31

STATISTICAL TESTS OF AGREEMENT BETWEEN OBSERVATION AND HYPOTHESIS

Author's Note (CMS 7.a)

Papers 31 & 34 are attempts to reconcile, with the aid of the new concept of degrees of freedom, the discrepant and anomalous results observed by different authors, in the first case when confronted by data of a fourfold table, and in the second with distributions requiring the fitting of parameters. The types of confusion which had arisen are of some historical interest.

Economica, 3: 139-147, (1923).

Statistical Tests of Agreement between Observation and Hypothesis

By R. A. Fisher.

1. Introductory.

In a recent number (January, 1922) of the *Journal of the Royal Statistical Society* (Ref. 1) I put forward a proof that the distribution of χ^2, the Pearsonian test of goodness of fit, is not known merely from the number of frequency classes. In cases where the population, with which the sample is compared in calculating χ^2 has been itself reconstructed from the sample, we must also take account of the number of degrees of freedom absorbed in this process of reconstruction. The two cases of widest application were (i) contingency tables in which the population is reconstructed by assigning to the margins the frequencies observed in the sample, and (ii) frequency curves constructed to agree with the sample in respect of one or more moments. The common sense of this correction lies in the fact that when the population with which the sample is compared has been artificially identified with the sample in certain respects, such as the marginal frequencies, or the moments, we shall evidently make an exaggerated estimate of the closeness of agreement between sample and population, if we regard the sample as an unselected sample of a population known *à priori*. It was possible to show that the distribution was in fact that which arises when from any population a large number of samples are taken, and only those samples chosen which agree with the population in (say) the marginal frequencies; these samples compared to the true population will give values of χ^2 distributed in the same manner as in the practical case in which we compare any sample with a population artificially constructed from it. In both these cases the value of n' with which Elderton's Table should be entered is found by adding unity to the number of degrees of freedom in which the sample and the population are free to differ.

The following scheme shows the various forms of sampling which have been considered in this discussion, and which have not always

	Random Sample.	Selected Sample.
True population	A.—No correction needed.	B.—Correction needed.
Reconstructed population	C.—Correction needed.	

been clearly distinguished. A is the type of sampling considered by Professor Pearson in his memoir of 1900 (2); when the distribution of a random sample is given by the frequencies $x_1 \ldots x_n$, while the distribution of the true population from which the sample was drawn is given by $m_1 \ldots m_n$, when $S(x)=S(m)$, then Pearson showed that

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

was distributed in the Type III distribution made available by Elderton's Table. The true distribution of χ^2 for any finite value of $S(x)$ is, of course, discontinuous, but when the frequencies in all the cells are fairly large, the distribution of χ^2 is nearly independent of $S(x)$, and tends to that of the Type III curve. The distribution still depends on the number of cells, and this number is equated to n^1 in entering Elderton's Table.

Although the validity of Pearson's method of testing goodness of fit (in case A) is not universally accepted, I know of no published criticism to which it has been exposed, and personally do not question its correctness when the frequencies are sufficiently large.

In case B there is also, so far as I am aware, no disagreement. Pearson (3), in 1916, showed that in this case a complete correction was possible by entering Elderton's Table with a value of n^1 less than that of the frequency classes by the number of linear restrictions imposed on the sample.

The really important case is C, in which the theoretical distribution is unknown, and is reconstructed from the marginal values of the sample; this case is by far the most frequent in actual practice. It was the contention of my paper of January, 1922, that this case is equivalent to B, for the same relations are established between the sample and the population to which it is compared, either by selecting such samples as agree with a known population, or by comparing a sample chosen at random with a population constructed to agree with it in the same respects.

2—Fourfold Tables.

In any fourfold table in which the marginal totals agree, with those of the population to which it is compared in the calculation of χ^2, the differences $(x-m)$, have the same value, positive or negative, in all four quadrants. The magnitude of this difference is clearly a measure of the departure of the sample from expectation, and if it is divided by its standard error of random sampling, we find the same value χ which appears in the Pearsonian test of goodness of fit. From this it would appear that, with large samples, χ tends to be normally distributed in the distribution

$$df=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\chi^2}d\chi$$

which is identical with the Type III distribution of Elderton's Table for $n^1=1$.

Using this fact, Bowley (5, p. 371) tests the significance of two contingency tables; this was so far as I know the only instance previous to my note of January, 1922, in which the value of χ^2 had been correctly used to calculate P for testing the independence of the variety in a fourfold table. It may easily be shown in the same way that if x be any measure of divergence from proportionality in the case of a fourfold table, such as the difference of the percentages, then

$$\frac{x^2}{\sigma_x^2} = \chi^2$$

For the differences of percentages see (1). For other measures I may quote Pearson (6, p. 29):

$$\chi = \frac{r_{hk}}{{}_0\sigma_{hk}} = \frac{Q}{{}_0\sigma_Q} = \frac{\phi}{{}_0\sigma_\phi}$$

where r_{hk}, Q, ϕ, are all measures of the departure of the observed table from independence, and ${}_0\sigma_Q$, etc., are the standard errors of random sampling for a population in which the variates are independent. Pearson, however, denies that χ is normally distributed, on the ground that from Elderton's Table with $n' = 4$, its distribution should be

$$df = \sqrt{\frac{2}{\pi}}\, \chi^2 e^{-\frac{1}{2}\chi^2} d\chi$$

If I am right, Pearson was misled by assuming that $n' = 4$ gave the correct distribution when the marginal frequencies of the population are reconstructed from the sample, into the wholly untenable view that r_{hk}, Q, ϕ, and other such measures, are not normally distributed even in large samples, but that their distribution is that given above. Since each of these quantities may be defined by a natural convention so as to be equally frequently positive and negative, the latter view involves the belief that their distribution is bimodal, with a zero frequency at the central point of the symmetrical distribution.

The proof that for large samples the distribution of these quantities is normal is not difficult; in the fourfold table

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	$a+b+c+d$

let $$x = \frac{a}{a+b} - \frac{c}{c+d}$$

I have shown (1) that $$\frac{x^2}{\sigma_x^2} = \chi^2.$$

It is necessary now to show that the distribution of χ will tend to normality as the sample is increased. Considering the sub-sample $a + b$, the distribution of a will be that of the bionomial

$$(p + q)^{a+b}$$

hence the distribution of $\frac{a}{a+b}$ tends to normality as $a + b$ is increased indefinitely. Similarly that of $\frac{c}{c+d}$ tends to normality with the same mean, but different standard deviation. Moreover, from a population in which the two attributes are independent, these will be two independent samples; and it is well known that the distribution of the difference of two independent normally distributed variates is itself normally distributed. Consequently, x tends to be normally distributed about mean at zero; if then

$$\chi = \frac{x}{\sigma_x}$$

χ must be normally distributed with standard deviation unity, while for its positive value

$$df = \sqrt{\frac{2}{\pi}}\, e^{-\frac{1}{2}\chi^2}\, d\chi,$$

and not

$$df = \sqrt{\frac{2}{\pi}}\, \chi^2 e^{-\frac{1}{2}\chi^2}\, d\chi.$$

The difference of opinion is a simple one of fact. If I am right, the mean value of $\chi^2 = 1$, if Professor Pearson is right, it is 3. Professor Bowley, on the other hand, regards two assumptions as possible: In speaking of a table showing Inoculated—Not inoculated, Recovered—Died (4, Doubtful Case, p. 4), he writes:

> " If we applied the method of Case III (red and black cards) we should be assuming the total number recovered, inoculated, etc., were given, and ask whether, if recovery and inoculation were totally unconnected, so large a number would be found by chance in the first compartment.
>
> " If we applied the method of Case IV we should be assuming that we were examining only a sample from a larger universe in which the proportions recovered : died and inoculated : not inoculated were not known."

It is difficult to know what meaning is to be attached to this distinction. Professor Bowley does not explain what difference there is between this case and that treated in his book (5, p. 372), where the divisions are Recovered—Died; Not Vaccinated—Vaccinated, for which, rightly in my opinion, he takes χ to be normally distributed. Nor does he explain why this case is considered doubtful, as against Case IV (Dull—Not Dull; With defects—Without) where he, wrongly in my opinion, uses the formula $n^1 = 4$. In all these

cases the marginal totals *in the sample* are given, and *in the population* are unknown, and since the latter in all these cases has been reconstructed from the former, the sample can only differ from the population in one degree of freedom. No assumption as to what we know or do not know can alter the consequences of our procedure in calculating χ^2, and this procedure is the same in all these cases. *

3—χ^2 AS FUNCTION OF FREQUENCIES.

That the distribution of χ^2 from random samples is determined solely by the procedure by which it is calculated may be emphasized by returning to first principles. In a fourfold table let the probabilities that an event shall fall into each of the four compartments be p_1, p_2, p_3, p_4. Then

$$p_1 + p_2 + p_3 + p_4 = 1,$$

and if the variates are independent

$$p_1 p_4 = p_2 p_3,$$

so that we may write

$$p_1 = pp^1, p_2 = qp^1, p_3 = pq^1, p_4 = qq^1.$$

If now a sample of n be taken, the chance that the number of observations in the four compartments are a, b, c, d will be the term of the multinomial expansion

$$\frac{n!}{a!\ b!\ c!\ d!}\ p_1^a\ p_2^b\ p_3^c\ p_4^d.$$

Since the simultaneous distribution of a, b, c, d is thus determinate in terms of p_1, p_2, p_3, p_4, and n, the distribution of any function of a, b, c, and d, is also determinate. Two such functions may be considered; in the first place let

$$\chi^2 = \frac{a^2}{p_1 n} + \frac{b^2}{p_2 n} + \frac{c^2}{p_3 n} + \frac{d^2}{p_4 n} - n = S\left\{\frac{(a - p_1 n)^2}{p_1 n}\right\} \quad (A)$$

this is the value of χ^2 used when p_1, p_2, p_3, p_4, are known by hypothesis and we desire to test if the observation a, b, c, d is in accordance with that hypothesis. It is agreed that this function tends to be distributed as n is increased, in a manner independent of p_1, p_2, p_3, p_4 and n, in the distribution given in Elderton's Table under $n^1 = 4$.

When, however, p_1, p_2, p_3, p_4 are not known, and it is required to test, not any particular hypothesis of their values, but whether the two variates are or are not distributed independently, then we make the substitution

$$p_1 = \frac{(a + b)\ (a + c)}{n^2}, \text{ etc.},$$

and so arrive at the formula

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

* See Note on page 146 - R.A.F.

This is a different function of a, b, c, d from that previously used, and the error of using the formula $n^1 = 4$, in testing independence, consists in assuming, when a, b, c, d take all their possible values with frequencies given by the multinomial formula, that this new function is distributed in the same manner as the function previously used.

It will scarcely be disputed, after what has been already said, that if χ is calculated by equation B, then it will be distributed in random samples in a normal curve with standard deviation unity. If necessary a formal and complete proof of this fact can be given. It is sufficient here to point out the difference between the functions A and B, and to emphasize the fact that the distribution found by Pearson for function A cannot be assumed to be correct if B is the function actually employed.

The differences between the two functions A and B are, in fact, very great. A is a function of the probabilities p_1, p_2, p_3, p_4, and can only be used if these are provided by the hypothesis to be tested; it is distributed in accordance with the formula $n^1 = 4$, and its mean value is 3. On the other hand, B is a function of the observed frequencies only, and is used to test independence, that is, when our hypothesis tells us no more than that $p_1p_4 = p_2p_3$; it is distributed in accordance with the formula $n^1 = 2$, and its mean value is 1. To enter Elderton's Table under $n^1 = 4$, with a value of χ^2 calculated by equation B, is not to make a test of independence, for such a value of χ^2 is not in fact distributed in the distribution for which the table $n^1 = 4$ was calculated. The effect of this error is greatly to over-estimate the agreement of the observed sample with expectation, and correspondingly to under-estimate the significance of discrepancies from expectation based on the hypothesis of independent variates. Thus for χ^2 calculated by equation B the value obtained from random samples with independent variates will exceed 4 in only 4·55 times to 100 trials. An observed value $\chi^2 = 4$, therefore, strongly suggests that the variates are not independent. But if χ^2 be obtained from equation A, its value from random samples with independent variates will exceed 4 as many as 26·15 times to 100 trials; no significant departure from expectation could be inferred from such a value; this shows how misleading it may be to calculate χ^2 by equation B, and at the same time to assume the distribution to be that of the function given in equation A.

4—Yule's Experiments.

It is the more surprising that Bowley should have reverted to the Pearsonian mode of testing fourfold tables since the actual distribution of χ^2 in this case has been determined experimentally by Yule. Yule's experiment was designed to settle the question of the distribution of χ^2 in the fourfold table in Case C, where the population is reconstructed from the sample. He also calculated χ^2 from the known population and verified that in this case the

Pearsonian formula $n^1 = 4$ was correct. No less than 350 observations were made. The distributions of the values of χ^2 calculated from the reconstructed populations were as follows (7, p. 100):

	Number Expected $n^1 = 2$.	Number Observed.	Number Expected $n^1 = 4$.
0 — ·25	134·02 +	122	10·80 —
·25 — ·50	48·15 —	54	17·58 —
·50 — ·75	32·56 —	41	20·13 —
·75 — 1·00	24·21 +	24	21·05 —
1 — 2	56·00 —	62	80·10 +
2 — 3	25·91 +	18	63·?7 +
3 — 4	13·22 +	13	45·56 +
4 — 5	7·05 +	6	31·38 +
5 — 6	3·86 —	5	21·07 +
6 —	5·01 +	5	39·06 +
	349·99	350	350·00

There can be no question that the expectation $n^1 = 4$ completely fails, while $n^1 = 2$ fits the observations well; calculating P from the 10 classes, the distribution being known *à priori*, $n^1 = 10$, $\chi^2 = 7{\cdot}53$, $P = {\cdot}583$, according to Yule, for $n^1 = 2$. For $n^1 = 4$ the fit is so bad that it is not worth while to calculate the exact value of χ^2. By no possibility could it be considered as fitting the observations; and it is to be emphasized that in this series the procedure of calculating χ^2 was to take a random sample, without limitation of its marginal frequencies, and compare it with a reconstructed population having the same marginal frequencies. That is to say, the procedure was that of Case C, which is the most important case in view of the frequency of its occurrence.

Yule's data thus affords a conclusive confirmation of my theoretical conclusions for the case of the fourfold table. Its generality for contingency tables will be readily conceded by those who will follow the reasoning of my paper of January, 1922. The application to the goodness of fit of curves fitted by moments or otherwise was only touched upon in that paper. For the present it will be sufficient to say that the correction is needed, whenever the population is reconstructed from the sample; and is exact whenever the Type III distribution of Elderton's Table is exact. In the original paper one cause of inexactitude was mentioned; others may be added, but for the practical application of tests of Goodness of Fit, it is sufficient that the method of fitting should be such that χ^2 does not depart far from its minimum value. In fitting the normal curve by moments it is easy to verify that Elderton's Table with corrected n^1 gives the exact distribution of χ^2.

REFERENCES

1. R. A. FISHER (1922).—On the interpretation of χ^2 from contingency tables, and on the calculation of *P.*—*J.R.S.S.*, LXXXV, pp. 87-94.

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.*, L, pp. 157, etc.

3. K. PEARSON (1916).—On the general theory of multiple contingency with special reference to partial contingency.—*Biom.*, XI, pp. 145-158.

4. A. L. BOWLEY and L. R. CONNOR (1923).—Test of correspondence between statistical grouping and formulæ.—ECONOMICA, No. 7, pp. 1-9.

5. A. L. BOWLEY (1920).—*Elements of Statistics.* P. S. King and Son, London.

6. K. PEARSON (1912).—On a novel method of regarding the association of two variates classed solely in alternate characters.—*Drapers Company Research Memoirs*, Biometric Series VII.

7. G. UDNY YULE (1922).—On the application of the χ^2 method to association and contingency tables, with experimental illustrations.—*J.R.S.S.*, LXXXV, pp. 95-104.

NOTE.—Professor Pearson has since admitted (*Biometrika*, XIV, p. 418) that Greenwood and Yule's tables (Inoculated—Not inoculated, Attacked—Not attacked) for Typhoid and Cholera are correctly treated by taking $n^1=2$. Presumably the same rule may now be allowed for other diseases. Professor Pearson has, however, opened a new and unexpected line of defence by claiming that these tables are not fourfold tables at all. It is difficult to be certain what distinction is in view; the only distinction mentioned is that " they " (Greenwood and Yule) " have arbitrarily fixed by the size of their inoculated and uninoculated groups two of the marginal totals." To avoid confusion of thought three points may be noted: (i) Greenwood and Yule did not in any sense fix the numbers inoculated and uninoculated, but accepted all suitable data reported; (ii) if the data had referred to experimental conditions in which the proportion of inoculated to uninoculated could be assigned at will, this circumstance would have made no difference to the distribution of χ^2, since the marginal proportion in the population with which the sample is compared is, in any case, identified with that of the sample; (iii) the proportion of inoculated to uninoculated involves only one degree of freedom; in order to diminish the degrees of freedom from 3 to 1 it would be necessary, on Professor's Pearson's argument, for Greenwood and Yule to fix, equally arbitrarily, the numbers attacked and not attacked by the epidemics. R. A. F.

[In regard to Mr. Fisher's Cases A and B no doubt has arisen. " A " is Prof. Pearson's original problem, and Case I in the article

in the last issue of ECONOMICA. " B," where the marginal totals (or the moments) are fixed and the same for all samples, is that treated by Mr. Fisher generally in the *Statistical Journal*, 1922, pp. 87-94, in the *Elements of Statistics*, pp. 371-2, and in Case III, ECONOMICA.

The whole difficulty lies in Case C, which corresponds to Cases II and IV. In Mr. Yule's experiment and in Mr. Fisher's treatment, the marginal totals are not kept constant (as they are in *Elements*, pp. 371-2), and the reconstructed population is adjusted for each sample (which is not done in Case IV) ; for in each of Mr. Yule's 350 samples the numbers occurring were compared with a population with the same marginal totals as in that sample. Mr. Fisher indicates, but does not give explicitly, a proof that the theoretical distribution of χ^2 in such a reckoning is very close to that found by Mr. Yule ; perhaps he should have emphasized that this is not merely a corollary to his important proof of the right treatment of Case B.

The problem is, to find the chance that so great a divergence from proportionality as is observed would be found in a random selection from an uncorrelated population. The solution given in ECONOMICA was that, if the proportions in the population were p_1, p_2, p_3, p_4, the result is that given above in connection with Mr. Fisher's formula (*A*). It was shown that the chance is greatest when $p_1 = (a + b)(a + c)/n^2$, etc., but it was not supposed that $p_1, p_2 \ldots$ varied from experiment to experiment as in connection with formula (*B*) ; the supposition was that many samples were taken from the unchanged selected population.

The dispute is not about the mathematics ; the doubt is whether the variation of samples supposed in ECONOMICA, or that supposed by Mr. Fisher and Mr. Yule, is appropriate to the problem. Prof. Pearson (*Biometrika*, XIV, p. 189) may be cited in favour of the ECONOMICA supposition.—A. L. BOWLEY.]