HW6 - Categorical Data

Shyam BV

November 2, 2016

6.6, 6.12, 6.20, 6.28, 6.44, 6.48

6.6 2010 Healthcare Law.

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. We are calculating the population proption.

(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. Confidence interval is calculated for point estimate.

(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%.

TRUE. If we perform many samples and the population is not skewed, then the proportions will be between 43% to 49%

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

FALSE. CI will get narrower and the margin of error will decrease.

6.12 Legalization of marijuana, Part I. 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 an sample statistic. We have to construct CI to get the population parameter

(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.

```r
p = .48
n=1259
se <- sqrt((p*(1-p))/n)

paste("confience interval is (",round(p - 1.96*se,4),",",round(p + 1.96*se,4),")")
```

```
## [1] "confience interval is ( 0.4524 , 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.

Yes, it has to be normal distribution. Sample can be considered as normal if there are atleast 10 successes and 10 failures.

```r
n*p
```

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

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

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

Also it can be repeated multiple times for normal distribution

(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?

Yes. It can be. Because the upperlimit is 50.75%. It is little more than 50%.

6.20 Legalize Marijuana, Part II.

As discussed in Exercise 6.12, 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 ?

```r
p =.48
me = 0.02

paste("We need to sample atleast",round((1.96^2*(p*(1-p)))/me^2))
```

```
## [1] "We need to sample atleast 2397"
```

6.28 Sleep deprivation, CA vs. OR, Part I.

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

```
me <- 1.96*sqrt(((.08*(1-0.08))/11545)+((0.088*(1-0.088))/4691))

paste("confience interval difference is (",round((0.08-0.088)-me,4),",",round((0.08-0.088)+me
,4),")")
```

```
## [1] "confience interval difference is ( -0.0175 , 0.0015 )"
```

The difference in confidence interval has zero. So is no such sleep deprivation between California and Oregon residents.

6.44 Barking deer.

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
20.45	62.62	168.70	174.23	426

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

H0: The barking deer does not prefer a certain habitat to forage. HA: The barking deer has certain habitats that it prefer to forage

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

A chi-square test. Since we have cases that can be classified into several groups, we can determine if the forage habitats proportion is representative of the land make up.

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

Independence. Each case that contributes a count to the table must be independent of all the other cases in the table. We assume that all the barking deer habitat variables are independent of each other

Sample size / distribution. Each particular scenario (i.e. cell count) must have at least 5 expected cases.

Both the conditions have been satisfied

(d) Do these data provide convincing evidence that barking deer prefer to forage in certain habitats over others? Conduct an appropriate hypothesis test to answer this research question.

```
habitats <- c(4, 16, 67, 345)

z_woods <- (4-(426*.048))^2/426*.048

z_grass <- (16-(426*.147))^2/426*.147

z_forest <- (67-(426*.396))^2/426*.396
```

```
z_other <- (345-(426*.409))^2/426*.409

chisquare <- z_woods+z_grass+z_forest+z_other

 pchisq(chisquare, 3, lower.tail=FALSE)
```

```
## [1] 2.335203e-08
```

Since the p-value for the chi square distribution is very small, we can strongly reject the null hypothesis that deers have no preference in the habitats were they forage.

6.48 Coffee and Depression. 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.

Depression	$\leq$ 1 cup/wk	2-6 cups/wk	1 cup/day	2-3 cups/day	$\geq$ 4 cups/day
Yes	670	**373**	905	564	95
No	11,545	6,244	16,329	11,726	2,288
Total	12,215	6,617	17,234	12,290	2,383

a) What type of test is appropriate for evaluating if there is an association between coffee intake and depression?

chi square test to test for independence in 2 way table

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

H0: There is no relationship between coffee consumption and clinical depression.

HA: There is a relationship between coffee consumption and clinical depression.

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

The overall proportion of women who do suffer from depression is 5.14%. The overall proportion of women who do not suffer from depression is 94.86%

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.

```
exp_cnt =  6617 * 0.0514; round(exp_cnt,digit=0)
```

```
## [1] 340
```

```
obs_cnt = 373

(obs_cnt - exp_cnt)^2/exp_cnt
```

```
## [1] 3.179824
```

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

```
df = (5-1)*(2-1)
1 - pchisq(20.93,df)
```

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

f) What is the conclusion of the hypothesis test?

Based on the p-value of ~ 0.0003 is less than 0.05, So we reject 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.

Yes, I agree with his statement. Based on this study, there is a very weak relationship between coffee consumption and depression among women.