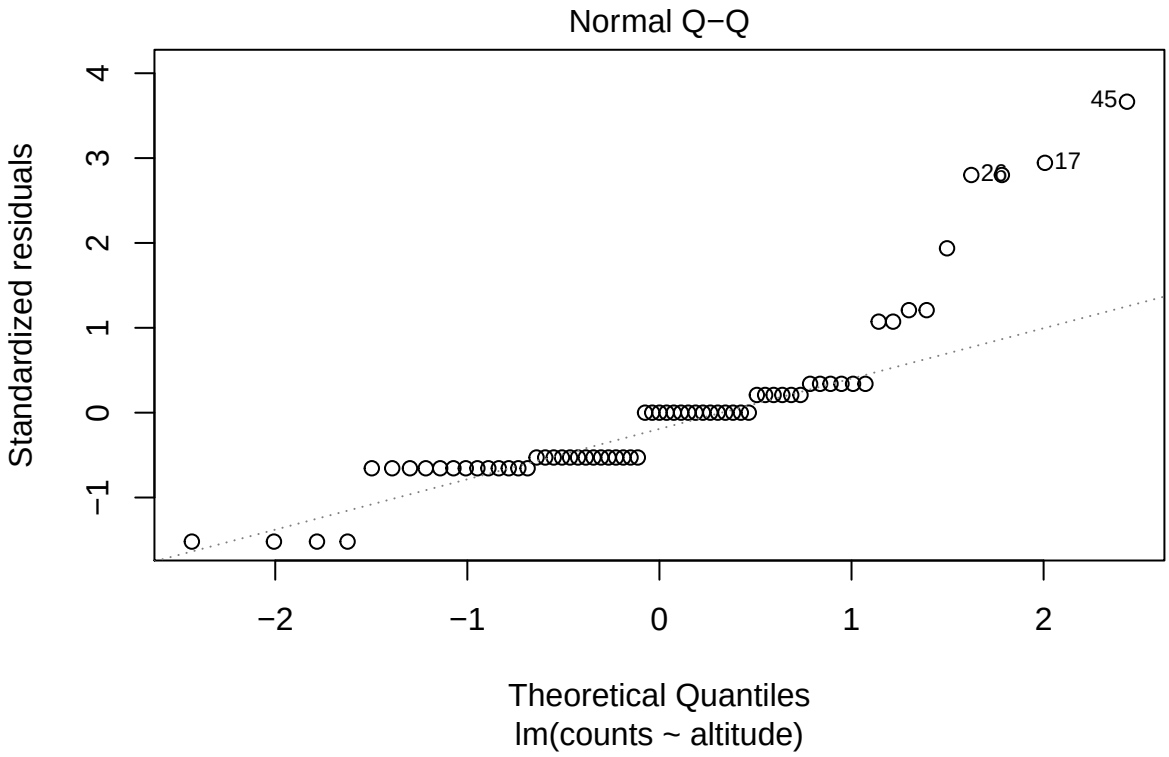
Example 2

We counted the number of occurences of the butterfly species *Schmetterling aurea* at three different altitudes. We want to know if the butterfly is distributed evenly among the three altitudinal levels.

```
df <- read.csv('butterfly_count.csv') #import the data
boxplot(counts ~ altitude, data = df) #explore the data
```




```
model <- lm(counts ~ altitude, data = df) #create linear model
plot(model, which = 2) #check the residuals
```




```
kruskal.test(counts ~ altitude, data = df) #Kruskal-Wallis test
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  counts by altitude
## Kruskal-Wallis chi-squared = 29.036, df = 2, p-value = 4.954e-07
```

In this case, the result of the non-parametric test are not different from results from a linear model (ANOVA):

```
model <- lm(counts ~ altitude, data = df)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: counts
##           Df Sum Sq Mean Sq F value    Pr(>F)
## altitude   2 35.152 17.5758  12.669 2.315e-05 ***
## Residuals 64 88.789  1.3873
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

If we want to know which altitudes are different from each other, we have to compare the occurences between altitudes using a U test between each pairings, for a total of three tests. This is called multiple testing or multiple comparison. Because we are considering several hypotheses together (the first group is different from the second + the first is different from the third + the second is different from the third), we have to adjust the significance level (and the p-value) of the test to not increase the overall probability of rejecting at least one true null hypothesis (type I error). The most common way to account for this is by using the Bonferroni correction, which can be stated as: *every time you perform multiple comparisons, divide you significane level (or multiply your p-value) for the number of comparisons.*

```r
subset_1 <- subset(df, df$altitude != "300(m)") # != means: not equal to
subset_2 <- subset(df, df$altitude != "150(m)") #subset to exclude "150(m)"
subset_3 <- subset(df, df$altitude != "0(m)") #subset to exclude "0(m)"
wilcox.test(counts ~ altitude, data = subset_1)
```

```
## Warning in wilcox.test.default(x = c(1L, 0L, 2L, 1L, 0L, 0L, 1L, 0L, 0L, :
## cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  counts by altitude
## W = 156.5, p-value = 0.0006357
## alternative hypothesis: true location shift is not equal to 0
```

```r
wilcox.test(counts ~ altitude, data = subset_2)
```

```
## Warning in wilcox.test.default(x = c(1L, 0L, 2L, 1L, 0L, 0L, 1L, 0L, 0L, :
## cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  counts by altitude
## W = 240, p-value = 0.007069
## alternative hypothesis: true location shift is not equal to 0
```

```r
wilcox.test(counts ~ altitude, data = subset_3)
```

```
## Warning in wilcox.test.default(x = c(0L, 0L, 5L, 3L, 1L, 1L, 5L, 3L, 1L, :
## cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  counts by altitude
## W = 405, p-value = 9.394e-07
## alternative hypothesis: true location shift is not equal to 0
```

In this case, all p-value after Bonferroni correction are still below the significance level of $\frac{0.05}{3} = 0.17$, and we can reject all null hypotheses.

Exercise 2

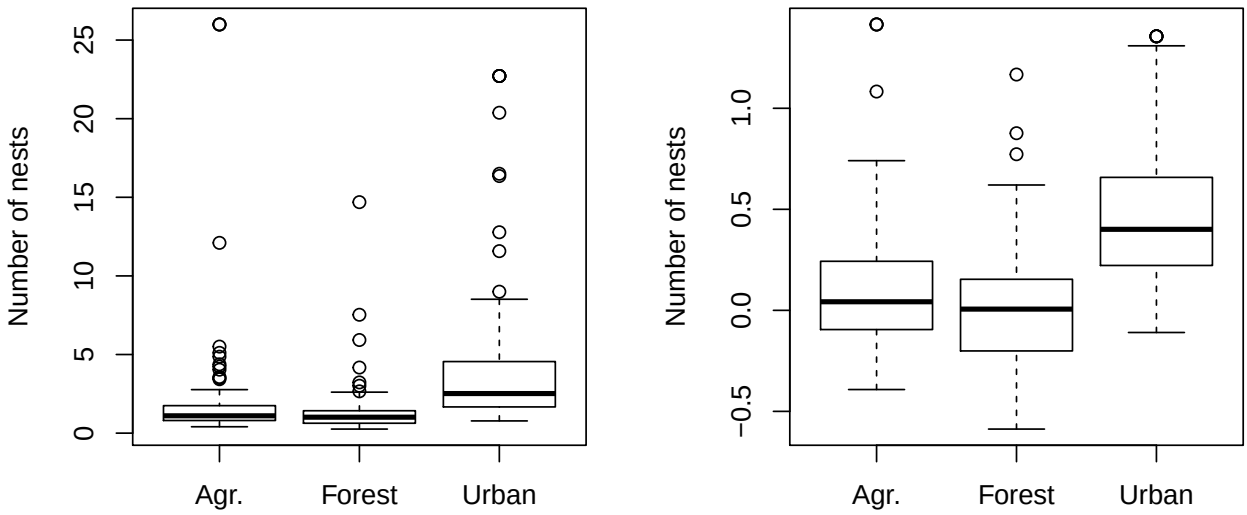
Background

We want to know what is the effect of urbanization on the reproduction of spiders. We have strong evidence to think that spiders have a higher fitness (number of offspring) in urban areas. To test this, we sample the number of spider nests in three different areas: agricultural area, forested area and urban area. The null hypothesis is that there is not difference between the three areas.

Test the hypothesis

First, navigate to your working directory with *setwd()* and load the dataset called *spider-nest.csv* into the environment using *read.csv()*. Familiarize with the dataset and display the boxplot of the count of number of nests per area. This can be done using the function *boxplot(y ~ x)*.

```
setwd('/home/GIT/BEHAVIOURAL-BIOLOGY-2019/non-par_ANOVA')
dataset <- read.csv('spider-nest.csv')
par(mfrow = c(1, 2)) # 1 row, 2 columns in the figure
boxplot(Number.of.nests ~ Area, data = dataset, ylab = 'Number of nests')
boxplot(log10(Number.of.nests) ~ Area, data = dataset, ylab = 'Number of nests')
```

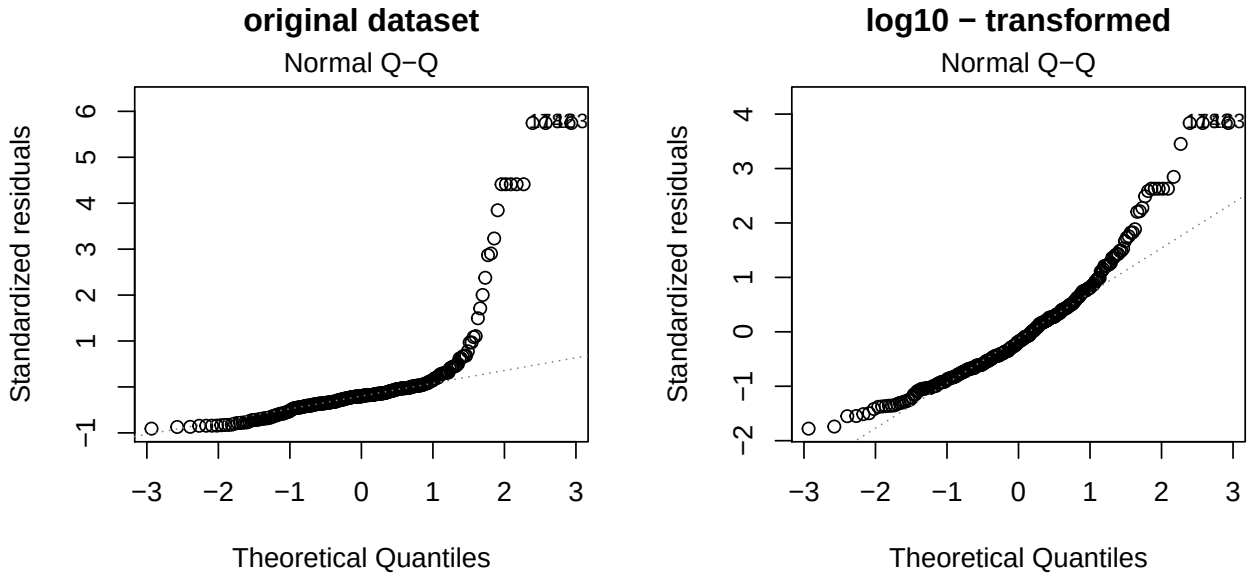


Counts of objects in an area are rarely normally distributed variables. To check if this is the case, we create a linear model and plot its residuals.

```
par(mfrow = c(1, 2))

model <- lm(Number.of.nests ~ Area, data = dataset)
plot(model, which = 2, main = 'original dataset')

model <- lm(log10(Number.of.nests) ~ Area, data = dataset)
plot(model, which = 2, main = 'log10 - transformed')
```



Question 1: are these data normally distributed?

Question 2: using the Kruskal-Wallis test, accept or reject the null hypothesis: there is not difference in the number of nests between areas.

```
kruskal.test(dataset$Number.of.nests, dataset$Area)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  dataset$Number.of.nests and dataset$Area
## Kruskal-Wallis chi-squared = 91.152, df = 2, p-value < 2.2e-16
```

Question 3: using the rank-sum test test, find the areas that differ in number of nests with an overall significance level of $\bar{\alpha} = 0.05$ (use Bonferroni correction). Which ones are different? In R the rank-sum test is implemented in the *wilcox.test()* function, which contains the parameters *paired* that determines if the rank-sum test is performed ($paired = F$), or the signed-rank test is performed ($paired = T$).

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  urban and forest
## W = 8645, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  urban and agr
## W = 7942, p-value = 6.609e-13
## alternative hypothesis: true location shift is not equal to 0
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  agr and forest
## W = 5973, p-value = 0.01749
## alternative hypothesis: true location shift is not equal to 0
```

All p-values are below 0.05, but we can not reject all null hypotheses and say that all areas have different number of nests. In this case, we did not make one test, but three. We need to apply the Bonferroni correction. After the Bonferroni correction, the p-values become:

```
## [1] 1.601616e-18
```

```
## [1] 1.982653e-12
```

```
## [1] 0.05247604
```

After Bonferroni correction we cannot reject all null hypotheses: agricultural and forested areas have the same number of spider nests.