Chapter 6 - Inference for Categorical Data

Kenan Sooklall

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, the confidence interval is for estimating the population parameter not sample parameters

(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

(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%.

```
smp = c(rep(c('support'), round(1012*0.46)), rep(c('no support'), round(1012*.54)))
data.frame(smp) %>%
specify(response = smp, success = "support") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.95)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.431    0.491
```

True

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

A higher confidence level widens the curve so that causes the margin of error to go up.

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.

True 48% is a sample statistic from the population of 1259.

(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.

```
n = 1259
p = 0.48
q = 1-p
smp = c(rep(c('support'), round(n*p)), rep(c('no support'), round(n*q)))
data.frame(smp) %>%
specify(response = smp, success = "support") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.95)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.452    0.508
```

```
m <- 1.96 * sqrt(p * q / n)
c(0.48 - m, 0.48 + m)
```

```
## [1] 0.4524028 0.5075972
```

Between 45.2% and 50.8% of the population will support the legal use of marijuana with a 95% confidence.

(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.

Requirement n * p = 604 > 10 n * q = 654 > 10

Both requirements are met by a factor larger than 10. The critic point is false.

(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?

Depending on how majority is defined the 95% confidence is at the tip but given the margin of error I don't think the news piece is justified. I think it would be accurate to say "Americans are divided on marijuana legalization" since the true value is probabily ~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?

$moe = sqrt(p * (1-p)/n) \rightarrow n = p * (1-p)/moe^2$

```
p = 0.48
p * (1-p) / 0.02^2
```

```
## [1] 624
```

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.

```
pc = 0.08
po = 0.088
nc = 11545
no = 4691
z = 1.96
se = sqrt(pc * (1-pc)/nc + po * (1-po) / no)
me = z * se
p = pc - po
c(p - me, p + me)
```

```
## [1] -0.017498128  0.001498128
```

The difference in sleep proportions between Californians and Oregonians is about 1.8%

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
4.8	14.7	39.6		

(a) Write the hypotheses for testing if barking deer prefer to forage in certain habitats over others.

H_o = barking deers preference for Woods, Cultivated grassplot and Deciduous forests are the same

H_a = barking deers preference differ between Woods, Cultivated grassplot and Deciduous forests

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

(c) Check if the assumptions and conditions required for this test are satisfied. Independence does hold; however each scenario should have at least 5 expected cases and the Woods category has only 4.

(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.

randomly

```
observed = c(4,16,61,345)
expected = 426 * c(0.048,0.147,0.396, 0.409)
df = 3
chi = sum((observed - expected)^2/expected)
pchisq(chi, 3, lower.tail = FALSE)
```

```
## [1] 2.799724e-61
```

That's a very small p meaning we reject the null hypothesis and accept the alternative, barking deers preference differ between Woods, Cultivated grassplot and Deciduous forests

---

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? Chi square test All conditions apply

- Random sample
- Each case only contributes to one cell in the table
- Each scenario has at least 5 expected cases

(b) Write the hypotheses for the test you identified in part (a). H_o - Coffee consumption **has no** affect on the risk of depression in women H_a - Coffee consumption **has an** affect on the risk of depression in women

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

```
clinical_depression = 2607/50739
no_clinical_depression = 1 - clinical_depression
```

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

```
expected = clinical_depression * 6617
cont = (373 - expected)^2/expected
```

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

```
r=2
c=5
pchisq(20.93, (r-1)*(c-1), lower.tail = FALSE)
```

```
## [1] 0.0003269507
```

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

With a threshold of 0.05 and a p-val of 0.0003 we 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.

Our conclusion is that coffee has an effect on a womens risk of clinical depression. So some caution would be fine; however given this is just one study more investigation will be required.