DISTRIBUTION OF THE MAXIMUM OF THE ARITHMETIC MEAN OF CORRELATED RANDOM VARIABLES

By John Gurland

Iowa State College

1. Summary. The initial distribution considered here is obtained from a multivariate analogue of the Pearson Type III distribution, and the value of the correlation is taken to be non-negative. There is obtained here the distribution of the maximum in samples of fixed size n from a random variable which is the arithmetic mean of k such correlated random variables. This distribution is obtained for large values of n and for large values of k. The appropriate expressions for the mode and scale parameters are also given.

2. Introduction. The mathematical model presented here is applicable when the aforementioned samples of size n consist of independent observations from a fixed population. The distribution function for this population, given by (4) below, is obtained by regarding each of the k correlated random variables as having the same fixed Pearson Type III distribution, and as having the same correlation with each of the remaining variables. Such a model may be relevant, for instance, in considering the tensile strength of a substance; in this case n is roughly proportional to the number of flaws. Another possible field of application is the investigation of maximum coincident loads in electrical engineering problems; but in this case there are difficulties involved in determining the appropriate value of n, and in assuming that the n observations of the sample are independent. The above mathematical model could be extended to cover such situations; however, the present article is confined to the theoretical problem stated in the summary.

3. Distribution of the arithmetic mean of correlated Pearson Type III variables. In many problems the primary distribution is skew and is well fitted by a Pearson Type III distribution

$$p_{Z_1}(x) = \frac{e^{-x/\theta}x^{\lambda-1}}{\theta\Gamma(\lambda)}, \qquad x > 0. \tag{1}$$

Before considering the problem concerning the maximum value of the mean, it is first required to ascertain the distribution of $(1/k)\sum_1^k Z_i$, where each Z_i has the probability density (1) and there is a constant correlation ρ between every pair of Z's. The assumption of constant correlation is indeed restrictive, but it is most convenient to have a single parameter which measures the over-all correlation of the population. It is possible to consider more general models by the same methods described below, but the present article considers only the case of constant correlation.

Received September 21, 1953, revised August 17, 1954.

The characteristic function of a multivariate analogue of (1) is obtained by a device, similar to that used by Krishnamoorthy and Parthasarathy [1], as follows. Consider the multivariate normal probability density

$$p_X(x) = \frac{|\Omega|^{1/2}}{(\sigma\sqrt{2\pi})^k} \exp\left(-\frac{x\Omega x'}{2\sigma^2}\right),$$

where $X = (X_1, X_2, \cdots, X_k)$ and Ω is a positive definite matrix. The characteristic function of $Y = (X_1^2, X_2^2, \cdots, X_k^2)$ is

$$\phi_Y(t) = \frac{|\Omega|^{1/2}}{|\Omega - 2i\sigma^2 T|^{1/2}},$$

where T is the diagonal matrix

$$T = \begin{bmatrix} t_1 & 0 & 0 & \cdots & 0 \\ 0 & t_2 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & t_k \end{bmatrix}.$$

Then $[\phi_Y(t)]^m$ is the characteristic function of a natural multivariate analogue of the χ^2 distribution. Now

$$[\phi_Y(t)]^m = |I - 2i\sigma^2 T\Omega^{-1}|^{-m/2}$$

where I is the unit $n \times n$ matrix. Take $\frac{1}{2}m = \lambda$ and $2\sigma^2 = \theta$ to obtain

$$\phi_Z(t) = |I - i\theta T\Omega^{-1}|^{-\lambda}. \tag{2}$$

This is the characteristic function of $Z = (Z_1, Z_2, \cdots, Z_k)$, where each Z_i has the probability density (1). Note that in (2) any positive definite covariance matrix Ω is permissible. The special case of interest here is

$$\Omega^{-1} = \begin{bmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho & \rho & \rho & \cdots & 1 \end{bmatrix}.$$

The characteristic function of $(1/k)\sum_1^k Z_i$ is then

$$f(t) = \begin{vmatrix} 1 - it\theta/k & -it\theta\rho/k & \cdots & -it\theta\rho/k \\ -it\theta\rho/k & 1 - it\theta/k & \cdots & -it\theta\rho/k \\ \cdots & \cdots & \cdots & \cdots \\ -it\theta\rho/k & -it\theta\rho/k & \cdots & 1 - it\theta/k \end{vmatrix}^{-\lambda} \tag{3}$$

$$= [1 - (it\theta/k)(1 - \rho)]^{-\lambda(k-1)}[1 - (it\theta/k)(1 - \rho + \rho k)]^{-\lambda}.$$

On referring to tables of Fourier integrals [2], one finds that the probability density is

$$p(x) = \frac{1}{c} x^{k\lambda-1} \exp\left[\frac{-kx}{\theta(1-\rho)}\right] {}_1F_1\left(\lambda, k\lambda, \frac{\rho k^2 x}{\theta(1-\rho)(1-\rho+\rho k)}\right) \tag{4}$$

where

$$c = (\theta/k)^{\lambda k}(1-\rho)^{\lambda(k-1)}(1-\rho+\rho k)^{\lambda}\,\Gamma(\lambda k) \tag{5}$$

and where ${}_1F_1(w, v, z)$ is the confluent hypergeometric function which is given by

$${}_1F_1(w, v, z) = 1 + \frac{w}{v}\frac{z}{1} + \frac{w(w+1)}{v(v+1)}\frac{z^2}{2!} + \frac{w(w+1)(w+2)}{v(v+1)(v+2)}\frac{z^3}{3!} + \cdots.$$

For $\rho = 0$ the density given by (4) reduces to that of an arithmetic mean of k independent random variables, each of which has the density given by (1).

4. Asymptotic solution of a certain transcendental equation. Before obtaining the large-sample distribution of the maximum value in samples from (4), it is necessary to consider the solution of the equation

$$x - A\log x + B = C \tag{6}$$

where A, B, and C are constants and the value of C is large. If $A < 0$ there is one solution but if $A > 0$ there are two solutions, as is easily seen by considering the intersection of $y = x + B$ and $y = A\log x + C$, and noting that $C \to \infty$. However the only solution required here is the one which increases indefinitely as C increases indefinitely.

Let $r(x) = x - A\log x + B - C$. Now $dr/dx = 1 - A/x > 0$ for x large, which shows that the function is monotonically increasing for x sufficiently large. As a first approximation to a solution, try

$$x_1 = C - B + A\log C. \tag{7}$$

This makes

$$r(x_1) = -A\log\left\{1 - \frac{B - A\log C}{C}\right\} \approx \frac{A\,(B - A\log C)}{C} \tag{8}$$

for C large. This approaches zero as $C \to \infty$. It can, in fact, be shown that the solution satisfies

$$x = C - B + A\log C + o(1/C^{1-\epsilon}) \tag{9}$$

where ϵ is an arbitrarily small number such that $0 < \epsilon < 1$. This is easily seen[1] by noting that for C large enough,

$$r(C - B + A\log C - 1/C^{1-\epsilon/2}) < 0,$$

$$r(C - B + A\log C + 1/C^{1-\epsilon/2}) > 0.$$

[1] The author is grateful to the referee for simplifying this proof.

If B as well as C is permitted to increase indefinitely, the asymptotic solution (7) remains valid, provided $B/C \to 0$. This requirement is obvious on referring to (8). However, the order of the approximation indicated in (9) may not be valid for this case. A sufficient condition for the validity of (9) is $B = O(\log C)$.

5. Large sample distribution of the maximum for the case $\rho > 0$. The main problem now is to find the distribution function of the maximum in samples of size n from the population characterized by the density (4). In this section the distribution will be obtained first for large values of n, then for large values of k and n. The two cases $\rho > 0$ and $\rho = 0$ will also be treated separately.

Let $F(x) = \int_0^x p(t)\,dt$ where $p(t)$ is the density (4). Then $[F(x)]^n$ is the distribution function of the maximum in samples of size n. The probability density of the maximum is

$$g(x) = nF^{n-1}(x)\,p(x). \tag{10}$$

Apply the transformation

$$y_n = n[1 - F(x)]. \tag{11}$$

Then, if X is the random variable with density (10), and Y_n is the random variable defined by (11), Y_n has the probability density function

$$p_{Y_n}(x) = (1 - x/n)^{n-1},$$

and it will be noted that $\lim_{n\to\infty} p_{Y_n}(x) = e^{-x}$, for $x > 0$. In solving (11) for x as a function of y_n, it will be shown below that $X = -\alpha \log Y_n + \beta$ for large values of n (and hence large values of x). Consequently, by a limit theorem of Mann and Wald [3], the limiting distribution function of $(X - \beta)/\alpha$ is the distribution function $\exp(-e^{-x})$.

We now proceed to solve (11) for x. The equation may be written

$$y_n = \frac{n}{c}\int_x^\infty t^{\lambda k-1} \exp\left[\frac{-kt}{\theta(1-\rho)}\right] {}_1F_1(\lambda, k\lambda, \delta t)\,dt \tag{12}$$

where $\delta = \rho k^2/\theta(1-\rho)(1-\rho+\rho k)$, and c is given by (5).

Since the distribution will be considered for large values of n it will suffice to find the solution of (12) which is valid for large values of x (cf. Fisher and Tippett [4], Cramér [5]). In evaluating the integral in (12), the following properties of the conflue t hypergeometric function will be required.

$$\frac{d}{dx}\,{}_1F_1(w, v, x) = \frac{w}{v}\,{}_1F_1(w+1,\ v+1,\ x); \tag{13}$$

$${}_1F_1(w, v, x) = \frac{e^x\Gamma(v)}{x^{v-w}\Gamma(w)}\left[1 + O\left(\frac{1}{x}\right)\right], \qquad x \to \infty. \tag{14}$$

Formula (13) follows from the definition of the confluent hypergeometric function (cf. [6]). A more general asymptotic expression than (14) is established by Barnes [7].

Integrate (12) by parts and apply (13) to obtain $cy_n/n = \alpha_1 + \alpha_2 + \alpha_3$, where

$$\alpha_1 = \frac{\theta(1-\rho)}{k} x^{k\lambda-1} \exp\left[\frac{-kx}{\theta(1-\rho)}\right] {}_1F_1(\lambda, k\lambda, \delta x),$$

$$\alpha_2 = \frac{\theta(1-\rho)(k\lambda-1)}{k} \int_x^\infty t^{\lambda k-2} \exp\left[\frac{-kt}{\theta(1-\rho)}\right] {}_1F_1(\lambda, k\lambda, \delta t)\, dt,$$

$$\alpha_3 = \frac{\delta\theta(1-\rho)}{k^2} \int_x^\infty t^{\lambda k-1} \exp\left[\frac{-kt}{\theta(1-\rho)}\right] {}_1F_1(\lambda+1, k\lambda+1, \delta t)\, dt.$$

Now define $\alpha = \theta(1-\rho+\rho k)/k$. In virtue of (14), and for $\rho > 0$, these expressions reduce to

$$\alpha_1 = \frac{\theta(1-\rho)}{k} \frac{x^{\lambda-1}}{\delta^{\lambda(k-1)}} \frac{\Gamma(k\lambda)}{\Gamma(\lambda)} e^{-x/\alpha}\left[1 + O\left(\frac{1}{\delta x}\right)\right],$$

$$\alpha_2 = \frac{\theta(1-\rho)\alpha}{k} \frac{k\lambda-1}{\delta^{\lambda(k-1)}} x^{\lambda-2} \frac{\Gamma(k\lambda)}{\Gamma(\lambda)} e^{-x/\alpha}\left[1 + O\left(\frac{1}{\delta x}\right)\right],$$

$$\alpha_3 = \frac{\theta(1-\rho)\alpha}{k} \frac{x^{\lambda-1}}{\delta^{\lambda(k-1)-1}} \frac{\Gamma(k\lambda)}{\Gamma(\lambda)} e^{-x/\alpha}\left[1 + O\left(\frac{1}{\delta x}\right)\right].$$

As a result of the above simplification it is now possible to write (12) in the form

$$\frac{cy_n}{n} = \frac{\theta(1-\rho)}{k} \frac{(1+\delta\alpha)x^{\lambda-1}}{\delta^{\lambda(k-1)}} \frac{\Gamma(k\lambda)}{\Gamma(\lambda)} e^{-x/\alpha}\left[1 + O\left(\frac{1}{\delta x}\right)\right].$$

A further simplification is afforded by referring to the representation of c in (5). Thus

$$\frac{c_1 y_n}{n} = x^{\lambda-1} e^{-x/\alpha}\left[1 + O\left(\frac{1}{\delta x}\right)\right], \tag{15}$$

$$c_1 = \frac{\alpha^{\lambda k-1}}{1+\alpha} (1-\rho)^{\lambda(k-1)-1}(1-\rho+\rho k)^{\lambda(1-k)-1}\Gamma(\lambda).$$

It now remains to solve (15) as a function of y_n. If the terms involving $O(1/\delta x)$ are neglected, and we take logarithms of the remaining terms, we obtain

$$x - \alpha(\lambda-1)\log x + \alpha(\log y_n + \log c_1) = \alpha \log n. \tag{16}$$

Define

$$A = \alpha(\lambda-1), \qquad B = \alpha(\log y_n + \log c_1), \qquad C = \alpha \log n. \tag{17}$$

If k is small and n is large, then the solution (9) is applicable. Hence

$$x = -\alpha \log y_n + \beta + o\,(1/[\alpha \log n]^{1-\epsilon}), \tag{18}$$

$$\beta = \alpha\,[\log (n/c_1) + (\lambda-1)\log(\alpha \log n)].$$

Thus, the limiting distribution of $(X-\beta)/\alpha$ is $\exp(-e^{-x})$, and α and β represent respectively the scale parameter and the mode for large values of n.

If k as well as n is allowed to increase indefinitely, then in order to apply the results of Section 4, the expressions for B and C in (17) must be altered. In fact, equation (16) may be written as

$$x - \alpha(\lambda - 1)\log x + \alpha \log y_n + \alpha \log \{\theta^{\lambda k-1}(1-\rho)^{\lambda(k-1)-1}(1-\rho+\rho k)^{\lambda}\delta^{\lambda(k-1)}\Gamma(\lambda)\} = \alpha \log [k^{\lambda k-2}\{k + \delta\theta(1-\rho+\rho k)\}] + \alpha \log n.$$

Now if we define

$$B = \alpha \log y_n + \alpha \log \{\theta^{\lambda k-1}(1-\rho)^{\lambda(k-1)-1}(1-\rho+\rho k)^{\lambda}\delta^{\lambda(k-1)}\Gamma(\lambda)\},$$

$$C = \alpha \log n + \alpha \log [k^{\lambda k-2}\{k + \delta\theta(1-\rho+\rho k)\}],$$

it follows, since δ is of the order of magnitude of k, that

$$B \approx \alpha\lambda k \log k, \qquad C \approx \alpha\lambda k \log k + \alpha \log n.$$

Hence, if $k \log k = o(\log n)$, it follows that $\lim_{k\to\infty, n\to\infty} B/C = 0$, in which case (9) is applicable. This gives $x = -\alpha \log y_n + \beta$, with $\alpha = \theta(1-\rho+\rho k)/k$ as in (16) and

$$\beta = \alpha \log (n/c_1) + \alpha(\lambda - 1)\log \{\alpha \log [n(k\lambda)^{k\lambda-2}(k + \delta\theta(1-\rho+\rho k)]\}. \tag{19}$$

6. Large sample distribution of the maximum for the case $\rho = 0$. The distribution of the maximum for the case $\rho = 0$ must be considered separately because δ now reduces to zero, and the approximations employed in the evaluation of α_1, α_2, and α_3 are no longer valid for large values of x. With $\rho = 0$ the density in (4) becomes

$$p(x) = \frac{1}{c} x^{k\lambda-1} e^{-kx/\theta}, \qquad x > 0$$

where now $c = (\theta/k)^{\lambda k}\Gamma(\lambda k)$. The same method as that described above is applied to

$$y_n = \frac{n}{c}\int_x^{\infty} t^{\lambda k-1} e^{-kt/\theta}\, dt.$$

For large values of x this yields

$$cy_n/n = (\theta/k)e^{-kx/\theta}x^{k\lambda-1}[1 + O(1/x)].$$

As before, neglect the terms involving $O(1/x)$ and take logarithms of the remaining terms to obtain

$$x - \alpha(k\lambda - 1)\log x + \alpha[\log y_n + \log \{\alpha^{k\lambda-1}\Gamma(\lambda k)\}] = \alpha \log n \tag{20}$$

where now $\alpha = \theta/k$. If k is small and n is large the solution becomes again $x = -\alpha \log y_n + \beta$, where now

$$\beta = \alpha \log \frac{n}{\alpha^{\lambda k-1}\Gamma(\lambda k)} + \alpha(k\lambda - 1)\log (\alpha \log n). \tag{21}$$

If k as well as n is allowed to increase indefinitely, then, as before, the expressions for B and C must be altered to apply the results of Section 4. Rewrite equation (20) as

$$x - \alpha(k\lambda - 1) \log x + \alpha \log y_n = \alpha \log n - \alpha \log [\alpha^{\lambda k-1}\Gamma(\lambda k)].$$

If we define

$$B = \alpha \log y_n, \qquad C = \log \frac{n}{\alpha^{\lambda k-1}\Gamma(\lambda k)},$$

then $\lim_{k\to\infty, n\to\infty} B/C = 0$ if $\alpha^{\lambda k-1}\Gamma(\lambda k) = o(n)$. We obtain, as before,

$$x = -\alpha \log y_n + \beta,$$

where now

$$\beta = \alpha \log \frac{n}{\alpha^{\lambda k-1}\Gamma(k\lambda)} + \alpha(k\lambda - 1) \log \left\{ \alpha \log \frac{n}{\alpha^{\lambda k-1}\Gamma(k\lambda)} \right\}. \tag{22}$$

6. Conclusion. A few remarks are in order regarding the results obtained in this paper. Firstly, the cases in which both k and n are allowed to increase indefinitely require that k should not increase too rapidly relative to n, if the results obtained are to remain valid. Further, the order of approximation for these cases need not be the same as for the cases in which only n is allowed to increase indefinitely.

Secondly, some extensions of the results obtained here require further research. For instance, as has already been stated, the correlation need not be the same for all pairs of Z_i's. Further, the initial distribution from which samples of size n are taken need not be fixed but might change during the course of the sampling. This might further be complicated by the fact that the observations in the sample could be correlated.

7. Acknowledgments. The author has had many stimulating discussions with Dr. Emil Jebe and Mr. Landy Altman, who first brought this general problem to his attention. He has also had the benefit of some illuminating comments by Dr. Ray Mickey and Dr. Harry Goheen.

REFERENCES

[1] A. S. Krishnamoorthy and M. Parthasarathy, "A multivariate gamma-type distribution", *Ann. of Math. Stat.*, Vol. 22 (1951), pp. 549–557.

[2] G. A. Campbell and R. M. Foster, "Fourier integrals for practical applications", *Bell Telephone Technical Publications*, 1942.

[3] H. B. Mann and A. Wald, "On stochastic limit and order relationships", *Ann. of Math. Stat.*, Vol. 14 (1943), pp. 217–226.

[4] R. A. Fisher and L. H. C. Tippett, "Limiting forms of the frequency distribution of the largest or smallest member of a sample", *Proc. Cambridge Philos. Soc.*, Vol. 24 (1928), pp. 180–190.

[5] H. Cramér, *Mathematical Methods of Statistics*, Princeton University Press, 1946.

[6] *Tables of the Confluent Hypergeometric Function $F(n/2, \frac{1}{2}, x)$ and Related Functions*, National Bureau of Standards Applied Mathematics Series, No. 3, U. S. Government Printing Office, Washington, D. C.

[7] E. W. Barnes, "On functions defined by simple types of hypergeometric series", *Trans. Cambridge Philos. Soc.*, Vol. 20 (1905), pp. 253–280.