DISTRIBUTIONS OF THE MEMBERS OF AN ORDERED SAMPLE

By Charles E. Clark and G. Trevor Williams

Booz, Allen and Hamilton, and The Johns Hopkins University

1. Introduction. Let the members of a random sample from a distribution $F(x)$ with probability density $F'(x) = f(x)$ be in order of magnitude $x_1, \cdots, x_m, \cdots, x_n, \cdots, x_N$, with $x_i \leqq x_{i+1}$, $i = 1, \cdots, N - 1$, and $m < n$. We shall compute the moments of the distribution of x_m and of the joint distribution of x_m and x_n.

The results are derived under the assumption that $F^{-1}(x)$, the inverse of $F(x)$, is a polynomial. Then we discuss the applicability of the results to any distribution for which $F^{-1}(x)$ is differentiable at $m/(N + 1)$ and $n/(N + 1)$. In this general case no restriction on $F(x)$ is imposed other than the differentiability; in particular, the interval on which $0 < F(x) < 1$ can be finite, semi-finite, or infinite.

2. Present status of the problem. This problem is handled through analyses of several specific distributions in reference [1] listed at the end of this paper. It is suggested that any one of the Pearson type frequency curves can be adequately approximated by one of the density functions handled in that paper. Although a general method is employed, there is no general development or general results; each distribution requires special, extensive computations. In contrast to these earlier results, the present paper contains a general development with results that are easily specialized to particular distributions.

Following [1] there have been discussions of asymptotic distributions. It is known that if m and N increase with m/N approaching a limit different from zero and one, under quite general conditions the distribution of x_m is asymptotically normal; see [2] or [3]. Also it was pointed out in [4] that with some restrictions on the distribution function the limiting distribution of x_m as N increases, but m is fixed, has the probability density

$$m^m \exp\,[my - \exp\,(-y)]/(m - 1)!$$

where y is a normalization of x_m; see [5]. However, it is suggested in [6] that in the case of the normal distribution if $m = 1$, one should have a sample of size 10^{12}, and Mr. Kendall concludes in [5], p. 221, that "For practical purposes, therefore, there is still no adequate general approximate form for the distribution of mth values." However, a contribution to the asymptotic case of this problem is made in [6]. In contrast to these asymptotic results, the present paper is concerned with the exact sampling distributions for any sample size. In the case of large samples, known approximations concerning moments are equivalent to the leading terms of some of the expansions of this paper.

Received August 20, 1957; revised December 26, 1957.

3. The moments of the distribution of x_m. The probability density function of x_m is

$$(1)\qquad [B(m, N - m + 1)]^{-1}[F(x)]^{m-1}[1 - F(x)]^{N-m}f(x)$$

where the coefficient is the reciprocal of the beta function.

We shall use the random variable $t = F(x_m)$ whose probability density function is $[B(m, N - m + 1)]^{-1}t^{m-1}(1 - t)^{N-m}$. We denote the central moments of this distribution by ν_i, $i = 0, 1, 2, \cdots$. Using p to denote the mean, we compute that $p = m/(N + 1)$.

At first we shall assume that the inverse of $F(x)$ is

$$(2)\qquad F^{-1}(x) = \sum_{i=0}^{r} a_i(x - p)^i.$$

Later we shall remove the restriction that $F^{-1}(x)$ is a polynomial.

The kth raw moment of the distribution of x_m immediately reduces to

$$\mu_k' = [B(m, N - m + 1)]^{-1} \int_0^1 [F^{-1}(t)]^k t^{m-1}(1 - t)^{N-m}\, dt, \qquad k = 0, 1, \cdots.$$

For each k we can write as a finite sum

$$[F^{-1}(t)]^k = \sum b_i(t - p)^i$$

where the coefficients b_i are functions of a_i and k. In this notation we have

$$(3)\qquad \mu_k' = \sum b_i\nu_i.$$

We calculate that

$$\nu_i = \sum_{j=0}^{i} \binom{i}{j} \frac{N!}{(m - 1)!(N - m)!} (-p)^j \int_0^1 t^{i-j+m-1}(1 - t)^{N-m}\, dt$$

$$= p^i \sum_{j=0}^{i-2} (-1)^j \binom{i}{j} \frac{(m + 1)\cdots(m + i - j - 1)}{p^{i-j-1}(N + 2)\cdots(N + i - j)} + (-1)^{i-1}ip^i + (-p)^i.$$

This expression will be reduced to a more convenient form. We use the identity

$$\frac{m + A}{p(N + A + 1)} = 1 + \frac{Aqp^{-1}}{N + A + 1}, \qquad q = 1 - p = \frac{N - m + 1}{N + 1}$$

and reduce ν_i to

$$\nu_i = p^i \sum_{j=0}^{i} (-1)^j \binom{i}{j} \prod_{A=1}^{i-j-1} \left(1 + \frac{Aqp^{-1}}{N + A + 1}\right), \qquad i = 0, 1, 2, \cdots.$$

In this formula

$$\prod_{A=1}^{-1} \left(1 + \frac{Aqp^{-1}}{N + A + 1}\right) = \prod_{A=1}^{0} \left(1 + \frac{Aqp^{-1}}{N + A + 1}\right) = 1.$$

From this result we get $\nu_0 = 1$, $\nu_1 = 0$, and

$$\nu_2 = \frac{pq}{N+2},$$

$$\nu_3 = \frac{2pq(q-p)}{(N+2)(N+3)},$$

$$\nu_4 = \frac{3p^2q^2N + 3pq(2-5pq)}{(N+2)(N+3)(N+4)},$$

$$\nu_5 = \frac{20p^2q^2(q-p)N + 4pq(q-p)(6+5pq)}{(N+2)(N+3)(N+4)(N+5)},$$

$$\nu = \frac{15p^3q^3N^2 + 10p^2q^2(13-40pq)N + 5pq(24-94pq+37p^2q^2)}{(N+2)(N+3)(N+4)(N+5)(N+6)}.$$

We shall use the notation $x_p = F^{-1}(p)$, and $f^{(i)} = f^{(i)}(x_p)$. We can express the a_i in (2) in terms of the derivatives of $F(x)$ at x_p by means of the relations between the derivatives of a function and its inverse. From the a_i we calculate the b_i, and with the use of (3) we get the raw moments. These include

$$\mu_1' = x_p - \frac{f'}{2f^3}\cdot\frac{pq}{N+2} + \frac{3f'^2 - ff''}{6f^5}\cdot\frac{2pq(q-p)}{(N+2)(N+3)} + \frac{10ff'f'' - f^2f''' - 15f'^3}{24f^7}\cdot\frac{3p^2q^2N + 3pq(2-5pq)}{(N+2)(N+3)(N+4)} + \cdots.$$

Here as elsewhere derivatives are denoted by primes and powers by arabic numerical exponents. Finally the central moments μ_k are obtained, such as the following.

$$\mu_2 = \frac{1}{f^2}\cdot\frac{pq}{N+2} - \frac{f'}{f^4}\cdot\frac{2pq(q-p)}{(N+2)(N+3)} + \left[\frac{5f'^2}{4f^6} - \frac{f''}{3f^5}\right]\frac{3p^2q^2N + 3pq(2-5pq)}{(N+2)(N+3)(N+4)} - \frac{f'^2}{4f^6}\cdot\frac{p^2q^2}{(N+2)^2} + \cdots,$$

$$\mu_3 = \frac{1}{f^3}\cdot\frac{2pq(q-p)}{(N+2)(N+3)} - \frac{3f'}{2f^5}\cdot\frac{3p^2q^2N + 3pq(2-5pq)}{(N+2)(N+3)(N+4)} + \frac{3f'}{2f^5}\cdot\frac{p^2q^2}{(N+2)^2} + \cdots,$$

$$\mu_4 = \frac{1}{f^4}\cdot\frac{3p^2q^2N + 3pq(2-5pq)}{(N+2)(N+3)(N+4)} + \cdots.$$

From these results we check the well known fact that if N increases with m/N fixed, the asymptotic distribution of x_m has the mean and variance x_p and pq/f^2N respectively (see [3]). Furthermore the known result that for large N the distribution is approximately normal is suggested by the following which are obtained from the leading terms of the above expressions.

$$\frac{\mu_3}{\mu_2^{3/2}} = N^{-1/2}\left[\frac{2(q-p)}{\sqrt{pq}} - \frac{3f'\sqrt{pq}}{f^2}\right] + \cdots,$$

$$\frac{\mu_4}{\mu_2^2} = 3\left[1 - \frac{5N+12}{(N+3)(N+4)}\right] + \cdots.$$

We next discuss the applicability of the results to distributions for which $F^{-1}(x)$ is not a polynomial. We note that the factor

$$[F(x)]^{m-1}[1 - F(x)]^{N-m}$$

in (1) assumes its maximum value at $(m - 1)/(N - 1)$. Hence (1) indicates that the probability density of x_m is practically zero except in a small neighborhood of $F^{-1}[(m - 1)/(N - 1)]$.[1] Hence the moments of the distribution of x_m can be determined with great accuracy from a knowledge of $F(x)$ in a small neighborhood of $F^{-1}[(m - 1)/(N - 1)]$. But this knowledge of $F(x)$ is given by a few derivatives of $F(x)$ at x_p because x_p is near

$$F^{-1}[(m - 1)/(N - 1)].$$

In other words, the first few terms of the Taylor expansion of $F^{-1}(x)$ at x_p should be enough to permit an accurate determination of the moments. Hence the above derivation holds with very little error if (2) is understood to be a few terms of the Taylor expansion.

4. The median. The results simplify in the case $N = 2m + 1$. We can compute that

$$\int_0^1 (t - 1/2)^j t^m (1 - t)^m \, dt,$$

which is clearly zero when j is odd, reduces when j is even to

$$\frac{m!}{2^{m+j}(j + 1)(j + 3)\cdots(j + 2m + 1)};$$

the reduction is achieved by the substitution of $t = \sin^2 \theta$ and use of a known integral (see [7]). This reduces, after multiplication by $B[(m + 1, m + 1)]^{-1}$, to

$$\nu_{2i} = \frac{1\cdot3\cdot5\cdots(2i - 1)}{4^i(2m + 3)(2m + 5)\cdots(2m + 2i + 1)}, \qquad i = 1, 2, \cdots.$$

5. The efficiency of the median. As a numerical illustration we shall compute the efficiency of the median as an estimator of the mean of a normal distribution. We consider $\varphi(x) = (2\pi)^{-1/2}e^{-x^2/2}$ and $\varphi = \varphi(0)$. The derivatives of $F^{-1}(x)$ at $x = 0$ are calculated from those of $\varphi(x)$. Using (3) with $k = 1, 2$ and the formulas of section 4, we obtain the variance of the median of a sample of size

[1] This statement is true even when $m - 1$ or $N - m$ is small. If, for example, $m - 1$ is small, $F(x) < (m - 1)/(N - 1)$ for $x < F^{-1}[(m - 1)/(N - 1)]$, and $[1 - F(x)]^{N-m}$ is clearly small if x is at least a little greater than $F^{-1}[(m - 1)/(N - 1)]$.

$N = 2n + 1$ in the form

$$\mu_2 = \frac{1}{4\varphi^2(2n+3)}\left\{1 + \frac{1}{4\varphi^2(2n+5)} + \frac{13}{96\varphi^4(2n+5)(2n+7)} + \frac{287}{2688\varphi^6(2n+5)(2n+7)(2n+9)} + \cdots\right\}.$$

Since the sample mean is efficient, and since the variance of the sample mean is $1/(2n + 1)$, if $E(2n + 1)$ is the efficiency of the median,

$$E(2n + 1) = [(2n + 1)\mu_2]^{-1}.$$

Evaluating φ we obtain

$$\frac{1}{E(2n+1)} = \frac{1.5707963(2n+1)}{2n+3} \cdot \left\{1 + \frac{1.5707963}{2n+5} + \frac{5.3460357}{(2n+5)(2n+7)} + \frac{26.484528}{(2n+5)(2n+7)(2n+9)} + \cdots\right\}.$$

A tabulation of this four term approximation appears in Table I.

The series for the reciprocal of the efficiency converges slowly for small

$$2n + 1.$$

In cases $n = 1, 2, 3$, the fourth term contributes 2.8%, 1.6%, 1.0%, respectively, of the tabulated value. To check the accuracy of the approximation we have calculated accurately (as described below) the reciprocal of the efficiency in cases $n = 1, 2, 3$. The values correct to three decimal places are given in the table. The relative errors are 5.6%, 2.2%, 1.1%, respectively.

TABLE I

Efficiency of the Median, Normal Distribution

$N = 2n + 1$	$[E(2n + 1)]^{-1}$, four term approximation	$[E(2n + 1)]^{-1}$, exact	$E(2n + 1)$
∞	1.571	1.571	.637
201	1.567		.638
101	1.564		.639
51	1.557		.642
31	1.549		.646
21	1.538		.650
11	1.503		.665*
9	1.486		.673*
7	1.457	1.473	.679
5	1.402	1.434	.697
3	1.270	1.346	.743

The third decimal places in E(11) and E(9) are in doubt.

The correct values of the reciprocal of the efficiency are obtained as follows. If $n = 1$, the reciprocal of the efficiency is, except for the factor

$$(2n + 1)/B(2, 2) = 18,$$

with $F'(x) = \varphi(x)$,

$$\begin{aligned}\int_{-\infty}^{\infty} x^2F(1 - F)\varphi\,dx &= \int_{-\infty}^{\infty} F\,d(-x\varphi + F) - \int_{-\infty}^{\infty} F^2\,d(-x\varphi + F)\\ &= 1 - \int_{-\infty}^{\infty} (-x\varphi + F)\varphi\,dx - 1 + \int_{-\infty}^{\infty} (-x\varphi + F)2F\varphi\,dx\\ &= -\left[\frac{\varphi^2}{2} + \frac{F^2}{2}\right]_{-\infty}^{\infty} + \left[\frac{2F^3}{3}\right]_{-\infty}^{\infty} + 2\int_{-\infty}^{\infty} F\,d\left(\frac{\varphi^2}{2}\right)\\ &= -1/2 + 2/3 - 2\int_{-\infty}^{\infty} \frac{\varphi^2}{2}\varphi\,dx\\ &= 1/6 - (2\pi)^{-\frac{3}{2}}\int_{-\infty}^{\infty} e^{-3x^2/2}\,dx\\ &= 1/6 - \frac{1}{2\pi\sqrt{3}}.\end{aligned}$$

Multiplying this last number by $3/B(2, 2)$ we get

$$\begin{aligned}\frac{1}{E(3)} &= 3 - \frac{3\sqrt{3}}{\pi}\\ &= 1.346\end{aligned}$$

as given above.

For $n = 2, 3$ the reciprocals of the efficiencies were calculated by numerical evaluation of

$$\frac{(2n + 1)}{B(n + 1, n + 1)}\int_{-\infty}^{\infty} x^2F^n(1 - F)^n\,dx.$$

6. The moments of the joint distribution of x_m and x_n, $m < n$. We consider next the joint distribution of x_m and x_n, $m < n$. The probability density is

$$\frac{N!}{(m - 1)!(n - m - 1)!(N - n)!} \cdot [F(x_m)]^{m-1}[F(x_n) - F(x_m)]^{n-m-1}[1 - F(x_n)]^{N-n}f(x_m)f(x_n).$$

The probability density of $t = F(x_m)$ and $u = F(x_n)$ is

$$\frac{N!}{(m - 1)!(n - m - 1)!(N - n)!}\,t^{m-1}(u - t)^{n-m-1}(1 - u)^{N-n}.$$

The expected values of t and u are $p_m = m/(N + 1)$ and $p_n = n/(N + 1)$ respectively. If $\nu_{\alpha\beta}$ is the expected value of $(t - p_m)^\alpha(u - p_n)^\beta$, we calculate that

$$\nu_{20} = \frac{p_m q_m}{N + 2}\cdot$$

$$\nu_{11} = \frac{p_m q_n}{N + 2},$$

$$\nu_{02} = \frac{p_n q_n}{N + 2},$$

$$\nu_{30} = \frac{2p_m q_m(q_m - p_m)}{(N + 2)(N + 3)},$$

$$\nu_{21} = \frac{2p_m q_n(q_m - p_m)}{(N + 2)(N + 3)},$$

$$\nu_{12} = \frac{2p_m q_n(q_n - p_n)}{(N + 2)(N + 3)},$$

$$\nu_{03} = \frac{2p_n q_n(q_n - p_n)}{(N + 2)(N + 3)},$$

$$\nu_{40} = \frac{3p_m^2 q_m^2 N + 3p_m q_m(2 - 5p_m q_m)}{(N + 2)(N + 3)(N + 4)},$$

$$\nu_{31} = \frac{3p_m^2 q_m q_n N + 3p_m q_n(2 - 5p_m q_m)}{(N + 2)(N + 3)(N + 4)},$$

$$\nu_{22} = \frac{p_m q_n[1 - (p_m + q_n) + 3p_m q_n]N + p_m q_n[1 + 5(p_m + q_n) - 15p_m q_n]}{(N + 2)(N + 3)(N + 4)},$$

$$\nu_{13} = \frac{3p_m p_n q_n^2 N + 3p_m q_n(2 - 5p_n q_n)}{(N + 2)(N + 3)(N + 4)},$$

$$\nu_{04} = \frac{3p_n^2 q_n^2 N + 3p_n q_n(2 - 5p_n q_n)}{(N + 2)(N + 3)(N + 4)}.$$

If $\mu'_{\alpha\beta}$ is the expected value of $x_m^\alpha x_n^\beta$,

$$\mu'_{\alpha\beta} = \frac{N!}{(m - 1)!(n - m - 1)!(N - n)!}\int_0^1 du$$
$$\cdot \int_0^u [F^{-1}(t)]^\alpha[F^{-1}(u)]^\beta\, t^{m-1}(u - t)^{n-m-1}(1 - u)^{N-n}\, dt.$$

Let the Taylor expansion

$$[F^{-1}(t)]^\alpha[F^{-1}(u)]^\beta = a_{00} + a_{10}(t - p_m) + a_{01}(u - p_n) + a_{20}(t - p_m)^2$$
$$+ a_{11}(t - p_m)(u - p_n) + a_{02}(u - p_n)^2 + \cdots$$

be finite. Then

$$\mu'_{\alpha\beta} = a_{00} + a_{20}\nu_{20} + a_{11}\nu_{11} + a_{02}\nu_{02} + a_{30}\nu_{30} + \cdots .$$

The coefficients a_{ij} are expressed in terms of the derivatives of $F(x)$ at $F^{-1}(p_m)$ and $F^{-1}(p_n)$.

As in the 1-dimensional case, if the Taylor expansion does not terminate, these results are approximations.

As an illustration of the results obtained in this manner, the covariance of x_m and x_n reduces to

$$\begin{aligned} V(x_m, x_n) = {} & \frac{1}{f_m f_n} \cdot \frac{p_m q_n}{N+2} - \frac{f'_m}{2f_m^3 f_n} \cdot \frac{2p_m q_n (q_m - p_m)}{(N+2)(N+3)} \\ & - \frac{f'_n}{2f_n^3 f_m} \cdot \frac{2p_m q_n (q_n - p_n)}{(N+2)(N+3)} \\ & + \frac{3f_m'^2 - f_m f_m''}{6f_m^5 f_n} \cdot \frac{3p_m^2 q_m q_n N + 3p_m q_n (2 - 5p_m q_m)}{(N+2)(N+3)(N+4)} \\ & + \frac{3f_n'^2 - f_n f_n''}{6f_n^5 f_m} \cdot \frac{3p_m p_n q_n^2 N + 3p_m q_n (2 - 5p_n q_n)}{(N+2)(N+3)(N+4)} \\ & + \frac{f'_m f'_n}{4f_m^3 f_n^3} \cdot \frac{p_m q_n[1 - (p_m + q_n) + 3p_m q_n]N + p_m q_n[1 + 5(p_m + q_n) - 15p_m q_n]}{(N+2)(N+3)(N+4)} \\ & - A_m A_n + \cdots \end{aligned}$$

where

$$f_m^{(i)} = f^{(i)}[F^{-1}(p_m)], \quad i = 0, 1, \cdots,$$

$$\begin{aligned} A = {} & -\frac{f'}{2f^3} \cdot \frac{pq}{N+2} + \frac{3f'^2 - f''}{6f^5} \cdot \frac{2pq(q-p)}{(N+2)(N+3)} + \frac{10ff'f'' - f^2 f''' - 15f'^3}{24f^7} \\ & \cdot \frac{3p^2 q^2 N + 3pq(2 - 5pq)}{(N+2)(N+3)(N+4)}, \end{aligned}$$

A_m is obtained from A by affixing the subscript m to every f, p, and q, and A_n is obtained similarly.

Using μ_2 as calculated above, we obtain from the last result the first two terms of the coefficient of linear correlation in the form

$$r(x_m, x_n) = \left(\frac{p_m q_n}{q_m p_n}\right)^{1/2} \left\{1 - \frac{A}{N+2}\right\}$$

in which

$$A = \frac{f_m'^2}{4f_m^4} p_m q_m - \frac{f'_m f'_n}{2f_m^2 f_n^2} p_m q_n + \frac{f_n'^2}{4f_n^4} p_n q_n .$$

The following special cases are easily obtained. If $f(x) = \exp(-x)$,

$$A = \tfrac{1}{4}[p_m q_m \exp(2x_m) - 2p_m q_n \exp(x_m + x_n) + p_n q_n \exp(2x_n)].$$

If $f(x) = (2\pi)^{-1/2} \exp(-x^2/2)$,

$$A = \frac{x_m^2}{4f_m^2} p_m q_m - \frac{x_m x_n}{2f_m f_n} p_m q_n + \frac{x_n^2}{4f_n^2} p_n q_n .$$

If $f(x) = \exp(-x)x^{r-1}/\Gamma(r)$,

$$\begin{aligned} A = \tfrac{1}{4}[\Gamma(r)]^2[(r - 1 - x_m)^2 x_m^{-2r} \exp(2x_m) p_m q_m \\ - 2(r - 1 - x_m)(r - 1 - x_n)(x_m x_n)^{-r} \exp(x_m + x_n) p_m q_n \\ + (r - 1 - x_n)^2 x_n^{-2r} \exp(2x_n) p_n q_n .\end{aligned}$$

REFERENCES

[1] K. Pearson, and M. V. Pearson, "On the mean character and variance of a ranked individual and on the mean and variance of the intervals between ranked individuals," Part 1, *Biometrika*, 23 (1931), pp. 364–397, and part 2, *Biometrika*, 24 (1932), pp. 203–279.

[2] N. Smirnoff, "Ueber die Verteilung des allgemeinen Gliedes in der Variationsreihe," *Metron*, 12 (1931), pp. 127–138.

[3] H. Cramér, "Mathematical methods of statistics," Princeton, 1946, pp. 367–378.

[4] E. J. Gumbel, "Les valeurs extremes des distributions statistiques," *Annales de l'Institut Henri Poincare*, 5, (1934), p. 115.

[5] M. G. Kendall, "The advanced theory of statistics," Volume 1, London, 1943, pp. 218–224.

[6] R. A. Fisher, and L. H. C. Tippett, "Limiting forms of the frequency distribution of the largest or smallest number of a sample," *Proc. Camb. Phil. Soc.*, 24 (1928), p. 180.

[7] E. T. Whittaker and G. N. Watson, "Modern analysis," Cambridge University Press, 1945, p. 256.