SRS and stratified sampling code

November 16, 2024

0.1 STAT 344 Group Project

[ ]:
```
# load the packages
library(dplyr)
```

```
Attaching package: 'dplyr'


The following objects are masked from 'package:stats':

    filter, lag


The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union


The following objects are masked from 'package:stats':

    filter, lag


The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

[ ]:
```
# read the data
data <- read.csv("Engineering_graduate_salary.csv", header=TRUE)

# set seed
set.seed(1)
```

[ ]:
```
# the head of the dataset
head(data)
```

```
# also the tail, if needed
tail(data)
```

A data.frame: 6 x 34

	ID <int>	Gender <chr>	DOB <chr>	X10percentage <dbl>	X10board <chr>	X12g <int>
1	604399	f	1990-10-22	87.80	cbse	2009
2	988334	m	1990-05-15	57.00	cbse	2010
3	301647	m	1989-08-21	77.33	maharashtra state board,pune	2007
4	582313	m	1991-05-04	84.30	cbse	2009
5	339001	f	1990-10-30	82.00	cbse	2008
6	609356	f	1989-12-02	83.16	icse	2007

A data.frame: 6 x 34

	ID <int>	Gender <chr>	DOB <chr>	X10percentage <dbl>	X10board <chr>	X12g <int>
2993	114364	m	1986-02-08	91.00	0	2003
2994	103174	f	1989-04-17	75.00	0	2005
2995	352811	f	1991-07-22	84.00	state board	2008
2996	287070	m	1988-11-24	91.40	bsemp	2006
2997	317336	m	1988-08-25	88.64	karnataka education board	2006
2998	993701	m	1992-05-27	77.00	state board	2009

[ ]:
```
# some important number
N <- nrow(data)  # Total population size
N
sample_size <- 300
```

2998

[ ]:
```
# SRS
srs_sample <- data %>% sample_n(sample_size)

# Calculate mean salary & se
mean_salary_srs <- mean(srs_sample$Salary, na.rm = TRUE)
se_salary_srs <- sd(srs_sample$Salary, na.rm = TRUE) / sqrt(sample_size)
se_salary_srs_fpc <- se_salary_srs * sqrt((N - sample_size) / (N - 1))



# Calculate proportion of students with A GPA & se
proportion_gpa_srs <- mean(srs_sample$collegeGPA >= 80, na.rm = TRUE)
se_proportion_gpa_srs <- sqrt(proportion_gpa_srs * (1 - proportion_gpa_srs) /␣
 ↪sample_size)
se_proportion_gpa_srs_fpc <- se_proportion_gpa_srs * sqrt((N - sample_size) /␣
 ↪(N - 1))



# results
```

```
cat("Mean Salary:", mean_salary_srs, "\n")
cat("Standard Error of Mean Salary with FPC:", se_salary_srs_fpc, "\n")

cat("Proportion of A GPA", proportion_gpa_srs, "\n")
cat("Standard Error of Proportion of A GPA with FPC:",␣
 ↪se_proportion_gpa_srs_fpc, "\n")
```

```
Mean Salary: 306816.7
Standard Error of Mean Salary with FPC: 10682.97
Proportion of A GPA 0.1133333
Standard Error of Mean Salary with FPC: 10682.97
Proportion of A GPA 0.1133333
Standard Error of Proportion of A GPA with FPC: 0.01736505
```

[ ]:
```
z <- 1.96

# Confidence Interval for Mean Salary (SRS)
ci_salary_srs_lower <- mean_salary_srs - z * se_salary_srs_fpc
ci_salary_srs_upper <- mean_salary_srs + z * se_salary_srs_fpc

# Confidence Interval for Proportion of A GPA (SRS)
ci_prop_gpa_srs_lower <- proportion_gpa_srs - z * se_proportion_gpa_srs_fpc
ci_prop_gpa_srs_upper <- proportion_gpa_srs + z * se_proportion_gpa_srs_fpc

cat("SRS Mean Salary CI: [", ci_salary_srs_lower, ",", ci_salary_srs_upper,␣
 ↪"]\n")
cat("SRS Proportion of A GPA CI: [", ci_prop_gpa_srs_lower, ",",␣
 ↪ci_prop_gpa_srs_upper, "]\n")
```

```
SRS Mean Salary CI: [ 285878.1 , 327755.3 ]
SRS Proportion of A GPA CI: [ 0.07929784 , 0.1473688 ]
SRS Proportion of A GPA CI: [ 0.07929784 , 0.1473688 ]
```

[ ]:
```
# Stratified Sampling by Gender
stratified_sample <- data %>%
  group_by(Gender) %>%
  sample_frac(0.1)  # Adjust fraction for 10%

stratified_stats <- stratified_sample %>%
  group_by(Gender) %>%
  summarise(
    N_h = n(),
    n_h = n(),
    mean_salary_h = mean(Salary, na.rm = TRUE),
    prop_A_gpa_h = mean(collegeGPA >= 80, na.rm = TRUE),
    var_salary_h = var(Salary, na.rm = TRUE),
    var_prop_A_h = prop_A_gpa_h * (1 - prop_A_gpa_h)
  ) %>%
```

```
  ungroup() %>%
  mutate(weight_h = N_h / N)  # Weight of each stratum

# Combined estimates using weights
stratified_mean_salary <- sum(stratified_stats$weight_h *␣
 ↪stratified_stats$mean_salary_h)
stratified_prop_A_gpa <- sum(stratified_stats$weight_h *␣
 ↪stratified_stats$prop_A_gpa_h)

# Calculate SE with FPC for the combined stratified estimates
stratified_se_salary <- sqrt(sum((stratified_stats$weight_h^2) *␣
 ↪(stratified_stats$var_salary_h / stratified_stats$n_h))) * sqrt((N -␣
 ↪nrow(stratified_sample)) / (N - 1))
stratified_se_prop_A <- sqrt(sum((stratified_stats$weight_h^2) *␣
 ↪(stratified_stats$var_prop_A_h / stratified_stats$n_h))) * sqrt((N -␣
 ↪nrow(stratified_sample)) / (N - 1))

cat("Stratified Mean Salary:", stratified_mean_salary, "\n")
cat("Stratified Standard Error of Mean Salary with FPC:", stratified_se_salary,␣
 ↪"\n")

cat("Stratified Proportion of A GPA:", stratified_prop_A_gpa, "\n")
cat("Stratified Standard Error of Proportion of A GPA with FPC:",␣
 ↪stratified_se_prop_A, "\n")
```

```
Stratified Mean Salary: 33183.79
Stratified Standard Error of Mean Salary with FPC: 1762.832
Stratified Proportion of A GPA: 0.01167445
Stratified Standard Error of Mean Salary with FPC: 1762.832
Stratified Proportion of A GPA: 0.01167445
Stratified Standard Error of Proportion of A GPA with FPC: 0.001748675
```

[ ]:
```
# Confidence Interval for Mean Salary (Stratified Sampling)
ci_salary_strat_lower <- stratified_mean_salary - z * stratified_se_salary
ci_salary_strat_upper <- stratified_mean_salary + z * stratified_se_salary

# Confidence Interval for Proportion of A GPA (Stratified Sampling)
ci_prop_gpa_strat_lower <- stratified_prop_A_gpa - z * stratified_se_prop_A
ci_prop_gpa_strat_upper <- stratified_prop_A_gpa + z * stratified_se_prop_A

cat("Stratified Mean Salary CI: [", ci_salary_strat_lower, ",",␣
 ↪ci_salary_strat_upper, "]\n")
cat("Stratified Proportion of A GPA CI: [", ci_prop_gpa_strat_lower, ",",␣
 ↪ci_prop_gpa_strat_upper, "]\n")
```

```
Stratified Mean Salary CI: [ 29728.64 , 36638.94 ]
Stratified Proportion of A GPA CI: [ 0.008247046 , 0.01510185 ]
Stratified Proportion of A GPA CI: [ 0.008247046 , 0.01510185 ]
```