2023

Stratified Sampling

A.MAYURI - 2348133

MAYURI.NARAYANAN2002@OUTLOOK.COM

Stratified Sampling

FINANCE -DATASET

A.Mayuri(2348133)

2023-09-12

Question

Consider a dataset as population of your choice and divide the population in various strata by choosing an appropriate variable of stratification. Give the estimates of population parameters(mean and total) by taking a random sample of adequate size using the proportional allocation and optimum allocation methods. Write a report on it.

Data Description:

The taken dataset gives a detailed description of loan defaults by customers on different kinds of auto-mobiles. the data is of large sample with N=200 customers in the past week

Objective:

To do a stratified analysis on the employee type ie(salaried/self employed)

Variables of Interest and their Definition:

UniqueID-Identifier for customers

loan_default– Payment default in the first EMI on due date

disbursed amount–Amount of Loan disbursed.

asset cost–Cost of the Asset

ltv–Loan to Value of the asset

branch_id–Branch where the loan was disbursed.

supplier_id–Vehicle Dealer where the loan was disbursed.

manufacturer_id–Vehicle manufacturer (Hero, Honda, TVS etc.)

Employment.Type–Employment Type of the customer (Salaried/Self Employed)

where we will do the analysis for 2 stratum ie, salaried=1, and stratum=0

ANALYSIS

Code

Step 1: IMPORT DATASET

To import the dataset that has been discussed above to initiat the analysis

```
library(readxl)
finb <- read_excel("C:/Users/mayur/Desktop/Mstat/tri sem 1/R/dataset/finb.xls
x")
View(finb)
attach(finb)
```

Step 2: IMPORT PACKAGE

We are downloading the samplingbook package to proceed with the stratified sampling data analysis

```
library(samplingbook)

## Loading required package: pps

## Loading required package: sampling

## Loading required package: survey

## Loading required package: grid

## Loading required package: Matrix

## Loading required package: survival

##
## Attaching package: 'survival'

## The following objects are masked from 'package:sampling':
##
##     cluster, strata

##
## Attaching package: 'survey'

## The following object is masked from 'package:graphics':
##
##     dotchart
```

Step 3: CREATE STRATUMS

Salaried =1 selfemployed =0

this helps us to create stratums by assigning a binary code for saalaried and self employed

```
# creating stratums . employment type= salaried
stratum1=finb[finb$Employment.Type==1
, ]
stratum1
```

```
## # A tibble: 115 × 8
##    UniqueID disbursed_amount asset_cost   ltv branch_id supplier_id
##       <dbl>            <dbl>      <dbl> <dbl>     <dbl>       <dbl>
##  1   420825            50578      58400  89.6        67       22807
##  2   529269            46349      61500  76.4        67       22807
##  3   510278            43894      61900  71.9        67       22807
##  4   510980            52603      61300  87.0        67       22807
##  5   486821            64769      74190  89.2        67       22807
##  6   478647            53278      61330  89.7        67       22807
##  7   479533            49478      57010  89.5        67       22807
##  8   600655            47549      61400  79.8        67       22807
##  9   467015            31184      57110  56.9        67       22807
## 10   586411            55213      68600  83.1        67       22807
## # ℹ 105 more rows
## # ℹ 2 more variables: manufacturer_id <dbl>, Employment.Type <dbl>
```

```
# creating stratums . employment type= self employed
stratum2=finb[finb$Employment.Type==0
, ]
stratum2
```

```
## # A tibble: 84 × 8
##    UniqueID disbursed_amount asset_cost   ltv branch_id supplier_id
##       <dbl>            <dbl>      <dbl> <dbl>     <dbl>       <dbl>
##  1   537409            47145      65550  73.2        67       22807
##  2   417566            53278      61360  89.6        67       22807
##  3   624493            57513      66113  88.5        67       22807
##  4   539055            52378      60300  88.4        67       22807
##  5   518279            54513      61900  89.7        67       22807
##  6   490213            53713      61973  89.6        67       22807
##  7   548567            53278      61230  89.8        67       22807
##  8   483869            49278      57080  89.4        67       22807
##  9   513916            57713      65750  89.3        67       22807
```

```
## 10   522020             53503      62100  87.3        67        22807
## # ℹ 74 more rows
## # ℹ 2 more variables: manufacturer_id <dbl>, Employment.Type <dbl>
```

Step 4: CALCULATE

mean and standard deviation for two variables disumbered amount and asset cost

```
# calculation for N-stratum population size
N1=sum(stratum1$Employment.Type==1)
N1

## [1] 115

N2=sum(stratum2$Employment.Type==0)
N2

## [1] 84

# mean calculation
M1_disamount=mean(stratum1$disbursed_amount)
M2_disamount=mean(stratum2$disbursed_amount)


M1_ascost=mean(stratum1$asset_cost)
M2_ascost=mean(stratum2$asset_cost)

#std deviation calculation


S1_disamount=sqrt(var(stratum1$disbursed_amount))
S1_ascost=sqrt(var(stratum1$asset_cost))


S2_disamount=sqrt(var(stratum2$disbursed_amount))
S2_ascost=sqrt(var(stratum2$asset_cost))

#output of mean and standard deviation for two variables disumbered amount an
d asset cost
M1_disamount

## [1] 49950.51

M1_ascost

## [1] 65339.81

S1_ascost

## [1] 5209.603
```

```
S1_disamount

## [1] 6736.428

M2_disamount

## [1] 51244.93

M2_ascost

## [1] 66002.29

S2_ascost

## [1] 7932.185

S2_disamount

## [1] 7561.868
```

Step 5: PROPORTIONAL ALLOCATION

let the sample of size n=10 has to be drawn using proportional allocation

```
#let the sample of size n=10 has to be drawn using proportional allocation fo
r asset cost

sample_size_ascost=stratasamp(n=10, Nh=c(N1, N2), Sh=c(S1_ascost, S2_ascost),
type="opt")
sample_size_ascost

##
## Stratum 1 2
## Size    5 5

#let the sample of size n=10 has to be drawn using proportional allocation fo
r disumbersed amount

sample_size_disamt=stratasamp(n=10, Nh=c(N1, N2), Sh=c(S1_disamount, S2_disam
ount), type="opt")
sample_size_disamt

##
## Stratum 1 2
## Size    5 5
```

Step 6: DETERMINATION OF SAMPLE SIZE

1)proportional

```
# determination of total sample size for given specified precision using prop
ortion
stratasize(e=.1, Nh=c(N1, N2), Sh=c(S1_ascost, S2_ascost), type='prop' )

##
## stratamean object: Stratified sample size determination
##
## type of sample: prop
##
## total sample size determinated: 199

stratasize(e=.1, Nh=c(N1, N2), Sh=c(S1_disamount, S2_disamount), type='prop'
)

##
## stratamean object: Stratified sample size determination
##
## type of sample: prop
##
## total sample size determinated: 199
```

2)Optimal

```
# determination of total sample size for given specified precision using prop
ortion
stratasize(e=.1, Nh=c(N1, N2), Sh=c(S1_ascost, S2_ascost), type='opt' )

##
## stratamean object: Stratified sample size determination
##
## type of sample: opt
##
## total sample size determinated: 191

stratasize(e=.1, Nh=c(N1, N2), Sh=c(S1_disamount, S2_disamount), type='opt' )

##
## stratamean object: Stratified sample size determination
##
## type of sample: opt
##
## total sample size determinated: 199
```

3) with precision

```
# determination of total sample size for given specified precision and confid
ence level

stratasize(e=.1, Nh=c(N1, N2), Sh=c(S1_ascost, S2_ascost),level=.99, type="op
t" )

##
## stratamean object: Stratified sample size determination
##
## type of sample: opt
##
## total sample size determinated: 191

stratasize(e=.1, Nh=c(N1, N2), Sh=c(S1_disamount, S2_disamount),level=.99, ty
pe="opt" )

##
## stratamean object: Stratified sample size determination
##
## type of sample: opt
##
## total sample size determinated: 199
```

Step 7: COLLECTION OF RANDOMN DATASET OF SIZE OF 5

```
#collect a random sample of size 5,5 from both strata
sample1=stratum1[sample(1:nrow(stratum1), 5, replace=FALSE), ]
sample1 # sample 1 collected from stratum 1

## # A tibble: 5 × 8
##   UniqueID disbursed_amount asset_cost   ltv branch_id supplier_id
##      <dbl>            <dbl>      <dbl> <dbl>     <dbl>       <dbl>
## 1   566809            48349      67650  72.4        67       22807
## 2   628750            48433      63896  80.2        78       17014
## 3   644762            51428      63306  86.9        78       17014
## 4   517611            46759      62577  78.3        78       17014
## 5   482553            48693      65500  77.9        78       17014
## # ℹ 2 more variables: manufacturer_id <dbl>, Employment.Type <dbl>

sample2=stratum2[sample(1:nrow(stratum2), 5, replace=FALSE), ]
sample2 # sample 1 collected from stratum 2
```

```
## # A tibble: 5 × 8
##   UniqueID disbursed_amount asset_cost   ltv branch_id supplier_id
##      <dbl>            <dbl>      <dbl> <dbl>     <dbl>       <dbl>
## 1   598020            51003      65687  78.7        34       15196
## 2   439084            50678      58300  89.9        67       22807
## 3   576901            49713      68000  77.9        78       17014
## 4   474338            44749      61865  73.4        34       15196
## 5   490213            53713      61973  89.6        67       22807
## # ℹ 2 more variables: manufacturer_id <dbl>, Employment.Type <dbl>
```

```
# total sample collected using stratified random sampling
total_sampled_data=rbind(sample1, sample2)
total_sampled_data
```

```
## # A tibble: 10 × 8
##    UniqueID disbursed_amount asset_cost   ltv branch_id supplier_id
##       <dbl>            <dbl>      <dbl> <dbl>     <dbl>       <dbl>
##  1   566809            48349      67650  72.4        67       22807
##  2   628750            48433      63896  80.2        78       17014
##  3   644762            51428      63306  86.9        78       17014
##  4   517611            46759      62577  78.3        78       17014
##  5   482553            48693      65500  77.9        78       17014
##  6   598020            51003      65687  78.7        34       15196
##  7   439084            50678      58300  89.9        67       22807
##  8   576901            49713      68000  77.9        78       17014
##  9   474338            44749      61865  73.4        34       15196
## 10   490213            53713      61973  89.6        67       22807
## # ℹ 2 more variables: manufacturer_id <dbl>, Employment.Type <dbl>
```

Step 8: ESTIMATIONS

```
#with optimum allocation

# Estimation of population mean using stratified random sample

nh2=as.vector(table(total_sampled_data$Employment.Type))
nh2

## [1] 5 5

wh=nh2/sum(nh2)
wh

## [1] 0.5 0.5
```

1) Disbursed amount

```
stratamean(y=total_sampled_data$disbursed_amount, h=as.vector(total_sampled_d
ata$Employment.Type),
wh=wh, eae=TRUE)
```

```
##              Mean        SE      CIu      CIo
## 0       49971.2 1464.6379 47100.56 52841.84
## 1       48732.4  754.9422 47252.74 50212.06
## overall 49351.8  823.8783 47737.03 50966.57
```

```
stratamean(y=total_sampled_data$disbursed_amount, h=as.vector(total_sampled_d
ata$Employment.Type),
wh=wh)
```

```
##
## stratamean object: Stratified sample mean estimate
## Without finite population correction.
## Mean estimate: 49351.8
## Standard error: 823.8783
## 95% confidence interval: [47737.03,50966.57]
```

Interpretation:

General interpretation::

Thus by using stratified sampling we have deduced the mean estimate of disbursed amount is 49351.8rupees with a standard deviation of 823.8783rupees. also, the mean estimate lied between [47737.03,50966.57] with a 95% confidence level.

Salaried professionals have a mean estimate of 48732.4 rupees as their disbursed amount with a std error of 754.9422 rupees.

Self employed have a mean estimate of 49971.2 rupees as their disbursed amount with a std error of 1464.6379 rupees.

Conclusion ::

Here we can observe that self-employed people will default more than that of salaried people. this could be due to the uncertainty in income, which is higher in self-employed individuals.

2) Asset cost

```
stratamean(y=total_sampled_data$asset_cost, h=as.vector(total_sampled_data$Em
ployment.Type),
wh=wh, eae=TRUE)
```

```
##            Mean        SE      CIu      CIo
## 0       63165.0 1681.0235 59870.25 66459.75
## 1       64585.8  904.8224 62812.38 66359.22
## overall 63875.4  954.5344 62004.55 65746.25
```

```
stratamean(y=total_sampled_data$asset_cost, h=as.vector(total_sampled_data$Em
ployment.Type),
wh=wh)
```

```
##
## stratamean object: Stratified sample mean estimate
## Without finite population correction.
## Mean estimate: 63875.4
## Standard error: 954.5344
## 95% confidence interval: [62004.55,65746.25]
```

Interpretation:

General interpretation::

Thus, by using stratified sampling we have deduced the mean estimate of asset cost is of 63,875.4 rupees with a standard deviation of 954.53 rupees. also the mean estimate lied between [62004.55,65746.25] with a 95% confidence level.

Salaried professionals have a mean estimate of 64,585.4 rupees as their asset amount with a std error of 904.82 rupees.

Self employed have a mean estimate of 63,165 rupees as their asset amount with a std error of 1681.0235 rupees.

Conclusion ::

Here we can observe that self employed people will take loan of lower value than that of salaried. this could be due to the uncertainty in income, which is higher in self employed individuals.