



www.alea.pt

Educational Dossiers



Statistical Surveys

An introduction to drafting questionnaires, sampling, and organising and displaying results

Maria João Ferreira

Pedro Campos

1. Foreword

The ALEA project - Local Applied Statistics Initiative - contributes toward the creation of new statistics teaching support media for primary and secondary students and teachers.

The project arose from an idea jointly fostered by Escola Secundária Tomaz Pelayo secondary school and Instituto Nacional de Estatística [National Statistics Institute of Portugal], founded on the requirements and structures that the intervening parties possess. The improvement of statistical literacy is thus a significant proviso in guaranteeing the provision of a service of public value. The teaching of statistics in lower and upper secondary education constitutes one of the most important instruments aimed at achieving this objective. ALEA's site on the internet is at the address: www.alea.pt.

Contents

1. Foreword
2. Why Perform Statistical Surveys?
3. Survey, Observation and Experimentation
4. How to Ask Questions? - General Rules for Constructing a Questionnaire
5. Choosing the Population to Survey and Data Collection Methods: Sampling
6. Collecting the Necessary Information on Sample Elements
7. Data Organisation and Display
8. See Also

The Educational Dossiers area was designed to support the creation of educational material on a range of topics.

Dossiers available in the English version of ALEA:

- Notes on the History of Statistics
- The Cinema in Numbers
- Graphical Representations
- Statistical Surveys

The **Statistical Surveys** dossier now follows. Professor Maria Eugénia Graça

Martins, from the Faculty of Sciences of Universidade de Lisboa [Lisbon University], contributed to this dossier through her role of scientific consultant to ALEA.

This dossier contains a small introduction to the phases of a survey performed by questionnaire, the rules for constructing a questionnaire, basic information on how to select sample elements and also how to draw up a report for the final presentation of the results. At the end of the dossier, the *See Also* section contains links to other studies of interest related to the topics covered (publications and websites).

2. Why Perform Statistical Surveys?

The **Survey** is one of the most used instruments in the field of applied research, particularly in social domains. The number of studies, ranging from market surveys to purely theoretical research and opinion **polls**, that are not totally or partially founded on data collected by means of **surveys**, is very few.

Poll:

A scientific study of one part of a population with the objective of studying attitudes, habits and preferences of the population relative to events, circumstances and subjects of public interest.

2.1. What is a Statistical Survey?

It is the need to find out about one or more characteristics of a **population** that leads us to undertake **surveys**.

Population:

A collection of individual units, which can be people or experimental results, with one or more common characteristics, that is to be studied.

The alternative to direct **observation**, in certain cases, even if it were feasible, would take too long or it would be impossible when the phenomena under analysis relate to the past (Ghiglione and Matalon, 1992).

The use of a **survey** is necessary each time we have the need for information on a wide variety of forms of behaviour of the same individual, or how much we intend to know about the same type of variable for many individuals.

A **survey** can be deemed to be specific questioning regarding a situation encompassing individuals, in order to generalise.

An example of a survey performed by INE:

*The objective of the **Survey of Household Economy** performed by INE is to obtain data on the origin and value of the income of households and how these are transformed into consumption expenditure. This survey provides the data to update the Consumer Price Index, to develop and construct a Poverty Indicator system, to analyse expenditure and income concentrations amongst households, as well as the data to perform other socio-economic studies.*

1º Dia do Inquérito

EXEMPLO

SEGUNDA-FEIRA

2.1.01. COMPRAS DO DIA Fez compras neste dia? Sim ☒ 1 Não ☐ 0

Nº de Linha	Tipo de Estabelecimento	Designação do produto	Quant.	Valor
001	Pisaria	Canapau fresco	1 Kg	1 51 0 01 0
002	Mercearia	Frijão Verde	0,5 Kg	1 51 0 01 0
003	"	Batatas	3 Kg	1 51 0 01 0
004	"	Agriões	0,750 Kg	1 51 0 01 0
005	"	Alface	1,2 Kg	1 51 0 01 0
006	Supermercado	Leite gordo - Longa duração	5 Lt	1 51 0 01 0
007	"	Ovos	12	1 51 0 01 0
008	Pronto-a-vestir	Camisola de Malha - Homem	1	1 51 0 01 0
009	"	Collants - Senhora	1	1 51 0 01 0
010	"	Fato de Treino - Criança	1	1 51 0 01 0
011				
012				
013				
014				
015				
TOTAL DE LINHAS				

2.2.01. AUTOCONSUMO Consumiu bens de produção própria? Sim ☒ 1 Não ☐ 0

Nº de Linha	Designação dos produtos de produção própria que consumiu	Quant.	Valor
001	Cebola	1,3 Kg	1 51 0 01 0
002	Vinho Maduro Branco	1 Lt	1 51 0 01 0
003			
004			
005			
TOTAL DE LINHAS			

2.3.01. AUTOABASTECIMENTO Retirou do seu estabelecimento algum produto? Sim ☒ 1 Não ☐ 0

Nº de Linha	Tipo de Estabelecimento	Produtos retirados do estabelecimento, sem pagar, para consumo do agregado	Quant.	Valor
001	Tabacaria	Português Suave	1 Maço	1 51 0 01 0
002	"	Marlboro	"	1 51 0 01 0
003				
004				
005				
TOTAL DE LINHAS				

Fig. 1 - Questionnaire used in the Survey of Household Economy (Source: INE)

Figure 1 contains one of the parts of the **questionnaire** that had to be completed every day by a member of the household, preferentially the person that did the shopping. The information (or data) **collection** method used in this **survey** combined collection via self-administered means (data completed by the surveyed individual) with data collection via **interview**. We shall tackle all of these data collection techniques further on in this dossier.

3. The Questionnaire and Survey Phases

3.1 Survey and Questionnaire

In this section we will broach the **survey** and **questionnaire** basics, covering the different data collection methods.

There are two types of data collection techniques: **documental** and **non-documental**. The objective of **documental**

Questionnaire:

It is one of the data registration media in a survey, which may or may not be done via an interview.

techniques is to collect data from already existing bibliographical media. Bibliographical research, text analysis, in addition to others, are an example. With **non-documental** techniques, the researcher makes direct observations (such as measuring the height of an athlete's leap or the number of push-ups per minute, for example) or indirect observations – which can be done, in this case, through the completion of a **questionnaire**.

The following figure comprises a flowchart regarding data collection techniques.

The **questionnaire** is one of the most used techniques in **surveys**. It is a non-documental technique, one of indirect **observation**, that can be implemented by means of an interview. The **survey** is often seen as a complete process (covering the stages from collection, to analysis, using various techniques). The **questionnaire** is the instrument of notation.

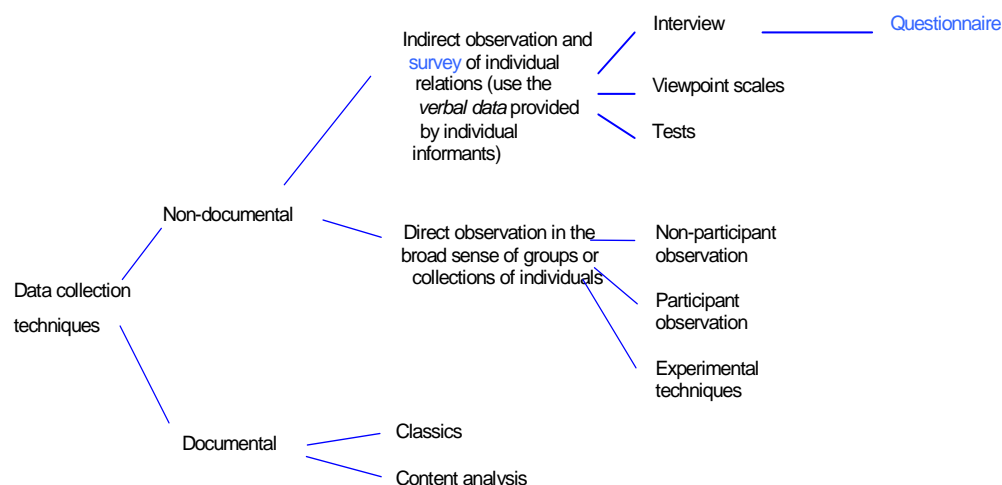


Fig. 2 (adapt. Lima, 1981)



We use a **survey**, as previously stated, to understand phenomena such as attitudes, opinions, preferences, etc., which are only practically accessible through language and which are only rarely spontaneously expressed. The **survey**, and often **observations**, provides us with data on what is occurring at a specific point in time. The use of a greater number of **questions** permits more in-depth analysis, allows the opinions and behaviour under analysis to be more perceptively described, allows more complex hypotheses to be verified, etc.

Notwithstanding all the advantages that a **questionnaire** has, there are always disadvantages, one of the most significant of which is that the **questionnaire** is totally dependent on language - everything we have is based on that which the respondent can or wanted to say.

The questions of a **questionnaire** must, therefore, use simple words and accessible, clear and precise language (eliminating the possibility of subjective interpretation by the respondent). The **questions** must be concise and direct (avoiding negatives, and double-negatives in particular).

We will explore the rules for the construction of **questionnaires** in more detail in the next chapter.

3.2. The stages of development of a survey

The stages of development of a **survey** do not follow a constant linear order. According to Ghiglione and Matalon (1992), before performing a **survey** we must know **who we want to survey** and what **we must ask**. We can say that we must take into consideration a number of issues when **drawing up a survey**: the population to be surveyed and what we want to find out about them, as well as the survey's objectives and how it will be performed and used must be defined during the **survey planning stage**; then, the **notation instrument (questionnaire) must be drafted**, in which special attention must be paid to the type of questions, their order, the language used and the final layout; lastly, there is the implementation of the **field work (data collection)**, where all the information required to achieve the survey's objectives is collected. We will discuss the various ways in which data can be collected further on.

4. How to Ask Questions? - General Rules for Constructing a Questionnaire

The **questionnaire** is one of the notation instruments most used to obtain information relative to a given **population**. The **questionnaire**'s construction and the manner in which the questions are formulated form a fundamental phase in the development of a **survey** (Ghiglione and Matalon, 1992). In order to construct a **questionnaire** we have to know exactly what it is we are looking for, to guarantee that the questions are interpreted in the same manner by all respondents, that all aspects of the questions have been well broached, etc. These conditions are verified by **interview** and testing the first versions of the **questionnaire** (**pre-test**).

Pre-test:

This is a trial run of the questionnaire amongst a small part of the population before its general use in the population.

4.1 The different types of questions

The first questions of a **questionnaire** are very important (Ghiglione and Matalon, 1992). These are the ones that indicate the general style of the **questionnaire** to the respondents, the type of response expected from them and the subject to be broached. These questions are also the basis for the establishment of the **interviewer-interviewee** relationship and determine how interviewees will react, namely if they feel that their private life is being invaded. It is usually preferable to start with questions that peak the interest of interviewees and do not scare them.

A **questionnaire**'s questions can be **closed-ended**, **open-ended** and **semi-open**.

4.1.1 Closed-ended questions

A **question** is said to be **closed-ended** if the answer is imposed (Grangé and Lebart, 1994). For example,

What is your marital status?

- [1] Single
- [2] Married or common-law union
- [3] Divorced or separated
- [4] Widowed

These types of questions authorise pre-coding, in other words, an immediate translation of the answer into an alpha-numerical code. These questions limit the respondents' answer solely to the answer options presented.

There are various types of **closed-ended questions**:

- Single choice questions (the respondent chooses just one answer from the list of options);
- Multiple choice questions (the respondent chooses a limited number, or limitless number, of answers from a list), for example:

In your opinion, what are the strong points of product X? (specify up to a maximum of 3 choices)

- | | |
|--------------------------|-------------------------|
| [1] general presentation | [6] solidity |
| [2] shape | [7] price |
| [3] easy to use | [8] length of warranty |
| [4] diversity of uses | [9] after-sales service |
| [5] effectiveness | |

- Classification (the respondent lists the different answer options in order of importance), for example:

Rank the following characteristics of product Y in order of importance as strengths - where 1 is the most important and 9 the least important characteristic.

- | | |
|--------------------------|-------------------------|
| [] general presentation | [] solidity |
| [] shape | [] price |
| [] easy to use | [] length of warranty |
| [] diversity of uses | [] after-sales service |
| [] effectiveness | |

Scaled questions are also a type of **closed-ended question**. These types of questions allow responses to be attenuated relative to agree/disagree types of question. We can establish a complete scale of responses in situations of this type:

I fully agree / I partially agree / I am indifferent / I do not much agree / I totally disagree

A **questionnaire** that is mainly composed of **closed-ended questions** must not surpass 45 minutes when completed in good conditions, in other words, in the

respondent's home or in a tranquil area (Ghiglione and Matalon, 1992). If this time limit is surpassed, the respondent loses interest, which can be noted by means of signs such as the speed with which responses were provided to questions, indicating the level of thought put in.

Closed-ended questions are, in principle, the most favourable from a results' analysis standpoint. In regard to a **survey** implemented and analysed quickly, such as an opinion **poll**, every effort is made to use only these types of questions.

Closed-ended Questions:

Questions presented to the respondent containing a pre-established list of responses. The respondents indicate the response that best corresponds to the response they wish to provide.

4.1.2 Open-ended questions

There are no constraints to the response to these types of question. The response is transmitted through the most reliable means.

Open-ended Questions:

The respondent can reply to the question as they deem fit, using their own vocabulary.

The space set aside must be measured beforehand to facilitate the elaboration of the response (Grangé, 1994).

An example of an open-ended question:

What kind of dishwasher powder do you use?

The reasons for providing **open-ended questions** are various. There is often not enough time to draw up a list of standardised responses to the question, therefore a blank space is left for the respondent to fill in. On the other hand, open-ended questions may be necessary when **questionnaire pre-tests** (see 4.5) were insufficient, or also when the responses provided in those **pre-tests** seemed to be too complex to be summarised in a list of acceptable size (Ghiglione and Matalon, 1992). Lastly, there is a compelling reason for us to leave a question **open-ended**: and that is that a **questionnaire** totally composed of closed-ended questions quickly becomes tedious. If people are provided with lists of responses, they can reflect less on their responses and take less care with what they say. Another reason for choosing the open-ended method is that it permits different coding, since after all the responses are analysed, they are coded through the construction of a code book.

4.1.3 Semi-open questions

Open and closed-ended responses can occur in the same question of a **questionnaire**:

With which company is your vehicle insured?

[1] company A
[2] company B
[...] ...
[10] other: _____

This combined form of question tends to solve the problems of pertinence and comprehensive nature of **closed-ended questions**, strongly reducing the post-inquiry coding costs due to a 'verbatim' response.

Sample Survey

Year _____ Class _____ Course _____

Gender: Male ☐ Female ☐

1. Do you have a telephone at home? Yes ☐ No ☐

2. Do you own a personal computer? Yes ☐ No ☐ *If the answer is no, please go to question 4*

2.1. Do you use or know how to use a computer?

2.2. Do you use it for:

Studying ☐ Doing school work ☐ Playing games ☐

"Surfing" the internet ☐

Other (please specify) _____

2.3. How many hours per week do you use a computer? _____

3. Do you use internet services? Yes ☐ No ☐

3.1. If yes, where?

At home ☐ At school ☐ At a friend's or family member's home ☐

Somewhere else (please specify): _____

4. What is your opinion of the "Internet at School" initiative?

Agree ☐ Disagree ☐ Don't know/don't wish to respond ☐

5. Do you consider your parents' income to be

High ☐ Medium ☐ Low ☐ Very low ☐

Thank you for participating in this study.

Closed-ended questions

Semi open ended questions

Fig. 3 - Example of a questionnaire designed by students of Tomaz Pelayo School



4.2 Question order

The concept of a start, middle and end must be taken into account when drawing up a **questionnaire**. There is no rule for question order, but there are some recommendations that should be followed. The start must contain a small introduction to the entity carrying out the survey, the **questionnaire**'s objective and the advantages that this study may have for the general public.

The first questions must be simple, since they will determine how the **questionnaire** goes. If the first questions are complicated, then the respondent may lose interest, which makes the **interviewer**'s task much harder. As the

questionnaire progresses the questions must become more specific, covering potentially embarrassing or private areas, such as "Do you brush your teeth

The first questions must be simple since they will determine how the questionnaire is handled.

every day?", as well as areas that may be more mentally taxing, such as asking respondents to rank in order of preference the products that they like most, etc. It does not matter if personal data comes at the start or at the end, it is up to the researcher. All questions must be clear, must never suggest a particular response and they must not express any expectation (Ghiglione and Matalon, 1992). One cannot ask everything in a **questionnaire**, since the different study areas can produce many questions; thus you must have enough perspicacity to choose the most important questions for the study.

A **questionnaire** should seem to be as natural a dialogue as possible.

A **questionnaire** must seem to be as natural a dialogue as possible. The **questions** must be concise and sequential, without repetitions or being out of context. For example, before asking if somebody liked film X they must be asked if they have ever seen film X, since then we can have a **filter question** that will assess the information the interviewee has on the film. If the interviewee's information is zero, in other words they have never seen film X, the following questions which may be about the film no longer mean anything to the interviewee, thus, this question has to be a filter, transferring the interviewee to another question on another subject.

Filter questions:
They filter people for whom certain questions do not mean anything or are not applicable.

Example of a filter question:

1. Have you ever seen film X?

☐ Yes ☐ No (go to question 2.)

1.1. Did you like the film?

☐ Yes ☐ No

1.2. Would you see film X again?

☐ Yes ☐ No

2. Have you ever seen television series Y?

Filter question

4.3. Other suggestions for drawing up questions

A **questionnaire** must not just contain open-ended questions or closed-ended questions. The questions must be alternated so that the **questionnaire** does not become boring. A **questionnaire** which only has closed-ended questions can, as already stated, cause the interviewee to become a little ‘upset’, as there is the

Double questions must not be used, i.e., not more than one idea should be introduced in each question.

sense that the responses are being imposed. **Double questions**, which are questions containing more than one idea, must not be used. Before inserting **questions that an interviewee may find embarrassing**, such as questions about religion, the consumption of certain products, etc., we must provide a small introduction to the interviewee, since many people may be scared of providing the wrong answer or admitting to ignorance. Therefore, a rule is that these questions should be tackled as follows:

"...in your own personal situation, could you tell me..."; "I would like to know your opinion on ...".

4.4 The different types of scales

If a questionnaire contains closed-ended questions, then it is necessary to always select a set of options for each question (according to Hill and Hill, 2000). In the question on gender, for

example, the options are *male* and *female*. It is worthwhile coding the responses (associating numbers to each response) so that these can be analysed later on using statistical means. The two most common types of scale in questionnaires are **nominal scales** and **ordinal scales**. Other types of scale, such as **interval scales** and **ratio scales**, are also used.

4.4.1 Nominal scale

This type of scale is used in questions like the following example:

What is the post you hold in the company in which you work?

Management	Technician	Administrative	General unskilled worker
1	2	3	4

Each category of such **questions** can be attributed a number to code the response. These numbers only identify the categories. The different options or categories can be coded using other symbols, not necessarily numerical, if required – for example the categories of the gender variable, male and female, may be represented by M and F, respectively. It makes no sense to calculate the mean of the variables on a **nominal scale**, instead the frequency of the options are calculated (Hill and Hill, 2000). To find out more about the calculation of frequencies on a **nominal scale**, see the Basic Statistics course available on ALEA's site (chapter III, page 2, Data, Tables and Graphs - 1. Types of Data, at:

http://www.alea.pt/english/html/nocoes/html/cap3_1_1.html

4.4.2 Ordinal scale

This type of scale is used in questions like the following example:

Indicate your level of agreement or disagreement with the following statements regarding product X:

	Totally disagree	Disagree	Neither agree or disagree	Agree	Totally agree
Product X has an attractive package.	1	2	3	4	5
Product X is very expensive.	1	2	3	4	5

Categories are also used for ordinal variables, in the same way that they are used for nominal variables, but they have an order between themselves. If a jury ranks five candidates on a scale from 1 – weakest to 5 – strongest, we can say that the candidate in fourth place is better than the candidate in third place. Nevertheless, we cannot say that the candidate classified with the number 4 is twice as good as the candidate classified with the number 2, as it is not possible to measure the magnitude of the differences between categories (Hill and Hill, 2000). The calculation of means also has no meaning for ordinal variables, but, as seen as they are ranked, the median can be calculated.

4.5 The pre-test

We mentioned the **pre-test** at the start of the chapter. But what, after all, is a **pre-test** for?

It is necessary, when a first version of a **questionnaire** is drafted, in other words, when all the **questions** have been drawn up and their order provisionally fixed, to guarantee that the **questionnaire** is, in fact, usable and that it effectively responds to the problems posed by the researcher (Ghiglione and Matalon, 1992). The **questionnaire** must be tested out on a small group of people, to see if they understand the meaning of the **questionnaire** and the questions. This allows us to know how the **questions** and responses are understood, it allows us to detect vocabulary errors and style errors and to highlight rejected, incomprehensible and unequivocal questions or parts of the same (Ghiglione and Matalon, 1992). The **pre-test** allows us to assess the rejection rate, discover people's reactions to the **questionnaire** and if there is any problem with the **question** order. We can also check whether there are any **questions** that produced the same response amongst nearly all the respondents, which makes them not very useful for more detailed analyses involving cross-referencing with other **questions**. In such an event, the wording of the **questions** must be rectified. If the analysis of the **pre-test** produces many alterations, then it is necessary re-test the **questionnaire** as many times as necessary.

5. How to Select the Elements for the Sample

Each time a **poll** is performed a **sample** of the **population** to be studied must be selected. This sample is then **surveyed** to permit the extrapolation of the results to the population as a whole (Vicente, Reis and Ferrão, 1996).

The need to discover more about one or more characteristics of a **population**, leads to the implementation of a data collection and analysis process. The difficulty and, in certain cases, the impossibility of studying the entire **population** has raised the significance of studies by **sampling**. It is impossible to guarantee the quality of a **poll** if little is known of the problems and the impact that these can have on the study results.

Sample:

A set of data or observations collected from a subset of the population, which is studied in order to provide conclusions regarding the population as a whole from which it was extracted.

5.1 Polls versus Censuses

It is not always possible to perform a census on a **population** to be studied, in other words,

Census:

The scientific study of a population of people, institutions or physical objects in order to obtain knowledge, analysing all elements, and make quantitative inferences regarding important characteristics of that population.

survey all of the elements and, even if it were possible, it is a process that would take a very long time, making the study very expensive and possibly even senseless, since it would become outdated. **Polls** are cheaper and quicker, since it is much easier to access

all of the elements of a **sample** than those of an entire **population**.

Censuses are important, as they are useful for updating databases for the performance of **polls**. In Portugal, **censuses** are performed every ten years, providing an exhaustive update of both the housing in existence as well as the characteristics of the resident population. This database becomes outdated as time progresses, since population and housing changes quickly occur, therefore the database is updated through **surveys** performed by **sampling**.



Fig. 4 - A census is an instant photograph of the population at a specified time.

5.2 The implementation phases of a poll

As usually occurs in a **poll**, a **sample** extracted from a **population** is **surveyed** (Vicente, Reis and Ferrão, 1996). The design and implementation of a study based around a **poll** is a complex process involving a number of interdependent phases.

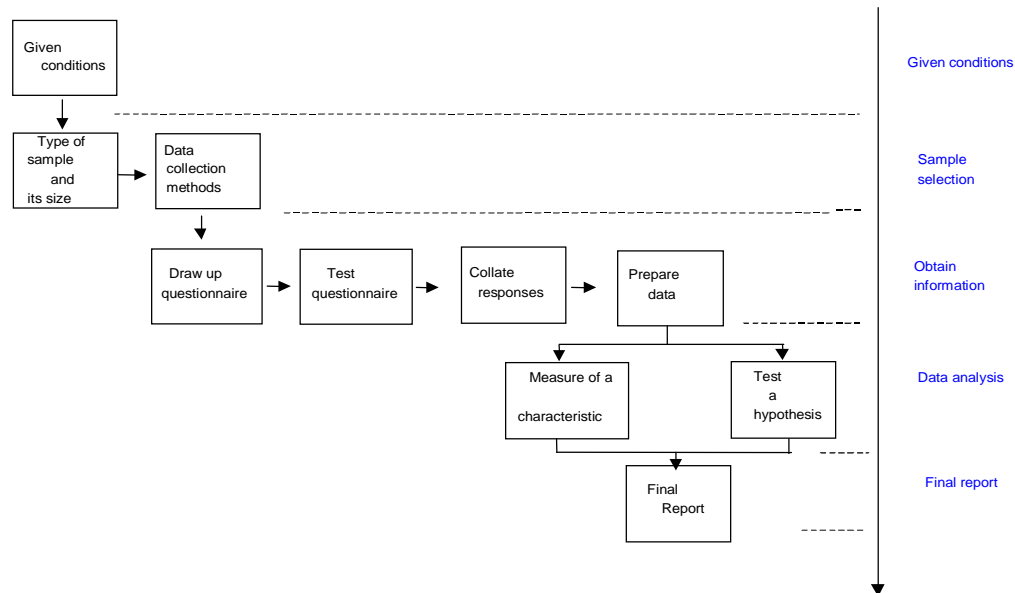


Fig. 5 - (adapt. Vicente, Reis and Ferrão, 1996)

The advantage of this scheme is that the **sampling** phase is clearly defined (at the *sample selection* level) in the **poll** process.

The section that follows is on **sampling methods**.

5.3 Sampling design stages

According to Vicente et al. (1996), “sampling design is the part of a **poll** where the selection of the elements from which the required data to be collected is made”.

Thus, the steps in collecting a **sample** can be described via the following sequence:

Sampling design steps

1. Define the target population
2. Identify the survey base
3. Select a sampling technique
4. Define the sample size
5. Select the sample elements
6. Collect the necessary information from the sample elements

5.3.1 Define the target population

The definition of the **target population** is one of the most important phases in undertaking a **poll**. Our study is focused on this **population**. The majority of authors (Cochran, 1963; Stuart, 1984; Barnett, 1991) define **target population** as all the elements on which our analysis is implemented and from which we aim to obtain data. In order to correctly define the **target population**, we first have

Target Population:

All the elements on which our analysis is implemented and from which we aim to obtain data.

Home:

A distinct and independent location constructed, reconstructed, extended or transformed for human housing and which, during the observation period, is not being totally used for any other purpose.

to be certain of the objective of our **survey**, and only then can we ask the question: who is encompassed by the **survey**? Which are the reference elements about which data we aim to obtain information? For example, let us suppose that the objective of our **survey** were to describe employment and unemployment in Portugal. This study has to be made amongst households, but as it is easier to detect families through their **home**, due to the existence of residence addresses, then our **target population** is that under **housing**.

5.3.2 Identifying the Survey Base

The **survey base** is a list of the elements from which the **sample** is selected (Vicente et al., 1996). It is necessary that the **sampling units** can be identified so that the **survey base** can be used as a source for the collection of the **sample**. The sampling units are elements or groups of elements of the **population**.

Survey Base:

Refers to the lists, charts or any other population register from which the sample will be extracted.

It is often impossible, due to the difficulty in constructing these listings, to ensure that the **target population** coincides with the

population to be surveyed or with the **survey base**. This is the case in regard to a very large population, making the successive selection of samples unacceptable. A very large sample that is very representative of the population is used in these cases. Such a sample is called a **survey base**. Then a number of samples, containing the characteristics of the initial target population, are extracted from this survey base. At Instituto Nacional de Estatística (INE), for example, the **mother sample** (used in a number of **surveys** such as employment surveys, for example) is a **large-scale sample** extracted from the **target population**, and from which other samples are extracted. Later, when the survey base starts to become saturated, this being when individuals have been surveyed a specified number of times, the survey base is updated by means of the substitution of the individuals for others. Gomes (1998) clearly explains that this strategy consists of updating a "representative" portion of the **target population**, which then takes on the role of **survey base**. In Portugal, INE updates the **mother sample** every five years and from 1998 onwards, a partial update has been performed yearly.

5.3.3 Selection of a sampling technique

Once the **target population** is defined, the **selection of the sample elements has to occur**. In this phase of the poll it is important to separate the probability-based or random-based methods (where a probability of inclusion in the **sample** is associated to the **population** elements) from the non-probability-based ones (in which such a probability is not calculated).

The probability based methods are associated to the selection of random samples. The entire **population** (or, where justified, the survey base) must be considered when a **random sample** is extracted.

Random sampling:

The procedure of selecting the elements or groups of elements in such a way that each population element has a calculable probability of inclusion in the sample, which is different to zero. In other words, each population element possesses a known probability of being selected.

A **sample** is considered to be **non-random** when specific elements of the **population** have no

Non-random sampling:

The procedure of selecting population elements that permits the choice of the individuals to be included in the sample, according to specific criteria that are more or less subjective. The probability of a specific element being selected is not known in this form of sampling.

possibility of being selected. Street interviews are an example of this, since, despite the fact that the people are chosen at random, the **sample** obtained is **non-random**, given that not all the individuals of a **population** have the same possibility of passing through the location at the time the interviews are being conducted.

It is important to note that the confidence level of the results (level of certainty relative to the precision of the estimate) can only be known with **random samples**, but, on the other hand, **non-random samples** provide for a study that is concluded much quicker and at lower cost (Vicente, Reis and Ferrão, 1996). No matter whether a **random or non-random sample** is chosen, it is important to obtain estimates close to the parameters to be estimated and this is only achieved if the **sample** is as representative as possible of the universe.



Fig. 6 - Door-to-door street interview

After a brief overview of the types of **samples**, we will take a more detailed look at the different sampling techniques. The main types of **random sampling** are: simple, systematic, stratified, by clusters, multi-stage and multi-phase.

A – Probabilistic-based methods

5.3.3.1 Simple random sampling

The most well-known type of **probability-based sampling** is **simple random sampling**. According to Stuart (1984), a **simple random sample** (s.r.s.) of size n is a **sample** selected by a process that gives each possible set of n elements the same probability of being selected.

In regard to this **sampling design**, all the elements of the **population** can be shown to have the same probability of being chosen to form part of the **sample**.

Sampling Design:
The method used to obtain the sample of the population.

A **simple random sample** can be obtained by the following means (Vicente, Reis and Ferrão, 1996):

Steps to obtaining a simple random sample:

1. Consecutively number the elements of the population from 1 to N ;
2. Select n elements using a random procedure such as the lottery method or using random number tables, which can be computer-generated. The numbers must be different and not greater than N ;
3. Once the numbers are chosen, the elements of the population to which they correspond form the sample.

The use of **s.r.s.** is not always the best option. The fact that all the individuals or objects of the **population** have the same chance of being in the **sample**, can result in geographically extensive **samples** and, if **face-to-face interviews** are necessary, then working with the

Face-to-face interview:

It can be considered to be a face to face conversation between two people, initiated and controlled by the interviewer, with the express purpose of obtaining relevant data, thereby permitting the attainment of the study's objectives.

sample obtained will prove to be expensive and time consuming. Such **samples** can be an excellent choice if the **population** is small and there are lists containing the **population** elements, thereby permitting the definition of a

survey base, and if the geographical spread of the elements is not a constraint.

An example application of simple random sampling:

A sample of 10 names has to be randomly selected from a population composed of 20 names. The researcher associates each name of the initial list to a number from 1 to 20, when in alphabetical order, for example. The numbers are represented by two digits - number 1 is 01, for example. Then the researcher uses a random number table (which can be found in practically all statistics books) to select two digit numbers until the required sample size is obtained. Note that it will be necessary select more than 10 numbers since some of the selected numbers will not form part of the population - for example, if the number 56 is generated, this will have to be ignored and a new number selected. Another process involves randomly generating ten numbers between 1 and 20 using a computer program (calculation worksheet, for example).

In a **population** composed of **N** items, the total number of possible **samples** of **n** items, extracted without replacement is given by :

$C_n^N = \frac{N!}{n!(N-n)!}$, thus, the probability¹ of any one being selected is given by :

$$\left(\frac{N!}{n!(N-n)!}\right)^{-1}.$$

5.3.3.2 Systematic random sampling

In a population of size **N**, ranked according to any criteria, a **systematic random sample** of size **n** is obtained by randomly selecting an element from amongst the first **K** of the **survey**

¹ The Portuguese version o ALEA has a “combinatory calculation” in the Basic Probabilities course of ALEA at: http://alea.ine.pt/html/probabil/html/cal_combinatorio/html/calcomb.html

base, where K is the whole number of the N/n quotient, and adding all of the following K -th elements (Vicente, Reis and Ferrão, 1996).

Steps to obtaining a systematic sample of size n :

1. Calculate the k interval of the sample (obtained by the N/n quotient, in which K represents the whole portion of that quotient).
2. Randomly choose a number j between 1 and k .
3. Starting from this number, successively add on the value k , thereby selecting the elements $j, j+k, j+2k, j+3k, \dots, j+(n-1)k$, covering a total of n observations selected for the sample.

The selection of an element with **systematic random sampling** depends on the previous selection. Only the first element is, in fact, randomly selected, since all the others are dependent on that first choice. The probability of selection in this type of **sample** is not identical for all elements.

An example application of systematic random sampling (a known population)

Extracted from Vicente, Reis and Ferrão (1996)

A systematic random sample of 100 individuals is required from a population of 5135 individuals. The sample interval is $5135/100$ or 51.35 in other words, producing $k=51$; then a number between 1 and 51 is randomly selected (2, for example) and all the 51-th values of the list are generated. In this case, the sample is composed of the following elements: 2, 53, 104, 155, ... , 5051.

Systematic random sampling is often preferable to **simple random sampling (s.r.s.)**, since it is easier to perform in the sense that it is less time-consuming than s.r.s. which uses the lottery method. Its disadvantages, on the other hand, include the difficulty of attributing

An example application of systematic random sampling (an unknown population):

Let us suppose that we want to extract a sample of 20 purchasers at a certain commercial outlet.

We cannot use s.r.s. since we do not know the population size, therefore we have to use systematic sampling. How do we obtain our sample?

We could choose to select every fifth shopper, therefore the 5th, 10th, 15th, 20th shopper, etc., are elements of our sample.

numbers by chance, when the population is unknown. In such cases, the value j is selected by chance, but all the other elements ($j+k, j+2k, \dots$) are selected by application of a fixed interval, and are



therefore not selected at random (Hill and Hill, 2000).

Another disadvantage is that repetitions, which may skew the sample, must be taken into consideration. In the case of the control of the punctuality and timekeeping of a certain

employee, for example, the population is composed of the daily entries of the entry and exit times in the record book. Let us suppose that the employee is authorised to arrive later on Wednesdays for family reasons. If we use systematic sampling to collect the sample and if $k=7$, where Wednesday is the first day, then we will have to select Wednesdays only, which will skew the sample. Problems of this type arise whenever repetitions are associated to the population, as occurs in this case with the days of the week.

5.3.3.3 Stratified random sampling

While the two previous **sampling** techniques consider the **population** as a whole, there are situations in which sub-domains or sub-groups, resulting from the division of the **population** into groups or **strata**, can be identified (Vicente, Reis and Ferrão, 1996). One such technique is

Strata:

A sub-group of elements of the population, that is intended to be as internally homogenous as possible in regard to the characteristic under analysis.

known as **stratified sampling**, in which each **strata** is taken to be a separate **population** and the selection of the elements in each one of the strata is performed individually.

Thus, the principle of **stratified sampling** is to divide the **population** into subsets called **strata**, so that each one can be **surveyed**.

Steps to obtaining a stratified sample:

1. Define the strata. Each stratum must be quite different from any other, and the elements within each stratum have to possess common characteristics (e.g. gender, age group).
2. Select the elements within each stratum, independently from one another.
3. The elements selected from each stratum jointly constitute the sample.

This type of **sampling** is widely used, given that the majority of **populations** can be divided into **strata** (for example, men/women, higher education students/non-higher education students, etc.) and it allows the analysis of sub-groups with lower variability than those produced by s.r.s. The disadvantage of this type of sampling is that it is very expensive and time-consuming when there are many strata.

The **population** of N units is divided into sub-populations or strata, N_1, N_2, \dots, N_k elements, where $N_1 + N_2 + \dots + N_k = N$. The **strata** formed in this way are mutually exclusive and exhaustive.

The logic behind the **stratification** of a **population**, as already referred to, is the

Parameter:

A quantitative indicator of an attribute or characteristic of the population (e.g. the average age of women, the total number of small companies, etc.).

identification of groups that vary from one another, or, in other words, from the **parameter** being studied, and possess very little variation inside each grouping, in other words, each one is homogenous (Vicente, Reis and Ferrão, 1996). Each **stratum** is a separate **population**,

from which the **sample** is extracted, that supplies an estimate. The estimates obtained from k strata are the basis for the construction of estimates of the population parameter being studied.

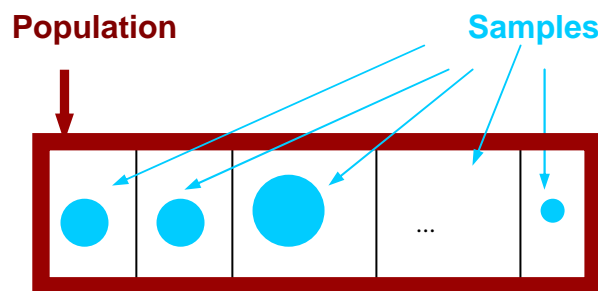


Fig. 7 - Schematic illustration of stratified random sampling

An example application of stratified random sampling:

Let us study the sales totals for services rendered of civil construction companies. We can divide the population of companies into three strata regarding the number of employees: small – 10 employees or less, medium – between 11 and 40 employees, large – more than 41 employees. Once the strata have been defined, the second stage is implemented - the collection of an s.r.s. relative to each strata. Assuming that the population under study is composed of 500 companies, 55% of which are small, 35% are medium-sized and 10% are large, and that the required sample size is 85, then we should select samples with size of 47, 30 and 8, respectively from the small, medium and large company sets. This selection maintained the ratio of sample size proportional to the set of strata sizes.



5.3.3.4 Random sampling by clusters

A **cluster** is an entity that occurs naturally associated to a reality. A school, for example (composed of different classrooms, students and teachers) can be considered to be a **cluster**.

Universities, hospitals, cities, countries and anyplace where there are replicas of the population

Cluster:

A group of elementary units of the population, ideally with the same population variability.

to study can all be deemed to be **clusters**. These groups are selected at random and all the **elements** of that group are included in the **sample**.

This type of **sampling** is preferred in many cases due to the fact that the costs are relatively lower compared to other sampling types.

Steps to obtaining a cluster sample:

1. Specify the clusters, in other words, the elements of clusters are generally close together in physical terms and therefore have very similar characteristics. Thus, the definition of very large clusters may not be in our interest.
2. Randomly select a sample of clusters and include in the sample all the elements belonging to the selected clusters.

As it is not always easy to obtain **survey bases**, the use of **cluster sampling** makes the process cheaper. **Cluster sampling** is extensively used when a **survey** of a large geographical area is required.

A good illustration of this type of **sampling** can be obtained from a bunch of grapes. If we remove one grape from the bunch we will be able to find out if the rest of the grapes of the bunch are of good quality or not, without having to eat the entire bunch. In this case, selecting all the elements of the bunch for the sample would prove to be a bit redundant.

The principle that makes **stratified sampling** efficient makes **cluster sampling** inefficient (Vicente, Reis and Ferrão, 1996). The greater the similarity between the elements of a **cluster**, the better the results generated by this **cluster** if it is used as a strata in **stratified sampling** will be, and if the same is used as a sample unit in **cluster sampling** then the results will be worse.

Example: differences between stratified and cluster sampling

Case 1: Stratified sampling

The employees of firm XYZ are grouped according to the departments in which they work (sales, marketing, research and manufacturing). Ten employees are selected at random from each group.

Case 2: Cluster sampling

Five hotels of the LÍrios chain (which is composed of 10 hotels) were selected at random. All of the employees of these five hotels were included in the sample.

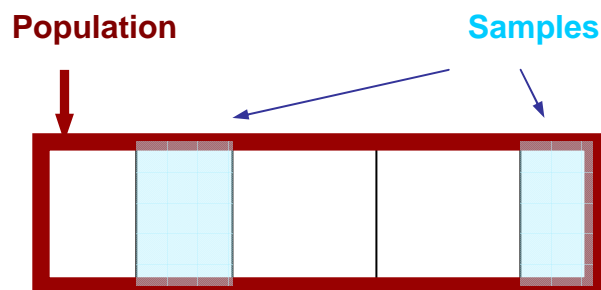


Fig. 8 - Schematic diagram of random sampling by clusters. Let us suppose that the clusters (represented by the cells) are the hotels of the abovementioned example. In this case only two of the hotels were selected from the total of five existing in the population.

5.3.3.5 Multi-stage random sampling

Multi-stage random sampling can be deemed to be an extension of **cluster sampling**, where only some of the clusters are selected and only a few elements are extracted from the groups or clusters by means of **simple random sampling**.

The table indicates some examples of **clusters** in **multi-stage sampling** (Vicente, Reis and Ferrão, 1996):

Clusters or primary sample unit	Secondary sample unit	Tertiary sample unit	Quaternary sample unit
Parish	Block	Building	Housing
Page	Line of text		
Country	Urban centre	Trading establishment	

The **sample** for the employment survey performed by INE is collected using a **multi-stage sampling** process. According to the methodology used (INE, 1998) the **population** is split into a certain number of **primary units** (parishes). Each primary unit is in turn broken down into statistical sections (adjacent geographical areas of a single parish with around 300 **homes**). Each statistical section forms a **secondary unit**. Each section lists all the home units contained in the same.

Sample unit:

An element or group of elements of the population. A sample comprises sample units based on probability methods.

B - Non-probability-based methods

Now that we have looked at some **random sampling** techniques, we shall now take a look at some **non-random sampling techniques**. According to Bacelar (1999), these techniques, contrary to **random techniques**, do not provide any "statistical guarantee" that the selected **sample** is representative. In relation to these techniques, there is no statistical theory supporting the generation of representative **samples**, but there may be a significantly high probability that a representative **sample** is obtained, even though this probability is often not very easy to ascertain. These **non-random sampling techniques** are extensively used and they are very useful when it is not possible to use **random samples**, generally in preliminary or exploratory studies.

5.3.3.6 Convenience sampling

A **convenience sample** consists of a group of individuals that are available at the time the survey is implemented. These **samples** are not representative of the **population** (Vicente et al., 1996). This type of **sampling** can be successfully used, despite its scientific fragility, in situations where ascertaining general ideas and identifying critical aspects can be more important than scientific objectivity, as is the case with **pre-tests** of a **questionnaire**. Due to the "opportunistic" nature of the sample, its elements may not be representative of the population.

An example application of convenience sampling:

In a study on the association between the income of households and access to mental health services (psycho-analysis, psychology services, etc.), a researcher placed five interviewers in front of five supermarkets and five churches in a rundown neighbourhood in the suburbs of New York.

5.3.3.7 "Snowball" sampling

This type of **sampling** involves the use of individuals previously identified as belonging to the **sample**. This technique is used in those cases where there is no available information on the **population**, or it has become impossible to make the information available. This type of sampling is used on small populations or those with very specific characteristics.

The **interviewer** will construct a **sample** based on this technique by asking the interviewee after the interview has finished to provide the names of other individuals that may also be surveyed (Vicente et al., 1996). An inconvenience with regard to this process is that the interviewees tend to indicate friends, which means that the **sample** is composed of people that think and act in the same way.

An example application of "Snowball" sampling:

To obtain a sample of drug addicts in Oporto, for which no list exists, we would have to find and interview a person corresponding to this profile. Then, on completing the interview, we would ask the interviewee to name other drug addicts living in Oporto and we would guarantee that the source of the information is not revealed.

5.3.3.8 Quota sampling

This is the most widely used **non-random sampling** method. It is very similar to **stratified random sampling**, but the elements comprising the **sample** are not **randomly selected**. This **sampling method** exists basically because of the non-existence of **population** lists (Vicente et al., 1996). **Quota sampling** generates a **sample** in which the proportion of elements possessing a certain characteristic is approximately equal to the proportion of individuals in the **population** that possess that same characteristic. For example, if the **population** has as many men as women, then the same will be true in the **sample**.

Steps to obtaining a quota sample:

1. Define the quotas, in other words, divide the population into categories. The choice of variables is mainly achieved using the population census as a basis, in a context of socio-demographic variables.
2. Select the elements. The interviewer decides who is chosen. The only obligation is that the quotas established in the sample design are kept. The selection of the elements is often planned, such as the use of pre-defined urban circuits or of formulas to find the floor and apartment numbers to be surveyed in an apartment building.

The quality of a **quota sample** depends on the way in which **interviewers** select individuals and enter into contact with them (Ghiglione and Matalon, 1992). **Interviewers** should be sent to zones selected by chance, in order to guarantee better representativeness. Once there, they can approach people passing by in the street or use the door-to-door method, or a combination of the two. The reproduction of **population** distributions must be deemed to be a necessary but not adequate pre-requisite regarding the quality of a **sample**.

The time taken to perform this type of sampling in the field is less than that needed for **random methods**, since there is no need to contact the interviewee more than once (Vicente et al., 1996). If the individual is not available on the first attempt at contact then the same is automatically substituted for another individual. This is a clear advantage in situations where there is great urgency in obtaining the data.

An example application of quota sampling:

In a survey on "who does physical exercise", we shall certainly have to take age, gender, free time, etc., into account.

The first step must involve finding out the proportions of these characteristics in the population. Let us suppose that 40% of the population is male and 60% female. Then 40% of the people interviewed must be male and 60% must be women, which is the "quota" relative to that characteristic.

There follows a table comparing some of the most widely used probability-based methods and non-probability-based methods.

Method/Description	Advantages	Disadvantages
Probabilistic methods		
Simple sampling (any set of n elements has the same probability of being selected, therefore the probability of the elements being selected is identical)	Easy to use	The members of some of the less representative groups of interest may not occur in the desired proportion.
Stratified sampling (the study population is grouped according to characteristics or strata)	Leads to analyses by sub-group with variances lower than those of simple sampling	Expensive and time-consuming when there are many strata
Systematic sampling (all x -th elements of the population are selected until the end of the ordered sample is reached, using a fixed interval. This interval is determined by dividing the size of the population by the size of the required sample)	Convenient when there is a list of names backing up the sample	Repetition patterns must be taken into consideration, as these may bias the sample.
Cluster and multi-stage sampling (all the elements of groups formed naturally that form part of the sample)	Convenient to use when there are statistical units that correspond to the specified groups (schools, hospitals, etc.)	
Non-probabilistic methods		
Convenience sampling (the use of individuals that are available)	A practical method since the research targets available units (students at school, patients in a doctor's waiting room, etc.)	The "opportunistic" nature of the sample means that its elements may not be representative of the population.
"Snowball" sampling (previously identified elements identify other members of the population)	Useful when no reference information on the population exists or when such information is hard to obtain	The resulting sample can be quite biased.
Quota sampling (the population is divided into groups, based on characteristics that are only identifiable via interview)	Practical when there is reliable information on the proportions of the attributes of interest in the population	In this process the interviewer can unwittingly cause bias in the selection of the interviewees.

Fig. 9 - Some of the most widely used probability-based and non-probability-based methods – summary table

5.3.4 How to determine sample size

The question of **sample** size is always an important decision in the **survey** process. Two very important factors must be taken into account at this phase: the required **precision** of the results (since there is always an error, which you want to be as small as possible) and **time and cost constraints** relative to the **survey**.

We also have to take into account the fact that increasing the size of the **sample** not only increases precision but also increases the cost. We must therefore equilibrate the two.

The size of the **sample** required to provide specific precision in the results can only be mathematically calculated if the **samples** are randomly selected. Otherwise, we have, according to Weiers (1998), three options: employ the size already successfully used in

previous studies with the same characteristics, take the available funds for the study into consideration as well as the costs involved and, lastly, assume that the **sample** is **random** and calculate the necessary dimension, in which case the value produced is merely for information purposes. A **sample** must be representative of the **population**, which means that it must contain the typical aspects, since the **sample** is a miniature model of the **population**. It must be noted that the size of the sample to collect is not directly proportional to the population size and that this dimension fundamentally depends on the variability existing in the population. In relation to a population composed of 10th grade students of a secondary school, for example, if we want to study the average age then the size of the sample to be collected need not be very large since the age variable has similar values in a restricted age group. However, if the characteristic to be studied is the average travel time of students from home to school, then the sample size will have to be much greater, since the population variability is much larger. Each student may take a different length of time. In the extreme situation where all the elements have the same value for the variable under study, then a sample composed of one element is all that is required to provide full information details for the population; if, however, the variable has different values for all of the elements, then it would be necessary investigate all of the elements to obtain the same level of information (Graça Martins, 2001).

Example: Determining sample size in an estimation problem relative to the proportion (p)

Ascertain the real proportion (p) of individuals with income below PTE 1,000,000 per year in a Portuguese region. The confidence interval for a proportion is as follows (assuming that sample size is greater than 100):

$$\left[f - c \sqrt{\frac{f(1-f)}{n}}; f + c \sqrt{\frac{f(1-f)}{n}} \right]$$

Thus, the sample size is ascertained by inserting the required amplitude (A) and confidence level.

$$n = \frac{4c^2 f(1-f)}{A^2}$$

where:

c = parameter determined by the required confidence level

n = sample size

f = relative frequency of the attribute in the sample (sample proportion)

Some notes on estimating the proportion p

1. In a population of size N , p is the (unknown) proportion of population elements that possess a certain characteristic. To estimate this proportion (p), a sample of size n is collected and the proportion p' of elements in that sample possessing the studied

characteristic is calculated. The estimator p' is a good estimator of p . It also has some interesting properties, the most significant of which is that its variance (a measure of the variability between p and p') is equal to $\frac{p(1-p)}{n} \left(\frac{N-n}{N-1} \right)$.

Note that if the size n of the sample is very small in comparison to the size of the population, $N-n$ is approximately equal to $N-1$, and it is thus the first factor of the expression that measures variability. This is the reason why it is said that “when the size of the population is very large in comparison to the size of the sample, it can be deemed to be infinite”.

2. Confidence interval for proportion p

Irrespective of how the confidence interval is arrived at, the shape of the confidence interval for p , with a confidence of $100(1-\alpha)\%$ (α is a value normally deemed to be around 0.05, and is therefore used to represent 95% confidence!) is

$$(p' - z_{\alpha} \sqrt{\frac{p'(1-p')}{n}}, p' + z_{\alpha} \sqrt{\frac{p'(1-p')}{n}})$$

$$\text{Interval amplitude} = 2 z_{\alpha} \sqrt{\frac{p'(1-p')}{n}}$$

The quantity $z_{\alpha} \sqrt{\frac{p'(1-p')}{n}}$ is called the margin of error or precision of the survey.

3. What sample size must be collected in order to obtain an interval with a specific precision d and a $100(1-\alpha)\%$ confidence interval?

We will have to solve the following equation with regard to n :

$$z_{\alpha} \sqrt{\frac{p'(1-p')}{n}} < d$$

$$n > \left(\frac{z_{\alpha}}{d} \right)^2 p'(1-p')$$

As p' is only known after we collect the sample, we have to provide for the maximum value of $p'(1-p')$ which occurs when $p'=1/2$, hence

$$n > \left(\frac{z_{\alpha}}{2d} \right)^2$$

The table below shows the values of z_{α} , for some values of α :

Confidence $100(1-\alpha)\%$	z_α
90%	1.645
95%	1.960
98%	2.326
99%	2.576

Example: Determine the level of trust that the general public has in teachers. The estimate of the level of trust must have a confidence level of 95% and a maximum margin of error of 0.05. What is the size of the sample that must be collected?

$$n > \left(\frac{1.96}{2 \times 0.05} \right)^2$$

$$n = 385$$

If a margin of error of 0.02 is required for the same confidence level, then you can see that the sample size is much greater - it must be equal to 2401!

5.3.5 Selecting the elements of the sample

There are different ways of selecting the **elements** of a **sample**, as we have seen in the previous sections. In **random samples** the selection scheme used objectively designates the **element** to be chosen. In such cases, the existence of prior lists containing the details of the elements included in the sample makes it possible to identify each interviewee and establish contact (face-to-face, via telephone or via post) in order to set the data collection process in motion. In INE's employment survey, for example, the interviewees are contacted by post, followed by a set of face-to-face visits from interviewers. If the **sample** is **not random**, the interviewer has to select the elements to be included and this is performed based on human judgement, due to the lack of a survey base (Vicente, Reis and Ferrão, 1996). Quota sampling, on the other hand, uses guides or plans, which serve as good aids since they help the interviewer to introduce a certain degree of randomness into the interviewee selection

process. These guides or plans contain formulas for selecting streets in a parish or for selecting homes in a building.

6. Collecting the Necessary Information from Sample Elements

Once the **sample elements** have been selected then they must be contacted in order to obtain the necessary data to permit the achievement of the study's objective. There are essentially three methods of collecting data in a **poll-based study**: **face-to-face interviews**, **telephone interviews** and **postal questionnaires**. Each one of these methods has its advantages and disadvantages, as set out below.

6.1 Face-to-face interview

The face-to-face interview can be considered to be a conversation between two people, initiated and controlled by the **interviewer**, with the express purpose of obtaining relevant data, thereby permitting the attainment of the study's objectives (Mayer, 1974). For a long time this method for collecting data was the most used, and it remains quite important in certain **surveys** performed by INE. This data collection method can be quite expensive since the **interviewer** must be trained beforehand and then the interviewer has to travel to the interviewee's town to conduct the interview. These trips often have to be done a number of times because the interviewees are not home or because they are not available to fill in the **questionnaire** at that particular time. A refusal often occurs, which makes this method more costly than the following two methods. According to Aaker and Day (1990) only 30% to 40% of the **interviewer's** time is spent on the interview itself, the rest of the time is spent travelling, locating interviewees, etc. This method does have advantages over the **postal questionnaire** method, such as the fact that the interview can be achieved in a few minutes while an interview by **postal questionnaire** can take weeks to complete. The response rate is much higher in a **face-to-face interview**, due to the fact that the **interviewer** provides greater incentive to the interviewee to respond.

Interviewer:

The person responsible for the collection of data in order to meet the specific objectives of each study, performing the interviews in accordance with the established rules.

6.2 Telephone interview

The **telephone interview** is an alternative to the **face-to-face interview**. The **interviewer** questions the interviewee by telephone and collects the data by this means, just as the name indicates. This method is often cheaper than the face-to-face method and it is more advantageous if we consider, for example, the fact that it is not necessary to visit **homes** several times to perform the interview. The time spent with a **telephone interview** is less than that of a **face-to-face interview**. But there are also some disadvantages. If the **questionnaire** is lengthy then telephone call costs will be closer to the costs of **face-to-face interviews**, in addition to the fact that the interviewee will become tired more quickly.

6.3 Postal questionnaire

This method's format means that the respondent to the questionnaire must provide the responses, after having read the questions and accompanying instructions, without the aid of an **interviewer** (Vicente, Reis and Ferrão, 1996). This method is recommended for geographically spread-out populations. The cost of data collection is reduced. The **questionnaires** are pre-tested a number of times to ensure that the questions are fully comprehended and understood in the same way by all respondents. Despite the reduced costs, the time variable is not always favourable. If a rapid response is required then this method is not recommended. In addition, the rate of no-responses in this type of survey may usually be higher than that of other methods.

7. Data Organisation and Display

After defining the problem to be studied, planning the survey and collection of data, the problem of **data organisation** is encountered. The organisation of data comprises ‘summarising’ the data obtained in a simple and clear way that best provides for their interpretation. Data can be displayed in a number of ways. In an initial approach, data can be displayed in frequency tables, bar charts, pie charts histograms, etc. Further information of data organisation in regard to introductory descriptive statistics can be obtained from the Dossiers on Statistics with Excel (only in the Portuguese version) and Graphical Representations. The latter is available at the webpage:

<http://www.alea.pt/english/html/statofic/html/dossier/html/dossier9.html>

Lastly, the presentation format of the final report must be taken into consideration. According to Hill and Hill (2000) there exist various types of report, such as the academic and internal report, for example. Both have similar layouts and contain the items set out below.

7.1 Some recommendations

Any report must possess a title stating the content presented in the report. The contents must contain all of the chapters contained in the report. These must be numbered and indicate the start page.

Even though the summary is the first part of the report, it is usually only written after all the other components of the report have been written, reviewed, "polished" and are available in their final version. (Hill and Hill, 2000). The summary must contain the reason that led to the investigation being conducted, how it was done, the most significant results and the conclusions drawn regarding the investigation and how these may aid in solving the problem. The objective of the introduction is to explain the type of investigation and the underlying reasons. It must provide a brief overview of the other report chapters.

7.2 The results

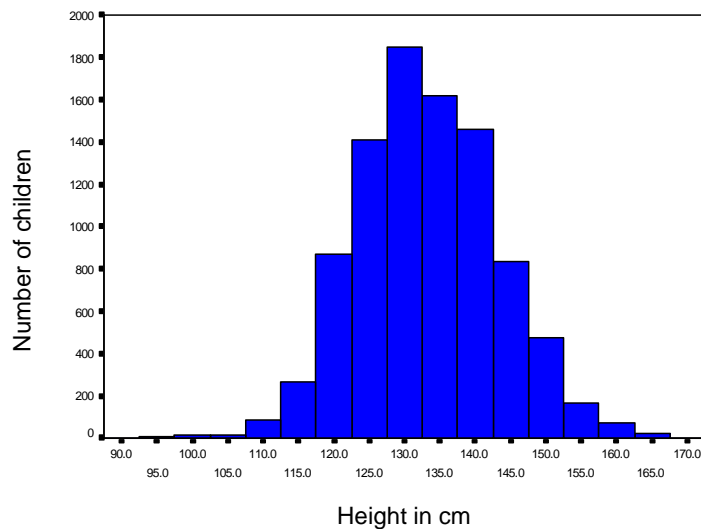
Numerical results can be displayed in a number of ways. A preliminary exploratory analysis of the results must always be made, with particular focus on a summary of the main variables that were analysed.

In the ‘mini-censuses’ performed by INE, for example, one of the analysed variables was the height of individuals.² In the report presenting the results of this work, one of the tables contains a descriptive summary of this variable:

Height

	N	Minimum	Maximum	Mean	Standard deviatio
HEIGHT	9171	92	170	133.21	9.917

The same data were also graphically displayed in a histogram³.



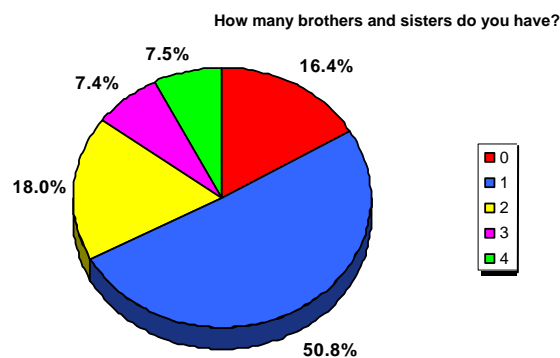
² One of ALEA’s main projects in 2001 was that of ‘mini-censuses’ performed on primary schools. Primary schools were surveyed and all the collected data was organised and processed by a joint team that included technicians of INE and Sociedade Portuguesa de Estatística [Portuguese Statistics Society]. The main purpose of the ‘mini-censuses’ were to familiarise students with what a census is, what it is used for and how it is performed. The report and results of the mini-census are available at the ALEA’s Portuguese version:
<http://www.alea.pt/Html/statofic/html/censos2001/html/censos2001.html>

³ See histogram construction rules in ALEA’s Statistical Concepts at
http://www.alea.pt/english/html/nocoes/html/cap3_2_10.html

For the ‘Number of siblings’ variable, the following frequency table and corresponding pie chart were generated.

	Absolute Frequency	Relative Frequency (%)	Accumulated Relative Frequency (%)
Number of siblings 0	1403	16.4	16.4
1	4356	50.8	67.1
2	1540	18.0	85.1
3	636	7.4	92.5
4 or more	643	7.5	100.0
Total	8578	100.0	
Did not respond	593		
Total	9171		

This table displays the number of siblings that each child has. We can see that close on half the children that responded to this question have more than one brother or sister and that 16% are only children. Some 18% of children have two siblings and the rest have three or more.



According to Hill and Hill (2000) we must have the target audience in mind when displaying the results, so that the most suitable display method is chosen. When the target audience is used to reading and interpreting tables then they must be used in a manner that facilitates their interpretation. When, on the other hand, the target audience is not used to reading and interpreting tables then graphs must be used to display the most important information. Both data display methods must have an associated text explanation to improve viewers’ understanding. Any graphs and tables used must be numbered and must possess a title.

8. See Also...

Publications

- ALEA, Estatística com Excel [*Statistics with Excel*], Educational Dossier no. IV, (only in Portuguese) available at:
http://www.alea.pt/html/statofic/html/dossier/html/meio_dossier4.html
- ALEA, *Graphical representation - Notes on the creation and presentation of several types of charts*, Educational Dossier, available at:
<http://www.alea.pt/html/statofic/html/dossier/html/dossier11.html>
- BACELAR, S. (1999), Relatório de Aula Teórico-Prática sobre Amostragem nas *Ciências Sociais* [Report of Theory-Practical Lesson on Sampling in Social Sciences], FEP - Economics Faculty, Oporto, Universidade do Porto [Oporto University];
- CAMPOS, P. (2000), *Módulo 2 - da Conceção ao Tratamento Estatístico de Questionários* - Apontamentos do curso de Análise Estatística de Dados com SPSS [Module 2 - From the Design to Statistical Processing of Questionnaires - Notes of the Statistical Analysis of SPSS Data course]. Escola Superior de Biotecnologia da Universidade Católica [Biotechnology School of the Portuguese Catholic University], Oporto;
- GHIGLIONE, R. and MATALON, B. (1992), *O Inquérito, Teoria e Prática* [Surveys, Theory and Practice], Oeiras, Celta Editora;
- GOMES, P. (1998), *Tópicos de Sondagens* [Survey Topics], (Course forming part of the 6th Congress of the Portuguese Statistics Society, in Tomar, 9 to 12 June 1998);
- GRANGÉ, D. and LEBART, L. (1994), *Traitements Statistiques des Ênquetes*, Paris, Edições Dunod;

- HILL, M. M. and Hill, A. (2000), *Investigação por Questionário* [Investigation by Questionnaire], Lisbon, Edições Sílabo;
- INE (1998), Employment Survey - Series - 1998; also available on the internet as part of the publication relative to the 1998 first quarter employment statistics;
- LIMA, M. P. (1981), *O Inquérito Sociológico - Problemas de Metodologia* [Sociological Surveys - Methodology Problems], 2nd Edition, Editorial Presença;
- MARTINS, E. G., (2001), *Noções Básicas sobre Amostragem - Introdução à Inferência Estatística* [Basic Sampling - Introduction to Statistical Inference], Department of Statistics and Operational Research, Faculty of Sciences, Universidade de Lisboa [Lisbon University];
- STUART, A., (1984), *The Ideas of Sampling, Monograph no. 4*, Charles Griffin and Company Ltd, London;
- VICENTE, P., REIS, E. and FERRÃO, F. (1996), *Sondagens - A amostragem como factor decisivo da qualidade* [Surveys - Sampling as a decisive factor in quality], Lisbon, Edições Sílabo;
- WEIERS, R.M. (1998), *Marketing Research*, 2nd Ed., Prentice-Hall, London.

Websites:



<http://www.socio-estatistica.com.br/>



<http://www.fecap.br/portal/index.asp>

These two sites contain a number of suggestions relative to the construction of questionnaires and some bibliographical references.