Jack Nelson, Data Science Graduate Student, Southern Methodist University

[[1]](#footnote-1)

MSDS 6370 Term Project: Designing a Stratification Sampling Plan(April 2018)

*Introduction* — Numerous methods and techniques have been used in research and in practice to determine optimal sampling plans when predicting a desired attribute(s). A sample dataset given for the purposes of this project was used to execute three stratification sampling design methods. Methods for determining strata include using cumulative values of the attribute, leveraging the square root of the frequency after separating values in a cumulative value fashion, and approximating weight and strata standard deviation so that they are equal to one another. Once strata are determined using each technique, all techniques were tested to determine the most accurate and precise method.

TABLE I

Basic Statistics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Total | Mean | Standard Deviation | Range |
| Sales | 1,388,810,308 | 142,267 | 665,959 | 52,156,303 |
| Inventory | 1,754,954,823 | 179,774 | 1,573,358 | 105,379,550 |

The data given is a simulated dataset reflecting a month’s worth of information for a particular industry. It contains three attributes: a unique identifier (coID), current month sales for a company (Sales), and current month inventory for a company (Inventory). Sales was used to carry out each stratification method, however, Inventory was used for a separate paper. Results based on both Sales and Inventory were evaluated to determine the best option to predict not only the attribute the sample design is based on, but also its counterpart’s attribute.

The paper will first provide basic descriptive statistics for the attribute being used in the stratification sampling design. Then strata will be determined using the three techniques that will be explained in better detail at that time. Neyman allocation will then be carried out to determine sample sizes per stratum. A series of samples are taken from the sampling design and run to receive predicted measures and precision for the predictions. The results are analyzed further to determine the recommended sampling plan to predict both attributes. Conclusions will follow noting key take a ways from the project.

# Descriptive Statistics

S

ales was the attribute of interest throughout the entirety of designing strata, however, Inventory was also leveraged in measuring accuracy and precision. A total of 9,762 simulated companies reported Sales and Inventory for a particular month. There is no specification for the measuring units or currency domination for either attribute, but it will be assumed U.S. dollars was used. Basic statistics for each variable, including population total, population mean, standard deviation, and range, are shown in Table I. In addition, histograms visualizing the distribution of each attribute are shown in Figure I. It is observed that both distributions are heavily skewed to the right. Log transformations to each variable were made and re-visualized in a histogram in Figure II. The log transformed variables show distributions much more consistent with the normal. Original variable skewness gives reasoning to develop a stratified sampling design as opposed to a simple random sample.

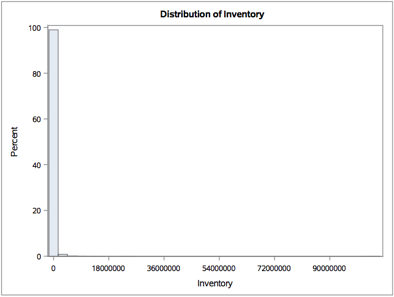
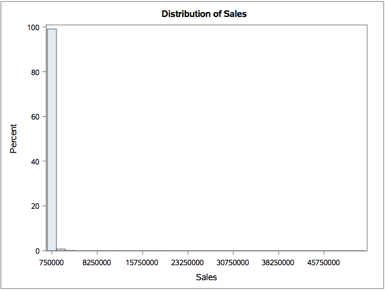


FIGURE I

Distribution Histogram: Sales & Inventory

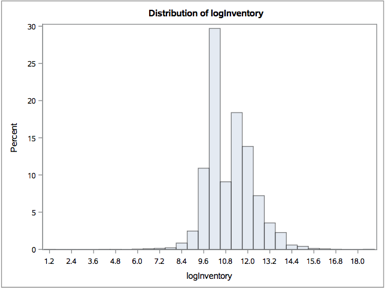
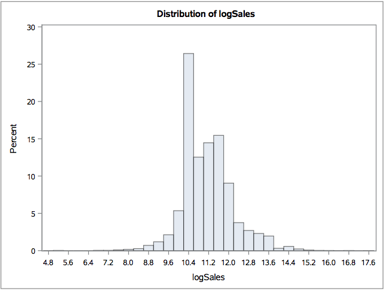


FIGURE II

Distribution Histogram: Log Transformed Sales & Inventory

A strong correlation is observed between the two variables (Pearson R = 0.82) presenting evidence that a well-designed sampling plan should have the ability to generalize both Sales and Inventory. The correlation can be observed when the variables are plotted on the log scale in Figure III. All statistical calculations and visual plots were generated in SAS using various PROC commands. The script and log can be found in the appendix.

# Stratification Design

Three separate methods were used to form strata. Each build upon one another and progress in complexity. Sections for each method briefly describe the procedure as well as provide results from the given dataset, dividing observations into each stratum. All calculations needed to determine strata for all three methods were carried out using Python script on a Jupyter Notebook, which is provided in the Appendix.

TABLE II

Cumulative Strata Results

|  |  |  |
| --- | --- | --- |
| Stratum | N | Cumulative Sales Threshold |
| 1 | 4285 | 138,881,031 |
| 2 | 1988 | 277,762,062 |
| 3 | 1308 | 416,643,092 |
| 4 | 919 | 555,524,123 |
| 5 | 554 | 694,405,154 |
| 6 | 314 | 833,286,185 |
| 7 | 201 | 972,167,216 |
| 8 | 125 | 1,111,048,246 |
| 9 | 57 | 1,249,929,277 |
| 10 | 11 | 1,388,810,308 |

TABLE II

Cumulative Strata Results

|  |  |  |
| --- | --- | --- |
| Stratum | N | Cumulative Sales Threshold |
| 1 | 4285 | 138,881,031 |
| 2 | 1988 | 277,762,062 |
| 3 | 1308 | 416,643,092 |
| 4 | 919 | 555,524,123 |
| 5 | 554 | 694,405,154 |
| 6 | 314 | 833,286,185 |
| 7 | 201 | 972,167,216 |
| 8 | 125 | 1,111,048,246 |
| 9 | 57 | 1,249,929,277 |
| 10 | 11 | 1,388,810,308 |

TABLE III

Subgroup Results

|  |  |  |
| --- | --- | --- |
| Subgroup |  |  |
| 1 | 2560 | 50.6 |
| 2 | 1725 | 41.5 |
| 3 | 1138 | 33.7 |
| 4 | 850 | 29.2 |
| 5 | 707 | 26.6 |
| 6 | 601 | 24.5 |
| 7 | 498 | 22.3 |
| 8 | 421 | 20.5 |
| 9 | 319 | 17.9 |
| 10 | 235 | 15.2 |
| 11 | 175 | 13.2 |
| 12 | 139 | 11.8 |
| 13 | 109 | 10.4 |
| 14 | 92 | 9.6 |
| 15 | 79 | 8.9 |
| 16 | 46 | 6.8 |
| 17 | 34 | 5.8 |
| 18 | 23 | 4.8 |
| 19 | 9 | 3.0 |
| 20 | 2 | 1.4 |

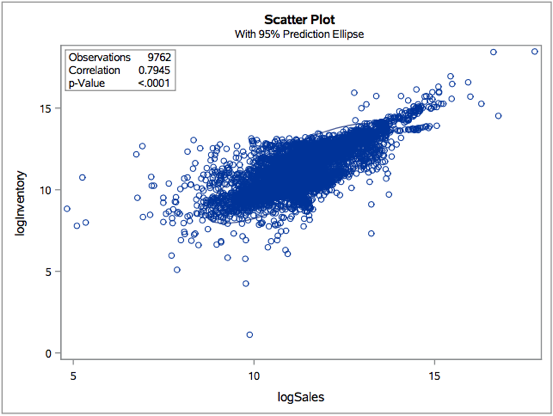


TABLE IV

Cumulative Strata Results

|  |  |  |
| --- | --- | --- |
| Stratum | N | Cumulative Threshold |
| 1 | 2560 | 35.79 |
| 2 | 0 | 71.58 |
| 3 | 1725 | 107.37 |
| 4 | 1988 | 143.16 |
| 5 | 707 | 178.95 |
| 6 | 1099 | 214.75 |
| 7 | 740 | 250.54 |
| 8 | 410 | 286.33 |
| 9 | 419 | 322.12 |
| 10 | 114 | 357.91 |

FIGURE III

Correlation Scatterplot: Log Transformed Sales & Inventory

## Cumulative

The cumulative method forms strata by sorting values in ascending order and developing stratum cutoff points when the cumulative sum crosses that threshold. The thresholds are equally spaced based on the desired number of strata.

The Sales variable was sorted to calculate the cumulative sum as the variable increased in value. Ten strata thresholds were made by taking the population sum divided by ten times the stratum number (1,388,810,308 / 10 \* stratum). Each observation was then binned in accordance to its threshold. Results are shown in Table II.

TABLE II

Cumulative Strata Results

|  |  |  |
| --- | --- | --- |
| Stratum | N | Cumulative Sales Threshold |
| 1 | 4285 | 138,881,031 |
| 2 | 1988 | 277,762,062 |
| 3 | 1308 | 416,643,092 |
| 4 | 919 | 555,524,123 |
| 5 | 554 | 694,405,154 |
| 6 | 314 | 833,286,185 |
| 7 | 201 | 972,167,216 |
| 8 | 125 | 1,111,048,246 |
| 9 | 57 | 1,249,929,277 |
| 10 | 11 | 1,388,810,308 |

## Cumulative

TABLE V

Equal Strata Results

|  |  |
| --- | --- |
| Stratum | N |
| 1 | 4208 |
| 2 | 3435 |
| 3 | 1212 |
| 4 | 529 |
| 5 | 280 |
| 6 | 85 |
| 7 | 11 |
| 8 | 2 |

In a very similar fashion as the cumulative method, the cumulative square root of frequency method begins by sorting a variable in ascending order and subdividing observations in accordance to equally separated cumulative sum thresholds. Typically, much more subdivisions are made than the desired number of strata. The square root of the N value, or frequency in each subgroup is calculated and used as a new measure to group subgroups into the desired number of strata in a cumulative fashion. This method determines strata by approximating a minimum of value for weights and standard deviation within subgroups.

For the given dataset, twenty subgroups were made based on the cumulative sum of Sales. The twenty subgroups were then used to determine ten strata. Results for the subgroups are shown in Table III and the strata results are shown in Table IV. Notice only nine strata were made in the results since no subgroup fell in the stratum 2 threshold.

## Equal

The goal of the Equal method is to find stratum sizes such that [1] where [2]. This is done through iterating between the two equations, updating stratum sizes, until a local minimum is found. At which point stratum sizes will have minimal deviations between iterations. In order to find these stratum sizes, a baseline needs to be developed for k using an existing stratification design. The closer the starting point is to a local minimum; the less iterations need to be executed.

A baseline was established using the cumulative square root of frequency method where . New sizes for each stratum was determined by plugging into equation [2] and solving for . This procedure was performed fifteen times in an attempt to find a local minimum. While performing each iteration, number of strata reduced twice to end up with only eight strata in the final strata size design. The results are shown in Table V.

# Neyman Allocation

A stratified sample size of 500 is calculated for each stratification design using the Neyman allocation method. Neyman allocation determines sample sizes per stratum based on the variance and size of a stratum relative to all strata variances and sizes. The following equation is applied to each stratum to approximate a sample size:

TABLE VIII

Cumulative Sample Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Seed | Total | Standard Deviation | 95% CI Total Min | 95% CI Total Max |
| 91119 | 1,379,235,591 | 10,269,590 | 1,359,057,634 | 1,399,413,368 |
| 91120 | 1,387,966,563 | 10,174,115 | 1,367,976,288 | 1,407,956,838 |
| 91121 | 1,387,599,506 | 10,849,340 | 1,366,282,538 | 1,408,916,475 |
| 91122 | 1,378,281,016 | 9,638,689 | 1,359,342,755 | 1,397,219,277 |
| 91123 | 1,392,994,198 | 9,848,483 | 1,373,643,730 | 1,412,344,667 |
| Actual | 1,388,810,308 |  |  |  |

Typically, rounding and minor adjustments are made to uphold the desired total sample size. In the case of this particular dataset, variance among smaller strata sizes at the top end of the sales distribution is causing Neyman allocation to designate sample sizes much bigger than the stratum size. This was done purposefully to determine the certainty set. Strata at the top where this occurs was clumped together to form a certainty set, leaving leftover samples to the next highest stratum. Results for all three stratification designs are shown from Table VI through Table VIII.

TABLE IX

Cumulative Sample Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Seed | Total | Standard Deviation | 95% CI Total Min | 95% CI Total Max |
| 91119 | 1,384,234,219 | 14,134,322 | 1,356,463,001 | 1,412,005,437 |
| 91120 | 1,347,254,321 | 12,663,776 | 1,322,372,443 | 1,372,136,199 |
| 91121 | 1,388,073,264 | 14,841,262 | 1,358,913,044 | 1,417,233,483 |
| 91122 | 1,386,713,823 | 13,908,040 | 1,359,387,206 | 1,414,040,440 |
| 91123 | 1,398,234,500 | 14,418,129 | 1,369,905,655 | 1,426,563,344 |
| Actual | 1,388,810,308 |  |  |  |

All Neyman allocation procedures were done in a Jupyter Notebook using a Python script. The file can be found in the Appendix.

TABLE X

Equal Sample Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Seed | Total | Standard Deviation | 95% CI Total Min | 95% CI Total Max |
| 91119 | 1,373,257,324 | 13,329,405 | 1,347,067,744 | 1,399,446,905 |
| 91120 | 1,390,733,372 | 13,183,445 | 1,364,830,574 | 1,416,636,171 |
| 91121 | 1,360,759,975 | 12,610,850 | 1,335,982,211 | 1,385,537,740 |
| 91122 | 1,391,320,921 | 12,541,286 | 1,366,679,834 | 1,415,962,007 |
| 91123 | 1,384,713,667 | 12,025,986 | 1,361,085,040 | 1,408,342,293 |
| Actual | 1,388,810,308 |  |  |  |

TABLE VII

Equal Sample Sizes

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 1 | 51 | 51 |
| 2 | 97 | 97 |
| 3 | 57 | 57 |
| 4 | 60 | 60 |
| 5 | 58 | 137 |
| 6 | 60 | 85\* |
| 7 | 58 | 11\* |
| 8 | 59 | 2\* |

\*certainty set

TABLE VI

Cumulative Sample Sizes

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 1 | 55 | 55 |
| 2 | 32 | 32 |
| 3 | 18 | 18 |
| 4 | 18 | 18 |
| 5 | 33 | 33 |
| 6 | 25 | 25 |
| 7 | 19 | 128 |
| 8 | 54 | 125\* |
| 9 | 43 | 57\* |
| 10 | 205 | 11\* |

\*certainty set

TABLE VII

Cumulative Sample Sizes

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 1 | 15 | 15 |
| 2 | 0 | 0 |
| 3 | 3 | 3 |
| 4 | 15 | 15 |
| 5 | 2 | 2 |
| 6 | 10 | 10 |
| 7 | 15 | 15 |
| 8 | 15 | 15 |
| 9 | 41 | 312 |
| 10 | 385 | 114\* |

\*certainty set

# Sampling Results

Sampling designs for each stratification design method were tested on the dataset provided. Samples and the samples’ metrics were determined using SAS’s PROC SURVEYSELECT and PROC SURVEYMEANS. Five samples were taken per method using the same five seed values to replicate the same random selection across methods for comparison. Mean, standard deviation, total, standard error, and 95% confidence intervals for both mean and total were calculated and outputted to compare to the actual population mean and total for Sales. Results are shown on Table VIII through X. Only statistics relative to totals are displayed in the tables for space and because the desired prediction statistics is population total.

Sample results reveal a surprising conclusion. The most simplistic stratification method yielded the most accurate and precise predictions of Sales population total. The cumulative method is consistently 2,000,000-4,000,000 less in standard deviation compared to the other more complex methods and all Sales population total predictions fell within the window of the actual population total. Both complex methods contained a sample result where the population total prediction range did not fall within the actual population total (highlighted in red).

# Simple Random Sample

A simple random sample sampling strategy with a sample size of 500 yields prediction values and precision ranges grossly off from the real Sales population total. The importance of stratified sampling is highlighted in these results. Simple random samples make a general assumption that the population distribution follows a normal distribution which is not the case in our dataset. Variable distributions are heavily skewed to the right which explains why predictions are approximately 5% of the actual Sales population total with a wide range of standard deviation precision. Results of the simple random sample with the same seed values are shown in Table XI.

# Comparing Inventory Stratification

In a complementary paper, the Inventory attribute was used to develop a stratified sampling design to predict the Inventory population total. The paper followed the same three methods in developing strata and assigning a Neyman allocation sampling design.

Its outstanding performer was the cumulative square root of frequency. It contained the Inventory population total in its 95% confidence interval for all five random seeds, as was done with the Sales variable. It also had comparable standard deviation with respect to its adversaries.

The choice between top sampling designs based on Sales or Inventory to predict both Sales and Inventory population totals is not as conclusive. As can be expected, the designs in which were modeled based on a certain attribute substantially does better predicting that attribute accurately and precisely. Table XII and Table III show results for each top performing sampling design, predicting both Sales and Inventory. The decision between the two would have to be determined based on the business question that needs to be answered. Full evaluation of all designs can be found in the Appendix.

TABLE XII

Inventory Cumulative Sample Results: Sales & Inventory

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Seed | Total | Standard Deviation | 95% CI Total Min | 95% CI Total Max |
| 91119 | 1,348,756,642 | 45,747,277 | 1,258,872,062 | 1,438,641,222 |
| 91120 | 1,424,313,035 | 62,541,391 | 1,301,431,258 | 1,547,194,812 |
| 91121 | 1,413,509,719 | 41,031,906 | 1,332,889,933 | 1,494,129,505 |
| 91122 | 1,394,883,264 | 47,801,906 | 1,300,961,734 | 1,488,804,795 |
| 91123 | 1,394,527,076 | 46,003,675 | 1,304,138,723 | 1,484,915,428 |
| Sales | 1,388,810,308 |  |  |  |
| 91119 | 1,754,954,823 | 10,238,880 | 1,739,327,957 | 1,779,562,806 |
| 91120 | 1,747,186,709 | 10,200,891 | 1,727,143,924 | 1,767,229,494 |
| 91121 | 1,756,459,181 | 10,649,488 | 1,735,534,989 | 1,777,383,373 |
| 91122 | 1,753,740,854 | 9,938,577 | 1,734,213,467 | 1,773,268,241 |
| 91123 | 1,773,269,453 | 10,218,702 | 1,753,191,673 | 1,793,347,233 |
| Inventory | 1,754,954,823 |  |  |  |

TABLE XI

Sales Cumulative Sample Results: Sales & Inventory

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Seed | Total | Standard Deviation | 95% CI Total Min | 95% CI Total Max |
| 91119 | 1,379,235,501 | 10,269,590 | 1,359,057,634 | 1,399,413,368 |
| 91120 | 1,387,966,563 | 10,175,115 | 1,367,976,288 | 1,407,956,838 |
| 91121 | 1,387,599,506 | 10,849,340 | 1,366,282,538 | 1,408,916,475 |
| 91122 | 1,378,281,016 | 9,638,689 | 1,359,342,755 | 1,397,219,277 |
| 91123 | 1,392,994,198 | 9,848,483 | 1,373,643,730 | 1,412,344,667 |
| Sales | 1,388,810,308 |  |  |  |
| 91119 | 1,813,713,699 | 108,514,858 | 1,600,501,847 | 2,026,925,550 |
| 91120 | 1,744,775,114 | 47,587,619 | 1,651,274,144 | 1,838,276,083 |
| 91121 | 1,734,157,057 | 50,555,055 | 1,634,825,619 | 1,833,488,495 |
| 91122 | 1,834,437,310 | 71,462,579 | 1,694,026,411 | 1,974,848,209 |
| 91123 | 1,687,322,220 | 45,968,445 | 1,597,002,631 | 1,777,641,809 |
| Inventory | 1,754,954,823 |  |  |  |

# Conclusion

TABLE XI

Equal Sample Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Seed | Total | Standard Deviation | 95% CI Total Min | 95% CI Total Max |
| 91119 | 58,310,200 | 4,206,700 | 50,045,173 | 66,575,227 |
| 91120 | 78,847,701 | 17,321,642 | 44,815,361 | 112,880,040 |
| 91121 | 68,804,062 | 6,413,818 | 56,202,646 | 81,405,479 |
| 91122 | 82,944,694 | 10,543,967 | 62,228,652 | 103,660,735 |
| 91123 | 78,621,608 | 13,088,070 | 52,907,093 | 104,336,124 |
| Actual | 1,388,810,308 |  |  |  |

In most cases, Sales trumps Inventory as it can be a better predictor for an industries financial health. In that light, the cumulative model based on Sales would be the best option. However, in some business cases the question of interest in total volume, in which case cumulative square root of frequency for Inventory would be the preferred sampling design.

The accuracy is not lost in either model for predicting either attribute. There is simply a significant lose in precision depending the attribute of interest. Further work can be done to find a combined metric that may do its best in predicting both variables with a similar level of precision. Separating strata based on a value the is a linear combination of the two variables, reducing the variation between the two, may be an optimal option worth investigation.

REFERENCES

[1] U. Kretzschmar. (2009, July). AES128 – A C Implementation for Encryption and Decryption. [Online].

[2] W. Stallings. (2014). Cryptography and Network Security – Principles and Practice, Sixth Edition.

1. [↑](#footnote-ref-1)