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.

FALSE.

We are 100% certain that 46% of people in this sample support the decision. No confidence interval is needed. I suppose, although this proportion exists within the confidence interval, we will not be incorrect 5% of the time.

(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.

$$SE = \sqrt{\frac{0.46 \times (1 - 0.46)}{1012}} = 0.01566699$$

$$ME = 1.96 \times SE = 0.0307073$$

$$CI = \hat{p} \pm ME = (0.4292927, 0.4907073)$$

(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%.

TRUE.

This is how we make assumptions about the population.

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

FALSE.

The value of z^* for a 90% confidence interval is 1.645 (by convention), and that will make the margin of error and the confidence interval narrower than the previous significance level.

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.

This is a sample statistic. It shows how many respondents within a random sample of a population answered a question in a specific way. We can use this to discover some things about the corresponding population parameter.

(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.

$$\hat{p} = 0.48 \qquad n = 1259 \qquad z^* = 1.96$$

$$CI = \hat{p} \pm z^* \times \sqrt{\frac{\hat{p} \times (1-\hat{p})}{n}} = (0.452, 0.508)$$

(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.

TRUE.

The sampling is assumed to be independent, and $n(\hat{p})$ and $n(1-\hat{p})$ are both greater than 10.

(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, I do not believe this statement is justified. The sample statistic is below $\frac{1}{2}$, and the null hypothesis can be found within the confidence interval. There is no evidence here suggesting the general population favors legalization.

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?

$$\hat{p} = 0.48 \qquad z^* = 1.96 \qquad ME < 0.02$$

$$ME > z^* \times \sqrt{\frac{\hat{p} \times (1-\hat{p})}{n}}$$

$$n > 1.96^2 \times \frac{0.48 \times 0.52}{0.0004}$$

$$n > 2398$$

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.

$$\hat{p}_1 = 0.080 \quad n_1 = 11545 \qquad \hat{p}_2 = 0.088 \quad n_2 = 4691 \qquad z^* = 1.96$$

$$H_0 : \hat{p_1} - \hat{p_2} = 0$$
$$H_A : \hat{p_1} - \hat{p_2} \neq 0$$

$$SE = \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}$$
$$SE = \sqrt{0.000006375054 + 0.00001710851} \approx 0.0048$$
$$CI = (\hat{p_1} - \hat{p_2}) \pm z^* \times SE$$
$$= (-0.017, 0.0014)$$

Since zero falls within the confidence interval, we fail to reject the null hypothesis. We can conclude that the variance in the random samples accounts for the difference between the two proportions.

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 : These regions are random samples, and the deer have no bias.

H_A : The dear are biased towards foraging in one type of terrain.

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

We should use a chi-square test.

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

The terrain types are independent, they do not overlap, however, the minimum number of expected cases 5 is not reached in the woods.

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

We cannot conduct a test because the conditions for the test are not satisfied.

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 depression	Yes	670	373	905	564	95	2,607
	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?

We should use a chi-square test!

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

H_0 : Caffeine consumption and depression are completely independent.

H_A : Depression is a function of caffeine consumption.

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

$\hat{p} = \frac{2607}{50739} \approx 0.0514 \qquad 1 - \hat{p} = \frac{48132}{50739} \approx 0.9486$

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

$E_2 = 6617 \times \hat{p} \approx 340 \qquad \frac{(O_2 - E_2)^2}{E_2} = \frac{1089}{340} \approx 3.203$

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

$df = k - 1 = 4 \qquad p = 0.000326951$

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

The large chi-square and nearly zero p-value suggest strong evidence to reject the null 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.

Although, it is reasonable to reject the null hypothesis, this study does not confirm H_A. There is nothing in the chi-square test that can qualify the magnitude of the value. Simply, there is probably a connection, but it's unclear what it could be.