PRINCIPLES OF MEDICAL STATISTICS

IX—FURTHER EXAMPLES AND DISCUSSION OF χ^2

THE χ^2 test has a wide applicability and forms a useful test of "significance" in many medical statistical problems, especially those in which the observations must be grouped in descriptive categories (as in the state of nutrition) and are not capable of being expressed quantitatively; some further examples of its use may, therefore, be supplied.

A TEST OF DIFFERENCES BETWEEN SERA

Let us suppose that batches of serum from different groups of donors are used for the prevention of measles and give the results shown in Table IX. (The actual figures are imaginary though the example is drawn from the London County Council report on the measles epidemic of 1931–32.) The value of serum in general is not here in question, the assumption being made, for the purposes of this illustration, that it has value. The proportion of the total 700 children who after exposure to infection and the administration of serum escaped attack was 73 per cent., but between the different serum groups the proportion varies between 53·8 and 90·0 per cent. Taking into account the numbers of children upon whom each serum was tested, are the different degrees of success in the prevention of measles more than would be likely to occur by chance? Or is it likely that the sera are all equally valuable and the departure from uniformity in the results would be quite likely to occur in groups of the size used?

TABLE IX—*Measles Incidence and Serum Treatment*

Serum No.	Number of children in group.	Number of children in whom measles was "prevented."		Number of children in whom measles was not "prevented."		Percentage of children in whom measles was "prevented."
		Observed number.	Expected number.	Observed number.	Expected number.	
1	120	80	(87·6)	40	(32·4)	66·7
2	150	135	(109·5)	15	(40·5)	90·0
3	90	56	(65·7)	34	(24·3)	62·2
4	90	68	(65·7)	22	(24·3)	75·6
5	110	87	(80·3)	23	(29·7)	79·1
6	60	42	(43·8)	18	(16·2)	70·0
7	80	43	(58·4)	37	(21·6)	53·8
Total	700	511	(511)	189	(189)	73·0

As a first step we presume that an equal degree of success should have been observed in each of the groups; we then measure the observed departure from this uniformity and see whether such departure is compatible with our hypothesis of all the tested sera being equally valuable. As the expected degree of success with each serum, on this hypothesis, we use the proportion of successes observed in the total—i.e., 73·0 per cent. Applying this proportion to each of the groups we find the number of children in whom we expect measles to have been "prevented" or "not prevented" (the italicised figures in parentheses, e.g., 73 per cent. of 120 is 87·6). We observe, for example, that with serum No. 2 considerably more children escaped attack (135) than we expect on our hypothesis of uniformity (109·5). On the other hand, with serum No. 7 fewer children escaped attack (43) than we expect on our hypothesis (58·4). Are these differences more than would be likely to arise by chance?

χ^2 is the sum of the fourteen values of (observed number–expected number)^2 ÷ expected number, e.g. $(80-87{\cdot}6)^2 \div 87{\cdot}6 = 0{\cdot}66$; this sum equals 47·42. Only six expected values have to be calculated independently, by simple proportion, from the 73 per cent. value in the total that we took as the degree of success anticipated with each serum; the remaining values can be calculated by subtraction from the totals at the side and bottom (e.g., if 87·6 of 120 children are expected to escape attack, 32·4 must be expected not to escape attack). Or by the formula $n = (c-1)(r-1)$, $n = (2-1)(7-1) = 6$. The published tables of χ^2 must therefore be entered (i.e., consulted) with $\chi^2 = 47{\cdot}42$ and $n = 6$.

From Fisher's table it will be found for these values that P must be considerably less than 0·01, since the table shows that when $n=6$ P is 0·01 when χ^2 is only 16·81. Here we have a χ^2 nearly three times as large (outside the range of this table). More extensive tables (Tables for Statisticians and Biometricians issued by the Biometric Laboratory of University College, London. Cambridge University Press. 2nd edition. 1924. 15*s.*)[1] show that P is less than 0·000001. In other words if our hypothesis that the sera are all equally valuable in prevention of measles is true, then less than once in a million times in groups of children of the size here tested should we reach merely by chance results which departed from that uniformity of success to the extent that we have observed with these children. We may therefore reject our hypothesis and conclude that these results differ by more than is likely to be due to chance, that, *all other revelant factors being equal*, some sera were more efficient than others.

Whether we need the more exact probability taken from the larger tables is rather a matter for the individual to determine. If P is less than 0·01—i.e., one in a hundred—then we may perhaps be content to say, without finding any more accurate probability, that the departure from uniformity is an unlikely event to occur by chance. Many statisticians, as pointed out in the last section, take $P=0{\cdot}05$ as a conventional level of "significance"—i.e., if P is greater than 0·05 then the observed values do not differ from the expected values by more than might reasonably be ascribed to chance, while if P is less than 0·05, then it is likely that they do differ by more than might be ascribed to chance. The smaller the value of P the smaller, clearly, is the probability that the differences noted are due to chance.

THE HOURLY DISTRIBUTION OF BIRTHS

In Table X distributions are given of a series of live and stillbirths according to the time of day at which they took place. With live births the figures suggest that a high proportion of the total take place during the night and a smaller proportion during the early afternoon and evening. With stillbirths rather the reverse appears to be the case; the proportion during the night is somewhat low, while in the morning and afternoon the number is

[1] These tables are so constructed that they must be entered with $n+1$ instead of with n.

rather high, though the differences are not very uniform. Are these differences between the numbers recorded at the various hours of the day likely to have occurred merely by chance in samples of the observed size? On the hypothesis that they have occurred by chance, that live and stillbirths are both distributed evenly over the day, the number of live births in each three-hourly period should be 4028 and the number of stillbirths 159 (32,224÷8, and 1272 ÷8). These are the expected values on the hypothesis of uniformity.

TABLE X—*Distribution of Live and Stillbirths over the Day*

Time interval.	Number of live births observed.	Observed number–expected number.	Number of still-births observed.	Observed number–expected number.
Midnight– ..	4,064	+36	126	−33
3 A.M.–	4,627	+599	157	− 2
6 A.M.–	4,488	+460	142	−17
9 A.M.–	4,351	+323	195	+36
12 noon–	3,262	−766	150	− 9
3 P.M.–	3,630	−398	178	+19
6 P.M.–	3,577	−451	144	−15
9 P.M.–midnight	4,225	+197	180	+21
Total ..	32,224	0	1272	0

For each distribution, live and stillbirths, eight values of ((observed–expected)^2 ÷ expected) have to be calculated. Their sum is χ^2. With live births χ^2 equals 413·0, with stillbirths it is 23·8. (It may be noted that as the expected value is the same throughout, the quickest way of calculating χ^2 is to sum the squares of the (observed–expected) values and divide this total by the expected number, instead of making separate divisions by the expected number in each instance.) In both sets of data $n=7$, for one value is dependent upon the total of the expected births having to equal the total of the observed births. In both cases the χ^2 table shows that P is less than 0·01. The differences from the uniformity that we presumed ought to be present are therefore more than would be ascribed to chance, and we conclude that neither live nor stillbirths are distributed evenly over the twenty-four hours in these records. The differences of the live births from uniformity are more striking than those of the stillbirths, for they show a systematic excess during the hours between 9 P.M. and 12 noon and a deficiency between 12 noon and 9 P.M.; with the stillbirths there is some change of sign from one period to another which makes the differences from uniformity less clearly marked—perhaps due to the relatively small number of observations. Inspection of these differences themselves adds considerably to the information provided by χ^2. The latter value tells us that the differences are not likely to be due to chance; the differences themselves show in what way departure from uniformity is taking place, and may suggest interpretations of that departure.

INTERPRETATION OF THE ASSOCIATIONS FOUND

It must be fully realised that χ^2 gives no evidence whatever of the *meaning* of the associations found. For instance in Table VIII the value of χ^2 was such that we concluded that the intelligence and the state of nutrition of a group of children were not independent. The interpretation of that association is quite another matter. We cannot say offhand that the state of nutrition *affected* the level of intelligence. Possibly those children who fell in the group with low intelligence had more instances of subnormal nutrition because intelligence is an inherited characteristic and parents of low intelligence may feed their children inefficiently. Similarly one measles serum may have "prevented" measles to a greater extent than another not because of its superior efficiency but because the children to whom it was administered had in fact been less exposed to risk of infection. We need to be satisfied that the children were effectively equivalent in other relevant respects. The value of the χ^2 test is that it prevents us from unnecessarily seeking for an explanation of, or relying upon, an "association" which may quite easily have arisen by chance. But if the association is not likely to have arisen by chance we are not thereby exonerated from considering different hypotheses to account for it. If we use some form of treatment on mild cases and compare the fatality experienced by those cases with that shown by severe cases not given that treatment, the χ^2 value will certainly show that there is an association between treatment and fatality. But clearly that association between treatment and fatality is only an indirect one. We should have reached just the same result if our treatment were quite valueless, for we are not comparing like with like and have merely shown that mild cases die less frequently than severe cases. Having applied the sampling test we must always consider with care the possible causes to which the association may be due.

It must be observed also that the value of χ^2 does not measure the *strength* of the association between two factors but only whether they are associated at all in the observations under study. Given sufficiently large numbers of observations the test will show that two factors are associated even though the degree of relationship may be very small.

THE "FOURFOLD" TABLE

With what is known as a fourfold table—i.e., one with four groups in it—χ^2 may be calculated by means of the expected numbers, in the way previously illustrated, or alternatively from the formula:

$$\chi^2 = \frac{(ad-bc)^2\,(a+b+c+d)}{(a+b)\,(c+d)\,(a+c)\,(b+d)},$$

where a, b, c, and d are the numbers falling in the groups and the additions are the totals in the margins, as in the Table below.

—	Number of individuals— Inoculated.	Not inoculated.	Total.
Number of individuals { attacked ..	a	c	$a+c$
not attacked ..	b	d	$b+d$
Total	$a+b$	$c+d$	$a+b+c+d$

n in this case is only 1 since when one expected value has been found the remainder can be found by subtraction. The χ^2 table is therefore consulted with $n=1$ and the value of χ^2 found from the observations.

If the numbers involved are small the value of χ^2 will be more accurately given by the formula:

$$\chi^2 = \frac{(ad - bc - \frac{1}{2}(a+b+c+d))^2 (a+b+c+d)}{(a+b)(c+d)(a+c)(b+d)},$$

where ad is the bigger of the two cross products (a correction due to F. Yates).

For instance suppose the Table is as follows:—

—	Inoculated.	Not inoculated.	Total.
Attacked	10	65	75
Not attacked ..	30	95	125
Total ..	40	160	200

The larger cross product is 30×65. For χ^2 we, therefore, have

$$\frac{(65 \times 30 - 10 \times 95 - \frac{1}{2}(200))^2 \times 200}{40 \times 160 \times 75 \times 125}$$

and $\chi^2 = 2{\cdot}7$. For $n = 1$ and $\chi^2 = 2{\cdot}7$, P is 0·10, so that the difference between the inoculated and uninoculated is not more than might be expected to occur by chance (10 times in 100 tests would we reach so large a difference by chance). Without the correction $\chi^2 = 3{\cdot}33$ and P is 0·07 which rather exaggerates the association. This is the value we should also reach if we made the calculation by means of the expected values. In this case we should argue that, according to these figures, the chance of being attacked is 75/200. The chance of being inoculated is 40/200. The chance of being an inoculated *and* an attacked person, if the two characteristics are independent (our testing hypothesis) is, then, $(75/200 \times 40/200)$. This is the chance for each of the 200 individuals observed. Therefore the number we should expect to see in the "inoculated-attacked" cell on this hypothesis of independence is 200 times the probability, or 200 $(75/200 \times 40/200) = 15$. It will be noted that the actual number (10) is rather less than the expected number. The difference between these observed and expected values is 5, its square is 25, and dividing by the expected number the contribution of this cell to χ^2 is 1·66. The other expected values can be similarly calculated (or by subtraction from the marginal totals), and their contributions to χ^2 computed. χ^2 then equals $1{\cdot}66 + 1{\cdot}00 + 0{\cdot}42 + 0{\cdot}25 = 3{\cdot}33$, as above.

These results could, of course, be equally well tested by means of the formula for the standard error of the difference of two proportions: 25 per cent. of the inoculated and 40·6 per cent. of the uninoculated were attacked, a difference of 15·6 per cent. The standard error of this difference is, as previously shown, $\sqrt{\frac{37{\cdot}5 \times 62{\cdot}5}{40} + \frac{37{\cdot}5 \times 62{\cdot}5}{160}} = 8{\cdot}6$ (where 37·5 is the percentage attacked in the total group). The difference is not twice its standard error and therefore cannot be regarded as an unlikely event to occur by chance.

THE ADDITIVE CHARACTERISTIC OF χ^2

One further characteristic of χ^2 is useful in practice. Suppose we had three such tables as the above showing the incidence of attacks upon different groups of inoculated and uninoculated persons, observed, say, in different places, and each table suggests an advantage to the inoculated but in no case by more than could fairly easily have arisen by chance—e.g., the χ^2 values are 2·0, 2·5, and 3·0, and, with n equal to 1 in each case, the P's are 0·157, 0·114, and 0·083. The systematic advantage of the inoculated suggests that some protection is conferred by inoculation. We can test this uniformity of result, whether taken together these tables show a "significant" difference between the inoculated and uninoculated, by taking the sum of the χ^2 values and entering the χ^2 table again with this sum and the sum of the n values—namely, $\chi^2 = 2{\cdot}0 + 2{\cdot}5 + 3{\cdot}0 = 7{\cdot}5$ and $n = 3$. P in this case is slightly larger than 0·05 so that we must still conclude that the three sets of differences, though very suggestive, are not quite beyond what might fairly frequently arise by chance in samples of the size observed.

Finally it must be noted that χ^2 must always be calculated from the *absolute* observed and expected numbers and never from percentages or any other proportions.

SUMMARY

The χ^2 test is particularly useful for testing the presence, or absence, of association between characteristics which cannot be quantitatively expressed. It is not a measure of the strength of an association, though inspection of the departure of the observed values from those expected on the no-association hypothesis will often give some indication, though not a precise numerical measure, of that degree. As with all tests of "significance" the conclusion that a difference has occurred which is not likely to be due to chance, does not exonerate the worker from considering closely the various ways in which such a difference may have arisen. In other words, a difference due to one special factor is not a corollary of the conclusion that a difference is not due to chance. There are a number of ways in which the value of χ^2 can be calculated, some speedier than others. An alternative method is given for fourfold tables and a correction in such instances for small samples. The calculation of χ^2 must always be based upon the absolute numbers. In this discussion the mathematical development of the test and the foundation of the table by means of which the value of χ^2 is interpreted in terms of a probability have been ignored. The test can be applied intelligently without that knowledge, provided the rules for calculation of the values of χ^2 and n are followed, and the usual precautions taken in interpreting a difference observed.

A. B. H.

ALVARENGA PRIZE.—The College of Physicians of Philadelphia will award the Alvarenga prize, amounting this year to $200, on July 14th to the author of the best essay submitted on any branch of medicine. Essays must be original unpublished contributions and should be typewritten in standard English, or be accompanied by an English translation. Each essay must be sent without signature, but must be plainly marked with a motto and be accompanied by a sealed envelope having on the outside the motto of the paper and within the name and address of the author. They must reach the secretary of the college, 19, South 22nd-street, Philadelphia, before May 1st.

The Alvarenga prize for 1936 has been awarded to Dr. Harry Eagle, passed assistant surgeon, United States Public Health Service, who is at present stationed at the Johns Hopkins Hospital. Dr. Eagle's essay dealt with the present status of the blood coagulation problem.