Advanced Business Data Analysis CA

Carl Flynn

x17347726

BSHCDA4

29/03/2021

**Introduction**

The report will be analysing a number of data samples taken from the Small Area Population Statistics 2016 and the Small Areas OSI Boundaries datasets provided by the Central Statistics Office and Ordinance Survey Ireland respectfully. The datasets provided give data on various areas of society in Ireland and their distribution throughout the small areas of Ireland. In the following report we will compare and run tests on five separate areas from the combined datasets and report on the results obtained from these statistical tests. The statistical tests used on the data will be non-parametric tests including: Mann-Whitney U Test, Kruskal-Wallis H Test and Wilcoxon Signed Ranks Test. Visualisations will include histograms, boxplots and frequency plots. All samples will also have descriptive statistics provided.

Both datasets were combined into one larger dataset using R Studio before testing began with this larger dataset being broken into 5 smaller datasets as csv files so they could be easily read by all software used for running tests and visuals. All of the statistical tests and data visualisations were completed using R Studio, MS Excel and SPSS.

The sample data that will be tested is as follows:

* Professional Working Male vs. Female in Ireland (Wilcoxon Signed Ranks Test).
* Skilled vs. Semi-skilled vs. unskilled males in areas throughout Ireland(Kruskal-Wallis H Test).
* No formal vs. Primary vs. Lower Secondary vs. Upper Secondary education across both genders throughout areas in Ireland(Kruskal-Wallis H Test).
* All Others Gainfully Occupied and Unknown in Dublin, Male vs. Female (Wilcoxon Signed Ranks Test).
* Managerial and Technical Employment in Ireland, Male vs. Female (Mann-Whitney U Test)

**Test 1:** Professional Working Males vs. Females in Small Areas Throughout Ireland

Null Hypothesis (*H0*)

*The number of professional working individuals in an area is the same across each gender*

Alternate Hypothesis (*H1*)

*The number of professional working individuals in an area is not the same across each gender*

Significance Level (*α*)

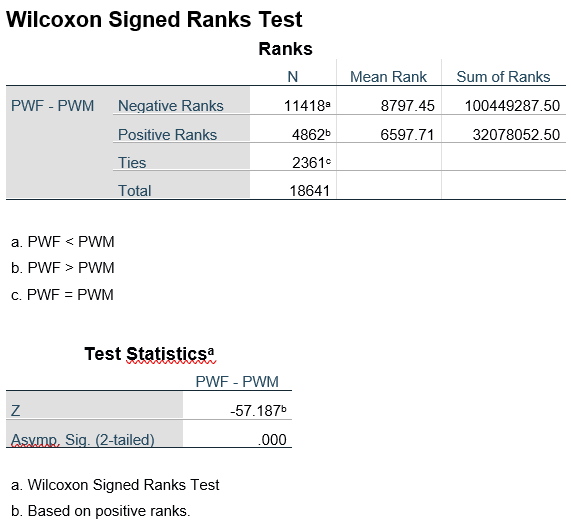
*0.05*

This test was conducted on professional working males vs. professional working females in the same area throughout Ireland taken from the Small Area Population Statistics dataset provided by the Central Statistics Office. The test performed on the sample data was a Wilcoxon Signed Ranks Test which is a non-parametric test to be conducted on two independent samples and was performed with a 95% confidence interval. The test was completed in SPSS with visualisations from Excel.



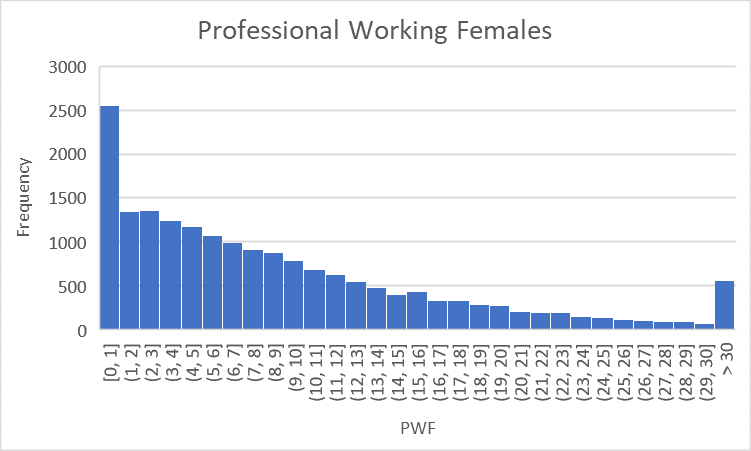
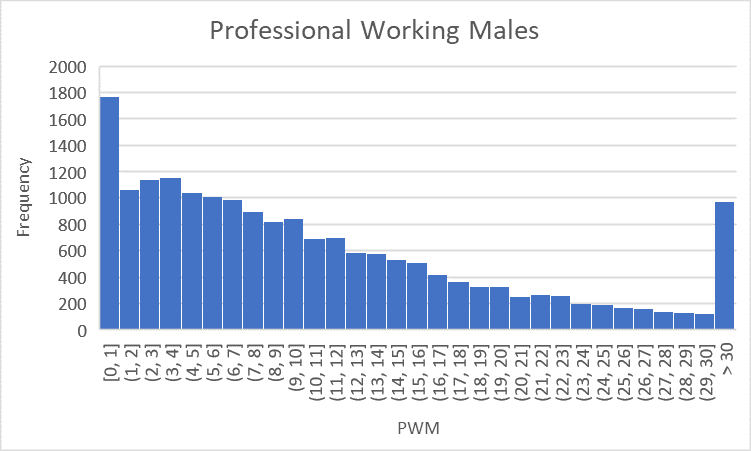
The male sample had a higher overall average but with a higher range (*M* = 11.36 *SD* = 9.934 *R* = 101) as opposed to the females who had a lower average but shorter range (*M* = 9.38 *SD* = 8.661 *R* = 83).

**Results:**



From the ranks result we can discern that the males have a higher mean rank of 8797.45 versus the females mean rank of 6587.71 this shows that overall there are more professional working males than females on average. With the p value being smaller than 0.000 it falls below our confidence interval of 0.05 so we can reject the null hypothesis.

**Visualisations:**



From the above histograms we can see the distribution of each sample. Both are heavily skewed to the right with more than half of the values falling below the mean of the sample. The female sample is slightly more skewed than the male sample with a large number of values falling in the 0-1 range.

**Test 2:** Skilled vs. Semi-skilled vs. Unskilled Males in Small Areas throughout Ireland

Null Hypothesis (*H0*)

*The number of males in each skill category is the same across all areas.*

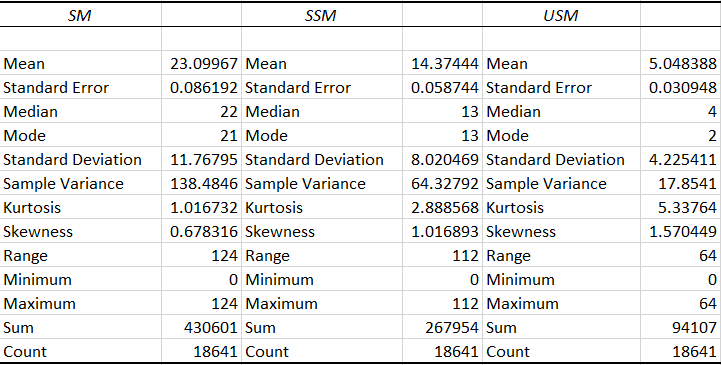
Alternate Hypothesis (*H1*)

*The number of males in each skill category is not the same across all areas.*

Significance Level (*α*)

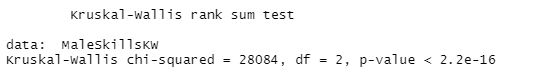
*0.05*

This test was conducted on skilled(*SM*)/semi-skilled(*SSM*)/unskilled(*USM*) males in the same areas throughout Ireland taken from the Small Area Population Statistics dataset provided by the Central Statistics Office. The test performed on the sample data was a Kruskal Wallis H test which is a non-parametric test to be conducted on more than two (in this case three) samples and was performed with a 95% confidence interval. The test was completed in R Studio with descriptive statistics coming from Excel.



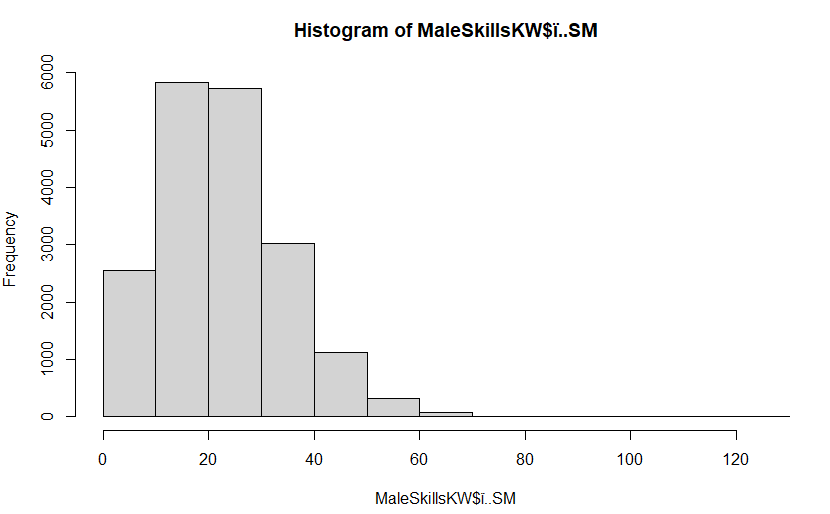
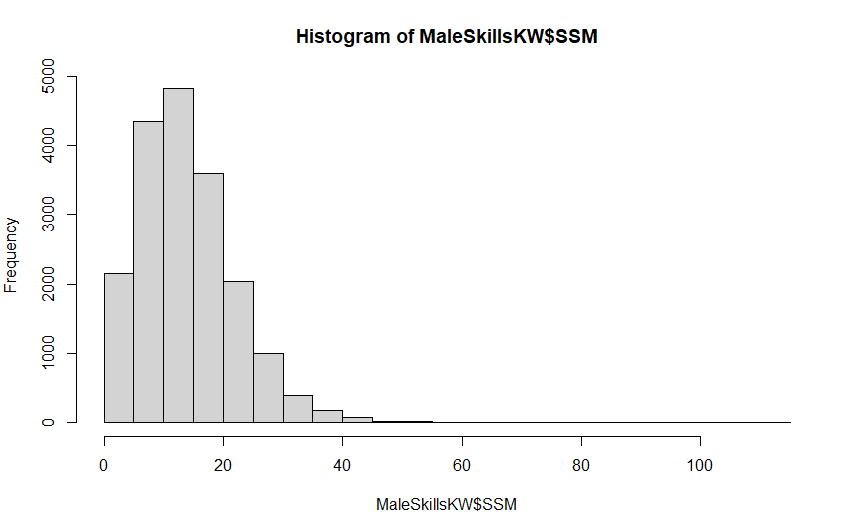
As we can see from the descriptive stats above the skilled males have the highest average of the three samples and the largest sum of all values (*M* = 23.09 *SD* = 11.76 *S* = 430601 *R* = 124)followed by the semi-skilled males sample statistics (*M* = 14.37 *SD* = 8.02 *S* = 267954 *R* = 112) and finally followed by the unskilled males sample which has significantly lower statistics than both other samples (*M* = 5.04 *SD* = 4.22 *S* = 94107 *R* = 64).

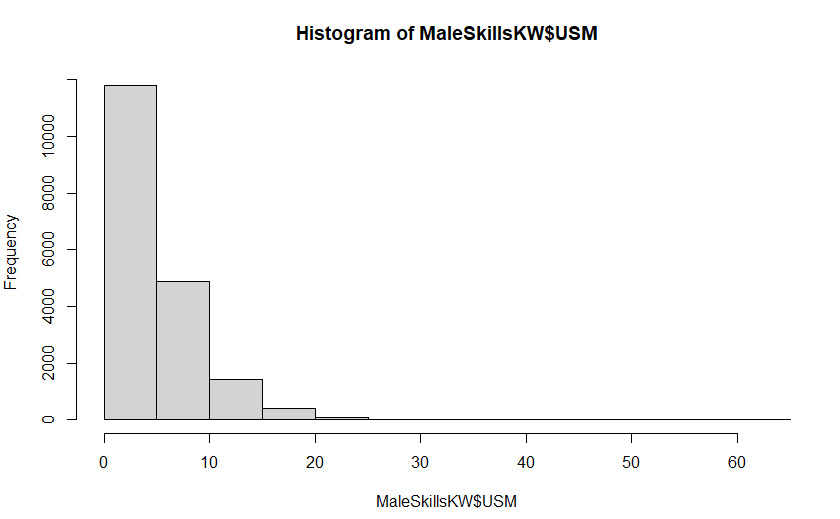
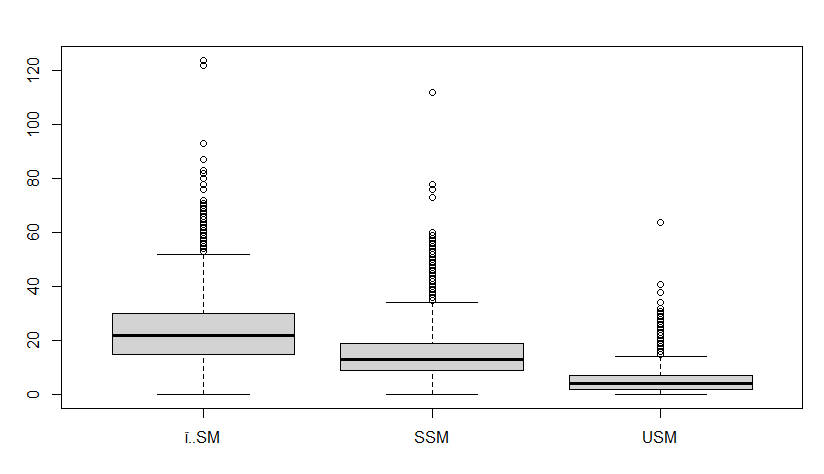
**Results:**



Our h value in this test came as 2.2e-16 which is in scientific notation, when this is converted to a decimal it is 0.00000000000000022 which falls well below our confidence interval of 0.05 so we can reject our null hypothesis.

**Visualisations:**

All 3 of the above histograms are not normally distributed samples and each has a skew to the right. The skilled and semi-skilled workers are more evenly distributed with the skilled males closely grouping around the 20 mark and the semi-skilled males closely grouping around the 15 mark. The unskilled males however are heavily skewed to the right with almost 12000 areas having between 0-5 unskilled male workers. These points are again shown in the boxplot as we can see the distribution for each sample.

**Test 3:** No formal vs. Primary vs. Lower Secondary (Junior Cert) vs. Upper Secondary (Leaving Cert) Education Across Both Genders Throughout Small Areas in Ireland.

Null Hypothesis (*H0*)

*The level of education achieved across Ireland is equal across all areas.*

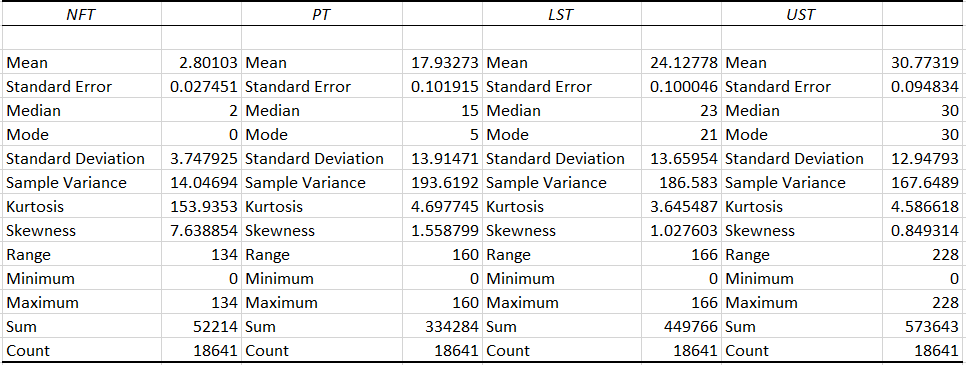
Alternate Hypothesis (*H1*)

*The level of education achieved across Ireland is not equal across all areas.*

Significance Level (*α*)

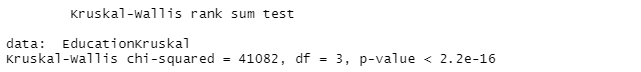
*0.05*

This test was conducted on No Formal (*NFT*) vs. Primary (*PT*) vs. Lower Secondary (*LST*) vs. Upper Secondary (*UST*) across both genders in the same areas throughout Ireland taken from the Small Area Population Statistics dataset provided by the Central Statistics Office. The test performed on the sample data was a Kruskal Wallis H test which is a non-parametric test to be conducted on more than two (in this case four) samples and was performed with a 95% confidence interval. The test was done in R Studio with descriptive statistics coming from Excel.



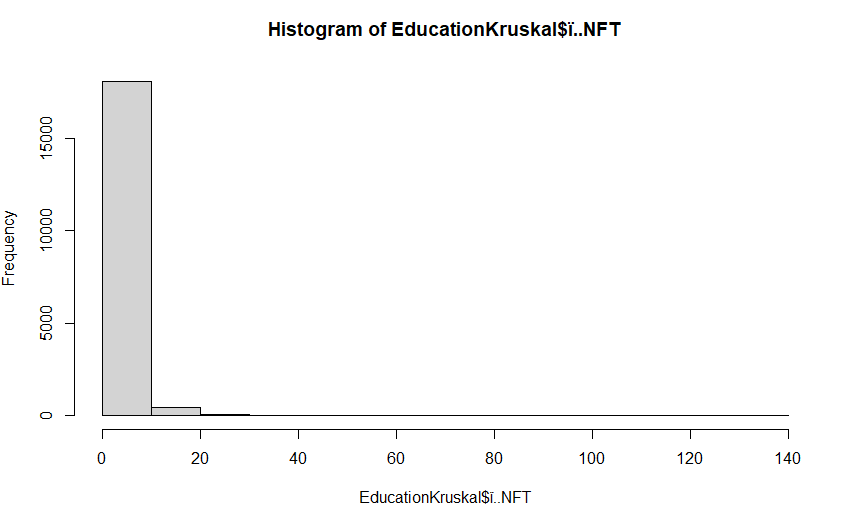
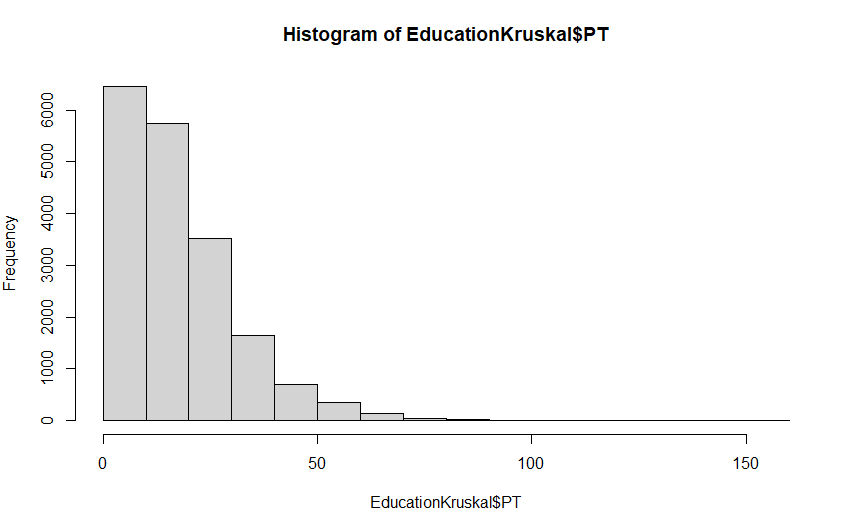
As we can see from the descriptive statistics above, the UST has the highest average of the four samples and the largest sum of all values (*M* = 30.77 *SD* = 12.94 *S* = 573643 *R* = 228)followed by the LST sample statistics (*M* = 24.12 *SD* = 13.65 *S* = 449766 *R* = 166) followed again by the PT sample which has slightly lower statistics than the previous samples (*M* = 17.93 *SD* = 13.91 *S* = 334284 *R* = 160), finally our NFT sample, which falls very significantly below all of our other samples across all of the measures (*M* = 2.80 *SD* = 3.74 *S* = 52214 *R* = 134).

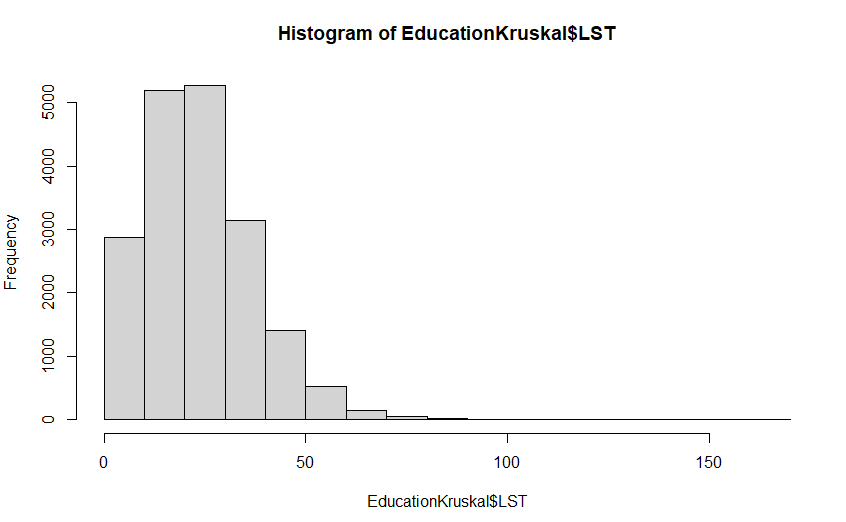
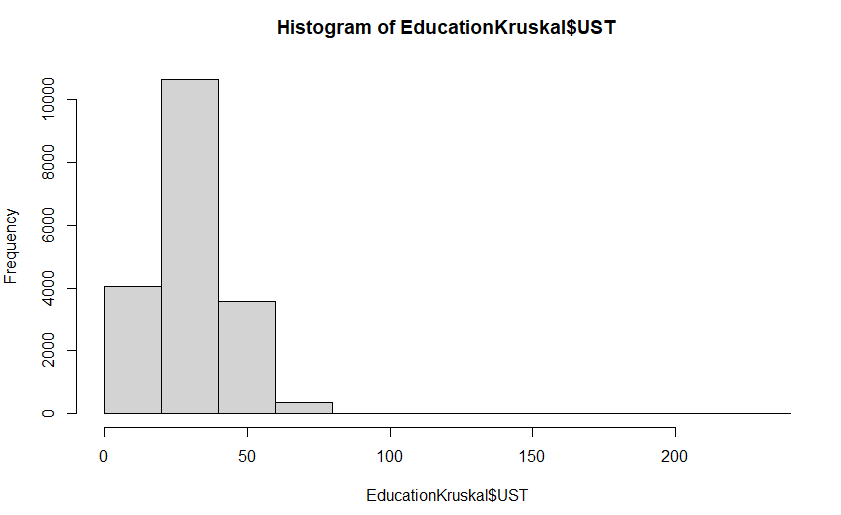
**Results:**



Our h value in this Kruskal Wallis H test was 2.2e-16 which is in scientific notation, when this is converted to a decimal it is 0.00000000000000022 which falls well below our confidence interval of 0.05 so we can reject our null hypothesis.

**Visualisations:**

All 4 of the above histograms are not normally distributed samples and each has a skew to the right. The LST and UST have a slightly more normal distribution opposed to NFT and PT which both have heavily skewed distributions. The NFT sample is extremely skewed with the vast majority of values lying within the 0-10 range. Even though the UST sample has a more normal distribution the modal range of 20-40 holds more than half of the values as can be seen with the high peak and the Kurtosis value of 4.58.

**Test 4:** All Others Gainfully Occupied and Unknown in Dublin, Males vs. Females

Null Hypothesis (*H0*)

*The number of Others Gainfully Occupied and Unknown in Dublin is equal for each gender.*

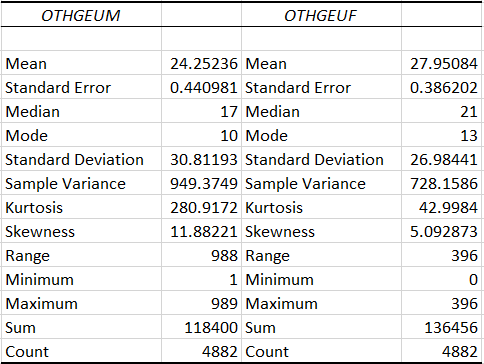
Alternate Hypothesis (*H1*)

*The number of Others Gainfully Occupied and Unknown in Dublin is not equal for each gender.*

Significance Level (*α*)

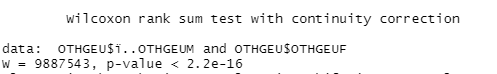
*0.05*

This test was conducted on others gainfully occupied and unknown across both males and females in the same areas throughout Dublin taken from the Small Area Population Statistics dataset provided by the Central Statistics Office. The test performed on the sample data was a Wilcoxon Signed Ranks Test which is a non-parametric test to be conducted on two independent samples and was performed with a 95% confidence interval. The test was completed in R Studio with visualisations and descriptive statistics from Excel.



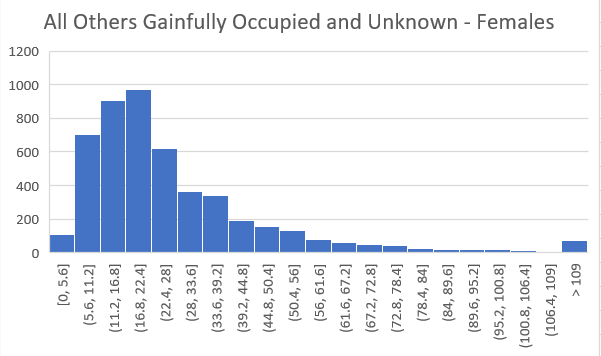
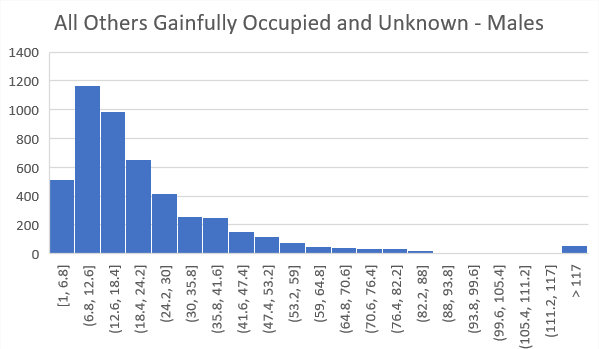
The female sample had a higher overall average and sum of values but with a lower range (*M* = 27.95 *SD* = 26.98 *S* = 136456 *R* = 396) as opposed to the male’s lower average and sum of values but significantly higher range (*M* = 24.25 *SD* = 30.81 *S* = 118400 *R* = 988). We can also see for both samples that the mean, median and modes are not equal and have high kurtosis’ so we can already tell that the data is not normally distributed before seeing the visualisations.

**Results:**



After converting our above p value of 2.2e-16 from scientific notation to decimal we get a value of 0.00000000000000022 which falls below our confidence interval of 0.05 so we reject the null hypothesis

**Visualisations:**



From the above visualisations we can see the distribution of both samples. Both are heavily skewed to the right with the males being slightly more skewed than the females. Both samples also have a number of outliers that we can see in both the histograms above and the range from the descriptive stats.

**Test 5:** Managerial and Technical Employment in Small Ares Throughout Ireland, Male vs. Female

Null Hypothesis (*H0*)

*The number of Managerial and technical workers is equal for both genders.*

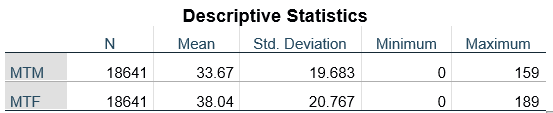
Alternate Hypothesis (*H1*)

*The number of Managerial and technical workers is not equal for both genders.*

Significance Level (*α*)

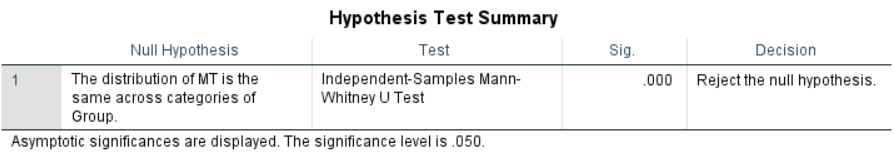
*0.05*

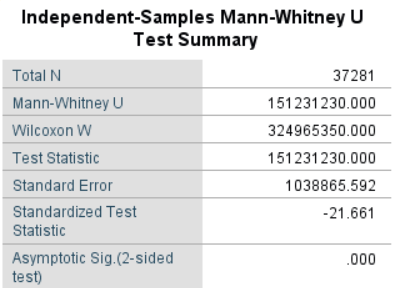
This test was conducted to see if the distribution of managerial and technical roles are equal for both males (*MTM*)and females(*MTF*) in the same areas throughout Ireland taken from the Small Area Population Statistics dataset provided by the Central Statistics Office. The test performed on the sample data was a Mann-Whitney U Test which is a non-parametric test to be conducted on two independent samples and was performed with a 95% confidence interval. The test and all descriptive statistics and visualisations were completed using SPSS.



From the above descriptive statistics we can see that the female sample has a higher average and higher range of the two samples (*M* = 38.04 *SD* = 20.76 *R* =189). However the male sample is not far off of the female average or range (*M* = 33.67 *SD* = 19.68 *R* = 159).

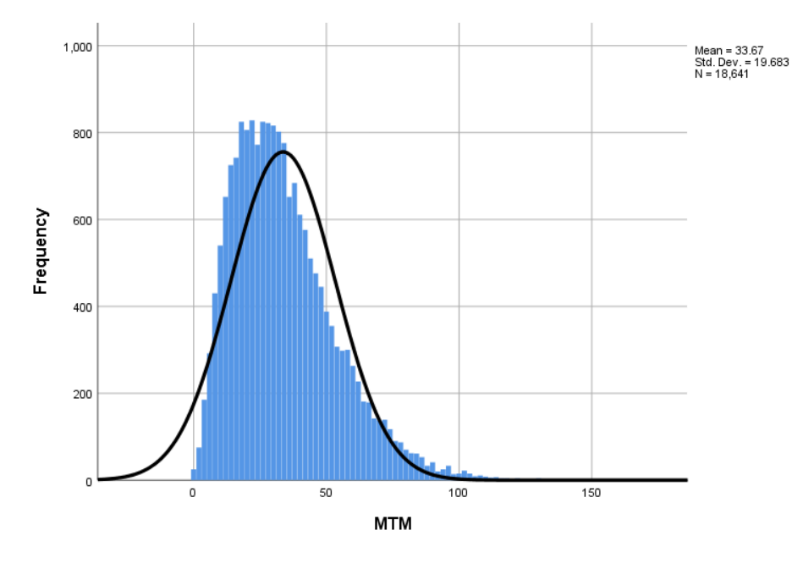
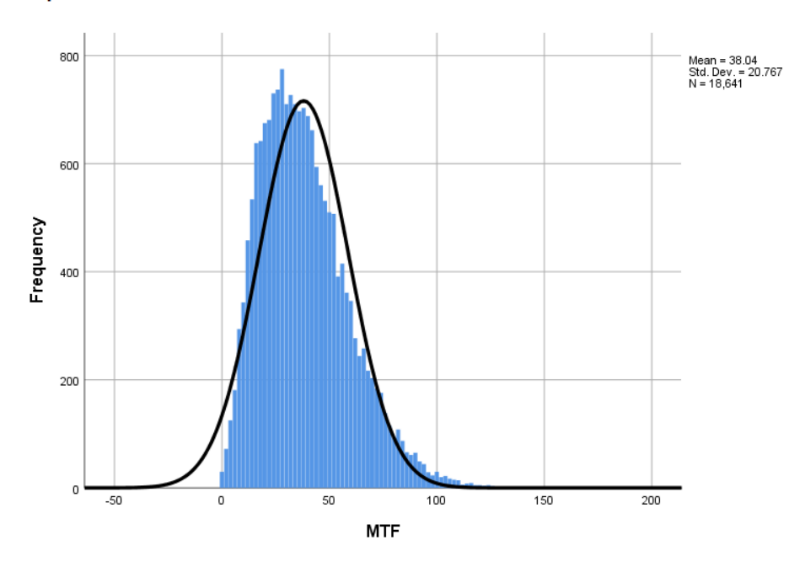
**Results:**

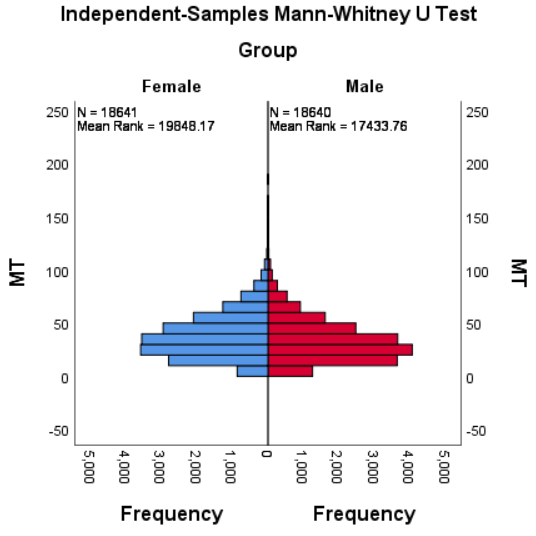




From our test results we can see that our p value returned 0.000, this falls below our confidence interval of 0.05 so we can agree with SPSS and reject the null hypothesis.

**Visualisations:**



From both of the histograms of the samples we can see that both samples skew to the right with the male sample skewing slightly more. We can also see that the female sample, although having a more normal distribution, has a higher peak. This can be further seen in the above graph with both histograms side by side and more easily comparable.

**Conclusions:**

1. **Test 1:** For comparing professional working male and female in small areas in Ireland we conducted a Wilcoxon Signed Ranks Test to test if there is a significant difference between our non normally distributed samples among the samples and to get our p value. We have strong support to reject the null hypothesis as the p value turned out to be less than our confidence interval of 0.05 so we can reject our null hypothesis and accept our alternate hypothesis for this test.
2. **Test 2:** For this test which compares skilled, semi-skilled and unskilled workers in areas throughout Ireland we ran a Kruskal-Wallis H Test on the samples to test if there is a significant difference between our three non normally distributed samples.. From our test result we have evidence to reject our null hypothesis as our h value was 0.00000000000000022 which falls far below the confidence interval of 0.05 so we reject the null hypothesis and accept our alternate hypothesis in this case.
3. **Test 3:** For test 3 we compare all individuals throughout the small areas in Ireland with no formal, primary, lower secondary and upper secondary levels of education. For these samples we again ran a Kruskal-Wallis H Test on the samples to test for significant difference among the four samples. This time from our test result we also have evidence to reject our null hypothesis as our h value was 0.00000000000000022 which is below the confidence interval of 0.05 so the null hypothesis is rejected and the alternate hypothesis is accepted.
4. **Test 4:** For the fourth test we are comparing all other gainfully occupied and unknown in Dublin for both males and females. The test used on these two samples was the Wilcoxon Signed Ranks Test to examine if there is significant difference between both samples. From running our test we got a return value of 0.00000000000000022 for our p value. This gives us sufficient evidence that we can reject our null hypothesis and accept our alternate hypothesis
5. **Test 5:** For the final test we are analysing managerial and technical roles across the small areas in Ireland for both males and females. With our dependant for the sample being ordinal (in this case our dependants are male and female) a Mann-Whitney U test is suitable to use to examine if there is significant difference among the two samples. After running the test the p value returned was 0.000 which provides suitable evidence to reject the null hypothesis as the p value falls below the 0.05 confidence interval, so again for this test we reject the null hypothesis and accept the alternate hypothesis.