Abstract:

Two samplings of the monthly industry data using methods based on two variables, Sales and Inventory. Stratification using cumulative square root method. Allocation of samples to strata based on Neyman Allocation. Comparison of the results of the sampling methods.

Purpose:

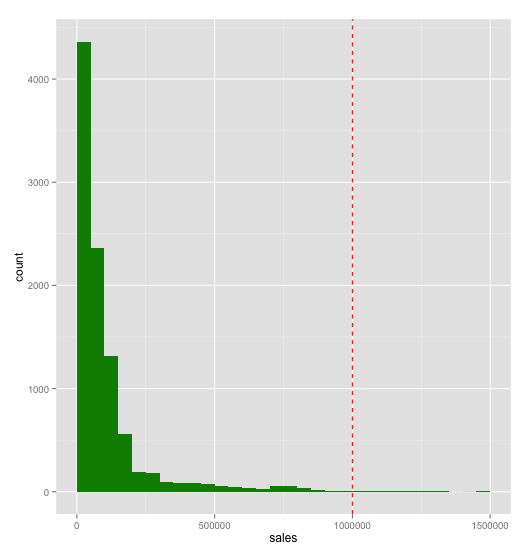
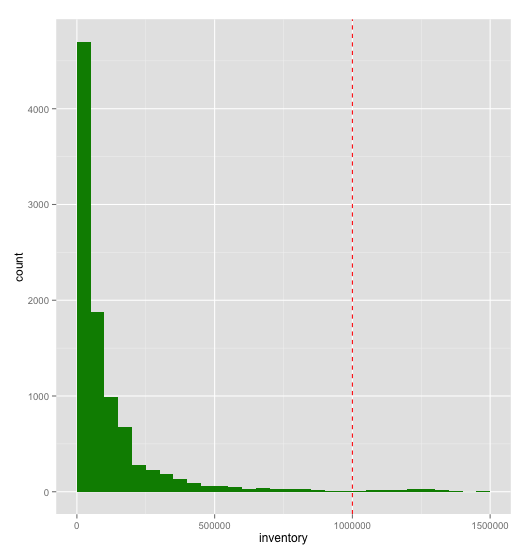
To create and execute a sampling plan for monthly data collected in Industry3. The purpose of the data collection is to estimate the total Sales and total Inventory for the entire population each month. The quality of the estimate of total Sales has priority over the estimate of total Inventory although both are important.

Background Analysis:

The monthly data has three variables: UnitID, Sales, Inventory.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | N | Std Dev | Mean |
| sales | 9762 | 665959.08 | 142266.99 |
| inventory | 9762 | 1573357.61 | 179774.11 |

Appendix: 1.1

Histograms of Inventory and Sales

There are very strong skews and long tails in both variables. The Pearson’s R correlation coefficient is .82 so we expect sales and inventory to have similar distributions.

Sampling Plan:

The tale of the data is very sparse with very large values. This is an important consideration when forming stratum for sampling. The tail of the data contains the largest companies in Industry3. To account for these characteristics and improve the accuracy of our monthly estimates we are going to make a certainty strata that we will sample with a probability of 1.

Stratification By Sales

We are going to start with the Sale variable.

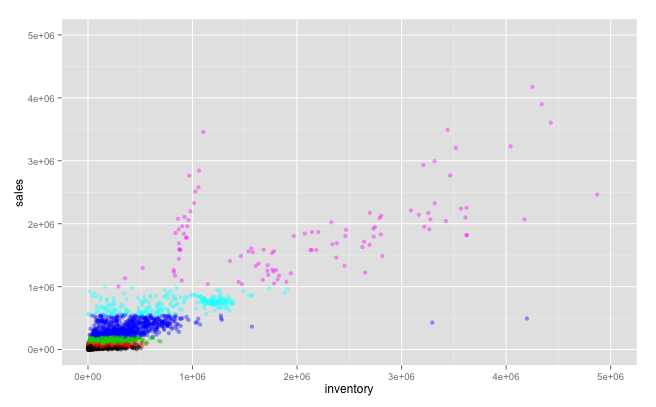
|  |
| --- |
| Cummulative Square Root F Table For Sales |
| lower upper frequency sqrtf csqrtf stratum |
| 1 0 50000 4362 66.045439 66.04544 1 |
| 2 50000 100000 2360 48.579831 114.62527 2 |
| 3 100000 150000 1313 36.235342 150.86061 3 |
| 4 150000 200000 557 23.600847 174.46146 3 |
| 5 200000 250000 189 13.747727 188.20919 4 |
| 6 250000 300000 183 13.527749 201.73694 4 |
| 7 300000 350000 92 9.591663 211.32860 4 |
| 8 350000 400000 86 9.273618 220.60222 4 |
| 9 400000 450000 87 9.327379 229.92960 4 |
| 10 450000 500000 76 8.717798 238.64739 4 |
| 11 500000 550000 54 7.348469 245.99586 5 |
| 12 550000 600000 42 6.480741 252.47660 5 |
| 13 600000 650000 36 6.000000 258.47660 5 |
| 14 650000 700000 28 5.291503 263.76811 5 |
| 15 700000 750000 51 7.141428 270.90954 5 |
| 16 750000 800000 55 7.416198 278.32573 5 |
| 17 800000 850000 38 6.164414 284.49015 5 |
| 18 850000 900000 16 4.000000 288.49015 5 |
| 19 900000 950000 7 2.645751 291.13590 5 |
| 20 950000 1000000 5 2.236068 293.37197 5 |
| Certainty Stratum > 1000000 125 6 |

Appendix 2.1

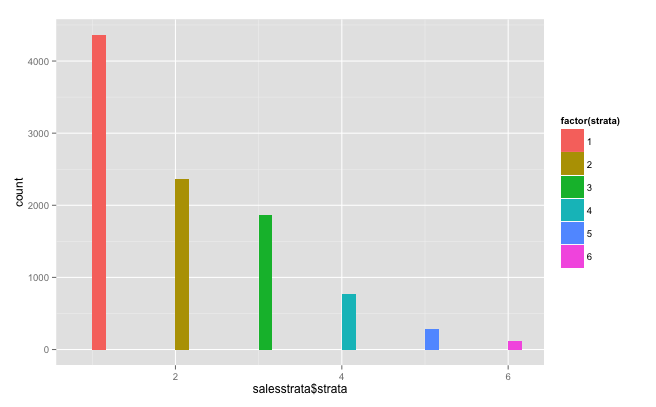
Cutoff points for strata:

|  |  |  |  |
| --- | --- | --- | --- |
| 58.67439 | 117.34879 | 176.02318 | 234.69757 |

Industry 3 Stratified By Sales



Histogram of Sales Strata

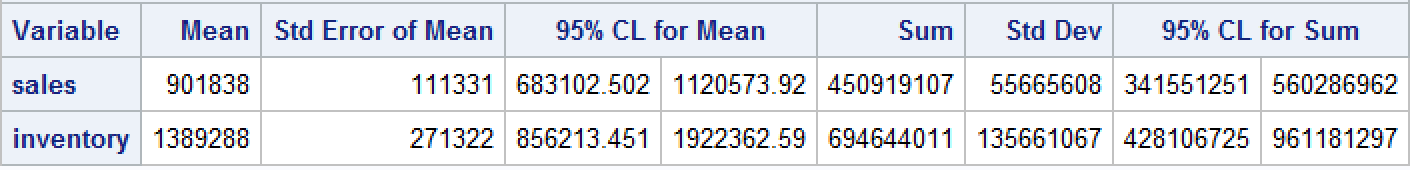


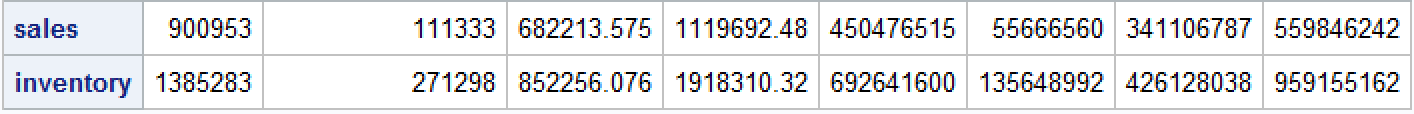
Using Neyman Allocation to find the how many we are selecting from each strata for a sample size of 500.

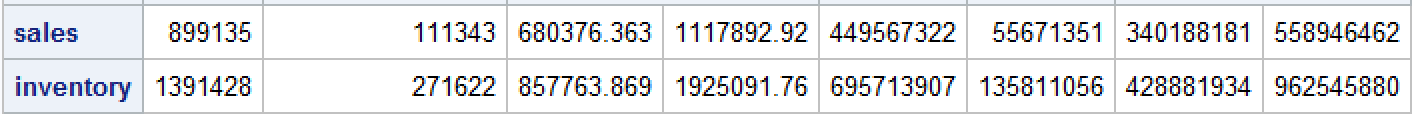
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Neyman Allocation |  |  |  |  |
| Stratum | Nh | Sh=StdDev | NhSh | nh | nh rounded |
| 1 | 4362 | 9838.31 | 42914708.22 | 70.44820185 | 70 |
| 2 | 2360 | 14274.97 | 33688929.2 | 55.30328838 | 55 |
| 3 | 1870 | 25762.06 | 48175052.2 | 79.08351104 | 79 |
| 4 | 767 | 98121.44 | 75259144.48 | 123.5443889 | 124 |
| 5 | 278 | 102157.3 | 28399729.4 | 46.62060986 | 47 |
| Certainty/6 | 125 |  |  |  | 125 |
| total | 9637 |  | 228437563.5 |  | 500 |

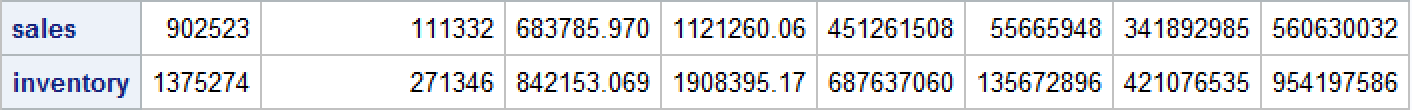
Solved in Excel

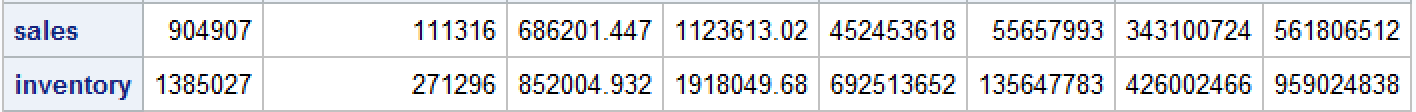
Five samples were selected using Neyman allocation to form estimates for Sales and Inventory.











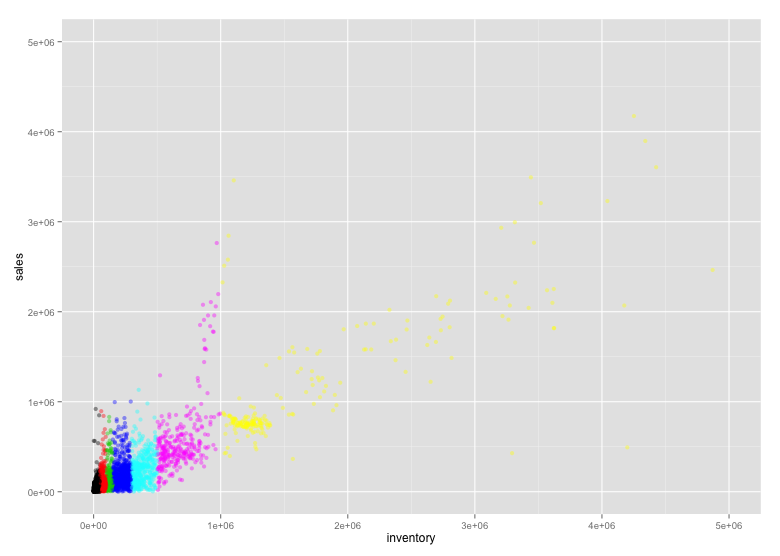
APPENDIX 1.2

Stratification By Inventory

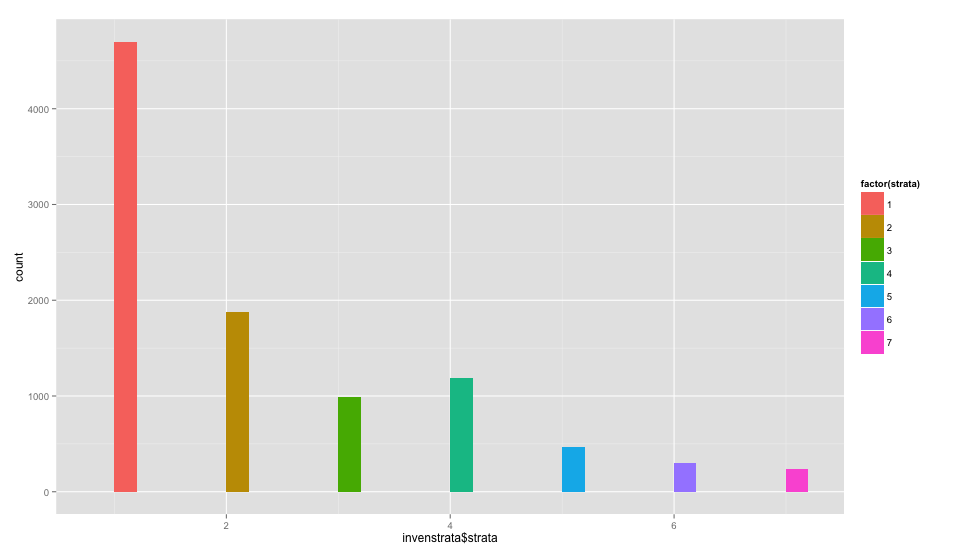
|  |
| --- |
| Cummulative Square Root F Table For Inventory |
| lower upper frequency sqrtf csqrtf stratum |
| 1 0 50000 4700 68.556546 68.55655 1 |
| 2 50000 100000 1876 43.312816 111.86936 2 |
| 3 100000 150000 988 31.432467 143.30183 3 |
| 4 150000 200000 680 26.076810 169.37864 4 |
| 5 200000 250000 281 16.763055 186.14169 4 |
| 6 250000 300000 232 15.231546 201.37324 4 |
| 7 300000 350000 187 13.674794 215.04803 5 |
| 8 350000 400000 128 11.313708 226.36174 5 |
| 9 400000 450000 87 9.327379 235.68912 5 |
| 10 450000 500000 61 7.810250 243.49937 5 |
| 11 500000 550000 62 7.874008 251.37338 6 |
| 12 550000 600000 52 7.211103 258.58448 6 |
| 13 600000 650000 31 5.567764 264.15225 6 |
| 14 650000 700000 37 6.082763 270.23501 6 |
| 15 700000 750000 29 5.385165 275.62017 6 |
| 16 750000 800000 29 5.385165 281.00534 6 |
| 17 800000 850000 24 4.898979 285.90432 6 |
| 18 850000 900000 17 4.123106 290.02742 6 |
| 19 900000 950000 13 3.605551 293.63297 6 |
| 20 950000 1000000 6 2.449490 296.08246 6 |
| Certainty Strata >1000000 7 |

Cutoff points for strata:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 49.34708 | 98.69415 | 148.04123 | 197.38831 | 246.73539 |

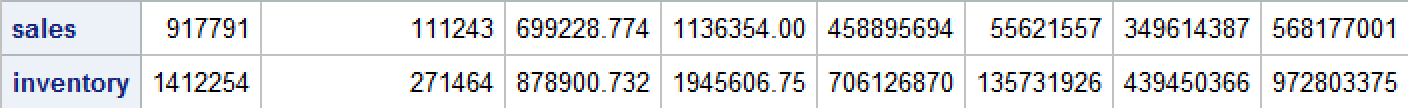
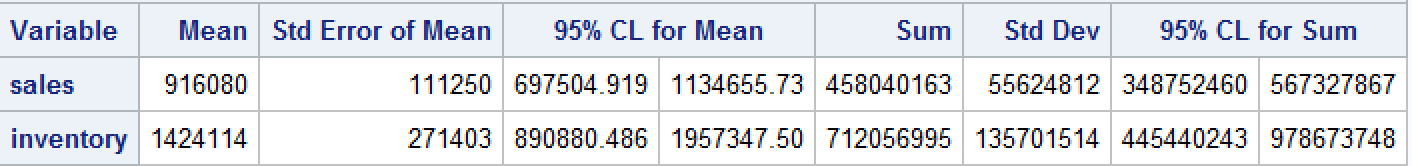


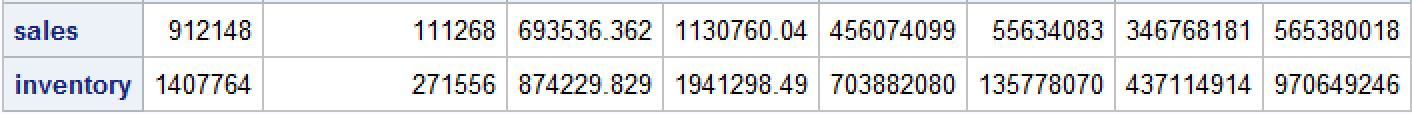
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Neyman Allocation |  |  |  |  |
| Stratum | Nh | Sh=StdDev | NhSh | nh | nh rounded |
| 1 | 4362 | 44128.75 | 192489607.5 | 75.03703396 | 75 |
| 2 | 2360 | 71714.87 | 169247093.2 | 65.97654826 | 66 |
| 3 | 1313 | 86730.68 | 113877382.8 | 44.3921163 | 44 |
| 4 | 929 | 125910.17 | 116970547.9 | 45.59790573 | 46 |
| 5 | 341 | 554237.26 | 188994905.7 | 73.67471594 | 74 |
| 6 | 332 | 543353.89 | 180393491.5 | 70.32167981 | 70 |
| Certainty/7 | 125 | 13036936.69 |  |  | 125 |
| total | 9762 |  | 961973028.6 |  | 500 |

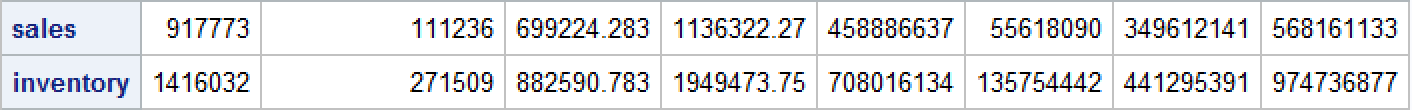
Using Neyman Allocation to find the how many we are selecting from each strata for a sample size of 500.

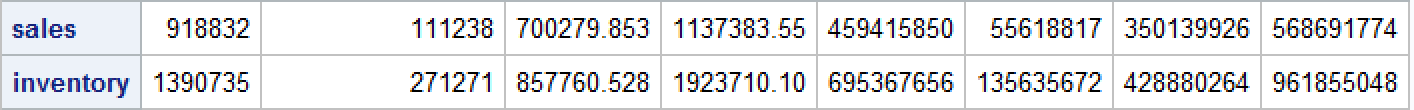
Solved in Excel

Five samples were selected using Neyman allocation to form estimates for Sales and Inventory.









APPENDEX 1.3

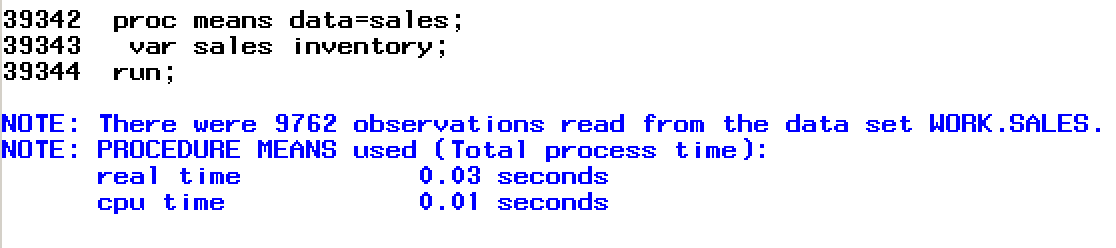
APPENDIX

1.1 Background Analysis

**proc** **means** data=sales;

var sales inventory;

**run**;



1.2 SURVEY SELECT & SURVEY MEANS- SALES

**PROC** **SORT** DATA=sales OUT=sales1 ;

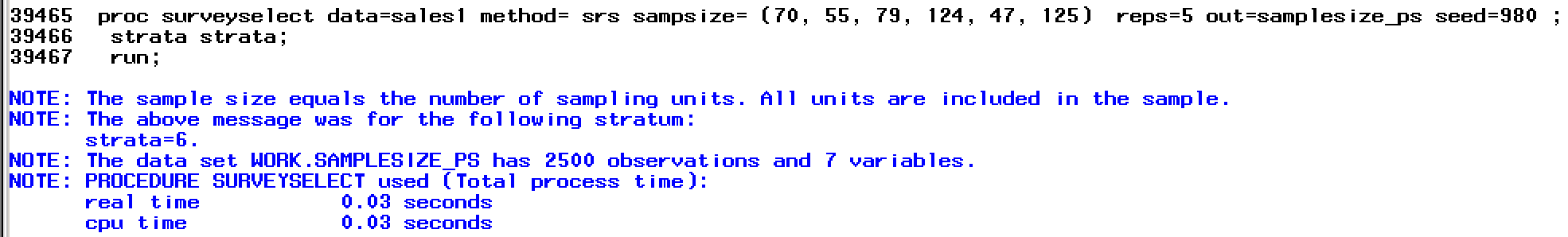
BY strata ;

**RUN** ;

**proc** **surveyselect** data=sales1 method= srs sampsize= (**70**, **55**, **79**, **124**, **47**, **125**) reps=**5** out=samplesize\_ps seed=**980** ;

strata strata;

**run**;



**PROC** **SORT** DATA=samplesize\_ps OUT=samplesize\_ps1 ;

BY Replicate;

**RUN** ;

**PROC** **PRINT** DATA=samplesize\_ps1 ;

**RUN** ;

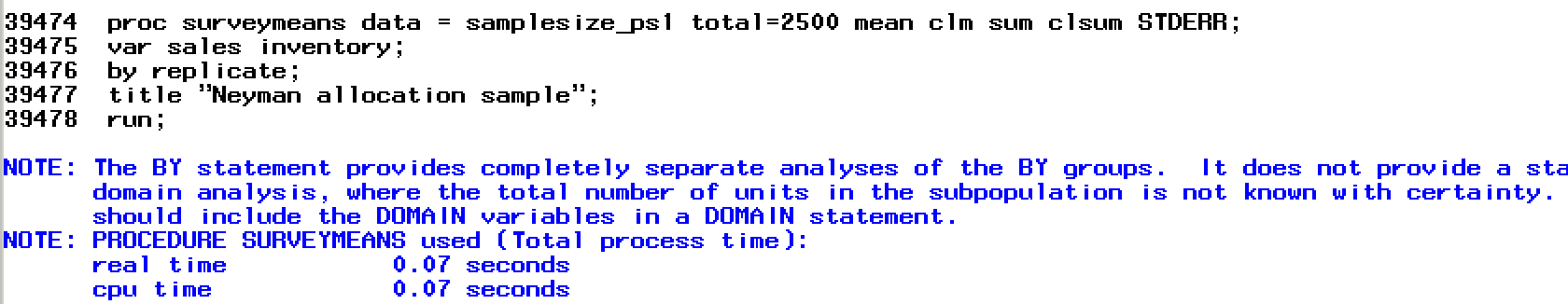
**proc** **surveymeans** data = samplesize\_ps1 total=**2500** mean clm sum clsum STDERR;

var sales inventory;

by replicate;

title "Neyman allocation sample";

**run**;



1.3 SURVEY SELECT & SURVEY MEANS-INVENTORY

**PROC** **SORT** DATA=inven OUT=inven1 ;

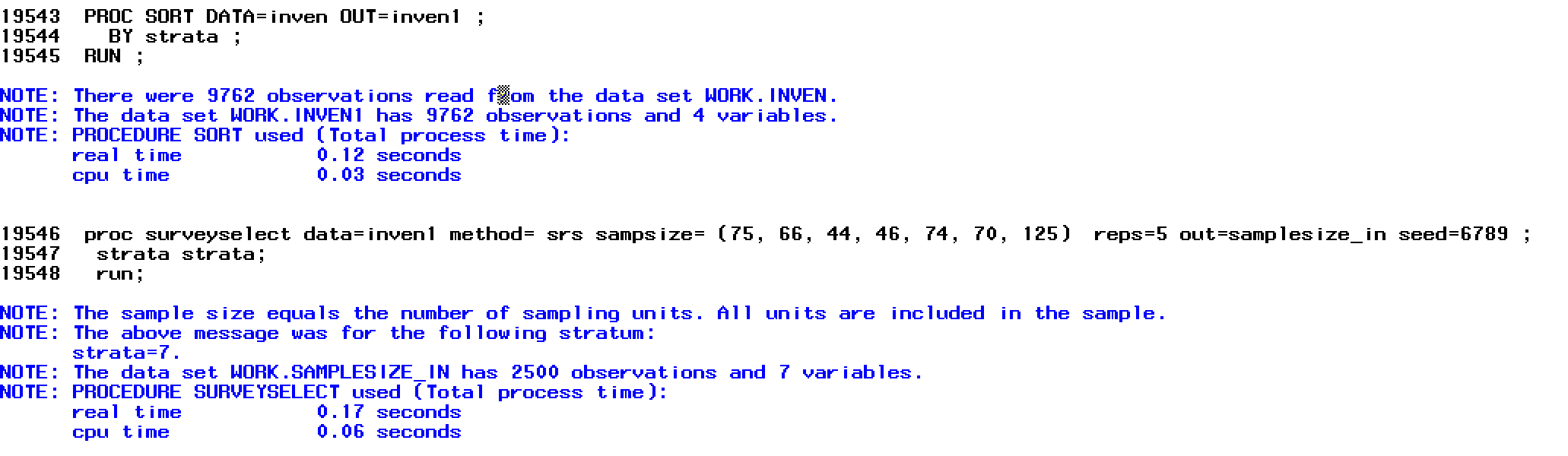
BY strata ;

**RUN** ;

**proc** **surveyselect** data=inven1 method= srs sampsize= (**75**, **66**, **44**, **46**, **74**, **70**, **125**) reps=**5** out=samplesize\_in seed=**6789** ;

strata strata;

**run**;



**PROC** **SORT** DATA=samplesize\_in OUT=samplesize\_in1 ;

BY Replicate;

**RUN** ;

**PROC** **PRINT** DATA=samplesize\_in ;

**RUN** ;

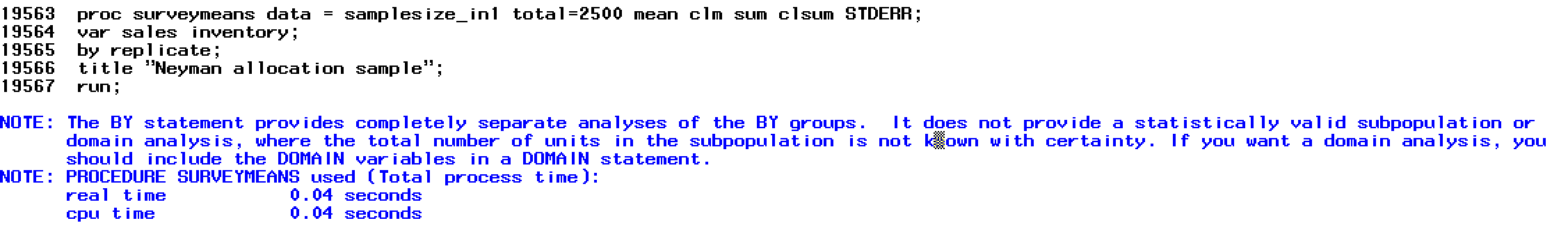
**proc** **surveymeans** data = samplesize\_in1 total=**2500** mean clm sum clsum STDERR;

var sales inventory;

by replicate;

title "Neyman allocation sample";

**run**;



2.1 Cum sqrt f implementations on output in R

#Cumulative Square- Root F method for selecting strata

#Function based on function created by Timothy R. Johnson

#https://rpubs.com/trjohns/stratify

stratify <- function (y, strata, breaks)

{

h <- hist(y, plot = FALSE, breaks = breaks)

g <- length(h$counts)

z <- data.frame(lower = rep(NA, g), upper = rep(NA, g), frequency = h$counts,

sqrtf = sqrt(h$counts), csqrtf = cumsum(sqrt(h$counts)),

stratum = NA)

k <- 1:(strata - 1) \* max(z$csqrtf)/strata

for (i in 1:g) {

z$lower[i] <- h$breaks[i]

z$upper[i] <- h$breaks[i + 1]

}

for (i in 1:(strata - 1)) {

tmp <- which(abs(z$csqrtf - k[i]) == min(abs(z$csqrtf -

k[i])))

z$stratum[c(1:g) <= tmp & is.na(z$stratum)] <- i

}

z$stratum[is.na(z$stratum)] <- strata

return(list(output = z, cutpoints = k))

}

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

#STRATIFICATION OF SALES

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

#Certainty Strata For Sales will be Instances Greater than 1 million

#Subset of the data we will be stratifying

subsetsale<- subset(dat, sales < 1000000)

#Use cum sqrt f to break sales into 20 groups (5% each) then divide into 6 strata

sales.cumsqurtftable<-stratify(subsetsale$sales, strata = 5, breaks = 20)

sales.cumsqurtftable

#applying stratum to sales

salesstrata<-dat

salesstrat<-function(x=2) {

if (x <= 50000){

return(1)

}else if (x <=100000 & x>50000){

return (2)

}else if (x <=200000 & x>100000){

return (3)

}else if (x <=550000 & x>200000){

return (4)

}else if (x <=1000000 & x>550000){

return (5)

}else{

return (6)

}

}

salesstrat(250909)

salesstrata$strata<-mapply(salesstrat, salesstrata$sales)

hist(salesstrata$strata)

write.csv(file="salesStrata", x=salesstrata)

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

#STRATIFICATION OF INVENTORY

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

#Certainty Strata For Inventory will be Instances Greater than 1 million

#Subset of the data we will be stratifying

subsetinven<- subset(dat, inventory < 1000000)

#Use cum sqrt f to break sales into 20 groups (5% each) then divide into 6 strata

inven.cumsqurtftable<-stratify(subsetinven$inventory, strata = 6, breaks = 20)

#applying stratum to sales cutoff take from output of above funtion ^^^

invenstrata<-dat

invenstrat<-function(x=2) {

if (x <= 50000){

return(1)

}else if (x <=100000 & x>50000){

return (2)

}else if (x <=150000 & x>100000){

return (3)

}else if (x <=300000 & x>150000){

return (4)

}else if (x <=500000 & x>300000){

return (5)

}else if (x <=1000000 & x>500000){

return (6)

}else{

return (7)

}

}

invenstrat(250909)

invenstrata$strata<-mapply(invenstrat, invenstrata$sales)

hist(invenstrata$strata)

write.csv(file="invenStrata", x=invenstrata)