Homework2

1. Freshwater guppy

A student was interested in determining whether there was a relationship between color and sex in a species of freshwater guppy. She captured 180 fish from a segment of a slow-moving stream. There were 70 red males and 30 grey males vs 25 red females and 55 grey females. Set up the appropriate null and alternate hypotheses, and run the test in R. Be sure to report your degrees of freedom, test statistic, P value, and whether you accept or reject the null.

Null hypothesis: there is no significant difference between sexes of fish based on color

Alternative hypothesis: there is a significant relationship between sex and color

Code

```
t <- cbind(c(70,30),c(25,55))

row.names(t)<-c("Red","Grey")
colnames(t)<-c("Males","Females")

row_sums <- rowSums(t)
col_sums <- colSums(t)
grand_total <- sum(t)

expected <- outer(row_sums, col_sums) / grand_total

is.table(t) # FALSE
```

```
## [1] FALSE
```

```
t<-as.table(t)
is.table(t) # TRUE
```

```
## [1] TRUE
```

```
t
```

```
##      Males Females
## Red     70      25
## Grey    30      55
```

```
chisq.test(t, corr=F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  t
## X-squared = 26.777, df = 1, p-value = 2.283e-07
```

Results

X-squared = 26.777, df = 1, p-value = 2.283e-07, n=180

Based on these results we can reject the null hypothesis and say that there is a significant relationship between sex and color

2. Individual data

Data for chi-square tests generally consists of counts of individuals within specified categories, so collecting data for chi square tests is often fairly quick and straightforward. Please collect data on ANY topic of interest to you and run either a chi-square goodness of fit test or a chi-square test for 2-way contingency tables.

Background: Data is tree species labels created by two people, Kanoa and Kamahao, with confidence values from 5=100% confident, 4=95% confident, 3=75% confident, 2=51% confident, and 1=0% confident. I'm interested in looking at if confidence differ between the two labelers.

Alternative hypothesis: Confidences differ among the two labelers

Null hypothesis: Confidences don't differ between labelers

Code

```
data <- read.csv("~/Downloads/data.csv")

data.kanoa <- data[data$created_by=="kanoa", ]
data.kamahao <- data[data$created_by=="kamahao", ]

levels <- 1:5
tab <- cbind(
  kanoa   = table(factor(data.kanoa$confidence, levels=levels)),
  kamahao = table(factor(data.kamahao$confidence, levels=levels))
)
chi <- chisq.test(tab)
```

```
## Warning in chisq.test(tab): Chi-squared approximation may be incorrect
```

```
chi$expected
```

```
##          kanoa     kamahao
## 1     6.497625   0.5023753
## 2    38.985748   3.0142518
## 3   108.603156   8.3968442
## 4   129.952494  10.0475059
## 5 10657.960977 824.0390227
```

Results

X-squared = 416.51, df = 4, p-value < 2.2e-16

Based on these results we can reject the null hypothesis and say that the confidence labeling strategies are statistically different.