MITx: 14.310x Data Analysis for Social Scientists

Help

Bookmarks

Questions 1 - 11

Bookmark this page

The following problems are based on the paper:

> Duflo, Esther, Rema Hanna, and Stephen P. Ryan. 2012. "Incentives Work: Getting Teachers to Come to School." American Economic Review, 102(4): 1241-78.

First, read the abstract of the paper in the following link: http://economics.mit.edu/files/5582. You can refer back to the paper as necessary.

You will complete the following exercise for the variable **open**, the proportion of times the school was open during a random visit.

Note: The dataset used to generate the Lecture 15 slides relating to this paper is slightly different than the dataset we have provided, so do not be alarmed if your answers are slightly different!

In order to complete this exercise we are providing you with the code *problem1.R*. The code has some missing parts that you have to fill in order to run it. The dataset that you will need is teachers_final.csv

---

Question 1

1.0/1.0 point (graded)

First, consider the case where we have 8 schools. Our aim is to calculate the Fisher's exact p-value. Under the assumption that we will have the same number of treated and control units, how many potential treatment assignments across these 8 units are possible?

- ○ 50
- ○ 60
- ◉ 70 ✔
- ○ 80

Explanation

As it was discussed in the lecture there are 8 units and 4 of them are going to be assigned to the treatment group. So in this case we will have that the total number of potential treatment assignments is $\binom{8}{4}$ which is equal to 70.

Submit You have used 1 of 2 attempts

Suppose that after the treatment has been assigned and the experiment has been carried out, the researcher has the following data. The variable **open** corresponds to the fraction of days that the school was opened when random visits were made.

Use of Randomization and Nonparametric Regression
Finger Exercises due Nov 21, 2016 at 05:00 IST

Module 8: Homework
Homework due Nov 14, 2016 at 05:00 IST

Exit Survey

treatment	open
0	0.462
1	0.731
0	0.571
0	0.923
0	0.333
1	0.750
1	0.893
1	0.692

Question 2

1.0/1.0 point (graded)

Assume that we wish to calculate the absolute difference in means by treatment status as our statistic. For this observed data, what would be the value of this statistic? *Please give your answer to 5 decimal places, i.e. if your answer is 0.582931, please round to 0.58293*

0.19425

✔ **Answer:** 0.19425

0.19425

Explanation

We have that:

$$\hat{\tau} = \overline{\mathbf{Y}}_T^{obs} - \overline{\mathbf{Y}}_C^{obs} = \frac{Y_2^{obs}+Y_6^{obs}+Y_7^{obs}+Y_8^{obs}}{4} - \frac{Y_1^{obs}+Y_3^{obs}+Y_4^{obs}+Y_5^{obs}}{4} = 0.19425$$

Submit You have used 1 of 2 attempts

Now use the R code we have provided and fill in the missing information. The code calculates the value of this statistic (the difference in means for the treatment vs control group) for all the potential treatment assignments.

Question 3

1.0/1.0 point (graded)

How many of these statistics are larger than or equal to the one from our observed data?

- 11
- 16 ✔
- 21
- 26
- 31
- 36

Explanation

We can use a conditional function in R to test whether the statistic in other assignments exceeds the value in the observed data. In particular we have that by running the code:

larger_than_observed $<-$ (test_statistic >= observed_test)
sum(larger_than_observed)

There are 16 assignments in which this is the case.

Submit You have used 1 of 2 attempts

Question 4

1.0/1.0 point (graded)

What would be the Fisher's Exact p-value in this case?

Please round your answer to three decimal places, i.e. if your answer is 0.23418, please round to 0.234, and if it is 0.23498, please round to 0.235.

0.229 ✔ **Answer:** 0.229

0.229

Explanation

In this case we know that the p-value is given by $16/70$ that equals ≈ 0.229.

How should we interpret this? In general, we want to know whether the camera intervention had an effect, and whether the treated schools were open more frequently than the control schools. The mean of the treatment group is higher than the mean of the control group, indicating the teacher camera intervention may have indeed had an effect. However, under the sharp null hypothesis that there is no treatment effect in any of the schools in our sample, we have that if we randomly allocate 4 units to treatment, 23% of the time, the treatment and control groups would have looked at least as different as what we observed here, or even more different.

Submit You have used 1 of 2 attempts

Question 5

1.0/1.0 point (graded)

Now load the data set in R. Suppose we want to test the sharp null hypothesis in this data, with 49 treatment schools. Is it the case that the number of possible assignments would be too large to test this sharp null hypothesis (at least with your laptop and less than an hour of computing time)?

- a. Yes ✔
- b. No

Explanation

As discussed in lecture, it is pretty hard to compute all the potential statistics if you have a large number of observations. In this case we will have about $\binom{100}{49}$ possible random assignments. This is about $9.891308e+28$ different combinations.

Submit You have used 1 of 1 attempt

Question 6

1.0/1.0 point (graded)

A solution to this problem with a large number of observations is to simulate different random assignments and calculate the proportion of simulations in which the statistic exceeds the value of the observed data. We have provided you with the code that performs this exercise on the data **teachers_final.csv** with 100,000 simulations. If you run this code, is the approximate Fisher's p-value similar to the one we got with our 8 schools example?

- a. Yes
- b. No ✔

Explanation

No, in this case we have that by performing the 100,000 simulations the p-value that we obtain is very close to 0.

Submit You have used 1 of 1 attempt

Question 7

1.0/1.0 point (graded)

Since we are working in a very large sample, we can now consider Neyman's methods of inference. What is the Average Treatment Effect (ATE) on the observed data set?

Please round your answer to four decimals places, i.e. if it is 0.23451, please round to 0.2345

0.1969 ✔ **Answer:** 0.1969

0.1969

Explanation

In this case we have that the ATE is given by:

$$\overline{\mathbf{Y}}_T^{obs} - \overline{\mathbf{Y}}_C^{obs}$$

$$\overline{\mathbf{Y}}_T^{obs} - \overline{\mathbf{Y}}_C^{obs} = 0.1969$$

Submit You have used 1 of 2 attempts

Question 8

1/1 point (graded)

What is the upper bound of the standard error of this point estimate using Neyman's method?

Please round your answer to the fourth decimal place, i.e. if your answer is 0.43128, please round to 0.4313, and if it is 0.43122, please round to 0.4312.

0.0306

✔ **Answer:** 0.0306

0.0306

Explanation

We need the estimated standard error, and will use the conservative estimator of sampling variance $\hat{\mathbb{V}}_{\mathbf{neyman}}$,

$$\hat{\mathbb{V}}_{\mathbf{neyman}} = \frac{s_c^2}{N_c} + \frac{s_t^2}{N_t}$$

where

$$s_c^2 = \frac{1}{N_c - 1} \sum_{i:W_i=0} (Y_i^{obs} - \overline{Y}_c^{obs})^2$$

and

$$s_t^2 = \frac{1}{N_t - 1} \sum_{i:W_i=1} (Y_i^{obs} - \overline{Y}_t^{obs})^2$$

So,

$$\hat{\mathbb{V}}_{\mathbf{neyman}} = \frac{s_c^2}{N_c} + \frac{s_t^2}{N_t} = \mathbf{0.03055^2}$$

This operation is also performed with the R code we have provided.

Submit You have used 2 of 2 attempts

Correct (1/1 point)

Question 9

1/1 point (graded)

What is the t-statistic if we want to test the null hypothesis the ATE is equal to zero?

PLease round your answer to the second decimal point, i.e. if your answer is 4.567, please round to 4.57, and if it is 4.562, please round to 4.56.

6.45

Answer: 6.45

6.45

Explanation

In this case we will have that the t-statistic is:

$$t = \frac{\overline{\mathbf{Y}}_T^{obs} - \overline{\mathbf{Y}}_C^{obs}}{\sqrt{\hat{\mathbb{V}}_{\mathbf{neyman}}}} = \frac{0.1969}{0.03055} = 6.44$$

Submit You have used 1 of 2 attempts

✔ Correct (1/1 point)

Question 10

1/1 point (graded)

Is the associated p-value to this test similar to the one we found for the sharp null hypothesis in question 6?

- a. Yes ✔
- b. No

Explanation

The associated p-value to this test is $2 * (1 - \Phi(6.44)) \approx 0$, which is the same as the one we found in question 6.

Submit You have used 1 of 1 attempt

Correct (1/1 point)

Question 11

1/1 point (graded)

What is the 95% confidence interval of this test?

- a. It is given by $(0.127, 0.267)$
- b. It is given by $(0.147, 0.247)$
- c. It is given by $(0.157, 0.237)$
- d.It is given by $(0.137, 0.257)$

Explanation

The 95% CI is given by:

$(0.1969 - 1.96 * 0.03055, 0.1969 + 1.96 * 0.03055) = (0.137, 0.257)$

Submit You have used 1 of 2 attempts

✔ Correct (1/1 point)