Save this homework file as: HW4\_LastName\_FirstName.docx

Save your R script file as: HW4\_LastName\_FirstName.R

You must use R for the analysis, but if you use Excel for anything else, save your file as: HW3\_LastName\_FirstName.xlsx

Be sure to include your name and homework 4 in the syntax comments. Add comments throughout your script so the grading team will be able to follow your process better.

*-0.5 points will be deducted for not using comments in your syntax file.*

Use the available R script file to help setup variables for select questions below.

You may copy/paste R code in this homework to help you complete the questions.

Any R code will appear in blue like this.

**Any Excel formulas will appear like this.**

**Grading Rubric**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Total** | **A** | **B** | **C** | **D** | **E** |
| **Q1** | **1.00** | **0.20** | **0.40** | **0.40** | **--** | **--** |
| **Q2** | **1.00** | **0.20** | **0.70** | **0.10** | **--** | **--** |
| **Q3** | **1.00** | **0.33** | **0.33** | **0.34** | **--** | **--** |
| **Q4** | **1.00** | **0.25** | **0.50** | **0.25** | **--** | **--** |
| **Q5** | **1.00** | **0.20** | **0.20** | **0.20** | **0.20** | **0.20** |

**Question 1**

1) Is there a relationship between gender and preference for ice cream flavor? The table below summarizes the preferences of Male and Female on their preferred flavor of ice-cream.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Strawberry | Vanilla | Chocolate |
| Male | 100 | 120 | 60 |
| Female | 350 | 200 | 90 |

1a) Which type of Chi-square test is appropriate for this analysis? (0.20 points)

**For this analysis, you should use the chi-square test of independence because we have two measurement variables, which are gender (male, female) and ice cream flavor (strawberry, vanilla, chocolate). We are measuring whether these two variables are related.**

1b) Conduct the appropriate Chi-square test in R and report the X2, degrees of freedom and P-value. You may choose the level of alpha. Be sure to report the values using APA citation. (0.40 points)

**X2 (2, n = 920) 28.362, p = 0.0000006938**

**X2 = 28.362**

**Degrees of freedom = 2**

**P-value = 6.938 x 10-7**

**-- R –**

**# entering data**

**icecream <- matrix(c(100, 120, 60, 350, 200, 90), ncol=3, byrow=T)**

**# naming rows/columns**

**rownames(icecream) <- c("Male", "Female")**

**colnames(icecream) <- c("Strawberry", "Vanilla", "Chocolate")**

**icecream**

**# chi-square test**

**chisq.test(icecream)**

**sum(icecream)**

1c) Based on the results above what would you conclude about ice cream preferences between male and female? (0.40 points)

**The p-value, which is 6.938 x 10-7, is less than 0.05 (α). This means the results are statistically significant. This low p-value also represents strong evidence against the null hypothesis. I would reject the null and conclude that there is a relationship between gender and ice cream flavor preference.**

**Question 2**

2) A research firm claims that the distribution of the days of the week that people are most likely to order food for delivery is different from the distribution seen in the past. You randomly select 494 people and record which day of the week each is most likely to order food for delivery. The table below also shows the results of your count. At alpha, α, = 0.05, test the research firm’s claim. *Note that history in the table below is expressed as a percentage.*

|  |  |  |
| --- | --- | --- |
|  | History (%) | Frequency (f) |
| Sunday | 7 | 47 |
| Monday | 4 | 15 |
| Tuesday | 5 | 27 |
| Wednesday | 12 | 45 |
| Thursday | 11 | 44 |
| Friday | 37 | 166 |
| Saturday | 24 | 150 |

2a) Which type of Chi-square test is appropriate for this analysis? State the null and alternate hypothesis with a word statement. (0.20 points)

*Hint: the null can either directly state the population information; if not directly stating it, then it should refer to the population distribution is some way.*

**The appropriate chi-square test for this analysis is the** **chi-square goodness of fit because it’s used to test whether a frequency distribution fits an expected distribution.**

**H0: The distribution people order food for delivery is 7% on Sundays, 4% on Mondays, 5% on Tuesdays, 12% on Wednesdays, 11% on Thursdays, 37% on Fridays, and 24% on Saturdays.**

**Ha: The distribution of days of the week that people are most likely to order food for delivery differs from the distribution seen in the past.**

2b) Perform A chi-square test in Excel. Report the degrees of freedom, X2 critical value, your calculated X2 statistic, and the corresponding p-value.   
You may express the p-value as > or < alpha; if you report the exact p-value, round to no more than 3 decimal places. *You do not need to report these values using APA.*  (0.70 points)

Use the following command in Excel to generate your p-value but replace Chi-Square with the Chi-Square Value you estimated and df with your degrees of freedom. You can enter the df number directly into the formula or click on an Excel cell that contains the df value.

**=CHISQ.DIST.RT(Chi-Square,df)**

**Degrees of freedom = 6**

**X2 critical value = 12.592**

**X2 statistic = 21.106**

**P-value = 0.002 < 0.05**

**(0.001755198)**

**The p-value is less than the alpha, which means the results are statistically significant and we can reject the null hypothesis. The data favors the alternative hypothesis.**

2c) Based on the results above what would you conclude about the research firms claim? (0.10 points)

**Based on the test results, there’s enough evidence at the 0.05 level of significance to conclude that the distribution of days that people are most likely to order food for delivery differs from the expected distribution seen in the past.**

**Question 3**

A program of pet therapy has been running for students during the week before final exams. Are the participants in the program calmer during finals than non-participants? The results from a random sample of students are reported below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Pet therapy participation | |  |
| Calmness | Participants | Non-participants | Totals |
| High level of calmness | 23 | 15 | 38 |
| Low level of calmness | 11 | 18 | 29 |
| Total | 34 | 33 | 67 |

3a) Is there a statistically significant relationship between participation in pet therapy and calmness during exams? Use Excel to calculate statistical significance. Use alpha = 0.05. Report the X2 test results using APA citation. Round your Chi-Square statistic to no more than two decimal places. (0.33 points)

**No, there is not a statistically significant relationship between pet therapy participation and calmness during exams.**

**X2 (1, n = 67) 2.52, p = 0.11266123**

3b) Compute column percentages for the table to determine the pattern of the relationship. Which group was more likely to be calm? Calculate this in Excel and create a table below to show your percentages. This does not have to be a PQ table. Round your percentages to whole numbers, no decimal places. (0.33 points)

**Those who are pet therapy are more likely to show a high level of calmness during exams.**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Pet therapy participation** | |  |
| **Calmness** | **Participants** | **Non-participants** | **Totals** |
| **High level of calmness** | **68%** | **45%** | **57%** |
| **Low level of calmness** | **32%** | **55%** | **43%** |
| **Total** | **100%** | **100%** | **100%** |

3c) We are interested to see the difference between those with high levels of anxiety. Create a PQ chart in Excel to show the pattern of the relationship using the percentages from the question above. Place a copy of the PQ chart below. (0.34 points)

Chart, bar chart

Description automatically generated

**Question 4**

There is an ever-raging debate in technical groups and across the Internet regarding the image file format .gif. Some people pronounce it with a hard ‘g’ like golf or going. Others pronounce it with a soft ‘g’ like giraffe or the name Geoffrey. A third, much smaller group pronounce it by stressing all the letters individually, “JEE EYE EFF.” Below is a portion data taken from a survey conducted on StackOverflow (a popular website for programmers and the tech-savvy) of more than 50,000 people spanning 200 countries.

The values in the table below are percentages but treat them as observed frequencies for this question. (I know the rows do not add up to 100. An ‘other’ response was not included here.)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Hard ‘g’ | Soft ‘g’ | All letters | Total |
| USA | 65 | 32 | 1 | 98 |
| Europe | 75 | 22 | 1 | 98 |
| Asia | 33 | 34 | 30 | 97 |
| Total | 173 | 88 | 32 | 293 |

4a) Demonstrate that you can recreate the table above in R. You must have the script in your script file to complete the rest of the question but copy and paste it below for credit. (0.25 points)

*Hint: There are several ways to do this. One likely approach was earlier in this homework.*

**-- R –**

**# entering data**

**gif <- matrix(c(65, 32, 1, 98, 75, 22, 1, 98, 33, 34, 30, 97, 173, 88, 32, 293), ncol=4, byrow=T)**

**# naming rows/columns**

**rownames(gif) <- c("USA", "Europe", "Asia", "Total")**

**colnames(gif) <- c("Hard 'g'", "Soft 'g'", "All letters", "Total")**

**gif**

4b) Calculate the Chi-Square statistic in both Excel and R. Determine statistical significance using alpha = 0.01. Report the results using APA. (0.50 points)

**X2 (9, n = 1172) 72.298, p = 0.000000000005385**

**5.385 x 10-12 < 0.01**

**X2 = 72.298**

**Degrees of freedom = 9**

**P-value = 5.385 x 10-12**

**-- R –**

**# chi-square test**

**chisq.test(gif)**

**sum(gif)**

4c) What can you conclude from these results? Your answer should say something about the statistical significance (or lack of) AND comment about how that finding relates back to region and pronunciation of gif. (0.25 points)

**Since the p-value is less than 0.01 (α), we know that the results are statistically significant, and we can reject the null hypothesis. In this case, the null hypothesis states there is no effect or relationship between pronunciation of “gif” and region. However, the data favors the alternative hypothesis, which states that an effect or relationship between pronunciation of “gif” and region exists.**

**Question 5**

Society changes over time and our values of work and family change as well. Using the GSS 2016 data, we will explore the relationship, or lack thereof, between gender and age groups and willingness to sacrifice good job opportunities for the benefit of family life.

Use R setup file to recode the age variable into age groups, where each age group spans five years with high and low age group caps. Most of it should look familiar but see if you can follow the process.

5a) Report the Chi-Square test results for men and women. Use APA citation to report. Are the results for both genders statistically significant? Use alpha of 0.05. (0.20 points)

**Men: X2 (27, n = 693) 43.52, p = 0.02316**

**Since the p-value is less than 0.05 (α), these results are statistically significant.**

**Women: X2 (27, n = 756) 35.01, p = 0.1386**

**The p-value is greater than 0.05 (α), which means these results are not statistically significant.**

5b) The Chi-Square results for each test should have given you a warning:

Warning message:

In chisq.test(tab.job.men) : Chi-squared approximation may be incorrect

*If you did not receive the warning, you likely did something wrong from above. Check with the instructional team.*

Why do you think you received this warning message? (0.20 points)

**I think I received this warning message because some of the expected values will be small, so the approximations may be off.**

5c) One way to resolve the error is to either eliminate some of the age groups; another strategy is to combine the age groups. We will do the second option. Instead of using 10 age groups, let’s recode them into 4 age groups. We will use quartiles to approximate 4 groups of relatively equal size for each gender.

Start with a summary of the previously created age variable:

summary(gss$age.r)

You should see a Min of 18, 1st Qtr of 34, Median & Mean of 49 (rounded), 3rd Qtr of 62 and Max of 89.

Modify the previous setup code that created 10 age groups so that it to creates 4 age groups using the following quartile information:

Group 1: 18 to 34,

Group 2: 35 to 49,

Group 3: 50 to 62, and

Group 4: 63 and older

When you are done recoding, I suggest you run the next two lines to first check your recode and see that it worked right and then see the four age group summary statistics.

table(gss$age.r, gss$age4) # confirm recode

Create a single PQ table that reports frequencies and percentages of your newly recoded four age groups with separate columns for each gender.

*Hint: if you are using the summary tools output—you only need the group categories, frequency, % valid, and totals for each of those. Be sure to not count any missing in your frequency total.* (0.20 points)

**\*\* I’m not entirely sure how to work through 5c-e, but I somewhat attempted 5c. Any guidance would be greatly appreciated. \*\***

5d) Lines 70 to 75 in the R setup file contains the code the bivariate tables and compute the Chi-Square statistic with 10 age groups. Modify it to change the old 10 age group variable to your new 4 age group variables. Calculate and report the Chi-Square test results for each gender using APA citation. Use alpha = 0.05. (0.20 points)

5e) Use R setup code to reproduce the bivariate tables by gender and change the frequencies to column percentages.

Look at the first response category of each table: “Yes, I have done so and probably would do so again.” Create a single PQ bar chart using the proportions from that response for each gender (convert them to percentages rounded to whole numbers). Use the age-groups on the x-axis, include some version of the variable response category as the title, a legend to distinguish bars for men and bars for women. Don’t worry about adding the sample size to this chart.

Then write two to four sentences about you can conclude from the Chi-Square results in 5d and this PQ chart. *One to two sentences to say if one/both tests were statistically significant or not and one to two sentences to interpret the chart.* (0.20 points)