**Secrets of the society**

- Who does not like to answer sensitive questions and which information we do not like to share?

Estera Wisniewska

HdSDa word count: 2117

Table of Contents

[**1.** **INTRODUCTION** 2](#_Toc112338063)

[**2.** **LOCATION - CHECKING IF ALL PARTS OF IRELAND ARE EQUALY SECRETIVE** 3](#_Toc112338064)

[**STEP I – Calculating the sample size and selecting sampling method.** 3](#_Toc112338065)

[**STEP II – Checking if each county has a normal distribution.** 4](#_Toc112338066)

[**STEP III – LOOKING for the variance BETWEEN AREAS** 6](#_Toc112338067)

[**3.** **NARROWING THE LOCATION** 8](#_Toc112338068)

[**STEP I - Checking if each county has a normal distribution.** 8](#_Toc112338069)

[**STEP II - LOOKING FOR THE VARIANCE BETWEEN DUBLIN’S COUNTIES** 9](#_Toc112338070)

[**4.** **GENDER** 11](#_Toc112338071)

[**STEP I – calculating sample size and choosing sampling method** 11](#_Toc112338072)

[**STEP II – Checking if gender samples have normal distribution** 11](#_Toc112338073)

[**STEP III – checking if there is significant difference between Male and Female regarding to not answering questions** 13](#_Toc112338074)

[**5.** **QUESTIONS – WHAT IS THE SUBJECT ARE WE TRYING TO HIDE THE MOST?** 14](#_Toc112338075)

[**sTEP I – CHECKING NORMALITY OF EACH SAMPLE** 15](#_Toc112338076)

[**STEP II – COMPARING IF ALL QUESTIONS ARE NOT BEEING ANSWERED AT THE SAME LEVEL** 16](#_Toc112338077)

[**6.** **TESTING – is it true that Wexford is more secretive than ‘the REST’?** 18](#_Toc112338078)

[**7.** **SUMMARY** 20](#_Toc112338079)

[Bibliography 21](#_Toc112338080)

# **INTRODUCTION**

We all have a secrets. It is in a human nature to keep some information to ourself and do not share with anybody else but close family or friends. It is true that we like to always keep a few cards in our sleeve and use them in right time just to gain advantage over someone.

We like to feel that we are better than others, we like to present ourselves in better light and keeping some information to ourself can make it happen. But, why are we doing the same to a census questions which are anonymous? We cannot do ‘better’ here? We are not able to score higher than the rest, so why are we doing it? Are we ashamed of something? Or scared to admit that we might be not as good as ‘the rest’? Or, maybe it is in our nature? If so, do some of us are more prone to keep secrets than others?

Let check what society is trying to hide the most and who might be the most secretive person of us all - Person X.

# **LOCATION - CHECKING IF ALL PARTS OF IRELAND ARE EQUALY SECRETIVE**

For the beginning of this analysis, we will have a closer look at census questions based on Small Area Population Statistics (SAPS) [1] which were answered by people living in Ireland.

Data for each county is divided into multiple ‘small areas’ which represent a different number of dwellings therefore for more accurate results we are going to use percentage of total responds (by question in each county) as our data sets.

In census from 2016 there was 18 questions where person had option ‘Other’ or could write answer by himself but did not do it. Both of this answers were recorded as ‘not stated’ and that is exactly what we will be focusing on.

### **STEP I – Calculating the sample size and selecting sampling method.**

We want to check how secretive is population of Ireland therefore will start our analysis by creating sufficient sample size.

Small Area Population Statistics recorded 31 areas/counties and this will be our population size – N;

Confidence level = 95%

Margin of Error = 5% =>

sample size = 29 (based on online calculator [2] )

Using simple random sampling method all areas were assignes to a unique number and 29 numbers were pulled at random.

DESCRIPTIVE STATISTICS:

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

### **STEP II –** **Checking if each county has a normal distribution.**

H0: all counties have approximately normal distribution

HA: at least one of the county has not normal distribution

α = 0.05 - most suitable for social studies

We will accept H0 if Sig > α and reject if H0 Sig < α.

RESULTS:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Tests of Normality** | | | | | | | |
|  | County | Kolmogorov-Smirnova | | | Shapiro-Wilk | | |
|  | Statistic | df | Significance | Statistic | df | Significance |
| scores | Carlow | .139 | 10 | .200\* | .924 | 10 | .388 |
| Cavan | .160 | 10 | .200\* | .962 | 10 | .807 |
| Clare | .249 | 10 | .078 | .899 | 10 | .214 |
| Cork City | .157 | 10 | .200\* | .964 | 10 | .832 |
| Donegal | .208 | 10 | .200\* | .854 | 10 | .064 |
| Dublin city | .170 | 10 | .200\* | .947 | 10 | .631 |
| Fingal | .211 | 10 | .200\* | .900 | 10 | .220 |
| Galway city | .173 | 10 | .200\* | .932 | 10 | .465 |
| Monaghan | .259 | 10 | .055 | .847 | 10 | .053 |
| Offlay | .181 | 10 | .200\* | .931 | 10 | .453 |
| Roscommon | .185 | 10 | .200\* | .938 | 10 | .529 |
| Sligo | .249 | 10 | .080 | .861 | 10 | .079 |
| Dublin South | .165 | 10 | .200\* | .934 | 10 | .483 |
| Tipperary | .127 | 10 | .200\* | .974 | 10 | .921 |
| Waterford | .127 | 10 | .200\* | .974 | 10 | .921 |
| Westmeath | .213 | 10 | .200\* | .897 | 10 | .201 |
| Wexford | .121 | 10 | .200\* | .967 | 10 | .864 |
| Wicklow | .185 | 10 | .200\* | .933 | 10 | .476 |
| Meath | .105 | 10 | .200\* | .986 | 10 | .989 |
| Galway Co | .200 | 10 | .200\* | .944 | 10 | .602 |
| Kerr | .158 | 10 | .200\* | .929 | 10 | .435 |
| Kildare | .175 | 10 | .200\* | .952 | 10 | .695 |
| Kilkenny | .192 | 10 | .200\* | .920 | 10 | .358 |
| Laois | .208 | 10 | .200\* | .901 | 10 | .224 |
| Leitrim | .224 | 10 | .170 | .877 | 10 | .121 |
| Limerick | .124 | 10 | .200\* | .962 | 10 | .808 |
| Longforf | .201 | 10 | .200\* | .937 | 10 | .519 |
| Louth | .105 | 10 | .200\* | .980 | 10 | .964 |
| Mayo | .223 | 10 | .174 | .870 | 10 | .101 |
| \*. This is a lower bound ... | | | | | | | |
| Chart  Description automatically generateda. Lilliefors Significance Correction | | | | | | | |

Looking at the Shapiro – Wilk test results we can see that significance for every county is **higher than α = 0.05** which means we accept H0.

CONCLUSION:

All counties have approximately normal distribution.

DISCLOSURE: to avoid Type II Error sample size within each county should be increased from 10 to a suitable proportion of county data in Small Area File.

### **STEP III – LOOKING for the variance BETWEEN AREAS**

Is any of the county more secretive than the other? In order to answer this question we will have to look at the difference between them. Is there any difference and if so, which county likes to keep secrets.

CHOOSING THE TEST:

We have 29, independent, random samples with normal distribution therefore best test would be One-Way Anova but we are failing variance requirement. Because our samples are small and variance between each sample is not equal we will proceed with Kruskal-Wallis test.

H0: sample median are equal

HA: sample median are not equal

α = .05 - social study

We will accept H0 if Sig. > α and reject if Sig. < α .

RESULTS:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Hypothesis Test Summary** | | | | |
|  | Null Hypothesis | Test | Sig.a,b | Decision |
| 1 | DefaultAllDistinctValuesUsed\_I\_KruskalWallis\_2 | Independent-Samples Kruskal-Wallis Test | .001 | Reject the null hypothesis. |
| a. The significance level is .050. | | | | |
| b. Asymptotic significance is displayed. | | | | |

Significance = .001

.001 < .05 Reject H0

Kruskal-Wallis Test results shows that not all medians are equal => there is a difference between counties median.

Chart, radar chart

Description automatically generated

Table

Description automatically generated

CONCLUSION:

Population of Ireland is not equally secretive and some counties prefer not to answer questions at higher level than others. Biggest difference is between Wexford and Donegal.

# **NARROWING THE LOCATION**

Results above shows that there is significant difference between counties in whole Ireland, but will it be the same withing one county? If one county will be divided into smaller groups will all parts have different level of secrecy?

To answer this question we will compare counties of Dublin, in more specific: Dublin South and Fingal.

DESCRIPTIVE STATISTICS:

Using Python code we can produce descriptive for both samples [3].

Dublin\_/south Fingal

count 18.000000 18.000000

mean 8.887222 7.852778

std 12.869790 12.600763

min 0.610000 0.920000

25% 2.767500 1.825000

50% 3.510000 2.475000

75% 6.390000 5.897500

max 50.050000 49.530000

### **STEP I - Checking if each county has a normal distribution.**

H0: all counties have approximately normal distribution

HA: at least one of the county has not normal distribution

α = 0.05 - most suitable for social studies

We will accept H0 if Sig > α and reject if H0 Sig < α.

CHOOSING THE TEST:

Best way to check normality of smaller samples is Shapiro Wilk test. We will accept H0 when p-value will be > α and reject when p-value will be < α.

Chart, scatter chart

Description automatically generatedRESULTS:

From the Python output we can see the result of Shapiro Wilk test for each county:

Dublin South:

Statistic = 0.6146

P-value = 9.476696504862048e-06 = **0.000009**

Chart, scatter chart

Description automatically generated

Fingal:

Statistic = 0.5853

P-value = 4.8568349484412465e-06 = **0.000005**

The p-values for both counties is smaller than α = .05 therefore we reject H0.

CONCLUSION :

Results show that none of the counties have a normal distribution.

### **STEP II -** **LOOKING FOR THE VARIANCE BETWEEN DUBLIN’S COUNTIES**

CHOOSING THE TEST:

We have 2 small, random, continuous samples with not normal distribution, taken from the same area, which I will assume, depends/pair together at some level – best test will be Wilcoxon Signed Rank Test.

H0: median difference between 2 counties is zero

HA: median difference between 2 counties is not zero

α = .05 - social study

n = 18

We will accept H0 if Wstat > Wcrit and reject if Wstat < Wcrit

RESULTS:

Table

Description automatically generated

T- = 26.5

T+ = 144.5

Wstat = 26.5

A screenshot of a computer

Description automatically generated with low confidence

Wcrit = 40

Wstat < Wcrit 26.5 < 40

Based on results we are going to reject H0.

CONCLUSION :

Our Wilcoxon Signed Rank Test results shows that there is a difference in medians between South Dublin and Fingal which means that there is a difference between South Dublin and Fingal regarding not answering questions.

# **GENDER**

In this part of analysis we will focus on gender, are Females more or less secretive than Males? Or are we prefer not to share some information at the same level?

### **STEP I – calculating sample size and choosing sampling method**

Small Area Population Statistics has 18600 records,

N = 18600

Confidence level = 95%

Margin of Error = 5% => **n = 376**

Since there is no gender related pattern, best method for this sampling will be systematic sampling where data is organised in order. One random small area was chosen as a starting point and next 375 small areas were choses every 50th row in data set.

We have 9 questions with possible answer ‘Not Stated’ or ‘Other’ which were divided by a gender. Answers from those questions were calculated as percentage of total responds for the question and calculated average for each of Small Area was taken.

DESCRIPTIVE STATISTICS:

Table

Description automatically generated

### **STEP II – Checking if gender samples have normal distribution**

H0: both genders have approximately normal distribution

HA: at least one gender has not normal distribution

α = 0.05 - most suitable for social studies

We will accept H0 if Sig > α and reject if H0 Sig < α.

RESULTS:

Chart, line chart

Description automatically generated

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Chart, line chart  Description automatically generated  **Tests of Normality** | | | | | | | |
|  | Gender | Kolmogorov-Smirnova | | | Shapiro-Wilk | | |
|  | Statistic | df | Significance | Statistic | df | Significance |
| Not\_answered | Male | .117 | 376 | .000 | .888 | 376 | .000 |
| Female | .111 | 376 | .000 | .862 | 376 | .000 |
| a. Lilliefors Significance Correction | | | | | | | |

Both gender’s p value / Significance is lower than α = 0.05.

CONCLUSION :

Both groups have not normal distribution.

**STEP III – checking if there is significant difference between Male and Female regarding to not answering questions.**

CHOOSING THE TEST:

We have 2 independent, big samples with not normal distribution therefore best test for checking difference between them will be Mann – Whitney U Test.

H0: there is no difference between gender

HA: there is difference between gender

α = .05 - social study

We will accept H0 if Sig > α and reject if H0 Sig < α.

RESULTS:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Hypothesis Test Summary** | | | | |
|  | Null Hypothesis | Test | Sig.a,b | Decision |
| 1 | DefaultAllDistinctValuesUsed\_I\_MannWhitneyU\_2 | Independent-Samples Mann-Whitney U Test | .935 | Retain the null hypothesis. |
| a. The significance level is .050. | | | | |
| b. Asymptotic significance is displayed. | | | | |

Mann-Whitney U Test Sig. > .05 therefore we will accept H0.

CONCLUSION :

Test results strongly indicates that there is no difference between gender regarding to not answering census questions.

# **QUESTIONS – WHAT IS THE SUBJECT ARE WE TRYING TO HIDE THE MOST?**

It is commonly known that there are some sensitive subjects which people are trying to avoid. We will not talk about them, ask questions and will not answer if any sensitive to us topic will arise. It is in our nature.

But what subject is sensitive to us in census?

In this part we are going to analyse each question where population of Ireland could select ‘Other’ or not answer at all. We are assuming that the person by choosing option ‘Other’ do not want to answer the question as a main reason - rather than lack of variety of answers.

DESCRIPTIVE STATISTICS:

Table

Description automatically generated

Table

Description automatically generated

### **sTEP I – CHECKING NORMALITY OF EACH SAMPLE**

H0: all questions have approximately normal distribution

HA: at least one of the question has not normal distribution

α = 0.05 - most suitable for social studies

We will accept H0 if Sig > α and reject if H0 Sig < α.

RESULTS:

Shapiro Wilks results from Python output:

Table

Description automatically generated

Chart, box and whisker chart

Description automatically generated

P-value for question 15 is < α therefore we will fail to accept H0.

CONCLUSION :

Not all samples have approximately normal distribution.

### **STEP II – COMPARING IF ALL QUESTIONS ARE NOT BEEING ANSWERED AT THE SAME LEVEL**

In Small Area Population Statistics file we have 18 questions where answers were logged as ‘Other’ or ‘Not stated’. We will be looking for a difference between questions.

CHOOSING THE TEST:

We have 18 samples, with not normal distribution – best test will be Kruskal Wallis H test.

H0: sample median are equal

HA: sample median are not equal

α = .05 - social study

We will accept H0 if Sig. > α and reject if Sig. < α .

RESULTS:

Python output for Kruskal Wallis H test shows:

H: 187.13831273199932 , p-value: 1.0853988671552265e-30

Significant difference between the Questions is found.

CONCLUSION:

Results of the test concludes that not all questions are ‘not answered’ at the same level. People decided not to answer to some questions more often than to other.

Chart, radar chart

Description automatically generated

Table

Description automatically generatedAfter proceeding with paired comparisons we can see that most often questions where people decide not to answer are:

* Question 10 – Population aged 15 years and over by field of study – Education
* Question 11 – Population aged 15 years and over by sex and highest level of education complete – Education
* Question 12 – Population by general health and sex – Health

# **TESTING – is it true that Wexford is more secretive than ‘the REST’?**

In our first test we concluded that county Wexford is more secretive than other. We took 10 random small areas from each county, calculated average from each and that was our sample for each county.

Now we will compare responses for each question. We are going to check if Wexford don’t answer particular question at similar level as the whole population.

DESCRIPTIVE STATISTICS:

As a representation of population we calculated sample size and took 376 small areas at random.

Table

Description automatically generated

Table

Description automatically generated

CHOOSING THE TEST:

We want to measure the discrepancy between Wexford and the population’s level of not answering questions - the best test to do is Chi Square test (Goodness of fit).

We start with population sample n = 376, calculating average percentage for each question, and this will be our EXPECTED value.

Wexford average percentage for each question will be OBSERVED value.

Table

Description automatically generated

RESULTS:

χ² = ∑(O - E)² / E = **35.48**

α = .05 ; df = 17

Table

Description automatically generated

χ² crit = **27.59**

χ² > χ² crit

Wexford do not follow population distribution and answers questions at different level than population.

CONCLUSION:

Wexford is truly more secretive than the rest of Irish population.

# **SUMMARY**

At the beginning of this analysis we stated that we all have secrets and we do not like to share some information in general.

Based on Small Area file, we conducted statistical tests and discovered that each county do not answer question at different level and Wexford is the most secretive area of Ireland. People living there, on average, are more likely to chose answer ‘Other’ or not answer to question at all, than the rest of population.

Similar results were founded after performing test on Dublin area. Different parts of Dublin have different level of secrecy.

We found out as well that the most sensitive subject is Education and Health. Questions relating to this subjects were left unanswered the most.

Another test results showed that there is no difference between female and male in secrecy level. Unfortunately data set did not allow us to perform any further analysis and we were unable to discover who is the most secretive person from us all. Characteristics on Person X remaining unknown.

(Unless, Person X is so secretive that, did not respond to census questions at all.)

# Bibliography

|  |  |
| --- | --- |
| [1] | C. S. Office, “Cencus 2016 Small Area Population Statistics,” [Online]. Available: https://www.cso.ie/en/census/census2016reports/census2016smallareapopulationstatistics/. [Accessed 2022]. |
| [2] | “Sample size calculator,” 2022. [Online]. Available: https://www.surveysystem.com/sscalc.htm. |
| [3] | “Data science made simple,” 2022. [Online]. Available: https://www.datasciencemadesimple.com/descriptive-summary-statistics-python-pandas/#:~:text=Descriptive%20or%20summary%20statistics%20in%20python%20%E2%80%93%20pandas%2C,columns%20and%20gives%20summary%20statistics%20of%20numeric%20columns. |