Estadística

U4 4B

David Aarón Ramírez Olmeda

Marzo 2023

6.16 Is college worth it? Part I. Among a simple random sample of 331 American adults who do not have a four-year college degree and are not currently enrolled in school, 48% said they decided not to go to college because they could not afford school.

(a) A newspaper article states that only a minority of the Americans who decide not to go to college do so because they cannot afford it and uses the point estimate from this survey as evidence. Conduct a hypothesis test to determine if these data provide strong evidence supporting this statement.

(b) Would you expect a confidence interval for the proportion of American adults who decide not to go to college because they cannot afford it to include 0.5? Explain.

(a) Let p be the true proportion of American adults who decide not to go to college because they cannot afford it. The null hypothesis is that p = 0.5, while the alternative hypothesis is that p < 0.5. This is a one-tailed test since we are interested in whether the proportion is less than 0.5.

We can use a normal approximation to the binomial distribution since np = 331 × 0.5 = 165.5 and n(1 - p) = 331 × 0.5 = 165.5 are both greater than 10.

The test statistic is:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.48 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{331}}} = -0.727736$$

Using a significance level of 0.05, the critical value is 0.2327. Since -1.47 > 0.2327, we do not have enough evidence to reject the null hypothesis. The data does not provide sufficient evidence to claim that only a minority of Americans who choose not to go to college do so because they cannot afford it

(b) Yes. Since the sample proportion is 0.48 and the sample size is 331, the standard error is:

$$SE = \sqrt{\frac{0.48(1 - 0.48)}{331}} = 0.02746$$

Using a 95% confidence level, the margin of error is:

$$ME = 1.96 \times SE = 0.0538$$

The confidence interval is:

$$CI = \hat{p} \pm ME = 0.48 \pm 0.0538 = (0.426, 0.533)$$

6.23 Social experiment, Part I. A "social experiment" conducted by a TV program questioned what people do when they see a very obviously bruised woman getting picked on by her boyfriend. On two different occasions at the same restaurant, the same couple was depicted. In one scenario the woman was dressed "provocatively" and in the other scenario the woman was dressed "conservatively". The table below shows how many restaurant diners were present under each scenario, and whether or not they intervened.

		Scenario		
		Provocative	Conservative	Total
Intervene	Yes	5	15	20
	No	15	10	25
	Total	20	25	45

Explain why the sampling distribution of the difference between the proportions of interventions under provocative and conservative scenarios does not follow an approximately normal distribution.

1. The sample sizes under the two scenarios are small, 20 and 25. According to the central limit theorem, the sampling distribution of the difference between two sample proportions is approximately normal when the sample sizes are large enough (at least 10) and the underlying population is not extremely skewed. In this case, the sample sizes are small, and the central limit theorem may not hold.

2. The two samples may not be independent of each other. The intervention of one diner in one scenario could influence the intervention of another diner and so on.

3. Each particular scenario must have at least 5 expected cases. Therefore, the condition/tip of having all expected cell counts greater than or equal to 5 is not satisfied.

Therefore, due to the small sample sizes and the probable lack of independence between the samples, the sampling distribution does not follow an approximately normal distribution.

6.43 Rock-paper-scissors is a hand game played by two or more people where players choose to sign either rock, paper, or scissors with their hands. For your statistics class project, you want to evaluate whether players choose between these three options randomly, or if certain options are favored above others. You ask two friends to play rock-paper-scissors and count the times each option is played. The following table summarizes the data: Rock Paper Scissors 43 21 35 Use these data to evaluate whether players choose between these three options randomly, or if certain options are favored above others. Make sure to clearly outline each step

We can use a chi-square goodness of fit test, this test compares the observed frequencies with the expected frequencies under the assumption of randomness.

Null hypothesis (H0): The players choose between rock, paper, and scissors randomly. Alternative hypothesis (Ha): The players do not choose between rock, paper, and scissors randomly.

Assuming a significance level of 0.05 and three categories, the degrees of freedom for this test are 2 (number of categories minus 1).

Under the null hypothesis, each category is expected to have an equal probability of being chosen, which would result in each category having an expected frequency of 33.33 (100/3).

Calculate the test statistic and p-value

```r
data <- c(43, 21, 35)
expected <- c(1/3, 1/3, 1/3)
result <- chisq.test(x = data, p = expected)
print(result)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  data
## X-squared = 7.5152, df = 2, p-value = 0.02334
```

Since the p-value is less than the significance level of 0.05, we can reject the null hypothesis that the players choose between rock, paper, and scissors randomly. Therefore, we have evidence to suggest that certain options may be favored over others. However, it's important to note that this conclusion is based on a small sample size of only two players, and we cannot generalize these findings to the larger population of players.

6.49 Shipping holiday gifts. A December 2010 survey asked 500 randomly sampled Los Angeles residents which shipping carrier they prefer to use for shipping holiday gifts. The table below shows the distribution of responses by age group as well as the expected counts for each cell (shown in parentheses).

		Age						
		18-34		35-54		55+		Total
Shipping Method	USPS	72	(81)	97	(102)	76	(62)	245
	UPS	52	(53)	76	(68)	34	(41)	162
	FedEx	31	(21)	24	(27)	9	(16)	64
	Something else	7	(5)	6	(7)	3	(4)	16
	Not sure	3	(5)	6	(5)	4	(3)	13
	Total	165		209		126		500

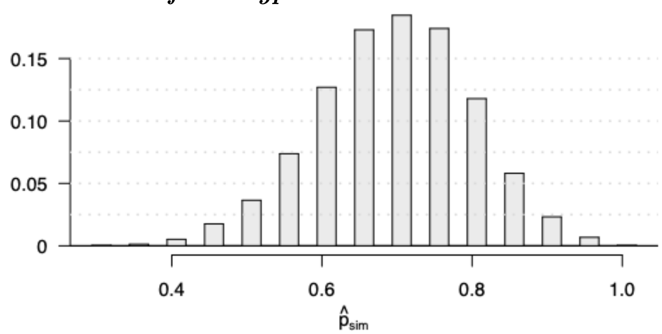
(a) State the null and alternative hypotheses for testing for independence of age and preferred shipping method for holiday gifts among Los Angeles residents. (b) Are the conditions for inference using a chi-square test satisfied?

The null hypothesis for testing the independence of age and preferred shipping carrier is that there is no association between these two variables among Los Angeles residents.
The alternative hypothesis is that there is an association between age and preferred shipping carrier.

To check the conditions for inference using a chi-square test, we need to ensure that the expected cell counts are all greater than or equal to 5. One of the expected cell counts is 3, therefore, that condition is not satisfied. So the conditions for inference using a chi-square test are **not** satisfied.

6.53 A popular uprising that started on January 25, 2011 in Egypt led to the 2011 Egyptian Revolution. Polls show that about 69% of American adults followed the news about the political crisis and demonstrations in Egypt closely during the first couple weeks following the start of the uprising. Among a random sample of 30 high school students, it was found that only 17 of them followed the news about Egypt closely during this time.

(a) Write the hypotheses for testing if the proportion of high school students who followed the news about Egypt is different than the proportion of American adults who did.
(b) Calculate the proportion of high schoolers in this sample who followed the news about Egypt closely during this time.
(c) Based on large sample theory, we modeled ^p using the normal distribution. Why should we be cautious about this approach for these data?
(d) The normal approximation will not be as reliable as a simulation, especially for a sample of this size. Describe how to perform such a simulation and, once you had results, how to estimate the p-value.
(e) Below is a histogram showing the distribution of ^psim in 10,000 simulations under the null hypothesis. Estimate the p-value using the plot and determine the conclusion of the hypothesis test.



- a

Null hypothesis: The proportion of high school students who followed the news about Egypt is the same as the proportion of American adults who did, p = 0.69.

Alternative hypothesis: The proportion of high school students who followed the news about Egypt is different from the proportion of American adults who did, p != 0.69.

- b

p = 17/30 = 0.567

- c

We should be cautious about using the normal distribution to model the proportion p for these data because the sample size is small (n = 30), and the normal approximation may not be accurate in such cases.

- d

1. Define the number of simulations to run.
2. For each simulation:

- Generate a random sample of 30 trials with a probability of success equal to 0.69.
- Calculate the proportion of successes in the simulated sample.

3. Calculate the proportion of simulations where the proportion of successes is as extreme or more extreme than our observed sample proportion of 17/30. This proportion represents the estimated p-value.

- e

We can visually estimate that there are only a few simulations where p_sim less than 0.57. Counting the bars that are less than or equal to 0.57 on the histogram, we can see that there are 5 bars that correspond to simulated proportions that are less than or equal to 0.57.

Therefore, the estimated p-value is approximately $5/10000 = 0.0005$. Since the p-value is less than 0.05, we reject the null hypothesis and conclude that there is evidence that the proportion of high school students who followed the news is different from the proportion of American adults who did.