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. Figure 1 below shows a scatterplot, and correlation analysis, both reveal a high level in interrelation between the two variables.

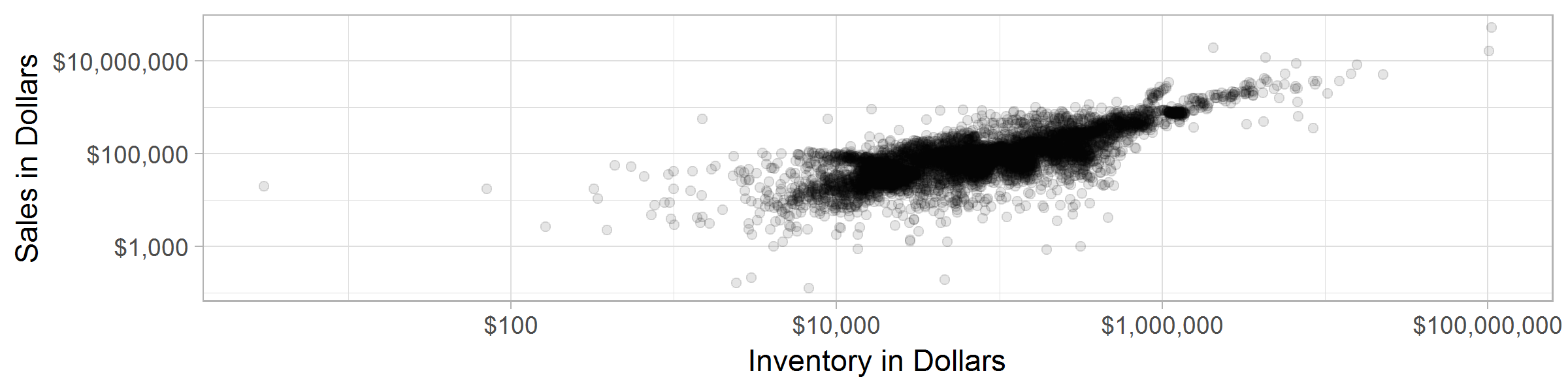


Figure 1: Scatterplot of Sales vs Inventory

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. Figure 2 is a histogram of the inventory by company on a log scale. It is hard to see with the width but there is a long tail after 1,000,000 and a small pocket at 100,000,000.

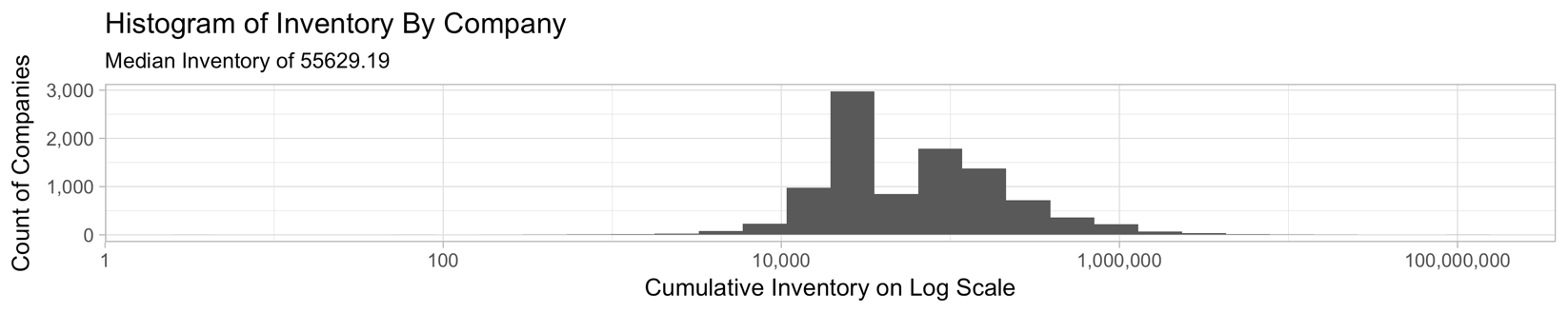


Figure 2: Histogram of cumulative inventory on a log scale

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.

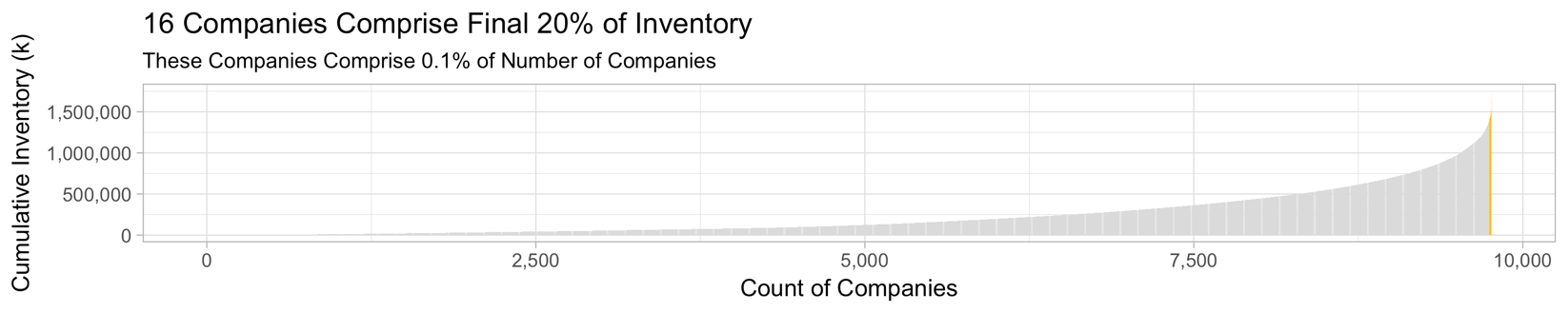


Figure 3: Distribution of the cumulative inventory (x1000) and the count of companies

In Figure 3, you may be able to see the gold sliver at the end to show the top 16 companies that 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 Figure 4 below.

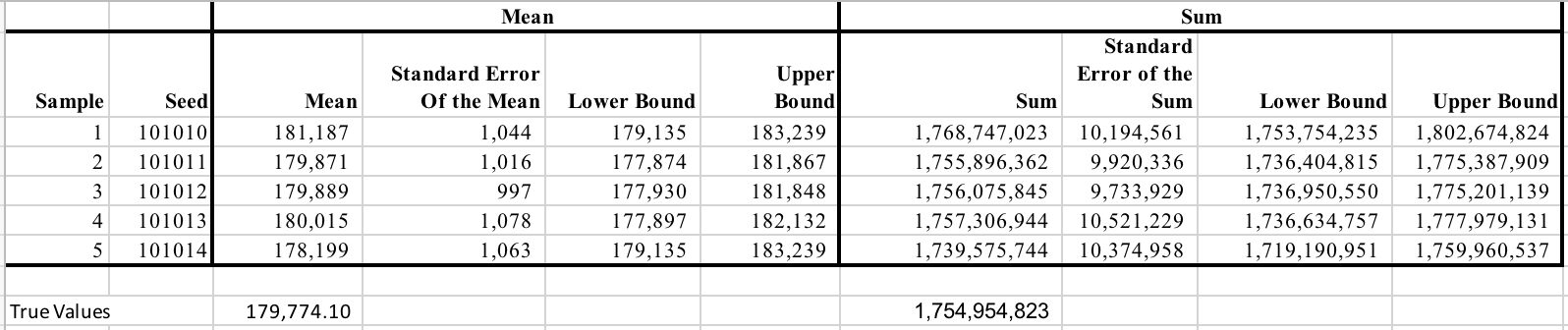


Figure 4: Results of stratification by inventory

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 in Figure 5 below:

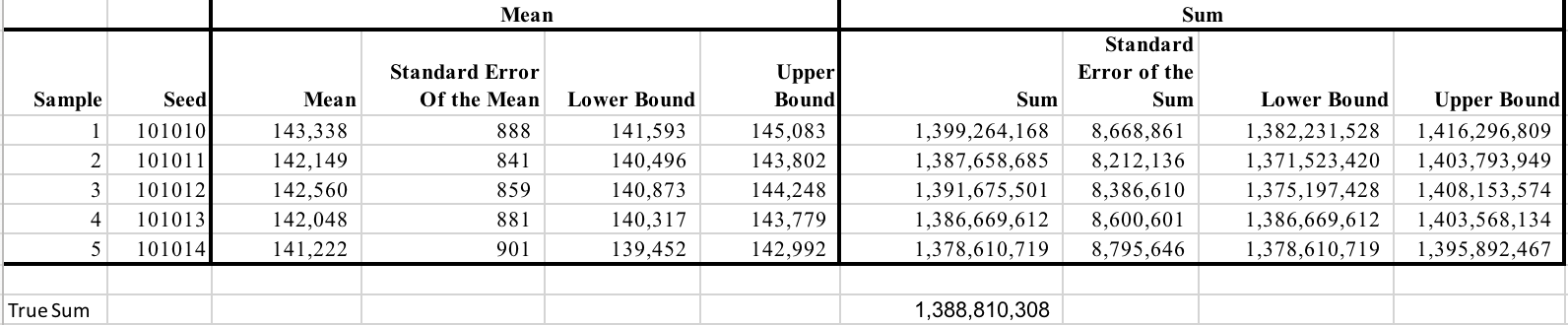


Figure 5: Results of stratification by sales

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.

We ran this one more time, this time removing the stratification and doing a plain simple random sample. The estimation for the sum of sales was 61,887,665 compared to the true sum of 1,388,810,308. This estimation is completely off due to the fact that the 500 randomly chosen samples were not verified as an accurate sample landscape of our entire population.

Conclusion

When deciding whether to stratify our data or not, we found that it is much better to stratify than not because our data was very skewed. When we did not stratify, the sum was underestimated and we found that the results from a stratified sample design typically captured a value close to the true sum. 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.

I gained so much hands-on experience with stratification in this project. Manually determining strata in excel was very difficult, especially with so many data points. By stratifying in R, I was able to feel confident that my results were correct and did not suffer from manual errors. Another key thing that I learned was how important it is to understand your data and variables. Knowing that our variables were highly correlated and had a few large outliers helped to pick a good number of stratum and comprise our certainty strata.

Appendix 1: R Code for Strata Construction

Note: All code is available in R Markdown Form, with visualizations in the Project Repositiory: <https://github.com/dpmurraygt/MSDS6370Project>

library(dplyr)

library(ggplot2)

library(tidyr)

library(magrittr)

library(stratification)

library(readxl)

library(ggthemes)

ProjectData <- read\_xls("Data/ProjectData.xls")

#Parameters

#Certainty Strata Size - In terms of percent of total sales

CertaintyPct <- 0.2

#Number of Non-Certainty Strata

NumberOfStrata <- 6

#Total Sample, including Certainty

TotalSampleSize <- 500

#designate the top 20% of cumulative sales as certainty strata

ProjectData %<>% arrange(inventory) %>% mutate(CumInv = cumsum(inventory)) %>% mutate(cumPctInv = CumInv/sum(inventory)) %>% mutate(Certainty=ifelse(cumPctInv>0.8, TRUE, FALSE))

CertaintyStrata <- ProjectData %>% mutate(Index=row\_number()) %>% filter(Certainty==TRUE) %>% select(Index)

#Remove the certainty part, what's left?

NonCertaintySampleSize <- 500 - nrow(CertaintyStrata)

StrataSizes <- c(5:10)

AllStratification <- list()

#make a filtered version of the data frame minus certainty strata that we will match up with the strata assignment

Limited <- ProjectData %>% filter(Certainty==FALSE) %>% select(coID, inventory, sales)

TheCertainty <- ProjectData%>% filter(Certainty==TRUE) %>% select(coID, inventory, sales) %>% mutate(strata="Certainty")

for (size in StrataSizes){

ThisStratification<-strata.cumrootf(x=ProjectData$inventory, n=TotalSampleSize, Ls = size, certain = CertaintyStrata$Index)

print(ThisStratification)

#stratum file output

filename <- paste("StratumAssign/InventoryWSalesStratumAssignment", size, ".csv", sep="")

temp<-cbind(Limited, strata=as.numeric(ThisStratification$stratumID))

TheCertainty <- ProjectData%>% filter(Certainty==TRUE) %>% select(coID, inventory, sales) %>% mutate(strata=size+1)

#Append Back in the certainty part

temp<- rbind(temp, TheCertainty)

write.csv(temp, file = filename)

}

Appendix 2: SAS Code

/\* 6 Stratum + Certainty\*/

PROC IMPORT

DATAFILE="\\Client\C$\Users\kkirasich\Documents\GitHub\MSDS6370Project\StratumAssign\InventoryStratumAssignment7.xlsx"

OUT = invData

REPLACE

DBMS = XLSx;

run;

PROC SURVEYSELECT DATA = invData out = srsSelect sampsize = (109, 54, 63, 84, 85, 89, 16) seed =101010 stats;

strata strata;

title "Seven Strata + certainty Sample Selection";

run;

proc print data=srsSelect;

run;

data strsizes;

input strata \_total\_;

datalines;

1 109

2 54

3 63

4 84

5 85

6 89

7 16

;

PROC SURVEYMEANS DATA = srsSelect sum clsum total = strsizes;

var inventory;

strata strata;

weight SamplingWeight;

title "6 Strata + Certainty, cum(f) and Neyman Allocation";

/\* 7 Stratum + Certainty\*/

PROC IMPORT

DATAFILE="\\Client\C$\Users\kkirasich\Documents\GitHub\MSDS6370Project\StratumAssign\InventoryStratumAssignment7.xlsx"

OUT = invData

REPLACE

DBMS = XLSx;

run;

PROC SURVEYSELECT DATA = invData out = srsSelect sampsize = (101, 53, 77, 51, 71, 68, 63, 16) seed =101010 stats;

strata strata;

title "Seven Strata + certainty Sample Selection";

run;

proc print data=srsSelect;

run;

data strsizes;

input strata \_total\_;

datalines;

1 5094

2 2005

3 1462

4 612

5 356

6 190

7 63

8 16

;

PROC SURVEYMEANS DATA = srsSelect sum clsum total = strsizes;

var inventory;

strata strata;

weight SamplingWeight;

title "7 Strata + Certainty, cum(f) and Neyman Allocation";

/\* Below this line is 8 plus Certainty \*/

DATAFILE="\\Client\C$\Users\kkirasich\Documents\GitHub\MSDS6370Project\StratumAssign\InventoryStratumAssignment8.xlsx"

OUT = invData8

REPLACE

DBMS = XLSx;

run;

PROC SURVEYSELECT DATA = invData8 out = srsSelect8 sampsize = (91, 52, 30, 49, 67, 67, 63, 65, 16) seed =101010 stats;

strata strata;

title "Eight Strata + certainty Sample Selection";

run;

proc print data=srsSelect;

run;

data strsizes;

input strata \_total\_;

datalines;

1 4848

2 1932

3 1012

4 834

5 587

6 296

7 172

8 65

9 16

;

PROC SURVEYMEANS DATA = srsSelect8 mean cl sum clsum total = strsizes;

var inventory;

strata strata;

weight SamplingWeight;

title "8 Strata + Certainty, cum(f) and Neyman Allocation";

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* SEED + 1 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

PROC SURVEYSELECT DATA = invData8 out = srsSelect8 sampsize = (91, 52, 30, 49, 67, 67, 63, 65, 16) seed =101011 stats;

strata strata;

title "Eight Strata + certainty Sample Selection";

run;

data strsizes;

input strata \_total\_;

datalines;

1 4848

2 1932

3 1012

4 834

5 587

6 296

7 172

8 65

9 16

;

PROC SURVEYMEANS DATA = srsSelect8 mean cl sum clsum total = strsizes;

var inventory;

strata strata;

weight SamplingWeight;

title "8 Strata + Certainty, cum(f) and Neyman Allocation";

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* SEED + 1 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

PROC SURVEYSELECT DATA = invData8 out = srsSelect8 sampsize = (91, 52, 30, 49, 67, 67, 63, 65, 16) seed =101012 stats;

strata strata;

title "Eight Strata + certainty Sample Selection";

run;

data strsizes;

input strata \_total\_;

datalines;

1 4848

2 1932

3 1012

4 834

5 587

6 296

7 172

8 65

9 16

;

PROC SURVEYMEANS DATA = srsSelect8 mean cl sum clsum total = strsizes;

var inventory;

strata strata;

weight SamplingWeight;

title "8 Strata + Certainty, cum(f) and Neyman Allocation";

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* SEED + 1 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

PROC SURVEYSELECT DATA = invData8 out = srsSelect8 sampsize = (91, 52, 30, 49, 67, 67, 63, 65, 16) seed =101013 stats;

strata strata;

title "Eight Strata + certainty Sample Selection";

run;

data strsizes;

input strata \_total\_;

datalines;

1 4848

2 1932

3 1012

4 834

5 587

6 296

7 172

8 65

9 16

;

PROC SURVEYMEANS DATA = srsSelect8 mean cl sum clsum total = strsizes;

var inventory;

strata strata;

weight SamplingWeight;

title "8 Strata + Certainty, cum(f) and Neyman Allocation";

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* SEED + 1 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

PROC SURVEYSELECT DATA = invData8 out = srsSelect8 sampsize = (91, 52, 30, 49, 67, 67, 63, 65, 16) seed =101014 stats;

strata strata;

title "Eight Strata + certainty Sample Selection";

run;

data strsizes;

input strata \_total\_;

datalines;

1 4848

2 1932

3 1012

4 834

5 587

6 296

7 172

8 65

9 16

;

PROC SURVEYMEANS DATA = srsSelect8 mean cl sum clsum total = strsizes;

var inventory;

strata strata;

weight SamplingWeight;

title "8 Strata + Certainty, cum(f) and Neyman Allocation";

/\*get descriptive stats, distributions of sales and inventory \*/

proc means data = invdata;

var sales inventory;

run;

proc univariate data = salesdata;

var inventory;

histogram sales/kernel;

inset n='Number of Companies';

label LoanType = 'Type of Loan';

options gstyle;

run;

/\*get descriptive stats, distributions of sales and inventory \*/

proc means data = salesdata;

var sales inventory;

run;

proc univariate data = salesdata;

var sales;

histogram sales/kernel;

inset n='Number of Companies';

label LoanType = 'Type of Loan';

options gstyle;

run;