Homework 2, Solutions

Problem #1

1.

```r
my.chisq.test <- function(tab){
  n1 <- nrow(tab)
  n2 <- ncol(tab)
  E <- matrix(0, n1, n2)
  ind <- as.matrix(expand.grid(1:n1,1:n2))
  E[ind] <- apply(ind, 1, function(x) return(sum(tab[x[1],]) * sum(tab[,x[2]])/sum(tab)))
  chisq <- sum((tab - E)^2/E)
  pval <- 1-pchisq(chisq, df=(n1-1)*(n2-1))
  return(c(chisq=chisq,
           pval=pval))
}
```

2.

```r
suppressMessages(library(tidyverse))
listings <- read.csv("~/Downloads/listings.csv")
```

a. We are interested in *neighborhood_group* (the NYC borough) and *room_type*.

```r
listings %>% select(neighbourhood_group, room_type) %>% head()
```

```
##   neighbourhood_group       room_type
## 1           Manhattan Entire home/apt
## 2           Manhattan    Private room
## 3            Brooklyn Entire home/apt
## 4           Manhattan Entire home/apt
## 5           Manhattan Entire home/apt
## 6            Brooklyn    Private room
```

b. Hypotheses are

$$H_0 : \{\text{Room types are independent of NYC borough}\} \;\; vs \;\; H_a : \{\text{Room types are dependent on NYC borough}\}$$

c.

```r
my.tab <- listings %>% select(neighbourhood_group, room_type) %>% table()
my.tab
```

```
##                    room_type
## neighbourhood_group Entire home/apt Private room Shared room
##       Bronx                     378          659          68
##       Brooklyn                 9565        10131         418
##       Manhattan               13054         7931         471
##       Queens                   2118         3489         204
##       Staten Island             181          187          10
```

```r
colSums(my.tab)
```

```
## Entire home/apt    Private room     Shared room
##           25296           22397            1171
```

```r
rowSums(my.tab)
```

```
##         Bronx      Brooklyn     Manhattan        Queens Staten Island
##          1105         20114         21456          5811           378
```

```r
sum(my.tab)
```

```
## [1] 48864
```

For cell $(1,1)$, Bronx borough & Entire home/apt:

$$E_{1,1} = \frac{\text{row 1 total } \times \text{ column 1 total}}{\text{total sample size}} = \frac{(378+659+68)\times(378+9565+13054+2118+181)}{48864} =$$

$$\frac{1105 \times 25296}{48864} = 572.0383$$

For cell $(5,3)$, Staten Island & Shared room:

$$E_{5,3} = \frac{\text{row 5 total } \times \text{ column 3 total}}{\text{total sample size}} = \frac{378 \times 1171}{48864} = 9.06$$

d. Proceed to apply your *my.chisq.test*() and interpret the results. As a sanity check, also run R's built-in *chisq.test*() function on that same data, make sure the outputted X^2 and p-values match with those provided by *my.chisq.test*().

```r
my.chisq.test(my.tab)
```

```
##    chisq     pval
## 1609.391    0.000
```

There's strong statistical evidence to claim dependence between NYC borough and the room type. The *chisq.test*() output confirms it.

```r
chisq.test(my.tab)
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  my.tab
## X-squared = 1609.4, df = 8, p-value < 2.2e-16
```

e. In case you end up claiming that variables are not independent, proceed to make a few comments on **strength** of the relationship (as was done for Income & Happiness example in class).

```
prop.table(my.tab, margin = 1)
```

```
##                     room_type
## neighbourhood_group Entire home/apt Private room Shared room
##       Bronx               0.34208145   0.59638009  0.06153846
##       Brooklyn            0.47553943   0.50367903  0.02078155
##       Manhattan           0.60840790   0.36964019  0.02195190
##       Queens              0.36448116   0.60041301  0.03510583
##       Staten Island       0.47883598   0.49470899  0.02645503
```

For example, Manhattan has a really high % of entire home/apt - 60%, which is at least 13% higher than in any other borough. Meanwhile, Queens and Bronx mostly provides private rooms - 60%, which is at least 10% higher than in other boroughs. Moreover, the proportion of shared rooms in Bronx (6%) is at least 1.5-2 times higher than those in other boroughs.

Problem #2

1.

11.84 Degrees of freedom explained

a) The order of the calculations is given in the table.

Political Views	Vote for Female President: Yes	No	Total
Extremely Liberal	56	1st: 58 – 56 = 2	58
Moderate	490	2nd: 509 – 490 = 19	509
Extremely Conservative	4th: 61 – 3 = 58	3rd: 24 – 2 – 19 = 3	61
Total	**604**	**24**	**628**

b) The order of the calculations is given in the table.

Political Views	Vote for Female President: Yes	No	Total
Extremely Liberal	2nd: 58 – 2 = 56	1st: 24 – 3 – 19 = 2	58
Moderate	3rd: 509 – 19 = 490	19	509
Extremely Conservative	4th: 61 – 3 = 58	3	61
Total	**604**	**24**	**628**

2.

11.9 Happiness and gender

a) H_0: Gender and happiness are independent.
H_a: Gender and happiness are dependent.

b) If the null hypothesis of independence is true, it is not unusual to observe a chi-square value of 1.04 or larger because the probability is 59% of this occurring. Hence, there is no evidence of an association between gender and happiness.

Expected counts:

- For $(1,1)$ cell, with $Gender = Female, Happiness = Not$:

$$E_{1,1} = \frac{\text{(row 1 total) x (column 1 total)}}{\text{total sample size}} = \frac{(154+592+336)\times(154+123)}{1964} = \frac{1082\times 277}{1964} = 152.6039$$

- For $(2,3)$ cell, with $Gender = Male, Happiness = Very$:

$$E_{2,3} = \frac{\text{(row 2 total) x (column 3 total)}}{\text{total sample size}} = \frac{(123+502+257)\times(336+257)}{1964} = \frac{882\times 593}{1964} = 266.3065$$

which are both pretty close to the observed counts, confirming the χ^2-test results of non-significance.

3.

11.16 Primary food choice of alligators

a)

	Primary Food					
Lake	**Fish**	**Invertebrates**	**Birds & Reptiles**	**Other**	**Total**	***n***
Hancock	54.5%	7.3%	14.6%	23.6%	100%	55
Trafford	24.5%	34.0%	22.6%	18.9%	100%	53

b) H_0: The distribution of primary food choice is the same for alligators caught in lakes Hancock and Trafford (homogeneity).

H_a: The distributions differ for the two lakes.

c) $df = (2-1)(4-1) = 3$, so we expect the chi-squared statistic to be 3 with a standard deviation of $\sqrt{2df} = \sqrt{2(3)} = \sqrt{6} = 2.5$. Since 16.79 is $(16.79-3)/2.5 = 5.5$, or 5.5 standard deviations about the expected value of 3, it is considered extreme.

d) Since the P-value is less than 0.001, there is strong evidence that the distribution of primary food choice of alligators differs in the two lakes.

Problem #3.

1. a.

```
tab.1 <- matrix(c(0.51, 0.49, 0.49, 0.51)*100, 2,2)
tab.2 <- matrix(c(0.51, 0.49, 0.49, 0.51)*200, 2,2)
tab.3 <- matrix(c(0.51, 0.49, 0.49, 0.51)*10000, 2,2)
tab.1
```

```
##      [,1] [,2]
## [1,]   51   49
## [2,]   49   51
```

```
tab.2
```

```
##      [,1] [,2]
## [1,]  102   98
## [2,]   98  102
```

```
tab.3
```

```
##      [,1] [,2]
## [1,] 5100 4900
## [2,] 4900 5100
```

```r
my.chisq.test(tab.1)
```

```
##     chisq      pval
## 0.0800000 0.7772974
```

```r
my.chisq.test(tab.2)
```

```
##     chisq      pval
## 0.1600000 0.6891565
```

```r
my.chisq.test(tab.3)
```

```
##       chisq        pval
## 8.000000000 0.004677735
```

b.
$$p_1 = \{\text{proportion of females going to religious services weekly}\}$$
$$p_2 = \{\text{proportion of males going to religious services weekly}\}$$

Then
$$\text{Difference in proportion} : p_1 - p_2 = 0.51 - 0.49 = 0.02$$
$$\text{Risk ratio} : \frac{p_1}{p_2} = \frac{0.51}{0.49} = 1.0408$$

c. As n increases, we can clearly see that the results become more **statistically significant**, while not changing at all from the **practical significance** standpoint. That's why for large sample sizes, it is critical to pay attention to both statistical significance and the practical effect size.

2. Solution below:

11.32 Marital happiness

a) Since the P-value is less than 0.001, there is strong evidence for an association between marital and general happiness.

b) No, not generally. Large X^2 values can occur even for weak (but still significant) associations.

c) The percentage of being not too happy is 20/588 – 11/26 = 0.389, or about 40 percentage points higher for those who are not too happy in their marriage compared to those who are very happy in their marriage.

d) Those who are not too happy in their marriage are about $(11/26)/(20/588) = 11.8$, or about 12 times as likely to be not too happy compared to those who are very happy in their marriage.