Statistics institution

Repeated sampling in Successive Survey

(RSSS)

Xiaolu Cao

15-högskolepoängsuppsats inom Statistik III, HT2011

Supervisor: Mikael Möller

Abstract:

In this thesis, the idea of sampling design for successive surveys is investigated. A good sampling design can influence the result of the survey. It can decrease the bias estimation of survey in real world. Author wants to compare a new sampling design with a classical sampling design. The thesis concerned on using repeated sampling within realm of stratified random sampling. Thesis shows this knowledge in theory and how the theory is applied. The data for application are generated by a Monte-Carlo method. The result from simulations shows that using repeated sampling to estimate the population mean gives higher precision and less variance than to use fixed sample

Thanks

I would like to express my gratitude to Mikael Möller who support and guidance during my writing this thesis.

Contents

1 Introduction **2**

1.1 Background 2

1.2 Related work 3

1.3 Data set 3

2 Theoretical Framwork **5**

2.1 Notations and concept of repeated sampling 5

2.2 Repeated sampling 5

2.3 Simple random sampling 6

2.4 Repeated sampling within realm of simple random sampling 7

2.5 Repeated sampling within realm of stratified random sampling 12

2.6 Estimation of sample size: 13

3 Experimental results **15**

3.1 Implementation issues 15

3.2 Experimental result for data set 1: 15

3.2.1 Sample size 15

3.2.2 Comparison for fixed sample and repeated sample 16

3.2.3 Estimated variance 16

3.3 Optimality 16

3.4 Experimental result for data set 2 17

3.4.1 Experiment result for testing scenario 17

3.4.2 Sample size 17

3.4.3 Comparison for fixed sample and repeated sample 17

3.4.4 Estimate variance 17

3.4.5 Summary of experiment 17

4 Conclusion and discussion **19**

References **20**

1 Introduction

1.1 Background

Successive surveys means survey on repeated occasions. Many studies have used survey method: e.g. sociological, economical, medical etc. The objective for these are to estimate characteristics of a population on 'repeated occasions' in order to measure time-trend as well as current values of the characteristics (see [Rao and Graham]). For example, the estimation of the yearly average income for people living in Stockholm county. When changes in time-dependent population values are to be examined, several sampling alternatives, as listed by Yate (1960), can be used. These involve: (i) A new sample on each occasion; (ii) A fixed sample used on all occasions; (iii) A subsample of the original sample on second occasion; (iv) A partial replacement of units from occasion to occasion. In this thesis we concentrate on comparing (ii) *a fixed sample used on all occasions* and (iv) *a partial replacement of units from occasion to occasion* (see [Manoussakis]). A partial replacement of units from occasion to occasion is usually known as repeated sampling; i.e. part of the sample is replaced at equidistance times. During this process on the h^{th} occasion we may have parts of the sample that are matched with the $(h-1)^{th}$ occasion, parts that are matched with both the $(h-1)^{th}$ and the $(h-2)^{th}$ occasion, and so on (see [Patterson]). Finding the optimal replacement strategy for two or more occasions is very important. Section 1 will compare the conventional method (ii) and repeated sampling within the realm of stratified random sampling.

Stratified sampling divide the population Ω_N into K homogeneous subpopulations of size $N_1, N_2, \ldots, N_K$. Each subpopulation is called a stratum. Draw from each stratum a sample, the drawing must be independent in different strata. If the process of drawing a sample is random, then the whole procedure is called stratified random sampling. Stratified sampling is a common technique and works well for populations with a variety of attributes. The principal reasons are (see [Cochran]):

1. If data of known precision are wanted for certain subdivisions of the population, it is advisable to treat each subdivision as a "population" in its own right.

2. Administrative convenience may dictate the use of stratification; for example, the agency conducting the survey may have field offices, each of which can supervise the survey for a part of the population.

3. Sampling problems may differ markedly in different parts of the population. With human populations, people living in institutions (e.g., hotels, hospitals, prisons) are often placed in a different stratum from people living in ordinary homes because a different approach to the sampling is appropriate for the two situations. In sampling businesses we may possess a list of the large firms, which are placed in a separate stratum. Some type of area sampling may have to be used for the smaller firms.

4. Stratification may produce a gain in precision in the estimates of characteristics of the whole population. If it is possible to divide a heterogeneous population into subpopulations, each of which is internally homogeneous, stratification may produce a gain in precision in the estimates of the characteristics. This is suggested by the name stratum,

> with its implication of a division into layers. If each stratum is homogeneous, in that the measurements vary little from one unit to another, a precise estimate of any stratum mean can be obtained from a small sample in that stratum. These estimates can then be combined into a precise estimate for the whole population.

For reasons above, this thesis will focus on using repeated sampling within the realm of stratified random sampling. The improvement by using repeated sample will be investigated by both theoretically derivations and data simulation.

1.2 Related work

Random sampling, stratified sampling, cluster sampling two-stage stratified sampling, etc usually use fixed samples. In these cases, every occasion use the same units.

However, measuring the same units on successive occasion surveys might not be a wise choice. This because using the same fixed sample may negatively influence investigated units (e.g. people) and this might influence the final result. This motivates new methods to load, for each occasion, a "new" sample. On the other hand using new sample on each occasion, the continuity of statistical features between current sample and new sample can not be guaranteed also new samples cost extra. Because of the reasons above the repeated sampling technique was invented. This technique consists of partial replacement of units from occasion to occasion.On the $(h+1)^{th}$ occasion, do a partial replacement of units.

The process of repeated sampling may thus be summarized as: The sample for the $(h+1)^{th}$ occasion include some units which match with the h^{th} occasion, parts which are matched with both h^{th} and $(h-1)^{th}$ occasion, and so on.

The method of repeated sampling within the realm of simple random sampling is mature. But there are only a few paper about the use of repeated sampling within the realm of more complex sampling, e.g. stratified random sampling, stratified cluster sampling, two-stage stratified sampling etc.

Stratified sampling is a method of sampling from a heterogenous population. Stratification is the process of dividing this population into homogeneous stratums before sampling. This thesis will use repeated sampling within stratified sampling. Here stratified sampling contains stratified random sampling, stratified cluster sampling and two-stage stratified sampling. In this thesis we focuses on repeated sampling within the realm of stratified random sampling.

1.3 Data set

Since real data was not available we decided to use the Monte-Carlo method to generate datasets. Hence we first simulate our population from which we will take a sample. Examples of characteristics are the income for all workers who work at Stockholm University during 2007, 2008, 2009 etc, or the BMI (body mass index) for all the people who live in Stockholm county during 2006, 2007, 2008 etc. We simulate income data since they are not possible for us to get individually from public official statistics.

The expression "Monte-Carlo method" is actually very general. Monte-Carlo methods also called MC methods, are stochastic techniques which are based on pseudo random generators and probability theory. You can find Monte Carlo methods used in many different areas; from economics to nuclear physics to regulating the traffic flow. Of course the way they are applied varies widely from field to field, and there are dozens of subsets of Monte Carlo methods even within chemistry. But, strictly speaking, to call something a "Monte Carlo" experiment, all you need to do is to use random numbers when examining your problem (see [Woller]).

Generally, we assume that the data distribution follows a Gaussian distribution. Thus, we use the function `normrnd` in MATLAB to generate our dataset. From Statistics Sweden we obtain the average income for males and females living in Stockholm county during 2006-2009:

Average income (tkr)

	2006	2007	2008	2009
male	222.5	233.1	242.3	242.5
female	172.0	179.4	186.8	190.4

With aid of the Matlab function `normrnd`(μ, σ, N), we generate N Gaussian random numbers with mean μ and standard deviation σ. In our case, we generate different datasets by using given mean values (as shown in table Average income) and varying standard deviation values. Common standard deviation value can be in range of 10 to 50.

2 Theoretical Framwork

2.1 Notations and concept of repeated sampling

In the sequal we are interested in measuring the individual income in Stockholm county. More exact, we want to estimate the mean income and its standard deviation for two different sampling schemes.

We introduce the following notation

$$Y_j(h) = \text{income individual } j \text{ at occasion } h \text{ in Stockholm county}$$

where $j \in \{1, 2, \ldots, N\}$. The size of the population N is the same at every occasion h. Let

$$\begin{aligned} \Omega &= \{Y_1(h), Y_2(h), \ldots, Y_N(h)\} \\ \mu(h) &= \bar{Y}(h) = \frac{1}{N}\sum_{j=1}^{N} Y_j(h) \\ \sigma^2(h) &= \frac{1}{N-1}\sum_{j=1}^{N}\left(Y_j(h) - \bar{Y}(h)\right)^2 \end{aligned}$$

Here occasion h stands for one of the years investigated ($h = 2006, 2007, 2008, 2009$).

Onwards we will use small y to denote a sample value and our samples will be of size n. For partial replacement of units, we use m to denote a keep part and u to denote a replacement part. $\hat{\mu}_m(h)$ is the estimate of the population mean use keep part from original sample at h^{th} occasion, and $\hat{\mu}_{u+m}(h)$ is the estimate of the population mean use repeated sampling at h^{th} occasion.

2.2 Repeated sampling

In a stratified random sampling process, we do simple random sampling in each stratum. Using repeated sampling, within realm of stratified random sampling, means using repeated sampling in each stratum. Repeated sampling is a process that changes the sample from occasion to occasion.

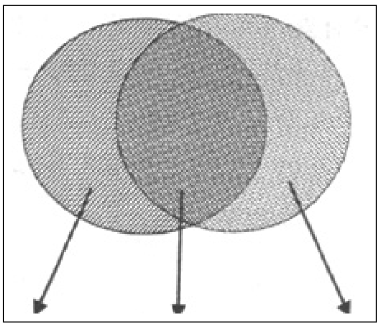

Figure 1:

dropped units kept units new units

We write $(h-1)^{th}$ occasion instead of previous occasion and h^{th} occasion instead of current occasion. On the h^{th} occasion , we need keep units in the non-replacement part of the sample used at $(h-1)^{th}$ occasion and drop the rest. We draw new observations to replace the dropped ones. This process will call repeated sampling and it is described by figure on previous page.

2.3 Simple random sampling

The idea behind simple random sampling is that the values of sample characteristics is approximetely the same as the corresponding values in the population. If we use a fixed sample for a survey then the accuracy of the measurements in this sample will decrease over time due to population changes. It is believed that a repeated sample is better than a fixed sample. In repeated sampling, we need take into account the correlation between current occasion h^{th} and previous occasion $(h-1)^{th}$. We use correlation coefficient ρ to denote this relationship. Generally, the value of ρ is between 0 and 1. When $\rho = 1$, the previous sample is not good for estimation of current population characteristics. The proportion of replacement is 100%.

Contrarily, if $\rho = 0$, that means data from the current occasion has no relationship with the previous occasion. This sample is though useful for estimation of population characteristics in the current occasion, we need not change any units in the previous sample. The proportion of replacement is 0%.

In the real world, we normally use standard deviation $\sigma(\bar{y})$ to measure how close the estimate $\bar{y}$ of a sample is to μ of the population. The bigger standard deviasion $\sigma(\bar{y})$, the flatter distribution curve and the more spread is data. In other words, the estimate of the mean has low precision. Contrarily, if $\sigma(\bar{y})$ is small that means the distribution curve is peaked, all data are close to the mean. Hence, the estimate of the mean has high precision.

In simple random sampling, the formula for standard deviation of the mean $\bar{y}$ is in sampling from

1. an infinite population:
$$\sigma(\bar{y}) = \frac{\sigma}{\sqrt{n}}$$

2. a finite population:
$$\sigma(\bar{y}) = \frac{\sigma}{\sqrt{n}}\left(\sqrt{1-\frac{n}{N}}\right)$$

From formulas above we see that standard deviation, sample size and population size will effect the standard deviation for the mean $\bar{y}$. But these formulas are not good enough to calculate the standard deviation of the mean $\bar{y}$ in repeated sampling. In repeated sampling, sample size and population size are the same at all occasions. But the sample is not the same from occasion to occasion. The correlation between the two occasions will have an effect on the sample estimate. Hence it will also have an effect on the sample characteristics for the current occasion. Further, the optimal proportion of replacement also effects the characteristics of the sample. The correlation will help us to decide how many units to replace to get an estimate of highest precision. If we want to estimate the characteristics of a population well, we need have a sample that resembles the population. These are the reasons why we should add the correlation coefficient ρ and the optimal proportion of replacement p into formulas above.

2.4 Repeated sampling within realm of simple random sampling

Firstly, only consider about a survey in two successive occasions h and $h-1$. The sample size n is same in all occasions and denote measurement variable by y. So $y\left(h\right)$ means variable at current occasion and $y\left(h-1\right)$ means variable at previous occasion.

Let m be the number of kept units ($m \leq n$) and u the number of dropped units ($u = n-m$). Then we have the following sampling schemes for occasion h given occasion $h-1$ (where we write $r \mid s$ for measures at time r given sample at time s)

1. All units are newly sampled (h) and their values are measured at the current occasion (h). In this case

$$\begin{aligned}\hat{\mu}\left(h \mid h\right) &= \bar{y}\left(h \mid h\right)\\ V(\hat{\mu}\left(h \mid h\right)) &= \frac{s^2\left(h \mid h\right)}{n}\end{aligned}$$

In the sequal we will write $\bar{y}\left(h\right)$ and $s^2\left(h\right)$ when sample time is same as measure time.

2. All units are from the previous occasion $(h-1)$ but their values are measured at the current occasion (h).

$$\begin{aligned}\hat{\mu}\left(h \mid h-1\right) &= \bar{y}\left(h \mid h-1\right)\\ V(\hat{\mu}\left(h \mid h-1\right)) &= \frac{s^2\left(h \mid h-1\right)}{n}\end{aligned}$$

3. The sample is regarded as a population in its own respect and we take a subsample, size m, of it.

Sample in current occasion is a subsample of original sample at previous occasion. It means $m < n$. The linear regression estimate is designed to increase precision by use of an auxiliary variate $Y_j\left(h-1\right)$ that is correlated with $Y_j\left(h\right)$. When the relation between $Y_j\left(h\right)$ and $Y_j\left(h-1\right)$ is examined, it may be found that although the relation is approximately linear, the line does not go through the origin. This suggests an estimate based on the linear regression of $y_j\left(h\right)$ on $y_j\left(h-1\right)$ rather than on the ratio of the two variables.

We suppose that values $y_j\left(h\right)$ and $y_j\left(h-1\right)$ are each obtained for every unit in the sample and that the population mean $\mu\left(h-1\right)$ of $\{Y_j\left(h-1\right)\}$ is known. The linear regression estimate of $\mu\left(h\right)$, the population mean of $\{Y_j\left(h\right)\}$ is for the kept part and some β (to be decided) (see [Cochran])

$$\hat{\mu}_m\left(h \mid h-1\right) = \bar{y}_m\left(h \mid h-1\right) + \beta\left[\bar{y}_n\left(h-1\right) - \bar{y}_m\left(h-1\right)\right] \tag{1}$$

Here $\hat{\mu}_m\left(h \mid h-1\right)$ is the estimate of $\mu(h)$ given by the kept part, at time h, with sample, at time $h-1$. We let $f=\frac{m}{n}$, then estimate of variance for population is

$$\begin{aligned}
V_{\min}\left(\hat{\mu}_m\left(h\mid h-1\right)\right) &= \frac{1-f}{m} \\
&\quad \times \frac{\sum_{j=1}^{n}\left\{\left[\left(y_j\left(h\right)-\bar{y}_m\left(h\mid h-1\right)\right)-\beta\left(y_j\left(h-1\right)-\bar{y}\left(h-1\right)\right)\right]\right\}^2}{n-1} \\
&= \frac{1-f}{m}\times\left[s^2\left(h\right)-2\beta s\left(h,h-1\right)+\beta^2 s^2\left(h-1\right)\right] \qquad (2)
\end{aligned}$$

Where $s\left(h,h-1\right)$ is the covariance between h^{th} and $(h-1)^{th}$ occasion.

Here, the regression coefficient is estimated as

$$\hat{\beta}=\frac{\sum_{j=1}^{n}\left(y_j\left(h\right)-\bar{y}_m\left(h\mid h-1\right)\right)\left(y_j\left(h-1\right)-\bar{y}\left(h-1\right)\right)}{\sum_{j=1}^{n}\left(y_j\left(h-1\right)-\bar{y}\left(h-1\right)\right)^2}=\frac{s\left(h,h-1\right)}{s^2\left(h-1\right)} \qquad (3)$$

The ρ is the population correlation coefficient between h^{th} and $\left(h-1\right)^{th}$ occasion, and is estimated as

$$\hat{\rho}^2=\frac{s^2\left(h,h-1\right)}{s^2\left(h\right)\times s^2\left(h-1\right)} \qquad (4)$$

here $\hat{\rho}$ depends on occasion h and $\left(h-1\right)$.

We combine equation (2), (3) and (4) and get

$$V_{\min}\left(\hat{\mu}_m\left(h\mid h-1\right)\right)=\frac{s^2\left(h\right)\left(1-\hat{\rho}^2\right)}{m}+\hat{\rho}^2\frac{s^2\left(h\right)}{n} \qquad (5)$$

4. The corresponding calculations for the replacement part give us

   All observations in replacement part at current occasion are new and have no correlation with same observations at previous occasion. So the sample mean is

$$\hat{\mu}_u\left(h\right)=\bar{y}_u(h)=\frac{1}{u}\sum_{j=1}^{u}y_j(h)$$

   and

$$V(\hat{\mu}_u\left(h\right))=\frac{s^2\left(h\right)}{u}$$

5. From 3 we get $\hat{\mu}_m\left(h\mid h-1\right)$ and from 4 we get $\hat{\mu}_u\left(h\right)$. Then, at time h, all units are measured, and combined into the estimate

$$\hat{\mu}_{u+m}\left(h\right)=\phi\times\hat{\mu}_u\left(h\right)+\left(1-\phi\right)\times\hat{\mu}_m(h\mid h-1.) \qquad (6)$$

Where ϕ is a constant weight factor. and observe that $\bar{y}_u(h \mid h)$ is independent from $\hat{\mu}_m(h \mid h-1.)$. In following equation, and $g\,(h)$ is shor for $g(h \mid h)$. From equation (1) into (6) give

$$\hat{\mu}_{u+m}\,(h) = \phi \cdot \hat{\mu}_u\,(h) + (1-\phi) \cdot \{\bar{y}_m(h \mid h-1) + \beta\,[\bar{y}_n\,(h-1) - \bar{y}_m\,(h-1)]\} \tag{7}$$

To calculate the variance we use equations (5), (9) and (8) to get

$$V(\hat{\mu}_u\,(h)) = \frac{s^2\,(h)}{u} = \frac{1}{w_u\,(h)} \tag{8}$$

$$\frac{s^2\,(h)\,(1-\hat{\rho}^2)}{m} + \hat{\rho}^2 \frac{s^2\,(h)}{n} = \frac{1}{w_m\,(h)} \tag{9}$$

$$\begin{aligned} V(\hat{\mu}_{u+m}\,(h)) &= \phi^2 \cdot V(\hat{\mu}_u\,(h)) + (1-\phi)^2 \cdot V_{\min}(\hat{\mu}_m(h \mid h-1.)) && (10)\\ &= \phi^2 \cdot \frac{s^2\,(h)}{u} + (1-\phi)^2 \cdot (\frac{s^2\,(h)\,(1-\hat{\rho}^2)}{m} + \hat{\rho}^2 \frac{s^2\,(h)}{n}) && (11)\\ &= \phi^2 \cdot \frac{1}{w_u\,(h)} + (1-\phi)^2 \cdot \frac{1}{w_m\,(h)} && (12) \end{aligned}$$

We want to get the minimum variance, and an estimate of ϕ. $V\left(\hat{\mu}_{u+m}\,(h)\right)$ we take derivative with respect to ϕ and put it equal to zero. Then we can get the optimal ϕ for minimum variance estimate. This gives the weight factor.

$$\phi = \frac{w_u\,(h)}{w_m\,(h) + w_u\,(h)}$$

It also means:

$$\phi = \frac{V(\hat{\mu}_u\,(h))}{V(\hat{\mu}_m(h \mid h-1)) + V(\hat{\mu}_u\,(h))} \tag{13}$$

Then we use (13) instead of ϕ in (12) to get optimal variance estimate by optimal ϕ

$$V_{optimal}\left(\hat{\mu}_{u+m}\,(h)\right) = \frac{1}{w_u\,(h) + w_m\,(h)} = \frac{s^2\,(h)\left(n - u\hat{\rho}^2\right)}{n^2 - u^2\hat{\rho}^2} \tag{14}$$

The optimal proportion of replacement is an important estimation value. We want to estimate an optimal proportion of replacement which can help us to get the minimum variance for the estimate of population mean. In other words, it can help us to estimate, the population mean, with highest precision. If we want to get the optimal proportion of replacement for the minimum variance, we derivate the $V(\hat{\mu}_{u+m}\,(h))$ with respect to u and put derivative equal to zero. We get after some calculations

$$p = \frac{u}{n} = \frac{1}{1+\sqrt{1-\hat{\rho}^2}}\ (\hat{\rho} \neq 1) \tag{15}$$

use (15) into equation (14) to get minimum variance estimate by optimal ϕ and p.

$$V_{\min}\left(\hat{\mu}_{u+m}(h)\right) = \frac{s^2(h)}{2n}\left(1+\sqrt{1-\hat{\rho}^2}\right)$$

Secondly, we consider partial replacement of units on occasions that are more than two. We assume that these are $(h-2)^{th}$, $(h-1)^{th}$ and h^{th} occasion. So on h^{th} occasion, we need use estimation of sample mean in $(h-1)^{th}$ occasion to be the auxiliary variable, and use it to help us to estimate on h^{th} occasion. In fourth consequence above, $(h-1)^{th}$ occasion is the first occasion, but in this case it is the second occasion. That is the reason for us to use $\hat{\mu}_m(h \mid h-1)$ instead of $\bar{y}_n(h-1)$ and $\phi(h)$ instead of ϕ in this case. We still assume that population variance S^2 is same for all occasions. The estimation of population mean on h^{th} occasion is

$$\hat{\mu}_{u+m}(h) = \phi(h)\cdot\hat{\mu}_u(h) + (1-\phi(h))\cdot\hat{\mu}_m(h \mid h-1) \ \ (h>2) \tag{16}$$

$$\phi(h) = \frac{w_u(h)}{w_u(h)+w_m(h)} \tag{17}$$

We get it from (13)

$$\hat{\mu}_u(h) = \frac{1}{u}\sum_{i=1}^{u} y_i(h)$$

$$V(\hat{\mu}_u(h)) = \frac{s^2(h)}{u(h)} = \frac{1}{w_u(h)} \tag{18}$$

$$\hat{\mu}_m(h \mid h-1) = \bar{y}_m(h) + \beta(\hat{\mu}_m(h-1 \mid h-2) - \bar{y}_m(h-1)) \tag{19}$$

So combine the (16) and (19) to get the formula for estimate of population mean at h^{th} occasion, when $h>2$,

$$\hat{\mu}_{u+m}(h) = \phi(h)\cdot\hat{\mu}_u(h) + (1-\phi(h))\cdot\{\bar{y}_m(h) + \beta\left[\hat{\mu}_m(h-1 \mid h-2) - \bar{y}_m(h-1)\right]\}$$

Here

$$\hat{\mu}_m(h-1 \mid h-2) = \bar{y}_m(h-1) + \beta(\hat{\mu}_m(h-2 \mid h-3) - \bar{y}_m(h-2))$$

so when we calculate for variance of $\hat{\mu}_m(h \mid h-1)$, it becomes

$$V_{\min}(\hat{\mu}_m(h \mid h-1)) = \frac{s^2(h)(1-\hat{\rho}^2)}{m(h)} + \hat{\rho}^2 V_{\min}(\hat{\mu}_m(h-1))\frac{s^2(h)}{n} = \frac{1}{w_m(h)} \tag{20}$$

We combine equation (12), (17), (18) and (20)

$$V(\hat{\mu}_{u+m}(h)) = \phi_h^2\cdot\frac{s^2(h)}{u(h)} + (1-\phi_h^2)\cdot\left[\frac{s^2(h)(1-\hat{\rho}^2)}{m(h)} + \hat{\rho}^2 V_{\min}(\hat{\mu}_m(h-1))\frac{s^2(h)}{n}\right]$$

Combine equation (18), (20) and use equation (17) instead of $\phi(h)$, then we get

$$V(\hat{\mu}_{u+m}(h)) = \frac{1}{w_u(h)+w_m(h)} = g(h)\frac{S^2}{n}$$

We assume that S^2 is same for all occasions, but for the sample $s^2\left(h\right)$ is changed by occasions. If we use S^2 to estimate for the $V(\hat{\mu}_{u+m}\left(h\right))$, we need put a weight $g\left(h\right)$ here, and $g\left(h\right)$ is the ratio for variance which estimated by (h^{th}) occasion and 1^{st} occasion. At the first occasion, we assume that $s^2\left(1\right)\approx S^2$.

$$g\left(h\right)=\frac{V(\hat{\mu}_{u+m}\left(h\right))}{\frac{S^2}{n}}$$

$g\left(1\right)=1$ when $h=1$.

$$\begin{aligned}Q &= \frac{S^2}{V(\hat{\mu}_{u+m}\left(h\right))}=\frac{n}{g\left(h\right)}\\ &= S^2(w_u\left(h\right)+w_m\left(h\right))\\ &= u\left(h\right)+\frac{1}{\frac{1-\hat{\rho}^2}{m(h)}+\frac{\hat{\rho}^2 g(h-1)}{n}} \qquad (21)\end{aligned}$$

We want to get maximum of Q, it means that a minimum variance estimate, so we take the derivative of $m\left(h\right)$ in equation (21) and put it equal to zero. This gives after some simplificationt that:

$$\frac{1-\hat{\rho}^2}{m^2\left(h\right)}=(\frac{1-\hat{\rho}^2}{m\left(h\right)}+\frac{\hat{\rho}^2 g\left(h-1\right)}{n})^2$$

For keep part and this equation has the solution.

$$\frac{\hat{m}\left(h\right)}{n}=\frac{\sqrt{1-\hat{\rho}^2}}{g\left(h-1\right)\left(1+\sqrt{1-\hat{\rho}^2}\right)} \qquad (22)$$

Now combining equation (21) and (22) we get

$$\frac{1}{g\left(h\right)}=1+\frac{1-\sqrt{1-\hat{\rho}^2}}{g\left(h-1\right)\left(1+\sqrt{1-\hat{\rho}^2}\right)} \qquad (23)$$

Now define $r\left(h\right)$ and b as,

$$r\left(h\right)=\frac{1}{g\left(h\right)}$$

$$b=\frac{1-\sqrt{1-\hat{\rho}^2}}{1+\sqrt{1-\hat{\rho}^2}}$$

We may then rewrite equation (23) as

$$r\left(h\right)=1+b\cdot r\left(h-1\right)$$

Then it is readily found that

$$r\left(h\right)=1+b+b^2+b^3+...+b^{h-1}=\frac{1-b^h}{1-b}$$

since

$$r\left(h\right)=\frac{1}{g\left(h\right)}=1+b+b^2+b^3+...+b^{h-1}=\frac{1-b^h}{1-b}$$

and hence

$$\begin{aligned} g\left(h\right) &= \frac{1-b}{1-b^h} \\ g\left(h-1\right) &= \frac{1-b}{1-b^{h-1}} \end{aligned} \tag{24}$$

combine equation (22) and (24)

$$\frac{\hat{m}\left(h\right)}{n}=\frac{(1-b^{h-1})\sqrt{1-\hat{\rho}^2}}{(1-b)(1+\sqrt{1-\hat{\rho}^2})}=\frac{1-b^{h-1}}{2}$$

from this follows

$$\frac{\hat{u}\left(h\right)}{n}=1-\frac{\hat{m}\left(h\right)}{n}=\frac{1+b^{h-1}}{2}$$

which gives us optimal proportion of replacement

$$p=\frac{\hat{u}\left(h\right)}{n}=\frac{1}{2}\left[1+\left(\frac{1-\sqrt{1-\hat{\rho}^2}}{1+\sqrt{1-\hat{\rho}^2}}\right)^{h-1}\right] \tag{25}$$

2.5 Repeated sampling within realm of stratified random sampling

With aid of equations above, we can derive formulas for stratified random sampling with replacement.

N is the size for whole population. K is the number of strata. $K=1,2,\ldots,K$, the size for each stratum is $N_1,N_2,\ldots,N_K$ and $\sum_{i=1}^{K}N_i=N$. The sample size for each stratum is n_i and m_i is the number of kept units, and u_i is the number of replacement units in strata i. So that $m_i+u_i=n_i$, $i=1,2,\ldots,K$. The total sample size is n, so $\sum_{i=1}^{K}n_i=n$. The value for j^{th} observation in i^{th} stratum is denoted by Y_{ij} and $Y_{ij}\left(h\right)$ is the value for j^{th} observation in i^{th} stratum on h^{th} occasion. The weight for each stratum is $W_i=\frac{N_i}{N}$ $i=1,2,\ldots,K$ and the sample size for each stratum are $n_i=n\times W_i$.

Estimate of sample mean for stratum i is $\bar{y}_i\left(h\right)$ where

$$\bar{y}_i\left(h\right)=\frac{1}{n_i}\sum_{j=1}^{n_i}y_{ij}\left(h\right)$$

Estimate of variance for stratum i is $S_i^2\left(h\right)$ where

$$S_i^2\left(h\right)=\frac{1}{N_i}\sum_{j=1}^{N_i}(Y_{ij}\left(h\right)-\mu_i\left(h\right))^2$$

For sample variance estimation for stratum i is $s_i^2(h)$ where

$$s_i^2(h) = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(y_{ij}(h) - \bar{y}_i(h))^2$$

Estimate of population mean in stratified random sampling is

$$\hat{\mu}_{u+m}(h) = \sum_{i=1}^{L} W_i \hat{\mu}_i(h) \tag{26}$$

Combine equation (7), (16) and (26)

$$\hat{\mu}_{u+m}(h) = \sum_{i=1}^{K} W_i \left\{\phi_i(h) \cdot \bar{y}_{ui}(h) + (1 - \phi_i(h)) \cdot [\bar{y}_{mi}(h) + \beta(\hat{\mu}_{mi}(h-1) - \bar{y}_{mi}(h-1))]\right\}$$

This is the formula for minimum variance linear unbiased estimator $\hat{\mu}_{u+m}(h)$ of the true population mean $\bar{Y}(h)$ on the h^{th} occasion.

If the proportion for replacement is not equal to zero, the estimation of mean at each stratum $\hat{\mu}_i(h)$ is a minimum variance linear unbiased estimator. So that in a stratified sampling case, if $\hat{\mu}_i(h)$ is a minimum variance linear unbiased estimator, $i = 1, 2, \ldots, K$, then $\hat{\mu}_{u+m}(h) = \sum_{i=1}^{K} W_i \hat{\mu}_i(h)$ is a minimum variance linear unbiased estimator of $\bar{Y}(h)$. (see [Prabhu Ajgaonkar])

Estimation of variance for $\hat{\mu}_{u+m}(h)$ is

$$V(\hat{\mu}_{u+m}(h)) = \sum_{i=1}^{K} W_i^2 V(\hat{\mu}_i(h)) \tag{27}$$

Combine equation (11) and (27) to get

$$V(\hat{\mu}_{u+m}(h)) = \sum_{i=1}^{K} W_i^2 \left\{\phi_i^2(h)\frac{S_i^2(h)}{u_i(h)} + (1 - \phi_i^2(h))\left[\frac{S_i^2(h)(1 - \hat{\rho}_i^2(h))}{m_i(h)} + \hat{\rho}_i^2(h)\frac{S_i^2(h)}{n_i}\right]\right\}$$

Using sample variance s^2 instead of true variance S^2 in equation (11) will get the equation for estimation of stratum's variance:

$$\begin{aligned} \hat{v}(\bar{y}_i(h)) &= \phi_i^2(h)\hat{v}(\bar{y}_{ui}(h)) + (1 - \phi_i(h))^2 \hat{v}(\bar{y}_{mi}(h)) \\ &= \phi_i^2(h)\frac{s_i^2(h)}{u_i(h)} + (1 - \phi_i(h))^2\left(\frac{s_i^2(h)\left(1 - \hat{\rho}_i^2(h)\right)}{m_i(h)} + \hat{\rho}_i^2(h)\frac{s_i^2(h)}{n_i}\right) \end{aligned} \tag{28}$$

2.6 Estimation of sample size:

Using stratified random sampling in the successive occasion, the sample size n and of population size N are the same for all occasions. We only need consider total sample size when

$h = 1$. Because when $h = 1$, the method of sampling is stratified random sampling. The estimation of sample size for repeated sampling, within the realm of stratified random sampling, is the same as in stratified random sampling. Here it is assumed that the estimate has a specified variance V. If instead, the margin of error d has been specified, $\mathrm{V} = (d/t)^2$, where t is the normal deviate corresponding to the allowable probability that the error will exceed the desired margin.(see [Cochran])

As a reminder $W_i = \frac{N_i}{N}$ and $w_i = \frac{n_i}{n}$ from this follows.

$$\mathrm{V} = \frac{1}{n} \sum \frac{W_i^2 S_i^2}{w_i} - \frac{1}{N} \sum W_i S_i^2$$

$$n = \frac{\sum\limits_i W_i^2 S_i^2 / w_i}{\mathrm{V} + \frac{1}{N} \sum\limits_i W_i S_i^2}$$

Presumed optimum allocation (for the fixed sample size n): $w_i \propto W_i S_i$

$$n = \frac{\sum\limits_i (W_i S_i)^2}{\mathrm{V} + \frac{1}{N} \sum\limits_i W_i S_i^2}$$

Proportional allocation: $w_i = W_i$

$$n = \frac{\sum\limits_i W_i S_i^2}{\mathrm{V} + \frac{1}{N} \sum\limits_i W_i S_i^2} \tag{29}$$

3 Experimental results

3.1 Implementation issues

This thesis focuses on a minimum variance estimation problem for a time-dependent population mean. We restrict our investigation to the case of linear unbiased estimators, and to fixed sample and population sizes at all occasions.

This section is experimental part. Firstly, we use Monte-Carlo method to generate data. This data simulation is based on the mean value of people's average yearly income in Stockholm county during 2006, 2007, 2008 and 2009. For population of income is all people living in Stockholm county, we write it $\Omega = \{Y_1, Y_2, \ldots, Y_N\}$. First, calculate the true population mean for 2006, 2007, 2008 and 2009. Value in 2006 is the base information, we only use it to decide the sample size. We use equation (29) to calculate the sample size n. Data separates into two strata by gender: one stratum is Male and one is Female. The stratum has N_1 and N_2 individuals. We denote them $\Omega_{male} = \left\{Y_1^m, Y_2^m, \ldots, Y_{N_1}^m\right\}$ and $\Omega_{female} = \left\{Y_1^f, Y_2^f, \ldots, Y_{N_2}^f\right\}$. Then we use random sampling at each stratum to get the sample and the sample sizes are n_1 and n_2. The samples are $n_1 = \left\{y_1^m, y_2^m, \ldots, y_{n_1}^m\right\}$ and $n_2 = \left\{y_1^f, y_2^f, \ldots, y_{n_2}^f\right\}$. After that, use fixed sample respectively to estimate the mean and variance for whole population in 2007, 2008 and 2009, these are $\hat{\mu}\,(07)_{fixed}$, $\hat{\mu}\,(08)_{fixed}$ and $\hat{\mu}\,(09)_{fixed}$. And then use repeated sampling to estimate the mean and variance for the population in 2008 and 2009, because 2007 is the first occasion. At the first occasion the mean and variance are the same as the fixed sample estimates. Use equation (25) to calculate the optimal size for replacement at each stratum. These are $u_{male}\,(08)$, $u_{female}\,(08)$, $u_{male}\,(09)$ and $u_{female}\,(09)$. The number of kept part is m and $n_1 - u_{male}\,(08) = m_{male}\,(08)$, $n_2 - u_{female}\,(08) = m_{female}\,(08)$, $n_1 - u_{male}\,(09) = m_{male}\,(09)$ and $n_2 - u_{female}\,(09) = m_{female}\,(09)$. Estimate the sample mean and sample variance for the new sample which is equal to replacement part plus kept part, at each stratum. After that use equation (26) and (27) to calculate the population mean and variance for 2008 and 2009. We can estimate population mean with fixed sample and estimate population mean with repeated sampling method. Which one is closer to the true mean, means use that method of sampling to get the estimate with highest precision. And the ratio of these two variances shows how much difference of precision there are between these two methods. During this experiment we only use data that is yearly income in 2007, 2008 and 2009. When we use repeated sampling to estimate the mean and variance for whole population, we estimate the optimal proportion for replacement.

3.2 Experimental result for data set 1:

3.2.1 Sample size

Calculate from base year 2006

N	N_{male}	N_{female}	n	n_{male}	n_{female}
10000	4900	5100	2312	1133	1179

Same for all occasions.

3.2.2 Comparison for fixed sample and repeated sample

	2007	2008	2009
true mean $\bar{Y}$ (tkr)	205.836	214.036	215.975
estimation of $\hat{\bar{Y}}_{fixed}$ (tkr)	205.865	213.950	215.746
estimation of $\hat{\bar{Y}}_{repeated}$(tkr)	205.865	213.999	216.031
estimation of $V(\hat{\bar{Y}}_{fixed})$	0.0567	0.06695	0.0762
estimation of $V(\hat{\bar{Y}}_{repeated})$	0.0567	0.0455	0.0561
$\left(\frac{u}{n}\right)_{male}$		0.7133	0.605
$\left(\frac{u}{n}\right)_{female}$		0.709835	0.6009
bias of estimation (fixed sample)	0.014%	0.04%	0.106%
bias of estimation (repeated sample)		0.017%	0.026%
$V(\hat{\bar{Y}}_{fixed})/V(\hat{\bar{Y}}_{repeated})$		1.213	1.1653

3.2.3 Estimated variance

From this table we see that the estimated bias is always smaller for repeated sampling when compared with "fixed" sampling. It is also found that $V(\hat{\mu}_{fixed}) > V(\hat{\mu}_{repeated})$ for both year 2008 and 2009. Hence repeated sampling is found to be a better choice.

3.3 Optimality

Here we check whether the optimal proportion of replacement is real optimum or not. The method for estimation of optimal proportion of replacement is the same in 2008 and 2009. So we a test whether the proportion estimated is optimal or not for 2008 ($h = 2$). If it is an optimal proportion in 2008, it will be also an optimal proportion in 2009. During this experiment we only change the proportion of replacement, since the correlation coefficient, regression coefficient and optimum stratum weight etc. will not change. We found estimation for optimum proportion of replacement is 0.7133 for male and 0.709835 for female.Use proportions 0.6 and 0.8 to check bias. If both 0.6 and 0.8 gives higher bias the proportion of replacement, which was gotten from last experiment, is better one. Then also check 0.65 and 0.75.

$\left(\frac{u}{n}\right)_{male}$ at second occasion	0.6	0.65	0.7133	0.75	0.8
bias of estimation	0.036%	0.022%	0.017%	0.027%	0.06%

From this experiment, change the proportion of replacement, lower and higher, which was estimated from last experiment. Follow the proportion changes closes to 0.71; the bias of estimation has decreased and closes to the 0.017% which estimated in last experiment. It means that the proportion of replacement, estimated in last experiment, is the optimum proportion of replacement.

3.4 Experimental result for data set 2

We used simulated data. Simulate again but with a bigger variance ($\sigma = 50$) and keep the total number, stratum and stratum size the same Use this new data to repeat experment above.

3.4.1 Experiment result for testing scenario

3.4.2 Sample size

Calculate from base year 2006.

N	N_{male}	N_{female}	n	n_{male}	n_{female}
10000	4900	5100	5475	2683	2792

Same for all occasions.

3.4.3 Comparison for fixed sample and repeated sample

	2007	2008	2009
true mean $\bar{Y}$ (tkr)	206.062	214.196	216.103
estimation of $\hat{\bar{Y}}_{fixed}$ (tkr)	235.808	213.750	215.703
estimation of $\hat{\bar{Y}}_{repeated}$ (tkr)	235.808	214.210	216.187
estimation of $V(\hat{\bar{Y}}_{fixed})$	0.10028	0.10881	0.12675
estimation of $V(\hat{\bar{Y}}_{repeated})$	0.10028	0.07671	0.08999
$\left(\frac{u}{n}\right)_{male}$		0.7077	0.6
$\left(\frac{u}{n}\right)_{female}$		0.71	0.6
bias of estimation (fixed sample)	14.436%	0.7037%	0.185%
bias of estimation (repeated sample)		0.0068%	0.039%
$V(\hat{\bar{Y}}_{fixed})/V(\hat{\bar{Y}}_{repeated})$		1.191	1.187

3.4.4 Estimate variance

The results also show that use repeated sampling has higher precision estimation than use fixed sampling. Hence repeated sampling if found to be a better choice. No matter the variance for data is larger or small.

3.4.5 Summary of experiment

From these experiments, the results show that using repeated sampling, within the realm of stratified sampling, to estimate mean of whole population is closer to the true population mean than when we use fixed sampling. In other words, use repeated sampling will give higher precision than fixed sampling. The formula, for repeated sampling that we derived in theory part, are useful to us to estimate population mean with higher precision. They also help us to get the optimal proportion of replacement.

4 Conclusion and discussion

This thesis has derived formulas to estimate population mean and variance when using repeated sampling within realm of stratified random sampling. After deriving the formulas we use data from SCB to simulate the whole process. The results from the simulation for data set 1 shows that, use of repeated sampling can improve the precision considerably (0.04%-0.017%) and lower the variance compared to fixed sampling when estimating the population mean at 2008. And for year 2009, we atfain compareable improvements.

For these simulations, the results are that stratified random sampling with replacement when estimating the population mean gives lower variance than to use stratified random sampling without replacement. This is due to that the sample with partial replacement on the h^{th} occasion include some units that match with $(h-1)^{th}$ occasion and some units that match both the $(h-1)^{th}$ and $(h-2)^{th}$ occasion. To use this sample to estimate population mean on h^{th} occasion will have higher precision than to use the sample which only matched $(h-2)^{th}$ occasion. During the whole process, the optimal proportion of replacement is a very important characteristic. In the section Optimality, we test if the estimate of proportion is optimal or not. We change the proportion of replacement around our estimate, from experiment for data set 1, since our estimate is 0.7133 we choose 0.6, 0.65 and 0.75, 0.8. The results shows that when proportion is closer to the estimated proportion 0.7133, then the bias for estimation of population mean do decrease. That means the estimation of proportion in experiment for data set 1 is at a real optimal proportion of replacement.

Using repeated sampling within realm of stratified random sampling is one of a sampling designs. Design of sampling is an art, a successful sampling design will have direct positive influence on the result of the survey. In this thesis the results from the experimental part shows that our assumption is correct. Using repeated sampling within realm of stratified random sampling can help us to get estimates with higher presicion than fixed sampling. This is a good use of sampling design.

But there is also alternatives to stratified random sampling such as single-stage cluster sampling, systematic sampling, double sampling, stratified cluster sampling and two-stage stratified sampling, etc. We are not sure if these methods also gives estimates with higher presicion. In this thesis we only studied that using repeated sampling within realm of stratified random sampling. So using repeated sampling within other sampling methods may give good sampling designs.

References

[Cochran] W.G. Cochran, *Sampling Techniques*, third edition, Wiley, New York

[Eckler] A.R. Eckler *Rotation Sampling*, Ann. Math. Statist. 26-664-685.

[Gbur and Sielken] E.E. Gbur and R.L. Sielken *Rotation Sampling Design,* Texas A&M University.

[Manoussakis] E. Manoussakis, *Repeated Sampling with Partial Replacement of Units*, The Annals of Statistics, Vol.5, No.4 (Jul.,1977), pp 795-802.

[Patterson] H.D.Patterson, *Sampling on successive occasions with partial replacement of units*, Journal of the Royal Statistical Society, Series B, 12 (1950), 241-55

[Prabhu Ajgaonkar] S.G.Prabhu-Ajgaonkar and B.D.Tikkiwal *On classes of estimators of the variance function of a linear estimator* Volume 18, Number 1, 15-20, DOI: 10.1007/BF02614232

[Rao and Graham] J. N. K. Rao and Jack E. Graham, *Designs for Sampling on Repeated Occasions*, Journal of the American Statistical Association, Vol.59, No.306 (Jun.,1964), pp 492-509.

[Ulam] N.M.S.Ulam, *The Monte Carlo Method*, Journal of the American Statistical Association, Vol.44, No.247. (Sep., 1949), pp.335-341

[Woller] J.Woller, *The basics of Monte Carlo Simulations*, university of Nebraska-Lincoln Physical Chemistry Lab, Spring 1996.