Chapter 6 - Inference for Categorical Data

Jose. Mawyin

10/26/2019

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.

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.

False

(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. This is the only statament that truly follows the concept of confidence interval:

"95% confidence" means that in about 95 percent of the samples the true value of of the observation is contained in the confidence interval obtained from the sample.

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

True. The lower the confidence level, the higher 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.

"A parameter is a value that describes a characteristic of an entire population." while "A statistic is a characteristic of a sample." In this case, the 48% is a statistic of the 1,259 population sample used in this study rather than a characteristic of the entire population.

(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.

We can calculate the confidence interval and Standard Error using the equations below:

$$\text{Confidence Interval} = \text{point estimate } \pm z^{\star} SE$$

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

$$\text{Margin of Error } = \pm z^{\star} SE$$

```
P.E <- 0.48
z.star <- 1.96
N <- 1259
S.E.p <- (sqrt(P.E*(1-P.E)/N)) %>% round(4)
lower <- (P.E - z.star * S.E.p) %>% round(2)
upper <- (P.E + z.star * S.E.p) %>% round(2)
z.star.S.E.p <- (z.star*S.E.p) %>% round(3)
error.margin <- z.star.S.E.p*100
cat("The upper 95% confidence interval is",upper ,"and the lower is",lower,
    "with an error margin of" , error.margin,"%")
```

```
## The upper 95% confidence interval is 0.51 and the lower is 0.45 with an error margin of 2.8 %
```

(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.

The population sample for this study is large and random so it does meet the requirements for a normal distribution.

> Conditions for $\bar{x}$ being nearly normal and SE being accurate
> Important conditions to help ensure the sampling distribution of $\bar{x}$ is nearly normal
> and the estimate of SE sufficiently accurate:
> · The sample observations are independent.
> · The sample size is large: $n \geq 30$ is a good rule of thumb.
> · The population distribution is not strongly skewed. This condition can be
> difficult to evaluate, so just use your best judgement.

(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?

It is not justified to just quote a number without mentioning the accuracy of the statement (i.e. margin of error)

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 ?

$$\text{Margin of Error } = \pm z^{\star} SE$$

$$\text{Margin of Error } = \pm z^{\star} \sqrt{\frac{p(1-p)}{n}}$$

We need to solve for n.

```
Margin.Error <- 0.02
p.Q3 <- 0.48
z.star <- 1.96
n.Q3 <- (p.Q3*(1-p.Q3)*(z.star/Margin.Error)^2) %>% round(0)
cat("To have a 2% margin of error we need to survey", n.Q3, "people.")
```

```
## To have a 2% margin of error we need to survey 2397 people.
```

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.

We will use the following definions:

The standard error of the difference in sample proportions is:

$$SE_{\hat{p}_1-\hat{p}_2} = \sqrt{SE^2_{\hat{p}_1} + SE^2_{\hat{p}_2}} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

$$\text{Confidence Interval} \ = \ \text{point estimate} \ \pm z^\star SE_{\hat{p}_1-\hat{p}_2}$$

```r
n.Cal <- 11545
Prop.Cal <- 0.088
n.Or <- 4691
Prop.Or <- 0.08
SE.Or.Cal <- sqrt(  (Prop.Cal*(1-Prop.Cal)/n.Cal)  + (Prop.Or*(1-Prop.Or)/n.Or) ) %>% round(4)
z.star <- 1.96
z.star.SE.Or.Ca <- (z.star*SE.Or.Cal) %>% round(3)
error.margin <- z.star.SE.Or.Ca
P.E.Or.Cal <- Prop.Cal-Prop.Or
Upper.Interval <- P.E.Or.Cal + error.margin
Lower.Interval <- P.E.Or.Cal - error.margin
cat("The difference betwee the two proportions is:",100*P.E.Or.Cal,"%. \nThe confidence
    interval in the difference of the two proportions (California and Oregon) is:",
    Upper.Interval,Lower.Interval, "\nThe error margin percentage is:",error.margin*100)
```

```
## The difference betwee the two proportions is: 0.8 %.
## The confidence
##     interval in the difference of the two proportions (California and Oregon) is: 0.017 -0.001
## The error margin percentage is: 0.9
```

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.

> Hypothesis test for a proportion:
> Set up hypotheses and verify the conditions using the null value, p_0, to ensure $\hat{p}$ is nearly normal under H_0. If the conditions hold, construct the standard error, again using p_0, and show the p-value in a drawing. Lastly, compute the p-value and evaluate the hypotheses.

```r
Total <- 426
Other <- 345
Ratio.Other.Total <- (Other/Total) %>% round(3)
Ratio.NoOther.Total <- 1 - Ratio.Other.Total

W.Per.Cent <- 4.8
CG.Per.Cent <- 14.7
DF.Per.Cent <- 39.6
Sum.Certain.Habitats <- W.Per.Cent + CG.Per.Cent + DF.Per.Cent
Other.Per.Cent <- 100 - Sum.Certain.Habitats
cat("The null hypotheses for testing if barking deer prefer to forage in certain habitats over
    others is that the proportion of certain habitats where the deer forage match the
    proportion of certain habitats area." )
```

```
## The null hypotheses for testing if barking deer prefer to forage in certain habitats over
##     others is that the proportion of certain habitats where the deer forage match the
##     proportion of certain habitats area.
```

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

We can use a null hypotheses test to answer this research question.

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

> **Conditions for the sampling distribution of $\hat{p}$ being nearly normal:**
> The sampling distribution for $\hat{p}$, taken from a sample of size n from a population with a true proportion p, is nearly normal when
> 1. the sample observations are independent and
> 2. we expected to see at least 10 successes and 10 failures in our sample, i.e.
> $np \geq 10$ and $n(1-p) \geq 10$. This is called the success-failure condition.

The conditions above are met for this sample.

(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.

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

First, lets calculate the standard error:

```r
p.certain_habitas <- Ratio.NoOther.Total
n.all.forage <- 426
SE.Certain.Habitats <- sqrt( (p.certain_habitas*(1-p.certain_habitas)/n.all.forage) ) %>% round(4)
SE.Certain.Habitats
```

```
## [1] 0.019
```

```r
z.star.SE.Certain.Habitats <- (z.star*SE.Certain.Habitats) %>% round(3)
error.margin.per.cent <- z.star.SE.Certain.Habitats*100
error.margin.per.cent
```

```
## [1] 3.7
```

```r
cat("\nThe land percentage of other habitas is:" ,Other.Per.Cent,
    "%. The percentage of certain habitas is:", Sum.Certain.Habitats, "%.")
```

```
##
## The land percentage of other habitas is: 40.9 %. The percentage of certain habitas is: 59.1 %.
```

```r
cat("\nThe percentage of other habitas where the deer forage is:" ,Ratio.Other.Total*100,
    "%. The percentage of certain habitas where the deer forage is:", Ratio.NoOther.Total*100, "%.")
```

```
##
## The percentage of other habitas where the deer forage is: 81 %. The percentage of certain habitas wh
```

```r
cat("\nThe error margin for the percentage of certain habitas where the deer forage is:",
    error.margin.per.cent, "%" )
```

```
##
## The error margin for the percentage of certain habitas where the deer forage is: 3.7 %
```

The null hypotheses for testing if barking deer prefer to forage in certain habitats over others is that the proportion of certain habitats where the deer forage match the proportion of certain habitats area.

The null hypotheses failed. However, we found that the barking deer prefer not "certain habitas" but what is considered as "other habitas"

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	2-6	1	2-3	≥ 4	
		cup/week	cups/week	cup/day	cups/day	cups/day	Total
Clinical	Yes	670	$\boxed{373}$	905	564	95	2,607
depression	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?

Since the data is binned, we can use a Chi-Squared Test.

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

H_0 : There is no association between coffee intake and depression.

H_A : There is an association between coffee intake and depression.

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

```
Total.Women <- 50736
Women.W.Depression <- 2607
Women.W.N.Depression <- 48132
Prop.Women.W.Depression <- (100*Women.W.Depression/Total.Women) %>% round(2)
Prop.Women.W.N.Depression <- (100*Women.W.N.Depression/Total.Women) %>% round(2)
cat(Prop.Women.W.Depression,"% of women suffer from depression while,",
    Prop.Women.W.N.Depression, "% of women do not.")
```

```
## 5.14 % of women suffer from depression while, 94.87 % of women do not.
```

(d) Identify the expected count for the highlighted cell, and calculate the contribution of this cell to the test statistic, i.e. $(Observed - Expected)^2/Expected$).

```
Cof.Addic.Expected <- (2607*(6244/48132)) %>% round(0)
Cof.Addic.Observed <- 373
T.Stat <- ((Cof.Addic.Observed - Cof.Addic.Expected)^2)/Cof.Addic.Expected
cat("The expected count is:",Cof.Addic.Expected, "\nThe observed count is:",Cof.Addic.Observed,
    "\nThe contribution to the test statistic is:",T.Stat)
```

```
## The expected count is: 338
## The observed count is: 373
## The contribution to the test statistic is: 3.62426
```

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

```r
chi.squared <- 20.93
df.caff <- 5
PVal.Caff <- pchisq(chi.squared, df=df.caff, lower.tail=FALSE) %>% round(5)
cat("The p value given the test statistic value is", PVal.Caff, "and is much
    less than the significance level of 0.05. Therefore, we discount the null hypotheses.")
```

```
## The p value given the test statistic value is 0.00084 and is much
##     less than the significance level of 0.05. Therefore, we discount the null hypotheses.
```

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

The null hypotheses is discounted.

$$H_A: \text{ There is an association between coffee intake and depression.}$$

(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.

I disagree with the author. There is a correlation between coffee intake and depression.