Chapter 6 - Inference for Categorical Data

Cameron Smith

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%.

ANSWERS

a) True in part, we are actually 100% confidence regarding this sample.

b) True, this is the point of statistical inference.

c) True, based on the 3% margin of error (plus/minus 3% of the 46% mean in this sample).

d) False, as the confidence level decreases so does the margin of error (bigger fishing net).

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?

ANSWERS

a) 48% is a sample statistic, in particular the mean of the sample.
b) The interval is 45% to 51% (see code block below).
c) Yes, it is mostly true. The CI can be used assuming the distribution is normal based on the central limit theorem, which requires that the sample is independent and that it has at least 10 observed successes and 10 observed failures.
d) No, it does not appear so. It’s certainly possible within the CI, but by no means guaranteed.

```r
# Answer for part b

n1 <- 1259
p1 <- .48

se <- sqrt( (p1*(1-p1)) / n1 )

low1 <- p1 - 1.96 * se
high1 <- p1 + 1.96 * se

paste("The 95% interval is", round(low1, 3), "to", round(high1, 3))
```

```
## [1] "The 95% interval is 0.452 to 0.508"
```

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

3,117 Americans would need to be surveyed.

```r
q <- qnorm(.01)

(q^2 * .48^2) / (.02^2)
```

```
## [1] 3117.251
```

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 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.

ANSWER

The CI is -.001 to .017 (see code block)

Comment / Note to self: How are we supposed to know which one is p1 and which one is p2 for these types of questions?

```r
# Need to find n and p for both proportions.  Formula on page 217 of book
p_california <- .08
n_california <- 11545

p_oregon <- .088
n_oregon <- 4691

pt_estimate <- p_oregon - p_california
z <- 1.96
se <- sqrt(
  (p_oregon*(1-p_oregon) / n_oregon) +
      (p_california*(1-p_california) / n_california)
)

low2 <- pt_estimate - z * se
high2 <- pt_estimate + z * se

paste("The 95% interval is", round(low2, 3), "to", round(high2, 3))
```

```
## [1] "The 95% interval is -0.001 to 0.017"
```

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.
(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 appropriate hypothesis test to answer this research question.

ANSWERS

a) The null hypothesis is that barking deer do not prefer to forage in certain habitats over others. The alternate hypothesis is that barking deer do prefer to forage in certain habitats over others.
b) A goodness of fit test using chi-square can be used to answer this question.
c) The conditions for a chi-square test are satisfied:

    (1) Independence (each case independent of others): Yes
    (2) Sample size / distribution (each w/ at least 5 cases): Yes, though the actual observations seem to differ from those expected.

d) The p-value is nearly 0 which suggests there is almost no chance that the null hypothesis is true and thus we can reject it.

```r
# Total observations
total_obs <- 426

# Variables for observed values
woods_obs <- 4
grass_obs <- 16
forest_obs <- 61
other_obs <- 345

# Variables for expected values
woods_exp <- total_obs * .048
grass_exp <- total_obs * .147
forest_exp <- total_obs * .396
other_exp <- total_obs * (1 - (.048 + .147 + .396))

# Check to make sure exp adds to 426
isTRUE(woods_exp + grass_exp + forest_exp + other_exp == total_obs)
```

```
## [1] TRUE
```

```r
# Do the chi-square test
results <-
  ((woods_obs - woods_exp)^2 / woods_exp) +
  ((grass_obs - grass_exp)^2 / grass_exp) +
  ((forest_obs - forest_exp)^2 / forest_exp) +
  ((other_obs - other_exp)^2 / other_exp)

deg_freedom <- 4 - 1
pvalue <- 1-pchisq(results, deg_freedom)

paste("The p-value for this chi-square test is", pvalue, "so the null hypothesisis rejected")
```

```
## [1] "The p-value for this chi-square test is 0 so the null hypothesisis rejected"
```

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	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. $(\mathit{Observed} - \mathit{Expected})^2/\mathit{Expected}$.
(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. Do you agree with this statement? Explain your reasoning.

ANSWERS

a) A chi square test would be appropriate here.
b) There is no association between coffee intake and depression.
c) Suffer = 5.1%, and Do not suffer = 94.9% (see code block below)
d) The expected count is 340 (see code block below)
e) The p-value is zero.
f) The null hypothesis can be rejected.
g) I agree. Although it is nearly certain that there is a relationship, correlation does not necessarily infer causation. There could be lots of other reasons that depressed women drink more coffee than others.

```r
# Answer for part b
num_suffer <- 670 + 373 + 905 + 564 + 95
num_nosuffer <- 11545 + 6244 + 16329 + 11726 + 2288
num_total <- num_suffer + num_nosuffer
p_suffer <- num_suffer / num_total
p_nosuffer <- num_nosuffer / num_total

# Answer for part d
num2to6_exp <- 6617 * p_suffer

# Answer for part e
test_stat2 <- 20.93
deg_freedom2 <- 5 - 4
pvalue2 <- 1-pchisq(20.93, deg_freedom2)
deg_freedom <- 5 - 1
pvalue <- 1-pchisq(test_stat2, deg_freedom)
paste("The p-value for this chi-square test is", round(pvalue2,3))
```

```
## [1] "The p-value for this chi-square test is 0"
```