AN APPROXIMATION TO "STUDENT'S" DISTRIBUTION*

By Walter A. Hendricks

I. Introduction

The function commonly known as "Student's" distribution occupies a prominent position among the classic contributions to the field of statistics, not only for its intrinsic value but also for the stimulus which it gave to statistical research at the time of its discovery.

The function, which may be written in the form,

$$dF_z = \frac{1}{B[\frac{1}{2}(n-1), \frac{1}{2}]} (1 + z^2)^{-\frac{1}{2}n} dz, \tag{1}$$

gives the distribution of the ratio, z, of the estimated arithmetic mean, $\bar{x}$, to the estimated standard deviation, s, for samples of n observations drawn from the normal universe specified by the arithmetic mean, zero, and the standard deviation, σ. This function, together with a table of values of its integral was given by "Student."[9,10]

In view of the fact that similar distributions were subsequently found by Fisher[2] to arise in a larger variety of practical problems than was originally supposed, a table of values of a new integral was later given by "Student"[11] in which the distribution of a variable, t, defined by the relation,

$$t = (n-1)^{\frac{1}{2}} z, \tag{2}$$

rather than the distribution of z itself, was considered. Another table giving the distribution of t, in a form intended to be more convenient for use by research workers wishing to apply statistical methods to experimental data, was later given by Fisher.[3]

The integration of functions of the type defined by equation (1) involves considerable labor, a fact which has been somewhat embarrassing to practical statisticians interested in the distributions of z and t for values of n larger than those included in the above-mentioned tables. The recent appearance of Tables of the Incomplete Beta-Function, prepared under the direction of Pearson,[7] has considerably alleviated the difficulty, but the requirements of certain practical problems are not easily satisfied even with the aid of these tables. Consequently, simple approximations to the distributions of z and t,

[*] A thesis submitted to the Faculty of the Columbian College of The George Washington University in part satisfaction of the requirements for the degree of Master of Arts.

which will be sufficiently accurate for most practical purposes, should be of some interest.

According to "Student,"[9] the distribution of z tends to approach a normal curve with a standard deviation of $(n - 3)^{-\frac{1}{2}}$ for values of n greater than 10. However, Deming. and Birge[1] have recently suggested that the distribution tends to approach a normal curve with a standard deviation of $(n - 1\frac{1}{2})^{-\frac{1}{2}}$.

This thesis presents a simple approximation to the distribution of z, which can be readily extended to the distribution of t and which will give more accurate results than either of the above approximations.

II. Approximation to the Distribution of Z

The approximation presented here is based upon the assumption that, for large values of n, the distribution of s tends to approach a normal curve with the arithmetic mean, $\bar{s}$, and the standard deviation, $\dfrac{\sigma}{2^{\frac{1}{2}}n^{\frac{1}{2}}}$, that is,

$$dF_s = \frac{n^{\frac{1}{2}}}{\pi^{\frac{1}{2}}\sigma} e^{-\frac{n}{\sigma^2}(s-\bar{s})^2} ds. \tag{3}$$

Since the distribution of the estimated arithmetic mean, $\bar{x}$, is known to be normal, with the standard deviation, $\dfrac{\sigma}{n^{\frac{1}{2}}}$, we have for the joint distribution of s and $\bar{x}$:

$$dF_{s,\,\bar{x}} = \frac{n}{2^{\frac{1}{2}}\pi\sigma^2} e^{-\frac{n}{\sigma^2}[\frac{1}{2}\bar{x}^2+(s-\bar{s})^2]} ds\, d\bar{x}. \tag{4}$$

$\bar{s}$ may be expressed in terms of n and σ by the well-known relation,

$$\bar{s} = c_n\sigma, \tag{5}$$

in which the factor, c_n, is defined by the formula,

$$c_n = \frac{2^{\frac{1}{2}}}{n^{\frac{1}{2}}} \frac{\Gamma(\frac{1}{2}n)}{\Gamma[\frac{1}{2}(n-1)]}. \tag{6}$$

If we write, $c_n\sigma$, in place of $\bar{s}$, in equation (4) and make the transformation,

$$\bar{x} = sz, \tag{7}$$

we have for the joint distribution of s and z:

$$dF_{s,z} = \frac{n}{2^{\frac{1}{2}}\pi\sigma^2} e^{-\frac{n}{\sigma^2}[\frac{1}{2}s^2z^2+(s-c_n\sigma)^2]} s\, ds\, dz. \tag{8}$$

To find the distribution of z, all that is necessary is to write:

$$dF_z = k\left[\int_{-\infty}^{+\infty} e^{-(as-b)^2} s\, ds\right] dz, \tag{9}$$

in which:

$$(10)\qquad \begin{aligned} k &= \frac{n}{2^{\frac{1}{2}}\pi\sigma^2}\, e^{-nc_n{}^2 \frac{z^2}{z^2+2}}, \\ a &= \frac{n^{\frac{1}{2}}}{2^{\frac{1}{2}}\sigma}\,(z^2 + 2)^{\frac{1}{2}}, \\ b &= 2^{\frac{1}{2}} n^{\frac{1}{2}} c_n\,(z^2 + 2)^{-\frac{1}{2}}. \end{aligned}$$

The integral in brackets in equation (9) can be evaluated without any difficulty. We have:

$$(11)\qquad \int_{-\infty}^{+\infty} e^{-(as-b)^2}\, s\, ds = \frac{b\pi^{\frac{1}{2}}}{a^2}.$$

Substituting this value in equation (9) and replacing k, a, and b by the quantities which they represent, we obtain the following expression for the distribution of z:

$$(12)\qquad dF_z = \frac{2n^{\frac{1}{2}}c_n}{\pi^{\frac{1}{2}}}\, e^{-nc_n{}^2 \frac{z^2}{z^2+2}}\,(z^2 + 2)^{-\frac{3}{2}}\, dz.$$

If we now define a new variable, u, by the relation,

$$(13)\qquad u^2 = 2nc_n{}^2\,\frac{z^2}{z^2 + 2},$$

and make the appropriate substitutions in equation (12), we have, for the distribution function of u:

$$(14)\qquad dF_u = \frac{1}{2^{\frac{1}{2}}\pi^{\frac{1}{2}}}\, e^{-\frac{1}{2}u^2}\, du.$$

Equation (14) is obviously a normal curve with unit standard deviation. We have thus deduced the interesting fact that, for values of n sufficiently large so that the distribution of s may be represented by a normal curve, the quantity, $2^{\frac{1}{2}}n^{\frac{1}{2}}c_n\,\dfrac{z}{(z^2 + 2)^{\frac{1}{2}}}$, is distributed as a normal deviate with unit standard deviation.

The accuracy of this approximation as compared with that of the approximation suggested by "Student"[9] and that of the more recent approximation suggested by Deming and Birge[1] may now be considered. As previously stated, the "Student" approximation is based on the assumption that the quantity, $(n - 3)^{\frac{1}{2}}z$, is distributed as a normal deviate with unit standard deviation for values of n greater than 10, while that suggested by Deming and Birge is based on the assumption that the quantity, $(n - 1\frac{1}{2})^{\frac{1}{2}}z$, is so distributed.

Table 1* gives values of the integral, I_z, defined by:

$$(15)\qquad I_z = \frac{1}{B[\frac{1}{2}(n - 1), \frac{1}{2}]} \int_{-\infty}^{z} (1 + z^2)^{-\frac{1}{2}n}\, dz,$$

* All tables and charts to which reference is made are to be found in the Appendix.

for the case, $n = 10$, together with the corresponding approximate values obtained by making use of the three approximations suggested by "Student," Deming and Birge, and the present author, respectively. The exact values and those obtained by the "Student" approximation were derived from values calculated by "Student"[9] and given by Pearson.[5] All other data in the table were calculated by the present author.

An inspection of Table 1 shows that the values of I_z based on the approximation presented in this thesis agree very well with the corresponding exact values. The agreement is better than that found in the case of either of the other two approximations. The Deming and Birge approximation gives better results than the "Student" approximation for values of z in the neighborhood of zero, but for other values of z the opposite is true.

III. Approximation to the Distribution of t

Since tables giving the distribution of the variable, t, have largely superseded those giving the distribution of z in practical statistical work, the feasibility of applying the above three approximations to the distribution of t is worthy of consideration.

The variable, t, has already been defined in terms of n and z by equation (2). If, in equation (12), we make the transformation,

$$z = (n - 1)^{\frac{1}{2}}t, \tag{16}$$

we have, for the distribution function of t:

$$dF_t = \frac{2n^{\frac{1}{2}}(n-1)c_n}{\pi^{\frac{1}{2}}} e^{-nc_n^2 \frac{t^2}{t^2+2(n-1)}} [t^2 + 2(n-1)]^{-\frac{3}{2}}\, dt. \tag{17}$$

If we now define a variable, v, by the relation,

$$v^2 = 2nc_n^2 \frac{t^2}{t^2 + 2(n-1)}, \tag{18}$$

we have, for the distribution function of v:

$$dF_v = \frac{1}{2^{\frac{1}{2}}\pi^{\frac{1}{2}}} e^{-\frac{1}{2}v^2}\, dv. \tag{19}$$

Equation (19) shows that, for values of n sufficiently large so that the distribution of s may be represented by a normal curve, the quantity,

$$2^{\frac{1}{2}}n^{\frac{1}{2}}c_n \frac{t}{[t^2 + 2(n-1)]^{\frac{1}{2}}},$$

is distributed as a normal deviate with unit standard deviation. On the other hand, if we assume with "Student" that, for large values of n, the quantity, $(n - 3)^{\frac{1}{2}}z$, is normally distributed about zero with unit standard deviation, we should expect to find that the quantity, $\frac{(n-3)^{\frac{1}{2}}}{(n-1)^{\frac{1}{2}}}t$, is also distributed as a normal deviate with unit standard deviation. If the Deming and Birge approximation to the distribution of z is assumed to be valid, we should expect to find that the quantity, $\frac{(n - 1\frac{1}{2})^{\frac{1}{2}}}{(n - 1)^{\frac{1}{2}}} t$, is distributed as a normal deviate with unit standard deviation.

To test the accuracy of each of these three approximations to the distribution of t, we may make use of the well-known table of values of t given by Fisher.[3] This table is so constructed that a value of t corresponding to a given number of "degrees of freedom" and a given value of "P" may be read from the table, where P is defined by the relation,

$$P = 1 - \frac{2}{(n - 1)^{\frac{1}{2}} B[\frac{1}{2}(n - 1), \frac{1}{2}]} \int_0^t \left(1 + \frac{t^2}{n - 1}\right)^{-\frac{1}{2}n} dt. \tag{20}$$

The entries in the last line of the table, corresponding to an infinite number of "degrees of freedom," are the deviates of a normal curve with unit standard deviation.

To test the accuracy of the "Student" approximation, we may calculate the entries for a line of this table, corresponding to $n - 1$ "degrees of freedom," by multiplying the entries in the last line of the table by $\frac{(n - 1)^{\frac{1}{2}}}{(n - 3)^{\frac{1}{2}}}$. These approximate values of t may then be compared with the exact values given in the table. The accuracy of the Deming and Birge approximation may be tested in the same manner, except that in this case the entries in the last line of the table should be multiplied by $\frac{(n - 1)^{\frac{1}{2}}}{(n - 1\frac{1}{2})^{\frac{1}{2}}}$. To test the accuracy of the approximation given by equation (19), we may calculate the values of t corresponding to $n - 1$ "degrees of freedom" by means of the relation,

$$t^2 = \frac{2(n - 1)v^2}{2nc_n^{\,2} - v^2}, \tag{21}$$

in which the entries in the last line of the table are to be taken as the values of v.

Table 2 gives the exact values of t corresponding to the values of P given in Fisher's table for $n = 10$, together with the approximate values calculated by means of each of the above three approximations. This comparison of the accuracies of the three approximations is equivalent to the comparisons presented in Table 1. The conclusions which may be drawn are in agreement with those which have already been drawn from that table.

In order to test the behavior of each of the approximations for a larger value of n, values of t corresponding to the different values of P were calculated for $n = 30$. The results are presented in Table 3. The rank of each of the three approximations, with regard to accuracy, for $n = 30$ is the same as for $n = 10$. Although all three give more accurate results for the larger value of n, the superiority of the approximation presented in this thesis is quite apparent.

For extremely large values of n, all three approximations evidently tend to become one-hundred percent accurate, for the distribution of t tends to become normal as n is increased indefinitely. In the case of the "Student" and Deming and Birge approximations, the ratios, $\frac{(n-1)^{\frac{1}{2}}}{(n-3)^{\frac{1}{2}}}$ and $\frac{(n-1)^{\frac{1}{2}}}{(n-1\frac{1}{2})^{\frac{1}{2}}}$, obviously approach unity, respectively, as n becomes very large. The approximate value of t given by equation (21) also tends to approach the normal deviate, v, as n is increased for we have:

$$\begin{aligned}\lim_{n\to\infty} t^2 &= \lim_{n\to\infty} \frac{2(n-1)v^2}{2nc_n^2 - v^2} = \lim_{n\to\infty}\left[\frac{2nv^2}{2nc_n^2 - v^2} - \frac{2v^2}{2nc_n^2 - v^2}\right] \\ &= \lim_{n\to\infty}\left[\frac{v^2}{c_n^2 - \frac{v^2}{2n}} - \frac{2v^2}{2nc_n^2 - v^2}\right] = v^2. \end{aligned} \tag{22}$$

IV. Discussion

The greater accuracy of the approximation to the distribution of z presented in this thesis apparently can not be explained by the hypothesis that the distribution of s becomes normal more rapidly than the distribution of z as n is increased. Table 4 presents values of the ordinates of the normal curve with unit standard deviation, together with the corresponding ordinates of the exact distributions of the quantities, $\frac{2^{\frac{1}{2}}n^{\frac{1}{2}}}{\sigma}(s-\bar{s})$, $(n-3)^{\frac{1}{2}}z$, $(n-1\frac{1}{2})^{\frac{1}{2}}z$, and $2^{\frac{1}{2}}n^{\frac{1}{2}}c_n\frac{z}{(z^2+2)^{\frac{1}{2}}}$, for $n = 10$. Although the distribution of $\frac{2^{\frac{1}{2}}n^{\frac{1}{2}}}{\sigma}(s-\bar{s})$ seems to follow the normal curve more closely than does the distribution of $(n-3)^{\frac{1}{2}}z$, the opposite seems to be true in the case of the distribution of $(n-1\frac{1}{2})^{\frac{1}{2}}z$. The distribution of $2^{\frac{1}{2}}n^{\frac{1}{2}}c_n\frac{z}{(z^2+2)^{\frac{1}{2}}}$, however, follows the normal curve quite closely.

The behavior of these distributions for $n = 10$ can be observed more easily in Figures 1, 2, and 3 in which the frequency curves of $\frac{2^{\frac{1}{2}}n^{\frac{1}{2}}}{\sigma}(s-\bar{s})$, $(n-3)^{\frac{1}{2}}z$, and $(n-1\frac{1}{2})^{\frac{1}{2}}z$ are respectively plotted together with the normal curve with unit standard deviation. The frequency curve of $2^{\frac{1}{2}}n^{\frac{1}{2}}c_n\frac{z}{(z^2+2)^{\frac{1}{2}}}$ was not plotted because of the fact that this curve follows the normal curve so closely that the two curves could not be distinguished when plotted on the scale used in the other three charts.

The most reasonable conclusion which can be drawn from Table 4 and Figures 1, 2, and 3 is that the departure of the exact distribution of s from the normal curve has very little effect in destroying the normality of the distribution of $2^{\frac{1}{2}}n^{\frac{1}{2}}c_n\frac{z}{(z^2+2)^{\frac{1}{2}}}$.

V. Values of the Factor, c_n

For the practical application of the approximations to the distributions of z and t presented in this thesis, a table of values of the factor, c_n, is required. Values of this factor, for values of n as high as 100, have been tabulated by Pearson[4, 6] and by Shewhart.[8] For values of n greater than 100, c_n may be calculated accurately to at least five significant figures by the following relation, given by Pearson[4] and by Deming and Birge[1]:

$$c_n = 1 - \frac{3}{4n} - \frac{7}{32n^2}. \tag{23}$$

Table 5 presents values of c_n for some large values of n, calculated by the present author. For values of n not included in this table, c_n may be calculated by means of equation (23) just as rapidly as by interpolation in the table.

VI. Summary and Conclusions

For values of n sufficiently large so that the distribution of s may be represented by a normal curve, the quantities,

$$2^{\frac{1}{2}} n^{\frac{1}{2}} c_n \frac{z}{(z^2 + 2)^{\frac{1}{2}}} \text{ and } 2^{\frac{1}{2}} n^{\frac{1}{2}} c_n \frac{t}{[t^2 + 2(n - 1)]^{\frac{1}{2}}},$$

are distributed as normal deviates with unit standard deviation. The results obtained by assuming a normal distribution of s are more accurate than those obtained by assuming that either $(n - 3)^{\frac{1}{2}} z$ or $(n - 1\frac{1}{2})^{\frac{1}{2}} z$ is distributed as a normal deviate with unit standard deviation. For extremely large values of n, the distribution of each of the above quantities tends to approach a normal curve with a mean of zero and unit standard deviation.

VII. References

(1) Deming, W. Edwards, and R. T. Birge, 1934. On the statistical theory of errors. Reviews of Modern Physics, 6: 119–161.

(2) Fisher, R. A., 1925. Applications of "Student's" distribution. Metron, 5: 90–104.

(3) Fisher, R. A., 1934. Statistical Methods for Research Workers, 5th ed. Oliver and Boyd, Edinburgh and London.

(4) Pearson, Karl, 1915. On the distribution of the standard deviations of small samples. Biometrika, 10: 522–529.

(5) Pearson, Karl, 1924. Tables For Statisticians And Biometricians, Part I, 2nd ed. Cambridge University Press, Cambridge.

(6) Pearson, Karl, 1931. Tables For Statisticians And Biometricians, Part II, 2nd ed. Cambridge University Press, Cambridge.

(7) Pearson, Karl, 1934. Tables Of The Incomplete Beta-Function. Cambridge University Press, Cambridge.

(8) Shewhart, W. A., 1931. Economic Control Of Quality Of Manufactured Product. D. Van Nostrand Co., New York.

(9) "Student," 1908. The probable error of a mean. Biometrika, 6: 1–25.

(10) "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. Biometrika, 11: 414–417.

(11) "Student," 1925. New tables for testing the significance of observations. Metron, 5: 105–120.

THE GEORGE WASHINGTON UNIVERSITY.

VIII. Appendix

TABLE 1

Exact values of I_z and approximate values, derived from tables of the normal probability integral, for $n = 10$

z	I_z			
	Exact value	"Student" approximation	Deming & Birge approximation	Hendricks approximation
−2.0	.0001	.0000	.0000	.0004
−1.8	.0002	.0000	.0000	.0006
−1.6	.0005	.0000	.0000	.0010
−1.4	.0011	.0001	.0000	.0018
−1.2	.0029	.0007	.0002	.0038
−1.0	.0075	.0041	.0018	.0086
− .8	.0199	.0171	.0098	.0211
− .6	.0527	.0562	.0401	.0535
− .4	.1304	.1448	.1218	.1307
− .2	.2816	.2984	.2799	.2817
.0	.5000	.5000	.5000	.5000
+ .2	.7184	.7016	.7201	.7183
+ .4	.8696	.8552	.8782	.8693
+ .6	.9473	.9438	.9599	.9465
+ .8	.9801	.9829	.9902	.9789
+1.0	.9925	.9959	.9982	.9914
+1.2	.9971	.9993	.9998	.9962
+1.4	.9989	.9999	1.0000	.9982
+1.6	.9995	1.0000	1.0000	.9990
+1.8	.9998	1.0000	1.0000	.9994
+2.0	.9999	1.0000	1.0000	.9996

TABLE 2

Exact values of t corresponding to different values of P and approximate values, derived from normal deviates, for n = 10

P	t			
	Exact value	"Student" approximation	Deming & Birge approximation	Hendricks approximation
.90	.129	.142	.129	.129
.80	.261	.287	.261	.261
.70	.398	.437	.396	.398
.60	.543	.595	.540	.544
.50	.703	.765	.694	.703
.40	.883	.954	.866	.884
.30	1.100	1.175	1.066	1.104
.20	1.383	1.453	1.319	1.386
.10	1.833	1.865	1.693	1.844
.05	2.262	2.222	2.017	2.290
.02	2.821	2.638	2.394	2.896
.01	3.250	2.921	2.650	3.389

TABLE 3

Exact values of t corresponding to different values of P and approximate values, derived from normal deviates, for n = 30

P	t			
	Exact value	"Student" approximation	Deming & Birge approximation	Hendricks approximation
.90	.127	.130	.127	.127
.80	.256	.263	.256	.256
.70	.389	.399	.389	.389
.60	.530	.543	.529	.530
.50	.683	.699	.680	.683
.40	.854	.872	.849	.854
.30	1.055	1.074	1.045	1.055
.20	1.311	1.328	1.293	1.312
.10	1.699	1.705	1.659	1.700
.05	2.045	2.031	1.977	2.047
.02	2.462	2.411	2.347	2.466
.01	2.756	2.670	2.598	2.764

TABLE 4

Ordinates of the normal curve with unit standard deviation and ordinates of the exact distribution functions of $\frac{2^{\frac{1}{2}}n^{\frac{1}{2}}}{\sigma}(s - \bar{s})$, $(n - 3)^{\frac{1}{2}}z$, $(n - 1\frac{1}{2})^{\frac{1}{2}}z$, *and* $2^{\frac{1}{2}}n^{\frac{1}{2}}c_n \frac{z}{(z^2 + 2)^{\frac{1}{2}}}$ *for* $n = 10$

Deviation from mean	Ordinates of distribution function				
	Normal deviate	$\frac{2^{\frac{1}{2}}n^{\frac{1}{2}}}{\sigma}(s - \bar{s})$	$(n - 3)^{\frac{1}{2}}z$	$(n - 1\frac{1}{2})^{\frac{1}{2}}z$	$2^{\frac{1}{2}}n^{\frac{1}{2}}c_n \frac{z}{(z^2 + 2)^{\frac{1}{2}}}$
−3.0	.0044	.0006	.0071	.0108	.0034
−2.5	.0175	.0085	.0181	.0254	.0156
−2.0	.0540	.0454	.0459	.0581	.0544
−1.5	.1295	.1356	.1092	.1234	.1306
−1.0	.2420	.2663	.2256	.2290	.2426
− .5	.3521	.3751	.3692	.3454	.3522
.0	.3989	.3999	.4400	.3991	.3990
+ .5	.3521	.3343	.3692	.3454	.3522
+1.0	.2420	.2245	.2256	.2290	.2426
+1.5	.1295	.1233	.1092	.1234	.1306
+2.0	.0540	.0560	.0459	.0581	.0544
+2.5	.0175	.0213	.0181	.0254	.0156
+3.0	.0044	.0068	.0071	.0108	.0034

TABLE 5

Values of c_n *for large values of* n

n	c_n	n	c_n
100	.99248	900	.99917
150	.99499	1000	.99925
200	.99624	2000	.99962
250	.99700	3000	.99975
300	.99750	4000	.99981
350	.99786	5000	.99985
400	.99812	10000	.99992
450	.99833	20000	.99996
500	.99850	30000	.99997
600	.99875	40000	.99998
700	.99893	50000	.99998
800	.99906	100000	.99999

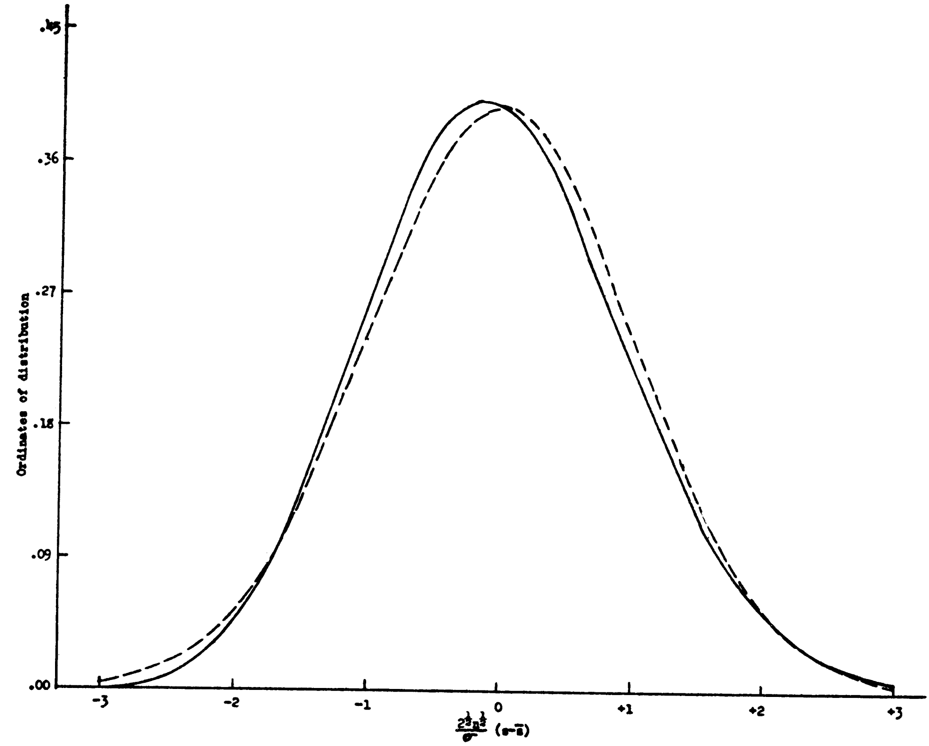

FIG. 1. EXACT DISTRIBUTION OF $\frac{2^{\frac{1}{2}}n^{\frac{1}{2}}}{\sigma}(s-\bar{s})$ FOR $n = 10$ AND NORMAL CURVE WITH UNIT STANDARD DEVIATION

————, exact distribution; ----------, normal curve

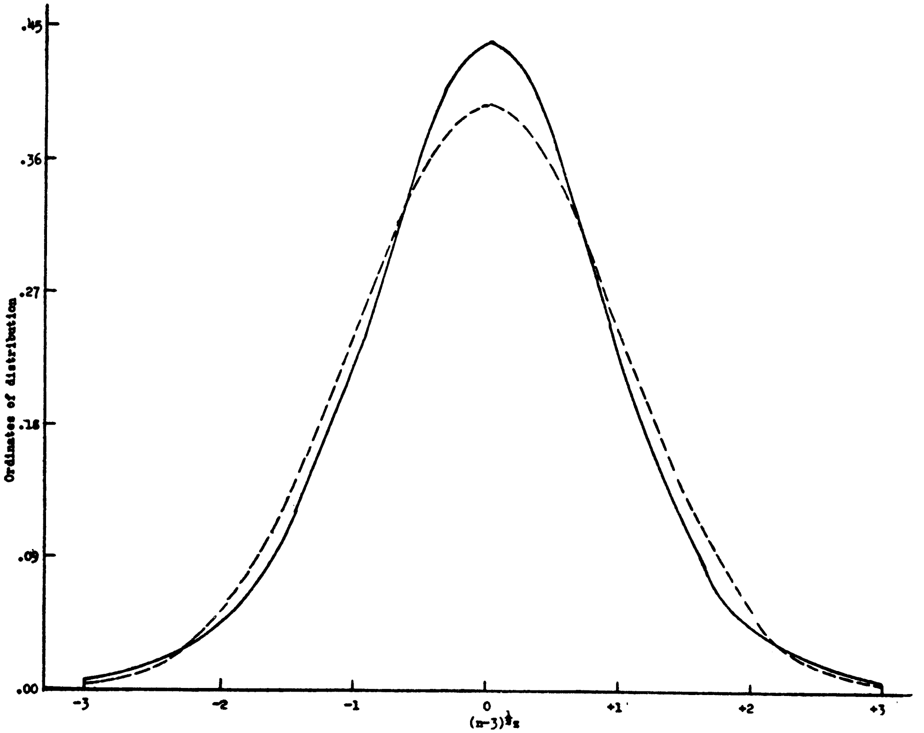

FIG. 2. EXACT DISTRIBUTION OF $(n-3)^{\frac{1}{2}}z$ FOR $n = 10$ AND NORMAL CURVE WITH UNIT STANDARD DEVIATION

————, exact distribution; ----------, normal curve

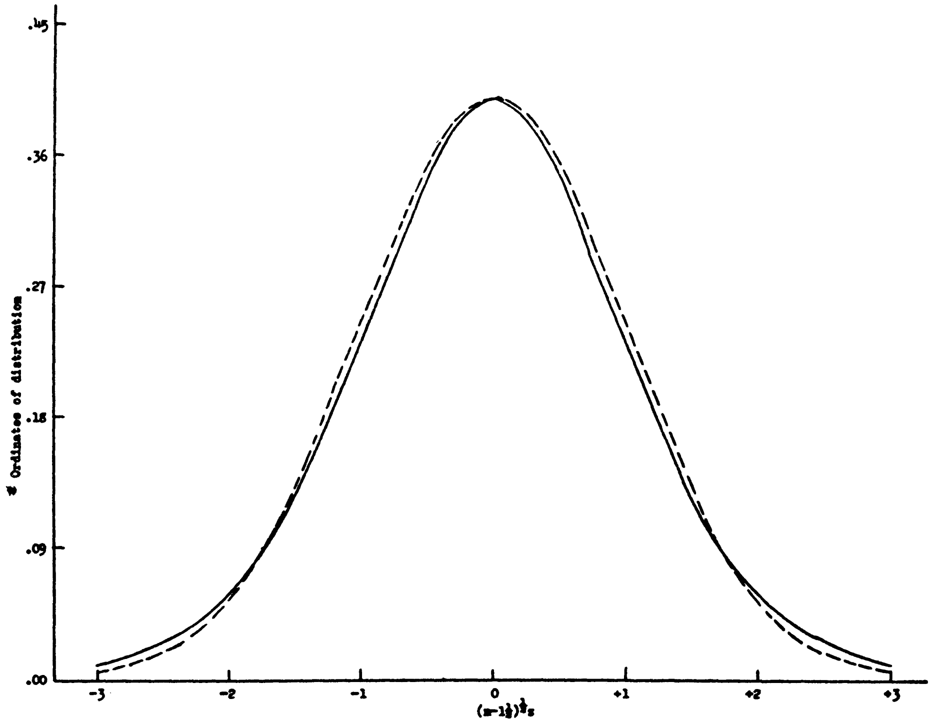

FIG. 3. EXACT DISTRIBUTION OF $(n - 1\frac{1}{2})^{\frac{1}{2}}z$ FOR $n = 10$ AND NORMAL CURVE WITH UNIT STANDARD DEVIATION

————, exact distribution; – – – – –, normal curve