THE TRANSFORMATION OF STATISTICS TO SIMPLIFY THEIR DISTRIBUTION*

By Harold Hotelling and Lester R. Frankel

1. **Introduction.** The custom of regarding a result as significant if it exceeds two or three times its standard error has now given way among informed statisticians to a consideration of the exact probabilities associated with the distribution of the statistic in question. For example, in such problems as that of examining the significance of the difference between the means of two samples, particularly small samples, it is no longer adequate to regard the difference of means, divided by the sample estimate of its standard error, as normally distributed. The significance of this ratio, "Student's ratio," is judged instead by the value of

$$P = 2 \int_t^\infty \phi_n(z)\, dz \tag{1}$$

where n is the number of degrees of freedom entering into the estimate of variance, and

$$\phi_n(z) = \frac{1}{\sqrt{\pi n}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \frac{1}{\left(1 + \frac{z^2}{n}\right)^{\frac{1}{2}(n+1)}}. \tag{2}$$

If the probability law underlying the observations themselves is normal, and they are independent, P is the exact probability of the value of t obtained being equalled or exceeded on the hypothesis that there is no real difference between the means.

Methods of approximating P have been studied by R. A. Fisher[1] and by W. A. Hendricks,[2] and tables have been presented by Student[3] and Fisher.[4] Nevertheless, the practical statistician will very frequently wish to make judgments of significance without stopping to consult a table, or laboriously to compute P, and will tend to revert to the former inaccurate but convenient practice of treating t as normally distributed with unit variance. The essential

* Presented at the joint meeting at Indianapolis of the American Mathematical Society and the Institute of Mathematical Statistics, December 30th, 1937.

[1] *Expansion of Student's Integral in Powers of n^{-1}.* Metron, vol. 5 (1925).

[2] Annals of Mathematical Statistics, vol. 7 (1936), pp. 210-221.

[3] *New Tables for Testing the Significance of Observations.* Metron, vol. 5 (1925).

[4] *Statistical Methods for Research Workers,* Oliver and Boyd, 1925–1936. Tables IV and VI.

reason for this is that the normal distribution to which that of t approximates for large values of n has only one parameter in the expression for the probability. Hence it is easy to remember a few important values, such as those corresponding to $P = .01$ and $.05$; and when values of P representing other levels of significance are in question, the single-entry tables of the normal probability integral are more easily available and easier to use than the double-entry table of Student's integral. Indeed, t is a more useful statistic than Student's original ratio of mean to sample standard deviation, to which it is in the simplest case proportional, partly because of the close approximation of t for large samples to a normally distributed variate of unit variance.

For more complicated statistics the practical need for something simpler than the exact distribution is even more urgent, on account of the larger number of parameters involved in the distributions. For example, the large class of problems giving rise to probabilities expressible as incomplete beta functions require for exactitude the use of Pearson's extensive triple-entry table,[5] and even this is inadequate for some ranges of the parameters. The shorter tables of R. A. Fisher[6] and of Snedecor[7] are helpful, but are also necessarily of triple entry.

It is a common practice, for example, among economists and psychologists, to select either by graphic methods or by preliminary calculation that one, out of many tests that might be applied to available data, for which P is the least. Such selection evidently introduces a bias, which is the more subtle because the tests giving high and therefore insignificant probabilities are likely to be forgotten. Often the only way to guard against such fallacies is to insist on a value of P lower than is easily determined from tables. Thus, if k independent tests of significance have been made, and only the smallest value P is reported, its significance should be judged not by this value P itself, but by the probability

$$P' = 1-(1 - P)^k = kP - \cdots$$

of the least value being so small. If we equate P' to some such standard value as .01, then P must, for this standard level of confidence, take only a fraction, approximately $1/k$, of this value. Such a small probability will often fall outside the range of existing tables.

Instead of relying on tables or direct computation from the exact distribution of a statistic, it will sometimes be desirable to use a modification of the statistic, selected so as to have the normal or some other standard distribution. We shall consider a type of transformation of a statistic such that the distribution becomes the limiting form of the original distribution as the sample size increases. Thus our transformation will reduce to the application to the statistic of a correction which will be small when the sample is large. We shall show how to make simple approximate corrections of this character for two cases.

[5] *Tables of the Incomplete Beta Function*, Biometrika Office, 1934.

[6] Loc. cit. Tables IV and VI.

[7] *Calculation and Interpretation of Analysis of Variance and Covariance.* Ames, Iowa. Collegiate Press. 1934.

The first of these is the Student ratio t, the lower limit of the integral in (1). Putting

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \tag{3}$$

and

$$P = 2 \int_x^\infty \phi(z)\, dz \tag{4}$$

which in view of (1) and the fact that the integral of each distribution from $-\infty$ to ∞ is unity is equivalent to

$$\int_0^x \phi(z)\, dz = \int_0^t \phi_n(z)\, dz \tag{5}$$

we shall show that x has an asymptotic expansion:

$$\begin{aligned} x \sim t \Big\{ 1 - \frac{t^2 + 1}{4n} + \frac{13t^4 + 8t^2 + 3}{96n^2} - \frac{35t^6 + 19t^4 + t^2 - 15}{384n^3} \\ + \frac{6271t^8 + 3224t^6 - 102t^4 - 1680t^2 - 945}{92{,}160n^4} - \cdots \Big\} \end{aligned} \tag{6}$$

It will frequently be a sufficient approximation to treat

$$t\left(1 - \frac{t^2 + 1}{4n}\right)$$

as normally distributed. These appear to be approximations of practical value when $n \geq t^2$.

The second statistic whose transformation to a function having its limiting distribution we shall consider is the generalized Student ratio T, appropriate to all the uses to which t can be put, but with a multiplicity of variates instead of one to serve as the basis of the test of significance.[8] This is defined with reference to variates $x_1, \cdots x_p$, together with a linear function of sample values (proportional for example to the difference between the means in two samples), such that if ξ_i is the value of this function of the sample values of x_i ($i = 1, \cdots, p$) then the variance of ξ_i in the population sampled is the same as that of x_i, and on the hypothesis to be tested, the population mean of each ξ_i is zero. In terms of unbiased quadratic estimates s_{ij} of the covariances σ_{ij} among $x_1, \cdots, x_p$, each based on n degrees of freedom, we may define l_{ij} as the cofactor of s_{ij} divided by the determinant of the statistics s_{ij}. Then T is defined by

$$T^2 = \Sigma\Sigma l_{ij} \xi_i \xi_j \tag{7}$$

[8] Harold Hotelling, *The Generalization of Student's Ratio.* Annals of Mathematical Statistics, vol. 2 (1931), pp. 360-378.

the summations running independently with respect to i and j from l to p. For independent samples from a multivariate normal population, the distribution of T has been shown[9] to be

$$\frac{2\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{n-p+1}{2}\right)n^{\frac{1}{2}p}} \frac{T^{p-1}\,dT}{\left(1+\frac{T^2}{n}\right)^{\frac{1}{2}(n+1)}}. \tag{8}$$

As n increases, the distribution of T approaches the χ distribution with p degrees of freedom:

$$\frac{\chi^{p-1}e^{-\frac{1}{2}\chi^2}\,d\chi}{2^{\frac{1}{2}(p-2)}\,\Gamma\left(\frac{p}{2}\right)} \tag{9}$$

By equating the probabilities derived from these two distributions, we shall define χ as a function of T, and obtain asymptotic expansions for the functions χ and χ^2 thus defined.

Since the probability associated with T is expressible in terms of the incomplete beta function, or the analysis of variance distribution integral, it follows that any of the many common statistics, of which simple functions have this distribution, can be expressed simply in terms of T. Tests of significance in a wide variety of cases may therefore be made with the help of the asymptotic expansion corresponding to T^2, together with a table of χ^2.

A further advantage of the transformation of a statistic into a normally distributed variate of unit variance and zero mean is that further statistical tests are possible with such variates. Since a great part of statistical theory is based on the assumption of such normal distributions, an extensive field of applications becomes available in this way. For example, if several independent tests give values of t based on various numbers of degrees of freedom, and it is desired to combine these tests so as to get a single probability, the corresponding values of the normally distributed variate x defined above may be squared and added. The sum will then have the χ^2 distribution, with a number of degrees of freedom equal to the number of values of t used. In a similar manner, the values of χ^2 corresponding to a number of independently determined values of T^2 may be added, and the sum will have the χ^2 distribution with a number of degrees of freedom equal to the sum of the various values of p involved.

The advantages of this type of what may be called "normalization" of a statistic have been brought out by R. A. Fisher for the particular case of the correlation coefficient. His use[10] of $z = \frac{1}{2}\log\frac{1+r}{1-r}$ facilitates such operations as the averaging of values obtained from independent samples, or taking the

[9] Harold Hotelling, loc. cit.

[10] *Statistical Methods for Research Workers*, Sec. 35.

difference between two values, with the testing of significance of the result in each case. This is because z, unlike r, has a nearly normal distribution, with variance nearly independent of the population value. We note in passing that this function is the same as $\tanh^{-1} r$, and may therefore be determined accurately and readily from the Smithsonian Institution Tables of Hyperbolic and Exponential Functions.

2. **Normalization of** t. The "duplication formula" in the theory of the Gamma function[11] shows that

$$\Gamma\left(\frac{n+1}{2}\right) = \frac{\sqrt{\pi}\,\Gamma(n)}{2^{n-1}\Gamma\left(\frac{n}{2}\right)}$$

Substituting this in (2) and taking logarithms we have:

$$\begin{aligned} \log \phi_n(z) &= -\tfrac{1}{2}\log n - (n-1)\log 2 + \log \Gamma(n) \\ (10) \qquad & \qquad - 2\log\Gamma\left(\frac{n}{2}\right) - \frac{n+1}{2}\log\left(1+\frac{z^2}{n}\right) \end{aligned}$$

The last logarithm may be expanded in a series of powers of z^2/n which not only converges uniformly on the interval $0 \leq z \leq t$ when $n > t^2$, but has the property of being a *uniformly asymptotic* representation of the function on this interval. This means that the sum of the first j terms of the series ($j = 0, 1, 2, \cdots$) differs from the function represented, by a quantity whose product by n^{j+1} has, for sufficiently large values of n, an upper bound independent of z, so long as z remains in this interval. Uniformly asymptotic series have a number of important properties, among which is[12] term by term integrability with respect to z. In this sense we have the uniform asymptotic representation:

$$(11) \quad -\frac{n+1}{2}\log\left(1+\frac{z^2}{n}\right) \sim -\frac{z^2}{2} - \frac{2z^2 - z^4}{4n} + \frac{3z^4 - 2z^6}{12n^2} - \frac{4z^6 - 3z^8}{24n^3} + \cdots$$

We shall obviously have another uniform asymptotic representation if we add to this, term by term, asymptotic series with terms independent of z, such as those for the gamma function logarithms in (10). Since[13]

$$(12) \quad \log\Gamma(n) \sim \tfrac{1}{2}\log 2\pi + (n - \tfrac{1}{2})\log n - n + \sum_{r=1}^{\infty} \frac{(-1)^{r-1}B_r}{2r(2r-1)n^{2r-1}},$$

where

$$B_1 = \tfrac{1}{6}, \qquad B_2 = \tfrac{1}{30}, \qquad B_3 = \tfrac{1}{42}, \qquad B_4 = \tfrac{1}{30}, \qquad B_5 = \tfrac{5}{66}, \cdots$$

[11] Whittaker and Watson, *Modern Analysis*, 4th ed., p. 240.

[12] H. Schmidt, *Beiträge zu eine Theorie der allgemeinen asymptotischen Darstellungen.* Math. Annalen, vol. 113 (1937), pp. 629-656. The property mentioned above is proved in Schmidt's Theorem 6.

[13] Whittaker and Watson, loc. cit., pp. 252, 125.

are the Bernoulli numbers, we obtain upon substituting in (10) this and the similar formula for $\log \Gamma\left(\frac{n}{2}\right)$, together with (11), and some simplification,

$$\begin{aligned}\log \phi_n(z) \sim -\tfrac{1}{2}\log 2\pi - \frac{z^2}{2} + \frac{-1 - 2z^2 + z^4}{4n} \\ + \frac{3z^4 - 2z^6}{12n^2} + \frac{1 - 4z^6 + 3z^8}{24n^3} + \frac{5z^8 - 4z^{10}}{40n^4} + \cdots\end{aligned} \tag{13}$$

Upon differentiating (5) we obtain:

$$\phi(x)\frac{dx}{dt} = \phi_n(t) \tag{14}$$

Since ϕ is simply the normal distribution function (3), this may be written:

$$-\frac{1}{2}\log 2\pi - \frac{x^2}{2} + \log\frac{dx}{dt} = \log \phi_n(t) \tag{15}$$

We shall always in this paper use the symbol "lim" to mean the limit as n approaches infinity. The functions of n and z, or of n and t, which we shall denote by R, R', R'', with or without subscripts, are to be such that the absolute value of each has an upper bound independent of n, z and t so long as $n \geq 1$, and z and t are confined to some fixed finite interval.

From (13) we have that $\lim \log \phi_n(z) = \log \phi(z)$, whence, by the continuity of the exponential function,

$$\lim \phi_n(z) = \phi(z)$$

This holds *uniformly* for $0 \leq z \leq t$. Subtracting $\int_0^t \phi(z)\,dz$ from both sides of (5) we therefore find that

$$\int_t^x \phi(z)\,dz = \int_0^t \{\phi_n(z) - \phi(z)\}\,dz \tag{16}$$

can by choosing n large enough be made as small as we please. Since $\phi(z) > 0$, it follows that the function x of t and n is such that

$$\lim x = t. \tag{17}$$

A parallel argument, proving slightly more than (17), is the following. From (13),

$$\log \phi_n(z) = \log \phi(z) + \frac{R'}{n}$$

where R' is a bounded function of the kind described above. Therefore

$$\phi_n(z) = \phi(z)\left(1 + \frac{R''}{n}\right).$$

Substituting this in (16) we have that

$$\int_t^x \phi(z)\,dz = \frac{1}{n}\int_0^t \phi(z)R''\,dz$$

From the mean value theorem of integral calculus it then follows that

$$x = t + \frac{R_1}{n} \tag{18}$$

An asymptotic series may be substituted in a power series, and the result is a valid asymptotic representation of the corresponding function. (Schmidt, loc. cit., Theorem 4.) This justifies taking the exponential of each side of (13) and arranging in a series of powers of n^{-1} to give

$$\phi_n(z) \sim \phi(z)\left\{1 + \frac{\alpha_1(z)}{n} + \frac{\alpha_2(z)}{n^2} + \cdots\right\} \tag{19}$$

This asymptotic development will, like the original one, hold uniformly in every finite interval, and may therefore be integrated term by term. Thus

$$\int_0^t \phi_n(z)\,dz = \int_0^t \phi(z)\left\{1 + \frac{\alpha_1(z)}{n} + \frac{\alpha_2(z)}{n^2} + \cdots + \frac{\alpha_j(z)}{n^j}\right\}dz + \frac{R_{j+1}}{n^{j+1}} \tag{20}$$

where $|R_{j+1}|$ has an upper bound independent of n and t when $n \geq 1$, and t is confined to a finite interval, $0 \leq t \leq T$. Substituting this in (16) we obtain:

$$\int_t^x \phi(z)\,dz = \int_0^t \phi(z)\left\{\frac{\alpha_1(z)}{n} + \cdots + \frac{\alpha_j(z)}{n^j}\right\}dz + \frac{R_{j+1}}{n^{j+1}} \tag{21}$$

In terms of a sequence of functions $f_1, f_2, \cdots$ of t to be defined below, let

$$x_j = t + \frac{f_1}{n} + \frac{f_2}{n^2} + \cdots + \frac{f_j}{n^j}. \tag{22}$$

Now $\int_t^{x_j} \phi(z)\,dz$ can be expanded in a series of powers of n^{-1} which converges for sufficiently large values of n; for the Taylor series

$$\phi(z) = \phi(t) + (z - t)\,\phi'(t) + \cdots \tag{23}$$

can be integrated to give a series of powers of $x_j - t$, which by (22) is a polynomial in n^{-1}. As a matter of fact we have from (22) that $x_j - t$ can be made arbitrarily small by taking n large enough; consequently the series (23) and that obtained by integration in this way will converge uniformly and absolutely. We thus have:

$$\begin{aligned}\int_t^{x_j} \phi(z)\,dz &= \frac{1}{n}f_1\phi + \frac{1}{n^2}\left(f_2\phi + \frac{1}{2}f_1^2\phi'\right)\\ &\quad + \frac{1}{n^3}\left(f_3\phi + f_1f_2\phi' + \frac{1}{6}f_1^3\phi''\right) + \cdots\end{aligned} \tag{24}$$

Now let us define f_1, $f_2 \cdots$, by equating the coefficient of each power of n in (24) to that of the same power of n in the right member of (21). This process gives a sequence of equations

$$\begin{aligned} f_1\phi &= \int_0^t \phi(z)\alpha_1(z)\,dz \\ f_2\phi + \tfrac{1}{2}f_1^2\phi' &= \int_0^t \phi(z)\alpha_2(z)\,dz \\ f_3\phi + f_1f_2\phi' + \tfrac{1}{6}f_1^3\phi'' &= \int_0^t \phi(z)\alpha_3(z)\,dz \\ f_4\phi + (f_1f_3 + \tfrac{1}{2}f_2^2)\phi' + \tfrac{1}{2}f_1^2f_2\phi'' + \tfrac{1}{24}f_1^4\phi''' &= \int_0^t \phi(z)\alpha_4(z)\,dz \end{aligned} \tag{25}$$

Since $\phi \neq 0$ the first of these equations defines f_1 for every value of t; when f_1 has been determined, the second equation defines f_2; then the third defines f_3, and so forth. It is to be observed that the functions f_1, f_2, $\cdots$ thus determined are not changed when the value of j appearing in (22) is increased; we have a unique sequence.

If for the right-hand member of (15) we substitute that of (13), replacing z by t, and on the left of (15) put

$$x = t + \frac{f_1}{n} + \frac{f_2}{n^2} + \cdots$$

$$\frac{dx}{dt} = 1 + \frac{f_1}{n} + \frac{f_2}{n^2} + \cdots$$

and then expand in a formal manner in powers of n^{-1}, we shall upon equating coefficients of like powers of n obtain a sequence of differential equations

$$\begin{aligned} f_1' - tf_1 &= \tfrac{1}{4}(-1 - 2t^2 + t^4) \\ f_2' - tf_2 &= \tfrac{1}{2}f_1'^2 + \tfrac{1}{2}f_1^2 + \tfrac{1}{4}t^4 - \tfrac{1}{6}t^6 \\ &\cdots\cdots\cdots\cdots\cdots\cdots \end{aligned} \tag{26}$$

These, with the initial condition $f_1 = f_2 = \cdots = 0$ for $t = 0$ determine the same sequence of functions as before. The equations (26) are in fact obtainable simply by differentiating (25) and cancelling out the factor $\phi(t)$. That this must be true follow from the equivalence of the various formal processes of manipulating series of powers of n^{-1}, whether convergent or divergent, to give equivalent results. The differential equations are easily solved; the solutions, at least for f_1, f_2, f_3, and f_4, are all polynomials. Why they should come out as polynomials is not immediately obvious; but their calculation is made easier if each f_j is replaced in the differential equations by a polynomial of degree $2j + 1$ with undetermined coefficients, involving only odd powers of t. The f's of lower order are replaced by values previously determined, and the coefficients are found by equating like powers of t. This process supplies at each stage more equations than unknown coefficients; their consistency verifies the assumption that f_j is a polynomial of the kind specified, at least for $j \leq 4$. These polynomials are the coefficients of the powers of n^{-1} in (6).

The series on the right of (24) not only converges but is an asymptotic series uniformly valid when t varies in any finite interval. Hence upon subtracting (24) from (21) and taking account of (25) we find that

$$\int_{x_j}^{x} \phi(z)\, dz = \frac{R'_{j+1}}{n^{j+1}}$$

where $| R'_{j+1} |$ is uniformly bounded. Upon applying the mean value theorem to the integral on the left we find that x differs from x_j, and thus from the first j terms (22) of the series (6), by a quantity whose product by n^{j+1} remains bounded when n approaches infinity. This proves the validity of the asymptotic expansion.

3. **Accuracy of the Approximation.** To follow through the above processes in such a way as to obtain useful limits for the error involved in using the first few terms of the series (6) in place of x would be excessively difficult. However, the magnitude of the error in taking the first two or three terms as an approximation to x may be judged from the tables below to be adequately small for practical purposes, provided $n \geq t^2$. The essential singularity of the normal distribution at infinity, in contrast with the algebraic nature of the Student distribution, means a poorer approximation of one to the other as t increases while n remains fixed, though a better approximation as n increases. This is illustrated in the following tables, where it will be observed that the approximations are better for large than for small values of n, and of

$$P = 2\int_{x}^{\infty} \phi(z)\, dz = 2\int_{t}^{\infty} \phi_n(z)\, dz$$

It will be seen that for $n = 10$ and $P < .001$, the utility of the asymptotic series, or at least of its first five terms, is vitiated by the rapid oscillation of consecutive terms, due to the high values of t^2 in relation to n.

	$P = .10$		$P = .05$		$P = .01$	
	$n = 10$	$n = 30$	$n = 10$	$n = 30$	$n = 10$	$n = 30$
t	1.812	1.697	2.228	2.042	3.169	2.750
x_1	1.618	1.642	1.896	1.954	2.294	2.554
x_2	1.650	1.645	1.980	1.960	2.754	2.579
x_3	1.643	1.645	1.953	1.960	2.446	2.575
x	1.645		1.960		2.576	

	$P = .001$		$P = .0001$		
	$n = 10$	$n = 30$	$n = 10$	$n = 30$	$n = 100$
t	4.587	3.646	6.22	4.482	4.052
x_1	2.059	3.212	.05	3.69	3.88
x_2	4.981	3.313	12.86	3.98	3.89
x_3	0.896	3.283	−20.44	3.85	3.89
x_4	7.163	3.293	75.66	3.91	3.89
x	3.291		3.891		

4. **Transformation of the Generalized Student Ratio.** The arguments and methods of calculation set forth in Section 2 may be applied with little or no change to the transformation of various other statistics in such a way that the limiting distribution for large samples is reached at once for the transformed statistic. In particular, to deal with the generalized Student ratio T, we may equate (8) to (9), represent χ as an asymptotic expansion with undetermined coefficients which are functions of T, and then by substituting and equating like powers of n^{-1} obtain as before a sequence of differential equations for determining the coefficients. This process gives

$$\chi \sim T - \frac{pT + T^3}{4n} + \frac{(8 - 5p^2)T + (4 + 4p)T^3 + 13T^5}{96n^2} \cdots \tag{27}$$

This reduces to the expansion of x in terms of t previously found if we put $p = 1$.

It is somewhat more convenient in practice to use χ^2 and T^2, to avoid extracting the square root of the latter expression, and to utilize the existing tables of χ^2. Ordinarily therefore we should not use (27), but the series

$$\chi^2 \sim T^2\left\{1 - \frac{p + T^2}{2n} + \frac{(4 - p^2) + (2 + 5p)T^2 + 8T^4}{24n^2} \cdots\right\},$$

which may be obtained in the same way, or by squaring (27) in a formal manner. That these are genuine asymptotic approximations follows by essentially the same argument as before.

Columbia University and Washington, D. C.