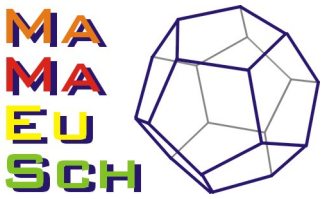
MaMaEuSch

Management **Ma**thematics for
European **Sch**ools
http://www.mathematik.uni-kl.de/~ mamaeusch



Population and sample. Sampling techniques

Paula Lagares Barreiro*
Justo Puerto Albandoz*

MaMaEuSch†
Management Mathematics for European Schools
94342 - CP - 1 - 2001 - 1 - DE - COMENIUS - C21

*University of Seville

†This project has been carried out with the partial support of the European Community in the framework of the Sokrates programme. The content does not necessarily reflect the position of the European Community, nor does it involve any responsibility on the part of the European Community.

Contents

Chapter 1

Population and sample. Sampling techniques

Let us extend in this chapter what we have already presented in the beginning of Descriptive Statistics, including now the definition of some sampling techniques and concepts in order to be able to decide which is the appropriate sampling technique for each situation.

Let us imaging, for instance, that your class has been chosen as a sample of a population. The study that is going to be made can be about different themes, for example:

1. The opinion about the possibility of organizing alternative activities in your city and a proposal of the activities that can be made.
2. A poll about the opinion on the different politic leaders.
3. The opinion about about the possible choices for a end-of year-trip with the students of your class.

Do you think that your class would be a good sample for any of these situations? The answer is that, for instance, for the second situation, the students of a class are not an appropriate sample. For the first situation, we may think that the students of a class can give us interesting information, though maybe the sample can be too "small" and we could have a lack of information (boys and girls of other ages, living in different quarters,...), while for the third situation, the sample can be very useful. Therefore, it is very important the choice of an appropriate sampling technique which assures us that we are choosing a "good" sample for the study we want to make.

1.1 Reasons to use sampling. Previous considerations

Let us imagine that we are going to make studies to get the following information:

- The percentage of Spanish population that has access to internet.
- The average lasting of a concrete trade of batteries.

For the first case, the population you would have to ask to is bigger than 40 million people. It is obvious that making an interview to more than 40 million people requires a big effort in many fields. First of all, there is a big need of time, and second, of money, because it is necessary to employ many people to make the interviews, pay their trips to let them go to every village, etc. Moreover, there is an additional difficulty: it is complicated to get to each and every Spaniard, because when we make the interviews, there will be people in hospitals, in a trip to a foreign country, etc. In this situation, for economic reasons, it will be convenient to interview a certain part of the population, a sample, chosen in an appropriate way so that we can obtain later conclusions for the whole population.

In the second situation we have a different difficulty. If we want to know the lasting of a certain battery, we have to use it until it is over. Therefore, somehow we "destroy" this element of the population. If we would have to try each and every battery of the population, we would keep none of them. Thus, what we should do in this situation is also to choose an appropriate sample and then we could take the appropriate general conclusions.

Due to the reasons we have just mentioned, it is convenient in many instances to use samples. But if we want to get really good conclusions from them, we need to assure that we make a right choice of our samples. For instance, for the case of the internet access in Spain, if we choose 10 people out of the 40 million of inhabitants, this is clearly not enough, it is not a representative sample. I will also not be representative if we choose 100 people from Madrid, or choosing all your friends and your family. There are some topics which should be clearly defined once we want to sample:

1. The selection method for the elements of the population (sampling method to be used).
2. Sample size.
3. Reliability degree of the conclusions that we can obtain, this is, an estimation of the error that we are going to have (in terms of probability).

As we have just said, a non appropriate selection of the elements of the sample can cause further errors once we want to estimate the corresponding parameters in the population. But we can find some more different types of errors: the interviewer can be partial, this is, he can promote some answers more than others. It can also happen that the person we are going to interview does not want to answer certain questions (or cannot answer). We classify all these possible errors in the following way:

1. **Selection error**: if any of the elements of the population has a higher probability of being selected than the rest. Let us imagine that we want to measure how satisfied the clients of a gymnasium are, and for that, we are going to interview some of them from 10 to 12 in the morning. This means that the people who go to the gymnasium in the afternoon will not be represented, and then the sample will not be representative of all the clients. A way to avoid this kind of errors is choosing the sample so that all the clients have the same probability of being selected.
2. **Non-answer error**: it is also possible that some of the elements of the population do not want or cannot answer certain questions. Or it can also happen, when we have a questionnaire including personal questions, that some of the members of the population do not answer honestly. This errors are generally very complicated to avoid, but in case that we want to

check honesty in answers, we can include some questions (filter questions) to detect if the answers are honest.

After what we have seen until now, we can say that we have a biased sample when it is not representative for the population.

1.2 Sampling techniques

We have already stressed the importance of a right choice for the elements of the sample so as to make it representative of our population but, how can we classify the different ways of choosing a sample? we can say that there are three types of sampling:

1. Probability sampling: it is the one in which each sample has the same probability of being chosen.
2. Purposive sampling: it is the one in which the person who is selecting the sample is who tries to make the sample representative, depending on his opinion or purpose, thus being the representation subjective.
3. No-rule sampling: we take a sample without any rule, being the sample representative if the population is homogeneous and we have no selection bias.

We will always make probability sampling, because in case we choose the appropriate technique, it assures us that the sample is representative and we can estimate the errors for the sampling. There are different types of probability sampling:

- Random sampling with and without replacement.
- Stratified sampling.
- Cluster sampling.
- Systematic sampling.
- Other types of sampling techniques.

Let us imagine now that we have already selected a sample. From a high school with 560 students, we have selected a sample of 28 students to know if they have internet connection at home. But, what does it mean to select 28 out of 560? Which proportion of the population are we selecting? And when we want to have conclusions about the population, how many of the students of the population does each one of the sample elements represent?

To calculate the proportion of students that we are interviewing, we divide the sample size by the population size, this is: $28/560 = 0.05$, and this means that we make the poll to 5% of the population.

Now we are going to calculate how many students represents each one of the elements of the sample. We make the other quotient, now we divide the number of elements of the population by the number of elements of the sample: $560/28 = 20$, which would mean that each of the students of the sample represents 20 students of the high school.

The two concepts that we have just presented have the following formal definition:

1. **Elevation factor:** it is the quotient between the size of the population and the size of the sample, $\frac{N}{n}$. It represents the number of elements existing in the population for each element of the sample.

2. **Sampling factor:** it is the quotient between the size of the sample and the size of the population, $\frac{n}{N}$. If this quotient is multiplied by 100, we get the percentage of the population represented in the sample.

1.3 Random sampling with and without replacement

We have already mentioned that if we want to sample in such a way that the sample we get is representative, we should choose a probabilistic sampling technique. How will you do to select 28 students out of 560 in a high school to get that all of them have the same probability of being in the sample? The easiest thing would be to make a draw to choose 28 of them, this is, to choose them randomly, so that they all have the same possibility of belonging to the sample.

This selection process corresponds to a random sampling. We will say that we are making random sampling when the process, through which we choose the sample, guarantees that all the possible samples that we can take from the population have the same probability of being chosen, this is, all the elements of the population have the same probability of being chosen to belong to the sample.

When a certain element is selected and we have measured the variables needed in a certain study and it can be selected again, we say that we make sampling with replacement. This sampling technique is usually called simple random sampling.

In the case that the element cannot be selected again after being selected once, we say that we have obtained the sample through a random sampling without replacement.

In our example, when we are going to select the sample out of the 560 students of the high school, if we are going to ask about the fact that they have internet connection at home or not, it is not interesting for us to ask twice the same person, so once we choose an element of the population we don't want to choose it again. So we would make random sampling without replacement.

Though these two methods are different, when the size of the population is infinite, or it is so big that we can consider that it is infinite, both methods will lead us to similar conclusions. Nevertheless, if the sampling fraction n/N is greater than 0.1 (we sample more than 10% of the population) the difference between the conclusions we get may be important.

When we ask in our example if the students have internet connection at home or not, we are interested not only in the number of students having the connection but also in the proportion that it represents in the high school. These two values and the average in some other cases (for instance, when we ask about the height of the students), are the parameters calculated more often and the ones we usually want to estimate. In the case of random sampling, with and without replacement, these estimators have the following expressions:

Total:

$$\widehat{X} = N \sum_{i=1}^{n} \frac{X_i}{n}.$$

Average:

$$\widehat{\overline{X}} = \sum_{i=1}^{n} \frac{X_i}{n}.$$

Proportion:

$$\widehat{P} = \sum_{i=1}^{n} \frac{P_i}{n}.$$

The proportion would be the average of a variable that only can be zero or one. In the expressions above:

X_i is the value of the variable we are studying.

N is the size of the population.

n is the size of the sample.

P_i is a variable that takes values 0 or 1.

The estimation of the error for these estimators would be:

Total:

For sampling with replacement:

$$\widehat{V}(\widehat{X}) = N^2 \frac{S^2}{n}.$$

For sampling without replacement:

$$\widehat{V}(\widehat{X}) = N^2 (1 - \frac{n}{N}) \frac{S^2}{n}.$$

Average:

For sampling with replacement:

$$\widehat{V}(\widehat{\overline{X}}) = \frac{S^2}{n}.$$

For sampling without replacement:

$$\widehat{V}(\widehat{\overline{X}}) = (1 - \frac{n}{N}) \frac{S^2}{n}.$$

Proportion:

For sampling with replacement:

$$\widehat{V}(\widehat{P}) = \frac{\widehat{P}\widehat{Q}}{n-1}.$$

For sampling without replacement:

$$\widehat{V}(\widehat{P}) = (1 - \frac{n}{N}) \frac{\widehat{P}\widehat{Q}}{n-1}.$$

1.4 Stratified sampling

Let us imagine now that we want to make a poll to know what do people in your city do in their spare time. We all know that ancient people do not have the same activities than middle-age people, as your parents, for instance. We would be interested in getting that all the information, that we already know, can help us to find a more representative sample. In fact, we are interested in getting that all these groups are represented in our sample. These groups that have been defined (in our

example, by ages) we will call them strata. What we will do now is to divide our sample in such a way that we have elements of all the strata. Let us define the way we sample in this case.

Let us consider that we have our population of size N divided into k subpopulations of sizes $N_1, N_2, \ldots, N_k$. These subpopulations are disjoint and verify that $N_1 + N_2 + \cdots + N_k = N$. Each of the subpopulations is called stratus. If we want to have a sample of n elements of the initial population, we select a sample of size n_i so that $n_1 + n_2 + \cdots + n_k = n$.

Which advantages and disadvantages presents stratified sampling? We present them now:

Advantages:

- We can have more precise information inside the subpopulations about the variables we are studying.
- We can raise precision of the estimators of the variables of the whole population.

Disadvantages:

- The choice of the size of the samples inside each stratus to let the sample size be n.
- It may be difficult in some populations to divide into strata.

As a general thing, stratified sampling provides better results than the random sampling when the strata are more different among them and more homogeneous internally.

We can consider 3 methods to distribute the size of the sample among the strata.

1. Proportionally to the size of each stratus, i.e., if we take the j-th stratus with size N_j, and then a sample of this stratus will have size $n \cdot (N_j/N)$, being N the size of the population and n the size of the sample.
2. Proportionally to the variability of the parameter we are considering in each stratus. For instance, if we know that the variance for the height in the male students is 15 cm and for the female students is 5 cm, the proportion of the male students to female students is 3 to 1 and the sample should keep that proportion.
3. We assign the same size to each stratus. As a consequence we promote the smaller strata and the contrary happens with the bigger ones in terms of precision.

For the case of stratified sampling, the main estimators are the following:

Total:

$$\widehat{X} = \sum_{h=1}^{k} N_h \overline{X}_h.$$

Average:

$$\widehat{\overline{X}} = \sum_{h=1}^{k} w_h \overline{X}_h = \sum_{h=1}^{k} \frac{N_h}{N} \overline{x}_h.$$

Proportion:

$$\widehat{P} = \sum_{h=1}^{k} w_h \widehat{P}_h,$$

where
$\overline{X}_h$ is the sample average for variable X in stratus h.
N_h is the size of stratus h.
N is the size of the population.
n_h is the sample size in stratus h.
n is the sample size.
$\widehat{P}_h$ is the sample proportion of the variable in stratus h,
and the estimation for the error we make when we estimate the population parameters is:
Total:

$$\widehat{V}(\widehat{X}) = \sum_{h=1}^{k} N_h^2 (1 - f_h) \frac{\widehat{S}_h^2}{n_h},$$

with

$$f_h = \frac{n_h}{N_h} \qquad \text{y} \qquad \widehat{S}_h^2 = \frac{n_h}{n_h - 1} \left[\frac{1}{n_h} \sum_{i=1}^{n_h} X_{hi}^2 - \overline{x}_h \right].$$

Average:

$$\widehat{V}(\widehat{\overline{X}}) = \sum_{h=1}^{k} w_h^2 (1 - f_h) \frac{\widehat{S}_h^2}{n_h},$$

where w_h, f_h y S_h^2 are the same as before.
Proportion:

$$\widehat{V}(\widehat{P}) = \sum_{h=1}^{k} w_h^2 (1 - f_h) \frac{\widehat{P}_h \widehat{Q}_h}{n_h - 1},$$

where $\widehat{Q}_h = 1 - \widehat{P}_h$.

1.5 Cluster sampling

We think now about making a poll to study the average height of the students of high schools in your city. Instead of sampling among each of the students of the city, we could consider the possibility of choosing some quarters because referring to the height, quarters are like "small populations" that we can compare to the city. In this case, can we simplify the choice of the sample so that we choose quarters without loosing accuracy? The answer is that in this case, we could choose quarters and analyze the height without loosing accuracy. Let us present the sampling method which allows that.

In cluster sampling, population is divided into units or groups, called strata (usually they are units or areas in which the population has been divided in), which should be as representative as

possible for the population, i.e., they should represent the heterogeneity of the population we are studying and they should be homogeneous among them.

The reason to make this sampling is that sometimes it is too expensive to make a complete list of all the elements of the population that we want to study, or that when we finish making the list it may have no sense to make the study.

The main disadvantage that we may have is that if the clusters are not homogeneous among them, the final sample may not be representative of the population.

If we suppose that the clusters are as heterogeneous as the population, referring to the variable we are considering, and that the clusters are homogeneous among them, then to get a sample we only have to choose some clusters. We say that we make cluster sampling in one stage.

This sampling method has the advantage that it simplifies the collecting of the sample information.

Let us see now the expressions of the estimators for this sampling technique:

Total:

$$\widehat{X} = M\frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} M_i}.$$

Average:

$$\widehat{\overline{X}} = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} M_i}.$$

Proportion:

$$\widehat{P} = \frac{\sum_{i=1}^{n} A_i}{\sum_{i=1}^{n} M_i},$$

where

$\widehat{X}_i$ is the total of variable X in cluster i.

$\widehat{\overline{X}}_i$ is the sample average of variable X in cluster i.

N is the number of clusters of the population.

M is the size of the population.

n is the number of clusters of the sample.

M_i is the size of cluster i.

A_i is the total of variable A, which takes values 0 or 1 in cluster i,

and the estimation of the errors we make when we estimate through these expressions are:

Total:

$$\widehat{V}(\widehat{X}) = \frac{N(N-n)}{n}\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}M_i)^2.$$

Average:

$$\widehat{V}(\widehat{\overline{X}}) = \frac{N(N-n)}{M^2 n}\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}M_i)^2.$$

Proportion:

$$\widehat{V}(\widehat{P}) = \frac{N(N-n)}{M^2 n} \frac{1}{n-1} \sum_{i=1}^{n} (P_i - \overline{P} M_i)^2.$$

1.6 Systematic sampling

We can think about a different way of sampling. Let us imagine that in your high school and we have decided to choose 28 people. In this case, the elevation factor would be $560/28 = 20$. We number students from 1 to 560. We then choose a number x randomly from 1 to 20 and this would be the first student selected. Then, we select number $x+20, x+2\cdot 20$ and so on. It is not a random sampling because all the samples are not equally probable. Let us define this sampling technique.

Let us suppose that we have a population of N elements ordered and numbered from 1 to N, and we want to get a sample with n elements. This population can be divided in n subsets, each of them with $v = \frac{N}{n}$ elements, i.e., each subset has as many elements as the elevation factor indicates.

We randomly choose a numbered element from 1, 2 until $\frac{N}{n}$ and we call it x_0, and then we take the following elements: $x_0 + v, x_0 + 2v, x_0 + 3v, x_0 + 4v, \ldots$

In case that v is not a natural number, we clear to the closer one (lower), so maybe some samples may have size $n-1$. This fact brings a small perturbation in the theory of systematic sampling, that we do not have to take into account, if $n > 50$.

This type of sampling needs that we have previously checked that the ordered elements present no periodicity in the variables we want to study, because if we can find periodicity and it is close to value v, the results that we obtain would have a big bias and would not be valid.

Systematic sampling is equivalent to random sampling if the elements are numbered in an random way.

Advantages of this method are:

1. Extends the sample to all the population.
2. It is very easy to apply it.

Disadvantages of the method are:

1. Increase of the variance if there is periodicity in the numbering of the elements, appearing a bias due to selection.
2. Problems when we want to estimate the variance.

We can consider an instance of cluster sampling, having each cluster the following elements we present by their number in the list:

First cluster: $1, 1+v, 1+2v, 1+3v, 1+4v, \ldots$

Second cluster: $2, 2+v, 2+2v, 2+3v, 2+4v, \ldots$

$\ldots$

v-th cluster: $v, 2v, 3v, 4v, \ldots\ nv$.

Selecting a systematic sample is equivalent to select randomly only one cluster. To do so, it is necessary that each of the clusters has a similar structure to the population.

We can also consider systematic sampling as a particular case of stratified sampling with n strata, each of them with v elements, so that we choose only one element of each stratus.

In stratified sampling the selected element is random, while in this technique we choose randomly the first element and the rest are determined by factor v.

The estimators for this type of sampling are:

Total:

$$\widehat{X} = v \sum_{i=1}^{n} X_i.$$

Average:

$$\widehat{\overline{X}} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Proportion:

$$\widehat{P} = \frac{1}{n} \sum_{i=1}^{n} P_i,$$

where P is a variable taking values 0 or 1.

1.7 Other sampling techniques

Two-stage sampling is a particular case of cluster sampling in which in the second stage we do not select all the elements of the cluster, but some elements chosen in a random way. Clusters in the first stage are called primary units and the ones in the second stage are secondary units.

Multistage sampling is a generalization of the previous technique, so that each cluster can be a group os clusters and so on in each stage.

In general, to make complicated studies concepts of stratifying, clusters and random sampling are used. For instance, the population of a country can be divided into clusters (provinces, cities, quarters) that can be heterogeneous inside (for instance, referring to consumption) but homogeneous among them. Afterwards it is necessary to divide these units in homogeneous strata (primary units, for instance, quarters). Each of these units is divided into new units (buildings) called secondary units, which are divided into flats (houses). We would choose our sample in the following way:

1. We select a stratified sample. We would take at least one stratus (one quarter).
2. We choose randomly some buildings of each of the selected quarters.
3. We take randomly one or several houses of each of the buildings selected.

Chapter 2

An example of the application of sampling techniques

We have decided to make a study in a high school. We want to have data about the number of left handed students, the number of students who have internet connection at home, the height of the students and the pocket money they receive weekly.

The usefulness of knowing the number of left handed students of a high school is easy to understand, because the high school should have an appropriate equipment for them, for instance adapted chairs.

Internet connection at home is an important information. It can be used not only to check wether it is possible to offer some material for the students through the internet, but also to know if they access to some other didactic information available on the web.

The study of height is classical. It is anyway interesting to know if height is changing with years and the population is getting taller.

Pocket money is a social relevant data. It is also interesting to know how much money the students deal with, and it is also interesting to know how they spend it to understand what they devote their time to.

Once we have fixed what we want to get, we decide to sample to get the conclusions about all the students of the high school without asking each of them. The information available for us is the one referred to the distribution of students in years and classes:

	A	B	C	D	E	Total
1st year	33	20				53
2nd year	20	15	30			65
3rd year	20	15	26	14		75
4th year	27	27	25			79
5th year	33	28	30	31	23	145
2th year	30	34	32	31		127

So we are working with a population of 544 students of a high school.

We start posing that we are going to use a sample size of around 60 students, which is the maximum allowed and that we think that may be enough for the study we are going to make.

We can get then the first information, our sampling fraction would be:

$$f = \frac{n}{N} = \frac{60}{544} = 0.1102,$$

this is, we are going to sample approximately 11% of the population. We can also calculate the elevation factor which would be:

$$E = \frac{N}{n} = \frac{544}{60} = 9.1,$$

or equivalently, each student interviewed represents 9 colleagues.

Now we have to decide which method we want to use to sample the different characteristics we are going to study. Let us denote them in the following way:

- X will represent the height.
- Y will represent the pocket money.
- Z will represent variable "being left handed", which will take value 1 if a student if left handed and 0 if he/she is not left handed.
- I will represent variable "having internet connection at home" which will take value 1 in affirmative case and 0 in negative case.

We will make a difference into 2 cases of the 4 variables. The first thing, we make to ourselves a question: we have our population divided into groups and levels, can we consider that this division has an influence in any of these variables? This is, can we consider that in each level, for instance, the average height can change? The answer to this question is that it is logic to think that it will change. A priori, we can suppose that the age has an important influence for the height. And for the pocket money? Then the age is also important, because we all could get more money from our parents while getting older. Does it happen the same for being left handed? Then, the answer is no because if you are left handed, this happens from the day you were morn, so age has no influence on this. And the same applies for the fact of having internet connection at home. So we choose different sampling techniques for these two cases.

Case I: variables pocket money and height

we have already mentioned that we have the population divided into levels and groups. For us, the division in levels is a division into *strata* because the levels are homogeneous inside them with respect to the age (and we can also think that it happens the same for the pocket money and the height), and as we have said before, age has a big influence on these variables and it makes sense that we are interested in having all these strata represented in our sample. So we choose for these cases *random stratified sampling.*

The next thing to be done is to decide the sample size inside each strata.

We have 6 strata with the following sizes:

Stratus	Size
1st level (stratus 1)	$N_1 = 53$
2nd level (stratus 2)	$N_2 = 65$
3rd level (stratus 3)	$N_3 = 75$
4th level (stratus 4)	$N_4 = 79$
5th level (stratus 5)	$N_5 = 145$
6th level (stratus 6)	$N_6 = 127$

The usual thing in this situation is to use sample size in the strata proportional to their size, so that the sizes of the samples keep the same proportion than the sizes of the strata. We calculate then the size of the sample in each stratus through the following expression:

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

and we get the following sample sizes:

$n_1 = 60 \cdot \frac{53}{544} = 5.84$ so we take $n_1 = 6$,
$n_2 = 60 \cdot \frac{65}{544} = 7.16$ so we take $n_2 = 8$,
$n_3 = 60 \cdot \frac{75}{544} = 8.27$ so we take $n_3 = 8$,
$n_4 = 60 \cdot \frac{79}{544} = 8.71$ so we take $n_4 = 8$,
$n_5 = 60 \cdot \frac{145}{544} = 15.99$ so we take $n_5 = 16$,
$n_6 = 60 \cdot \frac{127}{544} = 14.00$ so we take $n_6 = 14$,

where the clearing has been made to keep the sample size 60 that we had posed. So we have the sample sizes that we needed and we can make random sampling inside each stratus, to select the number of students that we have already decided.

Our data are the following:

for the height we got:

Stratus 1	165	161	153	150	151	153										
Stratus 2	157	161	168	162	165	171	169	164								
Stratus 3	168	165	175	175	165	163	165	165								
Stratus 4	164	171	177	163	170	165	160	175								
Stratus 5	175	173	161	158	175	164	158	161	158	171	175	170	187	168	170	185
Stratus 6	190	178	194	183	165	170	176	173	168	183	173	183	174	177		

and for the pocket money:

Stratus 1	10	0	3.5	0	0	3										
Stratus 2	0	5	0	15	0	3	2	0								
Stratus 3	5	8	8	0	20	5	10	10								
Stratus 4	12	6	5	12	12	6	0	0								
Stratus 5	5	10	12	15	10	12	30	12	30	10	6	5	10	21	40	15
Stratus 6	12	10	9	6	8	9.4	15	0	20	10	15	10	0	0		

We now proceed to the estimations The first thing to do is to calculate the average in the strata, which gives us information about the behavior of the variables in the strata. Later on, we will calculate the average of the height and the pocket money of the students of the high school and we

will give it together with an estimation of the error we get when we make such an estimation. We make the process independently for each of the variables:

For the height we have:

Stratus	Average	Std deviation
1	$\overline{x}_1 = 155.5$	$S^2_{x1} = 36.7$
2	$\overline{x}_2 = 164.625$	$S^2_{x2} = 21.4107$
3	$\overline{x}_3 = 167.625$	$S^2_{x3} = 22.5535$
4	$\overline{x}_4 = 168.125$	$S^2_{x4} = 36.6964$
5	$\overline{x}_5 = 169.3125$	$S^2_{x5} = 81.6958$
6	$\overline{x}_6 = 177.642857$	$S^2_{x6} = 67.478$

We can directly see that something curious. The average is increasing as the level increases. This leads us to think that the choice of stratified sampling has been right in this case.

We calculate now the same for pocket money:

Stratus	Average	Std deviation
1	$\overline{y}_1 = 2.75$	$S^2_{y1} = 4.026$
2	$\overline{y}_2 = 3.125$	$S^2_{y2} = 26.4107$
3	$\overline{y}_3 = 8.25$	$S^2_{y3} = 33.3571$
4	$\overline{y}_4 = 6.625$	$S^2_{y4} = 25.4107$
5	$\overline{y}_5 = 15.1875$	$S^2_{y5} = 101.2291$
6	$\overline{y}_6 = 8.8857$	$S^2_{y6} = 35.229$

Now we calculate the estimated average from the complete sample and the estimation of the error in terms of the estimation of the variance for the 2 variables we are studying. For the height:

$$\widehat{\overline{X}} = \sum_{h=1}^{6} w_h \overline{x}_h = \sum_{h=1}^{6} \frac{N_h}{N} \overline{x}_h = \frac{53}{544} \cdot 155.5 + \frac{65}{544} \cdot 164.625 + \frac{75}{544} \cdot 167.625 + \frac{79}{544} \cdot 168.125$$

$$+\frac{145}{544} \cdot 169.3125 + \frac{127}{544} \cdot 177.642857 = 168.9463.$$

The expression for the variance is

$$\widehat{V}(\widehat{\overline{X}}) = \sum_{h=1}^{k} w_h^2 (1 - f_h) \frac{\widehat{S}_h^2}{n_h},$$

and in our case we have:

Stratus	w_h	w_h^2	f_h	$1 - f_h$
1	$\frac{53}{544} = 0.095$	0.009	$\frac{6}{53} = 0.1132$	0.8868
2	$\frac{65}{544} = 0.1194$	0.014	$\frac{8}{65} = 0.123$	0.8769
3	$\frac{75}{544} = 0.1344$	0.018	$\frac{8}{75} = 0.1066$	0.8934
4	$\frac{79}{544} = 0.1415$	0.02	$\frac{8}{79} = 0.1012$	0.8988
5	$\frac{145}{544} = 0.2598$	0.0675	$\frac{16}{145} = 0.1103$	0.8897
6	$\frac{127}{544} = 0.2276$	0.0518	$\frac{14}{127} = 0.1102$	0.8898

Now we substitute these numbers in the previous expression and we get:

$$\widehat{V}(\widehat{\overline{X}}) = \sum_{h=1}^{k} w_h^2(1-f_h)\frac{\widehat{S}_h^2}{n_h} = 0.009 \cdot 0.8868 \cdot \frac{36.7}{6} + 0.014 \cdot 0.8769 \cdot \frac{21.4107}{8} + 0.018 \cdot 0.8934 \cdot \frac{22.5535}{8}$$

$$+0.02 \cdot 0.8988 \cdot \frac{36.6964}{8} + 0.0675 \cdot 0.8897 \cdot \frac{81.6958}{16} + 0.0518 \cdot 0.8898 \cdot \frac{64.478}{14} = 0.728.$$

So in the case of the height we already have our estimations. The estimated average height is 168.9463 and we calculate that we have an error of 0.728.

Now we make the same calculation for the pocket money. We start by calculating the estimated average:

$$\widehat{\overline{Y}} = \sum_{h=1}^{6} w_h \overline{y}_h = \sum_{h=1}^{6} \frac{N_h}{N} \overline{y}_h = \frac{53}{544} \cdot 2.75 + \frac{65}{544} \cdot 3.125 + \frac{75}{544} \cdot 8.25 + \frac{79}{544} \cdot 6.625$$

$$+\frac{145}{544} \cdot 15.1875 + \frac{127}{544} \cdot 8.8857 = 8.8633.$$

The estimation of the variance can be calculated directly because we have the same values for w_h and f_h:

$$\widehat{V}(\widehat{\overline{Y}}) = \sum_{h=1}^{k} w_h^2(1-f_h)\frac{\widehat{S}_h^2}{n_h} = 0.009 \cdot 0.8868 \cdot \frac{4.026}{6} + 0.014 \cdot 0.8769 \cdot \frac{26.4107}{8} + 0.018 \cdot 0.8934 \cdot \frac{33.3571}{8}$$

$$+0.02 \cdot 0.8988 \cdot \frac{25.4107}{8} + 0.0675 \cdot 0.8897 \cdot \frac{101.2291}{16} + 0.0518 \cdot 0.8898 \cdot \frac{35.229}{14} = 0.666.$$

Case II: Variables "being left handed" and "having internet connection at home"

Now we want to study variables "being left handed" and "having internet connection at home". It is easy to see that the division into strata is not useful in this case, so we should think about using some other sampling technique. We still want to get a sample of around 60 students. We could think that, with respect to these variables, the groups that the population is divided in behave like small populations, i. e., we can consider that the groups behave like the whole high school. Moreover it is interesting for us the possibility of sampling some groups because selecting a random sample of students, finding them and interviewing them is not an easy task.

But now, what are groups for us? We have already said that inside them, they behave like small populations with respect to our variables, while the groups are similar among them. This means that we have the population divided into clusters, so we will apply cluster sampling to this situation.

The next thing to be done is the number of groups to be sampled. We know that the groups do not have the same size, but 2 or 3 groups would assure us a sample of around 60 students. To avoid the possibility of having a sample of 2 small groups and then getting a too small sample for our purposes, we decide to select 3 groups from the high school.

So the data we have got are the following. For the variable "being left handed":

Cluster 1:

1 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0,

Cluster 2:

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,

Cluster 3:

0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0,

where 1 means being left handed and 0 means being not left handed. Now, for the variable "having internet connection at home", we got:

Cluster 1:

1 0 0 1 0 1 0 1 0 1 1 1 1 0 1 0 0 1 0 0,

Cluster 2:

1 1 1 0 1 1 0 1 1 0 1 0 0 1 1 1 1 1 1 1 1 1 0,

Cluster 3:

1 1 0 1 0 1 1 1 1 0 1 1 1 1 1 1 0 1 1 1 0 1 1 1 1,

where now 1 means having internet connection at home and 0 means having not.

We start now estimating the total amount of left handed students, and the proportion of left handed students, as well as the total amount of students having internet connection at home and the proportion that this represents in the whole high school.

We calculate the total and proportion for each group and variable:

	Left handed		Internet	
Cluster	Total	Proportion	Total	Proportion
1	3	0.15	10	0.5
2	0	0	17	0.7391
3	2	0.08	20	0.8

Now we can calculate the estimations for the proportion and total of variables Z and I. We start with variable Z:

$$\widehat{Z} = M \cdot \frac{\sum_{i=1}^{n} \widehat{Z}_i}{\sum_{i=1}^{n} M_i} = 544 \cdot \frac{\sum_{i=1}^{3} \widehat{Z}_i}{\sum_{i=1}^{3} M_i} = 544 \cdot \frac{3+0+2}{20+23+25} = 544 \cdot \frac{5}{68} = 40,$$

$$\widehat{P_Z} = \frac{\sum_{i=1}^{n} A_i}{\sum_{i=1}^{n} M_i} = \frac{3+0+2}{20+23+25} = \frac{5}{68} = 0.0735,$$

and we do the same for variable I

$$\widehat{I} = M \cdot \frac{\sum_{i=1}^{n} \widehat{I}_i}{\sum_{i=1}^{n} M_i} = 544 \cdot \frac{\sum_{i=1}^{3} \widehat{I}_i}{\sum_{i=1}^{3} M_i} = 544 \cdot \frac{10+17+20}{20+23+25} = 544 \cdot \frac{47}{68} = 376,$$

$$\widehat{P_I} = \frac{\sum_{i=1}^{n} A_i}{\sum_{i=1}^{n} M_i} = \frac{10+17+20}{20+23+25} = \frac{47}{68} = 0.6911.$$

We continue now estimating the error we have committed for the variable "being left handed":

$$\widehat{V}(\widehat{Z}) = \frac{N(N-n)}{n} \frac{1}{n-1} \sum_{i=1}^{n} (Z_i - \overline{Z} M_i) = \frac{21(21-3)}{3} \frac{1}{2} \left[(3 - 0.0735 \cdot 20)^2 + (0 - 0.0735 \cdot 23)^2 + (2 - 0.0735 \cdot 25)^2 \right]$$

$$= 329.18.$$

$$\widehat{V}(\widehat{P_Z}) = \frac{N(N-n)}{nM^2}\frac{1}{n-1}\sum_{i=1}^{n}(P_{Zi}-\widehat{P}M_i) = \frac{21(21-3)}{3\,(544)^2}\frac{1}{2}\left[(3-0.0735\cdot 20)^2+(0-0.0735\cdot 23)^2+(2-0.0735\cdot 25)^2\right]$$

$$= 0.00111,$$

and finally we calculate the estimated errors for "having internet connection at home",

$$\widehat{V}(\widehat{I}) = \frac{N(N-n)}{n}\frac{1}{n-1}\sum_{i=1}^{n}(I_i-\overline{I}M_i) = \frac{21(21-3)}{3}\frac{1}{2}\left[(10-0.6911\cdot 20)^2+(17-0.6911\cdot 23)^2+(20-0.6911\cdot 25)^2\right]$$

$$= 1464.123.$$

$$\widehat{V}(\widehat{P_I}) = \frac{N(N-n)}{nM^2}\frac{1}{n-1}\sum_{i=1}^{n}(P_{Ii}-\widehat{P}M_i) = \frac{21(21-3)}{3\,(544)^2}\frac{1}{2}\left[(10-0.6911\cdot 20)^2+(17-0.6911\cdot 23)^2+(20-0.6911\cdot 25)^2\right]$$

$$= 0.0038.$$