STATS 101/108 PRE-SEMESTER WORKSHOPS

for Semester Two, 2010

Students **MUST REGISTER** for <u>each workshop</u> with
The Student Learning Centre

Student Learning Centre

Topics we teach and can provide advice on include:

- ✓ Essay writing
- ✓ Computer skills
- ✓ Reading and notetaking
- ✓ Memory and concentration
- ✓ Report writing
- ✓ Test and examination skills
- ✓ Thesis and dissertation writing
- ✓ Tutorial skills
- ✓ Research skills
- ✓ Time and stress management
- ✓ Mathematics
- ✓ **Statistics**
- ✓ Oral presentation and seminar skills
- ✓ Language learning
- ✓ Specific learning disabilities
- ✓ Motivation and goal setting
- ✓ Survival skills (in the University system)

Programmes within SLC include:

- Te Puni Wananga
  Maori university tutors committed to enhancing Maori students' success

- Fale Pasifika
  Pacific Island tutors committed to enhancing success for Pacific Island students

- Language Exchange (LEX)
  Learn a new language, make a new friend

- English Conversation Groups
  Improve English, develop critical thinking, and express ideas and opinions

Statistical help available at the SLC

The Student Learning Centre (SLC) offers help for STATS 101/108 by offering a number of workshops

SLC STATS 101/108 Workshops

Any questions regarding STATS 101/108 workshops should be forwarded to:

Leila Boyle
SLC Undergraduate Statistics Co-ordinator
Email: l.boyle@auckland.ac.nz

Workshops are run in a relaxed environment, typically set at a pace for those students that find the Statistics Department's tutorials too fast. All workshops allow plenty of time for questions.

In fact, this is encouraged ☺

1) Saturday Workshops

These will be five Saturday workshops throughout the semester to help students with different sections of the course.

2) Computer Workshops: Excel / PASW (previously SPSS)

Three different 1-hour computer-based workshops introducing students to the *Excel* computer program and the statistics computer package PASW (previously SPSS) will be held throughout the semester. These classes will be helpful for each of the three assignments.

3) Pre-test Workshops

There will be three workshops to cover the basics that you need to know for the test.

4) Pre-exam Workshops

There will be six workshops to cover the basics that you need to know for the exam. Each will be repeated twice.

Note: All workshops concentrate on questions reviewing the basic concepts, rather than questions on finer details. They are designed to assist students to achieve a pass; they are not designed to cover all material.

Useful Websites

- SLC webpage: **www.slc.auckland.ac.nz**
- Online enrolment through the SLC site: **www.slc.auckland.ac.nz**, click on "**SLC Workshops**" and then click on "**Undergraduate Workshops**"
  - For STATS 101/108 workshops, view by "**Statistics**"
  - For other SLC workshops on skills appropriate for undergraduate students, scroll down the page or view by the appropriate category
  - For SLC workshops aimed at helping students learn basic computing skills (e.g. Excel, Word, PowerPoint), view by "**Computer skills**"
- Cecil: **https://cecil.auckland.ac.nz/**
- **Leila's website for STATS 101/108 SLC workshop handouts & information: www.stat.auckland.ac.nz/~leila**

FRACTIONS, DECIMALS AND PERCENTAGES

Exercise 1

1. **(a)** What is 50% of 8? **(b)** What is a half of a third?

2. **(a)** What is $\frac{24}{60}$ as a percentage? **(b)** What is $\frac{24}{60}$ as a decimal?

3. Convert the following numbers to both fractions and decimals:

(a) 50% **(b)** 0.5%

(c) 100% **(d)** 5%

(e) 1% **(f)** 99%

(g) 10% **(h)** 25%

(i) 3.5%

4. **(a)** What is .34 as a percentage? **(b)** What is .015 as a percentage?

ORDER OF OPERATIONS

In Mathematics we carry out operations such as +, -, ÷, ×, on **numbers** or **expressions**. **Expressions** involve numbers and symbols.

For example, to find:

7 + 6	we carry out the operation of addition on the two numbers 7 and 6
8^2	we carry out the operation of finding the square of the single number 8
$\sqrt{16+9}$	we carry out the operation of finding the square root on the expression 16 + 9

Calculate $\frac{16}{8}$ means divide the number 16 by the number 8, but $\frac{3+7}{4-1}$ means divide the expression 3 + 7 by the expression 4 - 1.

When each operation is carried out on numbers alone, that is, there are no expressions involved, the order of operations in mathematics is:

Step 1: Brackets

If there is more than one set of brackets, work from the innermost one outwards.

Step 2: Exponents

This is when a number is to "the power of" another number. This includes squaring, cubing etc.

Step 3: Division or Multiplication

These operations are done at the same time, working from left to right.

Step 4: Addition and Subtraction

These operations are also done at the same time, working from left to right.

For example to calculate: $(4+3^2)-6\times 2$

calculate the power first $=(4+9)-6\times 2$

simplify the brackets $=13-6\times 2$

carry out the multiplication $=13-12$

then do the subtraction $=1$

To describe the four steps, you may have learnt the following the acronym:

Brackets
Exponents
Division
Multiplication
Addition
Subtraction

The problem with BEDMAS is that it appears as if division takes precedence over multiplication and addition takes precedence over subtraction. This is not so:

To calculate: $8\times 2+16\div 4$

$=(8\times 2)+16\div 4$ (working from left to right)

$=16+16\div 4$

$=16+(16\div 4)$ (division before addition)

$=16+4$

$=20$

To calculate: $7-5+2\times 1-4$

$=7-5+(2\times 1)-4$ (multiplication first)

$=7-5+2-4$

$=(7-5)+2-4$ (again working from left to right)

$=2+2-4$

$=(2+2)-4$

$=4-4$

$=0$

Exercise 2

1) $5 + \frac{1}{2}$ of 6

2) $7 + 6 \div 2 \times 3$

3) $-4 \times [(3 + 7) \div 2] + 30$

4) $\sqrt{16} - 2^2 + \sqrt{9} \times \sqrt{4}$

Invisible brackets

When operations are carried out on expressions, the expression must be simplified first to a single number, then the order of operations can be carried out as above.

Consider this problem: $\frac{8+2}{2+3}$

You probably know that the long line means divide. So this says divide the numerator (the part above the line) by the denominator (the part below the line).

However, we cannot do the division first as we are dividing an expression by an expression.

We can do the addition in the numerator: $8 + 2 = 10$

and the addition in the denominator: $2 + 3 = 5$

Now we can do the division: $\frac{10}{5} = 2$

We could also write the problem like this:

$$(8 + 2) \div (2 + 3) = 10 \div 5 = 2$$

So, first look for operations on expressions.

Treat any **expressions** as though they are enclosed in **invisible brackets** and simplify them. Carry out operations on **numbers only** using the usual order of operations.

In the example $\frac{3+4}{\sqrt{12-8}}$ neither the division nor the square root operation can be carried out immediately as these are operations on expressions.

However, we can, do the:

addition	3 + 4 = 7, and
subtraction	12 - 8 = 4,
	$\frac{7}{\sqrt{4}}$
then the square root	$\sqrt{4}=2$
finally, the division:	$\frac{7}{2}$
or in another form	$7 \div 2 = 3\frac{1}{2}$

Exercise 3

Evaluate the following

1) $\sqrt{3^2+4^2}$

2) $\frac{6+3\times(5-7)}{10^2}$

3) $\frac{2^2+3^2}{\sqrt{81}-2^3}$

4) $\frac{10}{10+10}$

5) $\frac{(4-5)\times 5}{\sqrt{13^2-12^2}}$

CALCULATORS

ALSO SEE *CALCULATOR SKILLS* (LECTURE WORKBOOK, CHAPTER 2, PAGES 6-9)

1. Presume you know $\boxed{+}$ $\boxed{-}$ $\boxed{\times}$ $\boxed{\div}$

2. As most scientific calculators have an in-built order of operations the equals sign (=) indicates the end of data entry as well as an instruction to calculate up to this point.

2 $\boxed{+}$ 3 $\boxed{=}$ $\boxed{\times}$ 4 $\boxed{=}$ 20

This key sequence will do 2 + 3 = 5 then 5 × 4 = 20

3. Your aim in using a calculator is to be accurate and to be able to produce the answer **without** having to note figures down on paper part-way through the calculation.

4. Calculators vary. Make sure you understand the **colour** coding on the one you use as each key often has more than one use.

5. There are often several ways of doing the same problem.

Example: From the following sample data 12, 25, 19, 20, 35 use your calculator to determine the sample mean, standard deviation and sample size.

<u>For MY calculator:</u>

1. To put my calculator into **statistics mode** I use:
2. To **clear** the **statistics memories** I use:
3. The **data entry key** on my calculator is:
4. To check the **number of observations** entered I use:
5. I find the **sample mean**, $\bar{x}$, by:

   I find the **sample standard deviation**, σ_{n-1}, by:
6. I enter **frequency data** by using the following key/s between the number and its frequency:

(Example answers: n = 5, $\bar{x}$ = 22.2, σ_{n-1}= 8.5264)

Exercise 4

Calculate the sample mean and sample standard deviation for the following sets of numbers:

	sample size n	sample mean $\bar{x}$	sample std dev s / σ_{n-1}
(a) 16, 18, 20			
(b) 6.2, 7.1, 9.2, 8.1, 1.2			
(c) 1, -5, 6, -3, 2, 0			
(d) 12, 15.6, 2, -15, -15, 2			
(e) 15.2, -22.5, -62.5, 25.0, 0.0, 8.25			

Special Keys

(and) are keys to open and close brackets respectively. This is very useful where implied brackets need to be put in.

x^2 this calculates the square

$\sqrt{\ }$ this gives the positive square root of a number

+/− or (−) changes the sign of the number. This is how a number is made negative.

Examples

Use your calculator to work through the following examples. Note: you may have to use one of the following keys:

[2nd Function] , [SHIFT] or [INV] key to get the function you want.

1. $\mathbf{6.3^2}$

6.3 [x^2] [=] 39.69

2. $\mathbf{\sqrt{25}}$

[√] 25 [=] 5 **or** 25 [√] [=] 5

3. **–8.67 + 5.1**

Either

8.67 [+/–] [+] 5.1 [=] –3.57

or

[–] 8.67 [+] 5.1 [=] –3.57

4. $\mathbf{\frac{(3.1 + 8.6)}{2.1}}$

Either

[(] 3.1 [+] 8.6 [)] [÷] 2.1 [=] 5.571428571

or

3.1 [+] 8.6 [=] [÷] 2.1 [=] 5.571428571

5. $\mathbf{\frac{8.31}{(2.7 + 17.3)}}$

8.31 [÷] [(] 2.7 [+] 17.3 [)] [=] 0.4155

6. $\mathbf{6.1 + (3.2 - 2.65)^2}$

Either

6.1 [+] [(] 3.2 [−] 2.65 [)] [x^2] [=] 6.4025

or

3.2 [−] 2.65 [=] [x^2] [+] 6.1 [=] 6.4025

7. $\dfrac{\sqrt{\mathbf{7.29 - 3.68}}}{\mathbf{2.7}}$

Either

[√] [(] 7.29 [−] 3.68 [)] [÷] 2.7 [=] 0.703703

or

7.29 [−] 3.68 [=] [√] [ANS] [÷] 2.7 [=] 0.703703

8. $\dfrac{\mathbf{5.3}}{\mathbf{2.7 + 3.8}}$

Either

5.3 [÷] [(] 2.7 [+] 3.8 [)] [=] 0.8153846

or using the reciprocal key.

2.7 [+] 3.8 [=] [x^{-1}] [x] 5.3 [=] 0.8153846

9. $\sqrt{\dfrac{\mathbf{23.1 + 17.8}}{\mathbf{9.41 - 6.84}}}$

Either

[√] [(] [(] 23.1 [+] 17.8 [)] [÷] [(] 9.41 [−] 6.84 [)] [)] [=]

3.98928526

or

[(] 23.1 [+] 17.8 [)] [÷] [(] 9.41 [−] 6.84 [)] [=] [√] [ANS] [=]

3.98928526

10. $\dfrac{9}{13 - 5.6} + \dfrac{8.7 + 9.2}{7.3}$

9 [÷] [(] 13 [−] 5.6 [)] [+] [(] 8.7 [+] 9.2 [)] [÷] 7.3 [=]

3.668271011

Evaluate each of the following using the key indicated.

Exercise 5 [+/−] [(−)] **key**

(a) −6.4 + 3.8 =

(b) −7.3 − 5.16 =

(c) − 9.82 x 6.43 =

(d) −71.6 + 8.4 x −2.6 =

(e) −36.9 + 47.63 =

Exercise 6 Bracket keys

(a) (4.6 + 5.2) x 3.5 =

(b) (45.3 − 21.7) ÷ 0.72 =

(c) 26.3 x 2.8 x (4.9 − 3.75) =

(d) $\dfrac{61.3 + 17.2}{86.5}$ =

(e) $\dfrac{76.9}{36.2} + \dfrac{84.2}{19.7}$ =

(f) $\dfrac{76.9 - 7.83}{96.1 + 17.8}$ =

(g) $\dfrac{8.15 - 5.93}{7.2 \times 9.6}$ =

Exercise 7 $\boxed{\sqrt{\ }}$ key

(a) $\sqrt{86.3 + 17.2}$ =

(b) $\sqrt{2.6 \times (8.6 - 5.3)}$ =

(c) $\sqrt{\dfrac{4.73 + 8.96}{2.1}}$ =

(d) $\dfrac{\sqrt{7.4 + 8.95}}{6.8}$ =

(e) $\sqrt{\dfrac{31.2}{46.5} + \dfrac{71.3}{16.8}}$ =

(f) $\sqrt{\dfrac{1.2}{2.35 + 7.92}}$ =

(g) $\dfrac{643 - 756}{\sqrt{0.945}}$ =

Exercise 8 $\boxed{x^2}$ key

(a) $12.2^2 - 5.3$ =

(b) 4.2×3.8^2 =

(c) $\dfrac{79.6^2}{4.3 - 2.15}$ =

(d) $(1.2 + 5.78)^2$

(e) $(4.2 \times 3.8)^2$ =

(f) $1.2 + 6.9^2$ =

(g) $\dfrac{4.1}{7.9^2} + \dfrac{4.5^2}{2.6}$ =

ANSWERS

Exercise 1

1. **(a)** 4 **(b)** $\frac{1}{6}$

2. **(a)** 40% **(b)** 0.4

3. **(a)** $\frac{1}{2}$, 0.5 **(b)** $\frac{5}{1000} = \frac{1}{200}$, 0.005 **(c)** 1

(d) $\frac{1}{20}$, 0.05 **(e)** $\frac{1}{100}$, 0.01 **(f)** $\frac{99}{100}$, 0.99

(g) $\frac{1}{10}$, 0.1 **(h)** $\frac{1}{4}$, 0.25 **(i)** $\frac{7}{200}$, 0.035

4. **(a)** 34% **(b)** 1.5%

Exercise 2

1) 8
2) 16
3) 10
4) 6

Exercise 3

1) 5
2) 0
3) 13
4) ½
5) -1

Exercise 5 [+/–]

(a) –2.6
(b) –12.46
(c) –63.1426
(d) –93.44
(e) 10.73

Exercise 6 Bracket keys

(a) 34.3 **(b)** 32.777778
(c) 84.686 **(d)** 0.9075144
(e) 6.3984211 **(f)** 0.6064091
(g) 0.032118

Exercise 7 [√]

(a) 10.173495 **(b)** 2.9291637
(c) 2.5532426 **(d)** 0.5946343
(e) 2.2169834 **(f)** 0.3418262
(g) –116.2419

Exercise 8 [x^2]

(a) 143.54 **(b)** 60.648
(c) 2947.0512 **(d)** 48.7204
(e) 254.7216 **(f)** 48.81
(g) 7.8541561

Exercise 4

	sample size, n	sample mean, $\bar{x}$	sample std dev, s / σ_{n-1}
(a)	$n = 3$	$\bar{x} = 18$	$s = 2$
(b)	$n = 5$	$\bar{x} = 6.36$	$s = 3.0940$
(c)	$n = 6$	$\bar{x} = 0.1667$	$s = 3.8687$
(d)	$n = 6$	$\bar{x} = 0.2667$	$s = 12.9995$
(e)	$n = 6$	$\bar{x} = -6.0917$	$s = 31.9811$

THE LANGUAGE OF MATHEMATICS

Mathematical symbols

We can regard a mathematical statement as a sentence in a very simple foreign language.

We use	**numbers**	such as	1.2, $\frac{3}{4}$, -7
	symbols	such as	$+, \times, \div, -, \sqrt{\ }, ^2$
	letters	for example	x, X, a, y, s

In statistics you will need these special symbols:

$\bar{x}$ (read as "x bar")

x_i (read as "x subscript i") – e.g. x_1 (read as "x one"), x_2 (read as "x two"), etc

and some Greek letters:

μ	(mu - pronounced mew)
σ	(sigma)
β	(beta)
χ	(chi - pronounced kie)
Σ	(capital sigma)

Example 1:

Suppose we have a list of scores on a test, say the first score is 80, the second is 60, the third is 77, the fourth is 83, and the fifth is 79. We could write this as follows:

$x_1 = 80$, $x_2 = 60$, $x_3 = 77$, $x_4 = 83$, $x_5 = 79$.

The subscript tells us which score it is - and this notation has the advantage of not restricting us to 26 letters of the alphabet.

If we want to find the average of these scores by adding them up and dividing by 5 - and call our answer $\overline{X}$ we would write:

$$\overline{X} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}$$
$$= \frac{80 + 60 + 77 + 83 + 79}{5}$$
$$= 75.8$$

You may ask, "Why not write $\overline{X}$ in terms of the numbers in the first place instead of introducing all that notation?"

What makes algebra so useful and powerful, is that it gives us a concise way of expressing a general fact.

Example 2:

Suppose a group of newly born babies require a certain medication. The amount needed is 2ml plus half a ml for each kilogram of body weight. Thus a baby weighing 3kg needs 2 + 3 x ½ mL, i.e. 3½ mL. We can express the formulae mathematically as follows:

Let	W_i be the weight of the *i*th baby in kg
so that	W_1 is the weight of the 1st baby
	W_2 is the weight of the 2nd baby, and so on.
Let	A_i be the amount of medication needed by the *i*th baby in ml
Then	$A_i = 2 + ½ W_i$

As W_i changes from baby to baby, so will A_i.

Exercise 1

Express each of the following as an algebraic formula:

1) The range (R) of a set of scores is the difference between the largest score (L) and the smallest score (S).

2) The weights of n children are W_1, W_2, ...W_n.

a) Their total weight, T, is the sum of their weights.

b) The mean of their weights, $\overline{X}$, is their total weight divided by the number of children.

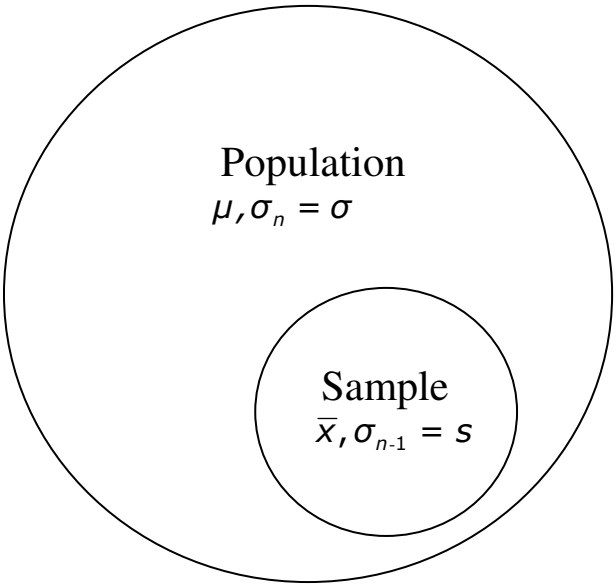
3) A Z-score (Z) is found by subtracting the mean ($\overline{X}$) from the raw score (X) and dividing the answer by the standard deviation (S).

Notation for Statistics

A commonly used convention in statistics is to use **Greek letters for population values** and **Roman letters (English alphabet) for sample variables.**

A **population** is the whole group we are interested in. E.g., the population of all employees in Australia, the population of students at Auckland University.

A **sample** is a portion of the population. You will be dealing mainly with simple random samples.



Inequalities

$a > 1$ means that a is greater than 1

$a < 3$ means that a is less than three.

$90 \leq a \leq 100$ means that a is bigger/greater than or equal to 90 but smaller/less than or equal to 100.

SIGMA Notation

The symbol $\sum$ (capital sigma) is often used as shorthand notation to indicate the sum of a number of similar terms. Sigma notation is used extensively in statistics.

Example 3:

Suppose we weigh five children. We will denote their weights by x_1, x_2, x_3, x_4 and x_5.

The sum of their weights $x_1 + x_2 + x_3 + x_4 + x_5$ is written more compactly as

$$\sum_{j=1}^{5} x_j .$$

- The symbol $\sum$ means "add up".
- Underneath $\sum$ we see "j=1" and on top of it "5". This means that j is replaced by whole numbers starting at the bottom number, 1, until the top number, 5, is reached.

Thus $\sum_{j=2}^{5} x_j = x_2 + x_3 + x_4 + x_5$

and $\sum_{j=2}^{4} x_j = x_2 + x_3 + x_4$

So the notation $\sum_{j=1}^{n} x_j$ tells us:

a) to add the scores x_j

b) where to start: x_1

c) where to stop: x_n (where n is some number).

Example 4:

Now take the weights of the children to be

x_1 = 10kg, x_2 = 12kg, x_3 = 14kg, x_4 = 8kg and x_5 = 11kg.

Then $\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5$

$= 10 + 12 + 14 + 8 + 11$

$= 55$

Notice that we have used i instead of j in the formulae above. The "*j*" is what we call a *dummy variable* - any letter can be used.

Example 5:

Now let us find $\sum_{i=1}^{4} 2x_i$ where x_1 = 2, x_2 = 3, x_3 = -3 and x_4 = 1.

Again, starting with i = 1 we replace the expression 2 x_i with its value and add up the terms until i = 4 is reached.

So $\sum_{i=1}^{4} 2x_i = 2x_1 + 2x_2 + 2x_3 + 2x_4$

$= 2(2) + 2(3) + 2(-3) + 2(1)$

$= 4 + 6 - 6 + 2$

$= 6.$

Example 6:

Let us find $\sum_{k=1}^{3} (x_k - 4)$ where x_1 = 7, x_2 = 4 and x_3 = 1.

Here, $\sum_{k=1}^{3} (x_k - 4) = (x_1 - 4) + (x_2 - 4) + (x_3 - 4)$

$= (7 - 4) + (4 - 4) + (1 - 4)$

$= 3 + 0 + (-3)$

$= 0$

Exercise 2

1) Evaluate $\sum_{i=1}^{4} x_i$ where $x_1 = 5$, $x_2 = 2$, $x_3 = 3$ and $x_4 = 8$.

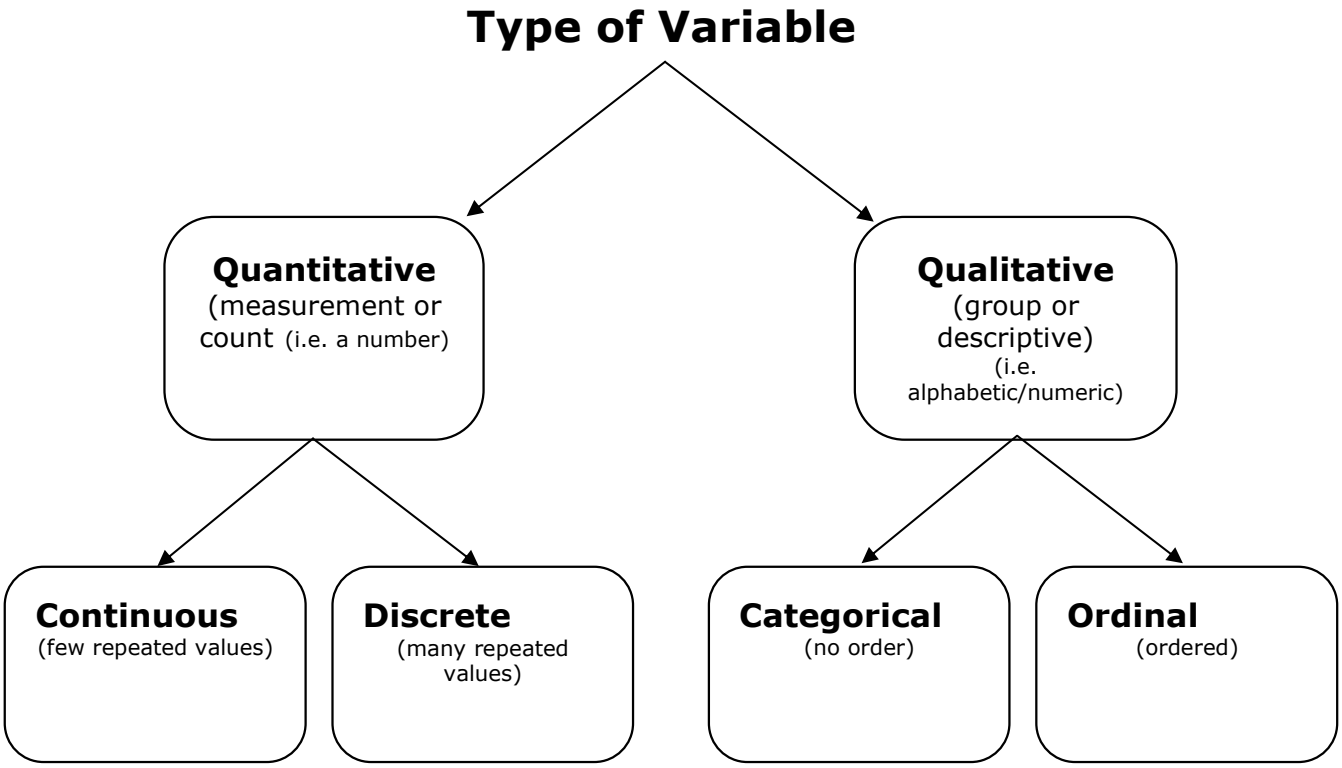
2) Evaluate $\sum_{k=1}^{n} 5x_k$ where $x_1 = 10$, $x_2 = 14$, $x_3 = -2$ and $n = 3$.

3) Find $\overline{X} = \frac{1}{5} \sum_{j=1}^{5} x_j$ where $x_1 = 10$kg, $x_2 = 12$kg, $x_3 = 14$kg, $x_4 = 8$kg and $x_5 = 11$kg. ($\overline{X}$ is the mean weight of the children)

4) Find the value of $\sum_{i=1}^{3} (x_i - \overline{x})^2$ where $x_1 = 105$, $x_2 = 100$, $x_3 = 95$ and $\overline{x} = 100$.

Basic Data Analysis

Our first step is to decide what type of data we have. Once we know this we can then decide what plots to generate and which tests to perform, etc.



Measures of Central Tendency

When given a set of raw data one of the most useful calculations we can make is finding the **centre** of that set of data.

There are three common ways of giving an idea of where the centre of a set of numbers is. They are the mean, the median and the mode.

- **The mean** - $\bar{x}$ (read *x* bar). Also known as the average or expected value. Add up all the numbers and divide by how many numbers there are, i.e. $\sum \frac{x_i}{n}$. Affected by outliers.
- **The median** - the middle number (= Med – also known as the 50th percentile). It is found by putting the numbers in order and taking the actual middle number if there is one, or the average of the two middle numbers if not. Not affected by outliers.
- **The mode** - most frequently occurring number/most common value, useful for qualitative data. Not affected by outliers.

Example 7:

The weight of luggage presented by airline passengers at the check-in (measured to the nearest kg)

18 23 20 21 24 23 20 20 15 19 24

$$\text{Mean} = \frac{18+23+20+21+24+23+20+20+15+19+24}{11}$$

$$= 20.64$$

Median = 20

15 18 19 20 20 20 21 23 23 24 24

↑

middle number

Mode = 20

The number 20 occurs here 3 times.

Here the mean, median, and mode are all appropriate measures of **central tendency**.

Central tendency describes the tendency of the observations to bunch around a particular value, or category. The mean, median and mode are all measures of central tendency. The best one to use in a given situation depends of the type of variable given.

Example 8:

Which measure of central tendency would be used with the following data?

Type of pets owned by a class of twenty students.

Type of Pets	Number of Pets
Cat	6
Dog	5
Goldfish	3
Rabbit	1
Bird	4
None	7
Total	26

The data above is called **qualitative** data (groups) and the average to use with such data is the **mode**. "No pets" would be described as the "modal group", as this is the group that occurs most often. Here it makes no sense to talk of mean or median. If, on the other hand, we were not interested in the type of pet kept, but the **number** of pets owned by students then we might have something like this.

Number of Pets	Tally	Frequency
0	𝍸 ||	7
1	𝍸 |	6
2	|||	3
3	||	2
4	||	2
Total		20

Now we are concerned with a **quantitative** variable and the **average** used most with quantitative variables is the mean. Here, in fact, the mean is 1.3.

$$\text{Mean} = \frac{(0\times 7)+(1\times 6)+(2\times 3)+(3\times 2)+(4\times 2)}{20} = 1.3$$

Note that (1 x 6) is really 1 + 1 + 1 + 1 + 1 + 1, since 6 students have 1 pet each, and (2 x 3) is really 2 + 2 + 2, since 3 students have 2 pets each.

Since there are 20 scores the median will occur between the tenth and eleventh score. The median is 1, since the tenth and the eleventh scores are both 1, and the mode is 0.

The mean has some advantages over the median as a measure of central tendency of quantitative variables. One of them is that when the mean is calculated all the observed values are used. However, to calculate the median, while the observed values are used in the ranking, only the middle or the middle two values are used in the calculation. Another is that it is fairly stable from sample to sample. This means that if we take several samples from the same population their means are less likely to vary than their medians.

However, the median is used as a measure of central tendency if there are a few extreme values observed. These are called outliers and their effect would be to pull the mean too far from the centre of the distribution.

Example 9:

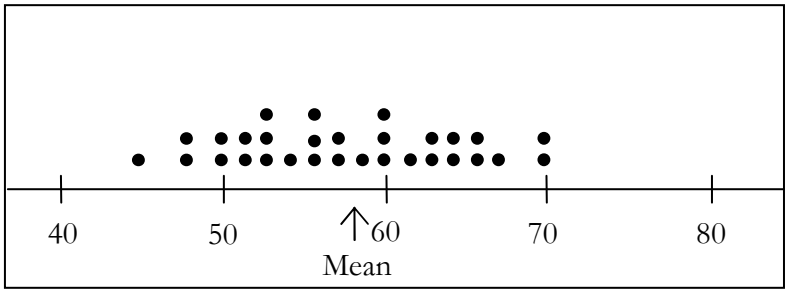
Let's look again at our pets example if one of the students kept 18 goldfish.

Number of Pets	Tally	Frequency
0	𝍸 ||	7
1	𝍸 |	6
2	|||	3
3	|	1
4	||	2
18	|	1
Total		20

The mean is now 2.05, but the mode is still 0 and the median is still 1. The effect of the outlier was to significantly increase the mean and this now means that the median is a more accurate measure of the centre of the distribution.

The mean is the value usually used to indicate the centre of a distribution. We can also think of the mean as the balance point of a distribution.

For example, consider the following distribution of students' marks on a test. Without doing any calculation, we would guess that the balance point of the distribution to be approximately 58. (Think of is as the centre of a seesaw.)



Exercise 3

1) Ten patients at a doctor's surgery wait for the following lengths of times to see their doctor.

5 mins 17 mins 8 mins 2 mins 55 mins

9 mins 22 mins 11 mins 16 mins 5 mins

What are the mean, median and mode? What measure of central tendency would you use here?

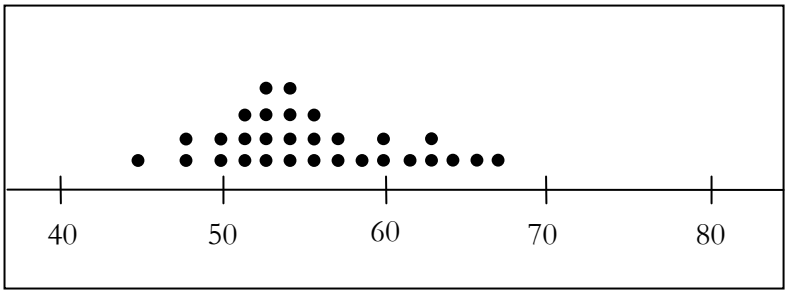
2)

Method of Transport	Number of Students
Walk	5
Car	4
Train	15
Bicycle	10
Motorbike	6
Bus	10
Total	50

What measure of central tendency is the most appropriate here? Find it.

3) Which measure of central tendency is best used to measure the average house price in Auckland?

4) Without doing any calculation, estimate the mean of the following distribution.



Measures of Dispersion

The mean is the value usually used to indicate the centre of a distribution. If we are dealing with **quantitative** variables our description of the data will not be complete without a measure of the extent to which the observed values are spread out from the mean.

We will consider two measure of dispersion and discuss the merits and pitfalls of each.

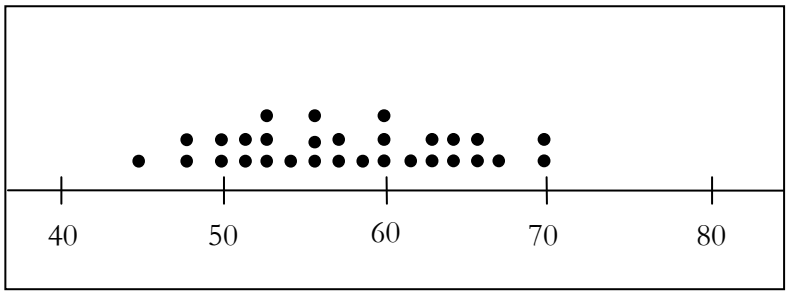
Range

One very simple measure of dispersion is the range. Affected by outliers.

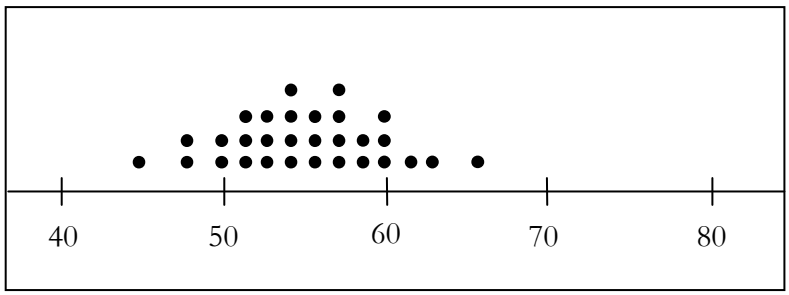
Range = Maximum - Minimum.

Let's consider the following two distributions.

The marks of a group of thirty students on two tests.



Marks on Test A



Marks on Test B

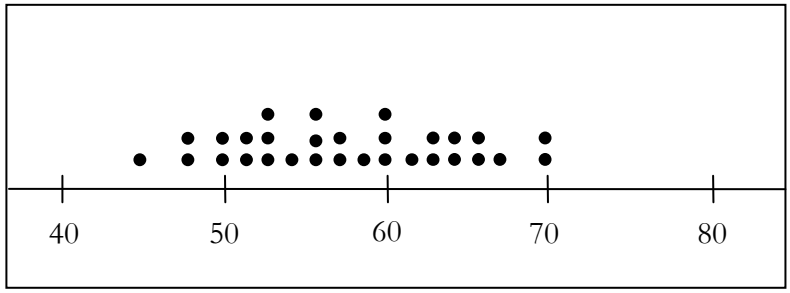
Here it is clear that the marks on test A are more spread out than the marks on test B, and we need a measure of dispersion that will accurately indicate this.

On test A, the range of marks is 70 - 45 = 25

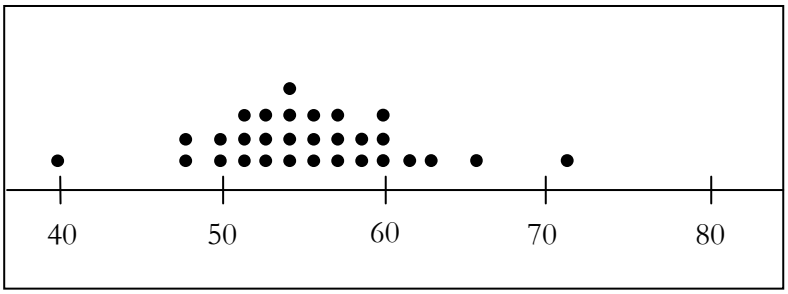
On test B, the range of marks is 65 - 45 = 20

Here the range gives us an accurate picture of the dispersion of the two distributions.

However, as a measure of dispersion the range is severely limited. Since it depends only on two observations, the lowest and the highest, we will get a misleading idea of dispersion if these values are outliers. This is illustrated very well if the students' marks are distributed as follows.



Marks on Test A



Marks on Test B

On test A, the range is still 70 - 45 = 25.

On test B, the range is now 72 - 40 = 32, but apart from the outliers the distribution of marks is clearly less spread out than that of A.

We want a measure of dispersion that will accurately give a measure of the variability of the observations. We will concentrate now on the measure of dispersion usually used, the standard deviation.

Standard Deviation

Suppose we have a set of data where there is no variability in the observed values. Each observation would have the same value and the mean would be that same value. There would be no deviation from the mean. Now suppose that we have a set of observations where there is variability. The observed values **would** deviate from the mean, some by only a little. The standard deviation is a kind of average amount of these deviations form the mean.

Sample Mean & Sample Standard Deviation

Up until this point we have been talking about the mean and standard deviation of a population. These have been written using the Greek letters μ and σ respectively. However, in statistics we usually analyse data from a sample taken from a population, in order to make inferences about what goes on in that population. Our data sets are usually random samples drawn from the population.

The **sample mean** of a sample of size n is written as $\bar{x}$ (read x bar).

We find the sample mean in the same way as the population mean, we add up the sample scores and divide by the number of sample scores.

The **sample standard deviation** of a sample of size n is written as s or s_x or σ_{n-1} or $x\sigma_{n-1}$. It is affected by outliers.

On our calculator, we can get both the population standard deviation (σor σ_n) and the sample standard deviation. We are only ever interested in the sample standard deviation unless we sample the entire population (e.g. do a census).

Example 10:

The ages of seven geography students who went on a field trip were 20, 19, 19, 25, 20, 18, 19.

Using the calculator the mean = 20, standard deviation = 2.3094 (4dp)

Thus the average age of the students on the geography field trip was 20 "give or take" 2.3 years. So, usually, the average age is between 17.7 years and 22.3 years old.

Inter-quartile range: IQR

The inter-quartile range, or IQR for short, is calculated by:

IQR = Upper quartile – Lower quartile

where the lower quartile (Q_1) is the upper boundary of the lower quarter of the data and the upper quartile (Q3) is lower boundary of the upper quarter of the data. The IQR is the middle 50% of data and it is not affected by outliers.

Five number summary

The five number summary is a list of five summary statistics, given in a particular order, in brackets:

(Min, Q1, Med, Q3, Max)

Exercise 4

1) Which of the following lists has the greater standard deviation?

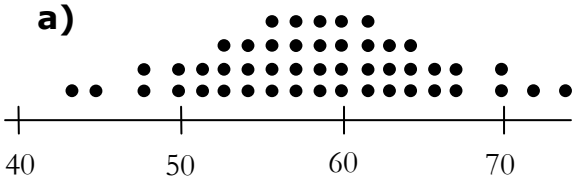
a) 98 99 100 101 102

b) 2 4 6 8 10

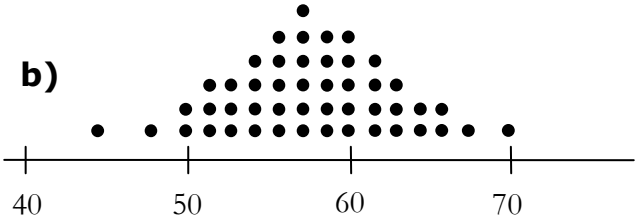
c) 2 10

2) Which of the following distributions has the greater standard deviation?

a)



b)



3) The following are the number of customers a restaurant served for lunch on a sample of ten consecutive days.

46 50 51 60 62 64 72 41 53 55

Find a suitable measure for the central tendency and a suitable measure for the measure of dispersion. Interpret the results.

4) The raw scores that a sample of eight students got on a history test were:

69 84 93 61 79 88 57 67

Find a suitable measure for the central tendency and a suitable measure for the measure of dispersion. Interpret the results.

ANSWERS

Exercise 1

1) $R = L - S$

2) **a)** $T = W_1 + W_2 + \ldots + W_n = \sum_{i=1}^{n} W_i$

b) $\overline{X} = \frac{T}{n} = \frac{\sum_{i=1}^{n} W_i}{n}$

3) $Z = \frac{X - \overline{X}}{S}$

Exercise 2

1) $\sum_{i=1}^{4} x_i = x_1 + x_2 + x_3$

$= 5 + 2 + 3 + 8$

$= 18$

2) $\sum_{k=1}^{n} 5x_k = \sum_{k=1}^{3} 5x_k = 5x_1 + 5x_2 + 5x_3$

$= 5(10) + 5(14) + 5(-2)$

$= 110$

3) $\overline{X} = \frac{1}{5} \sum_{j=1}^{5} x_j = \frac{1}{5}(x_1 + x_2 + x_3 + x_4 + x_5)$

$= \frac{1}{5}(10 + 12 + 14 + 8 + 11)$

$= \frac{1}{5}(55)$

$= 11$

4) $\sum_{i=1}^{3}(x_i - \bar{x})^2 = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2$

$= (105 - 100)^2 + (100 - 100)^2 + (95 - 100)^2$

$= (5)^2 + (0)^2 + (-5)^2$

$= 25 + 0 + 25$

$= 50$

Exercise 3

1) Mean $= \dfrac{5+17+8+2+55+9+22+11+16+5}{10} = 15$

Median = 10

2 5 5 8 9 11 16 17 22 55

↑

middle number

Mode = 5

The median would be the preferred measure of central tendency to use here and not the mean, since there is an outlier of 55 mins. This is making the assumption that the outlier is a freak value and should be disregarded. The mode would not be suitable, because it is just chance that two people waited for the same period of time, and all the others waited for different time periods.

2) The mode is the only possible measure of central tendency to use here, since we are dealing with qualitative data (groups). The modal group is "train".

3) The median is used to indicate average house prices in Auckland. The inclusion of the very expensive houses (those worth millions of dollars) in the calculation of the mean would make the "average" house price too high to be representative of the general market. Nor is the mode suitable because it could happen by chance that a very large number of houses all have the same non-representative value.

4) The actual value of the mean is 56. How close to this value did you get with your guess?

Exercise 4

1) **a)** Standard deviation = 1.5811 (4dp)

b) Standard deviation = 3.1623 (4dp)

c) Standard deviation = 5.6569 (4dp)

Therefore (c) has the biggest standard deviation.

2) a) has the greater standard deviation.

3) Mean = 55.4, Standard deviation = 9.2159 (4dp).

4) Mean = 74.75, Standard deviation = 13.1448 (4dp).

PROBABILITIES

A probability is a number between 0 and 1 that quantifies uncertainty. There are two main sources of probabilities that we will deal with. One source is from a model, the other source is from data.

Some models that may involve equally likely outcomes are: tossing a coin, rolling a die, ...

TERMINOLOGY

- A **random experiment** is an experiment whose outcome cannot be predicted.
- A **sample space** is the collection of all possible outcomes.
- **An event** is a collection of outcomes.
- An event that **occurs** is when any outcome making up that event occurs.

For **equally likely outcomes** and events *A*,

$$\text{pr}(A) = \frac{\text{no. of outcomes in } A}{\text{total no. of outcomes}}$$

Probabilities can refer to the:

- Whole Sample
    - pr (an event)
    - pr (one event and another event)
    - pr (one event or another event)
- Part of the Sample
    - pr (one event | another event) - conditional probability

Probability rules are as follows:

- $0 \leq \text{pr}(X = x) \leq 1$
- $\Sigma \text{ pr}(X = x) = 1$

Example 1

A person tosses a coin twice. What is the sample space?

S = {HH, HT, TH, TT}.

What is the probability that a person gets a head on the first toss?

$$\text{pr (head on first toss)} = \frac{2}{4}$$

What is the probability that a person gets a head?

$$\text{pr (Head)} = \frac{3}{4}$$

What is the probability that a person gets 2 tails?

pr (2 Tails) = 1 / 4

Exercise 1:

A person rolls a six-sided die.

1) What is the sample space?

2) What is the probability that the person gets a 6?

3) What is the probability that the person gets an even number?

4) What is the probability that the person gets a number less than 3?

Example 2

The following table gives the ages of the parents of nuptial living births, which were registered in 1980. In constructing the table multiple births have been taken into consideration (Source: N.Z. Official Yearbook, 1982).

		Age of Father			
		Under 25	25 - 34	35 and over	Total
Age of Mother	Under 25	6,620	7,436	311	14,367
	25 - 34	759	18,679	3,446	22,884
	35 and over	10	435	1,578	2,023
	Total	7,389	26,550	5,335	39,274

A birth is chosen at random from the 39,274 births. What is the probability that the baby chosen has:

(a) a mother who is less than 25 years old.

$$\frac{14,367}{39,274} = 0.3658 \text{ } (4dp)$$

(b) a father who is in the 25 - 34 age group.

$$\frac{26,550}{39,274} = 0.6760 \text{ } (4dp)$$

(c) a mother **and** a father who are both less than 25 years old.

$$\frac{6,620}{39,274} = 0.1686 \text{ } (4dp)$$

(d) a mother **or** a father who is less than 25 years old.

$$\frac{7,389 + 14,367 - 6,620}{39,274} = 0.3854 \text{ } (4dp)$$

(e) a father who is more than 25 years old.

$$\frac{26,550 + 5,335}{39,274} = 0.8119 \text{ } (4dp)$$

(f) a mother who is less than 25 years old, **given** that the father is less than 25 years old.

$$\frac{6,620}{7,389} = 0.8959 \text{ } (4dp)$$

(g) a father who is more than 35 years old, **given** that the mother is less than 25 years old.

$$\frac{311}{14,367} = 0.0216 \text{ } (4dp)$$

Example 3

The *Listener* (February, 1995) contained an article, which looked at changing attitudes towards abortion of people 15 years and over. It compared the results of two surveys, one conducted in 1985 and the other in 1994. Below is a table based on the 1985 data, giving the responses to the following question: "Do you approve or disapprove of abortion when the mother's health is at risk?".

		Age (years)				
		15 - 24	25 - 39	40 - 54	55 +	Total
Response	Approve	327	575	375	394	1671
	Disapprove	29	41	24	22	116
	Don't know	51	22	18	12	103
	No response	8	5	3	7	23
	Total	415	643	420	435	1913

A person is chosen at random from these 1913 people. What is the probability that the person selected:

(a) is in the age range 25 - 39 years?

$$\frac{643}{1913} = 0.3361 \text{ } (4dp)$$

(b) disapproves of abortion when the mother's health is at risk?

$$\frac{116}{1913} = 0.0606 \text{ } (4dp)$$

(c) is in the age range 15 - 24 years **and** approves of abortion when the mother's health is at risk?

$$\frac{327}{1913} = 0.1709 \text{ } (4dp)$$

(d) is more than 40 years old?

$$\frac{420 + 435}{1913} = 0.4469 \text{ } (4dp)$$

(e) is in the age range 55+ **or** disapproves of abortion when the mother's health is at risk?

$$\frac{435 + 116 - 22}{1913} = 0.2765 \text{ } (4dp)$$

(f) is in the age range 25 - 39 years, **given** that they approve of abortion when the mother's health is at risk?

$$\frac{575}{1671} = 0.3441 \text{ } (4dp)$$

(g) approves of abortion when the mother's health is at risk, **given** that they are in the age group 15 - 24 years.

$$\frac{327}{415} = 0.7880 \text{ } (4dp)$$

Exercise 2:

Time, 17 October 1994, reported on a sex survey in America conducted by a Chicago National Opinion Research Centre team. A team of highly trained interviewers interviewed and questioned 3452 subjects. The results of the question "How many sexual partners have you had since you were 18?" are show in the table below.

		Number of Sexual Partners						
		None	1	2 - 4	5 - 10	11 - 20	21+	Totals
Gender	Women	51	549	616	342	103	51	1712
	Men	52	348	365	401	278	296	1740
	Totals	103	897	981	743	381	347	3452

If one of the people in the survey is chosen at random:

(a) what is the probability that he or she has had more than 4 sexual partners since age 18?

(b) what is the probability that the person is a woman?

(c) what is the probability that the person is a woman who has had 5 - 10 sexual partners since age 18?

(d) what is the probability that the person is a man who has had more than 10 sexual partners since age 18?

(e) what is the probability that the person is a woman who has had no more than 1 sexual partner since age 18?

(f) what is the probability that the person is a man, given that they have had more than 20 sexual partners since age 18?

(g) what is the probability that the person has had 2 - 4 sexual partners since age 18, given that the person is a woman?

Exercise 3:

In the U.S. and in Europe, the presence of air bags in an automobile has become a key factor in deciding whether to purchase a particular model of automobile. A random sample of 93 automobiles were cross-classified by their size and the level of air bag installation. Below is the cross-classification of the two variables, TYPE-OF-CAR and AIR-BAGS, for the 93 automobiles.

		AIR-BAGS			
		None in the car	Driver's side only	Driver's and passenger's side	Row Totals
TYPE OF CAR	Small	16	5	0	21
	Sporty	3	8	3	14
	Compact	5	9	2	16
	Mid-size	4	11	7	22
	Large	0	7	4	11
	Van	6	3	0	9
	Column Totals	34	43	16	93

If a car was chosen at random, what is the probability that:

(a) the car is small **and** has no air bags installed?

(b) there are air bags installed?

(c) the car is sporty **or** has no air bags installed?

(d) there are no air bags installed, **given** that the car is small?

(e) the car is mid-size, **given** that there are air bags installed on the driver's side only?

(f) the car is large, **given** that there are air bags installed?

ANSWERS

Exercise 1

1) $S = \{1, 2, 3, 4, 5, 6\}$

2) *pr (getting a 6)* $= \frac{1}{6} = 0.1667$ (4dp)

3) *pr (getting an even number)* $= \frac{3}{6} = \frac{1}{2} = 0.5$

4) *pr (getting a number less than 3)* $= \frac{2}{6} = \frac{1}{3} = 0.3333$ *(4dp)*

Exercise 2

(a) $\frac{743 + 381 + 347}{3452} = 0.4261$ *(4dp)*

(b) $\frac{1712}{3452} = 0.4959$ *(4dp)*

(c) $\frac{342}{3452} = 0.0991$ *(4dp)*

(d) $\frac{278 + 296}{3452} = 0.1663$ *(4dp)*

(e) $\frac{51 + 549}{3452} = 0.1738$ *(4dp)*

(f) $\frac{296}{347} = 0.8530$ *(4dp)*

(g) $\frac{616}{1712} = 0.3598$ *(4dp)*

Exercise 3

(a) $\frac{16}{93} = 0.1720$ *(4dp)*

(b) $\frac{43 + 16}{93} = 0.6344$ *(4dp)*

(c) $\frac{34 + 14 - 3}{93} = 0.4839$ *(4dp)*

(d) $\frac{16}{21} = 0.7619$ *(4dp)*

(e) $\frac{11}{43} = 0.2558$ *(4dp)*

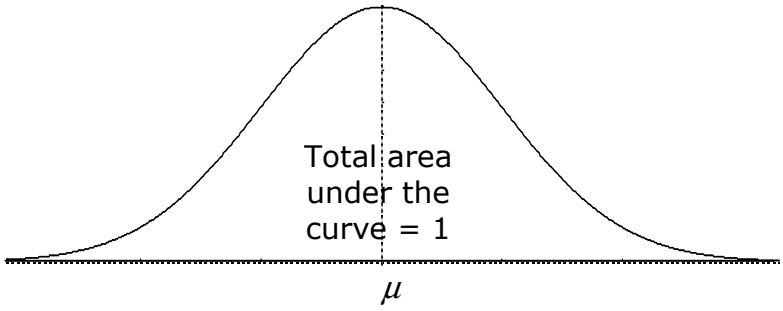
(f) $\frac{7 + 4}{43 + 16} = 0.1864$ *(4dp)*

RANDOM VARIABLES

A **random variable** is a type of measurement taken on the outcome of an experiment.

NORMAL DISTRIBUTION

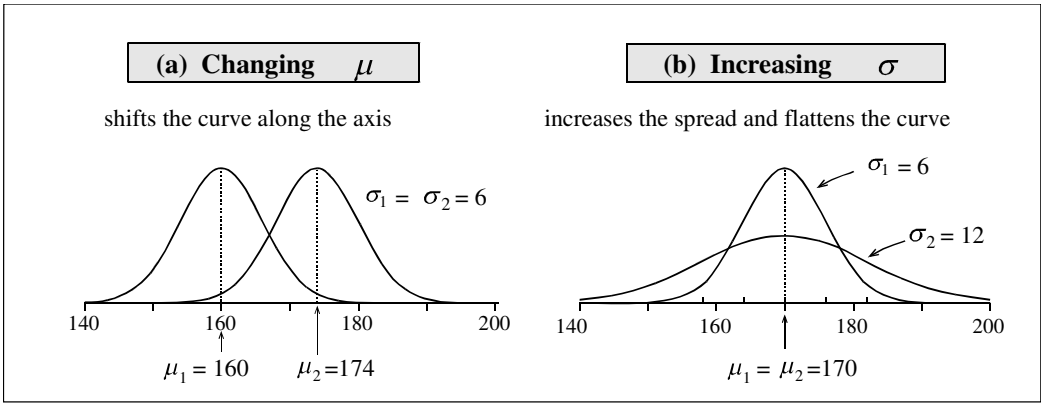
The Normal Distribution comes from a continuous random variable. It has a probability density function curve which is smooth, **bell-shaped**, and symmetric.



The Normal distribution is important because it:

- fits a lot of data particularly well
- can be used to approximate other distributions
- is very important in statistical inference

The shape of the curve is solely determined by the parameters of the Normal distribution, the mean μ, and the standard deviation σ.



Areas beneath the curve represent probabilities.

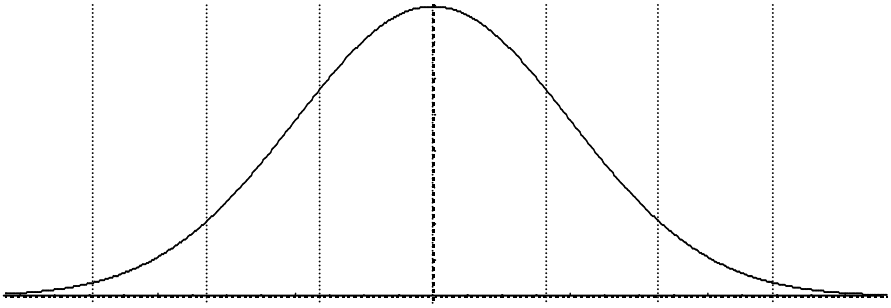
- The total area under the curve is 1.
- $\Pr(X \leq x)$ = the area below the point x.
- $\Pr(X \geq x)$ = the area above the point x.

A random observation from a Normal Distribution has approximately:

68% of observations are within 1σ of μ.

95% of observations are within 2σ of μ.

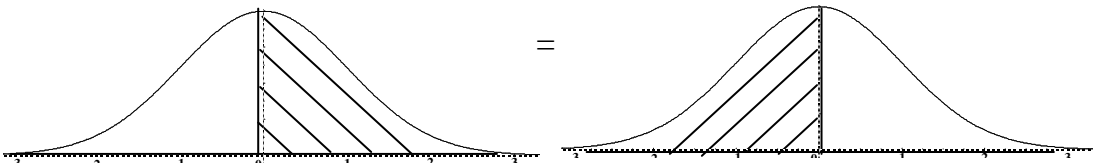
99.7% of observations are within 3σ of μ.



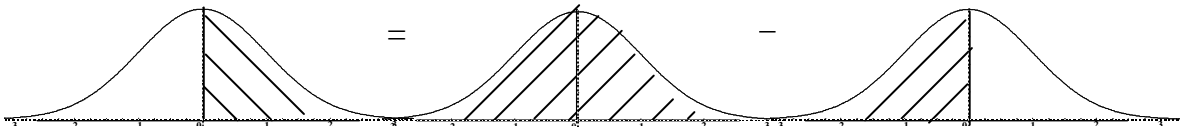
2 key concepts

1. area under the curve = probability
2. numbers on axis = data values

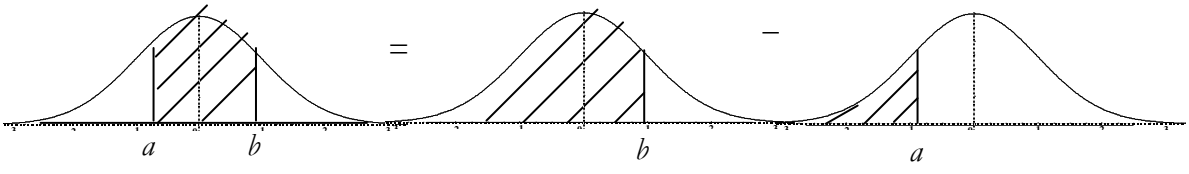
$\Pr(X > \mu) = \Pr(X < \mu) = 0.5$



$\Pr(X \geq x) = 1 - \Pr(X \leq x)$



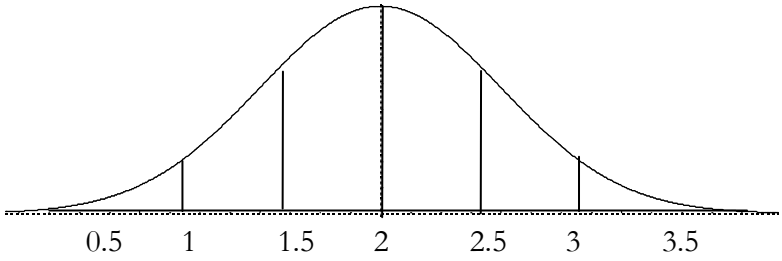
$\Pr(a \leq X \leq b) = \Pr(X \leq b) - \Pr(X \leq a)$



Example 1

Data was collected on the time taken to prepare personal income tax returns (including interview time and typing time) and was found to follow a Normal distribution with a mean of 2 hours and a standard deviation of 0.5 hours.

Approximately 95% of the personal income tax returns took between $2 - 2 \times 0.5 = 1$ hours and $2 + 2 \times 0.5 = 3$ hours.



Exercise 1

1) Work out the values corresponding to 1, 2 and 3 standard deviations above and below the mean for each of the following distributions. Mark your results under a diagram of a Normal curve.

   (a) The price of a suit of a certain type has mean = \$300 and sd = \$50.

   (b) The number of complaints per week at a large department store has a mean of 55 and an sd of 8.

   (c) The length of a television commercial has a mean of 1 minute and sd of 1.5 seconds

2) 3000 accounts receivable at a large appliance store are found to follow a Normal distribution with a mean of \$350 and a standard deviation of \$100.

   (a) What percentage of accounts is greater than \$350?

   (b) What percentage of accounts is less than \$150?

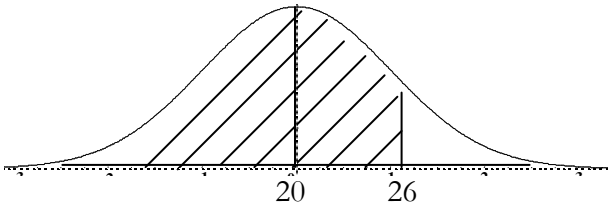
   (c) How many of these accounts would be between \$250 and \$550?

   (d) Almost all of accounts lie between \$_____ and \$_____.

   (e) The manager of the store wants to send reminders to the highest 2.5% of the accounts. This represents all accounts above \$_____.

Example 2

The total number of hours per week lost due to sickness in a particular small business was found to be Normally distributed with a mean of 20 hours and a standard deviation of 5 hours. Sketch a Normal graph relating to the probability that the number of hours lost due to sickness in a week does not exceed 26 hours. Write this probability in symbols.



Probability in symbols = $\Pr(X \leq 26)$

Exercise 2

1) The price of a suit of a certain type of material follows a Normal distribution with a mean of \$300 and a standard deviation of \$50.

   (a) Sketch a Normal graph relating to the probability that the suit costs less than \$200. Write this probability in symbols.

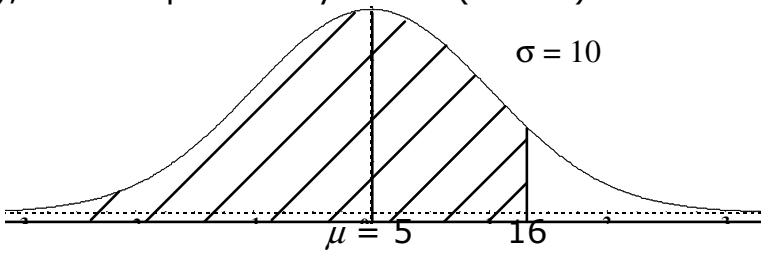
   (b) Sketch a Normal graph relating to the probability that the suit costs less than \$350. Write this probability in symbols.

Calculating Normal Probabilities using given Computer Output

Normal probabilities are calculated either with the SPSS or Excel computer programs or a graphics calculator. Following is an example with exercises of the type of computer output you that can expect to see in the STATS 10x course.

Example 3

If $X \sim$ Normal ($\mu = 5$, $\sigma = 10$), find the probability that Pr($X \leq 16$).



Step 1: $x = 16$

Step 2: $\mu = 5$

Step 3: Pr($X \leq 16$)

Step 4: Reading off the table: Pr($X \leq 16$) = 0.864334

= 0.8643 (4 dp)

OUTPUT FROM EXCEL

x	Pr($X \leq x$)	x	Pr($X \leq x$)	x	Pr($X \leq x$)
0.0	0.308538	7.5	0.598706	16.0	0.864334
1.5	0.363169	9.0	0.655422	17.5	0.894350
3.0	0.420740	11.5	0.742154	19.0	0.919243
4.5	0.480061	13.0	0.788145	21.5	0.950529
6.0	0.539828	14.5	0.828944	23.0	0.964070

OUTPUT FROM COMPUTER

```
Cumulative Distribution Function
Normal with mean = 5.00000 and standard deviation = 10.0000

x          P( X <= x)
0.0000     0.3085
1.5000     0.3632
3.0000     0.4207
4.5000     0.4801
6.0000     0.5398
7.5000     0.5987
9.0000     0.6554
11.5000    0.7422
13.0000    0.7881
14.5000    0.8289
16.0000    0.8643
17.5000    0.8944
19.0000    0.9192
21.5000    0.9505
23.0000    0.9641
```

Exercise 3

For the Normal distribution given in Example 3:

1) Find Pr($X \leq 3$)
2) Find Pr($X \leq 11.5$)
3) Find Pr($X \leq 19$)
4) Find Pr($X \geq 4.5$)
5) Find Pr($X \geq 13$)
6) Find Pr($X \geq 23$)
7) Find Pr($1.5 \leq X \leq 7.5$)
8) Find Pr($9 \leq X \leq 14.5$)
9) Find Pr($17.5 \leq X \leq 21.5$)

Exercise 4

1) An insurance company found that claims under its travel insurance for lost luggage was approximately Normally distributed with a mean of \$820 and a standard deviation of \$170.

x	$\Pr(X \leq x)$	x	$\Pr(X \leq x)$	x	$\Pr(X \leq x)$
200	0.000133	800	0.453174	1000	0.855160
400	0.006745	825	0.511732	1100	0.950227
600	0.097812	860	0.593010	1250	0.994287
750	0.340256	890	0.659744	1500	0.999968
780	0.406990	950	0.777777	1650	0.999999

(a) What is the probability of a claim being less than \$1000?
(b) What is the probability of a claim being more than \$1250?
(c) What is the probability of a claim being more than \$800?
(d) What is the probability of a claim being more than \$400?
(e) What is the probability of a claim being less than \$200?
(f) What is the probability of a claim being less than \$750?
(g) What is the probability of a claim being less than \$1650?
(h) What is the probability of a claim being between \$890 and \$1100?
(i) What is the probability of a claim being between \$750 and \$1000?

ANSWERS

Exercise 1

(1) (a) 1σ above the mean = $300 + 1 \times 50 = \$350$

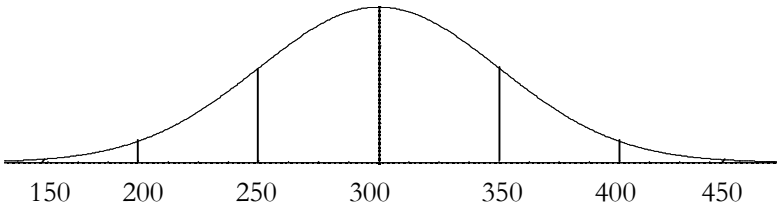
2σ above the mean = $300 + 2 \times 50 = \$400$

3σ above the mean = $300 + 3 \times 50 = \$450$

1σ below the mean = $300 - 1 \times 50 = \$250$

2σ below the mean = $300 - 2 \times 50 = \$200$

3σ below the mean = $300 - 3 \times 50 = \$150$



(b) 1σ above the mean = $55 + 1 \times 8 = 63$ complaints

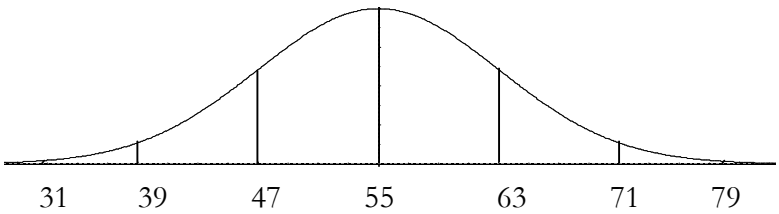
2σ above the mean = $55 + 2 \times 8 = 71$ complaints

3σ above the mean = $55 + 3 \times 8 = 79$ complaints

1σ below the mean = $55 - 1 \times 8 = 47$ complaints

2σ below the mean = $55 - 2 \times 8 = 39$ complaints

3σ below the mean = $55 - 3 \times 8 = 31$ complaints



(c) 1σ above the mean = 60 + 1 × 1.5 = 61.5 seconds

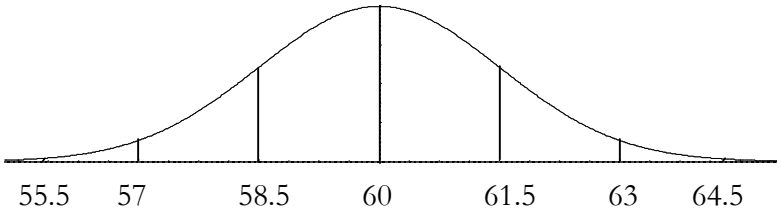
2σ above the mean = 60 + 2 × 1.5 = 63 seconds

3σ above the mean = 60 + 3 × 1.5 = 64.5 seconds

1σ below the mean = 60 – 1 × 1.5 = 58.5 seconds

2σ below the mean = 60 – 2 × 1.5 = 57 seconds

3σ below the mean = 60 – 3 × 1.5 = 55.5 seconds



(2) **(a)** 50% of accounts are above $350.

(b) 2.5% of accounts are below $150.

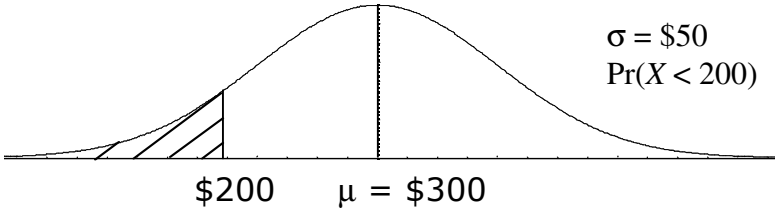
(c) 81.5% of accounts are between $250 and $550. This represents 3000 × .815 = 2445 accounts.

(d) $50 and $650. (3 standard deviations from the mean.)

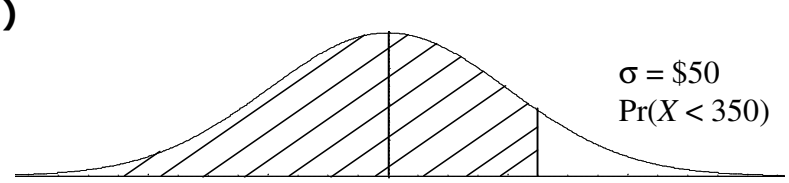
(e) $550 – 2 standard deviations above the mean.

Exercise 2

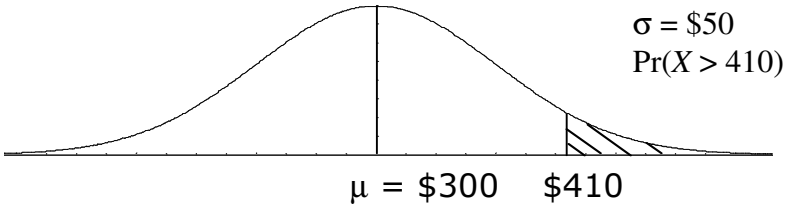
(1) (a)



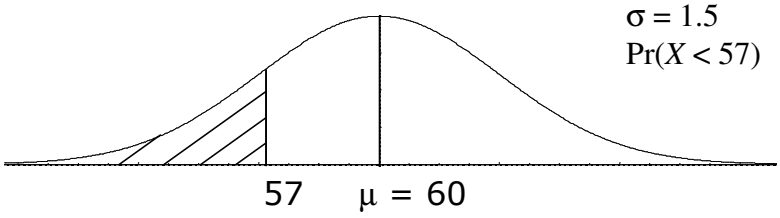
(b)



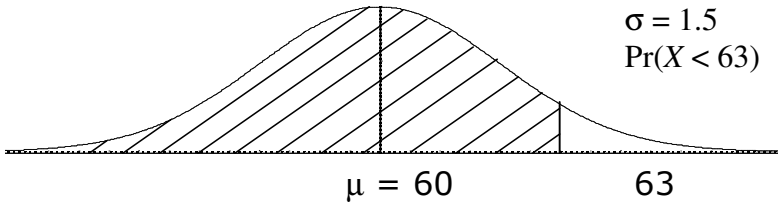
(c)


σ = $50
Pr(X > 410)
μ = $300
$410

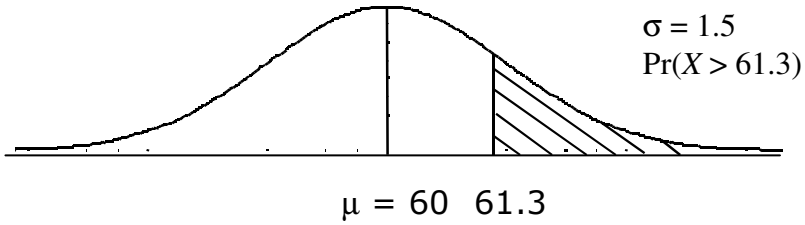
(2) (a)


σ = 1.5
Pr(X < 57)
57
μ = 60

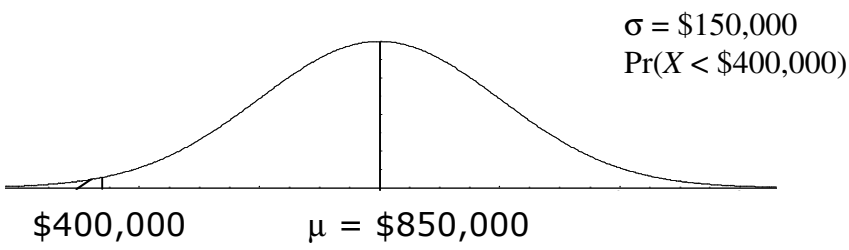
(b)


σ = 1.5
Pr(X < 63)
μ = 60
63

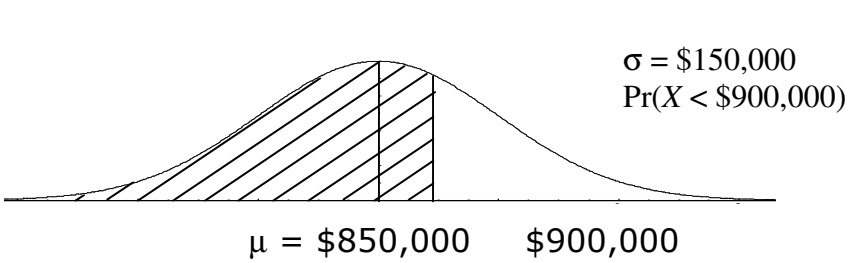
(c)


σ = 1.5
Pr(X > 61.3)
μ = 60
61.3

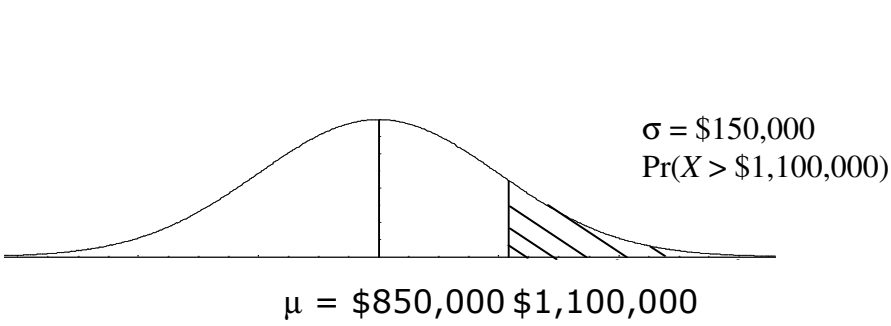
(3) (a)



(b)



(c)



Exercise 3

1) $\Pr(X \leq 3) = 0.4207$ (4dp)
2) $\Pr(X \leq 11.5) = 0.7422$ (4dp)
3) $\Pr(X \leq 19) = 0.9192$ (4dp)
4) $\Pr(X \geq 4.5) = 1 - \Pr(X \leq 4.5) = 1 - 0.4801 = 0.5199$ (4dp)
5) $\Pr(X \geq 13) = 1 - \Pr(X \leq 13) = 1 - 0.7881 = 0.2119$(4dp)
6) $\Pr(X \geq 23) = 1 - \Pr(X \leq 23) = 1 - 0.9641 = 0.0359$ (4dp)
7) $\Pr(1.5 \leq X \leq 7.5) = \Pr(X \leq 7.5) - \Pr(X \leq 1.5)$
   $= 0.5987 - 0.3632 = 0.2355$ (4dp)
8) $\Pr(9 \leq X \leq 14.5) = \Pr(X \leq 14.5) - \Pr(X \leq 9)$
   $= 0.8289 - 0.6554 = 0.1735$ (4dp)
9) $\Pr(17.5 \leq X \leq 21.5) = \Pr(X \leq 17.5) - \Pr(X \leq 21.5)$
   $= 0.9505 - 0.8944 = 0.0561$ (4dp)

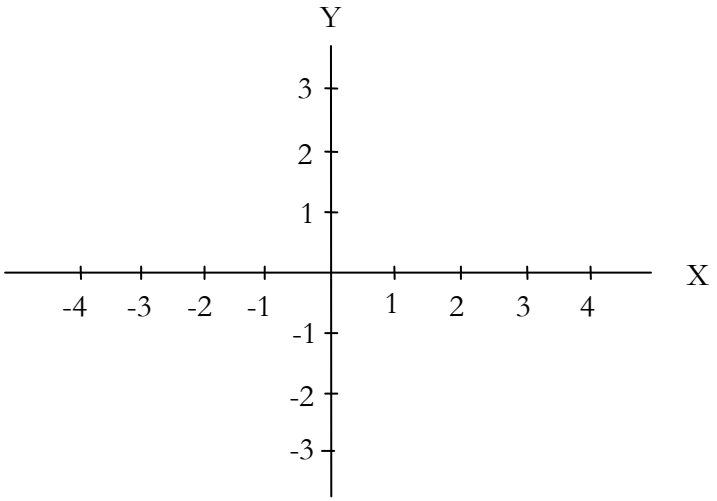
Exercise 4

1) $\Pr(X \leq 1000) = 0.8552$ (4dp)
2) $\Pr(X \geq 1250) = 1 - \Pr(X \leq 1250) = 1 - 0.9943 = 0.0057$ (4dp)
3) $\Pr(X \geq 800) = 1 - \Pr(X \leq 800) = 1 - 0.4532 = 0.5468$ (4dp)
4) $\Pr(X \geq 400) = 1 - \Pr(X \leq 400) = 1 - 0.0067 = 0.9933$ (4dp)
5) $\Pr(X \leq 200) = 0.0001$ (4dp)
6) $\Pr(X \leq 750) = 0.3403$ (4dp)
7) $\Pr(X \leq 1650) = 0.999999$ [or 1.0000 (4dp)]
8) $\Pr(890 \leq X \leq 1100) = \Pr(X \leq 1100) - \Pr(X \leq 890)$
   $= 0.9502 - 0.6597 = 0.2905$ (4dp)
9) $\Pr(750 \leq X \leq 1000) = \Pr(X \leq 1000) - \Pr(X \leq 750)$
   $= 0.8552 - 0.3403 = 0.5149$ (4dp)

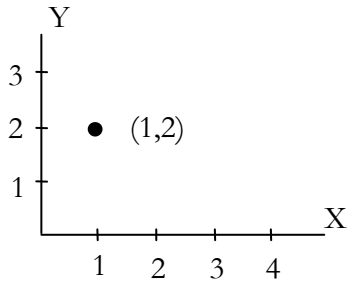
LINEAR EQUATIONS

Plotting Points

We represent points on a pair of axes. The figure below shows a horizontal axis, the *x*-axis, and a vertical axis, the *y*-axis.

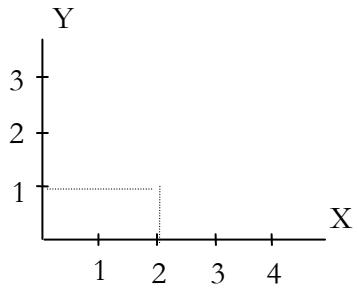


The point represented below is (1,2); that is, it has an *x* co-ordinate of 1 and a *y* co-ordinate of 2.



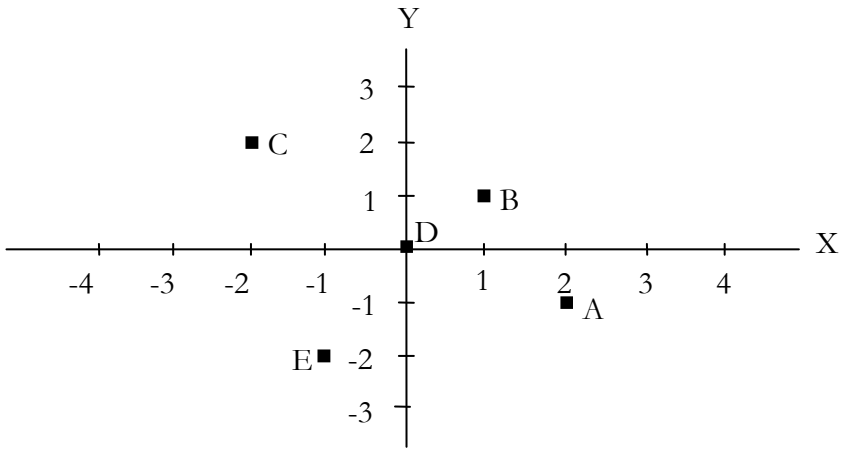
Now try plotting the point (2,1), i.e. $x = 2$, $y = 1$, on a pair of axes.

To do this we go 2 units along the *x*-axis and draw an imaginary vertical line. Then we go one unit up the *y*-axis and draw an (imaginary) horizontal line. The intersection of the two lines is the required point.



Exercise 1:

Write down the co-ordinates of the labelled points.



A =

B =

C =

D =

E =

Lines

A straight line is described as an equation of the form $y = \beta_0 + \beta_1 x$ where β_0 and β_1 are constants, that is, fixed values. Every equation of this form can be graphed as a straight line.

Exercise 2:

Which of the following equations are linear (describe straight lines)?

a) $y = -1 + 2x$

b) $y = x^2 + 1$

c) $y = \sqrt{3}x + 1$

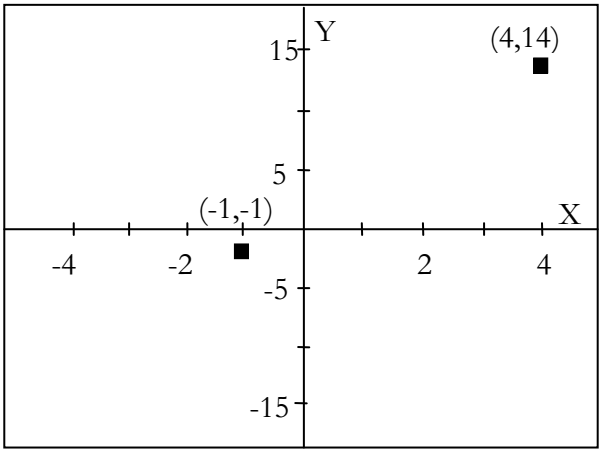
d) $y = 3\sqrt{x} + 1$

Consider the equation $y = 3x + 2$ ➔ $y = 2 + 3x$.

To sketch the graph of this line we need only two points on the line.

Choose any two values of x, say $x = -1$ and $x = 4$, then find the corresponding values of y:

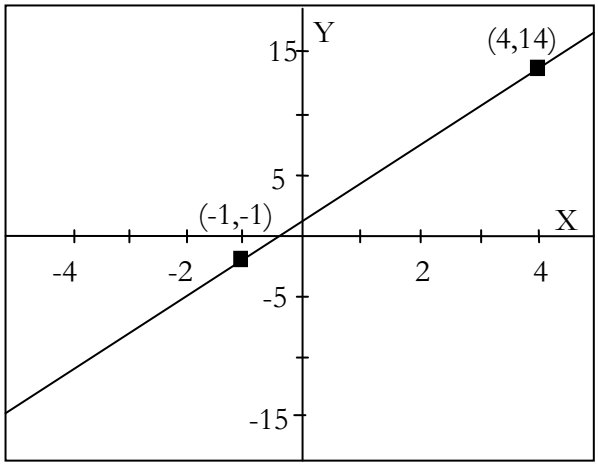
1. $x = -1$, $y = 2 + 3(-1) = -1$
2. $x = 4$, $y = 2 + 3(4) = 14$



Clearly if we join these points we will have a straight line.

Can you see why the two points are sufficient to describe a unique line?

Draw the line joining the two points.



Now choose any other value of x, and find the corresponding value of y by substituting for x in the equation.

Plot this point and check that it lies on the line.

Two useful attributes of the line are its:

1. "steepness" or **slope**. The **slope** or **gradient** of the line means how steep it is. Imagine yourself standing on a point on the line and walking

"up" it. The steeper it is the more you will have to walk up for the same amount along.

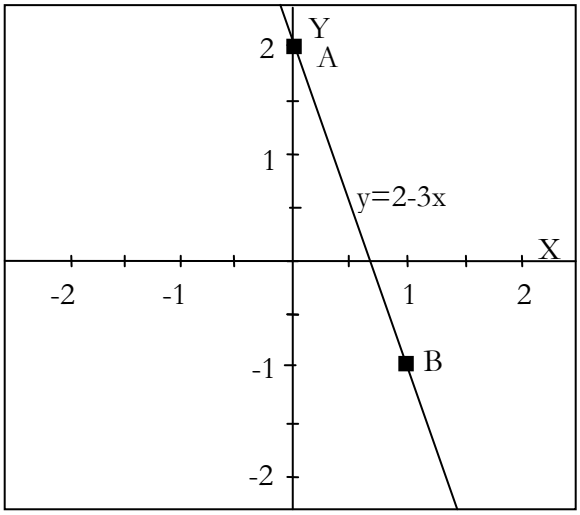
2. "position" or where it cuts the *y*-axis. This is called the **intercept** on the *y*-axis.

Thus we can define slope as the ratio $\frac{amount\ uphill}{amount\ along}$ from any point to any other on the line, which is to the right of the first point. Another definition is: for every one-unit increase in *x*, the slope is the amount that *y* changes by.

For example, suppose you are at the point (-1,-1) on the previous line $y = 3x + 2$. To get to the point (4,14) you have to walk 15 units up for 5 along, therefore the slope is $\frac{15}{5} = 3$.

Notice that if the equation of the line is $y = \beta_0 + \beta_1 x$, the "β_1" tells us the slope.

Now graph the line $y = 2 - 3x$ by choosing any two points on it. Then find the slope of the line

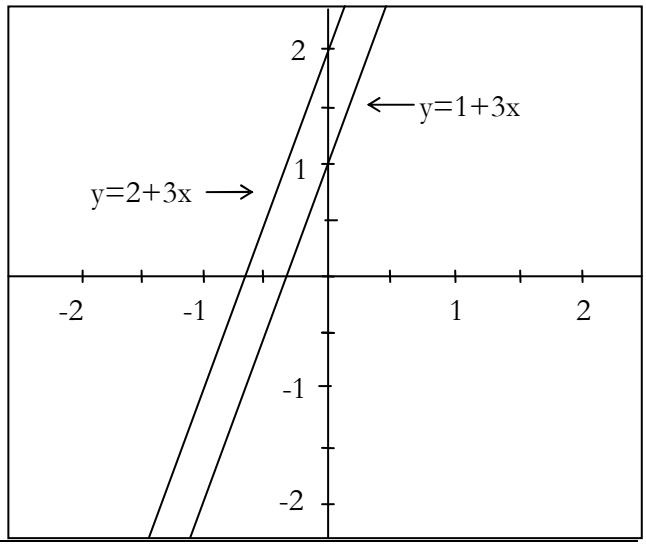


The point A is (0,2) and the point B is (1,-1). To get from A to B you would have walked 3 units downhill and one along to the right.

Hence the slopes is $\frac{-3}{1} = -3$, adopting the convention that "downhill" is negative.

The position of the line or intercept on the *y*-axis is given by "β_0", because when:

1. **$x = 0$,**
2. **$y = \beta_1 (0) + \beta_0$**
3. **$y = 0 + \beta_0$**
4. **$y = \beta_0$**



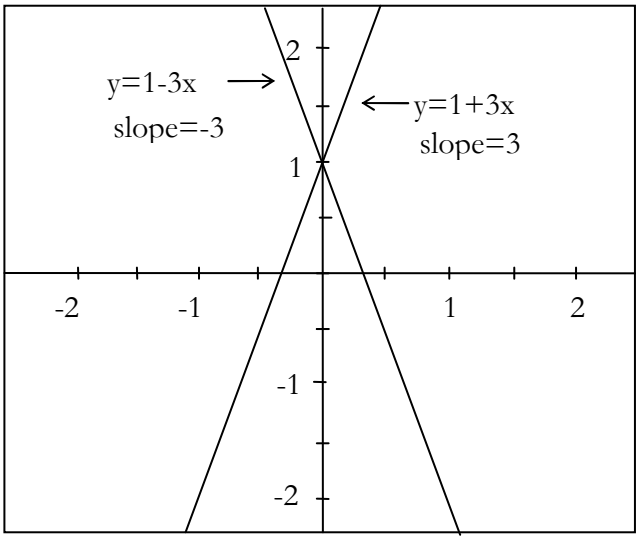
This means that the line $y = 1 + 3x$ will have the same slope but a different intercept on the *y*-axis to $y = 2 + 3x$.

We check this by sketching the line $y = 1 + 3x$ on the same axes as $y = 2 + 3x$.

Example 1

What is the slope of $y = 1 - 3x$?
In what way will this line differ from that of $y = 1 + 3x$?

The slope of $y = 1 - 3x$ is -3, while the slope of $y = 1 + 3x$ is +3.
Therefore the former has a downward slope while the latter has an upward slope. But the both have the same intercept on the y-axis, namely +1.



Example 2

Find the equation of a line, which has y-intercept of 7 and gradient -3.

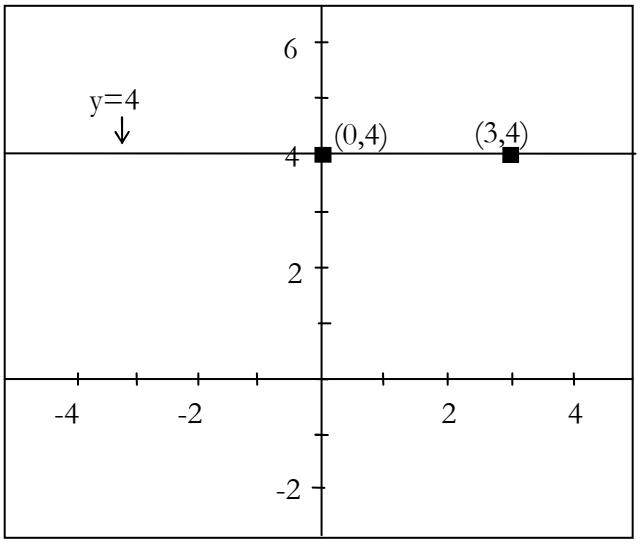
$y = \beta_0 + \beta_1 x$ where $\beta_0 = 7$ and $\beta_1 = -3$.
So the equation of the line is $y = 7 + -3x$.

Example 3

$y = 4$ is the equation of a straight line. Let us find its slope.

Choose any two x values, say $x = 0$ and $x = 3$. For both of these, $y = 4$ gives a point on the line.

Join the points (0,4) and (3,4). We have a horizontal line. Hence the slope of this line is zero - no amount up for three units along.

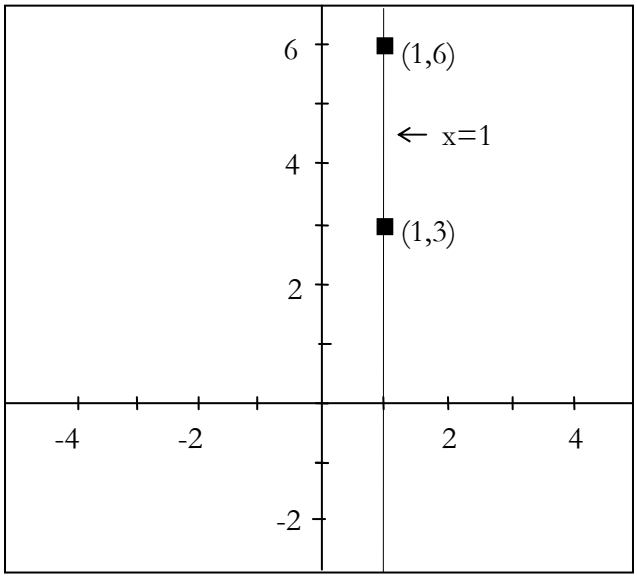


Example 4

Find the slope of the line through (1,2) and (1,6)

Clearly we have to go 3 units up, but no distance at all along. According to our rule, the slope should be $\frac{3}{0}$. But division by 0 is not possible, so we say the slope is undefined.

A line which is parallel to the vertical axis has equation $x = \beta_0$, for some β_0. In this case the equation is $x = 1$ as x is always 1, no matter what the value of y is.



Exercise 3:

1) Find the equations of the following lines:

   a) the gradient is -4 and the y-intercept is 2

   b) the line has the same slope as (is parallel to) $y = 3x - 2$ and has y-intercept 5

   c) the line is horizontal and passes through (1,2)

   d) the line is vertical and passes through (-1,3)

2) Match up the following:

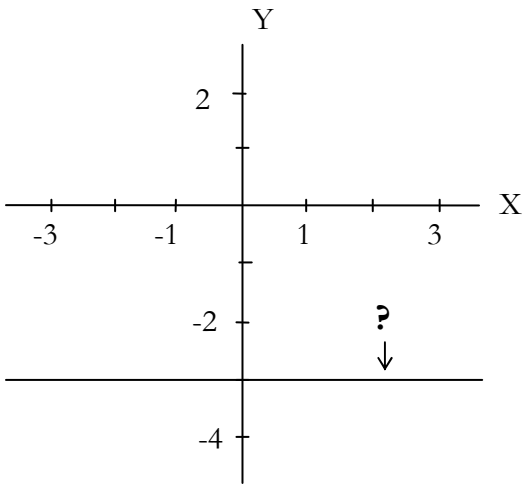
   The line $y = 3x + 7$

a)	is parallel to the line	i)	$2y = 14 + 6x$
b)	has the same y-intercept but different gradient to the line	ii)	$y = 3x - 7$
c)	has gradient of opposite sign to the line	iii)	$y = 7 + 5x$
d)	is identical to the line	iv)	$y = -7 - 3x$

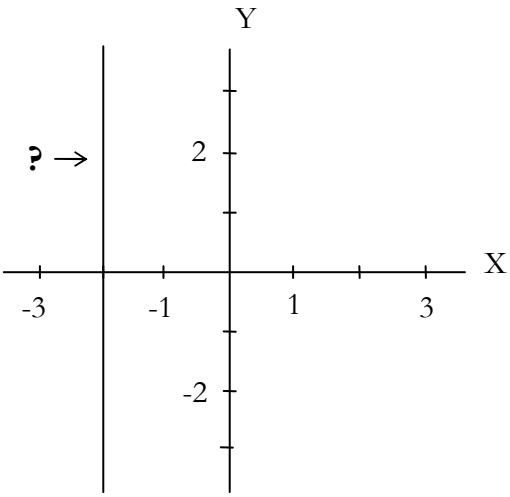
3) Use your knowledge of gradient and intercept to sketch the graphs of $y = x$ and $y = 2x$.

4 What are the equations of the following lines?

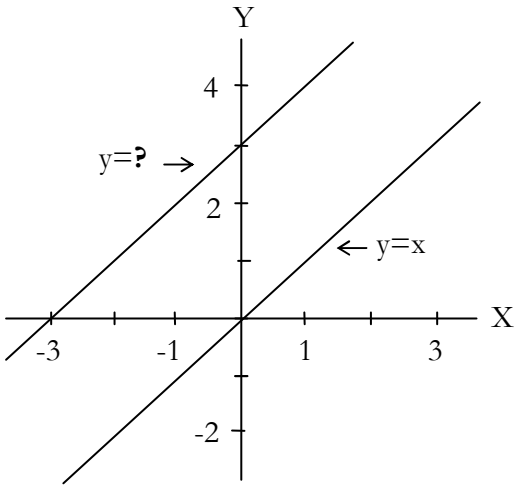
a)



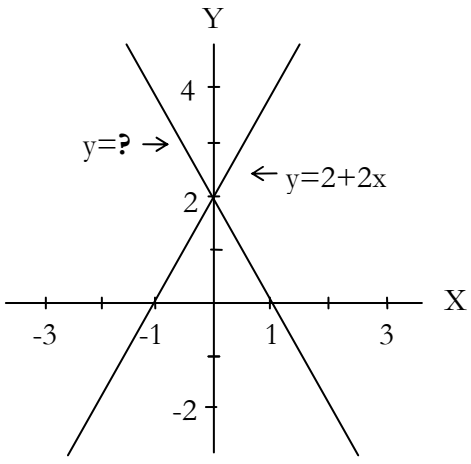
b)



c)

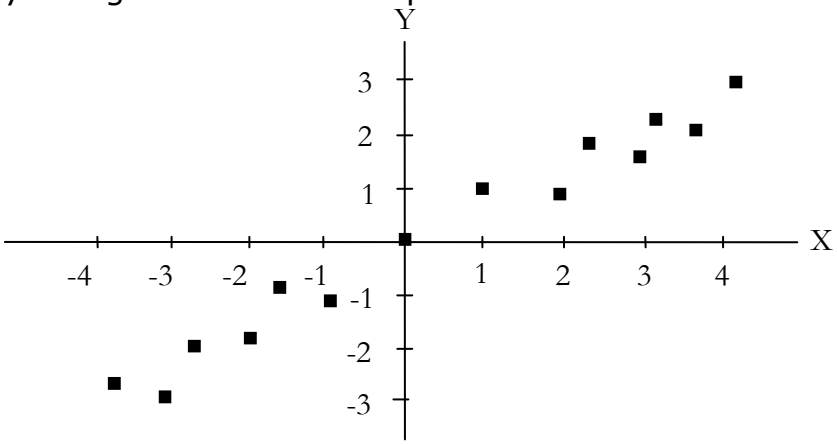


d)



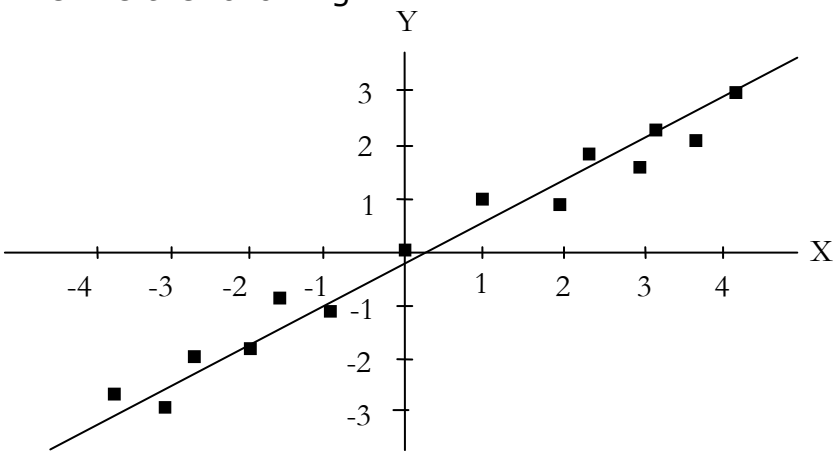
Regression

What do you do when you have collected data and plotted it, and it doesn't all fall on a perfectly straight line? For example:



In regression we use a method called "least squares". It fits the best line to the data while minimising the sum of the squared distances from the line.

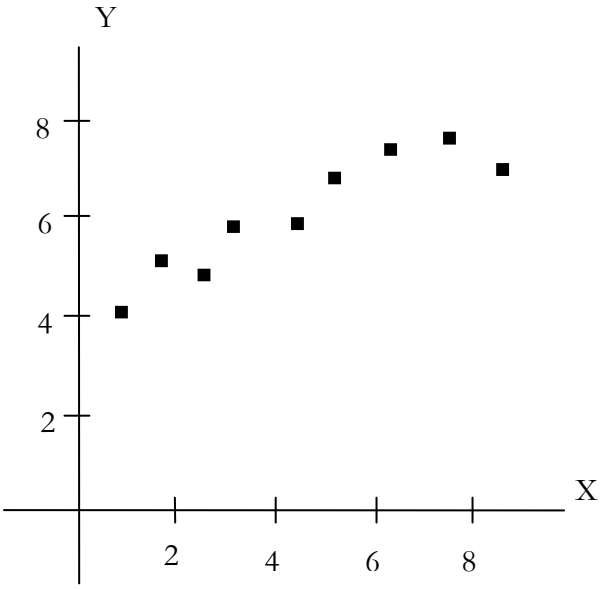
It will suggest a line like the following:



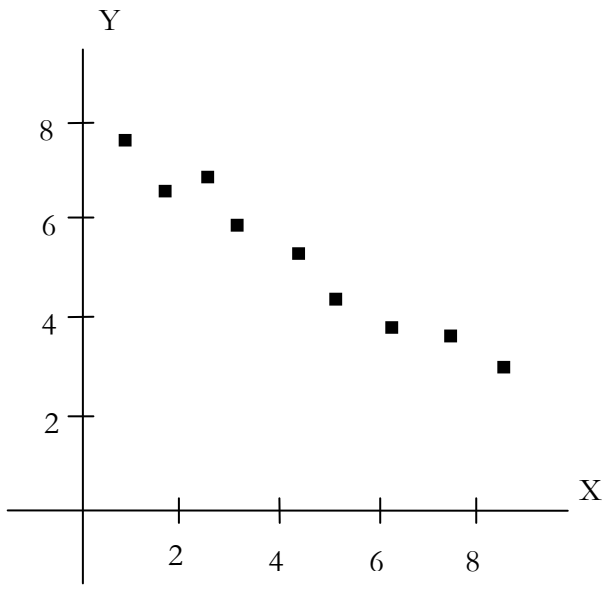
Exercise 4:

1) By hand fit a reasonable line of best fit through the following graphs.

a)

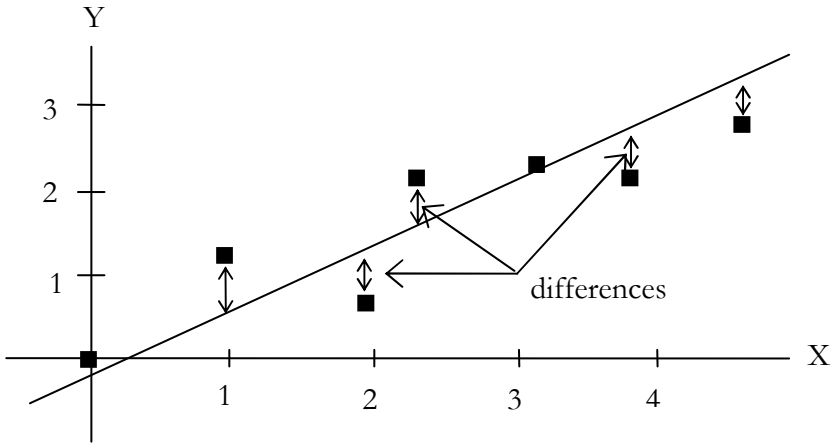


b)



There are two main components of a regression relationship: trend and scatter. Trend is the overall average "best fit" line. Scatter is the distance between the individual data value and the trend line.

For example:



The best-fit line represents the trend, and the differences represent the scatter.

ANSWERS

Exercise 1

A = (2, -1), B= (1, 1), C = (-2, 2), D = (0, 0), E = (-1, -2)

Exercise 2

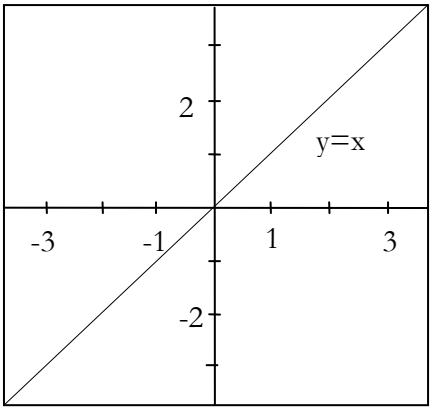
a) Yes
b) No, x^2 instead of x
c) Yes
d) No, $\sqrt{x}$ instead of x

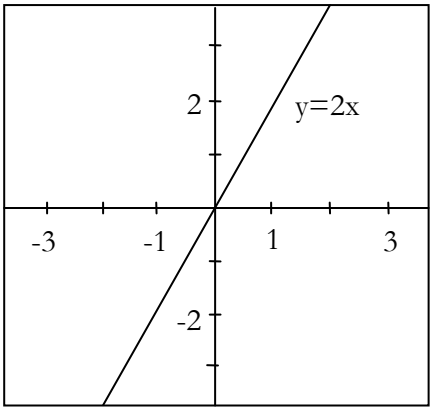
Exercise 3

1) a) $y = -4x + 2$
   b) $y = 3x + 5$
   c) $y = 2$
   d) $x = -1$

2) a) - ii) as they have the same slope (+3).
   b) - iii) as the have the same intercept (+7).
   c) - iv) as the slope has the opposite sign (-3).
   d) - i) as they are multiples of each other.

3)



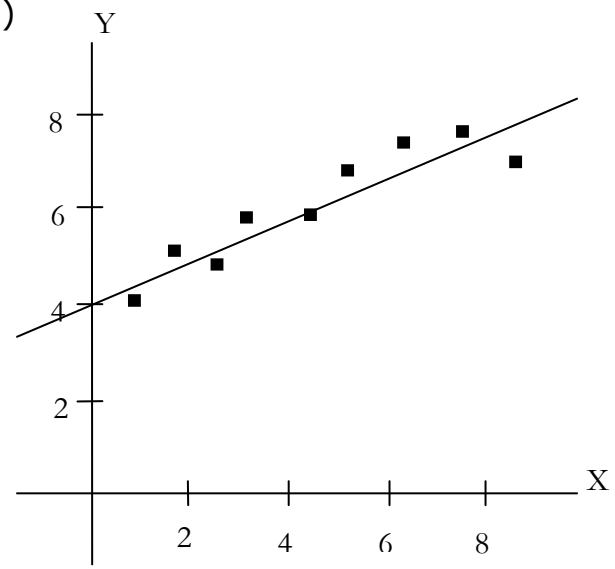


4) a) $y = -3$
   b) $x = -2$
   c) $y = 3 + x$
   d) $y = 2 - 2x$

Exercise 4

1)

a)



b)