WTS TUTORING

DATA HANDLING

GRADE : 10 TO 12

COMPILED BY : KWV"BABE'SWEMATHS/MASTERMATHS" SIBIYA

CELL NO. : 0826727928

EMAIL : KWVSIBIYA@GMAIL.COM

FACEBOOK P. : WTS MATHS & SCEINCE TUTORING

Data Handling

Data refers to the pieces of information that have been observed and recorded, from an experiment or a survey.

Ungrouped data

- Here we use the individual scores that are recorded.
- They need to be ranked in ascending order of size.
- However, if the data base is large, this is cumbersome,
- and it is difficult to analyse the data, so the stem and leaf is used to arrange the data.

The stem and leaf diagram

The data is listed in intervals that depend on the place value of the digits of each data.

Note:

- The leaf is the units digit-i.e. furthest to the right in the number.
- The stem is the tens/hundreds or thousands digit
- Back to back

Kwv 1

A farmer in the Free State has 32 cattle to sell. Their weights in kilograms (kg) are:
81; 81 ;82; 82; 83; 84; 84; 85; 85; 86; 86; 87; 87; 88; 89; 90; 92; 92; 93; 94; 96;
150; 152; 153; 154; 320; 375; 376; 380; 381; 390.
Consider the data above and arrange it in the stem and leaf.

Grouping Data

- One of the first steps to processing a large set of raw data is to arrange the data values together into a smaller number of groups,
- and then count how many of each data value there are in each group.
- The groups are usually based on some sort of interval of data values, so data values that fall into a specific interval, would be grouped together.

- The grouped data is often presented graphically or in a frequency table. (Frequency means "how many times")
- Note that $n = \varepsilon f$

Kwv 1

The height of 30 learners are given below. Fill in the grouped data below.

(Tally is a convenient way to count in 5's. We use llll to indicate 5.)

142 163 169 132 139 140 152 168 139 150 161 132 162 172 146 152 150 132 157 133

141 170 156 155 169 138 142 160 164 168

Group	Tally	Frequency	Midpoint(X)	F.X
$130 \leq h < 140$				
$140 \leq h < 150$				
$150 \leq h < 160$				
$160 \leq h < 170$				
$170 \leq h < 180$				

Graphical Representation of Data

- Once the data has been collected, it must be organised in a manner that allows for the information to be extracted most efficiently.
- One method of organisation is to display the data in the form of graphs.
- Bar graphs, histograms and pie charts will be drawn directly from the data.

Bar and Compound Bar Graphs

- A bar chart is used to present data where each observation falls into a specific category.
- The frequencies (or percentages) are listed along the y-axis and the categories are listed along the x-axis.

- The heights of the bars correspond to the frequencies.
- The bars are of equal width and should not touch neighbouring.
- A compound bar chart (also called component bar chart) is a variant: here the bars are cut into various components depending on what is being shown.
- If percentages are used for various components of a compound bar, then the total bar height must be 100%.
- The compound bar chart is a little more complex but if this method is used sensibly, a lot of information can be quickly shown in an attractive fashion.

Histograms and Frequency Polygons

- It is often useful to look at the frequency with which certain values fall in pre-set groups or classes of specified sizes.
- The choice of the groups should be such that they help highlight features in the data.
- If these grouped values are plotted in a manner similar to a bar graph, then the resulting graph is known as a histogram.
- The same data used to plot a histogram are used to plot a frequency polygon, except the pair of data values are plotted as a point and the points are joined with straight lines.
- Unlike histograms, many frequency polygons can be plotted together to compare several.
- frequency distributions, provided that the data has been grouped in the same way and provide a clear way to compare multiple datasets.

Pie Charts

- A pie chart is a graph that is used to show what categories make up a specific section of the data,
- and what the contribution each category makes to the entire set of data.
- A pie chart is based on a circle, and each category is represented as a wedge of the circle.

Method: Drawing a pie-chart

1. Draw a circle that represents the entire data set.

2. Calculate what proportion of 360 degrees each category corresponds to according to Angular Size

3. Draw a wedge corresponding to the angular contribution.

4. Check that the total degrees for the different wedges add up to close to 360 degrees.

The graphs drawn from the ungrouped or raw data

Line and Broken Line Graphs

- All graphs that have been studied until this point (bar, compound bar, histogram, frequency polygon and pie) are drawn from grouped data.
- Line and broken line graphs are plots of a dependent variable as a function of an independent variable, e.g. the average global temperature as a function of time, or the average rainfall in a country as a function of season.
- Usually a line graph is plotted after a table has been provided showing the relationship between the two variables in the form of pairs.
- Just as in (x; y) graphs, each of the pairs results in a specific point on the graph, and being a line graph these points are connected to one another by a line.
- Many other line graphs exist; they all connect the points by lines, not necessarily straight lines.
- Sometimes polynomials, for example, are used to describe approximately the basic relationship between the given pairs of variables, and between these points.

➢ More kwvs

Graphical Representation of Data

1. Represent the following information on a pie chart.

Walk	15
Cycle	24
Train	18
Bus	8
Car	35
Total	100

2. Represent the following information using a broken line graph.

Time	07h00	08h00	09h00	10h00	11h00	12h00
Temp (_C)	16	16,5	17	19	20	24

3. Represent the following information on a histogram. Using a coloured pen, draw a frequency polygon on this histogram.

Time in seconds	Frequency
16 - 25	5
26 - 35	10
36 - 45	26
46 - 55	30
56 - 65	15
66 - 75	12
76 – 85	10

4. The maths marks of a class of 30 learners are given below, represent this information using a suitable graph.

82 75 66 54 79 78 29 55 68 91 43 48 90 61 45 60 82 63 72 53 51 32 62 42 49 62 81 49 61 60 224

5. Use a compound bar graph to illustrate the following information

Year	2003	2004	2005	2006	2007
Girls	18	15	13	12	15
Boys	15	11	18	16	10

Summarising Data

If the data set is very large, it is useful to be able to summarise the data set by calculating a few quantities that give information about how the data values are spread and about the central values in the data set.

Measures of Central Tendency

Mean or Average

- The mean, (also known as arithmetic mean), is simply the arithmetic average of a group of numbers (or data set) and is shown using the bar symbol.
- So the mean of the variable x is $\bar{x}$ pronounced "x-bar".
- The mean of a set of values is calculated by adding up all the values in the set and dividing by the number of items in that set.
- The mean is calculated from the raw, ungrouped data.

Method: Calculating the mean

1. Find the total of the data values in the data set.
2. Count how many data values there are in the data set.
3. Divide the total by the number of data values.

FORMULA:

Kwv 1

What is the mean of x = {10, 20 , 30 , 40 ,50}?

Median

- The median of a set of data is the data value in the central position, when the data set has been arranged from highest to lowest or from lowest to highest.
- There are an equal number of data values on either side of the median value.

Ungrouped data

Method: Calculating the median

1. Order the data from smallest to largest or from largest to smallest.
2. Count how many data values there are in the data set.
3. Find the data value in the central position of the set.

Kwv 1

What is the median of {10, 14, 86, 2, 68, 99, 1}?

This example has highlighted a potential problem with determining the median. It is very easy to determine the median of a data set with an odd number of data values, but what happens when there is an even number of data values in the data set? When there is an even number of data values, the median is the mean of the two middle points.

Median

- An easy way to determine the central position or positions for any ordered data set is to take the total number of data values, add 1, and then divide by 2.
- If the number you get is a whole number, then that is the central position.
- If the number you get is a fraction, take the two whole numbers on either side of the fraction, as the positions of the data values that must be averaged to obtain the median.

Kwv 1

What is the median of {11, 10, 14, 86, 2 , 68 , 99 ,1 }?

Mode

- The mode is the data value that occurs most often, i.e. it is the most frequent value or most common value in a set.

Method: Calculating the mode

For ungrouped

- Count how many times each data value occurs.
- The mode is the data value that occurs the most.

For grouped

- Simple look at the interval with the higher frequency
- It is referred as modal class

.

Kwv 1

Find the mode of the data set x = {1, 2, 3, 4, 4, 4, 5, 6, 7, 8, 8, 9, 10 , 10}

Measures of Dispersion

- The mean, median and mode are measures of central tendency, i.e. they provide information on the central data values in a set.
- When describing data it is sometimes useful (and in some cases necessary) to determine the spread of a distribution. Measures of dispersion provide information on how the data values in a set are distributed around the mean value.
- Some measures of dispersion are range, percentiles and quartiles.

Range

- The range of a data set is the difference between the lowest value and the highest value in the set.

Method: Calculating the range

1. Find the highest value in the data set.
2. Find the lowest value in the data set.
3. Subtract the lowest value from the highest value. The difference is the range.

Quartiles

- Quartiles are the three data values that divide an ordered data set into four groups containing equal numbers of data values.
- Lower quartile Q_1 : P = $\frac{n+1}{4}$
- The lowest 25% of the data being found below the first quartile value
- The median is the second quartile Q_2 : P = $\frac{n+1}{2}$
- The median, or second quartile divides the set into two equal sections.
- Upper quartile Q_3 : P = $\frac{3(n+1)}{4}$
- The lowest 75% of the data set should be found below the third quartile
- Note: for the grouped data you must remove 1 for positions

Kwv 1

What are the quartiles of {3, 5, 1, 8, 9, 12, 25, 28, 24, 30, 41, 50}?

Inter-quartile Range

- The inter quartile range is a measure which provides information about the spread of a data Set.
- and is calculated by subtracting the first quartile from the third quartile, giving the range of the middle half of the data set, trimming off the lowest and highest quarters, i.e. Q3 − Q1.
- The semi-interquartile range is half the interquartile range, i.e. Q3−Q1

Kwv 1

Question: A class of 12 students writes a test and the results are as follows:
20, 39, 40, 43, 43, 46, 53, 58, 63, 70, 75, 91.
Find the range, quartiles and the Interquartile Range.

Percentiles

- Percentiles are the 99 data values that divide a data set into 100 groups.
- The calculation of percentiles is identical to the calculation of quartiles,
- except the aim is to divide the data values into 100 groups instead of the 4 groups required by quartiles.
- Position : $P = \frac{r(n+1)}{100}$
- r stand for the percentage given.

Method: Calculating the percentiles

1. Order the data from smallest to largest or from largest to smallest.
2. Count how many data values there are in the data set.
3. Divide the number of data values by 100. The result is the number of data values per group.
4. Determine the data values corresponding to the first, second and third quartiles using the number of data values per quartile.

Five number summary

1. Minimum value

It the smallest number that occurs in the data set.

2. Maximum value

It is a greatest number that occurs in the data set.

3. The median

The median is the middle most number when the data is arranged from smallest to greatest.

Note:

- For even data
- For odd data

4. The lower quartile

It is a lower halve of the data from median.

5. The upper quartile

It is an upper halve of the data from median

The upper and lower quartiles are the median of the upper

Note: the data must be ordered

Outliers:

- A point or score which is widely separated from the other points or scores, this is mostly applicable to the plot called scatter plot and Box and Whisker.
- To check the outlier : [$Q_1 - 1{,}5 \times IQR$; $Q_3 + 1{,}5 \times IQR$]

Kwv 1

Consider the set {12, 13, 18, 20, 22, 27, 29}.

Find the 5 number summary.

Box and Whisker Plot

- Box and Whisker Plots allow us to interpret the spread of the data more easily.
- The Box is the part from the lower quartile to the upper quartile
- and the whiskers are the lines on either end of the box.
- The end point of the whiskers gives us the minimum and maximum values.

NB; it focuses on the spread around the median

Percentages in the box plot

- It is very important to note that the first 25% (first quarter) of results lies between the minimum and the lower quartile.
- The next 25% (second quarter) of results lies between the lower quartile and the median.
- The third quarter lies between the median and the upper quartile and the last quarter of data lies between the upper quartile and the maximum value.

Kwv 1

Consider the data below:

8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31,

Draw box and whisker plot.

Interpreting the box and whisker plot

Shape of a data set; This describes how the data is distributed relative to the mean and median.

Positively skewed:

If the mean > median then the data is positively skewed (skewed to the right). This means that the median is close to the start of the data set.

Negatively skewed:

If the mean < median then the data is negatively skewed (skewed to the left). This means that the median is close to the end of the data set.

Symmetrically skewed:

- Symmetrical data sets are balanced on either side of the median (spread fairly evenly).
- If the mean, median, and mode are approximately equal to each other, the distribution can be assumed to be approximately symmetrical.
- With both the mean and median known, the following can be concluded:
  mean = median then the data is symmetrical

NB: The longer whisker shows the greater variability and/or spread.

Ogive Curves (Cumulative Frequency curves)

In mathematics, the name ogive is applied to any continuous cumulative frequency polygon.

Note:

- The Cumulative Frequency is the sum of all the frequencies within a specific interval or boundary.
- Use Z shape to write cumulative frequency from frequency
- Moving from cumulative to frequency use: $f_n = cf_n - cf_{n-1}$
- Every interval always starts at the lower band.
- The Cumulative Frequency table is obtained from the frequency table.
- The sum of all the frequencies is always equal to the Cumulative Frequency value.
- The last number on the cumulative will give the total frequency

Drawing of Ogive

- 7 shape is used to locate the points
- To plot the graph we plot the cumulative frequency value against the end point value (x-value) for each interval.
- Use a smooth, continuous curve.

Note:

One extra point is obtained by plotting (x; 0), x is the lower boundary of the lowest class interval. This is done because all the values must lie above x.

Frequency table

Intervals	Frequency	Cumulative

- ✓ **Finding the Lower Quartile, Median and Upper Quartile using an ogive curve**

Note;

- Find the position of each and then use the cumulative frequency.
- Cumulative frequency curves make it very simple to answer questions that involve "less than" or "more than".
- For box and whisker: the right lower interval indicate the minimum value and the right upper interval indicate the maximum value

✓ **Finding the mean and standard deviation using an ogive curve**

Add all the y-values of the cumulative frequency and the divide it by number of point.
Note: that the range can also be found.

Interpreting the Ogive

- To interpret you will need to use : $cf_f - cf_i$

Using the histogram

Note:
The bars 'touch' meaning that we are working with continuous data.
How to complete the Cumulative frequency table from the histogram given.

Standard deviation and Variance

- The variance and the standard deviation are measures of how spread out a set of data is.
- In other words, they are measures of variability. it is a measure of the average distance between the values of the data in the set and the mean.
- If the data values are all similar, then the standard deviation will be low (closer to zero).
- If the data values are highly variable, then the standard variation is high (further from zero).
- The standard deviation is always a positive number and is always measured in the same units as the original data.

- For example, if the data are distance measurements in metres, the standard deviation will also be measured in metres.
- Standard deviation is directly proportional to mean.
- If the data is more closed together the Standard deviation and mean

The variance: is the average squared deviation of each number from the mean.

The standard deviation: it is the square root of the variance.

NB; it focuses on the spread around the mean

➢ Calculations

✓ Two methods

1. Pen and paper method

2. A calculator method

Method 1

Manual calculation for finding σ – the standard deviation

For ungrouped:

Table:

X(score)	(x - x)	$(x - x)^2$

Steps

1. Find (The mean average).

2. Subtract the mean from each of your values. (Column2).

3. Square each of the results (Column 3).

4. Add all the values in column 3, and divide by the total number of original values. i.e.: find the average of column 3. This answer is the variance.

5. To find the standard deviation, σ, square root the answer found in step 4.

For grouped data

Table:

Intervals	f(frequency)	x(midpoint)	f.x	(x - x)	$(x - x)^2$	$f(x - x)^2$

The table used for a set of grouped data is slightly different as the frequency has to be taken into account now.

Steps

1. Find the mean:

- the midpoint from intervals
- multiply each midpoint by the frequency

2. Subtract the mean from each of your values (Column 4).

3. Square each of the results (Column 5).

4. Multiply each square with the frequency (Column 6).

5. Add all the values in column 6, and divide by the total number of original values. i.e.: find the average of column 6. This answer is the variance.

6. To find the standard deviation, σ, square root the answer found in step 5.

Method 2

Using the calculator

The variance and/or standard deviation can be calculated easily with a calculator:

Although you are encouraged to use a calculator to calculate the standard deviation, you must also be able to perform this calculation manually.

For ungrouped data

CASIO *fx* – ES PLUS to demonstrate this:

Steps

Step 1: Press "SET UP". Select 2: STAT

Step 2: Press 1: 1 – VAR

Step 3: Enter the numbers one by one followed by the equals after each number.

Once you have completed entering all the data as described in step 3, press the AC button once.

Step 4: Press the " shift" button and then the "1" button. (Notice that Mean is also an option here, so you can use your calculator to determine the mean.)

Select 4: VAR

Step 5: Select 3: $x\sigma n$ and then the "=" button.

Kwv 1

Finding the standard deviation manually for 5 test scores:

62% ; 80% ; 71% ; 51% ; 86%

For grouped data

- Note that when using the calculator be sure to put the frequency mode on.
- On the **CASIO *fx* – 82ES PLUS**, this is done by pressing "SHIFT" then "SET UP".

- Scroll down and select 3:STAT. Then select 1: ON.
- Use midpoints as the x- values
- Then use the frequency

Kwv 1

The shoe sizes of a group of 50 Grade 11 students were recorded and summarised in the table below:

Shoe size	4	5	6	7	8	9	10	11	12
Frequency	2	4	4	8	7	12	10	2	1

Calculate the standard deviation

1. Using a table and
2. Using a calculator

- **Variation**

Note:

- Within one/ two standard deviation intervals (max/min)
- How to calculate % of standard deviation intervals
- Range is directly proportional to standard deviation.
- The larger the standard deviation, the greater the variability of the data (the greater the spread of the data)
- A large standard deviation indicates that the data values are far from the mean and a small standard deviation indicates that they are clustered closely around the mean.
- Standard deviation with one $(x - \alpha ; x + \alpha)$
- Standard deviation with one $(x - 2\alpha ; x + 2\alpha)$

Scatter Plots

- In science, the scatter plot is widely used to present measurements of two or more related variables.
- We say this is **bivariate data.**

- It is particularly useful when the variables of the y-axis are thought to be dependent upon the values of the variable of the x-axis (usually an independent variable).
- Normal the first row indicate the x-axis

Note:

- The data points are plotted but joined the resulting pattern indicates the type and strength of the relationship between two or more variables

The line of the best fit / regression line

- is the line drawn with the aim of having the same number of points above the line as below the line in grade 11
- in grade 12 we use the equation $y = a + bx$
- a - represent y- intercept
- b – represent the gradient or slope
- if x is given then substitute to get y or visa verse

How to calculate the equation

Casio

- MODE 2
- PRESS 2: A + BX
- ENTER DATA POINTS
- THEN PRESS AC
- THEN PRESS SHIFT 1
- THEN PRESS 5 : REG
- THEN PRESS 1 : A
- THEN PRESS SHIFT 1
- THEN PRESS 5 : REG
- THEN PRESS 2 : B
- THEN PRESS AC
- THEN PRESS MODE 1 TO GET BACK TO NORMAL MODE

NOTE:

- Able to state whether a trend is linear, quadratic (parabola) or exponential.
- Outlier: a disable point(s) or incorrectly recorded

To draw the line

- Calculate the x and y mean to create the point
- And then use the value of a (0: a)
- Then join the two points

Interpretation

- Extrapolation : estimating outside the given domain
- Interpolation: estimating inside the given domain

Relationship (correlation)

- ✓ **Positive trend/gradient**: as the variable on the x-axis increases, the variable on the y-axis also increases.
- ✓ **Negative gradient/trend**: as the variable on the x-axis increases, the variable on the y-axis decreases.

Correlation (r)

- ✓ It is a value that give an indication of the strength of the association
- ✓ $r > 0$ means positive association
- ✓ $r < 0$ means negative association
- ✓ If the points on the scatter plot are close to the line of best fit, we have a strong correlation or association between the two variables.
- ✓ If the points are not so close to the line of best fit, we have a weak correlation or association between the two variables.