Problem Set 10

11/14/2019

This problem set is due at 8:05 am on 11/21

Please upload both the Rmd file and the PDF file on Sakai

Total points: 10

You have been hired by the Vietnam Chamber of Commerce and Industry (VCCI) to evaluate a training program on corporate governance and internal accounts management. The designers of the training course believe that businesses which have higher quality corporate governance and strong procedures for managing internal purchases will have enjoy greater productivity and business growth. Your job is to construct the research design for the experiment.

1. Load the dataset vietnamese_firms.dta. This data is the 2016 Enterprise Survey, annual survey of businesses in Vietnam conducted by the General Statistical Office. At the top of the file, set a seed (`set.seed(2019)`). This command is important as it will ensure that the randomization sequence is the same for all students. (1 point)
2. What type of experimental design (i.e. encouragement, over-subscription) do you believe is most appropriate for this study? Explain why? (1 point)
3. Using the data provided perform a power calculation in R or using Power Calculator(https://egap.shinyapps.io/Power_Calculator/). The managers of the training program believe that firms exposed to the management training program will experience a 10% increase in profitability above the control group. They estimate a standard deviation of 870 in profitability. Profitability is coded as *profit* in the dataset. Assuming individual level randomization, how many firms do you need to sample to have sufficient power to identify a statistically significant effect (.05 level) if one exists? (2 points)
4. For logistical reasons, VCCI has suggested performing the training at the provincial level. There are 63 provinces in Vietnam. Remove province 1 (`province == 1`). Then, we would like to randomly sample 31 provinces to be in the control group and 31 provinces for the treatment group. Using the **clus.rho** command, calculate the intracluster correlation in profitability before the experiment. Perform a power calculation that takes into account this clustered design suggestion. (2 points)

```
set.seed(1)
firm <- read_dta("vietnamese_firms.dta") %>%
      filter(province != "01")

prov_treat =  data.frame(
province = unique(firm$province),
rd_id = simple_ra(62,.5)
)
head(prov_treat)
```

```
##   province rd_id
## 1       02     0
## 2       04     0
## 3       06     1
## 4       08     1
## 5       10     0
## 6       11     1
```

```
sample = firm %>%  group_by(ownership) %>%
  sample_frac(.01)
```

```
head(sample)
```

```
## # A tibble: 6 x 4
## # Groups:   ownership [2]
##   province ownership  profit    id
##   <chr>        <dbl>   <dbl> <dbl>
## 1 62               1   -8528 61155
## 2 46               1    1468 49289
## 3 75               1   86227 71451
## 4 31               2      85 34134
## 5 56               2    1930 55102
## 6 79               2  114583 86376
```

5. Using the number of observations from your power calculation in question 3, draw a random sample of firms for the experiment, blocking on firm ownership (*ownership*). Use R's `sample_n()`, `sample_frac()` functions. (1 point)

6. Again, blocking on *ownership*, randomly assign an equal number of firms to the control and treatment group. You can use `block_ra()` (1 point)

```
# package for random assignment

set.seed(1)

z = block_ra(sample$ownership )

table(z, sample$ownership)
```

```
##
## z      1    2    3    4    5    6    7    8    9   10   11   12   13
##   0    1    3    5    1   51  232    1 1159  226    5   42    1    5
##   1    2    3    6    1   52  232    1 1159  226    5   43    1    5
```

```
sample$treat <- z

sample %>% arrange(ownership ,id, treat)
```

```
## # A tibble: 3,468 x 5
## # Groups:   ownership [13]
##    province ownership  profit    id treat
##    <chr>        <dbl>   <dbl> <dbl> <int>
##  1 46               1    1468 49289     1
##  2 62               1   -8528 61155     1
##  3 75               1   86227 71451     0
##  4 15               2     389 20674     1
##  5 31               2      85 34134     0
##  6 40               2     170 43586     0
##  7 56               2    1930 55102     1
##  8 79               2  114583 86376     1
##  9 79               2    5506 89003     0
## 10 40               3    4009 44448     0
## # ... with 3,458 more rows
```

```
head(sample)
```

```
## # A tibble: 6 x 5
```

```
## # Groups:   ownership [2]
##   province ownership  profit    id treat
##   <chr>        <dbl>   <dbl> <dbl> <int>
## 1 62               1   -8528 61155     1
## 2 46               1    1468 49289     1
## 3 75               1   86227 71451     0
## 4 31               2      85 34134     0
## 5 56               2    1930 55102     1
## 6 79               2  114583 86376     1
```

7. What sorts of limitations (i.e. LATE, attrition) do you expect for this experiment? What can be done to mitigate the effects of those problems? (2 points)