PRINCIPLES OF MEDICAL STATISTICS

VI.—FURTHER PROBLEMS OF SAMPLING: PROPORTIONS

In the previous section the concept of the standard error was developed, and was illustrated by the calculation of the standard error of the mean. In addition it was pointed out that every statistical value calculated from a sample must have its standard error—i.e., may differ more or less from the real value in the universe that is being sampled. For example, the standard deviation, or measure of the scatter, of the observations will vary from sample to sample, and its standard error will show how much variability this value is in fact likely to exhibit from one sample to another taken from the same universe. In practical statistical work a value which is of particular importance, owing to the frequency with which it has to be used, is the *proportion*. For example, from a sample of patients with pneumonia we calculate the proportion who die. Let us suppose that from past experience, *covering a very large body of material,* we know that the fatality-rate of patients with pneumonia is, let us say, 20 per cent. (the actual figure, from the point of view of the development of the argument, is immaterial). We take, over a chosen period of time, a randomly selected group of a hundred patients and treat them with serum. Then, presuming that our sample is a truly representative sample of all patients with pneumonia—e.g., in age and in severity—we should observe if serum treatment is valueless about 20 deaths (it may be noted that we are also presuming that there has been no secular change in the fatality-rate from pneumonia). We may observe precisely 20 deaths or owing to the play of chance we may observe more or less than that number. Suppose we observe only 10 deaths; is that an event that is likely or unlikely to occur by chance with a sample of 100 patients? If such an event is quite likely to occur by chance then we must conclude that serum *may* be of value but, so far as we have gone, we must regard the evidence as insufficient and the case unproven. Before we can draw conclusions safely we must increase the size of our sample. If, on the other hand, such an event is very unlikely to occur by chance we may reasonably conclude that serum is of value (that is, of course, having satisfied ourselves that our sample of patients is comparable with those observed in the past in all respects except that of serum treatment). Before we can answer the problem as to what is a likely or an unlikely event we must determine the standard error of a proportion—i.e., the variability of a proportion in samples of a given size taken from the same universe. Presuming the treatment is of no help, then the fatality-rate we should observe on a very large sample is 20 per cent. (or nearly that). How far is the rate likely to differ from that figure in samples of different size?

SAMPLE OF ONE

If our sample comprises only one patient the fatality-rate may be either 0 or 100 per cent.; if the patient dies the fatality-rate is greater than that of past experience; if the patient recovers this is obviously not very convincing evidence in favour of our treatment for, according to past experience, 4 out of 5 patients are likely to recover without our treatment (20 per cent., or only 1 in 5, die).

SAMPLE OF TWO

If the sample is increased to two patients, three events become possible, (i) both may recover, (ii) 1 may recover and 1 may die, (iii) both may die.

On the basis of past experience we can calculate the probability of each of these events occurring. (i) The chance that one may recover is 4/5; the chance that the other may recover is also 4/5; the chance that *both* will recover is the product of these two independent probabilities—i.e., $4/5 \times 4/5 = 16/25$. (ii) The chance that one patient will recover is again 4/5; the chance that the other will die is 1/5; the chance that *both* these events will occur is, therefore, $4/5 \times 1/5$; but this value must be multiplied by 2, for the event can happen in two different ways—viz., patient A may live and patient B die, or patient A may die and patient B live. The probability, therefore, of observing one recovery and one death is $2\ (4/5 \times 1/5) = 8/25$. (iii) Finally the probability of each patient dying is 1/5 and of both patients dying is $1/5 \times 1/5 = 1/25$. We can tabulate these values as follows:

Event.	Probability of event.	Fatality-rate per cent.
Both patients recover ..	16/25 = 0·64	0
One patient recovers, one dies	8/25 = 0·32	50
Both patients die	1/25 = 0·04	100
—	25/25 = 1·00	—

The total probability is 1, for there is no alternative to these three events. Clearly the only event that suggests that our treatment is of value is the recovery of both patients, when the fatality-rate is 0 compared with the 20 per cent. of past experience. The death of one patient in a sample of two gives a fatality-rate of 50 per cent., and of both patients one of 100 per cent., both rates being worse than past experience. But the more favourable event, the recovery of both patients, is obviously an event which is more likely than not to occur by chance; it may be expected to occur 64 times in 100 trials with 2 patients even if the treatment is ineffective. Therefore with a single sample of 2 patients and a normal fatality-rate of 20 per cent. the chance that both will recover is large, and if such a result is observed we cannot deduce from it that our special treatment is of value.

SAMPLE OF THREE

If we increase the sample to three patients four events become possible, (i) all 3 may recover, (ii) 2 may recover and 1 die, (iii) 1 may recover and 2 die, (iv) all 3 may die.

The probability of each event can be calculated as before. (i) The chance of the recovery of all three patients is $4/5 \times 4/5 \times 4/5 = 64/125$. (ii) The chance that two may recover and one die is $4/5 \times 4/5 \times 1/5$; this must be multiplied by 3, for this event can happen in three different ways since any one of the three patients may be the one to die; this equals 48/125. (iii) The chance that one may recover and two may die is $4/5 \times 1/5 \times 1/5$, also multiplied by 3 for this event can also happen in three ways; this equals 12/125. (iv) Finally, the chance that all three may die is $1/5 \times 1/5 \times 1/5$, an event which can happen only in one way, and equals 1/125. Tabulating we have :

Event.	Probability of event.	Fatality-rate per cent.
Three recover	64/125 = 0·512	0
Two recover, one dies ..	48/125 = 0·384	33·3
One recovers, two die ..	12/125 = 0·096	66·7
Three die	1/125 = 0·008	100·0
—	125/125 = 1·000	—

The only event that favours our treatment is, again, the recovery of all the patients. Any other event gives a higher fatality-rate than that of past experience—viz., 20 per cent. But the recovery of all 3 patients is an event which is quite likely to occur by chance ; it may be expected to occur 51 times in 100 trials with 3 patients even if the treatment is ineffective. With a single sample of 3 patients, therefore, the chance that they will all recover is large, and again we cannot deduce that our special treatment is of value.

SAMPLE OF FOUR

If we increase the sample to four patients five events become possible, (i) all four may recover, (ii) three may recover and one die, (iii) two may recover and two die, (iv) one may recover and three die, (v) all four may die.

What is the probability of each of these events on the basis of past experience ? (i) The chance that all four recover is 4/5×4/5×4/5×4/5 ; this event can happen in only one way, and the probability equals 256/625. (ii) The chance that three recover and one dies is 4/5×4/5×4/5×1/5, multiplied in this case by 4, for there are four different ways in which this event can happen ; any one of the four patients can be the one to die. The probability of this event is also, therefore, 256/625. (iii) The chance that two recover and two die is 4/5×4/5×1/5×1/5, multiplied in this case by 6, for there are six ways in which the event can happen. For if the patients are named A, B, C, and D, the following events are possible :

Recover.	*Die.*
AB	CD
AC	BD
AD	BC
BC	AD
BD	AC
CD	AB

The probability of this event is, therefore, 96/625. (iv) The chance that only one recovers and three die is 4/5×1/5×1/5×1/5, multiplied, as before, by 4 (for any one of the four may be the fortunate one to recover) ; this equals 16/625. (v) Finally, the chance that all 4 will die is 1/5×1/5×1/5×1/5=1/625. Tabulating :

Event.	Probability of event.	Fatality-rate per cent.
All four recover	256/625 = 0·4096	0
Three recover, one dies ..	256/625 = 0·4096	25
Two recover, two die ..	96/625 = 0·1536	50
One recovers, three die ..	16/625 = 0·0256	75
All four die	1/625 = 0·0016	100
—	625/625 = 1·0000	—

Once more the recovery of all the patients is the only result which gives a fatality-rate lower than that of past experience, but this, again, is an event quite likely to occur by chance ; it may be expected to occur nearly 41 times in 100 trials with 4 patients even if the treatment is ineffective.

SAMPLE OF TEN

We can with samples of any size calculate by these methods the probability of favourable results occurring merely by chance ; as the sample increases in size, however, the calculations become progressively more laborious. But clearly we need not calculate *all* the probabilities. If, for example, we treat ten patients then the only results which are better than that of past experience are those which give no patients at all dying or only 1 patient dying—i.e., fatality-rates of 0 or 10 per cent. If two of the ten patients die the fatality-rate is normal according to past experience, 20 per cent., and if three or more die then it is higher than that of past experience. The probability of all 10 patients recovering equals $(4/5\times 4/5\times 4/5\times 4/5\times 4/5\times 4/5\times 4/5\times 4/5\times 4/5\times 4/5)=(4/5)^{10}=0{\cdot}1074$; the chance of 9 patients recovering and 1 dying $=\{(4/5)^9\times(1/5)\}\times 10$ (for any one of the 10 may be the one to die)$=0{\cdot}2684$. Either of these events gives a result better than past experience, and the probability of observing by chance one *or* the other is the *sum* of the two probabilities, or $0{\cdot}1074+0{\cdot}2684=0{\cdot}3758$. In other words we should expect to get a better result than past experience nearly 38 times in 100 trials with 10 patients even if serum treatment were quite ineffective. Obviously in a single sample of 10 patients a result better than that of past experience is still not an unlikely event to occur by chance, and from such an observation we cannot deduce that serum has reduced our fatality-rate.

SAMPLE OF A HUNDRED

If now we return to our original problem—namely, a sample of 100 patients of whom only 10 die—the probability we need is that with which this result *or a better one* might be expected to occur even if our treatment with serum were quite ineffective—so that we ought to have observed 20 deaths on the basis of past experience. It is *possible* to calculate this by just the same means as were applied to smaller numbers. Tabulating we have the results shown on the top of the opposite page.

The sum of the probabilities will give the number of times we might expect to reach a result as favourable as the one we have observed, or even more favourable, merely as a result of chance. This sum is 0·0057, and we may conclude that only 57 times in 10,000 trials with a hundred patients would such a result turn up merely by chance. Such a result, therefore, suggests that our treatment favourably influenced the survival-rate. But this calculation is extremely heavy and some shorter method is in practice essential.

THE GENERAL CASE

Let us return to the tabulation regarding two patients. This shows, if we write it in percentage form, that if we had 2 patients in each of 100 hospitals the fatality-rate would (on the average) be 0 in 64 of these hospitals, 50 per cent. in 32 of the hospitals, and 100 per cent. in 4 of them. From these figures we can calculate the *mean* fatality-rate in the 100 hospitals and the *standard deviation* of the frequency distribution round that mean. The mean fatality-rate is $(64\times 0)+(32\times 50)+(4\times 100)\div 100=20$ per cent., and the standard deviation will be found, calculated by the ordinary methods, to be 28·3. A similar calculation

Event.	Probability of event.	Fatality-rate per cent.
All 100 recover	$(4/5)^{100} = 0{\cdot}00000000020$	0
99 recover, 1 dies	$(4/5)^{99} \times (1/5) \times 100^{*} = 0{\cdot}00000000509$	1
98 recover, 2 die	$(4/5)^{98} \times (1/5)^{2} \times 4{,}950 = 0{\cdot}000000063$	2
97 recover, 3 die	$(4/5)^{97} \times (1/5)^{3} \times 161{,}700 = 0{\cdot}0000005147$	3
96 recover, 4 die	$(4/5)^{96} \times (1/5)^{4} \times 3{,}921{,}225 = 0{\cdot}00000312$	4
95 recover, 5 die	$(4/5)^{95} \times (1/5)^{5} \times 75{,}287{,}520 = 0{\cdot}00001498$	5
94 recover, 6 die	$(4/5)^{94} \times (1/5)^{6} \times 1{,}192{,}052{,}400 = 0{\cdot}00005928$	6
93 recover, 7 die	$(4/5)^{93} \times (1/5)^{7} \times 16{,}007{,}560{,}800 = 0{\cdot}0001990$	7
92 recover, 8 die	$(4/5)^{92} \times (1/5)^{8} \times 186{,}087{,}894{,}300 = 0{\cdot}0005784$	8
91 recover, 9 die	$(4/5)^{91} \times (1/5)^{9} \times 1{,}902{,}231{,}808{,}400 = 0{\cdot}001478$	9
90 recover, 10 die	$(4/5)^{90} \times (1/5)^{10} \times 17{,}310{,}309{,}456{,}440 = 0{\cdot}003363$	10

* These multipliers are the number of different ways in which the event could happen. Clearly there are 100 ways in which 1 could die and 99 survive; there are 4,950 ways in which 2 could die and 98 survive, and so on.

for three patients gives a mean fatality-rate in the 100 hospitals of 20 per cent. and a standard deviation round it of 23·1; for four patients the mean is 20 per cent. and the standard deviation is also 20·0. The *mean* fatality-rate in samples of each size—viz., 20 per cent.—is the value, it will be noted, that we expect to reach according to past experience; but in the individual sample we shall not necessarily observe this mean value, for round it there will be a variability in the fatality-rate from sample to sample, due to the play of chance, measured by the standard deviation and decreasing as the size of the sample increases. If we could calculate this standard deviation *without* having to find the different probabilities for each event we should have a measure of the variability that will occur by chance in the fatality-rate in samples of different sizes. This calculation is, in fact, very simply made. If on the basis of past experience we expect 20 per cent. of patients to die and 80 per cent. to recover, then the standard deviation round that expected 20 per cent. will be in samples of 2 equal to the square root of $\frac{20\times 80}{2}=28{\cdot}3$, in samples of 3 equal to the square root of $\frac{20\times 80}{3}=23{\cdot}1$, and in samples of 4 equal to the square root of $\frac{20\times 80}{4}=20{\cdot}0$. These values are the same as those found above by the longer calculation. In more general terms the standard deviation, or as it is usually termed the standard error, of a percentage is $\sqrt{\frac{p\times q}{n}}$, where p is the percentage of individuals belonging to one category (e.g., alive), q is the percentage in the other category (e.g., dead), and n is the number of individuals in the sample. We can, therefore, readily find the standard error of the percentage—i.e., the variability it would show from sample to sample—in samples of 100, or more, patients. With 100 patients the standard error is $\sqrt{\frac{20\times 80}{100}}=4{\cdot}0$. In other words, on the basis of past experience we should expect 20 of the 100 patients to die, but in different samples of that size we should not always observe that proportion dying; the proportions observed in samples of one hundred will be scattered round 20 with a standard deviation of 4. We know (as was shown with the standard error of the mean in the previous section) that there will be relatively very few samples in which the proportion actually observed will differ by more than twice the standard error from the mean expected value. For instance, with 100 patients we expected 20 per cent. to die, but as this percentage has, in samples of this size, a standard error of 4, we might by chance observe a value in a single sample as high as $20+2\,(4)=28$ or as low as $20-2\,(4)=12$. Actually we observed a value of 10 per cent. This is beyond the value that might, *according to our criterion*, be likely to arise by chance and, *other things being equal*, we may deduce that it *appears likely* that serum treatment lowered the fatality-rate. The italicised words must be emphasised. It must be recognised that we are weighing probabilities, never, as is sometimes suggested by non-statistical authors of medico-statistical papers, reaching "mathematical proof." A difference between the observed and expected values *may* be a "real" difference (in the sense that the treatment was effective) even though it is not twice the standard error; but the calculation shows that the hypothesis that the difference has occurred by chance is equally valid. If, on the other hand, the difference between the observed and expected values is, say, four times the standard error, this does not "prove" that it is "real" difference; it may still be the result of chance. But the calculation shows that the hypothesis that it is due to chance is unlikely to be true, for such a chance difference is a rare event. The advantage of the calculation is that the investigator is thus enabled "critically to estimate the value of his own results; he may be prevented from wasting his time by erecting some elaborate superstructure of argument on a difference between two averages (or proportions) which is no greater than a difference that might well be obtained on drawing two random samples from one and the same record" (G. U. Yule: Industrial Health Research Board, Report No. 28, 1924, p. 6).

Finally, presuming that the difference recorded between the observed and expected values is more than would be expected from the play of chance, then we must consider carefully whether it is due to the factor we have in mind—e.g., serum—or to some other factor which differentiated our sample—e.g., age or severity of disease—from the general population of patients. Where, as Yule expresses it, "some particular interpretation is rather attractive," the investigator must be the more on his guard.

For the sake of clarity the standard error of the proportion has been deduced on the basis of a figure known from past experience. In actual practice such a figure is not often available or may be an unsatisfactory criterion of the expected level in the observed sample, owing to some secular change. The more usual procedure is the comparison of two percentages recorded over the same period of time in an experimental and a control group. The development of this test is discussed in the next section.

Summary

A statistical value which is of particular importance, from the frequency of its use, is the proportion, or percentage. By simple means the standard error of this value can be calculated, that is the amount of variability it will show from sample to sample for samples of different sizes. The relation of the difference between an expected percentage and an observed percentage to this standard error shows whether that difference is likely or unlikely to have arisen merely by chance. As a convention we take twice the standard error as a criterion. If the difference is more than twice the standard error it is said to be "significant"—i.e., unlikely to have arisen by chance; if it is less than twice the standard error the difference is said to be "not significant"—i.e., it may easily have arisen by chance. The test always involves weighing probabilities, and can never amount to proof in the logical sense. The test can give no information as to the *origin* of the difference beyond saying that chance is an unlikely explanation.

A. B. H.

SPECIAL ARTICLES

HEARING-AIDS

A REPORT TO THE MEDICAL RESEARCH COUNCIL

THE accurate investigation of deafness is a product of the recent advances in sound reproduction and recording which have followed telephone engineering and broadcasting. The application of these to the alleviation of deafness is not a simple one, for individual deaf persons vary in the extent to which they are deaf to different pitches of the auditory range. Dr. and Mrs. Ewing and Dr. Littler,[1] who have prepared a report which has been published by the Medical Research Council, are pioneer advocates in this country of group hearing-aids for classes of severely deaf children, and they show that these aids are safe to use and do not cause any deterioration in hearing.

They open their report by summing up the problem to be considered as follows:—

> (1) How does the ear of a partially or severely deaf patient behave when stimulated by loud sounds and to what extent can speech be made intelligible to him by amplification?
>
> (2) What are the characteristics of the most efficient type or types of aid?
>
> (3) What tests will ensure that the patient may be effectively advised as to what, if any, type of aid is suitable to his individual needs?

They emphasise the first of these three questions, since the deaf patient must be supplied with a portable instrument which will enable him to follow the speech of any speaker under conditions which are little, if at all, subject to his control and often most unsuitable for sound reproduction by any existing mechanical means. It is known that, no matter how short the duration of sound, there is an average upper limit of intensity at all audible frequencies for people with normal hearing. Intensities beyond this upper limit cause discomfort or pain, and the limit is therefore sometimes called the threshold of feeling. The area between this and the threshold of audibility is the auditory-sensation area. This area is restricted in the deaf patient in the sense that his defect prevents him from hearing a sound of intensities and often of frequencies which are audible to the normal ear. No hearing-aid can restore the capacity to respond to sound over the lost range. A highly sensitive aid may amplify a whisper to such an extent that it rises above the raised threshold of audibility, but that part of the area for which he is deaf remains dead ground for all time. The normal ear can respond effectively to sounds ranging from a whisper (something like 15–30 decibels) to that of speech so loud as to approach the threshold of feeling (130 decibels). Efficiency in the supply and use of hearing-aids requires, in the first place, the measurement of the patient's threshold of minimum audibility for speech. This is most conveniently expressed in terms of the standard intensities used to give levels of loudness for speech, music, and noise. For example, a patient who is so deaf that his threshold of audibility for speech is 60 decibels above that of normal listeners cannot be expected to follow average *mezzo forte* conversation at a distance of three feet, for such conversation has an intensity value of approximately 55 decibels above the normal threshold of audibility. A speech threshold 100 decibels above normal implies that there is little capacity to hear speech even uttered in a loud voice close to the ear. Thus the measurement of the threshold of audibility for speech shows the extent to which the deafness interferes with the ability to follow speech, and it also indicates the degree of amplification needed to bring sounds of different levels of loudness within the patient's range of hearing.

Speech is, of course, a mixed sound, but nevertheless the threshold of audibility for it can be reliably deduced from the threshold for pure tones, and these can be ascertained by means of a pure tone audiometer or beat-tone oscillator. Generally the speech threshold is within 5–10 decibels of the lowest reading for pure tones. Experiments were made both by air-conduction and by bone-conduction. The subject was seated in a sound-proof room and signalled to the operator in another room when he could hear the sound. When speech was being tested, and not pure tone only, the speaker sat inside the soundproof room with his lips twelve inches from a microphone. The word "bah" was used for pure hearing, and various single syllable words closely resembling each other, or meaningless double-syllable words were used to test intelligibility. Only two of the patients tested had sufficient acuity for average *mf* conversation at three feet. For one of them *mf* conversation was 10 decibels quieter than a whisper would be to a normal ear. He was the least deaf of the nine patients, but a course of lip-reading

[1] A. W. G. Ewing, M.A., Ph.D., I. R. Ewing, M.Sc., and T. S. Littler, M.Sc., Ph.D (Department of Education of the Deaf, Victoria University of Manchester): The Use of Hearing Aids. Medical Research Council, Special Report Series No. 219. 1937. Pp. 40. 9*d*.