Applied Microeconometrics II

Assignment 1 Solutions

1. Currently 75% of high school students score pass a high School Exit Exam. (This rate has been calculated using all students in the state, approximately 20,000 students.) A new online tutoring program is being proposed, to be taken the summer before the final year of high school. The claim is that program will increase the exam pass rate to 78%.

   You are being asked to design an experimental evaluation to test this claim. You will be forming a smaller "treatment" group who will be given the online program, and all other students will remain in the "control" group.

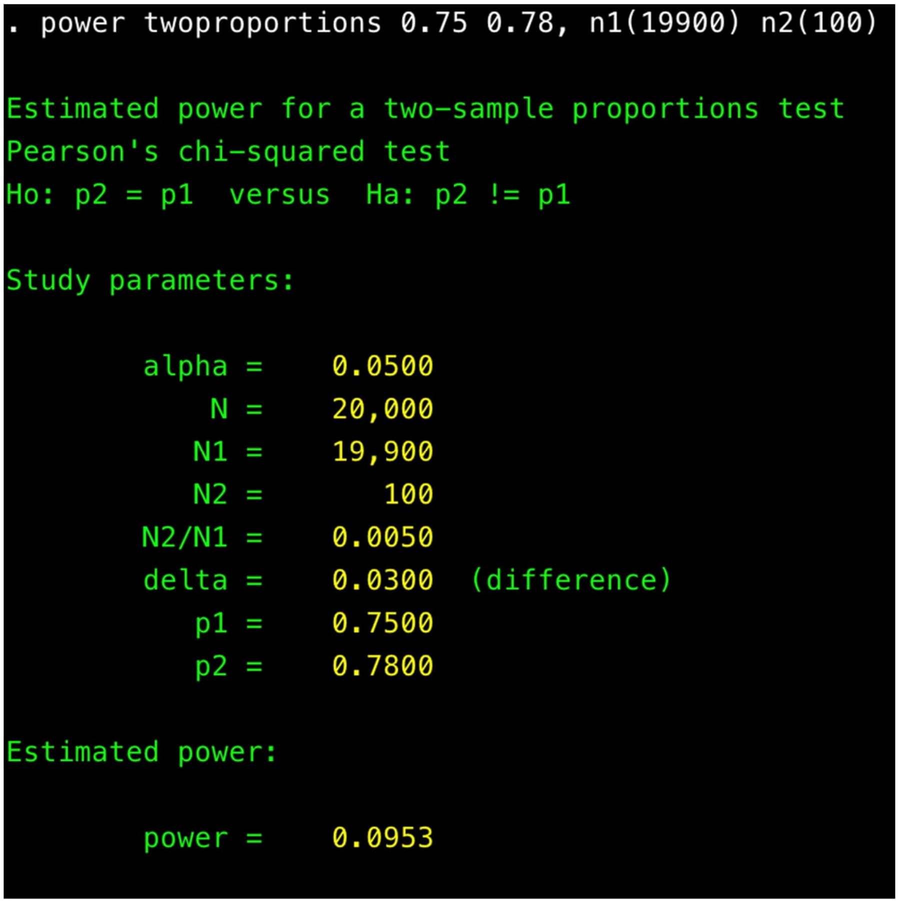
a. Suppose you allow the students to volunteer for the program. What would be the threat to internal validity? Argue what would be the likely direction of bias (positive/negative) for an estimate of the treatment effect? Suppose you ran a regression

$$Y_i = \alpha + ß\, T_i + e_i$$

Would the estimate of the treatment effect, ß, be over or underestimated?

Individuals who volunteer for the program likely differ from those who do not: they may be students with lower test scores who are using the treatment to improve their performance, or they may be overachievers, who can't pass on the opportunity to receive more training. In any case, self-selection into treatment would plausibly bias our estimation of the treatment effect, affecting internal validity. If we believe motivation is the key omitted variable, and it is positively correlated with the outcome, as well as with receiving the treatment, the bias will be positive, meaning the coefficient will be overestimated.

b.What is the probability of Type 2 Error (β) if your treatment sample size is 100? What is the power of your test? You can use the Stata **power** command, or perform approximate calculations manually, following lecture notes.

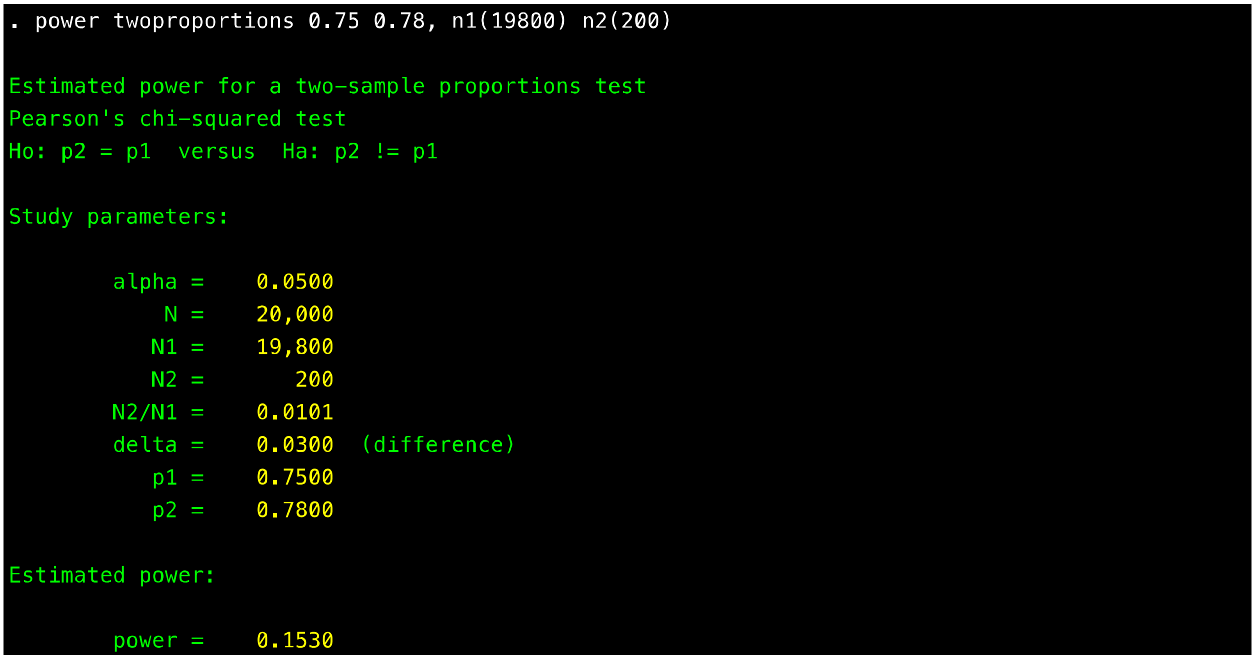


```
. power twoproportions 0.75 0.78, n1(19900) n2(100)

Estimated power for a two-sample proportions test
Pearson's chi-squared test
Ho: p2 = p1  versus  Ha: p2 != p1

Study parameters:

        alpha =    0.0500
            N =    20,000
           N1 =    19,900
           N2 =       100
        N2/N1 =    0.0050
        delta =    0.0300  (difference)
           p1 =    0.7500
           p2 =    0.7800

Estimated power:

        power =    0.0953
```

c. Would you say your design in part c is *underpowered*? Explain how having an underpowered experiment poses risks to statistical conclusion validity. What type of errors is your analysis more exposed to?

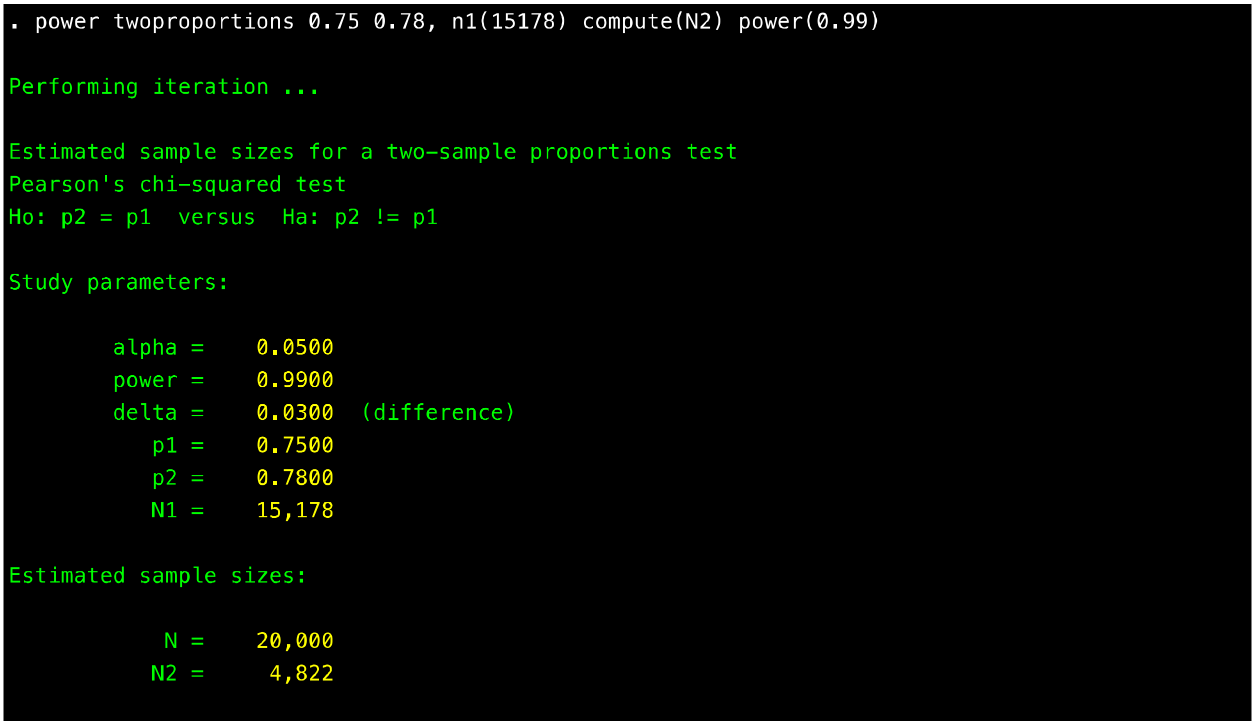
If a test is insufficiently powered, we have two few observations to generally detect an effect, so we will get false negatives (Type II errors). This is not necessarily because there isn't an effect, but simply because we don't have enough observations to detect one.

d. What is the probability of Type 2 Error (β) if your treatment sample size is 200? Now what is the Power of your test? You can use the Stata power command or perform approximate calculations manually, following lecture notes.



```
. power twoproportions 0.75 0.78, n1(19800) n2(200)

Estimated power for a two-sample proportions test
Pearson's chi-squared test
Ho: p2 = p1  versus  Ha: p2 != p1

Study parameters:

        alpha =    0.0500
            N =    20,000
           N1 =    19,800
           N2 =       200
        N2/N1 =    0.0101
        delta =    0.0300  (difference)
           p1 =    0.7500
           p2 =    0.7800

Estimated power:

        power =    0.1530
```

e) How large a sample do you need to insure that β = 0.01? Note that you can proceed by trial and error, starting with a control group of 16,000 observations, using the Stata power command.

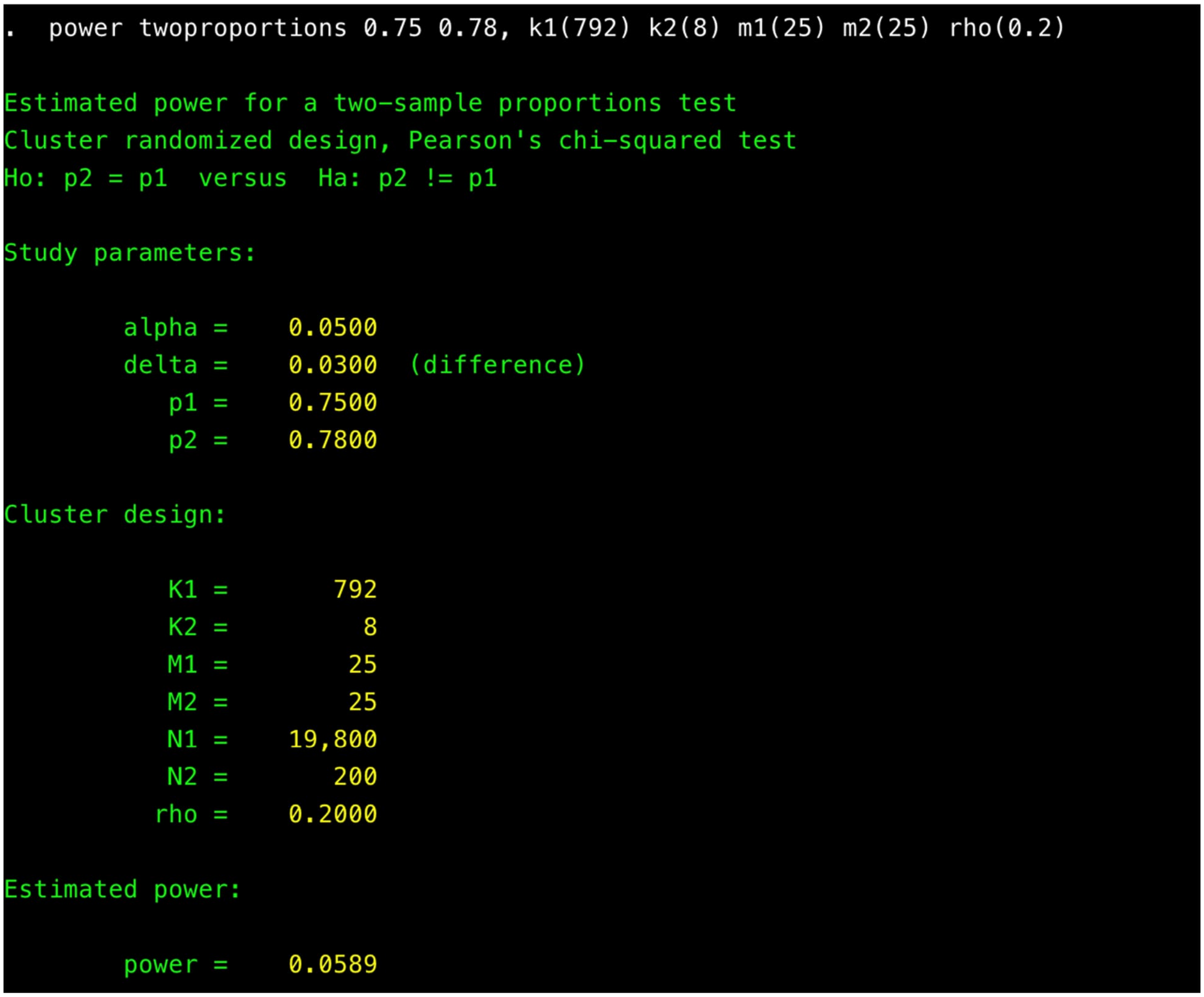


```
. power twoproportions 0.75 0.78, n1(15178) compute(N2) power(0.99)

Performing iteration ...

Estimated sample sizes for a two-sample proportions test
Pearson's chi-squared test
Ho: p2 = p1  versus  Ha: p2 != p1

Study parameters:

        alpha =    0.0500
        power =    0.9900
        delta =    0.0300  (difference)
           p1 =    0.7500
           p2 =    0.7800
           N1 =    15,178

Estimated sample sizes:

            N =    20,000
           N2 =     4,822
```

f) Suppose that instead of an online program, where students study independently at home, the program is implemented in classrooms. An evaluation finds a regression coefficient on the treatment effect of beta=0.02, t=2.6. These results however ignored intraclass correlation. Assuming positive intraclass correlation, how would the results above change?

The standard errors are likely too small, meaning the t coefficient is too large. The coefficient would stay the same, but the t stat might decrease to levels where we would no longer reject the null.

g) Suppose that all students are in classrooms of exactly 25 students. Your treatment group has 8 classrooms of 25 students, and all control students are also in classrooms of exactly 25 students. What is the probability of Type 2 Error (β)? How does the power of your test compare with the power you found in part d? Use the Stata **power** command.



```
.  power twoproportions 0.75 0.78, k1(792) k2(8) m1(25) m2(25) rho(0.2)

Estimated power for a two-sample proportions test
Cluster randomized design, Pearson's chi-squared test
Ho: p2 = p1  versus  Ha: p2 != p1

Study parameters:

        alpha =    0.0500
        delta =    0.0300  (difference)
           p1 =    0.7500
           p2 =    0.7800

Cluster design:

           K1 =       792
           K2 =         8
           M1 =        25
           M2 =        25
           N1 =    19,800
           N2 =       200
          rho =    0.2000

Estimated power:

        power =    0.0589
```

Question 2.
As the sample sizes feature in the denominator in the z score, and power increases with the z score, as shown in lecture 1, the problem of allocating a total sample between treatment and control is equivalent to maximizing a function such as:

$$\frac{1}{\frac{1}{x}+\frac{1}{n-x}}$$

, where n is the total number of observations and x is the size of the treatment sample. After taking first order conditions, you will find the solution to be x=n/2.

Question 3.

Begin by expressing the variance of $\bar{S}_j$ (the mean of S in cluster j) as the sum of the two variance components (individual and cluster-specific):

$$\sigma^2_{\bar{S}_j} = \sigma^2_\gamma + \frac{\sigma^2_\epsilon}{n} \tag{1}$$

Using the formula for ICC, you can express $\sigma^2_\epsilon = \sigma^2_S(1-\rho)$ and $\sigma^2_\gamma = \rho\sigma^2_S$ and plug them back into (1)

After some short algebraic manipulation, you can write (1) as:

$$\sigma^2_{\bar{S}_j} = \frac{\sigma^2_S}{n}(1+(n-1)\rho) \tag{2}$$

The variance of the overall mean $\bar{S}_{ij}$ can be expressed as

$$\begin{aligned} \sigma^2_{\bar{S}} &= Var\left(\frac{\sum_{j=1}^m \sum_{i=1}^n S_{ij}}{mn}\right) = Var\left(\frac{\sum_{j=1}^m n\frac{\sum_{i=1}^n S_{ij}}{n}}{mn}\right) \\ &= Var\left(\frac{\sum_{j=1}^m n\bar{S}_j}{mn}\right) = Var\left(\frac{\sum_{j=1}^m \bar{S}_j}{m}\right) = \frac{1}{m^2} Var\left(\sum_{j=1}^m \bar{S}_j\right) \end{aligned} \tag{3}$$

Noting that ϵ_{ij} and γ_j are not correlated across clusters and using (2), we finally have

$$\begin{aligned} \sigma^2_{\bar{S}} &= \frac{1}{m^2}\sum_{j=1}^m Var(\bar{S}_j) = \frac{1}{m^2}\sum_{j=1}^m \frac{\sigma^2_S}{n}(1+(n-1)\rho) \\ &= \frac{1}{m^2} m \frac{\sigma^2_S}{n}(1+(n-1)\rho) = \frac{\sigma^2_S}{mn}(1+(n-1)\rho) \end{aligned} \tag{4}$$

Note: To simplify calculations, all clusters are assumed to be of the same size: n_j=n.

4. Read the selections posted on the course website from "Experimental and Quasi-experimental Designs for Generalized Causal Inference", William R. Shadish, Thomas D. Cook, Donald Thomas Campbell . Answer the following questions:

a) What qualitative methods can help detect compensatory rivalry (John Henry) effects?

(Unstructured) interviews, direct observation.

b) Describe the problem of confounding constructs with levels of constructs. Why is it a problem for assessing the effects of treatment in experiments? How can this problem be addressed?

Making inferences about general constructs but failing to acknowledge that only limited levels of the construct were actually studied. For example, studies frequently use monetary incentives that are very small to infer the effect of monetary incentives on all sorts of outcomes. Large amounts of monetary incentives however may produce choking under pressure effects, but researchers fail to capture those effects if they pay their participants oh say 10 dollars. One way to address the problem is to use several levels of treatment.

c) What are floor and ceiling effects in the measurement of an independent variable? How do they affect statistical power?

Floor effects: all or most of the values cluster near the lowest possible value; ceiling effects: all values cluster close to the highest value. Floor and ceiling effects restrict the range of the variable, leading to lower power, and attenuating bivariate relationships.

d) How do meta analyses help answer concerns about of external validity? What are the main drawbacks of meta analyses?

Chapter 3, page 86: meta analyses show results for multiple studies, covering different sets of individuals, settings, outcomes and treatments- in this sense, they improve external validity and contextualize results. Metanalyses cover many papers, but often do not, and cannot control for the quality of the estimates in the original papers, particularly if the original data used in those papers is not available. Because of this problem, metanalyses often present a wide range of estimates, but end-of-the-range estimates are often just the consequence of poor quality research.

e) What is the Bonferroni correction? Why do we use it- what is the threat to statistical conclusion validity we are trying to address?

Research conclusions can be misleading if multiple tests are performed, because the Type I error rate for multiple tests is actually much larger than for a single test. The Bonferroni correction simply divides the target Type I error rate, which is typically 0.05, by the number of tests in the set, and then uses the Bonferroni-corrected α in individual tests. The resulting smaller Type I error rate is more conservative.