Kaitlin Kirasich and Dennis Murray

MSDS-6370 Project 1

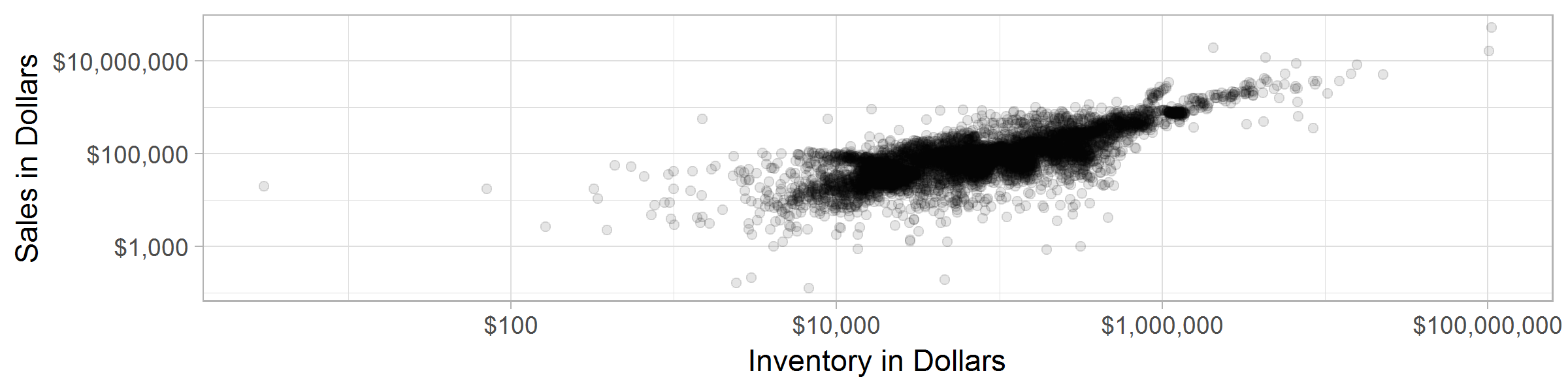
Stratification on Inventory

Introduction

For our project, we explored a simulated population of one month’s data for an industry. This data contained two variables for consideration, sales and inventory. For my portion of the project, I explored the stratification of the inventory variable to design a sampling plan for a sample size of 500 to be selected for sales and inventory data collection each month for the next two years.

Description of Variables

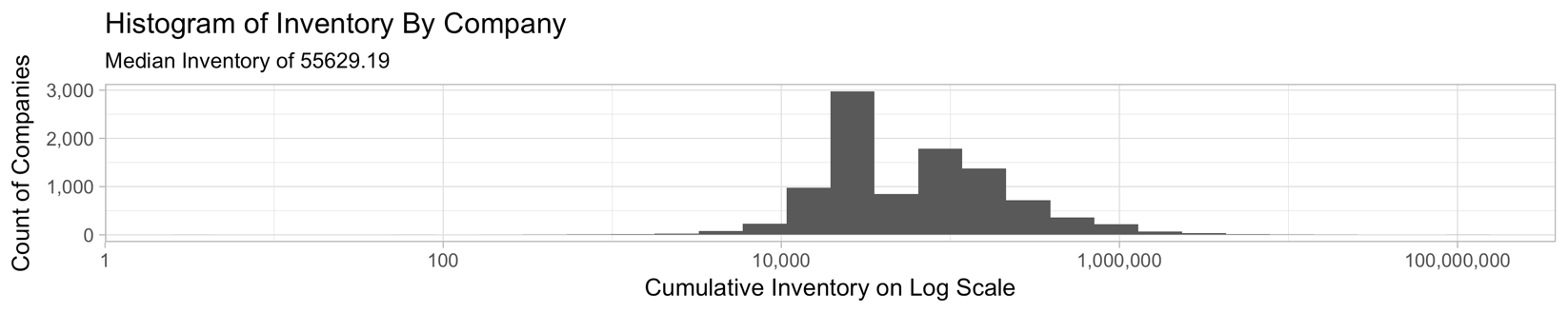
The data set contains the sales and inventory value for 9,762 companies. A scatterplot, and correlation analysis, both reveal a high level in interrelation between the two variables.



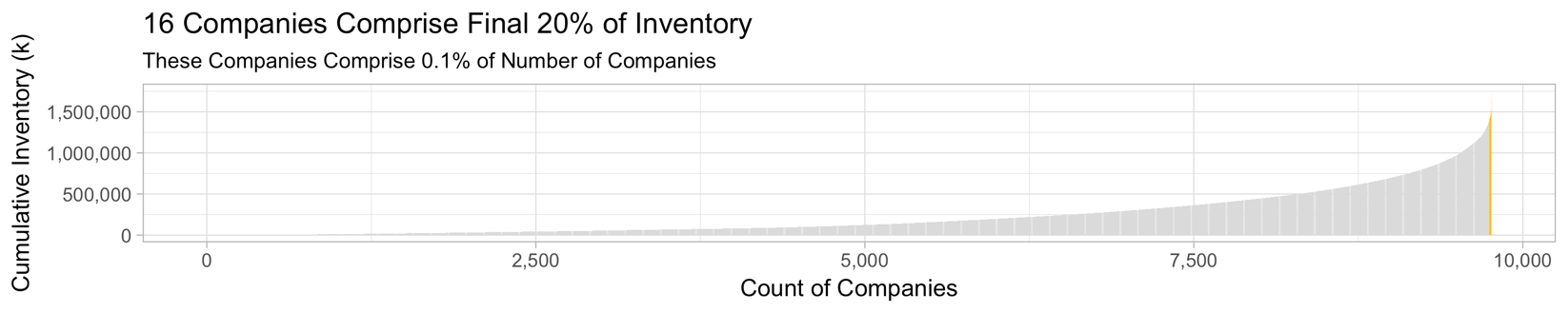
In fact, the correlation of the two variables is found to be 0.83. There is reason to believe that many of the observations would fall into similar strata no matter which variable is used, however, some of the examples will shift between strata.

Certainty Stratum – Methodology and Selection

Exploration of the data found useful information about the shape of the distribution of the data set for Inventory by Company. Among the 9,762 companies included, the Mean Inventory was 179,774 units, but with a much lower median of 55,629. This indicates a long tail distribution – with several outliers on the upper end that increase the mean and the standard deviation of the inventory distribution.



Considering this information, we investigated and found that a small number of the companies comprise a large share of the overall inventory. In this case, a certainty stratum that comprises a large share of the overall value might allow a better insight into the final result. In a non-probability study, this could be done as a cut-off sample – but with the proposed survey design for this study, we propose that the companies providing the final 20% of cumulative inventory to be taken as a certainty stratum.



We found that the top 16 companies make up 20% of the sum of total inventory in the population. This will allow for 484 of the 500 samples to be allocated to the remainder of the population, 9,693 companies. This gives an overall probability of selection for the remaining companies of 0.0499 from the overall probability of selection of 0.051 (500/9,762).

Stratum Creation and Break Points

With the certainty stratum determined, the next consideration is the creation of the stratum and the allocation of sampling units by stratum. The allocation method has already been determined: Neyman allocation, which will consider weight the stratum’s contribution to total variability of the sample, and then move more samples to the stratum with the largest contribution to variability. Based on our preliminary exploration, we would expect more of the allocation to be given to the upper end of the spectrum, as the large number of companies with inventory less than the mean offer a smaller share of the overall variability of the population.

By looking at the histogram of inventory, this helped to determine some idea of the number of stratum required. It is clear that the long tail at the upper end of the inventory spectrum is deserving of its own certainty stratum – and creating this certainty stratum provides for all companies with inventory greater than 100Mil to be included in the final sample. Considering the remainder of the companies in the histogram, there is a clear median value, but then three higher frequency bins to either side of the median. The lower tail and the space between the upper median range and the certainty stratum could be considered another group. In total, we might consider 8 stratum, including the certainty stratum. To test our theory, we tried out 6 through 11 stratum and eventually landed on 8 strata because it provided the lowest standard error.

Determination of Stratum

In order to determine the break points and number of stratum, we went with the Cumulative Method. This method sorts the value of inventory from least to greatest and then sums on a cumulative basis moving from least to greatest individual companies. The final cumulative sum is divided by our number of strata. At each 1/Nth cumulative sum, a break point occurs.

The R library *stratification* was used to assist in the calculation of strata boundaries, assignment to strata, and calculation of the Neyman allocation by strata. The function *strata.cumrootf()*  allows for creation of strata based on the cumulative root frequency schema devised by Dalenius and Hodges. The full dataset of sales values was loaded and executed in R, using a loop to consider all counts of strata between six and eleven (including the certainty strata.) The output of the process provided a strata assignment for each company in the dataset, and the resulting assignment was paired with the remainder of the data for processing of the selection via SAS.

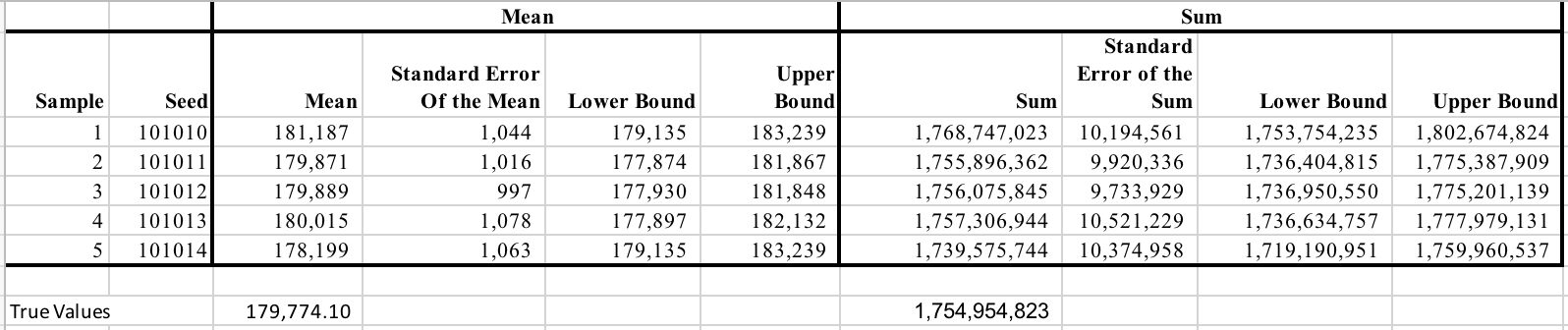
The values for the number of strata were tested by running a single random sample and the *PROC SURVEYMEANS* function in SAS. As stated above, the 8 strata sample returned the lowest standard error of the sum.

Survey Selection

Survey selection was performed in SAS using the *PROC SURVEYMEANS* function. This task was performed five times, each time with a new value for the random seed. The initial seed value for the first sample was set to 101010 and incremented by 1 for each additional seed.

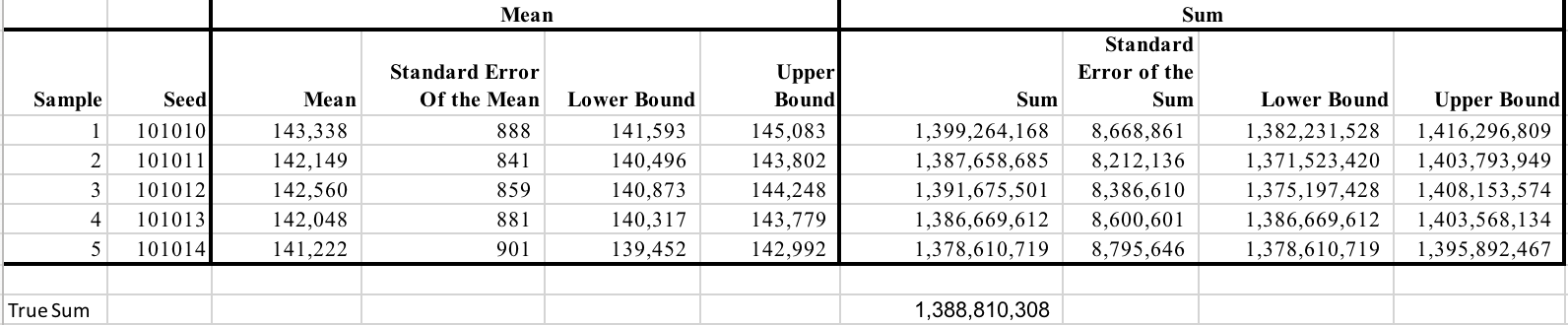
Results

As mentioned, the *PROC SURVEYMEANS* was run five times, each with a unique seed value for reproducibility. The results are shown in the table below.



Across all samples, the true sum, 1,754,954,823 was included in the confidence interval for the sum, and the absolute difference between the actual total inventory and the estimated total inventory from the sample is less than 1% in all cases.

We actually saw very similar results when running the exact same sampling design with the sales variable. Dennis’s results for the sales stratification is shown below:



In his sample, he also found 8 as the ideal number of strata (plus the certainty strata). This is not surprising considering the high correlation between the variables. In his sampling, we was also able to capture the true sum of sales in all 5 samples.

For both inventory and sales, seed 101011 proved to estimate a sum closest to the true sum value (less than 0.01% absolute difference). Because estimating the sales has a higher priority than estimating the inventory, I used my inventory stratification to try to estimate total sales (using our best seed 101011). In this case, the estimate of sales turned out to be 1,349,489,489. The confidence interval does capture the true sum of sales. However, the absolute difference between this estimate and the true values is 2%. This is by no means bad but it does indicate that stratifying by sales will provide a closer accuracy to true sum of sales than stratifying by inventory. Because correlation of variables is so high at 0.83, we believe that by stratifying by inventory or sales will give a good estimate. However, stratification by sales will provide slightly better estimates if the sales estimation is a higher priority than inventory estimation.