Statistical Tests on Various Different Cultural Aspects of Irish Life Based on Census 2016 Data

Paul Long
School of Computing
National College of Ireland
Dublin, Ireland

Table of Contents

Word Count: 2482 words

This report is based on a project undertaken regarding data used from the 2016 Irish Census data[1], completed by the Central Statistics Office. There are five insights that gleamed by manipulating the data and running various statistical tests, based on 15 categories of data about life in Ireland and its people.

The five questions investigated were as following:

1) Is there a significant interaction effect between the urbanity of the counties with the three largest cities regarding the proportion of people engaged in primary and secondary industries (agriculture, forestry, and fishing, building and construction, and manufacturing). That is, is there a significant relationship in being from the city or country and being from Cork, Galway, or Dublin in the proportion of working in these industries?
2) Is there a significant difference in the proportion of internet usage between people from Donegal (North-West) and people from Carlow (South-East)? Furthermore, has the proportion of internet usage changed since the 2011 census[2]?
3) Is there a significant difference in the proportion of people employed vs total in the counties with the 5 largest Irish Cities (Dublin, Cork, Galway, Waterford, and Limerick)?
4) How long are people's commutes? Additionally, after how long has everyone completed their commute?
5) Is gender independent of a student's choice to study in an Abstract Field, a Verbal Field, or an Other Field?

These questions will be answered using Kolmogorov-Smirnov and Shapiro-Wilk tests for Normality, Two Way ANOVA with Tukey post hoc test, Mann Whitney U Test, Wilcoxon Signed Rank Test, Kruskal Wallis Test with Tukey post hoc test, and Chi Square Test for Independence. Some of these calculations relied upon the fact that the test statistics -E(V)/std dev $\sim$ N(0,1), via the Central Limit Theorem.

The first step in this process was to read the 2016 Census Data and the 2016 Census Glossary into Python, matching the, via GUIDs. This dataframe was then inner joined with the 2011 Census Data, so that any new townlands were not included.

Following this, the data was manipulated so that there was an adequate format for the completion of the statistical test – joining counties and regions together, deleting irrelevant columns, creating new data via arithmetic. An example of this manipulation was in Insight 1, where the number of 'Total People in Agriculture, Fisheries and Farming' was added to 'Total people in Building and Construction' and 'Total People in Manufacturing Industries', and divided by the 'Total People', producing the 'Proportion of people in Primary and Secondary Industries'.

This process was done for five insights, creating the right format for each statistical test that would be carried out. These then formed five dataframes, which were saved.

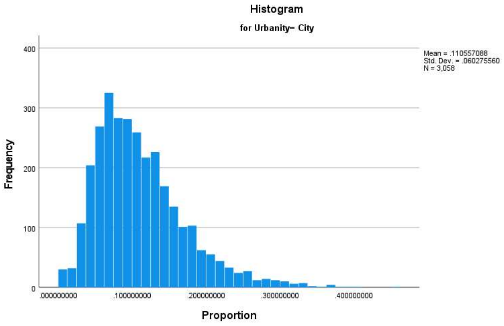
Insight 1

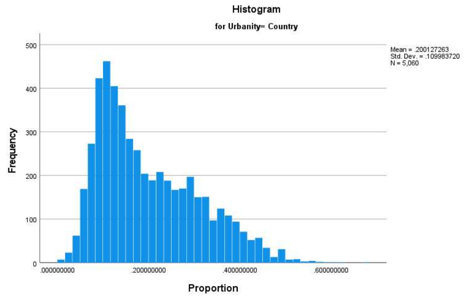
This goal was to investigate whether there was a significant relationship in being from the city or country and being from Cork, Galway, or Dublin in the proportion of being involved in either of these industries. This is done via a Two Way ANOVA Test. The data was processed such that there was one continuous variable (The proportion from each townland), and two groups of independent categorical data (Urbanity and County). Thus, a test for normality and variance was performed.

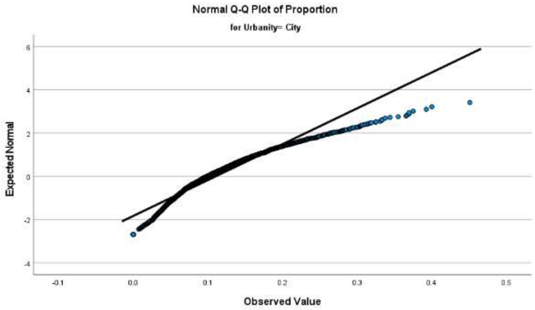
The data is quite highly positively skewed, however each cell was skewed in a similar way and each distribution were quite similar. In addition, all QQ plots approximated a straight line. IN addition, each cell had similar (low) variances, the statistics being 0.012 and 0.004, and 0.011, 0.009 and 0.002.

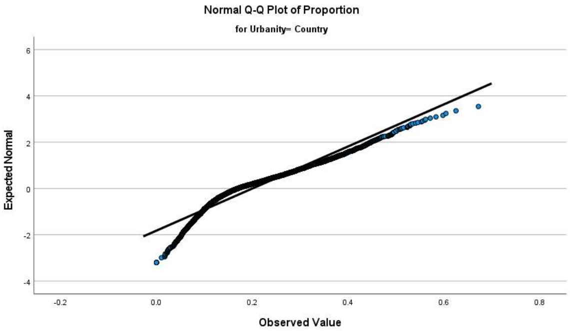
Furthermore, There were also no extreme outliers. There were some long tails, but the data looked too continuous to be an outlier. Whilst this data is not ideal data, Two Way ANOVA is robust in dealing with skewness. This is the justification behind computing the Two Way ANOVA Test. The test was also computed with alpha = 0.025, to minimize any potential rise in Type I error due to the small effect that the Two Way ANOVA may have had. The test was conducted in SPSS.

Urbanity

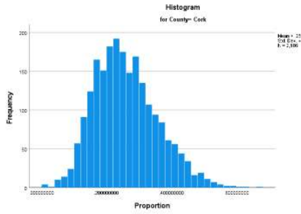


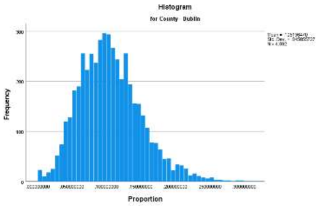


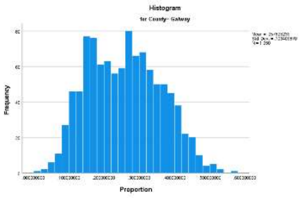


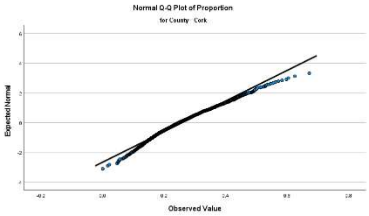


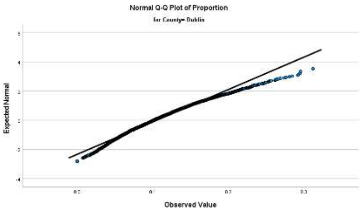
Counties

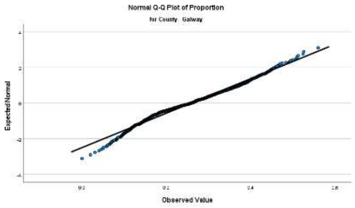












Statistical Report: Get Fa crit calculations.

H_0a: There was no difference in the mean proportion of workers in Primary and Secondary Industries across counties.

H_1a: There was a difference in the mean proportion of workers in Primary and Secondary Industries across counties.

H_0b: There was no difference in the mean proportion of workers in Primary and Secondary Industries across urbanity.

H_1b: There was a difference in the mean proportion of workers in Primary and Secondary Industries across urbanity.

H_0c: Urbanity and County are independent in the mean proportion of workers in Primary and Secondary Industries across urbanity. That is, there is no interaction effect.
H_1c: Urbanity and County are not independent in the mean proportion of workers in Primary and Secondary Industries across urbanity. That is, there is an interaction effect.

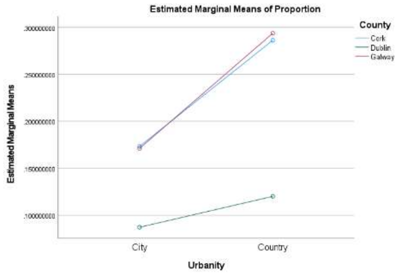
α **=** 0.025
DFa = 2, DFb = 1, and DFc = 2.

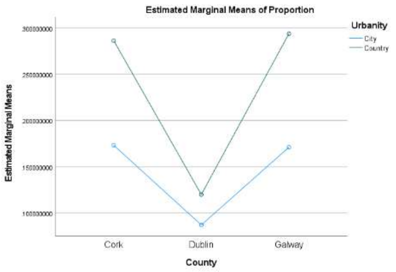
There was an Fa stat value of 3338, an Fb stat of 2299 and Fc stat 361.
Each hypothesis had a p value of: 0.000, 0.000 and 0.001 respectively.

Null hypotheses are rejected.
Hence, it appears there was a significant difference in:

- the means across counties,
- the means across urbanity
- there was a significant interaction effect.

From the Post Hoc pairwise comparison, there were p values of <0.001 between Dublin and Cork, and Dublin and Cork, suggesting there was a significant difference. There was a p value of 0.847 between Galway and Cork, suggesting there was no significant difference.

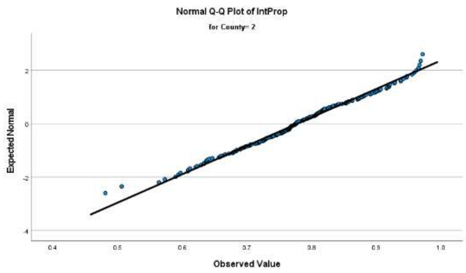


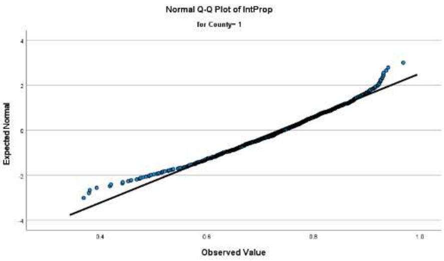


However, it should be noted that this is the case whilst there were some serious issues with the data – such as the skew and the normality. Thus, there shouldn't be much weight placed on this insight, if any at all.

Insight 2

This goal was to investigate whether there was a significant difference in the proportion of households with internet. This was done via a Mann Whitney U Test. A test for normality and homogeneity of variance was conducted. The Kolmogorov-Smirnov test for Normality concluded that data from Carlow appeared to be approximately normally distributed, but that the data from Donegal appeared to be not normally distributed, with significance levels of 0.20 and <0.001 respectively.





Hence, a Two Tailed Mann Whitney U Test was conducted to test for whether there was a significant difference. This test was conducted in SPSS:

Statistical Report:

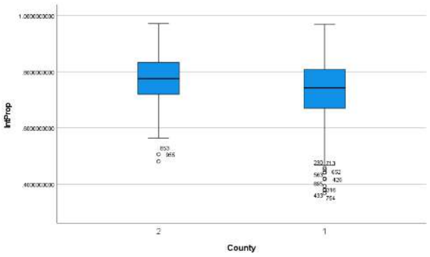
H_0: There was no significant statistical difference in the proportion of households with internet in Donegal and Carlow.
H_1: There was a significant statistical difference in the proportion of households with internet in Donegal and Carlow.

α **=** 0.05
$U_{donegal}$ = 63249.

Std dev = $(((760*213*(760+213+1))/12)^{1/2}$ = 3624.812
Z_{stat} = (63249-(760*213)/2)/ 3624.812= -4.880
Z_{crit} = -1.96
$Z_{stat} < Z_{crit}$

Null hypothesis is rejected.
It appears there is a statistical difference in the proportion of households with internet in Donegal and Carlow.

From the box plots and the descriptive statistics, Carlow appears to have a higher proportion of households with internet than Donegal (means of 0.736 and 0.778 respectively).



There was then a subsequent Wilcoxon Signed Rank Test completed to investigate whether overall households with internet had changed since 2011. Since the Donegal data was not normally distributed, this was sufficient to accept that a Wilcoxon Test was appropriate. Thus, a Two Tailed Wilcoxon Signed Rank Test was conducted in SPSS. The townlands that did not appear in both were not included in this test.

Statistical Report:

H_0: There was no significant statistical difference in the proportion of households with internet in Donegal and Carlow between 2011 and 2016.
H_1: There was a significant statistical difference in the proportion of households with internet in Donegal and Carlow between 2011 and 2016.

α **=** 0.05
n = 966-1=965 (due to tie)

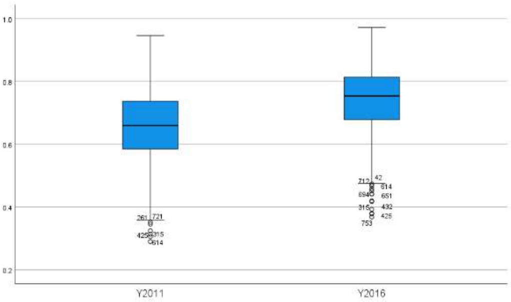
W+ = 68, Variable = 11511.5.
Std dev = $(((966*965)*(965*2+1))/24)^{1/2}$ = 8660.396

Z_{stat} = (11511.5-(966*965)/4)/ 8660.396 = -25.580
Z_{crit} = -1.96

$Z_{stat} < Z_{crit}$

Null hypothesis is rejected.
It appears there is a statistical difference in the mean proportion of households with internet in Donegal and Carlow between 2011 and 2016.

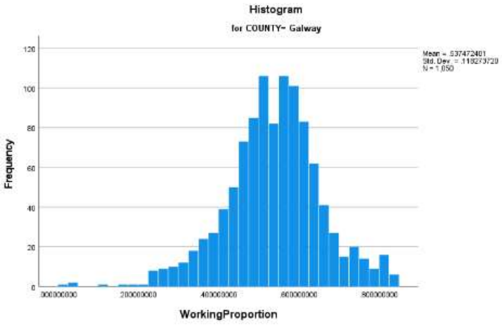
Furthermore, from choosing the W+ statistic, the 2016 has a higher proportion than 2011. This is further evidenced by means of 0.659 and 0.744 respectively.

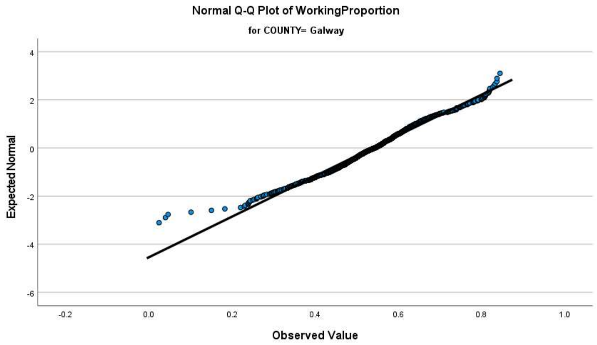


Insight 3

This goal was to investigate whether there is a significant difference in the proportion of people employed vs total people in the counties with the largest cities in Ireland. This was done via a Kruskal Wallis test, with a post hoc Tukey test.

A test for normality and homogeneity of variance was initially conducted. Both the Kolmogorov-Smirnov and Shapiro Wilk tests for Normality concluded that none of the data from each county appeared to be normally distributed. Hence, this data was appropriate for a Kruskal Wallis test. Hence, a Kruskal Wallis Test was conducted to test for whether there was a significant difference between counties. This test was conducted in SPSS.





Statistical Report:

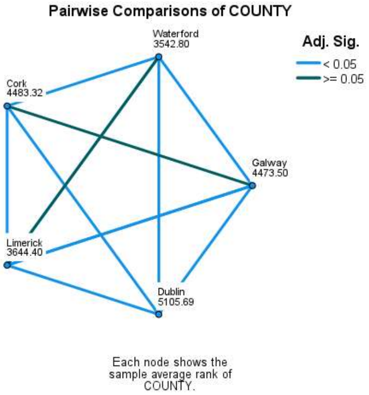
H_0: There was no significant statistical difference in the proportions of people working compared to total people in Dublin, Galway, Cork, Limerick, and Waterford.
H_1: There was no significant statistical difference in the proportions of people working compared to total people in Dublin, Galway, Cork, Limerick, and Waterford.
α **=** 0.05
DF = 4

$H_{stat} = 336.978$
$H_{crit} = 9.488$
$H_{stat} < H_{crit}$

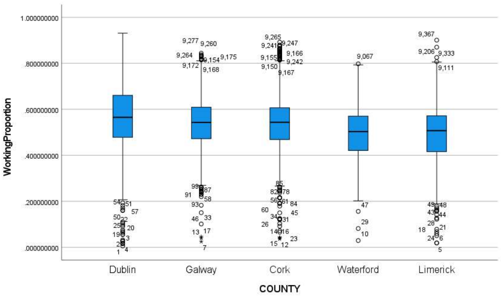
Null hypothesis is rejected.

It appears there is a statistical difference in the proportions of people working compared to the total people in Dublin, Galway, Cork, Limerick, and Waterford.

From the Tukey post hoc test, there are significant differences between every county except Limerick-Waterford and Cork-Galway.



Furthermore, it's clear that Dublin has the highest average proportion of working people, with Waterford and Limerick having comparable average proportions, those being the lowest, Cork and Galway having comparable averages in between Dublin and Waterford & Limerick.



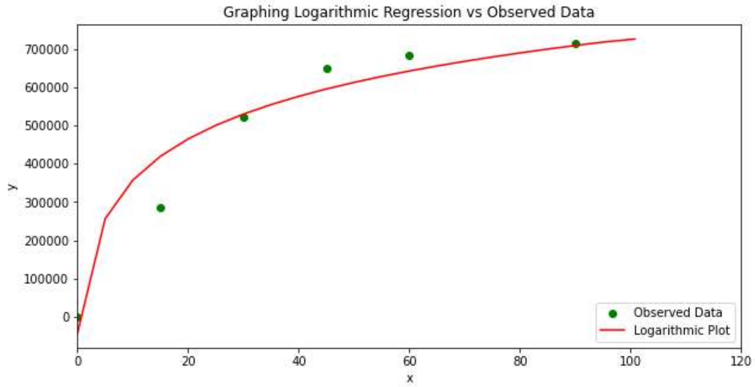
Insight 4

This goal was to investigate how many people finish their work, school, or college commute per minute in the counties of Munster. This was done by creating a logarithmic regression model. This involved taking the bins for each county (0-15 mins, 15-30 mins, 30-45 mins, 45-60 mins, 60-90 mins and 90+ mins) and cumulating the data so that the data approximated a logarithmic model. Then, the higher bound of each bin was taken and 1 was added to each (16 mins, 31 mins, etc.) to use as the dependent variable (ln(16), ln(31), etc.). This was done so that the regression line could start at 0 and give a better approximation. The regression line was computed.

The equation of the regression line, accounting for the extra minute was:

166337.395*ln(x+1)-41413.930.

From this, some calculations were made. For example, it's approximated that 73882 people's work commute in Munster took 1 minute. It was also calculated that after 99 minutes and 45 seconds, all the people in Munster will have completed their commute.



Insight 5

This goal was to investigate whether gender is independent of a student's decision to study different categories of subjects – Abstract, Verbal, and Other. This was done via a Chi Square Test for Independence.

Initially, a random sample of 100 townlands was chosen via a random number generator. Then, the fields were set out as: 'Science, Mathematics and Computing', 'Engineering, Manufacturing and Construction', and 'Art' were chosen as the 'Abstract' Fields, 'Social Studies, Humanities and Law', 'Education' and 'Humanities' were chosen as the 'Verbal' fields, and 'Health and Welfare', 'Agricultural and Veterinary', 'Services' and 'Other' were chosen as the 'Other' fields. These were then totalled for each of the 100 townlands.

Then the data was totalled into 6 cells – Male-Abstract, Male-Verbal, Male-Other, & Female-Abstract, Female-Verbal, Female-Other, before totals were calculated for each category. The Chi Square Test was then calculated by hand in Excel.

Statistical Report:

H_0: Students' Gender and Subject Category were independent.
H_1: Students' Gender and Subject Category were not independent.

α = 0.05
DF = 2

X^2_{stat}= 981.5264
X^2_{stat} = 5.99
$X^2_{stat} > X^2_{crit}$

Null hypothesis is rejected.
It appears that students' gender is not independent from their subject category.

Furthermore, we can see that from 7400 students (3493 males and 3907 females), 624 more males chose Abstract Fields than expected and were down in Verbal and Other, whereas 342 more females chose Verbal Fields than expected and females were up 282 in Other Fields but were significantly down in Abstract Fields.

These five insights were designed to gain interesting facts and pieces of information about the people in Ireland from the census data. However, there were some issues. It must be mentioned that the first insight had less than ideal data. The level of skew was imperfect, and the failing of some of the normality and variance tests were a cause of concern. Thus, the caveat at the end was important. The insight was designed to investigate an interesting question and also to use the Two Way ANOVA. However, the data probably would not hold up to scrutiny, however I felt that it was still worthwhile to do. Also, none of the data tested had perfectly normal data, and this was the best available, given the confines of the format of the test. Also, it would have been preferrable to have more data to use to create the regression model in Insight 4.

There were five questions posed from the 2016 census data:

1) is there a significant interaction effect between the county and urbanity for people in primary and secondary industries in Dublin, Cork, and Galway?
2) Is there a significant difference in proportion households with internet from Donegal and Carlow? Has this proportion of internet usage changed since 2011?
3) Is there a significant difference between the proportions of people employed vs total people in Dublin, Cork, Galway, Waterford, and Limerick?
4) How long are people's work commutes? Approximately how long is the longest work commute?
5) Are students' study choice category and gender independent in 100 randomly sampled townlands?

Each question, through the use of various statistical testing of the data, has been answered fully, and insights have been gained:

1) There is a significant interaction effect between county and urbanity for people in Dublin, Cork, and Galway in primary and secondary industries, though this must be caveated as the data was imperfect for the test.
2) There is a significant difference between Donegal and Carlow's internet proportion – Carlow has a higher proportion. Furthermore, there has been a substantial increase since 2011 in the internet usage of both places.
3) There is a significant difference in the proportion of people employed in Dublin, Cork, Galway, Waterford and Limerick, Dublin having the highest proportion, Waterford and Limerick having the lowest.
4) The equation 166337.395*ln(x+1)-41413.930 approximates the amount of people have completely their commute per minute. It's approximated that the everybody will have finished their commute by 1 hour 40 minutes.
5) Student's decision on whether to choose an Abstract field, a Verbal field or Another field was not independent from gender. From the data, more males chose abstract fields, and more females chose verbal and other fields.

Bibliography

1. CSO, (2016), Census Data 2016, Ireland, CSO
2. CSO, (2016), Census Data 2016, Ireland, CSO