Homework 6 : Inference for Categorical Data

Ramnivas Singh

04/20/2021

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

Answer:
(a) False, the confidence interval is a population, not a sample (b) This statement is true based on the definition of confidence interval (c) Close but false, we don't know what the true proportion is, which the true interval is dependent on (d) False,mMargin of error would increase as the confidence level goes down. As we are less confidence about our answer, we allow for more error.

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.
(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.
(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.
(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?

Answer:

(a) 48% sample is statistic. That is, it was calculated based on 1,259 US residents survey
(b)

```
N = 1259
p = 0.48
ci = 0.975

zscore = qnorm(ci)
se = sqrt((p * (1-p))/N)

lower_ci = p - round(zscore * se,2)
upper_ci = p + round(zscore * se,2)

cat('95% confidence interval is',lower_ci, 'to', upper_ci)
```

```
## 95% confidence interval is 0.45 to 0.51
```

The 95% confidence interval for the proportion of US residents who think marijuana should be made legal is from 45.24% to 50.76%.

(c) Yes. Because by the guidelines for a normal approximation if both proportions are greater than 10 and the data is independent. 0.48 * 1259 = 604 and 0.52 * 1259 = 654.

(d) No. The conficence interval spans between 45.2% and 50.8%. This means that we are 95% confident that the true mean falls betwen these values which are above (majority) and below (minority) 50%.

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?

Answer:

```
#me = zscore * sqrt((p(1-p))/N)
me = 0.02
survey_size = (zscore**2) * (p * (1-p)/me**2)
cat('Survey size for margin of error of 2% is', survey_size)
```

```
## Survey size for margin of error of 2% is 2397.07
```

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.

Answer:

```
pCA <- 0.08
nCA <- 11545
pOR <- 0.088
nOR <- 4691
cip <- 0.95 # Defining confidence interval

pDiff <- pOR - pCA

# Compute standard error and margin of error for the proportion difference.
SE <- ( ((pCA * (1 - pCA)) / nCA) +  ((pOR * (1 - pOR)) / nOR)) ^ 0.5
z <- qnorm(cip + (1 - cip) / 2)

me <-z * SE

# Construct the 95% confidence interval.
ci <- c(pDiff - me, pDiff + me )
```

The 95% confidence interval for the difference between the proportions of Californians and Oregonians who are sleep deprived is from -0.0015 to 0.0175. This interval overlaps 0, therefore we can conclude with a 95% confidence level that the proportions are not statistically different.

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.
(b) What type of test can we use to answer this research question?
(c) Check if the assumptions and conditions required for this test are satisfied.
(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.

Answer:

(a) H0: There is no preference for barking deer to forage in certain habitats over others. H1: There is a preference for barking deer to forage in certain habitats over others.
(b) The type of test we can use to answer this research question is a chi-squared test.
(c) The assumptions and conditions for this test are satisfied. We can assume each observation is independent. Also each scenario is expected to have at least 5 cases. We can see this for all scenarios.

```
0.048 * 426 # Woods
```

```
## [1] 20.448
```

```
0.147 * 426 # Grassplot
```

```
## [1] 62.622
```

```
0.396 * 426 # Forests
```

```
## [1] 168.696
```

```
(1 - (0.048 + 0.147 + 0.396)) * 426 # Other
```

```
## [1] 174.234
```

(d)

```
chisq.test(x = c(4, 16, 67, 345), p = c(0.048, 0.147, 0.396, 0.409))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  c(4, 16, 67, 345)
## X-squared = 272.69, df = 3, p-value < 2.2e-16
```

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?
(b) Write the hypotheses for the test you identified in part (a).
(c) Calculate the overall proportion of women who do and do 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. $(Observed - Expected)^2/Expected)$.
(e) The test statistic is $\chi^2 = 20.93$. What is the p-value?
(f) What is the conclusion of the hypothesis test?
(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: (a) The Chi-square test for two-way tables is appropriate for evaluating if there is an association between coffee intake and depression. (b)
H_0: There is no association between caffeinated coffee consumption and depression.
H_a: There is an association between caffeinated coffee consumption and depression. (c) The overall proportion of women who do suffer from depression is 5.13%. The overall proportion of women who do not suffer from depression is 94.86% (d)

```
k <- 5
df <- k - 1
Depressed <- 2607/50739
cup2.6.week.depressed <- Depressed * 6617

expCnt <- cup2.6.week.depressed
cellContrib <- (373 - expCnt)^2 / expCnt
```

The contribution to the test statistic for the highlighted cell is 3.2

(e)

```
value <- pchisq(20.93, df=df, lower.tail=FALSE)
value
```

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

(f) Based on the p-value of ~ 0.0003 is less than 0.05, I am rejecting null hypothesis and conclude there is an association between caffeinated coffee consumption and depression.
(g) Yes, I agree because not enough information is available regarding other confounding factors that could be influencing depression rates.