Mohammad Javad Ranjbar

810101173

Homework 5

Statistical Inference, Fall 2022

## Contents

## Question #1:

a) For 95% confidence, we have $\mathrm{ME} = 2$

$$82 - 2 < p < 82 + 2 \Rightarrow 80 < p < 84$$

A) False. The CI is constructed to estimate the population proportion, which is wrong. The CI is constructed to estimate the population proportion and not the sample proportion. The correct phrase would be: We are 95% confident that 82 % to 84% of **all** Americans think it's the government's responsibility to promote equality between men and women.

B) True. This is the correct detention of CI for this question. We are 95% confident that 82 % to 84% of **all** Americans think it's the government's responsibility to promote equality between men and women.

C) True. Confidence intervals provide us with an upper and lower limit around our sample proportion, and within this interval, we can then be confident we have captured the population proportion.

D) True. Based on the ME formula, we need to multiply the sample size by 4.

$$\frac{\mathrm{ME}_1}{\mathrm{ME}_2} = \frac{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1}}}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_2}}} = \frac{2}{1} \Rightarrow \sqrt{\frac{n_2}{n_1}} = 2 \Rightarrow n_2 = 4n_1$$

E) True. We are 95% confident that Americans think it's the government's responsibility to promote equality between men and women.

## Question #2:

$$H_0: p = 0.5$$

$$H_A: p > 0.5$$

First, we need to check for the conditions:

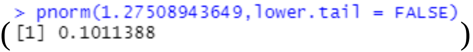
- Independence: Sampled observations must be independent.
  - The newspaper collects random samples. Therefore, the random sampling condition is true.
  - The sampling is without replacement and we are sure that the number of samples is less than 10% of all the population.
- Sample size/skew: There should be at least 10 successes and 10 failures in the sample:
  - $np \geq 10 \text{ and } n(1-p) \geq 10 \Rightarrow 450 * \frac{50}{100} > 10 \text{ and } 450 * \frac{50}{100} > 10.$

With these conditions verified, the normal model may be applied to $\hat{p}$.

$$\mathrm{SE} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{\frac{50}{100} * \frac{50}{100}}{450}} = 0.02357022603$$

$$Z = \frac{\text{observation} - \text{null}}{\mathrm{SE}} = \frac{0.53 - 0.5}{0.02357022603} = 1.27508943649$$

$P(z \geq 1.27) = 0.1011$( )



```
> pnorm(1.27508943649,lower.tail = FALSE)
[1] 0.1011388
```

Because the p-value is larger than 0.05, we do not reject the null hypothesis, and we do not find convincing evidence to support the campaign manager's claim.

## Question #3:

a)

| Name | Yes | No | Total | $\hat{p}$ | $p_{pool}$ |
|---|---|---|---|---|---|
| **Lopinavir** | 26 | 110 | 120 | $\frac{26}{36}$ | $\frac{36}{120}$ |
| **Nevaripine** | 10 | 94 | 120 | $\frac{10}{36}$ | |
| **Total** | 36 | 204 | 240 | 240 | |

b)

$H_0: p_{f_{Nevaripine}} = p_{f_{Lopinavir}}$: There is no significant difference in virologic failure rates between the two methods of using Lopinavir or Nevaripine.

$H_A: p_{f_{Nevaripine}} \neq p_{f_{Lopinavir}}$: There is a significant difference in virologic failure rates between the two methods of using Lopinavir or Nevaripine.

c)

First, we need to check for the conditions:

- Independence:
  - within groups: sampled observations are independent within each group
    - The women have been randomly assigned to each treatment group.
    - Sampling is without replacement, and the number of the sample is less than 10% population of women.
  - between groups: the two groups of treatment are independent of each other
- Sample size/skew: Each sample should meet the success-failure condition:
  - $n_1 p_{pool} \geq 10$ and $n_1(1 - p_{pool}) \geq 10 \Rightarrow 120 * \frac{36}{240} = 18 > 10$ and $120 * \frac{204}{240} = 102 > 10$
  - $n_2 p_{pool} \geq 10$ and $n_2(1 - p_{pool}) \geq 10 \Rightarrow 120 * \frac{36}{240} = 18 > 10$ and $120 * \frac{204}{240} = 102 > 10$

With these conditions verified, the normal model may be applied.

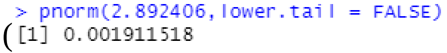
$\hat{p}_1 = \frac{26}{36}, \hat{p}_2 = \frac{10}{36}, p_{pool} = \frac{36}{240}$

$$SE = \sqrt{\frac{p_{pool}\left(1-p_{pool}\right)}{n_1} + \frac{p_{pool}\left(1-p_{pool}\right)}{n_2}} = \sqrt{\frac{\frac{36}{240}*(1-\frac{36}{240})}{120} + \frac{\frac{36}{240}*(1-\frac{36}{240})}{120}}$$

$$= 0.04609772$$

$$Z = \frac{observation - null}{SE} = \frac{\frac{26}{120} - \frac{10}{120} - 0}{0.04609772} = 2.892406$$

$p(|Z| > 2.892406) = 0.001911518$ ( )



```
> pnorm(2.892406,lower.tail = FALSE)
[1] 0.001911518
```

$p - value < \alpha = 0.05 \Rightarrow$ Therefore, There is strong evidence to reject $H_0$. This means, there is a difference in the proportion of virologic failures among those who take Nevaripine and those who take Lopinavir.

## Question #4:

| | Sample size | Approve law | Disapprove law | Other |
|---|---|---|---|---|
| **"people who cannot afford it will receive financial help from the government" is given second** | 771 | 47% | 49% | 3% |
| **"people who do not buy it will pay a penalty" is given second** | 732 | 34% | 63% | 3% |

First, we need to check for the conditions:

- Independence:
  - within groups: the observations are independent, both within the samples and between the samples.
    - the statements in brackets were randomly assigned to each respondent.
    - Sampling is without replacement, and the number of the sample is less than 10% population.
  - between groups: the two groups of the respondent are independent of each other
- Sample size/skew: Each sample should meet the success-failure condition:
  - $n_1\hat{p}_1 \geq 10$ and $n_1(1-\hat{p}_1) \geq 10 \Rightarrow 771 * 0.47 = 362.37 > 10$ and $771 * 0.53 = 408.63 > 10$
  - $n_2\hat{p}_2 \geq 10$ and $n_2(1-\hat{p}_2) \geq 10 \Rightarrow 732 * 0.34 = 248.88 > 10$ and $732 * 0.66 = 483.12 > 10$

With these conditions verified, the normal model can be used for the point estimate of the difference in support

$$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = \sqrt{\frac{0.47*(1-0.47)}{771} + \frac{0.34*(1-0.66)}{732}} = 0.02509$$

For 95% confidence, we have $z^* = 1.6448$ ( )

```
> qnorm(0.95)
[1] 1.644854
```

$\hat{p}_2 - \hat{p}_1 \pm z^* * SE = (0.49 - 0.34) \pm 1.6448 \; * 0.02509 \Rightarrow (0.08873, 0.1712)$

We are 90% confident that the approval rating for the 2010 healthcare law changes is between 9% and 17% due to the ordering of the two statements in the survey question.

## Question #5:

| Options | Rock | Scissors | Paper | Total |
|---|---|---|---|---|
| **Time played** | 43 | 21 | 35 | 99 |
| **Expected** | 33 | 33 | 33 | 99 |

Our Hypothesis would look like this:

$$H_0: p_{\text{Rock}} = p_{\text{Scissors}} = p_{\text{Paper}}\text{: All options are favored equally}$$

$$H_A\text{: Some options are favored more than others options.}$$

We need to check the conditions for the test:

- Independence: Sampled observations are independent. And each case only contributes to one cell.
- Sample size: Expected Value is equal to np, where the proportion for each is 0.33 assuming an equal chance to pick each one, and the sample size is 99 (43+21+35). All the expected values will be 33, which is greater than 5. Therefore, we have a large enough sample.

With the verification of the conditions. We Chi-Square Goodness of Fit test to test our hypothesis.

$$\chi^2 = \frac{(43-33)^2}{33} + \frac{(21-33)^2}{33} + \frac{(35-33)^2}{33} = 7.515152$$

$$\mathrm{df} = \mathrm{k} - 1 = 3 - 1 = 2$$

$p - value = 0.023$( )

```
> pchisq(7.515152, 2, lower.tail = FALSE)
[1] 0.02334025
```

p-value is less than $\alpha = 0.05$. Therefore, Therefore, There is strong evidence to reject $H_0$. This means, some of the options are favored more than others options.

## Question #6:

| | Ketchup | Mustard | Relish | Total |
|---|---|---|---|---|
| **Male** | 15 (19.2) | 23 (20.16) | 10 (8.64) | 48 |
| **Female** | 25 (20.8) | 19 (21.84) | 8 (9.36) | 52 |
| **Total** | 40 | 42 | 18 | 100 |

Our Hypothesis would look like this:

$$H_0\text{: Gender and condiment are independent.}$$

$$H_A\text{: Gender and condiment are dependent.}$$

We need to check the conditions for the test:

- Independence: Sampled observations are independent and without replacement and less than 10% of the population. And each case only contributes to one cell.
- Sample size: None of the expected counts in the table are less than 5.

With the above conditions verified, we can proceed with the Chi-Square test.

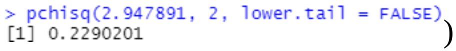
$$\text{Expected count} = \frac{(\text{row total}) \ * \ (\text{column total})}{\text{total table}}$$

$\mathrm{E}_{\text{male,Ketchup}} = \frac{40 * 48}{100} = 19.2$ , $\mathrm{E}_{\text{male,Musterd}} = \frac{42 * 48}{100} = 20.16$ , $\mathrm{E}_{\text{male,Relish}} = \frac{18 * 48}{100} = 8.64$

$\mathrm{E}_{\text{female,Ketchup}} = \frac{40 * 52}{100} = 20.8$ , $\mathrm{E}_{\text{female,Musterd}} = \frac{42 * 52}{100} = 21.84$ , $\mathrm{E}_{\text{female,Relish}} = \frac{18 * 52}{100} = 9.36$

$$\chi^2 = \frac{(15-19.2)^2}{19.2} + \frac{(23-20.16)^2}{20.16} + \frac{(10-8.64)^2}{8.64} + \frac{(25-20.8)^2}{20.8} + \frac{(19-21.84)^2}{21.84} + \frac{(8-9.36)^2}{9.36} = 2.947891$$

df = 3-1 = 2

p-value = 0.229 ( 

```
> pchisq(2.947891, 2, lower.tail = FALSE)
[1] 0.2290201
```

)

With a p-value greater than 5%, Since p-value is high we fail to reject $H_0$. This means, we conclude that there is not enough evidence in the data to suggest that gender and preferred condiment are dependent.

## Question #7:

I solved this question in two ways. The first method is by using the built-in chi-square test and the second method is without using the chi-square test function. The result of the two methods is exactly equal.

Our Hypothesis would look like this:

$H_0$: Absence occurs with equal frequencies during a five-day workweek.

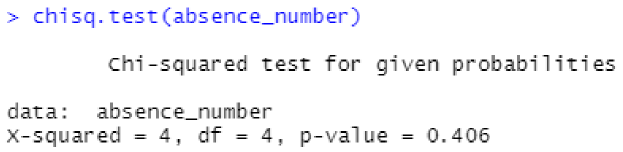
$H_A$: Absence does not occur with equal frequencies during a five-day workweek.

Check the conditions for the test:

- Independence: Sampled observations are independent. And each case only contributes to one cell.
- Sample size: Expected Value is 14 which is larger than 5.

| | Monday | Tuesday | Wednesday | Thursday | Friday |
|---|---|---|---|---|---|
| **Number of Absences** | 18 | 16 | 10 | 10 | 16 |
| **Expected** | 14 | 14 | 14 | 14 | 14 |

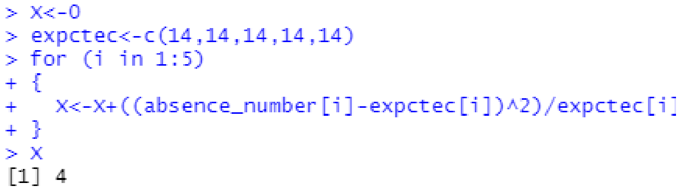
1- By using the chi-square function:



```
> chisq.test(absence_number)

        Chi-squared test for given probabilities

data:  absence_number
X-squared = 4, df = 4, p-value = 0.406
```

With a p-value greater than 5%, Since p-value is high we fail to reject $H_0$. Therefore, we conclude that there is not enough evidence in the data to suggest that absence does not occur with equal frequencies during a five-day workweek.

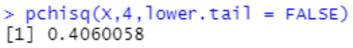
2- Without using the chi-square function:



```
> X<-0
> expctec<-c(14,14,14,14,14)
> for (i in 1:5)
+ {
+   X<-X+((absence_number[i]-expctec[i])^2)/expctec[i]
+ }
> X
[1] 4
```

$$\chi^2 = \frac{(18-14)^2}{14} + \frac{(16-14)^2}{14} + \frac{(10-14)^2}{14} + \frac{(10-14)^2}{14} + \frac{(16-14)^2}{14} = 4$$

$\text{df} = 5 - 1 = 4$

p-value $= 0.4060058$(  )

```
> pchisq(X,4,lower.tail = FALSE)
[1] 0.4060058
```

With a p-value greater than 5%, Since p-value is high we fail to reject $H_0$. Therefore, we conclude that there is not enough evidence in the data to suggest that absence does not occur with equal frequencies during a five-day workweek.

## Question #8:

I solved this question in two ways. The first method is by using the built-in chi-square test and the second method is without using the chi-square test function. The result of the two methods is exactly equal.

Our Hypothesis would look like this:

$H_0$: The students' smoking habit is independent of their exercise level.

$H_A$: The students' smoking habit is dependent on their exercise level.

Check the conditions for the test:

- Independence: Sampled observations are independent. And each case only contributes to one cell.
- Sample size: each particular scenario does not have at least 5 cases. Therefore this condition is False.

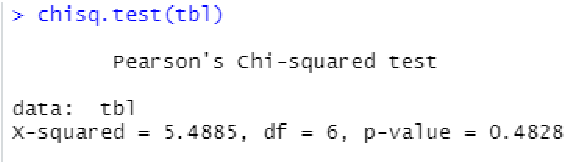
The second condition is not verified for doing the test. However, I asked one of the teaching assistants and they said to ignore this condition and do the test normally.

Hence, I assume the conditions are met:

| | | Smoke | | | |
|---|---|---|---|---|---|
| | | Freq | None | Some | Total |
| **Excer** | Heavy | 7 | 1 | 3 | 11 |
| | Never | 87 | 18 | 84 | 189 |
| | Occas | 12 | 3 | 4 | 19 |
| | Regul | 9 | 1 | 7 | 17 |
| | Total | 115 | 23 | 98 | 236 |

1- With The chi-square test:

We can simply use the built-in function of chi-square test:

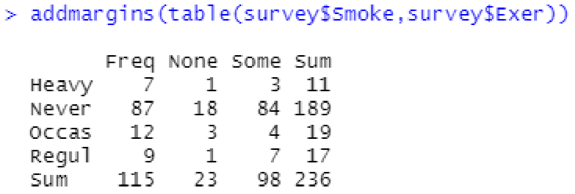


```
> chisq.test(tbl)

        Pearson's Chi-squared test

data:  tbl
X-squared = 5.4885, df = 6, p-value = 0.4828
```

p-value = 0.482. With a p-value greater than 5%, Since p-value is high we fail to reject $H_0$. Therefore, we conclude that there is not enough evidence in the data to suggest that the smoking habit is dependent on the exercise level of the students.

2- Without chi-square test

In R:

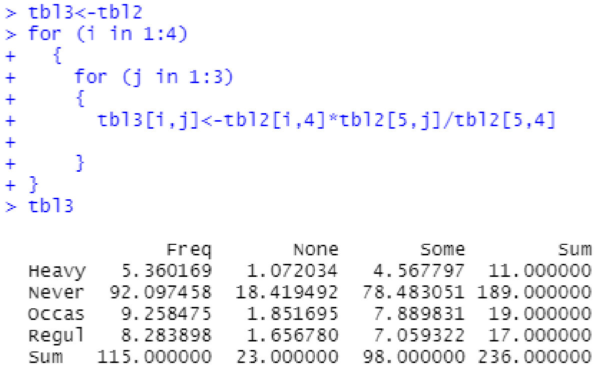


```
> addmargins(table(survey$Smoke,survey$Exer))

        Freq None Some Sum
  Heavy    7    1    3  11
  Never   87   18   84 189
  Occas   12    3    4  19
  Regul    9    1    7  17
  Sum    115   23   98 236
```

Now we need to calculate the excepted values:

| Smoke | | | | | |
|---|---|---|---|---|---|
| | | Freq | None | Some | Total |
| **Excer** | Heavy | 7 (5.36) | 1 (1.07) | 3 (4.56) | 11 |
| | Never | 87 (92.09) | 18 (18.41) | 84 (78.48) | 189 |
| | Occas | 12 (9.25) | 3 (1.85) | 4 (7.88) | 19 |
| | Regul | 9 (8.28) | 1 (1.65) | 7 (7.05) | 17 |
| | Total | 115 | 23 | 98 | 236 |

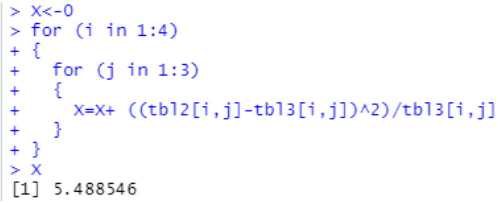
In R:



```
> tbl3<-tbl2
> for (i in 1:4)
+   {
+     for (j in 1:3)
+     {
+       tbl3[i,j]<-tbl2[i,4]*tbl2[5,j]/tbl2[5,4]
+
+     }
+ }
> tbl3

             Freq       None       Some        Sum
  Heavy   5.360169   1.072034   4.567797  11.000000
  Never  92.097458  18.419492  78.483051 189.000000
  Occas   9.258475   1.851695   7.889831  19.000000
  Regul   8.283898   1.656780   7.059322  17.000000
  Sum   115.000000  23.000000  98.000000 236.000000
```

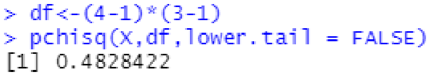
Now we calculate the chi-square static:

$$\begin{aligned}\chi^2 = {} & \frac{(7-5.36)^2}{5.36}+\frac{(1-1.07)^2}{1.07}+\frac{(3-4.56)^2}{4.56}+\frac{(87-92.09)^2}{92.09}+\frac{(18-18.41)^2}{18.41}\\ & +\frac{(84-78.48)^2}{78.48}+\frac{(12-9.25)^2}{9.25}+\frac{(3-1.85)^2}{1.85}+\frac{(4-7.88)^2}{7.88}+\frac{(9-8.28)^2}{8.28}\\ & +\frac{(1\ -1.65)^2}{1.65}+\frac{(7-7.05)^2}{7.05}=5.488546\end{aligned}$$



```
> x<-0
> for (i in 1:4)
+ {
+   for (j in 1:3)
+   {
+     x=x+ ((tbl2[i,j]-tbl3[i,j])^2)/tbl3[i,j]
+   }
+ }
> x
[1] 5.488546
```

df = 3-1 = 2



```
> df<-(4-1)*(3-1)
> pchisq(x,df,lower.tail = FALSE)
[1] 0.4828422
```

p-value $= 0.482$. With a p-value greater than 5%, Since p-value is high we fail to reject $H_0$. Therefore, we conclude that there is not enough evidence in the data to suggest that the smoking habit is dependent on the exercise level of the students.