**MSDS 6370 Sampling Statistics**

**Midterm Exam – Spring 2018**

**Submitted By Hari Narayan Sanadhya**

**This exam is due at 12:00 Noon CT on Thursday, February 22, 2018. You are to work independently on this exam. You may not consult other people. However, you may use course materials. (Please take SMU Honor code seriously)**

1. **(10 pts)** A SRSWOR is chosen from a population of 400 acres of land. Results are below.

|  |  |
| --- | --- |
| **Price/Acre** | |
|
| **Mean** | 15,050 |
| **Median** | 11,882 |
| **Sample Standard Deviation** | 10,548 |
| **Minimum** | 0 |
| **Maximum** | 53,862 |
| **Sum** | 602,000 |
| **Sample Size** | 40 |

1. Give an estimate of the standard error of the mean from the sample?

**Solution:**

*)*

1. Calculate a 95% confidence interval for the population total land value.

**Solution:**

Where n = sample size

= sample mean

zstat = z-statistic for 95% confidence

SE = Standard error of mean

95% CI = 40\*15050 ± 40\*1.96\*1582.2 = 602000 ± 124044.48 = (477955.52, 726044.485)

1. Suppose you want to retake the sample to achieve a margin of error for the mean of 2000 with a 90% confidence interval. How large a sample size would be needed to achieve this objective?

**Solution:**

MOE = 2000

MOE =

Solving for n, we get the equation as

Sample size of 41 would be required to achieve a margin of error for the mean of 2000 with a 90% confidence interval.

2. **True/False:** (**20 pts; 2 points for each**)

For each statement, select True or False. Make sure it is clear which one you are choosing. Think Carefully!

2.1 Simple random samples always will give a better estimate of the mean than judgement samples if implemented properly. (TRUE OR FALSE)

**FALSE (SRS if implemented properly yield great estimate but it not always the case. Judgement sample can also give good estimates depending on who is doing the judgement or how the judgement is been done. They may sometimes yield better estimates then the simple random samples but is not that likely to happen)**

2.2 The largest mean for a simple random sample is a subset of the highest values in the population. (TRUE OR FALSE)

**FALSE (Largest mean is derived using the highest values in the population but may not be a subset of the same as the set of highest values used are just some specific number but the mean obtained can be one of those numbers or a value between them)**

2.3 For a yes, no question on a survey you can find the overall proportion of yes answers in the survey by summing the proportion of yes answers in strata. (TRUE OR FALSE)

**FALSE (Because it will not take into account the different sizes that each strata has)**

2.4 Sampling error can be minimized, but never eliminated if the sample does not include the entire population. (TRUE OR FALSE)

**TRUE**

2.5 If one observation in a population has a probability of selection that is unknown, then you cannot use weighting to estimate the mean from a probability sample. (TRUE OR FALSE)

**TRUE (No more a probability sample)**

2.6 When using the probability of selection to find the mean of a sample, the formulas used include the probabilities of selection in functions in both the numerator and denominator. (TRUE OR FALSE)

**FALSE (Only in the numerator)**

2.7 From the Central Limit Theorem we know that the sample mean has a normal distribution. This allows us to construct meaningful confidence intervals for our sample estimates.

(TRUE OR FALSE)

**TRUE**

2.8 Proc surveyMeans in SAS is used to select a random sample. (TRUE OR FALSE)

**FALSE (PROC SURVEYSELECT does the task of random sample selection)**

2.9 When using sample data to estimate a population variance you do not have to include the finite population correction. (TRUE OR FALSE)

**TRUE (Population variance estimation would be done mostly when the population is big and it's not possible to collect all the data.)**

2.10 The standard error of the mean is the standard deviation of the sample. (TRUE OR FALSE)

**FALSE (Standard error is the standard deviation of the sampling distribution not the standard deviation of the sample)**

3. **(10 points)** The Chicago Fitness Company has five clubs in different parts of the city. The company wants to estimate the total number of customers served in a year by taking an SRSWOR sample of two clubs. Last year, Club1, Club2, Club3, Club4, and Club5 served 1000, 1200, 1400, 1800, and 1600 customers respectively:

(a) (2 points) Fill in the table below showing all samples.

Solution

|  |  |  |
| --- | --- | --- |
| **Sample** | **Clubs** | **Estimate Total Served** |
| 1 | club1, club2 | 5500 |
| 2 | club1, club3 | 6000 |
| 3 | club1, club4 | 7000 |
| 4 | club1, club5 | 6500 |
| 5 | club2, club3 | 6500 |
| 6 | club2, club4 | 7500 |
| 7 | club2, club5 | 7000 |
| 8 | club3, club4 | 8000 |
| 9 | club3, club5 | 7500 |
| 10 | club4, club5 | 8500 |
| Number of Samples | 10 |  |

(b) (3 points) Find the sampling distribution of the estimate of the total number of customers served and fill in the table below.

Solution

|  |  |  |
| --- | --- | --- |
| **Total Served** | **# of samples with Value** | **Proportion of samples with value** |
| Lowest Sample Value 5500 | 1 | 0.1 |
| 6000 | 1 | 0.1 |
| 6500 | 2 | 0.2 |
| 7000 | 2 | 0.2 |
| 7500 | 2 | 0.2 |
| 8000 | 1 | 0.1 |
| Highest Sample Value 8500 | 1 | 0.1 |

(C) (5 points) Suppose clubs 1, 2, and 3 are on the Southside and clubs 4 and 5 are on the Northside of the city.

1. Show the possible samples from taking a sample of one club from the Southside and one from the Northside and calculate an estimate of the population mean using these samples.

**Solution:**

|  |  |  |  |
| --- | --- | --- | --- |
| Club | Served\_customers | Weight | Served\_customers \* Weight |
| club1 | 1000 | 3 | 3000 |
| club2 | 1200 | 3 | 3600 |
| club3 | 1400 | 3 | 4200 |
| club4 | 1800 | 2 | 3600 |
| club5 | 1600 | 2 | 3200 |

|  |  |  |  |
| --- | --- | --- | --- |
| Sample | Clubs | Weighted Estimate of Total Served | Estimated Population Mean |
| 1 | club1, club4 | 1320 | 1400 |
| 2 | club1, club5 | 1240 |  |
| 3 | club2, club4 | 1440 |  |
| 4 | club2, club5 | 1360 |  |
| 5 | club3, club4 | 1560 |  |
| 6 | club3, club5 | 1480 |  |

2. Calculate the standard error of the mean using these samples and show how this standard error is different than the standard error from the simple random sampling without replacement design.

**Solution**

|  |  |  |  |
| --- | --- | --- | --- |
| STD ERR using weighted Sample | STD ERR using SRS | Mean Using Weighted Sample | Mean Using SRS |
| 115.931 | 182.5741858 | 1400 | 1400 |

Though in both the cases, the estimated population mean is exactly same but the standard error from stratified sample is less then the standard error obtained from SRS and hence the stratified design is better. The std error from SRS is approximately 1.575 times the standard error from stratified design.



4. **(10 points)** There are 12 peanut farms in a rural county in Florida. 4 farms are owned by a person less than or equal to 40 years of age and 8 are owned by a person older than 40.

A random sample without replacement of size 2 is taken from this population and the proportion older than 40 is estimated from the sample.

(a) Calculate the mean and standard deviation of the sampling distribution.

**Solution**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| farm 1 | farm 2 | proportion older than 40 | Mean | Std Dev |
| 1 | 2 | 0 | 0.666667 | 0.320256 |
| 1 | 3 | 0 |  |  |
| 1 | 4 | 0 |  |  |
| 1 | 5 | 0.5 |  |  |
| 1 | 6 | 0.5 |  |  |
| 1 | 7 | 0.5 |  |  |
| 1 | 8 | 0.5 |  |  |
| 1 | 9 | 0.5 |  |  |
| 1 | 10 | 0.5 |  |  |
| 1 | 11 | 0.5 |  |  |
| 1 | 12 | 0.5 |  |  |
| 2 | 3 | 0 |  |  |
| 2 | 4 | 0 |  |  |
| 2 | 5 | 0.5 |  |  |
| 2 | 6 | 0.5 |  |  |
| 2 | 7 | 0.5 |  |  |
| 2 | 8 | 0.5 |  |  |
| 2 | 9 | 0.5 |  |  |
| 2 | 10 | 0.5 |  |  |
| 2 | 11 | 0.5 |  |  |
| 2 | 12 | 0.5 |  |  |
| 3 | 4 | 0 |  |  |
| 3 | 5 | 0.5 |  |  |
| 3 | 6 | 0.5 |  |  |
| 3 | 7 | 0.5 |  |  |
| 3 | 8 | 0.5 |  |  |
| 3 | 9 | 0.5 |  |  |
| 3 | 10 | 0.5 |  |  |
| 3 | 11 | 0.5 |  |  |
| 3 | 12 | 0.5 |  |  |
| 4 | 5 | 0.5 |  |  |
| 4 | 6 | 0.5 |  |  |
| 4 | 7 | 0.5 |  |  |
| 4 | 8 | 0.5 |  |  |
| 4 | 9 | 0.5 |  |  |
| 4 | 10 | 0.5 |  |  |
| 4 | 11 | 0.5 |  |  |
| 4 | 12 | 0.5 |  |  |
| 5 | 6 | 1 |  |  |
| 5 | 7 | 1 |  |  |
| 5 | 8 | 1 |  |  |
| 5 | 9 | 1 |  |  |
| 5 | 10 | 1 |  |  |
| 5 | 11 | 1 |  |  |
| 5 | 12 | 1 |  |  |
| 6 | 7 | 1 |  |  |
| 6 | 8 | 1 |  |  |
| 6 | 9 | 1 |  |  |
| 6 | 10 | 1 |  |  |
| 6 | 11 | 1 |  |  |
| 6 | 12 | 1 |  |  |
| 7 | 8 | 1 |  |  |
| 7 | 9 | 1 |  |  |
| 7 | 10 | 1 |  |  |
| 7 | 11 | 1 |  |  |
| 7 | 12 | 1 |  |  |
| 8 | 9 | 1 |  |  |
| 8 | 10 | 1 |  |  |
| 8 | 11 | 1 |  |  |
| 8 | 12 | 1 |  |  |
| 9 | 10 | 1 |  |  |
| 9 | 11 | 1 |  |  |
| 9 | 12 | 1 |  |  |
| 10 | 11 | 1 |  |  |
| 10 | 12 | 1 |  |  |
| 11 | 12 | 1 |  |  |

(b) Show the distribution of the mean of the sample.

**Solution**

|  |  |  |
| --- | --- | --- |
| proportion older than 40 | # of samples with Value | Proportion of samples with value |
| 0 | 6 | 0.090909091 |
| 0.5 | 32 | 0.484848485 |
| 1 | 28 | 0.424242424 |

Since the number of farms that have owners older than 40 years of age exceeds the number of farms that have owners less than 40 years of age, the proportion is not showing left skewed behavior.

(c) A SRSWOR is taken resulting in the values of 3 people older than forty and 2 people younger than forty. Based on this sample, give an estimate of the proportion of farms owned by people over forty in the population and an estimate of the standard error of this mean.

**Solution**

|  |  |  |  |
| --- | --- | --- | --- |
| Farms in Random Sample | Owner older then 40 | proportion older than 40 | standard error of this mean |
| 2 | 0 | 0.6 | 0.547722558 |
| 4 | 0 |  |  |
| 9 | 1 |  |  |
| 12 | 1 |  |  |
| 7 | 1 |  |  |

5. **(20 points)** A school principal wants to estimate the average test score of its students in three tracts. You are given data with test scores from a proportional random sample of fifty students in the school from the three tracts (see sheet Q5Data in the file midtermDatSpring2018.xlsx). The principal thinks the simplest approach is to average all the scores and construct a confidence interval to demonstrate it is a good estimate. The entire first, second, and third tracts in the school contain 55, 80, and 65 students respectively.

When answering the questions below, do not use any statistical software and show your work to demonstrate you performed your calculations manually.

Part (1) (10 points) Show the principal what is the best design to calculate an estimate for the average score of students in the school using the sample data. Use the data and calculate average estimates and standard errors in making your argument.

**Solution**

Since its given that the marks are from 3 different tracts and the number of students in each of the tracts are different. Students in each tracts are highly likely to preform differently so the variance of marks within each tract should be different than one another. Considering tracts as strata, Stratified sampling should be ideal instead of SRS.





Mean is almost the same is both the approaches but StdErr from proportional allocation (1.516198) is almost 48.46% less then the StdErr from the principles approach (2.9416086) and hence proportional approach is the best approach.

Part (2) (10 points) The principal does some research and says to you he thinks a Neyman allocation may work best for estimating test scores. However, he is not sure how to do this allocation. Using the data you have from the proportional design, recommend a Neyman allocation for another sample.

**Solution**

To perform Neyman allocation, we have all the information required except for the availability of variance of population within each stratum. To get that information, either the entire population data is required, or an educated guess is required to be made. For educated guess, either the historical data, if available, can be checked or professor of the tracts can be contacted to get the guess from them. If it's not possible to determine the variance within strata, any other variable which is supposed to very in the same way as the variance does, can be use as proxy variable in place of variance.

For the given problem, we have a population of 200 with 3 strata's and the population of each strata is 55, 80, and 65 respectively. The data that is given to us is a proportional sample with 50 observations. Since the sample size is big and is selected based on the proportion of data within each stratum, we are assuming that the sample is representative of the population so we are using the standard error for each strata in the sample data given to determine the population size for Neyman allocation.



So if instead of proportional allocation, sample of size 50 corresponding to Neyman allocation has to be picked, then the sample size for items to be picked randomly from each stratum would be 12, 20 and 18 respectively.



**6. Lab exam (30 pts)**

An advertising firm wants to determine how much to emphasize television advertising in a given state county. The county has two towns A & B, and a rural area. The firm only has enough money to sample 100 households. The population data is given in sheet labExamData in midtermDatSpring2018.xlsx. The unit of observation is the household and the variable of interest is hours of TV viewing.

(a) (5 pts) Select a stratified random sample, using proportional allocation. What is the allocation of the sample for this design? Show your computations for determining this allocation.

**Solution**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Strata | # of Households (Nh) | N | (Nh/N) | n | nh = n\*(Nh/N) |
| A | 200 | 400 | 0.5 | 100 | 50 |
| B | 80 | 400 | 0.2 | 100 | 20 |
| R | 120 | 400 | 0.3 | 100 | 30 |

|  |  |  |
| --- | --- | --- |
| **Random Sample** | | |
| Observation Number | Hours | Place |
| 90 | 43 | A |
| 85 | 39 | A |
| 9 | 35 | A |
| 105 | 27 | A |
| 108 | 29 | A |
| 4 | 32 | A |
| 124 | 25 | A |
| 80 | 29 | A |
| 54 | 31 | A |
| 93 | 28 | A |
| 109 | 40 | A |
| 144 | 41 | A |
| 62 | 45 | A |
| 6 | 27 | A |
| 48 | 39 | A |
| 103 | 28 | A |
| 42 | 26 | A |
| 127 | 34 | A |
| 110 | 35 | A |
| 186 | 41 | A |
| 129 | 34 | A |
| 81 | 29 | A |
| 159 | 35 | A |
| 188 | 37 | A |
| 50 | 39 | A |
| 7 | 29 | A |
| 134 | 27 | A |
| 152 | 37 | A |
| 97 | 26 | A |
| 75 | 31 | A |
| 3 | 41 | A |
| 23 | 29 | A |
| 102 | 43 | A |
| 111 | 36 | A |
| 148 | 29 | A |
| 130 | 36 | A |
| 115 | 31 | A |
| 123 | 43 | A |
| 177 | 31 | A |
| 45 | 25 | A |
| 8 | 40 | A |
| 1 | 35 | A |
| 147 | 35 | A |
| 181 | 29 | A |
| 51 | 36 | A |
| 21 | 38 | A |
| 17 | 36 | A |
| 88 | 45 | A |
| 89 | 43 | A |
| 79 | 36 | A |
| 226 | 10 | B |
| 242 | 10 | B |
| 241 | 30 | B |
| 215 | 30 | B |
| 270 | 30 | B |
| 250 | 25 | B |
| 208 | 4 | B |
| 225 | 25 | B |
| 219 | 4 | B |
| 259 | 49 | B |
| 255 | 25 | B |
| 209 | 27 | B |
| 244 | 30 | B |
| 260 | 4 | B |
| 252 | 41 | B |
| 235 | 41 | B |
| 275 | 15 | B |
| 271 | 25 | B |
| 205 | 49 | B |
| 267 | 41 | B |
| 355 | 20 | R |
| 388 | 20 | R |
| 282 | 15 | R |
| 339 | 21 | R |
| 287 | 32 | R |
| 334 | 24 | R |
| 382 | 8 | R |
| 325 | 12 | R |
| 337 | 11 | R |
| 329 | 14 | R |
| 359 | 30 | R |
| 309 | 30 | R |
| 363 | 24 | R |
| 384 | 24 | R |
| 352 | 11 | R |
| 290 | 14 | R |
| 311 | 14 | R |
| 302 | 7 | R |
| 340 | 15 | R |
| 367 | 12 | R |
| 386 | 20 | R |
| 301 | 32 | R |
| 330 | 7 | R |
| 368 | 21 | R |
| 298 | 30 | R |
| 381 | 21 | R |
| 295 | 34 | R |
| 335 | 15 | R |
| 338 | 12 | R |
| 360 | 30 | R |

(b) (5 pts) Using SAS PROC SURVEYSELECT, select a stratified random sample from the population, using the allocation specified in (a). Use a random number seed of 1117 in choosing the sample. Give your SAS code, along with a list of the sample elements. (Hint: To make your life simpler, use the code below to recode place when you choose the sample).

data analDat; set inputData;   
 if place = "A" then strata = 1;  
 if place = "B" then strata = 2;  
 if place = "R" then strata = 3;  
run;

**Solution**

**SAS Code**

libname xl XLSX '/home/harisanadhya0/sasuser.v94/MSDS 6370/midtermDatSpring2018.xlsx';

libname xl list ;

PROC DATASETS LIBRARY=xl;

QUIT;

/\* Get the data into a SAS dataset \*/

data work.HouseHold;

set xl.labexamdata;

if place = "A" then strata = 1;

if place = "B" then strata = 2;

if place = "R" then strata = 3;

run;

/\* Sort the data by strata \*/

proc sort data=work.HouseHold;

by strata;

run;

/\* Compute the variance within strata (This is to be used in Neyman Allocation) \*/

proc means data=work.HouseHold var;

by strata;

output out=work.var var= variance;

run;

/\* View the dataset to verify that the import is successful \*/

proc print data=work.HouseHold(obs=10); run;

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

proc sql;

create table popStrsizes as

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

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

run;

/\* View the population strata record count table \*/

proc print data=work.popStrsizes; run;

/\* Select the sample using proportional allocation \*/

proc surveyselect data=HouseHold method = srs out=str1sample sampsize = 100 seed=1117;

strata strata/ alloc=prop allocmin=2;

title "Proportional allocation";

run;

/\* View the Proportional Sample \*/

proc print data=str1sample;

run;

List of Elements

| **Obs** | **strata** | **hours** | **place** | **Total** | **AllocProportion** | **SampleSize** | **ActualProportion** | **SelectionProb** | **SamplingWeight** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | 1 | 35 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **2** | 1 | 29 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **3** | 1 | 41 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **4** | 1 | 25 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **5** | 1 | 32 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **6** | 1 | 45 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **7** | 1 | 29 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **8** | 1 | 38 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **9** | 1 | 39 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **10** | 1 | 29 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **11** | 1 | 32 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **12** | 1 | 25 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **13** | 1 | 45 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **14** | 1 | 27 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **15** | 1 | 31 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **16** | 1 | 26 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **17** | 1 | 39 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **18** | 1 | 45 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **19** | 1 | 38 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **20** | 1 | 25 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **21** | 1 | 32 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **22** | 1 | 26 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **23** | 1 | 39 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **24** | 1 | 36 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **25** | 1 | 27 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **26** | 1 | 37 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **27** | 1 | 38 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **28** | 1 | 40 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **29** | 1 | 35 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **30** | 1 | 36 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **31** | 1 | 28 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **32** | 1 | 29 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **33** | 1 | 32 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **34** | 1 | 40 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **35** | 1 | 37 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **36** | 1 | 36 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **37** | 1 | 29 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **38** | 1 | 45 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **39** | 1 | 39 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **40** | 1 | 45 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **41** | 1 | 29 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **42** | 1 | 36 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **43** | 1 | 29 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **44** | 1 | 26 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **45** | 1 | 29 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **46** | 1 | 37 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **47** | 1 | 29 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **48** | 1 | 28 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **49** | 1 | 37 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **50** | 1 | 28 | A | 200 | 0.5 | 50 | 0.5 | 0.25 | 4 |
| **51** | 2 | 4 | B | 80 | 0.2 | 20 | 0.2 | 0.25 | 4 |
| **52** | 2 | 41 | B | 80 | 0.2 | 20 | 0.2 | 0.25 | 4 |
| **53** | 2 | 4 | B | 80 | 0.2 | 20 | 0.2 | 0.25 | 4 |
| **54** | 2 | 25 | B | 80 | 0.2 | 20 | 0.2 | 0.25 | 4 |
| **55** | 2 | 27 | B | 80 | 0.2 | 20 | 0.2 | 0.25 | 4 |
| **56** | 2 | 49 | B | 80 | 0.2 | 20 | 0.2 | 0.25 | 4 |
| **57** | 2 | 30 | B | 80 | 0.2 | 20 | 0.2 | 0.25 | 4 |
| **58** | 2 | 30 | B | 80 | 0.2 | 20 | 0.2 | 0.25 | 4 |
| **59** | 2 | 15 | B | 80 | 0.2 | 20 | 0.2 | 0.25 | 4 |
| **60** | 2 | 10 | B | 80 | 0.2 | 20 | 0.2 | 0.25 | 4 |
| **61** | 2 | 25 | B | 80 | 0.2 | 20 | 0.2 | 0.25 | 4 |
| **62** | 2 | 4 | B | 80 | 0.2 | 20 | 0.2 | 0.25 | 4 |
| **63** | 2 | 49 | B | 80 | 0.2 | 20 | 0.2 | 0.25 | 4 |
| **64** | 2 | 30 | B | 80 | 0.2 | 20 | 0.2 | 0.25 | 4 |
| **65** | 2 | 30 | B | 80 | 0.2 | 20 | 0.2 | 0.25 | 4 |
| **66** | 2 | 25 | B | 80 | 0.2 | 20 | 0.2 | 0.25 | 4 |
| **67** | 2 | 41 | B | 80 | 0.2 | 20 | 0.2 | 0.25 | 4 |
| **68** | 2 | 10 | B | 80 | 0.2 | 20 | 0.2 | 0.25 | 4 |
| **69** | 2 | 15 | B | 80 | 0.2 | 20 | 0.2 | 0.25 | 4 |
| **70** | 2 | 49 | B | 80 | 0.2 | 20 | 0.2 | 0.25 | 4 |
| **71** | 3 | 20 | R | 120 | 0.3 | 30 | 0.3 | 0.25 | 4 |
| **72** | 3 | 8 | R | 120 | 0.3 | 30 | 0.3 | 0.25 | 4 |
| **73** | 3 | 8 | R | 120 | 0.3 | 30 | 0.3 | 0.25 | 4 |
| **74** | 3 | 34 | R | 120 | 0.3 | 30 | 0.3 | 0.25 | 4 |
| **75** | 3 | 34 | R | 120 | 0.3 | 30 | 0.3 | 0.25 | 4 |
| **76** | 3 | 24 | R | 120 | 0.3 | 30 | 0.3 | 0.25 | 4 |
| **77** | 3 | 24 | R | 120 | 0.3 | 30 | 0.3 | 0.25 | 4 |
| **78** | 3 | 12 | R | 120 | 0.3 | 30 | 0.3 | 0.25 | 4 |
| **79** | 3 | 14 | R | 120 | 0.3 | 30 | 0.3 | 0.25 | 4 |
| **80** | 3 | 7 | R | 120 | 0.3 | 30 | 0.3 | 0.25 | 4 |
| **81** | 3 | 7 | R | 120 | 0.3 | 30 | 0.3 | 0.25 | 4 |
| **82** | 3 | 14 | R | 120 | 0.3 | 30 | 0.3 | 0.25 | 4 |
| **83** | 3 | 21 | R | 120 | 0.3 | 30 | 0.3 | 0.25 | 4 |
| **84** | 3 | 11 | R | 120 | 0.3 | 30 | 0.3 | 0.25 | 4 |
| **85** | 3 | 8 | R | 120 | 0.3 | 30 | 0.3 | 0.25 | 4 |
| **86** | 3 | 24 | R | 120 | 0.3 | 30 | 0.3 | 0.25 | 4 |
| **87** | 3 | 11 | R | 120 | 0.3 | 30 | 0.3 | 0.25 | 4 |
| **88** | 3 | 12 | R | 120 | 0.3 | 30 | 0.3 | 0.25 | 4 |
| **89** | 3 | 7 | R | 120 | 0.3 | 30 | 0.3 | 0.25 | 4 |
| **90** | 3 | 7 | R | 120 | 0.3 | 30 | 0.3 | 0.25 | 4 |
| **91** | 3 | 20 | R | 120 | 0.3 | 30 | 0.3 | 0.25 | 4 |
| **92** | 3 | 30 | R | 120 | 0.3 | 30 | 0.3 | 0.25 | 4 |
| **93** | 3 | 21 | R | 120 | 0.3 | 30 | 0.3 | 0.25 | 4 |
| **94** | 3 | 30 | R | 120 | 0.3 | 30 | 0.3 | 0.25 | 4 |
| **95** | 3 | 24 | R | 120 | 0.3 | 30 | 0.3 | 0.25 | 4 |
| **96** | 3 | 8 | R | 120 | 0.3 | 30 | 0.3 | 0.25 | 4 |
| **97** | 3 | 24 | R | 120 | 0.3 | 30 | 0.3 | 0.25 | 4 |
| **98** | 3 | 15 | R | 120 | 0.3 | 30 | 0.3 | 0.25 | 4 |
| **99** | 3 | 11 | R | 120 | 0.3 | 30 | 0.3 | 0.25 | 4 |
| **100** | 3 | 12 | R | 120 | 0.3 | 30 | 0.3 | 0.25 | 4 |

(c) (5 pts) Use SAS PROC SURVEYMEANS with the sample above to compute these statistics for the hours of TV viewing: The estimate of the population mean, the standard error of the estimate, the 95% confidence interval for the mean. Paste your SAS code and output below. (Do not include the graphics in your paste.)

**Solution**

SAS CODE

proc SURVEYMEANS data=str1sample total=popStrsizes

mean clm stderr;

var hours;

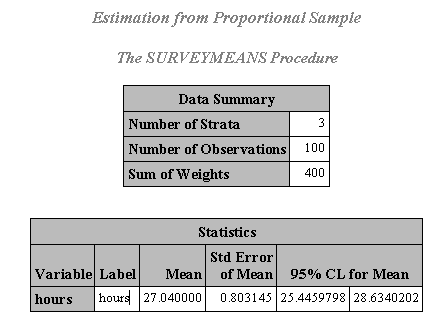
weight samplingweight;

strata Strata ;

title "Estimation from Proportional Sample" ;

run;

Output



(d) (5 pts) A Neyman allocation of the sample is 38, 33, and 29 for strata A, B, and R respectively. Using SAS PROC SURVEYSELECT, select a stratified random sample from the population, using the Neyman allocation. Use a random number seed of 1118. Give your SAS code.

**Solution**

SAS CODE

/\* Select the sample using Neyman allocation \*/

proc surveyselect data=HouseHold method = srs out=str2sample sampsize = (38, 33, 29) seed=1118;

strata strata;

title "Neyman allocation";

run;

/\* View the Neyman Sample \*/

proc print data=str2sample;

run;

Random Sample selected

| **Obs** | **strata** | **hours** | **place** | **SelectionProb** | **SamplingWeight** |
| --- | --- | --- | --- | --- | --- |
| **1** | 1 | 41 | A | 0.19000 | 5.26316 |
| **2** | 1 | 28 | A | 0.19000 | 5.26316 |
| **3** | 1 | 41 | A | 0.19000 | 5.26316 |
| **4** | 1 | 35 | A | 0.19000 | 5.26316 |
| **5** | 1 | 29 | A | 0.19000 | 5.26316 |
| **6** | 1 | 26 | A | 0.19000 | 5.26316 |
| **7** | 1 | 34 | A | 0.19000 | 5.26316 |
| **8** | 1 | 38 | A | 0.19000 | 5.26316 |
| **9** | 1 | 31 | A | 0.19000 | 5.26316 |
| **10** | 1 | 25 | A | 0.19000 | 5.26316 |
| **11** | 1 | 41 | A | 0.19000 | 5.26316 |
| **12** | 1 | 35 | A | 0.19000 | 5.26316 |
| **13** | 1 | 45 | A | 0.19000 | 5.26316 |
| **14** | 1 | 40 | A | 0.19000 | 5.26316 |
| **15** | 1 | 38 | A | 0.19000 | 5.26316 |
| **16** | 1 | 37 | A | 0.19000 | 5.26316 |
| **17** | 1 | 27 | A | 0.19000 | 5.26316 |
| **18** | 1 | 28 | A | 0.19000 | 5.26316 |
| **19** | 1 | 31 | A | 0.19000 | 5.26316 |
| **20** | 1 | 40 | A | 0.19000 | 5.26316 |
| **21** | 1 | 32 | A | 0.19000 | 5.26316 |
| **22** | 1 | 31 | A | 0.19000 | 5.26316 |
| **23** | 1 | 28 | A | 0.19000 | 5.26316 |
| **24** | 1 | 36 | A | 0.19000 | 5.26316 |
| **25** | 1 | 40 | A | 0.19000 | 5.26316 |
| **26** | 1 | 41 | A | 0.19000 | 5.26316 |
| **27** | 1 | 34 | A | 0.19000 | 5.26316 |
| **28** | 1 | 37 | A | 0.19000 | 5.26316 |
| **29** | 1 | 29 | A | 0.19000 | 5.26316 |
| **30** | 1 | 45 | A | 0.19000 | 5.26316 |
| **31** | 1 | 29 | A | 0.19000 | 5.26316 |
| **32** | 1 | 41 | A | 0.19000 | 5.26316 |
| **33** | 1 | 31 | A | 0.19000 | 5.26316 |
| **34** | 1 | 26 | A | 0.19000 | 5.26316 |
| **35** | 1 | 41 | A | 0.19000 | 5.26316 |
| **36** | 1 | 29 | A | 0.19000 | 5.26316 |
| **37** | 1 | 40 | A | 0.19000 | 5.26316 |
| **38** | 1 | 38 | A | 0.19000 | 5.26316 |
| **39** | 2 | 27 | B | 0.41250 | 2.42424 |
| **40** | 2 | 4 | B | 0.41250 | 2.42424 |
| **41** | 2 | 27 | B | 0.41250 | 2.42424 |
| **42** | 2 | 49 | B | 0.41250 | 2.42424 |
| **43** | 2 | 30 | B | 0.41250 | 2.42424 |
| **44** | 2 | 49 | B | 0.41250 | 2.42424 |
| **45** | 2 | 25 | B | 0.41250 | 2.42424 |
| **46** | 2 | 41 | B | 0.41250 | 2.42424 |
| **47** | 2 | 41 | B | 0.41250 | 2.42424 |
| **48** | 2 | 15 | B | 0.41250 | 2.42424 |
| **49** | 2 | 4 | B | 0.41250 | 2.42424 |
| **50** | 2 | 15 | B | 0.41250 | 2.42424 |
| **51** | 2 | 27 | B | 0.41250 | 2.42424 |
| **52** | 2 | 49 | B | 0.41250 | 2.42424 |
| **53** | 2 | 30 | B | 0.41250 | 2.42424 |
| **54** | 2 | 30 | B | 0.41250 | 2.42424 |
| **55** | 2 | 41 | B | 0.41250 | 2.42424 |
| **56** | 2 | 41 | B | 0.41250 | 2.42424 |
| **57** | 2 | 10 | B | 0.41250 | 2.42424 |
| **58** | 2 | 10 | B | 0.41250 | 2.42424 |
| **59** | 2 | 25 | B | 0.41250 | 2.42424 |
| **60** | 2 | 15 | B | 0.41250 | 2.42424 |
| **61** | 2 | 27 | B | 0.41250 | 2.42424 |
| **62** | 2 | 30 | B | 0.41250 | 2.42424 |
| **63** | 2 | 30 | B | 0.41250 | 2.42424 |
| **64** | 2 | 15 | B | 0.41250 | 2.42424 |
| **65** | 2 | 30 | B | 0.41250 | 2.42424 |
| **66** | 2 | 41 | B | 0.41250 | 2.42424 |
| **67** | 2 | 27 | B | 0.41250 | 2.42424 |
| **68** | 2 | 10 | B | 0.41250 | 2.42424 |
| **69** | 2 | 15 | B | 0.41250 | 2.42424 |
| **70** | 2 | 25 | B | 0.41250 | 2.42424 |
| **71** | 2 | 41 | B | 0.41250 | 2.42424 |
| **72** | 3 | 34 | R | 0.24167 | 4.13793 |
| **73** | 3 | 15 | R | 0.24167 | 4.13793 |
| **74** | 3 | 8 | R | 0.24167 | 4.13793 |
| **75** | 3 | 34 | R | 0.24167 | 4.13793 |
| **76** | 3 | 30 | R | 0.24167 | 4.13793 |
| **77** | 3 | 7 | R | 0.24167 | 4.13793 |
| **78** | 3 | 15 | R | 0.24167 | 4.13793 |
| **79** | 3 | 11 | R | 0.24167 | 4.13793 |
| **80** | 3 | 14 | R | 0.24167 | 4.13793 |
| **81** | 3 | 24 | R | 0.24167 | 4.13793 |
| **82** | 3 | 12 | R | 0.24167 | 4.13793 |
| **83** | 3 | 34 | R | 0.24167 | 4.13793 |
| **84** | 3 | 11 | R | 0.24167 | 4.13793 |
| **85** | 3 | 14 | R | 0.24167 | 4.13793 |
| **86** | 3 | 11 | R | 0.24167 | 4.13793 |
| **87** | 3 | 24 | R | 0.24167 | 4.13793 |
| **88** | 3 | 12 | R | 0.24167 | 4.13793 |
| **89** | 3 | 12 | R | 0.24167 | 4.13793 |
| **90** | 3 | 32 | R | 0.24167 | 4.13793 |
| **91** | 3 | 7 | R | 0.24167 | 4.13793 |
| **92** | 3 | 24 | R | 0.24167 | 4.13793 |
| **93** | 3 | 11 | R | 0.24167 | 4.13793 |
| **94** | 3 | 24 | R | 0.24167 | 4.13793 |
| **95** | 3 | 12 | R | 0.24167 | 4.13793 |
| **96** | 3 | 21 | R | 0.24167 | 4.13793 |
| **97** | 3 | 20 | R | 0.24167 | 4.13793 |
| **98** | 3 | 21 | R | 0.24167 | 4.13793 |
| **99** | 3 | 15 | R | 0.24167 | 4.13793 |
| **100** | 3 | 8 | R | 0.24167 | 4.13793 |

e) (5 pts) Use SAS PROC SURVEYMEANS with the Neyman sample compute these statistics for the hours of TV viewing: The estimate of the population mean, the standard error of the estimate, the 95% confidence interval for the mean. Paste your SAS code and output below. (Do not include the graphics in your paste.)

**Solution**

SAS CODE

proc SURVEYMEANS data=str2sample total=popStrsizes

mean clm stderr;

var hours;

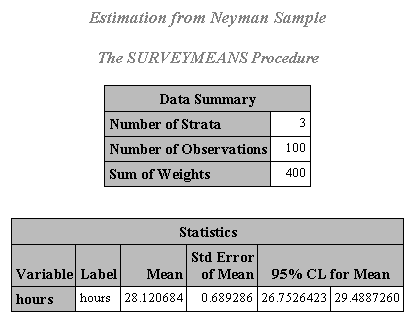
weight samplingweight;

strata Strata ;

title "Estimation from Neyman Sample" ;

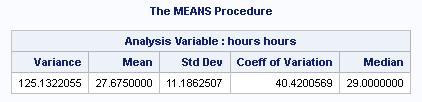
run;

OUTPUT



f) (5 pts) Comment on whether the Neyman allocation gives better or worse estimates.

**Solution**



Mean obtained from both Neyman and Proportional allocation are different and almost equidistant from the true population mean (27.04 from Proportional, 28.12 from Neyman and 27.65 is the population mean) but in case of Neyman, the standard error (0.689) of mean is almost 14.2% less then the standard error (0.803) of mean from proportional allocation. Also in both the cases, the true mean lies within the 95% CI of mean. Hence it can be said that Neyman allocation performs better then proportional allocation.