Illapani_Homework_HW6

Srini Illapani

October 30, 2015

Inference for categorical data.

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 100% confident that between 43% and 49% of the sample population agree with 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.

True. Based on the 3% margin eror and 46% of the population agree with 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%.

True, even if many random samples of 1,012 American have been taken.

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

False. At a lower confidence level, the margin of error would be smaller due to the use of a smaller Z score in the margin of error computation.

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 a sample statistic because it is not an inference of the entire population but a portion of the surveyed people.

(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
z <- qnorm(0.975)
SE <- sqrt( (p * (1-p) )/ n)

# Calculate the margin of error
moe <- z * (SE)
moe

## [1] 0.02759672

# Calculate the 95% Confidence Interval
CI <- data.frame(lower=p - moe, upper=p + moe)
CI

##       lower     upper
## 1 0.4524033 0.5075967
```

(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, a model like in this case follows a 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, the news piece is justified and the CI of 45% to 50% supports that.

6.20 Legalize Marijuana, Part II.

As discussed in Exercise ??, 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 ?

```
# Calculate the standard error SE
moe <- 0.02
z <- qnorm(0.975)
SE <- moe/z

# Calculate n
p <- 0.48
n <- (p * (1-p)) / SE^2
n

## [1] 2397.07
```

About 2,307 Americans need to be surveyed with a CI of 95% and 2% Margin of Error.

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.

```
# Data
pCa <- 0.08
pOr <- 0.088
Diff <- pOr - pCa
nCa <- 11545
nOr <- 4691

# Calculate standard error and margin of error
SE <- sqrt( ((.08 * (1 - .08)) / 11545) +  ((.088 * (1 - .088)) / 4691))
moe <- qnorm(0.975) * SE

# Calculate the 95% confidence interval
CI <- data.frame(lower=Diff - moe, upper=Diff + moe )
CI

##           lower      upper
## 1 -0.001497954 0.01749795
```

The CI interval depicts that the californians and Oregonians are sleep deprieved in similar proportion and the population size is not a factor.

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
4	16	67	345	426

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

Ho : The barking deer do not prefer to forage in certain habitats.
Ha : The barking deer prefer to forage in certain habitats.

(b) What type of test can we use to answer this research question?
chi-square goodness of fit test.

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

Woods	Cultivated grassplot	Deciduous Forests	Other	Total
20.45	62.62	168.70	174.23	426

The expected results of each type of bed site is above 5. We will assume these cases are independent and the test for inference is satisified.

(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.

We will use the set of hypothesis mentioned to answer the first question above.

```
hab <- c(4, 16, 67, 345)
exp <- c(20.45, 62.62, 168.70, 174.23)
k <- length(hab)
df <- k - 1

# Calculate the chi-square
chi <- 0
for(i in 1:k)
{
  chi <- chi + ((hab[i] - exp[i])^2 / exp[i])
}
chi
```

```
## [1] 276.6286
```

```
# Calculate p-val
p <- pchisq(chi, df=df, lower.tail=FALSE)
p
```

```
## [1] 1.135815e-59
```

The chi-square value is large and the p-value very small, hence we reject the null hypothesis. There is convincing evidence the barking deer prefer to forage in certain habitats over others.

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

		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 for two-way tables.

(b) Write the hypotheses for the test you identified in part (a).
Ho : There is no association between caffeinated coffee consumption and depression.
Ha : There is an association between caffeinated coffee consumption and depression.

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

Based on the table, the overall proportion of women who suffer from depression is 5.14% and the 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.

```
expcount <- 6617 * 0.0514

cellContribution <- (373 - expcount)^2 / expcount
paste("The The contribution to the test statistic for the highlighted cell
is:", cellContribution)
```

```
## [1] "The The contribution to the test statistic for the highlighted cell
is: 3.1798243718426"
```

(e) The test statistic is chi-square = 20.93. What is the p-value?

```
k <- 5
df <- k - 1
p <- pchisq(20.93, df=df, lower.tail=FALSE)
paste("The p-value is:", p)
```

```
## [1] "The p-value is: 0.000326950725917055"
```

(f) What is the conclusion of the hypothesis test? Since the p-value is less than 0.05, 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.64 Do you agree with this statement? Explain your reasoning.

Yes. The study was not done in a controlled environment and over the years there could have been many other factors that may have contributed to the clinical depression besides coffee including other medical and environmental conditions the population had.