Understanding Descriptive Statistics: A college based approach in Zimbabwe

Edwin Rupi

Department of Mathematics and Science
Masvingo Teachers' College
P.O Box 760
Masvingo, Zimbabwe
[1]*edwinrupi@gmail.com*, [2]*0775 333 523, 0714669997*

Abstract

The main focus of this paper was to describe how much, data values, spread out from one another and how they relate to each other. Calculations of variance and standard deviation helped describe the nature of scores recorded for classroom situations. Just like the stem and leaf plot we saw from the box and whisker plot that the data are either negatively or positively skewed depending on the length of whiskers to the left or right. The box and whisker plot was used to identify some outliers, that is, values more extreme than the whiskers. The box plot is thus useful in spotting data dispersion, screening data or singling out potential problem data before any analysis. The scatter diagram proved more important in trying to describe relationships between two variables measured on the same subject. This article also looked at the two major types of correlation coefficients (Pearson and Spearman) and their calculations. The meaning of each value was also scrutinised stemming from its size.

1.0. Population and Sample

A population is a totality of all observations under consideration for example all college students in Zimbabwe or all animals in Masvingo. A sample is a part of the population selected for study for example students at Masvingo teachers' college or pupils in a particular class and grade at a school. It is from the sample that conclusions or generalisations about a population can be made. For practical purposes any number of observations less than 30 is considered a sample unless stated otherwise. The reasons for working with a sample in an investigation include reduced costs and manageability of the smaller group. Note that the selected group should represent the population as much as possible.

1.1. Variance

In this paper I am going to look at the mean and variance as well as the standard deviation of data that is ungrouped. When a large data set has been collected say for marks of students of a particular stream interest is usually not in the individual marks but otherwise the particular descriptive quantities like the average (mean), median, variance or standard deviation.

In statistics, **variance** measures how far a set of numbers are spread out. A variance of zero indicates that all the values are identical. Variance is always non-negative. A small variance indicates that the data points tend to be very close to the mean (expected value) and hence to

each other, while a high variance indicates that the data points are so much spread out around the mean and from each other.

1.2. Standard deviation

An equivalent measure is the square root of the variance, called the Standard Deviation. The Standard Deviation is a measure of how spread out numbers are. Its symbol is **σ** (the Greek letter sigma), for a population and *s* for a sample. The standard deviation has the same dimension as the data, and hence is comparable to deviations from the mean.

If all possible observations of the system are present then the calculated variance is called the population variance (σ^2). Normally, however, only a subset is available, and the variance calculated from this is called the sample variance (s^2).The variance calculated from a sample is considered an estimate of the full population variance. There are multiple ways to calculate an estimate of the population variance. We calculate it from

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

$$= \frac{1}{N}\sum_{i=1}^{N}(x_i^2 - 2x\bar{x} + \bar{x}^2)$$

$$= \frac{1}{N}\left[\sum_{i=1}^{N} x_i^2 - 2\bar{x}\sum_{i=1}^{N} x_i + \sum_{i=1}^{N}\bar{x}^2\right]$$

$$= \frac{1}{N}\left[\sum_{i=1}^{N} x_i^2 - 2\bar{x}(N\bar{x}) + N\bar{x}^2\right]$$

Since $\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$ *and so* $\sum_{i=1}^{N} x_i = N\bar{x}$

$$= \boldsymbol{\frac{1}{N}\left[\sum_{i=1}^{N} x_i^2 - N\bar{x}^2\right]}$$

$$= \boldsymbol{\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{x}^2}$$

And the corresponding sample variance is given by

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x - \bar{x})^2$$

Note the difference between the two formulae and attention should be paid to the situation in which each will be used.

Example

The following marks were obtained by five students in a class of 25 in a Statistics test

45, 63, 76, 69 and 58. Calculate the standard deviation and comment on the data.

Solution: Note that this is a sample hence

$$\bar{x} = \frac{45 + 63 + 76 + 69 + 58}{5} = 62.2$$

$$s^2 = \frac{1}{5-1}\{(45 - 62.2)^2 + (63 - 62.2)^2 + (76 - 62.2)^2 + (69 - 62.2)^2 + (58 - 62.2)^2\}$$

$$= \frac{1}{4}\{295.84 + 0.64 + 190.44 + 46.24 + 17.64\}$$

$$= \frac{1}{4}\{550.8\}$$

$$= 137.7$$

The standard deviation is given by s = $\sqrt{137.7}$ = 11.73 which is a too high value indicating a greater difference in the marks obtained by the students.

2.0. Graphical displays

2.1. Stem and leaf diagram

A **Stem-and-leaf display** is a device for displaying quantitative data in a graphical way similar to a histogram in an attempt visualizes the shape of a distribution.

A basic stem-and-leaf display contains two columns separated by a vertical line. The left column contains the *stems* and the right column contains the *leaves*. The stems are listed to the left of the vertical line. . It is important that each stem is listed only once and that no numbers are skipped, even if it means that some stems have no leaves. Typically, the leaf contains the last digit of the number and the stem contains all of the other digits. The leaves are listed in increasing order in a row to the right of each stem.

In the case of very large numbers, the data values may be rounded to a particular place value (such as the hundreds place) that will be used for the leaves. The remaining digits to the left of the rounded place value are used as the stem.

To construct a stem-and-leaf display, the observations must first be sorted in ascending order: this can be done most easily if working by hand by constructing a draft of the stem-and-leaf display with the leaves unsorted, then sorting the leaves to produce the final stem-and-leaf display. Here is the sorted set of data values that will be used in the following example:

Example 1

44 48 67 41 59 73 69 76 84 66 73 92 86 76 81 94 88 97 94 102 69 69 52 106 68 78 69 70 90 105 66 42 82 54 64 88 68 66 100 73 41 109 56 84 79

You must be determined what the stems will represent and what the leaves will represent.

In this example, the leaf represents the ones place and the stem will represent the rest of the number (tens place and higher).

It is important to note that when there is a repeated number in the data (such as three 66s) then the plot must reflect such (so the plot would look like 6 | 6 6 6 7 when we have the list of numbers 66, 66, 66, 67)

4 | 1 1 2 4 4 8
5 | 2 6 9
6 | 4 6 6 6 7 8 8 9 9 9 9
7 | 0 3 3 3 6 6 8 9
8 | 1 2 4 4 6 8 8
9 | 0 2 4 4 7
10 | 0 2 5 6 9
key: 7|3 = 73

Stems	Leaves
15	1
14	0 0 5
13	2 5 7 9
12	1 6 8
11	4 5 6 7 8 9
10	1 3 3 5 8 8 8 8 9
9	0 0 0 4 4 7 8 8
8	1 1 4 7 8
Key: 16	7 represents a score of 167

The stem and leaf diagram above shows the heights of pupils in a class measured in centimetres. We can see the list has 81, 81, 84, 87, 88, 90, 90, 90, 94, 94, 97, 98, 98, 101, 103, 103, 105, 108, 108, 108, 109.114, 115, 116, 117, 118, 119, 121, 126, 128, 132, 135, 137, 139, 140, 140, 145, 151.

If the list was not arranged in order of size it would be a challenge to identify measures like the mode, median, range and inter-quartile range. The stem and leaf display is very useful in this regard. In the display above

(I) Mode = 108

(II) Median = 108
(III) Range =151 – 81 = 70

The distribution of the heights on the diagram goes even further to suggest the kind of symmetry of the scores.

2.1.1. Advantages of the stem and leaf display

(a). Unlike histograms, stem-and-leaf displays retain the original data with perfect integrity and puts the data in order.

(b). Stem-and-leaf displays are useful for displaying the relative shape of the data, giving the reader a quick overview of the distribution.

(c). They are also useful for highlighting outliers and finding the mode.

2.1.2. Disadvantages of stem and leaf display

(a). Stem-and-leaf displays are only useful for moderately sized data sets (around 15-150 data points). With very small data sets a stem-and-leaf displays can be of little use, as a reasonable number of data points is required to establish definitive distribution properties.

(b). With very large data sets, a stem-and-leaf display will become very cluttered, since each data point must be represented numerically.

2.2. Box and Whisker Plot

A Box and Whisker plot becomes more suitable for displaying data values as the data size increases. The plot is a convenient way of graphically depicting groups of numerical data through their quartiles. Box plots may also have lines (whiskers) extending horizontally or vertically from the boxes indicating variability outside the upper and lower quartiles, hence the terms **Box-and-Whisker Plot** and **Box-and-Whisker Diagram**. The whiskers are the lines that extend to the smallest and largest data point. Outliers may be plotted as individual points.

It is also possible to get a sense of the data's distribution by examining **five statistical summary measures** (or five number summary),

(1) The minimum,

(2) The maximum,

(3) The median (or second quartile or 50th percentile) = $Q_2 = \frac{1}{2}(n+1)^{th}\ term$,

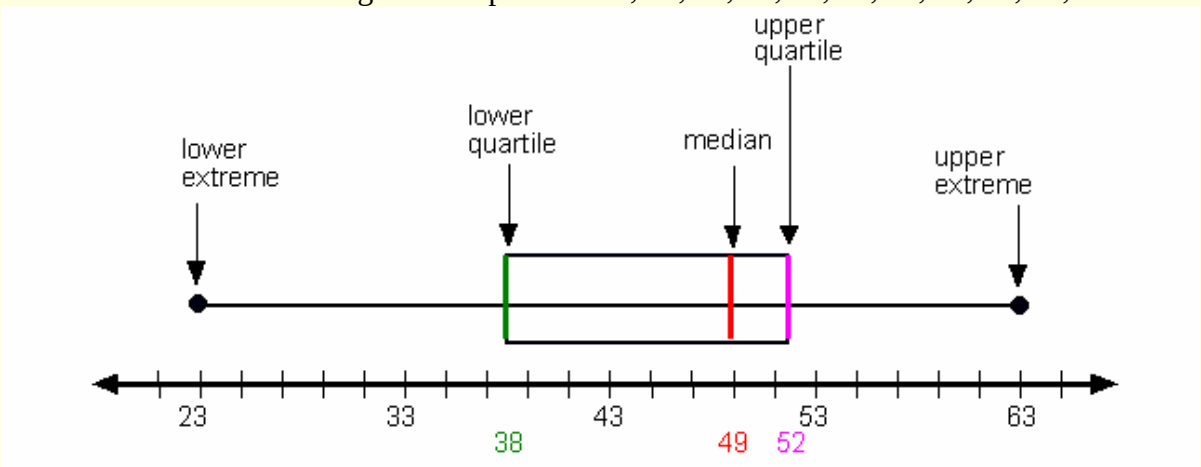
(4) The first quartile (25th percentile) =$Q_1 = \frac{1}{4}(n+1)^{th}\ term$, and

(5) The Upper quartile (75th percentile) = $Q_3 = \frac{3}{4}(n+1)^{th}\ term.$

(6) Inter-quartile range = $Q_3 - Q_1$

Such information will show the extent to which the data is located near the median or near the extremes. The first and third quartiles are at the ends of the box, the median is indicated with a vertical line in the interior of the box, and the maximum and minimum are at the ends of the whiskers.

Draw a box and whisker diagram to represent :23, 33, 38, 42, 45, 49, 51, 51, 52, 58, 63



In the diagram above the minimum value is 23, the maximum being 63 while Q_1 , Q_2 and Q_3 are 38, 49 and 52 respectively.

The inter-quartile range here is 52 – 38 = 14. Outliers are points lying beyond 1.5 x IQR and 3 x IQR from $Q_1$and Q_3 respectively while extreme outliers are points lying beyond 3 x IQR from Q_1 and Q_3. Spatz (2011) views outliers as scores that are unusually too small or too large in a given distribution. Such scores may reflect an error in measurement during recording or during data entry. In the example above scores beyond 21(1.5x14) and 42(3x14) from Q1 and Q3 respectively are the outliers. Hogan and Evalenko (2006) propose an easier way of identifying the lower and upper outliers ,that is,
Lower outlier =$Q_1 - (1.5xIQR)$ and
Upper outlier =$Q_3 + (1.5xIQR)$
In a research outliers may produce misleading results and so once identified these can be discarded. Therefore the researcher should decide on whether to retain or discard on the basis of desired results. Fortunately for this distribution there are no outliers.

3.0.Scatter diagram

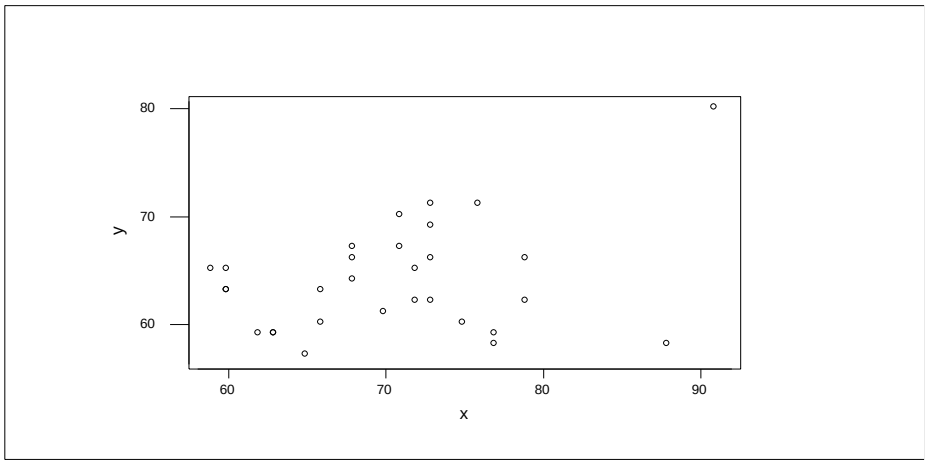
A scatter diagram , sometimes referred to as a scatter plot, or scatter-graph is a type of mathematical diagram that employs the Cartesian coordinates to display data values for two variables measured on the same set of individuals, that is, (x_1, y_1) , (x_2, y_2) , ... ,(x_n, y_n). The predictor variable (independent variable) is normally presented on the x-axis while the dependent variable is shown on the y-axis. The data is displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis (x-axis) and the value of the other variable determining the position on the vertical axis (y-axis). Any study of correlation, which will be discussed in next section, should always start with a scatter diagram. Thus it is much easier to visualise the relationship between the two variables on the diagram.

A scatter diagram helps figure out the possible kinds of correlations (relationships) that exists between variables as well as establishing whether there is a linear or non-linear relationship between the variables. For example, weight and height, weight would be on y axis and height would be on the x axis. The relationship could be positive (rising), negative (falling), or null (showing no relationship at all). If the points plotted on the Cartesian plane slope from lower left to upper right, the variables are positively correlated and the strength of the relationship depends on how close points are to the line of best fit. If the pattern of dots slopes from upper left to lower right, it indicates a negative correlation (relationship).

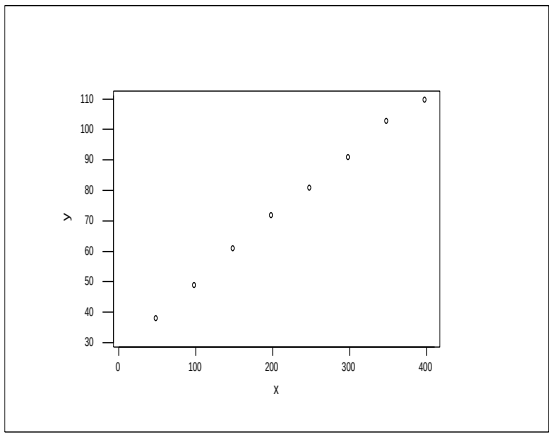
I will illustrate correlation with an example below.
In a study of student performance, 30 students had their coursework marks (x) recorded and examination marks (y) recorded in a particular year. The results are shown in the table below.

Student	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Examination mark (x)	80	65	62	59	59	69	66	71	59	66	70	66	61	63	67
Coursework mark (y)	91	70	73	62	63	74	73	79	63	71	68	71	70	75	60
Student	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
Examination mark (x)	59	67	65	58	65	62	60	57	63	62	64	58	71	63	60
Coursework mark (y)	68	71	72	60	77	72	72	65	72	60	69	68	78	66	66

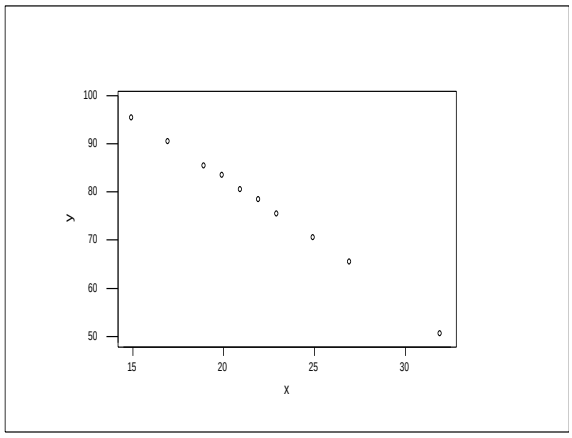
A scatter plot of the two variables is shown below



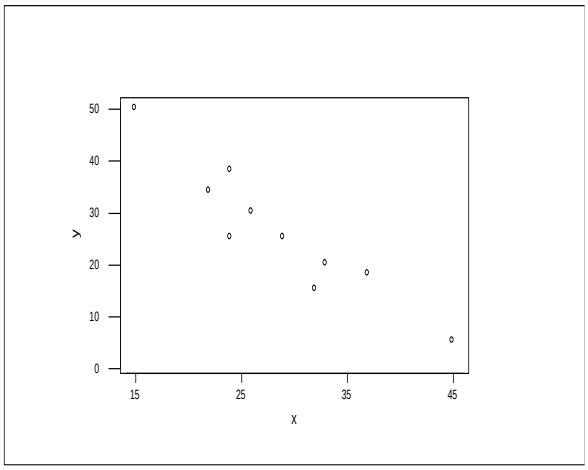
The scatter plot above was done using Minitab and can equally be done using other software or manually on graph paper. A first impression of the points shows some positive correlation between the examination mark and coursework mark. However we cannot specify with some degree of confidence the strength of the relationship. A scatter diagram makes it particularly easy to spot trends and correlations between the two variables



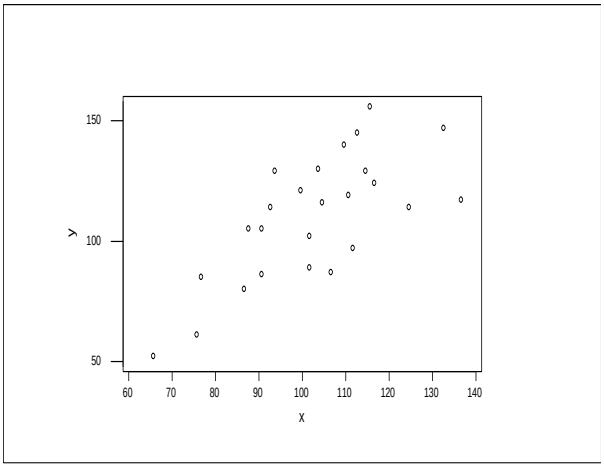
(a).Very strong positive correlation



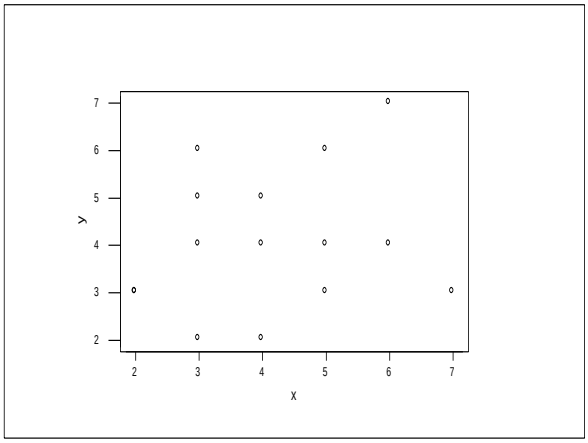
(b). Very strong negative correlation



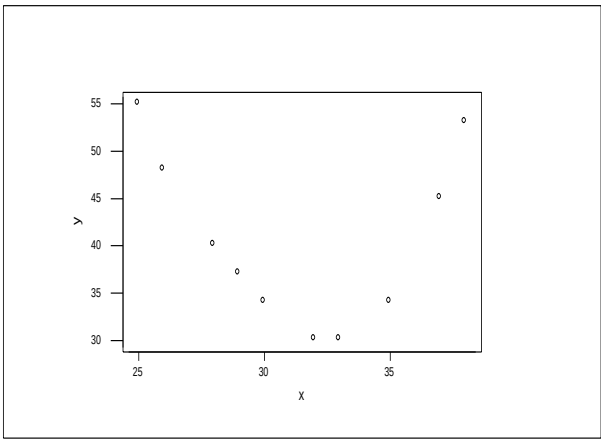
(c). Weak negative correlation



(d). Weak positive correlation



(e). Little or no correlation



(f).Non linear correlation

A random scatter indicates a weak to non-existent relationship between the two variables (e). In other words one cannot predict with some certainty what one variable would be when the other is known. In the case of a perfect or near perfect positive relationship (a), as on quantity increases the other increases in a almost predictable fashion. The relationship is negative when one variable decreases when the other increases (b). In some cases a perfect or near perfect negative relationship does exist. The quantities in (c) and (d) above have weak negative and weak positive relationships respectively.

4.0. Correlation

Correlation refers to a broad class of statistical relationships between two random variables or two sets of data. It shows whether and how strongly pairs of variables are related. James (2003) defines correlation as the 'go togetherness" of two given variables. The correlation coefficient gives the amount of information common to two variables. There are several correlation coefficients often denoted by ρ (rho) (for a population correlation) or *r* (for a sample) measuring the degree of correlation and the most common of these is the **Pearson Correlation Coefficient** measuring a linear relationship between two variables. The **Spearman' Rank Correlation** coefficient measures the extent to which , as one variable increases the other variable tends to increase without requiring that increase is represented by a linear relationship. It is based on the rank relationship between variables.

The value of the correlation lies between -1 and 1, that is $-1 \leq \rho \leq 1$.This means the coefficient can be either negative or positive. The correlation coefficient is a number with no units. The values -1 and 1 represent perfect negative and perfect positive correlation respectively. The closer the absolute value is to 1, the stronger the relationship. There is no correlation, that is, no linear relationship between the variables when $r = 0$ indicating a random non linear relationship. In such a case we say the two variables have nothing in common. Below is a summary of the correlation ranges and their respective comments

Correlation range	**Description**
0-0.19	Very weak
0.20-0.39	Weak
0.40-0.59	Moderate
0.60-0.79	Strong
0.80-1.00	Very strong

It is important to note that when two variables are correlated it does not mean one causes the other but in fact imply that when one variable changes the other should in some way get known before it happens. Howell (2011) argues that the sign of the value of the correlation only shows the direction of the of the relationship, that is, 0.9 and -0.9 portray the same degree of relationship but differ in their direction.

4.1. Pearson's Product Moment Correlation coefficient

I will illustrate the computation with an example below

The English PSB and Mathematics PSB marks for 15 college students were presented in a table as shown below.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
English	76	82	43	92	46	77	81	66	81	43	91	18	32	74	65
Mathematics	51	42	71	60	52	70	76	54	80	76	15	40	67	41	60

A scatter plot of the set of data can be prepared by taking the x and y axis to represent English and Mathematics (either way). More important is the calculation of the product moment correction coefficient and interpret the results. Note that it is simply Pearson's correlation coefficient required given by

$$r = r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad \text{.....(i)}$$

$$= \frac{\sum_{i=1}^{n}(x_iy_i-x_i\bar{y}-y_i\bar{x}+\bar{x}\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i^2-2x_i\bar{x}+\bar{x}^2)}\sqrt{\sum_{i=1}^{n}(y_i^2-2y_i\bar{y}+\bar{y}^2)}}$$

$$= \frac{\sum_{i=1}^{n}x_iy_i-\bar{x}\sum_{i=1}^{n}y_i-\bar{y}\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n}x_i^2-2\bar{x}\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}\bar{x}^2}\sqrt{\sum_{i=1}^{n}y_i^2-2\bar{y}\sum_{i=1}^{n}y_i+\sum_{i=1}^{n}\bar{y}^2}}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y}}{\sqrt{(\sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2)}\sqrt{(\sum_{i=1}^{n} y_i^2 - 2n\bar{y}^2 + n\bar{y}^2)}}$$

$$since\ \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n},\ \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\sqrt{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i - n\frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n(\frac{\sum_{i=1}^{n} x_i}{n})^2}\sqrt{\sum_{i=1}^{n} y_i^2 - n(\frac{\sum_{i=1}^{n} y_i}{n})^2}}$$

$$= \frac{\frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sqrt{\frac{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}{n}}\sqrt{\frac{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}{n}}}$$

$$= \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}} \quad \text{.....(ii)}$$

A computation of the correlation coefficient can be done using the table below

	x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
1	76	51	11.53333333	-6	-69.19999999	133.01777	36
2	82	42	17.53333333	-15	-263	307.41777	225
3	43	71	-21.4666666	14	-300.5333333	460.81777	196
4	92	60	27.53333333	3	82.599999999	758.08444	9
5	46	52	-18.4666666	-5	92.333333333	341.01777	25
6	77	70	12.53333333	-13	-162.9333333	157.08444	169
7	81	76	16.53333333	19	314.13333333	273.35111	361
8	66	54	1.533333333	-3	-4.599999999	2.3511111	9
9	81	80	16.53333333	23	380.26666666	273.35111	529
10	43	76	-21.4666666	19	-407.8666666	460.81777	361
11	91	15	26.53333333	-42	-1114.4	704.01777	1764
12	18	40	-46.4666666	-17	789.9333333	2159.1511	289
13	32	67	-32.4666666	10	-324.6666666	1054.0844	100
14	74	41	9.53333333	-16	-152.5333333	90.884444	256
15	65	60	0.53333333	3	1.5999999999	0.2844444	9
	$\sum_{i=1}^{15} x_i$ =967	$\sum_{i=1}^{15} y_i$ =855			$\sum_{i=1}^{15}(x-\bar{x})(y-\bar{y})$ = -1138.66666667	$\sum_{i=1}^{15}(x-\bar{x})^2$ =7175.7333	$\sum_{i=1}^{15}(y-\bar{y})^2$ = 4338

Hence $r = \frac{-1138.66666667}{\sqrt{7175.73333333}.\sqrt{4338}} = -0.203969082.$

or

$$r = r_{xy} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}}$$

(ii) is more convenient for use when you can easily retrieve results from your calculator , that is

$\sum_{i=1}^{n} x_i = 967$, $\bar{x} = 64.466666667$,$\sum_{i=1}^{n} x_i^2 = 69\,515$, $\sum_{i=1}^{n} y_i = 855$, $\bar{y} = 57$, $\sum_{i=1}^{n} y_i^2 = 53\,073$, $\sum_{i=1}^{n} x_i y_i = 54\,306.$

These can be computed as in the table below

	x	y	xy	x^2	y^2
1	76	51	3876	5776	2601
2	82	42	3444	6724	1764
3	43	71	3053	1849	5041
4	92	60	5520	8464	3600
5	46	52	2392	2116	2704
6	77	70	5390	5929	4900
7	81	76	6156	6561	5776
8	66	54	3564	4356	2916
9	81	80	6480	6561	6400
10	43	76	3268	1849	5776
11	91	15	1365	8281	225
12	18	40	720	324	1600
13	32	67	2144	1024	4489
14	74	41	3034	5476	1681
15	65	60	3900	4225	3600
	$\sum_{i=1}^{15} x_i = 967$	$\sum_{i=1}^{15} y_i = 855$	$\sum_{i=1}^{15} xy = 54306$	$\sum_{i=1}^{15} x_i^2$=69515	$\sum_{i=1}^{15} y_i^2$=53073

Substituting into (ii) gives $r = r_{xy} = \frac{15X54306-(967)(855)}{\sqrt{15X69515-(967)^2}\sqrt{15X53073-(855)^2}}$

= -0.145717806.

The result shows a weak negative linear correlation between the marks in English and those in Mathematics suggesting a mark in English does not seem to influence the result in Mathematics. Pfenning (2011) emphasises two points to note in the value of the correlation coefficient.

(i).The value of r is unaffected by interchanging the variables on the Cartesian plane. If the points are plotted on the Cartesian plane they will always slope up or down when x become y

(ii).The value of r is unaffected by a change of units of measurement for example dollars to cents or kilometres to metres.

4.2. The Coefficient of Determination.R^2

The Coefficient of Determination is simply the squared value of the correlation coefficient. Correlation coefficients can be explained much easier using the Coefficient of Determination which is very simple to calculate. It is a measure of how well the regression line represents the data, that is, it provides an estimate of the fraction of overlapping variance between two sets of numbers (i.e., the degree to which the two sets of numbers vary together). If the regression line passes exactly through every point on the scatter plot it would be able to explain all the variation. The further the line is away from the points, the less it is able to explain. In the example above $r = -0.145717806$, then the Coefficient of Determination $R^2 = -0.14571786^2 = 0.021233694$. This means the regression line cannot explain the data.

A coefficient of determination of 0.92 can be interpreted as a fraction, or as 92%. As an example if two sets of marks in Mathematics and Science correlate at 0.76 the coefficient of determination becomes 0.5776 implying 57.76% of the variance of scores in Mathematics is shared with scores in Science. The remaining 42.24% (100-57.76) is not accounted for, that is no one knows to whom that amount is related to.

4.3. Spearman' Rank Correlation Coefficient

In some cases it is necessary to rank items on two dimensions for reasons of down- weighting extreme scores or maybe because the researcher does not trust the nature of the underlying scale and then correlate the two sets of data. The ranking may be done with the smallest score assuming position 1 and second smallest getting position 2 and so on. This gives what we call Spearman's correlation coefficient for ranked data (r_s) such that the r_s will measure the linear relationship between the two sets of ranks. I will also illustrate with an example the calculation of Spearman's rank correlation coefficient below.

The table below shows ten sets of readings for BP (systolic and diastolic) recorded for 10 patients.

Patient	1	2	3	4	5	6	7	8	9	10
Systolic	33	25	26	28	29	30	32	33	36	38
diastolic	54	55	45	50	41	34	30	30	25	38

A scatter diagram can be drawn depicting the kind of relationship between the two scores. A calculation of the Spearman's rank correlation coefficient will further confirm the relationship and so help interpret the results. The computations follow in the table below.

Patient	1	2	3	4	5	6	7	8	9	10
Systolic	33	25	26	28	29	30	32	33	36	38
Rank	7.5	1	2	3	4	5	6	7.5	9	10
Diastolic	54	55	45	50	41	34	30	30	25	38
Rank	9	10	7	8	6	4	2.5	2.5	1	5
d_i	-1.5	-9	-5	-5	-2	1	3.5	5	8	5
d_i^2	2.25	81	25	25	4	1	12.25	25	64	25

Note that the rank of each score is its position when the scores are arranged in ascending or descending order. Here $d_i = systolic - diastolic$ .The same interpretation will be obtained for $d_i = diastolic - systolic$.

Now $r_{sk} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$

where $\sum_{i=1}^{n} d_i^2 = 2.25 + 81 + 25 + 25 + 4 + 1 + 12.25 + 25 + 64 + 25 = 264.5$.

Hence $r_k = 1 - \frac{6X264.5}{10(10^2-1)}$ = 1-$\frac{1587}{990}$ = 1-1.603030303 = -0.603030303 = -0.603.

The value suggests there is some negative correlation between the two readings.

5.0.Conclusion

A study of the nature of numbers proved very useful for a classroom practitioner. The different kinds of data displays help visualise marks or any scores for purposes of practical adjustments to the learning situation for attainment of better results. A correlation coefficient, regardless of whether it is a Pearson correlation or a Spearman correlation, can tell us directly only about the marks in two different subjects on which it is computed. In such situations educators should note that correlation cannot be expected to give very precise information about other subjects on which it was not computed. It is therefore important to look at the distribution of data for unusual scores before calculating the correlation coefficient. This is the reason why correlation analysis should always begin with a scatter diagram. The scatter will help identify such observations which we cannot simply exclude from the research but justify why we are omitting it from our study.

REFERENCES:

Crawshaw J, Chambers, J. (2013) A concise course in Advanced Level Statistics (4th edition) Nelson Thornes, United Kingdom.

Elliott, Jane; Catherine Marsh (2008). Exploring Data: An Introduction to Data Analysis for Social Scientists (2nd ed.). Polity Press. ISBN 0-7456-2282-8.

Hogan, T.P, Evalenko, K. (2006)The elusive definition of outliers in introductory statistics textbooks for behavioural sciences. Teaching of Psychology, 33, 252-256.

Howell, D. C (2011). Fundamental Statistics for the Behavioral Sciences (7th edition). Linda Schreiber ,Wadsworth.

Jessica M. (2005) Seeing Through Statistics 3rd Edition, Thomson Brooks/Cole pp 166-167. ISBN 0-534-39402-7

Mendenhall W, Scheaffer R. L, Wackerly D.D (1981) Mathematical Statistics with Applications. Duxbury Press Boston, Massachusetts

Moore, D. S. and McCabe G. P. (1999) Introduction to the Practice of Statistics. New York: W. H. Freeman.

Nancy R. Tague (2004). "Seven Basic Quality Tools". The Quality Toolbox. Milwaukee, Wisconsin: American Society for Quality. p. 15. Retrieved 2010-02-05

Pfenning, N. (2011). Elementary statistics: Looking at the big picture. Richard Stratton. ISBN-13:978-0-495-83145-7.

Spatz C (2011) Basic Statistics: Tales of distributions (10th edition) Belmont, Wadsworth

Stephen B. (1994). Basic Statistics (Special pre-publication ed.). Dubuque, Iowa: Wm. C.

Wild, C. and Seber, G. (2000) Chance Encounters: A First Course in Data Analysis and Inference pp. 49–54 John Wiley and Sons. ISBN 0-471-32936-3

William (1993). Visualizing data. Murray Hill, N.J. Summit, N.J: At & T Bell Laboratories Published by Hobart Press. ISBN 978-0963488404