ON THE ASYMPTOTIC DISTRIBUTION OF THE SUM OF POWERS OF UNIT FREQUENCY DIFFERENCES[1]

By Bradford F. Kimball

New York State Department of Public Service

1. Summary. Since the "unit" frequency differences (see (2.2) below) are dependent, the usual methods for establishing the normal character of the asymptotic distribution of the sum of random variables fail.

However, the essential character of the distribution is disclosed by the integral functional relationship (3.6). From this it is possible to show that for large samples the distribution approximates "stability" in the normal sense ([2] and Lemma 2).

Using the condition that the third logarithmic derivative of the characteristic function is uniformly bounded for all n on a neighborhood of $t = 0$ one can prove that the asymptotic distribution exists and is normal.

2. Introduction. Consider a one dimensional statistical universe characterized by a cumulative frequency function (*cdf*) $F(x)$ which is continuous. Consider an ordered random sample x_i of size N such that

$$x_i \leq x_{i+1}, \qquad i = 1 \text{ to } N - 1. \tag{2.1}$$

Consider frequency differences u_i defined by

$$\begin{aligned} u_1 &= F(x_1), \qquad u_{N+1} = 1 - F(x_N), \\ u_{i+1} &= F(x_{i+1}) - F(x_i), \qquad i = 1 \text{ to } N - 1. \end{aligned} \tag{2.2}$$

Thus

$$\sum_{N+1} u_i \equiv 1, \tag{2.3}$$

and the formal integral of the probability density function (*pdf*) of the u_i taken over the complete sample space of x_i can be written as

$$N! \int du_1 \, du_2 \cdots du_{h-1} \, du_{h+1} \cdots du_{N+1} = 1, \tag{2.4}$$

where u_h is any u_i which it is found convenient to omit, and the region of integration is the N-fold Euclidean space bounded by the coordinate hyperplanes

$$u_i = 0, \qquad i \neq h, \qquad i = 1, 2, \cdots N + 1,$$

and the hyperplane

$$u_1 + u_2 + \cdots + u_{h-1} + u_{h+1} + \cdots + u_{N+1} = 1. \tag{2.5}$$

(See [1]).

[1] This is the second paper in connection with the subject announced in Abstract No. 9, *Annals of Math. Stat.*, Vol. 17 (1946), p. 502; and Abstract No. 331, *Bull. Am. Math. Soc.*, Vol. 52 (1946), p. 827. For first paper, see [1].

Consider a test function y_M defined by

$$y_M = \sum_M u_i^p, \quad p > 0, \quad M \leqq N + 1, \tag{2.6}$$

where p is a real positive number, M is an integer less than or equal to $N + 1$ and such that if $M < N + 1$ the u_i which are to be omitted may be arbitrarily selected, but the subscripts indicating the order relation (2.2) are for the present retained.

Consider the case where N is odd and M is even, and set

$$N = 2n + 1, \quad M = 2m. \tag{2.7}$$

Divide the set of $N + 1$ frequency differences u_i defined by (2.2) into two subsets such that each subset contains $n + 1$ differences of which exactly m are included in the test function (2.6). Now let N become infinite over odd numbers $N_1, N_2, \cdots$. In other words the sample size is to increase without limit. For each sample size N_j in such a sequence let M_j be an even number such that

$$M_j \leq N_j + 1 \tag{2.8}$$

and such that the ratio M_j/N_j is controlled for large values of N by

$$\lim_{N\to\infty} M_j/N_j = \text{constant } c, \quad 0 < c \leq 1. \tag{2.9}$$

As above for each step in the sequence the set of $N_j + 1$ frequency differences u_i is divided into two subsets of $n_j + 1$ frequencies each with

$$N_j = 2n_j + 1, \quad M_j = 2m_j, \tag{2.10}$$

such that m_j frequencies of each subset are included in the test function

$$y_{M_j} = \sum u_i^p. \tag{2.11}$$

Now we note that for a random sample of size N taken from the above universe, the characteristic function $G_N(t; y_M)$ may be defined by

$$G_N(t; y_M) = N! \int e^{ity_M} \, du_1 \, du_2 \cdots du_N \tag{2.12}$$

taken over region in Euclidean space of N dimensions as indicated for the integral (2.4), taking index h equal to $N + 1$.

3. Proof of integral relationship—Lemma 1. For simplicity of notation drop subscripts from M_j, N_j, n_j and m_j. We separate the test function y_M into two parts y_m and $y_{m'}$ such that

$$y_M = y_m + y_{m'} = \sum_m u_i^p + \sum_{m'} u_i^p, \quad m = m' = M/2 \tag{3.1}$$

where the m frequency differences u_i in y_m are those included in first subset and those contained in $y_{m'}$ are those of the original M frequencies included in the second subset (see (2.10) and (2.11)).

The formal integral defining $G_N(t; y_M)$ may be written

$$(3.2)\quad G_N(t; y_M) = \Gamma(2n + 2) \int_{R_2} e^{ity_m}\, du_1 \cdots du_{n+1} \int_{R_1} e^{ity_{m'}}\, du_{n+2} \cdots du_{2n+1},$$

where

R_2 = $2n + 1$ dimensional Euclidean space bounded by coordinate hyperplanes and plane $\sum_{2n+1} u_i = 1$,

R_1 = n dimensional Euclidean space bounded by the coordinate hyperplanes and the plane

$$(3.3)\quad \begin{aligned} u_{n+2} + u_{n+3} + \cdots + u_{2n+1} &= 1 - w, \\ w = u_1 + u_2 + \cdots &+ u_{n+1}. \end{aligned}$$

Now introduce the transformation to u_i'

$$(3.4)\quad u_i'(1 - w) = u_i, \qquad i = n + 2, n + 3, \cdots, 2n + 1, 2n + 2.$$

Thus we have

$$\sum_{n+1} u_i' \equiv 1,$$

and the n u_i' involved in the integration are bounded above by the hyperplane $\sum_n u' = 1$. The Jacobian is $(1 - w)^n$.

Similarly under transformation

$$(3.5)\quad \begin{aligned} v_i w = u_i, \qquad i &= 1, 2, \cdots, n + 1, \\ \sum_{n+1} v_i &\equiv 1. \end{aligned}$$

Let v_i, $i = 1, 2, \cdots n$ and w replace the remaining variables of integration. Thus the region of integration of these v_i is $v_i \geq 0$ with the hyperplane $\sum_n v_i = 1$ furnishing the upper bound. The Jacobian of the transformation is w^n.

The regions of integration of these new variables u_i' and v_i are seen to be independent of each other and of w. Noting effect of above transformations on y_m and $y_{m'}$, the integral (3.2) will be found to reduce to the following form:

$$(3.6)\quad G_N(t; y_M) = \frac{\Gamma(2n + 2)}{\Gamma^2(n + 1)} \int_0^1 w^n(1 - w)^n G_n(tw^p; y_m) G_n(t(1 - w)^p; y_m)\, dw,$$

where

$$N = 2n + 1, \qquad M = 2m.$$

LEMMA 1. *This functional relationship holds for all values of N and M subject to the condition that N be an odd integer and M an even integer. One may note that a similar integral functional relationship will hold for any partition $(n_0 n_1)$ of the $N - 1$ free frequency differences such that*

$$n_0 + n_1 = N - 1, \qquad m_0 + m_1 = M,$$

with corresponding changes in the Gamma functions which precede the integral.

In order to find out what happens when N becomes large the partially normalized test function z_M is introduced. This is defined by

$$z_M = (y_M - \bar{y}_M)(N + 1)^p/\sqrt{M}, \tag{3.7}$$

where (cf. [1], formula (3.1))

$$\bar{y}_M = E(y_M) = \frac{M\Gamma(N + 1)\Gamma(p + 1)}{\Gamma(N + 1 + p)}. \tag{3.8}$$

I have referred to z_M as a partially normalized variable since

$$\begin{aligned} E(z_M) &= 0, \\ \lim_{N\to\infty} E(z_M^2) &= \Gamma(2p + 1) - \Gamma^2(p + 1) - cp^2\Gamma^2(p + 1), \end{aligned} \tag{3.9}$$

where this limit can be shown to be greater than zero for

$$\begin{aligned} p &\neq 1, \qquad 0 < c \leq 1, \\ p &= 1, \qquad 0 < c < 1. \end{aligned} \tag{3.10}$$

Recalling the separation of the test function into two parts (see (3.1)) we define $\bar{y}_m$ and $\bar{y}_{m'}$ by

$$\bar{y}_m = \bar{y}_{m'} = \frac{m\Gamma(n + 1)\Gamma(p + 1)}{\Gamma(n + 1 + p)} \tag{3.11}$$

with

$$M = 2m, \qquad N = 2n + 1.$$

From Stirling's formula it can then be shown that

$$(N + 1)^p\bar{y}_M/\sqrt{M} = (2^p/\sqrt{2})2[(n + 1)^p\bar{y}_m/\sqrt{m}] + o(1), \tag{3.12}$$

where $o(1)$ goes to zero as N and M become infinite subject to the condition (2.9). Thus if we define z_m and $z_{m'}$ by

$$z_m = (y_m - \bar{y}_m)(n + 1)^p/\sqrt{m}, \qquad z_{m'} = (y_{m'} - \bar{y}_{m'})(n + 1)^p/\sqrt{m}, \tag{3.13}$$

since

$$y_M = y_m + y_{m'}$$

and

$$(N + 1)^p/\sqrt{M} = (2^p/\sqrt{2})(n + 1)^p/\sqrt{m},$$

it follows that

$$z_M = (2^p/\sqrt{2})(z_m + z_{m'}) + o(1). \tag{3.14}$$

Hence if we denote the characteristic function of the distribution of the partially normalized test function z_M by $G_N(t;\, z_M)$ and proceed to develop an integral functional relationship similar to (3.6), one arrives at

$$G_N(t;\, z_M) = e^{ito(1)} \frac{\Gamma(2n+2)}{\Gamma^2(n+1)} \int_0^1 w^n (1-w)^n G_n[t(2w)^p/\sqrt{2};\, z_m] \cdot G_n[t2^p(1-w)^p/\sqrt{2};\, z_m]\, dw \tag{3.15}$$

with

$$N = 2n + 1, \qquad M = 2m.$$

4. Resulting functional relationship when N becomes large. The second lemma shows that the functional equation satisfied by the characteristic function of a normal distribution is approximated when N is large. Suppose we now set

$$w = (1+s)/2, \qquad 1 - w = (1-s)/2, \qquad dw = ds/2. \tag{4.1}$$

Substituting in (3.15) we have

$$G_N = \frac{e^{ito(1)}\,\Gamma(2n+2)}{2^{2n+1}\,\Gamma^2(n+1)} \int_{-1}^{+1} (1-s^2)^n G_n[t(1+s)^p/\sqrt{2};\, z_m] \, G_n[t(1-s)^p/\sqrt{2};\, z_m]. \tag{4.2}$$

Set

$$H(t,\, s) = G_n[t(1+s)^p/\sqrt{2};\, z_m] G_n[t(1-s)^p/\sqrt{2};\, z_m]. \tag{4.3}$$

Then

$$H_s = G_n' G_n tp(1+s)^{p-1}/\sqrt{2} - G_n G_n' tp(1-s)^{p-1}/\sqrt{2}. \tag{4.4}$$

Using law of mean write

$$H(t,\, s) = H(t,\, 0) + sH_s[t,\, h(s)], \qquad 0 < |\, h(s)\,| < s. \tag{4.5}$$

Substituting in (4.2) we have

$$e^{-ito(1)} G_N = H(t,\, 0) + \frac{\Gamma(2n+2)}{2^{2n}\,\Gamma^2(n+1)} \int_0^1 H_s[t,\, h(s)](1-s^2)^n s\, ds. \tag{4.6}$$

With $E(z_m) \equiv 0$, from the fact that the limiting variance of z_m is bounded (see (3.9)) it follows that the first derivative of its characteristic function remains bounded in any finite interval, for all n ([3], p. 90). Thus

$$|\, G_n'(t;\, z_m)\,| < A, \qquad 0 \le |\, t\,| \le D, \qquad \text{for all } n. \tag{4.7}$$

For case $p \ge 1$, by virtue of condition (4.7) H_s will remain bounded over interval of integration of (4.6) as N becomes infinite. Let B denote such upper bound of the absolute value of H_s. Then, carrying out the integration

$$\text{absolute value of integral} < \frac{B\Gamma(2n+2)}{2^{2n}\,\Gamma^2(n+1)} \frac{1}{2(n+1)} \tag{4.8}$$

for any value of t. This quantity approaches zero as N goes to infinity uniformly for t on any finite range. For the case that $0 < p < 1$ a similar argument may be used by including the factor $(1 - s)^{p-1}$ which appears in H_s in the integration, and placing the upper bound on the absolute value of the factor $G_n G_n'$.

Substituting back for $H(t, 0)$ in (4.6) one arrives at

LEMMA 2. *The characteristic function $G_n(t; z_m)$ satisfies the relationship*

$$G_N(t; z_M) = [G_n(t/\sqrt{2}; z_m)]^2 + o(1), \qquad N = 2n + 1, \qquad M = 2m, \tag{4.9}$$

where $o(1)$ goes to zero with increasing n, uniformly for t on any finite interval

$$0 \leq |t| \leq D. \tag{4.10}$$

The above lemma indicates that if the asymptotic *pdf* of z_m exists, it will be a "stable" distribution in the normal sense [2]. In order to set the stage for proving the existence of this asymptotic distribution we shall first investigate the third logarithmic derivative of $G_n(t; z_m)$.

5. Investigation of third logarithmic derivative. We shall now show that the third logarithmic derivative of G is uniformly bounded in some neighborhood of $t = 0$. We first prove that the absolute value of the third derivative of G is bounded for all t and n. Now the third derivative will have absolute value less than the third absolute moment which I denote by μ_3. Using Liapounoff's inequality

$$\mu_3^2 \leq \mu_2 \mu_4 \tag{5.1}$$

one asks whether the fourth moment μ_4 remains finite as n and m become infinite.

Computation of the fourth moment about the mean appears to be somewhat formidable. However it is not so difficult to show that it remains finite with increasing m and n. Referring to previous paper ([1] formulas (4.8)–(4.10)) we use quasi-moment generating function $g_0(x)$ such that

$$d^r g_0(0)/dx^r = \Gamma(pr + 1), \qquad g_0(0) = 1, \tag{5.2}$$

and it follows that

$$E(\sum_m u_i^p)^r = d^r[g_0(0)]^m/dx^r \Gamma(n + 1)/\Gamma(n + 1 + pr), \tag{5.3}$$

and one recalls that

$$y = \sum_m u_i^p, \qquad \bar{y} = \frac{m\Gamma(n + 1)\Gamma(p + 1)}{\Gamma(n + 1\ p)}$$

with

$$z = [(n + 1)^p/\sqrt{m}][y - \bar{y}].$$

The resulting fourth moment of z will be in the form of a fourth degree polynomial in m whose coefficients are of the type

$$\frac{(n + 1)^{4p}\Gamma(n + 1)}{\Gamma(n + 1 + 4p)}, \qquad \frac{(n + 1)^{3p}\Gamma(n + 1)}{\Gamma(n + 1 + 3p)}, \cdots,$$

combined with the first moment, with m^{-2} appearing as a factor. By expansion of the Gamma function in asymptotic series in $(n + 1)$ it is not difficult to show that the coefficient of m^4 becomes asymptotic like $(n + 1)^{-2}$, and that the coefficient of m^3 becomes asymptotic like $(n + 1)^{-1}$. It follows that as n and m go to infinity with $m \sim c(n + 1)$, that this fourth moment approaches a finite limit. Hence one concludes that the third derivative of G has bounded absolute value for all n and t.

Since the absolute value of the first derivative of G is uniformly bounded for finite t and all n it follows from the properties of a characteristic function that given a positive number K less than unity, it is possible to find a value of $t = t_0$ greater than zero such that

$$(5.4) \qquad 0 < K \leq | G_n(t, z) | \leq 1, \qquad 0 \leq | t | \leq t_0,$$

for all n.

From the above double inequality and the fact that the absolute values of the first three derivatives are uniformly bounded it follows that *the third logarithmic derivative of G is uniformly bounded for all n on the interval*

$$(5.5) \qquad 0 \leq | t | \leq t_0 .$$

6. Proof that the asymptotic distribution of z exists and is normal. Since absolute value of G is uniformly bounded away from zero on interval (5.5) one can write the functional relation (4.9) as

$$(6.1) \qquad \log G_N(t, z_M) = 2 \log G_n(t/\sqrt{2}, z_m) + o(1),$$

where $o(1)$ goes to zero with increasing n uniformly for t on interval (5.5).

Introduce the notation:

$\lambda(n)$ equals variance of z_m,

$q(t, n)$ equals third logarithmic derivative of $G_n(t, z_m)$,

$R(t, N)$ equals remainder defined by

$$(6.2) \qquad \log G_N(t, z_M) = -\lambda(N)t^2/2 + R(t, N).$$

Write

$$(6.3) \quad \log G_n(t/\sqrt{2}, z_m) = -\lambda(n)t^2/4 + q(t\theta/\sqrt{2}, n)t^3/(12\sqrt{2}), \quad 0 < \theta < 1.$$

Substituting (6.2) and (6.3) in (6.1)

$$(6.4) \quad R(t, N) = [\lambda(N) - \lambda(n)]t^2/2 + [1/\sqrt{2}]q(t\theta/\sqrt{2}, n)t^3/6 + o(1).$$

By (3.9)

$$(6.5) \qquad \lim \lambda(n) = \lim \lambda(N) = \text{positive number } \lambda.$$

We have proved that there exists an upper bound U such that

$$(6.6) \qquad | q(t, n) | \leq U$$

for all n and for t on interval

$$0 \leq |t| \leq t_0. \tag{6.7}$$

Hence from (6.4) one can reason that given a positive ϵ, a number N_0 can be found such that

$$|R(t, N)| \leq [1/\sqrt{2}]U|t^3/6| + \epsilon \tag{6.8}$$

for all t on (6.7) and for $N > N_0$.
By (6.1)

$$R(t, 2N + 1) = [\lambda(2N + 1) - \lambda(N)]t^2/2 + 2R(t/\sqrt{2}, N) + o(1). \tag{6.9}$$

Using (6.8)

$$|R(t/\sqrt{2}, N)| \leq [1/\sqrt{2}]U|t^3/(12\sqrt{2})| + \epsilon.$$

Hence for any positive number ϵ_2 a number N_2 can be found such that

$$|R(t, N)| \leq (1/2)U|t^3/6| + 2\epsilon + \epsilon_2, \qquad N > N_2,$$

for all t on (6.7). After k such operations, taking $\epsilon_i = \epsilon$

$$|R(t, N)| \leq (1/2)^{k/2}U|t^3/6| + (2^k - 1)\epsilon, \qquad N > N_k. \tag{6.10}$$

Thus given a positive number d one can determine k such that

$$(1/2)^{k/2}Ut_0^3/6 < d/2,$$

and ϵ such that

$$2^k\epsilon < d/2,$$

and therefore a number N_{k+1} such that

$$|R(t, N)| < d, \qquad N > N_{k+1} \tag{6.11}$$

for all t on interval (6.7).

It follows that $G_N(t, z_M)$ converges uniformly to exp. $(-\lambda t^2/2)$ on interval (6.7).

Convergence of $G_N(t, z_M)$ for a value $t = t_1$ outside the interval (6.7) may be proved by choosing integer k such that

$$0 < |t_1|/(\sqrt{2})^k \leq t_0, \tag{6.12}$$

and taking

$$t_3 = t_1/(\sqrt{2})^k.$$

Recalling that the functional relation (4.9) holds for all finite t, this can be applied k times, thus building up t_3 to t_1.

It follows from the continuity theorem that the distribution function of z_m converges to the normal distribution function.

7. Statement of theorem proved. The proof given above has involved the restriction that N be odd and M even (see (2.7)). This restriction is required for the integral relationship (3.6). However, if N were even one could take $n_0 = N/2$ and $n_1 = n_0 - 1$ and deal with, G_{n_0} and G_{n_1} in the integrand. Also if M were odd, one could take $m_0 = (M + 1)/2$, $m_1 = m_0 - 1$, and deal with $G_{n_0}(t, m_0)$ and $G_{n_1}(t, m_1)$ in the integrand. This would of course carry with it corresponding changes in the Gamma functions which precede the integral. As long as we require that

$$N = n_0 + n_1 + 1, \qquad M = m_0 + m_1,$$
$$\lim M/N = \lim m_0/n_0 = \lim m_1/n_1 = c > 0,$$

the arguments used in arriving at the asymptotic relations (3.15) and (4.9) will apply. Hence the theorem:

THEOREM[2]. *For a one dimensional statistical universe whose cdf is continuous, consider the function of the unit frequency differences* u_i

$$y = \sum_m u_i^p \tag{7.1}$$

taken from an ordered random sample of size n (see (2.2)) where p is any real positive number, and m is any positive integer less than or equal to n + 1. The selection of which m unit frequencies are to be included is arbitrary. Then with

$$\bar{y} = E(y) = \frac{m\Gamma(n + 1)\Gamma(p + 1)}{\Gamma(n + p + 1)} \tag{7.2}$$

consider the partially normalized variable

$$z = \frac{(n + 1)^p}{\sqrt{m}} (y - \bar{y}). \tag{7.3}$$

If n goes to infinity, with m becoming infinite so that

$$\lim m/n = c > 0, \tag{7.4}$$

then the asymptotic cumulative distribution of z exists and is normal, with

$$\lim E(z^2) = \Gamma(2p + 1) - \Gamma^2(p + 1) - cp^2\Gamma^2(p + 1), \tag{7.5}$$

except in the trivial case $p = 1$, $m = n + 1$, *in which case* $z \equiv 0$, *and in the case* $p = 1, c = 1$.

REFERENCES

[1] B. F. KIMBALL, "Some basic theorems for developing tests of fit for the case of the non-parametric probability distribution, I", *Annals of Math. Stat.*, Vol. 18 (1947), No. 4, pp. 540–548.

[2] P. LEVY, *Théorie de l'Addition des Variables Aléatoires*, Gauthier-Villars, Paris, 1937, Chapter V.

[3] H. CRAMÉR, *Mathematical Methods of Statistics*, Princeton University Press, Princeton, 1946.

[4] P. A. P. MORAN, "Random Division of an Interval", *Jour. Roy. Stat. Assoc., Suppl.*, Vol. 9 (1947), pp. 92–98.

[2] For the case $p = 2$, $m = n + 1$, an interesting proof was published by P. A. P. Moran in 1947, see [4].