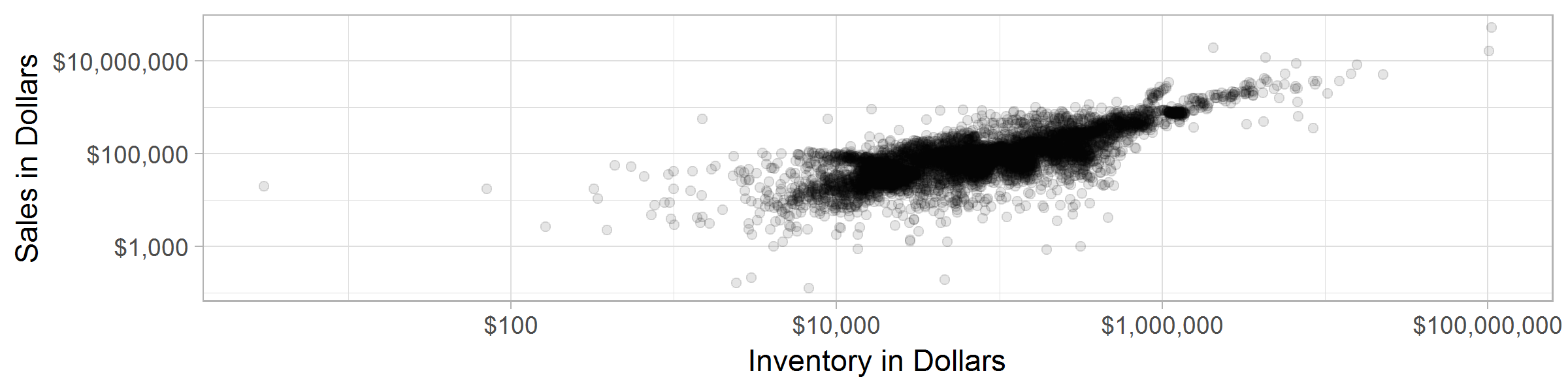
consideration of stratification variables: sales

Description of Variables

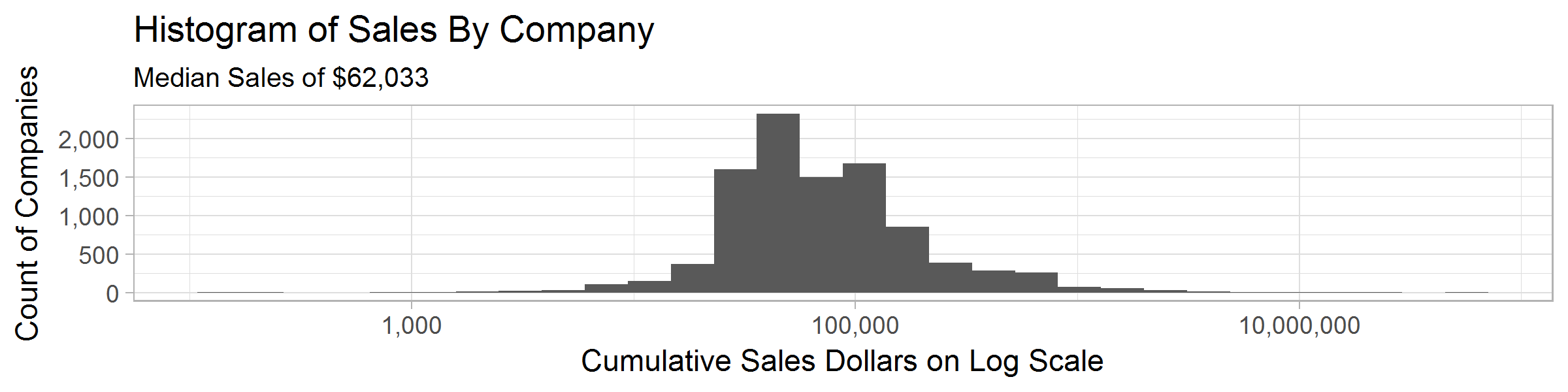
The data set contains the sales and inventory value for 9,762 companies. A scatterplot, and correlation analysis, both reveal a high level in interrelation between the two variables.



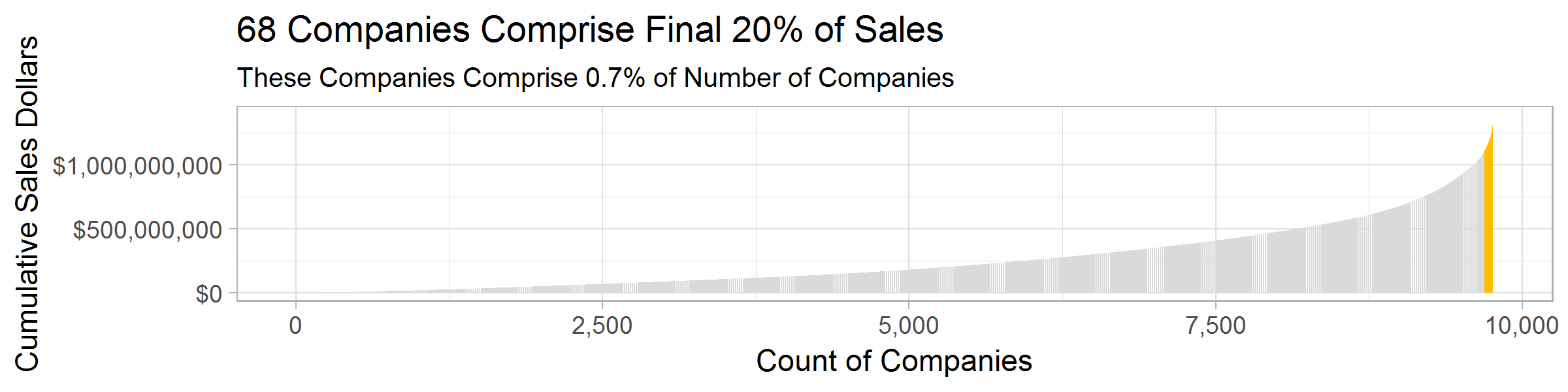
In fact, the correlation of the two variables is found to be 0.83. There is reason to believe that many of the observations would fall into similar strata no matter which variable is used, however, some of the examples will shift between strata.

Certainty Stratum – Methodology and Selection

Exploration of the data found useful information about the shape of the distribution of the data set for Sales by Company. Among the 9,762 companies included, the Mean Sales was $142,267, but with a much lower median - $62,033. This indicates a long tail distribution – with several outliers on the upper end of sales that are likely increasing the mean as well as the standard deviation of the sales value.



Considering this information, we investigated the possibility that a small number of the companies comprises a large share of the overall sales. In any scenario where the variables of interest – in this case, the total sales of the population – are of interest, 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 sales to be taken as a certainty stratum.



This comprises 68 companies, but which sell 20% of the overall value of all companies in the population. This will allow for 432 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 0.045. Overall probability of selection is 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 sales less than the mean offer a smaller share of the overall variability of the population.

Considering the histogram of sales helps to determine some idea of the number of stratum required. It is clear that the long tail at the upper end of the sales spectrum is deserving of its own certainty stratum – and creating this certainty stratum provides for all companies with sales greater than $1.818MM 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 – but will test two different counts of stratum and determine the best method based on the expected variance produced.

Two methods of the determining the break points, and number of stratum, were considered. The first method considered was the Cumulative Method. This method considers the value of Sales, arranged from least to greatest, and then summed on a cumulative basis moving from least to greatest individual companies. Then this is broken into equal size strata.

Determination of Stratum

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 al 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.

Two values for the number of strata were tested via running a single random sample and the *PROC SURVEYMEANS* function in SAS: seven and eight strata. The eight strata sample returned a slightly improved standard error of the sum, and the higher number of strata were selected for the full analysis process.

Stratification Breakpoints and Frequency

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Stratum | Low | High | Nh | nh | fh |
| 1 | $ 124 | $ 30,374 | 1,979 | 33 | 0.0167 |
| 2 | $ 30,409 | $ 60,689 | 3,286 | 63 | 0.0192 |
| 3 | $ 60,725 | $ 106,090 | 2,181 | 58 | 0.0266 |
| 4 | $ 106,177 | $ 166,597 | 1,187 | 48 | 0.0404 |
| 5 | $ 166,704 | $ 287,794 | 539 | 58 | 0.1076 |
| 6 | $ 287,963 | $ 499,802 | 338 | 64 | 0.1893 |
| 7 | $ 499,836 | $ 802,102 | 184 | 108 | 0.5870 |
| 8 (Certainty) | $ 803,927 | $ 52,156,428 | 68 | 68 | 1.0000 |

The largest stratum is #2, which covers companies from $30,409 to $60,689 (3,286 companies) but receives a Neyman allocation for a probability of selection of 0.0192. The highest probability of selection (besides the certainty) is also the largest absolute number of companies selected, from stratum 7, has a 0.587 probability of selection and selects 108 companies of the total 184 in the range. Stratum 7 runs from $499,836 to $802,102.

Survey Selection

Survey selection was performed in SAS using the *PROC SURVEYMEANS* function. This task was performend 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 mean, 1,388,810,308 was included in the confidence interval for the sum, and the absolute difference between the actual total sales and the estimate from the sample is less than 1% in all cases, with a maximum difference of 0.75% and a minimum difference of 0.21%.

Discussion and Conclusions

The outcomes of our inventory stratification was similar – with all five samples having confidence intervals that included the true mean. We also were able to sample the sales value based on the stratification of inventory, and still were within about 2% of the true value of the sum of sales.

This points to a conclusion – that with these two highly correlated variables, we could expect to receive equal outcomes no matter which variable is selected for stratification. With the scatterplot shown in the exploration section, most of the observations lie close to a line that would describe a fixed relationship between sales and inventory. In this case, both sales and inventory would probably be equally likely to provide an insight to the businesses in the population.

This leads to my first insight on building stratification levels and definitions: a thorough understanding and exploration of the data is needed to understand the data, and the population it describes, before attempting to build stratification levels. While it was not a part of this project, the thorough discussion of objectives, and the population, with stakeholders and business experts would help the statistician in designing strata that will fit the sample appropriately.

The second conclusion is that some aspect of simulation will help provide insight to the best stratification design. In our case, the decision was made between seven and eight strata, including certainty, was led by the simulation of one sample for seven and eight strata of the selection and the slight improvement in standard error. This is certainly a “big data” influenced method – the cost, in currency and time, of simulation is very low now and can provide substantial insight to the data set.

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 <- 5

#Total Sample, including Certainty

TotalSampleSize <- 500

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

ProjectData %<>% arrange(sales) %>% mutate(CumSales = cumsum(sales)) %>% mutate(cumPctSales = CumSales/sum(sales)) %>% mutate(Certainty=ifelse(cumPctSales>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, sales)

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

for (size in StrataSizes){

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

print(ThisStratification)

#stratum file output

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

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

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

#Append Back in the certainty part

temp<- rbind(temp, TheCertainty)

write.csv(temp, file = filename)

}

Appendix 2: SAS Code

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

/\*Import the data from local directory, place into a table called salesData\*/

PROC IMPORT

/\* DATAFILE="\\Client\C$\Users\dmurray\Documents\SMU\MSDS 6370 Sampling\Project\MSDS6370Project\Data\projectData.xls" \*/

DATAFILE="\\Client\C$\Users\dmurray\Documents\SMU\MSDS 6370 Sampling\Project\MSDS6370Project\StratumAssign\SalesStratumAssignment8.xlsx"

OUT = salesData

REPLACE

DBMS = XLSx;

run;

PROC SURVEYSELECT DATA = salesData out = srsSelect sampsize = (22, 55, 62, 52, 48, 53, 55, 85, 68) seed =101010 stats;

strata strata;

title "Eight Strata + certainty Sample Selection";

run;

proc print data=srsSelect;

run;

data strsizes;

input strata \_total\_;

datalines;

1 1282

2 3538

3 2167

4 1352

5 584

6 381

7 270

8 120

9 68

;

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

var sales;

strata strata;

weight SamplingWeight;

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

PROC SURVEYSELECT DATA = salesData out = srsSelect sampsize = (22, 55, 62, 52, 48, 53, 55, 85, 68) seed =101011 stats;

strata strata;

title "Eight Strata + certainty Sample Selection Run 2";

run;

proc print data=srsSelect;

run;

data strsizes;

input strata \_total\_;

datalines;

1 1282

2 3538

3 2167

4 1352

5 584

6 381

7 270

8 120

9 68

;

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

var sales;

strata strata;

weight SamplingWeight;

title "8 Strata + Certainty, cum(f) and Neyman Allocation Run 2";

PROC SURVEYSELECT DATA = salesData out = srsSelect sampsize = (22, 55, 62, 52, 48, 53, 55, 85, 68) seed =101012 stats;

strata strata;

title "Eight Strata + certainty Sample Selection Run 3";

run;

proc print data=srsSelect;

run;

data strsizes;

input strata \_total\_;

datalines;

1 1282

2 3538

3 2167

4 1352

5 584

6 381

7 270

8 120

9 68

;

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

var sales;

strata strata;

weight SamplingWeight;

title "8 Strata + Certainty, cum(f) and Neyman Allocation Run 3";

PROC SURVEYSELECT DATA = salesData out = srsSelect sampsize = (22, 55, 62, 52, 48, 53, 55, 85, 68) seed =101013 stats;

strata strata;

title "Eight Strata + certainty Sample Selection Run 4";

run;

proc print data=srsSelect;

run;

data strsizes;

input strata \_total\_;

datalines;

1 1282

2 3538

3 2167

4 1352

5 584

6 381

7 270

8 120

9 68

;

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

var sales;

strata strata;

weight SamplingWeight;

title "8 Strata + Certainty, cum(f) and Neyman Allocation Run 4";

PROC SURVEYSELECT DATA = salesData out = srsSelect sampsize = (22, 55, 62, 52, 48, 53, 55, 85, 68) seed =101014 stats;

strata strata;

title "Eight Strata + certainty Sample Selection Run 5";

run;

proc print data=srsSelect;

run;

data strsizes;

input strata \_total\_;

datalines;

1 1282

2 3538

3 2167

4 1352

5 584

6 381

7 270

8 120

9 68

;

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

var sales;

strata strata;

weight SamplingWeight;

title "8 Strata + Certainty, cum(f) and Neyman Allocation Run 5";

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;