Sampling Without Replacement and Logistic Regression in 2019 Canadian Election Data

Zijian Wang

2020/10/07

Abstract

In this report, we fit the logistic regression model to predict if respondents will vote Liberal Party or not. Then compare the accuracy of parameter estimate with the model with sampling in different sizes.
2019 Canadian Election Datasets were used. The conclusion of this essay can help us to choose the appropriate size for sampling.

1.Introduction

I work for the Liberal Party to work and used Election data in Ontario for research.
Target population: All Canadians who are living in ON. The population is the entire group of people or objects to which the researcher wishes to generalize the study findings.
Frame: A list of households registered in the government whit the connect method—for example, phone number, email address, or address. A sampling frame is a list of all the items in your population. It’s a complete list of everyone or everything you want to study. The difference between a population and a sampling frame is that the general population and the frame are specific.
Sample: 1000/2000/5000 Ontario household owner. The sample is the selected elements (people or objects) chosen for participation in a study.

2 survey methodology

2.1 sampling methods

In this report, we used sampling without replacement in the 2019 Canadian Election Dataset. The target is choosing an appropriate sampling size. ### 2.2 Way to reach

I prefer to send target people a email and every respondents will get a $5 gift card

2.3 Cost estimate

Cost: (5$ gift card)*sample size.

2.4 Non-response

Non-response will be seem as missing value. To avoid of affect by missing value I will input with median Values.

2.5 Respondent privacy

Every member of our team will sign a privacy protection agreement. And all the names and email addresses will be stored after Md5 encrypt.

3 Simulate and Test

3.1 Prepare

In this step, I filter respondents from Ontario and choose gender and education as a predictor to fit the model to predict if respondents will vote Liberal Party or not.

```
library(data.table)
library(haven)
library(ggplot2)
library(Metrics)
library(dplyr)
library(caret)
library(InformationValue)
setwd("D:/R_17")
dt <- read_dta(file = "2019+Canadian+Election+Study+-+Online+Survey+v1.0.dta")
dt <- data.table(dt)
dtBackup <- dt
dt <- dt[dt$cps19_province==22,
         c("cps19_gender","cps19_education","cps19_votechoice")]
names(dt) <- c("gender","education","vote")
dt[,":="(gender=as.numeric(gender),
         education=as.numeric(education),
         vote=as.numeric(vote))]
dt <- na.omit(dt)
dt[,":="(male=ifelse(gender==1,1,0),
         female=ifelse(gender==2,1,0),
         other=ifelse(gender==3,1,0),
         vote=ifelse(vote==1,1,0))]
dt <- dt[,c("male","female","other","education","vote")]
```

3.2 Logistic Regression with population

A fit logistic regression model with gender and education to predict if respondents will vote Liberal Party or not. I choose Misclassification error as metrics. Misclassification error is the percentage mismatch of predict value vs. actual value, irrespective of 1's or 0's. The lower the misclassification error, the better is the model. The misclassification error of Logistic Regression with the population is 0.3326309, indicating that the model correctly predicts about 66.74% of respondents.

```
lrMod <- glm(vote ~., data = dt, family = binomial)
votePrdProb <- predict(lrMod,dt, type = "response")
optCutOff <- optimalCutoff(dt$vote, votePrdProb)[1]
votePrdBin <- ifelse(votePrdProb>optCutOff,1,0)
mean(votePrdBin!=dt$vote)
```

```
## [1] 0.3326309
```

3.3 Logistic Regression with Sample

Define a function to implement logistic regression with the sample. The input of the function is a sample, and the output is a Misclassification error.

```
logisticReg <- function(dataS){
  lrMod = glm(vote ~., data = dataS, family = binomial)
  optCutOff = optimalCutoff(dataS$vote, lrMod$fitted.values)[1]
  votePrdProb = predict(lrMod,dt, type = "response")
  votePrdBin = ifelse(votePrdProb>optCutOff,1,0)
  classError = mean(votePrdBin!=dt$vote)
  return(classError)
}
```

The code below uses the first 50 respondents to fit the model then used a model to predict the population.

```
logisticReg(dt[c(1:50),])
```

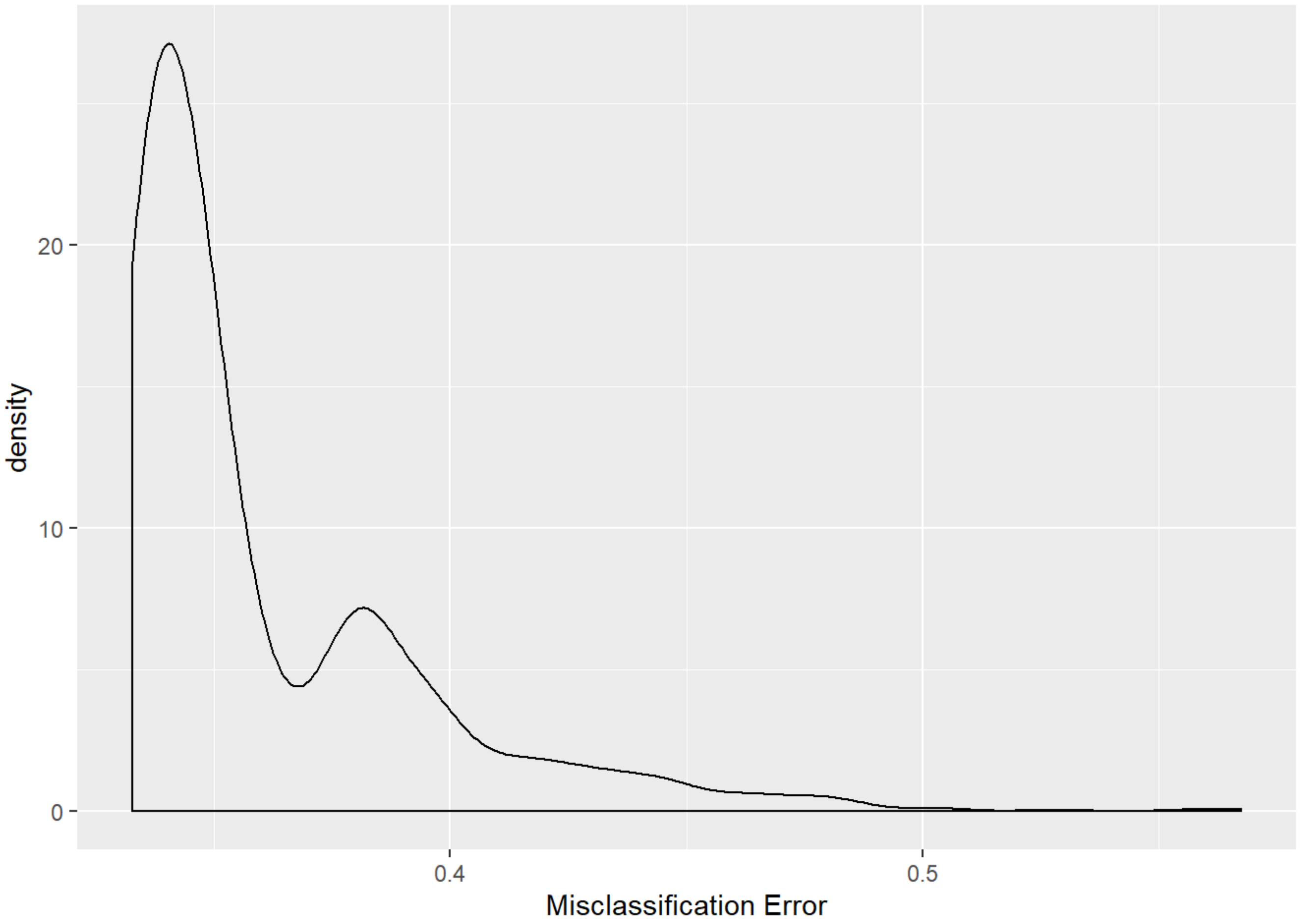
```
## [1] 0.6176041
```

Define a function to implement function above with select size and loop. The function’s input is the size loop number, and the output is the Misclassification error list.

```
getClassErroeWithSample <- function(size,N=1000){
  classErrorList = rep(0,N)
  for (i in c(1:N)) {
    indexS = sample(c(1:nrow(dt)),size,replace = F)
    dataS = dt[indexS,]
    classError = logisticReg(dataS)
    classErrorList[i] = classError
  }
  return(classErrorList)
}
```

The code below sample 100 respondents to fit the model and used the model to predict the population once repeat 1000 times, which is the default.
The figure below shows the distribution of Misclassification error.

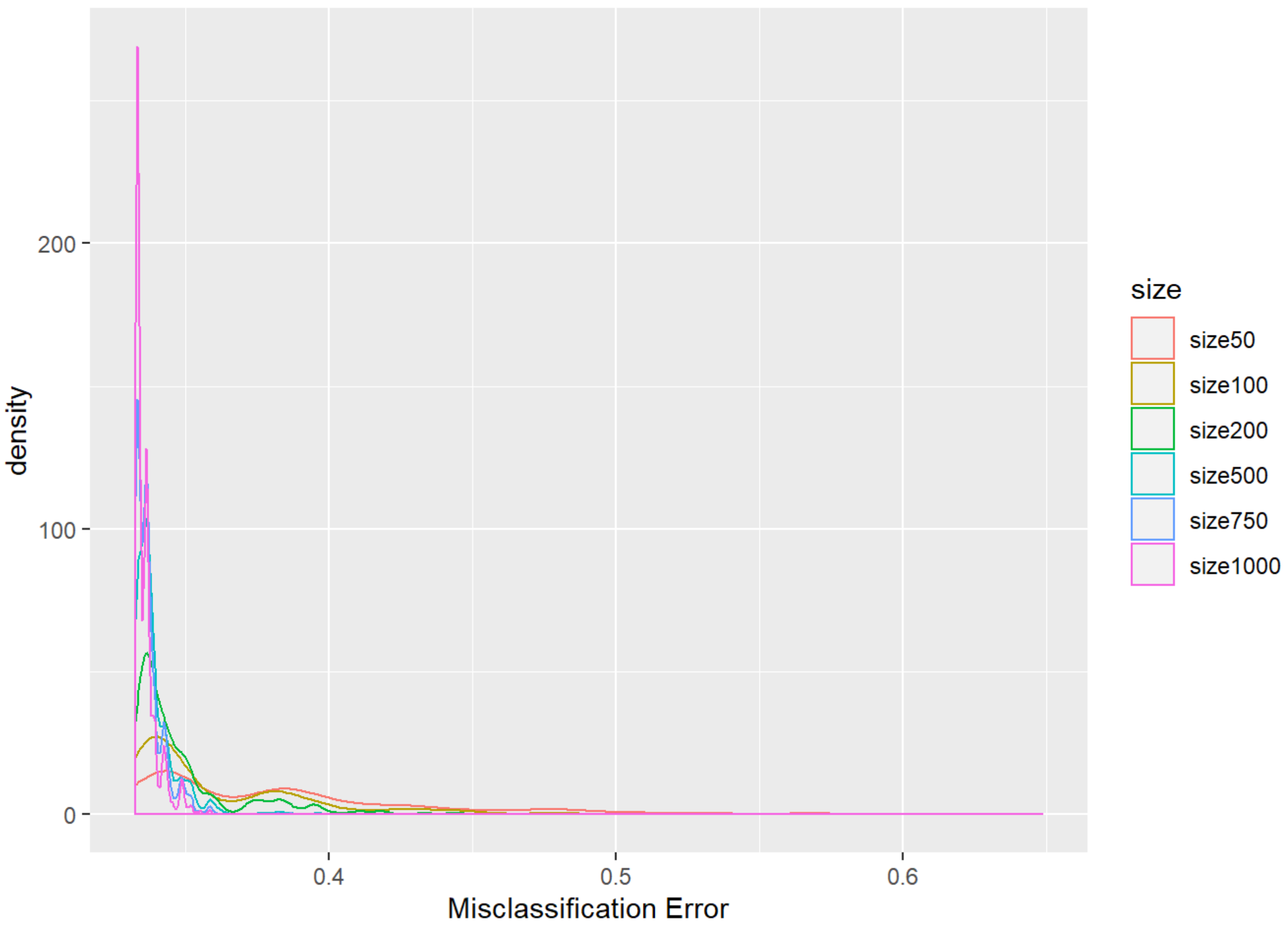
```
ce100 <- getClassErroeWithSample(100)
ggplot(data.table(ce100),aes(x=ce100))+
  geom_density()+
  labs(x="Misclassification Error")
```



Code below try sample size 50,100,200,500 and 1000.

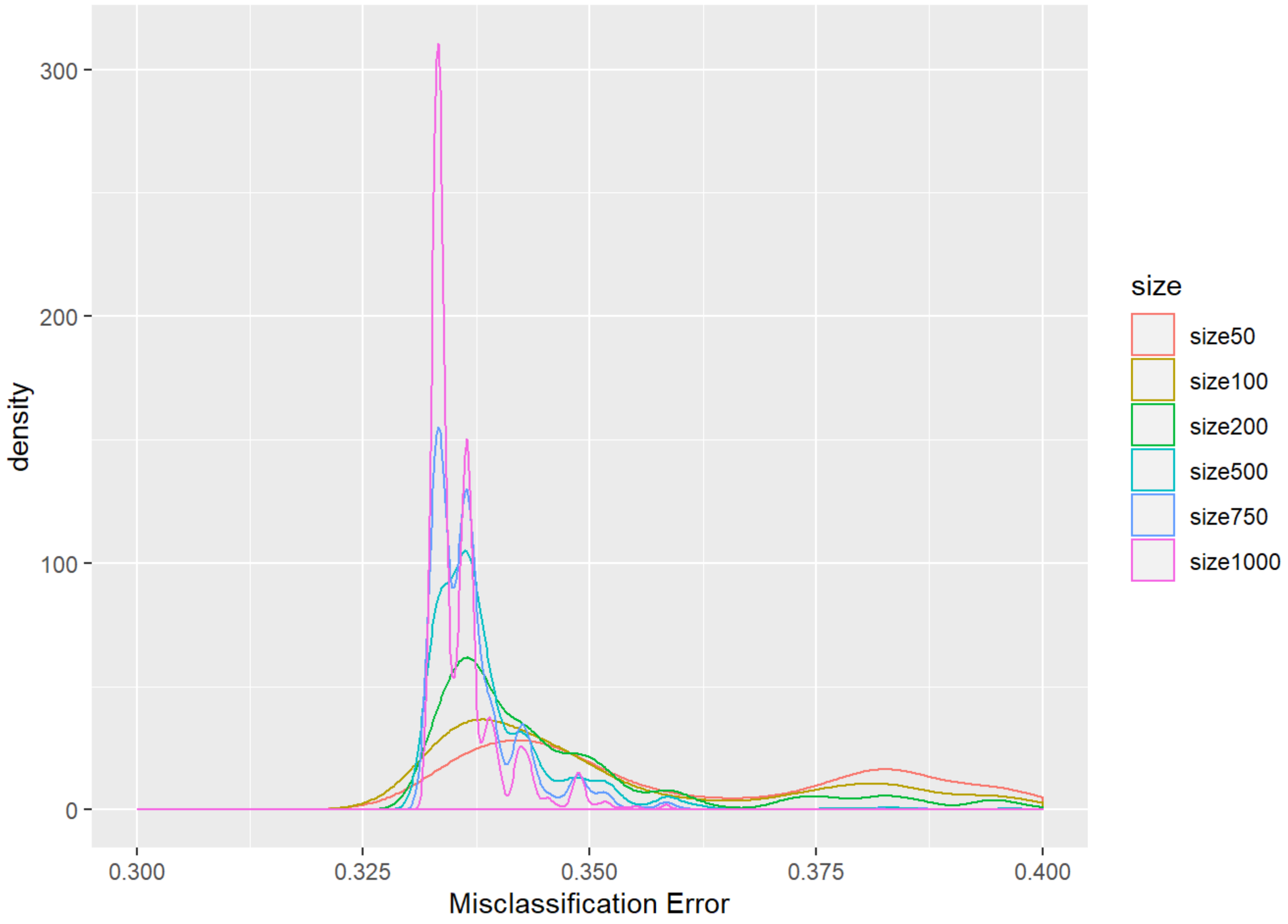
The figure below shows the distribution of the Misclassification error of different sample sizes.

```
classErrorDt <- data.table(size50=getClassErroeWithSample(50),
                           size100=getClassErroeWithSample(100),
                           size200=getClassErroeWithSample(200),
                           size500=getClassErroeWithSample(500),
                           size750=getClassErroeWithSample(750),
                           size1000=getClassErroeWithSample(1000))
classErrorDt2 <- melt(classErrorDt)
names(classErrorDt2)[1] <- "size"
ggplot(classErrorDt2,aes(x=value,color=size))+
  geom_density()+
  labs(x="Misclassification Error")
```



The figure below takes attention to the interval from 0.3 to 0.4.

```
ggplot(classErrorDt2,aes(x=value,color=size))+
  geom_density()+
  xlim(0.3,0.4)+
  labs(x=″Misclassification Error″)
```

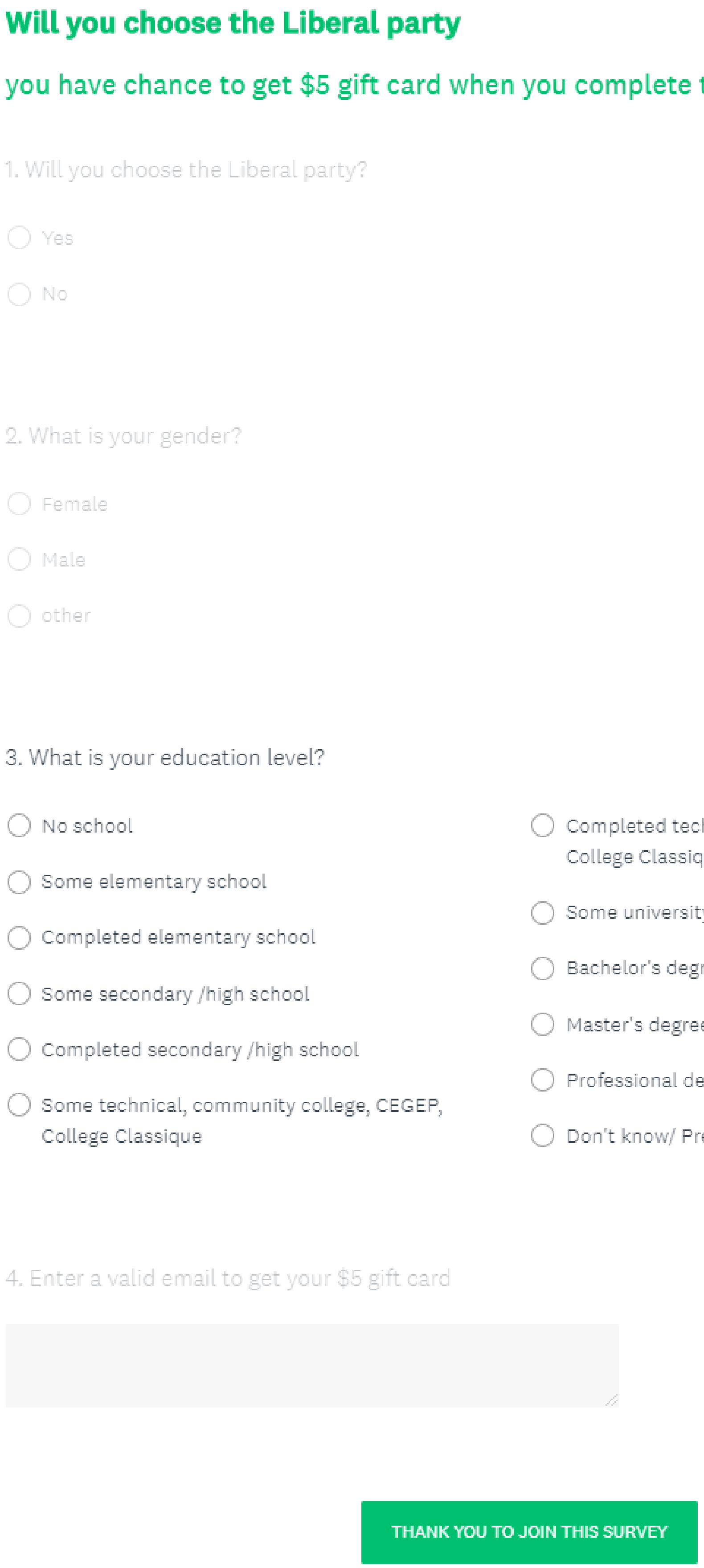


4 Results and Discussion

The survey needs to consider both accuracy and cost. So we need to average out for sample size.
As a result above, we suggest choosing a sampling size of 500, which can predict precise enough with cost acceptable. # 5 appendices

```
knitr::opts_chunk$set(echo = TRUE,
                      warning = FALSE,
                      message=FALSE)
library(data.table)
library(haven)
library(ggplot2)
library(Metrics)
library(knitr)
library(tidyverse)
Sys.setlocale("LC_TIME", "English")
options(scipen = 200)
library(data.table)
library(haven)
library(ggplot2)
library(Metrics)
library(dplyr)
library(caret)
library(InformationValue)
setwd("D:/R_17")
dt <- read_dta(file = "2019+Canadian+Election+Study+-+Online+Survey+v1.0.dta")
dt <- data.table(dt)
dtBackup <- dt
dt <- dt[dt$cps19_province==22,
         c("cps19_gender","cps19_education","cps19_votechoice")]
names(dt) <- c("gender","education","vote")
dt[,":="(gender=as.numeric(gender),
         education=as.numeric(education),
         vote=as.numeric(vote))]
dt <- na.omit(dt)
dt[,":="(male=ifelse(gender==1,1,0),
         female=ifelse(gender==2,1,0),
         other=ifelse(gender==3,1,0),
         vote=ifelse(vote==1,1,0))]
dt <- dt[,c("male","female","other","education","vote")]
lrMod <- glm(vote ~., data = dt, family = binomial)
votePrdProb <- predict(lrMod,dt, type = "response")
optCutOff <- optimalCutoff(dt$vote, votePrdProb)[1]
votePrdBin <- ifelse(votePrdProb>optCutOff,1,0)
mean(votePrdBin!=dt$vote)
logisticReg <- function(dataS){
  lrMod = glm(vote ~., data = dataS, family = binomial)
  optCutOff = optimalCutoff(dataS$vote, lrMod$fitted.values)[1]
  votePrdProb = predict(lrMod,dt, type = "response")
  votePrdBin = ifelse(votePrdProb>optCutOff,1,0)
  classError = mean(votePrdBin!=dt$vote)
  return(classError)
}
logisticReg(dt[c(1:50),])
getClassErroeWithSample <- function(size,N=1000){
  classErrorList = rep(0,N)
  for (i in c(1:N)) {
    indexS = sample(c(1:nrow(dt)),size,replace = F)
    dataS = dt[indexS,]
    classError = logisticReg(dataS)
    classErrorList[i] = classError
  }
  return(classErrorList)
}
ce100 <- getClassErroeWithSample(100)
ggplot(data.table(ce100),aes(x=ce100))+
  geom_density()+
  labs(x="Misclassification Error")
classErrorDt <- data.table(size50=getClassErroeWithSample(50),
                           size100=getClassErroeWithSample(100),
                           size200=getClassErroeWithSample(200),
                           size500=getClassErroeWithSample(500),
                           size750=getClassErroeWithSample(750),
                           size1000=getClassErroeWithSample(1000))
classErrorDt2 <- melt(classErrorDt)
names(classErrorDt2)[1] <- "size"
ggplot(classErrorDt2,aes(x=value,color=size))+
  geom_density()+
  labs(x="Misclassification Error")
ggplot(classErrorDt2,aes(x=value,color=size))+
  geom_density()+
  xlim(0.3,0.4)+
  labs(x="Misclassification Error")
```

All of code were show above.
github url: https://github.com/nerowangg/Sampling-Without-Replacement-and-Logistic-Regression-in-2019-Canadian-Election-Data.git (https://github.com/nerowangg/Sampling-Without-Replacement-and-Logistic-Regression-in-2019-Canadian-Election-Data.git) The survey url is https://www.surveymonkey.com/r/NFQJZ6J (https://www.surveymonkey.com/r/NFQJZ6J).
The screenshot was show below.



Will you choose the Liberal party

you have chance to get $5 gift card when you complete this survey

1. Will you choose the Liberal party?

○ Yes

○ No

2. What is your gender?

○ Female

○ Male

○ other

3. What is your education level?

○ No school

○ Some elementary school

○ Completed elementary school

○ Some secondary /high school

○ Completed secondary /high school

○ Some technical, community college, CEGEP, College Classique

○ Completed technical, community college, CEGEP, College Classique

○ Some university

○ Bachelor's degree

○ Master's degree

○ Professional degree or doctorate

○ Don't know/ Prefer not to answer

4. Enter a valid email to get your $5 gift card

# References

- J. N. K. Rao, 1966, On the Comparison of Sampling with and without Replacement
- 2019 Canadian Election Dataset were used.
- package: data.table haven ggplot2 Metrics knitr tidyverse