Statistical inference with the GSS data

Setup

Load packages

```
library(ggplot2)
library(dplyr)
library(statsr)
library(tidyr)
library(knitr)
library(forcats)
```

Load data

Make sure your data and R Markdown files are in the same directory. When loaded your data file will be called `gss`. Delete this note when before you submit your work.

```
load("gss.Rdata")
```

Part 1: Data

What is this data about?

- The GSS survey has been conducted since the year 1972.
- Since 1972, the General Social Survey (GSS) 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;.
- GSS questions cover a diverse range of issues including national spending priorities, marijuana use, crime and punishment, race relations, quality of life, confidence in institutions, and sexual behavior.

Notable changes to the survey

GSS frequency of survey:

- The survey used to be conducted annual from 1972-1994. Due to changes in funding, The surves has then been conducted bi-annually from 1996, 1998,2002 and so on, till date.
- The method of taking the survey has evolved considerably over time. And currently circa 2020:

GSS Administration:

- The vast majority of GSS data is obtained in face-to-face interviews.
- Computer-assisted personal interviewing (CAPI) began in the 2002 GSS.
- Under some conditions when it has proved difficult to arrange an in-person interview with a sampled respondent, GSS interviews may be conducted by telephone.

Structure of the survey?

- The items appearing on the survey are of 2 types:
- Permanent questions that occur in each and every survey
- Rotating questions that appear in 2 out of 3 surveys.

Sampling used in the GSS:

- Around 1500 people were sampled annually from 1972 - 1992. Since 1994 two-samples of size 1500 each are surveyed. However, the survey is conducted once every 2 years. There has been no oversampling done since 1994.
- Modified probablilty sampling used from 1972-1974 (random sampling).
- At the top most level, PSU (Primary Sampling Units) are: Metropolitan and non-Metropolitan areas. These are stratified, based on region, age and race before selection.
- Within the PSU, they are broken down into :Block Groups and Enumeration Districts. the BG and ED were stratified according to race and income before selection.
- The third stage of selection is the blocks, they had probablites assigned based on block size before the selection.
- Inside the selected blocks, the average cluster size is 5 respondents per cluster.
- So we have a sampling approach that uses: Stratified sampling at the PSU level -> Stratifies sampling at the BG and ED level -> selection of blocks by probablitly based on size -> finally we select clusters with (5 respondents) in a cluster.

Generalizabilty concerns with the GSS survey:

- The survey has been restricted only to the English speaking population of the US from: 1972- 2004.
- From 2006 spanish speakers were included in the survey.
- So , if a person knew neither English nor Spanish, at the moment, it looks like the survey does not cover those demographics. This seems to present a blind spot, that makes the survey not inclusive of the whole US population.

Answer :

- **Sampling** : Random sampling is used. It is organized in the following way: Stratified sampling at the PSU level -> Stratifies sampling at the BG and ED level -> selection of blocks by probablitly based on size -> finally we select clusters with (5 respondents **randomly selected**) in a cluster.
- **Generalizability** : Since the data is collected using a survey of randomly selected English speaking and Spanish speaking individuals from stratified sample at the PSU -> BG level and finally cluster samping at the lowest level. across the US. We can safely say the survey is generalizable **only** across the English/Spanish speaking US population.
- **Causality** : The survey is **NOT** useful for causality. As we have not randomly selected and randomly assigned individuals into experimental groups to draw any causual relationships among the variable of the data in which we seek to establish causality.
- **Potential biases**: There is a chance of participation bias , due to non-participation of randomly selected individuals, non-participation of Non-English and Non-Spanish speakers. Volunteer bias due to certain segments of individuals in each strata more willing to partiipate than other individuals.

In conclusion the survey is **Generalizable** across the English/Spanish US population. Any patterns we observe are **Associative** and we need further controlled experiments to establish **Causality**. There is a possibility of **Participation bias** as well as **Non-response** bias to be introduced into the survey.

Part 2: Research question

Is there a relationship between gun-ownership and race ?

1) I am really interested in exploring the relationship between gun-ownership and race.
2) **Warning :** Anything from this study can only be considered to be **Associative** in nature.
3) This research question is interesting, as it can serve as a launching pad to delve deeper and analyze other factors round the **"why"** of various factors that contribute to prevelance of guns in certain races over others.

It can open up our understanding into underlying factors or even discover whether there are **confounding** variables that might lead to any observed associative relationships between Race and Guns.

Part 3: Exploratory data analysis

Step1 :

Selecting the variables of interest:

We begin the analysis by selecting the following 2 variables from our dataset:

- owngun: Contains Respondent's answer to the question: *Do you happen to have in your home (or garage) any guns or revolvers?*
- race: Contains Respondent's answer to the question: *What race do you consider yourself?*

```
gunsandrace <- gss%>%select(c("race","owngun"))
str(gunsandrace)
```

```
## 'data.frame':    57061 obs. of  2 variables:
##  $ race  : Factor w/ 3 levels "White","Black",..: 1 1 1 1 1 1 1 1 2 2 ...
##  $ owngun: Factor w/ 3 levels "Yes","No","Refused": NA NA NA NA NA NA NA NA NA NA ...
```

Step2 :

Create a Summary Statistic table:

We start to organize the data into a wide format, by grouping the data bases on:

- OwnGun as the row of the table.
- Race as the column of the table.

```
contigencyTable <- gunsandrace%>%group_by(race)%>%count(owngun)
```

```
## Warning: Factor `owngun` contains implicit NA, consider using
## `forcats::fct_explicit_na`
```

```
# Our data is in a long format. In order to generate the contigency table of interest and for easy visu
# We are going to transform the data into a wide format:  Race will form the columns, and owngunss will

contigencyTable_WF <- contigencyTable%>%spread(race,n)%>%mutate(Column_Total =  rowSums(.[2:4]))%>%muta
contigencyTable_WF$owngun <- fct_explicit_na(contigencyTable_WF$owngun,na_level = "Not Answered")
kable(contigencyTable_WF)
```

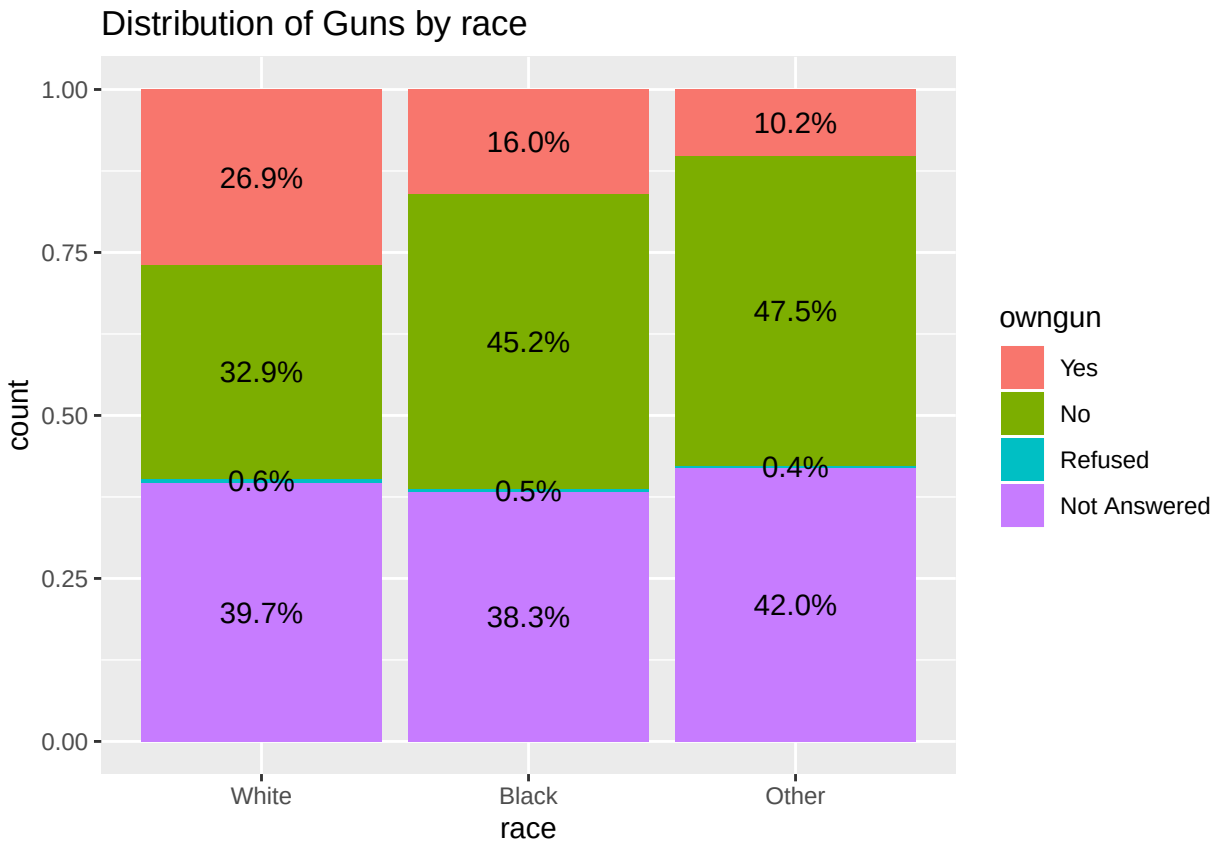
owngun	White	Black	Other	Column_Total	Row_Total
Yes	12448	1269	283	14000	46350
No	15235	3586	1323	20144	7926
Refused	266	39	10	315	2785
Not Answered	18401	3032	1169	22602	57061

- owngun: CATEGORICAL variable contain 4 levels: (Yes, No, Refused, Not Answered)
- Race: CATEGORICAL variable has been spread out in a wide format as: WHITE, BLACK and OTHER
- Column_Total: Contains the Row sum across all 3 race categories for each row.
- Row_Total: Contains the sum of all items in each column.
- Note: Not Answered is NOT ignored as this is a huge %, it serves as it's own level in owngun

Step3 :

Create visual representation of Guns vs race:

```
## # A tibble: 12 x 4
## # Groups:   race [3]
##    race  owngun           n ratio
##    <fct> <fct>        <int> <chr>
##  1 White Yes          12448 26.9%
##  2 White No           15235 32.9%
##  3 White Refused        266 0.6%
##  4 White Not Answered 18401 39.7%
##  5 Black Yes           1269 16.0%
##  6 Black No            3586 45.2%
##  7 Black Refused         39 0.5%
##  8 Black Not Answered  3032 38.3%
##  9 Other Yes            283 10.2%
## 10 Other No            1323 47.5%
## 11 Other Refused         10 0.4%
## 12 Other Not Answered  1169 42.0%
```



- By looking at the graphs, we can interpret that there seems to be variation across the races in terms of owning a gun. For example: The percentage of whites vs blacks who declare owning a gun drops from **26.9% to 16.0%.**
- Similarly the variation amongst the other categories of No, Refused and Not Answered also show a difference amongst the races.
- We see across the different race categories that , there indeed is a difference in percentage, of those who answer: Yes, No, Refuse, Not Answered for the question on guns.

- We need to establish, if the variations across races we see are purely due to chance in the sample, or if there really is a race to gun associavity.

Part 4: Inference

Step0 : Hypothesis

H0: There is no difference between GunOwnership amongst the Races.

HA: There is a difference between GunOwnership amgonst Race.

Step1 :

Decide on the inference methodology:

- Since we are dealing with 2 categorical variables: owngun - race.
- Atleast 1 or both categorical variables have more than 2 levels.
- We have decided to use the: Chi-Square test of Independence.

Step2 :

Check conditions to apply Chi-Square test of independence:

Independence:

- Since the respondents for our data have been randomly selected from the population. We can assume that they are independent of each other.
- The total number of respondents in our sample is **57061**. This is definitely less than 10% of the US population. Since our survey is done without replacement of respondents. We can safely assume that this condition for independence has also been met.
- Each case should contribute to only one cell inside the table: We are able to ensure this condition holds, because, each respondent falls into only one of the ownguns category and one race category.

Sample size:

- Each cell should have an expected values of atleast 5.
- In order to ensure this condition is met, we need to generate the expected values for these observations.

Step3 :

Create Expected Cases:

We begin by reprinting our summary statistic table:

```
kable(contigencyTable_WF)
```

owngun	White	Black	Other	Column_Total	Row_Total
Yes	12448	1269	283	14000	46350
No	15235	3586	1323	20144	7926
Refused	266	39	10	315	2785
Not Answered	18401	3032	1169	22602	57061

- To calculate the expected cases, we first need to calculate the total numbers in each of the 4 owngun categorical variable: YES/No/Refused/Not Available.
- Luckily, we already calculated this data and it is available in our **Column_Total**, Column in the table.
- We also need the Total of each column, as well as the Total sample size, both of these are available in out **Row_Total** column.
- Finally, to calculate the Expected Cases, we use the following formula:

Expected Cases = Row Total X Column Total / Total Observations

- For Example:
- In order to calculate the Expected_White cases across all owngunns levels: Yes, No, Refused, Not Answered.

Formula:

- Row_Total(White) X Column_Total(Yes/No/Reserved/Not Answered)
- **Row_Total(White):** 46350
- **Column_Total for Yes:** 14000,
- **Column_Total for No:** 20144,
- **Column_Total for Reserved:** 315,
- **Column_Total for Not Answered:** 22602

We have to repeat the above steps for: Black and Other in the following manner:

- Row_Total(Black) X Column_Total(Yes/No/Reserved/Not Answered)
- Row_Total(Other) X Column_Total(Yes/No/Reserved/Not Answered)

Applying this we are able to calculate the expected values, and add it to our table.

```
contigencyTable_WF <- contigencyTable_WF%>%mutate(Expected_White = (Row_Total[1]*Column_Total[1:4])/570
contigencyTable_WF$Expected_White <- round(contigencyTable_WF$Expected_White,digits = 0)
contigencyTable_WF$Expected_Black <- round(contigencyTable_WF$Expected_Black,digits = 0)
contigencyTable_WF$Expected_Other <- round(contigencyTable_WF$Expected_Other,digits = 0)
kable(contigencyTable_WF)
```

owngun	White	Black	Other	Column_Total	Row_Total	Expected_White	Expected_Black	Expected_
Yes	12448	1269	283	14000	46350	11372	1945	
No	15235	3586	1323	20144	7926	16363	2798	
Refused	266	39	10	315	2785	256	44	
Not Answered	18401	3032	1169	22602	57061	18359	3140	

- Expected_White: Gives the Expected number of cases for Whites, if there is no difference across races.
- Expected_Black: Gives the Expected number of cases for Blacks, if there is no difference across races.
- Expected_Other: Gives the Expected number of cases for Other, if there is no difference across races.

We see that all the expected values in all the cells are atleast 5 or greater

Therefore we can apply the Chi-Square Test of Independence.

Step4 :

Apply Chi-Square test of Independence:

Hypothesis:

H0: There is no difference between GunOwnership and Race.

HA: There is a difference between GunOwnership and Race.

Under the assumption that the null hyopthesis H0 is true. We use the following formula to calculate the Chi-Square value:

$$\sum \frac{(ObservedValue - ExpectedValue) * (ObservedValue - ExpectedValue)}{ExpectedValue}$$

We now apply this formula to by first calculating the individual : Observed-Expected/Expected values. first and add them to our table.

```
contigencyTable_WF$ChiWhite <- ((contigencyTable_WF$White[1:4]-contigencyTable_WF$Expected_White[1:4])^
contigencyTable_WF$ChiBlack <- ((contigencyTable_WF$Black[1:4]-contigencyTable_WF$Expected_Black[1:4])^
contigencyTable_WF$ChiOther <- ((contigencyTable_WF$Other[1:4]-contigencyTable_WF$Expected_Other[1:4])^
kable(contigencyTable_WF)
```

owngun	White	Black	Other	Column_Total	Row_Total	Expected_White	Expected_Black	Expected_
Yes	12448	1269	283	14000	46350	11372	1945	
No	15235	3586	1323	20144	7926	16363	2798	
Refused	266	39	10	315	2785	256	44	
Not Answered	18401	3032	1169	22602	57061	18359	3140	

We finally sum up all the newly calculated Chi-values to get the final Chi-square value.

```
chisquare <- as.double(contigencyTable_WF%>%summarise(ChiSquare = sum(sum(ChiWhite),sum(ChiBlack),sum(C
```

The calculated chi-square value is: **998.6877495**

We now use the R function **pchisq**

Our degrees of freedom are: (Number of Row Observation -1) X (Number of Column Column - 1)

```
dimensionsTable <- gunsandrace%>%select(c(race,owngun))
C1 <- length(levels(dimensionsTable$owngun))
C2 <- length(levels(dimensionsTable$race))
df <- (C1-1)*(C2-1)
```

The degrees of freedom is: 6

Plugging in all the values into the R function:

```
pvalue <- pchisq(chisquare,df,lower.tail = FALSE)
```

We get a p-value : $1.718777 \times 10^{-212}$

This p-value is really really tiny compared to our significance level of 5%.

Bonus : Re-check with MASS library

```
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
input_table <- table(dimensionsTable$owngun,dimensionsTable$race)
chisq.test(input_table)
```

```
##
##  Pearson's Chi-squared test
##
## data:  input_table
## X-squared = 998.62, df = 6, p-value < 2.2e-16
```

Note: Since we are usin Chi-Square test for independence, we are not going to calculate confidence intervals. As in this test we compare more than 2 categories. And not just 2 level proportions

Conclusion

- The p-value : $1.718777 \times 10^{-212}$ is smaller than 5% significance level.
- Therefore we **reject the null hypothesis H0**, in favour of the alternative HA.
- **We reject H0: There is no difference between GunOwnership and Race.**
- The alternative that OwningGuns and Race are dependent on each other , **the alternative HA is accepted.**

- Note: Just because owningGuns and Race are dependent, we cannot establish causality between them. It does open more interesting lines of thinking that maybe there are confounder variables like:income differences between races, belief.cultural differences between races that might have caused this **Associative** relationship.
- This has opened up more interesting questions that need to be explored further.