Data606 Chapter 6 - Inference for Categorical Data

Homework Completed By Chad Bailey

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

```
## Student response to problem 1

## 1.a True, we are 100% confident the sample rate is 46%, which is between 43% and 49%
## 1.b True, the margin of error +/- 3% around the sample statistic 46% would give a
##           confidence interval of 43% to 49% for the general population.
## 1.c True, by definition roughly 95% of the samples would be expected to fall within
##           the given confidence interval.
## 1.d True, as a confidence interval's level of certainty is increased the error increases
```

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?

```
## Student response to problem 2

## 2.a It is a sample statistic. It is only a description of the sample of the
##     1,259 US residents not a description of the general US population

## 2.b
## 2.b1 construct 95% confidence interval
          lowerci <- 0.48 - 1.96*sqrt(0.48*(1-0.48)/1259)
          upperci <- 0.48 + 1.96*sqrt(0.48*(1-0.48)/1259)
          c(lowerci, upperci)
```

```
## [1] 0.4524028 0.5075972
```

```
## 2.b2 Based on this sample We are 95% confident that between 45.2% and 50.8% of Americans
##      think the use of marijuana should be made legal

## 2.c1 The GSS is conducted with high standards and so it is safe to assume
##      observations are independent
## 2.c2 Both the number of successes (p) and number of failures (1-p) are 10 or greater
          0.48 * 1259
```

```
## [1] 604.32
```

```
          (1-0.48) * 1259
```

```
## [1] 654.68
```

```
## 2.c3 Since both conditions are met, yes, these data are normal or nearly normal and
##      the normal model is a good approximation.

## 2.d No, the statement that a "Majority of Americans think marijuana should be legalized"
##     is not juststified. It is possible that the true population proportion could be
##     above 50% but given that the upper limit of the 95% confidence interval is 50.7 it
##     is unlikely.
```

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 ?

```
## Student response to problem 3

## 3.a formula for margin of error: me = (critical z) * sqrt(p*(1-p)/n)
## 3.b margin of error formula solved for n: (critical z)^2 * p * (1-p) / (me)^2
## 3.c critical value for 95% is 1.96
## 3.d worked equation
        (1.96^2) * 0.48 * (1-0.48) / (0.02^2)
```

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

```
## 3.5 To get a margin of error, the sample size would need to be at least 2398
```

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.

```
## Student response to problem 4

## 4.a  prepare:  n_ca: 11,545
##                p_ca: 0.08
##                n_or: 4,691
##                p_or: 0.088
## 4.b  Check:  (1) yes, observaions are independent
##              (2) yes, both success (p) and failure (1-p) are 10 or greater
##                  for both the California and Oregon samples


## 4.c  calculate:
##      4.c1 California
              lowerci <- round((0.088-0.08) - (1.96 *sqrt(0.088*(1-0.088)/4691
                                                         +0.08*(1-0.08)/11545)), 3)
              upperci <- round((0.088-0.08) + (1.96 *sqrt(0.088*(1-0.088)/4691
                                                         +0.08*(1-0.08)/11545)), 3)

              c(lowerci, upperci)
```

```
## [1] -0.001  0.017
```

```
## 4.d  conclude: We are 95% confident that the rates for CA and OR have a difference
##                between -0.1% and 1.7%. Because 0% is contained within the
##                confidence interval there is not enough information to say whether
##                there is a true difference between the rates for these two states.
```

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	67	345	426

(a) Write the hypotheses for testing if barking deer prefer to forage in certain habitats over others.
(b) What type of test can we use to answer this research question?
(c) Check if the assumptions and conditions required for this test are satisfied.
(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.

```
## Student response to problem 5

## 5.a H0: Barking do not have a preference of foraging in certain habitats
## 5.b Chi-squares test for one-way table can be used to answer this question
## 5.c check chi-squares assumptions
##      (1) independence: yes, each of the events are independent
##      (2) each particular scenario must have at least 5 EXPECTED cases
              habitat_rates <- c(0.048, 0.147, 0.396, 1-0.048-0.147-0.396)
              expected_rates <- habitat_rates*426
              expected_rates
```

```
## [1]  20.448  62.622 168.696 174.234
```

```
##          Yes, each particular scenario has at least 5 expected cases
## 5.d hypothesis testing
      observed_rates <- c(4, 16, 67, 345)

      rbind(observed_rates, expected_rates)
```

```
##                  [,1]   [,2]    [,3]    [,4]
## observed_rates  4.000 16.000  67.000 345.000
## expected_rates 20.448 62.622 168.696 174.234
```

```
      cases <- (observed_rates - expected_rates)^2 / expected_rates

      chi_stat <- sum(cases)
      chi_stat
```

```
## [1] 276.6135
```

```
      1 - pchisq(q=chi_stat, df= 4-1)
```

```
## [1] 0
## p-value is < 0.05, therefore the NULL Hypothesis is rejected.
## The data support the alternate hypothesis that the barking
## deer do have a preference for some foraging habitats over others.
```

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	$\boxed{373}$	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?
(b) Write the hypotheses for the test you identified in part (a).
(c) Calculate the overall proportion of women who do and do not suffer from depression.
(d) Identify the expected count for the highlighted cell, and calculate the contribution of this cell to the test statistic, i.e. $(Observed - Expected)^2/Expected)$.
(e) The test statistic is $\chi^2 = 20.93$. What is the p-value?
(f) What is the conclusion of the hypothesis test?
(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.64 Do you agree with this statement? Explain your reasoning.

```
## Student response to problem 6

## 6.a Chi-squares two-way table test
## 6.b H0: The number of cups of coffee will have no effect on rates of depression
##     H1: The number of cups of coffee will have an effect on rates of depression
## 6.c Overall rates of depression/non-depression
        overall_dep <- 2607/50739
        overall_nondep <- 48132/50739
## 6.d expected rate for depressed and 2-6 cups/week
        expected_cell <- 6617*overall_dep
        expected_cell
```

```
## [1] 339.9854
```

```
        (373-expected_cell)^2 / expected_cell
```

```
## [1] 3.205914
```

```
## 6.e chi-squares p-value
##      df = (R-1) * (C-1)
##         = (2-1) * (5-1)
##         = 1 * 4
##         = 4
        1 - pchisq(q = 20.93, df = 4)
```

```
## [1] 0.0003269507
## 6.f The conclusion is that the null hypothesis is rejected. There is some
##     effect of coffee consumption and rates of depression.

## 6.g Yes, the cautionary statement is warrented. This test only shows that
##     that there is SOME effect. However, it does not isolate that effect.
```