Unit 4

Summarising Data: Measures of Central Tendency and Dispersion

Overview

In Unit 3, we examined a few formats for presenting data. In this context, we said that tables and charts are valuable because they can give us a quick impression of the general characteristics of the data. We have also mentioned that tables and charts are inherently limited in terms of presenting the data. However, there are a few mathematical procedures, which, when used alongside the tables and charts, will help us get a better understanding of the data. In this unit, we will be discussing two of these mathematical procedures: central tendency (the central values about which the data appear to be spread, such as the mean, the median and the mode) and the variability or dispersion of the data. We will discuss a few other mathematical procedures in Units 6, 7 and 8.

Unit 4 Learning Objectives

After completing this unit, you should be able to:

- discuss the mean, median and mode as measures of central tendency in statistics;
- calculate the mean, median, and mode of a given data set;
- describe the relationship among the measures of central tendency;
- explain the concept of 'skewness' in statistics;
- discuss the standard deviation and variance as measures of dispersion in statistics;
- determine the semi-interquartile range of a given set of data;
- determine standard deviation (SD) and variance from a set of data.

The sessions in this Unit are:

Session 1: Measures of central tendency.

Session 2: Measures of dispersion.

Session 3: Deriving measures from grouped data.

Note to students

This unit contains several activities and one practice assignment at the end of the unit. You are to work on the activities on your own. If you have any questions or concerns please post a message in the unit discussion forum so that your E-tutor can provide assistance to you. The assignment is to be uploaded in the practice assignment area.

Session 1

Measures of Central Tendency

Introduction

A measure of central tendency is a number which indicates the middle of the distribution of data values. The three main measures of central tendanceis are the mean, the median and the mode. We shall discuss these measures in this session as well as the relationship among them.

Objectives

After completing this session you should be able to:

- calculate the mean, median and mode of a given set of data;
- compare the measures of central tendency;
- explain "skewness" in statistics.

The Mean

It would be convenient if we could summarise all the information about a group of data in one figure. To an extent, we can do this by a measure that can be thought of as the core of a data set—the average. Most often, in our conversation, we use this term. But what do we really mean when we say, "he's an average batsman" or "she's an average student"? When we make such statements as these, what we are really talking about is a typical value. For example, if in three successive innings a batsman or batswoman scores 30, 25, and 11 runs, then the cricketing statistician reports that he or she has made an average of 22. The statistician simply added the three scores and divided them by three. This average value is the mean of the scores. You will have noticed that the mean value was not one of the scores actually obtained.

We will now describe the rules for determining the mean value of a set of data. Let us use the example given above and denote the number of runs scored in any one innings as X. To be more specific, let us define the runs scored as follows:

X_1 = the number of runs scored in the first innings;
X_2 = the number of runs scored in the second innings;
X_5 = the number of runs scored in the fifth innings, and so on.

Generally, we use X_i to refer to the amount scored in the i^{th} inning, where 'i' can be made equal to 1, 2, 5 or n! For three innings, the arithmetic mean of the scores would be:

$$(X_1 + X_2 + X_3) \div 3$$

while for five innings, it would be:

$$(X_1 + X_2 + X_3 + X_4 + X_5) \div 5$$

The symbol $\overline{X}$ (pronounced as X bar), represents the arithmetic mean of the scores. (Please note that $\overline{X}$ is the symbol for *sample* mean. The symbol μ is used for *population* mean. We pronounce this letter as *mu*, and it is a letter in the Greek alphabet.)

For *n* innings, for example, the mean would be:

$$= \quad \overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

It is convenient to use the $\sum$ (sigma) notation to represent a sum. Consider, for instance, the sum:

$$X_1 + X_2 + X_3$$

We may write this as:

$$\sum X_i = X_1 + X_2 + X_3$$

Often, when it is not clear from the context, this may be written as:

$$\sum_{i=1}^{3} X_i$$

As mentioned above, the letter 'i' is used as an index for 1, 2 or 3. For n values, the sum:

$$X_1 + X_2 + \ldots + X_n$$

may be represented as:

$$\sum_{i=1}^{n} X_i$$

We can then present the equation:

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

as:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Both equations say exactly the same thing. Should we not therefore consider them equivalent? Perhaps we should. However, the second equation does the job much more compactly and concisely, and is therefore more convenient to use and preferred by statisticians.

To reiterate, X_i is a collective expression for all the data points in the exercise, with 'i' taking on values from '1' to 'n'. For our example above, n = 3 since our cricketer played three innings. The total of all three innings, therefore, is:

$$\sum_{i=1}^{3} X_i = X_1 + X_2 + X_3$$

and the arithmetic mean is:

$$\overline{X} = \frac{X_1 + X_2 + X_3}{3}$$

That is, the mean is the sum of the scores divided by the number of innings.

Of course, in real statistical problems, *n* is not likely to be 3 but may well be in the order of hundreds. It will be very tedious then to manually calculate the mean value and, therefore, we can revert to the convenience of a statistical package like MINITAB.

Calculating the mean using MINITAB

Let us now move from this simple cricket example to the one we used in Unit 3, i.e., the Mathematics scores of 82 female students (Table 3.3). Suppose that we wish to determine the arithmetic mean of the 82 scores obtained by female students doing the course. Calculating the mean manually or with a pocket calculator will prove to be a tedious process. We may opt for computers, as they can do the computations much faster (and, in fact, can provide us with many more statistics that we will need). Consider the MINITAB output shown in Display 4.1:

Display 4.1
Descriptive Statistics of the
Set of Female Candidates in Mathematics (1994/95)

N	N*	MEAN	MEDIAN	TRMEAN	STDEV
82	10	65.73	70.00	66.78	22.71

SEMEN	MIN	MAX	Q1	Q3
2.51	13.00	98.00	48.00	85.25

Source: MINITAB Output (Adopted)

We have obtained this output using the *Descriptive Statistics* command in the MINITAB computer programme and, as you can see, this command provides a host of descriptive statistics besides the mean value. For example, besides the mean (65.73), Display 4.1 provides the minimum score (13) and the maximum score (98), which we have identified in Unit 3. We have not yet discussed a few other statistics shown in Display 4.1. 'TRMEAN' or **trimmed mean**, for example, is closely related to the mean. This is obtained by deleting approximately the largest 5% and the smallest 5% of the data and then finding the mean of the remaining values. But then, why should we trim the mean? There is often practical justification for trimming the mean in certain circumstances. For instance, it is possible that a data set includes a few extremely large or extremely small values, with the rest of the data clustered in a general range. If this is so, the mean value of the full data set could mislead us, as these untypical or *outlier* values could bias or **skew** the value in a particular direction. Let us illustrate this with a simple example. Suppose that during an academic year, a teacher gave 20 assignments to a class and he or she was to issue an annual score that should be representative of a student's performance over the year. The marks a student scored for the 20 assignments, if arranged in an ascending order, are as follows:

0, 89, 89, 89, 89, 89, 89, 90, 90, 90
90, 90, 90, 90, 91, 91, 91, 91, 91, 91

A quick inspection shows that the student has scored typically between 89 and 91 over the year. However, as you notice, there is also a single mark '0' that is not typical of the student's marks. If we denote a mark as X, then the average of all 20 marks would be:

$$\frac{X_1 + X_2 + X_3 + \ldots + X_{20}}{20}$$

$$= \frac{0 + 89 + 89 + \ldots + 91}{20}$$

$$= 85.5$$

You can see that this is not a typical mark. However, if we trimmed the first and last 5% (5% of 20 = 1), i.e., the first and last mark, we get a mean of 89.9, which is more typical of the marks scored.

By eliminating the outliers, the trimmed mean gives us a more useful indication of the mean value of the data. When using trimmed means, you should bear the following two points in mind:

- if the trimmed mean does not differ considerably from the mean, then we know that the extreme values of the data did not significantly bias the mean calculation;
- if they differ, however, then we know that our data set was characterised by untypical extreme values, which, if left, could lead us to draw erroneous conclusions.

ACTIVITY 4.1

Explain the statistical notion 'measure of central tendency'.

Note:
a) Write your answer in the space given here.
b) Post your answer in the Unit discussion forum.

Let us now look at the second measure of central tendency: the median.

The Median

Let us start with this question. What do we mean when we say that a student is performing below average? This can be interpreted in many ways. For instance, do we mean that the student is more likely to fail than to pass? To say that a student is more likely to fail than to pass, is another way of saying that there is a higher *probability* of the student failing, than passing. On the basis of the discussion so far, you could also take it to mean that when the class arithmetic mean is calculated, this student usually

falls below that standard. There is still another possible interpretation which is that the student's relative performance places him or her in the bottom half of the class. The value that separates the two halves of the sample or population is called the **median** (Me). In more immediate language, the *median* mark is that mark above which we find 50% of the class and below which we find the other 50%.

Suppose, for example, we want to find the median of the following data set:

22
27
8
5
13

The first thing we should do is to rank the data in ascending order, as follows:

5
8
13
22
27

In this manner the median, or middle value, of 13, is obvious. Half of the population (2) is above the median, and the other half is below.

DEFINITION 4.1

The ***median value*** is therefore that value which cuts the population into half.

In this example where we have an odd number of items, the separation is easy. However, it is slightly more complicated with even numbers of items such as in the following data set:

15
30
7
14
3
20

When we arrange them in ascending order the data will appear as under:

3
7
14
15
20
30

What is the median value here? We know that it will lie somewhere between the two

central values of 14 and 15. We assume it to lie halfway between these two values, thus obtaining a median value of 14.5.

As pointed out, the difference in these two examples is that one data set comprised an odd set of numbers, while the other was even. With odd-numbered data sets, the median value is simply the middle value after the data has been ranked in ascending order. With even-numbered data sets, however, the median is found by taking the mean value of the two middle values.

Calculating the median using MINITAB

The MINITAB output in Display 4.1 shows the median value of the marks obtained by all female candidates in the Mathematics examination in the academic year 1994/95 as 70. You may obtain the same result by manually ranking the 82 values and, since 82 is an even number, finding the mean of the two middle values. The two middle values are 69 and 71. By adding these values and dividing the sum by 2, we get the mean (70, here).

If you look at the stem and leaf display (Figure 3.7), you will see that the central value (or values, if an even number of items) can be easily identified from this display. That is to say, a stem and leaf display is another way of calculating the median value. Notice, the cumulative frequency count peaked at 41, half the total number of items. The last value of the first 41 (69) and the first value of the second 41 (71) are the two central values. The median value is therefore 70.

Having acquainted ourselves with the mean and the median, let us now discuss the other measure of central tendency, i.e., the mode. But first, do Activity 4.2.

ACTIVITY 4.2

Compare the mean and the median as measures of central tendency.

Note:
a) Write your answer in the space given here.
b) Post your answer in the Unit discussion forum to recieve feedback from your tutor.

The Mode

What do we imply when we say that 'he or she is of average height'? We will not measure everyone in a defined population (or in a sample), find the arithmetic mean and use this as a measure of the average. Instead, we would measure the most frequently recurring value in a population or sample and this is called the **mode**. We shall illustrate this in the following example:

Suppose that you have given a spelling test to 20 'Common Entrance' students and the scores obtained are:

17, 17, 13, 14, 20, 20, 19, 19, 18, 17, 17, 17, 16, 14, 9, 7, 13, 9, 14, 20

We can present the scores in the form of a frequency table as shown in Table 4.1 below:

Table 4.1
Frequency Distribution: Scores Obtained in a Spelling Test

Score	Frequency
7	1
9	2
13	2
14	3
16	1
17	5
18	1
19	2
20	3
MODE = 17	

The most frequently occurred mark, in this example, which we call the **modal mark**, is 17. It is the most typical of the marks obtained so that any student falling below this score might also be said to be performing below average.

A major limitation of the mode as a measure of central tendency (compared to the mean and median values) is that a data set may not have a unique modal value, and sometimes may not have a mode at all. For example, a data set with every value occurring only once has no mode. Consequently, it is possible to have any of the following:

- **Unimodal**: a data set which has only one value occurring with greater frequency than all others.
- **Bimodal:** a data set that has two values occurring with the same highest frequency.
- **Multimodal:** a data set which contains more than two modes.

This measure of central tendency (the mode) is of most value to us in unimodal situations. We should note, however, that whether the data set is unimodal, bimodal or multimodal, it is statistically inadvisable to use the mode as a measure of central tendency in small sample situations. Note that, unlike the mean and the median, we can determine the mode for qualitative as well as quantitative data.

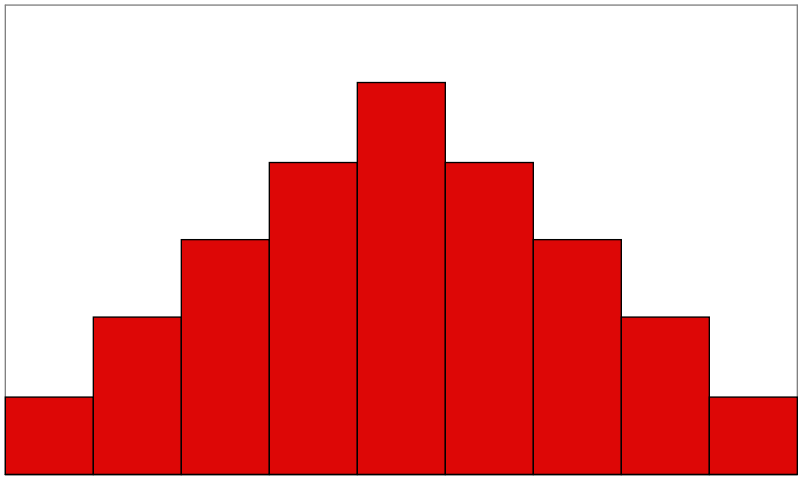
We have said one way of finding the mode is to order and rank the data, and simply read off the most frequently occurring value. Since the stem and leaf display also orders and ranks a data set, we can use it to read the modal value. For instance, if you look at the display in Figure 3.7, you will see that 85 is the modal mark (with a frequency of 4). By this time you may have guessed that there must be some kind of relationship among the mean, the median and the mode. Let us now see what this relationship is.

Relationship among 3M's

From our analysis of the Mathematics scores in Unit 3 (Table 3.3), we found a mean of 65.73, a median of 70 and a mode of 85.

The degree of agreement between these three measures of central tendency can tell us a lot about the degree of symmetry in a data set. Symmetry is probably most obvious with a pictoral display like a histogram, so let us talk a bit more about histograms first. In Unit 3, we discussed the use of the histogram to display a frequency distribution. A frequency distribution chart and a histogram give us an idea of the shape of the distribution of the data set. The word 'distribution', in fact, implies how we can spread a variable across a certain range. We can spread the data either in a symmetric or in an asymmetric manner. When we look at a histogram, we will know whether or not a data set is symmetric or asymmetric. Figure 4.1 below shows us a histogram representing a symmetric distribution:

Figure 4.1
Histogram: Symmetric Distribution

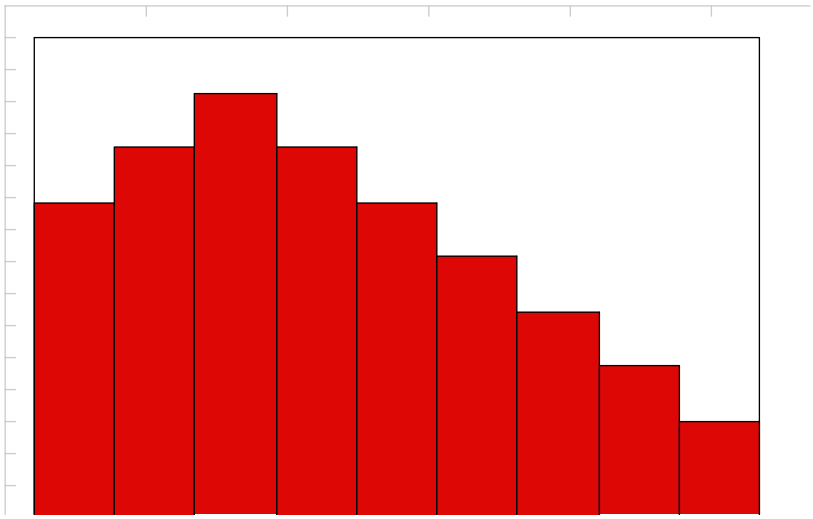


When a data set has a symmetric distribution, then the mean, the median and the mode of the data coincide. So, clearly, our Mathematics scores discussed in Unit 3 do not represent a symmetric data set. If you look again at Figure 3.3, you will notice this.

Skew

If a distribution is not symmetric (i.e., asymmetric), then we say that it is *skewed*, and the corresponding histogram will slant in a particular direction. Obviously, the distribution curve can lean either to the left or to the right. We can thus distinguish between a *right-* or *positively-skewed* distribution and a *left* - or *negatively-skewed* distribution. Figure 4.2 below illustrates a positively skewed distribution:

Figure 4.2
Histogram: Positively Skewed Distribution

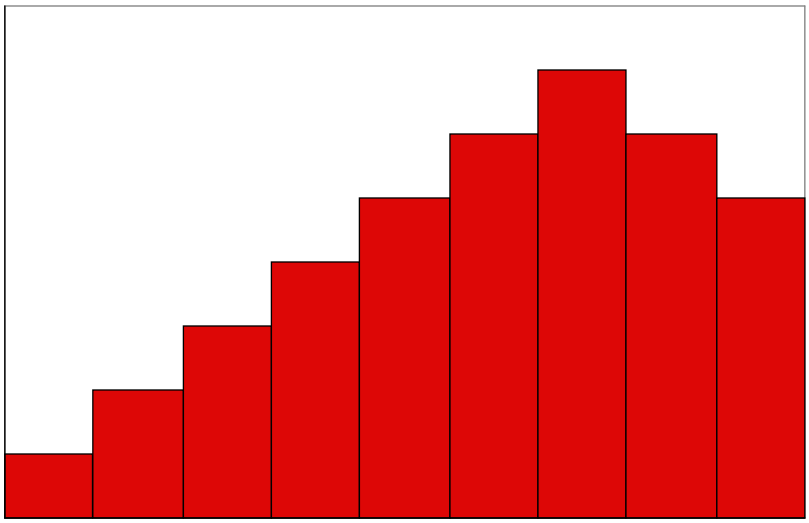


When a data set is positively-skewed, we can find the mean at the right side of the curve. It is easily verified in this case that the following holds:

mode < median < mean

This means, the mode *is less than* the median; the median *is less than* the mean. The mathematical notation < stands for is less than.

Now, let us look at a negatively skewed distribution:



When a data set is negatively-skewed, then the mean is at the left side of the curve. It is easily verified in this case that the following holds:

mean < median < mode

This means, the mean *is less than* the median; the median *is less than* the mode.

If you look again at the histogram in Figure 4.3, you will notice that it is negatively skewed and that the above relationship holds. That is:

mean = 65.73 < median = 70 < mode = 85

Having got some idea about the three measures of central tendency, i.e., the mean, the median and the mode, and the relationships among them, let us now move on to the other statistical measure. Put differently, now that we have located the centre of a distribution, we should also know the measure of the dispersion or variability of the data. We will study dispersion in the following session.

ACTIVITY 4.3

Explain the concept of skewness.

Note:
Write your answer in the space given here and then post the answer in the discussion forum.

Summary

We defined a measure of central tendency as a number which indicates the middle of the distribution of data values. We described the three main measures of central tendency, namely the mean, median and mode. We can therefore find a 'central ' value in a data set according to one of three interpretations. If by 'central' we mean arithmetic avreage then we must use the mean; if by 'central' we mean the typical value we must use the mode; and if by central we mean the middle value, we must use the median In this session we also stated that when a data set has symmetric distribution, then the measures of central tendency coincide. If the distribution is not symmetrical then we say it is skewed.

Session 2

Measures Of Dispersion

Introduction

In Session 1 we developed measures of central tendency. In this session, we change our focus to dispersion or spread of the data and seek to develop measures of that spread. There are several measures of dispersion or spread. We will start off the discussion with *range, quartile deviation* and move on to *standard deviation*.. In this session our focus will be on the dispersion of data.

Objectives

On completing this session students should be able to:

- explain range and quartile deviation,
- calculate the mean, varience and standard varation, given a set of numbers.

Using Measures of Dispersion

Consider two hypothetical villages, A and B, in each of which only three individuals live. Table 4.2 below shows the annual income of each inhabitant of both villages:

Table 4.2
Income Distribution in Villages A and B

	Annual Income ($1000)			
Village	Individual 1	Individual 2	Individual 3	Mean Income ($1000)
A	3	4	5	4
B	1	4	7	4

The mean income for both villages is $4000. (It is only co-incidental that the mean value is the same as one of the actual values.) Clearly, if we were to rely solely on the mean measure as an index of well-being in both the villages, we would conclude that this is identical. And while there is some truth in this, it is also quite clear that the two situations are significantly different. Indeed, it is obvious that there is a greater variability in the income in B than in A. We may conclude that B has a more unequal distribution of income than A. For a more accurate picture of the income distribution in the two villages we need to combine information about the mean value with

information about the dispersion or spread of income in each village. We shall do so next.

The Range

Perhaps the most rudimentary measure of dispersion (that is, the spread of the data values) is the **range**. In statistics, we calculate the range as the difference between the highest and the lowest recorded values in a data set. For village A, for example, the range is calculated as $2000 ($5000 - $3000), while for village B, it is calculated as $6000 ($7000 - $1000). The greater dispersion in village B indicates, among other things, that the mean value of $4000 is less reliable as an index of well-being in village B than it is in village A. So, we can express the above in the form of an equation as:

Range = Maximum value - Minimum value

It is important that you remember this equation. Now, let us apply this to our example: the scores of female candidates in Mathematics. Display 4.1 gives us the maximum and minimum values. What is the range? The range is 85 (maximum (98) - minimum (13)).

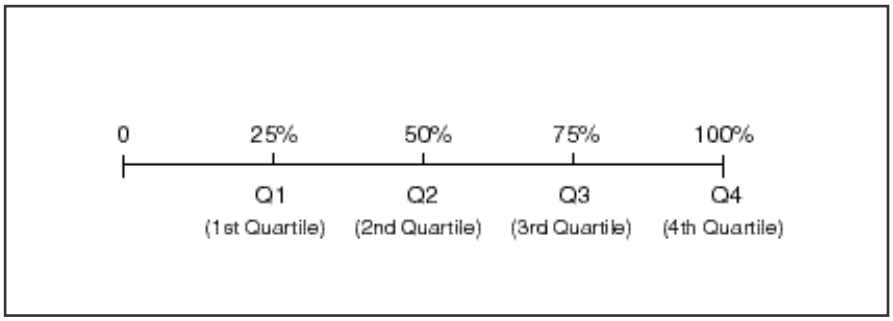
The Quartile Deviation

The presence of a few extreme outliers can distort the usefulness of the range as a measure of dispersion. Look at the example, introduced earlier, of 20 assignment marks. The range of the marks is 91 but you can see that all but one of the marks lie between 89 and 91. The single score of 0 has distorted the range and hence it does not give an accurate picture of the spread of marks. A possible solution to this dilemma is to calculate a trimmed range. A frequently employed adaptation of such a procedure, is to eliminate the lowest and highest 25% of the values and to consider only the range of the remaining values. We call the measure so obtained the **interquartile range** (IQR) because its lower and upper limits are called, respectively, the first and third quartiles: 25% of the distribution is to be found below the first quartile (Q1) and 25% above the third quartile (Q3).

$$IQR = Q_3 - Q_1$$

See the illustration below.

Figure 4.4
Quartiles



At times, we refer to the median as the *50th percentile* – for obvious reasons. We can also speak of the 10th percentile, the level below which 10% of the population is found, the 20th percentile, the 30th percentile, and so on. Nothing stops us from talking about all different percentiles. The more frequently used percentile measures are, however, the 25th percentile and the 75th percentile. You should remember that the 25th percentile or the level below which 25% of the data can be found, is called the *first quartile*, while the 50% percentile is known as the *second quartile* (which can be referred to as the median) and the 75% percentile is the *third quartile* and so on.

The interquartile range is not often used in statistics but the semi interquartile range, also known as the **quartile deviation** (QD), is used more frequently. This can be calculated as follows:

$$QD = \frac{Q_3 - Q_1}{2}$$

The quartile deviation (QD) is used as a dispersion measure when the median is used as the measure of central tendency. You should note that the median (Me) is used as a measure of central tendency with skewed distributions. Note also that [Me ± QD] accounts for 50% of the distribution. That is (Me - QD) to (Me + QD) should represent 50% of the distribution.

Let us now revisit our earlier example again and calculate these dispersion measures for the marks obtained by female candidates in the 1994/5 mathematics examination. First, we can read Q1 (first quartile) and Q3 (third quartile) directly from the output in Display 4.1, i.e., 48 and 85.25, respectively. To measure the interquartile range (IQR), we find the difference between these two values, IQR= Q3 – Q1 = 85.25 – 48 = 37.25. We can then determine that QD = 37.25/2 = 18.625, since:

$$\frac{Q_3 - Q_1}{2} = QD$$

The two quartiles (Q1 and Q3), the median (which, if you recall, is really the second quartile), together with the maximum and minimum values of the data set comprise what is known as the five number summary of a data set which you met in unit 3. If you recall, we refer to these measurements with regard to the boxplot in unit 3. We said that the two extreme points in the boxplot represent the minimum and maximum values of the data set. This we can corroborate by referring to Table 3.1 as well as the stem and leaf display (Figure 3.7).

ACTIVITY 4.4

Describe the range and the interquartile range.

Note:
Write your answer in the space given here and then post ithe answer in the discussion forum.

Deviations From the Mean

This suggests, as the name implies, that we are talking about how each value deviates or differs from the mean value. We said before that the mean is the most popular measure of central tendency and it would seem only reasonable that, for greater accuracy, it should be accompanied by a measure of the dispersion or spread of the data. To do this we determine the distance of each individual data-point from the mean value. Let us illustrate this using the income data for village A (Table 4.1), and derive the **deviations from the mean** as shown in Table 4.3 below:

Table 4.3
Deviations from Mean Income: Village A

Income ($1000)	Deviations from mean income ($1000) Mean Value = $4000
3	-1
4	0
5	1

Table 4.4 below shows a similar set of statistics for village B:

Table 4.4
Deviations from Mean Income: Village B

Income	Deviations from mean income
1	-3
4	0
7	3

You may have guessed that the deviation from the mean is calculated as

(Actual value - Mean value).

The deviations for village B are always greater than those for village A, confirming once again the greater inequality of income in village B. As the data set becomes larger, it becomes more tedious to make point by point comparisons, especially if there is no point by point superiority of one set over another. What we need is a summary measure, something like an average of the deviations from the mean. Such a measure is the **mean absolute deviation** (MAD) and this is described below.

The Mean Absolute Deviation

Remember that the deviations from the mean are also measures of dispersion. It is tempting to get a summary measure by adding them up and dividing by 3 to get an average value of this dispersion. But when you do so, do you notice what happens? The deviations sum to zero. This is, in fact, a consistent result of summing deviations from the mean. This could be used as a check to see if your calculations of deviations from the mean are correct. If they do not sum to zero then they are incorrect and should be corrected.

An obvious solution in this case would be to ignore the sign attached to the deviations and calculate the mean of the sum of the absolute values of the deviations from the mean. In our example, we would obtain for village A:

$$\frac{1+0+1}{3} = \frac{2}{3}$$

And for village B:

$$\frac{3+0+3}{3} = \frac{6}{3}$$

These figures are examples of a measure of dispersion known as the *mean absolute deviation.*

A more general expression for the mean absolute deviation (MAD) is:

$$\text{MAD} = \frac{|X_1 - \overline{X}| + |X_2 - \overline{X}| + \ldots + |X_n - \overline{X}|}{n}$$

where the expression |X| means 'the absolute value of X'.

There are at least two immediate problems associated with this measure:

- it is mathematically clumsy (for reasons which we cannot fully elaborate here);
- greater deviations from the mean are penalised only to the extent of the distance from the mean.

The second point needs further clarification. It means that a deviation of 4, for example, has attached to it a penalty that is four times larger than a deviation of 1. Note, however, that values above or below the mean are penalised to the same extent.

A word of caution. Despite these problems, the mean absolute deviation is a legitimate representative measure of dispersion. Its advantage is that, conceptually, it is easy to understand. It is useful when we are interested merely in representing dispersion. However, if we are interested in further statistical analysis, it is not sufficient and we use the more rigorous measures of variance and standard deviation.

The variance and standard deviation

By far the most popular measure of dispersion is the standard deviation which is derived from the variance. It overcomes the shortcomings associated with the mean absolute deviation we discussed above. Instead of the absolute value of the deviation from the mean, the variance is calculated from squared deviations. Table 4.5 below shows the deviations and squared deviations from the mean for villages A and B:

Table 4.5
Deviations and Squared Deviations from Mean Income: Villages A and B

Village A		Village B	
Deviations	Sq. Deviations	Deviations	Sq. Deviations
-1	1	-3	9
0	0	0	0
1	1	3	9

The variance for village A is calculated as:

$$V_A = \frac{1+0+1}{3} = \frac{2}{3}$$

(That is, the mean of the squares of deviations from the mean.)

And for village B as:

$$V_B = \frac{9+0+9}{3} = 6$$

Generally, for 'n' observations, we can calculate the variance of some variable X as:

$$\text{Var}(X) = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2}{n} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n}$$

By squaring the deviations from the mean, we eliminate the problem associated with possible negative values. Further, we also attach a greater penalty to greater deviations from the mean, a penalty equivalent to the square of the value.

To summarise, the variance is the average of the squares of the deviations from the mean. Mathematically, this is written as:

$$(X) = \frac{(X_1 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2}{n}$$

There is one major inconvenience in using the variance as a measure of dispersion: it is not expressed in the same 'unit measures' as that of the original variable under consideration. For example, in our examples involving villages A and B, the variance is measured in 'thousands of dollars squared' while the original income values are in 'thousands of dollars'. The unit of the variance is always the square of the unit of the individual data values. Taking the square root would give us a measure expressed in 'thousands of dollars', which is the same 'unit of measure' in which the original variable is measured. We define the statistic so derived as the *standard deviation* (SD). The SD is therefore the positive square root of the variance. The standard deviation of the income of village A is, consequently:

$$SD_A = \sqrt{\frac{2}{3}} = 0.8165$$

and village B is:

$$SD_B = \sqrt{6} = 2.449$$

We say that village A's income is distributed with a mean of $4000 and a standard deviation of $816.50 while village B's is distributed with a mean of $4000 and a standard deviation of $2449. The wider dispersion in village B is an indication that the mean is not so reliable as a measure of economic well being as it is for village A.

For a given variable X, the general formula for the standard deviation of X is:

$$SD(X) = \sqrt{\text{Var}(X)}$$

Before completing our discussion on the standard deviation, look again at Display 4.1. Under the head STDEV, the display shows the standard deviation of the marks obtained by female students in the Mathematics examination. Its value is 22.71.

Comparing dispersion of more than one data set

If given two or more datasets and you're required to compare their dispersions, the crude approach will be to compare the ranges but as previously mentioned, this is not reliable as the range focuses only on the MIN and MAX values in each dataset. THe more reliable approach will be to compare their standard deviations. The greater the standard deviation, the greater the dispersion within the dataset.

ACTIVITY 4.5

Find the mean, variance and standard deviation of the following numbers, using a pocket calculator:

4, 4, 7, 10, 10

Note:

Write your answer in the space given here.

Summary

There are several ways of measuring dispersion of data. In this session we reviewed the rudimentary as well as the more popular methods.

Session 3

Deriving Measures from Grouped Data

Introduction

We get raw or *ungrouped* data only from primary data. The data obtained from secondary sources may have already been processed and *grouped* in a particular format – as we have grouped the examination score data in Tables 3.3 and 3.4. In such circumstances, we have to rely on methods other than the ones we have already discussed to extract the relevant measures of central tendency and dispersion. This is the subject of the session.

A word of caution. (Note that neither MINITAB nor any other statistical package is programmed to use the techniques that we are about to discuss. However, you may use the pocket calculator for this.)

Objectives

On completing this session students should be able to:

- calculate the mean of a set of data by using frequencies;
- calculate the median from grouped data;
- calculate the mode from grouped data.

Calculating the Mean Using Frequencies

Let us introduce some of the tools for working with grouped data (such as those grouped in class intervals) by continuing, for the moment, to work with ungrouped data. Consider the following series:

7, 7, 9, 9, 9, 13, 14, 14, 16, 16, 17, 17, 21

We can find the mean of this data-set the usual way—by adding all the values and dividing by the total number of data points. The average we get is:

(7+7+9+9+9+13+14+14+16+16+17+17+21)÷13
= 169 ÷ 13
= 13

As you see, we have two sevens, three nines, one thirteen, two fourteens, two sixteens, two seventeens, and one twenty-one.

Suppose we group the identical digits as shown below and then calculate the average, should it be different from the one we calculated above?

[(7x2) + (9x3) + (13x1) + (14x2) + (16x2) + (17x2) + (21x1)]
÷ (2+3+1+2+1+3+1)

What have we done here? On the left hand side of the ÷ sign, rather than adding the individual values together, we have taken a short–cut: first, multiplying each value by the frequency with which the value occurs and then adding these products together. On the right hand side of the ÷ sign, we sum the frequencies of each variable which give us the total number of observations (here equal to 13). The following table summarises the procedure:

Table 4.6
Calculating the Mean Using Frequencies

Value (X)	Frequency (f)	f•X
7	2	14
9	3	27
13	1	13
14	2	28
16	2	32
17	2	34
21	1	21
TOTAL	13	169

Average = 169/13 = 13

In more general terms and for 'n' observations, let f_i be the frequency with which the value X_i occurs. Then the mean is calculated as:

$$\overline{X} = \frac{f_1 X_1 + f_2 X_2 + \ldots + f_n X_n}{f_1 + f_2 + \ldots + f_n} = \frac{\sum_{i=1}^{n} f_i X_i}{\sum_{i=1}^{n} f_i}$$

(Note that $\sum_{i=1}^{n} f_i = n$

To calculate the variance as shown in Table 4.7, we can also use a similar procedure.

Table 4.7
Calculation of the Variance Using Frequencies

X	f	$(X - \overline{X})^2$	$f \bullet (X - \overline{X})^2$
7	2	36	72
9	3	16	48
13	1	0	0
14	2	1	2
16	2	9	18
17	2	16	32
21	1	64	64
TOTAL	13		236

Variance = 236/13 = 18.15

In terms of a formulae, it can be written as:

$$\text{Var}(X) = \frac{f_1(X_1 - \overline{X})^2 + f_2(X_2 - \overline{X})^2 + \ldots + f_n(X_n - \overline{X})^2}{f_1 + f_2 + \ldots + f_n} = \frac{\sum_{i=1}^{n} f_i(X_i - \overline{X})^2}{\sum_{i=1}^{n} f_i}$$

The standard deviation is the square root of this value (18.15), which is 4.32.

The Mean from Grouped Data

Consider the data in Table 3.4 where the scores obtained by female students in the Mathematics examination are grouped in a particular way. We have reproduced this grouping in Table 4.7 but we also include the mid-point (or class mark) of each interval. Why should we do that? We assume that, in the absence of the detailed information shown in Table 3.2, all students who have a mark in this group have a mark equal to the mid-point mark. This allows us to apply the formula worked out above, i.e.,

$$\overline{X} = \frac{f_1 X_1 + f_2 X_2 + \ldots + f_n X_n}{f_1 + f_2 + \ldots + f_n} = \frac{\sum_{i=1}^{n} f_i X_i}{\sum_{i=1}^{n} f_i}$$

as is done in the last column of Table 4.8:

Table 4.8
Calculation of Mean From Grouped Data
(Marks Obtained by Females in the Mathematics Exam)

Class Interval	Midpoint (X)	Frequency (f)	f•X
10-20	15	4	60
20-30	25	2	50
30-40	35	4	140
40-50	45	12	540
50-60	55	9	495
60-70	65	10	650
70-80	75	14	1050
80-90	85	11	935
90-100	95	16	1520
TOTAL		82	5440

Average = 5440/82 = 66.34

In a similar fashion, we can calculate the variance from the information shown in Table 4.9 below:

Table 4.9
Calculation of Variance From Grouped Data
(Marks Obtained by Females in the Mathematics Exam)

Class Interval	X	f	$(X - \overline{X})^2$	$f \cdot (X - \overline{X})^2$
10-20	15	4	2635.946	10543.784
20-30	25	2	1709.116	3418.232
30-40	35	4	982.2871	3929.1484
40-50	45	12	455.4579	5465.4948
50-60	55	9	128.6287	1157.6583
60-70	65	10	1.799515	17.995150
70-80	75	14	74.97031	1049.58434
80-90	85	11	348.1411	3829.5521
90-100	95	16	821.3119	13140.9904
TOTAL		82		42552.439490

Variance = 42552.44/82 = 518.9
Standard Deviation = $\sqrt{518.93}$ = 22.8

For this data set, therefore, we have two sets of calculations for the mean and standard deviation. We have the values from Display 4.1, which are MINITAB's calculations from the *raw* or *ungrouped* data. And we have the above values, which are our manual calculations from the *grouped* data.

Table 4.10
Mean and Standard Deviations from Grouped and Ungrouped Data: A Comparison (Marks Obtained by Female Students in the Mathematics Exam)

Measures	Calculation from ungrouped data	Calculation from grouped data
Mean	66.1	66.34
Standard Deviation	22.7	22.8

You will notice that the extrapolation or estimation of the values from the grouped data is quite good, in that the results obtained are extremely close to the raw data calculations.

The Median from Grouped Data

Take a look at the grouped data displayed in Table 3.4 where we tabulated the cumulative frequency as well as the frequency. How do we determine the median value from this grouped data? Remember, the median is the 50th percentile of the data set – the value that divides the *population* into two equal groups. Clearly, then, the median will lie in the *class* 60-70, which includes the 41st value. But how do we extract a median *value* from this interval?

To do this, we can use the following formula (or a variant of it):

$$Me = L_1 + \left(\frac{\frac{n}{2} - \left(\sum f\right)_1}{f_{median}} \right) C$$

where:

L_1 = value of lower limit of median class interval

n = number of observations in complete data set

$\left(\sum f\right)_1$ = cumulative frequency up to but not including the median interval

= frequency in the median class

C = length of the class interval

From Table 3.4: $L_1 = 60$, $n = 82$, $\left(\sum f\right)_1 = 31$, $f_{median} = 10$ and $C = 10$.

Substituting these values in the previous equation, i.e.,

$$Me = L_1 + \left(\frac{\frac{n}{2} - \left(\sum f\right)_1}{f_{median}} \right) C$$

we get:

$$Me = 60 + \left(\frac{\frac{82}{2} - 31}{10} \right) x\ 10 = 70$$

If you look at Display 4.1, you will see that the median value obtained from the raw data is also 70.

The Mode from Grouped Data

Look again at Table 3.4. Clearly, the modal class is 90-100, as this is the class interval with the highest frequency. A formula similar to the one above can be used for the extraction of the modal value from this modal class:

$$M_o = L_1 + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) x\ C$$

where:

L_1 = lower limit of the class interval

Δ_1 = frequency of modal class—frequency of class just below (Δ is the Greek letter, delta, in upper case)

Δ_2 = frequency of modal class - frequency of class just above

C = length of the class interval

From Table 3.4, Δ_1 = 16-11 = 5, Δ_2 = 16-0 = 16 and, of course, C = 10. Applying this formula we get:

$$M_o = 90 + \left(\frac{5}{5 + 16} \right) x\ 10 = 92.4$$

Note that this is probably a better measure of the central tendency than would have been obtained from the raw data, that is, if we had counted up the individual marks, (in which case we will find the mode to be 85) since it involves 16 values rather than 3.

The Quartile Deviation from Grouped Data

We have said that the median is the 50th percentile. We can therefore calculate the 25th and 75th percentiles (which are used in the calculation of the quartile deviation) from grouped data in a similar fashion to the median. These are shown below.

For the first quartile:

$$Q_1 = L_1 + \left(\frac{\frac{n}{4} - \left(\sum f\right)_1}{f_{Q_1}} \right) \times C$$

where:

L_1 = lower value or class interval

n = number of observations in data set

$\left(\sum f\right)_1$ = cumulative frequency up to but not including the quartile interval

f_{Q_1} = frequency of class interval containing first quartile

and so:

$$Q_1 = 40 + \left(\frac{\frac{82}{4} - 10}{12} \right) \times 10 = 48.75$$

And, for the third quartile:

$$Q_3 = L_1 + \left(\frac{\frac{3n}{4} - \left(\sum f\right)_1}{f_{Q_3}} \right) \times C$$

where = frequency of interval containing third quartile.

and so:

$$Q_3 = 80 + \left(\frac{\frac{(82 \times 3)}{4} - 55}{11} \right) \times 10 = 85.91$$

We therefore obtain:

$$QD = \frac{Q_3 - Q_1}{2} = \frac{85.91 - 48.75}{2} = 18.58$$

Let us now see the following table:

Table 4.11
Quartiles and Quartile Deviation from Grouped and Ungrouped Data: A Comparison (Marks Obtained by Females in the Mathematics Exam)

Measures	Calculation from ungrouped data	Calculation from grouped data
Q_1	48	48.75
Q_3	85.25	85.91
QD	18.625	18.58

Again, notice the closeness of the two sets of values, reinforcing our conclusion that the extrapolation of these values from grouped data is quite accurate.

ACTIVITY 4.6

State why measures of dispersion are important in Statistics.

Note:
Write your answer in the space given here and in the unit discussion forum.

Wrap Up

In this unit, we have discussed two sets of measures in Statistics:

- measures of central tendency;
- measures of dispersion.

For central tendency, we discussed the mean, the median and the mode. We explained how these measures are related in terms of the skewness of the data and showed that we can use the MINITAB programme to calculate these measures of central tendency. We have used histograms, and stem and leaf displays to illustrate the relationship between central tendency and skewness. We have also studied measures of dispersion such as range, interquartile range, semi-interquartile range (or quartile deviation), deviation from the mean, mean absolute deviation, variance and standard deviation. Finally, we compared these measures for grouped and ungrouped data. We explained the concept of range with reference to semi-interquartile range and its use when the distribution is skewed.