Summary

For the comprehension of a series of figures tabulation is essential; a diagram (*in addition to tables but usually not in place of them*) is often of considerable aid both for publication and still more, for a preliminary study of the features of the data. In publication, however, it should always be remembered that quite simple data are perfectly clear in a table and therefore a graph is merely a waste of space. Alternatively with very complicated data a diagram may be equally of no assistance and a waste of space. In medical literature the amount of space wasted by useless or unreadable diagrams is quite astonishing.

Both tables and graphs must be entirely self-explanatory without reference to the text. As far as possible the original observations should always be reproduced (in tabulated form showing the actual numbers belonging to each group) and not given only in the form of percentages – i.e. the percentages of the total falling in each group. The exclusion of observations from the tabulated series on any grounds whatever must be stated, the criterion upon which exclusion was determined clearly set out, and usually the number of such exclusions stated. Conclusions should be drawn from graphs only with extreme caution and only after careful consideration of the scales adopted.

6 The Average

When a series of observations has been tabulated, i.e. put in the form of a frequency distribution, the next step is the calculation of certain values which may be used as descriptive of the characteristics of that distribution. These values will enable comparisons to be made between one series of observations and another. The two principal characteristics of the distribution which it is invariably necessary to place on a quantitative basis are (*a*) its average value and (*b*) the degree of scatter of the observations round that average value. The former, the average value, is a measure of the *position* of the distribution, the point around which the individual values are dispersed. For example, the average incubation period of one infectious illness may be 7 days and of another 11 days. Though individual values may overlap, the two distributions have different central positions and therefore differ in this characteristic of location. In practice it is constantly necessary to discuss and compare such measures. A simple instance would be the observation that persons following one occupation lose *on the average* 5 days a year per person from illness; in another occupation they lose 10 days. The two distributions differ in their position and we are led to seek the reasons for such a difference and to see whether it is remediable.

In the above discussion the term 'average' has been introduced according to common everyday usage, namely that the average is the mean, or, more accurately, arithmetic mean, of all the observations (the sum of all the observations divided by the number of observations). The three terms, *average, mean, arithmetic mean,* are customarily taken as interchangeable. There are, however, in common use, and sometimes of special value, two other 'averages' or measures of position, namely the *median* and the *mode.* The median of a series of observations is the value, or magnitude, of the central or middle observation when all the observations are listed in order from lowest to highest. In other words, half the observations lie below the median and half the observations lie above it, and the median therefore divides the distribution into two halves. It defines the position of the distribution in that way.

The mode, as its name indicates, is the most frequently occurring, or most fashionable, value observed in the series.

Both these measures of position have, on occasions, a special value. For instance, in discussing the length of an illness we may be interested not so much in the mean, or average, length but in the *usual* length. The average length may be high owing (in short series of observations) to a few unduly long cases which pull up its value. We are more concerned here with the most frequent length, or the mode. If, however, we have insufficient data to determine a most frequent length, the median may give a better indication of the central position than the mean. The mean may be affected by large outlying observations and the median is unaffected. For example, if most of the illnesses last between 10 and 15 days the median will not be appreciably altered by the addition of 2 patients ill for 3 and 6 months; they merely represent two more cases lying above the middle point, and how much above is immaterial. On the other hand, the mean might be increased by their addition to 20 days and be, therefore, a poor measure of the general location of the distribution.

In general it is a sound rule in practice invariably to calculate the arithmetic mean and to use in conjunction with it, when it seems necessary, the median and the mode.

Calculation of the Averages

With a short series of observations the calculation of the mean is quite simply made. Let us take the following figures relating to the recorded age of onset of disease (age last birthday) for a group of 27 patients suffering from some specific illness.

Age of Patient at Onset of Disease (in Years)

39	36
50	57
26	41
45	61
47	47
71	44
51	48
33	59
40	42
40	54
51	47
66	53
63	54
55	

The arithmetic mean is the sum of all the values divided by their number. For the above example this is 39 + 50 + 26 + 45 + etc., giving a total of 1320. The average age of the 27 persons at onset of disease is, therefore, 1320 ÷ 27, or 48·9 years (as the age taken at onset was the age of the person *last birthday*, strictly speaking this average should be increased by half a year; for *on the average* persons aged, say, 40 last birthday will at the time of onset be $40\frac{1}{2}$ – they may be anywhere between 40 and 41 – and this will be the case with each person in the sample, again on the average).

Writing, now, the values in order of magnitude from lowest to highest we have the following list: 26, 33, 36, 39, 40, 40, 41, 42, 44, 45, 47, 47, 47, (48), 50, 51, 51, 53, 54, 54, 55, 57, 59, 61, 63, 66, 71. The median being the central value will be the 14th observation, there being 13 lower values than this and 13 higher values. It is, therefore, 48 years, and half the patients had lower ages of onset than this and half had higher.

The most commonly occurring observation is 47 years (three instances), but, beyond a statement of fact regarding this particular distribution of values, it is clear that with so short and widely scattered a series such a value is not likely to be a reliable estimate of the mode, or age most likely to occur in general with this disease.

A simple method of determining which value is required for the median is to divide by 2 the number of observations plus 1, or $(n + 1) \div 2$; in the present instance $(27 + 1) \div 2 = 14$, and the 14th value is the median. If there were 171 values the median is $(171 + 1) \div 2 =$ the 86th value, there being then 85 lower values, the median, 85 higher values = 171 in all.

In these instances when the total number of observations is an odd number there is no difficulty in finding the median as defined – the central value with an equal number of observations smaller and greater than itself. Often, however, the definition cannot be strictly fulfilled, namely when the total number of observations is an even number. If in the above series an additional patient had been observed with an age of onset of 73 there would have been 28 observations in all. There could be no central value. In such a situation it is usual to take the mean, or average, of the *two central values* as the median. Thus we should have: 26, 33, 36, 39, 40, 40, 41, 42, 44, 45, 47, 47, 47, (48, 50), 51, 51, 53, 54, 54, 55, 57, 59, 61, 63, 66, 71, 73. The two central values are 48 and 50, with 13 values lying on either side of them, and the median is taken as $(48 + 50) \div 2 = 49$ years.

The method of finding which are the required observations will, again, be to divide by 2 the number of observations plus 1; and the me-

dian will be the average of the values immediately above and below. Thus with 28 observations we have $(28 + 1) \div 2 = 14{\cdot}5$ and the 14th and 15th values are required; with 172 observations we have $(172 + 1) \div 2 = 86{\cdot}5$ and the 86th and 87th values are required, there then being 85 below, 2 for the median calculation, and 85 above = 172.

Difficulties with the Median

This customary extension of the definition of the median presents no difficulty and is a reasonable procedure. Often, however, in a short series of observations the definition cannot be completely fulfilled. Thus we might have as our observations of ages of onset the following values: 26, 33, 36, 39, 40, 40, 41, 42, 44, 45, 47, 48, 48, (48), 48, 51, 51, 53, 54, 54, 55, 57, 59, 61, 63, 66, 71. With 27 observations the middle one, the 14th, can of course be found. But it will be seen that there are *not* 13 smaller observations than this value and *not* 13 larger, for three of them are equal to the 14th value. In strict terms of the definition the median cannot be found in this short series. Even with a large number of observations it may be impossible to find a median value if the characteristic under discussion changes discontinuously (Table 3). Take, for example, the following frequency distributions:—

TABLE 3

TWO FREQUENCY DISTRIBUTIONS

A		B	
Height of a Group of Children, in Centimetres	Number of Children	Number of Children in a Family	Number of Families observed
127–	96	0	96
129–	120	1	120
131–	145	2	145
133–	83	3	83
135–	71	4	71
137–	32	5	32
139–141	18	6 or more	18
Total	565	Total	565

In the left-hand distribution, A, we have measurements of the heights of a group of 565 children and in the right-hand distribution, B, the number of children observed in 565 families.

The median height will be that of the 283rd child, $(565 + 1) \div 2$, when the observations are listed in order. To a considerable extent they are already in order, in the frequency distribution. Adding up the observations from the lower end of the distribution $96 + 120 = 216$, and we therefore need 67 more beyond this point to reach the 283rd. The median value, accordingly, lies in the group 131–133 cm and merely putting the 145 observations in that group in exact height order the required 67th can be found. (It can be estimated more simply but accurately by a method described later, p. 69). Variation in stature is continuous and thus a median value can reasonably be calculated.

On the other hand, the number of children in a family cannot vary continuously but must proceed by unit steps. The 283rd family must again be the 'middle' one and it must have 2 children, but there can be no central value for family size which divides the distribution into two halves, half the families having fewer children and half having more children. There is no real median value. Sometimes, however, it may be reasonable to extend the definition and to accept for the median the value which divides the distribution in such a way that half the observations are *less than or equal to* that value and half are *greater than or equal to it.*

Calculations from Grouped Figures: the Mean

With a large number of observations it would be very laborious to calculate the mean by summing all the individual values. With no serious loss of accuracy this can be avoided by working from the frequency distribution. So long as the classes into which the observations have been grouped are not too wide we can presume that the observations in that group are located at its centre. Taking the distribution A above, we presume that the 96 children whose height lies between 127 and 129 cm were all 128 cm, that the 120 children whose height lies between 129 and 131 cm were all 130 cm. With small class-intervals and a distribution that is not grossly asymmetrical that will approximate quite closely to the truth. Thus to reach the mean stature we may proceed as in Table 4. The sum of all the 565 statures is computed to be 74 742 cm, and the mean stature is, therefore, this sum divided by 565, or 132·28 cm.

Unless we have calculating machines at our disposal, even this method of calculation is unduly laborious since large multiplications have to be performed. It can, fortunately, be simplified.

Suppose 7 children were measured and their statures were found to be, in centimetres, 116, 119, 124, 127, 132, 136, 145. The mean stature, sum-

TABLE 4

CALCULATION OF THE MEAN, USING TRUE UNITS OF MEASUREMENT

Height of Children in cm	Number of Children	Mid-point of Group	Sum of Statures of the Children measured
127–	96	128	128 × 96 = 12 288
129–	120	130	130 × 120 = 15 600
131–	145	132	132 × 145 = 19 140
133–	83	134	134 × 83 = 11 122
135–	71	136	136 × 71 = 9 656
137–	32	138	138 × 32 = 4 416
139–141	18	140	140 × 18 = 2 520
Total	565		74 742

ming the values, is $899 \div 7 = 128{\cdot}4$ cm. But instead of summing these values we might proceed, if it saved time, by seeing how far each of these children differed from, say 125 cm. Originally, indeed, we might merely have measured these differences from a mark 125 cm high, instead of taking the statures from ground-level, so that we would finally have the average level of stature above or below 125 cm instead of above 0, or ground-level. These differences from 125 cm are -9, -6, -1, $+2$, $+7$, $+11$, $+20$, and their sum, taking sign into account, is $+24$; their mean is then $24 \div 7$, or $+3{\cdot}4$. This is the mean difference of the children from 125, and their mean stature from ground-level will, therefore, be $125 + 3{\cdot}4$, or $128{\cdot}4$ cm as before.

The same process can be applied to the frequency distribution. A central group can be taken as base line and its stature called 0 in place of its real value. The groups below it become -1, -2, -3, etc., and those above become $+1$, $+2$, $+3$, etc. Using these smaller multipliers, we can more simply find the mean stature in these units and then convert the answer into the original real units. The figures previously used then become as set out in Table 5.

The sum of the statures measured from 0 in working units is $+393 - 312 = 81$; their mean is $81 \div 565$, or $+0{\cdot}14$. The base line 0 was placed, it will be noted, against the group 131–133 cm, or, in other words, against the mid-point 132 (the value used in Table 4 working with the real values). We have found, then, that in ***working units*** the average difference of the children's stature from this central point of 132 is $+0{\cdot}14$.

TABLE 5

CALCULATION OF THE MEAN, USING ARBITRARY UNITS

Height of Children in cm. Real Units	Number of Children	Height in Working Units	Sum of Statures in Working Units
127–	96	− 2	96 × − 2 = − 192
129–	120	− 1	120 × − 1 = − 120
131–	145	0	145 × 0 = 0 − 312
133–	83	+ 1	83 × 1 = + 83
135–	71	+ 2	71 × 2 = + 142
137–	32	+ 3	32 × 3 = + 96
139–141	18	+ 4	18 × 4 = + 72
			+ 393
Total	565		+ 81

Their real mean stature from ground-level must therefore be 132 plus *twice* $0{\cdot}14 = 132{\cdot}28$ cm, the same value as was found previously from the more complicated sums using real values. The multiplier 2 is arrived at thus: the mean is found to be $0{\cdot}14$ of a unit above the 0 value when the groups differ in their distances from one another's centres by unity, e.g. from 0 to -1 to -2. But in the real distribution their distance from one another's centres is 2, e.g. from 130 to 132. Therefore the mean in real units must be 2 times $0{\cdot}14$ above 132 cm.

While the method has been demonstrated on a distribution of statures it is, of course, perfectly general. Returning to the previous example of age at onset of a disease, the calculation of the mean from a frequency distribution will be made as in Table 6.

The mean age of the patients in working units is $-4015/7292$, or $-0{\cdot}5506$, and we have to translate this value into the real units. It lies, it will be seen, $0{\cdot}5506$ of a working unit *below* the 0 which was placed against the group 40–45. The centre of this group is $42{\cdot}5$ and the real mean is, therefore, $42{\cdot}5 - 5(0{\cdot}5506) = 39{\cdot}75$ years.

The formula is, then—

Mean in real units = *Centre value* in real units of the group against which 0 has been placed, plus or minus (the mean in working units × width of class adopted in the frequency distribution in real units). The plus or minus depends, of course, upon the sign of the mean in working units.

TABLE 6

CALCULATION OF THE MEAN, USING ARBITRARY UNITS

Age of Patient at onset of Disease (in Years)	Age in Working Units	Number of Patients	Sum of Ages of Patients in Working Units	
15–	− 5	14	− 70	
20–	− 4	163	− 652	
25–	− 3	861	− 2583	
30–	− 2	1460	− 2920	
35–	− 1	1466	− 1466	
40–	0	1269		− 7691
45–	1	953	+ 953	
50–	2	754	+ 1508	
55–	3	221	+ 663	
60–	4	103	+ 412	
65–69	5	28	+ 140	
				+ 3676
Total		7292		− 4015

Points especially to be remembered, on which beginners often go wrong, are these:—

(*a*) The 0 value corresponds in real units to the *centre* of the group against which it is placed and *not* to the start of that group.

(*b*) In translating the mean in working units into the real mean the multiplier, or width of the groups in real units, must be brought into play.

(*c*) The groups in real units should be, throughout, of the same width; if a group contains no observations, nevertheless the appropriate number in the working units must be allotted to it. Otherwise the size in working units of observations farther away from the 0 is not correctly defined.

(*d*) Values smaller in real units than the group taken as 0 should always be taken as minus in working units and higher values always taken as plus. Otherwise confusion arises in passing finally from the working mean to the real mean.

(*e*) The exact value for the centre of the group against which the 0 has been placed must be carefully considered. For simplification it has been taken above as the mid-point of the apparent group, but sometimes that will not be strictly correct. For instance, in taking the age at onset of disease the age of the patient might be recorded either as age last birthday or as age to nearest birthday. With age

last birthday persons placed in the group 40–45 may be any age between 40 years and 44 years + 364 days, and the centre point will be 42·5. But with age to nearest birthday a person aged 39½ is called 40 and one aged 44½ is called 45. The group 40–45 therefore runs from 39·5 to 44·5 and its centre point is 42 years.

Somewhat similar reasoning will apply to the stature of children, depending here upon the accuracy with which the original measurements were made, or, in other words, upon what measurements were, in fact, allocated to the group entitled, say, 127–129 cm. For instance, if the measurements were taken to the nearest ½ of a centimetre the group 127–129 would run from 126·5 to 128·5.

The working can be checked, it may be added, by changing the position of the 0 on the working unit scale, and this is a better check than a repetition of the previous arithmetic.

Calculation from Grouped Figures: the Median

For calculation of the median from the frequency distribution we may return to the statures of children, repeated in Table 7.

TABLE 7

Height of Children in cm	Number of Children
127–	96
129–	120
131–	145
133–	83
135–	71
137–	32
139–141	18
Total	565

As pointed out on p. 65, the median stature will be the height of the 283rd child when the observations are listed in order. In practice, however, we do not trouble to list them in order when dealing with large numbers but make an *estimate* of the median from the grouped values in the frequency distribution. The value to be found is that of the point which divides the distribution into exactly two halves, i.e. with 282·5

observations below and 282·5 above the mid-point. In other words, the value needed lies at the mid-point 565/2 or 282·5. Adding the numbers from the start of the distribution we have 96 + 120 = 216 and 216 + 145 = 361. The mid-line at 282·5, therefore, falls in the group with 145 children whose height lies between 131 and 133 cm. By simple proportion it will lie at a point 66·5/145 of 2 cm beyond the 131 cm at which this group of children starts (since 282·5 − 216 = 66·5). The median may, therefore, be calculated as 131 + (66·5/145) of 2 cm = 131·92 cm.

Similarly with the 7292 patients with a given age at onset (Table 6) we require the value which will divide the distribution into its two halves or 7292/2 = 3646 observations. Adding again from the start we have 14 + 163 + 861 + 1460 = 2498 and 2498 + 1466 = 3964. The mid-line falls, therefore, in the group which starts at 35 years of age and continues up to 40 years. By simple proportion it will lie (1148/1466 of 5 years) beyond the age of 35 at which this group starts (for 3646 − 2498 = 1148). The median is, therefore, calculated to be 35 + 1148/1466 of the span of 5 years which the group covers, i.e. 35 (the opening point of the group) +3·92 years (the further part of the group absorbed in reaching the required point), which equals 38·92 years.

The main point to recall in calculation is that the median must be computed from the *opening* point of the group in which it is located and *not* from the mid-point of that group. Again the actual starting-point of the group may have to be closely considered.

Calculation from Grouped Figures: the Mode

The mode, as previously noted, is the observation most frequently occurring. Like the other averages, the mean and the median, it has a single value (unless the distribution is bi-modal). In other words, we may say that most children in the sample above had a height between 131 and 133 cm and most patients had an age at onset of disease of 35–40 years. But these group-intervals do not specify the mode, for where the largest number of observations will fall depends partly upon the group-intervals that we choose to adopt. For instance, the most frequent stature might lie between 131·5 and 134·5 cm and the most frequent age between 35 and 37 if we chose to put the observations into such groups instead of those previously adopted. What we need for the value of the mode is that value at which the curve of the distribution would, when plotted, reach its highest point if we had a vast number of observations and could make the groups in which they are placed indefinitely small. In fact, an observed

curve is invariably irregular, owing to paucity of numbers. To find the highest point of it accurately it is therefore necessary to fit to the actual observations some ideal curve which well describes their trend and then calculate the highest point of this curve. This is a somewhat difficult process, but, fortunately for frequency distributions that are not very asymmetrical, we can find the mode with sufficient accuracy from the formula

$$\text{Mode} = \text{Mean} - 3(\text{Mean} - \text{Median})$$

or the mode equals the mean less 3 times the difference between the mean and the median.

In the previous examples we found:—

	Stature of Children	Age at Onset of Disease
Mean	132·28	39·75
Median	131·92	38·92

The modes are, therfore, calculated to be

$$132{\cdot}28 - 3(132{\cdot}28 - 131{\cdot}92) = 132{\cdot}28 - 3(0{\cdot}36) = 131{\cdot}20 \text{ cm}$$

and

$$39{\cdot}75 - 3(39{\cdot}75 - 38{\cdot}92) = 39{\cdot}75 - 3(0{\cdot}83) = 37{\cdot}26 \text{ years.}$$

Certain distributions may have more than one maximum, in which case they are bi-modal or multi-modal.

The Weighted Average

Let us suppose the following fatality-rates are observed:—

Age in Years	Fatality-rate per cent	Number of Patients
Under 20	47·5	40
20–39	15·0	120
40–59	22·4	250
60 and over	51·1	90

It would be wrong to compute the general fatality-rate at all ages by taking the average of these four rates, i.e. $(47{\cdot}5 + 15{\cdot}0 + 22{\cdot}4 + 51{\cdot}1) \div 4 = 34{\cdot}0$ per cent. The rate at all ages will depend upon the number of patients who fall ill at each age, as shown in the third column. To reach the rate at all ages the separate age-rates must be 'weighted' by the number of observations in each group. Thus we have $(47{\cdot}5 \times 40) + (15{\cdot}0 \times 120) + (22{\cdot}4 \times 250) + (51{\cdot}1 \times 90)$ divided by 500, the total number of patients, which equals 27·8 per cent, substantially lower than the erroneous figure. By such weighting we are in effect calculating the total number of deaths that took place in the total 500 patients. These deaths divided by the number of patients gives the required rate, and the unweighted average of the rates will not produce it unless either the number of patients at each age is the same or the fatality-rate remains the same at each age. In general it is, therefore, incorrect to take an unweighted average of rates or of a series of means.

With such a series of means it is equally simple to reach the grand mean. Thus, suppose we have the mean weight of three sets of children, namely 37·2 kg for 120 children, 34·0 kg for 82 children, and 41·3 kg for 126 children. Remembering that the mean is the sum of the original weights divided by the number of children, the total weight of the 120 children with a mean of 37·2 kg must have been 4464 kg. Similarly the total weight of the other groups must have been $82 \times 34{\cdot}0$ and $126 \times 41{\cdot}3$, or 2788 kg and 5204 kg. The total weight of all 328 children was therefore 12 456 kg and the mean of all children 38 kg.

Calculation of the Mean of a Few Observations

A point to remember in taking the average of a few observations is that the whole numbers need not invariably be added. Thus we might have observations of the number of cases of an infectious disease in each week of the year in each of 3 years and require the average annual number in each week.

Year	*Week 1*	*Week 2*	*Week 3*	*Week 4*	etc.
1974	126	132	163	182	
1975	121	126	159	191	
1976	128	120	161	190	

In week 1 each total has 120 as a common feature and it is necessary to add only $6 + 1 + 8 = 15$, and dividing by 3 the mean is 125. In week 2, again using 120 as a base line, we have $12 + 6 + 0 = 18$ and the mean is

126. In week 3 we may take 160 as base line and we have $+3 - 1 + 1 = 3$ and the mean is 161. In week 4 we may take 180 as base line and have $2 + 11 + 10 = 23$ and the mean is 187·7. With experience these small differences from a base line can be accurately noted and mentally added and much labour thus saved.

Summary

The general position of a frequency distribution on some scale is measured by an average. There are three averages in common use: (*a*) the arithmetic mean, (*b*) the median, and (*c*) the mode. The arithmetic mean, usually termed the mean or average, is the sum of all the observations divided by their number. The median is the central value when all the values are listed in order from the lowest to the highest. The mode is the most frequent observation, or, strictly speaking, the value at which the ideal curve to which the observations conform reaches its highest point. It can be found approximately from the formula Mode = Mean − 3(Mean − Median) so long as the shape of the distribution is not very skew. In taking the mean of a series of sub-means the 'weights' attached to the latter must be taken into account.