QATAR UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

EXACT OPTIMAL SAMPLE ALLOCATION FOR ESTABLISHMENTS HAVING MORE THAN TEN EMPLOYEES IN QATAR.

BY
YASAMIN MOHAMMAD R SHAMSI

A Project Submitted to

the College of Arts and Sciences

in Partial Fulfillment of the Requirements for the Degree of

Masters of Science in Master of Science in Applied Statistics

January 2020

COMMITTEE PAGE

The members of the Committee approve the Project of

Yasamin Mohammad Shamsi defended on 26/09/2019.

_____

Mohammad Salehi
Thesis/Dissertation Supervisor

_____

Faiz Elfaki
Committee Member

Approved:

_____

Ibrahim AlKaabi, Dean, College of Arts and Sciences

ABSTRACT

SHAMSI YASAMIN MOHAMMAD, Masters : January : [2020], Applied Statistics

Title: Exact Optimal Sample Allocation for Establishments having More Than Ten Employees in Qatar

Supervisor of Project: Mohammad, Salehi.

The primary aim of this study is to select a stratified random sampling with three different techniques which is optimal allocation, proportional allocation and the new method is exact optimal allocation which is a great method with more advantage than other methods. In addition, we use two different stratification method which is: optimal rule of stratification and stratification based on establishment activity. In each two stratification we apply the three methods, and the total sample size is chosen by optimal allocation technique and set for both other techniques, so that we can make a comparison among there variance.

The result we get is as following stratification by optimal rule of stratification give smaller variance in all three technique than stratification based on establishment activity. Among the three method we use, the powerful method is exact sampling allocation because it gives smaller variance than the other methods, followed by optimal allocation (Neyman) and the largest variance we get from proportional allocation in both method of stratification. In exact optimal allocation with predetermining variance, we get smaller number of sample size for stratification by establishment activity but with very large variance, otherwise with optimal rule of stratification with predetermining variance we get a large number of sample size with smaller variance.

Key words: Stratified random sample, Optimal allocation, Proportional allocation, Exact optimal allocation.

DEDICATION

This project is dedicated to

My family and to those who believe in me.

Thank you.

ACKNOWLEDGMENTS

I would like to thank Professor Mohammed Salehi; all stages of this work were performed under his supervision and guidance. A special thanks to our committee members Professor Ayman Bakleezi and Dr. Faiz Elfaki for their recommendations and comments, which raised my awareness and helped me improve my research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

This chapter provides the study background by introducing the stratified sampling method, statement of the problem, significance of the study, the general and specific objectives, and the definition of key terms.

1.1 GENERAL BACKGROUND AND KEY TERMS

Sampling method was used first time by a Persian king to estimate the number of his troops during an overrun of Greece (Thompson,2012, p.7). In the first two decades of the twentieth century the basic of simple random sampling were worked out and most standard random sampling like: stratified sampling, systematic sampling, cluster sampling and other have been introduced by the end of 1930s. In 1940, probability sampling was familiarized by the U.S. census, where unequal probability sampling was introduced in the end of 1940s and 1950s (Thompson,2012, p.7).

Sampling design plays a major role in majority studies, applying a correct method for choosing a sample will give representative sample for the population. In addition, the sample size and cost all depend on the sample technique that will be used. Of course, most of the researchers look for the best sampling technique in the sense of sample size, cost and the precision of estimators. Sample is a subset that is selected for research study and the purpose of sampling is to find representative samples.

One of the most popular sampling methods is stratified sampling which have one essential feature which is dividing the population into groups called strata each one of these strata have one especial character that differ from the other groups. The main advantage of stratified sampling over simple random sample is that stratified sampling

may give more precise estimate of the population parameters, for example the mean estimator of stratified sampling gives smaller standard error than simple random sample of the same size in almost all cases. The second advantage is the analysis will be related to the category. For the strata which will be more representative for the population (Butcher, 1966, p.8). We can as well control the cost of sampled population (Pfeffermann, 2009, p111).

For selecting sample in stratified sampling, the population should be divided into homogenous group, then random sample will be selected from each stratum depending on the population size in each stratum. The number of samples to be selected from each strata will depend in some method or some other sampling techniques like simple random sampling without replacement or with replacement or using some method like proportional allocation which will consider the size of strata to select a sample from it (Sampath, 2001, p. 79) , or the other method is optimal allocation (Neyman) which is aim to minimize the overall variance for the least cost and vice versa(Cochran, 1977, p.89).

The advantage of stratified sampling is studied by Padilla et al (2017), they try to find the best sampling method for some geographic events that is rarely occurs such as Biomass burning. The objective of the study is to improve the sampling design by: allowing researchers to define low and high burned area strata that differ by biome which is classification an area of plant according to animal and plant that live in it, and by year. Then they identify the best method for allocating the sample to strata. The stratification is constructed in three levels. The first stratification level consisted of assigning each sampling unit to a calendar year. The second stratification level consisted of assigning each sampling unit to the major biome for which the TSA (Thiessen Scene Areas) had the

maximum area, the third stratification level was based on thresholds of burned area. Sample unit is specifying as 100 unit per year. The allocation is determined by two type of equations: (1) $n_h \, \alpha \, N_h \sqrt{\overline{BA_h}}$, (2) $n_h \, \alpha \, N_h \overline{BA_h}$. where BA is the dates and location of burned surface and it cover multiple years periods, N_h is number of populations in stratum h and the sample to be selected denoted by n_h . Standard errors under simple random sampling were 2.5 to 8 times greater than those of the stratified sampling options indicating that the stratified designs offer high potential to reduce standard errors.

The stratified design with equation (1) yielded smaller standard errors than the stratified option with equation, (2) for Precision comparisons stratification for year by biome give the best accuracy estimates. The standard errors obtained from the stratification and allocation based on equation (1) were almost uniformly higher than the corresponding standard errors obtained based on equation (2).

In other field Pirzadeh et al (2013) studied and analyzed the traces which is specialized in records event that occur in an operating system, to record information about a program's execution which help to understand system behavior and for maintenance tasks. So that, researcher presents an approach for reducing the size of traces that is based on the sampling of the trace content. They used stratified sampling of execution traces. They divided the traces into group of similar events that together perform an essential step of the general execution. The number of events will be selected from each execution phase to yield a sample that is representative of the original trace. They applied two case studies one based on stratified sampling of execution traces, with the purpose of comparing its usefulness with random sampling. It is found that stratified approach provided better results in more than 80% of the cases. In all these cases, it led to

a more representative sampled trace because they use execution phases as strata. Furthermore, as the size of the sample decreases or increase, stratified sample keeps its representativeness because it is divided by exaction phases as strata while random sampling is not.

In this project, we will use stratified sampling method with optimal allocation (Neyman) and proportional to size allocation. The new path we add on stratified sampling is exact optimal allocation which we will use in our project (Wright 2017). Exact optimal allocation is a very powerful method that has advantages over other method like optimal allocation and proportional allocation. With exact optimal allocation we get smaller variance estimator. Furthermore, it gives the minimum number of sample size to represent the population with better variance. The advantage of exact optimal allocation will be shown when we made a comparison between it with other stratification method. This great method will be applied on the establishments having more than ten employees in Qatar. The data was taking from planning and statistics authority of Qatar from annual bulletin of industry and energy statistics 2017. Since the data have frame including all required establishment details the sample was selected using simple random sampling method the objective of this data is to estimate and study the GDP (Gross domestic profit) for Qatar, so by finding better sampling technique with smaller variance and more representative of population like exact optimal allocation that will give better estimation for the GDP.

Moreover, we will investigate the advantage of exact optimal allocation. Also, we will use a simulation study as well as their theoretical variance to study their efficiencies.

key terms:

Sampling design

If ω is the population of all subset of S and p(s) is the probability distribution defined on ω, then the probability distribution { p(s) , s ϵ ω) is sampling design (Sampath, 2001, p2).

Stratified sample

The procedure of partition the population into groups called strata, then selecting a sample independently from each stratum. (Singh et al, 1996, p.102)

Stratified Random sampling

If the sample selected from each stratum is random one, then it is stratified random sample. (Singh et al, 1996, p.102)

Proportional allocation

The number of sampled units in each stratum is proportional to strata size. (Lohr, 2009, p105)

Optimal allocation (Neyman 1934)

A method used to allocate sample to strata based on the strata variances and similar sampling costs in the strata.

1.2 PROBLEM STATEMENT

In most research the researcher cares about the sample size and how much it will cost and how much variation they will get from this sample. Stratified random sampling have two popular methods that can be used to allocate a sample to strata. One of this method is optimal allocation (Neyman) which give smaller variance than the proportional

allocation. But there is one problem may occur which is getting non-integer number of sample size. The second method is proportional allocation which gives larger variance estimator. So that we will use new powerful methods which is Exact Optimal allocation in stratified sampling and compare it with optimal allocation and proportional allocation throw their variance estimator and the sample size.

1.3 SIGNIFICANCE OF THE STUDY

The results of our great method: exact optimal allocation will make a different in the variance of stratified sampling, the variance will be smaller than the other two methods. It will give sample size with smallest variance, which will help in real life data to reduce the cost and get more representative sample of the population.

1.4 OBJECTIVE OF THE STUDY

The primary objective of this study is we learn and add an extra knowledge with helpful technique which is exact optimal allocation, and we implement this method on a local data. In addition, we will compare it with proportion allocation and ordinary optimal allocation to highlight the great difference between exact optimal allocation and optimal allocation and proportion allocation.

1.5 SPECEFIC OBJECTIVES

To achieve the main objective, we have the following specific objectives:

1. Use a real local data for the study which is obtained from Planning and Statistics Authority of Qatar, which is about Industry and Energy Statistics in 2017.

2. Stratified data by two way of stratification which is: optimal rule of stratification and stratification by establishment activity.
3. We apply the three methods of sampling allocation, optimal allocation, proportional allocation and new method exact optimal allocation.
4. Compare the stratification methods which give smaller variance among the three methods.

CHAPTER 2: STRATIFIED RANDOM SAMPLING WITH THREE DIFFERENT METHODS

2. STRATIFIED RANDOM SAMPLING

Stratified random sampling mostly used in large scale survey. The stratification helps allocate the sample in different strata and this can be done in a way to be more representative for the population (Arnab, 2017, p214). In stratified sampling the population (N) is divided into subpopulation ($N_1, N_2, N_3, \ldots, N_k$), these subpopulations are non-overlapping and contain the whole population (Lohr,2009, p.95). The subpopulation is called strata and each stratum is represented in the sample with probability equal 1(Chaudhuri, 2005, p.229). For obtaining any estimator of the population parameter such as the population mean, proportion and total can be estimated by computed the weighted average for each stratum (Lemeshow. et al, 2013). For creating the strata there is two condition should be fulfilled, first the population size in the strata should be known to select the sample proportionally to each stratum size. Second it should be possible to draw sample from each stratum separately. (Kalton , 1983, p.26). To achieve the maximum level of precision per unit in each stratum, the strata should be divided among the correct category (Deming, 1966, p213). The sample will be selected from each stratum and the sample size within each stratum is denoted by ($n_1, n_2, n_3, \ldots, n_k$). Selecting the sample from the strata can be done randomly with or without replacement (Cochran, 1977, p.89) or by any method like optimal allocation and proportional allocation which will be discuss latter. Often the method of selecting sample is one in all strata and the estimator that will be measure is same in all strata. (Sarndal et al, 1992, p.101)

Stratified sampling has many advantages over the other sampling method such as: (Singh et al, 1966 , p.103)

- Stratified random sample provide better precision estimator than simple random sample.
- stratified random sample require smaller cost.
- Stratified random sample ensure obtaining sufficient sample points to support a separate analysis of any subgroup.

2.1 OPTIMAL ALLOCATION (NEYMAN)

Neyman[1] was one of the principal architects of modern theoretical statistics. He proposed and studied randomized experiments in 1923. Furthermore, in 1934 his paper *"On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection*" was the pioneering research leading to modern scientific sampling (Neyman,1934).

Using Neyman allocation when there is problem of how to choose a sample at a time when there was controversy about purposive selection versus random sample. The idea of Neyman allocation is separation the population frame into groups which is called strata that differ from each other, then piking some unit from each stratum to represent the population. The limitation of Neyman allocation is requirement of knowing population variance within each stratum (Thompson,2012,p109). The main reason of using optimal allocation is to minimize variance of specific cost for the selected sample or to minimize cost for specific value of variance (MUKHOPADHYAY et al,2008,p.55) and it can be minimize by using Cauchy-Schwarz inequality , and within any stratum the cost will be dependent on the stratum size (Cochran, 1977,p97).

The formula for calculating the sample size in each stratum for Neyman

[1] *Statistics is the servant to all sciences.", Jerzy Neyman*

Allocation is as following:

$$n_h = n \cdot \frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h} \qquad \text{(If cost is unknown)} \qquad (1.1)$$

$$n_{h=} n * \left(\frac{{}^{N_h S_h}/_{\sqrt{c_h}}}{\sum_{h=1}^{H} {}^{N_h S_h}/_{\sqrt{c_h}}} \right) \qquad \text{(If cost is not known)} \qquad (1.2)$$

where

$$n = \frac{\left(\sum_{h=1}^{L} {}^{N_h S_h}/_{\sqrt{c_H}} \right)^2}{N^2 D + \sum_{i=1}^{L} N_i s_i^2} \qquad (2)$$

$$S_h = \sqrt{\frac{1}{N_h - 1} \sum_{i=1}^{N_h} (y_i - \bar{y}_i)^2} \qquad (3)$$

n_h –sample size in stratum h;

n – total sample size;

N_h – population size in stratum h;

y_i – variable of interest in stratum h.

S_h – variance in stratum h.

C_h – cost in stratum h.

And D is the bound of error and it is defined as

$$\mathrm{D} = \frac{\mathrm{B}^2}{4} \qquad (\text{when etimating mean } (\mu)) \qquad (4.1)$$

$$\mathrm{D} = \frac{\mathrm{B}^2}{4\mathrm{N}^2} \qquad (\text{when estimating total}) \qquad (4.2)$$

The scheme of Neyman allocation provides the most precision for estimating a population mean given a fixed total sample and vice-versa.

2.2 PROPORTIONAL ALLOCATION

Proportional allocation was first proposed by Bowely (1926) and it is method used in stratified sampling to assign sample size to strata, so the sample will be

proportion to stratum size, often proportion allocation used when stratum variance and cost are equal (Singh et al, 1966 , p.111). The best situation to use proportional analysis is when we want to focus on some stratum that have more unit than the other (Cochron, 1977,p109). Proportional allocation is identical to selecting the sample in each stratum with the same probability of selection (Groves, et al ,2004 ,p111).

The formula for calculating the sample size in each stratum for proportion allocation is as following:

$$n_h = n \cdot \left(\frac{N_h}{\sum_{h=1}^{H} N_h}\right) \qquad (5)$$

and:

n_h –sample size in stratum h;

n – total sample size;

N_h – population size in stratum h;

And the formula for n is:

$$n = \frac{N_h s_h^2}{ND + \frac{1}{N}\sum_{h=1}^{H} N_h s_h^2} \qquad (6)$$

2.3 EXACT OPTIMAL ALLOCATION

Exact optimal allocation is our main goal and method in this project, this method has been introduced recently (Wright, 2017) and it will be applied in a real-life data to investigate it advantages among other method. In this powerful method, we have two cases: first if the variance is not given and second case if the variance is fixed and we want to select minimum number of sample size.

2.3.1 EXACT OPTIMAL ALLOCATION WITH FIXED SAMPLE SIZE

In exact optimal allocation with fixed sample size we obtain an exact allocation for minimum variance and the number of the sample in each stratum is positive integer.

To apply exact optimal allocation method, we should have ($N_h S_h$) which is populations size of each stratum multiply by the standard deviation of the strata. We then need to find an upper limit and lower limit for each stratum by choosing adequate limit for each stratum. Next, we assign one value and we compute $\frac{N_h S_h}{\sqrt{1*2}}$ for each stratum and we chose the largest priority value. To interoperate the priority value let $\bar{y}_{mh}(\bar{y}_{mh}-1)$ be sample mean based on simple random sample of $m_h(m_h-1)$ unit from the hth strata. When the sample size in the *h*th strata increase from (m_h-1) to m_h the variance $\hat{T}_h$ decrease as well as the variance of $\hat{T}_y$, the amount of decrease will be by

$$var\ (N_h\ \bar{y}_{mh}-1)-var\ (N_h\ \bar{y}_{mh})=\ \frac{N_h^2 S_h^2}{(m_h-1)(m_h)}=\left(\frac{N_h S_h}{\sqrt{(m_h-1)(m_h)}}\right)^2$$

based on that each selection of the largest priority value will decrease the variance of $\hat{T}_y$ by an associated squared priority value from a stratum. Let then we assign two value for each stratum $\frac{N_h S_h}{\sqrt{1*2}}$ and $\frac{N_h S_h}{\sqrt{2*3}}$ and so on. The last value in the array will be the maximum upper limit plus one (max b_h+1)

Table 1. Array and largest priority values.

$\frac{\boldsymbol{N_1 S_1}}{\sqrt{\boldsymbol{1*2}}}$	$\frac{\boldsymbol{N_1 S_1}}{\sqrt{\boldsymbol{2*3}}}$	$\frac{\boldsymbol{N_1 S_1}}{\sqrt{\boldsymbol{3*4}}}$	**....**	$\frac{\boldsymbol{N_1 S_1}}{\sqrt{\boldsymbol{b_{h+1}*b_{h+2}}}}$
$\frac{\boldsymbol{N_h S_h}}{\sqrt{\boldsymbol{1*2}}}$	$\frac{N_h S_h}{\sqrt{2*3}}$	$\frac{N_h S_h}{\sqrt{3*4}}$	....	$\frac{N_h S_h}{\sqrt{\boldsymbol{b_{h+1}*b_{h+2}}}}$
$\frac{\boldsymbol{N_H S_H}}{\sqrt{\boldsymbol{1*2}}}$	$\frac{N_H S_H}{\sqrt{2*3}}$	$\frac{N_H S_H}{\sqrt{3*4}}$	....	$\frac{N_H S_H}{\sqrt{\boldsymbol{b_{h+1}*b_{h+2}}}}$

Next, we will remove all value that is smaller than or equal to $(a_h - 1)$ and all the value that is greater than $(b_h - 1)$ from each stratum depending on there limit, where a_h is the lower limit and b_h is the upper limit.

lower limit: $(a_h - 1)$ (7.1)

Upper limit: $(b_h - 1)$ (7.2)

From the remaining value we chose the largest priority values which can be calculated in equation (8) and at last we chose the numbers that satisfy sum of n from each stratum with the minimum variance.

$n - \sum_{h=1}^{H} a_h$ (8)

2.3.2 EXACT OPTIMAL ALLOCATION WITH FIXED VARIANCE:

To apply the second case of exact optimal allocation first it needs to meet a specified sampling variance which is $V_0 = var\left(\hat{T}_Y\right)$, and $V_0 < \sum_{h=1}^{H} N_h (N_h - 1) S_h^2$ which is the sampling variance when $n_h = 1$ for all h. After that $(N_h S_h)$ need to be found for the array then we assign one value for each stratum and find $\frac{N_h S_h}{\sqrt{1*2}}$ and we find the largest priority value and we increase the sample size by one value for the stratum that have largest priority value with keeping the other strata fixed, then we assign two value for each stratum $\frac{N_h S_h}{\sqrt{1*2}}$ and $\frac{N_h S_h}{\sqrt{2*3}}$ and so on. We keep increasing the sample size until we reach the first value of the variance that is smaller than V_0 (Ex: $V_x < V_0$) and we take one value before (Ex: V_{x-1}) which is greater than V_0 . If we have some strata with very large variance, we keep the sample size of it equal the population size ($n_h = N_h$).

To calculate the specified desired precision V_0 :

$V_0 = (\, coefficient\ of\ variation * predicted\ Total\ of\ Y\,)^2$ (9)

to calculate the variance for the first value we use the following formula:

$$V_1 = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\sigma_2^2}{n_h} \quad (10)$$

and to calculate the variance for the others value we use (11) and repeated until we reach variance smaller than V_0, we will calculate the sample size that have variance larger than V_0 only and that will be our sample size.

$$var_j(\hat{T}_y) = var_{j-1}(\hat{T}_y) - \left(N_h S_h \Big/ \sqrt{x * y}\right)^2 \quad (11)$$

CHAPTER 3: REAL DATA ANALYSIS

The data used in this project is taken from Planning and Statistics Authority of Qatar about Industry and Energy Statistics in 2017. The data considers 21 groups of different establishments activity having more than ten employees, and it includes two variables: number of employee and their salary in those establishments. The total number of establishments are 1452 (Appendix A). Depending on Planning and statistic authority they collect the sample using simple random sample with replacement and this data are using to get an estimation of the salary for the GDP (Gross domestic profit) for Qatar. So that when we get more representative sample from the population with less variation, we will achieve a better estimate for the GDP with more precision. Under some restrictions we couldn't get any information for each establishment independently on other words we couldn't get the frame with full or part details, because the name of the establishment, number of employee and their salary is confidential and cannot be release, so, we generate the population based on the available sample, so for each establishment the number of employee and their salary will generated by sensible method, we use R programme for generating the population. (Appendix B). The allocation is based on knowing the population variance and depend on census, and since the frame is updated each five years, the variance of population we will have in our project is applicable for four years and it should be updated in year five with the census. We also study the exact allocation method based on a pilot survey.

3.1 STRATIFICATION BY OPTIMAL RULE OF STRATIFICATION

We use two methods for stratification, first method is optimal rule for stratification, which is a mathematic method to create the strata, we stratified by number of employee in each stratum, so we divide the population into six strata based on number of employees in each establishment, then we find the frequency for each strata, next step is taking the square root of the frequency then the cumulative number -which is the added all the total together- will be divide by the square root of frequency as shown in table (2) note that: ($N_1 = \sqrt{229} = 15.13$ then $N_2 = 15.13 + \sqrt{146} = 27.22, \dots$)

Table 2. The step for calculating the strata

Strata	*Frequency*	$\frac{cumulative}{\sqrt{frequency}}$
N_1	229	15.13
N_2	146	27.22
N_3	171	40.29
N_4	518	63.05
N_5	95	72.80
N_6	293	89.92

We decide to have 4 strata because dividing in to 6 strata does not make any change especially in N_6 so we will keep the data with four strata as we get from the original source. So, we will divide the last number in the last column by 4 ($\frac{89.92}{4} =$ 22) to get the maximum number of employee in the first strata, then we multiply the result by 2 for the second strata, and that result we multiply by 3 for the third strata and last we multiply the last result by 4 for the fourth strata (to get the result as it shown in table (3):

Table 3. The limit for each stratum

Strata	Limit
$l_1 = {}^{89.92}/_{4}$	22
$l_2 = 22*2$	45
$l_3 = 22*3$	67
$l_4 = 22*4$	90

Now we can come out with the final 4 strata in table (4)

Table 4. the final four strata

Strata	Number of Employees	Frequency
L1	From 1 to 40	375
L2	From 41 to 60	171
L3	From 61 to 100	613
L4	101 and above	293

So, we have $N_1 = 375$, $N_2 = 171$, $N_3 = 613$ and $N_4 = 293$ and based on the strata we will compute the variance for the salary which will be $\sigma_1^2 = 623622.33$, $\sigma_2^2 = 585525.26$, $\sigma_3^2 = 521115.74$ and $\sigma_4^2 =$1195145717.41 based on equation (3).

After we have found our strata, we will apply three sampling methods in each type of stratification starting with Optimal sampling allocation (Neyman)

3.1.1 OPTIMAL ALLOCATION METHOD (NEYMAN)

First method will be applied is optimal allocation method, so that the stratum is divided according to number of employee and the estimator will be the salary. To find the total number of sample size (n) we need to computed it from equation (2) , we will consider the cost are equal one in all stratum and we have the following information $\sum_{i=1}^{4} N_i \sigma_i = 10998756.28$ and $\sum_{i=1}^{4} N_i \sigma_i^2 = 350831105242.46$

based on equation (2) we will have n = $\frac{(10998756.28)*(10998756.28)}{1452^2+350831105242.46} = 345$

We will find number of sample size for each stratum based on equation (1.1)

$$n_1 = 345 * \frac{296136.61}{10998756.28} = 9.28 \approx 9$$

$$n_2 = 345 * \frac{130837.38}{10998756.28} = 4.10 \approx 4$$

$$n_3 = 345 * \frac{442514.56}{10998756.28} = 13.87 \approx 14$$

$$n_4 = 345 * \frac{10129267.73}{10998756.28} = 317.56 \approx 318$$

We get $n_4 = 318$ and the total population in the forth strata = 293 and the result for the sample cannot exceed the population size, so in this case we will take the whole population, which is mean $n_4 = N_4 = 293$, we will now calculate the other sample again.

$$n_1 = 52 * \frac{296136.61}{869488.55} = 17.71 \approx 18$$

$$n_2 = 52 * \frac{130837.38}{869488.55} = 7.82 \approx 8$$

$$n_3 = 52 * \frac{442514.56}{869488.55} = 26.46 \approx 26$$

$$n_4 = 293$$

with total variance = (7043637007.81 + 3261088212.99 + 10556424510.74 + 1452451.69) = 13,885,250,367.15. Note that since the cost is unknown, we will consider it equal one in all strata as well as the bound of error.

3.1.2 PROPORTIONAL ALLOCATION METHOD

Here we will use fixed number of sample size in all following method with n=345 and we will compute the sample size in each stratum by equation (5) and we will have the following result:

$$n_1 = 345 * \frac{375}{1452} = 89.05 \approx 89$$

$$n_2 = 345 * \frac{171}{1452} = 40.61 \approx 41$$

$$n_3 = 345 * \frac{613}{1452} = 145.57 \approx 146$$

$$n_4 = 345 * \frac{293}{1452} = 69.58 \approx 70$$

with total variance = (750907722.31 + 321440953.25 + 1025718773.95 + 1124403328018.86) = 1126501395468.37

3.1.3 EXACT OPTIMAL ALLOCATION

3.1.3.1 EXACT OPTIMAL ALLOCATION WITH FIXED SAMPLE SIZE:

To apply exact optimal allocation, we will set the interval limit for each stratum as following:

$15 \leq n_1 \leq 20$

$5 \leq n_2 \leq 10$

$25 \leq n_3 \leq 30$

$289 \leq n_4 \leq 294$

Next step we need to find $\frac{N_h \sigma_h}{\sqrt{1*2}}$, and the array will end at 294+1= 295 which is $max\ b_h + 1$.

Table 5. Finding the priority value

$N_h\sigma_h$	$\frac{1}{\sqrt{1*2}}$	$\frac{1}{\sqrt{2*3}}$	$\frac{1}{\sqrt{3*4}}$	…	$\frac{1}{\sqrt{295*296}}$
296136.61	209400.20	120897.26	85487.28	…	1002.16
130837.38	92516.00	53414.14	37769.50	…	442.77
442514.56	312905.05	180655.81	127742.95	…	1497.51
10129267.73	7162473.90	4135256.23	2924067.72	…	34278.45

Now, we will delete the value that is smaller than or equal to the smaller limit minus $(a_h - 1)$ one and greater than upper limit minus one $(b_h - 1)$ by compute in (7.1 and 7.2) from each stratum.

Table 6. The upper and lower limit for each stratum

Strata	Lower limit	Upper limit
First strata	15– 1 = 14	20 – 1 = 19
Second strata	5 – 1 = 4	10 – 1= 9
Third strata	25 – 1 = 24	30 – 1 = 29
Fourth strata	289 – 1 = 288	294 – 1 = 293

From table (6) we can see in first stratum $(a_h - 1)$ = 14, so all the value that is smaller than or equal to 14 will be deleted from stratum 1 only and for the upper limit $(b_h - 1) = 19$ that mean all the value that is grater than 19 will be deleted from stratum 1 only, and same steps will be repeated for all strata we have.

After deleting all the value in all strata, we will be left with at least a_h unit and not more than b_h unit in each stratum:

Table 7. The sample size and priority value for each stratum

Stratum (1)			*Stratum (2)*		
Sample Size	$N_1\sigma_1 = 296136.60$	**Priority value**	Sample Size	$N_2\sigma_2 = 130837.37$	**Priority value**
15	1/√(15*16)	**19115.54**	5	1/√(5*6)	**23887.53**
16	1/√(16*17)	**17955.92**	6	1/√(6*7)	**20188.65**
17	1/√(17*18)	16929.00	7	1/√(7*8)	**17483.88**
18	1/√(18*19)	16013.23	8	1/√(8*9)	15419.33
19	1/√(19*20)	15191.49	9	1/√(9*10	13791.47
Stratum (3)			Stratum (4)		
Sample Size	$N_3\sigma_3$ = 4929	**Priority value**	Sample Size	$N_4\sigma_4$ = 30768	**Priority value**
25	1/√(25*26)	**17356.85**	289	1/√(289*290)	**34988.89**
26	1/√(26*27)	16701.64	290	1/√(290*291)	**34868.44**
27	1/√(27*28)	16094.10	291	1/√(291*292)	**34748.83**
28	1/√(28*29)	15529.22	292	1/√(292*293)	**34630.03**
29	1/√(29*30)	15002.65	293	1/√(293*294)	**34512.04**

We now need to find the largest priority number by subsisting in equation (8), 345 – (15+5+25+289) = 345 – 334= 11 which can be found in bold in table (7). Last step is finding all the sum that give the total number of sample size (n=345) with minimum variance by taking sample from each stratum and find their variance which in our case is $n_1 = 18, n_2 = 7, n_3 = 27,\ n_4 = 293$. and the variance will be 13,661,852,139.44.

3.1.3.2 EXACT OPTIMAL ALLOCATION WITH FIXED VARIANCE

For calculating V_0 we will use the predicted total salary which is equal 8,010,540,000 and the coefficient of variance equal 0.006% so, substitute in equation (9) we will get $V_0 = (0.00006\% * 8010540000)^2 = 231007503929.76$ and we will start our sampling with 1 unit with keeping the fourth strata fixed since it has

very large variance.

Table 8. Number of population and variance in each stratum

Strata	N	σ^2
1	375	623622.33
2	171	585425.26
3	613	521115.74
4	293	1195145717.41

To calculate the first variance, we will use equation (10)

Table 9. The step of calculation the first variance

Strata	Formula	Result
First strata	$375^2 * \left(1 - \frac{1}{375}\right) * \frac{623622.33}{1}$	8746302374.00
Second strata	$171^2 * \left(1 - \frac{1}{171}\right) * \frac{585425.26}{1}$	17018312192.00
Third strata	$613^2 * \left(1 - \frac{3}{613}\right) * \frac{521115.74}{3}$	195499695054.82
Fourth strata	$239^2 \left(1 - \frac{293}{293}\right) \frac{1195145717.41}{293}$	0
$V_1 =$	299,981,039,620.82	

Since V_1 is larger than V_0 we will increase the sample size by one for the largest priority value until we get variance less than V_0

We will construct our array to find the largest priority value, to increase our sample size. The array will not include stratum four because it has a very large variance and we set the sample size for it fixed $n_4 = 293$.

Table 10. Construct array to find the largest priority values

Stratum $N_h S_h$	$\frac{1}{\sqrt{1*2}}$	$\frac{1}{\sqrt{2*3}}$	$\frac{1}{\sqrt{3*4}}$	$\frac{1}{\sqrt{4*5}}$	$\frac{1}{\sqrt{5*6}}$	$\frac{1}{\sqrt{6*7}}$
Stratum 1 296136.61	209,400.20	120,897.26	34,900.03	66,218.16	12,089.73	45,694.87
Stratum 2 130837.38	92,516.00	53,414.14	15,419.33	29,256.13	5,341.41	20,188.65
Stratum 3 442565.21	312,940.86	180,676.49	52,156.81	98,960.59	18,067.65	68,289.29

In table (10) it shown that when we assign one value to all stratum the third stratum have the largest priority value, so we will increase the sample size in third stratum by one and keep stratum one and two fixed and we will calculate the variance for step one by equation (11) as it shown in table (11) then we will assign another value for all strata in the array and the largest priority value which is in stratum three so, we will increase the sample size in stratum three only by one and calculate the variance and so on until we reach the variance that is smaller than V_0.

Table 11. The calculation of sample size and variance

Step	n_1	n_2	n_3	n_4	Variance
1	1	1	1	293	299,981,039,620.82
2	1	1	2	293	267,344,516,453.99
3	1	1	3	293	251,026,254,870.57
4	1	1	4	293	241,235,297,920.52
5	1	1	5	293	234,707,993,287.16
6	1	1	6	293	230,045,632,834.75

The variance is become smaller than V_0 in the sixth step (V_6), so, we will stop in the fifth step (V_5) and we will get the following result, the total sample size (n) equal 300 with $n_1 = 1$, $n_2 = 1$, $n_3 = 5$, $n_4 = 293$, and the variance equal 192,297,032,941.57

3.2 STRATIFICATION BY ESTABLISHMENT ACTIVITY

The second method for stratification is stratification by establishment activity which already divided by the planning and statistics authority of Qatar**,** we put the establishments in 4 strata as following:

strata 1: non-metallic manufactures.

strata 2: metallic manufactures.

strata 3: food and beverage product.

strata 4: pharmaceutical and chemical.

and we get $N_1 = 864$, $N_2 = 121$, $N_3 = 36$ and $N_4 = 431$ with $\sigma_1^2 =$ 8,475,364.29 , $\sigma_2^2 = 9{,}270{,}080.55$, $\sigma_3^2 = 7{,}781{,}931{,}997.931$ and σ_4^2 =377,272,113.388 respectively.

We will apply the 3 method for stratification to compare which type of stratification is more accurate for the salary.

3.2.1 OPTIMAL ALLOCATION METHOD (NEYMAN)

First we will find total number of sample size (n) by substitute in equation (2) , and we will consider the cost equal one since we don't know it, and we have the following information $\sum_{i=1}^{4} N_i \, \sigma_i = 14430999.73$ and $\sum_{i=1}^{4} N_i \, \sigma_i^2 = 11872930755553.90$

so we will have n = $\frac{(14430999.73\ )*(14430999.73\ )}{1452^2+11872930755553.90} = 18$

know we will find number of sample size for each stratum

$$n_1 = 18 * \frac{2515318.18}{14430999.73} = 3.06 \approx 3$$

$$n_2 = 18 * \frac{368406.36}{14430999.73} = 0.45 \approx 1$$

$$n_3 = 18 * \frac{3175749.34}{14430999.73} = 3.86 \approx 4$$

$$n_4 = 18 * \frac{8371525.85}{14430999.73} = 10.18 \approx 10$$

with total variance = (21,016,191,133,041.82+ 134,601,569,653.00 + 1,736,927,221,938.19 + 7,624,334,058,579.72) = 11,597,481,983,212.70

3.2.2 PROPORTIONAL ALLOCATION METHOD

Here we will use fixed number of sample size in all following method which is n=18. So, we will calculate number of sample size in each stratum by substitute in equation (5)

$$n_1 = 18 * \frac{864}{1452} = 10.44 \approx 10$$

$$n_2 = 18 * \frac{121}{1452} = 1.46 \approx 2$$

$$n_3 = 18 * \frac{36}{1452} = 0.43 \approx 1$$

$$n_4 = 18 * \frac{431}{1452} = 5.21 \approx 5$$

with total variance = (598860501948.12 +66739944952.95 + 9805234317393.00+ 13297975490644) = 23,768,810,254,938.10

3.2.3 EXACT OPTIMAL ALLOCATION

3.2.3.1 EXACT OPTIMAL ALLOCATION WITH FIXED SAMPLE SIZE:

To apply exact optimal allocation, we will set the interval limit for each stratum as following:

$1 \leq \boldsymbol{n_1} \leq 5$

$1 \leq \boldsymbol{n_2} \leq 5$

$2 \leq \boldsymbol{n_3} \leq 6$

$8 \leq \boldsymbol{n_4} \leq 12$

Next step we need to find $\frac{N_h \sigma_h}{\sqrt{1*2}}$, and the limit of our array is 12+1=13 according to $max\ b_h + 1$

Table 12. Finding the priority value

$\boldsymbol{N_h \sigma_h}$	$\frac{\mathbf{1}}{\sqrt{\mathbf{1*2}}}$	$\frac{\mathbf{1}}{\sqrt{\mathbf{2*3}}}$	$\frac{\mathbf{1}}{\sqrt{\mathbf{3*4}}}$	…	$\frac{\mathbf{1}}{\sqrt{\mathbf{13*14}}}$
2515318.18	1778598.54	1026874.35	726109.81	…	186447.79
368406.36	260502.64	150401.27	106349.76	…	27308.10
3175749.34	2245593.89	1296494.24	916759.87	…	235402.20
8371525.85	5919562.70	3417661.12	2416651.35	…	620538.78

Now all values that are smaller than or equal $(a_h - 1)$ will be removed individually from all strata and the values that are greater than $(b_h - 1)$, by substitute in equation (7.1 and 7.2)

Table 13. The lower and upper limit for each stratum

Strata	*Lower limit*	*Upper limit*
First strata	1 – 1 = 0	5 – 1 = 4
Second strata	1 – 1 = 0	5 – 1= 4
Third strata	2 – 1 = 1	6 – 1 = 5
Fourth strata	8 – 1 = 7	12 – 1 = 11

After deleting values, we will be left with these values in each stratum:

Table 14. The sample size and priority value for each stratum

STRATUM 1			STRATUM 2		
SAMPLE SIZE	$N_1\sigma_1$ $= 2515318.17$	Priority value	Sample Size	$N_2\sigma_2$ $= 368406.36$	Priority value
1	*1/√(1*2)*	*1,778,599*	*1*	*1/√(1*2)*	*260,503*
2	*1/√(2*3)*	*1,026,874*	*2*	*1/√(2*3)*	*150,401*
3	*1/√(3*4)*	*726,110*	*3*	*1/√(3*4)*	*106,350*
4	*1/√(4*5)*	*562,442*	*4*	*1/√(4*5)*	*82,378*
STRATUM 3			Stratum 4		
SAMPLE SIZE	$N_3\sigma_3$= 3175749.34	Priority value	Sample Size	$N_4\sigma_4$= 8371525.85	Priority value
2	*1/√(2*3)*	*2,245,594*	*8*	*1/√(345*346)*	*986,594*
3	*1/√(3*4)*	*1,296,494*	*9*	*1/√(346*347)*	*82,436*
4	*1/√(4*5)*	*916,760*	*10*	*1/√(347*348)*	*798,194*
5	*1/√(5*6)*	*710,119*	*11*	*1/√(348*349)*	*728,648*

Now we need to find the largest priority number by substitute in equation (8), 18 – (1+1+2+8) = 18 – 12= 6 which can be found in bold in table (14).

last step is finding all the sum that give the total number of sample size (n=18) with minimum variance which will be 11,254,355,944,795.50 with $n_1 = 3$, $n_2 = 1$, $n_3 = 5$, $n_4 = 9$.

3.2.3.2 EXACT OPTIMAL ALLOCATION WITH FIXED VARIANCE

It is given that V_0 will be the total salary which is equal 8,010,540,000 and the coefficient of variance equal 5.90% so, substitute in equation (9) we will get $V_0 = (5.90\% * 8010540000) = 22337142254986$ and we will start our sampling with 3 unit with keeping the third strata fixed since it has the largest variance.

Table 15. Number of population and variance in each stratum

STRATA	N	$\boldsymbol{\sigma^2}$
1	3	8,475,364.29
2	3	9,270,080.55
3	36	7,781,931,997.93
4	3	377,272,113.39

To calculate the third variance, we will use equation (10)

Table 16. The step of calculation the first variance

STRATA	FORMULA	RESULT
FIRST STRATA	$864^2 * \left(1 - \frac{3}{864}\right) * \frac{8475364.29}{3}$	2,101,619,133,041.82
SECOND STRATA	$121^2 * \left(1 - \frac{3}{121}\right) * \frac{9270080.55}{3}$	44,119,403,386.26
THIRD STRATA	$36^2 * \left(1 - \frac{36}{36}\right) * \frac{7781931997.93}{36}$	0
FOURTH STRATA	$431^2 \left(1 - \frac{3}{431}\right) \frac{377272113.39}{3}$	23,198,210,737,479.50
$\boldsymbol{V_3} = 25343949273907.60$		

Since V_3 is larger than V_0 we will construct the array and increase the stratum that will have the largest priority value until we get variance less than V_0 as table (18). note that stratum three will not be included in the array because it has a very large variance, we will keep the sample size for it fixed which will be $n_3 = 36$.

Table 17. Construct array to find the largest priority values

Stratum $N_h S_h$	$\frac{1}{\sqrt{1*2}}$	$\frac{1}{\sqrt{2*3}}$	$\frac{1}{\sqrt{3*4}}$	$\frac{1}{\sqrt{4*5}}$	...	$\frac{1}{\sqrt{27*28}}$
Stratum 1					...	
2515318.17	1,778,598.54	1,026,874.35	296,433.09	562,442.24		3,452.74
Stratum 2					...	
368406.36	260,502.64	150,401.27	43,417.11	82,378.17		505.71
Stratum 4					...	
8371525.85	5,919,562.70	3,417,661.12	986,593.78	1,871,930.09		11,491.46

From the array we indicate which stratum have the largest priority value to increase the sample size by one in each step and calculate the variance by equation (8) until we reach variance that is smaller than V_0.

Table 18. The calculation of sample size and variance

Step	n_1	n_2	n_3	n_4	Variance
1	3	3	36	3	25,343,949,273,907.60
2	3	3	36	4	24,503,500,618,131.10
3	3	3	36	5	23,999,231,424,665.20
4	3	3	36	6	23,663,051,962,354.50
5	3	3	36	7	23,422,923,774,989.80
6	3	3	36	8	23,242,827,634,466.30
7	3	3	36	9	23,102,752,858,503.50
8	3	3	36	10	22,990,693,037,733.30
9	3	3	36	11	22,899,007,729,830.40
10	3	3	36	12	22,822,603,306,578.00
11	3	3	36	13	22,757,953,409,979.80
12	3	3	36	14	22,702,539,212,895.60
13	3	3	36	15	22,654,513,575,422.70
14	3	3	36	16	22,612,491,142,633.90
15	3	3	36	17	22,575,412,525,467.20
16	3	3	36	18	22,542,453,754,652.50
17	3	3	36	19	22,512,964,328,134.00
18	3	3	36	20	22,486,423,844,267.40
19	3	3	36	21	22,462,411,025,530.90
20	3	3	36	22	22,440,581,190,315.90
21	3	3	36	23	22,420,649,601,641.40
22	3	3	36	24	22,402,378,978,689.70
23	3	3	36	25	22,385,570,005,574.20
24	3	3	36	26	22,370,054,030,390.60

25	3	3	36	27	22,355,687,386,702.10
26	3	3	36	28	22,342,346,931,848.50
27	3	3	36	29	22,329,926,508,364.10

Since in step 27 the variance is smaller than V_0we stop in step 26, and will get the following result, the total sample size (n) equal 70 with $n_1 = 3$, $n_2 = 3$, $n_3 = 36$, $n_4 = 28$, and the variance equal 4486078721809.65

3.3 RESULT

Stratification by optimal rule of stratification give better result than establishment activity because it gives smaller variance in all three methods.

With optimal rule of stratification Optimal allocation variance (20,862,602,183.23) is smaller than proportional allocation variance (1,126,501,395,468.37) and the powerful method is exact optimal allocation which give the smallest variance (13,661,852,139.44) between all the three methods.

With stratification by establishments activity Optimal allocation variance (11,597,481,983,212.70) is smaller than proportional allocation (23,768,810,254,938.10) and the powerful method is exact optimal allocation which give the smallest variance (11,254,355,944,795.50) between all the three methods.

With optimal rule of stratification the exact optimal allocation with predetermined variance gives sample size $n = 300$ and the variance is 192,297,032,941.57 but in stratification by establishments activity the exact optimal allocation with predetermined variance gives sample size smaller than optimal rule of stratification $n = 70$ but the variance is 4,486,078,721,809.65 which is much larger than stratification by establishment activity

CHAPTER 4: SIMULATION

In simulation, we estimate number of employee and their salary in each establishment with the suitable distribution, the second step we select 10% sample which is reasonable sample and it is approximately n=145. units from all strata and calculate their variances (S^2). the selection was stratified random sampling, by repetition we select 10 sample to watch how the variance is change. for code see (Annexe B).

Table 19. Optimal allocation

ID	n1	n2	n3	n4	Variance
1	84	13	6	42	2,208,407,489,241.04
2	72	17	5	51	25,186,824,985,616.10
3	86	14	3	42	13,893,064,500,612.60
4	92	12	9	32	102,290,612,476.67
5	75	15	3	52	542,292,201,682.87
6	93	16	5	37	4,096,738,301,868.76
7	92	7	5	41	2,323,814,554,965.02
8	85	18	1	41	46,042,634,467.37
9	77	13	2	53	41,131,577,729.76
10	97	9	4	35	2,673,890,932,597.11

Now we use the same sample in each 10 cases to find the exact sample allocation we will use $\sigma^2 = S^2$.

Table 20. Exact sample allocation

ID	n1	n2	n3	n4	Variance
1	83	13	5	44	2,110,961,153,073.44
2	71	16	4	54	23,919,594,847,849.60
3	86	16	2	41	12,786,446,150,404.70
4	91	11	10	33	97,306,768,594.52
5	75	15	4	51	420,534,865,918.79
6	93	15	4	33	3,850,377,542,826.98
7	92	6	4	43	2,223,443,462,733.27
8	86	17	2	41	45,532,488,841.96
9	99	8	4	34	40,345,740,498.06
10	77	12	2	54	2,625,287,093,692.72

Output shows that using optimal sample give larger variance than exact sample allocation, the difference could be large in some sample, so we can conclude that exact sample allocation is more powerful method than optimal allocation since it give smaller variance with same sample size (n=145). Figure (1) show a line plot for the variance for both method in our simulation.

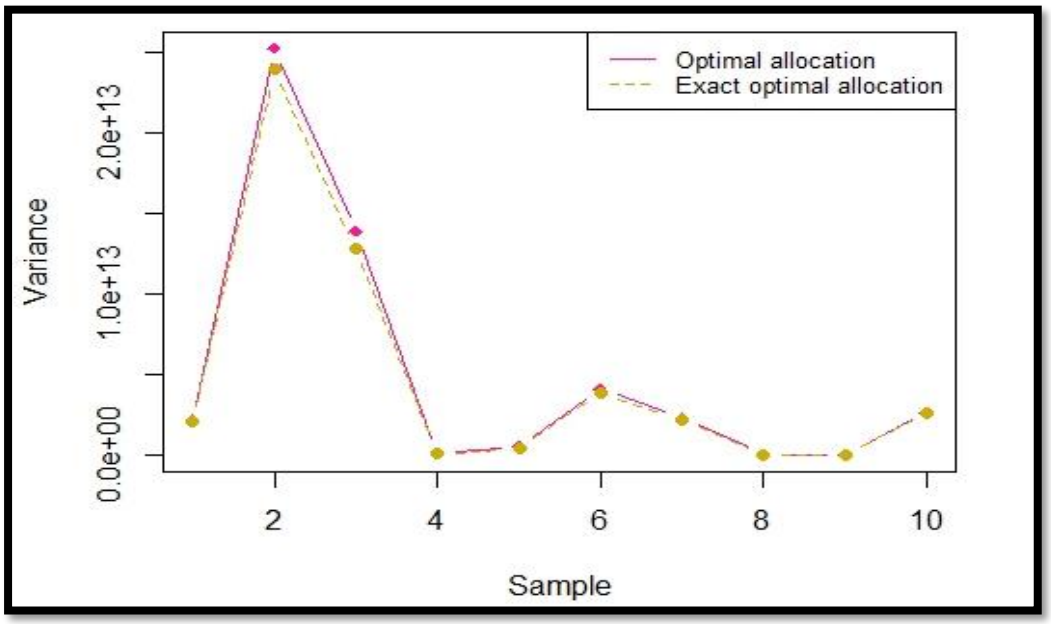

Figure 1. Optimal allocation VS Exact optimal allocation

CHAPTER 5: SUMMARY AND CONCLUSION

5.1 SUMMARY

In this study a real data from industry and Energy Statistics in 2017 have been used which is provided by Planning and Statistics Authority in Qatar. The data considers 21 groups of different establishments activity having more than ten employees. A simulation has been done to come up with all the establishments and their number of employees with their salary, the total number of establishments are 1452. Then, two methods of stratification have been applied. The first stratification method is optimal rule of stratification and the strata is obtained by employee number, the second stratification method is by establishment activity, so the strata was obtained dependent on the activity of each establishment, and both stratification methods consist of 4 strata. Allocating the sample size for each stratum was done by three allocation methods, which are optimal allocation, proportion allocation and exact optimal allocation. The results in both methods of stratification were in favor of the exact optimal allocation because it gives smaller variances compare to with optimal allocation and proportional allocation. Moreover, in exact optimal allocation with predetermining variance, we get smaller number of sample size for stratification by establishment activity except that the variance is very large. Otherwise with optimal rule of stratification we get an adequate number of sample size with smaller variance. Among the stratification methods, optimal rule of stratification gives better result compare to stratification by establishment activity.

5.2 CONCLUSION

Exact optimal allocation is a method of choosing sample in stratified random sampling that will guarantee that we will never get non-integer result for the sample size as optimal allocation (Neyman). It will also give smaller variance than the other method which are optimal allocation and proportional allocation. Based on the results, we can get better representative sample for the population with smaller variance and that will give less cost for the study by using exact optimal allocation, and this method will help to give a good estimation for the salary for the GDP (gross domestic product) for Qatar. Also, among choosing the number of sample size, the stratification by optimal rule of stratification get better result but the stratification by establishment activity gives smaller sample with large variances. So that the stratification of the population maybe changes to ensure getting the best-represented sample of the population.

In case the researcher has large number of strata, the matrix for finding the minimum variance will be very large, which needs a computer programming to calculate it. A further research can be writing such a computer program.

REFERENCES

Arnab, R., 2017, *Survey sampling theory and applications*, Academic press, United Kingdom, 930pp.

Butcher, H., 1966, *Sampling in Educational research*, Manchester university press, Manchester, 33pp.

Chaudhuri, A.,Stenger. H., 2005, *Survey sampling theory and methods*, Taylor and Ftancis Group, United State of America, 371pp.

Cochran. W., 1977, *Sampling techniques*, john Wiley & sons, United States of America A john Wiley & sons a, 428pp.

Deming, W., 1966, Some Theory of sampling, General Publishing, Canada, 602pp.

Groves, R., Fowler, F., Couper, M.,Lepkowski,J,.Singer,E.,Tourangeau,R., 2009, *Survey methodology*, A john Wiley & sons . Hoboken, New Jersey.450pp.

Kalton, G.,1983, *Introduction to survey sampling*, Sage publication, United State of America,96pp.

Levy, P., Lemeshow, S., 2013, *Sampling of populations: Method and applications*, John Wiley & sons, Hoboken, New Jersey, 616pp.

Lohr. S.,2009, *Sampling: Design and analysis*, Cengage learning, United State of America, 608pp.

Mukhopadhyay, P., 2008,*Theory and methods of survey sampling*, Asoke k ,New-Delhi, 580pp.

Neyman, J.,1934, *on the two different aspect of the representive method: The method of stratified sampling and the method of purposive selection*, Blackwell publishing for the royal statistical society, 97, 558-625.

Padilla, M., Olofsson, P., Stehman, S., Tansey, K., Chuvieco, E. 2017.

Stratification and sample allocation for reference burned area data. Remote Sensing of Environment. 203,240–255

Pfeffermann, D.,2009, Sample surveys design, methods and application, Diacri, India,722

Pirzadeh, H., Shanian, S. Hamou-lhadj, A., Alawneh, L., Shafiee, A. 2013. *Stratified sampling of execution traces: Execution phases serving as strata. Science of Computer Programming*, 78,1099–1118.

Planning and Statistics Authority of Qatar. https://www.mdps.gov.qa

Sampath. S., 2001, *Sampling theory and methods*, Narosa publishing house, New Delhi, 184pp.

Sarndal, C., Swensson, B.,Wretman, J., 1992, *Model assisted survey sampling*, Springer -verlag, New York , 694pp.

Scheaffer, R., Mendenhall, W., Ott, R. 2006. *Elementary survey sampling*. Thomson Brooks. United States of America. 464pp.

Singh, R., Mangat, N. S., 1996, *Elements of survey sampling*, Springer Science and Business Media, 390pp.

Som, R., 1995, *Practical Sampling Techniques*, CRC Press, United States of America. 672pp.

Thompson, Steven K., 2012, *Sampling*, A john Wiley & sons, Hoboken, New Jersey., 436pp.

Wright, T. 2017. *Exact optimal sample allocation: More efficient than Neyman*. Statistics and Probability Letters, 129, 50–57.

APPENDIX A: DATA

Num.	Activity Code	Main Economic Activity	Salary	Number of Employees	Number of Establishment.
Total			**8019839**	**110067**	**1452**
1	15	Manufacture of leather and related products	1465	104	4
2	21	Manufacture of basic pharmaceutical products and pharmaceutical preparations	2738	216	1
3	13	Manufacture of textiles	11655	489	9
4	29	Manufacture of motor vehicles, trailers and semi- trailers	12346	559	7
5	30	Manufacture of other transport equipment	62255	564	2
6	17	Manufacture of paper and paper products	19595	687	9
7	19	Manufacture of coke and refined petroleum products	426044	846	3
8	33	Repair and installation of machinery and equipment	117895	999	26
9	27	Manufacture of electrical equipment	91065	1459	20
10	28	Manufacture of machinery and equipment n.e.c.	58948	1840	3
11	11	Manufacture of beverages	140383	2630	12
12	31	Manufacture of furniture	143828	3804	58
13	18	Printing and reproduction of recorded media	297406	4203	38
14	24	Manufacture of basic metals	1160988	5142	8
15	16	Manufacture of wood and of products of wood and cork, except furniture, manufacture of aeticles of straw and plaiting materials plaiting materials	176765	5598	93
16	22	Manufacture of rubber and plastics products	265818	6132	95
17	14	Manufacture of wearing apparel	144481	6705	356
18	10	Manufacture of food products	254856	7016	109
19	20	Manufacture of chemicals and chemical products	2685331	8376	32
20	23	Manufacture of other non-metallic mineral products	1103196	26032	202

21	25	Manufacture of fabricated metal products, except machinery and equipment	834025	26525	365

APPENDIX B: R CODE

```
# generating population for establishment
# import package to read from excel then import the data
da=Data
po=numeric(0)
for(i in 1:21){
  sa=as.numeric(da[i,4])
  n=as.integer(da[i,5])
  m=as.integer(da[i,6])
  em=rmultinom(1,n,rep(1/m,m))
  sal=sa*(em/sum(em))
  print(sal)
  d=cbind(rep(i,m),sal,em)
  po=rbind(po,d)
}
dim(po)
sum(da[,6])
em=rmultinom(1,104,rep(1/4,4))
sa=2738*em/sum(em)

# generate Function to take sample from each stratum and calculate some estimators
N1=sum(da[,6]==1)
N2=sum(da[,6]==2)
N3=sum(da[,6]==3)
N4=sum(da[,6]==4)
```

```
numeric(0)
# take random sample of size 151 and calc the total variance by loop
sam151=rep(da,10000)
for ( i in 1:10000){
  samp=sample(da,151,replace = FALSE)
  sam[i]=mean(samp)
  }
sam151
# Split  the sample by the strata and calculate the variance for the salary
sam.split=split(sam151$salary,sam$strata2)
vart=function(x)
  x=  N[i]^2*(1-(n[i]/N[i])*(s[i]/N[i])
  sapply(sam.split, function(x))
         vart=sapply(sam.split,vart))
# To find mean and variance for each stratum
ave=tapply(da[,3],da[,6],mean)
var=tapply(da[,3],da[,6],var)
# To calculate optimall allocation
N=N1+N2+N3+N4
sumNsig2=(N1*var1)+(N2*var2)+(N3*var3)+(N4*var4)
sumNsig=(N1*sqrt(var1))+(N2*sqrt(var2))+(N3*sqrt(var3))+(N4*sqrt(var4))
n=((sumNsig*sumNsig)/(N^2+sumNsig2))
n1o=n*(N1*sqrt(var1)/sumNsig)
n2o=n*(N2*sqrt(var2)/sumNsig)
```

```
n3o=n*(N3*sqrt(var3)/sumNsig)

n4o=n*(N4*sqrt(var4)/sumNsig)

# To calculate proportional allocation

n1p=n*(N1/N)

n2p=n*(N2/N)

n3p=n*(N3/N)

n4p=n*(N4/N)

# Drow the chart

x=c(2208407489241.04,25186824985616.10,13893064500612.60,102290612476.67,
542292201682.87,4096738301868.76,2323814554965.02,46042634467.37,41131577
729.76,2673890932597.11)

y=c(2110961153073.44,23919594847849.60,12786446150404.70,97306768594.52,
420534865918.79,3850377542826.98,2223443462733.27,45532488841.96,40345740
498.06,2625287093692.72)

plot(x,pch=18,type = "b",col = "deeppink", xlab = "Sample", ylab = "Variance",
    main = "Exact optimal allocation VS Optimal allocation")

lines(y, pch=16,type = "b", col = "gold3",lty=2)

legend("topright", legend=c("Optimal allocation", "Exact optimal
allocation"),col=c("deeppink", "gold3"), lty=1:2, cex=0.8)
```