Stratified Random Sampling of Schools

Juda Kaleta

2024-06-28

Our study aims to employ stratified random sampling to select a representative sample of Czech schools at ISCED level 2. The sampling process ensures proportional representation based on school size, determined by the number of students.

Population Definition

The target population consists of Czech schools at ISCED level 2 that meet the following criteria:

- The type of the school is `B10` (normal school).
- The school is fully organized.
- The school had a non-zero number of students in the given year.

```
data <- read_excel(file_path, sheet = "v0324jc1")

population <- data %>%
  filter(
    TYP == "B10",                            # only normal schools
    org == 1,                                # only fully organized schools
    !!sym(col_total_students_count) > 0,     # filter out empty schools
  )
```

The table below provides a summary of the target population of schools based on the given criteria.

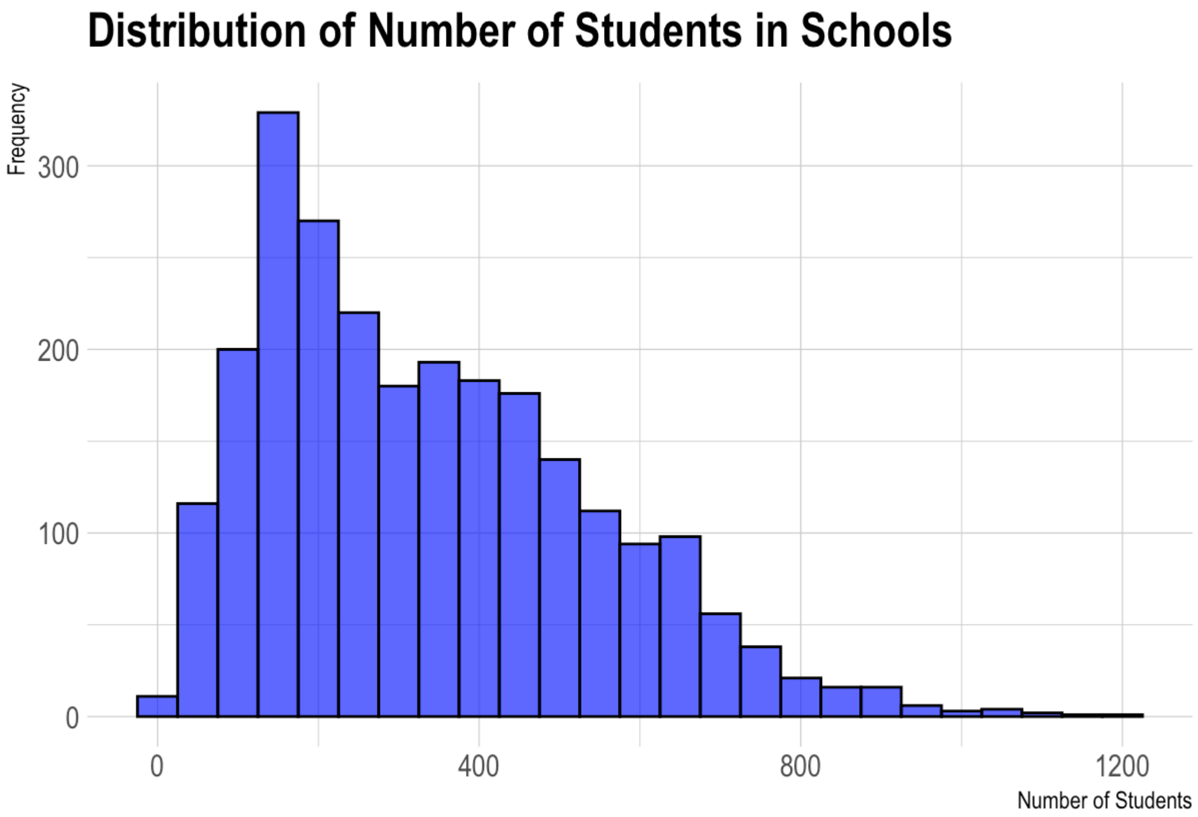
```
summary_population <- population %>%
  summarize(
    total_schools = n(),
    min_students = min(!!sym(col_total_students_count)),
    max_students = max(!!sym(col_total_students_count)),
    mean_students = mean(!!sym(col_total_students_count)),
    median_students = median(!!sym(col_total_students_count))
  )

kable(summary_population, caption = "Summary of the Target Population")
```

Summary of the Target Population

total_schools	min_students	max_students	mean_students	median_students
2486	14	1208	338.3797	302

The histogram below shows the distribution of the number of students in the target population of schools.

```
# Create a histogram of the number of students in schools
ggplot(population, aes(x = !!sym(col_total_students_count))) +
  geom_histogram(binwidth = 50, fill = "blue", color = "black", alpha = 0.7) +
  labs(title = "Distribution of Number of Students in Schools",
       x = "Number of Students",
       y = "Frequency") +
  theme_ipsum()
```



Stratification Criteria

The stratified sample is selected based on school size (number of students). Schools are divided into 3 groups using Jenks natural breaks. This method optimally partitions the data into homogeneous groups, ensuring that each stratum contains schools with similar sizes.

```
jenks_breaks <- classIntervals(population[[col_total_students_count]], n = n_strata,
style = "jenks")
breaks <- jenks_breaks$brks
```

To assign each school to a corresponding stratum based on the Jenks breaks, we create a new column in the population dataset.

```
# Create a new column for strata based on the breaks
population <- population %>%
  mutate(
    strata = cut(
      !!sym(col_total_students_count),
      breaks = breaks,
      include.lowest = TRUE,
      labels = (seq_along(breaks) - 1)[-1]
    )
  )
```

We can then summarize the population within each stratum to understand the distribution of school sizes across the different strata.

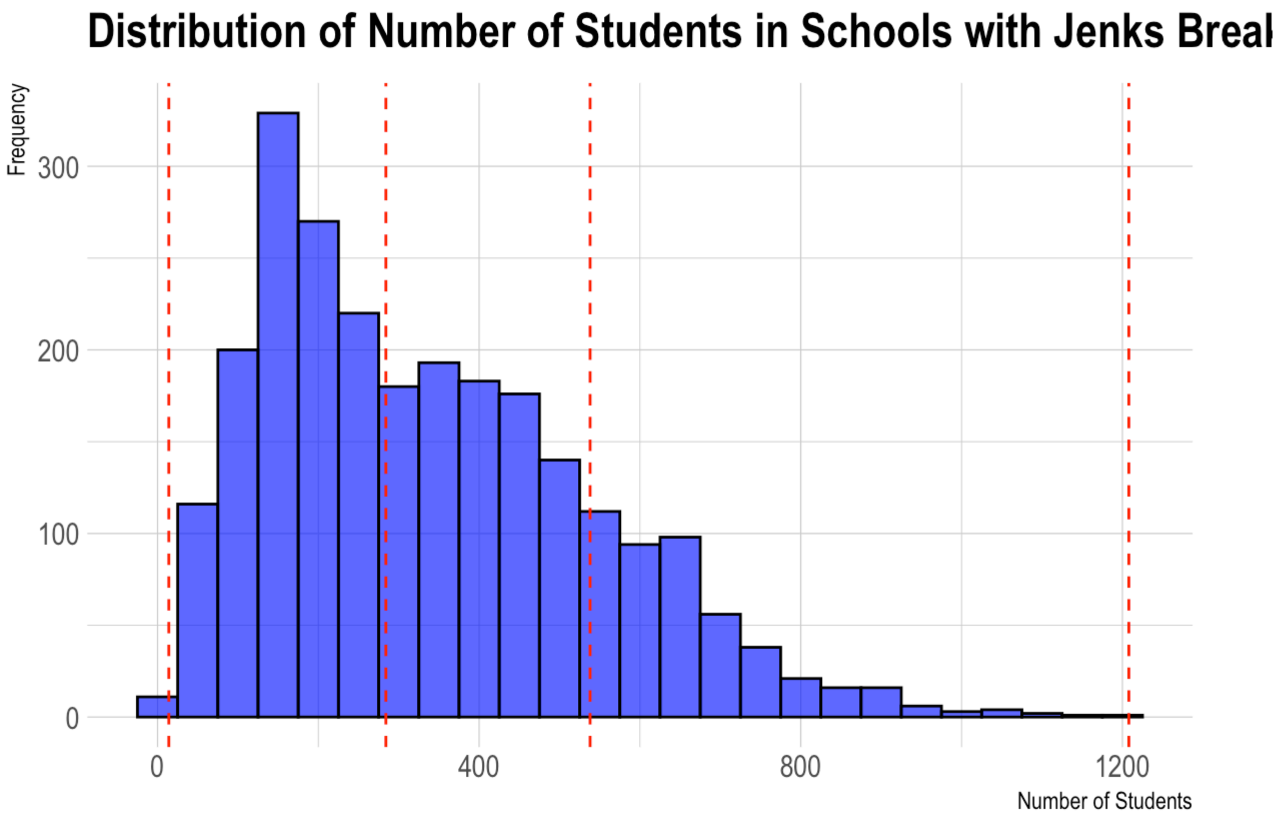
```
summary_population_strata <- population %>%
  group_by(strata) %>%
  summarize(
    total_schools = n(),
    min_students = min(!!sym(col_total_students_count)),
    max_students = max(!!sym(col_total_students_count)),
    mean_students = mean(!!sym(col_total_students_count)),
    median_students = median(!!sym(col_total_students_count))
  )

kable(summary_population_strata, caption = "Summary of the Target Population by Strat
a")
```

Summary of the Target Population by Strata

strata	total_schools	min_students	max_students	mean_students	median_students
1	1182	14	284	165.9010	164.0
2	864	285	538	403.8160	400.5
3	440	539	1208	673.2273	641.0

To visualize the distribution of the number of students in schools and the resulting Jenks breaks, we create a histogram with the break points indicated.

```
ggplot(population, aes(x = !!sym(col_total_students_count))) +
  geom_histogram(binwidth = 50, fill = "blue", color = "black", alpha = 0.7) +
  geom_vline(xintercept = breaks, linetype = "dashed", color = "red") +
  labs(title = "Distribution of Number of Students in Schools with Jenks Breaks",
       x = "Number of Students",
       y = "Frequency") +
  theme_ipsum()
```



Sample Size Determination

To determine the appropriate sample size, we use Cochran's equation. This method provides a way to calculate a statistically valid sample size based on the desired confidence level, margin of error, and estimated proportion of the population.

```
cochran_sample_size <- function(population_size, confidence_level = 0.95, margin_of_e
rror = 0.05, estimated_proportion = 0.5) {
  # Calculate Z-value for the desired confidence level
  z_value <- qnorm(1 - (1 - confidence_level) / 2)

  # Calculate Cochran's sample size
  n_0 <- (z_value^2 * estimated_proportion * (1 - estimated_proportion)) / (margin_of
_error^2)

  # Adjust sample size for finite population
  sample_size <- n_0 / (1 + (n_0 - 1) / population_size)

  return(round(sample_size))
}
```

Next, we calculate the total sample size needed for our study using the function defined above.

```
population_size <- nrow(population)
total_sample_size <- cochran_sample_size(population_size)
```

The total sample size required is 333 schools. We then compute the sample size for each stratum to ensure proportional representation.

```
strata_sample_sizes <- population %>%
  group_by(strata) %>%
  summarize(
    stratum_size = n(),
    stratum_sample_size = round((stratum_size / population_size) * total_sample_size)
  )

population <- population %>%
  left_join(strata_sample_sizes, by = "strata")

kable(strata_sample_sizes, caption = "Sample Sizes for Each Stratum")
```

Sample Sizes for Each Stratum

strata	stratum_size	stratum_sample_size
1	1182	158
2	864	116
3	440	59

Sampling Procedure

This section details the process of randomly selecting schools for each stratum, saving the sampled data, visualizing the results, and validating the sample.

We use stratified random sampling to select schools within each stratum based on the predetermined sample sizes.

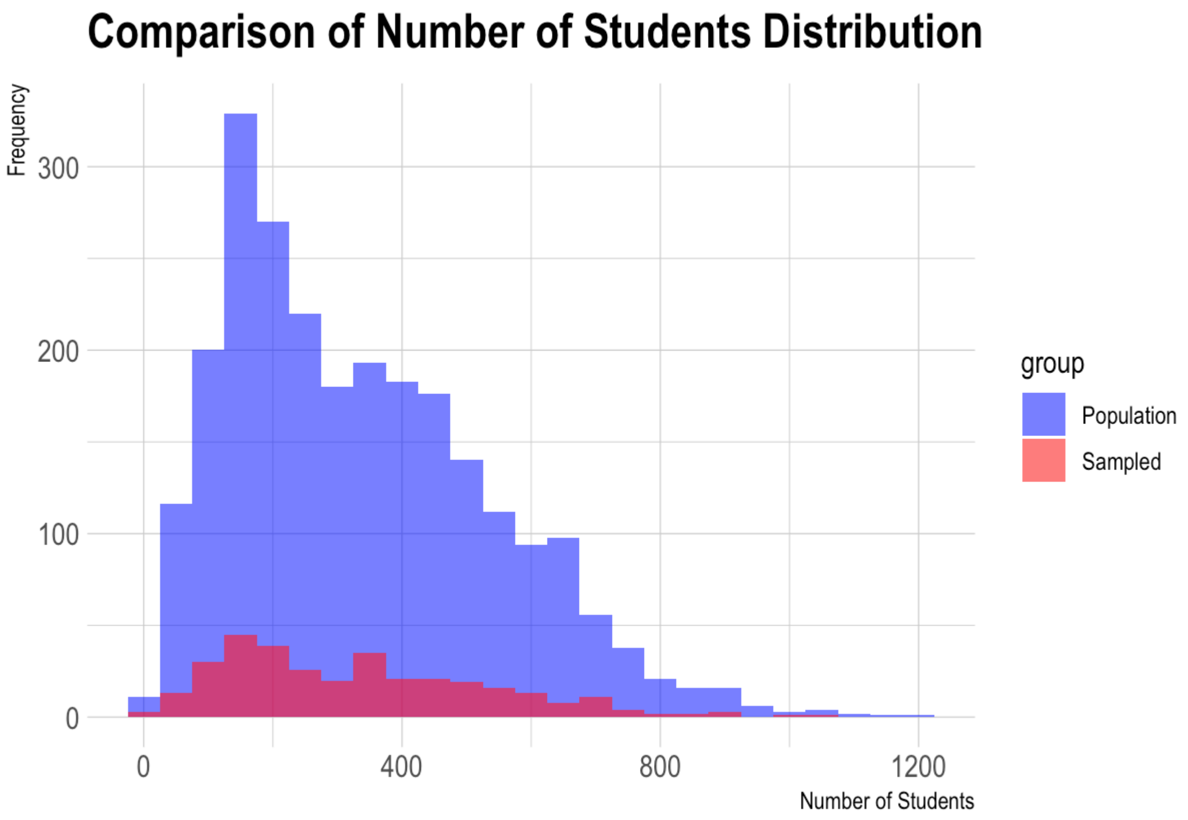
```
set.seed(123)  # For reproducibility
sampled_schools <- population %>%
  group_by(strata) %>%
  group_modify(~ slice_sample(.x, n = min(.x$stratum_sample_size[1], nrow(.x))))
```

We save the sampled schools to an RData file for future analysis and reference.

```
save(sampled_schools, file = "../outcomes/sampled_schools.RData")

# Save the sampled schools to a CSV file, including only specified columns
csv_export_columns <- sampled_schools %>%
  select(strata, izo, zar_naz, ulice, misto)

write.csv(csv_export_columns, file = "../outcomes/sampled_schools.csv", row.names = F
ALSE)
```

We visualize the distribution of the number of students in both the sampled schools and the entire population to compare their distributions.

```
# Combine population and sampled schools for visualization
population$group <- "Population"
sampled_schools$group <- "Sampled"

combined_data <- bind_rows(population, sampled_schools)

# Plot histogram
ggplot(combined_data, aes(x = !!sym(col_total_students_count), fill = group)) +
  geom_histogram(position = "identity", alpha = 0.6, binwidth = 50) +
  labs(title = "Comparison of Number of Students Distribution",
       x = "Number of Students",
       y = "Frequency") +
  scale_fill_manual(values = c("Population" = "blue", "Sampled" = "red")) +
  theme_ipsum()
```


Comparison of Number of Students Distribution
Frequency
Number of Students
group
Population
Sampled