Statistical inference with the GSS data

Setup

Load packages

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.3
```

```
library(statsr)
```

```
## Warning: package 'statsr' was built under R version 4.0.3
```

Load data

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

Part 1: Data

Background Of The Data

The GSS gathers data on contemporary American society in order to monitor and explain trends and constants in attitudes, behaviors, and attributes. Hundreds of trends have been tracked since 1972. In addition, since the GSS adopted questions from earlier surveys, trends can be followed for up to 70 years.

The GSS contains a standard core of demographic, behavioral, and attitudinal questions, plus topics of special interest. Among the topics covered are civil liberties, crime and violence, intergroup tolerance, morality, national spending priorities, psychological well-being, social mobility, and stress and traumatic events.

Altogether the GSS is the single best source for sociological and attitudinal trend data covering the United States. It allows researchers to examine the structure and functioning of society in general as well as the role played by relevant subgroups and to compare the United States to other nations.

The GSS aims to make high-quality data easily accessible to scholars, students, policy makers, and others, with minimal cost and waiting.

Method Of Collection Of Data

The GSS is conducted as an in-person interview by the team at National Opinion Research Center (NORC) - hosted at the University of Chicago - and targets adults living in the United States. The GSS sample is drawn using an area probability design that randomly selects respondents in households across the United States to take part in the survey. Respondents that become part of the GSS sample are from a mix of urban, suburban, and rural geographic areas. Participation in the study is strictly voluntary.

The survey is conducted face-to-face with an in-person interview by NORC at the University of Chicago. 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 survey takes about 90 minutes to administer. In this data set we have responses from 57,061 individuals for 114 variables.

Generalizability, Causality and Bias

Within a certain set of constraints that may introduce bias (see below) the GSS performs random sampling, so as to make it broadly generalizable to the US population.

The GSS is, however, an observational study - with no explicit random assignments to treatments - so all relationships indicated may indication association, but not causation.

As it relates to bias, there are a few potential concerns:

i. First, the changes in methodologies over the years may introduce bias. As an example, it was not until 2006 that Spanish-speaking adults were included in the survey.
ii. By using a voluntary in-person survey that takes approximately one hour, there is the possibility of under-reporting those that choose to not respond to the survey.
iii. Since the answers to the interview questions are not validated, respondents may alter their responses in a variety of ways, such as expressing desirable behaviors and traits, while under-reporting undesirable ones.

But, since there is also non-response sub-sampling which would help in reducing the non-response bias. Reducing this bias assures the statistical significance and makes the chosen samples more representative of the entire US population.

Part 2: Research question

We are interested in investigating whether there is any association between "Years served in armed forces" and "Attitude towards foreign aid".

It is a general belief that the training which the candidates have to go through, before joining the armed forces usually encourages hatred towards general citizens of other nations, thus foreign aid seems to be an unimportant matter to them. To test the truthfulness of this belief, we need to perform inferential procedures.

The variables of interest here are 'vetyears' and 'nataid'

Part 3: Exploratory data analysis

Before moving on for the inferential procedures, we must clean the data, that is, remove those entries which have missing values for our variables of interest.

```
stats <- gss %>% filter(!is.na(vetyears) & !is.na(nataid)) %>% select(vetyears, nataid)
```

Now, we must perform some exploratory data analysis to get a preliminary idea of the data set we are working with (factors of the variables of interest, general trend, etc.).

```
summary(stats)
```

```
##             vetyears                nataid
##  None            :11031   Too Little :  790
##  Less Than 2 Yrs :  591   About Right: 3268
##  2 To 4 Years    : 1635   Too Much   :10728
##  More Than 4 Yrs :  624
##  Some,Dk How Long:    5
```

We should remove the observations of the respondents who are not sure about the length of their services in the Armed forces, in order to avoid ambiguous inferential results.

```
stats <- stats %>% filter(vetyears != "Some,Dk How Long")
```

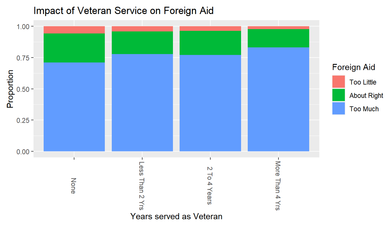
Now, we must create a contingency table with the different factors of Veteran years along the row and different factors of attitude towards Foreign Aid along the columns. (Note that we have removed all the observations who aren't sure about the length of their services as a Veteran)

```
table_data <- table(stats$vetyears, stats$nataid)
table_data
```

```
##
##                  Too Little About Right Too Much
##   None                  692        2755     8484
##   Less Than 2 Yrs        24         107      460
##   2 To 4 Years           60         313     1262
##   More Than 4 Yrs        14          92      518
##   Some,Dk How Long        0           0        0
```

Now, let us plot the poportion of these factors.

```
ggplot(stats) + geom_bar(aes(x=vetyears, fill= nataid), position="fill")+labs(x="Years served as Veteran", y="Propo
```



It is already pretty evident that there is some observable difference in the proportion of those who think that the Foreign aid is too little. We see that as the experience in Armed Forces is increasing, we see a dip in the proportion of respondents who believe that the Foreign aid is not enough, indicating the truthfulness of the general idea.

Part 4: Inference

While performing Exploratory Data Analysis(EDA), we report observable difference in the proportion, but is the difference significant? For that we have to call for Inferential procedures. Our objective is to check whether the Factors in study (Years served in Armed Forces and Attitude towards foreign aid) are independent or not.

Hypothesis:-

$$H_0 : \text{Years of veteran service is independent of Attitude towards Foreign aid}$$
$$vs\ H_1 : \text{Attitude towards Foreign aid is dependent on Years of veteran service}$$

Method:-

Here we have two categorical variables (vetyears and nataid), both having more than two factors. Thus in this scenario, the Chi-square test for independence is suitable in this case.

Conditions For The Test:-

The key conditions for the chi-square test of independence are as follows,

i. Independence of observations: This is assumed to be true because of the sampling methodology used by the GSS. Furthermore, the sample size is less than 10% of the population and each observation is recorded in a single cell.
ii. Cell count: As we can see from the contingency table below, all the cell counts are greater than 5, so there wouldn't be any issues with the continuity.

Note: We have removed the observations where the respondents are not sure about the term of service as a Veteran.

```
table_data
```

```
##
##                  Too Little About Right Too Much
##   None                  692        2755     8484
##   Less Than 2 Yrs        24         107      460
##   2 To 4 Years           60         313     1262
##   More Than 4 Yrs        14          92      518
##   Some,Dk How Long        0           0        0
```

Now, let us move on to the conducting of the test.

```
chisq.test(stats$vetyears, stats$nataid)
```

```
##
##  Pearson's Chi-squared test
##
## data:  stats$vetyears and stats$nataid
## X-squared = 77.523, df = 6, p-value = 1.159e-14
```

Here we have 4 levels of the Factor: Years in Veteran service and 3 levels of the Factor: Attitude towards Foreign Aid. Thus the degrees of freedom (df) for our test would be,

$$df = (4-1) \times (3-1) = 6$$

Conclusion and Interpretation:-

We have a very high value of the test-statistic and thus a very low p-value. Thus we reject the null hypothesis with a strong evidence. So, the claim of independence of the factors in consideration is false.

We can thus conclude that tenure of service in Armed Forces really has a significant effect on the attitude towards Foreign Aid. In fact, from the Exploratory Data Analysis we can say that there is significance evidence of the negative relationship between the tenure in armed forces and sympathy towards foreign nation.
