NOTES ON THE DISTRIBUTION OF THE GEOMETRIC MEAN[1]

By Burton H. Camp

There are two transformation theorems which apply particularly well to the distribution of a product and therefore to the distribution of the geometric mean of a sample. Both are implicit in the known theory of the transformation of integrals, but it is useful to state them in forms which are especially adapted to probability theory. Several examples will be considered in which distributions of the geometric mean will be derived by using these theorems.

The first theorem may be stated as

Theorem A: *Let the point set q in an N-dimensional u-space be defined so that in q a given function of the u's, $F(u_1, u_2 \cdots u_N)$ has the property that*

$$\xi \leq F < \xi + d\xi. \tag{1}$$

Let $\bar{q}$ be the elementary volume of the point set q defined as an N-tuple integral

$$\int_q du_1 \cdots du_N$$

taken over q, having a value of order $d\xi$. Let

$$u_i = \theta(t_i), \qquad i = 1, 2, \cdots N \tag{2}$$

be continuous and differentiable monotonic functions of the t's with unique inverses

$$t_i = \theta^{-1}(u_i). \tag{3}$$

Let r be the point set in t-space corresponding to q in u-space under the transformation (2) with elementary volume given by the integral

$$\bar{r} = \int_r dt_1 \cdots dt_N. \tag{4}$$

If $J(\xi)$ is defined as $\dfrac{dt_1}{du_1} \cdots \dfrac{dt_N}{du_N}$ at a point in q for which $F = \xi$, and if, for all points in q,

$$\left| \frac{dt_1}{du_1} \cdots \frac{dt_N}{du_N} - J(\xi) \right| < M \cdot d\xi. \tag{5}$$

When M is a constant, independent of q, then the volume $\bar{r}$, is, except for terms of order $(d\xi)^2$, given by

$$\bar{q}|J(\xi)|. \tag{6}$$

[1] Read at a joint meeting of the American Mathematical Society and the Institute of Mathematical Statistics, Indianapolis, December 30, 1937.

The proof is immediate for we have

$$\begin{aligned}\bar{r} &= \left| \int_q \frac{dt_1}{du_1} \cdots \frac{dt_N}{du_N}\, du_1 \cdots du_N \right| \\ &= \left| \int_q \left[\frac{dt_1}{du_1} \cdots \frac{dt_N}{du_N} - J(\xi) \right] du_1 \cdots du_N + \int_q J(\xi)\, du_1 \cdots du_N \right| \\ &\leq \left| \int_q \left[\frac{dt_1}{du_1} \cdots \frac{dt_N}{du_N} - J(\xi) \right] du_1 \cdots du_N \right| + \bar{q} \cdot J(\xi).\end{aligned}$$

But, by (5), the integral in the last line has a value less than $\bar{q}\, M \cdot d\xi$, and $\bar{q}$ is of order $d\xi$. Therefore $\bar{r}$ differs from $\bar{q}|J(\xi)|$ by terms of order $(d\xi)^2$.

Let us now apply this theorem to a simple case. The volume of the set q, where $\xi \leq u_1 + \cdots + u_N < \xi + d\xi$, $u_i < a$, $i = 1, \cdots, N$, can easily be shown to be

$$\bar{q} = C(Na - \xi)^{N-1}\, d\xi.$$

Let $u_i = \log t_i$. Then it follows from the theorem that

$$\bar{r} = K\, e^{\xi}(Na - \xi)^{N-1}\, d\xi,$$

$\bar{r}$ being the volume of the point set r, where

$$\xi \leq \log (t_1 \cdots t_N) < \xi + d\xi. \tag{7}$$

By the use of (7) one can now use the geometrical method of finding the probability distribution of the geometric mean,

$$x = (t_1 \cdots t_N)^{1/N}, \tag{8}$$

of samples of N from the universe $\phi(t)\, dt$, provided that $\phi\ (t_1) \cdots \phi(t_N)$ is a continuous function of ξ. Unfortunately there do not appear to be many such ϕ functions. One that is of interest is

$$\phi(t)\, dt = kt^{2s}\, dt, \qquad 0 \leq t \leq e^a.$$

Let $D(\xi)d\xi$ represent the distribution of ξ. We have

$$\begin{aligned}D(\xi)\, d\xi &= \int_r \phi(t_1) \cdots \phi(t_N)\, dt_1 \cdots dt_N = \int_r k^N (t_1, \cdots t_N)^{2s}\, dt_1 \cdots dt_N \\ &= \bar{r} k^N e^{2s\xi} = Ce^{\xi + 2s\xi}(Na - \xi)^{N-1}\, d\xi.\end{aligned}$$

Thence we obtain as the distribution of x:

$$f(x)\, dx = C_1 x^{2sN+N-1}(a - \log x)^{N-1}\, dx.$$

The form of $f(x)$ in the special case in which $s = 0$ and ϕ is a rectangle has been found by other authors,[2] and is

$$f(x)\, dx = C_1 x^{N-1}(a - \log x)^{N-1}\, dx.$$

[2] E.g. see S. Kullback, "An application of characteristic functions to the distribution problem of statistics," *Annals of Mathematical Statistics*, vol. 5 (1934), pp. 263–270.

The second transformation theorem to be used may be stated as.

THEOREM B: *Let $\psi(u)du$ be the probability element for a given universe and let the sample $(u_1, u_2. \cdots u_N)$ be taken. Let the statistic $\xi = \gamma(u_1, u_2, \cdots u_N)$ have the distribution $F(\xi)d\xi$. If the transformation* (2), *satisfying the conditions imposed on it in Theorem A be applied both to the universe and to the statistic, yielding $\varphi(t)dt$ and $\xi = g(t_1, \cdots t_N)$ respectively, then the element of distribution of ξ, as obtained from ϕ, is, as before, $F(\xi)d\xi$.*

The proof is straight forward, for the distribution of ξ, as obtained from $\psi(u)du$ is given by

$$\int_q \psi(u_1) \cdots \psi(u_N)\, du_1, \cdots du_N$$

and, as obtained from $\phi(t)dt$, it is

$$\int_r \phi(t_1) \cdots \phi(t_N)\, dt_1, \cdots dt_N$$

where q is the set in u-space where $\xi \leq \gamma < \xi + d\xi$ and r is the set in t-space where $\xi \leq g < \xi + d\xi$. It is clear that these two integrals have the same value because of the relation

$$\psi(u)\, du = \psi(\theta(t)) \frac{d\theta(t)}{dt} \cdot dt \equiv \varphi(t)\, dt$$

and the unique correspondence between the points of q and r set up by the transformation (2), with its unique inverse (3).

This theorem is particularly well adapted to the derivation of the distribution of the geometric mean because of the simple logarithmic transformation connecting the sum and the product of N numbers, and because several distributions of the sum are already known. Two of these cases will now be presented.

EXAMPLE 1. Let x be the geometric mean (8) of the sample of N from a universe with distribution law

$$\phi(t)\, dt = \frac{(\log t)^{\mathrm{p}-1}}{t^2\, \Gamma(\mathrm{p})} dt \qquad (t > 1). \tag{9}$$

Then the distribution of x is

$$f(x)\, dx = \frac{N^{N\mathrm{p}} (\log x)^{N\mathrm{p}-1}}{x^{N+1}\, \Gamma(N\mathrm{p})} dx \qquad (x > 1), \tag{10}$$

and it is to be noticed that x has the same type of distribution as t.

To prove (10), first let $\xi = (u_1 + \cdots + u_N)/N$, where the u's are a sample from a Type III universe,

$$\psi(u)\, du = \frac{e^{-u} u^{\mathrm{p}-1}}{\Gamma(\mathrm{p})} du \qquad (u > 0).$$

Irwin[3] has shown that the distribution of ξ is

$$F(\xi)\,d\xi = \frac{Ne^{-N\xi}(N\xi)^{Np-1}}{\Gamma(Np)}\,d\xi. \tag{11}$$

Making the transformation $u = \log t$, we have

$$\xi = \log (t_1 \cdots t_N)^{1/N}, \qquad \phi(t)\,dt = \frac{(\log t)^{p-1}}{t^2\Gamma(p)}\,dt, \qquad t > 1,$$

and $F(\xi)d\xi$ is unchanged. We now obtain $f(x)dx$ by substituting $\xi = \log x$ in (11).

EXAMPLE[4] 2. If x is the geometric mean (8) of a sample of N from a universe whose distribution is

$$\phi(t)\,dt = \frac{1}{tc\sqrt{2\pi}}\,e^{-\frac{1}{2c^2}\left(\log \frac{t}{G}\right)^2}\,dt, \qquad (c, t, G > 0), \tag{12}$$

the distribution of x is

$$f(x)\,dx = \frac{\sqrt{N}}{xc\sqrt{2\pi}}\,e^{-\frac{N}{2c^2}\left(\log \frac{x}{G}\right)}\,dx, \qquad (x > 0). \tag{13}$$

To prove this, one begins with the arithmetic mean ξ and the universe,

$$\psi(u)\,du = \frac{1}{c\sqrt{2\pi}}\,e^{-\frac{1}{2c^2}(u-\bar{u})^2}\,du. \quad \text{Here } F(\xi)\,d\xi = \frac{\sqrt{N}}{c\sqrt{2\pi}}\,e^{-\frac{N}{2c^2}(\xi-\bar{u})^2}\,d\xi.$$

Again using $u = \log t$, one obtains $\xi = \log (t_1, \cdots, t_N)^{1/N}$ and

$$\phi(t)\,dt = \frac{1}{tc\sqrt{2\pi}}\,e^{-\frac{1}{2c^2}\left(\log \frac{t}{G}\right)^2}\,dt, \text{ where } G = e^{\bar{u}} > 0,$$

and $F(\xi)\,d\xi$ is unchanged. To get (13) one substitutes $\xi = \log x$ in $F(\xi)\,d\xi$.

Again it follows that the geometric mean has the same distribution as the universe except for a change in one of the parameters (c). This frequency curve has other interesting features. It was developed by Galton and McAlister[5] by quite a different method and was called the curve of equal facility. They were seeking for a distribution $\phi(t)$ which would have the characteristic that, if t and t' were two observations differing from G by the same relative amount, $(G - t)/t = (t' - G)/G$, they would have equal probabilities. McAlister noted various properties of ϕ, including the fact that G was actually its geometric mean, and that it was not the same as the mode or the arithmetic mean. Certain properties which he did not mention are the following:

(*i*) If one draws a sample from a universe with the distribution ϕ in order to

[3] *Biometrika*, vol. 19 (1927), p. 229; see also A. Church, *Biometrika*, vol. 18 (1926), p. 336.

[4] This distribution can also be obtained by the method of A. T. Craig, *American Journal of Mathematics*, vol. 54 (1932), p. 362, but it would be difficult to evaluate his integral without the substitution which would be suggested if the distribution were known.

[5] *Proceedings of the Royal Society*, vol. 29 (1879), pp. 365, 367.

determine G, the geometric mean of the universe, the maximum likelihood solution is x, the geometric mean of the sample.

(*ii*) The modal point of the sampling distribution (f) approaches G as a limit as N becomes infinite.

(*iii*) One can devise a function s of the sample analogous to but different from Student's s, and show that x/s has a distribution independent of the parameters of G and c of the universe. To do this it is necessary first to extend the second transformation theorem so as to include cases where the number of statistics (functions of the sample) being obtained simultaneously is greater than one. This is not difficult, but since the analogous tests for significance have been developed for the normal universe it would not be particularly useful, for if the observations are distributed in accordance with $\phi(t)$ their logarithms are distributed normally, and their logarithms can equally well be used for testing significance.

(*iv*) If one uses the curve of equal facility instead of the normal curve as the distribution of biological lengths, then any power of such lengths, in particular the third power, which is supposed to be approximately proportional to weights, would also be distributed in the same manner, except for a change in the parameters. This is a property which the normal curve does not have. It raises the question: Can biological lengths be represented by the curve of equal facility? The remainder of this paper will be devoted to a discussion of this question and cognate matters.

The curve of equal facility may be made to approach as a limit the normal curve if the origin be moved indefinitely to the left. This is almost intuitively evident from a consideration of the hypotheses under which the two curves were derived by Galton and McAlister. It is also indicated by the behavior of the lower moments. Let ν_i refer to the ith moment of (12) relative to the origin of t, μ_i to the corresponding moment relative to the arithmetic mean. It is easy to show that

$$\nu_i = G^i e^{\frac{1}{2}i^2c^2}, \quad i = 0, 1, \cdots, \qquad \nu_1 = \bar{t} = Gh, \text{ where } h = e^{\frac{1}{2}c^2}, \tag{14}$$

$$\mu_2 = G^2h^2(h^2 - 1), \qquad \mu_3 = G^3h^3(h^6 - 3h^3 + 2),$$
$$\mu_4 = G^4h^4(h^{12} - 4h^6 + 6h^2 + 3), \tag{15}$$

$$\begin{cases} \alpha_3 = \mu_3/\mu_2^{3/2} = (h^2 + 2)\,(h^2 - 1)^{1/2}, \\ \alpha_4 = \mu_4/\mu_2^2 \;\; = (h^2 - 1)^4 + 6(h^2 - 1)^3 + 15(h^2 - 1)^2 + 16(h^2 - 1) + 3. \end{cases} \tag{16}$$

From (16) it follows that as h approaches unity α_3 and α_4 approach their normal values, 0 and 3, respectively. If at the same time μ_2 is kept constant, it follows from (15) that G^2 and therefore $\bar{t}$ become infinite. So the origin is moved an infinite distance to the left.

The question, then, whether the curve of equal facility may be used equally well with the normal curve to represent biological lengths depends on whether in practical cases the natural choice of origin, which is the position indicated by

zero length, is such as to make the two curves practically indistinguishable. This is apparently the situation in the case of human statures. For 8585 adult males born in the British Isles[6] the values of the several constants, obtained by so fitting $\phi(t)$ to the observations that the mean and standard deviations agree, are as follows: $\bar{t} = 67.46$ in., $G = 67.411$, $\sigma = 2.56$, $h = 1.00072$, observed $\alpha_3 = 0.0125$, α_3 for $\phi = 0.11$; observed $\alpha_4 = 3.149$, α_4 for $\phi = 3.02$. Thus for the curve of equal facility α_3 is further from the observed value than for the normal curve, but α_4 is nearer to its observed value. In both cases the difference is unimportant. A graph of both curves[7] would not make it clear to the eye which of the two fitted the data better.

It would be expected that the distribution of the cubes of these statures, being roughly proportional to the weights of the men, would not be normally distributed. This also can be verified easily, for the distribution of $(y = t^k)$ from $\phi(t)dt$ is $\phi(y)dy$ except that ck replaces c, and G^k replaces G. So the distribution of cubes is:

$$F(y)\,dy = \frac{1}{3cy\sqrt{2\pi}}\, e^{-\frac{1}{18c^2}\left(\log \frac{y}{G^3}\right)^2} dy.$$

If this curve is fitted to the cubes of the statures, $\alpha_3 = 0.23$, and $\alpha_4 = 3.21$. Both are considerably further from their normal values than before. For this case the corresponding value of h is 1.0064. It is the closeness of this quantity to unity, or in other words the smallness of the coefficient of variation, $100\ \sigma/\bar{t} = 100\ (h^2 - 1)^{1/2}$, which determines how close the curve is to the normal. For the statures $\sigma/\bar{t} = 0.0379$. For the cubes of the statures[8] $\sigma/\bar{t} = 0.269$. Its values in certain other cases[9] are: length of forearm 0.05, chest circumference 0.08, strength of grip 0.26, visual acuity 0.39. It appears to be evident, therefore, that for many types of biometric measurements, especially lengths, which we know can be represented well by the normal curve, the curve of equal facility is practically just as good. In a given case it may fit a little better or a little worse. If we wish the distribution of the arithmetic mean as obtained by sampling from such data we may find it by supposing the universe normal; if we wish the distribution of the geometric mean we may find it by supposing the universe of a curve of equal facility. This device of substituting for the normal curve another type of curve which is equally good in practical cases, in order to find the distribution of a statistic which cannot be found easily for the normal curve, may perhaps be useful also for other statistics than the geometric mean.

WESLEYAN UNIVERSITY.

---

[6] G. Udny Yule and M. G. Kendall, *An Introduction to the Theory of Statistics*, London, 1937, pp. 94, 116, 157, 163, 187.

[7] Such as on page 187, Yule and Kendall.

[8] For the weights of a similar group of men $\sigma/\bar{t} = 0.137$, and thus the two curves would be more nearly alike if fitted to weights than if fitted to the cubes of these statures.

[9] From a long list with values ranging from 0.0049 to 0.5058, compiled by Raymond Pearl, *Medical Biometry and Statistics*, Philadelphia (1930), pp. 347–9.