**Sample Selection from Real Estate Data on 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 and study the characteristics of the selected items to get tangible insights on the characteristics of the population where the sample was drawn from. This project on Statistical Sampling Techniques directs into drawing a sizable sample based on various sample designs to derive an idea about population characteristics. Here we follow Stratified sampling along with few other sampling methods to get estimates of population parameters. The derived estimates from various methods are compared to assess their performance in representing the entire population.

**1.2 Data**

The team decide to work on a real estate dataset sourced from Kaggle (<https://www.kaggle.com/c/fi-ames-housing-price-competition/data>). This dataset provides the housing sale prices in the different neighborhoods of Ames, Iowa in between 2006 and 2010 along with several other aspects on real estate sales like zone, neighborhood, character of the lot, house style, age, foundation, house features, garage quality, sale type, etc. The original raw dataset contains 81 variables, both categorical and continuous, across 1460 unique sales.

For sampling purpose, we set our focus on seven fields to derive an estimate for average sale prices of houses in Ames. The attributes considered in sampling exercise are – Neighborhood, Building type, House Style, Number of rooms in a house, Gross living area, Year of sale and above all Sale price. For this data we have,

Population size (N) = 1460, Population parameter of interest = Average sale price

Average sale price for population (µ) = $180,921 Standard deviation of sale prices for population (σ) = $79,415

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

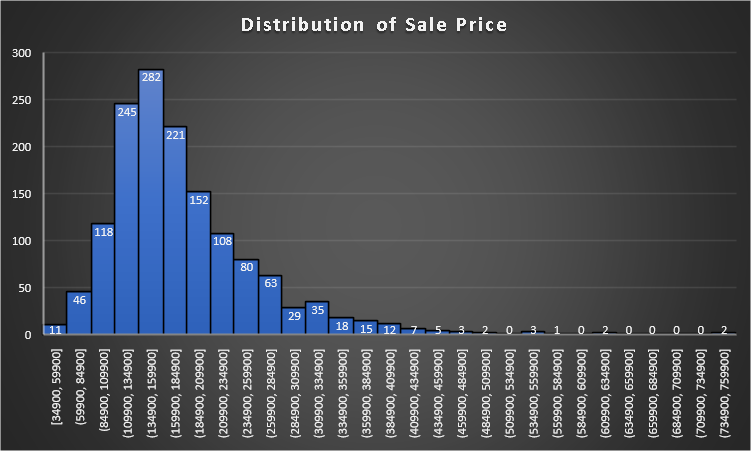


Fig. 1 – Population Distribution of House Sale Prices in Ames, Iowa

In our population, we observe sale price ranges from $35,000 to above $ 750,000 but more than 50% of sales cost $110,000 to $185,000. The above graph indicates a deviation from Normal distribution for sale prices due to an extreme right skewness which indicates presence of few costly sales taken place in scope of our dataset. With help of Central Limit Theorem and considering a sufficient large sample, despite the observed skewness in data, we safely assume Normality in population distribution.

**1.3 Sample Size**

To determine an optimal sample size for our design, we set a margin of error (MOE) for our estimates at $10,000, an approximated 5% level of average sale price in the population. Assuming normality in population 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 sample design, we continue to use 208 as 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 average sale price of houses sold.

**1.4.1 Simple Random Sampling (SRS)**

Simple random sampling is a sampling design where each sample unit is chosen randomly so that at any stage all the population units have 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).

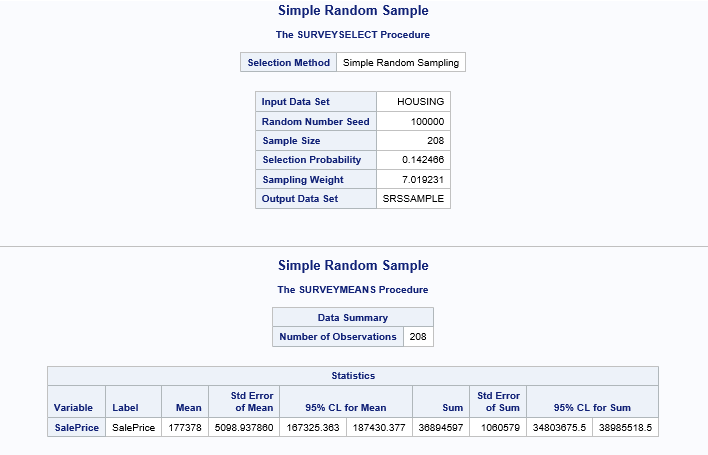
****

Fig. 2 – SRS output

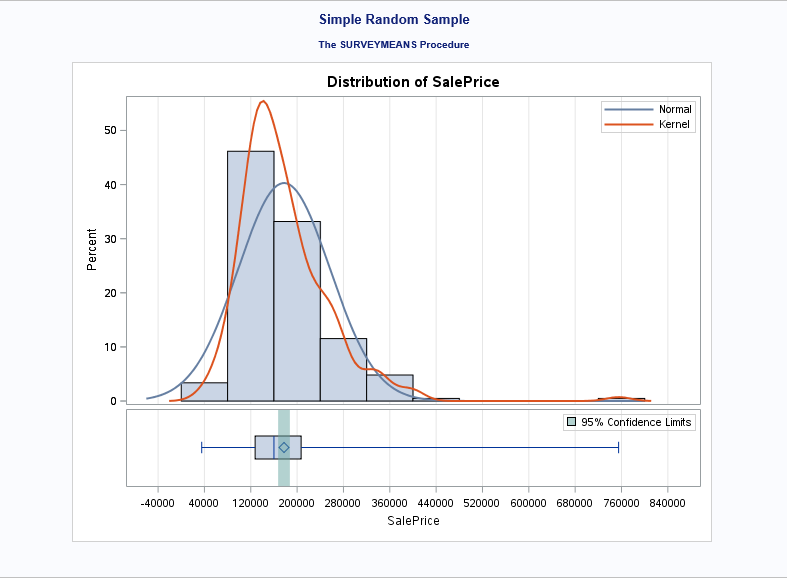
****

Fig. 3 – SRS output

From the above output we get estimates as,

Sample mean (x-bar) = 177378

S.E of sample mean (s) = 5098.94

95% C.I for sample mean = [167325, 187430]

**1.4.2 Stratified Sampling with Proportional Allocation**

Stratified sampling is a technique followed when population have several homogeneous subgroups, known as stratum and the sample is drawn in such a way that all strata in population have representatives in the sample. 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 | 143117 | 54278 |
| 1.5Unf | 14 | 110150 | 19036 |
| 1Story | 726 | 175985 | 77056 |
| 2.5Fin | 8 | 220000 | 118212 |
| 2.5Unf | 11 | 157355 | 63934 |
| 2Story | 445 | 210052 | 87339 |
| SFoyer | 37 | 135074 | 30481 |
| SLvl | 65 | 166703 | 38305 |
| **Grand Total** | **1460** | **180921** | **79443** |

Fig. 4 – Strata Count

Since we are to draw a composite sample of size 208, with proportional allocation the sample sizes for strata are calculated 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** |  | **1460** |  | **208** |

Fig. 5 – Sample size for strata

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

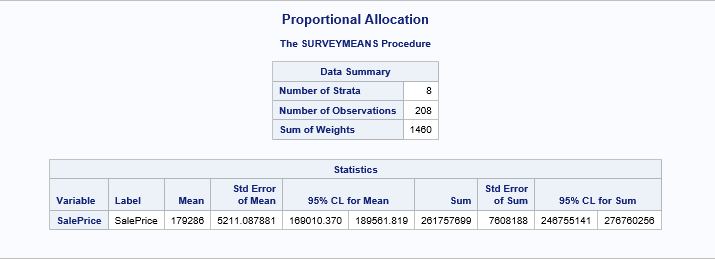
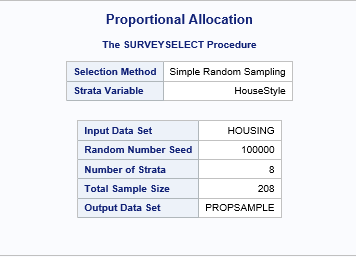
****

Fig. 6 – Stratified sampling (proportional allocation) output

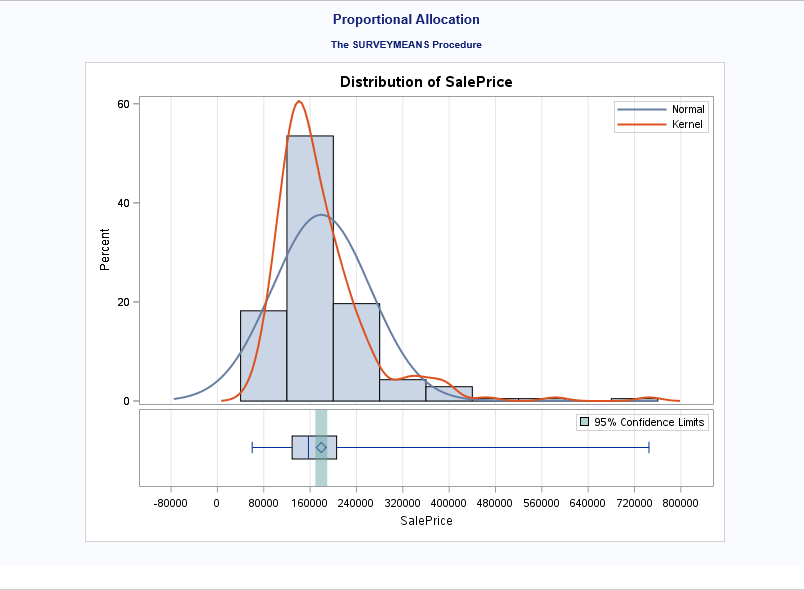
****

Fig. 7 – Stratified sampling (proportional allocation) output

From the above result we get estimates as,

Sample mean (x-bar) = 179286

S.E of sample mean (s) = 5211.09

95% C.I for sample mean = [169010, 189562]

**1.4.3 Stratified Sampling with Neyman Allocation**

This is another type of stratification method where the variation of an auxiliary variable between strata are considered for determining sample sizes from different strata to form the sample. Sample sizes are calculated in the tables below (Fig. – 8 & Fig. - 9).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Stratum** | **Count** | **Average SalePrice** | **StdDev of SalePrice** | **Average of GrLivArea** | **StdDev of GrLivArea** |
| 1.5Fin | 154 | 143117 | 54278 | 1565 | 445 |
| 1.5Unf | 14 | 110150 | 19036 | 896 | 110 |
| 1Story | 726 | 175985 | 77056 | 1309 | 381 |
| 2.5Fin | 8 | 220000 | 118212 | 2848 | 613 |
| 2.5Unf | 11 | 157355 | 63934 | 1908 | 445 |
| 2Story | 445 | 210052 | 87339 | 1887 | 528 |
| SFoyer | 37 | 135074 | 30481 | 973 | 280 |
| SLvl | 65 | 166703 | 38305 | 1374 | 387 |
| **Grand Total** | **1460** | **180921** | **79443** | **1515** | **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 | 1309 | 381 | 276606 | 92 |
| Stratum 2 | 1.5 Fin | 154 | 1565 | 445 | 68530 | 23 |
| Stratum 3 | 1.5 Unf | 14 | 896 | 110 | 1540 | 0 |
| Stratum 4 | 2 Story | 445 | 1887 | 528 | 234960 | 78 |
| Stratum 5 | 2.5 Fin | 8 | 2848 | 613 | 4904 | 2 |
| Stratum 6 | 2.5 Unf | 11 | 1908 | 445 | 4895 | 2 |
| Stratum 7 | Sfoyer | 37 | 973 | 280 | 10360 | 3 |
| Stratum 8 | SLvl | 65 | 1374 | 387 | 25155 | 8 |
| **Total** |  | **1460** |  |  | **626950** | **208** |

Fig. 9 – Sample size for strata

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

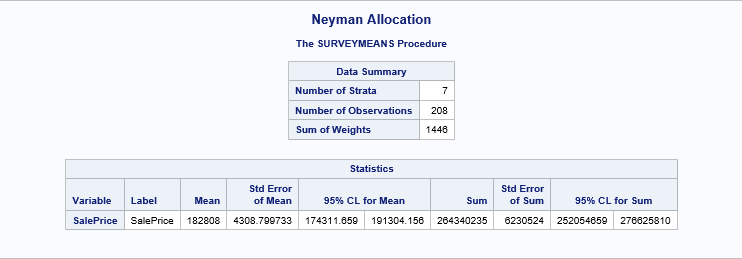
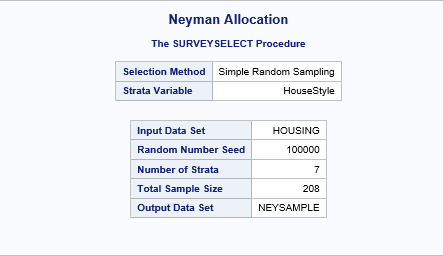
****

Fig. 10 – Stratified sampling (Neyman allocation) output

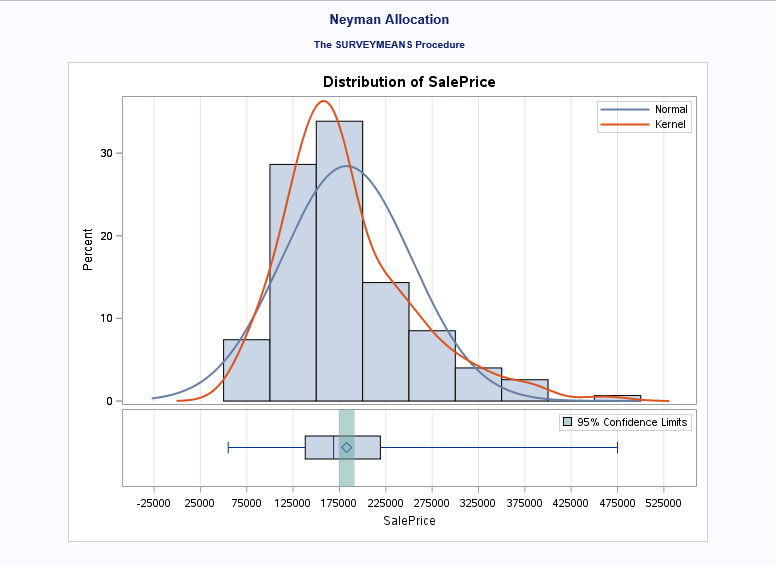
****

Fig. 11 – Stratified sampling (Neyman allocation) output

From the above result we get estimates as,

Sample mean (x-bar) = 182808

S.E of sample mean (s) = 4308.80

95% C.I for sample mean = [174311, 191304]

**1.4.4 Two-Stage Sampling**

Two-stage sampling is a specific type of sampling design where population is segregated in several clusters based on a characteristic and a specific number of clusters are selected from all the clusters. These selected clusters are called Primary Sampled Units (PSU). In second stage, a further sampling is followed where samples are chosen from all selected clusters based on a sampling design. And sample units selected at this stage are known Secondary Sampled Units (SSU).

Here in this exercise, we select 5 neighborhoods (20% of all neighborhoods) from the whole dataset. A stratification has been performed based on ‘HouseStyle’ to select sample units from the chosen neighborhoods.

Total number of PSU - Neighborhood (M) = 25

Selected number of PSU (m) = 5

Number of population units in selected neighborhoods (N) = 141

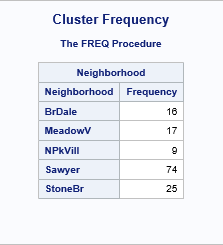
N1 = 16, N2 = 17, N3 = 9, N4 = 74, N5 = 25 where Ni = number of SSU in ith PSU

Total sample size (n) = 28

Then with a stratification based on ‘HouseStyle’ with proportional allocation we get the sample sizes for 5 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):

****

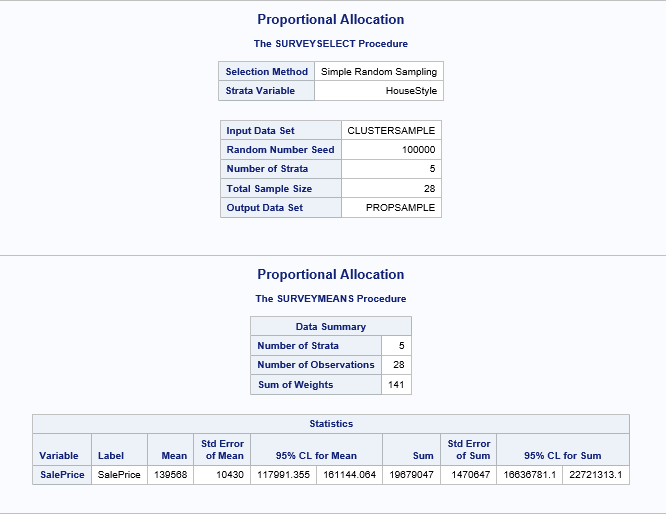
****

Fig. 12 – Two-stage sampling output

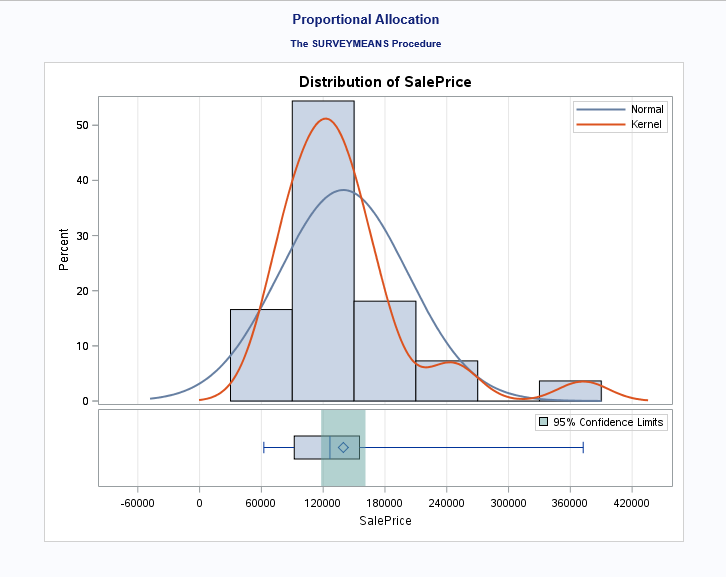
****

Fig. 13 – Two-stage sampling output

From the above result we get estimates as,

Sample mean (x-bar) = 139568

S.E of sample mean (s) = 10430.00

95% C.I for sample mean = [117991, 161144]

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

A comparison between sampling designs, based on the results we get, is given in the table below (Fig. 14).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sample Designs** | **Sample Mean** | **Standard Error of Sample Mean** | **Confidence Interval** | **Design Effect** |
| Simple Random Sampling | 177378 | 5099 | [167325, 187430] | - |
| Stratified Sampling (Proportional Allocation) | 179286 | 5211 | [169010, 189562] | 1.04 |
| Stratified Sampling (Neyman Allocation) | 182808 | 4309 | [174311, 191304] | 0.71 |
| Two-Stage Sampling | 139568 | 10430 | [117991, 161144] |  |

Fig. 14 – Comparison of sampling designs

It appears, from above comparison, that stratified sampling design results into the closest estimate for population mean (= 182921) as compared to estimates from other designs and have the minimum most standard error with the narrowest possible confidence interval. This indicates that stratification with Neyman allocation based on house style provides the most accurate and precise estimate for average sale price of the hoses for the given population.

Moreover, we see that design effect is greater than 1 when stratification done with proportional allocation which suggests that we need more sample units (almost 217 sample units) to achieve the same level of precision with SRS.But with stratified sampling design with Neyman allocation, design effect goes down significantly to 0.71 and this indicates that only 149 sample units would be sufficient to achieve an equal precision with a precision obtained from SRS.

Considering these factors, it can be concluded that stratified sampling design with Neyman allocation provides the best possible samples for estimating the population mean for house sale prices in Ames, Iowa, based on the given data.