CHAPTER 1  DESCRIPTIVE STATISTICS

Contents

<table>
<thead>
<tr>
<th>Contents</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Introduction</td>
<td>2</td>
</tr>
<tr>
<td>1.2 Some Basic Definitions</td>
<td>2</td>
</tr>
<tr>
<td>1.3 Method of Data Collection</td>
<td>3</td>
</tr>
<tr>
<td>1.4 Primary and Secondary Data</td>
<td>6</td>
</tr>
<tr>
<td>1.5 Graphical Descriptions of Data</td>
<td>7</td>
</tr>
<tr>
<td>1.6 Frequency Distribution</td>
<td>11</td>
</tr>
<tr>
<td>1.7 Central Tendency</td>
<td>16</td>
</tr>
<tr>
<td>1.8 Dispersion and Skewness</td>
<td>20</td>
</tr>
<tr>
<td>Exercise</td>
<td>29</td>
</tr>
</tbody>
</table>

Objectives: After working through this chapter, you should be able to:

(i) get familiar with some of the statistical terminology;
(ii) represent data graphically;
(iii) understand basic frequency distribution;
(iv) measure the central tendency of a given set of grouped or ungrouped data;
(v) evaluate the dispersion and skewness of a given set of ungrouped or grouped data
1.1 Introduction

Statistics is concerned with the scientific method by which information is collected, organised, analysed and interpreted for the purpose of description and decision making.

Examples using statistics are: *Hang Seng Index, Life or car insurance rate, Unemployment rate, Consumer Price Index*, etc.

There are two subdivisions of statistical method.
(a) Descriptive Statistics - It deals with the presentation of numerical facts, or data, in either tables or graphs form, and with the methodology of analysing the data.

(b) Inferential Statistics - It involves techniques for making inferences about the whole population on the basis of observations obtained from samples.

1.2 Some Basic Definitions

(a) Population - A population is the group from which data are to be collected.

(b) Sample - A sample is a subset of a population.

(c) Variable - A variable is a feature characteristic of any member of a population differing in quality or quantity from one member to another.

(d) Quantitative variable - A variable differing in quantity is called quantitative variable, for example, the weight of a person, number of people in a car.

(e) Qualitative variable - A variable differing in quality is called a qualitative variable or attribute, for example, color, the degree of damage of a car in an accident.

(f) Discrete variable - A discrete variable is one which no value may be assumed between two given values, for example, number of children in a family.

(h) Continuous variable - A continuous variable is one which any value may be assumed between two given values, for example, the time for 100-meter run.
1.3 Method of Data Collection

Statistics very often involves the collection of data. There are many ways to obtain data, and the World Wide Web is one of them. The advantages and disadvantages of common data collecting method are discussed below.

1.3.1 Postal Questionnaire

The principal advantages are:
- The apparent low cost compared with other methods although the cost per useful answer may well be high.
- No need for a closely grouped sample as in personal interviews, since the Post Office is acting as a field force.
- There is no interviewer bias.
- A considered reply can be given - the respondent has time to consult any necessary documents.

The principal disadvantages are:
- The whole questionnaire can be read before answering (which in some circumstances it is undesirable).
- Spontaneous answers cannot be collected. Only simple questions and instructions can be given.
- The wrong person may complete the form.
- Other persons' opinions may be given e.g. by a wife consulting per husband.
- No control is possible over the speed of the reply.
- A poor "response rate" (a low percentage of replies) will be obtained.

The fact that only simple questions can be asked and the possibility of a poor response rate are the most serious disadvantages and are the reasons why other methods will be considered. Only simple questions can be asked because there is nobody available to help the respondent if they do not understand the question. The respondent may supply the wrong answer or not bother to answer at all. If a poor response rate is obtained only those that are interested in the subject may reply and these may not reflect general opinion. The postal questionnaire has been used successfully on a number of topics by the Social Survey Unit, and in the U.S.A. there are a number of market research companies who specialise in this technique.

1.3.2 Telephone Interviewing

The main advantages are:
- It is cheaper than personal interviews but tends to be dearer on average than postal questionnaires.
- It can be carried out relatively quick.
- Help can be given if the person does not understand the question as worded.
• The telephone can be used in conjunction with other survey methods, e.g. for encouraging replies to postal surveys or making appointments for personal interviews.
• Spontaneous answers can be obtained.

The main disadvantages are:
• In some countries not everybody owns a telephone, therefore, a survey carried out among telephone owners would be biased towards the upper social classes of the community. But the telephone can be used in industrial market research anywhere since businesses are invariably on the telephone.
• It is easy to refuse to be interviewed on the telephone simply by replacing the receiver. The response rate tends to be higher than postal surveys but not as high as when personal interviews are used.
• As in the postal questionnaire, it is not possible to check the characteristics of the person who is replying, particularly with regard to age and social class.
• The questionnaire cannot be too long or too involved.

1.3.3 The Personal Interview

In market research this is by far the most commonly used way of collecting information from the general public.

Its main advantages are:
• A trained person may assess the person being interviewed in terms of age and social class and area of residence, and even sometimes assess the accuracy of the information given (e.g. by checking the pantry to see if certain goods are really there).
• Help can be given to those respondents who are unable to understand the questions, although great care has to be taken that the interview's own feelings do not enter into the wording of the question and so influence the answers of the respondents.
• A well-trained interviewer can persuade a person to give an interview who might otherwise have refused on a postal or telephone enquiry, so that a higher response rate, giving a more representative cross-section of views, is obtained.
• A great deal more information can be collected than is possible by the previous methods. Interviews of three quarters of an hour are commonplace, and a great deal of information can be gathered in this time.

Its main disadvantages are:
• It is far more expensive than either of the other methods because interviewers have to be recruited, trained and paid a suitable salary and expenses.
• The interviewer may consciously or unconsciously bias the answers to the question, in spite of being trained not to do so.
• Persons may not like to give confidential or embarrassing information at a face-to-face interview.
• In general, people may tend to give information that they feel will impress the interviewer, and show themselves in a better light, e.g. by claiming to read "quality" newspapers and journals.

• There is a possibility that the interviewer will cheat by not carrying out the interview or carrying out only parts of it. All reputable organisations carry out quality control checks to lessen the chances of this happening.

• Some types of people are more difficult to locate and interview than others, e.g. travellers. While this may not be important in some surveys, it will be on others, such as car surveys. One particular problem is that of the working housewife who is not at home during the day: hence special arrangements have to be made to carry out interviews in the evenings and at weekends.

1.3.4 Observation

This may be carried out by trained observers, cameras, or closed circuit television. Observation may be used in widely different fields; for example, the anthropologist who goes to live in a primitive society, or the social worker who becomes a factory worker, to learn the habits and customs of the community they are observing. Observation may also be used in "before and after" studies, e.g. by observing the "traffic" flow in a supermarket before and after making changes in the store layout. In industry many Work Study techniques are based upon observing individuals or groups of workers to establish the system of movements they employ with a view to eliminating wasteful effort. If insufficient trained observers are available, or the movements are complicated, cameras may be used so that a detailed analysis can be carried out by running the film repeatedly. Quality control checks and the branch of market research known as retail audits may also be regarded as observation techniques.

The advantages of the observational technique are:
• The actual actions or habits of persons are observed, not what the persons say they would do when questioned. It is interesting to note that in one study only 40% of families who stated they were going to buy a new car had actually bought one when called upon a year later.
• Observation may keep the system undisturbed. In some cases it is undesirable for people to know an experiment or change is to be made, or is taking place to maintain high accuracy.

The main disadvantages are:
• The results of the observations depend on the skill and impartiality of the observer.
• It is often difficult in practice to obtain a truly random sample of persons or events.
• It is difficult to predict future behaviour on pure observation.
• It is not possible to observe actions which took place before the study was contemplated.
• Opinions and attitudes cannot usually be obtained by observation.
• In marketing, the frequency of a person's purchase cannot be obtained by pure observation. Nor can such forms of behaviour as church-going, smoking and crossing roads, except by employing a continuous and lengthy (and hence detectable) period of observation.

1.3.5 Reports and Published Statistics

Information published by international organisations such as the United Nations Organisation gives useful data. Most governments publish statistics of population, trade, production etc. Reports on specialised topics including scientific research are published by governments, trade organisations, trade unions, universities, professional and scientific organisations and local authorities. The World Wide Web is also an efficient source of obtaining data.

1.4 Primary and Secondary Data

Before considering whether to investigate a data collection exercise at all it is wise to ascertain whether data which could serve the purpose of the current enquiry is already available, either within the organisation or in a readily accessible form elsewhere.

When data is used for the purpose for which it was originally collected it is known as primary data; when it is used for any other purpose subsequently, it is termed secondary data. For example, if a company buyer obtains quotations for the price, delivery date and performance of a new piece of equipment from a number of suppliers with a view to purchase, then the data as used by the Buyer is primary data. Should this data later be used by the Budgetary Control department to estimate price increases of machinery over the past year, then the data is secondary.

Secondary data may be faced with the following difficulties:
• The coverage of the original enquiry may not have been the same as that required, e.g. a survey of house building may have excluded council built dwellings.
• The information may be out of date, or may relate to different period of the year to that required. Intervening changes in price, taxation, advertising or season can and do change people's opinions and buying habits.
• The exact definitions used may not be known, or may simply be different from those desired, e.g. a company which wishes to estimate its share of the "fertilizer" market will find that the government statistics included lime under "fertilizers".
• The sample size may have been too small for reliable results, or the method of selecting the sample a poor one.
• The wording of the questions may have been poor, possibly biasing the results.
No control is possible over the quality of the collecting procedure, e.g. by seeing that measurements were accurate, questions were properly asked and calculations accurate.

However, the advantage of secondary data, when available and appropriate, is that a great deal of time and money may be saved by not having to collect the data oneself. Indeed in many cases, for example with import-export statistics, it may be impossible for a private individual or company to collect the data which can only be obtained by the government.

1.5 Graphical Descriptions of Data

1.5.1 Graphical Presentation

A graph is a method of presenting statistical data in visual form. The main purpose of any chart is to give a quick, easy-to-read-and-interpret pictorial representation of data which is more difficult to obtain from a table or a complete listing of the data. The type of chart or graphical presentation used and the format of its construction is incidental to its main purpose. A well-designed graphical presentation can effectively communicate the data's message in a language readily understood by almost everyone. You will see that graphical methods for describing data are intuitively appealing descriptive techniques and that they can be used to describe either a sample or a population; quantitative or qualitative data sets.

Some basic rules for the construction of a statistical chart are listed below:
(a) Every graph must have a clear and concise title which gives enough identification of the graph.
(b) Each scale must have a scale caption indicating the units used.
(c) The zero point should be indicated on the co-ordinate scale. If, however, lack of space makes it inconvenient to use the zero point line, a scale break may be inserted to indicate its omission.
(d) Each item presented in the graph must be clearly labelled and legible even in black and white reprint.

There are many varieties of graphs. The most commonly used graphs are described as below.
(a) Pie chart - Pie charts are widely used to show the component parts of a total. They are popular because of their simplicity. In constructing a pie chart, the angles of a slice from the center must be in proportion with the percentage of the total. The following example of pie charts gives the percentage of education attainment in Hong Kong.

(b) Simple bar chart - The horizontal bar chart is also a simple and popular chart. Like the pie chart, the simple horizontal bar chart is a one-scale chart. In constructing a bar chart, it is noted that the width of the bar is not important, but the height of the bar must be in proportion with the data. The following bar chart gives the monthly household income of Hong Kong.
(c) Two-directional bar chart - A bar chart can use either horizontal or vertical bars. A two-directional bar chart indicates both the positive and negative values. The following example gives the top 5 cities which have the highest/lowest recorded temperature.

(d) Multiple bar chart - A multiple bar chart is particularly useful if one desires to make quick comparison between different sets of data. In the following example, the marital status of male and female in Hong Kong are compared using multiple bar chart.
(e) Component bar chart - A component bar chart subdivides the bars in different sections. It is useful when the total of the components is of interest. The following example gives the nutritive values of food.

(f) Other type of graphs - Graphic presentations can be made more attractive through the use of careful layout and appropriate symbols. Sometimes information pertaining to different geographical area can even be presented through the use of so-called statistical map.

A pictograph illustrates statistical data by means of a pictorial symbol. It can add greatly to the interest of what might otherwise be a dull subject. The chosen symbol must have a close association with the subject matter, so that the reader can comprehend the subject under discussion at a glance.
Chapter 1: Descriptive Statistics

1.6 Frequency Distribution

Statistical data obtained by means of census, sample surveys or experiments usually consist of raw, unorganized sets of numerical values. Before these data can be used as a basis for inferences about the phenomenon under investigation or as a basis for decision, they must be summarized and the pertinent information must be extracted.

Example 1

A traffic inspector has counted the number of automobiles passing a certain point in 100 successive 20-minute time periods. The observations are listed below.

<p>| | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>20</td>
<td>16</td>
<td>18</td>
<td>30</td>
<td>22</td>
<td>26</td>
<td>15</td>
<td>5</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>17</td>
<td>11</td>
<td>37</td>
<td>21</td>
<td>6</td>
<td>10</td>
<td>20</td>
<td>22</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>19</td>
<td>20</td>
<td>12</td>
<td>23</td>
<td>24</td>
<td>17</td>
<td>18</td>
<td>16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>16</td>
<td>28</td>
<td>26</td>
<td>15</td>
<td>29</td>
<td>19</td>
<td>35</td>
<td>20</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>30</td>
<td>21</td>
<td>22</td>
<td>20</td>
<td>15</td>
<td>18</td>
<td>16</td>
<td>23</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>24</td>
<td>28</td>
<td>19</td>
<td>24</td>
<td>22</td>
<td>17</td>
<td>19</td>
<td>8</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>18</td>
<td>23</td>
<td>21</td>
<td>25</td>
<td>19</td>
<td>20</td>
<td>22</td>
<td>21</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>20</td>
<td>19</td>
<td>11</td>
<td>23</td>
<td>17</td>
<td>23</td>
<td>13</td>
<td>17</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>27</td>
<td>18</td>
<td>21</td>
<td>24</td>
<td>33</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>27</td>
<td>18</td>
<td>22</td>
<td>17</td>
<td>20</td>
<td>14</td>
<td>21</td>
<td>22</td>
<td>19</td>
<td></td>
</tr>
</tbody>
</table>

A useful method for summarizing a set of data is the construction of a frequency table, or a frequency distribution. That is, we divide the overall range of values into a number of classes and count the number of observations that fall into each of these classes or intervals.

The general rules for constructing a frequency distribution are

i) There should not be too few or too many classes.

ii) Insofar as possible, equal class intervals are preferred. But the first and last classes can be open-ended to cater for extreme values.

iii) Each class should have a class mark to represent the classes. It is also named as the class midpoint of the ith class. It can be found by taking simple average of the class boundaries or the class limits of the same class.

1. Setting up the classes

Choose a class width of 5 for each class, then we have seven classes going from 5 to 9, from 10 to 14, …, and from 35 to 39.
2. Tallying and counting

<table>
<thead>
<tr>
<th>Classes</th>
<th>Tally Marks</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 – 9</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>10 – 14</td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>15 – 19</td>
<td></td>
<td>36</td>
</tr>
<tr>
<td>20 – 24</td>
<td></td>
<td>35</td>
</tr>
<tr>
<td>25 – 29</td>
<td></td>
<td>12</td>
</tr>
<tr>
<td>30 – 34</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>35 – 39</td>
<td></td>
<td>2</td>
</tr>
</tbody>
</table>

3. Illustrating the data in tabular form

**Frequency Distribution for the Traffic Data**

<table>
<thead>
<tr>
<th>Number of autos per period</th>
<th>Number of periods</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 – 9</td>
<td>3</td>
</tr>
<tr>
<td>10 – 14</td>
<td>9</td>
</tr>
<tr>
<td>15 – 19</td>
<td>36</td>
</tr>
<tr>
<td>20 – 24</td>
<td>35</td>
</tr>
<tr>
<td>25 – 29</td>
<td>12</td>
</tr>
<tr>
<td>30 – 34</td>
<td>3</td>
</tr>
<tr>
<td>35 – 39</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
</tr>
</tbody>
</table>

In this example, the class marks of the traffic-count distribution are 7, 12, 17, ..., 32 and 37.