34

The Conditions under which χ^2 Measures the Discrepancy between Observation and Hypothesis. *

* See Author's Note, Paper 31.

1. *Introductory.*

The interesting series of experiments on the distribution of χ^2, reported by Dr. Brownlee [1] affords an opportunity, not only of clearing up such doubts as still remain as to the necessity of entering Elderton's table with a corrected, or reduced, value of n', but also of bringing the conditions under which χ^2 affords a measure of goodness of fit into relation with the general theory of statistical estimation.

If x is the frequency of observations in any compartment of a frequency distribution, and if m is the expectation in that compartment, Pearson introduced ([2] 1900) the statistic

$$\chi^2 = \mathrm{S}\left\{\frac{(x-m)^2}{m}\right\}$$

as a measure of the discrepancy between observation and expectation. He succeeded in calculating the distribution of χ^2, when the values of x were the frequencies in random samples from an infinite population in which the frequencies were proportional to m, and showed that the distribution of χ^2 depends, in the limit when the samples are large, only on the number of classes, n', into which the samples were divided. In the same paper Pearson considered the possibility that when the values of m are not *a priori* expectations, but are themselves calculated from the observed values, the distribution of χ^2 might be modified by this procedure. He concluded that this was not so, and applied the test without correction to several examples in which the expectations in the several classes had been calculated from the distribution in the sample.

In 1922 [3] I was able to show, in the case of contingency tables, for which the margins of the expected table are reconstructed from those of the observed table, that the distribution of χ^2 was given exactly by Pearson's formula if we take for n', not the number of classes in the table, but one more than the number of degrees of freedom in which the expected table might differ from the values observed. The number of degrees of freedom is the number of frequencies which may be given arbitrary values without conflicting with the condition that the marginal totals are already specified. Thus, for a contingency table with two variates having r rows and c columns, n' should be equated, not to cr, but to $1 + (c-1)(r-1)$.

 Journal of the Royal Statistical Society, 87: 442-450, (1924).

In the same paper I expressed the opinion that the same reasoning should be applied to testing the goodness of fit of frequency curves, but that some discrepancy would arise if the grouping used in calculating the theoretical distribution were different from that employed in testing the goodness of fit.

Dr. Brownlee, after verifying the accuracy of the distribution with corrected n', in several instances, considers a coin-tossing experiment, in which he has obtained 32 samples in each of which 256 observations are distributed in the five classes, 4 heads, 3 heads, 2 heads, 1 head, no head. He finds that when the observations are compared to the theoretical distribution, given by the expansion of

$$(\tfrac{1}{2} + \tfrac{1}{2})^4,$$

the values of χ^2 obtained agree well with expectation for 4 degrees of freedom ($n' = 5$); also that when compared to the theoretical distribution

$$(p + q)^4, \qquad p + q = 1,$$

where p is obtained from the observations, by making χ^2 a minimum, the observed values of χ^2 agree with expectation for 3 degrees of freedom ($n' = 4$); but when the comparison is made with areas of a normal curve calculated by moments, using Sheppard's correction, in which calculation 2 degrees of freedom, representing the mean and standard deviation, are involved, the values of χ^2 do not at all conform to expectation for 2 degrees of freedom ($n' = 3$), but are distinctly higher. In fact, 5 out of the 32 values of χ^2 observed exceed 6, for which, when $n' = 3$, P is about 0·05; so that in 5 individual samples we should be led to conclude that the observation significantly contradicted the hypothesis, and in the aggregate of 32 samples contradicted it conclusively.

2. *Reasons for abnormal distribution of* χ^2.

This example illustrates so well the different reasons for which χ^2 may be abnormally distributed that these reasons may be considered in turn. χ^2 will be abnormally distributed—

(A.) *If the hypothesis tested is not in fact true.*

The distribution in the population from which Brownlee's samples were drawn appears to have been in the ratio 1, 4, 6, 4, 1. This ratio is not reproducible by dissecting a normal curve at equal intervals of the abscissa. In terms of the standard deviation the distance from the mean of the limits of the central group would be from the Kelley-Wood table [4], 0·488777; the next limits, representing the points beyond which the tail of the curve is one-sixteenth of the total area, would be at + 1·534121; while to include the

whole area the next limits are at infinity. For the hypothesis to be true, these values should be in the ratio 1 : 3 : 5. If we suppose the whole tail to be included in the extreme classes, we must still recognize that the central class, including three-eighths of a normal curve, stands on a smaller length of abscissa than the adjoining areas each including one-quarter of the curve. If, therefore, the values of χ^2 are found to be excessive, they are only performing their prime function in indicating the inexactitude of the hypothesis tested. In fact, with increasing samples, the values of χ^2 in such a case should increase without limit, and cannot be expected to be distributed as in Elderton's table.

Even if a hypothesis be true, the value of χ^2 obtained will not measure the goodness of fit, if the method of fitting employed is inadequate ; for in such a case the hypothesis to be tested is not adequately represented by the series of "expected" frequencies obtained.

In the first place the distribution of χ^2 will be abnormal—

(B.) *If the method of estimation employed is Inconsistent.*

A method of fitting fails to fulfil the criterion of consistency if, when applied to an infinite sample, *i.e.* to the whole population, it fails to reproduce the exact form of the population. Let us suppose that the frequencies in the five classes of the population were proportional to the areas of a normal curve divided at $\pm 0{\cdot}5$, $\pm 1{\cdot}5$; the fractions in the five classes would then be—

0·0668072, 0·2417303, 0·3829250, 0·2417303, 0·0668072 ;

from these the second moment, using Sheppard's correction, is 0·934585, whereas the true standard deviation is unity, equal to the grouping unit. Thus, using an indefinitely large sample, our method of estimation introduces an error of about 3 per cent. into our estimate of the standard deviation. Consequently from this cause also the values of χ^2 obtained will increase indefinitely as the size of the sample is increased.

(C.) *If the method of estimation employed is Inefficient.*

In any problem of estimation innumerable statistics, all functions of the observations, may be devised for the estimation of the required parameter, such that in all cases the error tends to zero as the size of the sample is increased. Such statistics all satisfy the criterion of consistency, and may all be termed consistent ; for large samples the sampling distribution of each of them may tend to normality, with variance inversely proportional to the number in the sample from which it was calculated, but the variance

of different statistics derived from samples of the same size will generally be different. We are thus led to specify out of the mass of consistent statistics, a group characterized by the fact that as the sample is increased their distribution curves tend to normality with the least possible variance. Such statistics satisfy the criterion of efficiency, and may be called efficient statistics. The efficiency of any other statistic is defined so as to be inversely proportional to its variance in large samples, the efficiency of efficient statistics being 100 per cent. For example, it may be proved that in estimating the mean of a normal distribution, no statistic can be more efficient than the mean of the sample, and that this has a variance of σ^2/n, where n is the number in the sample. The variance of the median obtained from a large sample tends to the value $\pi\sigma^2/2n$; consequently, while the efficiency of the mean is 100 per cent., that of the median is only 63·66 per cent.

3. *Properties of efficient statistics.*

I have shown elsewhere [5] that a statistic satisfying the criterion of efficiency may be found by the Method of Maximum Likelihood, and that its variance in random samples may be calculated directly from the frequency distribution of the population. If m is the expected frequency in any class, and x is the frequency observed, then any parameter θ, of which the series of values of m are functions, may be estimated by maximizing

$$\mathrm{L} = \mathrm{S}\,(x \log m)$$

for variations of θ ; this leads to an equation of the form

$$\mathrm{S}\left(\frac{x}{m}\,\frac{\partial m}{\partial \theta}\right) = 0,$$

from which, in any special case, θ may be obtained. The variance in random samples of the value so obtained is given by

$$-\frac{1}{\sigma^2} = \mathrm{S}\left(m\,\frac{\partial^2}{\partial \theta^2}\log m\right),$$

or, since $\mathrm{S}(m)$ is independent of θ, by

$$\frac{1}{\sigma^2} = \mathrm{S}\left\{\frac{1}{m}\left(\frac{\partial m}{\partial \theta}\right)^2\right\}.$$

Before connecting these properties with the distribution of χ^2, we may prove two elementary propositions respecting statistics which satisfy the criterion of efficiency.

1. The correlation between any two estimates of the same parameter which satisfy the criterion of efficiency tends to $+1$, as the sample is increased indefinitely.

Let the variance of each estimate tend to σ^2/n as the sample is increased, and let the correlation between the two estimates be r. Then the variance of their mean will be

$$\frac{\sigma^2}{n} \cdot \frac{1+r}{2}.$$

But, by hypothesis, this cannot be less than σ^2/n, therefore r cannot tend to a value less than unity.

2. The correlation between any estimate which satisfies the criterion of efficiency, and any other consistent estimate of the same parameter, tends for increasingly large samples to a limit, r, given by

$$r = \sqrt{\mathrm{E}},$$

where E is the efficiency of the second statistic.

Let A be the efficient statistic with variance σ^2/n, and B the inefficient statistic with variance $\sigma^2/\mathrm{E}n$; from them compound a new statistic C, such that

$$(1 + \mathrm{E} - 2r\sqrt{\mathrm{E}})\,\mathrm{C} = (1 - r\sqrt{\mathrm{E}})\,\mathrm{A} + (\mathrm{E} - r\sqrt{\mathrm{E}})\,\mathrm{B};$$

then the variance of C is

$$\frac{\sigma^2}{n} \cdot \frac{1-r^2}{1+\mathrm{E}-2r\sqrt{\mathrm{E}}} = \frac{\sigma^2}{n} \cdot \frac{1-r^2}{1-r^2+(r-\sqrt{\mathrm{E}})^2};$$

if therefore r does not tend to $\sqrt{\mathrm{E}}$ as the samples are increased the variance of C will tend to be less in the limit than the variance of A, which is impossible. Therefore, in the limit $r = \sqrt{\mathrm{E}}$.

An easy corollary is that the correlation of A with $(B - A)$ is zero, so that the deviations of B from the population value may be regarded as made up of two parts: one, an error of random sampling, properly so called, is the deviation of A from the population value; the other, distributed independently of the first, is the error of estimation by which the inferior estimate, B, differs from the superior estimate, A.

4. *The minimum of* χ^2.

All statistics which satisfy the criterion of efficiency being equivalent for large samples, it is important in connection with the χ^2 test that the method of minimizing χ^2 is one of them. For

$$\chi^2 = \mathrm{S}\left\{\frac{(x-m)^2}{m}\right\},$$

and if this is a minimum for variation of a parameter θ, we find

$$\mathrm{S}\left(\frac{x^2-m^2}{m^2} \cdot \frac{\partial m}{\partial \theta}\right) = 0.$$

Now, for large samples this equation tends to equivalence with that obtained by the Method of Maximum Likelihood, for the latter may be written

$$\mathrm{S}\left(\frac{x-m}{m}\cdot\frac{\partial m}{\partial\theta}\right)=0,$$

and, for large samples, the factor,

$$\frac{x+m}{m}$$

tends in all classes to the constant value, 2. Hence, all methods of fitting involving only efficient statistics tend, for large samples, to minimize χ^2.

5. *The effect on χ^2 of substituting for the true value of any parameter an estimate of it derived from sample.*

Let m stand for the frequency in any class expected from the true value of the parameter, and m' the corresponding frequency calculated from an efficient estimate.

Let

$$\chi^2=\mathrm{S}\left\{\frac{(x-m)^2}{m}\right\}$$

and

$$\chi'^2=\mathrm{S}\left\{\frac{(x-m')^2}{m'}\right\};$$

then

$$\chi^2-\chi'^2=\mathrm{S}\left\{\frac{(x-m)^2}{m}-\frac{(x-m')^2}{m'}\right\}=\mathrm{S}\left\{x^2\left(\frac{1}{m}-\frac{1}{m'}\right)\right\}.$$

The difference of the reciprocals of m and m' will depend on the difference $\delta\theta$ between the true value of the parameter and its value derived from the sample. Since $\delta\theta$ decreases proportionately to $n^{-\frac{1}{2}}$, as the size of the sample is increased, we shall expand the above expression in powers of $\delta\theta$, noticing that since both x and m increase proportionately to n, we shall have to carry the expansion as far as the term in $(\delta\theta)^2$, while in that term factors which tend to unity with increasing sample, such as x/m', may be omitted. Now

$$\frac{1}{m}-\frac{1}{m'}=-\frac{1}{m'^2}\frac{\partial m'}{\partial\theta}\delta\theta+\left\{\frac{2}{m'^3}\left(\frac{\partial m'}{\partial\theta}\right)^2-\frac{1}{m'^2}\frac{\partial^2 m'}{\partial\theta^2}\right\}\frac{(\delta\theta)^2}{2};$$

but, since χ^2 has been made a minimum,

$$\mathrm{S}\left(\frac{x^2}{m'^2}\frac{\partial m'}{\partial\theta}\right)=0.$$

Hence

$$\chi^2 - \chi'^2 = \frac{(\delta\theta)^2}{2} \cdot S\left\{ \frac{2}{m'}\left(\frac{\partial m'}{\partial \theta}\right)^2 - \frac{\partial^2 m'}{\partial \theta^2} \right\}$$

$$= (\delta\theta)^2 \cdot S\left\{ \frac{1}{m'}\left(\frac{\partial m'}{\partial \theta}\right)^2 \right\}.$$

Moreover, for any efficient statistic,

$$\frac{1}{\sigma^2} = S\left\{ \frac{1}{m}\left(\frac{\partial m'}{\partial \theta}\right)^2 \right\};$$

consequently the amount, by which χ^2 is reduced, is the square of a quantity normally distributed with unit standard deviation. The substitution thus diminishes the average value of χ^2 by unity, and this alone shows that if χ^2 is still distributed in the type III distribution given by Elderton's table, the value of n' with which the table is entered must be reduced by unity. It is, however, apparent, since θ has been found by a process equivalent to making χ'^2 a minimum, that χ'^2 is distributed independently of the additional square, $(\delta\theta)^2/\sigma^2$, and since χ^2 is distributed as is the sum of the squares of a number of quantities distributed normally and independently each with unit standard deviation, it is necessary that χ'^2 should be distributed as in the sum of the squares of a number smaller by unity of such quantities, and consequently the type III distribution is always reproduced.

If, however, χ_1^2 is the value obtained by using an inefficient statistic of efficiency E, then we find as above

$$\chi_1^2 - \chi'^2 = \frac{(\delta\theta)^2}{\sigma^2},$$

where σ^2 is the variance of an efficient statistic, and $\delta\theta$ is the error of estimation by which the inefficient statistic differs from the efficient one. The mean value of $(\delta\theta)^2$ is

$$\sigma^2\left(\frac{1}{E} - 1\right);$$

consequently the mean value of χ_1^2 may be found from that of χ^2 by subtracting $2 - \frac{1}{E}$. In this case, however, the distribution is not the Type III characteristic of χ^2. It will be noticed that with efficiencies below 50 per cent., the mean value of χ^2 is less than that of χ_1^2, so that the reconstructed population is generally less like the sample than is the population from which the sample was drawn.

The effect of using statistics, therefore, which are inconsistent is to make the value of χ^2 increase indefinitely with the size of the

sample; consistent statistics which are somewhat inefficient disturb the distribution by altering the mean, and in other ways. When such are used the value of χ^2 obtained does not measure merely the deviations of observation from hypothesis, but includes also deviations due to errors in the estimation of the parameters. Consequently such values should not be entered in Elderton's table in testing goodness of fit, but, if such tests are intended, the small corrections should be applied by which efficient statistics may be obtained.

The cases in which Dr. Brownlee's experiments have verified the theoretical distribution of χ^2 have all been obtained, actually or by approximation, by making χ^2 a minimum. The theoretical distribution would equally have appeared if any other efficient method had been used. For example, the five frequencies, α, β, γ, δ, ε, might have been fitted with the binomial distribution

$$256\ \{(\tfrac{1}{2} + \eta) + (\tfrac{1}{2} - \eta)\}^4$$

by taking

$$1024\eta = 2(\alpha - \varepsilon) + (\beta - \delta).$$

It is not necessary for our purpose to push refinement in methods of fitting beyond the requirement that all statistics used should be fully efficient, for the χ^2 distribution is in any case only exact when the sample is increased without limit, and in these circumstances all efficient statistics are equivalent. Only in an enquiry into the accuracy of the χ^2 distribution for small samples would such further refinements be required, and it is by no means obvious with small samples (i) that the method of minimizing χ^2 possesses any advantage over other efficient methods, or (ii) that the form of χ^2, without modification, provides the ideal measure of discrepancy for small samples. The method of maximum likelihood, for example [5], minimizes the quantity

$$\begin{aligned} \mathrm{L} &= \mathrm{S}\left(x \log \frac{x}{m}\right) \\ &= \mathrm{S}\left\{\tfrac{1}{2}\frac{(x-m)^2}{m} - \tfrac{1}{6}\frac{(x-m)^3}{m^2} + \tfrac{1}{12}\frac{(x-m)^4}{m^3} - \ldots\right\} \end{aligned}$$

of which $\frac{1}{2}\chi^2$ is the limit as the samples are indefinitely increased.

References.

1. J. Brownlee (1924).—"Some Experiments to Test the Theory of Goodness of Fit." *J.R.S.S.*, Vol. LXXXVII, pp. 76–82.

2. K. Pearson (1900).—" On the Criterion that a given System, of Deviations from the Probable in the case of a Correlated System of Variables is such that it can be reasonably supposed to have arisen from random sampling." *Phil. Mag.*, Series 5, Vol. L, pp. 157–175.

3. R. A. Fisher (1922).—" On the Interpretation of χ^2 from Contingency Tables, and the calculation of *P*." *J.R.S.S.*, Vol. LXXXV, pp. 87–94.

4. T. L. Kelley (1923).— " Statistical Method," 390 pp. Macmillan Co., New York.

5. R. A. Fisher (1921).—" On the Mathematical Foundations of Theoretical Statistics." *Phil. Trans.*, A, Vol. 222, pp. 309–368.