Sampling_Project

Giorgia Treglia

2025-04-15

```
setwd("C:/Users/Armando/Downloads/Sampling")
library(TeachingSampling)
library(sampling)
library(dplyr)
data(BigLucy)
```

Question 1: Simple Random Sampling (SRS)

Select a random sample of size 3% of the population (aprox. n = 2562) and estimate the mean income (variable Income) of the population providing a 95% confidence interval (CI) for your calculation.

```
set.seed(123)
N <- nrow(BigLucy)
n <- round(0.03 * N)
srs_indices <- sample(1:N, n)
srs_sample <- BigLucy[srs_indices, ]

mean <- mean(srs_sample$Income)
se <- sd(srs_sample$Income) / sqrt(n)
lower <- mean - 1.96 * se
upper <- mean + 1.96 * se

cat(" Mean:", mean, "\n", "SE:", se, "\n", "95% CI:", lower, "-", upper)
```

```
##  Mean: 422.9281
##  SE: 5.084661
##  95% CI: 412.9622 - 432.894
```

The mean income estimation using the simple random sampling method, which assumes all units in the population have equal probability of being selected, is 422.93, with a standard error of 5.08, and a 95% Confidence Interval range that goes from 412.96 to 432.89. These results indicate that the average income is likely to fall in this range, while the relatively low standard error suggests that the estimate is precise.

Question 2: Systematic Sampling

Select a systematic sample of the same size and give the corresponding results.

```
set.seed(456)
k <- floor(N/n)
print(k)
```

```
## [1] 33
```

```
start <- sample(1:k, 1)
sys_indices <- seq(start, by = k, length.out = n)
sys_sample <- BigLucy[sys_indices, ]

mean_sys <- mean(sys_sample$Income)
se_sys <- sd(sys_sample$Income) / sqrt(n)
lower_sys <- mean_sys - 1.96 * se_sys
upper_sys <- mean_sys + 1.96 * se_sys

cat(" Mean:", mean_sys, "\n", "SE:", se_sys, "\n", "95% CI:",
    lower_sys, "-", upper_sys)
```

```
##  Mean: 431.1887
##  SE: 5.246041
##  95% CI: 420.9065 - 441.471
```

Here, instead of randomly picking each unit, the code selects every k company, where 33 is the step that implements equal spacing, starting from a randomly chosen index between 1 and 33, where $k = N/n$. The mean estimation, using systematic sampling is 431.19, with a standard error of 5.25, while the confidence interval ranges from 420.91 to 441.47. Compared to simple random sampling, the estimate is slightly higher and the confidence interval is a bit wider, suggesting more variability.

Question 3: Stratified Sampling (Proportional Allocation)

The population is separated into three strata, based on variable 'Level'. The separation is in three strata and is made according to declared taxes from the companies. Using this variable select a stratified sample of the same size as in srs, to estimate the mean income and its 95% CI. Implement proportional allocation.

```
set.seed(12)
levels <- unique(BigLucy$Level)
table <- table(BigLucy$Level)
Nh <- round(n * table / sum(table))
print(Nh)
```

```
##
##    Big Medium  Small
##     87    774   1698
```

```r
s_sample <- data.frame()

for (level in names(Nh)) {
  str_data <- BigLucy[BigLucy$Level == level, ]
  str_sample <- str_data[sample(1:nrow(str_data), Nh[level]), ]
  s_sample <- rbind(s_sample, str_sample)
}

mean_s <- mean(s_sample$Income)
se_s <- sd(s_sample$Income) / sqrt(n)
lower_s <- mean_s - 1.96 * se_s
upper_s <- mean_s + 1.96 * se_s

cat(" Mean:", mean_s, "\n", "SE:", se_s, "\n", "95% CI:", lower_s, "-", upper_s)
```

```
##  Mean: 426.4795
##  SE: 5.100918
##  95% CI: 416.4817 - 436.4773
```

Using stratified sampling, the aim is to divide the population into three strata based on the Level variable (Big, Medium and Small), then selected samples from each stratum in proportion to its size in the population. The estimated average income is 426.48, while the standard error is 5.10, and the 95% confidence interval goes from 416.48 to 436.47. The result is slightly higher than simple random sampling, but still quite precise. It is also a valid method because it make sure that company of all sizes are well represented.

Question 4: Probability Proportional to Size (PPS) Sampling

Revise q1 and instead of srs, i.e. equal probabilities, select the sample of size n = 2562 with unequal probabilities and more specifically, probabilities proportional to size, based on variable 'Employees'. Provide again with the estimate of the mean income and its 95% CI.

```r
set.seed(101)
pik <- inclusionprobabilities(BigLucy$Employees, n)
sampled <- UPsystematic(pik)
pps_sample <- BigLucy[sampled == 1, ]
weights <- 1 / pik[sampled == 1]

income <- pps_sample$Income
mean_pps <- sum(income * weights) / sum(weights)
var_pps <- sum(((income - mean_pps)^2) * (weights^2)) / sum(weights)^2
se_pps <- sqrt(var_pps)
lower_pps <- mean_pps - 1.96 * se_pps
upper_pps <- mean_pps + 1.96 * se_pps

cat(" Mean:", mean_pps, "\n", "SE:", se_pps, "\n", "95% CI:",
    lower_pps, "-", upper_pps)
```

```
##  Mean: 431.1527
##  SE: 5.464557
##  95% CI: 420.4421 - 441.8632
```

The average income using PPS sampling is estimated using the Horvitz-Thompson estimator and it is 431.15 with a standard error of 5.46. In this method companies with more employees have a higher chance of being picked, which can give a better picture if income depends on the company size. The confidence interval, from 420.44 to 441.86, shows where the real average income is likely to fall. The estimate is close to the others, but the wider interval means there's a bit more uncertainty in this result.

Question 5: Stratified PPS Sampling

Revise q3 and select the stratified sample according to proportional allocation, but implement sampling proportional to size within each strata. The variable based on which the 'sizes' will be calculated with is again variable 'Employees'. Provide with the new estimate of the mean income and its 95% CI.

```
set.seed(789)
pps_ss <- data.frame()

for (level in names(Nh)) {
  stratum <- BigLucy[BigLucy$Level == level, ]
  pik <- inclusionprobabilities(stratum$Employees, Nh[level])
  selected <- UPsystematic(pik)
  sub_sample <- stratum[selected == 1, ]
  sub_sample$weights <- 1 / pik[selected == 1]
  pps_ss <- rbind(pps_ss, sub_sample)
}

y <- pps_ss$Income
weights <- pps_ss$weights
mean_HT <- sum(y * weights) / sum(weights)
var_HT <- sum(((y - mean_HT)^2) * weights^2) / sum(weights)^2
se_HT <- sqrt(var_HT)
lower_HT <- mean_HT - 1.96 * se_HT
upper_HT <- mean_HT + 1.96 * se_HT

cat(" Mean:", mean_HT, "\n", "SE:", se_HT, "\n", "95% CI:",
    lower_HT, "-", upper_HT)
```

```
##  Mean: 429.8041
##  SE: 5.822787
##  95% CI: 418.3914 - 441.2167
```

Using stratified sampling with PPS inside each group, the estimated average income is 429.80. The 95% CI goes from 418.39 to 441.21, and the standard error is 5.82. This method gives similar results to the others but it divides the companies by size and gives more weight to companies with more employees. This is done by calculating the inclusion probabilities for each stratum using the inclusionprobabilities() function and then the UPsystematic() function to perform systematic sampling within each stratum. This can help make the sample more representative, especially if bigger companies tend to earn more. The confidential interval is a bit wide, so the estimate is good but not the most precise.

Question 6: Auxiliary Variable Techniques (Taxes)

Assume that variable 'Taxes' is available for the population since this is declared by each company. In this question the srs sample and the results obtained in Q1. Should be revised using variable 'Taxes' and in particular the techniques: Ratio, Regression, Post stratification with three strata.

Ratio Estimation

```
r <- sum(srs_sample$Income) / sum(srs_sample$Taxes)
print(r)
```

```
## [1] 37.18193
```

```
mean_r <- r * mean(BigLucy$Taxes)

residuals <- srs_sample$Income - r * srs_sample$Taxes
se_r <- sqrt(sum(residuals^2) / ((n - 1) * mean(srs_sample$Taxes)^2)) / sqrt(n)
lower_r <- mean_r - 1.96 * se_r
upper_r <- mean_r + 1.96 * se_r

cat(" Mean:", mean_r, "\n", "SE:", se_r, "\n", "95% CI:", lower_r, "-", upper_r)
```

```
##  Mean: 439.5895
##  SE: 0.6610301
##  95% CI: 438.2939 - 440.8851
```

The estimated mean income with the Ratio Estimation method is 439.59 with a very small standard error of 0.66. The 95% confidence interval goes from 438.29 to 440.89, which is very narrow. The ratio estimate r, calculated as the sum of Income divided by the sum of Taxes in the sample, is 37.18, representing the relationship between income and taxes in the sample (for every unit of tax, the income is approximately 37.18 times larger). Since the variable Taxes is closely related to Income, the ratio method worked really well in this case.

Regression Estimation

```
model <- lm(Income ~ Taxes, data = srs_sample)
beta1 <- coef(model)[2]
print(beta1)
```

```
##   Taxes
## 14.5559
```

```
mean_reg <- mean(srs_sample$Income) + beta1*(mean(BigLucy$Taxes) - mean(srs_sample$Taxes))

residuals <- model$residuals
se_reg <- sqrt(sum(residuals^2)/(n - 2))/sqrt(n)
```

```r
lower_reg <- mean_reg - 1.96 * se_reg
upper_reg <- mean_reg + 1.96 * se_reg

cat(" Mean:", mean_reg, "\n", "SE:", se_reg, "\n", "95% CI:",
    lower_reg, "-", upper_reg)
```

```
##  Mean: 429.4507
##  SE: 2.047419
##  95% CI: 425.4377 - 433.4636
```

The estimate from regression is 429.45 with a standard error of about 2.05, while the confidence interval goes from 425.44 to 433.46. In this case beta1 is 14.55, meaning that for every unit increase in Taxes, the income is expected to increase by 14.55 units. This method also improves the estimate compared to simple random sampling or other methods by using the relationship between Income and Taxes, though it's not as precise as the ratio estimation.

Post-Stratification

```r
BigLucy$TaxStrata <- cut(BigLucy$Taxes, breaks = c(0, 11, 48, Inf),
                         include.lowest = TRUE)
srs_sample$TaxStrata <- cut(srs_sample$Taxes, breaks = c(0, 11, 48, Inf),
                        include.lowest = TRUE)

Nh_post <- table(BigLucy$TaxStrata)
print(Nh_post)
```

```
##
##   [0,11]  (11,48] (48,Inf]
##    56876    25515     2905
```

```r
nh_post <- table(srs_sample$TaxStrata)
print(nh_post)
```

```
##
##   [0,11]  (11,48] (48,Inf]
##     1717      769       73
```

```r
Wh_post <- Nh_post / sum(Nh_post)
print(Wh_post)
```

```
##
##     [0,11]    (11,48]   (48,Inf]
## 0.66680735 0.29913478 0.03405787
```

```r
means_post <- tapply(srs_sample$Income, srs_sample$TaxStrata, mean)
mean_post <- sum(Wh_post * means_post)
sh2_post <- tapply(srs_sample$Income, srs_sample$TaxStrata, var)
print(sh2_post)
```

```
##   [0,11]  (11,48] (48,Inf]
## 15560.34 16177.51 66694.94
```

```r
se_post <- sqrt(sum((Wh_post^2) * (sh2_post / nh_post)))

lower_post <- mean_post - 1.96 * se_post
upper_post <- mean_post + 1.96 * se_post

cat(" Mean:", mean_post, "\n", "SE:", se_post, "\n", "95% CI:",
    lower_post, "-", upper_post)
```

```
##  Mean: 427.6657
##  SE: 2.640392
##  95% CI: 422.4905 - 432.8409
```

The distribution of companies into tax strata is:

- *[0, 11]*: 56876 companies (sampled 1717)
- (11, 48]: 25515 companies (sampled 769)
- (48, Inf]: 2905 companies (sampled 73)

The mean income in these strata is:

- *[0, 11]*: 15560.34
- (11, 48]: 16177.51
- (48, Inf]: 66694.94

The weights for each stratum,instead, are 0.67 for the first, 0.30 for the second, and 0.03 for the third. The estimate mean of the income with post-stratified method is 427.67, with a standard error of 2.64, while the 95% confidence interval goes from 422.49 to 432.84. This method adjusts the estimate by grouping companies based on their Taxes, and it helps make the sample better match the structure of the population, but still not precise as the ratio estimation.

Question 7: Cluster Sampling (6 Counties)

The country is divided into 99 counties and each company belongs to one county according to its geographical location. Variable 'Zone' is the variable that gives this information. Assume that the researcher has available only this information about the companies. He selects the sample by selecting first an srs sample of 6 counties out of the 99 and then contacts all companies that belong to the selected six counties. Give the sample size of companies for your sample and answer q1 and q2 using this last sample.

```r
set.seed(1234)
zones <- unique(BigLucy$Zone)
selected <- sample(zones, 6)
print(selected)
```

```
## [1] County38 County80 County33 County22 County19 County46
## 100 Levels: County1 County10 County100 County11 County12 County13 ... County99
```

```
cluster <- data.frame()
for (i in 1:6) {
  zone <- selected[i]
  companies <- BigLucy[BigLucy$Zone == zone, ]
  cluster <- rbind(cluster, companies)
}

n_cluster <- nrow(cluster)
mean_c <- mean(cluster$Income)
se_c <- sd(cluster$Income) / sqrt(n_cluster)
lower_c <- mean_c - 1.96 * se_c
upper_c <- mean_c + 1.96 * se_c

cat(" Mean:", mean_c, "\n", "SE:", se_c, "\n", "95% CI:", lower_c, "-", upper_c)
```

```
##  Mean: 413.3296
##  SE: 3.238022
##  95% CI: 406.9831 - 419.6761
```

```
r_cluster <- sum(cluster$Income) / sum(cluster$Taxes)
print(r_cluster)
```

```
## [1] 37.59855
```

```
mean_rc <- r_cluster * mean(BigLucy$Taxes)
residuals_r <- cluster$Income - r_cluster * cluster$Taxes
se_rc <- sqrt(sum(residuals_r^2) / ((nrow(cluster) - 1) *
         mean(cluster$Taxes)^2)) / sqrt(nrow(cluster))

lower_rc <- mean_rc - 1.96 * se_rc
upper_rc <- mean_rc + 1.96 * se_rc

cat(" Mean:", mean_rc, "\n", "SE:", se_rc, "\n", "95% CI:", lower_rc, "-", upper_rc)
```

```
##  Mean: 444.515
##  SE: 0.4829398
##  95% CI: 443.5685 - 445.4616
```

Using cluster sampling by selecting 6 counties (38, 80, 33, 22, 19, 46), the estimated average income is 413.33. The standard error is 3.24, and the 95% confidence interval goes from 406.98 to 419.68. This method gave a slightly lower estimate than the previous ones. Since only companies from a few areas were included, the result might not fully represent the whole population. Still, the confidence interval is narrow, so the estimate is reasonably precise.

Using ratio estimation with the cluster sample, the estimated average income is 444.52. The standard error 0.48 and the 95% confidence interval goes from 443.57 to 445.46. The calculated ratio (Income/Taxes) is 37.60, indicating a strong relationship between the two variables within the selected clusters.

Question 8: Summary Table of Results

Summarize all your results in a table, providing with the estimate and its CI per method/technique for Q1-Q6.

```
methods <- c("Simple Random Sampling", "Systematic Sampling",
             "Stratified Sampling", "PPS Sampling", "Stratified + PPS",
             "Ratio Estimation", "Regression Estimation",
             "Post-Stratification", "Cluster Sampling", "Cluster Sampling - Ratio")

mean <- c(422.93, 431.19, 426.48, 431.15, 429.8, 439.59, 429.45, 427.67, 413.33, 444.52)
se <- c(5.08, 5.25, 5.10, 5.46, 5.82, 0.66, 2.05, 2.64, 3.24, 0.48)
lower <- c(412.96, 420.9, 416.48, 420.44, 418.39, 438.29, 425.44, 422.49, 406.98, 443.56)
upper <- c(432.89, 441.47, 436.47, 441.86, 441.21, 440.89, 433.46, 432.84, 419.68, 445.46)

summary <- data.frame(Method = methods, Mean = mean, Standard_Error = se,
                      Lower_bound = lower, Upper_bound = upper)
print(summary)
```

```
##                      Method   Mean Standard_Error Lower_bound Upper_bound
## 1    Simple Random Sampling 422.93           5.08      412.96      432.89
## 2       Systematic Sampling 431.19           5.25      420.90      441.47
## 3       Stratified Sampling 426.48           5.10      416.48      436.47
## 4              PPS Sampling 431.15           5.46      420.44      441.86
## 5          Stratified + PPS 429.80           5.82      418.39      441.21
## 6          Ratio Estimation 439.59           0.66      438.29      440.89
## 7     Regression Estimation 429.45           2.05      425.44      433.46
## 8       Post-Stratification 427.67           2.64      422.49      432.84
## 9          Cluster Sampling 413.33           3.24      406.98      419.68
## 10 Cluster Sampling - Ratio 444.52           0.48      443.56      445.46
```

This project explored different sampling methods to estimate the average income of industrial companies. The ratio estimation method, using taxes as auxiliary information, provided the most accurate and stable estimate, especially when applied to the cluster sample, with the smallest confidence interval. Stratified, Regression and Post-Stratification methods also improved results by accounting for company differences. On the other hand, cluster sampling alone showed a lower estimate, likely because it only included a small part of the population.