E1 *Data Collection* Introduction

Learning objectives

Here we concentrate on the topic of *data collection* and also introduce *data presentation*. After completing Unit E1 you should be able to

- understand how to produce unbiased questions
- appreciate the need for random samples
- use and apply tables such as distance charts
- present data clearly in the form of pictograms and charts.

The British statesman, *Benjamin Disraeli*, once stated,

"There are lies, damned lies and statistics."

This might have been a bright remark to make in a debate, but the status of statistics was given a quite underrated currency which persists, even today! Disraeli's statement is in fact the reverse of the truth, and a more apt expression would be,

"There are lies, damned lies and numerical statements made by people ignorant of statistics!"

Statistical analysis makes possible the testing of numerical data for relevance, reliability and validity. Statisticians must present data in such a form that others can utilise the relevant information to enable them to make judgements.

The study of Statistics is reported to have started with the Englishman, *John Graunt* (1620 – 1674), who collected and studied the death records in various cities of Britain. He was fascinated by the patterns he found in the whole population.

Much of current day statistical analysis is of quite recent development, the availability of cheap computing power acting as a catalyst for the development of appropriate ways of presenting and analysing data. In fact, the more advanced statistical analyses and tests are based on probability theory, developed over the past few centuries, but put into a more modern context by mathematical statisticians such as

Karl Pearson (1857 – 1936), founded the world's first University Statistics Department at University College London.



The English statistician *Sir Ronald Fisher* (1890 – 1962), described as "a genius who almost single-handedly created the foundations for modern statistical science".



The Polish-American mathematician *Jerzy Neyman* (1894 – 1981), made large contributions in probability theory, testing hypothesis, confidence intervals and other areas of mathematical statistics.



You can find interesting bibliographies of these people on the internet.

© CIMT, Plymouth University

E1 *Data Collection* Introduction

One of the key topics introduced in the unit is that of a *pictogram*. A pictogram (also called a pictograph or pictogramme) is an ideogram (a graphic symbol that represents an idea or concept that conveys its meaning through its pictorial resemblance to a physical object). Earliest examples include ancient or prehistoric drawings or paintings found on rock walls. Pictograms are also used in writing and graphic systems in which the characters are to a considerable extent pictorial in appearance.

Early written symbols were based on pictographs (pictures which resemble what they signify) and ideograms (symbols which represent ideas). They were used by the ancient Chinese culture since around 5000 BC and began to develop into logographic writing systems around 2000 BC. Pictographs are still in use as the main medium of written communication in some non-literate cultures in Africa, The Americas, and Oceania. Pictographs are often used as simple, pictorial, representational symbols by most contemporary cultures.

Pictographs from the Great Gallery, Canyonlands National Park, Utah, USA



Pictographs can often transcend languages in that they can communicate to speakers of a number of tongues and language families equally effectively, even if the languages and cultures are completely different. This is why road signs and similar pictographic material are often applied as global standards expected to be understood by nearly all.

The use of pictograms in Statistics, as in this unit, can also take the form of diagrams to represent statistical data by pictorial forms, and can be varied in colour, size, or number to indicate change.

This pictogram is a bar chart that uses pictures rather than bars to represent data.

The pictogram represents the number of TV units each of the cities listed has in stock. As referenced in the footing of the report, each of the TV icons represents 200,000 units. This is a summary report meant to quickly depict information – so the numbers will be rounded slightly – but the same information is presented.



Key points and principles

- Questions should be straightforward and not biased.
- Care must be taken to ensure that samples are representative.
- Data presentation should be

  clear

  straightforward

  unambiguous

  not biased
- Pictograms and bar charts are simple and effective ways of presenting data.

© CIMT, Plymouth University

E1 *Data Collection* Introduction

Facts to remember

- Questionnaires should be straightforward and only have unbiased questions
- Samples should be unbiased.

Glossary of terms

Population - this is the variable (e.g. age, height, weight, etc.) that you are investigating

Sampling - a number of members taken from a population; you take a sample from the population to estimate a particular characteristic of the population

Random sample - here every member of the population has an equal chance of being in the sample

Systematic sample - here the population is numbered, e.g. 1 → 100, and you take, for example, every 10th number (e.g. 1, 11, . . . , 91) to give a sample of, in this instance, size 10

Stratified sample - here the population has distinct groups; you need to ensure that all groups are fairly represented in the sample

Distance chart - a matrix of values which shows distances between specified places

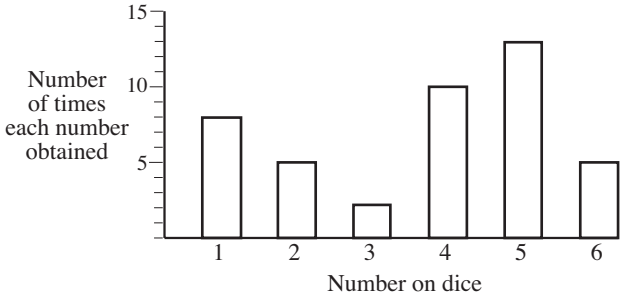
Pictogram - a pictorial representation of data. For example,

if  represents 10 balls,

then  will represent 5 balls, etc.

Bar chart - here the frequency of each event is illustrated on the *y*-axis. For example,

Bar chart showing the number of times each number, 1 - 6, was obtained when a dice was thrown



© CIMT, Plymouth University