The Generalization of Student's Ratio*

Harold Hotelling
Columbia University

The accuracy of an estimate of a normally distributed quantity is judged by reference to its variance, or rather, to an estimate of the variance based on the available sample. In 1908 "Student" examined the ratio of the mean to the standard deviation of a sample.[1] The distribution at which he arrived was obtained in a more rigorous manner in 1925 by R.A. Fisher,[2] who at the same time showed how to extend the application of the distribution beyond the problem of the significance of means, which had been its original object, and applied it to examine regression coefficients and other quantities obtained by least squares, testing not only the deviation of a statistic from a hypothetical value but also the difference between two statistics.

Let ξ be any linear function of normally and independently distributed observations of equal variance, and let s be the estimate of the standard error of ξ derived by the method of maximum likelihood. If we let t be the ratio to s of the deviation of ξ from its mathematical expectation, Fisher's result is that the probability that t lies between t_1 and t_2 is

$$\frac{1}{\sqrt{\pi n}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \int_{t_1}^{t_2} \frac{dt}{\left(1+\frac{t^2}{n}\right)^{(n+1)/2}} \tag{1}$$

where n is the number of degrees of freedom involved in the estimate s.

* Presented at the meeting of the American Mathematical Society at Berkeley, April 11, 1931.

[1] Biometrika, vol. 6 (1908), p. 1.

[2] Applications of Student's Distribution, Metron, vol. 5 (1925), p. 90.

S. Kotz et al. (eds.), *Breakthroughs in Statistics*

It is easy to see how this result may be extended to cases in which the variances of the observations are not equal but have known ratios and in which, instead of independence among the observations, we have a known system of intercorrelations. Indeed, we have only to replace the observations by a set of linear functions of them which are independently distributed with equal variance. By way of further extension beyond the cases discussed by Fisher, it may be remarked that the estimate of variance s^2 may be based on a body of data not involved in the calculation of ξ. Thus the accuracy of a physical measurement may be estimated by means of the dispersion among similar measurements on a different quantity.

A generalization of quite a different order is needed to test the simultaneous deviations of several quantities. This problem was raised by Karl Pearson in connection with the determination whether two groups of individuals do or do not belong to the same race, measurements of a number of organs or characters having been obtained for all the individuals. Several "coefficients of racial likeness" have been suggested by Pearson and by V. Romanovsky with a view to such biological uses. Romanovsky has made a careful study[1] of the sampling distributions, assuming in each case that the variates are independently and normally distributed. One of Romanovsky's most important results is the exact sampling distribution of L, a constant multiple of the sum of the squares of the values of t for the different variates. This distribution function is given by a somewhat complex infinite series. For large samples and numerous variates it slowly approximates to the normal form; for 500 individuals, Romanovsky considers that an adequate approach to normality requires that no fewer than 62 characters be measured in each individual. When it is remembered that all these characters must be entirely independent, and that it is usually hard to find as many as three independent characters, the difficulties in application will be apparent. To avoid these troubles, Romanovsky proposes a new coefficient of racial likeness, H, the average of the ratios of variances in the two samples for the several characters. He obtains the exact distribution of H, again as an infinite series, though it approaches normality more rapidly than the distribution of L. But H does not satisfy the need for a comparison between magnitudes of characters, since it concerns only their variabilities.

Joint comparisons of correlated variates, and variates of unknown correlations and standard deviations, are required not only for biologic purposes, but in a great variety of subjects. The eclipse and comparison star plates used in testing the Einstein deflection of light show deviations in right ascension and in declination; an exact calculation of probability combining the two least-square solutions is desirable. The comparison of the prices of a list of

[1] V. Romanovsky, On the criteria that two given samples belong to the same normal population (on the different coefficients of racial likeness), Metron, vol. 7 (1928), no. 3, pp. 3–46; K. Pearson, On the coefficient of racial likeness, Biometrika, vol. 18 (1926), pp. 105–118.

commodities at two times, with a view to discovering whether the changes are more than can reasonably be ascribed to ordinary fluctuation, is a problem dealt with only very crudely by means of index numbers, and is one of many examples of the need for such a coefficient as is now proposed. We shall generalize Student's distribution to take account of such cases.

We consider p variates $x_1, x_2, \ldots, x_p$, each of which is measured for N individuals, and denote by $X_{i\alpha}$ the value of x_i for the αth individual. Taking first the problem of the significance of the deviations from a hypothetical set of mean values $m_1, m_2, \ldots, m_p$, we calculate the means $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p$, of the samples, and put

$$\xi_i = (\bar{x}_i - m_i)\sqrt{N}.$$

Then the mean values of the ξ_i will all be zero, and the variances and covariances will be the same as for the corresponding x_i, since the individuals are supposed chosen independently from an infinite population.[1] In order to estimate them with the help of the deviations

$$x_i = X_{i\alpha} - \bar{x}_i$$

from the respective means, we call $n = N - 1$ the number of degrees of freedom and take as the estimates of the variances and covariances,

$$a_{ji} = a_{ij} = \frac{1}{n}\sum_{\alpha=1}^{N} x_{i\alpha}x_{j\alpha} \tag{2}$$

We next put:

$$a = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{vmatrix}$$

$$A_{ij} = A_{ji} = \frac{\text{cofactor of } a_{ij} \text{ in } a}{a}. \tag{3}$$

The measure of simultaneous deviations which we shall employ is

$$T^2 = \sum_{i=1}^{p}\sum_{j=1}^{p} A_{ij}\xi_i\xi_j. \tag{4}$$

For a single variate it is natural to take $A_{11} = 1/a_{11}$; then T reduces to t, the ordinary "critical ratio" of a deviation in a mean to its estimated standard error, a ratio which has "Student's distribution," (1). For examining the deviations from zero of two variates x and y,

[1] "Mean Value" is used in the sense of mathematical expectation; the variance of a quantity whose mean value is zero is defined as the expectation of its squares; the covariance of two such quantities is the expectation of their product. Thus the correlation of the two in a hypothetical infinite population is the ratio of their covariance to the geometric mean of the variances.

$$T = \frac{N}{L - r^2}\left\{\frac{\bar{x}^2}{s_1^2} - \frac{2r\bar{x}\bar{y}}{s_1 s_2} + \frac{\bar{y}^2}{s_2^2}\right\},$$

where

$$s_1^2 = \frac{\sum (X - \bar{x})^2}{N - 1}, \qquad s_2^2 = \frac{\sum (Y - \bar{y})^2}{N - 1},$$

$$r = \frac{\sum (X - \bar{x})(Y - \bar{y})}{\sqrt{\sum (X - \bar{x})^2 \sum (Y - \bar{y})^2}}.$$

For comparing the means of two samples, one of N_1 and the other of N_2 individuals, we distinguish symbols pertaining to the second sample by primes, and write

$$\xi_i = \frac{\bar{x}_i - \bar{x}_i'}{\sqrt{1/N_1 + 1/N_2}} \tag{5}$$

$$n = N_1 + N_2 - 2,$$

$$a = \frac{1}{n}[\sum (X_{i\alpha} - \bar{x}_i)(X_{j\alpha} - \bar{x}_j) + \sum (X_{i\alpha}' - \bar{x}_i')(X_{j\alpha}' - \bar{x}_j')]$$

$$= \frac{1}{n}[\sum X_{i\alpha}X_{j\alpha} - N_1\bar{x}_i\bar{x}_j + \sum X_{i\alpha}X_{j\alpha} - N_2\bar{x}_i'\bar{x}_j'] \tag{6}$$

and take as our "coefficients of racial likeness" the value (4) of T^2, in which the ξ_i are calculated from (5) and the A_{ij} from (6) and (3).

Other situations to which the measure T^2 of simultaneous deviations can be applied include comparisons of regression coefficients and slopes of lines of secular trend, comparisons which for single variates have been explained by R.A. Fisher.[1] In each case we deal for each variate with a linear function ξ_i of the observed values, such that the sum of the squares of the coefficients is unity, so that the variance is the same as for a single observation, and such that the expectation of ξ_i is, on the hypothesis to be tested, zero. Deviations $x_{i\alpha}$ of the observations from means, or from trend lines or other such estimates, are used to provide the estimated variances and covariances a_{ij} by (2). The number of degrees of freedom n is the difference between the number N of individuals and the number q of independent linear relations which must be satisfied by the quantities $x_{i1}, x_{i2}, \ldots, x_{iN}$ on account of their method of derivation. For all the variates, these relations and n must be the same.

The general procedure is to set up what may be called normal values $\bar{x}_{i\alpha}$ for the respective $X_{i\alpha}$, putting

$$x_{i\alpha} = X_{i\alpha} - \bar{x}_{i\alpha}. \tag{7}$$

The underlying assumption is that $X_{i\alpha}$ is composed of two parts, of which one,

[1] Metron, loc. cit., and Statistical Methods for Research Workers, Oliver and Boyd, third edition (1928).

$\varepsilon_{i\alpha}$, is normally and independently distributed about zero with variance σ_i^2 which is the same for all the observations on x_i. The other component is determined by the time, place, or other circumstances of the αth observation in some regular manner, the same for all the variates. Denoting this part by $\eta_{i\alpha}$, we have

$$X_{i\alpha} = \eta_{i\alpha} + \varepsilon_{i\alpha}.$$

Specifically, we take $\eta_{i\alpha}$ to be a linear function, with known coefficients $g_{\alpha s}$, of q unknown parameters $\zeta_{i1}, \ldots, \zeta_{iq}$ where $q < N$:

$$\eta_{i\alpha} = \sum_{s=1}^{q} g_{\alpha s}\zeta_{is}. \tag{8}$$

Thus in dealing with a secular trend representable by a polynomial in the time, we may take the g's as powers of the time-variable, the ζ's as the coefficients. For differences of means, the g's are 0's and 1's, and the ζ's the true means.

We estimate the ζ's by minimizing

$$2V_i = \sum_{\alpha=1}^{N} \varepsilon_{i\alpha}^2 = \sum_{\alpha=1}^{N} (X_{i\alpha} - \eta_{i\alpha})^2. \tag{9}$$

Substituting from (8), differentiating with respect to ζ_{is}, and replacing $\eta_{i\alpha}$ by $\bar{x}_{i\alpha}$ for the minimizing value, we obtain:

$$\sum_{\alpha=1}^{N} g_{\alpha s}(X_{i\alpha} - \bar{x}_{i\alpha}) = 0 \qquad (s = 1, 2, \ldots, q) \tag{10}$$

or by (7),

$$\sum_{\alpha=1}^{N} g_{\alpha s}x_{i\alpha} = 0 \qquad (s = 1, 2, \ldots, q). \tag{11}$$

Denoting also the minimizing values of ζ_{is} by z_{is}, we have made from (8),

$$\bar{x}_{i\alpha} = \sum_{s=1}^{q} g_{\alpha s}z_{is}.$$

Subtracting (8),

$$\bar{x}_{i\alpha} - \eta_{i\alpha} = \sum_{s=1}^{q} g_{\alpha s}(z_{is} - \zeta_{is}). \tag{12}$$

From (9),

$$\begin{aligned} 2V &= \sum_{\alpha=1}^{N} [(X_{i\alpha} - \bar{x}_{i\alpha}) + (\bar{x}_{i\alpha} - \eta_{i\alpha})]^2 \\ &= \sum_{\alpha=1}^{N} (X_{i\alpha} - \bar{x}_{i\alpha})^2 + 2\sum_{\alpha=1}^{N} (X_{i\alpha} - \bar{x}_{i\alpha})(\bar{x}_{i\alpha} - \eta_{i\alpha}) \\ &\quad + \sum_{\alpha=1}^{N} (\bar{x}_{i\alpha} - \eta_{i\alpha})^2. \end{aligned} \tag{13}$$

The middle term, by (12), equals

$$2\sum_{\alpha=1}^{N}\sum_{s=1}^{q} g_{\alpha s}(X_{i\alpha} - \bar{x}_{i\alpha})(z_{is} - \zeta_{is}),$$

this, by (10), is zero. Hence, by (7) and (13),

$$U_i = V_i + W_i,$$

where

$$2V_i = \sum_{\alpha=1}^{N} x_{i\alpha}^2$$

$$2W_i = \sum_{\alpha=1}^{N} (\bar{x}_{i\alpha} - \eta_{i\alpha})^2.$$

If the q equations (10) be solved for $\bar{x}_{i1}, \bar{x}_{i2}, \ldots, \bar{x}_{iN}$, the values of these quantities will be found to be homogeneous linear functions of the observations $X_{i\alpha}$. By (7), therefore, the quantities

$$\bar{x}_{i1}, \bar{x}_{i2}, \ldots, \bar{x}_{iN}$$

are homogeneous linear functions of the $X_{i\alpha}$. But they are not linearly independent functions, since they are connected by the q relations (11). Hence V is a quadratic form of rank

$$n = N - q.$$

Since V_i, by (9), is of rank N, W is of rank q.

This shows that Np new quantities $x'_{i\alpha}$, given by equations of the form

$$\begin{aligned} x'_{i\alpha} &= \sum_{\beta=1}^{N} c_{\alpha\beta}x_{i\beta} = \sum_{\beta=1}^{N} c_{\alpha\beta}x_{i\beta}, \qquad (\alpha = 1, 2, \ldots, n) \\ x'_{i\alpha} &= \sum_{\beta=1}^{N} c_{\alpha\beta}(\bar{x}_{i\beta} - \eta_{i\beta}) = \sum_{\beta=1}^{N} (c_{\alpha\beta}X_{i\beta} - c_{\alpha\beta}\eta_{i\beta}), \qquad (\alpha = n+1, \ldots, N) \end{aligned} \tag{14}$$

can be found such that

$$\begin{aligned} 2V_i &= \sum_{\alpha=1}^{N} x_{i\alpha}^2 = \sum_{\alpha=1}^{N} x_{i\alpha}'^2, \\ 2W_i &= \sum_{\alpha=n+1}^{N} x_{i\alpha}'^2, \end{aligned} \tag{15}$$

and therefore

$$2U_i = \sum_{\alpha=1}^{N} x_{i\alpha}'^2. \tag{16}$$

Substituting (14) in (15) and equating like coefficients,

$$\sum_{\alpha=1}^{n} c_{\alpha\beta}c_{\alpha\gamma} = \delta_{\beta\gamma} \tag{17}$$

where $\delta_{\beta\gamma}$ is the Kronecker delta, equal to 1 if $\beta = \gamma$, to 0 if $\beta \neq \gamma$.

The coefficients $c_{\alpha\beta}$ depend only on the $g_{\alpha s}$, which have been assumed to be the same for all the p variates. Thus (14) may be written

$$x'_{j\alpha} = \sum_{\gamma=1}^{N} c_{\alpha\gamma} x_{j\gamma}.$$

Multiplying by (14), summing with respect to α from 1 to n, and using (17),

$$\sum_{\alpha=1}^{n} x'_{i\alpha} x'_{j\alpha} = \sum_{\alpha=1}^{n} \sum_{\beta=1}^{N} \sum_{\gamma=1}^{N} c_{\alpha\beta} c_{\alpha\gamma} x_{i\beta} x_{j\gamma}$$

$$= \sum_{\beta=1}^{N} \sum_{\gamma=1}^{N} \delta_{\beta\gamma} x_{i\beta} x_{j\gamma} = \sum_{\beta=1}^{N} x_{i\beta} x_{j\beta}. \tag{18}$$

Just as in (2), we define a_{ij} in this generalized case by

$$a_{ij} = \frac{1}{n} \sum_{\alpha=1}^{N} x_{i\alpha} x_{j\alpha}. \tag{19}$$

Then by (18),

$$a_{ij} = \frac{1}{n} \sum_{\alpha=1}^{N} x'_{i\alpha} x'_{j\alpha}. \tag{20}$$

Of the last equation, (6) is a special case.

The random parts $\varepsilon_{i\alpha}$ of the observations on x_i have by hypothesis the distribution

$$\frac{1}{(\sigma_i \sqrt{2\pi})^N} e^{-U_i/2\sigma_i^2} \, d\varepsilon_{i1} \, d\varepsilon_{i2}, \ldots, d\varepsilon_{iN},$$

where V_i is given by (9). From what has been shown, it is clear that this may be transformed into

$$\frac{1}{(\sigma_i \sqrt{2\pi})^N} e^{-(x'^2_{i1} + x'^2_{i2} + \cdots + x'^2_{iN})/2\sigma_i^2} \, dx'_{i1}, \ldots, dx'_{iN},$$

showing that $x'_{i1}, \ldots, x'_{iN}$ are normally and independently distributed with equal variance σ_i^2.

The statistic ξ_i must be independent of the quantities $x'_{i1}, x'_{i2}, \ldots, x'_{in}$ entering into (20), its mean value must be zero, and its variance must be σ_i^2. These conditions are satisfied in the cases which have been mentioned, and are satisfied in general if ξ_i is a linear homogeneous function of $x'_{i,n+1}, \ldots, x'_{iN}$ with the sum of the squares of the coefficients equal to unity.

The measure of simultaneous discrepancy is

$$T^2 = \sum_{i=1}^{p} \sum_{j=1}^{p} A_{ij} \xi_i \xi_j,$$

A_{ij} being defined by (3) on the basis of (19). It is evident that

$$T^2 = -\frac{\begin{vmatrix} 0 & \xi_1 & \xi_2 & \cdots & \xi_p \\ \xi_1 & a_{11} & a_{12} & \cdots & a_{1p} \\ \xi_2 & a_{21} & a_{22} & \cdots & a_{2p} \\ \cdots\cdots & a & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ \xi_p & a_{p1} & a_{p2} & \cdots & a_{pp} \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{vmatrix}} \tag{21}$$

as appears when the numerator is expanded by the first row, and the resulting determinants by their first columns.

A most important property of T is that it is an absolute invariant under all homogeneous linear transformations of the variates $x_i, \ldots, x_p$. This may be seen most simply by tensor analysis; for ξ_i is covariant of the first order and A_{ij} is contravariant of the second order.

The invariance of T shows that in seeking its sampling distribution we may, without loss of generality, assume that the variates $x_1, \ldots, x_p$ have, in the normal population, zero correlations and equal variances for they may always by a linear transformation be replaced by such variates.

Let us now take

$$\xi_i, x'_{i1}, x'_{i2}, \ldots, x'_{in}$$

as rectangular coordinates of a point P_i in space V_{n+1} of $n+1$ dimensions. Since these quantities are normally and independently distributed with equal variance about zero, the probability density for P_i has spherical symmetry about the origin. Indefinite repetition of the sampling would result in a globular cluster of representative points for each variate. Actually the sample in hand fixes the points $P_1, P_2, \ldots, P_p$, which may be regarded as taken independently.

We shall now show that T is a function of the angle θ between the ξ-axis and the flat space V_p containing the points $P_1, P_2, \ldots, P_p$ and the origin 0. We shall denote by A the point on the ξ-axis of coordinates $1, 0, 0, \ldots, 0$, and by V_n the flat space containing the remaining axes. Since in V_{n+1} one equation specifies V_n and $n+1-p$ equations V_p, the intersection of V_n and V_p is specified by all these $n+2-p$ equations, and is therefore of $p-1$ dimensions. Call it V_{p-1}.

If $P_1, P_2, \ldots, P_p$ be moved about in V_p, θ will not change, and neither will T, since T is invariant under linear transformations, equivalent to such motions of the P_i. Hence T always has the value which it takes if all the lines OP_1, $OP_2, \ldots, OP_p$ are perpendicular, with the last $p-1$ of these lines lying in V_{p-1}. In this case the angle AOP_1 equals θ. Applying to the coordinates of A and of P_1 the formula for the cosine of an angle at the origin of lines to $(x_1, x_2, \ldots)$ and $(y_1, y_2, \ldots)$, namely,

$$\cos\theta = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}. \tag{22}$$

We obtain

$$\cos\theta = \frac{\xi}{\sqrt{\xi_1^2 + x_{11}'^2 + \cdots + x_{1n}'^2}}.$$

Since $x_{11}'^2 + \cdots + x_{1n}'^2 = na_{11}$, it follows that

$$n\cot^2\theta = \xi_1^2/a_{11}. \tag{23}$$

The fact that $P_2, P_3, \ldots, P_p$ lie in V_{p-1}, and therefore in V_n, shows that in this case

$$\xi_2 = \xi_3 = \cdots = \xi_p = 0.$$

Because $OP_1, OP_2, \ldots, OP_p$ are mutually perpendicular, (20) and (22) show that $a_{ij} = 0$ whenever $i \neq j$. Hence, by (21) and (23),

$$T = \xi_1/a_{11} = \sqrt{n}\cot\theta. \tag{24}$$

By this result the problem of the sampling distribution of T is reduced to that of the angle θ between a line OA in V_{n+1} and the flat space V_p containing p other lines drawn independently through the origin. The distribution will be unaffected if we suppose V_p fixed and OA drawn at random, with spherical symmetry for the points A.[1] Let us then, abandoning the coordinates hitherto used, take new axes of rectangular coordinates $y_1, y_2, \ldots, y_{n+1}$, of which the first p lie in V_p. A unit hypersphere about 0 is defined in terms of the generalized latitude-longitude parameters $\phi_1, \ldots, \phi_n$ if we put

$$\begin{aligned}
y_1 &= \sin\phi_1 \sin\phi_2 \sin\phi_3 \ldots \sin\phi_{p-1} \cos\phi_p \\
y_2 &= \cos\phi_1 \sin\phi_2 \sin\phi_3 \ldots \sin\phi_{p-1} \cos\phi_p \\
y_3 &= \cos\phi_2 \sin\phi_3 \ldots \sin\phi_{p-1} \cos\phi_p \\
y_4 &= \cos\phi_3 \ldots \sin\phi_{p-1} \cos\phi_p \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots
\end{aligned}$$

[1] This geometrical interpretation of T shows its affinity with the multiple correlation coefficient, whose interpretation as the cosine of an angle of a random line with a V_p enabled R.A. Fisher to obtain its exact distribution (Phil. Trans., vol. 213B, 1924, p. 91; and Proc. Roy. Soc., vol. 121A, 1928, p. 654). The omitted steps in Fisher's argument may be supplied with the help of generalized polar coordinates as in the text. Other examples of the use of these coordinates in statistics have been given by the author in The Distribution of Correlation Ratios Calculated from Random Data, Proc. Nat. Acad. Sci., vol. 11 (1925), p. 657, and in The Physical State of Protoplasm, Koninklijke Akademie van Wetenschappen te Amsterdam, verhandlingen, vol. 25 (1928), no. 5, pp. 28–31.

$$
\begin{aligned}
y_p &= \cos\phi_{p-1}\cos\phi_p \\
y_{p+1} &= \sin\phi_p\cos\phi_{p+1} \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
y_n &= \sin\phi_p\sin\phi_{p+1}\ldots\cos\phi_n \\
y_{n+1} &= \sin\phi_p\sin\phi_{p+1}\ldots\sin\phi_n,
\end{aligned}
$$

for the sum of the squares is unity. Since

$$y_{p+1}^2 + \cdots + y_{n+1}^2 = \sin^2\phi_p$$

we have

$$\phi_p = \theta.$$

The element of probability is proportional to the element of generalized area, which is given by

$$\sqrt{D}\, d\phi_1\, d\phi_2, \ldots, d\phi_n,$$

where D is an n-rowed determinant in which the element in the ith row and jth column is

$$\sum_{k=1}^{n+1} \frac{\partial y_k}{\partial \phi_i}\frac{\partial y_k}{\partial \phi_j}.$$

For $i \neq j$, this is zero. Of the diagonal elements, the first $p - 1$ contain the factor $\cos^2\phi_p$; the pth is unity; and the remaining $n - p$ elements contain the factor $\sin^2\phi_p$. Since ϕ is not otherwise involved, the element of area is the product of

$$\cos^{p-1}\phi_p \sin^{n-p}\phi_p\, d\phi_p$$

by factors independent of ϕ_p. The distribution function of θ is obtained by replacing ϕ_p by θ and integrating with respect to the other parameters. Since θ lies between 0 and $\pi/2$, we divide by the integral between these limits and obtain for the frequency element,

$$\frac{2\Gamma\left(\dfrac{n+1}{2}\right)}{\Gamma\left(\dfrac{p}{2}\right)\Gamma\left(\dfrac{n-p+1}{2}\right)} \cos^{p-1}\theta \sin^{n-p}\theta\, d\theta.$$

Substituting from (24) we have as the distribution of T:

$$\frac{2\Gamma\left(\dfrac{n+1}{2}\right)}{\Gamma\left(\dfrac{p}{2}\right)\Gamma\left(\dfrac{n-p+1}{2}\right)n^{p/2}} \frac{T^{p-1}\, dT}{\left(1 + \dfrac{T^2}{n}\right)^{(n+1)/2}}. \tag{25}$$

For $p = 1$ this reduces to the form of Student's distribution given by Fisher and tabulated in the issue of Metron cited; however, as T may be negative as well as positive in this case, Fisher omits the factor 2.

For $p = 2$ the distribution becomes

$$\frac{n-1}{n}\frac{T\,dT}{\left(1+\frac{T^2}{n}\right)^{(n+1)/2}}.$$

From this it is easy to calculate as the probability that a given value of T will be exceeded by chance,

$$P = \frac{1}{\left(1+\frac{T^2}{n}\right)^{(n-1)/2}}, \tag{26}$$

a very convenient expression.

The probability integral for higher values of p may be calculated in various ways, the most direct being successive integration by parts, giving a series of terms analogous to (26) to which, if p is odd, is added an integral which may be evaluated with the help of the tables of Student's distribution. If p is large, this process is laborious; but other methods are available.

The probability integral is reduced to the incomplete beta function if we put

$$x = (1 + T^2/n)^{-1},$$

for then the integral of (25) from T to infinity becomes

$$P = I_x\left(\frac{n-p+1}{2}, \frac{p}{2}\right),$$

the notation being

$$B_x(p, q) = \int_0^x x^{p-1}(1-x)^{q-1}\,dx,$$

$$B(p, q) = \int_0^1 x^{p-1}(1-x)^{q-1}\,dx,$$

$$I_x(p, q) = \frac{B_x(p, q)}{B(p, q)}.$$

Many methods of calculation have been discussed by H.E. Soper[1] and by V. Romanovsky.[2] An extensive table of the incomplete beta function being

[1] Tracts for Computers, no. 7 (1921).

[2] On certain expansions in series of polynomials of incomplete B-functions (in English), Recueil Math. de la Soc. de Moscou, vol. 33 (1926), pp. 207–229.

prepared under the supervision of Professor Karl Pearson has not yet been published.

Perhaps the most generally useful method now available is to make the substitution

$$z = \tfrac{1}{2}\log_e(n - p + 1)T^2 - \tfrac{1}{2}\log_e np,$$

$$n_1 = p,$$

$$n_2 = n - p + 1,$$

reducing (25) to a form considered by Fisher. Table VI in his book, "Statistical Methods for Research Workers," gives the values of z which will be exceeded by chance in 5 per cent and in 1 per cent of cases. If the value of z obtained from the data is greater than that in Fisher's table, the indication is that the deviations measured are real.

If the variances and covariances are known a priori, they are to be used instead of the a_{ij}; the resulting expression T has the well known distribution of χ, with p degrees of freedom. For very large samples the estimates of the covariances from the sample are sufficiently accurate to permit the use of the χ distribution for T. This is well shown by (25), in which, as n increases, the factor involving T approaches

$$T^{p-1}e^{-T^2/2}\,dT,$$

which is proportional to the frequency element for χ when χ is put for T.

As Pearson pointed out, the labor of calculating χ, which we replace by T, is prohibitive when forty or fifty characters are measured on each individual. With two, three, or four characters, however, the labor is very moderate, and the results far more accurate than any attainable with the Pearson coefficient. The great advantage of using T is the simplicity of its distribution, with its complete independence of any correlations among the variates which may exist in the population.

To means of a single variate it is customary to attach a "probable error," with the assumption that the difference between the true and calculated values is almost certainly less than a certain multiple of the probable error. A more precise way to follow out this assumption would be to adopt some definite level of probability, say $P = .05$, of a greater discrepancy, and to determine from a table of Student's distribution the corresponding value of t, which will depend on n; adding and subtracting the product of this value of t by the estimated standard error would give upper and lower limits between which the true values may with the given degree of confidence be said to lie. With T an exactly analogous procedure may be followed, resulting in the determination of an ellipse or ellipsoid centered at the point $\xi_1, \xi_2, \ldots, \xi_p$. Confidence corresponding to the adopted probability P may then be placed in the proposition that the set of true values is represented by a point within this boundary.