Leland Randles DATA606

<u>Homework, Chapter 6</u>

6.6 2010 Healthcare Law. 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.[39]

(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%.

(a) False. We know exactly what percentage of people agreed with the decision in the sample (46%), there is no need for a confidence interval.
(b) True. Another way to say it is that we are 95% confident that the "true" proportion of people who agreed with the decision was between 43% and 49%.
(c) True. This is what the 43-49% confidence interval and +- 3% margin of error tells us. If we took a mutitude of random samples, about 5% would have an approval percentage which is not between 43% and 46%.
(d) False. All things equal, the lower the confidence interval, the lower the margin of error.

6.12 Legalization of marijuana, Part I. 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.[44]

(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?

(a) It's a sample statistic. The 48% applies to the sample, not the population as a whole. The population percentage is an unknown.
(b) pe <- 0.48
se <- sqrt((0.48 * (1 - 0.48)) / 1059)
c(pe - (qnorm(0.975) * se), pe + (qnorm(0.975) * se))
[1] 0.44991 0.51009
We are 95% confident that the true proportion of US residents who think marijuana should

be legal is between 0.44991 and 0.51009.

(c) The poll is based on a simple random sample and consists of less than 10% of the population. Also, there are more than 10 successes and failures.

(d) No. Though some samples would likely consistute a majority, most samples would not.

6.20 Legalize Marijuana, Part II. As discussed in Exercise 6.12, 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 ?

Solve for *se* in pe + (z * *se*) = 0.50

0.48 + (qnorm(0.975) * se) = 0.50, equals 0.02 / 1.959964, or 0.01020427.

Solve for *n*: sqrt((0.48 * (1 - 0.48)) / n) = 0.01020427; n = 2397.071; so

Sample size must be 2,398 Americans.

6.28 Sleep deprivation, CA vs. OR, Part I. According to a report on sleep deprivation by the Centers for Disease Control and Prevention, the proportion of California residents who reported insufficient 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.[53]

```
pe <- 0.88 - 0.8
se <- sqrt(((0.088 * (1 - 0.088)) / 4691) + ((0.08 * (1 - 0.08)) / 11545))
c(pe - (qnorm(0.975) * se), pe + (qnorm(0.975) * se))
[1] 0.07050205 0.08949795
```

6.44 Barking deer. 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.[62]

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 prefer to forage in certain habitats over others? Conduct an appropriate hypothesis test to answer this research question.

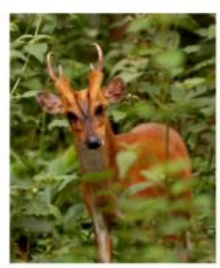

Photo by Shrikant Rao (http://flic.kr/p/4Xjdkk) CC BY 2.0 license

(a) H_0 = Barking deer do not prefer to forage in certain habitats over others.
H_A = Barking deer prefer to forage in certain habitats over others.
(b) Chi-square test for one-way table
(c) Each case that contibutes a count to the table must be independend of the other, and each particular scenario must have at least 5 expected cases (even the smallest has 0.048 * 426 caeses, which is 20+ cases).
(d) (4 – (0.048*426))^2 / (0.048*426) +
(16 – (0.147*426))^2 / (0.147*426) +
(67 – (0.396*426))^2 / (0.396*426) = 109.2465
Since there are 3 bins, df = 2. At this level of df, our X^2 test stat yields a very small p-value.

6.48 Coffee and Depression. 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.[63]

		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	$\boxed{373}$	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?
(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.

(a) Chi-square test for two-way tables
(b) H_0 = Caffeinated coffee consumption has no impact on incidences of depression.
H_A = Caffeinated coffee consumption has an impact on incidences of depression.
(c) Do suffer from depression: 2,607 / 50,739 = 0.05138059
Do not suffer from depression: 48,132 / 50,739 = 0.9486194
(d) Expected count is 0.05138059 * 6,617 = 339.9854, or 340
Contribution to X^2 is ((373 - 340)^2) / 340 = 3.202941

(e) There are 5 bins, so df = 4. This means the p-value is less than 0.001 according to the chi-suare table.

(f) At that p-value, we would reject the null hypothesis.

(g) I agree with the statement. We need to research further to determine the impact of caffeinated coffee consumption on incidents of depression.