Homework 6

Irene Jacob

2020-10-10

2010 Healthcare Law

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. Population proportion estimate is done using confidence intervals and not sample proportion.

(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. Population proportion estimate is done using confidence intervals and not sample proportion.

(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.

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

FALSE. It will be a lower margin.

Legalization of marijuana, Part I

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.

It is sample statistic. This is because it is estimated from the sample.

(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 <- 0.48
q <- 1-p
SE <- sqrt((p*q)/n)
lower = 0.48 - (qnorm(0.975)*SE)
upper = 0.48 + (qnorm(0.975)*SE)

c(lower, upper)

## [1] 0.4524033 0.5075967
```

The confidence interval is (0.4524033,0.5075967). The proportion of US residents who think marijuana should be made legal is between 45.24% and 50.76%.

(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.

```
s <- p*1259 >= 10
f <- q*1259 >= 10

c(s, f)
```

```
## [1] TRUE TRUE
```

it is a normal distribution as both the conditions are satisfied

(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?

```
c(lower, upper)
```

```
## [1] 0.4524033 0.5075967
```

Legalize Marijuana, Part II

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 <- 0.02
SE <- ME / 1.96
n <- ((p * q)) / (SE^2)
n
```

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

Sleep deprivation, CA vs. OR, Part I

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.

```
p1 <- 0.08
p2 <- 0.088
q1 <- 1-p1
q2 <- 1-p2
n1 <- 11545
n2 <- 4691

SE1 <- sqrt(((p1*q1)/n1) + ((p2*q2)/n2))

(p1 - p2) - 1.96 * SE1
```

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

```
(p1 - p2) + 1.96 * SE1
```

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

Barking deer

Micro habitat 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 HA: barking deer prefer foraging in a specific type of habitat

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

Chi-square goodness of fit test can be used here for hypothesis.

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

```
O <- c(4, 16, 61, 345, 426)
E_prop <- c(0.048, 0.147, 0.396, 1-0.048-0.147-0.396, 1)
E <- E_prop * 426
E
```

```
## [1]   20.448  62.622 168.696 174.234 426.000
```

(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.

Each expected value is above 5 so the behavior of the barking deer are likely independent.

```
k <- 4
df <- k-1
chi <- sum(((O - E)^2)/E)
( p_deer <- 1 - pchisq(chi, df))
```

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

The p value is) so the null hypothesis can be rejected.

Coffee and Depression

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	Yes	670	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?

Chi-square test for the two-way table can be used for evaluating if there is an association between coffee intake and depression.

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

H0: The risk of depression in women is the same regardless of amount of coffee consumed. HA: The risk of depression in women varies depending on amount of coffee consumed.

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

Proportion of women who suffer from depression: $2607/50739 = 0.051$ Proportion of women who do not suffer from depression: $48132/50739 = 0.949$

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

```
O1 <- 373
E1 <- (2607/50739)*6617
highlight <- sum(((O1 - E1)^2)/E1)
highlight
```

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

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

```
chi1 <- 20.93
df1 <-  (5-1)*(2-1)

p3 <- 1-pchisq(chi1, df1)
p3
```

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

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

Null hypothesis is rejected as the p value is very small.

(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 agree with this because this is an observational study.