Chapter 7

Handling Research Data with Descriptive Statistics

7 Handling Research Data with Descriptive Statistics

Felix Kutsanedzie[1]; Sylvester Achio[1]; Appiah Lewis Gyekye[1]

[1]Accra Polytechnic, Accra, Ghana

Abstract

Statistics is a tool used in research analysis and its classified into different branches such as descriptive statistics and inferential statistics. Each of these branches are used to achieve different purposes in research in terms of summary, data analysis and interpretations. The data type collected in the conduct of research determines the type of statistical tool. It can either fall under descriptive or inferential statistics, which must be deployed in data summary, analysis and interpretations. Most researchers and students do not know the tools that fall under descriptive and how to deploy them to handle data garnered during research studies. This chapter therefore identified the various statistical tools under descriptive statistics that can be used to for data processing.

Keywords

Research, Descriptive Statistics, Data, Tools, Analysis

7.1 Introduction

Descriptive statistics is one of the braches of statistics which involves the use of statistical tools to analysis an entire population given that all data sets representing the population are known. For descriptive statistics to be used for data analysis it presupposes that all values of all variables of every subject or entire subject of a population data set is known. Once the dada of the whole population is known, descriptive statistical tools can be used to summarize, analyse, and determine the pattern of the data set representing the population in order to give a decisive interpretation to the data collected in the conduct of a research study. The first stage in data analysis is to describe or summarize data, and the whole analysis may involve calculating and interpreting descriptive statistics. It should be made known that statistics refers to an index used to measure the performance of a sample while parameter is used to refer to a population. Hence for a descriptive statistic to be used for data analysis, all the values of the variables within the sample must be known. Likewise for descriptive statistics to be performed on a population all values of it variables must be known.

Descriptive statistics looks at the measures of central location or tendency; measures of dispersion or variation; distribution shapes; and measures of position. For each research study in when the values of variables for a population is known, all the following statistical tools has the potential of being used for its summary, analysis and interpretation based on the objective to be achieved. Thus in the conduct of a research, the data conducted may require the use of descriptive statistics or inferential depending on the data set that is collected. These statistical descriptive tools and procedures are explained to the readers in a way to enable them know how these tools are used for data analysis. Although these are vital statistical tools, the essence is not to explain the

statistics or mathematics but to emphasize their use and application for data analysis and interpretation in research studies.

7.2 Measures of Central Tendency or Location

The measures of central tendency or location are used to determine the average scores of the sample or population with values of variables known. Usually when instruments are used for taking measurements; and the measurements of the same variable is repeated they may be often times differ. Sometimes these values might be at their extremes. Taking for instance a class score of students for a test out of ten (10) in mathematic are given as follows: 0, 1, 5, 9, 9, 0, 0, 1, 0, 9, 8, 7. The measures of central tendency are used to determine whether the scores fall closer to the central point or the average so that a single value or figure or number is used representatively to describe the data collected. Thus the average of the collected score can be computed and used to describe the data; the median value; and the mode can all be used. However the most appropriate one depending on the type of data, level of measurement as well as the purpose of the study. For instance if the level of measurement of the data is nominal, the mode is appropriate; if however it is ordinal, the median and if interval or ratio, and the mean.

Mode as Measure of Central Location or Tendency

The mode is one of the measures of central location which means it is used as a single number or value or value label to describe or summarize a given set of collected data for a study. It is the number or a value label with the highest occurring frequency within a data set or the sample or population. For instance taking into consideration the test score in mathematic case given earlier - 0, 1, 5, 9, 9, 0, 0, 1, 0, 9, 8 and 7, the most frequent occurring value is 0, and so the modal

value is 0, meaning that majority of the students scored zero. However this is one of the least used measures of central location and it is the most unstable measure. It is appropriate to use this measure for data with nominal level of measurement. Assuming a researcher collects data on the colour of shirts worn in a class as: red, red, green yellow, blue, red, yellow, yellow and pink, yellow. In this case the data type is qualitative and the level of measurement that can be applied is nominal. Hence the use of the mean and the median cannot be computed or determined for the data and used as a single value or value label to represent or summarize, or describe the data appropriately. The only option available to obtain a single value or value that can be used to represent such data is the use of the mode, i.e. to find the modal value label or colour. Since yellow is the most occurring shirt colour, it becomes the modal colour which is used to describe or represent the data.

The mode can be found for both group and ungroup data that are collected.

However, it should be noted that the appropriate level of measurement of data for which the mean can be appropriately used to summarize as a single number is ratio or interval level.

The mean score for grouped and ungrouped data can be calculated as follows:

Data collected from a research study or an experiment can either be grouped or ungrouped. It is referred to as an ungrouped not classified and grouped when it is put into categories.

Taking for instance the weight of students taken using a weighing scale and recorded as follows:

50, 51, 60, 51, 67, 51, 68, 47, 75, 78, 81, 63, 64, 85, 46, 48, 77, 86, 66, 71. This dataset is ungrouped and therefore the determination of the measures of

central tendency would follow a different means of determination compared to the grouped data.

When the same dataset is put into categories with the following intervals: 40 – 49, 50 – 59, 60 – 69, 70 – 79, 80 – 89, then the data set would be referred to as grouped.

To find the mode of an ungrouped data, it is simply finding the frequency of the individual values or scores within the dataset and then taking the one with the highest frequency as the mode or the modal value or item.

Using the given ungrouped dataset - 50, 51, 60, 51, 67, 51 68, 47, 75, 78, 81, 63, 64, 85, 46, 48, 77, 86, 66, 71; the frequencies of the individual scores are derived below and used for the modal value determination.

Table 7.1 *Frequency Table of an Ungrouped Dataset.*

Weights (kg)	**47**	**48**	**46**	**50**	**51**	**60**	**63**	**64**	**66**	**67**	**68**	**71**	**75**	**77**	**78**	**81**	**85**	**86**
Frequencies	1	1	1	1	3	1	1	1	1	1	1	1	1	1	1	1	1	1

Thus the modal value or weight is 51kg because it is the weight measured with the highest frequency. It is a unimodal data, meaning it has one mode. The modal value 51kg indicates that the majority of the students weighed 51kg or the value that can be used to represent the weights of the students in the class is 51kg.

For the grouped data for the same dataset - 50, 51, 60, 51, 67, 51 68, 47, 75, 78, 81, 63, 64, 85, 46, 48, 77, 86, 66, 71 and 72, grouped into intervals such as 40 – 49, 50 – 59, 60 – 69, 70 – 79, 80 – 89. The frequency table for the grouped data is shown in table 7.2

Table 7.2 *Frequency Table of a Grouped Dataset.*

Weight Categories of Students	Frequencies
40 – 49	3
50 – 59	4
60 – 69	6
70 – 79	5
80 – 89	3

From the frequency table developed, the modal class of student weights is 60 – 69, but to determine the modal weight of the modal class or interval, the formula below is used:

$$Mode = L + \left(\frac{D_1}{D_1 + D_2}\right) C$$

L=Lower class boundary of the modal class. It is found by substracting 0.5 from the lower limit of the modal class, thus the lower class boundary of modal class =60-0.5=59.5, D_1=the difference in frequences between the modal class and the class before it,

$$i.e. D_1 = 6 - 4 = 2$$

D_2=the difference in frequencies between the modal class and the class after it,

$$i.e. D_2 = 6 - 5 = 1$$

C=the modal class size=upper class limit-lower class limit+1

$$C = 69 - 60 + 1 = 9 + 1 = 10$$

$$Modal\ Weights\ of\ students = 59.5 + \left(\frac{2}{2+1}\right) x\ 10$$

$$Modal\ weight\ of\ students = 59.5 + \left(\frac{2}{3}\right) x\ 10 = 59.5 + 6.67 = 66.16 \cong 66$$

Thus the mode is the most unstable measure of central tendency compared to all the other measures and thus most appropriate for data that is nominal level of measurement has been applied.

Median as Measure of Central Location or Tendency

The median is also another measure of central tendency used as a single value or value to describe or summarize or represent a given data set. The median is regarded as the value or value label that occupies the median position of the data. Again with the example given in the case of the test score - 0, 1, 5, 9, 9, 0, 0, 1, 0, 9, 8 and 7. The median is the value that occupies the middle position when the data set is arranged in an ascending or descending other. So in this case the data arranged in ascending order becomes: 0, 0, 0, 0, 1, 1, 5, 7, 8, 9, 9, 9. Counting from both sides, two numbers occupy the middle position. Thus the $median\ value = \frac{1+5}{2} = 3$. Therefore for the median value to be found required that the average of the two vales are computed and the value used as the median value. It should be noted here that just like the mode, the appropriate level of measurement of the data that the median can be applied to is the ordinal level.

The median score can be found for grouped and ungrouped data

To determine the median for an ungrouped data such -50, 51, 60, 51, 67, 51 68, 47, 75, 78, 81, 63, 64, 85, 46, 48, 77, 86, 66, 71; the values within the dataset are arranged in ascending or descending order and the number or value that occupies the central position is taken as the median.

46, 47, 48, 50, 51, 51, 51, 60, 64, 65, 66, 67, 68, 71, 75, 77, 78, 81, 85, 86

To find the position of the median value, divide the total frequency of the weight by two

$$median\ Position = \frac{20}{2} = 10th$$

However, if two numbers or values compete for the 9th position, the average of the two numbers is computed and then used as the median value or weight in

this case. Thus for the data above, the 10th position is occupied by two numbers i.e. 64 and 65 respectively counting from the left and right, therefore the Median mark or value $= \frac{65+66}{2} = 65.5$.

However when an odd number ungrouped dataset is used such as 2, 3, 3, 4, 5, 6, 2; the values within the dataset are arranged in either ascending or descending order such as 2, 2, 3, 3, 4, 5, 6 and their total frequency determined with one (1) added for the computation of the position of the median value.

Thus the position of the median value $= \frac{7+1}{2} = \frac{8}{2} = 4^{th}$, therefore the number or value occupying the 4th position becomes the median value.

Median value therefore =4

Using the dataset: 50, 51, 60, 51, 67, 51, 68, 47, 75, 78, 81, 63, 64, 85, 46, 48, 77, 86, 66, and 71 grouped into intervals such as 40 – 49, 50 -59, 60 – 69, 70 – 79, 80 – 89. The frequency table is developed and the median value calculated as follows:

Table 7.3 *Table Showing the Median Mark for a grouped Dataset.*

Weight Categories of Students	Frequencies	Cumulative frequency
40 – 49	3	3
50 – 59	4	7
60 – 69	6	13
70 – 79	5	18
80 – 89	3	20
	∑f = 20	

To compute the median value the total frequency of the dataset is determined and the median class is obtained by dividing the total frequency by 2, and tracing the class within which it falls.

$$Median\ Class\ position\ = \frac{20}{2} = 10th$$

Using the cumulative frequency of the dataset, the 10th position can be located within 60-69, thus the median class becomes 60 – 69.

Once the median class has been identified, the formula given below is used to compute for the median mark or value:

$$Median\ value\ = L + \left(\frac{\frac{\sum f}{2} - f_c}{f_m}\right)C$$

L=Lower class boundary of the median class. It is found by substracting 0.5 from the lower limit of median class, thus the lower class boundary of median class =60-0.5=59.5, f_c=cumulative frequency before the median class or sum of frequencies before the median class, f_c=3+4 = 7, f_m=frequency of the median class, f=Total frequency of the dataset, C=the modal class size=upper class limit-lower class limit+1

$$C = 69 - 60 + 1 = 9 + 1 = 10$$

$$Median\ value\ = 59.5 + \left(\frac{\frac{20}{2} - 7}{6}\right)x10$$

$$Median\ value\ = 59.5 + \left(\frac{20 - 7}{6}\right)x\ 10$$

$$Median\ value\ = 59.5 + \left(\frac{13}{6}\right)x\ 10 = 59.5 + \frac{130}{6} = 59.5 + 21.67$$

$$Median\ value\ = 59.5 + 21.67 = 81.17$$

Mean as Measure of Central Location or Tendency

The mean also referred to as average of a data set is considered as the most stable and most used measure of central location as it is used as a single number

or value to describe or summarize or represent the data set as a point of convergence where the true score lies. It is found to be the sum of all the scores and dividing it by the total number of occurrences. Thus using the test score - 0, 0, 0, 0, 1, 1, 5, 7, 8, 9, 9, 9;

$$\text{Thus the } mean\ value = \frac{0+0+0+0+1+1+5+7+8+9+9+9}{2} = 4.08$$

Taking an ungrouped dataset such as 50, 51, 60, 51, 67, 51, 68, 47, 75, 78, 81, 63, 64, 85, 46, 48, 77, 86, 66, and 71; the mean or average of the dataset can be determined by the formula

$$Mean\ (\bar{x}) = \frac{\sum x_i}{n}$$

$$Mean\ (\bar{x}) = \frac{50 + 51 + 51 + 51 + \cdots 71}{20}$$

$$Mean\ (\bar{x}) = \frac{1290}{20} = 64.5$$

For group data of the same dataset: 50, 51, 60, 51, 67, 51, 68, 47, 75, 78, 81, 63, 64, 85, 46, 48, 77, 86, 66, and 71 grouped into intervals such as 40 – 49, 50 – 59, 60 – 69, 70 – 79, 80 – 89.

A table is developed and the mean calculated as follows:

Table 7.4 *Table Showing Mean determined for a Grouped Dataset.*

Weight Class of Students	Frequencies (f)	Midpoint (x)	Fx
40 – 49	3	44.5	133.5
50 – 59	4	54.5	218
60 – 69	6	64.5	387
70 – 79	5	74.5	372.5
80 – 89	3	84.5	253.5
	∑f = 20		∑fx = 1364.5

The midpoint is found for each class by summing upper class limits and lower class limits for each class and dividing by 2. i.e. for class 40 – 49, the midpoint = (40 +49)/2 = 44.5

$$Mean\ (\bar{x}) = \frac{\sum fx}{n} = \frac{\sum fx}{\sum f} = \frac{1364.5}{20} = 68.23$$

7.3 Measures of Variation or Dispersion

These measures are single numbers or indices which are used to describe or summarize the variations in dataset collected in the conduct of a research study or experiment. Just as the measures of central location are single numbers or values or indices used to describe the points at which the dataset converges to a central point, the measures of dispersion or variation rather looks at the spread of values within the dataset collected from a research study. The measures of dispersion include the *range* of the dataset, *interquartile range or quartile deviation, variance, and standard deviation.*

Range as Measure of Variation

Range is the difference between the highest and lowest value recorded in a particular dataset collected from a research study. It gives an indication of the spread within the data though it is not a stable measure of dispersion.

For the dataset: 50, 51, 60, 51, 67, 51, 68, 47, 75, 78, 81, 63, 64, 85, 46, 48, 77, 86, 66, and 71; the range for this dataset is 86 - 46 = 40, and indication that the spread between or the variation between the lowest value and the highest value within the dataset is 40 units. Though it is not the best type of measure in terms of variations in a data collected, the lower the range the closer the differences or less variations in the data set. The level of measurement of data the range can be applied to is the ordinal level.

Quartile Deviation or semi-interquartile range

This is found by finding the difference between the third (3^{rd} or the 75^{th} percentile) or upper quartile and the first quartile (1^{st} or the 25^{th} percentile) or lower quartile and then dividing by 2. Before the quartile deviation or semi-interquartile range, the first and third quartile must be computed for the dataset.

To calculate the quartile deviation or semi-interquartile range for an ungrouped data such as:

50, 51, 60, 51, 67, 51, 68, 47, 75, 78, 81, 63, 64, 85, 46, 48, 77, 86, 66, and 71

The data is re-arranged in either ascending or descending order and the following formula are used to determine the positions of the first and the third quartile:

Position of the First (1^{st} or the 25^{th} Percentile) Mark $= \frac{1}{4}$ x N + 1

Position of the Second (2^{nd} or the 50^{th} Percentile) Mark $= \frac{2}{4}$ x N + 1 $= \frac{1}{2}$ x N + 1

Position of the Third (3^{rd} or the 75^{th} Percentile) Mark $= \frac{3}{4}$ x N + 1

Where N = size of the data in terms of number = 20.

The data re-arranged in ascending order becomes: 46, 47, 48, 50, 51, 51, 51, 60, 64, 65, 66, 67, 68, 71, 75, 77, 78, 81, 85, 86.

Therefore,

Position of the First (1st or the 25th Percentile) Mark $= \frac{1}{4}$ x $20 + 1 = 5 + 1 = 6^{th}$.

Thus the Mark that occupied the 6th position within the re-arranged data becomes the First Quartile Mark as show below:

The First Quartile Mark (Q1) = 51

Position of the Second (2nd or the 50th Percentile) Mark $= \frac{1}{2}$x $20 + 1 = 10 + 1 = 11^{th}$

46, 47, 48, 50, 51, 51, 51, 60, 64, 65, 66, 67, 68, 71, 75, 77, 78, 81, 85, 86

Thus the Second Quartile Mark (Q2) = 66

Position of the Third (3rd or the 75th Percentile) Mark $= \frac{3}{4}$ x $20 + 1 = 15 + 1 = 16^{th}$

46, 47, 48, 50, 51, 51, 51, 60, 64, 65, 66, 67, 68, 71, 75, 77, 78, 81, 85, 86

Thus the Third Quartile Mark (Q3) = 77

Quartile Deviation or semi-interquartile range mark$= \frac{Q_3 - Q_1}{2} = \frac{77-51}{2} = \frac{26}{2} = 13$

Therefore when the quartile deviation is small there is less variations within the dataset and vice versa.

For data such as: 50, 51, 60, 51, 67, 51, 68, 47, 75, 78, 81, 63, 64, 85, 46, 48, 77, 86, 66, and 71 grouped into intervals such as 40 – 49, 50 – 59, 60 – 69, 70 – 79, 80 – 89.

The quartile deviation can be computed as follows:

First, a frequency table such as done below is constructed for the dataset

Table 7.5 *Frequency Table of an Ungrouped Dataset used for the Determination of the Quartile Deviation for a Grouped Dataset.*

Weight Categories of Students	Frequencies	Cumulative frequency
40 – 49	3	3
50 – 59	4	7
60 – 69	6	13
70 – 79	5	18
80 – 89	3	20
	∑f = 20	

Use the following formula to identify the first, second and third quartile classes:

Position of the First (1st or the 25th Percentile) Class= $\frac{1}{4}$ x N + 1

Position of the Second (2nd or the 50th Percentile) Class= $\frac{2}{4}$ x N + 1 = $\frac{1}{2}$ x N + 1

Position of the Third (3rd or the 75th Percentile) Class= $\frac{3}{4}$ x N + 1

Position of the First (1st or the 25th Percentile) Class= $\frac{1}{4}$ x 20 + 1 = 5 + 1 = 6^{th}

Position of the Second (2nd or the 50th Percentile) Class= $\frac{1}{2}$ x 20 + 1 = 10 + 1 = 11^{th}

Position of the Third (3rd or the 75th Percentile) Class=$\frac{3}{4}$ x $20 + 1 = 15 + 1 = 16^{th}$

Therefore the 6th item in the data is supposed to be located within the First Quartile and using the frequency the 6th item would be located within the class 50 – 59, hence the First Quartile Mark can be located within 50 – 60; the Second Quartile Class is 60 – 69 since data item with position 11 falls within it; the Third Quartile Class is 70 – 79, since data item with the position 16 falls within it. Thus one can then proceed to determine the First, Second and Third Quartile Marks with the following respective formula:

$$Q_1 = L_1 + \left(\frac{\frac{N}{4} - f_c}{f_1}\right)c$$

Q_1=First quartile mark, N=Total frequency of data or data size=20, f_c=cumulative frequency before the first quartile class or total frequency before first quartile class=3, f_1=frequency of the first quartile class=4, L_1=lower class boundary of the first quartile class=49.5, c=first quartile class size=10

$$Q_1 = 49.5 + \left(\frac{\frac{20}{4} - 3}{4}\right) \text{x } 10 = 49.5 + \left(\frac{5-3}{4}\right) \text{x } 10 = 49.5 + (0.5)(10) = 54.5$$

$$Q_2 = L_2 + \left(\frac{\frac{N}{2} - f_c}{f_2}\right)c$$

Q_2=Second quartile mark, N=Total frequency of data or data size=20, f_c=cumulative frequency before the second quartile class or total frequency before second quartile class=7, f_1=frequency of the second quartile class=6, L_1=lower class boundary of the second quartile class=59.5, c=second quartile class size=10

$$Q_2 = 59.5 + \left(\frac{\frac{20}{2} - 7}{6}\right) x\ 10 = 59.5 + \left(\frac{10 - 7}{6}\right) x\ 10 = 59.5 + \left(\frac{3}{6}\right)(10) = 59.5 + 5$$
$$= 64.5$$

$$Q_3 = L_3 + \left(\frac{\frac{3N}{4} - f_c}{f_3}\right) c$$

Q_=Third quartile mark, N=Total frequency of data or data size=20, f_c=cumulative frequency before the third quartile class or total frequency before third quartile class=13, f_1=frequency of the third quartile class=5, L_1=lower class boundary of the third quartile class=69.5, c=third quartile class size=10

$$Q_2 = 69.5 + \left(\frac{\frac{3(20)}{4} - 13}{5}\right) x\ 10 = 69.5 + \left(\frac{15 - 13}{5}\right) x\ 10 = 69.5 + \left(\frac{2}{5}\right)(10)$$
$$= 69.5 + 4 = 73.5$$

Hence, $the\ Quartile\ Deviation\ or\ Semi - interquartile\ mark = \frac{Q_3 - Q_1}{2}$.

The Quartile Deviation or Semi-interquartile mark= $\frac{73.5 - 54.5}{2} = \frac{19}{2} = 9.5$.

Hence the value 9.5 is the measure of the variations within the given dataset.

Mean Deviation as a Measure of Variation or Dispersion

This is also a measure of variation, it is found by finding the summation of the differences between the mean of a dataset and the individual data values and dividing by the size of the dataset or the number of individual data it consist of. It thus suggests that the lower the mean deviation, the smaller the variation in the dataset and vice versa. Taking for instance the ungrouped data: 50, 51, 60,

51, 67, 51, 68, 47, 75, 78, 81, 63, 64, 85, 46, 48, 77, 86, 66, and 71. The mean deviation is computed using the formula below:

$$Mean\ Deviation = \frac{\sum(x_i - \bar{x})}{n}$$

Table 7.6 *Calculation of the Mean Deviation for an Ungrouped Dataset.*

Weights (x) kg	Mean ($\bar{x}$)	x-$\bar{x}$
47	64.5	-17.5
48	64.5	-16.5
46	64.5	-18.5
50	64.5	-14.5
51	64.5	-13.5
60	64.5	-4.5
63	64.5	-1.5
64	64.5	-0.5
66	64.5	1.5
67	64.5	2.5
68	64.5	3.5
71	64.5	6.5
75	64.5	10.5
77	64.5	12.5
78	64.5	13.5
81	64.5	16.5
85	64.5	20.5
51	64.5	-13.5
51	64.5	-13.5
86	64.5	21.5
		$\sum$(x-$\bar{x}$) = -5

$$Mean\ Deviation = \frac{-5}{20} = -0.25$$

For the data: 50, 51, 60, 51, 67, 51, 68, 47, 75, 78, 81, 63, 64, 85, 46, 48, 77, 86, 66, and 71 grouped into intervals such as 40 – 49, 50 – 59, 60 – 69, 70 – 79, 80 – 89. The mean deviation is computed by developing a table as shown in table 7.7.

Table 7.7 *Calculation of the Mean Deviation for a Grouped Dataset.*

Weights Classes	Midpoints (x)	Freq(f)	mean ($\bar{x}$)	x-$\bar{x}$	f(x − $\bar{x}$)
40 – 49	44.5	1	68.23	-23.73	-23.73
50 – 59	54.5	3	68.23	-13.73	-41.19
60 – 69	64.5	1	68.23	-3.73	-3.73
70 - 79	74.5	1	68.23	6.27	6.27
80 - 89	84.5	1	68.23	16.27	16.27
				∑(x − $\bar{x}$) =-18.65	∑f(x − $\bar{x}$) =-46.82

$$Mean\ Deviation = \frac{\sum f(x_i - \bar{x})}{n} = \frac{-46.82}{20} = -2.34$$

Variance and Standard Deviation as a Measure of Variation or Dispersion

It is a measure of the spread within a dataset. To calculate the variance for a data set, the mean for the collected data is computed, the differences in the mean and each data item is then calculated and each squared and then sum, and finally the outcome divided by the number or size of the dataset to obtain the variance. When the variance is small it means the variation within the dataset is also small and vice versa. Usually when sum of the differences between the mean and each individual value in the data is zero; the square of the differences in the mean and each individual value or item in the data set is rather determined. The square root of the variance is referred to as standard deviation. The standard deviation is therefore the most stable and the most used index of variability in a dataset. The appropriate level of measurement of data the standard deviation is applied to is the interval and ratio levels. The formula for the computation of variance and standard deviation are given as follows:

For a given ungrouped data such as: 50, 51, 60, 51, 67, 51, 68, 47, 75, 78, 81, 63, 64, 85, 46, 48, 77, 86, 66, and 71. The mean is first determined as follows:

$$Mean\ (\bar{x}) = \frac{\sum x_i}{n}$$

$$Mean\ (\bar{x}) = \frac{50 + 51 + 51 + 51 + \cdots 71}{20}$$

$$Mean\ (\bar{x}) = \frac{1290}{20} = 64.5$$

$$Variance\ (v) = \frac{\sum(\mathrm{x} - \bar{x})^2}{n}$$

$$Standard\ deviation\ (s) = \sqrt{\frac{\sum(\mathrm{x} - \bar{x})^2}{n}}$$

After which the table below can be prepared and used for the mean differences and their respective squares

Table 7.8.1 *Calculation of the Variance and Standard Deviation for an Ungrouped Dataset.*

Weights (x) kg	Mean ($\bar{x}$)	x-$\bar{x}$	(x-$\bar{x}$)^2
47	64.5	-17.5	306.25
48	64.5	-16.5	272.25
46	64.5	-18.5	342.25
50	64.5	-14.5	210.25
51	64.5	-13.5	182.25
60	64.5	-4.5	20.25
63	64.5	-1.5	2.25
64	64.5	-0.5	0.25
66	64.5	1.5	2.25
67	64.5	2.5	6.25
68	64.5	3.5	12.25
71	64.5	6.5	42.25
75	64.5	10.5	110.25
77	64.5	12.5	156.25
78	64.5	13.5	182.25
81	64.5	16.5	272.25
85	64.5	20.5	420.25
51	64.5	-13.5	182.25
51	64.5	-13.5	182.25
86	64.5	21.5	462.25
		$\sum$(x-$\bar{x}$) = -5	$\sum$(x-$\bar{x}$)^2 =3367

$$Varaince = \frac{\sum(\mathrm{x} - \bar{x})^2}{n} = \frac{3367}{20} = 168.35$$

$$Standard\ deviation\ (s) = \sqrt{\frac{\sum(\mathrm{x} - \bar{x})^2}{n}}$$

$$Standard\ deviation\ (s) = \sqrt{\frac{3367}{20}} = \sqrt{168.35} = 12.97$$

$$Therefore\ s = \sqrt{v}$$

For the data: 50, 51, 60, 51, 67, 51, 68, 47, 75, 78, 81, 63, 64, 85, 46, 48, 77, 86, 66, and 71 grouped into intervals such as 40 – 49, 50 – 59, 60 – 69, 70 – 79, 80 – 89. The variance and standard deviation is computed by developing a table as below:

Table 7.8.2 *Calculation of the Variance and Standard Deviation for a Grouped Dataset.*

Weights Classes	Midpoints (x)	Freq (f)	mean ($\bar{x}$)	x-$\bar{x}$	(x-$\bar{x}$)^2	f(x-$\bar{x}$)^2
40 – 49	44.5	1	68.23	-23.73	563.11	563.11
50 – 59	54.5	3	68.23	-13.73	188.51	565.53
60 – 69	64.5	1	68.23	-3.73	13.91	13.91
70 - 79	74.5	1	68.23	6.27	39.31	39.31
80 - 89	84.5	1	68.23	16.27	264.71	264.71
				$\sum(\mathrm{x} - \bar{x}) = -18.65$	$\sum(\mathrm{x} - \bar{x})^2 = 1069.57$	$\sum \mathrm{f}(\mathrm{x} - \bar{x})^2 = 1446.57$

For group data variance and standard deviation, the formula below are used:

$$Varaince = \frac{\sum f(\mathrm{x} - \bar{x})^2}{n}$$

$$Variance = \frac{1446.57}{20} = 72.33$$

$$Standard\ deviation\ (s) = \sqrt{\frac{\sum f(\mathrm{x} - \bar{x})^2}{n}}$$

$$Standard\ deviation\ (s) = \sqrt{\frac{1446.57}{20}} = \sqrt{72.33} = 8.50$$

For data to be considered as relatively normal each individual value of the dataset should fall within the range of $\bar{x} \pm 3$SD (3- Standard deviation from the mean of the dataset).

7.4 Measures of Position

It indicates the relative position of an individual score to the scores of others within the dataset or measured on the same variable. It is used when one wants to measure the performance of an individual among his or her colleagues measured on the same variable. The measures of relative position include percentile and the standard scores.

Percentile rank as a Measure of Position

Percentile rank is an indication of whether a percentage score falls on or below a given score. Oftentimes it is deceptive to think a student performed badly just looking at his or her scores in given subjects. For instance taking a student who scored 45 marks and 65 marks out of 100 respectively in Science and English, it would appear as if the student performed better in English in his or her class compared to science but the story might actually be different when subjected to statistically analysis. When the student's performance in English is compared to the scores or performance in the same subject, it might be realized that though the student scored 65 marks in English, his or her score could be the lowest in the class while the 45 marks scored in science could be the highest score in the class. In order to analyse this situation so that the relative position of the student in each of these subject among the classmates is established, the

percentile ranking is used. Thus when the percentile ranking is computed for the scores of the students in English and the student's score of 65 marks for instance falls on or corresponds to 15^{th} percentile, it means that 15 percent of the scores of the students in the class scored lower than 65 marks. If thus the score of 45 marks obtained by the student correspond to the 90^{th} percentile, it thus means 90 percent of the class scored below the student's score of 45 marks.

The percentile can be computed for ungrouped and grouped data as below:

Ungrouped data

To determine the percentile for an ungrouped data of marks such as: 50, 51, 60, 51, 67, 51, 68, 47, 75, 78, 81, 63, 64, 85, 46, 48, 77, 86, 66, and 71.

In order to compute any percentile for an ungrouped data, the scores must be arranged in the ascending order.

This formula determines the position of the mark that corresponds to a particular percentile.

For the position of the 1^{st} percentile score $= \frac{1}{100}\ x\ N + 1$

P^{th}=the percentile position, N=total frequency or size of data

Thus to calculate the position of the 40^{th} percentile score;

$$P^{th} = \frac{40}{100}\ X\ 20 + 1 = 9^{th}$$

It means when the scores are arranged in an ascending order, the score that occupies the 9^{th} position is the score corresponding to the 40^{th} percentile.

46, 47, 48, 50, 51, 51, 51, 60, 64, 65, 66, 67, 68, 71, 75, 77, 78, 81, 85, 86.

The 40th percentile mark thus correspond to 64 mark.

Grouped Data

For the data such as: 50, 51, 60, 51, 67, 51, 68, 47, 75, 78, 81, 63, 64, 85, 46, 48, 77, 86, 66, and 71 grouped into intervals such as 40 – 49, 50 – 59, 60 – 69, 70 – 79, 80 – 89. First, a frequency table such as done below is constructed for the dataset.

Table 7.9 *Calculation of the Percentiles for a Grouped Dataset.*

Weight Categories of Students	Frequencies	Cumulative frequency
40 – 49	3	3
50 – 59	4	7
60 – 69	6	13
70 – 79	5	18
80 – 89	3	20
	∑f = 20	

To determine the percentile in the case of grouped data, the percentile class is first identified by using the formula:

The first (1st) percentile class is given by

$$P^{th} = \frac{1}{100} X N + 1$$

$$P^{th} = \frac{1}{100} X 20 + 1 = 1.2 \cong 1^{st}$$

Thus the class in which the frequency of 1.2 falls becomes the first (1st) percentile class. Thus the percentile class is 40 – 49.

Therefore to calculate the first (1st) percentile mark

$$P_1 = L_1 + \left(\frac{\frac{N}{100} - f_c}{f_1} \right) c$$

P_1=1st percentile mark, N=Total frequency of data or data size=20, f_c=cumulative frequency before the percentile class or total frequency before the percentile class=0, f_1=frequency of the first (1st) percentile class=0, L_1=lower class boundary of the first percentile class= 39.5, c=first percentile class size=10.

$$P_1 = 39.5 + \left(\frac{\frac{20}{100} - 0}{4} \right) x\ 10 = 39.5 + \left(\frac{0.2 - 0}{4} \right) x\ 10 = \ 39.5 + (0.05)(10) = 40$$

Thus the first percentile corresponding to the score 40 marks and thus the lowest mark of the class.

Standard Scores as a Measure of Position

Standard score is a derived score that expresses how far a given raw score is from a referenced point. Some examples of standard scores that are known include z-score and t-score. They are used in converting averages or means in terms of standard deviation units. Sometimes it becomes difficult to have an appropriate scale on which various averages of different tests on performance of individual can be determined. For instance determining the averages of scores in terms of students performance in English and in mathematics on one scale, the standard scores becomes a means of handling that. It should be noted that standard scores are only accurate to a degree where the scores are normally distributed. Thus when one wants to convert raw scores that are not normally distributed, there is a need to transform such scores to what is referred to as normalized scores before their conversion to standard scores.

Z-scores

It is a basic standard score which expresses the mean score in terms of standard deviation units. When scores are transformed into z-score the new mean of the distribution becomes 0 and the standard deviation is 1. On the z-score the mean of the data becomes 0 and a score that is 1 standard deviation above the mean it correspond to z-score of +1; and if a score is 1 standard deviation below the mean it corresponds to z-score of -1. It allows different sets of test scores to be compared on the same scale.

Assuming a student score 40 marks in mathematics and 60 marks in English, from just this statement any one would be tempted to conclude that the student is better in English compared to mathematics. However if it is added to the statement that the average or mean class score in mathematics is 20 and English is 70, then the whole story takes a different dimension in that it indicates that though the student scored a low mark of 40 in mathematics, he had above the average score in his or her class in the same subject; and the contrary can be said of the student's score in English. It thus presupposes that the student performed better in Mathematics as compared to English in the class. Again if the idea of standard deviation of the class scored in both subject is given as 20, then more detailed information on the students score can be obtained. The given information indicates that the student's performance in mathematics would be (20 + 20 = 40).

$$Z = \frac{x - \bar{x}}{SD} = \frac{40 - 20}{20} = \frac{20}{20} = +1$$

z=z-score, x=raw score, SD=standard deviation of scores in the subject, $\bar{x}$=mean of class scores in subject.

Thus the student's performance in mathematics on the z-score would be z-score of +1, meaning his or her performance in mathematics is 1 SD (Standard Deviation) above the mean. In the case of the subject English, (70 - 10 = 60)

$$Z = \frac{x - \bar{x}}{SD} = \frac{60 - 70}{20} = \frac{-10}{20} = -0.5$$

Thus the student's performance in English on the z-score would be z-score of - 0.5; this means the student's performance is 0.5 SD (Standard Deviation) below the mean.

T- Scores

The t-scores is also handled like the z-scores but in the case of t-scores, the size of the data is less, normally it is less than 30, above that size it is treated as z-score.

7.5 Distribution of Shapes

This refers to the shapes assumed by a dataset. An example is skewness of dataset. A dataset is considered normal when the mean, median and mode are equal. The distribution of a normal dataset is given by the normal curve, which is symmetrical in nature or shape, meaning it can be divided into equal halves. However if a dataset or distribution of a dataset is not normal, then it is said to be skewed either to one side or the other, i.e. left or right. When a dataset is skewed, it means most of it elements or data values are more packed on one side than the other. The distribution is termed positively or right skewed if the extremes scores are the upper end of the distribution, that is when the mean is greater than the median and the mode (mean > median > mode). It is negatively or left skewed when the extreme scores are the lower end of the distribution, that is the mean is less than the median and the mode (mean < median < mode).

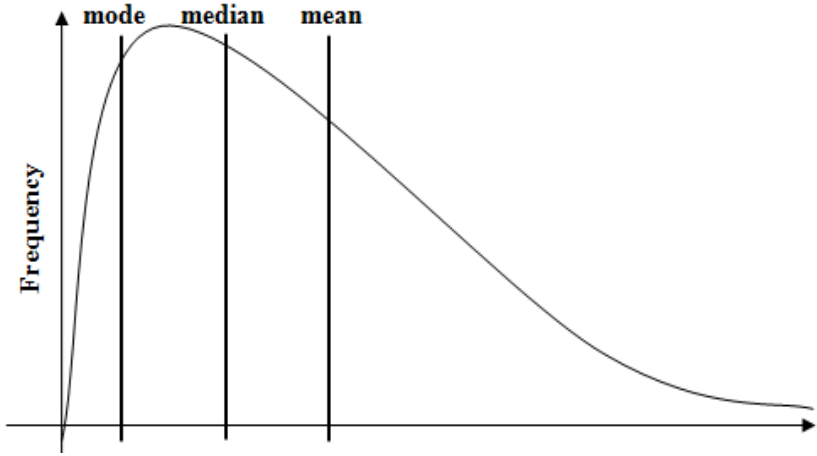

Figure 7.1 *Right skewed or positively data.*

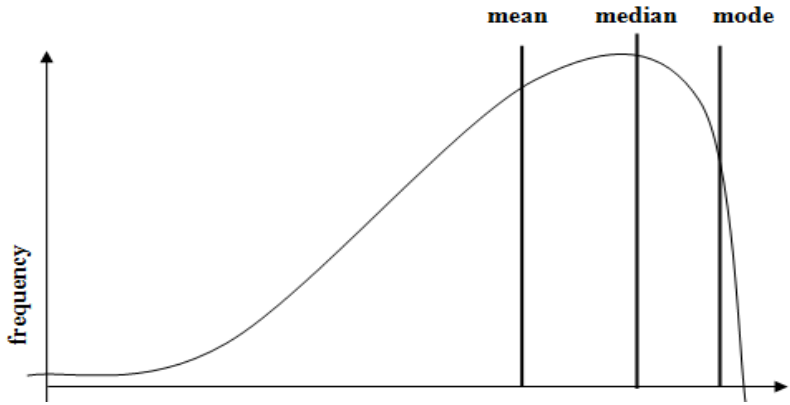

Figure 7.2 *Left skewed or negatively data.*

Table 7.9.1 *STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.*

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.0	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
-2.0	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
-1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
-1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
-1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
-1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
-1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
-1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
-1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08691	.08534	.08379	.08226
-1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
-1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
-1.0	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
-0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
-0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
-0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
-0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
-0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
-0.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
-0.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
-0.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
-0.1	.46017	.45620	.45224	.44828	.44433	.44038	.43644	.43251	.42858	.42465
-0.0	.50000	.49601	.49202	.48803	.48405	.48006	.47608	.47210	.46812	.46414

Table 7.9.2 *STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the RIGHT of the Z score.*

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
2.8	.99744	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861
3.0	.99865	.99869	.99874	.99878	.99882	.99886	.99889	.99893	.99896	.99900
3.1	.99903	.99906	.99910	.99913	.99916	.99918	.99921	.99924	.99926	.99929

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
3.2	.99931	.99934	.99936	.99938	.99940	.99942	.99944	.99946	.99948	.99950
3.3	.99952	.99953	.99955	.99957	.99958	.99960	.99961	.99962	.99964	.99965
3.4	.99966	.99968	.99969	.99970	.99971	.99972	.99973	.99974	.99975	.99976
3.5	.99977	.99978	.99978	.99979	.99980	.99981	.99981	.99982	.99983	.99983
3.6	.99984	.99985	.99985	.99986	.99986	.99987	.99987	.99988	.99988	.99989
3.7	.99989	.99990	.99990	.99990	.99991	.99991	.99992	.99992	.99992	.99992
3.8	.99993	.99993	.99993	.99994	.99994	.99994	.99994	.99995	.99995	.99995
3.9	.99995	.99995	.99996	.99996	.99996	.99996	.99996	.99996	.99997	.99997

Bibliography

[1] Altman, D. G., Boca, R. (1999). *Practical statistics for medical research.* London, New York, Washington D. C.: Chapman & Hall/CRC.

[2] Armitage, P., Berry, G. (1994). *Statistical Methods in Medical Research.* Oxford: Blackwell Scientific Publications.

[3] Bruning, J. L., and Kintz, B. L. (1977). *Computational Handbook of Statistics.* 2nd ed. Glenview, Illinois, Scott, Foresman.

[4] Greenfield, M. L. V. H., Kuhn, J. E., Wojtys, E. M. A. (1997) statistics primer: descriptive measures for continuous data. *Am J Sports Med.*, 25: 720-723.

[5] He, J., Jin, Z., Yu, D. (2009) Statistical reporting in Chinese biomedical journals. *Lancet* 373(9681): 2091-2093.

[6] McHugh, M. L. (2003). Descriptive statistics, part I: level of measurement. *JSPN*, 8: 35-37.

[7] Overholser, B. R., Sowinski, K. M. (2007). Biostatistics primer: part I. *Nutr Clin Pract.* 22: 629-635.

[8] Strasak, A. M., Zaman, Q., Pfeiffer, K. P., Gobel, G., Ulmer, H. (2007). Statistical errors in medical research - a review of common pitfalls. *Swiss Med Wkly 137* (3-4): 44-49.

[9] Strike, P. W. (1991). *Measurement and control, Statistical Methods in Laboratory Medicine.* Oxford: Butterworth-Heinemann.

[10] Wallis, W. A., Roberts, H. V. (1956). *Statistics: A New Approach.* Glencoe, Illinois, Free Press.