Chapter 6 - Inference for Categorical Data

```
library(tidyverse)
library(openintro)
```

Load functions I made for inference (ill include the code the first time I use one)

2010 Healthcare Law. (6.48, p. 248) On June 28, 2012 the U.S. Supreme Court upheld the much debated 2010 healthcare law, declaring it constitutional. A Gallup poll released the day after this decision indicates that 46% of 1,012 Americans agree with this decision. At a 95% confidence level, this sample has a 3% margin of error. Based on this information, determine if the following statements are true or false, and explain your reasoning.

(a) We are 95% confident that between 43% and 49% of Americans in this sample support the decision of the U.S. Supreme Court on the 2010 healthcare law.

```
se<-standard_error(.46,1012)
confidence_interval95(.46,se)
```

```
## 95% confidence interval is 0.4292927 - 0.4907073
```

Based on my function, (that I will include later, I missed this question when I first did the homework), this is true.

(b) We are 95% confident that between 43% and 49% of Americans support the decision of the U.S. Supreme Court on the 2010 healthcare law.

True, this also falls within the 95% confidence range

(c) If we considered many random samples of 1,012 Americans, and we calculated the sample proportions of those who support the decision of the U.S. Supreme Court, 95% of those sample proportions will be between 43% and 49%.

FALSE, a lower percentage would fall in that range, since the limits are tighter than the 95% range.

(d) The margin of error at a 90% confidence level would be higher than 3%.

FALSE, as Z decreases, the margin decreases, since Z is in the numerator of the equation for margin of error.

Legalization of marijuana, Part I. (6.10, p. 216) The 2010 General Social Survey asked 1,259 US residents: “Do you think the use of marijuana should be made legal, or not” 48% of the respondents said it should be made legal.

(a) Is 48% a sample statistic or a population parameter? Explain.

It is a sample statistic, because it is from a sample of the population of US residents.

(b) Construct a 95% confidence interval for the proportion of US residents who think marijuana should be made legal, and interpret it in the context of the data.

```
standard_error<-function(p_null_hypothesis,n){
  p<-p_null_hypothesis
  n<-n
  return(sqrt((p*(1-p)/n)))
}

n<-1259
p<-.48
SE<-standard_error(.48,1259)

confidence_interval95<-function(p,se){
  b<-p-1.96*se
  a<-p+1.96*se
  return(cat("95% confidence interval is" ,b,"-",a))
}


confidence_interval95(.48,SE)
```

```
## 95% confidence interval is 0.4524028 - 0.5075972
```

(c) A critic points out that this 95% confidence interval is only accurate if the statistic follows a normal distribution, or if the normal model is a good approximation. Is this true for these data? Explain.

This is true, but it is reasonable to assume the normal model because it passes the success, failure test.

```
success_failure(n,p)
```

```
## [1] "reasonable to construct confidence interval"
```

PERFORM TESTS:

1. INDEPENDENCE: It might not be independent, because it is not clear how they chose the people to survey, or if there was self selection. But if it is.
2. SUCCESS-FAILURE TEST:

```r
success_failure<-function(n,p){
  n<-n
  p<-p
  a<-(n*p>10)
  b<-n*(1-p)>10
  a
  b
  if(a==TRUE & b==TRUE){
    return("reasonable to construct confidence interval")
  }else{
    return("CANNOT use a confidence interval")
  }


}

success_failure(1259,.48)
```

```
## [1] "reasonable to construct confidence interval"
```

(d) A news piece on this survey's findings states, "Majority of Americans think marijuana should be legalized." Based on your confidence interval, is this news piece's statement justified?

No, because the confidence interval bounds the .5 mark. It could be higher or lower than 50% (but more likely lower)

Legalize Marijuana, Part II. (6.16, p. 216) As discussed in Exercise above, the 2010 General Social Survey reported a sample where about 48% of US residents thought marijuana should be made legal. If we wanted to limit the margin of error of a 95% confidence interval to 2%, about how many Americans would we need to survey?

```
n_size_margin_of_error<-function(z,margin_desired,estimate_of_proportion=.5){
  p<-estimate_of_proportion
  x<-margin_desired
  n<-z^2*p*(1-p)*1/x^2
  return(n)
}

n_size_margin_of_error(1.96,.02,.48)
```

```
## [1] 2397.158
```

Sleep deprivation, CA vs. OR, Part I. (6.22, p. 226) According to a report on sleep deprivation by the Centers for Disease Control and Prevention, the proportion of California residents who reported insuffient rest or sleep during each of the preceding 30 days is 8.0%, while this proportion is 8.8% for Oregon residents. These data are based on simple random samples of 11,545 California and 4,691 Oregon residents. Calculate a 95% confidence interval for the difference between the proportions of Californians and Oregonians who are sleep deprived and interpret it in context of the data.

```r
#insufficient rest cali
p1<-.08
#sample
n1<-11545

#insufficient rest Oregon

p2<-.088
#sample
n2<-4691

#SUCCESS-FAILURE TEST
p1p2_success_failure(n1,p1,n2,p2)
```

```
## [1] "reasonable to construct confidence interval"
```

```r
se<-p1p2_standard_error(p1,n1,p2,n2)

p1p2_confidence_interval95(p1,p2,se)
```

```
## [1] "95% confidence interval is -0.017 to 0.001"
```

Barking deer. (6.34, p. 239) Microhabitat factors associated with forage and bed sites of barking deer in Hainan Island, China were examined from 2001 to 2002. In this region woods make up 4.8% of the land, cultivated grass plot makes up 14.7% and deciduous forests makes up 39.6%. Of the 426 sites where the deer forage, 4 were categorized as woods, 16 as cultivated grassplot, and 61 as deciduous forests. The table below summarizes these data.

Woods	Cultivated grassplot	Deciduous forests	Other	Total
4	16	61	345	426

(a) Write the hypotheses for testing if barking deer prefer to forage in certain habitats over others.

H_0: the deer don't have a preference for foraging spots

H_A: there is a preference for one over another.

(b) What type of test can we use to answer this research question?

a chi-squared test

(c) Check if the assumptions and conditions required for this test are satisfied.

It is independent, and all categories of the probable values are above 5, as seen in the table below

```
counts<-c(4,16,61,345)
total<-sum(counts)
percents<-c(.048,.147,.396,.409)
table<-data.frame("counts"=counts,"percents"=percents)

table<-table%>%
  mutate(expected=percents*total)%>%
  select(-percents)

table
```

```
##   counts expected
## 1      4   20.448
## 2     16   62.622
## 3     61  168.696
## 4    345  174.234
```

(d) Do these data provide convincing evidence that barking deer pre- fer to forage in certain habitats over others? Conduct an appro- priate hypothesis test to answer this research question.

```
(chisq.test(table))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table
## X-squared = 145.37, df = 3, p-value < 2.2e-16
```

since the p-value is below .05 then we reject H_0 and accept H_A

Coffee and Depression. (6.50, p. 248) Researchers conducted a study investigating the relationship between caffeinated coffee consumption and risk of depression in women. They collected data on 50,739 women free of depression symptoms at the start of the study in the year 1996, and these women were followed through 2006. The researchers used questionnaires to collect data on caffeinated coffee consumption, asked each individual about physician-diagnosed depression, and also asked about the use of antidepressants. The table below shows the distribution of incidences of depression by amount of caffeinated coffee consumption.

		Caffeinated coffee consumption					
		≤ 1 cup/week	2-6 cups/week	1 cup/day	2-3 cups/day	≥ 4 cups/day	Total
Clinical	Yes	670	$\boxed{373}$	905	564	95	2,607
depression	No	11,545	6,244	16,329	11,726	2,288	48,132
	Total	12,215	6,617	17,234	12,290	2,383	50,739

(a) What type of test is appropriate for evaluating if there is an association between coffee intake and depression?

A two-way chi-squared test.

(b) Write the hypotheses for the test you identified in part (a).

H_0: there is no corrolation between coffe drinking and depression, the proportion of depressed people will be within the margin of error across the cups per week groups

H_A: there is a corrolation between coffee drinking and depression, the results will not be within the margin of error

(c) Calculate the overall proportion of women who do and do not suffer from depression.

```
#build table
coffee<-matrix(c(670,373,905,564,95,11545,6244,16329,11726,2288),ncol=5,byrow=TRUE)

colnames(coffee)<-c("cup/week","2-6 cup/week","1 cup/day","2-3 cup/day","4+ cup/day")
rownames(coffee)<-c("depression yes","depression no")

#column proportion
(expected_pcts<-margin.table(coffee,1)/margin.table(coffee))
```

```
## depression yes  depression no
##     0.05138059     0.94861941
```

(d) Identify the expected count for the highlighted cell, and calculate the contribution of this cell to the test statistic, i.e.

```
#expected count
dat<-coffee
row1<-lapply(margin.table(dat,2), FUN=function(x){x*expected_pcts[1]})
row2<-lapply(margin.table(dat,2),FUN=function(x){x*expected_pcts[2]})

expected_counts<-rbind(row1,row2)
expected_counts
```

```
##      cup/week 2-6 cup/week 1 cup/day 2-3 cup/day 4+ cup/day
## row1 627.614  339.9854     885.4932  631.4675    122.44
## row2 11587.39 6277.015     16348.51  11658.53    2260.56
```

(e) The test statistic is . What is the p-value?

```r
coffee<-as.table(coffee)
summary(coffee)
```

```
## Number of cases in table: 50739
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 20.932, df = 4, p-value = 0.0003267
```

the p-value is .0003 (f) What is the conclusion of the hypothesis test?

since the p value is below .05, we would REJECT the null hypothesis, and say there is a link between coffee and depression.

(g) One of the authors of this study was quoted on the NYTimes as saying it was "too early to recommend that women load up on extra coffee" based on just this study. Do you agree with this statement? Explain your reasoning.

```r
prop.table(coffee,2)
```

```
##                    cup/week 2-6 cup/week  1 cup/day 2-3 cup/day 4+ cup/day
## depression yes 0.05485059   0.05636996 0.05251248  0.04589097 0.03986572
## depression no  0.94514941   0.94363004 0.94748752  0.95410903 0.96013428
```

I agree that giving medical advice based on a single study is too early. Even though we are deep into the tail of our probability, You are able to say that it has a strong effect, but can you conclusively say it will always be (in this case) a positive effect?