Sampling Design and Survey Practice — Lab 2

October 13, 2021

Notice

- Midterm on October 20. Check the notice on eTL.

1 Systematic Sampling

- We will again use the `subway_snu2.csv` data. We wish to estimate μ (average number of passengers getting off at the SNU subway station per day, from March 1 to August 31) using the systematic sampling.
- Recall that $N = 184$, $\mu = 39450.47$ (true value).

```
subway <- read.csv("subway_snu2.csv", header = TRUE)
N <- nrow(subway)
mu <- mean(subway$number)
```

- There were two methods to select a systematic sample:
  - Method A: Randomly select a sampling unit from first k sampling units.
  - Method B: Randomly select a sampling unit from the population.
- Estimation of population mean:

$$\bar{y}_{\text{sy}} = \bar{y}_i,$$

where $i = 1, \cdots, k$. For example, if $i = 1$, $\bar{y}_1$ is the sample mean of the first item, $(1+k)$th item, $(1+2k)$th item, ...

```
set.seed(0)
k <- 4   # N = nk, 184 = 46 * 4
n <- N/k # 46

# Method A
j <- sample(k, 1)
index <- seq(j, N ,k)
y.bar.sy <- mean(subway$number[index])
y.bar.sy
```

```
## [1] 39073.02
```

```
# Method B
j <- sample(N, 1)
index <- seq(j %% k, N, k) # does not matter even if the remainder is 0
y.bar.sy <- mean(subway$number[index])
```

```
y.bar.sy
```

```
## [1] 39810.67
```

- We only consider the case $N = nk$. In this case, two methods are equivalent.
- Assuming $N = nk$, variance of $\bar{y}_{\text{sy}}$:

$$\text{var}(\bar{y}_{\text{sy}}) = \frac{1}{k}\sum_{i=1}^{k}(\bar{y}_i - \mu)^2 \tag{1}$$

$$= \frac{N-1}{N}S^2 - \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2 \tag{2}$$

$$= \frac{S^2}{n}\frac{N-1}{N}[1 + (n-1)\rho], \tag{3}$$

where

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \mu)^2, \quad \text{and}$$

$$\rho = \frac{E[(y_{ij} - \mu)(y_{ik} - \mu)]}{E[(y_{ij} - \mu)^2]}.$$

- Large variation within the systematic samples makes the variance of the estimator smaller (equation (2)). i.e. heterogeneous sample makes the estimation precise.
- Intraclass correlation coefficient ρ measures the homogeneity in a systematic sample, and (3) expresses the variance in terms of ρ.
- Use the following equation to compute ρ:

$$\rho = \frac{2}{n-1}\sum_{i=1}^{k}\sum_{j<k}^{n}[(y_{ij} - \mu)(y_{ik} - \mu)]\frac{1}{(N-1)S^2}$$

```
# y: vector of length N = nk
# k: number of systematic samples
intra.cor <- function(y, k){
  N <- length(y)
  n <- N/k
  mu <- mean(y)
  class <- rep(1:k, N/k)
  rho <- 0
  for(i in 1:k){
    comb.index <- t(combn(which(class==i),2))
    rho <- rho + sum((y-mu)[comb.index[,1]] * (y-mu)[comb.index[,2]])
  }
  rho*2/(n-1)/(N-1)/var(y)
}
```

```
rho <- intra.cor(subway$number, k=4)
rho
```

```
## [1] -0.02015526
```

- Another way to compute ρ is to use the ANOVA model.

$$\rho = 1 - \frac{n}{n-1}\frac{\text{SSW}}{\text{SST}},$$

where SSW and SSB are equivalent to SSE and SStr in the ANOVA model, respectively. Also, SST = SSW + SSB.

```
fit <- anova(lm(subway$number ~ as.factor(rep(1:k,n))))
fit$`Sum Sq` # SSB, SSW
```

```
## [1]    24530504 12107085004
```

```
SSW <- fit$`Sum Sq`[2]
SST <- sum(fit$`Sum Sq`)
1 - n/(n-1)*SSW/SST # = rho
```

```
## [1] -0.02015526
```

- Computing with and without ρ, the variance of $\bar{y}_{\text{sy}}$ is

```
# Eq (1)
y.bar <- tapply(subway$number, rep(1:k, N/k), mean)
sum((y.bar-mu)^2)/k
```

```
## [1] 133318
```

```
# Eq (3)
var(subway$number)/n*(N-1)/N*(1+(n-1)*rho)
```

```
## [1] 133318
```

- In practice, we assume $\rho = 0$ and estimate the variance by

$$\widehat{\text{var}}(\bar{y}_{\text{sy}}) = \frac{N-n}{N}\frac{s^2}{n},$$

where s^2 is the sample variance.

```
est.var.sy <- (N-n)/N*var(subway$number[index])/n
est.var.sy
```

```
## [1] 1022407
```

2 Ratio Estimator

- In the `subway_snu2.csv` data, the column `pass_in` denotes the number of passengers getting on at the SNU subway station per day.

- We wish to estimate the ratio of average number of passengers getting off at the station (y; `number`) to the average number of passengers getting on at the station (x; `pass_in`).
- Using simple random sampling, we estimate the population ratio R by

$$r = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{\bar{y}}{\bar{x}}.$$

```
n <- 30
index <- sample(N, n)
R <- sum(subway$number)/sum(subway$pass_in)
r <- sum(subway$number[index])/sum(subway$pass_in[index])
c(R, r)
```

```
## [1] 0.9830832 0.9911370
```

- The variance and its estimation is given as

$$\mathrm{var}(r) = \frac{1}{\mu_x^2}\frac{N-n}{Nn}\frac{1}{N-1}\sum_{i=1}^{N}(y_i - Rx_i)^2$$

$$\widehat{\mathrm{var}}(r) = \frac{1}{\bar{x}^2}\frac{N-n}{Nn}\underbrace{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - rx_i)^2}_{:=s_r^2}$$

```
var.ratio <- 1/mean(subway$pass_in)^2*(N-n)/(N*n)*
  var(subway$number-R*subway$pass_in)
est.var.ratio <- 1/mean(subway$pass_in[index])^2*(N-n)/(N*n)*
  var(subway$number[index]-r*subway$pass_in[index])
c(var.ratio, est.var.ratio)
```

```
## [1] 9.347845e-06 1.928301e-05
```

- The 95% confidence interval for R is approximately $(0.982, 1.000)$

```
r + c(-2,2)*sqrt(est.var.ratio)
```

```
## [1] 0.9823545 0.9999195
```

- Use packages to estimate the ratio R.

```
library(sampling)
library(survey)
subway.srs <- subway[index,]
subway.srs.des <- svydesign(id = ~1, data = subway.srs, fpc = ~rep(N,n))
res <- svyratio(~number, ~pass_in, design = subway.srs.des)
res
```

```
## Ratio estimator: svyratio.survey.design2(~number, ~pass_in, design = subway.srs.des)
## Ratios=
##         pass_in
## number 0.991137
## SEs=
##            pass_in
## number 0.004391242
```

```
confint(res)
```

```
##                     2.5 %    97.5 %
## number/pass_in 0.9825303 0.9997437
```

3 Regression Estimation

3.1 Regression Analysis (Review)

- Simple linear regression

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad (i = 1, \cdots, n)$$

where $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

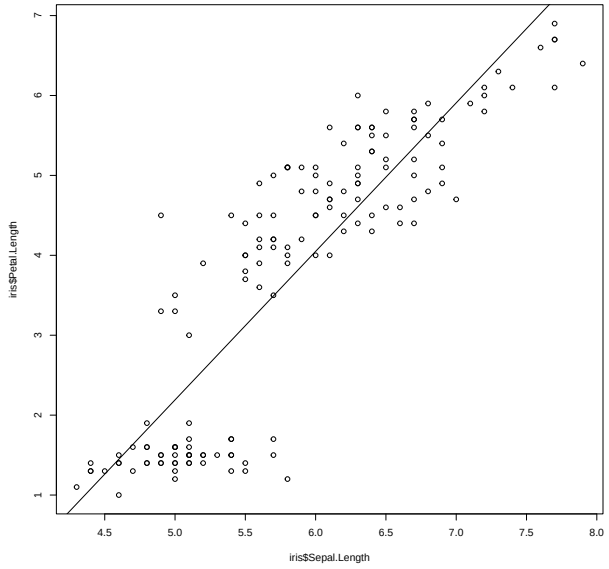
- Suppose we wish to see the relationship between `Petal.Length` (y) and `Sepal.Length` (x) in the `iris` data set.

```
data(iris)
fit <- lm(Petal.Length ~ Sepal.Length, data = iris)
summary(fit)
```

```
Call:
lm(formula = Petal.Length ~ Sepal.Length, data = iris)          => 모형

Residuals:                                                      => 잔차에 대한 요약통계량
     Min       1Q   Median       3Q      Max
-2.47747 -0.59072 -0.00668  0.60484  2.49512

Coefficients:                                                   => 회귀모수의 추정값, 표준오차, 검정통계량, p-val
             Estimate Std. Error t value Pr(>|t|)                  검정통계량 이용해 H1: beta1 =\= 0 검정
(Intercept)  -7.10144    0.50666  -14.02   <2e-16 ***            y.hat = -7.10 + 1.86 * x
Sepal.Length  1.85843    0.08586   21.65   <2e-16 ***            beta1에 대한 T 통계량: 21.65 (H0 기각)
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8678 on 148 degrees of freedom.      => 오차항의 표준편차 sigma 추정값(0.8678)
Multiple R-squared:   0.76, Adjusted R-squared:  0.7583.        => 결정계수(0.76)와 수정결정계수(0.7583)
F-statistic: 468.6 on 1 and 148 DF,  p-value: < 2.2e-16         => 직선관계의 유의성에 대한 F검정: 유의하다
```

- We can draw scatter plot and the fitted regression line.

```
plot(iris$Sepal.Length, iris$Petal.Length)
abline(fit)
```



- Use `plot` function to see the residual plot.

```
plot(fit)
```

- Multiple linear regression works similarly, plus sign (+) denoting the regression on multiple variables.

```
fit2 <- lm(Petal.Length ~ Sepal.Length + Sepal.Width, data = iris)
summary(fit2)
```

3.2 Regression Estimator

- As in the ratio estimation, suppose
  - y: average number of passengers getting off at the station (`number`)
  - x: average number of passengers getting on at the station (`pass_in`)
- It is known that $\mu_x = 40129.33$.

```
mu.x <- mean(subway$pass_in)
mu.x

## [1] 40129.33
```

- We wish to estimate μ_y, using both x and y. Use regression estimator $\hat{\mu}_{yL}$:

$$\hat{\mu}_{yL} = \bar{y} + \hat{\beta}_1(\mu_x - \bar{x}).$$

```r
n <- 30
subway.srs <- subway[sample(N,n),]
fit <- lm(number ~ pass_in, data = subway.srs)
beta1.hat <- coef(fit)[[2]]
beta1.hat
```

```
## [1] 0.9219317
```

```r
mu.hat.yL <- mean(subway.srs$number) + beta1.hat*(mu.x - mean(subway.srs$pass_in))
mu.hat.yL
```

```
## [1] 39592.29
```

- The variance of $\hat{\mu}_{yL}$ is estimated as:

$$\widehat{\text{var}}(\hat{\mu}_{yL}) = \frac{N-n}{Nn}\text{MSE}$$

```r
MSE <- sum(fit$residuals^2)/(n-2)
MSE
```

```
## [1] 491218.4
```

```r
est.var.reg <- (N-n)/(N*n)*MSE
est.var.reg
```

```
## [1] 13704.28
```