Statistical inference with the GSS data

kimnewzealand

29 June 2017

Setup

Load packages

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

Load data

```
setwd("~/InfStatsProject")
load("gss.Rdata")
```

Part 1: Data

BACKGROUND

As per the reference documentation http://gss.norc.org/ the data is an extract of the General Social Survey (GSS) Cumulative File 1972-2012 with the following modifications:

- All missing values have been removed from this data.
- Factor variables have been created on categorical variables.

This survey monitors societal change and monitors and explains trends and constants in attitudes, behaviors, and attributes in the United States and is conducted every two years through a 1:1 in person interview, from a random sample with all households from across the country having an equal chance of being selected, with 5,000 invited to respond to each survey.

SAMPLING METHOD

The sampling method is a simple random sample, selected randomly by household address with each having an equal probably of selection. The within sample responses can be assumed to be independent and less than 10% of the US population, which is in millions.

We can take a look at the actual number of respondents each year by looking at a summary table of the *year* variable against the calculated response rate (%).

```
yearsum <- summary(gss %>%
          group_by(year) %>%
          summarise(response.rate=n()/5000*100))
kable(yearsum, caption="Summary of Response Rate", align="c")
```

Table 1: Summary of Response Rate

year	response.rate
Min. :1972	Min. :27.44
1st Qu.:1980	1st Qu.:29.98
Median :1988	Median :31.98
Mean :1989	Mean :39.35
3rd Qu.:1998	3rd Qu.:40.88
Max. :2012	Max. :90.20

From this table summary the sample sizes varies from year to year. Since there are less responses than the 5,000 selected each year, there could be some voluntary or non-response bias in the survey.

The sample may be generalizable to the population if the random samples are a representation of the population.

This is an observational study, not an experimental study with random assignment, so it should not be used to derive causal statements.

SUMMARY OF DATASET

The gss.Rdata dataset contains 57061 observations and 114 variables in a long format.

The identifier variable for each survey is:
- *caseid: case identification number*

Part 2: Research question

We will explore any relationships confidence in the press and the news reported in media channels, such as newspapers and TV, in recent years.

This is topical at the moment with the recent elections in the US, France and the UK, where news is also being distributed through new channels, such as social media channels.

Part 3: Exploratory data analysis

Let's take a look at the relationship between the two variables:
- *year*: GSS year for this respondent
- *conpress*: confidence in press

```
# Produce summary statistics of the year and conpress variables
kable(t(as.matrix(summary(gss$year))),format='pandoc')
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1972	1983	1993	1992	2002	2012

```
kable(t(as.matrix(summary(gss$conpress))),format='pandoc')
```

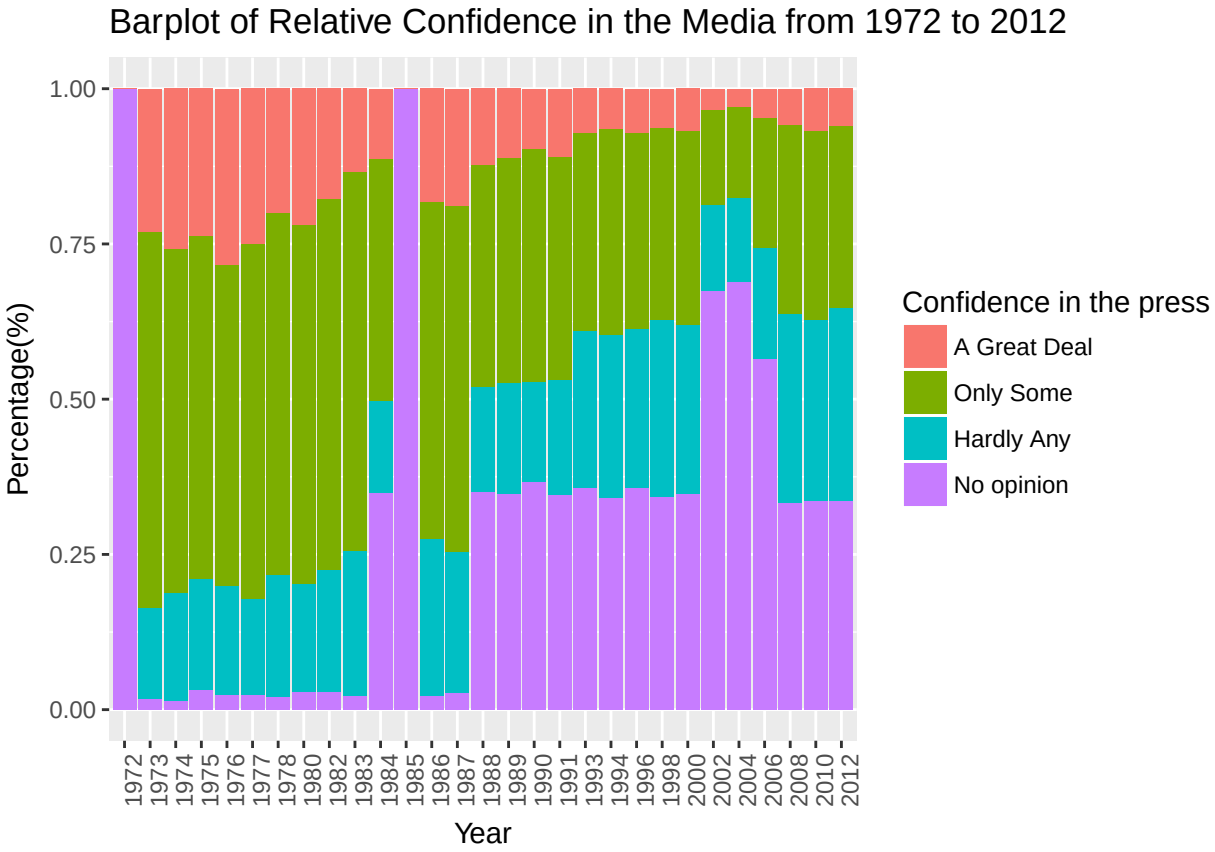
A Great Deal	Only Some	Hardly Any	NA's
6128	20346	11465	19122

A Great Deal	Only Some	Hardly Any	NA's

Although the *year* is an integer we will treat it as a categorical variable. The *conpress* is a factor with 3 levels and 34% NA's. As mentioned, all missing values have been recoded to missing (NA) in the dataset.

With two categorical variables so we will view a stacked barplot with percentages to view the relative values between years.

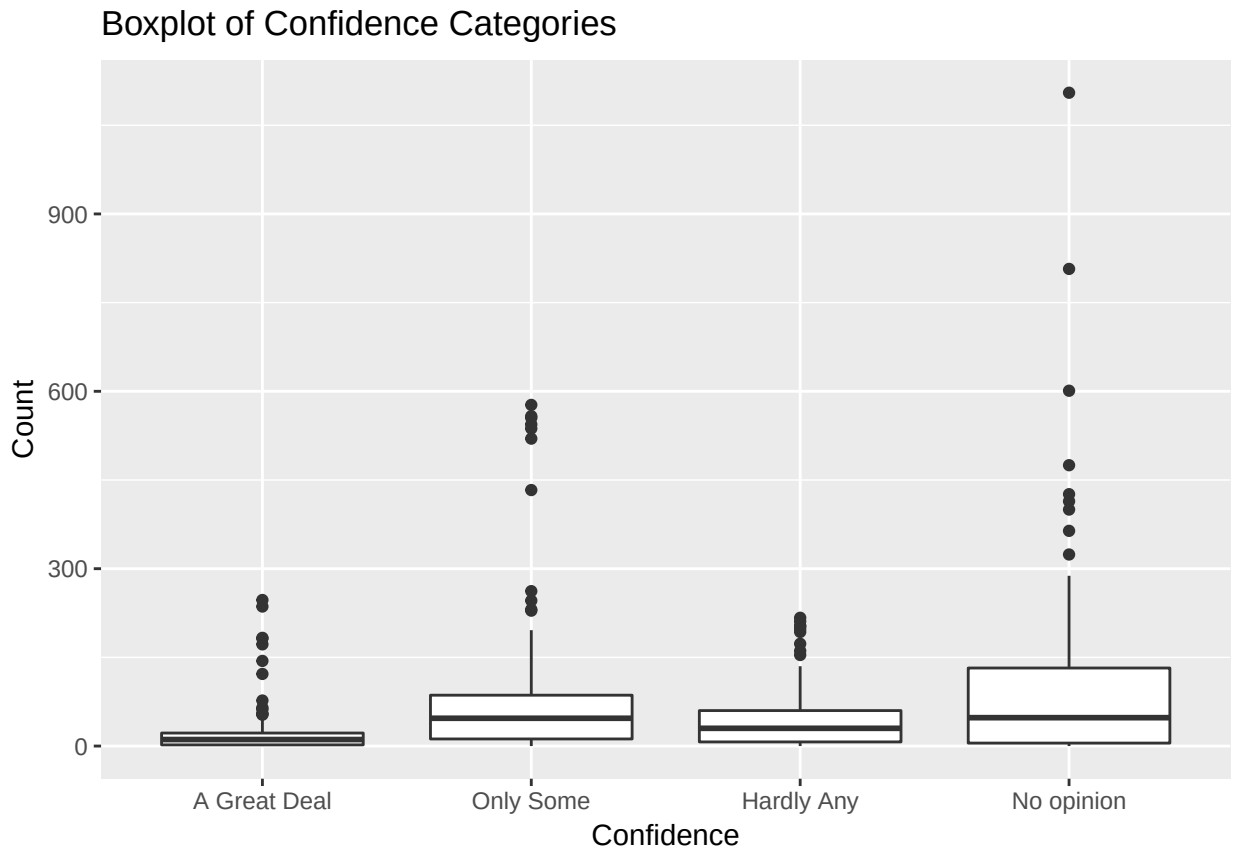
```
# Get levels and recode NA to "No opinion" as this variable has NA documented as such.
levels <- levels(gss$conpress)
levels[length(gss$conpress) + 1] <- "No opinion"

# Refactorconpress to include "No opinion" as a factor level
# and replace NA with "None"
gss$conpress <- factor(gss$conpress, levels = levels)
gss$conpress[is.na(gss$conpress)] <- "No opinion"

# Calculate a relative frequency table with table and prop.table and table functions
yearpress<- prop.table(table(gss$year,gss$conpress,useNA = "ifany"))
yearpress <- as.data.frame(yearpress)
names(yearpress) <- c("year","confidence","percentage")

# Plot using ggplot, plot a barplot
g <- ggplot(yearpress,aes(year,percentage))
g + geom_bar(stat="identity",position = position_fill(),aes(fill=confidence)) +
  theme(axis.text.x = element_text(angle=90)) +
  labs(title="Barplot of Relative Confidence in the Media from 1972 to 2012", x = "Year", y = "Percentag
  scale_fill_discrete("Confidence in the press")
```



Based on this summary it appears that the category "A Great deal" and "Only some" have decreased and that the category of "Hardly any" and "No opinion", has increased as a proportion over time.

We can also take a look at the confidence categories as a boxplot.

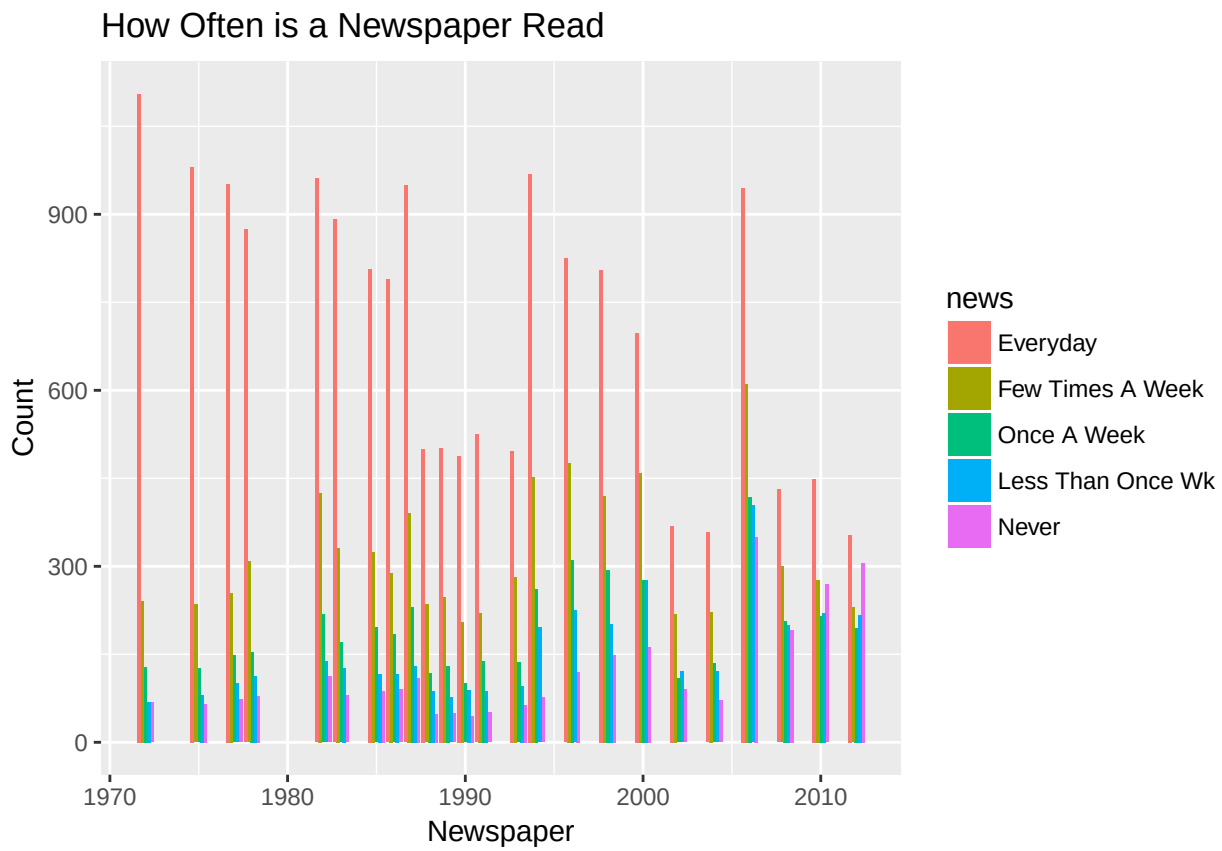
```
# Calculate a frequency table for year and conpress and news
yearpresstable <- table(gss$year,gss$conpress,gss$news)
yearpresstable <- as.data.frame(yearpresstable)
names(yearpresstable) <- c("year","confidence","newspaper","count")

# Plot boxplot of confidence categories
g <- ggplot(yearpresstable,aes(confidence,count))
g + geom_boxplot() +
  labs(title="Boxplot of Confidence Categories", x = "Confidence", y = "Count")
```



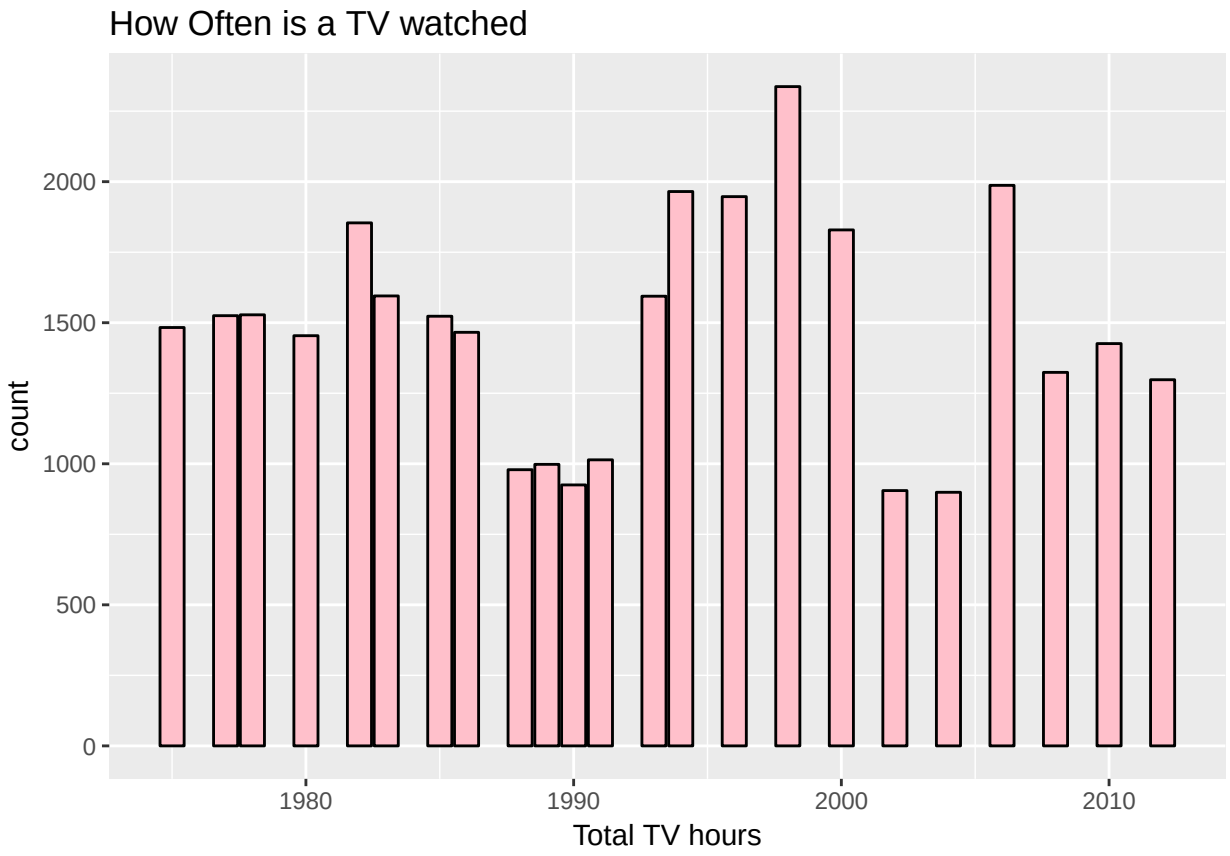
The variability is not constant across the year groups, and as we noted previously from the response rates, these year groups different sample sizes. By not meeting this constant variability condition, we would not be able to test means between the year groups using ANOVA.

Next let's review the media related variables visually using plots:
- *news*: How often does R read newspaper
- *tvhours*: Hours per day watching TV

```
# View the newspaper reading trends variable over time, removing missing value NA's from the plot
g <- ggplot(gss[!is.na(gss$news), ],aes(year))
g + geom_bar(aes(fill=news),position="dodge",na.rm=FALSE) +
  labs(title="How Often is a Newspaper Read",x= "Newspaper",y="Count")
```




```
# View the  tvhours trends over time, removing any missing value NAs
g <- ggplot(gss[!is.na(gss$tvhours), ],aes(year))
g + geom_bar(aes(fill=tvhours),fill="pink",colour="black",position="dodge") +
     labs(title="How Often is a TV watched",x= "Total TV hours",Y="Count")
```



These plots have similar patterns of decreasing over time, rising again in the 90's and both TV hours watched and newspapers read peaking in 2006, but decreasing to 2012.

Part 4: Inference

INFERENCE METHOD

We can look to perform a test using proportions to see if there is a relationship between confidence in the press *conpress* and how often a newspaper is read *news* for the most recent surveys, 2010 and 2012. As discussed, we are unable to test between group years due to non-constant variability of *conpress*.

Since the we are testing a categorical variables *conpress* and *news* both with more than two levels, we will use the chi-squared test for independence method.

We will not calculate a confidence interval as this calculation estimates one parameter, and here we have multiple levels.

HYPOTHESIS TEST

First set the hypothesis test:

The null hypothesis H_0 : that the two variables *conpress* and *news* are independent for 2010 and for 2012.

The alternative hypothesis H_A : that the two variables *conpress* and *news* are dependent for 2010 and for 2012.

```
# Summarise 2010 data in a contingency table to perform the test
yearpresstable2010 <- yearpresstable %>%
```

```
  filter(year=="2010") %>%
  select(-year)
n2010 <- sum(yearpresstable2010$count)
yearpresstable2010 <- with(yearpresstable2010,tapply(count,list(confidence,newspaper),mean))
df <- (dim(yearpresstable2010)[1]-1 )*(dim(yearpresstable2010)[2]-1)
kable(yearpresstable2010,caption="Contingency table for 2010", align  = c("c", "c"))
```

Table 4: Contingency table for 2010

	Everyday	Few Times A Week	Once A Week	Less Than Once Wk	Never
A Great Deal	36	13	16	11	11
Only Some	120	61	47	46	52
Hardly Any	97	63	44	58	74
No opinion	196	140	108	105	132

```
# Summarise 2012 data in a contingency table to perform the test
yearpresstable2012 <- yearpresstable %>%
  filter(year=="2012") %>%
  select(-year)
n2012 <- sum(yearpresstable2012$count)
yearpresstable2012 <- with(yearpresstable2012,tapply(count,list(confidence,newspaper),mean))
df <- (dim(yearpresstable2012)[1]-1 )*(dim(yearpresstable2012)[2]-1)
kable(yearpresstable2012,caption="Contingency table for 2012", align = c("c", "c"))
```

Table 5: Contingency table for 2012

	Everyday	Few Times A Week	Once A Week	Less Than Once Wk	Never
A Great Deal	14	5	3	7	13
Only Some	86	50	38	51	61
Hardly Any	99	49	44	51	78
No opinion	154	126	110	108	154

CHECK CONDITIONS

Next check the conditions for this test:

1) The observations in the sample are independent.
   As previously discussed, this is a simple random sample and since n (2010)= 1430 and n (2012) = 1301 which are both $< 10\%$ population in the US, therefore assume observations of within each year group are independent. Each case contributes to one cell in the table.

2) Expected counts for each cell should be at least 5.
   For 2010, there are 0 and for 2012 there are 1 counts under 5. Therefore this condition is met for 2010 and not met for 2012.

3) Degrees of freedom should be at least 2.
   In this case df = 12 for both years so this condition is met.

PERFORM INFERENCE

However for 2010 we can perform the chi-squared test for independence at the 5% significance level:

```
chisq.test(yearpresstable2010,df)
```

```
##
##  Pearson's Chi-squared test
##
## data:  yearpresstable2010
## X-squared = 18.048, df = 12, p-value = 0.1142
```

For 2012 we would need to perform Monte Carlo simulation as the conditions are not met above, and perform the chi-squared test for independence at the 5% significance level:

```
chisq.test(yearpresstable2012,simulate.p.value = TRUE, B = 10000)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 10000
##  replicates)
##
## data:  yearpresstable2012
## X-squared = 15.134, df = NA, p-value = 0.2294
```

INTERPRET RESULTS

2010 Chi-squared Test Since the p-value $>$ the 5% significance level, we fail to reject the null hypothesis, and conclude that the data does not provide evidence for the alternative hypothesis. In other words, does not provide evidence that the two variables *conpress* and *news* are dependent for 2010.

2012 Chi-squared Test Since the p-value $>$ the 5% significance level, we fail to reject the null hypothesis, and conclude that the data does not provide evidence for the alternative hypothesis. In other words, does not provide evidence that the two variables *conpress* and *news* are dependent for 2010.

CONCLUSION

We have seen for 2010 and 2012, we did not see evidence that there is a dependent relationship in the the confidence in the press and the frequency of newspapers read. Since this is an observational study, we would not be able to infer a causal relationship between these two variables as there may be other confounding factors that could be further investigated.

Additionally it would be useful to source social media usage data in the next GSS survey to perform further tests on media channels.