THE INVERSE HYPERBOLIC SINE TRANSFORMATION ON STUDENT'S t FOR NON-NORMAL SAMPLES

By A. K. GAYEN
Statistical Laboratory, Calcutta

1. The effect of parent non-normality on the sampling distribution of Student's t has been the subject of many experimental and theoretical investigations (E.S. Pearson, 1928, 1929; P. R. Rider, 1931; G. A. Baker, 1932; M. S. Bartlett, 1935; R. C. Geary, 1936, 1947; H. L. Rietz, 1939; A. K. Gayen, 1949; H. Hyrenius, 1950). It has been shown that for moderately non-normal, especially symmetrical populations, a good approximation to the distribution of t is provided by the normal theory Student's distribution. However, the effect of skewness in the parent distribution may be considerable as has been particularly pointed out by Bartlett (1935), Geary (1936, 1947) and Gayen (1949). Appropriate probability corrections for the population measure of $\sqrt{\beta_1}$ were given by Geary in 1936. His work was extended by Gayen (1949), who derived further corrective functions for β_2 and β_1 for any sample size and prepared tables for the corresponding probability corrections for a considerable range of values of the sample size. Geary (1947) suggested the preparation of such tables, but the corrective functions obtained by him were not satisfactory; also they were based on large sample assumptions.

2. In this paper will be studied the possibility of normalizing the usual test function, in cases of samples from a non-normal population by the inverse hyperbolic sine transformation which has been suggested (M.H. Quenouille and F. J Anscombe, 1950) for normalizing the normal theory t-distribution. Quenouille (1950) suggested the transformation $y=\pm\sinh^{-1}(t^2/n)^{\frac{1}{2}}$, where sign of y is to be taken as negative when t is negative and as positive when t is positive,* n being the number of degrees of freedom, which was modified later, as $y=\pm\sinh^{-1}(3t^2/2n)^{\frac{1}{2}}$, by Anscombe (1950) which improves upon Quenouille's form. Anscombe (1950) has claimed that for large n, this form of transformation normalizes closely the t-distribution in cases of normal samples. We shall here investigate how far Anscombe's form of transformation tends to normalize the t-distribution for Edgeworth population, including terms up to β_1 and β_2, which as we know represents satisfactorily moderately non-normal universes.

3. Following Fisher's (1929) methods of deriving the moments of a statistic for any population, the expressions for the first four raw moments of Student's t for any population were obtained by Geary (1936). The expressions for the fifth and the sixth raw moments have been derived for our purpose by using the expressions for the semi-invariants L_5 and L_6 given by Geary (1947). They are given below along with

* The convention is superfluous for the transformation can be written with advantages as,

$$y = \text{Sinh}^{-1}(1/n^{\frac{1}{2}})t$$

those obtained by Geary after correcting some errors in his expression for the fourth moment. Here the i-th semi-invariant of the parent population has been defined, as usual, by $\lambda_i = k_i/k_2^{\frac{i}{2}}$, k_i being the i-th cumulant of the universe, as defined by Fisher(1929).

We have thus, up to terms in n^{-2}, (n being sample size)

$$\mu'_1(t) = -\lambda_3/2n^{1/2} - 3(2\lambda_3 - 2\lambda_5 + 5\lambda_3\lambda_4)/16n^{3/2} + \ldots$$
$$\mu'_2(t) = 1 + 2(1+\lambda_3^2)/n + 2(3 - \lambda_4 - 3\lambda_3\lambda_5 + 6\lambda_3^2\lambda_4)/n^2 + \ldots$$
$$\mu'_3(t) = -7\lambda_3/2n^{1/2} - (210\lambda_3 - 66\lambda_5 + 105\lambda_3\lambda_4 + 210\lambda_3^3)/16n^{3/2} + \ldots$$
$$\mu'_4(t) = 3 + (18 - 2\lambda_4 + 28\lambda_3^2)/n + (102 - 30\lambda_4 + 4\lambda_6 + 120\lambda^2_3 - 108\lambda_3\lambda_5 - 6\lambda_4^2$$
$$+ 168\lambda_3^2\lambda_4 + 120\lambda_3^4)/n^2 + \ldots$$
$$\mu'_5(t) = -55\lambda_3/2n^{1/2} + (-1925\lambda_3/8 + 333\lambda_5/8 - 425\lambda_3\lambda_4/16 - 1225\lambda_3^3/4)/n^{3/2} + \ldots$$
$$\mu'_6(t) = 15 + (180 - 30\lambda_4 + 370\lambda^2_3)/n + (1680 - 540\lambda_4 + 76\lambda_6 - 1698\lambda_3\lambda_5 + 7155\lambda_3^2/2$$
$$+ 1860\lambda^2_3\lambda_4 - 90\lambda_4^2 + 4200\lambda_3^4)/n^2 \div \ldots \quad \ldots \quad (1)$$

The values of $\mu'_5(t)$ and $\mu'_6(t)$ obtained as above were found to check with the normal theory expressions when λ_3 and higher λ's were considered to be negligible. It is found difficult to have expressions for the moments of the transformed variable y calculated directly in a similar fashion.

Assuming n to be large, we can suppose $3t^2/2n < 1$, (t can vary through a sufficiently extended range of values) so that y can be expanded in the power series form:

$$y = \sqrt{\frac{3}{2}}\left\{\frac{t}{n^{1/2}} - \frac{t^3}{4n^{3/2}} + \frac{27}{160}\frac{t^5}{n^{5/2}} + o(n^{-5/2})\right\} \quad \ldots \quad (2)$$

up to terms in n^{-3}.

Utilising formulae (1) and (2) (here $n+1$ was taken as the sample size so that the degrees of freedom remains the same as Anscombe's) it has been possible to derive the first four raw moments of y in the forms:

$$\mu'_1(y) = \sqrt{3/2}\,[-\lambda_3/2n + \{3\lambda_3/4 + 3\lambda_5/8 - 15\lambda_3\lambda_4/16\}/n^2 + o(n^{-2})]$$
$$\mu'_2(y) = 3/2n + (3/4 + 3\lambda_3^2)/n^2 + (3/2 - 3\lambda_4/2 - 24\lambda_3^2 - 9\lambda_3\lambda_5 + 18\lambda_3^2\lambda_4)/n^3 + o(n^{-3})$$
$$\mu'_3(y) = \sqrt{3/2}\,[-21\lambda_3/4n^2 + (111\lambda_3/8 + 99\lambda_5/16 - 315\lambda_3\lambda_4/32 - 315\lambda_3^3/16)/n^3 + o(n^{-3})]$$
$$\mu'_4(y) = 27/4n^2 + (27/4 - 9\lambda_4/2 + 63\lambda_3^2)/n^3 + (-216 + 9\lambda_4/2 + 9\lambda_6 - 1251\lambda_3^2/2$$
$$-243\lambda_3\lambda_5 - 27\lambda_4^2/2 + 378\lambda_3^2\lambda_4 + 270\lambda_3^4)/n^4 + o(n^{-4}) \quad \ldots \quad (3)$$

giving the following expressions for the central moments:

$$\mu_2(y) = 3/2n + (3/4 + 21\lambda_3^2/8)/n^2 + (3/2 - 3\lambda_4/2 - 183\lambda_3^2/8 - 135\lambda_3\lambda_5/16$$
$$+ 531\lambda_3^2\lambda_4/32)/n^3 + o(n^{-3})$$
$$\mu_3(y) = \sqrt{3/2}\,\{-3\lambda_3/n^2 + (93\lambda_3/8 + 9\lambda_5/2 - 45\lambda_3\lambda_4/8 - 249\lambda_3^3/8)/n^3\} + o(n^{-3})$$
$$\mu_4(y) = 27/4n^2 + (27/4 - 9\lambda_4/2 + 405\lambda_3^2/8)/n^3 + (-216 + 9\lambda_4/2 + 9\lambda_6 - 9099\lambda_3^2/16$$
$$-3483\lambda_3\lambda_5/16 - 27\lambda_4^2/2 + 10611\lambda_3^2\lambda_4/32 + 13905\lambda_3^4/64)/n^4 + o(n^{-4}) \quad \ldots \quad (4)$$

The expressions for the first two γ-coefficients of y are found to be:

$$\gamma_1(y) = -2\lambda_3/n^{1/2} + (37/4\,\lambda_3 + 3\lambda_5 - 15\lambda_3\lambda_4/4 - 31\lambda_3^3/4)/n^{3/2} + o(n^{-3/2})$$
$$\gamma_2(y) = (12\lambda_3^2 - 2\lambda_4)/n + (-411/4 - 180\lambda_3^2 + 6\lambda_4 - 6\lambda_4^2 + 4\lambda_6 - 63\lambda_3\lambda_5 + 88\lambda_3^2\lambda_4$$
$$+ 363\lambda_3^4/8)/n^2 + o(n^{-2}) \quad \ldots \quad (5)$$

The above expressions for the moments of t or y hold for any population with a sufficiently large sample size n.

We shall now see how the γ-coefficients of y compare with those for the t-statistic. Also for the specific Edgeworth population up to terms in λ_4 and λ_3^2 the corresponding expressions will be compared.

R. C. Geary's (1947) expressions for the γ-coefficients of t are as follows:

$$\begin{aligned}\gamma_1(t) &= -2\lambda_3/n^{1/2} + (-9\lambda_3+3\lambda_5-15\lambda_3\lambda_4/4-83\lambda_3^3/8)/n^{3/2}+o(n^{-3/2})\\ \gamma_2(t) &= (6-2\lambda_4+12\lambda_3^2)/n+(54-18\lambda_4+4\lambda_6+75\lambda_3^2-63\lambda_3\lambda_5-6\lambda_4^2+81\lambda_3^2\lambda_4\\ &\qquad +699\lambda_3^4/8)/n^2+o(n^{-2}) \quad \ldots \quad (6)\end{aligned}$$

For normal populations, as has been observed by Anscombe (1950), $\gamma_1(y)$ is evanescent and so is $\gamma_1(t)$, but $\gamma_2(y)$ involves terms in n^{-2} whereas $\gamma_2(t)$ starts with a term in n^{-1}. Accordingly for large n, y is approximately normal, having a zero mean and a variance

$$\mu_2(y) = 3/2n+3/4n^2+3/2n^3+\ldots \cong 3/(2n-1) \qquad \ldots \quad (7)$$

In case of non-normal population comparison of (5) and (6) shows that skewness for y and t are almost of the same strength; thus there may not be any advantage in the use of the transformed variate y over that of the t-statistic. The kurtosis coefficient, on the other hand, is a bit reduced in the case of y, the term $6/n$ being absent in $\gamma_2(y)$.

Thus the normality of the inverse hyperbolic sine transformation y does not appear to hold good with some convenient expression for variance even in moderately large samples. The skewness of the population will be a particularly disturbing factor in the situation.

4. Confining now to the Edgeworth population upto terms in λ_4 and λ_3^2 we find,

$$\begin{aligned}\mu'_1(y) &= \sqrt{3/2}\,\{-\lambda_3/2n+3\lambda_3/4n^2+\ldots\}\\ \mu_2(y) &= 3/2n+(3/4+21\lambda_3^2/8)/n^2+(3/2-183\lambda_3^2/8-3\lambda_4/2)/n^3+\ldots\\ \gamma_1(y) &= -2\lambda_3/n^{1/2}+(37\lambda_3/4-15\lambda_3\lambda_4/4-31\lambda_3^3/4)/n^{3/2}+\ldots\\ \gamma_2(y) &= (12\lambda_3^2-2\lambda_4)/n+(-411/4+6\lambda_4-180\lambda_3^2-6\lambda_4^2+88\lambda_3^2\lambda_4+363\lambda_3^4/8)/n^2+\ldots\end{aligned}$$

which may be compared with the corresponding expressions for t itself given below:

$$\begin{aligned}\mu'_1(t) &= -\lambda_3/2n^{1/2}-3\lambda_3/8n^{3/2}+\ldots\\ \mu_2(t) &= 1+(8+7\lambda_3^2)/4n+2(3-\lambda_4)/n^2+\ldots\\ \gamma_1(t) &= -2\lambda_3/n^{1/2}+(-9\lambda_3-15\lambda_3\lambda_4/4-83\lambda_3^3/8)/n^{3/2}+\ldots\\ \gamma_2(t) &= (6-2\lambda_4+12\lambda_3^2)/n+(54-18\lambda_4+75\lambda_3^2-6\lambda_4^2+81\lambda_3^2\lambda_4+699\lambda_3^4/8)/n^2+\ldots\end{aligned}$$

Obviously, the situation would remain as before. The variance will decrease considerably in the case of y but except in cases of suitable λ-coefficients, or a very large sample size the normality of y cannot possibly be assumed. The use of the transformed variable y cannot help therefore in such non-normal situations; only the probability corrections (Geary, 1936 and Gayen, 1949) obtained for λ_3 and λ_4 can provide here adequate corrections for the tail probabilities.

5. *Summary*:—The Inverse Hyperbolic Sine transformation $y=\sinh^{-1}\sqrt{3/2n}\,t$ has been suggested for the normalization of a variable follwoing the normal-theory Student's t-distribution (Quenouille, 1950 and Anscombe, 1950), and it has been shown that for large n, y is approximately normal with the variance $3/(2n-1)$. The effect of non-normality on the transformed variate y has been studied here by deriving the moment functions of y for any population. The study shows that the use of y in cases of non-normal samples is not advisable as its distribution is considerably sensitive to the population form. For moderate non-normality the expressions for γ-coefficients, retaining terms upto λ_4 and λ_3^2 of Edgeworth series, do not also appear to be negligible.

In such cases supplementary tables of probability correction (Geary, 1936; Gayen, 1947) should be used in conjunction with the estimated values of β_1 and β_2 for adequate evaluation of the tail probabilities.

My sincere thanks are due to Mr. N. Bhattacharya for the assistance rendered in the course of working out the formulae of this paper.

REFERENCES

ANSCOMBE, F. J. (1948): The transformation of Poisson, Binomial and Negative-Binomial data. *Biometrika*, 35, 240.

ANSCOMBE, F. J. (1950): Table of the hyperbolic transformation $\sinh^{-1}\sqrt{x}$. *J. Roy. Stat. Soc.*, 113, 228.

BAKER, G. A. (1932): Distribution of the means divided by the standard deviations of samples from non-homogeneous populations. *Ann. Math. Stat.*, 3, 1.

BARTLETT, M. S. (1935): The effect of non-normality on the t-distribution. *Proc. Camb. Phil. Soc.*, 31, 223.

EDGEWORTH, F. Y. (1906): The generalised law of error, or law of great numbers. *J. Roy. Stat. Soc.*, 69, 497.

FISHER, R. A. (1929): Moments and product moments of sampling distributions. *Proc. Lond. Math. Soc.* (2) 30, 199.

GAYEN, A. K. (1949): The distribution of student's t in random samples of any size drawn from non-normal universes. *Biometrika*, 36, 353.

GEARY, R. C. (1936): The distribution of student's ratio for non-normal samples. *J. Roy. Stat. Soc. Suppl.*, 3, 178.

GEARY, R. C. (1947): Testing for normality. *Biometrika*, 34, 209.

HYRENIUS, H. (1950) : Distribution of Student Fisher's t in samples from compound normal function. *Biometrika*, 37, 286.

PEARSON, E. S. and ADYANTHAYA, N. K. (1928, 1929): The distribution of frequency constants in small samples from non-normal symmetrical and skew populations. *Biometrika*, 20A, 356 and 21, 259.

QUENOUILLE, M. H. (1950): Mentioned by Anscombe (1950).

RIDER, P. R. (1931): On small samples from certain non-normal universes. *Ann. Math. Stat.*, 2, 48.

RIETZ, H. L. (1939): On the distribution of the 'Student's ratio for small samples from certain non-normal populations. *Ann. Math. Stat.*, 10, 265.

Paper received: June, 1952.