DESCRIPTIVE STATISTICS

1 INTRODUCTION

Numbers and quantification offer us a very special language which enables us to express ourselves in exact terms. This language is called *Mathematics*. We will now learn the basic rules of Mathematics in order to communicate effectively with figures. A huge part of psychological research deals with statistical analysis so that one needs an adequate mathematical background to understand statistical computations.

1.1 Pocket calculator

For this course, you will need a *scientific* calculator, that is, one that has *statistical functions* and, more preferably, one having the *linear regression* (LR) *mode*. The most cost-effective calculator for the purpose of this section is the **CASIO FX-82 TL**. It will save you a tremendous amount of time – once statistical data entered, statistics like the number of observations, mean, standard deviation, correlation and regression coefficients can be readily obtained by just pressing buttons. Obviously, computer software like SPSS or SAS are much more powerful but the calculator can help you to determine basic statistics very quickly 'on the spot'.

1.2 Summation notation

The summation notation is used to summarise a *series*, that is, the sum of the terms of a *sequence*. It is denoted by Greek capital letter sigma, $\sum$, as opposed to small letter sigma, σ, which, in Statistics, stands for standard deviation.

Sigma is most of the time seen in the following form:

$$\sum_{r=a}^{b} f(r)$$

where r is known as the index, a and b are the lower and upper limits of summation respectively and $f(r)$ is known as the general term. r, just like a counter, starts at a and increases by steps of 1 until it reaches b. Each term of the series is obtained by substituting successive values of r in the general term. The following example illustrates the mechanism.

Example

$$\sum_{k=2}^{6}(2k+1) = [2(2)+1]+[2(3)+1]+...+[2(6)+1] = 5+7+9+11+13 = 45.$$

Here, the index (counter) is *k*. It can be observed that *k* takes on an initial value of 2 (the lower limit) and increases by steps of 1 until it reaches the upper limit 6. Every value that *k* assumes is substituted in the general term (2*k* + 1) in order to generate a term of the series. Obviously, the terms are added up since Sigma stands for summation.

In Statistics, however, we do not actually evaluate such expressions numerically but rather use the summation notation strictly for summarisation purposes. This is because the upper limit is generally non-numerical, that is, a variable. We deal mostly with expressions of the form $\sum_{i=1}^{n} x_i$. If expanded, this summation cannot be evaluated since it only gives the expression $x_1 + x_2 + x_3 + ... + x_{n-1} + x_n$.

Such expressions are found in the formulae for arithmetic mean and standard deviation. In this module, students are simply required to recognise the summation notation and understand its meaning so that they can at least use relevant statistical functions on calculators.

2 DISTRIBUTIONS

A distribution is a set of observations which have been classified and organised in an attempt to display information or calculate descriptive statistics. A *frequency distribution* of grouped data is a good example of a distribution.

2.1 Ungrouped data

This type of information occurs as individual observations, usually as a table or array of disorderly values (**Fig. 2.1.1**). These observations are to be firstly arranged in some order (ascending or descending if they are numerical) or simply 'grouped' together in the form of a *discrete frequency table* (**Fig. 2.1.2**), which is unlike a *continuous frequency table*, before proper presentation on diagrams is possible. We do not lose any information if the original data is arranged in an array or grouped as a discrete frequency table.

2	7	8	11	15
16	18	19	19	19
23	23	24	26	27
29	33	40	44	47
49	51	54	63	68

Fig. 2.1.1 Array

Age	Frequency
19	14
20	23
21	134
22	149
23	71
24	8
Total	**399**

Fig. 2.1.2 Discrete frequency table

2.2 Grouped data

When the range of *values* (not observations) is too wide, a discrete frequency table starts to become quite lengthy and cumbersome. Observations are then grouped into *cells* or *classes* in order to compress the set of data for more suitable tabulation. In this case, the data from **Fig. 2.1.2** would not be a good illustration, given the little variation in ages of students (from 19 to 24).

Age group	Real limits	Mid-class value	Frequency
21 – 25	20.5 – 25.5	23	5
26 – 30	25.5 – 30.5	28	12
31 – 35	30.5 – 35.5	33	23
36 – 40	35.5 – 40.5	38	39
41 – 45	40.5 – 45.5	43	32
46 – 50	45.5 – 50.5	48	21
51 – 55	50.5 – 55.5	53	9
56 – 60	55.5 – 60.5	58	2
Total			**143**

Fig. 2.2.1 Continuous frequency distribution

The main drawback in grouping of data is that the identity (value) of each observation is lost so that important descriptive statistics like the *mean* and *standard deviation* can only be *estimated* and not exactly calculated. For example, if the age group '21–25' has frequency 5 (**Fig. 2.2.1**), nothing can be said about the values of these 5 observations. Besides, a lot of new quantities have to be calculated in order to satisfy statistical calculations and analyses as will be explained in the following sections.

2.2.1 **Limits and real limits (or boundaries)**

A class is bounded by a lower and an upper limit – in the previous paragraph, the lower and upper *limits* of the age group '21–25' are 21 and 25 respectively. A real limit (**Fig. 2.2.1**) is obtained by making a *continuity correction* to a limit (explained below). In a frequency distribution, we differentiate between limits and real limits by the fact that *the upper limit of a cell can never be equal to the lower limit of the next cell*. Real limits are *fictitious* values if the values recorded are discrete. However, they are useful not only for the purpose of calculations but also for presentation of data on *histograms* as well as several other types of charts and diagrams.

For instance, if we have a frequency distribution of ages in which we have the two neighbouring cells '21–25' and '26–30', then drawing a histogram for this distribution will require that the limits 25 and 26 be equal, the reason being that there is no 'gap' between any two successive rectangles of a histogram! We therefore make a continuity correction of ± 0.5, the equivalent of half a 'gap'.

Note The 'gap' between any pair of successive cells in a frequency distribution is equal to the degree of accuracy to which the original observations were recorded.

In the above example, it is easy to deduce that age was recorded to the *nearest unit* since the 'gap' between the cells '21–25' and '26–30' is 1. The real limits of these 2 will now be '20.5–25.5' and '25.5–30.5'. Note that the following relationships hold:

Lower real limit = Lower limit – continuity correction
Upper real limit = Upper limit + continuity correction

2.2.2 **Mid-class values (MCV)**

The mid-class value, MCV, of a cell is defined as its *midpoint*, that is, the *average* of its limits or real limits. Thus, the MCV of the cell '21–25' is 23. The MCV of a cell is the representative of that cell in the sense that, since the values of all the observations in the cell are unknown individually, *it is assumed that they are all equal to the MCV*. This assumption is not fortuitous and neither is it unjustified. It has the logical implication that if observations are unknown, the best way of estimating statistics more accurately would be to assume that, at least, they are *uniformly distributed* within the cell (which could be untrue, of course!). Mathematically, the sum of the observations would be equal to the number of observations multiplied by the MCV (think about it!). The importance of the mid-class value can thus never be underestimated, especially for the calculation of the crucial statistics like the mean and standard deviation.

2.2.3 **Class interval or cell width**

The cell width is simply the length of the cell, that is, the difference between its lower and upper *real limits*.

Note Do not make the mistake of subtracting the lower limit from the upper limit since this will not give the exact cell width.

This can be easily verified by taking the cell '21–25'. Its cell width is 5 (21, 22, 23, 24 and 25), which is obtained by subtracting 20.5 from 25.5. We therefore use the following formula:

$$\textit{Cell width} = \textit{Upper real limit} - \textit{Lower real limit}$$

3 DESCRIPTION OF A DISTRIBUTION

A distribution is usually defined in terms of very precisely calculated *statistics* like the mean and standard deviation. The main objective of descriptive statistics is to be able to summarise an entire set of data, grouped or ungrouped, in terms of a few figures only. Summary statistics must be *powerful* and *explicit* enough to paint a global idea of a distribution, especially for the non-statistician. In general, a distribution is described in terms of four main characteristics:

1. Location
2. Dispersion
3. Skewness
4. Kurtosis

3.1 LOCATION (LOCALITY OR CENTRAL TENDENCY)

A measure of location, otherwise known as central tendency, is a point in a distribution that corresponds to a typical, representative or middle score in that distribution. The most common measures of location are the *mean* (arithmetic), *median* and *mode*.

3.1.1 Arithmetic mean

The arithmetic mean is the most common form of average. For a given set of data, it is defined as the sum of the values of all the observations divided by the total number of observations. The mean is denoted by $\bar{x}$ for a sample and by μ for a population. Its formula, however, differs for ungrouped and grouped data.

Ungrouped data

$$\bar{x} = \frac{\sum x}{n} \qquad \mu = \frac{\sum X}{N}$$

Grouped data

$$\bar{x} = \frac{\sum fx}{\sum f} \qquad \mu = \frac{\sum fX}{N}$$

n = sample size
N = population size
f = frequency of classes

Merits

1. It is widely understood.
2. Its calculation involves all observations.
3. It is suited to further statistical analysis.

Limitations

1. It cannot be located by inspection nor can it be found graphically.
2. Its value may be purely theoretical.
3. It is sensitive to extreme values.
4. It is not applicable for qualitative data.

3.1.2 **Geometric mean**

The geometric mean is a specialised measure of location. It is used to measure *proportional changes* in, for example, wages or prices of goods.

The geometric mean of n items is defined as the n^{th} root of their combined product. The general formula which is used to calculate the geometric mean is as follows:

$$\text{Geometric mean} = \sqrt[n]{x_1 \times x_2 \times x_3 \ldots \times x_n}$$

where n is the number of items to be averaged and
$x_1, x_2, x_3, \ldots, x_n$ are the individual values of the items to be averaged.

The best way to demonstrate the geometric mean when it is used to calculate proportional increases is by means of an example.

Example

The price of a particular commodity has been increasing over a four-year period as follows.

\$84 \$97 \$116 \$129

The proportional increases from each year to the next are

$$\frac{97-84}{84} = 0.155 = p_1$$

$$\frac{116-97}{97} = 0.196 = p_2$$

$$\frac{129-117}{117} = 0.112 = p_3$$

$$\text{Geometric mean} = \sqrt[n]{(1+p_1)(1+p_2)\ldots(1+p_n)}$$

$$= \sqrt[3]{1.155 \times 1.196 \times 1.112} = 1.154$$

Average proportional increase = 1.154 – 1 = 15.4%

Note $\$84 \times 1.154^3 = \$129$.

Merit

It takes little account of extreme values.

Limitation

It cannot be applied if the data contains zero values.

3.1.3 **Harmonic mean**

The harmonic mean is another specialised measure of location. It is used when the data consists of a set of rates such as prices ($/kg), speeds (km/hr) or production (output/man-hour).

The harmonic mean of n items is the number of items divided by the sum of the reciprocal of each individual item.

The general formula for calculating the harmonic mean is given as:

$$\text{Harmonic mean} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + ... + \frac{1}{x_n}}$$

with the usual notation.

Example

An organisation owns three lorries. Over a distance of 100 miles, one does 14 miles per gallon, one 18 miles per gallon and one 20 miles per gallon.

$$\text{Harmonic mean} = \frac{3}{\frac{1}{14} + \frac{1}{18} + \frac{1}{20}} = 16.95$$

Average consumption = 16.95 miles per gallon.

Merit

It takes little account of extreme values.

3.1.4 **Weighted mean**

A weighted mean is used whenever a simple average fails to give an accurate reflection of the relative importance of the items being averaged.

If a weight of w_i is assigned to an item x_i, then the formula for the weighted mean is given by

$$\bar{x}_{weighted} = \frac{\sum w_i x_i}{\sum w_i}$$

Example

In a certain institution, the year marks for modules are based upon a first-term test, a second-term test and a final exam at the end of the year. Given the number of topics to be covered for each assessment, they have a relative importance in the ratio 2:3:5. If a student obtained 74 marks in the first test, 63 in the second test and 55 in the final exams, what is his year mark?

The year mark is calculated as

$$\frac{(2\times 74)+(3\times 63)+(5\times 55)}{(2+3+5)} = 61.2$$

3.1.5 **Median**

The *median* is the middle observation of a distribution and is usually denoted by Q_2, given that it is also the second quartile. It is important to know that the median can only be determined after arranging numerical data in ascending (or descending) order. If n is the total number of observations, then the rank of the median is given by $\frac{1}{2}(n+1)$. For ungrouped data, if n is odd, the median is simply the middle observation but, if n is even, then the median is the mean of the two middle observations.

In the case of grouped data, the determination of the value of the median is slightly more complicated since the identity of individual observations is unknown. We proceed as follows:

1. Calculate the rank of the median.
2. Locate the cell in which the median is found (with the help of cumulative frequencies).
3. Determine the value of the median by linear interpolation (simple proportion).

The formula for calculating the median is given by

$$\text{Median} = LCB + \left(\frac{\frac{n+1}{2} - CF}{f} \right) c$$

where *LCB* is the lower real limit of the median class
f is the frequency of the median class
c is the class interval of the median class
CF is the cumulative frequency of the class preceding the median class

Note The 'median class' is the class containing the median.

Merits

1. It is rigidly defined.
2. It is easily understood and, in some cases, it can even be located by inspection.
3. It is not at all affected by extreme values.

Limitations

1. If *n* is even, the median is purely theoretical.
2. It is a *rank-based* statistic so that its calculation does not involve all the observations.
3. It is not suited to further statistical analysis.

Special note on percentiles

A *percentile* is a number or score-indicating rank which reveals the percentage of those being measured fall below that particular score. The *k*-th percentile is denoted by P_k and its rank is given by $\frac{k}{100}(n+1)$. For example, the median, Q_2 or P_{50}, is the 50^{th} percentile. The most widely used percentiles are the *quartiles*. Quartiles divide a distribution in four equal parts in terms of observations. The first or *lower quartile* is the value below which 25% of the distribution lies while the upper 25% of the distribution lies above the third or *upper quartile*. The median is also known as the second or *middle quartile*.

Quartiles are calculated in the same way as the median, that is using the same *formula* except, obviously, for the rank. (*Formula to be explained in detail.*)

3.1.6 **Mode**

The mode is the observation which occurs the most or with the highest frequency. Sometimes, it is denoted by $\hat{x}$. For ungrouped data, it may easily be detected by inspection. If there is more than one observation with the same highest frequency, then we either say that there is *no mode* or that the distribution is *multimodal*.

For grouped data, we can only *estimate* the mode – the class with the highest frequency is known as the *modal class*. Since we would prefer a single value for the mode (instead of an entire class), a rough approximation is the mid-class value of the modal class. However, there are two ways of estimating the mode quite accurately. Both should theoretically lead to the same result, the first one being *numerical* and the second, *graphical*.

The formula for a *numerical estimation* of the mode is given by

$$\text{Mode} = LCB + \left(\frac{f_1}{f_1 + f_2} \right) c$$

where f_1 is the difference between the frequencies of the modal class and that of the class preceding it and f_2 is the difference between the frequencies of the modal class and that of the class following it.

The mode may also be estimated by means of a frequency distribution *histogram*. We simply draw a histogram with the modal class and its two neighbouring classes, that is, found immediately before and after it.

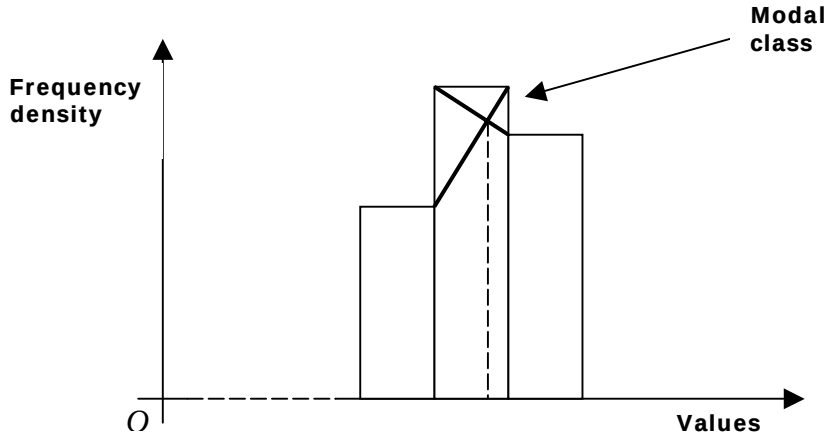

Fig. 3.1.6 Estimating the mode on a histogram

Merits

1. It is easy to understand and can sometimes be located by inspection.
2. It is not influenced by extreme values.
3. It may even be used for non-numerical data.

Limitations

1. Its calculation does not involve all the observations.
2. It is not clearly defined when there are several modes in a distribution.
3. It is not suited to further statistical analysis.

3.2 DISPERSION

A measure of dispersion shows the amount of variation or spread in the scores (values of observations) of a variable. When the dispersion is large, the values are widely scattered whereas, when it is small, they are tightly clustered. The two most well-known measures of dispersion are the *variance* and *standard deviation*.

3.2.1 Range

The range is simply the difference between the values of the maximum and minimum observations. It can only measure the extent to which the distribution spreads over the *x*-axis.

Merit

It is easy to calculate and understand.

Limitations

1. It is directly affected by extreme values.
2. It gives no indication of spread between the extremes.
3. It is not suited to further statistical analysis.

3.2.2 **Variance**

The variance is the most accurate way of determining the spread of a distribution as it qualifies for almost all the properties laid down for an ideal measure of dispersion. Sample and population variances are denoted by s^2 and σ^2 respectively. All statistical formulae, for ungrouped or grouped data, are given in terms of variance:

Ungrouped data

$$s^2 = \frac{\sum (x - \bar{x})^2}{n} \qquad \sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

Grouped data

$$s^2 = \frac{\sum f(x - \bar{x})^2}{\sum f} \qquad \sigma^2 = \frac{\sum f(X - \mu)^2}{N}$$

with the usual notations.

Note The formula for variance can be simplified using the laws of summation so that calculations may become shorter and less complicated.

$$s^2 = \frac{\sum x^2}{n} - \bar{x}^2 \qquad s^2 = \frac{\sum fx^2}{\sum f} - \bar{x}^2$$

3.2.3 **Standard deviation**

Standard deviation is defined as the *positive square root* of variance. It is as important as variance but is more commonly used due to its *linear* nature. The more widely the scores are spread out, the larger the standard deviation. We also use the term *standard error* in the case of an *estimate*.

The concept of standard deviation is so important that it can be treated as the foundation stone for *inferential statistics*, that is, estimation and hypothesis testing.

3.2.4 **Mean deviation**

The mean deviation is a measure of the average amount by which the values in a distribution differ from the arithmetic mean. Its formula is given by

$$\text{Mean deviation} = \frac{\sum f|x - \bar{x}|}{n}$$

Note Obviously, the frequency f falls off when there are no classes in the distribution, that is, only individual values.

Merits

1. It uses all values in the distribution to measure dispersion.
2. It is not greatly affected by extreme values.

Limitations

1. The distance from the mean does not reveal whether the observation is less than or greater than the mean.
2. It is not suitable for further statistical analysis.

3.2.5 **Quartile deviation and the inter-quartile range**

A measure of spread in a frequency distribution is the quartile deviation. This is equal to half the difference between the lower and upper quartiles and is sometimes called the *semi inter-quartile range*. Its formula is given by

$$\text{Quartile deviation} = \frac{Q_3 - Q_1}{2}$$

The quartile deviation shows the average distance between a quartile and the median. The smaller the quartile deviation, the less dispersed is the distribution. Just like the range, the quartile deviation can be misleading. If the majority of the data is towards the lower end of the range, then the third quartile will be considerably further above the median than the first quartile is below it. In such a case, when the two distances from the median are averaged, the difference is disguised. Then, it would be better to quote the actual values of the two quartiles rather than the quartile deviation.

It is customary to compare the efficiency of the median-quartile deviation pair with the mean-standard deviation in describing a distribution. Most of the time, the mean and the standard deviation are better since both their calculations involve all the observations. However, the median and the quartile deviation are hardly influenced by extreme values given that they are more *rank-based*.

3.2.6 **Coefficient of variation**

The coefficient of variation (*CV*) is mainly used to compare two distributions and is thus considered to be a *relative measure of dispersion*. When two distributions have the same mean but different standard deviations, it is easy to conclude which one is more dispersed – that would be the one with the higher standard deviation. However, if the means are not equal, it is somewhat difficult to compare the dispersions just by looking at the standard deviations.

The formula for the coefficient of variation is given by

$$\textit{Coefficient of variation} = \frac{s}{\bar{x}} \times 100$$

Example

Consider the two variables *A* and *B* in the following distributions.

	A	B
Mean	120	125
Standard deviation	50	51
Coefficient of variation	41.7	40.8

Table 3.2.6

At first glance, we would conclude that *B* has a greater variation (dispersion) since it has a higher standard deviation (51). We should also look at the values of the means – they are not equal. Thus, the only way to determine the degree of dispersion is by calculating the coefficient of variation for each distribution.

A has a *CV* of 41.7% while *B* has a *CV* of 40.8%. This shows the usefulness of the coefficient of variation. It is especially used in the comparison of rates of return in financial investments.

3.2.7 **Quartile coefficient of dispersion**

The quartile coefficient of dispersion measures the dispersion using quartiles. It differs from the quartile deviation because it is expressed as a proportion and not in units of the value of the variable. The lower the proportion, the less the dispersion.

Its formula is given by

$$\textit{Quartile coefficient of dispersion} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

3.2.8 **Coefficient of mean deviation**

The coefficient of mean deviation is simply the mean deviation expressed as a proportion of the arithmetic mean. This may be useful measure because it shows the *relative size* of the mean deviation.

Its formula is given as

$$\textit{Coefficient of mean deviation} = \frac{\sum f|x - \bar{x}|}{n\bar{x}}$$

Again, the frequency *f* falls off if there are no classes in the distribution.

3.3 **SKEWNESS**

Skewness is a measure of *symmetry* – it determines whether there is a concentration of observations somewhere in particular in a distribution. If most observations lie at the *lower end* of the distribution, the distribution is said to be *positively skewed* (or skewed to the right). If the concentration of observations is towards the *upper end* of the distribution, then it is said to display *negative skewness* (skewed to the left). A *symmetrical* distribution is said to have zero skewness.

Fig. 2.3.3 shows the various possible shapes of frequency distributions. The vertical bars on each diagram indicate the respective positions of the mean (bold), median (dashed) and mode (normal). In the case of a symmetrical distribution, the mean, median and mode are all equal in values (for example, the normal distribution).

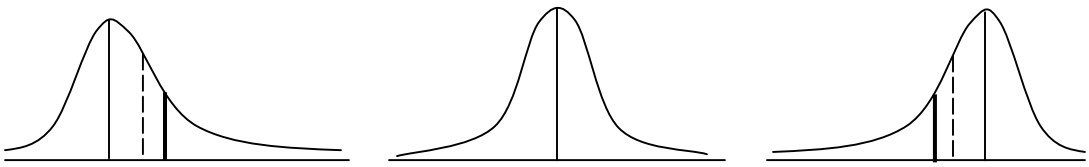

Fig. 3.3 Skewness

3.3.1 **Pearson's coefficient of skewness**

This is the most accurate measure of dispersion since its formula contains two of the most reliable statistics, the mean and standard deviation. The formula is given as

$$\alpha = \frac{3(\bar{x} - Q_2)}{s}$$

Note The validity of the formula can be verified by looking at the positions of the mean and median in **Fig. 3.3**.

3.3.2 **Quartile coefficient of skewness**

A less accurate but relatively quicker way of estimating skewness is by the use of quartiles of a distribution. The formula is given by

$$\alpha = \frac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1}$$

3.4 **KURTOSIS**

Kurtosis has a specific mathematical definition but, in the general sense, it indicates the degree of '*peakedness*' of a *unimodal* frequency distribution. It may be also considered as a measure of the relative concentration of observations in the centre, upper and lower ends and the 'shoulders' of a distribution. Kurtosis usually indicates to which extent a curve (distribution) departs from the bell-shaped or *normal* curve.

Kurtosis can be expressed numerically or graphically. The normal distribution has a kurtosis of 3 and is used as a reference in the calculation of the *coefficient of kurtosis* of any given distribution. If we observe the normal curve, we will see that its tails are neither too thick nor too thin and that there are neither too many nor too few observations concentrated in the centre. It is thus said to be *mesokurtic*.

If we start with the normal distribution and move scores from both centre and tails towards the shoulders, the curve becomes *flatter* and is said to be *platykurtic*. If, on the other hand, we move scores from the shoulders to the centre and tails, the curve becomes more *peaked* with thicker tails. In that case, it is said to be *leptokurtic*. **Fig. 3.4** shows the degree of peakedness for three types of distributions.

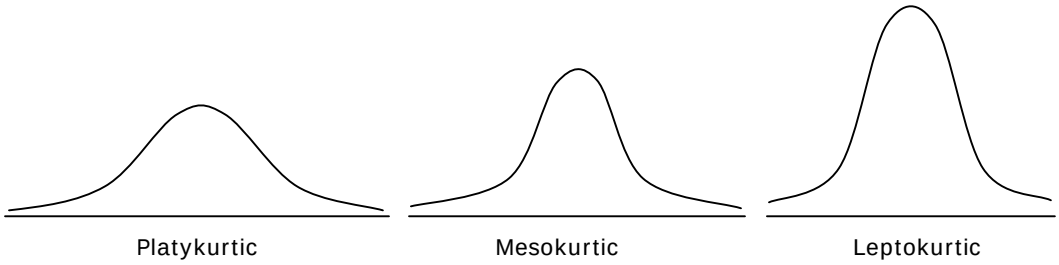

Fig. 3.4 Kurtosis

3.4.1 **Coefficient of kurtosis**

The formula for calculating kurtosis is given by

$$\beta = \frac{\sum (x - \bar{x})^4}{ns^4} \text{ or } \beta = \frac{\sum f(x - \bar{x})^4}{ns^4}$$

It is customary to subtract 3 from β for the sake of reference to the normal distribution. A negative value would indicate a *platykurtic* curve whereas a positive coefficient of kurtosis indicates a *leptokurtic* distribution.

4 EXAMPLES

We shall now illustrate the application of all the theory learnt in the previous sections by means of the following three examples (ungrouped and grouped data). The complete procedures for the calculations of descriptive statistics will be shown but it is generally advisable to use a **pocket calculator** to save computation time.

All three cases will be studied:

1. Ungrouped raw data
2. Ungrouped data in a discrete frequency distribution
3. Grouped (continuous) data

The full descriptive statistics have been calculated and given in **Tables 4.4**, **4.5** and **4.6**.

4.1 Example 1 (ungrouped raw data)

Data already arranged in ascending order:

2	7	8	11	15
16	18	19	19	19
23	23	24	26	27
29	33	40	44	47
49	51	54	63	68

Table 4.1

4.2 Example 2 (ungrouped data – discrete frequency table)

Age (*x*)	**Frequency (*f*)**	***cf***	***fx***	fx^2
19	14	14	266	5054
20	23	37	460	9200
21	134	171	2814	59094
22	149	320	3278	72116
23	71	391	1633	37559
24	8	399	192	4608
Total	**399**		**8643**	**187631**

Table 4.2

4.3 Example 3 (grouped data)

Age group	**Real limits**	**MCV (*x*)**	***f***	***cf***	***fx***	fx^2
21 – 25	20.5 – 25.5	23	5	5	115	2645
26 – 30	25.5 – 30.5	28	12	17	336	9408
31 – 35	30.5 – 35.5	33	23	40	759	25047
36 – 40	35.5 – 40.5	38	39	79	1482	56316
41 – 45	40.5 – 45.5	43	32	111	1376	59168
46 – 50	45.5 – 50.5	48	21	132	1008	48384
51 – 55	50.5 – 55.5	53	9	141	477	25281
56 – 60	55.5 – 60.5	58	2	143	116	6728
Total			**143**		**5669**	**232977**

Table 4.3

Table 4.4 Descriptive statistics for Example 4.1

Mean	$\bar{x} = \frac{\sum fx}{\sum f} = \frac{735}{25}$ = **29.4**
Median	Rank of median = $\frac{1}{2}(25+1) = 13$ Median = **24**
Mode	The observation with the highest frequency (3) is **19**
Lower Quartile	Rank of first quartile = $\frac{1}{4}(25+1) = 6.5$ $Q_1 = \frac{(15+16)}{2}$ = **15.5**
Upper Quartile	Rank of third quartile = $\frac{3}{4}(25+1) = 19.5$ $Q_3 = \frac{(44+47)}{2}$ = **67.5**
Maximum	Maximum observation = **68**
Minimum	Minimum observation = **2**
Range	Range = 68 – 2 = **66**
Quartile deviation	$QD = 0.5 \times (67.5 - 15.5)$ = **26**
Mean deviation	$MD = \frac{\sum f\lvert x - \bar{x}\rvert}{n} = \frac{368.8}{25}$ = **14.752**
Standard deviation	$s^2 = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2} = \sqrt{\frac{29351}{25} - (29.4)^2}$ = **17.598**
Quartile coefficient of dispersion	*Quart. coeff. of dis.* = $\frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{67.5 - 15.5}{67.5 + 15.5}$ = **0.313**
Coefficient of mean deviation	*Coeff. of MD* = $\frac{\sum f\lvert x - \bar{x}\rvert}{n\bar{x}} = \frac{14.752}{29.4}$ = **0.50**
Pearson's coefficient of skewness	$\alpha = \frac{3(\bar{x} - Q_2)}{s} = \frac{(3)(29.4 - 24)}{17.598}$ = **0.92**
Coefficient of kurtosis	$\beta = \frac{\sum (x - \bar{x})^4}{ns^4} = \frac{4226007.248}{(25)(17.598)^4}$ = **1.763**

Table 4.5 Descriptive statistics for Example 4.2

Mean	$\bar{x} = \frac{\sum fx}{\sum f} = \frac{8643}{399} = \mathbf{21.66}$
Median	Rank of median = $\frac{1}{2}(399+1) = 200$ *Median* = **22**
Mode	The observation with the highest frequency (149) is **22**
Lower Quartile	Rank of first quartile = $\frac{1}{4}(399+1) = 100$ $Q_1 = \mathbf{21}$
Upper Quartile	Rank of third quartile = $\frac{3}{4}(399+1) = 300$ $Q_3 = \mathbf{22}$
Maximum	*Maximum* observation = **24**
Minimum	*Minimum* observation = **19**
Range	*Range* = 24 – 19 = **5**
Quartile deviation	$QD = 0.5 \times (22-21) = \mathbf{0.5}$
Mean deviation	$MD = \frac{\sum f\lvert x-\bar{x}\rvert}{n} = \frac{328.38}{399} = \mathbf{0.823}$
Standard deviation	$s^2 = \sqrt{\frac{\sum fx^2}{\sum f} - \bar{x}^2} = \sqrt{\frac{187631}{399} - (21.66)^2} = \mathbf{1.013}$
Quartile coefficient of dispersion	*Quart. coeff. of dis.* = $\frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{22-21}{22+21} = \mathbf{0.023}$
Coefficient of mean deviation	*Coeff. of MD* = $\frac{\sum f\lvert x-\bar{x}\rvert}{n\bar{x}} = \frac{0.823}{21.66} = \mathbf{0.038}$
Pearson's coefficient of skewness	$\alpha = \frac{3(\bar{x} - Q_2)}{s} = \frac{(3)(21.66-22)}{1.013} = \mathbf{-1.007}$
Coefficient of kurtosis	$\beta = \frac{\sum f(x-\bar{x})^4}{ns^4} = \frac{468.7743}{(399)(1.103)^4} = \mathbf{1.116}$

Table 4.6 Descriptive statistics for Example 4.3

Mean	$\bar{x} = \frac{\sum fx}{\sum f} = \frac{5669}{143}$ = **39.64**.
Median	Rank of median = $\frac{1}{2}(143+1) = 72$ Median = $35.5 + \left(\frac{72-40}{39}\right)(5)$ = **39.60**
Mode	Modal class: 36 – 40 Mode = $35.5 + \left(\frac{16}{16+7}\right)(5)$ = **38.98**
Lower Quartile	Rank of first quartile = $\frac{1}{4}(143+1) = 36$ $Q_1 = 30.5 + \left(\frac{36-17}{23}\right)(5)$ = **34.63**
Upper Quartile	Rank of median = $\frac{3}{4}(143+1) = 108$ $Q_3 = 40.5 + \left(\frac{108-79}{32}\right)(5)$ = **45.03**
Maximum	Maximum observation = **60**
Minimum	Minimum observation = **21**
Range	Range = 60 – 21 = **39**
Quartile deviation	$QD = 0.5 \times (45.03 - 34.63)$ = **5.2**
Mean deviation	$MD = \frac{\sum f\lvert x-\bar{x}\rvert}{n} = \frac{879.6}{143}$ = **6.151**
Standard deviation	$s^2 = \sqrt{\frac{\sum fx^2}{\sum f} - \bar{x}^2} = \sqrt{\frac{232977}{143} - (39.64)^2}$ = **7.590**
Quartile coefficient of dispersion	*Quart. coeff. of dis.* = $\frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{45.03 - 34.63}{45.03 + 34.63}$ = **0.13**
Coefficient of mean deviation	*Coeff. of MD* = $\frac{\sum f\lvert x-\bar{x}\rvert}{n\bar{x}} = \frac{6.151}{39.64}$ = **0.155**
Pearson's coefficient of skewness	$\alpha = \frac{3(\bar{x} - Q_2)}{s} = \frac{(3)(39.64 - 39.60)}{7.590}$ = **0.016**
Coefficient of kurtosis	$\beta = \frac{\sum f(x-\bar{x})^4}{ns^4} = \frac{468.7743}{(399)(1.103)^4}$ = **1.116**