Teresa Ventaja

Higher diploma in science in data analytics |

Advanced Statistics

Non-parametric tests

A close up of text on a white background

Description automatically generated

[[1]](#footnote-2)

INDEX

[0. Introduction 2](#_Toc44765260)

[1. Mann-Whitney U Test – manually made in *Excel* 2](#_Toc44765261)

[2. Wilcoxon Signed Rank test - calculations in R 6](#_Toc44765262)

[3. Kruskal-Wallis H test – calculations in *SPSS* 10](#_Toc44765263)

[Bibliography 16](#_Toc44765264)

# 0. Introduction

First, before I select the tests I was going to do, I prepared the data, so I had it accessible and ready to work with it. I downloaded the 2 files[[2]](#footnote-3)[[3]](#footnote-4) and combined them with R code:

A screenshot of a cell phone

Description automatically generated

The regions in the tables are divided into 31 administrative counties and cities and 3409 Electoral Divisions (EDs). Legally there are 3440 EDs, but 32 of them have low population, so for privacy reasons they have been included into neighbouring EDs[[4]](#footnote-5) 3125. The document is divided in 15 themes, with a variety of tables by theme (see Glossary[[5]](#footnote-6)).

To perform the calculations in R for Section 2 “Wilcoxon Signed Rank test – calculations in R”, I used packages “pastecs”, “psych” and CircStats".

# 1. Mann-Whitney U Test – manually made in *Excel*

The first test I decided to do is a Mann-Whitney U Test to compare the means of transport used in a rural and an urban area. I picked two locations in Co. Louth by electoral areas, Dundalk Urban 4 and Stabannan. The reason why I selected those locations is because I suspected the data was not normally distributed, and I also suspected that the statistical test will find significant differences between the two areas. The original dataset contains multiple rows for each small area, and Theme 11 Table 1 (see glossary screenshot below) holds the data about methods of transport:



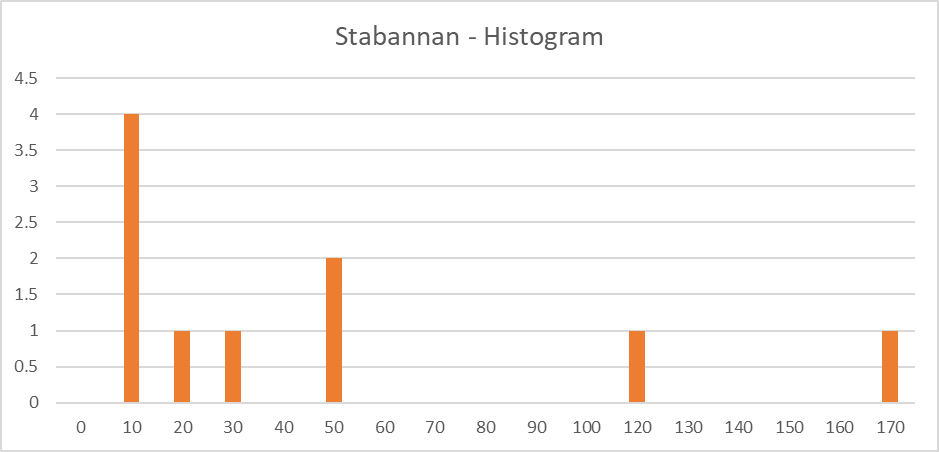
To do this test, I am following the assumptions described in the article of *Laerd Statistics* (2018). Since we can order the data hierarchically, we are going to **assume that we have ordinal data**, and therefore the first requirement for the selection of this test is met. I have not used “Total” and “Not stated – Total” data because it does not fit well into a hierarchy, and I tweaked the predefined order to make it more reasonable, moving “Work mainly at or from home – Total” to the top row. The pre-assumed order is No transport –> Individual non-motorised -> Collective -> Individual motorised -> Private motorised (other)[[6]](#footnote-7). We need 2 groups, and the original dataset contains multiple rows for area (2 for Stabannan and 18 for Dundalk Urban 4). To break this down into 2 groups, I simply summed the values up, ending up with 2 samples of 10 elements each, as follows:



I calculated descriptive statistics to understand better the data:



The distribution of the data is leptokurtic and positively skewed, more noticeably in Stabannan distribution than in Dundalk. Variances differ significantly and so do the median and mean values, but this is expected considering the different sums of totals. However, they are both still large numbers relative to the population size. The standard error suggests that Stabannan sample is more likely to represent more accurately the population than Dundalk sample. The following graphics will give a more visual sense of what the distribution look like:



We can see that the data does not seem to be normally distributed. Specially in the line chart we have an idea of the shape of the normal curve. Next step is checking if the data is normally distributed with more accurate resources, i.e. a normality test. In this case, I am using a Shapiro-Wilk test[[7]](#footnote-8):

A screenshot of a social media post

Description automatically generated

The test result confirms our **second assumption that the data is not normally distributed**, which justifies the use of this non-parametric test. Therefore, I will perform a Mann-Whitney U test as follows:

* H0 = The distribution of the means of transport used is the same across locations
* H1 = The distribution of the means of transport used is not the same across locations
* Level of significance (Alpha) = 0.05

A screenshot of a computer

Description automatically generated

We fail to reject the null hypothesis, so we have no evidence to suggest that the distribution of the means of transport used is not the same across locations.

# 2. Wilcoxon Signed Rank test - calculations in R

Now I am going to perform a Wilcoxon-Rank Test to compare the age of the population in 2011[[8]](#footnote-9) and in 2016 in small area “Dundalk Urban 4”. The format of the data collection made me think that the data was not normally distributed. The original dataset contains multiple rows for each small area, and Theme 1 Table 1 (see glossary screenshot below) holds the data about age:



In this case, I will check assumptions based on the article of *Statistics Solutions* (2020). I have interval data, so the first **assumption** (data measured at least at an interval scale) is met. I need 2 pair groups, and the original dataset contains 18 rows for Dundalk Urban 4, so I summed the values up, having 2 samples of 34 rows each. The population match and the data are paired (measured for 2011 and 2016 years, respectively). Lastly, each pair sample is random and independent. The data sample looks as follows:



I calculated descriptive statistics in R to understand better the data, using “pastecs” and “psych” packages and stat.desc() and describe() functions:

2011

nbr.val 3.400000e+01 34

nbr.null 0.000000e+00 0

nbr.na 0.000000e+00 0

min 3.700000e+01 37

max 3.960000e+02 396

range 3.590000e+02 359

sum 4.199000e+03 4199

median 5.700000e+01 57

mean 1.235000e+02 123.5

SE.mean 1.932143e+01 19.32

CI.mean.0.95 3.930975e+01 39.31

var 1.269280e+04 12692.8

std.dev 1.126623e+02 112.66

coef.var 9.122457e-01 0.912

skew 1.03

kurtosis -0.6

2016

nbr.val 3.400000e+01 34

nbr.null 0.000000e+00 0

nbr.na 0.000000e+00 0

min 2.400000e+01 24

max 3.770000e+02 377

range 3.530000e+02 353

sum 4.197000e+03 4197

median 5.600000e+01 56

mean 1.234412e+02 123.44

SE.mean 1.919783e+01 19.2

CI.mean.0.95 3.905827e+01 39.06

var 1.253092e+04 12530.92

std.dev 1.119416e+02 111.94

coef.var 9.068416e-01 0.907

skew 0.9

kurtosis -0.93

The distribution of the data is slightly platykurtic and positively skewed. Variances are similar and so are the median and mean values. The standard deviation and standard error values are also similar and large enough to suggest that we may have non-normally distributed data, but range and sum on both samples are quite homogeneous, indicating similarity between the two sets of data.

Next step is creating some visuals, first a P-P Plot using library "CircStats" and function pp.plot(), plotting the expected cumulative distribution (Y axe) against observed cumulative distribution (X axe).

**P-P Plot for 2011 sample:**

A screenshot of a cell phone

Description automatically generated

**P-P Plot for 2016 sample:**

A screenshot of a cell phone

Description automatically generated

We can see that none of them fit completely the expected cumulative distribution, but the 2016 data seems to be a bit closer to it. Now I am going to create 2 box plots to get another visual reference of the distribution of the samples.

**Whisker Plots:**

A screenshot of a cell phone

Description automatically generated

These plots clearly present a heavily asymmetric distribution. The distribution between the 1st quartile and the median is clustered, whereas in the region median-3rd quartile the data is severely disperse. The difference is even more pronounced in the 2016 data sample. All this clearly indicates a non-parametric distribution.

However, I am going to fully corroborate whether or not the data is normally distributed by performing a Kolmogorov-Smirnov test in R, using the ks.test() function for each sample separately. The result is 2.2 \*10e-16. Since the Alpha value in this test is equal to 0.001 and my p-value is smaller in both samples, I can reject the null hypothesis in favour of the alternative hypothesis, concluding that the data is not normally distributed:

**Kolmogorov-Smirnov test**

data: D2011

D = 1, p-value < 2.2e-16

alternative hypothesis: two-sided

data: D2016

D = 1, p-value < 2.2e-16

alternative hypothesis: two-sided

I reject the H0 in favor of the alternative hypothesis, concluding that **the data is not normally distributed**

Lastly, I will perform a Wilcoxon Signed Rank test in R as follows, using wilcox.test(x, y, paired = TRUE) formula:

* H0: M2011 = M2016 - The median difference between pairs of observations is zero
* H1: M2011 ≠ M2016 - The median difference between pairs of observations is not zero
* Level of significance (Alpha) = 0.05

**Wilcoxon signed rank test with continuity correction**

data: D2011 and D2016

V = 271, p-value = 0.8722

alternative hypothesis: true location shift is not equal to 0

data: D2016 and D2011

V = 290, p-value = 0.8722

alternative hypothesis: true location shift is not equal to 0

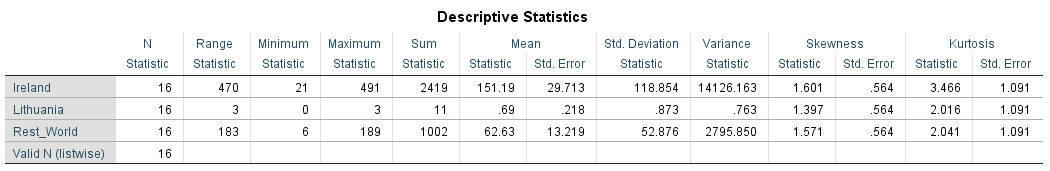
As a result, the Wstat is 271 and the p-value is 0.87. Since the p-value is greater than 0.05, I fail to reject the null hypothesis, so I have no evidence suggesting that the median difference between pairs of observations is not zero. In other words, the evidence suggests that there is no median difference between pairs.

# 3. Kruskal-Wallis H test – calculations in *SPSS*

Lastly, I am going to elaborate a Kruskal-Wallis H test to compare the birthplace of usually resident population in 2 Dublin areas (North City and Inns Quay C): those born in Ireland, in Lithuania and in the “Rest of the world” (Excluding the other EU28). The observation of values collected makes me think that the data is not normally distributed. The original dataset contains multiple rows for each small area (16 rows in total), and Theme 2 Table 1 (see glossary screenshot below) hold the data about usually resident population by place of birth:



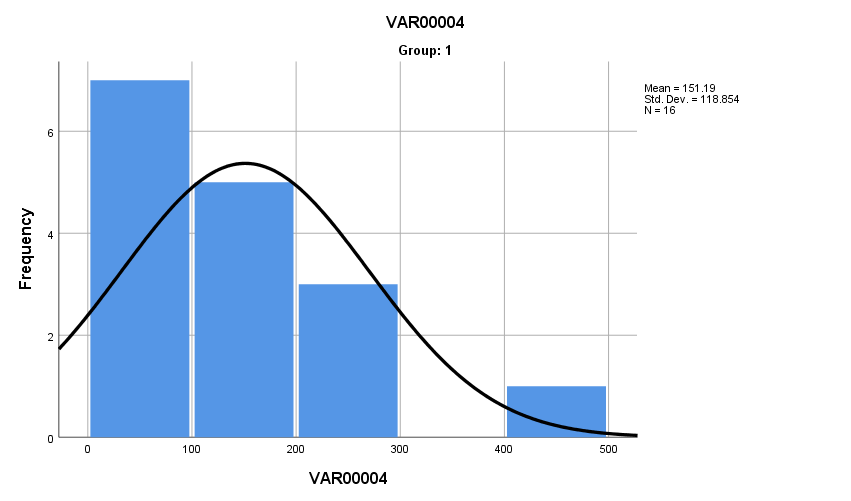
First, I need to check the requirements for using this test, and for that I will follow the article of *Statistics How To* (2016). I have 3 independent variables of equal sample size (n = 16), so the minimum number of groups (3 or more) appropriate for the selection of the Kruskal-Wallis H test is met. The dependent variable is at ordinal level (number of people), so this assumption is also met. I need to corroborate the assumption that the 3 samples are not normally distributed. As previously, I calculate descriptive statistics to have a better understanding of the data. This time I perform the calculations in SPSS:



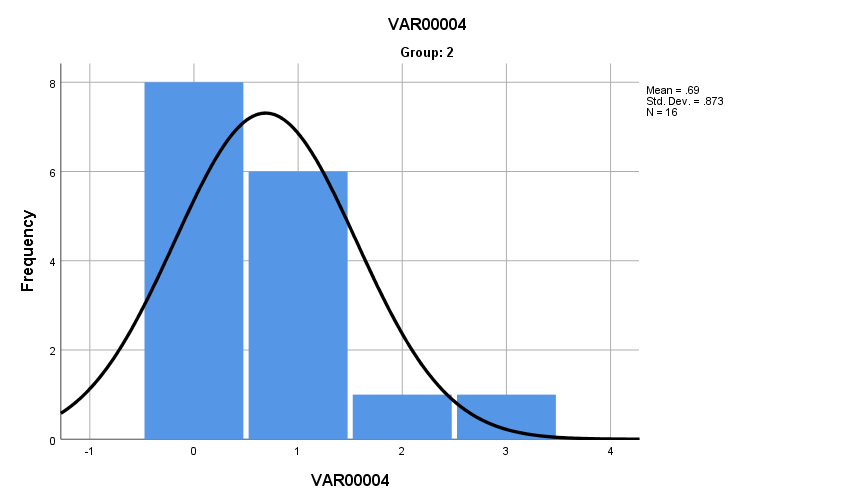
The distribution curve is leptokurtic and positively skewed. Variances differ significantly and so do the median and mean values. The standard deviation and standard error values are also different and suggest that the data is disperse. These values are large enough to make me suspect that we have non-parametric data, except for Lithuania sample, whose range is only 3, and so the distribution is more homogeneous. Nevertheless, for Lithuania 11 values and many 0 might also result in a non-normal distribution. Clearly, descriptive statistic indicate that we have 3 sets of data that are unlikely to be normal.

Next step I am going to visually represent the distribution of the data. To do so, and to perform the rest of calculations for this test, I have converted the 3 location groups into ordinal numerical values (1 for Ireland, 2 for Lithuania and 3 for “Rest of the World”):

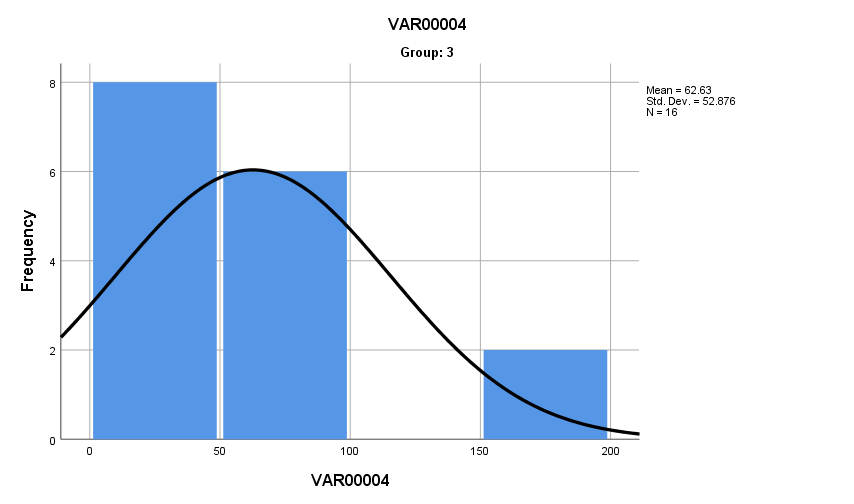
**Ireland Histogram**



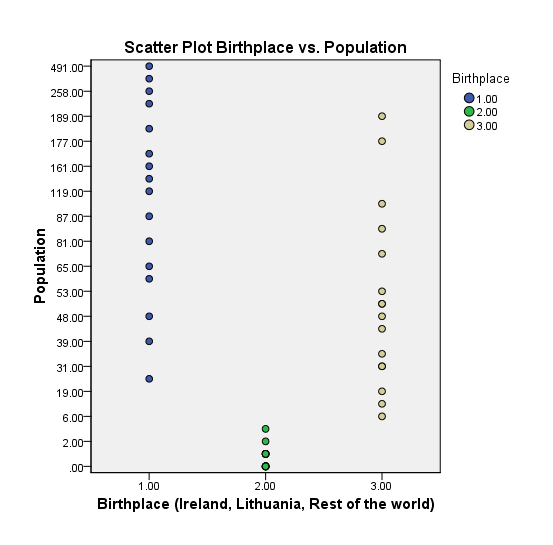
**Lithuania Histogram**



**Rest of the World Histogram**



As an alternative graphical representation, I have created a Scatter Plot in SPSS with colour labels. In this case, instead of having separate graphs, we can compare the distributions of the 3 groups just by looking at the plot:

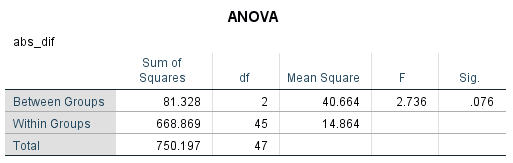


For the following calculations, I have been following *how2stats* (2011) channel in Youtube “Kruskal-Wallis - SPSS (part 1)” to part 5[[9]](#footnote-10)[[10]](#footnote-11)[[11]](#footnote-12)[[12]](#footnote-13)[[13]](#footnote-14). This tutorial references Keselman, Games, & Rogan (1979) and Green, S. B. & Salkind, N. J. (2005). The reason why I decided to follow this tutorial specifically is because of the relevance of the sources mentioned, and the completeness of the explanations.

Apart from the non-normal data, this test assumes that we have similar distributions of the data (assumption of homogeneity). Kurtosis and skewness are similar, which suggest we might meet this assumption, but I am going to test homogeneity to make sure. To do so, I am going to rank the data, calculate the rank mean for each group (Ireland, Lithuania and Rest of the world) and then calculate absolute difference between the rank mean and the dependent variable (count of people). On the result, I will perform an ANOVA test. For demonstration purposes, the following screenshots show part of the calculations and the result of ANOVA test:

A screenshot of a cell phone

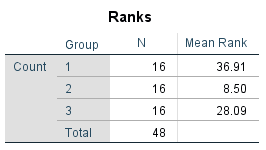
Description automatically generated

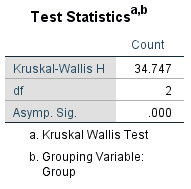


We can see that the significance value is 0.07, which is greater than 0.05. Therefore, I fail to reject the null hypothesis, so **homogeneity can be assumed** as a result of this test. Now I can go ahead and perform a Kruskal-Wallis H test in SPSS as follows:

* H0: MRIreland = MRLithuania = MRRest\_world
  + The mean rank difference between groups (birthplace) is zero
* H1: The mean rank difference between groups (birthplace) is not zero
* Level of significance (Alpha) = 0.05

**Kruskal-Wallis Test**





As a result, our Hstat is 34.75 and our Hcrit for a sample size of 16 is 5.91, but in addition to it, the P-value is smaller than 0.001. Since this P-value is less than 0.05 and my Hstat is greater than Hcrit, I can reject the null hypothesis of no differences between mean ranks for count of people residing in Ireland by birthplace. That means that **at least** there is a different between 2 groups (2 birthplaces). Having found that different, we can calculate an effect size estimate: Chi-Square value divided by n-1 (47). The result is 73.9, meaning that 73.9% of the variability in rank scores is accounted for by locations. More than 14% is considered a **large effect size**.

I am going to finalize this report by performing **post-hoc tests** to get more details into the different between specific groups.

A screenshot of a cell phone

Description automatically generated

Statistically significant P-value found for the difference between those born in Ireland and those born in Lithuania.

A screenshot of a cell phone

Description automatically generated

Statistically significant P-value found for the difference between those born in Lithuania and those born in the “Rest of the wold”.

A screenshot of a cell phone

Description automatically generated

Statistically significant P-value found for the difference between those born in Ireland and those born in the “Rest of the World”. In conclusion, there is a significant difference between the 3 birthplace groups pairwise, between the 3 of them.

# Bibliography

1. Central Statistics Office (2019):

* “*Census 2011 Small Area Population Statistics (SAPS)”* [Online]. Available at: <https://www.cso.ie/en/census/census2011smallareapopulationstatisticssaps/> (Accessed 4/7/2020)
* “*Census 2016 Boundary Files*” [Online]. Available at: <https://www.cso.ie/en/census/census2016reports/census2016boundaryfiles/> (Accessed 4/7/2020)

“*Census 2016 Small Area Population Statistics*” [Online]. Available at: <https://www.cso.ie/en/census/census2016reports/census2016smallareapopulationstatistics/>

“*Saps Tables and Shapefiles*” [Online]. Available at: <https://www.cso.ie/en/media/csoie/census/census2016/census2016boundaryfiles/SAPS_2016_Glossary.xlsx> (Accessed 4/7/2020)

“*SAPS 2016 Glossary*” [Online]. Available at: <https://www.cso.ie/en/media/csoie/census/census2016/census2016boundaryfiles/SAPS_2016_Glossary.xlsx> (Accessed 6/7/2020)

1. *how2stats* – Youtube (2011):

* “*Kruskal-Wallis - SPSS (part 1)”* [Online]. Available at: <https://www.youtube.com/watch?v=Md8rqQ-oUH0> (Accessed 6/7/2020)
* “*Kruskal-Wallis - SPSS (part 2)”* [Online]. Available at: https://www.youtube.com/watch?v=oWARSC31Rno (Accessed 6/7/2020)
* “*Kruskal-Wallis - SPSS (part 3)”* [Online]. Available at: https://www.youtube.com/watch?v=xLAUs4hL5Vs (Accessed 6/7/2020)
* “*Kruskal-Wallis - SPSS (part 4)”* [Online]. Available at: https://www.youtube.com/watch?v=Sloy2lbtPVc (Accessed 6/7/2020)
* “*Kruskal-Wallis - SPSS (part 5)”* [Online]. Available at: <https://www.youtube.com/watch?v=5UdaYXPeUYY> (Accessed 6/7/2020)
  + Keselman, Games, & Rogan (1979): “*Protecting the overall rate of Type I errors for pairwise comparisons with an omnibus test statistic.”* *Psychological Bulletin*, 86(4), 884-888, cited in *how2stats* – Youtube (2011)
  + Green, S. B. & Salkind, N. J. (2005): *“Using SPSS for Windows and Macintosh: Analyzing and understanding data (fourth edition)”.* New Jersey: *Pearson,* cited in *how2stats* – Youtube (2011)

Laerd Statistics (2018): “*Assumptions of the Mann-Whitney U test*” [Online]. Available at: <https://statistics.laerd.com/statistical-guides/mann-whitney-u-test-assumptions.php> (Accessed 4/7/2020)

1. Laerd Statistics (2018): “*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 4/7/2020)
2. Louth County Council (2019): “*Local Electoral Areas, Electoral Divisions & Municipal Districts Maps”* [Online]. Available at: <https://www.louthcoco.ie/en/services/electoral-register/electoral-area-maps-polling-scheme/maps-of-local-electoral-areas/> (Accessed 4/7/2020)
3. Ordnance Survey Ireland (2017): “*Small Areas Generalised 100m - OSi National Statistical Boundaries – 2015*” [Online]. Available at: <https://data.gov.ie/dataset/small-areas-generalised-100m-osi-national-statistical-boundaries-2015> (Accessed 4/7/2020)

Salkind, N. (2014): “Statistics for People Who (Think They) Hate Statistics”. *Sage Publications*. ISBN: 978-1-4522-7771-4.

1. Statistics How To (2016): “*What is the Kruskal Wallis Test?”* [Online]. Available at: <https://www.statisticshowto.com/kruskal-wallis#:~:text=Assumptions%20for%20the%20Kruskal%20Wallis,have%20three%20or%20more%20levels.> (Accessed 4/7/2020)
2. Statistics Solutions (2020): “*Assumptions of the Wilcoxon Sign Test*” [Online]. Available at:: <https://www.statisticssolutions.com/assumptions-of-the-wilcox-sign-test/#:~:text=The%20Wilcoxon%20Sign%20Test%20requires,both%20observations%20can%20be%20compared.&text=Independence%20%E2%80%93%20The%20Wilcoxon%20sign%20test,are%20randomly%20and%20independently%20drawn.> (Accessed 4/7/2020)

1. Image source: <https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Nonparametric/> [↑](#footnote-ref-2)
2. Data source: <https://www.cso.ie/en/census/census2016reports/census2016smallareapopulationstatistics/> [↑](#footnote-ref-3)
3. Data source: <https://data.gov.ie/dataset/small-areas-generalised-100m-osi-national-statistical-boundaries-2015> [↑](#footnote-ref-4)
4. <https://www.cso.ie/en/census/census2016reports/census2016boundaryfiles/> [↑](#footnote-ref-5)
5. https://www.cso.ie/en/media/csoie/census/census2016/census2016boundaryfiles/SAPS\_2016\_Glossary.xlsx [↑](#footnote-ref-6)
6. Note that the last 2 categories include unusual means of transport, such as vans or lorries, but the order still makes sense, because the most commonly thought to be usual methods of transport have a higher hierarchy. [↑](#footnote-ref-7)
7. I used the Real Statistics extension in Excel to perform the calculation (=SWTEST() function). [↑](#footnote-ref-8)
8. Data source: <https://www.cso.ie/en/census/census2011smallareapopulationstatisticssaps/> [↑](#footnote-ref-9)
9. Video source (part 1): <https://www.youtube.com/watch?v=Md8rqQ-oUH0> [↑](#footnote-ref-10)
10. Video source (part 2): <https://www.youtube.com/watch?v=oWARSC31Rno> [↑](#footnote-ref-11)
11. Video source (part 3): <https://www.youtube.com/watch?v=xLAUs4hL5Vs> [↑](#footnote-ref-12)
12. Video source (part 4): <https://www.youtube.com/watch?v=Sloy2lbtPVc> [↑](#footnote-ref-13)
13. Video source (part 5): <https://www.youtube.com/watch?v=5UdaYXPeUYY> [↑](#footnote-ref-14)