**Statistical Modeling**

**2024/2025/2**

**Simulation and Estimation Theory Homework**

|  |  |
| --- | --- |
| Name | Neptun |
| Adorjáni Jonathán | I4MJ4U |
| Medvegy Gábor | SZHNB1 |

**1. Description of the dataset**

In this analysis, we will analyze a dataset on California housing data. Firstly, we will extract 500 IID (independent and identically distributed), 500 SR (simple random), and 500 PS (proportionally stratified) samples of size 100, after selecting a quantitative variable. For each sample, we will calculate the mean and compare the theoretical and empirical MSE (mean squared error) across all three sampling methods. Finally, we will decide which method provides the most accurate estimation and what the results we got implicate.

Before we begin the analysis, it is necessary to thoroughly understand and describe the variables of our dataset to present a meaningful and precise analysis on the topic. It contains 10 variables and 20,640 observations. Data is about longitude, latitude, housing median age, total rooms, total bedrooms, population, households, median income, median house value, and ocean proximity.

|  |  |  |
| --- | --- | --- |
| Variable | Description | Measure of Scale |
| Longitude | Geographic coordinate representing the east-west position of a point. | Interval: Geographical measurement in degrees, with equal intervals but no true zero |
| Latitude | Geographic coordinate representing the north-south position of a point. | Interval: Geographical measurement in degrees, with equal intervals but no true zero |
| Housing Median Age | The median age of the housing units in the area | Ratio: Age starts from zero (new house), and differences in age are meaningful (e.g., a house 10-year-old vs 20). |
| Total Rooms | The total number of rooms in all housing units within a geographical area. | Ratio: The count of rooms starts at zero (no rooms), and the differences are meaningful (e.g., 100 rooms vs. 200 rooms) |
| Total Bedrooms | The total number of bedrooms in all housing units within a geographical area. | Ratio: Just like total rooms, this variable represents a count with a true zero point. |
| Population | The total number of people living in the area. | Ratio: Population counts start at zero (no people), and the differences are meaningful. |
| Households | The total number of households in the area. | Ratio: The count of households starts from zero, and differences are meaningful. |
| Median Income | The median income of households in the area, expressed in tens of thousands of dollars. | Ratio: Income has a true zero point (no income), and comparisons (e.g., double income) are meaningful. |
| Median House Value | The median market value of houses in the area. | Ratio: House values start from zero, and ratios or differences are meaningful (e.g., $100,000 is half as valuable as $200,000). |
| Ocean Proximity | The categorical representation of how close the area is to the ocean (e.g., “near bay”) | Nominal: This is a qualitative variable with categories that have no intrinsic ordering or measurable distance between them. |

The quantitative variable we chose to analyze is the median house value. This is because we assume that all the other variables are the ones that are influencing these values.

**2. Explaining the sampling process in R**

Firstly, we will need to install readxl external package to read in our dataset, which is in Excel format, and transform it into an R dataframe. Before we start the sampling process, we must set a seed of our choice to ensure the randomness of the sampling. After setting the seed, defining the number of samples and their sizes, we can begin the IID sampling (sampling with replacement). Firstly, we take one IID sample with the required size. Then with a for loop we generate the remaining number of samples, which will be binded to the first sample. The result is a list, which we converted into a dataframe. It is very important that inside the loop the seed must be incremented, otherwise we would get the same observations in each sample.

After completing the IID sampling, we moved on to the SR sampling. The sampling process itself is nearly the same as in case of the IID sampling. However, one very important distinction must be made. Since SR sampling is the equivalent of sampling WITHOUT replacement, we must set the replace parameter of the sample function to false.

**Finally, we moved to the PS sampling. Stratified sampling means that we divide the population into groups (strata) based on a certain characteristic. Proportional stratification means that the number of samples drawn from each group is proportional to its size in the population. We stratified the dataset based on our only qualitative variable, ocean proximity. This is because we follow the number one rule of real estate: “Location is key”. We assumed that the ocean proximity significantly influences the value of the chosen quantitative variable, median house value. To implement this type of sampling, we must count the number of houses in each group first, and then calculate their respective proportions in the population. Multiplying these values sample size)(gets us the number of houses needed in each sample from each group. After ensuring that the number of houses will be whole numbers, we initialize an empty matrix to store the samples. Then we generated the 500 separate samples, each representing a proportionally stratified sample of the population. Again, for reproducibility, we must increment the value of our seed in the first for loop generating the required number of samples, at each iteration. Then by using a second for loop inside, we iterated over all groups. For each group, we filtered the dataset to only include houses inside that group, randomly selected the required number of houses WITHOUT replacement, and stored the median house values in a vector. After processing all the groups, we stored these sample values as one row in the matrix we initialized before. Finally, we converted the result into a dataframe for further analysis.**

**3. Calculating estimators**

After fully completing the sampling process, we moved on to calculate estimators in our samples. We started with calculating the means in each sample, in all three dataframes storing the sampled data generated by the different sampling methods. We defined a new column in all the dataframes storing the sample means, then we applied the built-in function to calculate the specific values by applying this function row-wise.

We also calculated the population mean and population standard deviation using the built-in R functions on the population data of the chosen quantitative variable (the median house value).

**4. Calculating the Mean Squared Error**

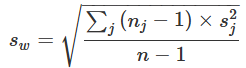
When calculating the Mean Squared Error of the mean estimation for all three samples, we utilized three different approaches, as asked for in the exercises:

* computing the theoretical Mean Squared Error
* computing the empirical Mean Squared Error defined as the variance of sample means around the population mean
* computing the empirical Mean Squared Error with estimating Bias and empirical Standard Error

***4.1. Calculating the theoretical Mean Squared Error***

The formula for the Mean Squared Error is MSE = SE2 + Bs2, where SE is Standard Error and Bs is Bias. Since in this case we are talking about the Mean Squared Error of Mean calculation, and since Mean (in theory) is an unbiased estimator, we can just assume that MSE = SE2, and then compute MSE using the formula of the theoretical Standard Error, as is SE = σ/√n in case of IID sampling, where σ means the standard deviation (in this case, the exercise lets use the previously computed population standard deviation) and n is the size of the sample (we have samples of 100 elements in all three samples of different sampling methods). Of course, we use SE2 here, so the formula we calculate with is σ2/n.

In case of Simple Random sampling, we should have a bit more precise sample than in case of IID since we do not take repetitions, and we incorporate this fact into the formula for Standard Error as well, by simply multiplying the SE formula of the IID sampling with a correction factor: √(1-n/N), where n is the sample size and N is the population size. Of course, when on the power, the square root disappears, so in R we calculated the theoretical MSE of the SR sampling as the MSE of IID sample multiplied by the 1-n/N correction factor.

The hardest task was by far to calculate the theoretical Mean Squared Error of the Proportionally Stratified sampling method, since here we had to use the within-strata standard deviation in the formula (which otherwise remained the same as in case of the SR sampling), as the between strata standard deviation is already taken care of due to how this sampling works. In order to calculate the within strata standard deviation we created a helper table, where we aggregated the mean and standard deviation of the median house value (our chosen quantitative variable) by ocean proximity (our chosen qualitative variable), while also noting how many of each observation type we have in the population data. Then we simply calculated the within-strata standard deviation using the formula we learnt, combining the variances of multiple groups, weighted by their sizes in the original population data: .

***4.2. Calculating the empirical Mean Squared Error, defined as the variance of sample means around the population mean***

In this task, we basically did the same calculation for all three sampling methods: we took the mean of the squared differences between each sample mean and the true population mean, as per the definition of the empirical Mean Squared Error in the task description. The value we got shows how much, on average, the sample means deviate from the population mean.

***4.3. Calculating the empirical Mean Squared Error using Bias and Standard Error***

For the last step in our analysis, we calculated the empirical MSE using the formula we already talked about: MSE = SE2 + Bs2, essentially decomposing the Mean Squared Error into the sum of the squares of Standard Error and Bias. In this case we also calculated everything the same way for the three different sampling methods, so we will explain our line of thought just once. As Standard Error is defined as the standard deviation of the sample means, we can simply use the built-in R function to calculate it – we need to use this one, and not a classical variance formula, since we want to have an unbiased estimate for our SE. As for Bias, we calculated how far our estimator of the mean is from the true population mean, by taking the mean of the means and subtracting from it the population mean. Then we just substituted our results into the formula of MSE, and we had the results.

**5. Analysis of the results**

Across our analysis, we calculated two versions of empirical Mean Squared Error: one version was based on the definition of MSE (variance of the sample means around the population mean), and the other one was based on decomposing MSE into Bias and Standard Error. The two versions differ very slightly only due to the calculation method we used: in the first case, the calculation’s basis is the classical variance, while in the second case, we used the built-in sd function – which uses the corrected variance. Had we used the classic variance in the SE calculation in the 2nd case as well, we would have got the same results for both calculations in the end.

1. *Difference between the empirical and theoretical standard errors*

The empirical and the theoretical standard errors are very close to each other – they differ by only at most about 1.5%. This means that our estimation is quite accurate, so we probably did a good job in sampling, and also the data we based our sampling on was of good quality. This is very little variation, due to the fact that we had many large samples (500 of size 100). We would probably never reach the exact same result – only in the theoretical case of generating every possible sample, which is almost impossible in case of IID sampling (due to repetitions).

1. *Which sampling method provides the most accurate estimates?*

We learned on classes that most of the times the order of efficiency in case of SE is: PS > SR > IID, as the SE\_PS < SE\_SR < SE\_IID. This was also the case in our analysis, as shown by the relative efficiencies we calculated. For example, in case of the relative efficiency when comparing Simple Random to IID (calculated using SE\_SR^2/SE\_IID^2), the result was about 98.6%, meaning that SR sampling reduced the standard error compared to IID sampling, although not that much – this is since the correction factor was nearly one, due to the fact that our selection ratio is 100/20640, so practically almost zero. However, in case of PS sampling, the relative efficiency when comparing it to SR sampling is about 72.7% - which means stratification made sense, since it reduces the Standard Error to 85.3% of the SR Standard Error.

1. *How do the standard errors compare to the theoretical standard error of IID sampling?*

Comparison of empirical and theoretical Standard Error of IID sampling indicates that the empirical SE is larger than the theoretical, although minimally (by about 1%) – this means that the sample estimates vary by a little bit smaller extent than what the mathematical theory expects them to do. We would also like to note, however, that we believe this difference is not substantial and is only caused by sampling.

When comparing the empirical SE of SR sampling to the theoretical SE of IID sampling, we can observe that they are almost the same. However, this can be considered an “accident”, since this is the result of the specific properties of our samples: our SR sampling reduces its empirical SE very little compared to the empirical SE\_IID, while the latter was only a bit higher than its theoretical form – resulting that they become almost identical.

The difference between empirical SE\_PS and theoretical SE\_IID only highlights the fact that PS sampling is much more efficient than IID sampling – as we expected and already explained before.

1. *Explain the extent of estimation differences between the SR and PS sampling methods.*

While SR sampling already provided quite accurate estimation, stratification still made sense because it reduced the Standard Error by almost 15%. By doing the stratification process we significantly reduced the error and improved efficiency as well. When using PS sampling, we have to consider only the within-strata variances since the sampling method already explains the between-strata variances – thus the smaller Standard Error and Mean Squared Error.