University of British Columbia

Sample Survey Project

Author:
Group leader-Yujie Chen(12391397)- Data Collection and Summaries, R code
Tianyi Wen(52924453)- Data Collection, Discussion and Conclusion
Tiffany Yang(57619181) Data Collection, Data Analysis
Mingrui Zhang(55850762) - Data Collection, Introduction, R code
Leah Zhao(22126593)- Data Collection

Supervisor:
Dr. Lang Wu

An Assignment submitted for the UBC

STAT 344

November 14, 2020

1 Introduction

We have observed that as courses have been moved online, the grading system and examination methods for many courses have changed. This can easily lead to changes in the amount of time students spend on study each day, and may make students experience additional anxiety from online courses compared to live ones. Our objective is to study whether students adapt to new online teaching scheme well, whether they make some adjustments on their study, and then whether the school should provide more support for students to help them reduce their anxiety and better complete their studies.

To that end, we estimate the true proportion of all current undergraduates at UBC Vancouver who prefer online courses to in-person courses. We first estimate this parameter based on a simple random sample. Then we estimate the true proportion for a second time using stratified sampling, stratifying the whole population into two strata: Male and Female. For each of two types of samples described above, we also estimate the true average difference in study time after and before the epidemic. This gives us another way to analyze the attitudes and reactions of students to the online learning scheme. Moreover, using two sampling methods allows us to better verify our analysis of true average study time difference and true proportion of preferences for online courses and draw a better conclusion for those.

To collect our data, we designed an online questionnaire open to all full-time undergraduates at UBC Vancouver and then we organized the data received into an Excel spreadsheet for further data analysis in R.

2 Data Collection and Data Summaries

Our targeted population is all full-time undergraduates at UBC Vancouver. As mentioned above,we designed an online questionnaire open to all full-time undergraduates at UBC Vancouver to collect our sample data. In the questionnaire, we got data for the Gender, Academic Year, Faculty, Study Time Before online, Study Time After online, Preference for online or in-person courses and Anxiety about online courses. We finally chose to focus on (1) Preference for online courses and (2) Difference in study time per day after and before courses move online.

Something to note about the data we collected: (1) Preference for online or in-person courses is binary, which can either be "online" or "in-person" and (2) Difference in study time per day after and before courses move online is continuous, which can take a value from 3 to 10 hours/day. We thought this range was reasonable, based on workload at UBC. In the following data analysis, the two parameters of interest are (1) population proportion of all full-time undergraduates at UBC Vancouver who prefer online courses to in-person courses and (2) population average study time difference after and before courses move online.

From this online questionnaire, we got 129 responses, consisting of 53 males and 76 females, which means that we had a SRS of size 129. Since we thought that there might be some difference in study habits between two genders which can affect their study time and preference for teaching methods, we decided to take Male and Female as two strata when performing the analysis of stratified sampling. We got the demographic enrollment information about all current undergraduates at UBC Vancouver from http://pair.ubc.ca/enrolment/enrolment-status/, from which we summarized that there are 31909 full-time undergraduates which consist of 13787 males (about 42 percent) and 18122 females (about 58 percent). Under the assumption that the within-stratum variance is the same for both strata (male and female) and the per-unit sampling costs for both strata (male and female) are the same, then we thought of proportional allocation as the optimal strategy. This means we need a sample, 42 percent of which are males and 58 percent of which are females. We found that our collected sample consisted of 53 males (about 41 percent) and 76 females (about 59 percent). It showed that our collected data satisfied our expectation for sample components.

We added the data we received into an Excel file which finally consisted of 129 observations and 7 variables. The specific data collected was in the attached file.

3 Data Analysis

We performed all the analysis in R. The results are summarized in the following table:

1st population: whether the student prefers online courses or not

Sampling Methods	Proportion	Standard Error	Confidence Interval
SRS	0.3100775	0.04079911	[0.2301113, 0.390043]
Stratified Sample	0.3070642	0.04028504	[0.2281055, 0.3860229]

2nd population: the change in study time before and after courses moved online

Sampling Methods	Mean	Standard Error	Confidence Interval
SRS	0.02325581	0.1688959	[-0.3077801, 0.3542917]
Stratified Sample	0.01290408	0.1683694	[-0.3170999, 0.3429080]

Our study has two parameter of interests, one is proportion of all students who prefer online courses, a binary random variable, and the other one is the average study time difference after and before courses moved online, a continuous random variable. It was assumed that all the students participating in this survey were randomly selected, and all their responses are independent to each other. It was also assumed that the population is normally distributed, and thus the central limit theorem was used to calculate the Confidence Intervals.

1st population: whether the student prefers online courses or not

For our simple random sampling, our sample size was 129, and we initially thought that it would be large enough for us to be confident in our results. Since population size was known, we decided to include the finite population correction in our calculation. By importing the Excel data set into R, we had our sample ready for analysis

- the code could be found in the appendix under the "Simple Random Sampling - R code". We calculated the Vanilla estimate of population proportion that preferred online courses by using the sample proportion. The value we obtained was 0.3100775, the standard error for this estimate was 0.04079911, and the 95% Confidence Interval is $[0.2301113, 0.390043]$.
The vanilla estimate is calculated by

$$\widehat{p}_{srs} = \frac{\text{number of students who prefer online in the sample}}{n} = \frac{40}{129} = 0.3100775$$

The standard error is calculated by

$$s.e(\widehat{p}_{srs}) = \sqrt{(1-\frac{n}{N})\frac{s^2}{n}} = \sqrt{(1-\frac{129}{31909})(\frac{0.2156008^2}{129})} = 0.04079911$$

The 95% CI is calculated by

$$\widehat{p}_{srs} \pm z_{0.975} \times s.e(\widehat{p}_{srs})$$
$$0.3100775 \pm 1.96 \times 0.04079911$$

$$[0.2301113, 0.390043]$$

In the Stratified Sample,

The value we obtained was 0.3070642, the standard error for this estimate was 0.04028504, and the 95% Confidence Interval is $[0.2281055, 0.3860229]$.
The vanilla estimate is calculated by

$$\begin{aligned}\widehat{p}_{str} =& \frac{\text{number of male students who prefer online in the sample}}{n} \times \frac{\text{number of male students}}{\text{N}} \\ &+ \frac{\text{number of female students who prefer online in the sample}}{n} \times \frac{numberofmalestudents}{N} \\ =&0.2264151 \times \frac{13787}{31909} + 0.3684211 \times \frac{18122}{31909} \\ =&0.3070642\end{aligned}$$

The standard error is calculated by

$$s.e(\widehat{p}_{str}) = \sqrt{\sum(\frac{N_h}{N})^2(1-\frac{n_h}{N_h})\frac{s_h^2}{n_h}} = 0.04028504$$

The 95% CI is calculated by

$$\widehat{p}_{str} \pm z_{0.975} \times s.e(\widehat{p}_{str})$$
$$0.3070642 \pm 1.96 \times 0.1683694$$

$$[0.2281055, 0.3860229]$$

2nd population: the change in study time before and after courses moved online
In the SRS, we obtained the Vanilla estimate of population average of study time difference by using the sample average. The value we obtained was 0.02325581, the standard error for this estimate was 0.1688959, and the 95% Confidence Interval is $[-0.3077801, 0.3542917]$.

The vanilla estimate is calculated by

$$\bar{y}_{srs} = \text{mean (study time after-study time before)} = 0.02325581$$

The standard error is calculated by

$$s.e(\bar{y}_{srs}) = \sqrt{(1-\frac{n}{N})\frac{s^2}{n}} = 0.1688959$$

The 95% CI is calculated by

$$\bar{y}_{srs} \pm z_{0.975} \times s.e(\bar{y}_{srs})$$
$$0.023255815 \pm 1.96 \times 0.1688959$$

$$[-0.3077801, 0.3542917]$$

In the stratified sample, the value we obtained was 0.01290408, the standard error for this estimate was 0.1683694, and the 95% Confidence Interval is $[-0.3170999, 0.3429080]$.

The vanilla estimate is calculated by

$$\bar{y}_{str} = \frac{n_{male}}{N_{male}} \times \bar{y}_{h,male} + \frac{n_{female}}{N_{female}} \times \bar{y}_{h,female} = 0.01290408$$

The standard error is calculated by

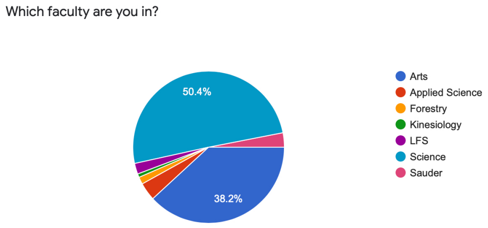
$$s.e(\bar{y}_{str}) = \sqrt{\sum(\frac{N_h}{N})^2(1-\frac{n_h}{N_h})\frac{s_h^2}{n_h}} = 0.1683694$$

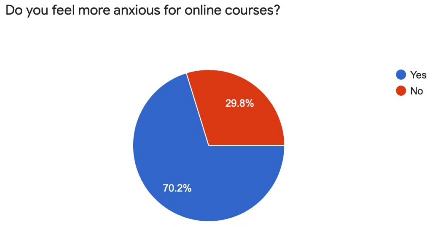
The 95% CI is calculated by

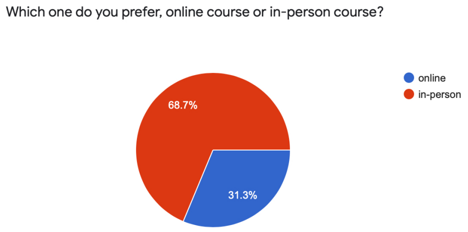
$$\widehat{y}_{str} \pm z_{0.975} \times s.e(\widehat{y}_{str})$$
$$0.01290408 \pm 1.96 \times 0.1683694$$

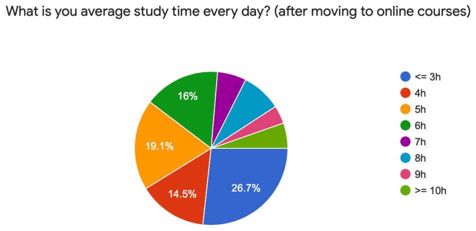
$$[-0.31709990.3429080]$$

The advantage of a stratified sample is that our standard error is usually less than that that computed from a simple random sample. This is because SRS considers the overall variance, including between-stratum variances.

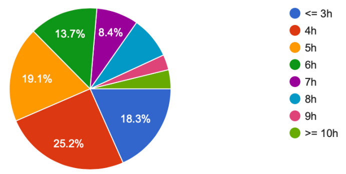




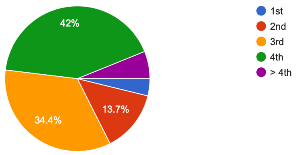




What is you average study time every day? (before moving to online courses)



Which year are you in? (Undergraduate)



4 Discussions and Conclusions

For our final result, we chose our stratified sample estimates, since our stratified sample estimates had a relatively smaller standard error than the simple random sample(SRS), but the difference is not that large.

Our final estimate of mean of the study time difference after and before online courses is 0.01290408 hours, with standard error 0.1683694 hours. We are 95% confident that the true mean of the study time difference after and before online courses is between -0.3170999 hours and 0.3429080 hours.

Our final estimate of proportion of the people that prefer online courses is 0.3070642, with standard error 0.04028504. We are 95% confident that the true proportion of the people that prefer online courses is between 0.2281055 and 0.3860229.

We assume that the within-stratum variance is the same for both strata (male and female) and the per-unit sampling costs for both strata (male and female) are the same, which is reasonable. Our sample size is 129, which is large enough for our result to be valid, but is still small for there to be limitations. Meanwhile, the existence of response bias and neighborhood bias, and we use CLT, which assume a normal distribution of our sample, these could also cause our result to be inaccurate.

Based on our result, there is only a small proportion of students who prefer online courses to in-class courses. But the study time doesn't changed that much before and after the online course, which indicate that students made some effective change to their study, and could adapt to online course well. In this case, UBC might consider giving student more online materials to study, and giving students more access to online libraries, and try their best to try to make courses in-class.

However, we cannot guarantee our results could be well generalized to the other population such as all UBC Vancouver graduate students or all full-time undergraduates in other universities, since the workload for online courses might vary a lot in different student bodies or universities. Further studies will have to be done to investigate the overall situation of all Canadian universities.

5 Appendix

Reference for population size and distribution:

http://pair.ubc.ca/enrolment/enrolment-status/

Code used:

Simple Random Sampling - R code

```
> library(readxl)
> X344_Project_Sample <- read_excel("C:/Users/huawei/Desktop/344 Project Sample.xlsx")
> View(X344_Project_Sample)
> t1=(X344_Project_Sample$‘Study Time_Before‘)
> t2=(X344_Project_Sample$‘Study time_After‘)
> change.time = t1-t2
> mean.timechange=mean(change.time)
> sd.timechange=sd(change.time)
> mean.timechange
[1] -0.02325581
> sd.timechange
[1] 1.922178
> se.timechange=sqrt((((sd.timechange)^2)/129)*(1-129/31909))
> se.timechange
[1] 0.1688959
> CI.time=c(mean.timechange - 1.96*se.timechange,mean.timechange + 1.96*se.timechange)
> CI.time
[1] -0.3542917  0.3077801
> samplesize.online=sum(X344_Project_Sample$Preference=="online")
> n=129
> p_hat.online=samplesize.online/n
> p_hat.online
[1] 0.3100775
> N=31909
> se.p_hat.online=sqrt((1-n/N)*p_hat.online*(1-p_hat.online)/n)
> se.p_hat.online
[1] 0.04064066
> CI.p_hat.online=c(p_hat.online - 1.96*se.p_hat.online,p_hat.online + 1.96*se.p_hat.online)
> CI.p_hat.online
[1] 0.2304218 0.3897332
> n1 =sum(X344_Project_Sample$Gender=="Male")
> n2 =sum(X344_Project_Sample$Gender=="Female")
> N1=13787
> N2=18122
> n1
[1] 53
> n2
[1] 76
> sample.male=X344_Project_Sample[X344_Project_Sample$Gender == "Male",][,1:7]
> change.time_male=sample.male$‘Study Time_Before‘ - sample.male$‘Study time_After‘
> sample_mean.male=mean(change.time_male)
> sample_mean.male
[1] 0.2641509
```

```
> sample.female=X344_Project_Sample[X344_Project_Sample$Gender == "Female",][,1:7]
> change.time_female=sample.female$‘Study Time_Before‘ - sample.female$‘Study time_After‘
> sample_mean.female = mean(change.time_female)
> sample_mean.female
[1] -0.2236842
> y_bar.str= (N1/N)*sample_mean.male+(N2/N)*sample_mean.female
> y_bar.str
[1] -0.01290408
> sd_timechange.male=sd(change.time_male)
> sd_timechange.female=sd(change.time_female)
> sd_timechange.male
[1] 1.913022
> sd_timechange.female
[1] 1.915541
> se.ybar_str= sqrt((N1/N)^2*(1-n1/N1)*(sd_timechange.male)^2/n1+(N2/N)^2*(1-n2/N2)*(sd_timechange.female)^2/n2)
> se.ybar_str
[1] 0.1683694
> CI.ybar_str=c(y_bar.str - se.ybar_str, y_bar.str + se.ybar_str)
> CI.ybar_str
[1] -0.1812734  0.1554653
> # Preference online online by gender of str.
> sample.male_online=sum(sample.male$Preference=="online")
> n1.online=sample.male_online
> n1.online
[1] 12
> sample.female_online=sum(sample.female$Preference=="online")
> n2.online=sample.female_online
> n2.online
[1] 28
> ybar_male.online=n1.online/n2
> ybar_female.online=n2.online/n1
> ybar_male.online
[1] 0.1578947
> ybar_female.online
[1] 0.5283019
> se.phat_online_str=sqrt((N1/N)^2*(1-n1/N1)*(1-ybar_male.online)*ybar_male.online/n1 + (N2/N)^2*(1-n2/N2)*(1-ybar_female.online)*ybar_female.online/n2)
> se.phat_online_str
[1] 0.03898343
> phat.online_str=(N1/N)*(n1.online/n1)+(N2/N)*(n2.online/n2)
> phat.online_str
[1] 0.3070642
```

```
> attach(STAT_344_Project_Sample)
> N_male <-13787
> N_female <- 18122
> N <- N_male + N_female
> # SRS
> n <- length(Gender)
> # 1st population: whether the student prefers online courses or not
> # Vanilla estimate: p_hat.srs
> prefer_online_or_not.sample <- rep(NA,length(Gender))
> for (i in 1:n){
```

```
+    if (STAT_344_Project_Sample[i,]$Preference=="online") {
+      prefer_online_or_not.sample[i] <- 1
+    } else {
+      prefer_online_or_not.sample[i] <- 0
+    }
+ }
> p_hat.srs <-mean(prefer_online_or_not.sample)
> p_hat.srs
[1] 0.3100775
> sample.variance <- var(prefer_online_or_not.sample)
> sample.variance
[1] 0.2156008
> # Since the population size N and sample size n are known, we
    choose to include the FPC
> se.srs <- sqrt((1-n/N)*sample.variance/n)
> se.srs
[1] 0.04079911
> # We would like to construct the 95% CI for population proportion
     of prefering online
> CI.srs <- c(p_hat.srs-1.96*se.srs,p_hat.srs+1.96*se.srs)
> CI.srs
[1] 0.2301113 0.3900438
> # 2nd population: the change in study time before and after
    courses move online
> #                 (after - before)
> # population parameter: mean change in study time >> mean(after-
    before)
> # Vanilla estimate: y_bar.srs
> study_time_change.sample <- STAT_344_Project_Sample$‘Study time_
    After‘-
+   STAT_344_Project_Sample$‘Study Time_Before‘
> y_bar.srs <- mean(study_time_change.sample)
> y_bar.srs
[1] 0.02325581
> sample2.variance <- var(study_time_change.sample)
> # Since the population size N and sample size n are known, we
    choose to include the FPC
> se.srs_ybar <- sqrt((1-n/N)*sample2.variance/n)
> se.srs_ybar
[1] 0.1688959
> # We would like to construct 95% CI for population mean change in
     study time
> CI2.srs <- c(y_bar.srs-1.96*se.srs_ybar,y_bar.srs+1.96*se.srs_
    ybar)
> CI2.srs
[1] -0.3077801  0.3542917
```

Stratified Sampling - R code

```
> # Stratified Sampling
> # We use the proportional allocation here
> # We know that N_male = 13787 and N_female = 18122 based on the
    enrollment data from
> # the school website.
> # Thus, we have a sample: n_male/N_male is approximately equal to
     n_female/N_female
> male.sample <-STAT_344_Project_Sample[Gender=="Male",]
> n_male <- nrow(male.sample)
> female.sample <- STAT_344_Project_Sample[Gender=="Female",]
> n_female <- nrow(female.sample)
> # 1st population: whether the student prefers online courses or not
> # Here are two strata: Male students and Female students
> # Vanilla estimate: p_hat.str
> male_prefer_online.sample <- rep(NA,n_male)
> for (i in 1:n_male) {
+   if(male.sample$Preference[i]=="online") {
+     male_prefer_online.sample[i] = 1;
+   } else {
+     male_prefer_online.sample[i] = 0;
+   }
+ }
> female_prefer_online.sample <- rep(NA,n_female)
> for (i in 1:n_female) {
+   if(female.sample$Preference[i]=="online") {
+     female_prefer_online.sample[i] = 1;
+   } else {
+     female_prefer_online.sample[i] = 0;
+   }
+ }
> p_hat.strata <- c(mean(male_prefer_online.sample),mean(female_prefer_online.sample))
> p_hat.strata
[1] 0.2264151 0.3684211
> # Since the population size N and sample size n are known, we choose to include the FPC
> se.strata <- c(sqrt((1-n_male/N_male)*var(male_prefer_online.sample)/n_male),
+                   sqrt((1-n_female/N_female)*var(female_prefer_online.sample)/n_female))
> se.strata
[1] 0.05792535 0.05558311
> wt.strata <- c(N_male/N,N_female/N)
> p_hat.str <- sum(wt.strata*p_hat.strata)
> p_hat.str
[1] 0.3070642
> se.p_hat.str <- sqrt(sum(wt.strata^2*se.strata^2))
> se.p_hat.str
[1] 0.04028504
> # We would like to construct 95% CI for population proportion of students who
> # prefer online courses
> CI.str <- c(p_hat.str-1.96*se.p_hat.str,p_hat.str+1.96*se.p_hat.str)
> CI.str
[1] 0.2281055 0.3860229
> # 2nd population: the change in study time before and after courses move online
> #                  (after - before)
> # population parameter: mean change in study time >> mean(after-before)
> # Vanilla estimate: y_bar.str
> # Here are two strata: Male students and Female students
> male_study_time_change.sample <- male.sample$‘Study time_After‘-
+                                  male.sample$‘Study Time_Before‘
```

```
> female_study_time_change.sample <- female.sample$‘Study time_After‘-
+   female.sample$‘Study Time_Before‘
> # Male and Female
> y_bar.strata <- c(mean(male_study_time_change.sample),
+                   mean(female_study_time_change.sample))
> y_bar.strata
[1] -0.2641509  0.2236842
> y_bar.str <- sum(wt.strata*y_bar.strata)
> y_bar.str
[1] 0.01290408
> # Since the population size N and sample size n are known, we choose to include the FPC
> se.y_bar.strata <- c(sqrt((1-n_male/N_male)*var(male_study_time_change.sample)/n_male),
+                      sqrt((1-n_female/N_female)*var(female_study_time_change.sample)/n_female))
> se.y_bar.strata
[1] 0.2622682 0.2192664
> se.y_bar.str <- sqrt(sum(wt.strata^2*se.y_bar.strata^2))
> se.y_bar.str
[1] 0.1683694
> # We would like to construct 95% CI for population mean change in study time
> CI2.str <- c(y_bar.str-1.96*se.y_bar.str,y_bar.str+1.96*se.y_bar.str)
> CI2.str
[1] -0.3170999  0.3429080
```