A screen shot of a document

AI-generated content may be incorrect.

|  |  |  |
| --- | --- | --- |
| ***Index*** | | |
| Problem No. | Problem Topics | Page No. |
| Problem 01 | Stratified Sampling | 1 − 2 |
| Problem 02 | Stratified Sampling | 3 − 5 |
| Problem 03 | Cluster Sampling | 5 − 6 |
| Problem 04 | Ratio Estimator | 7 − 9 |
| Problem 05 | PPS Sampling | 9 − 11 |
| Problem 06 | Double Sampling | 11 − 14 |
| Problem 07 | SRSWOR | 14 − 16 |
| Problem 08 | Systematic Sampling | 16 − 18 |

# Problem 01:

We are working with a population student of a High School. We use the sample size of 544 students are drawn from 6 strata of size 53, 65, 75, 79,145, 127 respectively. Then calculate the stratified mean, variance, and 95% confidence interval.

The sample strata are given as,

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 𝑆𝑡𝑟𝑎𝑡𝑎 1 | 𝑆𝑡𝑟𝑎𝑡𝑎 2 | 𝑆𝑡𝑟𝑎𝑡𝑎 3 | 𝑆𝑡𝑟𝑎𝑡𝑎 4 | 𝑆𝑡𝑟𝑎𝑡𝑎 5 | 𝑆𝑡𝑟𝑎𝑡𝑎 6 |
| 165 | 157 | 168 | 164 | 175 | 190 |
| 161 | 161 | 165 | 171 | 173 | 178 |
| 153 | 168 | 175 | 177 | 161 | 194 |
| 150 | 162 | 175 | 163 | 158 | 183 |
| 151 | 165 | 165 | 170 | 175 | 165 |
| 153 | 171 | 163 | 165 | 164 | 170 |
|  | 169 | 165 | 160 | 158 | 176 |
|  | 164 | 165 | 175 | 161 | 173 |
|  |  |  |  | 158 | 168 |
|  |  |  |  | 171 | 183 |
|  |  |  |  | 175 | 173 |
|  |  |  |  | 170 | 183 |
|  |  |  |  | 187 | 174 |
|  |  |  |  | 168 | 177 |
|  |  |  |  | 170 |  |
|  |  |  |  | 185 |  |

# Solution:

st1=c(165,161,153,150,151,153)

st2=c(157,161,168,162,165,171,169,164) st3=c(168,165,175,175,165,163,165,165) st4=c(164,171,177,163,170,165,160,175)

st5=c(175,173,161,158,175,164,158,161,158,171,175,170,187,168,170,185)

st6=c(190,178,194,183,165,170,176,173,168,183,173,183,174,177) Ni=c(53,65,75,79,145,127)

str=function(st1,st2,st3,st4,st5,st6,Ni){

n1=length(st1);n2=length(st2);n3=length(st3);n4=length(st4);n5=lengtPage | **4**h(st5);n6=length(st6) ni=c(n1,n2,n3,n4,n5,n6)

N=sum(Ni) wi=Ni/N fi=ni/Ni

strata.mean=c(mean(st1),mean(st2),mean(st3),mean(st4),mean(st5),mean(st6)) mean=sum(wi\*strata.mean)

si\_sq=c(var(st1),var(st2),var(st3),var(st4),var(st5),var(st6)) var=sum(wi^2\*(1-fi)\*si\_sq/ni)

ztab= qnorm(0.05/2,lower.tail=FALSE) LCL=mean-ztab\*sqrt(var)

UCL=mean+ztab\*sqrt(var)

return(cbind(Mean.st=mean,Variance.st=var,LCL,UCL))

}

str(st1,st2,st3,st4,st5,st6,Ni)

## **Output:**

### ***Mean.st Variance.st LCL UCL***

[1,] 169.5464 0.7762659 167.8195 172.6733

Comments: “The stratified mean indicates that the overall average value across all strata is approximately 169.54. The low variance suggests that the data points are relatively consistent around the mean. The 95% confidence interval (167.81 to 172.67) implies that we can be 95% confident that the true population mean falls within this range."

# Problem 02:

We are working with a population of 544 students of a High School. We use the sample size of 24 students are drawn from 6 strata of size 3, 3, 4, 3, 6, 5

respectively. Choose the sample from the population strata and then calculate the stratified mean, estimate of the variance, and 95% confidence interval.

The population strata are given as,

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 𝑆𝑡𝑟𝑎𝑡𝑎 1 | 𝑆𝑡𝑟𝑎𝑡𝑎 2 | 𝑆𝑡𝑟𝑎𝑡𝑎 3 | 𝑆𝑡𝑟𝑎𝑡𝑎 4 | 𝑆𝑡𝑟𝑎𝑡𝑎 5 | 𝑆𝑡𝑟𝑎𝑡𝑎 6 |
| 165 | 157 | 168 | 164 | 175 | 190 |
| 161 | 161 | 165 | 171 | 173 | 178 |
| 153 | 168 | 175 | 177 | 161 | 194 |
| 150 | 162 | 175 | 163 | 158 | 183 |
| 151 | 165 | 165 | 170 | 175 | 165 |
| 153 | 171 | 163 | 165 | 164 | 170 |
|  | 169 | 165 | 160 | 158 | 176 |
|  | 164 | 165 | 175 | 161 | 173 |
|  |  |  |  | 158 | 168 |
|  |  |  |  | 171 | 183 |
|  |  |  |  | 175 | 173 |
|  |  |  |  | 170 | 183 |
|  |  |  |  | 187 | 174 |
|  |  |  |  | 168 | 177 |
|  |  |  |  | 170 |  |
|  |  |  |  | 185 |  |

# **Solution:**

st1=c(165,161,153,150,151,153)

st2=c(157,161,168,162,165,171,169,164) st3=c(168,165,175,175,165,163,165,165)st4=c(164,171,177,163,170,165,160,175)

st5=c(175,173,161,158,175,164,158,161,158,171,175,170,187,168,170,185) st6=c(190,178,194,183,165,170,176,173,168,183,173,183,174,177) n1=3;n2=3;n3=4;n4=3;n5=6;n6=5

ni=c(n1,n2,n3,n4,n5,n6)

str=function(st1,st2,st3,st4,st5,st6,ni){

N1=length(st1);

N2=length(st2);

N3=length(st3);

N4=length(st4);

N5=length(st5);

N6=length(st6)

Ni=c(N1,N2,N3,N4,N5,N6)

N=sum(Ni) wi=Ni/N fi=ni/Ni

set.seed(1234)

sample.st1=sample(st1,n1) sample.st2=sample(st2,n2) sample.st3=sample(st3,n3) sample.st4=sample(st4,n4) sample.st5=sample(st5,n5) sample.st6=sample(st6,n6)

strata.mean=c(mean(sample.st1),mean(sample.st2),mean(sample.st3),mean(sample.st4),mean( sample.st5),mean(sample.st6))

mean=sum(wi\*strata.mean)

Si\_sq=c(var(st1),var(st2),var(st3),var(st4),var(st5),var(st6)) var.strata=sum(wi^2\*(1-fi)\*Si\_sq/ni)

si\_sq=c(var(sample.st1),var(sample.st2),var(sample.st3),var(sample.st4),var(sample.st5),var(sa mple.st6))

var.est=sum(wi^2\*(1-fi)\*si\_sq/ni)

ztab=qnorm(0.05/2,lower.tail=FALSE) LCL=mean-ztab\*sqrt(var.est)

UCL=mean+ztab\*sqrt(var.est)

return(cbind(Mean.strata=mean,Variance.strata=var.strata,Variance.est=var.est,LCL,UCL))

}

str(st1,st2,st3,st4,st5,st6,ni)

## **Output:**

### ***Mean.strata Variance.strata Variance.est LCL UCL***

[1,] 167.1356 1.403998 1.485277 164.32 169.46

Comments: "The stratified mean represents the average measurement derived from the stratified sample. The very low variance indicates minimal variability within the stratified data. The 95% confidence interval (164.32 to 169.46) suggests that the true population mean is likely to fall within this range with 95% confidence."

# Problem 03:

A city is to divided into 415 clusters. Twenty-four of the clusters will be sampled and interviews are conducted at every household in each of the 24 blocks sampled. The data on incomes are presented in the table below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 𝐶𝑙𝑢𝑠𝑡𝑒𝑟 (𝑖) | 𝑁𝑢𝑚𝑏𝑒𝑟 𝑜𝑓  𝑅𝑒𝑠𝑖𝑑𝑒𝑛𝑡𝑠(𝑚𝑖) | 𝑇𝑜𝑡𝑎𝑙  𝑖𝑛𝑐𝑜𝑚𝑒(𝑦𝑖) | 𝐶𝑙𝑢𝑠𝑡𝑒𝑟 (𝑖) | 𝑁𝑢𝑚𝑏𝑒𝑟 𝑜𝑓  𝑅𝑒𝑠𝑖𝑑𝑒𝑛𝑡𝑠(𝑚𝑖) | 𝑇𝑜𝑡𝑎𝑙  𝑖𝑛𝑐𝑜𝑚𝑒(𝑦𝑖) |
| 1 | 8 | 96 | 13 | 5 | 54 |
| 2 | 12 | 121 | 14 | 10 | 49 |
| 3 | 4 | 42 | 15 | 9 | 53 |
| 4 | 5 | 65 | 16 | 3 | 50 |
| 5 | 6 | 52 | 17 | 6 | 32 |
| 6 | 6 | 40 | 18 | 5 | 22 |
| 7 | 7 | 75 | 19 | 5 | 45 |
| 8 | 5 | 65 | 20 | 4 | 37 |
| 9 | 8 | 45 | 21 | 6 | 51 |
| 10 | 3 | 50 | 22 | 8 | 30 |
| 11 | 2 | 85 | 23 | 7 | 39 |
| 12 | 6 | 43 | 24 | 3 | 47 |

Use the data,

(i) Estimate the per-capita income in the city and

(𝑖𝑖) Place a bound on the error of estimation.

**Solution:**

#N=number of cluster in population

#M=number of elements in population

#m.vec=vector of cluster size in the population #y=either a vector of totals per cluster

cluster.mu=function(N,m.vec,y,M=NA){

n=length(m.vec)

if(is.na(M)){

m.bar=mean(m.vec)

}else{

m.bar=M/N

}

mu.hat=sum(y)/sum(m.vec)

s2.c=sum((y-(mu.hat\*m.vec))^2)/(n-1)

var.mu.hat=((N-n)/(N\*m.bar^2))\*s2.c B=2\*sqrt(var.mu.hat)

return(cbind(mu.hat,s2.c,var.mu.hat,B))

}

m=c(8,12,4,5,6,6,7,5,8,3,2,6,5,10,9,3,6,5,5,4,6,8,7,3)

y=c(96,121,42,65,52,40,75,65,45,50,85,43,54,49,53,50,32,22,45,37,51,30,39,47)

cluster.mu(415,m.vec=m,y,M=NA)

## **Output:**

### ***mu.hat s2.c var.mu.hat B***

[1,] 9.006993 636.602 16.8984 8.2300

Comments: "The estimated average income per person in the city is approximately 9.01 (in the given units). With 95% confidence, the true per-capita income is expected to fall within the range of 8.33 to 16.89."

# Problem 04:

We are interested to know the relative change over a two-year period in the assessed value of homes in a certain community. We are conducting a simple survey sampling of n = 20 homes from the N = 1000 total homes in the community. We obtain the values for this year(y) and the corresponding values from two years

ago(x) for each of the n = 20 homes include in the sample that given below.

|  |  |
| --- | --- |
| 𝑥 | 6.7, 8.2, 7.9, 6.4, 8.3, 7.2, 6.0, 7.4, 8.1, 9.3, 8.2, 6.8, 7.4, 7.5, 8.3, 9.1, 8.6, 7.9, 6.3, 8.9 |
| 𝑦 | 7.1, 8.4, 8.2, 6.9, 8.4, 7.9, 6.5, 7.6, 8.9, 9.9, 9.1, 7.3, 7.8, 8.9, 9.6, 8.7, 8.8, 7.0, 9.4 |

Estimate the relative change(R) and the ratio estimator of population mean in the assessed values for the N = 1000 homes by using X = 7.8 ; also find the

estimated variance of these estimators.

# Solution:

x=c(6.7,8.2,7.9,6.4,8.3,7.2,6.0,7.4,8.1,9.3,8.2,6.8,7.4,7.5,8.3,9.1,8.6,7.9,6.3,8.9)

y=c(7.1,8.4,8.2,6.9,8.4,7.9,6.5,7.6,8.9,9.1,7.3,8.8.3,8.9,9.6,8.7,8.8,8,7.0,9.4) N=1000

Xbar=7.8 ratio=function(x,y,N){

n=length(x)

r=sum(y)/sum(x)

var.r=(N-n)/(n\*(n-1)\*N\*Xbar^2)\*sum((y-r\*x)^2)

ztab=qnorm(0.05/2,lower.tail=FALSE)

LCL=r-ztab\*sqrt(var.r) UCL=r+ztab\*sqrt(var.r)

return(cbind(ratio=r,variance.ratio=var.r,LCL,UCL))

}

ratio(x,y,N)

# Output:

## ***ratio variance.ratio LCL UCL***

[1,] 1.066019 5.444505e-05 1.051557 1.080481

Comments: "The relative change (R ≈ 0.06601) suggests that, on average, current home values are 97.5% of their levels two years ago, indicating a slight decline in assessed values. The estimated population mean (≈ 7.84) represents the average current assessed value per home. Once calculated, the variance will aid in constructing confidence intervals, providing insight into the reliability of the estimate."

# Problem 05:

A list of 23 farmer agricultural districts of Bangladesh with areas in thousand

areas of lands (𝑥) is given in the accompanying table together with district wise production of rice (𝑦) (in 1000 metric tons) in 1998 − 99

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| District | 𝑥 | 𝑦 | District | 𝑥 | 𝑦 |
| 1.Banderban | 61 | 48 | 13.Tangail | 561 | 483 |
| 2.Chattogram | 1079 | 994 | 14.Barishal | 133 | 662 |
| 3.Khagrachari | 30 | 26 | 15.Jessore | 1452 | 1352 |
| 4.Cumilla | 1519 | 1313 | 16.Khulna | 1134 | 853 |
| 5.Noakhali | 1036 | 779 | 17.Kustia | 567 | 479 |
| 6.Rangamati | 48 | 40 | 18.Patuakhali | 1027 | 543 |
| 7.Sylhet | 2309 | 1512 | 19.Bogra | 1169 | 1093 |
| 8.Dhaka | 936 | 859 | 20.Dinajpur | 1573 | 1069 |
| 9.Faidpur | 1018 | 577 | 21.Pabna | 738 | 660 |
| 10.Jamalpur | 811 | 723 | 22.Rajshahi | 1799 | 1753 |
| 11.Kishorgonj | 1341 | 1121 | 23.Rangpur | 2243 | 1873 |
| 12.Mymensingh | 1715 | 928 | Total | 24299 | 19740 |

𝑎) Draw a sample of 5 districts without replacement using,

(𝑖) Simple random sampling

(𝑖) PPS method

𝑏) Estimate the average and total production of rice per district for both the samples.

𝑐) Compute the 95% confidence interval in each case.

***Solution:***

X=c(61,1079,30,1519,1036,48,2309,936,1018,811,1341,1715,561,133,1452,1134, 567,1027,1169,1573,738,1799,2243)

Y=c(48,994,26,1313,779,40,1512,859,577,723,1121,928,483,662,1352,853,479,54 3,1093,1069,660,1753,1873)

N=length(Y) n=5

***## a ##***

SRSWOR=function(X,Y,n){ #y=sample(Y,n,replace=FALSE) y=c(483,723,40,1093,660)

ybar=sum(y)/length(y)

#An estimate of the total population is Y.hat=N\*ybar

Ybar=sum(Y)/N S\_sq=sum((Y-Ybar)^2)/(N-1) var.Ybar=(N-n)/N\*(S\_sq/n) s\_sq=sum((y-ybar)^2)/(n-1) var.ybar=(N-n)/N\*(s\_sq/n) se.ybar=sqrt(var.ybar)

#95% CI for population mean ztab=qnorm(0.05/2,lower.tail=FALSE)

LCL=ybar-ztab\*se.ybar UCL=ybar+ztab\*se.ybar

return(rbind(ybar,Ybar,Y.hat,S\_sq,var.Ybar,s\_sq,var.ybar,se.ybar,LCL,UCL))

}

SRSWOR(X,Y,n)

######## PPS ###########

cbind(X,Cum.sum=cumsum(X),Probability=round(X/sum(X),3))

#ran=sample(1: length(X),5,replace=FALSE);ran ran=c(15,12,20,8,22)

yi=Y[ran] xi=X[ran] cbind(yi,xi) Xt=sum(X)

ybar\_pps=Xt/(N\*n)\*sum(yi/xi)

Yhat\_pps=N\*ybar\_pps

var.ybar\_pps=sum((Xt/N\*(yi/xi)-ybar\_pps)^2)/(n\*(n-1)) Se.ybar\_pps=sqrt(var.ybar\_pps)

ztab=qnorm(0.05/2,lower.tail=FALSE) LCL=ybar\_pps-ztab\*Se.ybar\_pps UCL=ybar\_pps+ztab\*Se.ybar\_pps

cbind(ybar\_pps,Yhat\_pps,var.ybar\_pps,LCL,UCL)

# Output:

[,1]

|  |  |
| --- | --- |
| ***ybar*** | 598.4900 |
| ***Ybar*** | 858.2609 |
| ***Y.hat*** | 14995.4000 |
| ***S\_sq*** | 252288.5652 |
| ***var.Ybar*** | 389019.0798 |
| ***s\_sq*** | 149466.7000 |
| ***var.ybar*** | 23050.4400 |
| ***se.ybar*** | 151.8237 |
| ***LCL*** | 301.9255 |
| ***UCL*** | 898.9745 |

### ***ybar\_pps Yhat\_pps var.ybar\_pps LCL UCL***

[1,] 859.49777 19783.091 7969.1839 677.36193 1039.0338

Comments : "The SRS sample yields a higher average production estimate but with a wider confidence interval, indicating greater variability. The PPS method provides a slightly lower estimate but may offer greater accuracy due to size-based sampling. The overlap in confidence intervals suggests that the estimates are consistent within the expected range of variability."

# Problem 06:

You are tasked with estimating the average daily calorie intake of individuals in a population across different age groups. To achieve this, you decided to employ double sampling, a cost-effective sampling technique.

a) Implement the double sampling technique with the following specifications:

* Select 50 individuals for each sampling day.
* Sample data for 5 days.

𝑏) Write a function to perform double sampling on the dataset, randomly

selecting a subset of individuals for each day and a subset of days to repeat the process.

c) Calculate the average daily calorie intake from the sampled data and estimate the overall average for the entire population.

𝑑) Discuss the implications of using double sampling in estimating population parameters and provide insights based on your findings.

# Solution:

***## a) #########***

data=read.csv(file.choose()) data

***### b) #######***

double\_sampling=function(data,i,d) {

set.seed(123)

individual=sample(data$ID,i,replace=FALSE) day=sample(data$ID,d,replace=FALSE)

sampled\_data=data[individual[day],]

return(sampled\_data)

}

sampled\_data=double\_sampling(data=data,i=50,d=5);

sampled\_data

## **Output:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **ID** | **Age** | **Calories** |
| 40 | 40 | 48 | 2300 |
| 41 | 41 | 40 | 2350 |
| 24 | 24 | 55 | 3000 |
| 32 | 32 | 47 | 2700 |
| 17  ## c) | 17 | 40 | 2500 |

ybar=mean(sampled\_data$Calories);ybar Ybar=mean(data$Calories);Ybar

## **Output:**

### ***ybar Ybar***

[1,] 2570 2550

Comments: Double sampling improves precision by refining estimates. It reduces costs compared to full population sampling. The confidence interval quantifies the uncertainty in the estimate. Results can be used to make policy recommendations on nutrition.

######## d) #############

paste("The average daily calorie intake from double sampling data is=",ybar)

paste("The overall average daily calorie intake for entire population is=",Ybar)

paste("which moderate distance is=",abs(ybar-Ybar))

## **Output:**

[1] "The average daily calorie intake from double sampling data is= 2570"

[1] "The overall average daily calorie intake for entire population is= 2550"

[1] "which moderate distance is= 20

Comment: The distance is small, Here the double sampling estimator is a good representative of population parameter.

# Problem 07:

From a population of size 500, there choose a sample of size 12 are

45,54,25,48,35,46,95,58,15,47,28,24

a) Find the mean, standard error, margin of error and 95% confidence interval.

b) The given data as a population, choose 5 sample using simple random sampling and find the mean, standard error, margin of error and 95% confidence interval.

# Solution:

### ##### a) ######

sample=c(45,54,25,48,35,46,95,58,15,47,28,24)

srswor=function(x){

N=500

n=length(x) Mean=mean(x) s\_sq=var(x) f=n/N

var.est=((1-f)/n)\*s\_sq S.E=sqrt(var.est)

M.e=1.96\*S.E

LCL=Mean-1.96\*S.E

UCL=Mean+1.96\*S.E

return(cbind(Mean,S.E,LCL,UCL,M.e))

}

srswor(sample)

## **Output:**

### ***Mean S.E LCL UCL M.e***

[1,] 42.32323 5.894394 31.98322 54.48344 11.95011

### ############ b) #########

pop=c(45,54,25,48,35,46,95,58,15,47,28,24)

n=5

SRSWOR=function(pop,n){ N=length(pop) set.seed(123)

x=sample(pop,n,replace=FALSE) Mean=mean(x)

s\_sq=var(x) f=n/N

var.est=((1-f)/n)\*s\_sq S.E=sqrt(var.est) me=1.96\*S.E

LCL=Mean-1.96\*S.E UCL=Mean+1.96\*S.E

print(x)

return(cbind(Mean,S.E,LCL,UCL,"Margin of Error"=me))

}

SRSWOR(pop,n)

## **Output:**

[1] 25 24 47 54 46

### ***Mean S.E LCL UCL Margin of Error***

[1,] 39.02 4.7104431 29.89932 48.48038 9.422684

Comments: The mean represents the central value of the sample. The standard error measures the expected variation of the sample mean from the population mean. The margin of error estimates the range within which the true population mean is likely to fall. The confidence interval provides an estimated range for the true population parameter based on the sample data.

# Problem 08:

Here a population of size 10 are given as, 12, 15, 25, 36, 37, 41, 45, 49, 50, 52 .

Draw a systematic random sampling of 4 patient from a list of 10 patients. Using sample data, estimate the mean, standard error and 95% confidence interval of this estimator and comment on it.

Solution:

p=c(12,15,25,36,37,41,45,49,50,52)

n=4

sys3=function(p,n){

N=length(p)

k=round(N/n,0)

set. Seed(123)

ran=sample(1:k,1)

s=c()

for(i in 1:n){

si=ran+k\*(i-1)

if(si>N)(si=si-N)

s[i]=p[si]

}

ybar=mean(s)

Ybar=mean(p)

s\_sq=(N/(N-1))\*var(p)

swsy2=(1/(k\*(n-1)))\*sum((s-ybar)^2)

swsy2=(1/(k\*(n-1)))\*sum((s-ybar)^2)

var.sys=((N-1)/N)\*s\_sq-((n-1)/n)\*swsy2

S.E=sqrt(var.sys)

LCL=ybar-1.96\*S.E

UCL=ybar+1.96\*S.E

M.e=1.96\*S.E

print(s)

return(cbind(Ybar,ybar,S.E,LCL,UCL,M.e))

}

sys3(p,n)

## **Output**:

### sample

[1] 12 25 37 45

### ***Ybar ybar S.E LCL UCL M.e***

[1,] 36.2 29.75 11.3774 7.450304 52.0497 22.2997

Comment: "The mean serves as an estimate of the population mean. The standard error (SE) indicates the expected variability among different samples drawn from the same population. A confidence interval, at a 95% confidence level, provides a range within which the true population mean is likely to fall."

End