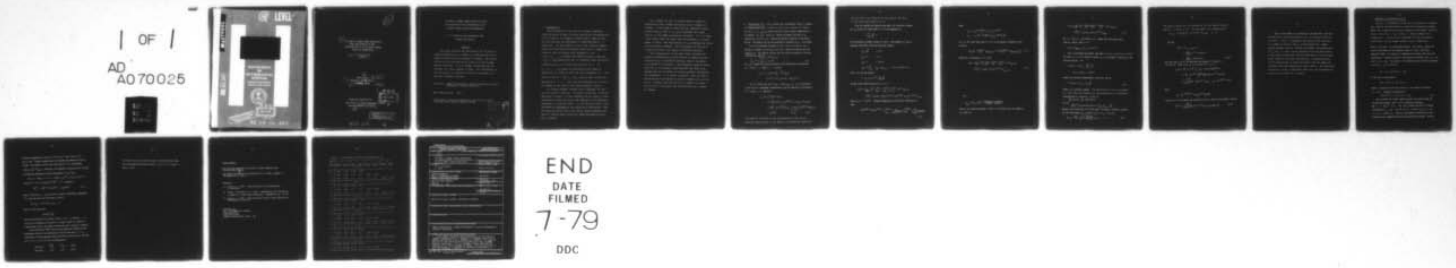
AD-A070 025 CLEMSON UNIV S C DEPT OF MATHEMATICAL SCIENCES F/G 12/1
ON SUMS OF GAMMA ORDER STATISTICS WITH APPLICATION TO THE DISTR--ETC(U)
SEP 78 K T WALLENIUS, K ALAM N00014-75-C-0451
UNCLASSIFIED N101 NL

| OF |
AD
A070025



END
DATE
FILMED
7-79
DDC

12 LEVEL II

DEPARTMENT OF MATHEMATICAL SCIENCES

CLEMSON UNIVERSITY
Clemson, South Carolina

(12)

(6) ON SUMS OF GAMMA ORDER STATISTICS WITH APPLICATION TO THE DISTRIBUTION OF R^2 IN BEST SUBSET MULTIPLE REGRESSION.

(10) K. T./Wallenius
K./Alam

(14) Report N101, TR-305
(9) Technical Report, #305
(11) September 1978
(12) 17p.

Research supported by
THE OFFICE OF NAVAL RESEARCH
Contract N00014-75-C-0451
(15)

ON SUMS OF GAMMA ORDER STATISTICS WITH APPLICATION TO THE DISTRIBUTION OF R^2 IN BEST SUBSET MULTIPLE REGRESSION

K. T. Wallenius and Khursheed Alam

Clemson University

ABSTRACT

This paper concerns the distribution of the sum of k largest observations in a sample of m observations from a gamma distribution with n degrees of freedom. The density and cdf of the distribution are given as a sum of gamma density functions. If n is integer valued then the sum consists of a finite number of terms. The distribution of the sum arises in a problem of selecting variables in a multiple regression analysis.

Key words: Gamma Distribution; Laplace Transform; Multiple Regression; Variable Selection.

AMS Classification: 62E15

*The authors' work was supported by The Office of Naval Research under Contract N00014-75-0451.

1. Introduction.

The distribution of the sum of k largest (smallest) order statistics arises in various statistical investigations. In life testing, for example, suppose that m items are put on trial and that the experiment is terminated when $k < m$ items fail. Let the length of life of the items be independently and identically distributed according to a gamma distribution with an unknown scale parameter θ, say. If $X_1 \leq X_2 \leq \ldots \leq X_k$ denote the observed failure times then $\sum_{i=1}^{k} X_i$, representing the sum of k smallest order statistics, together with X_k is a sufficient statistic for θ. If $Z_1 \leq Z_2 \leq \ldots \leq Z_k$ denote the observed failure times for another set of items for which the scale parameter is θ', say, then the ratio $R = \sum_{i=1}^{k} X_i / \sum_{i=1}^{k} Z_i$ may be used to test the hypothesis $H: \theta = \theta'$. Note that the distribution of R does not depend on the value of the scale parameter, under H.

For another example, suppose that m customers are waiting in a queue for service. In certain situations, it may be desirable to divert k of the customers who are likely to take individually longer servicing time compared to the remaining customers, to a special queue. Let the servicing time of the customers be independently and identically distributed. Then the total servicing time of the special queue represents the sum of k largest order statistics, whose distribution would be of interest.

Let Y_k denote the sum of k largest values in a sample of n observations from a gamma distribution with n degrees of freedom. In this paper we show that the cumulative distribution function (cdf) of Y_k can be expressed as a linear function of the gamma distribution functions. If n is a positive integer then the linear function consists of a finite number of terms. The distribution of Y_k is obtained by inverting its Laplace transform. The distribution of the sum of k smallest values in the sample is obtained similarly.

For an application of the given result consider the problem of determining the distribution of sample multiple correlation R_k in regression analysis, where k variables are selected for inclusion in the regression equation from a given set of m variables, which maximize the value of R_k. Suppose that the variables are jointly normally distributed and independent. It is shown in Section 3 that $(M-1)R_k^2$ is asymptotically distributed for large M (sample size) as the sum of k largest order statistics in a sample of m observations from a chi-square (χ_1^2) distribution with 1 degree of freedom.

2. Distribution of Y_k. Let X_r denote the r-th smallest value in a sample of m observations from a gamma distribution with n degrees of freedom, and let $Y_k = \sum_{r=m-k+1}^{m} X_r$ denote the sum of the m largest observations in the sample. Let $f_k(x)$ and $F_k(x)$ denote the density and cdf of Y_k, respectively, and let $L_k(\theta)$ denote the Laplace transform of the distribution. The density and cdf are obtained by inverting $L_k(\theta)$, as follows.

Let Y be distributed according to the gamma distribution with n degrees of freedom, and let $g_n(x)$ and $G_n(x)$ denote its density and cdf, respectively. The density function and the Laplace transform of the distribution are given by

$$g_n(x) = x^{n-1}e^{-x}/\Gamma(n), \qquad x > 0$$
$$\int_0^\infty e^{-\theta x}\, d\, G_n(x) = (1+\theta)^{-n}, \qquad \theta > 0. \tag{2.1}$$

Let $\phi_x(\theta)$ denote the Laplace transform of the conditional distribution of Y, given $Y \geq x$, where $x \geq 0$. We have

$$\phi_x(\theta) = (1-G_n(x))^{-1}\int_x^\infty e^{-\theta y}\, dG_n(y)$$
$$= (1+\theta)^{-n}\,(1-G_n((1+\theta)x))\,(1-G_n(x))^{-1}.$$

Let H(x) denote the cdf of X_{m-k}. Given $X_{m-k} = x$, Y_k is distributed as the sum of k independent observations from the conditional distribution of Y, given $Y \geq x$. Therefore

$$L_k(\theta) = \int_0^\infty \phi_x^k(\theta)\, d\, H(x)$$
$$= m\binom{m-1}{k}\int_0^\infty \phi_x^k(\theta)\, G_n^{m-k-1}(x)\,(1-G_n(x))^k d\, G_n(x)$$
$$= \begin{cases} m\binom{m-1}{k}(1+\theta)^{-nk}\int_0^\infty (1-G_n((1+\theta)x))^k G_n^{m-k-1}(x)\, dG_n(x) & 1 \leq k < m \\ (1+\theta)^{-mn} & k = m. \end{cases} \tag{2.2}$$

The Laplace transform of the distribution of the sum of k smallest observations in the sample is obtained by substitut-

ing $G_n((1+\theta)x)$ for $1-G_n((1+\theta)x)$ and $1-G_n(x)$ for $G_n(x)$ in the right hand side of (2.2).

First we consider the special case when n is a positive integer. Let c_{uv} denote the coefficient of x^u in the expansion of

$$(\Sigma_{\alpha=0}^{n-1} \frac{x^\alpha}{\alpha!})^v$$

for non-regative integer values of u and v. The numbers c_{uv} can be computed recursively from the following formula.

$$c_{uv} = \frac{v^u}{u!} , \qquad u \leq n-1$$

$$c_{ul} = o , \qquad u \geq n$$

$$c_{uv} = o , \qquad u > (n-1)v$$

$$c_{uv} = \Sigma_{\alpha=o}^{n-1} \frac{1}{\alpha!} c_{u-\alpha,v-1} , \qquad n \leq u \leq (n-1)v, \ v>1.$$

From (2.2) and the formula

$$1 - G_n(x) = e^{-x} \Sigma_{\alpha=o}^{n-1} \frac{x^\alpha}{\alpha!}$$

we have after simplification for $1 \leq k < m$

$$L_k(\theta) = \frac{m}{\Gamma(n)} \binom{m-1}{k} \Sigma_{r=o}^{m-k-1} \Sigma_{u=o}^{(n-1)k} \Sigma_{v=o}^{(n-1)r} (-1)^r c_{uk} c_{vr}$$

$$\binom{m-k-1}{r} \Gamma(u+v+n) (k+1+r)^{-u-v-n} (1+\theta)^{-nk+u} (1+\alpha_r\theta)^{-u-v-n} \qquad (2.3)$$

where $\alpha_r = k / (1+r+k)$. Through decomposition into partial fractions we have

$$(1+\theta)^{-nk+u} (1+\alpha_r\theta)^{-u-v-n} = \Sigma_{s=o}^{nk-u-1} \frac{a_s}{(1+\theta)^{nk-u-s}} + \Sigma_{s=o}^{u+v+n-1} \frac{b_s}{(1+\alpha_r\theta)^{u+v+n-s}}$$

(2.4)

where

$$a_s = (-\alpha_r)^s \binom{u+v+n+s-1}{s} (1-\alpha_r)^{-u-v-n-s}$$

$$b_s = (-1)^{nk-u} \binom{nk-u+s-1}{s} \left(\frac{1}{\alpha_r} - 1\right)^{-nk+u-s} \alpha_r^{-s}.$$

By (2.1) the right hand side of (2.4) is the Laplace transform of the function.

$$g^*_{ruv}(x) = \Sigma_{s=0}^{nk-u-1} a_s g_{nk-u-s}(x) + \alpha_r^{-1} \Sigma_{s=0}^{u+v+n-1} b_s g_{u+v+n-s}\left(\frac{x}{\alpha_r}\right). \quad (2.5)$$

Therefore, by inverting (2.3) we get

$$f_k(x) = \frac{m}{\Gamma(n)} \binom{m-1}{k} \Sigma_{r=0}^{m-k-1} \Sigma_{u=0}^{(n-1)k} \Sigma_{v=0}^{(n-1)r} (-1)^r c_{uk} c_{vr}$$

$$\binom{m-k-1}{r} \Gamma(u+v+n) (k+1+r)^{-u-v-n} g^*_{ruv}(x). \quad (2.6)$$

Let

$$p_{ruv\ell} = \Sigma_{t=0}^{\ell} \binom{\ell}{t} \alpha_r^t \frac{\Gamma(nk-u+\ell-t)}{\Gamma(nk-u)} \frac{\Gamma(n+u+v+t)}{\Gamma(n+u+v)}$$

where ℓ is a positive integer. From (2.3) we obtain the ℓ-th moment of Y_k, given by

$$E\, Y_k^{\ell} = \frac{m}{\Gamma(n)} \binom{m-1}{k} \sum_{r=0}^{m-k-1} \sum_{u=0}^{(n-1)k} \sum_{v=0}^{(n-1)r} (-1)^r c_{uk}\, c_{vr}$$

$$\binom{m-k-1}{r} \Gamma(u+v+n)\, (k+1+r)^{-u-v-n}\, p_{ruv\ell}\,. \qquad (2.7)$$

For $n = 1$ and $\ell = 1$, the formula (2.7) checks with the known result (see e.g. David (1970) 2.7.3)

$$E\, Y_k = \sum_{i=m-k+1}^{m} \sum_{j=1}^{i} (m-j+1)^{-1}.$$

Next, we consider the general case when n is not a positive integer. The case $n = 1/2$ is of special interest, as in the example, described in the previous section. Let

$$\Phi(a,b;x) = 1 + \sum_{r=1}^{\infty} \frac{(a)_r}{(b)_r} \frac{x^r}{r!}$$

$$(a)_r = a(a+1)\ldots(a+r-1)$$

denote the confluent hypergeometric function, and let

$$\Phi^s(a,b;x) = \sum_{r=0}^{\infty} d_{rs} \frac{x^r}{r!} \qquad (2.8)$$

where s is a positive integer. Differentiating (2.8) with respect to x and using the formula for the derivative of a confluent hypergeometric function, given by

$$\frac{d}{dx} \Phi(a,b;x) = \frac{a}{b}\, \Phi(a+1,b+1;x)$$

we get

$$\frac{sa}{b}\, \Phi^{s-1}(a,b;x)\Phi(a+1,b+1;x) = \sum_{r=0}^{\infty} d_{r+1\,s} \frac{x^r}{r!}\,.$$

Equating the coefficient of x^r both sides we obtain a recursive relation for the coefficients d_{rs}, given by $d_{0s} = 1$, $d_{r1} = (a)_r/(b)_r$

$$d_{r+1\,s} = \frac{sa}{b} \sum_{t=0}^{r} \binom{r}{t} \frac{(a+1)_{r-t}}{(b+1)_{r-t}} d_{t\,s-1}, \quad s > 1 \qquad (2.9)$$

The above formula will be used below for the special case in which $a = n$ and $b = n+1$. In this case (2.9) simplifies to

$$d_{r+1\ s} = ns\sum_{t=0}^{r}\binom{r}{t}\ (n+1+r-t)^{-1}d_{t\ s-1}, \quad s > 1.$$

We have

$$\begin{aligned} G_n(x) &= \sum_{r=1}^{\infty} g_{n+r}(x) \\ &= \frac{x^n}{\Gamma(n+1)}\, e^{-x}\ \Phi(1,n+1;x). \\ &= \frac{x^n}{\Gamma(n+1)}\ \Phi(n,n+1;-x). \end{aligned} \tag{2.10}$$

The last step in (2.10) follows from the relation $\Phi(b-a,b;-x) = e^{-x}\Phi(a,b;x)$. Using (2.10) in (2.2) we get for $1 \le k < m$

$$\begin{aligned} L_k(\theta) &= \binom{m}{k}\ (1+\theta)^{-nk}\int_0^{\infty}(1-G_n((1+\theta)x))^k dG_n^{m-k}(x) \\ &= k\binom{m}{k}\ (1+\theta)^{-nk}\int_0^{\infty} G_n^{m-k}\left(\frac{x}{1+\theta}\right)(1-G_n(x))^{k-1} g_n(x)\,dx \\ &= \frac{k\binom{m}{k}}{(\Gamma(n+1))^{m-k}}\ \sum_{r=0}^{\infty}(-1)^r d_{r\ m-k}\ \ell_{rk}\ (1+\theta)^{-mn-r} \end{aligned} \tag{2.11}$$

where

$$\ell_{rk} = \int_0^{\infty} x^{n(m-k)+r}\ (1-G_n(x))^{k-1} g_n(x)\,dx.$$

Inverting (2.11) we obtain the distribution of Y_k, given by the density function

$$f_k(x) = \frac{k\binom{m}{k}}{(\Gamma(n+1))^{m-k}}\ \sum_{r=0}^{\infty}(-1)^r d_{r\ m-k}\ \ell_{rk}\ g_{mn+r}(x). \tag{2.12}$$

Table I below shows, for illustration, the upper 90th and 95th percentiles of the distribution of Y_k for certain values of m,k and n. Since Y_1 represents the largest order statistic in a sample of size m, and Y_m is distributed as a gamma random variable with mn degrees of freedom, the percentiles of Y_k for k = 1 and m can be obtained from available tables of the gamma distribution and the distribution of the largest order statistic from that distribution. Percentage points of the distribution of order statistics from the gamma distribution have been tabulated by Gupta (1960). The percentage points of Y_1 given in the table agree with the corresponding points give in Table III of Gupta.

3. Asymptotic Distribution of R_k.

Let $X_1,\ldots,X_k$ denote a given set of m predictor variables and Y denote the predictand in a multiple regression problem, where the variables are jointly normally distributed. Specifically, let $(Y,X_1,\ldots,X_m)' \underset{\sim}{d} N(\mu,\Sigma)$, where $\underset{\sim}{d}$ means "distributed as". Since we are concerned with the correlation coefficient, we can assume without loss of generality

that $\underset{\sim}{\mu} = \underset{\sim}{0}$ and that Σ is a correlation matrix. Let $\underset{\sim}{Y}$ and $\underset{\sim}{X}_i$ denote the vector of deviations of the observed values of Y and X_i from their respective means in a sample of M observations obtained from the given distribution. Consider a subset of the predictor variables, say, $X_1,\ldots,X_k$. Let $\underset{\sim}{X} = (\underset{\sim}{X}_1,\ldots,\underset{\sim}{X}_k)$. The square of the sample multiple correlation between Y and $(X_1,\ldots,X_k)$ is given by

$$R^2 = (\underset{\sim}{Y}' \; \underset{\sim}{X} \; (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{Y}) \; / \; (\underset{\sim}{Y}'\underset{\sim}{Y}).$$

By the law of large numbers

$$(M-1) \; (\underset{\sim}{X}'\underset{\sim}{X})^{-1} \xrightarrow{p} \Sigma_1^{-1} \quad \text{as } M \to \infty$$

where Σ_1 denotes the correlation matrix of the predictor variables $X_1,\ldots,X_k$. Therefore, asymptotically

$$(M-1)R^2 \underset{\approx}{d} (\underset{\sim}{Y}' \; \underset{\sim}{X} \; \Sigma_1^{-1} \; \underset{\sim}{X}' \; \underset{\sim}{Y}) \; / \; (\underset{\sim}{Y}'\underset{\sim}{Y}) \qquad (3.1)$$

Let R_k denote the largest sample multiple co-relation among $\binom{m}{k}$ correlations between Y and k out of m predictor variables.
Let $V_i = (\underset{\sim}{Y}'\underset{\sim}{X}_i)^2 / (\underset{\sim}{Y}'\underset{\sim}{Y})$, and let S_k denote the sum of k largest values among $V_1,\ldots,V_m$. Let $\text{Cor}\,(Y,X_i) = \rho$ and $\text{Cor}\,(X_i,X_j) = \lambda$ for $i, j = 1,\ldots,m$ and $i \neq j$. That is, the predictor variables are equi-correlated with themselves and with the predictor variables. The two

distinct characteristic roots of Σ_1^{-1} are $(1-\lambda)^{-1}$ and $(1+(k-1)\lambda)^{-1}$. Let λ_* and λ^* denote, respectively, the minimum and maximum of the two values. The quantity on the right hand side of (3.1) lies between $\lambda_* \Sigma_{i=1}^{k} V_i$ and $\lambda^* \Sigma_{i=1}^{k} V_i$. Therefore, the asymptotic distribution of $(M-1)R_k^2$ is minorized (majorized) by the distribution of $\lambda_* S_k (\lambda^* S_k)$.

Let $\lambda = o$. Then $\lambda_* = \lambda^* = 1$. Given $\underset{\sim}{Y}$, $V_i^{1/2} = (\underset{\sim}{Y}' \underset{\sim}{X}_i)/(\underset{\sim}{Y}' \underset{\sim}{Y})^{1/2} \underset{\approx}{d}$ $N(\rho(\underset{\sim}{Y}' \underset{\sim}{Y})^{1/2}, 1-\rho^2)$ and $\mathrm{Cor}(V_i^{1/2}, V_j^{1/2}) = -\rho^2$. Therefore

$$(V_1^{1/2}, \ldots, V_m^{1/2}) \underset{\approx}{d} (U_1 + \rho W^{1/2}, \ldots, U_m + \rho W^{1/2}) \qquad (3.2)$$

where $W \underset{\approx}{d} \chi_m^2$ and $U_1, \ldots, U_m$ are jointly normally distributed independent of W, with mean zero and covariance, given by

$$\mathrm{Var}\ (U_i) = 1-\rho^2, \quad \mathrm{Cov}\ (U_i, U_j) = -\rho^2.$$

Thus for large M we have

$$(M-1)R_k^2 \underset{\approx}{d} S_k^2$$

with the distribution of S_k being given by (3.2). If moreover, $\rho = o$ then S_k^2 is distributed as the sum of k largest values in a sample of m observations from a chi-square distribution with 1 degree of freedom.

Diehr and Hoflin (1974) have given an empirical formula for the percentage points of the distribution of R_k^2 for the case $\lambda = \rho = o$. From Table 1 of their paper we obtain the 90% and 95% points of $(M-1)R_k^2$ for m = 5, k = 1,2,3, M = 106, as shown below

	k=1	2	3
90% point	5.25	7.56	8.50
95% point	6.51	9.14	10.18

The above figures are slightly larger (as they should be) than the corresponding percentage points of $2Y_k$ for $m = 4$, given in Table 1 below.

Acknowledgement

This work was supported by the Office of Naval Research under Contract N00014-75-C-0451.

The authors acknowledge the contribution of Dr. James S. Hawkes to the preparation of Table 1.

References

[1] David, H. A. (1970). Order Statistics, Wiley Publications in Statistics.

[2] Diehr, G. and Hoflin, D. R. (1974). Approximating the distribution of sample R^2 in best subset regressions. Technometrics 16, 317-320.

[3] Gupta, S. S. (1960). Order statistics from the gamma distribution. Technometrics 2, 243-262.

Khursheed Alam
Dept. of Mathematical Sciences
0-104 Martin Hall
Clemson University
Clemson, South Carolina 29631, USA

Table 1 - Percentiles of the distribution of Y_k

k	1		2		3		4	
	90%	95%	90%	95%	90%	95%	90%	95%
	n = ½							
m=2	1.90	2.51	2.30	3.00				
3	2.25	2.85	2.9	3.6	3.13	3.91		
4	2.46	3.09	3.3	4.0	3.7	4.7	3.89	4.74
	n = 1							
2	2.97	3.68	3.89	4.74				
3	3.37	4.08	4.7	5.6	5.33	6.30		
4	3.65	4.36	5.3	6.2	6.2	7.2	6.68	7.75
	n = 2							
2	4.71	5.56	6.68	7.75				
3	5.19	6.03	7.8	8.9	9.27	10.51		
4	5.53	6.36	8.6	9.7	10.6	11.8	11.77	13.15
	n = 3							
2	6.26	7.21	9.27	10.51				
3	6.80	7.73	10.6	12.0	12.99	14.43		
4	7.17	8.10	11.5	12.7	14.6	16.0	16.60	18.21
	n = 4							
2	7.71	8.75	11.77	13.15				
3	8.30	9.32	13.3	14.7	16.60	18.21		
4	8.71	9.71	16.3	15.6	18.4	20.1	21.30	23.10
	n = 5							
2	9.11	10.22	14.21	15.71				
3	9.74	10.83	15.9	17.4	20.13	21.89		

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER N101	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) On Sums of Gamma Order Statistics with Application to the Distribution of R^2 in Best Subset Multiple Regression.		5. TYPE OF REPORT & PERIOD COVERED 6. PERFORMING ORG. REPORT NUMBER Tech. Report #305
7. AUTHOR(s) K. T. Wallenius K. Alam		8. CONTRACT OR GRANT NUMBER(s) N00014-75-C-0451
9. PERFORMING ORGANIZATION NAME AND ADDRESS Clemson University Dept. of Mathematical Sciences Clemson, South Carolina 29631		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR 042-271
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Code 436 Arlington, Va. 22217		12. REPORT DATE September 1978 13. NUMBER OF PAGES 12
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Gamma Distribution; Laplace Transform; Multiple Regression; Variable Selection.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper concerns the distribution of the sum of k largest observations in a sample of m observations from a gamma distribution with n degrees of freedom. The density and cdf of the distribution are given as a sum of gamma density functions. If n is integer valued then the sum consists of a finite number of terms. The distribution of the sum arises in a problem of selecting variables in a multiple regression analysis.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)