

Lecture 3.3 - Practicing hypotheses tests

Student

2025-11-20



Hypothesis Testing

Setting Expectations

Included is a dataset regarding student drinking habits and other student characteristics from a 2008 survey of Portuguese high school students.

Some basic facts about Europe:

- According to a paper titled, “Youth Drinking Rates and Problems: A Comparison of European Countries and the United States” by Fries and Grube, the average problem drinking rate in Europe is about 19%.
- According to <https://www.statista.com/statistics/377585/household-internet-access-in-eu28/>, in 2008, the number of households in Europe with Internet access was about 60%
- According to <http://strongerfamilies.eu/about-us-2/one-parent-families-in-europe/>, about 10% of families in Europe are single-parent families
- According to <https://www.pewinternet.org/2015/10/01/basics-of-teen-romantic-relationships/>, about 19% of US students are in a romantic relationship

Setup and Data Exploration

1. What are the variables that map to these outcomes in your dataset?

- Problem drinking: `alcohol.use`
- Has internet at home: `internet.at.home`
- Single parent families: `parent.status`
- In a relationship: `has.romantic.partner`

For the following questions, make a small table:

2. What is the percentage of problem drinkers in the sample?

```
students <- students %>%  
  mutate(problem.drinking = case_when(alcohol.use < 4 ~ "SAFE",  
                                       alcohol.use >= 4 ~ "DANGEROUS"))
```

Think about what would happen if you reclassified this by defining 4 and above as problem drinking – make a note about this.

Obviously fewer students would qualify as a problem drinker. You need to think carefully about how the dataset values map to the hypothesis you want to test. It would be best to test both ways of classifying the data.

3. What is the percentage of students with Internet in their home in the sample?

4. What is the percentage of students who have single family homes in the sample?

Note that A is Apart and T is Together

5. What is the percentage of students who are in a romantic relationship in the sample?

Table 1: Distribution of key variables

| variable | value | n | proportion |
|----------------------|-----------------|-----|------------|
| Has romantic partner | no | 263 | 0.67 |
| Has romantic partner | yes | 132 | 0.33 |
| Internet at home | yes | 329 | 0.83 |
| Internet at home | no | 66 | 0.17 |
| Parent status | living.together | 354 | 0.90 |
| Parent status | living.apart | 41 | 0.10 |
| Problem drinker | SAFE | 316 | 0.80 |
| Problem drinker | DANGEROUS | 79 | 0.20 |

Planning

Pick one of the variables from questions 2-5. For this variable, consider the following questions:

- What, in your opinion, would be a substantively significant difference in this mean/proportion? How large would a difference need to be for you to consider it meaningful? Make some notes about this.

Let's say has romantic partner. I think maybe a 10% difference would seem meaningful.

- Are the conditions satisfied for conducting a hypothesis test? Why or why not?

Conditions:

- Randomization: Probably not met, but that has never stopped a study author before
- Independence: Depends on the sample design, but likely unless this is a snow-ball sample
- Success/failure: There are 263 students who no and 132 who yes. Condition met.

- What, in your opinion, should Portuguese policymakers do if you judge the the mean/proportion in your data are significantly different than world averages? Make some notes about this.

Hard to say. Maybe if there are too many students in a romantic relationship they should consider how much free time students have and how much they are focused on their studies.

Calculation

9. Write a fully specified hypothesis using a 95% cutoff.

Hypotheses

$$H_0 : p = 0.19 \quad H_A : p \neq 0.19$$

where p represents the true proportion of students who are in a romantic relationship. We are testing whether the proportion in our sample differs from the population value of 19%.

Significance Level

$$\alpha = 0.05$$

10. Calculate the p value for your hypothesis and indicate whether you reject or fail to reject the null hypothesis

$$\hat{p} = \frac{\text{yes count}}{n} = \frac{132}{395} \approx 0.334.$$

Under the null hypothesis $H_0 : p = 0.19$, the standard error is $SE_0 = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.19 \times 0.81}{395}} \approx 0.0197$.

$$\text{The test statistic is } z = \frac{\hat{p} - p_0}{SE_0} = \frac{0.334 - 0.19}{0.0197} \approx 7.304.$$

The two-sided p -value is $p\text{-value} = 2 \times P(Z > |z|) = 2 \times (1 - (|7.304|)) \approx 0$.

REJECT the null

Interpretation

11. Write a paragraph describing your results and also what kind of policy reforms you would recommend to policymakers based on the results of your hypotheses tests

It does appear many more students in Portugal are in a relationship, the difference with the population is both statistically and practically significant. I am not sure what kind of reforms would be necessary here, you would want to do some additional investigation as to whether the high rate of relationships is hurting the students' study habits.

Extra

Repeat the process for any of the other variables from questions 2-5 that you did not consider in the first round.