Lab 7

Salim M'jahad msm2243

March 8th, 2018

Instructions

Make sure that you upload an RMarkdown file to the canvas page (this should have a .Rmd extension) as well as the PDF or HTML output after you have knitted the file (this will have a .pdf or .html extension). Note that since you have already knitted this file, you should see both a **Lab7__UNI.pdf** and a **Lab7__UNI.Rmd** file in your UN2102 folder. Click on the **Files** tab to the right to see this. The files you upload to the Canvas page should be updated with commands you provide to answer each of the questions below. You can edit this file directly to produce your final solutions. The lab is due 11:59pm on Tuesday, March 20th.

Goal

The goal of this lab is to write a **R** function named **my.chi.squared.test** that performs the classic chi-squared test of association. Details of the chi-squared test are provided below.

Chi-Squared Test of Association

Consider testing whether two categorical variables are statistically associated with each other. A few examples follow:

- Is gender statistically associated with political affiliation?
- Is smoking status statistically associated with whether or not a respondent will have lung cancer?
- Is political affiliation statistically associated whether or not a respondent uses marijuana?
- Is the socioeconomic status of a respondent's parent/guardian statistically associated the respondent's education level?

The *levels* of a categorical variable are the different labels a variable can take on. For example, the variable **Gender** has levels **Male** and **Female** while political affiliation could have several levels. In the following example, political affiliation has three levels and marijuana usage also has three levels.

Example

Consider a study where each respondent in a random sample of high school and college students was cross-classified with respect to both political views and marijuana usage. The raw counts are displayed in the accompanying contingency table. Does the data support the hypothesis that political views and marijuana usage are associated with each other? Use a significance level of 1%. **Note**: The **expected counts** are calculated from the **observed counts**.

In this application we test the null alternative pair:

H_0 : Political views and marijuana usage are not associated with each other.

H_A : Political views and marijuana usage are associated with each other.

Observed counts

		Usage level			
		Never	Rarely	Frequently	**Total**
	Liberal	479	173	119	
Political views	Conservative	214	47	15	
	Other	172	45	85	
	Total				

Expected counts

		Usage level			
		Never	Rarely	Frequently	**Total**
	Liberal	494.38	151.56	125.17	
Political views	Conservative	176.98	54.22	44.81	
	Other	193.65	59.33	49.03	
	Total				

The Chi-Squared Statistic

Consider a $r \times c$ contingency table where r is the number of rows and c is the number of columns in the table. The observed count for cell i, j is denoted o_{ij}.

	column 1	column 2	$\cdots$	column c	Total
row 1	o_{11}	o_{12}	$\cdots$	o_{1c}	row total (1)
row 2	o_{21}	o_{22}	$\cdots$	o_{2c}	row total (2)
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
row r	o_{r1}	o_{r2}	$\cdots$	o_{rc}	row total (r)
Total	column total (1)	column total (2)	$\cdots$	column total (c)	sample size (n)

The **expected counts** are calculated using

$$e_{ij} = \frac{(\text{Row total}) \times (\text{Column total})}{\text{Grand total}}.$$

The random variable

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

has an approximate chi-squared distribution with degrees of freedom

$$df = (r - 1) \times (c - 1).$$

Test Statistic:

$$\chi^2_{calc} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

A large computed value of the test statistic supports the alternative H_A.

P-value and Rejection Rule

The P-value is computed using the **R** code

$$P(\chi^2 > \chi^2_{calc}) = 1 - \text{pchisq}(\chi^2_{calc}, df = (r-1) \times (c-1))$$

Reject the null hypothesis if

$$Pvalue \leq \alpha$$

Tasks

Part 1

Write a **R** function named **my.chi.squared.test** that performs the classic chi-squared test of association.

Requirements of the function **my.chi.squared.test** follow below:

- Inputs:
  - **data** is the raw data set. Note that the data can be in its original form or summarized by a contingency table.
  - **level** is the significance level as a decimal. It must strictly exist between 0 and 1 with default at $\alpha = .05$.
- The output should be a **list** with the following labeled elements:
  - **expected** is a table of expected counts corresponding to the observed counts.
  - **statistic** is the computed chi-squared test statistic.
  - **pvalue** is the computed P-value.
  - **conclusion** is the statistical conclusion. It should say either "Reject H0 at 5% significance" or "Fail to reject H0 at 5% significance." The level of the test (5%) should change depending on your initial input in the function.

```r
setwd("/Users/salimmjahad/Desktop/STAT_COMP/lab7")
smoking <- read.csv("Smoking.csv", as.is = TRUE)
viewsandpot <- read.csv("ViewsandPot.csv", as.is = TRUE)

my.chi.squared.test <- function(data,
                                level=0.05) {
  contig <- table(data)
  exp_c <- contig
  g_total <- sum(contig)

  for (row in rownames(contig)) {
    rowsum <- sum(contig[row,])
    for (col in colnames(contig)) {
      exp_c[row, col] = (rowsum*sum(contig[,col]))/g_total
    }
  }

  TS <- sum(((contig-exp_c)^2)/exp_c)
  PV <- 1-pchisq(TS,df=(length(rownames(contig))-1)*(length(colnames(contig))-1))
  concl = ifelse(PV>level,
                     paste0("Fail to reject H0 at ", level*100, "% significance"),
                     paste0("Reject H0 at ", level*100, "% significance"))

  result = list(expected= exp_c,
                statistic= TS,
                pvalue= PV,
                conclusion= concl)
  return(result)
}
```

Part 2

Run the function **my.chi.squared.test** on the political view versus marijuana usage example. Test the claim at 1% significance. Accompany the test with an appropriate graphic using **ggplot**. Note that the dataset **ViewsandPot** is posted on Canvas.

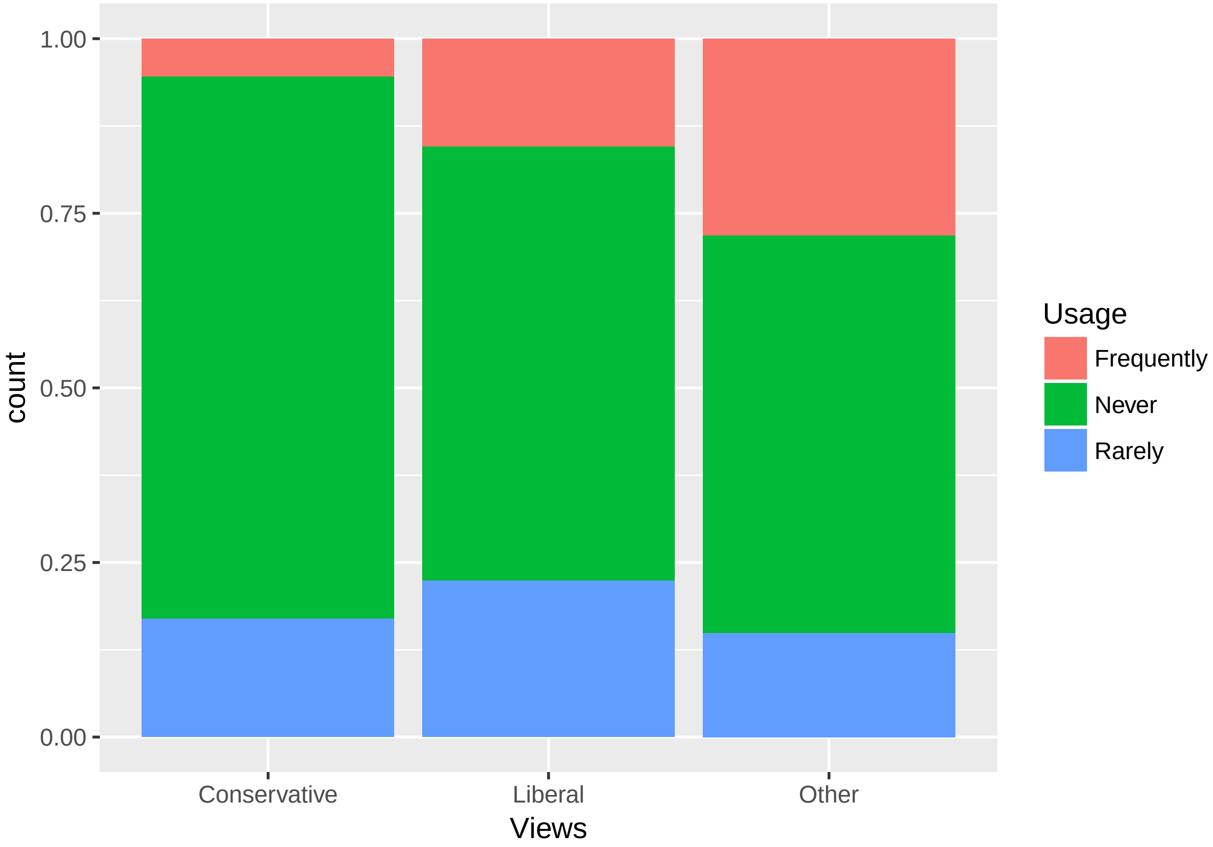
```
vap <- my.chi.squared.test(viewsandpot, 0.01)
halo <- table(viewsandpot)
vap
```

```
## $expected
##             Views
## Usage         Conservative   Liberal     Other
##   Frequently     44.80652 125.16605  49.02743
##   Never         176.97554 494.37732 193.64715
##   Rarely         54.21794 151.45663  59.32543
##
## $statistic
## [1] 64.65417
##
## $pvalue
## [1] 3.043121e-13
##
## $conclusion
## [1] "Reject H0 at 1% significance"
```

```
head(viewsandpot)
```

```
##   Usage   Views
## 1 Never Liberal
## 2 Never Liberal
## 3 Never Liberal
## 4 Never Liberal
## 5 Never Liberal
## 6 Never Liberal
```

```
library(ggplot2)
ggplot(viewsandpot, aes(x=Views, fill=Usage))+geom_bar(position="fill")
```



Part 3

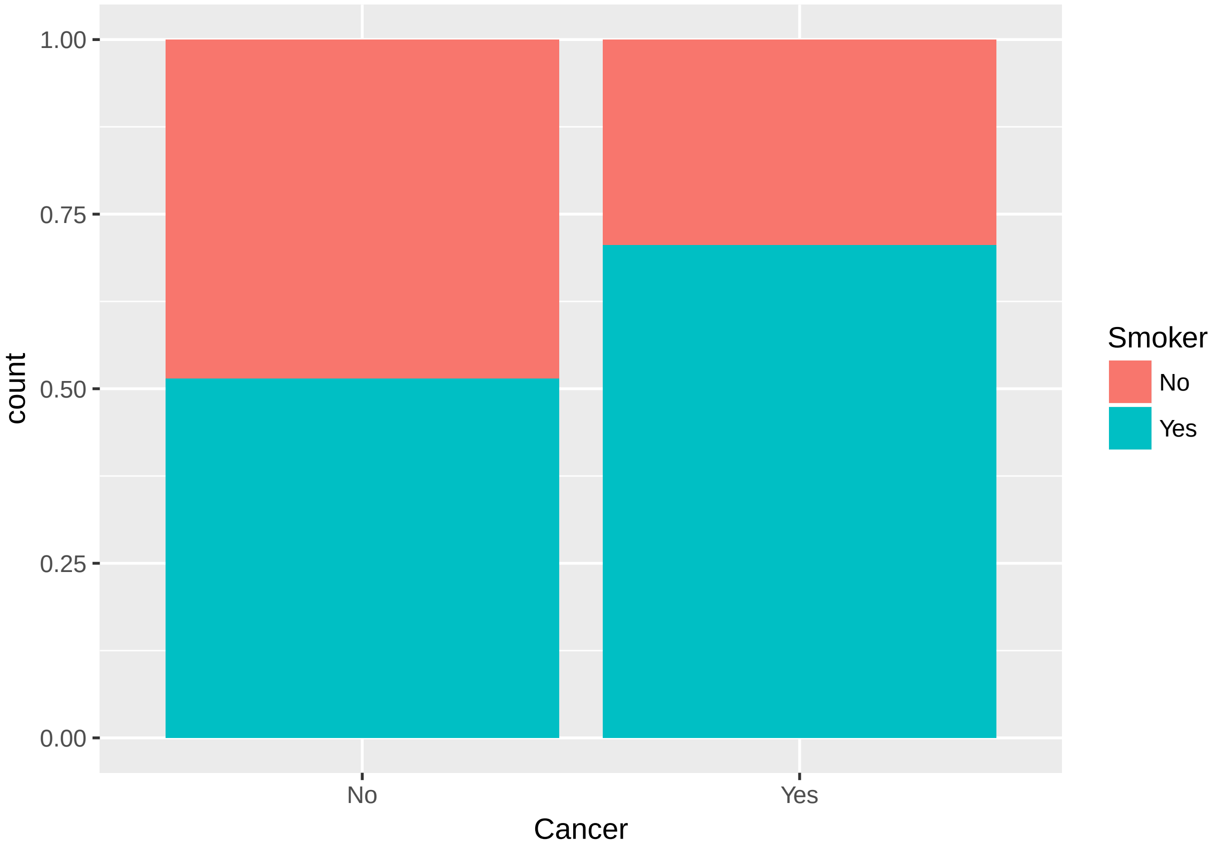
Consider a new application stated below:

The health histories of 11,900 middle-aged men were tracked over many years. During the study 126 of the men developed lung cancer, including 89 men who were smokers and 37 men who were former smokers. The data set **Smoking** is posted on Canvas. Run the function **my.chi.squared.test** on the above data set to see if smoking status is statistically associated to whether or not the men developed lung cancer. Accompany the test with a contingency table and an appropriate graphic using **ggplot**.

```
s <- my.chi.squared.test(smoking, 0.01)
s
```

```
## $expected
##       Smoker
## Cancer         No        Yes
##    No  5687.13882 6086.86118
##    Yes   60.86118   65.13882
##
## $statistic
## [1] 18.28929
##
## $pvalue
## [1] 1.897707e-05
##
## $conclusion
## [1] "Reject H0 at 1% significance"
```

```r
ggplot(smoking, aes(x=Cancer, fill=Smoker))+geom_bar(position="fill")
```


count
No
Yes
Cancer
Smoker
No
Yes