Learner Guide

Apply knowledge of statistics and probability to critically interrogate and effectively communicate findings on life related problems

SAQA ID 9015
Apply knowledge of statistics and probability to critically interrogate and effectively communicate findings on life related problems
NQF Level 4, 6 Credits

Table of Contents

Module 1
Critique and use techniques for collecting, organising and representing data

After completing this module, the learner will be able to critique and use techniques for collecting, organising and representing data, by successfully completing the following:

- Situations or issues that can be dealt with through statistical methods are identified correctly
- Appropriate methods for collecting, recording and organising data are used so as to maximise efficiency and ensure the resolution of a problem or issue
- Data sources and databases are selected in a manner that ensures the representativeness of the sample and the validity of resolutions
- Activities that could result in contamination of data are identified and explanations are provided of the effects of contaminated data
- Data is gathered using methods appropriate to the data type and purpose for gathering the data
- Data collection methods are used correctly
- Calculations and the use of statistics are correct
- Graphical representations and numerical summaries are consistent with the data, are clear and appropriate to the situation and target audience
- Resolutions for the situation or issue are supported by the data and are validated in terms of the context

Techniques for collecting, organising and representing data

The Concise Oxford Dictionary defines statistics as *"*Numerical facts systematically collected" and statistic as *"*Statistical fact or item*"*.

Statistics entails all aspects of information: collecting, organising, comprehending, communicating, and interpreting.

The word "Statistics" originally referred to collections of facts (not necessarily numerical) about the State. According to modern usage, however, it refers to collections of numerical facts or estimates and is not restricted to the State. The word "statistics", when used in the plural, refers to the figures themselves, suitably classified and tabulated together with any secondary statistics derived from them, such as percentages or averages. "Statistics", when used in the singular, refers to the study which deals with the collection, analysis and interpretation of figures. This is also called Statistical Method.

Statistical Method can be broadly divided into two categories, namely descriptive and mathematical statistics.

Descriptive Statistics compiles and presents data exactly as recorded. Descriptive statistics is the most basic form of statistics and lays the foundation for all statistical knowledge.

Mathematical Statistics- Based on the theory of probability, it attempts to draw precise general conclusions from the data.

1.1 Situations or issues that can be dealt with through statistical methods

Statistics are all around us. In fact it would be difficult to go through a full week without using statistics. Imagine watching a football game where no one kept score. The action itself might provide enough excitement to hold your attention for a while, but think of all the drama that would be lost if winning and losing weren't at issue.

Without statistics we couldn't plan our budgets, pay our taxes, enjoy games to their fullest, or evaluate performance...

Statistics are used when you want to know something about a collection or set of objects. This could be a group of people, the days in a month, or the set of times a cricketer was at bat. For example, you might want to know the heights of a group of people. If the group consists of the learners in class, we can simply ask them their heights. However, if we are interested in the heights of every person in South Africa, we cannot personally ask each one their height. In fact, even the South African **census**, which is mandated to count everyone in SA, cannot find everyone in the country.

For this reason, two words are important for the science of statistics: **population** and **sample**. The population is the group you are interested in (e.g., all people in SA). The sample is the group you can collect information about (e.g. the learners in class). A sample is always a subset of the population, i.e. every member of the sample is in the population.

Definitions:

Census: a comprehensive inquiry of all data obtainable.

Sample: a set of objects taken at random from a larger group

The Law of Statistical Regularity states that a set of objects (a sample) taken at random from a larger group (population) tends to reproduce the characteristics of that larger group.

Statistics can be used to:

- Determine trends in societal issues such as crime and health;
- Identify relevant characteristics of target groups such as age range, gender, socio-economic group, cultural belief, and performance;
- Consider the attitudes or opinions of people on issues.

Sometimes research involves answering questions that have to do with people's attitudes, beliefs and opinions – this is **qualitative research**. At other times the questions to be answered centre around the state of a certain situation; i.e. how much of a thing exists. That is called **quantitative research**.

Class Activity 1: Situations or issues that can be dealt with through statistical methods

In small groups, complete the formative activity in your Learner Workbook

1.2 Methods for collecting, organising and analysing data (EEK1; AC 2, 5 and 6)

Definition of data

Data (singular is *datum*): things known or assumed as the basis for inference (drawing conclusions) or calculation

Methods for collecting data

There are two types of data that may be collected: Primary data and Secondary data.

- **Primary data** is data collected first hand by the researcher specifically for the project being done.

  Primary data is relevant, specific and accurate but may be very costly in both money and time.

- **Secondary data** is data received from a previous study that may be useful for the purposes of the current study.

  Secondary data, on the other hand, may be cheaper and quickly accessible but may be out of date, vague or irrelevant for the purposes of the current study.

The method of data collection you select must suit the research question you wish to answer.

Questionnaires

First, the mode of data collection must be decided upon (*e.g., mail, telephone, or in person*). Once this has been determined a questionnaire can be developed.

The three most important things for any questionnaire designer to remember are simplicity, simplicity, and simplicity. Ideas need to be conveyed clearly and questions should be easy to comprehend. There must be no guesswork for the respondent when it comes to understanding exactly what information is being requested.

It is recommended that questionnaires be written at the 9th-grade (standard 7) reading level.

Questions must be clearly defined and unambiguously phrased; otherwise, the resulting data are apt to be misleading.

Attention must also be given to the length of questionnaires. Long questionnaires tend to lead to respondent fatigue and errors arising from inattention, refusals, and incomplete answers.

There are other factors to take into account when planning a questionnaire. These include such diverse considerations as: the order in which the questions are asked. their appearance, even such things as the questionnaire's physical size and format.

Respondents are more likely to cooperate if the questions are simple, clear, easy to answer, and personally relevant to them.

Types of questionnaires:

- **Dichotomous key**

The **dichotomous key** is constructed from a series of highly organised statements arranged into **couplets**. A couplet consists of two descriptions which should represent mutually exclusive choices so that the respondent can only answer “yes” or “no”. Once a decision is made, that selection directs you to another couplet (either the next in order or one further on in the key), and this process is repeated until a conclusion is reached.

- **Likert scale**

A Likert scale measures the extent to which a person agrees or disagrees with the question. The most common scale is 1 to 5. Often the scale will be 1=strongly disagree, 2=disagree, 3=not sure, 4=agree, and 5=strongly agree.

Interviews

In the interview, the researcher talks to the respondent and obtains information directly. This can be advantageous because it is flexible and in-depth. When a respondent speaks, the interviewer is able to redirect the questioning to deal with the unexpected. The researcher can ascertain why the respondent answers a certain way. People will more readily answer questions in an interview than they will in a questionnaire.

Interviewing is sometimes difficult because it costs a great deal in time and money. It can reach out to far fewer respondents than the questionnaire. It takes a great deal of

experience and expertise on the part of the interviewer to prepare an interview schedule (the list of questions to be asked) and to ask questions in a way that allows valid conclusions to be drawn from the responses.

Surveys

Conducting a credible survey entails scores of activities, each of which must be carefully planned and controlled. Taking shortcuts can invalidate the results and badly mislead the users of the information gathered.

Basic steps of the survey process:

1. **Organisation**-The survey taker determines who is to be sampled and what is to be learned about the sample.
2. **Questionnaire Design**-Based on the goal of the survey, questions for survey respondents are prepared and arranged in a logical order to create the survey questionnaire.
3. **Sampling**-A repeatable plan is developed to randomly choose a sample capable of meeting the survey's goals. Then a sample is selected.
4. **Data Collection**-A plan for contacting the sample and collecting information from participants is developed and carried out.
5. **Data Processing**-Collected data are entered into the computer and checked for accuracy.
6. **Analysis-**The results of the survey are compiled and disseminated.

A final problem that can be encountered during surveys is “Interviewer bias” , which can easily arise in highly-charged emotive or political inquiries, for example, when the interviewer misunderstands a reply, marks a wrong code on the answer sheet, or even interprets an answer incorrectly as a result of his/her own view on the topic.

Replies can also be biased through forgetfulness on the part of the people interviewed, by the desire to make a good impression on the interviewer, or by the fear that a truthful answer may result in something to their disadvantage. In such cases, it may be better to leave a form or questionnaire to be filled in and collected later.

Checklists

If a researcher has made provision for all possible alternative answers to each question, and if the respondent need merely ticks the answer that applies, the questionnaire is called a checklist. A checklist is structured and the questions are closed.

Advantages of checklists are that:

- they are easy and convenient to answer;
- the responses are easy to measure;
- the data can be processed in a uniform way.

The respondents can only choose which answer best describes them and because the respondent is offered little choice the researcher only needs to count up how many respondents agreed with a specific response.

Disadvantages of checklists are that:

- The respondent may be irritated by not being able to find his or her preferred answer among the given alternatives.
- The checklist may produce results that lack accuracy because the questions are answered at a very superficial level. The respondents may want to qualify or explain their responses but not have the opportunity to do so.

This problem may especially be faced by a respondent presented with a situation (or, incident) who is then asked to check which attitude he most identifies with, with regard to the incident. The checklist is not always the most practical method of gauging the opinions and beliefs of the respondent.

Observation[1]

The scientific method

In the classic scientific method of research there are five steps that together make up the proper technique:

Step one: Observe a natural phenomenon.
Step two: Draw a conclusion based on what you see.
Step three: Predict what might happen next or formulate a hypothesis for what should happen.
Step four: Design a process or experiment which can test the hypotheses.
Step five: Refine, extend and restructure the theory in the light of more evidence.

Observation refers to the recording of data using scientific instruments. Observation is the most obvious form of primary data collection. The term may also refer to any data collected during this activity.

The need for reproducibility requires that observations by different observers be comparable. Human sense impressions are subjective and qualitative making them difficult to record or compare. The idea of measurement evolved to allow recording and comparison of observations made at different times and places by different people. Measurement consists of using observation to compare the thing being measured to a standard: an artifact, process or definition which can be duplicated or shared by all observers, and counting how many of the standard units are comparable to the object. Measurement reduces an observation to a number which can be recorded, and two observations which result in the same number are equal within the resolution of the process.

Senses are limited, and are subject to errors in perception such as optical illusions. Scientific instruments were developed to magnify human powers of observation, such

[1] Retrieved from: http://en.wikipedia.org/wiki/Observation

as weighing scales, clocks, telescopes, microscopes, thermometers, cameras, and tape recorders, and also translate into perceptible form events that are unobservable by human senses, such as indicator dyes, voltmeters, spectrometers, infrared cameras, oscilloscopes, interferometers, geiger counters, x-ray machines, and radio receivers.

One problem encountered throughout scientific fields is that the observation may affect the process being observed, resulting in a different outcome than if the process was unobserved. This is called the *observer effect*. For example, it is not normally possible to check the air pressure in an automobile tyre without letting out some of the air, thereby changing the pressure. However, in most fields of science it is possible to reduce the effects of observation to insignificance by using better instruments.

Randomness, probability and association (EEK 4)

Random samples

A **sample** is a subject chosen from a population for investigation.

In statistical terms a **random sample** is a set of items that have been drawn from a population in such a way that each time an item was selected, every item in the population had an equal opportunity to appear in the sample. If, for example, numbered pieces of cardboard are drawn from a hat, it is important that they be thoroughly mixed, that they be identical in every respect except for the number printed on them and that the person selecting them be well blindfolded.

Second, in order to meet the equal opportunity requirement, it is important that the sampling be done with replacement. That is, each time an item is selected, the relevant measure is taken and recorded. Then the item must be replaced in the population and be thoroughly mixed with the other items before the next item is drawn. If the items are not replaced in the population, each time an item is withdrawn, the probability of being selected, for each of the remaining items, will have been increased.

For example, in a population of 9, the initial probability that a given item will be selected is 1/9. If, however, an item is drawn and not returned before drawing a second item, the probability that a given item will be drawn will have been increased to 1/8. Of course, this kind of change in probability becomes trivial if our population is very large, but it is important to recognise the principle illustrated here, to fully understand the concept of a random sample.

It is also important to recognise that when sampling with replacement, it is possible for the same item to appear more than once in a sample and it is possible to draw a random sample that is larger than the population from which it came. Notice also, that it is possible to draw as many random samples as we like from a give population. The key idea here is that we either sample with replacement or we draw our samples from a population that is so large that the withdrawal of successive items changes probability by an amount that is too small to be of concern.

- A hand of cards dealt from a properly shuffled pack of cards is a random sample
- Winning tickets in a lottery are chosen at random by taking numbers out of a revolving drum

- If we want to choose a random sample of a certain population or group, we can take every tenth one, for example every 10th invoice in the sample, or every employee whose ID number ends in a particular digit, for example everyone whose number ends with a 7.

Quota sampling

Quota sampling refers to choosing the sample that proportionally reproduces the representation of a certain group in the population chosen.

For example, in a factory employing 1300 women and 700 men, a random sample of workers might contain 73 women and 27 men. A quota sample would have to contain 65 women and 35 men, i.e. according to their proportional representation in the population.

Quota sampling can only be used if the composition of the population is known.

It is regarded as more accurate than random sampling and is used extensively in market research, where an interviewer selects sample people or households in a particular sub-group.

However, the researcher may, quite subconsciously, tend to avoid slum-type areas and aggressive-looking individuals, so bias can creep into the statistical survey.

Probability

Each year, millions of people travel to casinos hoping they will come away richer. Many more people visit their local supermarket each day to bet with lottery cards.

Why do we invest this money on chance? We do it because we believe we can beat the odds. We believe in the possibility of winning.

Mathematical principles can tell us more than whether it is possible to win. They can tell us how often we are likely to win. The mathematical concept that deals with the chances of winning a lottery draw or a poker game is **probability**.

If we can determine the probability that a certain event (such as winning the lottery) will occur, we can make a better choice about whether to risk the odds.

How do we determine probability?

Problem:

A spinner has 4 equal sectors coloured yellow, blue, green and red.

What are the chances of landing on blue after spinning the spinner?

What are the chances of landing on red?

Solution:

The chances of landing on blue are 1 in 4, or one fourth.

The chances of landing on red are 1 in 4, or one fourth.

This problem asked us to find some probabilities involving a spinner.

Let's look at some definitions and examples from the problem above:

Definition	Example
An **experiment** is a situation involving chance or probability that leads to results called outcomes.	In the problem above, the experiment is spinning the spinner.
An **outcome** is the result of a single trial of an experiment.	The possible outcomes are landing on yellow, blue, green or red.
An **event** is one or more outcomes of an experiment.	One event of this experiment is landing on blue.
Probability is the measure of how likely an event is.	The probability of landing on blue is one fourth.

We work out probability by dividing the number of successful outcomes by the total number of possible outcomes:

Probability Of An Event
$P(A) = \frac{\text{The Number Of Ways Event A Can Occur}}{\text{The total number Of Possible Outcomes}}$

Example 1

Every Saturday night the lotto draw takes place and the winner of the game show gets to draw a ball from a variety of balls in a round canister in order to win a car.

We want to work out what the probability is of the winner drawing the red ball, which will make him/her the winner of the car.

First we have to find out how many balls are in the canister:

- 5 green balls
- 6 yellow balls
- 1 red ball (the winning ball)

There are 5 + 6 + 1 = 12 balls in the canister, thus a total of 12 possible outcomes

Probability (P) = $\frac{\textit{number of successful outcomes}}{\textit{total number of possible outcomes}}$

Probability (Green ball) = 5/12
Probability (Yellow ball) = 6/12
Probability (Red ball) = 1/12

So the chance of the winner drawing a red ball is 1 out of 12.

Example 2

Choose a number at random from 1 to 5.

- What is the probability of each outcome?
- What is the probability that the number chosen is even?
- What is the probability that the number chosen is odd?

Outcomes:
The possible outcomes of this experiment are 1, 2, 3, 4 and 5.

Probabilities:

$$P(1) = \frac{\text{\# of ways to choose a 1}}{\text{total \# of numbers}} = \frac{1}{5}$$

$$P(2) = \frac{\text{\# of ways to choose a 2}}{\text{total \# of numbers}} = \frac{1}{5}$$

$$P(3) = \frac{\text{\# of ways to choose a 3}}{\text{total \# of numbers}} = \frac{1}{5}$$

$$P(4) = \frac{\text{\# of ways to choose a 4}}{\text{total \# of numbers}} = \frac{1}{5}$$

$$P(5) = \frac{\text{\# of ways to choose a 5}}{\text{total \# of numbers}} = \frac{1}{5}$$

$$P(\text{even}) = \frac{\text{\# of ways to choose an even number}}{\text{total \# of numbers}} = \frac{2}{5}$$

$$P(\text{odd}) = \frac{\text{\# of ways to choose an odd number}}{\text{total \# of numbers}} = \frac{3}{5}$$

The outcomes 1, 2, 3, 4 and 5 are equally likely to occur as a result of this experiment. However, the events even and odd are not equally likely to occur, since there are 3 odd numbers and only 2 even numbers from 1 to 5.

Summary:

The probability of an event is the measure of the chance that the event will occur as a result of an experiment.

The probability of an event A is the number of ways event A can occur divided by the total number of possible outcomes.

The probability of an event A, symbolised by P(A), is a number between 0 and 1, inclusive, that measures the likelihood of an event in the following way:

- If P (A) > P (B) then event A is more likely to occur than event B.
- If P (A) = P (B) then events A and B are equally likely to occur.

Association

In statistics, an **association** comes from two variables that are related. Many people confuse association with causation. Association **does not** imply causation.

For example, the United Nations did a study of government failure — when governments fall or are overthrown. The best indicator of a government about to fall was the infant mortality rate. The dead children do not cause the government to fall, rather they are joint effects of a common cause.

Another example is ice cream consumption and murder. The sales of ice cream and murder are strongly positively correlated. Which causes which: does eating ice cream cause murder or does murder make people eat ice cream? The answer is neither — increases in both ice cream consumption and murder correlate with hot weather.

Another perspective on the relationship between association and causality is that association does not imply a direct causal connection between the associated variables. If, however, association is non random (i.e., not due purely to chance), then it implies that some causal mechanism is operative.

Often, the nature of the causal mechanism underlying an association is the joint influence of one or more common causes operating on the variables in question. For example, both the increase in ice cream consumption and murder may occur during warm weather (a conclusion that would require further information to confirm or refute).

Methods for organising data

Data that has merely been collected is not very useful. It has to be processed, so that we may draw useful conclusions. This entails organising the raw data so that it makes sense.

A very simple way of organising data is to arrange it in ascending order. That is, write all the numbers down arranged in order from smallest to largest. An important part of organising data is to calculate aspects such as the averages (mean, median and mode) of sets of data. These averages are called “measures of central tendency” and enable us to see how close data points are to one another.

We can also look at “measures of dispersion”, which tell us how far apart the data points are (range, standard deviation and variance).

We will now look at these methods of organising data in more detail.

i. **Frequency distribution**

Before a large number of observations can be analysed, they must be sorted into a convenient number of groups, or **classes**.

The data must be sorted according to the numerical value of some characteristic, called a **variable.** Variables are things that we measure, control, or manipulate in

research. They differ in many respects, most notably in the role they are given in our research and in the type of measures that can be applied to them. Thus a number of people might be sorted according to their height, age, weight, or any other characteristic capable of being measured.

Variables differ in "how well" they can be measured, i.e. in how much measurable information their measurement scale can provide. There is obviously some measurement error involved in every measurement, which determines the "amount of information" that we can obtain. Another factor that determines the amount of information that can be provided by a variable is its "type of measurement scale." Specifically variables are classified as: (a) nominal, (b) ordinal, (c) interval or (d) ratio.

a. **Nominal variables** allow for only qualitative classification. That is, they can be measured only in terms of whether the individual items belong to some distinctively different categories, but we cannot quantify or even rank order those categories. For example, all we can say is that two individuals are different in terms of variable A (e.g. they are of different races), but we cannot say which one "has more" of the quality represented by the variable. Typical examples of nominal variables are gender and race.

b. **Ordinal variables** allow us to rank order the items we measure in terms of which has less and which has more of the quality represented by the variable, but still they do not allow us to say "how much more." A typical example of an ordinal variable is the socio-economic status of families. For example, we know that upper-middle is higher than middle, but we cannot say that it is, for example, 18% higher.

c. **Interval variables** allow us not only to rank order the items that are measured, but also to quantify and compare the sizes of differences between them. For example, temperature, as measured in degrees Celsius, constitutes an interval scale. We can say that a temperature of 40 degrees is higher than a temperature of 30 degrees, and that an increase from 20 to 40 degrees is twice as much as an increase from 30 to 40 degrees.

d. **Ratio variables** are very similar to interval variables; in addition to all the properties of interval variables, they feature an identifiable absolute zero point, thus they allow for statements such as x is three times more than y. Typical examples of ratio scales are measures of time or space. For example, as the Kelvin temperature scale is a ratio scale, not only can we say that a temperature of 200 degrees is higher than one of 100 degrees, but we can also correctly state that it is twice as high. Interval scales do not have the ratio property. Most statistical data analysis procedures do not distinguish between the interval and ratio properties of the measurement scales.

ii. **Continuous and discrete variables**

A variable that can take only discrete (whole) values, is called **discrete** or discontinuous- a man cannot have 6.3 children, there cannot be 5.2 houses in the street, or 3.5 rooms in a house.

A **continuous variable**, on the other hand, can take any value within a range- temperature need not be an exact number of degrees, but may be measured to several decimal places, e.g. 37.5℃ and every object whose temperature is rising or falling will, during the process, take every possible temperature between the final and initial values. Similar examples are the speed of a vehicle and the height of a growing plant. In other words a continuous variable can take an infinite number of values.

Measures of centre and spread (EEK 2, AC7)

The interpretation of data is very simple if you are able to work through it systematically. The most important features of data are:

- Frequency
- Average
- Median
- Mode
- Range

Frequency

Frequency is the number of times a certain value appears in a series of data.

Example:

In the series of data below, the number 5 appears 6 times, therefore the frequency of 5 is 6. It is the value that appears most often in the series:

3; 5; 3; 7; 5; 6; 5; 9; 5; 2; 4; 4; 5; 5; 8

If we put this series of data in a table, then the frequency would be much clearer:

Number	Frequency
0	0
1	0
2	1
3	2
4	2
5	6
6	1
7	1
8	1
9	1
10	0

When you are using the tally system to determine the frequency, you will draw a line for every time something occurs, i.e. I.

When it occurs four times, you draw four lines, i.e. I I I I, but when you reach the fifth occurrence, you do not draw the fifth line next to the other four, but you draw a line through the other four lines to show that you have reached 5, i.e. ~~IIII~~. It makes it much easier to count when you reach the end.

Example:

Defects	MONTH			
	1	2	3	Total
Type 1	~~////~~	//	//	9
Type 2	/		/	2
Type 3	///	//	/	6
Type 4	//	//	/	5
Type 5	~~////~~ /	////	~~////~~	15
Total	17	10	10	37

Average

Mean[2]

The mean or average is one of the more common statistics you will see, and it's easy to calculate: all you have to do is **add** up all the values in a set of data and then **divide** that sum by the number of values in the dataset.

That is[3]:

$$\text{Mean} = \frac{\text{Sum of all data values}}{\text{Number of data values}}$$

Symbolically,

$$\bar{x} = \frac{\sum x}{n}$$

where $\bar{x}$ (read as 'x bar') is the mean of the set of x values,
$\sum x$ is the sum of all the x values, and
n is the number of x values.

Example

The marks of seven students in a mathematics test with a maximum possible mark of 20 are given below:

15 13 18 16 14 17 12

Find the mean of this set of data values.

Solution:

$$\text{Mean} = \frac{\text{Sum of all data values}}{\text{Number of data values}}$$
$$= \frac{15+13+18+16+14+17+12}{7}$$
$$= \frac{105}{7}$$
$$= 15$$

So, the mean mark is 15.

Symbolically, we can set out the solution as follows:

[2] Adapted from www.robertniles.com
[3] Retrieved from: http://www.mathsteacher.com.au/year8/ch17_stat/02_mean/mean.htm

$$\bar{x} = \frac{\sum x}{n}$$
$$= \frac{15+13+18+16+14+17+12}{7}$$
$$= \frac{105}{7}$$
$$= 15$$

So, the mean mark is 15.

Median

Whenever you find yourself referring to "the average worker" this, or "the average household" that, you don't want to use the mean to describe those situations. You want a statistic that tells you something about the worker or the household in the middle. That's the **median**.

Again, this statistic is easy to determine because the median literally **is** the value in the middle: just line up the values in your set of data, from largest to smallest, or in descending order- the one in the dead-centre is the median.

Comparing the mean to the median for a set of data can give you an idea how widely the values in your dataset are spread apart.

Example:

The marks of nine students in a geography test that had a maximum possible mark of 50 are given below:

47 35 37 32 38 39 36 34 35

Find the median of this set of data values.

Solution:
Arrange the data values in order from the lowest value to the highest value:
32 34 35 35 36 37 38 39 47

The fifth data value, 36, is the middle value in this arrangement.

$\therefore$ Mediar $= 36$

Note:

The number of values, n, in the data set $= 9$

$$\text{Median} = \frac{1}{2}(9+1)\text{ th value}$$
$$= 5\text{th value}$$
$$= 36$$

In general:

$$\text{Median} = \frac{1}{2}(n+1)\text{ th value, where } n \text{ is the number of data values in the sample}$$

If the number of values in the data set is even, then the **median** is the average of the two middle values.

Example:
Find the median of the following data set:

12 18 16 21 10 13 17 19

Solution:

Arrange the data values in order from the lowest value to the highest value:

10 12 13 16 17 18 19 21

The number of values in the data set is 8, which is even. So, the median is the average of the two middle values.

$$\therefore \text{Median} = \frac{4\text{th data value} + 5\text{th data value}}{2}$$
$$= \frac{16+17}{2}$$
$$= \frac{33}{2}$$
$$= 16.5$$

Alternative method:

There are 8 values in the data set.

$$\therefore n = 8$$

$$\text{Now, median} = \left(\frac{n+1}{2}\right)\text{th value}$$
$$= \left(\frac{8+1}{2}\right)$$
$$= \frac{9}{2}$$
$$= 4.5\text{th value}$$

The fourth and fifth scores, 16 and 17, are in the middle. That is, there is no single middle value.

$$\therefore \text{Median} - \frac{16+17}{2}$$
$$= \frac{33}{2}$$
$$= 16.5$$

Note:

Half of the values in the data set lie **below the median** and half lie **above the median.**

The median is the most commonly quoted figure used to measure property prices. The use of the median avoids the problem of the mean property price which is affected by a few expensive properties that are not representative of the general property market.

Mode

The mode is the **number** that occurs most frequently in the series of data. In the series of data below, the mode is **5**.

3; 5; 3; 7; 5; 6; 5; 9; 5; 2; 4; 4; 5; 5; 8

Applications:

In printing: it is important to print more of the most popular books; because printing different books in equal numbers would cause a shortage of some books and an oversupply of others.

In manufacturing: it is important to manufacture more of the most popular items; because manufacturing different items in equal numbers would cause a shortage of some items and an oversupply of others.

Note:

- It is possible for a set of data values to have more than one mode.
- If there are two data values that occur most frequently, we say that the set of data values is **bimodal**.
- If there is no data value or data values that occur most frequently, we say that the set of data values has no mode.

Range

The range is the difference between the highest number and the lowest number in a set of data. The **range** in the following set of data will be as follows:

R500 000 - **highest**

R250 000

R250 000

R105 000

R105 000

R105 000

R105 000

R60 000

R60 000**- lowest**

Range = Highest Number – Lowest Number

= R500 000- R60 000

= R440 000

Methods for analysing data

The mean, median and mode of a data set are collectively known as **measures of central tendency** as these three measures focus on where the data is centred or clustered. To analyse data using the mean, median and mode, we need to use the most appropriate measure of central tendency. The following points should be remembered:

- The mean is useful for predicting future results when there are no extreme values in the data set. However, the impact of extreme values on the mean may be important and should be considered; e.g. the impact of a stock market crash on average investment returns.
- The median may be more useful than the mean when there are extreme values in the data set as it is not affected by the extreme values.
- The mode is useful when the most common item, characteristic or value of a data set is required.

Once the researcher has collected all the relevant figures, s/he has to sort them into a reasonably compact statement. Both the final report, as well as the intermediate stages will consist largely of **statistical tables**, as they are a convenient way of summarising the data in an orderly manner and of presenting the results concisely and intelligibly.

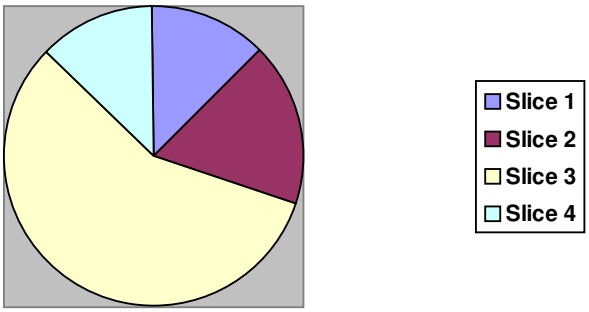
Worksheets are generally kept as a permanent record of the calculations performed on the data, both as a guide for future investigations and for reference in case figures are queried or further detail is required.

The great advantage of **charts** is that the important features stand out immediately, for example comparisons and trends, which would normally only be revealed by careful checking of figures. It is important to remember that charts must not depict too many items and the colours must also be clearly distinguishable, as confusion could arise. Most of the charts occurring in statistics are **graphs** or similar to graphs in that they represent **relations between two variables**, for example we have time-series charts, which show how one variable, such as output, sales, population or prices, varies with time.

Techniques for representing and evaluating statistics (EEK 3)

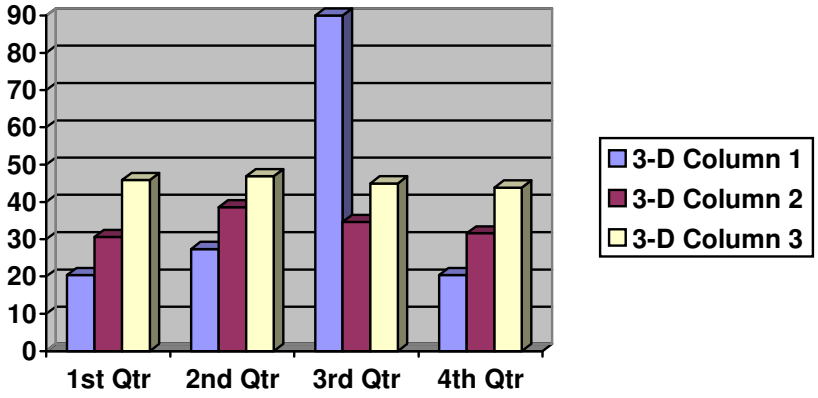
Pie chart

This type of chart depicts how an aggregate is divided into its principal components. The various items are proportional to the areas representing them in the circle, or to the angles of the various sectors. They can be converted into percentages by dividing the angles at the centre by 360 and multiplying by 100, i.e. if the angle is 45°, we can divide 45 by 360 X 100, which equals 12.5%. Alternatively we can determine the angle by multiplying 360° with the percentage and dividing by 100:



Bar chart

In the simplest form of bar chart, several items are shown graphically by horizontal or vertical bars of uniform width, with lengths proportional to the values they represent:



Histogram

In statistics, a **histogram** is a graphical display of tabulated frequencies. A histogram is the graphical version of a table which shows what proportion of cases fall into each of several or many specified categories.

We distinguish between true bar graphs and histograms. True bar graphs show category data (e.g. how many in each category: how many people have cats, how many people had a sandwich for lunch), and the bars in the bar graphs should be separate because the categories are separate.

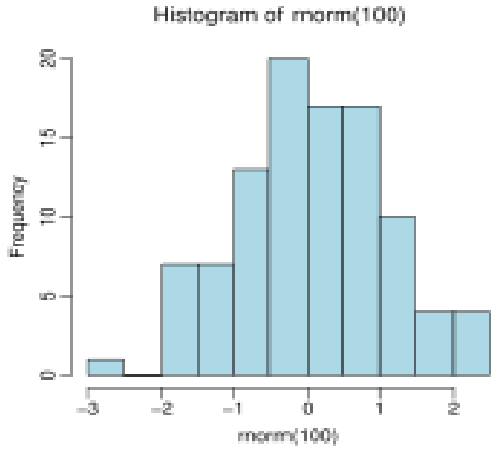
Histograms show numerical data, and generally data in a range (e.g. how many children are between 46 and 48 inches tall). In a histogram, the bars are supposed to touch to represent that the bar includes data from a range of values.

To represent data using a histogram:

- First, make sure that you are trying to represent the **frequency** of a certain occurrence.
- The independent variable will always be represented along the *x*-axis. In the case of a histogram this would always be a measure of **time** or **period**.
- Along the *y*-axis would be the **frequency**, that is, the number of occurrences of a particular event. This is the dependent variable.

The meaning of the terms “independent variable and “dependent variable”:

No matter what you are trying to investigate, you will find that one thing depends on another. For example, the state of your physical fitness depends on how you live. If you are too lazy to exercise, if you eat mostly junk foods, if you smoke, and if you drink alcohol to excess, you will most probably be extremely unfit. In other words, your lifestyle choices are the independent variable. Your state of physical fitness depends on your lifestyle choices, and is therefore the dependent variable.



Example of a histogram of 100 normally distributed random values.

The difference between a bar chart and a histogram is that **a histogram is used exclusively to represent frequency**.

Stem and leaf plots[4]

A stem-and-leaf plot is a display that organises data to show its shape and distribution .In a stem-and-leaf plot each data value is split into a "**stem**" and a "**leaf**".

The "**leaf**" is usually the last digit of the number and the other digits to the left of the "leaf" form the "**stem**". The number 123 would be split as:

stem 12

leaf 3

Constructing a stem-and-leaf plot:

The data:

Math test scores out of 50 points:

35, 36, 38, 40, 42, 42, 44, 45, 45, 47, 48, 49, 50, 50, 50.

[4] Retrieved from: http://regentsprep.org/Regents/Math/data/stemleaf.htm

<table>
<tr><td>Writing the data in numerical order may help to organise the data, but is NOT a required step. Ordering can be done later.</td><td>35, 36, 38, 40, 42, 42, 44, 45, 45, 47, 48, 49, 50, 50, 50</td></tr>
<tr><td>Separate each number into a stem and a leaf. Since these are two digit numbers, the tens digit is the stem and the units digit is the leaf.</td><td>For example:
The number 38 would be represented as:
<table><tr><th>Stem</th><th>Leaf</th></tr><tr><td>3</td><td>8</td></tr></table></td></tr>
<tr><td>Group the numbers with the same stems. List the stems in numerical order. (If your leaf values are not in increasing order, order them now.) Title the graph.</td><td><table><tr><th colspan="2">Math Test Scores
(out of 50 pts)</th></tr><tr><th>Stem</th><th>Leaf</th></tr><tr><td>3</td><td>5 6 8</td></tr><tr><td>4</td><td>0 2 2 4 5 5 7 8 9</td></tr><tr><td>5</td><td>0 0 0</td></tr></table></td></tr>
<tr><td>Prepare an appropriate legend (key) for the graph.</td><td>Legend: 3 | 6 means 36</td></tr>
</table>

A stem-and-leaf plot shows the shape and distribution of data. It can be clearly seen in the diagram above that the data clusters mostly around the row with a stem of 4.

- The leaf is the digit in the place farthest to the right in the number, and the stem is the digit, or digits, in the number that remain when the leaf is dropped.
- To show a one-digit number (such as 9) using a stem-and-leaf plot, use a stem of 0 and a leaf of 9.
- To find the median in a stem-and-leaf plot, count off half the total number of leaves.

Note:

If you are comparing two sets of data, you can use a back-to-back stem-and-leaf plot.

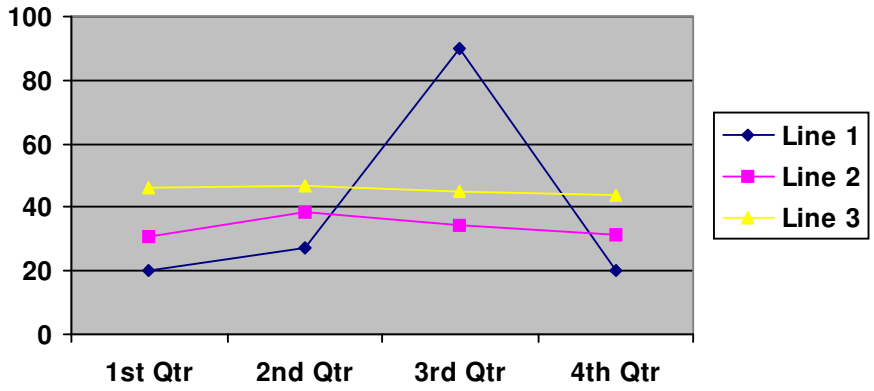
Data Set A		Data Set B
Leaf	**Stem**	**Leaf**
3 2 0	4	1 5 6 7

The numbers 40, 42, and 43 are from Data Set A. The numbers 41, 45, 46, and 47 are from Data Set B.

One advantage to the stem-and-leaf plot over the histogram is that the stem-and-leaf plot displays not only the frequency for each interval, but also displays all of the individual values within that interval.

Line chart

The most common type of chart in economic and commercial statistics is that of the time series, showing the progress of one or more quantities at successive times, usually at regular intervals.



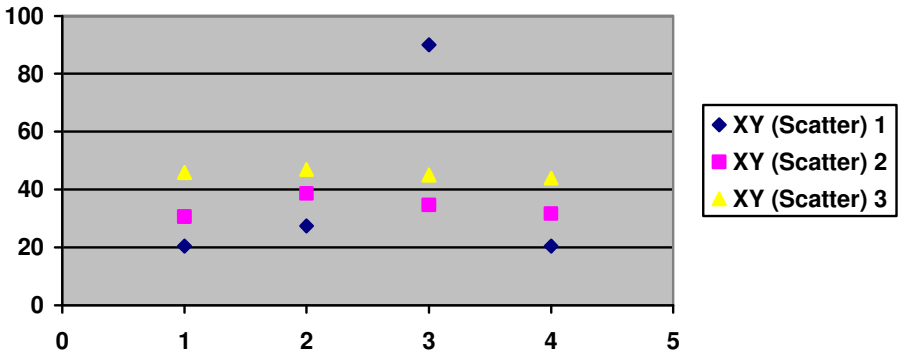
Scatter diagram/plot

The scatter plot can be used to find correlations (or connections) between certain phenomena. Two or more variables that tend to move in sympathy, are said to be **correlated**. The existence of correlation can be shown by means of a scatter diagram, in which each point corresponds to a pair of observations, one variable being plotted horizontally and the other vertically; for example, we can plot the correlation of the average duration of time off for injuries in one team, with the average duration of time off for the same types of injuries for another team.

When high values of one variable are associated with high values of another (for example the adult sons of men below average height will also tend to be short, although they will be nearer to the average than their fathers, therefore the sons of fathers 1,6m tall might, on average reach about 1,67m), they are said to be **directly or positively correlated**. When high values of one tend to accompany low values of the other, for example unemployment and labour turnover, they are **inversely or negatively correlated**.

To represent data on a scatter plot

- Arrange the data according to (*x*-values) independent variables and (*y*-values) dependent variables as you would for the bar chart.
- Plot the points directly onto the plot grid.



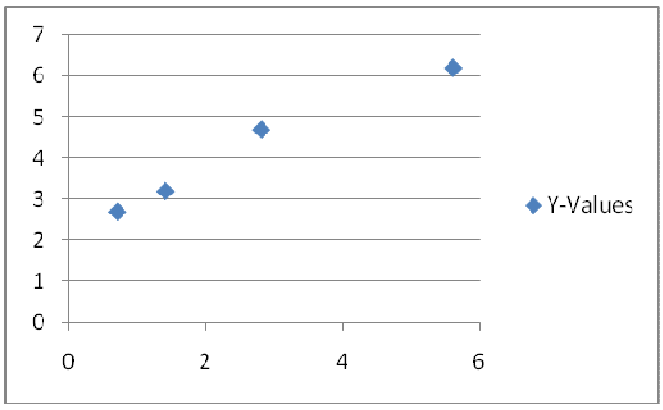
The correlation of two variables has to do with how closely related the two variables are. Let us call these variables *x* and *y*. The correlation coefficient is the measurement by which we determine the correlation. It ranges between –1 and 1. A coefficient of close to 1 or –1 means that the variables are very closely associated. The correlation coefficient has been defined by statisticians.

Therefore there is a very strong correlation or connection or relationship between the two variables. The weakest coefficient is equal to or near to 0, meaning there is little or no connection between the two variables. A positive correlation means that as one variable gets bigger, the other one also tends to get bigger.

It is beyond the scope of this unit standard to actually calculate correlation coefficients. But this information is provided to help you understand when the correlation coefficient is equal to 1 or -1 or 0.

- If the data points lie alongside one another in such a way that it is possible to draw a straight line through most of the points, and that straight line has a positive gradient, then the correlation coefficient is either 1 or close to 1.
- If the data points lie alongside one another in such a way that it is possible to draw a straight line through most of the points, and that straight line has a negative gradient, then the correlation coefficient is either -1or close to -1.
- If the data points lie on the scatter plot in such a way that you cannot draw a straight line through them, then there is little or no correlation (connection, or relationship) between the two variables. Then the correlation coefficient is 0 or close to 0.

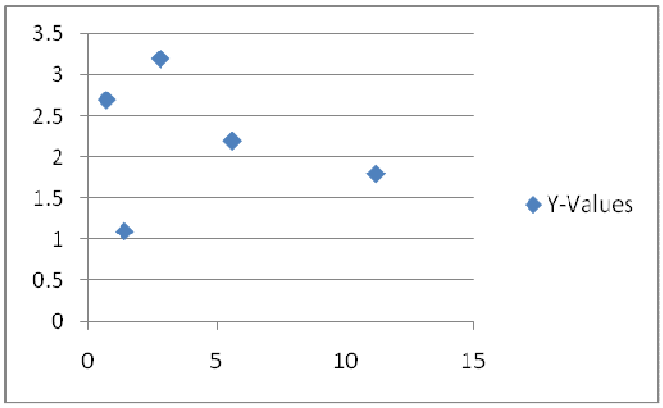
Example 1



Maximum positive correlation: (r = 1). This means you could draw a straight line through the data points marked on the scatter plot.

This means that as the value of the *x*-variable increases so does the value of the *y*-variable. Where the correlation is at its strongest: (r = 1,0) the data points form a straight line.

Example 2



Zero correlation: (r = 0). Here there is no connection between the *x*-variable and the *y*-variable. Where the data points form a random pattern on the scatter plot the correlation is equal to zero.

As the data points become more scattered the correlation becomes weaker. Scatter plots make the correlation of data easy to spot.

Summary

To choose an appropriate statistical graph, consider the set of data values. In general, use the following guidelines:

- Use a **bar chart** if you are not looking for trends (or patterns) over time and the items (or categories) are not parts of a whole.
- Use a **pie chart** if you need to compare different parts of a whole, there is no time involved and there are not too many items (or categories).
- Use a **line graph** if you need to see how a quantity has changed over time. Line graphs enable us to find trends (or patterns) over time.

Class Activity 2: Methods for collecting, organising and analysing data

In small groups, complete the formative activity in your Learner Workbook

1.3 Select data sources and databases to ensure representativeness of the sample and validity of resolutions

When conducting a study, a researcher selects a relatively small group of participants (a sample) from an entire population of all possible participants (for example,

selecting university students at a couple of universities from all university students in the world).

Ideally, the researcher would have a **Representative Sample**, which is when your participants closely match the characteristics of the population, which, in turn, helps you generalise your results from your small group of people to large groups of people. For example, imagine you are at the supermarket picking out grapes. There are red, green, small, large, and globe grapes. In a representative sample you would have an **equivalent number of each type of grape**. You could then taste them all and make generalisations about all grapes just from tasting these few because your sample represents the larger population.

Class Activity 3: Select data sources and databases to ensure representativeness of the sample and validity of resolutions

In small groups, complete the formative activity in your Learner Workbook

1.4 Contamination of data

Data integrity is the assurance that data is consistent, correct, and accessible.

Problems of **inaccuracy** or **contamination of data** can occur, for example, when survey respondents refuse to answer a question, or when they give inaccurate reports. Consider, for example, a researcher who draws a random sample of persons from a population of interest and queries them about some variables, hence constructing a survey. The purpose of such a survey can be to learn features of the distribution of a variable of interest, or features of the relation between several variables; for example, an economist may be interested in learning the effect of drug abuse during work hours on employment status, or the effect of child care subsidies on single mothers' labour market outcomes.

In this situation, several problems can appear:

- Some of the selected subjects may decide not to take part in the survey.
- Others may participate in the survey, but then either refuse to answer some of the questions or give inaccurate responses; for example, some respondents might refuse to answer questions related to drug abuse or grant recipiency because of the social stigma associated with them; other respondents might react differently to the social stigma and give inaccurate reports (e.g. erroneously reporting that they are not on welfare while they truly are).

Class Activity 4: Contamination of data

In small groups, complete the formative activity in your Learner Workbook

1.5 Problem-solving using statistics (AC 8 and 9)

Statistics is a problem-solving process that seeks answers to questions through data.

By asking and answering statistical questions, we can learn more about the world around us. Statistics is used every day to help us gain insight into questions that affect our lives: Is crime in our country increasing or decreasing? What is the safest

way to invest money? Will eating more fruits and vegetables really make us live longer?

Four things make a problem statistical: the way in which you ask the question, the role and nature of the data, the particular ways in which you examine the data, and the types of interpretations you make from the investigation. A statistics problem typically contains four components:

1. Ask a question

Asking a question gets the process started. It's important to ask a question carefully, with an understanding of the data you will use to find your answer.

For example, you might want to know the heights of a group of people. If the group consists of the learners in class, we can simply ask them their heights. However, if we are interested in the heights of every person in South Africa, we cannot personally ask each one their height. In fact, even the South African **census**, which is mandated to count everyone in SA, cannot find everyone in the country.

For this reason, two words are important for the science of statistics: **population** and **sample**. The population is the group you are interested in (e.g., all people in SA). The sample is the group you can collect information about (e.g. the learners in class). A sample is always a subset of the population, i.e. every member of the sample is in the population.

Census: a comprehensive inquiry of all data obtainable.

Sample: a set of objects taken at random from a larger group

The **Law of Statistical Regularity** states that a set of objects (a sample) taken at random from a larger group (population) tends to reproduce the characteristics of that larger group.

Identify the goal or objective of the statistical inquiry

Objectives are clear statements of the specific activities required to achieve the goal. In this case we want to determine something about a collection or set of data in order to make a decision.

Identify the scope of the inquiry

- Can the required information be given in numerical terms?

The incidence of drunkenness, for example, can be measured by numbers of warnings, which will depend on the alertness and efficiency of security staff, management or the team leader

- What is the precise definition of the object to be measured?

In a wage inquiry, for example, should the data be wage rates or actual earnings, should allowance be made for overtime and bonuses, should earnings be gross or net, before or after tax, etc.?

- What should be the field of inquiry? How wide do we spread our net?
- Is there any information already available from routine statistics or published sources?

2. Collect appropriate data

Collecting data to help answer the question is an important step in the process. You obtain data by measuring something, so your measurement methods must be chosen with care. Sampling is one way to collect data; experimentation is another.

Select a sample

The larger the sample, the more closely it will resemble the population from which it is taken, while too small a sample tends not to give a reliable result. In the insurance industry, for example, actuaries base their life tables and calculations of premiums to be charged on masses of data obtained from past experience, stretching back over many years. They will then make certain assumptions about, for example, mortality (death) rates based on the data they have obtained. Other insurance companies will probably come to more or less the same conclusions, even though the records of any one company are only a sample.

Forecasts and conclusions based on statistical data can sometimes be proved incorrect or invalid by political events, economic crises, new inventions and other unforeseen circumstances. During the past two decades we have seen that insurance companies have had to adjust their life tables as a result of a dramatic rise in the mortality rate due to HIV/Aids.

The advantages of taking a sample are that:

- It costs much less and takes less time than a census
- The results are known more quickly
- It is possible to ask more questions and obtain more information than would be possible in a full-scale census
- It can be done more frequently than a full-scale census

Disadvantage of sampling:

- The major disadvantage of taking a sample as opposed to taking a census is that there is almost always some loss of accuracy.

However, if a sample survey is carried out with due care, it can give better results than a census taken indifferently.

The decision whether to take a sample or a census depends mainly on the cost of the inquiry in time and money, the delay in getting the results if a census is chosen and the degree of accuracy required.

If we want a sample to be truly representative, it has to be chosen without bias, so that every item has the same chance of being chosen. A public opinion survey, for

example, should not be taken from one particular section of the population, like the readers of a particular newspaper.

A certain investigator was working on a survey of gambling habits, in particular betting on the horses. After interviewing several people in a queue, he was puzzled to find that his results were hopelessly biased, until he discovered these people were waiting for the bus to a horse racing track!

3. Analyse the data

Data must be organised, summarised, and represented properly in order to provide good answers to statistical questions. Also, the data you collect usually vary (i.e., they are not all the same), and you will need to account for the sources of this variation.

4. Interpret the results

After you have analysed your data, you must interpret them in order to provide an answer to the original question.

Class Activity 5: Problem-solving using statistics

In small groups, complete the formative activity in your Learner Workbook

Module 2
Use theoretical and experimental probability to develop models

After completing this module, the learner will be able to use theoretical and experimental probability to develop models, by successfully completing the following:

- Experiments and simulations are chosen and/or designed appropriately in terms of the situation to be modelled
- Predictions are based on validated experimental or theoretical probabilities
- The results of experiments and simulations are interpreted correctly in terms of the real context
- The outcomes of experiments and simulations are communicated clearly

Theoretical and experimental probability

When we use formulae to work out the probability of a certain event we are calculating mathematical probability. When we try to do it physically we end up with experimental results.

In this Module, we will look into experimental probability and theoretical probability, or "mathematical expectations" vs. "experimental results".

2.1 Experiments and simulations

Experimental Probability

Experimental probability refers to the probability of an event occurring when an experiment is conducted. In such a case, the probability of an event is being determined through an actual experiment.

Mathematically,

Experimental probability= $\frac{\text{Number of event occurrences}}{\text{Total number of trials}}$

In other words, if a die is rolled 6000 times and the number '5' occurs 990 times, then the experimental probability that '5' shows up on the die is 990/6000 = 0.165.

Example:

A bag contains 10 red marbles: 8 blue marbles and 2 yellow marbles. Find the experimental probability of getting a blue marble.

Solution:

- Take a marble from the bag.
- Record the colour and return the marble.
- Repeat a few times (perhaps 10 times).
- Count the number of times a blue marble was picked (Suppose it is 6).

The experimental probability of getting a blue marble from the bag is $\frac{6}{10} = \frac{3}{5}$

Theoretical Probability

We can also find the theoretical probability of an event. Theoretical probability is determined by noting all the possible outcomes theoretically, and determining how likely the given outcome is. Mathematically, the formula for theoretical probability of an event is

Theoretical probability = $\frac{\text{Number of favourable outcomes}}{\text{Total number of outcomes}}$

For example, the theoretical probability that the number '5' shows up on a die when rolled is 1/6 = 0.167. This is because of the 6 possible outcomes (die showing '1', '2', '3', '4', '5', '6'), only 1 outcome (die showing '5') is favourable.

Example:

A bag contains 10 red marbles, 8 blue marbles and 2 yellow marbles. Find the theoretical probability of getting a blue marble.

Solution:

There are 8 blue marbles. Therefore, the number of favourable outcomes = 8.

There are a total of 20 marbles. Therefore, the number of total outcomes = 20

$$P(\text{event}) = \frac{\text{Number of favorable outcomes}}{\text{Number of total outcomes}} = \frac{8}{20} = \frac{2}{5}$$

Example:

Find the probability of rolling an even number when you roll a die containing the numbers 1-6. Express the probability as a fraction, decimal, ratio and percentage.

Solution:

The possible even numbers are 2, 4, 6. Number of favourable outcomes = 3.

Total number of outcomes = 6

$$P(\text{event}) = \frac{\text{Number of favorable outcomes}}{\text{Number of total outcomes}} = \frac{3}{6} = \frac{1}{2}$$

The probability = $\frac{1}{2}$ (fraction) = 0.5 (decimal) = 1:2 (ratio) = 50% (percent)

Class Activity 6: Experiments and simulations

Individually, complete the formative activity in your Learner Workbook

2.2 Make predictions

Experimental probability is frequently used in research and experiments of social sciences, behavioural sciences, economics and medicine.

In cases where the theoretical probability cannot be calculated, we need to rely on experimental probability.

For example[5], to find out how effective a given cure for a pathogen in mice is, we simply take a number of mice with the pathogen and inject our cure. We then find out

[5] Kalla, Siddharth (2009). Experimental Probability. Retrieved 24 August 2010 from Experiment Resources: http://www.experiment-resources.com/experimental-probability.html

how many mice were cured and this would give us the experimental probability that a mouse is cured to be the ratio of number of mice cured to the total number of mice tested.

In this case, it is not possible to calculate the theoretical probability. We can then extend this experimental probability to all mice.

It should be noted that in order for experimental probability to be meaningful in research, the sample size must be sufficiently large.

In our above example, if we test our cure on 3 mice and all of these are cured, then the experimental probability that a mouse is cured is 1. However, the sample size is too small to conclude that the cure works in 100% of the cases

As the number of trials keeps increasing, the experimental probability tends towards the theoretical probability. Therefore, the trials should be sufficiently large in number.

2.3 Interpret results of experiments and simulations

Data analysis and interpretation is the process of assigning meaning to the collected information and determining the conclusions, significance, and implications of the findings. The steps involved in data analysis are a function of the type of information collected; however, returning to the purpose of the research and the key questions will provide a structure for the organisation of the data and a focus for the analysis.

The analysis of numerical (quantitative) data is represented in mathematical terms.

The most common statistical terms include:

- Mean – The mean score represents a numerical average for a set of responses.
- Standard deviation – The standard deviation represents the distribution of the responses around the mean. It indicates the degree of consistency among the responses. The standard deviation, in conjunction with the mean, provides a better understanding of the data. For example, if the mean is 3.3 with a standard deviation (StD) of 0.4, then two-thirds of the responses lie between 2.9 (3.3 – 0.4) and 3.7 (3.3 + 0.4).
- Frequency distribution – Frequency distribution indicates the frequency of each response. For example, if respondents answer a question using an agree/disagree scale, the percentage of respondents who selected each response on the scale would be indicated. The frequency distribution provides additional information beyond the mean, since it allows for examining the level of consensus among the data.

The analysis of narrative (qualitative) data is conducted by organising the data into common themes or categories. It is often more difficult to interpret narrative data since it lacks the built-in structure found in numerical data.

Initially, the narrative data appears to be a collection of random, unconnected statements. The reearch purpose and questions can help direct the focus of the data organisation.

The following strategies may also be helpful when analysing narrative data:

Focus groups and Interviews:

- Read and organise the data from each question separately. This approach permits focusing on one question at a time (e.g., experiences with tutoring services, characteristics of tutor, student responsibility in the tutoring process).
- Group the comments by themes, topics, or categories. This approach allows for focusing on one area at a time (e.g., characteristics of tutor – level of preparation, knowledge of content area, availability).

Documents:

- Code content and characteristics of documents into various categories (e.g., training manual – policies and procedures, communication, responsibilities).

Observations:

- Code patterns from the focus of the observation (e.g., behavioural patterns – amount of time engaged/not engaged in activity, type of engagement, communication, interpersonal skills).

The analysis of the data via statistical measures and/or narrative themes should provide answers to the research questions. Interpreting the analysed data from the appropriate perspective allows for determination of the significance and implications of the research.

Class Activity 7: Interpret results of experiments and simulations and make predictions

Individually, complete the formative activity in your Learner Workbook

2.4 Communicate the outcomes of experiments and simulations

The final stage in the research process is to report the findings. For learners doing small-scale research for their own purposes, communication may be quite informal. The learner may simply draw conclusions from what he or she gleans from the data analysis.

For more serious research projects, those conducting the research will prepare a written report outlining what was researched and offer results.

Research report

Generally, your research/ project report will include the following sections:

1. Title page
2. Table of contents
3. Heading
4. Introduction
5. Body
6. Conclusion
7. Recommendations
8. References

However, it's always best to consult a style manual for your discipline, to talk to other people in your discipline who have written reports, and to look at similar reports that

have been published in order to more fully understand the expectations for reports in your field.

Introduction
The introduction prepares readers for the discussion that follows by introducing the purpose, scope, and background of the report. The audience for your report largely determines the length of the introduction and the amount of detail included in it. You should include enough detail so that someone knowledgeable in your field can understand the subject and your research.

You should begin your introduction at the top of a new page, preceded on the page only by the report's full title. The title is followed by the word Introduction, which can be either a centre or side heading.

Most introductions contain three parts to provide context for the research done: purpose, scope, and background information. These parts often overlap one another, and sometimes one of them may be omitted simply because there is no reason for it to be included.

It is very important to consider the purpose of your research and your report in the introduction. If you do not completely understand what the purpose is, there is little chance that the reader will understand your purpose either.

The following questions will help you to think about the purpose of your research and your reason for writing a report:

- What kind of problem did you work on?
- Why did you work on this problem? If the problem was assigned, try to imagine why the facilitator assigned this particular problem; what were you supposed to learn from working on it?
- Why are you writing this report?
- What should the reader know or understand when he/she has finished reading the report?

Scope refers to the ground covered by the report and will outline the method of investigation used in the project. Considering the scope of your project in the introduction will help readers to understand the parameters of your research and your report. Scope may also include defining important terms.

These questions will help you to think about the scope of both your research and your report:

- How did you work on the research problem?
- Why did you work on the problem the way you did?
- Were there other obvious approaches you could have taken to this problem? What were the limitations you faced that prevented your trying other approaches?
- What factors contributed to the way you worked on this problem? What factor was most important in deciding how to approach the problem?

Background Information includes facts that the reader must know in order to understand the discussion that follows. These facts may include descriptions of conditions or events that caused the project to be authorised or assigned and details of previous work and reports on the problem or closely related problems.

You might also want to review theories that have a bearing on the project and references to other documents although if you need to include a lengthy review of other theories or documents, these should be placed in an appendix.

Ask yourself:

- What facts does the reader need to know in order to understand the discussion that follows?
- Why was the project authorised or assigned?
- Who has done previous work on this problem?
- What facts are already known that support the theory?
- What will the reader know about the subject already and what will you need to tell them so they can understand the significance of your work?

Body

The body is usually the longest part of the research report, and it includes all of the evidence that readers need to have in order to understand the subject. This evidence includes details, data, results of tests, facts, and conclusions.

Exactly what you include in the body and how it is organised will be determined by the context in which you are writing.

Be sure to check the specific guidelines under which you are working to see if your readers are expecting you to organise the body in a particular way.

However, bear in mind that all the techniques for clear, accurate, objective writing apply, no matter what the format chosen.

Conclusion

The conclusion of a research report is usually a very short section that introduces no new ideas.

You may ask, then, why include conclusions?

The conclusion is important because it is your last chance to convey the significance and meaning of your research to your reader by concisely summarising your findings and generalising their importance.

It is also a place to raise questions that remain unanswered and to discuss ambiguous data. The conclusions you draw are opinions, based on the evidence presented in the body of your report.

Recommendations

You may or may not need to include a section titled “Recommendations.” This section appears in a report when the results and conclusions indicate that further work needs to be done or when you have considered several ways to resolve a problem or improve a situation and want to determine which one is best.

You should not introduce new ideas in the recommendations section, but rely on the evidence presented in the results and conclusions sections.

The style of the report should be concise, formal, and written in the past tense. This is the style most appropriate to written reports in any scientific or technical environment. Your sentences should present ideas in a logical sequence. Do not give instructions (e.g. write ‘A was connected to B’ rather than ‘Connect A to B’).

Paragraphs should be used to introduce new topics. You are also expected to write legibly, with good grammar, and spell accurately. You should proof read reports.
Diagrams, charts and graphs should only be computer generated if the detail can be as complete as those drawn neatly by hand. Elaborate presentation is neither required nor encouraged, clarity in your writing and presentation is your main aim.
Where a report is short, it is acceptable to combine two or more sections under one heading, e.g. Results and Discussion.

Layout for a Project Report:

TITLE

Your Name

Date(s): duration of project from start to implementation

1. Summary or Abstract

A summary is not absolutely essential, but certainly desirable, especially when we bear in mind that the average reader does not have the time or inclination to read attentively right to the last detail.

This is usually the last section written, but should head the report. It should briefly explain what the project is about, and give a concise summary of the results and their significance. As it will probably be the only section read by most readers, it must be clear.

2. Introduction

This contains the statement of the problem, and aim of the project. The introduction should normally be no more than 20% of the total report in length.

3. Solutions identified

Describe the process you and your team followed to identify possible solutions to the problem.

4. Solution chosen

Briefly state which solution you decided upon and give the reasons why: cost, labour, time, etc.

5. Implementation

This is where you say how you went about implementing the identified solution. This is also where you put your data, without any significant analysis. Data should not normally be put in an appendix. Graphs and charts should be clearly labelled. State results for quantities measured and refer to budgets if applicable.

6. Results

Analyse, interpret and discuss each result in some detail. This should include a discussion and sensible analysis of the new problems arising during implementation and how you intend addressing them.

7. Conclusion

This is not just a rehash of the summary. Try to take an overview of the project, where you've succeeded, and where further investigation might be warranted.

8. Acknowledgements and references

Those who have contributed to the work deserve to be acknowledged. Give full details of references used and referred to in preparing your report.

Additionally, an oral presentation may be required in which the research is explained within a slide presentation.

Class Activity 8: Communicate the outcomes of experiments and simulations

Individually, complete the formative activity in your Learner Workbook

Module 3
Critically interrogate and use probability and statistical models

After completing this module, the learner will be able to critically interrogate and use probability and statistical models, by successfully completing the following:

- Statistics generated from the data are interpreted meaningfully and interpretations are justified or critiqued
- Assumptions made in the collection or generation of data and statistics are defined or critiqued appropriately
- Tables, diagrams, charts and graphs are used or critiqued appropriately in the analysis and representation of data, statistics and probability values
- Predictions, conclusions and judgements are made on the basis of valid arguments and supporting data, statistics and probability models
- Evaluations of the statistics identify potential sources of bias, errors in measurement, potential uses and misuses and their effects

Probability and statistical models

A **statistical model** is a set of mathematical equations which describe the behaviour of an object of study in terms of random variables and their associated probability distributions.

Three methods are sufficient to describe all statistical models:

1. Choosing a **statistical unit**, such as a person, to **observe directly**: Multiple observations of the same unit over time are a common way of studying relationships among the attributes of a single unit. Experiments on human behaviour have special concerns. The famous Hawthorne study examined changes to the working environment at the Hawthorne plant of the Western Electric Company. The researchers were interested in determining whether increased illumination would increase the productivity of the assembly line workers. The researchers first measured the productivity in the plant, then modified the illumination in an area of the plant and checked if the changes in illumination affected productivity. It turned out that productivity indeed improved (under the experimental conditions). However, the study is heavily criticized today for errors in experimental procedures, specifically for the lack of a control group. The Hawthorne effect refers to finding that an outcome (in this case, worker productivity) changed due to observation itself. Those in the Hawthorne study became more productive not because the lighting was changed but because they were being observed!
2. Observing a **statistical population** (or set) of similar units rather than one individual unit. **Survey sampling** offers an example of this type of modelling. An example of an observational study is one that explores the correlation between smoking and lung cancer. This type of study typically uses a survey to collect observations about the area of interest and then performs statistical analysis. In this case, the researchers would collect observations of both smokers and non-smokers, perhaps through a case-control study, and then look for the number of cases of lung cancer in each group.
3. Focusing on functional **subunits** of the statistical unit; for example, physiology modelling which studies the organs which make up the unit (person) in order to better understand the statistical unit.

3.1 Interpret statistics generated from the data

Before analysing and interpreting your results, it's very important to be sure you have an adequate number of responses.

Whether you are looking at results from "all respondents" or just a demographic slice, you need to be sure there are enough respondents to make the results statistically meaningful.

How many respondents do you need? There is no hard and fast rule, but more is better. If you have a smaller number of respondents, you need much stronger results in order to draw conclusions from the numbers. If you have just 10 respondents and they all said "strongly disagree" then you can probably trust that, but if just 7 out of 10 said "strongly disagree" then you might want to collect more data to be sure there is a trend there. On the other hand, if you have 1000 respondents and 700 of them said "strongly disagree" you can be pretty sure that this result is meaningful.

If you included demographic questions in your survey, you may have a fairly good idea of who your respondents are, but keep in mind the different ways in which the

respondents might not represent all of the people in your "population". For example, if you are surveying customers and your survey is internet-based, you are not likely to get much feedback from people who don't use computers.

Every situation is unique and you should spend a bit of time thinking about the kinds of people who might be under represented in your survey results.

Garbage In - Garbage Out

Your results will only be as good as the questions that you asked. If your questions were poorly worded, you are probably going to find that your data are not very useful.

Two common examples of this to be aware of:

1. The "easy" question - if a question is too softly worded or has an obvious answer, you will find that almost all respondents answered with the same response. This means your question has not effectively distinguished what it was intended to distinguish.
2. The "confusing" question - there are so many forms of confusing questions, but they generally all yield similar response patterns - you will see an unusually high number of "unable to rate" responses as well a larger than average spread of responses in the frequency distribution. What you are seeing in this is that people just did not understand the question or that different people interpreted the question differently.

Quantitative (Numeric) Data

Start by looking at the numbers. Generate a report for all respondents and look at the following:

1. Overall Average Scores - high or low? This is the obvious first place to start. Very high or very low scores mean either that you are doing really well or really poorly in an area - or they might mean that the question is poorly worded.
2. Relative Scores - how do the scores on each item compare to the scores on similar items in your survey?
3. Look at the results for the different demographic subgroups, especially focusing on the items where you had interesting things happening in the frequency distributions.

Qualitative (Text or Open Ended; Non-Numeric) Data

Some of your greatest opportunities to understand your results will come from the comments that people have provided. Remember that satisfied people often don't make comments or have little to say, so if you find a disproportionate number of negative comments, don't be too discouraged. Look at each of them as an opportunity. Just as with numeric data, you should look for trends in the qualitative data. You will probably need a much larger sample to spot trends, but they are important to identify so you don't get misguided by one or two comments that might not reflect the views of very many of your constituents.

Qualitative Data Analysis:

1. Start by reading through all the comments. Get a feeling for what people are saying.

2. Now go back and categorise the comments into different areas. The categories you put them into are up to you, but after having read through all the comments, you should have an idea of where to begin. Do your best to categorise all the comments, but don't be too concerned if you have a handful left over at the end which don't fit in any category.
3. Now look at each category separately. How many unique comments are in each? How detailed are those comments? How strongly are they stated? At this point, you should be able to identify which categories are more important and which are less important. It's not an exact process, but patterns almost always emerge if you have enough response data to work with. If you find that you have several categories which seem to be equally important, that's fine too.
4. Now, if your survey included demographic questions, look at the different subgroups to see if any relationships emerge between demographic groups and categories of comments. This might seem like a time consuming process, but the outcome will be worth the effort.

Qualitative and Quantitative Data		
	Quantitative	**Qualitative**
Objective	"The chip speed of my computer is 2 GHz"	"Yes, I own a computer"
Subjective	"On a scale of 1-10, my computer scores 7 in terms of its ease of use"	"I think computers are too expensive"

Class Activity 9: Interpret statistics generated from the data
Individually, complete the formative activity in your Learner Workbook

3.2 Define and critique assumptions

A **critique** is a systematic response, evaluation and careful analysis of an argument to determine:

- What is said
- How well the points are made
- What assumptions underlie the argument
- What issues are overlooked
- What implications are drawn from such observations.

What are assumptions?

Assumptions are accepted cause and effect relationships, or estimates of the existence of a fact from the known existence of other fact(s).

Although useful in providing a basis for action and in creating 'what if' scenarios to simulate different realities or possible situations, assumptions are dangerous when accepted as reality without thorough examination.

Given that the validity of conclusions drawn from a statistical analysis depend on the validity of any assumptions made, it is important that these assumptions should be reviewed at some stage. In some instances, for example, where data are lacking, this may have to be restricted to just making a judgement about whether an assumption is reasonable.

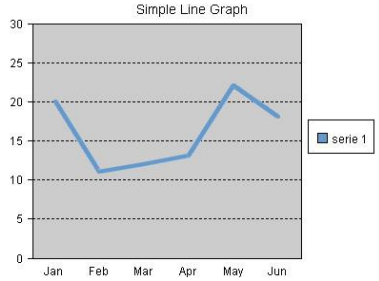
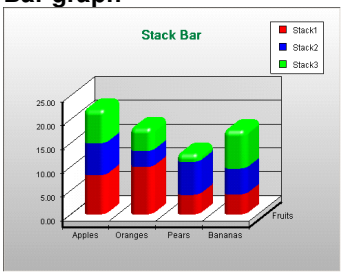
This can also be expanded slightly to trying to judge what effect a departure from the assumptions might have.

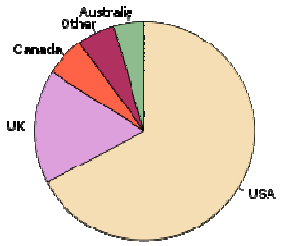
Class Activity 10: Define and critique assumptions
Individually, complete the formative activity in your Learner Workbook

3.3 Critique tables, diagrams, charts and graphs

As you saw in Module 1, there are many chart and diagram formats you can choose from when representing information graphically. Selecting the right type starts with a good understanding of how each is created and what each is intended to depict.

Using the correct type of chart in your analysis will add great value and improve the clarity and effectiveness of your communication.

Type	Most useful for
Line graph 	Connecting the data points that you plot. They are most useful for showing trends, and for identifying whether two variables relate to (or "correlate with") one another. Trend data: • How do sales vary from month to month? • How does engine performance change as its temperature increases? Correlation: • On average, how much sleep do people get, based on their age? • Does the distance a child lives from school affect how frequently he or she is late? You can only use line graphs when the variable plotted along the x-axis is continuous - for example, time, temperature or distance.
Bar graph 	Shows relationships between different data series. Here the height of the bar represents the measured value or frequency: The higher or longer the bar, the greater the value. When your x-axis variables represent discontinuous data (such as different products or sales territories), you can only use a bar graph.

	In general, line graphs are used to demonstrate data that is related on a continuous scale, whereas bar graphs are used to demonstrate discontinuous data. **Note:** A bar graph is not the same as a histogram. On a histogram, the width of the bar varies according to the range of the x-axis variable (for example, 0-2, 3-10, 11-20, 20-40 and so on) and the area of the column indicates the frequency of the data. With a bar graph, it is only the height of the bar that matters.
Pie chart  	The data you are measuring must depict a ratio or percentage relationship. You must always use the same unit of measure within a pie chart. **Tip 1:** Be careful not to use too many segments in your pie chart. More than about six and it gets far too crowded. Here it is better to use a bar chart instead. **Tip 2:** If you want to emphasise one of the segments, you can detach it a bit from the main pie. This visual separation makes it stand out. **Tip 3:** Pie charts are useful when a picture of the data makes meaningful relationships visible (patterns, trends, and exceptions) that could not be easily discerned from a table of the same data.
Table **Companies** / **Percentage** Company B 40% Company C 25% Company D 17% Company A 10% Company E 7% Company F 1% Total 100%	A properly designed table is easy to read:

Class Activity 11: Critique tables, diagrams, charts and graphs
Individually, complete the formative activity in your Learner Workbook

3.4 Make predictions, conclusions and judgements

Data analysis involves breaking down the data to draw significant insights. The steps involved in data analysis are:

- Validate (make sure that the data is good)
- Summarise (compute central tendency and dispersion)
- Assess
- Search for structure (relationship between variables with scatter plots, correlation, etc)
- Compare (mean, standard deviation, etc)
- Present results
- Draw conclusions.

How to make predictions from a graph

When looking at a spreadsheet of data, it can be difficult to notice trends. Converting that data into a graph provides an easier way of looking at the same information. A graph can make it much easier to notice trends that are not as obvious when viewing the raw data, which makes it much easier to make predictions. For example, if a graph is showing an upward trend in sales, then a person can make a reasonable prediction that sales will continue their upward trend as long as no variables change.

You can look for trends in the graph. For example, sales of ice cream are likely to go down in the winter months and then go back up during the summer months. Viewing a graph that covers several years of sales would likely show this trend. Once you see the trend repeated over several years, you can reasonably make a prediction from a graph that ice cream sales during the next year will be low in July and high in January.

You can also observe places where the graph deviates from the trend. For example, if an ice cream parlour had a fire and closed its door in July one year, then sales would be down during that time. This would be a deviation from the general trend shown on the graph. When you see a deviation, try to identify the variables that could account for the deviation. Then, factor in those variables when making predictions from the graph about future sales.

Analyse specific types of graphs. With a line graph, it is fairly easy to make predictions because line graphs show changes over a period of time. You can look at past performance in a line graph and make a prediction about future performance. With bar graphs and pie charts, you need to compare graphs from different periods of time and notice the changes between the two to make predictions.

Class Activity 12: Make predictions, conclusions and judgements
Individually, complete the formative activity in your Learner Workbook

3.5 Identify potential sources of bias, errors in measurement, potential uses and misuses and their effects

An error, in the statistical sense, is the difference between an estimate and the true or ideal value, and is not necessarily a mistake as we know it. Errors can be broadly classified as errors of origin, errors of inadequacy, errors of manipulation and errors of analysis:

- Errors of origin are due to the defects in the data collected, e.g. false information caused by forgetfulness or misunderstanding.
- Errors of inadequacy arise when the items observed are too few to provide a representative sample of the population.
- Errors of manipulation are mistakes in counting or measuring, or any mistakes made in observing and handling the data. Sampling errors occur due to variables such as measuring at different times of the day, or taking samples from different parts of a mixture.
- Errors of analysis will occur as a result of sending samples to different laboratories, or when using different methods of analysis.

In any statistical investigation, we can always attribute some of the variation in data to measurement error, part of which can result from the measurement instrument itself. But human mistakes, especially recording errors (e.g., misreading a dial, incorrectly writing a number, not observing an important event, misjudging a particular behaviour), can also often contribute to the variability of the measurement and thus to the results of a study.

Accuracy

Absolute accuracy is generally impossible, so very often all that required is a rough estimate where figures have been rounded off, for example to the nearest thousand (often indicated by a column heading such as R000), or a range within which the true figure is believed to lie, for example 267 000 can be indicated as lying between 265 000 and 269 000, or as 267 000 ± 1 500, meaning that the estimate may be out by 1 500 either way.

What is a Margin of Error?

When results of surveys are reported in the media, they often include a statement like "55 percent of respondents favour Ms. Smith in the upcoming mayoral election. There is a margin of error of 3 percentage points."

What does a statement like this mean?

Surveys are typically designed to provide an estimate of the true value of one or more characteristics of a population at a given time. The target of a survey might be:

- *The average value* of a measurable quantity, such as annual 2009 income.
- *A proportion*, such as the proportion of likely voters having a certain viewpoint in an election
- The *percentage* of children under three years of age immunised for polio in 2010

An estimate from a survey is unlikely to exactly equal the true population quantity of interest for a variety of reasons. For one thing, the questions maybe badly worded. For another, some people who are supposed to be in the sample may not be at home, or even if they are, they may refuse to participate or may not tell the truth. These are sources of "*non-sampling error*."

But the estimate will probably still differ from the true value, even if all non-sampling errors could be eliminated. This is because data in a survey are collected from only some-but not all-members of the population to make data collection cheaper or faster, usually both.

Suppose, in the election poll mentioned earlier, we sample 100 people who intend to vote and that 55 support Ms. Smith while 45 support Mr. Jones. This would seem to suggest that a majority of the voters, including people not sampled but who will vote in the election, would support Ms. Smith.

Of course, just by chance, a majority in a particular sample might support Ms. Smith even if the majority in the population supports Mr. Jones. Such an occurrence might arise due to "*sampling error*," meaning that results in the sample differ from a target population quantity, simply due to the "luck of the draw", *i.e. by which set of 100 people were chosen to be in the sample*.

Does sampling error render surveys useless? Fortunately, the answer to this question is "No." But how should we summarise the strength of the information in a survey? That is a role for the margin of error.

The "margin of error" is a common summary of sampling error, referred to regularly in the media, which quantifies uncertainty about a survey result. The margin of error can be interpreted by making use of ideas from the laws of probability or the "laws of chance," as they are sometimes called.

Surveys are often conducted by starting out with a list (known as the "sampling frame") of all units in the population and choosing a sample. In opinion polls, this list often consists of all possible phone numbers in a certain geographic area (both listed and unlisted numbers).

In a scientific survey every unit in the population has some known positive probability of being selected for the sample, and the probability of any particular sample being chosen can be calculated.

The beauty of a probability sample is twofold. Not only does it avoid biases that might arise if samples were selected based on the whims of the interviewer, but it also provides a basis for estimating the extent of sampling error. This latter property is what enables investigators to calculate a "margin of error."

Such intervals are sometimes called 95 percent confidence intervals and would be expected to contain the true value of the target quantity (in the absence of non-sampling errors) at least 95 percent of the time. An important factor in determining the margin of error is the size of the sample. Larger samples are more likely to yield

results close to the target population quantity and thus have smaller margins of error than more modest-sized samples.

In the case of the election poll in which 55 of 100 sampled individuals support Ms. Smith, the sample estimate would be that 55 percent support Ms. Smith- however, there is a margin of error of 10 percent. Therefore, a 95 percent confidence interval for the percentage supporting Ms. Smith would be (55%-10%) to (55%+10%) or (45 percent, 65 percent), suggesting that in the broader community the support for Ms. Smith could plausibly range from 45 percent to 65 percent. This implies-because of the small sample size-considerable uncertainty about whether a majority of townspeople actually support Ms. Smith.

Instead, if there had been a survey of 1,000 people, 550 of whom support Ms. Smith, the sample estimate would again be 55 percent, but now the margin of error for Ms. Smith's support would only be about 3 percent. A 95 percent confidence interval for the proportion supporting Ms. Smith would thus be (55%-3%) to (55%+3%) or (52 percent, 58 percent), which provides much greater assurance that a majority of the town's voters support Ms. Smith.

Three things that seem to affect the margin of error are sample size, the type of sampling done, and the size of the population.

Sample Size- the size of a sample is a crucial factor affecting the margin of error. In sampling, in order to estimate a population proportion-such as in telephone polls- a sample of 100 will produce a margin of error of no more than about 10 percent, a sample of 500 will produce a margin of error of no more than about 4.5 percent, and a sample of size 1,000 will produce a margin of error of no more than about 3 percent. This illustrates that there are diminishing returns when trying to reduce the margin of error by increasing the sample size. For example, to reduce the margin of error to 1.5% would require a sample size of well over 4,000.

The survey researcher also has control over the **design of the sample**, which can affect the margin of error. Three common types are simple random sampling, random digit dialling, and stratified sampling.

- A simple random sampling design is one in which every sample of a given size is equally likely to be chosen. In this case, individuals might be selected into such a sample based on a randomising device that gives each individual a chance of selection. Computers are often used to simulate a random stream of numbers to support this effort.
- Telephone surveys that attempt to reach not only people with listed phone numbers but also people with unlisted numbers often rely on the technique of random digit dialling.
- Stratified sampling designs involve defining groups, or strata, based on characteristics known for everyone in the population, and then taking independent samples within each stratum. Such a design offers flexibility, and, depending on the nature of the strata, they can also improve the precision of estimates of target quantities (or equivalently, reduce their margins of error).

Of the three types of probability sampling, stratified samples are especially advantageous when the target of the survey is not necessarily to estimate the proportion of an entire population with a particular viewpoint but instead is to estimate differences in viewpoints between different groups.

Size of Population-Perhaps surprising to some, one factor that generally has little influence on the margin of error is the size of the population. That is, a sample size of

100 in a population of 10,000 will have almost the same margin of error as a sample size of 100 in a population of 10 million.

Class Activity 13: Identify potential sources of bias, errors in measurement, potential uses and misuses and their effects

Individually, complete the formative activity in your Learner Workbook

Reflection

Individually, complete the formative activity in your Learner Workbook

References and Further Reading

1. Connor, L.R. and A.J.H. Morrell. 1962. ***Statistics in theory and practice.*** London: Sir Isaac Pitman & Sons.
2. http://www.johngalt.com/statistical_methods.shtml
3. www.statssa.gov.za
4. http://www.whatisasurvey.info
5. http://www.animatedsoftware.com/statglos/sgrandsa.htm
6. http://www.mathsteacher.com.au/year8/ch17_stat/02_mean/mean.htm
7. http://www.arts.cornell.edu/econ/fmolinari/dissertation.pdf
8. http://www.alleydog.com/glossary/definition.php?term=Representative%20Sample
9. http://www.learner.org/courses/learningmath/data/session1/part_a/index.html
10. http://www.mathgoodies.com/lessons/vol6/intro_probability.html
11. http://psychology.wikia.com/wiki/Association_(statistics)
12. http://www.onlinemathlearning.com/theoretical-probability.html
13. http://www.experiment-resources.com/experimental-probability.html
14. http://en.wikipedia.org/wiki/Observation
15. https://oira.syr.edu/oira/Assessment/AssessPP/Analyze.htm
16. http://en.wikipedia.org/wiki/Statistical_model
17. http://www.custominsight.com/articles/interpreting-survey-data.asp
18. http://en.wikipedia.org/wiki/Statistics
19. http://www.perceptualedge.com/articles/08-21-07.pdf
20. http://www.mindtools.com/pages/article/Charts_and_Diagrams.htm
21. http://www.userfocus.co.uk/articles/datathink.html
22. http://www.ehow.com/how_4481030_make-predictions-from-graph.html#ixzz0xj6ZFfhZ