Lab #4, Part #1 – Student Alcohol Consumption and the Chi-Square

In this lab, we will go through various ways to analyze and test categorical data in R. We will follow the lectures quite closely, starting with a test for a single proportion, a test of differences in a proportion, a 1-way table χ^2 test, and a 2-way table χ^2 test. You'll use R-Studio to both manually calculate χ^2 test statistics and apply functions to calculate and look up p-values for your test statistics. Throughout the lab, I'll ask you to explore and visualize your data with some functions we've used in the past as well.

For the lab, we'll be start by using the data set `schoolalcohol.csv`. You will use the raw data to create proportions and contingency tables for various categories and binary outcomes to answer questions in each activity below. These data include information on a sample of math class students in Portugal. The data come from here (used as a Kaggle data analysis competition data set), and I have reduced it to 7 columns (variables) and changed a few things. These include `sex`, `age`, `weekdayAlcohol` (a measure for very low to very high alcohol consumption during weekdays), `weekendAlcohol` (same measure, but for weekends), `health` (a measure representing health status as very poor to very healthy), number of `absences` from class, and number of `failures` of this class previously.

Please be sure to include all code and all visuals and tables in your submission on Canvas. Note that all responses to questions should be in complete sentences and provide details and explanations when asked.

Activity #1: Open R-Studio and Load in Data.
Go ahead and download the files (`schoolalcohol.csv`) from the Canvas Lab #4 page and save them to a new lab folder in your class folder directory (Lab 4 would be a good choice for your new folder) as .csv files. Then open up R-Studio, set your working directory with `setwd()`, and load the data. Name the data `schooldat`.

```
setwd("c:/Users/…")
schooldat <- read.csv(file = "schoolalcohol.csv", h = T)
head(schooldat)
```

Be sure that you've told R that the data has variable names (headers) with `h = T`, and confirm that they show up correctly.

Activity #2. Creating New Variables
Since we want to evaluate this data categorically, I want you to recode some variables with the `ifelse()` function For this, we'll re-evaluate absences in groups, rather than as individual counts. We'll start with 0-5 absences, 6-10 absences, 11-20 absences, and >20 absences and name these groups "`Present`", "`Somewhat Absent`", "`Absent`", and "`Very Absent`". Remember that with the `ifelse()` function, you're telling R that, if the value of `absences` is between 0 and 5, for example, then recode to a new variable called `absentCat`, and call this

observation "Present". There are other ways to recode variables, but we'll stick with this one for now. You'll have to do a new ifelse() command for each category, and note that after the first line, the "else" option is now just schooldat$absentCat (saying that, if it doesn't meet the "if" condition, then leave it as is). Use the code below, and complete the last 2 lines of it so that you have the Absent and Very Absent categories coded correctly.

```
schooldat$absentCat <- ifelse(schooldat$absences <= 5,
     "Present", NA)

schooldat$absentCat <- ifelse(schooldat$absences >= 6 &
     schooldat$absences <= 10, "Somewhat Absent",
     schooldat$absentCat)
```

Once you've done that, use the table() function to check the counts in each category. You should have 478 Present, 122 Somewhat Absent, 41 Absent, and 8 Very Absent students. We will return to this later.

```
table(schooldat$absentCat)
```

Now on your own do the same for the failures category. In this case, you'll make it binary. If a student has failed previously (a value larger than 0),then you want the new category to say Failed. If they have not, name the category First Time. Name this variable failPrev.

2.1 Include both tables you've made in your lab write up. Make sure to make it nice and neat using table functions in Microsoft Word, rather than copying and pasting from R. Given what you see in the table, describe absentee rates for the course overall in the data. Do the same for previous failures of the course among students in this sample.

Activity #3. Test for A Single Proportion.

For this activity, we'll start with testing a single proportion for two different variables. The first will be the split of male and female students in the course. This is a relatively simple question: are males and females equally represented in the class?

We want to set up the null hypothesis that tests equal representation in the class, or that 0.5 (50%) of students are female (F). Simply treat female as success (1) and male as failure (0). This is just like other tests we've done. You can just use the table() function to see the counts for each.

3.1 In your lab write-up, state the null and alternative hypotheses both with mathematical notation and in words in a complete sentence for each H_0 and H_A. Make it a two-tailed test. Here, report the proportion of Males and Females in a neat and tidy table.

Since we have a pretty big sample size here, and we'll assume sex is an independent variable, we can assume the success-failure condition holds for our data and use the normal approximation.

3.2 Given that, use R to calculate the standard error under the null, and then the z-score for the proportion. Remember that under the null, $SE = \sqrt{\frac{p_0(1-p_0)}{n}}$. Report each of these in your lab write-up and be sure to include this as part of your code submission.

Now that you have the z-score and sample size, go ahead and use the pnorm() function with the z-score you calculated. Note that it will show the area to the left of the z-score since it is positive, so you'll want to subtract it from 1 and multiply it by 2 for a two-tailed test.

3.3 Report the p-value here and explain in words what it means about the likelihood of observing the value you have in the data.

3.4 Do you reject or fail to reject the null hypothesis? Explain in a complete sentence and in the context of the data.

Activity #4. Test for Differences in Proportions.

The next topic for lab is looking at differences in proportions. For this, we'll again use the `sex` variable, but instead look at prior failure rates. In other words, we'll have a proportion of `Failure` in the `failPrev` variable for each sex (`M` and `F`).

To begin, calculated the occurrence of male and female students that have previously failed the course in the data set again with the `table()` function. Note that this time you'll include two variables – separated by a comma – in the parentheses. The result should be a table with 4 cells showing the counts for each of the categories for each `sex`.

4.1 Write the null and alternative hypotheses related to the difference in proportions for this question (two-tailed). Again, write it out in words and in complete sentences. Include this table in your lab write-up and make it neat and tidy. In the table, include the proportions of previously `Failed` for each the `F` and `M` groups.

4.2 Now use R to calculate the pooled proportion, the pooled standard error under the null, and the z-score associated with the difference in proportions. Remember that $SE = \sqrt{\frac{\hat{p}_{pooled}(1-\hat{p}_{pooled})}{n_{Male}} + \frac{\hat{p}_{pooled}(1-\hat{p}_{pooled})}{n_{Female}}}$. Since we are assuming we can use the normal approximation, use `pnorm()` to get the two-tailed p-value for your difference in proportions.

4.3 Do you reject or fail to reject the null hypothesis? Explain in a complete sentence and in the context of the data.

Activity #5. One-Way Tables and the χ^2 Test.
Now let's move onto understanding a bit more about the relationship between alcohol consumption and academic performance. We'll measure alcohol consumption first by weekday and then by weekend, and evaluate academic performance based only on whether or not the student has previously failed the course. Note here that we won't be able to say much about causation, since we don't know if alcohol caused previous failure, or if previous failure lead to more alcohol consumption now. But having an association between the two is still interesting.

Let's start with weekday alcohol consumption levels using the variable `weekdayAlcohol`. Remember that this varies from Very Low to Very High, and is self-reported. The first thing we want to do is see how previous failure rates, `failPrev`, show up across the different levels of alcohol consumption. So, let's may a one-way table with `weekdayAlcohol` as the columns and `failPrev` as the rows using the code below. For the next couple activities, we're going to name our table and save it as an object. Notice that it reports the columns in alphabetical order, but I've reported them in the table below from Very Low to Very High.

```
dayFail <- table(schooldat$failPrev, schooldat$weekdayAlcohol)
dayFail
```

	Very Low	Low	Moderate	High	Very High	Total
Failed	55	11	13	10	11	100
First Time	365	93	31	30	30	549

Note that our interest is in testing whether the `Failed` students are different from the `First Time` students. For this, we will assume the true distribution of drinking across the student population is given by the `First Time` students. We'll then base expected counts for Failed students based on these rates. So we'll need to use the rates for the last row above, and multiply by the column total in the first row.

	Very Low	Low	Moderate	High	Very High	Total
Failed	55	11	13	10	11	100
First Time Rate	0.665	0.169	0.056	0.055	0.055	1.00
Expected	66.5	16.9	5.6	5.5	5.5	100

5.1 In your lab write-up, state the null and alternative hypotheses in a complete sentence for each H_0 and H_A. Make it a two-tailed test.

5.2 Next, use R-Studio to calculate the z-scores for each of the column `Failed` counts relative to the expected counts. From there, square them and sum them up to find the χ^2 test statistic, then use `pchisq()` to find the p-value (note that we have $k = 5 - 1 = 4$ degrees of freedom). Try to do all this in R-Studio. Remember the following calculation for the test statistic:

$$\chi^2 = \frac{(ObservedCount_1 - NullCount_1)^2}{NullCount_1} + \cdots + \frac{(ObservedCount_k - NullCount_k)^2}{NullCount_k}$$

You should end up with code that looks something like the following:

```
zVL <- (55-66.5)/(sqrt(66.5))
zL <- (11-16.9)/(sqrt(16.9))
zM <- (13-5.6)/(sqrt(5.6))
zH <- (10-5.5)/(sqrt(5.5))
zVH <- (11-5.5)/(sqrt(5.5))

chiSqAll <- zVL^2 + zL^2 + zM^2 + zH^2 + zVH^2
chiSqAll

2*(1-pchisq(chiSqAll, 4))
```

5.3 Report the statistic and p-value and whether you reject or fail to reject the null hypothesis that Failed students have similar weekday drinking habits to the overall student population.

5.4 Now perform 5.1, 5.2, and 5.3 for the variable `weekendAlcohol`. Be sure to set up your null and alternative hypothesis (two-tailed) and write them out as we've done throughout lab. Rather than do the individual cell z-scores, you can just use the function for this one. Complete all other steps as above, and report the results neatly alongside a conclusion about rejecting or failing to reject the null hypothesis. Be sure to use complete sentences and report your conclusions in the context of the problem at hand.

5.5 Compare the two tests you did. Given what you see, describe differences (if any) and possible reasons for these observations.

Activity #6. Two-Way Tables and the χ^2 Test.

For this activity, we will perform 2 hypothesis tests using two-way tables and the χ^2 test. This time around, we'll introduce the `chisq.test()` function, which will do much of the work for us once we create a table.

First, create a contingency table for `absentCat` and `weekdayAlcohol`, and then do the same for `absentCat` and `weekendAlcohol`. Name these tables `dayAbsent` and `endAbsent`, respectively.

6.1 Report the tables in your lab write-up and set up a null and alternative hypothesis for each contingency table using a complete sentence for each (two-tailed).

Use the chisq.test() function to test the `absentCat` by `weekdayAlcohol` table with the following code:

```
chisq.test(dayAbsent)
```

Notice that this handy function reports the test statistic, degrees of freedom, and the p-value! Much easier this way, huh? As we continue forward in the semester, we'll begin using more complex functions like this that do a lot of the work for us. Note that R-Studio tells us with a warning message that the approximation may not be correct. The reason for this is that we have some very low counts in the `Very Absent` category. Technically, we should have combined that with the Absent category to make them a bit more populated, but we'll ignore that for today.

Now go ahead and do the same for the `endAbsent` table.

6.2 Report the test statistic, degrees of freedom, and p-value for each of your χ^2 tests. Do you reject or fail to reject each of your null hypotheses? Explain in words and in the context of the data what this means.

Activity #7. Carry Out a χ^2 Hypothesis Test with Weekday Consumption and Health.
Repeat what you did in Activity #6, but this time look at differences in weekday alcohol consumption and health.

7.1 Be sure to state the null and alternative hypotheses, report the contingency table, report your test statistic and p-value, and discuss whether you reject or fail to reject the null hypothesis as in Activity #6. Is this what you expected (note that health level is self-reported)?

7.2 To finish up, use 4-5 sentences to talk about what you found throughout this lab with respect to drinking, failure of courses, and overall health of students in Portugal from this sample.