Chapter 6 - Inference for Categorical Data

David Simbandumwe

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(infer)


rm(list=ls())
```

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 - we are 100% confident that 46% of Americans in the sample agree with the decision**

(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 - that is the definition of the 95% confidence interval**

(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 - expected outcome of taking randome samples of a population**

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

- **false - the margin of error would be lower since the z value for .90 confidence is lower than the z value for .95 confidence**

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.

- **48% is a sample statistic because it is a measure of the 1,259 sample respondents from the us 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.

```
n <- 1259
p_hat = 0.48
(se = round(sqrt(p_hat * ( 1 - p_hat) / n),4))
```

```
## [1] 0.0141
```

```
z <- round(-qnorm((1-.95)/2),2)


(ci <- c( round(p_hat - z * se,4),round(p_hat + z * se,4)))
```

```
## [1] 0.4524 0.5076
```

- **There is a 95% chance that the proportion of the american population that supports legalizing marijuana is between 0.452 and 0.508**

(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.

- **yes if you take a random sampling of the american population. the american population should follow a normal distribution**

(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?

- **The confidence interval includes p $<$ 0.5 as possible proportions for the population so the non qualified affirmative statement is not justified.**

H0 - a majority of Americans don not support legalization of marijuana (p <= 0.5) H1 - a majority of Americans support legalization of marijuana (p > 0.5)

```
p_hat <-  0.48
pt <- 0.5
n <- 1259

SE <- sqrt((0.5*0.5)/n)
ts <- (0.48 - 0.5) / SE
```

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? **- 2398**

```
p = 0.48
z <- 1.96
ME <- 0.02


(n <- ((ME/z)^2 / (p * (1-p)))^-1)
```

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

```
sqrt((p*(1-p))/n) * z
```

```
## [1] 0.02
```

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.

```
p_c <- .08
n_c <- 11545

p_o <- 0.088
n_o <- 4691

PE <- p_c - p_o

SE <- sqrt( ((p_c*(1-p_c))/n_c ) + ((p_o*(1-p_o))/n_o) )

z <- round(-qnorm((1-0.95)/2),4)


(CI <- c(round(PE-z*SE,4), round(PE+z*SE,4)))
```

```
## [1] -0.0175  0.0015
```

```
# pooled proportion
(p_pool <- ((p_c*n_c) + (p_o*n_o)) / (n_c + n_o))
```

```
## [1] 0.08231141
```

```
# p-value
z1 <- sqrt(-PE / SE)
p_value <- 2 * pnorm(z1)
```

- **We are 95% confident that the proportion of residents in California and Oregon who reported insufficient sleep in the past 30 days is between -0.0175 and 0.0015.**

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.

- **H0 - barking deer have no preference for certain habitats. the foraging proportions are equal to the land proportions**
- **H1 - barking deer have a preference for certain habitats. the foraging proportions are not equal to the land proportions**

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

- **Hypothesis test for proportions**

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

- **Independence - samples should be independent. taken at random**
- **Sample Size - sample size should be less that 10% of the population**
- **Successes / failurs > 10 - there should be at least 10 expected successes and 10 expected failures**

(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.

- ---

```
data <- tibble (
    habitat = c("wood","cultivate","forest","other"),
    habitat_num = c(4, 16, 61, 345),
    forage_p = c( 0.0480,    0.1470,     0.3960,     0.4090 )
)

data <- data %>%
    mutate(
        habitat_p = habitat_num / sum(habitat_num),
        forage_num = round(forage_p * sum(habitat_num),0)
    )

X2 <- sum((data$forage_num- data$habitat_num)^2/data$habitat_num)
df <- 4 - 1

p_value <- pchisq(X2, df, lower.tail = FALSE)

paste0("p_value = ", p_value)
```

```
## [1] "p_value = 2.74925329181142e-103"
```

```
# null_dist <- data %>%
#   specify(habitat_p ~ forage_p) %>%
#   hypothesize(null = "independence") %>%
#   generate(reps = 1000, type = "permute")
```

- **Since the p-value is lower than the level of significance we reject the null hypothesis in favour of the alternate 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	$\boxed{373}$	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

```
caffeine_df <- tibble (
    caffeine = c("less1_week", "2-6week", "1_day", "2-3_day", "more4_day"),
    yes = c(670, 373, 905, 564, 95),
    no = c(11545, 6244,16329,11726,2288)
)
```

(a) What type of test is appropriate for evaluating if there is an association between coffee intake and depression?

- **chi squared test of independence**

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

- **H0 - caffeine consumption does not reduce the rate of depression in women**
- **H1 - caffeine consumption does reduce the rate of depression in women**

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

```
# suffer from depression
dep_p <- round(sum(caffeine_df$yes) / sum(caffeine_df$no, caffeine_df$yes),4)


# do not suffer from depression
nodep_p <- round(sum(caffeine_df$no) / sum(caffeine_df$no, caffeine_df$yes),4)
```

- **5.14% of women suffer from depression**
- **94.86% of women 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. $(Observed - Expected)^2/Expected$).

```
obs <- 373
row_tot <- 6617
exp_val <- row_tot * dep_p

cont <- (obs - exp_val)^2 / exp_val
```

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

```
chi_2 <- 20.93
df <- (5-1) * (2 -1 )

pchisq(chi_2, df = df,lower.tail = FALSE)
```

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

- **The p-value is 0.0003**

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

- **p value of 0.007 is less than 0.05 confidence interval so we should reject the null hypothesis. caffeine consumption does not impact the rates of 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.

- **there is not enough evidence in the sample to reject the alternate hypothesis that caffeine consumption reduces the rate of depression**