62

ON A PROPERTY CONNECTING THE CHI-SQUARE MEASURE OF DISCREPANCY WITH THE METHOD OF MAXIMUM LIKELIHOOD

Author's Note (CMS 9.94a) *

A more general and systematic exposition than had previously been attempted of the connection between the χ^2 measure of discrepancy between fitted expectations and observational frequencies, and the method of fitting by maximum likelihood.

Atti del Congresso Internazionale dei Matematici, Bologna, 6: 95-100, (1928).

R. A. FISHER (Rothamsted-Herts - Inghilterra)

Reproduced from Atti del Congresso Internazionale dei Matematici, Bologna, 6: 95-100 (1928)

ON A PROPERTY CONNECTING THE χ^2 MEASURE OF DISCREPANCY WITH THE METHOD OF MAXIMUM LIKELIHOOD

1. - Introductory. — The measure of discrepancy, χ^2, between observation and hypothesis conforms to its well known series of distributions only in the limit when the number of observations tends to infinity; and in the theory of finite samples it is not obvious that this measure has any unique merit. In the theory of large samples it has been shown (1) that the test of Goodness of Fit based upon it is valid only if all statistics used in estimating adjustable parameters satisfy the criterion of efficiency, and further that all efficient statistics tend in large samples to equivalence. Recently, however (2), in the detailed investigation of a simple sampling problem arising in genetics the writer found that χ^2 is specially related to a particular type of efficient solution, namely to that obtained by the method of maximum likelihood. In view of the theoretical and practical importance of solutions obtained by this method in the exact theory of finite samples, it is of interest to trace how general the observed relationship may be, and to ascertain its bearing upon the interpretation of χ^2 derived from finite samples.

2. - A particular example. — In the particular case examined four types of offspring may occur, the expectations from a sample of n being

$$\frac{n}{4}(2+\theta,\ 1-\theta,\ 1-\theta,\ \theta)$$

in which θ is an adjustable parameter depending on the linkage between the two genetic factors concerned; if

$$a_1,\ a_2,\ a_3,\ a_4$$

are the numbers observed in the four classes, it is evident that the two expressions

$$x = a_1 + a_2 - 3(a_3 + a_4)$$
$$y = a_1 + a_3 - 3(a_2 + a_4)$$

(1) R. A. FISHER (1924): *The conditions under which χ^2 measures the discrepancy between observation and hypothesis.* Journal of the Royal Statistical Society, LXXXVII, pp. 442-449.

(2) R. A. FISHER (1928): *Statistical Methods for Research Workers.* Edinburgh, Oliver & Boyd, 2nd Edition, XI + 269 p.

have each an expectation zero for all values of θ, and that the random sampling distribution of each may be derived from the binomial expansion

$$\left(\frac{3}{4}+\frac{1}{4}\right)^n$$

in entire independence of the value of θ.

In the comparison of four observed frequencies with a series of expectations amounting to the same total, three degrees of freedom will be available for discrepancies; if, however, the expectations have been calculated from the observations to which they are to be compared by the use of one adjustable parameter we may anticipate that the degrees of freedom left available for discrepancies will be reduced to two. We may identify these with x and y since these represent discrepancies which are not affected by modifications of the value of θ.

For any value of θ the value of χ^2 may be written

$$\chi^2=\frac{4}{n}\left\{\frac{a_1^2}{2+\theta}+\frac{a_2^2+a_3^2}{1-\theta}+\frac{a_4^2}{\theta}\right\}-n$$

which, for a given value of n, is a quadratic function of the frequencies; if, for our two chosen components x and y we form the quadratic expression

$$\frac{1}{1-\varrho^2}\left\{\frac{x^2}{\sigma_1^2}-\frac{2\varrho xy}{\sigma_1\sigma_2}+\frac{y^2}{\sigma_2^2}\right\} \tag{1}$$

and substitute for σ_1^2 and σ_2^2 the mean values of x^2 and y^2, and for $\varrho\sigma_1\sigma_2$ the mean value of xy, we obtain

$$Q^2=\frac{3}{8n(1-\theta)(1+2\theta)}\left\{x^2+y^2-\frac{2}{3}(4\theta-1)xy\right\}$$

which is also, for a given value of n, a quadratic function of the frequencies. On comparing the two expressions term by term it appears that

$$\chi^2-Q^2=\left\{\frac{a_1}{2+\theta}-\frac{a_2+a_3}{1-\theta}+\frac{a_4}{\theta}\right\}^2\Big/\frac{(1+2\theta)n}{2\theta(1-\theta)(2+\theta)}.$$

The value of χ^2-Q^2 is therefore always positive, except for the special value of θ for which

$$\frac{a_1}{2+\theta}-\frac{a_2+a_3}{1-\theta}+\frac{a_4}{\theta}=0$$

which is the equation for the estimate of θ provided by the method of maximum likelihood. For this method maximises

$$L=a_1\log(2+\theta)+(a_2+a_3)\log(1-\theta)+a_4\log\theta$$

whence

$$\frac{\partial L}{\partial\theta}=\frac{a_1}{2+\theta}-\frac{a_2+a_3}{1-\theta}+\frac{a_4}{\theta}.$$

Moreover

$$\frac{\partial^2 L}{\partial\theta^2}=-\left\{\frac{a_1}{(2+\theta)^2}+\frac{a_2+a_3}{(1-\theta)^2}+\frac{a_4}{\theta^2}\right\},$$

of which the mean value is

$$-\frac{(1+2\theta)n}{2\theta(1-\theta)(2+\theta)},$$

which, with changed sign, is the quotient in our expression for χ^2-Q^2; or, otherwise interpreted, the quantity of information relative to θ which the data contain.

The use of χ^2 from finite samples is therefore exactly equivalent to the use of the bivariate expression (1) when the method of maximum likelihood is employed. The difference χ^2-Q^2 may be regarded as that part of the discrepancy between observation and hypothesis which is due to imperfect methods in the estimation of θ. This part will be large even in large samples if inefficient methods are used; but with efficient methods other than the method of maximum likelihood it may be expected to be sufficiently small for sufficiently large samples.

3. - The statistic defined by an equation linear in the frequencies, which is also efficient. — The particular instance examined is special in that the frequencies are expressible as linear functions of the unknown parameter. The connection between χ^2 and the maximum likelihood solution does not, however, flow from this fact, but from the fact, which is true in general, that the equation of maximum likelihood is linear in the frequencies. The maximum likelihood equation may indeed be derived from the conditions that it shall be linear in the frequencies, and efficient for all values of θ.

Consider any statistic T defined as the relevant root of an equation of the form

$$X=S(ka)=0$$

in which a stands for the frequency in any class, k for a coefficient, which is to be a functions of θ, and S for summation over all classes. Then the sampling variance of T from large samples may be equated to the sampling variance of X, for a given value of θ, divided by the square of the mean value of $\partial X/\partial\theta$.

But, if we let p denote the probability of any class,

$$V(X)=nS(pk^2);$$

and the mean value of $\partial X/\partial\theta$ is given by,

$$\partial\overline{X}/\partial\theta=nS\left(p\frac{\partial k}{\partial\theta}\right).$$

If now the statistic is consistent

$$S(kp)=0$$

for all values of θ, and hence

$$S\left(p\frac{\partial k}{\partial\theta}\right)+S\left(k\frac{\partial p}{\partial\theta}\right)=0;$$

so we may write

$$V(T)=\frac{S(pk^2)}{nS^2\left(k\frac{\partial p}{\partial\theta}\right)}.$$

If we minimise this for variations of k it appears that

$$\frac{kp}{S(pk^2)}=\frac{\partial p/\partial\theta}{S(k\partial p/\partial\theta)}$$

for each class, or

$$ka\frac{1}{p}\frac{\partial p}{\partial\theta}.$$

Using these relations we find

$$V(T)=1\div nS\left\{\frac{1}{p}\left(\frac{\partial p}{\partial\theta}\right)^2\right\}$$

and the equation for T becomes

$$S\left(\frac{a}{p}\frac{\partial p}{\partial\theta}\right)=0$$

which is in fact the equation of maximum likelihood.

4. - The resolution of χ^2 into its components. — In general with any number of parameters θ_1, θ_2,...., θ_r to be estimated, the equations of maximum likelihood will be

$$S\left\{\frac{a}{p}\frac{\partial p}{\partial\theta_1}\right\}=0$$
$$\cdots\cdots\cdots$$
$$S\left\{\frac{a}{p}\frac{\partial p}{\partial\theta_r}\right\}=0.$$

Any linear function of the frequencies,

$$x_1=S(ka),$$

will have a mean value zero provided that

$$S(kp)=0;$$

and this mean will be stationary for variations of θ_1,...., θ_r provided the r conditions,

$$S\left(k\frac{\partial p}{\partial\theta}\right)=0, \tag{2}$$

are fulfilled. If the number of classes is s it will then be possible to form $s-r-1$ quantities x fulfilling these conditions, and such that for any two of them x_c and x_d the condition

$$S(pk_ck_d)=0 \tag{3}$$

shall also be satisfied.

Considering now the quantities

$$\xi = \frac{a - pn}{\sqrt{pn}}$$

as rectangular coordinates of a point in s dimensions. Since

$$x = S(k\xi\sqrt{pn})$$

it follows that

$$\frac{x}{\sqrt{S(npk^2)}} = S\left\{\frac{k}{\sqrt{S(k^2)}}\xi\right\}$$

the $s-r-1$ values of which represent the coordinates of the same point with respect to $s-r-1$ new axes, while the condition (3) shows that these new axes are mutually orthogonal and (2) shows that they lie in the surfaces

$$S\left\{\frac{a}{p}\frac{\partial p}{\partial\theta}\right\} = 0,$$

and evidently also in

$$S(a) = n.$$

But

$$S(\xi^2) = \chi^2$$

is the square of the distance of the sample point from the origin, and

$$S\left\{\frac{x^2}{S(npk^2)}\right\}$$

which we may now write, Q^2, must represent the square of its distance from the generalised surface

$$x_1 = x_2 = \ldots = x_{s-r-1} = 0$$

and these quantities will necessarily be the same if the remaining r coordinates, which can be built up as linear functions of

$$S\left(\frac{a}{p}\frac{\partial p}{\partial\theta}\right)$$

are all zero. In fact $\chi^2 - Q^2$ must always be a positive quantity expressible as a homogeneous quadratic function of the quantities

$$S\left(\frac{a}{p}\frac{\partial p}{\partial\theta}\right),$$

Q^2 being the sum of the squares of $s-r-1$ linear functions of the frequencies the mean value of each of which is stationary for variations of the parameters.

It follows that, in the theory of large samples, we may always speak of χ^2 as made up of two parts, one of which is due to errors of estimation, and vanishes when estimation is efficient, while the other is distributed in random samples as is the sum of the squares of $s-r-1$ quantities each normally distributed about zero with unit standard deviation. In the theory of finite samples the former

portion vanishes only when the adjustable parameters are estimated by the method of maximum likelihood, while the latter is the sum of the squares of quantities distributed, generally discontinuously, but each with unit standard deviation, and without mutual correlation for the particular set of parametric values arrived at by the method of maximum likelihood, and such that the mean of each is stationary at zero for variations of the parameters in the neighbourhood of these values.