STAT 477/STAT 577
HW 4 - Module 2: Sections 1 through 3

1. (39 pts) The Women's Health Initiative conducted a randomized experiment to see if hormone therapy was helpful for post-menopausal women. The women were randomly assigned to receive the estrogen plus progestin hormone therapy or a placebo. After 5 years, the number of women who developed cancer in each group was determined. The data can be found in the **WHI.csv** file in Canvas.

   (a) (6 pts; 4 pts for table and 1 pt for each proportion) Obtain a contingency table of the two variables. What proportion of women developed cancer in each group?

```
whi.data<- read.csv(file.choose(), header = T)
```

We will choose to have the Hormone group be group 1 and the placebo group be group 2. We will also choose to have the Cancer = Yes outcome be the category of interest.

```
whi.data$Group<- factor(whi.data$Group,
                        levels = c("Hormone", "Placebo"))
whi.data$Cancer<- factor(whi.data$Cancer,
                         levels = c("Yes", "No"))
```

The contingency table is:

```
whi.table<- table(whi.data$Group, whi.data$Cancer)
whi.table

##
##              Yes   No
##   Hormone    107 8399
##   Placebo     88 8014
```

The proportion of women in the Hormone group with Cancer is:

```
whi.table[1,1]/(whi.table[1,1]+whi.table[1,2])

## [1] 0.01257936
```

and the proportion of women in the Placebo group with Cancer is:

```
whi.table[2,1]/(whi.table[2,1]+whi.table[2,2])

## [1] 0.01086152
```

   (b) (12 pts) Conduct a hypothesis test to determine if the proportion of women with cancer is different between the two groups.

Let p_1 be the proportion of women in the population taking hormone replacement therapy who get cancer and let p_2 be the proportion of women in the population taking a placebo who get cancer. The null and alternative hypotheses are:

$$H_0 : p_1 - p_2 = 0$$
$$H_a : p_1 - p_2 \neq 0$$

```
prop.test(whi.table[,1], margin.table(whi.table, 1),
          alternative = "two.sided", correct = F)

##
##  2-sample test for equality of proportions without continuity correction
##
## data:  whi.table[, 1] out of margin.table(whi.table, 1)
## X-squared = 1.0553, df = 1, p-value = 0.3043
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.001553781  0.004989461
## sample estimates:
##     prop 1     prop 2
## 0.01257936 0.01086152
```

The test statistic $X^2 = 1.0553$ (or you can use $z = \sqrt{1.0553} = 1.027$ with p-value = 0.3043. We do not evidence to conclude the proportion of women in the population who get cancer is different between the two groups.

(c) (7 pts; 3 pts for interval and 4 pts for interpretation) Calculate a 90% confidence interval for the difference in the proportion of women with cancer between the hormone therapy group and the placebo group. Interpret this confidence interval.

```
prop.test(whi.table[,1], margin.table(whi.table, 1),
          alternative = "two.sided", conf.level = 0.9, correct = F)

##
##  2-sample test for equality of proportions without continuity correction
##
## data:  whi.table[, 1] out of margin.table(whi.table, 1)
## X-squared = 1.0553, df = 1, p-value = 0.3043
## alternative hypothesis: two.sided
## 90 percent confidence interval:
##  -0.001027791  0.004463471
## sample estimates:
##     prop 1     prop 2
## 0.01257936 0.01086152
```

The 90% confidence interval for the difference between p_1 and p_2 is (-0.0010, 0.0045). We are 90% confident the proportion of women in the population taking hormone replacement therapy who get cancer is between 0.0010 less than to 0.0045 more than the proportion of women in the population taking a placebo who get cancer.

(Note: since the interval includes 0, it would also be reasonable to interpret the confidence interval by not interpreting the interval.)

(d) (7 pts; 3 pts for interval and 4 pts for interpretation) Calculate a 90% confidence interval for the relative risk of developing cancer when taking hormone therapy and interpret this confidence interval.

```
rr.ci(whi.table[,1], margin.table(whi.table, 1),
      conf.level = 0.9)

## Estimated Relative Risk =  1.158158
## Confidence Interval for Population Relative Risk =  0.915308 1.465442
```

The 90% confidence interval for the population relative risk is (0.9153, 1.4654). We are 90% confident the proportion of women taking hormone therapy and who get cancer is between 0.9153 and 1.4654 times the proportion of women taking a placebo and who get cancer.

(e) (7 pts; 3 pts for interval and 4 pts for interpretation) Calculate a 90% confidence interval for the odds ratio of developing cancer when taking hormone therapy and interpret this confidence interval.

```
or.ci(whi.table[,1], margin.table(whi.table, 1),
      conf.level = 0.9)

## Estimated Odds Ratio =  1.160173
## Confidence Interval for Population Odds Ratio =  0.9143634 1.472065
```

The 90% confidence interval for the population odds ratio is (0.9144, 1.4721). We are 90% confident the odds of a woman from this population who takes hormone therapy developing cancer is between 0.9144 to 1.4721 times the odds of a woman from this population who takes a placebo developing cancer.

2. (29 pts) On the night of April 14, 1912, the luxury liner *RMW Titanic* hit an iceberg and sank in the North Atlantic Ocean. In the popular movie from 1997 about this disaster, first class passengers appeared to be able to get to the life boats, while third class passengers were kept away. Is there truth to this appearance? Was the proportion of passengers rescued different for each class of ticket? The data containing information about the number of people with each class of ticket, including crew, and whether or not the person was rescued or lost can be found in the **titanic.csv** file in Canvas.

   (a) (8 pts; 5 pts for graph and 3 pts for interpretation) Obtain a mosaic plot that compares the proportion of passengers rescued among the four ticket classes. Interpret the mosaic plot.

   Read in the data:

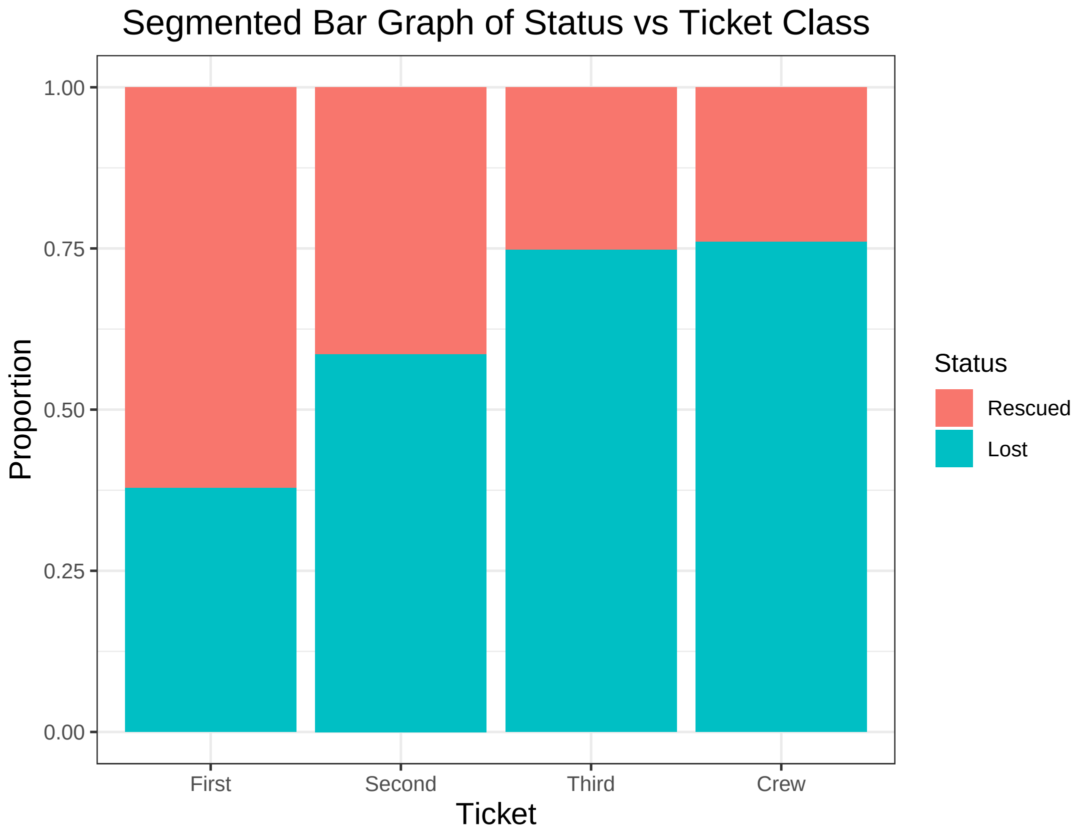
```
titanic.data<- read.csv(file.choose(), header = T)
```

Next, we will order the categories of the two variables: Ticket and Status.

```
titanic.data$Ticket<- factor(titanic.data$Ticket,
             levels = c("First", "Second", "Third", "Crew"))
titanic.data$Status<- factor(titanic.data$Status,
             levels = c("Rescued", "Lost"))
```

Now we will obtain the segmented bar graph to view the conditional distribution of Status given Ticket.

```
  ggplot(titanic.data, aes(x = Ticket, fill = Status))+
    geom_bar(position = "fill")+
    theme_bw()+
    theme(axis.title.y = element_text(size = rel(1.2)),
          axis.title.x = element_text(size = rel(1.2)),
          axis.text.x = element_text(size = rel(1)),
          axis.text.y = element_text(size = rel(1)),
          plot.title = element_text(hjust=0.5, size = rel(1.4)))+
    labs(y = "Proportion",
         title = "Segmented Bar Graph of Status vs Ticket Class")
```



We can see the proportion of rescued passengers was the largest in first class, then second class, while third class and the crew had about the same, much lower, proportion of rescued passengers.

(b) (12 pts) Conduct a hypothesis test to determine if the proportion of passengers rescued was the same across all ticket classes.

To start, we will need to obtain the contingency table for the two variables.

```
titanic.table<- table(titanic.data$Ticket, titanic.data$Status)
titanic.table

##
##          Rescued Lost
##   First      202  123
##   Second     118  167
##   Third      178  528
##   Crew       212  673
```

Let p_1 be the proportion of passengers rescued from first class, p_2 from second class, p_3 from third class, and p_4 from the crew.

$H_0 : p_1 = p_2 = p_3 = p_4$

$H_a :$ at least one p_i is different for $i = 1, 2, 3, 4$

```
titanic.test<- prop.test(titanic.table[,1],
                         margin.table(titanic.table, 1),
                         correct = F)
titanic.test

##
##  4-sample test for equality of proportions without continuity correction
##
## data:  titanic.table[, 1] out of margin.table(titanic.table, 1)
## X-squared = 187.79, df = 3, p-value < 2.2e-16
## alternative hypothesis: two.sided
## sample estimates:
##    prop 1    prop 2    prop 3    prop 4
## 0.6215385 0.4140351 0.2521246 0.2395480
```

Test Statistic: $X^2 = 187.79$

p-value < 0.0001

Conclusion: We have extremely strong evidence at least one of the ticket classes has a different proportion of rescued passengers.

(c) (7 pts) Determine the pairwise hypothesis tests for the proportion of passengers rescued for the four ticket classes. Which class(es) appear to have a significantly different proportion rescued?

The pairwise hypothesis tests have the following p-values:

```
pairwise.prop.test(titanic.table[,1],
                   margin.table(titanic.table, 1),
                   p.adjust.method = "BH")
```

```
##
##  Pairwise comparisons using Pairwise comparison of proportions
##
## data:  titanic.table[, 1] out of margin.table(titanic.table, 1)
##
##        First   Second  Third
## Second 7.0e-07 -       -
## Third  < 2e-16 8.3e-07 -
## Crew   < 2e-16 3.9e-08 0.6
##
## P value adjustment method: BH
```

All pairs of proportions have very small p-values except for the pair of third class and crew. This indicates we have extremely strong evidence the proportion of rescued passengers is different between first and second class, between first and third class, between second and third class, and between the crew and first class, and the crew and second class.

(d) (2 pts) Was the movie correct: Did the proportion of passengers rescued differ among ticket classes?

Yes, the proportion of passengers rescued differed among the ticket classes.

3. (32 pts) In 1996, in the General Social Survey of 1,895 adults in the United States conducted by the National Opinion Research Center, respondents were asked about their attitudes towards premarital sex. The question asked was **When is premarital sex wrong?** and the possible answers were **Always Wrong**, **Almost Always Wrong**, **Sometimes Wrong**, **Not Wrong at All**. People's attitudes about social behaviors tend to be related to other more general background variables about the individual. Among other questions, respondents were asked about one such variable, their religious affiliation. Possible answers were **Catholic**, **Protestant**, **Jewish**, **Other**, **None**. The data can be found in the **GSS.csv** file in Canvas.

Read in the data.

```
GSS.data<- read.csv(file.choose(), header = T)
```

Set the order of the categories for the two variables.

```
GSS.data$Religion<- factor(GSS.data$Religion,
                           levels = c("Catholic", "Protestant",
                                      "Jewish", "Other", "None"))
```

```
GSS.data$Wrong<- factor(GSS.data$Wrong,
                        levels = c("Always", "Almost.Always",
                                   "Sometimes", "Never"))
```

Obtain the contingency table for the two variables.

```
GSS.table<- table(GSS.data$Religion, GSS.data$Wrong)
GSS.table

##
##               Always Almost.Always Sometimes Never
##   Catholic        62            37       120   226
##   Protestant     355           117       227   384
##   Jewish           0             3        14    34
##   Other           15            13        23    40
##   None            20            13        45   147
```

(a) (4 pts) Calculate the conditional distribution of attitude towards premarital sex given religious affiliation is Catholic.
The conditional distribution is the first row in the following table:

```
round(prop.table(GSS.table, 1), 4)

##
##                Always Almost.Always Sometimes  Never
##   Catholic    0.1393        0.0831    0.2697 0.5079
##   Protestant  0.3278        0.1080    0.2096 0.3546
##   Jewish      0.0000        0.0588    0.2745 0.6667
##   Other       0.1648        0.1429    0.2527 0.4396
##   None        0.0889        0.0578    0.2000 0.6533
```

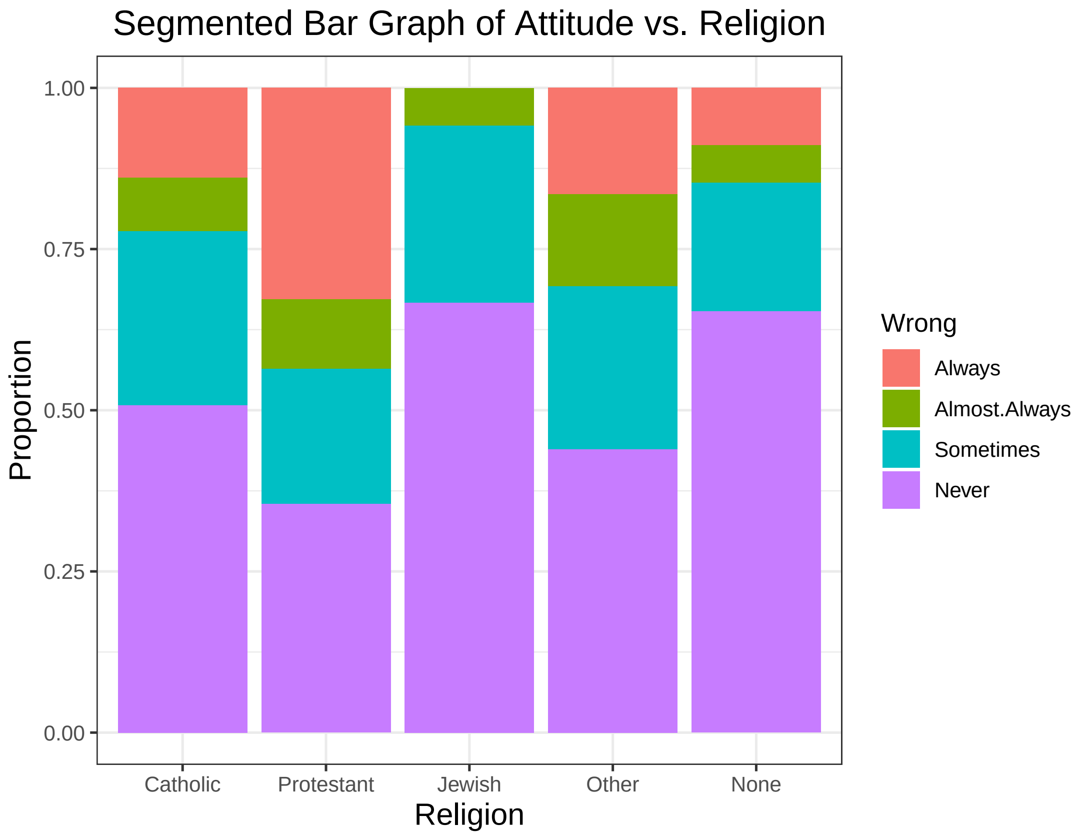
(b) (4 pts) Calculate the conditional distribution of attitude towards premarital sex given religious affiliation is Protestant.
The conditional distribution is the second row in the following table:

```
round(prop.table(GSS.table, 1), 4)

##
##                Always Almost.Always Sometimes  Never
##   Catholic    0.1393        0.0831    0.2697 0.5079
##   Protestant  0.3278        0.1080    0.2096 0.3546
##   Jewish      0.0000        0.0588    0.2745 0.6667
##   Other       0.1648        0.1429    0.2527 0.4396
##   None        0.0889        0.0578    0.2000 0.6533
```

(c) (10 pts; 4 pts for graph and 6 pts for interpretation) Obtain a mosaic plot that compares the attitudes towards premarital sex among the give religious affiliation groups. Interpret the mosaic plot.

```
ggplot(GSS.data, aes(x = Religion, fill = Wrong))+
  geom_bar(position = "fill")+
  theme_bw()+
  theme(axis.title.y = element_text(size = rel(1.2)),
        axis.title.x = element_text(size = rel(1.2)),
        axis.text.x = element_text(size = rel(1)),
        axis.text.y = element_text(size = rel(1)),
        plot.title = element_text(hjust=0.5, size = rel(1.4)))+
  labs(y = "Proportion",
       title = "Segmented Bar Graph of Attitude vs. Religion")
```



Interpretation: The heights of the Almost.Always and Sometimes categories are similar in each of the religion categories, but the Always and Never categories are very different among the religion categories.

(d) (12 pts) Conduct a hypothesis test to determine if attitudes towards premarital sex is the same for all five groups.

H_0: the distribution of attitudes towards premarital sex are the same for each religious affiliation.

H_a: at least one of the religious affiliations has a different distribution of attitudes towards premarital sex.

```
GSS.test<- chisq.test(GSS.table)

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

GSS.test

##
##  Pearson's Chi-squared test
##
## data:  GSS.table
## X-squared = 157.02, df = 12, p-value < 2.2e-16
```

Test Statistic: $X^2 = 157.02$

p-value < 0.0001

Conclusion: We have extremely strong evidence at least one of the religious affiliations has a different distribution of attitudes towards premarital sex.

(e) (2 pts) In conducting the hypothesis test, you will find at least one of the cells in the table has an expected value less than 5. Identify the cell(s).

```
GSS.test$expected

##
##                  Always Almost.Always Sometimes     Never
##   Catholic   106.14248     42.973615 100.74142 195.14248
##   Protestant 258.31979    104.585224 245.17520 474.91979
##   Jewish      12.16464      4.925066  11.54565  22.36464
##   Other       21.70554      8.787863  20.60106  39.90554
##   None        53.66755     21.728232  50.93668  98.66755
```

The cell of Jewish - Almost Always has an expected value of 4.925.