CASE STUDIES

Use of Statistical Tools in a development context. Analysis of Variance (ANOVA)

Maribel Ortego and Enric Lloret

PHOTO: Monduli District, Tanzania. Agustí Pérez-Foguet .

GDEE GLOBAL DIMENSION IN ENGINEERING EDUCATION

CASE STUDIES

Use of Statistical Tools in a development context. Analysis of Variance (ANOVA)

EDITED BY
Global Dimension in Engineering Education

COORDINATED BY
Agustí Pérez-Foguet, Enric Velo, Pol Arranz, Ricard Giné
and **Boris Lazzarini (***Universitat Politècnica de Catalunya)*
Manuel Sierra *(Universidad Politécnica de Madrid)*
Alejandra Boni and **Jordi Peris** *(Universitat Politècnica de València)*
Guido Zolezzi and **Gabriella Trombino** *(Università degli Studi di Trento)*
Rhoda Trimingham *(Loughborough University)*
Valentín Villarroel *(ONGAWA)*
Neil Nobles and Meadhbh Bolger *(Practical Action)*
Francesco Mongera *(Training Center for International Cooperation)*
Katie Cresswell-Maynard *(Engineering Without Border UK)*

DL B 10735-2015
ISBN 978-84-606-7546-4

Citation: Ortego, M. and Lloret, E. (2015) ' Use of Statistical Tools in a development context. Analysis of Variance (ANOVA)',
in *Case studies for developing globally responsible engineers*, GDEE (eds.), Global Dimension in Engineering Education, Barcelona.
Available from: http://gdee.eu/index.php/resources.html

Disclaimer: This document has been produced with the financial assistance of the European Union
The contents of this document are the sole responsibility of the authors and can under no circumstances be regarded as reflecting the position of the European Union

USE OF STATISTICAL TOOLS IN A DEVELOPMENT CONTEXT. ANALYSIS OF VARIANCE (ANOVA)

Dr. Maribel Ortego, Department of Applied Mathematics III, Universitat Politècnica de Catalunya

Enric Lloret, Civil engineer consultant, MSc in Technologies for Human Development candidate

INDEX

1 INTRODUCTION

Water, sanitation and hygiene (WASH) are essential for health, welfare and livelihoods. Increased access and better services lead to higher levels of school achievement and improved economic productivity. Yet too many people do not have these basic human rights.

Today in our world there are already 2.5 billion people lacking access to improved sanitation and 748 million people lacking access to an improved source of drinking water.

This situation has been recognized for many years and since 2000 when the Millennium Development Goals (MDG) were set up, important improvements have been made. To do so, the definition and monitoring of parameters and indicators was necessary through the years. Managing that information has been of huge importance for decision-making, since undesirable trends could be identified and measures against them could be set up.

Working with huge amounts of data requires the use of statistical tools in order to make the analysis affordable and to use these data as a decision-making tool.

In this context, the aim of this case study is to make students familiar with a statistical tool, Analysis of Variance (ANOVA) and specific software in statistics, working with real data from a survey in Mozambique.

1.1 DISCIPLINES COVERED

The main discipline covered by this case study is the analysis of real data through statistical techniques, including Analysis of Variance (ANOVA). The aim is that students are able to understand these techniques and apply them by using specialized software in order to analyze the data series and determine relationships between variables and trends.

This case study requires some basic knowledge of statistics. Finally the case study promotes teamwork since the class activity is to be completed in groups of 3 or 4 students and the conclusions shared and discussed in class.

1.2 LEARNING OUTCOMES

As a result of this case study, students are expected to be able to:

- Understand the problem of the lack of access to a safe drinking water source worldwide and its consequences on human development
- Be aware of the water quantity standards needed for human development and the relationship between consumption and distance to the source of water
- Know the difference between a safe and an unsafe drinking water source
- Know and use the Analysis of Variance (ANOVA) and Regression techniques as statistical tools to be applied to analyze data
- Apply statistical analysis to real survey data by using specialized statistics software

1.3 ACTIVITIES

ANOVA is one of thebasic statistical tools. ANOVA can be used in different ways (for a single or multiple factors). The proposed two activities allow students to understand how ANOVA works for a single factor. In the class activity students will work in groups of 3-4 people after some basic theoretical concepts are introduced by the lecturer. After this group work, a discussion will be set up and results and conclusions put in common. An individual activity is proposed as homework in order to consolidate learning.

2 DESCRIPTION OF THE CONTEXT

In this section a description of the context of the case study is given. First, the problem of the lack of access to a safe water source is briefly explained, as well as the goals to be accomplished following the MDGs and the efforts of monitoring its progress. The need of aid-decision tools is presented and the importance of statistics to do it explained. Some statistical techniques are introduced and applied to analyze real data series in the proposed activities.

2.1 INTRODUCTION: WATER ACCESS, JMP and WMP

In order to reduce the differences in access to drinking water sources around the world the Millennium Development Goals (MDG), set up in 2000, defined a goal to increase of the percentage of people with access to improved sources of drinking water from 76 % (total coverage in 1990) to 88% by 2015.

The indicator to be monitored (access to an improved source of water) was defined and data has been collected and analyzed annually in order to control its evolution. Working with huge amounts of data requires the use of statistical tools in order to make the analysis affordable and obtain a disaggregated view of the indicator (by country, by region, by gender, etc.) to be used as a decision-making tool.

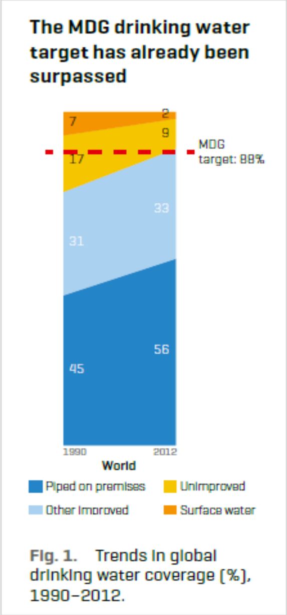


Figure 1. Trends in global drinking water coverage (%).

Disaggregated values of the indicator are essential to go beyond the global average and identify undesirable trends in specific regions or socio-economical classes, for example. Actually, this issue is relevant to the 2012 indicator value that shows 89% of the global population using an improved source of drinking water, which hides the real situation. Globally, this goal is reached (Figure 1) but the differences between regions are notable (Figures 2 and 3): nearly half of the 700 million people still lacking ready access to improved sources of drinking water are in sub-Saharan Africa (Figure 4).

Global country or regional values hide other types of inequalities (by social status, type of land, or gender, for example). As shown in Figure 5, having the disaggregated urban-rural values is necessary to fully understand the situation.

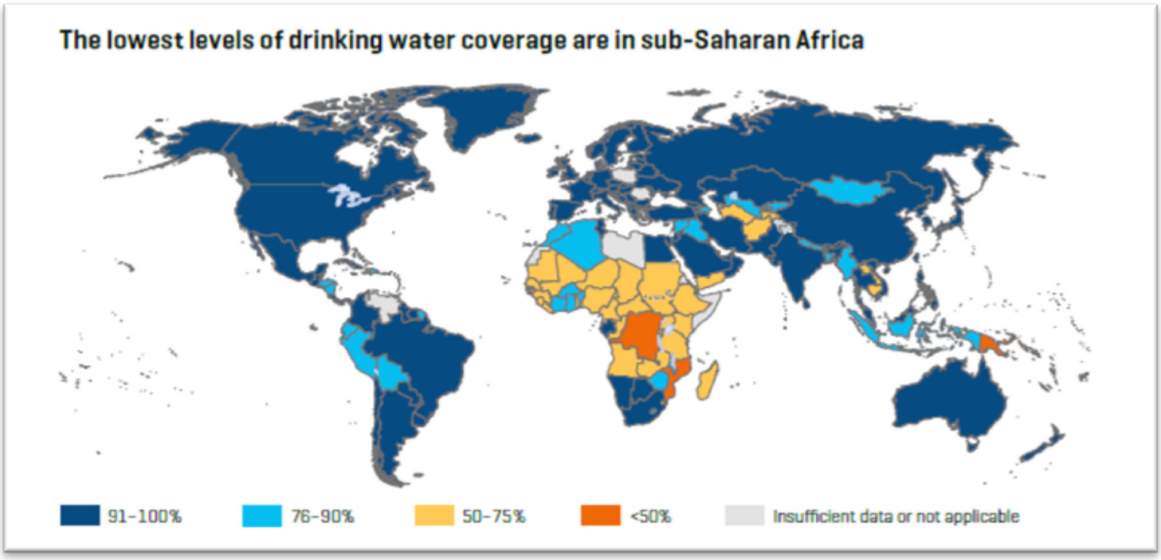

The lowest levels of drinking water coverage are in sub-Saharan Africa
91-100%
76-90%
50-75%
<50%
Insufficient data or not applicable

Figure 2. Proportion of the population using improved drinking water sources in 2012. Source : JMP 2014

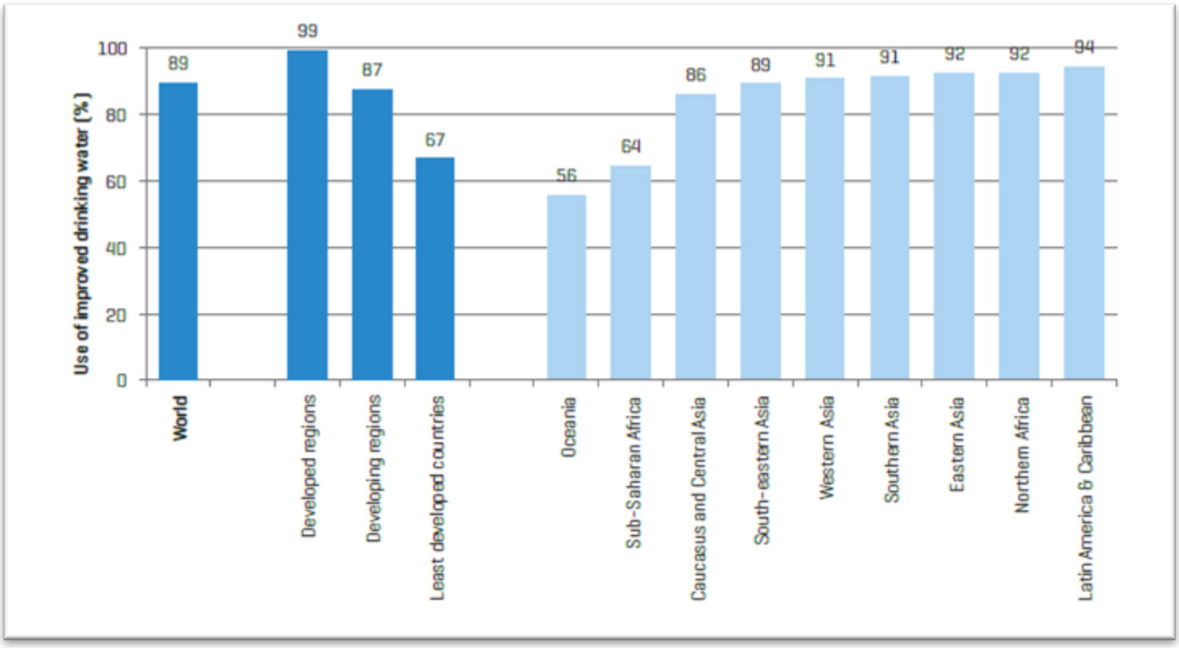


Figure 3. Use of improved drinking water sources in 2012 by regions. Source : JMP 2014

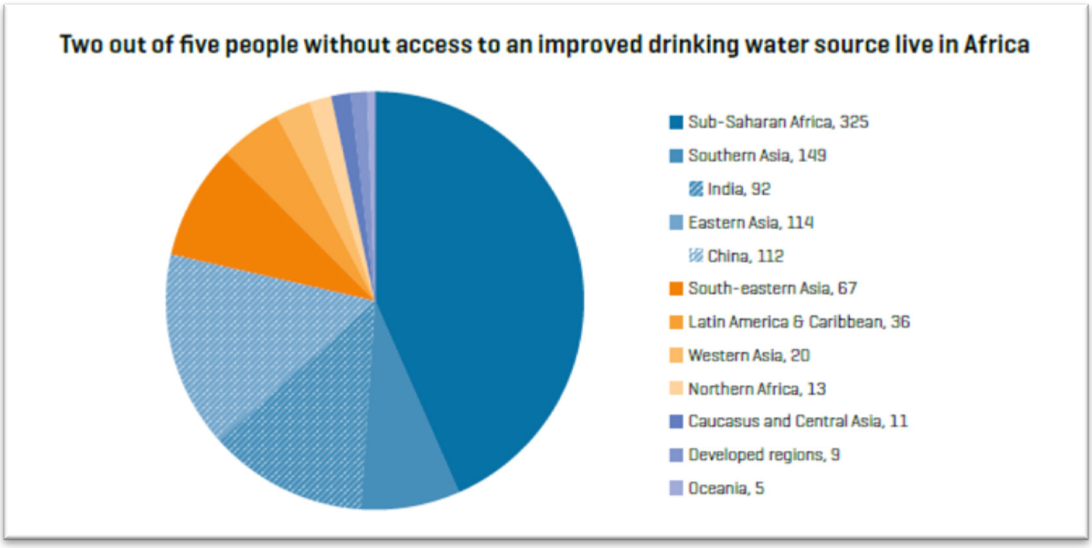


Figure 4. Number of people (in millions) without access to an improved drinking water source in 2012, by MDG region. Source: JMP 2014

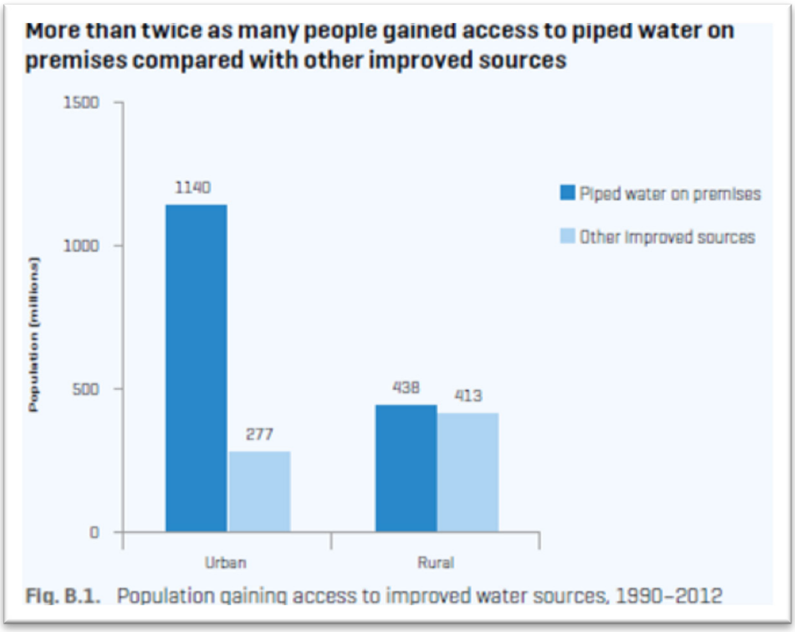


Figure 5. Population gaining access to improved water sources. 1990-2012.
Source: JMP 2014

2.2 JUSTIFICATION. Importance of STATISTICS in data analysis

To address the monitoring challenges in the run up to the MDG target year of 2015 and beyond, a strategy called Joint Monitoring Program (JMP) was formulated by WHO and UNICEF. The mission of the JMP is to be the trusted source of global, regional and national data on sustainable access to safe drinking water and basic sanitation for use by governments, donors, international organizations and civil society.

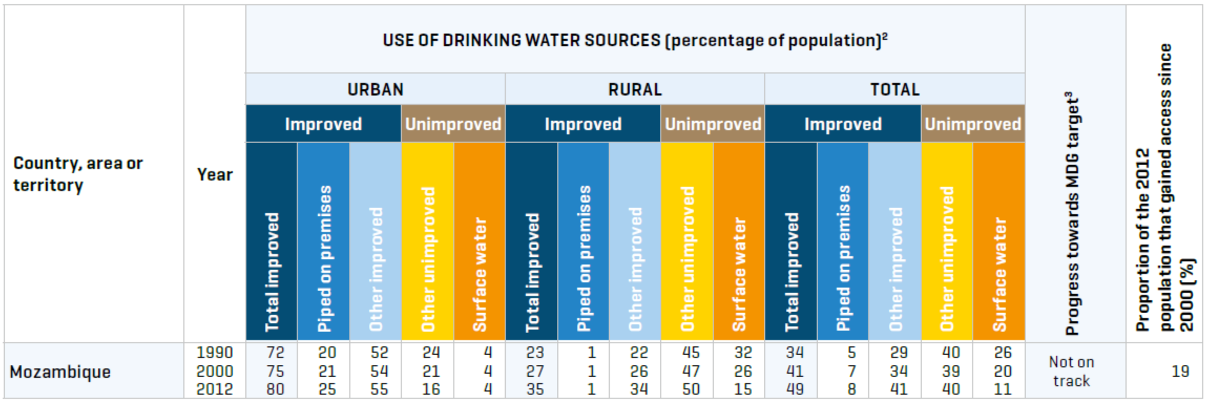
To fulfill its mission, the JMP has three strategic objectives. One of them is to compile, analyze and disseminate high quality, up-to-date, consistent and statistically sound global, regional and country estimates of progress towards internationally established drinking water and sanitation targets. This supports informed policy and decision-making by national governments, development partners and civil society. As seen above, the right use of statistical tools is necessary in order to reach the above objective.

One of these tools is the ANOVA (Analysis of Variance). Strictly, ANOVA is a collection of statistical models used to analyze the differences between group means. We can refer to ANOVA for a single factor or for multiple factors. In its simplest form, ANOVA provides a statistical test of whether or not the means of several groups are equal.

The aim of the current case study is for students to learn how to apply ANOVA for a single factor to a set of real data. The set of data is from a village in Mozambique, Southeast Africa.

Mozambique's estimated 2014 Human Development Index (HDI) was 0.393, ranked 178 (out of 187 countries) ordered by its HDI. That means Mozambique is in the set of countries classified as Low Human Development.

In relation to the use of drinking water sources, Mozambique has made a big effort to meet the MDG. Nevertheless, as seen in Figure 6, the country is not on track to meet these targets.



Country, area or territory	Year	USE OF DRINKING WATER SOURCES (percentage of population)[2]															Progress towards MDG target[3]	Proportion of the 2012 population that gained access since 2000 (%)
		URBAN					RURAL					TOTAL						
		Improved			Unimproved		Improved			Unimproved		Improved			Unimproved			
		Total improved	Piped on premises	Other improved	Other unimproved	Surface water	Total improved	Piped on premises	Other improved	Other unimproved	Surface water	Total improved	Piped on premises	Other improved	Other unimproved	Surface water		
Mozambique	1990	72	20	52	24	4	23	1	22	45	32	34	5	29	40	26	Not on track	19
	2000	75	21	54	21	4	27	1	26	47	26	41	7	34	39	20		
	2012	80	25	55	16	4	35	1	34	50	15	49	8	41	40	11		

Figure 6. Use of Drinking Water Sources in MOZAMNBIQUE. Source: JMP 2014

When the source of water is not inside the premises, people usually have to move (sometimes a long way to reach the source) to collect it. The task of collecting water is done in many cases by women and children, which exposes them to sexual or other forms of hazards. In addition, the time spent collecting water is time taken from studying or being at school (in the case of children).

In the proposed activity (next section), students will learn who (in which percentage) is the member of the family that collects the water, the quantities collected and the time taken for collection.

3 CLASS ACTIVITY

3.1 METHODOLOGY

This activity is prepared for a two hour session course plus 30 minutes of individual reading for the introduction and context. The activity is divided into three parts: a first introductory refresh of statistical concepts (30 min) to be developed by the lecturer, a second part of group work (1 hour) and a final part of gathering conclusions and discussion of results in class.

The session is to be carried out in a well-equipped computer classroom. Specific statistical software (called “R”) (R core team, 2014) will be introduced, understanding that the aim of the activity is not to manage this software but to show students the potentialities of this kind of resource.

In the following sections, we will introduce an “R” code (“R” commands –instructions-combination) that solves the activity. Instructions may vary if other statistical software is used.

3.2 PROPOSED STATEMENT

In a village of Mozambique, a survey was conducted to monitor the current WATSAN (Water and Sanitation) situation. Data was taken from 1229 households distributed across 18 districts. The available data are:

- The current water source situation of each household: piped on premises, or not.

- The person fetching water (fetching person) in the household: adult female, adult male, girl (age < 15 years old), or boy (<15 years old).

- The total number of members in each household.

- The amount of water consumption, in litres/person a day.

- The time to fetch water.

In this activity we will focus on the time to fetch water. The aim is to apply an ANOVA for a single factor (time of fetching water) and to determine if there is a marked difference between this time depending on who is the family member fetching it.

The referred dataset is available for the activity in a .csv format file that should be given to the students.

3.3 SOLUTION

In this section the steps to solve the activity are described. Each step is accompanied by a short explanation and the right instruction to be used in “R”.

Those steps are the followings:

1 Loading data:

```
#LOADING THE DATA FILE .CSV
#Clearing computer memory
rm(list=ls(all=TRUE))
#setting the working directory
directorio<-"..//"
setwd(directorio)
#loading the .CSV file
Datos <- read.table("MZB_HH_WaterSupply.csv", header=TRUE, sep=";", na.s
trings="NA", dec=".", strip.white=TRUE)
```

"R" code

2 Saving file to RData format (useful when data are gathered from different datasets)

```
save(Datos,file="DataMoz.RData")
```

"R" code

3 Plotting Fetching.person data frequency

```
plot (Datos$Fetching.person)
```

"R" code

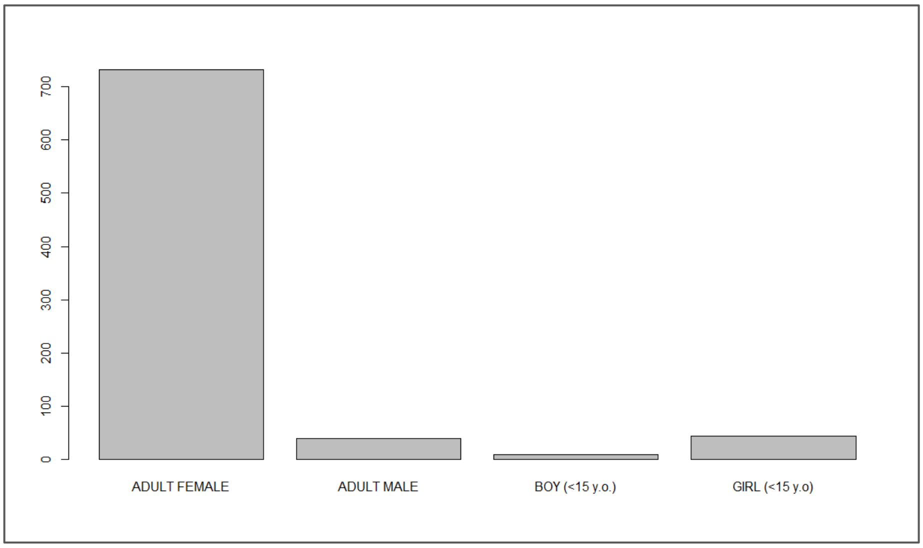


Figure 7. Fetching.person data frequency

In Figure 7 the different levels (four: adult female, adult male, boy (<15 y.o.) and girl (<15 y.o)) of the Fetching.person factor are shown. Clearly the majority of fetching people are female (either adult or girls).

4 Logarithmic transformation of "Time" variable

Fetching time is a variable with positive value. Its scale is relative, and therefore, it is better represented with a logarithmic scale. A new variable is created and included to the dataset:

```
Datos$logtime <- with(Datos, log(Time.to.fetch.water))
```

“R” code

5 Box-Plotting “logtime” for the several levels of “Fetching.person” factor.

In order to visually compare the mean values of (log)fetching time for the four levels of fetching person factor, we use a boxplot:

```
boxplot(logtime~Fetching.person, data=Datos, id.method="y",
main="Boxplot Fetching Person")
```

“R” code

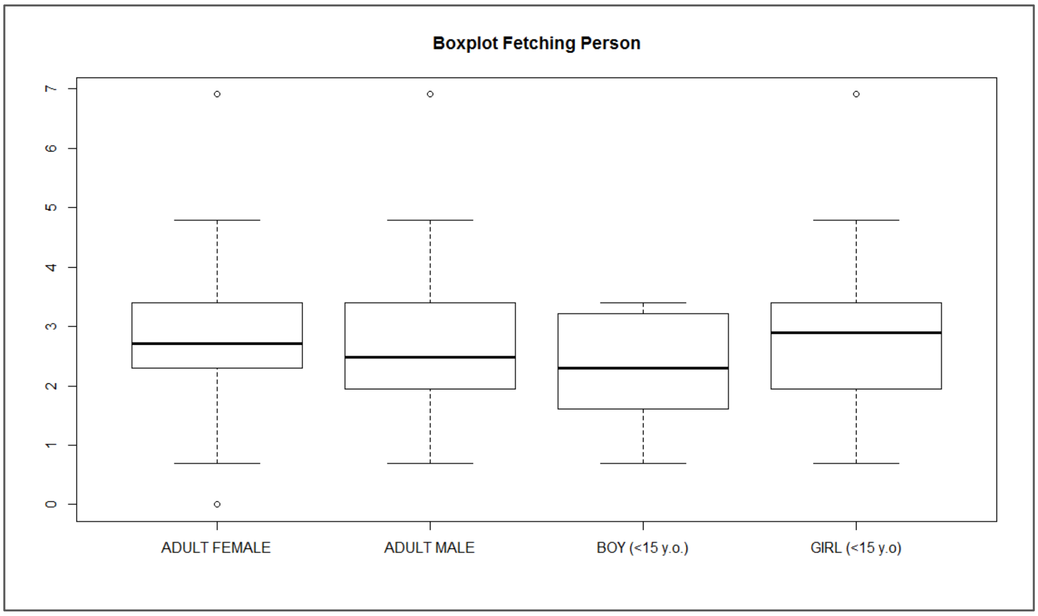


Figure 8. Boxplot of Fetching.person variable

In Figure 8, the data distribution ((log)-fetching.time) for each level of the factor is shown. The bold line represents the median and the whiskers give us an idea of the variability of data (variance or typical deviation). The size of the central box, determined by the first and third quantile, gives an idea of the variability of the central 50% of data.

As seen in the figure, the mean fetching times for each level of the factor do not seem very different, nor does the variability within each group. This means that the mean fetching time will not be different depending on which family member fetches the water (fetching person factor). Nevertheless, to confirm this some other statistical tests need to be carried out.

6 Homogeneity of variances

In this step a test of homogeneity of variances is performed. The null hypothesis is “all the variances are equal” and the alternative hypothesis is “at least one of the variances is different to the others”. That is:

$$\begin{cases} H_o\colon\ \sigma_1{}^2 = \sigma_2{}^2 = \sigma_3{}^2 = \sigma_4{}^2 \\ H_1\colon\ \quad at\ least\ one\ of\ them\ is\ \ different \end{cases}$$

The contrast is set up with an α value of 0.05.

```
bartlett.test(logtime~Fetching.person, data=Datos)
```

“R” code

```
Bartlett test of homogeneity of variances

data:  logtime by Fetching.person
Bartlett's K-squared = 3.9159, df = 3, p-value = 0.2707
```

The results show that the null hypothesis cannot be refused at level α , as the p-value > α, and, therefore the hypothesis of homogeneity of variances is accepted.

In order to compare means through an ANOVA procedure, variances must be homogeneous for all levels of the factor. Otherwise, a non-parametric alternative should be used, such as Kruskall-Wallis. In R, kruskal.test (variable ~ factor, data=nameofdataset).

7 Analysis of Variance (ANOVA)

The results of step 6 do not draw us to make any conclusion about the homogeneity of means. In order to do so, an ANOVA test will be carried out. The ANOVA procedure can be performed in R using the instruction aov(), but also lm().

ANOVA model assumptions need to be confirmed prior to any interpretation of the results. If any of the model assumptions is not satisfied, the results from the use of ANOVA technique may be unreliable. The model assumptions are:

a) Response variable residuals are normally distributed (or approximately normally distributed)
b) Samples are independent
c) Variances of populations are equal
d) Responses for a given group are independent and identically distributed normal random variables

In order to test the above assumptions, a graphical interpretation will be done, confirmed by a formal statistical test. .

```
AnovaModel.2 <- aov(logtime~Fetching.person, data=Datos)
```

"R" code

A visual interpretation can be performed by plotting the results of the ANOVA as follows.

```
par(mfrow=c(2,2))
plot(AnovaModel.2)
```

"R" code

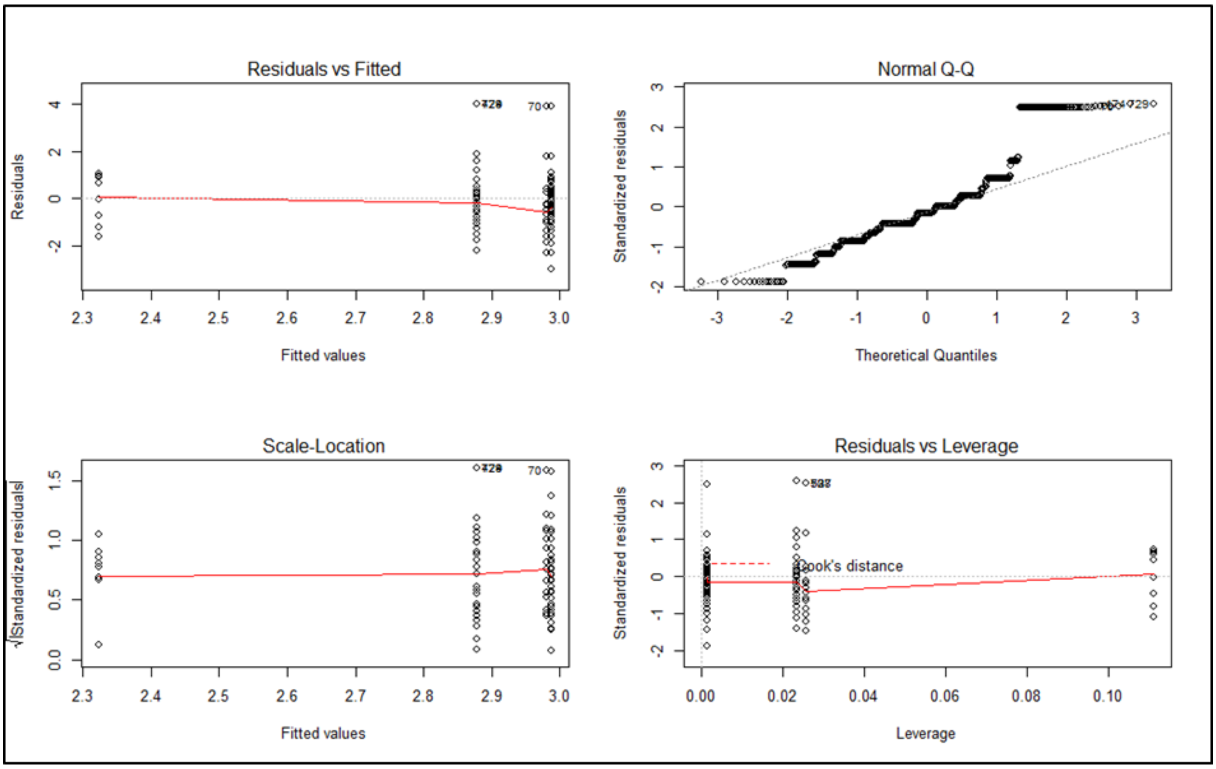


Figure 9. ANOVA assumptions plot

Apparently the assumed normal distribution of model residuals is not accomplished (Figure 9, top-right). A normality Shapiro test of goodness of fit is now carried out in order to complete this diagnostic.

```
> shap <- shapiro.test(AnovaModel.2$residuals);shap; shap$p.value
```

"R" code

```
  Shapiro-Wilk normality test

data:  AnovaModel.2$residuals
W = 0.8819, p-value < 2.2e-16

[1] 1.612136e-24
```

In this case p-value is less than α =0.05, and therefore the hypothesis of normality of the model residuals is rejected.

Normality of residuals is one of the assumptions of linear models. If this assumption is not accomplished, the decisions taken from tests should be made with care, as the distribution of the test statistic is not normal. For instance, the distribution of the ANOVA test statistic is not an F-distribution, and therefore the p-value should be interpreted with care.

For this dataset the error distribution is skewed by the presence of a few large outliers. These few extreme observations can have a great influence on parameter estimates of the model, as the estimation method is the minimization of the squared error. Outliers should be studied with care. There are a few possibilities for their appearance: error codes taken as values, transcription errors, non-adequacy of the linear model, or presence of two or more mixed populations, among others.

```
summary(AnovaModel.2)
```

“R” code

```
                 Df Sum Sq Mean Sq F value Pr(>F)
Fetching.person   3    4.3   1.439   0.575  0.632
Residuals       819 2050.2   2.503
406 observations deleted due to missingness
```

As normality of the model residuals has been rejected, the distribution of the test statistic is not an F-Fisher-Snedekor, but a distribution with heavier tails. A simulation study may be performed in order to obtain the exact value of the p-value for this unknown distribution. However, it is possible to make a decision about the equality of means between the different levels of the factor: the value of the p-value is substantially greater than α =0.05, and therefore the hypothesis of equality cannot be rejected at this α =0.05 level. That is, the mean values of fetching time are the same for all levels of the fetching person factor (women, girls, boys or other people in the family).

8 Anova using lm()

The homogeneity of means can also be assessed through the lm() procedure in R, as ANOVA can be rewritten as a linear model.

```
AnovaModel.lm <- lm(logtime~Fetching.person, data=Datos)
summary(AnovaModel.lm)
```

“R” code

```
Call:
lm(formula = logtime ~ Fetching.person, data = Datos)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.9867 -0.7986 -0.2787  0.4144  4.0287 

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                   2.986748   0.058479  51.074   <2e-16 *
**
Fetching.personADULT MALE    -0.006675   0.260012  -0.026    0.980    
Fetching.personBOY (<15 y.o.) -0.662524   0.530624  -1.249    0.212    
Fetching.personGIRL (<15 y.o) -0.108722   0.248265  -0.438    0.662    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.582 on 819 degrees of freedom
  (406 observations deleted due to missingness)
Multiple R-squared:  0.002101,	Adjusted R-squared:  -0.001554 
F-statistic: 0.5747 on 3 and 819 DF,  p-value: 0.6317
```

The result of the comparison of means for each level is the same using aov() or lm(), but the summary of results of the latter is useful when the diagnostic shows that means are different. In this case, the level (or levels) with a different mean show coefficients that cannot be considered null. Often, this information spares the need for a subsequent multiple comparison test. As for this dataset, the normality hypothesis has been rejected, the interpretation of the t-tests should therefore be taken with care. These t-tests assess if there is an increment/ decrement of the mean value due to each level of the factor. The distribution of the test statistic is not a t-Student distribution, but as p-values are substantially greater than α =0.05, the hypothesis of null coefficient cannot be rejected. Therefore, there is no difference in mean values of the fetching time for each level of the fetching person factor.

3.4 EVALUATION CRITERIA

The evaluation of the class activity is divided into two parts: the results of the activity performed by each group (marked up to 2.5 points) and the participation of the student in the class debate (up to 1.5 points).

The remaining 6 points are evaluated on the homework activity.

The evaluation criteria for the class activity are defined below:

- The group has not been capable of writing a code in “R”: 0 points

- The group has written a code in “R” related to the subject and the majority of sections are well structured: 1 point
- The group has written a code in “R” that strongly matches with the one developed in class: 2,5 points

4 HOMEWORK ACTIVITY

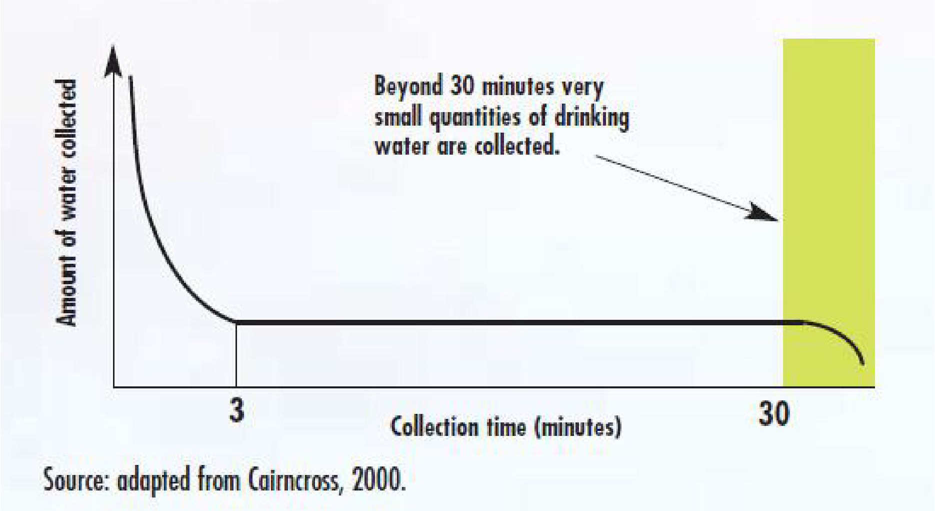
4.1 METHODOLOGY

Homework activity is prepared to be done at home as individual work in order to consolidate the learned concepts. A computer with “R” installed is needed to complete the activity.

The working time estimation is two hours. A lecturer feedback is to be given some weeks after the activity handed to the student.

4.2 PROPOSED STATEMENT

A relationship between the time of fetching water from the source and the amount water consumption has been pointed out in the specialized literature. The following figure shows that relationship:



Source: adapted from Cairncross, 2000.

Figure 10. Water consumption and time to source.

As seen, for a fetching time less than 3 minutes (source of water in premises, for example) the amount of water collected is high, decreasing as the time to reach the source of water increases. For a fetching time between 3 and 30 minutes the amount of water collected stabilizes and above 30 minutes the amount of water collected decreases.

The activity consists of:

- Represent (plot) the relationship between the fetching time and the amount of water collected for the real data from Mozambique.
- Perform an ANOVA test to know whether the mean amount of water collected is the same regardless of the fetching time, or whether it depends on it (as shown in Figure 10).

4.3 SOLUTION

4.3.1 LOGARITHMIC TRANSFORMATION OF “WATER CONSUMPTION_SCORE” VARIABLE

As in the case of the variable ‘Time needed to fetch water’, the variable ‘Water consumption’ has a positive value, and its scale can be considered as relative. Therefore, it is better represented if transformed by means of a logarithmic transformation. A new variable is created and included to the dataset:

```
Datos$logWaterCons <- with(Datos, log(Water.Consumption))
```

“R” code

4.3.2 FETCHING TIME vs WATER COLLECTED (plot)

The variables to be plotted are “ log Water Consumption” and “Fetching time 2”.

The “R” instructions to be used are: **plot()**, and **par()** to define the plotting-window distribution.

```
par(mfrow=c(1,1))
plot(Datos$Fetching.time3, Datos$logWaterCons, xlab = "Fetching_Time"
, ylab = "Q collected")
```

“R” code

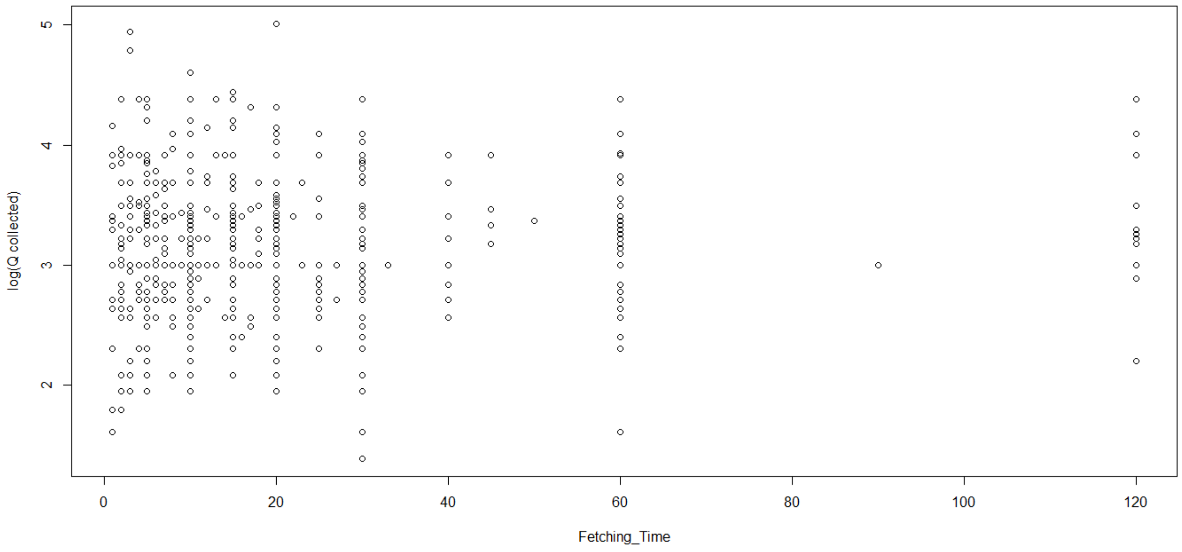


Figure 11. Fetching time vs Water collected

Figure 11 does not give us a clear idea about the relationship between these two factors in order us to assess if a greater time to fetch water implies a lower quantity of water collected. This is because, in general there is a great variability of water consumption among the households with the same fetching time

This variability among households with the same fetching time is better summarized in Figure 12. In order to illustrate this variability, fetching time has been considered as a factor and the boxplot of (log)-water consumption has been represented for each level of the factor.

```
boxplot(logWaterCons~Fetching.time3,data=Datos, id.method="y", main="
Boxplot Water Consumption vs Time", xlab = "Fetching_Time (min)", yla
b = "Q collected (l)")
```

“R” code

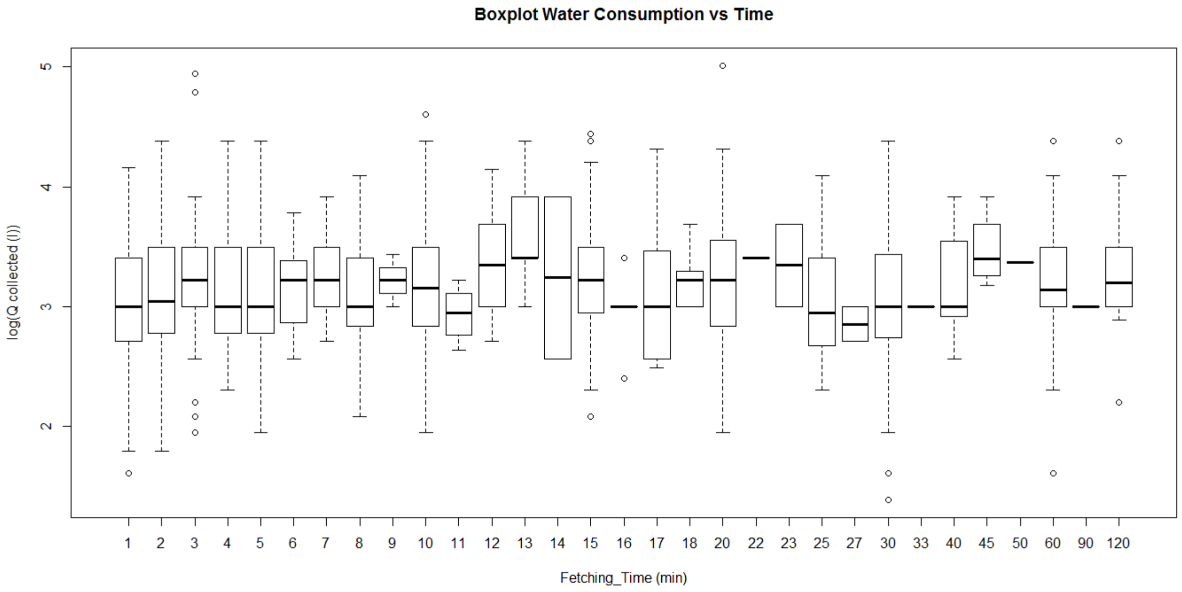


Figure 12. Boxplot Water collected vs Fetching time

The boxplots of (log)-water collected for each fetching time in Figure 12 show similar mean values for all levels of the factor, therefore it is not useful for giving a better insight into the suggested relationship between these two variables.

It should be noted that fetching times are very specific for small values (values < 20 minutes) but as they increase, times are usually rounded: half an hour, an hour, one and a half hours, for instance. This rounding should be taken into account in the interpretation of the results during further analysis.

4.3.3 ANOVA TEST

In order to practice the use of the ANOVA technique, the original statement of the problem has been slightly changed. Taking into account the rounding of values, a new variable has been defined from the fetching time variable. This variable is “time-factor” and classifies the fetching time in three levels: low (< 3 minutes), medium (time to fetch water is between 3 and 30 minutes) and high (for fetching time > 30 minutes).

In this section we will use the ANOVA test in order to analyze whether the average of water consumption depends on the level of the fetching time factor, or whether it is equal for each of them. For this exercise, α is set to α=0.01.

Figure 13 shows data frequency for each level of the time factor.

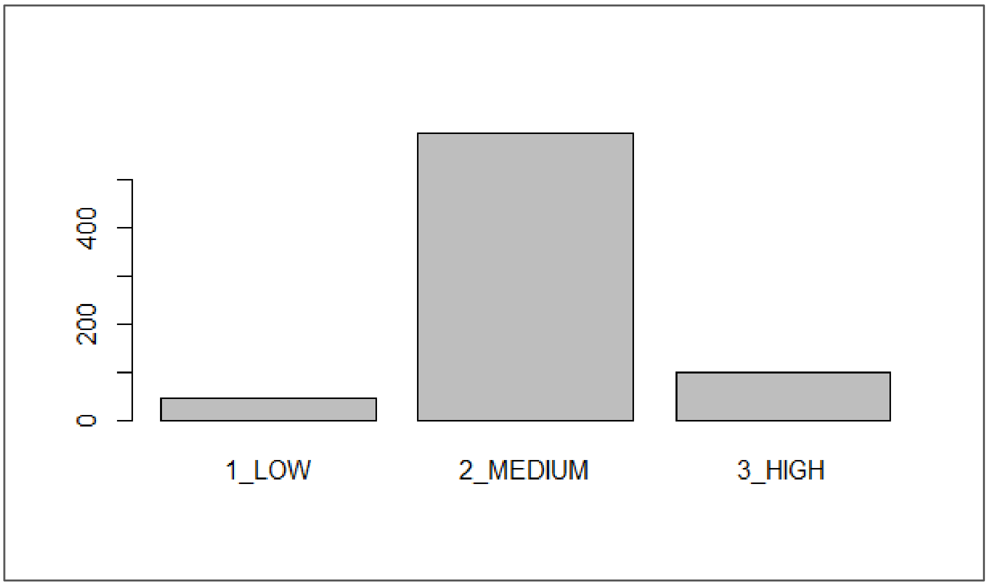


Figure 13. Time-factor variable frequency

A first visual representation of the differences in means for the amount of water collected for each level of the time factor is shown in the boxplot in Figure 14. Visually, mean values are not very different, but variability for the medium level of the factor seems greater than for the other categories. This medium level includes a wide range of situations, and the observed outliers add extra variability.

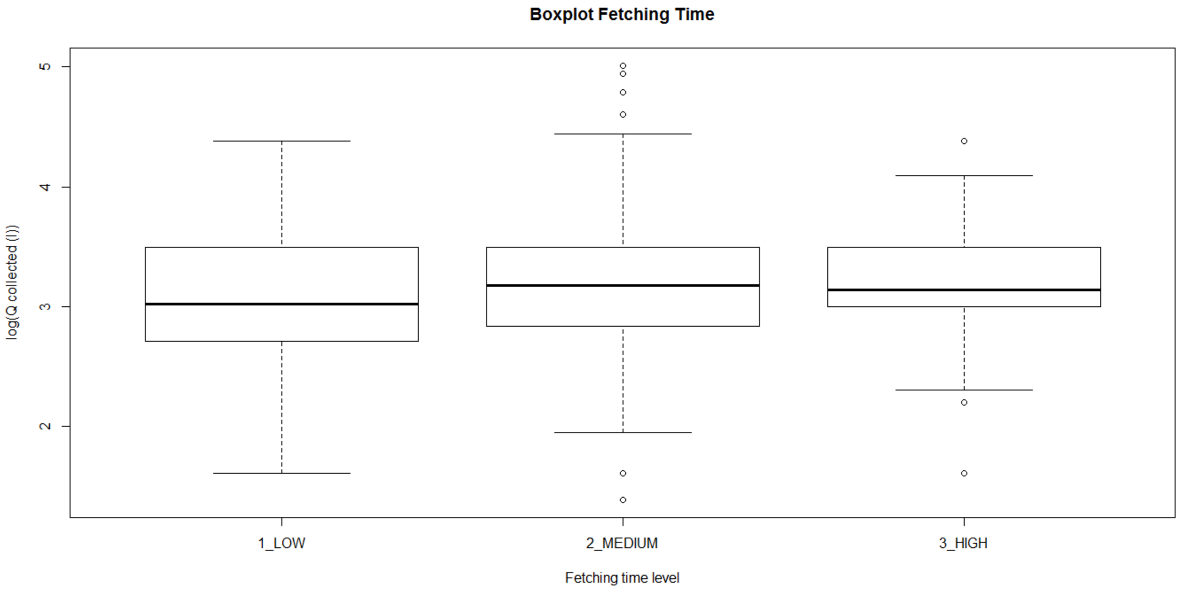


Figure 14. Boxplot of fetching time

4.3.3.1 VARIANCE HOMOGENEITY TEST

In order to assess the homogeneity of variances hypothesis, a test is performed:

$$\begin{cases} H_o\colon \ \sigma_1{}^2 = \sigma_2{}^2 = \sigma_3{}^2 = \sigma_4{}^2 \\ H_1\colon \qquad any\ of\ them\ is\ different \end{cases}$$

The contrast is set up with an α value of 0.01.

```
bartlett.test(logWaterCons~Time.factor, data=Datos)
```

“R” code

```
        Bartlett test of homogeneity of variances

data:  logWaterCons by Time.factor
Bartlett's K-squared = 6.4068, df = 2, p-value = 0.04062
```

The null hypothesis of homogeneity cannot be rejected at this α level, as p-value > α=0.01 , so variances can be considered equal.

4.3.3.2 TESTING ANOVA ASSUMPTIONS

Assuming that variances of the amount of water consumed are equal for all the levels of the time factor, we apply ANOVA in order to assess if the mean values for all levels are the same.

```
AnovaModel.3 <- aov(logWaterCons~Time.factor, data=Datos)
```

“R” code

Before drawing conclusions from the given results, ANOVA assumptions need to be confirmed.

A visual interpretation can be performed by plotting the diagnostics of the ANOVA as follows.

```
par(mfrow=c(2,2))
plot(AnovaModel.3)
```

“R” code

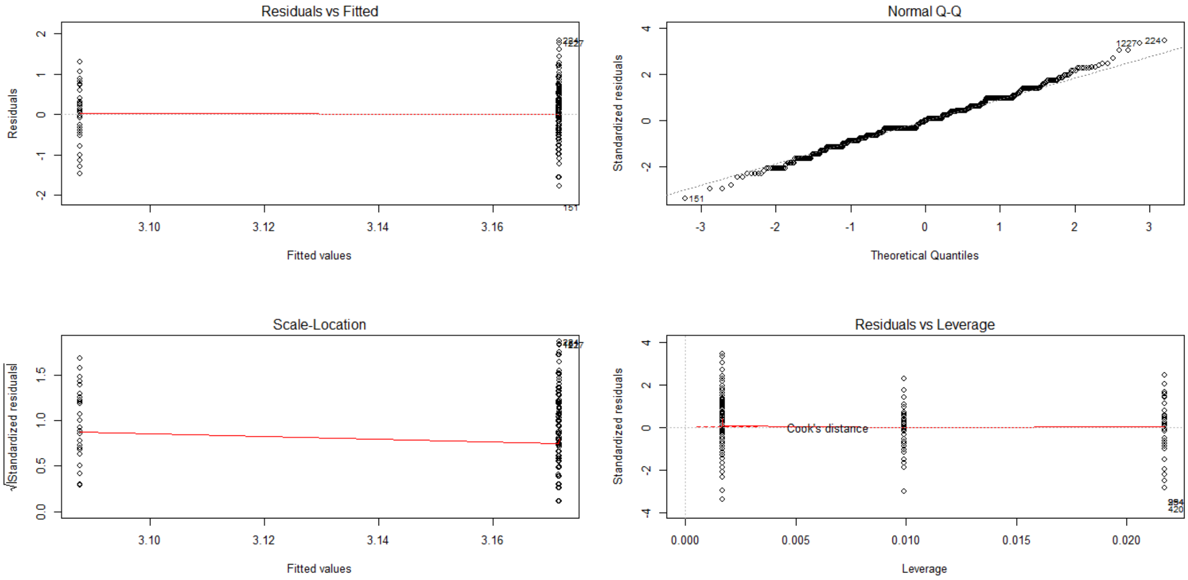


Figure 15. Anova results plotting

Apparently the normal distribution is not fitting data residuals (top-right figure), mainly for the higher values of water consumption. It seems that the suitable distribution for these residuals should have a heavier tail than the normal distribution.

4.3.3.3 NORMAL GOODNESS OF FIT FOR RESIDUALS

A Shapiro test is now carried out in order to complete the diagnostic of goodness of fit of the residuals.

```
> shap <- shapiro.test(AnovaModel.3$residuals);shap; shap$p.value
```

“R” code

```
        Shapiro-Wilk normality test

data:  AnovaModel.3$residuals
W = 0.9916, p-value = 0.0003219

[1] 0.0003219133
```

In this case p-value is less than α =0.01 and then the normality hypothesis is rejected. Model residuals do not follow a normal distribution.

4.3.3.4 ANOVA DIAGNOSTIC

The summary of results of the ANOVA test is:

```
summary(AnovaModel.3)
```

"R" code

```
                Df Sum Sq Mean Sq F value Pr(>F)
Time.factor      2    0.3  0.1513   0.542  0.582
Residuals      741  207.0  0.2793
485 observations deleted due to missingness
```

As model residuals are non-normal, the distribution of the ANOVA test is not F-Fisher-Snedekor.

A simulation study may be performed in order to obtain the exact value of the p-value for this unknown distribution. However, the p-value is substantially greater than α=0.01, and therefore the hypothesis of mean homogeneity should not be rejected. Therefore, the means of water consumption are the same for each level of the fetching time factor

4.3.3.5 What if α=0.05?

The value of α needs to be fixed prior to the performance of the test. Values such as α=0.01, 0.05 or 0.1 are common, but it should be noted that the power of the test could is also linked with this selection of α.

If a value of α= 0.05 is fixed for this ANOVA procedure, results vary from the ones presented in sections 4.3.2.1 to 4.3.2.4:

- Variance homogeneity test (section 4.3.2.1.):

For α= 0.05, the homogeneity of variances is rejected, as p-value = 0.04062 < α= 0.05. Therefore, variances should be considered as different.

- Equality of mean values for each level of the factor:

As variances are different, a non-parametric comparison procedure has to be used, instead of ANOVA

```
> kruskal.test(logWaterCons ~ Time.factor, data=Datos)
```

"R" code

```
	Kruskal-Wallis rank sum test

data:  logWaterCons by Time.factor
Kruskal-Wallis chi-squared = 0.6453, df = 2, p-value = 0.7242
```

For α= 0.05, the equality of means of water consumption for each of the levels of the time factor cannot be rejected. Although the variability of the amount of water consumed is different among these levels, the mean values cannot be considered as different.

4.4 EVALUATION CRITERIA

The evaluation criteria for this activity are specified below:

- The student has not sent any report for the activity : 0 points
- The activity is performed and a report has been sent but it does not follow a logical structure to solve the activity: 1 point
- The student has done the activity and sent a well-structured report to solve the activity 4 points
- The activity is performed, the report has been sent and is well logical structured in order to solve the activity and the results are mainly match the ones attended: 6 points

5 BIBLIOGRAPHY

- Arriaga Gómez, A. J.; Fernández Palacín, F.; López Sánchez, M. A.; Muñoz Márquez, M.; Pérez Plaza, S.; Sánchez Navas, A., 2008. “Estadística Básica con R y R-commander”. Servicio de Publicaciones de la Universidad de Cádiz
- Cairncross S and Feachem R, 1993, Environmental health engineering in the tropics: an introductory text (2nd edition). John Wiley and Sons, Chichester, UK.
- Practical Regression and Anova in R, JJ Faraway, 2002. URL: http://www.maths.bath.ac.uk/~jjf23/book/
- R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/.
- The effect of scale in daily precipitation hazard assessment. Egozcue, J. J., Pawlowsky-Glahn, V., Ortego, M. I. and Tolosana-Delgado, R. Natural Hazards and Earth System Science, Vol. 6, pp. 459-470 (2006) (www.nat-hazards-earth-syst-sci.net/6/459/2006/).
- WHO et UNICEF, 2014. “Progress on Drinking water and Sanitation. 2014 update.”

6 FURTHER WORK

The suggested activity allows for the introduction of the simplest form of ANOVA, 1-factor, to students. However, ANOVA with 2 or more factors or ANCOVA can also be introduced using the same dataset. Even linear and logistic regression methods can be introduced using this framework, to answer questions related to the problem.

GDEE
GLOBAL DIMENSION IN ENGINEERING EDUCATION

http://www.gdee.eu

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

United Nations Educational, Scientific and Cultural Organization
UNESCO Chair in Engineering for Human and Sustainable Development
UNIVERSITY OF TRENTO - Italy

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

POLITÉCNICA "Ingeniamos el futuro"
CAMPUS DE EXCELENCIA INTERNACIONAL

Loughborough University

ONGAWA
INGENIERÍA PARA EL DESARROLLO HUMANO

Centro per la Formazione alla Solidarietà Internazionale
Training Centre for International Cooperation

PRACTICAL ACTION
Technology challenging poverty

engineers without borders uk

This project is funded by