Chapter 6 - Inference for Categorical Data

Title: CUNY SPS MDS Data606_HW6"

Author: Charles Ugiagbe

Date: “10/16/2021”

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%.

Solution 1:

a. **FALSE:** We are 100% confidence that 46% of the sample 1,012 support healthcare law. The confidence interval of between 43% and 49% applies to the entire population and not the sample mean.

b. **TRUE:** Because the confident level of 95% tell us that 43% to 49% of Americans support Healthcare law.

c. **FALSE:** The statement we are 95% confident the 43% to 49% of Americans support healthcare law applys to the entire Population and not the sample mean.

d. **FALSE:** A Decrease in the confidence interval will also reduces the Margin of 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?

Solution 2:

a. The 48% is a sample statistic as it only capture the opinions from the sample respondent.

b. 95% C.I = 1.94

We are 95% confident that between 45.2% and 50.8% of US population thinks marijuana should be made legal

```
p <- 0.48
n <- 1259
SE <- sqrt(p*(1-p)/n)
ME <- 1.96 * SE
c(p-ME, p+ME)
```

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

c. Yes, its true for this data.

Reason: The sample size is large enough and were randomly drawn.

1. The sample's observations are independent and ramdomly drawn.
2. The success-failure condition of at least 10 successes and 10 failures in the sample is met. i.e. np ≥ 10 and n(1 − p) ≥ 10. When these conditions are met, then the sampling distribution of p(hat) is nearly normal

d. CI (0.4524, 0.5076). Since the confidence interval fall above 50%, We cannot reject the hypothesis that majority of American think marijuana should be legalized. Howerver, we can subject it to further testing.

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?

Solution 3:

$ME = 1.96\times SE = 1.96\times\sqrt{p(1-p)/n} \,.$

Our goal is to find the smallest sample size n so that this margin of error is reduce to 2%. For a 95% confidence level.

p = 0.48 ; The value z* corresponds to 1.96:

$ME = Z^{*} \times \sqrt{\frac{p(1-p)}{n}}$

$0.02 = 1.96 \times \sqrt{\frac{p(1-p)}{n}}$

$0.02 = 1.96 \times \sqrt{\frac{0.48(1-0.48)}{n}}$

$0.02^{2} = 1.96^{2} \times {\frac{0.48(1-0.48)}{n}}$

$n = 1.96^{2} \times{\frac{0.48(1-0.48)}{0.02^{2}}}$

$n = 2397.16$

From the answer, we need to 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.

Solution 4:

$Point\ Estimate(PE) = P\hat{}_{1}-P\hat{}_{2}$

$Point\ Estimate(PE) \pm Z^{*} \times SE$

$P\hat{}_{1}-P\hat{}_{2}\pm Z^{*} \times \sqrt{\frac{p_{1}(1-p_{1})}{n}+\frac{p_{2}(1-p_{2})}{n}}$

95% of $Z^{*} = 1.96$

```
p1 <- 0.08
p2 <- 0.088
n1 <- 11545
n2 <- 4691
PE <- p1-p2
SE <- sqrt((p1*(1-p1))/n1 + (p2*(1-p2))/n2)
ME <- 1.96*SE
c(PE-ME, PE+ME)
```

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

CI = (0-0.0174, 0.0014)

Since the 95% confidence interval is approximately 0, there is no statistically significant difference between the proportions of Californians and Oregonians who are sleep deprived.

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.

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.

Solution 5:

a. H0: Barking deer does not prefer to forage in certain habitats over others.

HA: Barking deer prefer to forage in certain habitats over others.

b. We can use a chi-squared test to answer this research question.

c. **Independence:** Each case that contributes a count to the table must be independent of other cases in the table

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

Wood has 4 observed cases but over 20 expected cases. so, the condition is satisfied.

d. H0: Barking deer does not prefer to forage in certain habitats over others.

HA: Barking deer prefer to forage in certain habitats over others.

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

# Compute the chi2 test statistic
chi <- sum((habitats - region ) ^ 2 / region)
chi
```

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

```
# check the chi2 test statistic and find p-val
p_Val <- 1 - pchisq(chi, df = df)
p_Val
```

```
## [1] 0
```

The chi-Square value is large enough that the p-value is 0. Hence, we reject the null hypothesis and accept the alternative hypothesis.

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.

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.

Solution 6:

a. A Chi-square test is appropriate.

(b)

\(H_{0}:\) There is no association between coffee intake and depression.

\(H_{A}:\) There is association between coffee intake and depression.

c. 5.14% of women suffers from depression while 94.86% of women do not suffer from depression.

```
# Proportion of women who suffer from depression
    2607/50739
```

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

```
# Proportion of women who do not suffer from depression
    48132/50739
```

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

d. The expected count for the highlighted cell (373)

```
expected = ((2607/50739 )*6617)
expected
```

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

Contribution of the cell to the test statistics

```
#(Observed-Expected)^2/Expected)
(373-expected)^2 /expected
```

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

e. If test statistic is \(\chi ^{2}=20.93\) What is the p-value?

```
chisq <- 20.93
    df <- (5-1)*(2-1)
    pval <- 1- pchisq(chisq, df)
    pval
```

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

f. The p-value is very small (0.0003). Since P-value < 0.05, we reject the null hypothesis that There is no association between caffeinated coffee consumption and depression. Then accept the alternative hypothesis.

g. I agree it is too early to recommend that women load up on extra coffee. Based on this study, there is a very weak relationship between coffee consumption and depression among women. further tests would need to be conducted a before we can explicitly state that coffee effectively treats depression in women.