Public Policy 529
Fall 2023: Problem Set #7

Due Monday, November 6, end of the day

1. Facing claims that city police were engaging in racial profiling, the city of Grand Rapids hired a consulting firm to perform a study on traffic stops in the city. The results of this study were released in April 2017, and the consulting firm's report is posted on Canvas in the Problem Sets folder. In short, the study found that Black motorists were stopped at "close to twice the rate that would be expected given their presence in the traffic."

   It is useful to examine the study's methodology. First, the consulting firm collected benchmark data on the race of drivers at particular intersections in the city. Thus, for each location, we have a sample with information on the percentage (i.e. proportion) of drivers that are Black. Second, the consulting firm collected data from the police department on the race of people who were stopped near those same locations, providing a second sample that measures the proportion of drivers that are Black. Under the null hypothesis of no racial profiling, the percentages in these independent samples are the same.

   (a) Earlier in the course, we talked about measurement. Examine how the consulting firm measured the benchmark data on the race of drivers (pp. 30-39 of the study). Assess the reliability and validity of this measurement strategy.

   (b) According to the data (p. 56), at the corner of Alpine & Leonard, 13.8% of the 3,042 drivers were Black. Out of 487 traffic stops made in that vicinity, 27.5% of the drivers were Black. Construct a 95% confidence interval for the difference of proportions. Be sure to use the correct standard error for a confidence interval.

   (c) Now perform a significance test (α = .01) in which the null hypothesis is that there is no difference between the proportion of drivers who are Black and the proportion of traffic stops that involve Black drivers. Perform all the steps and report all relevant statistics.

2. One important measure of development in a country is the rate of life expectancy. Suppose that, across a sample of democracies, the mean of this variable is 71.2 (n=77; s=10.2), and in a sample of non-democracies the mean is 64.8 (n=15; s=11.3).

   (a) Are these independent or dependent samples? Explain.

(b) Do you think we should make the assumption that life expectancy rates have the same variance in the populations of democracies and non-democracies? Explain.

(c) Suppose we cannot assume that these samples come from populations with equal variances, find the standard error of the difference for mean years of life expectancy and estimate degrees of freedom using the shortcut method.

(d) Using the standard error you just calculated, perform a significance test for the difference of means ($\alpha = .05$).

(e) Suppose instead that we can assume these samples come from populations with equal variances. What will be the standard error and degrees of freedom for the difference of means test?

(f) Perform a significance test for the difference of means under the assumption of equal variances ($\alpha = .05$).

3. In the `anes2020subset` dataset, the variable `BAplus` is a dichotomous variable in which 1 indicates the person has a BA degree or higher, and 0 indicates the person does not. In this question, you will test whether the mean of `PoliceTherm` is different for the two populations represented by these samples. `PoliceTherm` is a person's feeling thermometer score for the police.

   (a) Make a table that shows the means of `PoliceTherm` for each category of `BAplus`. In Stata, the command is `tab BAplus, summarize(PoliceTherm)`.

   In R, the most simple command is `aggregate(PoliceTherm ~ BAplus, data = anes2020, FUN = mean)`. This does not give you standard deviations or frequencies, however. So, I encourage you to go to page 10 of the R help document on Canvas and use the example in the middle of the page as the basis for your command.

   (b) Use software to perform a significance test for the difference of means. In Stata, the command in Stata is `ttest PoliceTherm, by(BAplus) unequal`.

   In R, the command is `t.test(anes2020$PoliceTherm ~ anes2020$BAplus)`

   (c) Report the difference between the means and state the conclusion of this significance test.

4. Continuing with the `anes2020subset` dataset, this question will have you test whether the proportion of people who have health insurance is different for people who have a BA degree versus those who do not have a BA degree.

   (a) Make a joint frequency distribution table with row and column totals for these two variables. Have `HealthIns` make the rows and `BAplus` make the columns. You have made this kind of table before.

   (b) Do this question by hand using the formulas. Using the data from the table, perform a significance test for whether the proportion of people that have health insurance is different across the categories of `BAplus`. Be sure to use the correct formula for the standard error.

(c) Now, use software to perform this test. In Stata, the command in Stata is `prtest HealthIns, by(BAplus)`.

In R, the process is more involved. Follow the steps outlined in Section 8.2 of the R Help document.

5. Use the dataset `LifeExpectancy` for this question. In this dataset, the variable `LifeExp2000` is the level of life expectancy for each country in the year 2000. The variable `LifeExp2010` is the level of life expectancy for the same set of countries in the year 2010. These are dependent samples.

(a) To measure the difference in these two life expectancy rates for each country, use the `generate` command to create a new variable called `LEdiff`.

In Stata, use the `generate` command:

```
gen LEdiff = LifeExp2010 - LifeExp2000
```

In R, the command is:

```
life$LEdiff <- life$LifeExp2010 - life$LifeExp2000
```

Browse your results with either `browse` in Stata or `View(life)` in R. When done, use commands to obtain the mean, standard deviation, and sample size for `LEdiff`.

Report your findings in your answers. How do you interpret the mean of `LEdiff`?

(b) We learned that the mean of the differences is the same as the difference of the means. Is that true? Use commands to find the means of `LifeExp2010` and `LifeExp2000`, then calculate the difference between them. Compare this to the mean of `LEdiff` that you found above.

(c) This insight from part (b) tells us that testing whether the mean of `LEdiff`=0 is the same as testing whether `LifeExp2010` and `LifeExp2000` have different means. By hand, perform the test of whether the mean of `LEdiff`=0. You have the information you need to calculate the standard error from your summary statistics in part (a). It's a one-sample test of statistical significance. Produce a *t* statistic and *p*-value for this test.

(d) Now use your software to perform a dependent samples (i.e. paired) *t*-test for the difference of means. Report the results and compare them to the test that you performed in part (c).

In Stata, the command is: `ttest LifeExp2010=LifeExp2000`

In R, the command is: `t.test(life$LifeExp2000, life$LifeExp2010, paired = TRUE)`