TRANSFORMATION OF NON-NORMAL FREQUENCY DISTRIBUTIONS INTO NORMAL DISTRIBUTIONS*

By

G. A. Baker

This investigation is undertaken for two reasons: (1) there has been a demand on the part of some statisticians for an analytic method of transforming non-normal distributions into normal distributions; and, (2) a non-normal distribution and the transformation necessary to transform it into a normal distribution serve to specify the distributions in random samples of estimates of the parameters of the original distribution in terms of the distributions of estimates of the parameters of a normal distribution in random samples. In this way valuable approximations to the distributions of the parameters of the original non-normal population may be secured.

PART I

TRANSFORMATIONS OF FREQUENCY DISTRIBUTIONS

Consider a non-normal frequency distribution represented by $f(x)\,dx$ where the origin is taken at some central point, say the mode, mean, or median, or near one of these points, and the scale is, or approximately is, the standard deviation of the distribution. We seek a function, φ, such that $f(x)\,dx$ transformed by the transformation, $\quad x = \varphi(u),$

becomes a normal distribution of total area $\sqrt{2\pi}$, mean zero and standard deviation unity, i.e.

* Presented at the May, 1932 meeting of the Illinois section of the American Mathematical Association.

$$(1) \qquad f[\varphi(u)]\cdot\varphi'(u)\,du = e^{-\frac{1}{2}u^2}du.$$

In a previous paper[1] expressions similar to (1) were regarded as differential equations which can be solved exactly in certain special cases. In the case of (1) it seems preferable to regard

$$(2) \qquad f[\varphi(u)]\,\varphi'(u) = e^{-\frac{1}{2}u^2}$$

as an identity in u. If it is assumed that f and φ are functions that can be represented by Maclaurin's expansions, which is a reasonable assumption regarding f and φ if f is near normal, the two members of (2) can be expanded and the coefficients of corresponding powers of u can be equated thus determining φ.

Suppose that

$$(3) \qquad f(x) = \sum_{n=0}^{\infty} A_n x^n$$

$$(4) \qquad \varphi(x) = \sum_{n=1}^{\infty} B_n x^n$$

$$(5) \qquad \varphi'(x) = \sum_{n=1}^{\infty} n\cdot B_n\cdot x^{n-1}.$$

Then (2) becomes

$$(6) \qquad \sum_{m=0}^{\infty} A_m\left[\sum_{n=1}^{\infty} B_n u^n\right]^m \cdot \sum_{n=1}^{\infty} n\cdot B_n\cdot u^{n-1} = 1-\frac{u^2}{2}+\frac{u^4}{2\cdot4}-\frac{u^6}{2\cdot4\cdot6}+\frac{u^8}{2\cdot4\cdot6\cdot8}-\frac{u^{10}}{2\cdot4\cdot6\cdot8\cdot10}+\cdots.$$

Hence

$$B_1 = \frac{1}{A_0}, \qquad B_2 = \frac{-A_1B_1^3}{2}, \qquad B_3 = -\frac{B_1}{6} - A_1B_1^2B_2 - \frac{A_2B_1^4}{3},$$

[1] Transformations of Bimodal Distributions, *Annals of Mathematical Statistics*, Vol. I, No. 4, Nov. 1930.

$$B_4 = -A_1\left(B_1^2 B_3 + \frac{B_1 B_2^2}{2}\right) - \frac{A_2 B_1^3 B_2}{2} - \frac{A_3 B_1^5}{4}$$

$$B_5 = \frac{B_1}{40} - A_1\left(B_1^2 B_4 + B_1 B_2 B_3\right) - A_2\left(B_1^3 B_3 + B_1^2 B_2^2\right) - A_3 B_1^4 B_2 - \frac{A_4 B_1^6}{5}$$

$$\begin{aligned} B_6 = &-A_1\left(B_1^2 B_5 + B_1 B_2 B_4 + \frac{B_1 B_3^2}{2}\right) \\ &- A_2\left(B_1^3 B_4 + \frac{B_1 B_2^3}{3} + 2 B_1^2 B_2 B_3\right) \\ &- A_3\left(B_1^4 B_3 + B_1^3 B_2^2\right) - A_4 B_1^5 B_2 - \frac{A_5 B_1^7}{6} \end{aligned}$$

$$\begin{aligned} B_7 = &-\frac{B_1}{336} - A_1\left(B_1^2 B_6 + B_1 B_2 B_5 + B_1 B_3 B_4\right) \\ &- A_2\left(B_1^3 B_5 + B_1 B_2^2 B_3 + B_1^2 B_3^2 + 2 B_1^2 B_2 B_4\right) \\ &- A_3\left(B_1^4 B_4 + \frac{B_1^2 B_2^3}{7} + 3 B_1^3 B_2 B_3\right) \\ &- A_4\left(B_1^5 B_3 - \frac{8}{7} B_1^4 B_2^2\right) - A_5 B_1^6 B_2 - \frac{A_6 B_1^8}{7} \end{aligned}$$

$$\begin{aligned} B_8 = &-A_1\left(B_1^2 B_7 + B_1 B_2 B_6 + B_1 B_3 B_5 + \frac{1}{2} B_1 B_4^2\right) \\ &- A_2\left(B_1^3 B_6 + B_1 B_2^2 B_4 + B_1 B_2 B_3^2 + 2 B_1^2 B_2 B_5 + 2 B_1^2 B_3 B_4\right) \\ &- A_3\left(B_1^4 B_5 + \frac{B_1 B_2^4}{4} + 3 B_1^3 B_2 B_4 + \frac{9}{8} B_1^3 B_3^2 + \frac{15}{8} B_1^2 B_2^2 B_3\right) \\ &- A_4\left(B_1^5 B_4 + 4 B_1^4 B_2 B_3 + \frac{3}{2} B_1^3 B_2^3\right) \\ &- A_5\left(B_1^6 B_3 + \frac{5}{2} B_1^5 B_2^2\right) \\ &- A_6 B_1^7 B_2 - \frac{A_7 B_1^9}{8}. \end{aligned}$$

The corresponding formulas for determining a function to transform a normal distribution of total area $\sqrt{2\pi}$, mean at zero and standard deviation of unity, into a given non-normal distribution are as follows. (The A's are the coefficients in the expansion of the given non-normal distribution and the B's are the coefficients of the transforming function.)

$$B_1 = A_0\,, \quad B_2 = \frac{A_1}{2}\,, \quad B_3 = \frac{A_2}{2} + \frac{B_1^3}{6}\,,$$

$$B_4 = \frac{1}{4} A_3 + \frac{1}{4} B_1^2 B_2$$

$$B_5 = \frac{1}{5} A_4 + \frac{1}{2} B_1^2 B_3 + \frac{1}{2} B_1 B_2^2 - \frac{1}{40} B_1^5$$

$$B_6 = \frac{1}{6} A_5 + \frac{1}{2} B_1^2 B_4 + \frac{1}{6} B_2^3 + B_1 B_2 B_3 - \frac{1}{8} B_1^4 B_2$$

$$\begin{aligned} B_7 = {} & \frac{1}{7} A_6 + \frac{1}{2} B_1^2 B_5 + \frac{1}{2} B_2^2 B_3 + \frac{1}{2} B_1 B_3^2 + B_1 B_2 B_4 \\ & - \frac{1}{8} B_1^4 B_3 - \frac{1}{7} B_1^3 B_2^2 + \frac{B_1^7}{336} \end{aligned}$$

$$\begin{aligned} B_8 = {} & \frac{1}{8} A_7 + \frac{1}{2} B_1^2 B_6 + \frac{1}{2} B_2^2 B_4 + \frac{1}{2} B_2 B_3^2 \\ & + B_1 B_2 B_5 + B_1 B_3 B_4 - \frac{1}{8} B_1^4 B_4 \\ & - \frac{1}{2} B_1^3 B_2 B_3 - \frac{3}{16} B_1^2 B_2^3 + \frac{1}{48} B_1^6 B_2 \,. \end{aligned}$$

These formulas give very simple results for the expression of the first few terms of the transforming function, φ, in terms of the coefficients of the given function. If the coefficients in the expansion of φ rapidly approach zero so that only a few terms are needed for a good approximation the method outlined should be effective. Edgeworth[2] has discussed at some length the transformation or "translation" of normal distributions into non-normal distributions and has given several methods of determining the coefficients of the transforming function. The formulas presented here are more simple but their practicability can be demonstrated only by numerical results in special cases. For practical purposes the left-hand member of (6) need represent the right-hand member accurately only in the interval, say $-2 \leq u \leq 2$.

ILLUSTRATION

For example, consider

$$f(x) = .9929\left(1+\frac{x}{10}\right)^{99} \cdot e^{-10x},$$

which is skewed noticeably in the positive direction but which is of a type that approaches a normal distribution as the skewness approaches zero. Then

$A_0 =$	$.9929$	$B_1 =$	1.0072
$A_1 =$	$-.1000$	$B_2 =$	$.0511$
$A_2 =$	$-.4887$	$B_3 =$	$.0050$
$A_3 =$	$.0823$	$B_4 =$	$-.0080$
$A_4 =$	$.1142$	$B_5 =$	$.0004$
$A_5 =$	$-.0270$		

[2] Bowley, A. L.-F. Y. Edgeworth's Contributions to Mathematical Statistics, pp. 65-78.

TABLE I

Comparison of the ordinates of the normal function, function with skewness of .2, and the skewed function transformed by the transformation $x = 1.0072u + .0511u^2 + .0050u^3 - .0080u^4$.

u	Normal curve*	Function with Skew .2†	Transformed skew curve†	Normal minus skew curve	Normal minus transformed skew curve
2.0	.053991	.049243	.0576	.0047	.0036
1.8	.078950	.076810	.0910	.0021	.0120
1.6	.110921	.112956	.1327	.0020	.0118
1.4	.149727	.157043	.1715	.0073	.0218
1.2	.194186	.206951	.2099	.0128	.0157
1.0	.241971	.259120	.2505	.0171	.0085
0.8	.289692	.308958	.2897	.0193	.0058
0.6	.333225	.351538	.3366	.0183	.0033
0.4	.368270	.382453	.3776	.0142	.0093
0.2	.391043	.398583	.3907	.0075	.0004
0.0	.398942	.398859	.3989	.0001	.0000
0.2	.391043	.383157	.3906	.0079	.0005
0.4	.368270	.354545	.3688	.0137	.0005
0.6	.333225	.316273	.3299	.0170	.0033
0.8	.289692	.272360	.2842	.0173	.0055
1.0	.241971	.226714	.2323	.0153	.0097
1.2	.194186	.182641	.1803	.0115	.0139
1.4	.149727	.142563	.1319	.0072	.0178
1.6	.110921	.107939	.0908	.0030	.0202
1.8	.078950	.079354	.0717	.0004	.0073
2.0	.053991	.056702	.0452	.0027	.0088

* These columns were taken from Luis R. Salvosa's tables, *Annals of Mathematical Statistics*, Vol. I, No. 2, p. 64 et seq.

† These values were calculated by interpolating in the above mentioned tables.

The ordinates of the normal curve, $f(x)$, and $f(x)$ transformed by the transformation determined by the first four β's are compared in Table I.

The ordinates of the transformed distribution are much nearer those of the normal curve over an interval that includes seventy-five per cent of the frequency but for the rest of the range considered the agreement is not so good. These facts indicate that more terms of the transforming function must be taken in order to secure close results for large values of $|u|$.

It is difficult to set up a rigorous criterion as to the number of $B's$ necessary to define adequately the transforming function, but the following considerations are of value in this connection.

Suppose that $f(x)$ may be adequately represented in the interval $a < x < b$ by m terms, i.e.

$$f(x) = A_0 + A_1 x + A_2 x^2 + \cdots + A_m x^m,$$

and that m is large enough so that the first m terms of the expansion of the normal function give an adequate representation of it. This is clearly possible since the expansions with which we are dealing converge uniformly in the open interval. Then the first m $B's$ may be determined so that the first m terms of $f(x)\,dx$ transformed by the transforming function determined by the m $B's$ are identical with the first m terms in the expansion of the normal function. In addition there will remain certain terms which may cause a serious discrepancy. For $f(x)\,dx$ becomes

$$A_0(B_1 + 2B_2 u + \cdots) + A_1(B_1 u + B_2 u^2 + \cdots)(B_1 + 2B_2 u + \cdots) +$$
$$\cdots + A_m (B_1 u + B_2 u^2 + \cdots)^m \cdot (B_1 + 2B_2 u + \cdots).$$

Let us assume that all $B_i's = 0$, $i > m$, and investigate the terms in u of degree higher than m.

Since the first terms of $f(x)\,dx$ transformed contribute few terms involving u^n, $n > m$, and the higher order terms have small coefficients, it is to be expected that if m $B's$ are used a good result will be obtained, at least for moderate values of u.

Some skewed distributions that differ considerably from normal may yield a rapidly converging sequence of $B's$, that is in case there is a natural relation of this kind existing between the non-normal and normal distributions.

The main reason for investigating the possibility of an easily determined transforming function that will transform a non-normal distribution into a normal distribution is the fact that the distributions in random samples of estimates of the parameters of the non-normal distribution can be expressed in terms of the transformation and the sampling distributions of the parameters of the normal distribution into which the non-normal distribution is transformed. This proposition is developed in Part II.

PART II

DISTRIBUTION OF THE ESTIMATES OF THE PARAMETERS OF NON-NORMAL DISTRIBUTIONS

Suppose that a variable x is distributed as $f(x)\,dx$ where $f(x)$ is such that it can be transformed into a normal distribution by means of a quadratic transformation,

$$x = ay + by^2 \tag{1}$$

Then $f(x)\,dx$ becomes

$$f(ay + by^2)(a + 2by)\,dy, \tag{2}$$

where y is normally distributed.

The total of a sample of n x's drawn at random from $f(x)$ is

$$(x_1 + x_2 + x_3 + \cdots + x_n), \tag{3}$$

which by virtue of (1) becomes

$$a(y_1 + y_2 + \cdots + y_n) + b(y_1^2 + y_2^2 + \cdots + y_n^2). \tag{4}$$

The coefficient of a in (4) is an estimate of the total of a sample of n of a normally distributed variable and the coefficient of b is an estimate of the second moment about a fixed point of a normally distributed variable which can be written as an estimate of the standard deviation squared plus the estimate of the mean squared. Thus (4) can be written as

$$n \cdot a \cdot \bar{m} + n \cdot b \cdot (\bar{\sigma}^2 + \bar{m}^2),$$

where the bar over m and σ denotes estimates of these parameters by means of samples. The distributions of $\bar{m}$ and $\bar{\sigma}$ are known and are independent. If the mean of distribution (2) is taken to be zero, then $\bar{m}$ is distributed as proportional to $e^{-\frac{n\cdot\bar{m}^2}{2}}$ and $y=a\bar{m}+b\bar{m}^2$ is distributed as proportional to

$$(5) \qquad \frac{e^{-\frac{1}{2}\cdot\frac{n(a^2+2by+a\sqrt{a^2+4by})}{2b^2}}}{\sqrt{a^2+4by}}, \qquad -\frac{a}{2b} \leq y \leq \infty.$$

The distribution of $b\bar{\sigma}^2$ is, except for a constant factor,

$$(6) \qquad z^{\frac{n-3}{2}} e^{-\frac{nz}{2b}}, \qquad 0 \leq z \leq \infty$$

if $n \geq 2$.

If two variables, x and y, are distributed as $f(x,y)\,dx\cdot dy$, then the probability that a value of $\varphi(x,y)=v$ is in dv is given as the surface area of the cylinder $\varphi(x,y)=v$ between $z=f(x,y)$ and $z=0$ times dv.[3]

In this case the probability function of $v=y+z$ is proportional to

$$(7) \qquad \int_{-\frac{a^2}{4b}}^{v} \frac{e^{-\frac{1}{2}\cdot\frac{n(a^2+2by+a\sqrt{a^2+4by})}{2b^2}}}{\sqrt{a^2+4by}} \cdot (v-y)^{\frac{n-3}{2}} \cdot e^{-\frac{n(v-y)}{2b}}\,dy.$$

Put $y=ax+bx^2$ and (7) becomes

$$(8) \qquad e^{-\frac{n}{2b}v} \int_{\frac{-a-\sqrt{a^2+4bv}}{2b}}^{\frac{-a+\sqrt{a^2+4bv}}{2b}} e^{\frac{na}{b}x} \cdot [v-ax-bx^2]^{\frac{n-3}{2}}\,dx, \qquad -\frac{a^2}{4b} \leq v \leq \infty.$$

[3] Baker, G. A.—Random Sampling from Non-Homogeneous Populations, *Metron*, Vol. VIII, No. 3, Feb. 1930.

If $f(x)$ can be transformed into a normal distribution by means of a cubic transformation.

(9) $$x = ay + by^2 + cy^3$$

then (3) becomes

(10) $$a(y_1 + y_2 + \cdots + y_n) + b(y_1^2 + y_2^2 + \cdots + y_n^2) + c(y_1^3 + y_2^3 + \cdots + y_n^3)$$

which can be written as $n(a\bar{m} + b\bar{m}^2 + c\bar{m}^3 + 3c\bar{\sigma}^2\bar{m} + b\bar{\sigma}^2)$.

Hence the means of samples of n are distributed as proportional to

(11) $$\int_\gamma^\beta \left[\frac{\nu - ax - bx^2 - cx^3}{3cx + b}\right]^{\frac{n-3}{2}} \cdot e^{-\frac{nx^2}{2} - \frac{n}{2}\left[\frac{\nu - ax - bx^2 - cx^3}{3cx + b}\right]} \frac{\sqrt{(3cx+b)^4 + (6c^2x^3 + 6bcx^2 + 2b^2x + ab + 3c\nu)^2}}{(3cx+b)^2}\,dx,$$

where γ/β represents the interval or intervals for which $\nu - ax - bx^2 - cx^3$ and $3cx + b$ have the same signs and ν varies from $-\infty$ to $+\infty$.

Suppose that the given frequency distribution can be transformed into a normal distribution by the transformation

$$x = ay + by^2,$$

and consider the expression for the estimation of the standard deviation squared of the x-distribution from a sample of n,

(12) $$\frac{(x_1^2 + x_2^2 + \cdots + x_n^2)}{n} - \left[\frac{(x_1 + x_2 + \cdots + x_n)}{n}\right]^2$$

which becomes

$$(13)\quad \frac{\left[(ay_1+by_1^2)+\cdots+(ay_n+by_n^2)\right]}{n}-\left[\frac{(ay_1+by_1^2)+\cdots+(ay_n+by_n^2)}{n}\right]^2$$

where y is normally distributed. In terms of the estimates of the mean and standard deviation of the y's (13) can be written as

$$2b^2\bar{\sigma}^4+a^2\bar{\sigma}^2+4ab\bar{\sigma}^2\bar{m}+4b^2\bar{\sigma}^2\bar{m}^2.$$

Hence the estimates of the standard deviations of the original population will be distributed as proportional to

$$\int_{-\infty}^{\infty} e^{-\frac{n}{2}\left[\frac{-(a^2+4abx)+\sqrt{(a^2+4abx+4b^2x^2)^2+8b^2\nu}}{4b^2}\right]}$$

$$(14)\qquad \cdot\left[\frac{-(a^2+4abx+4b^2x^2)+\sqrt{(a^2+4abx+4b^2x^2)^2+8b^2\nu}}{4b^2}\right]^{\frac{n-3}{2}}$$

$$\sqrt{1+\left[-\frac{(4ab+8b^2x)}{2b^2}+\frac{(a^2+4abx+4b^2x^2)(4ab+8b^2x)}{4b^2\sqrt{(a^2+4abx+4b^2x^2)^2+8b^2\nu}}\right]^2}\,dx,$$

where ν varies from 0 to $+\infty$.

The distributions of the estimates of other parameters and the distributions of the estimates of the mean and standard deviation for different transformations can be expressed in terms of the distributions of the mean and standard deviation of the resulting transformed normal distribution but it is obvious that the process becomes complicated.