Chapter 6 - Inference for Categorical Data

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.
p**=.46 ,** n**=1,012 ,** $Z - score$**=1.96**

$SE = \sqrt(\frac{p(1-p)}{n})$ **or** $SE = \sqrt(\frac{.46(1-.46)}{1,012})$

Critical Interval = $p \pm (Z \times SE)$ **or Critical Interval** = $.46 \pm (1.96 \times 0.016)$

Which equates to a CI of 0.4292933-0.4907067 or roughly 43%-49%

(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.
ANSWER: FALSE. The Critical Intervals of 43%-49% are a reflection of the **POPULATION** but **NOT** the sample.

(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.
ANSWER: TRUE. This defines the meaning of CI.

(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%.
ANSWER: FALSE Only because the CI again reflects the population proportion which is unknown and not the sample proportion so this assumption cannot be made.

(d)

The margin of error at a 90% confidence level would be higher than 3%.
ANSWER:FALSE Confidence levels and margin of error have a **DIRECT RELATIONSHIP**, so as confidence levels decrease, so does 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.
p**=.48 ,** n**=1,259 ,** $Z - score_{(b)}$**=1.96**

SE=$\sqrt{(}\frac{p(1-p)}{n})$ **or** SE=$\sqrt{(}\frac{.48(1-.48)}{1{,}259})$

Critical Interval $= p \pm (Z_{(b)} \times SE)$ **or Critical Interval** $= .48 \pm (1.96 \times 0.01408022)$

Which equates to a CI of 0.4524033 - 0.5075967 or roughly 45%-51%

(a)

Is 48% a sample statistic or a population parameter? Explain.
ANSWER: 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.
ANSWER: CI **= 0.4524033 - 0.5075967 or roughly 45%-51%**

(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.
ANSWER: NO. Both INDEPENCE and SUCCESS-FAILURE conditions are met, therefore we can treat the distribution as nearly normal.

Independence. There are n = *1000* observations for each sample proportion $\hat{p}$, and each of those observations are independent draws. The most common way for observations to be considered independent is if they are from a simple random sample.
Success-failure condition. We can confirm the sample size is sufficiently large by checking the success-failure condition and confirming the two calculated values are greater than 10:
np $= 1,259 \times 0.48 = 604.32 \geq 10$ **n(1-p)**$= 1,259 \times (1 - 0.48) = 654.68 \geq 10$

(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?

ANSWER:NO. I would say the critical intervals fall just short of in favor, with the Upper CI just etching over 50%. The majority of the range staying below implying its close but **MOST** likely not representative of the **MAJORITY** of the true population.

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 = 2% or 0.02
Solving for n **in** $n = Z^2 \times \frac{p(1-p)}{ME^2} < n$ **or** $1.96^2 \times \frac{0.48(1-0.48)}{0.02^2} < n$
$n =$ **2397.158**

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.

n_{cali}**=11,545 ,** n_{Oregon}**=4,691 ,** p_{cali}**=.08 ,** p_{oregon}**=.088 ,** $p_{combined}$**=.08-0.88 = -0.008**

SE=$\sqrt(\frac{p_{cali}(1-p_{cali})}{n_{cali}} + \frac{p_{oregon}(1-p_{oregon}}{n_{oregon}})$ **or SE=**$\sqrt(\frac{.08(1-.08)}{11{,}545} + \frac{.088(1-.088)}{4{,}691})$

ME= $Z \times SE$ **or** 1.96×0.009498128

Finally CI=$p_{combine} \pm ME$ **which equals to:**
ANSWER

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

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

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.
ANSWER:
H_0: **Barking Deers do not prefer forage in certain habitats over others**
H_a: **Barking Deers show preference in habitats**

(b)

What type of test can we use to answer this research question?
ANSWER:
Chi-square test

(c)

Check if the assumptions and conditions required for this test are satisfied.
ANSWER: All necessary conditions met (refer below)
Independence: We assume yes since the forage and habitat are distinct.
Sample size/distribution: Expected cases ≥ 5.

(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.
ANSWER: p-value of the data is small enough to support that deer favor certain habitats for foraging

```
## [1] "n = 426 and Habitats = "
```

```
## [1] "4"   "16"  "61"  "345"
```

$E(X)_{regions} = n \times p_{region}$ **or:**

```
## [1]  20.448  62.622 168.696 174.234
```

```
## [1] "Chi Square is equal to 284.06094954246"
```

```
## [1] "p-value equals to 2.79972424857738e-61"
```

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 depression	Yes	670	$\boxed{373}$	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?
ANSWER: Chi Square test

(b)

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

H_0**: No relationship between coffee and depression**
H_a**: Relationship between coffee and depression**

(c)

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

$Depressed_{women}$**=2607** $NotDepressed_{women}$**=48132**

```
## [1] "The overal proportion equals to 0.0541635502368487 or 5.14%"
```

(d)

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

```
## [1] "Expected count is 3.20591443200495"
```

(e)

The test statistic is $\chi^2 = 20.93$. What is the p-value? **Degrees of freedom** $(rows-1)(columns-1)$ **or** $(5-1)(2-1)$ **or just** $4 \times 1 = 4$
ANSWER:

```
## [1] "When using 1-pchisq I find the p-value to be 0.000326950725917041"
```

(f)

What is the conclusion of the hypothesis test?
p-vlue $<$ 0.05 $\therefore$ we reject the null hypothesis

(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.

Agreed, more testing with other variables taken into account should be conducted