CHAPTER 1

FREQUENCY DISTRIBUTIONS

1·01. Introduction

In this chapter we deal with some general principles from the viewpoint of the theory of statistics in preparation, to some extent at least, for the more specialized study of the theory of errors and related subjects. Statistics is a branch of mathematics concerned, in its simplest stages, with the study of the detailed information relating to some specific characteristic such as the weight measures of national service recruits or the marks scored in an examination; in some of the subsequent chapters the emphasis is on the study of observations or measures in which the appearance of unavoidable errors is fully recognized.

As a typical example, illustrating the initial methods of statistics, suppose that the school authorities of a city provide the individual measures of the heights of a thousand boys belonging to a particular age group. This mass of information is not readily comprehended as it stands, and the first task of the statistician is to have the information arranged in a suitably manageable or condensed form. This is done most conveniently by deriving the numbers of boys with heights, for example, between 40 and 41 in., between 41 and 42 in., and so on; there is now a much clearer picture of the numerical distribution according to height groups. This condensed information can now be displayed by means of a graphical representation from which the distribution in the several groups can be seen at a glance.

Associated with this distribution are certain calculable quantities which specify the chief characteristics of the distribution (with these we deal later), and so the original mass of information relating to the heights of a thousand boys, perhaps arranged initially in the most haphazard fashion, is finally replaced by the graphical representation and these calculable quantities. It is thus possible to interpret the statistics as a whole and, possibly, to suggest uses to which this information can be applied.

A second example of a less simple nature, taken from astronomy, concerns the stars of the *main sequence* for which detailed information had been accumulated relating, in particular, to luminosity (or intrinsic brightness), effective temperature (or, more crudely, surface temperature) and mass. In the early years of the present century it was found that there existed a definite relationship between luminosity and spectral type, later translated in terms of temperature. Further,

it was seen that the most luminous stars of the main sequence were also the most massive, and the feeblest the least massive. Here was a challenge to the theoretical investigator of the physical conditions within a star, culminating (in 1924) in Eddington's discovery† of the 'mass-luminosity relationship'.

Not infrequently, the characteristics of a statistical distribution relating to a series of observations or measures have led to a greater insight into the particular problem under consideration and to new discoveries such as Bradley's discovery of aberration and nutation referred to in § 2·04 (p. 39).

1·02. Frequency distributions

We consider as an example the statistics relating to the heights of a thousand men in a regiment. To condense the information conveniently, we derive the number of men in each of seven groups according to height; the first group relates to heights between 62 and 64 in., with 63 in. as the 'middle height'; the second group relates to heights between 64 and 66 in., with 65 in. as the middle height; and so on.

The number in any group is called the *frequency* for that group, usually denoted in statistical theory by f, with suffices $1, 2, \ldots$ to indicate the group concerned;‡ thus the frequency for the ith group is denoted by f_i. The statistics for the seven groups are shown in Table 1, the middle heights being shown in the second row and the corresponding frequencies in the third row.

Table 1. *Distribution of heights of* 1000 *soldiers*

Group ...	1	2	3	4	5	6	7
Middle height (in.)	63	65	67	69	71	73	75
Frequency (f)	20	80	190	250	280	170	10

One method of displaying the data in table 1 is shown in Fig. 1, in which heights are indicated horizontally and frequencies are indicated vertically. Consider the first group; the range is 62–64 in. (A to C in the figure), and the frequency is 20. Erect a rectangle $ABDC$ on AC as base and height AB equal to the frequency. Repeat this construction for each of the groups, thus obtaining a

† A. S. Eddington, *Mon. Not. R. Astr. Soc.* **84**, 308, 1924; see also *Internal Constitution of the Stars* (Cambridge University Press, 1926), p. 116.

‡ When we are dealing with observations of, say, a particular characteristic of a star or of the measurement of a physical quantity, the term corresponding to frequency is called the *weight*, usually denoted by w, with the appropriate suffix.

series of rectangles. Since the bases of the rectangles are equal the ratios of the rectangular areas are equivalent to the ratios of the frequencies.

Such a diagram is called a *histogram* and it gives a condensed, although not wholly complete, picture of the distribution of heights amongst the thousand soldiers.

A second method of illustrating the distribution of heights consists in drawing, in Fig. 1, straight lines between the successive mid-points P_1, P_2, ... of the upper horizontal sides BD, FG, ... of the rectangles

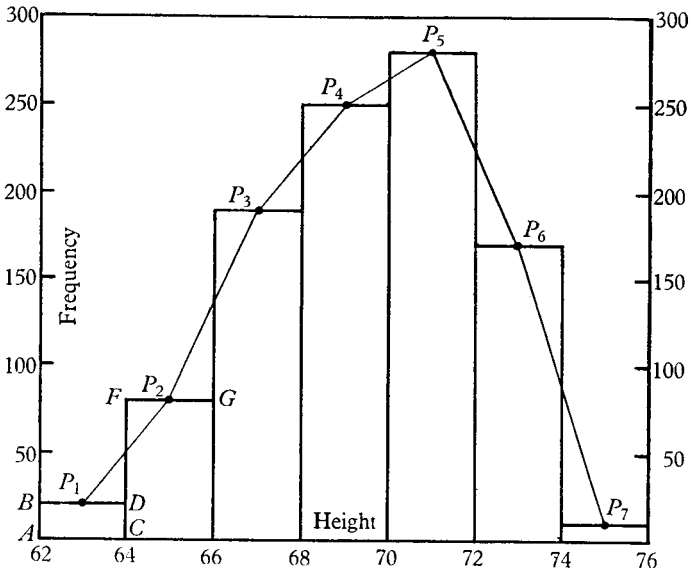

Fig. 1

forming the histogram. The succession of straight lines P_1P_2, P_2P_3, ... is called a *frequency polygon*; the coordinates of each of the vertices P_1, P_2, ... relate the frequency in a group to the middle height for that group. Like the histogram, the frequency polygon gives a reliable, although not wholly complete, representation of the distribution of heights.

In Fig. 1 the width of each rectangle is 2 in.; this is called the *class interval*, which we denote by c.

For any set of statistics the choice of class interval is dictated in practice according to circumstances, the principal factor being the total frequency—in our example this is the total number of soldiers furnishing the statistics, namely, one thousand. If we had required

a more detailed picture of the distribution of heights, we could have taken the class interval to be 1 in., in which event we would derive the frequencies corresponding to height groups 62–63, 63–64, ... The histogram and the frequency polygon would then be constructed according to the principles illustrated in Fig. 1.

1·03. Illustration of the practical application of statistics

The mathematical developments in the theory of statistics have become so recondite in recent years that there is a danger of overlooking the ultimate aims of the collector of statistics, which are, first, the interpretation of the data and, secondly, the possibility of using the results for some well-defined purpose.

As an example which is instructive—and may be illuminating and comforting to the student—we consider the results of an examination taken by 800 candidates and for which the maximum mark is 200; for simplicity, we shall assume that no candidate has scored full marks.

In Table 2† the numbers, denoted by f, of candidates with marks between 0 and 24 (both inclusive), between 25 and 49 (both inclusive), ..., are shown in the fourth row, the third row containing the middle marks for the several groups.

Table 2. *Distribution of marks in an examination*

Group ...	1	2	3	4	5	6	7	8
Range of marks	0–24	25–49	50–74	75–99	100–124	125–149	150–174	175–199
Middle mark	12	37	62	87	112	137	162	187
Frequency (f)	24	64	120	168	184	160	64	16
U	24	88	208	376	560	720	784	800
V (% of U)	3	11	26	47	70	90	98	100

Instead of drawing a histogram or frequency polygon we adopt a third method which consists of constructing a curve, known as an *ogive* curve, based on the numbers, U, of candidates who obtain less than 25, 50, 75, ... marks; these numbers are shown in the fifth row of the table. Thus, the candidates who *fail* to score 50 marks consist of those in groups 1 and 2, and the number, U, in this case is $24+64$ or 88; similarly, the candidates who fail to score 75 marks consist of those in groups 1, 2 and 3, and the number is $88+120$ or 208; and so on.

† To enable the reader to concentrate on principles rather than on arithmetical details the frequencies, f, in an actual examination have been rounded off to the nearest multiple of 8 so that the percentages in the last line of the table are integers.

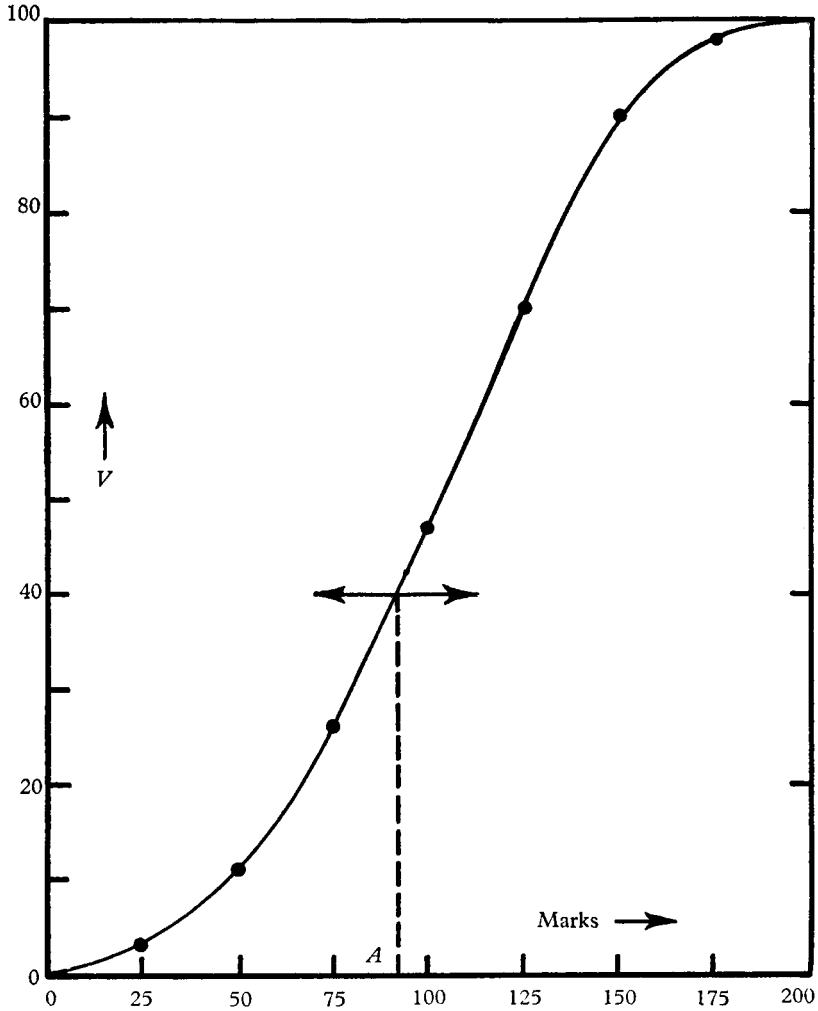

Fig. 2

The last row gives the percentage, V, of candidates failing to score 25, 50, ... marks.

In Fig. 2 marks are indicated on the horizontal axis and the values of V on the vertical axis; the values of V in Table 2 are plotted against the corresponding marks 25, 50, 75, ...

Suppose that the statistics in Table 2, together with Fig. 2, refer to a final degree examination in chemistry for the year 1957; we assume, further, that a similar treatment of marks was made for the

corresponding examination in 1956. Now, as a general rule, examiners attempt to set papers of equal difficulty from year to year and, with large numbers of candidates taking the examination annually, it may be anticipated that under normal circumstances† the percentage of failures would remain substantially constant from year to year.

Suppose that, in 1956, 40 % of the candidates failed and that it is desirable to keep the standard of performance—as reckoned in this way—uniform from year to year. It is seen at once from Fig. 2 that a failure of 40 % in 1957 corresponds to the abscissa indicated by the point A, that is, to a mark of 92. If the standard is to be blindly followed, all candidates with marks not greater than 92 would be adjudged to fail. In actual practice the examiners would give careful attention to all 'border-line' candidates with marks, say, between 87 and 99, and, following a detailed scrutiny of their scripts, they might feel disposed to readjust the marks of several candidates, increasing the marks in some cases and diminishing the marks in others.

Alternatively, if the pass mark in 1956 was 100 with 40 % of the candidates failing, the curve in Fig. 2 shows that in 1957 the percentage of candidates failing to reach 100 marks is 47 %. This suggests either that the paper (or papers) set in 1957 was harder than the paper (or papers) set in 1956, or that the candidates in 1957 were, as a whole, somewhat inferior in ability to the candidates in 1956. The business of the examiners in such an event is to effect some compromise in the light of relevant information available to them, such as the general quality of the work, authentic reports of serious epidemics in schools, teachers' estimates, and so on.

1·04. Characteristics of a frequency distribution

(i) *The mean.* We take a simple example. If f_1 men each earn x_1 shillings per week and f_2 men each earn x_2 shillings per week, the total number of shillings earned is $f_1x_1+f_2x_2$ and the *mean* wage is $(f_1x_1+f_2x_2)/(f_1+f_2)$ shillings per week. Generally, if $x_1, x_2, \ldots, x_n$ are the weekly wages of $f_1, f_2, \ldots, f_n$ men respectively, the mean, $\bar{x}$, is given by

$$\bar{x}=\frac{f_1x_1+f_2x_2+\ldots+f_nx_n}{f_1+f_2+\ldots+f_n}=\frac{\sum_{i=1}^{n} f_ix_i}{\sum_{i=1}^{n} f_i}. \qquad (1)$$

† *Abnormal* circumstances would include, for example, the partial dislocation of education, to different degrees, in different parts of the country due to large-scale illness or, as was found between 1939 and 1945, to war-time conditions.

Here we regard x_1, x_2, ... as the particular values of a *variable* x, usually referred to in the theory of statistics as the *variate*.

Let N denote the total number of men concerned—or, in statistical language, the total frequency. Then

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

and (1) becomes

$$\bar{x} = \frac{1}{N} \sum_{1}^{n} f_i x_i . \qquad (2)$$

This formula is applicable in all such problems where the frequencies associated with particular values of the variable are given.

(ii) *The median.* The median is the value, m, of the variable dividing the distribution of statistics such that the frequency of values less than m is equal to the frequency of values greater than m.

Referring to Table 2 (p. 4) we see from the values of U that 376 candidates have marks less than 100—that is, up to 99—and that 560 candidates have marks less than 125—that is, up to 124. It is evident that the median, m, lies between 99 and 124 marks. Moreover, we could ascertain the median mark by examining the individual marks of the 184 candidates in group 5 (marks between 100 and 124) by arranging these in increasing numerical sequence and noting the mark not exceeded by the first 24 candidates in the group, 24 being the balance between 376 and 400.

This procedure is tedious and in practice we proceed as follows. If it is assumed that the marks in group 5 are uniformly distributed in the range 100–124, we have to find the mark, m, such that 24 out of the 184 candidates in the group have this mark. Then

$$m = 99 + \tfrac{24}{184} . 25 = 102{\cdot}3,$$

or, to the nearest integer, $m = 102$.

This result is verifiable from Fig. 2, being the mark associated with 50 % of the candidates.

(iii) *The quartiles.* The quartiles are the three values of the variable—q_1, q_2 and q_3—such that the frequencies for values of the variable between 0 and q_1, between q_1 and q_2, between q_2 and q_3, and between q_3 and N (the total frequency over the whole range) are all equal, each being $\frac{1}{4}N$. It is evident that q_2 is the median.

The practical method of calculating q_1 and q_3 is the same as that for calculating the median.

In the theory of errors the quartiles q_1 and q_3 have a special significance (see § 4·09).

(iv) *The mode.* The mode is the value of the variable for which the frequency is a maximum.

Table 1 shows that the maximum frequency in the various groups is 280—in group 5—corresponding to the middle height 71; this last number is the abscissa of the point, P_5, in the frequency polygon (Fig. 1) corresponding to the frequency 280; accordingly, the mode is 71.

If we had taken the class interval to be, say, a half of that to which Table 1 refers, the value of the mode found in this way may be expected to differ slightly from the first value 71; the value of the mode is thus dependent to some extent on the selection of the class interval.

1·05. Calculation of the mean

By 1·04 (2), the mean $\bar{x}$ of the n discrete values $x_1, x_2, \ldots, x_n$ of a variable x, with frequencies $f_1, f_2, \ldots, f_n$, is given by

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{n} f_i x_i. \tag{1}$$

Considering our example in §1·03 we can find $\bar{x}$ *accurately* by forming the sum in (1) by means of the frequencies associated with the individual marks 0, 1, 2, ..., 199. This would be an intolerably long and tedious calculation.

We can condense the calculation and obtain a sufficiently reliable value of $\bar{x}$ by assuming that, in the first group (Table 2, p. 4), the average mark of the 24 candidates in the group is the middle mark 12 for the range 0–24 marks; then $x_1 = 12$ and $f_1 = 24$. Similarly, we assume that the average mark of the 64 candidates in group 2 is the middle mark 37; then, $x_2 = 37$ and $f_2 = 64$. The remaining groups are treated in a similar way. Our calculation for $\bar{x}$, by means of (1), would then be

$$\bar{x} = \frac{1}{N}[24.12 + 64.37 + \ldots + 16.187],$$

where $N = 800$. It can be verified that $\bar{x} = 100\frac{3}{4}$, or, to the nearest integer, $\bar{x} = 101$. The calculation is still long and tedious.

To simplify the arithmetical work still further we introduce the following device. Let a denote a convenient value of the variable which we estimate to be in the neighbourhood of $\bar{x}$. Let

$$x_i = a + \xi_i, \tag{2}$$

from which the values of ξ_i (which can be positive or negative) are readily derived. Then,† by (1) and (2),

$$N\bar{x} = \Sigma f_i(a + \xi_i) = a\Sigma f_i + \Sigma f_i \xi_i. \tag{3}$$

† In (3) and elsewhere the limits 1 to n in the summations will be omitted for simplicity when no confusion is likely to be caused.

Let $\bar{\xi}$ denote the mean of the quantities ξ_i with frequencies f_i; then $N\bar{\xi}=\Sigma f_i\xi_i$, and (3) becomes

$$N\bar{x}=Na+N\bar{\xi}$$

or

$$\bar{x}=a+\bar{\xi}. \tag{4}$$

This formula is the appropriate one when the class interval is unity. When the class interval, c, is different from unity, we can simplify the computations still further by writing

$$\xi_i=cu_i, \tag{5}$$

from which $\Sigma f_i\xi_i=c\Sigma f_iu_i$, so that

$$\bar{\xi}=c\bar{u}, \tag{6}$$

where $\bar{u}$ is the mean of the quantities u_i with associated frequencies f_i. Formula (4) then becomes

$$\bar{x}=a+c\bar{u}. \tag{7}$$

We use the statistics of Table 2 to calculate $\bar{x}$ first by means of (4) and, secondly, by means of (7); the details are found in Table 3, the second column of which gives the middle mark of each of the groups and the third column gives the corresponding frequencies.

Table 3. *Calculation of the mean*

(1)	(2)	(3)	(4)	(5)	(6)	(7)
Group	x_i (middle mark)	f_i	ξ_i	$f_i\xi_i$	u_i	f_iu_i
1	12	24	−100	− 2,400	−4	− 96
2	37	64	− 75	− 4,800	−3	−192
3	62	120	− 50	− 6,000	−2	−240
4	87	168	− 25	− 4,200	−1	−168
5	112	184	0	**−17,400**	0	**−696**
6	137	160	+ 25	+ 4,000	+1	+160
7	162	64	+ 50	+ 3,200	+2	+128
8	187	16	+ 75	+ 1,200	+3	+ 48
		800		**+ 8,400**		**+336**

Summary: $N=800$; $\Sigma f\xi=-9000$; $\Sigma fu=-360$.

As a rough guess the mean mark $\bar{x}$ is between 87 and 112 (groups 4 and 5) and, since from column 2 the class interval, c, is 25, it is evident from (5) that it would be convenient to have the values of ξ_i given by multiples of 25, for then the values of u_i will be integers. A value of a satisfying these desiderata is clearly 112.

In column 4 are to be found the values of ξ_i and in the next column the values of $f_i\xi_i$; the sum of the negative values of $f_i\xi_i$ is shown near the middle of column 5 and the sum of the positive values at the bottom of the column; these values are $-17{,}400$ and $+8{,}400$ respectively; the final sum is then -9000, as shown in the summary at the foot of the table. Hence, by (4), since the total frequency, N, is 800,

$$\bar{x} = 112 + \tfrac{1}{800}(-9000) = 100\tfrac{3}{4},$$

or, to the nearest integer, $\bar{x} = 101$, agreeing with the result previously stated.

In column 6 the values of u_i, based on (5), are given and in the last column the values of $f_i u_i$ are found, with the negative and positive sums shown as in column 5. The value of $\bar{u}$ is given by

$$\bar{u} = \tfrac{1}{800}(-696 + 336) = -\tfrac{9}{20},$$

and hence, by (7),

$$\bar{x} = 112 - 25 . \tfrac{9}{20} = 100\tfrac{3}{4},$$

or, to the nearest integer, $\bar{x} = 101$.

The advantages of the second method (that involving $\bar{u}$) over the first method (involving $\bar{\xi}$) as regards simplicity and economy of calculation are sufficiently obvious to require no further emphasis.

1·06. Moments

Consider the frequency polygon in Fig. 3 with n vertices $P_1, P_2, \ldots, P_n$, and, in particular, the ith vertex, P_i, corresponding to the value, x_i, of the variable; in the figure, $OQ_i = x_i$ and $Q_iP_i = f_i$. Let AB be any line parallel to OY and denote the abscissa of A by a.

The r-th moment of the frequency distribution about AB—that is, about the line $x = a$—is denoted by $\mu_r(a)$ and defined by

$$\mu_r(a) = \frac{1}{N}\Sigma f_i(x_i - a)^r, \tag{1}$$

r being a positive integer, including zero.

Write, as before,

$$\xi_i = x_i - a. \tag{2}$$

Then, as in 1·05 (4)

$$\bar{\xi} = \bar{x} - a. \tag{3}$$

From (1) we then have, by means of (2),

$$\mu_r(a) = \frac{1}{N}\Sigma f_i \xi_i^r. \tag{4}$$

We refer specifically to the algebraical quantity $\xi_i \equiv x_i - a$ as the *deviation* of x_i from a; the values of ξ_i may be positive or negative.