KARL PEARSON'S APPROACH TO χ^2

CHAPTER X.

TESTS OF CORRESPONDENCE BETWEEN DATA AND FORMULÆ.

In the general method of the representation of observations by a mathematical formula, the question must arise how the adequacy of the formula is to be tested, or, as it is frequently phrased, a test of the goodness of fit is required.

Consider for example the table used above (p. 310) of the weekly expenditure on food per "unit" in 970 families.

	m'	m			
Expenditure.	number of	calculated	$e = m \sim m'$	Standard	$\frac{e^2}{m}$
	cases.	numbers.	difference.	deviations.	
Not exceeding $5.5s$	18	22	4	4.6	.7
5.5	107	123	16	10.4	2.1
7.5	255	234	21	13.3	1.9
9.5	245	249	4	13.6	.1
11.5	173	168	5	11.8	.1
13.5	101	89	12	9.0	1.6
15.5	38	51	13	7.0	3.3
17.5	17	22	5	4.6	1.1
19.5	9	11	2	3.3	.4
Over 21.5	7	1	6	?	36.0
Totals	970	970	88	—	47.3

The calculated numbers are from the second approximation to the Law of Great Numbers. A rough method formerly used was to add the differences between the calculated numbers and the numbers observed in each compartment, irrespective of sign, and to express this total as a percentage of the number of cases. The "percentage misfit" thus calculated is $88 \div 9.70 = 9.1$ per cent.

The weakness of this method is that it is not related to any measurement of probability, and one cannot tell at sight whether the fit is good or not. Of two competing formulæ, the presumption is that that which gives the lower percentage misfit is the better; also when we have several sets of similar observations we can tell roughly by this method which is nearest to the formula, and in some cases in which set the observations are most regular.

The percentage misfit is generally diminished if compartments are merged together.

As regards the contents of individual compartments, we already have a simple test. If m_t is the calculated number in a compartment when there are N observations in all, the chance of finding $m_t + e_t$ observations in this compartment in

$$\frac{1}{\sigma\sqrt{\pi}}e^{-\frac{1}{2}\cdot\frac{e_t^2}{\sigma^2}} \text{ (formula (19)) where } \sigma^2 = \frac{m_t}{N}\left(1 - \frac{m_t}{N}\right)N,$$

and the probability of exceeding any assigned multiple or sub-multiple of σ is given by the table (p. 271). The standard deviation for each grade in the above example except the last is given, and it is seen that four out of nine errors are less than σ, their standard deviation, two are between σ and $\frac{3\sigma}{2}$ and the remaining three less than 2σ. No separate measurement is improbable, and therefore the whole grouping may be presumed to be not improbable, except the final number, 7 above $21.5s$.

That numbers in extreme grades should be discontinuous in relation to middle grades is common in many classes of observations.

The deviations are not independent, however, since their total must be zero; and even if the deviation in one compartment taken by itself is improbably large, it may yet not be improbable when all the compartments are considered. A measurement which allows for this modification has been devised by Professor Pearson, and part of the analysis in a simplified form, a brief table of the results, and some applications are given in the following paragraphs (see *The Philosophical Magazine*, No. 302, July, 1900, pp. 157–175).

Suppose that a formula, which is presumed to represent the distribution of observations, leads to the expectation of m_1, $m_2 \dots m_n$ observations in n grades or compartments, when $N, = m_1 + m_2 + \cdots + m_n$, is the whole number of observations.

In an experiment or group of observations. suppose that $(m_1 + e_1) \dots (m_t + e_t) \dots (m_n + e_n)$ are found in the compartments, so that $e_1 + \cdots + e_t + \cdots + e_n = 0$.

Write $p_1 = \frac{m_1}{N} \dots p_t = \frac{m_t}{N} \dots$.

Then p_t is the chance that an observation from a group satisfying perfectly the formula will fall into the t^{th} grade.

The chance that $m_t + e_t$ will fall into this grade when N are chosen at random from an indefinitely large universe is

$$\frac{1}{\sigma\sqrt{\pi}} e^{-\frac{1}{2} \cdot \frac{e_t^2}{\sigma_t^2}},$$

where $\quad \sigma_t^2 = p_t(1 - p_t)N = p_t q_t N$, where $q_t = 1 - p_t$.

It can be shown that the joint chance of the errors named is

$$Ke^{-\frac{1}{2}\mathrm{X}^2}, \text{ where } X^2 = \mathrm{S}.\frac{e_t^2}{m_t}, \text{ and } Se_t = 0,$$

K being a constant.

For, if there were only two compartments, $e_1 + e_2 = 0$, and the joint chance equals the chance of either.

Then $\quad p = \frac{m_1}{N}, q = \frac{m_2}{N}, m_1 + m_2 = N.$

The chance is

$$\frac{N^{\frac{1}{2}}}{\sqrt{2\pi m_1 m_2}} e^{-\frac{1}{2}\left(\frac{e_1^2}{m_1} + \frac{e_2^2}{m_2}\right)}, \text{ since } \frac{e_1^2 N}{m_1 m_2} = \frac{e_1^2(m_1 + m_2)}{m_1 m_2}, \text{ and } e_1^2 = e_2^2.$$

If there are *three* compartments

$$e_1 + e_2 + e_3 = 0, \quad m_1 + m_2 + m_3 = N, \quad \sigma_1^2 = \frac{m_1}{N}.\frac{m_1 + m_2}{N}.N,$$

and similarly for σ_2^2 and σ_3^2.

$$2e_1e_2 = e_3^2 - e_1^2 - e_2^2.$$

$$r\sigma_1\sigma_2 = \text{mean } e_1e_2 = \tfrac{1}{2}(\sigma_3^2 - \sigma_1^2 - \sigma_2^2)$$

$$= \frac{1}{2N}\{m_3(m_1+m_2) - m_1(m_2+m_3) - m_2(m_1+m_3)\}$$

$$= -\frac{m_1m_2}{N}. \quad \text{(Compare p. 419.)}$$

The chance of the concurrence of e_1 and e_2, and therefore of e_3 also, is given by the normal correlation surface as

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} e^{-\frac{1}{2(1-r^2)}\left(\frac{e_1^2}{\sigma_1^2} + \frac{e_1^2}{\sigma_1^2} - \frac{2re_1e_2}{\sigma_1\sigma_2}\right)}.$$

Now

$$\sigma_1^2\sigma_2^2(1-r^2) = \frac{m_1m_2(m_2+m_3)(m_1+m_3)}{N^2} - \frac{m_1^2m_2^2}{n^2} = \frac{m_1m_2m_3}{N},$$

since $m_1 + m_2 + m_3 = N$.

Hence the index of e is

$$-\frac{N}{2m_1m_2m_3}(e_1^2\sigma_2^2 + e_2^2\sigma_1^2 - 2r\sigma_1\sigma_2e_1e_2)$$

$$= -\frac{N}{m_1m_2m_3}\left\{\frac{e_1^2m_2(m_1+m_3)}{N} + \frac{e_2^2m_1(m_2+m_3)}{N} + \frac{2e_1e_2m_1m_2}{N}\right\}$$

$$= -\frac{1}{2m_1m_2m_3}\left\{(e_1+e_2)^2m_1m_2 + e_1^2m_2m_3 + e_2^2m_1m_3\right\}$$

$$= -\frac{1}{2}\left(\frac{e_1^2}{m_1} + \frac{e_2^2}{m_2} + \frac{e_3^2}{m_3}\right), \text{ since } e_1 + e_2 = -e_3.$$

Now if the second and third compartments had been merged into one containing $M + E$ observations, where $M = m_2 + m_3$ and $E = e_2 + e_3$, the chance would have been

$$K_1 e^{-\frac{1}{2}\left(\frac{e_1^2}{m_1} + \frac{E^2}{M}\right)},$$

where K_1 is a constant.

The effect, therefore, of dividing the second compartment without changing the first is to alter the constant and to replace $\frac{E^2}{M}$ by $\frac{e_2^2}{m_2} + \frac{e_3^2}{m_3}$ in the index.

Similarly if two compartments are given, the effect of dividing the third compartment without changing the first two must be to alter the constant and to replace $\frac{e_3^2}{m_3}$ by $\frac{e_3^2}{m_3} + \frac{e_4^2}{m_4}$ in the index, and so on.

Hence for n compartments the chance, P, of errors e_l, $e_2 \dots e_n$. is

$$Ke^{-\frac{1}{2}X^2}, \text{ where } X^2 = \frac{e_1^2}{m_1} + \frac{e_2^2}{m_2} + \cdots + \frac{e_n^2}{m_n},$$

and $$e_1 + e_2 + \dots e_n = 0 \dots\dots\dots\dots\dots\dots (130).$$

Notice that X^2 is the same expression as is used in obtaining the coefficient of contingency.

[A proof of the formula, without the above method of induction is given by Pearson, by the use of the multiple correlation equation.]

If the selections in the compartments had been independent and without the condition that $e_1 + e_2 + \cdots = 0$, the chance would have been

$$Ke^{-\frac{1}{2}X^2} \times e^{-\frac{1}{2}\left(\frac{e_1^2}{S-m_1}+\frac{e_2^2}{S-m_2}+\cdots\right)}$$

for the index would have been

$$-\frac{1}{2}\left(\frac{e_1^2 N}{m_1(N-m_1)}+\ldots\right) = -\frac{1}{2}\left(\frac{e_1^2}{m_1}+\frac{e_1^2}{N-m_1}+\ldots\right).$$

If there are many compartments and the largest of the fractions $\frac{m_t}{N}$ is small, the second part of the index is negligible N compared with the first, and the two expressions tend to equality, and the effect of the correlation is small.

The chance of the occurrences if there is no correlation is less than that when there is correlation, since the last factor, if not negligible, is less than 1. (The constant is eliminated in further processes.) Hence the aggregation of uncorrelated chances, which is simpler than the present method, gives, an unduly unfavourable view of the appropriateness of a formula.

The chance of every system of errors that gives a particular value of X^2 is the same. Now, when the probability of.a deviation from the mean in normal frequency is in question, it is customary to measure the probability that so great a deviation to left or right should have occurred, viz.,

$$2\int_z^\infty \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}dz.$$

Similarly here we may measure the chance of the occurrence of the system of errors or a less probable system by evaluating

$$2\iint \ldots Ke^{-\frac{1}{2}X^2}d_X,\ \text{where } d_X \text{ is written for } de_l.de_2\ldots de_{n-1}$$

and the integral is $\overline{n-1}$ fold and extended from X to ∞, with the condition $e_1 + e_2 + \cdots + e_n = 0$, K being so chosen that

$$\int_{-\infty}^{\infty} Ke^{-\frac{1}{2}X^2}d\ = 1.$$

The existence of this condition makes the integration complicated, and reference should be made to Pearson's original analysis for its working out.

The result is that

$$P = \sqrt{\frac{2}{\pi}}\int_X^\infty e^{-\frac{1}{2}X^2}.d_X + \sqrt{\frac{2}{\pi}}e^{-\frac{1}{2}X^2}\left(\frac{X}{1}+\frac{X^2}{1.3}+\cdots+\frac{X^{n-3}}{1.3.5-\overline{n-3}}\right)$$

when n is even, and

$$P = e^{-\frac{1}{2}X^2}\left(1 + \frac{X^2}{2} + \cdots + \frac{X^{n-2}}{2.4\ldots n-3}\right) \text{ when } n \text{ is odd.} \tag{131}$$

A table of the values of P for various values of x2 and n is given in *Biometrika*, Vol. 1, pp. 155 *seq*. We can, in a very brief form, obtain a working rule for determining whether a formula does or does not adequately represent an observed group by picking out values of x^2 which for a given n make $P = \frac{1}{2}$ or slightly more, or, further up the scale of improbability, make $P = .0455$ or slightly less, which corresponds to twice the standard deviation in the normal curve.

n.	X .	P.	X^2.	P.
3	1	.61	6	.050
4	2	.57	8	.046
5	3	.56	10	.040
6	4	.55	12	.035
7	5	.54	13	.043
8	6	.54	15	.036
9	7	.54	16	.042
10	8	.53	18	.035
11	9	.53	19	.040
12	10	.53	20	.045
13	11	.53	22	.038
14	12	.53	23	.042
15	13	.53	24	.046
16	14	.526	26	.038
17	15	.525	27	.041
18	16	.524	28	.045
19	17	.523	30	.037
20	18	.522		
25	23	.520		
30	28	.518		

If $X^2 < n - 2$, it is at least an even chance—as likely as not—that the observations would be found from a group represented by the formula.

If $X^2 > 2n$, the improbability is considerable.

Strictly, the test should be applied using as many compartments as are given by the observations, for the merging of compartments affects the resulting value of P; but it is often difficult to get back to ungraded observations, and in the case of continuous variables, such as height, the original grades would be as fine as the measurements could be made.

A more serious difficulty is that in any compartment the observed $m_t + e_t$ must be integral, while m_t is in general not integral, and some value of e_t would be found in the most perfect representation. In consequence, the number to be expected in the least occupied compartment must be reasonably large, or we obtain spurious contributions to X^2. This in practice rules out detailed extreme compartments, and in their rejection or fusion an element of arbitrariness is introduced and no fine measurement is possible.

On the other hand, when we are testing the applicability of the normal curve of error, or the general law of great numbers, based on Edgeworth's hypothesis (p. 298–9), there is no expectation of closeness of fit on abscissæbeyond a small multiple of the standard deviation—the smaller as the number of independent elements that contribute to the measurement diminishes—so that the test is only applicable to the well-occupied central compartments ; but in choosing the extent over which the test is made, the fineness of the method is lost.

Hence, only a broad, but often sufficiently definite, result can be obtained.

Illustrations.

If we neglect the extreme grade in Example 7, on p. 310, $X^2 = 11.3$, $n = 9$, $P = .18$, and the formula "2nd approx." is adequate.

If we take the Pearsonian formula, on the same page, $X^2 = 21.4$, $n = 9$, $P = .006$, but if we exclude the lowest as well as the highest grade, $X^2 = 4.1$, $n = 8$, $P = .77$; hence this formula expresses the central eight grades but not either extreme.

The same conclusions are reached if we simply take the standard deviations of the grades separately.

In the table on p. 309 relating to the ages of school children, $n = 8$. The normal curve gives $X^2 = 16.7$ and $P = .02$, which is not satisfactory. The second approximation, however, gives $X^2 = .47$ and P is indistinguishable from 1.

In the experiment on the numbers of letters in words (pp. 305–6), the sum of 10 words, graded by 5 letters, gives $n = 13$, and with the normal curve $X^2 = 33$, $P = .001$, or omitting the lowest and two highest extreme grades, $n = 10$, $X^2 = 6.1$, $P = .73$. The second approximation, however, including all grades, gives $X^2 = 8.4$, $P = .74$.

The sums of 100 words graded by 20 letters give $n = 10$, $X^2 = 2.96$, $P = .965$ with the normal curve, and no further approximation can improve on this.

An example of a different kind is found, when a distribution found by sample is compared with the whole group from which the sample is taken, to verify the rules of sampling or the adequacy of the method.

NUMBER OF COMPANIES PAYING DIVIDENDS AT VARIOUS RATES.

	Number in sample m.	Relative numbers in all companies m.	Standard deviation.	$\frac{e^2}{m}$
Below 3 per cent.	34	30	5.3	.53
3 per cent.	108	108.8	8.9	0
4 " "	117	124.4	9.3	.44
5 " "	60	70.8	7.4	1.65
6 per cent. to 8 per cent.	48	43.2	6.2	.53
8 per cent.	33	22.8	4.6	4.57
	400	400		7.72

Here $n = 6$, $X^2 = 7.72$, $P = .185$. The result is fairly good, but spoilt by the highest grade.

This test has been applied to the distribution in two dimensions, in the experiment tabulated on p. 394.

The 24 squares, .3 to left and right of centre, and 2 above and below it, which contain in theory 11 or more observations, were taken as separate compartments. Outlying squares were grouped in the 9 regions shown by the thick lines, rather arbitrarily, so as to get contiguous squares which aggregated to at least 9 expected observations in the second approximation. The results are as follows:—

	Normal surface.		2nd approximation	
	X^2.	P.	X^2.	P.
24 central squares	20.3	.59	17.5	.79
9 outlying regions	27.8		10.1	
33 regions	48.6	.035	27.6	.59

The improvement in the outlying regions by the use of the second approximation is very marked.

From: A L Bowley, *Elements of Statistics* (4th edn), London: P S King 1920.