**Lab 4. Allocating sample to strata in stratified sampling**

**MSDS 6370**

**Submitted by Hari Narayan Sanadhya**

**Exercises – Part I**

**The excel below shows how this entire exercise (Part 1) was solved using excel.**

****

**Solution Ex 1**

**Table 1. Proportional allocation of sample of size 50 from Industry 2 population (Table created in data tab of the excel attached)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| stratum | Number of units | Nh/N | 50\*Nh/N | sample size |
| 1 | 228 | 0.608 | 30.4 | 29 |
| 2 | 101 | 0.269333 | 13.46667 | 13 |
| 3 | 30 | 0.08 | 4 | 4 |
| 4 | 10 | 0.026667 | 1.333333 | 2 |
| 5 | 6 | 0.016 | 0.8 | 2 |

For stratum 4 and 5, since the value of the sample size was coming out to be less than 2, so replaced those with 2 and then for stratums 1, 2 and 3, recalculate the sample size using proportional allocation.

**Solution Ex 2**

**Table 2. Neyman allocation of sample of size 50 from Industry 2 population (Table created in data tab of the excel attached)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| stratum |  |  | 50\* | sample size |
| 1 | 13422.57456 | 0.316488767 | 15.82443834 | 17 |
| 2 | 9363.570336 | 0.220782147 | 11.03910735 | 12 |
| 3 | 4898.794194 | 0.115507895 | 5.775394751 | 6 |
| 4 | 7443.45462 | 0.175508042 | 8.775402076 | 9 |
| 5 | 7282.509835 | 0.17171315 | 8.585657492 | 6 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| stratum | Nh | Sh | |  | | --- | |  | | | |  | | --- | |  | | | |  | | --- | | 50\* | | | sample size |
|
| 1 | 228 | 58.87094104 | 13422.57456 | 0.316488767 | 15.82443834 | 17 |
| 2 | 101 | 92.70861719 | 9363.570336 | 0.220782147 | 11.03910735 | 12 |
| 3 | 30 | 163.2931398 | 4898.794194 | 0.115507895 | 5.775394751 | 6 |
| 4 | 10 | 744.345462 | 7443.45462 | 0.175508042 | 8.775402076 | 9 |
| 5 | 6 | 1213.751639 | 7282.509835 | 0.17171315 | 8.585657492 | 6 |

For Stratum 5, the value of the sample size was greater than the actual samples available in the sample. So, replaced the sample size of strata 5 by the size of that strata and recalculate the sample size using Neyman's allocation.

**Solution Ex 3**

**Table 3. Neyman allocation of sample of size 50 from Industry 2 population when substituting the sum of measure of size for product of stratum size and std. deviation (Table created in data tab of the excel attached)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| stratum | MOSh | MOSh/total MOS | 50\* MOSh/total MOS | sample size |
| 1 | 5968.472499 | 0.343023791 | 17.15118955 | 18 |
| 2 | 4147.081261 | 0.238343653 | 11.91718263 | 12 |
| 3 | 2086.905223 | 0.119939924 | 5.996996223 | 6 |
| 4 | 2774.350415 | 0.159449205 | 7.972460263 | 8 |
| 5 | 2422.778202 | 0.139243427 | 6.96217134 | 6 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| stratum | Nh | Sh | MOSh | MOSh/total MOS | 50\* MOSh/total MOS | sample size |
|
| 1 | 228 | 26.17751096 | 5968.472499 | 0.343023791 | 17.15118955 | 18 |
| 2 | 101 | 41.06021051 | 4147.081261 | 0.238343653 | 11.91718263 | 12 |
| 3 | 30 | 69.56350742 | 2086.905223 | 0.119939924 | 5.996996223 | 6 |
| 4 | 10 | 277.4350415 | 2774.350415 | 0.159449205 | 7.972460263 | 8 |
| 5 | 6 | 403.7963671 | 2422.778202 | 0.139243427 | 6.96217134 | 6 |

For Stratum 5, the value of the sample size was greater than the actual samples available in the sample. So, replaced the sample size of strata 5 by the size of that strata and recalculate the sample size using Neyman's allocation.

**Solution Ex 4 (Table created in "Proportional allocation" tab of the excel attached)**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Mean   |  | | --- | |  | | Std Error of Mean | Degree of Freedom | tstat | 95% CI of Mean | | Sum | Std Error of Sum | 95% CL for Sum | |
| 420.3485 | 13.0116607 | 45 | 2.014103 | 394.1417 | 446.5553 | 157630.7 | 4879.372762 | 147803.1 | 167458.3 |

**Solution Ex 5 (Table created in "Neyman Allocation" tab of the excel attached)**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Mean   |  | | --- | |  | | Std Error of Mean | Degree of Freedom | tstat | 95% CI of Mean | | Sum | Std Error of Sum | 95% CL for Sum | |
| 476.7079 | 12.39295318 | 45 | 2.014103389 | 451.7472 | 501.6686 | 178765.5 | 4647.357 | 169405.2 | 188125.7 |

**Solution Ex 6**

|  |  |  |
| --- | --- | --- |
| **Population mean and standard Deviation** | | |
| **Mean** | **Std Dev** | **Sum** |
| 452.418 | 1159.364725 | 169656.8684 |

When compared the output of the Proportional allocation with Neyman allocation sample, Nayman allocation gives better result. The reason being

1. From Neyman, the mean obtained is closer to the population mean when compared with the mean obtained from the Proportional allocation.
2. From Neyman, the standard error obtained is smaller as compared with the standard error obtained from the Proportional allocation.
3. For sum, though values obtained in both the cases were not that good when compared with the sum of sales in the population, Neymans allocation looks better in this case as well since the sum of sales in population lies in 95% CI of sum which is not the case for proportional allocation.

**Exercises – Part II**

**Solution Ex 1-6 (SAS Code)**

**\*\*\*\* This is code to extract sample from population and do analysis for**

**stratified sample \*\*\*\*;**

**\*\*\*\* input population below \*\*\*\*\*\*;**

**libname xl XLSX '//home/harisanadhya0/sasuser.v94/MSDS 6370/Unit 4/lab4Dat.xlsx';**

**data industryA;**

**set xl.data;**

**run;**

**\*\*\*\* input population above \*\*\*\*\*\*;**

**/\* View the data to verify the import \*/**

**proc print data=industryA(obs=10);**

**title "Sample input data to verify if the import is successful";**

**run;**

**/\* Population parameter values for Sales field \*/**

**proc means data=industryA mean stddev sum;**

**var sales;**

**title "Population parameter values for sales field using proc Means";**

**run;**

**\*\*\*\* Create table with population strata sizes below \*\*\*;**

**proc sql;**

**create table popStrsizes as**

**select Strata , count(\*) as \_total\_ from industryA group by Strata order by Strata;**

**\*\*\*\* "order by" guarantees order in popStrsizes is consistent sample \*\*\*\*;**

**run;**

**/\* Print the dataset contains the strata counts \*/**

**proc print data=popStrsizes;**

**title "Table with population strata sizes";**

**run;**

**/\* Exercises – Part II Question 1 \*/**

**\*\*\*\* select a sample of size 50 using the proportional allocation \*\*\*\*;**

**proc surveyselect data=industryA**

**out = propsample**

**sampsize = (29, 13, 4, 2, 2)**

**seed=91118 noprint; \*\*\* For proportional sample use 91118, Neyman sample 91119, MOS**

**sample 91119 \*\*\*;**

**strata Strata;**

**title "Sample Selected using Proportional allocation";**

**run;**

**proc sort data = propsample;by Strata; run;**

**\*\*\* this guarantees order in sample is consistent with popStrsizes \*\*\*;**

**proc print data=propsample;**

**run;**

**/\* Exercises – Part II Question 2 \*/**

**\*\*\*\* select a sample of size 50 using the proportional allocation \*\*\*\*;**

**proc surveyselect data=industryA**

**out = neysample**

**sampsize = (17, 12, 6, 9, 6)**

**seed=91119 noprint; \*\*\* For proportional sample use 91118, Neyman sample 91119, MOS**

**sample 91119 \*\*\*;**

**strata Strata;**

**title "Sample Selected using Neyman's allocation";**

**run;**

**proc sort data = neysample;by Strata; run;**

**\*\*\* this guarantees order in sample is consistent with popStrsizes \*\*\*;**

**proc print data=neysample;**

**run;**

**/\* Exercises – Part II Question 3 \*/**

**\*\*\*\* select a sample of size 50 using the proportional allocation \*\*\*\*;**

**proc surveyselect data=industryA**

**out = mossample**

**sampsize = (18, 12, 6, 8, 6)**

**seed=91119 noprint; \*\*\* For proportional sample use 91118, Neyman sample 91119, MOS**

**sample 91119 \*\*\*;**

**strata Strata;**

**title "Sample Selected using Neyman's allocation with the sum of MOS ";**

**run;**

**proc sort data = mossample;by Strata; run;**

**\*\*\* this guarantees order in sample is consistent with popStrsizes \*\*\*;**

**proc print data=mossample;**

**run;**

**/\* Exercises – Part II Question 4 \*/**

**proc surveymeans data=propsample total=popStrsizes**

**mean sum clsum clm t;**

**var sales;**

**weight samplingweight;**

**strata Strata ;**

**title "Estimation from Proportional Sample" ;**

**run;**

**/\* Exercises – Part II Question 5 \*/**

**proc surveymeans data=neysample total=popStrsizes**

**mean sum clsum clm t;**

**var sales;**

**weight samplingweight;**

**strata Strata;**

**title "Estimation from Neyman's Sample" ;**

**run;**

**/\* Exercises – Part II Question 6 \*/**

**proc surveymeans data=mossample total=popStrsizes**

**mean sum clsum clm t;**

**var sales;**

**weight samplingweight;**

**strata Strata;**

**title "Estimation from Neyman's Sample using MOS" ;**

**run;**

SAS Output



**Solution Ex 7**

Comparing the results of Proportional allocation and Neyman's allocation using Sales and MOS, its found that the closest value of Mean, Standard Deviation and Sum to the population is obtained in the case of Neyman's allocation using MOS. The mean of sales from the population is 452.4183158 and mean of sales from the three samples obtained is 412.372281, 426.411487 and 430.619864 respectively. Though the means obtained from all the three samples are not equal to the population mean, but still the closest one is from the Neymans allocation using MOS. Smallest value obtained for standard error of mean is 10.133078 making it the best. But if we look at the 95% confidence interval of the mean obtained in the three cases, only in case of proportional allocation, the population mean lies within it. The same happens in case of estimation of sum of sales i.e. only in case of proportional allocation, the population sum of sales lies within the 95% confidence interval. So as far as best results are concerned, results from Neyman's allocation using MOS appear to the best but the best of the 3 samples for prediction with 95% confidence appears to be the sample from proportional allocation.