The case/control method, however, is likely to give an answer more speedily than the cohort with its prolonged follow-up and, in some circumstances, it is likely to be the only possible approach. For example, with a relatively rare condition like multiple sclerosis it might be quite impossible to categorise a sufficiently large population to give incidence rates from future occurrences within a reasonable span of time. It would not be at all difficult to accumulate a large group of cases and retrospectively to explore their past.

The cohort method does not, however, always involve a subsequent waiting period. It can be applied so long as a population can be defined *at any specific time* and then its subsequent events noted, e.g. in records already accumulated. For instance, for the live births that took place in a given hospital over some earlier years it might be possible to determine from the available clinical records whether or not the mother had an X-ray of the abdomen. One might then determine by inquiry or other already available records, the health of the child 5 years later. In other words, the cohort method has thus been applied to existing records.

In some circumstances one might well choose to make a pilot case/control inquiry before embarking upon a more arduous cohort investigation.

In conclusion, though the prospective approach must usually be the 'method of choice,' there can certainly be no *one* right way in which to make every investigation.

Summary

One of the most decisive and difficult tasks in any inquiry is the construction of an appropriate form of record. Care must be taken to ensure that the questions are clear and unambiguous and, as far as possible, self-explanatory. Each question should require some answer and the standard of accuracy necessary for the purpose in hand should be considered. To ensure a high rate of return a form may need to be kept short. On the other hand, to ensure an unbiased return there may be occasions when extra questions are useful. Pilot inquiries can be invaluable in revealing the difficulties and defects of a proposed large-scale investigation.

Many inquiries follow one of two forms of approach – the case/control or retrospective (looking backwards) and the cohort or prospective (looking forwards). The latter has much in its favour but with rare events may be impossible. There can be no one right or wrong way in all circumstances.

5 Presentation of Statistics

Once a number of observations or measurements has been made, or collected, the first object must be to express them in some simple form which will permit, directly or by means of further calculations, conclusions to be drawn. The publication, for instance, of a long series of the responses of patients to a specific treatment is not particularly helpful (beyond providing material for interested persons to work upon), for it is impossible to detect, from the unsorted mass of raw material, relationships between the various factors at issue. The worker must first consider the questions which he believes the material is capable of answering and then determine the form of presentation which brings out the true answers most clearly. For instance, let us suppose the worker has amassed a series of after-histories of patients treated for gastric ulcer and wishes to assess the value of the treatments given, using as a measure the amount of incapacitating illness suffered in subsequent years. There will be various factors, the influence of which it will be of interest to observe. Is the age or sex of the patient material to the upshot? Division of the data must be made into these categories and tables constructed to show how much subsequent illness was in fact suffered by each of these groups. Is the after-history affected by the type of treatment? A further tabulation is necessary to explore this point. And so on. The initial step must be to divide the observations into a relatively small number of groups, those in each group being considered alike in that characteristic for the purpose in hand. To take another example, relevant to the remarkable and fascinating history of scarlet fever with its fluctuating virulence, Table 1 shows some past fatality-rates from scarlet fever in hospital; for this purpose children within each year of age up to 10 and in each five-year group from 10 to 20 are considered alike with respect to age. It is, of course, possible that by this grouping we are concealing real differences. The fatality-rate at 0–6 months may differ from the fatality-rate at 6–12 months, at 12–18 months it may differ from the rate at 18–24 months. To answer that question, further subdivision – if the number of cases justifies it – would be necessary. In its present form (accepting the figures of hospital cases at their face value) the grouping states that fatality declines nearly steadily with age, a conclusion which it would be impossible to draw from the 11 526 original unsorted and ungrouped records. The construction of a *frequency distribution* is the first desideratum – i.e. a table showing the frequency with which there are present individuals with some defined characteristic or characteristics.

TABLE 1

THE HISTORY OF SCARLET FEVER

Fatality-rate of Hospital Cases in the Years 1905–14

Age in Years	Number of Cases	Number of Deaths	Fatality-rate per cent
(1)	(2)	(3)	(4)
0–	46	18	39·1
1–	383	43	11·2
2–	881	50	5·7
3–	1169	60	5·1
4–	1372	36	2·6
5–	1403	24	1·7
6–	1271	22	1·7
7–	986	21	2·1
8–	864	6	0·7
9–	673	5	0·7
10–	1965	14	0·7
15–19	513	3	0·6

The fatality or case-mortality rate is the proportion of patients with a particular disease who die.

The Frequency Distribution

In constructing the frequency distribution from the original unsorted records, the first point to be settled is the number of classes or groups to be used. As one of the main objects of the resulting table is to make clear to the eye the general tenor of the records, too many groups are not desirable. Otherwise, with as many as, say, 50 groups, the tabulation will itself be difficult to read and may fail to reveal the salient features of the data. On the other hand, a very small number of groups may equally fail to bring out essential points. Also, in subsequent calculations made from the frequency distribution, we shall often need to suppose that all the observations in a group can be regarded as having the value of the middle of that group, e.g. if in a frequency distribution of ages at death there are 75 deaths at ages between 40 and 45 (i.e. 40 or over but less than 45) we shall presume that each can be taken as 42½. In fact, of course, some will be less than 42½, some more, but so long as the groups are not made unduly wide and the numbers of observations are not too few, no serious error is likely to arise; 42½ will be the mean age of the 75 deaths, or very near to it.

In general, therefore, some 10 to 20 groups is usually an appropriate number to adopt. Also it is usually best to keep the class- or group-interval a constant size. For instance, in Table 1 the class-interval is 1 year of age up to age 10, and it is easy to see from the figures that the absolute number of cases per year of age rises rapidly to a maximum at age 5–6 and then declines. There is, however, an abrupt and large rise in the absolute number at age 10 merely because the class-interval has been changed from 1 year to 5 years – the mean number of cases per year of age would here be only 1965 ÷ 5, or 393. This change of interval makes the basic figures (not the rates) more difficult to read and sometimes makes subsequent calculations more laborious. Generally, therefore, the class-interval should be kept constant. Also, as a general rule, the distribution should initially be drawn up on a fine basis – i.e. with a considerable number of groups, for if this basis proves too fine, owing to the numbers of observations being few, it is possible to double or treble the group-interval by combining the groups. If, on the other hand, the original grouping is made too broad, the subdivision of the groups is impossible without retabulating much of the material.

As an example of the construction of the frequency distribution we may use the following 88 death-rates which were taken from one of the Occupational Mortality Supplements of the Registrar-General of England and Wales. The rates, as set out in four columns, have been copied merely in the order of occupations as adopted in the report and it is desired to tabulate them (see page 46).

The first step is to find the upper and lower limits over which the tabulation must extend. The lowest rate is 3·9, the highest is 19·3. We have therefore a range of 15·4. A class interval of 1 will give 16 groups and clearly will be convenient to handle. On this basis we may take the groups, or classes, as 3·5 to 4·4, 4·5 to 5·4, 5·5 to 6·4, and so on. Setting out these groups, we may enter each rate by a stroke against the appropriate group. It is convenient to mark each fifth by a diagonal line as shown. Addition is then simple and errors are less likely to occur. (In making this tabulation the groups may be set out as above, 3·5 to 4·4, 4·5 to

5·4, etc., or as in the table, 3·5–, 4·5–, etc. (see page 47). It is most undesirable to have them in the form 3·5–4·5, 4·5 to 5·5, etc., since observations precisely on a dividing line, e.g. 4·5, will then sometimes be absentmindedly put in one group and sometimes in the other.)

The Annual Death-rate per 1000 at ages 20–64 in each of 88 Occupational Groups (untabulated material)

(1)	(2)	(3)	(4)
7·5	10·3	7·7	6·8
8·2	10·1	12·8	7·1
6·2	10·0	8·7	6·6
8·9	11·1	5·5	8·8
7·8	6·5	8·6	8·8
5·4	12·5	9·6	10·7
9·4	7·8	11·9	10·8
9·9	6·5	10·4	6·0
10·9	8·7	7·8	7·9
10·8	9·3	7·6	7·3
7·4	12·4	12·1	19·3
9·7	10·6	4·6	9·3
11·6	9·1	14·0	8·9
12·6	9·7	8·1	10·1
5·0	9·3	11·4	3·9
10·2	6·2	10·6	6·0
9·2	10·3	11·6	6·9
12·0	6·6	10·4	9·0
9·9	7·4	8·1	9·4
7·3	8·6	4·6	8·8
7·3	7·7	6·6	11·4
8·4	9·4	12·8	10·9

This method is satisfactory if the number of observations is not very large. But it always has the disadvantage that the only check of accuracy is to repeat the work, and if some differences are found it is not always easy to locate the error or errors. A better method is to enter the observations on cards, one to each card, and then to 'deal' the cards into their packs. These packs can then be checked, that they contain only the correct components, and added.

The final figures resulting from this construction of the frequency distribution are given in Table 2, from which it can be clearly seen that the majority of the death-rates lie between 6·5 and 11·5 per 1000 and that they are fairly symmetrically spread round the most frequent rate of 8·5–9·4. To avoid unduly lengthening the table by the inclusion of 4 groups with no entries against them, the final group has been termed 13·5 and over. As a rule this is an undesirable procedure unless the entries can

Process of Tabulation of 88 Death-rates

Death-rate		
3·5–	|	1
4·5–	||||	4
5·5–	𝍸	5
6·5–	𝍸 𝍸 |||	13
7·5–	𝍸 𝍸 ||	12
8·5–	𝍸 𝍸 𝍸 |||	18
9·5–	𝍸 𝍸 |||	13
10·5–	𝍸 𝍸	10
11·5–	𝍸 |	6
12·5–	||||	4
13·5–	|	1
14·5–		. .
15·5–		. .
16·5–		. .
17·5–		. .
18·5–	|	1
19·5 +		. .
Total		88

also be precisely specified, as is done here in a footnote. Without that specification full information on the spread of the rates has not been given to the reader and he may be hampered if he wishes to make calculations from the distribution. A similar caution relates to the lowest group.

TABLE 2

THE ANNUAL DEATH-RATE PER 1000 AT AGES 20–64 IN 88 DIFFERENT OCCUPATIONAL GROUPS

Death-rate per 1000	Number of Occupational Groups with given Death-rate
3·5–	1
4·5–	4
5·5–	5
6·5–	13
7·5–	12
8·5–	18
9·5–	13
10·5–	10
11·5–	6
12·5–	4
13·5 and over	2*
Total	88

* 1 death-rate of 14·0 and 1 death-rate of 19·3.

Statistical Tables

Returning to Table 1 (p. 44), this may be used in illustration of certain basic principles in the presentation of statistical data.

(i) The contents of the table as a whole and the items in each separate column should be clearly and fully defined. For lack of sufficient headings, or even any headings at all, many published tables are quite unintelligible to the reader without a search for clues in the text (and not always then). For instance, if the heading given in column (1) were merely 'age,' it would not be clear whether the groups refer to years or months of life. The unit of measurement must be included.

(ii) If the table includes rates, as in column (4), the base on which they are measured must be clearly stated – e.g. death-rate per cent, or per thousand, or per million, as the case may be (a very common omission in published tables). To know that the fatality-rate is '20' is not helpful unless we know whether it is 20 in 100 patients who die (1 in 5) or 20 in 1000 (1 in 50).

(iii) Whenever possible the frequency distributions should be given in full, as in columns (2) and (3). These are the basic data from which conclusions are being drawn and their presentation allows the reader to check the validity of the author's arguments. The publication merely of certain values descriptive of the frequency distribution – e.g. the arithmetic mean or average – severely handicaps other workers. For instance, the information that for certain groups of patients the mean age at death from cancer of the lung was 54·8 years and from cancer of the stomach was 62·1 years is of very limited value in the absence of any knowledge of the distribution of ages at death in the two classes.

(iv) Rates or proportions should not be given alone without any information as to the numbers of observations upon which they are based. In presenting experimental data, and indeed nearly all statistical data, this is a fundamental rule (which, however, is constantly broken). For example, the fatality-rate from smallpox in England and Wales (ratio of registered deaths to notified cases) was 42·9 per cent in 1917 while in the following year it was only 3·2 per cent. This impressive difference becomes less convincing of a real change in virulence at that time when we note that in 1917 there were but 7 cases notified, of whom 3 died, and in 1918 only 63 of whom 2 died. (Though the low rate of 1918 *may* mark the presence of variola minor.) 'It is the essence of science to disclose both the data upon which a conclusion is based and the methods by which the conclusion is attained.' By giving only rates or proportions, and by omitting the actual numbers of observations or frequency distributions, we are excluding the basic data. In their absence we can draw no valid conclusion whatever from, say, a comparison of two, or more, percentages. The news 'media' are great offenders in this respect. They will glibly report that the influenza epidemic has risen 5-fold in a week but rarely are we told whether the basic figures are 10 and 50 cases or 100 and 500.

It is often stated that it is wrong to calculate a percentage when the number of observations is small, e.g. under 50. *But so long as the basic numbers are also given* it is difficult to see where the objection lies and how such a presentation can be misleading – except to those who disregard the basic figures and who, therefore, in all probability will be misled by any presentation. If, for example, we have 9 relapses in 23 patients in one group and 4 relapses in 15 patients in another, it is difficult for the mind to grasp the difference (if any). Some common basis would seem essential, and percentages are a convenient one. In this comparison they are, in fact, 39 and 27. They will thus be seen, in view of the small numbers involved, to be not very different. The fundamental rules should be that the writer gives no percentages without adding the scale of events underlying them and the reader accepts no percentages without considering that scale. Inevitably with small numbers care will be needed in drawing conclusions; but that is true whatever the basis from which they are viewed.

(v) On a point of detail it is sometimes helpful in publishing results to use one decimal figure in percentages to draw the reader's attention to the fact that the figure is a percentage and not an absolute number. An alternative, especially useful in tables, is to give percentages (or rates) in italics and absolute numbers in bold type. This variation in type, too, often makes a large table simpler to grasp. As a general principle two or three small tables are to be preferred to one large one. Often the latter *can* be read but its appearance may well lead to it going unread.

(vi) Full particulars of any deliberate exclusions of observations from a collected series must be given, the reasons for and the criteria of exclusion being clearly defined. For example, if it be desired to measure the success of an operation for, say, cancer of some site, it might, from one aspect, be considered advisable to take as a measure the percentage of patients surviving at the end of 5 years, *excluding those who died from the operation itself* – i.e. the question asked is 'What is the survival-rate of patients upon whom the operation is successfully carried out?' It is obvious that these figures are not comparable with those of observers who have included the operative mortality. If the exclusion that has been made in the first case is not clearly stated no one can necessarily deduce that there is a lack of comparability between the records of different observers, and misleading comparisons are likely to be made. Similarly one worker may include among the subsequent deaths only those due to cancer and exclude unrelated deaths – e.g. from accident – while another includes all deaths, irrespective of their cause. Definition of the exclusions will prevent unjust comparisons.

Sometimes exclusions are inevitable – e.g. if in computing a survival-rate some individuals have been lost sight of so that nothing is known of their fate. The number of such individuals must invariably be stated, and it must be considered whether the lack of knowledge extends to so many patients as to stultify conclusions (*vide* p. 212).

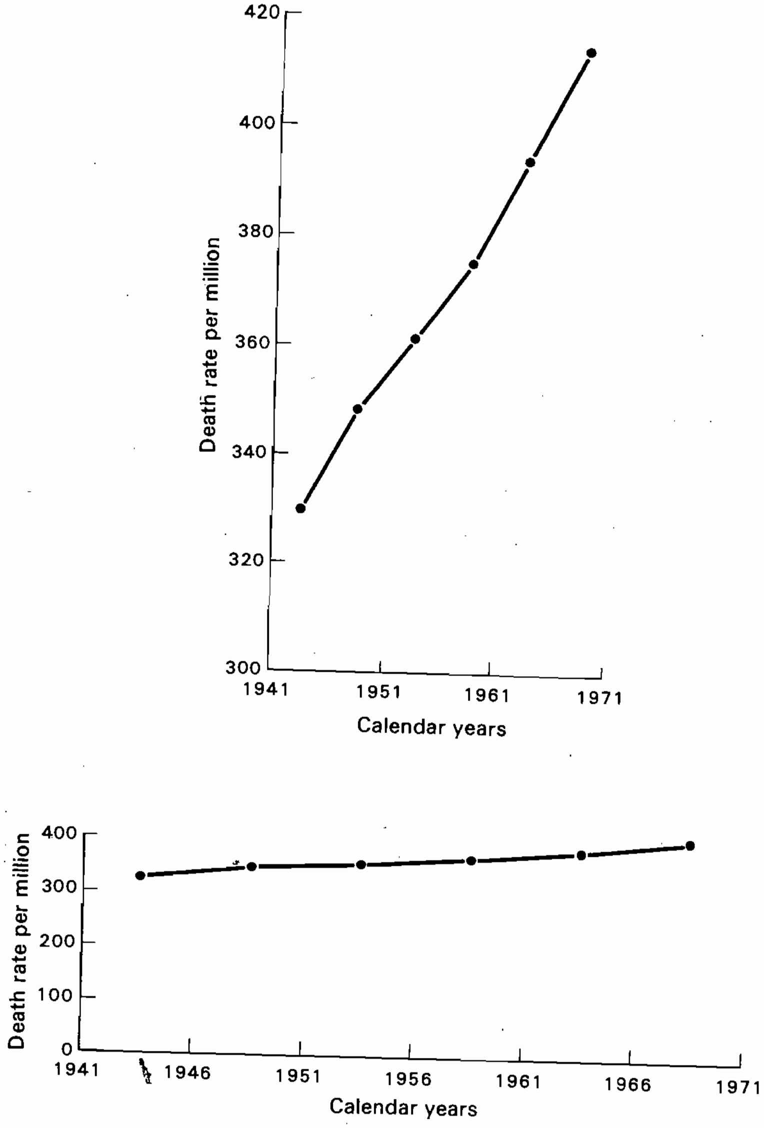
Beyond these few rules it is very difficult, if not impossible, to lay down rules for the construction of tables. The whole issue is the arrangement of data in a concise and easily read form. In acquiring skill in the construction of tables probably the best way is to consider published tables critically with such questions as these in mind: 'What is the *purpose* of this table? What is it *supposed* to accomplish in the mind of the reader? . . . wherein does its failure of attainment fall?' Study of the tables published by the professional statistician – e.g. in the Annual Reports of the Registrar-General of England and Wales – will materially assist the beginner.

Graphs

Even with the most lucid construction of tables such a method of presentation will often give difficulties to the reader, especially to the non-numerically-minded reader. The presentation of the same material diagrammatically often proves a very considerable aid and has much to commend it if certain basic principles are not forgotten.

(i) The sole object of a diagram is to assist the intelligence to grasp the meaning of a series of numbers by means of the eye. If – as is unfortunately often the case – the eye itself is merely confused by a criss-cross of half a dozen, or even a dozen, lines, the sole object is defeated. The criterion must be that the eye can with reasonable ease follow the movements of the various lines on the diagram from point to point and thus observe what is the change in the value of the ordinate (the vertical scale) for a given change in the value of the abscissa (the horizontal scale). The writer should always remember, too, that he is familiar with the data, and what may be obvious to him is not necessarily obvious to the reader. The object of the graph is to make it obvious, or at least as clear as possible, and simplicity is invariably the keynote.

(ii) The second point to bear in mind in constructing *and in reading* graphs is that by the choice of scales the same numerical values can be made to appear very different to the eye. Figs. 1 and 2 are an example. Both show the same data – namely the trend of the death rate in women from cancer of the breast in England and Wales between 1941 and 1970. In Fig. 1 the increase in mortality that has been recorded appears to have been exceedingly rapid and of serious magnitude while in Fig. 2 a slow and far less impressive rise is suggested. This difference is, of course, due to the differences in the vertical and horizontal scales and to the fact that in Fig. 1 the vertical scale does not start at zero (see below). In reading graphs, therefore, the scales must be carefully observed and the magnitude of the changes interpreted by a rough translation of the points



Figs. 1 and 2. Death rates per million women from cancer of the breast. England and Wales 1941 to 1970.

into actual figures. In drawing graphs undue exaggeration or compression of the scales must be avoided.

It must be considered also whether a false impression is conveyed, as quite frequently happens, if the vertical scale does not start at zero but at some point appreciably above it. Figures 3 and 4 show what can be done in this way with infant mortality-rates of 29, 28, 27 and 26 per 1000 in four consecutive years and the aid of a little ingenuity, i.e. change of scale and omission of the zero base line. Thus the reader should always look with care to see whether the vertical scale does start at zero and at the nature of the scale chosen. Indeed, he may, as a safeguard, well need to translate the points on the graph back to the original figures, e.g. the 29, 28, 27, and 26 of Fig. 3. In other words he would be safer with the data tabulated and not in a diagram.

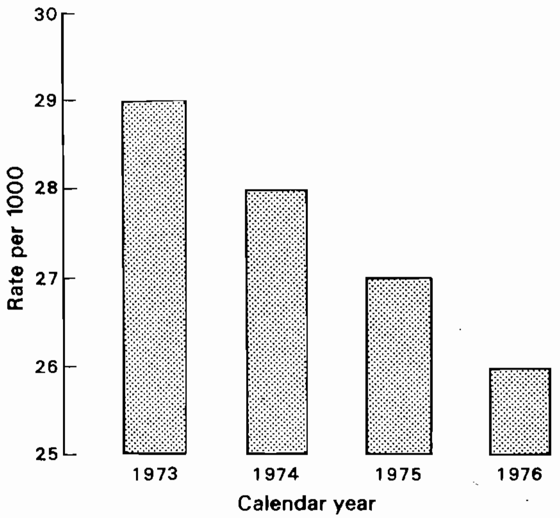


Fig. 3. The infant mortality-rate in four consecutive years in the City of X, drawn on an exaggerated scale and omitting the zero base.

(iii) Graphs should always be regarded as subsidiary aids to the intelligence and *not* as the evidence of associations or trends. That evidence must be largely drawn from the statistical tables themselves. It follows that graphs are an unsatisfactory ***substitute*** for statistical tables. A deaf ear should be turned to such editorial pleading as this: 'If we print the graphs would it not be possible to take the tables for granted? Having

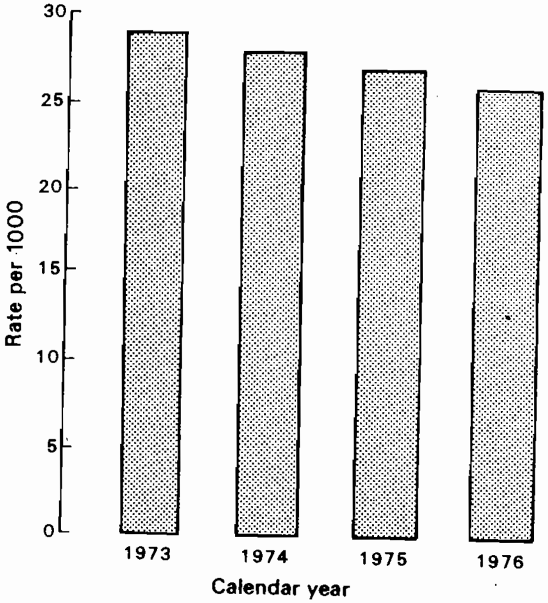


Fig. 4. The infant mortality-rate in four consecutive years in the City of X, drawn on a reasonable scale and including the zero base.

given a sample of the process by which you arrive at the graph is it necessary in each case to reproduce the steps?' The retort to this request is that statistical tables are *not* a step to a diagram, they are the basic data. Without these basic data the reader cannot adequately consider the validity of the author's deductions, and he cannot make any further analysis of the data, if he should wish, without laboriously and inaccurately endeavouring to translate the diagram back into the statistics from which it was originally constructed (and few tasks are more irritating). There are, of course, some occasions when the statistical data are not worth setting out in detail and a graph may be sufficient. But careful thought is advisable before that procedure is followed.

(iv) The problem of scale illustrated in Figs. 1 and 2 is also an important factor in the comparison of trend lines. Thus Fig. 5 shows the trend of the infant mortality rate and of the death rate of young children in England and Wales between 1931 and 1971. Unless the scale of the ordinate (the vertical scale) is carefully considered, the inference drawn from this graph might well be that over this period the infant mortality rate declined relatively more than the death rate in young children. Actually the precise reverse is the case – relatively the death rate at ages 1–4 years declined considerably more than the infant mortality rate. In 1966–70 the infant mortality rate was 29 per cent of the rate recorded in 1931–35 while the death rate at ages 1–4 years was only 12 per cent of its earlier level. ***Absolutely*** infant mortality shows the greater improvement (from 62 per 1000 to 18 per 1000 compared with 6·6 and 0·8 in the death rate at ages 1–4); but relatively the death rate of young children shows the advantage. If it is the ***relative*** degree of improvement that is at issue, Fig. 5 is insufficient. For this purpose the rates in each quinquennium may be converted into percentages based upon the rates in 1931–35 as shown on the next page in Fig. 6. (or plotted on semi-logarithmic paper which automatically shows the rate of change in the rates).

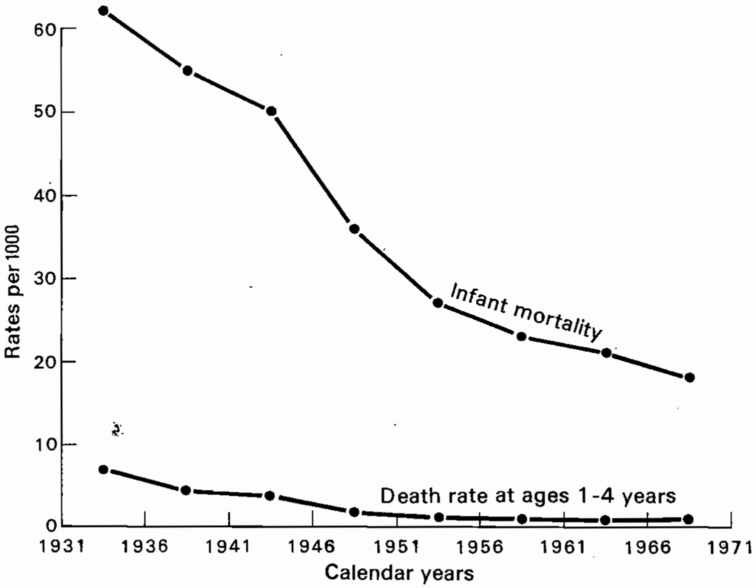


Fig. 5. Infant mortality per 1000 and death-rate of children aged 1–4 years per 1000 in England and Wales from 1931–35 to 1966–70.

(v) It is a ***sine qua non*** with graphs, as with tables, that they form self-contained units, the contents of which can be grasped without reference to the text. For this purpose inclusive and clearly stated headings must be given, the meaning of the various lines indicated, and a statement made against the ordinate and abscissa of the characteristics to which these scales refer (see Figs. 1 to 8).

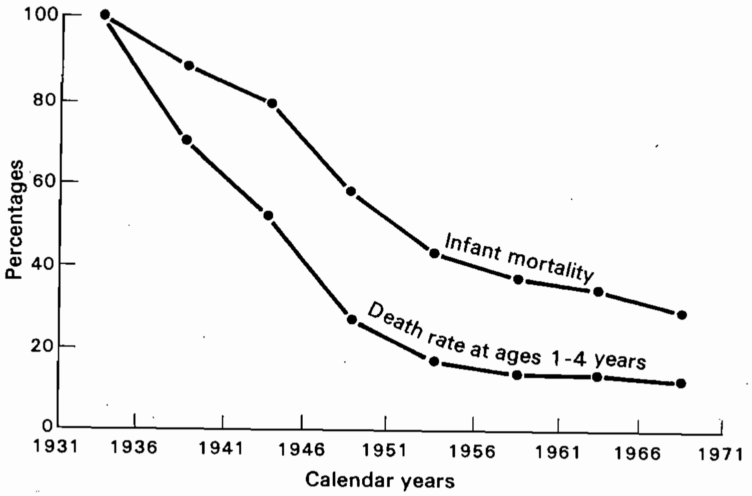


Fig. 6. The rates in each 5-year period expressed as percentages of the corresponding rates in 1931–35.

Frequency Diagrams

Many types of diagrams have been evolved to bring out the main features of statistical data. In representing frequency distributions diagrammatically two types are primarily used: (1) the *frequency polygon* and (2) the ***histogram***; the latter is usually the better. In both of these the base line denotes the characteristic which is being measured and the vertical scale reveals the frequency with which it occurs. In Table 1 (page 44) the numbers of deaths from scarlet fever in a certain study were as follows:—

Age in years	*Number of Deaths*
0–	18
1–	43
2–	50
3–	60
4–	36
5–	24
6–	22
7–	21
8–	6
9–	5
10–	14
15–19	3

Here we have a frequency distribution of some deaths from scarlet fever in relation to age. To graph this distribution the base line is divided into single years of age and the vertical scale is made to relate to the number of deaths. To draw the frequency polygon, or line diagram, a point representing the observed frequency is made against the ***middle*** of the age group concerned. Thus there were 18 deaths at ages 0–1, and against the middle of the 0–1 age interval on the base-line a point is made against 18 on the vertical frequency scale. Similarly a point is made against the middle of the 1–2 age group against 43 on the vertical scale. Finally these points are joined one to another to complete the diagram (see Fig. 7).

Some difficulty, it will be noted, occurs with the final figures since the scale is here changed from one year to five years. An erroneous picture will result if the 14 deaths are plotted against the mid-point of the age group 10–15. It may then appear that the number of deaths is again rising though in terms of *per year of age* they are still declining. It will be better, therefore, to find this average number of deaths per year of age by dividing the 14 by the 5 years to which they relate, and plotting this point

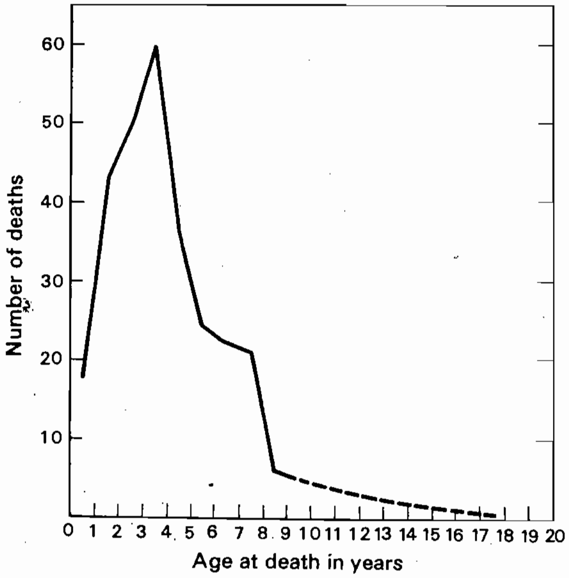


Fig. 7. The frequency of some deaths from scarlet fever in relation to age. Frequency polygon.

(2·8) against the mid-point of the age group to which it relates.

This difficulty is more clearly overcome, and the figures more clearly shown diagrammatically, by the use of the histogram, in which the frequency is represented by an area corresponding to the number of observations (or to the proportion of the total falling in each group). Thus in the present example a point is placed against 18 on the vertical scale both against age 0 and age 1 on the base line. A rectangle is then drawn to show this frequency. Similarly a point is placed against 43 both above age 1 and age 2, and this rectangle is completed. The area of each rectangle is thus proportional to the number of deaths in the year of age concerned. To maintain a correct area when the scale of age grouping changes, we must divide the recorded deaths by the new unit of grouping, i.e. the 14 and the 3 in the two final groups by 5, giving, on the average, 2·8 and 0·6 deaths per year of age. These figures then relate to the whole of the 5-year age group and the rectangles will extend over the ages 10–15 and 15–20. In other words, in the absence of more detailed data,

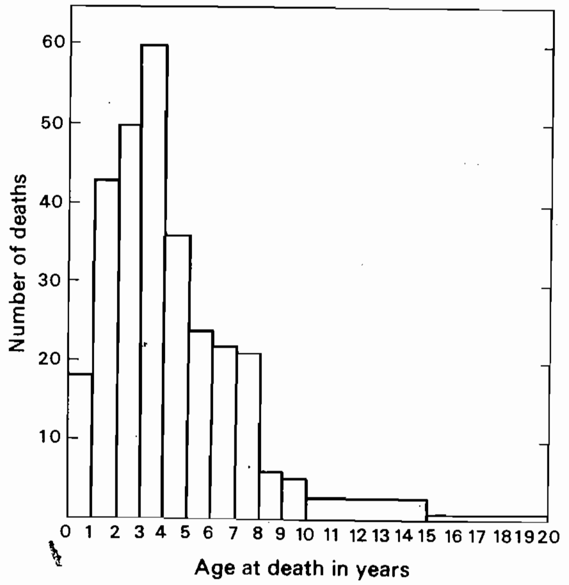


Fig. 8. The frequency of some deaths from scarlet fever in relation to age. Histogram.

we plot the figures as if there were 2·8 deaths at ages 10–11, 2·8 at ages 11–12, 2·8 at ages 12–13, and so on; 0·6 at ages 15–16, 0·6 at ages 16–17, 0·6 at ages 17–18, and so on (see Fig. 8). If thought desirable the actual number of deaths reported can be written inside, or just above, the rectangles.

Another very simple form of diagram is the *bar chart*, which can be used to show pictorially the absolute, or relative, frequency of events, e.g. the numbers of deaths due to specified causes, the percentage of patients with a particular disease showing certain symptoms, as illustrated in Fig.9.

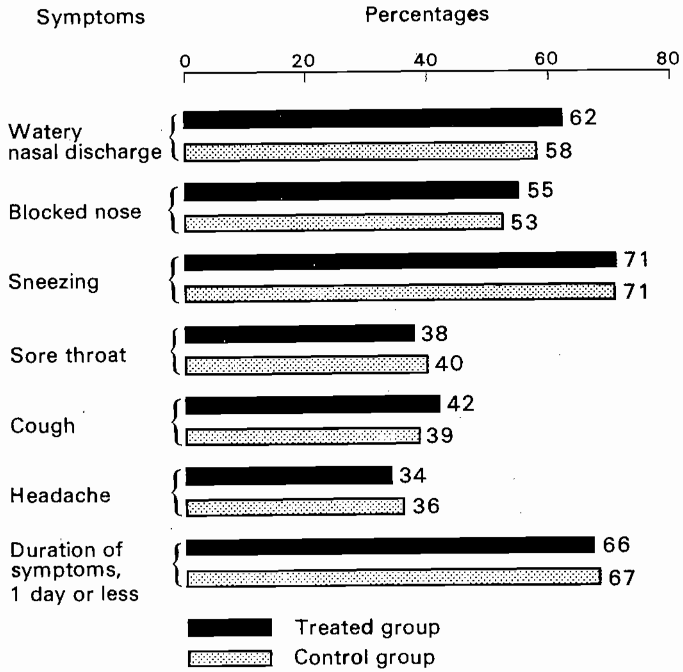


Fig. 9. A bar chart showing the percentage of patients with defined *presenting symptoms* of the common cold in two groups, T to be treated with an antihistamine compound and C to be treated with a placebo.

In a preliminary exploration of the degree of association between two characteristics, e.g. body temperature and the erythrocyte sedimentation rate, a *scatter diagram* is valuable (see p. 162).

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.