

## Chapter 6 - Inference for Categorical Data

Donald Butler

10/17/2021

**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.

I suppose this is technically true, because we are 100% confident that the sample mean is 46%.

- (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, the confidence interval is  $\hat{p} \pm ME$ .

- (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%.

False, other sample proportions could be outside of this range and yet still contain the population proportion.

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

False, the margin of error would decrease with the confidence level.

---

**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.

Sample Statistic

- (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.

```
ME = 1.96 * sqrt((.48*.52)/1259)
```

```
(.48 - ME)
```

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

```
(.48 + ME)
```

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

We are 95% confident that the percentage of Americans that support legalizing marijuana is between 45.2% and 50.7%.

- (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.

A normal distribution is a good approximation because the responses are independent and both  $pn$  and  $(1-p)n$  are greater than 10.

- (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?

While the statement may be true, the confidence interval doesn’t suggest that it is.

**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?

$$ME = 1.96 \times \sqrt{\frac{.48 * .52}{n}} \leq .02$$
$$48.96 \leq \sqrt{n}$$

So, we would need to survey about 2400 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 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.

```
ME = 1.96 * sqrt((.08 * .92)/11545 + (.088 * .912)/4691)
((.008) - ME)
```

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

```
((.008) + ME)
```

```
## [1] 0.01749813
```

We are 95% confident that the difference in population percentage of sleep deprivation between California and Oregon to be between -.15% and 1.7%.

**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.

$H_0$  : Barking deer have no preference for foraging in certain habitats over others.

$H_A$  : Barking deer do have a preference for foraging in certain habitats over others.

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

We can perform a chi-squared test.

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

Each observation is independent of the others and calculating the expected values based on land percentages, we expect at least 5 for each category.

```
# expected woods
(.048 * 426)
```

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

```
# expected cultivated grassplot
(.147 * 426)
```

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

```
# expected deciduous forests
(.396 * 426)
```

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

```
# expected other
((1-.048-.147-.396) * 426)
```

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

(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.

```
chisq.test(x = c(4,16,61,345),p = c(.048,.147,.396,1-.048-.147-.396))
```

```
##  
## Chi-squared test for given probabilities  
##  
## data:  c(4, 16, 61, 345)  
## X-squared = 284.06, df = 3, p-value < 2.2e-16
```

The p-value is less than .05, so we reject  $H_0$ , and conclude that barking deer prefer to forage in certain habitats over others.

---

**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.

		<i>Caffeinated coffee consumption</i>					Total
		$\leq 1$	2-6	1	2-3	$\geq 4$	
		cup/week	cups/week	cup/day	cups/day	cups/day	
<i>Clinical depression</i>	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?

We can perform a chi-squared test.

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

$H_0$  : There is no association between coffee consumption and depression.

$H_A$  : There is an association between coffee consumption and depression.

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

```
(2607/50739)
```

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

```
(48132/50739)
```

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

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

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

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

```
observed = 373
```

```
((observed - expected)^2/expected)
```

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

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

```
(1 - pchisq(q = 20.93, df = (2-1)*(5-1)))
```

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

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

Since the p-value of .0003 is less than 5%, we reject  $H_0$ , and conclude that there is an association between coffee consumption and depression.

- (g) One of the authors of this study was quoted on the NY Times 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.

Yes, because this was an observational study instead of clinical trial, causation cannot be determined.