

**National College of Ireland**

**Project Submission Sheet – 2017/2018**

**School of Computing**

|  |  |  |  |
| --- | --- | --- | --- |
| **Student Name:** | JORDI BATLLE | | |
| **Student ID:** | 17133246 | | |
| **Programme:** | HDSDAICTSEP3 | **Year:** | 2017/2018 |
| **Module:** | ADVANCED BUSINESS DATA ANALYSIS | | |
| **Lecturer:** | GIOVANI ESTRADA | | |
| **Due Date:** | 8/04/2018 | | |
| **Project Title:** | CA2: NON-PARAMETRIC TESTS | | |
|  |  | | |
| **Word Count:** | Aprox. 1,300 words | | |
|  |  | | |

**I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.**

**ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.**

|  |  |
| --- | --- |
| **Signature:** | ……………………………………………………………………………………………………………… |
| **Date:** | ……………………………………………………………………………………………………………… |

**PLEASE READ THE FOLLOWING INSTRUCTIONS:**

1. Please attach a completed copy of this sheet to each project (including multiple copies).

2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.

|  |  |
| --- | --- |
| **Office Use Only** | |
| Signature: |  |
| Date: |  |
| Penalty Applied (if applicable): |  |

NON-PARAMETRIC TESTS

Table of Contents

[1- Introduction 3](#_Toc510973779)

[2- South and East Vs. Border, Midlands and West 5](#_Toc510973780)

[2.1- Checking the Data 7](#_Toc510973781)

[2.2- Running the Statistical Test 11](#_Toc510973782)

[3- Cities Comparison 12](#_Toc510973783)

[3.1- Checking the Data 16](#_Toc510973784)

[3.2- Running the Statistical Test 28](#_Toc510973785)

[References 31](#_Toc510973786)

# Introduction

In this report, I will run a study on a survey performed on the population aged over 5 years old with relation to several aspects of their commuting journey to school, college or work place in Ireland (Maynooth University, 2011).

The file presents 18,488 records on a wide range of different categories; this study will focus only on the population who commute on foot.

The file shows as well different decompositions of the areas of study; this statistical analysis will consider only the following fields:

* County
* NUTS\_II (SE or BMW region)

Taking into account the characteristics of Ireland, were the population is concentrated in Dublin region, which makes this region completely different from the rest. Therefore, making a statistical analysis including Dublin would probably lead to false conclusions with regard to the rest of Ireland, That is the reason why I have excluded from the report the areas from Dublin region (Dublin City, Dún Laoghaire-Rathdown, Fingal and South Dublin).

I end up then with a report with 13,682 records; and checking the file we see there is a good balance between the number of observations for the 2 regions I will firstly study: South and East versus Border, Midlands and West (as per NUTS\_II column).

Later I will perform an analysis to compare the averages of 5 of the most representative cities in Ireland (Dublin excluded).

Below, an overview of the final file subject to study:

> # Structure of the file

> str(Foot\_Journey)

'data.frame': 13682 obs. of 3 variables:

$ County : Factor w/ 29 levels "Carlow","Cavan",..: 1 1 1 1 1 1 1 1 1 1 ...

$ NUTS\_II: Factor w/ 2 levels "BMW","SE": 2 2 2 2 2 2 2 2 2 2 ...

$ Foot : int 51 42 19 22 31 26 40 14 13 40 ...

> # Sample with first lines

> head(Foot\_Journey)

County NUTS\_II Foot

1 Carlow SE 51

2 Carlow SE 42

3 Carlow SE 19

4 Carlow SE 22

5 Carlow SE 31

6 Carlow SE 26

> # Summary of the file

> summary(Foot\_Journey)

County NUTS\_II Foot

Cork County :1650 BMW:6937 Min. : 0.00

Donegal : 761 SE :6745 1st Qu.: 5.00

Galway County: 741 Median : 12.00

Kildare : 731 Mean : 18.44

Kerry : 701 3rd Qu.: 26.00

Mayo : 643 Max. :281.00

(Other) :8455

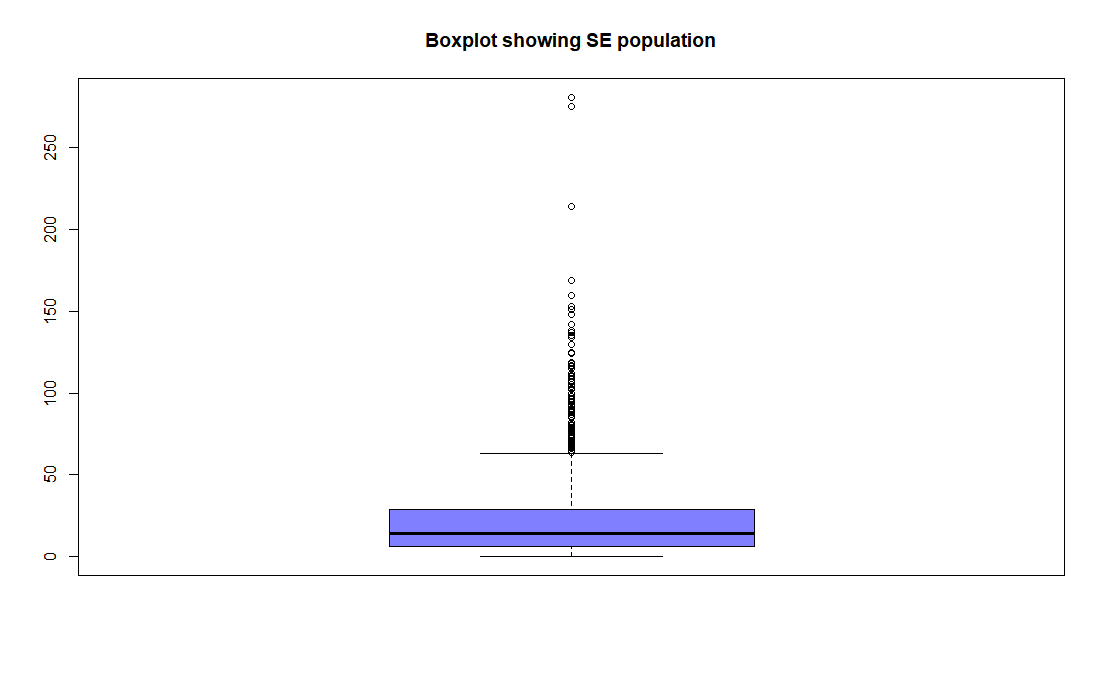
# 2- South and East Vs. Border, Midlands and West

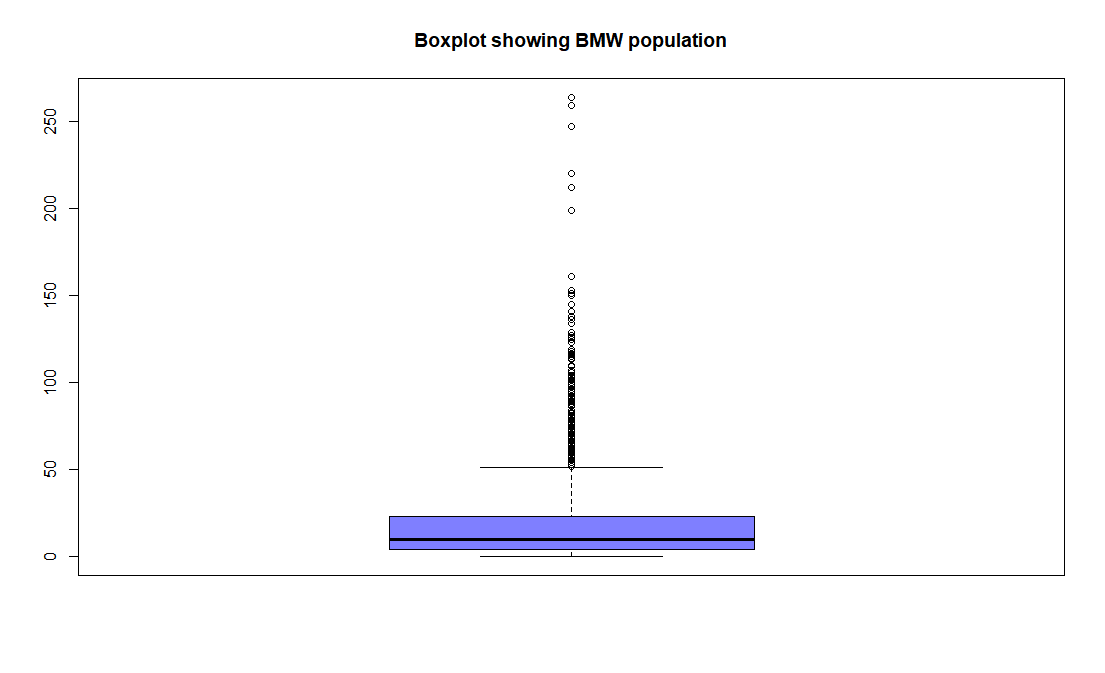
In the first analysis, I am going to compare the population commuting on foot in the South and East regions of Ireland against the population in the Border, Midlands and West regions.

Below a brief statistics summary of the 2 populations subject to study run in Excel:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***BMW REGION*** | |  | ***SE REGION*** | |
|  |  |  |  |  |
| Mean | 16.67911201 |  | Mean | 20.25455893 |
| Standard Error | 0.230632779 |  | Standard Error | 0.243264135 |
| Median | 10 |  | Median | 14 |
| Mode | 4 |  | Mode | 4 |
| Standard Deviation | 19.20909389 |  | Standard Deviation | 19.97878443 |
| Sample Variance | 368.9892883 |  | Sample Variance | 399.1518272 |
| Kurtosis | 21.63385684 |  | Kurtosis | 14.24451839 |
| Skewness | 3.293937657 |  | Skewness | 2.506885816 |
| Range | 264 |  | Range | 281 |
| Minimum | 0 |  | Minimum | 0 |
| Maximum | 264 |  | Maximum | 281 |
| Sum | 115703 |  | Sum | 136617 |
| Count | 6937 |  | Count | 6745 |

Below boxplots showing the distribution of the observations for each region run in R Studio:





# Checking the Data

Before to start, in order to determine what test needs to be run for my analysis, I check whether the following assumptions are met:

* **Assumption #1:**

The dependent variable, population commuting on foot, is continuous.

* **Assumption #2:**

The independent variable, NUTS\_II, consists of two categorical, independent groups. In this case, every observation is part of the SE region or the BMW region (and never both at the same time).

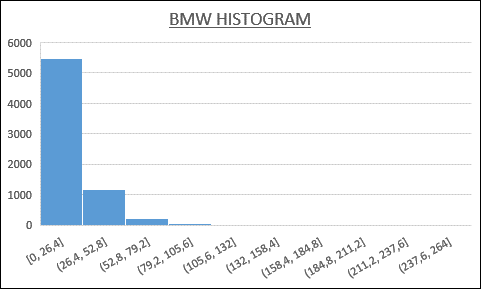
* **Assumption #3:**

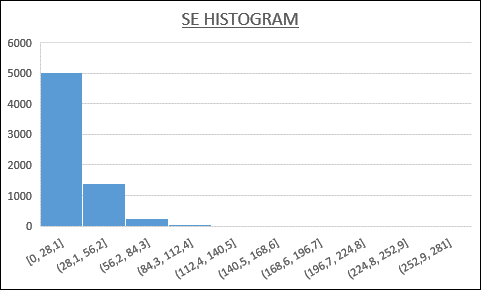
The observations are independent. There is no relationship between the observations in each category or between the categories themselves. In this case, every person is unique and therefore independent from each other.

* **Assumption #4:**

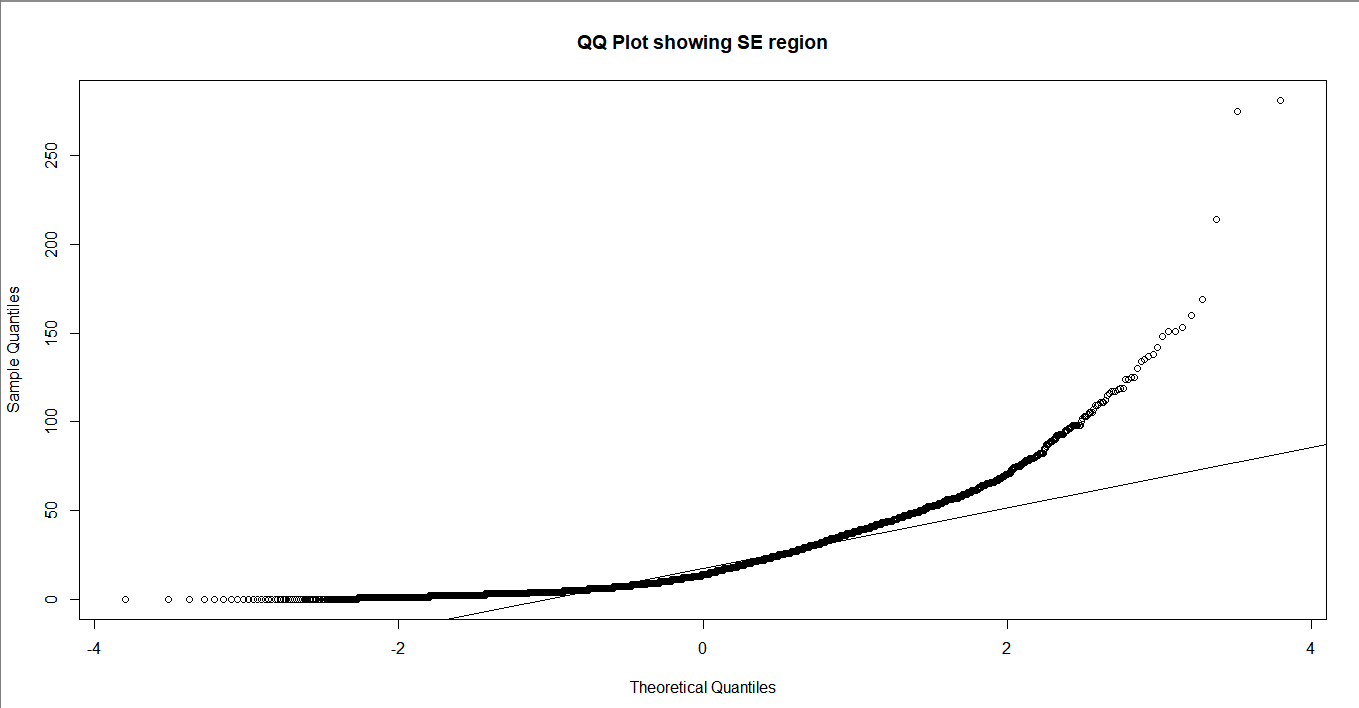
In order to decide what test is the most appropriate, I checked whether the dependent variables, population commuting on foot, are approximately normally distributed for each category of the independent variable, BMW and SE regions.

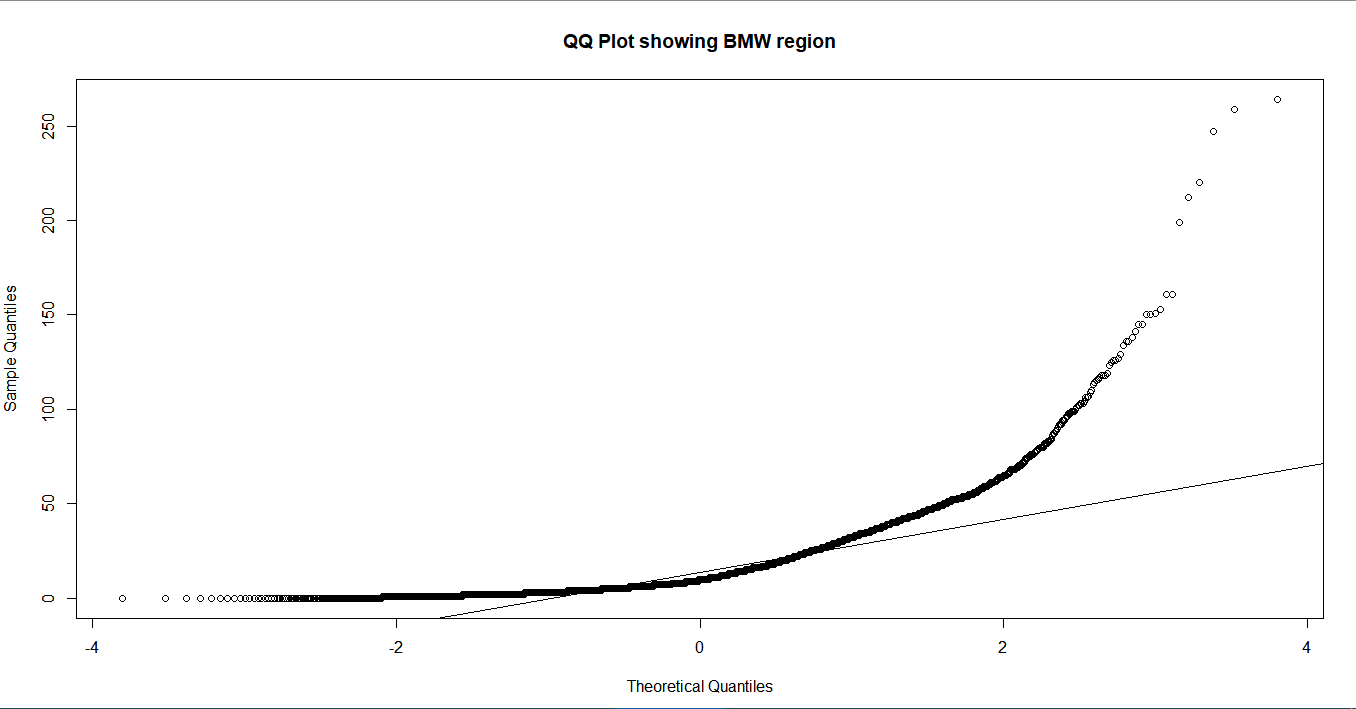
Below, Excel histograms showing the distribution of the dependent variable for each region:





Below, QQ plots displaying the values observed against normally distributed data run in R Studio:





From the previous charts, we can conclude that the data is not normally distributed; in order to confirm the results, I run the Kolmogorov-Smirnov test of normality:

> # Kolmogorov-Smirnov test for normality

> ks.test(SE$Foot, BMW$Foot)

Two-sample Kolmogorov-Smirnov test

data: SE$Foot and BMW$Foot

D = 0.1072, p-value < 2.2e-16

alternative hypothesis: two-sided

If I set the hypotheses:

H0: The observations are normally distributed

H1: The observations are not normally distributed

As we can see, the result of the test (p-value) is nearly 0.

If we set the level of significance α at 5% (the standard value), the p-value is smaller than α and, therefore, we fail to accept the null hypothesis; we cannot confirm the samples are normally distributed.

# Running the Statistical Test

As per previous checks, we can conclude that we need to run a test to compare 2 non-normally distributed independent populations. The test that that best fits is therefore the Mann-Whitney U Test.

1. **State the Null and Alternate hypotheses**

If we state the following:

SE: South and East region observations

BMW: Border, Midlands and West region observations

H0: µSE = µBTW

there is no difference between the means of commuters on foot in the SE and BMW regions.

H1: µSE ≠ µBTW

there is a difference between the means of commuters on foot in the SE and BMW regions.

1. **I set the level of significance α at 0.05 as the standard practice.**

The probability of rejecting the null hypothesis when it is true (type I error) is 5%.

1. **As per the above-mentioned data checks, I will run a Mann-Whitney U Test**
2. **Compute the test statistic value using appropriate formulas**

> # Mann-Whitney U test to compare non normal populations

> wilcox.test(SE$Foot, BMW$Foot)

Wilcoxon rank sum test with continuity correction

data: SE$Foot and BMW$Foot

W = 26741000, p-value < 2.2e-16

alternative hypothesis: true location shift is not equal to 0

1. **Conclusion**

We see that: **p < α**

We fail to accept H0. At a 95% level of confidence, we can conclude that there is a significant difference in the population means and therefore, there is a difference between the population that commutes on foot in the SE and BMW regions.

# 3- Cities Comparison

In this analysis, I am going to compare the populations commuting on foot in 5 of the biggest cities in Ireland (Dublin excluded):

* Cork
* Galway
* Kildare
* Meath
* and Limerick

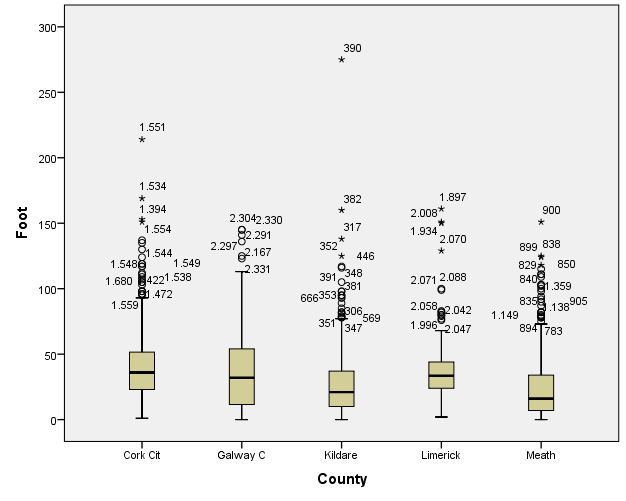
Below a brief statistics summary from SPSS of the 5 populations subject to study:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Case Processing Summary** | | | | | | | |
|  | County | Cases | | | | | |
|  | Valid | | Missing | | Total | |
|  | N | Percent | N | Percent | N | Percent |
| Foot | Cork Cit | 519 | 100,0% | 0 | 0,0% | 519 | 100,0% |
| Galway C | 307 | 100,0% | 0 | 0,0% | 307 | 100,0% |
| Kildare | 731 | 100,0% | 0 | 0,0% | 731 | 100,0% |
| Limerick | 258 | 100,0% | 0 | 0,0% | 258 | 100,0% |
| Meath | 636 | 100,0% | 0 | 0,0% | 636 | 100,0% |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Descriptives** | | | | | |
|  | County | | | Statistic | Std. Error |
| Foot | Cork Cit | Mean | | 40,74 | 1,151 |
| 95% Confidence Interval for Mean | Lower Bound | 38,48 |  |
| Upper Bound | 43,00 |  |
| 5% Trimmed Mean | | 38,15 |  |
| Median | | 36,00 |  |
| Variance | | 687,011 |  |
| Std. Deviation | | 26,211 |  |
| Minimum | | 1 |  |
| Maximum | | 214 |  |
| Range | | 213 |  |
| Interquartile Range | | 29 |  |
| Skewness | | 1,949 | ,107 |
| Kurtosis | | 6,347 | ,214 |
| Galway C | Mean | | 36,75 | 1,690 |
| 95% Confidence Interval for Mean | Lower Bound | 33,43 |  |
| Upper Bound | 40,08 |  |
| 5% Trimmed Mean | | 34,28 |  |
| Median | | 32,00 |  |
| Variance | | 876,416 |  |
| Std. Deviation | | 29,604 |  |
| Minimum | | 0 |  |
| Maximum | | 145 |  |
| Range | | 145 |  |
| Interquartile Range | | 43 |  |
| Skewness | | 1,114 | ,139 |
| Kurtosis | | 1,238 | ,277 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Kildare | Mean | | 26,41 | ,875 |
| 95% Confidence Interval for Mean | Lower Bound | 24,69 |  |
| Upper Bound | 28,13 |  |
| 5% Trimmed Mean | | 24,08 |  |
| Median | | 21,00 |  |
| Variance | | 559,499 |  |
| Std. Deviation | | 23,654 |  |
| Minimum | | 0 |  |
| Maximum | | 275 |  |
| Range | | 275 |  |
| Interquartile Range | | 27 |  |
| Skewness | | 2,827 | ,090 |
| Kurtosis | | 18,880 | ,181 |
| Limerick | Mean | | 37,11 | 1,345 |
| 95% Confidence Interval for Mean | Lower Bound | 34,46 |  |
| Upper Bound | 39,76 |  |
| 5% Trimmed Mean | | 34,99 |  |
| Median | | 33,50 |  |
| Variance | | 466,821 |  |
| Std. Deviation | | 21,606 |  |
| Minimum | | 2 |  |
| Maximum | | 161 |  |
| Range | | 159 |  |
| Interquartile Range | | 20 |  |
| Skewness | | 2,454 | ,152 |
| Kurtosis | | 10,113 | ,302 |
| Meath | Mean | | 23,82 | ,921 |
| 95% Confidence Interval for Mean | Lower Bound | 22,02 |  |
| Upper Bound | 25,63 |  |
| 5% Trimmed Mean | | 21,18 |  |
| Median | | 16,00 |  |
| Variance | | 539,619 |  |
| Std. Deviation | | 23,230 |  |
| Minimum | | 0 |  |
| Maximum | | 151 |  |
| Range | | 151 |  |
| Interquartile Range | | 27 |  |
| Skewness | | 1,803 | ,097 |
| Kurtosis | | 4,010 | ,194 |

Below a boxplot from SPSS showing the distribution of the observations for each of the cities of analysis:



# Checking the Data

Before to start, in order to determine what test needs to be run for my analysis, I check whether the following assumptions are met:

* **Assumption #1:**

The dependent variable, population commuting on foot, is continuous.

* **Assumption #2:**

The independent variable, county, consists of 5 categorical, independent groups. In this case, every observation relates to one and only one of the cities subject to analysis.

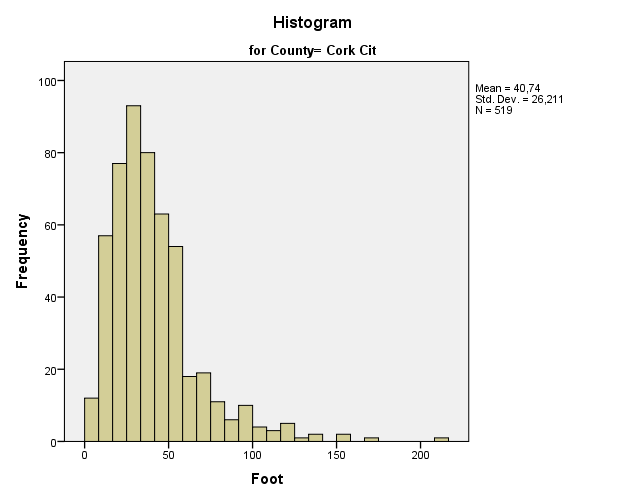
* **Assumption #3:**

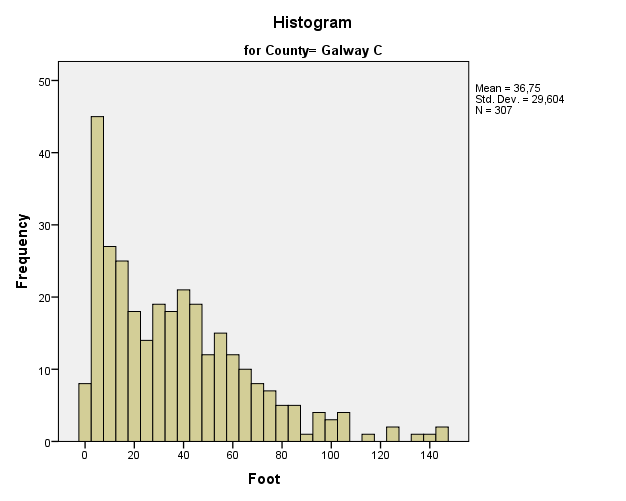
The observations are independent. There is no relationship between the observations in each category or between the categories themselves. In this case, every person is unique and therefore independent from each other.

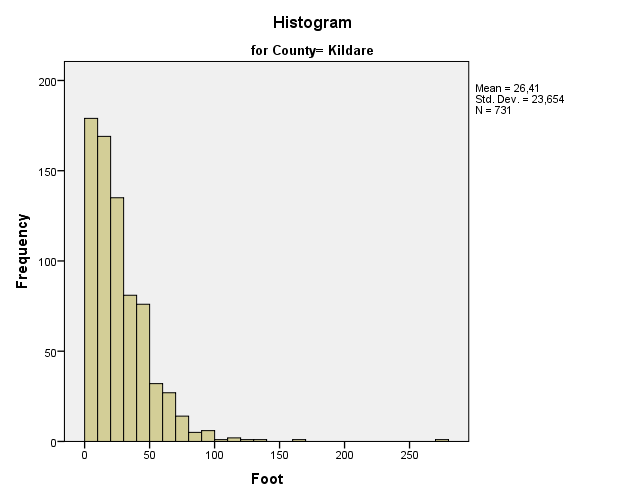
* **Assumption #4:**

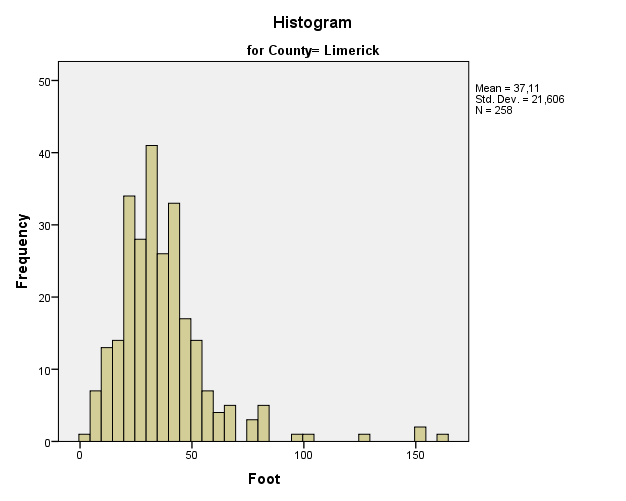
In order to decide what test is the most appropriate, I checked whether the dependent variables, population commuting on foot, are approximately normally distributed for each category of the independent variable (Cork, Galway, Kildare, Meath and Limerick.

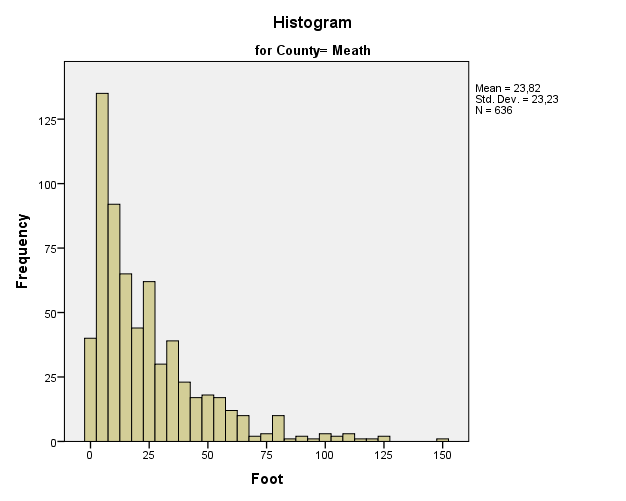
Below, histograms from SPSS showing the distribution of the dependent variable for each category:



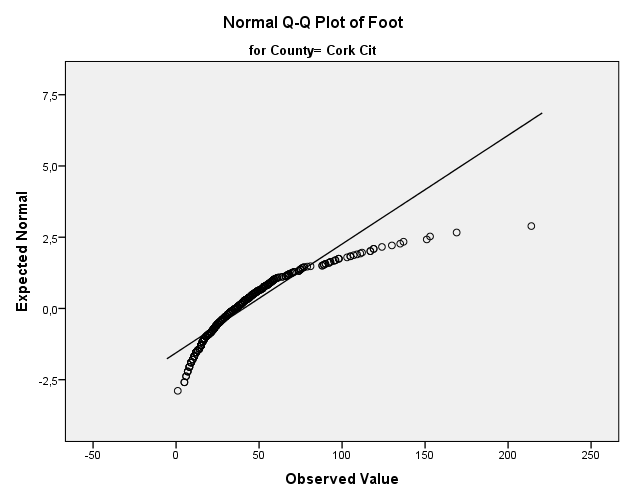


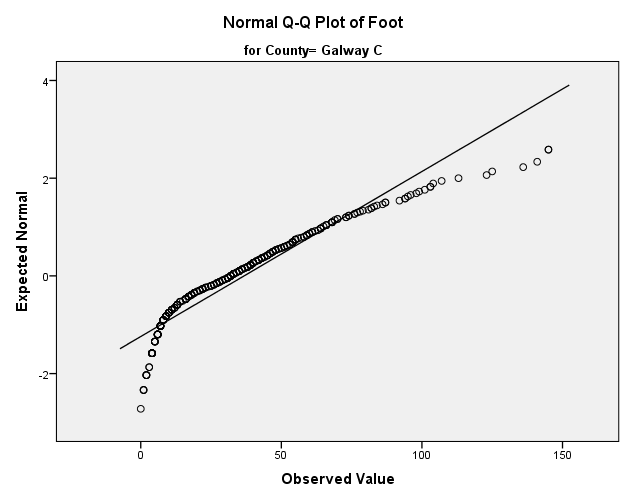


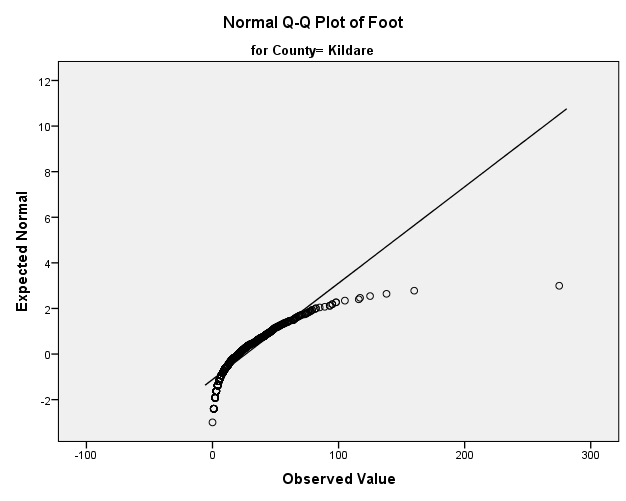


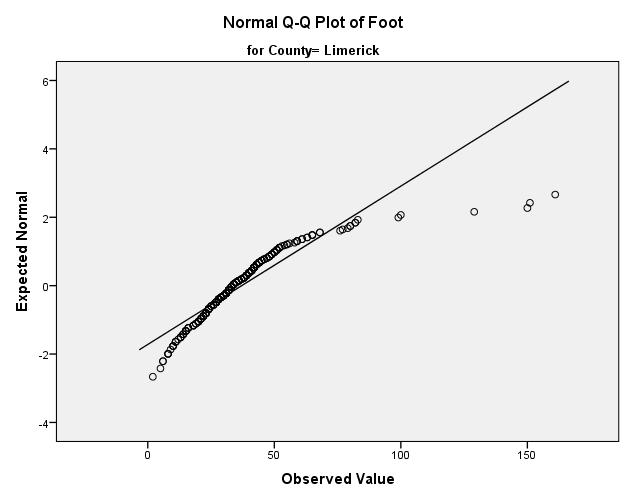


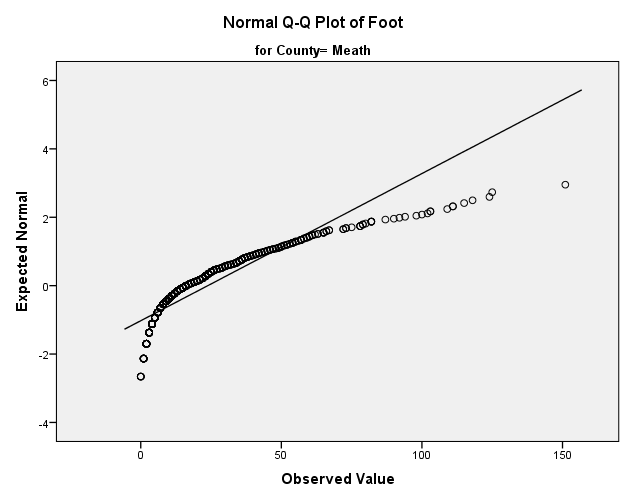
Below, QQ plots from SPSS displaying the values observed for each category against normally distributed data:











From the previous charts, we can conclude that the data is not normally distributed; in order to confirm the results, I run the Kolmogorov-Smirnov test of normality in SPSS:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Tests of Normality** | | | | | | | |
|  | County | Kolmogorov-Smirnova | | | Shapiro-Wilk | | |
|  | Statistic | df | Sig. | Statistic | df | Sig. |
| Foot | Cork Cit | ,116 | 519 | ,000 | ,855 | 519 | ,000 |
| Galway C | ,110 | 307 | ,000 | ,905 | 307 | ,000 |
| Kildare | ,140 | 731 | ,000 | ,803 | 731 | ,000 |
| Limerick | ,138 | 258 | ,000 | ,812 | 258 | ,000 |
| Meath | ,157 | 636 | ,000 | ,822 | 636 | ,000 |
| a. Lilliefors Significance Correction | | | | | | | |

If I set the hypotheses:

H0: The observations are normally distributed

H1: The observations are not normally distributed

As we can see, the result of the test (p-value/significance) are 0 (rounded at 3 decimals).

If we set the level of significance α at 5% (the standard value), the p-value is smaller than α and, therefore, we fail to accept the null hypothesis; we cannot confirm the samples are normally distributed.

# Running the Statistical Test

As per previous checks, we can conclude that we need to run a test to compare 5 non-normally distributed independent populations. The test that that best fits is therefore the Kruskal-Wallis H Test.

1. **State the Null and Alternate hypotheses**

If we state the following:

1 = Cork

2 = Galway

3 = Kildare

4 = Meath

5 = Limerick

H0: µ1 = µ2 = µ3 = µ4 = µ5

there is no difference between the means of commuters on foot in the cities checked.

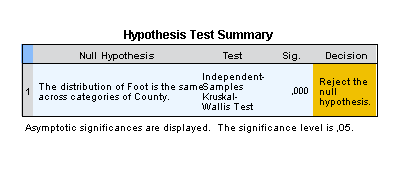
H1: µ1 ≠ µ2 ≠ µ3 ≠ µ4 ≠ µ5

there is a difference between the means of commuters on foot in the cities checked.

1. **I set the level of significance α at 0.05 as the standard practice.**

The probability of rejecting the null hypothesis when it is true (type I error) is 5%.

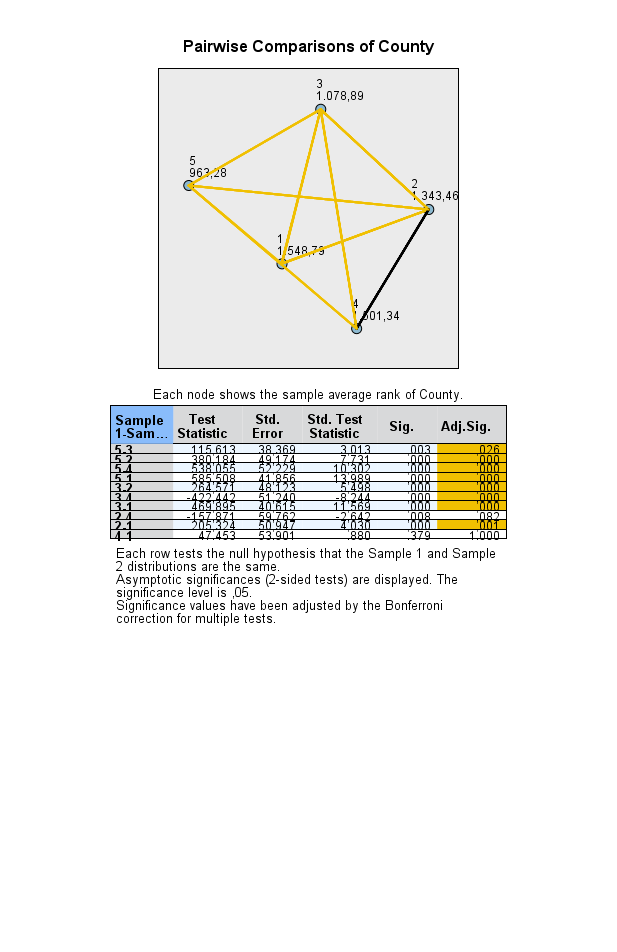
1. **As per the above-mentioned data checks, I will run a Kruskal-Wallis H Test**
2. **Compute the test statistic value using appropriate formulas**



1. **Conclusion**

We see that: **p < α**

We fail to accept H0. At a 95% level of confidence, we can conclude that there is a significant difference between the population means and therefore, there is a difference between the population that commutes on foot for each of the cities.



Running post hoc test, we see that p < α for all group pairs except one, Galway-Meath.

We can conclude that we only fail to reject to null hypothesis for Galway and Meath populations; while for the other populations there is a significant difference between themselves.

# References

LAERD STATISTICS, 2013. *Kruskal-Wallis H Test using SPSS Statistics.* [Online]   
Available at: https://statistics.laerd.com/spss-tutorials/kruskal-wallis-h-test-using-spss-statistics.php  
[Accessed April 2018].

LAERD STATISTICS, 2013. *Testing for Normality using SPSS Statistics.* [Online]   
Available at: https://statistics.laerd.com/spss-tutorials/testing-for-normality-using-spss-statistics.php  
[Accessed April 2018].

LAERD STATISTICS, n.d. *Mann-Whitney U test in SPSS.* [Online]   
Available at: https://statistics.laerd.com/premium-sample/mwut/mann-whitney-test-in-spss.php  
[Accessed April 2018].

Maynooth University, 2011. *theme\_11\_small\_areas.* [Online]   
Available at: http://airo.maynoothuniversity.ie/files/dDATASTORE/small\_areas/theme\_11\_small\_areas.csv  
[Accessed April 2018].

Salkind, N. J., 2017. *Statistics for People Who (Think They) Hate Statistics: Microsoft Excel 2016.* 4th ed. Los Angeles: Sage Publications.