

Simulation of data

Yuze Kang, Yijie Zhao, Shuyu Duan, Rachel Oh

2020/10/8

Data Simulation

We should have total 6 variables obtained from the results of survey. We used R and tidyverse to simulate data for estimation and analysis.(Wickham et al. 2019; R Core Team 2020) Some real data from the past are referenced in order to make the simulation results closer to the real situation.

Age

- (1) For age, from Statistics Canada we've known the number of people of each age group in Ontario population, in this case we divide population into 4 groups(0–18, 19–49, 50–89, 90 and over), and by calculation the percentages are 21%, 41%, 37% and 1% respectively.(see Statistics Canada, n.d.)

```
# Size of population(Number of simulations)
N<-100000
set.seed(538)
# Age
# We assume in each age group, ages appears at equal probability, so it follows a uniform distribution.
Age_0_18<-runif(N* 0.21, min = 0, max = 18)
Age_19_49<-runif(N*0.41, min = 19, max = 49)
Age_50_89<-runif(N*0.37, min = 50, max = 89)
Age_90_over<-runif(N*0.010, min = 90, max = 110)
#This is the simulated Age data
Age_all<-c(Age_0_18, Age_19_49, Age_50_89, Age_90_over)
# Shuffle the data
Age<-sample(Age_all)
```

Gender

- (2)For Gender, we assume the ratio of female and male is 1:1.

```
set.seed(233)
#Gender
gender_types<-c("Female","Male")
Gender<-sample(gender_types, N, replace = T)
```

District

- (3) For district the person is living in, there are total 124 districts in Ontario, we assume each person has the equal probability to live in any district.

```
set.seed(304)
#District
District<-sample(1:124, N, replace = T)
```

Political party

- (4) We referenced the results of election in 2019 and will simulate according to the vote share in Ontario.(CBCNews, n.d.)

*LIB: 41.4%

*CON: 33.2%

*NDP: 16.8%

*GRN: 6.2%

*IND: 0.4%

*PP: 1.6%

*OTH: 0.4%

```
set.seed(347)
#Political Party supported
#Simulate according to the past vote share
Poli_Parties<-c("LIB","CON","NDP","GRN","IND","PP","OTH")
PoliticalP_supported<-sample(Poli_Parties,N,prob = c(0.414, 0.332, 0.168, 0.062, 0.004, 0.016, 0.004), replace = T)
```

Education Level

- (5) We referenced from the 2016 education attainment status in Canada, and will simulate according to this.(see Canada, n.d.a)

*Bachelor's degree or higher: 28.5%

*University below bachelor's 3.1%

*College diploma 22.4%

*Apprenticeship or other trades certificate 10.8%

*High school diploma 23.7%

*No certificate, diploma or degree 11.5%

```
set.seed(302)
#Education level
Edu_level_types<-c("Bachelor or higher","University below bachelor's", "College diploma","Apprenticeship or other trades certificate","High school diploma","No certificate, diploma or degree")
Edu_level<-sample(Edu_level_types, N, prob = c(0.285, 0.031, 0.224, 0.108, 0.237, 0.115),replace = T)
```

Income

(6) We simulate income in the same way we simulate age. We referenced an after-taxed income results in 2018, from Statistics Canada. (Canada, n.d.b)

```
set.seed(303)
#2018 after taxed
Income_under_15000<- runif(N*0.2, min = 0, max = 14999)
Income_15000_24999<- runif(N*0.2, min = 15000, max = 24999)
Income_25000_39999<- runif(N*0.2, min = 25000, max = 39999)
Income_40000_59999<- runif(N*0.2, min = 40000, max = 59999)
Income_60000_over<- runif(N*0.2, min = 60000, max = 500000)
Income_all<-c(Income_under_15000, Income_15000_24999, Income_25000_39999, Income_40000_59999, Income_60000_over)
#shuffle
Income<-sample(Income_all)
```

Now we combine all variables together to for a new data frame, then sample from these observations.

```
set.seed(305)
#Combine together
All_together.df<-tibble(Age, Gender, PoliticalP_supported, Edu_level, Income, District)
#Sampling, using SRS, randomly pick from the data frame.
sample_selcet_rows<-sample(N,10000)
sample_dataset<-All_together.df[sample_selcet_rows,]
```

Simulation

By using tidyverse and R functions and references of several real-world data-sets, we acquired a sample data set for further estimation and analysis.

```
glimpse(sample_dataset)

## Rows: 10,000
## Columns: 6
## $ Age                <dbl> 74.103903, 46.069665, 27.272283, 11.542934, 24...
## $ Gender              <chr> "Female", "Female", "Male", "Female", "Female"...
## $ PoliticalP_supported <chr> "CON", "LIB", "LIB", "LIB", "LIB", "LIB", "GRN...
## $ Edu_level           <chr> "University below bachelor's", "Bachelor or hi...
## $ Income              <dbl> 31417.3809, 383507.5589, 41790.8137, 14467.685...
## $ District            <int> 67, 116, 2, 54, 60, 24, 113, 122, 111, 117, 61...
```

Firstly we filter out people that is under 18, who is not eligible to vote for election.

```
Sample_clean<-sample_dataset %>% filter(Age>=18)
sample_size<-nrow(Sample_clean)
```

Then, we look at the vote share in the remaining voters:

References

Canada, Statistics. n.d.a. “Chart 1 Educational Attainment¹ for the Population Aged 25 to 64, Canada, 2016.” <https://www150.statcan.gc.ca/n1/daily-quotidien/171129/cg-a001-eng.htm>.

———. n.d.b. “Table 11-10-0238-01 Distribution of Market, Total and After-Tax Income of Individuals, Canada, Provinces and Selected Census Metropolitan Areas.” <https://doi.org/https://doi.org/10.25318/1110023801-eng>.

CBCNews. n.d. “Federal Election 2019 Live Results.” <https://doi.org/https://newsinteractives.cbc.ca/elections/federal/2019/results/>.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.

Statistics Canada. n.d. “Table 17-10-0005-01 Population Estimates on July 1st, by Age and Sex.” <https://doi.org/https://doi.org/10.25318/1710000501-eng>.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. “Welcome to the tidyverse.” *Journal of Open Source Software* 4 (43): 1686. <https://doi.org/10.21105/joss.01686>.