SOC 5050: Problem Set 06

Christopher Prener, Ph.D.

October 24th, 2016

Directions

Please complete all steps below. Your final do-file, log-file, and markdown file with answers should be uploaded to your GitHub assignment repository by 4:20pm on Monday, October 31st, 2016.

The data given here for calculation by hand are drawn from a teaching data-set of fictional panel data. They are structured as a typical panel design with a variable (`pre`) measuring initial, baseline scores and a variable (`post`) measuring follow-up scores.

Part 1: One-Sample T-test

```
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     pre |     100         9.9    .5301991    5.301991     8.84797    10.95203
------------------------------------------------------------------------------
```

1. Using the above data from the fictional panel data-set, test (by hand) how the baseline variable (`pre`) compares with a prior study that found that the baseline score in this population was 12. Be sure to provide a full interpretation of this test.
2. Using the NHIS dataset, clean the variable `LCDURA4` so that the "unknown units" category is coded as missing. Test whether this variable compares to a previous study that found the average length of asthma diagnosis to be 8 years. Be sure to provide a full interpretation of this test.
3. Create a well laid-out plot showing the mean and 95% confidence interval for the cleaned version of `LCDURA4`.

Part 2: Independent Samples T-test

```
Variance ratio test
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    male |      48    9.395833    .6904488    4.783569     8.00683    10.78484
  female |      52    10.36538    .7968778    5.746367    8.765586    11.96518
---------+--------------------------------------------------------------------
combined |     100         9.9    .5301991    5.301991     8.84797    10.95203
------------------------------------------------------------------------------
    ratio = sd(male) / sd(female)                                 f =   0.6930
Ho: ratio = 1                                 degrees of freedom =   47, 51

    Ha: ratio < 1               Ha: ratio != 1                 Ha: ratio > 1
  Pr(F < f) = 0.1028         2*Pr(F < f) = 0.2057           Pr(F > f) = 0.8972

// ============================================================

Difference of Means Data
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    male |      48    9.395833    .6904488    4.783569     8.00683    10.78484
  female |      52    10.36538    .7968778    5.746367    8.765586    11.96518
---------+--------------------------------------------------------------------
combined |     100         9.9    .5301991    5.301991     8.84797    10.95203
---------+--------------------------------------------------------------------
```

4. Using the above data from the fictional panel data-set, test (by hand) how the baseline variable (`pre`) differs between male and female participants. Does the baseline variable vary by gender? Be sure to adjust the test for heterogeneous variance values if appropriate. Also be sure to provide a full interpretation of this test.

5. Using the NHIS dataset, recode the variable `MRACBPI2` so that "0" represents white respondents and "1" represents black respondents. Test whether African Americans (who are more likely to suffer from asthma than whites) have longer average length of asthma diagnoses than whites in this study. Be sure to adjust the test for heterogeneous variance values if appropriate. Also be sure to provide a full interpretation of this test.

6. Calculate and interpret the appropriate effect size for the result in the previous question.

7. Create a well laid-out plot showing the means and 95% confidence intervals for both white and black respondents in terms of the length of asthma diagnoses.

Part 3: Dependent Samples T-test

```
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     pre |     100         9.9    .5301991    5.301991     8.84797    10.95203
    post |     100        8.05    .4482007    4.482007    7.160673    8.939327
---------+--------------------------------------------------------------------
    diff |     100        1.85    .6613996    6.613996    .5376396     3.16236
------------------------------------------------------------------------------
```

8. Using the above data from the fictional panel data-set, test (by hand) how the baseline variable (`pre`) differs from the follow-up scores. Does the (unknown) intervention have an impact on these participants? Be sure to provide a full interpretation of this test.

9. Using the dataset `bpwide.dta`, which comes pre-installed with Stata[1], test whether mean blood pressure scores differ between the baseline and follow-up time periods. Be sure to provide a full interpretation of this test.

10. Create a well laid-out plot showing the means and 95% confidence intervals for average blood pressure socres at both the pre and post time periods.

[1] *Hint:* Use the `sysuse` command!

Part 4: Power Analysis for Difference of Means

11. You are trying to design a study that tests for differences in in a battery of questions designed to assess a student's likelihood of dropping out of college. The scale generated by these questions is a single variable that contains scores from 0 to 50. Assume that this scale is roughly normally distributed with mean scores probably lying somewhere between 20 and 30 (you have no pilot data to base your power analysis on). In order to detect a small difference in mean scores between students who are first generation college students and students with at least one parent who attended college, what size sample would you need?

12. In order to detect a large difference in mean scores, what size sample would you need?
13. In order to detect a moderate difference in means scores, what size sample would you need?

Grading Rubric

Part 1 This section is worth six points.

Part 2 This section is worth seven points.

Part 3 This section is worth six points.

Part 4 This section is worth three points.

Stata Do-File The overall quality of the Stata do-file stack is worth eight points. This grade will be based on the clarity, organization, and layout of your do-files.

Document Details

Document produced by Christopher Prener, Ph.D. for the Saint Louis University course SOC 5050 - QUANTITATIVE ANALYSIS: APPLIED INFERENTIAL STATISTICS. See the course wiki and the repository `README.md` file for additional details.