Chapter 6 - Inference for Categorical Data

Jai Jeffryes

10/20/2019

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. Confidence intervals are assertions about a population, not a sample.**

(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. The margin of error is the measure which, plus or minus, spans the interval in which the population proportion is likely to fall.**

(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%. **False. Confidence intervals refer to populations, not samples.**

(d) The margin of error at a 90% confidence level would be higher than 3%. **False. It would be narrower, reflecting a lower standard error defining a more narrow range of the sample distribution.**

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. **A sample statistic. The US population is too large to measure its parameters, which is why we collect samples in order to estimate them.**

(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.

```r
# What do I know?
n <- 1259
p <- .48

# What do I need?
# SE
se <- sqrt((p * (1 - p)) / n)
```

The confidence interval is the point estimate of .48 plus or minus z (1.96) times the standard error. (0.4659198, 0.4940802). We are 95% confident the population proportion of US residents who think marijuana should be legal falls within that interval.

(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. **What we need are enough samples to include at least 10 successes and 10 failures. This sample appears large enough for that.**

(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, the point estimate is .48. Indeed, the population proportion may be higher than the point estimate, but it is 95% likely to fall between .47 and .49, not a majority.**

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

```
n <- 1.96^2 * ((.48 * (1 - .48)) / .02^2)
n
```

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

(Round up now.) We need to survey at least 2398 Americans.

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.

```r
se <- sqrt(((.08 * (1 - .08)) / 11545) + ((.088 * (1 - .088)) / 4691))
me <- 1.96 * se

(.08 - .088) - me
```

```
## [1] -0.01749813
```

```r
(.08 - .088) + me
```

```
## [1] 0.001498128
```

We are 95% confident that Californians' sleep deprivation ranges from 1.7% lower than Oregonians' to being the same (.0015 rounded down).

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.

ANSWER

H_0: The deer have no preference of habitat for foraging.

H_A: The deer prefer some habitats over others.

(b) What type of test can we use to answer this research question? **Chi-squared test.**

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

ANSWER

Each expected count must be at least 5. The total sites number 426. Expected counts are:

Woods: 4.8% = 20.448

Grass plot: 14.7% = 62.622

Deciduous forests: 39.6% = 168.696

Other: 40.9% = 174.234

These expected counts satisfy the conditions of a chi-squared test.

(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.

```
chi2 <- ((4 - 20.448)^2 / 20.448) + ((16 - 62.622)^2 / 62.622) + ((67 - 168.696)^2 / 168.696) + ((0 - 1
deg <- 3
p_val <- pchisq(chi2, deg, lower.tail = FALSE)
```

Did I do that right?. I get a p value of $3.7386902 \times 10^{-61}$. That's pretty close to zero. I reject the null hypothesis. Barking deer have foraging preferences.

I don't know if I did it right, so I'll describe my thinking, at least. The land proportions tell you how many sites out of the 426 fall into each habitat. If foraging were random, the land proportions times the 426 sites would be the expected value for each land category. The problem didn't state the proportion for "other" sites, so I subtracted what was stated from 1 to get that. Chi-squared is the observed minus expected squared all divided by expected. Then you add up the terms for each site. There are 4 categories of site, and degree of freedom is 1 less than that, so 3 for degree of freedom.

I used the function `pchisq()` to get the p value. I believe you have to disregard the lower tail. Did I do this correctly?

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	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? **Chi-squared test for two-way table.**

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

H_0: Coffee consumption has no bearing on depression.

H_A: Coffee consumption does affect incidence of depression.

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

Depression: 2,607 / 50,739 = 0.0513806

No depression: 48,132 / 50,739 = 0.9486194

(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 count: (row 1 * column 2) / table total.

```r
expected <- (2607 * 6617) / 50739
cell <- (373 - expected)^2 / expected
```

Expected: 339.9853958.

Cell contribution: 3.2059144.

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

ANSWER

```r
df <- 1 * 4
p_val <- pchisq(20.93, df, lower.tail = FALSE)
```

p-value: 3.2695073×10^{-4}.

(f) What is the conclusion of the hypothesis test? **We reject the null hypothesis. Coffee consumption does affect incidence of depression in women.**

(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. **I agree. Doesn't this test merely demonstrate that coffee has an affect on incidence of depression? It doesn't tell you yet if it decreases it or increases it. Isn't that so?