Chapter 6 - Inference for Categorical Data

Chunjie Nan

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.

Answer: False. It says that 46% of the sample of 1012 Americans agrees the decision.

(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.

Answer: True. With the 95% of confidence level and 3% of margin error means 43% and 49% of Americans support the decision.

(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%.

Answer: False. The confidence interval gives the true population proportion.

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

Answer: False. The margin error is increasing as the confidence level goes down.

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.

Answer: It is 48% of the sample statisitic for 1259 US residents.

(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.

Answer:

```
z <-1.96
n <-1259
p <- 0.48
se <- sqrt(p*(1-p)/n)
se
```

```
## [1] 0.01408022
```

```
up <- p + z*se
low <- p - z*se
confidence <- c(up, low)
confidence
```

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

so, the confidence interval is between 0.51 and 0.45 with the the standard error of 0.014.

(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.

True. The estimation is is good if the distribution is normal and the observation is independent.

(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?

False. Because the 95% of confidence interval is between 0.45 and 0.51.

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?

```
ME <- 0.02
z <-qnorm(0.975)
se <- ME/z
se
```

```
## [1] 0.01020427
```

```
P <- 0.48
n <- (p * (1-p)) / se^2
n
```

```
## [1] 2397.07
```

So, we need to survey at leat 2398 US residents for the survey.

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.

```
z <- 1.96
prop_CA <- 0.08
n_CA <- 11545
prop_Oreg <- 0.088
n_Oreg <-4691

# standard error for california
se_CA <- sqrt(prop_CA*(1-prop_CA)/n_CA)
se_CA
```

```
## [1] 0.002524887
```

```
# standard error for Oregon
se_Oreg <- sqrt(prop_Oreg*(1-prop_Oreg)/n_Oreg)
se_Oreg
```

```
## [1] 0.004136243
```

```
sepro <- sqrt(se_CA^2+se_Oreg^2)
sepro
```

```
## [1] 0.004845984
```

```
Low<-(prop_Oreg-prop_CA)-z*sepro
Up<-(prop_Oreg-prop_CA)+z*sepro
CI <- c(Up, Low)
CI
```

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

Answer: The confidence interval contains 0, so we fail to reject the null hypothesis.

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.

Answer: null hypotheis: There is no difference between each forage for deer. alternative hypothesis: Deer prefers at least one of the forage.

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

Answer: We can use a Chi-square test.

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

Answer: The case is independent for each case and it has 5 expected cases at least.

(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.

```
n<-426
obs<-c(4,16,67,345)
expect<-c(n*0.048,n*0.147,n*0.396,n*0.409)
chisq<-sum((obs-expect)^2/expect)
df<-3
pchisq(chisq,df,lower.tail=FALSE)
```

```
## [1] 1.144396e-59
```

Answer: We cannot reject the null hypothesis because there's no difference between each forage for 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?

Answer: Chi square test is used.

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

Null Hypothesis : Coffee intake and depression are independent. Alternative Hypothesis : Coffe intake and depression are dependent.

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

```
prop_women_with_dep = 2607/50739
prop_women_with_dep
```

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

```
prop_women_without_dep = 1 - prop_women_with_dep
prop_women_without_dep
```

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

Answer: About 5.1% women suffer from depression, and 94.9% of women 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. $(Observed - Expected)^2/Expected$).

```
groups = 5
df = 4
expected = prop_women_with_dep*6617
cell = (373 - expected)^2/expected
cell
```

```
## [1] 3.205914
```

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

```
p_value = pchisq(20.93,df,lower.tail = FALSE)
p_value
```

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

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

Answer: We reject the null hypothesis.Therefore, we can conclude that coffee intake and depression are dependent.

(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.

Answer: I agree with the author's statement because there might be other factors that affect depression as well.