CHAPTER 1

Basic Statistics

"The word 'statistic' is derived from the Latin status, which, in the middle ages, had become to mean 'state' in the political sense. 'Statistics', therefore, originally denoted inquiries into the condition of a state."

— Wynnard Hooper (1854–1935) - English author

Expected Learning Outcomes

- Explain the common terms used in statistics.
- Describe, interpret and calculate descriptive statistics.
- Distinguish the differences between confidence interval, standard error and standard deviation.
- Outline common study designs.
- Describe the parts of a research proposal.

Data

- Without data there would be no statistics!
- Data is Information that can be analysed, interpreted and presented statistically.
- Data can take many forms, digital data, personal data, sample data, laboratory experimental data, field data, population data.
- In statistics we categorize these into numerical and non-numerical data.

What is Statistics?

The earliest use of statistics came from rulers and governments, who wanted information (**data**), such as the number of people, resources (e.g. food, gold, land) in order to set taxes, fund infrastructure (building projects), raise and maintain armies, and go to war (Appendix 1). To make accurate decisions, ideally you would want to

Essential Statistics for Bioscientists, First Edition. Mohammed Meah.

collect all the information or data available about a defined group or category. This is called the **population (**the entire group of individuals or observations).

The modern equivalent of this is called a census or survey of the population (usually every 10 years) of a country, which collects information such as the total number of people, ethnicity, age, and gender. The data obtained is called the population data. Some examples include, those with a disease or condition (e.g. diabetes or hypertension), smoking, animals, or plants.

However, it is not practical or possible to get the population data most of the time, so we take a random sample which can be representative of the population. The **sample** is a defined group of individuals or observations such as smoking habits taken from an identified and specific population. The sample should be representative of the population and is chosen by setting inclusion and exclusion criteria. These criteria define the characteristics or features of the sample. For example, inclusion criteria could be healthy, males, aged 20–30; exclusion criteria participants (or subjects) not suitable for selection might be smokers, not on any medication or have a medical condition.

> **"By a small sample, we may judge of the whole piece."**
> *Miguel de Cervantes (1547–1616), from his novel "Don Quixote"; Spanish novelist, poet and playwright*

Statistics can also be defined as: a branch of mathematics which involves data collection, data presentation, data analysis, and interpretation of data which comes from a population or sample of the population. In statistics, we usually take data from population samples. A **population sample** consists of a certain proportion/percentage of the total population determined by the researcher. The bigger and more representative the sample size, the more valid would be the results of data analysis. For example, we take height measurements (an example of data collection) of 50 male and 50 female level 5 bioscience students. This sample of 100 students would be taken from a total population of 645 level 5 bioscience students, to determine the average (mean) height of both male and female students. We would then see whether this average height is representative of the average height of all the level 5 bioscience students.

Types of Statistical Methods

In statistics, we analyse sample data to make predictions and generalisations about the population data. The sample data collected is described according to the amount of data (total data), centre of data (average, middle and most commonly occurring) and spread of data (e.g. lowest value and highest value). These are described numerically and called **descriptive statistics**. The statistical terms used in descriptive statistics (e.g. mean, mode, median, standard deviations) are described below.

The statistical terms used in descriptive statistics using a variety of graphs (Chapter 2), this is called **exploratory statistics**. Further analysis to look for differences between sample data and population data, differences between two or more groups of sample data, or looking for associations or relationships between sample groups is called **inferential statistics.** The types of inferential

statistical tests used will be determined by data size, data distribution, data type and number of groups of data. Inferential tests can be classified into two types: **parametric** or tests which are based on known probability **distributions** of the population (Chapters 3 and 4), and **non-parametric** which do not follow a distribution (Chapters 3 and 5). A distribution in statistics describes the possible values and likely occurrences of these values (e.g. in tossing a coin, getting a head or tail, and the likelihood of getting heads or tails with further attempts) in experiments (Chapter 2).

Types of Data

Data can be observations or measurements. Data can be classified into **quantitative** (numerical – numbers) or **qualitative** (categorical – non-numerical).

Quantitative Data

This numerical data is split into continuous (lots of data collected in very small steps), and discrete (number of specific events occurring). **Discrete data** take specific number values (e.g. number of pregnancies, number of vaccinations) and give less information. **Continuous data** (e.g. height, weight, cholesterol, cell counts, concentration of substances in fluids, decimals, percentages, ratios) – very small steps, give more information; this type of data are used mainly in parametric tests (Chapter 4).

Qualitative Data

This non-numerical data can be split into **discontinuous data**, such as whole numbers, ranks, scales, gender, colours, species, classes, position in a race, and blood groups. This is also called **categorical data**, which can also be divided into **nominal data** (data which cannot be ranked in size, e.g. blood groups, gender, ethnicity) and **ordinal data** (data which can be ranked, e.g. position in a race, anxiety scale, pain scale, intensity of exercise). This type is predominantly used in non-parametric tests (Chapter 5).

Both quantitative and qualitative data can be expressed by the term **variable**. A variable is a specific factor, property, or characteristic of a population or a sample. It is the name given to the data that is collected, e.g. height, weight, gender, colour, and size.

Collection of Data

Data can be collected from observations (e.g. epidemiological study), surveys (e.g. questionnaires and interviews) or experiments (e.g. clinical trials, pilot study).

Framingham Heart Study (1948)

This was a famous medically important long term epidemiological study, initiated by the National Institute of Health (NIH). It collected a large amount of data on the epidemiology and risk factors of cardiovascular disease in 5209 adults in 1948. They identified hypertension (elevated blood pressure), high cholesterol (fat) levels, and cigarette smoking, as major risk factors for cardiovascular disease (Mahmood *et al.* (2014)).

Surveys

Ask a series of questions where the answers are subjective (based on personal opinion). These can be non-numerical or numerical by adding a number scale to the responses

e.g. (i) anxiety or pain – could be classified into none, mild, moderate, or severe

(ii) Likert Scale – agree, disagree, neither agree or disagree, strongly agree, e.g. module evaluation or describing a product you have bought

(iii) Visual Analogue Scale – an increasing number scale (e.g. 1 is low and 10 is high). An example of this is the Borg Rate of Perceived Exertion which scales the intensity of exercise.

Clinical Trial

A clinical trial compares the effect of one treatment with another. It involves patients, healthy people, or both in four stages (Phase 1 to Phase 4).

The rapid increase in infections and mortality caused by the coronavirus (Covid-19) worldwide in 2019 led to the need to develop effective vaccines. The two most popular vaccines in the UK are the Oxford–AstraZeneca and the Pfizer Biontech. Both of these underwent clinical trials after ethical approval. Each trial involved two groups, in which half the volunteers were given the vaccine and the other half were given a placebo. The groups were randomly selected and were matched (e.g. for age and gender). Volunteers did not know whether they were receiving the vaccine or the placebo, nor did the researchers know (double blind). Prior to using human volunteers, the vaccines were tested on animals.

The Oxford–AstraZeneca trial sampled 23,848 people across the UK, Brazil and South Africa between April and November 2020.

The Pfizer Biontech trial sampled 46,331 people from 153 sites around the world between July and December 2020.

Observations, hypotheses, theories

Before the 20th century, science was based on induction from observations from which theories were formed (e.g. Newton's laws of motion in 1664). Karl Popper, a philosopher in 1935, suggested that theories and laws should be put first as a hypothesis, which would then be tested by experimental hypothesis falsified by observations and experiments (e.g. Darwin's theory of evolution in 1859).

Experiments

An **experiment or study** gathers data from a sample (assumed to be drawn randomly from the population). Experiments test a hypothesis (or prediction) of something happening. For example, suppose you wanted to know if there was a difference between an old treatment and a new treatment for a disease condition. Two predictions which we can make about this are: (i) there will be no difference between the old treatment and the new treatment (called the **null hypothesis**) on the disease condition, or (ii) there will be a difference between the old and new treatments (called the **alternative hypothesis**) on the disease condition.

In experiments, two groups of variables need to be identified. One group is the **dependent variables,** which are the outcome measures of the experiment (e.g. cell size, rate of growth, and cholesterol levels). The other group is the **independent variables,** which are the variables being manipulated in the experiment (e.g. type of treatment (drug, surgery), type of activity (exercise, training) and type of nutrient (protein, vitamin)).

Research Proposal

The experiment could be part of a **research project (**a short- or long-term study**),** and can be done in the laboratory or field (natural environment). In starting a research project or applying for research funding, it is common to write a **research proposal.** The proposal describes the subject area of the research and has various parts which answer the following questions:

- What is the research area of interest and justification for doing the research?
- What is the aim of the research?
- How will the aim be investigated?
- What are the expected outcomes?

Parts of a Research Proposal

> **"The formulation of the problem is often more essential than its solution, which may be merely a matter of mathematical or experimental skill."**
> *-Albert Einstein (1879–1955) - Physicist*

1. **Background literature search**: to see what is known, what is unknown, and whether this study has been done before. This would involve doing literature searches, using key words from the title of the project and **databases** (e.g. in bioscience the most popular are PubMed, Science Direct, Google Scholar), reading primary sources (e.g. journal papers) and secondary sources (e.g. textbooks). It is very important to identify what is unknown or controversial, or where there is a gap of knowledge from previous studies, an observation, or initial investigation (pilot study). This would then be the basis of the rationale (reason) for doing the study. For example, suppose you were interested in doing a research project on exercise training and haemoglobin levels. You would do a literature search with your key words being exercise, training, and haemoglobin. Your search may produce a large number of studies, which

you will then need to filter, e.g. human studies, type of exercise, duration of training, gender, year of publication. You then need to identify gaps in knowledge or controversy, i.e. levels of haemoglobin between gender, ethnicity, types, and duration of exercise training.

2. **A clear aim**: typically, this would be a question you want to answer based on the unknown or controversy identified in the background literature search. Ideally, you would want to state a narrow aim (say in the form of a question). For our example, suppose you found in the literature a controversy about the duration of exercise affecting the levels of haemoglobin. Some studies showed there was a difference after three weeks and others showed that the difference was only found after six weeks. You could have an aim such as: Is there a difference in haemoglobin levels after three and six weeks of aerobic exercise training?
3. **Hypothesis**: a prediction of outcomes. This is stated as two parts:
   - **i.** Null hypothesis: there is no difference (in the mean variables) between the control and experimental conditions. For our example, the null hypothesis would be 'there is no difference in mean haemoglobin levels after three and six weeks of training'.
   - **ii.** Alternative hypothesis: there is a difference (in the mean variables) between the control and experimental conditions: 'there is a difference in mean haemoglobin levels after three and six weeks of training'.
4. **Methodology**: methods used to investigate the aim. There are four major parts in methods. An outline of the subjects (human or animal or material), equipment/techniques used, protocol (procedures or steps of the experiment), and analysis of data (statistical tests to be used). These parts would incorporate the following:
   - **i.** Health and safety issues (for the researchers, participants, environment).
   - **ii.** Ethical considerations (e.g. using humans or animals for data collection in the study/experiment).
   - **iii.** How large should the sample size be?
   - **iv.** How many replicates (repeat measurements)?
   - **v.** What are the dependent and independent variables? Independent variables are those being manipulated in experiments and dependent variables are those we measure to see the effect of the manipulation.
   - **vi.** How do you exclude confounding variables (factors that could adversely affect an experiment or other data collection procedure)?
   - **vii.** Control and experimental conditions (what is the baseline or reference and the interventions?)
   - **viii.** Analysis of results (descriptive and inferential statistics).
   - **ix.** Precautions (to ensure you collect relevant study data).

Example of Biomedical experiment methodology

Consider our example: an experiment investigating the effect of exercise training on haemoglobin levels in a group of healthy adults

i. Health and safety issues (for the researchers, participants, environment): taking blood, preventing infections, laboratory safety, disposal of used materials, e.g. needles, equipment safety will depend on type of equipment used (e.g. cycle ergometer, treadmill), injury during training, first aider availability.

ii. Ethical considerations (e.g. using humans or animals for data collection in the study/experiment): ethics application to Ethics Committee for approval.
iii. How large the sample size should be: depends on the power of experiment required, calculation of sample size.
iv. How many replicates (repeat measurements): depends on protocol and study design, e.g. could have two groups (one group does no training, one does training for six weeks, measure haemoglobin at the beginning and then week by week in both groups). Also measure blood pressure and heart rate to monitor training intensity.
v. What are the dependent and independent variables (variables being manipulated in experiment are independent variables and dependent variables are the variable we measure to see the effect of the manipulation): dependent variable would be haemoglobin levels, independent variable would be exercise training duration.
vi. How do you exclude confounding variables (factors that could adversely affect an experiment or other data collection procedure): environmental temperature constancy, pressure kept constant, time and day for experiment consistent, stronger design of experiment, e.g. matching subjects.
vii. Control and experimental conditions: this depends on the design, so no training for controls and training for experimental group. The two groups would be matched for state of health, fitness, and age.
viii. Analysis of results: differences in mean haemoglobin levels between controls and experimental group would be tested using an independent t-test.
ix. Precautions: subjects familiar with protocol, i.e. procedures, use of equipment, calibration of equipment, questionnaires to monitor food and fluid intake.

5. **Expected Results:** making predictions on outcomes based on existing literature and the proposal design.
6. **Costs**: of materials, consumables, equipment, animals, facilities, payments for volunteers, and investigators.
7. **Timeline**: breakdown of when various parts of the project will be completed typically as a **Gantt chart** (tabular display of breakdown of the project into various parts and associated timelines to completion of the parts (Figure 1.1)).

As you can see in the Gantt chart below, an estimate has been made of the time allocated to complete the research proposal, ethics application, learn and practice the protocol, collect data, analyse the results, and finally submit the final research

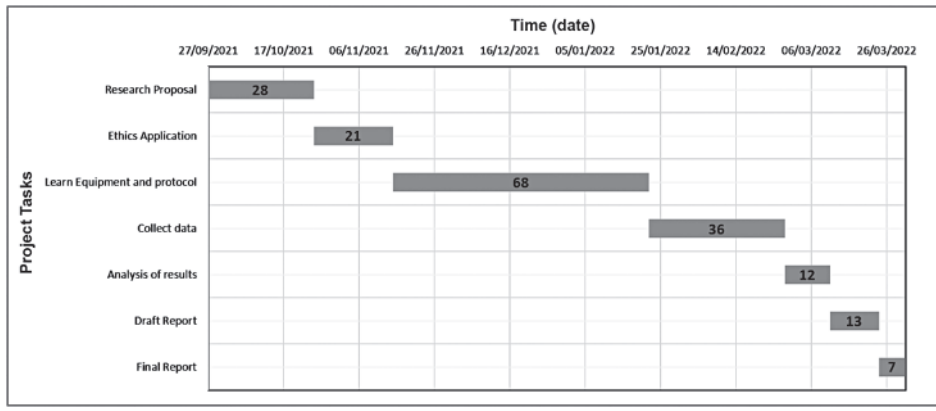

FIGURE 1.1 Gantt chart of research project tasks versus time (days).

report. The times may need to be altered due to problems encountered in the project, e.g. equipment and technique failure, availability of materials, volunteers and funding, and even unforeseen circumstances such as corona virus infections preventing research.

Deciding on the Research Experiment

To investigate the aim of the research and test the hypothesis, one or more experiments are needed. The effectiveness of experiments depends on several factors, including experimental design, sample size, and the statistical tests used for analysis.

Experimental Design It is crucial to think of the design of the experiment as this will determine whether the aim can be answered, and the type of statistical test you will be using to analyse your results (Chapter 3). How do you decide on the design? There are well known established designs (see section below on study design) which could fit your study.

Steps in choosing the design

Step 1: What is (are) the aim (aims) of your study? The clearer this is, the better, particularly if you can reduce the aim to a specific question.

Step 2: What will you be measuring (what variables) to see an effect, e.g. size, growth, symptoms, heart rate, cholesterol, absorbance?

Step 3: How will you measure your variables? What equipment or techniques and materials will you use?

Step 4: What will you be altering, or manipulating, or intervening (e.g. treatment, training, stress, concentration, dose of drug, vaccines, temperature)? So, you want to know the effect of this change on the measured variable.

Step 5: How many groups will there be, and how do you assign groups (see study designs below)? If you have one group, are they going to act as their own control, and then measure them again following the intervention? You might have two or more separate groups not related to each other.

Step 6: Design the protocol so that errors are reduced, by minimizing all the factors that could affect the experiment (e.g. environment, time of day, temperature, within subject factors (e.g. relaxed or not), between subject factors (type, age, gender, health).

Step 7: If possible, keep your design simple, and ask if it will allow you to collect data that may answer your aim. If your study doesn't fit the well-known study designs or has a complex design, then you may need further advice from your research supervisor or statistician.

Step 8: What analysis will you do: e.g. descriptive statistics, plot graphs, inferential tests to look for differences or associations in the means?

To illustrate the above steps in the design of a study, consider the biomedical study in the text box below.

An example of a biomedical study

Step 1: What is (are) the aim/aims of your study? The effect of arm exercise induced muscle soreness on creatine kinase levels in fit and unfit healthy adults.

Step 2: What will you be measuring (what are the dependent variables)? Creatine kinase (a quantitative marker of muscle damage), soreness scores (subjective scale of soreness).

Step 3: How will you measure your variables? Creatine kinase measured from 20 μL of blood from the finger and blood analyser (Reflotron); Visual Analogue Scale (0 for no soreness to 10 for severe soreness).

Step 4: What will you be altering or manipulating or intervening (independent variables)? Exercise of the upper arms (biceps contraction and relaxation).

Step 5: How many groups will there be, and how do you assign groups? Two groups of 10 healthy subjects, matched for age (age range 18–25 years) and fitness, of any gender. A fitness test on a treadmill (measurement of maximal oxygen uptake, the higher the value, the fitter you are) will be used to split volunteers into fit and unfit groups.

Step 6: What is the protocol? A week before the experiment, volunteers will have their fitness assessed by having their oxygen levels measured (using a metabolic analyser: Cosmed K5) whilst running to exhaustion (speed increased by 1 km hr^{-1} every minute) on a treadmill (Cosmed Mercury). They will then be split into two groups based on their maximal oxygen consumption ($VO2_{max}$), group A (fit) and group B (unfit). One week later, both groups will have their creatine kinase and soreness level measured at rest (control condition). Then they will perform arm exercise (biceps curl) at an intensity of 50% of their maximum in the standing posture until exhaustion. Creatine kinase and soreness levels will be measured at the end of the exercise and 30 minutes after exercise.

Precautions: No food 1 hour before the test, volunteers will be familiarized with the procedures; volunteers will do warm up exercise for 4 minutes at low intensity for both treadmill and arm exercise; safety procedures will be followed for use of equipment and taking blood samples.

Step 7: What is your design? Matched pairs design (groups have been matched for fitness).

Step 8: What analysis will be done? Descriptive statistics (N, mean, SEM, SD) will be calculated using Microsoft Excel. Mean creatine kinase and soreness levels will be plotted against conditions (control, exercise, recovery). Diferences in mean creatine kinase between the fit and unfit groups will be compared using one-way ANOVA (analysis of variance), with significance set at the 95% level ($P < 0.05$).

Note: Prior to developing the study aim, you would have researched the literature for the background and rationale, e.g. muscle soreness, a characteristic pain associated with exercise induced muscle damage (EIMD) in eccentric elbow flexor exercise that has a delayed onset. There is controversy about the levels of EIMD in trained and untrained subjects, and the levels of serum creatine released.

Having obtained an idea about the steps in choosing a study design, let us consider some additional factors to take into consideration in the study design.

Additional Factors to Consider in Deciding the Study Design

i. Control for the experiment: the baseline to which the intervention in the experiment will be compared to. Will one control be enough, or do you need more (e.g. positive and negative controls)?
ii. Intervention/ manipulation/ treatment: this is the input you are using to see if there is any effect.
iii. Type of sample, e.g. humans, animals, cells.
iv. If using humans: do not inform the subject about intervention (blind), i.e. subject does not know whether receiving placebo (looks like the treatment but lacks active ingredient or effect) or treatment.
v. Putting groups of subjects with similar characteristics together in a block.
vi. Subjects put into blocks by random selection: randomized experimental design.
vii. Controlled design: subjects carefully chosen.
viii. Replication: repetition of an experiment.
ix. Large enough sample size.

> **"If your experiment needs statistics, you ought to have done a better experiment."**
> *Ernest Rutherford (1871–1937), 1st Baron Rutherford of Nelson, New Zealand - born physicist*

We will now expand on the most important areas, namely controls, sample size, repeating experiments, and common study designs.

Controls Essential in the study design of any experiment is deciding on a control condition or group. The control group receives no treatment or receives the standard treatment. It is this group which is compared to the experimental or intervention group. The control is the baseline for comparing any changes caused by the intervention or experiment.

There are different types of controls: positive, negative, and blank. A positive control shows the expected result; they would receive an intervention where the expected result is known and would not receive the experimental intervention. A negative control shows the expected result does not occur, so the group or condition would receive something to show no response (e.g. placebo such as a sugar pill in place of the drug), and would not receive the experimental intervention. Sometimes a blank control (which does not have the substance you are testing for, e.g. distilled water) is used (particularly in calibrating instruments and dealing with solutions), before and after an experiment to check the baseline or standard condition.

An assay is a biological method to estimate the concentration of a substance (e.g. protein). A microbiological assay might compare the inhibition of microorganism growth by an antibiotic compared to concentrations of standard antibiotics. Thus, our main experiment is to treat the bacteria with the new antibiotic. Our negative control is treating with a non-antibiotic. Our positive control is to use a well-known established antibiotic.

There are different types of assay methods. One popular method is called ELISA (enzyme-linked immunosorbent assay) which is used to measure hormones (e.g. cortisol) or cytokines which are a variety of small proteins important in cell signalling, immunity and inflammation (e.g. IL6). The concentrations of these substances are very small (e.g. nanograms and picograms per ml). In this method, the positive control is a soluble sample containing the protein; the negative control is the soluble sample which does not have the protein, and a standard sample that has known concentration of the protein.

Sample Size for Studies Before an experiment is carried out it is useful to calculate the power of the experiment. The **power** is the probability (likelihood of an event occurring) that an experiment would detect a real difference, i.e. what sample size do you need? In clinical trials, the power is essential to see a treatment difference. A 90% power requirement means we want a 90% chance of detecting (value of variable measured) and require a sample size to be calculated (See Appendix 5). The power of a study is usually 80%, i.e. in 20% of cases we will miss the real difference and say there is no effect of the intervention in the experiment.

The power of a trial increases with increase in sample size. Estimating sample size is important because (i) if you use too large a sample you could waste time, money, and resources, e.g. doing a survey, (ii) if you use too small a sample you could get inaccurate results; there is less reliability on the sample data conclusions; you have to use less powerful statistical techniques for analysis, and it is difficult to infer conclusions about the population. Ideally you want a larger sample in order to make predictions about the population.

Repeating Measurements It is essential to make sure that measurements are reliable (can be repeated – precision) and are valid (accurate). **Reliability** is the consistency or precision or repeatability of a measuring test. **Validity** is a measure of the accuracy of the test. A test is valid if it measures what it claims to measure. Consider the aim of using a dart to hit a target number of 20 on a dart board. Then our validity is how often we hit the number 20. The more times we hit 20, the more accurate the result. Our precision would also be high because we are repeatedly hitting the target. However, if we hit another number like 1 a high number of times, then our accuracy would be low but our precision would be high. The worst scenario is a low precision and a low accuracy (dart hits are all over the dart board).

So regular calibration (checking what is being measured with a reference or known values) and repeat measurements are essential.

In sampling data, how many replicates are used needs to be considered. **Replicates** are repeat measurements. Typically common is to do three, e.g. three samples of the same or three repeats of the same measurement. The analysis would then take the highest, the median, or the average of the replicates for further analysis. For example, in assays, measuring the concentrations or absorbances three times. In lung function tests, it is common to do three repetitions of the blowing manoeuvres.

Another factor to consider in repeating measurements is **order effects**. For example, if the experiment involves a physical activity, fatigue may occur, and so adequate rest needs to be included in the protocol. Increasing familiarity with an experiment or **learning effect** may also affect learning outcome, occur, and so **randomization** of the task needs to be considered. These should be discussed in choosing a study design. Listed below are some familiar designs reported in publications.

Common Examples of Study Design

1. **Repeated measures design**: each subject performs in both the control and experimental condition which could be more than one (e.g. three different treatments or conditions). For example, to see the effect of three different drugs in treating blood pressure, each subject receives a drug, and the effect on blood pressure is measured after 1 week. So, placebo (looks like the drug but does not have the active ingredient, e.g. sugar substitute) in one week, drugs A, B, and C in subsequent weeks. So, all subjects have received the same drugs. You could just have one drug, so the design would be paired, no treatment (no drug) or after treatment (receives drug). This is a strong design as the subject is acting as their own control and ruling out inter-subject variation.
2. **Independent subjects design**: some subjects perform in condition A and others in condition B. Subjects must be allocated randomly to each group. For example, two separate (unrelated or independent) groups of patients take drug A for a disease condition. Or two independent groups, one group taking drug A and the other taking drug B for a condition. You want to know which group shows a bigger response to the drug. The design is prone to within-subject and inter-subject variation.
3. **Matched pairs design**: from the results of a pre-test, the subjects are sorted into matched pairs (pairs of equal abilities on the task to be measured). One person from each pair performs in the experimental condition and one in the control. Suppose you had a group of patients with different levels of severity (e.g. difficulty in breathing). You would measure their lung function and then allocate them to groups based on severity (mild, moderate, and severe). Then each group would receive the same treatment or drug. This is a strong design due to matching the prior condition of the subjects.
4. **Cross-over design**: subjects are allocated randomly to either group A or group B. Group A does test 1 and group B does test 2. Then they cross over, Group B does test 1 and group A does test 2. If you had a group of subjects all with a similar condition, you would just randomly (e.g. assign a number to each subject) split them into two groups A and B, if you want to see the effect of two drugs D1 and D2. Then group A would be given D1 and group B would be given D2. They would then be measured to see the effect of the drug. Then group A would be given the other drug (D2) and group B would be given D1, and you would repeat the measurement. This is a strong design.
5. **Single blind design**: subjects are not aware whether they received the treatment or the placebo. This is common in clinical trials.
6. **Double blind design**: neither the subjects nor the researchers are aware who has taken the intervention (e.g. drug) and who has taken placebo, until after the experiment is completed. This is an extremely strong design and is considered the gold standard in clinical trials.

Carrying Out the Experiment

The above designs are incorporated into studies which vary in duration, size, and type.

Types of Studies A study conducted over a short time period with limited data is called a pilot study. The preliminary findings of this often lead to a longer study. Commonly, studies can be split into cross-sectional or longitudinal studies. **Cross-sectional studies** are studies in which subject data is collected once or more but in a short time period. **Longitudinal studies** are those in which subject data is collected over a long time period (months to years, e.g. epidemiological studies). Observation studies (with no intervention), involve data being collected over time to see how factors affecting disease vary. **Prospective studies** (forward looking) observe groups over time. **Retrospective studies** (backward looking) look back at groups who have been exposed to a condition. **Clinical trials** or experimental trials are studies where groups receive an intervention (e.g. therapy, test, or procedure) and they are compared to groups who did not receive the intervention.

In addition to studies collecting new data, studies are also published from use of existing data from published articles or analysis of accumulated data, e.g. hospital databases of patient data. Two examples of these are **systematic reviews** and **meta-analysis**. These studies are more common in clinical fields to review treatments and techniques. However, final year undergraduate projects are increasingly using these types of studies.

Systematic reviews identify, select, and critically appraise a specific aim or clearly defined question based on clear inclusion and exclusion criteria. It attempts to find all available information available and then to clearly state reasons for including and excluding studies. It then assesses the quality and bias of included studies and provides recommendations and suggestions for addressing any knowledge gaps in the area. For example, to review the situation about the current Corona virus, a systematic review was conducted by Park *et al.* (2020) called 'A Systematic Review of COVID-19 Epidemiology Based on Current Evidence'. For more information on conducting a systematic review please see the article by Wright *et al.* (2007).

Meta-analysis is a statistical technique to combine the findings of selected studies (based on inclusion and exclusion criteria, contains quantitative data) to critically answer a clearly defined aim or question. It uses statistical methods to evaluate, synthesize and summarize the study findings objectively. An example of a meta-analysis study can be seen in the paper by Bjordal *et al.* (2004). It can be done independently or as part of a systematic review. Examples of the types of plots (Forest and Funnel) produced in meta-analysis are shown in chapter 3.

After the Experiment

After collecting sample data which has been stored on computer software files and memory sticks, in laboratory notebooks or printed out, the data would then be displayed in tables and graphs.

Display of Data This is explored to look for any patterns or trends. These include seeing if the data are close together or spread out, and whether there is an increase, a decrease, or no change. However, these would be difficult to see if you had a large amount of data. To make this easier, data is displayed in tables and different types of graphs (see Chapter 2). The tables and graphs can display both the

collected (raw) data and the calculated statistical data (which are also called summary statistics or descriptive statistics).

Summary Statistics The summary statistics describe the data using statistical terms in two ways. One is the **central tendency** (middle), for example, using the word 'average' or 'mean'. The average will replace the raw data with one value. The second way is to describe the data in terms of how **spread out** it is, for example, using the word 'range' to describe the lowest (minimum) and highest (maximum) values. In this chapter we will explain mean and range, and also consider other terms to describe central tendency and spread of data in a sample taken from the population.

Let us look at some raw and statistical summary data to illustrate some of the terms described in this chapter (Table 1.1).

As you can see, the data is organized into three areas: the **variables**, the individual or **raw data**, and the **descriptive statistics**. The columns show the variables measured with their units. The rows show the individual data. The descriptive statistics in the last four rows show the number of subjects (**N**), the mean and the terms used to describe variation in the data, standard deviation (**±SD**) and the variation of the mean (**±SEM**) which show the error of the mean. The next section will describe and explain these descriptive statistics including the symbols used (see Appendix 2 & 3 for a list of common statistical terms & symbols) and show examples of how to calculate them.

> "***He who loves practice without theory is like the sailor who boards ship without a rudder and compass and never knows where he may be cast.***"
> – Leonardo da Vinci (1452–1519) – artist, engineer

Descriptive Statistics

We have explained above that descriptive statistics summarize raw data. We will now describe more terms used to summarize data, i.e. the size of a sample, the shape of distribution (e.g. normal or bell shaped), central tendency (e.g. mean, mode, median) and the variability of the sample data (e.g. standard deviation, variance, range)

Sample Size

This is the total number of observations or data, the common symbol used is ***N***. This is very important, particularly if it is small because: (i) if *N* is small it is difficult to make conclusions about the population, (ii) it is difficult to use more powerful statistical tests such as parametric tests, (iii) it is difficult to make conclusions of clinical importance, (iv) it is difficult to show reproducibility or precision, (v) when high percentages are mentioned, the sample size could be very small (e.g. in surveys, advertising, polling). If *N* is large (typically > 30), then the sample distribution (see Chapter 3) will approximate more to the population distribution (sample mean and variance similar to population mean and variance).

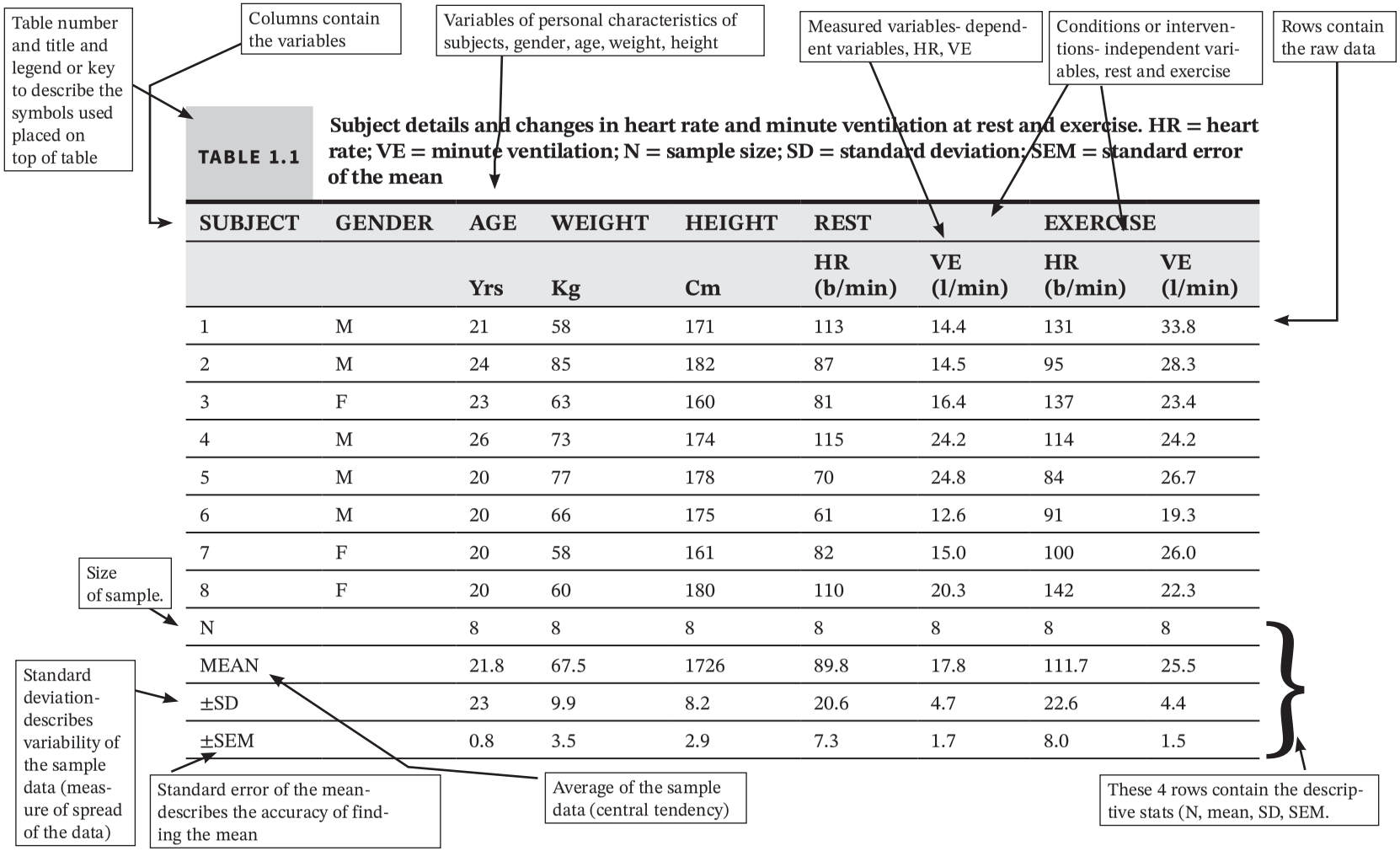

TABLE 1.1 **Subject details and changes in heart rate and minute ventilation at rest and exercise. HR = heart rate; VE = minute ventilation; N = sample size; SD = standard deviation; SEM = standard error of the mean**

SUBJECT	GENDER	AGE	WEIGHT	HEIGHT	REST		EXERCISE	
		Yrs	Kg	Cm	HR (b/min)	VE (l/min)	HR (b/min)	VE (l/min)
1	M	21	58	171	113	14.4	131	33.8
2	M	24	85	182	87	14.5	95	28.3
3	F	23	63	160	81	16.4	137	23.4
4	M	26	73	174	115	24.2	114	24.2
5	M	20	77	178	70	24.8	84	26.7
6	M	20	66	175	61	12.6	91	19.3
7	F	20	58	161	82	15.0	100	26.0
8	F	20	60	180	110	20.3	142	22.3
N		8	8	8	8	8	8	8
MEAN		21.8	67.5	1726	89.8	17.8	111.7	25.5
±SD		23	9.9	8.2	20.6	4.7	22.6	4.4
±SEM		0.8	3.5	2.9	7.3	1.7	8.0	1.5

Measures of Central Tendency

Mean Is the arithmetic average of all the values in a sample of data, i.e. the sum of all the data divided by the number of data. The symbol used is $\bar{X}$ (x bar). If we knew the mean of the data from the total population, then the symbol used would be μ (mu).

Mean ($\bar{X}$) = sum of data/number of data = $\sum x/N$

where x = data value; $\sum x$ = sum of data; N = total number of data.

Importance of the mean

- The most common and frequently used term to describe the central or middle view of data is the mean. The median and mode are not used as frequently.
- In inferential tests (hypothesis tests), differences in the mean between two or more groups are tested using probability.
- Experiments report the accuracy of the sample mean as an estimate of the population mean using standard errors and confidence intervals.

The symbol $\sum$ is called sigma. It is the sum or total from adding several items of data. Suppose we have three data values called x_1, x_2, x_3, then the sum would be: $\sum = x_1 + x_2 + x_3$

It is a measure of the central tendency (middle or centre) of set of data. It is markedly affected by very high or low values which are called outliers. It doesn't give information about the spread of the data. There are errors in calculating the mean as above if the scale of the measurements is not linear (e.g. pH is a log scale), if you are using ratios or percentages, finding the overall mean of sample means if the sample sizes are different. For example, mean blood pressure (pressure exerted in arterial blood vessels by the heart) is often quoted as 120/80, where the top figure is the systolic pressure (highest pressure generated by the heart) and the bottom figure is diastolic pressure (resting pressure of the heart).

Franklin D. Roosevelt

An American president who died in 1945 aged 63 years from cerebral haemorrhage (stroke) had a blood pressure of 300/190, an extreme case of hypertension!

EXAMPLE 1.1 | Calculation of the mean

Students were asked to hold their breath as long as possible after filling their lungs with air to maximum. The breath hold times in seconds measured with a stopwatch were:

55, 60, 50, 67, 71, 78, 75, 80, 84, 52, 65

Step 1: Count the number of sample data: here $N = 11$
Step 2: Find the sum of the sample data: here $\sum x = 737$
Step 3: Divide the sum of the data by the sample size

Thus, mean ($\bar{x}$) = sum of data/number of data = $\sum x/N$
or mean = 737/11 = 67 s

Median This is the central or middle value of a group of data, after the data is ordered (low to high). Unlike the mean it is not affected by extreme values (very low or very high values), however, it doesn't give the spread of the data. If there is an even number of data, then there will be two middle numbers, in which case the average of the two numbers will give the median.

EXAMPLE 1.2 | Calculation of the median for an odd sample

For the above breath data: 55, 60, 50, 67, 71, 78, 75, 80, 84, 52, 65

Step 1: First put the data in order of increasing size
50, 52, 55, 60, 65, 67, 71, 75, 78, 80, 84

Step 2: Find the middle number: in this case it is 67 s (with five numbers to the left of 67 and five numbers to the right of 67).

In an odd sample, the middle value can be found easily.

EXAMPLE 1.3 | Calculation of the Median for an even sample

If we added an extra student's breath hold time (e.g. 65 s) to the above data, we will now have an even sample of data:

55, 60, 50, 67, 71, 78, 75, 80, 84, 52, 65, 65

Step 1: Put sample data in order
50, 52, 55, 60, 65, 65, 67, 71, 75, 78, 80, 84

Step 2: In an even sample there will be two middle values, Identify the two middle numbers, i.e. 65 and 67

Step 3: Calculate the median by taking the average of the two middle numbers,
i.e. median = $(65 + 67)/2 = 66$ s.

Mode The mode is the value that appears most frequently (largest frequency). In a sample of data, the number of times a number occurs is counted.

EXAMPLE 1.4 | Calculation of the Mode

Volunteers in a research study had their weight in kg measured as shown below:

50, 50, 52, 55, 60, 65, 65, 65, 67, 71, 75, 78, 80, 84

EXAMPLE 1.4 *(continued)*

Step 1: Examine the data and count the number of times a number appears in the sample data. For the above data, we can see that the number 50 occurs twice, whilst the number 65 occurs three times, the rest of the numbers appear once.

Step 2: Therefore

The mode = the value which appears most frequently is = 65 kg.

Standard Error of the Mean (SEM) The mean of a sample is usually stated with a number called the standard error of the mean (SEM), which is an estimate of the precision of estimating the population mean. It is a measure of the accuracy of the mean. It would be written as mean ± SEM, where the ± SEM defines the range, interval, or band where the mean could be.

For example, if the mean red blood cell count from eight healthy subjects was 5.8 ± SEM 0.8 mill μL^{-1}.
The lower limit would be 5.8 – 0.8 = 5 mill μL^{-1}
The upper limit would be 5.8 + 0.8 = 6.6 mill μL^{-1}

Thus, the mean could be between 5 mill μL^{-1} to 6.6 mill μL^{-1} for this sample. You can compare your mean red blood cell count to published population means for red blood cell counts to see if it was normal or abnormal.

To understand SEM, consider the measurement of the weights of students in one classroom. We can calculate the mean for this classroom. This mean value will be in a range with an upper and lower limit just like the red blood cell count above. We can take a second sample by repeating the weights of students in another classroom and again we will have a mean and a range where the mean might be defined by an upper limit and a lower limit. This range where the mean could be found is called the standard error of the mean for that sample. If we now do this for all the classrooms in the school, we will have an overall mean. This overall mean is the population mean (the population would be all the classes in the school). So having more samples and larger samples would give us a mean which more accurately represents the population mean.

If the SEM range is small, then there is more accuracy in determining the mean. A smaller range is favoured if the sample size is large and there is a small variation of data (small standard deviation or variance). Another

SEM

- Is a measure of the precision or accuracy of the sample mean as an estimate of the population mean.
- Is equal to the standard deviation/square root of sample size.
- Is useful for calculating the confidence interval
- Decreases as sample size increases.
- If you repeat an experiment and each time calculate the sample mean, then the SD of the sample means is called the SEM.

term which is used to give an estimate of finding the mean is the confidence interval (see below). The standard deviation and variance are measures of the spread of data (see below). The standard error of the mean is equal to the standard deviation divided by the square root of the sample size.

$$\mathbf{SEM = SD/\sqrt{N}}$$

EXAMPLE 1.5 | Calculation of the standard error of the mean (SEM)

Consider the breath hold data from Example 1.1.

55, 60, 50, 67, 71, 78, 75, 80, 84, 52, 65

Step 1: Calculate the standard deviation (SD) of the sample data
Step 2: Find the square root of *N*
Step 3: Divide the SD by the square root of *N*

$$\text{SEM} = 11.704/\sqrt{11} = 3.53 \text{ s}$$

Step 4: Thus the ± SEM is = ± 3.53 s

Step 5: Since the mean is 67 s, then the mean breath hold ± SEM = 67 ± 3.53 s

The mean is the most common measure of central location or central tendency. However, as we have mentioned it is affected by large or small values and so the median avoids this, but it doesn't show the spread of the data. You will need to see which best describes your data, by exploring your data by graphical methods (see Chapter 2).

Variation of Data

The variation of the data (also known as dispersion or spread) can be described by the range, standard deviation, variance, coefficient of variation, quartiles, interquartile range, and percentiles. These will be described below.

Range This is a measure of the spread of the data, from the smallest to the largest size of your sample data. It is calculated by the largest value subtracting the smallest value. The disadvantage of the range is that it doesn't show the scatter of the data and extreme values affect it.

In biosciences, a special form of range called the **reference range** is very important particularly in disease diagnosis and to see whether you are in the normal range. These reference ranges are obtained from collating data from large studies. Consider the reference range for blood pressure in normal and hypertension (sustained and elevated blood pressure) conditions. The World Health Organization (WHO) define the following reference ranges for blood pressure:

EXAMPLE 1.6 | Calculation of the Range

For the sample data of age (years) of subjects:

18, 27, 45, 23, 27, 30

Step 1: Examine the data and identify the largest and smallest values; in this case it is 45 and 18 respectively

Step 2: Calculate the range

Range = highest value – lowest = 45 – 18 = 27 years.

115/75 to 120/80 mmHg is normal
120/80 to 139/89 mmHg is pre-hypertension
140/90 to 159/99 mmHg is stage 1 hypertension
>160/ > 100 mmHg is stage 2 hypertension

Regular measurements of blood pressure can then be used to monitor and assess the effect of the interventions of anti-hypertensive drugs or lifestyle changes (e.g. diet, exercise) on the blood pressure.

Blood Glucose (fasting) levels (from the National Institute for Clinical Excellence)

Normal: 4–5.6 mmol L^{-1} (72–100 mg/100 mL blood)
Pre-diabetic: 5.6–6.9 mmol L^{-1} (100–125 mg/100 mL
Diabetic: > 7 mmol L^{-1} (126 mg/100 mL)

In disease diagnosis, body fluids (e.g. urine, saliva, cerebrospinal fluid) are sampled. One of the most common body fluids which is regularly sampled is blood. A blood sample will have many reference ranges for a wide number of variables (Figure 1.2).

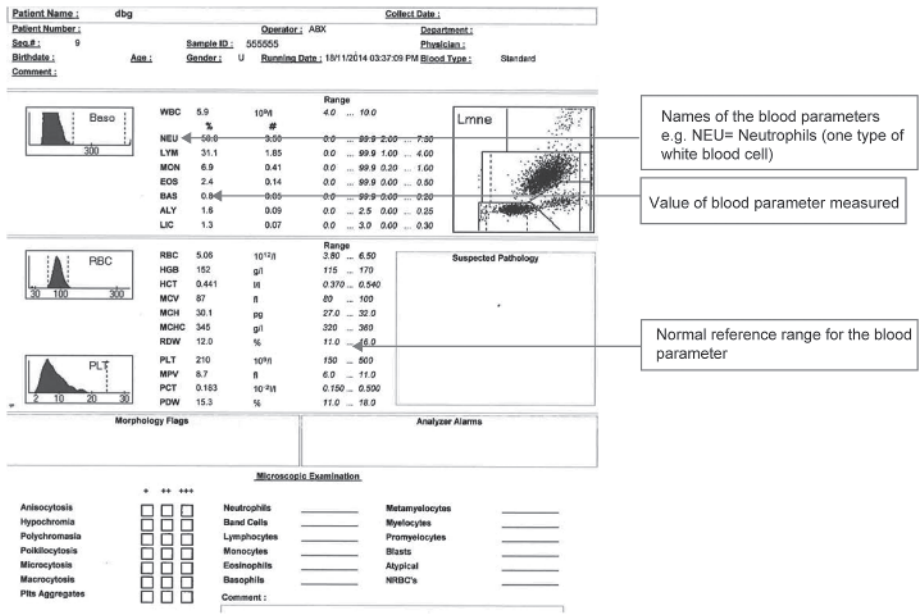

FIGURE 1.2 Blood parameters and reference ranges (from one person) using Pentra 60 Blood Analyser.

SD

- Is a measure of variability of sample data.
- It is an estimate of the variability of the population from which the sample is taken.
- SD^2 = variance.
- In a normal distribution, 95% of data will be within ± 2 SD.

Note: if the population is sampled, then the (N–1) is replaced by N in the formula for SD and the symbol used for population standard deviation is σ (sigma)

Standard Deviation (SD) This is a measure of the distribution of data around the mean; i.e. how far away from the mean an individual value lies. If the data is normally distributed, then SD tells us what proportion of the scores falls within certain limits: e.g. 68% of scores lie within ± 1 SD of the mean; 95% between ± 2 SD of the mean, and 99.7% lie within ±3 SD.

The larger the SD, the larger the range of values (variation) in the data.

The SD can be calculated for a sample dataset using the formula below.

SD = Square root of (sum of squares/ degrees of freedom)
$= \sqrt{[(\sum x_i - \bar{x}) / N - 1)]}$.

Where, $\bar{x} = mean$; X_i = each data value; $\sum(x_i - \bar{x})^2$ = sum of squares; N = sample size; $N - 1$ = degrees of freedom

Sum of squares

- Suppose we have three data values called $X1, X2, X3$ and the mean of these three values was $\bar{x}$.
- The difference between each data value and the mean would be $(X1 - \bar{x})$, $(X2 - \bar{x})$, $(X3 - \bar{x})$.
- The square of each difference would be $(X1 - \bar{x})^2$, $(X2 - \bar{x})^2$, $(X3 - \bar{x})^2$.
- The sum of these squared differences would be the sum of squares.

EXAMPLE 1.7 | Calculation of the standard deviation (SD) manually

Consider the body mass index (BMI) in kg m^{-2} measured in six women:
18, 27, 45, 23, 27,30

Step 1: Collate the data (Table 1.2).

Step 2–Step 7: These steps are shown in Table 1.2.

The SD = 9.16

TABLE 1.2 **Manual calculation of the standard deviation (SD) for values (x)**

	x	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
	18	-10.33	106.71
	27	-1.33	1.77
	45	16.67	277.89
	23	-5.33	28.41
	27	-1.33	1.77
	30	1.67	2.79
Mean	$\bar{x} = 28.33$		
Sum of squares			$\Sigma (x_i - \bar{x})^2$ 419.34
SD = Square root of (sum of squares/degrees of freedom)			$(\Sigma(x_i - \bar{x})^2) / (N-1) = \sqrt{(419.34/5)} = \sqrt{83.87} = 9.16$

Step 2: Find the mean of your data

Step 3: Subtract the mean from each value i.e. x_i-$\bar{x}$

Step 4: Square the answer in step 3

Step 5: Add the answers of step 3 (i.e. sum of squares)

Step 6: Divide the sum of squares by the degrees of freedom

Step 7: Take the square root of the answer for step 6

Variance This is an estimate of variability. It measures how far every data is from the mean and from each other. Thus, a small variance indicates that the data is close to the mean and to other values. It is very closely related to standard deviation, since the square of the standard deviation is equal to variance, i.e. **SD^2 = variance**. It is easier to work with mathematically, since SD has positive and negative values.

Variance = sum of squares/degrees of freedom

$$= (\sum(x_i - \bar{x})^2/(N-1)$$

Where, $\bar{x}$ = *mean*, N-1 = degrees of freedom, x_i = each value of sample, $(x_i - \bar{x})^2$ = sum of squares

EXAMPLE 1.8 | Calculation of the variance manually

Consider the BMI data from Example 1.7.
18, 27, 45, 23, 27, 30

Step 1: Tabulate the data (Table 1.3) and find the mean of the data
Step 2: Subtract the mean from each value,, i.e. x_i-$\bar{x}$
Step 3: Square the answer in step 2
Step 4: Add the answers of step 3 (i.e. sum of squares)
Step 5: Dividing the sum of squares by the degrees of freedom gives the variance

Thus, the variance is 83.87. If we take the square root of this, we will obtain the SD, which is 9.16.

TABLE 1.3 **Manual calculation of variance for data (x)**

	x	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
	18	-10.33	106.71
	27	-1.33	1.77
	45	16.67	277.89
	23	-5.33	28.41
	27	-1.33	1.77
	30	1.67	2.79
Mean	$\bar{x} = 28.33$		
Sum of squares		$\left(\sum (x_i - \bar{x})^2 = 419.34\right)$	
Variance (sum of squares/degrees of freedom)			$\left(\sum (x_i - \bar{x})^2\right) / (N - 1)$ $= 419.34 / 5 = 83.87$

Coefficient of Variation (CV%) The coefficient of variation (CV) is a measure of the variation of data. It is the standard deviation divided by the mean and expressed as a percentage. The higher the CV, the more variable the data.

$$\textbf{CV (\%) = (SD/mean) } \times \textbf{100}$$

Variance

- Is a measure of variability of sample data.
- Variance = square of the standard deviation (SD^2).
- Equal or unequal variance between two groups has to be tested before an independent t-test.
- Variance is compared between multiple groups in parametric tests such as ANOVA.

EXAMPLE 1.9 | Calculation of the Coefficient of variation (CV)

Consider the BMI data below:
18, 27, 45, 23, 27, 30

Step 1: Calculate the mean of the data, i.e. mean = 28.33
Step 2: Calculate the standard deviation of the data, i.e. SD = 9.16
Step 3: Divide the standard deviation by the mean and convert to a percentage by multiplying by 100, i.e. CV = (SD/Mean) x 100 = (9.16/28.33) x 100 = 32%

Quartiles The spread of sample data can also be described by splitting the data into four quartiles after putting them in order. The quartiles are split between the minimum value and the maximum value. Each quartile represents 25% of the data. The **first quartile (Q1)** describes 25% of the data below and 75% above it. The **second quartile (Q2)** represents 50% of the data below and 50% of the data above it. And the **third quartile** represents 75% of the data below it and 25% of the data above it. The **fourth quartile** has all the data below it. The **interquartile range (Q3–Q1)** represents the middle 50% of the data. These quartiles are displayed graphically in a box and whisker plot (see Chapter 2) and below (Figure 1.3). The box shows the middle part of the data. Quartiles are useful to see if data is skewed either to the lower or to the higher values, so it is a measure of variability.

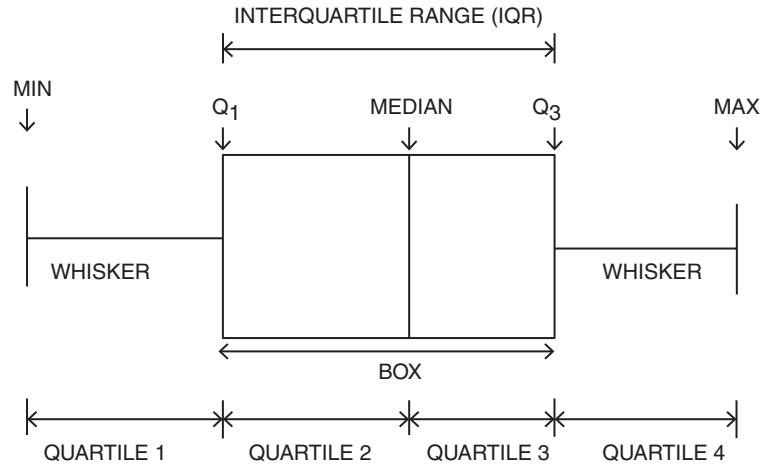

FIGURE 1.3 Sample data split into quartiles on a box and whisker plot.

If we measured the cholesterol levels of a group of subjects and displayed the results as quartiles in a box and whisker plot, we would know the following information: (i) what the lowest (minimum) and highest (maximum) cholesterol for the group is; (ii) the middle 50% of the group's cholesterol (from the interquartile range); the middle value (median) cholesterol for the group. For any individual we would know whether they were in the high cholesterol, low cholesterol, or middle range of values and where the value lies in relation to others in the group.

Percentiles Sample data can also be described by percentiles which are very similar to quartiles. Data which has been put in order are divided into 100ths or percentages. A percentile describes a score which describes what percentage of the other values are below it. For example, if your score is 68 and this is ranked as the 75th percentile, then 75% of the data values in the sample are below 68. Percentiles can be calculated using the formula: ***Percentile = (number of values below score)× (total number of values)/ 100***

The range and the interquartile range indicate overall variation in a dataset, but they don't show the detailed spread of the data. Standard deviation and variance show this detail. Variance is a more convenient way of looking at spread of data as it

doesn't involve the negative numbers of standard deviation. Plotting data graphically (see Chapter 2) for example box and whisker plots or frequency distributions will not only show the spread of data but also the central location of data.

EXAMPLE 1.10 | Calculation of quartiles

Ten students produced a blood smear glass slide. The slides were assessed for their quality and the following scores (out of 100) were given. Find the first, second, third quartiles and the interquartile range.

55, 30, 48, 51, 37, 66, 86, 73, 68, 75

Step 1: Put the numbers in order

30, 37, 48, 51, 55, 66, 68, 73, 75, 86

Step 2: Find the median
The two middle numbers are 55 and 66. Thus the median = (55 + 66)/2 = 60.5

Step 3: Find the first quartile (Q1)
From the numbers below 60.5, the middle number is = 48

Step 4: Find the middle number for the values above 60.5, the middle number = 73

Step 5: Find the interquartile range
The interquartile range = Q3-Q1 = 73–48 = 25

EXAMPLE 1.11 | Calculation of the percentile

In a reaction time test on a computer to the following scores (out of 100) were given. Calculate the percentile of the person who scored 73.

55, 30, 48, 51, 37, 66, 86, 73, 68, 75

Step 1: Put the scores in order
30, 37, 48, 51, 55, 66, 68, 73, 75, 86

Step 2: Count the number of values below the score of 73.
This is 7.

Step 3: Count the total number of scores.
This is 10.

Step 4: Calculate the percentile
Percentile = (number of values below the score of 73) x (total number of scores)/100)
Or = 7/ (10 x100)
= (7 x 10)/100
= 0.7
Or 70%.
This means that 70% of the scores were below 73.

Confidence Interval

To understand the concept of confidence interval better, the reader should read about probability (frequency) distributions like the normal distribution (*t* and *Z* distributions) in Chapter 2 and look at Chapter 3 to see how it relates to probability.

If studies were repeated on different random groups from a population, a range of results would be obtained (each random group will have different means and standard deviations). The confidence interval (CI) represents the intervals or range of values in which the mean occurs, i.e. a measure of reliability of the mean occurring. The CI can also be considered as a measure of the uncertainty in finding the mean.

Confidence Interval

- Is a measure of the certainty of finding the mean at different confidence levels.
- The most common CI is the 95%.
- For a large sample, 95% CI = mean ± 1.96 SEM.
- For small samples, 95% CI = mean ± t_{value} SEM.
- A 95% CI implies that if you did an experiment 100 times, in 95 experiments you will include the mean.
- CI takes into account sample size, standard error and probability.
- CI decrease with increase in sample size as we obtain a better estimate of the population mean.
- Both CI and SEM describe the error of estimating the population mean, however the CI shows the error based on different confidence levels.

The CI is expressed by two parts:

i. **A range or interval**: an upper estimate and lower estimate, e.g. the blood pressure measured from a sample of data could have a CI between 115 and 125 mmHg.
ii. **Level of the confidence interval:** as a percentage (e.g. 99%, 95%, 90%). A 95% CI states that we have a 95% chance of finding the true population mean within the interval.

If we want a higher CI to find the mean, then the probability or reliability of finding the mean decreases. For example, a 99% CI is wider than a 95% CI which is wider than a 90% CI. But the probability of finding the mean is more difficult the wider the confidence interval as you are increasing the margin of error. So, the best compromise and most widely used confidence interval is the 95% CI. The narrower the confidence interval, the more accurately we can predict the true mean value.

The CI can be calculated for means, medians, proportions, correlations, and regression coefficients. It is affected by the sample size, standard deviation (SD), and the level of confidence.There are slightly different formulas for the CI depending on the size of the sample and whether the population standard deviation is known.

Formula 1. If the sample size is less than 30 and the population SD is unknown then use:

$$\mathbf{CI = mean \pm t_{value} \times SEM}$$

Or,

$$CI = \text{mean} \pm t_{value} \times SD/\sqrt{N}$$

where t_{value} = value corresponding to $p = 0.05$ and degrees of freedom (N-1).
SD = sample standard deviation
Mean = sample mean
Note: $SD/\sqrt{N}$ is also the standard error (SEM);
t_{value} x $SD/\sqrt{N}$ is also called the margin of error (ME)

Sample size affects SEM and the width of the confidence interval. The smaller the sample the less accurate is the standard error and the wider the confidence interval width, i.e. there is less likelihood of finding the true mean. So, a sample size of 100 will have a wider CI compared to a sample size of 1000. If zero is within the CI, this suggests a non-significant result; if zero is outside the CI, this suggests significance.

Formula 2. If the sample size is greater than 30, and we know the population SD, then use:

$$CI = mean \pm Z\ SEM$$

Where mean = sample mean,
SEM = standard error (or $SD/\sqrt{N}$);
Z = confidence level value

t_{value}

- Are standardized values calculated from sample data which are used in hypothesis tests.
- A t_{value} is calculated by dividing the difference between the means of two samples by the standard error.
- Are used in hypothesis tests to show whether differences in means between two samples occurred by chance.
- The t_{values} form a probability distribution very similar to the normal distribution.
- Every t_{value} is associated with a P value and degree of freedom.
- P value is a number (between 0 and 1 or 0 and 100%) which states the probability of an event happening. A P of 5% (=0.05) means there is a 1 in 20 chance of an event happening.
- Degree of freedom is related to sample size, it is your sample size minus 1. It is measure of the number of values in a data sample that have the freedom to vary.

$$CI = mean \pm t_{value} \times SD/\sqrt{N}$$

Z value

- Z values (or Z scores) are standardized values obtained from sample data values which come from a population with a normal distribution, with known population mean and standard deviation.
- Z = (sample data value − population mean)/population standard deviation.
- Z values describe the variation of sample values in SD from the mean.
- Z values form a standard normal distribution (frequency of Z against SD) with mean of zero.
- Z values are negative below the mean and positive above the mean.
- The importance of Z values are (i) they allow us to find the probability of a sample data occurring in a normal distribution, and (ii) compare two sample data values from two different normal distributions.
- Advantage of Z scores is that they show how an individual compares with other people. For example your marks for a biochemistry and microbiology module test could be compared more accurately with other people if the marks were converted to Z scores, i.e. we are all using the same scale.

If we know the mean and standard deviation of a group of sample data, we can convert every data point in the sample to a standard score or Z score. Z is known as a standard score or Z score, has a distribution, and is a measure of the standard deviation above and below the population mean. It is used to compare sample results to the normal population. Z can be defined by the following formula.

Z = (observed value – sample mean)/standard deviation of sample.
Or,
$$\mathbf{Z = (X - \bar{x})/SD}$$

Note if the population mean or standard deviation is known, then these would be used in the formula. We could replace Z score with a t score from the t distribution, to calculate the confidence interval. The sample size affects t, but when the sample size is > 30 then t is quite close to the value of Z, we can then use, $Z = 1.96$, as the critical value for a 95% confidence interval.

EXAMPLE 1.12 | Calculation of the confidence interval from sample data

Consider the breath hold times in seconds from example 1.1.
55, 60, 50, 67, 71, 78, 75, 80, 84, 52, 65

Step 1: Calculate the standard deviation: SD = 11.704 s
Step 2: Calculate the mean: 67 s
Step 3: Calculate the standard error of the mean (SEM): here SEM = 3.529 s
Step 4: Note the sample size and degree of freedom, i.e,. $N = 11$; degree of freedom $(N\text{-}1) = 10$
Step 5: Using the tables for t values (see Appendix 8) and the degree of freedom of 10, find the t value, i.e. 2.228
Step 6: Calculate the margin of error (ME)
Where, ME = (t value for n–1 degrees of freedom x SD)/square root of N
Or ME = t value for n–1 degrees of freedom x SEM
Or ME = 2.228 x 3.529 = 7.86

Step 7: Calculate the confidence interval
CI = mean ± ME
CI = 67 ± ((2.228 x 11.704)/ $\sqrt{N}$) = 67 ±7.86 = 59.14 to 74.86

EXAMPLE 1.13 | Calculation of the confidence interval with known mean and SD

Forty students had their systolic blood pressure measured and the mean was 123 mmHg with a SD of 8 mmHg.
a) Find the 95% confidence interval

EXAMPLE 1.13 *(continued)*

b) Find the 99% confidence interval

Step 1: The formula for the 95% confidence interval is shown below, state the values for the parts of the formula

Or confidence interval is between mean + t (SD/$\sqrt{N}$) and mean – t (SD/$\sqrt{N}$)
SD = 8, mean = 123, N = 40, df = 40-1 = 39, t = 2.02 (from table of critical values df = 39, and P = 0.05), SEM = (SD/$\sqrt{N}$) = 8/$\sqrt{40}$ = 1.24

Step 2: Put values into formula
123 + 2.02 (8/$\sqrt{40}$) and 123–2.02 (8/$\sqrt{40}$)

Or 95% confidence interval is 120.4 to 125.6 mmHg

Step 3: For the 99% confidence interval, substitute a new t value of 2.70 (from df of 39 and P = 0.01 and critical table)
Thus, confidence interval is

123 ± 2.70(8/$\sqrt{40}$)

Or 99% confidence interval is 119.6 to 126.4 mmHg.

CI = mean ± 1.96 SEM This gives the values between which 95% confidence of finding the mean, i.e. we can be 95% certain that the true mean value lies within ± 2 standard deviations.

Interpretation: in theory, our population mean would be between 120.4 mmHg and 125.6 mmHg. However, this is not correct, since one sample is not representative of the population. You would have to take many samples and work out CI for each sample, and then from these estimated CI you could more accurately predict the population mean. We can see that the 99% CI is wider than the 95% CI, which means that the uncertainty in finding the population mean is greater in the 99% than in the 95% CI.

Degrees of Freedom

The degrees of freedom are a measure of the ability (or 'freedom') of values or variables to vary.

Suppose you have a bag containing five different coloured balls, where the variable is the colour of the ball. If you take out one of the balls from the bag, then you have decreased your degree of freedom by one to choose from the five colours. If you take out another ball you have decreased your degree of freedom even further, and so on.

Suppose we have N observations in a sample, then each is free to have any value, but if you calculate the sample mean, then you restrict the value of the test measurement, i.e. your degree of freedom is restricted. Hence the degree of freedom mathematically = N–1.

> **"... the actual and physical conduct of an experiment must govern the statistical procedure of its interpretation."**
> *Ronald Aylmer Fisher (1890–1962)- British mathematician, Biologist*

Summary

- Data can be obtained from population samples or total populations.
- Data types: quantitative (continuous, discrete); qualitative (ordinal, nominal).
- Data samples are described by descriptive statistics as measures of central location or central tendency (mean, mode, median) and the spread (variation) of the data (standard deviation, variance, quartiles, interquartile range, percentiles).
- Sample data is used to make predictions about the population data.
- Sample data are used to produce probability (frequency) distributions (e.g. normal, t, Z).
- Population data have a population mean and population standard deviation.
- Sample data have a sample mean and sample standard deviation.
- Standard error describes the error of the sample mean from the actual mean of the population and standard deviation is a measure of the variability of sample data around the mean.
- The accuracy of the mean is described by the confidence interval (usually the 95%) and standard error.
- The mean is a very important statistic as it describes the central tendency of data and is used in comparing differences between samples in hypothesis testing.
- The variation of the data is most commonly described by standard deviation and variance.
- The reference range is an important guide in defining normal and abnormal values, e.g. in diagnosis and treatment of disease, accuracy of instrumentation.
- Experiments need a clear aim, experimental design, a control, and adequate sample size.
- Data can be collected from different types of studies such as epidemiological, clinical trials, and experiments.

Sample Problems

Answers to all sample problems are found in Appendix 7.

Type of data

1) Describe the following data using the terms, quantitative, qualitative, discrete, continuous, ordinal, and nominal:
(i) Ethnicity of volunteers, (ii) number of red blood cells counted on a haemocytometer, (iii) Borg's perceived rate of exertion, (iv) metabolic rate measured from a wet spirometer, (v) size of Bowman's capsule in a kidney

Descriptive Statistics

2) The following weights in grams were obtained from male Wistar rats. Calculate the descriptive statistics (mean, median, mode, range, SD, SEM) for the following data:
200, 250, 400, 350, 520, 324, 568, 425, 300, 236, 185, 520, 250, 300, 350, 400, 300

Confidence Interval and Coefficient of Variation

3) Calculate the 95% confidence interval (CI) and the coefficient of variation for the measurements of a volume of 4000 μL^{-1} of red blood cell diluting fluid measured with a Gilson pipette (P5000) by 12 students.
4100, 3980, 3990, 4200, 3700, 4300, 4500, 4000, 3995, 4005