Chapter 6 - Inference for Categorical Data

Samuel I Kigamba

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.

Response: Confidence interval is about the range of population parameters and not the sample. Thus this statement is false.

(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.

REsponse: This is the true interpretation of a confidence interval and thus the statement is 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%.

Confidence interval relates to range of population paramenters and not the sample. The statement is False.

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

The ME at 90% confidence interval would be would be equivalent of z-score*SE and the z-score at 90% is 1.96. Thus the ME would be less than 3%.

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.

Response: This is a sample statistic since it relates to a sample proportion.

(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.

```
uppr <- 0.48 + 1.96*sqrt(0.48*(1-0.48)/1259)
lwr <- 0.48 - 1.96*sqrt(0.48*(1-0.48)/1259)

c(lwr, uppr)
```

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

Response: We are confident that the true proportion of Americans who think Marijuana should be legalized is between 45.2% and 50.8%.

(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.

Test for success-failure condition

```
success <- 0.48*1259 >= 10
failure <- (1-0.48)*1259 >= 10

c(success, failure)
```

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

Both conditions are satisfied, and the statistic follows a normal distribution

(d) A news piece on this surveys findings states, Majority of Americans think marijuana should be legalized. Based on your confidence interval, is this news pieces statement justified?

```
c(lwr, uppr)
```

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

Response: From a confidence interval of (45.24028%, 50.75972%) we can conclude that since the lower interval falls below 50% we cannot justify the statement above.

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 ?

```
#0.02 = 1.96 * sqrt(0.48*(1-0.48)/n)

n = 0.48*0.52/(0.02/1.96)^2
n
```

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

We should survey 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.

- California : sample size, c_n1 = 11545 and sample proportion, pˆ=0.08.
- Oregon : sample size, n2 = 4691 and sample proportion, pˆ=0.088.
- 95% confidence interval : ??=1???0.95=0.05. From the normal distribution table, the required Z(0.05) value for 95% confidence level is 1.96.
- The 95% confidence interval for the difference between the proportions of Californians and Oregonians who are sleep deprived is

```
n_Cal <- 11545
n_Ore <- 4691
mean_Cal <- .08
mean_Ore <- .088
mean_diff <- mean_Cal - mean_Ore
SE <- sqrt((((mean_Cal)*(1-mean_Cal)/n_Cal)+mean_Ore)*(1-mean_Ore)/(n_Ore))

z <- 1.96
Lower_CI <- mean_diff-(z*SE)
Upper_CI <- mean_diff+(z*SE)
CI <- c(Lower_CI, Upper_CI)
CI
```

```
## [1] -0.0161073298  0.0001073298
```

There is no significant difference to conclude that there was difference in sleep deprivation between oregon and california.

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.

Response: H0: There is no difference between each forage category. HA: There is difference with at least one of the forage categories.

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

Response: A Chi-Square test can be used to answer this research question.

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

Response: Independence is met, since the cases are not dependent on each other. The sample is not randomly distributed since there are only 4 categories noted as woods.

(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
microhabitats <- c(4,16,67,345)
expctd <- c(n*0.048,n*0.147,n*0.396,n*0.409)
Chisq <- sum((microhabitats-expctd)^2/expctd)

df <- 4-1
pchisq(Chisq,df,lower.tail=FALSE)
```

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

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?

Response: A Chi-Square test can be used to answer this research question.

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

Response: H0: There is no association between coffee consumption and depression. HA: There is an association between coffee consumption and depression.

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

```
pdepressed <- (2607/50739)
pdepressed
```

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

```
pnon_depressed <- (48132/50739)
pnon_depressed
```

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

```
c(pdepressed, pnon_depressed)
```

```
## [1] 0.05138059 0.94861941
```

```
paste('only about', pdepressed*100, '% suffer from depression. The other', pnon_depressed*100, '% do not suffer fro
```

```
## [1] "only about 5.1380594808727 % suffer from depression. The other 94.8619405191273 % do not suffer from depres
```

(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 <- pdepressed*6617
Expected
```

```
## [1] 339.9854
```

```
Observed <- 373
Test_Stat <- ((Observed-Expected)^2)/Expected
Test_Stat
```

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

Response: only abut 3.2% are a contribution of the highlighted cells.

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

```
pvalue <- pchisq(20.93, 4)
pvalue <- 1-pvalue
paste('The pvalue is', pvalue)
```

```
## [1] "The pvalue is 0.000326950725917041"
```

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

Response: Since the pvalue is less than 0.05 we reject the null hypothesis (H0) in favor of the alternative hypothesis (HA).

(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.

Response: Yes i agree. We need further analysis before we can come to a conclusion. There might be other variables that affect the studies conclusion that might need to be taken into consideration.