DATA606_HW6_RJM

RJM

2019-11-01

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.

6.48(a)

(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 By using this sample, we are trying to estimate the confidence interval for the whole population. 46% Americans from the sample agree with the decision, but when we consider the population then there is a 3% margin of error.

6.48(b)

(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 With a 3% margin of error, the poll result point toward 43% and 49% of Americans supporting the decision with a 95% condifdence interval.

6.48(c)

(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 95% confidence interval means that the population has the range between 43% and 49%. This might be true for some samples or not.

6.48(d)

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

False The margin of error will be lower as the confidence interval is 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.

6.10 (a)

(a) Is 48% a sample statistic or a population parameter? Explain.

It is a sample statistic which could help us determine the mean for the population.

6.10 (b)

(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.

```
# Collecting the required data:

n_a <- 1259

prob <- 0.48

# Based on 95% CI, z is:

z_val_a <- 1.96

std_err_a <- sqrt((prob * (1 - prob)) / n_a)
low_CI_a <- prob - (z_val_a * std_err_a)
up_CI_a <- prob + (z_val_a * std_err_a)

low_CI_a
```

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

```
up_CI_a
```

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

```
# The CI is between 45.24% and 50.76%.
```

6.10 (c)

(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.

Since the observations in the sample are independent ($< 10\%$ of the population) and have adequate representation (> 10), the assumption of normal distribution is reasonable.

6.10 (d)

(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?

With 50.76% as the upper limit of the CI, there is a possibility of majority of Americans agreeing with the legalization of marijuana.

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?

6.16

```
me_b <- 0.02
z_val_b <- qnorm(0.975)

std_err_b <- me_b/z_val_b

n_b <- (prob * (1 - prob)) / std_err_b^2
ceiling(n_b)
```

```
## [1] 2398
```

```
# 2398 Americans will have to be surveryed for a 2% margin of error.
```

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.

6.22

```
# California residents
n_c <- 11545
prob_c <- 0.08

# Oregon residents
n_o <- 4691
prob_o <- 0.088

prob_d <- prob_o - prob_c


std_err_c <- sqrt( ((prob_c * (1 - prob_c)) / n_c) +  ((prob_o * (1 - prob_o)) / n_o))
me_c <- qnorm(0.975) * std_err_c

low_CI_c <- prob_d - me_c
low_CI_c
```

```
## [1] -0.001497954
```

```
up_CI_c <- prob_d + me_c
up_CI_c
```

```
## [1] 0.01749795
```

```
# The CI is between -0.15% and 1.75%. The CI could be interpreted to mean that the difference
# between the proportions could be .15% at the lower end and 1.75% at the upper end.
# Since CI has 0% in it, there is a possibility that the Californians and Oregonians have the same
# levels of sleep deprivation.
```

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

6.34(a)

(a) Write the hypotheses for testing if barking deer prefer to forage in certain habitats over others.

H_0: Barking deer have no preference of habitats for forage. H_A: Barking deer have prefer certain habitats over others for forage.

6.34(b)

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

A chi-squred test is best in this situation where we have to assess the difference between the expected and observed frequences of habitat preference amongst multiple choices.

6.34(c)

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

We make the assumption that the observations are independent of each other, and each category has at least 5 recorded instances. This is true in the case of the lowest percentage of expected occurences for woods, which is at 20.45 (4.8% * 426) based on the ratio of sites. The assumptions and conditions are satisfied.

6.34(d)

(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.

```
obs_reported <- c(4, 16, 67, 345)
obs_ratio <- c(0.048*426, 0.147*426, 0.396*426, 0.409*426)

chi_deer <- sum((obs_reported - obs_ratio ) ^ 2 / obs_ratio)

p_value_d <- 1 - pchisq(chi_deer, 3)
p_value_d
```

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

P-value is dropping to 0 means which the null hypothysis will be rejected that barking deer have no preference for habitats to forage.

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

6.50(a)

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

A chi-square test is best suited for this situation to determine the association between coffee intake and deppression.

6.50(b)

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

H_0: There is no association between coffee intake and depression. H_A: There is an association between coffee intake and depression.

6.50(c)

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

```
# The proportion of women not sufferring from depression is:
prop_w <- (48132/50739) * 100

prop_w
```

```
## [1] 94.86194
```

```
# The proportion of women sufferring from depression is:


prop_d_w <- (2607/50739) * 100

prop_d_w
```

```
## [1] 5.138059
```

```
# The proportion of women not suffering from depression is 94.86%.

# The proportion of women suffering from depression is 5.14%.
```

6.50(d)

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

```
# The exoected count is 340.

exp_count <- 6617 * 0.0514

floor(exp_count)
```

```
## [1] 340
```

```
obs_count <- 373

contribution_d <- ((obs_count - exp_count)^2 / exp_count)

contribution_d
```

```
## [1] 3.179824
```

```
# The contribution of this cell to the test statistics is 3.18%.
```

6.50(e)

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

```
# The degree of freedom is:

deg_free <- (5-1) * (2-1)
deg_free
```

```
## [1] 4
```

```
# The degree of freedom is 4.


# The p-value is:
p_value_w <- 1 - pchisq(20.93, deg_free)

p_value_w
```

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

```
# The p-value is 0.0003.
```

6.50(f)

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

The null hypothese is rejected (due to the small value of p) that there is no association between the coffee intake and depression.

6.50(g)

(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.

Yes, as the rejection of the null hypothesis does not necessite the confirmation of alternative hypothesis. We will need more data to establish that relationship.