Chapter 6 - Inference for Categorical Data

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.

ANSWER : FALSE. The confidence interval applies to the entire population and not just to the sample taken.

(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.

ANSWER : FALSE .This sample allows us to make an inference about the population

(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%.

ANSWER : False: confidence interval will give us us a range of possible values for the true population proportions

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

ANSWER : False: it would be lower, since we are lowering our confidence

---

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.
(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.
(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.
(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?

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?

formula ME=z*SE

```
# using data from above
ME <- 0.02
p<-0.48

#for 95 confidence

z<-qnorm(.975)
z
```

```
## [1] 1.959964
```

```
n <- (p * (1 - p ) * z ^ 2) / ME ^ 2
cat("Americans needed for survey is : " , n)
```

```
## Americans needed for survey is :  2397.07
```

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.

Answer : Since the interval contains 0 we can state with a 95% confidence that the proportions of Californians and Oregonians #are not statistically different.

```
pCA <- 0.08
nCA <- 11545
pOR <- 0.088
nOR <- 4691
cip <- 0.95 # Defining confidence interval

pdifference<- pOR-pCA

# Compute standard error and margin of error for the proportion difference.
SE <- ( ((pCA * (1 - pCA)) / nCA) +  ((pOR * (1 - pOR)) / nOR)) ^ 0.5

#for 95 confidence
Z<-qnorm(.975)
Z
```

```
## [1] 1.959964
```

```
ME<-Z * SE

# Construct the 95% confidence interval.
ci <- c(pdifference - ME, pdifference + ME )

cat("95 percent ci value will be :",ci)
```

```
## 95 percent ci value will be : -0.001497954 0.01749795
```

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.

Answer : Null Hypothesis: Barking deer prefers to forage in certain habitats over others. Alternative Hypothesis: Barking deer doesn't prefer to forage in certain habitats over others.

(b) What type of test can we use to answer this research question? Answer : Chi-Squared test can be used to answer this research question.

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

1. We assume each case is independent of each other.
2. A condition requires that each cell-count must have at least 5 expected cases.

```
habitat_types <- c("Woods", "Cultivated Grass", "Deciduous Forests", "Others", "Total")
habitat_ratio <- c(0.048, 0.147, 0.396, 1-(0.048+0.147+0.396), 1)
habitat_expected <- c(round(426*0.048), round(426*0.147), round(426*0.396), 426-(round(426*0.048)+round
habitat_used <- c(4,16,61, 426-(4+16+61), 426)
formulate_table <- array(data=c(habitat_ratio, habitat_expected, habitat_used),
                 dim = c(5,3),
                 dimnames = list(habitat_types, c("Ratio", "Expected", "Used")))
formulate_table
```

```
##                   Ratio Expected Used
## Woods             0.048       20    4
## Cultivated Grass  0.147       63   16
## Deciduous Forests 0.396      169   61
## Others            0.409      174  345
## Total             1.000      426  426
```

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

Answer : If p<0.05, we reject the null hypothesis. based on below test we reject H0 and accept HA in this case.

```
#chisq.test(x=habitat_used[1:4], p=habitat_ratio[1:4])

habitats <- c(4, 16, 67, 345)
region <- c(20.45, 62.62, 168.70, 174.23)
k <- length(habitats)
df <- k - 1
# Compute the chi2 test statistic
chi <- (habitats - region ) ^ 2 / region
```

```r
chi <- sum(chi)

# check the chi2 test statistic and find p-val
p_Val <- 1 - pchisq(chi, df = df)

cat("the calculated p-value is : " , p_Val)
```

```
## the calculated p-value is :  0
```

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	2-6	1	2-3	≥ 4	
		cup/week	cups/week	cup/day	cups/day	cups/day	Total
Clinical	Yes	670	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?

Answer : Chi-squared test is appropriate for evaluating if there is an association between coffee intake and depression.

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

Answer:

H0: There is no association between coffee intake and depression in women.

HA: There is association between coffee intake and depression in women.

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

Answer:

The overall proportion of women who suffer from depression = 2607/50739 = 5.138%

The overall proportion of women who do not suffer from depression = 48132/50739 = 94.862%

```
coffee_consumption <- c("<=1 cup/week", "2-6 cups/week", "1 cup/day",
                        "2-3 cups/day", ">=4 cups/day", "TOTAL")
depression_yes <- c(670, 373, 905, 564, 95, 2607)
depression_no <- c(11545, 6244, 16329, 11726, 2288, 48132)
depression_total <- depression_yes + depression_no
table_coffee <- array(data=c(depression_yes , depression_no , depression_total),
                         dim=c(6,3),
                         dimnames=list(coffee_consumption,
                                       c("Depression_Yes","Depression_No", "Total")))
table_coffee <- t(table_coffee)
table_coffee
```

```
##                 <=1 cup/week 2-6 cups/week 1 cup/day 2-3 cups/day >=4 cups/day
## Depression_Yes           670           373       905          564           95
## Depression_No          11545          6244     16329        11726         2288
## Total                  12215          6617     17234        12290         2383
##                TOTAL
## Depression_Yes  2607
## Depression_No  48132
## Total          50739
```

(d) Identify the expected count for the highlighted cell, and calculate the contribution of this cell to the test statistic, i.e. $(Observed - Expected)^2/Expected$.

expected<-5.138*6617/100 cat("The expected count for the highlighted cell "373" is: ", expected)

contributionofthiscall <- (373-expected)ˆ2 /expected cat("The contribution of this cell to the test statistic is:" , contributionofthiscall)

(e) The test statistic is $\chi^2 = 20.93$. What is the p-value?

```
n <- 5
k <- 2

df <- (n-1)*(k-1)
chi2 <- 20.93

p_value <- 1 - pchisq(chi2, df)
```

(f) What is the conclusion of the hypothesis test?

Answer : As the $p<0.05$, we reject the null hypothesis and accept the alternative hypothesis.

(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.

Answer : Yes, as the rejection of the null hypothesis does not necessarily means the confirmation of alternative hypothesis. More data is needed to establish conclusive decision.