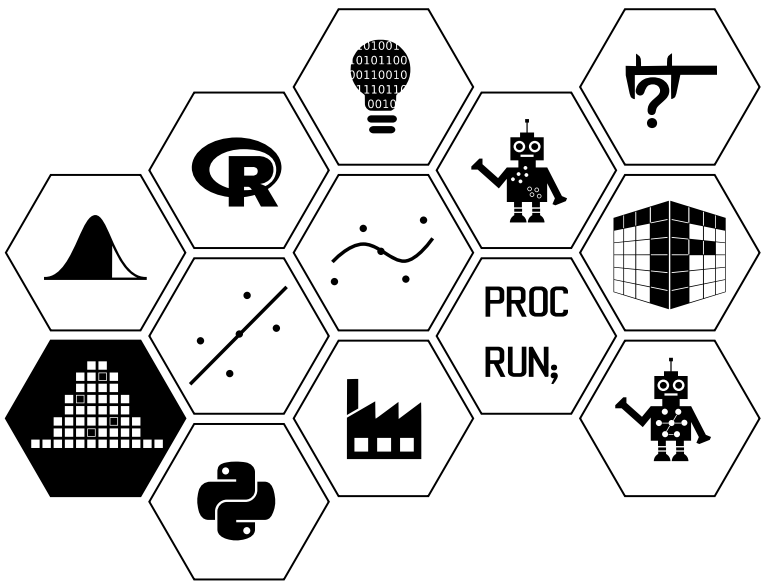
Probability and Sampling Fundamentals

Week 11: Stratified and cluster sampling



DATA ANALYTICS
GLASGOW

Week 11 learning material aims

In Week 9 and Week 10 we learned about different methods of sampling, with a particular focus on methods of **non-probability sampling** and **probability sampling** respectively. This week, we will describe additional methods of probability sampling that may improve precision relative to simple random sampling.

The material in week 11 covers:

- stratified sampling;
- cluster sampling;
- two-stage sampling.

Stratified sampling

If a population is homogeneous with respect to the characteristic, or variable, of interest then simple random sampling will yield a homogeneous sample and the sample mean will be a precise and accurate estimator of population mean and the sample drawn through simple random sampling will be representative of the population. In general, we can increase the precision of an estimator by using a sampling scheme which can reduces the heterogeneity. If the population is heterogeneous with respect to the characteristic of interest, then one such sampling procedure that can increase precision is stratified sampling.

Stratified random sampling, or stratified sampling, is a probability-based sampling method that divides a population into non-overlapping groups, or strata, with respect to a relevant characteristic, or a stratification characteristic. Within each stratum simple random sampling is used to sample units. This method ensures that units for each relevant strata are included in the sample.

Definition 1

Stratified sampling

Partitioning a population into non-overlapping groups and sampling within each group is known as stratified sampling. Each group is called a stratum.

Definition 2

Stratified random sampling

Partitioning a population into non-overlapping strata and randomly sampling within each stratum is known as stratified sampling.

The basic idea behind the stratified sampling is to divide a heterogeneous population into smaller groups or strata, such that the sampling units are homogeneous with respect to a stratification characteristic within strata and heterogeneous between strata. Strata should be non-overlapping and so we may treat each as an independent population. We may then draw a sample by simple random sampling from within each stratum. The results obtained in week 10 can be applied to each stratum. There are general principles that one should adhere to, namely

1. Strata should be non-overlapping. That is, each unit within a population should belong to exactly one stratum.

2. Strata should form a partition of the total population.

3. Units within a stratum should be more similar to each other in comparison to units between strata with respect to the characteristics of interest.

4. We should aim for homogeneity within strata relative to the population

5. The success of this sampling method depends on the choice of characteristic used to partition the population.

Suppose a population contains N units that we wish to partition into k strata such that stratum i has N_i units with

$$\sum_{i=1}^{k} N_i = N.$$

We draw a sample of size n_i from the ith stratum using simple random sampling. The total stratified sample consists of

$$\sum_{i=1}^{k} n_i = n$$

units.

Definition 3

Sampling fraction

The size of a sample stratum n_i divided by the size of the population stratum N_i is the sampling fraction of stratum i, f_i.

If we use simple random sampling within each stratum, then the inclusion probability of unit j within stratum i is $\pi_{ij} = f_i$.

Proportionate stratification

A sample with proportionate stratification is sampled such that the sampling fraction within each stratum is equivalent to the population fraction within each corresponding stratum.

Example 1

Children height

In Scotland, children complete seven years of primary school beginning in primary 1 until primary 7. Children usually begin primary school aged 4 or 5. Therefore children in primary 1 are aged 4 to 5 and children in primary 7 are aged 11 or 12.

Suppose we wanted to estimated the average height of school children in Scotland.

Of course children aged 4 or 5 will be shorter than children aged 11 or 12. Therefore, a reasonable stratification characteristic may be age or primary. Luckily, the Scottish government releases **school statistics here** and we know that the average size of a primary class lies between 20 to 26 children. In other words, we could assume equal numbers of children across primary ages.

Since we know that height depends on age and there are roughly equal numbers of children across primary ages, we could use proportionate stratification with primary as a stratification characteristic to sample from this population.

Disproportionate stratification

A sample with disproportionate stratification is sampled such that the sampling fraction within each stratum is not equivalent to the population fraction within each corresponding stratum.

Example 2

Children learning outcomes

Suppose now we are interested in the learning experience in children from different ethnic backgrounds.

If 80% of children are Caucasian, 19% Asian and 1% African and we sample 100 children proportional to this distribution, then we would sample 80 Caucasian children, 19 Asian children and 1 African child. It is highly unlikely 1 African child would represent the African population of school children. In this scenario we may choose to better represent each ethnicity by under-sampling Caucasian children and over-sampling African children

Notice in this example, we want to compare learning outcome between ethnic groups. Therefore, ethnic group is the stratification characteristic. Using disproportionate stratification seems reasonable since we do not want to under-represent any ethnicity. However, if we wished to estimate a population characteristic, such as height in the previous example, the disproportionate sample would not evenly represent the population and could potentially produce a biased estimator if there were differences in height between ethnic groups.

As discussed in Week 9 and Week 10 one of the most important areas of application for probability is statistical inference where we draw conclusions about populations based on information collected from samples. Let's define some notation specifically for stratified random sampling.

- k = the number of strata
- N_i = the total number of units within stratum i ($i = 1, \ldots, k$).
- $W_i = \frac{N_i}{N}$ = the proportion of the population within stratum i ($i = 1, \ldots, k$).
- n_i = the number of sampled units within stratum i ($i = 1, \ldots, k$).
- f_i = the sampling fraction of stratum i ($i = 1, \ldots, k$).
- y = characteristic/variable of interest.
- y_{ij} = observation j from stratum i for $j = 1, \ldots, N_i, i = 1, \ldots, k$).
- $\bar{y}_i$ = the average value of y within stratum i ($i = 1, \ldots, k$).
- $\bar{y}$ = population mean value of y.
- $\sigma_i^2 = \frac{1}{N_i} \sum_{j=1}^{N_i} (y_{ij} - \bar{y}_i)^2$ = the variance of y within stratum i ($i = 1, \ldots, k$).

We will concentrate on estimating $\bar{y}$ using stratified random sampling. Recall that within each stratum, we draw samples using simple random sampling. More importantly, samples are drawn independently from each stratum. Therefore, within each stratum, the results derived in week 10 are applicable within

each stratum. For example, $\bar{y}_i$ is an unbiased estimator of the population mean value of y within the ith stratum.

Proposition 1

$$\bar{y}_{\text{strat}} = \sum_{i=1}^{k} W_i \bar{y}_i$$

is an unbiased estimator of population mean $\bar{y}$.

Proposition 2

$$Var(\bar{y}_{\text{strat}}) = \sum_{i=1}^{k} W_i^2 (1 - f_i) \frac{S_i^2}{n_i}$$

where S_i is the population standard deviation within stratum i.

Supplement 1

Stratified unbiased estimator

Let $\bar{y}_{\text{strat}}$ define the sample mean value of y from a stratified random sample. Using the defined notation

$$\begin{aligned}
\bar{y}_{\text{strat}} &= \sum_{i=1}^{k} W_i \bar{y}_i \\
&= \sum_{i=1}^{k} \frac{N_i}{N} \bar{y}_i \\
&= \sum_{i=1}^{k} \frac{\sum_{j=1}^{n_i} y_{ij}}{N} \\
&= \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} \\
&= \bar{y}
\end{aligned}$$

Example 3

Children learning outcomes

Suppose now we are interested in the learning experience of children from different ethnic backgrounds. We want to sample school children in Scotland and ask them to rate their experience from 1 to 10 with 1 being least satisfied and 10 being most satisfied.

684,415 children are registered in Scottish primary schools where 80% of children are Caucasian, 19% Asian and 1% African and we sample 50 children proportional to this distribution. Specifically, we sampled 40 Caucasian children, 9 Asian children and 1 African child.

The results are given in the table below

Caucasian	Asian	African
10 10 7 6	10 10 10 5	6
8 2 8 4	8 4 10 10	
7 5 10 10	5	
8 6 7 1		
8 8 3 5		
6 10 10 9		
5 9 8 7		
2 5 8 10		
9 2 10 5		
7 9 10 4		

We can compute summary statistics for each group as follows.

Caucasian	Asian	African
$n_1 = 40$	$n_2 = 9$	$n_3 = 1$

Caucasian	Asian	African
$N_1 = 547,532$	$N_2 = 130,039$	$6,844$
$W_1 = 0.8$	$W_2 = 0.19$	$W_3 = 0.01$
$\bar{y}_1 = 6.95$	$\bar{y}_2 = 8$	$\bar{y}_3 = 6$
$\hat{\sigma}^2_{y_1} = 6.97$	$\hat{\sigma}^2_{y_2} = 6.75$	$\hat{\sigma}^2_{y_3} = 0$

Therefore,

$$\begin{aligned}\hat{\bar{y}} = \bar{y}_{\text{strat}} &= \sum_{i=1}^{k} W_i \bar{y}_i \\ &= 0.8 \times 6.95 + 0.19 \times 8 + 0.01 \times 6 \\ &= 7.14\end{aligned}$$

Therefore on average Caucasian and African school children score less than the overall average and Asian children score higher than the overall average. This suggests that, on average, Asian children may enjoy their learning experience more in comparison to Caucasian and African children.

As discussed previously, it is highly unlikely 1 African child would represent the African population of school children. In this scenario we may choose to better represent each ethnicity by under-sampling Caucasian children and over-sampling African children.

We can now repeat the experiment taking into account the previous statement Thus, we now sample 18 Caucasian children, 16 Asian children and 16 African children. The results are presented below.

Caucasian	Asian	African
10 1 8 10	9 9 10 8	10 4 1 8
9 6 6 4	7 9 7 10	8 2 8 9
10 2 1 4	10 10 10 9	10 5 9 1
10 3 5 6	9 10 10 10	6 2 5 9
3 10		

We can compute summary statistics for each group as follows.

Caucasian	Asian	African
$n_1 = 18$	$n_2 = 16$	$n_3 = 16$
$N_1 = 547,532$	$N_2 = 130,039$	$6,844$
$W_1 = 0.8$	$W_2 = 0.19$	$W_3 = 0.01$
$\bar{y}_1 = 6$	$\bar{y}_2 = 9.1875$	$\bar{y}_3 = 6.0625$
$\hat{\sigma}^2_{y_1} = 10.94$	$\hat{\sigma}^2_{y_2} = 1.10$	$\hat{\sigma}^2_{y_3} = 10.60$

Therefore,

$$\begin{aligned}\hat{\bar{y}} = \bar{y}_{\text{strat}} &= \sum_{i=1}^{k} W_i \bar{y}_i \\ &= 0.8 \times 6 + 0.19 \times 9.1875 + 0.01 \times 6.0625 \\ &= 6.61\end{aligned}$$

Notice that our population mean value is lower than the first estimator ($\hat{\bar{y}} = 7.14$). In this case, we over represent African and Asian children and under estimate Caucasian children.

The cost of sampling

Since populations are often large and costly to investigate, it is more common to conduct research on a sample with the total sample size being restricted by time cost and the desired level of statistical power. On the other hand, the sample size within each stratum is left to the discretion of the surveyor/researcher relative to the total sample size. For a fixed total sample size, we want to allocate samples within each stratum such that we maximise precision of our sample estimate.

Suppose each unit with stratum i comes at a cost c_i. The total cost of a sample can be described as

$$\text{Cost} = \sum_{i=1}^{k} c_i n_i$$

where n_i is the sample size within stratum i. The problem then becomes on minimising this cost function. Of course there are several ways to define a cost function but we will focus on the simple cost function defined above.

Equal sample size

If we decided to sample the same number of units from each stratum, then

$$n_i = \frac{n}{k}$$

where n is the total sample size and k the total number of strata.

Example 4

Study Time

Suppose we are interested in the average time spent per week students spend studying on the three year MSc data analytics course. We know that 100 students are enrolled in year 1, 80 students are enrolled in year 2 and 40 students are enrolled in year 3 and believe that the number of hours depends on the year of study. We will use stratified random sampling to collect data from an equal number of students across the three years . We predict a cost of £2 to contact each individual student to ask how many hours they spend studying (on average) per week and we have a total budget of £100 to collect data.

How many students can we sample in total?

$$\begin{aligned}
\text{Cost} &= \sum_{i=1}^{k} c_i n_i \\
100 &= \sum_{i=1}^{k} 2n_i \\
&= \sum_{i=1}^{3} 2\frac{n}{3} \\
&= \sum_{i=1}^{3} 2\frac{n}{3} \\
&= 2n
\end{aligned}$$

$$n = 50$$

We would sample 16 students for each of the three years, or we could sample 17 students from two years and 16 from the remaining year. Now suppose that I tell you that you need to pay the surveyor £20 for their time. How many students can we sample in total?

$$\text{Cost} = 20 + \sum_{i=1}^{k} c_i n_i$$

$$\begin{aligned} 80 &= \sum_{i=1}^{k} 2n_i \\ &= \sum_{i=1}^{3} 2\frac{n}{3} \\ &= \sum_{i=1}^{3} 2\frac{n}{3} \\ &= 2n \end{aligned}$$

$$n = 40.$$

We would sample 13 students from each of the three years, or we could sample 14 students from one year and 13 students from the remaining two years.

Proportional allocation

If we decided to sample using proportionate sampling, then

$$n_i = \frac{N_i n}{N}$$

where n is the total sample size, N the population size and N_i the population size within stratum i.

Example 5

Study Time

Suppose we are interested in the average time spent per week students spend studying on the three year MSc data analytics course. We know that 100 students are enrolled in year 1, 80 students are enrolled in year 2 and 40 students are enrolled in year 3 and believe that the number of hours depends on the year of study. We will use stratified random sampling to collect data from a number of students across the three years proportional to the size of each year. We predict a cost of £2 to contact each individual student to ask how many hours they spend studying (on average) per week and we have a total budget of £100 to collect data.

How many student can we sample in total?

In this example, we are fortunate to know the total number of students enrolled in each year with $N_1 = 100, N_2 = 80, N_3 = 40$ and so $N = 220$.

$$
\begin{aligned}
n_1 &= \frac{100n}{220} \\
&= \frac{5n}{11}
\end{aligned}
$$

$$
\begin{aligned}
n_2 &= \frac{80n}{220} \\
&= \frac{4n}{11}
\end{aligned}
$$

$$
\begin{aligned}
n_3 &= \frac{40n}{220} \\
&= \frac{2n}{11}
\end{aligned}
$$

$$
\begin{aligned}
\text{Cost} &= \sum_{i=1}^{k} c_i n_i \\
100 &= \sum_{i=1}^{k} 2n_i \\
&= 2\left[\frac{5n}{11} + \frac{4n}{11} + \frac{2n}{11}\right] \\
&= 2n
\end{aligned}
$$

$$
n = 50
$$

$$
\begin{aligned}
n_1 &= \frac{5n}{11} \\
&\approx 23
\end{aligned}
$$

$$
\begin{aligned}
n_2 &= \frac{4n}{11} \\
&\approx 18
\end{aligned}
$$

$$
\begin{aligned}
n_3 &= \frac{2n}{11} \\
&\approx 9
\end{aligned}
$$

Now suppose that I tell you that you need to pay the surveyor £20 for their time. How many students can we sample in total?

$$\text{Cost} = 20 + \sum_{i=1}^{k} c_i n_i$$

$$\begin{aligned} 80 &= \sum_{i=1}^{k} 2n_i \\ &= 2\left[\frac{5n}{11} + \frac{4n}{11} + \frac{2n}{11}\right] \\ &= 2n \end{aligned}$$

$$n = 40.$$

$$\begin{aligned} n_1 &= \frac{5n}{11} \\ &\approx 18 \end{aligned}$$

$$\begin{aligned} n_2 &= \frac{4n}{11} \\ &\approx 15 \end{aligned}$$

$$\begin{aligned} n_3 &= \frac{2n}{11} \\ &\approx 7 \end{aligned}$$

Task 1

Suppose we can divide a population into four strata with population sizes 1000, 500, 1500 and 750 respectively and we wish to sample 200 units. Using proportional allocation, complete the following table

Stratum	**1**	**2**	**3**	**4**
Population size	1000	500	1500	750
W				
Sample size				

Optimal allocation

Not only should we consider the size of each stratum but we need to also consider the precision of our estimate. In other words, we could also aim to minimise the variance of an estimator subject to our cost constraint. Recall that we are aiming to reduce the amount of variability within each stratum such that units within each stratum are homogeneous. In reality, this may not be possible. Therefore, we may want to consider sampling more units from strata that are more variable.

Neyman's allocation, a special case of optimal allocation, is

$$n_i = \frac{nN_iS_i}{\sum_{i=1}^{k} N_iS_i}$$

where n is the total sample size, N_i the population size within stratum i and S_i the population standard deviation within stratum i.

Under this construction n_i refers to the size of stratum i, N_i, and it's proportional to the standard deviation of stratum i, S_i. The amount of variation is similar across strata, then Neyman's allocation is equivalent to proportional allocation.

Proposition 3

Sample sizes $n_1, n_2, \ldots, n_k$ that miniminse $Var(\bar{y}_{\text{strat}})$ such that $\sum_{i=1}^{k} n_i = n$ are

$$n_i = \frac{nN_iS_i}{\sum_{i=1}^{k} N_iS_i}$$

In addition to the population size of each stratum N_i, we also need to know the population standard deviation of the variable of interest y within each stratum. In practice, it is likely that we would not have this information. We could however use previous studying to gauge the amount of variability in each stratum or use another variable with known variability as a proxy.

Example 6

Study Time

Suppose we are interested in the average time spent per week students spend studying on the three year MSc data analytics course. We know that 100 students are enrolled in year 1, 80 students are enrolled in year 2 and 40 students are enrolled in year 3 and believe that the number of hours depends on the year of study. We will use stratified random sampling to collect data from a number of students across the three years with Neyman's optimised

allocations. We predict a cost of £2 to contact each individual student to ask how many hours they spend studying (on average) per week and we have a total budget of £100 to collect data.

In year 1, the standard deviation of hours spent studying is 4 hours, in year 2 the standard deviation is 2 hours and in year 3 the standard deviation is 6 hour.

How many student can we sample in total?

In this example, we are fortunate to know the total number of students enrolled in each year with $N_1 = 100, N_2 = 80, N_3 = 40$ and so $N = 220$.

$$
\begin{aligned}
n_1 &= \frac{nN_1S_1}{\sum_{i=1}^{3} N_iS_i} \\
&= \frac{n \times 100 \times 4}{(100 \times 4 + 80 \times 2 + 40 \times 6)} \\
&= \frac{n400}{(800)} \\
&= \frac{5n}{10}
\end{aligned}
$$

$$
\begin{aligned}
n_2 &= \frac{nN_2S_2}{\sum_{i=1}^{3} N_iS_i} \\
&= \frac{n \times 80 \times 2}{(100 \times 4 + 80 \times 2 + 40 \times 6)} \\
&= \frac{n160}{(800)} \\
&= \frac{2n}{10}
\end{aligned}
$$

$$
\begin{aligned}
n_3 &= \frac{nN_3S_3}{\sum_{i=1}^{3} N_iS_i} \\
&= \frac{n \times 40 \times 6}{(100 \times 4 + 80 \times 2 + 40 \times 6)} \\
&= \frac{n240}{(800)} \\
&= \frac{3n}{10}
\end{aligned}
$$

$$
\begin{aligned}
\text{Cost} &= \sum_{i=1}^{k} c_i n_i \\
100 &= \sum_{i=1}^{k} 2n_i \\
&= 2\left[\frac{5n}{10} + \frac{2n}{10} + \frac{3n}{10}\right] \\
&= 2n
\end{aligned}
$$

$$
n = 50
$$

$$
\begin{aligned}
n_1 &= \frac{5n}{10} \\
&\approx 25
\end{aligned}
$$

$$
\begin{aligned}
n_2 &= \frac{2n}{10} \\
&\approx 10
\end{aligned}
$$

$$
\begin{aligned}
n_3 &= \frac{3n}{10} \\
&\approx 15
\end{aligned}
$$

Now suppose that I tell you that you need to pay the surveyor £20 for their time. How many students can we sample in total?

$$
\begin{aligned}
\text{Cost} &= 20 + \sum_{i=1}^{k} c_i n_i \\
80 &= \sum_{i=1}^{k} 2n_i \\
&= 2\left[\frac{5n}{10} + \frac{2n}{10} + \frac{3n}{10}\right] \\
&= 2n
\end{aligned}
$$

$$
n = 40.
$$

$$
\begin{aligned}
n_1 &= \frac{5n}{10} \\
&\approx 20
\end{aligned}
$$

$$
\begin{aligned}
n_2 &= \frac{2n}{10} \\
&\approx 8
\end{aligned}
$$

$$
\begin{aligned}
n_3 &= \frac{3n}{10} \\
&\approx 12
\end{aligned}
$$

Notice that in the example, year 3 had more variability both in the number of hours spent and the least number of students. With Neyman's allocation, we would sample more individuals from year 3 in comparison to year 2 although year 2 contains 80 students.

Advantages of stratified sampling

The key to stratified random sampling is to capture population characteristics in the sample through grouping of similar units. This sampling method works well if a population contains a wide variety of characteristics that may be used to group units. Stratification gives smaller standard errors and greater precision than simple random sampling. In particular, the more heterogeneity between strata the greater the precision in parameter estimates.

Disadvantages of stratified sampling

In order to optimise stratified sampling, we need information on all units within a population in order to form suitable strata. In reality, obtaining a fully well defined population is challenging. In some cases we may find strata to overlap in that some units within a population may reasonably fall into more than one stratum. For example, defining ethnicity or nationality may not be straightforward.

When should we use stratified sampling

In summary, stratified sampling is most useful when

- The target population is heterogeneous.
- Subgroups can be defined that successfully group similar units.
- We want a sample to be representative of these subgroups.

Task 2

Suppose a local school contains 1000 children from year 1 to year 5 as outlined in the table below

year	population size	standard deviation
1	250	2
2	150	2
3	300	2
4	100	5
5	200	3

We want to sample 100 students and are interested in finding out which subject students prefer.

Define a suitable stratification and determine the number of units within each stratum.

Task 3

A local office building with 136 employees wishes to gauge job satisfaction. In total, there are four large office spaces with a mixture of senior and junior staff such that

Office	Senior	Junior	Total
1	25	12	37
2	10	20	30
3	3	25	28
4	6	35	41
Total	44	92	136

We have a total budget of £50 excluding all overhead costs with a cost of £1.50 to survey a junior member of staff and £3 to survey a senior member of staff.

1. How could we use stratified sampling to conduct this survey?

2. How many member of staff could we sample?

Cluster sampling

In many practical situations, it is not always possible to access all units within a population. With stratified random sampling, we need to be able to group all units together according to a stratification characteristic and then sample within each stratum. If it is not feasible to access all units then clustering sampling may be used.

The basic principles of clustering sampling are as follows.

1. Divide the population in natural clusters based on some rule. A commonly used rule is geographical area.

2. Treat each cluster as a sampling unit.

3. Sample clusters based on a sampling method. For example, we may use simple random sampling to sample clusters.

4. Collect the necessary information from all sampling units within each sampled cluster.

Definition 4

Cluster sampling

Cluster sampling is a sampling procedure where we sample units within a population using a sampling method where sampling units are clusters.

Definition 5

Cluster

A cluster is a set, or group, of population units.

In order to demonstrate the difference between stratified sampling and cluster sampling. Suppose we wish to collect data from a population arranged in streets. Within each street, households can be distinguished by shape as illustrated below. Sampled units are coloured red.

Figure 1

If we choose to use stratified sampling, we may group households together based on shape. We would have eight strata with households within each stratum of the same shape. We may then use simple random sampling to sample within each stratum based on some pre-determined sample size allocations.

If we choose to use cluster sampling, we may define each street to be a cluster. We would have five clusters with households within each cluster on the same street. We may then use simple random sampling to sample clusters.

Notice that in a stratified sample, we would ideally assume that units within strata are homogeneous and units between strata are heterogeneous. On the other hand, in cluster sampling we would ideally assume that units within clusters are heterogeneous and cluster are homogeneous. The overall success and efficiency of cluster sampling depends on heterogeneity of units within clusters and homogeneity of clusters. Once clusters are defined, any method of sampling can be applied treating each clusters as a sampling unit. We will focus on using simple random sampling to sample clusters.

Advantages of cluster sampling

The most immediate benefit of cluster sampling is that we do not need to access all units within the population. Once we have determined the clustering structure, for example geographical location, then we need to only be concerned with units within the sampled clusters. Therefore, we may prefer cluster sampling when we have no reliable information on all units within a population or if it is expensive to access certain units within a population.

Disadvantages of cluster sampling

Since we do not consider the entire population, clusters may not reflect the true diversity of the population.

Cluster notation

Let's re-define some of the notation introduced previously for stratified sampling. In this case, we need notation that can define clusters and also the number of units within a cluster.

- N = number of clusters.
- M_i = number of units in cluster i.
- $M = \sum_{i=1}^{N} M_i$ = number of units in the population.
- $\bar{M} = \frac{M}{N}$ = the average number of units in clusters.
- n = the number of sampled clusters.
- y_{ij} = the jth observation of variable y in cluster i.
- $\bar{y}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} y_i j$ = the average value of y in the ith cluster.
- $\bar{y}_N = \frac{1}{N} \sum_{i=1}^{N} \bar{y}_i$ = the average value of y across clusters.
- $\bar{y} = \frac{1}{M} \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}$ = the population average value of y.

Constructing clusters

We have two possible options to consider when constructing clusters.

Clusters are of equal size

If we have a population that can be divided into N clusters each of size M then we may sample n clusters using simple random sampling.

Proposition 4

Equally sized cluster mean estimator

If we sample n clusters from N equally sized clusters and $\bar{y}i$ is the sample mean in cluster i, the mean of cluster means

$$\bar{y}_{clust_1} = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_i$$

is an unbiased estimator the population mean $\bar{y}$.

Supplement 2

Proof

$$\begin{aligned} E(\bar{y}_{clust_1}) &= \frac{1}{n}\sum_{i=1}^{n}E(\bar{y}_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\bar{y} \\ &= \bar{y} \end{aligned}$$

since we have used simple random sampling to sample within clusters.

Clusters are of unequal size.

In practical situations, clusters will often be unequal in size. In this case assume we are using simple random sampling without replacements of n clusters from a total of N clusters.

Proposition 5

Weighted cluster mean estimator

If we sample n clusters and $\bar{y}i$ is the sample mean in cluster i, the weighted cluster mean

$$\frac{N}{nM}\sum_{i=1}^{n} M_i\bar{y}_i = \frac{1}{n}\sum_{i=1}^{n}\frac{M_i}{\bar{M}}\bar{y}_i$$

is an unbiased estimator the population mean $\bar{y}$.

Supplement 3

Bias of weighted cluster mean estimator

Let $\bar{y}^*_{\text{clust}} = \frac{N}{nM}\sum_{i=1}^{n} M_i\bar{y}_i$. Then

$$\begin{aligned}
bias(\bar{y}^*_{\text{clust}}) &= E(\bar{y}^*_{\text{clust}}) - \bar{y} \\
&= \frac{N}{M}E\left(\frac{1}{n}\sum_{i=1}^{n} M_i\bar{y}_i\right) - \bar{y} \\
&= \frac{N}{M}E\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{M_i} y_{ij}\right) - \bar{y} \\
&= NE\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{M_i}\sum_{j=1}^{M_i} y_{ij}\right) - \bar{y} \\
&= NE\left(\frac{1}{n}\sum_{i=1}^{n}\bar{y}_i\right) - \bar{y} \\
&= \sum_{i=1}^{N}\bar{y}_i - \bar{y} \\
&= \sum_{i=1}^{N}\frac{1}{M_i}\sum_{j=1}^{M_i} y_{ij} - \bar{y} \\
&= \sum_{i=1}^{N}\frac{1}{M_i}\sum_{j=1}^{M_i} y_{ij} - \bar{y} \\
&= \frac{1}{M}\sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij} - \bar{y} \\
&= 0
\end{aligned}$$

Supplement 4

Variance of weighted cluster mean estimator

If we sample n clusters and $\bar{y}i$ is the sample mean in cluster i, the weighted cluster mean would be

$$\bar{y}^*_{\text{clust}} = \frac{N}{nM}\sum_{i=1}^{n} M_i \bar{y}_i$$

and

$$Var(\bar{y}^*_{\text{clust}}) = \left(\frac{N-n}{Nn\bar{M}}\right)\frac{1}{N-1}\sum_{i=1}^{N}(y_{i\cdot} - \bar{y}^*_{\text{clust}})^2$$

This of course assumes that we know M, the total number of units in the population. Recall that an advantage of cluster sampling is that we do not need to be able to access all units within a population. In practice, we would not be able to estimate M.

Proposition 6

Cluster mean estimator

If we sample n clusters and $\bar{y}i$ is the sample mean in cluster i, the mean of cluster means

$$\bar{y}_{\text{clust}} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_i$$

is a biased estimator the population mean $\bar{y}$.

Supplement 5

Variance of cluster mean estimator

If we sample n clusters and $\bar{y}i$ is the sample mean in cluster i, the mean of cluster means

$$\bar{y}_{\text{clust}} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_i$$

$$Var(\bar{y}_{\text{clust}}) = \left(\frac{N-n}{Nn}\right)\frac{1}{n-1}\sum_{i=1}^{n}(\bar{y}_i - \bar{y}_{\text{clust}})^2.$$

Supplement 6

Bias of mean of cluster means

Let $\bar{y}_{\text{clust}} = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_i$. Then

$$
\begin{aligned}
bias(\bar{y}_{\text{clust}}) &= E(\bar{y}_{\text{clust}}) - \bar{y} \\
&= \frac{1}{N}\sum_{i=1}^{N}\bar{y}_i - \frac{1}{M}\sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij} \\
&\quad \text{since we are using simple random sampling} \\
&= \frac{1}{N}\sum_{i=1}^{N}\bar{y}_i - \frac{1}{M}\sum_{i=1}^{N} M_i\bar{y}_i \\
&= \frac{1}{M}\left[\frac{M}{N}\sum_{i=1}^{N}\bar{y}_i - \sum_{i=1}^{N} M_i\bar{y}_i\right] \\
&= \frac{1}{M}\left[\frac{1}{N}\sum_{i=1}^{N} M_i\sum_{i=1}^{N}\bar{y}_i - \sum_{i=1}^{N} M_i\bar{y}_i\right] \\
&= \frac{-1}{M}\sum_{i=1}^{N}(M_i - \bar{M})(\bar{y}_i - \bar{y}) \\
&\approx Cov(M_i, \bar{y}_i)
\end{aligned}
$$

Example 7

A company wishes to estimate the number of phone calls made to an office in a typical day. In total there are 25 offices each with between 5 to 20 phones with an average of 12 phones per office. They decide to randomly sample 4 offices and calculate the average number of calls on each phone within the 4 offices across two weeks. The results are given below

Cluster	M_i	daily number of calls
1	21	29 32 34 28
		29 34 28 30 27
		29 25 25 22 26
		40 18 34 39

Cluster	M_i	daily number of calls
		27 26 23
2	7	12 11 11 5
		11 8 13
3	16	42 47 53 40
		31 33 41 34
		44 31 33 41
		35 49 35 31
4	5	27 29 27 2
		27

Provide an estimate of the average number of call made to a phone on a typical day.

In this case, we have four cluster of different sizes. We need to first estimate the average number of calls within each cluster.

Cluster	M_i	daily number of calls	$\bar{y}_i$
1	21	29 32 34 28	28.81
		29 34 28 30 27	
		29 25 25 22 26	
		40 18 34 39	
		27 26 23	
2	7	12 11 11 5	10.14
		11 8 13	
3	16	42 47 53 40	38.75
		31 33 41 34	

Cluster	M_i	daily number of calls	$\bar{y}_i$
		44 31 33 41	
		35 49 35 31	
4	5	27 29 27 2	27.6
		27	

Therefore, $N = 25$, $n = 4$, $M = \bar{M}N = 12 \times 25 = 300$.

$$
\begin{aligned}
\hat{\bar{y}} &= \frac{N}{nM} \sum_{i=1}^{n} M_i \bar{y}_i \\
&= \frac{25}{4 \times 300} \sum_{i=1}^{4} M_i \bar{y}_i \\
&= \frac{25}{1200} [21 \times 28.81 + 7 \times 10.14 + 16 \times 38.75 + 5 \times 27.6] \\
&= \frac{25}{1200} [1433.99] \\
&= 29.87 \\
&\approx 30 \text{ calls}
\end{aligned}
$$

Using the biased estimator $\frac{1}{n} \sum_{i=1}^{n} \bar{y}_i$,

$$
\begin{aligned}
\hat{\bar{y}} &= \frac{1}{n} \sum_{i=1}^{n} \bar{y}_i \\
&= \frac{1}{4} \sum_{i=1}^{4} \bar{y}_i \\
&= \frac{1}{4} [28.81 + 10.14 + 38.75 + 27.6] \\
&= \frac{1}{4} [105.3] \\
&= 26.32 \\
&\approx 26 \text{ calls}
\end{aligned}
$$

Task 4

A train company wishes to estimate the number of trains that arrive late to their terminal station in a typical day. In the city of interest, there are 50 terminal stations with an average

of 20 trains terminating at each station during an average day. We decide to monitor 5 terminal stations over the course of one day. For each train that arrives at the station, we note down the difference between the arrival and scheduled times. The data are given in the table below

Cluster	M_i	**Arrival time - Scheduled time (minutes)**
1	15	-7.95 -5.22 5.34 0.25
		1.03 1.71 -3.32 0.22
		5.18 -4.08 0.95 -5.79 -5.56
		2.82 -0.73
2	8	-0.38 -1.01 -0.58 1.00
		-0.09 -0.06 1.48 0.32
3	6	2.23 -1.94 0.06 1.71 -1.48 -0.61
4	32	1.79 0.59 0.70 -1.76
		0.94 1.22 0.41 0.42 -0.86
		1.30 1.29 -0.16 1.53 1.22
		0.53 -0.72 0.92 -0.89 0
		-0.76 -0.65 0.03 -1.51
		-0.51 1.16 -0.61 -0.01 -0.07
		-1.08 1.36 1.95 -1.43
5	28	4.70 -0.06 15.05 3.60 11.25
		-4.00 1.16 16.45 -3.04 -4.28 -17.37
		-19.02 26.10 0.54 -8.55 -9.07 -10.77 11.40
		1.25 -10.36 11.25 13.26 -0.77 0.49
		0.82 3.77 9.56 25.09

Estimate the average difference between arrival time and scheduled time.

Two-Stage Sampling

Notice in cluster sampling that once we have selected a cluster, all units within that cluster are sampled. On the other hand, with stratified random sampling, we sample within strata to form our sample. In two-staged sampling, we divide the population into cluster and then sample within each cluster.

Definition 6

Two-stage sampling

Two-stage sampling is a sampling procedure where units within a population are sampled using a sampling method (stage 1) and then sub-sampling occurs within each sampled unit (stage 2). It is useful to consider units at stage 1 to be clusters. Therefore at stage 2, we sample within each cluster.

In order to demonstrate the difference between stratified sampling, cluster sampling and two-stage sampling we may add an extra layer to our street example. Suppose we wish to collect data from a population arranged in streets. Within each street, households can be distinguished by shape as illustrated below. In the second layer we sample within each selected cluster. Sampled units are coloured red.

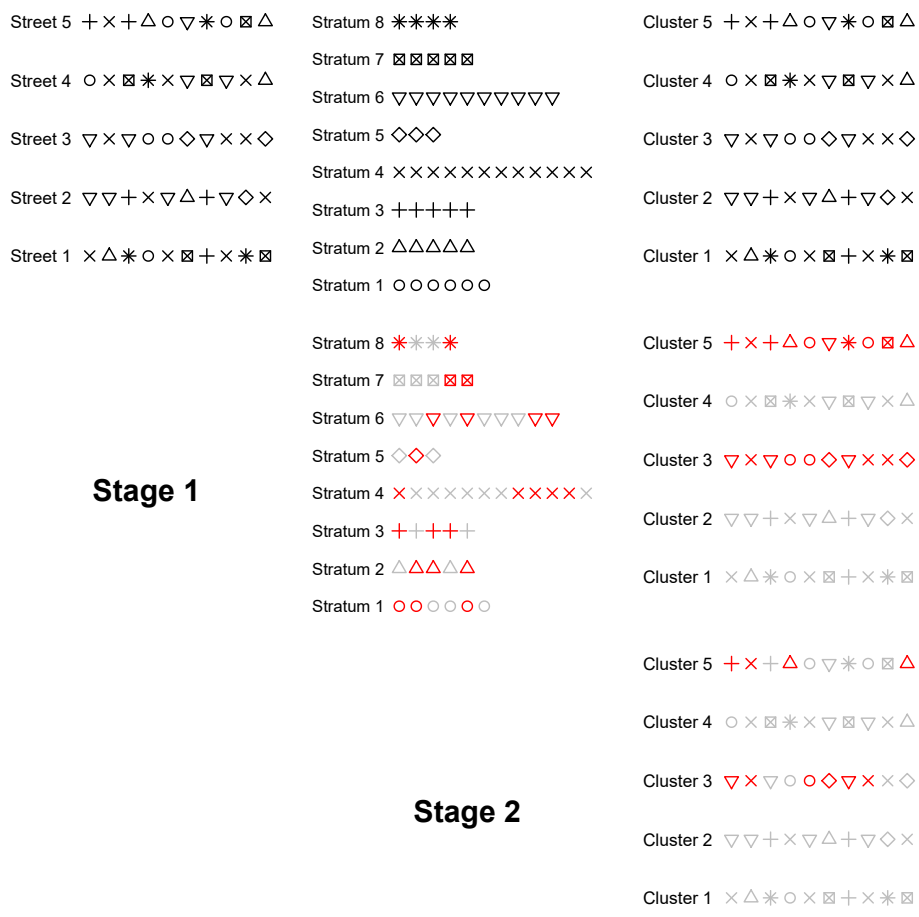


Figure 2

The clusters which form the units of sampling at the first stage are called the **first stage units** and the sampled units within clusters are called the **second stage units** or **sub-units**.

We will consider the case of two-stage sampling with simple random sampling without replacement is used at the first stage to sample clusters and simple random sampling without replacement used to sub-sample within clusters at the second stage. We will further assume that clusters are of unequal size.

Relation to stratified and cluster sampling

Both stratified and cluster sampling are special cases of two-stage sampling.

If we choose to sample all units within each sampled cluster then this is equivalent to cluster sampling.

Suppose we choose to sample all clusters (such that $n = N$ using the notation defined in the cluster sampling section) and then sub-sample within each cluster then this is equivalent to stratified sampling if the clusters were defined through a variable suitable for stratified sampling.

Advantages of two-stage sampling

Recall that we are trying to balance estimator precision and the cost of sampling. If units within clusters, defined as stage one, are homogeneous that sub-sampling fewer units within each cluster will save time and reduce cost without loss in precision.

Notation

Let's add to the notation introduced previously for cluster sampling. In this case, we need notation that can define clusters and also the number of units within a cluster.

- N = number of clusters.
- M_i = number of units in cluster i.
- $M = \sum_{i=1}^{N} M_i$ = number of units in the population.
- $\bar{M} = \frac{M}{N}$ = the average number of units in clusters.
- n = the number of sampled clusters at stage one.
- m_i = the number of units sampled at stage two from the ith cluster sampled at stage one.
- $m = \sum_{i=1}^{n} m_i$ = the total number of units sampled at stage two.
- y_{ij} = the jth observation of variable y in cluster i.
- $\bar{y}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} y_{ij}$ = the average value of y in the ith cluster.
- $\bar{y}_{im_i} = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij}$ = the average value of second stage sampled units within the ith first stage sampled cluster.
- $\bar{y}_n = \frac{1}{n} \sum_{i=1}^{n} \bar{y}_{im_i}$ = average value of the first stage sampled clusters.
- $\bar{y}_N = \frac{1}{N} \sum_{i=1}^{N} \bar{y}_i$ = the average value of y across clusters.

Two-stage expectation

We have seen already with cluster sampling that we can take the expected value of units within clusters at stage 2 and then average over all clusters sampled at stage 1. In other words for any parameter θ

$$E(\theta) = E_1(E_2(\hat{\theta})).$$

More precisely, if we estimate $\bar{y}$

$$E(\hat{\bar{y}}) = E_1(E_2(\hat{\bar{y}})).$$

That is, we average samples units within each cluster and then average across our cluster averages.

Following from the results already obtained in the previous section, we will present unbiased estimators of the population mean value firstly assuming we sample an equal number of units from equally sized

clusters and secondly assuming we sample an unequal number of units from unequally sized clusters.

Proposition 7

Cluster mean estimator

If we sample m units from n equally sized clusters and $\bar{y}_{im_i}$ is the sample mean in cluster i, the weighted cluster mean

$$= \frac{1}{n}\sum_{i=1}^{n}\bar{y}_{im_i}$$

is an unbiased estimator the population mean $\bar{y}$.

Proposition 8

Weighted cluster mean estimator

If we sample m_i units from n unequally sized clusters and $\bar{y}_{im_i}$ is the sample mean in cluster i, the weighted cluster mean

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{M_i}{\bar{M}}\bar{y}_{im_i}$$

is an unbiased estimator the population mean $\bar{y}$.

The details of both these estimators have been omitted from these notes. You can find more details **in this book chapter**.

Example 8

Let's revisit the company call's example from the previous section. Instead we will use two-stage sampling to estimate the number of phone calls made to an office in a typical day.

A company wishes to estimate the number of phone calls made to an office in a typical day. In total there are 25 offices each with between 5 to 20 phones with an average of 12 phones per office. They decide to randomly sample 4 offices and sub-sample within each office proportional to the number of phone's in the respective office. The results are given below

Cluster	M_i	m_i	daily number of calls
1	19	12	32 28 29 34
			29 25 28 27
			23 26 18 34
2	7	4	11 13 8 5
3	16	10	42 47 53 40 41
			31 33 44 35 49
4	5	3	27 29 2

Provide an estimate of the average number of calls made to a phone on a typical day.

In this case, we have four clusters of different sizes. We need to first estimate the average number of calls within each cluster.

Cluster	M_i	m_i	daily number of calls	y_{im_i}
1	19	12	32 28 29 34	28.58
			29 25 28 27	
			23 26 18 34	
2	7	4	11 13 8 5	9.25
3	16	10	42 47 53 40 41	41.5
			31 33 44 35 49	
4	5	3	27 29 2	19.33

Therefore, $N = 25$, $n = 4$,
$\bar{M} = 12, M_1 = 19, M_2 = 7, M_3 = 16, M_4 = 5, m_1 = 12, m_2 = 4, m_3 = 10$ and $m_4 = 3$.

$$
\begin{aligned}
\hat{\bar{y}} &= \frac{1}{n}\sum_{i=1}^{n}\frac{M_i}{\bar{M}}\bar{y}_{im_i} \\
&= \frac{1}{4}\sum_{i=1}^{4}\frac{M_i}{12}\bar{y}_{im_i} \\
&= \frac{1}{4\times 12}[19\times 28.58 + 7\times 9.25 + 16\times 41.5 + 5\times 19.33] \\
&= 28.50 \\
&\approx 29 \text{ calls}
\end{aligned}
$$

Example 9

The Academic Performance Index (API) is a measurement of academic performance in schools in California, USA. It ranges from a low score of 200 to a high score of 1000.

We have data from 6194 schools detailing API in the year 2000. Each school belongs to one of 757 districts and fall into three categories namely elementary, middle or high school.

Each district contains a different number of schools. In fact, the majority of districts contain 1 - 3 schools. The largest district contains 552 schools.

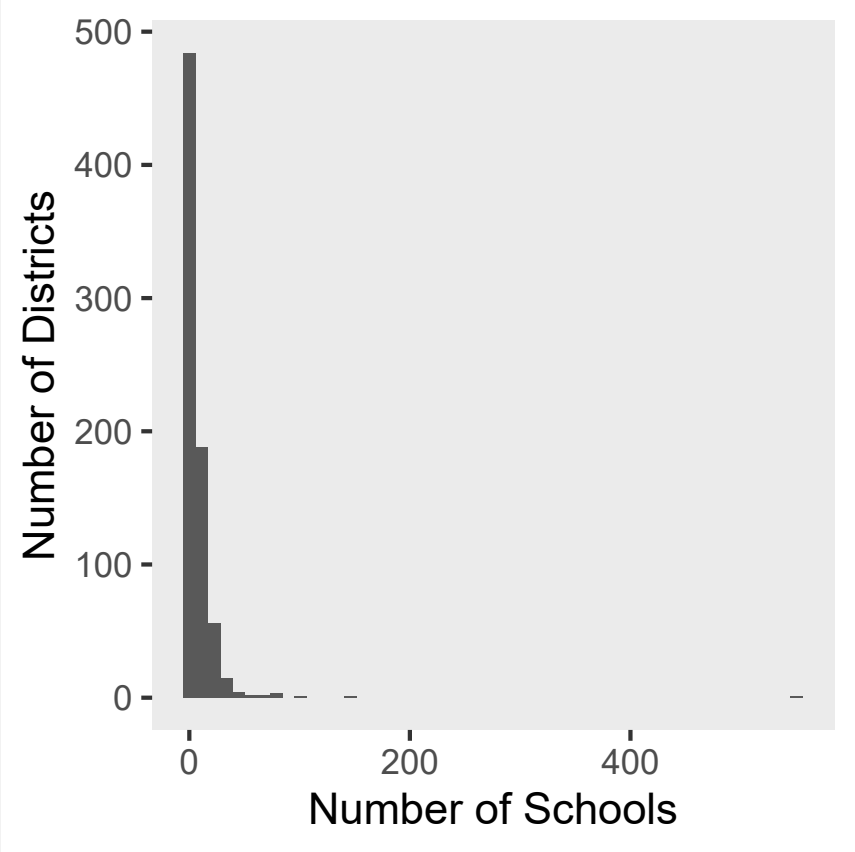


Figure 3

In addition, there is an unequal number of elementary, middle and high schools with most of the schools elementary.

Type	Number of Schools
Elementary	4421
Middle School	755
High School	1018

We are interested in the average API score across the 6194 schools. We will estimate the average score using four methods

1. Proportional stratified sampling using type of school (elementary, middle or high school) as our grouping variable. 100 elementary schools, 50 middle schools and 50 high schools were randomly sampled without replacement within the respective stratum.

2. Cluster sampling using district to define clusters. 15 districts were randomly sampled without replacement.

3. Two-stage clustering using districts at the first stage and then sampling schools within each district at the second stage. 40 districts were randomly sampled without replacement. Within each district, between 1 and 72 were randomly sampled without replacement.

4. Simple random sampling without replacement sampling 200 schools.

Here is the R code used in this example

```
library(survey)
data(api)
mean(apipop$api00)
sum(apipop$enroll, na.rm=TRUE)

#stratified sample
dstrat<-svydesign(id=~1,strata=~stype, weights=~pw,
data=apistrat, fpc=~fpc)
summary(dstrat)
svymean(~api00, dstrat)
svytotal(~enroll, dstrat, na.rm=TRUE)
```

```
# one-stage cluster sample
dclus1<-svydesign(id=~dnum, weights=~pw, data=apiclus1,
fpc=~fpc)
summary(dclus1)
svymean(~api00, dclus1)
svytotal(~enroll, dclus1, na.rm=TRUE)

# two-stage cluster sample
dclus2<-svydesign(id=~dnum+snum, fpc=~fpc1+fpc2, data=apiclus2)
summary(dclus2)
svymean(~api00, dclus2)
svytotal(~enroll, dclus2, na.rm=TRUE)
```

Let's compare estimates based on these sampling approaches

Approach	Estimate	Standard Error
1	662.29	9.41
2	644.17	23.54
3	670.81	30.10
4	676.32	8.34
True API	664.71	

In order to maximise precision in stratified random sampling, we need units within strata to have similar values.

In order to maximise precision in cluster sample, we need heterogeneous clusters and representative of the population. In this example, some districts had only one school and so this illustrates how reduced heterogeneity within clusters can result in decreased precision (or inflated standard errors).

Or course, in estimating our population parameter, we need to know the size of each cluster. The example below illustrates a method of two-stage clustering in the case where the population size of each cluster is unknown.

Example 10

A company wishes to estimate the number of man-hours lost due to employee sickness. In total the company has 500 office buildings with various numbers of employees in each one.

The decide to firstly randomly sample (without replacement) 50 offices.

Office	Number of employees	Office	Number of employees
1	31	26	380
2	92	27	40
3	15	28	218
4	134	29	313
5	309	30	71
6	63	31	114
7	342	32	343
8	157	33	220
9	164	34	325
10	230	35	245
11	298	36	3
12	198	37	118
13	208	38	186
14	86	39	43
15	104	40	260
16	383	41	240
17	315	42	99

Office	Number of employees	Office	Number of employees
18	368	43	98
19	205	44	381
20	273	45	95
21	254	46	231
22	101	47	34
23	389	48	249
24	252	49	122
25	301	50	378

Using this sample, the company decide to estimate the average number of employees per office. In total, across these 50 offices, there are 10,078 employees with an average of 201.56 employees per office.

Sampling 10,078 people within these offices came at too high a cost and so the company decided to randomly sample without replacement 20 offices from this list of 50.

Office	Number of employees	man-hours lost in one year
2	92	7.58
24	252	136.11
29	313	295.25
35	245	166.21
15	104	85.73
18	368	253.08
48	249	133.06
43	98	53.35

Office	Number of employees	man-hours lost in one year
9	164	99.40
28	218	100.64
4	134	92.96
26	380	259.81
14	86	42.54
19	205	116.31
41	240	45.50
38	186	158.28
37	118	97.91
39	43	134.05
8	157	176.50
12	198	37.60

Based on these 20 offices, the average number of man-hours lost in one year is 124.60 hours.

However, notice the relationship between the number of employees in an office and the number of hours lost.

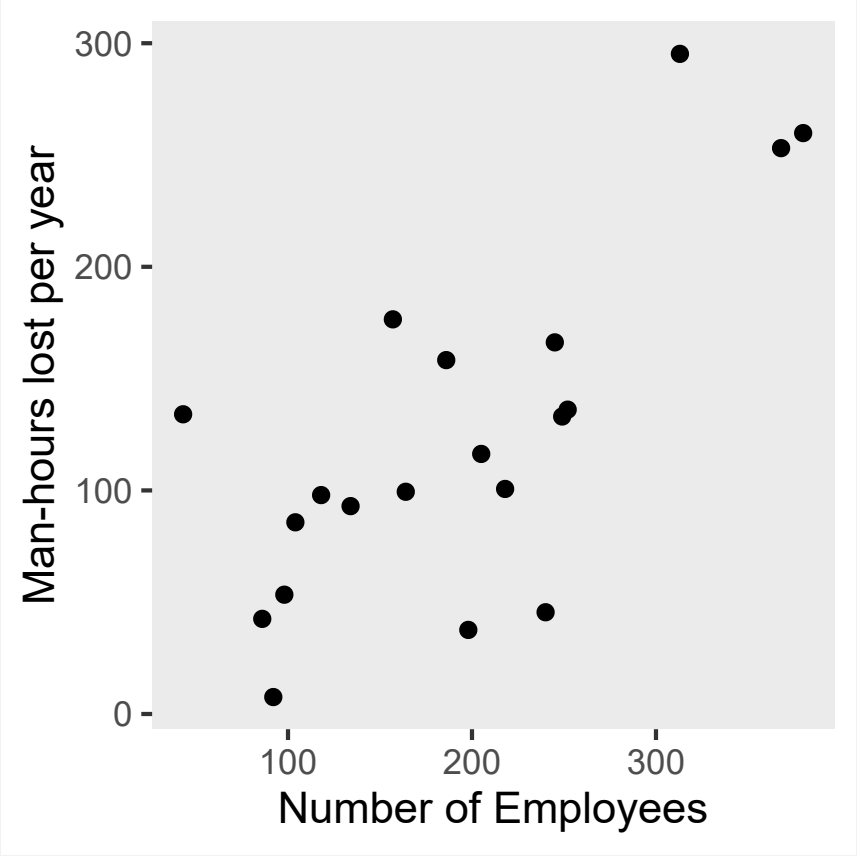


Figure 4

Therefore, our average value of 124.60 hours depends on the size of each office sampled. We can adjust for the size of each office in our estimate using our initial estimator of the average number of employees per office (201.56) and also the average number of employees based on our sample of 20 offices (192.5).

Let's first introduce some notation.

$$
\begin{aligned}
y &= \text{the total number of man-hours lost per year} \\
\bar{y} &= \text{the average number of man-hours lost per year} \\
\hat{\bar{y}} &= \text{the average number of man-hours lost per year estimated from 20 offices} \\
\bar{x}_1 &= \text{the average number of employees based on 50 offices} \\
\bar{x}_2 &= \text{the average number of employees based on 20 offices}
\end{aligned}
$$

$$
\begin{aligned}
\hat{\bar{y}} &= 124.69 \\
\bar{x}_1 &= 201.56 \\
\bar{x}_2 &= 192.5
\end{aligned}
$$

$$\begin{aligned}\text{average man-hours lost per year} &= \frac{\hat{\bar{y}}}{\bar{x}_2}\bar{x}_1 \\ &= \frac{124.69}{192.5}201.56 \\ &= 130.56\end{aligned}$$

Learning outcomes for week 11

Now you have reached the end of week 11 you should be able to:

- Describe stratified sampling;
- Describe the advantages and disadvantages of stratified sampling;
- Estimate a population parameter using a stratified sample;
- Describe cluster sampling;
- Describe the advantages and disadvantages of cluster sampling;
- Estimate a population parameter using a cluster sample;
- Describe two-stage sampling;
- Describe the advantages and disadvantages of two-stage sampling;
- Estimate a population parameter using a two-stage sample;

Summary of results from week 11

Stratified sampling

Partitioning a population into non-overlapping groups and sampling within each group is known as stratified sampling. Each group is called a stratum.

Unbiased estimator or population mean using stratified sample

$$\bar{y}_{\text{strat}} = \sum_{i=1}^{k} W_i \bar{y}_i$$

is an unbiased estimator of population mean $\bar{y}$ where $\bar{y}_i$ is the sample mean in stratum i and W_i is the proportion of the population in stratum i.

Cluster sampling

Sampling units within a population using a sampling method where sampling units are clusters.

Weighted cluster mean estimator

If we sample n clusters and $\bar{y}i$ is the sample mean in cluster i, the weighted cluster mean

$$\frac{1}{n}\sum_{i=1}^{n}\frac{M_i}{\bar{M}}\bar{y}_i$$

is an unbiased estimator the population mean $\bar{y}$ where M_i is the total number of units in clusters i and $\bar{M}$ the average number of units across clusters.

Two-stage sampling

Sampling units within a population using a sampling method where sampling units are clusters and then sub-sampling within each selected cluster.

Weighted cluster mean estimator

If we sample m_i units from n unequally sized clusters and $\bar{y}_{im_i}$ is the sample mean in cluster i, the weighted cluster mean

$$=\frac{1}{n}\sum_{i=1}^{n}\frac{M_i}{\bar{M}}\bar{y}_{im_i}$$

is an unbiased estimator the population mean $\bar{y}$ where M_i is the total number of units in clusters i and $\bar{M}$ the average number of units across clusters.

Answer 1

The total population size is $1000+500+1500+750=3750$.

Stratum	1	2	3	4
Population size	1000	500	1500	750

Stratum	1	2	3	4
W	$\frac{4}{15}$	$\frac{2}{15}$	$\frac{6}{15}$	$\frac{3}{15}$
Sample size	53	27	80	40

Answer 2

Given only the information given, then we could use year as a stratification

stratum	1	2	3	4	5
population size	250	150	300	100	200
standard deviation	$\frac{1}{4}$	$\frac{15}{100}$	$\frac{3}{10}$	$\frac{1}{10}$	$\frac{1}{5}$
W	2	2	2	5	3
equal allocation	20	20	20	20	20
proportional allocation	25	15	30	10	20
Neyman's allocation	20	12	24	20	24

Given the large amount of variability in year 4 in comparison to the other years, Neyman's sampling may be the most appropriate.

Answer 3

In order to perform stratified sampling, we could have two strata containing senior and junior staff, we could have 4 strata with one for each office or we could have eight strata with one for each combination of office and staff level. We need to worry if job satisfactions may be determined by staff level and/or office space. Notice that it is more expensive to sample a senior member of staff.

Let n be the total sample size and let's use proportional sampling with

Office	Senior	Junior	Total
1	0.18	0.09	37
2	0.07	0.15	30
3	0.02	0.18	28
4	0.04	0.27	41
Total	44	92	136

$$
\begin{aligned}
\text{Cost} &= \sum_{i=1}^{8} c_i n_i \\
50 &= 3[0.18n + 0.07n + 0.02n + 0.04n] + 1.5[0.09n + 0.15n + 0.18n + 0.27n] \\
&= 0.93n + 1.035n \\
&= 1.965n \\
\\
n &= 25
\end{aligned}
$$

Equating to sample sizes

Office	Senior	Junior	Total
1	4	2	6
2	1	4	5
3	0	4	4
4	3	7	10
Total	8	17	25

Can you derive a better sampling solution?

Answer 4

Using the unbiased estimator $\frac{1}{n\bar{M}} \sum_{i=1}^{n} M_i \bar{y}_i$ with $n = 5$ and $\bar{M} = 20$

We can estimate the average time difference for each station/cluster separately.

Cluster	M_i	**Arrival time - Scheduled time (minutes)**	
1	15	-7.95 -5.22 5.34 0.25	-1.01
		1.03 1.71 -3.32 0.22	
		5.18 -4.08 0.95 -5.79 -5.56	
		2.82 -0.73	
2	8	-0.38 -1.01 -0.58 1.00	0.085
		-0.09 -0.06 1.48 0.32	
3	6	2.23 -1.94 0.06 1.71 -1.48 -0.61	-0.005
4	32	1.79 0.59 0.70 -1.76	0.20
		0.94 1.22 0.41 0.42 -0.86	
		1.30 1.29 -0.16 1.53 1.22	
		0.53 -0.72 0.92 -0.89 0	
		-0.76 -0.65 0.03 -1.51	
		-0.51 1.16 -0.61 -0.01 -0.07	
		-1.08 1.36 1.95 -1.43	
5	28	4.70 -0.06 15.05 3.60 11.25	2.45
		-4.00 1.16 16.45 -3.04 -4.28 -17.37	
		-19.02 26.10 0.54 -8.55 -9.07 -10.77 11.40	

Cluster	M_i	Arrival time - Scheduled time (minutes)	
		1.25 -10.36 11.25 13.26 -0.77 0.49	
		0.82 3.77 9.56 25.09	

$$
\begin{aligned}
\hat{\bar{y}} &= \frac{1}{5 \times 20} \sum_{i=1}^{5} M_i \bar{y}_i \\
&= \frac{1}{5 \times 20} [15 \times -1.01 + 8 \times 0.085 + 6 \times -0.005 + 32 \times 0.20 + 28 \times 2.45] \\
&= \frac{1}{100} [60.5] \\
&= 0.605 \\
&\approx 36 \text{ seconds}
\end{aligned}
$$

Therefore, the city trains on average arrive just over half an minute late and so the city trains on average have impeccable timing!