# PRINCIPLES OF MEDICAL STATISTICS

## VIII—FURTHER PROBLEMS OF SAMPLING : $\chi^2$

IN the previous section attention was devoted to the appropriate test of the "significance" between two proportions. Occasions frequently arise, however, when we need to compare the characteristics of more than two groups. For instance instead of comparing the proportion of vaccinated persons who are attacked with small-pox with the proportion of unvaccinated persons who are attacked, we wish to see whether the proportions attacked vary with the duration of time that has elapsed since vaccination. We shall then have a series of differences between groups to interpret. For the comparison of such distributions the $\chi^2$ test has been developed (originally by Karl Pearson).

### Interpretation of a Series of Proportions

The statistical procedure may best be discussed by means of a concrete example. Table VII shows the distribution of intelligence quotients in a group of nearly a thousand children and the number in each intelligence group that were clinically assessed as having normal or subnormal nutrition.

TABLE VII—*Intelligence and Nutritional State* *

| — | Intelligence quotients— | | | | |
|---|---|---|---|---|---|
| | Under 80 %. | 80–89 %. | 90–99 %. | 100 % and over. | Total. |
| (i) Number of children with satisfactory nutrition .. | 245 | 228 | 177 | 219 | 869 |
| (ii) Number of children with unsatisfactory nutrition .. | 31 | 27 | 13 | 10 | 81 |
| (iii) Total number of children observed .. .. .. | 276 | 255 | 190 | 229 | 950 |
| (iv) Percentage in each intelligence group that had unsatisfactory nutrition.. | 11·2 | 10·6 | 6·8 | 4·4 | 8·5 |

* From the Relation between Health and Intelligence in School Children by N. J. England (1936) *J. Hyg., Camb.* **36**, 74.

The percentages in the last line of the table show that amongst the more intelligent children there were proportionately fewer instances of subnormal nutrition, and the regular progression of the percentages, from 11·2 to 4·4 as the intelligence quotient rises, suggests that this relationship is unlikely to have arisen merely by chance. It is clear, however, that if another sample of 950 children were taken at random from the same universe of children we should not necessarily find in the different intelligence groups exactly 11·2, 10·6, 6·8, and 4·4 per cent. with unsatisfactory nutrition. Each of these percentages is certain to vary from one sample to another, and the smaller the sample the more, as has been previously shown, they are likely to vary. The question at issue then becomes this: is it likely that the magnitude of the differences between these percentages, and also their orderly progression, could arise merely by chance in taking samples of the size given in line three? Is it likely, in other words, that if we had observed the whole universe of children from which the sample was taken the percentage with unsatisfactory nutrition would be just the same in each intelligence group? To answer that question the assumption is made that the percentage with unsatisfactory nutrition ought to be identical in each of these groups—i.e., that intelligence and nutrition are unrelated. We then seek to determine whether that assumption is a reasonable one by measuring whether the differences actually observed from the uniform figure might frequently or only infrequently arise by chance in taking samples of the recorded size. If we find that the departure from uniformity is of the order that might frequently arise by chance, then we must conclude that the varying percentage of children with unsatisfactory nutrition in the different intelligence groups *suggests* that intelligence and nutrition are associated, but that these differences are not more than might have arisen by chance and might vanish if we took another sample of children. We must, therefore, be cautious in drawing deductions from them. If, on the other hand, we find that the differences from our assumed uniformity are such as would only arise by chance very infrequently, then we may reject our original hypothesis that each of the groups ought to show the same percentage. For if that hypothesis were true an unlikely event has occurred, and it is reasonable to reject the unlikely event and say that we think the differences observed are real, in the sense that they would not be likely to disappear (though they might be modified or increased) if we took another sample of children of equal size.

#### THE TECHNIQUE INVOLVED

The statistical technique, therefore, involves—

(1) Calculating how many children in each intelligence group would fall in the satisfactory nutrition category, and how many in the unsatisfactory nutrition category, on the assumption that intelligence and nutrition are unrelated, so that the proportions in the nutrition categories ought to be the same in each group.

(2) Calculating the differences between these numbers expected on the hypothesis of no relationship and the numbers actually observed.

(3) Calculating whether these differences are of a magnitude likely or unlikely to be due to chance.

The first step is to determine what is the uniform percentage with unsatisfactory nutrition that we should expect to observe in each intelligence group if intelligence and nutrition are unassociated. Clearly that figure must be the percentage of malnourished children in the universe from which our sample was taken; that is the figure we should, apart from sampling errors, expect to obtain in each intelligence group, if this characteristic is not associated with nutrition. We do not know that figure but as an estimate of it we may take the proportion of malnourished children in our sample as a whole—namely, 8·5 per cent. Our assumption then is that the proportion of malnourished children ought to be 8·5 per cent. in each intelligence group. If that assumption is true should we be likely to observe, in a sample of the size given, percentages of 11·2, 10·6, 6·8, and 4·4 merely by chance? We must first calculate the number of children expected in each intelligence group on this hypothesis that 8·5 per cent. of them ought to have belonged to the malnourished category and 91·5 per cent. to the satisfactorily nourished category. These figures are given in italics in Table VIII, being calculated by simple proportion—e.g., there were 276 children whose intelligence quotient was below 80 per cent.; we expect 8·5 per cent. of these 276, or 23, to be malnourished and 91·5 per cent., or the remaining 253, to be satisfactorily nourished.

TABLE VIII—*Calculation of* $\chi^2$

| — | Intelligence quotients— Under 80 %. | 80–89 %. | 90–99 %. | 100 % and over. | Total. |
|---|---|---|---|---|---|
| (i) Observed number with satisfactory nutrition .. | 245 | 228 | 177 | 219 | 869 |
| (ii) Expected number on hypothesis that nutrition and intelligence are unrelated.. | *253* | *233* | *174* | *209* | *869* |
| (iii) Difference, observed minus expected .. .. .. | −8 | −5 | +3 | +10 | .. |
| (iv) (Difference)$^2$ .. .. | 64 | 25 | 9 | 100 | .. |
| (v) (Difference)$^2$ ÷ Expected number .. .. .. | 0·25 | 0·11 | 0·05 | 0·48 | .. |
| (i) Observed number with unsatisfactory nutrition .. | 31 | 27 | 13 | 10 | 81 |
| (ii) Expected number on hypothesis that nutrition and intelligence are unrelated.. | *23* | *22* | *16* | *20* | *81* |
| (iii) Difference, observed minus expected .. .. .. | +8 | +5 | −3 | −10 | .. |
| (iv) (Difference)$^2$ .. .. | 64 | 25 | 9 | 100 | .. |
| (v) (Difference)$^2$ ÷ Expected number .. .. .. | 2·78 | 1·14 | 0·56 | 5·00 | .. |
| *Total*— | | | | | |
| Observed.. .. .. | 276 | 255 | 190 | 229 | 950 |
| Expected.. .. .. | *276* | *255* | *190* | *229* | *950* |

The next step is to calculate the differences between the observed and expected entries. These, given in the third line of each half of the table, show that there are in fact more children of low intelligence with subnormal nutrition than would be expected on our hypothesis of equality and fewer with subnormal nutrition in the higher intelligence groups than would be expected on that hypothesis—e.g., of children in the under 80 per cent. intelligence group 31 had unsatisfactory nutrition whereas we expected only 23; of children in the over 100 per cent. intelligence group there were only 10 with unsatisfactory nutrition whereas we expected 20. These differences if added together according to their sign must come to zero. To get rid of this difficulty of sign, and for mathematical reasons underlying the test, each difference is squared (as shown in line iv). Finally the following point must be observed. If the number expected in a group is, say, 75, and the number actually observed is 100, the difference of 25 is clearly a relatively large and important one, since it amounts to one-third of the expected number. On the other hand, if the number expected in the group is 750 and the number actually observed is 775, the difference of 25 is a relatively small and unimportant one, since it amounts to only 3·3 per cent. of the expected number. To assess the importance of the difference, its square is therefore divided by the number expected in that category (in the first case above this gives a value of $(25)^2/75$ or 8·3, in the second case $(25)^2/750$ or only 0·83). The sum of these values is known as $\chi^2$. In this table it equals $0{\cdot}25 + 0{\cdot}11 + 0{\cdot}05 + 0{\cdot}48 + 2{\cdot}78 + 1{\cdot}14 + 0{\cdot}56 + 5{\cdot}00 = 10{\cdot}37$.

In other words $\chi^2$ is equal to the sum of all the values of

$$\frac{(\text{Observed number} - \text{Expected number})^2}{\text{Expected number}}.$$

### THE INTERPRETATION OF THE $\chi^2$ VALUE

$\chi^2$, it will be observed, will be zero if the observations and expectations are identical; in all other cases it must be a positive value, and the larger the relative differences between the observed and expected values the larger must the value of $\chi^2$ be. Also the more sub-groups there are in the table the larger $\chi^2$ may become, since each sub-group contributes a quota to the sum. Therefore in interpreting the value of $\chi^2$ account must be taken both of the value itself and of the number of sub-groups contributing to it. In fact, although its mathematical foundation is complex, the interpretation of $\chi^2$ is simple by means of published tables—e.g., Table III in Fisher's "Statistical Methods for Research Workers" (Sixth edition. Edinburgh and London: Oliver and Boyd. 1936. 15*s*.), of which an illustrative extract is given at the end of this article. These tables show whether the various differences found between the observed and expected values, as summed up in $\chi^2$, are sufficiently large to be opposed to our hypothesis that there ought to be a uniform proportion of malnourished children in each intelligence group. For instance in the present example $\chi^2 = 10{\cdot}37$ and the number of independent sub-groups is 3 (this number will be referred to in a moment); for these values Fisher's table gives a probability of somewhat less than 0·02—see table below. (The exact probability can be obtained from larger tables but is not always important, the main question being whether this value of $\chi^2$ is one which is likely or unlikely to have arisen by chance.) The meaning of this probability is as follows. If our hypothesis that we ought to have observed the *same* percentage of malnourished children in each of the intelligence groups is true —i.e., we are sampling a universe in which malnutrition and intelligence are not associated—then in the different intelligence groups of the size shown we might have reached merely by chance the differing malnourished proportions actually observed (or even larger differences from the uniformity expected) about once in fifty times (0·02 equals 2 in 100). In other words, if we had 50 separate samples of 950 children with this distribution of intelligence and ought to observe within each the same percentage of malnourished children in the intelligence groups, then in approximately only one of these 50 samples should we expect the actual proportions of malnourished children in the intelligence groups to differ between one another, by chance, by as much as (or more than) the proportions we have observed here. Once in fifty we may take to be an unlikely event and may therefore conclude that our original testing-hypothesis of equality is wrong, and that it is more likely that the intelligence groups really differ in the incidence of malnourishment. If on the other hand our $\chi^2$ value had turned out to be 4·64, then the probability figure would have been 0·2. In other words, once in five trials we might reach differences of the observed magnitude between observation and expectation merely by chance. Once in five trials is a relatively frequent event, and we should have to conclude that the differences between the proportions of malnourished children in the intelligence groups might easily have arisen by chance, and we can draw from them no more than very tentative conclusions.

As with all tests of "significance" it must be observed that the final conclusion turns upon probabilities. There is no point at which we can say the differences could not have arisen by chance or that they must have arisen by chance. In the former case we say chance is an unlikely cause, in the latter that it could easily be the cause. It may not have been the cause in the latter—real differences may exist—but our data are insufficient for us to rule out chance as a valid hypothesis.

### HOW TO USE FISHER'S $\chi^2$ TABLE

Finally we must consider the number of sub-groups, or "cells," in the table contributing to $\chi^2$.

In Table VII (or VIII which shows the calculations on the data of Table VII) there are 8 such values contributing to the sum. In finding $P$, the probability, from Fisher's table we took the number, $n$, to be only 3. The rule to be followed in tables of the type taken above as an example is that $n$ is the number of cells which can be filled up independently of the totals in the margins at the side and bottom of the table. In Table VII the expected numbers of malnourished children in three of the intelligence groups must be found, by simple proportion, by applying 8·5 per cent. to the total number of children in each of those three groups. Having found those three values all the other expected values can, however, be found by simple subtraction, since the total expected values vertically and horizontally must be the same as the observed values—e.g., having found the expected values of 23, 22, and 16=61 in three intelligence groups, the expected number in the fourth group must be 20, since the total must be 81 (8·5 per cent. of 229 is nearly 20). Similarly if 23 malnourished children are expected in the intelligence group of "under 80 per cent." then the well-nourished in that intelligence group must be 253, for the total children of that type is 276. Hence only 3 values in this table need to be found independently by applying proportions, and the remainder can be found by subtraction.

In any practical problem the number of expected values which need to be found independently by simple proportion can easily be ascertained on inspection of the table; alternatively they can be found by means of the rule $n=(c-1)\ (r-1)$, where $c$ is the number of columns, excluding the "total" columns, and $r$ the number of rows in the table—e.g., in Table VII, $n=(4-1)\ (2-1)=3$. It is with this $n$ that Fisher's table must be consulted.

It may be pointed out that in the above example the expected numbers were taken to the nearest whole number for the sake of simplification, but it would have been more accurate to retain one decimal and in practice this should be done. A warning may be added that if the expected number in any cell is less than 5 the $\chi^2$ value is liable to be exaggerated and the probability derived from it may be inaccurate. In such cases adjacent columns should, if possible, be amalgamated to give larger expected numbers.

The following table is a short extract, given for illustrative purposes, from Fisher's table. In the latter the $\chi^2$ values are tabulated for each value of $n$ from 1 to 30 and for 13 probability values.

*Table of* $\chi^2$

| $n$ | $P$=0·90 | 0·70 | 0·50 | 0·20 | 0·05 | 0·02 | 0·01 |
|---|---|---|---|---|---|---|---|
| 1 | 0·0158 | 0·148 | 0·455 | 1·642 | 3·841 | 5·412 | 6·635 |
| 3 | 0·584 | 1·424 | 2·366 | 4·642 | 7·815 | 9·837 | 11·341 |
| 10 | 4·865 | 7·267 | 9·342 | 13·442 | 18·307 | 21·161 | 23·209 |
| 15 | 8·547 | 11·721 | 14·339 | 19·311 | 24·996 | 28·259 | 30·578 |

$n$ is the number of cells that needed to be filled independently (as defined above) in the table of observations under study. In the centre of Fisher's table are set out the values of $\chi^2$. At the top is the probability arising from the observed value of $\chi^2$ for a given value of $n$. For instance, in the example taken above $n$ was 3 and $\chi^2$ was 10·37. Glancing along the $\chi^2$ values against $n=3$, we see that a $\chi^2$ of 9·837 gives a $P$, or probability, of 0·02 and one of 11·341 gives a $P$ of 0·01. Our value of 10·37 must therefore give a $P$ of somewhat less than 0·02. If on the other hand $\chi^2$ had been 4·642 we see that the $P$ against that value is only 0·2. If $n$ had been 10 and the value of $\chi^2$ found were 18·307, then $P$ would be 0·05. As a conventional level a $P$ of 0·05 is usually taken as "significant"—i.e., the proportions observed in the different groups under examination would only show by chance a departure from the assumed uniformity as great as (or greater than) that actually observed once in twenty times. But, as pointed out in the last section, the worker is entitled to take any level of "significance" he wishes so long as he makes his standard clear.

## Summary

In practical statistics occasions constantly arise on which we wish to test whether persons who are characterised in some particular way are also differentiated in some second way—e.g., whether persons of different hair colour have also a different incidence of, say, tuberculosis. Suppose we take a random sample of the population, divide them into groups according to their hair colour, and then compute for each of these groups the percentage of persons with active tuberculosis. These percentages will probably not be identical even if hair colour and incidence of tuberculosis are *not* associated. Owing to chance they will vary between themselves. We need a measure to show whether such observed differences in incidence are in fact likely or unlikely to be the result of chance. To this type of problem the $\chi^2$ test is applicable.

A. B. H.

## INFECTIOUS DISEASE

### IN ENGLAND AND WALES DURING THE WEEK ENDED FEB. 6TH, 1937

*Notifications.*—The following cases of infectious disease were notified during the week: Small-pox, 0; scarlet fever, 1666; diphtheria, 1253; enteric fever, 61; pneumonia (primary or influenzal), 3140; puerperal fever, 59; puerperal pyrexia, 148; cerebro-spinal fever, 37; acute poliomyelitis, 3; encephalitis lethargica, 6; dysentery, 30; ophthalmia neonatorum, 78. No case of cholera, plague, or typhus fever was notified during the week.

Notifications of acute pneumonia in England and Wales were 95 per cent. above expectation, in London itself 17 per cent. below.

The number of cases in the Infectious Hospitals of the London County Council on Feb. 12th was 3437, which included: Scarlet fever, 809; diphtheria, 1006; measles, 22; whooping-cough, 597; puerperal fever, 20 mothers (plus 12 babies); encephalitis lethargica, 284; poliomyelitis, 3. At St. Margaret's Hospital there were 17 babies (plus 7 mothers) with ophthalmia neonatorum.

*Deaths.*—In 122 great towns, including London, there was no death from small-pox, 4 (1) from enteric fever, 3 (0) from measles, 5 (1) from scarlet fever, 40 (4) from whooping-cough, 52 (5) from diphtheria, 37 (13) from diarrhœa and enteritis under two years, and 976 (101) from influenza. The figures in parentheses are those for London itself.

Of the 788 deaths from influenza in the great towns of England and Wales outside Greater London the largest totals were reported from Manchester 51, Birmingham 50, Sheffield 37, Liverpool 35, Leeds 20, Plymouth 15, Derby 12, Bristol 11, Ipswich 10. Of these only Birmingham, Sheffield, and Leeds showed a drop from the previous week. The corresponding figures for Scottish and Irish towns were Edinburgh 37, Glasgow 39, Dublin 43, Belfast 18.

Darlington reported 1 death from enteric fever, Liverpool the other 2. Bolton had 4 deaths from whooping-cough, Middlesbrough 3, Willesden, Wood Green, and Manchester each 2. Fatal cases of diphtheria were reported from 24 great towns, Birmingham 7, Liverpool 5, Hull and Manchester each 4, Swansea 3.

The number of stillbirths notified during the week was 292 (corresponding to a rate of 42 per thousand total births), including 39 in London.