Introduction to parametric tests

Miguel Alvarez

March 15, 2014

Contents

1 Introduction

This is a detailed explanations of an R-session carried out in the context of an internal workshop of the Crop Science Research Group at the University of Bonn (Germany). This session is an introductory exercise on Analysis of Variance (ANOVA).

As requirements for this session you may have an actual version of R in your computer or optional the editor RStudio. Additionally you will need a version of the package `SWEApack`, which is provided by the tutor.

The data set used in this session belongs to the results of the SWEA project (Agricultural Use and Vulnerability of Small Wetlands in East Africa). To load the data set in your `R Workspace` you may type the following command lines.

```
> library(SWEApack)
> data(Africa.env)
```

The table `Africa.env` contains observations of 36 plots in four localities (two localities from Kenya and two localities from Tanzania). Plots are classified into three land uses (Figure 1), namely "unused fields" (semi-natural vegetation), "grazing lands" and "fallows" (long- and short-term abandoned crop fields).

Additional to country, locality and land use membership, `Africa.env` contains soil chemical variables for each plot, namely organic carbon content (in g kg^{-1}), total nitrogen content (in g kg^{-1}), plant available phosphorous (in mg kg^{-1}), exchangeable potassium (in cmol kg^{-1}), electric conductivity (in dS m^{-1}) and pH (Kamiri , 2010).

2 Analysis of variance

The analysis of variance (ANOVA) assumes that the variability of measurements is influenced by determined factors. Additional effects by unknown factors are considered as random (non-systematic) and called "experimental errors". For an easy explanation we will consider the effect of just one factor (independent variable) on a dependent variable, the so called one-way ANOVA. Just to contextualize this example in a scientific problematic, our question will be *"Is the electric conductivity of the soils depending on their country of origin?"* or in other words *"Is the country of origin influencing the electric conductivity*

Figure 1: Classification of land uses in the context of the SWEA project.

of the soils?" I do not really know, how interesting or logic this question can be but it works as a nice example. Then we may formulate a null hypothesis H_0: *"electric conductivity of soils is not depending on their country of origin"*. The alternative hypothesis will be H_1: *"electric conductivity is depending on country of origin and therefore different between countries"*.

To carry out the analysis of variance we will use the function `aov`, which is internally calling a linear model (function `lm`). Alternatively you can use the function `anova`, but then you may previously fit the linear model.

```
> ANOVA <- aov(EC~Country, data=Africa.env)
> summary(ANOVA)

            Df Sum Sq Mean Sq F value   Pr(>F)
Country      1 11.447  11.447   39.94 3.32e-07 ***
Residuals   34  9.745   0.287
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

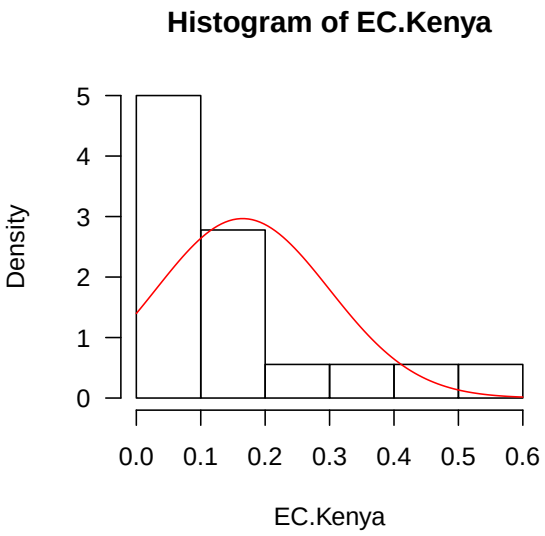
According to the result of `aov` there is a singnificant difference in the soil electric conductivity between countries. But is this result good enough?

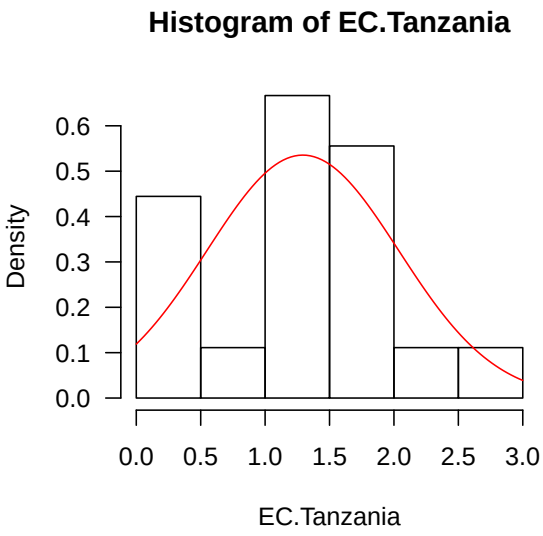
The analysis of variance is a parametric test, that is to say, the values may be normal distributed within populations (in this case different countries). Additionally, the variances between compared populations may be homogeneously distributed (homoscedasticity).

2.1 Tests of normality

Previous to start a test of normality, we look a graphic overview of the distribution of values for the variable electric conductivity. For it, we first split the variable by country of origin and then we check their histograms.

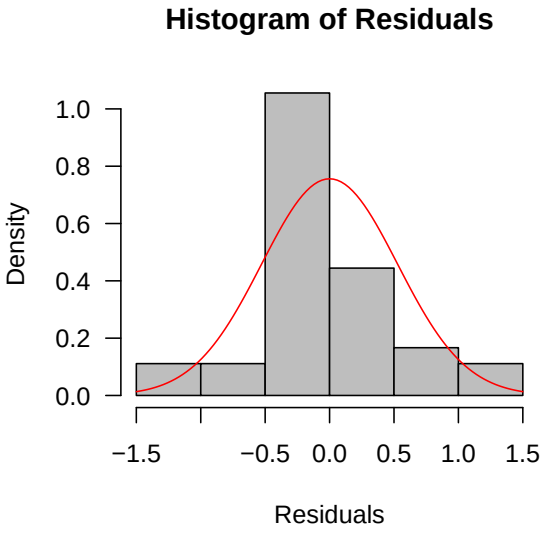
```
> ## Electric conductivity in Kenyan soils
> EC.Kenya <- subset(Africa.env, Country=="Kenya")$EC
> ## Electric conductivity in Tanzanian soils
> EC.Tanzania <- subset(Africa.env, Country=="Tanzania")$EC
> ## Histograms
> par(mfrow=c(1,2), las=1)
> ## Electric conductivity in Kenyan soils
> hist(EC.Kenya, freq=FALSE)
> curve(dnorm(x, mean=mean(EC.Kenya), sd=sd(EC.Kenya)), col="red", add=TRUE)
> ## Electric conductivity in Tanzanian soils
> hist(EC.Tanzania, freq=FALSE)
> curve(dnorm(x, mean=mean(EC.Tanzania), sd=sd(EC.Tanzania)), col="red", add=TRUE)
```

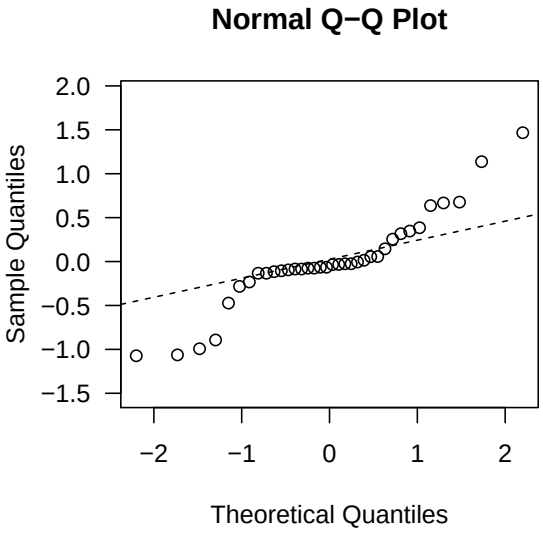




To make life easy, we can alternatively check the distribution fo the residuals in the models. Here, we will use an additional graphic option showing the deviance of observed distribution to a theoretical normal distribution, namely the normal quantile-quantile-plot (Q-Q-plot).

```
> ## Extraction of residuals
> Residuals <- resid(ANOVA)
> par(mfrow=c(1,2), las=1)
> ## Plotting histogram of residuals
> hist(Residuals, freq=FALSE, col="grey")
> curve(dnorm(x, mean=mean(Residuals), sd=sd(Residuals)), col="red", add=TRUE)
> ## Plotting q-q plots
> qqnorm(Residuals, asp=1)
> qqline(Residuals, lty=2)
```





Which is your opinion? Is the distribution of residuals a normal one? While some authors recommend a visual assessment as a honest alternative to decide, if the distribution will be considered as normal or not, other authors advice that visual assessment cannot substitute a test.

To check normal distribution of values we have two alternatives, among others, the Kolmogorov-Smirnov test (function `ks.test`) and the Shapiro-Wilk test (function `shapiro.test`). In this session we will use the second option (for applications of `ks.test` look in Dormann & Kühn 2011).

```
> shapiro.test(Residuals)
```

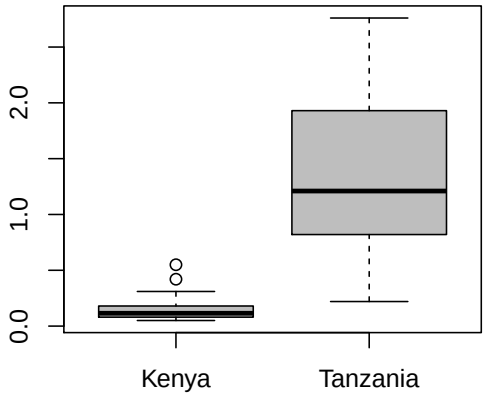
```
        Shapiro-Wilk normality test

data:  Residuals
W = 0.9026, p-value = 0.004017
```

In the Shapiro-Wilk test our null hypothesis is *H_0: "the distribution of residuals is not different from a normal one"*. Considering a confidence interval of 95%, we will find significant differences when the p-value is lower or equal to 5% ($P < 0.05$), which is the case in this example.

2.2 Tests of homoscedasticity

The second condition for an ANOVA is an homogeneous distribution of variances comparing populations (homoscedasticity). As in the previous example, we will first look for a graphic option, in this case a boxplot of the soil electric conductivity by country.

```
> boxplot(EC ~ Country, data=Africa.env, col="grey")
```



Obviously the variance in the Tanzanian samples is much higher than the variance for the Kenyan samples. Herewith we may have also more than one alternative for statistical tests but we will just apply the Bartlett test (function `bartlett.test`).

```
> bartlett.test(EC ~ Country, data=Africa.env)

        Bartlett test of homogeneity of variances

data:  EC by Country
Bartlett's K-squared = 34.7121, df = 1, p-value = 3.823e-09
```

Now our null hypothesis was *H_0: "the variances of soil electric condutivity compared between countries are homogeneous"*. Once again, a p-value lower than 5% will indicate a significant difference between variances, which is the case here.

In conclusion, the variable electric conductivity does not fulfill the requirements for a parametric test and the results of the first ANOVA are not valid. What to do now?

2.3 Data transformations

It is not everything lost for the moment. The salvation can be provided by data transformation. The most popular methods are relatively simple (see Table 1). Such transformations change not only the scale of the variables but they also modify the distribution of the values. Numerical variables are frequently transformed using the logarithmic, reciprocal and squared root transformations. The arcsine-squared root transformation is used to transform percentages and proportions.

If the mentioned transformations did not succeed, it remains the alternative of a Box-Cox transformation. For it the function `boxcox` is available in the package MASS.

Some special cases are the dummy-transformation, applied for the analysis of nominal variables, and the standardization required mainly by multivariate statistics.

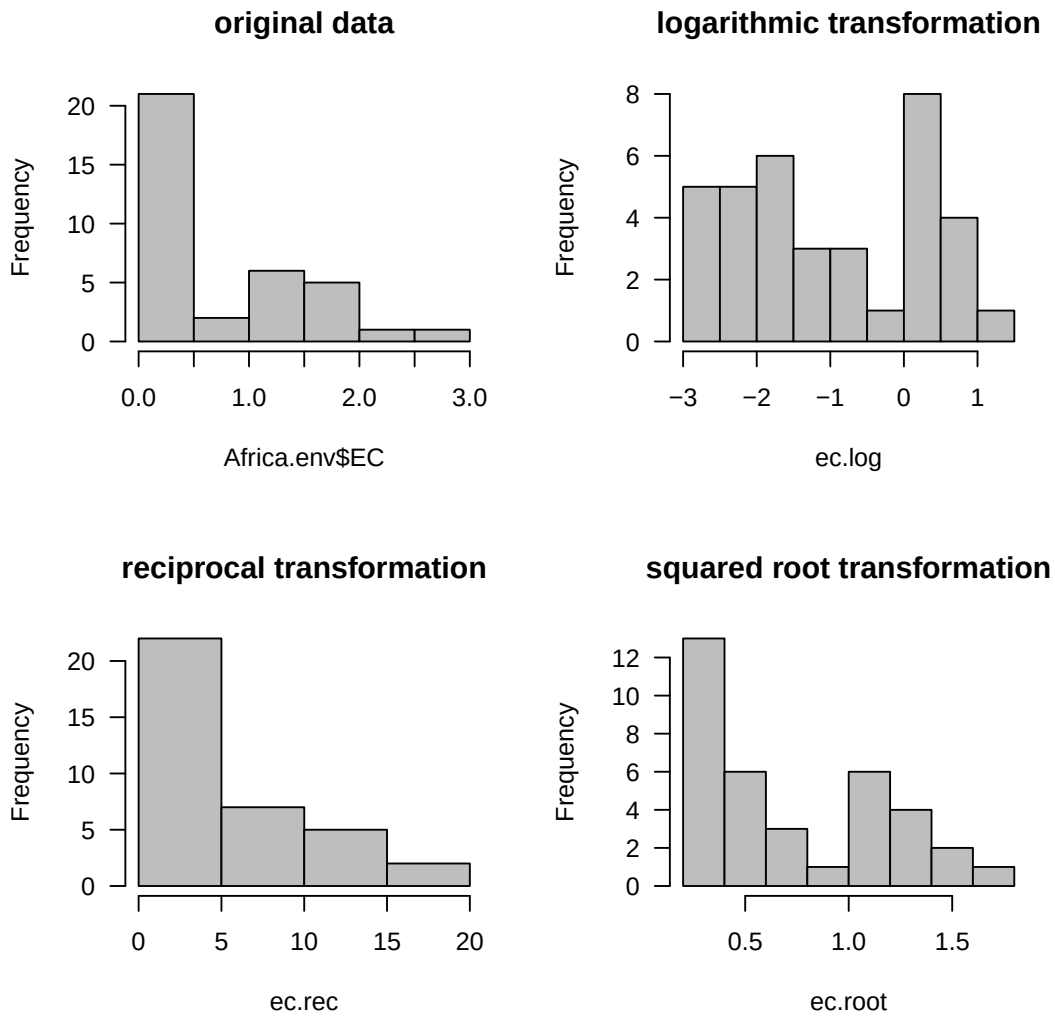
Table 1: Formulas and functions used for data transformation.

Function or formula in R	description
`log(x+c, base=exp(1))`	logarithmic transformation
`1/(x+c)`	reciprocal transformation
`sqrt(x+c)`	squared root transformation
`asin(sqrt(x/c))*180/pi`	arcsine-squared root transformation
`(x-mean(x)/sd(x)), scale(x)`	standardisation

```
> # logarithmic transformation
> ec.log <- log(Africa.env$EC, base=exp(1))
> # reciprocal transformation
> ec.rec <- 1/Africa.env$EC
> # squared root transformation
> ec.root <- sqrt(Africa.env$EC)
```

Let us take a look on the effects of the previous data transformations in the distribution of values.

```
> par(mfrow=c(2,2), las=1)
> hist(Africa.env$EC, col="grey", main="original data")
> hist(ec.log, col="grey", main="logarithmic transformation")
> hist(ec.rec, col="grey", main="reciprocal transformation")
> hist(ec.root, col="grey", main="squared root transformation")
```



We test first the normality of residuals and homoscedasticity of variances in the logarithmic transformation of soil electric conductivity.

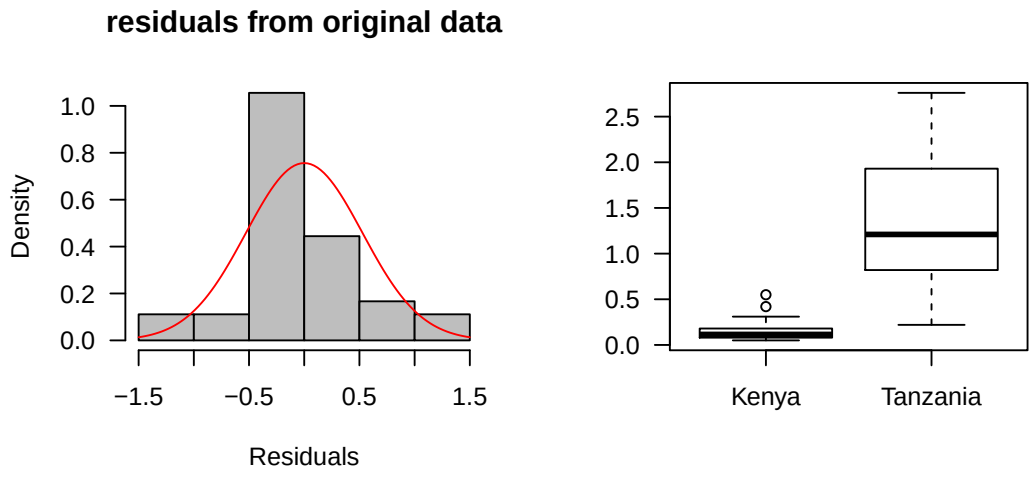
```
> # Test to normality
> shapiro.test(resid(aov(ec.log~Country, data=Africa.env)))

        Shapiro-Wilk normality test

data:  resid(aov(ec.log ~ Country, data = Africa.env))
W = 0.9831, p-value = 0.8442

> # Test to homoscedasticity
> bartlett.test(ec.log~Country, data=Africa.env)

        Bartlett test of homogeneity of variances

data:  ec.log by Country
Bartlett's K-squared = 0.5208, df = 1, p-value = 0.4705
```

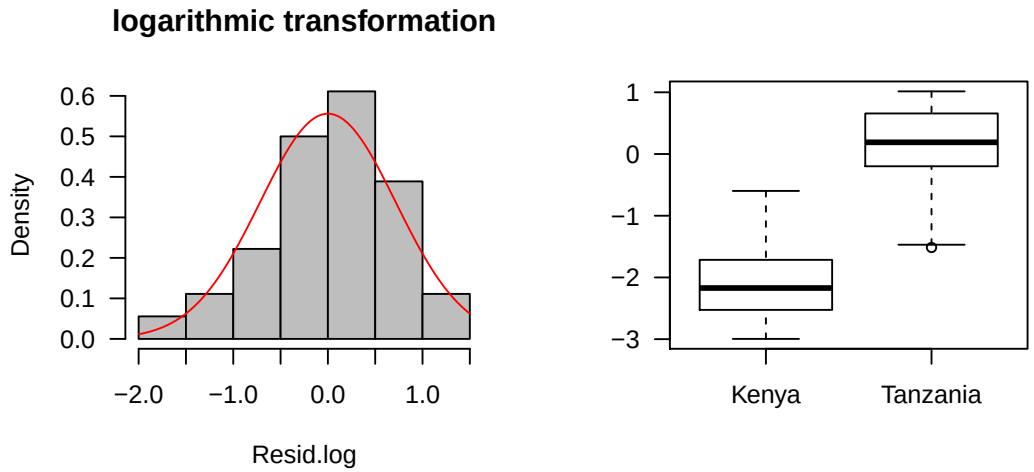
Here we obtain for the logarithmic transformation of soil electric conductivity p-values higher than 0.05 in both, the Shapiro-Wilk test and the Bartlett test. That means, the residuals are normally distributed and the variances are homogeneous. But, what about the other transformation alternatives?

```
> # Tests for reciprocal transformation
> shapiro.test(resid(aov(ec.rec~Country, data=Africa.env)))

        Shapiro-Wilk normality test

data:  resid(aov(ec.rec ~ Country, data = Africa.env))
W = 0.9368, p-value = 0.04055

> bartlett.test(ec.rec~Country, data=Africa.env)

        Bartlett test of homogeneity of variances

data:  ec.rec by Country
Bartlett's K-squared = 22.5029, df = 1, p-value = 2.098e-06

> # Tests for squared root transformation
> shapiro.test(resid(aov(ec.root~Country, data=Africa.env)))

        Shapiro-Wilk normality test

data:  resid(aov(ec.root ~ Country, data = Africa.env))
W = 0.9436, p-value = 0.06584

> bartlett.test(ec.root~Country, data=Africa.env)

        Bartlett test of homogeneity of variances

data:  ec.root by Country
Bartlett's K-squared = 12.7333, df = 1, p-value = 0.0003592
```

For the reciprocal transformation none of the requirements are fulfilled (considering the same significance interval of 95%), while for the squared root transformation only the normality of residuals can be achieved. Therefore, the logarithmic transformation is the only one making the data suitable for a parametric test. We can compare original and transformed data in following graphics.

```
> # Calculation of residuals for logarithmic transformation
> Resid.log <- resid(aov(ec.log ~ Country, data=Africa.env))
> # Plotting distributions
> par(mfrow=c(2,2), las=1)
> hist(Residuals, freq=FALSE, col="grey", main="residuals from original data")
> curve(dnorm(x, mean=mean(Residuals), sd=sd(Residuals)), col="red", add=TRUE)
> boxplot(EC ~ Country, data=Africa.env)
> hist(Resid.log, freq=FALSE, col="grey", main="logarithmic transformation")
> curve(dnorm(x, mean=mean(Resid.log), sd=sd(Resid.log)), col="red", add=TRUE)
> boxplot(ec.log ~ Country, data=Africa.env)
```





After logarithmic transformation the histogram seems to be better adjusted to a theoretical normal distributiona and the boxplots are more similar in their amplitudes.

2.4 ANOVA by the proper way

Once we find a proper transformation of the data, we can carry out the ANOVA and assure correct results.

```
> ANOVA <- aov(ec.log ~ Country, data=Africa.env)
> summary(ANOVA)

            Df Sum Sq Mean Sq F value   Pr(>F)
Country      1  38.21   38.21   72.11 6.44e-10 ***
Residuals   34  18.02    0.53
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After transformation there is still a highly significant difference between countries in the soil electric conductivity. And we can publish it.

3 Post-hoc contrasts

Previous to finish the parametric test I like to make some advices regarding ANOVA. In the example we applied the ANOVA to a factor with just two levels (Kenya and Tanzania). In such case may be enough the use of a paired two-sample t-test (function `pairwise.t.test`). We can apply the one-way ANOVA to factors with more than two levels instead, for example comparing soil reaction in different localities.

```
> ANOVA <- aov(pH ~ Locality, data=Africa.env)
> summary(ANOVA)

            Df Sum Sq Mean Sq F value   Pr(>F)
Locality     3  17.17   5.724   14.15 4.82e-06 ***
Residuals   32  12.94   0.404
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In other cases we may require the addition of a new explanatory (independent) variable, for example the land use type.

```
> ANOVA <- aov(pH ~ Locality + LandUse, data=Africa.env)
> summary(ANOVA)

            Df Sum Sq Mean Sq F value  Pr(>F)
Locality     3 17.172   5.724  14.410 5.4e-06 ***
LandUse      2  1.024   0.512   1.289    0.29
Residuals   30 11.916   0.397
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Or even more, we may also need to know the interaction between factors.

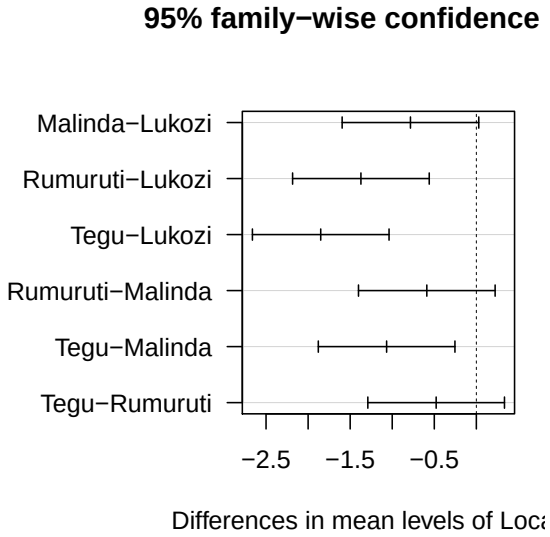
```
> ANOVA <- aov(pH ~ Locality * LandUse, data=Africa.env)
> summary(ANOVA)

                Df Sum Sq Mean Sq F value   Pr(>F)
Locality         3 17.172   5.724  14.865 1.12e-05 ***
LandUse          2  1.024   0.512   1.329    0.283
Locality:LandUse  6  2.675   0.446   1.158    0.361
Residuals        24  9.242   0.385
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We come back to the first example and we test the effect of locality in the soil reaction (measured as pH). Of course, you may check previously the normality of residuals and the homoscedasticity of variances, but this is just one example. Here the ANOVA tell us "there is a difference in the soil reaction depending on sampling locality" but we do not know, which localities have soil pH values differetn to which ones. To know it, we require a *post-hoc* contrast. In this session we will use the Tukey's honest significant difference (function `TukeyHSD`).

```
> ANOVA <- aov(pH ~ Locality, data=Africa.env)
> summary(ANOVA)

            Df Sum Sq Mean Sq F value   Pr(>F)
Locality     3  17.17   5.724   14.15 4.82e-06 ***
Residuals   32  12.94   0.404
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> Tukey <- TukeyHSD(ANOVA)
> Tukey

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = pH ~ Locality, data = Africa.env)

$Locality
                       diff       lwr         upr     p adj
Malinda-Lukozi   -0.7833333 -1.595514  0.02884776 0.0619263
Rumuruti-Lukozi  -1.3722222 -2.184403 -0.56004113 0.0003764
Tegu-Lukozi      -1.8500000 -2.662181 -1.03781891 0.0000038
Rumuruti-Malinda -0.5888889 -1.401070  0.22329220 0.2225102
Tegu-Malinda     -1.0666667 -1.878848 -0.25448558 0.0062137
Tegu-Rumuruti    -0.4777778 -1.289959  0.33440331 0.3963421
```

In the output of `TukeyHSD` you may look for those pairwise comparissons where both, the upper and the lower limit of the difference are either negative or positive. If the mentioned values have different symbols, the difference is not significant. That can be also checked in the column showing the p-values (`p adj`). There is also a graphic option to show the results of `TukeyHSD`.

```
> # Plotting Tukey output
> par(las=1, mar=c(5,10,5,1))
> plot(Tukey)
```



4 Disclaimer

This handout is a preliminary version done by a non-statistician, therefore use carefully and at own risk. Since this text will be improved for further R-courses, any comment, correction and suggestion is most welcome (send them to malvarez@uni-bonn.de).

The content of this handout is based on German references (Ligges , 2008; Köhler et al. , 2002; Dormann & Kühn , 2011).

Some English references can be Zar (2009) and the tutorials by King (2014), Schumacher (2007) and Everitt & Hothorn (2005).

5 Excercises

- You ask "is there a difference in the soil pH between sampling localities?", but this time you will check first the normal distribution of residuals and the homogeneity of variances between localities.
- Is there differences between applied insecticides in the amount of insects (survivors) found in experimental plots? Use the data set `InsectSprays`.
- Apply a factorial ANOVA to the data set `ToothGrowth`.
- Now try the same with your own data set.

6 Bibliography

Dormann CF, Kühn I (2011). *Angewandte Statistik für die biologischen Wissenschaften.* UFZ, Leipzig-Halle. PDF online.

Everitt BS, Hothorn T (2011). *A handbook of statistical analyses using R.* UFZ, Leipzig-Halle. PDF online.

Kamiri HW (2010). *Effects of land use dynamics on attributes of wetland soils in Africa.* INRES, Bonn. Bonner Agrikulturchemische Reihe 41.

King WB (2014). *R Tutorials*. Coastal Carolina University. URL

Köhler W, Schachtel G, Voleske P (2002). *Biostatistik*. Springer, Berlin.

Ligges U (2008). *Programmieren mit R*. Springer, Leipzig.

Schumacher J (2007). *Analysis of variance (ANOVA) in R*. PDF online.

Zar JH (2009). *Biostatistical analysis*. Prentice Hall.