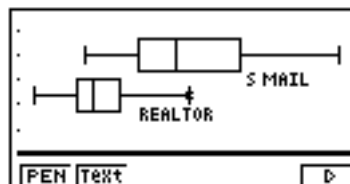


Making decisions through the analysis of data

Instalment Two – Data measured on an interval scale (plus)



A product of the Noel Baker Centre for School Mathematics.
WIP (Work in progress)

*LUMAT-NSW (2003) is the initiative of the
Noel Baker Centre for School Mathematics, Prince Alfred College and
CASIO AUSTRALIA.*



*Noel Baker Centre
for School Mathematics*

Contents

<u>Stenduser – What kids do and like.</u>	page 3
<u>Introduction</u>	page 5
Investigation 1 : Pulse Rates	page 5
Populations and Samples : concept	page 6
<u>Types of Variables : Nominal; Ordinal; Interval</u>	page 8
SI Exercise 1a	page 11
<u>The Tools of our Trade</u>	page 13
Stemplots	page 14
SI Exercise 1b	page 16
Using the Stemplot : outliers; shape; centre; spread	page 18
shape	page 19
inference using statistics - GOSCSC	page 20
SI Exercise 1c	page 22
Histograms	page 24
SI Exercise 1d	page 27
<u>CASIO 9850GB PLUS : Histograms</u>	page 28
SI Exercise 1e	page 31
<u>Measures of Centre</u>	page 32
Median	page 32
Mean	page 33
Mode	page 34
Resistance	page 34
SI Exercise 1f	page 38
<u>Measures of Spread</u>	page 39
Quartiles	page 39
Five number summary	page 41
Box and whisker plots	page 41
SI Exercise 1g	page 44
CASIO 9850GB PLUS and boxplots	page 45
Interpreting boxplots	page 46
SI Exercise 1h	page 48
Putting it all together	page 49
SI Exercise 1i	page 53
<u>The Project : Student's own investigation</u>	page 54

Stenduser: What kids do and like.

In March 2003, over 21000 Australian students (mainly SA students with a few from WA and NT) from Years 8 to Adult re-entry took part in a survey containing 34 questions designed by a group of students as a part of the *SeniorSchoolCensus-online* project.

In 2004, many thousands (data still be collected at time of writing) of Australian students from Years 8 to Adult re-entry from all states and territories took part in a survey containing 40 questions as a part of the *SeniorSchoolCensus-online* project.

The questions fell into three categories:

- About the students school
- About the students home and
- About the student

The result of each stage of this project can be considered as a **population** that consists of thousands of actual school students from Years 8 - Adult re-entry.

Are you curious about what the students had to say? Do you wonder if you have similar ideas and feelings as others?

It is rare to have all data associated with a population. It is often far too expensive, or simply impossible to collect data from all members of a population (the ABS attempts to do this in their Census). Statisticians usually resort to gathering data from a **sample** of the population and use the results from the analysis of the data to predict what is happening in the population. There is much to learn about how to do this properly. Many times, inadequate or inappropriate analysis takes place and hence any conclusions that made are of questionable value.

In the name of good learning, the responses from the populations are being withheld from the world, *but* you can sample from each population in an appropriate manner in order to predict the goings on of each population using our *sampler*. The sampler will choose a *simple random sample* (SRS) from the population. The sampler also allows you to select SRS's, it is highly likely that each one will be different to the last.

Challenge One

Go to the *SeniorSchoolCensus-online* project website: <http://www.censusonline.net> and look through the questions to which the students responded in each year (2003 and 2004). Choose five of the questions that are of most interest to you **from one of the years**.

Challenge Two

Go to the *SeniorSchoolCensus-online* project website: <http://www.censusonline.net> and navigate your way to the *sampler* for the year of your choice.

Leave all of the characteristics tagged as '**All**' and select a SRS of 20 individuals.

Determine the mean height of the 20 individuals in your sample.

Do you think the sample mean would be close to the mean of the population (ie. the thousands of heights in our database)?

Challenge Three

Return to the sampler and leave all of the characteristics tagged as '**All**' and select another SRS of 20 individuals. Determine the mean height of the 20 individuals in your sample.

Compare the sample mean you determined in Challenge Two to those of your class members. Describe the variation of the sample means determined by you and your class members.

Challenge Four

Repeat the procedure from Challenge 3 but with samples of 100 and then 255.

Describe the variation of the sample means collected by your class members.

Challenge Five

Investigate the physical activity, homework and computer game playing habits of the population. You may choose to investigate the habit of girls, or compare Year 8 girls to Year 12 girls or some other combination.

Be sure to use your statistical skills to construct an argument that clearly supports any hypotheses you make.

INTRODUCTION

Statistics is essentially the science of **collecting, analysing and interpreting the analysis of information in order to** investigate a problem that has arisen, a question that has been asked or simply to investigate a situation of interest.

In this unit you will learn the basic skills required to take some information that has been collected in an appropriate manner, analyse it and then use the results of this analysis to build an argument that supports a conjecture.

Information is often called *data*. In the study of Statistics we use the word *data* for the information we collect.



Decisions through Data - Unit 1 'What is Statistics ?' (12 minutes).



INVESTIGATION ONE: PULSE RATES

If we sit still for a reasonable time and do not exert ourselves in any way, and then measure our pulse rate, we are said to have measured our **resting pulse rate**.

If we then carry out some form of exercise for a given time and measure our pulse rate it should be very different. This we will call our **exercise pulse rate**.

Both of these pulse rates can be an indicator of a persons fitness level. Another quantity that is an indicator of fitness is the **time taken** for a persons pulse rate to return to their resting pulse rate after exercise. This is hard to ascertain unless a person is hooked up to a 'techo' machine. A reasonable measure of this can be found by measuring pulse rate at some nominated time after the exercise has finished.

THE QUESTION: Do year 12 students or year 9 students generally **have more impressive** pulse rates?



THOUGHTS

Start a new page in your problem book - title it 'STATISTICAL INVESTIGATIONS'.

T1.1 Do you think that the pulse rates of year 12's will be different, in general, than year 9's. Why?

T1.2 In order to be able to answer this question::

How much information may you need to gather?

How would you analyse it?

T1.3 When measuring a pulse rate, would it matter if you used a 15 second time interval or a 1 minute time interval to arrive at the 'number of beats per minue'?

INVESTIGATING THE PULSE RATES A LITTLE FURTHER

Consider the following hypothesis made by a seasoned teacher.

Hypothesis: Year 12 students will have generally lower resting pulse rates, lower exercise pulse rates and, five minutes after completing exercise will have returned closer to their resting pulse rate than year 9 students.

Carry out the following experiment to collect some data that will help you to test this hypothesis.



EXPERIMENT

1. Sit and rest for five minutes.
2. Measure your pulse rate (beats per minute) and record this value in your workbook.
3. Do 20 steps up using a chair to step up on. Have a partner hold the chair.
4. Measure your pulse rate immediately after you finish the step ups. . Record this in your workbook.
5. Now rest for exactly 5 minutes and measure your pulse rate again. Record this in your workbook.
6. Record each of your classmates data in your workbook

Twenty five year 12 students carried out the same steps as given above. Your teacher will now supply you with their data.

7. Analyse the data you have and draw a conclusion about the accuracy of the hypothesis stated above.

POPULATIONS AND SAMPLES

Be aware that you have not collected data from all of year 12 and year 8 students that exist. Hence you have not collected data from what is called the *population* of year 8 and year 12 students You have just a *sample* of the data that could be collected. Hence you will be **unable to prove or disprove the hypothesis**. At best you will be able to give support to the opinion that the hypothesis is right or wrong.

You should be wondering if you can give valid support to the opinion that the hypothesis is right or wrong with the number of pieces of data you have.

Whether the support will be valid mainly depends on two things:

- how you collected the data and
- how much data you have collected.

Your work in the Stenduser should help you to begin to understand why the size of the sample is important.



RESEARCH

R1.1 Go to the library, surf the internet or perform any other form of research to discover how it is we should have selected the students from which we collected the data on pulse rate.

WHAT TYPE OF PROBLEMS CAN WE SOLVE USING STATISTICS?

The type of problems that can be solved using statistics fall into two broad categories .

A : Problems that involve **MEASURABLE QUANTITIES** that VARY due to HUMAN MANIPULATION or due to CHANCE (ie. due to nature or the physical structure of the situation). By **MEASURABLE** we mean measured using a scale of equal numeric units.

eg. the length of heart operations,
the petrol consumption of a car,
the value of people's phone bills.

B : problems that involve **CHARACTERISTICS**, the options for which may fall into several **CATEGORIES (or classifications)** .

eg. the marital status of a person (married, single, divorced etc.)
the type of clothing a person generally wears,
the type of worker a person is (blue collar, white collar, no collar!)
the football team a person supports.

The MEASURABLE QUANTITIES and CHARACTERISTICS are called **VARIABLES**. They are called **VARIABLES** because there are sometimes many, but always at least a few, different **RESPONSES** for the **quantity** or **characteristic**.

The **RESPONSES** are traditionally called **LEVELS**.

Hence a variable is said to have a **NUMBER OF LEVELS**.

A variable is therefore any measurable or categorical characteristic that changes for one reason or another .

TYPES OF VARIABLES

We therefore have two main types of variables , **A : MEASURED**
B : CATEGORICAL

Examples of such variables are :

A : MEASURED: height , age , time , weight , fat content , pressure .

B : CATEGORICAL: sex , age , name , breed/ race , brand .

Note that some variables can be considered as both measured and categorical - how can this be the case?

The key is to consider the possible responses or levels of the variable.

AGE for example could have levels of 15 years, 22 years, 39 years, 89 years etc.. Many levels exist in this case. Clearly in this case we would consider AGE as a *measured variable*.

Alternatively AGE could have levels of YOUNG, MIDDLE-AGED or OLD. Only three levels exist in this case. Clearly in this case we would consider AGE as a *categorical variable*.

As we proceed you will discover that it is critical to first *define the variable(s)* that you wish to investigate, and equally as important, to decide what TYPE of variable(s) it is.

The type of variable is best determined by considering the levels that are possible for this variable.

FINE TUNING THE TYPES OF VARIABLES.

Consider the following problems that could be investigated using Statistics.

Is a machine producing bolts of an acceptable diameter , with respect to size and cost ?

What size bolts should be kept in stock in large quantities in Harries Hardware shop?

In each of these problems the variables to be investigated, DIAMETER OF BOLT, BOLT SIZE are both measured variables. The method of analysis for each of these problems is different. Hence there is a need to refine further our variable types. Our aim is to be able to classify each variable as a certain type so that a standard method of analysis can be prescribed for that type of variable.

It should be clear that all variables fall into TWO BROAD CATEGORIES namely:

MEASURED VARIABLES and CATEGORICAL VARIABLES .

eg.

VARIABLE	POSSIBLE LEVELS	TYPE
weight of humans	56 kg, 132.4 kg, etc <i>(infinite levels exist)</i>	measured
number of sixes rolled on one die in four rolls	0, 1, 2, 3 or 4 <i>(finite number of levels)</i>	measured
ability level	below average, average, above average <i>(note the ORDER here)</i>	categorical
eye colour	dark, light <i>(note there is no order here)</i>	categorical

Note that even though weight and number of sixes rolled on one die in four rolls are both measured they are intrinsically different. One has an *infinite number of levels and a large range* while the other has a *finite number of levels and a small range*. Data from each variable requires different analysis.

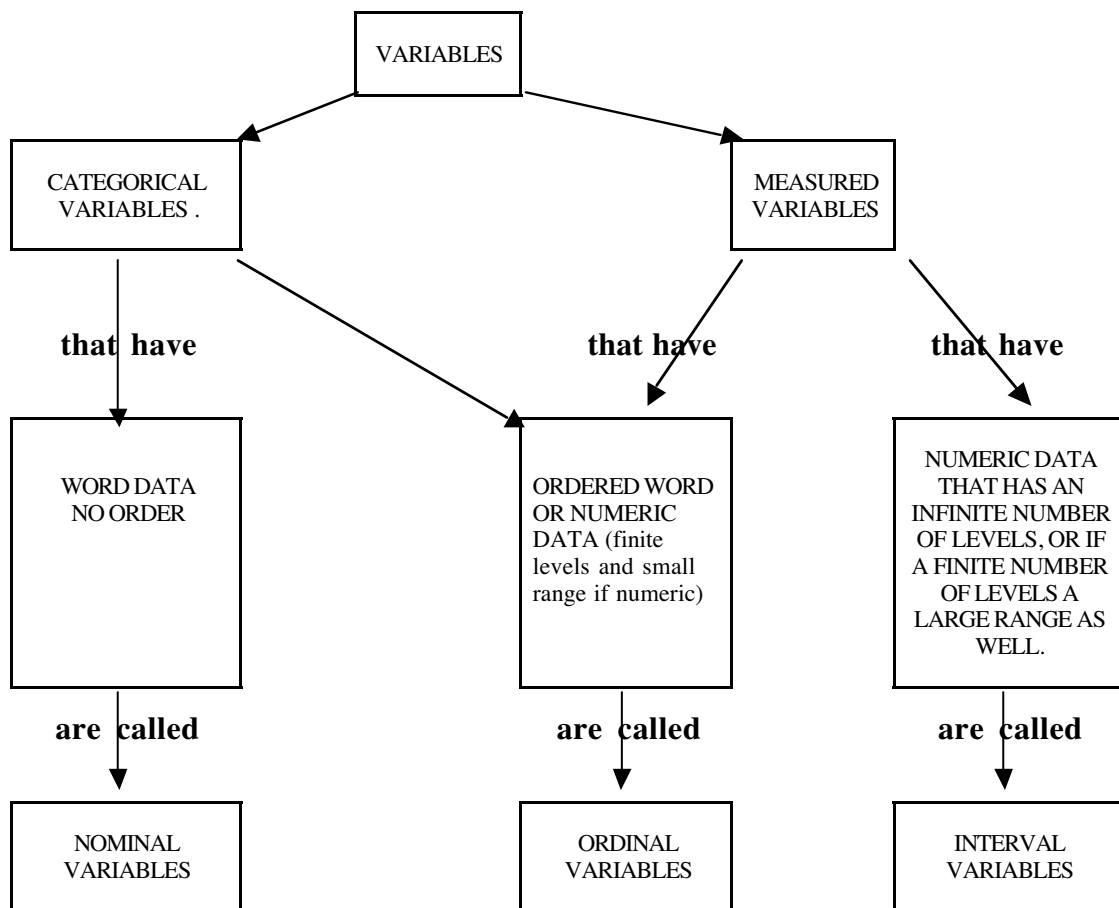
Similarly, ability level and eye colour are both categorical but intrinsically different. The ability level responses has an ORDER whereas the eye colour responses are unordered. Data from each variable requires different analysis.

Hence a finer categorisation of variables is required.

All variables can be categorised into three different types:

- NOMINAL: word data that has no order.
- ORDINAL: word or numeric data that has an ordered nature. If numeric, ordinal data has finite levels and a small range.
- INTERVAL: numeric data that has an infinite number of levels or if a finite number of levels a large range as well.

The following flow chart will help you to establish whether a variable is nominal, ordinal or interval.



Examples:

GENDER male/female, categorical word data, no order which implies **NOMINAL**

HEIGHT (measured in cm)
 numeric data, infinite levels which implies **INTERVAL**

QUALITY (rated as poor/average/good/excellent)
 word data, with order which implies **ORDINAL**.

**SI Exercise 1a**

- 1) For each of the following variables give
 - i) a list of possible levels (ie. possible responses), if there are only a few, or state that there are many or infinitely many levels.
 - ii) measurement units (where appropriate)
 - iii) the variable type (as nominal, ordinal or interval) . More than one solution may be possible!

eg the variable height could have many levels from 30 cm to 210 cm . It could therefore be considered an INTERVAL VARIABLE , but it could also have levels of 'short, medium and tall', which means it could be considered an ORDINAL VARIABLE.

- a) season
 - b) distance between cities
 - c) animal type
 - d) country of birth
 - e) temperature
 - f) weight
 - g) pulse rate
 - h) fan speed
 - i) age
 - j) brand
 - k) the difference between the two values of the upper most face of two fair dice that are rolled.
- 2) State 5 variables, different to those listed above , possible levels, measurement units where appropriate and the variable type .
 - 3) Various studies have attempted to rank cities in terms of how desirable they are in which to live and work . State five variables that you would collect data on if you were to design the study . State measurement units where appropriate and the variable type .

- 4) In America the number of deaths from cancer has steadily increased over time . In 1985 about 462000 people died which was an increase from the 1970 total of 331000. The medical fraternity claim they have made progress in saving people once cancer is diagnosed . A member of parliament is bewildered at this claim as the numbers show that an increase has occurred .
- a) Explain how the **number of deaths** can increase even if progress has genuinely been made (ie. less people are dying ??) .
 - b) State at least one variable that would be more appropriate to measure the effectiveness of the medical treatment for cancer . State this variable's type .
- 5) A women's magazine carried out a survey that attempted to guage what dishwasher was most reliable out of two brands . All they published was the following table and stated that BRAND A was the most reliable .
- a) Are they correct? State an appropriate variable that could determine if their claim is correct .
 - b) Do you feel that the magazine has acted responsibly in only publishing this data? Explain your point of view.

BRAND	NO. OF OWNERS	NO. OF SERVICE CALLS IN THE PAST YEAR
A	13 376	2942
B	480	192

AN INDEPTH LOOK AT THE TOOLS OF OUR TRADE .

In this unit we are going to explore the analysis of problems involving INTERVAL DATA.

From the introduction to this unit you should be aware that we are aiming to test claims and / or solve problems and hence make decisions by using statistics . This course is not an exception to the *old toolbox analogy* . We need to have a selection of tools at hand in order to be able to solve the problem. We not only need the tools in the box but must be fully aware of how to use them .The following section is designed to look in depth and separately at each of the tools you are required to use when solving problems involving interval data.

DISTRIBUTIONS

If we are supplied with a set of interval data the very first thing that we should be interested in is how the data **VARIES**, or to put another way, how the data is **DISTRIBUTED** .

We do this because understanding how the data changes may give clues to solving the problem .

DEFN : The pattern of variation of a set of data is called its DISTRIBUTION . The distribution shows

- 1) the numerical values of the variable (data)and
- 2) how often each value occurs.(frequency)

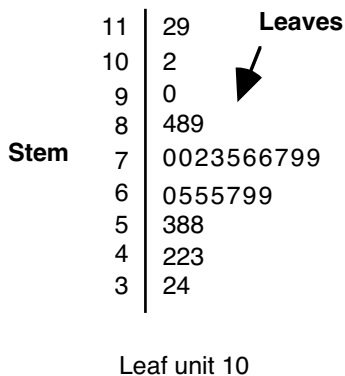
A distribution can be represented in two basic ways .

- 1) a frequency distribution table (to be seen later)
- 2) some type of graph .

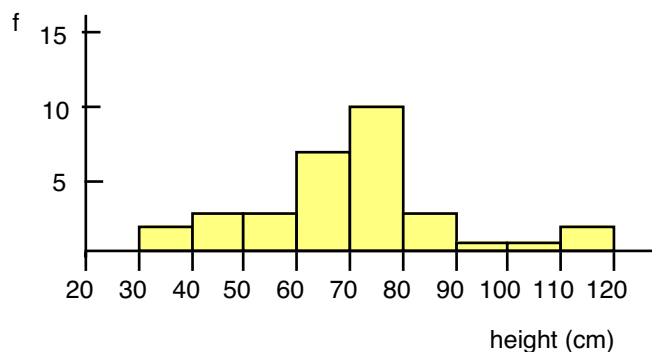
The easiest and quickest way to see what the distribution is like is to employ the use of a graph . **For INTERVAL DATA we use only two types**

- 1) a stem and leaf plot
- 2) a histogram

A STEMPLOT



A FREQUENCY HISTOGRAM





Decisions through Data - Unit 2 'Stemplots?'(11 minutes).

THE STEM AND LEAF PLOT (or STEMPLOT)

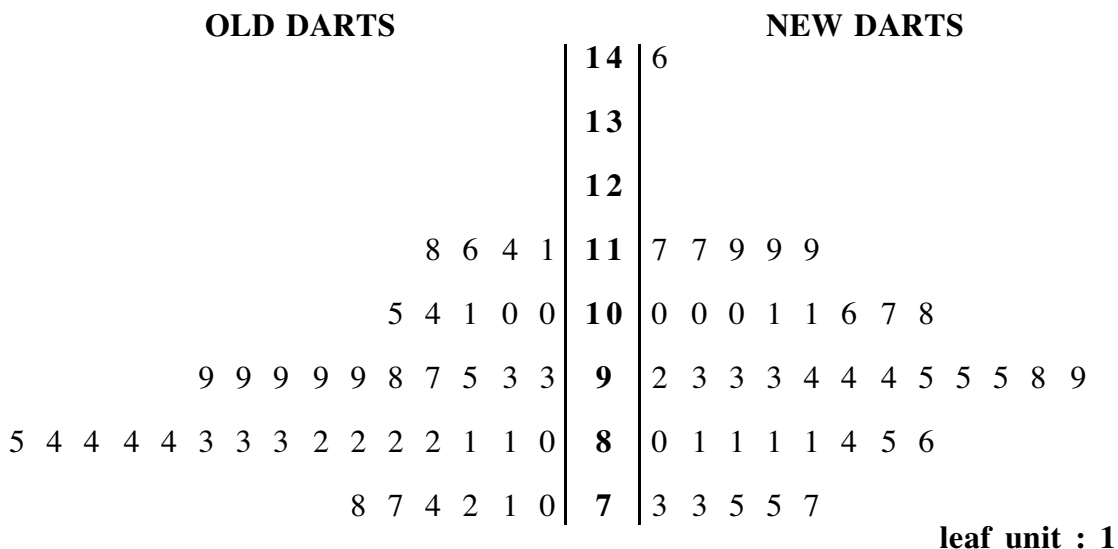
The stemplot offers us a quick way to gain a VISUAL picture of the distribution . The stemplot is only useful for **small sets of data**.

Case Sudy One: There are darts and there are darts!

Some years ago I set up an experiment to test a claim of one of my students. He claimed that the quality of dart used affected the accuracy of your throws. To test this claim he was asked to throw an expensive, new set of three darts 40 times. Each set of three darts had their score tallied with the aim being to score as many points as possible. He was then asked to repeat this process but with a set of old cheap darts I had found in a secondhand shop. The scores he achieved are shown below. The new darts were only thrown 39 times.

OLD DART scores					NEW DART scores				
84	82	85	81	83	81	95	94	93	95
82	84	99	84	99	81	75	99	119	94
83	84	82	80	82	81	93	92	117	119
83	74	72	77	70	146	101	100	106	86
81	93	71	99	93	101	100	100	98	85
99	97	100	98	105	95	94	93	80	119
104	99	114	118	116	75	77	73	107	108
110	95	111	101	78	73	84	117	81	

Here we have **two distributions** that we need to compare and contrast. To get an idea of how the data varies we will draw **BACK TO BACK STEM** and **LEAF PLOT** (actually two stem and leaf plots).



HOW TO CONSTRUCT ONE OF THESE

The leaf is formed from the last digit of each piece of data. The remaining digits of the data form the stem of the stem and leaf plot.

Firstly determine what the stem will be formed from by locating the lowest and highest number and dropping off the last digit. In this case we require stems from 7 to 14, which of course represents 70 to 140.

Then draw the stem as shown and attach the leaves, ie. the last digit of each piece of data. Note that the leaves are in RANK order from lowest to highest. It is easiest to achieve this by first drawing a stemplot with leaves that occur as read from the data and then another with the leaves in rank order.

In the case of having a single data set to investigate only one side of the plot would be drawn (as seen on page 10).

LEAF UNITS

It is common to attach a statement of what the leaf unit is. It is simplest to state the PLACE VALUE of the leaf, ie 1000, 100, 10, 1, 0.1, 0.01, 0.001 etc.

WHAT ELSE THEY SHOW US

Note that, as well as showing how the data varies, it shows the actual data, groups the data (in this case in groups of 'the 70's, the 80's etc.) and shows the FREQUENCY of each of these groups. The '70's group for the old darts has a frequency of 6. In layman's terms this means that the thrower threw 70 or over but under 80 six times out of the forty shots he had.



SI Exercise 1b

*We will revisit each of the problems below in **SI Exercise 1c**. Start each question on a clean page leaving another page free before starting the next question.*

- It is a commonly held belief that people who have suffered a heart attack have generally higher blood pressure than those people who have not and are not likely to have a heart attack. To investigate this claim 25 adult males who had suffered a heart attack and 25 adult males who had not were randomly selected from people at a heart week fundraising event. Their systolic blood pressures are given below.

HEART ATTACK					NO HEART ATTACK				
84	129	111	146	143	94	105	87	111	98
123	102	138	117	104	104	84	123	99	123
146	144	116	139	132	123	84	97	148	105
114	127	105	116	126	134	94	115	93	100
106	77	131	141	129	105	111	123	98	93

Produce a back to back stemplot for this data. Remember the leaves must be in rank order.

- Brain surgery is very tricky business. Time is considered to be a critical factor when operating. The shorter the operation the better. A new procedure has been developed to carry out a task for which an old procedure has been used for years. Forty operations using the new technique are timed and forty operations using the old technique are timed. The time is measured to the nearest one tenth of an hour. The data is given below.

NEW TECHNIQUE					OLD TECHNIQUE				
9.5	12.4	9.6	9.7	8.6	10.1	9.4	10.4	11.3	8.9
9.9	9.5	10.9	9.6	7.8	10.5	9.3	10.7	12.4	9.5
8.8	10.2	10.7	10.8	11.0	10.3	10.3	11.2	12.7	10.2
9.0	11.7	9.1	7.7	7.3	11.3	9.8	12.3	11.5	12.3
12.2	9.6	10.7	9.6	8.6	12.4	9.8	12.4	9.6	11.3
10.2	12.7	12.4	8.9	9.6	15.1	10.2	11.6	12.8	9.8
11.3	9.5	9.7	8.8	10.1	11.3	10.4	11.8	13.6	10.3
12.4	8.5	9.2	10.9	9.4	9.3	10.7	9.8	12.4	9.4

Produce a back to back stemplot for this data

3. In your job as JUNIOR STATISTICAL CONSULTANT for Harradine and Co. you are contacted by WHEELS magazine to compare the performance of German and Japanese cars. The variable chosen to measure performance is the time taken for the car to accelerate from 0 - 100 km/h (shortest time of course). You contact the manufacturers of a number of each type of car and request data on 0 - 100 km/h times. You receive the following.

GERMAN CARS						JAPANESE CARS				
10.0	8.5	6.9	5.5	6.4		9.4	9.3	8.0	6.2	7.7
6.4	7.9	8.8	5.1	6.0		2.0	9.1	6.5	9.3	9.5
8.5	6.9	7.1	10.9	4.9		8.9	6.8	12.5	8.6	12.0
7.5	9.2	8.7	8.6	8.9		6.7	7.1	8.2	10.0	
5.4	6.7	9.7	8.3			7.2	10.5	9.7	8.8	
						8.5	5.7	11.7	9.2	
						9.5	8.3	6.3	6.6	

Produce a back to back stemplot for this data

USING THE STEM PLOT

Now that we have the stemplot what do we do with it ? Remember that its purpose is to give us a visual representation of the distribution . There are some very useful features of the distribution to now compare and contrast.

When using the stemplot we must :

- 1) Look for any significant deviations from the overall shape . These may be
 - a) significant gaps between clusters of data
 - b) **OUTLIERS** - individual observations that fall **well outside** the overall pattern. Well outside is not easily defined and each situation must be taken on its merits.

It should be noted that outliers distort the overall picture and as such , seeing they are in an extreme minority it is tempting to cut them out . This is quite common practice in statistics BUT not before the reason for the outlier being there has been investigated. If no reason can be found for their presence that appears unusual then we must leave them in the analysis. Normally they are the result of faulty equipment or some human error . In some cases however they open up a whole new area to explore that could be critical to the study. The case study below outlines this fact.

In 1985 British scientists reported a hole in the ozone layer of the earth's atmosphere over the South Pole. This is disturbing, since ozone protects us from cancer-causing ultraviolet radiation. The British report was at first disregarded, since it was based on ground instruments looking up. More comprehensive observations from satellite instruments looking down had shown nothing unusual. Then, examination of the satellite data revealed that the South Pole ozone readings were so low that the computer software used to analyze the data had automatically suppressed these values as erroneous outliers. Readings dating back to 1979 were reanalyzed and showed a large and growing hole in the ozone layer that is unexplained and possibly dangerous. Computers analyzing large volumes of data are often programmed to suppress outliers as protection against errors in the data. As the example of the hole in the ozone layer illustrates, suppressing an outlier without investigating it can keep valuable information out of sight. (Source Moore & McCabe)

- 2) Examine the overall shape of the distribution: is it symmetrical or approximately symmetrical about a horizontal line passing through the centre or is it skewed to the high or low; is it uni or bimodal ? Examples of these 'shapes' can be seen overleaf. When deciding on shape be flexible - err on the side of symmetry or approximate symmetry. To be consider skewed, a distribution must posses a considerable 'tail'. *It should be noted that small samples may often have irregular shapes that do not reflect the shape of the population from which they were taken.*
- 3) Locate the centre of the distribution . This can be done by finding the **MEDIAN**. The median is the centre score when the data is in ascending rank order (which the stemplot does !) if there is an odd number of pieces of data , or the average of the two centre pieces if there is an even number of pieces of data .
- 4) Determine and comment on the **SPREAD** of each distribution. For now the tool we will use for this is the **RANGE** - which is simply calculated by subtracting the lowest score from the highest score. It is often useful to quote the lowest and highest score as well.

AN APPROXIMATELY SYMMETRICAL DISTRIBUTION

11	29
10	2
9	0
8	489
7	0023566799
6	0555799
5	388
4	233
3	25

leaf unit : 1

A SKEWED - HIGH DISTRIBUTION

11	2
10	2
9	0
8	489
7	0023
6	05557
5	344455778
4	2345567779999
3	25

Leaf unit 100

A SKEWED - LOW DISTRIBUTION

11	29
10	000122335577799
9	045778999
8	233338
7	000
6	05
5	38
4	23
3	2

Leaf unit 0.1

A BIMODAL DISTRIBUTION

11	29
10	233456
9	0113344455678
8	489
7	00
6	0555799999999
5	388
4	233
3	25

Leaf unit 0.1

INFERENCE USING STATISTICS

The process of inference is to arrive at a conclusion/hypothesis about a situation based on observations made from a number of cases (not all). In our dart problem we have 40 and 39 cases respectively. To infer with some measure of validity we must use all the information at hand. The following summary table will help us to do this.

	OLD DART	NEW DART
outliers	none	one at 146
shape	skewed to the high	approx. symmetrical
centre	84.5 (<i>average of the 20 and 21st piece</i>)	94 (<i>the 20th piece</i>)
spread	41 (<i>70 to 118</i>)	69 (<i>73 to 146 inc. outlier</i>) 42 (<i>73 to 119 exc outlier</i>)

TIME TO PLAY LAWYER.

Now we must use the information in the table to build an argument to support the conclusion to which we have come. It is no good simply saying the new darts are better. We need to be able to use the analysis we have carried out to argue our point.

You need to build a **picture of the situation in the reader's head**, and use this to convince the reader of your stance. The following suggests a standard approach that is useful in most cases. The following acronym will help you to remember how to do it:

G.O.S.C.S.C

1. Produce an appropriate **G**raphic.
2. Look for **O**utliers and treat them appropriately
3. Describe/compare the **S**hape of each distribution.
4. Describe/compare the **C**entre of each distribution. [median]
5. Describe/compare the **S**pread of each distribution. [range]
6. Draw your **C**onclusion.

If you follow this line of analysis then you will be able to form a conclusion based on **firm knowledge** rather than simply **speculation**.

A NOTE ABOUT CONCLUSIONS

One needs to be careful when drawing conclusions. You need to consider whether you have all the data (a census) and hence can make a **factual conclusion in a general sense** (ie about the population) or whether you only have a sample of the data and as a result can only conclude with a **hypothesis** about the characteristics of the **population** or support for a hypothesis being tested.

A MODEL ARGUMENT FOR CASE STUDY ONE.

OLD DARTS		NEW DARTS
	14	6
	13	
	12	
8 6 4 1	11	7 7 9 9 9
5 4 1 0 0	10	0 0 0 1 1 6 7 8
9 9 9 9 9 8 7 5 3 3	9	2 3 3 3 4 4 4 5 5 5 8 9
5 4 4 4 4 3 3 3 2 2 2 2 1 1 0	8	0 1 1 1 1 4 5 6
8 7 4 2 1 0	7	3 3 5 5 7


leaf unit : 1

One outlier is present in the new dart scores - a score of 146, 27 points higher than the next highest score. This score was investigated and found to be a true score and as such will remain in the rest of the analysis. The old dart distribution is skewed to the high while the distribution of new dart scores is approximately symmetrical. The centre of the new dart distribution is 10 points higher than the old dart distribution (medians of 94 points and 84.5 points respectively). The spreads of the distributions are very similar if we ignore the outlier (46 (73 to 119) for new darts and 48 (70 to 118) for old darts).

Hence I can conclude that the analysis of our samples **support the hypothesis** that the scores from the new dart will be, **on the whole**, better than those that will be returned from the old darts.

Note that, unless you have carried out the experiment, it may not be possible to comment, as has been above, about the outlier. In this unit you will be instructed what to do. In your own project, however, you will treat outliers as previously discussed.

THE TERMS 'SUGGEST', 'ON THE WHOLE' OR 'ON AVERAGE' and LIMITATIONS / EXPERIMENTAL ERRORS.



THOUGHTS

T1.5 Note the use of the terms, 'support the hypothesis', 'on the whole' or 'on average' are very important.

Why?

T1.6 The experimental set up of this dart investigation may well be less than perfect. It may even make the conclusions somewhat suspect. Discuss areas of this experimental setup you feel may be suspect.



SI Exercise 1c

Do this work in part of each of the blank spaces you were asked to leave in SI Ex. 1b.

1. *Treat all outliers as correct pieces of information.*

Revisit question 1 of **SI Exercise 1b**

- a) Copy and complete the following table in your problem book.

	HEART ATTACK	NO HEART ATTACK
outliers		
shape		
centre (median)		
spread (range)		

- b) Write an argument that supports the conclusion you have drawn about the blood pressures of people who have had heart attacks compared to those who have not (ie the populations from which the samples were drawn).

2. *Treat all outliers as mistakes and disregard them once they have been discussed.*

Revisit question 2 of **SI Exercise 1b**

- a) Copy and complete the following table in your problem book.

	NEW TECHNIQUE	OLD TECHNIQUE
outliers		
shape		
centre (median)		
spread (range)		

- b) Write an argument that supports the conclusion you have drawn about the different brain surgery techniques.

3. *Treat all outliers as mistakes and disregard them once they have been discussed.*

Revisit question 3 of **SI Exercise 1b**

- a) Copy and complete the following table in your problem book.

	GERMAN CARS	JAPANESE CARS
outliers		
shape		
centre (median)		
spread (range)		

- b) Write an argument that supports the conclusion you have drawn about the times of the population of German cars compared to the population of Japanese cars.

HISTOGRAMS



Decisions through Data - Unit 3 'Histograms and Distributions' (11 minutes).

Consider the following data . It is the amount of money (\$) spent by 50 randomly chosen shoppers in a grocery store .

2.32 6.61 6.90 8.04 9.45 10.26 11.34 11.63 12.66 12.95 13.67 13.72 14.35
 14.52 14.55 15.01 15.33 16.55 17.15 18.22 18.30 18.71 19.54 19.55 20.58
 20.89 20.91 21.13 23.85 26.04 27.07 28.76 29.15 30.54 31.99 32.82 33.26
 33.80 34.76 36.22 37.52 39.28 40.80 43.97 45.58 52.36 61.57 63.85 64.30
 69.49

THOUGHTS

1.7 *Why would it be difficult to draw a stemplot for this data set? If you can not see why, try it out - it will be obvious then.*

1.8 *Go back to the data you collected on the frogs from investigation one. Can you draw a stemplot that will show you how the data are distributed? Why or why not?*

We need to modify the way we draw a stemplot or employ the use of another type of graphic. In this course we are going to look at another, more universally acceptable graphic - the HISTOGRAM.

Histograms are the formal graphic that replaces the stemplot for displaying the distribution of an **INTERVAL VARIABLE**.

A stemplots is used as an informal tool to get an initial and quick idea of what the distribution is like. Hence they are only ever used with relatively small data sets.

One good point about a stemplot is that it does show the actual data - a histogram does not. However this is probably the only factor it has over a histogram.

Recall that the stemplots 'grouped' the data. We actually need to do this before a histogram is produced. This is done with the help of a **FREQUENCY DISTRIBUTION TABLE**. They are called this as they show the pattern of variation in the data albeit without a picture. This will be seen in the FREQUENCY COLUMN.

Case Sudy Two: My mate the Great White Shark.

While Greg Norman was here for the S.A. Open, I decided to ask for his cooperation in a data gathering exercise. He agreed. I asked him to hit 30 balls in succession with his Driver. I then measured how far each ball travelled. The data was as follows:

251.2, 245.1, 248.0, 251.1, 254.6, 248.8, 263.2, 262.9, 265.0, 254.5
 264.3, 257.0, 262.8, 264.4, 260.6, 255.9, 269.7, 263.2, 277.5, 267.4
 270.5, 265.5, 270.7, 272.9, 275.6, 266.5, 265.5, 244.6, 253.9, 250.0

This type of data must be GROUPED before a histogram can be drawn.

In forming groups, find the lowest and highest values, and then make the **group width** such that you achieve about **6 to 10 groups**.

In this case the lowest is 244.6m while the largest is 277.5m. This gives a range of approximately 35m, hence a group width of 5 will give eight groups.

We will use the following method of grouping. The group '240 - ' actually means that any piece of data ≥ 240 but <245 can fit in this group. Similarly in the group '260 - ' will contain data ≥ 260 but <265 . This technique creates a home **for every number** ≥ 240 but <280 . Groups should all be of the same width.

The tally column is used to count the data that falls in a given group in an efficient way. Do not try to determine the number of pieces of data in the 240 - group first off. Simply place a stroke in the tally column to register an entry as you work your way through the data from start to finish as it is presented to you.

The frequency column summarise the number of pieces of data in each group. The relative frequency column measures the percentage of the total number of pieces of data in each group. Here, percentages offer an easier way to compare the 'amount' of balls Greg hit over 270 m but under 275 m to the number he hit under 245 m.

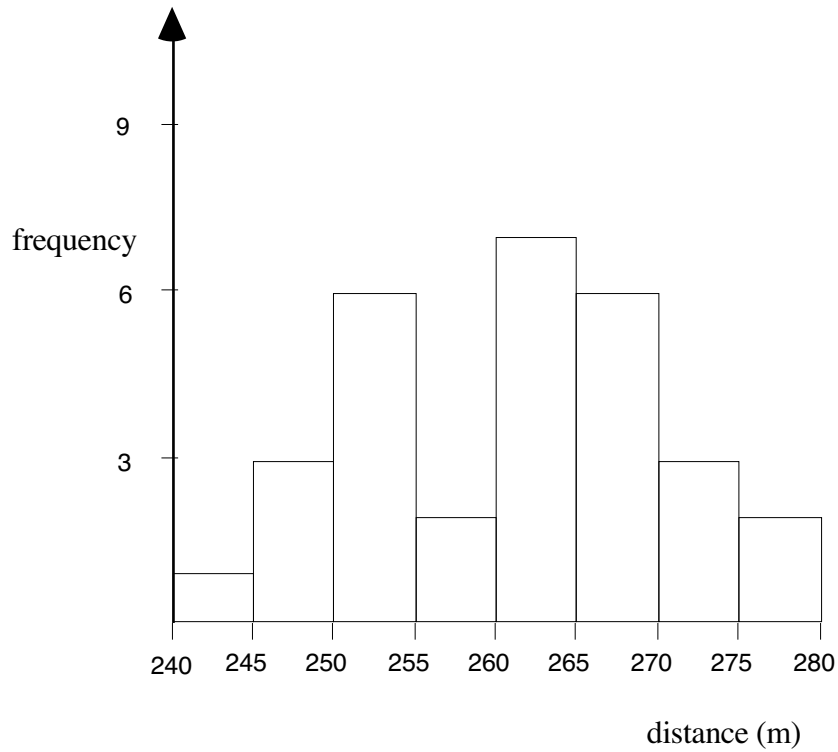
GREG NORMAN'S 30 DRIVES.

Scores (Subject Scores)	Tally	Frequency (f)	Relative Frequency (rf)
240 -		1	3.3
245 -		3	10
250 -	###	6	20
255 -		2	6.6
260 -	###	7	23.3
265 -	###	6	20
270 -		3	10
275 - (but < 280)		2	6.6
		$\Sigma f = 30$	$\Sigma rf = 100$

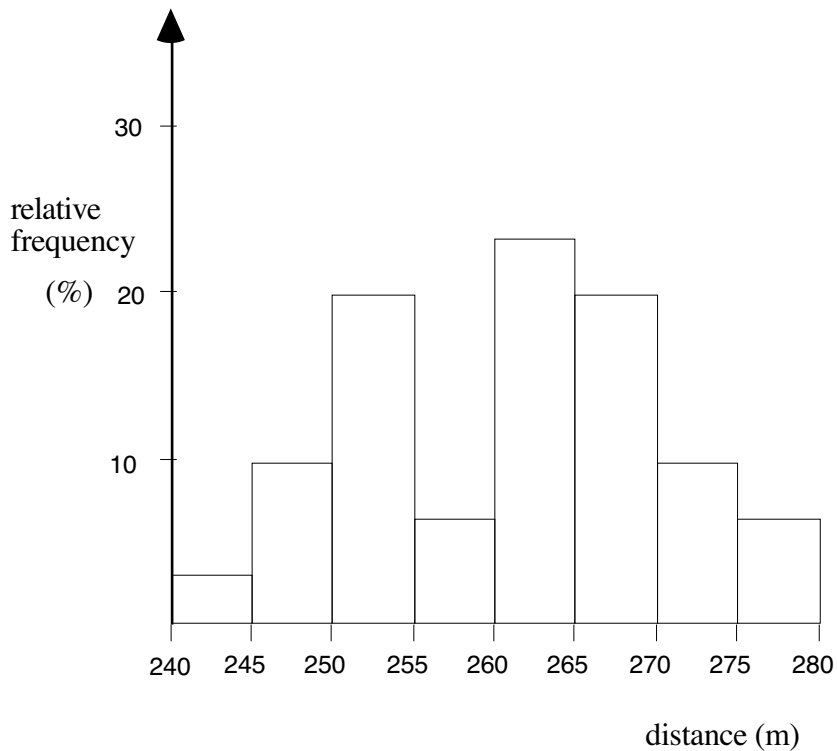
From this table two histograms can be drawn, firstly a **FREQUENCY HISTOGRAM**, and secondly a **RELATIVE FREQUENCY HISTOGRAM**. They look as follows

Note all histograms require a title.

A **FREQUENCY HISTOGRAM** displaying the distribution of 30 of Greg Norman's drives.



A **RELATIVE FREQUENCY HISTOGRAM** displaying the distribution of 30 of Greg Norman's drives.



The advantage of the relative frequency histogram is best seen when you wish to compare distributions with different numbers of pieces of data. Using percentages allows for a fair comparison.

Notice how the horizontal axis is labelled. **The left edge of each bar is the first possible entry for that group.**

Given that a histogram is simply an elaborate stemplot, all of the theory learned in the stemplot section applies to histograms as well. For example, all the shape theory holds. **G.O.S.C.S.C** obviously still applies but the G is now a histogram and not a stemplot.



SI Exercise 1d

In each of the following questions compare the look of the histograms produced to the stemplots produced earlier. As we are wanting to compare the data sets using the histograms produced, be sure that your groups are the same for each set and that the frequency or relative frequency scales are identically incremented.

1. Review the data on the blood pressures of people who had previously suffered heart attacks and those who had not. It is on page 13.
 - a) Determine the highest and lowest score for each data set.
 - b) Produce groups of width that you will give you between 6 and 10 groups in which to place all data points. **Use the same groups for each of the variables so that you can compare the distributions using your histograms. Also, make sure you use the same horizontal scale.**
 - c) Produce two Frequency Distribution tables, as seen on page 22, one for the blood pressures of those who have had a heart attack and one for those who have not.
 - d) Draw two frequency histograms for this data. Be sure to attach all the correct labels.
 - e) Draw two relative frequency histograms for this data. If you think about it you should be able to do it on the same histograms as you have drawn in d). If you can not see how to do this draw two separate ones.

2. Review the data on the times for two techniques of a brain surgery procedure. It is on page 13.
 - a) Determine the highest and lowest score for each data set.
 - b) Produce groups of width that you will give you between 6 and 10 groups in which to place all data points.
 - c) Produce two Frequency Distribution tables, as seen on page 22.
 - d) Draw two frequency histograms for this data. Be sure to attach all the correct labels.
 - e) Draw two relative frequency histograms for this data.

3. Review the data on the times to 100 km/hr for a number of German and Japanese cars. It is on page 14.
 - a) Determine the highest and lowest score for each data set.
 - b) Produce groups of width that you will give you between 6 and 10 groups in which to place all data points.
 - c) Produce two Frequency Distribution tables, as seen on page 22.
 - d) Draw two frequency histograms for this data. Be sure to attach all the correct labels.
 - e) Draw two relative frequency histograms for this data.



Producing a frequency histogram and frequency distribution table from raw data

with the

CASIO 9850GB PLUS



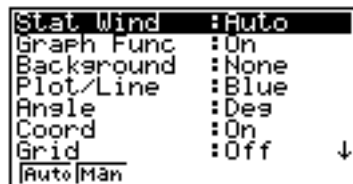
With the calculator turned on and the main menu visible, use the arrow key to highlight the STAT menu (shown opposite).



Then press the blue EXE key (alternatively, simply press 2). The following screen will result:



Now press SHIFT and MENU to arrive at the preferences for this module. Be sure each option is set as shown opposite. Use the down arrow key to scroll down to the lower options.



Press the EXIT key to return to the window containing the lists.

To enter data into the lists, simply use the number key pad and press the blue EXE key after each entry.



Enter the distance data from Greg Norman seen on page 22.

If you make an error in typing BEFORE you press the EXE key, use the back arrow key to repair it. If you have pressed the EXE key simply use the up arrow to select the wrong data point and re-enter it.

The title for each list can not be changed.

Access $\overline{\text{F6}}$ (F6) to reveal the options at the screen base seen opposite.



Access GRPH (F1) to reveal the options shown opposite. We are able to set up three different types of graphs.



Access SET (F6). This allows us to set up from one to three graphs.

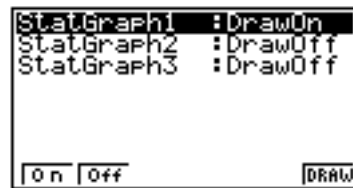
You can select which graph you wish to set up by accessing GPH1 (F1) or GPH2 (F2) or GPH3 (F3). Graph 1 should be selected now.



Use the down arrow key and the F keys to set the options as shown opposite.

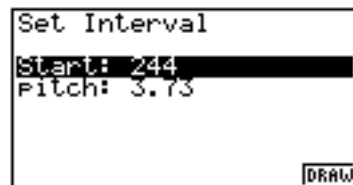
Press the EXIT key.

Access SEL (F4). This allows you to select which of the three graphs you want the calculator to draw.

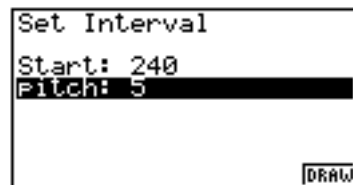


Have graph 1 turned on and the other two turned off.

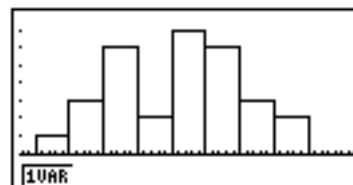
Access DRAW (F6) and the screen opposite will appear. The calculator has suggested the first lower group limit (Start) and the group width (pitch) for us.



We can change this to what ever we want. Use the arrow keys to select each and then change them to 240 and 5 respectively. You should always choose these yourself.

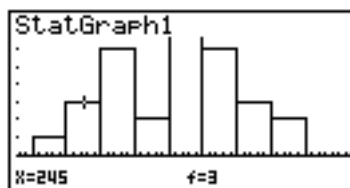


Access DRAW (F6) and the frequency histogram seen opposite will result.



Note the 1VAR option at the bottom left of the screen. We will want to use this later.

Press SHIFT and access TRCE (F1). This allows you to 'arrow' across the tops of the frequency histogram's bars to determine the frequency values for each group. The X value is the lower group limit, even though it reads it as in the centre of the bar.



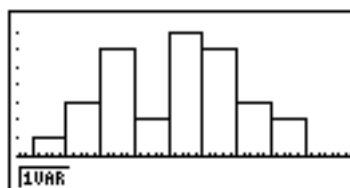
Compare this frequency histogram to the one seen on page 24.

Note that the tracing of the histogram allows us to produce a frequency distribution table without the manual tallying – the machine does it for us.

Calculating the MEDIAN and other summary statistics with the CASIO 9850GB PLUS

Recall the 1 VAR option seen opposite.

It will tell the calculator to calculate a series of summary statistics for the data you are analysing.



It must, however, be set up first.

Press the MENU key and re-enter the STAT module. Access CALC (F2) and then access SET (F6). The window opposite will be result. Set the options as shown and press the EXE key.

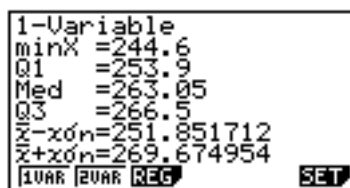


It is here that you tell the calculator which list the data is in that you wish for it to calculate summary statistics.

Now access 1 VAR (F1) and the screen(s) opposite will result. Arrow up and down to see all calculations.



We will discuss most of these later, but for now, observe the mean quoted as \bar{x} and the median quoted as Med.



Because we have only 1 variable's data listed, we need only worry about the top

setting. Be sure it is set on **List 1**, it can be changed as desired by accessing any of the lists shown at the screen base.

You must re-set this window if you have multiple lists of data and require the statistics for, say List 3.



If you have two lists of data, and set the window as shown, it will calculate the statistics for both sets and record them separately, one as X and the other as Y .



SI Exercise 1e

Note, so that you do not have to delete data, use the ‘file’ structure on your calculator. In List mode, go to the SET UP (Shift then MENU) and choose File 1 for question 1. Then use File 2 for Question 2 and so on.

1. Enter the 'Brain Surgery' data into the Casio 9850 GB Plus. Put the old technique data in list 1 and the new techniques data into list 2.
 - a) Produce a histogram for the old techniques data. Compare them with the histograms that you drew for SI Exercise 1d on page 26
 - b) Find the median of each distribution according to the calculator. Record these values and compare them to your own calculations, which you should be able to find in the tables from **SI Exercise 1c**
 - c) Also record the 'mean' of each distribution as calculated by the calculator. We will be studying this statistic later and will come back to these values that you have recorded.
 - d) By using the trace feature of the graphic calculator produce a frequency distribution table for this data.

2. Enter the ‘Car’ data into the Casio 9850 GB Plus. Put the German car data in list 1 and Japanese Car data into list 2.

Repeat parts a) - d) of question 1.

3. Enter the ‘Dart’ data into the Casio 9850 GB Plus. Put the old dart data in list 1 and the new dart data into list 2.

Repeat parts a) - d) of question 1.

4. Go to the *SeniorSchoolCensus-online* website. Go to the ‘sampler’ and choose a sample of 255 Year 12 girls from private schools and another sample of 255 Year 12 boys from private schools Using the FA-123 Computer Link, transfer the data pertaining to amount of money earned last week into the Casio 9850GB Plus.

Repeat parts a) - d) of question 1.

FINE TUNING OUR ARGUMENTS

SOME NEW STATISTICS

So far we have been concentrating on the skill of building an argument to prove a point. We have essentially used the *STATISTICS* that measure centre, spread and a description of the shape of a distribution to build our arguments. It should be noted how weak an argument is if you use just one statistic. It is important to use G.O.S.C.S.C Believe it or not our argument can still be made stronger.

Other STATISTICS that measure the centre and spread of a distribution exist in addition to what we have learned. The *range* especially is a rather poor measure of spread when used on its own, so the following measures will be used to strengthen our arguments. In some situations one measure may be better than another, hence you will also learn when to use and when not to use a certain measure.



**Decisions through Data - Unit 4 'Measures of Centre'
(11 minutes).**

POPULATION vs SAMPLE and SYMBOL CONVENTIONS

Statistics are calculated for a set of data. If the data is gathered from *every possible member* of the group being studied, we say that the data is taken from the **POPULATION** being studied. Normally the populations being studied are so large that this is either not possible or uneconomical.

More commonly a **SUBSET** of the members of the population will have data collected from them. Such a subset is called a **SAMPLE**, and its members are normally chosen in some sort of **RANDOM** fashion so that the sample is essentially identical to the population in all respects except for size. The aim is to make the sample an accurate *model* of the population so that the features of the **sample's distribution** are close to the features of the **population's distribution**. Hence, if the method of sample selection was successful, statistics such as mean and range should have similar values for both the sample and the population.

Hence two types of STATISTICS exist:

1. SAMPLE STATISTICS
2. POPULATION STATISTICS (more commonly called *population parameters*)

eg. it is possible to have a *sample mean* and a *population mean*. To distinguish between these we use different symbols. These will be introduced in the following sections.

STATISTICS THAT MEASURE A DISTRIBUTION'S CENTRE

Three measures of centre are commonly employed. They are the *mean, median and mode*. These are often called **AVERAGES**.

Both the *mean and median are only useful statistics for INTERVAL data sets*. The mode is most appropriately applied to nominal or ordinal data sets, this will be discussed further later on.

Median

The median is the only **ALWAYS** appropriate measure of centre. As we saw earlier, it is simply the central piece of data in the set when it is **placed in rank order**. Hence it can be thought of as the value which *about half of the data in the data set is above and about half of the data is below*.

Should the data set contain an odd number of pieces of data then the process of locating the median is simple.

eg. If there is 25 pieces of data, rank them in order and then realise that it is not possible to split the data into two equal sized sets. The best effort results in two sets of 13 and one left over. Hence the 14th piece of data will be the median.

69 pieces of data would result in thinking that we could have two sets of 34 with one left over and hence the 35th piece of data will be the median.

Should the set contain an even number of pieces of data then it could be split into two equal subsets. Hence in this case there is really *two central values*. We say that in cases such as this the median is the average of the two central pieces.

eg. If there is 24 pieces of data, rank them in order and then realise that it is possible to split the data into two equal sized sets, both of 12. Hence the average of the 12th and 13th pieces of data will be the median.

70 pieces of data would result in thinking that we could have two sets of 35 and hence the average of the 35th and 36th pieces of data will be the median.

Mean

The mean is calculated by adding all the pieces of data together and dividing by the number of pieces of data in the set. Hence it takes into account all the values in the sample or population, unlike the median.

If we let x be the variable representing the data pieces, n be the variable representing the number of pieces of data in the sample, Σ represents "the sum of", \bar{x} represent the mean of a **sample** and μ represent the mean of a **population** then we have:

$$\mu = \frac{\Sigma x}{n} \quad \text{or} \quad \bar{x} = \frac{\Sigma x}{n}$$

In essence the mean is the value that *each piece* of data in a given data set could be to achieve the same Σx as in the actual data set.

eg. the numbers 4,6,16,4 and 10 have a mean of 8. Hence the data set 8,8,8,8,8 sums to 40 just like the original.

In an applied sense we could consider the batting average of Sir Donald Bradman. He had an average of 99 runs from 70 innings. The 99 tells us that if he had scored 99 in each of his 70 innings then he would have accumulated the same number of runs overall as he did anyway. It may or may not have been a good measure of the centre of The Don's distribution of runs.

THE MERITS OF THE MEAN AND MEDIAN USED AS A MEASURE OF CENTRE

Because of the way that the mean is calculated, for some data sets it is a very POOR measure of a distributions centre.

If a data set has a few extremely high values compared to the other set members, then the mean value is increased dramatically. Similarly if the set contains a few extremely low values, the mean value will be decreased dramatically. Hence the mean values will not be really measuring the data sets centre.

Should there be about the same number of extremely high and extremely low values then their effects nullify each other.

*Hence if the data set has symmetry
the mean should accurately measure the centre of the distribution.*

It should be noted that the MEDIAN is not affected by extremely high or low scores in the same way that the mean is, since it is not the result of any addition.

As a result of this the **median** is called a **RESISTANT MEASURE OF CENTRE**, while the *mean* is called a *NON-RESISTANT* measure of centre as it has little resistance to the effects that a few large or small values have on it.

Mode

The mode is quite simply the most frequently occurring LEVEL in a data set. This is next to useless in an interval data set due to the fact that in most cases the interval variable will have MANY levels, and it is usual that a few of these levels will have a similar frequency. Hence the concept has little meaning. However, in a nominal or ordinal set where the number of levels is far less, the most frequent level is likely to be useful to report. It is called a measure of centre as quite often the mode or modal group in the case of grouped data is around the centre of the distribution.

A SCENARIO TO CONSIDER.

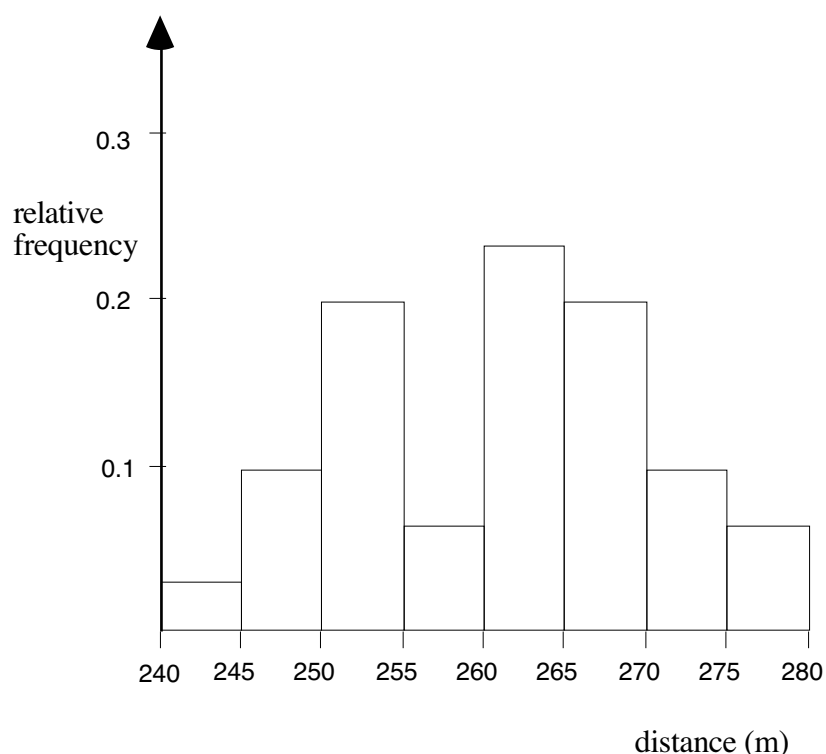
Recall the data gained from Greg Norman while he was here for the S.A. Open. The data was as follows:

251.2, 245.1, 248.0, 251.1, 254.6, 248.8, 263.2, 262.9, 265.0, 254.5

264.3, 257.0, 262.8, 264.4, 260.6, 255.9, 269.7, 263.2, 277.5, 267.4

270.5, 265.5, 270.7, 272.9, 275.6, 266.5, 265.5, 244.6, 253.9, 250.0

A relative frequency histogram illustrating the distribution of the length of his 30 drives is given below.



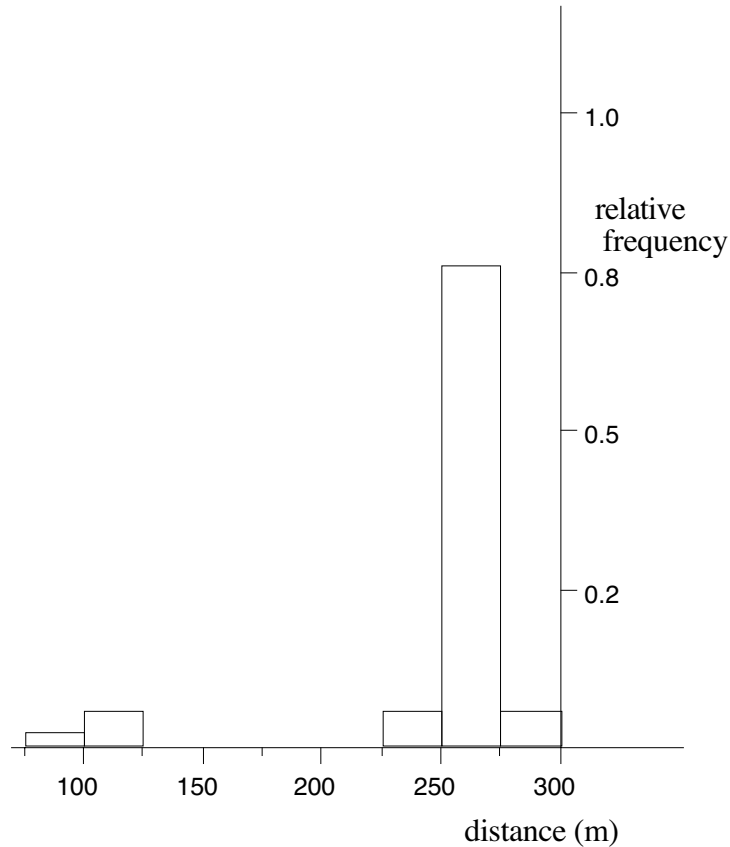
Notice that Greg's drives had no abnormally high or low values and the distribution is approximately symmetrical.

The median is 263.05m while the mean is 260.76m. Hence the mean is an accurate measure of centre in this case.

Now what would have happened if Greg had hit a few DUFFERS. Let us say that his three shortest balls were very short! The data set would change as follows:

251.2, **82.1**, **111.0**, 251.1, 254.6, 248.8, 263.2, 262.9, 265.0, 254.5
 264.3, 257.0, 262.8, 264.4, 260.6, 255.9, 269.7, 263.2, 277.5, 267.4
 270.5, 265.5, 270.7, 272.9, 275.6, 266.5, 265.5, **103.2**, 253.9, 250.0

A relative frequency histogram would look as follows:

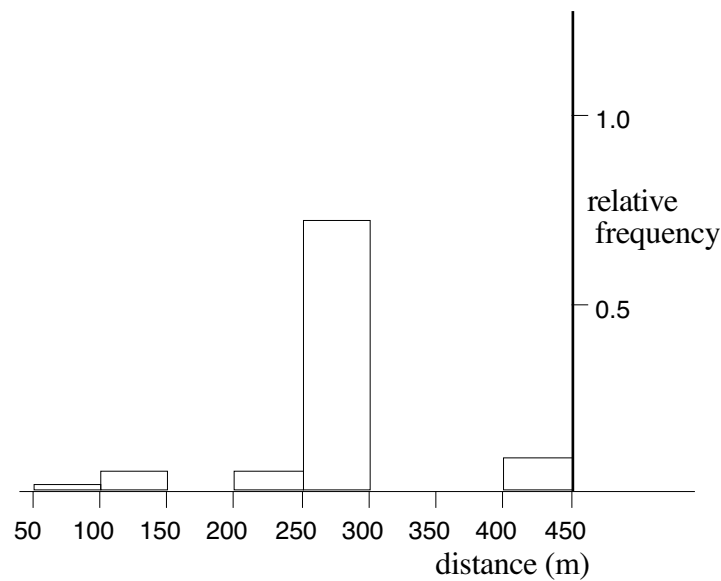


The median obviously does not change however the MEAN will. It actually becomes 245.83m, about 15 metres less than before. Hence the mean is no longer an accurate measure of the centre of this distribution.

What would happen should Greg have hit a few ROCKET balls in addition to the duffers? Let us imagine that the longest balls he hit were very long. The data could have become:

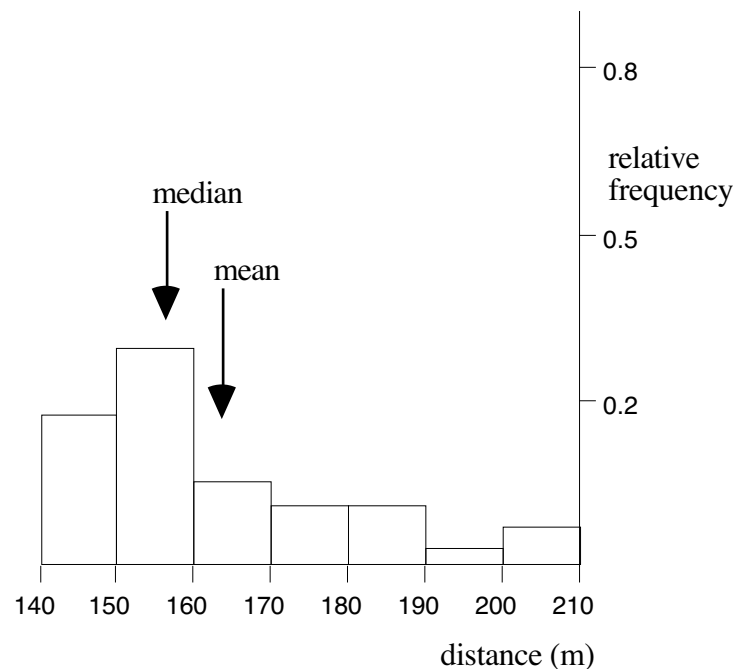
103.2, **82.1**, **111.0**, 251.1, 254.6, 248.8, 263.2, 262.9, 265.0, 254.5
 264.3, 257.0, 262.8, 264.4, 260.6, 255.9, 269.7, 263.2, **415.5**, 267.4
 270.5, 265.5, **403.9**, **420.0**, 275.6, 266.5, 244.6, 244.6, 253.9, 250.0

A relative frequency histogram reveals:



The mean of this distribution is 261.49 metres while the median has of course remained unchanged at 263.05m. Hence the extra large and extra small values have nullified each other and the mean is again an accurate measure of the centre of this distribution. **Note that the distribution is again approximately symmetrical.**

Now I fancied myself as a bit of a GREY SHARK. So I went and hit 30 golf balls with my driver. The relative frequency histogram reveals the results below.



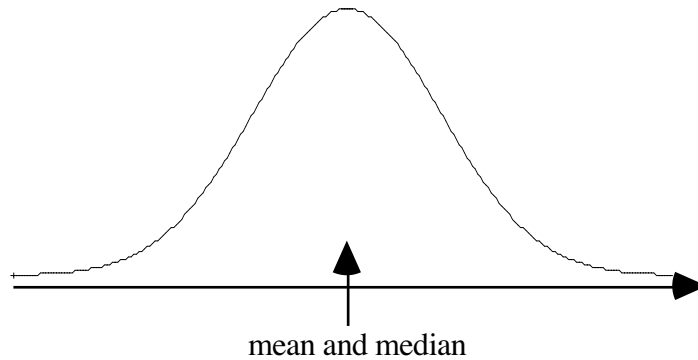
This distribution is clearly skewed to the high. This suggests that the bulk of my hits were around the 140 to 170 metres but I did get a few out around the 170 to 210 region. This would suggest that the mean would not be an accurate measure of the centre of this distribution due to the few higher scores. Indeed the mean is 163.66m compared to the median of 157.50m.

Note that you have not been supplied with the data in this case. The idea is for you to get a feel for the data set from the histogram.

SUMMARY OF MEDIAN AND MEAN

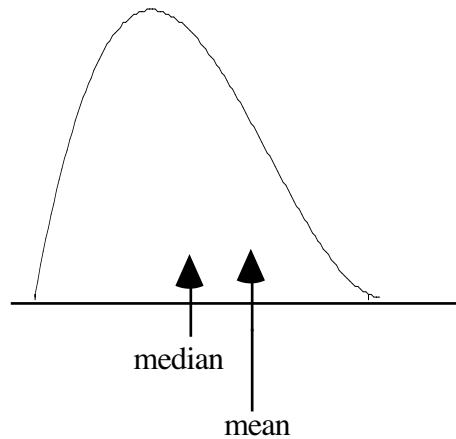
- All measures of centre are useless to build an argument with if quoted on their own.
- The median is the only measure of centre that will locate the true centre regardless of the data set's features. It is unaffected by the presence of higher or lower than normal values. It is called a **resistant measure of centre**.
- The mean is an accurate measure of centre if the distribution is symmetrical or approximately symmetrical. If it is not then the unbalanced high or low values will DRAG the mean toward them and hence cause the mean to be an inaccurate measure of centre. It is called a non-resistant measure of centre. *It should not be used in discussion if it is considered inaccurate.*
- The following diagrams show the approximate relative positions of the mean and median for the more common shaped distributions.

SYMMETRICAL OR APPROXIMATELY SYMMETRICAL DISTRIBUTIONS

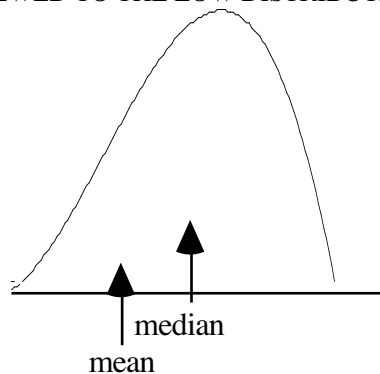


NOTE: other symmetrical shapes exist other than this 'BELL' type shape. Any symmetrical distribution will have approximately equal means and medians.

SKEWED TO THE HIGH DISTRIBUTIONS



SKEWED TO THE LOW DISTRIBUTIONS





SI Exercise 1f

1. a) Go back to the data sets from questions 1,2 and 3 from SI Exercise 1b. Determine, **without** the help of the CASIO 9850GB Plus, the mean for each of the set.
- b) Consult the tables you completed as a part of SI Exercise 1c and verify that the distribution's shape has caused relative positions of the mean and median to be as expected.
2. Go back to your solutions to SI Exercise 1e and find where you recorded the mean and median for the data from questions 1,2 and 3 from SI Exercise 1b. Check to see that your calculation of the mean corresponds to that of the calculator.
3. Choice magazine decides to compare two brands of electric light globes. Manufacturer A claims to produce globes that last as long as Manufacturer B. The interesting point is that Manufacturer A charges only half the price of Manufacturer B for the globes. Forty globes from each manufacturer are randomly selected and their life time (to the nearest hour) is recorded. The data is given below.

Manufacturer A									
460	943	984	1080	909	721	922	939	799	911
1041	926	720	1016	831	793	748	905	684	852
1005	926	835	868	773	1052	971	977	1014	927
972	870	938	954	1093	697	876	946	852	859
Manufacturer B									
1088	897	1072	943	1053	992	836	1038	918	819
1000	986	1018	888	1113	907	994	1020	1077	952
1230	1015	1022	1096	1082	1005	1174	1007	942	959
1034	1004	903	1077	1012	1154	1116	962	1016	1513

- a) Using the CASIO 9850GB Plus, produce histograms with group widths of 100 hours and summary statistics for each data sets and complete the table below. Copy it into your workbook.

	Manufacturer A	Manufacturer B
outliers		
shape		
centre (median)		
centre (mean)		
mean accurate (Y or N)		
spread (range)		

- b) Write an argument that concludes whether or not Manufacturer A's claim seems valid based on the 80 globes tested. Go back and review the format of a statistical argument on page 18 and page 19. For this argument you can attach the more universally acceptable histogram rather than a stemplot and, if appropriate, discuss the mean as a measure of centre.

IMPROVED MEASURES OF SPREAD



Decisions through Data - Unit 5 'Boxplots' (10 minutes).

1. QUARTILES AND THE INTERQUARTILE RANGE.

Recall that thus far we have only used the range to give us an idea of spread of a distribution. This has some huge inadequacies.

	<p><i>T1.10 List the obvious inadequacies of using the range as a measure of spread.</i></p>
<p>THOUGHTS</p>	

The range is clearly a NON-RESISTANT measure of spread.

Statisticians realised that if they were to invent a measure that in some way avoided the use of the possible extreme values that they would have a more reliable measure of spread. They did so and called it the **INTERQUARTILE RANGE**.

It is essentially the range of the middle 50% of the data.

This means that we need to work out the **quarter** and **three quarter** markers of a data set when in **rank order**.

The quarter marker is called QUARTILE 1 (Q_1) and the three quarter marker is called QUARTILE 3 (Q_3).

Q_1 is simply the median of the lower half of the data when in rank order.

Q_3 is simply the median of the upper half of the data when in rank order.

To find these values we simply rank the data set from lowest to highest, find the actual median and then find the median of the two halves. We have a little difficulty depending on how many data points there are in the set. For example:

80, 93, 94, 94, 94, 97, 98, 98, 104, 108, 110, 111, 114, 134
--

Here we have 14 data points and hence the median is NOT one of the data points, so we can divide this data into halves exactly. Therefore, we have the median being 98; the average of the two central scores - both 98:

80, 93, 94, 94, 94, 97, 98, 98, 104, 108, 110, 111, 114, 134
--

We then see we have 7 data points either side and hence we now find the median of these as we normally would, 3 either side making $Q_1 = 94$ and repeating the process for the upper part we see that $Q_3 = 110$.

Now for a different number of data points we will have a slightly different method.

81, 84, 85, 88, 95, 97, 100, 101, 102, 104, 106, 118, 125

Here we have 13 data points, and hence our median is one of the data points - 100. Now to divide this set into two halves we wonder what to do with the 100 - we ignore it exists, and calculate the median of the lower six scores and the upper six scores.

81, 84, 85, 88, 95, 97, 100, 101, 102, 104, 106, 118, 125

With six points we see we must find that Q_1 will be the mean of the third and fourth values ie 86.5 and that Q_3 will be the mean of the tenth and eleventh values, ie 105.

So:


$$Q_1 = 86.5$$

$$Q_3 = 105$$

The INTERQUARTILE RANGE (IQR) is now simply calculated by subtracting Q_1 from Q_3 .

$$\begin{aligned} \text{IQR} &= Q_3 - Q_1 \\ &= 105 - 86.5 \\ &= 13.5 \end{aligned}$$

Clearly this gives us a better idea about the spread of the data, BUT it has some flaws.



THOUGHTS

T1.11 List two possible flaws in this statistic

T1.12 Describe how to find the quartiles if 1999 pieces of data were present.

T1.13 Describe how to find the quartiles if 2000 pieces of data were present.

T1.14 Describe how to find the quartiles if 2001 pieces of data were present.

THE FIVE NUMBER SUMMARY
AND ITS GRAPHICAL REPRESENTATION -
THE BOX AND WHISKER PLOT

We now have knowledge of a simple set of STATISTICS that measure the centre and spread of an interval distribution. In order to convey a simple but useful picture of any interval distribution to a person, the FIVE NUMBER SUMMARY can be used. This is simply the following five statistics:

LOWEST SCORE, Q1, MEDIAN, Q3 and the *HIGHEST SCORE*.

These five numbers can also be represented graphically using a Box and Whisker Plot, sometimes simply called a BOXPLOT. When two or more interval data sets are to be compared SIDE BY SIDE BOXPLOTS are drawn. Previously we said we would learn about a tool that was better to use when comparing distributions than histograms; the boxplot is the tool.

Consider the following scenario:

Mr. Smith's class and Mr. Jackson's class have a competition based around their pulse rates after exercise. Each class member exercises in the same way for five minutes and then takes their pulse. Mr. Jackson's class maintains that they will have a generally lower pulse rate after exercise. The data is given below.

Name	Pulse Rate (beats/min)	Class
HUW	9 4	SMITHS
ANDREW M.	9 4	SMITHS
DAMIEN	8 0	SMITHS
CARLO	1 1 0	SMITHS
PETER	9 7	SMITHS
MICHAEL.C.	1 0 4	SMITHS
CHRIS.P.	9 8	SMITHS
MICHAEL.H.	1 0 8	SMITHS
JOE.V.	9 4	SMITHS
DAMIAN.B.	1 1 1	SMITHS
MATTHEW	9 9	SMITHS
ANDREW	1 1 4	SMITHS
AARON	9 3	SMITHS
Mr Smiths	1 3 4	SMITHS
LEIGH	8 1	JACKSON
CRAIG	1 0 1	JACKSON
ADAM	1 0 6	JACKSON
JOE.C.	8 4	JACKSON
ZELKO	8 5	JACKSON
CHRIS	1 0 4	JACKSON
GUISEPPE	9 5	JACKSON
MARK	1 1 8	JACKSON
VU	9 7	JACKSON
CHAD	1 0 0	JACKSON
COREY	1 0 2	JACKSON
Mr Jackson	8 8	JACKSON

Now dealing with the Smiths class first we should rank them in order:

80, 93, 94, 94, 94, 97, 98, 98, 104, 108, 110, 111, 114, 134

Given that there are 14 pieces of data the median will be the average of the 7th and 8th item, ie 98 beats per minute. Now Q1 will be the median of the lower half of the data (note that the median is not removed here!) ie. 94 beats per minute and Q3 will be the median of the upper half ie. 110 beats per minute.

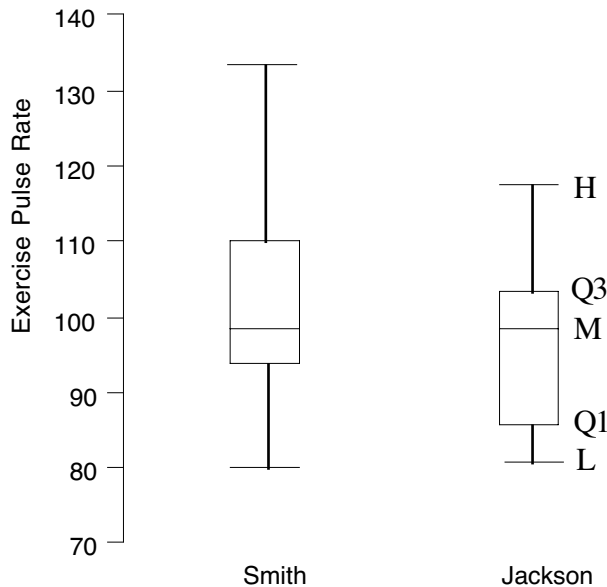
Now the Jackson class gives a median of 98.5 beats per minute, Q1 of 86.5 beats per minute and a Q3 of 103 beats per minute. As follows:

81, 84, 85, 88, 95, 97, 100, 101, 102, 104, 106, 118

Hence the FIVE NUMBER SUMMARIES for each are as follows:

Smith: 80, 94, 98, 110, 134
 and Jackson: 81, 86.5, 98.5, 103, 118

Now these five number summaries can be represented graphically with side by side boxplots as follows. Note they are drawn using a common scale:



Now these can tell an interesting story but at the same time can be a little misleading. Notice that in this case the median of the distributions are about the same, but Q1 and Q3 values differ a lot. Also notice that the lowest values are very close. With a little thought we could say the following:

The range of the first half of each data set are almost identical. However the third 25 % are spread quite differently. Smith's class from 98 to 110 beats per minute while Jackson's class ranges from 98.5 to 103. This suggests there MAY be more consistent pluse rates in this quarter in Jackson's class.

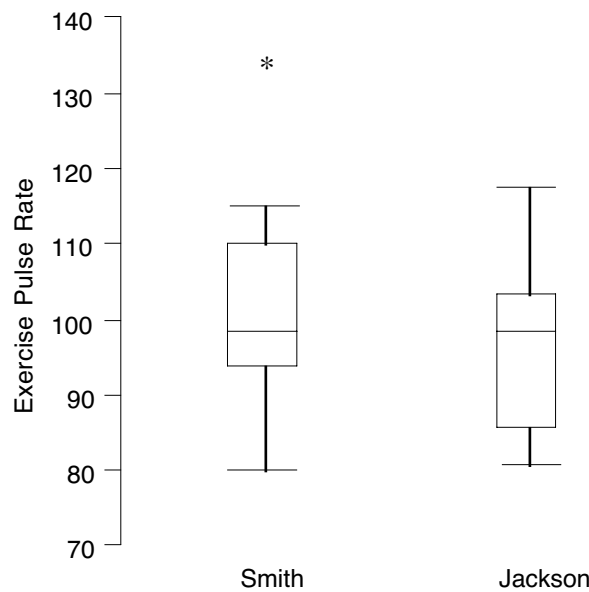
Note that just from the BOXPLOT we cannot say for sure that this is the case as it **does not show the distribution of data in between the 5 numbers** but ONLY THE 5 NO. SUMMARY. Hence to be sure we need to look at the data.

With a little more thought I could say that:

The upper 50% of the Smiths' class data APPEARS to be a lot more widely dispersed than Jackson's class. Smiths' 98 to 134 beat per minute compared to 98.5 to 118 beats per minute in Jackson's class.

However note the word 'appears'. *A boxplot in ISOLATION can be misleading.*

Looking at the data we see that the highest score in the Smiths class could be considered for outlier status - it was in fact Mr. Smith - **an unfit fat old bloke!** The next lowest was 114 beats per minute. If we considered it to be an outlier we could draw a **MODIFIED BOXPLOT**. These are exactly the same only with the outliers marked with an * and the whisker reaching to the next lowest value. eg.



This alters the picture a little! We can see that each quarter of the data sets have different spreads but the overall spread is about the same. To be sure about the quarter spreads however we would need to consult the data as, as seen before ONE data point could be the cause of what we are observing!

The moral of the story is that BOXPLOTS give us QUICK and USEFUL information about the distribution of the data that may or may not be completely correct and hence should always be substantiated or otherwise by consulting the data.



SI Exercise 1g

1. Determine the five number summary and the interquartile range for each of the following data sets that have already been placed in rank order. Then draw a boxplot for each data set
 - a) 40, 101, 134, 142, 150, 165, 222, 224, 231, 300
 - b) 5, 40, 65, 65, 65, 65, 80, 90
 - c) 12, 12, 12, 18, 18, 18, 18, 18, 30, 30, 42
 - d) 4, 8, 10, 14, 16, 18, 18, 18, 20, 40, 42, 42
2. Look closely at the boxplots you produced in Qu 1 and compare them to the data. You should be able to see how boxplots, especially with small data sets can be misleading.
2. Return to the three problems posed on pg 13 in **SI Exercise 1b**. The stemplots that you produced in that exercise can be used here as it was ranked. For each of the three problems:
 - a) determine the five number summary and interquartile range
 - b) draw side by side boxplots so that the distributions can be compared.



Producing side by side BOX PLOTS with the CASIO 9850GB PLUS



We will produce side by side boxplots on the CASIO 9850GB for the pulse data given on page 40.

Enter the STAT module.



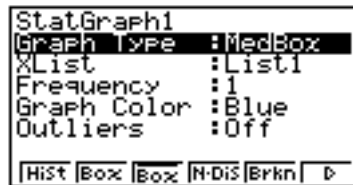
Enter Smith's classes data in List 1 and Jackson's classes data in List 2.

List 1	List 2	List 3	List 4
1	94	81	
2	94	101	
3	80	106	
4	110	84	
5	97	85	

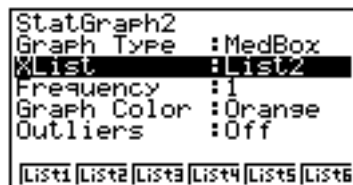
Access $\overline{\text{F6}}$ (F6) and then access GRPH (F1) and then access SET (F6). The screen opposite should be visible.



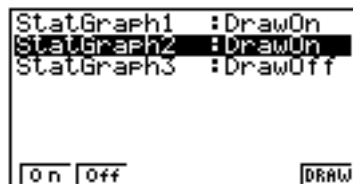
Set StatGraph 1 as shown opposite. Use the arrow keys to select each option and then the F keys to select the setting required. Note the **OUTLIER OPTION** you may choose to have it on or off.



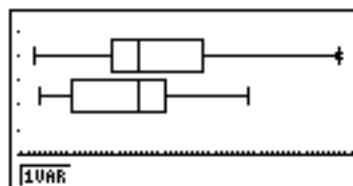
Arrow up to highlight StatGraph 1 and then select StatGraph 2 by accessing GPH2 (F2) and set up this as shown opposite.



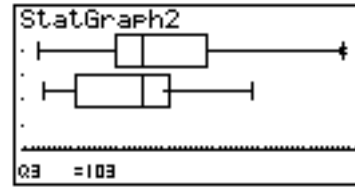
Press the black EXIT key and access SEL (F4) and make both StatGraph's 1 and 2 ON and the third one off.



Access DRAW (F6) and the following boxplots should result.

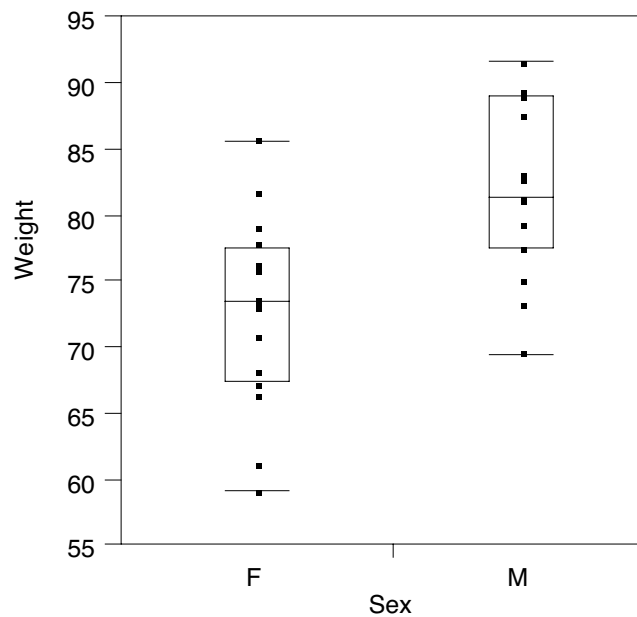


Press SHIFT and then access TRCE (F1) and then use the arrow keys to navigate the boxplots and see that the calculator will give you the values of Q1, median and so on.



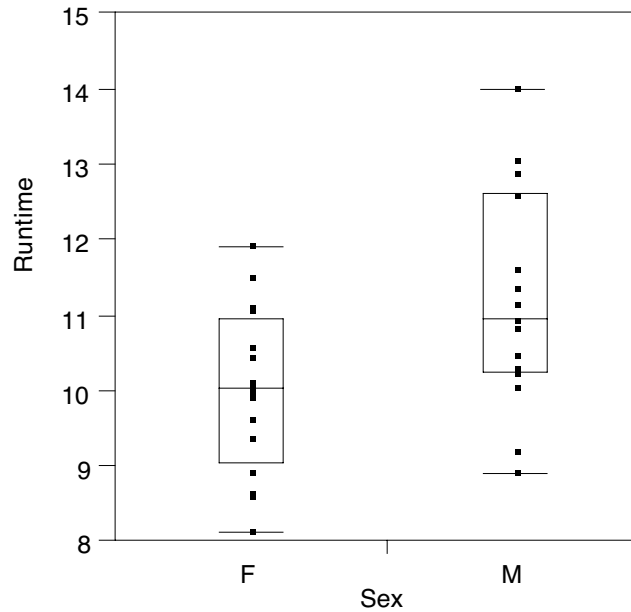
SOME COMPUTER GENERATED SIDE BY SIDE BOXPLOTS AND THE STORY THAT THEY TELL?

The following BOXPLOTS were constructed from a data set produced from a local gymnasium. Thirty one people enrolled into a fitness course (16 females). They first had their weight measured and recorded. The next activity was a 1 mile run (which was a walk for many). The people's time (minutes) for the run was measured and recorded. THE TWO GRAPHICS ARE AS FOLLOWS:



The most obvious piece of information that is revealed here is that **about 75% of the males in this group are heavier than 75% of the females in the group (as the Q3 of the females is approximately equal to the Q1 of the males)**. Given the data is illustrated as well and the distribution appears to be fairly even in each quarter we can tighten this up by saying that:

about 75% of the males in this group are CLEARLY heavier than 75% of the females in the group



Now on this graphic we see that the Q3 of the females corresponds to the median of the males and that the data is reasonably evenly distributed throughout (except perhaps for the males third quarter?) and hence we could say that :

about 50% of the males recorded clearly slower times than about 75% of the females.

In closing we should note that the statements made here are rather powerful with respect to proving a point. However, it should be stressed that on their own they are not sufficient and hence they should be incorporated with the use of other statistics and methods of descriptions learned.

The key is to put all we have learned together in an appropriate manner. Using the tools that are appropriate and leaving out those that are not.

As a general rule remember:

**GRAPHIC(S)
OUTLIERS
SHAPE
CENTRE
SPREAD.**



SI Exercise 1h

1. Recall the Brain Surgery investigation.
 - a) Produce side by side boxplots of this data. Compare their appearance to the hand drawn versions you produced in the last exercise.
 - b) Write a short paragraph in which you compare these distributions using just the boxplot as an aid.
2. Recall the 'Car' investigation
Repeat parts a) and b) of question 1 for the Car data.
3. Recall the 'Darts' investigation
Repeat parts a) and b) of question 1 for the Darts data.



Putting it all together
with the
CASIO 9850GB PLUS



It is now time to put it all together. It is first worth noting that in real situations where problems are investigated the data sets that are analysed are much larger than the ones we have worked with so far. Statisticians can not base their conclusions on only a very small sample from a population. Some data set contain thousands of pieces of data! You will learn more about this if you study Statistics in the senior school. Consider the following scenario:

A home owner was interested in whether the Sunday Mail and The Realtor (a real estate sales paper produced weekly) contained houses that were for sale for reasonably similar prices in general or whether one paper contained houses of generally higher prices. To investigate this he randomly selected 100 homes advertised in the Sunday Mail and 101 from the Realtor over a period of three weeks and recorded the asking price for each house.

I think you would agree that analysing this much data by hand would be painful, not to mention unnecessary when we have the CASIO 9850GB PLUS.

It would also be painful for everyone to type this data into their machine. Your teacher has this data entered into their calculator and if you follow the steps below, you can simply transfer the data from one calculator to another.

Join the calculators together via the cable provided by your teacher. Each end is placed in the jack at the calculators base.

On each calculator enter the LINK mode.

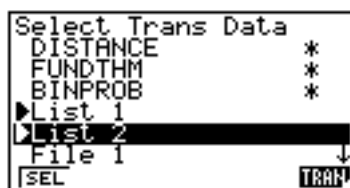


On the calculator to receive the data, simply access RECV (F2).

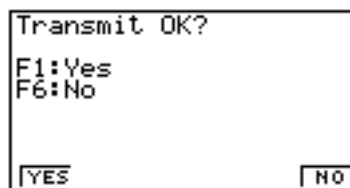
On the calculator with the data already entered, access TRAN (F1). The screen opposite will appear.



Then access SEL (F1) and then arrow down to highlight List 1, select it by accessing SEL (F1) and then repeat this for List 2.



Then access TRAN (F6) and then access YES (F1) and the transfer should occur.



The calculator that you transmitted the data to should now have the real estate data within it, Sunday Mail in list 1 and Realtor in list 2.

TIME TO PLAY LAWYER AGAIN!

Recall from our earlier work that we based our arguments around:

G.O.S.C.S.C

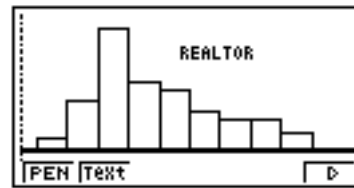
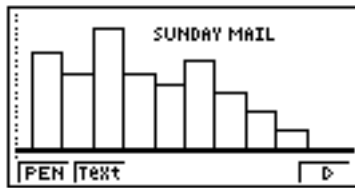
1. Produce an appropriate **G**raphic.
2. Look for **O**utliers and treat them appropriately
3. Describe/compare the **S**hape of each distribution.
4. Describe/compare the **C**entre of each distribution. [mean/median]
5. Describe/compare the **S**pread of each distribution. [IQR/range]
6. Draw your **C**onclusion.

When using the CASIO 9850GB PLUS we produce HISTOGRAMS, BOXPLOTS and SUMMARY STATISTICS for centre and spread as an initial step and then study these and finally prepare a report using the

G.O.S.C.S.C

structure.

Produce all the necessary items using the CASIO 9850GB PLUS and then compare them to the ones overleaf. Your histograms may differ a little depending on what values you gave to START and PITCH.



```
1-Variable
x̄ = 136685
Σx = 1.3668E+07
Σx² = 1.9921E+12
x̄n = 35188.534
x̄n-1 = 35365.8073
n = 100
[1VAR] [2VAR] [REG] [SET]
```

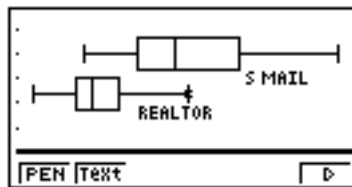
```
1-Variable
x̄ = 88506.9306
Σx = 8.9392E+06
Σx² = 8.2727E+11
x̄n = 18903.3485
x̄n-1 = 18997.6301
n = 101
[1VAR] [2VAR] [REG] [SET]
```

```
1-Variable
minX = 79950
Q1 = 109950
Med = 129950
Q3 = 164900
x̄-x̄n = 101496.466
x̄+x̄n = 171873.534
[1VAR] [2VAR] [REG] [SET]
```

```
1-Variable
minX = 51950
Q1 = 74975
Med = 84950
Q3 = 99725
x̄-x̄n = 69603.5821
x̄+x̄n = 107410.279
[1VAR] [2VAR] [REG] [SET]
```

```
1-Variable
Med = 129950
Q3 = 164900
x̄-x̄n = 101496.466
x̄+x̄n = 171873.534
maxX = 219000
Mod = 89950
[1VAR] [2VAR] [REG] [SET]
```

```
1-Variable
Med = 84950
Q3 = 99725
x̄-x̄n = 69603.5821
x̄+x̄n = 107410.279
maxX = 136950
Mod = 77500
[1VAR] [2VAR] [REG] [SET]
```



A summary table like the one below helps to give an over view before we start to write the report.

	Sunday Mail	Realtor
outliers	none	none
shape	slight skew to the high	approximately symmetrical
centre (median - as slight skewness)	\$129 950	\$84 950
spread (IQR)	\$54 950	\$24 750
boxplot story	Over three quarters of Sunday Mail prices higher than three quarter of realtor prices, around 50% Of Realtor prices less than 90% of the Sunday Mail Prices.	

A MODEL ARGUMENT FOR THE REAL ESTATE INVESTIGATION.

No abnormally high or low house prices were found in the asking prices of homes collected from either the Sunday Mail or Realtor. The Sunday Mail's price distribution is slightly skewed to the high, while that of the Realtor is approximately symmetrical. The median asking price in the Sunday Mail sample was \$129 950 compared to a much lower \$84 950 for the Realtor sample. The asking prices in the Sunday Mail sample showed far more variation than the Realtor. The interquartile ranges were \$54 950 and \$24 750 respectively. It is also worth noting that over three quarters of Sunday Mail prices higher than three quarter of realtor prices and around 50% Of Realtor prices less than 90% of the Sunday Mail Prices.

Hence we can conclude that the analysis of our samples support the hypothesis that the asking prices of houses in the Sunday Mail are, in general, considerably higher than houses advertised in the Realtor.

Remember that the aim is to build a picture of the situation in the reader's head.

It should be understood that we **have not proven** that the asking prices in the Sunday Mail are generally higher than the asking prices found in the Realtor. It is possible to prove this STATISTICALLY. This is what you will learn how to do if you study Statistics in the senior school.



SI Exercise 1i

Access your teachers calculator and transfer the data sets, 'Our pulse rates', 'Tissues' and 'Rtimes' as you need them.

1. 'Our pulse rates' is the data that we collected at the very start of this unit.
 - a) Use the CASIO 9850GB PLUS to produce histograms, boxplots and summary statistics and use them to complete as many tables similar to that one seen below as is required to analyse this data.

	Year 9	Year 12
outliers		
shape		
centre (mean or median)		
spread (IQR)		
boxplot story		

- c) Construct an argument that supports your hypothesis in regard to the differences in pulse rates of year 9 and 12 students.

2. 'Tissues' is a data set that was collected by a year 12 student investigating the strength of two different brands of tissue. One brand is an expensive brand and the other a cheap brand. The strength of each tissue tested was measured by a homemade machine that attempted to rip the tissue when weights were added. The data is the amount of weight that was added to the machine at the time the tissue ripped. Obviously the larger the weight value, the stronger the tissue.
 - a) Use the CASIO 9850GB PLUS to produce histograms, boxplots and summary statistics.
 - b) Copy and complete a table in your work book similar to the one in 1 b).
 - c) Construct an argument in which you compare the strength distributions of each type of tissue and make a hypothesis regarding the strengths of the populations of tissues that were studied.

3. 'Rtimes' is a data set that was collected by a university student investigating the reaction times of school students in different years at a given school. She tested the reaction time of reception, year 7 and 12 girls. The reaction times are measured in milliseconds and were gathered using the car simulator. The students had to put their foot on the brake pedal as soon as possible after a light was triggered.
 - a) Use the CASIO 9850GB PLUS to produce comparable histograms, boxplots and summary statistics.
 - b) Copy and complete a table in your work book similar to the one in 1 b).
 - c) Construct an argument in which you compare the reaction times of each year level and make a hypothesis regarding the reaction times of the populations of students that were studied.



INVESTIGATING A PROBLEM OF YOUR OWN

This unit of work has been designed to equip you with the skills required to carry out a simple investigation of your own design.

YOUR TASK

You are required to do the following:

1. Identify a problem/issue of interest to you that you are able to collect data about easily. The problem must involve two or more interval data sets that can be compared in order to draw a conclusion/make a hypothesis. You may like to investigate a problem/issue that arises from the wonderful *SeniorSchoolCensus-online* project – <http://www.censusonline.net>
2. After some discussion with your teacher, collect the necessary data.
3. Enter the data into the CASIO 9850GB PLUS.
4. Analyse the data appropriately.
5. Construct an argument that supports the hypothesis that you have drawn.
6. Comment on any weaknesses in the method you have employed that may cause your hypothesis to be suspect.

FORMAT OF PRESENTATION

You may choose how you wish to present your findings. It may be a newspaper or magazine article, a video (eg. a News item), a poster, a Power Point presentation or simply a traditional presentation on A-4 paper. There are some things, however, that must appear in the presentation, no matter the medium you choose. They are:

1. A description of the problem or issue your are investigating.
2. A simple account of the method you have employed to collect your data.
3. The analysis you carried out. This includes a copy of your data, any graphics and summary statistics you produced and the argument that you wrote to support your hypothesis.
4. Your hypothesis.
5. A discussion of any weaknesses in your method that may cause your conclusion to be suspect.

Above all else enjoy your investigation.

DEADLINES

Problem to be identified by:

Data to be collected by:

Data to be entered and analysed by:

Presentation to be completed by: