NOVEMBER 2019

STATS 344: SAMPLE SURVEY PROJECT REPORT

Report by:

1)Kewei Li (82433020): Team Leader, researcher, programmer

2)Qing Cong (54058243): researcher, programmer, coordinator

3)Isabella Zhong (21988167): researcher, analyst, author

Sample Survey: An Investigation of UBC Book Collections

Introduction

As a group of undergraduate statistics students, our team is interested in implementing our skills in daily life. We decided to estimate how many pages we are expected to read when we get a book from the UBC libraries, where we usually spend time studying. By investigating in this topic, we aim to help students to develop a feasible study plan in consideration of the best way to allocate their time between reading the books and doing the works.

Background

Our study has two parameters of interest, one of them is the average number of pages each book has, a continuous random variable; and the other one is the proportion of books that were published in the last ten years, a binary random variable. The target population in our research is all the books from all the libraries on the UBC Vancouver campus, including: Asian Library, David Lam Management Research Library, Education Library, Music, Art and Architecture Library, Koerner Library, Law Library, and Woodward Library. To increase the readability of our research and the R code, we labeled them from 1 to 7. Moreover, to obtain the most representative sample, we made use of two sampling methods, simple random sampling and two-stage cluster sampling and all the samples are drawn from the population.

To perform the analysis, there are some assumptions we made and they are listed below:

1) All the samples are drawn randomly and our selection of books shows no personal preference over any specific types of books.
2) The number of pages each book has in each library is independent.
3) The year in which the book is published does not have an effect on which library the book will be stored in the UBC library.
4) The population is large enough, so we don't need a finite population corrector in our calculation.
5) Each cluster (library) is distinct and subpopulations (books) are different within each cluster. This means the seven libraries are mutually exclusive from one another and there are no overlapping books in each library.
6) Each cluster (library) has the same amount of book collections. (An important assumption in Standard Error computation for two-stage cluster method)

Data Collection

We obtained the SRS sample by randomly sampling every library, and for two-stage cluster sampling, we randomly selected the clusters by R, which resulted in our clusters being libraries 1,3,4,5,6, and 7. To determine a reasonable sample size, we run a pre-study program in R, which we set seed to 666 to create reproducible random sampling results. For the whole study, we set the desired accuracy as ±1100 percentage points, and z score as 1.28, which is 80% confidence interval. Following that, we used the formula n=(z^2*variance)/accuracy to calculate the number of samples we should collect, 110 books. Thus, we obtained a total of 220 samples which half of them are drawn by SRS and the other half by two-stage cluster sampling. The data are displayed in the appendix.

Data Analysis

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

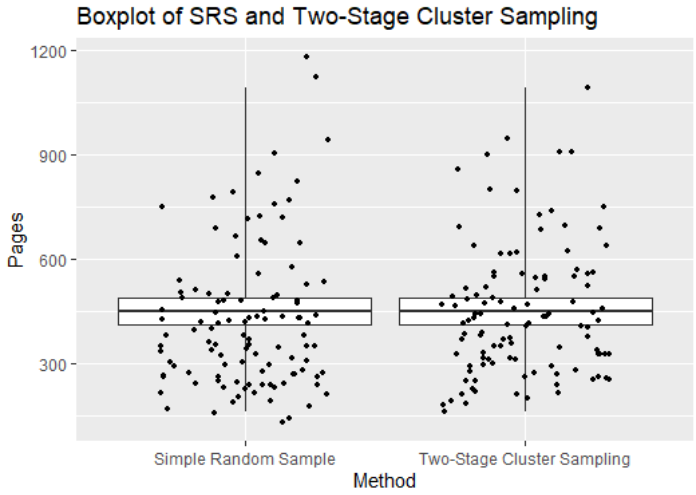
Method\Statistical Summary for Continuous Random Variable	Mean	Standard Error	Confidence Interval
SRS Sample	428.3818	19.76626	(403.0810, 453.6826)
Two-stage Cluster Sample	438.3144	12.63611	(422.1402, 454.4886)

Method\Statistical Summary for Binary Random Variable	Proportion	Standard Error	Confidence Interval
SRS Sample	0.3455	0.0453	(0.2875, 0.4035)
Two-stage Cluster Sample	0.1818	0.0367	(-0.2879, 0.6515)

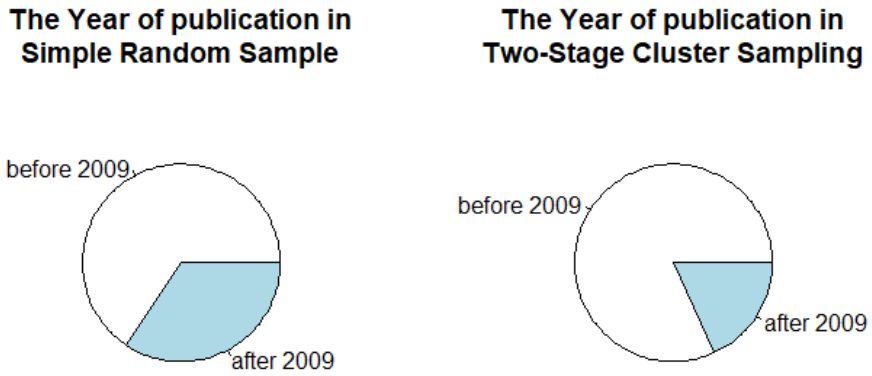
Comparing to the SRS sampling method, the means of the continuous random variable calculated for the cluster sampling method differ by 2.318%, and standard errors differ by -36.072%, which is a significant difference. Nevertheless, it is expected that the SRS method would perform better in terms of minimizing the variations within the sample because its sampling process is simpler. For the binary random variable, the means calculated for the two methods differ by -47.38% and standard error differs by --18.98%. The estimated proportion from SRS almost doubled the proportion from Two-stage cluster, so it turns out that the difference between the two methods is quite large in this case.

For the continuous random variable, we interpret the resulting confidence interval as we are 80% confident that the true mean number of pages of each book found in UBC

libraries falls between ((403.0810, 453.6826) by SRS sampling method, or, the true mean lies between (422.1402,454.4886) by two-stage sampling method. The comparison of the values from the two methods is illustrated below by boxplot. It clearly demonstrates that although there are a few outliers in each sample, the values on the 25, 50, 75th percentiles are quite close.



For the binary random variable, we interpret the resulting confidence interval as we are 80% confident that the true proportion of books published in last 10 years in UBC library falls between (0.2875, 0.4035) by SRS sampling method, or, the true proportion lies between (-0.2879, 0.6515) by two-stage sampling method. To demonstrate the result graphically, we use a pie chart for the categorical variable that whether the book is published in the last ten years.



Calculation

Explanation of Computation for Variance of Two-stage Cluster Sample (Continuous)

$$Var(\hat{\bar{y}}) \approx \frac{1}{\bar{M}^2}\left(1-\frac{n}{N}\right)\frac{s_r^2}{n}$$

$$s_r^2 = \frac{1}{n-1}\sum_{i\in S}(M_i\bar{y}_i - M_i\hat{\bar{y}}_r)^2$$

The variance formula for Two-stage Cluster Sampling on the left-hand side was obtained from the reference textbook: **Sampling: Design and Analysis, by Sharon L. Lohr** page 186.

Notations are explained as follow:

- $\hat{\bar{y}}$: The estimate of average page numbers in a book.
- $\bar{M}$: The average book volume in (i.e. number of books) each library.
- M_i : Book volume in (i.e. number of books) in a specific library.
- N : Total number of library.
- n : Number of library in this sample.
- s_r^2 : Sample variance within each library.

We perform the following transformation on the Formula.

$$Var(\hat{\bar{y}}) \approx \frac{1}{\bar{M}^2}\left(1-\frac{n}{N}\right)\frac{\frac{1}{n-1}\sum_{i\in S}(M_i\bar{y}_i - M_i\hat{\bar{y}}_r)^2}{n}$$

$$Var(\hat{\bar{y}}) \approx \frac{1}{\bar{M}^2}\left(1-\frac{n}{N}\right)\frac{1}{n(n-1)}M_i^2\sum_{i\in S}(\bar{y}_i - \hat{\bar{y}}_r)^2$$

Given our Assumption 6 (assume all library have the same amount of book collections), here we have: $M_i = \bar{M}$

Now we can simplify as:

$$Var(\hat{\bar{y}}) \approx \frac{N-n}{N}\frac{1}{n(n-1)}\sum_{i\in S}(\bar{y}_i - \hat{\bar{y}}_r)^2$$

$$Var(\hat{\bar{y}}) \approx \frac{7-6}{7}\frac{1}{6(6-1)}\sum_{i\in S}(\bar{y}_i - \hat{\bar{y}}_r)^2$$

$$Var(\hat{\bar{y}}) \approx \frac{1}{7}\frac{1}{30}\sum_{i\in S}(\bar{y}_i - \hat{\bar{y}}_r)^2$$

Hence we code the formula as follows: (See Appendix – R Code (code line 78))

```
sqrt((1/210)*sum((mean.ssu-mean(mean.ssu))^2))
[1] 12.63611
```

Discussion

Obviously, the above graphs have depicted that the SRS method and two-stage cluster sampling method can result in completely different values, so we will talk about the advantages and disadvantages of each method in this section.

For the SRS method, the first advantage we noticed is that the sampling process is easier because all we did were randomly drawing the book from every library, while for two-stage cluster sampling, we need to randomly select the clusters to be sampled on the first stage. Furthermore, since we assumed that the stock in each library is unique, the sample we got by SRS might be more representative than the other method which only samples some of the libraries. However, SRS is more time-consuming (i.e. we need to visit all libraries in this study to obtain the SRS sample), which is the main drawback of our experience since we spend more time sampling all the libraries.

For the two-stage cluster sampling method, the main advantages are that this method is more practical since it required fewer resources (i.e. only need to visit some of the libraries) and took less time for us to complete the sampling process. Most of the time, the SRS method can result in more accurate estimates because the standard error, the measure of the preciseness of an estimate, is usually lower than that of the cluster sampling method.

Overall, the concern lies in the phenomenon that the house effect might be involved in both methods. It can arise as the systematic bias inherent to the team member who was collecting the samples. The member might have preferences over some books or that one might choose the books that are within his or her approach and easily accessible. Therefore, sampling bias cannot be ignored in our study.

Conclusion

This section will focus on discussing the limitations and broader implications of our study. Admittedly, the estimates for the binary random variable differ significantly, and the standard errors of both random variables obtained from SRS are larger than that from two-stage cluster sampling. The reasons why these unusual results happened in our study might be subject to the following four reasons.

First, the sampling process may not be completely random. As mentioned above, although we tried to make sure each book was drawn with the same probability, we cannot eliminate all the biases in our sample. Secondly, our confidence interval used in this survey is 80%, which is relatively narrower comparing to the commonly used 95% confidence interval. As such, the probability of us capturing the true value of random variables might be lower. The third possible reason might be the sample size, since the one we used here is 220 books in total, which may not be large enough to draw a statistically significant conclusion. Besides, despite all the three reasons discussed above,

the result of having a slightly lower standard error for the cluster sample could still happen just by random chance, even no biases or design failure was involved.

Notwithstanding the limitations of our research, students can make use of our estimate of the expected number of pages each book has in the UBC library calculated based on the two-stage cluster sampling method, which is about 439 pages. It can be considered as a reliable result because the standard error is reasonably small, suggesting that the estimate is quite accurate.

In summary, if we were able to sample more books from the libraries, it's more likely that we will capture the true value of the two random variables. Even so, our estimates can be applied as a sensible inference for students who would like to make study plans based on how much time they would spend on reading books.

Reference

Lohr, S.L. (1999 1st ed., 2010 2nd ed.) Sampling: Design and Analysis. Duxbury Press. (p.186)

Appendix

Appendix I: R Code

```
> # STAT344 PROJECT
> # map index number to Library locations (on UBCV campus)
> # 1 - Asian Library
> # 2 - David Lam Management Research Library (Sauder)
> # 3 - Education Library
> # 4 - Music, Art and Architecture Library (IKB)
> # 5 - Koerner Library
> # 6 - Law Library
> # 7 - Woodward
>
> set.seed(666)
>
> index<-c(1,2,3,4,5,6,7)
> set.seed(666)
> # pre sample: two stage cluster
> PSU.pre<-sample(index, 3, replace = FALSE)
> PSU.pre
[1] 6 2 5
> SSU.pre<-sample(PSU.pre, 15, replace = TRUE)
> SSU.pre
 [1] 6 2 5 5 2 6 6 5 6 6 6 6 5 6 5
> table(SSU.pre)
SSU.pre
2 5 6
2 5 8
>
> csv.pre<-read.csv("projectprestudy.csv")
> mean(csv.pre$pages)
[1] 503.2667
> # subset different clusters
> sub.pre.2<-subset(csv.pre, csv.pre$library.index == 2)
> sub.pre.5<-subset(csv.pre, csv.pre$library.index == 5)
> sub.pre.6<-subset(csv.pre, csv.pre$library.index == 6)
> # get mean of pages in clusters
> mean.pre.2<-mean(sub.pre.2$pages)
> mean.pre.5<-mean(sub.pre.5$pages)
> mean.pre.6<-mean(sub.pre.6$pages)
> mean.pre.ssu <- c(mean.pre.2, mean.pre.5, mean.pre.6)
> mean.pre <- mean(csv.pre$pages)
> mean.pre
[1] 503.2667
> var.pre <- (1-3/7)*(sum((mean.pre.ssu-mean.pre)^2)/2)
> var.pre
[1] 9567.289
```

```
> #(1-6/7)*sum((mean.ssu-mean(mean.ssu))^2)/5
>
> # pre study ends, calculate sample size.
> accuracy_desired<-11.95
> z_desired<-1.28     #for 80% C.I.
> n_sample<-(z_desired^2*var.pre)/(accuracy_desired^2)
> n_sample<-round(n_sample)
> n_sample
[1] 110
>
> # two stage cluster study
> PSU<-sample(index, 6, replace = FALSE)
> PSU
[1] 4 3 5 6 1 7
> SSU<-sample(PSU, n_sample, replace = TRUE)
> table(SSU)
SSU
 1  3  4  5  6  7
19 19 18 10 22 22
> # SRS
> SRS <- sample(index, n_sample, replace = TRUE)
> table(SRS)
SRS
 1  2  3  4  5  6  7
18 19 18 10 21 15  9
>
> #import data
> csv.srs <- read.csv("project-srs.csv")
> csv.2.cluster <- read.csv("project-2-cluster.csv")
>
> # split clusters
> cluster.1<-subset(csv.2.cluster, csv.2.cluster$library.index == 1)
> cluster.3<-subset(csv.2.cluster, csv.2.cluster$library.index == 3)
> cluster.4<-subset(csv.2.cluster, csv.2.cluster$library.index == 4)
> cluster.5<-subset(csv.2.cluster, csv.2.cluster$library.index == 5)
> cluster.6<-subset(csv.2.cluster, csv.2.cluster$library.index == 6)
> cluster.7<-subset(csv.2.cluster, csv.2.cluster$library.index == 7)
>
> mean.ssu.1<-mean(cluster.1$pages)
> mean.ssu.3<-mean(cluster.3$pages)
> mean.ssu.4<-mean(cluster.4$pages)
> mean.ssu.5<-mean(cluster.5$pages)
> mean.ssu.6<-mean(cluster.6$pages)
> mean.ssu.7<-mean(cluster.7$pages)
> mean.ssu <- c(mean.ssu.1,mean.ssu.3,mean.ssu.4,mean.ssu.5,mean.ssu.6,mean.ssu.7)
>
> #report mean and se for continuous r.v. (2-stage cluster)
> mean.2.cluster <- mean(mean.ssu)
> mean.2.cluster
[1] 438.3144
> var.2.cluster <- (1/210)*sum((mean.ssu-mean(mean.ssu))^2)
> se.2.cluster <- sqrt(var.2.cluster)
> se.2.cluster
[1] 12.63611
>
>
> # check sample size
> n_sample
[1] 110
> nrow(csv.srs) == n_sample
[1] TRUE
> nrow(csv.2.cluster) == n_sample
[1] TRUE
>
> #report mean and se for continuous r.v. (srs)
> mean.srs <- mean(csv.srs$pages)
> mean.srs
[1] 428.3818
> var.srs <- var(csv.srs$pages)/nrow(csv.srs)
> se.srs<-sqrt(var.srs)
> se.srs
[1] 19.76626
>
> # report mean and se for discrete r.v.
> #pre-study
> year.pre <- table(csv.pre$public.year <=2009)
> year.pre
FALSE  TRUE
    3    12
> books.after.2009.pre <- as.data.frame(year.pre)[1,2]
> prop.pre <- books.after.2009.pre/15
> prop.pre
[1] 0.8
> se.prop.pre <- sqrt(prop.pre*(1-prop.pre)/n_sample)
> se.prop.pre
[1] 0.103279556
```

```
>
> #srs
> year.srs <- table(csv.srs$public.year <=2009)
> year.srs

FALSE  TRUE
   38    72
> books.after.2009.srs <- as.data.frame(year.srs)[1,2]
> prop.srs <- books.after.2009.srs/n_sample
> prop.srs
[1] 0.3454545
> se.prop.srs <- sqrt(prop.srs*(1-prop.srs)/n_sample)
> se.prop.srs
[1] 0.0453387
>
> #2-stage cluster
> year.2.cluster <- table(csv.2.cluster$public.year <=2009)
> year.2.cluster
FALSE  TRUE
   20    90
> books.after.2009.2.cluster <- as.data.frame(year.2.cluster)[1,2]
> prop.2.cluster <- books.after.2009.2.cluster/n_sample
> prop.2.cluster
[1] 0.1818182
> se.prop.2.cluster <- sqrt(prop.2.cluster*(1-prop.2.cluster)/n_sample)
> se.prop.2.cluster
[1] 0.03677454
```

Appendix II:

Pre-study			760	2010	7	226	2010	5
pages	public year	library index	272	2018	7	348	2016	5
290	2004	2	341	2012	7	439	1995	5
577	1996	2	437	2009	6	351	2007	5
435	2019	5	349	2019	6	447	1999	5
409	2001	5	266	2013	6	296	1998	5
490	1987	5	793	1961	6	479	2017	5
204	2007	5	237	2015	6	263	1982	5
303	2001	5	903	2008	6	212	2004	5
1096	2002	6	943	1988	6	432	2003	5
904	2008	6	477	2018	6	261	2005	5
388	2017	6	1125	2018	6	379	1989	5
318	2009	6	472	2018	6	362	2017	5
433	2008	6	500	2000	6	646	2013	5
601	2017	6	322	2009	6	449	1992	4
530	2007	6	556	2016	6	268	1971	4
571	2002	6	539	2016	6	498	1970	4
Simple Random Sample			1183	2018	6	279	2009	4
pages	public year	library index	298	2006	5	349	2017	4
294	2001	7	456	1999	5	414	1993	4
428	2008	7	423	1991	5	192	1960	4
847	2014	7	414	2015	5	205	1953	4
339	2001	7	435	2007	5	380	1972	4
316	2008	7	268	1981	5	215	2010	4
241	2014	7	770	2010	5	261	2012	3

159	1960	3	578	2003	1	444	2009	4
419	2002	3	752	1984	1	550	1986	4
715	1976	3	176	2012	1	184	2002	4
247	2088	3	306	2019	1	269	1928	4
336	1992	3	431	2016	1	726	1980	4
232	2003	3	431	2016	1	425	1970	4
825	1985	3	230	2002	1	415	1956	4
665	2010	3	646	2009	1	251	1975	4
189	1972	3	368	2017	1	260	1951	4
504	2002	3	237	1991	1	486	2010	4
142	1994	3	528	2018	1	901	2008	4
216	1984	3	399	2016	1	337	1991	5
327	1987	3	607	2007	1	239	2005	5
170	1997	3	*Two-Stage Cluster Sampling*			390	2011	5
482	1983	3	pages	public year	library index	358	2017	5
502	1967	3	477	2008	3	201	2009	5
382	2005	3	273	2002	3	292	1987	5
534	2018	2	368	1976	3	181	1988	5
278	2002	2	218	1997	3	312	1978	5
304	2001	2	552	1983	3	384	1905	5
480	1985	2	215	1974	3	328	2017	5
352	2004	2	406	2009	3	739	2009	6
490	2016	2	512	1986	3	459	2013	6
242	2014	2	493	2001	3	299	2010	6
418	2001	2	250	2003	3	328	1960	6
490	2000	2	228	1978	3	563	1891	6
272	2013	2	433	1996	3	447	2006	6
129	1999	2	191	2008	3	750	2005	6
352	2004	2	352	2006	3	469	2019	6
512	2003	2	317	1997	3	516	2013	6
689	1988	2	256	2017	3	441	2005	6
654	1994	2	345	1997	3	638	2015	6
779	1996	2	328	1982	3	315	2001	6
237	1988	2	1093	2002	3	261	2001	6
248	2003	2	432	1970	4	379	2005	6
302	2001	2	552	1968	4	252	2007	6
237	1983	1	472	1945	4	412	2012	6
724	1982	1	614	1963	4	465	2007	6
395	1980	1	640	1931	4	624	2005	6
429	2002	1	684	2008	4	468	2013	6
719	2011	1	330	1947	4	313	2008	6

416	2017	6	326	1987	7	367	2008	1
378	2009	6	477	2005	7	281	1992	1
457	2006	7	425	2004	7	212	2016	1
254	1997	7	434	2008	7	908	2000	1
948	2005	7	858	1976	7	802	2008	1
212	2012	7	371	2006	7	691	1992	1
490	2000	7	349	2015	7	444	2004	1
291	1994	7	523	1993	7	162	2008	1
402	1990	7	407	2008	7	561	1976	1
558	2003	7	693	1999	1	545	2004	1
329	2006	7	541	2002	1	616	2001	1
498	2003	7	278	1992	1	295	1983	1
518	1990	7	559	1999	1	619	2010	1
698	1984	7	570	2012	1			
796	2013	7	910	2015	1			