

## MISCELLANEA.

### I. On the $\chi^2$ test of Goodness of Fit.

BY KARL PEARSON, F.R.S.

IN a paper published in the *Philosophical Magazine* for July 1900, pp. 157—175, I dealt with the following problem : A very large population is sampled, say, the population  $n_1, n_2, \dots, n_p$ , with total  $N$ , and any individual sample is  $m_1, m_2, \dots, m_p$ , total  $M$ . The "probable constitution" is given by :

$$m_1' = \frac{M}{N} n_1, \quad m_2' = \frac{M}{N} n_2, \quad \dots \quad m_p' = \frac{M}{N} n_p.$$

If a large number of samples of size  $M$  are taken, what is the distribution of variations from the "probable constitution" in these samples ?

I showed that if the distribution of categories were such that no category contained a few isolated units, then the distribution depended on the calculation of  $\chi^2 = \sum_1^p \frac{(m_i - m_i')^2}{m_i'}$ , and provided a value for the probability  $P$  that samples would not diverge more than any given sample from the "probable constitution." This process is now familiar to statisticians as the  $\chi^2$ ,  $P$  test.

The sole limiting conditions were that the samples should be random, and each should be of the same size  $M$ .

In some cases the "probable constitution" ( $m'$  series) can be found at once because the distribution of the sampled population is known *a priori*. In other cases the values of the  $m'$  series have to be approximated to, and such approximations are the general rule in all discussions of probable error.

We say for example that the standard deviation of the mean of a sample taken from an indefinitely large population of size  $N$  and standard deviation  $\sigma$  is  $\sigma/\sqrt{n}$ , where  $n$  is the size of the sample.

We say that the standard deviation of second moment-coefficients of samples of size  $n$  is

$$\frac{\sqrt{\mu_4 - \mu_2^2}}{\sqrt{n}},$$

where  $\mu_2 (= \sigma^2)$  and  $\mu_4$  are the second and fourth moment-coefficients of the population sampled. In fact every constant of the sample has a probable error determinable in terms of the constants of the sampled population. All these distributions of deviations from "probable constitution" are true for perfectly general but random samples of size  $n$  drawn from our indefinitely large population.

But unfortunately in a considerable number of cases that sampled population is unknown to us ; we have no direct means of finding  $\mu_2, \mu_4$ , etc. What accordingly do we do ? Why we replace the constants of the sampled population by those calculated from the sample itself, as the best information we have. And the justification of this proceeding is not far to seek.  $\mu_2$  as found for the sample will only differ from the  $\mu_2$  of the sampled population by terms of the order  $1/\sqrt{n}$  ; for example if we are not dealing with *small* samples, and  $\sigma'$  be the standard deviation of the sample,  $\sigma'$  differs from  $\sigma$  by terms of the order  $\sigma/\sqrt{2n}$  and accordingly the standard deviation of the mean is written  $\sigma'/\sqrt{n}$  when it is really  $\sigma/\sqrt{n}$ . This method of treating probable errors is universal in the case of fair sized samples to-day and scarcely needs justification. In writing the

sample values of the constants for those of the sampled population, we do not in any way alter our original supposition that we are considering the distribution of random samples of size  $n$ . We have still  $p-1$  degrees of freedom, if we have  $p$  categories of frequency.

The process of substituting sample constants for sampled population constants does *not* mean that we select out of possible samples of size  $n$ , those which have precisely the same values of the constants as the individual sample under discussion. Clearly the given sample has definite moment-coefficients, and if there be  $p$  frequency categories the first  $p-1$  moment-coefficients together with the size  $n$  of the sample would suffice to fix all the frequencies of the  $p$  categories\*. Hence no deviations from the "probable constitution" would be possible if we confined our attention to samples of  $n$  tied to the constants of the given sample! In using the constants of the given sample to replace the constants of the sampled population, we in no wise restrict the original hypothesis of free random samples tied down only by their definite size. We certainly do not by using sample constants reduce in any way the random sampling degrees of freedom.

What we actually do is to replace the accurate value of  $\chi^2$ , which is unknown to us, and cannot be found, by an approximate value, and we do this with precisely the same justification as the astronomer claims, when he calculates his probable error on his observations, and not on the mean square error of an infinite population of errors which is unknown to him. The whole of this matter was very fully discussed (pp. 164-7) in my original paper dealing with the  $\chi^2$ ,  $P$  test.

The above re-description of what seem to me very elementary considerations would be unnecessary had not a recent writer in the *Journal of the Royal Statistical Society*† appeared to have wholly ignored them. He considers that I have made serious blunders in not limiting my degrees of freedom by the number of moments I have taken; for example he asserts (p. 93) that if a frequency curve be fitted by the use of four moments then the  $n'$  of the tables of goodness of fit should be reduced by 4. I hold that such a view is entirely erroneous, and that the writer has done no service to the science of statistics by giving it broad-cast circulation in the pages of the *Journal of the Royal Statistical Society*.

What he would obtain if he placed this restriction on his samples is not the  $\chi^2$  for the distribution of samples of size  $n$ , but of samples which give definite moments. The absurdity of this manner of approach is at once obvious, if as I have suggested, we consider the  $p$  first-moments, as there is no reason why we should not do,—for these are just as much "fixed" as the first four—and the conclusion must be that we can learn nothing at all about variation from our sample; for we have  $p$  frequency groups and  $p$ -tying conditions.

When we wish to find the probable error of a mean or a standard deviation, we do *not* start by fixing down these characters to their values in the individual sample; we suppose them to take all the possible values they could take by sampling, and after we have reached our measure of variation we then put into our formula the sampled values, to give an approximate value to the functions reached, because we are in ignorance of the real values in the sampled population.

The writer in the *Journal of the Royal Statistical Society* speaks as if I applied  $\chi^2$  to a contingency table *starting* by fixing the marginal totals. As far as I am aware I am not guilty of this. My conception of contingency is very different from my conception of  $\chi^2$ . I started my conception of contingency with the idea not of a random sample, but with the idea that some function of frequencies alone without regard to their relation to the measured characters would lead to the value of the correlation. Naturally I started from the deviation of the individual cell contents from the same cell contents on the basis of independent probability, as determined by the marginal totals. There was no question of sampling in the matter. In now fairly usual notation I termed

$$m_{uv} = \frac{m_{u.} m_{.v}}{M}$$

\* This is Thiele's method of representing frequency distributions.

† Vol. LXXXV. p. 87, 1922.

the cell contingency and after playing about with such cell contingencies for a time succeeded in finding a function  $\phi^2$  of them which for indefinitely fine grouping for a bi-variate normal frequency distribution gave the correlation  $r$  as :

$$r = \sqrt{\frac{\phi^2}{1 + \phi^2}},$$

where

$$\phi^2 = \frac{1}{M} S \frac{\left( m_{xy} - \frac{m_x m_y}{M} \right)^2}{\frac{m_x m_{x'}}{M}} \dots\dots\dots (a).$$

I see no reason for confusing this  $\phi^2$  as a measure of correlation with the  $\chi^2$  which is a measure of variability in the samples of constant size drawn from an indefinitely large population. It was different in its origin, as far as I am concerned, and different in its use. It is only when we come to consider the probable error of  $\phi^2$  that we have to distinguish between (a) the actual marginal totals of the sample and (b) the probable constitution of the marginal totals as deduced from an indefinitely large sampled population.

There are, as those who have read *Biometrika*\* will recognise, considerable difficulties about determining the probable error of  $\phi^2$ , where

$$1 + \phi^2 = S \left( \frac{m_{xy}^2}{m_x m_{x'}} \right),$$

and the determination of the mean  $\phi^2$  and of the standard deviation of  $\phi^2$  involves very troublesome analysis.

So laborious is the arithmetic involved that for ordinary statistical use it became doubtful whether it would not be better to define  $\phi^2$  as the mean squared contingency measured not from the marginal totals of the sample, but from the "probable constitution" of the marginal totals of the sample as deduced from the sampled population. In this case if

$$m'_{xy} = \frac{M}{N} n_{xy}, \quad m'_{x.} = \frac{M}{N} n_{x.}, \quad m'_{.y} = \frac{M}{N} n_{.y},$$

$$\phi^2 = S \frac{\left( m_{xy} - \frac{m'_{x.} m'_{.y}}{M} \right)^2}{\frac{M}{M'} \cdot \frac{m'_{x.} m'_{.y}}{M}} \dots\dots\dots (\beta)$$

or,

$$1 + \phi^2 = S \left( \frac{m_{xy}^2}{m'_{x.} m'_{.y}} \right);$$

with this change of definition the probable error and mean of  $\phi^2$  are more easily obtainable, and in this case for the first time,  $M\phi^2$  can be looked upon as equivalent to a  $\chi^2$ .

The form (a) from my standpoint cannot be treated as a  $\chi^2$ , because it is not the deviation-measure of a given sample from the sampled population. Nor again is (β) the deviation-measure of the sample from the sampled population, unless we assume that population to have zero contingency, i.e.  $m'_{xy} = m'_{x.} m'_{.y} / M$ .

But  $\chi^2$  may in the form (β) be treated as a deviation-measure of the actual sample from an artificial sampled population, which differs from the actual population in having no correlation or contingency, but having the same marginal distributions of the two characters.

The moment, however, we assume form (β) for our contingency we are giving, what we clearly must give, absolute freedom to the marginal totals of our samples. The sole limit on our sample is its total size  $M$ . But when we come to actually calculating  $\phi^2$  for the individual sample, or the mean value or the standard deviation (i.e. probable error) of  $\phi^2$  for a series of samples, we have only one course open to us, if we do not know the constants of the sampled population, we must insert the marginal totals of the individual sample of which we have cognizance in place of the

\* Vol. v. p. 191, Vol. x. p. 570, Vol. xi. p. 570, and Vol. xii. p. 259.

unknown values of the sampled population. Thus (a) and (β) provide ultimately the same  $\phi^2$ , but the probable error of  $\phi^2$  and the mean value of  $\phi^2$  will be different in the two cases. In the first case we vary our marginal totals with the sample as they obviously would vary in practice. In the second case we define our  $\phi^2$  to be a deviation from the independent probability of an artificial population, we do not keep the marginal totals of the sample fixed any more than in (a). But if we think in terms of  $\chi^2$  (and not  $\phi^2$ ) we appear to do so because ultimately we have to take our marginal probabilities as those of the sample in default of a knowledge of any better values.

This point seems to me well illustrated in what my critic in the *Journal of the Royal Statistical Society* has to say on p. 90 of his paper about Messrs Greenwood and Yule's use of  $\chi^2$  for a fourfold table. He asserts that they ought to have entered the table of goodness of fit with  $n'=2$ . The problem before them was whether their fourfold tables could possibly be samples of bi-variate independent probability distributions. Each sample from such a distribution would have perfectly free cell frequencies  $m_{11}, m_{12}, m_{21}, m_{22}$ , subject to the sole binding condition that

$$m_{11} + m_{12} + m_{21} + m_{22} = M.$$

The proper  $\chi^2$  is given by

$$\chi^2 = \frac{\left(m_{11} - \frac{m'_{1.}m'_{.1}}{M}\right)^2}{\frac{m'_{1.}m'_{.1}}{M}} + \frac{\left(m_{12} - \frac{m'_{1.}m'_{.2}}{M}\right)^2}{\frac{m'_{1.}m'_{.2}}{M}} + \frac{\left(m_{21} - \frac{m'_{2.}m'_{.1}}{M}\right)^2}{\frac{m'_{2.}m'_{.1}}{M}} + \frac{\left(m_{22} - \frac{m'_{2.}m'_{.2}}{M}\right)^2}{\frac{m'_{2.}m'_{.2}}{M}} \dots (\gamma),$$

and this has three degrees of freedom and is what Messrs Yule and Greenwood desired to find, and they properly used the value of  $P$  for  $n'=4$ .

Then like the astronomer, who finding the probable error of his mean to be  $.67449\sigma/\sqrt{M}$  and not knowing the  $\sigma$  of his sampled population, puts it equal to the  $\sigma$  of his observations, so Messrs Yule and Greenwood very properly replaced the marginal totals of their unknown population by those of their sample, but very properly did not replace  $n'=4$  by  $n'=2$ !

But says my critic\*, if they had, they would have got the same measure of improbability as if they had compared the difference of percentages! Quite so, and obviously so; for in taking percentages they have actually fixed their marginal totals taking 100 of each class and thus for the first time confined their attention to a limited class of samples, not the random sample of size  $M$ , which has not its marginal totals fixed. We have, indeed, reduced our degrees of freedom by two in taking ratios.

When we consider generally the  $\chi^2$  for a fourfold table to measure the improbability of a sample we are really comparing the special sample

$a$	$b$	$a+b$	with	$a'$	$b'$	$a'+b'$
$c$	$d$	$c+d$		$c'$	$d'$	$c'+d'$
$a+c$	$b+d$	$M$		$a'+c'$	$b'+d'$	$M$

the general population, where in the latter case  $a'd' = c'b'$ .

Now the mean square contingency of the first of these tables is

$$\begin{aligned} \phi^2 &= \frac{1}{M} \left\{ \frac{\left(a - \frac{(a+b)(a+c)}{M}\right)^2}{\frac{(a+b)(a+c)}{M}} + \frac{\left(b - \frac{(a+b)(b+d)}{M}\right)^2}{\frac{(a+b)(b+d)}{M}} + \frac{\left(c - \frac{(a+c)(c+d)}{M}\right)^2}{\frac{(a+c)(c+d)}{M}} + \frac{\left(d - \frac{(c+d)(b+d)}{M}\right)^2}{\frac{(c+d)(b+d)}{M}} \right\} \\ &= \left\{ \frac{a^2}{(a+b)(a+c)} + \frac{b^2}{(a+b)(b+d)} + \frac{c^2}{(a+c)(c+d)} + \frac{d^2}{(c+d)(b+d)} - 1 \right\} \\ &= \frac{(ab - cd)^2}{(a+b)(a+c)(b+d)(c+d)}. \end{aligned}$$

\* Loc. cit. p. 90.

But the  $\chi^2$  is

$$\begin{aligned} &= \frac{\left(a - \frac{(\alpha' + b')(\alpha' + c')}{M}\right)^2}{\frac{(\alpha' + b')(\alpha' + c')}{M}} + \frac{\left(b - \frac{(\alpha' + b')(b' + d')}{M}\right)^2}{\frac{(\alpha' + b')(b' + d')}{M}} + \frac{\left(c - \frac{(\alpha' + c')(c' + d')}{M}\right)^2}{\frac{(\alpha' + c')(c' + d')}{M}} + \frac{\left(d - \frac{(c' + d')(b' + d')}{M}\right)^2}{\frac{(c' + d')(b' + d')}{M}} \\ &= M \left\{ \frac{a^2}{(\alpha' + b')(\alpha' + c')} + \frac{b^2}{(\alpha' + b')(b' + d')} + \frac{c^2}{(\alpha' + c')(c' + d')} + \frac{d^2}{(c' + d')(b' + d')} - 1 \right\}, \end{aligned}$$

there being *three* degrees of freedom or we must take  $n' = 4$  in calculating the probability  $P$ , this may be written

$$\chi^2 = \frac{1}{M} \left\{ \frac{a^2}{p'_{.1}p'_{.1.}} + \frac{b^2}{p'_{.1}p'_{.2.}} + \frac{c^2}{p'_{.2}p'_{.1.}} + \frac{d^2}{p'_{.2}p'_{.2.}} - 1 \right\} \dots\dots\dots(\delta),$$

where  $p'_{.1}$ ,  $p'_{.2}$ ,  $p'_{.1.}$  and  $p'_{.2.}$  are the four percentage numbers of the marginal categories in the sampled population. Now we do not know these percentages in that population and we do what every physicist, every astronomer, and—till I saw the paper by my critic in the *Journal of the Statistical Society* I should have said—every statistician does, supply the unknown constants from the sample, which leads us to

$$\chi^2 = \frac{M(ab - cd)^2}{(a + b)(a + c)(b + d)(c + d)} = M\phi^2$$

as used in my memoir of 1912\*.

The problem I had and still have in view is the variability in samples of definite size—with no other restriction than sample size. The solution of that problem is absolutely comparable with that of any discussion of the probability of an observed result in the theory of probable errors. We have in the bulk of such cases constants involved which concern the distribution in an unknown population, and we supply those constants from the sample itself.

As I have already noted the probable error of a mean is

$$\frac{.67449 \sqrt{\mu_2' - \mu_1'^2}}{\sqrt{M}}.$$

By this we understand that the means of samples restricted solely by their size  $M$  from an indefinitely large population of moment-coefficients  $\mu_1'$ ,  $\mu_2'$  about a fixed origin will have a variability determined by the above formula. But when we proceed to give both  $\mu_1'$  and  $\mu_2'$  the values determined from the sample we know, we do *not* add in the manner of my Royal Statistical Society critic, "but in doing so the type of samples is reduced to those having the mean and standard deviation of the sample." If we did, this selection of samples would clearly have no variation of mean or standard deviation at all! In fact probable errors would be meaningless, unless we drew our samples from a population already fully known to us, in which case we should not in 99% of cases want to sample it at all.

In the same way when we use the marginal totals of the sample in formulae like (δ) we do not thereby reduce our samples to those having constant marginal totals, we merely take the best approximation available to the proper value of  $\chi^2$ , and the fact that  $\chi^2$ , as found from the sample, is only an approximation to the true  $\chi^2$  was fully recognised and discussed in my original memoir in the *Philosophical Magazine*.

It only remains to say that the following sentence of my critic's paper seems to me based upon a fallacious principle and apparently flows from a disregard of the nature of probable errors in general.

"It should be pointed out that certain of Pearson's *Tables for Statisticians and Biometricians*, namely Tables XVII, XIX and XX, together with XXII (*Abac* to determine  $r_p$ ) are all calculated

\* On a novel method of regarding the association of two variates classed solely in alternative categories. *Drapers' Company Research Memoirs*, Cambridge University Press.

on the assumption that  $p'=4$  in fourfold tables, and consequently should not be used when, as is almost always the case, the marginal totals are obtained from the data" (*loc. cit.* p. 91).

I hold those tables are quite correctly calculated for  $n'=4$ , and those who attempt to modify them by assuming  $n'=2$  will be dealing with an entirely different problem. Namely, they will be considering not the improbability of the given sample as one of all possible samples of the given size, which it really is, but one of the indefinitely smaller number of samples that have fixed marginal totals. We do not find the probable error of  $r$  for a tetrachoric table\* on the assumption that the marginal totals are fixed. We find it on the assumption that the marginal totals also vary from sample to sample, and when we have found it, then we substitute in the result the values of not only the marginal totals, but the cell-contents,  $a, b, c, d$ , of the sample itself for those of the unknown population. With  $\chi^2$  we go through an exactly similar process of reasoning. If by this procedure we in some mysterious manner tied our degrees of freedom down to the values of the cell-contents used in our formula and adopted from our sample there could be no probable error for  $r$ , for the values of  $a, b, c$ , and  $d$  are all required and used. I trust my critic will pardon me for comparing him with Don Quixote tilting at the windmill; he must either destroy himself, or the whole theory of probable errors, for they are invariably based on using sample values for those of the sampled population unknown to us. For example here is an argument for Don Quixote of the simplest nature: In the  $s$ th category of a population  $N$  the frequency is  $n_s$ , a sample shows  $m_s$  in a total  $M$ . The standard deviation of this frequency is

$$\sqrt{M \frac{n_s}{N} \left(1 - \frac{n_s}{N}\right)}.$$

But we don't know the population sampled and accordingly obtain an approximate value of the above standard deviation by writing for  $\frac{n_s}{N}$ ,  $\frac{m_s}{M}$  and taking for the standard deviation of  $m_s$ ,  $\sqrt{m_s \left(1 - \frac{m_s}{M}\right)}$ . In doing this it is not a question even of using a marginal total, we have used a cell frequency found from our sample. We have therefore according to our critic reduced our possibilities of freedom by selecting out of all possible samples those with  $m_s$  in the  $s$ th cell—this is exactly parallel to our reducing our freedom by "fixing" marginal proportions or moment-coefficients. But if  $m_s$  be fixed, it is ridiculous to talk of a variation of the  $m_s$  frequency. Therefore either  $m_s=0$  or  $m_s=M$ , or the usual theory and practice of probable errors are wholly at fault. I think this will illustrate what I mean by Don Quixote and the windmill.

## II.

*Is Tuberculosis to be regarded from the Aetiological Standpoint as an acute disease of Childhood?* By Dr KR. F. ANDVORD (Christiania). *Tubercle*, Vol. III. No. 3, December, 1921.

This paper is, we must confess, unconvincing. The author holds that in a community that has long been subject to tuberculosis the time of infection should be fixed in the infantile years for the great majority of cases and consequently we should protect children for the first three or four years from infection.

As evidence of his views he takes a graph of what he calls a "population frame" which is really the well-known "number living in a stationary population" ( $L_x$ ) and represents within this graph the numbers dying from tuberculosis and the numbers who have suffered from it at each age. We are doubtful if his graphs for deaths are correctly drawn. They are made to rise suddenly for about a year and then fall till age 7 but we suspect that they should fall from birth till age 7. We cannot justify his chart (No. VIII) which gives the whole population and the

\* *Phil. Trans.* Vol. 195 A, p. 14.

tubercular population. The non-tubercular found by this chart actually increase after age 17 for many years so that the non-tubercular not only have no mortality but are increased by some process of resurrection! Admittedly the chart is hypothetical but as it stands it calls for amendment.

Dr Andvord's remark that "one would hardly gather from these per-thousand curves," i.e. from rates of mortality for various ages, "that, as is really the case, more persons die from tuberculosis in the first and second years of life than in any subsequent age period" seems to betray an inexperience in matters related to a life table: this weakness is shown elsewhere, e.g. p. 102, where deaths are stated without populations and without reference to age distributions.

Dr Andvord may have other evidence in support of his views but the article under review does not justify them statistically; we think every point he brings out could be explained as well on other hypotheses. He cannot, moreover, completely prove his case till he has studied communities which become subject to infection after having been kept free from it. For if his theory be correct, the measures he proposes would necessarily produce such a community.

W. PALIN ELDERTON.