Data Analysis and Statistical Inference

Final Project Report

Introduction and Aim of the Study

The main target of this study is to investigate any possible relation between religion and financial income in the US in the last decade. More precisely I decided to focus on Protestants, Catholics and the ones who claimed to belong to no religious community at all (identified as None). These three categories, according to the data, have been the most common ones in the United States in the period 2000-2012.
Therefore the target of the investigation may be summarized by the following question: **"Is there a relationship between the religious faith of a US citizen (Protestant, Catholic or None) and his/her total family income?"**
My personal interest derives from a general conviction that there could be some religious communities wealthier than others, due to historical, social or political reasons and the exploration of this kind of aspects may lead to underline some specific features going on behind the scenes. More generally I thing that digging this matters may enlighten some subtle pattern hiding behind the data, such as religious discrimination at work resulting in people belonging to a specific community getting higher, more qualified and more paid job positions. Highilighting this kind of aspects is a starting point for a broader research about social and financial conditions among and within different religious communities.

General Discussion about the Data of Interest

The research project was based on the data collected in the online-available database of the General Social Survey, 1972-2012 [Cumulative File] (ICPSR 34802)[1](GSS), which since 1972 has been monitoring societal change and studying the growing complexity of American society. The GSS aims to gather data on contemporary American society in order to monitor and explain trends and constants in attitudes, behaviors, and attributes; to examine the structure and functioning of society in general as well as the role played by relevant subgroups; to compare the United States to other societies in order to place American society in comparative perspective and develop cross-national models of human society.

The dataset is composed by 57061 cases corresponding to an equal number of interviewed citizens. Each person may be considered as a single case, as there are several pieces of information (variables) which were recorded for each case. In particular, cases are single individuals respecting the following characteristics: all non institutionalized, English and Spanish speaking persons 18 years of age or older, living in the United States. As just mentioned each candidate was asked several questions about a number of aspects of his own life, his family, his community, the society he lives in.
The data were collected by three main methods:

[1] Citation: Smith, Tom W., Michael Hout, and Peter V. Marsden. General Social Survey, 1972-2012 [Cumulative File]. ICPSR34802-v1. Storrs, CT: Roper Center for Public Opinion Research, University of Connecticut /Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributors], 2013-09-11. doi:10.3886/ICPSR34802.v1

- *computer-assisted personal interview* (CAPI). Data are inserted directly into an electronic sheet over a PC and the interviewer and the respondent are both present at the moment of the survey, in front of the computer. The difference with CASI is that in the latter the interviewed is left alone in order to answer the questions more privately.

- *face-to-face interview.*

- *telephone interview.*

The two variables I took into account from the dataset are the following:

- **income06**: categorical variable. The interviewed was asked which of the proposed groups he/she thought his/her total-tax-less-family income of the previous year would fall. There are 25 possibile interval varying from a minimum of less than 1000$ to a maximum of more than 150.000$, plus a category named "Refused", including all the cases who did not accept to reveal their finacial condition.

- **relig**: categorical variable. The interviewed was asked by which of the proposed communities his/her religious faith would better be identified. A more proper description of the variable is going to be provided during the exploratory data analysis.

The study is observational because researchers recorded data "in a way that does not directly interfere with how the data arise". The structure of the survey and the data collection methods are clearly not typical of an experimental setup. In the latter case, in fact, researchers would have sampled individuals and divided them into groups organizing an experiment in order to investigate the possibility of a causal connection between two or more variables. From the point of view of the generalizability of the study it is crucial to focus on the population of interest, whic,h in this very case, includes all non institutionalized, English and Spanish speaking persons 18 years of age or older, living in the United States. According to 2011 American Community Survey Data on Language Use 79.2205% of american families speak English at home, while the 20.7794% speak Spanish which added up result in a global 99.9999%.

This means that we can reasonably generalize the results to the totality of US population 18 years of age or older. Furthermore the used data collection methods compensate each other in terms of any potential source of sampling bias. For instance, CAPI is mainly addressed to computer friendly persons. This bias may be prevented by phone interviews which enable researchers to reach and convince less technology-friendly people. The last but not the least is the face-to-face survey which compensates the unavoidable bias introduced by a phone call. The latter gives for granted the connection to a phone line which may not always be respected. In addition to that it is necessary to consider that generally children, youngsters or more generally minors do not have a clear financial overview over the family. Despite their belonging to a particular religious community they may have great insights into the total family income, which means that their contribution to the survey, at least on this very aspect, would have been pointless.

All this considerations lead us to the conclusion that the results of the study may be generalized to all US families. However, since the survey is observational, the findings do not imply causal relationships.

Exploratory Data Analysis

In the present section a brief exploratory data analysis is performed. The relevant statistics is provided together with the associated R code.

The first two used functions R are *summary* and *str*, which help to get a broader and in the mean time synthetic view over the data. As ir is clear the *gss.after.subsetting* dataset is composed by only two variables , *Income* and *Religion.* Both are factors consisting respectively in 26 (actually 25 as I did not ttake into account the *Refused* category) and 13 levels. In particular *Religion* mantains all the original 13 levels despite only three of them have been selected (Protestant, Catholic and None).

```
> summary(gss.after.subsetting)

        Income          Religion
 60-74.999 : 826   Protestant:4606
 40-49.999 : 759   Catholic  :2154
 50-59.999 : 674   None      :1622
 75-89.999 : 628   Jewish    :   0
 90-109.999: 517   Other     :   0
 150-Over  : 517   Buddhism  :   0
 (Other)   :4461   (Other)   :   0

> str(gss.after.subsetting)

'data.frame': 8382 obs. of  2 variables:
 $ Income  : Factor w/ 26 levels "Below-1","1-2.999",..:
             16 21 18 17 18 18 21 9 4 8 ...
 $ Religion: Factor w/ 13 levels "Protestant","Catholic",..:
             4 2 1 1 4 2 1 1 2 1 ...
```

In order to visualize the data in a cleaner way a plot is provided too. The whole dataset has been converted into a contingency table, which has been properly plotted in the figure below.

Figure 1 shows pretty clearly the distribution of incomes among and within the three investigated communities. Nevertheless it is quite hard to identify any particular pattern hiding behind the data. It is necessary to proceed with a more complete and rigorous analysis in order to draw any conclusion concerning a possible correlation between religious community and family financial income. For further details about the data see the Appendix at the end of the report.

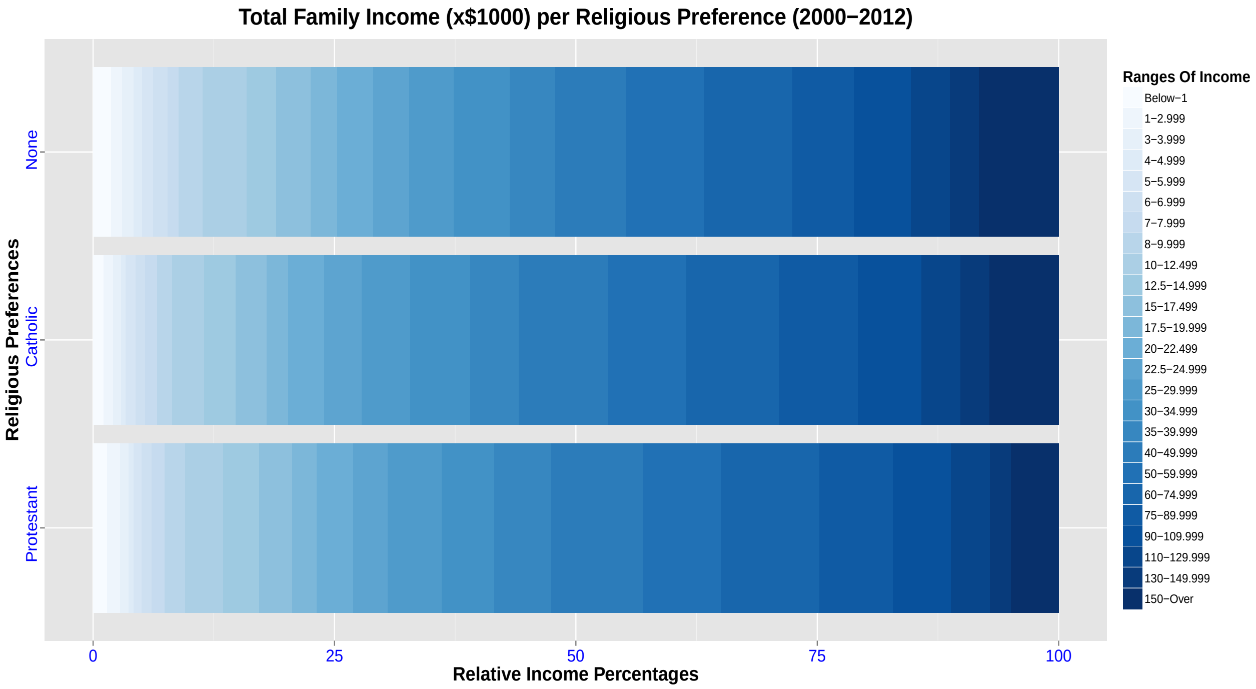

Figura 1: Graphical representation of the contingency table. In order to properly visualize and compare the data, the whole dataset has been converted to relative probabilities.

Inference

As stated at the end of the previous section in order to end up with a proper conclusion and answer the original question at the base of the study it is necessary to perform a rigorous statistic test on the dataset. First a purpose of clearness, first of all we recap the main target of the project which is to answer the following question: **"Is there a relationship between the religious faith of a US citizen (Protestant, Catholic or None) and his/her total family income?"**

As we are dealing with two categorical variables (*Income* and *Religion*), both of which with more than two levels (respectively 25 and 3), only an hypothesis test is admittable. In particular, as no defined parameter of interest can be highlighted, I have performed a theroretical Chi-square test for independence, which is allowed by the fact that each particular scenario (i.e. cell count) has at least 5 expected cases. The proof of this condition being met is provided in the following table, which summarizes the whole dataset with each case joined by its expected value. As you can see all scenarios have expected value well above 5.

The total number of degrees of freedom is $df = (R-1) \times (C-1)$ which is equal to $df = (25-1) \times (3-1) = 48$, well above the minimum allowed of 2.
As for the independence issue, the GSS sampling has been randomic and in any case the number of cases in each scenario as well as the total amount of cases is below the 10% of the population of the US.

Income	*Protestant*		*Catholic*		*None*	
(x$1000)	**Real**	Expected	**Real**	Expected	**Real**	Expected
Below-1	**67.00**	65.90	**23.00**	30.80	**30.00**	23.20
1-2.999	**62.00**	56.60	**22.00**	26.50	**19.00**	19.90
3-3.999	**40.00**	41.80	**17.00**	19.50	**19.00**	14.70
4-4.999	**26.00**	28.00	**11.00**	13.10	**14.00**	9.90
5-5.999	**36.00**	42.30	**22.00**	19.80	**19.00**	14.90
6-6.999	**48.00**	51.10	**21.00**	23.90	**24.00**	18.00
7-7.999	**63.00**	59.90	**27.00**	28.00	**19.00**	21.10
8-9.999	**98.00**	94.00	**33.00**	43.90	**40.00**	33.10
10-12.499	**181.00**	179.70	**72.00**	84.00	**74.00**	63.30
12.5-14.999	**172.00**	160.50	**70.00**	75.00	**50.00**	56.50
15-17.499	**156.00**	156.10	**70.00**	73.00	**58.00**	55.00
17.5-19.999	**118.00**	114.80	**47.00**	53.70	**44.00**	40.40
20-22.499	**174.00**	173.60	**81.00**	81.20	**61.00**	61.10
22.5-24.999	**166.00**	169.80	**83.00**	79.40	**60.00**	59.80
25-29.999	**256.00**	241.80	**109.00**	113.10	**75.00**	85.10
30-34.999	**252.00**	263.20	**133.00**	123.10	**94.00**	92.70
35-39.999	**270.00**	250.00	**109.00**	116.90	**76.00**	88.00
40-49.999	**440.00**	417.10	**199.00**	195.00	**120.00**	146.90
50-59.999	**369.00**	370.40	**175.00**	173.20	**130.00**	130.40
60-74.999	**471.00**	453.90	**206.00**	212.30	**149.00**	159.80
75-89.999	**349.00**	345.10	**176.00**	161.40	**103.00**	121.50
90-109.999	**279.00**	284.10	**142.00**	132.90	**96.00**	100.00
110-129.999	**185.00**	185.70	**87.00**	86.90	**66.00**	65.40
130-149.999	**100.00**	116.50	**64.00**	54.50	**48.00**	41.00
150-Over	**228.00**	284.10	**155.00**	132.90	**134.00**	100.00

Given that, we can state our hypothesis:

- H_0 : (nothing going on): *Religion* and *Total Family Income* are **independent**, meaning that the amount of money earned by a US family per year **does not vary** by belonging to either the Protestant or the Catholic community, or no religious community at all.
- H_A : *Religion* and *Total Family Income* are **dependent**, meaning that the amount of money earned by a US family per year **does vary** by belonging to either the Protestant or the Catholic community, or no religious community at all.

Let's recall that applying the Chi-square test for independence means that we are to evaluate whether there is convincing evidence that a set of observed counts O_{11}, O_{12},O_{13} ..., O_{RC} in $R \times C$ categories are unusually different from what might be expected under a null hypothesis. Call the expected counts that are based on the null hypothesis, E_{11}, E_{12},E_{13} ..., E_{RC}, computed as

$$E_{row\ i,\ col\ j} = \frac{(row\ i\ total) \times (column\ j\ total)}{table\ total}$$

If certain conditions are met, then the test statistic below follows a chi-square distribution with $(R-1) \times (C-1)$ degrees of freedom:

$$\chi^2 = \frac{(O_{11} - E_{11})^2}{E_{11}} + \frac{(O_{12} - E_{12})^2}{E_{12}} + \cdots + \frac{(O_{RC} - E_{RC})^2}{E_{RC}}$$

The p-value for this test statistic is found by looking at the upper tail of this Chi-square distribution. We consider the upper tail because larger values of χ^2 would provide greater evidence against the null hypothesis.

The result of the Chi-square test for independence over the dataset of interest is the following:

```
> chisq.test(gss.after.subsetting)

        Pearson's Chi-squared test

data:  gss.after.subsetting
X-squared = 77.695, df = 48, p-value = 0.004269
```

Because we typically test at a signifcance level of $\alpha = 0.05$ and the p-value is less than 0.05, the null hypothesis is rejected. That is, the data provide convincing evidence that there is some association between the amount of money earned by a US family per year and belonging to either the Protestant or the Catholic community, or no religious community at all.

Conclusions

The aim of the research project was to investigate whether there could be any association between the total tax-free income of an American family and their belonging to the Protestant, Catholic or none religious community at all. The dataset was taken from the General Social Survey 1972-2012 (GSS), which since 1972 has been monitoring societal change and studying the growing complexity of American society. the original database was subset in order to take into account only the two variables of interest, *Income* and *Religion*; the latter has been subset itself to select only three inner levels, Protestant, Catholic and None. Due to the tipology of the investigated data, only a hypothesis test based on the Chi-square test for independence can be performed.

The result of the statistical analysis leads us to reject the null hypothesis and then to state that there is some association between the amount of money earned by a US family per year and belonging to either the Protestant or the Catholic community, or no religious community at all.

This could be only the beginning of a wider study about correlation between religion and financial condition in the US. Deeper insights must be get into the matter and more complex statistical tools and techniques must be used in order to infer complete and satisfying conclusions.

APPENDIX - Attached Dataset

```
> xtabs(~ Income + Religion, data=gss.after.subsetting) -> gen.adm.table
> addmargins(gen.adm.table)

              Religion
Income        Protestant Catholic None  Sum
  Below-1             67       23   30  120
  1-2.999             62       22   19  103
  3-3.999             40       17   19   76
  4-4.999             26       11   14   51
  5-5.999             36       22   19   77
  6-6.999             48       21   24   93
  7-7.999             63       27   19  109
  8-9.999             98       33   40  171
  10-12.499          181       72   74  327
  12.5-14.999        172       70   50  292
  15-17.499          156       70   58  284
  17.5-19.999        118       47   44  209
  20-22.499          174       81   61  316
  22.5-24.999        166       83   60  309
  25-29.999          256      109   75  440
  30-34.999          252      133   94  479
  35-39.999          270      109   76  455
  40-49.999          440      199  120  759
  50-59.999          369      175  130  674
  60-74.999          471      206  149  826
  75-89.999          349      176  103  628
  90-109.999         279      142   96  517
  110-129.999        185       87   66  338
  130-149.999        100       64   48  212
  150-Over           228      155  134  517
  Sum               4606     2154 1622 8382
```

```
> addmargins(prop.table(table(gss.after.subsetting[,"Income"],
             gss.after.subsetting[,"Religion"]),2)*100,1)

                Protestant Catholic None
  Below-1             1.45     1.07   1.85
  1-2.999             1.35     1.02   1.17
  3-3.999             0.87     0.79   1.17
  4-4.999             0.56     0.51   0.86
  5-5.999             0.78     1.02   1.17
  6-6.999             1.04     0.97   1.48
  7-7.999             1.37     1.25   1.17
  8-9.999             2.13     1.53   2.47
  10-12.499           3.93     3.34   4.56
```

```
12.5-14.999       3.73     3.25   3.08
15-17.499         3.39     3.25   3.58
17.5-19.999       2.56     2.18   2.71
20-22.499         3.78     3.76   3.76
22.5-24.999       3.60     3.85   3.70
25-29.999         5.56     5.06   4.62
30-34.999         5.47     6.17   5.80
35-39.999         5.86     5.06   4.69
40-49.999         9.55     9.24   7.40
50-59.999         8.01     8.12   8.01
60-74.999        10.23     9.56   9.19
75-89.999         7.58     8.17   6.35
90-109.999        6.06     6.59   5.92
110-129.999       4.02     4.04   4.07
130-149.999       2.17     2.97   2.96
150-Over          4.95     7.20   8.26
Sum             100.00   100.00 100.00
```