STAT 477/STAT 577
HW 3 - Solutions

1. Many people are superstitious about the number 13. But does triskaidekaphobia (fear of the number 13) have economic implications for large high-rise hotels? A USA Today/Gallup poll conducted February 9-11, 2007 asked 1,006 randomly selected people 18 years old and older in telephone interviews "Suppose you checked into a hotel and were given a room on the thirteenth floor. Would this bother you or not?" Their responses are given in the file **floor13.csv** in Canvas.

   (a) (10 pts) Use R to give the summary table and bar graph of the sample data.

   Read in the data:

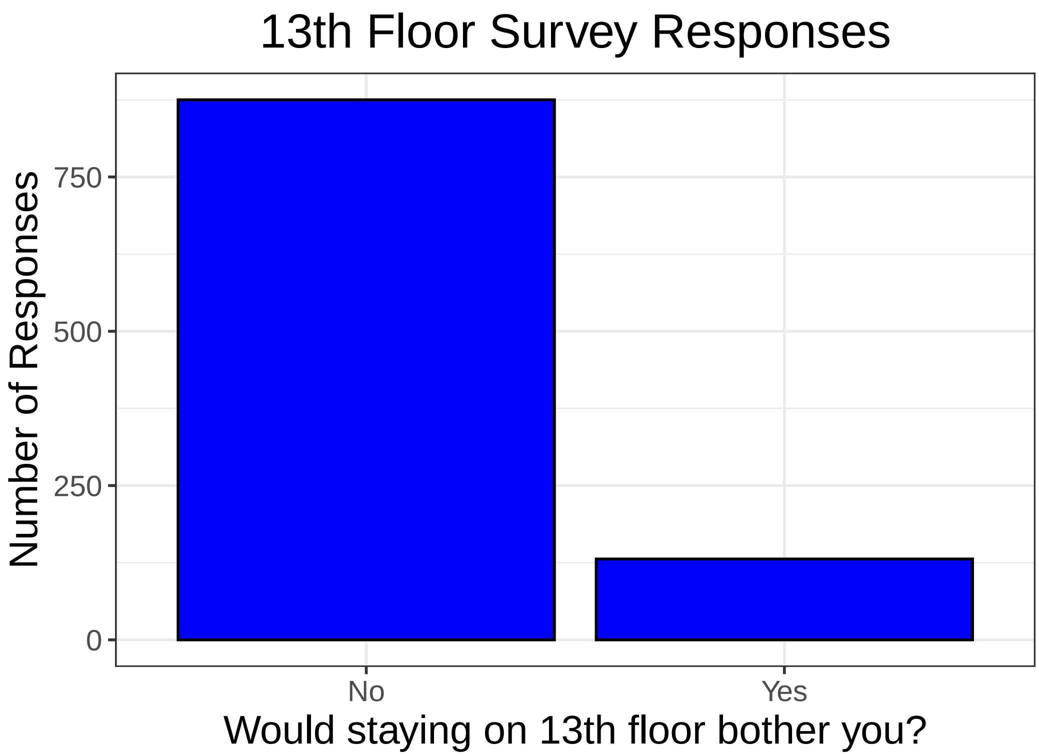
```
floordata <- read.csv(file.choose(), header = T)
```

```
floor.counts<- count(floordata, var = 'Bothered')
floor.table<- mutate(floor.counts,
                    prop = freq/sum(floor.counts[2]))
floor.table<- rbind(floor.table, data.frame(Bothered='Total',
                                t(colSums(floor.table[, -1]))))
floor.table

##   Bothered freq      prop
## 1       No  875 0.8697813
## 2      Yes  131 0.1302187
## 3    Total 1006 1.0000000
```

   Make the bar graph for the survey variable:

```
ggplot(floordata, aes(x=Bothered))+
  geom_bar(fill = "blue", colour = "black")+
  labs(x = "Would staying on 13th floor bother you?",
       y = "Number of Responses",
       title = "13th Floor Survey Responses")+
    theme_bw()+
    theme(axis.title.y = element_text(size = rel(1.4)),
          axis.title.x = element_text(size = rel(1.4)),
          axis.text.x = element_text(size = rel(1.2)),
          axis.text.y = element_text(size = rel(1.2)),
          plot.title = element_text(hjust=0.5, size = rel(1.6)))
```



(b) (6 pts) Our category of interest is being bothered by staying on the 13th floor. Use R to calculate a 95% confidence interval for the population proportion p using the normal approximation method.

Using the `prop.ci()` function, we get

```
prop.ci(131, 1006, type = "normal", conf.level = 0.95)

## 0.1094222 0.1510152
```

The 95% confidence interval is (0.1094, 0.1510).

(c) (5 pts) Give the interpretation of the 95% confidence interval you calculated in part (b) in context.

We are 95% confident the proportion of adults in the United States you would be bothered by staying on the 13th floor is between 0.1094 and 0.1510.

(d) (6 pts) Use R to calculate a 95% confidence interval for the population proportion p using Wilson's score method. Compare this interval to the one you calculated in part (b).

Using the `prop.ci()` function, we get

```
prop.ci(131, 1006, type = "score", conf.level = 0.95)

## 0.1108208 0.1524299
```

The 95% confidence interval is (0.1108, 0.1524).

(e) (6 pts) If the USA Today/Gallup Poll were conducted again using this question, what sample size would be needed in order to **guarantee** a 90% confidence interval would have a margin of error of **no more than** 2%?

In order to guarantee the 90% confidence interval will have a margin of error of no more than 2%, we will need to use the worst case scenario formula and use $p = 0.5$.

```
nprop.ci(0.5, 0.02, 0.9)

## [1] 1691
```

So the sample size should be 1,691.

2. According to a Gallup poll, of 1,019 randomly selected adults aged 18 or older in the United States, 662 believe that global warming is more a result of human actions than natural causes.

(a) Describe the population proportion of interest p in words.

The population proportion of interest is the proportion of adults in the United States who believe that global warming is more a result of human actions than natural causes.

(b) Give the value of the sample proportion $\hat{p}$.

$\hat{p} = 662/1019 = 0.6497$

(c) Calculate a 95% confidence interval for the population proportion of interest using Wilson's score method.

```
prop.ci(662, 1019, type = "score", conf.level = 0.95)

## 0.6198521 0.6783369
```

The 95% confidence interval is (0.6199, 0.6783).

(d) Give the interpretation of the 95% confidence interval you calculated in part (c) in context.

We are 95% confident the proportion of adults in the United States who believe that global warming is more a result of human actions than natural causes is between 0.6199 and 0.6783.

(e) Gallup is planning to conduct another poll on global warming. They would like to have a 95% confidence interval with a margin of error of no more than 2.5%. What sample size do they need to obtain this margin of error?

In order for the 95% confidence interval to have a margin of error of no more than 2.5%, we will need to use the worst case scenario formula and use $p = 0.5$.

```
nprop.ci(0.5, 0.025, 0.95)

## [1] 1537
```

So the sample size should be 1,537.

3. (10 pts) In lecture, we discussed issues with the confidence interval for the population proportion p. Unlike confidence intervals for other parameters, several methods have been developed for calculating this confidence interval. For confidence intervals, you want to have a coverage rate, the percentage of confidence intervals containing the true population parameter from a large number of samples, close to the stated confidence level for the confidence intervals. For example, if you generate 100,000 samples from a population and calculate a 95% confidence interval from each sample's data, you want approximately 95,000 of the 100,000 confidence intervals (or 95%) to contain the population parameter.

   In this problem, we will study the coverage rate of 95% confidence intervals for the two methods from lecture: the normal approximation and Wilson's score method. Since the binomial distribution depends on the sample size n and the population proportion p, I simulated 100,000 samples from the binomial distribution with each combination of three values of sample size $n = 25, 500, 1000$ and three values of probability of success $p = 0.05, 0.25, 0.5$, for a total of 9 conditions. For each of the 100,000 simulated samples, I determined whether or not the value of p is located within the confidence interval and used this information to calculate the coverage rate. The table below contains the coverage rates for each of these 9 trials for the two methods.

	Normal Approximation			Score		
p	$n = 25$	$n = 500$	$n = 1000$	$n = 25$	$n = 500$	$n = 1000$
0.05	72.133%	93.283%	94.091%	96.664%	94.938%	95.027%
0.25	89.397%	94.312%	94.559%	93.853%	94.386%	94.813%
0.50	95.695%	94.598%	94.645%	95.661%	94.620%	94.642%

   (a) (4 pts) Does the Normal approximation method for calculating the confidence interval for p have any coverage rates different than the expected 95%? If so, for which combinations of n and p?

   There are 2 coverage rates noticeably different from 95%: the ones for $n = 25$ and $p = 0.05$ and $p = 0.25$.

   (b) (3 pts) Does Wilson's score method for calculating the confidence interval for p have any coverage rates different than the expected 95%? If so, for which combinations of n and p?

   No, all of the coverage rates are within 2% of 95%.

   (c) (3 pts) In this simulation, I included two values of p below 0.5. Would we gain any new information by adding the values of $p = 0.75$ and $p = 0.95$ to this simulation? Explain your answer.

   No, since the binomial distributions for p and $1 - p$ are mirror images of each other, we would get the same information from $p = 0.25$ as $p = 0.75$ and from $p = 0.05$ as $p = 0.95$.

4. (32 pts) In astrology, people are assigned one of 12 zodiac signs based on their birthday. For example, my birthday is May 26 and so I am assigned the zodiac sign Gemini, which is for people born on May 21 through June 20. Does your zodiac sign have any additional meaning for, influence on, or ability to predict your life path? In one small study, Fortune magazine collected the zodiac signs of 265 heads of the largest 400 companies. The data are given in the file **zodiac.csv** in Canvas. Based on these data, does it appear that some zodiac signs are more likely to be represented in heads of these types of companies than others?

   (a) (10 pts) Use R to give the summary table and bar graph of the sample data.

   Read in the data:

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

   I am going to order the signs by the appearance in the calendar. (This is not required).

```
signdata$Sign<- factor(signdata$Sign, levels = c("Aries", "Taurus",
                          "Gemini", "Cancer", "Leo", "Virgo",
                          "Libra", "Scorpio", "Sagittarius", "Capricorn",
                          "Aquarius", "Pieces"))
```

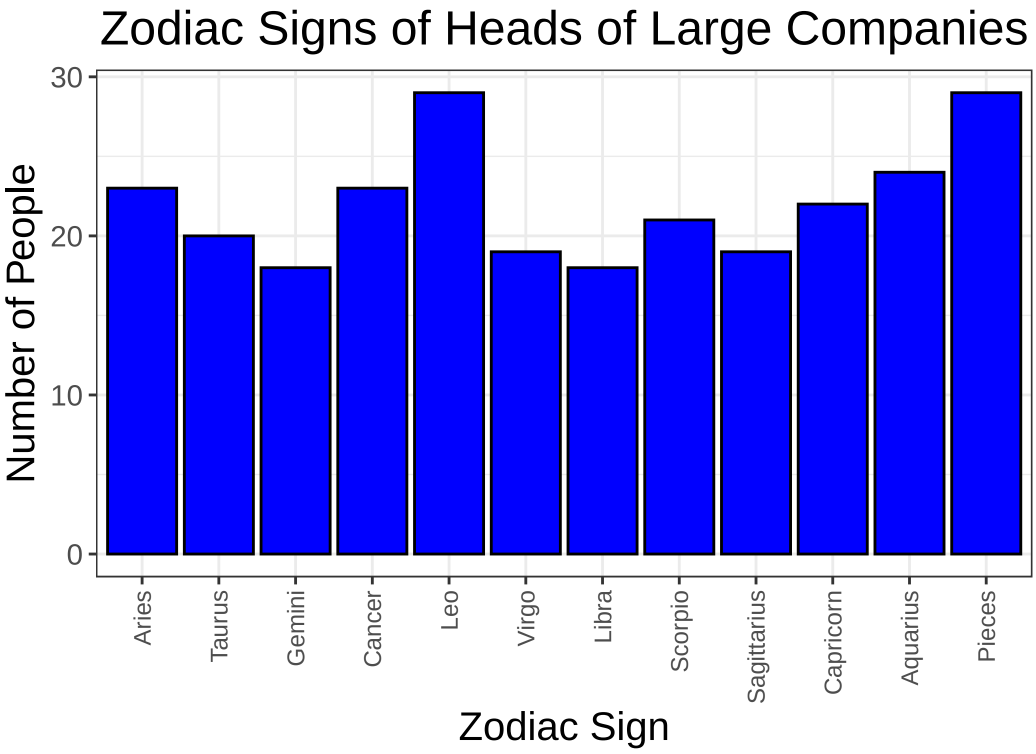
   Obtain the summary table:

```
sign.counts<- count(signdata, var = 'Sign')
sign.table<- mutate(sign.counts,
                    prop = freq/sum(sign.counts[2]))
sign.table<- rbind(sign.table, data.frame(Sign='Total',
                                t(colSums(sign.table[, -1]))))
sign.table

##           Sign freq       prop
## 1        Aries   23 0.08679245
## 2       Taurus   20 0.07547170
## 3       Gemini   18 0.06792453
## 4       Cancer   23 0.08679245
## 5          Leo   29 0.10943396
## 6        Virgo   19 0.07169811
## 7        Libra   18 0.06792453
## 8      Scorpio   21 0.07924528
## 9  Sagittarius   19 0.07169811
## 10   Capricorn   22 0.08301887
## 11    Aquarius   24 0.09056604
## 12      Pieces   29 0.10943396
## 13       Total  265 1.00000000
```

   Make the bar graph:

```r
ggplot(signdata, aes(x=Sign))+
  geom_bar(fill = "blue", colour = "black")+
  labs(x = "Zodiac Sign",
       y = "Number of People",
       title = "Zodiac Signs of Heads of Large Companies")+
    theme_bw()+
    theme(axis.title.y = element_text(size = rel(1.4)),
          axis.title.x = element_text(size = rel(1.4)),
          axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1),
          axis.text.y = element_text(size = rel(1.2)),
          plot.title = element_text(hjust=0.5, size = rel(1.6)))
```



(b) (6 pts; 3 pts each) Give the null and alternative hypotheses for answering the question posed. Assume each zodiac sign covers an equal number of birthdays in the year.

If there is no association between zodiac sign and being the head of this type of company, the probability of each zodiac sign is the same = 1/12.

$H_0 : p_1 = p_2 = \cdots = p_{12} = 1/12$

H_a : at least one $p_j \neq 1/12$

(c) (2 pts) Calculate the expected number of births in each zodiac sign under the null hypothesis.

```r
modelp<- c(rep(1/12, 12))
signgood.test<- chisq.test(sign.counts[2], p = modelp)
```

The expected value is $E(Y_j) = np_j = 265(1/12) = 22.083$ and can also be found in R with the output of the goodness of fit test variable:

```r
signgood.test$expected

##  [1] 22.08333 22.08333 22.08333 22.08333 22.08333 22.08333 22.08333 22.08333
##  [9] 22.08333 22.08333 22.08333 22.08333
```

(d) (4 pts) Calculate the contribution of the category Scorpio to the test statistic X^2. Only calculate this value for the category Scorpio, not for all the other categories.

We can calculate this value using the expected value above and the summary table information. This is:

$$\frac{(21 - 22.0833)^2}{22.0833} = 0.0531$$

You can also find this in the R output of the goodness of fit test variable:

```r
(signgood.test$residuals)^2

##  [1] 0.0380503145 0.1965408805 0.7550314465 0.0380503145 2.1663522013
##  [6] 0.4305031447 0.7550314465 0.0531446541 0.4305031447 0.0003144654
## [11] 0.1663522013 2.1663522013
```

The Scorpio category is the 8th in the list.

(e) (2 pts) Determine the value of the test statistic X^2 for this hypothesis test.

```r
signgood.test

##
##  Chi-squared test for given probabilities
##
## data:  sign.counts[2]
## X-squared = 7.1962, df = 11, p-value = 0.783
```

From the output of the goodness of fit test, the test statistic $X^2 = 7.1962$.

(f) (1 pt) What is the number of degrees of freedom for the test statistic X^2?

There are $J = 12$ categories, so the degrees of freedom is $J - 1 = 11$.

(g) (2 pts) Determine the p-value for this hypothesis test.

From the output of the goodness of fit test, the p-value is 0.783.

(h) (5 pts) Write a conclusion for this hypothesis test.

There is no evidence of lack of model fit. Some zodiac signs are no more likely to be represented in heads of these types of companies than others.