AD-A048 127   CLEMSON UNIV S C DEPT OF MATHEMATICAL SCIENCES   F/G 12/1

DISTRIBUTION OF A SUM OF ORDER STATISTICS FROM THE GAMMA DISTRI--ETC(U)

SEP 77 K T WALLENIUS, K ALAM   N00014-75-C-0451

UNCLASSIFIED   N77   NL

1 OF 1

ADA048 127

END

DATE FILMED

1-78

DDC

AD A048127

DDC
RECEIVED
JAN 3 1978
B

DISTRIBUTION OF A SUM OF ORDER STATISTICS FROM THE GAMMA DISTRIBUTION

K. T. WALLENIUS
AND
KHURSHEED ALAM

TECHNICAL REPORT #262

SEPTEMBER, 1977

D D C
RECEIVED
JAN 5 1978
B

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

ACCESSION for
NTIS White Section
DDC Buff Section
UNANNOUNCED
JUSTIFICATION
BY
DISTRIBUTION/AVAILABILITY CODES
Dist. AVAIL. and/or SPECIAL
A

DISTRIBUTION OF A SUM OF ORDER STATISTICS FROM THE GAMMA DISTRIBUTION

K. T. Wallenius & 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. If n is an integer, the density and cdf of the distribution are given as a linear function of gamma density functions. If n is not an integer, an approximate distribution of the same form is obtained. The distribution of the sum arises in a problem of selecting variables in a multiple regression analysis.

Key words: Gamma Distribution; Laplace Transform; Linear Regression.

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

C

1. Main result. Let X_r denote the r-th smallest observation in a smaple of size m 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 k largest observations in the sample. First we obtain the Laplace transform of the distribution of Y_k. By inverting the transform we derive the density and cdf of the distribution. If n is a positive integer then the density and the cdf of Y_k will be given as a linear function of gamma density functions.

Let $g_n(x)$ and $G_n(x)$ denote the density and cdf, respectively, of the gamma distribution with n degrees of freedom. The Laplace transform of the distribution is given by

$$\int_0^\infty e^{-\theta x}\, d\, G_n(x) = (1+\theta)^{-n}\ , \quad \theta > 0\ .$$

If Y is distributed according to the gamma distribution, the Laplace transform of the conditional distribution of Y, given $Y \geq x$, is given by

$$\phi_x(\theta) = (1-G_n(x))^{-1} \int_x^\infty e^{-\theta y}\, d\, G_n\ (y)$$

$$= (1+\theta)^{-n}\ (1-G_n((1+\theta)x))\ (1-G_n(x))^{-1}\ .$$

Let $\phi_k(\theta)$ denote the Laplace transform of Y_k, and 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

$$(1.1)\quad L_k(\theta) = \int_0^\infty \phi_x{}^k(\theta)\, dH(x)$$

$$= m\binom{m-1}{k} \int_0^\infty \phi_x{}^k(\theta)\, G_n{}^{m-k-1}(x)(1-G_n(x))^k dG_n(x)$$

$$= \begin{cases} m\binom{m-1}{k} (1+\theta)^{-nk} \int_0^\infty (1-G_n((1+\theta)x))^k G_n{}^{m-k-1}(x)dG_n(x) & 1 \le k < m \\ (1+\theta)^{-nm} & k = m\,. \end{cases}$$

Let n be a positive integer. Integrating by parts we get

$$(1.2)\qquad 1 - G_n(x) = g_1(x) + g_2(x) + \ldots + g_n(x)$$

$$= e^{-x} \sum_{\alpha=0}^{n-1} \frac{x^\alpha}{\alpha!}\ .$$

Let c_{uv} denote the coefficient of x^u in the expansion of

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

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

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

$$c_{u1} = 0\ ,\qquad u \ge n$$

$$c_{uv} = 0 \quad , \quad u>(n-1)v$$

$$c_{uv} = \sum_{\alpha=0}^{n-1} \frac{1}{\alpha!} c_{u-\alpha\ v-1} \ , \ n \le u \le (n-1)v, \ v>1.$$

From (1.1), using (1.2), we have after simplification

$$(1.3) \qquad L_k(\theta) = \frac{m}{\Gamma(n)} \binom{m-1}{k} \sum_{r=0}^{m-k-1} \sum_{u=0}^{(n-1)k} \sum_{v=0}^{(n-1)r} (-1)^r$$

$$c_{uk} c_{vr} (k+1+r)^{-u-v-n}\, \Gamma(u+v+n) \binom{m-k-1}{r}$$

$$(1+\theta)^{-nk+u} (1+\alpha_r \theta)^{-u-v-n}$$

for $1 \le k < m$, where $\alpha_r = k/(1+r+k)$.

Let $W = U + \alpha_r V$, where U and V are random variables independently distributed according to the gamma distribution with $nk - u$ and $n + u + v$ degrees of freedom, respectively. The cdf of W is given by

$$(1.4) H_{ruv}(x) = \int_0^{x/\alpha_r} G_{nk-u}(x-\alpha_r y)\ g_{n+u+v}(y)dy$$

$$= G_{n+u+v}(x/\alpha_r) - \sum_{s=1}^{nk-u} \int_0^{x/\alpha_r}$$

$$g_s(x-\alpha_r y) g_{n+u+v}(y) dy$$

$$= G_{n+u+v}(x/\alpha_r) - \sum_{s=1}^{nk-u} \sum_{t=0}^{s-1} \frac{(-\alpha_r)^t\ \Gamma(n+u+v+t)}{t!\ \Gamma(n+u+v)}$$

$$(1-\alpha_r)^{-n-u-v-t}\ G_{n+u+v+t} \left(\frac{1}{\alpha_r} - 1\right)x\ \ g_{s-t}(x) \ .$$

Using (1.2) we see that $H(x)$ is given as a linear function of the gamma density functions.

The Laplace transform of the distribution of W is equal to $(1+\theta)^{-nk}(1+\alpha_r\theta)^{-n-u-v}$. Therefore, from (1.3) we obtain by inversion the cdf of Y_k, given by

$$(1.5)\quad F_k(x) = \frac{m}{\Gamma(n)}\binom{m-1}{k}\sum_{r=0}^{m-k-1}\sum_{u=0}^{(n-1)k}\sum_{v=0}^{(n-1)r}(-1)^r\binom{m-k-1}{r}$$

$$\Gamma(u+v+n)(k+1+r)^{-u-v-n}\, c_{uk}\, c_{vr}\, H_{ruv}(x), 1\leq k<m\ .$$

For $k = m$ we have $F_m(x) = G_{nm}(x)$. Thus $F_k(x)$ is given as a linear function of the gamma density functions. By differentiation we obtain the density function of Y_k of the same form.

From (1.3) we obtain the ℓ-th moment of Y_k, given by

$$(1.6)\quad \mu_k^{\ell} = \frac{m}{\Gamma(n)}\binom{m-1}{k}\sum_{r=0}^{m-k-1}\sum_{u=0}^{(n-1)k}\sum_{v=0}^{(n-1)r}(-1)^r\binom{m-k-1}{r}$$

$$\Gamma(u+v+n)(k+1+r)^{-u-v-n}\, c_{uk}\, c_{vr}\, E\,(W^{\ell})$$

where

$$E(W^{\ell}) = \sum_{t=0}^{\ell}\binom{\ell}{t}\alpha_r^t\ \frac{\Gamma(nk-u+\ell-t)\ \Gamma(n+u+v+t)}{\Gamma(nk-u)\ \Gamma(n+u+v)}\ .$$

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

$$(1.7)\qquad E(Y_k) = \sum_{i=m-k+1}^{m}\sum_{j=1}^{i}(m-j+1)^{-1}\ .$$

Now we consider the case in which n is not an integer. Let $n = n^* + \nu$, where n^* denotes the integral part of n and $0 < \nu < 1$. For any positive integer $t > n^*$, let

$$(1.8) \qquad A_t(x) = \sum_{r=n^*+1}^{t} g_{r+\nu}(x) - \sum_{r=1}^{t-1} g_r(x) + 1$$

$$= G_n(x) - G_{t+\nu}(x) + G_{t+1}(x) .$$

We show that $A_t(x)$ is a probability distribution function for sufficiently large values of t. The derivative of $A_t(x)$ with respect to x is given by

$$(1.9) \quad A_t'(x) = g_n(x) - g_{t+\nu}(x) + g_{t+1}(x)$$

$$= x^{n-1} e^{-x} \left[\frac{1}{\Gamma(n)} - \frac{x^{t-n^*}}{\Gamma(t+\nu)} + \frac{x^{t-n+1}}{\Gamma(t+1)}\right] .$$

Let $P_t(x)$ denote the quantity inside the square bracket on the right side of (1.9). The derivative of $P_t(x)$ with respect to x changes sign from negative for positive as x varies from 0 to ∞. Hence $P_t(x)$ is minimized for $x = x_0$, say, given by

$$x_0^{1-\nu} = \frac{(t-n^*)\,\Gamma(t+1)}{(t-n+1)\,\Gamma(t+\nu)}$$

$$\simeq t^{1-\nu} e^{\nu-1} \quad \text{for large t.}$$

We have

$$(1.10) \qquad P_t(x_0) = \frac{1}{\Gamma(n)} - \frac{(1-\nu)x_0^{t-n^*}}{(t-n+1)\,\Gamma(t+\nu)}$$

$$= \frac{1}{\Gamma(n)} - \frac{(1-\nu)(t-n+1)^{-1}}{(t+\nu)}$$

$$\left(\frac{(t-n^*)\ \Gamma(t+1)}{(t-n+1)\quad \Gamma(t+\nu)}\right)^{\frac{t-n^*}{1-\nu}}$$

$$\simeq \frac{1}{\Gamma(n)} - \frac{1-\nu}{\sqrt{2\pi}}\, e^n\ t^{-n-\frac{1}{2}} > 0 \qquad \text{for large } t.$$

Therefore, there exists a vâlue of $t = t_0$, say, depending on n such that $P_t(x) > 0$ for all x and $t \geq t_0$. Since $A_t(0) = 0$ and $A_t(\infty) = 1$, it follows that $A_t(x)$ is a probability distribution function on $[0,\infty)$ for $t \geq t_0$.

Since $G_{t+\nu}(x) \to G_{t+\nu}(x)$ uniformly in x, as $t\to\infty$, it is seen from (1.8) that $G(x) \to A_t(x)$ uniformly in x. Also, $A_t(x) \leq G_n(x)$, since $G_{t+\nu}(x) \geq G_{t+1}(x)$. We have shown the following result.

Theorem 1.1. Let $t > n$ be a positive integer. Then $A_t(x) \leq G_n(x)$ for all $x \geq 0$, and $A_t(x) \to G_n(x)$ uniformly in x, as $t\to\infty$. There exists a value of t depending on n, such that, $A_t(x)$ is a probability distribution function on $[0,\infty)$ for $t \geq t_0$.

The above theorem shows that when n is not an integer we can approximate $G_n(x)$ by the distribution function $A_t(x)$, given as a linear function of gamma density functions $g_r(x)$ where r is integer valued. Therefore, when n is not an integer we approximate the distribution of y_k by the distribution of the sum of k largest order statistics from a sample from the distribution $A_t(x)$. The Laplace transform of the distribution of the sum is obtained by substituting $A_t(x)$ for $G_n(x)$ in (1.1) and is given by

$$(1.11) \qquad L_k^*(\theta) = m\binom{m-1}{k}(1+\theta)^{-nk}\int_0^\infty \Big(\sum_{r=1}^{t+1} g_r((1+\theta)x)$$

$$- \sum_{r=n^*+1}^{t} g_{r+\nu}((1+\theta)x)\Big)^k\Big(1+\sum_{r=n^*+1}^{t} g_{r+\nu}(x)\Big)$$

$$- \sum_{r=1}^{t+1} g_r(x)\Big)^{m-k-1} g_n(x)\, dx .$$

The right side of (1.11) is seen after simplification to be of the same form as (1.3). Inverting the transform we obtain the distribution function in the same form as (1.5).

If a few parameters of the distribution of Y_k are required, such as the quantiles , as in the application considered below, where n is not an integer, an alternative method is to interpolate from the corresponding values given for adjacent integer values of n.

Table I below shows the 90% and 95% upper points of the distribution of Y_k for certain values of k,m and n. The figures given in the table for n = ½ were obtained by the Monte-Carlo method. The table is not comprehensive. It is given only for illustration.

2. <u>Application</u>. The problem of selecting a subset of independent or predictor variables in regression analysis has been of long interest to applied statisticians, and because of the current availability of high-speed computation facility, this problem has received added attention in the recent statistical literature, Recently, Hocking (1976) has published an expository paper on the subject wherein he has described various aspects of the problem. The paper includes an extensive list of references to important publications in the area.

The following situation arises in a problem of selecting a subset from a given set of predictor variables in multiple regression analysis. There are given m predictor variables $X_1, \ldots, X_m$ and a dependent variable Y. The predictor variables and the dependent variable are jointly distributed according to a multivariate normal distribution. It is required to select a subset of k variables from the set of predictor variables which has "most" prediction value. We call it the best subset. There are $\binom{m}{k}$ subsets to choose from. Suppose that the $\binom{m}{k}$ multiple correlations between Y and each subset of the predictor variables are computed from a sample of M observations. Let R_k denote the largest among them. It is a common practice to select the subset associated with R_k as the best subset. The distribution of R_k which is required for a test of significance for example, is mathematically intractable. Theorem 2.1 below shows that if $Y, X_1, \ldots, X_m$ are independently distributed then $(M-1)\ R_k^2$ is asymptotically distributed for large M, as the sum of k largest observations in a sample of m observation from a chi-squared distribution with one degree of freedom (χ_1^2). The distribution of the sum is given by the results of the preceeding section.

Let $\underset{\sim}{Y}$ and $\underset{\sim}{X}_i$ denote the vectors of the deviations of the observed values of Y and X_i from their respective mean values in the sample. Consider a subset of predictor variables, say, $X_1, \ldots, X_k$. Let $X = (\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_k)$. The square of the sample multiple correlation coefficient between Y and $(X_1, \ldots, X_k)$ is given by

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

Let $(Y, X_1, \ldots, X_m) \overset{d}{\sim} N(\underset{\sim}{\mu}, \Sigma)$. Without loss of generality we can assume that $\underset{\sim}{\mu} = \underset{\sim}{0}$ and that Σ is a correlation matrix. Furthermore, suppose that $\Sigma = I$, the identity matrix, that is, the variables are independently distributed. Then, by the law of large numbers

$$(M-1)(X'X)^{-1} \overset{p}{\to} I \text{ as } M \to \infty.$$

Therefore, asymptotically for large M

$$(2.1) \qquad (M-1)R^2 \overset{d}{\sim} \sum_{i=1}^{k} (\underset{\sim}{Y}'X_i)^2/(\underset{\sim}{Y}'\underset{\sim}{Y}) \; .$$

$$\overset{d}{\sim} \sum_{i=1}^{k} V_i$$

where

$$V_i = (\underset{\sim}{Y}'\underset{\sim}{X}_i)^2/(\underset{\sim}{Y}'\underset{\sim}{Y}) \; .$$

Now $V_1, \ldots, V_m$ are random variables independently and identically distributed as χ_1^2. From (2.1) it follows that $(M-1)R_k^2$ asymptotically is distributed as the sum of k largest order statistics in a sample of m observations from χ_1^2 distribution. This result is stated in the following theorem.

Theorem 2.1. If $Y, X_1, \ldots, x_m$ are normally and independently distributed then $(M-1)R_k^2$ is asymptotically distributed as the sum of k largest order statistics in a sample of size m from χ_1^2 distribution.

References

[1] Anderson, T. W. (1957). An Introduction to Multivariate Statistical Analysis, Wiley Publication.

[2] David, H. A. (1970). Order Statistics, Wiley Publication.

[3] Hocking, R. R. (1976). The analysis and selection of variables in linear regression. Biometrics 32, 1-49.

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	1.9	2.4	2.3	3.0				
3	2.2	2.8	2.9	3.6	3.1	3.9		
4	2.3	3.1	3.3	4.0	3.7	4.7	3.9	4.7
				n = 1				
2	2.9	3.6	3.9	4.7				
3	3.3	4.0	4.7	5.6	5.3	6.3		
4	3.6	4.2	5.3	6.2	6.2	7.2	6.7	7.8
				n = 2				
2	4.7	5.5	6.7	7.8				
3	5.1	6.0	7.8	8.9	9.3	10.5		
4	5.5	6.3	8.6	9.7	10.6	11.8	11.8	13.1
				n = 3				
2	6.2	7.2	9.3	10.5				
3	6.7	7.7	10.6	12.0	13.0	14.4		
4	7.1	8.0	11.5	12.7	14.6	16.0	16.6	18.2
				n = 4				
2	7.7	8.7	11.8	13.1				
3	8.3	9.3	13.3	14.7	16.6	18.2		
4	8.7	9.7	16.3	15.6	18.4	20.1	21.2	23.0
				n = 5				
2	9.1	10.2	14.2	15.7				
3	9.7	10.8	15.9	17.4	20.1	21.9		

(14) N77, TR-262

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: N-77
2. GOVT ACCESSION NO.:
3. RECIPIENT'S CATALOG NUMBER:
4. TITLE (and Subtitle): (6) Distribution of a sum of order statistics from the gamma distribution.
5. TYPE OF REPORT & PERIOD COVERED: (9) Technical rept.,
6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): (10) K. T./Wallenius, Khursheed/Alam
8. CONTRACT OR GRANT NUMBER(s): (15) 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: (11) Sep 77
13. NUMBER OF PAGES: (12) 16 p.
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

407 183

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

Gamma Distribution, Laplace Transform, Selection of Variables.

Gm

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

The distribution of the sum of K largest order statistics in a sample from the gamma distribution with n degrees of freedom is expressed as a linear function of gamma density functions if n is integer valued and is approximated by a probability distribution of the same form if n is not integer valued.