Advanced Business Data Analysis

CA1

Student number - 17125286

Table of Contents

Introduction

Project is made for ‘Advanced Business Data Analysis’ module and it consist 3 separate statistical tests:

- Student's t-Test (Paired or Unpaired) – Chapter 1
- One-way ANOVA– Chapter 2
- Two-way ANOVA – Chapter 3

Environment: Microsoft Excel, SPSS and R studio.

1. Student T-Test – Airbnb cost of entire apartment per day, Dublin v Paris

Data is coming from http://insideairbnb.com/get-the-data.html and it is publicly available data from Airbnb.com. Datasets are dataset published on 18 February 2017 containing data of available apartments on that date for Dublin. And second dataset is published on 4th April 2017 containing data of available apartments on that date for Paris.

Both datasets has been merging together for the purpose of this assignment to perform Student T – Test. Sample of 50 each entire houses for rent has been randomly selected. We perform student T test unpaired as this is simple comparison of means of 2 groups.

Price is a price for entire house per day in euro.

1.1 Hypotheses

Unpaired t-Test – we will compare means of two samples of rental price in Dublin and Paris. Two categories of cities, 50 apartments in Dublin and different 50 apartments in Paris.

The Null Hypothesis is:

$H_0: \mu_1 = \mu_2$ (No difference between populations, the means of rental price are equal for both Dublin and Paris)

The Alternate Hypothesis is:

$H_1: \bar{X}_1 \neq \bar{X}_2$ (There is a difference between samples, the means of sample of rental price are not equal for both Dublin and Paris)

1.2 F-Test

- $F_{stat} = 5.93$
- Degrees of freedom:
- $DF_1 = 49$
- $DF_2 = 49$
- $\alpha = 0.05$
- $F_{crit} = 1.61$

$F_{stat} > F_{crit}$

Therefore sample variances are “unequal”

F-Test Two-Sample for Variances

	Dublin	*Paris*
Mean	161.46	84
Variance	8092.58	1365.265
Observations	50	50
df	49	49
F	5.927478	
P(F<=f) one-tail	2.46E-09	
F Critical one-tail	1.607289	

Table 1. Excel – F – Test

1.3 Determine α value

α value has been set up at 0.05, as the prices of entire apartment are not very sensitive data and 95% assurance that results are correct and type I error are not made will be enough.

1.4 T-Test

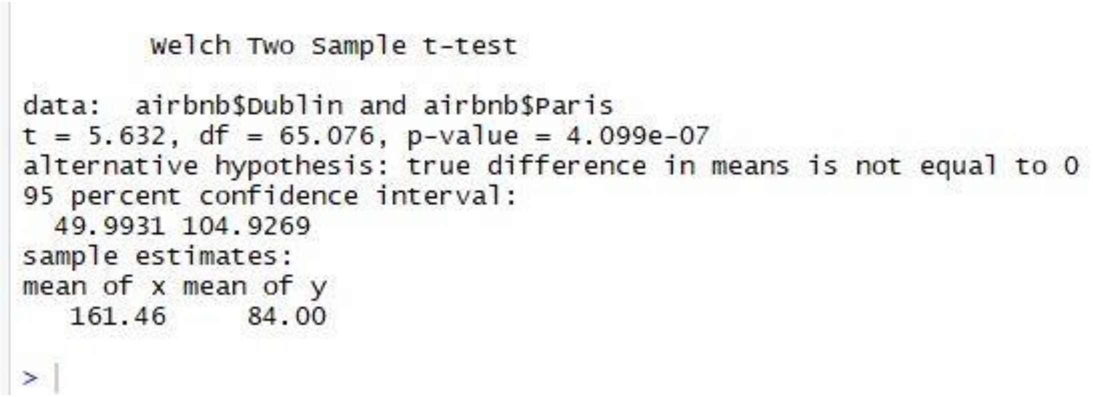
- Excel Result:

t-Test: Two-Sample Assuming Unequal Variances

	Dublin	*Paris*
Mean	161.46	84
Variance	8092.58	1365.265
Observations	50	50
Hypothesized Mean Difference	0	
df	65	
t Stat	5.632049	
P(T<=t) one-tail	2.06E-07	
t Critical one-tail	1.668636	
P(T<=t) two-tail	4.11E-07	
t Critical two-tail	1.997138	

Table 2 – Excel T -Test

- R studio Result:



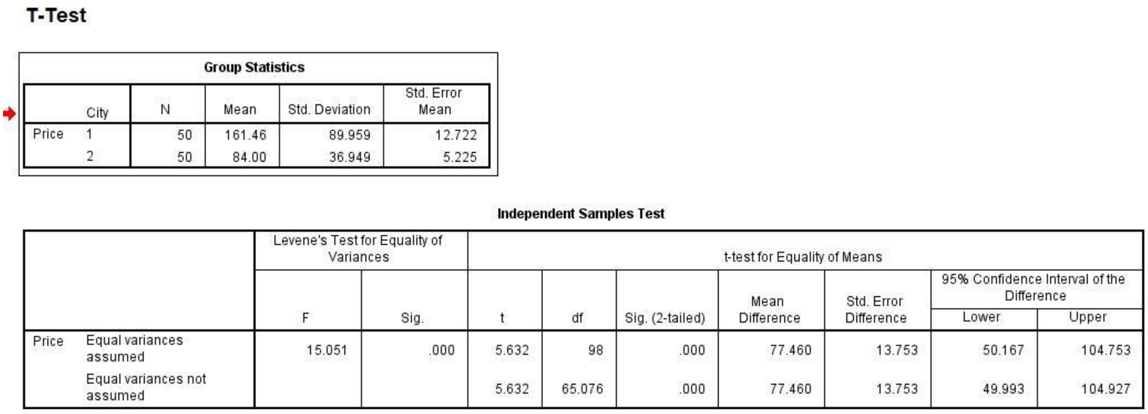
```
        Welch Two Sample t-test

data:  airbnb$Dublin and airbnb$Paris
t = 5.632, df = 65.076, p-value = 4.099e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  49.9931 104.9269
sample estimates:
mean of x mean of y
   161.46     84.00

>
```

Test 1 – R studio T - test

- SPSS result:



T-Test

Group Statistics

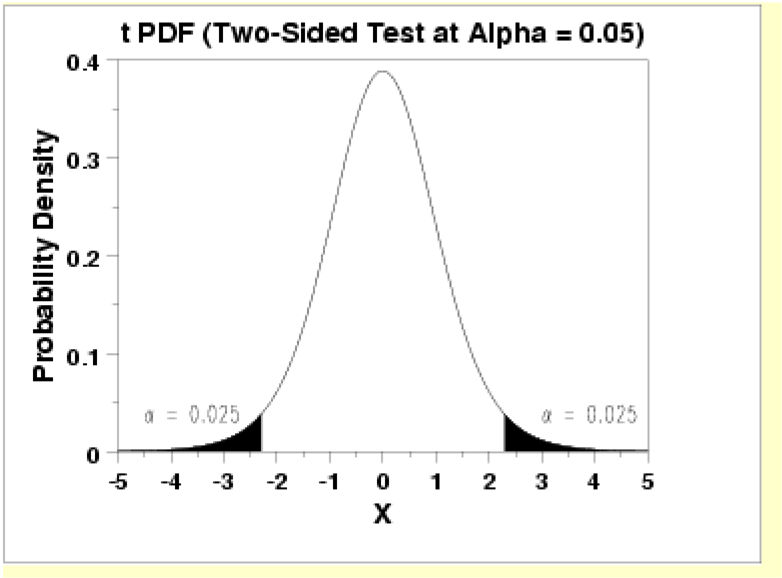
	City	N	Mean	Std. Deviation	Std. Error Mean
Price	1	50	161.46	89.959	12.722
	2	50	84.00	36.949	5.225

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Price	Equal variances assumed	15.051	.000	5.632	98	.000	77.460	13.753	50.167	104.753
	Equal variances not assumed			5.632	65.076	.000	77.460	13.753	49.993	104.927

Table 3 – SPSS T - TEST

1.5 Report:



Graph1 the significance level - α[1]

Result:

t = 5.632

α = 0.05

DF = 65

Report: t(65) = 5.632, p > 0.05

T critical = 1.668636

[1] http://www.itl.nist.gov/div898/handbook/eda/section3/eda3672.htm [Accessed 07/03/2018]

tstat > t crit – therefore t is significant and I did find a difference between prices of entire apartment for Dublin and Paris on Airbnb website.

1.6 Conclusions

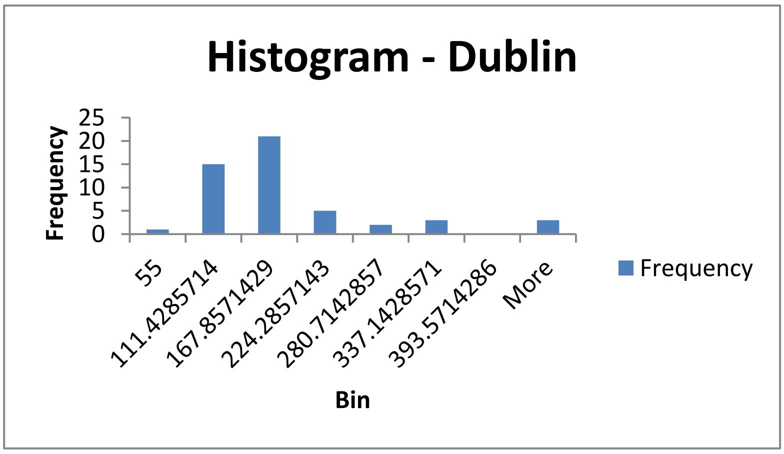
We reject null hypothesis in favour of Alternative Hypothesis. That means are not equal for both cities for rental price on Airbnb portal.

We find significant difference so we can say that prices in Dublin and Paris are not this same. I am afraid that Dublin is much more expensive for tourist than Paris nowadays. We can see that in descriptive statistics above:

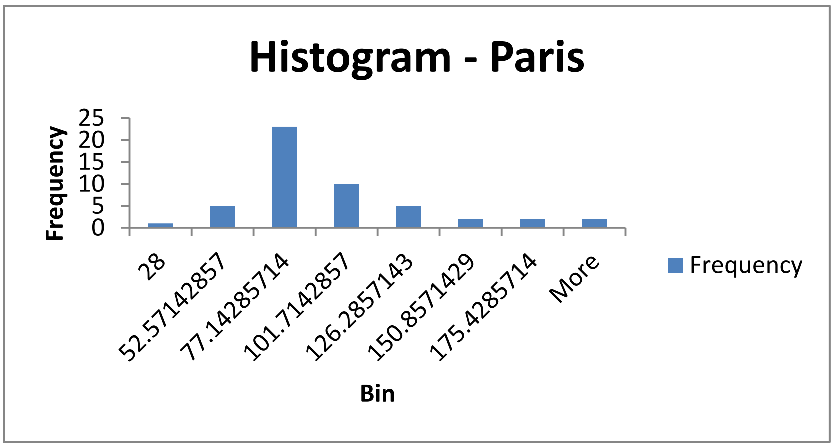
Dublin		*Paris*	
Mean	161.46	Mean	84
Standard Error	12.72209102	Standard Error	5.225447935
Median	127	Median	70
Mode	120	Mode	60
Standard Deviation	89.95876833	Standard Deviation	36.9494967
Sample Variance	8092.58	Sample Variance	1365.265306
Kurtosis	3.427772241	Kurtosis	1.609741502
Skewness	1.929422392	Skewness	1.359175716
Range	395	Range	172
Minimum	55	Minimum	28
Maximum	450	Maximum	200
Sum	8073	Sum	4200
Count	50	Count	50

Table 4 – Excel Descriptive statistics

We can see that difference also in histograms and distribution of data.



Histogram1 – Distribution of data of price of apartment per day in Dublin.



Histogram2 - Distribution of data of price of apartment per day in Paris.

On the histogram 1 and 2 we can see distribution of data. Distribution is normal - right – skewed in both cities.

Of course that's only a data from one booking portal. I would check it on hotel rates and available rates further. But that would explain the big demand for apartments for rent for Airbnb website in Dublin.

2. One-way ANOVA - Prices of 3 kinds on Lentil in the one of poorest state of India – Bihar

Dataset is a data downloaded from https://data.humdata.org/dataset/wfp-food-prices it was gathered by World Food Programme (http://www1.wfp.org/). It was published 05/12/2017.

'WFP is the world's largest humanitarian agency fighting hunger worldwide, delivering food assistance in emergencies and working with communities to improve nutrition and build resilience. Each year, WFP assists some 80 million people in around 75 countries.'[2]

It contains new and old data about price of different food in poor countries. In this work we will focus on Bihar in India and very popular food in this region – lentil. We have 3 kinds of lentil and we will check is the price depends from the lentil in Bihar. Lentil 1 is a masur, lentil 2 – moong, lentil 3 is an urad. Only these 3 kinds were checked by World Food programme in India. Possible that only these are available in this region – further research is needed.

We extracted randomly selected sample of 34 entries for each lentil. 104 records together. Prices are in Indian rupee and per kilogram.

Lentil is a big export product and also a main ingredient in many Indian dishes. It is a basis for Indian cuisine.

[2] http://www1.wfp.org/ [Accessed: 01/03/2018]

One-way ANOVA is perfect for this test as we have more than 2 groups for mean comparison and we have only 1 factor – kind of lentil.

1.7 Hypotheses

Null Hypothesis - $H_0 : \mu_l = \mu_2 = \mu_3$

There is no difference between prices of different kind of lentil in Bihar.

Alternative Hypothesis - $H_1 : \mu_l \neq \mu_2 \neq \mu_3$

There is a difference between prices of different kind of lentil in Bihar.

1.8 One – way ANOVA

Microsoft Excel result:

Anova: Single Factor

SUMMARY

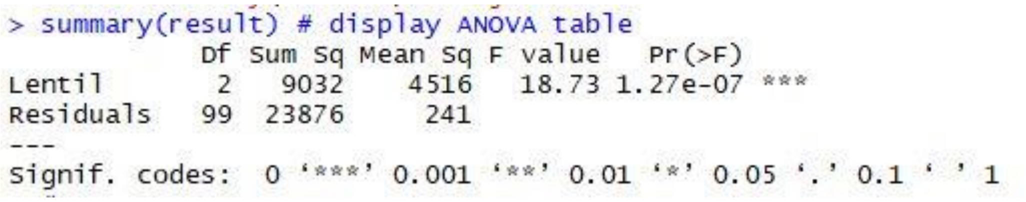
Groups	*Count*	*Sum*	*Average*	*Variance*
Lentil1	34	2465.22	72.50647	396.6288
lentil2	34	2417.39	71.09971	132.1314
Lentil 3	34	3118.75	91.72794	194.7432

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	9032.304	2	4516.152	18.72618	1.27E-07	3.08824
Within Groups	23875.61	99	241.1678			
Total	32907.91	101				

Table 5 – Excel ANOVA

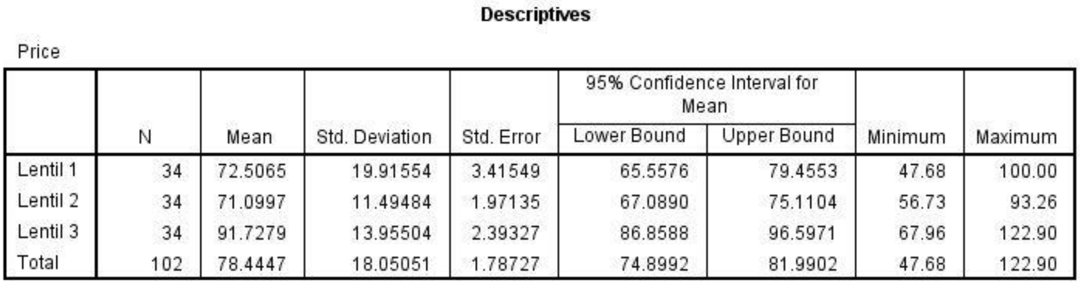
R studio result:



```
> summary(result) # display ANOVA table
             Df Sum Sq Mean Sq F value   Pr(>F)    
Lentil        2   9032    4516   18.73 1.27e-07 ***
Residuals    99  23876     241                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test2 – R studio ANOVA

SPSS result:



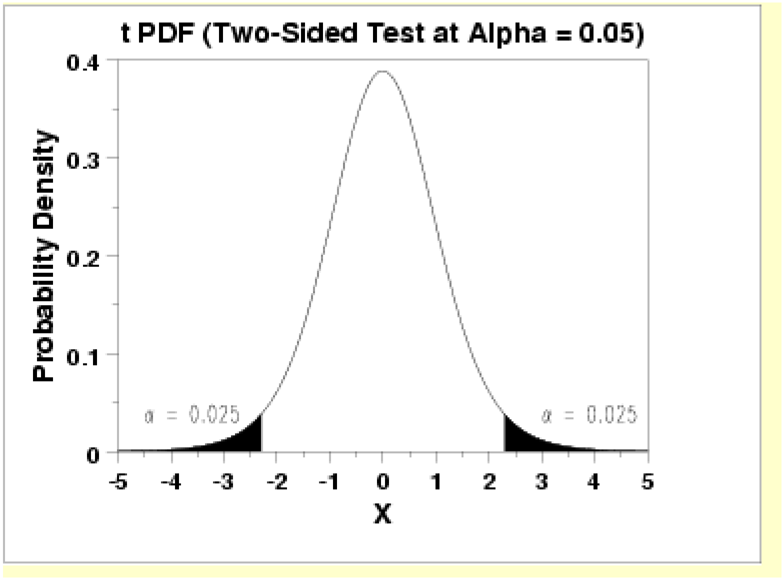
Descriptives

Price

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Lentil 1	34	72.5065	19.91554	3.41549	65.5576	79.4553	47.68	100.00
Lentil 2	34	71.0997	11.49484	1.97135	67.0890	75.1104	56.73	93.26
Lentil 3	34	91.7279	13.95504	2.39327	86.8588	96.5971	67.96	122.90
Total	102	78.4447	18.05051	1.78727	74.8992	81.9902	47.68	122.90

Table 6 – Descriptives – SPSS

1.7 Report



Graph1 the significance level - α[3]

[3] http://www.itl.nist.gov/div898/handbook/eda/section3/eda3672.htm [Accessed 07/03/2018]

Result:

$F = 18.73$

$\alpha = 0.05$

$DF_1 = 2$

$DF_2 = 99$

Report: $F(2, 99) = 18.73$, $p < 0.05$

F critical = 3.08

1.8 Determine α value

α value has been set up at 0.05, as the prices of entire apartment are not very sensitive data and 95% assurance that results are correct and type I error are not made will be enough.

1.9 Conclusions

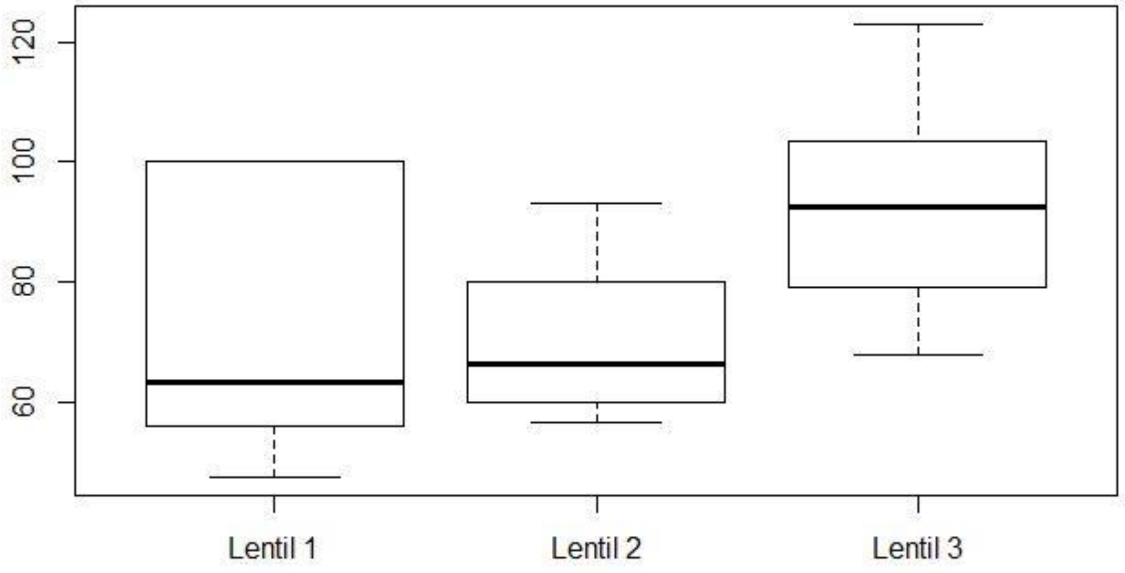
A one-way ANOVA to compare the three groups was performed. This analysis produced a statistically significant result: $F(2, 99) = 18.73$, $p < 0.05$

Reject H0 Null Hypothesis ($\alpha = 0.05$), Fstat > Fcrit

At least two of the three groups are significantly different.

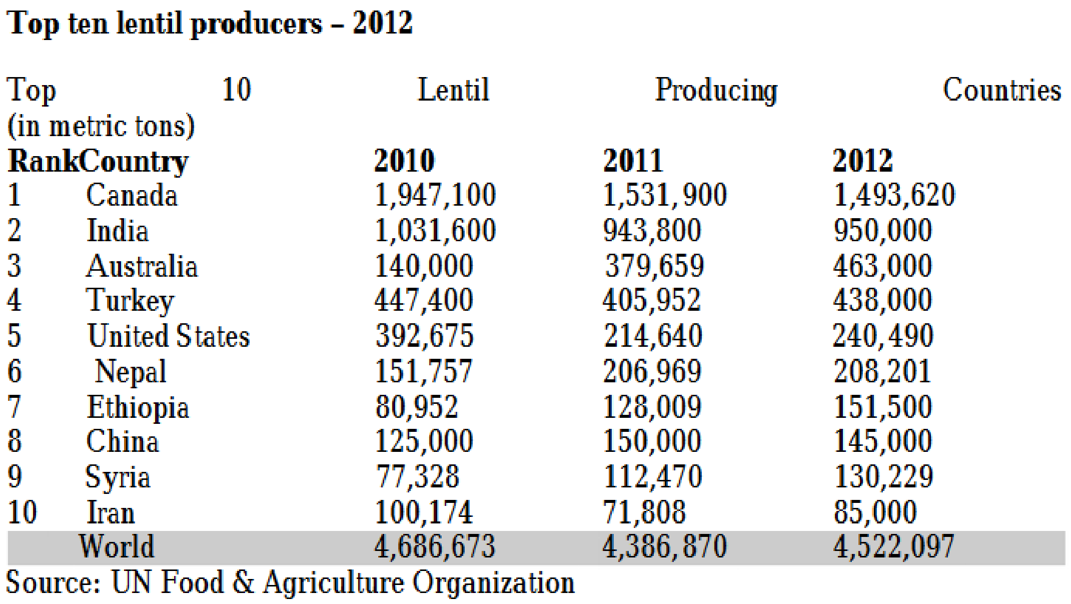
It means that not all kind of lentils in Bihar have this same price.

Post hoc Tukey tests revealed that the only significant difference between groups was found between group 1 and group 3.



Boxplot1 - Lentil 1 - masur, lentil 2 - moong, lentil 3 - urad.

Box and whisker diagram above shows 3 kinds of lentil and their prices in Bihar, India. The mean, max, and minimum price seems to be the most expensive for Lentil 3. The biggest difference in minimum and maximum price have first lentil – Masur.



Top ten lentil producers – 2012

Top 10 Lentil Producing Countries (in metric tons)

Rank	**Country**	**2010**	**2011**	**2012**
1	Canada	1,947,100	1,531,900	1,493,620
2	India	1,031,600	943,800	950,000
3	Australia	140,000	379,659	463,000
4	Turkey	447,400	405,952	438,000
5	United States	392,675	214,640	240,490
6	Nepal	151,757	206,969	208,201
7	Ethiopia	80,952	128,009	151,500
8	China	125,000	150,000	145,000
9	Syria	77,328	112,470	130,229
10	Iran	100,174	71,808	85,000
	World	4,686,673	4,386,870	4,522,097

Source: UN Food & Agriculture Organization

Picture1. Top ten lentil producers – 2012 , http://www.fao.org/home/en/ [Accessed 07/03/2018]

As we can see on Picture 1 the India is second producer on lentil on the world. The poor regions of India depend mostly on that production. They produce it mostly on their own consumption in India.

Prices of lentil are different in every shop and every region and depend on many factors. Knowledge of which lentil is the most expensive in which region gives a power to redistribution of good to eliminate hunger at least in some aspects.

If we will push production or make subvention for correct plants that could help logistically put the prices down of this important food. Private companies will use that analysis to choose the most expensive lentil (of course other factors need to be considered, like for example cost of production and transport fees). But main aim of that work is to eliminate hunger not the profits. Of course assume full cooperation from Indian government.

3. Two-way ANOVA – Prices of Sorghum in Sudan

This analysis is even more important than previous one. Two-way ANOVA give is possibility to check prices considered 2 factors not only one for more than 2 groups.

This statistical test will be performed on data a data downloaded from https://data.humdata.org/dataset/wfp-food-prices it was gathered by World Food Programme (http://www1.wfp.org/). It was published 05/12/2017. We will check means of prices in few poor regions of Sudan and per Sorghum on free market and Sorghum from food aid.

Sorghum is very important and main grain in many parts of Africa. We do not know too much about that, but most people eat it in some product ingredients. In Africa it is a substitute to our wheat flour. Also you can make adhesives and paper from sorghum. In USA 30% production of sorghum is for ethanol.

'However, sorghum, a cereal grain, is the fifth most important cereal crop in the world, largely because of its natural drought tolerance and versatility as food, feed and fuel. In Africa and parts of Asia, sorghum is primarily a human food product'[4]

Below you can find analysis of prices of Sorghum in some regions of Sudan with and without food aid.

Prices are per 3 kilograms and in Sudanese pound.

Region 1 - Northern Darfur

[4] https://wholegrainscouncil.org/whole-grains-101/grain-month-calendar/sorghum-june-grain-month [Accessed 07/03/2018]

Region 2 – South Darfur

Region 3 – Southern Kodorfan

1.1 Hypotheses

Two-way ANOVA we will compare means of 3 samples of price of Sorghum in Sudan with additional factor of type of market with food aid Sorghum and free market Sorghum. 68 records have been tested.

H1: All the Sudan regions have equal price of Sorghum on the average

H2: Both the Sorghum groups (free market and food aid) have equal price on the average

Third hypothesis is also tested:

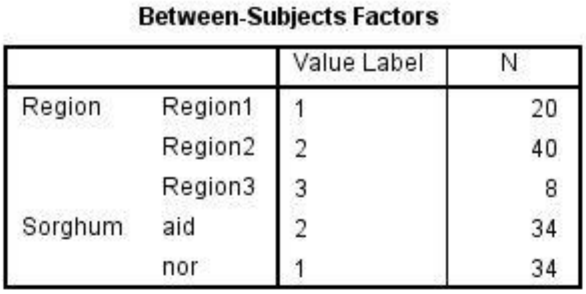
H3: The two factors (region and kind of Sorghum) are independent or that interaction effect is not present.

1.2 Two-way ANOVA

- Excel results:

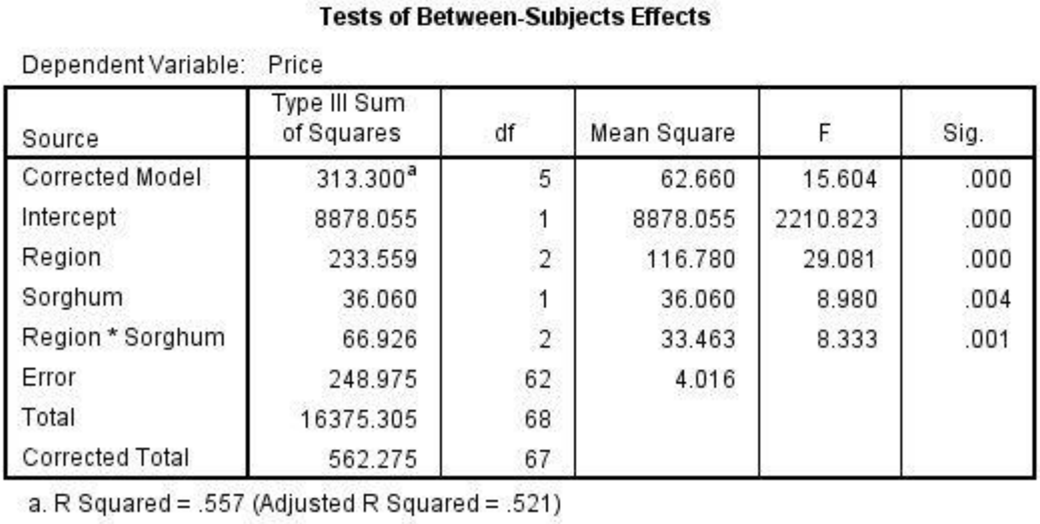
As the samples are not equal for each region, we are not able to perform test in Excel due to Excel limitations.

- SPSS results:



Between-Subjects Factors

		Value Label	N
Region	Region1	1	20
	Region2	2	40
	Region3	3	8
Sorghum	aid	2	34
	nor	1	34

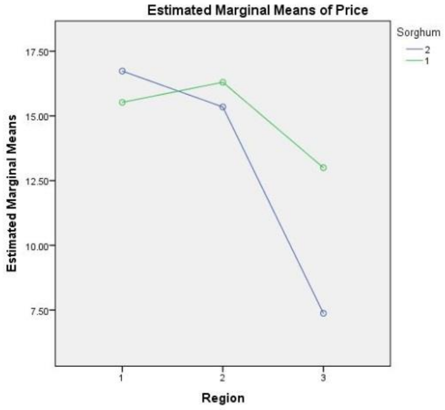


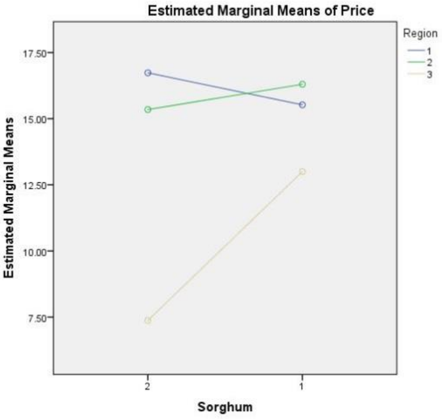
Tests of Between-Subjects Effects

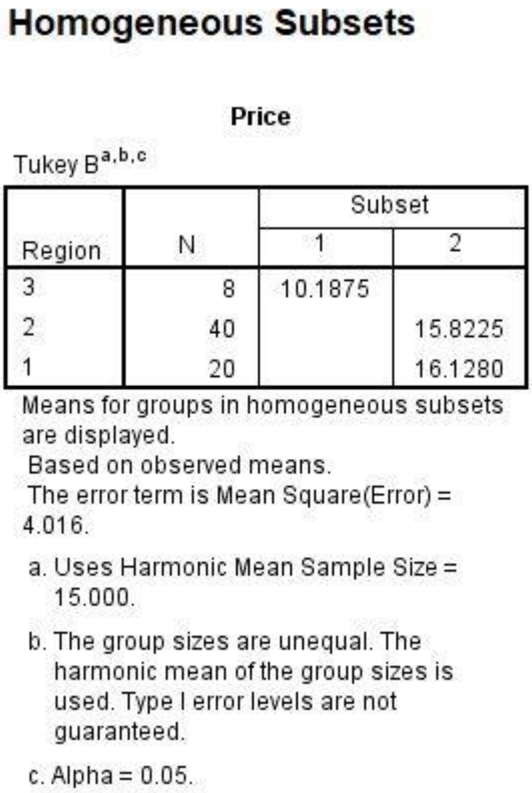
Dependent Variable: Price

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	313.300[a]	5	62.660	15.604	.000
Intercept	8878.055	1	8878.055	2210.823	.000
Region	233.559	2	116.780	29.081	.000
Sorghum	36.060	1	36.060	8.980	.004
Region * Sorghum	66.926	2	33.463	8.333	.001
Error	248.975	62	4.016		
Total	16375.305	68			
Corrected Total	562.275	67			

a. R Squared = .557 (Adjusted R Squared = .521)







Homogeneous Subsets

Price

Tukey B[a,b,c]

Region	N	Subset	
		1	2
3	8	10.1875	
2	40		15.8225
1	20		16.1280

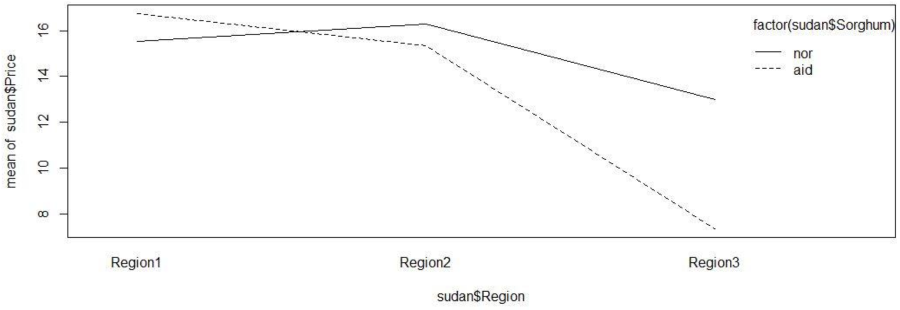
Means for groups in homogeneous subsets are displayed.
Based on observed means.
The error term is Mean Square(Error) = 4.016.

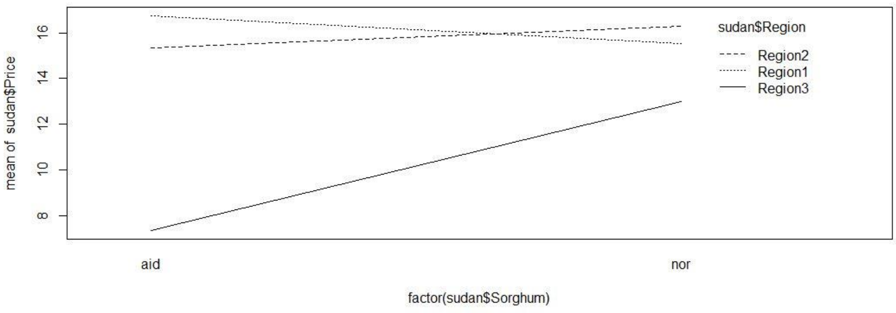
a. Uses Harmonic Mean Sample Size = 15.000.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

c. Alpha = 0.05.

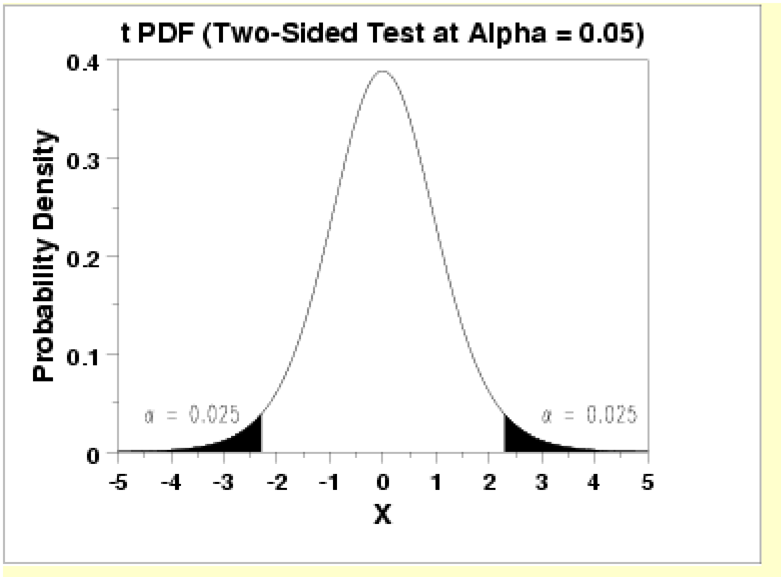
- R studio results:





1.3 Determine α value

α value has been set up at 0.05, as the prices of Sorghum are not very sensitive data and 95% assurance that results are correct and type I error are not made will be enough.



Graph3 the significance level - α[5]

1.4 Report

A two-way ANOVA was conducted to examine the effects of type of market for Sorghum and region of Sudan on price of Sorghum.

α = 0.05

Degrees of freedom:

DF_1 = 2

DF_2 = 62

INTERACTIONS:

p = 0.001 (Region*Sorghum) This is less than .05 (i.e., it satisfies p < .05), which means that there is a statistically significant interaction effect.

[5] http://www.itl.nist.gov/div898/handbook/eda/section3/eda3672.htm [Accessed 07/03/2018]

Plots in R studio and SPSS show that lines are not parallel, indicating that we might have an interaction effect. Interactions that do crossover.

F(2, 62) = 8.333, p = .001

The effect of Region of Sudan on Price of Sorghum depends of food aid.

The effect of food aid in Sudan on Price of Sorghum depends of region in Sudan.

There was a statistically significant interaction between food aid and region of Sudan for price of Sorghum.

MAIN EFFECT of Region:

F(2, 62) = 29.081, p < .05 There was a statistically significant main effect of region of Sudan.

MAIN EFFECT of type of market of Sorghum (free market or food aid):

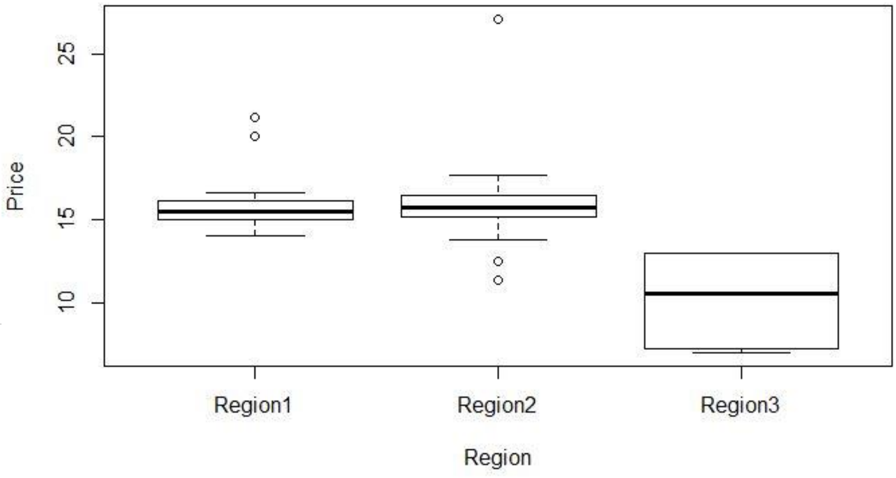
F(2, 62) = 8.98, p < .05 There was a statistically significant main effect of type of market of Sorghum.

Pairwise Comparisons:

1.5 Conclusions

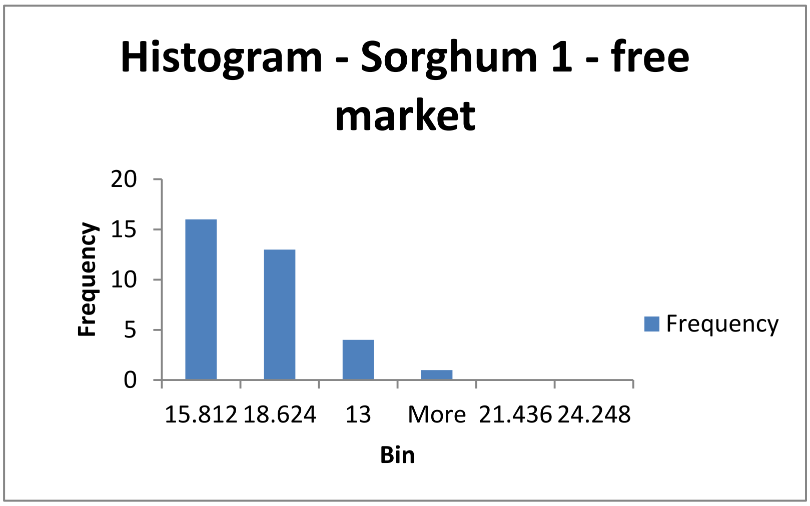
Basically we conducted test and rejected all hypothesis. Its mean that price of Sorghum in Sudan depends from Region of Sudan and from food aid too. Also both factors interact with each other and food help is different in each region and that cause difference in prices.

The Boxplot 2 present and 3 regions and we can see that Region 1 and 2 is similar but the thir one is different with mean, min and max price. Region 3 - Kodorfan is located on south part of Sudan and have borders with South Sudan. Region 1 and 2 is Darfur North and South. It is a western part of Sudan and it is free from wars only from few years. Furthermore it is still unsafe place even for humanitarian organisations. The food aid may lower the prices but still is not

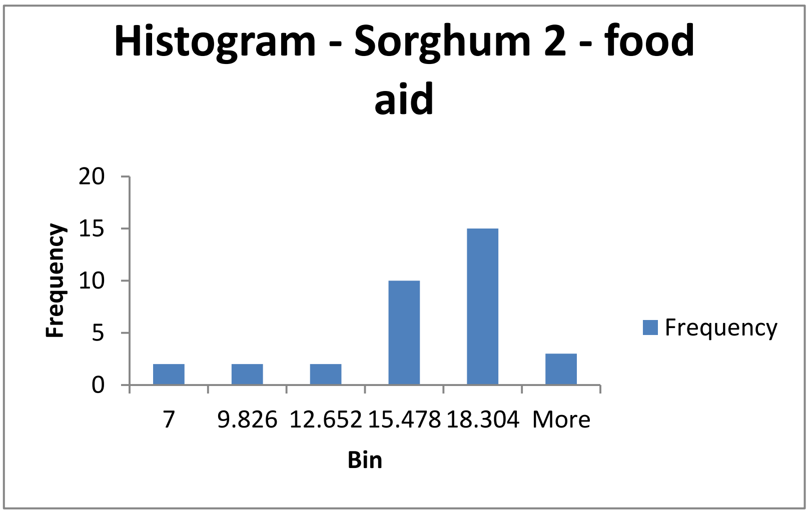


enough.

Boxplot2 – Region of Sudan – Prices of Sorghum



Histogram – Sorghum 1



Histogram – Sorghum 2

Distribution of both data with Sorghum 1 and Sorghum 2 are normal the Sorghum 1 – free market it is right – skewed and for Sorghum 2 is a left - skewed.

Bibliography

1. http://insideairbnb.com/get-the-data.html [Accessed: 01.03.2018]
2. http://www1.wfp.org/ [Accessed: 01/03/2018]
3. https://data.humdata.org/dataset/wfp-food-prices [Accessed: 01/03/2018]
4. http://www.fao.org/home/en/ [Accessed 07/03/2018]

5. https://wholegrainscouncil.org/whole-grains-101/grain-month-calendar/sorghum-june-grain-month [Accessed 07/03/2018]
6. http://www.itl.nist.gov/div898/handbook/eda/section3/eda3672.htm [Accessed 07/03/2018]

Appendices

1. Dataset with data of available apartments for Dublin and Paris.
2. Dataset with data of prices of 3 kinds of lentils in Bihar, India.
3. Dataset with data of prices of Sorghum in Sudan
4. R code file