Chapter 6 - Inference for Categorical Data

Zachary Palmore

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. CI always references the population, not the 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. This uses correct wording and references the population as a whole.

(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. There is no guarantee that those sample proportions will produce the true population proportion nor can we confirm that after many random samples, the confidence intervals will remain exactly 43% and 49%. In fact, with many random samples, the results should be slightly different (or at least more accurate).

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

False. If the margin of error at a 95% confidence level is 3% then at a 90% confidence level with the same population the margin of error would be lower than 3% because the interval that could include the population parameter shrinks and the confidence decreases. As the confidence level decreases, the margin of error should also decrease.

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.

48% is a sample statistic. The true population parameter is unknown.

(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.

Based on the formula for CI:

$$CI = p \pm (z * SE)$$

Where SE is:

$$\sqrt{(\frac{p(1-p)}{n})}$$

We are given the sample size, n, at 1259 and the proportion of respondents, 0.48. The zscore that corresponds with a 95% confidence interval is 1.96. Now, we can calculate.

```
n <- 1259
p <- 0.48
z <- 1.96
SE <- sqrt((p*(1-p))/n)
loci <- p - (z * SE)
upci <- p + (z * SE)
CI_gss_legal <- c(loci, upci)
CI_gss_legal
```

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

The lower confidence interval is 0.452 and the upper is 0.5076.

(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.

There are two conditions that need to be met for the distribution to be nearly normal. Nearly normal is enough to be a good approximation of 95% confidence. The first condition is that the observations are independent of one another and second is the success failure condition. If the residents were randomly asked in this study, then we can assume they are independent of another. Then there is the success failure condition which states that $n * (1 - p) \geq 10$ and $n * p \geq 10$. We can test this.

```
n*(1-p)
```

```
## [1] 654.68
```

```
n*p
```

```
## [1] 604.32
```

In this case, the success failure condition is met because these results are greater than 10. The distribution is nearly normal.

(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 95% confidence interval is a best estimate of the population parameter and it is not clearly over 50%; which is what would be needed to secure a majority. The interval needs to be at or over 50 at its lowest interval to say safely, that a majority of people support it.

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?

By rearranging the formula for margin of error we can solve for n, the sample size.

$$n = \frac{p(1-p)}{\frac{me}{z}^2}$$

```
me <- 0.02
(p*(1-p))/(me/z)^2
```

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

We round up to be safe and determine that we would need to survey at least 2398 individuals.

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.

$$CI = (p_1 - p_2) \pm z * \sqrt{(\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2})}$$

```
p1 <- 0.080
p2 <- 0.088
n1 <- 11545
n2 <- 4691
z <- 1.96
# In 5 steps:
se1 <- (p1*(1-p1))/(n1)
se2 <- (p2*(1-p2))/(n2)
diff <- p1-p2
diff+z*sqrt((se1+se2))
```

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

```
diff-z*sqrt((se1+se2))
```

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

```
# In 2 steps
(p1-p2)+z*sqrt(((p1*(1-p1))/n1)+(p2*(1-p2))/n2)
```

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

```
(p1-p2)-z*sqrt(((p1*(1-p1))/n1)+(p2*(1-p2))/n2)
```

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

We are 95% confident that the difference between the proportion of Californians and Oregonians who are sleep deprived is between -1.75% and 0.15%.

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.

Null hypothesis: Barking deer have no preference in the microhabitats they choose to forage

Alternative hypothesis: Barking deer have a preference in the microhabitats they choose to forage

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

A chi-squared test could be used to answer this question because in this example we want to evaluate whether there is convincing evidence that a set of observed counts in several categories are unusually different from what might be expected under a null hypothesis. The observations, in this example, are the bed sites of barking deer and the expected counts are the proportions of land multiplied by the population of deer that make up the area.

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

There are two conditions for a chi-squared test. Independence and sample size or sample distribution. In this example, the deer bed site observations are independent of all other cases in the table. In the sample size and distribution, each count must have at least 5 expected cases. This can been viewed in the data frame.

```
# To view the expected bed counts
deer <- as.vector("")
deer$categories <- c("Woods", "Grassplots", "Deciduous", "Other", "Total")
```

```
## Warning in deer$categories <- c("Woods", "Grassplots", "Deciduous", "Other", :
## Coercing LHS to a list
```

```
deer$areas <- c(0.048, 0.147, 0.396, 1-(sum(0.048, 0.147, 0.396)), sum(0.048, 0.147, 0.396, 1-(sum(0.048
deer$beds <- c(4, 16, 61, 345, 426)
deer <- as.data.frame(deer)
deer <- deer %>%
  mutate(expected_beds = areas*426) %>%
  mutate(chi_values = ((beds-expected_beds)^2)/expected_beds )
deer
```

```
##   X.. categories areas beds expected_beds chi_values
## 1          Woods 0.048    4        20.448   13.23047
## 2     Grassplots 0.147   16        62.622   34.71002
## 3      Deciduous 0.396   61       168.696   68.75343
## 4          Other 0.409  345       174.234  167.36703
## 5          Total 1.000  426       426.000    0.00000
```

Although there is an observed bed count of 4, the lowest expected count of deer beds is 20.448. Since 20.448 is greater than 5, this condition has also been met.

(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.

Yes, based on the chi-squared test statistic and a p-value of near 0, we can conclude that there is significant evidence to reject the null hypothesis in favor of the alternative, that barking deer have preference in where they forage.

```
# Using stats calculator
k <- 426
obs <- c(4,16,67,345)
exp <- c(n*0.048,n*0.147,n*0.396,n*0.409)
chisquared <- sum((obs-exp)^2/exp)
df <- 3
pchisq(chisquared,df,lower.tail=FALSE)
```

```
## [1] 1.060239e-137
```

```
# Chi-squared By hand
o1 <- 4
o2 <- 16
o3 <- 61
o4 <- 345
e1 <- 0.048*426
e2 <- 0.147*426
e3 <- 0.396*426
e4 <- (1-(sum(0.048, 0.147, 0.396)))*426 # 174.234
x1 <- ((o1-e1)^2)/e1
x2 <- ((o2-e2)^2)/e2
x3 <- ((o3-e3)^2)/e3
x4 <- ((o4-e4)^2)/e4
Xsquared <- sum(x1,x2,x3,x4)
# Or in one linear step
X_a <- (((o1 - e1)^2)/e1)+(((o2 - e2)^2)/e2)+(((o3 - e3)^2)/e3)+(((o4 - e4)^2)/e4)
X_a
```

```
## [1] 284.0609
```

```
Xsquared
```

```
## [1] 284.0609
```

This is a large chi-squared value (far from 0) indicating a weak correlation between the observed and expected bed counts. Thus a weak correlation is present in the null hypothesis, that barking deer forage sites are random or have no preference. Using the most common $df = 1$ we can see the chi-squared value is 284.0609 and corresponding p-value is close to zero. At a significance level of $p = 0.01$, the results are significant and we should reject the null hypothesis as there is a preference in forage site for barking deer.

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?

A chi-squared test could be used to answer this question as well. We are evaluating if there is an association between the categorical variables of coffee intake and depression.

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

Null hypothesis: Caffeinated coffee consumption has no effect on the proportion of women who are diagnosed with clinical depression

Alternative hypothesis: Caffeinated coffee consumption either increases or decreases the proportion of women who are diagnosed with clinical depression

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

Given that the number of women surveyed and the total of those who were diagnosed with clinical depression, we can find their proportions like this:

```
2607/50739
```

```
## [1] 0.05138059
```

```
48132/50739
```

```
## [1] 0.9486194
```

Where 0.9486 were not diagnosed with clinical depression and 0.0513 were diagnosed with clinical depression.

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

The expected count of women to be diagnosed with clinical depression by drinking 2-6 cups of caffeinated coffee per week is about 339.

```
# Clinically depressed
6617*.0513
```

```
## [1] 339.4521
```

```r
# Not clinically depressed
6617*.09486
```

```
## [1] 627.6886
```

Their contribution to the test statistic is 3.4100

```r
((373-339)^2) / 339
```

```
## [1] 3.410029
```

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

Using degrees of freedom = (# of row - 1)*(# of columns - 1) we find that df = 4 and the p-value is about 0.00033.

```r
df <- (2-1)*(5-1)
chi <- 20.93
# Calculate p-value
pchisq(chi, df, lower.tail = FALSE)
```

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

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

The results are significant at a level of 0.01 and we should reject the null hypothesis. There appears to be a correlation between caffeinated coffee consumption and the proportions of clinical 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. Do you agree with this statement? Explain your reasoning.

Yes, correlation does not equal causation. There may be many other factors influencing the proportions that were not accounted for. To say that if women drink extra coffee it will decrease their likelihood of clinical depression is not true and does not properly report the results of the study.