*J. roy. Army med. Cps.* 1976. **122,** 41-46

# MILITARY MEDICAL PRACTICE

## BASIC NOTES ON STATISTICAL METHODS IN MEDICINE—3

Major D. A. MOORE, M.B., B.S., D.P.H., D.I.H., D.T.M.&H., R.A.M.C.

(*continued*)

### $\chi^2$ test

Often in medicine, as a result of observation of our patients or populations, we find ourselves asking this sort of question: " Is there a link between diet and gastric carcinoma?"; " What is the relationship between residence in high rise flats and the incidence of upper respiratory tract infection in children?"; " What effect, if any, on speech development in children does admission to hospital have?", and so on. We are in fact attempting to demonstrate the presence of an *association* between two apparently independent factors.

The statistical test used to establish such associations, subject to the usual rules of probability, is the $\chi^2$ test. The working of the test is best illustrated by means of an example but before we get that far, two important points should be noted.

First: the $\chi^2$ test must be calculated with " real figures ", that is, it will not work using percentages or rates.

Second: although the test will show the existence of an association within certain bounds of probability, it gives no guide to the true *strength* or *degree* of association present.

*Example*

**Table I**

**Haemoglobin estimation of 1,000 parous women.**
**Relationship of parity with anaemia**

| | Part I | | | Part II | |
|---|---|---|---|---|---|
| | Observed | | | Expected | |
| Parity | <13 g/100 ml | >13 g/100 ml | Totals | <13 g/100 ml | >13 g/100 ml |
| 1 | 44 | 220 | 264 | $\frac{255}{1000} \times 264 = 67.32$ | $\frac{745}{1000} \times 264 = 196.68$ |
| 2 | 86 | 267 | 353 | $\frac{255}{1000} \times 353 = 90.02$ | $\frac{745}{1000} \times 353 = 262.99$ |
| 3 | 60 | 122 | 182 | $\frac{255}{1000} \times 182 = 46.41$ | $\frac{745}{1000} \times 182 = 135.59$ |
| 4 | 27 | 83 | 110 | $\frac{255}{1000} \times 110 = 28.05$ | $\frac{745}{1000} \times 110 = 81.95$ |
| 5 | 38 | 53 | 91 | $\frac{255}{1000} \times 91 = 23.21$ | $\frac{745}{1000} \times 91 = 67.80$ |
| Totals | 255 | 745 | 1000 | | |

A randomly selected group of 1,000 parous women drawn from Social Classes IV and V (according to their husband's occupation) had their haemoglobin estimated. With a criterion of normality of not less than 13 g/100 ml an attempt was made to relate parity with anaemia so defined. The results of the investigation are in Part I of Table I.

The expected figures in Part II of the table are obtained, as can be seen, by simple proportion from the observed results and are the expected figures assuming that there is no real association between parity and haemoglobin levels. In other words; we have assumed a null hypothesis.

The next step is to calculate the difference observed—expected. These can be seen in Table II. It can be seen that the sum of the columns and rows under " Totals " in the

**Table II**

**Haemoglobin estimation of 1,000 parous women.**
**Calculation of the difference observed—expected in Table I**

| < 13 g/100 ml | | | > 13 g/100 ml | | | Totals | |
|---|---|---|---|---|---|---|---|
| | | | | | | < 13 g/100 ml | > 13 g/100 ml |
| Observed | | Expected | Observed | | Expected | Observed | Expected |
| 44 | — | 67.32 | 220 | — | 196.70 | —23.32 | 23.30 |
| 86 | — | 90.01 | 267 | — | 263.00 | — 4.01 | 4.00 |
| 60 | — | 46.40 | 122 | — | 135.60 | 13.60 | —13.60 |
| 27 | — | 28.05 | 83 | — | 81.96 | — 1.05 | 1.04 |
| 38 | — | 23.20 | 53 | — | 67.79 | 14.80 | —14.79 |

table is zero (allowing for a small discrepancy due to rounding up or down of decimal places). This is how it should be and is a useful check of arithmetic accuracy.

Now, the formula for the calculation of $\chi^2$ is:—$\chi^2 = \dfrac{\Sigma (O-E)^2}{E}$

So, substituting our calculated figure we get:— $\dfrac{-23.32^2}{67.32} + \dfrac{-4.01^2}{90.01} + \dfrac{13.60^2}{46.40}$

$$+ \frac{-1.05^2}{28.05} + \frac{14.8^2}{23.20} + \frac{23.30^2}{196.70} + \frac{4.00^2}{263.00} + \frac{-13.60^2}{135.60} + \frac{1.04^2}{81.96} + \frac{-14.79^2}{67.79} = 29.15$$

The corresponding probability to this value of $\chi^2$ must be looked for in the appropriate table using the degrees of freedom calculated from the formula (columns − 1) + (rows − 1). Therefore in our case this will be (2 − 1) + (5 − 1) = 5 degrees of freedom. Consulting the table we find (you must take this on trust) that with 5 degrees of freedom our calculated value of $\chi^2$ −29.15 indicates $p < 0.001$. We have thus " disproved " our null hypothesis and shown at odds of better than 999 : 1 that an association does exist between parity and haemoglobin levels in our group of women.

*The $\chi^2$ test of " goodness of fit "*

A very useful application of the $\chi^2$ test is to assess the " goodness of fit " of

observed results in the situation where it is possible to calculate theoretically expected results. An example will make this clearer.

In an experiment 1,000 observations were made to the nearest 0.1 cm. If there is no observer terminal digit bias it would be reasonable to calculate theoretically that there will be 100 observations for each possible terminal digit. Obviously there will be some variation due to the play of chance but can we dismiss the variations shown in Table III in the observed figures as chance?

**Table III**

**Example of an experiment of 1,000 observations made to the nearest 0.1 cm**

| Terminal digit | Number of observations | Expected number of observations | (O—E) | $\frac{(O—E)^2}{E}$ |
|---|---|---|---|---|
| .0 | 157 | 100 | 57 | 32.49 |
| .1 | 52 | 100 | —48 | 23.04 |
| .2 | 106 | 100 | 6 | 0.36 |
| .3 | 60 | 100 | —40 | 16.00 |
| .4 | 98 | 100 | — 2 | 0.04 |
| .5 | 189 | 100 | 89 | 79.21 |
| .6 | 107 | 100 | 7 | 0.49 |
| .7 | 61 | 100 | —39 | 15.21 |
| .8 | 113 | 100 | 13 | 1.69 |
| .9 | 57 | 100 | —43 | 18.49 |
| | 1000 | 1000 | | $187.02 = \Sigma \frac{(O—E)^2}{E}$ |

So, with 9 degrees of freedom (1 column−1) + (10 rows−1) and looking up the tables of $\chi^2$ values (again on trust) we find that $\chi^2 = 187.02$ corresponds to $p < 0.001$. This shows plainly that the variation we have observed would occur by chance less than 1 in 1000 times. It may be safely assumed that observer bias has been involved.

*Yates correction*

Where there is a small number of columns and rows in the $\chi^2$ test and certainly every time the test is done as a 2 x 2 table (i.e. with 4 cells), in order to avoid results leading to a false overstatement of probability, Yates correction is used. The correction is very simple and consists of the reduction by 0.5 of each expression (O−E).

## Correlation and regression

We have already seen in the calculation of $\chi^2$ a statistical demonstration of the existence of an association between sets of variables representing two independent factors. At the time it was noted that although the association can be established by the $\chi^2$ test the result gives no indication of its *degree*. The statistical techniques we are now to deal with can give a single figure expression of this degree of association (the correlation coefficient) and a prediction of the change in one characteristic that will follow unit change in the other (the regression coefficient).

As they will be given here both techniques depend upon the association under examination being approximately linear and the two sets of variables to be random.

*Correlation coefficient*

In order to express correlation mathematically the coefficient of correlation, r, is calculated. The value of r cannot exceed 1 or $-1$; both of these values implying perfect positive and negative functional relationships respectively. $r = 0$ implies a state of complete dissociation. It follows that the closer the value of r approaches 1 or $-1$ the stronger the correlation but the observer must beware of the situation where an apparent high degree of correlation is due to both factors under study being strongly correlated to a missed third factor or that both are varying with time.

A formula for the calculation of r is:—$r_{xy} = \dfrac{\Sigma (x-\bar{x}) \quad (y-\bar{y})}{\sqrt{\Sigma (x-\bar{x})^2 \, \Sigma (y-\bar{y})^2}}$ but for ease of calculation this expression is expanded so:— $r_{xy} = \dfrac{\Sigma xy - \dfrac{\Sigma x \Sigma y}{n}}{\sqrt{\Sigma x^2 - \dfrac{(\Sigma x)^2}{n} \cdot \Sigma y^2 - \dfrac{(\Sigma y)^2}{n}}}$

*Example*

A sample of 9 soldiers of varying age were examined to determine their upper limit of hearing in kilohertz (1000's cycles per second). The results are seen in the first two columns of Table IV. The question now is: What is the correlation between age and the upper limit of hearing—if any.

**Table IV**

**Sample of 9 soldiers varying age examined to determine their upper limit of hearing in kilohertz.**

| Age | Kilohertz | | | |
|---|---|---|---|---|
| x | y | $x^2$ | $y^2$ | xy |
| 41 | 13.5 | 1681 | 182.25 | 553.5 |
| 34 | 14.4 | 1156 | 207.36 | 489.6 |
| 17 | 16.3 | 289 | 265.69 | 277.1 |
| 19 | 17.0 | 361 | 289.00 | 323.0 |
| 25 | 15.6 | 625 | 243.36 | 390.0 |
| 37 | 14.7 | 1369 | 216.09 | 543.9 |
| 48 | 12.7 | 2304 | 161.29 | 609.6 |
| 17 | 17.4 | 289 | 302.76 | 295.8 |
| 22 | 15.9 | 484 | 252.81 | 349.8 |
| $\Sigma x = 260$ | $\Sigma y = 137.5$ | $\Sigma x^2 = 8558$ | $\Sigma y^2 = 2120.61$ | $\Sigma yx = 3832.3$ |

$n = 9$

Therefore, substituting in the expanded form of the equation in Table IV we get:

$$\frac{3832.3 - \dfrac{260 \times 137.5}{9}}{\sqrt{8558 - \dfrac{(260)^2}{9} \times 2120.61 - \dfrac{(137.5)^2}{9}}} = \frac{-139.9}{144.41} \quad \therefore r_{xy} = -0.97$$

Thus, we have demonstrated a near perfect (because the example was so "fixed") negative correlation between age and the upper limit of hearing.

*The standard error of the correlation coefficient*

The correlation coefficient only truly represents the association existing in the sample of the universe from which it has been calculated. If further samples were taken the correlation coefficients calculated from them would tend to fall in an approximately normal distribution around the true correlation coefficient of the universe. The standard error of the sample correlation coefficient can be calculated from:

$\frac{1}{\sqrt{n-1}}$ so, using our sample above we would calculate $SE_r = \frac{1}{\sqrt{9-1}} = 0.354$

Twice this calculated $SE_r$ is 0.708. Our calculated $r_{xy}$ at 0.97 is therefore more than $+ 2\ SE_r$ from the expected correlation coefficient of 0 (assuming a null hypothesis) and is significant.

*Regression*

It has already been mentioned that in the situation where an association exists between the two sets of characteristics it is very useful to be able to predict the change in one characteristic consequent upon unit change in the others. To do this we must construct a line that interprets the trend of the association with minimal error. Figure 1 shows two such lines drawn with only one point shown.

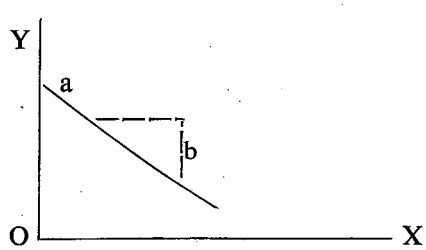


Fig. 1 (a)

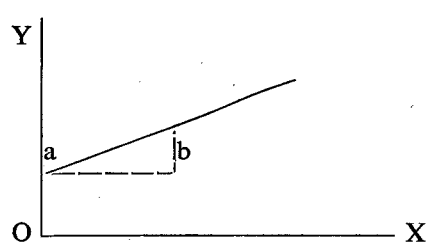


Fig. 1 (b)

Figure 1 (a) indicating a inverse relationship and Fig. 1 (b) a positive relationship. In these figures a is the *intercept* of the line on the y axis and represents the value of y when n = o and b is the *regression coefficient* of y on x. Of course another line could be drawn for the regression of x on y but normally y on x is more important (where y is the dependent, and x, the independent variable). These are the *lines of best fit* and are determined by the method of least squares: that is, to make the sum of the squares of the distances from the plotted points to the best line as small as possible. The regression equation for the regression of y on x is $y = a + b\,x$ where $b = \frac{\Sigma\,(x - \bar{x})\,(y - \bar{y})}{\Sigma\,(x - x)^2}$ or, expanded $\frac{\Sigma\,xy - \frac{\Sigma\,x\,\Sigma y}{n}}{\Sigma\,x^2 - \frac{(\Sigma\,x)^2}{n}}$ and $a = \bar{y} - b\,\bar{x}$

Therefore, using the figures from our example on correlation we can calculate:

$$b = \frac{3832.3 - \frac{260 \times 137.5}{9}}{8558 - \frac{(260)^2}{9}} = \frac{-139.9}{1046.9} = -0.134$$

and, $\Sigma y = 137.5$, $n = 9 \therefore \bar{y} = 15.28$. $\Sigma x = 260$, $n = 9 \therefore \bar{x} = 28.89 \therefore a = 15.28 - (-0.134 \times 28.89) = 19.15$.

Substituting for both a and b in the regression equation we get: $y = 19.15 - 0.134x$ thus for any value of x we can readily calculate y.

*Warning*

Beware the assumption that any *association* defined and quantified by the statistical methods described above means *causation*. This is not necessarily so as a moments thought will confirm. Secondly, beware extending the line beyond the limits of your data. Extrapolation is a useful technique in many fields but may lead to completely erroneous and sometimes ridiculous conclusions in clinical and epidemiological work.

## Conclusion

As was said in the beginning, these notes provide a bare basis of statistical methods useful in medicine. The reader is urged to "read around" each method so as to thoroughly understand the concept behind the method. Memorising of formulae and refurbishing of perhaps rusty algebraic and arithmetical knowledge may be useful for examination purposes but is of secondary importance in the practical application of statistical methods.

## Acknowledgement

I am indebted to Mr. Machin of the Royal Army Medical College for execution of the diagram in Part 1 of this article.