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.

1. 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.
2. 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.
3. 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%.
4. The margin of error at a 90% confidence level would be higher than 3%.

**ANSWER**

1. False. The sample shows a confirmed approval rate of 46%. Based on this, we are 95% confident that the approval rate for the entire US population lies between 43% and 49%.
2. True. A confidence interval is created to estimate the population proportion, which is 46%, with a margin of error of ±3%, resulting in a range from 43% to 49%.
3. True. This also describes a confidence interval, which provides an estimate of the population proportion within a certain range.
4. FALSE. With the same standard error (SE), a 90% confidence interval will have a smaller margin of error because it uses a narrower range. The z-score for a 90% confidence interval is 1.645, compared to 1.96 for a 95% confidence interval.

**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.

1. Is 48% a sample statistic or a population parameter? Explain.
2. 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.
3. 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.
4. 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?

**ANSWER**

1. This is a sample statistic because the 48% figure was obtained from a sample of 1,259 US residents, rather than representing the entire US population.
2. Given:

n=1259

p=0.48

z=1.96 (*A 95% confidence interval corresponds to an alpha level of 0.05. On the Z table, this gives a critical value of z=1.96z)*

First, calculating the standard error (SE):

SE = sqrt {p(1−p)/n} =sqrt { 0.48(1−0.48)/1259 } = sqrt { (0.48 × 0.52) / 1259 } = sqrt{0.2496/1259} ≈ 0.0141

Next, calculating the lower and upper bounds of the confidence interval:

Lower CI = p − ( z×SE ) = 0.48−(1.96×0.0141) = 0.48 − 0.0276 ≈ 0.4524

Upper CI = p+(z×SE) = 0.48+(1.96×0.0141) = 0.48 + 0.0276 ≈ 0.5076

Thus, the 95% confidence interval is approximately:

[0.4524,0.5076]

This indicates that we are 95% confident the approval rate for the legal use of marijuana in the US population falls between 45.24% and 50.76%.

1. Since the sample size exceeds 1,200 residents, it is considered quite large. Therefore, using a normal model allows us to calculate a 95% confidence interval.
2. This statement is false. The claim made in the news is not supported. While the confidence interval suggests the population probability could exceed 50%, it is equally likely to be significantly lower than 50%, as the majority of the confidence interval lies below 50%.

**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?

**ANSWER**

**Given Values:**

* + P=0.48P (Proportion)
  + ME=0.02 (Margin of Error)
  + Z = qnorm (0.975) = 1.96 (Z-score for 95% confidence interval)

**Calculating ZE:**

ZE = ME/Z = 0.02/1.96≈0.0102

**Calculating Size:**

N = P(1−P)/(ZE)^2 = 0.48×(1−0.48)/(0.0102)^2 = (0.48×0.52) / 0.00010404 ≈ 2395.

**Round and Add 1:**

Sample Size = round(n, 0) + 1 = 2395 + 1 = 2396

Thus, the final sample size is n=2396

**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**

**Given Values:**

* + nCal=11545 (Sample size for California)
  + nOre=4691 (Sample size for Oregon)
  + meanCal=0.08 (Mean for California)
  + meanOre=0.088 (Mean for Oregon)

**Calculate Mean Difference:**

mean\_diff = meanCal – meanOre = 0.08−0.088 = - 0.008

**Calculate Standard Error (SE):**

= sqrt { (meanCal⋅(1−meanCal)nCal)+(meanOre⋅(1−meanOre)nOre) }

= sqrt { (0.08⋅(1−0.08)/11545) + (0.088⋅(1−0.088)/4691) }

= sqrt {0.00000637+0.00001713​} ≈ sqrt {0.0000235}​ ≈ 0.0048

Thus, the mean difference is approximately −0.008 and the standard error SE is approximately 0.0048

**There is no substantial evidence to suggest a difference in sleep deprivation levels between residents of California and Oregon.**

**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

1. Write the hypotheses for testing if barking deer prefer to forage in certain habitats over others.
2. What type of test can we use to answer this research question?
3. Check if the assumptions and conditions required for this test are satisfied.
4. 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.

**ANSWER**

1. The null hypothesis states that the barking deer does not show a preference for any particular habitat when foraging. Conversely, the alternative hypothesis suggests that the barking deer does have specific habitats it prefers for foraging.
2. We can employ a Chi-square test.
3. Based on the data, we assume that the deer are independent and not influencing one another. In the woods section, there are 4 observed cases. However, when we calculate the expected number of cases for the woods, we find that 0.048×426=20.448, which is greater than or equal to 5, thus satisfying one of the conditions for the Chi-square test. Before performing this test, we must check two conditions: independence and sample size/distribution. While the cases appear to be independent, the sample distribution criterion is not met since the woods category only has 4 occurrences.

D. **Given Values:**

N=426 (Total from Table)

Observation=[4 16 67 345]

**Calculate Expected Values:**

Expected=[Nx0.048, Nx0.147, Nx0.396, Nx0.409]

=[426x0.048 426x0.147 426x0.396 426x0.409]

=[20.448 62.622 168.696 174.654]

**Calculate Chi-square Statistic:**

Chisq=∑((Observation−Expected)^2/Expected)

=(4−20.448)^2/20.448 + (16−62.622)^2/62.622 + (67−168.696)^2/ 168.696 + (345−174.654)^2/174.654

=276.61

**Degrees of Freedom:**

DF=4−1=3

**Calculating p-value:**

p-value = P (χ2 > Chisq) = pchisq(Chisq,DF,lower.tail=FALSE) = pchisq(276.61, 3, lower.tail=FALSE) ≈ 1.144×10^−59

Because the p-value for the Chi-square distribution is quite low, we can decisively reject the null hypothesis, suggesting that deer have distinct preferences for the habitats where they choose to forage.

**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 670 905 564 95 2,607

373

*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

1. What type of test is appropriate for evaluating if there is an association between coffee intake and depression?
2. Write the hypotheses for the test you identified in part (a).
3. Calculate the overall proportion of women who do and do not suffer from depression.
4. 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*).

1. The test statistic is *χ*2 = 20*.*93. What is the p-value?
2. What is the conclusion of the hypothesis test?
3. 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**

1. The Chi-square test for a two-way table can be employed to assess whether there is a relationship between coffee consumption and depression.
2. Null hypothesis: There is no association between coffee intake and clinical depression.   
   Alternative hypothesis: There is an association between coffee intake and clinical depression.
3. Depressed = 2607/50739 = 0.0514  
   Not.Depressed = 48132/50739 = 0.9486
4. **Given Values:**

Total observations: N=50739

Count of interest: 2607

Total count for the other group: 6617

Observed count: obs\_cnt=373

**Calculating Expected Count:**

Expected = [ (2607× 6617) / 50739 ] = 339.9854

**Calculating Test Statistic:**

test\_stat = (obs\_cnt−expected)^2/Expected

= (373 – 339.98)^2 /339.9854= 3.20

1. **Calculating p-value:**

p-value=P(χ2 ≤ 20.93, DF=4) = pchisq(20.93,4)

**Adjust for upper tail:**

p-value=1−p-value = 0.0003269507

1. I discard the null hypothesis and assert that there is a statistically significant difference among the different categories of coffee drinkers concerning clinical depression.
2. I concur with the author's assertion because this was an observational study, which means it cannot establish causation. This does not inherently indicate any causal relationship.