CHAPTER I

FREQUENCY DISTRIBUTIONS

1. Arithmetic mean. Partition values

Consider a group of N persons in receipt of wages. Let x shillings be the wage of an individual on some specified day. Then, in general, x will be a variable whose value changes with the individual. Possibly the N values of x will not all be different. Suppose there are only n different values $x_1, x_2, \ldots, x_n$ which occur respectively $f_1, f_2, \ldots, f_n$ times. The numbers f_i, in which the subscript i takes the positive integral values 1, 2, ..., n, are called the *frequencies* of the values x_i of the variable x; and the assemblage of values x_i, with their associated frequencies, is the *frequency distribution* of wages for that group of persons on the day specified. The sum of the frequencies is clearly equal to the number of persons in the group, so that

$$N = f_1 + f_2 + \ldots + f_n = \sum_{i=1}^{n} f_i, \tag{1}$$

where, as the equation indicates, $\sum_{i=1}^{n} f_i$ denotes the sum of the n frequencies f_i, i taking the integral values 1 to n. When the range of values of i is understood, the sum will be denoted simply by $\sum_i f_i$ or $\sum f_i$. The number N is the *total frequency*. The *mean* of the distribution is the arithmetic mean $\bar{x}$ of the N values of the variable, and is therefore given by

$$\bar{x} = \frac{1}{N}(f_1 x_1 + f_2 x_2 + \ldots + f_n x_n) = \frac{1}{N}\sum_{i=1}^{n} f_i x_i, \tag{2}$$

since the value x_i occurs f_i times. This formula expresses what is meant by saying that $\bar{x}$ is the *weighted mean* of the different values x_i, whose weights are their frequencies f_i.

Frequency distributions of many different variables will occur in the following pages. Thus the values of the variable x may be the heights, or the weights, or the ages of a group of persons, or the

yields of grain per acre from a number of plots of land. For each finite distribution f_i will denote the frequency of the value x_i. The total frequency N is then given by (1), and the mean $\bar{x}$ of the distribution by (2).

Example. The student to whom the above summation notation is new, may profitably verify the following relations. If a is a constant,

$$\sum_i af_ix_i = a\sum_i f_ix_i,$$

$$\sum_i f_i(x_i+a) = \sum_i f_ix_i + Na,$$

$$\sum_i f_i(x_i+a)^2 = \sum_i f_ix_i^2 + 2a\sum_i f_ix_i + Na^2.$$

If (x_i, y_i) is a pair of corresponding values of two variables, x and y, with frequency f_i,

$$\sum_i f_i(x_i+y_i) = \sum_i f_ix_i + \sum_i f_iy_i.$$

The symbol used as a subscript in connection with summation is immaterial; but i, j, r, s, t are perhaps most commonly employed.

Suppose that the frequency distribution of x consists of k partial or component distributions, $\bar{x}_j$ being the mean of the jth component and n_j its total frequency, so that

$$N = \sum_{j=1}^{k} n_j.$$

Then that part of the sum $\sum_i f_ix_i$ which belongs to the jth component has the value $n_j\bar{x}_j$, and the relation (2) is equivalent to

$$\bar{x} = \frac{1}{N}\sum_{j=1}^{k} n_j\bar{x}_j. \tag{3}$$

Consequently the mean of the whole distribution is the weighted mean of the means of its components, the weights being the total frequencies in those components.

Again, let u and v be two variables with frequency distributions in which a value of v corresponds to each value of u. Then the values of the variables occur in pairs. Let N be the number of pairs of values (u_i, v_i), and let

$$x_i = u_i + v_i.$$

The arithmetic mean of the N values of x is equal to that of the N values of the second member, which is $\bar{u}+\bar{v}$. Consequently

$$\bar{x} = \bar{u}+\bar{v}, \tag{4}$$

which expresses that the mean of the sum of two variables is equal to the sum of their means; and the result can be extended to the sum of any number of variables. The reader can prove similarly that, if a and b are constants, and

$$x = au+bv,$$

then

$$\bar{x} = a\bar{u}+b\bar{v}. \tag{4'}$$

Suppose the N values of the variable in the distribution to be arranged in ascending order of magnitude. Then the *median* is the middle value, if N is odd; while, if N is even, it is the arithmetic mean of the middle pair, or, more generally, it may be regarded as any value in the interval between these middle values. Similarly, the *quartiles*, Q_1, Q_2, Q_3, are those values in the range of the variable which divide the frequency into four equal parts, the second quartile being identical with the median; and the difference between the upper and lower quartiles, Q_3-Q_1, is the *interquartile range*. The *deciles* and *percentiles* are those values which divide the total frequency into ten and one hundred equal parts respectively. The median, quartiles, deciles and percentiles are often spoken of collectively as *partition values*, since each set of values divides the frequency into a number of equal parts. Sometimes they are referred to as *quantiles*.

That value of the variable whose frequency is a maximum is called a *mode*, or *modal value*, of the distribution. When, as usually happens, there is only one mode, the distribution is said to be *unimodal*.

We shall presently consider continuous frequency distributions. But it should be pointed out at once that the variable may be either continuous or discrete. A continuous variable is one which is capable of taking any value between certain limits; for example, the stature of an adult man. A discrete variable is one which can take only certain specified values, usually positive integers; for example, the

number of heads in a throw of ten coins, or the number of accidents sustained by a worker exposed to a given risk for a given time. Of course an *observed* frequency distribution can only contain a finite number of values of the variable, and in this sense all observed frequency distributions are discrete. Nevertheless, the distinction between continuous and discrete variables will be found to be of importance when we come to study populations and probability distributions. In the next few sections we shall assume that the values of the variables are discrete.

2. Change of origin and unit

The following graphical representation will be found helpful. Taking the usual x-axis with origin O, we may represent the variable x by the abscissa of the current point P. Then $\bar{x}$ is the abscissa of a

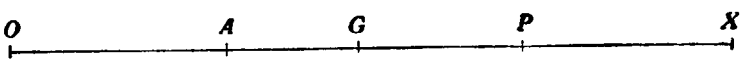

FIG. 1

fixed point G. It is frequently convenient to take a new origin at some point A, whose abscissa is a. Let ξ be the abscissa of P relative to A as origin. Then, since $OP = OA + AP$, we have

$$x = a + \xi.$$

Thus ξ is the excess of x above a, or the *deviation* of x from that value. Taking the mean of each member of this equation we have

$$\bar{x} = a + \bar{\xi}. \tag{5}$$

Thus $\bar{x} - a$, which is the deviation of the mean value of x from a, is equal to $\bar{\xi}$, which is the mean of the deviations of the values x_i from a. In particular, by taking a as $\bar{x}$, we have the result that *the sum of the deviations of the values x_i from their mean is zero*. This is also easily proved directly; for

$$\sum_i f_i(x_i - \bar{x}) = \sum_i f_i x_i - N\bar{x} = 0 \tag{6}$$

in virtue of (2).

In addition to choosing A as origin it may be convenient to use a different unit, say c times the original unit. Then, if u is the deviation of P from A measured in terms of this new unit,

$$u = (x-a)/c$$

or

$$x = a+cu. \tag{7}$$

Taking the mean value of the variable represented by either side we have, in virtue of (4′),

$$\bar{x} = a+c\bar{u}, \tag{8}$$

$\bar{u}$ being the mean value of u for the distribution.

Example. Eight coins were tossed together, and the number x of heads resulting was observed. The operation was performed 256 times; and the frequencies that were obtained for the different values of x are shown in the following table. Calculate the mean, the median and the quartiles of the distribution of x.

x	f	ξ	$f\xi$	$f\xi^2$	$f\xi^3$
0	1	−4	− 4	16	− 64
1	9	−3	−27	81	−243
2	26	−2	−52	104	−208
3	59	−1	−59	59	− 59
4	72	0	0	0	0
5	52	1	52	52	52
6	29	2	58	116	232
7	7	3	21	63	189
8	1	4	4	16	64
Totals	256	—	− 7	507	− 37

The different values of x are shown in the first column, and their frequencies in the second. The calculation is simplified by taking the value $x = 4$ as origin of ξ. The values of ξ corresponding to those of x are given in the third column, and those of the product $f\xi$ in the fourth. The remaining columns are not needed for the present. Totals for the various columns are given in the bottom row. Hence

$$\bar{\xi} = \frac{1}{N}\sum_i f_i\xi_i = -7/256 = -0{\cdot}027.$$

The mean value of x is therefore

$$\bar{x} = a+\bar{\xi} = 4-0{\cdot}027 = 3{\cdot}973.$$

The mode, being the value of x with the largest frequency, is clearly 4. To find the median we observe that the values of x are arranged in ascending order, and that the 128th and 129th are both 4. Hence the median is also 4. Similarly the 64th and 65th values are both 3, so that the lower quartile is 3. In the same way we find that the upper quartile is 5.

3. Variance. Standard deviation

The *mean square deviation* of the variable x from the value a is, as the name implies, the mean value of the square of the deviation of x from a. It is therefore given by $\frac{1}{N}\sum_i f_i \xi_i^2$. The positive square root of this quantity is the *root-mean-square deviation* from a. In the important case in which the deviation is taken from the mean of the distribution, the mean square deviation is called the *variance* of x, and is denoted by μ_2. The reason for the notation will appear in the next section. The positive square root of the variance is called the *standard deviation* (S.D.) of x, and is denoted by σ. Thus

$$\mu_2 = \sigma^2 = \frac{1}{N}\sum_i f_i (x_i - \bar{x})^2. \tag{9}$$

The variance (or the S.D.) may be taken as an indication of the extent to which the values of x are scattered. This scattering is called *dispersion*. When the values of x cluster closely round the mean, the dispersion is small. When those values, whose deviations from the mean are large, have also relatively large frequencies, the dispersion is large. The concepts of variance and S.D. will play a prominent part in the following pages.

When the mean square deviation from any value a is known, and also the deviation $\bar{\xi}$ of the mean from that value, the variance is easily calculated. For

$$\begin{aligned}\sigma^2 &= \frac{1}{N}\sum_i f_i (\xi_i - \bar{\xi})^2 = \frac{1}{N}\sum_i f_i \xi_i^2 - \frac{2}{N}\bar{\xi}\sum_i f_i \xi_i + \bar{\xi}^2 \\ &= \frac{1}{N}\sum_i f_i \xi_i^2 - \bar{\xi}^2. \end{aligned} \tag{10}$$

This formula is of great importance, and will be constantly employed. On multiplying by N we have an equivalent relation, which may be expressed

$$N\sigma^2 = \sum_i f_i \xi_i^2 - \frac{1}{N}\left(\sum_i f_i \xi_i\right)^2. \tag{11}$$

Thus $N\sigma^2$ is less than $\sum_i f_i \xi_i^2$, showing that *the sum of the squares of the deviations of the values x_i is least when the deviations are measured from the mean.*

Another possible measure of dispersion is the mean value of the absolute deviation from the mean of the distribution, commonly called the *mean deviation* from the mean. This quantity, however, does not lend itself readily to algebraical treatment, and is therefore not nearly so important as the variance and the S.D. The semi-interquartile range is also sometimes taken as an indication of the magnitude of the dispersion.

The significance of the magnitude of the standard deviation clearly depends upon the values of the variable. Thus a S.D. of 6 in. in the measurements of the height of a tower, is much less significant than an equal S.D. in the measurements of the height of a man. The ratio of the S.D. to the mean value of the variable is called the *coefficient of variation*. It is an absolute measure of dispersion in the sense that it is independent of the unit employed. And by means of this coefficient we are able to compare the variabilities of distributions of different characteristics, such as weight and height. Sometimes the coefficient of variation is defined as 100 times the above value, i.e. as the percentage of the mean which is equal to the S.D.

Example 1. In the example of the preceding section the mean square deviation of x from the value 4 is $507/256 = 1{\cdot}98$. Hence the variance is given by

$$\sigma^2 = 1{\cdot}98 - \bar{\xi}^2 = 1{\cdot}98 - (0{\cdot}027)^2 = 1{\cdot}98,$$

and

$$\sigma = 1{\cdot}407 = 1{\cdot}41 \text{ nearly.}$$

From the $f\xi$ column it is clear that the sum of the absolute deviations from $x = 4$ is 277. By measuring deviations from the mean, instead of from $x = 4$, we increase the absolute deviations of 161 values, and decrease those of 95 values, by 0·027. Hence the sum of the absolute deviations from the mean is $277 + 66(0{\cdot}027) = 278{\cdot}78$. The mean deviation is therefore 1·09 approximately.

Example 2. Find the mean and the variance for the distribution in which the values of x are the positive integers 1, 2, 3, ..., N, the frequency of each being unity.

Here

$$\bar{x} = \frac{N(N+1)}{2N} = \tfrac{1}{2}(N+1).$$

The mean square deviation from $x = 0$ is

$$(1^2 + 2^2 + \ldots + N^2)/N = \tfrac{1}{6}(N+1)(2N+1).$$

Hence

$$\sigma^2 = \tfrac{1}{6}(N+1)(2N+1) - \tfrac{1}{4}(N+1)^2$$

$$= \tfrac{1}{12}(N^2 - 1).$$

Example 3. For the distribution expressed by

$x =$	5	6	7	8	9	10	11	12	13	14	15
$f =$	18	25	34	47	68	90	80	62	38	27	11

the total frequency is 500. Show that the mean value of x is 10·054, the variance 5·58, the S.D. 2·36, the median 10, and the lower and upper quartiles 9 and 12 respectively. Also calculate the mean deviation from the mean as in Ex. 1.

Example 4. A distribution consists of several component distributions. Express the variance of the whole distribution in terms of those of the components and the deviations of the means of the components from the general mean.

Let n_j be the frequency in the jth component, σ_j its S.D., and $d_j = \bar{x}_j - \bar{x}$ the deviation of its mean from the general mean. Then the mean square deviation of this component from the general mean is $\sigma_j^2 + d_j^2$, and the sum of the squares of its deviations from the general mean is $n_j(\sigma_j^2 + d_j^2)$. Hence the variance σ^2 of the whole distribution is given by

$$N\sigma^2 = \Sigma n_j(\sigma_j^2 + d_j^2),$$

where N is the total frequency $\sum_j n_j$.

4. Moments

In the notation of the preceding sections the mean value of the rth power of the deviation of the variable from the value a is $\sum_i f_i \xi_i^r / N$. This is usually called the rth *moment* of the distribution about the value a, or the moment of order r. The term 'moment' is borrowed from Mechanics. Since f_i/N is the *relative frequency* of the value x_i in the distribution, and the deviation ξ from a is represented by the distance AP, the above expression can be regarded as the sum of the rth moments of the relative frequencies about A. The rth moment about the mean of the distribution is denoted by μ_r. The corresponding moment about a specified value other than the mean, will be denoted* by μ'_r. Thus

$$\mu_r = \frac{1}{N}\sum_i f_i(x_i - \bar{x})^r \tag{12}$$

is the rth moment about the mean, while

$$\mu'_r = \frac{1}{N}\sum_i f_i(x_i - a)^r = \frac{1}{N}\sum_i f_i \xi_i^r \tag{13}$$

* An alternative notation, with ν_r instead of μ'_r, has some advantages.

is the rth moment about the value a. Putting $r = 0$ we see that

$$\mu_0 = \mu'_0 = 1. \tag{14}$$

Similarly, in virtue of (2) and (6), we have

$$\mu'_1 = \bar{\xi}, \quad \mu_1 = 0. \tag{15}$$

The second moment about the mean is clearly the variance already discussed.

By means of the binomial expansion, moments about the mean of the distribution may be expressed in terms of moments about any other value, $x = a$. Thus

$$\begin{aligned}\mu_r &= \frac{1}{N}\sum_i f_i(\xi_i - \bar{\xi})^r \\ &= \mu'_r - \binom{r}{1}\bar{\xi}\mu'_{r-1} + \binom{r}{2}\bar{\xi}^2\mu'_{r-2} - \dots,\end{aligned} \tag{16}$$

$\binom{r}{s}$ denoting the binomial coefficient, often written ${}^rC_s$ or C^r_s. In particular, in virtue of (14) and (15),

$$\mu_2 = \mu'_2 - 2\bar{\xi}^2 + \bar{\xi}^2 = \mu'_2 - \bar{\xi}^2 \tag{17}$$

in agreement with (10). Similarly

$$\left.\begin{aligned}\mu_3 &= \mu'_3 - 3\bar{\xi}\mu'_2 + 2\bar{\xi}^3, \\ \mu_4 &= \mu'_4 - 4\bar{\xi}\mu'_3 + 6\bar{\xi}^2\mu'_2 - 3\bar{\xi}^4,\end{aligned}\right\} \tag{18}$$

and so on.

In calculating moments it is frequently convenient to change the unit. As in § 2, let u be the measure of the deviation from $x = a$, in terms of a unit c times the original unit, so that $\xi = cu$. Then the rth moment of x about a is

$$\mu'_r = \frac{1}{N}\sum_i f_i\xi_i^r = \frac{c^r}{N}\sum_i f_i u_i^r. \tag{19}$$

Thus the rth moment of the variable x is c^r times the corresponding moment of the variable u.

A distribution is said to be *symmetrical* when the frequencies are symmetrically distributed about the mean, that is to say, when

values equidistant from the mean have equal frequencies. For example, the distribution expressed by

$x =$	0	1	2	3	4	5	6	7	8
$f =$	1	8	28	56	70	56	28	8	1

is symmetrical about its mean $\bar{x} = 4$. In the case of a symmetrical distribution there is the simplification that all the moments of odd order about the mean are equal to zero, since the terms of the sum in (12) cancel in pairs. In the case of an unsymmetrical distribution, the degree of departure from symmetry is called its *skewness*. More than one measure of this property has been proposed. One of the simplest is μ_3/σ^3, while another is half this expression. These are clearly independent of the unit chosen for the variable, and they vanish if the distribution is symmetrical. Another measure of skewness, proposed by Karl Pearson, will be given later.

Example 1. For the distribution of x in the example of § 2, the third moment about $\xi = 0$ is

$$\mu_3' = -37/256 = -0{\cdot}145.$$

Hence the third moment about the mean is given by

$$\begin{aligned}\mu_3 &= \mu_3' - 3\bar{\xi}\mu_2' + 2\bar{\xi}^3 \\ &= -0{\cdot}145 - 3(-0{\cdot}027)(1{\cdot}98) + 2(-0{\cdot}027)^3 \\ &= 0{\cdot}018 \quad \text{nearly.}\end{aligned}$$

The skewness, calculated from the formula μ_3/σ^3, is

$$0{\cdot}018/2{\cdot}8 = 0{\cdot}0064,$$

which is very small.

Example 2. For the distribution in § 3, Ex. 3, show that $\mu_3 = -1{\cdot}92$, and deduce that $\mu_3/\sigma^3 = -0{\cdot}146$.

5. Grouped distribution

Frequently the number of different values of the variable represented in the distribution is so large that, for convenience in calculating the moments, it becomes necessary to approximate by grouping the values. In such cases the range of variation of x is usually divided into a number of equal intervals. The group of values falling in a given interval constitutes a *class*; and the number of such values is the *class frequency*. The magnitude of an interval is called the *class interval*. For simplicity of calculation the number