**Exercise 4**

**Hypothesis testing of categorical data in R**

Today, the data we'll be looking at are strictly categorical such as Male/Female, Yes/No, Blue/Red/Green, Chocolate/Vanilla/Strawberry.

For categorical data, the chisq.test() in R performs both

chi-squared tests for independence and

chi squared tests for goodness-of-fit

**For each task, show me the code, results, and CLEARLY STATE THE NULL, THE ALTERNATIVE, THE P VALUE, ACCEPTANCE OR REJECTION OF THE NULL, AND THE REASON FOR ACCEPTING OR REJECTING THE NULL.**

**Task 1. Chi square test of independence**

You want to understand if males prefer certain ice cream flavours over females. You stand outside Braums and do a survey of different people coming out. You record sex and the flavor of ice cream they bought. You note it down in a notepad on your computer as shown below and save it as a CSV file called IceCream.csv

Sex, Flavor

M,C

F,V

M,C

F,C

F,C

M,V

F,V

M,V

F,S

F,S

F,S

F,S

F,S

M,S

F,S

F,V

You now go home, open up your computer and fire up R Studio. Import your CSV file into R studio. You want to learn if sex and ice-cream flavor are related. What would you do ?

chisq.test(IceCream$sex,IceCream$flavour)

Simple as that. Now you need to interpret it though. The Chi-squared distribution is only an approximation to the sampling distribution, and the approximation is not very good when the expected cell counts are too small. This is the reason for the warning. Small sample sizes are not good for Chi Square.

**Task 2 Binomial test**

Using the same dataset, we know that the population distribution of males and females in Oklahoma is 0.5 to 0.5. Is the distribution of males and females sampled in your experiment fair ?

binom.test(“number of males you get”, “number of people sampled”, 1/2)

**Task 3 Chi square test of goodness of fit**

Using the same dataset, you know expected that the distribution of ice cream flavors would be 0.33, 0.33 and 0.33 for vanilla, strawberry, and chocolate. You want to do a goodness of fit between this expected distribution and the distribution of flavors you got.

observed = c(10,14,8) # observed frequencies

expected = c(0.33, 0.33,0.34) # expected proportions

chisq.test(x = observed, p = expected)

observed = c(3,4,2) # observed frequencies

expected = c(0.33, 0.33,0.34) # expected proportions

chisq.test(x = observed, p = expected)

**Task 4 Building a contingency table and doing Chi square test of independence on it**

Use the built in data set called survey that is in the library called MASS. We will focus on only 2 categorical variables, exercise and smoking. In the built-in data set [survey](http://www.r-tutor.com/node/61), the Smoke column records the students smoking habit, while the Exer column records their exercise level. We can tally the students smoking habit against the exercise level with the table function in R. The result is called the contingency table of the two variables.

library(“MASS”)

print(survey)

print(survey$Smoke)

print(survey$Exer)

chisq.test(survey$Smoke, survey$Exer)

# to build a contingency table, do the following

mytable = table(survey$Smoke, survey$Exer)

print(mytable)

chisq.test(mytable)

**Task 5 Chi square test of independence**

titanic <- read.csv(url("<http://whitlockschluter.zoology.ubc.ca/wp-content/data/chapter09/chap09f1.1Titanic.csv>"))

# the above 3 lines should all be typed as one line in R

print(titanic)

titanicTable <- table(titanic$survival, titanic$sex)

print(titanicTable)

#adds summation in the margins of the table

addmargins(titanicTable)

plot(titanicTable, col = c("firebrick", "goldenrod1"), xlab = "Sex", ylab = "Relative frequency", main = "Titanic")

chisq.test(titanic$survival, titanic$sex)

chisq.test(titanicTable)

=====================

ChiSquare test for Goodness of fit and Chi square test for independence.

These 2 tests are not the same even though they have the same name (chi-square) and approximately the same distribution.

• The "goodness-of-fit test" is a way of determining whether a set of categorical data came from a claimed discrete distribution or not. The null hypothesis is that they did and the alternate hypothesis is that they didn't. It answers the question: are the frequencies I observe for my categorical variable consistent with my theory?

• The "test of independence" is a way of determining whether two categorical variables are associated with one another in the population, like race and smoking, or education level and political affiliation.

• Note that in the test of independence, two variables are observed for each observational unit. In the goodness-of-fit test there is only one observed variable.