Chapter 6 - Inference for Categorical Data

Subhalaxmi Rout

03/15/2010

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:** Confidence interval is about a population not a sample.

(b) **True**

(c) **False:** Confidence Interval is about population propertion not about sample statistic.

(d) **False:** If the confidence interval will be narrow then the margin of error will be smaller.

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*) Yes 48% is a sample statistics becasuse it estimates the population propertion (p).

(*b*)

```
p <- 0.48
n <- 1259
SE <- sqrt(p *(1 -p)/n)
margin_error <- SE * 1.96
upper_bound <- round(p + margin_error,3)
lower_bound <- round(p - margin_error,3)
upper_bound
```

```
## [1] 0.508
```

```
lower_bound
```

```
## [1] 0.452
```

95% CI range is (.452, .508)

45.2% to 50.8% US residents think marijuana should be made legal

(*c*) Central Limit Theorem (CLT) tells us

- The observations in the sample are independent
- The sample size is sufficiently large (checked using the success/failure condition: np 10 and n(1−p) 10)

We can assume residents were selected in a random process because we do not have that information. Given sample size (n) = 1259 success: 1259 * 0.48 = 604.32 failure: 1259 * (1 - 0.48) = 654.68 Both success and failure 10.

So, this is TRUE i.e normal model is a good approximation.

(*d*) No, news piece statement is not justified. The confidence interval is between 45.2% to 50.8%. So, 45.2% to 50.8% of Americans think that marijuana should be legalized which is just barely covers half of Americans at the top of the range. So we cannot accurately say that majority of people support legalization.

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

Given

```
p <- 0.48
margin_error <- 0.02
z <- 1.96
SE <- margin_error / z
sample <- (p * (1-p) ) / (SE ^ 2)
sample
```

```
## [1] 2397.158
```

Almost 2398 Americans need for survey where margin of error of a 95% confidence interval to 2%.

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

Given

```
p_cal <- 0.08
p_ore <- 0.088

p_prop <- p_cal - p_ore

sample_cal <- 11545
sample_ore <- 4691

SE_cal <- (p_cal * (1-p_cal)) / sample_cal
SE_ore <- (p_ore * (1-p_ore)) / sample_ore

SE_prop <- sqrt(SE_cal + SE_ore)
margin_error <- 1.96 * SE_prop
margin_error
```

```
## [1] 0.009498128
```

```
upper_bound <- round(p_prop + margin_error,3)
lower_bound <- round(p_prop - margin_error,3)

upper_bound
```

```
## [1] 0.001
```

```
lower_bound
```

```
## [1] -0.017
```

CI range is (-0.017,0.001).

The proportion of Californians and Oregonians who are sleep deprived is between -0.017 and 0.001.

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.
(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

Creat a table to understand the given data.

- Expected value for Woods: 426 * (0.048)
- Cultivated grassplot: 426 * (0.147)
- Deciduous forests: 426 * (0.309)
- Other: 426 * (1 - (0.048 + 0.147 + 0.309))

```
library("DT")
observed <- c(4, 16, 67, 345, 426)
expected_prop <- c(0.048, 0.147, 0.396, 0.409, 1)
expected <- expected_prop * 426
deer <- rbind(observed, expected)
colnames(deer) <- c("Woods", "Cultivated grassplot", "Deciduous forests", "Other", "Total")

datatable(deer)
```

Show 10 entries Search:

	Woods	Cultivated grassplot	Deciduous forests	Other	Total
observed	4	16	67	345	426
expected	20.448	62.622	168.696	174.234	426

Showing 1 to 2 of 2 entries Previous 1 Next

(*a*) Hypothesis conditions

- H0 (null hypothesis): Barking deer has no preference of habitats for foraging.
- HA (alternative hypothesis): Barking deer has preference of certain habitats over others for foraging.

(*b*) I think, chi-square test can be the best fit for answer this research question.

(*c*) Condition for chi-square test:

- Independence:

1. Sampled observation must be independence

2. Random sample
3. n < 10% of population

- Sample size:

1. Each perticular scinario must have atleast 5 expected case.

Given table, each case that shows a count to the table is independent of all the other cases in the table.

Sample size: we can see in the `deer` table expected row all values greater than 5.

```
datatable(deer)
```

Show 10 entries Search:

	Woods	Cultivated grassplot	Deciduous forests	Other	Total
observed	4	16	67	345	426
expected	20.448	62.622	168.696	174.234	426

Showing 1 to 2 of 2 entries Previous 1 Next

Above assumtion we can say Chi-squared goodness of fit test.

(d) degree of freedom = k - 1 = 4 -1 = 3

```
df <- 3
chi_square <- sum(((observed - expected) ^ 2)/expected)

p_value <- 1 - pchisq(chi_square, df)
p_value
```

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

The `p` value is 0, so we can reject the null 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.

		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*) Chi-square test can be appropriate for evaluate if there is an association between coffee intake and depression.

(*b*) Hypothesis test:

- H0 (null hypothesis): There is no relationship between coffee intake and clinical depression
- HA (alternate hypothesis): There is a relationship between coffee intake and clinical depression.

(*c*) Proportion of women who suffer from depression:

```
prop_depres <- 2607 / 50739
prop_depres
```

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

Proportion of women who do not suffer from depression:

```
prop_no_depres <- 48132 / 50739
prop_no_depres
```

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

(*d*) Expected Count for highlighted cell and test statistics:

```
observed <- 373
expected <- (2607/50739)*6617
expected
```

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

```r
hightlighted_cell <- sum(((observed - expected) ^ 2)/expected)
hightlighted_cell
```

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

Expected Count for highlighted cell: 340

Test statistics = 3.2

(e) The test statistic is $\chi^2 = 20.93$

Degree of freedom: (5 - 1) * (2 - 1) = 4

```r
chisq <- 20.93
df <-  4

p_value <- 1-pchisq(chisq, df)
p_value
```

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

The `p value` is 0.0003269507.

(f) Due to low `p-value` we reject the null hypothesis that there is no relationship between clinical depression and coffee intake.

(g) I do agree with author's statement because this test was an observational study and therefore we cannot assume causation. To conclude there is a relation, experiments need to be conducted.