Data Analysis

Statistics and Demography Programme
Secretariat of the Pacific Community
Noumea, New Caledonia
2009

TOPIC 1
INTRODUCTION

Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.

H.G. Wells

Introduction

understanding statistics

This course is designed primarily to provide a basic understanding of statistical analysis and how to present the results of analysis in a report. Professional literature regularly includes results that are based on statistical analysis. Regardless of your areas of professional interest, you will be unable to understand many relevant areas unless you have some basic knowledge of statistical analysis. The methods underlying statistical analysis are not just a catalogue of statistical procedures and formulae, rather they offer the rationale upon which scientific research is often based.

As we study the subject of statistics, we need to establish definitions for particular words and concepts. In everyday use, these words may have fairly wide meanings and may be used in very different situations. In statistics, words need to be defined quite carefully so that everyone agrees with what we are talking about.

Why do we analyse data?

The **survey population** is that set of units from which information is to be collected. The survey population is also known as the **coverage** of the collection. Naturally it is desirable for the survey population to coincide with the target population, but this is sometimes not possible. For example, the target population of a population census may be all people in a country, but in many countries this may not be practical since people on remote islands may be too expensive to survey. Therefore, the survey population may be reduced to those persons on islands which cover a very high percentage of the total (or target) population. The results of the collection can then only be claimed to relate to the survey population and it is therefore important that users of the survey results are informed what the survey population covers.

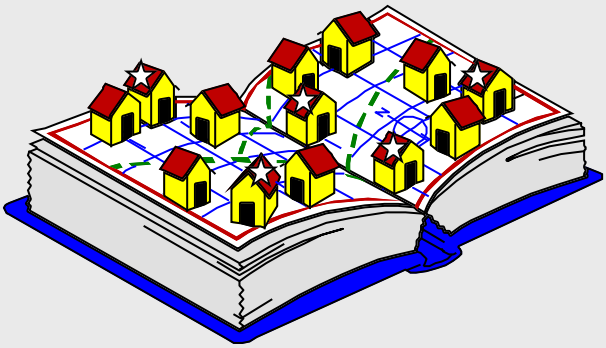
The **target population** is the scope of the survey …
the set of units we want to collect information from.

The **survey population** is the coverage of the collection …
the set of units we do collect information from.

two types of surveys

The terms **census** and **sample** refer to the extent of data collection undertaken. In a census, the data is collected from all units in the survey population. Therefore, as long as the population is well defined and the data is collected accurately from every unit, an accurate result will be obtained. In a sample, the required data is collected from a subset of units in the population and this data is used to estimate results for the population. This introduces an error called sampling error, but sampling errors can be

controlled and sampling is often preferable due to reduced costs and increased timeliness.



In a census, all of these houses would be surveyed. In a sample survey, however, only the ones with the star on the roof would be surveyed.

ADVANTAGES OF SAMPLE SURVEYS COMPARED WITH CENSUSES

- Reduces cost — both in terms of money and personnel requirements;
- Reduces time needed to produce results;
- Enables data to be collected more accurately (less non-response, collection/processing error); and
- Enables more detailed questions to be asked.

DISADVANTAGES OF SAMPLE SURVEYS COMPARED WITH CENSUSES

- Results (or estimates) are subject to sampling error;
- Detailed cross-tabulations and results for small geographic areas and sub-populations may be too inaccurate to be useful;
- Does not provide sampling frames for use in the future; and
- Difficulty in communicating results to users.

INFORMATION BOX 1: Censuses — A Global History

BC	3800	**BABYLON**	❒ carried out every six or seven years; it counted asses, oxen, butter, milk, honey and wool.
	2500	**EGYPT**	❒ carried out to assess the labour force available for building pyramids.
	1491	**ISRAEL**	❒ carried out to count people liable for military service and for taxation purposes.
	550	**CHINA**	❒ carried out by Confucius to obtain information on the nation's agricultural, industrial and commercial state.
AD	1719	**PRUSSIA**	❒ Europe's first systematic census.
	1790	**USA**	❒ America's first census.
	1801	**ENGLAND FRANCE**	❒ England and France's first censuses.

The decision whether to undertake a census or a sample often involves some difficult choices.

anaylysing survey data

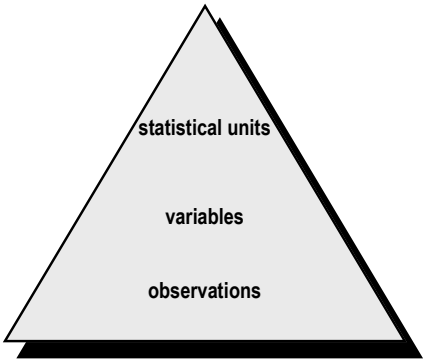
Once you have conducted your survey the data must be analysed. Statistical analysis is often about

simplification and summarisation of data. Methods for describing sets of measurements fall into two categories: graphs and numbers. In this course we learn about both of these techniques.

The statistical process

its an ordered world

Statistical data are the results of making observations of variables from statistical units. The three terms observation, variables and statistical units all have a special meaning in statistics. We shall define each of them in turn and then have a look at some examples.



observe and record information

An **observation** is the result of observing and recording information from a particular source. Usually an observation will be carried out by a person, but it could also be carried out by a machine (for example, a machine may automatically record the temperature at a particular location at a particular time each day). Some examples of the more common ways of making observations are:

measuring – measuring the heights of school children with a measuring tape;

counting – counting visitors to a tourist attraction;

recording answers to interviews – asking questions of people and recording their responses;

mailing out questionnaires – sending a form to businesses to determine their employment numbers;

copying from existing records – referring to customs records and copying import figures; and

using self-recording instruments – measuring temperatures on an outer island using an automatic machine.

and then there are variables

An observation will be made of the value of a variable for a unit in the population. The term **variable** means a feature or property of the units in the population and, of course, a unit may have many different variables, all or some of which we may observe. For example, if a police force is recruiting people, the police force must observe the values of several variables of each applicant in order to select the right people. Thus:

Statistical Units	*Variables*	*Observations*
Police Applicants	Age	years lived
	Height	cm tall
	Weight	kg heavy
	Eyesight	vision accuracy
	Freedom from disease	vital signs e.g. blood tests
	Marital status	legal marital status
	Academic record	pass rate
	Criminal record	crimes committed
	Previous work experience	employment history

observations from statistical units combine to form statistical data

The values of variables are observed from **statistical units**. Statistical units are not only people, they are any item in the population of interest. Examples of statistical units are: objects, groups of objects, people, families, households, geographical areas, buildings, periods of time and countries.

Values for one or more variables on a number of statistical units from the same population make up a series of statistical data. Thus, the total value of exports of a country for a number of years will form a series, with years as statistical units, while the value of exports of different commodities of one year will form another series with commodities as the statistical units.

TIP

In any statistical investigation, it is very important to clearly define the population of statistical units about which data will be collected, the values of what variables will be observed and how observations will be made.

... Exercises ...

To better understand some of the definitions and concepts that have been introduced, consider these examples. Write your answers in the spaces provided. The answers are provided below the examples.

1. In your own words define:

a) data _____

_____

b) information _____

_____

c) statistics _____

_____

_____

2. For each of the following situations try to identify how the observations are made, the variable or variables being observed, and the statistical unit.

(a) The sales manager of a local brewery finds out from copies of invoices the total volume in gallons and the value in dollars of the sales of beer for each month during 1995.

Method of observation	**Variable(s) observed**	**Statistical unit**

(b) An interviewer visits a village and selects three households. For each of these households, he uses a tape measure and compass to measure the total area of coconut plantation they own.

Method of observation	**Variable(s) observed**	**Statistical unit**

(c) An agricultural researcher uses his instruments to record the total rainfall in millimetres, and the maximum and minimum temperature every day.

Method of observation	**Variable(s) observed**	**Statistical unit**

(d) A statistical investigator visits the same shops at the same time each month and records the prices of the same products.

Method of observation	Variable(s) observed	Statistical unit

(e) People arriving at every point of entry into a country complete an arrival card. On it they state their sex, age, citizenship, country of residence, reason for the visit and their expected length of stay.

Method of observation	Variable(s) observed	Statistical unit

… Self-Review …

Each topic has a self review and these are designed to help you understand the topic and apply it to your project work. You should complete the self review and then check your answers against those provided.

1. In Kiribati[1] over a period of two weeks, the Department of Labour randomly selected 90 households in Tarawa and interviewed the permanent residents aged 15 years and over about their labour force status and main activities in a Tarawa Household Labour Force Survey. A total of 355 adults were surveyed, and 139 were working (full- or part-time), 14 were unemployed and seeking work and 202 were not in the labour force.

(a) What would the Department of Labour report to the Minister regarding the structure of the labour force in Kiribati?

_____

_____

_____

_____

(b) Is the 355 adults a sample or a population?_____

(c) What is the target population in this problem?_____

(d) What is the survey population in this problem? _____

(e) What is the method of observation? _____

(f) What variables were likely to be observed? _____

(g) What is the statistical unit? _____

2. Identify three current political, economic or social issues for which information is necessary. Then describe the information needed for each issue.

(a) _____

_____

_____

_____

[1] *Fictional data only.*

(b)

(c)

3. Suggest reasons why data would be collected on the following topics:

(a) burglaries

(b) causes of death

(c) climate

(d) fish stocks

(e) immigration

(f) schools