ST205 Coursework

Authors

The candidate numbers of our group are as follows:

83751, 34144, 32246, 30085

Introduction

We have used coding in order to complete the exercises in this coursework, as we feel that this is our strength as a group. We have chosen to submit the project in this format as this is the best way to display our methods.

At the end of the project under Final answers you will find a table containing all of our answers for estimated values, standard errors and confidence intervals. These values are also displayed in their respective question sections along with written/other

Part a)

Loading the required library: We will use the dplyr library, designed to facilitate data manipulation and analysis.

```
library(dplyr)
```

Reading the csv file into a dataframe:

```
df <- read.csv("C:/Users/User/Documents/ST205/collegedataset.csv")
```

Generating 20 random numbers between 1 and the last row on the dataset:

```
set.seed(123)  # Setting seed for reproducibility
SRS_random_numbers <- sample(1:nrow(df), 20, replace = FALSE)
# replace = FALSE because we don't want to get the same number twice
```

Here are the random numbers that we generated:

```
print(SRS_random_numbers)
```

```
 [1]  415  463  179  526  195  938 1142 1323 1253 1268 1038  665  602  709 1011
[16] 1115  953  348 1017  840
```

Here are the selected colleges:

```
print(df$Institution[SRS_random_numbers])
```

```
 [1] "Thomas More University"
 [2] "Johns Hopkins University"
 [3] "Embry-Riddle Aeronautical University-Daytona Beach"
 [4] "Western New England University"
 [5] "Saint Leo University"
 [6] "Langston University"
 [7] "Union University"
 [8] "University of Wisconsin-La Crosse"
 [9] "Randolph College"
[10] "Central Washington University"
[11] "Saint Francis University"
[12] "University of Nebraska at Omaha"
[13] "Millsaps College"
[14] "Stevens Institute of Technology"
[15] "La Roche University"
[16] "Christian Brothers University"
[17] "University of Tulsa"
[18] "Coe College"
[19] "Lincoln University"
[20] "Johnson C Smith University"
```

Creating a dataframe to include only the selected Institutions:

```
SRS_df <- df[SRS_random_numbers, ]
```

Part b)

i)

Means

We will use the formula:

$$\widehat{\overline{y_U}} = \overline{y_s} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Where n = 20, as this is our sample size.

```
tuitionfee_in_mean <- sum(SRS_df$tuitionfee_in)/20
print(tuitionfee_in_mean)
```

```
[1] 29435.25
```

```
tuitionfee_out_mean <- sum(SRS_df$tuitionfee_out)/20
print(tuitionfee_out_mean)
```

```
[1] 31930.25
```

Standard errors

The formula for the variance of the sample mean is:

$$var(\bar{y}_s) = \frac{s^2(1-f)}{n}$$

Where:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{N}(y_i - \bar{y}_s)^2$$

is the sample variance, and $f = n/N$.

```
sample_var_tuitionfee_in <- 1/(20-1)*sum((SRS_df$tuitionfee_in - tuitionfee_in_mean)^2)
print(sample_var_tuitionfee_in)
```

```
[1] 222742967
```

```
sample_var_tuitionfee_out <- 1/(20-1)*sum((SRS_df$tuitionfee_out - tuitionfee_out_mean)^2)
print(sample_var_tuitionfee_out)
```

```
[1] 134271747
```

We can now use these values to calculate the variances of the estimated means of tuitionfee_in and tuitionfee_out:

```
var_tuitionfee_in_est <- sample_var_tuitionfee_in*(1-20/1372)/20
print(var_tuitionfee_in_est)
```

```
[1] 10974799
```

```
var_tuitionfee_out_est <- sample_var_tuitionfee_out*(1-20/1372)/20
print(var_tuitionfee_out_est)
```

```
[1] 6615722
```

We can find the respective standard errors by taking the square root of the variances:

```
SE_tuitionfee_in_est <- var_tuitionfee_in_est^(1/2)
print(SE_tuitionfee_in_est)
```

```
[1] 3312.823
```

```
SE_tuitionfee_out_est <- var_tuitionfee_out_est^(1/2)
print(SE_tuitionfee_out_est)
```

```
[1] 2572.105
```

Hypothesis test

Let x = tuitionfee_in and y = tuitionfee_out.

We will test the hypotheses:

$$H_0: \mu_x - \mu_y = 0 \qquad \text{vs} \qquad H_1: \mu_x - \mu_y \neq 0$$

Note that:

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}} \sim N(0,1)$$

Under the null hypothesis, we have:

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}} \sim N(0,1)$$

At the 95% significance level, we reject the null hypothesis if:

$$|t| > z_{\alpha/2} = z_{0.025} = 1.96$$

We have:

```
t <- ((20+20-2)/(1/20+1/20))^(1/2)*
  ((tuitionfee_in_mean - tuitionfee_out_mean)/
     ((20-1)*var_tuitionfee_in_est + (20-1)*var_tuitionfee_out_est))
print(t)
```

```
[1] -0.0001455225
```

Interpretation

The modulus of our t value is 0.000146, which is far less than 1.96. Therefore, we do not reject the null hypothesis and conclude that there is no evidence to suggest a significant difference between the in-state and out-state tuition and fees across all colleges in the United States.

ii)

Proportion

To estimate the proportion of Black/African American students across all colleges in the US, we will use the same formula as before:

$$\widehat{\overline{y_U}} = \overline{y_s} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

```
ugds_black_mean <- mean(SRS_df$ugds_black)
print(ugds_black_mean)
```

```
[1] 0.203875
```

Standard error

The formula for the variance of the sample mean is:

$$var(\overline{y}_s) = \frac{s^2(1-f)}{n}$$

Where:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{N} (y_i - \overline{y}_s)^2$$

is the sample variance, and $f = n/N$.

```
SRS_var_ugds_black <- (1/(20-1))*sum((SRS_df$ugds_black-ugds_black_mean)^2)
print(SRS_var_ugds_black)
```

```
[1] 0.06597262
```

The variance of the sample mean can now be calculated:

```
var_ugds_black_mean <- SRS_var_ugds_black*(1-20/1372)/20
print(var_ugds_black_mean)
```

```
[1] 0.003250546
```

The standard error is the square root of this variance:

```
SE_ugds_black <- var_ugds_black_mean^(1/2)
print(SE_ugds_black)
```

```
[1] 0.05701356
```

Confidence interval

A 95% confidence interval for the proportion of Black/African American undergraduate students across all colleges is given by:

$$\bar{y}_s \pm 1.96 \times SE(\bar{y}_s)$$

```
SRS_upper <- ugds_black_mean + 1.96*SE_ugds_black
print(SRS_upper)
```

```
[1] 0.3156216
```

```
SRS_lower <- ugds_black_mean - 1.96*SE_ugds_black
print(SRS_lower)
```

```
[1] 0.09212842
```

So, the 95% confidence interval for the proportion of Black/African American students across all colleges in the US is (0.0921, 0.3156).

Part c)

Stratified sampling

We would ideally use stratified sampling for data where we observe heterogeneity between strata and homogeneity within strata. We would chose our strata based on the variable of interest. For instance, if our research question was 'Are universities with more gender diversity more expensive?', we would divide our strata by proportion of women (e.g. strata would be [0 - 0.2] [0.2-0.4] [0.4-0.6] [0.6-0.8] and [0.8-1]). Additionally, the stratification factor should be more evenly distributed rather than following, say, a normal distribution. The proportion of women across universities seems to be normally distributed, hence why we have not used this as our stratification factor for part d).

To determine the number of elements to be selected from each stratum, we could use proportional allocation, Neyman allocation, equal allocation or cost allocation:

- **Proportional allocation**: Best for representation. The number of elements chosen are proportional to the size of the stratum in the overall population. Insert formula from lecture/class
- **Neyman allocation**: Used to maximise precision. More elements are selected from strata with higher variance to reduce the sensitivity of the data. It can only be calculated if we know the variance in each stratum.
- **Equal allocation**: Select the same number of elements from each stratum. We would choose 4 universities through SRS from each stratum.
- **Cost allocation**: If the cost of surveying each stratum is different. Select the most cost effective strata according to budget allocation.

Cluster sampling

We would use cluster sampling for data where we observe homogeneity between groups and heterogeneity within groups. In the context of this dataset, we would probably perform two-stage

cluster sampling; grouping institutions into clusters (e.g. geographically, based on state or region), then randomly selecting 5 clusters followed by 4 institutions from each cluster to make a sample of 20.

Pros and cons

	Pros	Cons
Stratified	Increased precision. The variance of the resulting mean would be lower, leading to more confidence over our estimate due to the division of our population into homogeneous groups.	In this sample, we do not observe much potential homogeneity within the groups. Especially considering the homogeneity within each stratum differs according to each variable. We would have to know the statistic of interest before choosing our sample. Stratified sampling can be inconvenient and/or expensive, given that members of the same strata could be spread out over a large physical area. This issue does not apply to this project, as this is only relevant during data collection, whereas we have been provided with all the data.
Cluster	Usually, cluster sampling increases efficiency, especially when the population of interest is spread over a large area physically, as it can reduce the costs and the time associated with travel and data collection. In this sense, it is often more convenient and, in the case of large-scale studies, more feasible than other forms of sampling as well.	However, since we are dealing with data that has already been collected, cluster sampling does not offer any benefits to us with regards to this project. Some of the other disadvantages of cluster sampling are as follows: • Reduced precision because of the variability within clusters. • Intra-cluster homogeneity: the sample may not be proportionally representative of the population which can lead to biased results. • If there is a significant variation between clusters, there is a risk of introducing bias into the sample.

Part d)

We are going to stratify our sample based on region, using Neyman allocation. We believe that geographical stratification factors will provide the most homogeneity of tuition fees within strata. There are too many states for us to be able to select a sample of only 20 if we were to stratify by state, so we have chosen to stratify by region. We have chosen to use Neyman allocation as this will account for the variability within regions.

First, we will find the number of institutions in each region:

```
region_tally <- df %>%
  count(region)
```

```
region_tally %>% knitr::kable()
```

region	n
1	121
2	251
3	215
4	148
5	356
6	107
7	39
8	135

Then, we will calculate the sample size for each stratum, using the formula:

$$n_h = n \times \frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h}$$

In order to do this, we need to calculate the standard deviation of tuitionfee_in within each stratum. We can do this using the formula:

$$s_h^2 = \frac{1}{(n_h - 1)} \sum_{j=1}^{n_h} (y_{hj} - \bar{y}_h)^2$$

Now, we will create a dataframe which summarises the data to tell us the sample size needed for each strata (region) using the above formula. We have also calculated the mean and standard error for each statum.

```
STS_df <- df %>%
  group_by(region) %>%
  summarise(
    n_institutions = n(),
    mean = mean(tuitionfee_in),
```

```
    sd = sd(tuitionfee_in)) %>%
  mutate(strata_size = 20 * (n_institutions * sd / sum(n_institutions * sd))) %>%
  mutate(strata_n = round(strata_size))
```

```
STS_df %>% knitr::kable()
```

region	n_institutions	mean	sd	strata_size	strata_n
1	121	33407.30	16255.45	2.0289550	2
2	251	30511.71	15981.07	4.1377807	4
3	215	26395.84	12999.82	2.8831258	3
4	148	23390.91	13068.98	1.9952228	2
5	356	20455.13	12498.59	4.5898561	5
6	107	18074.40	12908.86	1.4248188	1
7	39	15912.26	13503.47	0.5432478	1
8	135	26192.91	17212.54	2.3969928	2

So, we have found that, under Neyman allocation, we will need to select 2 units from region 1, 4 from region 2, 3 from region 3, 2 from region 4, 5 from region 5, 1 each from regions 6 and 7 and 2 from region 8. We will select these units using simple random sampling. The following code groups the population data in terms of region, then randomly selects the correct sample size from each region.

```
set.seed(234) # for reproducibility again

# joining the two dataframes based on the "region" column
merged_df <- df %>%
  inner_join(STS_df, by = "region")

# sampling strata_n units from each region
stratified_sample <- merged_df %>%
  group_by(region) %>%
  group_modify(~ sample_n(.x, size = first(.x$strata_n), replace = FALSE))
```

```
stratified_sample %>%
  select(region, Institution) %>%
  knitr::kable()
```

region	Institution
1	Plymouth State University
1	Saint Joseph's College of Maine
2	McDaniel College
2	Cedar Crest College
2	The King's College
2	Goucher College
3	Carthage College

region	Institution
3	Baldwin Wallace University
3	Saint Norbert College
4	University of Northwestern-St Paul
4	Concordia College at Moorhead
5	Norfolk State University
5	Sweet Briar College
5	Point University
5	Erskine College
5	Williams Baptist University
6	North American University
7	University of Colorado Denver/Anschutz Medical Campus
8	California State University-Monterey Bay
8	California State University-Fresno

As you can see in the table above, we have selected the correct number of intsitutions for each region.

Here are the selected institutions under stratified sampling with Neyman allocation:

```
print(stratified_sample$Institution)
```

```
 [1] "Plymouth State University"
 [2] "Saint Joseph's College of Maine"
 [3] "McDaniel College"
 [4] "Cedar Crest College"
 [5] "The King's College"
 [6] "Goucher College"
 [7] "Carthage College"
 [8] "Baldwin Wallace University"
 [9] "Saint Norbert College"
[10] "University of Northwestern-St Paul"
[11] "Concordia College at Moorhead"
[12] "Norfolk State University"
[13] "Sweet Briar College"
[14] "Point University"
[15] "Erskine College"
[16] "Williams Baptist University"
[17] "North American University"
[18] "University of Colorado Denver/Anschutz Medical Campus"
[19] "California State University-Monterey Bay"
[20] "California State University-Fresno"
```

Part e)

To find an estimate for the mean in-state tuition fees for all intitutions in the United States, we can take the mean of tuitionfee_in for our stratified sample, using the same formula that we used for simple

random sampling in part b)i).

```
mean_STS_tuitionfee_in <- mean(stratified_sample$tuitionfee_in)
print(mean_STS_tuitionfee_in)
```

```
[1] 26806.55
```

Standard error

We can compute the variance of our estimated mean tuitionfee_in using the formula:

$$var(\bar{y}_{str}) = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2\left(1 - \frac{n_h}{N_h}\right)\frac{s_h^2}{n_h}$$

```
var_STS_mean <- (1/(1372^2))*
  sum((STS_df$n_institutions^2)*
        (1-STS_df$strata_n/STS_df$n_institutions)*
        ((STS_df$sd^2)/STS_df$strata_n))
print(var_STS_mean)
```

```
[1] 10092386
```

We then square root this value to get the standard error:

```
STS_SE <- var_STS_mean^(1/2)
print(STS_SE)
```

```
[1] 3176.852
```

Comparing with b)i)

The standard error for the estimate of the mean tuitionfee_in for part b)i) was 3312.82. This is higher than the standard error of 3176.85 given by the stratified sample. This is likely due to the fact that we used Neyman allocation to obtain our stratified sample, which accounts for the variability within strata (regions) leading to increased precision.

Part f)

Estimated ratio

Using the sample selected in part a) using simple random sampling, we wish to estimate the ratio:

$$B = \frac{\bar{y}_U}{\bar{x}_U}$$

We can do this using the formula for the *ratio estimator*:

$$\widehat{B} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$$

The code below divides the sum of tuitionfee_out by the sum of tuitionfee_in for the sample we selected under SRS in part a) in order to find an estimate for the ratio of in-state to out-of-state tuition and fees in all colleges in the US. In this scenario, y = tuitionfee_in and x = tuitionfee_out.

```
in_out_est_ratio <- sum(SRS_df$tuitionfee_in) / sum(SRS_df$tuitionfee_out)
print(in_out_est_ratio)
```

```
[1] 0.9218609
```

Standard error

We can find the variance of the ratio estimator using the following formula:

$$Var(\widehat{B}) = \frac{(1-f)}{n\bar{x}_s^2} \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \widehat{B}x_i)^2$$

Where *f* is *n*/*N*.

```
var_ratio <- ((1-20/1372)/
                (20*(mean(SRS_df$tuitionfee_out))^2))*
  (1/(20-1))*sum((SRS_df$tuitionfee_in-in_out_est_ratio*SRS_df$tuitionfee_out)^2)
print(var_ratio)
```

```
[1] 0.001387993
```

We then square root this variance to find the standard error:

```
SE_ratio <- var_ratio^(1/2)
print(SE_ratio)
```

```
[1] 0.03725577
```

95% confidence interval

A 95% confidence interval for the ratio is given by:

$$\widehat{B} \pm 1.96 \times SE(\widehat{B})$$

```
upper <- in_out_est_ratio + 1.96*SE_ratio
print(upper)
```

```
[1] 0.9948822
```

```
lower <- in_out_est_ratio - 1.96*SE_ratio
print(lower)
```

```
[1] 0.8488396
```

So the 95% confidence interval for B (the ratio of in-state to out-of-state tuition and fees in all colleges in the US) is (0.8488, 0.9949).

Part g)

The first step in taking a systematic sample is to reorder the sampling frame in a way that makes sense. We want to select a representative sample based on average faculty salary. Therefore, we will reorder the datframe based on this variable, in descending order, so that our sample consists of salaries spanning as close to the full range as possible. The risk of using systematic sampling without reordering the sampling frame is the potential introduction of bias into the sample, particularly if there is an underlying pattern or structure in the data that aligns with the sampling interval.

```
df_reordered <- arrange(df, desc(avgfacsal))
```

We have noticed that there is a unit at the end of our data frame that has a value of -99 for avgfacsal. It is clear that data NA points have been coded in this way, so we will remove this unit from our sampling frame before we select our sample. This is done by the code below:

```
df_reordered <- df_reordered[df_reordered$avgfacsal != -99, ]
```

It is important to note that our population size has now gone down to 1371.

Next we have to compute our step. To obtain a sample size of 20 from a population of 1371, our step will be:

```
step <- 1371/20
print(step)
```

```
[1] 68.55
```

Since this is not an integer, we will round it up to the next integer:

```
k <- ceiling(step)
print(k)
```

```
[1] 69
```

Next, we randomly select a random number (R) between 1 and k.

```
set.seed(345) # for reproducability again
R <- sample(1:k, 1)
```

```
print(R)
```

```
[1] 21
```

Then, we select our units:

```
units <- seq(R, 1371, by = k)
print(units)
```

```
 [1]   21   90  159  228  297  366  435  504  573  642  711  780  849  918  987
[16] 1056 1125 1194 1263 1332
```

Here are the selected colleges under systematic sampling:

```
print(df_reordered$Institution[units])
```

```
 [1] "Dartmouth College"
 [2] "Bowdoin College"
 [3] "Drexel University"
 [4] "Clemson University"
 [5] "Embry-Riddle Aeronautical University-Daytona Beach"
 [6] "Fairleigh Dickinson University-Metropolitan Campus"
 [7] "Lincoln University"
 [8] "Texas Woman's University"
 [9] "Texas A & M University-Commerce"
[10] "University of Mary Hardin-Baylor"
[11] "Purdue University Fort Wayne"
[12] "Saint Ambrose University"
[13] "Trinity Washington University"
[14] "Oklahoma Baptist University"
[15] "University of Maine at Fort Kent"
[16] "Alfred University"
[17] "Lewis-Clark State College"
[18] "University of Valley Forge"
[19] "Montana State University-Northern"
[20] "Kentucky Christian University"
```

```
SYS_df <- df_reordered[units, ]
```

Part h)

To calculate an estimate for the average faculty salary per month across all colleges in the United States, we use the same formula we have used for SRS:

$$\widehat{\overline{y}_U} = \overline{y}_s = \frac{1}{n}\sum_{i=1}^{n} y_i$$

```
est_mean_salary <- mean(SYS_df$avgfacsal)
print(est_mean_salary)
```

```
[1] 8283.5
```

Standard error

We will use the same formula we used for SRS to calculate the standard error of this estimate:

$$var(\bar{y}_s) = \frac{s^2(1-f)}{n}$$

and:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{N}(y_i - \bar{y}_s)^2$$

```
sample_var_salary <- (1/(20-1))*sum((SYS_df$avgfacsal-mean(SYS_df$avgfacsal))^2)
print(sample_var_salary)
```

```
[1] 6258982
```

```
var_est_mean_salary <- sample_var_salary*(1-20/1372)/20
print(var_est_mean_salary)
```

```
[1] 308387.2
```

So the standard error of our estimate of the mean salary is:

```
SE_est_mean_salary <- var_est_mean_salary^(1/2)
print(SE_est_mean_salary)
```

```
[1] 555.3262
```

Part i)

To compare simple random sampling and stratified sampling, we can compare the standard errors of our estimates, since we were estimating the same variable in both of these cases (tuitionfee_in). The standard error for the estimate of the mean tuitionfee_in under simple random sampling was 3312.82, whereas the standard error for the estimate obtained via stratified sampling was 3176.85. This is an example of how stratified sampling can be an effective method to increase precision. In addition, we used Neyman allocation to obtain our stratified sample, which accounts for the variability within strata leading to further increased precision.

We can also compare the biases of these two estimates, given that we have access to the full population data:

```
SRS_bias <- mean(df$tuitionfee_in) - tuitionfee_in_mean
print(SRS_bias)
```

```
[1] -4500.636
```

```
STS_bias <- mean(df$tuitionfee_in) - mean_STS_tuitionfee_in
print(STS_bias)
```

```
[1] -1871.936
```

The estimate via stratified sampling gave a smaller bias than SRS. This shows that, in this case, stratified sampling is not only more precise than SRS, but also more accurate.

Since we estimated a different variable using systematic sampling (avgfacsal), we cannot compare the standard error obtained from this estimate with the others. One thing that we can do is compare the mean absolute percentage error (MAPE). The lower the MAPE score, the more accurate the estimate. This can be done using the formula:

$$MAPE = 100 \times \frac{\bar{y} - \hat{y}}{\bar{y}}$$

```
SRS_mape <- 100 * SRS_bias/mean(df$tuitionfee_in)
print(SRS_mape)
```

```
[1] -18.04975
```

```
STS_mape <- 100 * STS_bias/mean(df$tuitionfee_in)
print(STS_mape)
```

```
[1] -7.507377
```

```
SYS_mape <- 100 * (mean(df$avgfacsal) - est_mean_salary)/mean(df$avgfacsal)
print(SYS_mape)
```

```
[1] -1.282446
```

As expected, the stratified sample has a lower MAPE score (-7.5%) than the random sample (-18.0%). The systematic sample has an extremely low MAPE score of -1.3%. This is an indication that our sample was particularly representative, which could be due how we reordered our sampling frame.

Sampling method	Pros	Cons
Simple random sampling	SRS was the most convenient sampling method for us, as it required the fewest calculations.	SRS was the least precise sampling method, giving the highest variance and bias. This sampling method has the highest risk of not being representative of the population.

Sampling method	Pros	Cons
Systematic sampling	This sampling method provided even coverage of the population due to reordering the dataframe and selecting sample elements at equal intervals. This was also a fairly easy and convenient method.	Systematic sampling is more time consuming than SRS.
Stratified sampling	Stratified sampling can increase precision if the stratification factor relates to the variable of interest and there in homogeneity within strata. In our case, we observed an increase in precision when using this method as opposed to SRS,	This was the most complex sampling method. Neyman allocation provides the most precision within stratified sampling, however, it can only be used when the population variance of each stratum is known.

In this project, the most accurate estimate was given by systematic sampling. However, in general, we feel that stratified sampling provides more advantages than any other sampling method. We do not feel that any other sampling designs would have been more appropriate.

Final answers

Question	Sampling method	Estimate	Value	Standard error	Confidence interval
Part b)i)	SRS	Mean in-state tuition fees	29435	3313	NA
	SRS	Mean out-state tuition fees	31930	2572	NA
ii)	SRS	Proportion of Black/African American undergraduates	0.2039	0.0570	(0.0921, 0.3156)
Part e)	Stratified sampling	Mean in-state tuition fees	26807	3177	NA
Part f)	SRS	Ratio of in-state to out-of-state tuition and fees	0.9219	0.0373	(0.8488, 0.9949)
Part h)	Systematic sampling	Average faculty salary	8284	555	NA