**Sample Selection from Real Estate Data from Ames, Iowa (2006-2010)**

*Sudip Bhattacharyya and Daniel Freeman*

April 4, 2018

**Task 1**

**1.1 Objective**

Statistical sampling is a scientific method of selecting a subset of items from a population to study the characteristics of the selected items to get tangible insights on the characteristics of the population from which the sample was drawn. The goal of this project is to draw a sufficiently sized sample based on various sample designs in order to obtain information about population characteristics. Here we use simple random sampling, stratified sampling and two-stage designs to produce estimates of population parameters. The derived estimates from various methods are compared to assess their performance in representing the entire population.

**1.2 Data**

We work with a real estate dataset sourced from Kaggle (<https://www.kaggle.com/c/fi-ames-housing-price-competition/data>). This dataset provides the house sale prices in 25 different neighborhoods of Ames, Iowa between 2006 and 2010 along with several other aspects such as zone, character of the lot, house style, age, foundation, house features, garage quality, sale type, among others. The original raw dataset contains 81 variables, both categorical and continuous, across 1,460 unique sales.

For sampling purposes, we set our focus on seven fields to derive an estimate for average sale prices of houses in Ames. The attributes considered in our sampling exercise are as follows: neighborhood, building type, house style, number of rooms, gross living area, year of sale and above all sale price. For this data, we have:

Population size (N) = 1,460, Population parameter of interest = Average sale price

Population average sale price (µ) = 180,921 Population standard deviation of sale price(σ) = 79,415

Population distribution of sale prices are displayed in Fig. 1.

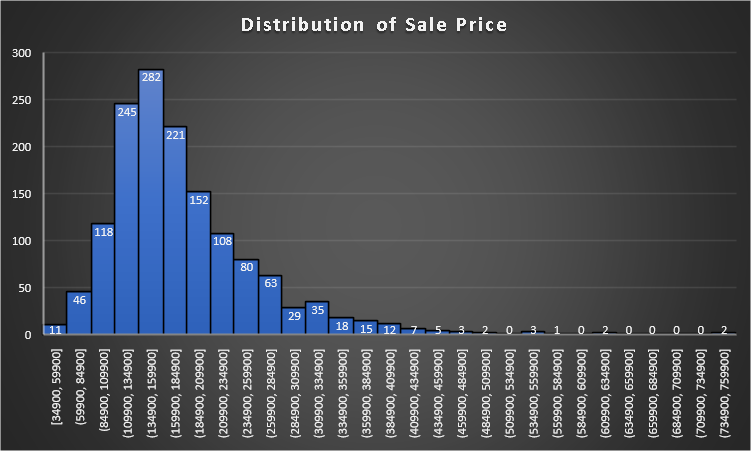


Fig. 1 – Population Distribution of House Sale Prices in Ames, Iowa

In our population, we observe a range of sale prices from $35,000 to above $ 750,000 but more than 50% of houses cost $110,000 to $185,000. The above graph indicates a deviation from a normal distribution for sale prices due to an extreme right skewness which indicates the presence of a few rather expensive houses. By way of the Central Limit Theorem and considering a sufficiently large sample, despite the observed skewness in our data, we safely assume normality in the population distribution.

**1.3 Sample Size**

To determine an optimal sample size for our design, we set a margin of error for our estimates at $10,000, approximately 5% of the average sale price in the population. Assuming normality in our sampling distribution, as stated above, we arrive at a sample size for simple random sampling with 95% confidence as described below.

Initial sample size (n0,srs) = (z95%)2 \* s2 / (I95%)2 = (1.96)2\*(79415)2/(10000)2 = 242

where, z95% = critical value for standard normal distribution at 95% level of significance

Final sample size (nsrs) = n0,srs / (1 + n0,srs /N) = 242 / (1+ 242/1460) = 208

Even though this optimal sample size largely depends on the complexity of the sample design, we continue to use 208 as the size of the sample for all the designs considered in this exercise.

**1.4 Sample Designs**

Four different sample designs – simple random sampling (SRS), stratified sampling with proportional allocation, stratified sampling with Neyman allocation and two-stage sampling design are chosen to estimate the average sale price of houses sold.

**1.4.1 Simple Random Sampling (SRS)**

Simple random sampling is a sampling design in which each sample unit is chosen randomly so that at any stage all the population units have the same probability of being chosen in the sample.

Using SAS for a simple random sampling design we get the output below (Fig. 2 & Fig. 3).

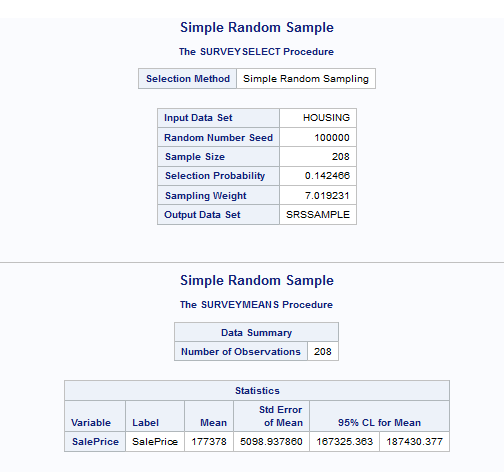


Fig. 2 – SRS output

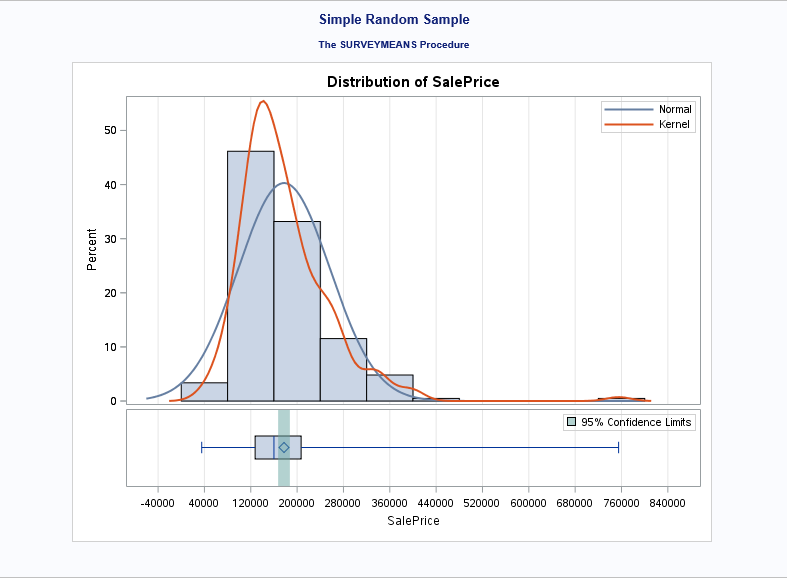
****

Fig. 3 – SRS output

From the above output, we get estimates as follows:

Sample mean (x-bar) = 177,378

SE of sample mean (s) = 5,098.94

95% CI for sample mean = [167,325, 187,430]

**1.4.2 Stratified Sampling with Proportional Allocation**

Stratified sampling is a technique followed when the population has several homogeneous subgroups, known as strata, and the sample is drawn in such a way that each stratum is representative of the population. A simple random sampling method is used to select sample units from each stratum. For a proportional allocation, sample units from each stratum maintain the same ratio of the number of units in each stratum in the population.

In this project, the variable *HouseStyle* has been used for stratification which produces 8 strata as shown in the table below (Fig. 4).

|  |  |  |  |
| --- | --- | --- | --- |
| **Stratum** | **Count** | **Average *SalePrice*** | **StdDev of *SalePrice*** |
| 1.5Fin | 154 | 143,117 | 54,278 |
| 1.5Unf | 14 | 110,150 | 19,036 |
| 1Story | 726 | 175,985 | 77,056 |
| 2.5Fin | 8 | 220,000 | 118,212 |
| 2.5Unf | 11 | 157,355 | 63,934 |
| 2Story | 445 | 210,052 | 87,339 |
| SFoyer | 37 | 135,074 | 30,481 |
| SLvl | 65 | 166,703 | 38,305 |
| **Grand Total** | **1,460** | **180,921** | **79,443** |

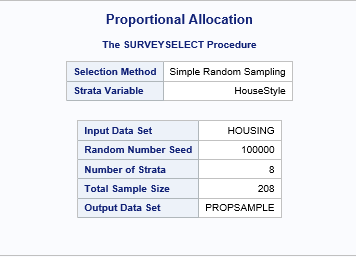
Fig. 4 – Strata Count

Drawing a sample of size 208 with proportional allocation, we calculate the sample sizes for each stratum in the table below (Fig. 5).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Stratum** | **House Style** | **No. of Houses (Nh)** | **Proportion (ph)** | **Sample Size (nh)** |
| Stratum 1 | 1 Story | 726 | 0.4973 | 103 |
| Stratum 2 | 1.5 Fin | 154 | 0.1055 | 22 |
| Stratum 3 | 1.5 Unf | 14 | 0.0096 | 2 |
| Stratum 4 | 2 Story | 445 | 0.3048 | 64 |
| Stratum 5 | 2.5 Fin | 8 | 0.0055 | 1 |
| Stratum 6 | 2.5 Unf | 11 | 0.0075 | 2 |
| Stratum 7 | Sfoyer | 37 | 0.0253 | 5 |
| Stratum 8 | SLvl | 65 | 0.0445 | 9 |
| **Total** |  | **1,460** |  | **208** |

Fig. 5 – Strata Sample Sizes (Proportional Allocation)

SAS produces the results for stratification with proportional allocation as displayed below (Fig. 6 & Fig. 7).

****

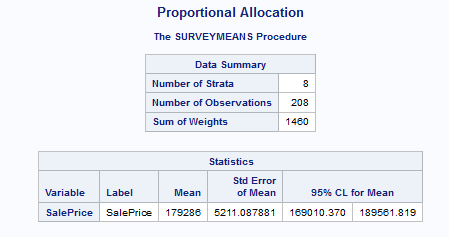


Fig. 6 – Stratified Sampling (Proportional Allocation) SAS Output

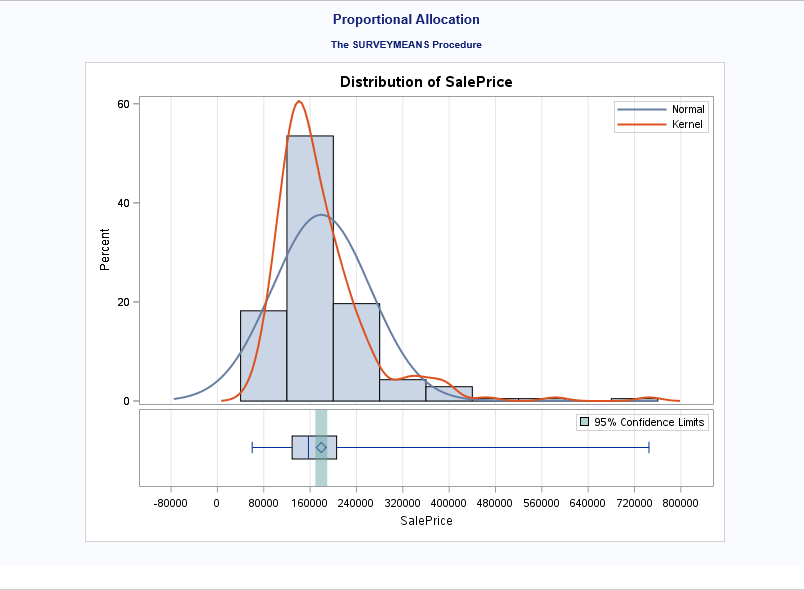
****

Fig. 7 – Stratified Sampling (Proportional Allocation) SAS Output

From the above result, we get estimates as follows:

Sample mean (x-bar) = 179,286

SE of sample mean (s) = 5,211.09

95% CI for sample mean = [169,010, 189,562]

**1.4.3 Stratified Sampling with Neyman Allocation**

This is another type of stratification method in which the variation of an auxiliary variable between strata are considered for determining the sample sizes for different strata to form the sample. Taking the variance of gross living area (*GrLivArea*) into account, we calculate the sample sizes for each stratum in the table below (Fig. 8 & Fig. 9).

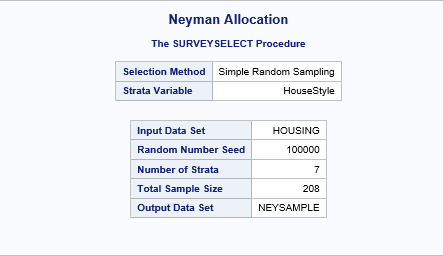
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Stratum** | **Count** | **Average *SalePrice*** | **StdDev of *SalePrice*** | **Average of *GrLivArea*** | **StdDev of *GrLivArea*** |
| 1.5Fin | 154 | 143,117 | 54,278 | 1,565 | 445 |
| 1.5Unf | 14 | 110,150 | 19,036 | 896 | 110 |
| 1Story | 726 | 175,985 | 77,056 | 1,309 | 381 |
| 2.5Fin | 8 | 220,000 | 118,212 | 2,848 | 613 |
| 2.5Unf | 11 | 157,355 | 63,934 | 1,908 | 445 |
| 2Story | 445 | 210,052 | 87,339 | 1,887 | 528 |
| SFoyer | 37 | 135,074 | 30,481 | 973 | 280 |
| SLvl | 65 | 166,703 | 38,305 | 1,374 | 387 |
| **Grand Total** | **1,460** | **180,921** | **79,443** | **1,515** | **525** |

Fig. 8 – Strata Count

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Stratum** | **House Style** | **No. Of Houses (Nh)** | **Average Gross Living Area** | **Std Dev Gross Living Area (sh)** | **Nh\*sh** | **Sample Size (nh)** |
| Stratum 1 | 1 Story | 726 | 1,309 | 381 | 276,606 | 92 |
| Stratum 2 | 1.5 Fin | 154 | 1,565 | 445 | 68,530 | 23 |
| Stratum 3 | 1.5 Unf | 14 | 896 | 110 | 1,540 | 0 |
| Stratum 4 | 2 Story | 445 | 1,887 | 528 | 234,960 | 78 |
| Stratum 5 | 2.5 Fin | 8 | 2,848 | 613 | 4,904 | 2 |
| Stratum 6 | 2.5 Unf | 11 | 1,908 | 445 | 4,895 | 2 |
| Stratum 7 | Sfoyer | 37 | 973 | 280 | 10,360 | 3 |
| Stratum 8 | SLvl | 65 | 1,374 | 387 | 25,155 | 8 |
| **Total** |  | **1,460** |  |  | **626,950** | **208** |

Fig. 9 – Strata Sample Sizes (Neyman Allocation)

SAS output for stratification with Neyman allocation is provided below (Fig. 10 & Fig. 11).

****

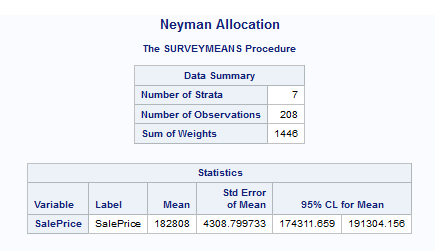


Fig. 10 – Stratified Sampling (Neyman Allocation) SAS output

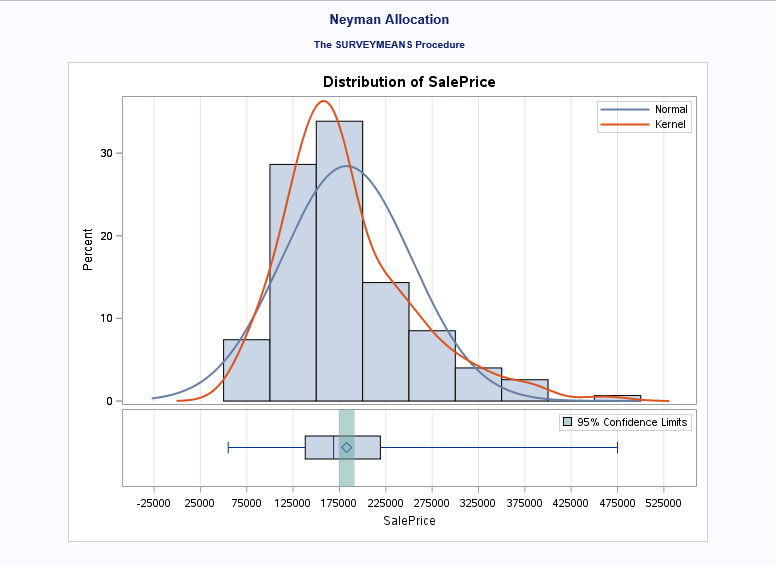
****

Fig. 11 – Stratified sampling (Neyman allocation) output

From the above results, we get estimates as follows:

Sample mean (x-bar) = 182,808

SE of sample mean (s) = 4,308.80

95% CI for sample mean = [174,311, 191,304]

**1.4.4 Two-Stage Sampling**

Two-stage sampling is a type of sampling design in which the population is segregated into several clusters based on a characteristic and only a certain number of clusters are selected for sampling. These selected clusters are called Primary Sampling Units (PSU’s). In the second stage, a further sampling is followed in which samples are chosen from all selected clusters based on a sampling design. Sample units selected at this stage are known as Secondary Sampling Units (SSU’s).

In this exercise, we select five neighborhoods (20% of all neighborhoods) from the populatoin. We perform a stratification based on *HouseStyle* to select sample units from the chosen neighborhoods.

Total number of PSU’s (M) = 25

Selected number of PSU’s (m) = 5

Number of population units in selected PSU’s (N) = 141

N1 = 16, N2 = 17, N3 = 9, N4 = 74, N5 = 25 where Ni = number of SSU’s in the ith PSU

Total sample size (n) = 28

Then using a stratification based on *HouseStyle* with proportional allocation, we get the sample sizes for the 5 selected neighborhoods as follows:

n1 = 1, n2 = 15, n3 = 8, n4 = 3, n5 = 1 where ni = selected number of SSU from ith PSU

We use SAS to perform this sampling design and the results are shown below (Fig. 12 and Fig. 13):

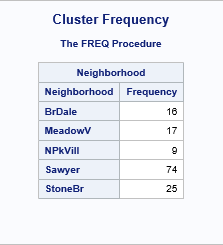
****

Fig. 12 – Two-stage sampling output

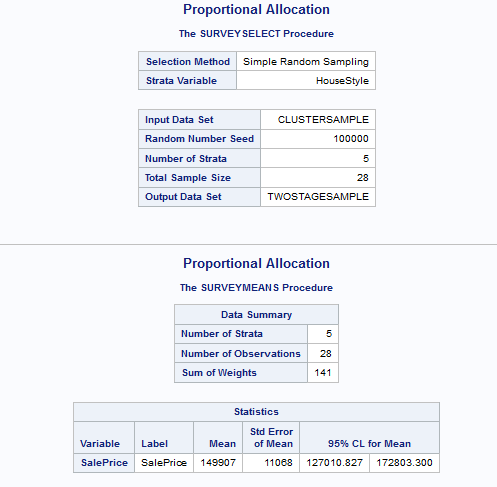


Fig. 12 – Two-stage sampling output

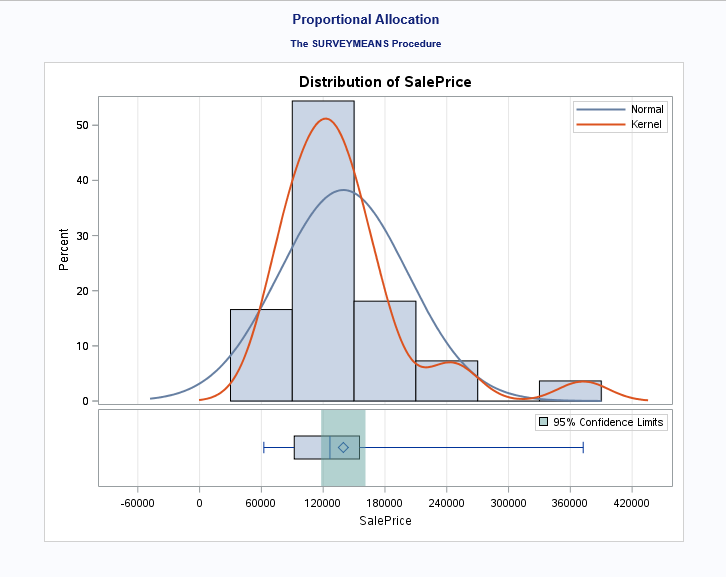
****

Fig. 13 – Two-stage sampling output

From the above results, we get estimates as follows:

Sample mean (x-bar) = 139,568

SE of sample mean (s) = 10,430.00

95% CI for sample mean = [117,991, 161,144]

**1.5 Design Effect & Comparison of Sampling Designs**

We provide a comparison of the results between the four sampling designs in the table below (Fig. 14).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sample Designs** | **Sample Mean** | **Standard Error of Sample Mean** | **Confidence Interval** | **Design Effect** |
| Simple Random Sampling | 177,378 | 5,099 | [167,325, 187,430] | - |
| Stratified Sampling (Proportional Allocation) | 179,286 | 5,211 | [169,010, 189,562] | 1.04 |
| Stratified Sampling (Neyman Allocation) | 182,808 | 4,309 | [174,311, 191,304] | 0.71 |
| Two-Stage Sampling | 139,568 | 10,430 | [117,991, 161,144] |  |

Fig. 14 – Comparison of Sampling Designs

From the above comparison, the stratified sampling design with Neyman allocation has the closest estimate for population mean ($182,921) as compared to the estimates from the other designs as well as the smallest standard error and narrowest confidence interval. This indicates that stratification with Neyman allocation based on house style provides both the most accurate (minimal bias) and the most precise (minimal variance) estimate for average house sale price in our given population of houses in Ames, Iowa.

Moreover, we see that the design effect is greater than 1 when stratification is done with proportional allocation, which suggests that we need more houses (almost 217) to achieve the same level of precision as an SRS.On the other hand, with stratified sampling design with Neyman allocation, the design effect decreases significantly to 0.71, and this indicates that only 149 houses are needed to achieve an equal precision as that obtained from an SRS.

Considering these factors, we conclude that stratified sampling design with Neyman allocation provides the best possible samples for estimating the population mean house sale price in Ames, Iowa, based on the given data.

**Task 2**

**2.1 Objective**

The objective of our second task is to select five random samples from our population of 1,460 houses for each of the four sampling techniques we use in Section 1.4. For each sample, we calculate point estimates of the mean house sale price along with their corresponding 95% confidence intervals. We then count the number of samples that contain the population mean sale price of 180,921 for each sampling technique.

**2.2 Random Sampling Results**

Using Excel’s *randbetween* function, we randomly choose twenty seeds between 0 and 999,999. We use one of these seeds in SAS’s *Proc Surveyselect* procedure for each of the five samples we select using each of the aforementioned four methods. The sample size for the simple random sample, proportional allocation and Neyman allocation samples is 208, as determined in Section 1.3. The sample for the two-stage samples is 28, as determined in Section 1.4.4.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Design** | **Iteration** | **Mean** | **Standard  Error** | **Lower**  **95% CL** | **Upper**  **95% CL** | **Percentage of CI's**  **Containing Population Mean** |
| Simple Random Sample | 1 | 175,784 | 4,614 | 166,689 | 184,880 | 80% |
| 2 | 187,675 | 5,572 | 176,690 | 198,659 |
| 3 | 176,136 | 4,282 | 167,694 | 184,578 |
| 4 | 169,466 | 4,013 | 161,555 | 177,377 |
| 5 | 178,916 | 4,601 | 169,845 | 187,988 |
| Proportional | 1 | 184,572 | 5,776 | 173,183 | 195,961 | 100% |
| 2 | 182,511 | 5,104 | 172,448 | 192,575 |
| 3 | 188,214 | 5,180 | 178,000 | 198,428 |
| 4 | 176,537 | 4,556 | 167,554 | 185,520 |
| 5 | 186,368 | 5,344 | 175,831 | 196,906 |
| Neyman | 1 | 188,179 | 5,111 | 178,100 | 198,257 | 100% |
| 2 | 181,245 | 5,092 | 171,204 | 191,286 |
| 3 | 177,541 | 4,758 | 168,160 | 186,923 |
| 4 | 184,498 | 4,910 | 174,816 | 194,180 |
| 5 | 176,185 | 4,481 | 167,348 | 185,021 |
| Two-stage (Cluster & Proportional) | 1 | 201,383 | 20,516 | 158,942 | 243,825 | 60% |
| 2 | 187,035 | 16,396 | 153,117 | 220,952 |
| 3 | 137,012 | 8,782 | 118,844 | 155,180 |
| 4 | 184,621 | 17,095 | 149,256 | 219,986 |
| 5 | 139,561 | 8,519 | 121,938 | 157,184 |

Fig. 15 – Random Sampling Results

Using SAS’s *Proc Surveymeans* procedure, we obtain the results in Fig. 15 above. The 95% confidence intervals for the mean house sale price for all five random samples via each of proportional stratified allocation and Neyman stratified allocation methods contain the population mean house sale price. However, only four of the simple random samples and only three of the two-stage cluster and proportional stratified allocation samples have 95% confidence intervals that contain the population mean.

|  |  |  |
| --- | --- | --- |
| **Design** | **Average  Mean** | **Average  Standard Error** |
| Simple Random Sample | 177,595 | 4,616 |
| Proportional | 183,640 | 5,192 |
| Neyman | 181,530 | 4,871 |
| Two-stage (Cluster & Proportional) | 169,922 | 14,262 |

Fig. 16 – Average Results by Sampling Design

The Neyman stratified allocation method comes the closet to predicting the population mean house sale price with an average of $181,530 across the five samples, merely $609 greater than the population mean (see Fig. 16). The average mean house sale price under the SRS method is more than $3,000 lower than the population mean while the average mean house sale price under the proportional stratified method is almost $2,000 more than population mean. . However, the SRS method has the lowest average standard error at $4,616 while the Neyman method comes in second at an average of $4,871. Even though the SRS has the lowest average standard error, only 80% of the confidence intervals under SRS capture the true mean. Due to the fact that only 5 of the 25 neighborhoods were used as clusters, the two-stage samples have a much higher average standard error at 14,262. The average sale price is also significantly lower at 169,922, 11,000 lower than both the population mean (180,921) and the average of the fifteen samples from the other three methods (180,922). As a result, the two-stage method not only significantly has a wide variance in its predictions but also significantly underestimates the true mean sale price. As we conclude in Section 1, we find the stratified Neyman allocation method to yield the best balance of accurately predicting the mean sale price (minimal bias) as well as minimizing the variance in its predictions.

**Appendix**

**SAS Code**

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Task 1 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

libname mydata "/home/freemand0/Data Sets/MSDS 6370 Statistical Sampling/Class Project/";

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Import Data \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

proc import datafile = "/home/freemand0/Data Sets/MSDS 6370 Statistical Sampling/Class Project/Sampling\_proj\_data.xlsx"

out = mydata.housing

dbms = xlsx replace;

getnames = Yes;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Simple Random Sampling \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

proc surveyselect data = mydata.housing method = srs n = 208 seed = 100000 out = mydata.srssample;

title "Simple Random Sample";

run;

proc surveymeans data = mydata.srssample total = 1460

mean clm;

var SalePrice;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Stratified Sampling with House Style for Stratification \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

proc sort data = mydata.housing;

by HouseStyle;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Proportion Allocation \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

proc surveyselect data = mydata.housing method = srs out = mydata.propsample

sampsize = (22, 2, 103, 1, 2, 64, 5, 9) seed = 100000;

strata HouseStyle;

title "Proportional Allocation";

run;

data mydata.strsizes;

informat HouseStyle 6.;

format HouseStyle 6.;

input HouseStyle 6. \_total\_;

datalines;

1.5Fin 154

1.5Unf 14

1Story 726

2.5Fin 8

2.5Unf 11

2Story 445

SFoyer 37

SLvl 65

;

run;

proc surveymeans data = mydata.propsample total = mydata.strsizes

mean clm;

var SalePrice;

weight SamplingWeight;

strata HouseStyle;

title "Proportional Allocation";

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Neyman Allocation \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

proc surveyselect data = mydata.housing method = srs out = mydata.neysample

sampsize = (23, 0, 92, 2, 2, 78, 3, 8) seed = 100000;

strata HouseStyle;

title "Neyman Allocation";

run;

proc surveymeans data = mydata.neysample total = mydata.strsizes

mean clm;

var SalePrice;

weight SamplingWeight;

strata HouseStyle;

title "Neyman Allocation";

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Two Stage Sampling (Neighborhood for Cluster & House Style for Stratification) \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

proc surveyselect data = mydata.housing noprint sampsize = 5 seed = 100000 out = mydata.ClusterSample;

cluster Neighborhood;

run;

proc freq data = mydata.ClusterSample;

tables Neighborhood / nocum nopercent;

title "Cluster Frequency";

run;

proc sort data = mydata.ClusterSample;

by HouseStyle;

run;

proc surveyselect data = mydata.ClusterSample method = srs out = mydata.twostagesample

sampsize = (1, 15, 8, 3, 1) seed = 100000;

strata HouseStyle;

title "Proportional Allocation";

run;

data mydata.strsizes\_cluster;

informat HouseStyle 6.;

format HouseStyle 6.;

input HouseStyle 6. \_total\_;

datalines;

1.5Fin 5

1Story 77

2Story 41

SFoyer 12

SLvl 6

;

run;

proc surveymeans data = mydata.twostagesample total = mydata.strsizes\_cluster

mean clm;

var SalePrice;

weight SamplingWeight;

strata HouseStyle;

title "Proportional Allocation";

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Task 2 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* True Population Mean \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

proc means data = mydata.housing;

var SalePrice;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Simple Random Sampling \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

proc surveyselect data = mydata.housing method = srs n = 208 seed = 965617 out = mydata.srssample1;

title "Simple Random Sample";

run;

proc surveymeans data = mydata.srssample1 total = 1460

mean clm;

var SalePrice;

run;

proc surveyselect data = mydata.housing method = srs n = 208 seed = 662888 out = mydata.srssample2;

title "Simple Random Sample";

run;

proc surveymeans data = mydata.srssample2 total = 1460

mean clm;

var SalePrice;

run;

proc surveyselect data = mydata.housing method = srs n = 208 seed = 869404 out = mydata.srssample3;

title "Simple Random Sample";

run;

proc surveymeans data = mydata.srssample3 total = 1460

mean clm;

var SalePrice;

run;

proc surveyselect data = mydata.housing method = srs n = 208 seed = 147603 out = mydata.srssample4;

title "Simple Random Sample";

run;

proc surveymeans data = mydata.srssample4 total = 1460

mean clm;

var SalePrice;

run;

proc surveyselect data = mydata.housing method = srs n = 208 seed = 90873 out = mydata.srssample5;

title "Simple Random Sample";

run;

proc surveymeans data = mydata.srssample5 total = 1460

mean clm;

var SalePrice;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Stratified Sampling with House Style for Stratification \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

proc sort data = mydata.housing;

by HouseStyle;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Proportion Allocation \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

proc surveyselect data = mydata.housing method = srs out = mydata.propsample1

sampsize = (22, 2, 103, 1, 2, 64, 5, 9) seed = 818956;

strata HouseStyle;

title "Proportional Allocation";

run;

proc surveymeans data = mydata.propsample1 total = mydata.strsizes

mean clm;

var SalePrice;

weight SamplingWeight;

strata HouseStyle;

title "Proportional Allocation";

run;

proc surveyselect data = mydata.housing method = srs out = mydata.propsample2

sampsize = (22, 2, 103, 1, 2, 64, 5, 9) seed = 93297;

strata HouseStyle;

title "Proportional Allocation";

run;

proc surveymeans data = mydata.propsample2 total = mydata.strsizes

mean clm;

var SalePrice;

weight SamplingWeight;

strata HouseStyle;

title "Proportional Allocation";

run;

proc surveyselect data = mydata.housing method = srs out = mydata.propsample3

sampsize = (22, 2, 103, 1, 2, 64, 5, 9) seed = 496581;

strata HouseStyle;

title "Proportional Allocation";

run;

proc surveymeans data = mydata.propsample3 total = mydata.strsizes

mean clm;

var SalePrice;

weight SamplingWeight;

strata HouseStyle;

title "Proportional Allocation";

run;

proc surveyselect data = mydata.housing method = srs out = mydata.propsample4

sampsize = (22, 2, 103, 1, 2, 64, 5, 9) seed = 342872;

strata HouseStyle;

title "Proportional Allocation";

run;

proc surveymeans data = mydata.propsample4 total = mydata.strsizes

mean clm;

var SalePrice;

weight SamplingWeight;

strata HouseStyle;

title "Proportional Allocation";

run;

proc surveyselect data = mydata.housing method = srs out = mydata.propsample5

sampsize = (22, 2, 103, 1, 2, 64, 5, 9) seed = 154257;

strata HouseStyle;

title "Proportional Allocation";

run;

proc surveymeans data = mydata.propsample5 total = mydata.strsizes

mean clm;

var SalePrice;

weight SamplingWeight;

strata HouseStyle;

title "Proportional Allocation";

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Neyman Allocation \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

proc surveyselect data = mydata.housing method = srs out = mydata.neysample1

sampsize = (23, 0, 92, 2, 2, 78, 3, 8) seed = 8509;

strata HouseStyle;

title "Neyman Allocation";

run;

proc surveymeans data = mydata.neysample1 total = mydata.strsizes

mean clm;

var SalePrice;

weight SamplingWeight;

strata HouseStyle;

title "Neyman Allocation";

run;

proc surveyselect data = mydata.housing method = srs out = mydata.neysample2

sampsize = (23, 0, 92, 2, 2, 78, 3, 8) seed = 357761;

strata HouseStyle;

title "Neyman Allocation";

run;

proc surveymeans data = mydata.neysample2 total = mydata.strsizes

mean clm;

var SalePrice;

weight SamplingWeight;

strata HouseStyle;

title "Neyman Allocation";

run;

proc surveyselect data = mydata.housing method = srs out = mydata.neysample3

sampsize = (23, 0, 92, 2, 2, 78, 3, 8) seed = 457196;

strata HouseStyle;

title "Neyman Allocation";

run;

proc surveymeans data = mydata.neysample3 total = mydata.strsizes

mean clm;

var SalePrice;

weight SamplingWeight;

strata HouseStyle;

title "Neyman Allocation";

run;

proc surveyselect data = mydata.housing method = srs out = mydata.neysample4

sampsize = (23, 0, 92, 2, 2, 78, 3, 8) seed = 732870;

strata HouseStyle;

title "Neyman Allocation";

run;

proc surveymeans data = mydata.neysample4 total = mydata.strsizes

mean clm;

var SalePrice;

weight SamplingWeight;

strata HouseStyle;

title "Neyman Allocation";

run;

proc surveyselect data = mydata.housing method = srs out = mydata.neysample5

sampsize = (23, 0, 92, 2, 2, 78, 3, 8) seed = 202192;

strata HouseStyle;

title "Neyman Allocation";

run;

proc surveymeans data = mydata.neysample5 total = mydata.strsizes

mean clm;

var SalePrice;

weight SamplingWeight;

strata HouseStyle;

title "Neyman Allocation";

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Two Stage Sampling (Neighborhood for Cluster & House Style for Stratification) \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

proc surveyselect data = mydata.housing noprint sampsize = 5 seed = 100000 out = mydata.ClusterSample;

cluster Neighborhood;

run;

proc freq data = mydata.ClusterSample;

tables Neighborhood / nocum nopercent;

title "Cluster Frequency";

run;

proc sort data = mydata.ClusterSample;

by HouseStyle;

run;

proc surveyselect data = mydata.ClusterSample method = srs out = mydata.twostagesample1

sampsize = (1, 15, 8, 3, 1) seed = 606631;

strata HouseStyle;

title "Proportional Allocation";

run;

proc surveymeans data = mydata.twostagesample1 total = mydata.strsizes\_cluster

mean clm;

var SalePrice;

weight SamplingWeight;

strata HouseStyle;

title "Proportional Allocation";

run;

proc surveyselect data = mydata.ClusterSample method = srs out = mydata.twostagesample2

sampsize = (1, 15, 8, 3, 1) seed = 632472;

strata HouseStyle;

title "Proportional Allocation";

run;

proc surveymeans data = mydata.twostagesample2 total = mydata.strsizes\_cluster

mean clm;

var SalePrice;

weight SamplingWeight;

strata HouseStyle;

title "Proportional Allocation";

run;

proc surveyselect data = mydata.ClusterSample method = srs out = mydata.twostagesample3

sampsize = (1, 15, 8, 3, 1) seed = 841676;

strata HouseStyle;

title "Proportional Allocation";

run;

proc surveymeans data = mydata.twostagesample3 total = mydata.strsizes\_cluster

mean clm;

var SalePrice;

weight SamplingWeight;

strata HouseStyle;

title "Proportional Allocation";

run;

proc surveyselect data = mydata.ClusterSample method = srs out = mydata.twostagesample4

sampsize = (1, 15, 8, 3, 1) seed = 13977;

strata HouseStyle;

title "Proportional Allocation";

run;

proc surveymeans data = mydata.twostagesample4 total = mydata.strsizes\_cluster

mean clm;

var SalePrice;

weight SamplingWeight;

strata HouseStyle;

title "Proportional Allocation";

run;

proc surveyselect data = mydata.ClusterSample method = srs out = mydata.twostagesample5

sampsize = (1, 15, 8, 3, 1) seed = 939001;

strata HouseStyle;

title "Proportional Allocation";

run;

proc surveymeans data = mydata.twostagesample5 total = mydata.strsizes\_cluster

mean clm;

var SalePrice;

weight SamplingWeight;

strata HouseStyle;

title "Proportional Allocation";

run;