Relationship between highest degree of US residents and their confidence in scientific community between 1990 and 2000

12/04/15

Introduction:

With the following project we would like to answer this simple question: is there a relationship between the highest degree of US residents and their confidence in scientific community between 1990 and 2000?

We believe this is a meaningful question because research, together with innovation and technology growth should be the keys to build a better future for any country. With this regard, the confidence in the scientific community is a facet that cannot be neglected.

We are going to test if the highest degree of US residents and their confidence in the scientific community are two independent variables.

Data:

Data are taken from the General Social Survey (GSS) website. This is a sociological survey used to collect data on demographic characteristics and attitudes of residents of the United States. The survey is conducted face-to-face with an in-person interview by the National Opinion Research Center at the University of Chicago, of adults (18+) in randomly selected households. The survey was conducted every year from 1972 to 1994 (except in 1979, 1981, and 1992). Since 1994, it has been conducted every other year. The data collected about this survey includes both demographic information and respondents' opinions on various matters.

The cases in the data are the respondents, i.e. resident of the United States. There are a total of 57061 cases in the survey. However, I will focus only on the time window (1990,2000) which will reduce the number of cases to 16039. The two variables I am going to study are the highest degree of the respondent and the confidence in the scientific community. They are referred to as "degree" and "consci" respectively. They are both categorical. Degree has 5 levels: "Lt High School", "High School", "Junior College", "Bachelor", "Graduate" "A Great Deal", "Only Some" and "Hardly Any" are the three level for the variable consci.

Since data are taken from a survey the study is an observational study. The sampling method is a stratified sampling with several stages of selection. The population was first stratified by region, age and race before selection.

The population of interest is the US residents. The findings can be generalized to the population since the data are random samples.

Potential source of bias:

- GSS selected only English and Spanish speakers. However, only a few percentage of the US population do not speak these two languages so we believe that this bias can be neglected.
- It is possible that a part of this sample has a degree or expect to get one in a scientific subject. It's hard to avoid this type of bias, therefore we will take it into account when we draw our conclusions.

Exploratory data analysis:

Let' first select a subset containing only data from 1990 to 2000

```
load(url("http://bit.ly/dasi_gss_data"))
gss_sub <- subset(gss, year >= 1990 & year <=2000, select = c(caseid, year, degree, consci))
```

Now let's have a look at the summary of the data.

```
degree <- gss_sub$degree
consci <- gss_sub$consci
summary(degree)
```

```
## Lt High School    High School Junior College       Bachelor       Graduate 
##           2546           8545           1023           2555           1149 
##           NA's 
##            222 
```

```
summary(consci)
```

```
## A Great Deal    Only Some   Hardly Any         NA's 
##         4197         4915          768         6160 
```

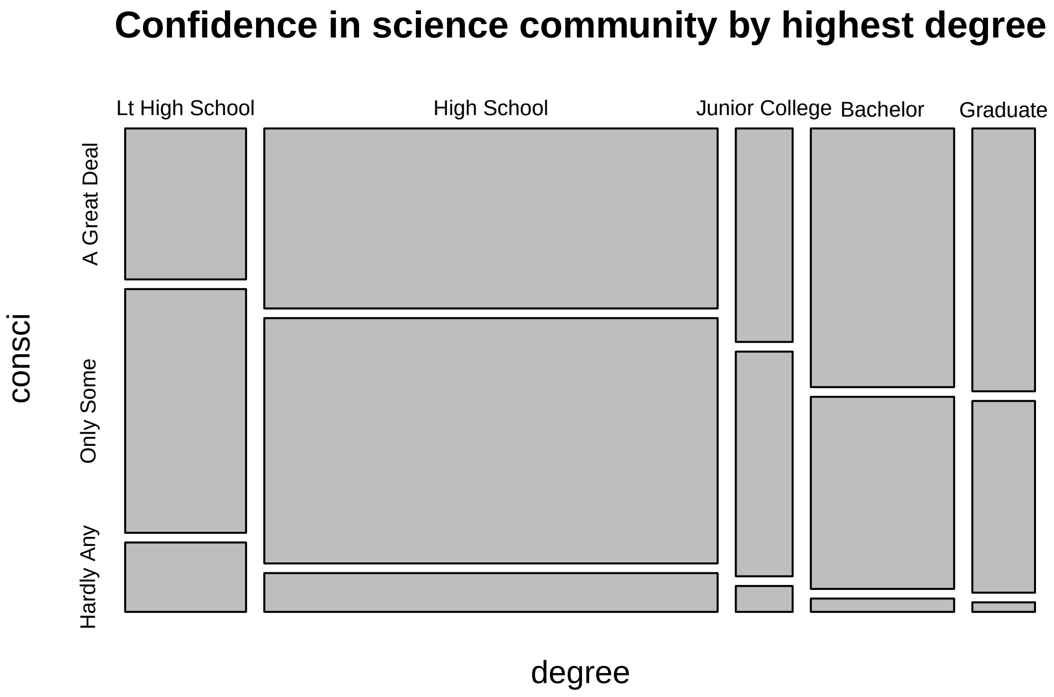
As you can see there are many NAs for both variables. We are going to remove them.

```
gss_sub <- gss_sub[complete.cases(gss_sub), ]
my_table <- table(degree, consci)
```

From the table, roughly the 24% of the US resident between 1990 and 2000 whose highest degree was High School have "A Great Deal" confidence on science. This percentage increases with the degree: 29% of US residents whose highest degree is Junior College, 36% of those with Bachelor and 36% of those Graduate.

A mosaic plot will help to have a better picture of the situation.

```
mosaicplot( degree ~ consci, data = gss_sub, main = "Confidence in science community by highest degree"
```



Inference:

The inference for this data set will be done via hypothesis testing. Since we are dealing with two categorical variables, both with more than two levels, we are going to use a chi-square independence test.

The two hypotheses are described below:

- Null hypothesis (H_0): Highest degree of US residents and their confidence in scientific community are independent.
- Alternative hypothesis (H_a): Highest degree of US residents and their confidence in scientific community are dependent. The confidence in the scientific community varies by highest degree.

To evaluate the hypotheses we are going to quantify how different the observed counts are from the expected counts. A large deviation from what would be expected based on sampling variation alone provide strong evidence for the alternative hypothesis.

Before going on, let's first check if the conditions for the chi-square test are met:

- Independence: the data were collected as random sample (see paragraph "Data") and the size of the sample is definitely less than 10% of the population.
- Sample size: each scenario has at least 5 expected cases.

Now, we can define the overall rates for each of the different highest degrees. First let's have a look at the sum of all the levels:

```
mt1 <- margin.table(my_table,1)
mt2<- margin.table(my_table,2)
mt1
```

```
## degree
## Lt High School    High School Junior College       Bachelor       Graduate
##           1411           5290            663           1672            731
```

```
mt2
```

```
## consci
## A Great Deal    Only Some   Hardly Any
##         4162         4852          753
```

```
lt_high_school_rate <- 1411/9767
lt_high_school_rate
```

```
## [1] 0.1444661
```

```
high_school_rate <- 5290/9767
high_school_rate
```

```
## [1] 0.5416197
```

```
jr_college_rate <- 663/9767
jr_college_rate
```

```
## [1] 0.06788164
```

```
bach_rate <- 1672/9767
bach_rate
```

```
## [1] 0.1711887
```

```
grad_rate <- 731/9767
grad_rate
```

```
## [1] 0.07484386
```

If the null hypothesis is true (highest degree and confidence in scientific community are independent), how many of the people with a confidence in scientific community "A Great Deal" we expect to have a less than high school degree? How many for those with a confidence "Only Some" and "Hardly Any"? And what about those US residents with a high school degree? And so on…

To obtain the expected number of US residents with a highest degree "less than high school" and a confidence in scientific community "A great deal", we take the total number of people with "A Great deal" confidence (4162) and multiply it by the overall rate of US residents with "less than high school" as highest degree (0.14).

```
my_table <- addmargins(my_table)

exp_great_deal_lt <- my_table["Sum", 1] * lt_high_school_rate
exp_great_deal_lt
```

```
## [1] 601.2677
```

Similarly, for the other values:

```
exp_great_deal_hs <- my_table["Sum",1] * high_school_rate
exp_great_deal_hs
```

```
## [1] 2254.221
```

```
exp_great_deal_jr <- my_table["Sum",1] * jr_college_rate
exp_great_deal_jr
```

```
## [1] 282.5234
```

```
exp_great_deal_ba <- my_table["Sum",1] * bach_rate
exp_great_deal_ba
```

```
## [1] 712.4874
```

```
exp_great_deal_gra <- my_table["Sum",1] * grad_rate
exp_great_deal_gra
```

```
## [1] 311.5002
```

```
exp_only_some_lt <- my_table["Sum",2] * lt_high_school_rate
exp_only_some_lt
```

```
## [1] 700.9493
```

```
exp_only_some_hs <- my_table["Sum",2] * high_school_rate
exp_only_some_hs
```

```
## [1] 2627.939
```

```
exp_only_some_jr <- my_table["Sum",2] * jr_college_rate
exp_only_some_jr
```

```
## [1] 329.3617
```

```
exp_only_some_ba <- my_table["Sum",2] * bach_rate
exp_only_some_ba
```

```
## [1] 830.6076
```

```
exp_only_some_gra <- my_table["Sum",2] * grad_rate
exp_only_some_gra
```

```
## [1] 363.1424
```

```
exp_hardly_any_lt <- my_table["Sum",3] * lt_high_school_rate
exp_hardly_any_lt
```

```
## [1] 108.7829
```

```
exp_hardly_any_hs <- my_table["Sum",3] * high_school_rate
exp_hardly_any_hs
```

```
## [1] 407.8397
```

```
exp_hardly_any_jr<- my_table["Sum",3] * jr_college_rate
exp_hardly_any_jr
```

```
## [1] 51.11488
```

```
exp_hardly_any_ba<- my_table["Sum",3] * bach_rate
exp_hardly_any_ba
```

```
## [1] 128.9051
```

```
exp_hardly_any_gra<- my_table["Sum",3]* grad_rate
exp_hardly_any_gra
```

```
## [1] 56.35743
```

The significance level of our test is going to be 5%. To compute the chi-square statistics we need to subtract the expected value to the observed value of each row, square the value, and divide by the expected. The number of degrees of freedom is given by (R-1) x (C-1) where R and C are the numbers of rows and columns of "my_table" respectively.

```
chi1 <- (((459-601)**2 )/601)+(((741-701)**2 )/701)+(((211-109)**2) / 109)+(((2052-2254)**2 )/2254)
chi2 <- (((2796-2628)**2 )/2628)+(((442-408)**2 )/408)+(((414-311)**2 )/ 311)
chi3 <- (((302-363)**2 )/363)+(((15-56)**2 )/56)

chi <- chi1 + chi2 + chi3
chi
```

```
## [1] 237.3398
```

```
df = (3-1) * (5-1)
```

The p-value can be easily computed using the function pchisq in R.

```
pchisq(chi,df,lower.tail=FALSE)
```

```
## [1] 8.2834e-47
```

The p-value is extremely small, and since it is also smaller than our significance level we reject the null hypothesis.

Conclusion:

These data provide convincing evidence that the highest degree of US residents and their confidence in scientific community are associated. As we already mentioned, we cannot extend this result to the entire US population since this is an observational study. It is still possible that there is a casual relationship between these two variables, however with this type of study we cannot be certain. We need also to consider possible source of bias such as a degree in a scientific subject.

Appendix:

print(gss_sub[1:60,])