43

R. A. FISHER, M. A.
Rothamsted Experimental Station

Applications of "Student's„ Distribution.

1. · Introductory.

The Theory of Errors may be said to have taken its origin in the fact that the accuracy of the mean of a number of observations may be estimated from the discrepancies observed among the individual values used in obtaining the mean. In the simple theory appropriate to samples drawn from a normal population, of which the variance (mean squared deviation) is σ^2, it is easy to show that the mean of n' observations will be distributed normally with variance

$$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n'}$$

consequently, if σ^2 were known *a priori*, the sampling distribution of the mean would be fully known. To test any hypothesis respecting the mean of the population, as for example that the mean of the population was any assigned value m, we should merely need to calculate

$$t = (\bar{x} - m)\frac{\sqrt{n'}}{\sigma}$$

and the probability integral

$$P = \frac{1}{\sqrt{2\pi}}\int_t^\infty e^{-\frac{1}{2}t^2}\,dt$$

would give, with any degree of accuracy required, the probability, on that hypothesis, that a greater discrepancy should occur than that actually observed. If the value of P so calculated turned out to be a small quantity such as 0·01, we should conclude with some confidence that the hypothesis was not in fact true of the population actually sampled.

2. - " Student's „ Distribution.

In the majority of cases in which such tests are required we have no *a priori* knowledge of the variance of the population, or indeed of whether its distribution is normal or not. The first point (which alone we shall consider in detail) was met by estimating the variance of the population from the sample itself; if $x_1, x_2, \ldots . x_{n'}$, be our observations, we may take

$$s^2 = \frac{S(x - \bar{x})^2}{n' - 1}$$

as an estimate of the unknown variance, σ^2.

s, so calculated, is a perfectly good estimate of σ, but it is seldom or never equal to σ, and, as was first pointed out by « Student » in his fundamental paper of 1908, (1), if in testing significance we substitute s for σ, and calculate

$$t = (\bar{x} - m)\frac{\sqrt{n'}}{s}$$

we have no right to assume that t will still be distributed in the normal curve, or that the significance of an observation can be accurately tested by the normal probability integral. In order to obtain an accurate test « Student » investigated the distribution of s^2 in random samples; from the relation connecting the first four moments he inferred that it was probably of the Pearsonian Type III, and so obtained the exact distribution, which may be written

$$df = \frac{1}{\left(\frac{n'-3}{2}\right)!}\left(\frac{n'-1}{2\sigma^2}\right)^{\frac{n'-1}{2}} \cdot (s^2)^{\frac{n'-3}{2}} e^{-\frac{(n'-1)s^2}{2\sigma^2}} d(s^2)$$

$$= \frac{1}{\left(\frac{n-2}{2}\right)!}\left(\frac{n}{2\sigma^2}\right)^{\frac{1}{2}n} (s^2)^{\frac{n-2}{2}} e^{-\frac{ns^2}{2\sigma^2}} d(s^2)$$

where $n = n' - 1$.

Assuming that the distribution of s^2 is independent of that of $\bar{x}$ this expression may be used to deduce the exact distribution of

$$t = (\bar{x} - m)\frac{\sqrt{n'}}{s} \quad ;$$

for the distribution of $\bar{x}$ is given by

$$df = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{n'(\bar{x}-m)^2}{2\sigma^2}} \cdot \frac{\sqrt{n'}}{\sigma}\, d\bar{x} \quad ,$$

or, for a given value of s by,

$$df = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{t^2 s^2}{2\sigma^2}} \cdot \frac{s}{\sigma}\, dt \quad ;$$

so that for all values of s

$$df = \frac{1}{\left(\frac{n-2}{2}\right)!\,\sqrt{2\pi}} \left(\frac{n}{2\sigma^2}\right)^{\frac{1}{2}n} \cdot \frac{dt}{\sigma} \int_0^\infty (s^2)^{\frac{n-1}{2}}\, e^{-\frac{(n+t^2)s^2}{2\sigma^2}}\, d(s^2)$$

$$= \frac{\left(\frac{n-1}{2}\right)!}{\left(\frac{n-2}{2}\right)!\,\sqrt{\pi n}} \left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}} dt \,.$$

This is equivalent to the form given by « Student » in 1908 (1). The result obtained at that time was partly intuitive, since on two points the demonstration was incomplete; (*i*) the distribution obtained for s^2 agrees with the true distribution in the first four moments, but might conceivably have differed from it in the higher moments. (*ii*) « Student » demonstrated that s^2 was not correlated with $(\bar{x}-m)^2$, but did not show that the two distributions were entirely independent.

3. - Proof of the exactitude of " Student's „ Distribution for Normal Samples.

One method of proving these two points, which has been found to be valuable in other sampling problems, is to consider the observations $x_1, \ldots, x_{n'}$ as rectangular coordinates in Euclidian space of n' dimensions; the volume element in this space will be

$$dv = dx_1\, dx_2 \ldots\ldots,\ dx_{n'},$$

and since the individual observations are independently distributed so that

$$df = \frac{1}{\sigma\ \sqrt{2\pi}}\, e^{-\frac{(x-m)^2}{2\sigma^2}}\, dx \quad ,$$

the frequency element for the sample will be

$$df = \frac{1}{(\sigma\sqrt{2\pi})^{n'}} e^{-\frac{1}{2\sigma^2} S(x-m)^2} dv \ .$$

The density with which samples occur in any region is therefore proportional to

$$e^{-\frac{1}{2\sigma^2} S(x-m)^2} = e^{-\frac{n'}{2\sigma^2}(\bar{x}-m)^2} \cdot e^{-\frac{(n'-1)s^2}{2\sigma^2}} .$$

To find the simultaneous distribution of $\bar{x}$ and s^2 it is necessary to express the element of volume dv in terms of these two statistics. Observing that $\bar{x}$ is proportional to the distance of the sample point from a fixed « plane » region,

$$S(x) = 0 \ ,$$

and that s is proportional to the distance from the fixed line,

$$x_1 = x_2 = x_3 = \ldots\ldots = x_{n'} \ ,$$

it follows that the element of volume, dv, varies as

$$s^{n'-2} \, ds \, d\bar{x},$$

so that df varies as

$$e^{\frac{n'}{2\sigma^2}(\bar{x}-m)^2} d\bar{x} \ . \ s^{n'-2} \ e^{-\frac{(n'-1)s^2}{2\sigma^2}} ds \ .$$

Since this falls into two factors involving $\bar{x}$ and s respectively, the two distributions must be wholly independent, and completing each with its necessary constant factors, we have

$$df = \frac{1}{\sqrt{2\pi}} e^{-\frac{n'}{2\sigma^2}(\bar{x}-m)^2} \frac{\sqrt{n'}}{\sigma} d\bar{x}$$

$$\times \frac{1}{\left(\frac{n'-3}{2}\right)} \left(\frac{n'-1}{2\sigma^2}\right)^{\frac{n'-1}{2}} (s^2)^{\frac{n'-3}{2}} e^{-\frac{(n'-1)s^2}{2\sigma^2}} d(s^2) \ ,$$

the simultaneous distribution from which the distribution of t has already been derived.

4. - Conditions for the wider application of " Student's „ distribution.

Although, for the solution of the specific problem attacked by « Student », the simultaneous distribution of $\bar{x}$ and s^2 is all that is required, yet the fact that this distribution is compounded of two *independent* distributions, (i) that of $\frac{(\bar{x} - m)\sqrt{n'}}{\sigma}$, distributed normally about zero with unit standard deviation, and (ii) that of

$$\frac{(n' - 1)\, s^2}{\sigma^2} \; (= \chi^2),$$

in the distribution

$$df = \frac{1}{\frac{n' - 3}{2}!} \left(\frac{\chi^2}{2}\right)^{\frac{n' - 3}{2}} e^{-\frac{1}{2}\chi^2} d\left(\tfrac{1}{2}\chi^2\right)$$

in such a way that

$$t = \frac{(\bar{x} - m)\sqrt{n'}}{\sigma} \div \sqrt{\frac{\chi^2}{n' - 1}} \quad ,$$

shows that « Student's » formula for the distribution of t is applicable to all cases which can be reduced to a comparison of the deviation of a normal variate, with an independently distributed estimate of its standard deviation, derived from the sums of squares of homogeneous normal deviations, either from the true mean of the distribution, or from the means of samples.

5. - Significance of differences between means.

This statistical situation occurs very frequently in connection with experimental work; and, consequently. « Student's » distribution affords the solution of a variety of problems beyond that for which it was originally prepared. Of these, one that appears continually under one form or another is the comparison of two mean values. If $\bar{x}_1$ and $\bar{x}_2$ are the means of two samples of n_1 and n_2 values respectively, and we wish to test if the two means are sufficiently alike to warrant the belief that the samples are drawn from the same population, or, on the other hand if the means are significantly different, we may suppose the hypothetical population to have a standard deviation, σ. Then

$$\frac{\bar{x}_1 - \bar{x}_2}{\sigma} \quad . \quad \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

will be normally distributed with unit standard deviation; further,

$$\frac{\overset{n_1}{S}(x_1 - \bar{x}_1)^2 + \overset{n_2}{S}(x_2 - \bar{x}_2)^2}{\sigma^2} = \frac{S_1 + S_2}{\sigma^2}$$

will be distributed in the χ^2 distribution for $n = n_1 + n_2 - 2$, (or $n' = n_1 + n_2 - 1$); moreover, these two distributions will be wholly independent. Consequently, if we write

$$t = \frac{(\bar{x}_1 - \bar{x}_2)\sqrt{n_1 + n_2 - 2}}{\sqrt{S_1 + S_2}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

then t will be distributed in « Student's » distribution, specified by

$$n = n_1 + n_2 - 2 \ .$$

Example 1.

As an example of the application of this method in experimental work we may take a portion of the data of an electro-culture experiment carried out at Rothamsted in 1922. Eight pots growing three barley plants each, were exposed to the action of a high tension discharge, while nine similar pots were enclosed in an earthed wire cage. The numbers of tillers in each pot were as follows: —

Electrified	16, 16, 20, 16, 20, 17, 15, 21	mean $17.625 = \bar{x}_1$
Caged	17, 27, 18, 25, 27, 29, 27, 23, 17	mean $23.333 = \bar{x}_2$

The difference between the means is therefore 5.708; also

$$S_1 = 37.875, \ S_2 = 184, \ S_1 + S_2 = 221.875$$

multiplying $(S_1 + S_2)$ by 17 and dividing successively by 15, 8 and 9, we have as the estimated variance of the difference between the means, 3.4925; and for the estimated standard deviation, 1.8688. The value of t, which is the ratio of the difference to its estimated standard deviation is therefore 3.054. For $n = 15$, Table I shows that this value will be exceeded by chance about 41 times in 10,000. In other words a difference, positive or negative, greater than that observed will occur by chance only about 8 times in a thousand trials. The difference must therefore be judged significant. The two series are definitely unlike in their tillering; the possibility, however, that unlikeness in the variability as well as unlikeness in the mean, has contributed to the result is not excluded by this test. The possibility that samples from populations alike in their mean, but unlike in

their variability should give significant values of t is scarcely of importance in relation to the practical use of the test in experimental work.

In cases where the two samples are equal in number, and in which each individual of one sample corresponds in some way to a particular individual of the second sample, we may test the significance of the difference of the means, either by the above method, or by the method as originally set forth by « Student ». In the latter case the differences between corresponding values are considered as a single sample, and the test shows if their mean differs significantly from zero. When both methods are available, sometimes the one and sometimes the other is the more sensitive; if either shows a significant deviation its testimony cannot be ignored. If, as frequently happens in experimental work the corresponding values of the two samples are positively correlated, the standard deviation of the differences will be reduced by this circumstance; against this advantage we must set off the fact that in treating the results as a single sample, the value of n is only half as great as if the two samples had been treated separately. The results of applying both tests to the same data supply a direct statistical measure of the efficacy of the system of « controls » which has been utilised.

6. - Significance of regression coefficients.

The second class of tests for which « Student's » distribution provides an exact solution, lies in testing the significance of the large class of statistics known as regression coefficients; and also in testing the significance of differences between regression coefficients obtained in different samples.

Consider a simple linear regression formula

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

in which the coefficients a and b have been calculated by the equations

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

if, for a given value of x, y is distributed normally with variance σ^2 then *confining attention to samples having the same values of* x, it is evident that b will be distributed normally with variance

$$\sigma^2{}_b = \frac{\sigma^2}{S\,(x - \bar{x})^2}$$

moreover, the mean of this distribution will be the value of the regression, say, β, obtained from an infinitely large sample. In other words

$$\frac{(b-\beta)\sqrt{S(x-\bar{x})^2}}{\sigma}$$

will be normally distributed about zero with unit standard deviation.

This expression involves the unknown parameter, σ, and it is by substituting for σ an estimate of its value, s, derived from the observations, that the distribution is changed from the normal form to that of « Student's » distribution. The estimate which we shall consider is ,

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

where n' is the number of observations. We shall prove that $\frac{1}{\sigma^2} S(y-Y)^2$ is distributed as is the sum of the squares of $n'-2$ quantities distributed independently and normally with unit standard deviation. It is perhaps worth while to give, at length, an algebraical method of proof, since analogous cases have hitherto been demonstrated only geometrically, by means of a construction in Euclidian hyperspace, and the validity of such methods of proof may not be universally admitted.

If $x_1, x_2, \ldots x_{n'}$ be distributed normally and independently with unit standard deviation and if,

$$\begin{aligned}
\zeta_1 &= p_{11}\,x_1 + p_{12}\,x_2 + \ldots\ldots\ldots p_{1n'}\,x_{n'}\\
\zeta_2 &= p_{21}\,x_1 + p_{22}\,x_2 + \ldots\ldots\ldots p_{2n'}\,x_{n'}\\
\zeta_{n'} &= p_{n'1}\,x_1 + p_{n'2}\,x_2 + \ldots\ldots\ldots p_{n'n'}\,x_{n'}
\end{aligned}$$

then $\zeta_1, \zeta_2, \ldots \zeta_{n'}$, will be distributed normally and independently with unit standard deviation, provided for all values of i

$$p^2_{i1} + p^2_{i2} + \ldots\ldots p^2_{in'} = 1$$

and for all unequal values of i and j

$$p_{i1}\,p_{j1} + p_{i2}\,p_{j2} + \ldots\ldots + p_{in'}\,p_{jn'} = 0$$

Now, if $\zeta_1, \zeta_2, \ldots \zeta_h$ be any h linear functions of $x_1, x_2 \ldots x_{n'}$ fulfilling these conditions, the set of h homogeneous equations,

$$\begin{aligned}
p_{11}\,p_{i1} + p_{12}\,p_{i2} + \ldots\ldots\ldots p_{1n'}\,p_{in'} &= 0,\\
p_{h1}\,p_{i1} + p_{h2}\,p_{i2} + \ldots\ldots\ldots p_{nh'}\,p_{in'} &= 0
\end{aligned}$$

involving the n' unknowns $p_{i_1}, p_{i2} \ldots\ldots p_{in'}$, can always be solved if h does not exceed $n' - 1$, and every such solution will yield a solution of

$$p^2_{i_1} + p^2_{i_2} + \ldots\ldots p^2_{in'} = 1 ;$$

consequently we can always find in succession, variates

$$\zeta_{h+1}, \zeta_{h+2} \ldots\ldots \zeta_{n'}$$

fulfilling the required conditions.

From this it follows that

$$\overset{n'}{\underset{1}{S}}(x^2) - \overset{h}{\underset{1}{S}}(\zeta^2)$$

can always be expressed in the form

$$\overset{n'}{\underset{h+1}{S}}(\zeta^2)$$

and is therefore distributed as is the sum of the squares of $n' - h$ quantities normally and independently distributed with unit standard deviation. Moreover this distribution will be wholly independent of $\zeta_1, \ldots\ldots \zeta_h$

Now if $$\alpha + \beta(x - \bar{x})$$

is the regression line of the population sampled, the quantities

$$\frac{1}{\sigma}\left\{y - \alpha - \beta(x - \bar{x})\right\}$$

are normally and independently distributed with unit standard deviation; and, since a and b have been chosen to make

$$\frac{1}{\sigma^2} S\left\{y - a - b(x - \bar{x})\right\}^2$$

a minimum, it may be reduced to the form

$$\frac{1}{\sigma^2} S\left\{y - \alpha - \beta(x - \bar{x})\right\}^2 - \frac{n'}{\sigma^2}(a - \alpha)^2 - (b - \beta)^2 \frac{S(x - \bar{x})^2}{\sigma^2}$$

and will, consequently, be distributed as is the sum of the squares of $(n' - 2)$ quantities, each distributed independently and normally with unit standard deviation, provided that

$$\frac{(a - \alpha)\sqrt{n'}}{\sigma} \text{ and } \frac{(b - \beta)}{\sigma}\sqrt{S(x - \bar{x})^2}$$

are so distributed. Now

$$\frac{(a-\alpha)\sqrt{n'}}{\sigma} = S\left\{\frac{1}{\sqrt{n'}} \cdot \frac{y-\alpha-\beta(x-\bar{x})}{\sigma}\right\}$$

and is of the form required for ζ_1 having

$$p_{11} = p_{12} \quad . \quad . \quad . \quad . \quad . \quad = p_{1n'} = \frac{1}{\sqrt{n'}}$$

satisfying the equation

$$p^2_{11} + p^2_{12} + \quad . \quad . \quad . \quad . \quad . \quad + p^2_{1n'} = 1 \quad ;$$

also

$$\frac{b-\beta}{\sigma} S(x-\bar{x})^2 = S\left\{\frac{x-\bar{x}}{S(x-\bar{x})^2} \cdot \frac{y-\alpha-\beta(x-\bar{x})}{\sigma}\right\} \quad ,$$

which is of the form required for ζ_2, having

$$p_{2i} = \frac{x_i - \bar{x}}{\sqrt{S(x-\bar{x})^2}} \quad ,$$

satisfying both the equations

$$p^2_{21} + p^2_{22} + \quad . \quad . \quad . \quad . \quad . \quad p^2_{2n'} = 1$$

and

$$p_{11}p_{21} + p_{12}p_{22} + \ldots\ldots + p_{1n'}p_{2n'} = S\left\{\frac{1}{n'} \frac{x_i - \bar{x}}{\sqrt{S(x-\bar{x})^2}}\right\} = 0$$

Consequently

$$\frac{1}{\sigma^2} S(y-Y)^2$$

must be distributed in random samples as is the sum of $(n'-2)$ quantities, each distributed independently and normally with unit standard deviation; and this distribution will be wholly independent of that of a and b.

Substituting for σ its estimated value s, we see that

$$t = \frac{(b-\beta)\sqrt{n'-2}\ \sqrt{S(x-\bar{x})^2}}{\sqrt{S(y-Y)^2}}$$

will be distributed in « Student's » distribution for $n = n'-2$. The quantity t involves no hypothetical quantities, being calculable wholly from the observations. It is the point of the method, as of « Student's » original treatment of the probable error of the mean, to obtain a quantity of known distribution expressible in terms of

the observations only. If we had found the distribution of b for samples varying in the values of x observed, we should have been obliged to express the distribution in terms of the unknown standard deviation σ_x in the population sampled; moreover, since σ_x is unknown, we should have been obliged to substitute for it, an estimate based upon $S\,(x - \bar{x})^2$; the inexactitude of the estimate would have vitiated our solution, and required us to make allowance for the sampling variation of $S\,(x - \bar{x})^2$; finally, this process, when allowance had been accurately made would lead us back to the « Student's » distribution found above. The proof given above has, however, the advantage that it is valid whatever may be the distribution of x, provided that y is normally and equally variable in each array, and that regression of y on x is linear in the population sampled.

7. Non-linear Regression.

The same distribution is adequate for the coefficients of a non linear regression formula, still provided that y is normally and equally variable in each array. For suppose the regression equation is of the form

$$Y = \alpha_0 + \alpha_1 X_1, + \alpha_2 X_2 + \ldots\ldots + \alpha_p X_p$$

where $X_o\,(= 1)$, X_1, X_2, $\ldots\ldots$ X_p are orthogonal functions of x, so that for all unequal values of i and j

$$S\,(X_i\,X_j) = 0$$

the summation being taken over the observed values of x. In the most important case X_i will be a polynomial in x of degree i. In any case the coefficient, α_i, will be estimated by means of the equation

$$a_i = \frac{S\,(y\,X_i)}{S\,(X_i^2)}$$

and

$$\sigma^2 a_i = \frac{\sigma^2}{S\,(X^2{}_i)}$$

consequently, by an easy extension of the reasoning used above,

$$t = \frac{(a_i - \alpha_i)\sqrt{n' - p - 1}\ \sqrt{S\,(X^2{}_i)}}{\sqrt{S\,(y - Y)^2}}$$

will be distributed in « Student's » distribution, for $n = n' - p - 1$.

In practice it is quickest to calculate $S(y - Y)^2$ from the relation

$$S(y - Y)^2 = S(y^2) - n' a^2_0 - a^2_1 S(X^2_1) - \ldots\ldots a^2_p S(X^2_p).$$

Alternatively this relation may be used to check the values of a_i obtained.

Example 2.

The difference in yield of total grain between the two dunged plots of Broadbalk field for the years 1901-1923 are as follows. The yield of plot 2B exceeded that of plot 2A by

	lb. per acre.		lb. per acre.		lb. per acre.		lb. per acre.		lb. per acre.
1901	124	1906	133	1911	207	1916	118	1921	106
1902	340	1907	315	1912	196	1917	221	1922	185
1903	146	1908	175	1913	211	1918	272	1923	142
1904	36	1909	321	1914	— 8	1919	410		
1905	127	1910	127	1915	418	1920	140		

Fitting a straight line to these differences, of the form

$$Y = a_0 + a_1 x,$$

where x is the date measured in years from the central year, 1912, we find $a_0 = 194$, $a_1 = 1.247$. Are these significant?

We have $S(y - Y)^2 = 259{,}872$, $S(x^2) = 1012$ so that for the significance of the mean

$$t = \frac{194\sqrt{21}.\sqrt{23}}{\sqrt{259{,}872}} = 8.36;$$

and for the significance of the linear rate of increase

$$t = \frac{1.247\sqrt{21}\sqrt{1012}}{\sqrt{259{,}872}} = .36$$

For each test $n = 21$. It is evident at once that whereas the difference in yield of the plots is clearly significant, there is no sign, during the present century, that the difference has been increasing or decreasing. In view of the latter fact the significance of the mean yields could have been tested satisfactorily without fitting the regression line.

8. Multiple Regression.

The same distribution is equally applicable to the coefficients of a multiple regression surface, plane or curved. In general the surface may be represented by

$$Y - \bar{y} = b_1 x_1 + b_2 x_2 \ldots\ldots + b_p x_p$$

where $x_1, \ldots\ldots, x_p$ are either variates, or functions of variates, measured from their means, in terms of which the regression is expressed. Since $x_1, \ldots\ldots, x_p$ may now be correlated, we shall require the determinant

$$\Delta = \begin{vmatrix} S(x_1^2) & S(x_1 x_2) & \ldots & S(x_1 x_p) \\ S(x_1 x_2) & S(x_2^2) & \ldots & S(x_2 x)_p \\ \ldots & \ldots & \ldots & \ldots \\ S(x_1 x_p) & S(x_2 x_p) & \ldots & S(x_p^2) \end{vmatrix}$$

where the summation is taken over the observed n' values.

Then

$$\sigma^2_{b_i} = \frac{\sigma^2 \Delta_{ii}}{\Delta}$$

where Δ_{ii} is the cofactor of $S(x_i^2)$.
Consequently if

$$t_i = \frac{(b_i - \beta_i)\sqrt{n' - p - 1}\sqrt{\Delta}}{\sqrt{S(y - Y)^2}\sqrt{\Delta_{ii}}}$$

where β_i is the population value corresponding to the observed b_i, then t_i will be distributed in «Student's» distribution for $n = n' - p - 1$.

9. Distributions related to "Student's", as that of χ^2 is to the normal curve.

Finally the probability integral with which we are concerned is of value in calculating the probability integral of a wider class of distributions which is related to «Student's» distribution in the same manner as that of χ^2 is related to the normal distribution. This wider class of distributions appears (i) in the study of intraclass correlations (2) (ii) in the comparison of estimates of the variance, or of the standard deviation from normal samples (3, p. 142) (iii) in testing the goodness of fit of regression lines (4) (iv) in testing the signi-

ficance of a multiple correlation (5), or (v) of a correlation ratio (3, p. 218).

For example, the distribution in random samples of a multiple correlation, R, obtained by correlating n_1 independent variates with a dependent variate, having no real correlation with them, is

$$df = \frac{\frac{n_1+n_2-2}{2}!}{\frac{n_2-2}{2}!\,\frac{n_1-2}{2}!}\,(R^2)^{\frac{n_1-2}{2}}\,(1-R^2)^{\frac{n_2-2}{2}}\,d(R^2)$$

where $n_1 + n_2 + 1$ stands for the number of the sample. If n_1 is even, the probability that R should exceed any specified value is the partial sum of a binomial expansion,

$$P = (1-R^2)^{\frac{1}{2}n_2}\left\{1 + \frac{n_2}{2}R^2 + \frac{n_2(n_2+2)}{2.4}R^4 \ldots\ldots \right.$$
$$\left. \ldots \frac{n_2(n_2+2)\ldots(n_2+n_1-4)}{2.4\ldots(n_1-2)}R^{n_1-2}\right\},$$

whereas when n_1 is odd

$$P = \frac{2}{\sqrt{\pi}}\,\frac{\frac{n_2-1}{2}!}{\sqrt{\frac{n_2-2}{2}}}\int_{\sqrt{\frac{n_2R^2}{1-R^2}}}^{\infty}\left(1+\frac{t^2}{n_2}\right)^{-\frac{1}{2}(n_2+1)}dt$$

$$+\frac{2}{\sqrt{\pi}}\cdot\frac{\frac{n_2-1}{2}!}{\frac{n_2-2}{2}!}\,(1-R^2)^{\frac{1}{2}n_2}\left\{R + \frac{n_2+1}{3}R^3 + \frac{(n_2+1)(n_2+3)}{3.5}R^5 + \ldots\right.$$
$$\left. \ldots \frac{(n_2+1)\ldots(n_2+n_1-4)}{3.5\ldots(n_1-2)}R^{n_1-2}\right\};$$

the analogy of these expressions with those given by PEARSON (6) for the χ^2 distribution is obvious. They become identical when

$$n_2 \to \infty,\ n_1 = n' - 1.$$

In the second form it will be noticed that the probability integral of the normal curve has been replaced by an integral of « Student's »

* For $\sqrt{\pi}\sqrt{\frac{n_2-2}{2}}$, read $\sqrt{\pi n_2}\sqrt{\frac{n_2-2}{2}}$!

distribution, of which the approximate value may be obtained from the tables. The multiple correlation must be judged significant only if the value of P obtained is too small to allow us to admit the hypothesis that the dependent variate is really uncorrelated with the independent variates.

REFERENCES.

1. STUDENT, (1908). *The probable error of a mean.* « Biometrika », vol. VI, pp. 1-25.
2. R. A. FISHER, (1921). *On the « Probable Error » of a coefficient of correlation deduced from a small sample.* « Metron », vol. I, Pt. 4, pp. 1-32.
3. R. A. FISHER, (1925). *Statistical Methods for Research Workers.* Oliver & Boyd. Edinburgh.
4. R. A. FISHER, (1922). *The goodness of fit of regression formulae, and the distribution of regression coefficients* « Journal of the Royal Statistical Society », vol. LXXXV, pp. 597-612.
5. R. A. FISHER, (1924). *The influence of rainfall on the yield of wheat at Rothamsted.* « Phil. Trans. », vol. CCXIII. pp. 89-142.
6. K. PEARSON, (1900). *On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling.* « Phil. Mag. », Series V, L, pp. 157-175.