CHAPTER 1 DESCRIPTIVE STATISTICS

Contents

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

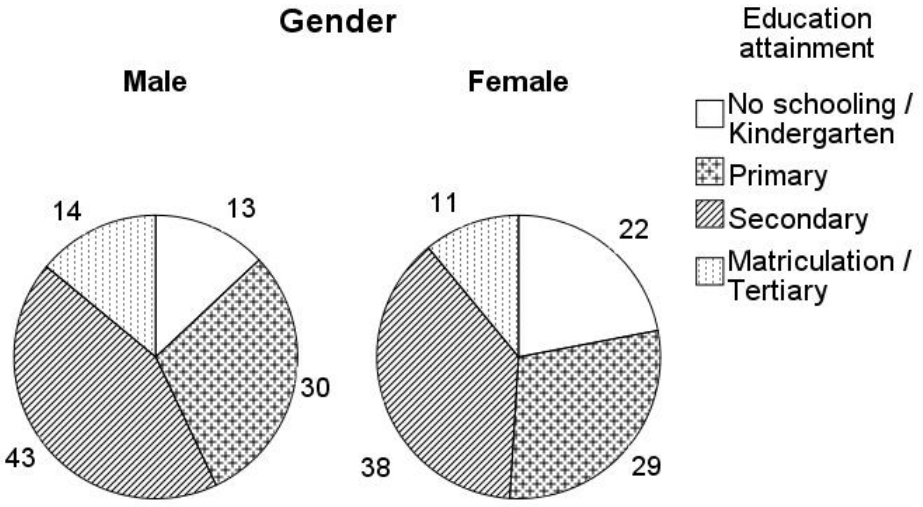
A graph is a method of presenting statistical data in visual form. The main purpose of any chart is to give a quick, easy-to-read-and-interpret pictorial representation of data which is more difficult to obtain from a table or a complete listing of the data. The type of chart or graphical presentation 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:

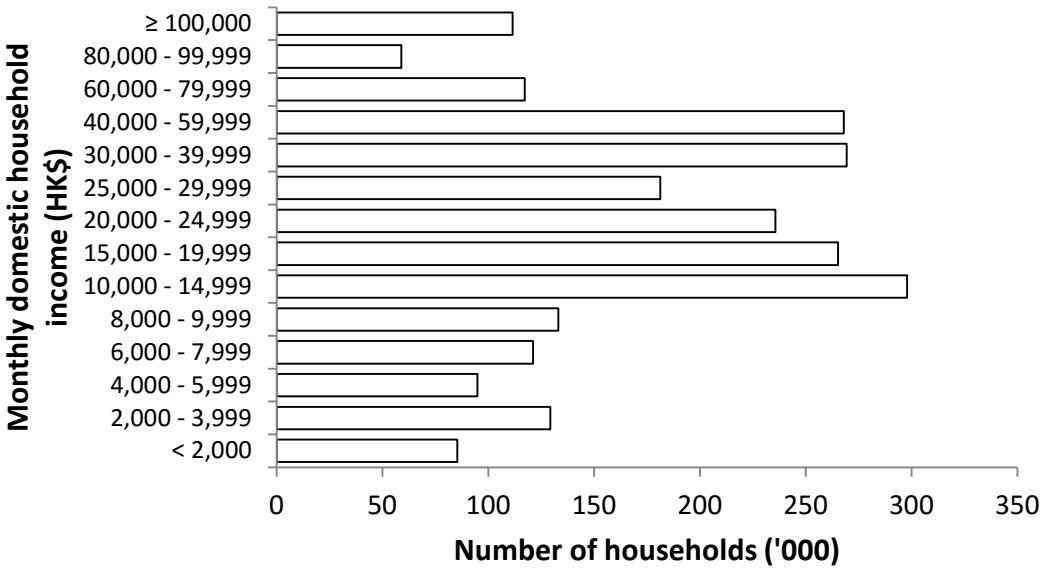
(a) Every graph must have a clear and concise title which gives enough identification of the graph.
(b) Each scale must have a scale caption indicating the units used.
(c) The zero point should be indicated on the co-ordinate scale. If, however, lack of space makes it inconvenient to use the zero point 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.

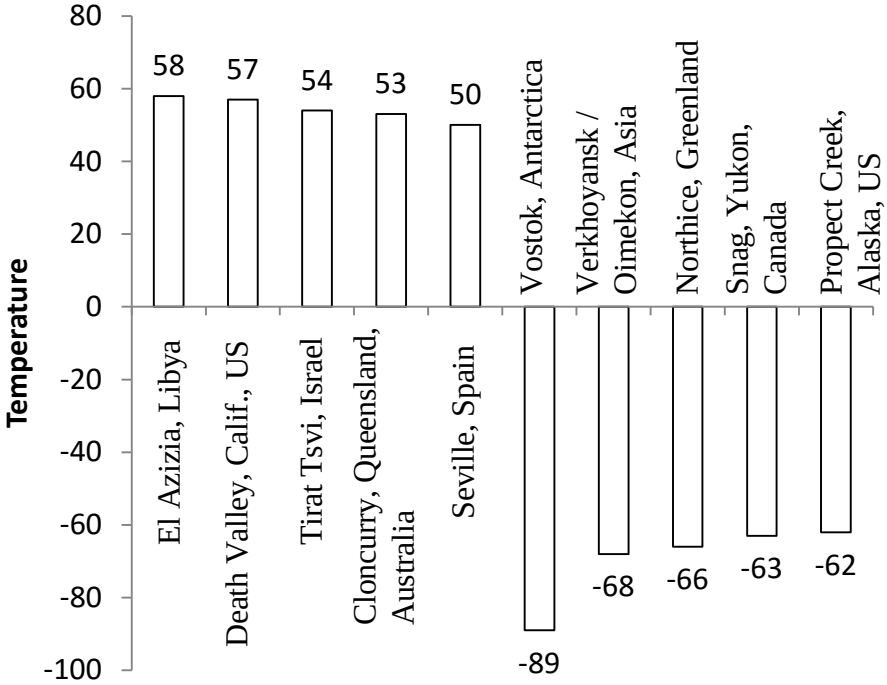
(a) Pie chart - Pie charts are widely used to show the component parts of a total. They are popular because of their simplicity. In constructing a pie chart, the angles of a slice from the center must be in proportion with the percentage of the total. The following example of pie charts gives the percentage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



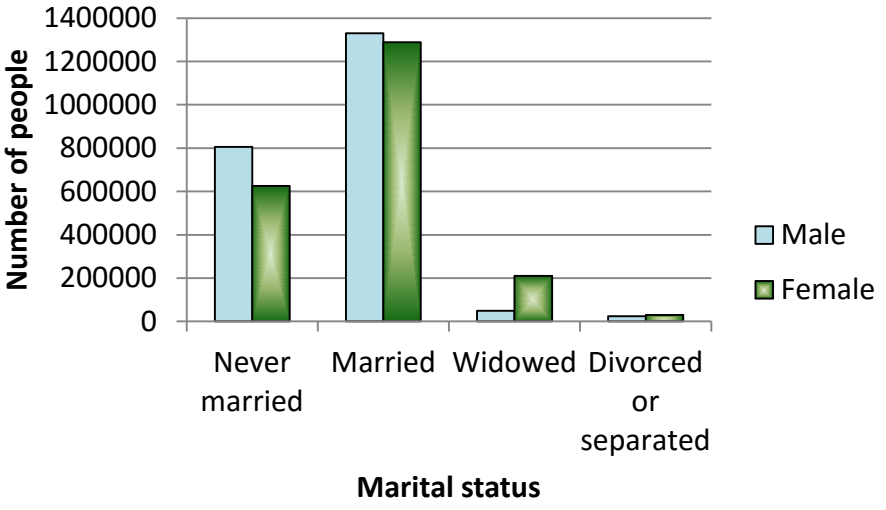
(b) Simple bar chart - The horizontal bar chart is also a simple and popular chart. Like the pie chart, the simple horizontal bar chart is a one-scale chart. In constructing a bar chart, it is noted that the width of the bar is not important, but the height of the bar must be in proportion with the data. The following bar chart gives the monthly household income of Hong Kong.



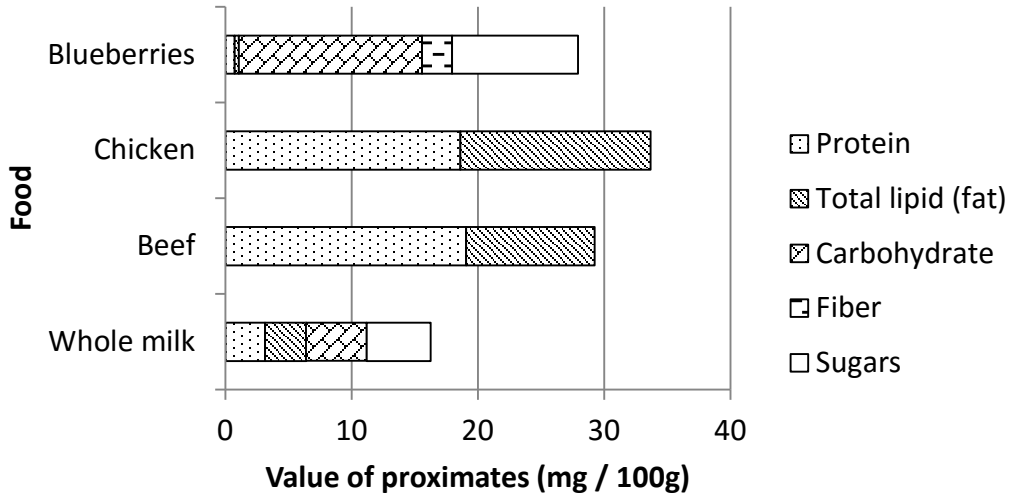
(c) Two-directional bar chart - A bar chart can use either horizontal or vertical bars. A two-directional bar chart indicates both the positive and negative values. The following example gives the top 5 cities which have the highest/ lowest recorded temperature.



(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 char.



(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.

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.

23	20	16	18	30	22	26	15	5	18
14	17	11	37	21	6	10	20	22	25
19	19	19	20	12	23	24	17	18	16
27	16	28	26	15	29	19	35	20	17
12	30	21	22	20	15	18	16	23	24
15	24	28	19	24	22	17	19	8	18
17	18	23	21	25	19	20	22	21	21
16	20	19	11	23	17	23	13	17	26
26	14	15	16	27	18	21	24	33	20
21	27	18	22	17	20	14	21	22	19

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

Classes	Tally Marks	Count
5 – 9		3
10 – 14		9
15 – 19		36
20 – 24		35
25 – 29		12
30 – 34		3
35 – 39		2

3. Illustrating the data in tabular form

Frequency Distribution for the Traffic Data

Number of autos per period	Number of periods
5 – 9	3
10 – 14	9
15 – 19	36
20 – 24	35
25 – 29	12
30 – 34	3
35 – 39	2
Total	100

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

1.6.1 Histogram

A histogram is usually used to present frequency distributions graphically. This is constructed by drawing rectangles over each class. The **area** of each rectangle should be proportional to its frequency.

Notes :

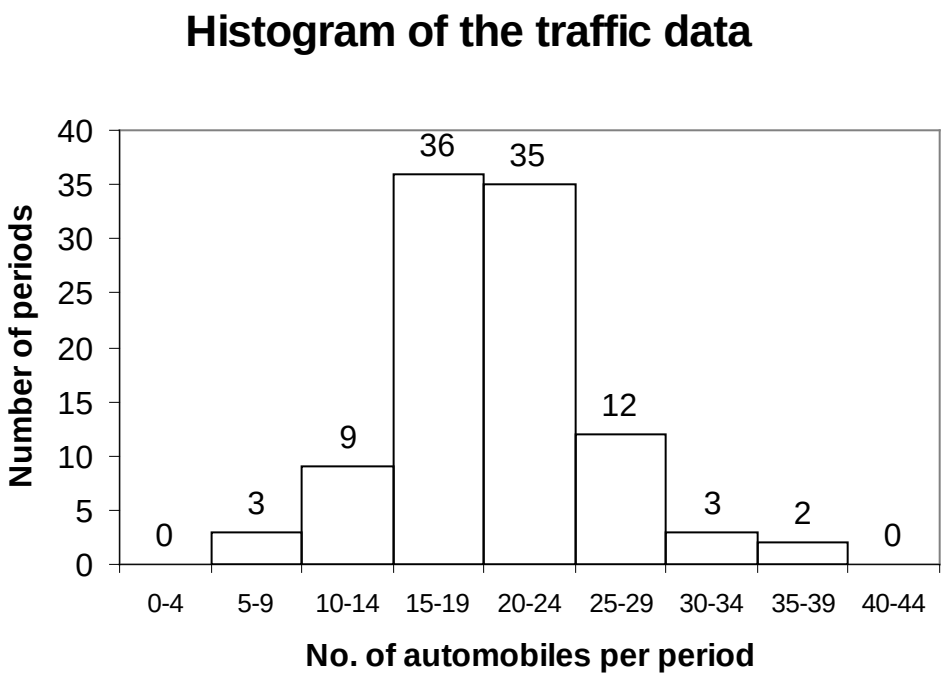
1. The vertical lines of a histogram should be the **class boundaries**.
2. The range of the random variable should constitute the major portion of the graphs of frequency distributions. If the smallest observation is far away from zero, then a 'break' sign ( ) should be introduced in the horizontal axis.

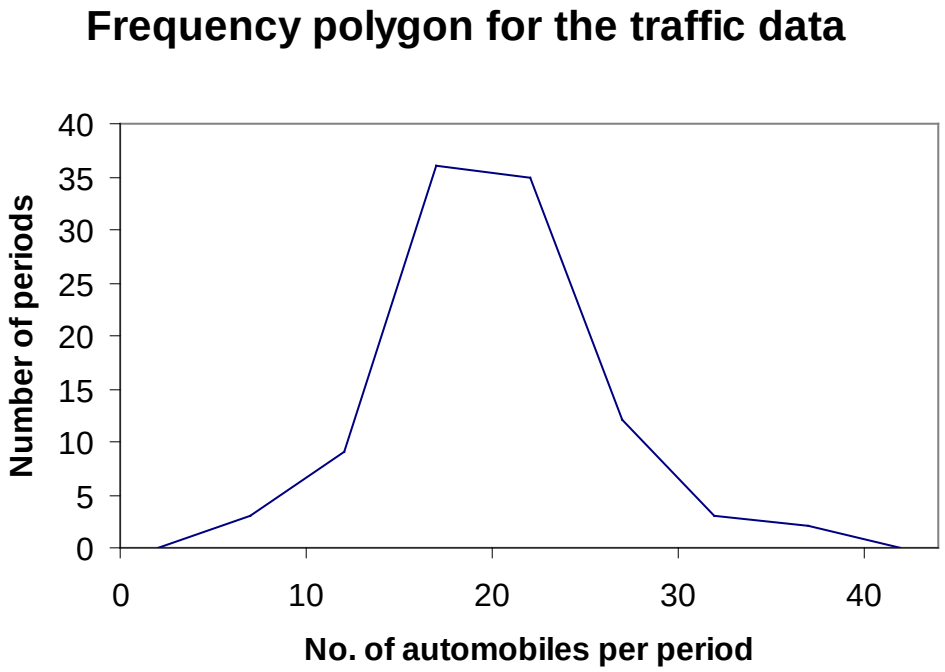
1.6.2 Frequency Polygon

Another method to represent frequency distribution graphically is by a frequency polygon. As in the histogram, the base line is divided into sections corresponding to the class-interval, but instead of the rectangles, the points of successive class marks are being connected. The frequency polygon is particularly useful when two or more distributions are to be presented for comparison on the same graph.

Example 2

Construct a histogram and a frequency polygon for the traffic data in Example 1.





1.6.3 Frequency Curve

A frequency curve can be obtained by smoothing the frequency polygon.

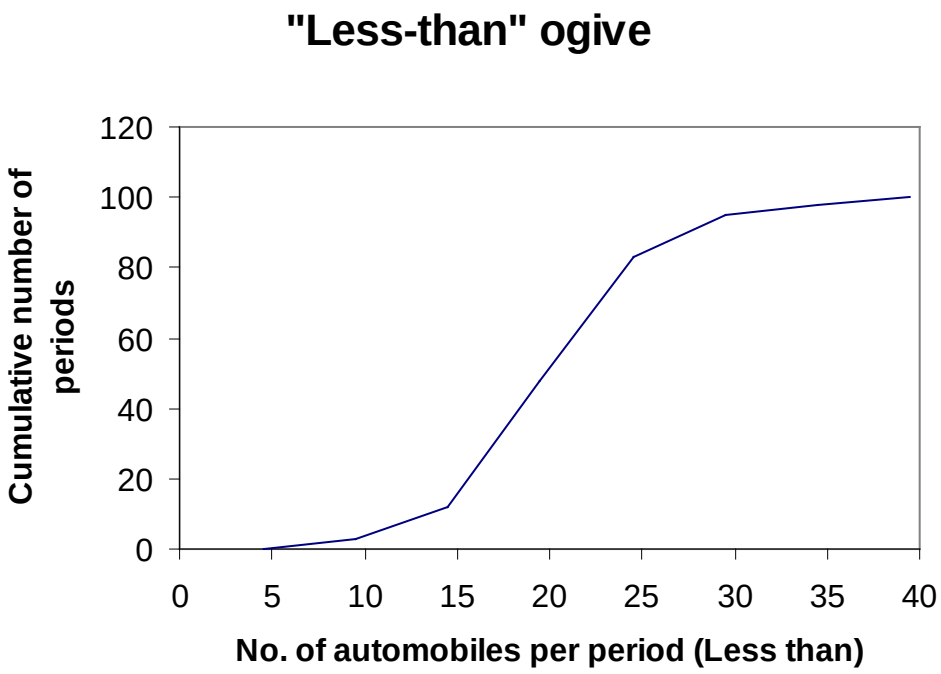
1.6.4 Cumulative Frequency Distribution and Cumulative Polygon

Sometimes it is preferable to present data in a cumulative frequency distribution, which shows directly how many of the items are less than, or greater then, various values.

Less than	Cumulative frequency
4.5	0
9.5	3
14.5	12
19.5	48
24.5	83
29.5	95
34.5	98
39.5	100

Example 3

Construct a "Less-than" ogive of the distribution of traffic data.



1.6.5 Cumulative Frequency Curve

A cumulative frequency curve can similarly be drawn.

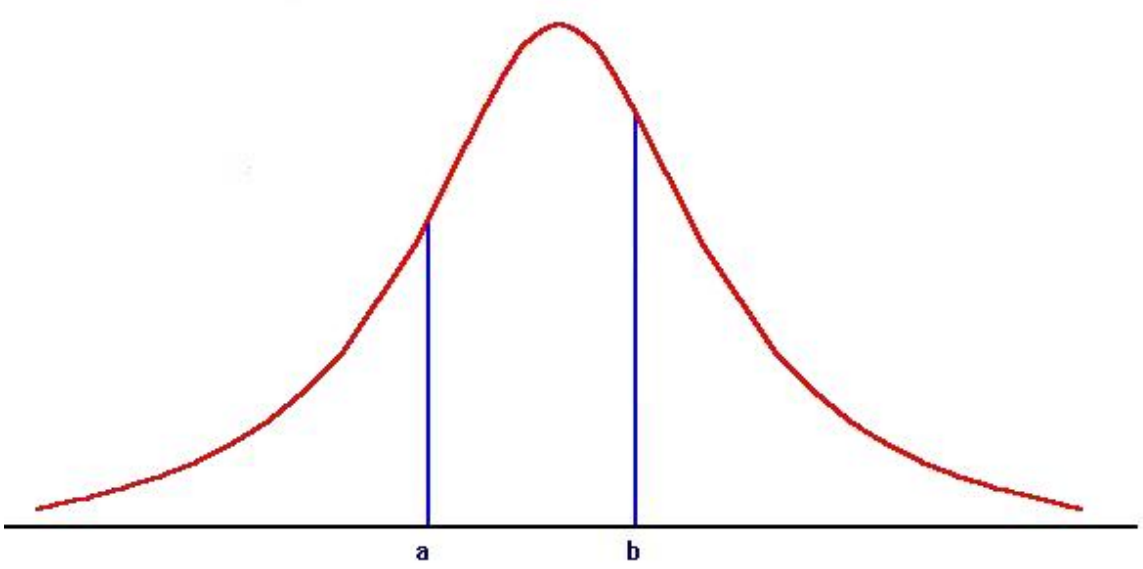
1.6.6 Relative Frequency

Relative frequency of a class is defined as:

$$\frac{\text{Frequency of the Class}}{\text{Total Frequency}}$$

If the frequencies are changed to relative frequencies, then a relative frequency histogram, a relative frequency polygon and a relative frequency curve can similarly be constructed.

Relative frequency curve can be considered as probability curve if the total area under the curve be set to 1. Hence the area under the relative frequency curve between a and b is the probability between interval a and b.



1.7 Central Tendency

When we work with numerical data, it seems apparent that in most set of data there is a tendency for the observed values to group themselves about some interior values; some central values seem to be the characteristics of the data. This phenomenon is referred to as central tendency. For a given set of data, the measure of location we use depends on what we mean by middle; different definitions give rise to different measures. We shall consider some more commonly used measures, namely arithmetic mean, median and mode. The formulas in finding these values depend on whether they are ungrouped data or grouped data.

1.7.1 Arithmetic Mean

The arithmetic population mean, μ, or simply called mean, is obtained by adding together all of the measurements and dividing by the total number of measurements taken. Mathematically it is given as

$$\mu = \frac{\sum x_i}{N}$$

Arithmetic mean can be used to calculate any numerical data and it is always unique. It is obvious that extreme values affect the mean. Also, arithmetic mean ignores the degree of importance in different categories of data.

$$\overline{x} = \frac{\sum x_i}{n}$$

Example 4

Given the following set of ungrouped data:

20, 18, 15, 15, 14, 12, 11, 9, 7, 6, 4, 1

Find the mean of the ungrouped data.

$$mean = \frac{20+18+2\cdot 15+14+12+11+9+7+6+4+1}{12}$$

$$= \frac{132}{12}$$

$$= 11$$

1.7.2 Median

Median is defined as the middle item of all given observations arranged in order. For ungrouped data, the median is obvious. In case of the number of measurements is even, the median is obtained by taking the average of the middle.

Example 5

The median of the ungrouped data: 20, 18, 15, 15, 14, 12, 11, 9, 7, 6, 4, 1 is

$$\frac{12+11}{2}$$

$$= 11.5$$

1.7.3 Mode

Mode is the value which occurs most frequently. The mode may not exist, and even if it does, it may not be unique.

For ungrouped data, we simply count the largest frequency of the given value. If all are of the same frequency, no mode exits. If more than one values have the same largest frequency, then the mode is not unique.

Example 6

The value for the mode of the data in Example 5 is 15 (unimodal)

Example 7

{2, 2, 2, 4, 5, 6, 7, 7, 7}

Mode = 2 or 7 (Bimodal)

Note that the mode is independent of extreme values and it may be applied in qualitative data.

1.7.5 Conclusion

For symmetrically distributed and unimodal data, the mean, median and mode can be used almost interchangeably.

Physically, mean can be interpreted as the center of gravity of the distribution. Median divides the area of the distribution into two equal parts and mode is the highest point of the distribution.

1.8 Dispersion and Skewness

Sometimes mean, median and mode may not be able to reflect the true picture of some data. The following example explains the reason.

Example 8

There were two companies, Company A and Company B. Their salaries profiles given in mean, median and mode were as follow:

	Company A	Company B
Mean	$30,000	$30,000
Median	$30,000	$30,000
Mode	(Nil)	(Nil)

However, their detail salary ($) structures could be completely different as that:

Company A	5,000	15,000	25,000	35,000	45,000	55,000
Company B	5,000	5,000	5,000	55,000	55,000	55,000

Hence it is necessary to have some measures on how data are scattered. That is, we want to know what is the dispersion, or variability in a set of data.

1.8.1 Range

Range is the difference between two extreme values. The range is easy to calculate but cannot be obtained if open ended grouped data are given.

1.8.2 Deciles, Percentile, and Fractile

Decile divides the distribution into ten equal parts while percentile divides the distribution into one hundred equal parts. There are nine deciles such that 10% of the data are $\leq D_1$; 20% of the data are $\leq D_2$; and so on. There are 99 percentiles such that 1% of the data are $\leq P_1$; 2% of the data are $\leq P_2$; and so on. Fractile, even more flexible, divides the distribution into a convenient number of parts.

1.8.3 Quartiles

Quartiles are the most commonly used values of position which divides distribution into four equal parts such that 25% of the data are $\leq Q_1$; 50% of the data are $\leq Q_2$; 75% of the data are $\leq Q_3$. It is also denoted the value $(Q_3 - Q_1) / 2$ as the Quartile Deviation, Q_D, or the semi-interquartile range.

1.8.4 Mean Absolute Deviation

Mean absolute deviation is the mean of the absolute values of all deviations from the mean. Therefore it takes every item into account. Mathematically it is given as:

$$\frac{\sum |x_i - \mu|}{N}$$

where: x_i is the value of the i^{th} item;
μ is the population arithmetic mean;
N is the population size.

1.8.5 Variance and Standard Deviation

The variance and standard deviation are two very popular measures of variation. Their formulations are categorized into whether to evaluate from a population or from a sample.

The population variance, σ^2, is the mean of the square of all deviations from the mean. Mathematically it is given as:

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

where: x_i is the value of the i^{th} item;
μ is the population arithmetic mean;
N is the population size.

The population standard deviation σ is defined as $\sigma = \sqrt{\sigma^2}$.

The sample variance, denoted as s^2 gives:

$$\frac{\sum (x_i - \bar{x})^2}{n-1}$$

The sample standard deviation, s, is defined as $s = \sqrt{s^2}$.

For ungrouped data, $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum x_i^2 - \left(\sum x_i\right)^2 / n}{n-1}}$

where: x_i is the value of the i^{th} item;
$\bar{x}$ is the sample arithmetic mean;
n is the sample size.

Note that when calculating the sample variance, we have to subtract 1 from the sample size which appears in the denominator.

1.8.6 Measures of Grouped Data

Example 9

Gas Consumption	Frequency (f_i)	Class boundary	Class mark (x_i)	$f_i x_i$	$f_i x_i^2$
10 – 19	1	9.5 – 19.5	14.5	14.5	210.25
20 – 29	0	19.5 – 29.5	24.5	0	0
30 – 39	1	29.5 – 39.5	34.5	34.5	1190.25
40 – 49	4	39.5 – 49.5	44.5	178	7921
50 – 59	7	49.5 – 59.5	54.5	381.5	20791.75
60 – 69	16	59.5 – 69.5	64.5	1032	66564
70 – 79	19	69.5 – 79.5	74.5	1415.5	105454.8
80 – 89	20	79.5 – 89.5	84.5	1690	142805
90 – 99	17	89.5 – 99.5	94.5	1606.5	151814.3
100 – 109	11	99.5 – 109.5	104.5	1149.5	120122.8
110 – 119	3	109.5 – 119.5	114.5	343.5	39330.75
120 – 129	1	119.5 – 129.5	124.5	124.5	15500.25
	100			7970	671705

1. $$\bar{x} = \frac{\sum x_i f_i}{n}, n = \sum f_i$$
$$= \frac{1\times 14.5 + 0\times 24.5 + \cdots + 1\times 124.5}{100}$$
$$= 79.7$$

2. For grouped data, the median can be found by first identify the class containing the median, then apply the following formula:

$$median = L_1 + \frac{\frac{n}{2} - C}{f_m}(L_2 - L_1)$$

where: L_1 is the lower class boundary of the median class;
n is the total frequency (i.e. the sample size);
C is the cumulative frequency just before the median class;
f_m is the frequency of the median class;
L_2 is the upper class boundary containing the median.

It is obvious that the median is affected by the total number of data but is independent of extreme values. However if the data is ungrouped and numerous, finding the median is tedious. Note that median may be applied in qualitative data if they can be ranked.

$$median = 79.5 + \frac{50-48}{20} \times 10 = 80.5$$

$$Q_1 = 59.5 + \frac{25-13}{16} \times 10$$
$$\approx 67$$

$$Q_3 = 89.5 + \frac{75-68}{17} \times 10$$
$$\approx 93.6$$

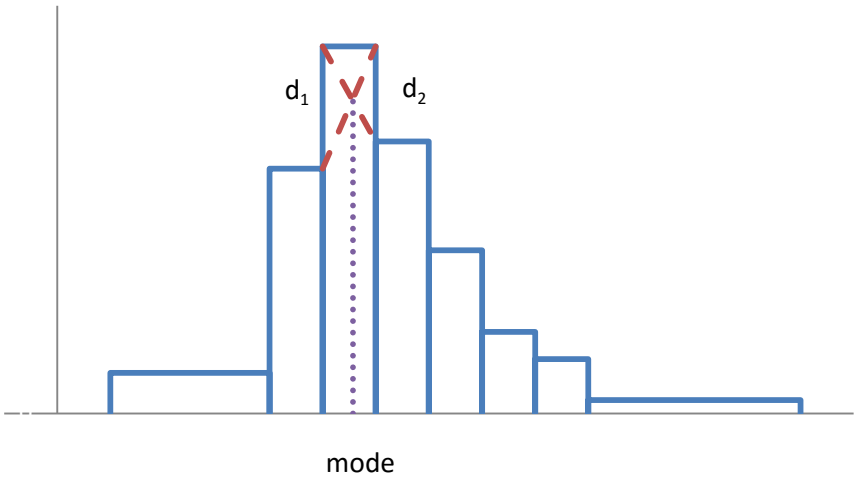
3.

For grouped data, the mode can be found by first identify the largest frequency of that class, called modal class, then apply the following formula on the modal class:

$$\text{mode} = L_1 + \frac{d_1}{d_1 + d_2}\left(L_2 - L_1\right)$$

where: L_1 is the lower class boundary of the modal class;
d_1 is the difference of the frequencies of the modal class with the previous class and is always positive;
d_2 is the difference of the frequencies of the modal class with the following class and is always positive;
L_2 is the upper class boundary of the modal class.

Geometrically the mode can be represented by the following graph and can be obtained by using similar triangle properties. The formula can be derived by interpolation using second degree polynomial.



$$mode = 79.5 + \frac{20-19}{(20-19)+(20-17)} \times 10$$
$$= 82$$

4.

$$s = \sqrt{\frac{\sum f\left(x-\bar{x}\right)^2}{\sum f - 1}} = \sqrt{\frac{\sum fx^2 - \left(\sum fx\right)^2 / \sum f}{\sum f - 1}} \quad \text{where } \sum f = n$$

$$sample \text{ s.d., } s = \sqrt{\frac{n\left(\sum x_i^2 f_i\right) - \left(\sum x_i f_i\right)^2}{n(n-1)}}$$
$$= \sqrt{\frac{100\left(671705\right) - \left(7970\right)^2}{100\left(100-1\right)}}$$
$$= 19.2$$

1.8.7 Coefficient of Variation

The coefficient of variation is a measure of relative importance. It does not depend on unit and can be used to make comparison even two samples differ in means or relate to different types of measurements.

The coefficient of variation gives:

$$\frac{\text{Standard Deviation}}{\text{Mean}} \times 100\%$$

Example 10

	$\bar{x}$	s
Salesman salary	\$916.76/month	\$286.70/month
Clerical salary	\$98.50/week	\$20.55/week

$$CVs = \frac{286.70}{916.76} \times 100\% = 31\%$$
$$CVc = \frac{20.55}{98.50} \times 100\% = 21\%$$

Example 11

We are going to use Example 8 to evaluate the different measurements of variation.

As stated above, the salary (\$) scales of the two companies are:

Company A:	5,000	15,000	25,000	35,000	45,000	55,000
Company B:	5,000	5,000	5,000	55,000	55,000	55,000

Range

Company A: \$55,000 - \$5,000
= \$50,000

Company B: \$55,000 - \$5,000
= \$50,000

Mean absolute deviation

Company A: \$ (|5,000 - 30,000| + |15,000 - 30,000| + |25,000 - 30,000| + |35,000 - 30,000| + |45,000 - 30,000| + |55,000 - 30,000|) / 6
= \$15,000

Company B: $ (|5,000 - 30,000| + |5,000 - 30,000| + |5,000 - 30,000| + |55,000 - 30,000| + |55,000 - 30,000| + |55,000 - 30,000|) / 6
= $25,000

Variance

Company A: $\{(5,000 - 30,000)^2 + (15,000 - 30,000)^2 + (25,000 - 30,000)^2 + (35,000 - 30,000)^2 + (45,000 - 30,000)^2 + (55,000 - 30,000)^2 \} / 6$
= 291,666,667 (dollar square)

Company B: $\{(5,000 - 30,000)^2 + (5,000 - 30,000)^2 + (5,000 - 30,000)^2 + (55,000 - 30,000)^2 + (55,000 - 30,000)^2 + (55,000 - 30,000)^2 \} / 6$
= 625,000,000 (dollar square)

Standard deviation

Company A: $\$\sqrt{291,666,667}$
= $17,078

Company B: $\$\sqrt{625,000,000}$
= $25,000

Coefficient of variation

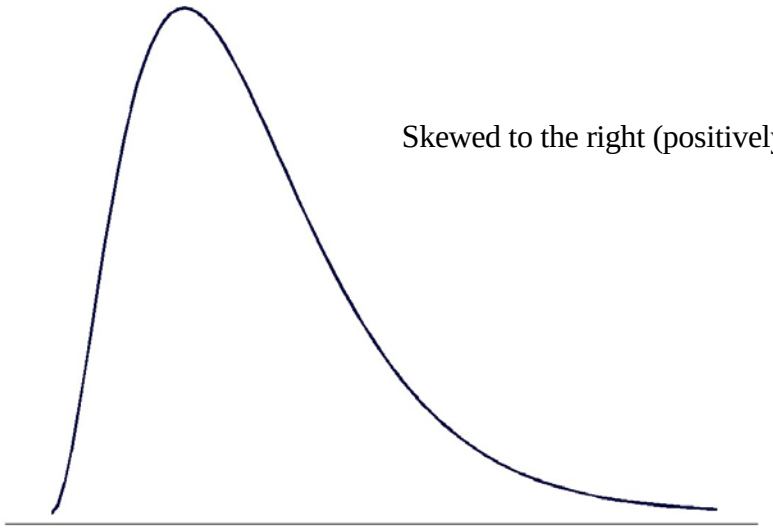
Company A: $17,078 / $30,000 × 100%
= 56.93%

Company B: $25,000 / $30,000 × 100%
= 83.33%

1.8.8 Skewness

The skewness is an abstract quantity which shows how data piled-up. A number of measures have been suggested to determine the skewness of a given distribution.

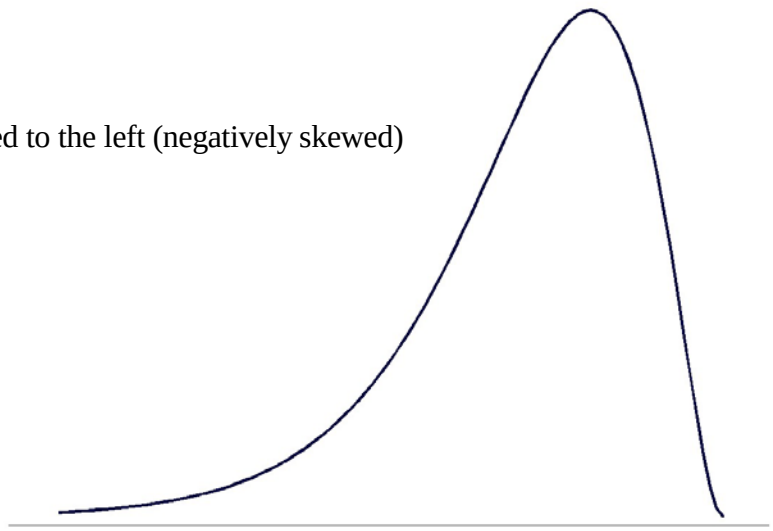
If the longer tail is on the right, we say that it is skewed to the right, and the coefficient of skewness is positive.



Skewed to the right (positively skewed)

If the longer tail is on the left, we say that is skewed to the left and the coefficient of skewness is negative.

Skewed to the left (negatively skewed)



Coefficient of Skewness

Pearson's 1st coefficient of skewness,

$$SK_1 = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}}$$

Pearson's 2nd coefficient of skewness

$$SK_2 = \frac{3(\text{Mean} - \text{Median})}{\text{Standard deviation}}$$

For moderately skewed distribution data, their relationship can be given by

$$\text{Mean - Mode} \approx 3 \cdot (\text{Mean - Median})$$

$$\text{Skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

$$\approx \frac{3\,(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

Chebyshev's Theorem

For any set of data, the proportion of data that lies between the mean plus and minus k standard deviations is at least $1 - \frac{1}{k^2}$

i.e. $\Pr\left(\mu - k\sigma \le x \le \mu + k\sigma\right) \ge 1 - \frac{1}{k^2}$

Symbols

	Population	Sample
Size	N	n
Mean	μ	$\overline{x}$
Standard deviation	σ	s
Variance	σ^2	s^2

EXERCISE: DESCRIPTIVE STATISTICS

1. In the following list, post a D for the situations in which statistical techniques are used for the purpose of description and an I for those in which the techniques are used for the purpose of inference.

   _____ (a) The price movements of 50 issues of stock are analysed to determine whether stocks in general have gone up or down during a certain period of time.

   _____ (b) A statistical table is constructed for the purpose of presenting the passenger-miles flown by various commercial airlines in the United States.

   _____ (c) The average of a group of test scores is computed so that each score in the group can be classified as being either above or below average.

   _____ (d) Several manufacturing firms in a particular industry are surveyed for the purpose of estimating industrywide investment in capital equipment.

2. No matter how few elements are included in a statistical population, however, a sample taken from that population (can/cannot) be larger than the population itself.

3. Thus any descriptive measurement of a population is considered to be a (statistics/parameter), and a descriptive measurement of a sample is a sample _____.

4. The word "statistics" has at least three distinct meanings, depending on the context in which it is used. It may refer to:

   (i) the procedure of statistical analysis
   (ii) descriptive measures of a sample
   (iii) the individual measurements, or elements, that make up either a sample or a population.

   (a) When one becomes "an accident statistics" by being included in some count of accident frequency, the term is used in the sense of definition ____.

(b) According to the definitions in a course of study called "Business Statistics" the term "statistics" is usually used in the sense of definition ____.

(c) According to the definitions when such sample statistics as the proportion of a sample in favour of a proposal and the average age of those in the sample are determined, the term "statistics" is being used in the sense of definition ____.

5. The two major applications of the tools of statistical analysis are directed toward the purposes of statistical _____ and statistical _____.

6. When all the elements in a statistical population are measured, the process is referred to as "taking a _____". If only a portion of the elements included in a statistical population is measured, the process is called _____.

7. Which of the following measures of variability is not dependent on the exact value of each observation?

(a) range
(b) variance
(c) standard deviation
(d) coefficient of variation

8. A measure of dispersion which is insensitive to extreme values in the data set is the:

(a) Quartile deviation
(b) Standard deviation
(c) Average absolute deviation
(d) All of the above

9. An absolute measure of dispersion which expresses variation in the same units as the original data is the:

(a) Standard deviation
(b) Coefficient of variation
(c) Variance
(d) All of the above

10. How does the computation of a sample variance differ from the computation of a population variance?

    (a) μ is replaced by $\overline{x}$
    (b) N is replaced by n –1
    (c) N is replaced by n
    (d) a and c but not b
    (e) a and b but not c

11. Which measure of central tendency would be most useful in each of the following instances?

    (a) The production manager for a manufacturer of glass jars, who is concerned about the proper jar size to manufacture, has sample data on jar sizes ordered by customers. Would the mean, median, or modal jar size be of most value to the manager?

    (b) The sales manager for a quality furniture manufacturer is interested in selecting the regions most likely to purchase his firm's products. Would he be most interested in the mean or median family income in prospective sales areas?

    (c) A security analyst is interested in describing the daily market price change of the common stock of a manufacturing company. Only rarely does the market price of the stock change by more than one point, but occasionally the price will change by as many as four points in one day. Should the security analyst describe the daily price change of the stock in terms of the mean, median, or modal daily market price change?

12. Why isn't an average computed from a group frequency distribution exactly the same as that computed from the original raw data used to construct the distribution?

13. For which type of distribution (positively skewed, negatively skewed, or symmetric) is:

    (a) The mean less than the median?

    (b) The mode less than the mean?

    (c) The median less than the mode?

14. The following scores represent the final examination grade for an elementary statistics course:

23	60	79	32	57	74	52	70	82	36
80	77	81	95	41	65	92	85	55	76
52	10	64	75	78	25	80	98	81	67
41	71	83	54	64	72	88	62	74	43
60	78	89	76	84	48	84	90	15	79
34	67	17	82	69	74	63	80	85	61

Using 10 class intervals with the lowest starting at 9:

(a) Set up a frequency distribution.
(b) Construct a cumulative frequency distribution.
(c) Construct a frequency histogram.
(d) Construct a smoothed cumulative frequency polygon.
(e) Estimate the number of people who made a score of at least 60 but less than 75.
(f) Discuss the skewness of the distribution.

15. Classify the following random variables as discrete or continuous.

(a) The number of automobile accidents each year in Hong Kong.
(b) The length of time to do problem 1 above.
(c) The amount of milk produced yearly by a particular cow.
(d) The number of eggs laid each month by 1 hen.
(e) Numbers of shares sold each year in the stock market.
(f) The weight of grain in kg produced per acre.

16. An electronically controlled automatic bulk food filler is set to fill tubs with 60 units of cheese. A random sample of five tubs from a large production lot shows filled weights of 60.00, 59.95, 60.05, 60.02 and 60.01 units. Find the mean and the standard deviation of these fills.

17. In four attempts it took a person 48, 55, 51 and 50 minutes to do a certain job.

(a) Find the mean, the range, and the standard deviation of these four sample values.
(b) Subtract 50 minutes from each of the times, recalculate the mean, the range, and the standard deviation, and compare the results with those obtained in part (a).
(c) Add 10 minutes to each of the times, recalculate the mean, the range, and the standard deviation, and compare the results with those obtained in part (a).

(d) Multiply each of the sample values by 2, recalculate the mean, the range, and the standard deviation, and compare the results with those obtained in part (a).

(e) In general, what effect does (1) adding a constant to each sample value, and (2) multiplying each sample value by a positive constant, have on the mean, the range, and the standard deviation of a sample?

18. Find the mean, median and mode for the set of numbers

(a) 3, 5, 2, 6, 5, 9, 5, 2, 8, 6;
(b) 51.6, 48.7, 50.3, 49.5, 48.9.

19. The lengths of a large shipment of chromium strips have a mean of 0.44 m and standard deviation of 0.001 m. At least what percentage of these lengths must lie between

(a) 0.438 and 0.442 m?
(b) 0.436 and 0.444 m?
(c) 0.430 and 0.450 m?

20. The 1971 populations and growth rates for various regions are given below. Find the growth rate for the world as a whole

Region	Population (millions)	Annual Growth Rate (%)
Europe	470	0.8
USSR	240	1.1
N. America	230	1.3
Oceania	20	2.1
Asia	2,100	2.3
Africa	350	2.6
S. America	290	2.9

21. Suppose that the annual income of the residents of a certain country has a mean of $48,000 and a median of $34,000. What is the shape of the distribution?

22. In a factory, the time during working hours in which a machine is not operating as a result of breakage or failure is called the 'downtime". The following distribution shows a sample of 100 downtimes of a certain machine (rounded to the nearest minute):

Downtime	Frequencies
0 – 9	3
10 – 19	13
20 – 29	30
30 – 39	25
40 – 49	14
50 – 59	8
60 – 69	4
70 – 79	2
80 – 89	1

With reference to the above distribution, calculate

(a) the mean.
(b) the standard deviation.
(c) the median.
(d) the quartiles Q_1 and Q_3.
(e) the deciles D_1 and D_9.
(f) the percentiles P_5 and P_{95}.
(g) Pearson's first and second coefficients of skewness.
(h) the modal downtime of the distribution by the empirical formula (using the results obtained in part (a) and part (c) only). Compare this result with the mode obtained in part (g).

23. Consider the following frequency distribution of weights of 150 bolts:

Weight (grams)	Frequency
5.00 and less than 5.01	4
5.01 and less than 5.02	18
5.02 and less than 5.03	25
5.03 and less than 5.04	36
5.04 and less than 5.05	30
5.05 and less than 5.06	22
5.06 and less than 5.07	11
5.07 and less than 5.08	3
5.08 and less than 5.09	1

(a) Calculate the mean and standard deviation of the weights of bolts to three decimal places.

(b) Estimate from the frequency distribution, the number of bolts which are within one standard deviation of the mean.

(c) Suppose that each bolt has a nut attached to it to make a nut-and-bolt. Nuts have a distribution of weights with a mean of 2.043 grams and standard deviation 0.008. Calculate the standard deviation of the weights of nut-and-bolts.

24. A random sample of 11 vouchers is taken from a corporate expense account. The Voucher amounts are as follows:

$276.72	194.17	259.83	249.45
201.43	237.66	199.28	211.49
240.16	261.10	226.21	

Compute:
(a) the range;
(b) the interquartile range;
(c) the variance (definitional and computational);
(d) the standard deviation;
(e) the coefficient of variation.

25. A hardware distributor reports the following distribution of sales from a sample of 100 sales receipts.

Dollar Values of Sales	Number of Sales (f)
$ 0 but less than 20	16
20 but less than 40	18
40 but less than 60	14
60 but less than 80	24
80 but less than 100	20
100 but less than 120	8
Total	100

Find:
(a) the variance (definitional and computational);
(b) the standard deviation;
(c) the coefficient of variation.

26. The National Space Agency requires that all resistors used in electronic packages assembled for space flight have a coefficient of variation less than 5 percent. The following resistors made by the Mary Drake Company have been tested with results as follows:

Resistor	Mean Resistance (K-ohms)	Standard Deviation (K-ohms)
A	100	4
B	200	12
C	300	14
D	400	16
E	500	18
F	600	20

Which of the resistors meets specifications?

27. Salaries paid last year to supervisors had a mean of $25,000 with a standard deviation of $2000. What will be the new mean and standard deviation if all salaries are increased by $2500?