STAT130 Project 2

Xiaohu Ma

University of California Riverside

STAT 130: Sampling Surveys

Jericho Lawson

November 19, 2024

Introduction

In statistics, designing methods for data collection is crucial. Statisticians need to choose the most appropriate sampling method from a variety of options. In this paper, I will apply four different sampling methods including “simple random sampling”, “Stratified sampling”, “One-stage cluster sampling”, “Two-stage cluster sampling” to the “sunroof” dataset for exploring how these different methods influence the representative of the data. The dataset “sunroof” mainly records the information about solar panels currently in use in the United States over the course of a year. It contains 909 observations and 31 variables, which includes region name, total solar energy generation for different directions of roof, number of panels in different directions, carbon offset in tons and so on. Among them, carbon offset in tons is selected as the variable of interest in this paper, and four different sampling methods are applied to measure the total carbon offset in the U.S. by using solar panels in one year. Besides, the variable “state name” will also be used for both strata and cluster when doing stratified and cluster sampling. After designing a sampling plan, analysis will be conducted for finding difference and accuracy of each method through graphs and calculations .

Designs

Simple Random Sampling:

- Sampling frame: The comprehensive list of all data about solar panel installations in the U.S. from U.S. Energy Information Administration (EIA).
- Identifiers: GPS, system ID
- The process to conduct srs in the real life:

    - **Access the data**: Collect the GPS of every solar panel across the U.S. from different sources which may include EIA, state energy agencies or utility companies.
        - To control selection bias, the range of sample selection should be all solar panels in the U.S., but not be a specific state by using convenience sampling.
    - **Define the sample size**: setting the acceptable margin of error, and calculate desired sample size.
    - **Random selection**: assign each collected GPS a number, and randomly pick some sample from those numbers.
        - To ensure randomness, a computer program like R can be used to randomly choose a set of numbers from 1 to the total number of GPS.
    - **Collect data**: collect all the detailed information on each solar panel with selected GPS.

- The process to conduct srs in the code:
    - **Process the data**: load and clean the data, ensuring no Null or NA value exists in the interest column.
    - **Decide the sample size**: Control the sample size is between 20-40% of the population. 35% might be a good choice.
    - **Create a sample**: assign each collected data a number and let R randomly select a set of numbers from 1 to the total number of data, extracting the rows that are represented by these data from the cleaned data.

        - To ensure randomness, using sample() function in R to randomly sample a group of data.

Stratified Sampling:

- Sampling frame: Separated list of solar panel systems within each state. Each list may come from state or local energy agencies.
- Identifiers: state code + GPS of solar panel

- The process to conduct stratified sampling in the real life:
    - **Identify stratification variables**: use states as stratification variables, separating the whole U.S. into 50 strata based on state.
    - **Access the data**: Obtain the list of current GPS of solar panels in each state from state energy agencies.
    - **Define the sample size**: Use proportional allocation to determine the specific numbers sampled from each stratum.
        - Proportional allocation might avoid bias. Because, the amount of sunlight received varies greatly within different states. For example, California receives a lot of sunlight, which means there will be more solar panels and each of them produces a higher carbon offset. If most of the data is sampled from California, its influence would be amplified, leading to sample bias.
    - **data selection:** in each stratum, randomly select the specific number of GPS by using srs.

        - To ensure randomness, a computer program like R can be used to randomly choose a set of numbers from 1 to the total number of GPS in each state.
    - **Collect data**: in each stratum, collect all the detailed information on each solar panel with selected GPS.

- The process to conduct stratified sampling in the code:
    - **Decide the sample size**:
        - Calculate the proportion: number of solar panels in each state/number of solar panels in the whole U.S.
        - Getting number of data sampled from each stratum h: Multiply each states proportion to the designed data size.(35% of the population, which is 318)
    - **Create a sample**: using group_by() and group_modify in package “dplyr” to randomly extract a specific number of data from each state.
        - To ensure randomness, use slice_sample() function in R to randomly sample data from specific stratum.
    - **Summary the sample data**: using summarise() function to make a table showing the information for each state, including its mean and sd of carbon offset. This is convenient for further calculations.

One-stage Cluster Sampling:

- Sampling frame: Separated list of solar panel systems within each state. Each list may come from state or local energy agencies.
- Identifiers: state code + GPS of solar panel

- The process to conduct one-stage cluster sampling in the real life:
    - **Identify cluster variables**: use states as cluster variables, separating the whole U.S. into 50 clusters based on state.
        - To control bias: ensure that clusters are homogeneous within and heterogeneous between.
    - **Define the sample size:** determine how many clusters will be selected from the total 50 clusters.
    - **Cluster selection:** by using SRS, randomly select the number of clusters from 50 states.
        - To ensure randomness, a computer program like R can be used to randomly choose a set of clusters from 1 to the total number of clusters.
    - **Collect data:** collect all the data from selected clusters.

- The process to conduct one-stage cluster in the code:
    - **Decide the sample size**: set the number of cluster sampled be 17, which ensures that total sampled data is between 20-40% of the population data.
    - **Create a sample**:
        - Using sample() function to randomly select 17 states.
            - To ensure randomness, use sample() function in R to randomly sample several clusters.
        - Extract all the rows that contain the name of selected states.

- **Summary the sample data**: using summarise() function to make a table showing the information for each state, including Mi, its mean and total of carbon offset. This is convenient for further calculations.

Two-stage Cluster Sampling:

- Sampling frame: sampled data from each separate list of solar panel systems within each state. Each list may come from state or local energy agencies.

- Identifiers: state code + GPS of solar panel

- The process to conduct two-stage cluster sampling in the real life:
  - **Identify cluster variables**: use states as cluster variables, separating the whole U.S. into 50 clusters based on state.
  - **Define the sample size:** determine how many clusters will be selected from the total 50 clusters. Also, determine how much data will be selected from each selected cluster.
  - **Cluster selection:** by using SRS, randomly select the number of clusters from 50 states.
  - **Sample selection:** using SRS again, randomly select the number of data from each selected cluster.
  - **Collect data: c**ollect the data from selected data in selected clusters.

- The process to conduct two-stage cluster sampling in the code:

- **Decide the sample size**: set the number of cluster sampled be 27, which ensures that total sampled data is between 20-40% of the population data.
- **Create a sample**:
  - Using sample() function to randomly select 27 states.
    - To ensure randomness, use sample() function in R to randomly sample several clusters.
  - Random select 65% of data in each selected cluster by using sample_frac function.
- **Summary the sample data**: using summarise() function to make a table showing the information for each state, including Mi, mi, its mean and total of carbon offset. This is convenient for further calculations.

Calculations

Simple Random Sampling:

- estimating the total number of carbon offset

$$\hat{t} = N\bar{y}$$
$$= 909 * 658266.9$$
$$\approx 598364594$$

- the standard error of estimator

$$SE(t) = \sqrt{N^2 (s^2/n)(1 - (n/N))}$$

$$= \sqrt{(909^2) * (2043964000000^2/318) * (318/909)}$$

$$\approx 58762298$$

- 95% confidence interval of real t

95%CI = $(\hat{t} - t_{\alpha/2, n-1} * SE(t), \hat{t} + t_{\alpha/2, n-1} * SE(t))$

= (598364594 - $t_{0.025, 318-1}$ * 58762298 , 598364594 + $t_{0.025, 318-1}$ * 58762298)

=(482751203, 713977986)

Stratified Sampling:

- estimating the total number of carbon offset

$\hat{t} = \sum_{h=1}^{H} N_h \bar{y}_h$

=(29 * 163536.2743) + (2 * 100166.7021) + (10 * 5976545.2769) + … + (5 * 96000.6889)

≈583360009

- the standard error of estimator

SE(t) = $\sqrt{\sum_{h=1}^{H} (1 - (n_h/N_h)) * (N_h^2) * (s_h^2/n_h)}$

=

$\sqrt{[(1 - (10/29)) * (29^2) * (634174^2/10)] + \ldots + [(1 - (2/5)) * (5^2) * (93838.325^2/2)]}$

≈49971376

- 95% confidence interval of real t

95%CI = $(\hat{t} - t_{\alpha/2, n-H} * SE(t), \hat{t} + t_{\alpha/2, n-H} * SE(t))$

= (583360009 - $t_{0.025, 316-49}$ * 49971376, 583360009 + $t_{0.025, 316-49}$ * 49971376)

≈ (484971936, 681748083)

One-stage Cluster Sampling:

- estimating the total number of carbon offset

$\hat{t}$ = (N/n) * $\sum_{1}^{n} t_i$

= (50/17) * (192197 + 54478519 + 385639 + … + 341438)

≈ 684262789

- the standard error of estimator

$s_t^2 = (1/(n-1)) * \sum(t_i - \hat{t})^2$

= (1/(17 - 1)) * [(192197 - 684262789)^2 + (54478519 - 684262789)^2 + … + (341438 - 684262789)^2]

≈ 4.782228e+17

$SE(t) = \sqrt{N^2 * (1 - (n/N)) * (s_t^2/n)}$

$=\sqrt{50^2 * (1 - (17/50)) * ((4.782228e + 17)/17)}$

≈6812910198

- 95% confidence interval of real t

$95\%CI = (\hat{t} - t_{\alpha/2, n-1} * SE(t), \hat{t} + t_{\alpha/2, n-1} * SE(t))$

= (684262789 -$t_{0.025, 17-1}$ * 6812910198, 684262789 + $t_{0.025, 17-1}$ * 6812910198)

≈ (-13758461642 15126987220)

Two-stage Cluster Sampling:

- estimating the total number of carbon offset

$\hat{t} = (N/n) \sum_{i \in S} (M_i * \bar{y}_i)$

= (50/27) * [(29 * 455263) + (10 * 666618) + … (5 * 64214.04)]

≈518223322

- the standard error of estimator

$s_t^2 = (1/(n-1)) * \sum_{i \in S} (t_i - (\hat{t}/N))^2$

= (1/(27 - 1)) * [(8649995 - (518223322/50))^2 + (3999709 - (518223322/50))^2 + … + (192642.1 - (518223322/50))^2]

≈7.342231e+13

SE(t) =

$$\sqrt{N^2 * (1 - (n/N)) * (s_t^2/n) + (N/n) \sum_{i=1}^{N} (1 - (m_i/M_i)) * (M_i^2) * (s_i^2/m_i)}$$

=

$$\sqrt{50^2 * (1 - (27/50)) * ((7.342231e + 13)/27) + (50/27) * [(1 - (19/29)) * 29^2 * \sqrt{(5.63e11/19) + \ldots + ((1 - 3/5) * 5^2 * (1.221713e + 10)/3)}}$$

≈ 60548676

- 95% confidence interval of real t

95%CI = $(\hat{t} - t_{\alpha/2, n-1} * SE(t), \hat{t} + t_{\alpha/2, n-1} * SE(t))$

= (518223322 -$t_{0.025, 27-1}$ * 60548676, 518223322 + $t_{0.025, 27-1}$ * 60548676)

≈(393763737 642682908)

Analysis

Stratified Sampling:

- Carbon offset between Strata:
    - Minimum: 836
    - Maximum: 19330690
- Distribution of the carbon offset:
    - Between group:
        - The median carbon offset varies significantly between states.
        - Some states, such as Arizona have notably higher carbon offset values (indicated by larger medians) compared to others like Delaware,which have much lower carbon offset.
    - Within group:
        - Spread: for most of the state, the variance of carbon offset is not very high.
        - Outlier: many states show outliers, which indicate unusually high or low value of carbon offsets.

One-stage Cluster Sampling:

- Carbon offset between Cluster:
    - Minimum: 182002
    - Maximum: 58527433
- Distribution of the carbon offset:
    - Between group:

        - California, Florida, Texas have much higher median, reflecting their larger contribution to carbon offset.
        - Variability between each cluster is apparent. Some states(California) have a wider range than others.
    - Within group:
        - For most states, their interquartile ranges are small, which shows that the variance of carbon offset is not very high within each group.

Two-stage Cluster Sampling:

- Carbon offset between Cluster:
    - Minimum: 192642
    - Maximum: 35929352
- Distribution of the carbon offset:
    - Between group:
        - California, Florida, Texas have much higher median, reflecting their larger contribution to carbon offset.
        - Variability between each cluster is apparent. Some states(California) have a wider range than others.
    - Within group:
        - For most states, their interquartile ranges are small, which shows that the variance of carbon offset is not very high within each group.

According to the bar-chart for stratified sampling, one-cluster sampling, and two-stage cluster sampling, we observe that the variance between states is significantly large. This makes sense because the amount of sunlight each state receives differs greatly. For instance, states like

California and Arizona receive much more sunlight compared to other states, which leads to higher solar panel usage and carbon offset contributions.

From the results, stratified sampling produces a smaller standard error (SE) and narrower confidence interval compared to the other methods. This is due to the substantial variance between states. Stratified sampling takes samples proportionally from each state, ensuring that data from every state is included in the analysis. In contrast, with simple random sampling (SRS), data is randomly selected from the entire population. If most of the data happens to come from states with high carbon offsets, such as California or Arizona, the results will likely be overestimated. Similarly, in cluster sampling, a few states are randomly selected as clusters for investigation. This introduces bias because if states like California or Arizona, which receive abundant sunlight, are selected, the estimated total carbon offset will be overestimated when applied to the entire population.

Therefore, when there is significant variance between groups (in this case, states), stratified sampling is likely the best method as it accounts for differences across all groups proportionally.

Reference

Boysen, J. (2017). *Google Project Sunroof*. Kaggle.com. https://www.kaggle.com/datasets/jboysen/google-project-sunroof

Alexander. (2015, July 28). *Identifier Variables*. Statistics How To. https://www.statisticshowto.com/identifier-variables/

Research. (2024, September 9). SEIA. https://seia.org/research/