Data Analysis and Statistical Inference Project

Arindam Biswas

01 Nov 2015

Introduction

It is often said that number of children in the household is dependent on the race of the household. The goal of this project is to analyze ***Is there a relationship between race of respondent and number of children in the household?***.

Number of children in the household affects the pool of economic resources available to the children within the household, and hence plays an important role in the economic and social well-being of children in the household. So studying the relationship between race of respondent and number of children in the household will help us understand these aspects.

Further the number of children in the household also have a broader impact on the overall demand for economic and social services, thereby acting as a measure for Government for better preparedness to meet such demands in future.

Data:

The data has been collected by *NORC at the University of Chicago*. The data was collected through a sociological survey *General Social Survey (GSS)*, used to collect data on demographic characteristics and attitudes of residents of the United States. The data has been collected cumulatively from 1972 to 2012.

The data has a total of 57,061 cases and 114 variables. However the number of variables differ over the years as the number of variables in the initial years were fewer and over the years new questions has been added to the survey and hence new variables are added to the data set.

The cases for this project are adult resident of the United States, who in their household have provided responses for the GSS survey questions pertaining to **race: Race of respondent** and **childs: Number of children**

The data variables to be studied are: ‘

- `race`: Race of respondent. A categorical and non-ordinal variable with levels `White`, `Black` and `Other`
- `childs`: Number of children. A numerical and discrete variable

As a sociological survey has been used to collect data on demographic characteristics and attitudes of residents of the United States. So the data was collected in a way that does not directly interfere with how the data arises, hence the study is observational.

As the survey was designed to administered randomly to any adult resident of United States, that indicates that random sampling was used to administer the survey, hence the results can be generalized to the entire the population.

While there was potential source of bias arising from non-response to the survey. However broadly the non-response to the survey has been also random and hence we can establish that the survey was responded by a sample set of respondents who are representative of the population.

As the study is observational, so at best we can only establish association between the variables of interest. Whereas if the study was an experiment, the data could have been used to establish causal links between the variables of interest.

Exploratory Data Analysis:

First we load the libraries required for the exploratory data analysis

```
library(ggplot2)
```

Then we use the following code to load the GSS dataset into R

```
load(url("http://bit.ly/dasi_gss_data"))
```

The GSS survey is being conducted for more than 40 years, however to be relevant in our analysis we are interested only in current data (surveys collected in 21st century), so we subset the data to include data only from year 2000 to 2012.

```
gss<-subset (gss,year %in% 2000:2012)
```

We obtain a sample size of **18945** observations from year 2000 to 2012.

Then we study the list of the variable names in the dataset

```
names(gss)
```

Then we analyse the structure of the two variables of interest - `race` and `childs`. We observe that `race` is categorical and non-ordinal. While `childs` is numerical and discrete

```
str(gss$race)
```

```
##  Factor w/ 3 levels "White","Black",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
str(gss$childs)
```

```
##  int [1:18945] 0 2 5 1 0 0 0 2 1 3 ...
```

Then we create a subset of the data using dataframe gss_subset with only the two variables of interest - `race` and `childs` and set the column headers as Race and ChildCount

```
gss_subset <- data.frame(Race=gss$race, ChildCount=gss$childs)
```

Now we remove those rows from the data frame gss_subset where either `race` or `childs` have missing (NA) values

```
gss_subset <- gss_subset[!(is.na(gss_subset$Race)) & !(is.na(gss_subset$ChildCount)),]
```

We observe that the number of cases/observations in the gss_subset data frame is **18898**

Then we summarize the race (categorical variable) using a table and a bar plot. We also summarize the childs (numerical variable) using a summary, a histogram and a box plot.

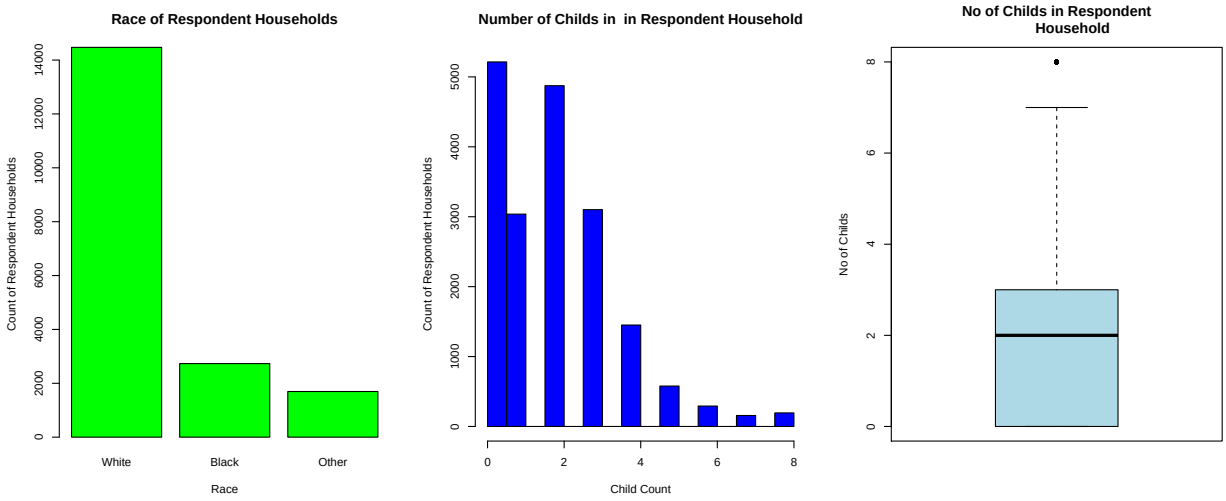
```
table(gss_subset$Race)
```

```
##
## White Black Other
## 14472  2731  1695
```

```
summary(gss_subset$ChildCount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   2.000   1.862   3.000   8.000
```

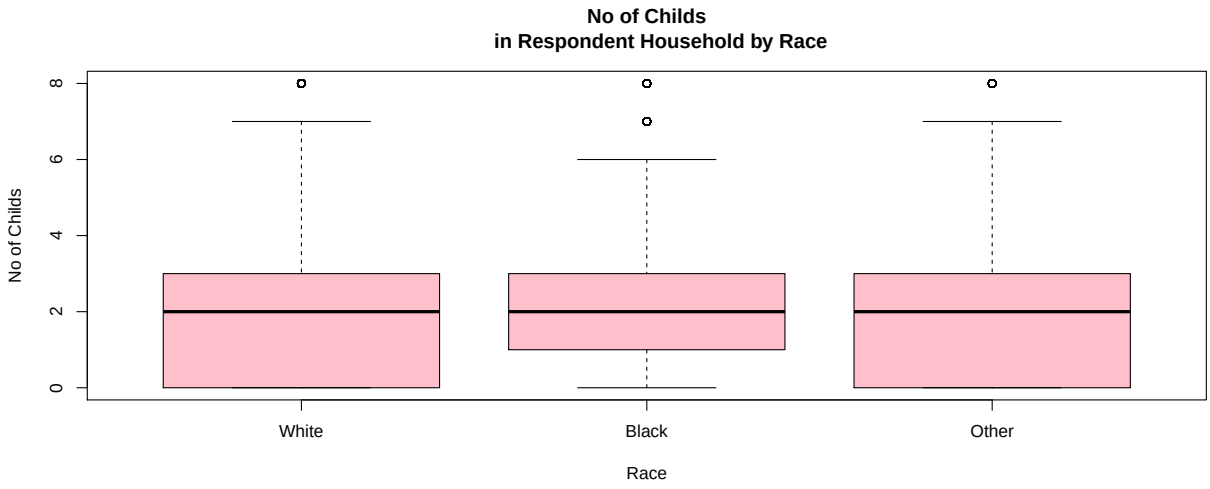
```
par(mfrow = c(1, 3))
plot(gss_subset$Race, main="Race of Respondent Households",
     xlab="Race", ylab="Count of Respondent Households", col="green")
hist(gss_subset$ChildCount, main="Number of Childs in  in Respondent Household",
     col="blue", xlab="Child Count", ylab="Count of Respondent Households")
boxplot(gss_subset$ChildCount, main = "No of Childs in Respondent
        Household", ylab = "No of Childs",col = "lightblue")
```



We observe that White race has maximum number of respondents while Other race has least. We also observe that number of childs in the respondent household is highly right skewed. We also observe that the median number of childs in the respondent household is two.

Now we analyze if there is a relationship between the two variables `race` and `childs` using boxplot and summary by `Race`.

```
plot(gss_subset$ChildCount ~ gss_subset$Race, main="No of Childs
     in Respondent Household by Race", xlab="Race", ylab="No of Childs", col="pink")
```




```
# Race: White
summary(gss_subset$ChildCount[gss_subset$Race == "White"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   2.000   1.805   3.000   8.000
```

```
# Race: Black
summary(gss_subset$ChildCount[gss_subset$Race == "Black"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   2.000   2.149   3.000   8.000
```

```
# Race: Other
summary(gss_subset$ChildCount[gss_subset$Race == "Other"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   2.000   1.884   3.000   8.000
```

We observe that while median value of no of childs is same across the three races, there is significant difference in the overall distribution and `Inter Quartile Range (IQR)` for number of childs across the three races. The IQRs for `White` and `Other` races are wider than the `IQR` for `Black` race.

Inference:

Hypothesis In this study we want to establish if there is a statistical significant difference between the mean number of children in the household by the race of the household.

The hypothesis for this study are as stated below:

- **Null Hypothesis (H_0)**: The mean number of childs in respondent household is the equal across all races of respondents.

$H_0 : \mu_{White} = \mu_{Black} = \mu_{Other}$

- **Alternative Hypothesis (H_A)**: At least one pair of mean number of childs in respondent household are different from each other.

H_A : *the average number of childs in respondent household* (μ_i) *varies across some* (*or all*) *races*

In statistical inference terms, we test a null hypothesis *(H_0)* where the mean number of childs in respondent household is equal for all races, and an alternative hypothesis *(H_A)* where at least one pair of means are different from each other.

Method to be Used We can test the hypothesis by performing a pairwise comparison of means across the various groups. But by doing so, we may find a difference just by chance even if there is no actual difference in the population.

So to address this, we will perform `Analysis of Variance (ANOVA)`, which uses single hypothesis test to check whether the mean number of children in the respondent household across the various races are equal.

`ANOVA` uses `F` test statistic, which represents a standardized ratio of variability in the sample means relative to the variability within the group. The larger the observed variability in the sample means relative to the within group observations, the larger the F will be and the stronger the evidence against the null hypothesis.

If the null hypothesis using `ANOVA` is rejected, then the results of pairwise comparison across the races are more significant.

Checking of Conditions `ANOVA` gives significant results if below three conditions are met:

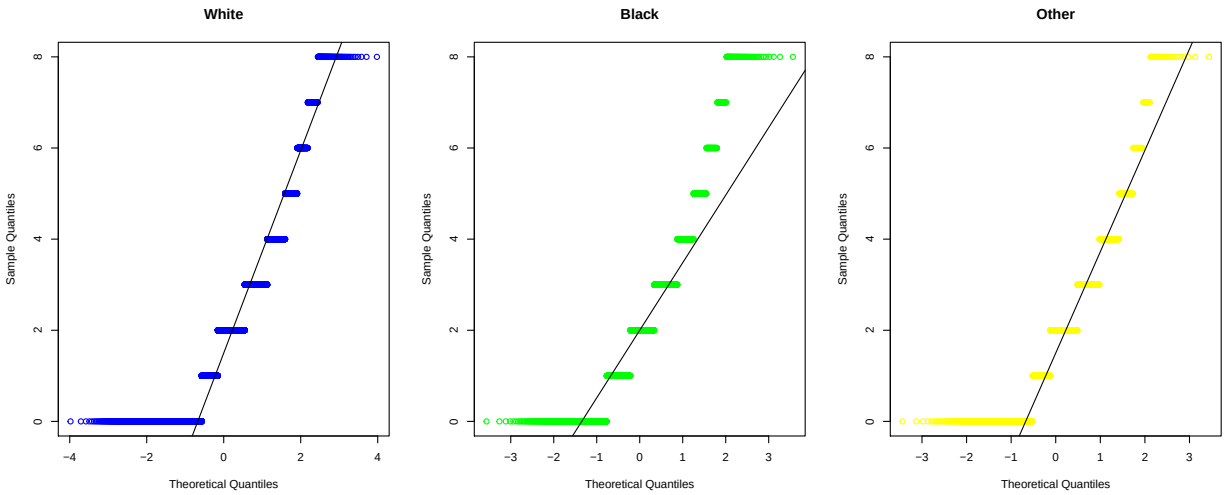
- Independence:

Within Groups - As the GSS data consists a random sample with sample size less than 10% of the population, hence the observations can be considered independent.

Between Groups - As the GSS data is taken randomly from respondents of different races with no linkage between themselves, hence the groups can be considered independent of each other or non-paired.

- Approximate Normality:

```
par(mfrow = c(1,3))
qqnorm(gss_subset$ChildCount[gss_subset$Race == "White"], main = "White", col="Blue")
qqline(gss_subset$ChildCount[gss_subset$Race == "White"])
qqnorm(gss_subset$ChildCount[gss_subset$Race == "Black"], main = "Black", col="Green")
qqline(gss_subset$ChildCount[gss_subset$Race == "Black"])
qqnorm(gss_subset$ChildCount[gss_subset$Race == "Other"], main = "Other", col="Yellow")
qqline(gss_subset$ChildCount[gss_subset$Race == "Other"])
```



From the normal probability plots for each group, we visualize the difference among observations distribution and standard distribution and see that there is some deviation from normality in each group.

- Constant Variance

From the box plot for `No of Childs in Respondent Household by Race`, we see that the total range and the interquantile range of the groups are different, with the lowest variability for `Black` race.

Hence we observe that the conditions for normality and constant variance are not fully met. To address this, a non-parametric test such as the `Kruskal-Wallis` test can be used. However as this is not covered under the class syllabus, we will proceed with the `ANOVA` analysis.

```
anova(lm(gss_subset$ChildCount ~ gss_subset$Race, data=gss_subset))
```

```
## Analysis of Variance Table
##
## Response: gss_subset$ChildCount
##                     Df Sum Sq Mean Sq F value    Pr(>F)
## gss_subset$Race      2    273 136.429  48.687 < 2.2e-16 ***
## Residuals        18895  52947   2.802
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation The `ANOVA` test reports a F test statistic of 48.687 and a p-value of approximately zero. This mean that the probability of observing a F value of 48.687 or higher, if the null hypothesis were true, is very low.

So we can reject the null hypothesis and we can say that the average number of child in the respondent household varies across some (or all) groups in a statistically significant way.

Since the null hypothesis has been rejected, we can do a pairwise comparison to find out which groups have different means.

For every possible pair of groups (3 pairs), we use a `t test` to confirm the null hypothesis that the means of the two groups are equal ($H_0 : \mu_{Group1} = \mu_{Group2}$) or the alternative hypothesis that they are different ($H_A : \mu_{Group1} \neq \mu_{Group2}$).

To avoid the increase of Type I error rate (rejecting a true null hypothesis), we apply a *Bonferroni* correction to the p-values which are multiplied by the number of comparison. With this correction, the difference of the means has to be bigger to reject the null hypothesis.

```
pairwise.t.test(gss_subset$ChildCount, gss_subset$Race,  p.adj="bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  gss_subset$ChildCount and gss_subset$Race
##
##       White   Black
## Black < 2e-16 -
## Other 0.19    9.6e-07
##
## P value adjustment method: bonferroni
```

We can see that for two group pairs *White-Black* and *Black-Other*, the p-value is lower than the significance level of 0.05 and so the null hypothesis is rejected for these group pairs. This means that the difference of the means of these two group pairs are statistically significant.

The null hypothesis is not rejected for the group pair *White-Other*. The difference of the means of this pair is not statistically significant and can be attributed to chance.

Conclusion:

The study establishes a positive correlation between the number of childs in respondent household and race of the respondent.

We tested our hypothesis with `ANOVA` and pairwise `t tests` and find out that the mean number of childs of the groups are significantly different from one another. The only exception being the group pair of *White-Other* race. It seems that the mean number of childs in household were not different for `White` and `Other` races.

As some of the conditions for the statistical inference methods used were not fully met, so we need to be cautious in interpreting the results and these results can not be considered as definitive.

Future research could address these shortcomings by analyzing the interaction of other variables and by using more sophisticated statistical techniques.

References:

The data set to be used for this project is General Social Survey Cumulative File, 1972-2012 Coursera Extract, which is modified for Data Analysis and Statistical Inference course (Duke University).

The data set can be downloaded from http://bit.ly/dasi_gss_data

Data 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

Persistent URL: http://doi.org/10.3886/ICPSR34802.v1

Other References The codebook is available at https://d396qusza40orc.cloudfront.net/statistics%2Fproject%2Fgss1.html lists all variables, the values they take, and the survey questions associated with them.

Appendix:

We list out the first 50 rows of the data frame gss_subset containing the columns `race` and `childs`

```
head(gss_subset, 50)
```

```
##     Race ChildCount
## 1  White          0
## 2  White          2
## 3  White          5
## 4  White          1
## 5  White          0
## 6  White          0
## 7  White          0
## 8  White          2
## 9  White          1
## 10 White          3
## 11 White          1
## 12 White          1
## 13 White          1
## 14 White          0
## 15 White          2
## 16 Black          4
## 17 White          1
## 18 White          0
## 19 White          1
## 20 Black          4
## 21 White          2
## 22 White          2
## 23 White          0
## 24 White          2
## 25 White          2
## 26 White          3
## 27 White          1
## 28 White          4
## 29 White          2
## 30 White          4
## 31 White          2
## 32 White          1
## 33 Black          5
## 34 White          0
## 35 White          0
## 36 Other          0
## 37 White          0
## 38 White          0
## 39 White          0
## 40 White          0
## 41 White          0
## 42 White          3
```

```
## 43 Black          2
## 44 White          0
## 45 White          0
## 46 White          1
## 47 White          3
## 48 White          2
## 49 White          3
## 50 White          1
```