University of British Columbia

STAT344 Project Proposal

Yufei Cai # 25616533

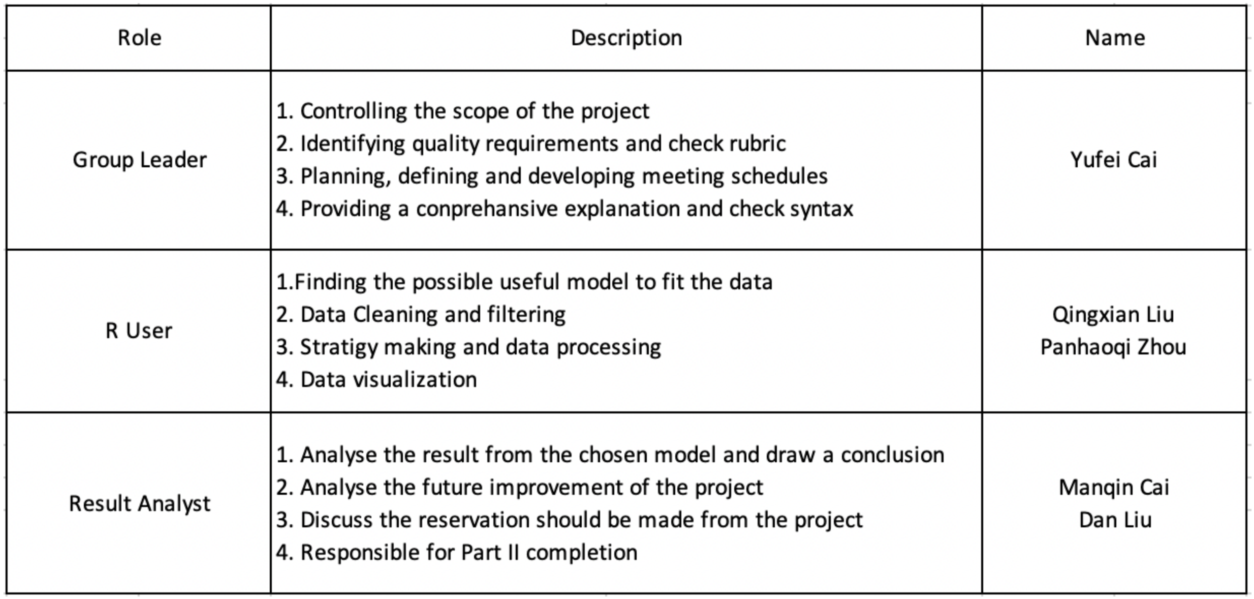
Manqin Cai # 59000448

Dan Liu # 14118566

Qingxian Liu # 89451330

Panhaoqi Zou # 70069018

Exploring Facebook Blog Lifetime Impression Quantity



Role	Description	Name
Group Leader	1. Controlling the scope of the project 2. Identifying quality requirements and check rubric 3. Planning, defining and developing meeting schedules 4. Providing a conprehansive explanation and check syntax	Yufei Cai
R User	1.Finding the possible useful model to fit the data 2. Data Cleaning and filtering 3. Stratigy making and data processing 4. Data visualization	Qingxian Liu Panhaoqi Zhou
Result Analyst	1. Analyse the result from the chosen model and draw a conclusion 2. Analyse the future improvement of the project 3. Discuss the reservation should be made from the project 4. Responsible for Part II completion	Manqin Cai Dan Liu

University of British Columbia

STAT344 Project Proposal

PART I

1 Introduction

With the development of social media, there is a growing concern on how to attract people's attention when making each post online. A social media post with high total impressions numbers (number of times that a post has been seen on social media) will cause external benefit and cost to the society. However, users have difficulty to access their own post quality due to the limited visibility of the average post quality. The previous study giving several performance metrics without discussing the details statistics behind the scenes.In this study, by exploring the average impression numbers, estimating the proportion impression number larger than 20000 and comparing the impression number between the different categories, will contribute to people post decision and the future social media metrics revised

2 Data Analysis

2.1 Methodology

The target populations are the number of impressions of all 500 users' posts(continuous) and whether the post has more than 20000 impressions(binary). Thus, the parameters of interest are the average number of impressions of all 500 users' posts and the proportion of the posts whose number of impressions is above 20000, respectively.

Since the standard deviation of total impressions is too large, we decided to determine the sample size by using proportion of posts with impressions over 20000 (we guess it's 20%). We assume that the ME of 95% CI is about 12%. By using study planning, according to (1) we can get $n_0 \geq \frac{1.962^2 \times 0.2 \times 0.8}{0.11^2} = 50.8$, and then according to (2), we got the sample size is 46. Thus, the sample size is taken as 50 to control the sample size smaller than 10% of the population size (population size = 500).

$$\delta = z_{\alpha/2}\sqrt{\frac{s_{guess}^2}{n}}.$$

(1)

Where δ is the half-width of the CI we desired, and s_{guess}^2 is the standard deviation we assumed

(2) $$n > \frac{n_0 \times N}{N + n_0}$$

Where n_0 is the sample size we calculated without FPC, and N is the population size

By selecting 50 posts randomly, we did the SRS. During SRS, we get a sub-table that contains 50 posts. We got the sample mean and sample standard deviation of number of post impressions and then calculated the vanilla estimator of mean of number of post impressions and its standard error by using the formula.

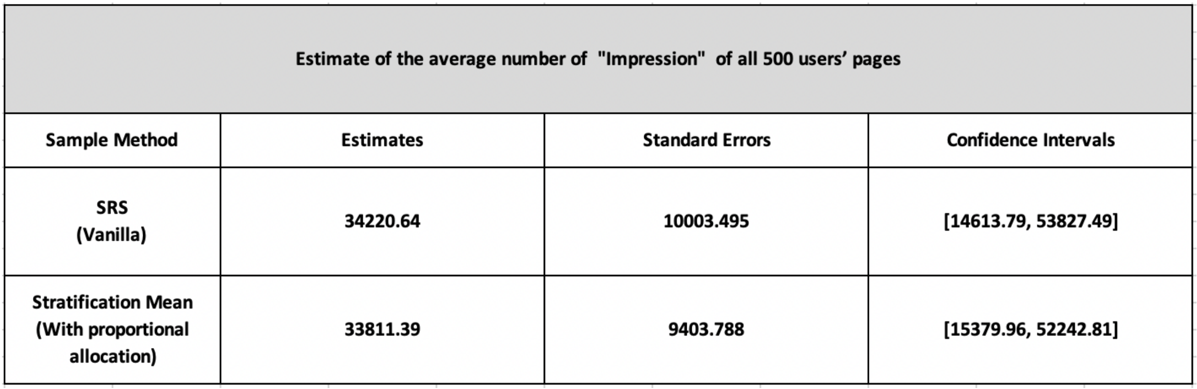
As for the stratified sampling, in our chosen dataset, two categorical variables “Type” and “Category” are regarded as potential candidates for subpopulation-division-principle. However, after the group discussion, we suspect that the number of records of each unique post type may significantly differ from each other. As intuitively the type “Photo” may account for the dominant proportion among all kinds of types as more users are more willing to post photos on Facebook compared with the other types. In that case, the variable “Category” is decided to be taken as the division principle. The strata are selected from each subpopulation according to their types in the column “Category” and their sizes are allocated by proportional allocation.

2.2 Assumption

The population size N is known.

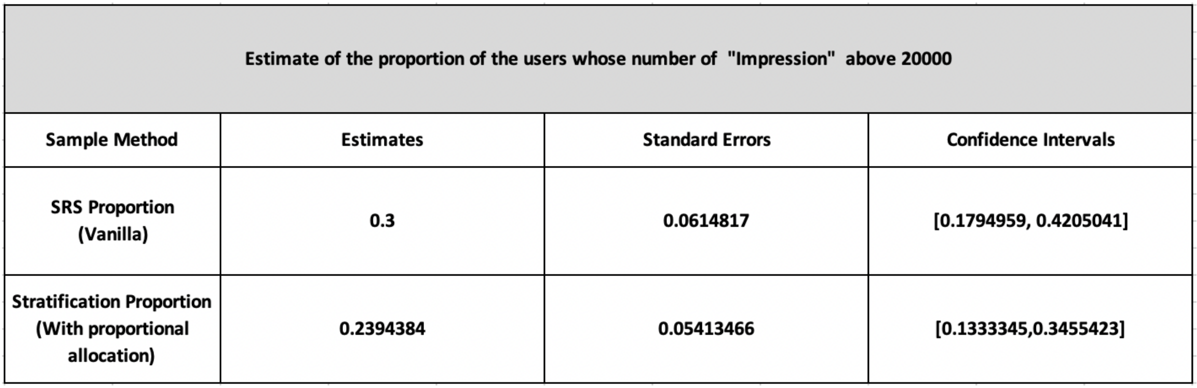
The guessed variances $s_{h,guess}^2$’s and the costs of sampling are all equal across strata.

2.3 Statistics



Estimate of the average number of "Impression" of all 500 users’ pages			
Sample Method	**Estimates**	**Standard Errors**	**Confidence Intervals**
SRS (Vanilla)	**34220.64**	**10003.495**	**[14613.79, 53827.49]**
Stratification Mean (With proportional allocation)	**33811.39**	**9403.788**	**[15379.96, 52242.81]**

Figure 1.1



Estimate of the proportion of the users whose number of "Impression" above 20000			
Sample Method	**Estimates**	**Standard Errors**	**Confidence Intervals**
SRS Proportion (Vanilla)	**0.3**	**0.0614817**	**[0.1794959, 0.4205041]**
Stratification Proportion (With proportional allocation)	**0.2394384**	**0.05413466**	**[0.1333345,0.3455423]**

Figure 1.2

Using simple random sampling analysis and estimation, the average number of impressions is 34220.64, the standard deviation is 10003.495, and the 95% confidence interval is [14613.79, 53827.49]. We estimate that the proportion of users with more than 20000 impression is 0.3000000, the standard deviation is 0.06148170, and the 95% confidence interval is [0.1794959, 0.4205041]. The average number of impressions analyzed and estimated by stratified sampling is 33811.39, the standard deviation is 9403.788, and the 95% confidence interval is [15379.96, 52242.81]. We estimate that the proportion of users with more than 20000 impressions is 0.2394384, the standard deviation is 0. 05413466, and the 95% confidence interval is [0.1333345, 0.3455423].

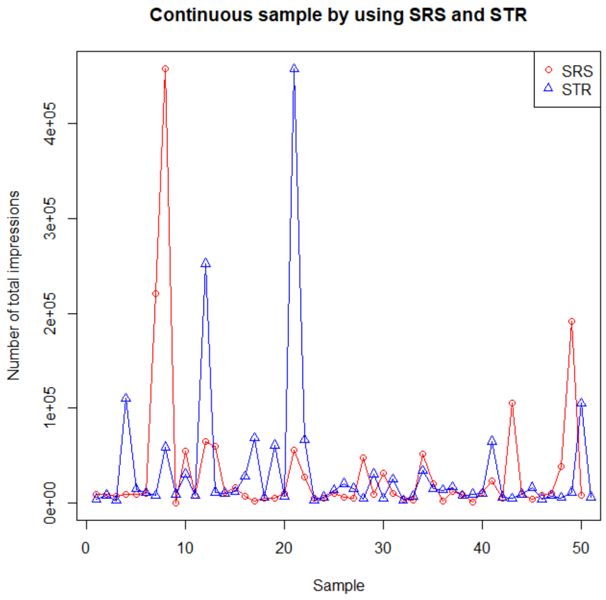


Figure 1.3

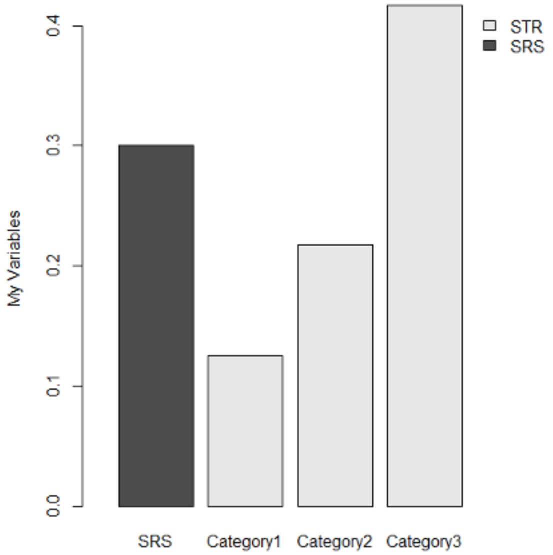


Figure 1.4

In Figure 1.3, x-axis means 50 posts that we sampled, and each number is a sample item, and y-axis represents the number of total impressions each post got. The red line shows the samples by using SRS and the blue line shows the samples by using STR.

In Figure 1.4, x-axis means the domain of the samples, and y-axis represents the percentage of the posts with over 20000 impressions in each domain. The black box shows the proportion of samples by using SRS and the grey boxes show the proportion of samples by using STR.

Based on the results shown above, for both target populations, it is clear to see that the standard error calculated by stratified sampling is smaller than the corresponding standard error calculated by SRS.

Moreover, the width of 95% confidence interval calculated by stratified sampling is narrower than that calculated by SRS as well. Let $\bar{y}_s$ be the sample mean when a single SRS (of equal size) is obtained from the whole population without stratified sampling.

Let $\overline{y_{str}}$ be estimator using stratified sampling with proportional allocation (assume assumptions hold). As the $\frac{Var(\overline{y_{str}})}{Var(\overline{y_s})} = \frac{s_{P,W}^2}{s_P^2} < 1$ and according to (3), it is reasonable that the standard error of stratified estimator will smaller than that of SRS estimator (such as Vanilla estimator), which may also indicate that either the between-strata variance is comparatively large or the within-strata variance is comparatively small.

$$s_P^2 = s_{P,W}^2 + s_{P,B}^2 \tag{3}$$

$Where\ s_{P,W}^2$ is the within-strata variance, and $s_{P,B}^2$ is the between-strata variance

2.4 Comparison

2.4.1 SRS

Advantage: Simple and easy to use

Limitations: If between-strata variance is large, $Var(\overline{y_s})$ will be larger than stratified sampling

2.4.2 Stratified

Advantage: $Var(\overline{y_{str}})$ will be small since the standard error of stratified estimator is only controlled by within-strata variance. Moreover, as we select sample from each sub-population, it avoids “rare” samples that aren’t representative.

Limitations: If the sample within-strata variance is large, then $Var(\overline{y_{str}})$ will also be large and it’s more complex than SRS.

3 Conclusion

We use simple random sampling and stratified sampling to obtain samples and estimates the average number of impressions of all 500 user pages and the proportion of users with more than 20000 impressions. The results obtained by the two different sampling methods are accurate to the actual population parameters within an acceptable range. At the same time as data analysis, we also compare the two sampling methods. Compared with the simple random sampling method, stratified random sampling has a minor estimation error and higher accuracy. However, it is more complicated.

In conclusion, A high number of impressions is strong proof that a post has enough appeal. Yet how to make posts of high quality and interest is difficult. This study explores the average number of likes, which allows people to look at the quality of their posts. It also shows that catogory3 "non-explicit brand related post" posts have more impressions, which also provides a reference for users who want to gain benefits from their posts.

4 Discussion

One of the limitations of this study is that it may involve biased choices. This leads to deviations. We must identify each member of the population under study and classify them into a subgroup and only one subgroup. A user does not necessarily correspond to exactly one "category". Therefore, there may be deviations in our layering.

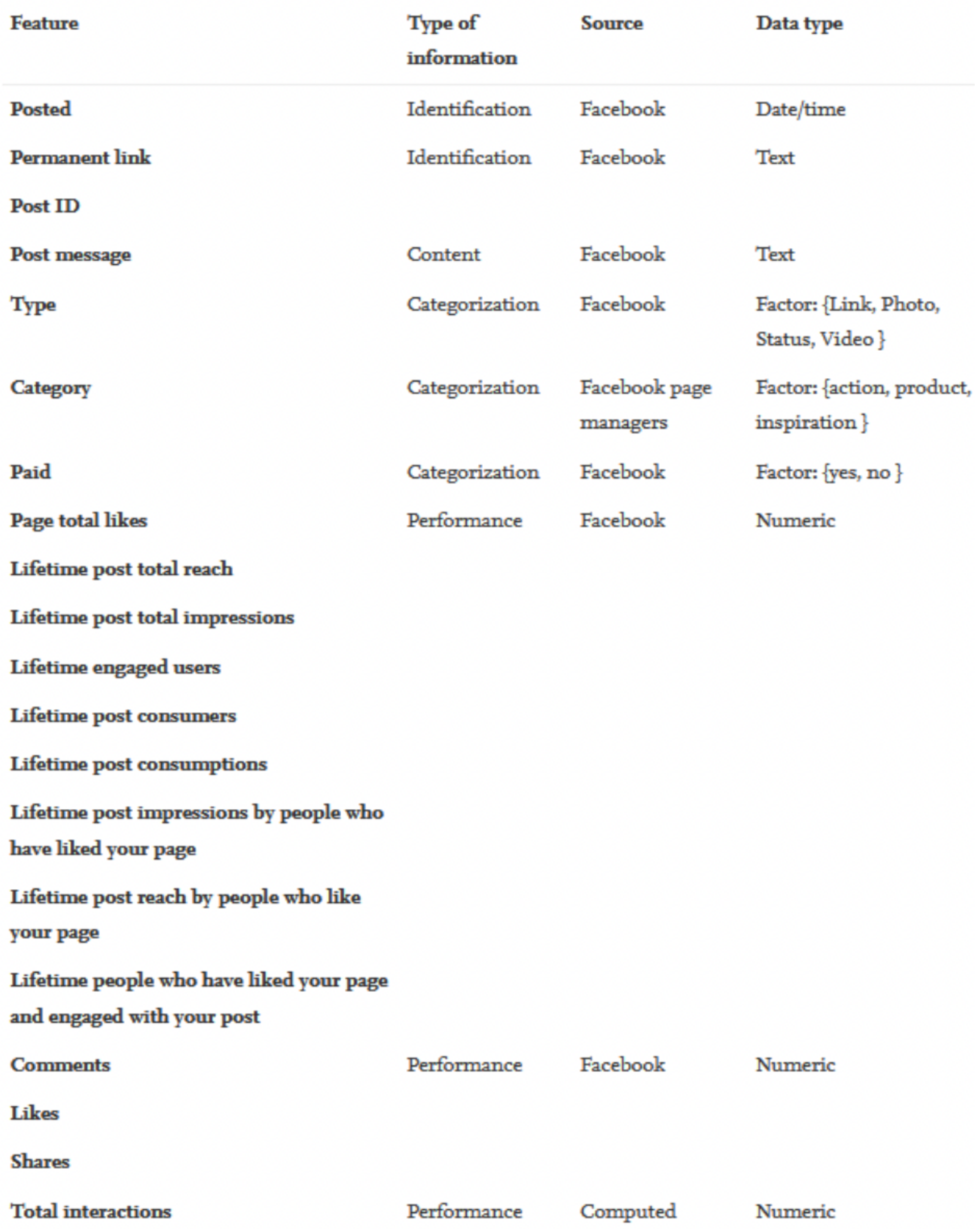
The most accurate conclusion can only be reached with an appropriate sample size. Smaller samples will give results that may not be representative of the whole. Fifty user samples are not enough to represent global Facebook users. Therefore, we cannot confidently say that our conclusions can be extended to more extensive or other populations

PART II

This article by Perlman and Wu concerns about the appearance of striking examples of allegedly inferior likelihood ratio tests (LRT) in the statistical literature. These examples usually appear in multi-parameter hypothesis testing problems and have common characteristics which dominate the LRT because it is more powerful anywhere. And the article claims this conclusion is wrong, the LR criterion is not inferior. In each case, the so-called superiority test violates statistical intuition, it will draw unreasonable conclusions in the worst case, and it is only applicable to some restrictive prior distribution at most. As the authors put, scientific intuition is extremely important in scientific studies in general therefore tests that get conclusions against scientific intuition are scientifically inappropriate tests. Therefore, although LRT is not absolutely correct for hypothesis testing, statisticians believe that LRT is still the first choice for non-Bayesian parameter hypothesis testing.

APPENDIX

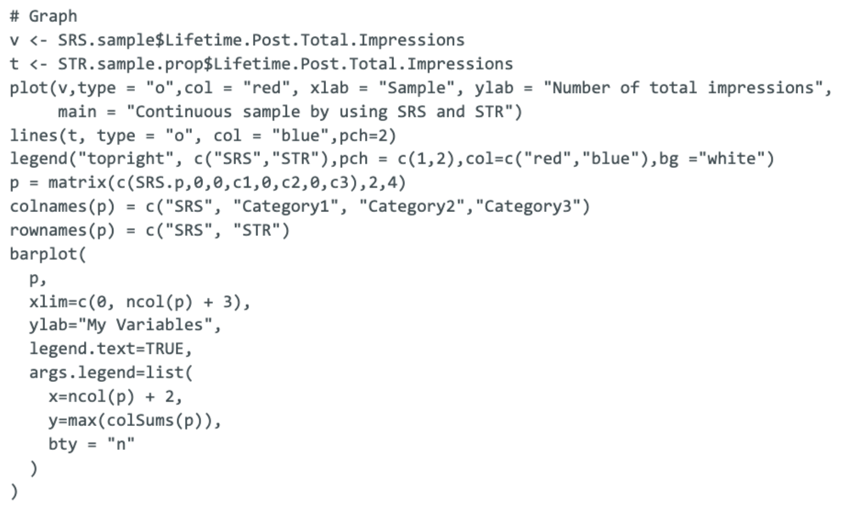
APPENDIX 1 – Data



Feature	Type of information	Source	Data type
Posted	Identification	Facebook	Date/time
Permanent link	Identification	Facebook	Text
Post ID			
Post message	Content	Facebook	Text
Type	Categorization	Facebook	Factor: {Link, Photo, Status, Video }
Category	Categorization	Facebook page managers	Factor: {action, product, inspiration }
Paid	Categorization	Facebook	Factor: {yes, no }
Page total likes	Performance	Facebook	Numeric
Lifetime post total reach			
Lifetime post total impressions			
Lifetime engaged users			
Lifetime post consumers			
Lifetime post consumptions			
Lifetime post impressions by people who have liked your page			
Lifetime post reach by people who like your page			
Lifetime people who have liked your page and engaged with your post			
Comments	Performance	Facebook	Numeric
Likes			
Shares			
Total interactions	Performance	Computed	Numeric

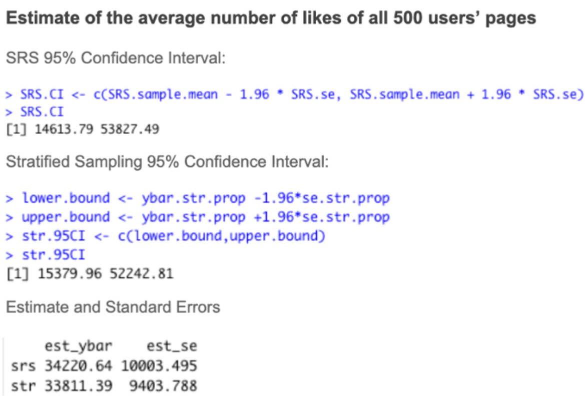
APPENDIX 2 – R Code

```
#load the dataset

fb_data = read.csv("/dataset_Facebook.csv", sep = ";",header = T)

attach(fb_data)
#population size for different Categories
N.h <- tapply(Lifetime.Post.Total.Impressions, Category, length)
#name of the Categories
Categories <- names(N.h)
Categories
N.h
detach(fb_data)

N <- sum(N.h)

#The code for SRS
#Population mean of Lifetime.Post.Total.Impressions
true.value <- mean(fb_data$Lifetime.Post.Total.Impressions)
true.value
#Calculate suitable sample size

#Total sample size
n <- 50
#Initialize the SRS sample
set.seed(10)
SRS.indices <- sample.int(N,  n, replace = F)
SRS.sample <- fb_data[SRS.indices , ]
# sample mean
SRS.sample.mean <- mean(SRS.sample$Lifetime.Post.Total.Impressions)
SRS.sample.mean

# sample variance
SRS.sample.variance = var(SRS.sample$Lifetime.Post.Total.Impressions)

# sample standard deviation
SRS.sd = sqrt(SRS.sample.variance)

# calculate 95% CI for population mean
SRS.se <- sqrt( (1-n/N)/n )*SRS.sd
SRS.se
SRS.CI <- c(SRS.sample.mean - 1.96 * SRS.se, SRS.sample.mean + 1.96 * SRS.se)
SRS.CI
srs <- c(est_ybar=SRS.sample.mean, est_se=SRS.se)

#The code for Stratified Sampling with proportional allocation
#Initialize the stratified sample
#h subpopulation sample size
n.h.prop <- round( (N.h/N) * n)
STR.sample.prop <- NULL
for (i in 1: length(Categories))
{
  row.indices <- which(fb_data$Category == Categories[i])
  sample.indices <- sample(row.indices, n.h.prop[i], replace = F)
  STR.sample.prop <- rbind(STR.sample.prop, fb_data[sample.indices, ])
}

ybar.h.prop <- tapply(STR.sample.prop$Lifetime.Post.Total.Impressions, STR.sample.prop$Category, mean)
var.h.prop <- tapply(STR.sample.prop$Lifetime.Post.Total.Impressions, STR.sample.prop$Category, var)
se.h.prop <- sqrt((1 - n.h.prop / N.h) * var.h.prop / n.h.prop)
rbind(ybar.h.prop, se.h.prop)

#Stratified sample mean of Lifetime.Post.Total.Impressions
ybar.str.prop  <- sum(N.h / N * ybar.h.prop)
#Standard error of ybar.str.prop
se.str.prop <- sqrt(sum((N.h / N)^2 * se.h.prop^2))
str.prop <- c(est_ybar=ybar.str.prop , est_se=se.str.prop)
#95% CI for population mean
lower.bound <- ybar.str.prop -1.96*se.str.prop
lower.bound
upper.bound <- ybar.str.prop +1.96*se.str.prop
upper.bound
str.95CI <- c(lower.bound,upper.bound)

rbind(srs=srs, str=str.prop)

#estimate the proportion of posts whose likes are more than 20000
#SRS
SRS.sample2 = SRS.sample[SRS.sample$Lifetime.Post.Total.Impressions >=20000,]
SRS.p=length(SRS.sample2$Lifetime.Post.Total.Impressions)/length(SRS.sample$Lifetime.Post.Total.Impressions)
SRS.sample.p.mean = SRS.p
SRS.sample.p.mean
SRS.se.p = sqrt((1 - n / N)*SRS.p*(1-SRS.p)/n)
SRS.se.p
SRS.p.CI <- c(SRS.sample.p.mean - 1.96 * SRS.se.p, SRS.sample.p.mean + 1.96 * SRS.se.p)
SRS.p.CI
srsp <- c(est_prop=SRS.p,est_se=SRS.se.p)

#The code for Stratified Sampling with proportional allocation
STR.sample.c1 <- STR.sample.prop$Lifetime.Post.Total.Impressions[fb_data$Category==1]
STR.sample.c2 <- STR.sample.prop$Lifetime.Post.Total.Impressions[fb_data$Category==2]
STR.sample.c3 <- STR.sample.prop$Lifetime.Post.Total.Impressions[fb_data$Category==3]

c1 <- length(na.omit(STR.sample.c1[STR.sample.c1>=20000]))/length(na.omit(STR.sample.c1))
c2 <- length(na.omit(STR.sample.c1[STR.sample.c2>=20000]))/length(na.omit(STR.sample.c2))
c3 <- length(na.omit(STR.sample.c1[STR.sample.c2>=20000]))/length(na.omit(STR.sample.c3))

phat.h.prop <- c(c1,c2,c3)
phat.var.h.prop <- c(c1*(1-c1),c2*(1-c2),c3*(1-c3))
phat.se.h.prop <- sqrt((1 - n.h.prop / N.h) * phat.var.h.prop / n.h.prop)
rbind(phat.h.prop, phat.se.h.prop)

phat.str.prop <- sum(N.h / N * phat.h.prop)
phat.str.prop
phat.se.str.prop <- sqrt(sum((N.h / N)^2 * phat.se.h.prop^2))
#phat.se.str.prop <- sqrt(sum((N.h / N)^2 *(1 - n.h.prop / N.h) * phat.var.h.prop / n.h.prop))
phat.se.str.prop
str.prop <- c(phat.str.prop, phat.se.str.prop)
lower.bound <- phat.str.prop-1.96*phat.se.str.prop
lower.bound
upper.bound <- phat.str.prop+1.96*phat.se.str.prop
upper.bound
str.95CI <- cbind(lower.bound,upper.bound)
strp.prop <- c(phat.str.prop,phat.se.str.prop)
rbind(srs=srsp, str=strp.prop)
```



```
# Graph
v <- SRS.sample$Lifetime.Post.Total.Impressions
t <- STR.sample.prop$Lifetime.Post.Total.Impressions
plot(v,type = "o",col = "red", xlab = "Sample", ylab = "Number of total impressions",
     main = "Continuous sample by using SRS and STR")
lines(t, type = "o", col = "blue",pch=2)
legend("topright", c("SRS","STR"),pch = c(1,2),col=c("red","blue"),bg ="white")
p = matrix(c(SRS.p,0,0,c1,0,c2,0,c3),2,4)
colnames(p) = c("SRS", "Category1", "Category2","Category3")
rownames(p) = c("SRS", "STR")
barplot(
  p,
  xlim=c(0, ncol(p) + 3),
  ylab="My Variables",
  legend.text=TRUE,
  args.legend=list(
    x=ncol(p) + 2,
    y=max(colSums(p)),
    bty = "n"
  )
)
```

APPENDIX 3 – R Result



Estimate of the average number of likes of all 500 users' pages

SRS 95% Confidence Interval:

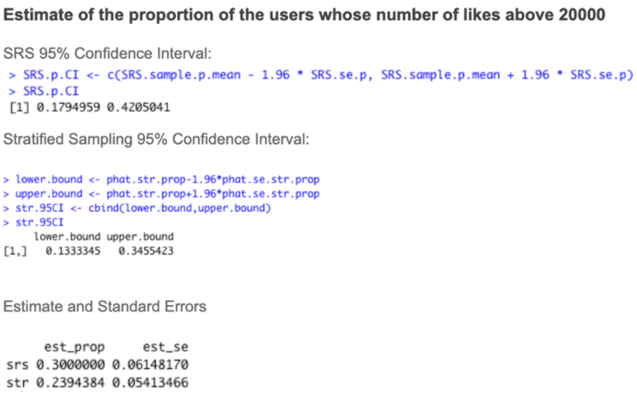
```
> SRS.CI <- c(SRS.sample.mean - 1.96 * SRS.se, SRS.sample.mean + 1.96 * SRS.se)
> SRS.CI
[1] 14613.79 53827.49
```

Stratified Sampling 95% Confidence Interval:

```
> lower.bound <- ybar.str.prop -1.96*se.str.prop
> upper.bound <- ybar.str.prop +1.96*se.str.prop
> str.95CI <- c(lower.bound,upper.bound)
> str.95CI
[1] 15379.96 52242.81
```

Estimate and Standard Errors

```
    est_ybar    est_se
srs 34220.64 10003.495
str 33811.39  9403.788
```



Estimate of the proportion of the users whose number of likes above 20000

SRS 95% Confidence Interval:

```
> SRS.p.CI <- c(SRS.sample.p.mean - 1.96 * SRS.se.p, SRS.sample.p.mean + 1.96 * SRS.se.p)
> SRS.p.CI
[1] 0.1794959 0.4205041
```

Stratified Sampling 95% Confidence Interval:

```
> lower.bound <- phat.str.prop-1.96*phat.se.str.prop
> upper.bound <- phat.str.prop+1.96*phat.se.str.prop
> str.95CI <- cbind(lower.bound,upper.bound)
> str.95CI
     lower.bound upper.bound
[1,]   0.1333345   0.3455423
```

Estimate and Standard Errors

```
     est_prop     est_se
srs 0.3000000 0.06148170
str 0.2394384 0.05413466
```

REFERENCES

(Moro et al., 2016) S. Moro, P. Rita and B. Vala. Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach. Journal of Business Research, Elsevier, In press. Available at: http://dx.doi.org/10.1016/j.jbusres.2016.02.010

Michael D. Perlman, Lang Wu. "The Emperor's new tests." Statistical Science, 14(4) 355-369 November 1999. Available at: https://doi.org/10.1214/ss/1009212517