**Sampling Estimation of Tuition Expense for 4-Year Institutions**

**MSDS 6370**

**Laurie Harris, Tim McWilliams**

**April 4, 2018**

**1. Introduction**

The Federal Office of Post-Secondary Education compiles an annual study on College Affordability and Transparency. Several data points are collected including tuition expense, non-tuition expense, institutional budgets and expenditures, etc. It is interesting to note that participation by institutions in this study is mandatory and a 100% collection rate is expected.

**2. Objective**

Apply two different sampling techniques to calculate mean tuition for 4-year universities and compare results. Using five independent samples for each technique, determine how many 95% confidence intervals contain the true mean.

**3. Data Description**

Data were obtained from the College Affordability and Transparency List Explanation Form, 2016-17. [<https://catalog.data.gov/dataset/college-affordability-and-transparency-list-explanation-form-201617> ]. The data contained measurements related to rates for the 2014-2015 school year. Data is accumulated for all post-secondary institutions. These include 4-year universities, 2-year programs, and institutions that offer less than 2-year programs. In addition, institutions are further categorized as public, private not-for-profit, and private for-profit.

This analysis was limited to the measurement of Tuition and Fees for 4-year institutions. Therefore, we considered information for 2,640 institutions over three categories: 4-year public, 4-year private not-for-profit, and 4-year private for-profit.

**4. Sample Size Calculation**

With the variable of interest (mean tuition) and population of 4-year institutions (2,640) known, we prepare to calculate the required sample size. For this study, we felt that a margin of error of $2,000 and 95% confidence interval were reasonable. The standard deviation from a prior study for the 2010-2011 school year was calculated to be $9,894 therefore we will use this value as our estimate of the standard deviation.

The sample size is calculated in Figure 1 using these parameters and the result is then adjusted to with the finite population correction.

|  |
| --- |
| Population = 2,640  Variable of interest = mean tuition  Margin of error = $2,000  CI = 95%  Standard deviation estimated = 9894 |

Figure 1. Simple Random Sample Calculations

Based on the above calculation, a sample size of 91 would be required to achieve a margin of error of $2,000 with 95% confidence.

**5. Sampling Methods**

Two separate sampling methods were used in this report. Frist, a simple random sample was conducted. Then, a stratified sample was conducted. For the stratified sample, the Neyman allocation was used. The estimates from the two sampling methods will be compared. To further compare, there will be five different samples selected from both the simple random sample and the Neyman allocation.

**5.1 Simple Random Sample**

After loading the data into SAS, a simple random sample was selected using PROC SURVEYSELECT. The mean, standard error and 95% confidence interval were calculated using PROC SURVEYMEANS and the result is shown below in Figure 2.

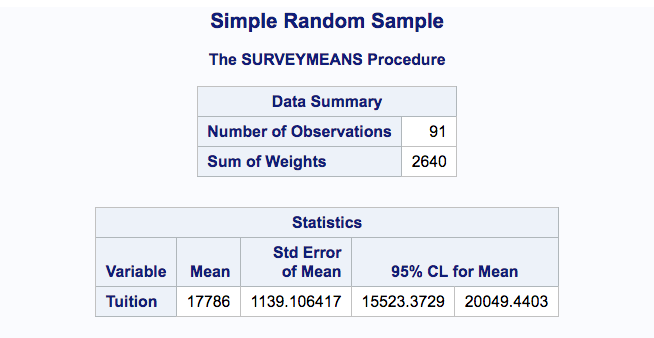


Figure 2. Simple Random Sample

**5.2 Neyman Allocation**

A stratified sample was then selected using Neyman allocation. The Neyman allocation takes the sample size proportional to the stratum size multiplied by the standard deviation in the stratum. This requires knowledge of the standard deviation of each stratum. The sample size is shown in Table 1.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | Stratum | Description | Population | Sh | Nh \* Sh | nh | Sample | | 1 | 4-year public | 676 | 103.39 | 69891.64 | 13.54 | 14 | | 2 | 4-year private not-for-profit | 1322 | 234.46 | 309956.1 | 60.05 | 60 | | 3 | 4-year private for-profit | 642 | 139.97 | 89860.74 | 17.40 | 17 | | Total |  | **2640** |  |  |  | **91** | |

Table 1. Stratified Sample – Neyman Allocation Sample Calculation

**5.2.1 Neyman Allocation Results**

Neyman allocation results in the smallest variance for the estimator of the mean, in this case the mean tuition, of any stratified sample with the same *n*. Therefore, this is the best method of stratification there is to use. Results are shown in Figure 3.

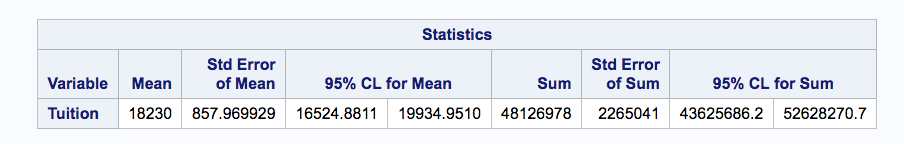


Figure 3. Stratified Sample – Neyman Allocation Results

**5.3 Comparison of Initial Sampling Methods**

When comparing the sampling methods, there are slight differences with the mean, standard deviation and the 95% confidence interval. As shown in Table 2, the Neyman allocation has a smaller standard deviation as well as a smaller confidence interval.

Table 2. Comparison of SRS and Stratified Sample – Neyman Allocation

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | |  | SRS | Neyman Allocation | | Mean | $17, 786 | $18, 230 | | Stdv | $1, 139.11 | $857.97 | | 95% CI | [ $15, 523.27, $20 049.44 ] | [ $16, 524.88, $19, 934.95 ] | |

**5.4 Comparison of 5 samples from each method**

To further examine the difference between the two sampling methods, SRS and Stratified with Neyman Allocation, five independent samples using each method were taken. The means, standard errors, and confidence intervals were calculated. Because the population known and relatively small, we are able to compute the true population mean to be $18,257. The independent samples were then reviewed to identify how many of their confidence intervals contained the population true mean. Results are shown in Table 3.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | Method | Instance | Mean | Standard Error | LCL | UCL | True Mean Contained | | SRS | 1 | 18,788 | 1,202.91 | 16,398.52 | 21,178.10 | Yes | | SRS | 2 | 17,885 | 1,153.64 | 15,592.68 | 20,176.51 | Yes | | SRS | 3 | 20,501 | 1,156.82 | 18,202.39 | 22,798.85 | Yes | | SRS | 4 | 18,287 | 1,276.32 | 1,5751.21 | 20,822.49 | Yes | | SRS | 5 | 17,969 | 1,204.27 | 15,576.43 | 20,361.40 | Yes | | Neyman | 1 | 17,438 | 779.56 | 15,888.89 | 18,987.81 | Yes | | Neyman | 2 | 17,426 | 865.47 | 15,706.40 | 19,146.28 | Yes | | Neyman | 3 | 18,871 | 889.20 | 17,104.00 | 20,638.18 | Yes | | Neyman | 4 | 18,001 | 830.71 | 16,350.06 | 19,651.79 | Yes | | Neyman | 5 | 18,116 | 871.80 | 16,383.73 | 19,848.77 | Yes | |

Table 3. Comparison of independent samples to population true mean

**5.5 Observations:**

All confidence intervals for the independent samples contained the true mean. The five stratified Neyman allocation samples had smaller standard errors and more precise confidence intervals. Which would suggest that using the Neyman allocation for sampling would be wiser.

**6. Conclusion**

The stratification sampling method using the Neyman allocation appears to provide the most precise confidence intervals in estimating the mean Tuition and Fees. Although both sampling methods give mean values in a similar range, the stratification method using the Neyman allocation would be the preferred sampling method for this analysis. Therefore, we recommend that when one needs to sample this data they are to use Neyman allocation.

**Appendix – Analysis Code**

/\* Bring in data \*/

data tuition;

set statsamp.TUITION;

run;

/\*

proc print data = tuition;

run;

\*/

/\*SRS select sample\*/

proc surveyselect data = tuition

out = srssample1

sampsize = 91

seed = 16 stats;

title "Simple Random Sample";

run;

/\*

proc print data = srssample1;

run;

\*/

/\*mean for simple random sample\*/

proc surveymeans data = srssample1 total = 2640 mean clm;

var Tuition;

weight SamplingWeight;

title "Simple Random Sample";

run;

/\*Stratified Sample data step\*/

data tuition\_st;

set tuition;

if Sector\_name = "4-year public" then stratum = 1;

else if Sector\_name = "4-year pnfp" then stratum = 2;

else stratum = 3;

run;

/\*

proc print data = tuition\_st;

run;

\*/

/\*select stratified sample based on proportional allocation\*/

proc surveyselect data = tuition\_st method = srs out = strsample1 sampsize = (23,46,22) seed = 816;

strata stratum;

title "Proportional Allocation";

run;

/\*

proc print data = strsample1;

run;

\*/

/\*set up totals for strata\*/

data strsizes;

input stratum \_total\_;

datalines;

1 676

2 1322

3 642

;

run;

/\*mean for stratified design\*/

proc surveymeans data = strsample1 total = strsizes mean clm;

var Tuition;

weight SamplingWeight;

title "Proportional Allocation";

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\* Stratified Sampling: Neyman Allocation \*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\* Create Strata \*/

data tuition\_stn;

set tuition\_st;

run;

proc print data=tuition\_stn;

run;

/\* STD for Neyman Allocation \*/

/\* Sort data first \*/

proc sort data=tuition\_stn;

by Sector\_name;

run;

proc surveymeans data=tuition\_stn total=2640 mean std clm;

var Tuition;

by Sector\_name;

run;

/\* Select sample size based on Neyman Allocation \*/

/\* Re-sort the data by stratum\*/

proc sort data=tuition\_stn;

by stratum;

run;

proc surveyselect data=tuition\_stn method=SRS out=NeymanAllocation sampsize=(14, 60, 17) seed=1111;

strata Stratum;

title "Neyman Allocation";

run;

/\*

proc print data=NeymanAllocation;

run;

\*/

/\* Set the totals for each strata \*/

data strsizes;

input stratum \_total\_;

datalines;

1 676

2 1322

3 642

;

run;

/\* Analyzing the sample \*/

proc surveymeans data=NeymanAllocation mean clm sum clsum total=strsizes;

var Tuition;

weight SamplingWeight;

strata Stratum;

title "Neyman Allocaiton";

run;

/\*\*\*\*\*\*\*\*\*\*5 iterations of Neyman\*\*\*\*\*\*\*\*\*\*\*\*\*/

proc surveyselect data=tuition\_stn method=SRS out=NeymanAllocation1 sampsize=(14, 60, 17) seed=2222;

strata Stratum;

title "Neyman Allocation1";

run;

proc surveymeans data=NeymanAllocation1 mean clm sum clsum total=strsizes;

var Tuition;

weight SamplingWeight;

strata Stratum;

title "Neyman Allocaiton1";

run;

proc surveyselect data=tuition\_stn method=SRS out=NeymanAllocation2 sampsize=(14, 60, 17) seed=487;

strata Stratum;

title "Neyman Allocation2";

run;

proc surveymeans data=NeymanAllocation2 mean clm sum clsum total=strsizes;

var Tuition;

weight SamplingWeight;

strata Stratum;

title "Neyman Allocaiton2";

run;

proc surveyselect data=tuition\_stn method=SRS out=NeymanAllocation3 sampsize=(14, 60, 17) seed=892;

strata Stratum;

title "Neyman Allocation3";

run;

proc surveymeans data=NeymanAllocation3 mean clm sum clsum total=strsizes;

var Tuition;

weight SamplingWeight;

strata Stratum;

title "Neyman Allocaiton3";

run;

proc surveyselect data=tuition\_stn method=SRS out=NeymanAllocation4 sampsize=(14, 60, 17) seed=5832;

strata Stratum;

title "Neyman Allocation4";

run;

proc surveymeans data=NeymanAllocation4 mean clm sum clsum total=strsizes;

var Tuition;

weight SamplingWeight;

strata Stratum;

title "Neyman Allocaiton4";

run;

proc surveyselect data=tuition\_stn method=SRS out=NeymanAllocation5 sampsize=(14, 60, 17) seed=1683;

strata Stratum;

title "Neyman Allocation5";

run;

proc surveymeans data=NeymanAllocation5 mean clm sum clsum total=strsizes;

var Tuition;

weight SamplingWeight;

strata Stratum;

title "Neyman Allocaiton5";

run;

/\*\*\*\*\*\*\*\*\*\*5 iterations of SRS\*\*\*\*\*\*\*\*\*\*\*\*\*/

proc surveyselect data = tuition

out = srssample2

sampsize = 91

seed = 456 stats;

title "Simple Random Sample1";

run;

/\*mean for simple random sample\*/

proc surveymeans data = srssample2 total = 2640 mean clm;

var Tuition;

weight SamplingWeight;

title "Simple Random Sample1";

run;

proc surveyselect data = tuition

out = srssample3

sampsize = 91

seed = 789 stats;

title "Simple Random Sample2";

run;

/\*mean for simple random sample\*/

proc surveymeans data = srssample3 total = 2640 mean clm;

var Tuition;

weight SamplingWeight;

title "Simple Random Sample2";

run;

proc surveyselect data = tuition

out = srssample4

sampsize = 91

seed = 6042 stats;

title "Simple Random Sample3";

run;

/\*mean for simple random sample\*/

proc surveymeans data = srssample4 total = 2640 mean clm;

var Tuition;

weight SamplingWeight;

title "Simple Random Sample3";

run;

proc surveyselect data = tuition

out = srssample5

sampsize = 91

seed = 9734 stats;

title "Simple Random Sample4";

run;

/\*mean for simple random sample\*/

proc surveymeans data = srssample5 total = 2640 mean clm;

var Tuition;

weight SamplingWeight;

title "Simple Random Sample4";

run;

proc surveyselect data = tuition

out = srssample6

sampsize = 91

seed = 7742 stats;

title "Simple Random Sample5";

run;

/\*mean for simple random sample\*/

proc surveymeans data = srssample6 total = 2640 mean clm;

var Tuition;

weight SamplingWeight;

title "Simple Random Sample5";

run;