20

THE GOODNESS OF FIT OF REGRESSION FORMULAE AND THE DISTRIBUTION OF REGRESSION COEFFICIENTS

Author's Note (CMS 6.596a)

Paper 19 had shown how χ^2 could be used correctly to test the goodness of fit of frequencies. It was natural to follow it by an investigation of the goodness of fit of regression lines. This is a more difficult problem, and a more maturely written paper. It is shown that the χ^2 distribution supplies only an approximation, the true distribution being that later known as z or F in the analysis of variance. Before its general applicability was recognised the z distribution kept turning up unexpectedly. Its relationships and uses were first summarised (Paper 34) in 1924. It is treated here as a modified χ^2. It is shown that the method may be extended to non-linear regression, and enables a correct interpretation to be put on the "correlation ratio."

Section 6 takes up a second topic, connected with the first only by arising also in regression data. It is shown that the significance of the coefficients of regression formulae, linear or non-linear, simple or multiple, may be treated exactly by "Student's" t-test.

 Journal of the Royal Statistical Society, 85: 597-612, (1922).

The Goodness of Fit of Regression Formulæ, and the Distribution of Regression Coefficients.

By R. A. Fisher, M.A.

Introduction.

The widespread desire to introduce into statistical methods some degree of critical exactitude has led to the employment, now general in careful work, of the two types of quantity which characterize modern statistics, namely, the "probable error" and the test of "goodness of fit." The test of goodness of fit was devised by Pearson, to whose labours principally we now owe it, that the test may readily be applied to a great variety of questions of frequency distribution. It is an essential means of justifying *a posteriori* the methods which have been employed in the reduction of any body of data. Slutsky and Pearson have extended the test to apply also to the fitness of regression formulæ, Pearson's correlation ratio having also been employed for this purpose.

It has been shown in a previous communication [2 Fisher, 1922] that the χ^2 test of goodness of fit can be accurately applied only if allowance is made for the number of constants fitted in reconstructing the theoretical population. This correction is particularly important in contingency tables, but is necessary in all cases; and the fact that it has not been recognized has led to the adoption of erroneous values in almost all the cases in which tests of goodness of fit have been employed. The values of P have been exaggerated, and it is to be feared that in many cases wrong conclusions have been drawn from the values of P obtained.

It has, therefore, been necessary to extend the examination to the tests of goodness of fit of regression lines. The errors due to neglecting the number of constants fitted are here very pronounced; but in addition other points have to be taken into consideration, which did not arise in our previous investigation. In the most important class of cases the curve of distribution of χ^2 is now no longer of the Pearsonian Type III, which is the basis of Elderton's tables, but of the neighbouring Type VI. Certain misconceptions also exist as to the form of the distribution of the correlation

ratio, η, which we hope to have cleared up. We have also taken the opportunity of solving the outstanding problem of the distribution of the regression coefficients in small samples.

1. *The accurate application of Elderton's tables.*

With any two variables x and y we shall suppose that the number of observations for which $x = x_p$ is n_p, and the number of these for which $y = y_q$ is n_{pq}; also that $\bar{y}_p$ is the mean of the observed values of y for a given value of x, so that

$$n_p \bar{y}_p = S_p(n_{pq} y_q).$$

We may regard the group n_p as a random sample from a population in which the value of x is constant; but the value of y varies freely about a certain mean, m_p, with a certain standard deviation, σ_p.

For such samples of n_p, therefore, the mean, $\bar{y}_p$, will vary about the same mean m_p, and since this mean of $\bar{y}_p$ is independent of the number of the array, m_p will be the mean of all values of $\bar{y}_p$ from random samples, however the number n_p may vary.

Any opinion put forward by Professor Pearson is worthy of respect; but it is impossible to agree with his statement [1, p. 240] that "This result cannot be taken as obvious, as the size of the "array in the sample varies." The *fact*, however, Pearson has verified for large samples as far as the third order of approximation. The difference in principle is of some importance, since the simplicity of many of the results here obtained is a consequence of the fact that we have not attempted to eliminate known quantities, given by the sample, from the distribution formulæ of the statistics studied, but only the unknown quantities—parameters of the population from which the sample is drawn—which have to be estimated somewhat inexactly from the given sample.*

Next, for arrays of any given size, the standard deviation of $\bar{y}_p$ is $\sigma_p / \sqrt{n_p}$, and it will be normally distributed if the population-array be normal, and approximately so in most cases if n_p be large. Pearson rightly points out that the values of $\bar{y}_p$ for arrays of different sizes will not be normally distributed, but the distribution will be markedly leptokurtic even for considerable arrays. This result follows from the fact that the distribution is a mixture of

* Statistics whose sampling distribution depends upon other statistics given by the sample cannot, in the strict sense, fulfil the Criterion of Sufficiency. In certain cases evidently no statistic exists which strictly fulfils this criterion. In these cases statistics obtained by the Method of Maximum Likelihood appear to fulfil the Criterion of Efficiency; the extension of this criterion to finite samples thus takes a new importance.

normal distributions, having the same mean, but different standard deviations. This mixed distribution need not concern us, however, for in applying tests of fitness we do not in practice ignore the size of the array. The simple fact is, that, when the population arrays are normal, the quantity

$$z_p = \sqrt{n_p}\,(\bar{y}_p - m_p)$$

is normally distributed about zero, with a standard deviation σ_p, and this distribution is independent of the size of the array.

In the case when the population arrays are equally variable, σ_p is constant [$= \sigma$], and if there are a arrays, the quantity

$$S(z_p^2) = S\{n_p(\bar{y}_p - m_p)^2\}$$

is the sum of the squares of a independent, normally and equally variable quantities, and consequently, if we write

$$\chi^2\sigma^2 = S(z_p^2),$$

χ^2 will be distributed as is the ordinary measure of goodness of fit. In applying Elderton's tables we must, of course, put n' equal to one more than the number of degrees of freedom, as I have demonstrated elsewhere [2]. If the values of m_p were known *a priori*, we should take $n' = a + 1$, but for regression formulæ fitted to the data by equations linear in y_p we merely reduce the number of degrees of freedom by the number of constants fitted. Thus, if m_p is a linear function of x, and a straight line is fitted, we have $n' = a - 1$, and the value of χ^2 then constitutes a test of whether or not m_p is in reality adequately represented by a linear function of x. Similarly, if a cubic polynomial in x be fitted, we have $n' = a - 3$.

2. *The exact distribution of χ^2 when σ is determined from the data.*

So far the results are exact on the assumption that σ is known; but as in practice σ must usually be obtained from the data, errors will be introduced from this source which necessarily influence the distribution of χ^2. It is true that σ may be estimated from the whole data, and is therefore known with accuracy of a higher order than the quantities which contribute to χ^2; nevertheless it is necessary to determine what aberrations are to be expected when the data are not very numerous.

From each array we can directly calculate the second moment s_p^2, and it has been shown [3] that the second moment of a normal sample of n_p is so distributed that the frequency with which it falls into the range ds_p^2 is proportional to

$$\sigma^{-(n_p-1)}\,(s_p^2)^{\frac{n_p-3}{2}}\;e^{-\frac{n_p s_p^2}{2\sigma^2}}\,d\,(s_p^2);$$

the chance that all the observed values of s_p^2 fall in assigned ranges is the product of a such quantities, for all are distributed independently; consequently to find the optimum value of σ, which will also be the value with the least probable error, we must make this product a maximum for variations of σ.

Taking logarithms and differentiating, we have

$$\frac{\delta L}{\delta \sigma} = \frac{S(n_p s_p^2)}{\sigma^3} - \frac{S(n_p - 1)}{\sigma},$$

whence the optimum value of σ^2 is s^2 where

$$(N - a)\, s^2 = S(n_p s_p^2).$$

We shall, therefore, suppose that σ is estimated by this method, and that

$$\chi^2 = \frac{S(z_x^2)}{s^2};$$

we must now find the distribution of this statistic.

The distribution of s^2 is of the same kind as those with which we have been concerned. For

$$S(n_p s_p^2) = S(y - \bar{y}_p)^2,$$

and may be regarded as the sum of the squares of N equally variable quantities, independent save for a linear restrictions of the form

$$S_p(y) = n_p \bar{y}_p.$$

If, therefore, we specify the distribution of s^2 in such a way as to express the frequency element, df, in terms of the variate element within which it occurs, we shall have

$$df \propto t^{\frac{1}{2}(N-a-2)}\, e^{-\frac{t}{2\sigma^2}}\, dt,$$

where t stands for $s^2(N - a)$. In the same way if τ stand for $\chi^2 s^2$ we have, if $p + 1$ constants have been used in fitting,

$$df \propto \tau^{\frac{1}{2}(a-p-3)}\, e^{-\frac{\tau}{2\sigma^2}}\, d\tau;$$

and these two distributions are independent, for the one depends only on the deviations from the means of normal samples, and the other only on the means.

The distribution of χ^2 will now be that of $(N - a)\frac{\tau}{t}$, so, substituting

$$\tau = \frac{\chi^2 t}{N - a}$$

in

$$t^{\frac{N-a-2}{2}}\,\tau^{\frac{a-p-3}{2}}\,e^{-\frac{t}{2\sigma^2}(t+\tau)}\,dt\,d\tau,$$

we obtain, ignoring constants,

$$(\chi^2)^{\frac{a-p-3}{2}}\,t^{\frac{N-p-3}{2}}\,e^{-\frac{t}{2\sigma^2}\left(1+\frac{\chi^2}{N-a}\right)}\,dt\,d\chi^2,$$

and so, on integrating from 0 to ∞ with respect to t,

$$(\chi^2)^{\frac{a-p-3}{2}}\left(1+\frac{\chi^2}{N-a}\right)^{-\frac{N-p-1}{2}}d\chi^2.$$

The variation in s^2, therefore, changes the exact form of the distribution curve for χ^2 from Type III to Type VI. The change is, however, very small if N be large, for as N increases

$$\left(1+\frac{\chi^2}{N-a}\right)^{-\frac{N-p-1}{2}} \to e^{-\frac{1}{2}\chi^2},$$

and so reproduces the Type III distribution.

3. *The nature of the approximation of the Type VI curve*

$$df = \frac{(N-a)^{-\frac{a-p-1}{2}}\cdot\frac{N-p-3}{2}!^*}{\frac{N-a-2}{2}!\,\frac{a-p-3}{2}!}\,x^{\frac{a-p-3}{2}}\left(1+\frac{x}{N-a}\right)^{-\frac{N-p-1}{2}}dx$$

to the Type III curve

$$df = \frac{2^{-\frac{a-p-1}{2}}}{\frac{a-p-3}{2}!}\cdot x^{\frac{a-p-3}{2}}e^{-\frac{1}{2}x}dx.$$

When x is small, the two curves have closely similar forms, the latter being the distribution of χ^2, as given by Elderton's table, when $n' = a - p$. The ratio of the ordinates at the terminus of the curve is obtained by expanding the constant multiplier of the first curve in powers of N^{-1}. It reduces to

$$1+\frac{(n'-1)(n'-3)}{4N}$$

for high values of P; therefore, 1 — P, as given by Elderton's table,

* The symbol $x!$ is used throughout this paper as equivalent to $\Gamma(x+1)$, whether x is an integer or not.

may be corrected by multiplying by this factor. Near the centre of the curve we may observe the position of the mean and mode.

	Mean.	Mode.
Type III ...	$a - p - 1$	$a - p - 3$
Type VI ...	$(a-p-1)\frac{N-a}{N-a-2}$	$(a-p-3)\frac{N-a}{N-a+2}$

The mean, therefore, is raised and the mode lowered in about the same proportion. For the higher values of x the curves are not closely similar, and since it is for these especially that the value of P is required, we shall obtain the necessary correction in P, as far as the terms in N^{-1}. The ratio of the ordinates is

$$1 + \frac{1}{4N}\{x^2 - 2(n'-1)x + (n'-1)(n'-3)\};$$

but, since

$$2^{\frac{n'-1}{2}} \cdot \frac{n'-3}{2}! \, P_{n'}(x) = \int_x^\infty x^{\frac{n'-3}{2}} e^{-\frac{1}{2}x} dx,$$

we have the correction

$$\frac{1}{4N}\{(n'-1)(n'+1)P_{n'+4} - 2(n'-1)^2 P_{n'+2} + (n'-1)(n'-3)P_{n'}\}$$

$$= \frac{n'-1}{4N}\{(n'+1)P_{n'+4} - 2(n'-1)P_{n'+2} + (n'-3)P_{n'}\},$$

which, in the absence of tables of the Type VI curve, will usually be found adequate.

4. *The correlation ratio.*

We are now in a position to make an accurate use of the correlation ratio, as a test of the fitness of regression formulæ. Let Y be the function of x used as regression formula, and let

$$NR^2 s_y^2 = S\{n_p(Y_p - \bar{y})^2\}, \quad Ns_y^2 = S(y-\bar{y})^2$$

where $\bar{y}$ is the mean of all the observed values of y; then it is easy to see that, provided Y has been fitted to the data so that

$$S\{n_p(\bar{y}_p - Y_p)^2\}$$

is a minimum for proportional variations of $Y - \bar{y}$, then

$$N(1-R^2)s_y^2 = SS\{n_{pq}(y-Y_p)^2\}$$

But the correlation ratio is given by the parallel formula,

$$N(1-\eta^2)\,s_y^2 = SS\{n_{pq}(y-\bar{y}_p)^2\} = (N-a)\,s^2;$$

hence, by subtraction,

$$N(\eta^2-R^2)\,s_y^2 = S\{n_p(\bar{y}_p-Y_p)^2\} = \chi^2 s^2$$

$$= \chi^2 \frac{N}{N-a}(1-\eta^2)\,\sigma_y^2.$$

In other words

$$\chi^2 = (N-a)\,\frac{\eta^2-R^2}{1-\eta^2};$$

and to test the significance of $\eta^2 - R^2$ we enter Elderton's table with $n' = a - p$, where $p + 1$ is the number of constants fitted to the regression line. Thus, for a linear regression formula,

$$\chi^2 = (N-a)\frac{\eta^2-r^2}{1-\eta^2},$$

and

$$n' = a-1,$$

using, if necessary, the correction for Type VI as before.

The exact form of the distribution of η itself would be difficult to obtain, but in practice η is usually employed to test the validity of a linear or other regression formula. For this purpose it is not the distribution of η but of the more variable quantity $(\eta^2-R^2)/(1-\eta^2)$ that is required, and the above expressions show it is approximately represented by a Type III curve, and that the probability of a greater discrepancy occurring by chance may be obtained from Elderton's table.

5. *Comparison with previous formulæ.*

Slutsky, in his method [4, p. 83] of treating homoscedastic data, has used a process analogous to that arrived at above, but with four deviations :—

(i) He averages the standard deviations of the arrays, and not their squares, in estimating the value of σ^2.

(ii) He divides his total by N instead of $N - a$.

(iii) He enters Elderton's table with $n' = a + 1$, instead of $n' = a - p$.

(iv) He takes the Type III distribution to be exact.

(i) Pearson [1, pp. 249–51] has criticized the first point, but his practice is not quite explicit. In his opinion evidently, if the surface is homoscedastic, we must take $s_y^2(1-\eta^2)$, but in the special case when the regression is also linear he replaces $1 - \eta^2$

by $1 - r^2$. The point is not one of importance, and I am not convinced that any material difference would be made by replacing $1 - \eta^2$ by $1 - R^2$ in general, when the regression is well fitted. There would seem to be no reason for treating *linear* differently from other regression formulæ. In dealing with Slutsky's price data, where the regression is doubtfully linear, Pearson prefers to use $1 - r^2$.

(ii) The second point is, strictly, a matter of convenience, for when we know the distribution of χ^2, calculated by one method, we also know its distribution in the second case. Since neither of these distributions is exactly the Type III tabled by Elderton, we are free to use whichever we please. The form we have chosen has the advantage of involving the best estimate of σ, and we have chosen it for this reason; but as in the Type VI distribution errors of estimation are completely eliminated, this choice has only the force of a convention. The close agreement of the curve we have obtained with the corresponding Type III in the neighbourhood of the median is a practical advantage; it should in any case be noted that the corrections which we have obtained for P refer only to our own form of the statistic χ^2.

Although strictly a matter of convenience, there is a real advantage when the matter is approached from other points of view, in the use of the best estimates. Thus, for example, when the arrays are undifferentiated, with respect to the distribution of y, we naturally take

$$\frac{1}{N-1} S (y-\bar{y})^2$$

as the best estimate of the variance of the whole of the observation; and as the arrays are undifferentiated this should agree on the average with our estimate of the variance in each array,

$$\frac{1}{N-a} SS \{n_{pq}(y-\bar{y}_p)^2\}.$$

Now

$$(1-\eta^2) S (y - \bar{y})^2 = SS \{n_{pq} (y-\bar{y}_p)^2\};$$

whence it follows that the mean value of $1 - \eta^2$ is

$$\frac{N-a}{N-1},$$

and that of η^2, therefore,

$$\frac{a-1}{N-1}.$$

Pearson has discussed the distribution of η in this case [5].

Observing that, even if the arrays are wholly undifferentiated, $\bar{\eta}$ will necessarily be positive, he points out that, in testing whether η differs significantly from zero, it is not only necessary to know the standard error of η, but also the mean value about which it varies. The standard error of η for undifferentiated arrays he had previously [6] evaluated at $1/\sqrt{N}$, and he then by a somewhat intricate method finds for the mean value of η^2 the value

$$\frac{a-1}{N},$$

and deduces that the mean value of η will be

$$\sqrt{\frac{a-1}{N}},$$

the latter deduction being clearly a slip.

In the case under consideration we have $p=0$, $R=0$, the regression line fitted being $Y=\bar{y}$. Then

$$(N-a)\frac{\eta^2}{1-\eta^2}$$

will be distributed in the Type VI curve

$$df = \frac{(N-a)^{-\frac{a-1}{2}} \cdot \frac{N-3}{2}!}{\frac{N-a-2}{2}!\,\frac{a-3}{2}!}\, x^{\frac{a-3}{2}}\left(1+\frac{x}{N-a}\right)^{-\frac{N-1}{2}} dx;$$

whence substituting for x, we find that η^2 is distributed in the Type I curve

$$\frac{\frac{N-3}{2}!}{\frac{N-a-2}{2}!\,\frac{a-3}{2}!}\,(\eta^2)^{\frac{a-3}{2}}(1-\eta^2)^{\frac{N-a-2}{2}}\,d\eta^2.$$

For large values of N the distribution of η does not tend to normality as Pearson supposed, but that of η^2 tends to a Type III curve. For the mean values of η and η^2 we have

$$\bar{\eta} = \frac{\frac{a-2}{2}!}{\frac{a-3}{2}!} \cdot \frac{\frac{N-3}{2}!}{\frac{N-2}{2}!},$$

or, approximately,

$$\bar{\eta} = \frac{\frac{a-2}{2}!}{\frac{a-3}{3}!}\sqrt{\frac{2}{N}}\left(1+\frac{3}{4N}\right);$$

while

$$\overline{\eta^2} = \frac{a-1}{N-1}$$

in agreement with our previous value.

The mean value for η^2 thus agrees sufficiently with that obtained by Pearson, but the accurate values for the mean and the standard deviation differ from his values. There is no purpose for pressing further a comparison on these lines, since, unless the number of arrays be large, the distribution of η is far from normal, and the significance of an observed value of η may be tested with some accuracy by the use of χ^2.

It may be noticed that, when the number of arrays is large,

$$\sigma_{\eta}{}^2 = \overline{\eta^2} - \bar{\eta}^2 = \frac{1}{2N}\left(1 - \frac{a}{N}\right)$$

to a first approximation, of which the second factor may usually be ignored.

(iii) The third point of difference between my method and those of Slutsky and Pearson, whereby I have made allowance for the number of constants involved in fitting the regression formula, has been more fully explained in a recent paper [2].

It is there shown that if

$$\chi^2 = S\left\{\frac{(n_p - \bar{n}_p)^2}{\bar{n}_p}\right\},$$

where $\bar{n}_p$ is the number of observations expected, and n_p the number observed in any cell, then the value of n' with which Elderton's table should be entered is not the total number of cells, but one more than the number of values of

$$n_p - \bar{n}_p$$

which can be independently specified. That is to say, that when the values of $\bar{n}_p$ are reconstructed from the data of the sample, $(n' - 1)$ is the number of degrees of freedom left after making this reconstruction.

In the same way for regression lines

$$\chi^2 = \frac{1}{s^2} S\{n_p (\bar{y}_p - Y_p)^2\},$$

and, if a is the number of arrays, $n' - 1 = a$, only if the values of Y_p are assigned independently of the sample. If, as more usually is the case, the values of Y_p are those of a regression formula fitted to the sample, the number of values of

$$\bar{y}_p - Y_p$$

which can be independently specified is reduced by the number of constants fitted. For example, if a cubic polynomial has been fitted, the number of degrees of freedom is $(a - 4)$, so that $n' = a - 3$.

6. *The distribution of regression coefficients.*

Hitherto we have only considered data in which a number of values of y are observed corresponding in groups to identical values of x; little statistical or physical data is strictly of this form, although the former may in favourable cases be confidently grouped, so as to simulate the kind of data for which the fitness of regression lines may be tested. The limitation of our methods to data of this form constitutes one of the most serious deficiencies in the statistical methods so far available. The position is well stated by Pearson [1, p. 258]:—

> "Of course it is needful for a test of this kind that the "number of measurements of A, 'the dependent variable,' "should considerably exceed the number of values of B tested. "It would fail entirely if only one value of A were taken for "each value of B, however numerous the latter might be. "We must have some basis on which to determine the error "made in a single determination of A. This is a point, "I think, often overlooked by the physicist. A fairly good "determination—I mean a quantitation determination—of the "goodness of fit of theory to observation could be made from "ten series of eight observations of A corresponding to ten "values of B. But no measure of goodness of fit could be "obtained from eighty observations of A corresponding to "eighty values of B, yet the latter system would probably "make the greater appeal to most physicists. I do not see "how quantitively to obtain any measure of the goodness of fit "of theory to observation in the latter method of procedure."

It appears to the writer that the problem is one rather for the statistician than for the physicist; for, given equally variable arrays, and a regression line of known form, the problem is perfectly objective. I emphasize it here as a problem awaiting solution—a manageable solution of which would be of great practical utility. That it is an objective problem is clear from the confidence with which very bad fits will be rejected at sight, as also from the fact that rough and common-sense methods of testing have been developed for some purposes. [9, Fisher, 1921.]

Although exact methods of testing the goodness of fit of regression lines are not available for the extended class of data, we are in a

position to give an exact solution of the distribution of the regression coefficients. This problem has been outstanding for many years; but the need for its solution was recently brought home to the writer by correspondence with " Student," whose brilliant researches [7] in 1908 form the basis of the exact solution.

For consider a simple linear regression formula

$$Y = a + b(x - \bar{x}),$$

of which the coefficients a and b are calculated by the equations

$$a = \bar{y}, \qquad b = \frac{S\,(y\,(x - \bar{x}))}{S\,(x - \bar{x})^2};$$

we note first that a and b are orthogonal functions, in that given the series of values of x observed, their sampling variation is independent.

Now " Student " [7] has shown how the probable error of a may be calculated; for if for a given value of x the standard deviation of y is σ, then a will be normally distributed, so that

$$\sigma_a^2 = \frac{\sigma^2}{n}.$$

So that if α is the population value of a, and $\tau = \frac{a-\alpha}{\sigma}\sqrt{n}$, then τ is normally distributed about zero with standard deviation unity. If σ^2 is unknown, the best estimate that can be made of it from the sample is

$$s^2 = \frac{1}{n-2} S\,(y-Y)^2$$

where the sum is divided by $(n - 2)$ to allow for the two constants, used in fitting the regression line. Then the distribution of s^2 is, if $\chi^2 = (n-2)\frac{s^2}{\sigma^2}$,

$$df = \frac{1}{\frac{n-4}{2}!}\left(\frac{\chi^2}{2}\right)^{\frac{n-4}{2}} e^{-\frac{1}{2}\chi^2}\, d\left(\frac{\chi^2}{2}\right).$$

The distribution of the two quantities s and a are wholly independent; hence, following " Student," we find the distribution of a quantity completely calculable from the sample, namely,

$$z = \frac{\tau}{\chi} = \frac{(a-\alpha)\sqrt{n}}{\sqrt{S\,(y-Y)^2}}.$$

For

$$df = \frac{1}{\frac{n-4}{2}!}\left(\frac{\chi^2}{2}\right)^{\frac{n-4}{2}} e^{-\frac{\chi^2}{2}} d\frac{\chi^2}{2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{\chi^2 z^2}{2}} \cdot \chi \cdot dz$$

$$= \frac{1}{\sqrt{\pi}} \cdot \frac{1}{\frac{n-4}{2}!} \cdot \left(\frac{\chi^2}{2}\right)^{\frac{n-3}{2}} e^{-\frac{1}{2}\chi^2(1+z^2)} d\left(\frac{\chi^2}{2}\right) \cdot dz;$$

and integrating with respect to χ^2 from 0 to ∞, we have

$$\frac{1}{\sqrt{\pi}} \cdot \frac{\frac{n-3}{2}!}{\frac{n-4}{2}!} \cdot \frac{dz}{(1+z^2)^{\frac{n-1}{2}}},$$

the Type VII curve obtained by "Student," with n reduced by unity, since we have fitted a regression line of the first degree.

Similarly, for b,

$$\sigma_b^2 = \frac{\sigma^2}{S(x-\bar{x})^2};$$

and if

$$z = \frac{(b-\beta)\sqrt{S(x-\bar{x})^2}}{\sqrt{S(y-Y)^2}}$$

we arrive at the same distribution as before, β being the population value of the regression coefficient.

The above argument immediately extends itself to regression lines of any form and involving any number of coefficients. For, suppose the regression equation is of the form

$$Y = a + bX_1 + cX_2 + \ldots\ldots kX_p,$$

where X_1, X_2 ..... X_p are orthogonal functions of x for the observed values, so that

$$S(X_a X_b) = 0$$

—in the most important case X_p will be polynomial in x, of degree p, orthogonal to the polynomials of lower degree [9]—then, for example,

$$k = \frac{S(yX_p)}{S(X_p^2)}$$

and

$$\sigma_k^2 = \frac{\sigma^2}{S(X_p^2)}.$$

Also, if

$$s^2 = \frac{1}{n-p-1} \mathrm{S}(\eta - \mathrm{Y})^2$$

the distribution of s is given by

$$df = \frac{1}{\frac{n-p-2}{2}!} \cdot \left(\frac{\chi^2}{2}\right)^{\frac{n-p-2}{2}} e^{-\frac{1}{2}\chi^2} d\left(\tfrac{1}{2}\chi^2\right)$$

where

$$\chi^2 = (n-p-1)\frac{s^2}{\sigma^2}.$$

Consequently, if

$$z = \frac{(k-\kappa)\sqrt{\mathrm{S}(\mathrm{X}_p{}^2)}}{\sqrt{\mathrm{S}(y-\mathrm{Y})^2}}$$

the distribution of z is the Type VII curve

$$df = \frac{1}{\sqrt{\pi}} \cdot \frac{\frac{n-p-2}{2}!}{\frac{n-p-3}{2}!} \cdot \frac{dz}{(1+z^2)^{\frac{n-p}{2}}};$$

and in this case, when $p+1$ constants have been fitted, all the other regression coefficients will be distributed in like manner, only substituting the corresponding function of x for X_p.

Tables of the Probability Integral of the above Type VII distribution have been prepared by "Student" [8], for values of $n-p$ from 0 to 30. These tables are in a suitable form for testing the significance of an observed regression coefficient. For larger samples the curve will be sufficiently normal for most purposes, the variance of z being

$$\frac{1}{n-p-3}.$$

The utility of "Student's" curve for the distribution of errors in the mean of a sample, in terms of the standard deviation, as estimated from the same sample, is increased by the circumstance that the same distribution also gives that of differences between such means. Thus, if $\bar{x}$ and $\bar{x}'$ are the means of samples of n and n', and we wish to test if the means are in sufficient agreement to warrant the belief that the samples are drawn from the same population, we may calculate

$$z = \frac{\bar{x}-\bar{x}'}{\sqrt{\mathrm{S}(x-\bar{x})^2+\mathrm{S}'(x'-\bar{x}')^2}} \cdot \sqrt{\frac{nn'}{n+n'}},$$

* For η, read y.

then z will be distributed so that

$$df = \frac{1}{\sqrt{\pi}} \cdot \frac{\frac{n+n'-3}{2}!}{\frac{n+n'-4}{2}!} \cdot \frac{dz}{(1+z^2)^{\frac{n+n'-1}{2}}}.$$

This method of comparison may be applied directly to regression coefficients, when the same series of values of x is observed in each case.

The above problem in which the errors of the coefficients of a regression of any form are considered, is in reality a special case of the multiple regression surface—special in the sense that with a single variable we can conveniently choose the terms of the regression equation, so that the several terms consist of uncorrelated functions. When this is not the case we have such a regression system as

$$Y = b_1 x_1 + b_2 x_2 + \ldots\ldots + b_p x_p$$

when x_1, $x_2 \ldots\ldots x_p$ are p independent variables, with certain mutual correlations. The accuracy of the regression coefficients is only affected by the correlations which appear *in the sample*, so that if we construct the determinant

$$\Delta = \begin{vmatrix} S(x_1^2) & S(x_1x_2) & \ldots\ldots & S(x_1x_p) \\ S(x_1x_2) & S(x_2^2) & \ldots\ldots & S(x_2x_p) \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ S(x_1x_p) & S(x_2x_p) & \ldots\ldots & S(x_p^2) \end{vmatrix}$$

from the values of the sample, then

$$\sigma_{b_1}^2 = \frac{\sigma^2 \Delta_{11}}{\Delta}$$

where Δ_{11} is the minor of $S(x_1^2)$.

Consequently, if

$$z = \frac{(b_1 - \beta_1)\sqrt{\Delta}}{\sqrt{S(y-Y)^2}\sqrt{\Delta_{11}}},$$

then, as before, z will be distributed in the Type VII distribution

$$df = \frac{1}{\sqrt{\pi}} \cdot \frac{\frac{n-p-2}{2}!}{\frac{n-p-3}{2}!} \cdot \frac{dz}{(1+z^2)^{\frac{n-p}{2}}}.$$

Conclusions.

(1) In testing the fitness of regression lines account must be taken of the number of degrees of freedom which have been absorbed in the process of fitting.

(2) The Type III distribution of Elderton's tables is not exact for testing regression lines, but the tables may be used as a basis of a useful approximation.

(3) The exact distribution of χ^2 is given by a curve of the Pearsonian Type VI, which for large samples approaches the Type III distribution.

(4) For undifferentiated arrays the distribution of η^2 is given by a curve of the Pearsonian Type I; for large samples this curve approaches the Type III distribution.

(5) The distribution in random samples of a great variety of regression coefficients may be treated by the method introduced by " Student " for the distribution of the mean of a normal sample, and as in that case lead to a distribution curve of the Pearsonian Type VII, which for large samples rapidly approaches normality.

The importance of the last result is considerable. It shows that a number of regression coefficients may be safely calculated from a sample of moderate size. Thus, in studying relations of a complex kind, such as occur in agricultural meteorology, it is useful to know that we may as accurately determine thirty coefficients from a sample of sixty sets of observations as we may calculate a single coefficient, or mean, from a sample of thirty-one observations.

References.

1. K. Pearson (1916).—" On the Application of Goodness of Fit Tables to Test Regression Curves and Theoretical Curves used to describe Observational or Experimental Data." *Biom.*, XI, 239–61.

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

3. R. A. Fisher (1915).—" Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population." *Biom.*, X, 507–21.

4. E. Slutsky (1913).—" On the Criterion of Goodness of Fit of the Regression Lines, and on the best Method of fitting them to the Data." *J.R.S.S.*, LXXVII, 78–84.

5. K. Pearson (1911).—" On a Correction to be made to the Correlation Ratio." *Biom.*, VIII, 254–6.

6. K. Pearson (1905).—" On the General Theory of Skew Correlation and Non-linear Regression." *Drapers' Company Research Memoirs*: Dulau and Co.

7. Student (1908).—" The Probable Error of a Mean." *Biom.*, VI, pp. 1–25.

8. Student (1917).—" Tables for Estimating the Probability that the Mean of a unique Sample of Observations lies between $-\infty$ and any given Distance of the Mean of the Population from which the Sample is drawn." *Biom.*, XI, 414–17.

9. R. A. Fisher (1921).—" An Examination of the Yield of Dressed Grain from Broadbalk." *Journal of Agricultural Science*, XI, 107–35.