Sasha Annan

Stat 422 – Final Project Report

This project aims at estimating the average house prices of a recent real estate listings in San Luis Obispo County and around it using simple random sampling and ratio estimation. The houses dataset comprises of information like multiple listing service number for the house, city where the house is located, the most recent listing price of the house, number of bedrooms and bathrooms, size of the house in square feet and the price of the house per square foot as listed on the on the website **https://wiki.csc.calpoly.edu/datasets/wiki/Houses .**

To estimate the mean house price, random sample of 30 house price data were obtained from the population of 782 house prices using the table of random numbers. A frame which is a list of sampling units used is the list of house prices and the sampling units is the list of the different house prices available.

N = 781 house prices

n = 30 sampled house prices

Sample variance, $S^2 = (792112.1667)^2$

Simple Random Sampling

To estimate the ***population total***

$$\hat{\tau} = N\bar{y} = \frac{N\sum_{i=1}^{n} y_i}{n} = \frac{781(18366000)}{30} = 478{,}128{,}200$$

$$\hat{V}\left(\hat{\tau}\right) = N^2\left(1-\frac{n}{N}\right)\frac{s^2}{n} = (781)^2\left(1-\frac{30}{781}\right)\frac{(792112.1677)^2}{30} = 1.226713332E16 = 1226713332000000$$

This enables us to determine the bound on the error of estimation which is given by;

$$B = 2\sqrt{\hat{V}\left(\bar{y}\right)} = 2\sqrt{1226713332000000} = 2\left(110757091.5\right) = 221{,}514{,}183$$

Therefore, the estimate of the total house price in San Luis Obispo County and around it is 478,128,200. However, the 95% confident interval approximation of the total house price is

$$478128200 \pm 221514183 \text{ or } (256614017, 699642383)$$

To estimate the ***population mean;***

$$\hat{\mu} = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{18366000}{30} = 612200$$

To estimate the ***variance***;

$$\hat{V}\left(\bar{y}\right) = \left(1 - \frac{n}{N}\right)\frac{s^2}{n} = \left(1 - \frac{30}{781}\right)\frac{\left(792112.1677\right)^2}{30} = 2.011134043E10 = 20111340430$$

This enables us to determine the ***bound on the error of estimation*** which is given by;

$$B = 2\sqrt{\hat{V}\left(\bar{y}\right)} = 2\sqrt{20111340430} = 2\left(141814.4578\right) = 283628.9155$$

Therefore the estimate of the mean or average house prices in San Luis Obispo County and around it is 612,200 with an error of estimation less than 283,628.92.

Ratio Estimation

The ration estimator is an alternative method that can be used to estimate the total and average house price since the prices of these houses in the San Luis Obispo County and around it may be determined by factors like the size of the house, number of bedrooms and bathrooms. As a result we can estimate the price of the house, y by making use the size of the house in square feet as a subsidiary variable x.

τ;

To estimate the ***ratio estimator*** of the ***<u>population total</u>***

The total house price $\hat{\tau}_y$

$$\tau_x = 1370701$$

$$r = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{18366000}{57094} = 321.6800364$$

$$\hat{\tau}_y = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} \tau_x = r\tau_x = (321.6800364)(1370701) = 440927147.6$$

τ;

This enables us to determine the bound on the error of estimation which is given by;

$$B = 2\sqrt{\hat{V}\left(\bar{y}\right)} = 2\sqrt{16430305870} = 2\left(128180.7547\right) = 256361.5094$$

Therefore the estimate of the total house price in San Luis Obispo County and around it is 440,927,147.6. However, the 95% confident interval estimate is $440927147.6 \pm 256361.5094$

To estimate the ***ratio estimator*** of the ***population mean***;

We can use $\bar{x}$ to approximate μ_x therefore $\mu_x = \bar{x} = 1903.133333$

$$\hat{\mu}_y = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}(\mu_x) = (321.6800364)(1903.133333) = 612199.9998$$

To estimate the ***variance*** of

s_r is the standard deviation of the deviations

$$\hat{V}\left(\hat{\mu}_y\right) = \left(1 - \frac{n}{N}\right)\frac{s^2{}_r}{n} = \left(1 - \frac{30}{781}\right)\frac{(702074.9078)^2}{30} = 1.579991817E10 = 15799918170$$

This produces a bound on the error of estimation which is given by;

$$B = 2\sqrt{\hat{V}\left(\hat{\mu}_y\right)} = 2\sqrt{15799918017} = 2(125694.7897) = 251389.5795$$

Compare Bounds for simple random sample and ratio estimation

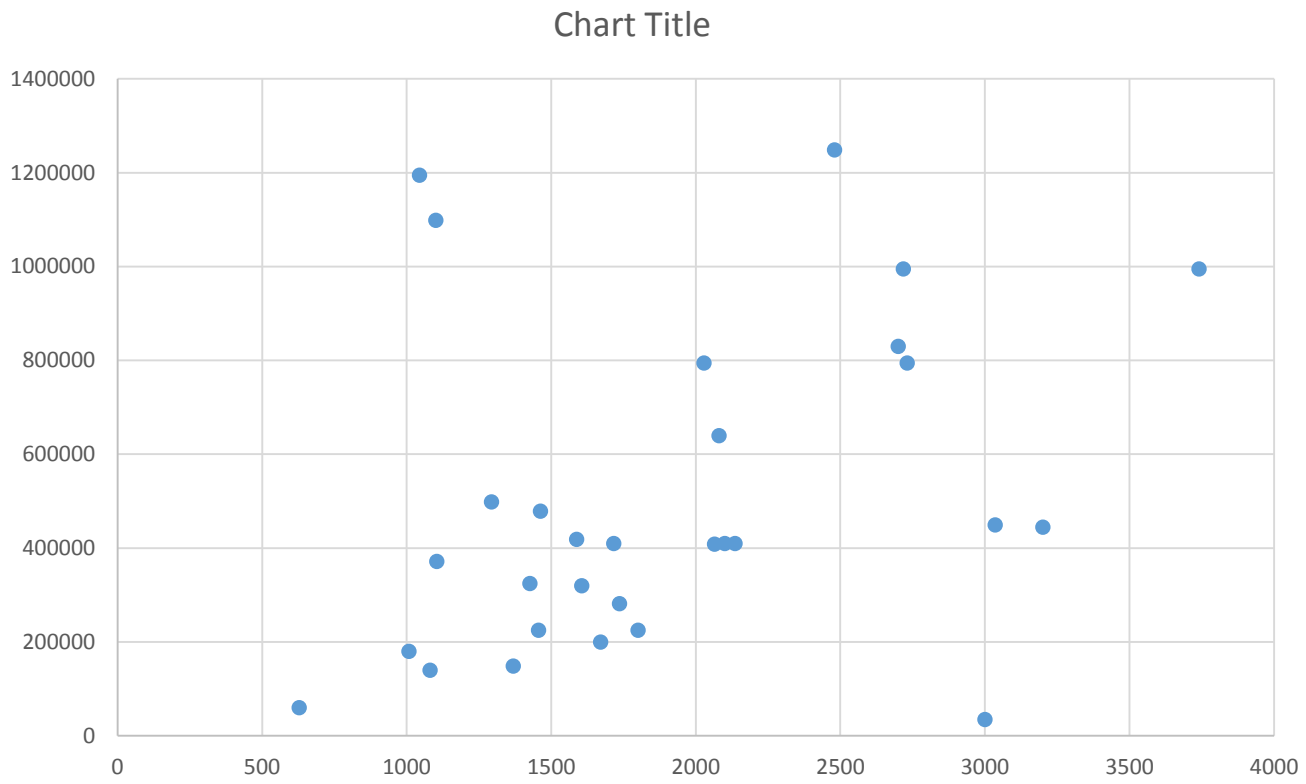
The results of the bounds for the simple random sampling and the ratio estimation shows that using a ratio estimation yield a lower bound. This may be as a result of using a subsidiary variable size in addition to the prices of the houses that was used in the simple random sampling method.

Relative efficiency of ratio estimation to simple random sampling;

$$\hat{RE}\left(\frac{\hat{\mu}}{\hat{\mu}_y}\right)=\left(\frac{283628.9155}{251389.5795}\right)^2=(1.12824452)^2=1.272935697=1.3$$

The relative efficiency results show that using the ratio estimation method is slightly better than simple random sampling. Therefore we need to use just little data to get the same variance estimate if simple random sampling in used instead of ratio estimation.

Scatterplot of House Price, *y* versus Size, *x* in square feet



Attached is the dataset of the sampled house prices and a scatterplot of the house prices versus the size. This graph shows that there is positive linear trend although there are some few unusual data points that are far away from the rest.