Vihanga Ratnasinghe (Bee127) | Nila Nahar (Bee70) | K-M Samiul Haque (Bee18)

STAC53 Course Project

Data set options:

Canadian Alcohol and Drug Use Monitoring Survey 2012

Background

Previously, periodic surveys were used to keep track of the use of alcohol and substances. The data helps in various ways, such as monitoring trends and check seasonal biases related to the use of alcohol and substances. During consultations during the year 2008 to 2012, stakeholders establish a need for reliable and accessible data. The data would help to keep track of trends and increase the evidence base for development in various departments. Since the first survey implementation in 2008, every summer, the published results are of the previous year's Canadian Alcohol and Drug Use Monitoring Survey (CADUMS). CADUMS updated with the current standard measures for testing alcohol/drug use and abuse. The survey allowed for continuous monitoring and provided rich information about alcohol and drug use, behavioral changes and tendencies, and seasonal biases. The responses reflected on policies and programs to identify places of improvement.

Survey Methodology

CADUMS is a random digit-dialed telephone-based general population survey regarding alcohol and substance use, including illicit drugs, as well as prescription drugs. CADUMS targeted Canadian residents aged 15 years and older with a household telephone. An electronic inventory consisting of active telephone area codes and exchanges in Canada made up the sampling frame. The provinces acted as strata, and the first stage of selection included a random sample of household telephone numbers. The sampling unit was each household. For the second stage of selection, the chosen respondent was a member who had the next birthday in the house. To ensure maximum participation, the data collection firm scheduled callbacks and reconducted unwilling respondents. Northwest Territories, Nunavut, and Yukon residents, including permanent residents of various institutions, no telephone households, and cellphones-only households, were excluded from the sample.

Method of Data Collection:

CADUMS used RDD methods via Computer Assisted Telephone Interviewing (CATI) to hold phone interviews to collect data. A content pre-test held to recognize potential problems with the questionnaire like bad flow of questions and modules, sensitive questions or modules affecting participation and responses, etc. Average completion time is approximately 21 minutes and data were collected over ten months. One of the limitations of the sampling design was an inability to reach residents due to homelessness, incarceration, households without telephone and cell-phone only houses. Research has shown homeless and incarcerated population consists of heavy substance users, and an inability to interview them could be a potential source of bias in the data collection process. However, since considering the excluded population makes up very little of the total Canadian population, the exclusion has minimal effect on the reliability of estimates. Additionally, language was an interview barrier, which restricted responses from residents not speaking or understanding either English or French. Since CADUMS tap into a sensitive subject like alcohol and substance use, under-reporting of socially unacceptable behaviors could potentially contribute to the bias of self-report. However, reviews have shown that despite limitations, self-reports are the best available means of estimating such intimate behaviors and should have less impact as long as bias estimate remains constant. Additionally, CADUMS incorporates ensuring confidentiality of data, using non-judgmental phrases, transparency between researchers and participants, and avoiding fear of punishment, all of which are proved to be appropriate approaches to obtain accurate responses through self-report. CADUMS sample represented 27,767,855 Canadians older than 14 years old. Response rate was the highest in Quebec with 55.8%, followed by Nova Scotia with 41.1% and Manitoba with 40.1%.

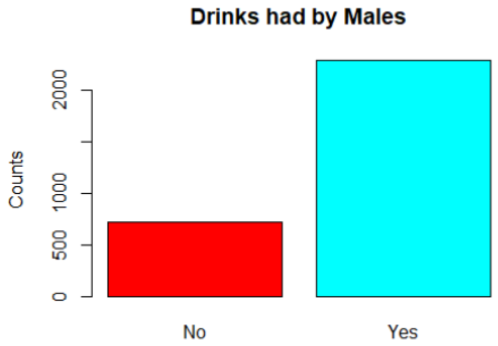
Research questions:

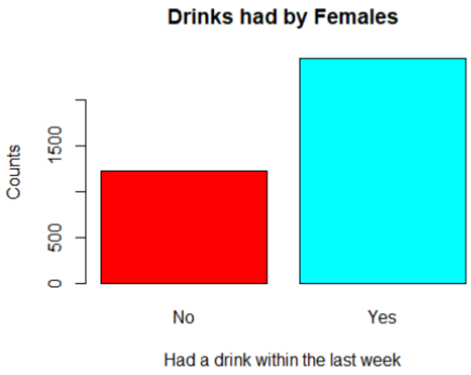
1. Research question 1: Check for association between drinking alcohol within the past week and their sex. Sex is taken as a discrete explanatory variable, whereas alcohol and substance use are taken as a discrete response variable.
2. Research question 2: Compare the mean smoking time between male and female respondents. Sex is taken as a discrete explanatory variable and smoking time is taken as a discrete response variable.

Brief summary of statistical methods:

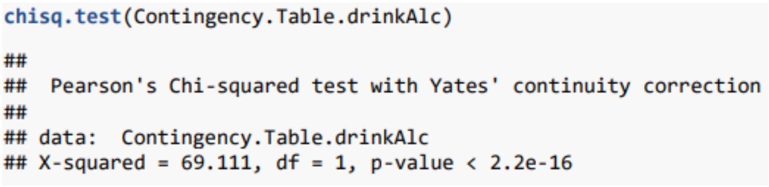
The statistical methods we will be using for our two questions will be a significance test and the use of a confidence interval. Significance tests are used to assess evidence about a claim based on the population from which the sample has been drawn. More specifically, we'll be using the Chi-Square Test of Independence to assess association between two variables. This test is used to see if there's a significant relationship between two variables which fits perfectly with our research questions. A t-test will also be used to compare means between our data. We will also use confidence intervals to help us in our statistical analysis. The null hypothesis for this test would be that there is no association between the two variables we're testing from the sample while the alternate hypothesis for this test would be that there is in fact some sort of association between the two variables.

Our first question in our statistical analysis is whether there was an association between if people had a drink within the past week and their sex. We can start off by doing some exploratory data analysis and creating boxplots to see the distribution of people who drank and their sex.



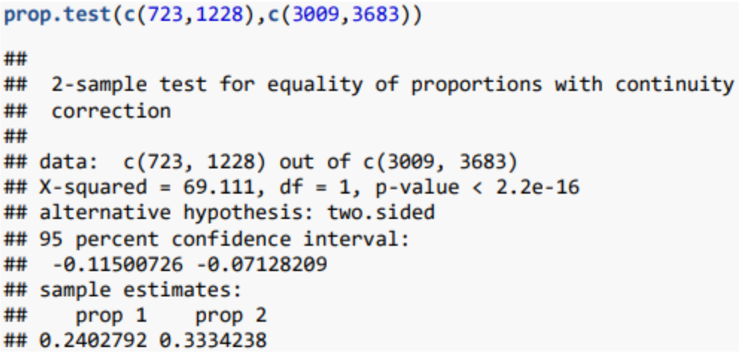


In terms of the test, we'll conduct a chi-squared test of independence. This test is used to check if there's a relationship between two variables. The null hypothesis, H_0 is that there is no association between having a drink within the past week and the alternate hypothesis, H_a is that there is an association between the two variables.



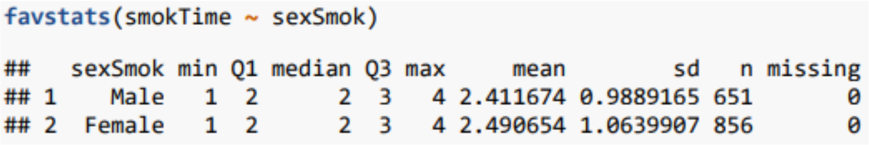
```
chisq.test(Contingency.Table.drinkAlc)

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  Contingency.Table.drinkAlc
## X-squared = 69.111, df = 1, p-value < 2.2e-16
```

The X^2 value is 69.111 and the p-value $< 2.2e - 16$. This means that we can reject the null hypothesis and so there's strong evidence to indicate that having a drink within the past week is associated with the sex of the respondents. We can also calculate a 95% confidence interval.

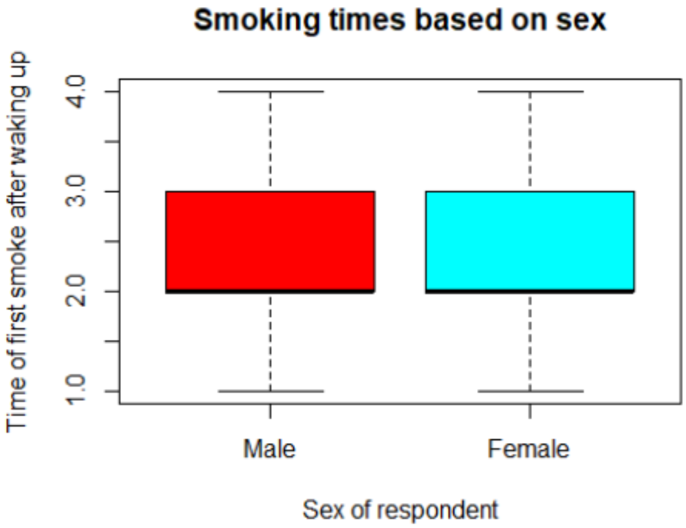


```
prop.test(c(723,1228),c(3009,3683))

##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(723, 1228) out of c(3009, 3683)
## X-squared = 69.111, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.11500726 -0.07128209
## sample estimates:
##    prop 1    prop 2
## 0.2402792 0.3334238
```

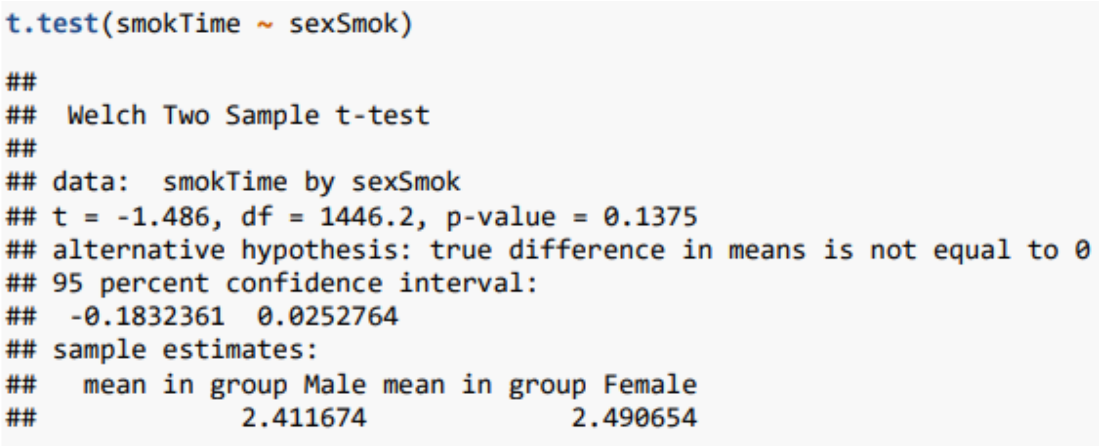
We can interpret this as that we're 95% confident that the percent of people that don't drink for females is between 7.1% and 11.5% higher than for the males. Conversely, it can be interpreted as that we're 95% confident that the percent of people that drink for males is between 7.1% and 11.5% higher than for the females. For our second question, we'll be comparing the mean initial smoking time between males and females. For some initial data analysis, we can create a side by side boxplot which is very helpful as it shows how the values are spread out. We can also create some summary statistics to figure out the mean and standard deviation for the two variables.



```
favstats(smokTime ~ sexSmok)

##   sexSmok min Q1 median Q3 max     mean        sd   n missing
## 1    Male   1  2      2  3   4 2.411674 0.9889165 651       0
## 2  Female   1  2      2  3   4 2.490654 1.0639907 856       0
```



As you can see, the mean and standard deviation is around the same for both variables. From the boxplot, we can also see that the variables are spread out evenly. A better way to compare the mean initial smoking time between males and females would be to use a t-test. This is a perfect test since a t-test is a specific test to determine if there's a significant difference between the means of two groups.



```
t.test(smokTime ~ sexSmok)

##
##  Welch Two Sample t-test
##
## data:  smokTime by sexSmok
## t = -1.486, df = 1446.2, p-value = 0.1375
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1832361  0.0252764
## sample estimates:
##   mean in group Male mean in group Female
##             2.411674             2.490654
```

Since our p value > 0.05, we fail to reject the null hypothesis and can't accept the alternative hypothesis. In the context of our data, the null hypothesis, H_0 is that the mean for males and females are equal, while the alternative hypothesis H_a is that the means aren't equal. Ultimately, this just means that there's a good chance the mean smoking times for males and females are equal. We can go further and use the 95% Confidence interval that's also given to us. Since the value 0 is within the range of -0.18 to 0.025, we know that there's a possibility that the difference between the means could be 0, which is the same as saying that the means are equal.