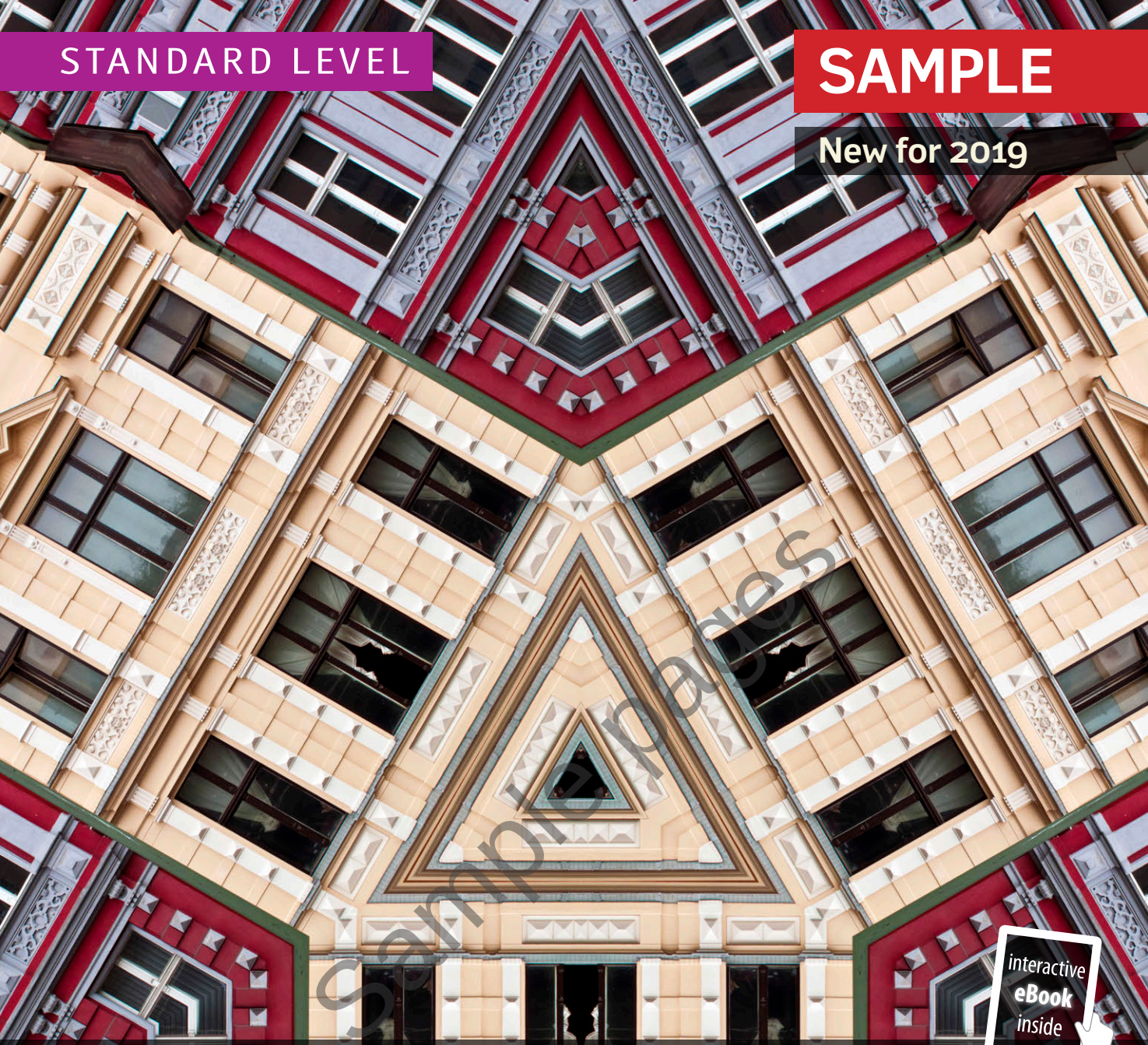
STANDARD LEVEL

SAMPLE

New for 2019



interactive eBook inside

Mathematics

Analysis and Approaches

For the IB Diploma

IBRAHIM WAZIR
TIM GARRY

Pearson

Contents

Introduction

1 Number, algebra and function basics

2 Functions

3 Sequences and series

4 Exponential and logarithmic functions

5 Trigonometric functions and equations

6 Geometry and trigonometry

7 Statistics

8 Probability 1

9 Differential calculus 1

10 Differential calculus 2

11 Integral calculus

12 Probability 2

The mathematical exploration - Internal assessment

Theory of knowledge

Answers

Index

Learning objectives

By the end of this chapter you should be familiar with...

- concepts of population, sample, random sample, and frequency distribution of discrete and continuous data
- reliability of data sources and bias in sampling
- sampling techniques and their effectiveness
- interpretation of outliers
- presentation of data using frequency tables and diagrams and box-and-whisker plots
- working with grouped data: mid-interval values, interval width, upper and lower interval boundaries, and frequency histograms
- calculating and interpreting the mean, median, mode, quartiles, and percentiles
- calculating and interpreting the range, interquartile range, variance, and standard deviation
- calculating and interpreting cumulative frequency graphs and using them to find the median, quartiles, and percentiles
- understanding and interpreting linear correlation of bivariate data
- working with linear regression.

Statistics are a part of everyday life, and you are likely to encounter them in one form or another on a daily basis.

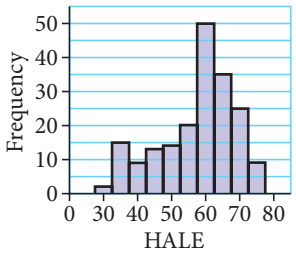


Figure 7.1 HALE data

For example, the World Health Organization (WHO) collects and reports data about worldwide population health on all 192 UN-member countries. Among the indicators reported is the health-adjusted life expectancy (HALE). This is based on life expectancy at birth, but includes an adjustment for time spent in poor health. It is most easily understood as the equivalent number of years in full health that a newborn can expect to live, based on current rates of ill-health and mortality. According to WHO rankings, lost years due to disability are substantially higher in poorer countries. Several factors contribute to this trend, including injury, blindness, paralysis, and the debilitating effects of tropical disease.

Of the 192 countries ranked by WHO, Japan has the highest healthy life expectancy (75 years) and Sierra Leone has the lowest (29 years).

Reports like this are commonplace in business publications, newspapers, magazines, and on the internet. There are some questions that come to mind as we read such a report. How did the researchers collect the data? How can we be sure that these results are reliable? What conclusions should be drawn from this report? The increased frequency with which statistical techniques are used in all fields, from business to agriculture to social and natural sciences, leads to the need for statistical literacy – familiarity with the goals and methods of these techniques – to be a part of any well-rounded educational programme.

Since statistical methods for summary and analysis provide us with powerful tools for making sense out of the data we collect, in this chapter we will first

start by introducing two basic components of most statistical problems – population and sample – and then delve into the methods of presenting and making sense of data. This will include some basic techniques in **descriptive statistics** – the branch of statistics concerned with describing sets of measurements, both samples and populations.

7.1 Graphical tools

Once you have collected a set of measurements, how can you display this set in a clear, understandable, and readable form? First, you must be able to define what is meant by measurement or 'data' and to categorize the types of data you are likely to encounter. We begin by introducing some definitions of the new terms in the statistical language that you need to know.

In the language of statistics, one of the most basic concepts is **sampling**. In most statistical problems, we draw a specified number of measurements or data – a **sample** – from a much larger body of measurements, called the **population**. On the basis of our observation of the data in the well-chosen sample, we try to describe or predict the behaviour of the population.

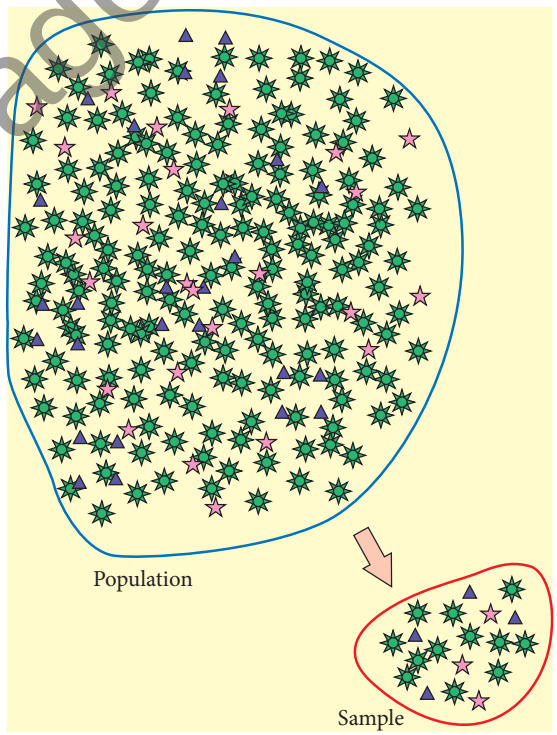


Figure 7.2 A sample is drawn from a population

A population is any entire collection of people, animals, plants, or things from which we may collect data. It is the entire group we are interested in, which we might wish to describe or draw conclusions about.

In order to make generalizations about a population, a sample is often studied. The sample should be representative of the population. For each population there are many possible samples.

For example, a study about the usage of resources in the households of an EU country stated that:

'… in the sample of 1674 households surveyed, the amount of water used by each washing cycle is given in the following…, The average time for each cycle was reported to be 42 minutes… . It was also discovered that the amount of laundry done by a household every year is related in some way to the household's income…'

In this example, the population is households' usage of water for washing, the average time spent on laundry, income, etc. The sample is the set of measurements of 1674 households that took part in the study. Notice that the population and sample are the measurements and not the people! The households are 'experimental units' or subjects in this study.

Height is a variable that changes with time for an individual and from person to person. If you gather the heights of the students at your school, the set of measurements you get is called a **data set**.

A **variable** is a characteristic that might vary over time or for different objects under consideration. When a variable is measured, the set of measurements obtained is called the **data** about that variable.

When a large amount of data is collected it becomes difficult to see what it means. The statistician's job is to summarize the data succinctly, bringing out the important characteristics of the numbers so that a clear and accurate picture emerges. There are several ways of summarizing and describing data, including tables, graphs, and numerical measures.

In everyday life, the terms 'reliability' and 'validity' are often used interchangeably. In statistics, however, these terms have specific meanings relating to different properties of the statistical or experimental method.

When looking at statistical results, we must be aware of how the data has been collected by assessing its **reliability** and **validity**.

Reliability, or **reproducibility**, is another word for consistency. It refers to the capacity of a test or method to produce the same result for two identical states or, more operationally, the closeness of the initial estimated values to the subsequent estimated values. For example, if one person takes the same personality test several times and always receives the same results, the test is reliable.

A test is valid if it measures what it is supposed to measure. If the results of the personality test claimed that a very shy person was in fact outgoing, the test would be invalid.

Reliability and validity are independent of each other. A measurement may be valid but not reliable, or reliable but not valid. Suppose your bathroom scale was reset to read 5 kg lighter than the actual weight. The reading it gave would be reliable, as it would be the same every time, but it would not be valid, since it would be lower than your correct weight.

Classification of variables

Numerical or categorical

Data can be classified into two main types: **numerical** (or **quantitative**) and **categorical** (or **qualitative**) data.

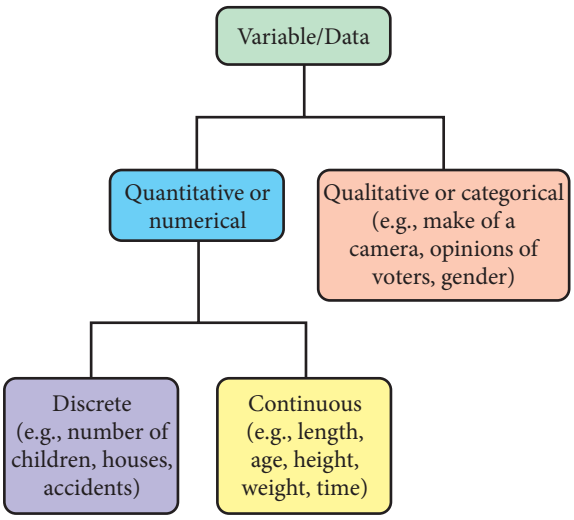


Figure 7.3 Data classifications

Numerical, or quantitative, variables measure a numerical quantity or amount on each experimental unit. This type of data always yields a numerical response.

There are two types of numerical data.

- **Discrete** data can take only particular values. For example, if you are counting the number of students that take a particular class the values will all be integers. It makes no sense to have 0.5 students.
- **Continuous** data can take any value, subject to the accuracy with which you can measure it. For example, the time it takes a student to travel from home to school could potentially be measured to the nearest second, although it might not be appropriate to measure to this level of accuracy.

Examples of numerical data include yearly income of company presidents, heights of students at school, time taken for students to finish their lunch at school, and total score received on exams.

There are two types of continuous variable.

- **Interval** variables can be measured along a continuum and the difference between two values on the continuum is meaningful.
- **Ratio** variables are interval variables with the additional condition that a value of 0 indicates that there is none of that variable. The name ratio reflects the fact that you can use the ratio of the measurements. So, for example, a distance of 20 m is twice as large as a distance of 10 m.

Temperature measured in degrees Celsius or Fahrenheit is an example of an interval variable. The difference between 20 °C and 30 °C is the same as the difference between 30 °C and 40 °C, but a temperature of 40 °C is not twice as hot as a temperature of 20 °C, because 0 °C does not mean there is no temperature.

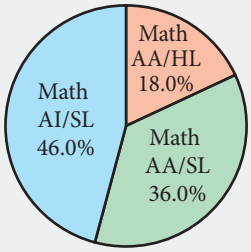
However, temperature measured in Kelvin is a ratio variable, because 0 Kelvin (often called absolute zero) indicates that there is no temperature whatsoever. A temperature of 100 Kelvin is twice as hot as 50 Kelvin. Other examples of ratio variables include height, mass, distance, and many more.

Categorical, or qualitative, variables measure a quality or characteristic of the experimental unit. Categorical data yields a qualitative response, such as colour.

We often use **pie charts** to summarize categorical data or to display the different values of a given variable (for example, percentage distribution). This type of chart is a circle divided into a series of segments. Each segment represents a particular category. The ratio of the area of each segment to the area of the circle is the same as the ratio of the corresponding category to the total data set.

Pie charts usually show the component parts of a whole. Often you will see a segment of the drawing separated from the rest of the pie in order to emphasize an important piece of information.

This pie chart shows how students in a large school are categorized into the IB Mathematics class they are taking. This is an example of qualitative data. There are 230 students in the Math AI/SL class, 180 students in the Math AA/SL class and 90 students in the Math AA/HL class. The pie chart shows what percentage of students take each class.



Frequency distributions

After collecting data, you should try to organize it so that it can be read easily. Methods for organizing data include **ordered arrays** and **stem-and-leaf diagrams** – not required.

In its raw form, your data may be listed in the order you collected it:

24, 26, 24, 21, 27, 27, 30, 41, 32, 38

Ordering the data in a ordered array, in either ascending or descending order, makes it easier to spot patterns and to start to understand the data:

21, 24, 24, 26, 27, 27, 30, 32, 38, 41

Suppose a consumer organization is interested in studying weekly food and living expenses of college students. A survey of 80 students yielded the data shown in Table 7.1. Expenses are given to the nearest euro.

38	50	55	60	46	51	58	64	50	49	48	65	58	61	65	53
39	51	56	61	48	53	59	65	54	54	54	59	65	66	47	49
40	51	56	62	47	55	60	63	60	59	59	50	46	45	54	47
41	52	57	64	50	53	58	67	67	66	65	58	54	52	55	52
44	52	57	64	51	55	61	68	67	54	55	48	57	57	66	66

Table 7.1 Weekly food and living expenses of college students

In its raw form, it is difficult to find any patterns or draw conclusions from this data. The first step in analysing data is to create a summary. This should show the following information:

- What values of the variable have been measured?
- How often has each value occurred?

Such summaries can be done in many ways. The most useful are **frequency distributions** and **histograms**. There are other methods of presenting data, some of which we will discuss later.

A **frequency distribution** is a table used to organize data. The left column, called **classes** or **groups**, includes numerical intervals on the variable being studied. The right column is a list of the **frequencies**, or **number of observations**, for each class. Intervals are normally of equal size. They must cover the range of the sample observations and they must not overlap.

Consider a frequency distribution for the living expenses of 80 college students. If the frequency distribution contained the intervals '35–40' and '40–45', to which of these two classes would a person spending €40 belong? More appropriate intervals would be '35 or more but less than 40' and '40 or more but less than 45'.

If classes are described with discrete limits such as '30–34', '35–39', then the boundaries are midway between the neighbouring endpoints. That is, the classes will be considered as '29.5 or more but less than 34.5', '34.5 or more but less than 39.5', Here the boundaries are 29.5, 34.5, 39.5, and each class width is 5..

Construction of a frequency distribution

There are some general rules for preparing frequency distributions that make it easier to summarize data and to communicate results.

Rule 1: Classes must be **inclusive** and **non-overlapping**. Each observation must belong to one, and only one, class interval. The **boundaries**, or **endpoints**, of each class must be clearly defined.

Rule 2: Determine k, the number of classes. Practice and experience are the best guidelines for deciding on the number of classes. In general, it is reasonable to have between 5 and 10 classes, but this is not an absolute rule. Practitioners use their judgement in these issues. If there are too few classes, some characteristics of the distribution will be hidden. If there are too many, some characteristics will be lost with the detail.

Rule 3: Intervals should be the same width. The width is determined by the formula

$$\text{interval width} = \frac{\text{largest number} - \text{smallest number}}{\text{number of intervals}}$$

Both the number of intervals and the interval width should be rounded up, possibly to the next integer. The above formula can be used when there are no natural ways of grouping the data. If this formula is used, the interval width is generally rounded to a convenient integer for easy interpretation.

Example 7.1

Organize the data from Table 7.1 into a frequency distribution, using appropriate class intervals.

Solution

Start by putting the data in ascending order.

38	39	40	41	44	45	46	46	47	47	47	48	48	48	49	49
50	50	50	50	51	51	51	51	52	52	52	52	53	53	53	54
54	54	54	54	54	55	55	55	55	55	56	56	57	57	57	57
58	58	58	58	59	59	59	59	60	60	60	61	61	61	62	63
64	64	64	65	65	65	65	65	66	66	66	66	67	67	67	68

With the data in order, we can immediately see that the smallest value is €38 and the largest value is €68. A reasonable grouping with nice round numbers is '35 or more but less than 40' and '40 or more but less than 45', etc. This gives a class width of 5.

Living expenses (l)	Number of students	Percentage of students
$35 \leqslant l < 40$	2	2.50
$40 \leqslant l < 45$	3	3.75
⋮	⋮	⋮
$65 \leqslant l < 70$	13	16.25
Total	80	100.00

Frequency and percentage frequency distributions of weekly expenses

Grouping the data in a table, as in Example 7.1, allows us to see some of its characteristics. For example, we can observe that there are few students who spend as little as €35 to €45, while the majority of the students spend more than €45. Grouping the data also causes some loss of detail, as we cannot see from the table what the real values in each class are.

In the table in Example 7.1, the **class midpoint**, also known as the **mid-interval value**, can be used to represent the data in that interval. For example, 37.5 can represent the data in the first class, while 42.5 can represent the data in the 40 to 45 class. This will be discussed in more detail later in the chapter. The values at the ends of each class, such as 35 and 40, are known as the **interval boundaries**.

Histograms

We can visualize a frequency distribution graphically using a **histogram**.

A histogram is a graph that consists of vertical bars constructed on a horizontal line that is marked off with intervals for the variable being displayed. The intervals correspond to the class intervals in a frequency distribution table. The height of each bar is proportional to the number of observations in that interval. The number of observations can also be displayed above the bars.

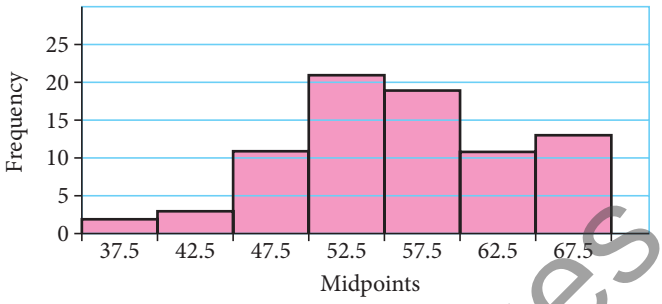
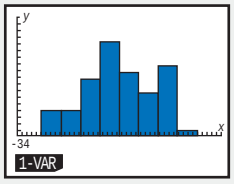


Figure 7.4 Histogram for the data in Example 7.1

Figure 7.4 shows a histogram for the data in Example 7.1. By looking at the histogram, it becomes visually clear that our previous observation is true. From the histogram we can also see that the distribution is not **symmetric**. You will find out more about the shape of frequency distributions later in this chapter.

Your GDC allows you to draw histograms. Different models will have different procedures. Here is a sample.



Cumulative and relative cumulative frequency distributions

A **cumulative frequency distribution** contains the total number of observations whose values are less than the upper limit for each interval. It is constructed by adding the frequencies of all the intervals up to and including the present interval. A **relative cumulative frequency distribution** converts all cumulative frequencies to cumulative percentages.

Table 7.2 shows a cumulative distribution and a relative cumulative distribution for the data in Example 7.1.

Living expenses (l)	No. of students	Cumulative number of students	Percentage of students	Cumulative Percentage of students
$35 \leqslant l < 40$	2	2	2.50	2.50
$40 \leqslant l < 45$	3	5	3.75	6.25
$45 \leqslant l < 50$	11	16	13.75	20.00
$50 \leqslant l < 55$	21	37	26.25	46.25
$55 \leqslant l < 60$	19	56	23.75	70.00
$60 \leqslant l < 65$	11	67	13.75	83.75
$65 \leqslant l < 70$	13	80	16.25	100.00
Total	80		100.00	

Table 7.2 Cumulative frequency and cumulative relative frequency distributions of weekly expenses

Notice how every cumulative frequency is added to the frequency in the next interval to give you the next cumulative frequency. The same is true for the relative frequencies..

As we will see later, cumulative frequencies and their graphs help in analysing data given in group form.

Cumulative frequency graphs

A **cumulative frequency graph**, sometimes called a **cumulative line graph** or an **ogive**, is a line that connects points that are the cumulative percentage of observations below the upper limit of each class in a cumulative frequency distribution. Figure 7.5 shows a cumulative frequency graph for the data in Example 7.1.

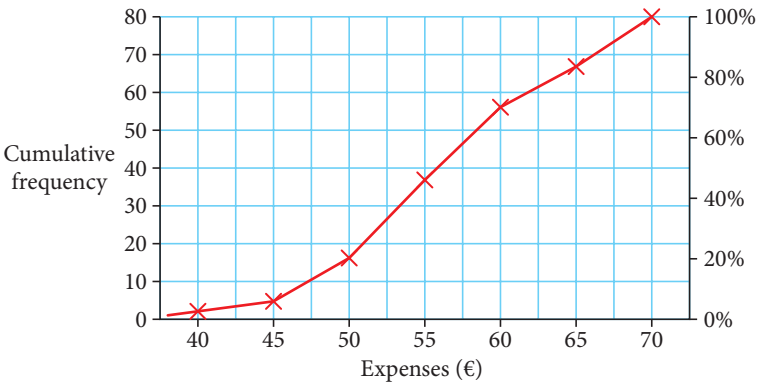


Figure 7.5 Cumulative frequency graph for the data in Example 7.1

Notice how the height of each line at the upper boundary represents the cumulative frequency for that interval. For example, at 50 the height is 16 and at 60 it is 56.

Example 7.2

The WHO data discussed in the introduction is given here in raw form.

(a) Prepare a frequency table, starting with a lower class boundary of 20 and a class interval of 5.

(b) Draw a histogram to represent the data.

(c) Draw a cumulative frequency graph to represent the data.

29	36	40	44	48	52	54	56	59	60	61	61	62	63	64	66	68	71	72	73	63	64	66	68
31	36	41	44	49	52	54	57	59	60	61	62	62	64	64	66	68	71	72	75	63	64	66	68
33	36	41	44	49	52	55	57	59	60	61	62	62	64	65	66	69	71	72	35	38	43	47	71
34	37	41	45	49	53	55	58	59	60	61	62	63	64	65	66	69	71	73	36	40	44	48	71
34	37	42	45	50	53	55	58	59	60	61	62	63	64	65	67	70	71	73	50	54	56	59	72
35	37	42	45	50	53	55	58	59	60	61	62	63	64	65	67	70	71	73	51	54	56	59	72
35	37	43	46	50	54	55	58	59	60	61	62	63	64	65	67	70	71	73	60	60	61	62	73
35	38	43	46	50	54	55	58	59	60	61	62	63	64	65	67	70	72	73	60	61	61	62	73

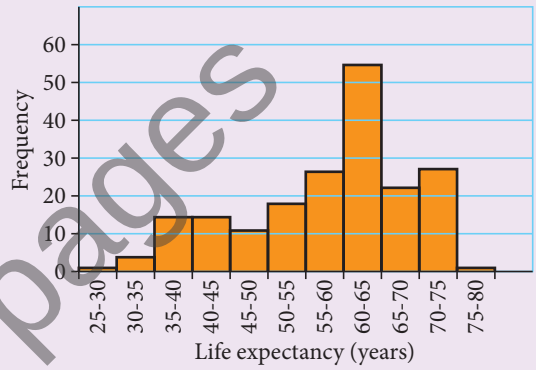
$25 \leqslant l < 30$ contains all observations larger than or equal to 25 but less than 30.

Solution

(a) First sort the data, then count every number in each class.

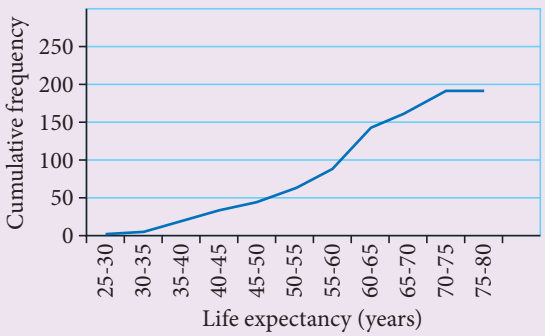
Life expectancy, l	Number of countries	Life expectancy	Number of countries
$25 \leqslant l < 30$	1	$55 \leqslant l < 60$	26
$30 \leqslant l < 35$	4	$60 \leqslant l < 65$	54
$35 \leqslant l < 40$	14	$65 \leqslant l < 70$	22
$40 \leqslant l < 45$	14	$70 \leqslant l < 75$	27
$45 \leqslant l < 50$	11	$75 \leqslant l < 80$	1
$50 \leqslant l < 55$	18		

(b) The histogram is shown on the right. Since all classes have equal width, the height and the area give the same impression about the frequency of the class interval. For example, the 60–65 class contains almost twice as many countries the 55–60 class, and the heights of the bars in the histogram reflect this, as do the areas. Similarly, the height of the 65–70 class is double that of the 45–50 class.



(c) In order to construct a cumulative frequency graph, we must first construct a cumulative frequency table.

Life expectancy	Number of countries	Cumulative number of countries	Life expectancy	Number of countries	Cumulative number of countries
$25 \leqslant l < 30$	1	1	$55 \leqslant l < 60$	26	88
$30 \leqslant l < 35$	4	5	$60 \leqslant l < 65$	54	142
$35 \leqslant l < 40$	14	19	$65 \leqslant l < 70$	22	164
$40 \leqslant l < 45$	14	33	$70 \leqslant l < 75$	27	191
$45 \leqslant l < 50$	11	44	$75 \leqslant l < 80$	1	192
$50 \leqslant l < 55$	18	62			



Sampling

Any study concerning populations needs data to be collected. Usually we do not collect data from the entire population. For statistical studies, data from samples is used. The method used to conduct a study is usually something like this:

1. Specify the population of interest.
2. Choose an appropriate sampling method.
3. Collect the sample data.
4. Analyse the pertinent information in the sample.
5. Use the results of the sample analysis to make an inference about the population.
6. Provide a measure of the inference's reliability.

Reasons for sampling

Taking a sample instead of conducting a census offers several advantages.

A **census** is a survey of the entire population.

A sample can save money and time. If an eight-minute interview is being undertaken, conducting the interviews with a sample of 100 people rather than with a population of 100 000 is obviously less expensive. In addition to the cost savings, the significantly smaller number of interviews usually requires less total time.

For given resources, the sample can broaden the scope of the study. With fixed resources, more detailed information can be gathered by taking a sample than by gathering information from the whole population. Concentrating on fewer individuals or items, the study can be broadened in scope to allow for more specialized questions.

Some research processes are destructive to the product or item being studied. For example, if light bulbs are being tested to determine how long they burn or if candy bars are being taste tested to determine whether the taste is acceptable, the product is destroyed.

If accessing the entire population is impossible, using a sample is the only option.

If sampling is deemed to be appropriate, it must be decided how to select a sample. Since the sample will be employed to draw conclusions about the entire population, it is crucial that the sample is **representative** of that population. It should reflect the relevant parameter of the population under consideration as closely as possible.

A **representative sample** is a sample that represents the characteristics of the population as closely as possible.

Random and non-random sampling

The two main types of sampling are **random** and **non-random**. In random sampling, every unit of the population has the same probability of being selected into the sample. Random sampling implies that chance enters into the process of selection.

In non-random sampling, not every unit of the population has the same probability of being selected into the sample.

Non-random sampling methods are not appropriate techniques for gathering data to be analysed by most of the statistical methods presented in this book.

Random sampling is also called **probability sampling** and non-random sampling is called **non-probability sampling**. Because every unit of the population is not equally likely to be selected in non-random sampling, assigning a probability of occurrence is impossible. The statistical methods presented and discussed in the IB syllabus assume that the data comes from random samples.

However, several non-random sampling techniques are described in this section, primarily to alert you to their characteristics and limitations.

Random sampling

We will discuss three basic random sampling techniques: **simple random sampling**, **stratified random sampling**, and **systematic random sampling**. Each technique offers advantages and disadvantages. Some techniques are simpler to use, some are less costly, and others show greater potential for reducing **sampling error**.

Sampling error occurs when, by chance, the sample does not represent the population.

Generally, all samples selected from the same population will give different results because they contain different elements of the population. Additionally, the results obtained from any one sample will not be exactly the same as those obtained from a census. The difference between a sample result and the result we would have obtained by conducting a census is called the **sampling error**, assuming that the sample is random and no non-sampling error has been made.

The sampling error is the difference between the result obtained from a sample survey and the result that would have been obtained if the whole population had been included in the survey.

Non-sampling errors can occur in both a sample survey and a census. Such errors occur because of human mistakes and not chance.

Simple random sampling

The most elementary random sampling technique is **simple random sampling**. Simple random sampling can be viewed as the basis for the other random sampling techniques. With simple random sampling, each unit of the sampling frame is numbered from 1 to N (where N is the size of the population). Next, a **random number generator** (or a **table of random numbers**) is used to select n items into the sample.

A **sampling frame** is a list of all the elements of a population from which a sample can be taken, such as a register of all the students in a college.

Example 7.3

Suppose it has been decided to interview 20 students from a school of 659 to form an understanding of their views of a new block-scheduling the school wants to adopt.

To find a simple random sample, number the students from 001 (or simply 1) to 659 and have a random generator choose 20 numbers. The students allocated to the chosen numbers form the sample.

The screenshot shows the first five from a list of random numbers generated by a GDC. Computer programs may be more efficient.

```
RanInt#(1,659,20)
{217,100,191,518,252▸
```

Stratified random sampling

In **stratified random sampling**, the population is divided into non-overlapping subpopulations called **strata**. The researcher then carries out simple random

sampling on each of the subpopulations. The main reason for using stratified random sampling is that it has the potential for reducing sampling error.

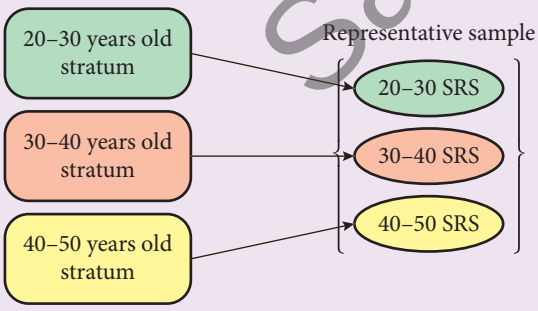
With stratified random sampling, the potential to match the sample closely to the population is greater than it is with simple random sampling because portions of the total sample are taken from different population subgroups. However, stratified random sampling is generally more costly than simple random sampling because each unit of the population must be assigned to a stratum before the random selection process begins.

Strata selection is usually based on available information. Such information may have been gleaned from previous censuses or surveys. The more different the strata are, the greater the benefits of using stratification. Internally, a stratum should be relatively homogeneous; externally, strata should contrast with each other. The process is demonstrated in Example 7.4.

Example 7.4

In FM radio markets, 'age of listener' is an important determinant of the type of programming used by a station.

The figure shows a stratification by age with three strata, based on the assumption that age makes a difference in preference of programming. This stratification assumes that listeners of 20 to 30 years of age tend to prefer the same type of programming, which is different from that preferred by listeners of 30 to 40 and 40 to 50 years of age. Within each age subgroup (stratum), **homogeneity** or alikeness is present; between each pair of subgroups, **heterogeneity** or difference, is present. An simple random sample is taken from each stratum. Together, the samples constitute a representative sample of the whole population.



An advantage of stratified random sampling is that, in addition to collecting information about the entire population, we can also compare different strata. In Example 7.4, the information we get will also help us compare the different age groups among each other.

Systematic random sampling

With **systematic random sampling**, every kth item is selected to produce a sample of size n from a population of size N. The value of k, sometimes called the sampling cycle, can be determined by the formula

$$k = \frac{N}{n}$$

If k is not an integer value, it should be rounded to the nearest integer.

Unlike stratified random sampling, systematic sampling is not done in an attempt to reduce sampling error. Rather, it is used because of its convenience and relative ease of administration.

Example 7.5

Given the data in Example 7.3, suppose we need to take a sample of 20 students using systematic sampling. First find k.

$$k = \frac{659}{20} \approx 32$$

From the list of 659 students, we randomly choose a starting number between 1 and 32. This might be 11, for example. After that we choose every 32nd number: 43, 75, 107, … .

Systematic sampling has other advantages besides convenience. Because systematic sampling is evenly distributed across the population, a knowledgeable person can easily determine whether a sampling plan has been followed in a study.

Non-random sampling

Sampling techniques used to select elements from the population by any mechanism that does not involve a random selection process are called **non-random sampling techniques**. Because chance is not used to select items from the samples, these techniques are **non-probability** techniques and are not desirable for use in gathering data to be analysed by standard methods of inferential statistics. Sampling error cannot be determined objectively for these sampling techniques. Two non-random sampling techniques are presented here: convenience sampling, and quota sampling.

Convenience sampling

In **convenience sampling**, elements for the sample are selected for the convenience of the researcher. The researcher typically chooses elements that are readily available, nearby, or willing to participate. The sample tends to be less variable than the population because in many environments the extreme elements of the population are not readily available. The researcher will select more elements from the middle of the population. For example, a convenience sample of homes for door-to-door interviews might include houses where people are at home, houses with no dogs, houses near the street, first-floor apartments, and houses with friendly people. In contrast, a random sample would require the researcher to gather data only from houses and apartments that have been selected randomly, no matter how inconvenient or unfriendly the location. If a research firm is located in a mall, a convenience sample might be selected by interviewing only shoppers who pass the shop and look friendly.

Quota sampling

Quota sampling appears to be similar to stratified random sampling at first glance. However, instead of selecting a simple random sample from each stratum, a non-random sampling method is used to gather data from one stratum until the desired quota of samples is filled. Quotas are described by setting the sizes of the samples to be obtained from the subgroups. Generally, a quota is based on the proportions of the subclasses in the population.

For example, a company is test marketing a new soft drink and is interested in how age groups react to it. An interviewer goes to a shopping mall and interviews shoppers of age group 16–20, for example, until enough responses are obtained to fill the quota.

In quota sampling, an interviewer starts by asking a few filter questions. If the respondent represents a subclass whose quota has been filled, the interviewer stops the interview.

Quota sampling can be useful if no previous information is available for the population. For example, suppose we want to stratify the population into cars using different types of winter tyres but we do not have lists of users of the 'Continental' brand of tyres. Through quota sampling, we would proceed by interviewing all car owners and casting out non-Continental users until the quota of Continental users is filled.

Quota sampling is less expensive than most random sampling techniques because it is a technique of convenience. Another advantage of quota sampling is the speed of data gathering. We do not have to call back or send out a second questionnaire if we do not receive a response; we just move on to the next element.

The problem with quota sampling is that it is a non-random sampling technique. Some researchers believe that a solution to this issue can be achieved if the quota is filled by randomly selecting elements and discarding those not from a stratum. This way quota sampling is essentially a version of stratified random sampling. The object is to gain the benefits of stratification without the high costs. However, it remains a non-probability sampling method.

Exercise 7.1

1. Identify the experimental units, sensible population and sample on which each of the following variables is measured. Then indicate whether the variable is quantitative or qualitative.

(a) Gender of a student.

(b) Number of errors on a final exam for 10th grade students.

(c) Height of a newly born child.

(d) Eye colour for children aged less than 14.

(e) Amount of time it takes to travel to work.

(f) Rating of a country's leader: excellent, good, fair, poor.

(g) Country of origin of students at international schools.

2. State what you expect the shapes of the distributions of the following variables to be: uniform, unimodal, bimodal, symmetric, etc. Explain why.

(a) Number of goals shot by football players during the last season.

(b) Weights of newborn babies in a major hospital during the course of 10 years.

(c) Number of countries visited by a student at an international school.

(d) Number of emails received by a high school student at your school per week.

3. Identify each variable as quantitative or qualitative.

(a) Amount of time to finish your extended essay.

(b) Number of students in each section of IB Maths SL.

(c) Rating of your textbook as excellent, good, satisfactory, terrible.

(d) Country of origin of each student in Maths SL courses.

4. Identify each variable as discrete or continuous.

(a) Population of each country represented by SL students in your session of the exam.

(b) Weight of IB Maths SL exams printed every May since 1976.

(c) Time it takes to mark an exam paper by an examiner.

(d) Number of customers served at a bank counter.

(e) Time it takes to finish a transaction at a bank counter.

(f) Amount of sugar used in preparing your favourite cake.

5. Grade point averages (GPA) in several colleges are on a scale of 0–4. Here are the GPAs of 45 students at a certain college.

1.8	1.9	1.9	2.0	2.1	2.1	2.1	2.2	2.2	2.3	2.3	2.4	2.4	2.4	2.5
2.5	2.5	2.5	2.5	2.5	2.6	2.6	2.6	2.6	2.6	2.7	2.7	2.7	2.7	2.7
2.8	2.8	2.8	2.9	2.9	2.9	3.0	3.0	3.0	3.1	3.1	3.1	3.2	3.2	3.4

Draw a histogram, a relative frequency histogram, and a cumulative frequency graph. Describe the data in two to three sentences.

6. The following are the grades of an IB course with 40 students on a 100-point test. Use the graphical methods you have learned so far to describe the grades.

61	62	93	94	91	92	86	87	55	56
63	64	86	87	82	83	76	77	57	58
94	95	89	90	67	68	62	63	72	73
87	88	68	69	65	66	75	76	84	85

7. The length of time (in months) between repeated speeding violations of 50 young drivers are given in the table below.

2.1	1.3	9.9	0.3	32.3	8.3	2.7	0.2	4.4	7.4
9	18	1.6	2.4	3.9	2.4	6.6	1	2	14.1
14.7	5.8	8.2	8.2	7.4	1.4	16.7	24	9.6	8.7
19.2	26.7	1.2	18	3.3	11.4	4.3	3.5	6.9	1.6
4.1	0.4	13.5	5.6	6.1	23.1	0.2	12.6	18.4	3.7

(a) Construct a histogram for the data.

(b) Would you describe the shape as symmetric?

(c) The law in this country requires that the driving licence be taken away if the driver repeats the violation within a period of 10 months. Use a cumulative frequency graph to estimate the fraction of drivers who may lose their licence.

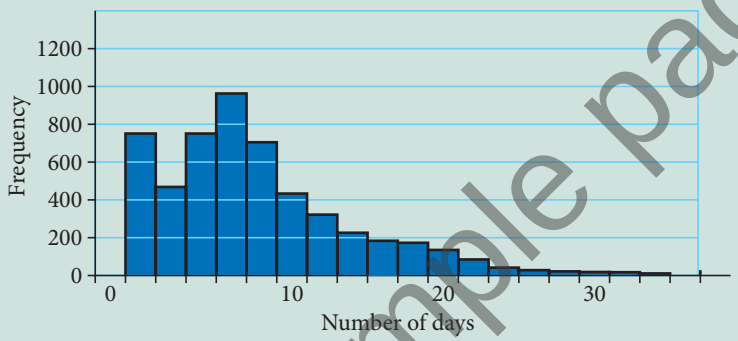
8. To decide on the number of counters needed to be open during busy times in a supermarket, the management collected data from 60 customers for the time they spent waiting to be served. The times in minutes are given in the following table.

3.6	0.7	5.2	0.6	1.3	0.3	1.8	2.2	1.1	0.4
1	1.2	0.7	1.3	0.7	1.6	2.5	0.3	1.7	0.8
0.3	1.2	0.2	0.9	1.9	1.2	0.8	2.1	2.3	1.1
0.8	1.7	1.8	0.4	0.6	0.2	0.9	1.8	2.8	1.8
0.4	0.5	1.1	1.1	0.8	4.5	1.6	0.5	1.3	1.9
0.6	0.6	3.1	3.1	1.1	1.1	1.1	1.4	1	1.4

(a) Construct a relative frequency histogram for the times.

(b) Construct a cumulative frequency graph and estimate the number of customers who have to wait 2 minutes or more.

9. The histogram below shows the number of days spent in hospital by heart patients in a certain country's hospitals in the 2015–2017 period.

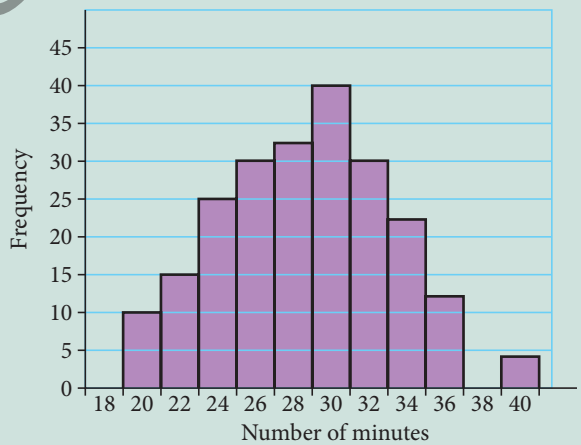


(a) Describe the data in a few sentences.

(b) Draw a cumulative frequency graph for the data.

(c) What percentage of the patients stayed less than 6 days?

10. One of the authors exercises on almost a daily basis. He records the length of time of his exercise on most of the days. Here is what he recorded for 2017.



(a) What is the longest time he has spent doing his exercises?

(b) What percentage of the time did he exercise more than 30 minutes?

(c) Draw a cumulative frequency graph for his exercise time.

11. Radar devices are installed at several locations on a main highway. Speeds, s, in km h^{-1} of 400 cars travelling on that highway are measured and summarised in the following table.

Speed	$60 \leqslant s < 75$	$75 \leqslant s < 90$	$90 \leqslant s < 105$	$105 \leqslant s < 120$	$120 \leqslant s < 135$	Over 135
Frequency	20	70	110	150	40	10

(a) Construct a frequency table for the data.

(b) Draw a histogram to illustrate the data.

(c) Draw a cumulative frequency graph for the data.

(d) The speed limit in this country is $130\ \text{km h}^{-1}$. Use your graph in **(c)** to estimate the percentage of the drivers driving faster than this limit?

12. Electronic components used in the production of computers are manufactured in a factory and their measures must be very accurate. Here are the lengths of a sample of 400 such components.

Length, l (mm)	< 5.00	$5.00 \leqslant l < 5.05$	$5.05 \leqslant l < 5.10$	$5.10 \leqslant l < 5.15$	$5.15 \leqslant l < 5.20$	More than 5.20
Frequency	16	100	123	104	48	9

(a) Construct a cumulative relative frequency graph for the data.

(b) The components must have a length between 5.01 and 5.18 mm, and any component with a length above 5.18 mm has to be scrapped. Use your graph to estimate the percentage of components that must be scrapped from this production facility.

13. The time, t, in seconds, that 300 customers wait at a supermarket checkout are recorded in the table below.

Time	$t < 60$	$60 \leqslant t < 120$	$120 \leqslant t < 180$	$180 \leqslant t < 240$	$240 \leqslant t < 300$	$300 \leqslant t < 360$	$t > 360$
Frequency	12	15	42	105	66	45	15

(a) Draw a histogram of the data.

(b) Construct a cumulative frequency graph of the data.

(c) Use the cumulative frequency graph to estimate the waiting time that is exceeded by 25% of the customers.

7.2 Measures of central tendency

When a data set is large, **summary measures** can help us to understand it. This section presents several ways to summarize quantitative data by calculating a **measure of central tendency**, also called a **measure of location** (a value that is representative of a typical data item), and a **measure of spread** (a value that indicates how well the typical value represents the data). These measures can be used in addition to or instead of tables and graphs.

The farthest we can reduce a set of data, and still retain any information at all, is to summarize the data with a single value. Measures of location do just that: they try to capture with a single number what is typical of the data. What single number is most representative of an entire list of numbers? We cannot say without defining 'representative' more precisely.

We will study three common measures of location: the **mean**, the **median**, and the **mode**. The mean, median, and mode are all 'most representative', but for different, related notions of representativeness.

- The arithmetic mean is commonly called the average. It is the sum of the data, divided by the number of data:

$$\text{mean} = \frac{\text{sum of data}}{\text{number of data}} = \frac{\text{total}}{\text{number of data}}$$

- The median of a set of measurements is the value that falls in the middle position when the data are sorted in ascending order. In a histogram, the median is the value that divides the histogram into two equal areas.
- The mode of a set of data is the most common value among the data. It is rare that several data coincide exactly, unless the variable is discrete, or the measurements are reported with low precision.

When these measures are computed for a population they are called **parameters**. When they are computed from a sample they are called **statistics**.

A **statistic** is a descriptive measure computed from a sample of data.
A **parameter** is a descriptive measure computed from an entire population of data.

Measures of central tendency provide information about a typical observation in the data or locate the data set.

Mean, median, mode

The mean

The most common measure of central tendency is the arithmetic mean, usually referred to simply as the 'mean' or the 'average.'

Example 7.6

The five closing prices of the NASDAQ Index for the first business week in November 2007 are given below. This is a sample of size $n = 5$ for the closing prices from the entire population.

2794.83 2810.38 2795.18 2825.18 2748.76

Find the average closing price.

Solution

$$\text{Average} = \frac{2794.83 + 2810.38 + 2795.18 + 2825.18 + 2748.76}{5} = 2794.87$$

Because this average was calculated from a sample, it is called the **sample mean**.

A second measure of central tendency is the **median**, which is the value in the middle position when the measurements are ordered from smallest to largest. The median of this data can only be calculated if we first sort them in ascending order.

2748.76 2794.83 **2795.18** 2810.38 2825.18

The **arithmetic mean** or **average** of a set of n measurements is equal to the sum of the measurements divided by n.

The sample mean: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$, where n is the sample size.

This is a **statistic**.

The population mean: $\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + x_3 + \cdots + x_N}{N}$ where N is the population size.

This is a **parameter**.

It is important to observe that you normally do not know the population mean, μ. It is usually estimated with the sample mean, $\bar{x}$.

The median

The **median** of a set of n measurements is the value that falls in the middle position when the data are sorted in ascending order.

In Example 7.6 we calculated the sample median by finding the third measurement to be in the middle position. If the number of measurements is even, the process is slightly different.

Let us assume that you took six tests last term and that your marks were, in ascending order,

52, 63, 74, 78, 80, 89.

When the data are arranged in order, there are two 'middle' observations, 74 and 78.

52 63 (74) (78) 80 89

To find the median, choose a value halfway between the two middle observations. This is done by calculating the mean of the two middle vales:

$$m = \frac{74 + 78}{2} = 76$$

The position of the median can be given by $\frac{n+1}{2}$. If this number ends with a decimal, you need to find the mean of the adjacent values.

In the NASDAQ Index case, we have five observations. The position of the median is then at $\frac{5+1}{2} = 3$.

In the marks example, the position of the median mark is at $\frac{6+1}{2} = 3.5$, hence we find the mean of the numbers at positions 3 and 4.

Although both the mean and median are good measures for the centre of a distribution, the median is less sensitive to **extreme values** or **outliers**. For example, the value 52 in the previous example is lower than all the other test scores and is the only failing score. The median, 76, would not be affected by