1. Introductory

1. The ordinary treatment of probability begins with the assumption that the chance that a certain event will occur is known, and proceeds to solve the problems that arise from the combination of events or the repetition of a particular experiment; it proves that a certain result is more likely to occur from experiment than any other, that a result based on a limited number of trials is unlikely to differ greatly from the expected result, and that the proportional deviation from the most probable result will generally decrease as the number of trials is increased.

Experiments can easily be made to show that the theoretical method leads to results which can be realized in practice when the probabilities can be estimated accurately beforehand; for example, various trials have been made with coin tossing in which it has been found that if five coins are tossed together and the number of them coming down 'heads' is recorded, then the distribution of the cases will agree with the binomial expansion $(\frac{1}{2}+\frac{1}{2})^5$ as the ordinary theory leads us to expect. Sequences of 'heads' or 'tails' form a series approximating to the geometrical progression with a common ratio of $\frac{1}{2}$, and the drawing of cards from a pack gives a result closely agreeing with the numbers that theoretical work suggests.

2. It frequently happens, however, that the probabilities are not known, and it is impossible to tell whether we are dealing with an experiment like coin tossing or sequences or card-drawing; in fact, the only thing known is the distribution of the number of cases into certain groups, and in these circumstances the inverse problem of tracing the theoretical series to which the statistics approximate may become an important matter. The difficulty of the subject is increased because statistics do not give the theoretical distribution exactly, and it is impossible to tell where the differences between the actual and theoretical results lie. To make the position clearer it will be well to restate the problem and ask whether it is possible to find the theoretical series to which a series, resulting from a statistical experiment, approximates. It may be difficult, perhaps impossible, to trace the probabilities corresponding to a given case, but yet practicable to form a reasonable opinion of the series of numbers that might be

reached if the experiment could be repeated an infinite number of times. On turning to the reasons which make it advisable to find this ideal result to which statistics approach, it will be seen that the exact elementary probabilities are not of supreme importance, and a reasonable representation of the series is of far greater practical value. We notice that one of the first objects of a statistician or an actuary dealing with statistical work is to express the observations in a simple form so that practical conclusions can be easily drawn from the figures that have been collected. If the available statistics fall naturally into fifty or sixty groups, he has to decide how they can be arranged to bring out the important features of the problem on which he is working; whereas if he can find a few numbers closely connected with the original series which can be used as an index to the whole, he can then give the result in a way that might assist comparison with similar statistics, and enable others who have to deal with the facts to appreciate the whole distribution more readily than they could do if it remained in its original form. The statistician has also to supply approximate values for intermediate terms when only a few can be obtained from his experience, or complete or continue a series when only a part of it is known. In many cases he has to keep the same terms as in his original series, but remove the roughnesses of material due to limitations in the number of cases available for his investigation; that is, he has to graduate his data.

3. In reality these objects are much alike, for if the statistical tables can be represented by an algebraic or transcendental formula, we can replace the whole series of numbers by a few values (the constants in the formula) which, if we deal systematically with the distributions we meet, facilitate comparison or enable us to supply missing terms, while the roughness of the original material can be removed by making a suitable formula represent the original statistics as nearly as possible. If a formula is based on theoretical considerations, it may also give a solution of the problem in probabilities mentioned at the outset, and we see that both the practical and theoretical requirements can be dealt with at the same time, for the smooth series sought by the theoretical student is the same thing as the formula required for practical work.

4. The advantages of any system of curves depend on the simplicity of the formulae and the number of classes of observations that can be dealt with satisfactorily, for a complicated expression is very little

improvement on the original groups of statistics, and a system which is not capable of general application leaves the statistician in difficulties whenever it breaks down. One other thing is necessary; if a formula is known to be a suitable one, there must be some method of finding the arithmetical constants that will give a good agreement in the particular case. Such a method, if it is to be of practical use, must be simple, reliable and capable of general and systematic application.

A broad idea of the objects to be accomplished ought to be kept clearly before the mind; they are likely to be forgotten because of the large amount of detail necessarily connected with the subject. It is also important because the advantages of systematic treatment are often overlooked, and short cuts and rough and ready methods are adopted to the detriment of the work, and formulae having no scientific basis and having no connection with others suitable to similar cases are sometimes used in rather haphazard fashion by statisticians. The consequence is that generalization is impossible, and where a law might be found one can see little but a great variety of attempts by energetic workers to reach their own conclusions regardless of the value of comparative statistics.

2. Frequency distributions

1. If statistics are arranged so as to show the number of times, or frequency with which, an event happens in a particular way, then the arrangement is a frequency distribution. Although some of our results will be of wider applicability, we shall generally confine our attention to these distributions.

It is necessary to have a name for the formula used to describe such distributions, and the term 'frequency-curve' has been adopted for the purpose.

2. Some distributions give the number of cases falling in a certain group of values of the independent variable, while others (e.g. Example 5 of Table 1) give the number of cases for an exact value. In the former case the exact values of the independent variable to which the groups correspond must be considered; for instance, 'exposed to risk at age x' includes those from $x-\frac{1}{2}$ to $x+\frac{1}{2}$, but the number of deaths at duration n those from n to $n+1$. When statistics are represented graphically, effect should be given to these differences, and, to bring out the points a little more clearly, the diagrams on pages 6 and 7 have been prepared. Note that 'curtate duration' in Example 1 is naturally represented by a frequency polygon, as it takes only integer values, but if the data are regarded as giving actual duration they should be represented by a histogram, since generally 'curtate duration n' means 'actual duration n to $(n+1)$'. The drawings of distributions, such as those in the diagrams, are called frequency polygons or histograms (Examples 1, 3, 5).

3. When statistics give the number of cases for an exact value of the independent variable, it is simple to plot them in a diagram by drawing ordinates and joining their tops (frequency polygons). In the case of groups of values there is a little complication, for we can either draw a rectangle standing on the entire base (histograms) or put in ordinates at the middle points of the bases and then join their tops (Example 3). The former method gives the correct idea of the amount of information conveyed by the statistics, but, for some purposes (e.g. for seeing the possible shape of the curve), the latter is more convenient, though it is open to technical objection. Cases such as Examples 1 and 4 are best expressed by the kind of drawing

given, while Example 3 though open to technical objection gives a better indication to most people of the shape of the actual distribution than a block diagram.

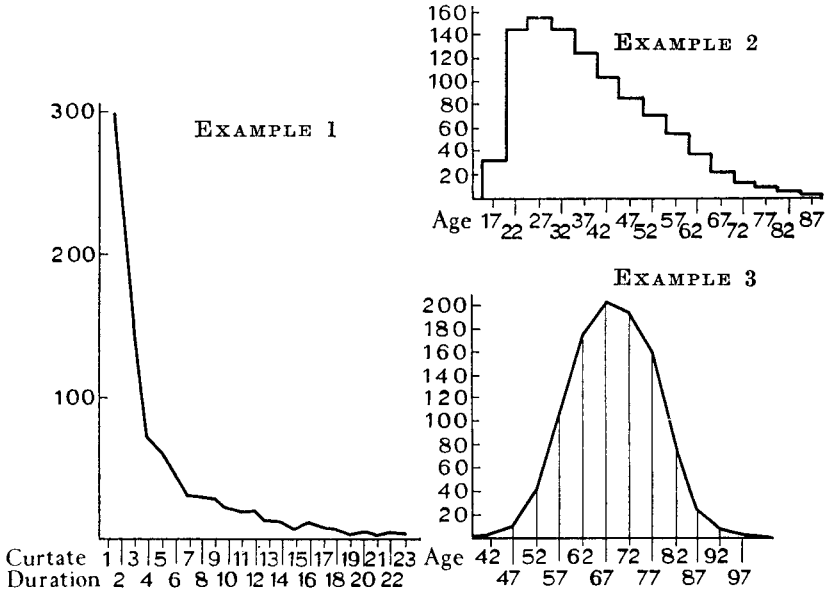
TABLE 1

EXAMPLE 1		EXAMPLE 2	EXAMPLE 3		EXAMPLE 4	EXAMPLE 5	
Curtate durations	Withdrawals with monthly incidence '0' in year of exit, *Principles and Methods* (p. 92)	Ages	Exposed to risk of sickness (Watson, *M.U. Tables*, p. 19)	Existing at close of observations Without Profit 'Old' Assurances	Existing at close of observations 'Old' Annuities (females)	Terms of the expansion of $1000(\frac{3}{4}+\frac{1}{4})^{12}$	No. of term
1	308	–19	34	...	...	32	1
2	200	20–24	145	...	...	127	2
3	118	25–29	156	...	...	232	3
4	69	30–34	145	...	...	258	4
5	59	35–39	123	...	...	194	5
6	44	40–44	103	3	...	103	6
7	29	45–49	86	9	...	40	7
8	28	50–54	71	42	...	11	8
9	26	55–59	55	111	29	2	9
10	21	60–64	37	176	23	1	10
11	18	65–69	21	200	81	...	11
12	18	70–74	13	193	151	...	...
13	12	75–79	7	160	192	...	...
14	11	80–84	3	73	239	...	...
15	5	85–89	1	26	157	...	...
16	11	90–94	...	6	93	...	...
17	7	95–99	...	1	29	...	...
18	6	100–	...	...	6	...	...
19	1	...	...	...	...	...	...
20	3	...	...	...	...	...	...
21	1	...	...	...	...	...	...
22	3	...	...	...	...	...	...
23	2	...	...	...	...	...	...
...	1000	...	1000	1000	1000	1000	...
True total	1308	...	2 995 724	2674	172	...	...
Mean	4·182	...	37·8750	68·485	79·400	3·998	...
Standard deviation	4·1996	...	2·76810	1·771288	1·774894	1·46215	...
Type	I	...	I	II	VII	...	...

4. The reader is no doubt already familiar with the fact that statistics tend towards a smooth series as the total number of cases is increased, and from this it can be seen how naturally practical statistics lead to the conception of a frequency-curve to describe the smooth distribution that would be obtained if an infinite supply of

homogeneous material were available for investigation. In other words, such curves would give an approximation to the total 'population' of which the particular case investigated was a sample.

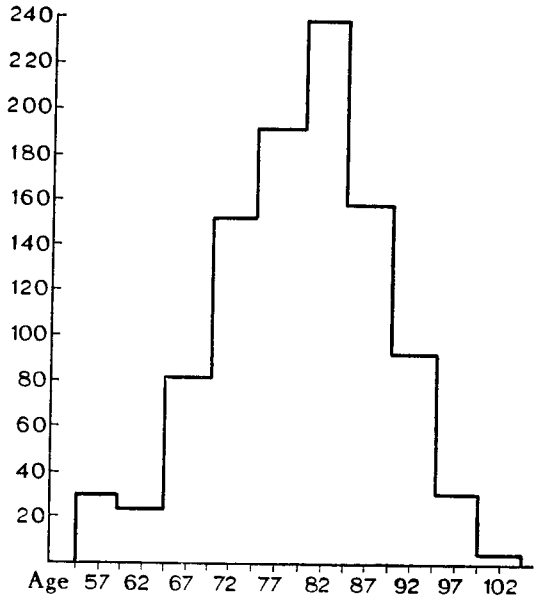
5. It may be noticed that a frequency-curve can be interpreted to give a frequency corresponding to every value of the independent variable along the whole range of the distribution, and will not restrict us to a few more or less arbitrary groups as is necessary with



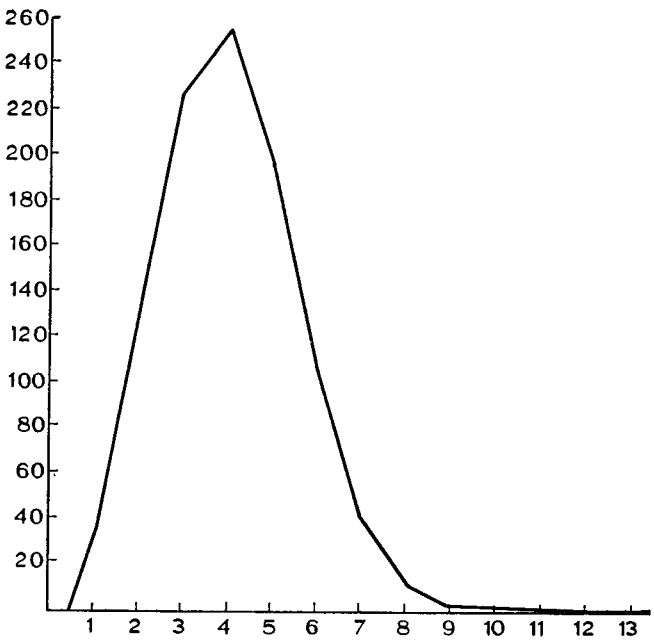
actual statistics. The binomial series and geometrical progression do the same when we imagine we are dealing with something that can be divided into a very large number of groups. Thus, if we mix a large quantity of sand of two colours and take out a fixed quantity of the mixture and record the number of grains of sand of either colour in each drawing, we should obtain a continuous curve from a large number of trials.

6. We will now define some important functions. When a distribution is arranged according to the progressive values of a variable characteristic, e.g. duration, age, etc., the average value of that characteristic (not the average of the frequencies) is called the *mean* of the distribution, and is given by

$$\frac{f_a \times a + f_b \times b + f_c \times c + \ldots + f_n \times n}{f_a + f_b + f_c + \ldots + f_n}$$



EXAMPLE 4



EXAMPLE 5

where f_r is the frequency corresponding to the value r of the variable; thus, in Example 1, 200 is the frequency corresponding to 2. If we assume infinitesimal increments, the mean is given by

$$\int f_x \times x\,dx \Big/ \int f_x\,dx$$

where the limits of the integral will be such as to cover the whole distribution. The mean could also be described as the position of the ordinate through the centre of gravity of the distribution (centroid vertical); this may be of help to some readers.

The *mode* is the characteristic that occurs most frequently; in other words, it is the position of the maximum ordinate. We cannot tell from the rough statistics which ordinate is greatest and the mode can therefore only be determined approximately until the law connecting the various groups, i.e. the frequency-curve, is known.

7. The *standard deviation* measures the way the frequencies are distributed in terms of the unit of measurement. It is given by

$$\sqrt{\left\{\frac{f_a a'^2 + f_b b'^2 + \ldots + f_n n'^2}{f_a + f_b + \ldots + f_n}\right\}}$$

where a', b', ... n' are the distances from the mean. In the form of integrals the standard deviation is

$$\sqrt{\left\{\int f_x \times x^2\,dx \Big/ \int f_x\,dx\right\}}$$

where x is measured from the mean.

As the frequencies farthest from the mean are multiplied by the largest values of x, a large standard deviation shows that the frequency distribution spreads out from the mean, while a small standard deviation shows that the frequency is closely concentrated about the mean. In considering the relative sizes of standard deviations, it is necessary to bear in mind the unit of measurement, because, if a given distribution is arranged in two series, first, according to years of age, and then in quinquennial age groups, the standard deviation will be five times as large in the former case as it is in the latter. This can be seen at once by comparing the two expressions

$$\sqrt{\left\{\int f_x x^2\,dx \Big/ \int f_x\,dx\right\}} \quad \text{and} \quad \sqrt{\left\{\int f_x (\tfrac{1}{5}x)^2\,dx \Big/ \int f_x\,dx\right\}}$$

The former is obviously five times the latter. The values of the standard deviations are given in Table 1 for each case. The diagram on page 10 shows two curves having the same mean B and approximately the same area, but the dotted curve has the larger standard deviation because it spreads out more on each side of the mean.

The *mean deviation* is the average of the absolute values of deviations from the mean. It is given by

$$(f_a|a'|+f_b|b'|+\ldots+f_n|n'|)/(f_a+f_b+\ldots+f_n)$$

or, in the form of integrals by

$$\int f_x|x|\,dx \Big/ \int f_x\,dx$$

where x is measured from the mean.

The reader will notice from the algebraic expressions given above that the mean, mode, standard deviation and mean deviation are not dependent on the number of cases (i.e. on the absolute size of the curve), but merely on the way they are distributed (i.e. on the proportionate numbers or the shape of the curve). The standard deviation measures the 'spread' or 'scatter' of the statistics from the mean. Its square is called the *variance*, and is often denoted var (x), x being the variable concerned.

8. An examination of frequency distributions (see Table 1 and pp. 6 and 7) shows that most of them start at zero, gradually rise to a maximum, and then fall sometimes at a very different rate. If the rise and fall are at the same rate, distribution will be symmetrical about the mean, which must then coincide with the mode. The difference between the mean and mode is therefore a function of the *skewness* or deviation from symmetry. In order to get a satisfactory measure, the spread of the material must be taken into account, and this leads us to measure skewness by the distance between mean and mode divided by standard deviation. If the mean is on the left-hand side of the mode when the statistics are plotted out in diagram, this function will be negative, and to remember the sign it is convenient to write:

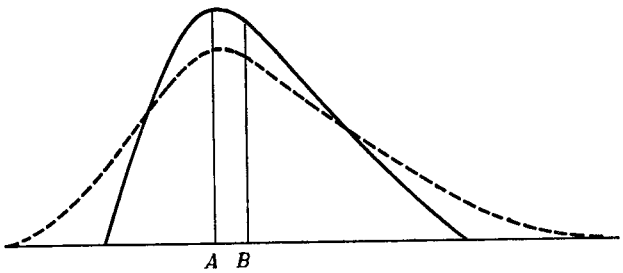
$$\text{Skewness} = \frac{\text{Mean}-\text{Mode}}{\text{S.D.}}$$

The diagram on page 10 will help to show the rationale of the measure for skewness. It gives two curves having the same mean B

and the same mode A, but with different standard deviations, and it is clear that the dotted curve, with its larger standard deviation, is more nearly symmetrical than the other curve. There are other measures of skewness, such as $\sqrt{\beta_1}$ (to be defined in section 4 of the next chapter).

9. We may summarize these functions by saying that the mean and mode fix the position of the curve on the axis; the standard deviation shows how the material is distributed about the mean, and the skewness shows the amount of the deviation from symmetry exhibited by the material.

These preliminary definitions will be sufficient for our present purpose, but the functions defined will be more easily understood when their actual connection with the practical work of curve-fitting has been studied. A student working at the subject for the first time should plot out several distributions on cross-ruled paper, in order to familiarize himself with their nature and appearance. He should also calculate means, standard deviations and mean deviations.



10. Up to this point we have defined our statistics as frequencies, that is, as a number of cases grouped together as alike either because they are actually alike in the sense of Example 5 or because the statistics throw them up in comparatively narrow groupings as in Examples 2, 3 and 4. When, however, we are tabulating our experience we have to deal with individual observations and they are grouped subsequently. From this point of view if there are N observations we may call them $o_1, o_2, o_3, \ldots o_N$, where o_1 may stand for the first observation and may be (see Example 3) one of the 200 existing in the 65–69 group. It might be a case 'existing' at age